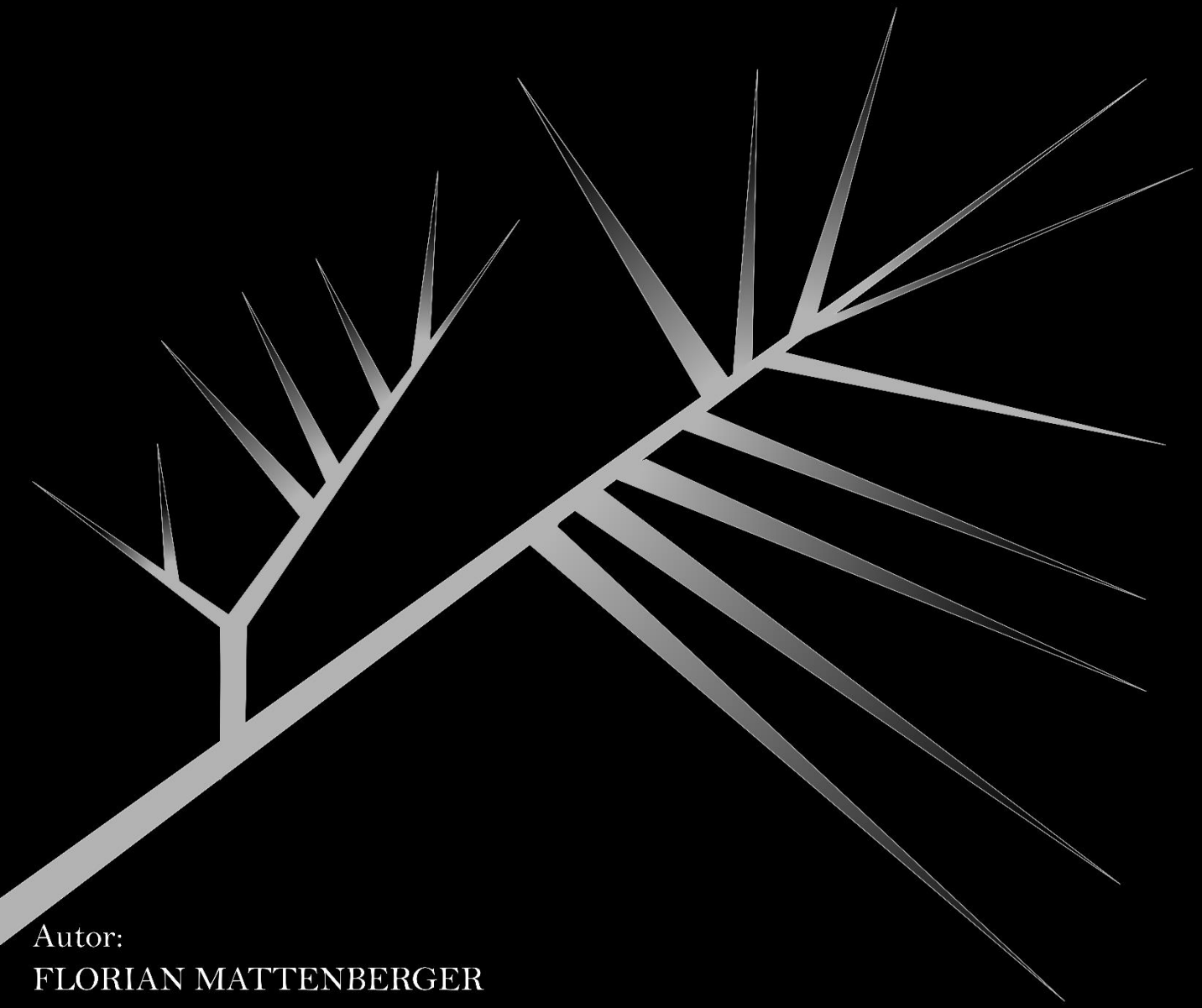
Unveiling adaptative mechanisms through experimental evolution:

The Role of duplicated genes and phenotypic plasticity in yeast, and the genetic variability in Coxsackievirus.



Autor:
FLORIAN MATTENBERGER

Directores:
Beatriz Sabater Muñoz
Ron Geller
Director honorífico:
Mario Ali Fares Riaño

Diciembre 2020

VNIVERSITAT DE VALÈNCIA

PROGRAMA DE DOCTORADO DE BIODIVERSIDAD Y BIOLOGÍA EVOLUTIVA

Unveiling adaptive mechanisms through experimental evolution: the role of duplicated genes and phenotypic plasticity in yeast, and the genetic variability in Coxsackievirus.

Memoria presentada por:

FLORIAN MATTENBERGER

para optar al

GRADO DE DOCTOR

por la Universitat de València, programa de Doctorado en

BIODIVERSIDAD Y BIOLOGÍA EVOLUTIVA

Directores:

Dra. Beatriz Sabater Muñoz

Dr. Ron Geller

Director Honorífico:

Dr. Mario Ali Fares Riaño

Tutor:

Dr. David Martínez Torres

Valencia, Diciembre 2020

Trabajo original presentado por **Florian Mattenberger** para optar al

GRADO DE DOCTOR

por la Universitat de València, programa de Doctorado en

BIODIVERSIDAD Y BIOLOGÍA EVOLUTIVA

MATTENBERGER FLORIAN - X1881214K
Firmado digitalmente por MATTENBERGER FLORIAN - X1881214K
Fecha: 2020.12.17 16:04:47 +01'00'

Fdo: **Florian Mattenberger**

Valencia, Diciembre 2020

La Doctora **Beatriz Sabater Muñoz** y el Doctor **Ron Geller**, como directores, y el Doctor **David Martínez Torres**, como tutor de esta tesis doctoral, autorizan su depósito en la Universitat de Vàlencia y su presentación en el Instituto de Biología Integrativa de Sistemas, instituto mixto CSIC – Universitat de València para su lectura y defensa.

SABATER MUÑOZ BEATRIZ - 09194430L
Digitally signed by SABATER MUÑOZ BEATRIZ - 09194430L
DN: c=ES, serialNumber=IDCES-09194430L, givenName=BEATRIZ, sn=SABATER MUÑOZ, cn=SABATER MUÑOZ BEATRIZ - 09194430L
Date: 2020.12.21 17:18:17 +01'00'

Fdo: Dra. **Beatriz Sabater Muñoz**
Científica titular-CSIC
Directora de Tesis

GELLER --- RON - Y2746411H
Digitally signed by GELLER --- RON - Y2746411H
Date: 2020.12.17 16:13:24 +01'00'

Fdo: Dr. **Ron Geller**
Investigador Ramón y Cajal-UV
Director de Tesis

MARTINEZ TORRES DAVID - 17219363E
Firmado digitalmente por MARTINEZ TORRES DAVID - 17219363E
Fecha: 2020.12.23 14:51:31 +01'00'

Fdo: Dr. **David Martínez Torres**
Profesor Titular-UV
Tutor Tesis

Florian Mattenberger ha disfrutado de un contrato FPI de formación predoctoral (BES-2016-076677) otorgado por el ministerio de Ciencia Innovación y Universidades asociado al proyecto BFU2015-66073-P.

Este trabajo de tesis ha sido realizado con el apoyo económico de los proyectos de investigación BFU2015-66073-P (Dr. Mario Ali Fares) y BFU2017-86094-R (Dr. Ron Geller). Realizando la primera parte de la investigación en Instituto de Biología Molecular y Celular de Plantas (IBMCP) bajo la dirección de los doctores M.A. Fares y B. Sabater-Muñoz, y la segunda parte en el Instituto de Biología Integrativa de Sistemas (I2SysBio), bajo la dirección del Dr. R. Geller.

“Hakuna matata”

The lion king

Agradecimientos

En primer lugar me gustaría recordar a **Mario,** que por desgracia no ha tenido la oportunidad de ver terminada esta tesis. Agradecerle que con él empezó toda esta aventura. Mario fue un gran científico, un gran jefe y sobre todo una gran persona. De él admiraba muchas cosas, pero especialmente su capacidad para pensar nuevos proyectos, nuevas ideas y materializarlas en el laboratorio como si fuese pan comido. Por desgracia nos dejó mucho antes de tiempo, ojalá hubiese podido leer el manuscrito completo, llegar al día de la defensa con un buen traje y una corbata bonita y sentirse orgulloso del trabajo que hemos hecho. También quiero agradecer esta tesis a **Beatriz**, quien también ha sido un pilar fundamental. Gracias a tu paciencia, tu experiencia y a todas tus enseñanzas. Sé que cuando Mario nos dejó tuvimos nuestros más y menos, pero al final estoy muy contento de que hayas sido mi directora y hayamos podido acabar este proyecto juntos. Of course, I also want to thank **Ron**, I met him when I was an undergraduate student at the Cavanilles. Who would have thought that after going hiking together I would end up finishing the thesis in your lab? Thanks for adopt me in your lab, for all your help in the harder times and for your good vibes and wise advices during the last three years. I would also like to thank **Christina** for her help and her wise advice.

También a **David Martínez**, con él comenzaron mis andaduras en la ciencia cuando dirigió mi TFG y mi TFM sobre los pulgones y sus artropsinas, y ha acabado siendo tutor de mi tesis doctoral. Al seu laboratori vaig conèixer a **Quelo**, a qui vull agrair la seua amistat i la seua paciència per a ensenyar a un xiquet com jo el funcionament d'un laboratori. ¡Ara que jo també he tingut alumnes…gracias y lo siento! Jeje. Por aquella época también conocí a **Mariano** ¿Quién no conoce a Mariano? Te agradezco habernos conocido, que me llevases a MusSol y que sigamos siendo amigos a día de hoy. También me gustaría mencionar a **David Coll** y a **Victor Luque** recordando los buenos momentos que hemos pasado en el máster y las ideas locas que nunca hemos llegado a acabar…pero cuando lo hagamos, ¡que tiemble el mundo!

Y como no, también agradecer a mis grandes amigos que he conocido en el IBMCP. **Afri, Pepe, Roser, Ana, Vero, MA, Clara, Pra, Rubén** y **Anamarija** porque el laboratorio nos ha unido y nuestra amistad ha florecido llenando esta etapa de mi vida de momentos muy bonitos y muy buenos recuerdos, pero también se ha convertido en un muro de contención para los malos momentos, que con vosotros han sido menos malos. Gracias a cada uno de vosotros por formar parte de mi vida, y que sepáis que una parte de esta tesis es también es vuestra. Por supuesto cuando hablamos de amistades que se forjan a golpe de pipeta no quisiera olvidarme de las dos chicas más "chicharacheras" del I2SysBio, **María Adelaida Concepción** y **Alejandra Fernanda Larrieux**, y también del gran señor **Dr. Víctor Oswaldo Latorre**. Gracias a los tres por los buenos y grandes momentos que hemos vivido juntos y que han hecho mucho más divertidos estos últimos años. ¡¡Y chicas, muchos ánimos y mucha suerte con lo que os queda!! De esos buenos momentos también forman parte **María D., Layla e Inés,** que con su borderia y su sarcasmo nos han alegrado muchos momentos de desesperación y de desahogo. También agradecer a todos mis compañeros del I2SysBio por los momentos vividos entre pipetas, en el P2, en el comedor o donde fuese que hayamos coincidido. Finalmente, a mis alumnas **Marina (la otra)** y **Cristina**. Marina, gracias por el buen rollo y los golpes de kung-fu que trajiste cuando viniste a hacer tu TFM con nosotros, mucha suerte y mucho ánimo con tu tesis y ¡a seguir partiendo tablones de madera con la mano! Cristina, m'alegre molt que hages vingut al nostre laboratori a fer les practiques i et desitge molta sort, estic segur que tindràs un futur brillant.

Pero cuando hablo de amigos, como no hablar de esas personas que siempre han estado ahí contra viento y marea. **Nuel, Vane** y **Pedrito** esta tesis es tanto mía como vuestra, vosotros me habéis visto crecer como persona y siempre habéis estado conmigo a pesar de la distancia. De verdad que muchas gracias, estar con vosotros ha sido siempre una vía de escape de los problemas y una lección de vida para ser feliz. ¡¡Brindo por muchos más momentos a vuestro lado!!

También mencionar a todo **MusSol**, especialmente a **Félix, Jordi** y a **Javi Tramoya** (y a Vero y a Mariano, que repiten mención especial), que con vosotros he vivido muchas risas encima y debajo del escenario, y espero que sean muchas más.

Vielen Dank auch an meine Eltern, **Peter** und **Cecile**, für eure Unterstützung und euer Vertrauen in mich. Ich weiß, dass ihr euch sehr bemüht habt das ich heute hier sein kann, und dass ich es euch nicht immer leicht gemacht habe. Trotzdem habt ihr mich immer unterstützt und wart immer da, wenn ich euch gebraucht habe. Dieses Doktorat gehört sowohl euch als auch mir, und ich hoffe, ihr seid genauso stolz auf mich wie ich auf euch. Es ist ein wunderschönes Geschenkt euch als Eltern zu haben.

A mi hermana **Saskia** y a **Rafa**, y a la recién llegada **Aurora**, no me olvido de vosotros. A pesar de vernos mucho menos de lo que querríamos (a día de hoy aun no he podido tener en brazos a mi pequeña ahijada por culpa del coronavirus, pero ya verás cómo nos vamos a reír de esto cuando seas mayor) siempre estais ahí. Esta tesis también va por vosotros.

Vull incloure també en aquests agraïments a la meua família en Bocairent, en especial a **Paqui, Juan** i **Joana**, per haver-me obert les portes de la vostra casa des del moment en el que ens vam conèixer. Gràcies per fer-me sentir com u més de la família, per el vostre recolzament i per els vostres ànims.

He querido dejar para el final a **Marina**. No tengo palabras para expresar todo mi agradecimiento. Hace 10 años que llevas apoyándome incondicionalmente en toda esta aventura. Has sido un pilar fundamental en mi vida, porque sin tu apoyo esta tesis doctoral no hubiera salido adelante. Gracias por tu paciencia, tu sabiduría, tus consejos e incluso tus broncas y enfados. Con todos tus consejos, valoraciones y discusiones has hecho que esta tesis sea mejor. Gracias por compartir conmigo tantas aventuras y tantos buenos momentos. Gracias por tu amor y cariño. Gracias por estar a mi lado y andar conmigo el camino de la vida. Gracias por todo lo que me has dado y por todos los años y todas las aventuras que nos quedan por vivir juntos.

INDEX

ABSTRACT – RESUMEN – RESUM

- *Abstract*

Living beings face changing and usually stressful environmental conditions, aggravated by climate change, which tests their ability to survive. Changes in the genetic composition of the population, in the form of mutations, is the source for evolution and adaptation to environmental changes. This genetic diversity is driven by two major evolutionary forces that change the genetic composition in a population, allowing access to new phenotypes: genetic drift and natural selection. On one hand, genetic drift randomly fixes mutations in the population independent of their effect. On the other, natural selection either selects for beneficial mutations or purges deleterious mutations in a given environment. Hence, the effect of the mutations is closely linked to the environment and, as a consequence, populations exhibiting high genetic diversity can evolve faster and adapt better to environmental fluctuations. In addition to natural selection and genetic drift, gene duplication is also of great importance to evolution, as it is the main source of new genetic material. Not surprisingly, gene duplication has been related to major leaps in evolution, such as the radiation of angiosperm plants or large morphological innovations in animals. However, the molecular mechanisms that underlie the preservation of duplicated genes for long periods of time remain unknown. To better understand these mechanisms, experimental systems enabling rapid evolution are needed as the natural time scale for natural evolution can be extremely long. For this reason, experimental evolution approaches using viruses and microorganisms have become a valuable tool in evolutionary biology. On one hand, viruses show extremely high mutation rates, especially RNA viruses, conferring them the ability to rapidly adapt to a changing environment, thus representing an ideal model to study the effect of mutations. On the other hand, the yeast *Saccharomyces cerevisiae*, which has its origin in a whole genome duplication that occurred more than 100 million years ago, is a good model for studying gene duplication and its role in adaptation. Moreover, beyond their use as models for understanding evolutionary processes, the rapid evolutionary capacity of RNA viruses poses a challenge for treating and preventing infections and *S. cerevisiae* is currently one of the species with the largest biotechnological and economic impact. Therefore, a comprehensive analysis of the effect of mutations in large virus populations and, a deep knowledge about how genetic duplication influences adaptation and biological innovation is essential for gaining a better understanding of evolutionary processes and can help maximize our use of the full biomedical and biotechnological potential of RNA viruses and the budding yeast.

This doctoral thesis aims to shed light on two fundamental evolutionary biology questions: What are the molecular mechanisms that determine the genomic stability of

duplicated genes that are maintained in the genome for enough time to acquire evolutionary relevance? And, how does genetic variability contribute to evolution and adaptation? To address these questions we used two different models: the yeast *S.cerevisiae* and the coxsackievirus B3 (CVB3). In the first part of this thesis, we show that the level of gene expression of duplicated genes, as well as transcriptional and functional divergence, are key for the stability of duplicated genes. In addition, we find that the transcriptional plasticity of duplicated genes plays a key role in adaptation to new and stressful environments like high concentrations of ethanol, glycerol, and lactate or for oxidative stress conditions. In the second part, we performed a deep mutational scanning of the CVB3 capsid to generate highly genetically diverse populations and captured the mutational fitness effect of >90% of all possible single amino acid mutations in the viral capsid. We then used these highly diverse populations to study the contribution of genetic variability to adapt to thermal inactivation, observing that increasing the initial genetic variability in the population helps evolution even in RNA viruses with extremely high mutation rates.

- *Resumen*

Los seres vivos se enfrentan a condiciones ambientales cambiantes y habitualmente estresantes, agravadas por el cambio climático, que ponen a prueba su capacidad de supervivencia. El cambio en la composición genética de las poblaciones reside en las mutaciones, que son la fuente para la evolución y la adaptación a los cambios. La diversidad genética intrapoblacional está regulada por dos grandes fuerzas evolutivas que cambian la composición genética permitiendo así el acceso a nuevos fenotipos: la deriva genética y la selección natural. Por un lado, la deriva genética fija mutaciones en la población de manera aleatoria e independiente del efecto que suponga dicha mutación para la población. Por otro lado, la selección natural si bien favorece la fijación de mutaciones beneficiosas también elimina mutaciones perjudiciales en un determinado ambiente. Por lo tanto, el efecto de las mutaciones está fuertemente ligado al ambiente y, en consecuencia, aquellas poblaciones que exhiben una mayor diversidad genética serán capaces de evolucionar más rápido y adaptarse mejor. Además de la evolución por selección natural y deriva genética, la duplicación genética también es de especial importancia para la evolución, pues es la principal fuente de nuevo material genético y de innovaciones biológicas. Tanto es así que las grandes transiciones evolutivas, como la radiación de las plantas angiospermas o las grandes innovaciones morfológicas en animales se han relacionado con eventos de duplicación. Sin embargo, los mecanismos moleculares que permiten mantener los genes duplicados durante largos periodos de tiempo siguen siendo desconocidos. Para intentar ampliar el conocimiento respecto a estos mecanismos, y dado que los efectos de la evolución en la naturaleza tarda mucho tiempo en poder observarse, se necesitan sistemas biológicos que sean capaces de evolucionar rápido. En este contexto, te tiempo adecuado al experimentador, se han llevado a cabo estudios de evolución experimental con virus y microorganismos que han supuesto una herramienta muy valiosa en el estudio de la biología evolutiva. Por un lado, los virus presentan tasas de mutación extremadamente elevadas, especialmente los virus de ARN, lo que les confiere la capacidad de adaptarse de manera muy rápida a un cambio ambiental. Por otro lado, la levadura *Saccharomyces cerevisiae*, cuyo origen se debe a una duplicación genómica acaecida hace más de 100 millones de años, es un buen modelo para estudiar la duplicación genética y su papel en la adaptación. Además, más allá de ser útil para la evolución experimental y el estudio de la biología evolutiva, la elevada capacidad evolutiva de los virus de ARN supone un desafío importante para la medicina y en la prevención de enfermedades emergentes y, la levadura *S.cerevisiae* es una de las especies con mayor impacto económico en la industria biotecnológica. Por lo tanto, llevar a cabo un análisis exhaustivo del efecto de

las mutaciones en poblaciones virales y estudiar en profundidad como la duplicación genética influye la adaptación y la innovación biológica en la levadura, es fundamental para generar un marco de conocimiento que permitirá maximizar el potencial biomédico y biotecnológico de los virus de ARN y de la levadura.

Esta tesis doctoral trata de abordar dos cuestiones fundamentales en biología evolutiva: ¿Cuáles son los mecanismos moleculares que determinan la estabilidad de los genes duplicados en el genoma durante el tiempo suficiente para que sean capaces de adquirir relevancia evolutiva? Y, ¿Cómo contribuye la variabilidad genética a la evolución y a la adaptación a nuevos ambientes? Para intentar responder a estas preguntas hemos utilizado dos modelos experimentales diferentes: la levadura *S. cerevisiae* y el coxsackievirus B3 (CVB3). En la primera parte de esta tesis, con la levadura hemos visto que el nivel de expresión génica de los genes duplicados, así como la divergencia transcripcional y funcional, son fundamentales para la estabilidad de los genes duplicados en el genoma. Además, hemos observado que la plasticidad transcripcional de los genes duplicados juegan un papel clave en la adaptación a nuevos ambientes desfavorables, tales como condiciones de estrés oxidativo o altas concentraciones de etanol, glicerol o ácido láctico. Y en la segunda parte de la tesisi, utilizando el virus CVB3, hemos realizado una aproximación de *Deep mutational scanning* sobre las proteínas de la cápside viral y hemos generado poblaciones virales con una elevada variabilidad genética. Gracias a ello hemos podido evaluar el efecto de las mutaciones en la cápside caracterizando alrededor del 90% de los cambios de amino ácidos. Además, hemos empleado estas poblaciones virales altamente diversas y hemos estudiado como esta variabilidad genética contribuye a la adaptación contra la inactivación térmica. Nuestros resultados muestran que, incluso en virus de ARN con tasas de mutación extremadamente elevadas, un aumento de la diversidad genética de la población al inicio de la evolución experimental acelera el proceso evolutivo y facilita la adaptación al nuevo ambiente.

- *Resum*

Els éssers vius s'enfronten a condicions ambientals canviants i habitualment estressants, agreujades pel canvi climàtic, que posen a prova la seua capacitat de supervivència. El canvi en la composició genètica de les poblacions residix en les mutacions, que són la font per a l'evolució i l'adaptació als canvis. La diversitat genètica intrapoblacional està regulada per dues grans forces evolutives que canvien la composició genètica permetent així l'accés a nous fenotips: la deriva genètica i la selecció natural. D'una banda, la deriva genètica fixa mutacions en la població de manera aleatòria e independent de l'efecte que suposa aquesta mutació per a la població. D'altra banda, la selecció natural si bé afavorix la fixació de mutacions beneficioses també elimina mutacions perjudicials en un determinat ambient. Per tant, l'efecte de les mutacions està fortament lligat a l'ambient i, en conseqüència, aquelles poblacions que exhibixen una major diversitat genètica seran capaces d'evolucionar més ràpid i adaptar-se millor. A més de l'evolució per selecció natural i deriva genètica, la duplicació genètica també és d'especial importància per a l'evolució, ja que és la principal font de nou material genètic i d'innovacions biològiques. Tant és així que les grans transicions evolutives, com la radiació de les plantes angiospermes o les grans innovacions morfològiques en animals s'han relacionat amb esdeveniments de duplicació. No obstant això, els mecanismes moleculars que permeten mantindre els gens duplicats durant llargs períodes de temps continuen sent desconeguts. Per intentar ampliar el coneixement respecte a aquests mecanismes, i atès que poder observar els efectes de l'evolució en la naturalesa soposa molt temps, es necessiten sistemes biològics capaços d'evolucionar rapidament. En aquest context s'han dut a terme estudis d'evolució experimental amb virus i microorganismes que han suposat una eina molt valuosa en l'estudi de la biologia evolutiva. D'una banda, els virus presenten taxes de mutació extremadament elevades, especialment els virus d'ARN, el que els conferix la capacitat d'adaptar-se de manera molt ràpida a un canvi ambiental. D'altra banda, el llevat *Saccharomyces cerevisiae*, l'orige del qual es deu a una duplicació genòmica esdevinguda fa més de 100 milions d'anys, és un bon model per estudiar la duplicació genètica i el seu paper en l'adaptació. A més, més enllà de ser útil per a l'evolució experimental i l'estudi de la biologia evolutiva, l'elevada capacitat evolutiva dels virus d'ARN suposa un desafiament important per a la medicina i en la prevenció de malalties emergents i, el llevat *S.cerevisisae* és una de les espècies amb major impacte econòmic a la indústria biotecnològica. Per tant, dur a terme una anàlisi exhaustiva de l'efecte de les mutacions en poblacions virals i estudiar en profunditat com la duplicació genètica influeix l'adaptació i la innovació biològica en el llevat, és fonamental per generar un

marc de coneixement que permetrà maximitzar el potencial biomèdic i biotecnològic dels virus d'ARN i del llevat.

Aquesta tesi doctoral tracta d'abordar dues qüestions fonamentals en biologia evolutiva: Quins són els mecanismes moleculars que determinen l'estabilitat dels gens duplicats en el genoma, durant el temps suficient, per a que siguen capaços d'adquirir rellevància evolutiva? I, Com contribuix la variabilitat genètica a l'evolució i l'adaptació a nous ambients? Per intentar respondre a aquestes preguntes hem utilitzat dos models experimentals diferents: el llevat *S. cerevisiae* i el coxsackievirus B3 (CVB3). En la primera part d'aquesta tesi, amb el llevat hem vist que el nivell d'expressió gènica dels gens duplicats, així com la divergència transcripcional i funcional, són fonamentals per a l'estabilitat dels gens duplicats en el genoma. A més, hem observat que la plasticitat transcripcional dels gens duplicats juguen un paper clau en l'adaptació a nous ambients desfavorables, com ara condicions d'estrès oxidatiu o altes concentracions d'etanol, glicerol o àcid làctic. I en la segona part de la tesi, utilitzant com a model experimental el virus CVB3, hem realitzat una aproximació de *Deep mutational scanning* sobre les proteïnes de la càpside viral i hem generat poblacions virals amb una elevada variabilitat genètica. Gràcies a això hem pogut avaluar l'efecte de les mutacions en la càpside caracteritzant al voltant del 90% dels canvis d'aminoàcids. A més, hem utilitzat aquestes poblacions virals altament diverses i hem estudiat com aquesta variabilitat genètica contribuix a l'adaptació contra l'inactivació tèrmica. Els nostres resultats mostren que, fins i tot en virus d'ARN amb taxes de mutació extremadament elevades, un augment de la diversitat genètica de la població a l'inici de l'evolució experimental accelera el procés evolutiu i facilita l'adaptació a nous ambient.

RESUMEN EXTENDIDO

En la naturaleza los seres vivos cambian conforme cambia el medio ambiente que los rodea. Este proceso de cambio se conoce como evolución, descrito por Darwin hace más de 160 años en una elegante teoría que todavía se estudia hoy en día. Según la teoría de la evolución de Darwin los organismos viven en un mundo de recursos limitados en el cual han de competir, y solo aquellos individuos que estén mejor adaptados serán los que ganen la batalla por la supervivencia, se reproducirán más, y por lo tanto dejarán más descendencia en la siguiente generación. De este modo, sus caracteres hereditarios son seleccionados y fijados en la población por un proceso conocido como selección natural. En este escenario, las mutaciones genéticas son la base de los cambios que permite a los organismos modificar su fenotipo en función del ambiente, y de este modo combatir en la lucha por la supervivencia y adaptarse al mundo que los rodea. Con el tiempo, la teoría de la evolución de Darwin se ha conceptualizado y se ha convertido en la piedra angular de la biología actual, sin embargo, con el avance de la biología molecular y la genética contemporánea se ha podido profundizar y entender mejor los modelos mecanísticos que subyacen a los procesos evolutivos.

Como ya se ha mencionado anteriormente, la base del cambio genético es la mutación. Esta ocurre a nivel de nucleótido y debido a la redundancia del código genético una mutación puede ser sinónima o no, dependiendo de si codifica para el mismo amino acido o no. En consecuencia, la mayoría de mutaciones sinónimas no tendrán *a priori* ningún efecto en la población mientras que las mutaciones no sinónimas pueden ser beneficiosas, perjudiciales o neutras. Sin embargo, las mutaciones ocurren de manera aleatoria en el genoma de los organismos durante toda su vida, la mayoría de veces durante la replicación genómica. Por lo tanto, la variación genotípica y fenotípica de una población cambia de manera aleatoria y constante, adaptándose a un medio ambiente también en constante cambio. De hecho, los procesos evolutivos se estudian a nivel poblacional ya que dependen de una manera muy directa del tamaño poblacional efectivo, es decir, del número de individuos de la población que dejan descendientes en la siguiente generación y perpetúan así la especie. Por tanto, teniendo en cuenta el tamaño poblacional, el cambio de las frecuencias en los genotipos de la población están gobernadas por dos fuerzas evolutivas de sentido opuesto. Por un lado, en poblaciones pequeñas, la fuerza evolutiva principal es la deriva genética, según la cual las mutaciones aparecen y se fijan en la población de manera completamente aleatoria sin tener en cuenta si su efecto es beneficioso o perjudicial para la población. Por otro lado, en poblaciones más grandes, la fuerza evolutiva que conduce la fijación de nuevos genotipos es la selección natural, estando por lo tanto el efecto de la variación

genotípica y fenotípica fuertemente ligado al ambiente. En este contexto, la selección natural puede ser positiva o negativa dependiendo del efecto que ejerza sobre la fijación de nuevas mutaciones. La selección natural positiva o beneficiosa selecciona y permite la fijación de mutaciones beneficiosas para el individuo, acelerando la evolución y favoreciendo la variabilidad genética y la aparición de nuevos fenotipos. La selección natural negativa o purificadora tiene el efecto contrario, ralentizado la evolución porque purga de la población todas aquellas mutaciones que aparecen y suponen una desventaja para el individuo porque su efecto es perjudicial en el ambiente en el que aparece. Como se ha podido comprobar, el efecto de una mutación, de un genotipo o de un fenotipo no es un valor absoluto, sino que está fuertemente ligado al ambiente en el que ocurren. Teniendo esto en cuenta, las mutaciones beneficiosas, a pesar de ser las que ocurren en menor frecuencia son la base para los procesos evolutivos y la adaptación. En un modelo clásico de evolución adapatativa se postula que la adaptación se produce por la fijación de mutaciones muy próximas a la neutralidad, pero ligeramente beneficiosas, y la acumulación secuencial de estas incrementan gruadualmente la adaptación al medio ambiente, llevando en a la población a un estado óptimo de adaptación al ambiente. Sin embargo, como hemos visto anteriormente, las mutaciones beneficiosas también pueden ser eliminadas por la deriva genética, por lo tanto, encontrar el equilibrio para fijar una nueva mutación en la población no depende únicamente de la fuerza con la que actúa la selección natural sino también del tamaño población efectivo en la que esta mutación aparece.

Basado en estos modelos clásicos, hace ya más de medio siglo que Kimura reformuló la teoría de la evolución introduciendo la idea de la neutralidad y poniendo como primera fuerza evolutiva la deriva genética. Propuso que la mayoría de mutaciones aleatorias que ocurren en una población no tienen ningún efecto sobre el fenotipo, por lo que pasan desapercibidas para la selección natural y en consecuencia su fijación es aleatoria y completamente dependiente de la deriva genética. Esta idea ha ido madurando y hoy en día está bien instaurada en la biología evolutiva, es decir, que las mutaciones neutrales que pasan desaparecidas para la selección natural pueden ser clave para la innovación biológica.

Esta teoría de evolución neutral abre la puerta a nuevos conceptos como la robustez mutacional, es decir, la capacidad de un sistema biológico para mantener su fenotipo a pesar de acumular mutaciones. Como cabe esperar, la neutralidad de las mutaciones es de gran importancia para incrementar la robustez. Dado que, en la naturaleza, los organismos están continuamente confrontados con condiciones alejadas del óptimo que pueden llevar la población a la extinción, la capacidad de mantener un fenotipo

competitivo (es decir, un fenotipo que le permita sobrevivir mejor y reproducirse más en un determinado ambiente) a pesar de las perturbaciones ambientales y genéticas es beneficioso. De la robustez mutacional se observan inevitablemente dos consecuencias complementarias. En primer lugar, cuanta mayor robustez mutacional presente una población menor será el número de fenotipos observables en la misma. En segundo lugar, la variabilidad genotípica de la población puede variar rápidamente a través de las mutaciones neutrales favoreciendo el acceso a nuevos fenotipos en pocas mutaciones. Para los sistemas biológicos el segundo punto es fundamental, porque esa variabilidad genética críptica (indetectable para la selección natural) es la clave para la evolución ya que es la principal fuente de innovación biológica. No obstante, incrementar la variabilidad genética críptica es un proceso evolutivo que lleva mucho tiempo en la naturaleza, ya que la mayoría de mutaciones que aparecen suelen tener un efecto negativo y la selección natural las elimina rápidamente. En consecuencia, estudiar de manera empírica como contribuye la variabilidad genética a la evolución es laborioso y lleva mucho tiempo. A pesar de ello, se han llevado a cabo diferentes aproximaciones experimentales utilizando agentes mutagénicos físicos y/o químicos, así como PCR propensa al error para acelerar el proceso. Sin embargo, todos estos sistemas dependen fuertemente del genotipo inicial y en el caso de la PCR de la preferencia de la polimerasa a cometer unos errores más que otros. Recientemente se ha desarrollado una nueva técnica DMS (del inglés *Deep mutational scanning*) para llevar a cabo experimentos de mutagénesis a gran escala. Una forma de hacer DMS consiste en utilizar cebadores específicos cubriendo toda la secuencia que se quiera mutagenizar e introduciendo mediante los cebadores nucleótidos aleatorios en cada uno de los codones, evitando así los sesgos de las técnicas utilizadas previamente.

Además de la robustez también existe la plasticidad fenotípica como otro mecanismo para sobrevenir a un cambio ambiental. Esta se refiere a la capacidad de un organismo de cambiar su fenotipo en respuesta a un cambio ambiental brusco sin necesidad de cambiar su genotipo. La plasticidad fenotípica más estudiada es la plasticidad del desarrollo que se puede observar tanto en plantas como en animales. Sin embargo, la plasticidad fenotípica transcripcional es igual de importante y ha recibido menos atención por el momento. Este segundo tipo de plasticidad se define como la capacidad de un gen de cambiar su expresión en respuesta al ambiente.

Hoy en día, cualquier estudio que trate sobre la robustez mutacional, la plasticidad fenotípica o las innovaciones biológicas ineludiblemente ha de tratar la duplicación génica. Esta es el resultado de un error ocurrido durante la replicación del ADN y que resulta en la generación de dos copias genéticas idénticas. La duplicación génica se ha

considerado una de las mayores fuerzas evolutiva para la innovación biológica, ya que sirve como una gran fuente de robustez mutacional y plasticidad fenotípica: Mientras una de las copias puede explorar el espacio genotípico disponible la otra copia es capaz de mantener la función original gracias a la redundancia ganada en la duplicación. En 1970 Susumu Ohno propuso un modelo para la evolución de los genes duplicados proponiendo tres posibles destinos evolutivos: La *pseudogenización,* la *subfuncionalización* y la *neofuncionalización*. Por un lado, la *pseudogenización* es la más frecuente y ocurre que después de la duplicación una de las copias comienza a acumular mutaciones deletéreas hasta que pierde la función y es purgada por la selección natural negativa o purificadora. Por otro lado, si los genes duplicados se mantienen en el genoma el tiempo suficiente pueden ocurrir dos posibles desenlaces. Primero, la *subfuncionalización* ocurre cuando la función ancestral del gen se comparte entre las dos nuevas copias originadas. Por lo tanto, ambas copias son mantenidas por la selección natural ya que ambas son necesarias para mantener la función ancestral. En segundo lugar, la *neofuncionalización* es la que más importancia tiene a efectos de la evolución a pesar de ser la que ocurre en menor frecuencia. Consiste en que, después de un evento de duplicación, una de las dos copias mantiene la función ancestral mientras que la otra copia desarrolla una nueva función.

Sin embargo y a pesar del alto potencial evolutivo de la duplicación génica, este es un evento con poca estabilidad genómica porque tiene un alto coste de eficacia biológica e incurre en serias limitaciones como por ejemplo la pérdida del balance de dosis génica y la pérdida de la estequiometria proteica en proteínas multidominio. De hecho, la inestabilidad de la duplicación genética se refleja perfectamente en el genoma de la levadura cervecera *Saccharomyces cerevisiae*, que duplicó su genoma hace más de 100 millones de años, perdiendo, casi al mismo tiempo, el 92% de los genes originados de esa duplicación ancestral.

En la actualidad no existe consenso en la comunidad científica sobre qué ocurre después de la duplicación génica y que factores determinan el destino evolutivo de los genes duplicados. El modelo presentado por Ohno está ampliamente aceptado, pero la disponibilidad de genomas completos y los avances en biología y genética molecular han permitido profundizar y completar este modelo clásico. Se ha visto, entre otros, la importancia de la *subfuncionalización*. También se ha postulado que existen genes que previo a la duplicación ya tienen una función secundaria, que son capaces de explotar y perfeccionar una vez se han duplicado, y finalmente, también se ha planteado que la duplicación génica es una vía para escapar de un conflicto adaptativo de un gen para llevar a cabo dos funciones similares. Además, otro punto importante en el

entendimiento de la evolución de los genes duplicados es el mecanismo por el que se duplican, siendo este un factor clave que determina la evolución y el destino evolutivo de los genes duplicados. Es decir, los genes duplicados se pueden originar por duplicaciones completas del genoma (en adelante WGDs del inglés *Whole Genome Duplications*) o por duplicaciones de una pequeña región genómica o un único gen (en adelante SSDs del inglés *Small Scale Duplications*). Se ha visto que los WGDs son más propensos a *subfuncionalizar* porque al mismo tiempo que el gen se duplica también lo hace todo su contexto genómico, por lo que se mantienen todas las interacciones funcionales con otros genes manteniendo así el balance de dosis génica. Por otro lado, los SSDs tienden a *neofuncionalizar* dado que su origen unitario en el contexto genómico supone que puedan establecer nuevas interacciones funcionales con otros genes con mayor facilidad dado a la redundancia funcional del gen duplicado.

En consecuencia, parece evidente que la duplicación génica contribuye a la robustez mutacional, a la plasticidad fenotípica y a la plasticidad transcripcional y por ende a la innovación biológica por antonomasia y con todo, a la adaptación de los organismos a nuevos ambientes. Sin embargo, como se ha mencionado anteriormente, la gran inestabilidad de la duplicación ha llevado a que la levadura *S. cerevisiae* perdiera casi la totalidad de los genes duplicados originados por WGDs hace más de 100 millones de años. No obstante, a pesar de la gran pérdida de genes, el número de genes duplicados presentes actualmente en el genoma de la levadura en bastante superior al que cabría esperar si la perdida de genes se produjese por puro azar. Entonces, ¿Qué factores han determinado que genes se mantienen en el genoma y cuales se pierden? ¿Se puede establecer un modelo mecanístico y unitario para explicar la supervivencia de los genes duplicados en el genoma?

Los resultados presentados en la primera parte de esta tesis doctoral demuestran que uno de los factores que favorecen el mantenimiento de los genes duplicados, tanto WGDs como SSDs, en el genoma de la levadura *S. cerevisiae* es el nivel de expresión génica. Esta observación concuerda con la hipótesis de la *subfuncionalización* de dosis génica (DSH del inglés *Dosage subfunctionalization Hypothesis*) según la cual la divergencia entre las dos copias génicas se produce mediada por el nivel de expresión, siendo seleccionados por la selección natural aquellos genes con altos niveles de expresión. Analizando el perfil transcripcional de la levadura hemos observado que aquellos genes con mayores niveles de expresión génica están más conservados después de un evento de duplicación. Además, los genes duplicados con altos niveles de expresión no solamente están más conservados en el genoma de la especie en cuestión, sino que además presentan una mayor estabilidad filogenética en el subfilo

Saccharomicotina que incluye a la mayoría de levaduras ascomicetas. Este resultado se podría explicar por el balance de dosis génica, ya que aquellos genes que forman parte de un complejo proteico o de una proteína multidominio cabría esperar que tuviesen mayor estabilidad que las que no, puesto que la duplicación de todas las partes del complejo se verá favorecido por la selección natural en pro de mantener el equilibrio estequiométrico. No obstante, nuestros resultados muestran que los genes duplicados que hemos estudiado en esta tesis doctoral, tanto si proceden de una duplicación genómica completa (WGDs) como si provienen de la duplicación de una región genómica en concreto (SSDs), no están en ningún caso enriquecidos de genes que codifican para proteínas que forman parte de un complejo proteico. Por otro lado, también se ha visto que el ruido estocástico en la expresión de genes con altos niveles de expresión favorece la divergencia transcripcional y funcional después de un evento de duplicación, de manera que aumenta la probabilidad de que estos genes puedan resultar en una ventaja adaptativa para el organismo en cuestión.

Los resultados presentados en el segundo capítulo de esta tesis doctoral demuestran que los genes duplicados retenidos en el genoma de *S. cerevisiae* exhiben una mayor plasticidad transcripcional que los genes de copia única cuando la levadura se somete a condiciones de estrés ambiental como por ejemplo altas concentraciones de alcohol, estrés osmótico y nutricional por la presencia de glicerol en el medio de cultivo, a estrés ácido y nutricional por la adición de lactato como principal fuente de carbono o, finalmente, a estrés oxidativo por la adición de agua oxigenada al medio de cultivo que origina especies reactivas del oxígeno (ROS). Estos resultados concuerdan con la posibilidad de que en las regiones reguladoras de los genes en cuestión se hayan acumulado polimorfismos que podrían ser consideradas pre-adaptaciones a un cambio ambiental imprevisto. De hecho, cuando las células de la levadura *S. cerevisiae* son sometidas a un estrés osmótico y nutricional, al sustituir la glucosa por glicerol como principal fuente de carbono, se observa una gran reprogramación transcripcional que no solo afecta al cambio metabólico (de fermentación a respiración).

A pesar de que la transcriptómica de la levadura cervecera sometida a estrés se ha estudiado en profundidad, todos estos estudios han hecho poco énfasis en los diferentes patrones de expresión que se pueden observan entre los genes duplicados y genes de copia única. En esta tesis doctoral se observan dos puntos clave en este aspecto. En primer lugar los genes duplicados presentes en el genoma de *S. cerevisiae* muestran una alta plasticidad transcripcional cuando la levadura es sometida a condiciones de estrés. En segundo lugar, la plasticidad transcripcional observada entre los genes

duplicados y los genes de copia única es muy diferente, probablemente como resultado de la selección natural a una respuesta adaptativa de las células.

Como ya se ha mencionado, queda en evidencia la posibilidad de que la plasticidad transcripcional observada en la levadura sea resultado de la acumulación de polimorfismos en las regiones reguladoras de los genes dando origen a pre-adaptaciones a un nuevo ambiente, desconocido y estresante para la célula. Después de haber realizado una evolución experimental neutral (con un cuello de botella lo suficientemente amplio para que el efecto de la deriva genética sea mínimo, pero lo suficientemente pequeño para que la selección natural no seleccione las mutaciones por su efecto beneficioso en la población) durante 660 generaciones permitiendo a la población de levaduras incrementar su variabilidad genética. Cuando esta población evolucionada se confrontó con las condiciones de estrés osmótico y nutricional ocasionadas por sustituir la glucosa del medio de cultivo por glicerol o lactato, se observó que presentaba una mayor eficacia biológica (medida como la tasa de crecimiento) y una respuesta transcripcional completamente diferente a la población ancestral. En conclusión, y de acuerdo con estudios previos, este resultado es fácilmente atribuible a la presencia de una mayor variabilidad genética en la población evolucionada. Además, en esta tesis doctoral se evidencia que existen dos tipos de respuesta al estrés. Una primera respuesta rápida y general a cualquier cambio brusco del ambiente y otra respuesta más lenta y más específica a un estrés en concreto con el que se ha visto confrontado el organismo. La respuesta rápida y general al estrés tiene como propósito minimizar daños y maximizar las probabilidades de supervivencia, mientras que la respuesta especifica trata de adaptarse a la nueva situación con una reprogramación celular completa y haciendo uso de todas las posibilidades metabólicas y funcionales que la célula tiene a su alcance y así mejorar su supervivencia.

Además de todo lo mencionado anteriormente, el nivel de expresión también influye en el patrón de divergencia transcripcional después de ocurrir un evento de duplicación génica. Los genes duplicados con un patrón de expresión divergente entre las copias, es decir, que una copia aumenta su expresión mientras que la otra copia se silencia, o los genes duplicados donde solo una de las dos copias se expresa en condiciones de estrés celular, son más propensos a acabar divergiendo funcionalmente. En esta tesis doctoral se propone y se establece un vínculo entre la divergencia transcripcional y la divergencia funcional en el contexto de la hipótesis *misfolding-mistranslation*. Esta hipótesis postula que los genes con altos niveles de expresión presentan una evolución más lenta puesto que la acumulación de mutaciones deletéreas implica el mal plegamiento de las proteínas, resultando así en un alto coste de eficacia biológica. En

consecuencia, se establece una fuerte presión de selección negativa o purificadora sobre los genes altamente expresados para reducir el coste que supone el mal plegamiento de las proteínas. Por otro lado, la divergencia funcional necesita de un perfeccionamiento y de una regulación muy específica del nivel de expresión génica para llevar a cabo una nueva función en el momento adecuado. En definitiva, los datos presentados en la primera parte de esta tesis doctoral no demuestran si la divergencia transcripcional favorece y resulta en divergencia funcional o viceversa, pero sí que se puede observar que las diferentes categorías transcripcionales de los genes duplicados en condiciones de estrés para la levadura *S. cerevisiae* establecen un fuerte vínculo entre ambos tipos de divergencia.

En efecto, los diferentes patrones de expresión observados en *S. cerevisiae* cuando esta es sometida a condiciones de estrés ambiental concuerdan de muy buen grado con el modelo clásico de la evolución de los genes duplicados de Ohno. Nuestros resultados muestran que los genes duplicados en los que solo una copia altera su expresión en condiciones de estrés ambiental, son los que presentan una mayor capacidad de explorar el espacio genotípico puesto que la mayor redundancia genética obtenida por la duplicación génica permite que una de las copias pase desapercibida para la selección natural, ya que su copia hermana sigue llevando a cabo la función ancestral. Este planteamiento abre la puerta a que la copia que no está sujeta a selección natural tenga una mayor probabilidad de encontrar un nuevo fenotipo o una nueva función que sea favorable en un determinado ambiente y en consecuencia dar origen a una innovación biológica. Además, observamos que los genes duplicados que aumentan su expresión en condiciones de estés tienen una mayor probabilidad de *neofuncionalizar*, es decir, de encontrar una nueva función celular. Esto puede ser debido a la rápida evolución de las dos copias génicas y a la baja dependencia funcional, sugiriendo que después de ocurrir un evento de duplicación génica las dos copias divergen muy rápidamente generando una presión selectiva para mantener ambos genes en el genoma y por tanto confiriendo estabilidad a la duplicación. Por otro lado, los genes que disminuyen su nivel de expresión en condiciones de estrés muestran evidencias de haber sido fruto de la *subfuncionalización*, es decir, que la función ancestral que el organismo llevaba a cabo, antes de la duplicación, con un único gen ahora está dividida entre las dos copias hermanas. Esto podría ser debido a que la evolución de los genes con bajo nivel de expresión después de un evento de duplicación ha permitido la acumulación de mutaciones deletéreas en ambas copias por igual, resultando en una degeneración de los genes que acaba en la partición de la función ancestral y, por ende, se establece una presión de selección natural para el mantenimiento de ambas copias

en el genoma de la levadura. Finalmente, aquellos genes con niveles de expresión discordante, es decir que una copia aumenta el nivel de expresión mientras que su copia hermana lo reduce, presentan un bajo nivel de divergencia genética y también un bajo nivel de divergencia funcional, sugiriendo que la divergencia transcripcional que se observa es debida a cambios en las regiones promotoras para llevar a cabo una función similar a la ancestral, pero en un contexto diferente.

En vista de los resultados obtenido en esta tesis doctoral no cabe duda de la importancia de los genes duplicados para la evolución de los organismos y su relación con los niveles de expresión, no obstante, también puede explicar un conflicto entre el ruido transcripiconal de los genes altamente expresados y la plasticidad transcripcional. En esta tesis doctoral resolvemos este conflicto observando que los genes duplicados están más enriquecidos en cajas TATA en sus regiones reguladores. Estos motivos TATA se han relacionado tanto con un mayor ruido en los niveles de expresión génica pero también con una mayor plasticidad transcripcional de los genes que las presentan. De esta manera, y en base a los resultados obtenidos, sugerimos que son los motivos TATA en las regiones reguladoras de los genes duplicados los que les permiten exhibir una mayor plasticidad transcripcional, reduciendo así el conflicto de coste/beneficio que existen entre el ruido en el nivel de expresión y la plasticidad transcripcional.

En conclusión, en la primera parte de esta tesis doctoral se ha observado que los genes con altos niveles de expresión presentan una mayor estabilidad genómica en la levadura y una mayor estabilidad filogenética en el subfilo *Saccharomycotina*. Además, en base a la hipótesis de la *subfuncionalización* de dosis, los genes duplicados con altos niveles de expresión presentan un mayor grado de variabilidad transcripcional, y también que los genes con altos niveles de expresión presentan una mayor plasticidad transcripcional en la levadura *Saccharomyces cerevisiae* sometida a condiciones de estrés ambiental. Se ha visto que la respuesta de estrés en la levadura altera la expresión de muchos genes, en su mayoría duplicados, sugiriendo que la plasticidad transcripcional se ha obtenido después de un evento de duplicación. También se ha relacionado la presencia de cajas TATA en las regiones promotoras de los genes duplicados contribuyendo a la plasticidad transcripcional. En cuanto a la respuesta al estrés, se ha observado que, en términos generales, más genes duplicados alteran su expresión, especialmente WGDs, y que estos genes contribuyen en mayor grado a la eficacia biológica del organismo. También, de acuerdo con los modelos de evolución de los genes duplicados, hemos observado que los genes en los que solo se altera una copia permite a la otra copia buscar nuevas funciones en el espacio genotípico, y que los genes duplicados que aumentan su expresión son más probables de

neofuncionalizar mientras que los genes duplicados que reducen su expresión son más propensos a *subfuncionalizar*.

En la segunda parte de esta tesis doctoral retomamos la cuestión de analizar la importancia y la contribución de la variabilidad genética para la evolución y la adaptación a nuevos ambientes. Utilizando el virus coxsackievirus B3 como modelo experimental de un virus de ARN hemos realizado una aproximación DMS sobre las proteínas de la capside viral y hemos caracterizado el efecto de más del 90% de las mutaciones en la estabilidad de la capside y en la eficacia biológica de las poblaciones virales. Nuestros resultados muestran que la gran mayoría de mutaciones tienen un efecto negativo, de hecho, solo el 1.2% tiene un efecto ligeramente beneficioso. Es interesante que este resultado es extrapolable a otros virus de la familia ya que hemos observado una fuerte correlación entre el efecto mutacional observado experimentalmente con el predicho a partir de los alineamientos de secuencia de otros virus de ARN. Además, nuestra aproximación nos ha permitido caracterizar y estudiar diferentes fenómenos sobre los mecanismos que confieren estabilidad a la capside y que son importantes para el correcto ensamblaje de la misma. En esta tesis doctoral aportamos datos empíricos sobre el efecto de las mutaciones sobre la estructura secundaria, la predisposición a formar agregados proteicos, o la conservación de regiones importantes para la unión de anticuerpos y el escape de la respuesta inmune del huésped. Estas observaciones han servido para implementar mejoras en los modelos de reconstrucción filogenética, aportando información adicional al alineamiento de secuencias y mejorando así la precisión del modelo. Finalmente, también hemos utilizado la cuantificación del efecto mutacional para establecer un modelo predictivo de proteínas citoplasmáticas del huésped que pueden tener un papel importante para el proceso infectivo y, por tanto, ser candidatas a tener un efecto antiviral desde una perspectiva biomédica.

Además, en esta segunda parte de la tesis doctoral hemos utilizado las poblaciones virales generadas por DMS para estudiar el efecto de incrementar la variabilidad genética en un proceso de evolución experimental. Hemos visto que, después de 10 pases seleccionando la fracción superviviente de la población después de haberse inactivado térmicamente, aquellas poblaciones que parten inicialmente de una variabilidad genética mayor son capaces de alcanzar estados óptimos con mayor rapidez y presentan más resiliencia a cambios de temperatura en el ambiente. Además, los resultados presentados en esta tesis doctoral han permitido caracterizar mutaciones importantes en la capside viral de CVB3 para resistir la inactivación térmica.

En conclusión, en el conjunto de esta tesis doctoral mostramos la importancia que tiene la variabilidad genética para la evolución, sirviendo de puente entre diferentes picos óptimos en el paisaje evolutivo. También mostramos que mediante una aproximación de DMS se puede obtener una gran variabilidad genética y que esta puede ser de utilidad para experimentos de evolución dirigida. Además, nuestros resultados muestran claras evidencias, tanto en virus como en levadura, que un aumento en la variabilidad genética es beneficioso para sobrevenir y adaptarse a un cambio brusco en el ambiente. De hecho, nuestros resultados concluyen en la importancia de los genes duplicados en la levadura para acumular esta variabilidad genética y contribuir así a los procesos adaptativos.

GENERAL INTRODUCTION

1. How organisms survive and change over time

More than 160 years ago, Charles Darwin proposed in his book "*On the origin of the species by means of Natural Selection, or the preservation of favoured races in the struggle for life*" a new and revolutionary theory of evolution that completely changed the understanding of how organisms inhabit Earth. He presented a model in which resources are limited in the environment and, only those individuals that can win the fight for resources can reproduce and produce offspring for the next generation. In his evolutionary model, Darwin proposed that species are gradually changing their phenotypes to adapt to a continuously changing environment, and only the best-performing phenotypes are selected and fixed in the population to continue into further generations. By this, organisms can change and adapt to the environment over time. Darwin's theory of evolution and its conceptualization of natural selection became key in understanding current biology and life sciences. However, the rise of molecular biology and contemporary genetics improved the understanding of the mechanisms underlying evolutionary processes (Nielsen 2005; Fares 2015a; Salas 2019).

It is well known that the building block for evolution is mutation. It is also well known that this occurs at the nucleotide level, and because of the redundancy in the genetic code, a mutation can be non-synonymous or synonymous, depending on if the encoding amino acid does change or not, respectively. Indeed, most synonymous mutations will have no effects in the population *a priori*, while non-synonymous mutations can be beneficial, detrimental, or neutral. Notwithstanding, mutations occur randomly across the genome of the organisms during their lifetime, mostly during genome replication, hence the genetic variation in the population changes randomly and so do new phenotypes. Indeed, evolutionary processes are studied at the population level, with changes in genotypic frequencies being driven by two opposing evolutionary forces that depend on the effective population size (i.e. the number of individuals in a population that can reproduce and contribute offspring to the next generation). On one hand, in small populations, the fixation of new genotypes occurs randomly by genetic drift. On the other hand, in bigger populations, the fixation of mutations is governed by natural selection and therefore has a strong environmental determinant. In this context, natural selection can be positive or negative depending on whether the effect of a new mutation is respectively more favorable or less favorable to the genotype established in the population in a given environment. In fact, despite being less frequent, beneficial mutations form the base for adaptation.

A classical model for adaptive evolution is Fisher's model (Fisher 1930), in which mutations with small effects have a higher chance of being beneficial compared to mutations with large effects, resulting in a gradual adaptation by the cumulative fixation of slightly beneficial mutations, driving the population to a phenotypically optimal state for a given environment. Consequently, it is expected that mutations with small effects that are close to neutrality are fixed sequentially in the population, gradually increasing the adaptation of the population (Muller 1964). Nevertheless, beneficial mutations can also be lost by genetic drift, and therefore the ability to reach high frequency in the population depends on the strength of selection as well as population size.

Based on Fisher's and Muller's models, Kimura reformulated the theory of evolution by introducing the idea of neutrality and establishing genetic drift as the major force in evolution. He proposes that neutral mutations arise in the population and consequently are unseen by natural selection, with their fixation being randomly driven by genetic drift (Kimura 1968, 1983). This concept has gained force, with the idea of neutrality being well established in evolutionary biology. Hence, neutral mutations without phenotypic effect are neither purged by purifying selection nor selected by positive natural selection and can become key in evolutionary innovations (Wagner 2005, 2012; Fares 2015c; Zhang 2018; Wideman *et al.* 2019). However, over the last few years, controversy about what drives evolution has arisen, with some authors completely rejecting neutrality (Kern and Hahn 2018) and others defending Kimura's theory of neutral evolution (Jensen *et al.* 2019). Although the debate is still open, the importance of considering neutralism and natural selection as important concepts for evolution seems clear: Neutralism allows accelerating the rate of fixation of mutations in the population that can later be selected by natural selection (Wagner 2008, 2012; Draghi *et al.* 2010; Fares 2015b; Baier *et al.* 2019). The combination of these different evolutionary forces makes populations more resilient to changing environments.

2. *Mutational robustness and phenotypic plasticity as mechanisms for increasing genotypic variability and its consequences for evolution*

Neutralism in evolution is especially important for increasing mutational robustness, which is defined as the ability of a biological system to maintain its phenotype despite accumulating mutations (Waddington 1942; Félix and Wagner 2008; Wagner 2012; Fares 2015c). As mentioned above, organisms must deal with a large range of non-optimal situations that can founder the population and lead it to extinction. Therefore, the

ability to maintain the fittest phenotype despite environmental or genetic perturbations is beneficial for biological systems. Thus, the more robust is a system the more mutations it can tolerate without a phenotypic effect. From this, two complementary phenomena are observed. First, in robust populations, there are fewer different phenotypes present. Second, genotypic variability accumulates rapidly through the population, hence, increasing neutral genetic variability (i.e. cryptic genetic variation) that is unseen by natural selection and confers the potential to access new phenotypes by small changes. For evolving systems the second point is key, because cryptic genetic variability enhances the capacity to evolve (i.e. evolvability), serving as a source for new adaptations and evolutionary innovations (Wagner 2011; Zheng *et al.* 2019; Wideman *et al.* 2019). Nevertheless, mutational robustness and evolvability reach a threshold above which an increase in the neutral genotypic network (i.e the accessible genotypes for a population without changing the phenotype) is detrimental for evolvability because it becomes challenging to access new phenotypes and evolutionary innovations. (Figure I-1A) (Fares 2015b; c).

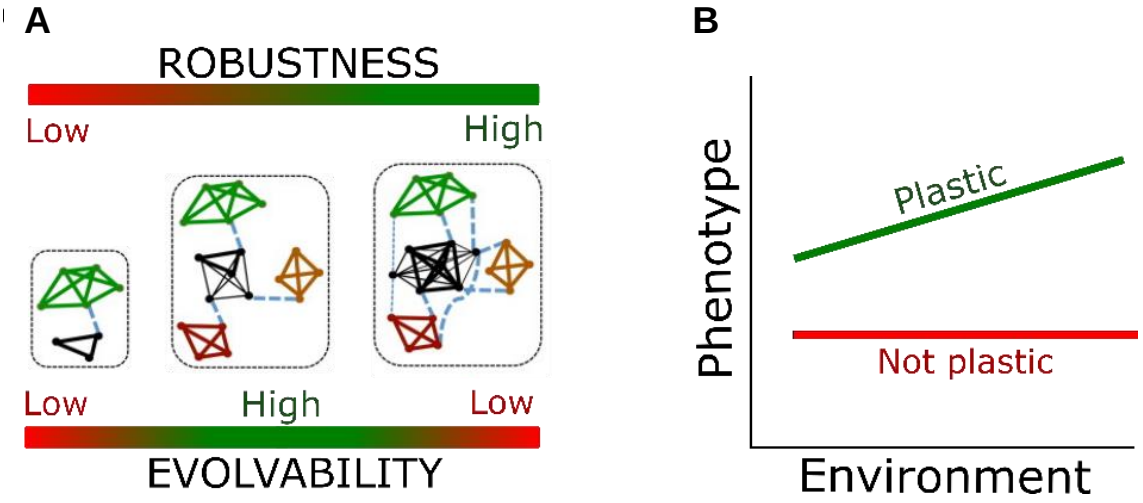


***Figure I-1* - Robustness and phenotypic plasticity are important pillars of evolution**. **A)** Robustness is defined by genotypic networks in which the different genotypes (nodes) are connected by neutral mutations that result in the same phenotype (network color). The lower the genotypic network (black network in the left panel), the lower potential for evolution (i.e. evolvability) of new phenotypes (green network). By increasing robustness, the genotypic network (black network in the middle panel) also increases the number of accessible new phenotypes, thus increasing the evolvability of the system. Notwithstanding, excessively large genotypic networks (i.e. extreme robustness, as in the black network in the right panel) decreases evolvability because different genotypes overlap in their accessible phenotypes. Adapted from Fares 2015c. **B)** Phenotypic plasticity is the ability of a biological system to change its phenotype in response to an environmental change. The red line represents an organism without phenotypic plasticity, in which the phenotype remains unaltered despite a changing environment. On the contrary, the green line represents an organism with phenotypic plasticity, in which the phenotype varies according to change in the environment.

Yet, increasing cryptic genotypic variability in a population is an evolutionary process that takes a long time, since neutral mutations appear at low frequencies. Indeed, most mutations that arise in a population are deleterious and are rapidly purged by purifying natural selection. Hence, empirical studies to unveil the contribution of genetic variability to the evolution of organisms are laborious and time-consuming. Notwithstanding, different approaches to artificially increase genetic variability have been performed using, for example, mutagenic agents like ethyl methanesulfonate (EMS) in animals (Flibotte *et al.* 2010), plants (Jander *et al.* 2003; Ke *et al.* 2019), or more recently in bacteriophages (Favor *et al.* 2020); UV exposure in *E.coli* (Shibai *et al.* 2017); or error-prone PCR of an enzyme in different organisms (Armetta *et al.* 2019; Guan *et al.* 2020; Thompson *et al.* 2020). However, these approaches have a strong bias as a result of the genetic background of the population (Stoltzfus and Norris 2016; Lyons and Lauring 2017; Couce and Tenaillon 2019; Svensson and Berger 2019; Canoid and Payne 2020) and the preferences for misincorporation of the polymerase (Acevedo *et al.* 2014; Geller *et al.* 2016). In the last years, deep mutational scanning (DMS) has been developed as a new tool for high-throughput mutagenesis experiments (Araya and Fowler 2011; Fowler and Fields 2014). Among the different DMS approaches, one experimental setup used to do codon level mutagenesis consists of the usage of primers spanning all codons of the desired target gene encoding random bases (Bloom 2014; Dingens *et al.* 2017; Bornscheuer and Höhne 2018). Thus, this DMS approach overcomes the biases of the previously described methods allowing the screening of almost the complete genotypic network of the population. Deep mutational scanning has been widely used in different systems: in viruses DMS has been used, for example, to define the protein stability, host tropism, or antibody neutralization sites. (Thyagarajan and Bloom 2014; Doud and Bloom 2016; Haddox *et al.* 2016, 2018; Ashenberg *et al.* 2017; Doud *et al.* 2017; Hom *et al.* 2019; Sourisseau *et al.* 2019; Lee *et al.* 2019); in bacteria and yeast DMS has been used to score the mutational effect on essential genes in *E.coli* (Choudhury *et al.* 2020), or to analyze the specificity of RNA recognition motifs in yeast (Melamed *et al.* 2013). DMS has also been used to study the fitness effect of mutations in a specific enzyme affecting, for example, its solubility (Klesmith *et al.* 2017).

In addition to robustness, another mechanism to overcome the challenge of environmental perturbations and expand genotypic variability within a population is phenotypic plasticity, which refers to the capability of the organisms to rapidly adapt their phenotype to a sudden fluctuation in the environment (Figure I-1B) (Price *et al.* 2003; Kelly *et al.* 2012). The most noticeable phenotypic plasticity is the developmental plasticity observed in plants and animals (Schlichting 1986; West-Eberhard 1989;

Lafuente and Beldade 2019). In sessile organisms like plants, phenotypic plasticity plays an important role in surviving in a changing environment. For example, the shape and color of leaves in a plant (the phenotype) can change depending on light, humidity, and temperature (the environment). As leaves in plants are responsible for photosynthesis and thermoregulation, fine regulation of their phenotype is important for survival (Chitwood and Sinha 2016; Fritz *et al.* 2018). Also in animals, phenotypic plasticity can be observed. A good example of phenotypic plasticity in animals, in particular of developmental plasticity, is observed in the pea aphid *Acyrthosiphon pisum* that induces the transition from parthenogenetic females to sexual morphs. This transition occurs when winter is coming and a shortening in the day length is perceived by the insect. Indeed, the phenotypic plasticity that allows the arising of sexual morphs is of special importance for generating a diapause egg that supposes the survival of the pea aphid in middle and higher latitudes (Simon *et al.* 2002, 2011; Richards *et al.* 2010).

Although developmental plasticity is widely observed, phenotypic plasticity can also be transcriptional. Transcriptional plasticity is defined here as the ability of the gene to change its expression when the environment changes without changing the genotype. An example of such transcriptional plasticity is presented by the curculionid *Tribolium castaneum*, which in the presence of pesticide diflubenzuron can control ABC transporter's efflux, hence contributing to the drug resistance of the insect (Rösner and Merzendorfer 2019).

3. Evolution by gene duplication

An important mechanism to increase robustness and phenotypic plasticity is gene duplication. Gene duplication is the result of an error occurring during DNA replication that leads to the generation of two identical gene copies. It is considered one of the major forces for evolutionary innovations. Already in 1932 Haldane suggested that gene duplication might be important for evolution because this would allow for accumulating mutations in one of the copies while the other one remains unaltered, thus avoiding natural selection (Haldane 1932). Indeed, gene duplication has been related to major leaps in evolution taking place in most organisms from the unicellular to the multi-cellular (Otto and Whitton 2000). In plants, the expansion and radiation of Angiosperms (i.e. flowering plants) has been related to a whole-genome duplication event that occurred roughly 200 million years ago (Kim *et al.* 2004; Cui *et al.* 2006; Soltis *et al.* 2009; Freeling 2009; Rensing 2014; Panchy *et al.* 2016; Clark and Donoghue 2018; Clark *et al.* 2019;

Defoort *et al.* 2019; Larson *et al.* 2019; Si *et al.* 2019). Also in animals, important evolutionary landmarks have been related to gene and genome duplication events, including increased synapse and behavior complexity, the formation and plasticity of the neural crest in vertebrates, the increased glycolytic fluxes, or the evolution of circadian clocks (Gibson and Spring 1998; Ohno 1999; Dermitzakis and Clark 2001; Lespinet *et al.* 2002; Hoegg *et al.* 2004; Looby and Loudon 2005; Freeling and Thomas 2006; Hoffmann *et al.* 2011; Storz *et al.* 2011, 2013; Hoffmann, Opazo, and Storz 2012; Hoffmann, Opazo, Hoogewijs, *et al.* 2012; Green and Bronner 2013; Dennis and Eichler 2016; Morgan *et al.* 2016; Zhou *et al.* 2019).

In 1970 Ohno proposed a simple but elegant model for the evolution of gene duplication with three possible outcomes after gene duplication (Figure I-2). On one hand, *Pseudogenization,* which is the most common outcome after duplication, occurs when one of the copies accumulates deleterious mutations until it loses the function and is purged by purifying natural selection, returning to the single-copy gene model. On the other hand, if the duplicated gene persists, two possible results are expected. First, *Sub-functionalization* occurs when the function of the ancestral gene is shared among both gene copies and both copies become selected by natural selection. Second, *Neo-functionalization,* which is the most important scenario for evolution, entails one gene copy retaining the ancestral function while the other evolve a new function. Nevertheless, despite the evolutionary potential for gene duplication as a source of new genetic material, this event has a high fitness cost and there exist several constraints such as the loss of the gene dosage balance or proteic stoichiometry in multimeric and multidomain proteins (Adler *et al.* 2014).

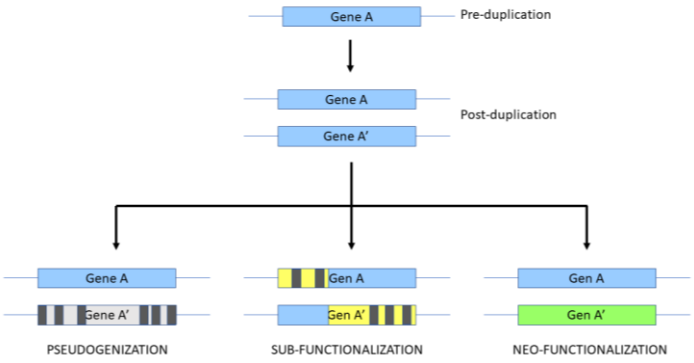


***Figure I-2* - Ohno's model for the evolution of gene duplication.** After a duplication event, the duplicated gene copies can evolve into three different fates. *Pseudogenization*: One copy accumulates deleterious mutations (dark grey squares) until it loses the function and is purged by purifying selection. *Sub-functionalization*: Both gene copies accumulate deleterious mutations retaining different parts or subfunctions of the ancestral function, and so it becomes essential to retaining both copies to perform the ancestral function (represented in blue). *Neo-functionalization*: One gene copy evolves into a new function (represented in green) and both genes are selected by natural selection.

Currently, there is no consensus in the scientific community about what occurs after a gene duplication event, and several models that attempt to explain the evolution of duplicated genes exist (Innan and Kondrashov 2010). Ohno's model is widely accepted and used in other conceptual frameworks. However, the availability of complete genome sequences and advances in molecular biology led to an improvement and an update of Ohno's theory (Zhang 2003). First, *sub-functionalization* was revised and amplified in the Duplication-Degeneration-Complementation model (DDC) (Force, Lynch, and Postlethwait 1999; Force, Lynch, Pickett, *et al.* 1999). The DDC model states that after a duplication event the resulting copies are neutral, so they are unseen by natural selection and maintained at low frequency in the population accumulating mutations due to genetic drift. During this degeneration period, each of the copies will lose some ability to perform the ancestral function reaching a point in which *sub-functionalization* is complete and, natural selection will maintain both copies since both copies are now necessary to perform the original function. Second, the Innovation-Amplification-Divergence model (IAD) was introduced, revising and amplifying *Neo-functionalization* fate (Bergthorsson *et al.* 2007). The IAD model states that the 'new function' is already present in the pre-duplication ancestral gene (i.e. by enzymatic promiscuity, in which a protein can perform a secondary less effective function), facilitating the rise of the new function by sequence divergence of the duplicates favored by natural selection. This model was experimentally tested in the enterobacteria *Salmonella enterica*, studying the evolution of a bifunctional parental gene able to biosynthesize principally Histidine but also Tryptophan at a lower level, yet sufficient for cells to grow in a media lacking both essential amino acids. After a few hundred generations in media lacking both amino acids, the parental bifunctional gene was duplicated, and each copy specialized in the biosynthesis of one amino acid (Nasvall *et al.* 2012). Finally, Escape from Adaptive Conflict (EAC) is a model that merges the DDC and IAD models described before. EAC model states that a gene can evolve two different beneficial functions simultaneously. This situation generates an 'adaptive conflict' because it is unlikely for the gene to perform both functions at maximum efficiency. Consequently, after a gene duplication event, both genes undergo *sub-functionalization* similar to what occurs in the DDC model but with the difference that the ancestral pre-duplication gene is already multifunctional, similar to what occurs in the IAD model (Hittinger and Carroll 2007). A good example of the EAC model was observed in plants with the flowering inducing gene FT and its paralog TFL1, being the later responsible for flowering repression. The pre-duplication gene ancestor was performing both opposite functions and, after gene duplication, each paralog specialized in one function resolving the adaptive conflict and performing a better and more finely controlled regulation of flowering in plants (Moraes *et al.* 2019; Jin *et al.*

2020). In addition to the different models of functional and transcriptional evolution of gene duplication, it has been observed that the mechanism of gene duplication also matters (i.e. Whole Genome Duplications, hereinafter WGDs, versus Small Scale Duplications, hereinafter SSDs), with WGDs being more prone to *sub-functionalization* while SSDs generally leading to *neo-functionalization* (Figure I-3) (Carretero-Paulet and Fares 2012; Fares *et al.* 2013). On one hand, WGDs are more likely to persist in the genome because the dosage balance is maintained in the organisms. Thus, after WGDs all the functions and interactions are doubled in the system. This redundancy means that a gene originated by WGDs is more robust to the loss of some of its interactors and can maintain better the ancestral function. Hereafter, the unbalanced loss of genes after a WGDs drives the system to be more prone to *sub-functionalization* (Figure I-3A). On the other hand, SSDs establish more functions and stronger interactions than WGDs because the system is more fragile to lose any of the interactors of a duplicated gene originated by SSDs, and therefore SSDs are often followed by *neo-functionalization* of one of the copies (Figure I-3B).

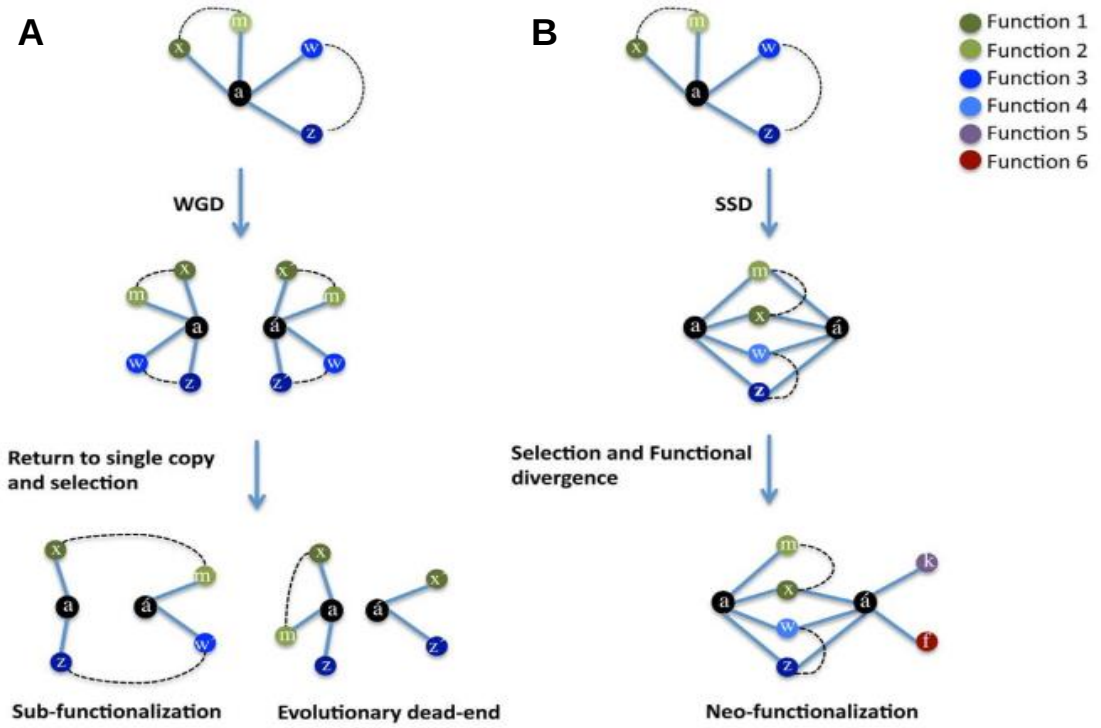


***Figure I-3*- Model of evolution after gene duplication about the origin of duplicated genes.** **A)** Whole-genome duplication (WGD) generates a complete doubling of genes (nodes: a, x, m, w, z), functions (colors), and interactions (solid edges). The partners of the duplicated gene (a) also interact functionally with one another (dashed edges) and are stoichiometrically balanced. Genetic redundancy resulting from WGD causes stochastic gene loss and drives the system to *sub-functionalization* (i.e. the ancestral function is yet divided between the two resulting gene copies). **B)** Small scale duplication events (SSD) resulting in the duplication of one or few genes in the genome, increase considerably genetic robustness. Thus, retaining both gene copies in the genome for enough time entails the functional divergence of one of the gene copies (á) and the evolution of biological and functional innovation (nodes k and f). Extracted from Fares et al. 2013.

4. *Experimental evolution is a valuable tool for studying evolutionary processes in the laboratory.*

One problem of studying evolutionary mechanisms and evolutionary events is that evolution takes a long time to happen in nature. To shed light on the molecular biology and mechanisms underlying evolutionary processes, experimental evolution has been developed as a valuable tool to reduce time and study evolution live. This consists of the setup of laboratory-controlled experiments, mostly with microorganisms, to explore evolutionary dynamics (Rainey *et al.* 2017). One of the best known experimental evolution experiments was done with the enterobacteria *Escherichia coli* by Richard Lenski's laboratory. It consists of a long-term evolution experiment that was set up in 1988 and continues today, having reached over 73,000 generations (http://myxo.css.msu.edu/ecoli/). One interesting result they observed was the nascent ability of *E.coli,* at roughly 30,000 generations, to aerobically metabolize citrate from media causing an increase in population size and diversity (Blount *et al.* 2008). More recently, a clonal experimental evolution (i.e. reducing drastically the population size at each passage, hence allowing the accumulation of deleterious mutations by genetic drift) was done with the baker's yeast *Saccharomyces cerevisiae* for over 2,000 generations. This experiment showed how ancient gene duplication contributes to genetic robustness and its role in adaptation to stress (Keane *et al.* 2014). Many other experimental evolution experiments have been performed in a vast number of different organisms, ranging from viruses (Grubaugh and Andersen 2017; Doud *et al.* 2017; Arribas *et al.* 2018; Gutierrez *et al.* 2019; Leeks *et al.* 2019) to plants (Gervasi and Schiestl 2017), animals (Andersson 2012; Hoang *et al.* 2016) or just enzymes (Baier *et al.* 2019).

5. *RNA viruses as a good model for experimental evolution*

Picornaviruses have been widely used for studying different aspects of viral biology, viral replication, and interaction with the host because these viruses are among the simplest of vertebrate viruses, are easy to grow and, medically and economically relevant (Tuthill *et al.* 2010; Zell 2018). In particular, Coxsackievirus B3 (CVB3) is an important human pathogen belonging to the enterovirus B genus of the picornavirus family, being the major pathogen of human viral myocarditis. (Liu *et al.* 2014; Garmaroudi *et al.* 2015). Indeed, CVB3 can be detected in the heart muscle of ~50% of patients with dilated cardiomyopathy (Liu *et al.* 2014). Aside from the medical and economic impact of CVB3,

this virus possesses a short replication time and its replication, structure, and pathogenesis has been well described (Muckelbauer *et al.* 2004; Yoder *et al.* 2012; Liu *et al.* 2014; Garmaroudi *et al.* 2015; Sin *et al.* 2015). CVB3 is a positive single-stranded RNA virus with a 30nm nonenveloped icosahedral capsid composed of four proteins (VP1-4). A molecule composed of one of each protein forms a protomer, then five protomers compose a pentamer, and finally, twelve pentamers make the whole capsid where the ~7.5kb viral genome is packaged. Like other RNA viruses, CVB3 has a high mutation rate which is orders of magnitude above those of most DNA-based organisms (Sanjuán and Domingo-Calap 2016; Peck and Lauring 2018). Hence, RNA viruses in general, and CVB3 in particular, show an elevated capacity for evolution. These high mutation rates together with short replication time, relatively simple genomic structure, and large population sizes make them ideal models for experimental evolution.

In this doctoral thesis, we use CVB3 to perform a comprehensive deep mutational scanning on the whole viral capsid to expand and define the complete genotypic and phenotypic space of a multimeric protein complex: the viral capsid. Additionally, we have used these highly diverse viral populations to analyze the effect of artificially increasing genetic variability during an experimental evolution experiment aimed at adapting to a new environment. We find that an increase of the genetic variability of the startling population results in improved adaptation even in RNA viruses with already high mutation rates.

6. Saccharomyces cerevisiae, a simple eukaryotic model organism

The budding yeast *Saccharomyces cerevisiae* is a widely used research model microorganism. This is because it is a simple unicellular eukaryotic system, slightly more complex than bacteria, with short generation time (doubling time ~1.5 hours at 28ºC) and ease of manipulation. *S. cerevisiae* has a well-studied genome of roughly 12Mb, organized in 16 chromosomes and containing around 6,300 genes (Goffeau *et al.* 1996). Additionally, genetic engineering can be performed easily through homologous recombination and/or cell transformation by a lithium acetate protocol (Gietz and Schiestl 2007; Gietz 2014). A large number of resources are available at the *Saccharomyces Genome Database* (https://www.yeastgenome.org/). Moreover, an interesting and valuable genome deletion project has been started in the year 2000 and to date, 90% of the *S. cerevisiae* genome has been disrupted, and over 20,000 knock-out strains are available (Giaever and Nislow 2014).

7. *Adaptation and gene duplication in the budding yeast*

The budding yeast *Saccharomyces cerevisiae* is widely distributed around the world, colonizing a large number of habitats and ecological niches (Cray, Bell, *et al.* 2013). Most importantly, yeast utilizes nutrients not only as a source of energy but also as signals which control the developmental, metabolic, and transcriptional activities of the cell (Broach 2012). The robust and versatile stress response of yeast allowed them to spread and adapt rapidly, and to dominate large microbial communities (Cray, Bell, *et al.* 2013).

Besides*, S. cerevisiae* duplicated its genome ~100 million years ago resulting from the possible hybridization between different yeast species (Wolfe and Shields 1997; Marcet-Houben and Gabaldón 2015; Wolfe 2015). Nevertheless, due to the instability of gene duplication, around 92% of the duplicated genes returned to being a single copy. Despite this big purge of duplicated genes, the actual number of duplicated genes present in the yeast genome (around 30% of all the genes) is significantly higher than expected if the gene loss would have occurred randomly. This raises the question of what factors influence the preservation of duplicated genes in the yeast genome. As mentioned before, numerous models have been proposed to explain the evolution and fate of duplicated genes, but a general mechanistic model for evolution by gene duplication is still needed to explain the retention of duplicated genes.

In this doctoral thesis, we propose that the expression level of genes is key for their retention after duplication, with highly expressed genes being more likely to be preserved. We also show strong evidence that gene duplication in the yeast *S. cerevisiae* contributes to adaptation to new stressful environments. This observation links mutational robustness gained through gene duplication with transcriptional plasticity and functional divergence. This link is in agreement with the hypothesis that, after gene duplication, natural selection on the gene copies is relaxed. As a consequence, the gene copies can explore the genotypic space and, following the classic model of evolution by gene duplication proposed by Ohno, eventually find a biological innovation.

OBJECTIVES

The main aim of this doctoral thesis is to unveil the contribution of gene duplication and genetic variability in biological innovation through experimental evolution. On one hand, the differential fixation of mutations between gene copies can lead to the functional and transcriptional divergence of the copies that might confer an adaptative advantage for the organism. Yet, the importance of this phenomenon, especially of transcriptional plasticity, has been poorly studied. For this, performing a quantitative analysis of transcriptional plasticity will help to provide a better understanding of the factors that determine the evolutionary fate of duplicated genes. On the other hand, defining the viable sequence space of multimeric proteins will help to understand better the constraints underlying the evolution of protein complexes. Finally, generating a comprehensive analysis and characterization of highly diverse populations can expand the knowledge about the contribution of genetic variability to the adaptation to environmental changes.

This doctoral thesis has been divided into two parts, one for each experimental model. In the first part, the main objective is to study the adaptive landscape of the yeast *Saccharomyces cerevisiae*, performing an integrative approach to understand the evolutionary fate of duplicated genes and phenotypic plasticity. In the second part, the main objective is to define the complete viable sequence space of the coxsackievirus B3 capsid and study the contribution to adaptation of the genotypic variability within the population. To achieve this, four specific objectives were proposed:

- **Part I: Biological innovation through gene duplication: adaptative evolution in *Saccharomyces cerevisiae***
    1. Determine the regulatory and genomic bases for the stability of duplicated genes and their relevance in transcriptional plasticity.
    2. Study the role of phenotypic plasticity and the contribution of duplicated genes in the cellular response of the yeast to environmental stress, adaptation, and biological innovation.
    3. Analyze in detail the transcriptional reprogramming in short stress response, and the adaptation to chronic environmental stress.

- **Part II: Biological innovation through mutation: functional sequence space of a viral capsid**
    4. Determine the viable sequence space of coxsackievirus B3 capsid proteins by deep mutational scanning and experimental evolution.

PART I – Biological innovation through gene duplication: adaptative evolution in *Saccharomyces cerevisiae*.

Objectives to achieve in this part:

a. Determine the regulatory and genomic bases for the stability of duplicated genes and their relevance in transcriptional plasticity. **Chapter I**.
b. Study the role of phenotypic plasticity and the contribution of duplicated genes in the cellular response of yeast to environmental stresses, adaptation, and biological innovation. **Chapter II.**
c. Analyze in detail the transcriptional reprogramming in short stress response, and the adaptation to chronic environmental stress. **Chapter III, IV and V.**

<u>CHAPTER I</u> – The regulatory and genomic bases for the stability of duplicated genes and their relevance in transcriptional plasticity.

A version of this chapter has been published as:

Mattenberger F., Sabater-Muñoz B., Toft C., Sablok G., Fares M.A. (2017) *Expression properties exhibit correlated patterns with the fate of duplicated genes, their divergence, and transcriptional plasticity in Saccharomycotina*, **DNA Research, 24(6): 559–570**.

1. Abstract

Gene duplication is an important source of novelties and genome complexity. What genes are preserved as duplicated through long evolutionary times can shape the evolution of innovations. Identifying factors that influence gene duplicability is therefore an important aim in evolutionary biology. Here, we show that in the yeast *Saccharomyces cerevisiae* the levels of gene expression correlate with gene duplicability, its divergence, and transcriptional plasticity. Genes that were highly expressed before duplication are more likely to be preserved as duplicates for longer evolutionary times and wider phylogenetic ranges than genes that were lowly expressed. Duplicates with higher expression levels exhibit greater divergence between their gene copies. Duplicates that exhibit higher expression divergence are those enriched for TATA-containing promoters. These duplicates also show transcriptional plasticity, which seems to be involved in the origin of adaptations to environmental stresses in yeast. While the expression properties of genes strongly affect their duplicability, divergence and transcriptional plasticity are enhanced after gene duplication. We conclude that highly expressed genes are more likely to be preserved as duplicates due to their promoter architectures, their greater tolerance to expression noise, and their ability to reduce the noise-plasticity conflict.

2. *Introduction*

Gene duplication is believed to be a rich source of novel functions and adaptations (Ohno 1970, 1999; Lynch and Conery 2000). This belief is supported by evidence coming from innovations following gene duplications in yeast, plants and animals. Indeed, protein families expanded after whole-genome and small-scale duplications yielding an unprecedented morphological diversity in plants (Wendel 2000; Otto and Whitton 2000; Holub *et al.* 2001; Lespinet *et al.* 2002; Van De Peer 2004; Cui *et al.* 2006; Soltis *et al.* 2009; Carretero-Paulet and Fares 2012). Other major innovations in animals have also been achieved through gene duplication (Hoegg *et al.* 2004), including increased synapse and behavior complexity (Grant 2016) and the neural crest formation and plasticity in vertebrates (Green and Bronner 2013). In yeast, gene duplication has contributed to metabolic innovation through the alteration of regulatory and transcriptional networks (Huminiecki and Conant 2012) or the increased glycolytic fluxes (Conant and Wolfe 2007). However, it remains unclear why certain duplicates have been preferred over others to persist in the genomes and be the source of innovations.

Since duplication is immediately followed by relaxed selection constraints on one or the two gene copies, the survival time of each gene copy is a limiting factor in the determination of its functional fate. In the majority of cases, duplication is resolved by the non-functionalization of one of the gene copies and its subsequent erosion from the genome (Ohno 1970, 1999; Lynch and Conery 2000). Accordingly, 92% of all genes that were duplicated through whole-genome duplication (WGD) > 100 MYA in *Saccharomyces* returned to single-copy genes 'shortly' after duplication (Wolfe and Shields 1997). Nonetheless, in many species, including yeast, the number of duplicated genes is larger than predicted by theory ranging between 30% of the genes in yeast (Fares *et al.* 2013) and more than 50% in plants (Blanc and Wolfe 2004a; Cui *et al.* 2006). Determining what genes remain in the genome as duplicates, and consequently lead to evolutionary leaps, is an important aim in evolutionary biology. However, this objective remains to be achieved.

A number of hypotheses have been proposed to explain the persistence of certain genes in duplicate. Rapid sequence divergence between gene copies can lead to their functional divergence followed by strong selective constraints on each copy, which could contribute to the preservation of duplicates in the genomes (Blanc and Wolfe 2004b; Fares *et al.* 2006; Scannell and Wolfe 2008; Conant and Wolfe 2008). Functional divergence requires, nevertheless, long evolutionary times, and given that selection relaxes after gene duplication, selective pressures are unlikely to retain both gene copies during the first million years following duplication. Preservation of duplicates can also be

selectively favored by the need to maintain gene–dosage balance (Conant and Wolfe 2008; Freeling 2009; Carretero-Paulet and Fares 2012; Conant *et al.* 2014), or provide genetic robustness against deleterious mutations (Keane *et al.* 2014; Fares 2015b). However, all these scenarios do not provide a general mechanistic explanation for what makes duplicates persist or alternatively perish.

Recently, it has been proposed dosage sub-functionalization as a plausible hypothesis to explain the fate of whole-genome duplicates (Gout and Lynch 2015). According to this hypothesis, highly expressed genes are more likely to be preserved as duplicates than lowly expressed genes. This is because stochastic variations in the levels of expression of the gene copies of highly expressed duplicates would not lead to copies with a lower expression level than that required for purifying selection to act upon them. Therefore, highly expressed duplicates are less likely to return to single-copy genes by drift. Whether this hypothesis could be applied to all duplicates regardless of the mechanism that originated them and whether such dosage sub-functionalization could also determine the patterns of divergence between gene copies has not been explored before.

Here, we present evidence that the levels of gene expression are correlated with the fates of whole-genome and small-scale duplicates, with highly expressed genes being more likely to be retained in double copy after duplication for longer periods of time than lowly expressed genes. Such duplicates are also more phylogenetically stable. We also show that the ancestral levels of gene expression are correlated with the evolution of duplicates expression. Retained duplicated genes evolve strong patterns of transcriptional (also known as phenotypic) plasticity, which are also correlated with the levels of gene expression. Finally, while the levels of gene expression are correlated with the duplicability of genes, duplicates phenotypic plasticity is manifested only after gene duplication; and this plasticity is proportional to the expression divergence between the copies of duplicated genes.

3. Material and methods

a. Identification of duplicated genes

Paralogs pairs of duplicated genes were identified as the resulting best reciprocal hits from all-against-all BLAST searches using BLASTP with an E-value cutoff of $1E^{-5}$ and a 50-bit score (Altschul *et al.* 1997). Paralogs were then divided into two groups according to the mechanism of their origin: WGDs and SSDs. WGDs are those extracted from the reconciled list provided by the Yeast Gene Order Browser (YGOB, http://wolfe.gen.tcd.ie//ygob; Byrne and Wolfe, 2005) (555 pairs of genes), and these were not subjected to subsequent SSD. All other paralogs were considered to belong to the category of SSDs (560 pairs of genes).

b. Growth of S. cerevisiae and gene expression analyses

The transcriptomic profiling was performed in the *S. cerevisiae* Y06240 haploid *msh2* deletion strain (BY4741; *Mat**a** his3D1 leu2DO met15DO ura3DO msh2::kanMX4*), with three technical replicates for each biological stress condition (3% lactic acid {YPL}, 3% ethanol {YPE}, 3% glycerol {YPG}, 0.25mM H_2O_2 + 1.5% dextrose {YPOxD}) in comparison with the normal growth condition (YPD media) (Figure ChI-1). Total RNA extractions were performed with RNeasy kit (Qiagen) following manufacturer instructions. Ribosomal RNA was removed by using Ribo-Zero Gold rRNA removal yeast (Illumina) depletion kit. Stranded RNA libraries were constructed using TruSeq stranded mRNA (Illumina) from oligo-dT captured mRNAs from depleted samples. Libraries were run in NextSeq 500 (Illumina) at 75nt single read by using High Output 75 cycles kit v2.0 (Illumina).

RNA libraries were sequenced at Genomic core facility at Servicio Central de Soporte a la Investigación Experimental (SCSIE) from University of Valencia, Spain. Raw reads were analyzed using FastQC report and cleaned with CutAdapt as implemented in RobiNA software package v 1.2.4 (Lohse *et al.* 2012). Low-quality reads were filtered and trimmed (Phred score inferior to 20 and size less than 40 nt were discarded). The reads were then aligned with Bowtie (up to two mismatches accepted) to the reference transcriptome (PRJNA290217) from the reference S288c strain. The normalization and statistical evaluation of differential gene expression has been performed using edgeR (Robinson *et al.* 2010) or DESeq (Anders and Huber 2010) with a P-value cut-off of 0.05, using the Benjamini–Yekutieli (Benjamini and Yekutieli 2001) method for multiple testing correction of P-value, and setting the Log-fold change at min = 1 to determine differential expression. The raw data (reads counts) was normalized according to the default

procedure of the differential expression analysis package used (edgeR or DESeq), being the dispersion estimated using the pooled setting, and RPKM (Reads Per Billion) expression values estimated as implemented in RobiNA software 3.1 All newly sequenced RNA sequences are available from the Sequence Read Archive with the following accession number (SRP074821). Expression data for each of the *S. cerevisiae* genes under YPD and each of the four stress conditions as well as the adjusted probabilities to identify significant fold changes are available in Supplementary Tables S1–S4.

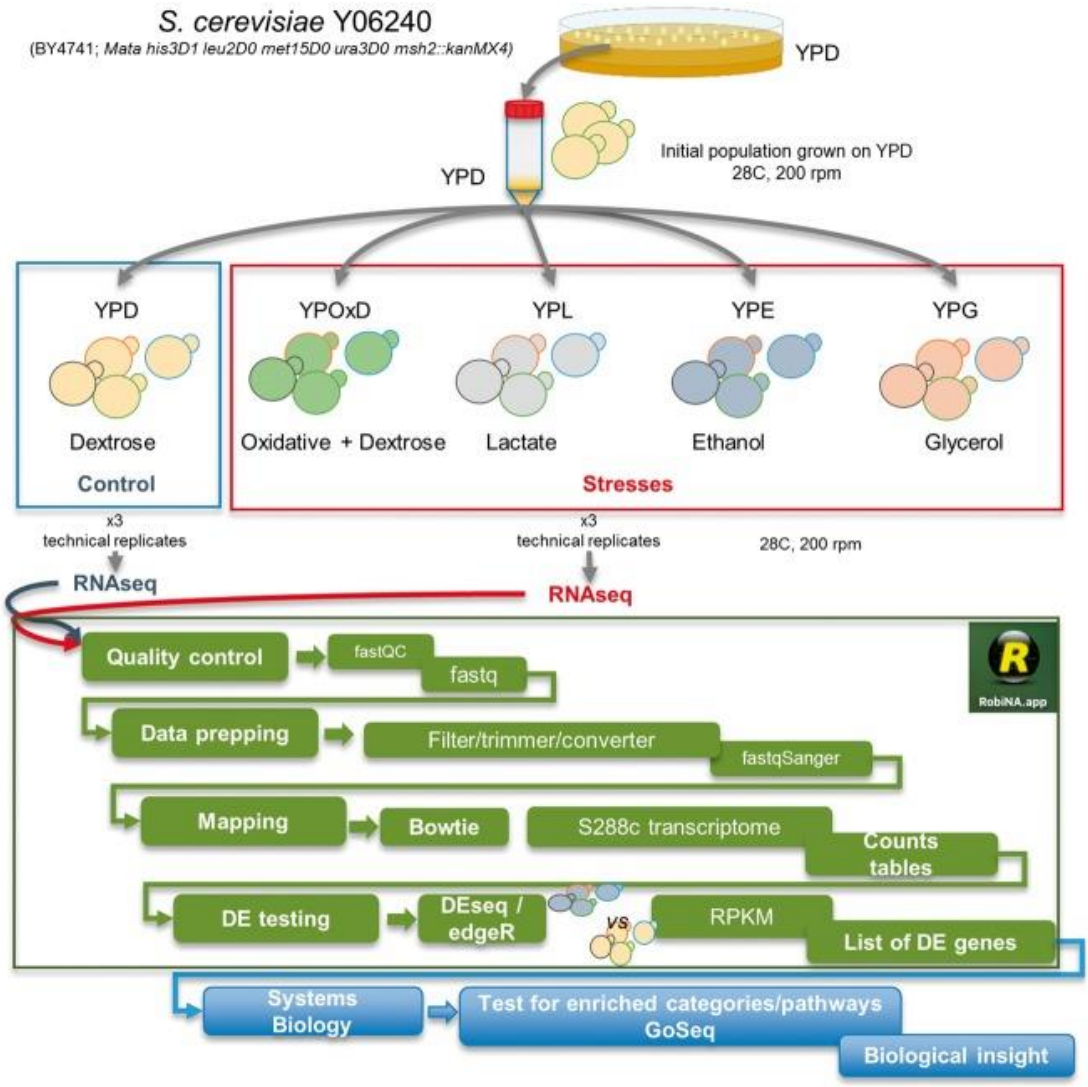


***Figure ChI-1* - Growth experiments of *S. cerevisiae* under stress conditions**. An initial isogenic population of *S. cerevisiae* (strain Y06240) was grown for 24h in Yeast extract, peptone, dextrose medium (YPD) at 28°C. This grown population was then subjected to four different stress conditions: ethanol, glycerol, lactate and oxidative stress in a medium supplemented with dextrose (each stress is indicated in a different color). Growth experiments were performed in triplicate. Samples of each of the populations, in addition to the control population grown in YPD, were processed for RNA sequencing. The technical steps for the extraction and analysis of RNA sequences are also shown.

c. Expression data for Lachancea kluyveri

Growth conditions, RNA extraction, and sequencing are specified in a previous study (Anders and Huber 2010). Briefly, authors performed the analyses on *L. kluyveri* reference strain CBS 3082_a (*MAT**a***). Transcriptomic data were obtained from growth cultures at mid exponential phase and for 20 different media, including YPD and 19 other stress conditions (listed in Ref. Anders and Huber, 2010). RNA sequencing was performed using Illumina HiSeq2000 platform with 50-base pair non-oriented single reads (Supplementary Table S5).

d. Expression data for Candida glabrata

Candida glabrata ATC2001 strain expression data (in the form of RPKMs) were obtained from a previous study. (Linde *et al.* 2015). Briefly, authors grown *C. glabrata* in normal YPD media and in the M199 medium at different pH values for pH shift. Total RNA was isolated by hot phenol-chloroform method. Libraries were subjected to ribosomal RNA depletion and sequenced using Illumina HiSeq2000 platform with 100-bp paired-end strand-specific. Reads were mapped with TOPHAT2, and counted using htseq (-m union, -t exon conditions) and normalized by the number of reads per kilobase of exon per million mapped reads (Supplementary Table S6). Differentially expressed genes were identified with raw counts by DeSeq and EdgeR, using the same cut-off parameters than the ones used in this study.

e. Software

Calculations and statistics were performed using MS Excel and R 3.2.1. Data management was possible using in-house built PERL scripts.

4. Results

a. *Duplicates preservation and phylogenetic stability are correlated with the levels of gene expression*

Highly expressed genes are more likely to be preserved as duplicates after whole-genome duplication (WGD) than lowly expressed genes in *S. cerevisiae* (Seoighe and Wolfe 1999; Aury *et al.* 2006; Gout *et al.* 2010; McGrath, Gout, Doak, *et al.* 2014; McGrath, Gout, Johri, *et al.* 2014; Gout and Lynch 2015). We searched for the orthologs of *S. cerevisiae* genes in *L. kluyveri* (strain CBS 3082; synonymous of *Saccharomyces kluyveri*) and their expression in Yeast Extract Peptone containing dextrose medium (YPD) (Brion *et al.* 2016) (Supplementary Table S5). *Lachancea kluyveri* is a respiratory yeast species pre-dating the WGD that took place in *S. cerevisiae* > 100 million years ago (Wolfe and Shields 1997). We found *L. kluyveri* orthologs for 5643 *S. cerevisiae* genes. Of the 5643 genes, 1469 genes were orthologs of *S. cerevisiae* duplicates, including WGDs and small-scale duplicates (SSDs), and 4174 were orthologs of *S. cerevisiae* singletons (Supplementary Table S5). The expression of *L. kluyveri* orthologs of *S. cerevisiae* duplicates (Median: 10.43; measured as the log2-transformed Reads Per billion, RPKM) was significantly greater than that of *L. kluyveri* orthologs of *S. cerevisiae* singletons (Median: 9.52) (Wilcoxon rank test: $P < 2.2 \times 10^{-16}$). We compared the transcription levels of genes in *S. cerevisiae* obtained in our study with those from another study (Albert *et al.* 2014) that used ribosomal profiling, a technique that measures ribosome occupancy and translation genome wide and provides an accurate measure of the translatable mRNA. RPKMs correlated strongly and significantly with the data of ribosome profiling (Spearman's correlation: $\rho = 0.77$, $P < 2.2 \times 10^{-16}$, Supplementary Data S1), indicating that RPKMs are indicative of the levels of gene expression and also the translatable mRNAs.

Previous studies concluded that highly expressed genes were more likely to be preserved as duplicates after WGDs because of absolute dosage constraints and constraints on dosage balance (Seoighe and Wolfe 1999; Papp *et al.* 2002; Gout *et al.* 2009, 2010; Qian *et al.* 2010; Birchler and Veitia 2012; Gout and Lynch 2015). Indeed, we found that this trend is true for *L. kluyveri* orthologs of *S. cerevisiae* WGDs ($N = 561$, Median expression: 10.64, Wilcoxon rank test: $P < 2.2 \times 10^{-16}$) and also for orthologs of *S. cerevisiae* SSDs ($N = 908$, Median expression: 10.28, Wilcoxon rank test: $P < 2.2 \times 10^{-16}$). The level of expression of orthologs of WGDs was, nevertheless, higher than that for SSDs (Wilcoxon rank test: $P = 7.79 \times 10^{-5}$).

The higher expression of duplicates compared to singletons can be due to a greater presence of genes encoding protein-complex proteins among duplicates than singletons.

We extracted the list of protein complexes from a previous study, using the table of annotated yeast high-throughput complexes available at (http://wodaklab.org/cyc2008/downloads) (Pu *et al.* 2009). Genes encoding proteins that are part of protein complexes (N = 1913) (Supplementary Table S7) do exhibit greater expression (Median expression: 11.56) than genes encoding complex-free proteins (N = 3960) (Median expression: 10.61, Wilcoxon rank test: $P < 2.2 \times 10^{-16}$). However, neither WGDs were more enriched for complex-encoding genes than singletons in *S. cerevisiae* (Fisher's exact test: F = 1.02, P = 0.76) nor SSDs showed significant difference in terms of enrichment for complex-encoding genes when compared to singletons (Fisher's exact test: F = 1.11, P = 0.15).

One caveat in this analysis is that gene expression in *L. kluyveri* may not reflect gene expression immediately after WGD. Against this prediction, gene expression in *L. kluyveri*, a species predating the WGDs and SSDs used in this study, was strongly and significantly correlated with gene expression in *S. cerevisiae* (Spearman correlation: $\rho = 0.59$, $P < 2.2 \times 10^{-16}$, Figure ChI-2A).

Duplicated genes exhibit different patterns of gene retention and phylogenetic stability in the different post-WGD Saccharomyces species (Scannell *et al.* 2006). We classified *S. cerevisiae* duplicated genes according to the presence of the two copies in each of the twelve available species post-dating the WGD (Figure ChI-2B). We first asked whether the expression of duplicates generated before *Saccharomycetales* speciation (including WGDs and SSDs) correlates with their phylogenetic stability, measured as the mean number of post-WGD species in which each copy was present (Figure ChI-2B).

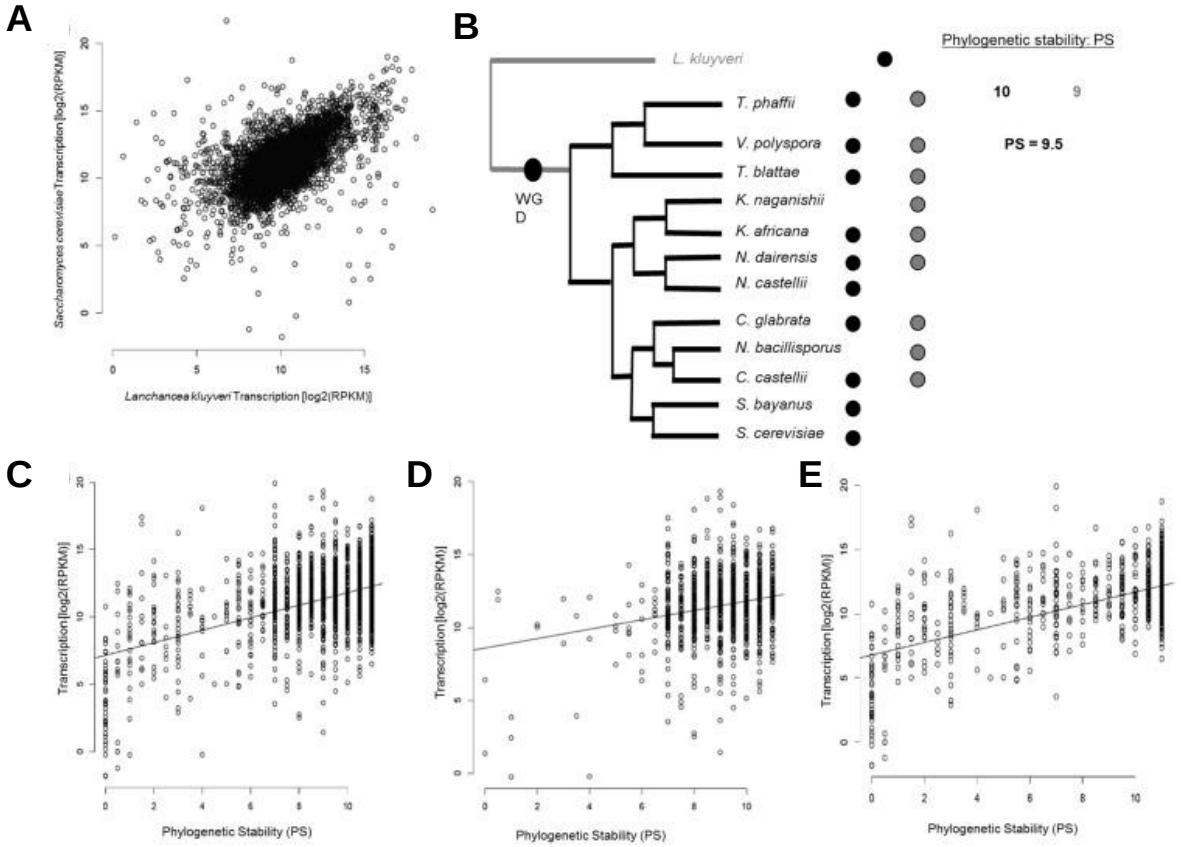


Figure ChI-2 - The levels of gene expression determine gene duplicability. **A)** Expression of genes in the pre-WGD species *Lachancea kluyveri* correlates positively with gene expressions in the post-WGD species *Saccharomyces cerevisiae*. Axes represent the transcription levels of genes as the logarithmic transformation of the Fraction of Reads Per billion. **B)** Estimation of the phylogenetic stability (i.e. number of species) of duplicated genes after the whole genome duplication in yeast. Black and grey circles refer to gene copies. A missing circle in one of the post-WGD species represents a gene copy loss, hence the return of that duplicate to single-copy gene in that species. The phylogenetic stability (PS) of that duplicate is calculated as the mean number of species in which gene copies are present. **C)** Phylogenetic stability correlates with the levels of gene expression of duplicates. **D)** Correlation of PS of WGDs and their expression levels. **E)** Correlation of PS of SSDs and their expression levels.

To determine the number of species postdating WGD in which each gene copy is present we used the Pillars information available from the Yeast Gene Order Information (Byrne and Wolfe 2005), which provides gene order and annotation for 12 post-WGD yeast species (Supplementary Table S8). For each gene copy, we counted the number of species in which it is found and averaged this number for the two sister gene copies in a duplicated pair (Figure ChI-2B). There was a positive and significant correlation between the mean gene copies expression in YPD and their phylogenetic stability (Spearman correlation: $\rho = 0.27$, $P < 2.2 \times 10^{-16}$, Figure ChI-2C). We then repeated the analysis for WGDs and SSDs separately. WGDs exhibited positive weak but significant correlation between gene expression and phylogenetic stability (Spearman correlation: $\rho = 0.13$, $P = 1.15 \times 10^{-5}$, Figure ChI-2D). In contrast, SSDs showed strong and

significant correlation between gene expression and phylogenetic stability (Spearman correlation: $\rho = 0.40$, $P < 2.2 \times 10^{-16}$, Figure ChI-2E).

b. The magnitude of divergence of duplicates expression correlates with the level of gene expression

A pivotal hypothesis to the dosage sub-functionalization proposed by Gout and Lynch (2015) is that highly expressed duplicates should exhibit more expression variation despite the action of purifying selection than lowly expressed genes. This is because noise in the expression of highly expressed genes is unlikely to compromise the selective constraints on these genes. That is, genes with higher expression levels should be more 'noisy' in their expression when duplicated, and thus they should generate more expression polymorphism in the population than lowly expressed genes. Accordingly, highly expressed duplicates are more likely to yield gene copies with diverged expressions than lowly expressed duplicates. To test this hypothesis, we first measured the fold expression difference (D) between the gene copies i and j when *S. cerevisiae* was grown under YPD as

$$D_{i,j} = 1 - \left[Min\left(E_i, E_j\right)\right] / \left[Max\left(E_i, E_j\right)\right]$$

with E referring to the expression of the gene under normal conditions (Supplementary Table S9). $D_{i,j}$ is normalized by the level of gene expression (i.e. the value is $0 \leq D_{i,j} \leq 1$), and thus it is an unbiased measure of the expression divergence between the gene copies. In support of our hypothesis, there was a weak but very significant correlation between the average expression of the gene copies of duplicates and $D_{i,j}$ (Spearman correlation: $\rho = 0.18$, $P = 4.23 \times 10^{-8}$). This correlation was also maintained when we analyzed separately WGDs (Spearman correlation: $\rho = 0.18$, $P = 6.98 \times 10^{-5}$) and SSDs (Spearman correlation: $\rho = 0.17$, $P = 2.89 \times 10^{-4}$).

c. The expression levels and promoter architecture correlate with patterns of expression divergence of duplicates and their transcriptional plasticity

Because higher expression can increase the chance for expression divergence after gene duplication, we sought to investigate if genes with higher expression can also evolve greater transcriptional plasticity under stress. Transcriptional plasticity is defined here as the ability of the gene to change its expression, while keeping its genotype, when the environment changes. For all four stresses with which *S. cerevisiae* was challenged (see section Material and methods), transcriptionally altered duplicates belonged to a set of genes with significantly higher expression in YPD growth media than transcriptionally unaltered duplicates (Figure ChI-3A-D). This was also true, with the exception of ethanol-induced stress, for singletons, albeit the effect was more pronounced in duplicates than

it was in singletons. Noticeably, for all four-stress conditions in *S. cerevisiae*, the levels of expression of duplicates with no altered transcription under stress were significantly higher than that for transcriptionally altered singletons (Figure ChI-3A-D).

We explored other mechanistic explanations for this expression difference between unaltered duplicates and altered singletons. One important factor that contributes to transcriptional plasticity is the existence of the TATA-box motif in the gene promoter, with TATA-containing genes being more sensitive to regulatory changes than TATA-less genes (Landry *et al.* 2007)

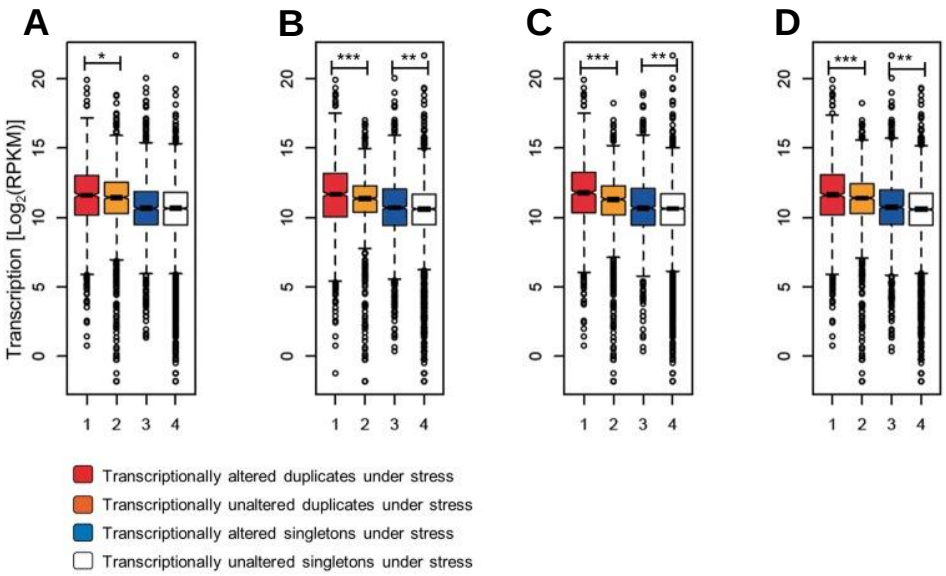


Figure ChI- 3 - The levels of gene expression determine the transcriptional plasticity of genes. We present the levels of expression as the logarithmic transformation of the Reads Per billion (RPKM) for duplicates (red boxes) and singletons (blue boxes) with transcriptional alterations and duplicates (orange boxes) and singletons (white boxes) without altered expressions when *S. cerevisiae* is faced with four different stress conditions **A)** Ethanol, **B)** glycerol, **C)** lactate and **D)** oxidative stress in a medium supplemented with dextrose. Significant differences in the expression levels of sets of genes are indicated as *, **, ***, when the probabilities are $P < 0.05$, $P < 0.01$ and $P < 10^{-6}$, respectively, using a Wilcoxon rank test.

Importantly, the level of expression of TATA-containing genes was higher than that of TATA-less genes, and this was true for duplicates and singletons in *S. cerevisiae* (Figure ChI-4A). The set of transcriptionally altered genes under stress conditions was enriched for TATA-containing genes when compared to the set of genes with no transcriptional plasticity, being this true for duplicates (Figure ChI-4B) and singletons (Figure ChI-4C). Generally, TATA-containing genes also exhibit expression noise, which can be coupled with transcriptional plasticity provided that noise and plasticity are not in conflict (Lehner 2010). Since gene duplication relaxes noise-plasticity conflict (Lehner 2010), we expected duplicated genes to be enriched for TATA motifs when compared to singletons. Of the 1090 genes containing TATA-motifs, 558 belonged to duplicates (281 were WGDs and 277 were SSDs) (25% of all duplicates) and 532 to singletons (12% of

all singletons) (Supplementary Table S10). Indeed, duplicated genes were more enriched for TATA-containing genes than singletons (Fisher's exact test: $F = 2.52$, $P < 2.2 \times 10^{-16}$), and this was the case for both transcriptionally plastic genes and genes with no transcriptional plasticity (Supplementary Figure S1A). Remarkably, duplicates with no transcriptional plasticity were slightly more enriched for TATA-containing genes than singletons with transcriptional plasticity, with the difference being significant in the case of *S. cerevisiae* grown under oxidative stress (Supplementary Figure S1B).

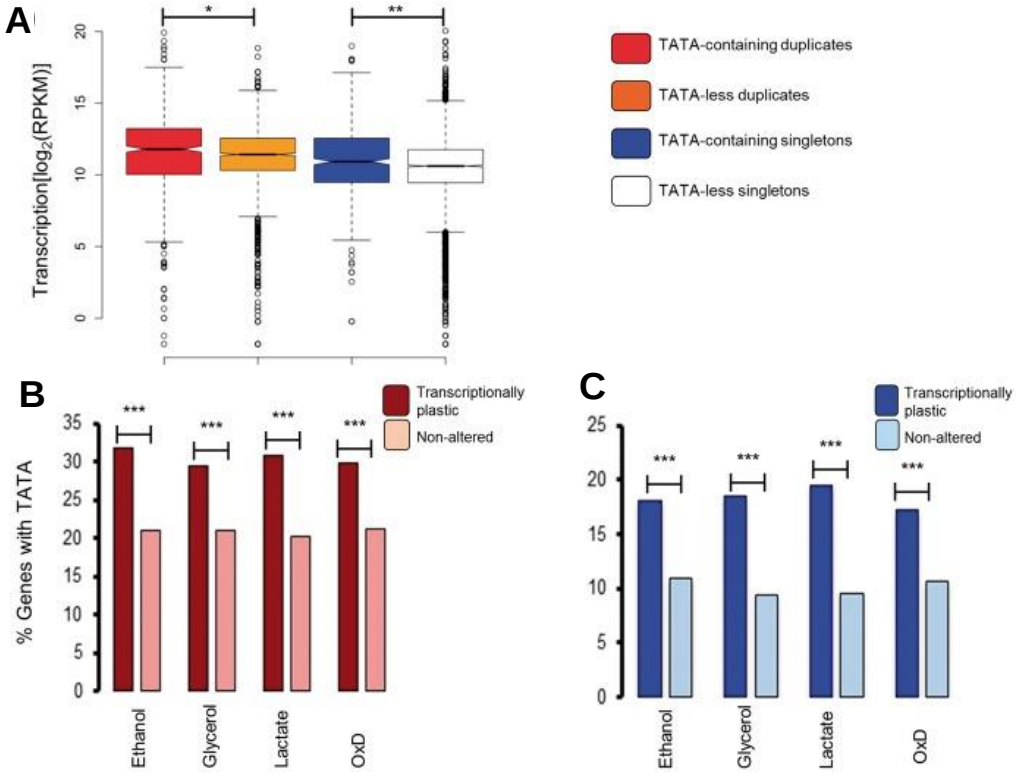


Figure ChI- 4 - TATA-containing genes exhibit greater expression levels and transcriptional plasticity than TATA-less genes. **A)** Comparison of the expression levels (measured as the logarithmic transformation of the Fragment Reads Per billion) between TATA-containing and TATA-less genes for duplicates (red and orange boxes, respectively) and singletons (blue and white boxes, respectively). Differences were tested using the Wilcoxon rank test and significant differences are identified as * and **, for probabilities of $P < 0.05$ and $P < 0.01$, respectively. The percentage of genes with TATA-containing promoters for transcriptionally plastic and transcriptionally unaltered genes when *S. cerevisiae* is growing under four different stress conditions (ethanol, glycerol, lactate and oxidative stress in a medium supplemented with dextrose) was compared for duplicates **B)** and for singletons **C)**. Significant differences between plastic and unaltered genes were identified using Fisher's exact test and are identified as ***, to indicate a probability of $P < 2.2 \times 10^{-16}$.

Finally, the magnitude of expression divergence between duplicates gene copies was correlated with the magnitude of transcriptional plasticity (measured as the fold change in expression of the most altered gene copy between YPD and stress) (Table ChI-1), both of which are in turn correlated with the levels of gene expression.

***Table ChI- 1* - Expression divergence between gene copies correlates with transcriptional plasticity of duplicates**

Stress source	Correlation (Spearman)	Probability
Ethanol	0.17	4.22×10^{-8}
Glycerol	0.20	1.78×10^{-10}
Lactate	0.16	4.13×10^{-7}
Oxidative + Dextrose	0.15	7.11×10^{-7}

d. The levels of gene expression correlate with the patterns of duplicates transcriptional plasticity

A prediction of the dosage sub-functionalization hypothesis is that the patterns of duplicates transcriptional plasticity should be dependent on the levels of gene expression. For instance, transcriptionally plastic genes that are lowly expressed under normal conditions should only be able to over-express under stress because a decline in their expression could drive one of the gene copies to non-functionalization due to relaxed selective constraints. Because plasticity is often correlated with expression noise (Lehner 2010) and, since surviving duplicates are those whose expression noise falls within the range of expression detectable by selection, expression noise should depend on the levels of duplicates expression. We divided duplicated genes according to the patterns of transcriptional plasticity they show when *S. cerevisiae* is grown under stress: (i) up-regulated: when the two gene copies were up-regulated under stress; (ii) down-regulated: when the two gene copies were down-regulated under stress; (iii) discordant: when one copy was up-regulated and the other was down-regulated under stress; and (iv) one-altered: when one copy was not altered but its paralogous copy was either up-regulated or down-regulated under stress (Supplementary Tables S11–S26). In all four stress conditions, duplicates that were down-regulated were also those that exhibited the highest expression levels under normal conditions, being these followed by duplicates in which only one copy exhibits transcriptional plasticity, then discordant duplicates and finally up-regulated duplicates (Figure ChI-5A-D).

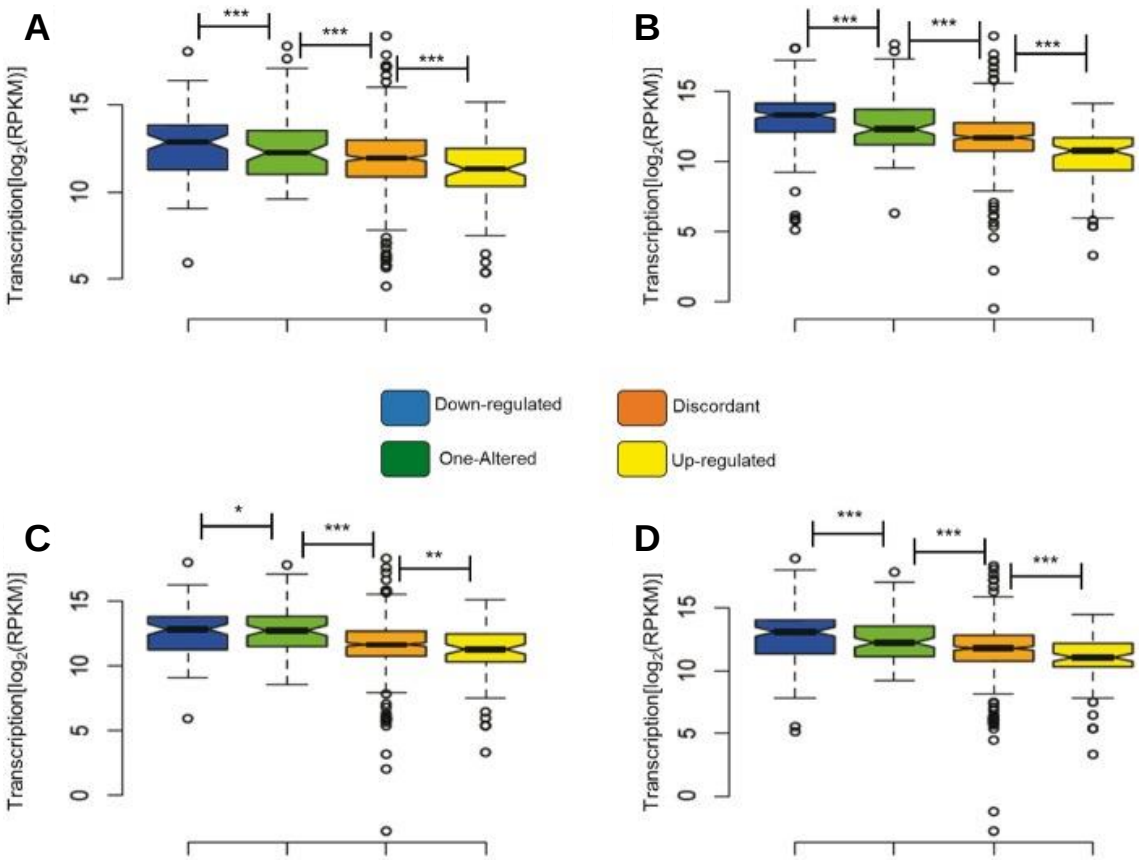


Figure ChI- 5 - The expression level of duplicates correlated with their patterns of expression divergence. Expression was measured as the logarithmic transformation of the Reads Per billion (RPKM). Significant differences in the expression levels of gene sets are indicated as *, ** and ***, referring to probabilities of $P < 0.05$, $P < 0.01$ and $P < 10^{-6}$, using a Wilcoxon rank test. **A)** Expression under ethanol stress, **B)** expression under glycerol stress, **C)** expression under lactate stress and **D)** expression under oxidative stress.

Interestingly, the mean level of expression under normal conditions of duplicates gene copies was correlated with the level of expression divergence of the gene copies under normal conditions in those duplicates that belong to the category discordant and one-altered, those categories with the highest expression divergence between gene copies, but not in those pairs in which both copies were either up-regulated or down-regulated (Figure ChI-6).

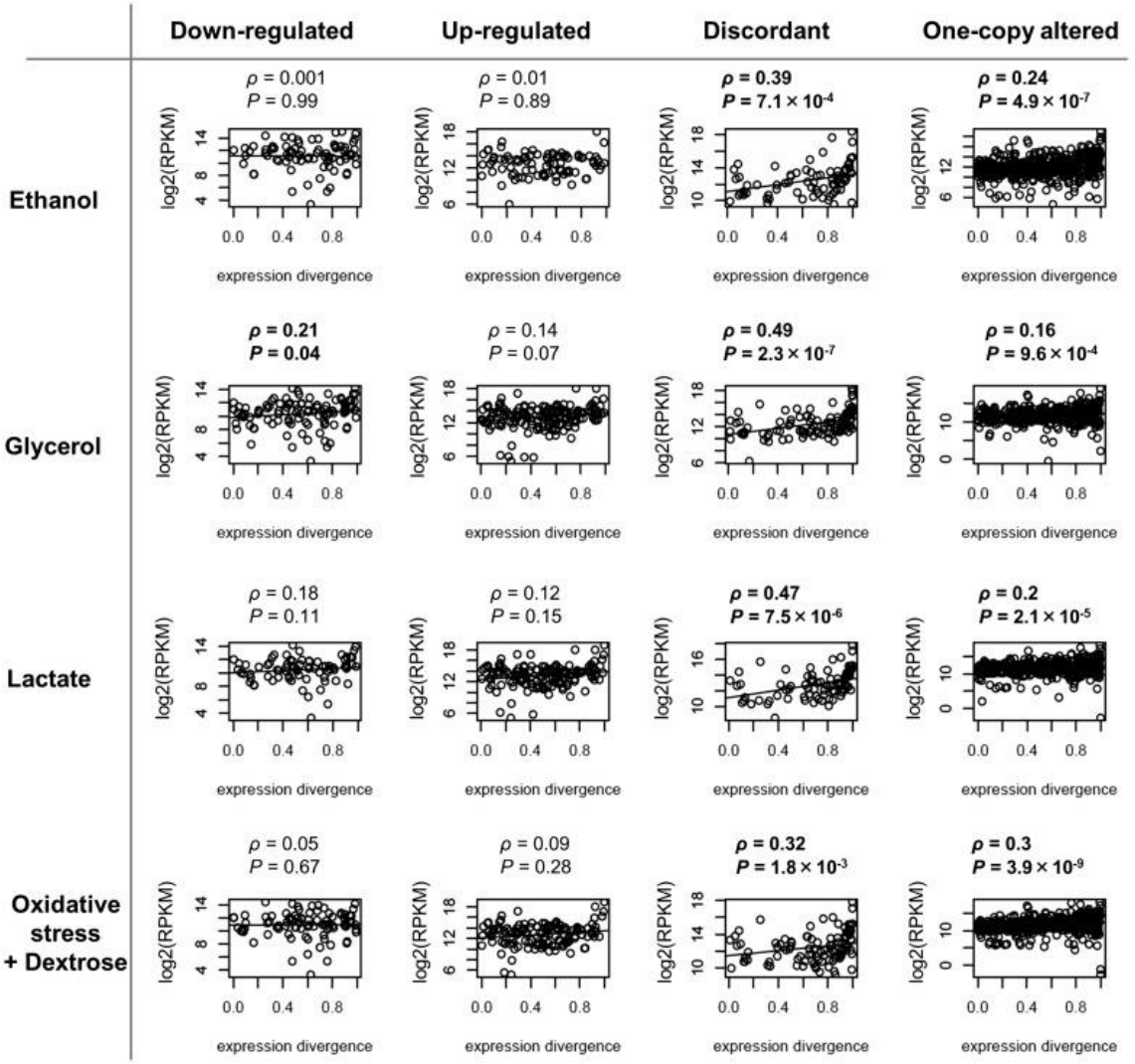


Figure Chl- 6 - The levels of gene expression correlate with the level of expression divergence in duplicates in which gene copies are either discordant (one copy up-regulated and the other copy down-regulated) or one copy only altered under stress. Transcriptional alteration patterns of duplicates were measured in four stress conditions (ethanol, glycerol, lactate and oxidative stress in a medium supplemented with dextrose). Expression divergence of duplicates was measured as 1 – (the expression of the lowest expressed copy divided by the expression of the highest expressed copy). Spearman correlation (ρ) between the levels of gene expression (measured as the logarithmic transformation of the Reads Per billion (RPKM)) and the levels of expression divergence was tested for significance. Significant plots are indicated in bold. We found significant associations between the levels of gene expression and the expression divergence between gene copies of discordant duplicates and those with one copy altered but not for up-regulated or down-regulated duplicates.

e. Gene duplication has contributed to increased transcriptional plasticity in yeast

We examined the link between gene duplication and phenotypic plasticity by comparing the transcriptomes of *S. cerevisiae* grown under a number of key stress conditions that this species faces in nature to those transcriptomes of *S. cerevisiae* grown under normal YPD media (section Material and methods). Transcriptionally

altered genes were more enriched for duplicates than for singleton genes over all stress conditions (Figure ChI-7). This trend was also true when we compared WGDs to singletons and SSDs to singletons (Figure ChI-7). To determine whether transcriptional plasticity is directly linked to gene duplication, we examined the transcriptional plasticity of *S. cerevisiae* duplicates and singletons orthologs in the pre-WGD species *L. kluyveri*. The transcription of *L. kluyveri* genes was previously assessed in 19 different stress conditions (Brion *et al.* 2016). For each condition, we sought the percentage of genes that were orthologs to *S. cerevisiae* duplicates (N = 1469) and singletons (N = 4174) that exhibited transcriptional alteration. In 18 of the 19 conditions, there was no significant difference in the percentage of transcriptionally altered genes under stress between the orthologs of *S. cerevisiae* duplicates and those of singleton (Table ChI-2). The only exception was SDS stress, in which the percentage of transcriptionally altered orthologs for *S. cerevisiae* duplicates was higher than that for transcriptionally altered singleton orthologs (Fisher's exact test: odds ratio F = 1.22, $P = 5 \times 10^{-3}$, Table ChI-2). In all other stresses that were equivalent to the ones used in our experiments (e.g. glycerol, ethanol), there was no significant difference in the number of transcriptionally altered genes between orthologs of *S. cerevisiae* duplicates and singletons (Table ChI-2). These data indicate that the high transcriptional plasticity of duplicates in *S. cerevisiae* was acquired after gene duplication.

To shed more light on the role of gene duplication in the acquisition of transcriptional plasticity, we examined the transcriptional patterns of a post-WGD species, *Candida glabrata*, in which some orthologs of *S. cerevisiae* duplicates are in single gene copy in *C. glabrata*, while others are preserved as duplicates and for which we had transcriptional information under acidic stress similar to our lactate stress dataset (Linde *et al.* 2015). *Saccharomyces cerevisiae* orthologs in *C. glabrata* were identified using synteny information available in the Pillars of the Yeast Gene Order Browser (YGOB) (Byrne and Wolfe 2005).

Table ChI- 2 Transcriptional alterations of L. kluyveri genes orthologs of S. cerevisiae under nineteen different stress conditions.

Stress	# duplicates orthologs (%)	# singletons orthologs (%)	Odds´ratio *F*	*P*
Galactose	588 (40.1)	1586 (37.6)	0.93	0.24
Glycerol	865 (58.8)	2378 (56.9)	1.08	0.21
23 °C	178 (12.1)	540 (12.9)	0.92	0.44
37 °C	458 (31.2)	1282 (30.7)	1.02	0.74
YNB	487 (33.2)	1366 (32.7)	1.02	0.77
Ethanol	609 (41.5)	1645 (39.4)	1.09	0.17
Methanol	291 (19.8)	865 (20.7)	0.95	0.49
SDS	410 (27.9)	1008 (24.1)	1.22	0.005
DMSO	654 (44.5)	1758 (42.1)	1.10	0.11
NaCl	234 (15.9)	757 (18.1)	0.86	0.06
$CaCl_2$	874 (59.5)	2499 (59.9)	0.98	0.80
$NiSO_4$	129 (8.8)	332 (7.9)	1.11	0.32
LiCl	253 (17.2)	737 (17.7)	0.97	0.72
$CoSO_4$	825 (56.2)	2362 (56.6)	0.99	0.85
BME	375 (25.6)	1102 (26.4)	0.96	0.53
5FU	298 (20.3)	922 (22.1)	0.89	0.15
Arsenic	127 (8.6)	343 (8.2)	1.06	0.62
6AU	230 (15.7)	674 (16.1)	0.96	0.68
Fluconazole	437 (29.7)	1195 (28.6)	1.06	0.42

In total, we identified 4844 reliable *S. cerevisiae: C. glabrata* orthologs. Of these 4844 genes, 788 genes were duplicated in *C. glabrata*, of which 123 were orthologs of *S. cerevisiae* singletons. These 4844 genes included 1659 out of the 2240 duplicated *S. cerevisiae* genes and 3185 singletons. Of the 2240 *S. cerevisiae* duplicates, orthologs for 1019 of them were in single gene copy in *C. glabrata* (Figure ChI-7B). Importantly, these 1019 genes exhibited as much transcriptional plasticity under acidic stress in *C. glabrata* (N = 599, 58.7% of the *C. glabrata* singletons) as *C. glabrata* singletons that had no duplicates orthologs in *S. cerevisiae* (1725 out of 3062 singletons, 56.3%) (Fisher's exact test: odd's ratio F = 1.10, P = 0.17). In conclusion, transcriptional plasticity seem to have been acquired after gene duplication because orthologs of duplicates that are in single copy genes in other species exhibit no evidence for transcriptional plasticity in these species.

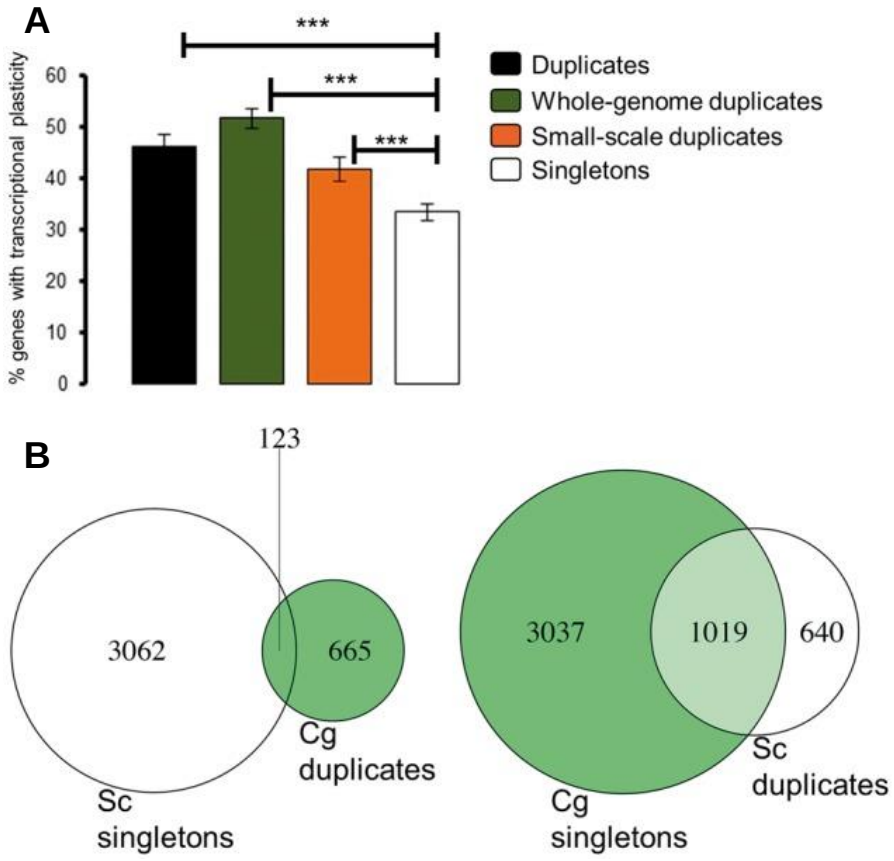


Figure ChI- 7 - Duplicates are more enriched for transcriptional plasticity than singletons, regardless of the mechanism of duplication. **A)** Transcriptional plasticity of duplicates and singletons in *S. cerevisiae*. Y-axis represents the mean percentage of genes with altered transcription across four stress conditions (ethanol, glycerol, lactate and oxidative stress in a medium supplemented with dextrose). We also compared WGDs and SSDs to singletons. Significant differences were found using Fisher's exact test, with *** indicating a probability of $P < 2.2 \times 10{-}16$. **B)** Venn diagram representing the overlap in the number of duplicates and singletons between *C. glabrata* and *S. cerevisiae*.

f. Transcriptionally plastic duplicates contribute to the response of S. cerevisiae to stress

Among the transcriptionally plastic duplicates, many had a significant biological role in the response to stress. For instance, analyses of duplicated genes up-regulated when *S. cerevisiae* is grown in ethanol identified a number of genes as largely up-regulated that are involved in ethanol metabolism (Supplementary Table S27). Heading the list of up-regulated duplicates is the one encoding the alcohol dehydrogenase ADH2 and ADH1, directly involved in the quick metabolism of ethanol inside the cell into acetaldehyde (Figure ChI-8). A number of other duplicated genes that are essential in the metabolism of Ethanol, including transmembrane transporters such as YAT1 to start the tricarboxylic cycle or the enzyme MSL1 that is essential for malate production, among

others (Figure ChI-8), they are all listed as duplicates with the highest up-regulation. This also applies to other stress conditions used in this study (Figure ChI-8).

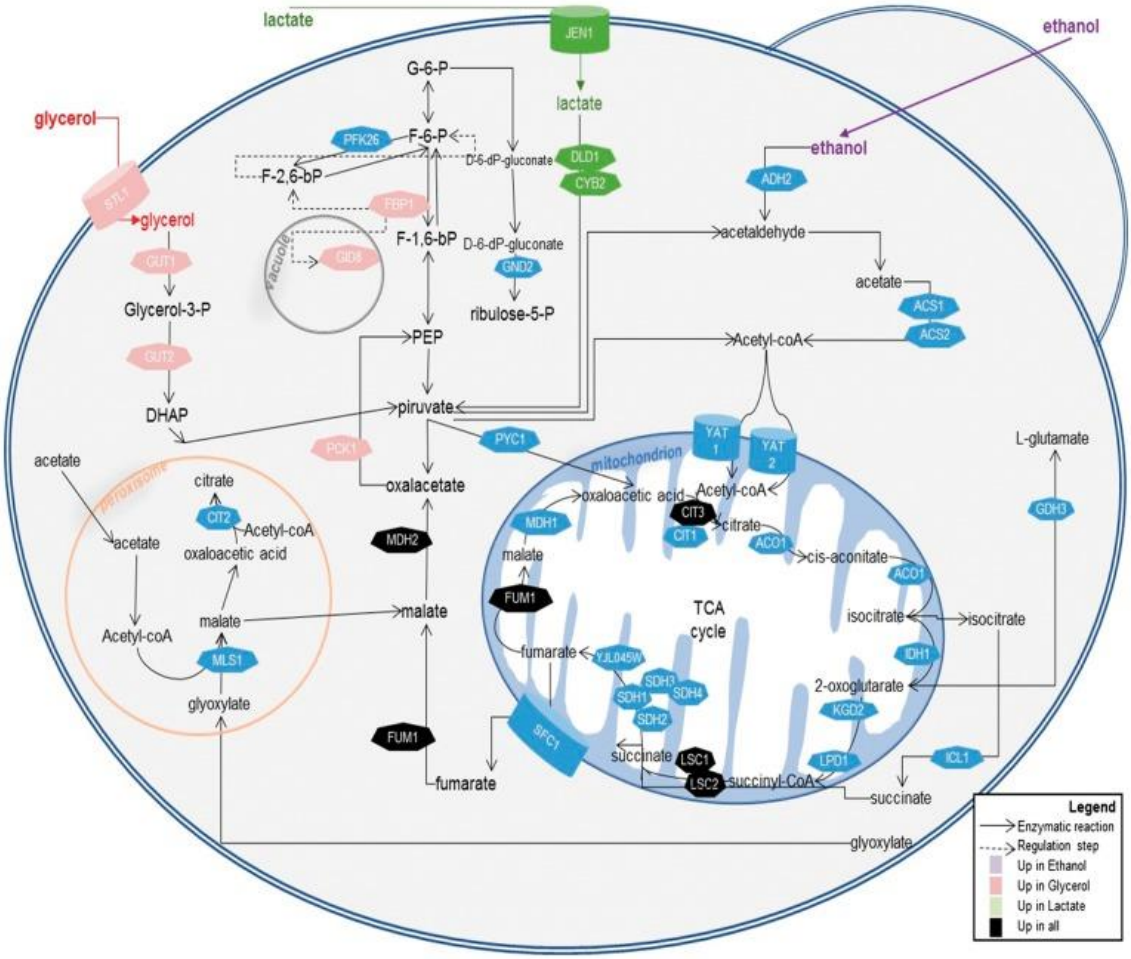


Figure ChI- 8 - Schematic representation of S. cerevisiae central metabolic pathways for the utilization of non-fermentable carbon sources (glycerol, lactate and ethanol). Solid arrows correspond to enzymatic reactions while dashed arrows correspond to regulatory steps. Proteins encoded by duplicated genes that are transcriptionally altered and are involved in each process are indicated (transporters are shown with a can form, and enzymes exhibit heptagonal forms).

5. Discussion

A number of factors have been considered key players in determining the fates of duplicated genes. However, an explanation that provides a general and inherent mechanism that strongly determines the duplicability of genes has been poorly investigated. Here, we present strong evidence that the levels of gene expression determine the likelihood of a gene to persist duplicated in the genome. We find that genes that are highly expressed in species pre-dating the whole-genome duplication event in *Saccharomyces* are more likely to be preserved in duplicate in species that originated after duplication. This observation is in agreement with the dosage sub-functionalization hypothesis (DSH), as high levels of gene expression would ensure purifying selection on the gene copies despite the stochastic variation in their expression levels (Gout and Lynch 2015). We, nevertheless, show that this is not a link exclusively seen in WGDs because the preservation of SSDs is also dependent on the expression levels of genes, however to a lesser extent than WGDs.

We also show that highly expressed genes exhibit greater phylogenetic stability (i.e. they are preserved in a greater number of post-duplication species) than lowly expressed genes, perhaps due to the higher likelihood of functional divergence between gene copies of highly expressed duplicates, and thus the emergence of purifying selection to maintain the functions of gene copies. Therefore, we conclude that the levels of gene expression determine the survival of duplicates across long evolutionary times. Our observations are not confounded by other factors such as dosage balance, mainly an issue in the case of genes encoding proteins that are part of protein complexes (Gout *et al.* 2009, 2010; Qian *et al.* 2010; Birchler and Veitia 2012; Gout and Lynch 2015), as neither WGDs nor SSDs are enriched for complex-encoding genes when compared to singletons in *S. cerevisiae*.

A pivotal conclusion derived from the link between gene expression and duplicability is that gene copies with higher expression levels can be noisier in terms of expression without undergoing an imbalance in their selective pressures. This noise could eventually lead to higher divergence between the gene copies of these duplicates when such divergence becomes adaptive. Our results are in agreement with this prediction and show that the magnitude of expression divergence, and perhaps functional divergence, is dependent on the levels of gene expression. Therefore, duplicates with higher expression are more likely to lead to higher divergence between the gene copies. This higher divergence between gene copies has been likely important for the origin of novel adaptations to stressful environments in the yeast *Saccharomyces cerevisiae* (Keane et al. 2014; Fares 2015c; Mattenberger et al. 2017a; Mattenberger et al. 2017b). In

agreement with this hypothesis, when *S. cerevisiae* was faced with a number of stress conditions, the levels of expression divergence between gene copies was positively correlated with their expression plasticity. Importantly, we show that the levels of gene expression also influence the patterns of transcriptional divergence between the gene copies of duplicates. Those duplicates with copies exhibiting discordant expression patterns or with only one copy altered under stress are likely those in which each copy encodes a function under different conditions compared to the other copy.

Expression noise is generally coupled with plasticity if noise and plasticity do not present a cost-benefit conflict (i.e. a conflict emerges when plasticity is important for adaptation but noisy expression can be detrimental) (Raser and O'Shea 2005; Blake *et al.* 2006; Lehner 2010). We show that duplicates are both more enriched for transcriptionally plastic genes and for genes with TATA motifs in their promoters when compared with singletons. TATA-motifs have been shown to be associated with higher noise and plasticity in TATA-containing genes (Newman *et al.* 2006; Tirosh *et al.* 2006; Landry *et al.* 2007). The expression properties of genes that have been preserved as duplicates are, therefore, remarkably different from those that returned to single copy genes. Noticeably, non-plastic duplicates that showed higher expression levels than transcriptionally plastic singletons also exhibited greater enrichment for TATA-containing genes than plastic singletons, linking TATA-containing genes to higher levels of gene expression. The question that remains is then whether it is the expression properties or duplication itself that have determined the fate of duplicates and the origin of adaptations. Our results lead to the conclusion that duplicated genes exhibit significantly different expression properties than singleton genes but also that gene duplication is mainly responsible for the origin of plasticity, as the transcriptional plasticity observed in *S. cerevisiae* originated after the duplication of genes which were not transcriptionally more plastic before duplication than other non-duplicable genes. Therefore, gene duplication provides the appropriate genetic and selective opportunity for the evolution of transcriptional plasticity. A remarkable result is that the percentage of singletons transcriptionally altered in *S. cerevisiae* was significantly lower than that of their orthologs altered in *L. kluyveri*. However, the absolute percentages of altered genes cannot be compared between *L. kluyveri* and *S. cerevisiae* as both species represent completely different metabolisms and the conditions of stress under which they were subjected were different. It is, therefore, likely that duplication itself may relax the cost–benefit conflict between noise and plasticity, as previously suggested (Lehner 2010), allowing for the emergence of plasticity and adaptation to environmental perturbations.

In conclusion, our result point to a strong correlation between the expression properties of genes, their duplicability, transcriptional plasticity, and ability to give rise to novel adaptations.

6. Accession numbers

All newly sequenced RNA sequences are available from the Sequence Read Archive with the following accession number (SRP074821).

7. Supplementary data

Supplementary data are available at *DNA research* online. Data set 1 and supplementary figure S1 are also available in the appendix of this manuscript. **Data Set 1:** Transcription levels correlate with expression levels in *S.cerevisiae*. **Figure S1:** Duplicates contain a greater proportion of genes with TATA motifs in their promoters than singletons. **Table S1:** Expression data (presented as Read per billion: RPKM) for *Saccharomyces cerevisiae* growing under either YPD or Ethanol. **Table S2:** Expression data (presented as Read per billion: RPKM) for *Saccharomyces cerevisiae* growing under either YPD or Glycerol. **Table S3:** Expression data (presented as Read per billion: RPKM) for *Saccharomyces cerevisiae* growing under either YPD or Lactate. **Table S4:** Expression data (presented as Read per billion: RPKM) for *Saccharomyces cerevisiae* growing under either YPD or under oxidative stress in a growth medium supplemented with dextrose. **Table S5:** Expression data (presented as the log2(Reads Per billion)) for *Lachancea kluyveri* in Glucose and 19 different stress growth conditions. The orthologs for *L. Kluyveri* genes in *S. cerevisiae* are also presented. **Table S6:** Fold change (FC) of expression of *Candida glabrata* genes when comparing expression in glucose and M199 medium. Only *C. Glabrata* genes for which orthologs exist in *Saccharomyces cerevisiae* are presented. **Table S7:** Open Reading Frames (ORFs) of *S. cerevisiae* encoding proteins known to be part of protein complexes. **Table S8:** Phylogenetic distribution of duplicated genes among the 12 post-whole genome duplication species in *Saccharomycotina*. **Table S9:** Expression divergence (D_{ij}) between the gene copies (A and B) of Duplicates *S. cerevisiae* growing under YPD. **Table S10:** Genes containing TATA motifs in the promoters in any of the post-WGD species and their nature as duplicates from whole-genome duplication (WGD) or small-scale duplication (SSD). **Table S11:** Expression of duplicates in which the two gene copies are up-regulated when *Saccharomyces cerevisiae* is faced with ethanol stress and the expression divergence between the two gene copies (a and b). **Table S12:** Expression of duplicates in which the two gene copies are down-regulated when *Saccharomyces cerevisiae* is faced with ethanol stress and the expression divergence between the two gene copies (a and b). **Table S13:** Expression of duplicates in which one of the gene copies is altered in

expression when *Saccharomyces cerevisiae* is faced with ethanol stress and the expression divergence between the two gene copies (a and b). **Table S14:** Expression of duplicates in which one of the gene copies is up-regulated and the other down-regulated when *Saccharomyces cerevisiae* is faced with ethanol stress and the expression divergence between the two gene copies (a and b). **Table S15:** Expression of duplicates in which the two gene copies are up-regulated when *Saccharomyces cerevisiae* is faced with glycerol stress and the expression divergence between the two gene copies (a and b). **Table S16:** Expression of duplicates in which the two gene copies are down-regulated when *Saccharomyces cerevisiae* is faced with glycerol stress and the expression divergence between the two gene copies (a and b). **Table S17:** Expression of duplicates in which one of the gene copies is altered in expression when *Saccharomyces cerevisiae* is faced with glycerol stress and the expression divergence between the two gene copies (a and b). **Table S18:** Expression of duplicates in which one of the gene copies is up-regulated and the other down-regulated when *Saccharomyces cerevisiae* is faced with glycerol stress and the expression divergence between the two gene copies (a and b). **Table S19:** Expression of duplicates in which the two gene copies are up-regulated when *Saccharomyces cerevisiae* is faced with lactate stress and the expression divergence between the two gene copies (a and b). **Table S20:** Expression of duplicates in which the two gene copies are down-regulated when *Saccharomyces cerevisiae* is faced with lactate stress and the expression divergence between the two gene copies (a and b).**Table S21:** Expression of duplicates in which one of the gene copies is altered in expression when *Saccharomyces cerevisiae* is faced with lactate stress and the expression divergence between the two gene copies (a and b). **Table S22:** Expression of duplicates in which one of the gene copies is up-regulated and the other down-regulated when *Saccharomyces cerevisiae* is faced with lactate stress and the expression divergence between the two gene copies (a and b).**Table S23:** Expression of duplicates in which the two gene copies are up-regulated when *Saccharomyces cerevisiae* is faced with oxidative stress and the expression divergence between the two gene copies (a and b).**Table S24:** Expression of duplicates in which the two gene copies are down-regulated when *Saccharomyces cerevisiae* is faced with oxidative stress and the expression divergence between the two gene copies (a and b). **Table S25:** Expression of duplicates in which one of the gene copies is altered in expression when *Saccharomyces cerevisiae* is faced with oxidative stress and the expression divergence between the two gene copies (a and b). **Table S26:** Expression of duplicates in which one of the gene copies is up-regulated and the other down-regulated when *Saccharomyces cerevisiae* is faced with oxidative stress and the expression divergence between the two gene copies (a and b). **Table S27:** Duplicated

genes in which at least one gene copy up-regulated when *Saccharomyces cerevisiae* is faced with ethanol stress.

CHAPTER II – The roles of phenotypic plasticity and duplicated genes in the cellular response to environmental stress, adaptation, and biological innovation.

A version of this chapter has been published as:

Mattenberger F., Sabater-Muñoz B., Toft C., Fares M.A. (2017) *The phenotypic plasticity of duplicated genes in Saccharomyces cerevisiae and the origin of adaptations.* ***G3: Genes, Genomes, Genetics, 7(1): 63-75.***

1. Abstract

Gene and genome duplication are the major sources of biological innovations in plants and animals. Functional and transcriptional divergence between the copies after gene duplication has been considered the main driver of innovations. However, here we show that increased phenotypic plasticity after duplication plays a more major role than thought before in the origin of adaptations. We perform an exhaustive analysis of the transcriptional alterations of duplicated genes in the unicellular eukaryote *Saccharomyces cerevisiae* when challenged with five different environmental stresses. Analysis of the transcriptomes of yeast shows that gene duplication increases the transcriptional response to environmental changes, with duplicated genes exhibiting signatures of adaptive transcriptional patterns in response to stress. The mechanism of duplication matters, with whole-genome duplicates being more transcriptionally altered than small-scale duplicates. The predominant transcriptional pattern follows the classic theory of evolution by gene duplication; with one gene copy remaining unaltered under stress, while its sister copy presents large transcriptional plasticity and a prominent role in adaptation. Moreover, we find additional transcriptional profiles that are suggestive of neo- and subfunctionalization of duplicate gene copies. These patterns are strongly correlated with the functional dependencies and sequence divergence profiles of gene copies. We show that, unlike singletons, duplicates respond more specifically to stress, supporting the role of natural selection in the transcriptional plasticity of duplicates. Our results reveal the underlying transcriptional complexity of duplicated genes and their role in the origin of adaptations.

2. *Introduction*

Gene duplication has been a major driving force of biological innovation in plants (Wendel 2000; Otto and Whitton 2000; Holub *et al.* 2001; Lespinet *et al.* 2002; Kim *et al.* 2004; Cui *et al.* 2006; Carretero-Paulet and Fares 2012) and animals (Otto and Whitton 2000; Hoegg *et al.* 2004). Arguably, understanding how gene duplication gives origin to novel functions and adaptations is a fundamental aim of evolutionary biology. The functional and transcriptional divergence between the gene copies of a duplicated gene has been proposed to facilitate the origin of novel functions (Ohno 1970, 1999; Lynch and Conery 2000; Conant and Wolfe 2008). However, the tempo and mode of each divergence kind and the interplay between both remains largely unexplored.

Ohno proposed that after the duplication of a gene, the emerging genetic redundancy leads to relaxed selection against one of the gene copies while the other copy remains under strong purifying selection (Ohno 1970, 1999). The selectively relaxed gene copy explores novel genotypes, many of which will be deleterious and lead to the loss of the rapidly evolving gene copy (Lynch and Conery 2003). A less likely scenario is the preservation of both copies by purifying selection after a period of relaxed selection leading to novel functions in the form of sub- or neo-functionalization (Ohno 1970, 1999; Lynch and Conery 2003; Taylor and Raes 2004). Particular scenarios to this general model for the functional divergence of gene copies have been proposed (Force, Lynch, Pickett, *et al.* 1999; Des Marais and Rausher 2008; Innan and Kondrashov 2010). The classic theory has also given credit to the expression divergence between gene copies as a pre-requisite for the preservation of genes in duplicate and the eventual finding of new functions (Ohno 1970; Ferris and Whitt 1979; Force, Lynch, Pickett, *et al.* 1999). Moreover, previous studies have found a genome-wide transcriptional response of *S. cerevisiae* to a wide range of environmental perturbations (Ferea *et al.* 1999; Causton *et al.* 2001; Ideker *et al.* 2001; Landry *et al.* 2006; Stern *et al.* 2007; Cormier *et al.* 2010).

The rapid evolution of gene expression after duplication (Li *et al.* 2005; Thompson *et al.* 2013) suggests an adaptive role for the transcriptional plasticity of duplicates. However, it remains open the question of whether duplicates follow the general response patterns to stresses that are shown by singleton genes or, alternatively, they have allowed the origin of stress-specific adaptations that have been favored by natural selection. It also remains obscure whether the transcriptional plasticity of duplicates has driven their functional specialization. Understanding this plasticity through studies like the one conducted here provides a much wider picture of the role of gene duplication in the origin of adaptations and ecological diversification.

Gene duplication in plants has been followed by rapid expression divergence between gene copies (Blanc and Wolfe 2004b; Ha *et al.* 2007, 2009; Wang *et al.* 2012). Since most duplicated genes are thought to mediate the interaction between the organism and environment, their expression changes have been suggested to be strongly linked to generating adaptations rather than responding to developmental perturbations (Ha *et al.* 2007). Most importantly, expression divergence has been seen to correlate with the sequence divergence between duplicates gene copies in plants (Blanc and Wolfe 2004b) and, although less clearly (Wagner 2000), in yeast (Gu *et al.* 2002). Two questions remain unexplored: (a) are duplicated genes more transcriptionally plastic than anticipated? And (b) does transcriptional plasticity determine the functional fates of gene copies? Answering these questions would reveal the potential of gene duplicates to expedite adaptations.

The Baker's yeast *S. cerevisiae* has duplicated its genome roughly 100 MYA (Wolfe and Shields 1997) triggered by the possible hybridization between different yeast species (Marcet-Houben and Gabaldón 2015; Wolfe 2015). Only 1120 pairs of duplicates have been retained, of which 554 belong to the whole-genome duplication event and the remaining are classified as duplications of small scale (Fares *et al.* 2013). Many of the yeast-duplicated genes enable the growth of *S. cerevisiae* under stressful conditions, the genetic basis of which has enabled the exploitation of the biotechnological benefits of yeast in the multimillionaire wine industry. The genetic and biotechnological properties of this yeast offer a unique opportunity to study the role of gene duplication in innovation. In this study, we explore whether the transcriptional plasticity of duplicated genes in *S. cerevisiae* has contributed to the origin of adaptations to stress and functional specialization of duplicates gene copies. We address this question by exhaustively and extensively analyzing the expression pattern dynamics of duplicated genes in the yeast *S. cerevisiae* after subjecting it to a number of stress conditions. Here, we find that not only duplicates are more transcriptionally polymorphic as concluded before (Ha *et al.* 2009) but that they are more transcriptionally plastic than singletons under environmental stress. This transcriptional plasticity increases after gene duplication and it is strongly correlated with the functional divergence of duplicates gene copies. The study of the patterns of sequence divergence, functional interactions, and transcriptional plasticity of duplicates makes it possible the identification of stress-specific as well as general transcriptional response patterns. We show that, unlike singleton genes, duplicates have given origin to stress-specific adaptations. Our data describe a complex dynamic of transcriptional evolution following the gene and genome duplications of a simple eukaryotic organism and reveal the origins of yeast adaptations.

3. Material and Methods

a. Identification of duplicated genes

Paralogs pairs of duplicated genes were identified as the resulting best reciprocal hits from all-against-all BLAST searches using BLASTP with an E-value cutoff of $1E^{-5}$ and a 50-bit score (Altschul *et al.* 1997). Paralogs were then divided into two groups according to the mechanism of their origin: WGDs and SSDs. WGDs are those extracted from the reconciled list provided by the YGOB (Yeast Gene Order Browser, http://wolfe.gen.tcd.ie//ygob, Byrne and Wolfe 2005) (555 pairs of genes), and these were not subjected to subsequent SSD. All other paralogs were considered to belong to the category of SSDs (560 pairs of genes). The duplicates used in this study have been estimated to have their origin on the time point of the whole genome duplication that took place 100 MYA (Wolfe and Shields 1997). Also, in this study we have used the SSDs that exhibit a similar distribution of synonymous substitutions as those of WGDs, so roughly belonging to the same age (Fares *et al.* 2013; Keane *et al.* 2014).

b. Sequence Alignments and analysis of divergence

For each protein-coding gene of *S. cerevisiae* we searched for its orthologue in the closely related species *Saccharomyces paradoxus* using the program blastP. Pairwise sequence alignments were built using the program ClustalW. To calculate the distance between *S. cerevisiae* and *S. paradoxus* for each of the genes, we estimated the number of non-synonymous nucleotide substitutions per non-synonymous site (d_N), synonymous substitutions per synonymous site (d_S), and the non-synonymous-to-synonymous rates ratio ($\omega = d_N/d_S$) using the maximum-likelihood approach under the Goldman and Yang model (Goldman and Yang 1994) as implemented in the PAML package version 4.7 (Yang 2007).

c. Analysis of Gene Expression in S. cerevisiae

The transcriptomic profiling was performed in the *S. cerevisiae* Y06240 haploid *msh2* deletion strain (BY4741; *Mat**a***; *his3D1*; *leud2DO*; *met15DO*; *ura3DO*; *msh2::kanMX4*) (Fares *et al.* 2013), with three technical replicates for each biological stress condition (3% lactic acid {YPL}, 3% ethanol {YPE}, 3% glycerol {YPG}, 0.25mM H_2O_2 {YPOx}, 0.25mM H_2O_2 + 1.5% dextrose {YPOxD}) in comparison with the normal growth condition (YPD media). Total RNA extractions were performed with RNeasy kit (Qiagen) following manufacturer instructions. Ribosomal RNA was removed by using Ribo-Zero Gold rRNA removal yeast (Illumina) depletion kit. Stranded RNA libraries were constructed using TruSeq stranded mRNA (Illumina) from oligo-dT captured mRNAs from depleted

samples. Libraries were run in NextSeq 500 (Illumina) at 75nt single read by using High Output 75 cycles kit v2.0 (Illumina).

The treatment of the RNA libraries was done following a previous study in which different methods of differential expression analyses were compared (Zhang *et al.* 2014). RNA libraries were sequenced at the Genomic core facility at Servicio Central de Soporte a la Investigación Experimental (SCSIE) from University of Valencia, Spain. Raw reads were analyzed using FastQC report and cleaned with CutAdapt as implemented in RobiNA software package v 1.2.4 (Lohse *et al.* 2012). Low-quality reads were filtered and trimmed (Phred score inferior to 20 and size less than 40nt were discarded). Since we had a reference transcriptome from the S288c strain, reads were then aligned with Bowtie (up to 2 mismatches accepted) to the reference transcriptome (PRJNA290217) from the reference S288c strain. Statistical assessment of differential gene expression was done either with edgeR (Robinson *et al.* 2010) and with DESeq (Anders and Huber 2010) as implemented in RobiNA. A previous study compared the different expression analysis methods, concluding that edge R and DeSeq were the best-performing methods when the objective is to analyze differential expression (Zhang *et al.* 2014). Comparison of logarithmic fold change of our expression data between edgeR and DESeq provided very strong correlation (Spearman correlation coefficient: $\rho = 0.995$, $P < 2.2 \times 10^{-16}$, Figure 1 of File S8). Significant expression changes were identified using a false discovery rate (FDR < 0.05). These results indicate that our quantification of expression data is robust to the method used. All newly sequenced RNA sequences are available from the Sequence Read Archive with the following accession number (SRP074821).

d. Genetic interaction data

We used the latest update of the genetic functional chart of *S. cerevisiae* (Costanzo *et al.* 2010) (Supplemental files S4 and S5 from http://drygin.ccbr.utoronto.ca/~costanzo/). The genetic map is based on the synthetic genetic array methodology (Tong *et al.* 2001). In this methodology, synthetic lethal genetic interactions are systematically mapped to single and double mutants. In this study, two genes are considered to interact genetically if the double knockout mutant of the two genes has a significantly larger or smaller effect than the multiplicative effects of simple knockouts.

e. Software

Calculations and statistics were performed using MS Excel and R 3.2.1. Data management was possible using in-house built PERL scripts.

4. Results

To test the role of duplicated genes of *S. cerevisiae* in the origin of adaptations, we sequenced the transcriptome of a haploid *msh2* deletion strain after growing it in normal YPD medium and in five different stress conditions: (a) Ethanol, (b) Glycerol, (c) Lactate, (d) Oxidative stress, and (e) Oxidative stress in a medium supplemented with dextrose (Methods). Subsequently, we compared the transcriptional modifications of *S. cerevisiae msh2::kanMX4* under each of the conditions and sought to investigate the role of duplicated genes in displaying transcriptional plasticity under stress. We used this strain because it has been evolved for hundreds of generations in YPD medium, hence is adapted to this medium, and allows the maintenance of population genetic polymorphism due to its higher mutation rate compared to the wild type strain.

a. Duplicated genes exhibit significant transcriptional plasticity under stress

After growing biological replicates of the yeast populations under normal and each of the five different stress conditions, we extracted total RNA for RNAseq library construction and identified differentially expressed (DE) genes based on the comparison of their expression levels relative to normal conditions (Material and Methods). In total, we obtained reliable RNA sequence data for 5825 genes in the YPD medium (normal conditions) and each of the five stress conditions (Files S1 to S5), of which an important fraction was significantly altered under stress conditions (Table ChII-1).

In all stress conditions, there was a significant transcriptomic response affecting 2248, 2827, 2509, 101, and 2638 genes under stress induced by ethanol, glycerol, lactate, oxidative stress, and oxidative stress in a medium supplemented with dextrose, respectively (Tables S1 to S5). The response was more significantly affecting duplicates than singletons in all stress conditions; although the low number of altered genes limited the power of the test in the case of oxidative stress, (Table ChII-1 and Figure ChII-1A). The mechanism of duplication, including whole-genome duplication (WGD) and small-scale duplication (SSD), also made a difference, with WGDs being more significantly enriched for altered-expression genes than SSDs (Table ChII-1 and Figure ChII-1B).

Table ChII- 1 - Transcription alterations under stress condition.

Stress	Comparison	Number of Genes of first type (%)	Number of genes of second type (%)	Odds ratio (F)	Probability
Ethanol	D[a] vs S[b]	907(40.5%)	1341(29.3%)	1.64	$< 2.2x10^{-16}$
	WGDs[c] vs S	515(47.6%)	1341(29.3%)	2.12	$< 2.2x10^{-16}$
	SSDs[d] vs S	392(33.9%)	1341(29.3%)	1.27	$8.1x10^{-4}$
	WGDs vs SSDs	515(47.6%)	392(33.9%)	1.68	$2.7x10^{-9}$
Glycerol	D vs S	1134(50.6%)	1693(36.9%)	1.75	$< 2.2x10^{-16}$
	WGDs vs S	617(56.9%)	1693(36.9%)	2.18	$< 2.2x10^{-16}$
	SSDs vs S	517(44.7%)	1693(36.9%)	1.41	$2.3x10^{-7}$
	WGDs vs SSDs	617(56.9%)	517(44.7%)	1.10	0.21
Lactate	D vs S	1038(46.3%)	1471(32.1%)	1.83	$< 2.2x10^{-16}$
	WGDs vs S	571(52.7%)	1471(32.1%)	2.28	$< 2.2x10^{-16}$
	SSDs vs S	467(40.4%)	1471(32.1%)	1.47	$2.6x10^{-8}$
	WGDs vs SSDs	571(52.7%)	467(40.4%)	1.56	$2.3x10^{-7}$
Oxidative	D vs S	42(1.9%)	59(1.3%)	1.46	0.06
	WGDs vs S	29(2.7%)	59(1.3%)	2.07	0.002
	SSDs vs S	12(1.1%)	59(1.3%)	0.82	0.65
	WGDs vs SSDs	29(2.7%)	12(1.1%)	2.54	0.007
Oxidative + Dextrose	D vs S	1064(47.5%)	1574(34.4%)	1.73	$< 2.2x10^{-16}$
	WGDs vs S	589(54.4%)	1574(34.4%)	2.20	$< 2.2x10^{-16}$
	SSDs vs S	475(41.1%)	1574(34.4%)	1.36	$5.1x10^{-6}$
	WGDs vs SSDs	589(54.4%)	475(41.1%)	1.61	$2.2x10^{-8}$

[a]Duplicated genes; [b] Singleton genes; [c] Whole-genome Duplicates; [d] Small-scale Duplicates

Taking all non-redundant transcriptomic responses together for all stress conditions, duplicated genes showed significantly larger increments of transcription under stress than singleton genes (Figure ChII-1B). Indeed, on average 837 duplicated genes out of the 2240 duplicates (37.4%) in *S. cerevisiae* exhibited significant increments of expression against 1227 out of 4580 singletons (26.8%) (Fisher´s exact test: odds ratio

F = 1.63, P < 2.2 x 10^{-16}). We identified an average over all the stresses of 464 duplicates out of 1100 WGDs, 42.2%, to be transcriptionally plastic, a proportion significantly higher than that for singletons (Fisher´s exact test: Odds ratio: F = 1.99, P < 2.2 x 10^{-16}). Likewise, the proportion of transcriptionally plastic SSDs (an average of 372 out of 1140 SSDs, 32.6%) was significantly higher than that for singletons (Fisher´s exact test: Odds ratio: F = 1.32, P = 1.1 x 10^{-4}). WGDs also presented significantly higher proportion of transcriptionally plastic genes than SSDs (Fisher´s exact test: odds ratio F = 1.51, P = 3.5 x 10^{-6}).

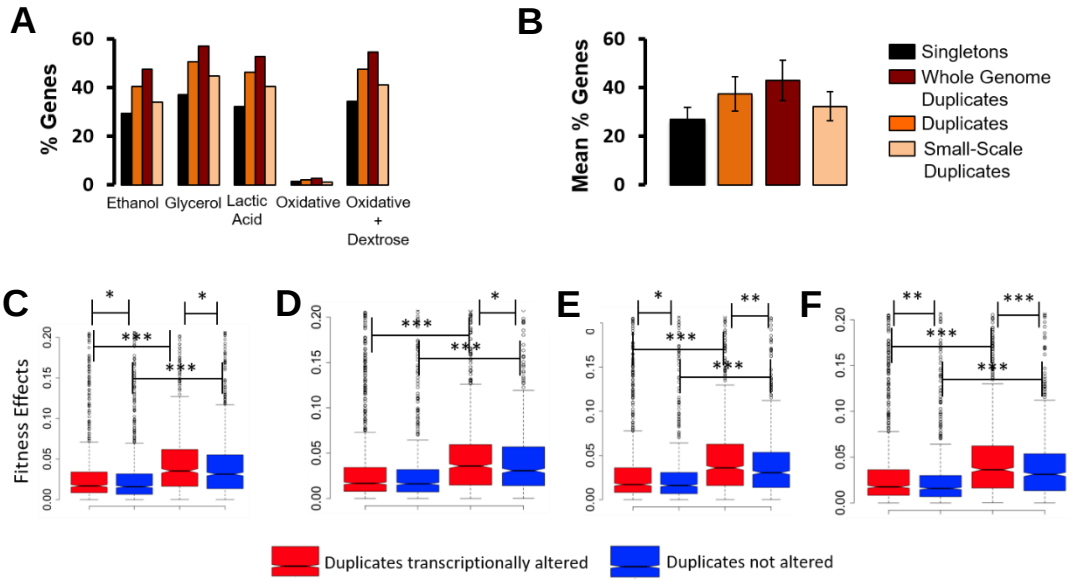


Figure ChII- 1- Duplicated genes exhibit higher transcriptional plasticity than singleton genes and are involved in adaptation. **A)** Percentage of genes with transcriptional flexibility when *S. cerevisiae* is grown under each of the five stress conditions tested in this study: ethanol stress, glycerol stress, acidic stress by lactate, oxidative stress, and oxidative stress in a medium supplemented with dextrose. **B)** The mean percentage of genes of the categories' singletons (black bar), duplicates (orange bar), duplicates generated by whole genome duplication (red bar), and duplicates generated by small-scale duplication (light yellow bar) with transcriptional alterations in the five stress conditions tested in this study. **C-F)** We measured the contribution of transcriptionally altered duplicates to the fitness of *S. cerevisiae* under YPD and stress growth conditions using knock-down gene data from Steinmetz et al. (2002). We then compared the fitness contribution of these altered duplicates (red boxes) with that of duplicates with no evidence for transcriptional plasticity under stress (blue boxes). These comparisons were performed for the sets of altered and not-altered duplicates identified under ethanol stress **(C)**, glycerol stress **(D)**, lactate stress **(E)**, and oxidative stress supplemented with dextrose **(F)**. Significant differences are indicated with *, **, and *** when the difference was significant at the levels of 0.01, 0.001, and 0.0001, respectively, using a Mann–Whitney U-test.

An alternative explanation to the adaptive value of duplicates transcriptional plasticity under stress is a low contribution of duplicates to fitness, with their response being the reflection of transcriptional noise caused by environmental perturbations. To test this possibility, we calculated the contribution of each duplicated gene to the fitness of *S. cerevisiae* under normal and stress conditions taking previously published fitness data (Steinmetz *et al.* 2002). To this end, we subtracted the normalized fitness values of a

strain after a gene was deleted or knocked down under stress from the fitness of its ancestral strain (i.e., 1). Therefore, large fitness absolute increment values indicate that the contribution of the gene to fitness is high under those conditions.

We first compared the contribution to fitness of duplicates with and without altered transcriptomic profiles in YPD. Transcriptionally altered duplicates exhibited a higher contribution to fitness than unaltered duplicates (Figure ChII-1C-F), discarding the possibility that altered duplicates may have less contribution to fitness than non-altered duplicates consequently being less selectively constrained to change. Both transcriptionally altered and non-altered duplicates showed a significant increase in their contribution to fitness under stress (Figure ChII-1C-F). However, this increase was sharper in transcriptionally altered duplicates than in unaltered duplicates.

b. Increased transcriptional plasticity after gene duplication

The higher transcriptional plasticity of duplicated genes in *S. cerevisiae*, when compared to singletons, may be the result of a biased preservation in duplicate of highly transcriptionally plastic genes. To test whether or not gene duplication increases transcriptional plasticity we examined the patterns of transcriptional plasticity of duplicates and singletons in the post-WGD yeast *Candida glabrata*, a phylogenetically close species to *S. cerevisiae*. To this end, we asked the question of whether duplicates of *S. cerevisiae* had singleton orthologs in *C. glabrata* that were not more transcriptionally plastic than expected when compared to other *C. glabrata* singletons under stress and vice versa. We obtained RNA sequence data from a previous publication in which transcriptomic data were available under YPD conditions and under acidic stress (Linde *et al.* 2015), similar to our data on lactic acid stress. *S. cerevisiae* orthologs from *C. glabrata* were identified using synteny information available in the pillars of YGOB (Byrne and Wolfe 2005). In total, we identified 4844 reliable *S. cerevisiae*: *C. glabrata* orthologs. Of these 4844 orthologs, 788 genes in *C. glabrata* were duplicated genes (394 pairs, File S6), of which 123 were duplicated in *C. glabrata* but not in *S. cerevisiae*. Of the 2240 duplicates of *S. cerevisiae*, we found 1019 orthologs that were singletons in *C. glabrata*. We first asked whether singletons in *C. glabrata* that are orthologs of duplicates in *S. cerevisiae* exhibit higher transcriptional plasticity than singletons in *C. glabrata* with no duplicates orthologs in *S. cerevisiae*. If this were the case, then gene duplication would have no role in transcriptional plasticity in *S. cerevisiae*. Notwithstanding that the transcriptional plasticity for a particular gene may vary among species, we found that the percentage of singletons with significant transcriptional alterations under stress in *C. glabrata* that are orthologs to *S. cerevisiae* duplicates (599 out of a total of 1019 genes,

58.7%) was not significantly higher than that of transcriptionally altered singletons in *C. glabrata* that had no duplicates orthologs in *S. cerevisiae* (1725 out of a total of 3062 singleton genes, 56.3%) (Fisher's exact test: odd's ratio F = 1.10, P = 0.17). Conversely, duplicates in *C. glabrata* that were orthologs to singletons in *S. cerevisiae* exhibited a percentage of their transcriptionally altered genes under stress (82 out of a total of 123 genes, 66.7%) significantly higher than singletons in *C. glabrata* (Fisher's exact test: odd's ratio F = 1.55, P = 0.02).

c. Differential patterns of transcriptional alterations within duplicated genes

We sought to investigate the different transcriptional profiles of pairs of duplicated genes and their contributions to the fitness of *S. cerevisiae*. We divided duplicated genes that underwent transcriptional alterations after stress into five different categories (Figure ChII-2A and Table ChII-2): (a) Duplicates in which both of the gene copies were up-regulated under stress (called herein Up pattern); (b) duplicates with both copies down-regulated under stress (Down pattern); (c) Duplicates with one copy up-regulated and one copy down-regulated under stress (Discordant pattern), (d) duplicates with one copy showing non-altered transcription under stress while its sister copy shows either up-regulation or down-regulation under stress (Only-one pattern), and (e) duplicates that remained unchanged under stress (Not-altered pattern).

Table ChII- 2 - Categories of altered expression of duplicates

Stress	**Number of pairs both copies Concordant**		**Number of pairs Discordant**	**Number of pairs One altered**	**Number of pairs Not altered**
	Down	**Up**			
Ethanol	89	76	74	438	677
Glycerol	159	103	102	413	777
Lactic acid	147	76	83	433	739
Oxidative	0	0	1	41	42
Oxidative + Dextrose	129	87	96	448	760

In each of the stress conditions, the category "Only-one" comprised the largest number of duplicates with altered transcriptional profiles, with this category including 53% to 97% of the altered duplicates in the five stresses (Figure ChII-2B). These results

support the classical view of evolution by gene duplication, according to which following gene duplication one copy undergoes rapid divergence while the other copy keeps the ancestral function. Here we show that this pattern of evolution by gene duplication is also true for the regulatory evolution of duplicated genes.

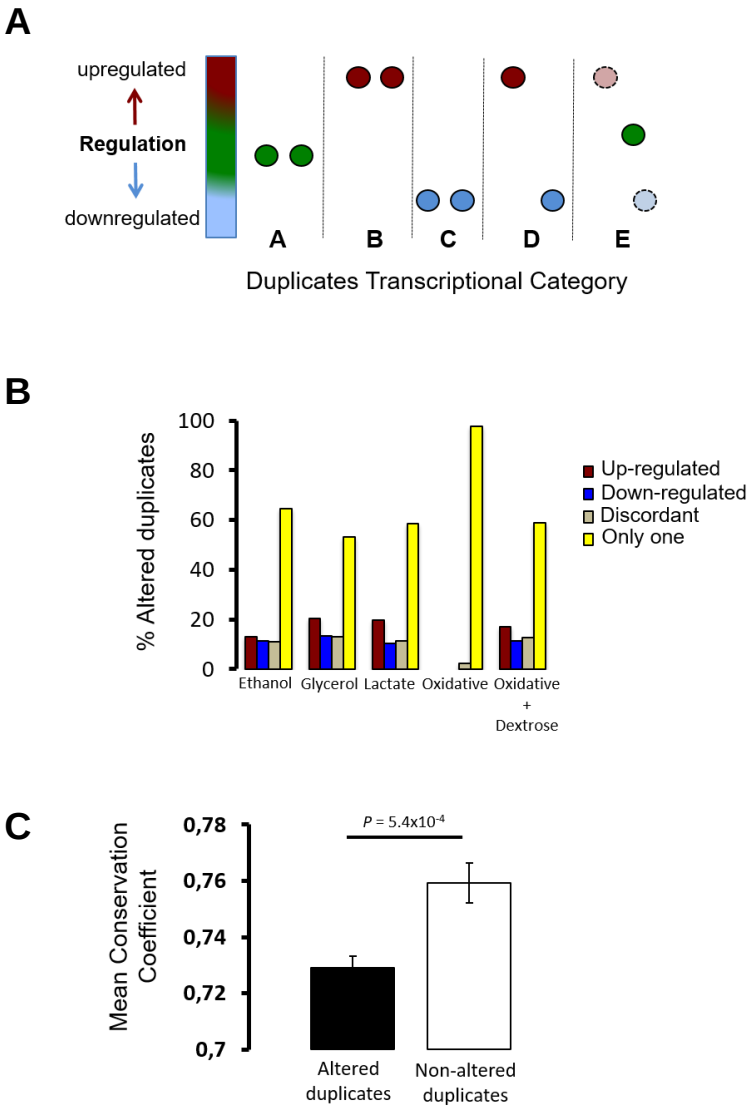


Figure ChII- 2 - Duplicated genes exhibit differential patterns of transcriptional plasticity under stress. We identified five patterns of transcriptional plasticity for the duplicates of *S. cerevisiae* growing under stress conditions **A)**, including duplicates in which neither copy has been altered (category A), those with both copies up-regulated (category B), those with both copies down-regulated (category C), those with gene copies showing discordant transcriptional plasticities (category D), and those in which only one gene copy is altered while its sister copy is unaltered under stress (category E). Calculating the percentage of the duplicated genes belonging to each of the transcriptional categories **B)**, we found that the category with only one copy altered (yellow bar) is the one showing the highest percentage of the altered duplicates under stress. **C)** We measured the conservation of the promoter regions for altered and not-altered duplicates and found that altered duplicates (black bar) exhibit lower conservation than not-altered duplicates under stress (white bar).

Duplicate gene copies with higher transcriptional divergence under stress should show higher sequence divergence when compared to orthologous sequences from other phylogenetically related species if the basis for this transcriptional plasticity was encoded in the gene sequence. To test this hypothesis, we were able to obtain 537 reliable promoter alignments for duplicates in *S. cerevisiae* and at least four additional phylogenetically related yeast species (File S7). The main *Saccharomyces* species we compared S. cerevisiae to were *S. bayanus, S. castellii, S. mikatae, S. paradoxus, S. kluyveri*, and *S. kudriavzevii*. Not all species presented annotated intergenic regions but we used all those alignments that at least included four of the species.

We aligned the 600 nucleotide sequence regions upstream of duplicated genes and their orthologs, as these are likely to include most if not all the regulatory elements of the genes (Ohler and Niemann 2001). We then measured the coefficient of conservation (CC) for each nucleotide site using the entropy equation (Cover and Thomas 2005; Halabi *et al.* 2009; Ruiz-González and Fares 2013):

$$CC = f_k^{(a)} \ln \frac{f_k^{(a)}}{q^{(a)}} + \left(1 - f_k^{(a)}\right) \ln \frac{1 - f_k^{(a)}}{1 - q^{(a)}};\ \forall a\ \in\ [A, T, G, C]$$

In this equation, CC of a nucleotide (*a*) at position (*k*) in an alignment is defined as the entropy of the observed frequency of *a* at *k* ($f_k^{(a)}$) relative to the background frequency of *a* in all sequences of the alignment ($q^{(a)}$). Therefore, the more conserved the site the higher is its CC value. CC was averaged for each promoter and then these averages were used to compare altered duplicates (those belonging to the categories “Up”, “Down”, “Discordant” and “Only-one”) with not-altered duplicates (those belonging to the category “Not-altered”). For each of the stress conditions we estimated the CC values for altered and not-altered duplicates. We then pulled all the data together from all stress conditions and compared the CC values of altered to that of not-altered duplicates. The CC values of duplicates with constant transcriptional profiles under stress (Mean ± SE = 0.76 ± 0.005) was significantly larger than those of duplicates with altered transcriptional profiles (Mean ± SE = 0.72 ± 0.01) (Figure ChII-2C), and the difference was significant using a parametric test (t-test: t = 3.47, *d.f.* = 1140.9, *P* = 5.4 x 10^{-4}) and a non-parametric test (Mann-Whitney U test: *P* = 0.003), indicating that higher transcriptional plasticity of duplicates may be due to a divergence in their promoter sequences from the ancestral pre-duplication state.

d. Duplicates with different transcriptional divergence patterns exhibit different functional dependencies

To determine whether the transcriptional plasticity of duplicated genes is accompanied by a functional divergence of gene copies, we analyzed the genetic interaction network of *S. cerevisiae* and asked how many of the duplicated genes show genetic interactions between their gene copies, hence are functionally dependent on one another (Costanzo *et al.* 2010), within each of the transcriptional categories (i.e., Up, Down, Discordant, One-altered, and Not-altered). To this end, we used the genetic interaction map of *S. cerevisiae* as a proxy to the functions of each of the genes (Costanzo *et al.* 2010). This map contains roughly 6.5 million genetic interactions and the functional chart for 75% of the *S. cerevisiae* genes. The number of genetic interactions for a particular gene is a proxy to the number of functions it performs, as the deletion of both of the genes identified as interacting produces significantly different fitness effects than the multiplicative effect of single gene deletions (Costanzo *et al.* 2010). We identified 762768 significant genetic interactions (i.e., epistasis, ε) in *S. cerevisiae*, of which 52% were synergistic (i.e., the double mutant exhibited significantly lower fitness W_{12} than the multiplicative effects of individual mutants: $\varepsilon = W_{12} - W_1W_2$; $\varepsilon < 0$) and 48% were antagonistic interactions ($\varepsilon = W_{12} - W_1W_2$; $\varepsilon > 0$). However, duplicated genes were largely biased regarding the sign epistasis, with the majority of the epistasis (89.5%) being synergistic (binomial test: $P < 2.2 \times 10^{-16}$). This pattern was also true for transcriptionally altered duplicates (89.74% synergistic epistasis). Dividing transcriptionally altered duplicates into the different categories provides similar results, with all such categories being equally enriched for duplicates with synergistic epistasis: up-regulated duplicates presented largely synergistic epistasis (varying between 86% in ethanol and 93% under oxidative stress supplemented with dextrose), and so did the one-altered category (ranging between 87% of the interactions being synergistic under glycerol stress and 92.7% in ethanol stress). These percentages were of the same order in the "down" and "Discordant" categories.

In all stress conditions the category "Down" showed the highest enrichment for those duplicates with interacting gene copies (Figure ChII-3A). On average over all stress conditions, the genetically interacting duplicates enrichment followed the same pattern, with a distribution among the categories in the following decreasing manner: the category "Down", followed by the category "Not-altered", then the category "One-altered", then the category "Discordant", and finally followed by the category "Up" (Table ChII-3 and Figure ChII-3A emerging box).

Table ChII- 3 - Number of duplicates with genetically interacting gene copies for each transcriptional category of duplicates

Stress	Number of pairs (Total) Down-regulated	Number of pairs (Total) Up-regulated	Number of pairs (Total) Discordant	Number of pairs (Total) One altered	Number of pairs (Total) Not altered
Ethanol	27(49)	6(57)	8(55)	61(301)	67(249)
Glycerol	38(85)	3(77)	15(77)	54(281)	59(190)
Lactate	36(81)	3(54)	10(56)	65(313)	55(207)
Oxidative + Dextrose	36(64)	5(67)	12(71)	72(319)	44(189)

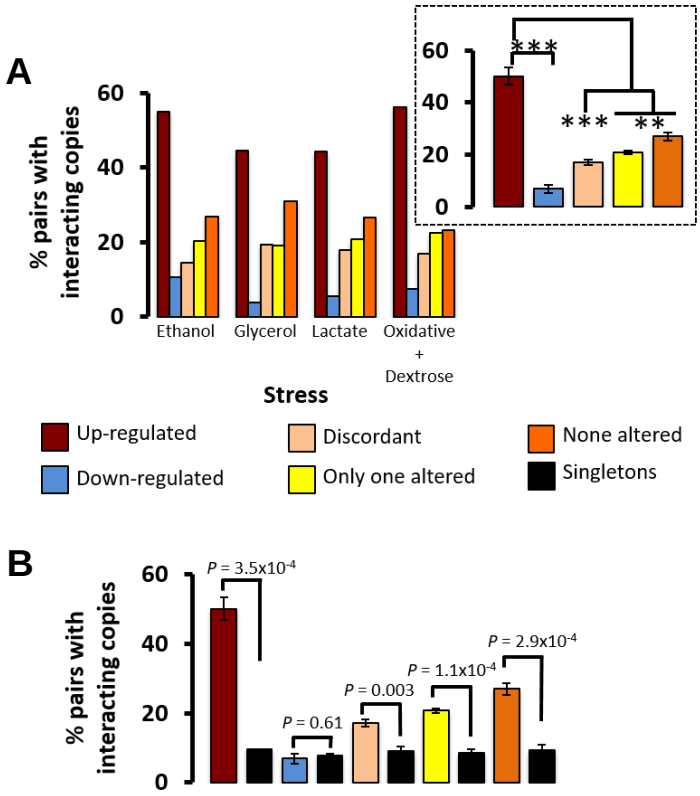


Figure ChII- 3 - The genetic dependencies between gene copies of transcriptionally plastic duplicates. We measured the number of pairs within each of the transcriptional categories with evidence of genetic interactions between duplicate gene copies using the functional landscape of *S. cerevisiae* (Costanzo *et al.* 2010). **A)** The percentage of pairs with interacting gene copies was very high in the category of down-regulated duplicates (red bars), very low in the category of up-regulated duplicates (blue bars), and intermediate in the other three categories under all the stress conditions examined in this study. The mean percentage of duplicates with interacting copies across the stress conditions for each transcriptional category is presented in the inset box. The category "Down" presented a larger mean percentage of duplicates whose gene copies are functionally dependent upon one another than any of the other categories. Significant differences are indicated with ** and *** when the probabilities are $P < 0.01$, $P < 0.001$, and $P < 10^{-4}$, respectively. **B)** The average proportion of duplicates with interacting gene copies was compared to the proportion of genetic interactions for sets of randomly sampled pairs of singletons with altered transcription profiles under stress. Each transcriptional profile for duplicates was compared to an equivalent set of random pairs of singletons with similar transcriptional profiles. For example, up-regulated duplicates were compared to random pairs of up-regulated singletons under stress.

The strong genetic interaction between the gene copies can be due to either each gene copy having a large fitness effects such that deleting both magnifies such effect or, alternatively, each gene copy having very low fitness effects due to genetic redundancy but deleting both significantly magnifies this effect (i.e., functional compensation of a gene deletion or both are needed to perform the function because gene copies have sub-functionalized). The category "Up" is the one with the lowest number of gene copies interactions, therefore is likely to contain very little genetic compensation, perhaps because gene copies have diverged in their function from the ancestral pre-duplication gene, and as such the multiplicative effect of deleting single gene copies may be as important in their contribution to fitness as the double gene deletions. The category "Discordant" shows higher levels of genetic interactions between gene copies than the category "Up", but lower levels than "Down". Since all discordant duplicates exhibit synergistic epistasis, this suggests certain functional redundancy under normal conditions for transcriptionally discordant duplicates, which also applies to the categories of "One-altered" (average percentage of synergistic epistasis among all stresses: 89.9%; binomial test: $P<3.61x10^{-7}$) and "Not-altered" duplicates.

To determine whether duplicate gene copies are more dependent upon each other's functions than expected, we built sets of singleton genes for each of the duplicates sets according to their transcriptional profiles. Each of the singleton transcriptional categories was built by taking random pairs of singleton genes. For example, for the "Up" category, both of the singleton genes were sampled from the set of up-regulated singleton genes under stress. We built sets of 1000 pairs and compared each of the duplicates transcriptional categories with the corresponding singleton transcriptional categories. Results show that all the categories, with the exception of the one including duplicates with both gene copies up-regulated, exhibit a significant proportion of their duplicates with interacting gene copies when compared to singletons of the same transcriptional category (Figure ChII-3B). Therefore, up-regulated duplicates seem to exhibit evidence of independent evolution of their gene copies likely due to the finding of novel functions by each copy under stress.

e. Functional divergence and genetic redundancy of duplicated genes

The differences in the functional dependencies between gene copies found in the duplicates transcriptional categories hint at a different mode of evolution by gene duplication for these categories. We hypothesize, based on the patterns of genetic interactions, that the functional fate of duplicates in terms of neo- or sub-functionalization is dependent on the transcriptional category they belong to and the genetic redundancy

between gene copies. Genetic redundancy has been shown to correlate with evolvability because it provides mutational robustness, which in turn increases the evolvability of genes (Wagner 2000, 2005; Draghi *et al.* 2010).

To test whether a given transcriptional category of duplicates is more likely to have evolved neo- or sub-functionalization, we examined three parameters linked to genetic interactions: (a) the number of shared interactions between the gene copies, (b) the number of total interactions of the gene copies. Neo-functionalized duplicates involve those in which one of the gene copies have lost all ancestral functions and acquired new functions, hence likely reduced the number of genetic interactions. Conversely, sub-functionalization should affect duplicates with many functions in which each copy has become specialized in a set of ancestral functions while sharing common functions with its sister gene copy, hence likely to be over-represented among highly interactions duplicates. Neo-functionalization should also lead to a lower sharing of genetic interactions between the gene copies as one of the copies has acquired novel, and perhaps independent, functions than sub-functionalization. In agreement with our hypotheses, duplicates from the down-regulated category exhibited greater number of genetic interactions (Mean ± SE: 393.39 ± 14.17) than those of the up-regulated category (Mean ± SE: 313.95 ± 13.95; t-test: t = 3.99, *d.f.* = 551.94, $P = 7.36 \times 10^{-5}$). The index of shared interactions was calculated as:

$$Sh_{A,B} = \frac{1}{2}\left(\frac{Sh}{N_A} + \frac{Sh}{N_B}\right)$$

with $Sh_{A,B}$ referring to the mean number of shared interactions between gene copies A and B, *Sh* referring to the number of shared interactions, and *N* being the total number of interactions. Duplicates from the down-regulated category shared more interactions (Mean ± SE: 0.15 ± 0.005) than those of the up-regulated category (Mean ± SE: 0.12 ± 0.004; t-test: t = 3.29, *d.f.* = 209.27, $P = 1.1 \times 10^{-3}$). These results indicate that while up-regulated duplicates may have neo-functionalized, down-regulated duplicates have likely sub-functionalized.

f. Sequence divergence levels of duplicates correlate with their transcriptional profiles

To determine whether the transcriptional duplicates categories included specific functional divergence profiles between gene copies, we inferred the amino acid distances between duplicate gene copies for all transcriptional categories and stress conditions. Divergence between duplicates gene copies was calculated using Poisson-corrected distances. Under all four stresses, the duplicates of category “Down” presented

the lowest gene copies distance (Figure ChII-4A), followed by the category "Discordant", then the category "One-altered", then the category "Up", and finally the category "Not-altered" (Figure ChII-4A). Taking all stresses together, we found three groups of transcriptional categories according to the divergence values between gene copies of duplicates (Figure ChII-4B). The first category is "Down": this category exhibited the lowest divergence levels between gene copies and significantly smaller than the following group that included "Discordant" category duplicates (median divergence values for "Down": 0.05, median for "Discordant": 0.12, Wilcoxon rank test: $P<2.2x10^{-16}$). The following group included duplicates belonging to the transcriptional categories of "One-altered" (median divergence between gene copies: 0.22), "Up" (median divergence between gene copies: 0.24) and the category "Not-altered" (median divergence of gene copies: 0.33). The category "Discordant" exhibited significantly lower divergence between gene copies than the categories "Up", "One-altered" and "Not-altered" (Wilcoxon rank test: $P < 2.2 \times 10^{-16}$).

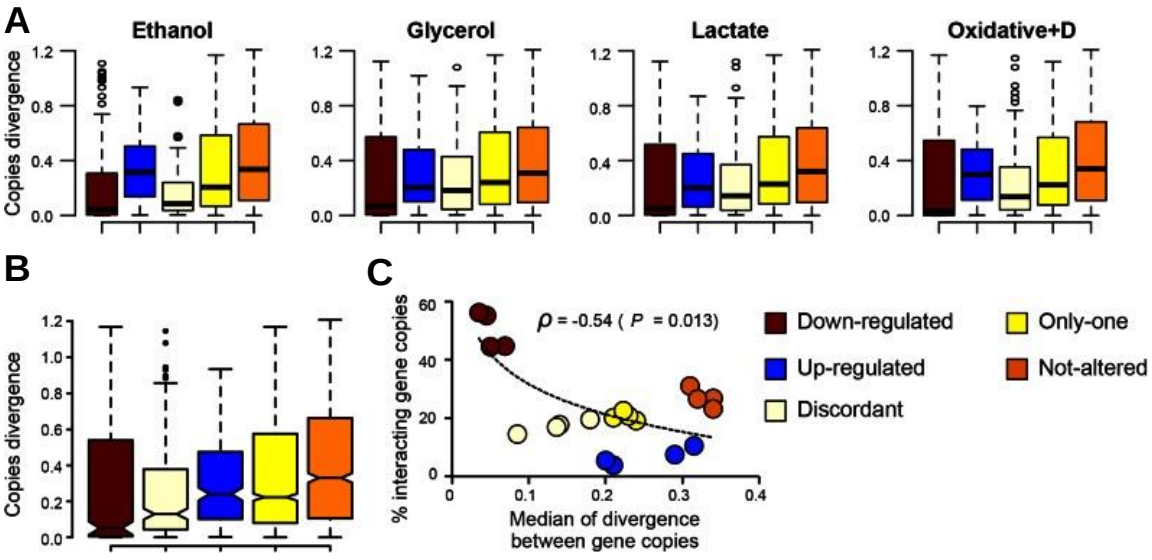


Figure ChII- 4 - Functional divergence analysis of duplicates with different patterns of transcriptional plasticity. **A)** Poisson-corrected amino acid distance between gene copies of duplicates for each of the transcriptional plasticity profiles ("Up," "Down," "Discordant," "Only one," and "Not-altered"). **B)** Comparison of the divergence levels between gene copies for the different transcriptional profiles. **C)** Correlation analysis between the percentage of duplicates with interacting gene copies and the divergence levels between the copies (red, blue, light yellow, yellow, and orange circles refer to the duplicate class of "Down," "Up," "Discordant," "One-altered," and "Not-altered," respectively).

Importantly, the mean sequence divergence levels between the duplicate gene copies correlated negatively with the mean percentage of duplicates in which both gene copies interacted genetically (Pearson correlation: ρ = -0.54, P = 0.013; Figure ChII-4C), indicating that the larger the divergence between the gene copies the lower is their functional dependency. This correlation became more significant when taking only those duplicates for which at least one gene copy has shown changing transcriptional patterns

under stress (Pearson correlation: ρ = -0.78, P = 3.4 x 10^{-4}). This result is in agreement with a previous study in which they analyzed the differences between pairs of WGDs in which both gene copies interacted genetically and those in which gene copies did not (Musso *et al.* 2008).

These results strongly suggest that gene copies that become up-regulated (category "Up") under stress have undergone accelerated evolution and divergence from their ancestral, pre-duplication, functions perhaps allowing the adaptation to stress conditions that are often encountered by the cell in nature.

g. The origin of specific and general adaptations in S. cerevisiae

To determine whether the transcriptional plasticity of duplicates is the result of an adaptive process to face environmental perturbations, we sought to investigate whether this plasticity is stress-specific (i.e., the result of adaptive processes) or a general response to stress. We examined common transcriptionally altered duplicates for each of the transcriptional categories among stress conditions. We found that a substantial proportion of duplicates showed stress-specific transcriptional plasticity (Figure ChII-5A). This pattern was the inverse in the case of transcriptionally altered singletons, with many common such singletons responding to all four stress conditions (Figure ChII-5B). Comparison of the proportion of duplicates in each of the categories for stress response (i.e., stress-specific, common genes response to 2, 3, or 4 stresses) revealed a more significant stress-specific transcriptional alterations in duplicates than in singletons (a mean of 34.4% of duplicates with transcriptional flexibility were stress-specific against 26% of singletons, Fisher's exact test: Odds ratio F = 1.47, P = 3.3 x 10^{-4}), while singletons showed more common responses to all stresses than duplicates (a mean of 32.8% of singletons responded to all stress conditions against 23% of duplicates, Fisher's exact test: Odds ratio F = 1.64, P =1.1x10^{-5}) (Figure ChII-5c). These results reveal a fundamental difference in the transcriptional plasticity of duplicates and singletons, with evidence for the role of natural selection in duplicates transcriptional differences as an adaptive mechanism.

Because of the fundamental different transcriptional plasticities between whole-genome (WGDs) and small-scale duplicates (SSDs) (Figure ChII-1B), we split the duplicates dataset into these two kinds and conducted the same comparison as above. Both the WGDs and SSDs showed very similar transcriptional flexibility patterns as the entire dataset: WGDs and SSDs had their largest transcriptional plasticity in genes that responded in a stress-specific manner (Figure ChII-5D).

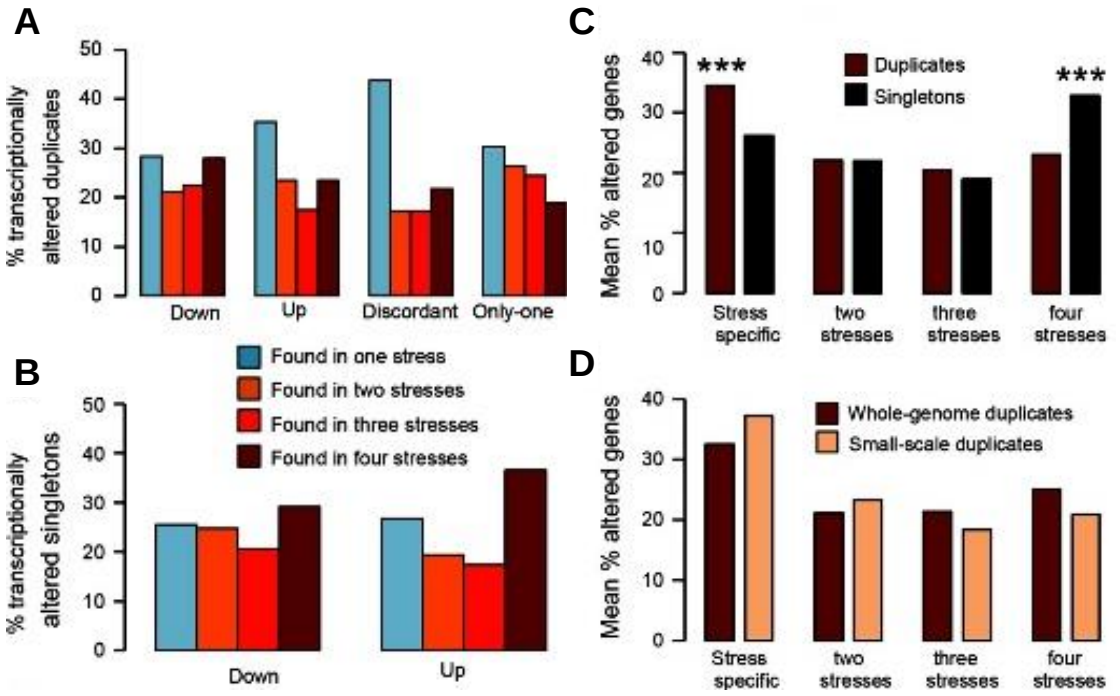


Figure ChII- 5 - Transcriptional plasticity of duplicates is stress-specific. We analyzed the distribution of transcriptionally altered duplicates in the different stress conditions. **A)** Proportion of the transcriptionally altered duplicates from each of the duplicates classes ("Up," "Down," "Discordant," and "Only one") that are altered in one stress only, two stresses, three stresses, or in all four stresses tested in this study. **B)** Proportion of singletons that are up- or down-regulated that respond to specifically one stress only, two stresses, three stresses, or all four stresses. **C)** The mean percentage of altered genes across the different transcriptional classes in duplicates and singletons that are altered under one or more types of stress. **D)** The mean percentage of whole-genome duplicates and small-scale duplicates that are transcriptionally altered when *S. cerevisiae* is faced with one or more stresses. *** indicates $P < 0.001$ under a Fisher's exact test.

To understand the relationship between adaptation to stress and transcriptional plasticity, we analyzed how the different transcriptional alterations in duplicates may have an important role in the adaptation to oxidative stress supplemented with Dextrose (Figure ChII-6). Oxidation generates reactive oxidative molecules or species (ROS) including peroxide, superoxide, hydroxyl radicals, and single oxygen in the cell. Increasing ROS in the cell can lead to important cell structure damages. We found a number of important duplicates that are involved in mitochondrial respiration (Icl1/Icl2, *Shh4/Shh1*, and *Sdh4/Sdh1* duplicated genes, among others) as well as duplicated genes encoding transcriptional gluconeogenesis activators (*Cat8/Sip4*, and *Csr2/Ecm21*) to be up-regulated under oxidative stress, perhaps to reduce the generation of ROS. Moreover, duplicated genes involved in NADH metabolism and oxidative processes in the glycolysis pathway (*Gdp1/Gdp2*, *Gpp1/Gpp2*) are down-regulated under stress (Figure ChII-6). Interestingly, as previously noticed (Bellí *et al.* 2004), the gene copies (*Pug1/Rta1*), involved in heme transport and iron ion homeostasis, showed discordant expression patterns, while the gene copies (*Fit3/Fit1*), involved in ion transport, showed transcriptional alterations only for *Fit3* (Figure ChII-6).

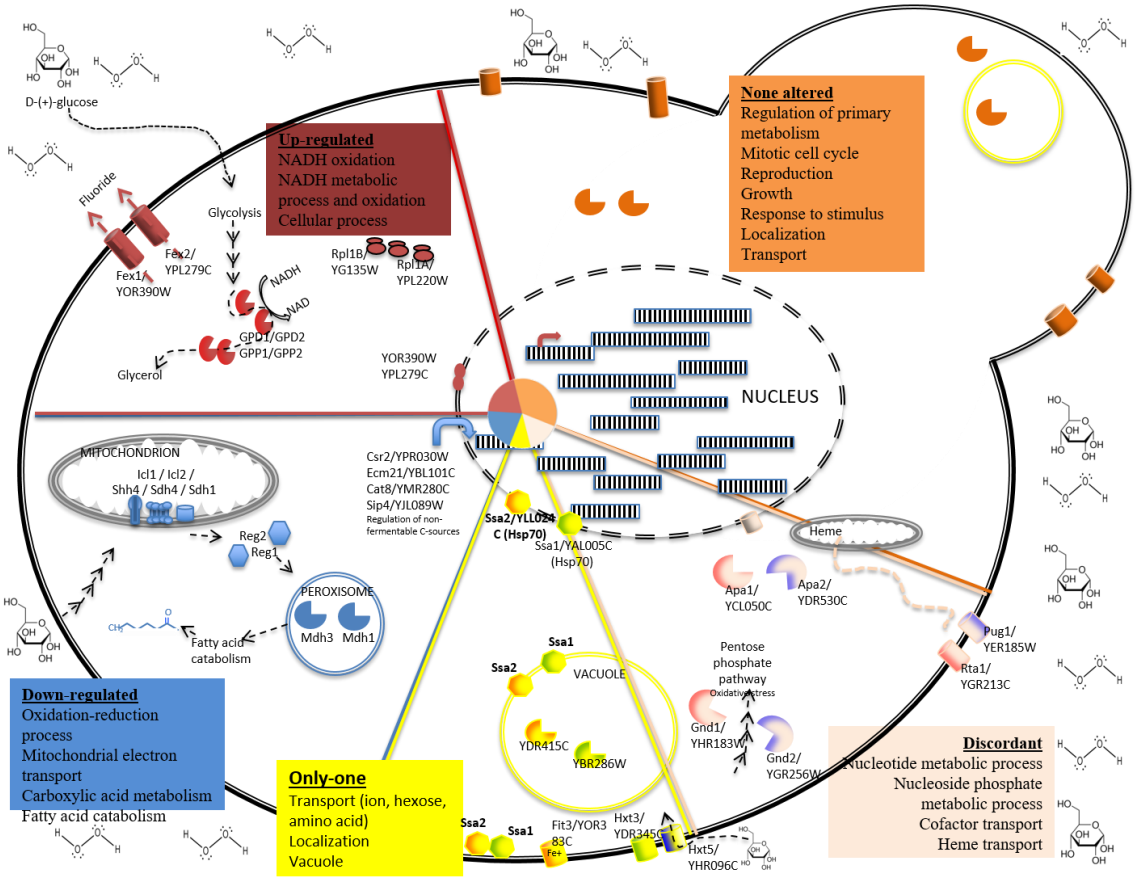


Figure ChII- 6 - Schematic representation of main GO processes and selected duplicate gene pairs affected by oxidative stress in the presence of dextrose. The schematic *S. cerevisiae* cell has been partitioned in a pie shape with the proportions being in accordance to the distribution of duplicates in categories ("Up," "Down," "Discordant," "One-altered," and "Not-altered"). Color codes follow the same categorization. For those categories, "One-altered" and "Discordant," on which each gene in the pair show a differential fold-change, it has been shadowed with green if not differentially expressed, with blue if down-regulated, or with red if up-regulated. Some of the pathways affected are indicated with the corresponding genes names and ID tags, except for the most common gene names, or when not available a common name and the ID tag is provided. Curved arrows in the nucleus represent some transcription factors affected; color indicates the corresponding category. Subcellular localization of proteins has been retrieved from *Saccharomyces* Genome Database following Gene Ontology term analysis.

The transcriptional plasticity of the duplicates belonging to the category "One-altered" was very noticeable under oxidative stress affecting functional classes required to minimize ROS, including many duplicated genes involved in ion transport (*Fit1/Fit3*), hexose transport (*Hxt3/Hxt5*) and heat stress response (*Ssa2/Ssa1*). Similarly, the transcriptional category "Discordant" showed also a prominent response pattern to oxidative stress, including duplicates involved in nucleotide/nucleoside metabolism (*Gnd1/Gnd2, Apa1/Apa2*), and in heme transport and iron ion homeostasis (*Pug1/Rta1*). Most of these duplicates have important roles in DNA replication and stress.

Interestingly, some of these genes are involved in many stresses but their transcriptional plasticity exhibits different patterns under different stresses. For example, the duplicates (*Gpd1/Gpd2*, and *Gpp1/Gpp2*) that drive glycerol production using dextrose through the glycolysis pathway (NADH metabolism and oxidative processes) behave transcriptionally different under the different stresses.

Both of the gene copies of these duplicates are down-regulated under oxidative stress supplemented with dextrose, while only one gene copy is down-regulated when the cell is subjected to stress by lactate or glycerol, and none of the gene copies showed any differential expression level with the wild type when the cell grew under ethanol stress. Similarly, the gene copies of the duplicates (*Shh4/Shh1* and *Sdh4/Sdh1*), involved in the oxidation of succinate and electrons transfer to ubiquinone are up-regulated under oxidative stress. However, under ethanol stress, only gene copies *Shh4* and *Sdh4* are up-regulated while their corresponding paralogs *Shh1* and *Sdh1* show wild-type expression levels (Figure ChII -6).

5. Discussion

In this study, we demonstrate that ancient duplicates of *S. cerevisiae* exhibit a large transcriptional plasticity when subjected to stress. This transcriptional plasticity may be the result of pre-adaptations to environmental stress. Such pre-adaptations may have been generated through an increase in the polymorphism of the regulatory sequence regions of duplicated genes, perhaps the same sequence changes that have led to the regulatory divergence between the gene copies of duplicates. The transcriptional divergence between gene copies has been studied in plants (Blanc and Wolfe 2004b; Ha *et al.* 2007, 2009; Wang *et al.* 2012) and animals (Huminiecki and Wolfe 2004). In *S. cerevisiae*, while the genome-wide transcriptional plasticity has been reported under stress (Ferea *et al.* 1999; Causton *et al.* 2001; Ideker *et al.* 2001; Landry *et al.* 2006; Stern *et al.* 2007; Cormier *et al.* 2010), the differential patterns in this plasticity between duplicates and singletons have received little attention. Findings from these studies led authors to conclude that the transcriptional plasticity of the genes in *S. cerevisiae* is the result of a general response to a wide range of stresses, sparking the possibility that this plasticity is an emerging property resulting from a universal feature of the underlying regulatory network. In this study we show that: (i) ancient duplicates of *S. cerevisiae* exhibit a large transcriptional plasticity when subjected to stress, and (ii) the transcriptional plasticity of duplicated genes differs from that of singletons, is more complex than thought before, and is likely the result of selection for an adaptive response to specific environmental challenges.

The transcriptional changes affecting one or both of the gene copies resulting from gene duplication may be selectively advantageous in unicellular organisms because the absence of tissue-specific transcriptional sub-functionalization precludes a relief of the genetic redundancy of duplicated genes. Therefore, in unicellular eukaryotes, such as *S. cerevisiae*, the efficiency of purifying selection or positive selection must be a strong force driving the fate of duplicated genes. In agreement with this prediction, the genetic redundancy generated in *S. cerevisiae* after the whole-genome duplication event that took place roughly 100MYA was erased by purifying selection, as 92% of duplicated genes returned to single-copy genes (Wolfe and Shields 1997). Despite this, the number of duplicates in *S. cerevisiae* (roughly 30% of all the genes) is higher than predicted by theory, raising the possibility that most retained genes have either become functionally specialized, hence less redundant, shortly after duplication (Force et al. 1999; Lynch and Katju 2004; He and Zhang 2005; Conant and Wolfe 2006; Des Marais and Rausher 2008; Barkman and Zhang 2009), transcriptionally divergent (Blanc and Wolfe 2004b; Francino 2005; Ha *et al.* 2007, 2009; Wang *et al.* 2012), preserved due to their higher

mutational robustness (Wagner 2000, 2005; Fares *et al.* 2013; Keane *et al.* 2014; Fares 2015b), maintained owing to a selective advantage for higher gene dosage (Conant and Wolfe 2008) or kept to preserve stoichiometric balances in duplicates encoding protein complexes (Gibson and Spring 1998; Veitia 2003a; b).

In this study, the transcriptional plasticity identified in *S. cerevisiae* is likely the result of population polymorphism at the regulatory regions of duplicates (Figure ChII-3C), which were selectively relaxed after gene duplication. This polymorphism has likely given rise to pre-adaptations to environments never before faced by the yeast and became fixed in the populations after facing such environmental perturbations. It is therefore likely that duplicates that show transcriptional plasticity, in particular those that become up-regulated under stress, are usually performing important functions in the cell and are hence maintained by purifying selection. However, under stress, such genes may encode new functions that provide the yeast with the ability to survive stress, a property encapsulated within the term exaptation (Gould and Vrba 1982). The question that remains is: how important of an adaptive force is the transcriptional divergence against the functional divergence of duplicates in *S. cerevisiae*?

Functional divergence after gene and genome duplication has been the subject of intense scrutiny and a number of examples unequivocally correlate the origin of important gene families and functional specialization with the divergence between gene paralogues. Indeed, key globin proteins that specialized in different aspects of oxygen metabolism have originated through whole-genome duplication events (Hoffmann *et al.* 2011; Storz *et al.* 2011, 2013; Hoffmann, Opazo, and Storz 2012; Hoffmann, Opazo, Hoogewijs, *et al.* 2012). Functional divergence has also been observed in a number of studies and has been correlated with an asymmetric increase in the rates of sequence evolution in the duplicates gene copies (Blanc and Wolfe 2004b), in good agreement with the fundamental tenet of the molecular evolution theory (Dermitzakis and Clark 2001; Gu *et al.* 2002). Expression divergence, but not transcriptional plasticity, between the gene copies of duplicated genes has also been demonstrated in a number of organisms.

We propose the hypothesis supporting a link between expression and functional divergence—that is, one level of divergence necessarily drives the other level. Indeed, gene expression levels largely determine the rates of evolution of the proteins they encode (Pál *et al.* 2001; Rocha and Danchin 2004; Drummond *et al.* 2006; Wilke and Drummond 2006; Drummond and Wilke 2008). The theoretical justification for this link between gene expression and its rate of evolution can be found in the misfolding-

mistranslation hypothesis, according to which highly expressed genes evolve slower constrained by the need to maintain low levels of misfolded or mistranslated proteins bearing destabilizing mutations (Drummond *et al.* 2005). On the other hand, functional divergence, or acquisition of novel functions, may involve a fine-tuning of the expression of the encoding gene to perform the required function at the right rate. Whether expression divergence came first than functional divergence or vice versa remains to be investigated but our data suggest a link between these two levels of divergence because different transcriptional categories exhibit different patterns of sequence evolution and divergence between gene copies (Figure ChII-4C).

We hypothesize that genetic redundancy has allowed the transcriptional divergence between gene copies due to relaxed selective constraints. This has allowed the divergence at the coding level driven by changes in gene expression, as gene expression is a strong determinant of sequence evolution (Drummond *et al.* 2005). Such functional divergence may have led to the acquisition of functions that enabled the adaptation to stress conditions (Figure ChII-5C).

In this study we determine the plasticity that each of the gene copies has at the regulatory level, the link of this plasticity with the functional dependencies among gene copies, and the role of such a link in the response to stress. Our study reveals different modes of evolution for the different transcriptional categories. Most responsive duplicates to stress present only one copy altered, following the classic view of evolution by gene duplication. The genetic dependencies and low sequence divergence between gene copies for these duplicates also reveal the mode of evolution and innovation: these duplicates exhibit the highest proportion of cases with synergistic epistasis between gene copies, which summed to the low sequence divergence between the gene copies indicates higher genetic redundancy (Vandersluis *et al.* 2010). This non-trivial pattern of evolution of novel functions is in agreement with previous predictions, according to which higher genetic redundancy allows the functional compensation between gene copies, the neutral exploration of genotypic space, and eventual finding of additional novel functions (Wagner 2005; Fares *et al.* 2013; Keane *et al.* 2014; Fares 2015c). The category of up-regulated duplicates exhibits evidence of neo-functionalization based on the rapid evolution of the gene copies when compared to their ancestor and the low functional dependency of each copy on its sister copy, suggesting the acquisition of novel functions. Duplicates with both copies being down-regulated under stress present low divergence between the gene copies and significant functional dependencies among the gene copies, suggesting the sub-functionalization of the gene copies through the partition of ancestral functions. The partition of ancestral functions in these duplicates is not

complete, as the gene copies share more functions than expected. This greater sharing may reflect a selective advantage for gene dosage, particularly in the ancestral state immediately post-dating genome duplication (Ihmels *et al.* 2007; Vandersluis *et al.* 2010). Finally, the category in which gene copies exhibit discordant transcriptional alterations under stress ("Discordant") present low number of genetic interactions between gene copies, low sequence divergence of the gene copies when compared to a pre-duplication ancestral gene, suggesting that both of the gene copies may be performing very similar functions under different conditions (Force et al. 1999). The category of discordant duplicates may include cases in which copies have diverged in their regulation such that one copy is active under stress and its sister copy is active under normal conditions, thereby avoiding the costly evolutionary optimization of the encoded function under two different conditions (Conant and Wolfe 2006). Remarkably, duplicates with no evidence for expression alteration under stress conditions exhibit greater number of interactions between gene copies than expected. Because these duplicates are not affected by environmental perturbations, their enrichment for genetic interactions supports genetic buffering between the gene copies that are independent of the environment, a phenomenon important for genetic robustness (DeLuna *et al.* 2008; Dean *et al.* 2008; Vandersluis *et al.* 2010; Keane *et al.* 2014; Fares 2015b).

Our results point to the expression divergence between the gene copies of duplicated genes as a strategy to yield adaptive responses to stress conditions. This divergence is the result of differential accumulation of mutations (polymorphism) in the promoters of the duplicates gene copies, such that one copy exhibits high expression polymorphism under normal environmental conditions (Keane *et al.* 2014). Since yeast generally undergoes stress because of the changes in the osmotic characteristics of the medium resulting from the metabolic byproducts generated, such stress conditions provide a selective advantage to those strains with polymorphic expression divergence of duplicated genes that allow responding to stress. A number of findings disregard expression noise as an alternative hypothesis to the adaptive value of the observed transcriptional alterations in duplicates. First, the profiles indicating adaptations to stress (i.e., the transcriptional categories including duplicates with at least one gene up-regulated) include duplicated genes that under normal conditions show as much contribution to fitness as those with no response to stress (Figure ChII-1C-F). Second, under stress conditions, the contribution of up-regulated duplicates to fitness is significantly higher than their contribution to fitness under normal conditions. Finally, duplicates with altered expression profiles are mostly stress-specific and affect functions that are strictly related to the interaction with the environment and signal transduction.

In conclusion, we reveal the underlying transcriptional plasticity of duplicates in *S. cerevisiae* and the potential of this plasticity to give origin to adaptations (by means of specialization of duplicates) to environmental perturbations. This study sets a new research endeavor mainly aiming at finding the yet unexplored metabolic capabilities resulting from the evolution of duplicated genes.

6. Data availability and Accession numbers

Strains are available upon request. All RNA sequences are available from the Sequence Read Archive (accession number SRP074821).

7. Supplementary data

Supplementary data are available at G3 online. File S8 is also available in the appendix of this manuscript. **Table S1:** Significant transcriptionally altered genes in *S. cerevisiae* upon growing in ethanol stress. **Table S2:** Significant transcriptionally altered genes in *S. cerevisiae* upon growing in glycerol stress. **Table S3:** Significant transcriptionally altered genes in *S. cerevisiae* upon growing in acidic stress. **Table S4:** Significant transcriptionally altered genes in *S. cerevisiae* upon growing in oxidative stress. **Table S5** Significant transcriptionally altered genes in *S. cerevisiae* upon growing in oxidative stress in a growth medium supplemented with Dextrose. **Table S6:** List of duplicated genes in *Candida glabrata*. **Table S7:** Conservation indices of promoter alignments for duplicated genes with altered transcriptional profiles. **File S8:** compares the methods edgeR and DESeq for the calculations of reads mapped to each gene.

CHAPTER III – Analysis of the transcriptional reprogramming in yeast due to short and chronic exposure to glycerol stress: Cellular responses and evolved adaptations.

A version of this chapter has been published as:

Mattenberger, F., Sabater-Muñoz, B., Hallsworth, J.E. and Fares, M.A. (2017) *Glycerol stress in Saccharomyces cerevisiae: Cellular responses and evolved adaptations.* ***Environmental Microbiology, 19: 990-1007.***

1. Abstract

Glycerol synthesis is key to central metabolism and stress biology in *Saccharomyces cerevisiae*, yet the cellular adjustments needed to respond and adapt to glycerol stress are little understood. Here, we determined the impacts of acute and chronic exposures to glycerol stress in *S. cerevisiae*. Glycerol stress can result from an increase of glycerol concentration in the medium due to the *S. cerevisiae* fermenting activity or other yeast metabolic activities. Acute glycerol-stress led to a 50% decline in growth rate and altered transcription of more than 40% of genes. The increased genetic diversity in *S. cerevisiae* population, which had evolved in the standard nutrient medium for hundreds of generations, led to an increase in growth rate and altered transcriptome when such population was transferred to stressful media containing a high concentration of glycerol; 0.41 M (0.990 water activity). Evolution of *S. cerevisiae* populations during a 10-day period in the glycerol-containing medium led to transcriptome changes and readjustments to improve control of glycerol flux across the membrane, regulation of cell cycle, and more robust stress response; and a remarkable increase of growth rate under glycerol stress. Most of the observed regulatory changes arose in duplicated genes. These findings elucidate the physiological mechanisms, which underlie glycerol-stress response, and longer-term adaptations, in *S. cerevisiae*; they also have implications for enigmatic aspects of the ecology of this otherwise well-characterized yeast.

.

2. *Introduction*

The specialist yeast *Saccharomyces cerevisiae*, known as an archetypal microbial weed in sugar-rich habitats, can also form part of microbial communities in diverse types of environments including soils, plant surfaces, and saline substrates (Botha 2011; Cray, Bell, *et al.* 2013; Lievens *et al.* 2015). Whereas *S. cerevisiae* is a copiotroph, there are no reports of strains capable of biotic activity at water activities below 0.900, e.g. dried fruits, honey, or sugar-saturated beet juice (Lievens *et al.* 2015). By contrast, these biomass-dense populations of this species are commonly found in sugar-rich substrates with intermediate water-activity values (0.990 to 0.920), including floral nectar, fruit juices, and the various substrates used to produce bioethanol (Cray, Russell, *et al.* 2013; Lievens *et al.* 2015; Cray *et al.* 2015). In its natural habitats, *S. cerevisiae* can be exposed to glycerol as a so-called byproduct of yeast and fungal metabolism (Hohmann 2015) and in fermenting substrates, *S. cerevisiae* is known to produce and release glycerol to extracellular concentrations as high as 0.60 M (Basso *et al.* 2008). Highly xerotolerant *S. cerevisiae* strains able to grow down to 0.880-0.900 water activity (see Hallsworth 1998) must accumulate approximately 3.7 M glycerol in order to reduce the water activity of the cytosol to that of the extracellular milieu (de Lima Alves *et al.* 2015). However, most strains can retain metabolic activity only down to 0.940-0.920 water activity, a value equivalent to 2.4-3.2 M glycerol.

Most importantly, and unlike metazoans, yeast utilizes nutrients not only as the source of energy to propel biosynthetic activity, but as the signals which control developmental, metabolic, and transcriptional activities of the cell (Broach 2012). The robust and versatile stress biology of *S. cerevisiae* has been implicated in its ability to dominate the microbial communities within specific habitats (Cray, Bell, *et al.* 2013); elucidating the mechanisms that underlie the dynamic responses, and give rise to the adaptive plasticity, of yeast is imperative to underlying aspects of its molecular and cellular biology, environmental microbiology, ecology and evolution, and biotechnology. Although much of the microbial genome may be implicated in its stress biology, many fundamental aspects of the stress biology of *S. cerevisiae*, particularly the regulatory mechanisms that enable it to respond and adapt to the surrounding environment, remain only partially understood.

Recent developments in biophysical techniques have provided insights into chaotrope-induced stress mechanisms and responses, competitive ability and ecology of *S. cerevisiae* (Bhaganna *et al.* 2010; Cray, Russell, *et al.* 2013; Cray, Bell, *et al.* 2013; Cray *et al.* 2015; de Lima Alves *et al.* 2015). Whereas high extracellular glycerol concentrations can cause a transient turgor change to the *S. cerevisiae* cell, glycerol

penetrates the plasma membrane within 1-2 minutes and does not therefore act as an osmotic stressor (Alemohammad and Knowles 1974; Kiyosawa 1991; Vilhelmsson and Miller 2002; de Lima Alves *et al.* 2015). Its primary mode-of-action as a stressor, therefore, is the depression of water activity and, at molar concentrations, an inhibitory level of chaotropic activity (Williams and Hallsworth 2009; Cray, Russell, *et al.* 2013; de Lima Alves *et al.* 2015). Some of the most xerotolerant *S. cerevisia*e strains can grow at 3-4 M glycerol (Cray, Bell, *et al.* 2013); these strains may be inhibited by both the low water-activity and high chaotropicity of the stressor (de Lima Alves *et al.* 2015). It should be noted that the impact of chaotropic solutes on the flexibility of cellular macromolecules is dependent on the temperature as well as kosmotropic substances present (Hallsworth *et al.* 2007; Bhaganna *et al.* 2010; Yakimov *et al.* 2015; Ball and Hallsworth 2015; Cray *et al.* 2015; de Lima Alves *et al.* 2015). *S. cerevisiae* utilizes trehalose, which is most abundant in the cell during the stationary phase, a highly kosmotropic compatible solute, to protect its macromolecular systems from the worst excesses of the combined chaotropicity of ethanol, acetaldehyde and high glycerol concentrations (Cray, Bell, *et al.* 2013; de Lima Alves *et al.* 2015).

Important advances made in RNA sequencing technology have also made it possible to reveal additional, novel aspects of genetic and molecular mechanisms of stress response (see Taymaz-Nikerel et al. 2016). Such developments, for instance, have enabled levels of experimental standardization and reproducibility, which elucidate biological mechanisms involved in molecular and cellular plasticity under stress and the sensitivity of a number of pathways to abiotic and biogenic stressors (Nagalakshmi *et al.* 2008; Van Dijk *et al.* 2011; Nookaew *et al.* 2012). Four pathways are instrumental in the effective adaptation of yeast to changes in environmental properties, including the signaling networks PKA, TORC1, Snf1, and Pho85. The PKA-signaling network controls cell growth, autophagy, glycogen synthesis, gluconeogenesis, and entry into quiescence (Taymaz-Nikerel *et al.* 2016). TORC1 is implicated in cell growth control and stress response, with the deletion of the *torc1* gene leading to increased thermotolerance and oxidative stress resistance (Cardenas *et al.* 1999; Bjornsti and Houghton 2004; Martin and Hall 2005; Aramburu *et al.* 2014). The Snf1 kinase-signaling pathway is involved in energy homeostasis and response to glucose or carbon limitations (Turcotte *et al.* 2010; Ghillebert *et al.* 2011; Crozet *et al.* 2014; Emanuelle *et al.* 2016). Under carbon-substrate limitations, there is massive reprogramming of gene expression characterized by the altered expression of genes involved in gluconeogenesis, the glyoxylate cycle, and the tricarboxylic acid cycle (Turcotte *et al.* 2010). Finally, the Pho85 signaling pathway responds to phosphate limitations and starvation, which is central to the biosynthesis of

nucleotides, phospholipids, and metabolites (Huang *et al.* 2007; Yadav *et al.* 2016). This pathway is also involved in the response to compromised protein folding and oxide-reductive stress, and other challenges (DeRisi *et al.* 1997; Carroll and O'Shea 2002).

Variation in the levels of glycerol is among the most important challenges which require a cellular response because glycerol is often involved in cellular homeostasis and in adjustment to changes in extracellular osmolarity (Hohmann *et al.* 2007; Hubmann *et al.* 2011). The levels of glycerol are maintained in *S. cerevisiae* by a fine-tuned regulation of glycerol-proton symport via the Stl1 transporter (Tulha *et al.* 2010; Dušková *et al.* 2015). In brief, the yeast Hot1 transcription factor binds physically to the promoter of *stl1*, and in response to osmostress, binds active stress-activated protein (Hog1), a transcription factor from the mitogen-activated protein kinase (MAPK) pathway. Once Hog1 is bound to the promoter of *stl1*, this transcription factor increases the rate of transcription by recruiting the chromatin-remodeling protein Rpd3 and associating with RNA PolIII and components of the mediator complex (Alepuz *et al.* 2003; De Nadal *et al.* 2004). Interestingly, the Hot1 transcription factor seems to bind specifically the promoter of *stl1* and is critical for its transcription (Bai *et al.* 2015). Glycerol is also produced in *S. cerevisiae* through glycolysis and the reduction of the glycolytic intermediate dihydroxyacetone phosphate to glycerol-3-phosphate and the subsequent oxidation of NADH to NAD^+. What transcriptomic changes are involved in responding and adapting to variations in the levels of glycerol remains little understood. Most importantly, whether such transcriptomic responses can evolve sufficiently to improve cell growth under glycerol changes has not been explored.

Yeast exhibits a genome-wide transcriptomic response to stress produced by glucose limitation (Ferea *et al.* 1999). The fact that many of the transcriptional alterations are not stress-specific (Causton *et al.* 2001; Ideker *et al.* 2001; Stern *et al.* 2007; Cormier *et al.* 2010) suggests the possibility that these transcriptional alterations are not the result of adaptive evolutionary changes in the response to stress but is a property emerging from a universal feature underlying regulatory networks. On the other hand, the expression divergence between duplicate gene copies under standard nutrient medium and stress conditions, suggests that response to stress may be the result of adaptive changes in the regulation of duplicated genes (Blanc and Wolfe 2004b; Li *et al.* 2005; Conant and Wolfe 2006; Thompson *et al.* 2013). Gene duplication is universally recognized as a source of novel functions and adaptations. This is because, after the duplication of a gene, the resulting identical gene copies generate genetic redundancy and relax natural selection against one gene copy. The relaxed gene copy can explore novel genotypes and eventually access new phenotypes while its sister copy gene

maintains the ancestral function (Ohno 1970, 1999). Accordingly, gene duplication has been linked to major evolutionary leaps in plants (Wendel 2000; Otto and Whitton 2000; Holub *et al.* 2001; Lespinet *et al.* 2002; Kim *et al.* 2004; Cui *et al.* 2006; Carretero-Paulet and Fares 2012) and animals (Otto and Whitton 2000; Hoegg *et al.* 2004). Evolution of gene expression and expression plasticity through duplication is stronger of a force than evolution of function, as shown in recent experimental analyses (Keane *et al.* 2014).

To both determine whether the response to stress is adaptive or not and address the associate knowledge gaps, here we determine the transcriptomic changes in standard medium-adapted *S. cerevisiae* growing under glycerol-induced stress. The specific aims were to characterize three phenomena: (i) transcriptomic initial response of *S. cerevisiae* growing in glycerol; (ii) impact of the genetic background of *S. cerevisiae* population on the response to exposure to glycerol-induced stress, and (iii) transcriptomic changes underlying the adaptation of *S. cerevisiae* during exposure to glycerol-induced stress for hundreds of generations (during a 10-day period). The sudden exposure to glycerol-induced stress triggers a genome-wide transcriptomic response, and we identified these transcriptomic responses and the associated changes in cellular processes. We show that the genetic variation in the population of *S. cerevisiae* can affect the transcriptomic response to glycerol-induced stress. The study thereby reveals the fine-tuning of *S. cerevisiae* regulatory re-programming during its adaptation and growth improvement when growing in glycerol for a long period. Finally, we present evidence that transcriptionally altered duplicated genes during glycerol-induced stress are the core genes in the immediate response (i.e., short-term) and in the response to alterations in glycerol levels during the evolution of *S. cerevisiae* during a 10-day exposure to glycerol.

3. Materials and Methods

a. Strains, culture media and culture conditions

The *S. cerevisiae* Y06240 haploid *msh2* deletion strain (BY4741; *Mat**a**; his3D1; leud2DO; met15DO; ura3DO; msh2::kanMX4*) (Fares et al. 2013) was used as study subject. The experimental evolution procedure is summarized in Figure ChIII-1. Briefly, from the glycerol stock, a colony was selected (t_{-1}) to start a liquid culture (t_0), from which a set of five populations were established in rich media (YPD: 2% (w/v) bacto peptone, 1% (w/v) yeast extract, 2% (w/v) dextrose; supplemented with kanamycin) and evolved through daily bottlenecks (1%) for 100 days. Populations were allowed to grow at 28°C for 24h, each in 5 ml of media in 50 ml Corning tubes, and subjected to serial passages (1%) in a daily manner. Every ten passages (10 days), a glycerol stock (25%) of the population was stored creating a fossil record. From passage 100 (t_{100}), populations were split in two, one half was grown in normal rich media (YPD) as control populations, and the remaining half in media containing 3% (w/v) glycerol (YPG: 0.41M glycerol, 2% (w/v) bacto peptone, 1% (w/v) yeast extract; also supplemented with kanamycin) as sole carbon source. Split populations were subjected to serial passages (at 10% dilution of the original population; bottleneck of 10% of the population) in a daily manner for an additional ten days (t_{110}), as described above.

b. Quantitation of water activity

Water activity of culture media was quantified at 28°C, using a Novasina Humidat-IC-II water-activity machine (Novasina, Pfäffikon, Switzerland) as described by Stevenson *et al.* (2017). A number of precautions were taken to ensure that the volatility of glycerol did not interfere with quantification and to minimize any other potential error to maintain a level of accuracy consistent with the sensitivity of the microbial cell (Hallsworth and Nomura 1999; Stevenson, Burkhardt, *et al.* 2015; Stevenson, Cray, *et al.* 2015). Calibration was carried out between each measurement of culture medium, using saturated salt solutions of known water activity (Winston and Bates 1960). The water activity of each medium type was determined three times, and the variation was within 0.001.

c. Determination of growth rates under YPD as well as glycerol-induced stress

Growth parameters were evaluated using the BioScreen C plate-reader system (Oy Growth Curves Ab Ldt., Hensinki, Finland) at t_0, $t_{100,}$ and t_{110}. Each time point was pre-cultured overnight at 28°C in 5ml of the corresponding media. Pre-cultures were used to inoculate 200 µl of fresh media (YPD or YPG) to an initial OD_{595} of 0.06-0.07, distributed

in 100-well Honeycomb plates (Oy Growth Curves). Each time point was tested at least in triplicate in the same plate with the two media. Each experimental run was conducted with negative (blank fresh media) and positive (ancestral, t_0, lines) controls. Plates were incubated at 28°C, with continuous shaking (medium force) in the Bioscreen C. Growth was monitored for a period of 92-120h taking OD_{595} measurement every 15 minutes. Maximum growth at exponential phase (μ_{max}) and lagging time (Lagt) were determined with GrowthRates software version 2.1 (Baty and Delignette-Muller 2004; Hall *et al.* 2014) (http://bellingham researchinstitute.com/software/index.html) across replicated cultures. Growth curves were constructed by plotting corrected OD versus time. OD was corrected for linearity by applying the following formula after blank subtraction to each time point: $OD_{cor}=OD_{obs}+0.449*OD_{obs}^2+0.191*OD_{obs}^3$ (Warringer and Blomberg 2003).

d. RNA extractions and transcriptomic analyses

The transcriptomic profiling was performed in the t_0, $t_{100,}$ and t_{110}, with three technical replicates for biological stress condition (3% glycerol (YPG)) in comparison with the normal growth condition (YPD media). Total RNA extractions were performed with RNeasy kit (Qiagen) following manufacturer instructions. Ribosomal RNA was removed by using Ribo-Zero Gold rRNA removal yeast (Illumina) depletion kit. Stranded RNA libraries were constructed using TruSeq stranded mRNA (Illumina) from oligo-dT captured mRNAs from depleted samples. Libraries were run in NextSeq 500 (Illumina) at 75nt single read by using High Output 75 cycles kit v2.0 (Illumina). RNA libraries were sequenced at the Genomic core facility at Servicio Central de Soporte a la Investigación Experimental (SCSIE) from University of Valencia, Spain.

The treatment of the RNA libraries was done following a previous study in which different methods of differential expression analyses were compared (Zhang *et al.* 2014). Raw reads were analyzed using FastQC report and cleaned with CutAdapt as implemented in RobiNA software package v 1.2.4 (Lohse *et al.* 2012). Low-quality reads were filtered and trimmed (Phred score inferior to 20 and size less than 40nt were discarded). Reads were then aligned with Bowtie (up to 2 mismatches accepted) to the reference transcriptome (PRJNA290217) from the reference S288c strain. Statistical assessment of differential gene expression was done with edgeR (Robinson *et al.* 2010) and with DESeq (Anders and Huber 2010) as implemented in RobiNA. A previous study compared the different expression analysis methods, concluding that edgeR and DeSeq were the best-performing methods when the objective is to analyze differential expression (Zhang *et al.* 2014). Comparison of the logarithmic fold change of our expression data between edgeR and DESeq provided a very strong correlation

(Spearman correlation coefficient: $\rho = 0.995$, $P < 2.2 \times 10^{-16}$, Figure 1 of File S8 from the previous chapther II). Significant expression changes were identified using a false discovery rate (FDR < 0.05). These results indicate that our quantification of expression data is robust to the method used. RNA raw reads are available from the Sequence Read Archive with accession number SRP074821.

Genes with significantly higher reads per billion (RPKM) under YPG than YPD (with a false discovery rate for the fold change of expression FDR < 0.05) were considered transcriptionally altered in YPG. Because RNA molecules can undergo degradation before being translated, we examined the correlation between the RPKMs of our transcriptomic analyses and those obtained by other groups using ribosomal profiling, a technique that measures ribosome occupancy and translation genome-wide (Albert *et al.* 2014). Despite the large number of data available for both of the methods ($N = 4682$), we found a very strong and significant correlation between the counts of both of the methods (Spearman´s correlation: $\rho = 0.77$, $P < 2.2 \times 10^{-16}$). Hence, highly transcribed genes are also highly translated and vice versa. Our data, therefore, are indicative of the levels of gene expression. RNA-seq technology is sensitive to biases in expression detection, such that often it becomes difficult distinguishing genes with very low read counts from background noise. However, it has recently been shown that RPKM metric is robust to the low expression filtering strategies (Lin *et al.* 2016). Indeed, we could identify differentially expressed genes at read counts as low read as RPKM = 0.001 (logRPKM = -2.8).

e. Identification of duplicated genes

Paralogous pairs of duplicated genes were identified as the resulting best reciprocal hits from all-against-all BLAST searches using BLASTP with an E-value cutoff of $1E^{-5}$ and a 50-bit score (Altschul *et al.* 1997). Paralogs were then divided into two groups according to the mechanism of their origin: WGDs and SSDs. WGDs are those extracted from the reconciled list provided by the YGOB (Yeast Gene Order Browser, http://wolfe.gen.tcd.ie//ygob, Byrne and Wolfe 2005) (555 pairs of genes), and these were not subjected to subsequent SSD. All other paralogs were considered to belong to the category of SSDs (560 pairs of genes). The duplicates used in this study have been estimated to have their origin on the time point of the whole genome duplication that took place 100 MYA (Wolfe and Shields 1997). Also, in this study, we have used the SSDs that exhibit a similar distribution of synonymous substitutions as those of WGDs, so roughly belonging to the same age (Fares *et al.* 2013; Keane *et al.* 2014).

f. Gene ontology -functional categories classification and visualization

For each differential expressed gene list, gene ontology (GO) term was associated with Gene Ontology Term Finder as implemented in the Saccharomyces Genome Database (http://www.yeastgenome.org/cgi-bin/GO/goTermFinder.pl), which also include a GO enrichment analysis, with a p-value cutoff of <0.01 (Supplementary tables S1 to S17). A semantic similarity score, simRel (Schlicker *et al.* 2006) was used to summarize and remove redundant GO terms in the list, as implemented in REVIGO software with medium (0.7) allowed similarity, using *S. cerevisiae* GO term database and the p-values from enrichment analysis (Tomislav *et al.* 2011).

g. Measure of metabolic distance

To calculate the metabolic distance between the three *S. cerevisiae* populations (t_0, t_{100}, and t_{110}) we compared the list of GO process terms enriched for transcriptionally altered genes between two populations (*i* and *j*) by calculating the number of shared process terms ($SP_{i,j}$) and the number of enriched terms for transcriptionally altered genes only in one of the populations but not the other (P_i and P_j). The metabolic distance between the two populations ($MD_{i,j}$) was calculated as:

$$MD_{i,j} = 1 - \frac{SP_{i,j}}{Min[P_i, P_j]}.$$

Here $Min[P_i, P_j]$ is the number of cellular processes enriched for transcriptionally altered genes for the population with minimum number of such processes. Following this equation, metabolic distance varies between 0, when the transcriptionally altered genes from both populations affect the same processes (i.e., $P_{i,j} = Min[P_i, P_j]$), and 1 when there is no overlap in the process terms.

h. Software

Calculations and statistics were performed using MS Excel and R 3.2.1. (R Core team 2013), except as indicated above for differential expression, and growth parameters.

4. Results

a. Glycerol acts as a potent cellular stressor of S. cerevisiae

From an initial *S. cerevisiae* colony (ancestral colony) we grew a culture in our standard nutrient medium; yeast extract, peptone, and dextrose medium (YPD; 0.998 water activity). This was the control population of *S. cerevisiae*, which included genetically related *S. cerevisiae* cells growing in a rich medium to which *S. cerevisiae* was adapted. Another culture was initiated from the ancestral colony in yeast extract, peptone, and 0.41 M glycerol (YPG), a stressful environment (0.990 water activity) to which *S. cerevisiae* was not adapted (Figure ChIII-1). We measured cell density of the populations of *S. cerevisiae* in the standard medium (YPD), and stressful medium (YPG), using optical density (OD_{595nm} was determined for five biological replicates at 15-min intervals during a 92-hour period) to provide a measure of cell number and metabolic activity under these media. The *S. cerevisiae* growth curve in YPD was a sigmoidal curve with a maximum growth rate of (μ_{max} ± s.d.m.= 0.33 h^{-1} ± 0.01), as calculated by the program GrowthRates (*Materials and Methods*). The growth rate of *S. cerevisiae* declined in YPG (Figure ChIII-2A), with a maximum growth rate in YPG of (μ_{max} ± s.d.m.=0.16 h^{-1} ± 0.003), a rate that was significantly lower than the growth rate in YPD (Normal test: z = 24.28, P << 0.001; Figure ChIII-2B). Importantly, the lag time in YPG (Lagt = 21.8 h) was approximately ten times longer than that in YPD (Lagt = 2.7 h) (Figure ChIII-2A). The difference in lagging time may be required for the regulatory re-programming of the cell to respond to YPG. Interestingly, cell number was greater in YPG than in YPD when their respective cultures reached the stationary phase (Figure ChIII-2A).

b. Transcriptome-wide, cellular stress response to glycerol

The response to glycerol stress was further characterized by comparing the transcriptome of the control population (standard medium) to that of the glycerol-stressed population (see *Material and Methods*; Figure ChIII-1). To this end, we compared the transcriptional level of each of the 5825 genes for which we obtained reliable RNA read counts between the control and the stressed populations (*Materials and Methods*).

Glycerol-induced stress altered significantly the expression (false discovery rate FDR < 0.05) of 2748 out of the 5842 (47.04% of the total) analyzed genes in *S. cerevisiae* (Figure ChIII-3). Out of the 2748 differentially expressed genes, 1331 were up-regulated (i.e., $\log_2FC > 1$) a proportion similar to that of down-regulated genes (Binomial test: P = 0.13). The extent of transcriptomic response was of the same order as previously reported for a number of stressful environmental conditions (Gasch *et al.* 2000).

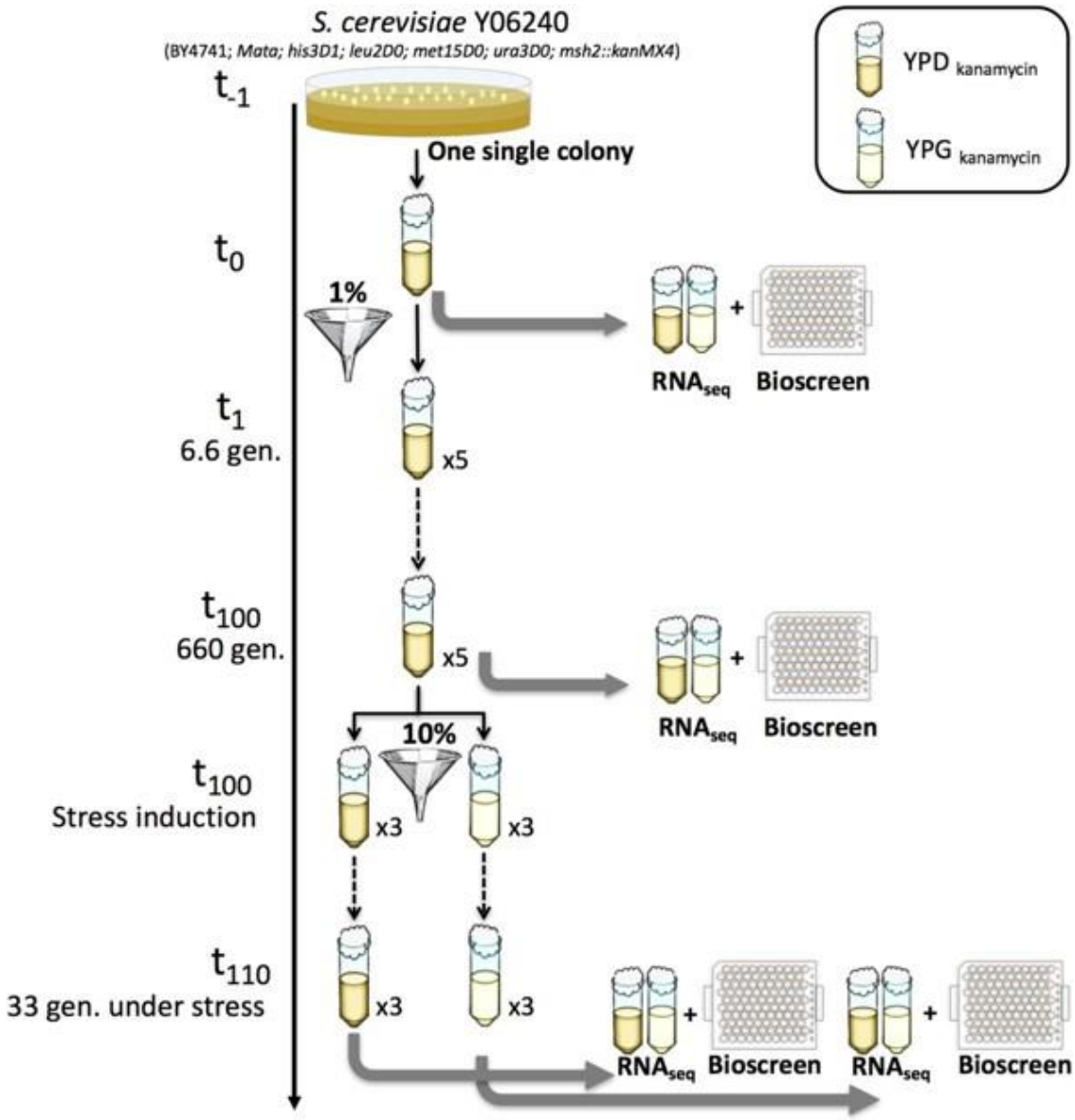


Figure ChIII- 1 - Experimental procedure to test the transcriptomic response of S. cerevisiae populations to glycerol-induced stress. From a single colony of *S. cerevisiae* strain YO6240, we derived a population which was used in the rest of the experiments (and named this population t_0). This population was subjected to glycerol-induced stress and the transcriptomic changes from growing in glucose (YPD) to growing in glycerol (YPG) were quantified. From the population at t_0, we evolved a population in YPD for 100 passages (approximately 660 generations of *S. cerevisiae*) by transferring a 1% dilution daily to a tube with a fresh YPD medium. Then, we subjected this population at t_{100} to glycerol-induced stress and quantified its transcriptomic response. Finally, the population at t_{100} was exposed to long-term glycerol induced stress by evolving for 10 passages (66 generations) under glycerol, with a control population growing under YPD. This evolution took place by transferring a 10% of the population to a new tube containing YPG or YPD media. All populations were grown in the presence of kanamycin to minimize the possibility of bacterial contamination.

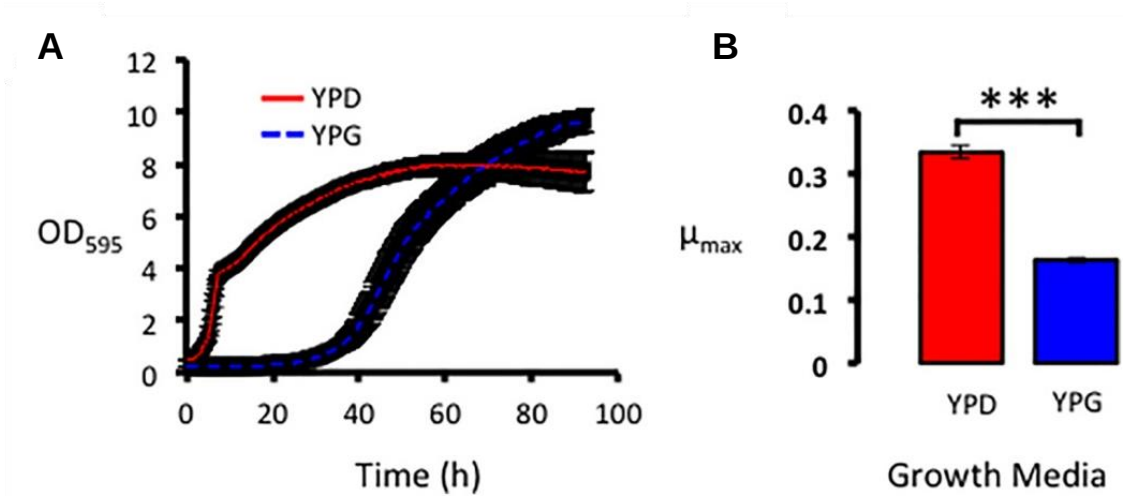


Figure ChIII- 2 - Environmental changes that involve a replacement of glucose with glycerol impact negatively the ability of S. cerevisiae to grow. We grew *S. cerevisiae* populations in normal conditions (YPD) or in growth media containing 3% glycerol (YPG). **A)** Growth curves show that *S. cerevisiae* grows at a higher rate in YPD (red sigmoidal curve) than in YPG (blue sigmoidal curve). The shaded area throughout the curve represents the standard deviations for each time point. **B)** The maximum growth rate of *S. cerevisiae* in YPD (red column) was significantly greater than its growth rate in YPG (blue column) using a Wilcoxon rank test (***P < 0.001).

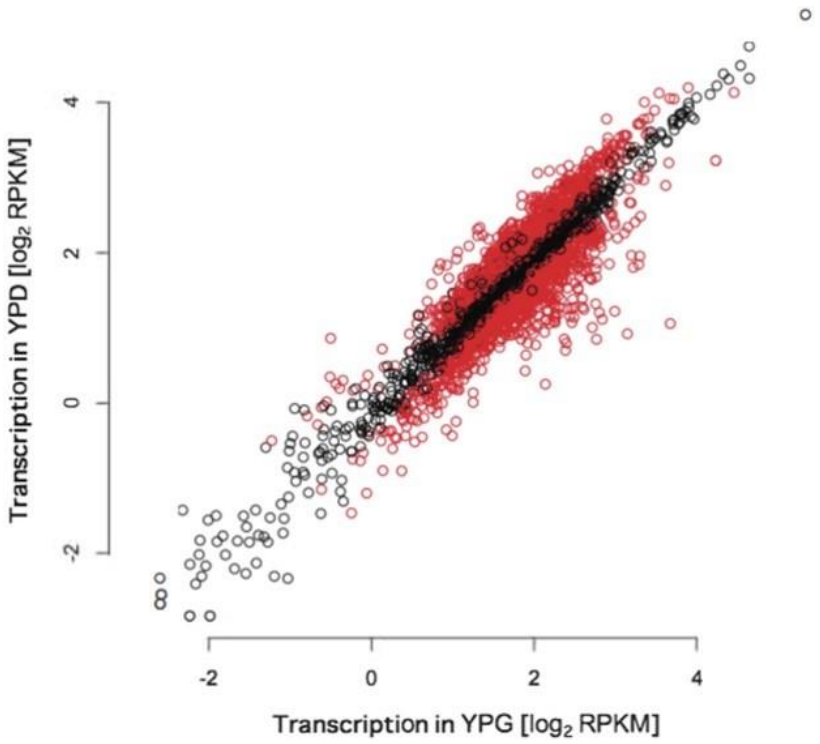


Figure ChIII- 3 - Growth of S. cerevisiae in glycerol unfolds a genome-wide transcriptional response. The plot represents the comparison of the expression level (measured as the logarithm of the number of reads per billion (RPKM)) of each *S. cerevisiae* gene under YPD medium (Y-axis) and YPG medium (X-axis). Expression levels of genes are represented as Reads per billion (RPKM) logarithmically transformed. Black dots in the plot represent genes with no evidence of transcriptional alterations when comparing their expression in YPD with this in YPG. Red dots represent' genes exhibiting significant transcriptional changes when comparing their expression in the two growth media and these are either up-regulated in YPG compared with YPD (red dots above the diagonal) or down-regulated (red dots below the diagonal).

We analyzed the distribution of de-regulated genes among the different cell process categories classified according to Gene Ontology (GO) terms using the *Saccharomyces* Genome Database (SGD: http://www.yeastgenome.org). De-regulation of genes affected fundamental processes in the cell (Table S1), including translation, mainly ribosomal genes and genes involved in ribosomal biogenesis, oxidation-reduction, and cellular respiration, as a consequence of shifting from a fermentative (in YPD) to an oxidative growth (in YPG), genes from the mitochondrial respiratory chain complex assembly, and genes encoding proteins involved in the transport of substances across the membrane (Table S1). Importantly, genes involved in deriving energy by oxidation from organic compounds were also transcriptionally altered (Table S1). A number of environmental stress response genes were transcriptionally altered, including genes from the pathway of positive and negative regulators of protein kinase A (PKA) such as *pkh1* and *pkh2*, *ras1*, *cdc25*, and *gpa2*. Only the gene encoding one β-subunit of the SNF1 complex, *sip2*, was transcriptionally altered. Several transcription factors were also altered, including *cat8* and *rds*2, both of which are potent activators of gluconeogenesis (Schüller 2003; Turcotte *et al.* 2010).

Identification of the genes that were up-regulated or down-regulated in cells grown on YPG indicated the primary physiological changes that had taken place (Figure ChIII-4). In general, growth in YPG up-regulated genes involved in respiration, including redox and genes from the mitochondrial respiration chain assembly. It also up-regulated genes involved in deriving energy from oxidative organic compounds, TCA cycle, ATP synthesis, fatty acid oxidation, phosphorylation, and transmembrane transport, among others (Table S2 and Figure ChIII-4A). In contrast, translation, represented by the biogenesis of ribosomes and maturation of ribosomal subunits, transcription, protein folding, cellular biosynthesis, glycosylation, and nucleic acid metabolism, among others were down-regulated when growing in YPG relative to YPD (Table S3 and Figure ChIII-4B).

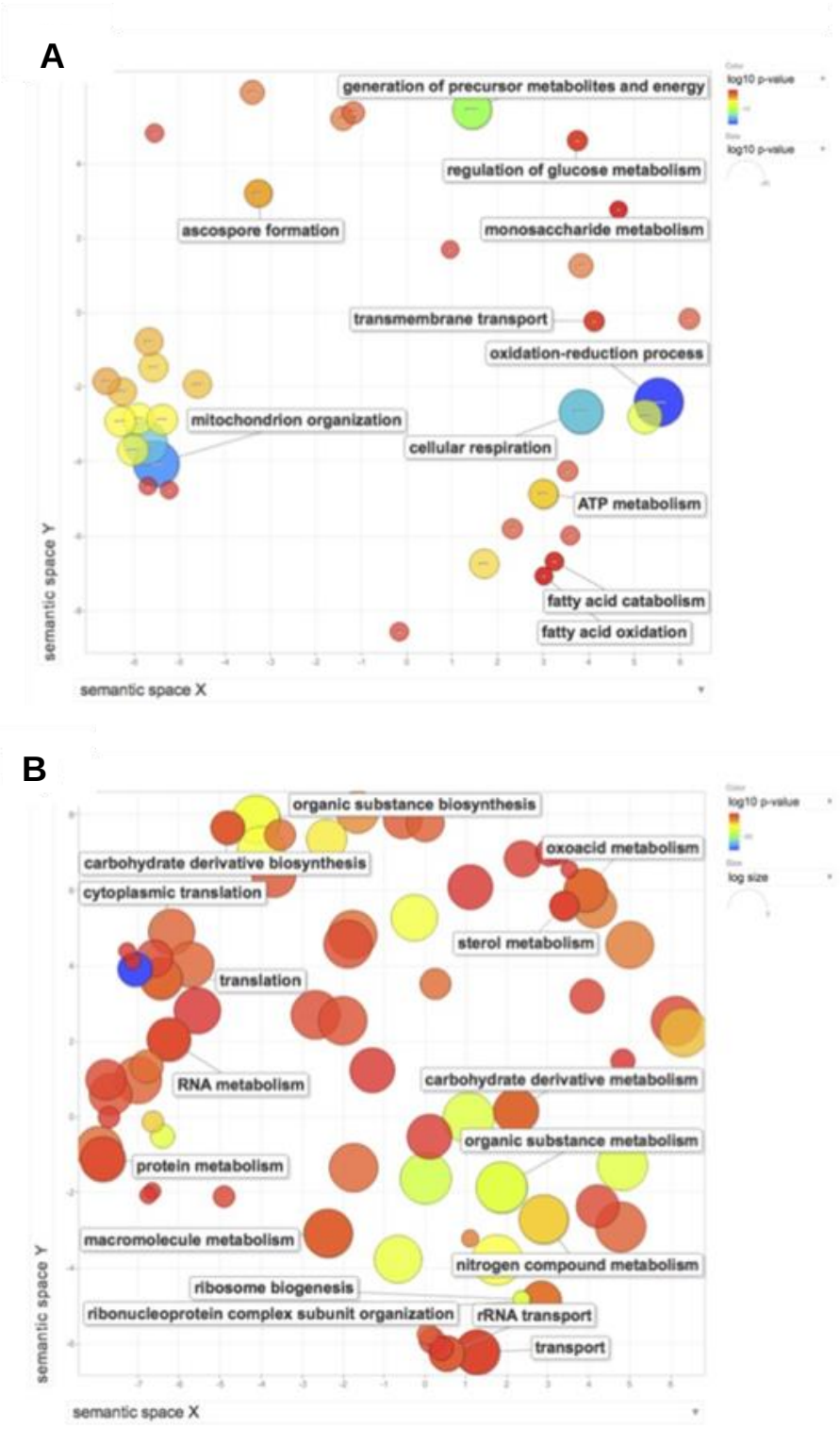


Figure ChIII- 4 - Growth in YPG induces up-regulation and down-regulation of genes involved in a large number of cellular processes. Significantly transcribed genes under glycerol stress were classified according to their gene ontology (GO) term by using Panther GO sorting system. The scatterplot shows the cluster representatives in a two-dimensional space, on which bubble color indicate the logarithm of GO term P-value (based on the number of genes belonging to this category), whereas the bubble size indicates the frequency of the GO term in the organism database (the more general term, the bigger bubble size). Only the most relevant cluster descriptors are shown. **A)** Processes enriched for up-regulated genes under YPG growth conditions. **B)** Processes enriched for down-regulated genes under YPG growth conditions.

c. Increased genetic diversity increases the phenotypic plasticity of S. cerevisiae

Population t_0 was derived from a single colony, with this population being very limited in terms of genetic diversity, and thus the observed transcriptomic response may not represent a natural population. To test how the genetic variation of a population may influence the transcriptomic response when such population is challenged with an environmental change (for example, growth in YPG instead of YPD), we first increased the genetic variation in the initial population by propagating this population of *S. cerevisiae* growing in YPD during 100 passages, with each passage involving the transfer of a hundredth of the population to a new flask with fresh YPD medium (*Material and Methods*) (Figure ChIII-1). In each passage, we transferred a hundredth of the population, corresponding to $\log_2 100 = 6.6$ generations of the yeast, to a new flask containing fresh YPD medium. Taking into account a mutation rate of 10^{-8} mutations/nucleotide for *S. cerevisiae Δmsh2* strain used in this study, on average each passage involves the fixation of one mutation per cell every day. Most of the variation accumulated during the experimental evolution of the propagated population was neutral (i.e., there were no deleterious mutations fixed by genetic drift) since the maximum growth rate of the propagated populations (μ_{max} ± s.d.m. = 0.35 ± 0.02) was not significantly different from that of the original population (Normal test: z = 1, P = 0.16).

After 100 passages (hereafter t_{100}), we challenged the propagated population by growing it in YPG and analyzed its transcriptomic changes at the exponential phase by comparing its transcriptomes in YPG and YPD. We then compared the transcriptomic changes at t_{100} with those at t_0. The number of transcriptionally altered genes (N = 2429) at t_{100} was 12% lower than that at t_0. However, only 58% of the genes transcriptionally altered at t_{100} when yeast had been cultured in YPG (N = 1426) were also altered at t_0, while 1003 genes that were differentially regulated in YPG compared to YPD at t_{100} were not altered at t_0 (Figure ChIII-5A). The fact that roughly half of the transcriptomic response at t_{100} to glycerol was unaltered at t_0 is intriguing (Figure ChIII-5A). We believe that the set of genes transcriptionally altered in both of the populations, at t_0 and t_{100} (N = 1426), represents the core genes of response to glycerol-induced stress, while the set of genes altered in YPG at t_0 but not at t_{100} (N = 1321), or those altered at t_{100} but not at t_0 (N = 1003) must represent transcriptomic variation that is dependent on the genetic composition of the population. Identification of the cellular processes significantly enriched for the set of altered genes in the population challenged with glycerol at t_0 and at t_{100} and the set of altered genes in only one of these populations revealed meaningful differences between the two sets of altered genes. The processes of translation, respiration, nucleotide metabolism, energy derivation by respiration, carboxylic acid

metabolism, ion transmembrane transport, and ATP metabolism were enriched for the set of transcriptionally altered genes in both t_0 and t_{100} populations (Figure ChIII-5B; Table S4). Processes mostly concerned with regulation of transcription and gene expression, protein export to the mitochondrion, mitochondrial translation, and transcription were significantly enriched for genes transcriptionally altered at t_0 but not t_{100} (Figure ChIII-5C and Table S5). Finally, cellular processes involved in the regulation of transcription and a number of other fundamental biological processes were significantly enriched for genes altered at t_{100} but not t_0 (Table S6).

Out of the genes up-regulated in cells cultured in YPG (in comparison with genes' expressions in cells cultured in YPD) at t_{100}, 55% (N = 660) were also up-regulated in cells cultured in YPG at t_0. This percentage is 6% greater than the down-regulated genes that were common to the t_{100} and t_0 populations (49%, N = 602, Fisher's exact test: F = 1.2, P = 0.01) (Figure ChIII-5D). Only 82 genes that were up-regulated at t_{100} (i.e., approximately 6.7% of those up-regulated) were among the down-regulated genes found at t_0, pointing to a high consistency between the transcriptomic data of *S. cerevisiae* isolated at t_0 and t_{100}. Therefore, in general terms, up-regulated genes at t_{100} that are transcriptionally altered at t_0 are also up-regulated at t_0, and down-regulated genes at t_{100} that are altered at t_0 are also down-regulated at t_0. The low overlap in the transcriptional profiles of the populations at t_0 and t_{100} was paralleled with differences in the growth rate of these populations in YPG. Indeed, the population evolved for 100 passages in YPD exhibited higher maximum growth rate in YPG (μ_{max} ± s.d.m. = 0.19 h^{-1} ± 0.015) than t_0 populations (μ_{max} ± s.d.m. = 0.16 h^{-1} ± 0.003; Normal test: z = 3.33, P < 0.025).

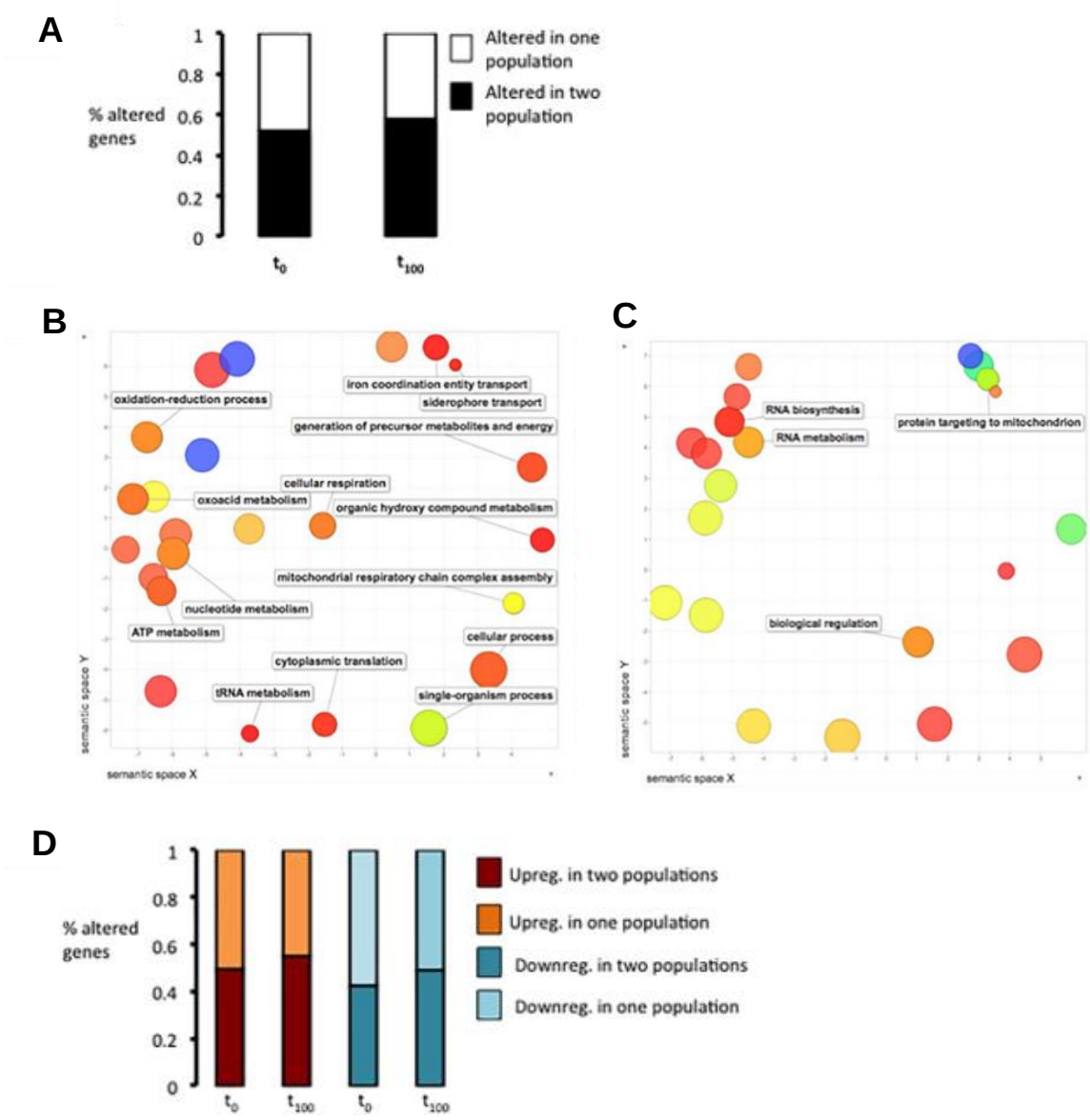


Figure ChIII- 5 - Populations of S. cerevisiae isolated at different times of their evolution exhibit overlapping and non-overlapping transcriptional responses to glycerol. A) Identification of genes that are transcriptionally altered in the ancestral population (t_0) and the population evolved for 100 days in YPD (t_{100}). Black areas of the columns represent the percentage of genes transcriptionally altered in that population when grown in YPG that are also altered in the other population, while white represents genes altered in one population but not in the other. **B)** Semantic clustering of cellular processes enriched for genes that were transcriptionally altered in both of the populations (at t_0 and t_{100}). The diameter of the circle represents the proportion of genes in a particular cellular process found transcriptionally altered, while the different processes are color-coded. The color of the bubbles represents the proportion of genes in a particular cellular process found transcriptionally altered (log P value), while the size indicates the frequency of the GO term in the organism. **C)** Semantic clustering of cellular processes enriched for genes that were transcriptionally altered in t_0 but not t_{100}. **D)** Proportion of genes that are up-regulated or down-regulated in one and/or both population(s).

d. Transcriptomic-wide, adaptive evolution after successive generations under glycerol stress

To determine what transcriptomic changes mediate adaptive growth in glycerol, we evolved in triplicate the t_{100}-evolved population (i.e., genetically diverse population) for 10 passages separately in YPG and in YPD by transferring a tenth of the population to a new fresh YPG-containing flask every 24 hours, with each passage allowing the transfer of $\log_2 10 = 3.33$ generations of the yeast. We used a lower dilution in the transfers in YPG than that used during the 100 passages of evolution in YPD to increase the population size transferred and consequently the efficacy of selection in the adaptation to YPG. After 10 passages of evolution, the evolved population exhibited a significant increase in its maximum growth rate in YPG (μ_{max} ±s.d.m. = 0.26 h^{-1} ± 0.013) compared to the populations isolated at t_{100} (Normal test: $z = 4.67$, $P < 0.01$) and t_0 (Normal test: $z = 6.66$, $P < 0.001$). The growth rate of the t_{110}-population evolved in YPG was greater than that of t_{110}-population grown in YPD (μ_{max} ±s.d.m. = 0.23 h^{-1} ± 0.01) (Normal test: $z = 3$, $P<0.025$).

Analysis of the transcriptomic changes in the t_{110}-population evolved in YPG in comparison with the t_{110}-population evolved in YPD revealed that adaptation to YPG involved a massive regulatory re-programming of *S. cerevisiae* with 2640 genes exhibiting differential expression in YPG compared to YPD. Out of the transcriptionally altered genes, 1274 genes (48.3% of the total) were up-regulated and 1366 were down-regulated. Comparison of the list of altered genes in the t_{110}-population evolved in YPG compared to YPD with that list in the t_{100}-population yielded 1383 genes out of the 2640 as transcriptionally altered in YPG in the t_{100} and t_{110} populations (52.4%), and 1453 genes (55%) were coincidentally altered in YPG compared to YPD in t_{110} and t_0 populations. We identified core-altered genes, those that were transcriptionally altered in the t_0, t_{100}, and t_{110} populations, and thus were genes responsive to glycerol-induced stress regardless of the length of exposure of the population to glycerol. This list of genes included 902 core genes (table S7) that were transcriptionally altered in YPG compared to YPD. We also identified the genes that were uniquely altered in cells cultured in YPG in each population and not in the others (Figure ChIII-6A).

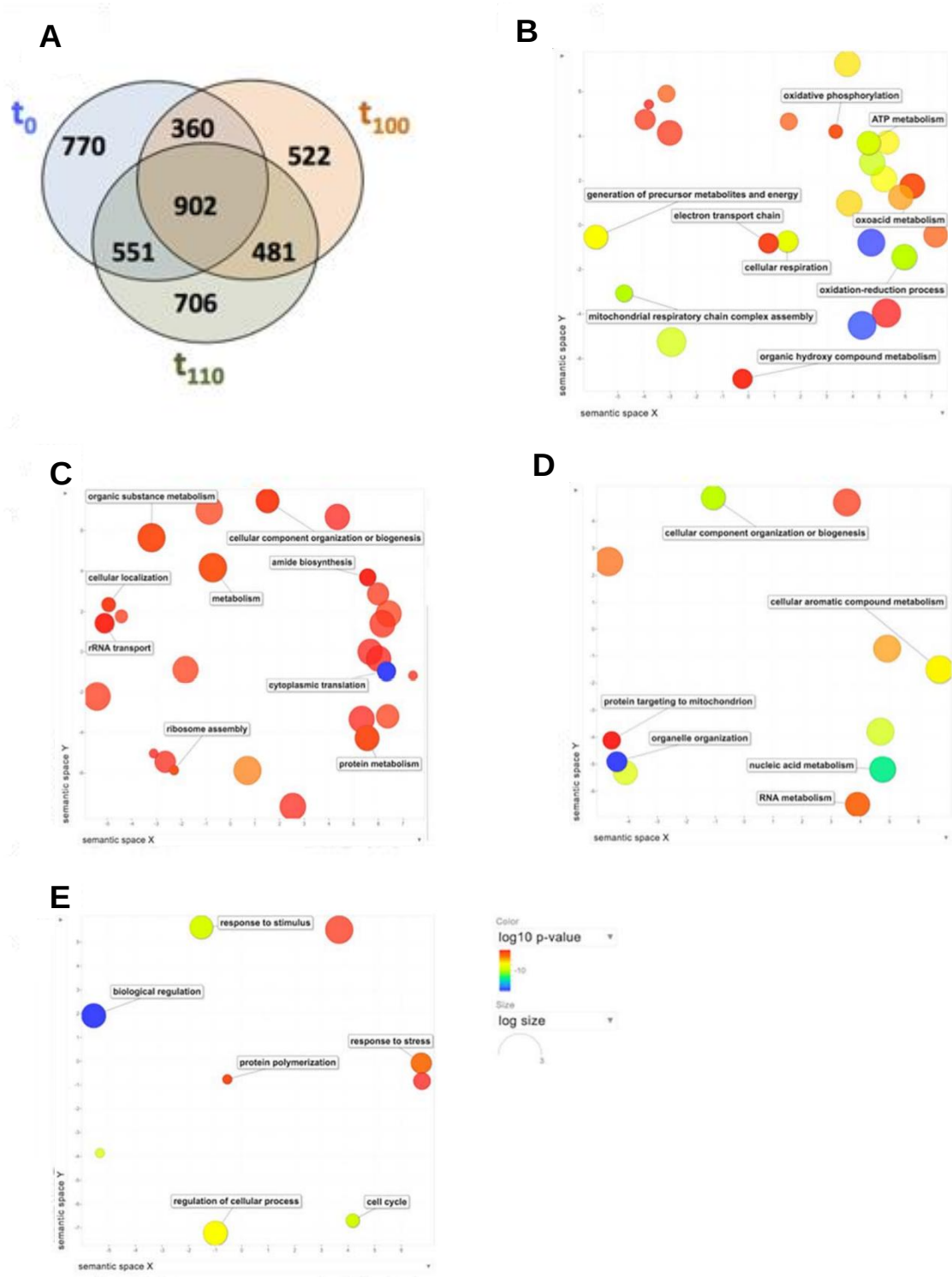


Figure ChIII- 6- Growth in glycerol induces two sets of genes, those that are common to all three populations (t_0, t_{100}, and t_{110}), also known as core genes, and those that are unique to one of the populations. **A)** Venn diagram identifying core genes and unique genes transcriptionally altered in one population. **B)** Semantic clustering of cellular processes enriched for genes that were transcriptionally altered in all three populations. The color of the bubbles represents the proportion of genes in a particular cellular process found transcriptionally altered (log P value), while the size indicates the frequency of the GO term in the organism. **C)** Semantic clustering of cellular processes enriched for genes that were transcriptionally altered in the population at t_0 growing in YPG. **D)** Semantic clustering of cellular processes enriched for genes that were transcriptionally altered in the population at t_{100} growing in YPG. **E)** Cellular processes enriched for genes that were transcriptionally altered in the population at t_{110} growing in YPG.

Altered core-genes and population-specific genes affected different cellular processes. Processes that are involved in the transition from fermentative to oxidative metabolism, controlling osmotic stress through transporters at the environment-cell interface, and energy production were enriched for core-altered genes. Examples of such genes include those encoding proteins from the mitochondrial respiratory chain complex, proteins involved in oxidative-reductive processes, cellular respiration, ion and substrates transport, and ATP metabolism (Figure ChIII-6B, Table S8). Cellular processes concerned with energy-saving such as translation, biosynthesis processes, protein metabolism, intracellular transport, and protein cellular localization genes were enriched for genes that were transcriptionally altered only in the t0 population (Figure ChIII-6C, Table S9). A subset of these processes was also enriched for genes transcriptionally altered in the t_{100} population (Table S10). Interestingly, the trafficking of proteins across the mitochondrial membrane, localization of proteins to the mitochondrion, and nucleic acid metabolism were particularly enriched for altered genes in the t_{100} population (Figure ChIII-6D). Finally, altered genes in t_{110}-population affected processes involved in stress stimuli response, cell cycle, and regulation, which were not enriched for altered genes from the t_0 and t_{100} populations (Figure ChIII-6E, Table S11).

e. Transcriptional response to glycerol-induced stress is mainly mediated by duplicated genes

Since *S. cerevisiae* bears 2240 duplicated genes (1120 pairs, roughly 32.8% of its genome), and gene duplication is often linked to the origin of novel adaptations, we investigated if duplicated genes have driven the transcriptomic responses to glycerol-induced stress. Of all the transcriptionally altered genes in the t_0 population, 1113 genes, corresponding to 40.5% of all transcriptionally altered genes, were duplicated genes, a proportion significantly greater than expected by chance (Binomial test: $P < 2.2 \times 10^{-16}$). Complementary to duplicated genes, 1634 of the transcriptional altered genes in this population (59.5% of all altered genes) were singletons, a proportion lower than expected by chance (binomial test: $P < 2.2 \times 10^{-16}$). Therefore, duplicated genes mostly drove transcriptional response to glycerol in the t_0 population. Duplicated genes in *S. cerevisiae* were generated through two different mechanisms, including whole-genome duplication (*N* = 554 pairs, corresponding to 49.5% of all duplicates), also known as WGDs (Wolfe and Shields 1997; Marcet-Houben and Gabaldón 2015) and small-scale duplications (*N* = 566 pairs, 50.5% of all duplicates), also known as SSDs (Fares *et al.* 2013; Keane *et al.* 2014). Since the evolution of duplicated genes is strongly dependent on the mechanism of duplication (Carretero-Paulet *et al.* 2013; Fares *et al.* 2013; Keane *et al.* 2014), we investigated whether differences in the presence of WGDs and SSDs existed

in the transcriptomic responses to glycerol. We found that WGDs were significantly more represented among transcriptionally altered genes in the t_0 population cultured in YPG than expected (binomial test: $P = 4 \times 10^{-4}$), while the opposite was true for SSDs. Similar results were observed for t_{100} and t_{110} populations (Table S12), although at t_{110} the population exhibited no differences between WGDs and SSDs (binomial test: $P = 0.34$).

We identified the core-altered duplicated genes, those duplicated genes that were altered in the t_0, $t_{100,}$ and t_{110} populations ($N = 369$, Figure ChIII-7A). Noticeably, the proportion of core-altered duplicated genes (369/1079 = 34.2%) was significantly greater than the proportion of core-altered singletons (533/1821 = 29.3%) (Fisher´s exact test: odds ratio $F = 1.26$, $P = 6.2 \times 10^{-3}$). Interestingly, a differential pattern was found between the transcriptional response of singletons and duplicates: while the majority of duplicates were down-regulated in cells cultured in YPG for all three populations, the opposite pattern was observed for singletons (Figure ChIII-7B). Indeed, the number of up-regulated duplicates in cells cultured in YPG was lower than expected at t_0, $t_{100,}$ and t_{110} (Figure ChIII-7B). Conversely, singleton genes exhibited a higher proportion of up-regulation than expected in all three populations (Figure ChIII-7B).

Examination of the cellular processes enriched for altered duplicated genes identified important differences between populations adapted to YPD and those adapted to YPG. Core-altered duplicates, those that were transcriptionally responsive to glycerol in all three populations, were distributed among cellular processes concerned with the transport of carbohydrates and organic substrates and respiration, including oxidation-reduction, energy derivation by oxidation of organic compounds (Figure ChIII-7C and Table S12). In the t_0 population challenged with glycerol most of the processes enriched for duplicated genes that were transcriptionally altered only in the t_0 population but not in the t_{100} or t_{110} populations were concerned with translation and biosynthetic processes (Figure ChIII-7D and Table S13). We found no cellular processes enriched for altered duplicated genes specifically in the t_{100} population but not in any of the other two time points. Finally, the t_{110} population adapted to YPG exhibited transcriptional alterations in a number of duplicates mostly involved in response to stress and stimuli as well as in the regulation of biological processes (Figure ChIII-7E and Table S14).

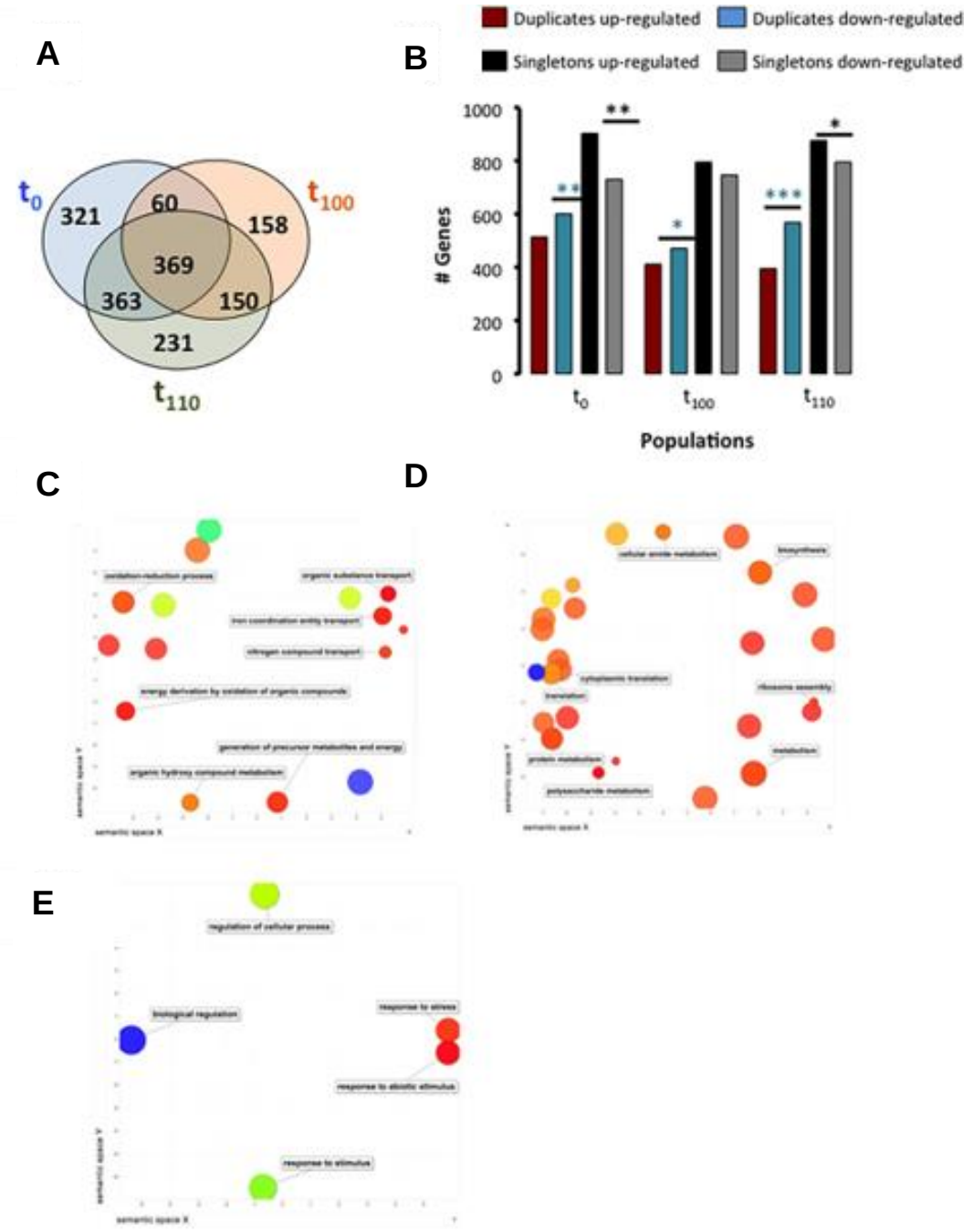


Figure ChIII- 7- Duplicated genes are more enriched than singletons for transcriptionally altered genes and show different response patterns to those of singletons under YPG growth conditions. **A)** Venn diagram identifying common duplicates response in the three populations (t_0, t_{100}, and t_{110}) and duplicates uniquely responsive to glycerol in each of the populations. **B)** Duplicates show more down-regulation than up-regulation, while singletons exhibit more up-regulation than down-regulation (*, **, *** refer to $P < 0.05$, $P < 0.01$, $P < 0.001$ under a binomial test, and colors identifies the most abundant pattern in the comparison between up- and down-regulated genes). **C)** Semantic clustering of cellular processes enriched for duplicated genes that were transcriptionally altered in all three populations. The color of the bubbles represents the proportion of genes in a particular cellular process found transcriptionally altered (log P value), while the size indicates the frequency of the GO term in the organism. **D)** Semantic clustering of cellular processes enriched for duplicated genes that were transcriptionally altered in in the population at t_0. **E)** Semantic clustering of cellular processes enriched for duplicated genes that were transcriptionally altered in in the population at t_{110}

It is worth noticing that analyses of up-regulated duplicates in the three populations also yielded very different outcomes for populations cultured in YPD (t_0 and t_{100} populations) compared to the population cultured in YPG: up-regulated duplicates in the t_0 and t_{100} populations were preferentially distributed in cellular processes of sexual reproduction, sporulation, transport, oxidative-reductive processes, and ascospore wall biogenesis, among others (Table S15 and S16). In contrast to this, in the t_{110} population many catabolic processes, including energy derivation by oxidation, TCA, and lipid catabolic processes were enriched for duplicated genes that were up-regulated in cells cultured in YPG (Table S17).

f. Adaptations of cellular metabolism of experimentally evolved S. cerevisiae populations

How metabolically divergent is the population adapted to glycerol from those responsive to glycerol but not adapted to it? We measured the 'metabolic' distance between the t_0, $t_{100,}$ and t_{110} populations using a simple measure of the distance between cellular processes terms (*Materials and Methods*). We analyzed separately the metabolic distance between pairs of the t_0, $t_{100,}$ and t_{110} populations for core-altered genes (i.e., those genes transcriptionally altered in all three populations), up-regulated and down-regulated genes in each of the populations (Table ChIII-1). Altered genes in cells from the t_0 and t_{100} populations, were metabolically closer to each other than each was to t_{110} population. Up-regulated genes that included singletons and duplicates, in the t_{100} population exhibited greater metabolic distance to the t_0 population than it did to the t_{110} population adapted to glycerol. For down-regulated genes, on the other hand, the distance between t_0 and t_{100} populations was shorter than the distance of each to t_{110} population. Using only core duplicated genes, those duplicated genes transcriptionally altered in all three populations, the ancestral population showed low metabolic distance to t_{100} population, and t_{100} population was metabolically close to t_{110} population, but t_0 was distant from t_{110} population. Both the up-regulated and down-regulated sets of duplicated genes exhibited a greater distance of the t_{110} population adapted to glycerol to any of the other two populations than this distance between t_0 and t_{100} populations.

Table ChIII- 1 - Metabolic distances between populations t_0, t_{100} and t_{110}

Genes	**Metabolic Distance ($MD_{i,j}$)**								
	Core altered			**Up-regulated**			**Down-regulated**		
	t_0 *vs* t_{100}	t_0 *vs* t_{110}	t_{100} *vs* t_{110}	t_0 *vs* t_{100}	t_0 *vs* t_{110}	t_{100} *vs* t_{110}	t_0 *vs* t_{100}	t_0 *vs* t_{110}	t_{100} *vs* t_{110}
All	0.27	0.37	0.29	0.51	0.34	0.28	0.39	0.44	0.46
Duplicates	0.09	0.38	0.09	0.37	0.51	0.52	0.30	0.63	0.60

5. Discussion

a. *Exposure to glycerol-induced stress triggers a genome-wide transcriptomic response*

Unicellular organisms, in particular non-motile ones, must have developed a number of metabolic strategies to sense micro-changes in the environment and prepare to combat environmental fluctuations. Here, we show that even small increases in the environmental concentrations of glycerol lead to dramatic transcriptional re-programming of the yeast *S. cerevisiae*. This re-programming affects a great proportion of the genes and it extends to cellular processes affecting all the organelles in the cell. Such genome-wide re-programming has been shown before to take place under a number of stress conditions. For example, in an experiment of adaptation of *S. cerevisiae* through experimental evolution under glucose-limited conditions, authors observed a large expression re-programming affecting several hundreds of genes (Ferea *et al.* 1999). Following these experiments, there are two kinds of transcriptomic responses, which vary with regards to the time of response, from very quick responses to very slow ones (Yosef and Regev 2011). These two types of responses have been shown to involve a large number of genes (Taymaz-Nikerel *et al.* 2016), in good agreement with the data presented in this study. Indeed, our results show that subjecting an initial population of *S. cerevisiae* to mild glycerol-induced stress leads to the alteration of the expression of hundreds of genes in a very short time. Most such genes are linked to a number of pathways that include positive and negative regulators of protein kinase A and other transport pathways that are involved in buffering the effects of glycerol variation around the membrane, including genes encoding transport proteins for degradation of unfolded proteins such as SSA1, ERP2, PMT2, etc.

Two well-differentiated patterns are observed during the transcriptional alterations under stress. On the one hand, genes involved in responding directly to exposure of cells to glycerol and the activation of the respiratory metabolism of the yeast cell are up-regulated. On the other hand, genes involved in the ribosomal biogenesis or in energy-dependent processes, such as translation and transcription are down-regulated. These patterns are in agreement with the hypothesis that such a transcriptional response does not counterbalance the effects of stress on fitness but are a first step towards adapting the cell to the new perturbed environment. Accordingly, despite this large re-programming of the cell, the growth rate of the population declines under glycerol in the ancestral (t_0) population when compared to its growth rate in YPD. This result is concordant with the lack of correlation between the transcriptional response and fitness gain or buffering of the effects of stress on cell's growth rate (Giaever *et al.* 2002; Giaever

and Nislow 2014). On the contrary, it seems that genes involved in environmental stress response are correlated with a decline in cell growth rate (Gasch *et al.* 2000; Regenberg *et al.* 2006; Castrillo *et al.* 2007; Fazio *et al.* 2008; Brauer *et al.* 2008).

b. The genomic background of a S. cerevisiae population impacts glycerol stress responses

An important question is how does the genetic background of a cell influence its response capacity to environmental perturbations? In this study, the experimental evolution of a population of *S. cerevisiae* for a greater number of generations in YPD (t_{100} population) demonstrates that the genetic composition of the population largely influences the transcriptional profile of the cells under stress. Indeed, we observed that only a fraction of the genes responding to stress in the evolved population are also responsive to stress in the ancestral non-evolved population. Since the ancestral population was founded from a single clone, hence lacked genetic variability, the difference in the transcriptional profile between the evolved and the ancestral population is likely the result of a change in the genotype of the population. An important result deriving from our analyses is that the population evolving in YPD, but not in the presence of glycerol, exhibits higher fitness when cultured in the presence of glycerol than its parental non-evolved population. A possible explanation for this finding is that the evolved population has explored a wide range of genotypes some of which may be adaptive to the environment containing glycerol, a phenomenon encapsulated within the term exaptation (Fares 2015c). The high dependence of the transcriptional plasticity of yeast on the genetic composition of the population supports previous suggestions that the transcriptional reprogramming of the cell does not occur in a stress-specific manner (Causton *et al.* 2001; Ideker *et al.* 2001; Stern *et al.* 2007; Cormier *et al.* 2010). Instead, the transcriptional response to stress may be a biological property emerging from a universal feature underlying regulatory networks. Since biological systems are known to bear distributed robustness—i.e., the product of a specific metabolic pathway could be achieved through many other alternative and unrelated pathways (Wagner 2015), the differential transcriptional profiles of the evolved and non-evolved populations in response to glycerol may be due to this distributed robustness, which in turn can lead to exaptations.

c. Is glycerol-stress response a combination of adaptive responses and system-level emerging properties?

The evolution of the yeast population in the presence of glycerol (i.e., the t_{110} population) yields results supporting the regulatory potential of yeast to improve its adaptation to changing environments. We distinguish between two main responses,

those generated against instantaneous environmental perturbations (i.e., quick responses) and those emerging from the selective fixation of adaptive regulatory changes that emerge when a population is subjected to constant stress. Accordingly, the quick response involves a set of genes that are de-regulated in the three populations (t_0, t_{100} and t_{110} populations). However, the adaptive response generated in the population evolved in YPG (t_{110} population) mostly includes genes involved in stress response and regulation processes. This late induction of genes involved in stress response has been previously observed in populations of yeast subjected to high temperatures (Gasch *et al.* 2000; Causton *et al.* 2001).

d. Transcriptional response to glycerol stress is driven mainly by duplicated genes

Determining whether or not there is potential to adapt to novel environments through rapid evolution of regulatory programs is an important aim in evolutionary biology, yet such an aim has remained obscure owing to the difficulty of mapping phenotypes to genotypes or to assign transcriptional changes to phenotypic variations. A clear pattern observed in this study links transcriptional variation to the phenotypic adaptation to glycerol, namely that transcriptionally altered genes in the presence of glycerol are enriched for duplicated genes. The classic theory of evolution by gene duplication states that after the duplication of a gene, one of the copies explores novel genotypes freed from selection constraints because its sister copy performs the ancestral well-adapted function (Ohno 1970, 1999; Lynch and Conery 2000; Conant and Wolfe 2006). This genotypic exploration affects both the functional and regulatory features of genes. We hypothesize that increasing the regulatory plasticity after gene duplication is a more likely scenario than increasing the functional plasticity, as the effects of the former are more likely to be the subject of rapid selection in changing environments while the latter requires longer evolutionary times to be selected for (Keane *et al.* 2014; Fares 2015c). Eventually, the expression plasticity may lead to functional plasticity because the rate of evolution of a gene is strongly determined by its expression level, a link that has been observed in all organisms examined so far from viruses to mammals (Krylov *et al.* 2003; Rocha and Danchin 2004; Drummond *et al.* 2005, 2006; Drummond and Wilke 2008; Pagan *et al.* 2012; Zhang and Yang 2015). Testament to this is the rapid expression divergence between the copies of a duplicated gene (Blanc and Wolfe 2004b; Li *et al.* 2005; Conant and Wolfe 2006; Thompson *et al.* 2013). The selective enrichment of glycerol-responsive duplicate genes raises the possibility that duplicated genes have led to major specializations in the regulatory response to changing environments, perhaps through the acquisition of novel functions or novel interactions in the cell (Fares *et al.* 2013).

e. Implications for S. cerevisiae ecology, and concluding remarks

Our data reveal a strong link between the regulatory re-programming of the cell and the environmental change in glycerol concentration. These data uncover a rapid genome-wide de-regulation of genes involved in fundamental metabolic processes in the cell, with the up-regulation of genes that allow a rapid shift from a fermentative to a respiratory metabolism as well as genes encoding membrane transporters of solutes and ions. In contrast, down-regulated genes in cells exposed to glycerol-induced stress affect energetic-costly processes, such as translation and transcription. We also identify a fine-tuned re-programming of the transcriptome as an adaptive response to 10-day exposure to glycerol-induced stress. Most of the transcriptomic responses to glycerol are driven by duplicated genes, revealing an unprecedented fundamental role of these genes in the evolution of adaptive responses to environmental perturbations.

In natural habitats of *S. cerevisiae* (which may include soils, plant surfaces, saline environments, or sugar-rich milieu), the yeast cell – like those of many microbes – is likely to experience multiple, concomitant stresses. In high-sugar substrates of >0.900 water activity, where *S. cerevisiae* can thrive, some of these stresses are self-imposed (ethanol stress, acetaldehyde stress, and organic-acid stress) and others are not (sugar-induced reduction of water activity, antimicrobials produced by competitors, sub- or supra-optimal temperatures, etc). The interplay between such parameters, community dynamics, responses and adaptations of *S. cerevisiae*, and its ability to grow and/or remain metabolically active has been the focus of recent studies by Cray et al. (2013 and 2015). The findings of the current study demonstrate that glycerol, which is produced by the yeast cell as a stress protectant, can also cause collateral damage, and also shows how *S. cerevisiae* can respond physiologically and can also evolutionarily adapt to this additional stress burden. A series of intriguing scientific questions remain unanswered: What is the role of duplicated genes in allowing quick switches of regulatory programs in *S. cerevisiae* subjected to environmental perturbations? What is the ecological cost of a rapid adaptation to glycerol stress? What is the impact of the regulatory re-programming of *S. cerevisiae* under stress on the evolution of protein-coding genes and the origin of novel functions? How plastic is the *S. cerevisiae* transcriptome to fluctuating environments?

6. Accession numbers

RNA raw reads are available from the Sequence Read Archive with accession number SRP074821.

7. Supplementary data

Supplementary data sets are available at *Environmental microbiology* online. Data Set S1 is also available in the appendix of this manuscript. **Data Set 1:** Role of duplicated genes in the response to glycerol-induced stress. **Table S1:** Enrichment of GO-terms of cellular processes for genes that are significantly transcriptionally altered under glycerol-induced stress compared to glucose. **Table S2:** Enrichment of GO-terms of cellular processes for genes that are significantly up-regulated under glycerol-induced stress compared to glucose. **Table S3:** Enrichment of GO-terms of cellular processes for genes that are significantly down-regulated under glycerol-induced stress compared to glucose. **Table S4:** Enrichment of GO terms linked to cellular processes for genes transcriptionally altered in *S. cerevisiae* populations isolated at t_0 and t_{100}. **Table S5:** Enrichment of GO terms linked to cellular processes for genes transcriptionally altered in *S. cerevisiae* populations isolated at t_0 but not t_{100}. **Table S6:** Enrichment of GO terms linked to cellular processes for genes transcriptionally altered in *S. cerevisiae* populations isolated at t_{100} but not t_0. **Table S7:** List of *S. cerevisiae* genes that are transcriptionally altered under glycerol-induced stress in *S. cerevisiae* populations isolated at t_0, t_{100}, and t_{110}.**Table S8:** GO terms linked to cellular processes that are enriched for transcriptionally altered genes under glycerol-induced stress in *S. cerevisiae* populations isolated at t_0, t_{100}, and t_{110}. **Table S9:** GO terms linked to cellular processes that are enriched for transcriptionally altered genes under glycerol-induced stress in *S. cerevisiae* populations isolated at t0, but not in populations isolated at t_{100} or t_{110}. **Table S10:** GO terms linked to cellular processes that are enriched for transcriptionally altered genes under glycerol-induced stress in *S. cerevisiae* populations isolated at t_{100}, but not in populations isolated at t_0 or t_{110}. **Table S11:** GO terms linked to cellular processes that are enriched for transcriptionally altered genes under glycerol-induced stress in *S. cerevisiae* populations isolated at t_{110}, but not in populations isolated at t_0 or t_{100}. **Table S12:** GO terms linked to cellular processes that are enriched for transcriptionally altered duplicated genes under glycerol-induced stress in *S. cerevisiae* populations isolated at t_0, t_{100}, and t_{110}. **Table S13:** GO terms linked to cellular processes that are enriched for transcriptionally altered duplicated genes under glycerol-induced stress in *S. cerevisiae* populations isolated at t0, but not in populations isolated at t_{100} or t_{110}. **Table S14:** GO terms linked to cellular processes that are enriched for transcriptionally altered duplicated genes under glycerol-induced stress in *S. cerevisiae* populations isolated at t_{110}, but not in populations isolated at t_{100} or t_0. **Table S15:** GO terms linked to cellular processes enriched for up-regulated duplicated genes under glycerol-induced stress in *S. cerevisiae* populations isolated at t_0. **Table S16:** GO terms linked to cellular processes enriched for up-regulated duplicated genes under glycerol-induced stress in *S. cerevisiae* populations isolated at t_{100}. **Table**

S17: GO terms linked to cellular processes enriched for up-regulated duplicated genes under glycerol-induced stress in *S. cerevisiae* populations isolated at t_{110}.

CHAPTER IV – Analysis of the transcriptional reprogramming in yeast due to short and chronic exposure to Ethanol stress: Cellular responses and evolved adaptations.

A version of this chapter has been published as:

Sabater-Muñoz, B., Mattenberger, F., Fares, M.A., Toft, C. (2020) *Transcriptional rewiring, adaptation, and the role of gene duplication in the metabolism of ethanol of Saccharomyces cerevisiae.* ***mSystems, 5(4): e00416-20.***

1. Abstract

Ethanol is the main by-product of yeast sugar fermentation that affects microbial growth parameters, being considered a dual molecule, a nutrient and a stressor. Previous works demonstrated that the budding yeast arose after an ancient hybridization process resulted in a tier of duplicated genes within its genome, many of them with implications in this ethanol “produce-accumulate-consume” strategy. The evolutionary link between ethanol production, consumption, and tolerance versus ploidy and stability of the hybrids is an ongoing debatable issue. The implication of ancestral duplicates in this metabolic rewiring, and how these duplicates differ transcriptionally, remains unsolved. Here, we study the transcriptomic adaptive signatures to ethanol as a nonfermentative carbon source to sustain clonal yeast growth by experimental evolution, emphasizing the role of duplicated genes in the adaptive process. As expected, ethanol was able to sustain growth but at a lower rate than glucose. Our results demonstrate that in asexual populations a complete transcriptomic rewiring was produced, strikingly by downregulation of duplicated genes, mainly whole-genome duplicates, whereas small-scale duplicates exhibited significant transcriptional divergence between copies. Overall, this study contributes to the understanding of evolution after gene duplication, linking transcriptional divergence with duplicates’ fate in a multigene trait as ethanol tolerance.

2. *Introduction*

Sensing and responding to the environment are central parts of metabolism of almost all unicellular organisms. During evolution, some budding yeasts (*Saccharomycotina*) faced a new source of carbon (sugars) in a new niche (nectar or fruits from the recently emerged [100 million years ago {MYA}] Angiosperms), a fact that has been postulated at the origin of fermentative metabolism with ethanol as the main end product. Behind this biological innovation has been unveiled gene duplication at two scales, whole-genome (WGD) and small-scale (SSD) duplication, and genome shrinkage after it, as evolutionarily driving genomic changes (reviewed in Dittmar and Liberles 2010; Wolfe and Shields 1997; Fares et al. 2013). The baker's yeast *Saccharomyces cerevisiae* is one of the most biotechnologically important species, being able to tolerate higher ethanol levels during fermentation than any other microbe (Ding *et al.* 2009; Voordeckers *et al.* 2015; Snoek *et al.* 2016).

Under sugar scarcity, yeast can switch from fermentative to respiratory metabolism using ethanol and glycerol as nonfermentative carbon sources to support growth (Gancedo 1998; Schüller 2003). This ethanol "make-accumulate-consume" strategy (Crabtree effect) has been partially linked to the yeast evolutionary origin history. However, ethanol in particular endangers the yeast metabolic activity, survival, cell morphology, growth ability, and biomass production. Ethanol also exhibits a general cell toxicity that yeasts used to control competitors' growth. This duality (nutrient and stressor) generates great concerns in the biotechnological industries (by its applications) and in the scientific community (by its molecular basis), highlighting the importance of systems biology studies (reviewed in references Haas et al. 2019; Dashko et al. 2014; Mullis et al. 2019; Yang et al. 2011).

Experimental evolution, in particular with *Escherichia coli* and *S. cerevisiae*, has been of unprecedented relevance to unveil evolutionary pathways underlying the origin of adaptations, including as examples the adaptation of *E. coli* to citrate in the known Lenski evolution experiment (Blount *et al.* 2008, 2012; Barrick *et al.* 2009; Turner *et al.* 2015; Lenski *et al.* 2015) and heat stress, nutrient limitations, antibiotic treatment, or tolerance to glycerol in *S. cerevisiae* (Gresham *et al.* 2008; Toprak *et al.* 2012; Yona *et al.* 2012; Oz *et al.* 2014; Mattenberger, Sabater-Muñoz, Hallsworth, *et al.* 2017; Strauss *et al.* 2019; Sandberg *et al.* 2019). Its use to understand the adaptation to ethanol has been addressed in only a few studies, while using ethanol as additional carbon source (Avrahami-Moyal *et al.* 2012; Voordeckers *et al.* 2015; Snoek *et al.* 2016). Indeed, only one work revealed the genomic dynamics including point mutations, copy number variation (gene duplication), ploidy changes, and clonal interference mix in a complex

evolutionary pathway that increases tolerance to ethanol (Voordeckers *et al.* 2015). Nonetheless, the transcriptional rewiring occurring during this response to ethanol and its importance in comparison with the contribution of genomic changes have not been explored. Indeed, the implication in ethanol response and adaptation of duplicates, from a transcriptional perspective, have been only marginally explored (Oz *et al.* 2014; Strauss *et al.* 2019). The interplay between duplicates and transcriptional rewiring remains unknown. It also remains elusive whether and how *S. cerevisiae* could optimize the use of ethanol as nonfermentative carbon source.

In this study, we undertake the challenge of elucidating the role of transcriptional rewiring to the response and adaptation to ethanol (as sole carbon source) in *S. cerevisiae* and revealing the link between gene duplication and ethanol usage. As already mentioned, previous studies revealed an unprecedented complexity in the genomic dynamics underlying adaptation to ethanol but, however, did not address the implication of transcriptional reprogramming of the ancestral duplicates (Avrahami-Moyal *et al.* 2012; Voordeckers *et al.* 2015; Snoek *et al.* 2016). Here, we evolved clonal populations of *S. cerevisiae* using glucose as carbon source and challenged them to use ethanol as sole carbon source in short and long (ethanol adaptive laboratory evolution) responses. We reveal the transcriptional reprogramming basis and the interplay of this with gene duplication in the response and adaptation to ethanol.

3. Materials and Methods

a. Yeast culture and experimental evolution

The *Saccharomyces cerevisiae* strain Y06240 (BY4741: *Mat**a***; *his3D1*; *leud2D0*; *met15D0*; *ura3D0*; *msh2*::*kanMX4*) was used as described previously (Mattenberger, Sabater-Muñoz, Toft, and Fares 2017; Mattenberger, Sabater-Muñoz, Hallsworth, *et al.* 2017). Briefly, a homogeneous population founded by growing a colony in a liquid culture of rich medium (YPD: 2% (w/v) bacto peptone, 1% (w/v) yeast extract, 2% (w/v) dextrose; supplemented with kanamycin) (t_0) was evolved through daily bottlenecks (1%) for 100 days (t_{100}; ~660 generations), in 5 ml of YPD medium in 50-ml Corning tubes, at 28°C and 220 rpm. From passage 100 (t_{100}), the population "a1" was divided into two sublines, each with three biological replicates. One subline was grown in YPD medium as control (lines "Da1"), whereas the second subline (lines "Ea1") was grown in a medium containing 3% ethanol as the sole carbon source (YPE: 3% (v/v) ethanol, 2% (w/v) Bacto peptone, 1% (w/v) yeast extract; supplemented with 100 µg/ml kanamycin). The populations were evolved for another 10 passages, with a daily bottleneck of 10% of the population, in 5 ml of the corresponding medium, as indicated previously. Every 10 passages, a fossil record of each line was established by preserving the entire population in 25% glycerol solution at −80°C (Figure ChIV-1A).

b. Growth characterization

Growth parameters for t_0, t_{100}, and t_{110} were obtained using the Bioscreen C plate-reader system (Oy Growth Curves Ab Ltd., Helsinki, Finland) as described in Mattenberger, Sabater-Muñoz, Toft, et al. (2017). Briefly, each time point was precultured overnight at 28°C, from the corresponding fossil record, and used to inoculate 200µl of fresh medium (YPD and/or YPE) to an initial optical density at 600 nm (OD_{600}) of 0.06 to 0.07, distributed in 100-well honeycomb plates, with 6 to 7 technical replicates. The experiment was run for 78 h at 28°C with continuous shaking (high level) and taking OD_{600} measurements (brown filter) every 15 min. Each run contained at least 3 controls for each medium (uninoculated fresh medium). The data were analyzed with Growthcurver v.0.3.0 under R-studio (Sprouffske and Wagner 2016).

c. RNA extraction and transcriptomic analysis

The RNA profiling was performed at the t_0, t_{100}, and t_{110} time points as indicated in Figure ChV-1A, following the same procedures as previously used (Mattenberger, Sabater-Muñoz, Toft, and Fares 2017; Mattenberger, Sabater-Muñoz, Hallsworth, *et al.* 2017). rRNA-depleted RNA (Illumina) libraries were constructed and sequenced at the

Genomic Core Facility at Servicio Central de Soporte a la Investigacion Experimental (SCSIE) from the University of Valencia, Spain. Reads (trimmed) were aligned with Bowtie2 (up to two mismatches accepted) to the reference S288c strain genome (only coding sequences [CDS]). Statistical assessment of differential gene expression was done with edgeR (Robinson *et al.* 2010), setting false-discovery rate (FDR) at <0.005, and applying BY correction for P-value (0.005).

d. Identification of duplicated genes, functional classification, and visualization

Paralogous pairs of duplicated genes were divided into two groups according to their origin mechanism: whole-genome duplicates (WGDs) or small-scale duplicates (SSDs). WGDs (555 pairs) were extracted from the reconciled YGOB list (Yeast Gene Order Browser, last accessed March 2018; http://wolfe.gen.tcd.ie//ygob (Byrne and Wolfe 2005)). SSDs (560 pairs) were identified after best reciprocal hits from all-against-all BLAST searches using BLASTP with an E value cutoff of $1E^{-5}$ and a 50-bit score (Altschul *et al.* 1997), selecting only those that exhibit a distribution of synonymous substitutions similar to WGDs (Fares *et al.* 2013; Keane *et al.* 2014). Differential expressed genes were further classified according to their gene ontology (GO) term as implemented in the R package clusterProfiler (Yu *et al.* 2012), followed by an enrichment analysis with a P-value cutoff of <0.01 and with the P-value being adjusted with the Benjamini and Hochberg method (Benjamini and Hochberg 1995).

e. Software

Unless otherwise indicated, statistics were performed using the appropriate packages in R v3.5.1 (R Core Team, 2018).

4. Results

a. Phenotypic changes of S. cerevisiae in response and adaptation to ethanol

At time points t_0, t_{100}, and t_{110}, we characterized growth parameters of *S. cerevisiae* populations in the standard medium (yeast extract-peptone-dextrose [YPD]) and stressful medium (yeast extract-peptone-ethanol [YPE]), using optical density measurements (Figure ChIV-1A). The mean maximum growth rate (μ_{max}) was significantly lower at time t_0 in YPE (μ_{max} ± standard deviation of the mean [SDm] = 0.1303 ± 0.0093 h^{-1}) than in YPD (μ_{max} ± SDm = 0.2096 ± 0.0343 h^{-1}; Wilcoxon rate test, $P = 5.8 \times 10^{-4}$). This difference between growth rates was also observed in all the evolved lines at all time points (Figure ChIV-1B; see also Figure S1 in the supplemental material). Diversifying the population for approximately 660 generations (t_{100}) increased the growth rate in YPD for all lines. Furthermore, the populations also increase their carry capacity after diversification (Figure ChIV-1C; Figure S2). However, only one of the lines increased its growth rate in YPE; the other two retained a similar growth rate as the ancestral population (Figures S1 and S2). All lines, except one, at time t_{110} reduced the growth rate after just 33 generations. The populations evolved in YPE and when challenged to grow in YPD showed a higher growth rate than the control population evolved in YPD. This difference comes from the growth rate recovery of one of the evolved populations in YPE. The rest of the populations in t_{100} perform similarly. Overall, the evolved population in YPE reduced their growth rate in YPE compared to the evolved populations in YPD but increased their carry capacity (Figures S1 and S2).

b. Up- and downregulation in response and adaptation to ethanol

Transcriptome sequencing (RNAseq) was conducted in populations t_0, t_{100}, and t_{110}, in YPD and/or challenged with ethanol (YPE) (Figure ChIV-1A). The exposure to ethanol led to the upregulation (fold change [FC] in the expression of the genes >25%, false-discovery rate [FDR] <0.005) of 833 and 1,389 genes compared to the same population grown in YPD for t_0 and t_{100}, respectively. Of the 833 genes upregulated in t_0, 557 (66.9%) were also upregulated in t_{100} (Figure S3A). Adaptation to ethanol stress for 10 passages led to the upregulation of 1,694 genes, of which 437 were also upregulated in t_0 and t_{100} (we call these core upregulated genes) (Figure S3A). The exposure to ethanol led to the downregulation of 751 and 940 genes compared to the same population grown in YPD for t_0 and t_{100}, respectively. Of the 751 downregulated genes in t_0, 326 (43.4%) were also downregulated in t_{100} (Figure S3B). The adaptation to ethanol stress led to the downregulation of 1,391 genes, of which 222 were also downregulated in t_0 and t_{100}.

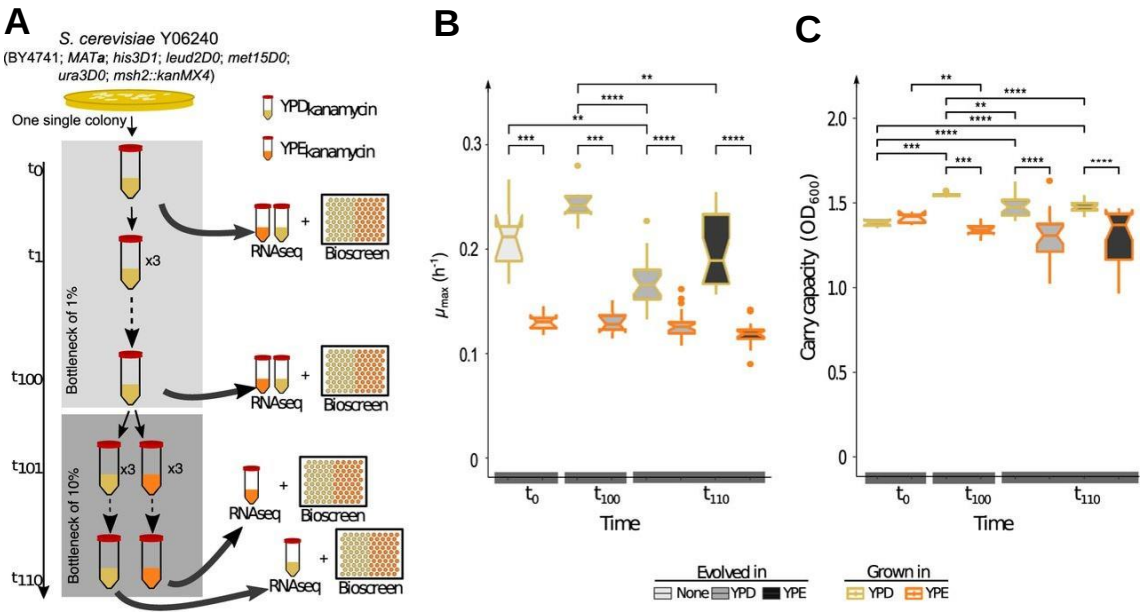


Figure ChIV- 1 - Experimental adaptive laboratory evolution scheme and phenotypic characterization in Saccharomyces cerevisiae Y06240. **A)** Experimental layout. From a single *S. cerevisiae* Y06240 (*Mat**a***; *his3D1; leud2D0; met15D0; ura3D0; msh2::kanMX4*) colony, we derived a population in liquid YPD medium (called population t_0). This population was split into 3 replicates and evolved in YPD for 100 passages (approximately 660 generations) by daily transferring 1% (0.5 ml) to a new tube (50 ml) with fresh YPD medium (4.5 ml) (we called this population t_{100}). After passage 100, we started the adaptive evolution, by splitting up the evolutionary experiment into two: one continuing to evolve in YPD (with 2% glucose as carbon source) and the other replacing the glucose with 3% ethanol (medium YPE). Populations were evolved for 10 passages with a daily 10% bottleneck (approximately 33 generations; we called this population t_{110}). In the experimental scheme, the points at which phenotypic characterization and transcriptome changes (RNAseq) were carried out are indicated. **B)** Phenotypic characterization was performed by characterization of population growth curves. The maximum growth rate (h^{-1}) of each population was determined at each control time point (t_0, t_{100}, and t_{110}) in their evolving medium (YPD or YPE) and in the challenging medium (YPE or YPD). **C)** "Carry capacity" of each population (OD_{600}) was also determined for each population and each control time point in their evolving medium and in the challenge one. Significant differences of each growth parameter are indicated as *, **, ***, and ****, when the probabilities are $P < 0.05$, $P < 0.005$, $P < 10^{-3}$, and $P < 10^{-4}$, respectively, using a Wilcoxon rank test.

c. Low overlap in the transcriptomic response and adaptation to ethanol

The number of upregulated genes among the populations t_0 and t_{100} (67% of t_0 upregulated genes are also upregulated in t_{100}) was high for the t_0 population, but t_{100} showed twice as many upregulated genes as t_0, perhaps indicating that experimental evolution in YPD for 100 passages has involved significant polymorphism in the transcriptomic reprogramming of cells in this population. Only 437 genes were core upregulated genes in all three populations. Populations t_{110}, the populations derived from t_{100} and evolved for 10 days in ethanol, showed an overlap of only 883 upregulated genes with their parental t_{100} populations despite the low number of passages separating them (Figure S3A).

To determine whether the functions affected by the transcriptomic responses have changed among populations, we performed an analysis of Gene Ontology (GO) terms of the set of upregulated genes. Populations at t_0 and t_{100} exhibited enrichment for upregulated genes ($P < 0.01$) in similar functional categories, affecting mainly the "oxidation-reduction process," "drug metabolic process," "aerobic respiration," "proton transmembrane transport," "mitochondrion organization," "small-molecule catabolic process," "oxidoreductase activity," "cofactor binding," and "proton transmembrane transporter activity" (Figure ChIV-2). The analysis, on the other hand, of GO term enrichment for upregulated genes in t_{110} population led to a somewhat different result. There was some overlap of enriched GO terms from t_0 and t_{110}, but more importantly, a number of GO term enrichments were specific for t_{110}. They include the terms "energy derivation by oxidation of organic compounds," "response to oxidative stress," and "cellular response to oxidative stress and response to inorganic substances" (Figure ChIV-2).

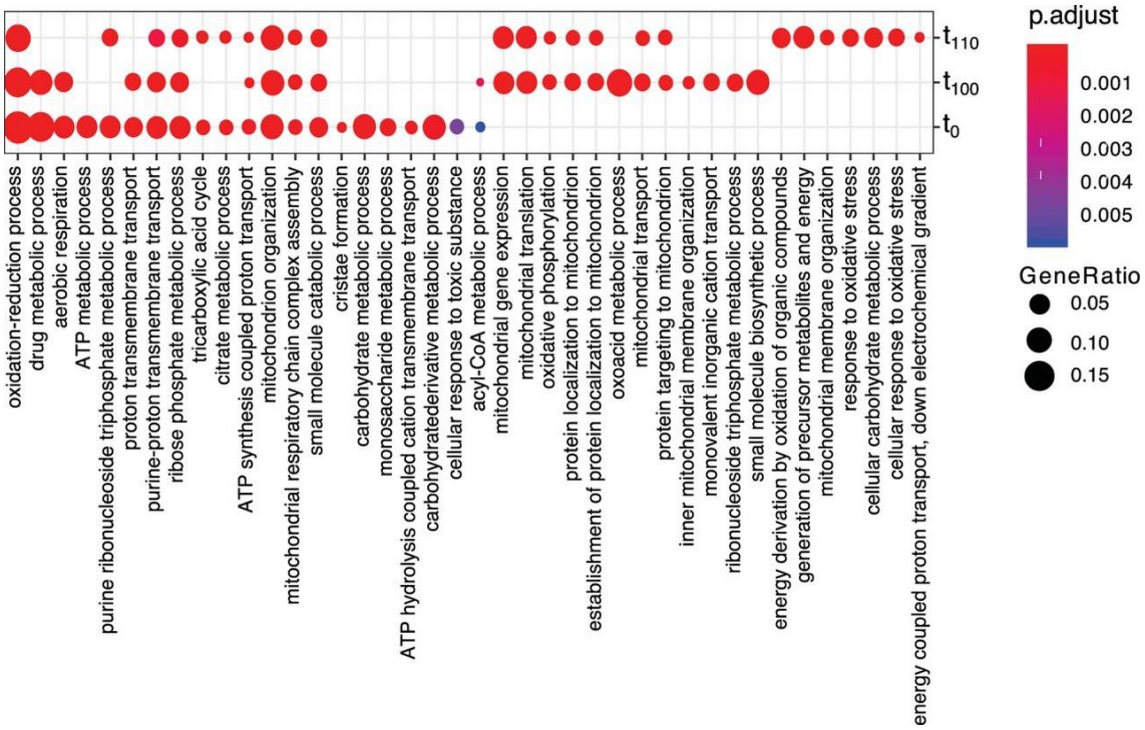


Figure ChIV- 2 - Biological processes enriched due to the use of 3% ethanol as the sole carbon source. Enrichment analysis of functional categories (biological process) for upregulated genes in YPE compared to YPD, at the three time points (t_0, t_{100}, and t_{110}), was performed with clusterProfiler.

The GO term analysis for downregulated genes showed even less overlap between the three populations (Figure S4). Population t_0 had enrichment in "cytoplasmic translation and ribosome biogenesis," which was also enriched at population t_{110}. Furthermore, population t_{110} had "ncRNA processing and methylation" enriched for

downregulated genes. In contrast, population t_{100} had only a few GO terms enriched, including “carbohydrate transport” and “nucleic acid phosphodiester bond hydrolysis,” with none of them overlapping the other two populations.

d. Duplicated genes encode rapid responses and adaptations to ethanol

Since duplicated genes are involved in the origin of new functions (Ohno 1970; Hakes *et al.* 2007; Van Hoek and Hogeweg 2009; Fares *et al.* 2013; Keane *et al.* 2014; Wolfe 2015), we sought to investigate if the response to ethanol, as the sole carbon source, was mainly driven by duplicates, differentiating between WGDs (Wolfe 2015) and SSDs (Hakes *et al.* 2007) in our analyses.

Population t_0 exhibited 312 duplicate (14.2%) and 524 singleton (11.6%) genes out of the 833 upregulated genes. The proportion of upregulated duplicates was higher than that of singletons (Fisher’s exact test: odds ratio $F = 1.22$, $P = 0.0071$) (Figure ChIV-3A). We also observed a higher expression fold change (FC) difference in duplicates (median FC = 1.343) than in singletons (median FC = 1.244) (Wilcoxon rank test: $P = 0.0215$) (Figure ChIV-3B). We found no difference in the response of duplicates when analyzing their origin (312 = 156 WGDs + 156 SSDs) (Fisher’s exact test: odds ratio $F = 1.002$, $P = 1$). Likewise, no difference was observed in the expression fold change between WGDs (median FC = 2.92) and SSDs (median FC = 3.13) (Wilcoxon rank test: $P = 0.65$).

Population t_{100} showed no difference in the response to ethanol between upregulated duplicates (494) and singletons (1,000) (Fisher’s exact test: odds ratio $F = 1.01$, $P = 0.807$). Remarkably, while the proportion of duplicates that were upregulated increased 58% after 100 passages of evolution, the proportion of singletons increased 92% in comparison with the parental population t_0 (Figure ChIV- 3A). No difference in expression fold change was observed between duplicates and singletons, nor between WGDs and SSDs.

Population t_{110} exhibited 578 upregulated duplicates and 1,116 upregulated singletons, not being significantly different (Fisher’s exact test: odds ratio $F = 1.06$, $P = 0.285$). Interestingly, upregulated duplicates (median FC = 1.346) saw a higher expression fold change than singletons (median FC = 1.298) (Wilcoxon rank test: $P = 0.0288$). We found significantly more WGDs (319) responding to ethanol than SSDs (259) (Fisher’s exact test: odds ratio $F = 1.23$, $P = 0.0248$), but no difference of expression fold change between the two (WGDs: median FC = 1.408; SSDs: median FC = 1.274; Wilcoxon rank test: $P = 0.0586$).

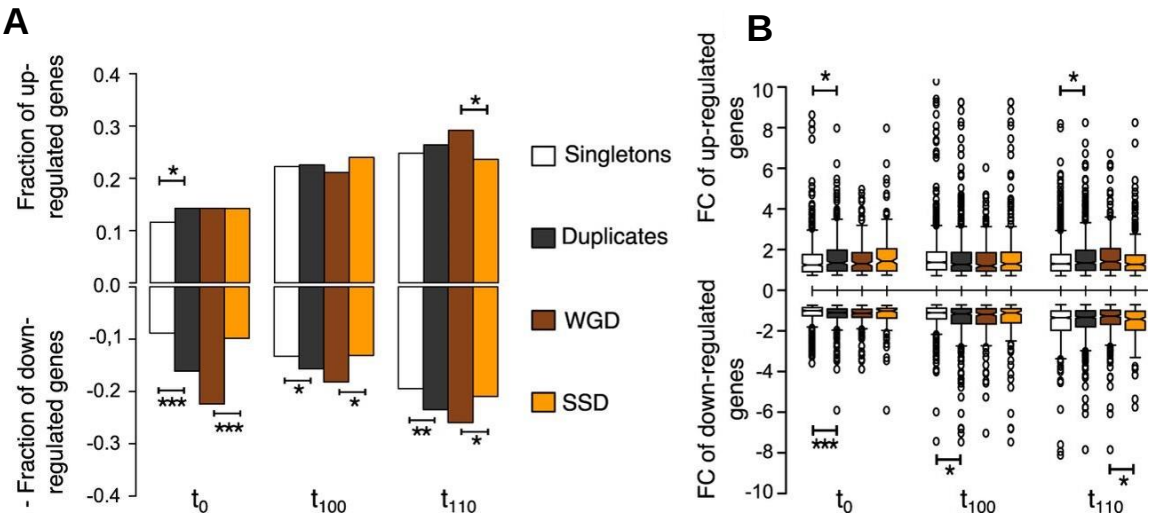


Figure ChIV- 3 - Genes responding transcriptionally to glucose replacement by ethanol, as carbon source, after adaptive evolution. **A)** Proportion of responding genes (showing transcriptional divergence [TD]) distributed in four categories (singletons, duplicates, WGDs, and SSDs). Upregulated genes are on the positive part of the y axis, whereas downregulated genes are on the negative part of the axis (after being made negative for representation purposes). Fisher's exact test has been used to test if the observed fractions of TD genes are significantly different from those expected. **B)** Expression difference (log fold change) in the two media, YPD and YPE. A Wilcoxon rank test has been used to test the difference in expression levels of sets of genes. Significant differences are indicated as *, **, and ***, when the probabilities are $P < 0.05$, $P < 0.005$, and $P < 10^{-3}$, respectively.

The response of duplicates to ethanol was even more apparent when looking at downregulated genes. Population t_0 showed a larger proportion of the duplicates (353) being downregulated than singletons (398) (Fisher's exact test: odds ratio $F = 1.82$, $P = 2.33 \times 10^{-14}$). Not only were there more ethanol-responding duplicates, but the response was also higher (duplicates: median FC = −1.103; singletons: median FC = −1.004; Wilcoxon rank test: $P = 3.75 \times 10^{-4}$). Most of this response was coming from WGDs (WGDs: 245; SSDs: 108; Fisher's exact test: odds ratio $F = 2.27$, $P = 8.31 \times 10^{-12}$). No difference in the expression fold change was found between the two types of duplicates (WGDs: median FC = −1.130; SSDs: median FC = −1.019; Wilcoxon rank test: P = 0.1804).

Population t_{100} showed similar results as population t_0: 343 of the downregulated genes were duplicates and 597 were singletons (Fisher's exact test: odds ratio $F = 1.18$, $P = 0.024$). The expression fold change was also higher in duplicates (median FC = −1.158) than in singletons (median FC = −1.108) (Wilcoxon rank test: P = 0.0174). More WGDs (199) were responding to ethanol stress than SSDs (144) (Fisher exact test: odds ratio $F = 1.39$, $P = 6.28 \times 10^{-3}$), but no difference in the expression fold change was observed (WGDs: median FC = −1.186; SSDs: median FC = −1.113; Wilcoxon rank test: P = 0.484).

Population t_{110} had no difference in duplicated genes (514) being downregulated compared to singletons (877) (Fisher exact test: odds ratio F = 1.20, P = 2.71×10^{-3}). However, duplicated genes had a higher expression fold change than singletons (duplicates: median FC = −1.332; singletons: median FC = −1.349; Wilcoxon rank test: $P = 0.338$). Interestingly, WGDs (284) were more abundant than SSDs (230) (Fisher's exact test: odds ratio $F = 1.23$, $P = 0.031$), but SSDs (median FC = −1.421) showed a higher expression fold change than WGDs (median FC = −1.271) (Wilcoxon rank test: $P = 0.0266$). The core gene of the downregulated genes consisted of more duplicates (4.3%) than singletons (2.84%) (Fisher's exact test: odds ratio $F = 2.36$, $P = 1.27 \times 10^{-4}$), with WGDs as the most affected duplicates (WGD 6.02%; SSD 2.5%: Fisher's exact test: odds ratio $F = 2.24$, $P = 1.63 \times 10^{-8}$).

e. Transcriptional divergence between duplicates gene copies is linked to the response and adaptation to ethanol of S.cerevisiae

If duplicates were linked to the response and adaptation of *S. cerevisiae* to ethanol, then we should expect the transcriptional divergence (TD) between gene copies of a duplicate to be correlated with its transcriptional patterns in ethanol. We identified those duplicated genes that exhibited a fold change expression difference between their gene copies of more than 25%. Of the 1,090 duplicated gene pairs (analysis contained both copies), 867 showed transcriptional divergence between gene copies in YPD. In the populations t_0, 274 of the 312 upregulated duplicates in ethanol belonged to duplicates with evidence of TD, a proportion greater than expected by chance (binomial test: $P = 1.846 \times 10^{-4}$) (Figure ChIV-4A). The mean expression fold change (in logarithmic scale) of the 274 duplicates was 1.59. We compared this mean to a null distribution of means built by sampling 274 duplicates from the population of the 867 duplicates with evidence of expression divergence (Figure ChIV-4B). The mean fold change of these duplicates was greater than expected by chance ($P = 3.0 \times 10^{-6}$). The fold change of the gene copy with the highest expression in ethanol divided by that of the least expressed gene copy is also correlated with the expression fold change of the duplicate (Pearson correlation: r = 0.66, $P < 2.2 \times 10^{-12}$). Importantly, among the most highly divergent and upregulated duplicates, we identified the plasma membrane H+-ATPase (*PMA2*), translational elongation factor (*HEF3*), plasma membrane permeases (*GIT1* and *SEO1*), and a gene involved in the metabolism under respiratory conditions (*RGI2*), among others (Table S1).

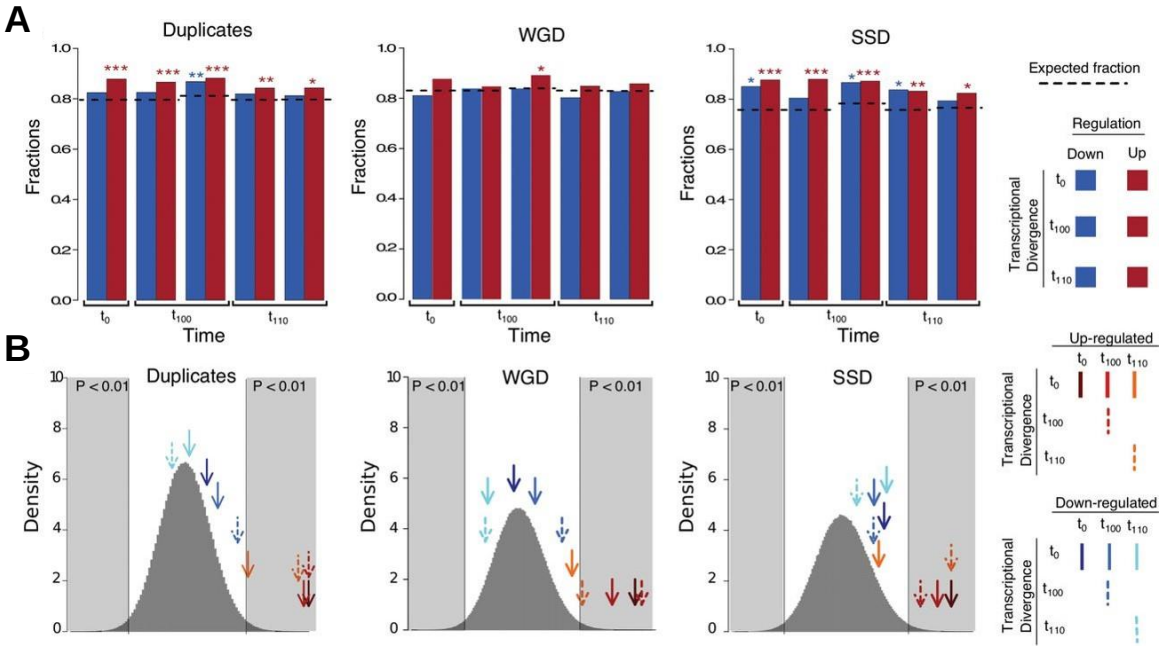


Figure ChIV- 4 - Proportion and mean transcriptional divergence of transcriptionally responding duplicates. **A)** The proportion of TD duplicates of transcriptionally responding genes, along with the expected proportions, marked with dashed lines. Significant differences in the observed number of TDs compared to the expected number are indicated as *, **, and ***, when the probabilities are $P < 0.05$, $P < 0.005$, and $P < 10^{-3}$, respectively, using a binomial test. **B)** Mean transcriptional divergence levels between duplicated genes within each of the categories (marked with arrows) are mapped onto a normal distribution build by random sampling, without replacement, of the same size from the corresponding gene pools. The gray blocks are indicating the significant part of the distribution ($P < 0.01$).

In the t_{100} population 426 duplicates of the 492 showed evidence of upregulation and belonged to the set of duplicates with evidence of expression divergence, a proportion greater than expected by chance (binomial test: $P = 6.81 \times 10^{-5}$). Like in the t_0 population, upregulated duplicates exhibited greater mean expression divergence between gene copies than expected by chance (mean = 1.45, $P = 1.05 \times 10^{-3}$) (Figure ChIV-4B). The phenotypic plasticity (expression fold change of duplicates in ethanol compared to YPD) was correlated with the expression divergence between the gene copies (Pearson correlation: r = 0.58, $P < 2.2 \times 10^{-16}$).

Population t_{110} also presented enrichment of upregulated duplicates for duplicates with evidence of expression divergence, with 485 out of the 576 upregulated duplicates exhibiting expression divergence (binomial test: $P = 5.22 \times 10^{-3}$). Interestingly, upregulated duplicated genes in t_{110} do not show higher transcriptional divergence than expected when calculated from the t_0 population (mean = 1.29, $P = 0.150$) but do show such when calculated from the t_{110} population (mean = 1.38, $P = 0.00120$), indicating that transcriptional divergence and upregulation are highly dependent on the current transcriptional background. Similar to the t_0 and t_{100} populations, the population of t_{110}

shows a correlation between phenotypic plasticity and expression divergence of duplicated genes (Pearson correlation: r = 0.5, $P < 2.2 \times 10^{-16}$).

In contrast to this pattern for the upregulated genes, we see no correlation between TD and downregulated genes, at any of the time points (Figure ChIV-4). The only correlation that is also present for the downregulated genes is the phenotypic plasticity in YPD and YPE (Pearson correlations: t0, r = 0.68, $P < 2.2 \times 10^{-16}$; t_{100}, r = 0.67, $P < 2.2 \times 10^{-16}$; t_{110}, r = 0.61, $P < 2.2 \times 10^{-16}$).

To understand why up- and downregulated genes show different patterns with respect to TD, we first checked the overall TD of up- and downregulated genes (Figure ChIV-5B). Downregulated duplicates had significantly lower TD than upregulated genes (P = 0.00244). The duplicated genes we are looking at are TD, which means we have one copy with a lower expression than the other (Figure ChIV-5A). Dividing up- and downregulated genes into low and high transcriptional diverged copy (TDC), we observe, as expected, higher TDC in downregulated genes (binomial test: $P = 0.0018$) and lower TDC in upregulated genes (binomial test: $P = 1.473 \times 10^{-9}$) (Figure ChIV-5C). All groups showed similar TD except for downregulated and low TDC, which had the lowest TD of all groups (Figure ChIV-5C).

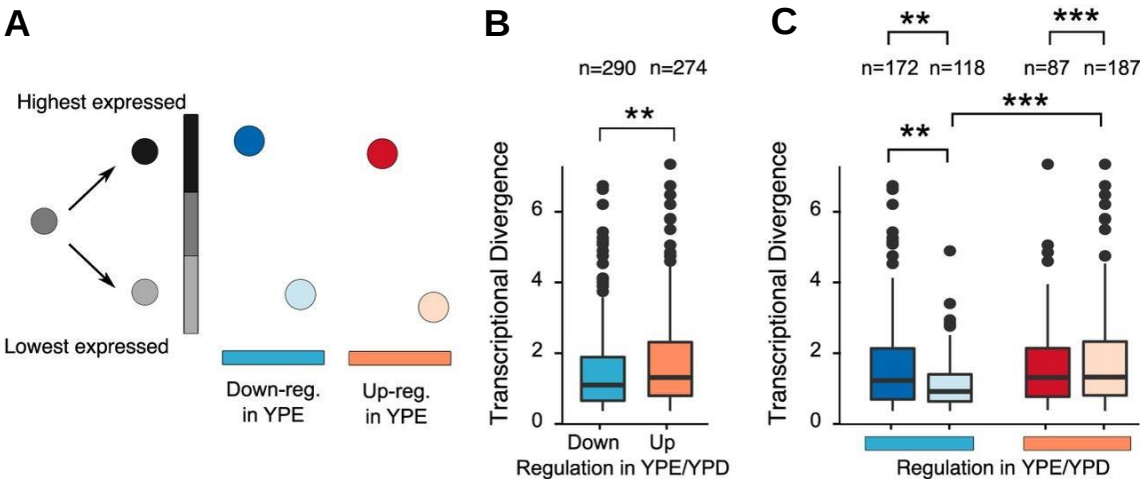


Figure ChIV- 5 - Transcriptional divergence of transcriptionally responding duplicates. **A)** Transcriptionally divergent (TD) duplicate characterization. TD duplicates are classified according to their sign of expression (high expression shown in dark colors and low expression shown in light colors), as the responding gene can be either the gene with high expression or the gene with low expression of the duplicated pair. Blue and red indicate the downregulated and upregulated pairs in YPE, respectively. **B)** Comparison of TDs of up- and downregulated genes at t_0. **C)** Comparison of TD pairs at t_0, differentiating each up-expressed gene into high- and low-expression gene categories as indicated in the scheme depicted in panel A. Significant differences are indicated as *, **, and ***, when the probabilities are P < 0.05, P < 0.005, and $P < 10^{-3}$, respectively. A Wilcoxon rank test was used for testing the significance between TDs of the different categories, whereas a binomial test was used for testing the number of TDs in the different categories.

Looking at GO enrichments of the four categories of Figure ChIV-5C, no overlap is observed between the upregulated and downregulated categories (Figure S5). Upregulated (highest transcribed copy) genes were enriched for “drug metabolic process,” “energy derivation by oxidation of organic compounds,” and “small molecule metabolic process,” and downregulated duplicates were enriched for “cytoplasmic transition” and different ribosome processes. To determine if the behavior of a gene has an influence on the response of the other duplicated copies, we further divided the groups into categories of the two copies having the same regulation profile, the two copies having different regulation profiles, or only one of the copies showing up- or downregulation in ethanol (Figure ChIV-6). Interestingly, we observed more duplicated copies which were both downregulated than expected ($P < 10^{-12}$), but this group also had the lowest TD. Inspecting the function of these genes, we see that a majority (48 out of the 60 genes) are ribosomal proteins. As would be expected, a lot of overlap of enriched GOs was observed between the different up- and downregulated categories. In the case of both duplicates being upregulated, we saw an enrichment for “carbohydrate metabolic process” and “oxidative phosphorylation.” Furthermore, these categories are the only ones that observe an enrichment of a pathway, namely, “superpathway of TCA cycle and glyoxylate cycle.” For the discordant duplicates, enrichment categories included “glucose 6-phosphate metabolic process,” “NADP metabolic process,” and “oxidoreduction coenzyme metabolic process” (Figure S5).

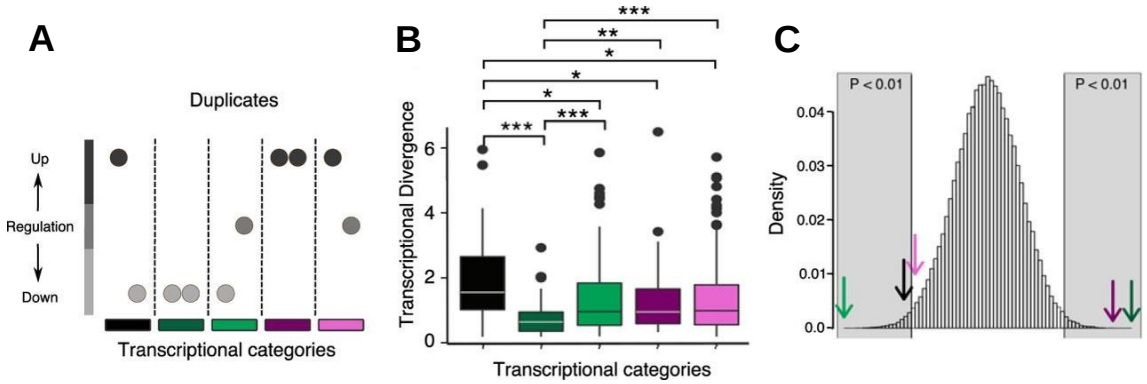


Figure ChIV- 6 - Characterization of TD transcriptionally responding duplicated pairs. **A)** TD duplicate categorization scheme (five categories): black, one gene copy is upregulated and the other is downregulated; dark green, both duplicated genes are downregulated; light green, one duplicate is downregulated and the other is unaltered; purple, both duplicates are upregulated; violet, one duplicated gene is upregulated and the other is unaltered. **B)** Comparison of TDs of the five categories described. A Wilcoxon rank test was used to determine significant differences indicated as *, **, and ***, when the probabilities are $P < 0.05$, $P < 0.005$, and $P < 10^{-3}$, respectively. **C)** Mean number of genes within each of the TD categories (marked with arrows with the coloring code described for panel A), mapped onto a normal distribution build by random sampling, without replacement, of the same size from the corresponding gene pools. Gray blocks over the normal distribution indicate the significant part of the distribution ($P < 0.01$).

Changing the carbon source from glucose to ethanol implies that the yeast goes from fermentation to aerobic respiration. Combining this with the fact that the tricarboxylic acid (TCA) cycle was enriched for upregulated duplicated pairs, we map the categorized proteins onto the two pathways (Figure ChIV-7). At least one of the proteins involved in each of the steps was upregulated, and in most cases the duplicated pairs were upregulated (i.e., *CIT1* and *CIT2*, *MDH1* and *MDH3*, and *ACS2* and *ACS1*). Interestingly, in cases where only one of a duplicated pair was within this pathway, we saw upregulation of just one of the proteins (i.e., *ACO1*, *LDP1*, or *KGD2*), namely, the one within the TCA cycle.

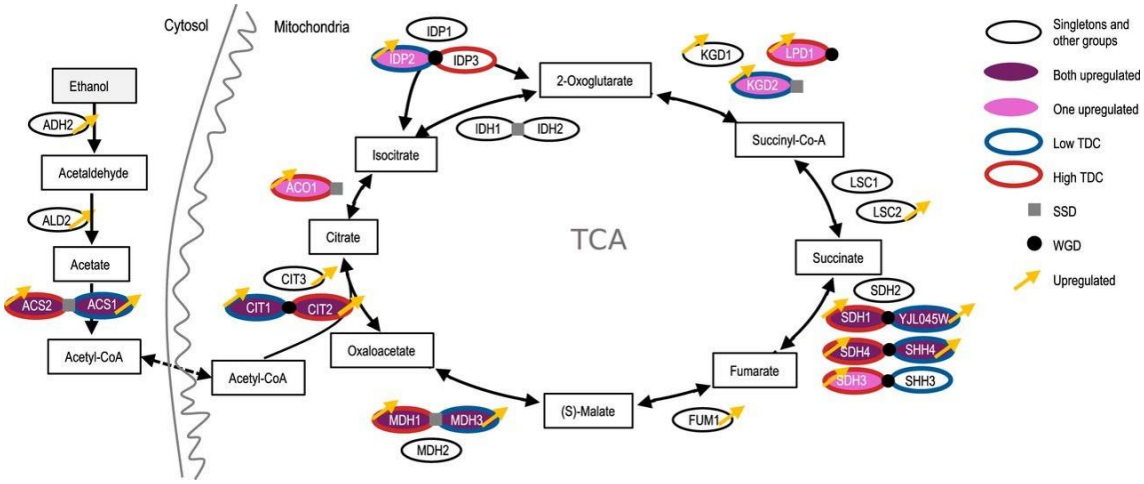


Figure ChIV- 7 - Pathway of nonfermentative C_2 metabolism in S. cerevisiae Y06240, from ethanol to TCA. Only ethanol degradation and the TCA pathway have been shown. Pathway information was taken from the KEGG pathway database. The proteins are colored after the duplicate categories set out in Figure ChIV-5 and ChIV-6, with indication of duplicate origin (SSD or WGD) and transcription-diverged copy level (TDC, in blue or red for low or high, respectively) or upregulation (yellow arrow).

f. Transcriptional divergence between duplicated genes play different roles in WGDs and SSDs

If TD of duplicated genes plays the same role in WGDs and SSDs, the same patterns should be observed. It has previously been noted that WGDs are more transcriptionally divergent than SSDs (Van Hoek and Hogeweg 2009). Using our data, we find 910 WGDs to be TD in YPD at t_0, compared to 824 SSDs, not a significant difference (Fisher's exact test: $P = 0.1389$). However, when looking at the magnitude of the TD between duplicates, we observed a significant difference between the two types, with WGDs showing a higher TD than SSDs (Wilcoxon rank sum test: P = 0.01281) (Figure ChIV-5B). It is worth noting that this difference of magnitude, between WGD and SSD, disappears if we look only at TD gene copies (Wilcoxon rank sum test: $P = 0.1575$).

To determine if the TD between the gene copies of WGDs and SSDs had different influences on the response to ethanol, we looked at how many TD duplicates were up- or downregulated in YPE compared to YPD at all three time points. For the duplicates *per se*, we had seen in the section above that upregulated genes contained more TD genes than expected (Figure ChIV-4A). When separating out the two types of duplicates, it was seen that SSDs contained more TD upregulated genes than expected at all three time points (binomial test: t_0, $P = 2.399 \times 10^{-4}$; t_{100}, $P = 7.239 \times 10^{-7}$; t_{110}, $P = 3.622 \times 10^{-3}$), as well as for downregulated genes at t_0 and t_{110} (binomial test: t_0, $P = 0.02408$; t_{110}, $P = 3.349 \times 10^{-3}$). This is the opposite pattern from what we observed in WGDs, where neither up- nor downregulated genes, at any of the time points, had more TD genes than expected (Figure ChIV-4A). To rule out that the limit set for a duplicated gene pair to be TD was not affecting our results, we redid the analysis for the TD limit going from equal expression to a 4-fold difference (Figure S6). In general, the pattern did not change much as the TD limit was changed, in particular at the lower TD limits.

As we saw a difference between WGDs and SSDs with respect to the quantity of TD genes that reacted to the ethanol stress, we wanted to see if there was a difference with respect to the magnitude of the TD and ethanol response. First, we compared the TDs of up- and downregulated genes. The general pattern observed was that the WGDs show a statistical difference between the magnitudes of TD of up- and downregulated genes (Wilcoxon rank sum test: t_0, $P = 6.5 \times 10^{-5}$; t_{100}, $P = 0.05$; t_{110}, $P = 5.3 \times 10^{-3}$) (Figure ChIV-8). In contrast, SSD showed no difference between the magnitudes of the TD for up- and downregulated genes. Second, we wanted to see if the observed mean TD of differentially expressed genes was higher or lower than expected by chance. We compared the observed mean TD with the normal distribution build by random sampling of the same size from the corresponding pools (WGDs and SSDs). In this case, we observed similar patterns for both WGDs and SSDs. The TD of upregulated genes exhibits the expected mean of TD. One interesting thing for both was that the mean TD at t_{110} was significant only when calculated from t_{110} but not from t_0, indicating that the transcriptional background has an influence on the response of the duplicated genes in ethanol stress. Furthermore, neither WGDs nor SSDs showed a significantly higher mean of TD of downregulated genes than expected, at all time points (Figure ChIV-4B).

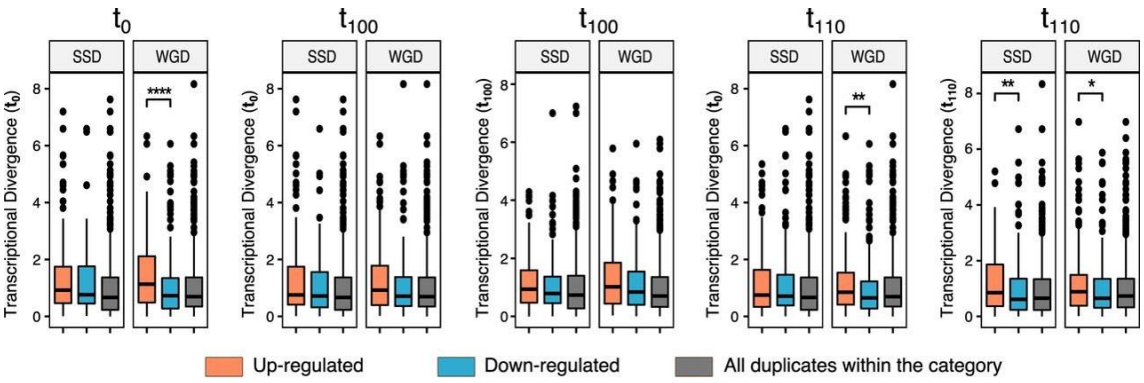


Figure ChIV- 8 - Distribution of transcriptional divergence per time point of the experimental evolution and per duplicate origin (SSDs and WGDs). A Wilcoxon rank test was performed to determine significant differences in the TD, indicated as *, **, ***, and ****, when the probabilities are $P < 0.05$, $P < 0.005$, and $P < 10^{-3}$, and $P < 10^{-4}$, respectively.

5. Discussion

a. Large transcriptional response to ethanol stress

One of the central mechanisms of unicellular organisms, and particularly nonmobile ones, is sensing and responding to changes in the environment. This is especially essential when the organism endures stress. Here, we show how changing the carbon source from glucose to ethanol leads to large transcriptional changes in the yeast *S. cerevisiae*. These changes are observed in a large percentage of the *S. cerevisiae* genes and encode a wide range of functions. Such genome-wide transcriptional changes have been shown before to take place in the response to numerous stresses, including glucose restriction (Guan *et al.* 2007), glycerol as the only carbon source (Oz *et al.* 2014), oxidative stress (Ferea *et al.* 1999; Mattenberger, Sabater-Muñoz, Toft, Sablok, *et al.* 2017; Anand *et al.* 2020), environmental estrogen (Chapal *et al.* 2019), acid tolerance (Mira *et al.* 2010; Bereketoglu *et al.* 2017), and thermal resilience (Geng *et al.* 2017), among others (Jarolim *et al.* 2013). Indeed, when the initial population was switched from a glucose-containing medium to one that contained only ethanol as carbon source, hundreds of genes altered their expression. With the medium not containing glucose, the yeast is performing aerobic respiration (Gancedo 1998; Voordeckers *et al.* 2015; Snoek *et al.* 2016), and the presence of ethanol induces oxidative stress of the cells (Costa *et al.* 1993; Alexandre *et al.* 2001; van Voorst *et al.* 2006; Gibson *et al.* 2008; Izawa and Inoue 2009; Talavera *et al.* 2018), which is reflected in our results, where we have an enrichment of “oxidation-reduction process” and “aerobic respiration” gene terms in the upregulated gene classes. Hence, two clear transcriptional patterns were observed during exposure to ethanol: (i) an upregulation of genes involved in stress response and (ii) a downregulation of ribosomal biogenesis or energy-dependent processes. Our observations agree with the suggested tradeoff cellular response (transcriptional rewiring of central metabolism) to environmental stress on yeast growth rate (López-Maury *et al.* 2008; Teixeira *et al.* 2011; Jarolim *et al.* 2013; Bereketoglu *et al.* 2017).

b. The genetic background influences the transcriptional response to ethanol.

The genetic background of a population influences its capability to grow and respond to stress (Zakrzewska *et al.* 2011; Ho and Zhang 2014; Bergström *et al.* 2014). In this study, evolving a population of *S. cerevisiae* for a large number of generations had huge influences on the transcriptional response to ethanol, as well as increasing the growth rate in the evolved medium (YPD). We observed transcriptional changes between the evolved and the nonevolved ancestral population. This change is likely due to a genetic change in the population, as the original ancestral population had low variability as it originated from a single clone and the evolved population gained genetic variability

through the evolution experiment, as described recently (Galardini *et al.* 2019). An interesting result from our study is that the evolved population improved its fitness in the evolved medium (YPD) but showed no change in the fitness when grown in ethanol (YPE). It has previously been shown that diversification can lead to exaptation in nonevolved environments (Oz *et al.* 2014; Mattenberger, Sabater-Muñoz, Toft, Sablok, *et al.* 2017; Nguyen Ba *et al.* 2019; Galardini *et al.* 2019). There are multiple possible reasons for us not observing this in our populations. First, increasing fitness to ethanol is hard. Many of the studies which have observed increased tolerance to ethanol see an increase of the ploidy of chromosome III (Fares 2015c; Kaboli *et al.* 2016); this occurs only in diploid and polyploid *S. cerevisiae*, and we are evolving a haploid population. Second, our ancestral population might have been at a local maximum in the ethanol fitness landscape of the population. Last, despite evolving our populations for approximately 660 generations, it might not have been long enough to acquire any exaptation to ethanol, deserving further study of the mutational landscape of these populations.

c. Going from acute to chronic exposure of ethanol rewires the transcriptome

The evolution of the yeast population in the presence of ethanol (the t_{110} populations) uncovered the regulatory changes that occur as the population reacts to acute and chronic exposure. It has previously been suggested that reducing the growth rate can lead to increased stress tolerance by redirecting the resources (López-Maury *et al.* 2008). The chronic-exposure population (evolved population in ethanol, YPE_t_{110}) showed an enrichment of upregulation of genes involved in oxidative stress, so overall these populations were upregulating more genes involved in stress response, an indication of higher allocation of resources to stress tolerance. This agrees with the fact that we observed a lower growth rate on ethanol for the evolved population in ethanol than for the population that evolved in YPD.

d. Duplicated genes play an important role in the response to ethanol

The first response, of an organism to stress, is through regulatory reprogramming; hence, the plasticity of the transcriptome will determine the potential for adapting to a new environment (Landry *et al.* 2006; Stern *et al.* 2007; Morard *et al.* 2019). However, this link is still not fully understood, enthralling scientists for the past 40 years and becoming of great importance recently (Lehner 2010), and that is predominately down to the difficulty of mapping phenotypes to genotype and assigning transcriptional changes to phenotypic variations. In this work, we clearly see a link between transcriptional variations, phenotypic response to ethanol, and gene copy number

(referring here to duplicates), as the transcriptionally altered genes are enriched for duplicated genes. The classical theory behind the evolution of duplicated genes states that one gene copy is able to evolve without or with reduced selection constraints as the other gene copy is performing the ancestral function (Lynch and Conery 2000; Hakes *et al.* 2007; Hallin and Landry 2019). Diversification of the gene copies happens not only at the functional level but also at the expression level (Papp *et al.* 2002; Conant and Wolfe 2008; Anand *et al.* 2020). The diversification at the expression level could open up the possibility to diverge functionally, as the rate of evolution is highly linked to its expression, although recently it has been shown that low-expression transcription factors adapt through cooperation rather than functional divergence (Zhang 2003; Anand *et al.* 2020). It has been suggested that the whole-genome duplication even facilitated the *Saccharomyces* yeast to evolve the ability to ferment sugars under anaerobic conditions, which is not the case for other yeasts (reviewed in Gout, Kahn, and Duret (2010); Hagman *et al.* (2013); and Escalera-Fanjul *et al.* (2019)). Here, we are forcing the yeast to use ethanol as the sole carbon source, meaning it has to perform respiration instead for fermentation, not requiring the Crabtree effect-implicated genes. In correlation with this, we observe that most of the changes of the duplicated genes were downregulation of WGD, consistent with the hypothesis that WGDs were providing the raw material for conservation of dosage-sensitive genes involved in both rewiring of rapid growth elements (ribosomal protein genes) and divergent regulation and specialization of gluconeogenesis-ethanol consumption phase versus glycolysis-ethanol production (Schüller 2003; Hagman *et al.* 2013). Taken as a whole, the rewiring of the transcriptome, and in particular the duplicated genes, indicates that the yeast cell goes into energy preservation when the carbon source is switched to ethanol.

e. The transcriptional background and the response to ethanol

In plants, it has been observed that duplicated genes diverge transcriptionally soon after duplication (Blanc and Wolfe 2004b; Ha *et al.* 2007, 2009; Escalera-Fanjul *et al.* 2019). Furthermore, a correlation between the divergence from the ancestral expression level and stress response has also been observed in plants (Wang *et al.* 2012). These all indicate that duplication and expression divergence are linked to adaptation and stress response (Blanc and Wolfe 2004b; Zou *et al.* 2009). In yeast, duplicated genes have also been shown to be transcriptionally diverged, particularly in WGD (Oz *et al.* 2014). In a wider study looking at transcriptional changes of duplicated genes under different stress conditions, it was observed that one of the gene copies was more transcriptionally plastic than the other (Strauss *et al.* 2019). These all indicate that transcriptional divergence plays an important role in maintaining duplicated genes in the

genome and expanding the phenotypic plasticity of the organism. Here, we observe that transcriptional divergence between gene copies is correlated with response to ethanol. In particular, responding duplicates have higher transcriptional divergence than expected. However, WGD and SSD have different parameters by which the TD influences the response to ethanol. The magnitude of the TD is important for WGD, where in contrast the number of genes with TD is important for SSD. One interesting thing that we observed in this study is the change of TD of the duplicated genes throughout our experiment and that this change was correlated with the response to ethanol, indicating that the transcriptional background is important for the actual stress response and this can change relatively quickly.

f. Concluding remarks

The recent advances in next-generation sequencing technologies coupled with the decrease of their prices have increased general interest in determining the role of polyploidy and transcriptional plasticity in ecological shifts or lifestyles. The switch to using ethanol as sole carbon source implied a yeast cell reprogramming to energy preservation with low growth rate but with similar biomass production due to transcriptional reprogramming of duplicates, especially those of the TCA cycle. In this work, we have unveiled that TD between duplicates and the transcriptional background affect duplicates' response to ethanol, with the magnitude of the TD being especially important for WGDs.

6. Data availability

Raw reads are available from the Sequence Read Archive (SRA) with accession numbers PRJNA321113 (t_0 in YPD and YPE), PRJNA610243 ($a1t_{100}$ in YPD), PRJNA610541 ($a1t_{100}$ in YPE), PRJNA610474 ($Da1t_{110}$ in YPD), and PRJNA610515 ($Ea1t_{110}$ in YPE).

7. Supplementary data

Supplementary data are available at *mSystems* online. Supplementary Figures S1, S2, S3, S4, S5, and S6 are also available in the appendix of this manuscript. **Figure S1:** Maximum growth rate as the phenotypic characterization of populations through experimental evolution. **Figure S2:** Evolution of the carrying capacity of populations under ethanol stress. **Figure S3:** Distribution of transcriptomic profiles under ethanol

stress. **Figure S4:** Biological processes enriched for downregulated genes, due to the use of 3% ethanol as the sole carbon source. **Figure S5:** Biological processes enriched for transcriptionally divergent duplicated genes. **Figure S6:** Implication of transcriptional divergence limit on transcriptional response of duplicates. **Table S1:** Divergent levels and transcriptional difference between YPE and YPD of duplicated genes.

CHAPTER V – The role of gene duplication in adaptation: experimental evolution of *Saccharomyces cerevisiae* under acidic stress.

Unpublished results.

1. *Abstract*

The baker yeast *Saccharomyces cerevisiae* is a widely used microorganism in biotechnology gaining a high relevance in the industry and having a large economic impact. Also, lactic acid is a very valuable chemical product in society. Thus, *S. cerevisiae* is one of the most used cell factories to produce lactic acid since lactate –the conjugate base of lactic acid- is a waste product of glycolysis. However, the accumulation of lactate/lactic acid in the growth culture challenges the survival of the yeast. In addition, lactate can also be converted into pyruvate under fasting conditions, yet compromising the yields of obtained lactic acid in industry and deriving important economic costs. Therefore, understanding the genetic and cellular bases of *S. cerevisiae* to deal with this stressful situation is key. Recently, the role of duplicated genes has been widely demonstrated for adaptation. Here we show that yeast growing capability under lactic acid stress conditions is improved in experimental evolution after a few passages. Indeed, we analyzed the transcriptional response of populations exposed to acute and chronic exposure to lactate as a unique carbon source and observed a genome-wide transcriptional response. Finally, we investigated the role of duplicated genes in the adaptation of *S. cerevisiae* to short and long-term exposure to lactic acid and observed that a huge cellular process reprogramming is carried out by the transcriptional response of duplicated genes.

2. *Introduction*

In nature, organisms have to deal with a huge range of non-optimal and stressful situations that can improve fitness and compromise their survival. Over time, the organisms have gained in complexity and the evolution has selected for developing several mechanisms to stress response that has allowed not only to survive in a constantly changing environment but also to spread into a vast number of different ecosystems. Especially in the case of *Saccharomyces cerevisiae*, the robust stress biology of the yeast has made it possible to conquer a huge variety of ecological niches (Botha 2011; Cray, Bell, *et al.* 2013; Lievens *et al.* 2015; Jeffares 2018; Chappell and Fukami 2018). Hence, a good understanding of the genetic clues that give rise to such adaptive plasticity is indispensable to understand its biology. In recent years, a big effort has been done to elucidate the pathways that underlie the challenge and the adaptation of the yeast to stressful environments (Dhar *et al.* 2013; Markiewicz-Potoczny and Lydall 2016; Lopandic 2018). Different approaches have been done in yeast to shed light on the adaptation to different challenges, such as tolerance to weak acids that drop down the pH in the growing medium (Mira *et al.* 2010; Narayanan *et al.* 2016), like acetic acid (González-Ramos *et al.* 2016; Geng *et al.* 2017) or lactic acid (Dato *et al.* 2014; Berterame *et al.* 2016; Fletcher *et al.* 2017; Ortiz-Merino *et al.* 2017), but also for other stressors like glycerol (Babazadeh *et al.* 2017; Mattenberger, Sabater-Muñoz, Hallsworth, *et al.* 2017) or ethanol (Sabater-Muñoz *et al.* 2020), or for starving glucose limiting conditions (Hosseini and Wagner 2016; Tamari *et al.* 2016). Additionally, the budding yeast *Saccharomyces cerevisiae* is one of the most important and widely used microorganisms in biotechnology, becoming a keystone in many industrial applications and economy. Moreover, lactic acid is a very valuable chemical product in society that is used from being one of the most common food preservatives in the world to be a precursor for biodegradable plastics (Dato *et al.* 2014). Thus, *S. cerevisiae* is one of the most used cell factories to produce lactic acid; notwithstanding this biochemical process challenges the survival of the yeast, so it is important to extend our knowledge on molecular and cellular biology, as well as, on the genetic basis that underlies the yeast response to the lactic acid in order to improve the biotechnological capability of *S.cerevisiae*.

Lactic acid, more concrete its conjugate base the lactate, is a waste product of glycolysis, however, it can be converted into pyruvate in a reverse reaction under fasting conditions. That gives the possibility to use lactate as a carbon source to obtain energy in glucose deprivation and non-fermentable aerobic conditions (Dejean *et al.* 2000; Lodi *et al.* 2002). Nevertheless, the use of lactate as a carbon source is challenging for the

yeast for two principal reasons: First, the lactate is unable to enter the cell without an active transporter that consumes energy (Andrade and Casal 2001; Casal *et al.* 2008) and second, the pKa of the lactic acid (pKa =3.86) is below the optimal pH (pH = 5.0-5.5) for the yeast to grow, meaning that the strength of the lactic acid drives the media to a low pH, a stressful situation that the yeast has to face. Therefore, understanding the genetic basis of the yeast's capability to overcome this stressful challenge is key, and important advances in RNA sequencing technologies has allowed to go further in the transcriptional response to environmental changes from a holistic point of view (Taymaz-Nikerel *et al.* 2016). This high throughput approaches evidenced the general sensitivity of a number of pathways to environmental stresses, in particular, *Saccharomyces cerevisiae* shows a complete transcriptomic response to stress by starvation (Ferea et al. 1999).

During the last years, several pieces of evidence have revealed the importance of duplicated genes for adaptation, not even in yeast (Keane *et al.* 2014; Mattenberger, Sabater-Muñoz, Toft, and Fares 2017; Mattenberger, Sabater-Muñoz, Hallsworth, *et al.* 2017; Mattenberger, Sabater-Muñoz, Toft, Sablok, *et al.* 2017; Sabater-Muñoz *et al.* 2020), but also in plants (Wendel 2000; Otto and Whitton 2000; Lespinet *et al.* 2002; Carretero-Paulet *et al.* 2013) and animals (Otto and Whitton 2000; Hoegg *et al.* 2004). Gene duplication is known as a major force in evolution and biological innovation (Ohno, 1970; Zhang, 2003; Wagner, 2011), because one of the copies can accumulate mutations while the other remains unchanged, thus avoiding the purge by natural selection. This grants the redundant gene copies to explore the available genetic space, hence increasing the capacity of the organism to accumulate cryptic genetic variation (i.e. mutations that are unseen by natural selection). Consequently, cryptic genetic variability improves the accessibility to new phenotypes, and by this the capacity of the population to evolve and to adapt to new environments. In addition, cryptic genetic variability also contributes phenotypic plasticity (i.e. the capacity to change the phenotype without changing the genotype) making the organisms more robust to fluctuating and stressful environmental changes. Under different stress conditions, duplicated genes have been shown to be crucial (Blanc and Wolfe 2004b; Li *et al.* 2005; Conant and Wolfe 2006; Thompson *et al.* 2013; K Zhang *et al.* 2017; Steenwyk and Rokas 2017). The divergence of duplicated genes that give origin to new features can occur at different levels, but important changes at the expression level have been reported (Tirosh and Barkai 2007; Tirosh *et al.* 2009).

In this study, we did an experimental evolution with the yeast *S. cerevisiae* and studied the transcriptomic response to lactate when this is present as a unique non-

fermentable carbon source. Our results show an improvement of the yeast population's growing capability after a very brief time challenged with lactic acid. We also examined the stress response of the yeast to acute and chronic exposure to lactic acid/lactate. On one hand, the accurate exposure to lactate as a unique carbon source drives the yeast to a genome-wide transcriptional response and cellular reprogramming. On the other, the long term evolution under chronic lactate exposure reveals that the yeast population can adapt to the new environment undergoing a huge cellular process re-programing carried out principally by the transcriptional plasticity of duplicated genes.

3. Materials and Methods

a. Biological samples, yeast culture and experimental evolution

In this study, the yeast *Saccharomyces cerevisiae* strain Y06240, a haploid *msh2* deletion strain (BY4741; *Mat**a**; his3D1; leud2D0; met15D0; ura3D0; msh2::kanMX4*), was used as described previously (Fares *et al.* 2013; Mattenberger, Sabater-Muñoz, Hallsworth, *et al.* 2017; Sabater-Muñoz *et al.* 2020). Briefly, a homogeneous population was founded (t_0) by growing a single colony into a liquid culture of rich complete media (YPD: 2% (w/v) bacto peptone, 1% (w/v) yeast extract, 2% (w/v) dextrose) and evolved through daily bottlenecks (1%) for 100 days (t_{100}; roughly 660 generations) to increase genetic variability. The evolution was performed in 5mL of rich YPD media in 50mL Corning tubes, at 28°C and 220 rpm. During the experimental evolution, a fossil record was constructed by storing a glycerol stock (25%; v/v) every 10 passages and freezing at -80°C. From passage 100 (t_{100}) the population was divided into two, with three biological replicates. One half was grown in YPD medium as control, whereas the second half was grown in medium containing 3% of lactic acid as the sole carbon source (YPL: 3% (v/v) lactic acid, 2% (w/v) bacto peptone, 1% (w/v) yeast extract, pH 5.5; supplemented with 100 µg/mL kanamycin). The two populations (YPD and YPL) were evolved for another 10 passages for generating the adapted populations. We used the same experimental setup previously described, with the difference that in this we used a daily bottleneck of 10% of the population.

b. Growth rates determination

Growth parameters for t_0, t_{100}, and t_{110} were obtained using Bioscreen C plate-reader system (Oy Growth Curves Ab Ldt., Helsinki, Finland) as described in Mattenberger et al. (2017). Briefly, each time point was pre-cultured, from the corresponding fossil record, overnight at 28°C in the corresponding evolving media, and then used to inoculate 200µL of fresh media to an initial OD_{595} of 0.06 to 0.07 and distributed in 100-well Honeycomb plates. The experiment was run for 2 to 3 days at 28°C with continuous shaking and taking OD_{595} measurements every 15 minutes. The collected data was analyzed with GrowthRates software version 3.0 (Baty and Delignette-Muller 2004; Hall *et al.* 2014).

c. RNA extraction and transcriptomic analysis

The transcriptomic profiling was performed in the t_0, $t_{100,}$ and t_{110} time points of lines evolved in YPD or YPL, with three technical replicates, being challenged to the stress or control media as appropriate. Briefly, time points from the fossil record were grown overnight in their evolving media, and then switched to YPL or YPD, or kept in their

evolving media until OD_{595} ~ 0.6, and cells were pelleted afterwards by centrifugation. Total RNA was extracted from with RNeasy kit (Qiagen) following manufacturer instructions.

Ribosomal RNA was removed by using Ribo-Zero Gold rRNA removal yeast (Illumina) depletion kit. Stranded RNA libraries were constructed using TruSeq stranded mRNA (Illumina) from oligo-dT captured mRNAs from depleted samples. Libraries were run in NextSeq 500 (Illumina) at 75nt single read by using High Output 75 cycles kit v2.0 (Illumina). RNA libraries were sequenced at the Genomic core facility at Servicio Central de Soporte a la Investigacion Experimental (SCSIE) from University of Valencia, Spain.

Raw reads were analyzed using FastQC report, cleaned with CutAdapt, and trimmed for quality and length (Pred score inferior to 20 and size less than 40 nt were discarded). Reads were aligned with Bowtie2 (up to two mismatches accepted) to the reference S288c strain genome (only CDS). Statistical assessment of differential gene expression was done with edgeR (Robinson *et al.* 2010), setting false discovery rate (FDR) at < 0.005, and applying BY correction for p-value (0.005).

d. Identification of duplicated genes, functional classification and visualization

Paralogous pairs of duplicated genes were divided into two groups according to their origin mechanism (WGDs or SSDs). WGDs (555 pairs) were extracted from the reconciled YGOB list (Yeast Gene Order Browser, last accessed March 2018; http://wolfe.gen.tcd.ie//ygob; Byrne and Wolfe, 2005). While SSDs (560 pairs) were identified after best reciprocal hits from all-against-all BLAST searches using BLASTP with an E-value cutoff of $1E^{-5}$ and a 50-bit score (Altschul *et al.* 1997), selecting only those that exhibit a distribution of synonymous substitutions similar to WGDs (Fares *et al.* 2013; Keane *et al.* 2014).

Differentially expressed genes were further classified according to their gene ontology (GO) term as implemented in the GO Term Finder of *Saccharomyces* Genome Database (http://www.yeastgenome.org/cgi-bin/GO/goTermFinder.pl), followed by an enrichment analysis with a P-value cutoff of <0.01 and semantic similarity summarization (simRel; Schlicker et al., 2006) as implemented in REVIGO (Tomislav *et al.* 2011).

e. Software

Unless otherwise indicated, calculations and statistics were performed using MS Excel and R 3.5.0 (R Core Team (2018))

3. Results

a. Lactic acid/lactate affects S.cerevisiae growth, while experimental adaptation recovers growth rate level

To test the ability of the yeast to overcome a challenge with a non-fermentable carbon source as lactate, the yeast *Saccharomyces cerevisiae* strain Y06240 – a haploid *msh2* deletion strain (BY4741; *Mat**a**; his3D1; leud2D0; met15D0; ura3D0; msh2::kanMX4*) – was evolved first through a daily 1% bottlenecking in YPD control media (containing glucose as carbon source) for 100 passages, followed by another 10 passages with only 10% population bottleneck in YPL stressing media (containing 3% lactic acid/lactate as non-fermentable carbon source), keeping also control populations evolving in YPD (Figure ChV-1A). Considering the high mutation rate of 10^{8} mutations/nucleotide for the *Δmsh2* yeast strain used in this study, we were able to fix on average one mutation per cell every passage.

The maximum growth rate of the population was used as a measurement of its fitness. In both media, the starting yeast population showed a sigmoidal curve with a maximum growth rate (μ_{max} ± s.d.m.) of 0.335 ± 0.01h^{-1} in YPD and of 0.161 ± 0.005h^{-1} in YPL, being significantly lower in the second (two-tailed t-test: p-value = 2.87 x 10^{-7}) (Figure ChV-1B)

As expected, most of the genetic variability generated during the evolution was neutral and does not affect the ability of the yeast to grow in YPD. The maximum growth rates of the evolved populations (t_{100} and t_{110}) (μmax ± s.d.m. = 0.349 ± 0.018h^{-1} and 0.334 ± 0.06 h^{-1} respectively) were not significantly different from the initial population (two-tailed t-test: p-value = 0.126 and 0.057 respectively). Nevertheless, when the evolved population t_{100} was challenged with YPL, the population exhibited a growth rate of 0.259 ± 0.053 h^{-1}, being significantly higher than the initial population (two-tailed t-test: p-value = 0.01) (Figure ChV-1B)

Interestingly, phenotypic adaptation to the new environment is produced very fast, as after 10 passages (~ 33 generations) the maximum growth rate of the YPL-adapted population (μmax ± s.d.m. = 0.316 ± 0.059 h^{-1}) was higher than the t_{100} population (two-tailed t-test: p-value = 3.7 x 10^{-4}) but was also higher than the control population challenged with YPL (μmax ± s.d.m. = 0.237 ± 0.019 h^{-1}, two-tailed t-test: p-value = 4.3 x 10^{-4}).

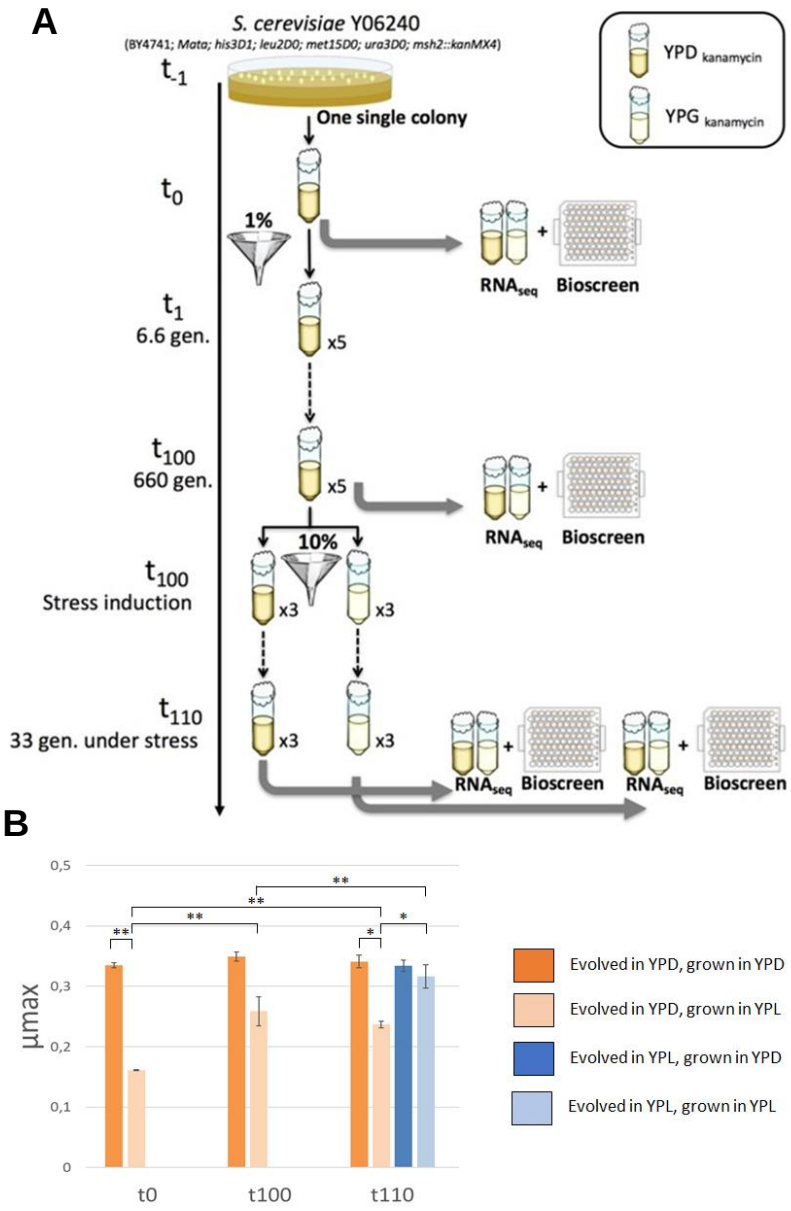


Figure ChV- 1 Experimental procedure to test the transcriptomic response and the growth capability of S.cerevisiae populations faced with lactic acid-induced stress. **A)** A single colony of *S.cerevisae* Y06240 was used to inoculate a founder population t_0 in YPD rich media. This population (t_0) was evolved for 100 passages (~660 generations) by transferring 1% of the population to fresh YPD rich media every 24 hours. This population (t_{100}) was then subjected to long term lactic acid stress (YPL) by evolving for other 10 passages (~33 generations) by transferring 10% of the population to fresh media. Additionally, a control population was also evolved in YPD. Population t_0, t_{100}, t_{110} were tested for their ability to grow in YLP compared with YPD using a Bioscreen C plate reader. Finally, populations t_0, t_{100}, t_{110} were subjected to lactic acid induced stress and the transcriptomic changes were analyzed by RNAseq (figure adapted from Mattenberger, Sabater-Muñoz, Hallsworth, & Fares, 2017b). **B)** The maximum growth rate of a genetically homogenous population of *S.cerevisiae* (t_0) is significantly higher in YPD (yellow bar) than in YPL (light yellow bar). After 100 passages of neutral evolution in YPD (t_{100}) the genetically heterogenic population has higher phenotypic plasticity and therefore it exhibits a significantly higher maximum growth rate in YPL than the ancestral population, meanwhile, there is no difference in its ability to grow in YPD. At the beginning of the adaptation process (t_{110}) the population evolving in YPL shows a significantly higher growth rate than the control evolving in YPD and growing in YPL (light blue bar), which maintains a similar growth rate than the plastic population t_{100} growing in YPL. Significant differences are indicated as * and ** when p-value for two-tailed t-test with inequivalent variance are < 0.05 and <0.01, respectively.

b. Transcriptional response of the yeast S.cerevisiae to lactic acid/lactate as the unique carbon source

The genetic transcriptional response of the yeast cells to a non-fermentable carbon source, lactic acid/lactate, was analyzed by comparing the transcriptomic profiles by RNA sequencing. Transcripts were mapped to a total of 6692 genes. At t_0, the challenge to YPL induced a de-regulation of 1283 genes (FDR< 0.005) comprising 19.17% of the total analyzed genes. Of these 1283 genes, roughly one half was up-regulated and the other half was down-regulated (628 with $\log_2$FC>1 and 655 with $\log_2$FC<-1; Exact binomial test: p-value =0.47). The transcriptomic profile of the evolved population t_{100} challenged with YPL, showing 1015 de-regulated genes out of 6692 (15.17%), was 4% lower than at t_0. Despite this decrease, the population t_{100} showed a higher number of genes that were up-regulated (N=615, 60.59%) than down-regulated (N=400, 39.41%, Exact binomial test: p-value = 1.59×10^{-11}), being only 431 (42.46% out of the 1015 altered genes in this population) also altered at t_0 (Figure ChV-2A). Interestingly the transcriptomic profiles between these populations are very different not only by the genes that are altered but also by the type of alteration they undergo. Out of the genes up-regulated in t_{100} challenged in YPL (N=615), 39.5% (N=243) were also up-regulated at t_0 (Figure ChV-2B). This is significantly higher than the number of down-regulated genes at both populations (34.5%, N=138, Fisher's exact test: Odds Ratio = 1.81, p-value = 4.23×10^{-7}). Indeed, to point out the high reliability between the transcriptomic data obtained here, we inquire how many genes that were up-regulated at t_{100} were down-regulated at t_0, observing only 27. Observing only 27 genes indicates that, in general, de-regulation sense is kept under our evolutionary experiment, meaning that the up-regulated genes at t_{100} are also up-regulated at t_0 and that the down-regulated genes at t_{100} are also down-regulated at t_0.

To understand what transcriptomic changes drive the adaptation process, we analyzed the transcriptome of the population t_{110} adapted to lactic acid/lactate as a carbon source. In this early adaptation, 2075 genes (31.01% of the total genes analyzed) altered their expression profile when compared to the growth in the YPD control media, showing a huge cellular reprogramming. Contrary to what is observed in the populations t_0 and t_{100}, this adapted population showed 1157 (55.76%) down-regulated genes, being this number significantly higher than the number of up-regulated genes (N=918, 44.24% Exact binomial test: p-value = 1.69×10^{-7}).

Consequently, by comparing the lists of altered genes we identified a core set of 317 transcriptionally altered genes responding to lactic acid/lactate stress (Figure ChV-2A).

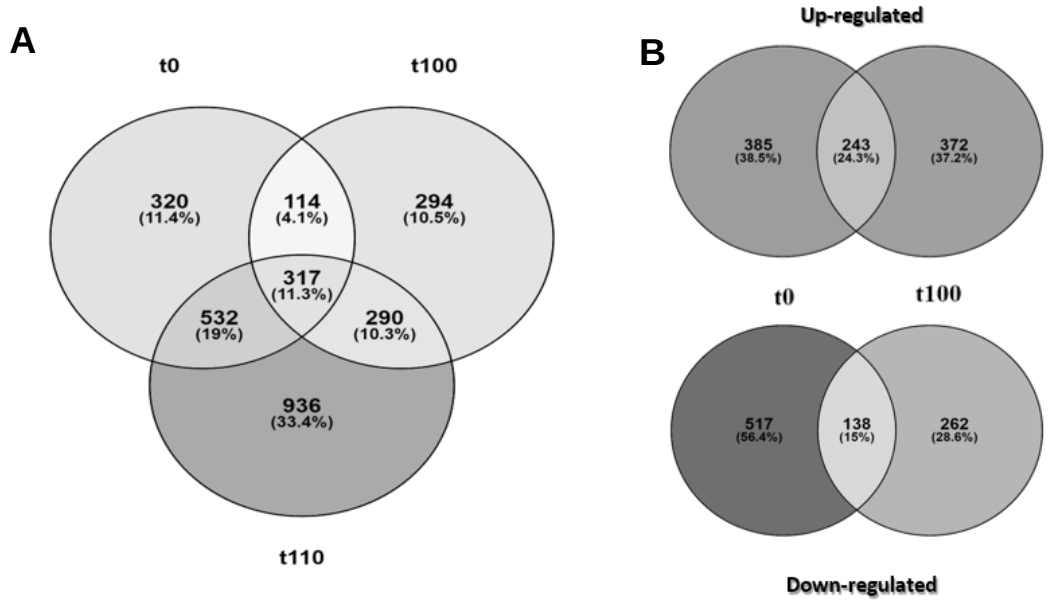


Figure ChV- 2 Changing the carbon source to lactic acid, a non-fermentable carbon source, induces a change in the transcriptome of the yeast population. From all the transcriptionally altered genes we can point out 3 sets of genes: those that are common to all three populations (t_0, t_{100}, and t_{110}), those that are common to two different populations, and those that are unique to one of the populations. **A)** Venn diagram identifying the number of transcriptionally altered genes identifying the level of sharing between different populations (done with http://bioinfogp.cnb.csic.es/tools/venny/index.html). **B)** Venn diagram identifying the number of up- and down-regulated genes shared between populations t_0 and t_{100} (done with http://bioinfogp.cnb.csic.es/tools/venny/index.html).

c. Many cellular processes are altered when the yeast S.cerevisiae is challenged with lactic acid/lactate

To shed light on how the yeast changed its cellular response when the environment drastically changes the carbon source, we analyzed what cellular processes were altered in the different populations according to the Gene Ontology terms (GO) using *Saccharomyces* Genome Database (SGD: https://www.yeastgenome.org/). At t_0 population fundamental cell processes were altered, including “cytoplasmic protein translation”, “mitochondrion organization”, “ribosome assembly”, “carbohydrate derivate metabolism”, “RNA and proton transport”. Also, “oxidation-reduction processes”, “cellular respiration”, “generation of precursor metabolites and energy” mainly from “cellular amino acid and organic acid metabolism” are overrepresented in the list of altered genes (Figure ChV-3A, Supplementary Table S1).

For the population t_{100}, we observed that important cellular processes such as “cellular respiration” and “generation of precursor metabolites and energy” are altered. Similarly, mainly mitochondrial processes that include “mitochondrial genome replication”, “mitochondrial morphogenesis”, “mitochondrial translation” and

"mitochondrial transport" are altered. Furthermore, many metabolic processes are altered, like the Krebs cycle and the pentose phosphate cycle. Finally, the transmembrane transport of carbohydrates was also de-regulated (Figure ChV-3B, Supplementary Table S2).

We observed that in general terms the yeast *Saccharomyces cerevisiae,* when challenged with lactic acid, up-regulates important cellular processes like the "cycle of carboxylic acids", "cellular respiration" and "oxidation-reduction" processes. Moreover, "ion transport", especially "hydrogen transmembrane transport", "mitochondrial organization" and "other small molecule metabolism" were up-regulated (Supplementary Table S8). While, "nucleoside diphosphate metabolism", "nucleotide phosphorylation", and glycolysis were down-regulated (Supplementary Table S9).

In the population $t_{110,}$ many cellular processes are de-regulated including "cellular component biogenesis", "ribosome assembly", "mitochondrion organization", "cytoplasmic and mitochondrial translation", "amino acid activation and RNA processing". Also, important metabolic processes are altered including "generation of precursor metabolites and energy", "oxidation-reduction processes" and "metabolism of organic acids" (Figure ChV-3C, Supplementary Table S3).

Interestingly, the cellular processes that are altered from the core genes set (those genes that are altered in all three populations, hence representing the main cellular response of the yeast challenged with acid lactic as unique non-fermentable carbon source) are mainly "ion and cation transport", "metabolism of organic acids" "cellular respiration", "oxidation-reduction processes" and "mitochondrion organization" (Figure ChV-3D, Supplementary Table S4).

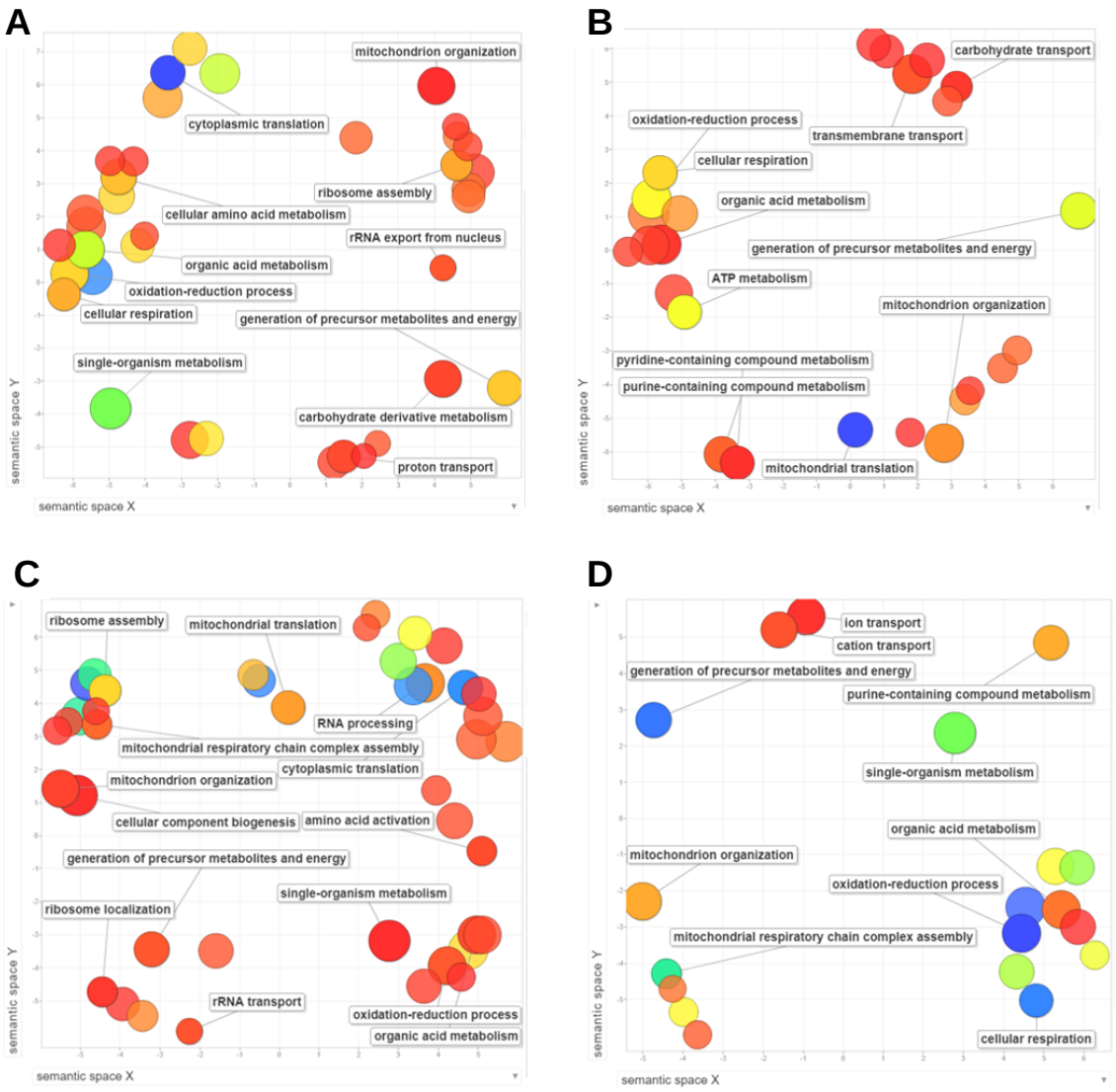


Figure ChIV- 3 Semantic clustering of cellular processes enriched for genes that were transcriptionally altered in the populations t_0 (A), t_{100} (B), and t_{110} (C) growing in YPL. The color of the bubbles represents the proportion of genes in a particular cellular process (log P-value), while the size indicates the frequency of the GO term in the organism. **D)** Semantic clustering of the cellular processes enriched for genes that were transcriptionally altered in all three populations, the core gene set in the cellular response to lactic acid.

d. The implication of duplicated genes in the transcriptional response to lactic acid/lactate

Since the budding yeast, *S. cerevisiae* has roughly a 32.8% of its genome duplicated (2202 duplicated genes, 1101 pairs), and gene duplication has often been proposed as one of the major sources for new genes and the origin of novel adaptations, we analyzed if duplicated genes are involved in the transcriptional and cellular response to lactic acid/lactate.

For the population t_0 we observed 578 altered duplicated genes, corresponding to 45.05% of all the transcriptionally altered genes. This is roughly 12% higher than the expected number of duplicated genes that alter their expression (Exact binomial test: p-value < 2.2×10^{-16}), indicating that duplicated genes play an important role in non-fermentable carbon source utilization, as the one tested here. Indeed, analyzing the fold-change of altered genes, we observed that duplicated genes exhibited a higher fold-change (median = 1.123) than singletons (median = 1.034, Wilcoxon test: p-value = 9.65×10^{-4}). It is interesting that despite the similar numbers of up- and down-regulated genes for population t_0, we observed a significantly higher number of duplicated genes that were down-regulated (N=349, 53.28%, Fisher's exact test: Odds ratio = 2.331, p-value < 2.2×10^{-16}) while we didn't observe a difference among the up-regulated genes (N=229, 36.46%, Fisher's exact test: Odds ratio = 1.173, p-value = 0.069).

Depending on the mechanism through which duplicated genes were generated these can be whole-genome duplicates (Wolfe and Shields 1997; Marcet-Houben and Gabaldón 2015) or small-scale duplicates (Fares *et al.* 2013; Keane *et al.* 2014), also known as WGD and SSD respectively. In the yeast *S. cerevisiae* used in this study 553 pairs of duplicated genes are WGD corresponding to 50.2% of all duplicated genes, meanwhile, 548 are SSD. It is well known that the evolution of duplicated genes depends on the mechanism through which these arose (Carretero-Paulet *et al.* 2013; Fares *et al.* 2013; Keane *et al.* 2014), therefore we analyzed the distribution of WGDs and SSDs among all the altered duplicated genes in the population t_0. We observed that 60.38% of the altered duplicated genes were SSD (N=349), a proportion significantly higher than expected by chance (Exact binomial test: p-values = 6.81×10^{-7}). It is also surprising that out of the up-regulated 229 duplicates we observe no significant differences between SSDs (N=121, 52.84%) and WGDs (N=108, 47.16%, Exact binomial test: p-value = 0.43), whereas when considering down-regulated duplicates, we observe roughly the double of SSDs (N=228, 65.33%) duplicates compared to WGDs (N=121, 34.67%, Exact binomial test: p-value = 1.08×10^{-8}). Despite more SSDs altered the expression compared to WGDs, we didn't observe any difference in fold-change between SSDs (median 1.132) and WGDs (median = 1.116, Wilcoxon test: p-value = 0.27).

For the evolved population t_{100}, 37.14% (N=377) of all de-regulated genes were duplicates, a proportion higher than expected by chance (Exact binomial test: p-value = 3.61×10^{-3}). These results are similar to the results obtained in population t_0, since the main difference between these populations is a higher genetic variability. Likewise, despite that at the population t_{100} we observed more genes that were up-regulated, there was a lower proportion of up-regulated duplicated genes (N=175, 28.46% out of all up-

regulated genes) than duplicated genes that were down-regulated (N=202, 50.5% out of all the down-regulated genes, Fisher's exact test: Odds ratio = 2.56, p-value = 2.201 x 10^{-12}). Contrary, in the population t_{100} the proportion of SSDs (N=198, 52.52%) was similar to the number of WGDs (N=179, 47.48%, Exact binomial test: p-value = 0.35), and analyzing the distribution of WGDs and SSDs for up- and down-regulated genes, we observed a higher proportion of WGDs that were up-regulated (N=100, 57.14%), while the opposite was true for down-regulated duplicated genes (N=79, 39.11%, Fisher's exact test: Odds ratio = 2.07, p-value = 6.26 x 10^{-4}). Surprising is that analyzing the fold-change of the altered genes at population t_{100} we observed no difference between the expression levels of duplicated genes (median = 1.07) and the expression levels of singletons (median = 1.118, Wilcoxon test: p-value = 0.57), however, we observed the opposite that at t_0, with SSDs showing higher fold-change differences (median = 1.167) than WGDs (median = 1.024, Wilcoxon test: p-value = 0.048).

For the population t_{110} 37.06% of the duplicated genes altered their expression, being this proportion also significantly higher than expected (Exact binomial test: p-value = 4.25 x 10^{-5}), and also the fold-change in the expression of duplicated genes (median = 1.192) was significantly higher than for singletons (median = 1.132, Wilcoxon test: p-value = 0.031). In agreement to the results observed in the populations t_0 and t_{100}, more duplicated genes were down-regulated (N=451, 38.98%) than up-regulated (N=318, 34.64%, Fisher's exact test: Odds ratio = 0.83, p-value = 0.044). Similarly, when we analyzed the distribution of WGD and SSD, we observed more SSDs (N=440, 57.22%) that altered the expression than WGDs (N=329, 42.78%, Exact binomial test: p-value = 7.09 x 10^{-5}). As observed in the other populations, at population t_{110} the same pattern was observed for up- and down-regulated genes. In both cases, we observed more SSDs that altered the expression than WGDs (UP: Exact binomial test: p-value = 0.037; DOWN: Exact binomial test: p-value = 6.81 x 10^{-4}) Furthermore, we observed that WGDs showed the same level of fold-change (median = 1.254) than SSDs (median = 1.136, Wilcoxon test: p-value = 0.157).

e. A huge cellular re-programming is driven through duplicated genes

We identified the duplicated genes that were altered at each population and the core gene set. We found that 148 genes belong to the core category (Figure ChV-4A) and are mainly involved in the generation of precursor metabolites and energy.

However, in the population t_0 altering the expression of duplicated genes results in an important cellular response to lactic acid. Important processes for the cell are altered such as "gene expression and cytoplasmic translation", "ribosome assembly", "cellular

component biosynthesis", "nitrogen catabolic processes", "RNA metabolism", "processing and transport" (Figure ChV-4B). In the population t_{100} fewer processes are altered, but de-regulation mainly concerns "generation of metabolites and energy", "cellular respiration", "oxidation-reduction processes" and "hexose transport" (Figure ChV-4C). Finally, the adapted population t_{110} altered "gene expression and translation", "ribosome assembly", "nitrogen compound and other small molecule metabolism", and "RNA processing and transport" (Figure ChV-4D).

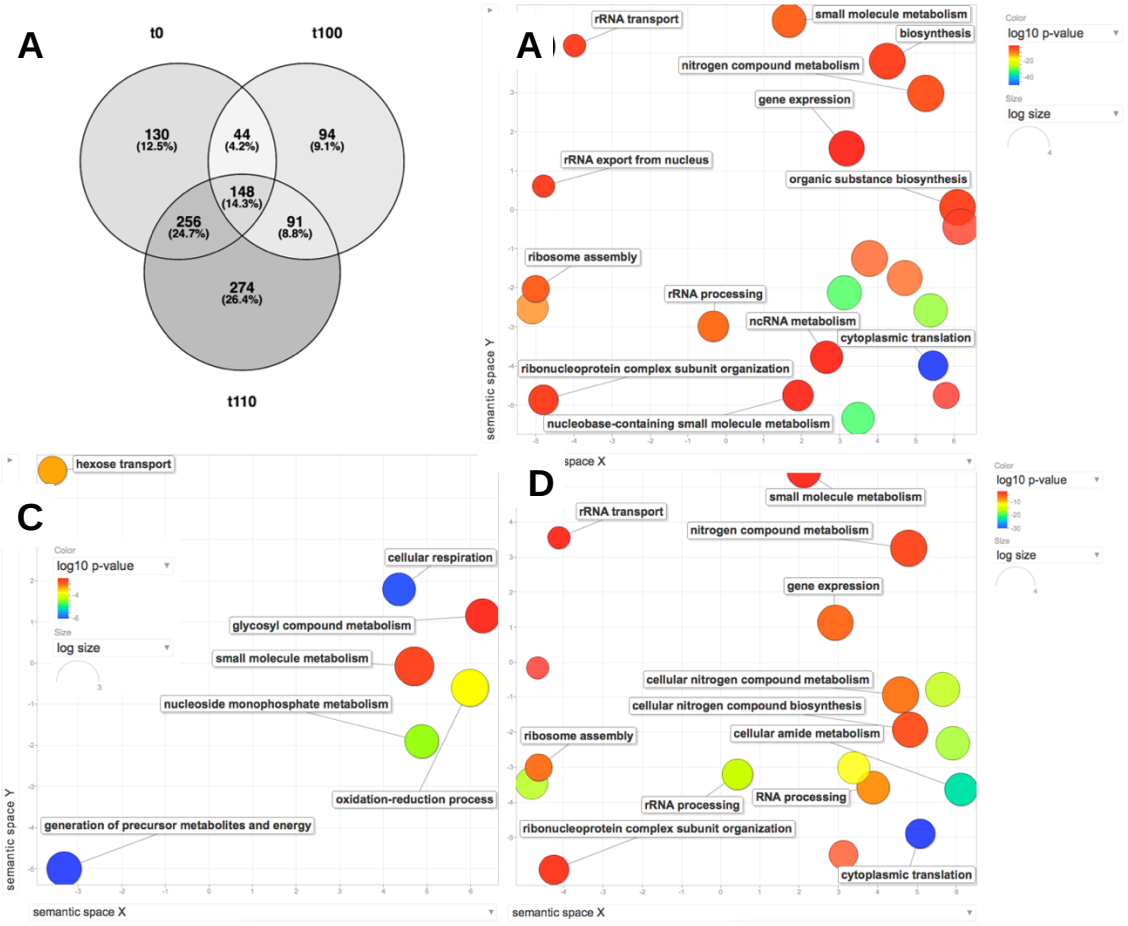


Figure ChIV- 4 – Duplicated genes undergo a major transcriptional change when the yeast S.cerevisiae populations challenged to lactic acid/lactate induced stress conditions. **A)** Venn diagram identifying the number of transcriptionally altered duplicated genes analyzing the level of sharing between the different populations (t_0, t_{100}, t_{110}) (done with http://bioinfogp.cnb.csic.es/tools/venny/index.html)**. B, C, D)** Semantic clustering of cellular processes enriched for duplicated genes that are transcriptionally altered in the populations t_0, t_{100} and t_{110}, respectively, growing in YPL. The color of the bubbles represents the proportion of genes in a particular cellular process (log P-value), while the size of the bubble indicates the frequency of the GO term in the organism.

f. Metabolic evolution of lactic acid/lactate adapted S.cerevisiae populations

To determine how similar are the different populations after the experimental evolution we measured the metabolic distance among the three populations, being this the grade of sharing of cellular GO processes to determine how close are the populations in terms of metabolism (see *Materials and Methods*). We performed the calculation of the metabolic distance for all the transcriptionally altered genes (including duplicates and singletons), but also for the up- and down-regulated genes between pairs of the t_0, $t_{100,}$ and t_{110} populations. As expected we observed that the metabolic most related populations for the whole altered gene set were t_0 and t_{100} since the population t_{100} is neutrally evolved from the population t_0 to increase the genetic variability. The same was true for the up-regulated gene set but surprisingly not for the down-regulated (Table 1). When we analyzed only the duplicated genes we observed that the closest metabolic distance for all comparisons was between populations t_0 and t_{110}.

Table ChIV- 1 - Metabolic distance between populations t_0, t_{100} and t_{110}

Genes	**Metabolic Distance ($MD_{i,j}$)**								
	Altered genes			**Up-regulated**			**Down-regulated**		
	t_0 *vs* t_{100}	t_0 *vs* t_{110}	t_{100} *vs* t_{110}	t_0 *vs* t_{100}	t_0 *vs* t_{110}	t_{100} *vs* t_{110}	t_0 *vs* t_{100}	t_0 *vs* t_{110}	t_{100} *vs* t_{110}
All	0.268	0.456	0.536	0.255	0.311	0.309	0.682	0.388	1.000
Duplicates	0.967	0.176	0.967	0.500	0.333	0.875	1.000	0.125	1.000

4. Discussion

In nature, organisms have to deal with a huge range of environmental fluctuations that generate a stressful situation that can, eventually, kill the organism. Therefore, evolution has selected rapid metabolic mechanisms to detect small environmental perturbations and react to overcome them. In this study, we show that changing the media composition substituting the dextrose by lactic acid drives to a complete transcriptomic re-wiring in the yeast *S.cerevisiae.* The addition of lactic acid to media, and the removal of dextrose as the principal carbon source, leads to a vast change in the composition of the media complicating the access to a carbon source and metabolic precursors. When the budding yeast is challenged against the stress media a high gene de-regulation and a huge cellular re-programming occurs affecting mainly the metabolic rate and energy consumption. Other experimental evolution experiments performed in yeast under glucose limiting conditions have shown that it occurs a huge transcriptional switch (Ferea *et al.* 1999) and that this response can range from being a very fast transcriptomic response, when the yeast encounters an acute stress situation, to very slow responses under chronic stress situations, involving in both cases a large number of genes (Taymaz-Nikerel *et al.* 2016). In good agreement, our results show thousands of genes altered in a very short time when the yeast was challenged to grow in the harsh media. Similar results were observed in the hybrid yeast *Zygosaccharomyces parabailii* when lactic acid was exogenously added to the media (Ortiz-Merino *et al.* 2017), and also in the yeast *Saccharomyces cerevisiae* when it was challenged against glycerol (Mattenberger, Sabater-Muñoz, Hallsworth, *et al.* 2017) or ethanol (Sabater-Muñoz *et al.* 2020). In this study, we observe that at all time points at which we analyzed the transcriptomic response the Krebs cycle was up-regulated, and glycolysis was down-regulated. This might be since the huge biochemical change occurring in the media forces the yeast to open new ways to obtain energy. Indeed, in the ancestor population (t_0) we observe roughly half of the altered genes to be up-regulated involving TCA, cellular respiration, and hydrogen transmembrane transport in concordance with the expected biochemical alterations in the media due to the lactic acid.

In addition, our results show that the genetic composition of the population is important when an environmental fluctuation occurs. An isogenic population – with low genetic variability – as occurs in t_0 because of being founded by a single yeast colony, has a strong effect on the capability of the population to overcome the stressful situation. According to the results that have been previously observed in glycerol and ethanol, the growth rate of the population t_0 under stress was significantly lower than for the evolved population t_{100} (Mattenberger, Sabater-Muñoz, Hallsworth, *et al.* 2017; Sabater-Muñoz

et al. 2020). Neutral evolution in standard YPD media during 660 generations allows the population to increase genetic variability by accumulating cryptic neutral variation, hence increasing the phenotypic plasticity of the population (Wagner 2005, 2011; Mattenberger, Sabater-Muñoz, Toft, and Fares 2017). Once achieved enough variability within the populations, we observe that the adaptation process to glucose deprivation – substituting the glucose of the media with 3% of lactic acid – is produced very fast. Indeed, in only 33 generations a significant effect on the maximum growth rate of the population was observed. However, despite that the adaptation is produced very fast, the response to adding lactic acid to the media is mainly a transcriptional response rather than a genotypic switch of the population, since the evolution time to select and fix fitter genotypes in the population during this experiment was too short (~33 generations). According to this hypothesis, our results did not show any detrimental effect on the capability for the evolved yeast populations to grow in the standard rich media with glucose, contrary to what would be expected if there were a genotypic switch to a new media with an obligated different principal carbon source (Tamari *et al.* 2016).

It is well known that gene duplication is a major force in the evolution of new genetic material and novel functions (Zhang 2003). Since gene duplication has been related from the radiation of flowering plants (Cui *et al.* 2006; Panchy *et al.* 2016) to an ongoing increase of complexity in animals (Freeling and Thomas 2006; Jaillon *et al.* 2009), in this study we were interested in deciphering the role of the duplicated genes during the adaptation to a challenging environment. Our results show that at the t_0 isogenic population, there were more duplicated genes that alter the expression than expected, indicating the importance of gene duplication for stress response and phenotypic plasticity. According to this, we observe also in the population t_{100} and t_{110} that there is a higher number of duplicated genes that alter the expression compared to singletons, pointing out the importance of duplicated genes in adaptation. Indeed, not only more duplicated genes alter the expression but those also show a higher fold-change than singletons. Additionally, previous work has shown that the origin of the duplication plays a role in its evolutionary fate, being small-scale duplicates (SSD) more prone to neo-functionalization while whole-genome duplicates (WGD) generally end up in sub-functionalization (Carretero-Paulet and Fares 2012; Fares *et al.* 2013). Our results show that most of the duplicated genes that alter the expression were originated from SSD, hence neo-functionalized since they arose. We also observed that more duplicates were down-regulated, and out of the down-regulated the majority were SSD. A possible explanation of these findings is that the SSD that are downregulated in YPL are important for glucose metabolic innovation history, hence neo-functionalized when they originated

to maximize glucose usage. The same was observed at the population t_{100} and in the population t_{110}, in which we also found more down-regulated duplicated genes, and in both those were SSD. Accordingly, in all three populations we found that out of the up-regulated duplicates there were more WGD probably because WGD are more prone to sub-functionalize (Carretero-Paulet and Fares 2012; Fares *et al.* 2013). Hence, since all functions and interactions of the duplicated genes are doubled in the organisms after a WGD makes these genes more robust to maintain the ancestral function despite losing some of its interactors. Thus, the unbalanced loss of genes after WGDs encourages the system to be more prone to sub-functionalize the duplicated genes. Our results stand by the fact that more WGD were observed among the up-regulated duplicates because those genes are the result of a sub-functionalization simplifying the metabolic swift from one carbon source to another one.

With our findings, we show that the metabolic switch among different carbon sources, particularly from dextrose to lactic acid/lactate, is conducted by a complete transcriptional rewiring mainly invoking duplicated genes that enhance phenotypic plasticity and drive adaptation to new unexplored environments. Despite many questions still need to be answered in further approaches, characterizing the genetic and cellular response to the addition of lactic acid/lactate to the media is of great interest for the industry since it has become a very valuable chemical product in society.

5. Supplementary data

Supplementary data can be found in the annex of this thesis. **Table S1** GO process obtained for de-regulated genes at t_0 **Table S2** GO process obtained for de-regulated genes at t_{100} **Table S3** GO process obtained for de-regulated genes at t_{110} **Table S4** GO process obtained for the de-regulated genes shared between t_0, t_{100} and t_{110}. **Table S5** GO process obtained for the de-regulated duplicated genes at t_0. **Table S6** GO process obtained for the de-regulated duplicated genes at t_{100}. **Table S7** GO process obtained for the de-regulated duplicated genes at t_{110}. **Table S8** GO process obtained for the up-regulated genes shared between t_0 and t_{100}. **Table S9** GO process obtained for the down-regulated genes shared between t_0 and t_{100}.

PART II – Biological innovation through mutation: functional sequence space of a viral capsid.

Objectives to achieve in this part:

4. Determine the viable sequence space of coxsackievirus B3 capsid proteins by Deep Mutational Scanning and experimental evolution. **Chapter VI and VII.**

CHAPTER VI – Defining the complete sequence space of the Coxsackievirus B3 capsid by Deep Mutational Scanning.

A version of this chapter has been published as:

Mattenberger, F., **Latorre, V., Tirosh, O., Stern, A. & Geller, R. (2021)**. *Globally defining the effects of mutations across a picornavirus capsid*. ***eLife, 10:e64256*.**

1. Abstract

The capsids of non-enveloped viruses are highly multimeric and multifunctional protein assemblies that protect the viral genome between infection cycles, dictate host and cell tropism, and mediate evasion of humoral immune responses. As such, capsids play key roles in viral biology and pathogenesis. Despite their importance, a comprehensive understanding of how mutations affect viral fitness across different structural and functional attributes of the capsid is lacking. To address this limitation, we globally define the effects of mutations in the capsid of a human picornavirus, generating a comprehensive dataset encompassing >90% of all possible single amino acid mutations. Moreover, we use this information to identify structural and sequence determinants that accurately predict mutational fitness effects, refine evolutionary analyses, and define the sequence specificity of key capsid encoded motifs. Finally, capitalizing on the sequence requirements identified in our dataset for capsid encoded protease cleavage sites, we implement and validate a bioinformatic approach for identifying novel host proteins targeted by viral proteases. Our findings present the most comprehensive investigation of mutational fitness effects in a picornavirus capsid to date and illuminate important aspects of viral biology, evolution, and host interactions.

2. Introduction

The capsids of non-enveloped viruses are among the most complex of any viral protein. These highly multimeric structures must correctly assemble around the genome from numerous subunits, at times numbering in the hundreds, while avoiding aggregation (Harrison 2013; Hunter 2013; Perlmutter and Hagan 2015). Moreover, the assembled structure must be both sufficiently stable to protect the viral genome during its transition between cells yet readily disassemble upon entry to initiate subsequent infections. For these functions to be achieved, viral capsids must encode the information for interacting with numerous cellular factors that are required to correctly fold and assemble around the genome (Macejak and Sarnow 1992; Callaway *et al.* 2001; Geller *et al.* 2007; Fields *et al.* 2013; Jiang *et al.* 2014). Viral capsids also play key roles in pathogenesis, dictating host and cell tropism by encoding the determinants for binding cellular receptors (Rossmann *et al.* 2002; Helenius 2013) and mediating escape from humoral immune responses (Heise and Virgin 2013; Cifuente and Moratorio 2019). As a result, viral capsids show the highest evolutionary rates among viral proteins.

The picornaviruses constitute a large group of single-stranded, positive-sense RNA viruses, and include several pathogens of significant medical and economic impact (Racaniello 2013). Their relative simplicity and ease of culture have made picornaviruses important models for understanding virus biology. Among the many breakthroughs achieved with these viruses was the determination of the first high-resolution structure of the capsid of an animal virus, making the picornavirus capsid the prototypical non-enveloped, icosahedral viral capsid (Racaniello 2013). Picornavirus capsid genesis initiates with the co-translational release of the P1 capsid precursor protein from the viral polyprotein via the proteolytic activity of the viral encoded 2A protease (Racaniello 2013; Jiang *et al.* 2014). Subsequently, the viral encoded 3CD protease ($3CD^{pro}$) cleaves the P1 capsid precursor to liberate three capsid proteins (VP0, VP3, and VP1), generating the capsid protomer. Five protomers then assemble to form the pentamer, twelve of which assemble around the viral genome to yield the virion. Finally, in some picornaviruses, VP0 is further cleaved into two subunits, VP4 and VP2, following genomic encapsidation to generate the infectious, 240 subunit particle (Racaniello 2013; Jiang *et al.* 2014). Work over the years has identified numerous host factors that help support capsid formation (Macejak and Sarnow 1992; Geller *et al.* 2007; Thibaut *et al.* 2014; Qing *et al.* 2014; Corbic Ramljak *et al.* 2018), defined antibody neutralization sites (Cifuente and Moratorio 2019), and identified numerous host receptors for many members of this viral family (Rossmann *et al.* 2002).

Despite significant progress in understanding the structure and function of picornavirus capsids, a comprehensive understanding of how mutations affect viral fitness across different structural and functional attributes is lacking. To address this, we perform a comprehensive analysis of mutational fitness effects (MFE) across the complete capsid region of the human picornavirus coxsackievirus B3 (CVB3), analyzing >90% of all possible single amino acid mutations. Furthermore, using this data, we develop models to predict the effect of mutations with high accuracy from available sequence and structural information, improve evolutionary analyses of CVB3, and define the sequence preferences of several viral encoded motifs. Finally, we use the information obtained in our dataset for the sequence requirements of capsid encoded 3CD protease cleavage sites to identify host targets of this viral protease. Overall, our data comprise the most comprehensive survey of MFE effects in a picornavirus capsid to date and provide important insights into virus biology, evolution, and interaction with the host.

3. Materials and Methods

a. Viruses, cells, and plaque assays

HeLa-H1 (CRL-1958) and HEK293 (CRL-1573) cells were obtained from ATCC. All work with CVB3 was based on the Nancy infectious clone (kind gift of Dr. Marco Vignuzzi, Institute Pasteur). Cells were cultured in culture media (DMEM with 10% heat-inactivated FBS, Pen-Strep, and L-Glutamine) with FBS concentrations of 2% during infection. For plaque assays, serial dilutions of the virus were used to infect confluent HeLa-H1 cells in 6 well plates for 45 minutes, followed by overlaying the cells with a 1:1 mixture of 56°C 1.6% Agar (Arcos Organics 443570010) and 37°C 2x DMEM with 4% FBS. Two days later, plates were fixed with formaldehyde (2% final concentration) after which the agar was removed and the cells stained with crystal violet to visualize plaques.

b. Deep mutational scanning (DMS)

The infectious clone was modified by site-directed mutagenesis to remove an XhoI site present in the capsid region (P1) and introduce an XhoI site at position 692 as well as a Kpn2I site at position 3314, generating a pCVB3-XhoI-P1-Kpn2I clone (Bou *et al.* 2019). In addition, a pCVB3-XhoI-ΔP1Kpn2I plasmid was generated by replacing the region between the XhoI and Kpn2I sites in pCVB3-XhoI-P1-Kpn2I with a short linker. To generate the template for DMS, the capsid region was amplified by PCR from pCVB3-XhoI-P1-Kpn2I with Phusion™ polymerase (Thermo Scientific) and primers HiFi-F (CTTTGTTGGGTTTATACCACTTAGCTCGAGAGAGG) and HiFi-R (CCTGTAGTTCCCCACATACACTGCTCCG) and gel purified (Zymoclean™ Gel DNA Recovery Kit). Primers spanning the full coding region of the capsid region were designed using the CodonTilingPrimers software from the Bloom lab (https://github.com/jbloomlab/CodonTilingPrimers) with the default parameters and synthesized by IDT (Supplementary Table S1). These primers were used to perform the mutagenesis PCR on the capsid template together with the HiFi-F or HiFi-R primers in triplicate following published protocols (Dingens *et al.* 2017) with the exception that 10 rounds of mutagenesis were performed for libraries 1 and 2, while a second round of 7 mutagenesis cycles was performed for library 3 to increase the number of mutation per clone. The products were gel purified and ligated to an XhoI and Kpn2I digested and gel purified pCVB3-XhoI-ΔP1Kpn2I using NEBuilder® HiFi DNA Assembly reaction (NEB) for 25 minutes. Mutagenesis efficiency was evaluated by the transformation of the assembled plasmids into NZY5α competent cells (NZY Tech), Sanger sequencing of 18-23 clones per library, and mutation analysis using the Sanger Mutant Library Analysis script (https://github.com/jbloomlab/SangerMutantLibraryAnalysis).

Subsequently, the assembled plasmid reactions were purified using a Zymo DNA Clean & Concentrator-5 kit (Zymo Research) and used to electroporate MegaX DH10B™ $T1^R$ Electrocomp™ cells (ThermoFisher) using a Gene Pulser XCell™ electroporator (BioRad) according to the manufacturer's protocol. Cells were then grown overnight in a 50 mL liquid culture at 33°C and DNA purified using the PureLink™ HiPure plasmid midiprep kit (Invitrogen). Transformation efficiency was estimated by plating serial dilutions of the transformation on agar plates. In total, $4.44x10^5$, $1.46x10^5$, and $2.19x10^5$ transformants were obtained for lines 1, 2, and 3, respectively.

Viral genomic RNA was then transcribed from SalI linearized, gel-purified full-length plasmids using the TranscriptAid T7 kit (ThermoScientific), and four electroporations were performed using $4x10^6$ HeLa-H1 cells in a 4mm cuvette in 400µL of calcium and magnesium-free PBS using with 8µg of RNA in a Gene Pulser XCell™ electroporator (BioRad) set to 240V and 950µF. Electroporated cells were then pooled, and one fourth was cultured for 9 hours to produce the passage 0 virus (P0). Following three freeze-thaw cycles, $2x10^6$ plaque forming units (PFU) were used to infect a 90% confluent 15cm plate in 2.5mL of infection media for 1 hour. Cells were then washed with PBS and incubated in 12 mL of infection media for 9 hours. Finally, cells were subjected to 3 freeze-thaw cycles, debris removed by centrifugation at 500xg and the supernatants collected to generate P1 virus stocks. All infections produced > $2.38x10^6$ PFU in P0 and > $1.2x10^7$ PFU in P1 as judged by plaque assay.

c. NGS analysis

Libraries were prepared following published protocols (Kennedy *et al.* 2014) and each library was run on a NovaSeq6000 2x150 at a maximum of 30G per lane to reduce potential index hopping. Reads trimming was performed using fastp (Chen *et al.* 2018) (command: -max_len1 150 --max_len2 150 --length_required 150 -x -Q -A), unsorted bam files were generated from fastq files using Picard tools FastqToSam (version 2.2.4) and merged into a single bam using the cat command of Samtools (version 1.5). The duplex pipeline was then implemented (https://github.com/KennedyLabUW/Duplex-Sequencing/UnifiedConsensusMaker.py) using the UnifiedConsensusMaker.py script and a minimum family size of 3, a cutoff of 0.9 for consensus calling, and an N cutoff of 0.3. The single-stranded consensus files (SSCS) were then aligned using BWA mem (version 0.7.16), sorted using Samtools, size selected to be 133 bp long using VariantBam (Wala *et al.* 2016), unaligned reads were discarded (Samtools view command with -F 4), and the resulting bam file indexed with Samtools.

Subsequently, fgbio (http://fulcrumgenomics.github.io/fgbio/; version 1.1.0) was used to hardclip 10 bp from each end and upgrade all clipping to hard-clip (-c Hard --upgrade-clipping true --read-onefive-prime 10 --read-one-three-prime 10 --read-two-five-prime 10 --read-two-three-prime 10). Variant bam was then used to keep all reads that were between 50-150bp, well-mapped, and had either no indels and less than 5 mutations (command –r {"":{"rules":[{"ins":[0,0],"del":[0,0],"nm":[0,4], "mate_mapped":true,"fr":true,"length":[50,150]}]}}'). Finally, the codons in each read were identified using the VirVarSeq (Verbist *et al.* 2015) Codon_table.pl script using a minimum read quality of 20. A custom R script was then used to generate a codon counts table for each codon position by eliminating all codons containing ambiguous nucleotides and codons with a strong strand bias (StrandOddsRatio > 4), as well as all codons that are reached via a single mutation (available at https://github.com/RGellerLab/CVB3_Capsid_DMS).

Amino acid preferences and mutational fitness effects were determined using DMStools2 (Bloom 2015) with the Bayesian option and the default settings.

d. Structural analyses

The crystal structure PDB:4GB3 (Yoder *et al.* 2012) was used for all structural analyses. The effects of mutations on aggregation were determined using TANGO version 2.3.1 (Fernandez-Escamilla *et al.* 2004) using the default settings and the effect on stability on the monomer and pentamer was determined using FoldX 4 (Schymkowitz *et al.* 2005) using the default settings. For the latter, the pentamer subunits were renamed to unique letters, all mutations between the reference sequence and the structure sequence were introduced using the BuildModel command, the structure was optimized using the RepairPDB command 5 or 10 times for the pentamer or monomer, respectively, and then the effects of the mutations were predicted using the BuildModel command (modified PDB files can be found at https://github.com/RGellerLab/CVB3_Capsid_DMS). Secondary structure and RSA were obtained from DSSP (http://swift.cmbi.ru.nl/gv/dssp/) using the dms_tools2.dssp function of dms_tools2, while interface, surface, and core residues as well as residue contact number, and presence in the two, three, and five-fold axes were obtained from ViprDB (http://viperdb.scripps.edu/) (Carrillo-Tripp *et al.* 2009). Distance from the center was calculated with Pymol using the Distancetoatom.py script on the monomer or pentamer.

e. Generation and evaluation of CVB3 capsid mutants

The PCR of the capsid region used as a template for DMS was phosphorylated and cloned into a SmaI digested pUC19 vector for use in the mutagenesis reactions (pUC19-HiFi-P1). For each mutant, non-overlapping primers containing the mutation in the middle of the forward primer were used to introduce the mutation with Phusion polymerase, followed by DpnI (Thermo Scientific) treatment, phosphorylation, ligation, and transformation of chemically competent bacteria. Successful mutagenesis was verified by Sanger sequencing. Subsequently, the capsid region was subcloned into pCVB3-XhoI-ΔP1-Kpn2I using XhoI and Kpn2I sites. Plasmids were then linearized with MluI and 2µg of plasmid was transfected into $5x10^5$ HEK293 cells together with a plasmid encoding the T7 polymerase (Yun *et al.* 2015) (Addgene 65974) using calcium phosphate. Briefly, an equal volume of 2x HBS (274mM NaCl, 10mM KCl, 1.4mM Na_2HPO_4) was added dropwise to DNA containing 0.25M $CaCl_2$ while mixing, incubated 15 minutes at RT, and then added dropwise to cells. Following 48 hours, passage 0 (P0) virus was collected and titered by plaque assay. From this, 10^5 PFU were used to infect 90% confluent 6 well HeLa-H1 cells (Multiplicity of infection [MOI] 0.1) for 1 hour at 37°C, after which the cells were washed twice with PBS and 2mL of infection media added. Cells were then incubated until CPE was observed. Emerging viral populations were titered by plaque assay and the capsid region sequenced to ensure no compensatory mutations or reversions arose during replication. The fitness of these mutants was then tested by direct competition with a marked reference virus using a TaqMan RT-PCR method (Moratorio *et al.* 2017). Briefly, in quadruplicates, confluent HeLa-H1 cells in a 24 well plate were infected with 200µL of a 1:1 mixture of $4x10^3$ PFU (MOI 0.01) of the test and marked reference viruses for 45 minutes. Subsequently, the inoculum was removed, the cells were washed twice with PBS, 200µL of infection media was added, and the cells were incubated for 24 hours at 37°C. Finally, cells were subjected to 3 freeze-thaw cycles, debris removed by centrifugation at 500xg, the supernatants collected and treated with 20µL of RNase-Free DNaseI (ThermoFisher) for 15 minutes at 37°C, and viral RNA extracted using the Quick-RNA™ Viral Kit (Zymo Research), eluting in 20µl. Quantification of the replication of each mutant versus the reference was performed using Luna® Universal Probe One-Step RT-qPCR kit (New England BioLabs) containing 3µL of total RNA, 0.4µM of each qPCR primers and 0.2µM of each probe. The standard curve was performed using 10-fold dilutions of RNA extracted from 10^7 PFU of wild-type and reference viruses. All samples were performed with three technical replicates. The relative fitness (W) of each mutant versus the common marked reference virus was calculated using the formula $W = [R(t)/R(0)]^{1/t}$, where R(0) and R(t) represents

the ratio of the mutant to the reference virus genomes in the initial mixture used for the infection and after 1 day (t=1), respectively (Carrasco *et al.* 2007; Moratorio *et al.* 2017).

f. Sequence variability and phylogenetic analyses

Amino acid variability was assessed using Shannon entropy. Briefly, all available, non-identical, full-genome CVB3, CVB, or Enterovirus B sequences were downloaded from Virus Pathogen Resource (Pickett *et al.* 2012) (www.viprbrc.org) and codon-aligned using the DECIPHER package in R (available at https://github.com/RGellerLab/CVB3_Capsid_DMS). All alignment positions not present in our reference strain were removed, and a custom R script was used to calculate Shannon entropy. For phylogenetic and differential selection analyses, PhyDMS was run using the default settings on an alignment of CVB3 genomes that was processed with the phydms_prealignment module and using the average preferences from the three DMS replicates.

g. Identification of 3CD^{pro} cleavage sites in the human proteome

The amino acid preferences (the relative enrichment of each amino acid at each position standardized to 1) was used to generate *in silico* 1000 peptides spanning the 10 amino acid region surrounding each cleavage site using a custom R script (available at https://github.com/RGellerLab/CVB3_Capsid_DMS). Specifically, for each peptide position, 100 peptides were generated that encoded each amino acid at a frequency corresponding to its preference observed in the DMS results, with the remaining positions unchanged. The resulting 1000 peptides from each cleavage site were uploaded to PSSMSearch (Krystkowiak *et al.* 2018) (http://slim.icr.ac.uk/pssmsearch/) using the default setting (psi_blast IC). Results were filtered to remove proteins indicated to be secreted, lumenal, or extracellular in the Warnings column. To test whether proteins were cleaved by the viral 3CD protease, the corresponding region was PCR amplified from the Nancy infectious clone (primers 3C-For: TATTCTCGAGACCATGGGCCCTGCCTTTGAGTTCG and 3D-Rev: TATTGCGGCCGCCTAGAAGGAGTCCAACCATTTCCT) and cloned into the pIRES plasmid (Clonetech) using the restriction sites XhoI and NotI (pIRES-3CD^{pro}). For analysis of fusion proteins, HEK293 cells were transfected with GFP-PLEKHA4 (kind gift of Dr. Jeremy Baskin, Cornell University), GFP-PLSCR1 (kind gift of Dr. Serge Benichou, Institut Cochin), FLAG-NLCR5 (Addgene #37521), HA-ZC3HAV1 (Addgene #45907), or the control plasmid FLuc-eGFP (Addgene #90170) together with the pIRES-3CD^{pro} plasmid using Lipofectamine™ 2000. Following 24 hours, proteins were collected by lysing in lysis buffer (50mM TRIS-HCl, 150mM NaCl, 1% NP40 and protease inhibitor

cocktail [Complete Mini EDTA-free, Roche]) and subjected to western blotting with the corresponding antibody (anti-GFP, Santa Cruz sc-9996; Anti FLAG, Santa Cruz sc-166335; anti-HA, Santa Cruz, sc-7392). For analysis of endogenous proteins, $3CD^{pro}$ was expressed for 48 hours before cell lysis, and western blotting using antibodies against WDR33 (Santa Cruz sc-374466), TSG101 (Santa Cruz sc-136111), GAK (Santa Cruz sc-137053), and MAGED1 (Santa Cruz sc-393291). When indicated, the $3C^{pro}$ inhibitor rupintrivir (Tocris Biosciences) was added at a concentration of 2 µM for the last 24 hours before collection. The predicted molecular weight of cleaved fragments was calculated using the mw function of the Peptides R package (version 2.4.2).

h. Statistical analyses

All statistical analyses were performed in R and were two-tailed. For random forest prediction, the R RandomForest package (version 4.6-14) was employed using the default setting with an mtry of 10, and for the linear model, the formula lm(MFE ~ enterovirus B entropy + WT amino acid * mutant amino acid + predicted effect of mutations on stability in the pentamer + relative surface exposure) was used (available at https://github.com/RGellerLab/CVB3_Capsid_DMS). Sequence logoplots were producing using Logolas (Dey *et al.* 2018)

4. *Results*

a. *Deep mutational scanning of a CVB3 capsid*

To generate CVB3 libraries encoding a large amount of diversity in the capsid region, we used a codon level PCR mutagenesis method (Bloom 2014). The mutagenesis protocol was performed on the capsid precursor region P1 in triplicate to generate three independent mutagenized libraries (Mut Library 1-3; Figure ChVI-1A). From these, three independent viral populations (Mut Virus 1-3) were derived by electroporation of in vitro transcribed viral RNA into HeLa-H1 cells (Figure ChVI-1A). High-fidelity next-generation sequencing (Bloom 2014) was then used to analyze the mutagenized libraries and resulting viruses, unmutagenized virus populations (WT virus 1-2), as well as controls for errors occurring during PCR (PCR) and reverse transcription (RT-PCR). High coverage was obtained for all samples (>10^6 per codon across all experimental conditions and >$6.5x10^5$ for the controls; Supplementary Table S2). Due to the high rate of single mutations within codons observed in the RT-PCR control compared to the mutagenized virus populations (Supplementary Table S2), all single mutants were omitted from our analysis to increase the signal-to-noise ratio. While this resulted in an inability to analyze 83.4% of synonymous codons in the capsid region (1746/2094) only 2.8% of non-synonymous mutations were lost to analysis (458/16169). Upon removing single mutations within codons, we obtained a large signal-to-noise ratio in the average mutation rate of 510x (range 449–572) and 245x (range 174–285) for the mutagenized libraries and viruses, respectively, compared to their error controls (Figure ChVI-1B and Supplementary Table S2). On average, 0.9 (range 0.8–1.02) codon mutations were observed per genome, which was in agreement with Sanger sequencing of 59 clones (range 18–23 per library; Figure S1 and Supplementary Table S3). As expected, the rate of stop codons, which should be invariably lethal in the CVB3 capsid, decreased significantly following growth in cells to <0.5% of that observed in the corresponding mutagenized libraries ($p < 0.005$ by paired t-test on log-transformed data; Supplementary Table S2). No major bias was observed in the position within a codon where mutations were observed (Figure S2A) nor in the type of mutation (Figure S2B), except for the WT virus, which had a high rate of A to G transitions in the two independent replicates analyzed. Of all 16169 possible amino acid mutations in the capsid region (851 AA x 19 AA mutation = 16169), a total of 14839 amino acid mutations were commonly observed in all three mutagenized libraries, representing 91.8% of all possible amino acid mutations in the capsid region, allowing us to globally assess the effects of the vast majority of amino acid mutations on the capsid (Figure ChVI-1C).

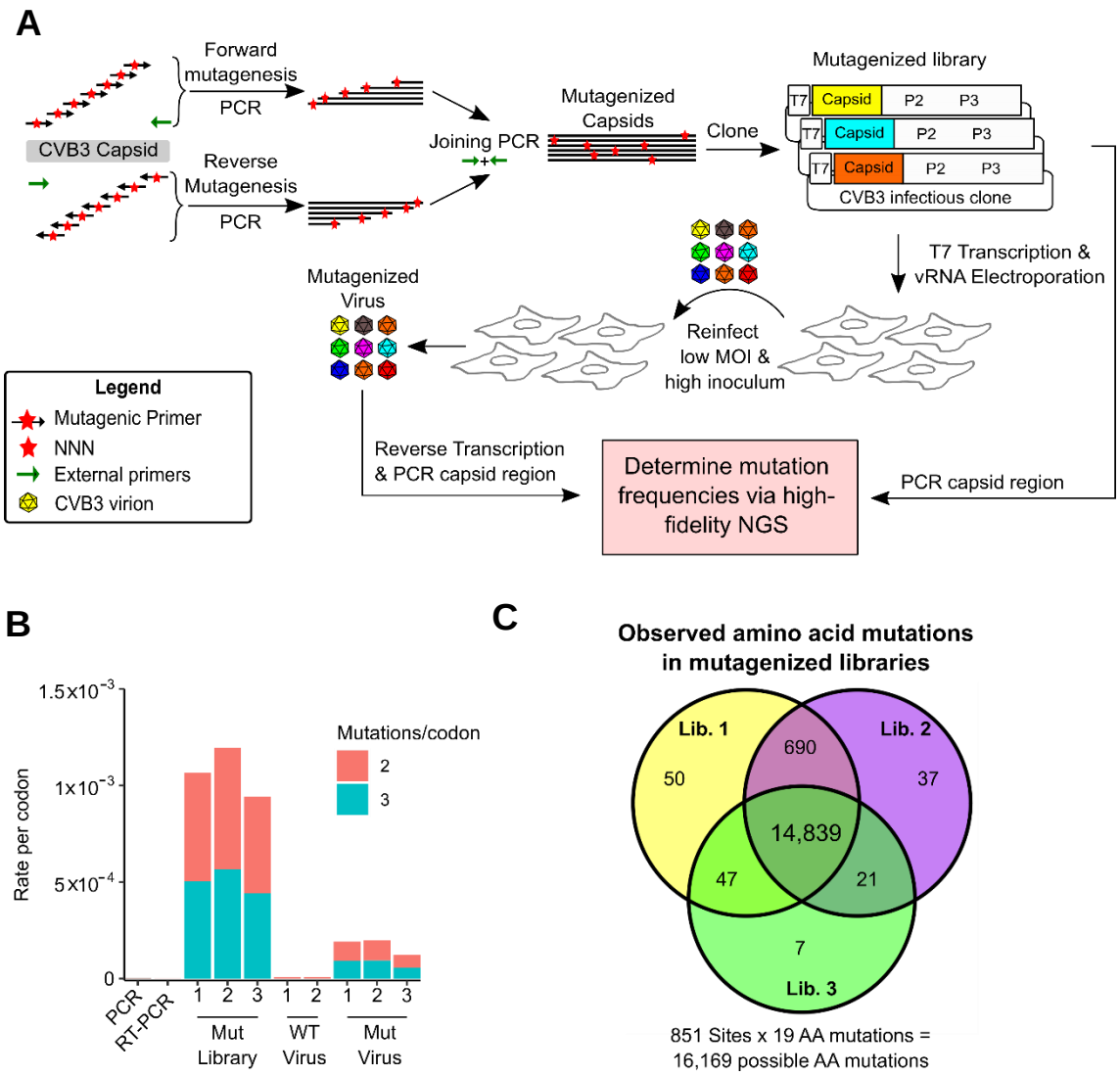


***Figure ChVI- 1* – Deep mutational scanning (DMS) of the CVB3 capsid. A)** Overview of the deep mutational scanning experimental approach. A forward mutagenesis PCR reaction was performed using a single external reverse primer and a pool of forward mutagenic primers targeting each codon in the capsid, each primer encoding degenerate nucleotides (NNN) at the codon matching position and template-complementary sequences upstream and downstream of the codon site. Similarly, a separate reverse mutagenesis PCR reaction was performed. The products of these PCRs were joined using the external primers and cloned into the CVB3 infectious clone to generate the mutagenized libraries. This process was performed in triplicate, generating 3 libraries (Mut Library 1-3). Viral genomic RNA (vRNA) was then produced from the mutant libraries via *in vitro* transcription and electroporated into cells to generate high diversity CVB3 populations (Mut Virus 1-3). The relative frequency of each mutation relative to the WT was then determined in both the mutagenized libraries and the resulting virus populations via high-fidelity duplex sequencing. **B)** The average rate of double or triple mutations per codon observed in the mutagenized libraries (Mut Library 1-3), the resulting mutagenized virus (Mut virus 1-3), as well as controls for the error rate of the amplification and sequencing process (PCR and RT-PCR) and the WT unmutagenized virus (WT Virus 1-2). Single mutations per codon were omitted from the analysis to increase the signal-to-noise ratio. **C)** Venn diagram showing the number of amino acid mutations observed in the mutagenized libraries.

b. Mutational fitness effects across the CVB3 capsid

We next derived the mutational fitness effects (MFE) of each observed mutation by examining how its frequency changed relative to that of the WT sequence following growth in cells. The preferences for the different amino acids at each position (amino acid preferences (Bloom 2015)) showed a high correlation between biological replicates (Spearman's $\rho > 0.83$; Supplementary Figure S3 and Supplementary Table S4 MFE). Overall, most mutations in the capsid were deleterious, with only 1.2% of mutations increasing fitness relative to the WT amino acid (Figure ChV-2A and Supplementary Table S4). Hotspots where mutations were tolerated were observed at several regions across the capsid (Figure ChVI-2A). These hotspots largely overlapped with highly variable regions in natural sequences, as measured by Shannon entropy in the enterovirus B family, indicating that lab measured MFE reflect natural evolutionary processes (Figure ChVI-2A, top). Indeed, a strong correlation was observed between MFE and sequence variability for the enterovirus B genus (Spearman's $\rho=0.59$, $p<10^{-16}$; Figure ChVI-2B). Similarly, antibody neutralization sites overlapped with hotspots for mutations (Figure ChVI-2A, top) and were significantly less sensitive to mutations ($p < 10^{-16}$ by Mann-Whitney test; Figure ChVI-2C). As expected, mutations were also less deleterious in loops compared to β-strands ($p < 10^{-16}$ by Kruskal-Wallis test; Figure ChVI-2D), at surface residues compared to core residues ($p < 10^{-16}$ by Kruskal-Wallis test; Figure ChVI-2E), and for mutations predicted to be destabilizing ($p < 10^{-16}$ by Mann-Whitney test; Figure ChVI-2F) or aggregation-prone ($p < 10^{-16}$ by Mann-Whitney test; Figure ChVI-2G). Importantly, independent validation of the MFE of 10 different mutants using a sensitive qPCR method (Moratorio *et al.* 2017) showed a strong correlation with the DMS results (Spearman's $\rho = 0.9$, $p < 0.001$; Supplementary Table S5).

c. Prediction of MFE from available structural and sequence information

As MFE correlated with natural sequence variation and different structural features of the capsid (Figure ChVI-2), we next investigated if MFE could be predicted from available structural and sequence information. For this, we obtained a dataset of 52 parameters, including structural information derived from the crystal structure of the CVB3 capsid (PDB:4GB3), amino acid properties, natural variation in available enterovirus sequences (Shannon entropy), and predicted the effects of mutation on stability and aggregation propensity using FoldX (Schymkowitz *et al.* 2005) and TANGO (Fernandez-Escamilla *et al.* 2004), respectively (Supplementary Table S6).

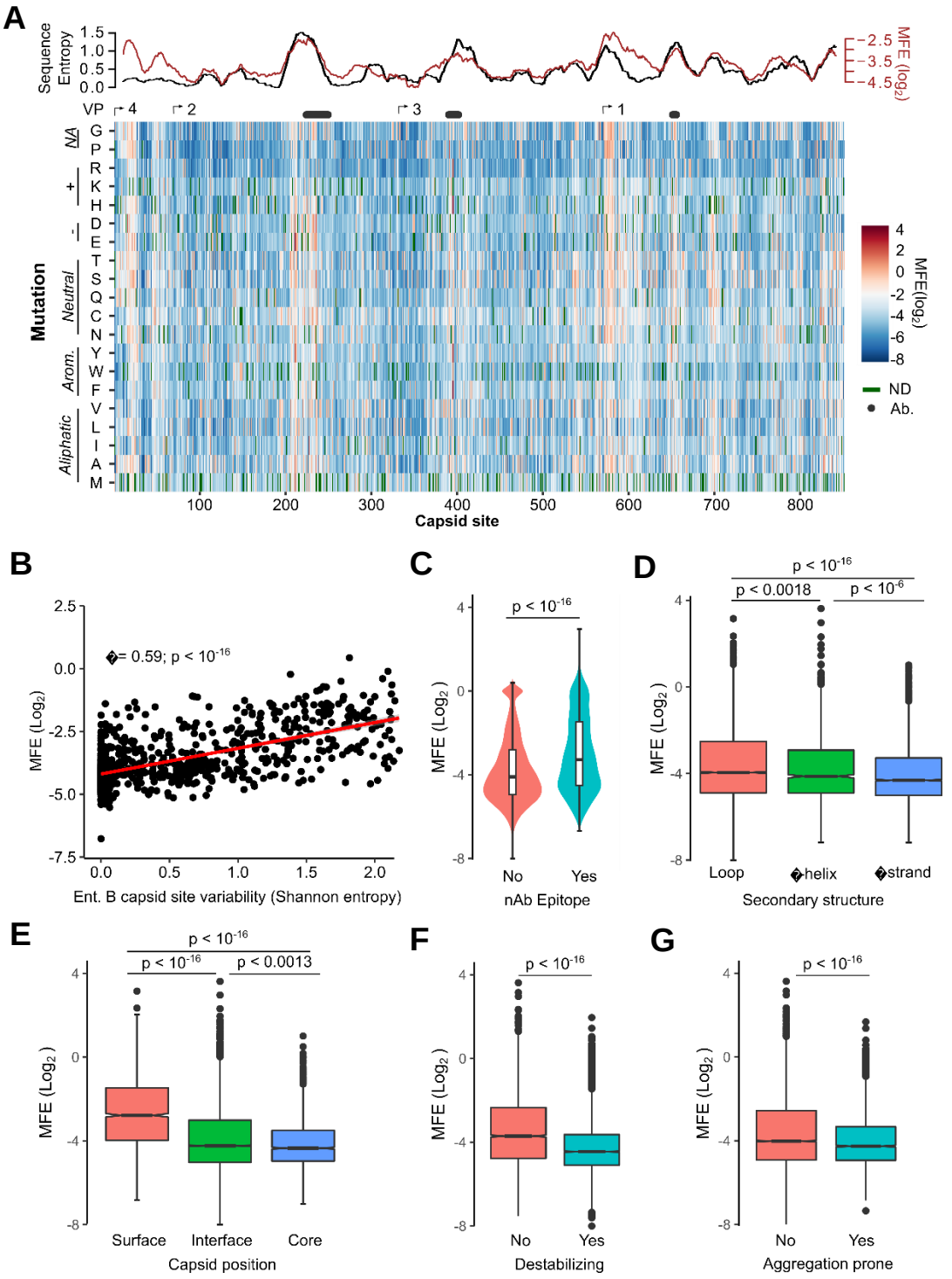


***Figure ChVI- 2* – Mutational fitness effects across the CVB3 capsid and their correlation with structural, evolutionary, and immunological attributes. A)** Overview of mutational fitness effects (MFE) across the CVB3 capsid. Bottom: a heatmap of MFE of all mutations observed at each position. Green indicates no data available (ND), and the position of the mature viral proteins (VP1-4) or antibody neutralization sites (nAb) are indicated above. Top: A 21 amino acid sliding window analysis of the average sequence variation in CVB3 genomes (Shannon entropy; black) or average MFE (red line). **B)** Correlation between derived MFE and variation in enterovirus B sequence alignments (Shannon entropy). **C)** Violion plot of MFE in antibody neutralization sites versus other capsid sites. **D-G)** Boxplots of MFE as a fluctuation of secondary structure (D), capsid €, or predicted effect of mutations on stability (F) or aggregation propensity (G) based on the 4GB3 capsid structure. Two-sided Mann-Whitney or Kruskal-Wallis tests were used for 2 or 3 category comparisons, respectively.

We then employed a random forest algorithm to identify the parameters that can best predict MFE, limiting our analysis to sites that present in the crystal structure and where mutations were observed in at least 2 replicates to improve accuracy (total of 9685 mutations). Overall, a model trained on 70% of the dataset was able to predict the remaining 30% of the data (2905 mutations) with high accuracy (Spearman's $\rho > 0.75$, Pearson's $r = 0.76$; $p < 10^{-16}$; Figure S4 A,B). Surprisingly, a random forest model trained on the top five predictors alone showed similar accuracy (Spearman's $\rho = 0.73$, Pearson's $r = 0.73$; $p < 10^{-16}$; Figure ChVI-3). Excluding natural sequence variation, amino acid identity, or structural attributes reduced model predictability significantly (>20%; data not shown), suggesting a combination of evolutionary, sequence, and structural information best explains MFE. Using an alternative approach, we were able to predict the data with slightly lower accuracy using a linear model with the same five predictors ($p < 10^{-16}$, Spearman's $\rho = 0.67$, Pearson's $r = 0.67$; Figure S4C). Together, these results suggest that the prediction of MFE in the CVB3 capsid can be achieved at relatively high accuracy based on available structural and sequence information. Due to the high conservation of capsid structure in picornaviruses, as well as the availability of numerous capsid sequences and structures, these findings are likely generalizable to related picornaviruses.

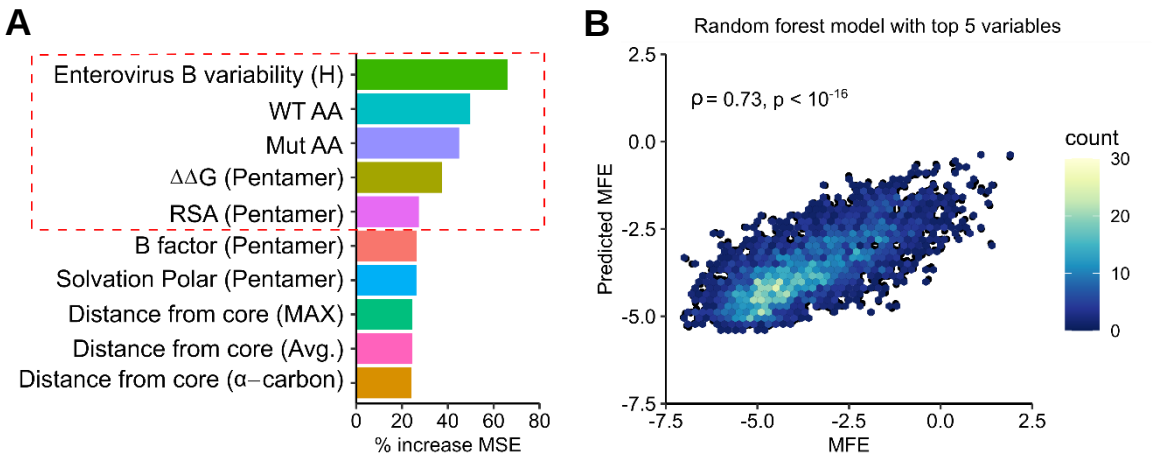


***Figure ChVI- 3* – Prediction of MFE based on structural and sequence information. A)** The top 10 predictors identified in a random forest model for existing MFE in the CVB3 capsid based on the percent of mean squared error (MSE) increase. **B)** Hexagonal plot showing the correlation between MFE predicted using a random forest algorithm trained on the top 5 predictors versus observed MFE. The random forest model was trained on 70% of the data, and then tested on the remaining 30% (shown).

d. Experimentally measured MFE inform of natural evolutionary processes

We next examined if our experimentally measured MFE could improve phylogenetic models of CVB3 evolution by incorporating site-specific amino acid preferences using PhyDMS (Hilton *et al.* 2017). Indeed, significant improvement in model fit was observed (Table ChVI-1 PHY; $p < 10^{-16}$ 148 using a log-likelihood test compared to non-site-specific codon models), supporting the relevance of our results to understanding evolutionary processes in nature. Nevertheless, selection in nature was significantly more stringent than in the lab ($\beta = 2.18$), indicating the presence of additional selection pressures.

***Table ChVI- 1* – Incorporation of DMS results in evolutionary models better describes natural CVB3 evolution compared to standard codon models.**

Model	∆AIC	LogLikelihood	Parameters	Parameter Values
ExpCM	0.00	-14580.51	6	Beta=2.18, kappa=7.47 omega=0.16
Goldman-Yang M5	4187.56	-16668.29	12	Alpha_omega=0.30, beta_omega=10.00, kappa=7.15
Averaged ExpCM	4303.74	-16732.38	6	Beta=0.61, kappa=7.55, omega=0.02
Goldman-Yang M0	4371.26	-16761.14	11	Kappa=7.14, omega=0.02

As laboratory conditions lack selection from antibodies, we used the sum of the absolute differential selection observed at each site (Bloom 2017) to examine whether known antibody neutralization sites show differential selection between the two environments (Supplementary Table S7). Indeed, antibody neutralization sites showed significantly higher differential selection values compared to other residues ($p < 10^{-6}$ by Mann-Whitney test; Figure ChVI-4A). Moreover, the three sites showing the strongest overall differential selection were found in known antibody neutralization sites: position 226 and 242 in the EF loop (residues 157 and 173 of VP2) and position 650 in the BC loop (residue 80 of VP1; Figure ChVI-4B-D and Supplementary Table S7). In summary, incorporation of our experimentally derived amino acid preferences into phylogenetic analyses significantly improved model fit and identified residues in antibody neutralization sites that show differential selection, suggesting these may play important roles in immune evasion *in vivo*.

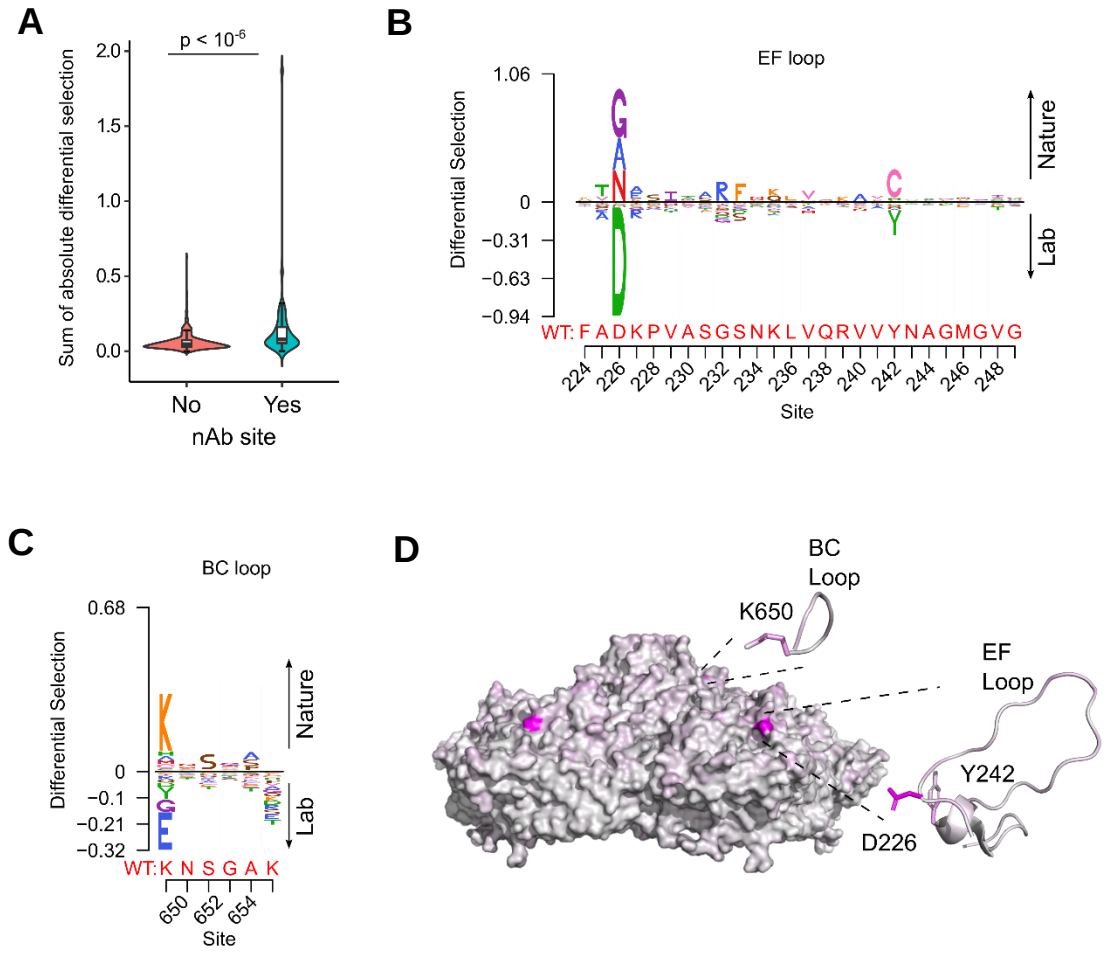


***Figure ChVI- 4* – Antibody neutralization sites show differential selection between laboratory and natural conditions. A)** Violin plot showing the sum of absolute selection observed at capsid sites comprising antibody neutralization sites (nAb) versus all other capsid sites. **B-C)** Logoplots showing the observed differential selection of sites in the EF loop or BC loop. The WT sequence is indicated in red. **D)** The CVB3 capsid pentamer (PDB:4GB3), colored according to the amount of differential selection. The BC and EF loops are shown next to the structure together with the sidechains for sites showing the highest differential selection.

e. Insights into capsid encoded motifs: Myristoylation and protease cleavage

Picornavirus capsids undergo a complex assembly path to generate the infectious particle. These include myristoylation, cleavage by the viral proteases 2A and 3CD^{pro}, as well as interaction with cellular chaperones and glutathione (Geller *et al.* 2007; Thibaut *et al.* 2014; Jiang *et al.* 2014; Qing *et al.* 2014; Corbic Ramljak *et al.* 2018) (Figure ChVI-5A). Having obtained a comprehensive dataset for MFE across the capsid, we next examined the sequence requirements for several of these capsid encoded motifs. Specifically, myristoylation of the N-terminal glycine is essential for virion assembly (Corbic Ramljak *et al.* 2018). In agreement with this, the N-terminal glycine in the CVB3 capsid showed the strongest average fitness cost upon mutation in the capsid (Figure S5 and Supplementary Table S4). The remaining sites in the myristoylation motif agreed with the canonical myristoylation motif in cellular proteins (Prosite pattern PDOC00008)

(Bologna *et al.* 2004), albeit with increased selectivity at three of the six positions (Figure S5A). On the other hand, a conserved WCPRP motif in the C-terminal region of VP1 that was shown to be important for $3CD^{pro}$ cleavage of the related foot and mouth disease virus capsid (FDMV; YCPRP motif) (Kristensen and Belsham 2019) was found to be intolerant to mutations compared to other capsid residues ($p < 0.05$ versus all other positions by Mann-Whitney test; sites 815-819 in CVB3). Moreover, within this motif, the sites showing the highest average fitness cost in our DMS dataset were identical to analogous positions in FMDV that resulted in a loss of viability upon mutation to alanine (Figure S5B) (Kristensen and Belsham 2019), highlighting the conservation of this motif across different picornaviruses.

The viral 3C protease ($3C^{pro}$) cleaves the picornavirus capsid at two conserved glutamine-glycine (QG) pairs to liberate the viral capsid proteins VP0, VP3, and VP1 (Figure ChVI-5A). Previous work has defined the sequence specificity of several picornavirus 3Cpro enzymes by examining both natural sequence variation and *in vitro* cleavage assays using synthetic peptides (Laitinen *et al.* 2016). However, unlike other $3C^{pro}$ mediated cleavage events in the viral polyprotein, the capsid is only efficiently cleaved by the precursor protein $3CD^{pro}$ (Ypma-Wong *et al.* 1988). To gain insights into the sequence specificity of $3CD^{pro}$, we examined the amino acid preferences for a 10 amino acid region surrounding the protease cleavage site (P5-P5'). As expected based on the known specificity of the 3C protease (Laitinen *et al.* 2016), a strong preference for the presence of QG was observed at both $3CD^{pro}$ cleavage sites in our dataset (positions P1 and P1' in the cleavage site; Figure ChVI-5B, C). Interestingly, significant correlation in amino acid preferences between the two cleavage sites was observed only at P1-P1' (Pearson's $\rho > 0.99$, $p < 10^{-16}$) and P4 (Pearson's $\rho > 0.49$, $p < 0.05$), as was the case in the enterovirus B alignments (Pearson's $\rho > 0.84$ and $p < 10^{-6}$ for positions P4, P1, and P'1; data not shown). Hence, the low agreement in amino acid preferences observed for most positions across the two $3CD^{pro}$ cleavage sites suggests cleavage is strongly dictated by positions P4, P1, and P1'.

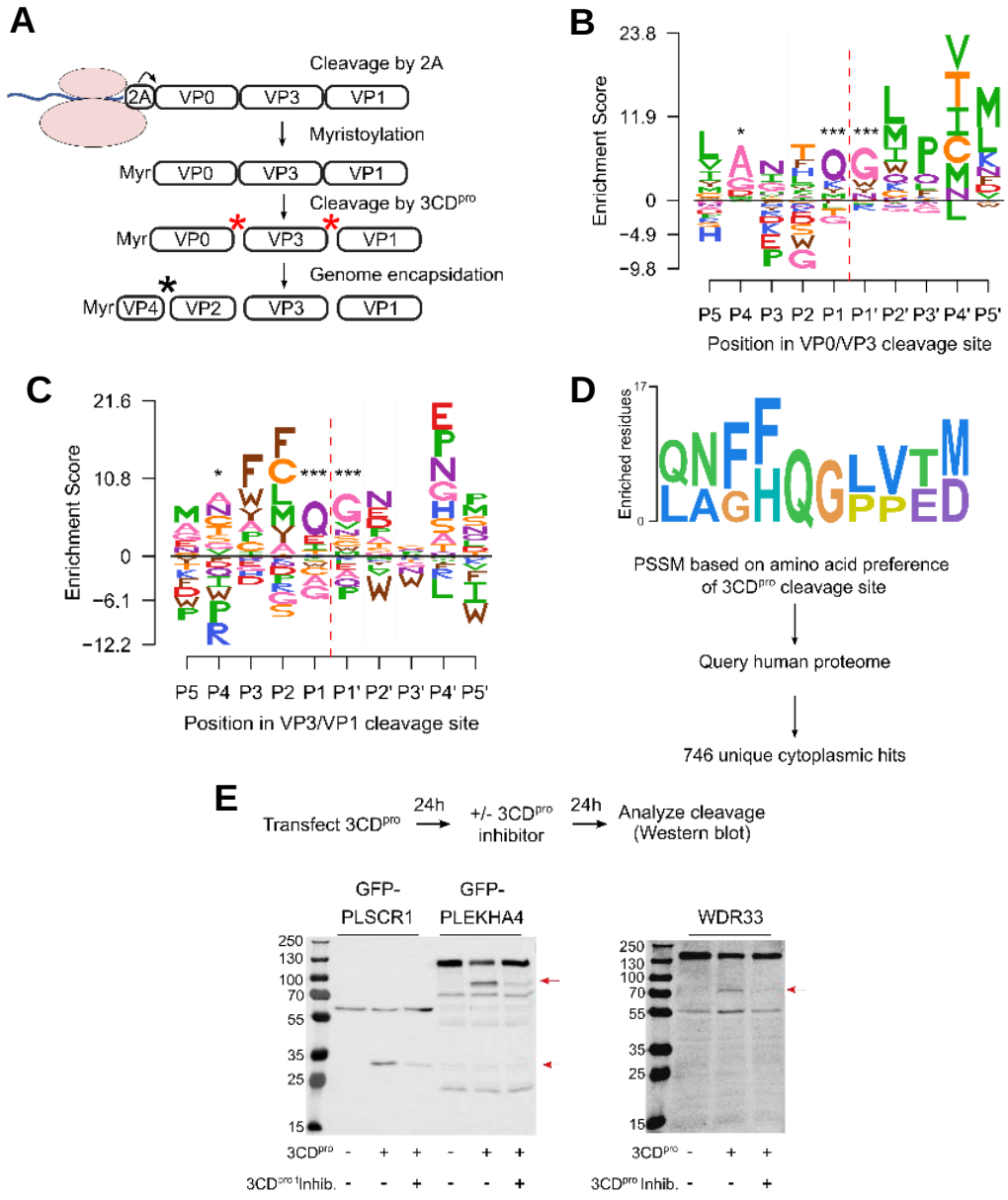


***Figure ChVI- 5* – Sequence preference of the capsid 3CD^{pro} cleavage sites and their use for the identification of novel cellular targets of the viral protease. A)** Overview of the CVB3 capsid maturation pathway. The CVB3 capsid precursor P1 is co-translationally cleaved by the viral 2A protease. P1 is then myristoylated and cleaved by the viral 3CD^{pro} to generate the capsid proteins VP0, VP3, and VP1. Finally, upon assembly and genome encapsidation, VP0 further cleaved into VP4 and VP2 in the protease independent manner to generate the mature capsid. **B-C)** Logoplots showing amino acid preferences for the 10 amino acid region spanning the 3CD^{pro} cleavage sites (P1-P'1) of both VP0/VP3 and VP3/VP1 in the DMS dataset. **D)** Overview of the bioinformatics pipeline for identification of novel 3CD^{pro} cellular targets using the amino acid preferences for the capsid cleavage sites from our DMS study. A position-specific scoring matrix (PSSM) was generated based on the amino acid preferences for the 10 amino acid region spanning the two 3CD^{pro} cleavages sites. This PSSM was then used to query the human genome for potential cellular targets, and non-cytoplasmic proteins were filtered out, yielding 746 proteins. **E)** The cellular proteins PLSCR1, PLEKHA4, and WDR33 are cleaved by 3CD^{pro}. Western blot analysis of cells cotransfected with 3CD^{pro} and GFP-PLSCR1 or GFP-PLEKHA4 and probed with a GFP antibody or transfected with 3CD^{pro} and probed using a WDR33 antibody. When indicated, the 3CD^{pro} inhibitor rupintrivir was included to ensure cleavage was meditated by the viral protease. Red arrows indicate cleavage products of the expected size (GFP-PLSCR1 full length =64kDa, cleaved N-terminus =36kDa; GFP-PLEKHA4 full length= 118kDa, cleaved N-terminus =72kDa; WDR33 full length =146kDA, cleaved N-terminus =72kDa). *p<0.05, *** p<0.001.

f. Identification of 3CD^{pro} cellular targets based on the sequence preferences of the capsid encoded protease cleavage site

In addition to cleaving the viral polyprotein, the picornavirus proteases cleave cellular factors to facilitate viral replication, including both antiviral factors and cellular factors that favor viral IRES-driven translation mechanism over cellular cap-dependent translation (e.g. DDX58, eIF4G, and PABP) (Sun *et al.* 2016; Laitinen *et al.* 2016). As the canonical 3C/3CD^{pro} QG cleavage site occurs on average 1.6 times per protein in the human proteome (~33000 times), we sought to examine if the rich dataset we obtained for the amino acid preferences of the capsid 3CD^{pro} cleavage sites can be used to identify novel cellular factors that are targeted by the viral protease. Specifically, a position-specific score matrix (PSSM) was generated for the 10 amino acid region spanning the two protease cleavage sites in the CVB3 capsid (P5-P5') based on the amino acid preferences identified in our study (Figure ChVI-5D). This PSSM was then used to query the human proteome for potential cleavage sites, yielding a total of 746 cytoplasmic proteins (Figure ChVI-5D; Supplementary Table S8). Eleven cellular factors that are known to be cleaved during enterovirus infection were identified using this approach, including the viral sensor Probable ATP-dependent RNA helicase DDX58 (RIG1), the immune transcription factors p65 (RELA) and interferon regulatory factor 7 (IRF7), and polyadenylate-binding protein 1 (PABPC1), an important factor in translation initiation and mRNA stability (Supplementary Table S8) (Laitinen *et al.* 2016; Jagdeo *et al.* 2018).

To evaluate whether our approach can identify novel cellular targets for the viral protease, we examined the ability of 3CD^{pro} to cleave eight different proteins found in the data set, focusing on those with cellular functions of potential relevance to CVB3 biology and which could be readily detected in our cell culture assay (e.g. availability of antibodies or tagged-variants, cleavage fragments of observable size, and high expression level). These included four interferon-inducible proteins (Pleckstrin Homology Domain Containing A4, PLEKHA4; Phospholipid Scramblase 1, PLSCR1; NOD-like receptor family CARD domain containing 5, NLRC5; Zinc Finger CCCH-Type Containing, Antiviral 1, ZC3HAV1) and four proteins involved in various cellular functions, namely apoptosis (MAGE Family Member D1, MAGED1), RNA processing (WD repeat domain 33, WDR33), and vesicle transport (Cyclin G Associated Kinase, GAK; Tumor Susceptibility 101, TSG101). Of these, three proteins were cleaved upon expression of the viral protease to generate fragments of the expected size (PLSCR1, PLEKHA4, and WDR33; Figure ChVI-5 E and Supplementary Table S8). Of note, while WDR33 was predicted to harbor two potential cleavage sites, only a single cleavage event was observed. Treatment with a specific 3CD^{pro} inhibitor, rupintrivir (Dragovich *et al.* 2003),

blocked the cleavage of these proteins, indicating the effect was due to the viral protease (Figure ChVI-5 D). In contrast, five of the proteins were found to not be cleaved upon $3CD^{pro}$ expression, suggesting additional determinants are involved in the cleavage of host factors (Figure S6). Hence, our approach correctly identified 30% of the predicted cleavage sites (3 of the 9 different cleavage sites), indicating a strong enrichment of cellular targets of the $3CD^{pro}$ in the dataset.

5. Discussion

The picornavirus capsid is a highly complex structure that plays key roles in viral biology and pathogenesis. In the current study, we employ a comprehensive approach to define the effects of single amino acid mutations in the CVB3 capsid, measuring the effects of >90% of all possible mutations. We find that most mutations in the capsid are deleterious, with very few mutations showing higher fitness than the WT sequence (1.2% of all mutations). Similar results have been reported in other non-enveloped capsid proteins (Acevedo *et al.* 2014; Hartman *et al.* 2018; Ogden *et al.* 2019) as well as non-capsid viral proteins (Bloom 2014; Thyagarajan and Bloom 2014; Wu *et al.* 2015; Doud and Bloom 2016; Du *et al.* 2016; Haddox *et al.* 2016; Ashenberg *et al.* 2017; Hom *et al.* 2019). In light of these results, it is likely that the large population sizes of RNA viruses help maintain viral fitness in the face of high mutation rates and strong mutational fitness costs.

Investigation of the factors that influence MFE in the capsid revealed a strong correlation with various structural and functional attributes. These included computationally predicted effects on stability and aggregation propensity, secondary structure, and surface exposure (Figure ChVI-2). Surprisingly, we find that MFE can be predicted with relatively high accuracy using only five parameters: natural sequence variation, the identity of the original and mutant amino acid, the predicted effect on protein stability, and relative solvent accessibility (Figure ChVI-3). A recent study examined the ability of 46 different variant effect prediction tools to predict MFE from 31 different DMS datasets of both viral and non-viral proteins (Livesey and Marsh 2020). Overall, viral proteins showed the lowest predictability (Spearman's correlation of <0.5). In contrast, we were able to predict MFE using a random forest model using these above-mentioned five parameters with an accuracy similar to the best prediction obtained in this analysis for any viral or non-viral protein (Pearson's r= 0.73; Spearman's ρ = 0.73; Figure ChVI-3B). Interestingly, SNAP2 (Hecht *et al.* 2015), a neural network-based classifier of mutational effects that was shown to correlate well with MFE in other studies (Gray *et al.* 2017; Reeb *et al.* 2020; Livesey and Marsh 2020), correlated poorly with our data (R^2 = -0.26). Overall, considering the relative conservation of capsid structure in picornaviruses as well as the availability of both capsid sequences and high-resolution structures for numerous members of this family, it is likely that these findings can be extrapolated to additional picornaviruses.

Incorporating site-specific amino acid preferences obtained from our DMS results into phylogenetic models was found to significantly improve model accuracy. This has been observed in DMS studies with other RNA viruses (Doud and Bloom 2016; Bloom 2017;

Haddox *et al.* 2018) and indicates that our laboratory-measured MFE captures additional information that cannot be obtained from sequence analysis alone. In addition, this approach allowed us to assess which sites show differential selection patterns as a result of the distinct environments encountered in nature and the laboratory. As expected, pressure from the adaptive immune system was found to be the major difference between these environments, with residues in antibody neutralization sites showing higher differential selection compared to other sites in the capsid (Figure ChVI-4A). Moreover, the sites showing the highest degree of differential selection were found in known antibody neutralization sites (Figure ChVI-4B-D). However, why these particular residues within antibody neutralization sites show differential selection, while others do not, remain to be elucidated. It has been shown that one, or a few, sites within antibody binding regions can have strong effects on escape from antibody neutralization (Lee *et al.* 2019), potentially explaining these findings. Interestingly, while the top three sites showing differential selection were in antibody neutralization sites, the mutation showing the fourth-highest differential selection was found in the HI loop of VP1. While not classically considered an antibody epitope, this loop has been shown to interact with an antibody fragment in the picornavirus coxsackievirus A6 (Xu *et al.* 2017), is known to mediate receptor binding in different picornaviruses (Belnap *et al.* 2000; Xing *et al.* 2000), and to interact with host cyclophilin A to facilitate uncoating (Qing *et al.* 2014). Whether these factors or others are responsible for the observed differential selection remains to be elucidated.

The CVB3 capsid encodes the information for directing myristoylation, protease cleavage, and interaction with host factors. We took advantage of our data to examine the sequence specificity and mutational tolerance of several known capsid encoded motifs. First, we examined the amino acid preferences of the CVB3 capsid myristoylation motif. We observe a strong correlation with the canonical myristoylation pattern (Prosite pattern PDOC00008), although with greater intolerance to mutations in three of the six residues in the capsid (Figure S5). This is likely to stem from additional constraints imposed by the capsid structure. On the other hand, we examined the amino acid preference of a conserved motif in VP1 that is required for $3CD^{pro}$-mediated cleavage of picornavirus capsids (Kristensen and Belsham 2019). Our data showed a higher cost to mutation in this motif relative to other capsid positions (Figure S5), highlighting its importance for capsid function. Finally, we examined the sequence preferences surrounding the two $3CD^{pro}$ cleavage sites. We find a strong dependence on the cleavage site residues (positions P1 and P1'; Figure ChVI-5) and to a lesser degree position P4, with large variation in the sequence preferences across the remaining positions between

the two cleavage sites. Overall, our experimentally measured MFE are congruent with existing information regarding the sequence preferences of the examined capsid motifs, yet provide in-depth insights into sequence specificity that cannot be obtained from examining natural sequence variation.

Finally, we used the amino acid preferences observed in $3CD^{pro}$ cleavage sites within the capsid to query the human genome for potential cellular targets of this protease (Figure ChVI-5D). Using this approach, we identify 746 cytoplasmic proteins that harbor a potential $3CD^{pro}$ target sequence, including 11 proteins previously shown to be cleaved by different picornavirus 3C proteases. We then validated our approach using eight proteins, comprising nine predicted cleavage sites. Six of the predicted cleavage sites were not affected by $3CD^{pro}$ expression (Figure S6). On the other hand, three proteins were observed to be specifically cleaved by the viral protease (Figure ChVI-5E): WD Repeat Domain 33 (WDR33), an important factor for polyadenylation of cellular pre-mRNAs (Chan *et al.* 2014) that has been shown to act as a restriction factor during influenza infection (Brass *et al.* 2009); the interferon-induced protein Phospholipid scramblase 1 (PLSCR1), which is involved in the replication of numerous viruses, likely due to its ability to enhance the expression of certain interferon-stimulated genes (Kodigepalli *et al.* 2015); and the interferon-induced Pleckstrin Homology Domain Containing A4 (PLEKHA4), a plasma membrane-localized signaling modulator (Shami Shah *et al.* 2019) that is currently not known to play a role in viral infection. Overall, our approach correctly predicts 30% of the identified cleavage sites. It is likely that incorporating additional selection criteria, such as accessibility of the cleavage peptide in the folded structure can be used to further reduce false positives. Nevertheless, extrapolating our validation results to the larger dataset suggests >200 new host targets of the protease are identified, many of which could play key roles in viral biology and pathogenesis.

6. Data availability and accession numbers

Unaligned bam files have been uploaded to SRA (Accession SAMN15437545-SAMN15437555; SRA 15437545-15437555). The scripts and data required to obtain the codon count tables for all samples, to perform the random forest and linear model predictions, to generate the peptides for use with PSSMsearch, as well as the sequence alignments and modified structure files for FoldX analysis can be found on Github (https://github.com/RGellerLab/CVB3_Capsid_DMS).

7. Supplementary data

Supplementary data are available online at bioRxive and also in the appendix of this manuscript. **Figure S1:** Sanger analysis of DMS libraries. **Figure S2:** Results of high-fidelity duplex sequencing. **Figure S3:** Correlation of amino acid preferences observed in experimental replicates. **Figure S4:** Prediction of mutational fitness effects using random forest or linear models. **Figure S5** Sequence preferences of capsid encoded motifs. **Figure S6:** Evaluation of select hits identified as potential $3CD^{pro}$ target proteins.

CHAPTER VII – Increased RNA virus population diversity improves adaptability.

Unpublished results.

1. Abstract

The replication machinery of most RNA viruses lacks proofreading mechanisms. As a result, RNA virus populations harbor a large amount of genetic diversity that confers them the ability to rapidly adapt to changes in their environment. In this work, we investigate whether further increasing the initial population diversity of a model RNA virus can improve adaptation to a single selection pressure. For this, we experimentally increased coxsackievirus B3 (CVB3) population diversity in the capsid region using deep mutational scanning and compared the ability of these high diversity CVB3 populations to achieve resistance to thermal inactivation relative to standard CVB3 populations. We find that high diversity viral populations are better able to achieve resistance to thermal inactivation at both the temperature employed during experimental evolution as well as at a more extreme temperature. Moreover, we identify mutations in the CVB3 capsid that confer resistance to thermal inactivation, finding significant mutational epistasis. Our results indicate that even naturally diverse RNA virus populations can benefit from experimental augmentation of population diversity for optimal adaptation and support the use of such viral populations in directed evolution efforts that aim to select for viruses with desired characteristics.

2. Introduction

RNA viruses are characterized by extreme mutation rates that are orders of magnitudes higher than those of most DNA based organisms (Sanjuán and Domingo-Calap 2016; Peck and Lauring 2018). Together with their short replication times and large population sizes, these high mutation rates confer RNA viruses an extreme capacity for rapid evolution. This, in turn, poses a significant challenge for treating and preventing infections by RNA viruses as it allows for subverting the immune system, gaining resistance to antiviral drugs, and jumping to new hosts. On the other hand, the capacity of RNA viruses to rapidly adapt to new environments can be capitalized upon to select for viruses with desired characteristics. In such directed evolution experiments, virus populations are grown under conditions that favor either the emergence or further optimization of the desired phenotype. For example, this process has been used to obtain live attenuated vaccines for numerous viruses by repeated growth at suboptimal conditions (Minor 2015), improving *in vivo* models by serial infection of the desired host (Roberts *et al.* 2007; Ilyushina *et al.* 2010; Sutton and Subbarao 2015; Li *et al.* 2017; C Zhang *et al.* 2017; Gorman *et al.* 2018), isolating RNA viruses with high-fidelity polymerases by growth under condition of increased mutational load (Bordería *et al.* 2016; Kautz and Forrester 2018), selection of viruses with improved oncolytic properties (Sanjuán and Grdzelishvili 2015; Svyatchenko *et al.* 2017; Seegers *et al.* 2019), or isolation of viruses and virus-like particles of increased stability (Shiomi *et al.* 2004; Adeyemi *et al.* 2017; Nguyen *et al.* 2018). To date, these studies have relied upon the natural mutational processes of RNA viruses to generate the diversity upon which selection is applied.

As a consequence of degeneracies in the genetic code, single mutations within a codon can on average reach 5.8 (CI_{95}: 5.61–6.00) other amino acids while double and triple mutations are required to reach the remaining 9.61 (CI_{95}: 9.28–9.93) and 3.59 (CI_{95}: 3.24–3.94) amino acids, respectively (Table S1). While RNA virus mutation rates are high, the probability that multiple mutations occur in the same codon remains low. For example, for a virus with a protein-coding region of 6.5 kb (e.g. the picornavirus coxsackievirus B3) and a mutation rate of $1x10^{-4}$ (Gnädig *et al.* 2012; Graci *et al.* 2012), the probability that two or three mutations occur in the same codon is $6.8x10^{-5}$ and $6.8x10^{-9}$, respectively (Binomial distribution). Hence, extremely large population sizes are required to sample the full spectrum of amino acid mutations across the viral protein-coding region. Moreover, inherent preferences in base misincorporation by the polymerase, such as the high transition to transversion bias frequently observed for RNA viruses (Acevedo *et al.* 2014; Geller *et al.* 2016), can further reduce the probability of

certain mutations from occurring. While the sequential acquisition of mutations over several replication cycles can potentially allow for multiple mutations within a codon to accumulate, selection against intermediate genotypes can limit such evolutionary trajectories. The viral genotype can also influence the potential for sampling particular non-synonymous changes during replication as synonymous codon choice can dictate the likelihood of reaching particular non-synonymous changes (Maeshiro and Kimura 1998) and dictate evolutionary trajectories (Cambray and Mazel 2008; Lauring *et al.* 2012). In sum, despite the high mutation rate of RNA viruses, the ability of RNA virus populations to reach particular non-synonymous mutations that may be beneficial for adaptation can require extremely large population sizes and can depend on the initial viral genotype. This raises the question of whether directed evolution experiments can benefit from the use of viral populations with experimentally increased population diversity rather than solely relying on the spontaneous emergence of mutations during virus replication.

In this work, we apply a codon-level mutagenesis protocol to the entire capsid region of the human picornavirus coxsackievirus B3 (CVB3) and generate viral populations with increased diversity across the capsid protein. Using these CVB3 populations, we examine if RNA viruses can further benefit from an initial increase in diversity to adapt to a single selection pressure, capsid thermal inactivation. We find that in an experimental evolution setting, high diversity CVB3 populations achieve greater adaptation, with a significantly improved ability to resist thermal inactivation compared to standard populations. Additionally, we identify several mutations in the CVB3 capsid that allow for resistance to thermal inactivation and find that mutational epistasis plays an important role in adaptation. Overall, our results indicate that even RNA viruses with extreme mutation rates can further benefit from augmentation of population diversity for adaptation, and support the use of viral population with experimentally expanded population diversity in directed evolution experiments.

3. Materials and Methods

a. Viruses, cells, and plaque assays

HeLa-H1 (CRL-1958) and HEK293 (CRL-1573) cells were obtained from ATCC. CVB3 was generated from the Nancy infectious clone (kind gift of Dr. Marco Vignuzzi, Institute Pasteur). For this work, the infectious clone was modified by site-directed mutagenesis to remove an XhoI site present in the capsid region (P1) and introduce an XhoI site at position 692 as well as a Kpn2I site at position 3314, generating a pCVB3-Xho-P1-Kpn2I clone. All mutations in protein-coding regions were synonymous. Cells were cultured in culture media (DMEM with 10% heat-inactivated FBS, Pen-Strep, and L-Glutamine). For infections, FBS concentrations were reduced to 2%. For plaque assays, serial dilutions of the virus were used to infect confluent HeLa-H1 cells in 6 well plates for 45 minutes, followed by overlaying the cells with a 1:1 mixture of 56°C 1.6% Agar (Arcos Organics 443570010) and 37°C 2x DMEM with 4% FBS. Two days later, a 10% formaldehyde solution was added to reach a final concentration of 2% to fix the cells and inactivate the virus, followed by staining with crystal violet and counting of plaques.

b. Codon level mutagenesis protocol

Mutagenesis was performed using degenerate primers as previously described (Mattenberger *et al.* 2020) with the exception that two rounds of mutagenesis were performed using 10 cycles and then 7 cycles for all samples. The products were gel purified and ligated to an XhoI and Kpn2I digested and gel purified pCVB3-Xho-P1-Kpn2I (Mattenberger *et al.* 2020) using NEBuilder® HiFi DNA Assembly reaction (NEB) for 60 minutes. Mutagenesis efficiency was evaluated by the transformation of the assembled plasmids into NZY5α competent cells (NZY Tech), Sanger sequencing of several clones and analysis using the Sanger Mutant Library Analysis script (https://github.com/jbloomlab/SangerMutantLibraryAnalysis). Subsequently, the assembled plasmid reactions were purified using a Zymo DNA Clean & Concentrator-5 kit (Zymo Research) and used to electroporate MegaX DH10B $T1^R$ Electrocomp™ cells (ThermoFisher) using a Gene Pulser XCell™ electroporator (BioRad) according to the manufacturer's protocol. Cells were then grown overnight in a 10mL liquid culture at 33°C. Transformation efficiency was estimated by plating a small amount of the transformation on agar plates. In total, $4.75x10^5$, $1.4x10^5$, and $7.1x10^5$ transformants were obtained for lines 1, 2, and 3, respectively. Of these, 75% contained full-length inserts as judged by colony PCR of 52 colonies using external primers 659F (TTGGATTGGCCATCCGGT) and 3450R (GTGCTGTGGTCGTGCTCACTAA).

Subsequently, plasmid DNA was isolated in duplicate using a miniprep kit (Macherey-Nagel NucleoSpin Plasmid).

c. Analysis of mutagenized libraries

To analyze library diversity, the mutagenized region was amplified from 10ng of each library or the unmutagenized control plasmid by performing 25 PCR cycles using Phusion™ polymerase and the HiFi-F (CTTTGTTGGGTTTATACCACTTAGCTCGAGAGAGG) and HiFi-R (CCTGTAGTTCCCCACATACACTGCTCCG) primers, followed by gel purification of the correct band (Zymoclean™ Gel DNA Recovery Kit). Libraries were then prepared following published protocols (Kennedy *et al.* 2014) and each library was run on a NovaSeq6000 2x150 at a maximum of 30G per lane to reduce potential index hopping. Analysis of mutations was performed as previously published (Mattenberger *et al.* 2020).

d. Production of WT and High Diversity viral populations

To produce viral populations, viral genomic RNA was transcribed from SalI linearized plasmids (TranscriptAid T7, ThermoScientific) and 10µg were electroporated into $4x10^6$ cells in a 4mm cuvette in 400µL of calcium and magnesium-free PBS using a Gene Pulser XCell™ electroporator (BioRad) set to 240V and 950uF. Two electroporations were performed for each mutagenized library. Cells were then placed in culture media for 9 hours to produce the passage 0 virus (P0). Following two freeze-thaw cycles, $2x10^6$ plaque-forming units (PFU) were used to infect a 90% confluent 15cm plate in 3mls of infection media for 45 minutes. Cells were then washed with PBS and incubated in 13mL of infection media for 8 hours. Finally, cells were subjected to 3 freeze-thaw cycles, debris removed by centrifugation at 500xg and the supernatants collected to generate P1 virus stocks. All infections produced > $3.75x10^7$ PFU in P0 and > $2x10^8$ PFU in P1 as judged by plaque assay. Wildtype virus was similarly produced by electroporation of viral vRNA into cells. Viral infections were allowed to continue until CPE was observed to generate the P0 population. The cultures were then subjected to two freeze-thaw cycles, followed by titration. Two additional passages at low MOI were performed to enable the natural accumulation of diversity generated by the viral polymerase and avoid any biases that may be introduced by the T7 RNA polymerase used in the *in vitro* transcription of viral RNA.

e. Experimental evolution for thermal resistance

Viral populations were adjusted to ~10^7 PFU/mL and two aliquots of 100µl were subjected to the indicated temperature for 30 minutes in a 0.2mL PCR tube using a

thermocycler. Both an unheated virus population and one of the heated aliquots were titered using a plaque assay to assess the degree of infectivity loss. The remaining aliquot was used to infect cells in a 6 well plate until cytopathic effect (CPE) was reached in most wells to ensure a similar number of replication cycles between different conditions. The emerging virus was titered and used for the next round of inactivation. To control for mutations conferring adaptation to cells, a WT virus population was blindly passaged for 10 passages in a single well of a 6 well plate. For this, viral titers of 10^8 PFU/mL were assumed upon CPE and an estimated 1,000 PFU from each infection was used to initiate the subsequent passage in order to mimic the amount of virus surviving heat treatment. Sanger sequencing was performed for all populations following 10 passages by extracting RNA from 100µL of virus supernatant (Zymo Quick-RNA™ Viral kit), generating cDNA using a gene-specific primer (TCTCTTGGACCTCTACTA) with M-MLV reverse transcriptase (NZY Tech), amplifying the P1 region with Phusion™ polymerase and primers 659F and 3548R, and sequencing the purified PCR product using primers HiFi-F, 2045F (TCGAGTGTTTTTAGTCGGACG), 2143R (GGCCGAACCACAGAACATAA) and 3450R (GTGCTGTGGTCGTGCTCACTAA). Sequences were analyzed using the Staden 2.0.0 package.

f. Evaluation for thermal resistance

To evaluate the thermal resistance of the virus populations following 10 passages, we performed a heat inactivation experiment at 45°C and 47°C. For this, we adjusted the viral populations to 10^6 PFU/ml and took 100µL in 0.2mL PCR tubes. The virus populations were subjected to heat inactivation at the indicated temperature for 30 minutes in a thermocycler. Both the starting virus inoculum and the surviving virus titer were obtained by plaque assay, and the fraction of the surviving virus was calculated. Experiments were performed in quadruplicate for each population and the input populations were titered twice to ensure no major bias in the initial amount of virus.

g. Generation and evaluation of CVB3 capsid mutants

The PCR of the capsid region used as a template for mutagenesis was phosphorylated and cloned into a SmaI digested pUC19 vector for use in the mutagenesis reactions. For each mutant, non-overlapping primers were designed that incorporated the mutation in the middle of the forward primer. PCR was performed using Phusion™ polymerase and mutagenic primers, followed by DpnI treatment, phosphorylation, ligation, and transformation of chemically competent bacteria. Successful mutagenesis was verified by Sanger sequencing. Subsequently, the capsid region was subcloned into the infectious clone using XhoI and Kpn2I sites. Plasmids

were then linearized with MluI and 1µg of plasmid was transfected into $2x10^6$ HEK293 cells together with a plasmid encoding the T7 polymerase (Yun *et al.* 2015) (Addgene 65974). Viruses were titered by plaque assay and tested for thermal resistance as indicated above. For mutants that did not yield any virus, the transfection was repeated 2 more times to ensure the lethality of the mutant. Emerging viral populations were sequenced to ensure no compensatory mutations or reversions arose during replication. For assessment of viral fitness, confluent HeLa-H1 cells were infected in triplicate in a single well of a 6 well plate with $3x10^6$ PFU for 45 minutes. Subsequently, the inoculum was removed, 2 mL of infection media was added and the cells were incubated for 9 hours. The cells were then subjected to two freeze-thaw cycles and the amount of virus produced was titered using a plaque assay.

h. Bioinformatics and statistical analyses

The crystal structure PDB:4gb3 was used to obtain secondary structure assignment using DSSP and to identify residues present in the core, interface, or surface via the VIPERdb (Carrillo-Tripp *et al.* 2009) (http://viperdb.scripps.edu/). Missing residues in the structure were assumed to be flexible loops. Amino acid variability was assessed using Shannon entropy. Briefly, all available, non-identical, full-genome CVB sequences from Virus Pathogen Resource (Pickett *et al.* 2012) (www.viprbrc.org) were downloaded (available at https://github.com/RGellerLab/dms_thermal_selection) and codon-aligned using the DECIPHER package in R. All alignment positions not present in our reference strain were removed, and a custom R script was used to calculate Shannon entropy. All statistical tests were performed in R (version 4.0.1), using a two-tail t-test on log-transformed data, in which case a two-sample Mann–Whitney U test was performed using the wilcox.test function.

4. Results

a. Generation of CVB3 populations with increased diversity across the capsid

Our aim was to examine how initial population diversity influences adaptation to a single selection pressure, comparing natural CVB3 populations with those harboring experimentally increased diversity. We chose to increase diversity across the 851 amino acid capsid region (P1) of CVB3 as it is not involved in viral genome replication and therefore will not introduce any bias in the capacity of the virus to evolve. For this, a PCR-based method for introducing mutations at the codon level (Bloom 2014; Mattenberger *et al.* 2020) was performed in triplicate to produce three independent libraries of the CVB3 infectious clone harboring increased diversity in the capsid region (see Methods; Figure ChVII-1A). Sanger sequencing of 40 clones from the three libraries (34,040 codons; range 7,659-15,318 per library) indicated an average mutation rate of 1.1 codon mutations per clone (range 1.06-1.23; Table S2) while sequencing of four clones from a control, non-mutagenized library revealed no mutations ($p < 0.05$ by Fisher's exact test). The majority of clones showed either no mutation (30%) or a single codon mutation (45%), while only 25% of clones had >1 mutation (Figure S1A), and the number of mutations per codon showed a nearly even distribution (Figure S1B). We next employed a high-fidelity next-generation sequencing technique to better define library diversity (Schmitt *et al.* 2012). A high rate of background mutations was observed for single mutations within codons compared to a non-mutagenized library (WT Lib; Figure ChVII-1B; Table S3), while two or three mutations per codon showed >400 fold higher rates in the mutagenized libraries compared to the WT library (Figure ChVII-1B; Table S3). Hence, we chose to exclude single mutations per codon from our dataset as these could not be readily distinguished from background errors. Analyzing only double and triple mutations per codon, we observe an average of 0.9 codon mutations per capsid region and a total of 92% of all possible single amino acid mutations represented in all three libraries (14,855 of 16,169 possible mutations; Figure ChVII-1C). Hence, our mutagenized libraries capture the vast majority of possible single amino acid mutations in the capsid region.

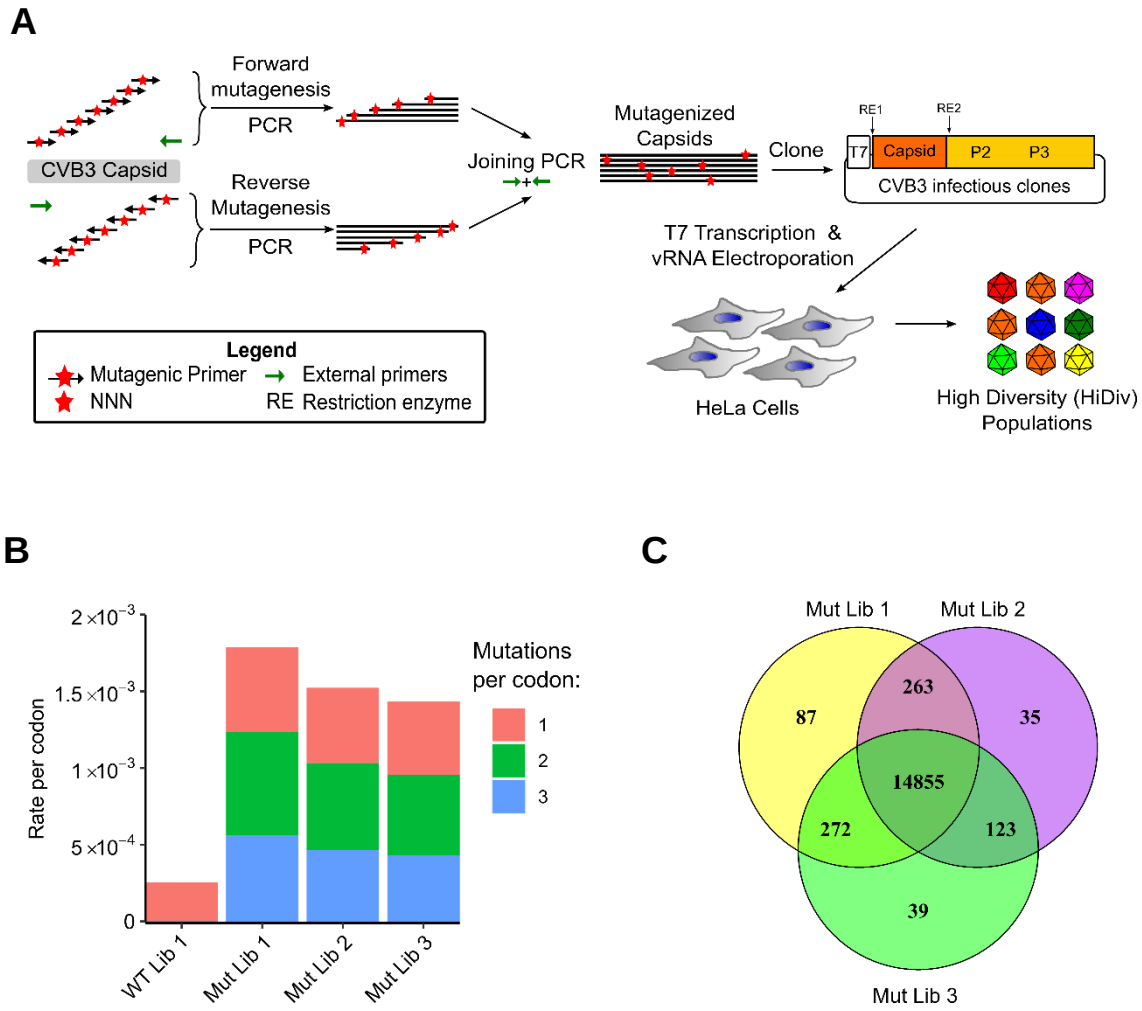


Figure ChVII-1 **Deep mutational scanning (DMS) of the CVB3 capsid**. **A)** Schematic representation of the mutagenesis protocol. A forward mutagenesis PCR reaction was performed using a single external reverse primer and a pool of forward mutagenic primers targeting each codon in the capsid, each primer encoding degenerate nucleotides (NNN) at the codon matching position. Similarly, a reverse mutagenesis PCR reaction was performed using a mix of reverse mutagenic primers targeting all codon sites in the capsid and a single external forward primer. The product of these PCRs was joined using the external primers and cloned into the CVB3 infectious clone to generate the mutagenized libraries. Viral genomic RNA (vRNA) was produced via in vitro transcription and electroporated into cells to generate high diversity CVB3 populations (HiDiv). **B)** The mutation rates for single, double, and triple mutations within codons observed in the control, unmutagenized library (WT Lib 1) or the mutagenized libraries (Mut Lib 1-3). **C)** Venn diagram showing the number of amino acid mutations observed in the mutagenized libraries.

b. Selection for viral populations with increased resistance to thermal inactivation

The mutagenized libraries were then used to produce high diversity (HiDiv) viral populations by electroporation of *in vitro* transcribed viral genomic RNA into HeLa-H1 cells (vRNA, HiDiv populations; Figure ChVII-1A). Infection was stopped following nine hours to allow for only a single infection cycle. As the initial multiplicity of infection was unknown, a second round of infection was performed to link each capsid to the genome it encapsidates. A detailed analysis of the mutational diversity found in the similarly

produced mutagenized populations following growth in cells is outside the scope of the current work and has been submitted elsewhere (Mattenberger *et al.* 2020).

To compare the effect of increased initial population diversity on adaptation, the three HiDiv and three WT viral populations were then subjected to an experimental evolution regimen to select for thermal resistance. For each passage, approximately one million plaque-forming units (PFU) were heated for 30 minutes and the surviving viruses were used to inoculate HeLa cells. Infections were stopped once significant cytopathic effect was observed in most conditions in order to minimize differences in the number of replication cycles between the different conditions. This protocol was repeated for a total of 10 passages (Figure ChVII-2A). An initial inactivation temperature of 43°C was chosen to minimize bottlenecking, followed by a passage at 44°C, and then eight passages at 45°C. At each step, the titers of the initial inoculum, the surviving population, and the amplified population were determined and used to calculate the fraction of surviving viruses (Figure ChVII-2A and Table S4). Overall, no major bottlenecks were observed for any of the passages except for passage 4 of the WT line 3 virus population, where only 5 PFU survived the inactivation (Table S4).

To examine whether viral populations with increased initial population diversity were better able to gain resistance to thermal inactivation, we compared the ability of the HiDiv and WT virus populations to survive heat treatment following 10 rounds of adaptation. Specifically, we examined the ability of the populations to survive either the temperature used during the selection regimen (45°C) or an increased temperature (47°C), in order to assess the ability of the populations to confront a stronger selection pressure. At both temperatures, the HiDiv populations showed increased thermal resistance, yielding 3.7 and 260 times more viruses on average at 45°C and 47°C, respectively ($p < 0.05$ and $p < 0.001$ by a two-tailed t-test; Figure ChVII-2B and Table S5). Hence, the increased initial diversity of the HiDiv populations improved adaptation to the applied selection pressure, as well as the capacity to confront stronger selection pressures.

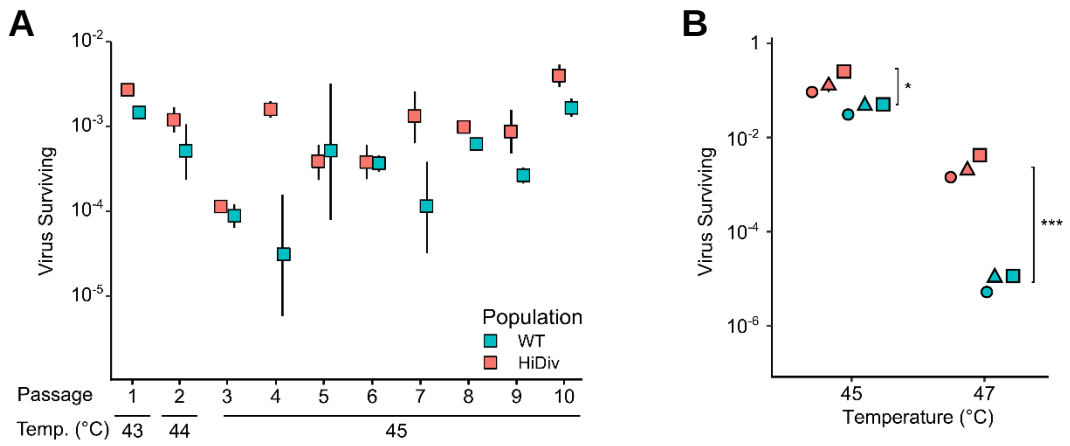


***Figure ChVII-2* Experimental evolution of resistance to thermal inactivation. A)** Experimental evolution of HiDiv and WT populations under conditions of thermal inactivation. HiDiv or WT populations were subjected to heat treatment at the indicated temperature for 30 minutes. The surviving fraction was then amplified and used for the subsequent passage. This process, representing a single passage, was repeated a total of 10 times at the indicated temperature. The mean fraction of the surviving virus and standard deviation for the three replicates are plotted. **B)** HiDiv populations show improved resistance to thermal inactivation compared to WT populations. The evolved passage 10 HiDiv and WT populations were tested for their ability to resist thermal inactivation following incubation at 45°C or 47°C for 30 minutes. Each population was tested in quadruplicate (Table S6) and the average and standard deviation of the fraction of surviving virus is indicated. Squares represent line 1, triangles line 2, and circles line 3. * $p<0.05$, *** $p<0.001$ by two-tailed t-test on log-transformed values.

c. Identification of novel mutations conferring thermal resistance

To identify mutations that confer resistance to thermal inactivation, the capsid sequence of all adapted populations were obtained by Sanger sequencing. Numerous mutations were observed in the populations subjected to thermal selection, none of which were observed in a mock selected WT population (Table ChVII-1 and Table S6). Overall, HiDiv populations had a larger number of mutations compared to WT populations ($p < 0.05$ by the Mann-Whitney test; Table ChVII-1 and Table S6). Mutations were found largely at variable positions in the capsid ($p < 0.005$ by Mann-Whitney test; Figure S2 and Table S7). In total, 9 mutations were observed in at least 2 independent lines, indicating these may contribute to thermal stability. Most of these mutations were in loops and subunit interfaces, as expected for capsid stabilizing mutations (Table S6 and Table S7). A previously described mutation conferring capsid stability, A512T, was observed in all 3 WT lines and 2 of 3 HiDiv lines (Carson *et al.* 2019). In addition, different mutation at position 581 was present in all evolved lines, indicating this position may be relevant for thermal stability.

Table ChVII-1 High frequency mutations observed in the evolved lines at the end of experimental evolution.

Viral Protein	Wildtype AA	CDS site	Mutant AA	Protein Position	Mock	WT			HiDiv		
					1[a]	1	2	3	1	2	3
VP0	G	51	S	51							x
	G	162	R/W	162						x	
	H	187	R	187						x	
	D	207	S	207					x	x	
VP3	N	395	H	63		x	x				
	A	512	T	180		x	x	x		x	x
	Q	566	L	234	x						
VP1[b]	**A**	**576**	**T**	**6**				x			
	A	**576**	**V**	**6**					x	x	
	I	**581**	**K**	**11**		x					
	I	**581**	**M**	**11**				x			
	I	**581**	**T**	**11**			x		x	x	x
	E	596	G	26					x	x	
	F	646	Y	76		x	x				
	I	711	V	141				x		x	
	Q	824	G	254				x			x
	K	827	R	257	x						
	K	829	Q	259					x		

[a]WT viral population passaged 10 times in HeLa-H1 cells without heat treatment.
[b]Bold font indicates a position where multiple non-synonymous mutations were observed.

Due to the higher thermal resistance of capsids from the HiDiv populations, we chose to experimentally test all mutants observed in more than one HiDiv line for their ability to confer thermal stability. However, we did not evaluate the effect of A512T, as this has already been demonstrated to enhance thermal stability (Carson *et al.* 2019). No single mutation was sufficient to mirror the phenotype of the evolved HiDiv population at 45°C, indicating optimal thermal resistance requires multiple synergistic mutations ($p<0.05$ by two-tailed t-test versus the HiDiv evolved population; Figure ChVII-3A and Table S8). At the higher temperature, the evolved HiDiv population showed increased thermal resistance, but this was not statistically significant versus D207S or the I581T mutations ($p>0.05$ by two-tailed t-test), and only marginally significant for the I581K mutation ($p=0.05$ by two-tailed t-test; Figure ChVII-3A). From the individual mutations analyzed, D207S showed the strongest effect on thermal inactivation, yielding 137 (SD 24) and 56 (SD 5) times more surviving PFU compared to the WT virus following incubation at 45°C and 47°C, respectively($p<1x10^{-5}$ by two-tailed t-test for both temperatures). Mutations at position 581 to threonine (I581T; present in all HiDiv and 1 WT evolved populations)

showed a more modest increase in thermal resistance compared to the WT virus, conferring an increase in the number of surviving PFU following heat treatment of 14.9 (SD 5.7) or 28.9 (SD 5.7) times for 45°C and 47°C, respectively ($p < 0.001$ using a two-tailed t-test for both temperatures). Interestingly, the E596G mutation failed to produce viable virus following 3 independent attempts, indicating a strong fitness cost for this mutation in the absence of other mutations, highlighting the role for epistasis in the evolved lines.

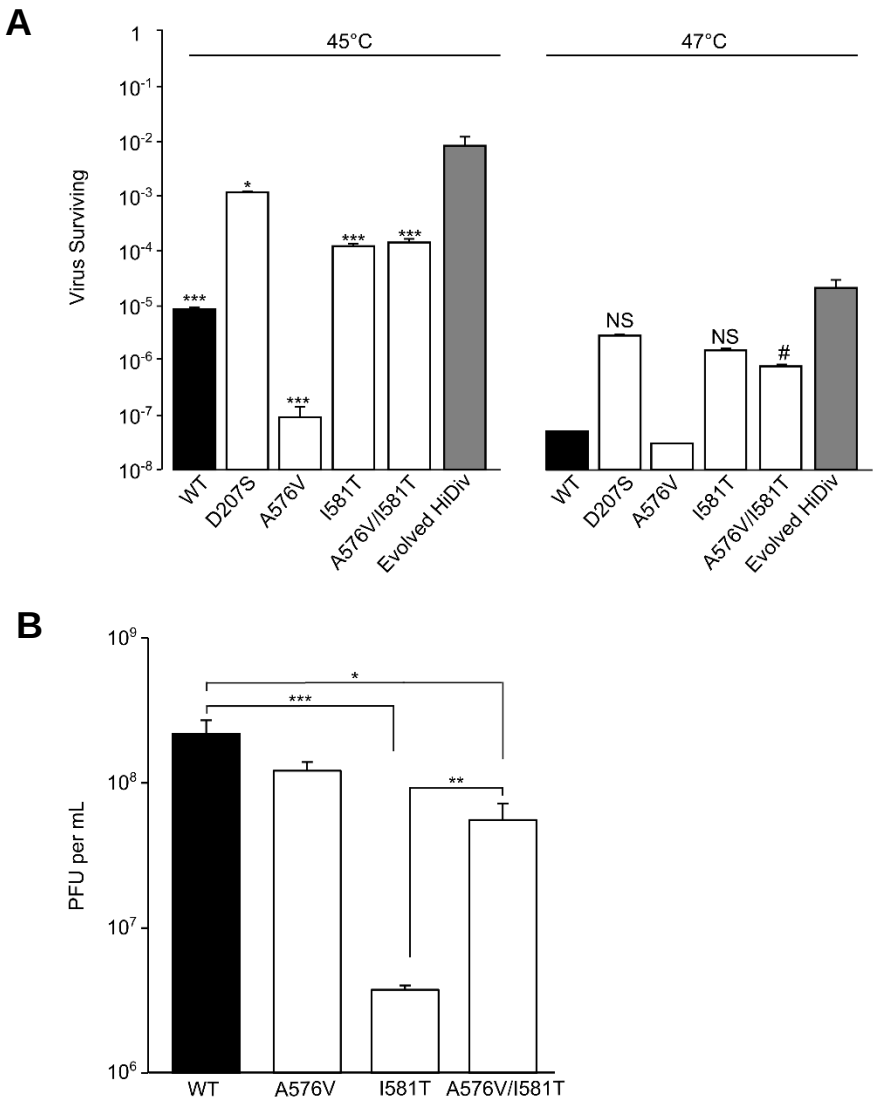


***Figure ChVII*-3 Resistance to thermal inactivation of select mutants from the evolved populations. A)** Fraction of surviving virus following thermal treatment for the different mutants. Viral populations were incubated for 30 minutes at the indicated temperature and the fraction of the surviving virus was calculated as previously described. **B)** The thermosensitive A576V mutant compensates for the fitness cost of the thermostable I581T mutation. The mean number of viruses produced for each population following a single round of infection is shown. Results indicate the mean and SEM of at least 3 independent replicates. For both the WT virus and A576V, no plaques were observed at the lowest dilution, and a minimum quantity of 1 plaque was added to enable graphical and statistical analysis. N = 3 for all experiments. # p= 0.05, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$ by a two-tailed t-test on log-transformed data.

Surprisingly, the A576V mutation observed in 2 of the HiDiv populations was found to reduce the number of viruses surviving incubation at 45°C by ~90 times compared to the WT populations ($p < 0.005$ using a two-tailed t-test; Figure ChVII-3A and Table S8). As this mutation always occurs in the context of the I581T mutation, we speculated that it is likely to either increase thermal stability or improve the fitness of the I581T mutation. To test these possibilities, we compared the ability of the double mutant (A576V/I581T) to resist thermal inactivation and evaluated the fitness of both the single mutations as well as the double mutants using one-step growth curves. The A576V/I581T double mutant did not display increased resistance to thermal inactivation compared to the I581T mutant, showing the same inactivation rate at 45°C and a modestly lower rate at 47°C ($p > 0.05$ and $p < 0.05$ at 45°C and 47°C using a two-tailed t-test, respectively; Figure ChVII-3A and Table S8). On the other hand, the double mutant had significantly improved fitness compared to the single I581T mutation, showing a modest three-fold reduction in virus titer versus the WT virus compared to a 52-fold reduction observed for the I581T mutant ($p < 0.05$ and $p < 0.001$ for the A576V/I581T and I581T versus WT using a two-tailed t-test, respectively; Figure ChVII-3B and Table S9). In contrast, the A576V mutation on its own did not significantly affect virus production ($p > 0.05$ using a two-tailed t-test; Figure CHVII-3B and Table S9). Together, these results indicate that the optimal adaptation of resistance to thermal inactivation is complex, requiring multiple changes that either increase thermal resistance or help compensate for reductions in fitness.

5. Discussion

RNA viruses have an extreme capacity for rapid evolution due to low-fidelity replication machinery, large population sizes, and small genomes. Indeed, RNA viruses replicate at the brink of error catastrophe, where even small increases in mutation rates can lead to the extinction of the population as a result of too many mutations (Perales and Domingo 2016). Previous work using high-fidelity polymerase mutants have shown that RNA virus populations with reduced diversity are less able to adapt to new environments compared to standard virus populations (Vignuzzi *et al.* 2008; Bordería *et al.* 2016; Kautz and Forrester 2018). In the current work, we examine the opposite scenario, testing whether RNA viruses can further benefit from experimental augmentation of initial population diversity during experimental evolution. We chose to increase diversity using a codon-level mutagenesis protocol which, unlike error-prone mutagenesis, allows for the introduction of all possible single amino acid mutations away from the original viral genotype. We find that experimentally increasing initial population diversity enables viral populations to achieve better adaptation to the selection regime (HiDiv versus WT population at 45°C; Figure ChVII-2B). Interestingly, not only were HiDiv populations better adapted to survive inactivation at the temperature confronted during the experimental evolution regime, but they were also better able to survive a stronger selection pressure (HiDiv versus WT population at 47°C; Figure ChVII-2B). Overall, these results indicate that even RNA viruses with high mutation rates can benefit from augmentation of population diversity to increase their capacity to adapt. While we investigated adaptation to a single selection pressure, increasing initial population diversity may be of particular relevance for adaptation to more complex environments such as changes in host species. Overall, these findings have implications for directed evolution experiments aimed at the selection of desired traits in RNA viruses, such as increased stability of live attenuated vaccines, alterations of tissue tropism in gene therapy, or optimization of oncolytic activity.

Analysis of mutations present in the adapted populations revealed numerous non-synonymous changes in both the WT and HiDiv populations (Table CHVII-1 and Table S8), suggesting that optimal adaptation to thermal stability requires multiple mutations in the capsid. In addition, we observed mutations that conferred thermal sensitivity (A576V) or which failed to yield infectious virus when present individually (E596G), indicating strong epistatic interactions that are common in RNA viruses (Sanjuán *et al.* 2005) (Figure ChVII-3). This is exemplified by mutations at positions 576 and 581. Mutation A576V was more sensitive to thermal inactivation relative to the WT virus but had no significant cost to fitness. The I581T mutation, on the other hand, was more resistant to

thermal inactivation but incurred a significant fitness cost. In combination, the A576V/I581T double mutant showed similar resistance to thermal inactivation as the I581T mutation alone but with a significantly lower cost to fitness. Interestingly, the I581M mutation, which did not yield viable virus when in isolation, occurs in the context of a different mutation at position 576 (A576T), suggesting an even stronger epistatic interaction between these two mutations.

From the mutations shown to confer resistance to thermal inactivation, the D207S mutation found in two of the three HiDiv populations conferred the greatest resistance to thermal inactivation (Figure ChVII-3A). This mutation requires three transversions to occur within the same codon (GAC to TCG; Table S8), an extremely unlikely event during natural viral replication due to both the low probability of multiple mutations occurring in the same codon and the lower rate of transversions compared to transitions in RNA viruses (Acevedo *et al.* 2014; Geller *et al.* 2016). This mutation is also unlikely to have appeared if random mutagenesis methods (e.g. error-prone PCR) were used to enrich the initial population diversity instead of the codon-based mutagenesis strategy employed in this work because the probabilistic nature of mutations inserted by such methods is unlikely to yield multiple mutations within the same codon. Interestingly, the D207S mutation can also be reached by two transition mutations (GAC to AGC), yet despite this simpler evolutionary pathway, this mutation was not observed in the WT evolved populations. Other mutations that increase thermal stability were reached by single mutations. This includes the I581T mutation, which was present in all HiDiv populations and conferred a more modest resistance to thermal inactivation (Figure ChVII-3A), as well as the A512T mutation, which was present in five of the six lines and was previously shown to confer thermal resistance (Carson *et al.* 2019). Hence, these mutations are likely to represent more frequent paths for adaptation in natural CVB3 populations, yet may not provide the optimal degree of resistance to thermal inactivation that can be reached by the incorporation of multiple mutations within a single codon.

The high rate of recombination in positive-strand RNA viruses such as CVB3 enables the shuffling of mutations from different genomes, promoting the joining of beneficial mutations in the same genotype while helping to purge deleterious mutations (Simon-Loriere and Holmes 2011). The large number of mutations observed in all lines after 10 passages is likely to arise at least in part from recombination between different genomes. In this sense, recombination is likely to play a key role in the adaptation of these highly diverse populations, especially considering the epistatic nature of the mutations observed in the evolved populations. The availability of mutants in RNA polymerases that show reduced recombination rates (Kempf *et al.* 2016; Xiao *et al.* 2017) can provide

an interesting tool for dissecting the role of recombination in the ability of highly diverse populations to better adapt to new environments compared to standard populations.

6. Data availability

Unaligned bam files have been uploaded to SRA (Accession SAMN16492169-SAMN16492172, SRX9320721-SRX9320724). The scripts and data required to obtain the codon count tables for all samples, calculating the number of amino acids reached by each type of mutation, calculating the probability of different mutations occurring in the same codon, and the alignment for entropy calculations can be found on Github (https://github.com/RGellerLab/dms_thermal_selection).

7. Supplementary data

Supplementary data are available in the appendix of this manuscript. **Figure S1:** Analysis of library diversity by Sanger sequencing. **Figure S2:** Positions mutated in evolved populations occur in variable positions. **Table S1:** Codons and amino acids reached by single, double and triple mutations **Table S2:** DMS mutagenesis primers **Table S3:** Mutations in libraries identified by Sanger sequencing **Table S4:** Mutations in libraries identified by high-fidelity next-generation sequencing **Table S5:** Viral titers during experimental evolution **Table S6:** Comparison of resistance to thermal inactivation of the evolved passage 10 populations **Table S7:** Mutations identified evolved passage 10 populations **Table S8:** Comparison of resistance to thermal inactivation of the mutants identified in evolved passage 10 populations **Table S9:** Comparison of titers for mutants identified in evolved passage 10 populations **Data set S1:** All available non-identical full-genome CVB sequences from Virus Pathogen Resource.

GENERAL DISCUSSION

This thesis is intended to unveil the contribution of gene duplication and genetic variability in biological innovation through experimental evolution. It seems clear that gene duplication contributes to biological robustness, phenotypic and transcriptional plasticity, and thereby to biological innovation and adaptation to new environments (Innan and Kondrashov 2010; Wagner 2011; Fares *et al.* 2013; Keane *et al.* 2014; Fares 2015b; Zheng *et al.* 2019). However, gene duplication is highly unstable because of genetic redundancy and high fitness costs (Krakauer and Plotkin 2002; Adler *et al.* 2014). Following a whole-genome duplication event that occurred 100 million years ago in the budding yeast *Saccharomyces cerevisiae,* about 92% of the duplicated genes returned to single copies (Wolfe and Shields 1997; Keane *et al.* 2014). Despite the large loss of duplicated genes, the number of duplicated genes in *S. cerevisiae* (roughly 30% of all genes) is higher than expected. But, the question about what factors have determined which genes are retained in the genome after a duplication event, and which are lost, still unanswered. While numerous models and hypotheses exist to explain the preservation of duplicated genes, the factors that play a key role in determining the fate and evolution of duplicated genes are yet to be deciphered (Innan and Kondrashov 2010). Can a general mechanistic model be established to explain the survival of duplicated genes in the genome?

Our results show that, in agreement with the Dosage Sub-functionalization Hypothesis (DSH), gene expression determines the preservation of duplicated genes in the yeast genome following both Whole-Genome Duplications (WGDs) and Small-Scale Duplications (SSDs). The DSH hypothesis is based on the Duplication-Degeneration-Complementation (DDC) model for evolution after gene duplication (Force, Lynch, and Postlethwait 1999; Force, Lynch, Pickett, *et al.* 1999). According to DSH, the degeneration step (i.e. the period during which the gene copies accumulate mutations) is driven by gene expression levels. Hence, natural selection acts on the total level of gene expression (i.e. maintaining the dosage balance) rather than on the individual level of each paralog, conserving those duplicated genes that by the sum of the expression levels of both copies maintain the total amount of the final product. Therefore, if expression divergence increases expression levels of one paralog and reduces the expression levels of its sister copy, natural selection will select for the highly expressed one while relaxing selection on the low expression copy (Gout and Lynch 2015). Although other works suggest that variance in gene expression undergoes neutral evolution, the same authors suggest that using another type of expression data could change their statement, since they used microarray data and intraspecific analysis (Ho *et al.* 2017). The results presented in the first chapter of this doctoral thesis show that

highly expressed genes before duplication are more conserved after a duplication event and show larger phylogenetic stability in the subphylum *Saccharomicotina* (e.g. most of the ascomycete yeasts). Although this observation could also be explained by dosage balance, since genes encoding for proteins forming part of a protein complex will be conserved by purifying selection to maintain interaction stoichiometry, we can reject this hypothesis because neither WGD nor SSD are enriched for genes encoding proteins forming part of a complex.

Additionally, the noise in the expression of higher expressed genes in the genome of *S. cerevisiae* favors functional and transcriptional divergence after a gene duplication event becoming adaptive to new environments (Keane *et al.* 2014; Fares 2015b). Indeed, this observation has been modeled on genes with noisy expression level with intermediate levels of environmental stress (i.e. the fitness trade-off become more evident), where after a duplication event one copy reduces gene expression to adjust the phenotype, yielding a selective advantage for gene duplication (Rodrigo and Fares 2018). In this vein, a recent study done in the bacteria *E.coli* has shown that expression noise is crucial for protecting alleles from extinction and promotes the accumulation of beneficial mutations, revealing an important link between noise in gene expression levels and adaptation to new environments (Schmutzer and Wagner 2020).

Our results in Chapter II show that ancient duplicates in the yeast *S. cerevisiae* present larger transcriptional plasticity under stress conditions. This can be explained by the existence of pre-adaptations to environmental stress by accumulating polymorphisms on regulatory regions of anciently duplicated genes, leading to a complete transcriptional divergence and hence to transcriptional plasticity. Similar results were obtained in a recent study made with the fungus *Epichloë,* where the authors observed clear sequence divergence in regulatory regions and greater expression variation in the retained duplicated genes compared with singletons (Wu and Cox 2019). Transcriptional divergence of duplicated genes has been widely studied in plants and animals (Huminiecki and Wolfe 2004; Blanc and Wolfe 2004b; Ha *et al.* 2007, 2009; Huminiecki and Conant 2012; Wang *et al.* 2012; Defoort *et al.* 2019; Jiang and Assis 2019; Lafuente and Beldade 2019), yet in yeast it has received little attention. Although the transcriptomic profile of the yeast *S. cerevisiae* has widely been studied under stress (Ferea *et al.* 1999; Causton *et al.* 2001; Ideker *et al.* 2001; Landry *et al.* 2006; Stern *et al.* 2007; Cormier *et al.* 2010), little attention has been given to the different expression patterns of duplicates and singletons. Here we show that when the budding yeast *S. cerevisiae* is subjected to moderate concentrations of glycerol (an osmotic stressor for the cell), ethanol, or lactic acid, a dramatic transcriptional re-programming is observed

that is mainly driven through duplicated genes (as explained in Chapters III to V). Our results underline the importance of transcriptional plasticity in *S.cerevisiae*, agreeing with similar observations made in other yeast species like *Saccharomyces kudriavzevii* under ethanol stress or *Zygosaccharomyces parabaili* under acidic stress (Ortiz-Merino *et al.* 2017; Macías *et al.* 2019). Indeed, from the results presented in this thesis, two main observations can be derived. First, ancient duplicates in *S. cerevisiae* present large transcriptional plasticity under stress. Second, the transcriptional plasticity observed in duplicates is higher than in singletons, likely being the result of selection for an adaptive response.

Furthermore, expression level influences the transcriptional divergence pattern between the gene copies of duplicated genes (see Chapters I and II), with gene duplicates having a discordant expression pattern among the copies (e.g. one copy up-regulated and the sister copy down-regulated) or with only one copy altered under stress being more prone to functionally diverge. In this thesis, we propose the hypothesis that there is a link between transcriptional divergence and functional divergence. This link is in agreement with the misfolding-mistranslation hypothesis, which postulates that highly expressed genes evolve slower because the accumulation of detrimental mutations favors misfolding of the protein that compromises fitness (i.e. protein expression supposes high energy cost, hence detrimental mutations on highly expressed genes that compromise the correct folding of the protein entails a big waste of energy), and this augmented fitness cost drives an increase of the action of purifying selection on highly expressed genes (Drummond *et al.* 2005). Conversely, functional divergence requires fine-tuning of the expression to perform the new function optimally. Thus, our data do not reveal if expression divergence drives functional divergence or the other way around, but the different transcriptional patterns observed under stress suggest a strong link between functional and transcriptional divergence. In fact, the different expression patterns of duplicated genes observed in the yeast *S. cerevisiae* under stress are in good agreement with the classical model for evolution by gene duplication proposed by Ohno (Ohno 1970, 1999; Zhang 2003). We observe in Chapter II that duplicates in which only one copy is altered under stress are the most responsive to explore genotypic space because the higher genetic redundancy after gene duplication allows one copy to be unseen by natural selection and this relaxation from selection provides the opportunity to find biological innovation (Wagner 2005, 2011; Fares *et al.* 2013; Keane *et al.* 2014; Fares 2015c). We also observe that up-regulated duplicates are more prone to *neo-functionalize* based on the rapid evolution and low functional dependency among gene copies. This suggests that after gene duplication both copies diverge rapidly, generating

a selective pressure for both, and allowing the acquisition of a novel function. On the other hand, duplicates that are down-regulated under stress show evidence of *sub-functionalization* based on the low sequence divergence and high functional dependency between the sister copies, suggesting that evolution after gene duplication was slower and the accumulation of deleterious mutation was occurring simultaneously on both copies. This drives to the partition of the ancestral function, forcing natural selection to keep both copies in the genome. Finally, those duplicates with discordant expression patterns (e.g. one copy is up-regulated and the other one is down-regulated) present low sequence divergence but also low functional divergence, suggesting transcriptional divergence for performing similar functions under different conditions.

As mentioned before, transcriptional plasticity in *S. cerevisiae* seems to be the result of polymorphisms in regulatory elements of duplicated genes that become fixed in the population after the gene duplication event, thereby providing pre-adaptations to novel environments (Tamari *et al.* 2016; Escalera-Fanjul *et al.* 2019). Indeed, this was observed after a neutral experimental evolution of 660 generations followed by subjecting the population to different stressors (e.g. glycerol, lactic acid, ethanol, or peroxide). The evolved population (t_{100}) showed both different stress responses and an increased fitness (e.g. growth rate) compared with the ancestral population (t_0), suggesting that the experimental evolution enabled increased genetic variability in the yeast population. Hence, the larger genetic variability acquired by neutral evolution for 100 passages (~660 generations) allowed the evolved yeast population to affront better a sudden environmental change. (see Chapters III to V). Indeed, the importance of increased genetic variability in the population to allow for better adaptation to challenging environments has also been observed in the second part of this thesis. We show that after the experimental evolution of highly diverse CVB3 viral populations, these achieved higher resistance to thermal inactivation compared with the wild-type populations (see Chapter VII). All the results observed in this thesis regard the benefit of genetic variability for evolution and adaptation agree with the recent work showing that YFP protein populations with cryptic genetic variation achieved faster color change (from yellow to green) and higher fluorescence (Zheng *et al.* 2019). Nevertheless, a conflict between noisy genes and transcriptional plasticity can exist (Raser and O'Shea 2005; Blake *et al.* 2006; Lehner 2010). Here we show in Chapter I that preserved duplicates are more enriched for TATA boxes in their regulatory regions, with these TATA motifs being associated with higher noise but also with higher transcriptional plasticity (Newman *et al.* 2006; Tirosh *et al.* 2006; Landry *et al.* 2007). Interestingly, not only plastic duplicates (i.e. showing high transcriptional plasticity) but also the non-plastic ones present more TATA

motifs in their regulatory regions compared to plastic singletons. Thus, taking together the different expression pattern and the enrichment in TATA boxes in duplicated genes suggest that plasticity is gained by gene duplication. Thus, gene duplication can reduce the cost-benefit conflict existing between noise and transcriptional plasticity, allowing adaptation to new environments. Indeed, our data reveal a strong link between cellular reprogramming of the cell and environmental stresses (e.g. glycerol, ethanol, or lactic acid), with the majority of the transcriptomic responses to cellular stressors being driven by duplicated genes.

In fact, an in-depth analysis of the stress response of the yeast *S.cerevisiae* shows that two types of stress responses can be observed: a quick one and a slow one. On one hand, the quick stress response is more general to all the stresses, resulting in the upregulation of important cellular functions like genes involved in cellular respiration, according to the shift from a fermentative carbon source (glucose) to a non-fermentative one (glycerol, lactate, or ethanol). Besides, specific stress response genes are also upregulated (e.g. transporters to efflux glycerol from the cytoplasm when the population was challenged with 3% glycerol, or hydrogen transmembrane transport in populations that have been challenged with lactic acid, which lowers the pH of the media). In contrast, energetically costly cellular processes (e.g. ribosome biosynthesis and protein translation) are generally down-regulated in all tested stresses. On the other hand, the slow stress response requires transcriptional fine-tuning, as can be observed in the populations evolved (t_{110}) to chronic stresses, including glycerol, ethanol, or lactic acid. Similar results have been observed subjecting the yeast to other stress conditions like temperature (Gasch *et al.* 2000; Causton *et al.* 2001; García-Ríos *et al.* 2017), low pH (Narayanan *et al.* 2016; Fletcher *et al.* 2017; Ortiz-Merino *et al.* 2017), different carbon sources (Bradley *et al.* 2019; Duan *et al.* 2019) or other stresses that play an important role in the industry (Legras *et al.* 2018; Lopandic 2018).

In the second part of this doctoral thesis, using a different experimental system, the coxsackievirus B3, we address the question about the important role of genetic variability in evolution (Zheng *et al.* 2019; Wideman *et al.* 2019). For this, we performed a PCR-based deep mutational scanning approach (DMS) to study the contribution of genotypic variability to adaptation, assessing the mutational fitness effect of almost all possible single amino acid mutations in the CVB3 viral capsid. As expected, the vast majority of mutations that we were able to characterize showed a deleterious fitness effect compared with the wild-type sequence, with only 1.2% out of ~92% of all possible mutations in the capsid showing higher fitness (see Chapter VI). Our exhaustive analysis of almost all possible amino acid mutations contributes to a better understanding of the

CVB3 viral capsid. Our results can be extended to other picornaviruses considering the conserved structure of their capsids. Indeed, we observe signatures of differential selection driven by the immune system of the host selecting for major genetic variability on antibody neutralization sites. Identify the antibody neutralization sites and try to establish a general model for identifying antibody and therapeutic targets has previously been addressed in other viruses (Burton 2002; Kalia *et al.* 2005; Haddox *et al.* 2016; Dingens *et al.* 2017; Doud *et al.* 2017), becoming a valuable tool in medicine. Similar to other DMS studies, our results show that the incorporation of site-specific information derived from DMS improves the accuracy of phylogenetic models. Using our dataset we were able to detect information about the detrimental fitness effect of a wider range of mutations that in nature are hard to observe, since these low fitness variants are rapidly purged by natural selection (Doud and Bloom 2016; Bloom 2017; Haddox *et al.* 2018). Besides, we could also use the obtained information to infer putative host factors that have a direct effect on the virus. Specifically, using the amino acid preferences for the viral protease cleavage site we were able to identify a list of human proteins that are potential targets for the CVB3 3CD protease. Constructing a position-specific score matrix for the $3CD^{pro}$ cleavage we queried the human protein and obtained 746 hits, including 11 previously described to be cleaved by the viral protease. (Ypma-Wong *et al.* 1988; Sun *et al.* 2016; Laitinen *et al.* 2016; Jagdeo *et al.* 2018). Interestingly, among the best best-scored hits we identified proteins involved in the interferon-inducible response, apoptosis, RNA processing, and vesicle transport (see Chapter VI). These hits are to be expected to be targeted by the viral protease since those cellular functions might have potential relevance to the CVB3 biology. Because several of these proteins can be readily detected in our culture cell assay we validated by Western Blot the cleavage of 3 hits out of the 7 testest proteins. Our results reveal the power of prediction of our method and our MFE dataset, however, a furthermore specific approach need to be performed including additional criteria to reduce false-positive hits. Nevertheless, the results presented in chapter VI correctly predict 30% of host cellular targets, many of which can play key roles in viral biology and pathogenesis.

Finally, subjecting these highly diverse populations to an experimental evolution for thermal resistance (i.e. selecting only the mutants that are resistant to thermal inactivation) showed that initial genetic variability allows reaching more complex mutations (i.e. double and triple mutations within the same codon) and strong epistatic interactions. These complex and almost inaccessible genotypes have a low probability to arise in the naturally diverse viral population but enhanced not only the capacity to perform better at the temperature the population was selected for (45°C) but also at even

higher temperatures (47°C; see Chapter VII). Our results show that artificially increasing genetic diversity in experimental evolution allows the population to evolve faster and adapt better to the selection pressure, even in RNA viruses with extremely high mutation rates. Even though using deep mutational scanning (DMS) to analyze the evolutionary relevance of cryptic genetic variation seems clear in evolutionary biology, our observations suggest that the DMS approach has also further applications of special relevance in biomedical and biotechnological research. For example, our findings can contribute significantly to improve the performance of directed evolution experiments in RNA viruses. Currently, directed experimental evolution setups rely upon, and take advantage of, the high mutation rates of the RNA viruses and have been used by the scientific community to develop live attenuated vaccines (Buynak and Hilleman 1966; Barrett *et al.* 1990; Minor 2015), selection to improve oncolytic characteristics (Ammayappan *et al.* 2013; Svyatchenko *et al.* 2017; Seegers *et al.* 2019) or isolate more stable virus-like particles (Adeyemi *et al.* 2017; Carson *et al.* 2019). However, using DMS to perform codon mutagenesis, and artificially increase the genetic diversity of the viral population, prior to experimental evolution will accelerate the experiments and result in a better, and more precise, adaptation (e.g. arising adaptative mutations in the populations that are naturally inaccessible).

CONCLUSIONS

Early in this manuscript, the main objective of unveiling the contribution of gene duplication and genetic variability in biological innovation through experimental evolution was proposed. As mentioned, the working plan to address this big question in evolutionary biology has been to divide this doctoral thesis into two parts, one for each experimental model. On one hand, the yeast *Saccharomyces cerevisiae* was used to perform a deep analysis of the contribution of gene duplication to adaptation by analyzing the evolutionary fate of duplicated genes and their role in conferring phenotypic plasticity. On the other hand, a deep mutational scanning was performed on the viral capsid of Coxsackievirus B3. By this, we were able to perform a comprehensive analysis of the mutational fitness effect of almost all single amino acid changes in the capsid and to study the contribution of genetic variability to adaptation.

To achieve all these questions, we proposed four specific objectives for which here I will now present the main conclusions for each one:

For **Objective 1** (*Determine the regulatory and genomic bases for the stability of duplicates genes and their relevance in transcriptional plasticity*), by performing genome comparisons analysis of yeast species from pre- and post- whole-genome duplication event, focusing on the structure of the duplicated genes promoters, determining expression difference between duplicates copies under different stresses, and determining the expression level of duplicates and singletons, we conclude that:

1. A direct correlation between expression level and phylogenetic stability of duplicates is observed, independently of the molecular mechanisms that originate the duplicates (either WGDs or SSDs) in *Saccharomycotina*.

2. Promoter structure, specifically TATA motifs, contributes to transcriptional plasticity.

3. As a general rule, transcriptional level determines the fate of duplicates, being highly expressed genes more likely to be preserved as duplicates (either WGDs or SSDs).

For **Objective 2** (*Study the role of phenotypic plasticity and the contribution of duplicated genes in the cellular response of yeast to environmental stress, adaptation, and biological innovation*), by performing an in deep characterization of *S. cerevisiae* Y06240 transcriptome response to different stresses with special emphasis on the contribution of duplicates, duplication mechanism, sequence divergence among duplicates, gene interactions and protein-protein interactions, we conclude that:

4. Duplicates present a higher transcriptional response to environmental stresses (switch from fermentation to respiration, pH change, osmotic pressure, or response to oxidative damage) than singletons, with a differential response depending on the mechanism of duplication (WGDs more transcriptionally altered than SSDs).

5. Transcriptional plasticity increases after gene duplication may be due to promoter sequence divergence since the duplication event/s.

6. Transcriptional profiling of duplicates copies under stress (both up, both down, only one altered, or discordant) is linked to duplicate evolvability and fate, being able to distinguish between *neo-* and *sub- functionalization* fate depending on duplication mechanism. WGDs are more prone to *neofunctionalization* and SSDs are more prone to *subfunctionalization*.

7. The transcriptional divergence between duplicates copies yields adaptive responses to stress conditions and allows for functional divergence of the copies.

8. Transcriptional divergence correlates with functional divergence, functional dependencies, and sequence divergence of duplicates copies.

For **Objective 3** (*Analyze in detail the transcriptional reprogramming in short stress response, and the adaptation to chronic environmental stress*), employing experimental evolution under adaptive conditions using different non-fermentative carbon sources such as glycerol (also an osmotic stressor), ethanol, or lactic acid (also lowering the pH of the media) and determining transcriptional profiles through the experimental evolution, we conclude that:

9. Challenging the yeast with a non-fermentative carbon source, induces a quick genome-wide transcriptional rewiring, affecting more than 40% of *S. cerevisiae* genes, and inducing more than a 50% decline in growth rate, compared with the use of glucose as carbon source.

10. A short adaptation period induced a massive regulatory reprogramming mostly affecting duplicated genes by downregulation. While under glycerol and ethanol stress more WGD were affected the opposite was true for lactic acid. Also under glycerol and lactic acid stress conditions, a mild recovery of growth rate was observed.

11. Roughly 30% of the altered duplicates belong to a core set of general response involved mostly in the transition from fermentation to respiration, control of stress response, and energy production.

For **Objective 4** (*Determine the viable sequence space of coxsackievirus B3 capsid proteins by deep mutational scanning and experimental evolution*) by performing a comprehensive deep mutational scanning approach to generate highly diverse CVB3 virus populations comprising almost all possible amino acid mutations in the proteins of the viral capsid, and subjecting these populations to an adaptive experimental evolution setup, we conclude that:

12. Most mutations in the CVB3 capsid are deleterious, as judged by assessing the mutational fitness effect of ~92% of all possible single amino acid mutations and observing only 1.2% of the mutations increasing fitness.

13. Mutational fitness effect in the CVB3 capsid shows a strong correlation with structural and functional attributes, such as aggregation propensity, secondary structure, and surface exposure.

14. Mutational fitness effect in the CVB3 capsid can be predicted with high accuracy using natural sequence variation, the identity of the original amino acid, the predicted effect on protein stability, and the relative solvent accessibility.

15. The incorporation of laboratory-measured mutational fitness effect into phylogenetic models improves the model accuracy, providing additional information that cannot be obtained from sequence analysis like signatures of selection from the adaptive immune system.

16. Using experimentally measured mutational fitness effect for the analysis of amino acid preferences and mutational tolerance of known capsid encoded motifs provide in-depth insights into sequence specificity. Indeed, using this information to query the human genome for potential cellular targets for the viral 3CD protease we correctly predict 30% of the identified cleavage sites.

17. Increased genotypic variability in the populations accelerates evolutions and favors adaptation, even in RNA viruses that naturally show extremely high mutation rates.

18. Optimal adaptation of resistance to thermal inactivation is a complex process requiring multiple mutations within the same codon and more than one mutation per genome, pointing out the importance of epistasis.

In summary, in this doctoral thesis we show evidence for the importance of cryptic genetic variability for evolution, serving as a source to bridge different optimal fitness peaks in the evolutionary landscape, and how this genetic variability can be achieved experimentally by a deep mutational scanning approach. In addition, we use experimental evolution to take advantage of increased genetic variability in viruses and

in yeast to show that this favors adaptation to environmental stress. In fact, we show experimental evidence for the role and importance of gene duplication for adaptation and source for biological innovation, supporting the general idea that natural selection relaxes after gene duplication, allowing the accumulation of genetic variability on one of the copies. Thus, our results show strong evidence for the existence of a link between duplicated genes and phenotypic and transcriptional plasticity through transcriptional and functional divergence after gene duplication.

REFERENCES

Acevedo A, Brodsky L, Andino R. **2014**. Mutational and fitness landscapes of an RNA virus revealed through population sequencing. *Nature* **505**: 686–690.

Adeyemi OO, Nicol C, Stonehouse NJ, Rowlands DJ. **2017**. Increasing Type 1 Poliovirus capsid stability by thermal selection. *Journal of Virology* **91**: e01586-16.

Adler M, Anjum M, Berg OG, Andersson DI, Sandegren L. **2014**. High fitness costs and instability of gene duplications reduce rates of evolution of new genes by duplication-divergence mechanisms. *Molecular Biology and Evolution* **31**: 1526–1535.

Albert FW, Muzzey D, Weissman JS, Kruglyak L. **2014**. Genetic influences on translation in yeast. *PLoS Genetics* **10**(10): e1004692.

Alemohammad MM, Knowles CJ. **1974**. Osmotically induced volume and turbidity changes of *Escherichia coli* due to salts, sucrose and glycerol, with particular reference to the rapid permeation of glycerol into the cell. *Journal of General Microbiology* **82**: 125–142.

Alepuz PM, De Nadal E, Zapater M, Ammerer G, Posas F. **2003**. Osmostress-induced transcription by Hot1 depends on a Hog1-mediated recruitment of the RNA Pol II. *EMBO Journal* **22**: 2433–2442.

Alexandre H, Ansanay-Galeote V, Dequin S, Blondin B. **2001**. Global gene expression during short-term ethanol stress in *Saccharomyces cerevisiae*. *FEBS Letters* **498**: 98–103.

Altschul SF, Madden TL, Schäffer AA, *et al.* 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research* **25**: 3389–3402.

Ammayappan A, Peng K-W, Russell SJ. **2013**. Characteristics of oncolytic vesicular stomatitis virus displaying tumor-targeting ligands. *Journal of Virology* **87**: 13543–13555.

Anand A, Chen K, Catoiu E, *et al.* 2020. OxyR is a convergent target for mutations acquired during adaptation to oxidative stress-prone metabolic states. *Molecular Biology and Evolution* **37**: 660–666.

Anders S, Huber W. **2010**. Differential expression analysis for sequence count data. *Genome Biology* **11**(10):R106.

Andersson L. **2012**. How selective sweeps in domestic animals provide new insight into biological mechanisms. *Journal of Internal Medicine* **271**: 1–14.

Andrade RP, Casal M. **2001**. Expression of the lactate permease gene JEN1 from the yeast *Saccharomyces cerevisiae*. *Fungal Genetics and Biology* **32**: 105–111.

Aramburu J, Ortells MC, Tejedor S, Buxadé M, López-Rodríguez C. **2014**. Transcriptional regulation of the stress response by mTOR. *Science Signaling* **7**(332):re2.

Araya CL, Fowler DM. **2011**. Deep mutational scanning : assessing protein function on a massive scale. *Trends in Biotechnology* **29**: 435–442.

Armetta J, Berthome R, Cros A, *et al.* 2019. Biosensor-based enzyme engineering approach applied to psicose biosynthesis. *Synthetic Biology* **4**(1):ysz028.

Arribas M, Aguirre J, Manrubia S, Lázaro E. **2018**. Differences in adaptive dynamics determine the success of virus variants that propagate together. *Virus Evolution* **4**: 1–11.

Ashenberg O, Padmakumar J, Doud MB, Bloom JD. **2017**. Deep mutational scanning identifies sites in influenza nucleoprotein that affect viral inhibition by MxA. *PLoS*

Pathogens **13**: 1–23.

Aury JM, Jaillon O, Duret L, *et al.* 2006. Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. *Nature* **444**: 171–178.

Avrahami-Moyal L, Engelberg D, Wenger JW, Sherlock G, Braun S. **2012**. Turbidostat culture of *Saccharomyces cerevisiae* W303-1A under selective pressure elicited by ethanol selects for mutations in SSD1 and UTH1. *FEMS Yeast Research* **12**: 521–533.

Babazadeh R, Lahtvee PJ, Adiels CB, Goksör M, Nielsen JB, Hohmann S. **2017**. The yeast osmostress response is carbon source dependent. *Scientific Reports* **7**: 1–11.

Bai C, Tesker M, Engelberg D. **2015**. The yeast Hot1 transcription factor is critical for activating a single target gene, STL1. *Molecular Biology of the Cell* **26**: 2357–2374.

Baier F, Miton CM, Pabis A, *et al.* 2019. Cryptic genetic variation shapes the adaptive evolutionary potential of enzymes. *eLife* **8**: 1–20.

Ball P, Hallsworth JE. **2015**. Water structure and chaotropicity: Their uses, abuses and biological implications. *Physical Chemistry Chemical Physics* **17**: 8297–8305.

Barkman T, Zhang J. **2009**. Evidence for escape from adaptive conflict? *Nature* **462**: E1–E1.

Barrett ADT, Monath TP, Cropp CB, *et al.* 1990. Attenuation of wild-type yellow fever virus by passage in HeLa cells. *Journal of General Virology* **71**: 2301–2306.

Barrick JE, Yu DS, Yoon SH, *et al.* 2009. Genome evolution and adaptation in a long-term experiment with *Escherichia coli*. *Nature* **461**: 1243–1247.

Basso LC, De Amorim H V., De Oliveira AJ, Lopes ML. **2008**. Yeast selection for fuel ethanol production in Brazil. *FEMS Yeast Research*.1155–1163.

Baty F, Delignette-Muller ML. **2004**. Estimating the bacterial lag time: Which model, which precision? *International Journal of Food Microbiology* **91**: 261–277.

Bellí G, Molina MM, García-Martínez J, Pérez-Ortsín JE, Herrero E. **2004**. *Saccharomyces cerevisiae* glutaredoxin 5-deficient cells subjected to continuous oxidizing conditions are affected in the epression of specific sets of genes. *Journal of Biological Chemistry* **279**: 12386–12395.

Belnap DM, McDermott BM, Filman DJ, *et al.* 2000. Three-dimensional structure of poliovirus receptor bound to poliovirus. *Proceedings of the National Academy of Sciences of the United States of America* **97**: 73–78.

Benjamini Y, Hochberg Y. **1995**. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **57**: 289–300.

Benjamini Y, Yekutieli D. **2001**. The control of the false discovery rate in multiple testing under depency. *The Annals of Statistics* **29**(4):1165-1188.

Bereketoglu C, Arga KY, Eraslan S, Mertoglu B. **2017**. Genome reprogramming in *Saccharomyces cerevisiae* upon nonylphenol exposure. *Physiological Genomics* **49**: 549–566.

Bergström A, Simpson JT, Salinas F, *et al.* 2014. A high-definition view of functional genetic variation from natural yeast genomes. *Molecular Biology and Evolution* **31**: 872–888.

Bergthorsson U, Andersson DI, Roth JR. **2007**. Ohno's dilemma: Evolution of new genes under continuous selection. *Proceedings of the National Academy of Sciences of the United States of America* **104**: 17004–17009.

Berterame NM, Porro D, Ami D, Branduardi P. **2016**. Protein aggregation and membrane lipid modifications under lactic acid stress in wild type and OPI1 deleted *Saccharomyces cerevisiae* strains. *Microbial Cell Factories* **15**: 1–12.

Bhaganna P, Volkers RJM, Bell ANW, *et al.* 2010. Hydrophobic substances induce water stress in microbial cells. *Microbial Biotechnology* **3**: 701–716.

Birchler JA, Veitia RA. **2012**. Gene balance hypothesis: Connecting issues of dosage sensitivity across biological disciplines. *Proceedings of the National Academy of Sciences of the United States of America* **109**: 14746–14753.

Bjornsti MA, Houghton PJ. **2004**. The TOR pathway: A target for cancer therapy. *Nature Reviews Cancer* **4**: 335–348.

Blake WJ, Balázsi G, Kohanski MA, *et al.* 2006. Phenotypic consequences of promoter-mediated transcriptional noise. *Molecular Cell* **24**: 853–865.

Blanc G, Wolfe KH. **2004a**. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *The Plant Cell* **16**: 1667–1678.

Blanc G, Wolfe KH. **2004b**. Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution. *The Plant Cell* **16**: 1679–1691.

Bloom JD. **2014**. An experimentally determined evolutionary model dramatically improves phylogenetic fit. *Molecular Biology and Evolution* **31**: 1956–1978.

Bloom JD. **2015**. Software for the analysis and visualization of deep mutational scanning data. *BMC Bioinformatics* **16**(168): 1–13.

Bloom JD. **2017**. Identification of positive selection in genes is greatly improved by using experimentally informed site-specific models. *Biology Direct* **12**: 1–24.

Blount ZD, Barrick JE, Davidson CJ, Lenski RE. **2012**. Genomic analysis of a key innovation in an experimental *Escherichia coli* population. *Nature* **489**: 513–518.

Blount ZD, Borland CZ, Lenski RE. **2008**. Historical contingency and the evolution of a key innovation in an experimental population of *Escherichia coli*. *Proceedings of the National Academy of Sciences of the United States of America* **105**: 7899–7906.

Bologna G, Yvon C, Duvaud S, Veuthey AL. **2004**. N-terminal myristoylation predictions by ensembles of neural networks. *Proteomics* **4**: 1626–1632.

Bordería A V., Rozen-Gagnon K, Vignuzzi M. **2016**. Fidelity variants and RNA Quasispecies In: Domingo E, Schuster P, eds. *Quasispecies: From Theory to Experimental Systems*. Springer, Cham, 303–322.

Bornscheuer UT, Höhne M. **2018**. *Protein engineering - Methods and protocols*.

Botha A. **2011**. The importance and ecology of yeasts in soil. *Soil Biology and Biochemistry* **43**: 1–8.

Bou JV, Geller R, Sanjuán R. **2019**. Membrane-associated enteroviruses undergo intercellular transmission as pools of sibling viral genomes. *Cell Reports* **29**: 714-723.e4.

Bradley PH, Gibney PA, Botstein D, Troyanskaya OG, Rabinowitz JD. **2019**. Minor isozymes tailor yeast metabolism to carbon availability. *mSystems* **4**: 1–19.

Brass AL, Huang IC, Benita Y, *et al.* 2009. The IFITM proteins mediate cellular

resistance to influenza a H1N1 virus, west nile virus, and dengue virus. *Cell* **139**: 1243–1254.

Brauer MJ, Huttenhower C, Airoldi EM, *et al.* 2008. Coordination of growth rate, cell cycle, stress response, and metabolic activity in yeast. *Molecular Biology of the Cell* **19**: 352–367.

Brion C, Pflieger D, Souali-Crespo S, Friedrich A, Schacherer J. **2016**. Differences in environmental stress response among yeasts is consistent with species-specific lifestyles. *Molecular Biology of the Cell* **27**: 1694–1705.

Broach JR. **2012**. Nutritional control of growth and development in yeast. *Genetics* **192**: 73–105.

Burton DR. **2002**. Antibodies, viruses and vaccines. *Nature Reviews Immunology* **2**: 706–713.

Buynak EB, Hilleman MR. **1966**. Live attenuated mumps virus vaccine. 1. vaccine development. *Proceedings of the Society for Experimental Biology and Medicine* **123**: 768–775.

Byrne KP, Wolfe KH. **2005**. The Yeast Gene Order Browser: Combining curated homology and syntenic context reveals gene fate in polyploid species. *Genome Research* **15**: 1456–1461.

Callaway A, Giesman-Cookmeyer D, Gillock ET, Sit TL, Lommel SA. **2001**. The multifunctional capsid proteins of plant RNA viruses. *Annual Review of Phytopathology* **39**: 419–460.

Cambray G, Mazel D. **2008**. Synonymous genes explore different evolutionary landscapes (HS Malik, Ed.). *PLoS Genetics* **4**: e1000256.

Canoid A V, Payne JL. **2020**. Mutation bias interacts with composition bias to influence adaptive evolution. *PLoS Computational biology* **16**(9):e1008269

Cardenas ME, Cutler NS, Lorenz MC, Di Como CJ, Heitman J. **1999**. The TOR signaling cascade regulates gene expression in response to nutrients. *Genes and Development* **13**: 3271–3279.

Carrasco P, Daròs JA, Agudelo-Romero P, Elena SF. **2007**. A real-time RT-PCR assay for quantifying the fitness of tobacco etch virus in competition experiments. *Journal of Virological Methods* **139**: 181–188.

Carretero-Paulet L, Albert VA, Fares MA. **2013**. Molecular evolutionary mechanisms driving functional diversification of the HSP90A family of heat shock proteins in eukaryotes. *Molecular Biology and Evolution* **30**: 2035–2043.

Carretero-Paulet L, Fares MA. **2012**. Evolutionary dynamics and functional specialization of plant paralogs formed by whole and small-scale genome duplications. *Molecular Biology and Evolution* **29**: 3541–3551.

Carrillo-Tripp M, Shepherd CM, Borelli IA, *et al.* 2009. VIPERdb2: An enhanced and web API enabled relational database for structural virology. *Nucleic Acids Research* **37**: D436-42.

Carroll AS, O'Shea EK. **2002**. Pho85 and signaling environmental conditions. *Trends in Biochemical Sciences* **27**: 87–93.

Carson SD, Tracy S, Kaczmarek ZG, Alhazmi A, Chapman NM, Carson SD. **2019**. Three capsid amino acids notably influence coxsackie B3 virus stability. *Journal of General Virology* **97**: 60–68.

Casal M, Paiva S, Queirós O, Soares-Silva I. **2008**. Transport of carboxylic acids in yeasts. *FEMS Microbiology Reviews* **32**: 974–994.

Castrillo JI, Zeef LA, Hoyle DC, *et al.* 2007. Growth control of the eukaryote cell: A systems biology study in yeast. *Journal of Biology* **6**.

Causton HC, Ren B, Sang Seok Koh, *et al.* 2001. Remodeling of yeast genome expression in response to environmental changes. *Molecular Biology of the Cell* **12**: 323–337.

Chan SL, Huppertz I, Yao C, *et al.* 2014. CPSF30 and Wdr33 directly bind to AAUAAA in mammalian mRNA 3′ processing. *Genes and Development* **28**: 2370–2380.

Chapal M, Mintzer S, Brodsky S, Carmi M, Barkai N. **2019**. Resolving noise-control conflict by gene duplication. *PLoS Biology* **17**.

Chappell CR, Fukami T. **2018**. Nectar yeasts: a natural microcosm for ecology. *Yeast* **35**: 417–423.

Chen S, Zhou Y, Chen Y, Gu J. **2018**. Fastp: An ultra-fast all-in-one FASTQ preprocessor In: *Bioinformatics*. Oxford University Press, i884–i890.

Chitwood DH, Sinha NR. **2016**. Evolutionary and environmental forces sculpting leaf development. *Current Biology* **26**: R297–R306.

Choudhury A, Fenster JA, Fankhauser RG, Kaar JL, Tenaillon O, Gill RT. **2020**. CRISPR / Cas 9 recombineering-mediated deep mutational scanning of essential genes in *Escherichia coli. Molecular Systems Biology* **16**.

Cifuente JO, Moratorio G. **2019**. Evolutionary and structural overview of human picornavirus capsid antibody evasion. *Frontiers in Cellular and Infection Microbiology* **9**: 283.

Clark JW, Donoghue PCJ. **2018**. Whole-genome duplication and plant macroevolution. *Trends in Plant Science* **23**: 933–945.

Clark JW, Puttick MN, Donoghue PCJ. **2019**. Origin of horsetails and the role of whole-genome duplication in plant macroevolution. *Proceedings. Biological sciences* **286**: 20191662.

Conant GC, Birchler JA, Pires JC. **2014**. Dosage, duplication, and diploidization: Clarifying the interplay of multiple models for duplicate gene evolution over time. *Current Opinion in Plant Biology* **19**: 91–98.

Conant GC, Wolfe KH. **2006**. Functional partitioning of yeast co-expression networks after genome duplication. *PLoS Biology* **4**: 545–554.

Conant GC, Wolfe KH. **2007**. Increased glycolytic flux as an outcome of whole-genome duplication in yeast. *Molecular Systems Biology* **3**.

Conant GC, Wolfe KH. **2008**. Turning a hobby into a job: How duplicated genes find new functions. *Nature Reviews Genetics* **9**: 938–950.

Corbic Ramljak I, Stanger J, Real-Hohn A, *et al.* 2018. *Cellular N-myristoyltransferases play a crucial picornavirus genus-specific role in viral assembly, virion maturation, and infectivity*. Public Library of Science.

Cormier L, Barbey R, Kuras L. **2010**. Transcriptional plasticity through differential assembly of a multiprotein activation complex. *Nucleic Acids Research* **38**: 4998–5014.

Costa V, Reis E, Quintanilha A, Moradas-Ferreira P. **1993**. Acquisition of ethanol

tolerance in *saccharomyces cerevisiae*: the key role of the mitochondrial superoxide dismutase. *Archives of Biochemistry and Biophysics* **300**: 608–614.

Costanzo M, Baryshnikova A, Bellay J, *et al.* 2010. The genetic landscape of a cell. *Science* **327**: 425–431.

Couce A, Tenaillon O. **2019**. Mutation bias and GC content shape antimutator invasions. *Nature Communications* **10**.

Cover TM, Thomas JA. **2005**. *Elements of Information Theory*. Wiley-Interscience, New York.

Cray JA, Bell ANW, Bhaganna P, Mswaka AY, Timson DJ, Hallsworth JE. **2013**. The biology of habitat dominance; can microbes behave as weeds? *Microbial Biotechnology* **6**: 453–492.

Cray JA, Russell JT, Timson DJ, Singhal RS, Hallsworth JE. **2013**. A universal measure of chaotropicity and kosmotropicity. *Environmental Microbiology* **15**: 287–296.

Cray JA, Stevenson A, Ball P, *et al.* 2015. Chaotropicity: A key factor in product tolerance of biofuel-producing microorganisms. *Current Opinion in Biotechnology* **33**: 228–259.

Crozet P, Margalha L, Confraria A, *et al.* 2014. Mechanisms of regulation of SNF1/AMPK/SnRK1 protein kinases. *Frontiers in Plant Science* **5**:190.

Cui L, Wall PK, Leebens-Mack JH, *et al.* 2006. Widespread genome duplications throughout the history of flowering plants. *Genome Research* **16**: 738–749.

Dashko S, Zhou N, Compagno C, Piškur J. **2014**. Why, when, and how did yeast evolve alcoholic fermentation? *FEMS Yeast Research* **14**: 826–832.

Dato L, Berterame NM, Ricci MA, *et al.* 2014. Changes in SAM2 expression affect lactic acid tolerance and lactic acid production in *Saccharomyces cerevisiae*. *Microbial Cell Factories* **13**: 1–18.

Dean EJ, Davis JC, Davis RW, Petrov DA. **2008**. Pervasive and persistent redundancy among duplicated genes in yeast. *PLoS Genetics* **4**(7):e1000113.

Defoort J, Van De Peer Y, Carretero-Paulet L, Golding B. **2019**. The evolution of gene duplicates in angiosperms and the impact of protein-protein interactions and the mechanism of duplication. *Genome Biology and Evolution* **11**: 2292–2305.

Dejean L, Beauvoit B, Guérin B, Rigoulet M. **2000**. Growth of the yeast *Saccharomyces cerevisiae* on a non-fermentable substrate: Control of energetic yield by the amount of mitochondria. *Biochimica et Biophysica Acta - Bioenergetics* **1457**: 45–56.

DeLuna A, Vetsigian K, Shoresh N, *et al.* 2008. Exposing the fitness contribution of duplicated genes. *Nature Genetics* **40**: 676–681.

Dennis MY, Eichler EE. **2016**. Human adaptation and evolution by segmental duplication. *Current Opinion in Genetics and Development* **41**: 44–52.

DeRisi JL, Iyer VR, Brown PO. **1997**. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278**: 680–686.

Dermitzakis ET, Clark AG. **2001**. Differential selection after duplication in mammalian developmental genes. *Molecular Biology and Evolution* **18**: 557–562.

Dey KK, Xie D, Stephens M. **2018**. A new sequence logo plot to highlight enrichment and depletion. *BMC Bioinformatics* **19**: 473.

Dhar R, Sägesser R, Weikert C, Wagner A. **2013**. Yeast adapts to a changing stressful environment by evolving cross-protection and anticipatory gene regulation. *Molecular Biology and Evolution* **30**: 573–588.

Van Dijk EL, Chen CL, Daubenton-Carafa Y, *et al.* 2011. XUTs are a class of Xrn1-sensitive antisense regulatory non-coding RNA in yeast. *Nature* **475**: 114–119.

Ding J, Huang X, Zhang L, Zhao N, Yang D, Zhang K. **2009**. Tolerance and stress response to ethanol in the yeast *Saccharomyces cerevisiae*. *Applied Microbiology and Biotechnology* **85**: 253–263.

Dingens AS, Haddox HK, Overbaugh J, Bloom JD. **2017**. Comprehensive mapping of HIV-1 escape from a broadly neutralizing antibody. *Cell Host and Microbe* **21**: 777-787.e4.

Dittmar K, Liberles D. **2010**. *Evolution after gene duplication*. Wiley-Blackwell, Hoboken, NJ.

Doud MB, Bloom JD. **2016**. Accurate measurement of the effects of all amino-acid mutations on influenza hemagglutinin. *Viruses* **8**: 1–17.

Doud MB, Hensley SE, Bloom JD. **2017**. Complete mapping of viral escape from neutralizing antibodies. *PLoS Pathogens* **13**: 1–20.

Draghi JA, Parsons TL, Wagner GP, Plotkin JB. **2010**. Mutational robustness can facilitate adaptation. *Nature* **463**: 353–355.

Dragovich PS, Prins TJ, Zhou R, *et al.* 2003. Structure-based design, synthesis, and biological evaluation of irreversible human rhinovirus 3C protease inhibitors. 8. Pharmacological optimization of orally bioavailable 2-pyridone-containing peptidomimetics. *Journal of Medicinal Chemistry* **46**: 4572–4585.

Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH. **2005**. Why highly expressed proteins evolve slowly. *Proceedings of the National Academy of Sciences of the United States of America* **102**: 14338–14343.

Drummond DA, Raval A, Wilke CO. **2006**. A single determinant dominates the rate of yeast protein evolution. *Molecular Biology and Evolution* **23**: 327–337.

Drummond DA, Wilke CO. **2008**. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* **134**: 341–352.

Du Y, Wu NC, Jiang L, *et al.* 2016. Annotating protein functional residues by coupling high-throughput fitness profile and homologous-structure analysis.*mBio* **7**(6):e01801-16

Duan SF, Shi JY, Yin Q, *et al.* 2019. Reverse evolution of a classic gene network in yeast offers a competitive advantage. *Current Biology* **29**: 1126-1136.e5.

Dušková M, Ferreira C, Lucas C, Sychrová H. **2015**. Two glycerol uptake systems contribute to the high osmotolerance of *Zygosaccharomyces rouxii*. *Molecular Microbiology* **97**: 541–559.

Emanuelle S, Doblin MS, Stapleton DI, Bacic A, Gooley PR. **2016**. Molecular insights into the enigmatic metabolic regulator, SnRK1. *Trends in Plant Science* **21**: 341–353.

Escalera-Fanjul X, Quezada H, Riego-Ruiz L, González A. **2019**. Whole-genome duplication and yeast's fruitful way of life. *Trends in Genetics* **35**: 42–54.

Fares MA. **2015a**. *Natural Selection. Methods and applications*. CRC Press, London.

Fares MA. **2015b**. The origins of mutational robustness. *Trends in Genetics* **31**: 373–

381.

Fares MA. **2015c**. Survival and innovation: The role of mutational robustness in evolution. *Biochimie* **119**: 254–261.

Fares MA, Byrne KP, Wolfe KH. **2006**. Rate asymmetry after genome duplication causes substantial long-branch attraction artifacts in the phylogeny of *Saccharomyces species*. *Molecular Biology and Evolution* **23**: 245–253.

Fares MA, Keane OM, Toft C, Carretero-Paulet L, Jones GW. **2013**. The roles of whole-genome and small-scale duplications in the functional specialization of *saccharomyces cerevisiae* genes. *PLoS Genetics* **9**.

Favor AH, Llanos CD, Youngblut MD, Bardales JA. **2020**. Optimizing bacteriophage engineering through an accelerated evolution platform. *Scientific Reports* **10**.

Fazio A, Jewett MC, Daran-Lapujade P, *et al.* 2008. Transcription factor control of growth rate dependent genes in *Saccharomyces cerevisiae*: A three factor design. *BMC Genomics* **9**: 341.

Félix MA, Wagner A. **2008**. Robustness and evolution: Concepts, insights and challenges from a developmental model system. *Heredity* **100**: 132–140.

Ferea TL, Botstein D, Brown PO, Rosenzweig RF. **1999**. Systematic changes in gene expression patterns following adaptive evolution in yeast. *Proceedings of the National Academy of Sciences of the United States of America* **96**: 9721–9726.

Fernandez-Escamilla AM, Rousseau F, Schymkowitz J, Serrano L. **2004**. Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nature Biotechnology* **22**: 1302–1306.

Ferris SD, Whitt GS. **1979**. Evolution of the differential regulation of duplicate genes after polyploidization. *Journal of Molecular Evolution* **12**: 267–317.

Fields BN, Knipe DM, Howley PM. **2013**. *Fields virology*. Wolters Kluwer Health/Lippincott Williams & Wilkins.

Fisher RA. **1930**. *The genetical theory of Natural Selection*. Oxford University Press, Oxford.

Fletcher E, Feizi A, Bisschops MMM, *et al.* 2017. Evolutionary engineering reveals divergent paths when yeast is adapted to different acidic environments. *Metabolic Engineering* **39**: 19–28.

Flibotte S, Edgley ML, Chaudhry I, *et al.* 2010. Whole-genome profiling of mutagenesis in *Caenorhabditis elegans*. *Genetics* **185**: 431–441.

Force A, Lynch M, Pickett F., Amores A, Yan YL, Postlethwait J. **1999**. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151**: 1531–1545.

Force A, Lynch M, Postlethwait J. **1999**. Preservation of duplicate genes by subfunctionalization. *American Zoologist* **39**: 0.

Fowler DM, Fields S. **2014**. Deep mutational scanning: A new style of protein science. *Nature Methods* **11**: 801–807.

Francino MP. **2005**. An adaptive radiation model for the origin of new gene functions. *Nature Genetics* **37**: 573–578.

Freeling M. **2009**. Bias in plant gene content following different sorts of duplication:

tandem, whole-genome, segmental, or by transposition. *Annual Review of Plant Biology* **60**: 433–453.

Freeling M, Thomas BC. **2006**. Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. *Genome Research* **16**: 805–814.

Fritz MA, Rosa S, Sicard A. **2018**. Mechanisms underlying the environmentally induced plasticity of leaf morphology. *Frontiers in Genetics* **9**.

Galardini M, Busby BP, Vieitez C, Dunham AS, Typas A, Beltrao P. **2019**. The impact of the genetic background on gene deletion phenotypes in *Saccharomyces cerevisiae* . *Molecular Systems Biology* **15**.

Gancedo JM. **1998**. Yeast carbon catabolite repression. *Microbiology and Molecular Biology Reviews* **62**: 334–361.

García-Ríos E, Morard M, Parts L, Liti G, Guillamón JM. **2017**. The genetic architecture of low-temperature adaptation in the wine yeast *Saccharomyces cerevisiae*. *BMC genomics* **18**: 159.

Garmaroudi FS, Marchant D, Hendry R, *et al.* 2015. Coxsackievirus B3 replication and pathogenesis. *Future Microbiology* **10**: 629–652.

Gasch AP, Spellman PT, Kao CM, *et al.* 2000. Genomic expression programs in the response of yeast cells to environmental changes. *Molecular Biology of the Cell* **11**: 4241–4257.

Geller R, Estada Ú, Peris JB, *et al.* 2016. Highly heterogeneous mutation rates in the hepatitis C virus genome. *Nature microbiology* **1**: 16045.

Geller R, Vignuzzi M, Andino R, Frydman J. **2007**. Evolutionary constraints on chaperone-mediated folding provide an antiviral approach refractory to development of drug resistance. *Genes and Development* **21**: 195–205.

Geng P, Zhang L, Shi GY. **2017**. Omics analysis of acetic acid tolerance in *Saccharomyces cerevisiae*. *World Journal of Microbiology and Biotechnology* **33**.

Gervasi DDL, Schiestl FP. **2017**. Real-time divergent evolution in plants driven by pollinators. *Nature Communications* **8**.

Ghillebert R, Swinnen E, Wen J, *et al.* 2011. The AMPK/SNF1/SnRK1 fuel gauge and energy regulator: Structure, function and regulation. *FEBS Journal* **278**: 3978–3990.

Giaever G, Chu AM, Ni L, *et al.* 2002. Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* **418**: 387–391.

Giaever G, Nislow C. **2014**. The yeast deletion collection: A decade of functional genomics. *Genetics* **197**: 451–465.

Gibson BR, Lawrence SJ, Boulton CA, *et al.* 2008. The oxidative stress response of a lager brewing yeast strain during industrial propagation and fermentation. *FEMS Yeast Research* **8**: 574–585.

Gibson TJ, Spring J. **1998**. Genetic redundancy in vertebrates: Polyploidy and persistence of genes encoding multidomain proteins. *Trends in Genetics* **14**: 46–49.

Gietz RD. **2014**. Yeast transformation by the LiAc/SS carrier DNA/PEG method. *Methods in Molecular Biology* **1163**: 33–44.

Gietz RD, Schiestl RH. **2007**. Quick and easy yeast transformation using the LiAc/SS carrier DNA/PEG method. *Nature Protocols* **2**: 35–37.

Gnädig NF, Beaucourt S, Campagnola G, *et al.* 2012. Coxsackievirus B3 mutator strains are attenuated in vivo. *Proceedings of the National Academy of Sciences of the United States of America* **109**: E2294-303.

Goffeau A, Barrell G, Bussey H, *et al.* 1996. Life with 6000 genes. *Science* **274**: 546–567.

Goldman N, Yang Z. **1994**. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular Biology and Evolution* **11**: 725–36.

González-Ramos D, Gorter De Vries AR, Grijseels SS, *et al.* 2016. A new laboratory evolution approach to select for constitutive acetic acid tolerance in *Saccharomyces cerevisiae* and identification of causal mutations. *Biotechnology for Biofuels* **9**: 1–18.

Gorman MJ, Caine EA, Zaitsev K, *et al.* 2018. An immunocompetent mouse model of Zika virus infection. *Cell host & microbe* **23**: 672-685.e6.

Gould SJ, Vrba ES. **1982**. Exaptation—a missing term in the science of form. *Paleobiology* **1**: 4–15.

Gout JF, Duret L, Kahn D. **2009**. Differential retention of metabolic genes following whole-genome duplication. *Molecular Biology and Evolution* **26**: 1067–1072.

Gout JF, Kahn D, Duret L. **2010**. The relationship among gene expression, the evolution of gene dosage, and the rate of protein evolution. *PLoS Genetics* **6**: 20.

Gout JF, Lynch M. **2015**. Maintenance and loss of duplicated genes by dosage subfunctionalization. *Molecular Biology and Evolution* **32**: 2141–2148.

Graci JD, Gnadig NF, Galarraga JE, Castro C, Vignuzzi M, Cameron CE. **2012**. Mutational robustness of an RNA virus influences sensitivity to lethal mutagenesis. *Journal of Virology* **86**: 2869–2873.

Grant SGN. **2016**. The molecular evolution of the vertebrate behavioural repertoire individual behavioural responses was articulated in the nineteenth century. *Philosophical Transactions of the Royal Society B: Biological Sciences* **371**: 0–8.

Gray VE, Hause RJ, Fowler DM. **2017**. Analysis of large-scale mutagenesis data to assess the impact of single amino acid substitutions. *Genetics* **207**: 53–61.

Green SA, Bronner ME. **2013**. Gene duplications and the early evolution of neural crest development. *Seminars in Cell and Developmental Biology* **24**: 95–100.

Gresham D, Desai MM, Tucker CM, *et al.* 2008. The repertoire and dynamics of evolutionary adaptations to controlled nutrient-limited environments in yeast. *PLoS Genetics* **4**(12):e1000303.

Grubaugh ND, Andersen KG. **2017**. Experimental evolution to study virus emergence. *Cell* **169**: 1–3.

Gu Z, Nicolae D, Lu HHS, Li WH. **2002**. Rapid divergence in expression between duplicate genes inferred from microarray data. *Trends in Genetics* **18**: 609–613.

Guan Y, Dunham MJ, Troyanskaya OG. **2007**. Functional analysis of gene duplications in *Saccharomyces cerevisiae*. *Genetics* **175**: 933–943.

Guan L, Gao Y, Li J, *et al.* 2020. Directed evolution of pseudomonas fluorescens lipase variants with improved thermostability using error-prone PCR. *Frontiers in Bioengineering and Biotechnology* **8**.

Gutierrez B, Escalera-Zamudio M, Pybus OG. **2019**. Parallel molecular evolution and

adaptation in viruses. *Current Opinion in Virology* **34**: 90–96.

Ha M, Kim ED, Chen ZJ. **2009**. Duplicate genes increase expression diversity in closely related species and allopolyploids. *Proceedings of the National Academy of Sciences of the United States of America* **106**: 2295–2300.

Ha M, Li WH, Chen ZJ. **2007**. External factors accelerate expression divergence between duplicate genes. *Trends in Genetics* **23**: 162–166.

Haas R, Horev G, Lipkin E, *et al.* 2019. Mapping ethanol tolerance in budding yeast reveals high genetic variation in a wild isolate. *Frontiers in Genetics* **10**.

Haddox HK, Dingens AS, Bloom JJD, *et al.* 2016. Experimental estimation of the effects of all amino-acid mutations to HIV's envelope protein on viral replication in cell culture. *PLOS Pathogens* **12**: e1006114.

Haddox HK, Dingens AS, Hilton SK, Overbaugh J, Bloom JD. **2018**. Mapping mutational effects along the evolutionary landscape of HIV envelope. *eLife* **7**: e34420.

Hagman A, Säll T, Compagno C, Piskur J. **2013**. Yeast "Make-Accumulate-Consume" life strategy evolved as a multi-step process that predates the whole genome duplication. *PLoS ONE* **8**(7):e68734.

Hakes L, Pinney JW, Lovell SC, Oliver SG, Robertson DL. **2007**. All duplicates are not equal: The difference between small-scale and genome duplication. *Genome Biology* **8**.

Halabi N, Rivoire O, Leibler S, Ranganathan R. **2009**. Protein sectors: evolutionary units of three-dimensional structure. *Cell* **138**: 774–786.

Haldane JB. 1932. *The causes of evolution*. New York: Longmans, Green and Co.

Hall BG, Acar H, Nandipati A, Barlow M. **2014**. Growth rates made easy. *Molecular Biology and Evolution* **31**: 232–238.

Hallin J, Landry CR. **2019**. Regulation plays a multifaceted role in the retention of gene duplicates. *PLoS Biology* **17**(11):e3000519.

Hallsworth JE. **1998**. Ethanol-induced water stress in yeast. *Journal of Fermentation and Bioengineering* **85**: 125–137.

Hallsworth JE, Nomura Y. **1999**. A simple method to determine the water activity of ethanol-containing samples. *Biotechnology and Bioengineering* **62**: 242–245.

Hallsworth JE, Yakimov MM, Golyshin PN, *et al.* 2007. Limits of life in MgCl2-containing environments: Chaotropicity defines the window. *Environmental Microbiology* **9**: 801–813.

Harrison SC. **2013**. Principles of virus structure In: Knipe DM, Howley PM, eds. *Field Virology*. Wolters Kluwer Health/Lippincott Williams & Wilkins, 52–86.

Hartman EC, Jakobson CM, Favor AH, *et al.* 2018. Quantitative characterization of all single amino acid variants of a viral capsid-based drug delivery vehicle. *Nature Communications* **9**: 1–11.

He X, Zhang J. **2005**. Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics* **169**: 1157–1164.

Hecht M, Bromberg Y, Rost B. **2015**. Better prediction of functional effects for sequence variants. *BMC Genomics* **16**: S1.

Heise MT, Virgin HW. **2013**. Pathogenesis of viral infection In: Knipe DM, Howley PM,

eds. *Fields Virology*. Wolters Kluwer Health/Lippincott Williams & Wilkins, 254–285.

Helenius A. **2013**. Virus entry and uncoating In: Knipe DM, Howley PM, eds. *Fields Virology*. Wolters Kluwer Health/Lippincott Williams & Wilkins, 87–104.

Hilton SK, Doud MB, Bloom JD. **2017**. Phydms : software for phylogenetic analyses informed by deep mutational scanning. *PeerJ* **5**: e3657.

Hittinger CT, Carroll SB. **2007**. Gene duplication and the adaptive evolution of a classic genetic switch. *Nature* **449**: 677–681.

Ho W-C, Ohya Y, Zhang J. **2017**. Testing the neutral hypothesis of phenotypic evolution. *Proceedings of the National Academy of Sciences* **114**: 12219–12224.

Ho WC, Zhang J. **2014**. The genotype-phenotype map of yeast complex traits: Basic parameters and the role of natural selection. *Molecular Biology and Evolution* **31**: 1568–1580.

Hoang KL, Morran LT, Gerardo NM. **2016**. Experimental evolution as an underutilized tool for studying beneficial animal-microbe interactions. *Frontiers in Microbiology* **7**.

Hoegg S, Brinkmann H, Taylor JS, Meyer A. **2004**. Phylogenetic timing of the fish-specific genome duplication correlates with the diversification of teleost fish. *Journal of Molecular Evolution* **59**: 190–203.

Van Hoek MJA, Hogeweg P. **2009**. Metabolic adaptation after whole genome duplication. *Molecular Biology and Evolution* **26**: 2441–2453.

Hoffmann FG, Opazo JC, Hoogewijs D, *et al.* 2012. Evolution of the globin gene family in deuterostomes: Lineage-specific patterns of diversification and attrition. *Molecular Biology and Evolution* **29**: 1735–1745.

Hoffmann FG, Opazo JC, Storz JF. **2011**. Differential loss and retention of cytoglobin, myoglobin, and globin-E during the radiation of vertebrates. *Genome Biology and Evolution* **3**: 588–600.

Hoffmann FG, Opazo JC, Storz JF. **2012**. Whole-genome duplications spurred the functional diversification of the globin gene superfamily in vertebrates. *Molecular Biology and Evolution* **29**: 303–312.

Hohmann S. **2015**. An integrated view on a eukaryotic osmoregulation system. *Current Genetics* **61**: 373–382.

Hohmann S, Krantz M, Nordlander B. **2007**. Yeast osmoregulation In: *Methods in Enzymology*. Academic Press Inc., 29–45.

Holub EB, Houb EB, Holub EB, Houb EB. **2001**. The arms race is ancient history in Arabidopsis, the wildflower. *Nature Reviews Genetics* **2**: 516–527.

Hom N, Gentles L, Bloom JD, Lee KK. **2019**. Deep mutational scan of the highly conserved influenza A virus M1 matrix protein reveals substantial intrinsic mutational tolerance. *Journal of Virology* **93**: 1–16.

Hosseini SR, Wagner A. **2016**. The potential for non-adaptive origins of evolutionary innovations in central carbon metabolism. *BMC Systems Biology* **10**: 1–14.

Huang D, Friesen H, Andrews B. **2007**. Pho85, a multifunctional cyclin-dependent protein kinase in budding yeast. *Molecular Microbiology* **66**: 303–314.

Hubmann G, Guillouet S, Nevoigt E. **2011**. Gpd1 and Gpd2 fine-tuning for sustainable reduction of glycerol formation in *Saccharomyces cerevisiae*. *Applied and Environmental*

Microbiology **77**: 5857–5867.

Huminiecki L, Conant GC. **2012**. Polyploidy and the evolution of complex traits. *International Journal of Evolutionary Biology* **2012**: 1–12.

Huminiecki L, Wolfe KH. **2004**. Divergence of spatial gene expression profiles following species-specific gene duplications in human and mouse. *Genome Research* **14**: 1870–1879.

Hunter E. **2013**. Virus assembly In: Knipe DM, Howley PM, eds. *Fields Virology*. Wolters Kluwer Health/Lippincott Williams & Wilkins, 127–152.

Ideker T, Thorsson V, Ranish JA, *et al.* 2001. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* **292**: 929–934.

Ihmels J, Collins SR, Schuldiner M, Krogan NJ, Weissman JS. **2007**. Backup without redundancy: Genetic interactions reveal the cost of duplicate gene loss. *Molecular Systems Biology* **3**.

Ilyushina NA, Khalenkov AM, Seiler JP, *et al.* 2010. Adaptation of pandemic H1N1 influenza viruses in mice. *Journal of Virology* **84**: 8607–8616.

Innan H, Kondrashov F. **2010**. The evolution of gene duplications: Classifying and distinguishing between models. *Nature Reviews Genetics* **11**: 97–108.

Izawa S, Inoue Y. **2009**. Post-transcriptional regulation of gene expression in yeast under ethanol stress. *Biotechnology and Applied Biochemistry* **53**: 93.

Jagdeo JM, Dufour A, Klein T, *et al.* 2018. N-Terminomics TAILS identifies host cell substrates of poliovirus and coxsackievirus B3 3C proteinases that modulate virus infection. *Journal of Virology* **92**: e02211-17.

Jaillon O, Aury JM, Wincker P. **2009**. “Changing by doubling”, the impact of Whole Genome Duplications in the evolution of eukaryotes. *Comptes Rendus - Biologies* **332**: 241–253.

Jander G, Baerson SR, Hudak JA, Gonzalez KA, Gruys KJ, Last RL. **2003**. Ethylmethanesulfonate saturation mutagenesis in Arabidopsis to determine frequency of herbicide resistance. *Plant Physiology* **131**: 139–146.

Jarolim S, Ayer A, Pillay B, *et al.* 2013. *Saccharomyces cerevisiae* genes involved in survival of heat shock. *G3: Genes, Genomes, Genetics* **3**: 2321–2333.

Jeffares DC. **2018**. The natural diversity and ecology of fission yeast. *Yeast* **35**: 253–260.

Jensen JD, Payseur BA, Stephan W, *et al.* 2019. The importance of the Neutral Theory in 1968 and 50 years on: A response to Kern and Hahn 2018. *Evolution* **73**: 111–114.

Jiang X, Assis R. **2019**. Rapid functional divergence after small-scale gene duplication in grasses. *BMC Evolutionary Biology* **19**: 97.

Jiang P, Liu Y, Ma H-C, Paul A V., Wimmer E. **2014**. Picornavirus morphogenesis. *Microbiology and Molecular Biology Reviews* **78**: 418–437.

Jin S, Nasim Z, Susila H, Ahn JH. **2020**. Evolution and functional diversification of FLOWERING LOCUS T/TERMINAL FLOWER 1 family genes in plants. *Seminars in Cell and Developmental Biology*.

Kaboli S, Miyamoto T, Sunada K, Sasano Y, Sugiyama M, Harashima S. **2016**. Improved stress resistance and ethanol production by segmental haploidization of the

diploid genome in *Saccharomyces cerevisiae*. *Journal of Bioscience and Bioengineering* **121**: 638–644.

Kalia V, Sarkar S, Gupta P, Montelaro RC. **2005**. Antibody Neutralization escape mediated by point mutations in the intracytoplasmic tail of human immunodeficiency virus type 1 gp41. *Journal of Virology* **79**: 2097–2107.

Kautz TF, Forrester NL. **2018**. RNA virus fidelity mutants: a useful tool for evolutionary biology or a complex challenge? *Viruses* **10**: 1–17.

Ke C, Guan W, Bu S, *et al.* 2019. Determination of absorption dose in chemical mutagenesis in plants. *PLoS ONE* **14**.

Keane OM, Toft C, Carretero-Paulet L, Jones GW, Fares MA. **2014**. Preservation of genetic and regulatory robustness in ancient gene duplicates of *Saccharomyces cerevisiae*. *Genome Research* **24**: 1830–1841.

Kelly SA, Panhuis TM, Stoehr AM. **2012**. Phenotypic plasticity: Molecular mechanisms and adaptive significance. *Comprehensive Physiology* **2**: 1417–1439.

Kempf BJ, Peersen OB, Barton DJ. **2016**. Poliovirus polymerase Leu420 facilitates RNA recombination and ribavirin resistance. *Journal of Virology* **90**: 8410–8421.

Kennedy SR, Schmitt MW, Fox EJ, *et al.* 2014. Detecting ultralow-frequency mutations by Duplex Sequencing. *Nature Protocols* **9**: 2586–2606.

Kern AD, Hahn MW. **2018**. The neutral theory in light of natural selection. *Molecular Biology and Evolution* **35**: 1366–1371.

Kim S, Yoo M-J, Albert VA, Farris JS, Soltis PS, Soltis DE. **2004**. Phylogeny and diversification of B-function MADS-box genes in angiosperms: evolutionary and functional implications of a 260-million-year-old duplication. *American Journal of Botany* **91**: 2102–2118.

Kimura M. **1968**. Evolutionary Rate at the Molecular Level by. *Nature* **217**: 624–626.

Kimura M. **1983**. *The neutral theory of molecular evolution*. Cambridge Univ. Press, Cambridge, MA.

Kiyosawa K. **1991**. Volumetric properties of polyols (ethylene glycol, glycerol, meso-erythritol, xylitol and mannitol) in relation to their membrane permeability: Group additivity and estimation of the maximum radius of their molecules. *BBA - Biomembranes* **1064**: 251–255.

Klesmith JR, Bacik JP, Wrenbeck EE, Michalczyk R, Whitehead TA. **2017**. Trade-offs between enzyme fitness and solubility illuminated by deep mutational scanning. *Proceedings of the National Academy of Sciences of the United States of America* **114**: 2265–2270.

Kodigepalli KM, Bowers K, Sharp A, Nanjundan M. **2015**. Roles and regulation of phospholipid scramblases. *FEBS Letters* **589**: 3–14.

Krakauer DC, Plotkin JB. **2002**. Redundancy, antiredundancy, and the robustness of genomes. *Proceedings of the National Academy of Sciences* **99**: 1405–1409.

Kristensen T, Belsham GJ. **2019**. Identification of a short, highly conserved, motif required for picornavirus capsid precursor processing at distal sites (BL Semler, Ed.). *PLOS Pathogens* **15**: e1007509.

Krylov DM, Wolf YI, Rogozin IB, Koonin E V. 2003. Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in

eukaryotic evolution. *Genome Research* **13**: 2229–2235.

Krystkowiak I, Manguy J, Davey NE. **2018**. PSSMSearch: A server for modeling, visualization, proteome-wide discovery and annotation of protein motif specificity determinants. *Nucleic Acids Research* **46**: W235–W241.

Lafuente E, Beldade P. **2019**. Genomics of developmental plasticity in animals. *Frontiers in Genetics* **10**.

Laitinen OH, Svedin E, Kapell S, Nurminen A, Hytönen VP, Flodström-Tullberg M. **2016**. Enteroviral proteases: structure, host interactions and pathogenicity. *Reviews in Medical Virology* **26**: 251–267.

Landry CR, Lemos B, Rifkin SA, Dickinson WJ, Hartl DL. **2007**. Genetic properties influencing the evolvability of gene expression. *Science* **317**: 118–121.

Landry CR, Oh J, Hartl DL, Cavalieri D. **2006**. Genome-wide scan reveals that genetic variation for transcriptional plasticity in yeast is biased towards multi-copy and dispensable genes. *Gene* **366**: 343–351.

Larson RT, Dacks JB, Barlow LD. **2019**. Recent gene duplications dominate evolutionary dynamics of adaptor protein complex subunits in embryophytes. *Traffic (Copenhagen, Denmark)*: 1–13.

Lauring AS, Acevedo A, Cooper SB, Andino R. **2012**. Codon usage determines the mutational robustness, evolutionary capacity, and virulence of an RNA virus. *Cell Host and Microbe* **12**: 623–632.

Lee JM, Eguia R, Zost SJ, *et al.* 2019. Mapping person-to-person variation in viral mutations that escape polyclonal serum targeting influenza hemagglutinin. *eLife* **8**:e49324.

Leeks A, Sanjuán R, West SA. **2019**. The evolution of collective infectious units in viruses. *Virus Research* **265**: 94–101.

Legras JL, Galeote V, Bigey F, *et al.* 2018. Adaptation of *S. cerevisiae* to fermented food environments reveals remarkable genome plasticity and the footprints of domestication. *Molecular Biology and Evolution* **35**: 1712–1727.

Lehner B. **2010**. Conflict between noise and plasticity in yeast. *PLoS Genetics* **6**(11):e1001185.

Lenski RE, Wiser MJ, Ribeck N, *et al.* 2015. Sustained fitness gains and variability in fitness trajectories in the long-term evolution experiment with *Escherichia coli. Proceedings of the Royal Society B: Biological Sciences* **282**(1821):20152292.

Lespinet O, Wolf YI, Koonin E V., Aravind L. **2002**. The role of lineage-specific gene family expansion in the evolution of eukaryotes. *Genome Research* **12**: 1048–1059.

Li K, Wohlford-Lenane CL, Channappanavar R, *et al.* 2017. Mouse-adapted MERS coronavirus causes lethal lung disease in human DPP4 knockin mice. *Proceedings of the National Academy of Sciences of the United States of America* **114**: E3119–E3128.

Li WH, Yang J, Gu X. **2005**. Expression divergence between duplicate genes. *Trends in Genetics* **21**: 602–607.

Lievens B, Hallsworth JE, Pozo MI, *et al.* 2015. Microbiology of sugar-rich environments: Diversity, ecology and system constraints. *Environmental Microbiology* **17**: 278–298.

de Lima Alves F, Stevenson A, Baxter E, *et al.* 2015. Concomitant osmotic and

chaotropicity-induced stresses in Aspergillus wentii: compatible solutes determine the biotic window. *Current Genetics* **61**: 457–477.

Lin Y, Golovnina K, Chen Z-X, *et al.* 2016. Comparison of normalization and differential expression analyses using RNA-Seq data from 726 individual *Drosophila melanogaster*. *BMC Genomics* **17**: 28.

Linde J, Duggan S, Weber M, *et al.* 2015. Defining the transcriptomic landscape of *Candida glabrata* by RNA-Seq. *Nucleic Acids Research* **43**: 1392–1406.

Liu B, Li Z, Xiang F, Li F, Zheng Y, Wang G. **2014**. The whole genome sequence of Coxsackievirus B3 MKP strain leading to myocarditis and its molecular phylogenetic analysis. *Virology Journal* **11**: 33.

Livesey BJ, Marsh JA. **2020**. Using deep mutational scanning to benchmark variant effect predictors and identify disease mutations. *Molecular Systems Biology* **16**.

Lodi T, Fontanesi F, Guiard B. **2002**. Co-ordinate regulation of lactate metabolism genes in yeast: The role of the lactate permease gene JEN1. *Molecular Genetics and Genomics* **266**: 838–847.

Lohse M, Bolger AM, Nagel A, *et al.* 2012. RobiNA: A user-friendly, integrated software solution for RNA-Seq-based transcriptomics. *Nucleic Acids Research* **40**: 622–627.

Looby P, Loudon ASI. **2005**. Gene duplication and complex circadian clocks in mammals. *Trends in Genetics* **21**: 46–53.

Lopandic K. **2018**. Saccharomyces interspecies hybrids as model organisms for studying yeast adaptation to stressful environments. *Yeast* **35**: 21–38.

López-Maury L, Marguerat S, Bähler J. **2008**. Tuning gene expression to changing environments: From rapid responses to evolutionary adaptation. *Nature Reviews Genetics* **9**: 583–593.

Lynch M, Conery JS. **2000**. The evolutionary fate and consequences of duplicate genes. *Science* **290**: 1151–1155.

Lynch M, Conery JS. **2003**. The evolutionary demography of duplicate genes. *Journal of Structural and Functional Genomics* **3**: 35–44.

Lynch M, Katju V. **2004**. The altered evolutionary trajectories of gene duplicates. *Trends in Genetics* **20**: 544–549.

Lyons DM, Lauring AS. **2017**. Evidence for the selective basis of transition-to-transversion substitution bias in two RNA viruses. *Molecular biology and evolution* **34**: 3205–3215.

Macejak DG, Sarnow P. **1992**. Association of heat shock protein 70 with enterovirus capsid precursor P1 in infected human cells. *Journal of virology* **66**: 1520–7.

Macías LG, Morard M, Toft C, Barrio E. **2019**. Comparative genomics between *Saccharomyces kudriavzevii* and *S. cerevisiae* applied to identify mechanisms involved in adaptation. *Frontiers in Genetics* **10**: 1–11.

Maeshiro T, Kimura M. **1998**. The role of robustness and changeability on the origin and evolution of genetic codes. *Proceedings of the National Academy of Sciences of the United States of America* **95**: 5088–5093.

Des Marais DL, Rausher MD. **2008**. Escape from adaptive conflict after duplication in an anthocyanin pathway gene. *Nature* **454**: 762–765.

Marcet-Houben M, Gabaldón T. **2015**. Beyond the whole-genome duplication: Phylogenetic evidence for an ancient interspecies hybridization in the baker's yeast lineage. *PLoS Biology* **13**: 1–26.

Markiewicz-Potoczny M, Lydall D. **2016**. Costs, benefits and redundant mechanisms of adaption to chronic low-dose stress in yeast. *Cell Cycle* **15**: 2732–2741.

Martin DE, Hall MN. **2005**. The expanding TOR signaling network. *Current Opinion in Cell Biology* **17**: 158–166.

Mattenberger F, Latorre V, Tirosh O, Stern A, Geller R. **2020**. Globally defining the effects of mutations in a picornavirus capsid. *bioRxiv*: 2020.10.06.327916.

Mattenberger F, Sabater-Muñoz B, Hallsworth JE, Fares MA. **2017**. Glycerol stress in *Saccharomyces cerevisiae*: Cellular responses and evolved adaptations. *Environmental Microbiology* **19**: 990–1007.

Mattenberger F, Sabater-Muñoz B, Toft C, Fares MA. **2017**. The phenotypic plasticity of duplicated genes in *Saccharomyces cerevisiae* and the origin of adaptations. *G3 Genes|Genomes|Genetics* **7**: 63–75.

Mattenberger F, Sabater-Muñoz B, Toft C, Sablok G, Fares MA. **2017**. Expression properties exhibit correlated patterns with the fate of duplicated genes, their divergence, and transcriptional plasticity in *Saccharomycotina*. *DNA Research* **24**: 559–570.

McGrath CL, Gout JF, Doak TG, Yanagi A, Lynch M. **2014**. Insights into three whole-genome duplications gleaned from the *Paramecium caudatum* genome sequence. *Genetics* **197**: 1417–1428.

McGrath CL, Gout JF, Johri P, Doak TG, Lynch M. **2014**. Differential retention and divergent resolution of duplicate genes following whole-genome duplication. *Genome Research* **24**: 1665–1675.

Melamed D, Young DL, Gamble CE, Miller CR, Fields S. **2013**. Deep mutational scanning of an RRM domain of the *Saccharomyces cerevisiae* poly (A) -binding protein. *RNA* **19**: 1537–1551.

Minor PD. **2015**. Live attenuated vaccines: Historical successes and current challenges. *Virology* **479–480**: 379–392.

Mira NP, Teixeira MC, Sá-Correia I. **2010**. Adaptive response and tolerance to weak acids in *Saccharomyces cerevisiae*: A genome-wide view. *OMICS A Journal of Integrative Biology* **14**: 525–540.

Moraes TS, Dornelas MC, Martinelli AP. **2019**. FT/TFL1: Calibrating plant architecture. *Frontiers in Plant Science* **10**: 97.

Morard M, Macías LG, Adam AC, *et al.* 2019. Aneuploidy and ethanol tolerance in *Saccharomyces cerevisiae*. *Frontiers in Genetics* **10**.

Moratorio G, Henningsson R, Barbezange C, *et al.* 2017. Attenuation of RNA viruses by redirecting their evolution in sequence space. *Nature Microbiology* **2**.

Morgan AP, Matthew Holt J, McMullan RC, *et al.* 2016. The evolutionary fates of a large segmental duplication in mouse. *Genetics* **204**: 267–285.

Muckelbauer JK, Kremer M, Minor I, *et al.* 2004. The structure of coxsackievirus B3 at 3.5 å resolution. *Structure* **3**: 653–667.

Muller HJ. **1964**. The relation of recombination to mutational advance. *Mutation Research - Fundamental and Molecular Mechanisms of Mutagenesis* **1**: 2–9.

Mullis A, Lu Z, Zhan Y, *et al.* 2019. Parallel concerted evolution of ribosomal protein genes in fungi and its adaptive significance. *Molecular Biology and Evolution* **37**(2):455-468

Musso G, Costanzo M, Huangfu MQ, *et al.* 2008. The extensive and condition-dependent nature of epistasis among whole-genome duplicates in yeast. *Genome Research* **18**: 1092–1099.

De Nadal E, Zapater M, Alepuz PM, Sumoy L, Mas G, Posas F. **2004**. The MAPK Hog1 recruits Rpd3 histone deacetylase to activate osmoresponsive genes. *Nature* **427**: 370–374.

Nagalakshmi U, Wang Z, Waern K, *et al.* 2008. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320**: 1344–1349.

Narayanan V, Sanchez i Nogue V, van Niel EWJ, Gorwa-Grauslund MF. **2016**. Adaptation to low pH and lignocellulosic inhibitors resulting in ethanolic fermentation and growth of *Saccharomyces cerevisiae*. *AMB Express* **6**.

Nasvall J, Sun L, Roth JR, Andersson DI. **2012**. Real-time evolution of new genes by innovation, amplification, and divergence. *Science* **338**: 384–387.

Newman JRS, Ghaemmaghami S, Ihmels J, *et al.* 2006. Single-cell proteomic analysis of *S. cerevisiae* reveals the architecture of biological noise. *Nature* **441**: 840–846.

Nguyen Ba AN, Cvijović I, Rojas Echenique JI, *et al.* 2019. High-resolution lineage tracking reveals travelling wave of adaptation in laboratory yeast. *Nature* **575**: 494–499.

Nguyen Y, Jesudhasan PR, Aguilera ER, Pfeiffer JK. **2018**. Identification and characterization of a poliovirus capsid mutant with enhanced thermal stability. *Journal of Virology* **93**.

Nielsen R. **2005**. Molecular signatures of natural selection. *Annual Review of Genetics* **39**: 197–218.

Nookaew I, Papini M, Pornputtapong N, *et al.* 2012. A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: A case study in *Saccharomyces cerevisiae*. *Nucleic Acids Research* **40**: 10084–10097.

Ogden PJ, Kelsic ED, Sinai S, Church GM. **2019**. Comprehensive AAV capsid fitness landscape reveals a viral gene and enables machine-guided design. *Science* **366**: 1139–1143.

Ohler U, Niemann H. **2001**. Identification and analysis of eukaryotic promoters: Recent computational approaches. *Trends in Genetics* **17**: 56–60.

Ohno S. **1970**. *Evolution by gene duplication*. Springer-Verlag New York Inc.

Ohno S. **1999**. Gene duplication and the uniqueness of vertebrate genomes circa 1970-1999. *Seminars in Cell and Developmental Biology* **10**: 517–522.

Ortiz-Merino RA, Kuanyshev N, Byrne KP, *et al.* 2017. Transcriptional response to lactic acid stress in the hybrid yeast *Zygosaccharomyces parabailii*. *Applied and Environmental Microbiology* **84**: AEM.02294-17.

Otto SP, Whitton J. **2000**. Polyploid incidence and evolution. *Annu Rev Genet* **34**: 401–437.

Oz T, Guvenek A, Yildiz S, *et al.* 2014. Strength of selection pressure is an important parameter contributing to the complexity of antibiotic resistance evolution. *Molecular*

Biology and Evolution **31**: 2387–2401.

Pagan I, Holmes EC, Simon-Loriere E. **2012**. Level of gene expression is a major determinant of protein evolution in the viral order mononegavirales. *Journal of Virology* **86**: 5253–5263.

Pál C, Papp B, Hurst LD. **2001**. Highly expressed genes in yeast evolve slowly. *Genetics* **158**: 927–931.

Panchy N, Lehti-Shiu MD, Shiu S-HH. **2016**. Evolution of gene duplication in plants. *Plant Physiology* **171**: pp.00523.2016.

Papp B, Pál C, Hurst LD. **2002**. Dosage sensitivity and the evolution of gene families in yeast. *Nature* **424**(6945): 194-197.

Peck KM, Lauring AS. **2018**. Complexities of viral mutation rates. *Journal of Virology* **92**.

Van De Peer Y. **2004**. Computational approaches to unveiling ancient genome duplications. *Nature Reviews Genetics* **5**: 752–763.

Perales C, Domingo E. **2016**. Antiviral strategies based on lethal mutagenesis and error threshold In: Domingo E, Schuster P, eds. *Quasispecies: From Theory to Experimental Systems*. Cham: Springer International Publishing, 323–339.

Perlmutter JD, Hagan MF. **2015**. Mechanisms of virus assembly. *Annual Review of Physical Chemistry* **66**: 217–239.

Pickett BE, Sadat EL, Zhang Y, *et al.* 2012. ViPR: An open bioinformatics database and analysis resource for virology research. *Nucleic Acids Research* **40**: D593.

Price TD, Qvarnström A, Irwin DE. **2003**. The role of phenotypic plasticity in driving genetic evolution. *Proceedings of the Royal Society B: Biological Sciences* **270**: 1433–1440.

Pu S, Wong J, Turner B, Cho E, Wodak SJ. **2009**. Up-to-date catalogues of yeast protein complexes. *Nucleic Acids Research* **37**: 825–831.

Qian W, Liao BY, Chang AYF, Zhang J. **2010**. Maintenance of duplicate genes and their functional redundancy by reduced expression. *Trends in Genetics* **26**: 425–430.

Qing J, Wang Y, Sun Y, *et al.* 2014. Cyclophilin A associates with Enterovirus-71 virus capsid and plays an essential role in viral infection as an uncoating regulator. *PLoS Pathogens* **10**.

Racaniello VR. **2013**. Picornaviridae: The viruses and their replication In: Knipe DM, Howley PM, eds. *Fields Virology*. Wolters Kluwer Health/Lippincott Williams & Wilkins, 453–489.

Rainey PB, Remigi P, Farr AD, Lind PA. **2017**. Darwin was right: where now for experimental evolution? *Current Opinion in Genetics and Development* **47**: 102–109.

Raser JM, O'Shea EK. **2005**. Noise in gene expression: Origins, consequences, and control. *Science* **309**: 2010–2013.

Reeb J, Wirth T, Rost B. **2020**. Variant effect predictions capture some aspects of deep mutational scanning experiments. *BMC Bioinformatics* **21**:107.

Regenberg B, Grotkjær T, Winther O, *et al.* 2006. Growth-rate regulated genes have profound impact on interpretation of transcriptome profiling in *Saccharomyces cerevisiae*. *Genome Biology* **7**:R107.

Rensing SA. **2014**. Gene duplication as a driver of plant morphogenetic evolution. *Current Opinion in Plant Biology* **17**: 43–48.

Richards S, Gibbs RA, Gerardo NM, *et al.* 2010. Genome sequence of the pea aphid *Acyrthosiphon pisum*. *PLoS Biology* **8**: e1000313.

Roberts A, Deming D, Paddock CD, *et al.* 2007. A mouse-adapted SARS-Coronavirus causes disease and mortality in BALB/c mice. *PLoS Pathogens* **3**: e5.

Robinson MD, Mccarthy DJ, Smyth GK. **2010**. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**: 139–140.

Rocha EPC, Danchin A. **2004**. An analysis of determinants of amino acids substitution rates in bacterial proteins. *Molecular Biology and Evolution* **21**: 108–116.

Rodrigo G, Fares MA. **2018**. Intrinsic adaptive value and early fate of gene duplication revealed by a bottom-up approach. *eLife* **7**:e29739.

Rösner J, Merzendorfer H. **2019**. Transcriptional plasticity of different ABC transporter genes from *Tribolium castaneum* contributes to diflubenzuron resistance. *Insect Biochemistry and Molecular Biology*: 103282.

Rossmann MG, He Y, Kuhn RJ. **2002**. Picornavirus-receptor interactions. *Trends in Microbiology* **10**: 324–331.

Ruiz-González MX, Fares MA. **2013**. Coevolution analyses illuminate the dependencies between amino acid sites in the chaperonin system GroES-L. *BMC Evolutionary Biology* **13**.

Sabater-Muñoz B, Mattenberger F, Fares MA, Toft C. **2020**. Transcriptional rewiring, adaptation, and the role of gene duplication in the metabolism of ethanol of *Saccharomyces cerevisiae* . *mSystems* **5**:e00416-20.

Salas A. **2019**. The natural selection that shapes our genomes. *Forensic Science International: Genetics* **39**: 57–60.

Sandberg TE, Salazar MJ, Weng LL, Palsson BO, Feist AM. **2019**. The emergence of adaptive laboratory evolution as an efficient tool for biological discovery and industrial biotechnology. *Metabolic Engineering* **56**: 1–16.

Sanjuán R, Cuevas JM, Moya A, Elena SF. **2005**. Epistasis and the adaptability of an RNA virus. *Genetics* **170**: 1001–1008.

Sanjuán R, Domingo-Calap P. **2016**. Mechanisms of viral mutation. *Cellular and Molecular Life Sciences* **73**: 4433–4448.

Sanjuán R, Grdzelishvili VZ. **2015**. Evolution of oncolytic viruses. *Current Opinion in Virology* **13**: 1–5.

Scannell DR, Byrne KP, Gordon JL, Wong S, Wolfe KH. **2006**. Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. *Nature* **440**: 341–345.

Scannell DR, Wolfe KH. **2008**. A burst of protein sequence evolution and a prolonged period of asymmetric evolution follow gene duplication in yeast. *Genome Research* **18**: 137–147.

Schlichting CD. **1986**. The evolution of phenotypic plasticity in plants. *Annual Review of Ecology and Systematics* **17**: 667–693.

Schlicker A, Domingues FS, Rahnenführer J, Lengauer T. **2006**. A new measure for functional similarity of gene products based on gene ontology. *BMC Bioinformatics* **7**: 1–16.

Schmitt MW, Kennedy SR, Salk JJ, Fox EJ, Hiatt JB, Loeb LA. **2012**. Detection of ultra-rare mutations by next-generation sequencing. *Proceedings of the National Academy of Sciences* **109**: 14508–14513.

Schmutzer M, Wagner A. **2020**. Gene expression noise can promote the fixation of beneficial mutations in fluctuating environments. *PLOS Computational Biology* **16**: e1007727.

Schüller HJ. **2003**. Transcriptional control of nonfermentative metabolism in the yeast Saccharomyces cerevisiae. *Current Genetics* **43**: 139–160.

Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, Serrano L. **2005**. The FoldX web server: An online force field. *Nucleic Acids Research* **33**.

Seegers SL, Frasier C, Greene S, Nesmelova I V., Grdzelishvili VZ. **2019**. Experimental evolution generates novel oncolytic vesicular stomatitis viruses with improved replication in virus-resistant pancreatic cancer cells. *Journal of Virology* **94**.

Seoighe C, Wolfe KH. **1999**. Yeast genome evolution in the post-genome era. *Current Opinion in Microbiology* **2**: 548–554.

Shami Shah A, Batrouni AG, Kim D, *et al.* 2019. PLEKHA4/kramer attenuates dishevelled ubiquitination to modulate wnt and planar cell polarity signaling. *Cell Reports* **27**: 2157-2170.e8.

Shibai A, Takahashi Y, Ishizawa Y, *et al.* 2017. Mutation accumulation under UV radiation in Escherichia coli. *Scientific Reports* **7**: 1–12.

Shiomi H, Urasawa T, Urasawa S, Kobayashi N, Abe S, Taniguchi K. **2004**. Isolation and characterisation of poliovirus mutants resistant to heating at 50ºC for 30 min. *Journal of Medical Virology* **74**: 484–491.

Si W, Hang T, Guo M, *et al.* 2019. Whole-genome and transposed duplication contributes to the expansion and diversification of TLC genes in Maize. *International Journal of Molecular Sciences* **20**.

Simon-Loriere E, Holmes EC. **2011**. Why do RNA viruses recombine? *Nature Reviews Microbiology* **9**: 617–626.

Simon JC, Pfrender ME, Tollrian R, Tagu D, Colbourne JK. **2011**. Genomics of environmentally induced phenotypes in 2 extremely plastic arthropods. *Journal of Heredity* **102**: 512–525.

Simon JC, Rispe C, Sunnucks P. **2002**. Ecology and evolution of sex in aphids. *Trends in Ecology and Evolution* **17**: 34–39.

Sin J, Mangale V, Thienphrapa W, Gottlieb RA, Feuer R. **2015**. Recent progress in understanding coxsackievirus replication, dissemination, and pathogsenesis. *Virology* **484**: 288–304.

Snoek T, Verstrepen KJ, Voordeckers K. **2016**. How do yeast cells become tolerant to high ethanol concentrations? *Current Genetics* **62**: 475–480.

Soltis DE, Albert VA, Leebens-Mack J, *et al.* 2009. Polyploidy and angiosperm diversification. *American Journal of Botany* **96**: 336–348.

Sourisseau M, Lawrence DJP, Schwarz MC, *et al.* 2019. Deep mutational scanning

comprehensively maps how Zika envelope protein mutations affect viral growth and antibody escape. *Journal of Virology* **93**.

Sprouffske K, Wagner A. **2016**. Growthcurver: An R package for obtaining interpretable metrics from microbial growth curves. *BMC Bioinformatics* **17**.

Steenwyk J, Rokas A. **2017**. Extensive copy number variation in fermentation-related genes among *Saccharomyces cerevisiae* wine strains. *G3: Genes|Genomes|Genetics* **7**: 1475–1485.

Steinmetz LM, Scharfe C, Deutschbauer AM, *et al.* 2002. Systematic screen for human disease genes in yeast. *Nature Genetics* **31**: 400–404.

Stern S, Dror T, Stolovicki E, Brenner N, Braun E. **2007**. Genome-wide transcriptional plasticity underlies cellular adaptation to novel challenge. *Molecular Systems Biology* **3**.

Stevenson A, Burkhardt J, Cockell CS, *et al.* 2015. Multiplication of microbes below 0.690 water activity: Implications for terrestrial and extraterrestrial life. *Environmental Microbiology* **17**: 257–277.

Stevenson A, Cray JA, Williams JP, *et al.* 2015. Is there a common water-activity limit for the three domains of life. *ISME Journal* **9**: 1333–1351.

Stevenson A, Hamill PG, Dijksterhuis J, Hallsworth JE. **2017**. Water-, pH- and temperature relations of germination for the extreme xerophiles *Xeromyces bisporus* (FRR 0025), *Aspergillus penicillioides* (JH06THJ) and *Eurotium halophilicum* (FRR 2471). *Microbial Biotechnology* **10**: 330–340.

Stoltzfus A, Norris RW. **2016**. On the causes of evolutionary transition:transversion bias. *Molecular Biology and Evolution* **33**: 595–602.

Storz JF, Opazo JC, Hoffmann FG. **2011**. Phylogenetic diversification of the globin gene superfamily in chordates. *IUBMB Life* **63**: 313–322.

Storz JF, Opazo JC, Hoffmann FG. **2013**. Gene duplication, genome duplication, and the functional diversification of vertebrate globins. *Molecular Phylogenetics and Evolution* **66**: 469–478.

Strauss SK, Schirman D, Jona G, *et al.* 2019. Evolthon: A community endeavor to evolve lab evolution. *PLoS Biology* **17**(3):e3000182.

Sun D, Chen S, Cheng A, Wang M. **2016**. Roles of the picornaviral 3c proteinase in the viral life cycle and host cells. *Viruses* **8**(3):82.

Sutton TC, Subbarao K. **2015**. Development of animal models against emerging coronaviruses: From SARS to MERS coronavirus. *Virology* **479**–**480**: 247–258.

Svensson EI, Berger D. **2019**. The role of mutation bias in adaptive evolution. *Trends in Ecology and Evolution* **34**: 422–434.

Svyatchenko VA, Ternovoy VA, Kiselev NN, *et al.* 2017. Bioselection of coxsackievirus B6 strain variants with altered tropism to human cancer cell lines. *Archives of Virology* **162**: 3355–3362.

Talavera D, Kershaw CJ, Costello JL, *et al.* 2018. Archetypal transcriptional blocks underpin yeast gene regulation in response to changes in growth conditions. *Scientific Reports* **8**.

Tamari Z, Yona AH, Pilpel Y, Barkai N. **2016**. Rapid evolutionary adaptation to growth on an “unfamiliar” carbon source. *BMC Genomics* **17**: 1–7.

Taylor JS, Raes J. **2004**. Duplication and divergence: the evolution of new genes and old ideas. *Annual Review of Genetics* **38**: 615–643.

Taymaz-Nikerel H, Cankorur-Cetinkaya A, Kirdar B. **2016**. genome-wide transcriptional response of *Saccharomyces cerevisiae* to stress-induced perturbations. *Frontiers in Bioengineering and Biotechnology* **4**.

Teixeira MC, Mira NP, Sá-Correia I. **2011**. A genome-wide perspective on the response and tolerance to food-relevant stresses in *Saccharomyces cerevisiae*. *Current Opinion in Biotechnology* **22**: 150–156.

Thibaut HJ, van der Linden L, Jiang P, *et al.* 2014. Binding of glutathione to enterovirus capsids is essential for virion morphogenesis. *PLoS Pathogens* **10**: e1004039.

Thompson VC, McGuire BE, Frier MS, *et al.* 2020. Temperature-sensitive recombinant subtilisin protease variants that efficiently degrade molecular biology enzymes. *FEMS Microbiology Letters* **367**(19):fnaa162.

Thompson DA, Roy S, Chan M, *et al.* 2013. Evolutionary principles of modular gene regulation in yeasts. *eLife* **2013**: 1–37.

Thyagarajan B, Bloom JD. **2014**. The inherent mutational tolerance and antigenic evolvability of influenza hemagglutinin. *eLife* **2014**: 1–26.

Tirosh I, Barkai N. **2007**. Comparative analysis indicates regulatory neofunctionalization of yeast duplicates. *Genome Biology* **8**:R50.

Tirosh I, Barkai N, Verstrepen KJ. **2009**. Promoter architecture and the evolvability of gene expression. *Journal of biology* **8**: 95.

Tirosh I, Weinberger A, Carmi M, Barkai N. **2006**. A genetic signature of interspecies variations in gene expression. *Nature Genetics* **38**: 830–834.

Tomislav S, Supek F, Bošnjak M, Škunca N, Šmuc T, Tomislav S. **2011**. REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS ONE* **6**(7):e21800.

Tong AHY, Evangelista M, Parsons AB, *et al.* 2001. Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science* **294**: 2364–2368.

Toprak E, Veres A, Michel JB, Chait R, Hartl DL, Kishony R. **2012**. Evolutionary paths to antibiotic resistance under dynamically sustained drug selection. *Nature Genetics* **44**: 101–105.

Tulha J, Lima A, Lucas C, Ferreira C. **2010**. *Saccharomyces cerevisiae* glycerol/H+symporter Stl1p is essential for cold/near-freeze and freeze stress adaptation. A simple recipe with high biotechnological potential is given. *Microbial Cell Factories* **9**.

Turcotte B, Liang XB, Robert F, Soontorngun N. **2010**. Transcriptional regulation of nonfermentable carbon utilization in budding yeast. *FEMS Yeast Research* **10**: 2–13.

Turner CB, Blount ZD, Lenski RE. **2015**. Replaying evolution to test the cause of extinction of one ecotype in an experimentally evolved population. *PLoS ONE* **10**(11):e0142050.

Tuthill TJ, Groppelli E, Hogle JM, Rowlands DJ. **2010**. Picornaviruses In: *Current topics in microbiology and immunology*. NIH Public Access, 43–89.

Vandersluis B, Bellay J, Musso G, *et al.* 2010. Genetic interactions reveal the evolutionary trajectories of duplicate genes. *Molecular Systems Biology* **6**:429.

Veitia RA. **2003a**. Nonlinear effects in macromolecular assembly and dosage sensitivity. *Journal of Theoretical Biology* **220**: 19–25.

Veitia RA. **2003b**. A sigmoidal transcriptional response: Cooperativity, synergy and dosage effects. *Biological Reviews of the Cambridge Philosophical Society* **78**: 149–170.

Verbist BMP, Thys K, Reumers J, *et al.* 2015. VirVarSeq: A low-frequency virus variant detection pipeline for Illumina sequencing using adaptive base-calling accuracy filtering. *Bioinformatics* **31**: 94–101.

Vignuzzi M, Wendt E, Andino R. **2008**. Engineering attenuated virus vaccines by controlling replication fidelity. *Nature Medicine* **14**: 154–161.

Vilhelmsson O, Miller KJ. **2002**. Humectant permeability influences growth and compatible solute uptake by Staphylococcus aureus subjected to osmotic stress. *Journal of Food Protection* **65**: 1008–1015.

Voordeckers K, Kominek J, Das A, *et al.* 2015. Adaptation to high ethanol reveals complex evolutionary pathways. *PLoS Genetics* **11**: 1–31.

van Voorst F, Houghton-Larson J, Jønson L, Kielland-Brandt MC, Brandt A. **2006**. Genome-wide identification of genes required for growth of *Saccharomyces cerevisiae* under ethanol stress. *Yeast* **23**: 351–359.

Waddington CH. **1942**. Canalization of development and the inheritance of acquired characters. *Nature* **150**: 563–563.

Wagner A. **2000**. Robustness against mutations in genetic networks of yeast. *Nature Genetics* **24**: 355–361.

Wagner A. **2005**. Robustness, evolvability, and neutrality. *FEBS Letters* **579**: 1772–1778.

Wagner A. **2008**. Neutralism and selectionism: a network-based reconciliation. *Nature reviews. Genetics* **9**: 965–74.

Wagner A. **2011**. *The Origins of Evolutionary Innovation.* United States: Oxford University Press Inc., New York.

Wagner A. **2012**. The role of robustness in phenotypic adaptation and innovation. *Proceedings of the Royal Society B: Biological Sciences* **279**: 1249–1258.

Wagner A. **2015**. Causal drift, robust signaling, and complex disease. *PLoS ONE* **10**: 8–13.

Wala J, Zhang CZ, Meyerson M, Beroukhim R. **2016**. VariantBam: Filtering and profiling of nextgenerational sequencing data using region-specific rules. *Bioinformatics* **32**: 2029–2031.

Wang Y, Wang X, Paterson AH. **2012**. Genome and gene duplications and gene expression divergence: A view from plants. *Annals of the New York Academy of Sciences* **1256**: 1–14.

Warringer J, Blomberg A. **2003**. Automated screening in environmental arrays allows analysis of quantitative phenotypic profiles in *Saccharomyces cerevisiae*. *Yeast* **20**: 53–67.

Wendel JF. **2000**. Genome evolution in polyploids. *Plant Molecular Evolution* **42**: 225–249.

West-Eberhard MJ. **1989**. Phenotypic plasticity and the origins of diversity. *Annual*

review of ecology and systematics. Vol. 20: 249–278.

Wideman JG, Novick A, Muñoz-Gómez SA, Doolittle WF. **2019**. Neutral evolution of cellular phenotypes. *Current Opinion in Genetics and Development* **58**–**59**: 87–94.

Wilke CO, Drummond DA. **2006**. Population genetics of translational robustness. *Genetics* **173**: 473–481.

Williams JP, Hallsworth JE. **2009**. Limits of life in hostile environments: No barriers to biosphere function? *Environmental Microbiology* **11**: 3292–3308.

Winston PW, Bates DH. **1960**. Saturated solutions for the control of humidity in biological research. *Ecology* **41**: 232–237.

Wolfe KH. **2015**. Origin of the yeast whole-genome duplication. *PLoS Biology* **13**: 1–7.

Wolfe KH, Shields DC. **1997**. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* **387**: 708–713.

Wu B, Cox MP. **2019**. Greater genetic and regulatory plasticity of retained duplicates in *Epichloë endophytic* fungi. *Molecular Ecology* **28**: 5103–5114.

Wu NC, Olson CA, Du Y, *et al.* 2015. Functional constraint profiling of a viral protein reveals discordance of evolutionary conservation and functionality. *PLoS Genetics* **13**(4):e1006709.

Xiao Y, Dolan PT, Goldstein EF, *et al.* 2017. Poliovirus intrahost evolution is required to overcome tissue-specific innate immune responses. *Nature Communications* **8**: 375.

Xing L, Tjarnlund K, Lindqvist B, *et al.* 2000. Distinct cellular receptor interactions in poliovirus and rhinoviruses. *EMBO Journal* **19**: 1207–1216.

Xu L, Zheng Q, Li Shaowei, *et al.* 2017. Atomic structures of Coxsackievirus A6 and its complex with a neutralizing antibody. *Nature Communications* **8**: 1–12.

Yadav KK, Singh N, Rajasekharan R. **2016**. Responses to phosphate deprivation in yeast cells. *Current Genetics* **62**: 301–307.

Yakimov MM, La Cono V, Spada GL, *et al.* 2015. Microbial community of the deep-sea brine Lake Kryos seawater-brine interface is active below the chaotropicity limit of life as revealed by recovery of mRNA. *Environmental Microbiology* **17**: 364–382.

Yang Z. **2007**. PAML 4: Phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution* **24**: 1586–1591.

Yang J, Bae JY, Lee YM, *et al.* 2011. Construction of *Saccharomyces cerevisiae* strains with enhanced ethanol tolerance by mutagenesis of the TATA-binding protein gene and identification of novel genes associated with ethanol tolerance. *Biotechnology and Bioengineering* **108**: 1776–1787.

Yoder JD, Cifuente JO, Pan J, Bergelson JM, Hafenstein S. **2012**. The crystal structure of a coxsackievirus B3-RD variant and a refined 9-angstrom cryo-electron microscopy reconstruction of the virus complexed with decay-accelerating factor (DAF) provide a new footprint of DAF on the virus surface. *Journal of Virology* **86**: 12571–12581.

Yona AH, Manor YS, Herbst RH, *et al.* 2012. Chromosomal duplication is a transient evolutionary solution to stress. *Proceedings of the National Academy of Sciences of the United States of America* **109**: 21010–21015.

Yosef N, Regev A. **2011**. Impulse control: Temporal dynamics in gene transcription. *Cell*

144: 886–896.

Ypma-Wong MF, Dewalt PG, Johnson VH, Lamb JG, Semler BL. **1988**. Protein 3CD is the major poliovirus proteinase responsible for cleavage of the p1 capsid precursor. *Virology* **166**: 265–270.

Yu G, Wang LG, Han Y, He QY. **2012**. ClusterProfiler: An R package for comparing biological themes among gene clusters. *OMICS A Journal of Integrative Biology* **16**: 284–287.

Yun T, Park A, Hill TE, *et al.* 2015. Efficient reverse genetics reveals genetic determinants of budding and fusogenic differences between nipah and hendra viruses and enables real-time monitoring of viral spread in small animal models of henipavirus infection. *Journal of Virology* **89**: 1242–1253.

Zakrzewska A, Van Eikenhorst G, Burggraaff JEC, *et al.* 2011. Genome-wide analysis of yeast stress survival and tolerance acquisition to analyze the central trade-off between growth rate and cellular robustness. *Molecular Biology of the Cell* **22**: 4435–4446.

Zell R. **2018**. Picornaviridae—the ever-growing virus family. *Archives of Virology* **163**: 299–317.

Zhang J. **2003**. Evolution by gene duplication: An update. *Trends in Ecology and Evolution* **18**: 292–298.

Zhang J. **2018**. Neutral theory and phenotypic evolution. *Molecular Biology and Evolution* **35**: 1327–1331.

Zhang K, Fang YH, Gao KH, Sui Y, Zheng DQ, Wu XC. **2017**. Effects of genome duplication on phenotypes and industrial applications of *Saccharomyces cerevisiae* strains. *Applied Microbiology and Biotechnology* **101**: 5405–5414.

Zhang ZH, Jhaveri DJ, Marshall VM, *et al.* 2014. A comparative study of techniques for differential expression analysis on RNA-seq data. *PLoS ONE* **9**(8):e103207.

Zhang J, Yang JR. **2015**. Determinants of the rate of protein sequence evolution. *Nature Reviews Genetics* **16**: 409–420.

Zhang C, Zhao Z, Guo Z, *et al.* 2017. Amino acid substitutions associated with avian H5N6 influenza A virus adaptation to mice. *Frontiers in Microbiology* **8**: 1763.

Zheng J, Payne JL, Wagner A. **2019**. Cryptic genetic variation accelerates evolution by opening access to diverse adaptive peaks. *Science* **365**: 347–353.

Zhou Y, Li X, Katsuma S, *et al.* 2019. Duplication and diversification of trehalase confers evolutionary advantages on lepidopteran insects. *Molecular Ecology* **28**: 5282–5298.

Zou C, Lehti-Shiu MD, Thomashow M, Shiu SH. **2009**. Evolution of stress-regulated gene expression in duplicate genes of *Arabidopsis thalianae*. *PLoS Genetics* **5**(7):e1000581.

APPENDIX

Chapter I.

Data S1: **Transcription levels correlate with expression levels in *S. cerevisiae.*** We compared the level of reads per billion (RPKM) obtained under our study to that of gene expression corresponding to a previous study (Albert et al. 2014) in which they used ribosomal profiling to determine the number of mRNA molecules that are translated genome wide. In total, we compared the levels of expression for 4682 genes. We found a very strong correlation between both measures of gene expression (Fig DS1, Spearman´s correlation: $\rho = 0.77$, $P < 2.2 \times 10^{-16}$).

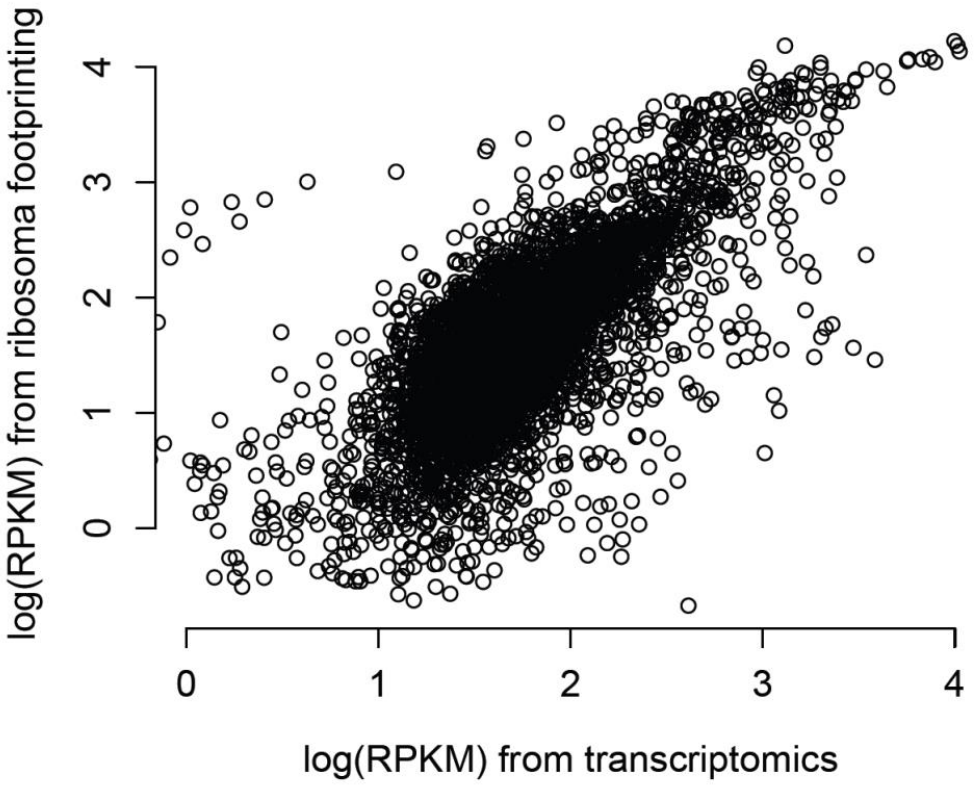


Figure DS 1 *The number of reads per billion (RPKM) obtained from transcriptomic data correlates significantly with that from ribosomal profiling data. The RPKM values are represented in logarithmic scale. We obtained ribosomal profiling RPKMs for 4682 genes from a previous study for yeast growing in rich growth media. Comparison of these with the transcriptomic data revealed a significant correlation (Spearman´s correlation: ρ = 0.77, P < 2.2 x 10-16), such that genes with high RPKMs in the transcriptomic data correspond to those with high RPKMs in the ribosomal profiling data, and vice versa.*

Figure ChI-S1. Duplicates contain a greater proportion of genes with TATA motifs in their promoters than singletons.

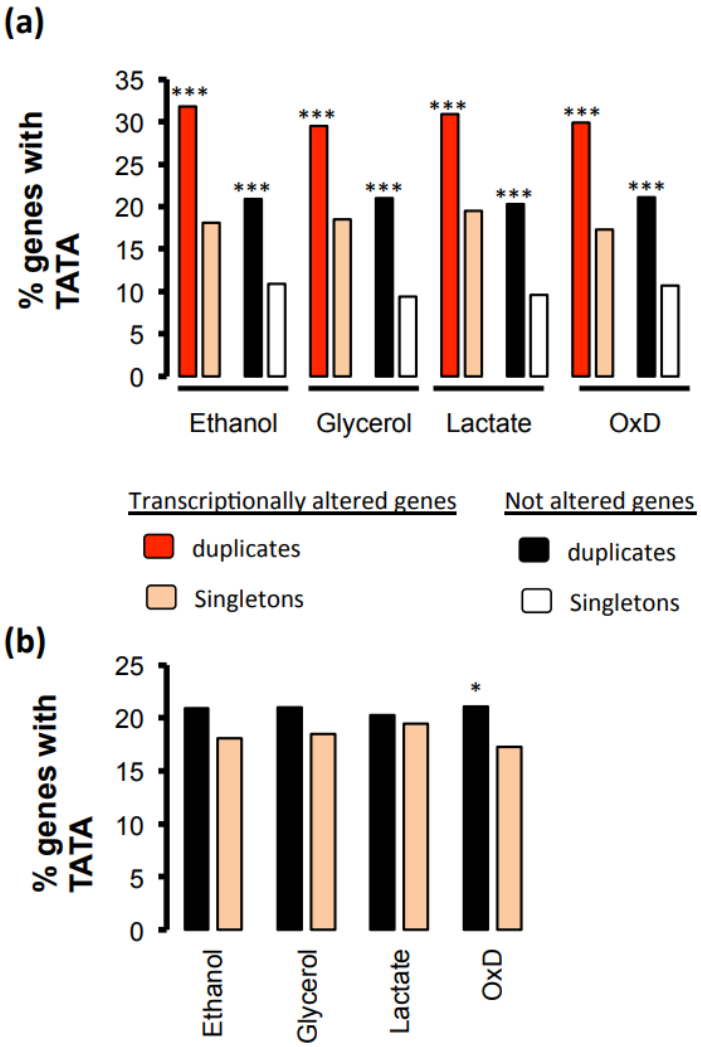


Figure ChI-S1 **Duplicates contain a greater proportion of genes with TATA motifs in their promoters than singletons. A)** The proportion of TATA-containing genes was greater for duplicates than singletons for transcriptionally altered genes and non-altered genes in four different stress conditions (Ethanol, Glycerol, Lactate and Oxidative stress in a medium supplemented with Dextrose). **B)** The proportion of transcriptionally non-altered duplicates containing TATA-motifs in their promoters was higher than the proportion of transcriptionally altered singletons that contained TATA-motifs in their promoters. Data were compared using Fisher´s exact test, with * and ***, indicating probabilities of $P < 0.05$ and $P < 10^{-10}$, respectively.

Chapter II.

File S8. Comparison of different methods of differential expression through the analysis of RNA sequence data. We compared three main methods of differential expression by calculating the correlation in the fold change estimated by these methods. The three methods used were edgeR, DESeq, and Cufflinks. Analysis of the correlation in the logarithm of fold change between edgeR and DESeq supports that the data are very robust to the use of these two methods, with a spearman´s correlation of 0.995, $P < 2.2 \times 10^{-16}$ (Fig FS8-1).

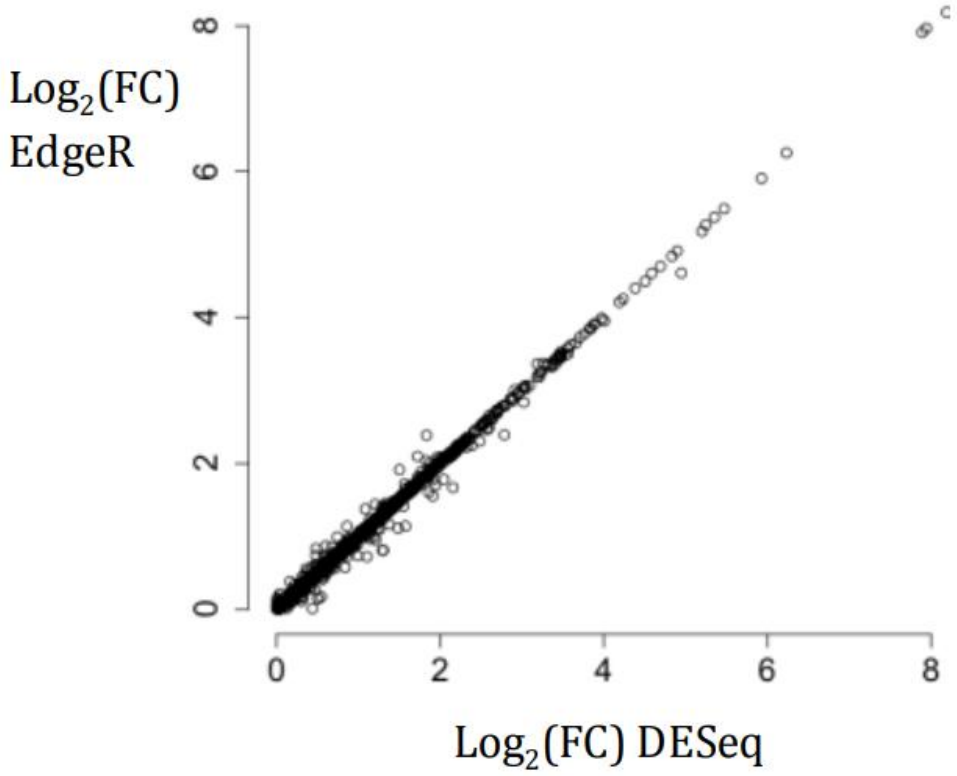


Figure FS8- 1 **Correlation analysis in the fold change expression using a reference transcriptome from strain S288c between two methods edgeR and DESeq.** All numbers have been log-transformed for the comparison.

Chapter III.

Data Set 1. Role of duplicated genes in the response to glycerol-induced stress.

We investigated if duplicated genes have driven the transcriptional response in the populations at t_{100} and t_{110} once these populations were challenged with 0.41 M glycerol in the growth media. Analysis of the t_{100} population yielded similar results than the t_0 population, with transcriptionally altered genes when such population as grown in YPG exhibiting significantly greater number of duplicates altered than expected (36.3% of altered genes were duplicates, binomial test: $P = 3.06 \times 10^{-4}$). Whole genome duplicates (WGDs) contributed more significantly than expected by chance to the transcriptional response of this population to glycerol-induced stress (N = 470 out of 881 duplicates, binomial test: $P = 0.023$). Finally, we observed a very similar pattern in the t_{110} population, which was adapted to YPG, with 36.7% of the altered genes belonging to the category of duplicated genes (a proportion greater than expected by chance. Binomial test: $P = 2.33 \times 10^{-5}$). Remarkably, unlike the t_0 and t_{100} populations, the t_{110} population that evolved in YPG exhibited no differences between the contribution of WGDs and SSDs to the transcriptional response to glycerol-induced stress. Indeed the number of WGDs involved in the transcriptional response to glycerol-induced stress (N = 495, out of 969) was not significantly greater than expected by chance (Binomial test: $P = 0.34$). These results suggest the possibility that WGDs evolved to respond to stress but the fine-tuning and adaptation to glycerol-induced stress required the evolution of the regulation of WGDs and SSDs.

Chapter IV.

Figure ChIV-S1. Maximum growth rate as phenotypic characterization of populations through experimental evolution.

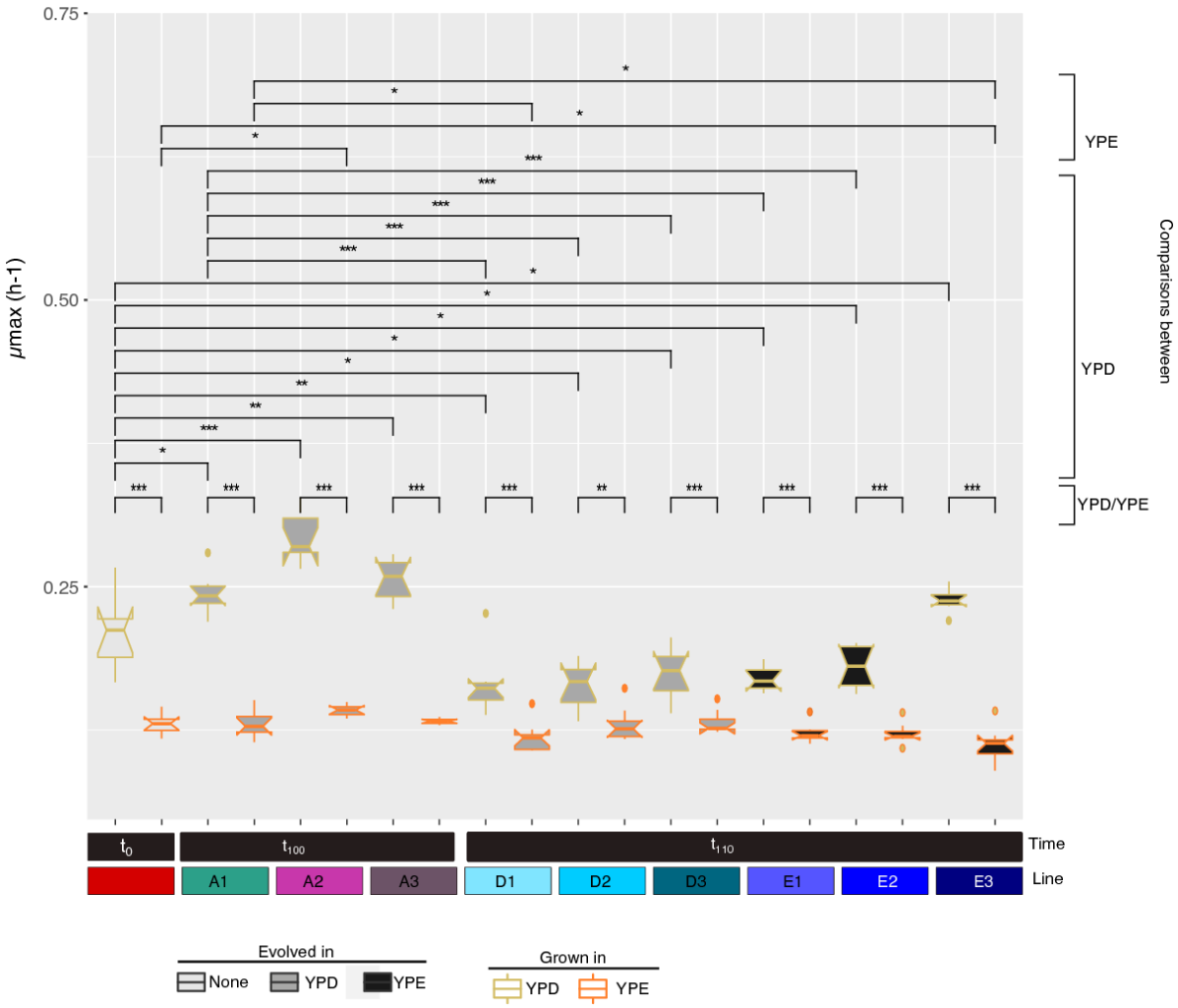


Figure ChIV-S1 **Phenotypic characterization was performed by characterization of population growth curves.** "Maximum growth rate" (h^{-1}) of each population, split out into the different lines, was determined at each control time point (t_0, t_{100}, and t_{110}) in their evolving medium (YPD or YPE) and in the challenge medium (YPE or YPD). Significant differences of each growth parameter are indicated as *, **, and ***, when the probabilities are $P < 0.05$, $P < 0.005$, and $P < 10^{-3}$, respectively, using a Wilcoxon rank test. Only biologically meaningful comparisons were added to the figure.

Figure ChIV-S2. Evolution of carry capacity of populations under ethanol stress.

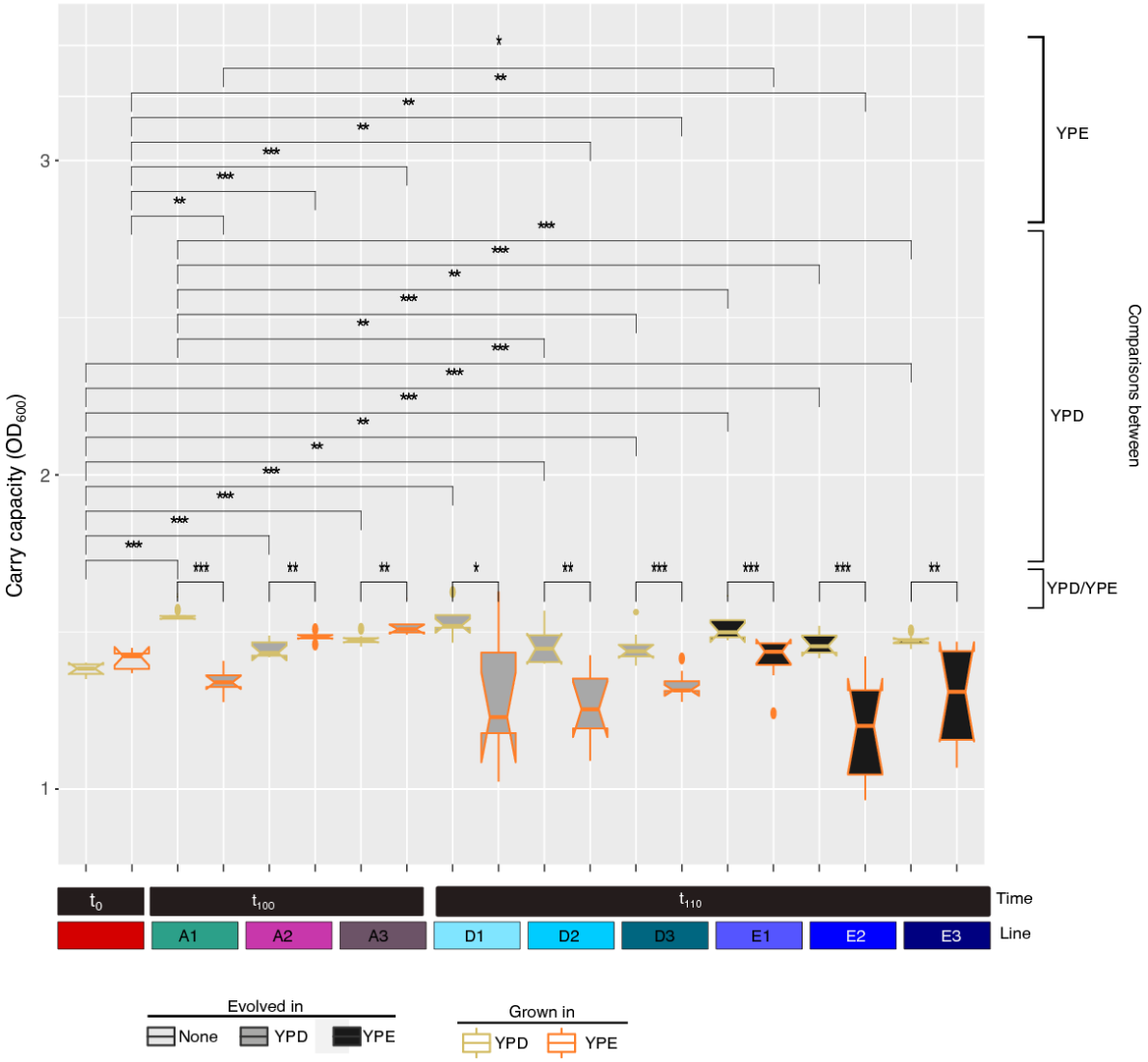


Figure ChIV-S2 **Phenotypic characterization was performed by characterization of population growth curves.** "Carry capacity" (OD600) of each population, split out into the different lines, was determined at each control time point (t_0, t_{100}, and t_{110}) in their evolving medium (YPD or YPE) and in the challenge medium (YPE or YPD). Significant differences of each growth parameter are indicated as *, **, and ***, when the probabilities are $P < 0.05$, $P < 0.005$, and $P < 10^{-3}$, respectively, using a Wilcoxon rank test. Only biologically meaningful comparisons were added to the figure.

Figure ChIV-S3. Distribution of transcriptomic profiles under ethanol stress.

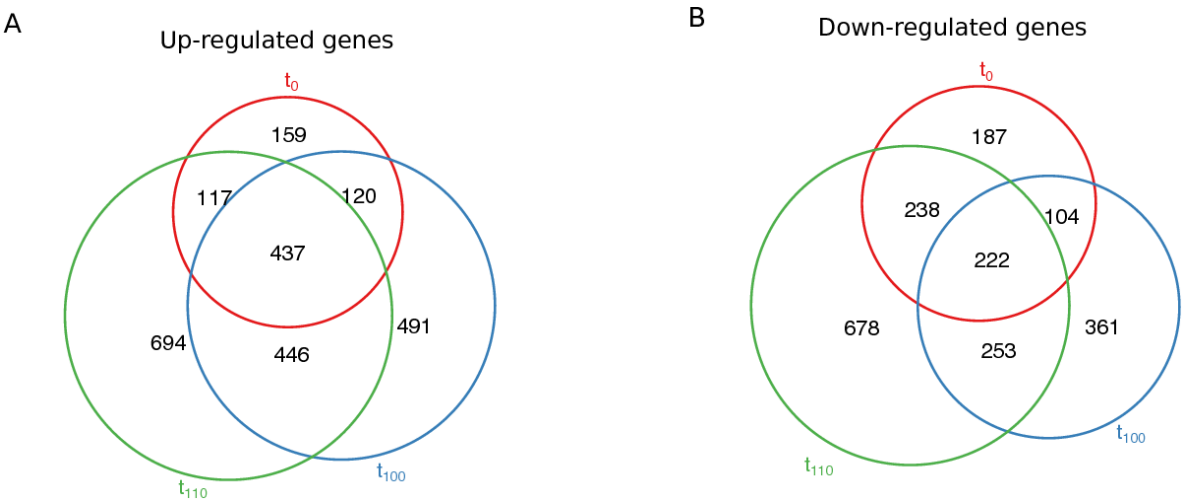


***Figure ChIV-S3* Distribution of transcriptomic profiles under ethanol stress.** Venn diagrams of up- and downregulated genes in YPE compared to YPD.

Figure ChIV-S4. Biological processes enriched for downregulated genes, due to the use of 3% ethanol as the sole carbon source.

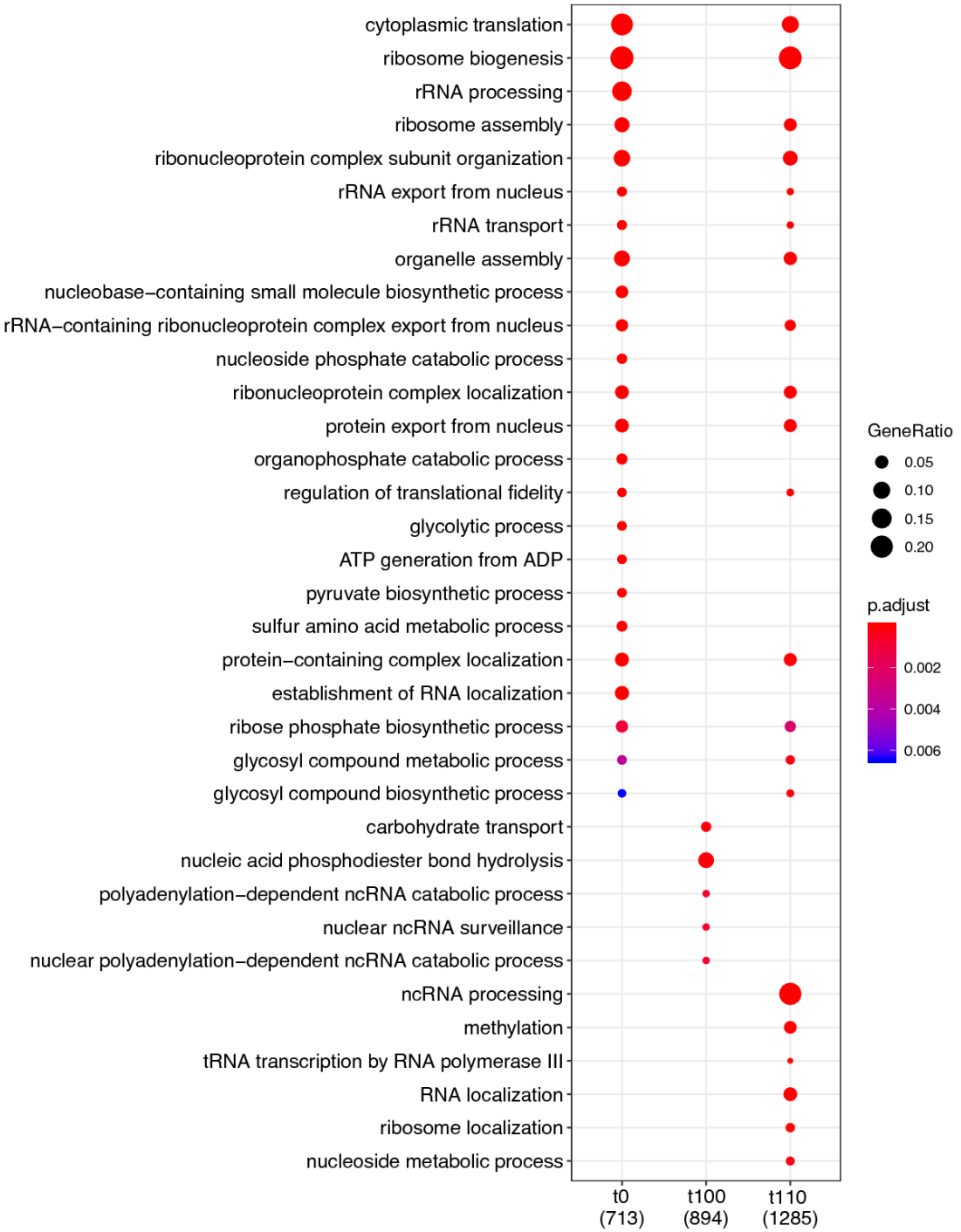


***Figure ChIV-S4* Biological processes enriched for downregulated genes, due to the use of 3% ethanol as the sole carbon source.** Enrichment analysis of functional categories (biological process) for downregulated genes in YPE compared to YPD, at the three time points (t_0, t_{100}, and t_{110}), was performed with clusterProfiler.

Figure ChIV-S5. Biological processes enriched for transcriptionally divergent duplicated genes.

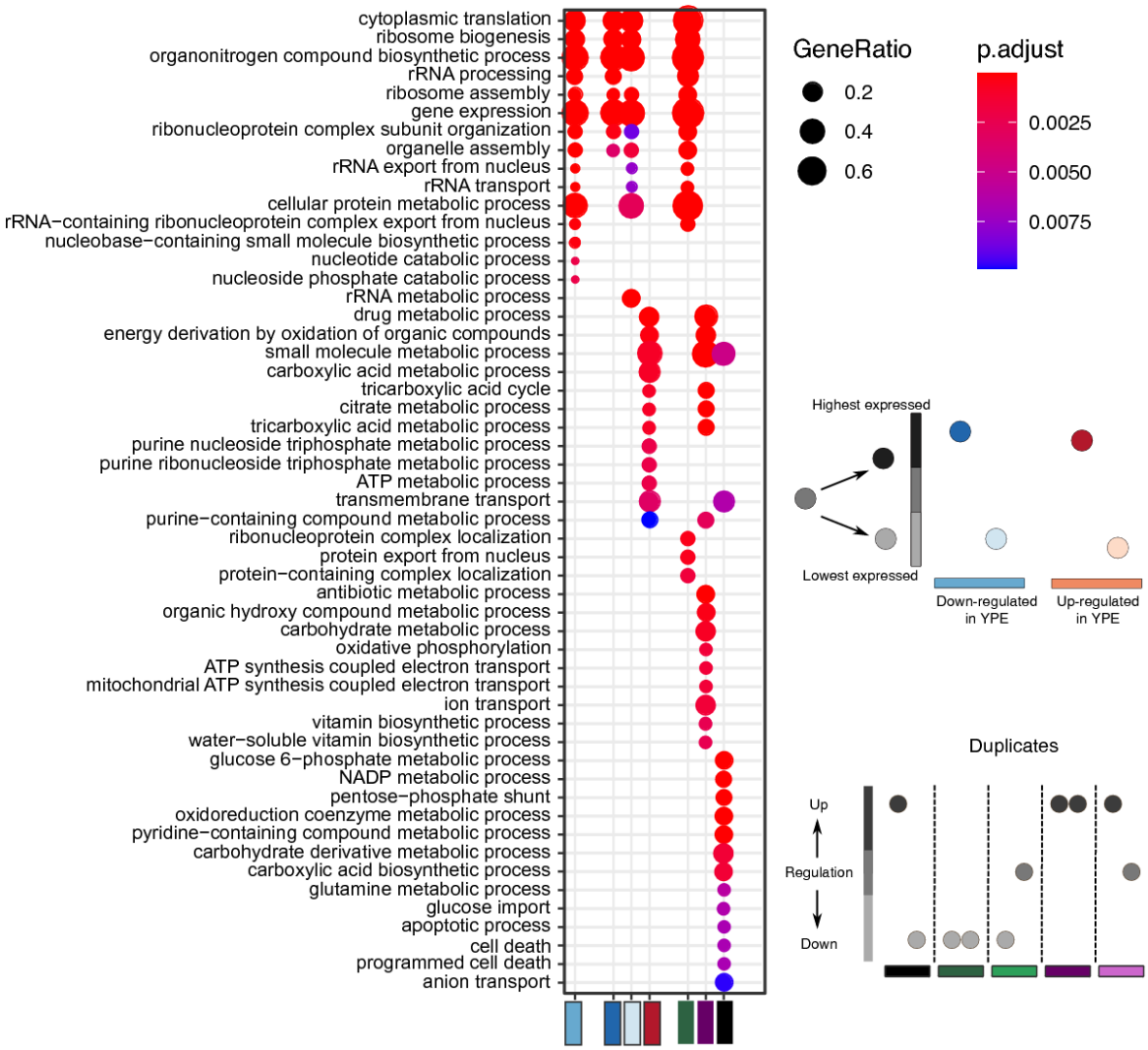


Figure ChIV-S5 **Biological processes enriched for transcriptionally divergent duplicated genes.** Enrichment analysis of functional categories (biological process) for down- and upregulated transcriptionally divergent duplicated genes in YPE compared to YPD was performed with clusterProfiler.

Figure ChIV-S6. Implication of transcriptional divergence limit on transcriptional response of duplicates.

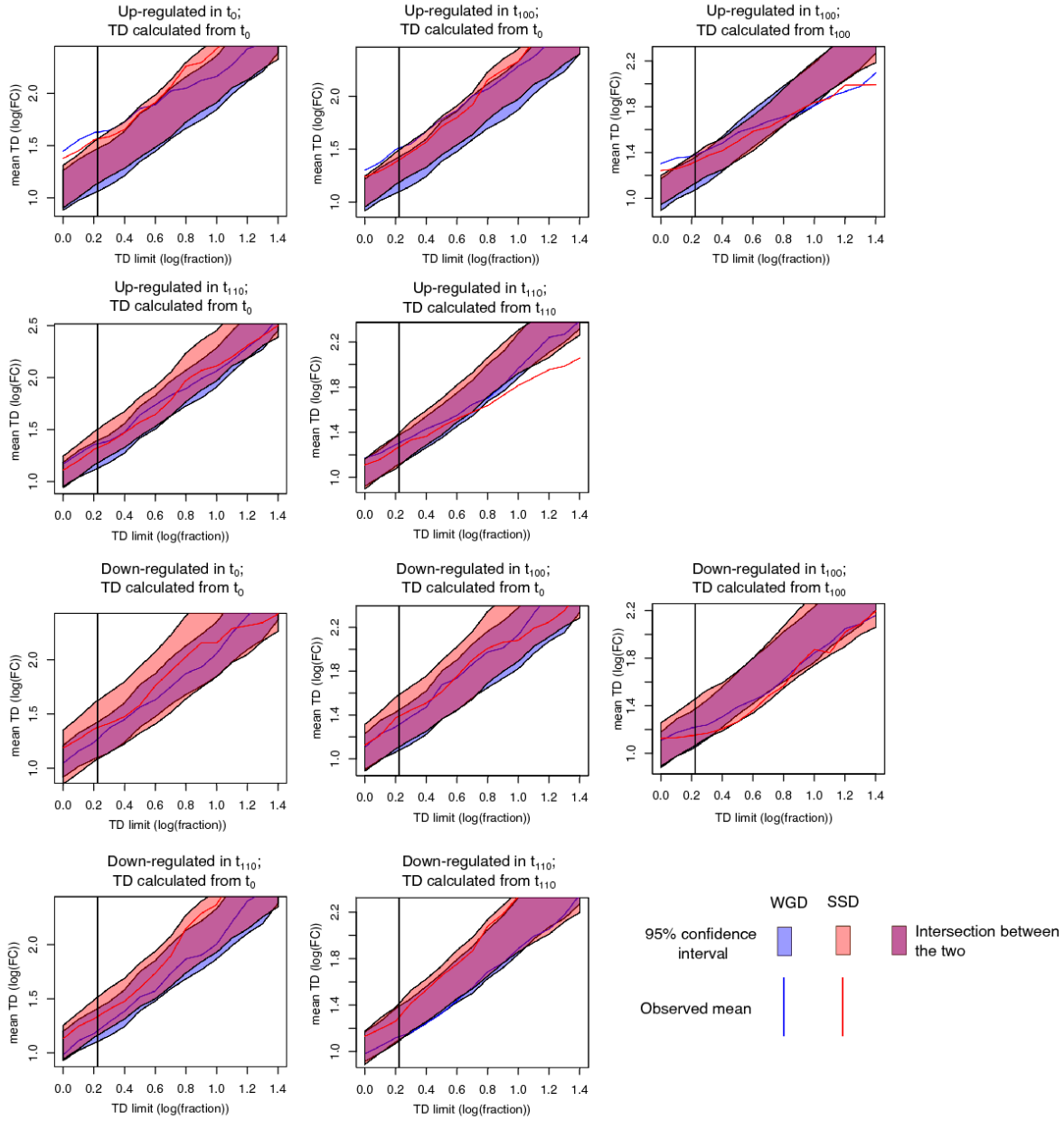


Figure ChIV-S6 **Implication of transcriptional divergence limit on transcriptional response of duplicates.** Shown is the influence of changing the TD limit, for which a duplicated pair is seen as being TD, on the enrichment analysis of responding duplicates which are TD. The analysis for the TD limit goes from equal expression to a fourfold difference.

Chapter VI.

Figure ChVI-S1. Sanger analysis of DMS libraries.

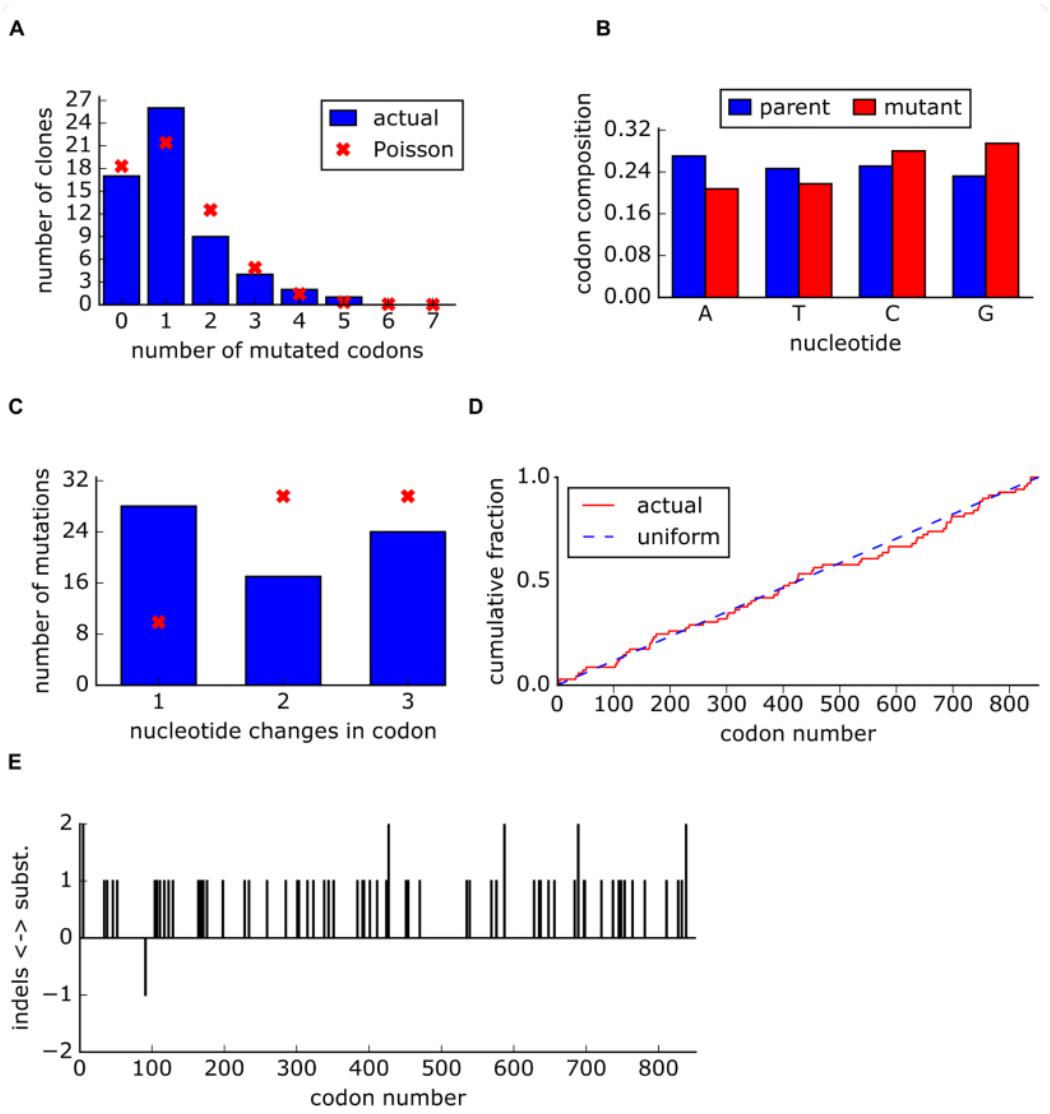


Figure ChVI-S 1 **Sanger analysis of DMS libraries. A)** The number of mutated codons per clone. **B)** Original and mutated base for each mutation. **C)** The number of nucleotide changes per codon. **D)** Cumulative fraction of mutations versus the codon position. **E)** Location of both mutations and indels across the capsid sequence.

Figure ChVI-S2. Results of high-fidelity duplex sequencing.

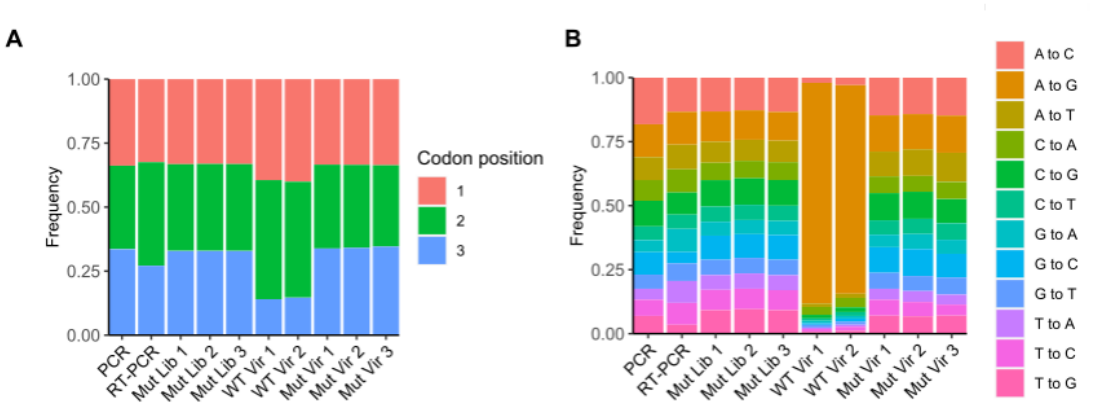


***Figure ChVI-S 2* Resuls of high-fidelity duplex sequencing. A)** The relative frequency of the mutated base within each mutated codon. **B)** The relative frequency of each mutation type.

Figure ChVI-S3. Correlation of amino acid preferences observed in experimental replicates.

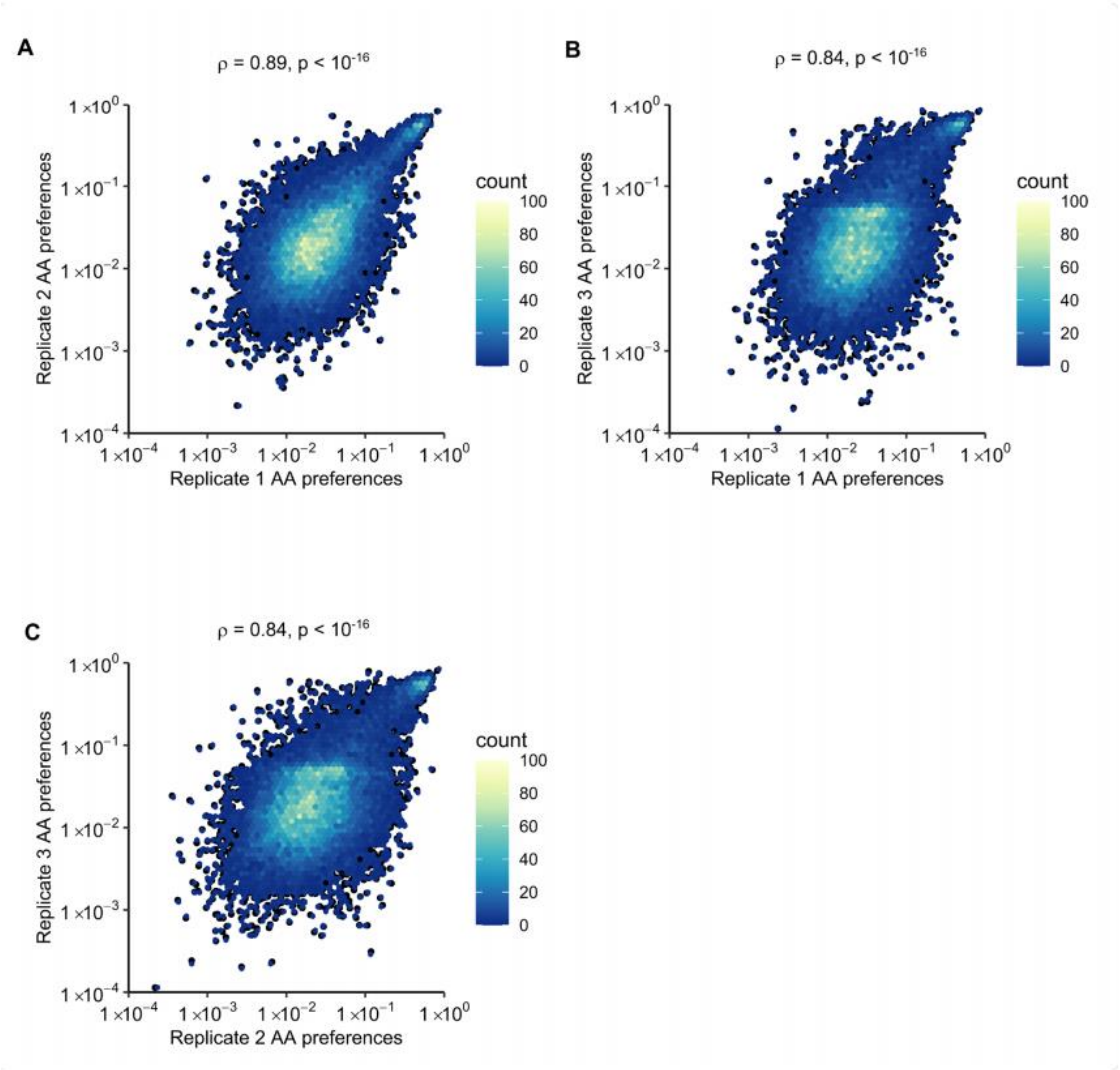


Figure ChVI-S 3 **Correlation of amino acid preferences observed in experimental replicates.** Hexagonal bin plots showing the correlation of amino acid preferences between the three experimental replicates. Spearman's correlation coefficient and p-value are shown above each plot.

Figure ChVI-S4. Prediction of mutational fitness effects using random forest or linear models.

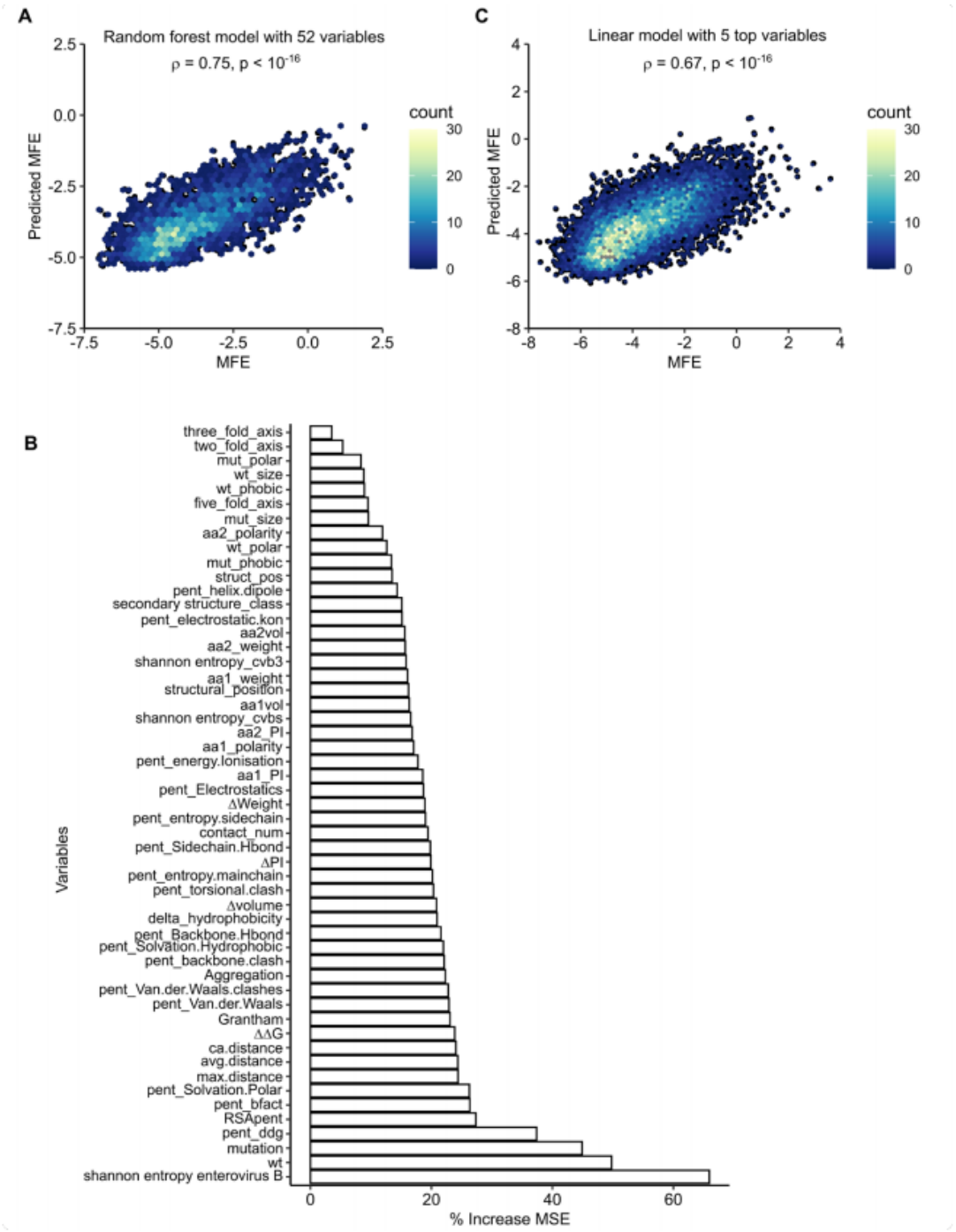


Figure ChVI-S 4 **Prediction of mutational fitness effects using random forest or linear models. A)** Hexagonal bin plot showing the correlation between actual and predicted MFE derived from a random forest model using all 52 variables. The model was trained on 70% of the data and tested on the remaining 30% of the data (shown). **B)** Variable importance obtained from the random forest model. **C)** Linear model using the top five parameters of the random forest model. See supplementary Table S6 for parameter description.

Figure ChVI-S5. Sequence preferences of capsid encoded motifs.

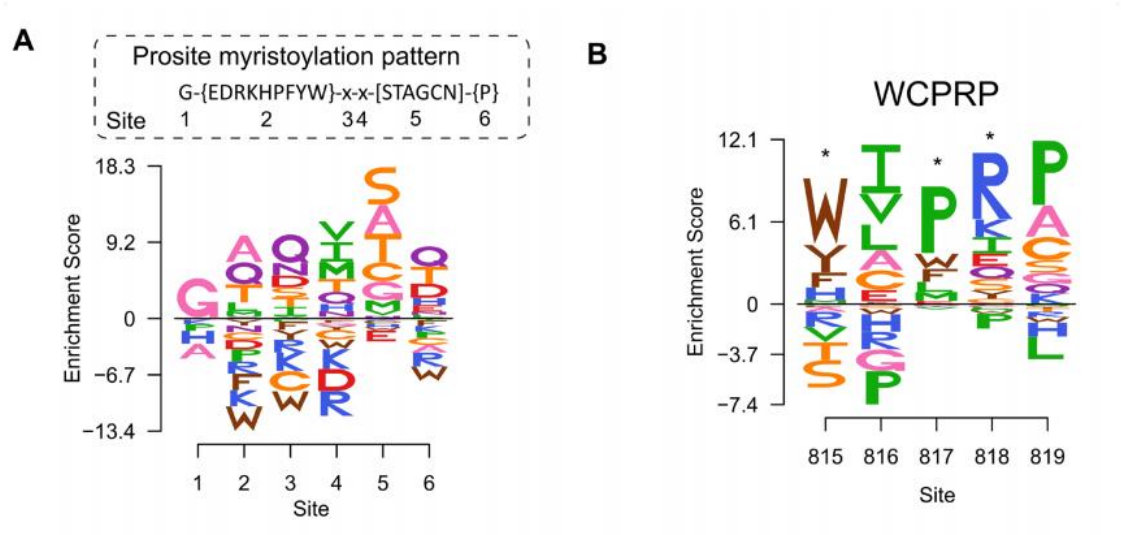


***Figure ChVI-S 5* Sequence preferences of capsid encoded motifs. A)** Amino acid preferences of the CVB3 myristoylation motif. The canonical Prosite myristoylation motif is indicated above, with curly brackets indicating disfavored amino acids and square brackets indicating tolerated amino acids. **B)** WCPRP motif required for $3CD^{pro}$ cleavage of P1. Asterisks indicate analogous positions in FMDV shown to be essential for viability (Kristensen andBelsham, 2019).

Figure ChVI-S6. Evaluation of select hits identified as potential 3CD^{pro} target proteins.

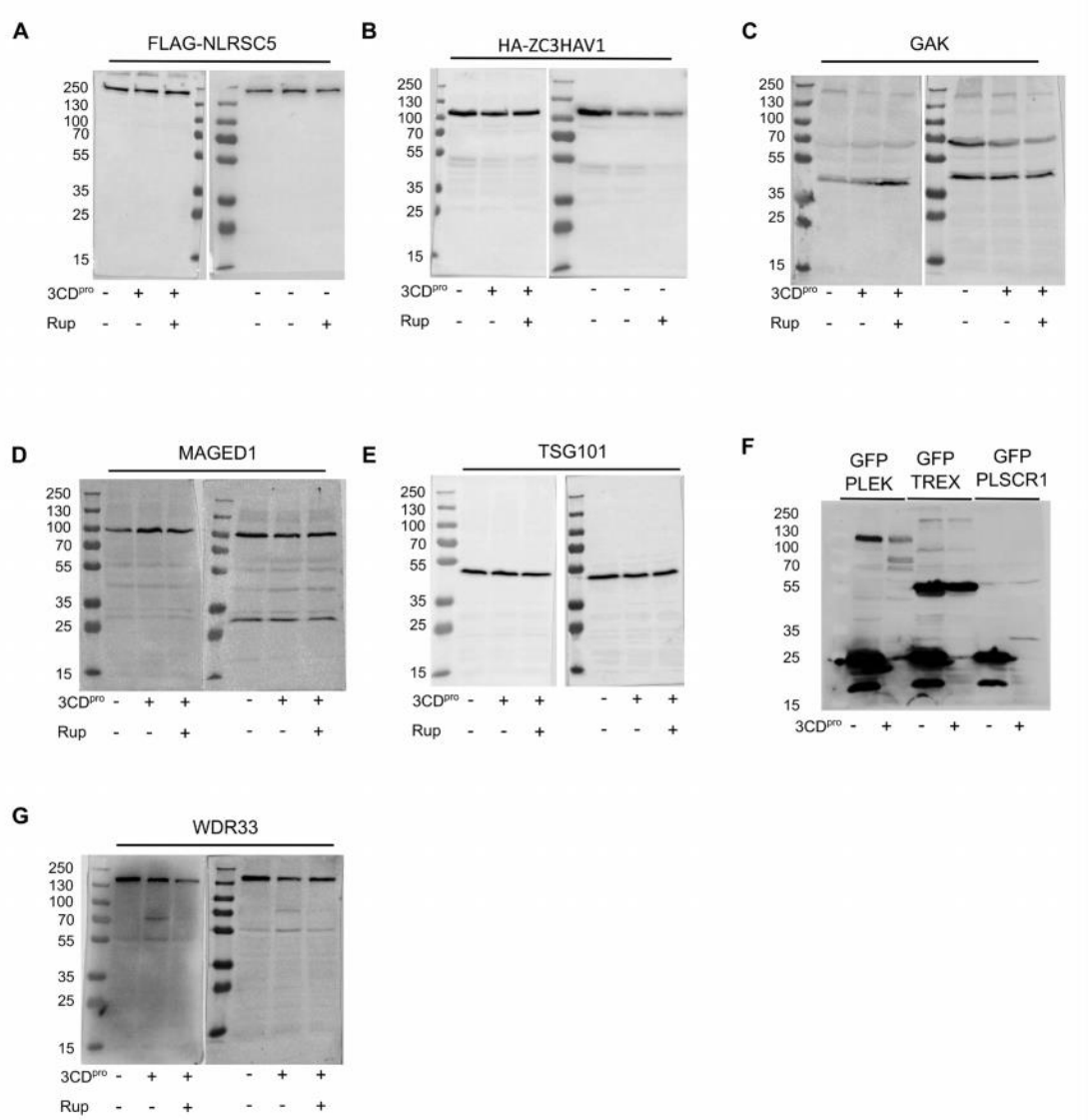


Figure ChVI-S 6 **Evaluation of select hits identified as potential 3CD^{pro} target proteins.** Western blots of cells transfected with 3CD^{pro} and probed for the indicated endogenous protein, or cotransfected with 3CD^{pro} and the indicated fusion protein and blotted for the tag. Each experiment was performed twice. When indicated, the 3C^{pro} inhibitor rupintrivir was added.

Chapter VII.

Figure ChVII-S1. Analysis of library diversity by Sanger sequencing.

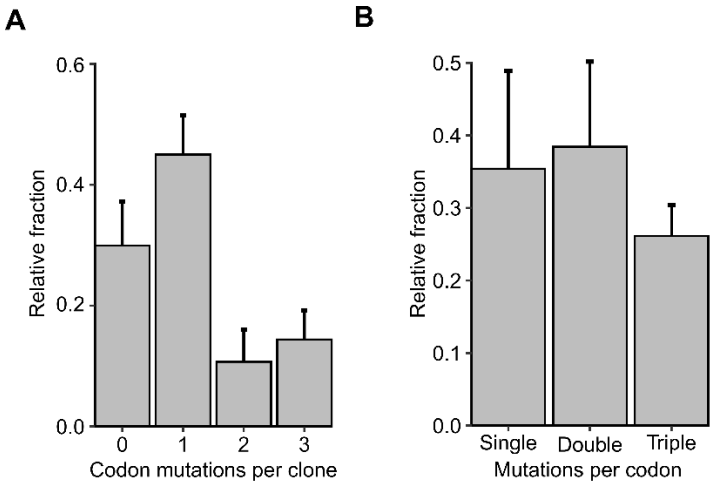


Figure ChVII-S1 **Analysis of library diversity by Sanger sequencing. A)** The average fraction of the number of codon mutations observed per clone in the three mutagenized libraries. **B)** The average fraction of single, double, and triple mutations within each codon mutation in the three mutagenized libraries.

Figure ChVII-S2. Positions mutated in evolved populations occur in variable positions.

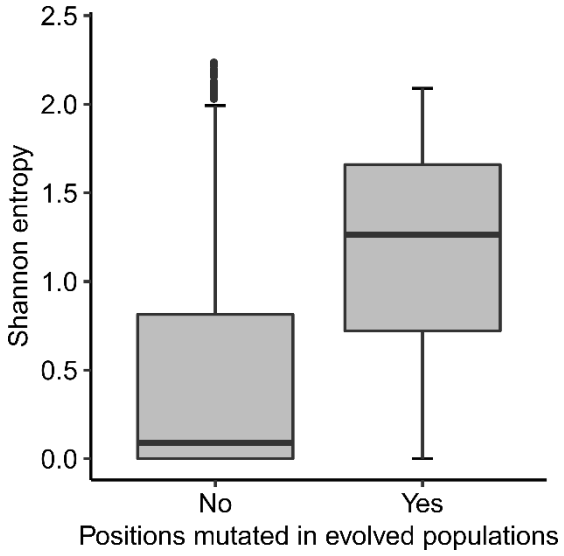


Figure ChVII-S2 **Positions mutated in evolved populations occur in variable positions.** Positions where mutations were observed in the thermal selected populations occur at more variable positions in enterovirus B sequences versus all other positions, as judged by Shannon entropy.

Dna Research, 2017, 0(0), 1–12
doi: 10.1093/dnares/dsx025
Full Paper

Full Paper

Expression properties exhibit correlated patterns with the fate of duplicated genes, their divergence, and transcriptional plasticity in *Saccharomycotina*

Florian Mattenberger[1,2,†], Beatriz Sabater-Muñoz[1,2,3,*,†], Christina Toft[4,5], Gaurav Sablok[6], and Mario A. Fares[1,2,3,*]

[1]Department of Abiotic Stress, Instituto de Biología Molecular y Celular de Plantas, Consejo Superior de Investigaciones Científicas-Universidad Politécnica de Valencia, Valencia 46022, Spain, [2]Systems Biology of Molecular Interactions and Regulation Department, Institute for Integrative Systems Biology (I2SysBio), Consejo Superior de Investigaciones, Científicas-Universitat de Valencia (CSIC-UV), Valencia 46980, Spain, [3]Department of Genetics, Smurfit Institute of Genetics, University of Dublin, Trinity College, Dublin, Ireland, [4]Department of Genetics, University of Valencia, Burjasot, Valencia 46100, Spain, [5]Department of Biotechnology, Instituto de Agroquímica y Tecnología de los Alimentos, Consejo Superior de Investigaciones Científicas (CSIC), Burjasot, Valencia, Spain, and [6]Plant Functional Biology and Climate Change Cluster (C3), University of Technology Sydney, Sydney, NSW 2007, Australia

*To whom correspondence should be addressed. Tel. +34 96 3879934. Email: mfares@ibmcp.upv.es, faresm@tcd.ie; sabaterb.tcd@gmail.com

†These authors contributed equally to this work and should be considered co-first author.
Edited by Dr. Katsumi Isono

Received 10 February 2017; Editorial decision 5 May 2017; Accepted 11 May 2017

Abstract

Gene duplication is an important source of novelties and genome complexity. What genes are preserved as duplicated through long evolutionary times can shape the evolution of innovations. Identifying factors that influence gene duplicability is therefore an important aim in evolutionary biology. Here, we show that in the yeast *Saccharomyces cerevisiae* the levels of gene expression correlate with gene duplicability, its divergence, and transcriptional plasticity. Genes that were highly expressed before duplication are more likely to be preserved as duplicates for longer evolutionary times and wider phylogenetic ranges than genes that were lowly expressed. Duplicates with higher expression levels exhibit greater divergence between their gene copies. Duplicates that exhibit higher expression divergence are those enriched for TATA-containing promoters. These duplicates also show transcriptional plasticity, which seems to be involved in the origin of adaptations to environmental stresses in yeast. While the expression properties of genes strongly affect their duplicability, divergence and transcriptional plasticity are enhanced after gene duplication. We conclude that highly expressed genes are more likely to be preserved as duplicates due to their promoter architectures, their greater tolerance to expression noise, and their ability to reduce the noise-plasticity conflict.

Key words: gene expression, gene duplication, transcriptional plasticity, duplicability, *Saccharomyces cerevisiae*

1. Introduction

Gene duplication is believed to be a rich source of novel functions and adaptations.[1–3] This belief is supported by evidence coming from innovations following gene duplications in yeast, plants and animals. Indeed, protein families expanded after whole genome and small-scale duplications yielding an unprecedented morphological diversity in plants.[4–11] Other major innovations in animals have also been achieved through gene duplication,[12] including increased synapse and behaviour complexity[13] and the neural crest formation and plasticity in vertebrates.[14] In yeast, gene duplication has contributed to metabolic innovation through the alteration of regulatory and transcriptional networks[15] or the increased glycolytic fluxes.[16] However, it remains unclear why certain duplicates have been preferred over others to persist in the genomes and be the source of innovations.

Since duplication is immediately followed by relaxed selection constraints on one or the two gene copies, the survival time of each gene copy is a limiting factor in the determination of its functional fate. In the majority of cases, duplication is resolved by the non-functionalization of one of the gene copies and its subsequent erosion from the genome.[1,3] Accordingly, 92% of all genes that were duplicated through whole-genome duplication (WGD) > 100 MYA in *Saccharomyces* returned to single copy genes 'shortly' after duplication.[17] Nonetheless, in many species, including yeast, the number of duplicated genes is larger than predicted by theory ranging between 30% of the genes in yeast[18] and more than 50% in plants.[6,19] Determining what genes remain in the genome as duplicates, and consequently lead to evolutionary leaps, is an important aim in evolutionary biology. However, this objective remains to be achieved.

A number of hypotheses have been proposed to explain the persistence of certain genes in duplicate. Rapid sequence divergence between gene copies can lead to their functional divergence followed by strong selective constraints on each copy, which could contribute to the preservation of duplicates in the genomes.[20–23] Functional divergence requires, nevertheless, long evolutionary times and given that selection relaxes after gene duplication, selective pressures are unlikely to retain both gene copies during the first million years following duplication. Preservation of duplicates can also be selectively favoured by the need to maintain gene–dosage balance,[5,21,24,25] or provide genetic robustness against deleterious mutations.[26,27] However, all these scenarios do not provide a general mechanistic explanation for what makes duplicates persist or alternatively perish.

Recently, it has been proposed dosage sub-functionalization as a plausible hypothesis to explain the fate of whole-genome duplicates.[28] According to this hypothesis, highly expressed genes are more likely to be preserved as duplicates than lowly expressed genes. This is because stochastic variations in the levels of expression of the gene copies of highly expressed duplicates would not lead to copies with a lower expression level than that required for purifying selection to act upon them. Therefore, highly expressed duplicates are less likely to return to single copy genes by drift. Whether this hypothesis could be applied to all duplicates regardless of the mechanism that originated them and whether such dosage sub-functionalization could also determine the patterns of divergence between gene copies has not been explored before.

Here, we present evidence that the levels of gene expression are correlated with the fates of whole-genome and small-scale duplicates, with highly expressed genes being more likely to be retained in double copy after duplication for longer periods of time than lowly expressed genes. Such duplicates are also more phylogenetically stable. We also show that the ancestral levels of gene expression are correlated with the evolution of duplicates expression. Retained duplicated genes evolve strong patterns of transcriptional (also known as phenotypic) plasticity, which are also correlated with the levels of gene expression. Finally, while the levels of gene expression are correlated with the duplicability of genes, duplicates phenotypic plasticity is manifested only after gene duplication; and this plasticity is proportional to the expression divergence between the copies of duplicated genes.

2. Material and methods

2.1. Identification of duplicated genes

Paralogs pairs of duplicated genes were identified as the resulting best reciprocal hits from all-against-all BLAST searches using BLASTP with an *E*-value cutoff of 1E−5 and a 50 bit score.[29] Paralogs were then divided into two groups according to the mechanism of their origin: WGDs and SSDs. WGDs are those extracted from the reconciled list provided by the Yeast Gene Order Browser (YGOB, http://wolfe.gen.tcd.ie//ygob[30]) (555 pairs of genes), and these were not subjected to subsequent SSD. All other paralogs were considered to belong to the category of SSDs (560 pairs of genes).

2.2. Growth of *S. cerevisiae* and gene expression analyses

The transcriptomic profiling was performed in the *S. cerevisiae* Y06240 haploid *msh2* deletion strain (BY4741; *Mata his3D1 leu2DO met15DO ura3DO msh2::kanMX4*), with three technical replicates for each biological stress condition (3% lactic acid (YPL), 3% ethanol (YPE), 3% glycerol (YPG), 0.25mM H_2O_2 + 1.5% dextrose (YPOxD)) in comparison with the normal growth condition (YPD media) (Fig. 1). Total RNA extractions were performed with RNeasy kit (Qiagen) following manufacturer instructions. Ribosomal RNA was removed by using Ribo-Zero Gold rRNA removal yeast (Illumina) depletion kit. Stranded RNA libraries were constructed using TruSeq stranded mRNA (Illumina) from oligo-dT captured mRNAs from depleted samples. Libraries were run in NextSeq 500 (Illumina) at 75nt single read by using High Output 75 cycles kit v2.0 (Illumina).

RNA libraries were sequenced at Genomic core facility at Servicio Central de Soporte a la Investigación Experimental (SCSIE) from University of Valencia, Spain. Raw reads were analyzed using FastQC report and cleaned with CutAdapt as implemented in RobiNA software package v 1.2.4.[31] Low-quality reads were filtered and trimmed (Phred score inferior to 20 and size less than 40 nt were discarded). The reads were then aligned with Bowtie (up to two mismatches accepted) to the reference transcriptome (PRJNA290217) from the reference S288c strain. The normalization and statistical evaluation of differential gene expression has been performed using edgeR[32] or DESeq[33] with a *P*-value cut-off of 0.05, using the Benjamini–Yekutieli[34] method for multiple testing correction of *P*-value, and setting the Log-fold change at min = 1 to determine differential expression. The raw data (reads counts) was normalized according to the default procedure of the differential expression analysis package used (edgeR or DESeq), being the dispersion estimated using the pooled setting, and RPKM (Reads Per Billion) expression values estimated as implemented in RobiNA software.[31] All newly sequenced RNA sequences are available from the Sequence Read Archive with the following accession number (SRP074821). Expression data for each of the *S. cerevisiae* genes under YPD and each of the four stress conditions as well as the adjusted probabilities

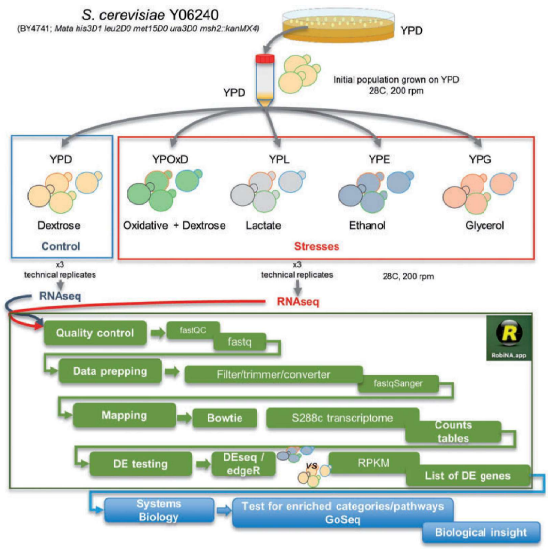


Figure 1. Growth experiments of *S. cerevisiae* under stress conditions. An initial isogenic population of *S. cerevisiae* (strain Y06240) was grown for 24 h in Yeast extract, peptone, dextrose medium (YPD) at 28 °C. This grown population was then subjected to four different stress conditions: ethanol, glycerol, lactate and oxidative stress in a medium supplemented with dextrose (each stress in indicated in a different colour). Growth experiments were performed in triplicate. Samples of each of the populations, in addition to the control population grown in YPD, were processed for RNA sequencing. The technical steps for the extraction and analysis of RNA sequences are also shown.

to identify significant fold changes are available in Supplementary Tables S1–S4.

2.3. Expression data for *Lachancea kluyveri*

Growth conditions, RNA extraction and sequencing are specified in a previous study.[35] Briefly, authors performed the analyses on *L. kluyveri* reference strain CBS 3082_a (*MAT*a). Transcriptomic data were obtained from growth cultures at mid exponential phase and for 20 different media, including YPD and 19 other stress conditions (listed in Ref. [35]). RNA sequencing was performed using Illumina HiSeq2000 platform with 50-base pair non-oriented single reads (Supplementary Table S5).

2.4. Expression data for *Candida glabrata*

Candida glabrata ATC2001 strain expression data (in the form of RPKMs) were obtained from a previous study.[36] Briefly, authors grown *C. glabrata* in normal YPD media and in the M199 medium at different pH values for pH shift. Total RNA was isolated by hot phenol–chlorophorm method. Libraries were subjected to ribosomal RNA depletion, and sequenced using Illumina HiSeq2000 platform with 100-bp paired-end strand-specific. Reads were mapped with TOPHAT2, and counted using htseq (-m union, -t exon conditions) and normalized by the number of reads per kilobase of exon per million mapped reads (Supplementary Table S6). Differentially expressed genes were identified with raw counts by DeSeq and EdgeR, using the same cut-off parameters than the ones used in this study.

2.5. Software

Calculations and statistics were performed using MS Excel and R 3.2.1. Data management was possible using in-house built PERL scripts.

3. Results

3.1. Duplicates preservation and phylogenetic stability are correlated with the levels of gene expression

Highly expressed genes are more likely to be preserved as duplicates after whole-genome duplication (WGD) than lowly expressed genes in *S. cerevisiae*.[37,28,38–41] We searched for the orthologs of *S. cerevisiae* genes in *L. kluyveri* (strain CBS 3082; synonymous of *Saccharomyces kluyveri*) and their expression in Yeast Extract Peptone containing dextrose medium (YPD)[35] (Supplementary Table S5). *Lachancea kluyveri is* a respiratory yeast species pre-dating the WGD that took place in *S. cerevisiae* > 100 million years ago.[17] We found *L. kluyveri* orthologs for 5643 *S. cerevisiae* genes. Of the 5643 genes, 1469 genes were orthologs of *S. cerevisiae* duplicates, including WGDs and small-scale duplicates (SSDs), and 4174 were

orthologs of *S. cerevisiae* singletons (Supplementary Table S5). The expression of *L. kluyveri* orthologs of *S. cerevisiae* duplicates (Median: 10.43; measured as the $\log_2$-transformed Reads Per billion, RPKM) was significantly greater than that of *L. kluyveri* orthologs of *S. cerevisiae* singletons (Median: 9.52) (Wilcoxon rank test: $P < 2.2 \times 10^{-16}$). We compared the transcription levels of genes in *S. cerevisiae* obtained in our study with those from another study[42] that used ribosomal profiling, a technique that measures ribosome occupancy and translation genome wide and provides an accurate measure of the translatable mRNA. RPKMs correlated strongly and significantly with the data of ribosome profiling (Spearman's correlation: $\rho = 0.77$, $P < 2.2 \times 10^{-16}$, Supplementary Data S1), indicating that RPKMs are indicative of the levels of gene expression and also the translatable mRNAs.

Previous studies concluded that highly expressed genes were more likely to be preserved as duplicates after WGDs because of absolute dosage constraints and constraints on dosage balance.[28,37,39,43–46] Indeed, we found that this trend is true for *L. kluyveri* orthologs of *S. cerevisiae* WGDs ($N = 561$, Median expression: 10.64, Wilcoxon rank test: $P < 2.2 \times 10^{-16}$) and also for orthologs of *S. cerevisiae* SSDs ($N = 908$, Median expression: 10.28, Wilcoxon rank test: $P < 2.2 \times 10^{-16}$). The level of expression of orthologs of WGDs was, nevertheless, higher than that for SSDs (Wilcoxon rank test: $P = 7.79 \times 10^{-5}$).

The higher expression of duplicates compared to singletons can be due to a greater presence of genes encoding protein-complex proteins among duplicates than singletons. We extracted the list of protein complexes from a previous study, using the table of annotated yeast high-throughput complexes available at (http://wodaklab.org/cyc2008/downloads).[47] Genes encoding proteins that are part of protein complexes ($N = 1913$) (Supplementary Table S7) do exhibit greater expression (Median expression: 11.56) than genes encoding complex-free proteins ($N = 3960$) (Median expression: 10.61, Wilcoxon rank test: $P < 2.2 \times 10^{-16}$). However, neither WGDs were more enriched for complex-encoding genes than singletons in *S. cerevisiae* (Fisher's exact test: $F = 1.02$, $P = 0.76$) nor SSDs showed significant difference in terms of enrichment for complex-encoding genes when compared to singletons (Fisher's exact test: $F = 1.11$, $P = 0.15$).

One caveat in this analysis is that gene expression in *L. kluyveri* may not reflect gene expression immediately after WGD. Against this prediction, gene expression in *L. kluyveri*, a species predating the WGDs and SSDs used in this study, was strongly and significantly correlated with gene expression in *S. cerevisiae* (Spearman correlation: $\rho = 0.59$, $P < 2.2 \times 10^{-16}$, Fig. 2a).

Duplicated genes exhibit different patterns of gene retention and phylogenetic stability in the different post-WGD *Saccharomyces* species.[48] We classified *S. cerevisiae* duplicated genes according to the presence of the two copies in each of the twelve available species post-dating the WGD (Fig. 2b). We first asked whether the expression of duplicates generated before *Saccharomycetales* speciation (including WGDs and SSDs) correlates with their phylogenetic stability, measured as the mean number of post-WGD species in which each copy was present (Fig. 2b). To determine the number of species post-dating WGD in which each gene copy is present we used the Pillars information available from the Yeast Gene Order Information,[30] which provides gene order and annotation for 12 post-WGD yeast species (Supplementary Table S8). For each gene copy, we counted the number of species in which it is found and averaged this number for the two sister gene copies in a duplicated pair (Fig. 2b). There was a positive and significant correlation between the mean gene copies expression in YPD and their phylogenetic stability (Spearman correlation: $\rho = 0.27$, $P < 2.2 \times 10^{-16}$, Fig. 2c). We then repeated the analysis for WGDs and SSDs separately. WGDs exhibited positive weak but significant correlation between gene expression and phylogenetic stability (Spearman correlation: $\rho = 0.13$, $P = 1.15 \times 10^{-5}$, Fig. 2d). In contrast, SSDs showed strong and significant correlation between gene expression and phylogenetic stability (Spearman correlation: $\rho = 0.40$, $P < 2.2 \times 10^{-16}$, Fig. 2e).

3.2. The magnitude of divergence of duplicates expression correlates with the level of gene expression

A pivotal hypothesis to the dosage sub-functionalization proposed by Gout and Lynch[28] is that highly expressed duplicates should exhibit more expression variation despite the action of purifying selection than lowly expressed genes. This is because noise in the expression of highly expressed genes is unlikely to compromise the selective constraints on these genes. That is, genes with higher expression levels should be more 'noisy' in their expression when duplicated, and thus they should generate more expression polymorphism in the population than lowly expressed genes. Accordingly, highly expressed duplicates are more likely to yield gene copies with diverged expressions than lowly expressed duplicates. To test this hypothesis, we first measured the fold expression difference (D) between the gene copies i and j when *S. cerevisiae* was grown under YPD as

$$D_{i,j} = 1 - [Min(E_i, E_j)]/[Max(E_i, E_j)]$$

with E referring to the expression of the gene under normal conditions (Supplementary Table S9). $D_{i,j}$ is normalized by the level of gene expression (i.e. the value is $0 \leq D_{i,j} \leq 1$), and thus it is an unbiased measure of the expression divergence between the gene copies. In support of our hypothesis, there was a weak but very significant correlation between the average expression of the gene copies of duplicates and $D_{i,j}$ (Spearman correlation: $\rho = 0.18$, $P = 4.23 \times 10^{-8}$). This correlation was also maintained when we analyzed separately WGDs (Spearman correlation: $\rho = 0.18$, $P = 6.98 \times 10^{-5}$) and SSDs (Spearman correlation: $\rho = 0.17$, $P = 2.89 \times 10^{-4}$).

3.3. The expression levels and promoter architecture correlate with patterns of expression divergence of duplicates and their transcriptional plasticity

Because higher expression can increase the chance for expression divergence after gene duplication, we sought to investigate if genes with higher expression can also evolve greater transcriptional plasticity under stress. Transcriptional plasticity is defined here as the ability of the gene to change its expression, while keeping its genotype, when the environment changes. For all four stresses with which *S. cerevisiae* was challenged (see section Material and methods), transcriptionally altered duplicates belonged to a set of genes with significantly higher expression in YPD growth media than transcriptionally unaltered duplicates (Fig. 3a–d). This was also true, with the exception of ethanol-induced stress, for singletons, albeit the effect was more pronounced in duplicates than it was in singletons. Noticeably, for all four-stress conditions in *S. cerevisiae*, the levels of expression of duplicates with no altered transcription under stress were significantly higher than that for transcriptionally altered singletons (Fig. 3a–d).

We explored other mechanistic explanations for this expression difference between unaltered duplicates and altered singletons. One important factor that contributes to transcriptional plasticity is the existence of the TATA-box motif in the gene promoter, with TATA-containing genes being more sensitive to regulatory changes than

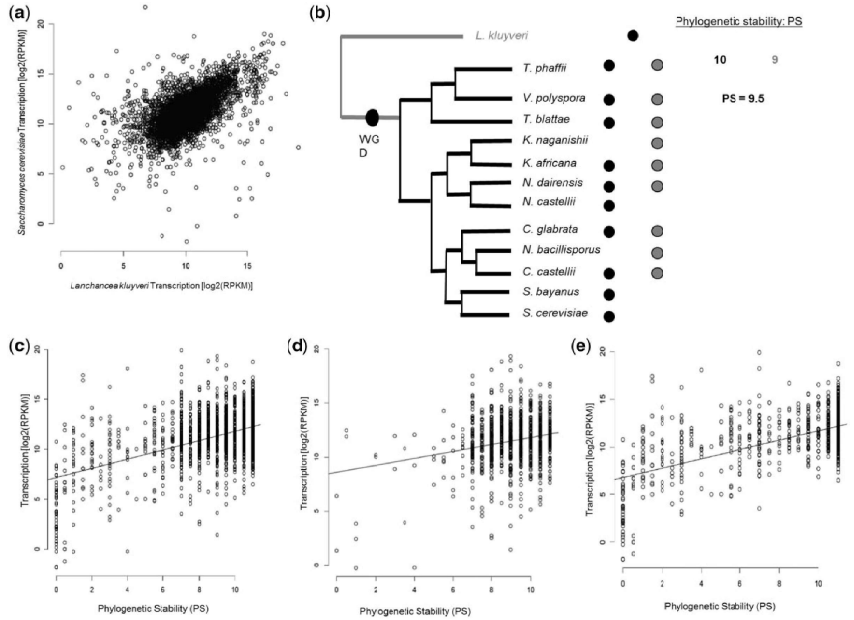


Figure 2. The levels of gene expression determine gene duplicability. (a) Expression of genes in the pre-WGD species *Lachancea kluyveri* correlate positively with genes expressions in the post-WGD species *Saccharomyces cerevisiae*. Axes represent the transcription levels of genes as the logarithmic transformation of the Fraction of Reads Per billion. (b) Estimation of the phylogenetic stability (i.e. number of species) of duplicated genes after the whole genome duplication in yeast. Black and gray circles refer to gene copies. A missing circle in one of the post-WGD species represent a gene copy loss, hence return of that duplicate to single-copy gene in that species. The phylogenetic stability (PS) of that duplicate is calculated as the mean number of species in which gene copies are present. (c) Phylogenetic stability correlates with the levels of gene expression of duplicates. (d) Correlation of PS of WGDs and their expression levels. (e) Correlation of PS of SSDs and their expression levels.

TATA-less genes.[49] Importantly, the level of expression of TATA-containing genes was higher than that of TATA-less genes, and this was true for duplicates and singletons in *S. cerevisiae* (Fig. 4a). The set of transcriptionally altered genes under stress conditions was enriched for TATA-containing genes when compared to the set of genes with no transcriptional plasticity, being this true for duplicates (Fig. 4b) and singletons (Fig. 4c). Generally, TATA-containing genes also exhibit expression noise, which can be coupled with transcriptional plasticity provided that noise and plasticity are not in conflict.[50] Since gene duplication relaxes noise-plasticity conflict,[50] we expected duplicated genes to be enriched for TATA motifs when compared to singletons. Of the 1090 genes containing TATA-motifs, 558 belonged to duplicates (281 were WGDs and 277 were SSDs) (25% of all duplicates) and 532 to singletons (12% of all singletons) (Supplementary Table S10). Indeed, duplicated genes were more enriched for TATA-containing genes than singletons (Fisher's exact test: $F = 2.52$, $P < 2.2 \times 10^{-16}$), and this was the case for both transcriptionally plastic genes and genes with no transcriptional plasticity (Supplementary Fig. S1a). Remarkably, duplicates with no transcriptional plasticity were slightly more enriched for TATA-containing genes than singletons with transcriptional plasticity, with the difference being significant in the case of *S. cerevisiae* grown under oxidative stress (Supplementary Fig. S1b).

Finally, the magnitude of expression divergence between duplicates gene copies was correlated with the magnitude of transcriptional plasticity (measured as the fold change in expression of the most altered gene copy between YPD and stress) (Table 1), both of which are in turn correlated with the levels of gene expression.

3.4. The levels of gene expression correlate with the patterns of duplicates transcriptional plasticity

A prediction of the dosage sub-functionalization hypothesis is that the patterns of duplicates transcriptional plasticity should be dependent on the levels of gene expression. For instance, transcriptionally plastic genes that are lowly expressed under normal conditions should only be able to over-express under stress because a decline in their expression could drive one of the gene copies to non-functionalization due to relaxed selective constraints. Because plasticity is often correlated with expression noise[50] and, since surviving

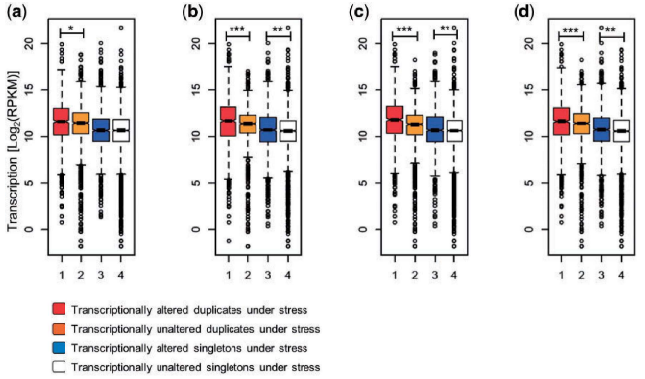


Figure 3. The levels of gene expression determine the transcriptional plasticity of genes. We present the levels of expression as the logarithmic transformation of the Reads Per billion (RPKM) for duplicates (red boxes) and singletons (blue boxes) with transcriptional alterations and duplicates (orange boxes) and singletons (white boxes) without altered expressions when *S. cerevisiae* is faced with four different stress conditions (a) Ethanol, (b) glycerol, (c) lactate and (d) oxidative stress in a medium supplemented with dextrose. Significant differences in the expression levels of sets of genes are indicated as *, **, ***, when the probabilities are $P < 0.05$, $P < 0.01$ and $P < 10^{-6}$, respectively, using a Wilcoxon rank test.

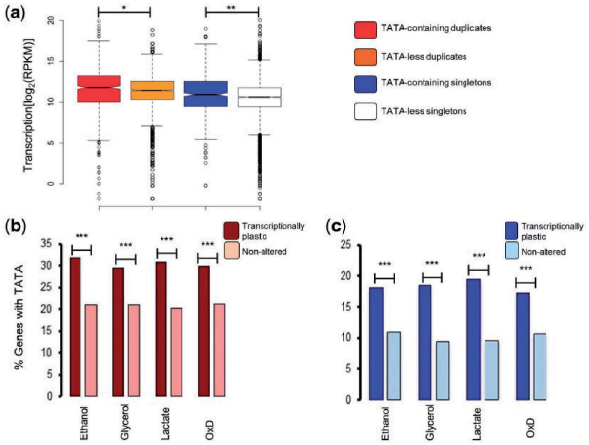


Figure 4. TATA-containing genes exhibit greater expression levels and transcriptional plasticity than TATA-less genes. (a) Comparison of the expression levels (measured as the logarithmic transformation of the Fragment Reads Per billion) between TATA-containing and TATA-less genes for duplicates (red and orange boxes, respectively) and singletons (blue and white boxes, respectively). Differences were tested using the Wilcoxon rank test and significant differences are identified as * and **, for probabilities of $P < 0.05$ and $P < 0.01$, respectively. The percentage of genes with TATA-containing promoters for transcriptionally plastic and transcriptionally unaltered genes when *S. cerevisiae* is growing under four different stress conditions (ethanol, glycerol, lactate and oxidative stress in a medium supplemented with dextrose) was compared for duplicates (b) and for singletons (c). Significant differences between plastic and unaltered genes were identified using Fisher's exact test and are identified as ***, to indicate a probability of $P < 2.2 \times 10^{-16}$.

duplicates are those whose expression noise falls within the range of expression detectable by selection, expression noise should depend on the levels of duplicates expression. We divided duplicated genes according to the patterns of transcriptional plasticity they show when *S. cerevisiae* is grown under stress: (i) up-regulated: when the two gene copies were up-regulated under stress; (ii) down-regulated: when the two gene copies were down-regulated under stress; (iii) discordant: when one copy was up-regulated and the other was

down-regulated under stress; and (iv) one-altered: when one copy was not altered but its paralogous copy was either up-regulated or down-regulated under stress (Supplementary Tables S11–S26). In all four stress conditions, duplicates that were down-regulated were also those that exhibited the highest expression levels under normal conditions, being these followed by duplicates in which only one copy exhibits transcriptional plasticity, then discordant duplicates and finally up-regulated duplicates (Fig. 5a–d).

Interestingly, the mean level of expression under normal conditions of duplicates gene copies was correlated with the level of expression divergence of the gene copies under normal conditions in those duplicates that belong to the category discordant and one-altered, those categories with the highest expression divergence between gene copies, but not in those pairs in which both copies were either up-regulated or down-regulated (Fig. 6).

3.5. Gene duplication has contributed to increased transcriptional plasticity in yeast

We examined the link between gene duplication and phenotypic plasticity by comparing the transcriptomes of *S. cerevisiae* grown under a number of key stress conditions that this species faces in nature to those transcriptomes of *S. cerevisiae* grown under normal YPD media (section Material and methods). Transcriptionally altered genes were more enriched for duplicates than for singleton genes over all stress conditions (Fig. 7). This trend was also true when we compared WGDs to singletons and SSDs to singletons (Fig. 7). To determine whether transcriptional plasticity is directly linked to gene duplication, we examined the transcriptional plasticity of *S. cerevisiae* duplicates and singletons orthologs in the pre-WGD species *L. kluyveri*. The transcription of *L. kluyveri* genes was previously assessed in 19 different stress conditions.[35] For each condition, we sought the percentage of genes that were orthologs to *S. cerevisiae* duplicates ($N=1469$) and singletons ($N=4174$) that exhibited transcriptional alteration. In 18 of the 19 conditions, there was no significant difference in the percentage of transcriptionally altered genes under stress between the orthologs of *S. cerevisiae* duplicates and those of singleton (Table 2). The only exception was SDS stress, in which the percentage of transcriptionally altered orthologs for *S. cerevisiae* duplicates was higher than that for transcriptionally altered singleton orthologs (Fisher's exact test: odds ratio $F=1.22$, $P=5\times10^{-3}$, Table 2). In all other stresses that were equivalent to the ones used in our experiments (e.g. glycerol, ethanol), there was no significant difference in the number of transcriptionally altered genes between orthologs of *S. cerevisiae* duplicates and singletons (Table 2). These data indicate that the high transcriptional plasticity of duplicates in *S. cerevisiae* was acquired after gene duplication.

To shed more light on the role of gene duplication in the acquisition of transcriptional plasticity, we examined the transcriptional patterns of a post-WGD species, *Candida glabrata*, in which some orthologs of *S. cerevisiae* duplicates are in single gene copy in *C. glabrata*, while others are preserved as duplicates and for which we had transcriptional information under acidic stress similar to our lactate stress dataset.[36] *Saccharomyces cerevisiae* orthologs in *C. glabrata* were identified using synteny information available in the Pillars of the Yeast Gene Order Browser (YGOB).[30] In total, we identified 4844 reliable *S. cerevisiae*: *C. glabrata* orthologs. Of these 4844 genes, 788 genes were duplicated in *C. glabrata*, of which 123 were orthologs of *S. cerevisiae* singletons. These 4844 genes included 1659 out of the 2240 duplicated *S. cerevisiae* genes and 3185 singletons. Of the 2240 *S. cerevisiae* duplicates, orthologs for 1019 of them were in single gene copy in *C. glabrata* (Fig. 7b). Importantly, these 1019 genes exhibited as much transcriptional plasticity under acidic stress in *C. glabrata* ($N=599$, 58.7% of the *C. glabrata* singletons) as *C. glabrata* singletons that had no duplicates orthologs in *S. cerevisiae* (1725 out of 3062 singletons, 56.3%) (Fisher's exact test: odd's ratio $F=1.10$, $P=0.17$). In conclusion, transcriptional plasticity seem to have been acquired after gene duplication because orthologs of duplicates that are in single copy genes in other species exhibit no evidence for transcriptional plasticity in these species.

Table 1. Expression divergence between gene copies correlates with transcriptional plasticity of duplicates

Stress source	Correlation (Spearman)	Probability
Ethanol	0.17	4.22×10^{-8}
Glycerol	0.20	1.78×10^{-10}
Lactate	0.16	4.13×10^{-7}
Oxidative + dextrose	0.15	7.11×10^{-7}

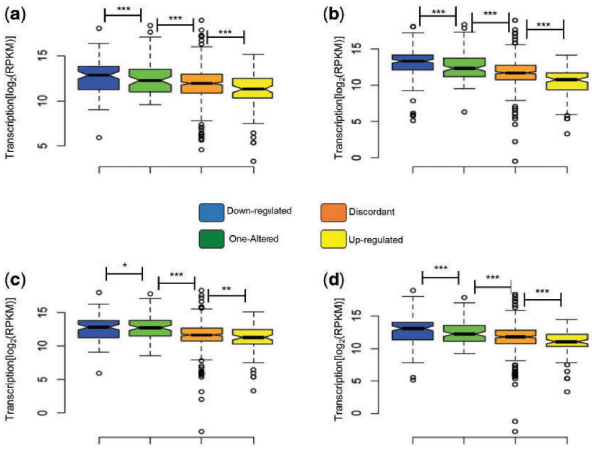


Figure 5. The expression level of duplicates correlated with their patterns of expression divergence. Expression was measured as the logarithmic transformation of the Reads Per billion (RPKM). Significant differences in the expression levels of gene sets are indicated as *, ** and ***, referring to probabilities of $P<0.05$, $P<0.01$ and $P<10^{-6}$, using a Wilcoxon rank test. (a) Expression under ethanol stress, (b) expression under glycerol stress, (c) expression under lactate stress and (d) expression under oxidative stress.

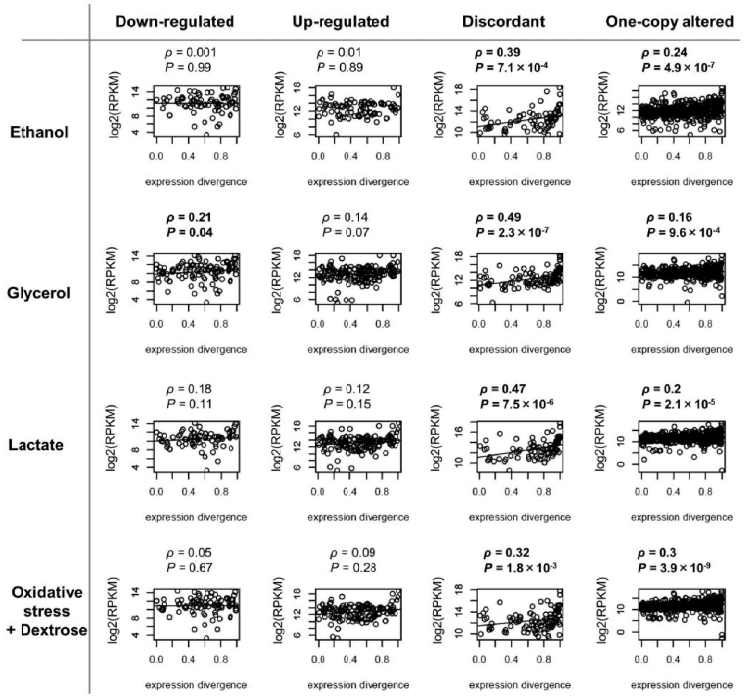


Figure 6. The levels of gene expression correlate with the level of expression divergence in duplicates in which gene copies are either discordant (one copy up-regulated and the other copy down-regulated) or one copy only altered under stress. Transcriptional alteration patterns of duplicates were measured in four stress conditions (ethanol, glycerol, lactate and oxidative stress in a medium supplemented with dextrose). Expression divergence of duplicates was measured as 1 – (the expression of the lowest expressed copy divided by the expression of the highest expressed copy). Spearman correlation (ρ) between the levels of gene expression (measured as the logarithmic transformation of the Reads Per billion (RPKM)) and the levels of expression divergence was tested for significance. Significant plots are indicated in bold. We found significant associations between the levels of gene expression and the expression divergence between gene copies of discordant duplicates and those with one copy altered but not for up-regulated or down-regulated duplicates.

3.6. Transcriptionally plastic duplicates contribute to the response of *S. cerevisiae* to stress

Among the transcriptionally plastic duplicates, many had a significant biological role in the response to stress. For instance, analyses of duplicated genes up-regulated when *S. cerevisiae* is grown in ethanol identified a number of genes as largely up-regulated that are involved in ethanol metabolism (Supplementary Table S27). Heading the list of up-regulated duplicates is the one encoding the alcohol dehydrogenase ADH2 and ADH1, directly involved in the quick metabolism of ethanol inside the cell into acetaldehyde (Fig. 8). A number of other duplicated genes that are essential in the metabolism of Ethanol, including transmembrane transporters such as YAT1 to start the tricarboxylic cycle or the enzyme MSL1 that is essential for malate production, among others (Fig. 8), they are all listed as duplicates with the highest up-regulation. This also applies to other stress conditions used in this study (Fig. 8).

4. Discussion

A number of factors have been considered key players in determining the fates of duplicated genes. However, an explanation that provides a general and inherent mechanism that strongly determines the duplicability of genes has been poorly investigated. Here, we present strong evidence that the levels of gene expression determine the likelihood of a gene to persist duplicated in the genome. We find that genes that are highly expressed in species pre-dating the whole-genome duplication event in *Saccharomyces* are more likely to be preserved in duplicate in species originated after duplication. This observation is in agreement with the dosage sub-functionalization hypothesis (DSH), as high levels of gene expression would ensure purifying selection on the gene copies despite the stochastic variation in their expression levels.[28] We, nevertheless, show that this is not a link exclusively seen in WGDs because the preservation of SSDs is also dependent on the expression levels of genes, however to a lesser extent than WGDs.

We also show that highly expressed genes exhibit greater phylogenetic stability (i.e. they are preserved in greater number of post-duplication species) than lowly expressed genes, perhaps due to the higher likelihood of functional divergence between gene copies of highly expressed duplicates, and thus emergence of purifying selection to maintain the functions of gene copies. Therefore, we conclude that the levels of gene expression determine the survival of duplicates across long evolutionary times. Our observations are not confounded by other factors such as dosage balance, mainly an issue in the case of genes encoding proteins that are part of protein complexes,[28,39,43,45,46] as neither WGDs nor SSDs are enriched for complex-encoding genes when compared to singletons in *S. cerevisiae*.

A pivotal conclusion derived from the link between gene expression and duplicability is that gene copies with higher expression levels can be noisier in terms of expression without undergoing an imbalance in their selective pressures. This noise could eventually lead to higher divergence between the gene copies of these duplicates when such divergence becomes adaptive. Our results are in agreement with this prediction and show that the magnitude of expression divergence, and perhaps functional divergence, is dependent on the levels of gene expression. Therefore, duplicates with higher expression are more likely to lead to higher divergence between the gene copies. This higher divergence between gene copies has been likely important for the origin of novel adaptations to stressful environments in the yeast *Saccharomyces cerevisiae*.[26,27,51–53] In agreement with this hypothesis, when *S. cerevisiae* was faced with a number of stress conditions, the levels of expression divergence between gene copies was positively correlated with their expression plasticity.

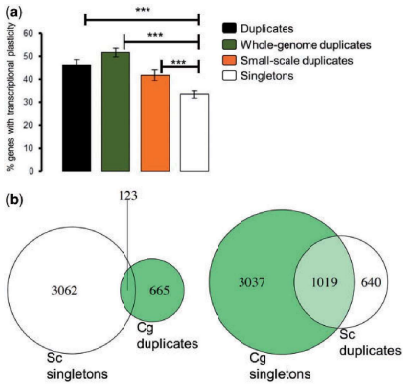


Figure 7. Duplicates are more enriched for transcriptional plasticity than singletons, regardless of the mechanism of duplication. (a) Transcriptional plasticity of duplicates and singletons in *S. cerevisiae*. *Y*-axis represents the mean percentage of genes with altered transcription across four stress conditions (ethanol, glycerol, lactate and oxidative stress in a medium supplemented with dextrose). We also compared WGDs and SSDs to singletons. Significant differences were found using Fisher's exact test, with *** indicating a probability of $P < 2.2 \times 10^{-16}$. (b) Venn diagram representing the overlap in the number of duplicates and singletons between *C. glabrata* and *S. cerevisiae*.

Table 2. Transcriptional alterations of *L. kluyveri* genes orthologs of *S. cerevisiae* under nineteen different stress conditions

Stress	# duplicates orthologs (%)	# singletons orthologs (%)	Odds' ratio *F*	*P*
Galactose	588 (40.1)	1586 (37.6)	0.93	0.24
Glycerol	865 (58.8)	2378 (56.9)	1.08	0.21
23 °C	178 (12.1)	540 (12.9)	0.92	0.44
37 °C	458 (31.2)	1282 (30.7)	1.02	0.74
YNB	487 (33.2)	1366 (32.7)	1.02	0.77
Ethanol	609 (41.5)	1645 (39.4)	1.09	0.17
Methanol	291 (19.8)	865 (20.7)	0.95	0.49
SDS	410 (27.9)	1008 (24.1)	1.22	0.005
DMSO	654 (44.5)	1758 (42.1)	1.10	0.11
NaCl	234 (15.9)	757 (18.1)	0.86	0.06
$CaCl_2$	874 (59.5)	2499 (59.9)	0.98	0.80
$NiSO_4$	129 (8.8)	332 (7.9)	1.11	0.32
LiCl	253 (17.2)	737 (17.7)	0.97	0.72
$CoSO_4$	825 (56.2)	2362 (56.6)	0.99	0.85
BME	375 (25.6)	1102 (26.4)	0.96	0.53
5FU	298 (20.3)	922 (22.1)	0.89	0.15
Arsenic	127 (8.6)	343 (8.2)	1.06	0.62
6AU	230 (15.7)	674 (16.1)	0.96	0.68
Fluconazole	437 (29.7)	1195 (28.6)	1.06	0.42

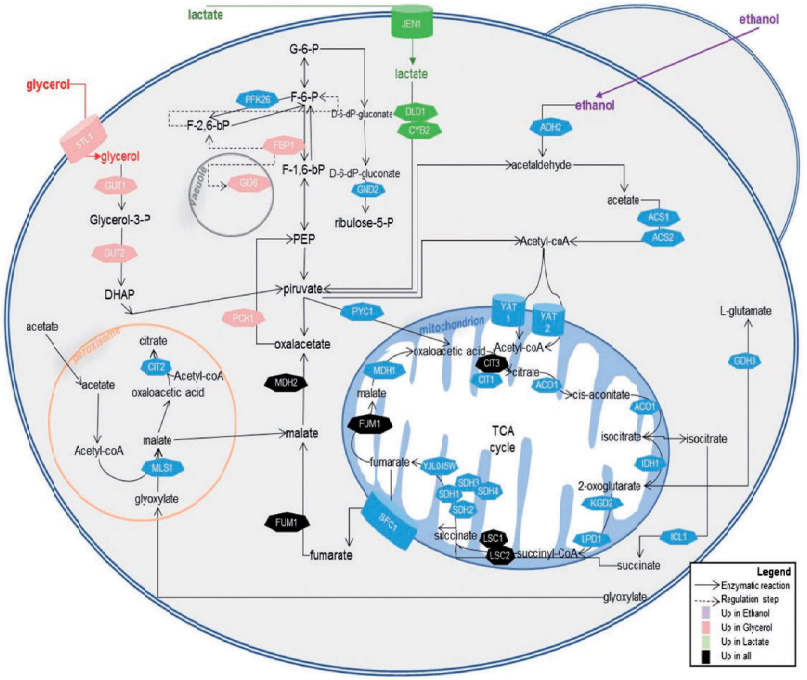


Figure 8. Schematic representation of *S. cerevisiae* central metabolic pathways for the utilization of non-fermentable carbon sources (glycerol, lactate and ethanol). Solid arrows correspond to enzymatic reactions while dashed arrows correspond to regulatory steps. Proteins encoded by duplicated genes that are transcriptionally altered and are involved in each process are indicated (transporters are shown with a can form, and enzymes exhibit heptagonal forms).

Importantly, we show that the levels of gene expression also influence the patterns of transcriptional divergence between the gene copies of duplicates. Those duplicates with copies exhibiting discordant expression patterns or with only one copy altered under stress are likely those in which each copy encodes a function under different conditions compared to the other copy.

Expression noise is generally coupled with plasticity if noise and plasticity do not present a cost-benefit conflict (i.e. a conflict emerges when plasticity is important for adaptation but noisy expression can be detrimental).[50,54,55] We show that duplicates are both more enriched for transcriptionally plastic genes and for genes with TATA motifs in their promoters when compared with singletons. TATA-motifs have been shown to be associated with higher noise and plasticity in TATA-containing genes.[49,56,57] The expression properties of genes that have been preserved as duplicates are, therefore, remarkably different from those that returned to single copy genes. Noticeably, non-plastic duplicates that showed higher expression levels than transcriptionally plastic singletons also exhibited greater enrichment for TATA-containing genes than plastic singletons, linking TATA-containing genes to higher levels of gene expression. The question that remains is then whether it is the expression properties or duplication itself that have determined the fate of duplicates and the origin of adaptations. Our results lead to the conclusion that duplicated genes exhibit significantly different expression properties than singleton genes but also that gene duplication is mainly responsible for the origin of plasticity, as the transcriptional plasticity observed in *S. cerevisiae* originated after the duplication of genes which were not transcriptionally more plastic before duplication than other non-duplicable genes. Therefore, gene duplication provides the appropriate genetic and selective opportunity for the evolution of transcriptional plasticity. A remarkable result is that the percentage of singletons transcriptionally altered in *S. cerevisiae* was significantly lower than that of their orthologs altered in *L. kluyveri*. However, the absolute percentages of altered genes cannot be compared between *L. kluyveri* and *S. cerevisiae* as both species represent completely different metabolisms and the conditions of stress under which they were subjected were different. It is, therefore, likely that duplication itself may relax the cost–benefit conflict between noise

and plasticity, as previously suggested,[50] allowing for the emergence of plasticity and adaptation to environmental perturbations. In conclusion, our result point to a strong correlation between the expression properties of genes, their duplicability, transcriptional plasticity, and ability to give rise to novel adaptations.

Acknowledgements

We would like to thank members of Fares' Lab for a careful reading and discussion of the results in the manuscript. We are also grateful to colleagues at Trinity College for helpful discussions. This work was supported by a grant from the Spanish Ministerio de Economía y Competitividad (MINECO-FEDER; BFU2015-66073-P) to M.A.F. F.M. is supported by a PhD grant from the Spanish Ministerio de Economía y Competitividad (reference: BES-2016-076677). C.T. was supported by a grant Juan de la Cierva from the Spanish Ministerio de Economía y Competitividad (reference: JCA-2012-14056).

Conflict of interest

None declared.

Accession numbers

All newly sequenced RNA sequences are available from the Sequence Read Archive with the following accession number (SRP074821).

Supplementary data

Supplementary data are available at *DNARES* online.

References

1. Ohno, S. 1970, *Evolution by Gene Duplication*. Springer Verlag, New York.
2. Ohno, S. 1999, Gene duplication and the uniqueness of vertebrate genomes circa 1970–1999, *Semin. Cell Dev. Biol.*, **10**, 517–22.
3. Lynch, M. and Conery, J.S. 2000, The evolutionary fate and consequences of duplicate genes, *Science*, **290**, 1151–5.
4. Otto, S.P. and Whitton, J. 2000, Polyploid incidence and evolution, *Annu. Rev. Genet.*, **34**, 401–37.
5. Carretero-Paulet, L. and Fares, M.A. 2012, Evolutionary dynamics and functional specialization of plant paralogs formed by whole and small-scale genome duplications, *Mol. Biol. Evol.*, **29**, 3541–51.
6. Cui, L., Wall, P.K., Leebens-Mack, J.H., et al. 2006, Widespread genome duplications throughout the history of flowering plants, *Genome Res.*, **16**, 738–49.
7. Holub, E.B. 2001, The arms race is ancient history in Arabidopsis, the wildflower, *Nat. Rev. Genet.*, **2**, 516–27.
8. Lespinet, O., Wolf, Y.I., Koonin, E.V. and Aravind, L. 2002, The role of lineage-specific gene family expansion in the evolution of eukaryotes, *Genome Res.*, **12**, 1048–59.
9. Wendel, J.F. 2000, Genome evolution in polyploids, *Plant Mol. Biol.*, **42**, 225–49.
10. Soltis, D.E., Albert, V.A., Leebens-Mack, J., et al. 2009, Polyploidy and angiosperm diversification, *Am. J. Bot.*, **96**, 336–48.
11. Van de Peer, Y. 2004, Computational approaches to unveiling ancient genome duplications, *Nat. Rev. Genet.*, **5**, 752–63.
12. Hoegg, S., Brinkmann, H., Taylor, J.S. and Meyer, A. 2004, Phylogenetic timing of the fish-specific genome duplication correlates with the diversification of teleost fish, *J. Mol. Evol.*, **59**, 190–203.
13. Grant, S.G. 2016, The molecular evolution of the vertebrate behavioural repertoire, *Philos. Trans. R. Soc. Lond. B: Biol. Sci.*, **371**, 20150051.
14. Green, S.A. and Bronner, M.E. 2013, Gene duplications and the early evolution of neural crest development, *Semin. Cell Dev. Biol.*, **24**, 95–100.
15. Huminiecki, L. and Conant, G.C. 2012, Polyploidy and the evolution of complex traits, *Int. J. Evol. Biol.*, **2012**, 292068.
16. Conant, G.C. and Wolfe, K.H. 2007, Increased glycolytic flux as an outcome of whole-genome duplication in yeast, *Mol. Syst. Biol.*, **3**, 129.
17. Wolfe, K.H. and Shields, D.C. 1997, Molecular evidence for an ancient duplication of the entire yeast genome, *Nature*, **387**, 708–13.
18. Fares, M.A., Keane, O.M., Toft, C., Carretero-Paulet, L. and Jones, G.W. 2013, The roles of whole-genome and small-scale duplications in the functional specialization of *Saccharomyces cerevisiae* genes, *PLoS Genet.*, **9**, e1003176.
19. Blanc, G. and Wolfe, K.H. 2004, Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes, *Plant Cell*, **16**, 1667–78.
20. Blanc, G. and Wolfe, K.H. 2004, Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution, *Plant Cell*, **16**, 1679–91.
21. Conant, G.C. and Wolfe, K.H. 2008, Turning a hobby into a job: how duplicated genes find new functions, *Nat. Rev. Genet.*, **9**, 938–50.
22. Fares, M.A., Byrne, K.P. and Wolfe, K.H. 2006, Rate asymmetry after genome duplication causes substantial long-branch attraction artifacts in the phylogeny of Saccharomyces species, *Mol. Biol. Evol.*, **23**, 245–53.
23. Scannell, D.R. and Wolfe, K.H. 2008, A burst of protein sequence evolution and a prolonged period of asymmetric evolution follow gene duplication in yeast, *Genome Res.*, **18**, 137–47.
24. Freeling, M. 2009, Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition, *Annu. Rev. Plant Biol.*, **60**, 433–53.
25. Conant, G.C., Birchler, J.A. and Pires, J.C. 2014, Dosage, duplication, and diploidization: clarifying the interplay of multiple models for duplicate gene evolution over time, *Curr. Opin. Plant Biol.*, **19**, 91–8.
26. Keane, O.M., Toft, C., Carretero-Paulet, L., Jones, G.W. and Fares, M.A. 2014, Preservation of genetic and regulatory robustness in ancient gene duplicates of *Saccharomyces cerevisiae*, *Genome Res.*, **24**, 1830–41.
27. Fares, M.A. 2015, The origins of mutational robustness, *Trends Genet.*, **31**, 373–81.
28. Gout, J.F. and Lynch, M. 2015, Maintenance and loss of duplicated genes by dosage subfunctionalization, *Mol. Biol. Evol.*, **32**, 2141–8.
29. Altschul, S.F., Madden, T.L., Schaffer, A.A., et al. 1997, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res.*, **25**, 3389–402.
30. Byrne, K.P. and Wolfe, K.H. 2005, The Yeast Gene Order Browser: combining curated homology and syntenic context reveals gene fate in polyploid species, *Genome Res.*, **15**, 1456–61.
31. Lohse, M., Bolger, A.M., Nagel, A., et al. 2012, RobiNA: a user-friendly, integrated software solution for RNA-Seq-based transcriptomics, *Nucleic Acids Res.*, **40**, W622–7.
32. Robinson, M.D., McCarthy, D.J. and Smyth, G.K. 2010, edgeR: a Bioconductor package for differential expression analysis of digital gene expression data, *Bioinformatics*, **26**, 139–40.
33. Anders, S. and Huber, W. 2010, Differential expression analysis for sequence count data, *Genome Biol.*, **11**, R106.
34. Benjamini, Y. and Yekutieli, D. 2001, The control of the false discovery rate in multiple testing under dependency, *Ann. Stat.*, **29**, 24.
35. Brion, C., Pflieger, D., Souali-Crespo, S., Friedrich, A. and Schacherer, J. 2016, Differences in environmental stress response among yeasts is consistent with species-specific lifestyles, *Mol. Biol. Cell*, **27**, 1694–705.
36. Linde, J., Duggan, S., Weber, M., et al. 2015, Defining the transcriptomic landscape of *Candida glabrata* by RNA-Seq, *Nucleic Acids Res.*, **43**, 1392–406.
37. Seoighe, C. and Wolfe, K.H. 1999, Yeast genome evolution in the post-genome era, *Curr. Opin. Microbiol.*, **2**, 548–54.
38. Aury, J.M., Jaillon, O., Duret, L., et al. 2006, Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*, *Nature*, **444**, 171–8.

39. Gout, J.F., Kahn, D., Duret, L. and Paramecium Post-Genomics, C. 2010, The relationship among gene expression, the evolution of gene dosage, and the rate of protein evolution, *PLoS Genet.*, **6**, e1000944.
40. McGrath, C.L., Gout, J.F., Doak, T.G., Yanagi, A. and Lynch, M. 2014, Insights into three whole-genome duplications gleaned from the *Paramecium caudatum* genome sequence, *Genetics*, **197**, 1417–28.
41. McGrath, C.L., Gout, J.F., Johri, P., Doak, T.G. and Lynch, M. 2014, Differential retention and divergent resolution of duplicate genes following whole-genome duplication, *Genome Res.*, **24**, 1665–75.
42. Albert, F.W., Muzzey, D., Weissman, J.S. and Kruglyak, L. 2014, Genetic influences on translation in yeast, *PLoS Genet.*, **10**, e1004692.
43. Gout, J.F., Duret, L. and Kahn, D. 2009, Differential retention of metabolic genes following whole-genome duplication, *Mol. Biol. Evol.*, **26**, 1067–72.
44. Papp, B., Pal, C. and Hurst, L.D. 2003, Dosage sensitivity and the evolution of gene families in yeast, *Nature*, **424**, 194–7.
45. Qian, W., Liao, B.Y., Chang, A.Y. and Zhang, J. 2010, Maintenance of duplicate genes and their functional redundancy by reduced expression, *Trends Genet.*, **26**, 425–30.
46. Birchler, J.A. and Veitia, R.A. 2012, Gene balance hypothesis: connecting issues of dosage sensitivity across biological disciplines, *Proc. Natl. Acad. Sci. USA.*, **109**, 14746–53.
47. Pu, S., Wong, J., Turner, B., Cho, E. and Wodak, S.J. 2009, Up-to-date catalogues of yeast protein complexes, *Nucleic Acids Res.*, **37**, 825–31.
48. Scannell, D.R., Byrne, K.P., Gordon, J.L., Wong, S. and Wolfe, K.H. 2006, Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts, *Nature*, **440**, 341–5.
49. Landry, C.R., Lemos, B., Rifkin, S.A., Dickinson, W.J. and Hartl, D.L. 2007, Genetic properties influencing the evolvability of gene expression, *Science*, **317**, 118–21.
50. Lehner, B. 2010, Conflict between noise and plasticity in yeast, *PLoS Genet.*, **6**, e1001185.
51. Mattenberger, F., Sabater-Munoz, B., Hallsworth, J.E. and Fares, M.A. 2017, Glycerol stress in Saccharomyces cerevisiae: cellular responses and evolved adaptations, *Environ. Microbiol*, **19**(3), 990–1007.
52. Mattenberger, F., Sabater-Munoz, B., Toft, C. and Fares, M.A. 2017, The phenotypic plasticity of duplicated genes in *Saccharomyces cerevisiae* and the origin of adaptations, *G3 (Bethesda)*, **7**, 63–75.
53. Mattenberger, F., Sabater-Munoz, B., Hallsworth, J.E. and Fares, M.A. 2017, Glycerol stress in Saccharomyces cerevisiae: cellular responses and evolved adaptations, *Environ. Microbiol.*, **19**, 990–1007.
54. Blake, W.J., Balazsi, G., Kohanski, M.A., et al. 2006, Phenotypic consequences of promoter-mediated transcriptional noise, *Mol. Cell*, **24**, 853–65.
55. Raser, J.M. and O'Shea, E.K. 2005, Noise in gene expression: origins, consequences, and control, *Science*, **309**, 2010–3.
56. Newman, J.R., Ghaemmaghami, S., Ihmels, J., et al. 2006, Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise, *Nature*, **441**, 840–6.
57. Tirosh, I., Weinberger, A., Carmi, M. and Barkai, N. 2006, A genetic signature of interspecies variations in gene expression, *Nat. Genet.*, **38**, 830–4.

The Phenotypic Plasticity of Duplicated Genes in *Saccharomyces cerevisiae* and the Origin of Adaptations

Florian Mattenberger,*,[1] Beatriz Sabater-Muñoz,*,†,[1] Christina Toft,‡,§ and Mario A. Fares*,†,[2]

*Department of Abiotic Stress, Instituto de Biología Molecular y Celular de Plantas, Consejo Superior de Investigaciones Científicas, Universidad Politécnica de Valencia, 46022 Spain, †Department of Genetics, Smurfit Institute of Genetics, University of Dublin, Trinity College, Dublin 2, Ireland, ‡Department of Genetics, University of Valencia, 46100 Burjassot, Spain, and §Departamento de Biotecnología, Instituto de Agroquímica y Tecnología de los Alimentos, Consejo Superior de Investigaciones Científicas, Valencia, 46980 Paterna, Spain

ORCID IDs: 0000-0002-2727-0284 (F.M.); 0000-0002-0301-215X (B.S.-M.); 0000-0003-1714-6703 (C.T.); 0000-0002-4345-3013 (M.A.F.)

ABSTRACT Gene and genome duplication are the major sources of biological innovations in plants and animals. Functional and transcriptional divergence between the copies after gene duplication has been considered the main driver of innovations . However, here we show that increased phenotypic plasticity after duplication plays a more major role than thought before in the origin of adaptations. We perform an exhaustive analysis of the transcriptional alterations of duplicated genes in the unicellular eukaryote *Saccharomyces cerevisiae* when challenged with five different environmental stresses. Analysis of the transcriptomes of yeast shows that gene duplication increases the transcriptional response to environmental changes, with duplicated genes exhibiting signatures of adaptive transcriptional patterns in response to stress. The mechanism of duplication matters, with whole-genome duplicates being more transcriptionally altered than small-scale duplicates. The predominant transcriptional pattern follows the classic theory of evolution by gene duplication; with one gene copy remaining unaltered under stress, while its sister copy presents large transcriptional plasticity and a prominent role in adaptation. Moreover, we find additional transcriptional profiles that are suggestive of neo- and subfunctionalization of duplicate gene copies. These patterns are strongly correlated with the functional dependencies and sequence divergence profiles of gene copies. We show that, unlike singletons, duplicates respond more specifically to stress, supporting the role of natural selection in the transcriptional plasticity of duplicates. Our results reveal the underlying transcriptional complexity of duplicated genes and its role in the origin of adaptations.

KEYWORDS
evolutionary biology
gene function
small-scale duplicates
whole-genome duplicates
transcriptional profiles

Gene duplication has been a major driving force of biological innovation in plants (Cui *et al.* 2006; Carretero-Paulet and Fares 2012; Holub 2001; Lespinet *et al.* 2002; Otto and Whitton 2000; Wendel 2000; Kim *et al.* 2004) and animals (Otto and Whitton 2000; Hoegg *et al.* 2004). Arguably, understanding how gene duplication gives origin to novel functions and adaptations is a fundamental aim of evolutionary biology. The functional and transcriptional divergence between the gene copies of a duplicated gene has been proposed to facilitate the origin of novel functions (Conant and Wolfe 2008; Lynch and Conery 2000; Ohno 1999, 1970). However, the tempo and mode of each divergence kind and the interplay between both remains largely unexplored.

Ohno proposed that after the duplication of a gene, the emerging genetic redundancy leads to relaxed selection against one of the gene copies while the other copy remains under strong purifying selection (Ohno 1970, 1999). The selectively relaxed gene copy explores novel genotypes, many of which will be deleterious and lead to the loss of the rapidly evolving gene copy (Lynch and Conery 2003). A less likely scenario is the preservation of both copies by purifying selection after a period of relaxed selection leading to novel functions in the form of sub- or neo-functionalization (Ohno 1970, 1999; Lynch and Conery 2003; Taylor and Raes 2004). Particular scenarios for this general model of the functional divergence of gene copies have been proposed (Des Marais and Rausher 2008; Force *et al.* 1999; Innan and Kondrashov 2010). Classic theory has also given credit to the expression divergence between gene copies as a prerequisite for the preservation of genes in duplicate and the eventual finding of new functions (Ferris and Whitt 1979; Force *et al.* 1999; Ohno 1970). Moreover, previous studies have found a genome-wide transcriptional response of *Saccharomyces cerevisiae* to a wide range of environmental perturbations (Ferea *et al.* 1999; Causton *et al.* 2001; Cormier *et al.* 2010; Ideker *et al.* 2001; Landry *et al.* 2006; Stern *et al.* 2007).

The rapid evolution of gene expression after duplication (Li *et al.* 2005; Thompson *et al.* 2013) suggests an adaptive role for the transcriptional plasticity of duplicates. However, the question remains open whether duplicates follow the general response patterns to stresses that are shown by singleton genes or, alternatively, they have allowed the origin of stress-specific adaptations that have been favored by natural selection. It also remains obscure whether the transcriptional plasticity of duplicates has driven their functional specialization. Understanding this plasticity through studies like the one conducted here provides a much wider picture of the role of gene duplication in the origin of adaptations and ecological diversification.

Gene duplication in plants has been followed by rapid expression divergence between gene copies (Blanc and Wolfe 2004; Ha *et al.* 2007, 2009; Wang *et al.* 2012). Since most duplicated genes are thought to mediate the interaction between the organism and environment, their expression changes have been suggested to be strongly linked to generating adaptations rather than responding to developmental perturbations (Ha *et al.* 2007). Most importantly, expression divergence has been seen to correlate with the sequence divergence between duplicate gene copies in plants (Blanc and Wolfe 2004) and, although less clearly (Wagner 2000a), in yeast (Gu *et al.* 2002). Two questions remain unexplored: (a) are duplicated genes more transcriptionally plastic than anticipated?; and (b) does transcriptional plasticity determine the functional fates of gene copies? Answering these questions would reveal the potential of gene duplicates to expedite adaptations.

The Baker's yeast *S. cerevisiae* duplicated its genome roughly 100 MYA (Wolfe and Shields 1997) triggered by the possible hybridization between different yeast species (Marcet-Houben and Gabaldon 2015; Wolfe 2015). Only 1120 pairs of duplicates have been retained, of which 554 belong to the whole-genome duplication event and the remaining are classified as duplications of small scale (Fares *et al.* 2013). Many of the yeast-duplicated genes enable the growth of *S. cerevisiae* under stressful conditions, the genetic basis of which has enabled the exploitation of the biotechnological benefits of yeast in the multimillionaire wine industry. The genetic and biotechnological properties of this yeast offer a unique opportunity to study the role of gene duplication in innovation. In this study, we explore whether the transcriptional plasticity of duplicated genes in *S. cerevisiae* has contributed to the origin of adaptations to stress and functional specialization of duplicate gene copies. We address this question by exhaustively and extensively analyzing the expression pattern dynamics of duplicated genes in the yeast *S. cerevisiae* after subjecting it to a number of stress conditions. Here, we find that not only duplicates are more transcriptionally polymorphic as concluded before (Ha *et al.* 2009) but that they are more transcriptionally plastic than singletons under environmental stress. This transcriptional plasticity increases after gene duplication and it is strongly correlated with the functional divergence of duplicate gene copies. The study of the patterns of sequence divergence, functional interactions, and transcriptional plasticity of duplicates makes possible the identification of stress-specific as well as general transcriptional response patterns. We show that, unlike singleton genes, duplicates have given origin to stress-specific adaptations. Our data describe a complex dynamic of transcriptional evolution following the gene and genome duplications of a simple eukaryotic organism and reveal the origins of yeast adaptations.

MATERIALS AND METHODS

Identification of duplicated genes

Paralog pairs of duplicated genes were identified as the resulting best reciprocal hits from all-against-all BLAST searches using BLASTP with an E-value cutoff of 1E-5 and a 50 bit score (Altschul *et al.* 1997). Paralogs were then divided into two groups according to the mechanism of their origin: whole-genome duplications (WGDs) and small-scale duplications (SSDs). WGDs are those extracted from the reconciled list provided by the Yeast Gene Order Browser (YGOB, http://wolfe.gen.tcd.ie//ygob; Byrne and Wolfe 2005) (555 pairs of genes), and these were not subjected to subsequent SSD. All other paralogs were considered to belong to the category of SSDs (560 pairs of genes). The duplicates used in this study have been estimated to have their origin on the time point of the WGD that took place 100 MYA (Wolfe and Shields 1997). Also, in this study we have used the SSDs that exhibit similar distribution of synonymous substitutions as those of WGDs, so roughly belonging to the same age (Fares *et al.* 2013; Keane *et al.* 2014).

Sequence alignments and analysis of divergence

For each protein-coding gene of *S. cerevisiae* we searched for its ortholog in the closely related species *S. paradoxus* using the program blastP. Pairwise sequence alignments were built using the program ClustalW. To calculate the distance between *S. cerevisiae* and *S. paradoxus* for each of the genes, we estimated the number of nonsynonymous nucleotide substitutions per nonsynonymous site (d_N), synonymous substitutions per synonymous site (d_S), and the nonsynonymous-to-synonymous rates ratio ($\omega = d_N/d_S$) using the maximum-likelihood approach under the Goldman and Yang model (Goldman and Yang 1994) as implemented in the PAML package version 4.7 (Yang 2007).

Analysis of gene expression in S. cerevisiae

The transcriptomic profiling was performed in the *S. cerevisiae* Y06240 haploid *msh2* deletion strain (BY4741; *Mata*; *his3D1*; *leud2DO*; *met15DO*; *ura3DO*; *msh2*::*kanMX4*) (Fares *et al.* 2013), with three technical replicates for each biological stress condition [3% lactic acid (YPL), 3% ethanol (YPE), 3% glycerol (YPG), 0.25 mM H_2O_2 (YPOx), 0.25 mM H_2O_2 + 1.5% dextrose (YPOxD)] in comparison with the normal growth condition (Yeast extract, Peptone, Dextrose media). Total RNA extractions were performed with RNeasy kit (Qiagen) following manufacturer instructions. Ribosomal RNA was removed using the Ribo-Zero Gold rRNA removal yeast (Illumina) depletion kit. Stranded RNA libraries were constructed using TruSeq stranded mRNA (Illumina) from oligo-dT captured mRNAs from depleted samples. Libraries were run in NextSequation 500 (Illumina) at 75 nt single read using High Output 75 cycles kit v2.0 (Illumina).

The treatment of the RNA libraries was done following a previous study in which different methods of differential expression analyses were compared (Zhang *et al.* 2014). RNA libraries were sequenced at the

doi: 10.1534/g3.116.035329
Manuscript received September 7, 2016; accepted for publication October 23, 2016; published Early Online October 31, 2016.

Supplemental material is available online at www.g3journal.org/lookup/suppl/doi:10.1534/g3.116.035329/-/DC1.
[1]These authors contributed equally to this work.
[2]Corresponding author: Instituto de Biología Molecular y Celular de Plantas, C/Ingeniero Fausto Elio s/n, 46022 Valencia, Spain. E-mail: mfares@ibmcp.upv.es

Table 1 Transcription alterations under stress conditions

Stress	Comparison	Number of Genes of First Type (%)	Number of Genes of Second Type (%)	Odds Ratio (F)	Probability
Ethanol	D[a] *vs.* S[b]	907 (40.5%)	1341 (29.3%)	1.64	$<2.2 \times 10^{-16}$
	WGDs[c] *vs.* S	515 (47.6%)	1341 (29.3%)	2.12	$<2.2 \times 10^{-16}$
	SSDs[d] *vs.* S	392 (33.9%)	1341 (29.3%)	1.27	8.1×10^{-4}
	WGDs *vs.* SSDs	515 (47.6%)	392 (33.9%)	1.68	2.7×10^{-9}
Glycerol	D *vs.* S	1134 (50.6%)	1693 (36.9%)	1.75	$<2.2 \times 10^{-16}$
	WGDs *vs.* S	617 (56.9%)	1693 (36.9%)	2.18	$<2.2 \times 10^{-16}$
	SSDs *vs.* S	517 (44.7%)	1693 (36.9%)	1.41	2.3×10^{-7}
	WGDs *vs.* SSDs	617 (56.9%)	517 (44.7%)	1.10	0.21
Lactate	D *vs.* S	1038 (46.3%)	1471 (32.1%)	1.83	$<2.2 \times 10^{-16}$
	WGDs *vs.* S	571 (52.7%)	1471 (32.1%)	2.28	$<2.2 \times 10^{-16}$
	SSDs *vs.* S	467 (40.4%)	1471 (32.1%)	1.47	2.6×10^{-8}
	WGDs *vs.* SSDs	571 (52.7%)	467 (40.4%)	1.56	2.3×10^{-7}
Oxidative	D *vs.* S	42 (1.9%)	59 (1.3%)	1.46	0.06
	WGDs *vs.* S	29 (2.7%)	59 (1.3%)	2.07	0.002
	SSDs *vs.* S	12 (1.1%)	59 (1.3%)	0.82	0.65
	WGDs *vs.* SSDs	29 (2.7%)	12 (1.1%)	2.54	0.007
Oxidative + dextrose	D *vs.* S	1064 (47.5%)	1574 (34.4%)	1.73	$<2.2 \times 10^{-16}$
	WGDs *vs.* S	589 (54.4%)	1574 (34.4%)	2.20	$<2.2 \times 10^{-16}$
	SSDs *vs.* S	475 (41.1%)	1574 (34.4%)	1.36	5.1×10^{-6}
	WGDs *vs.* SSDs	589 (54.4%)	475 (41.1%)	1.61	2.2×10^{-8}

[a] Duplicated genes.
[b] Singleton genes.
[c] Whole-genome duplicates.
[d] Small-scale duplicates.

Genomic core facility at Servicio Central de Soporte a la Investigación Experimental (SCSIE), University of Valencia, Spain. Raw reads were analyzed using FastQC report and cleaned with CutAdapt as implemented in RobiNA software package v 1.2.4 (Lohse *et al.* 2012). Low-quality reads were filtered and trimmed (Phred score <20 and size <40 nt were discarded). Since we had a reference transcriptome from S288c strain, reads were then aligned with Bowtie (up to two mismatches accepted) to the reference transcriptome (PRJNA290217) from the reference S288c strain. Statistical assessment of differential gene expression was done either with edgeR (Robinson *et al.* 2010) or with DESeq (Anders and Huber 2010) as implemented in RobiNA. A previous study compared the different expression analysis methods, concluding that edgeR and DESeq were the best-performing methods when the objective is to analyze differential expression (Zhang *et al.* 2014). Comparison of logarithmic fold-change of our expression data between edgeR and DESeq provided very strong correlation (Spearman correlation coefficient: $\rho = 0.995$, $P < 2.2 \times 10^{-16}$, Figure 1 of Supplemental Material, File S8). Significant expression changes were identified using a false discovery rate (FDR <0.05). These results indicate that our quantification of expression data is robust to the method used. All newly sequenced RNA sequences are available from the Sequence Read Archive with the following accession number: SRP074821.

Genetic interaction data

We used the latest update of the genetic functional chart of *S. cerevisiae* (Costanzo *et al.* 2010) (File S4 and File S5 from http://drygin.ccbr.utoronto.ca/~costanzo/). The genetic map is based on the synthetic genetic array methodology (Tong *et al.* 2001). In this methodology, synthetic lethal genetic interactions are systematically mapped to single and double mutants. In this study, two genes are considered to interact genetically if the double knockout mutant of the two genes has significantly larger or smaller effect than the multiplicative effects of simple knockouts.

Software

Calculations and statistics were performed using MS Excel and R 3.2.1, unless otherwise indicated. Data management was possible using in-house built Perl scripts.

Data availability

Strains are available upon request. All RNA sequences are available from the Sequence Read Archive (accession number SRP074821). File S1 contains the significant transcriptionally altered genes in *S. cerevisiae* upon growing in ethanol stress. File S2 contains the significant transcriptionally altered genes in *S. cerevisiae* upon growing in glycerol stress. File S3 contains the significant transcriptionally altered genes in *S. cerevisiae* upon growing in acidic stress. File S4 contains the significant transcriptionally altered genes in *S. cerevisiae* upon growing in oxidative stress. File S5 contains the significant transcriptionally altered genes in *S. cerevisiae* upon growing in oxidative stress in a growth medium supplemented with dextrose. File S6 contains the list of duplicated genes in *Candida glabrata*. File S7 contains the conservation indices of promoter alignments for duplicated genes with altered transcriptional profiles. File S8 compares the methods edgeR and DESeq for the calculation of reads mapped to each gene.

RESULTS

To test the role of duplicated genes of *S. cerevisiae* in the origin of adaptations, we sequenced the transcriptome of a haploid *msh2* deletion strain after growing it in normal YPD medium and under five different stress conditions: (a) ethanol, (b) glycerol, (c) lactate, (d) oxidative stress, and (e) oxidative stress in a medium supplemented with dextrose (see *Materials and Methods*). Subsequently, we compared the transcriptional modifications of *S. cerevisiae* *msh2*::*kanMX4* under each of the conditions and sought to investigate the role of duplicated genes in displaying transcriptional plasticity under stress. We used this strain because it has been allowed to evolve for hundreds of generations in YPD medium, hence is adapted to this medium, and allows the maintenance of population genetic polymorphism due to its higher mutation rate compared to the wild-type strain.

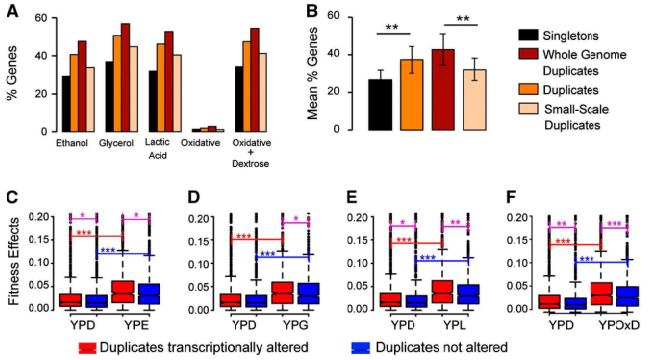


Figure 1 Duplicated genes exhibit higher transcriptional plasticity than singleton genes and are involved in adaptation. (A) Percentage of genes with transcriptional flexibility when *S. cerevisiae* is grown under each of the five stress conditions tested in this study: ethanol stress, glycerol stress, acidic stress by lactate, oxidative stress, and oxidative stress in a medium supplemented with dextrose. (B) The mean percentage of genes of the categories' singletons (black bar), duplicates (orange bar), duplicates generated by whole genome duplication (red bar), and duplicates generated by small-scale duplication (light yellow bar) with transcriptional alterations in the five stress conditions tested in this study. (C–F) We measured the contribution of transcriptionally altered duplicates to the fitness of *S. cerevisiae* under YPD and stress growth conditions using knock down gene data from Steinmetz *et al.* (2002). We then compared the fitness contribution of these altered duplicates (red boxes) with that of duplicates with no evidence for transcriptional plasticity under stress (blue boxes). These comparisons were performed for the sets of altered and not-altered duplicates identified under ethanol stress (C), glycerol stress (D), lactate stress (E), and oxidative stress supplemented with dextrose (F). Significant differences are indicated with *, **, and *** when the difference was significant at the levels of 0.01, 0.001, and 0.0001, respectively, using a Mann–Whitney *U*-test.

Duplicated genes exhibit significant transcriptional plasticity under stress

After growing biological replicates of the yeast populations under normal and each of the five different stress conditions, we extracted total RNA for RNAseq library construction and identified differentially expressed (DE) genes based on the comparison of their expression levels relative to normal conditions (see *Materials and Methods*). In total, we obtained reliable RNA sequence data for 5825 genes in the YPD medium (normal conditions) and each of the five stress conditions (File S1, File S2, File S3, File S4, and File S5), of which an important fraction was significantly altered under stress conditions (Table 1).

In all stress conditions, there was a significant transcriptomic response affecting 2248, 2827, 2509, 101, and 2638 genes under stress induced by ethanol, glycerol, lactate, oxidative stress, and oxidative stress in a medium supplemented with dextrose, respectively (File S1, File S2, File S3, File S4, and File S5). The response affected duplicates significantly more than singletons in all stress conditions, although the low number of altered genes limited the power of the test in the case of oxidative stress (Table 1 and Figure 1A). The mechanism of duplication, including WGD and SSD, also made a difference, with WGDs being more significantly enriched for altered expression genes than SSDs (Table 1 and Figure 1B). Taking all nonredundant transcriptomic responses together for all stress conditions, duplicated genes showed significantly larger increments of transcription under stress than singleton genes (Figure 1B). Indeed, on average 837 duplicated genes out of the 2240 duplicates (37.4%) in *S. cerevisiae* exhibited significant increments of expression against 1227 out of 4580 singletons (26.8%) (Fisher's exact test: odds ratio $F = 1.63$, $P < 2.2 \times 10^{-16}$). We identified an average of 42.2%, over all the stresses of 464 duplicates out of 1100 WGDs, to be transcriptionally plastic, a proportion significantly higher than that for singletons (Fisher's exact test: odds ratio $F = 1.99$, $P < 2.2 \times 10^{-16}$). Likewise, the proportion of transcriptionally plastic SSDs (an average of 372 out of 1140 SSDs, 32.6%) was significantly higher than that for singletons (Fisher's exact test: odds ratio $F = 1.32$, $P = 1.1 \times 10^{-4}$). WGDs also presented a significantly higher proportion of transcriptionally plastic genes than SSDs (Fisher's exact test: odds ratio $F = 1.51$, $P = 3.5 \times 10^{-6}$).

An alternative explanation for the adaptive value of the transcriptional plasticity of duplicates under stress is a low contribution of duplicates to fitness, with their response being the reflection of transcriptional noise caused by environmental perturbations. To test this possibility, we calculated the contribution of each duplicated gene to the fitness of *S. cerevisiae* under normal and stress conditions taking previously published fitness data (Steinmetz *et al.* 2002). To this end, we subtracted the normalized fitness values of a strain after a gene was deleted or knocked down under stress from the fitness of its ancestral strain (*i.e.*, 1). Therefore, large fitness absolute increment values indicate that the contribution of the gene to fitness is high under those conditions.

We first compared the contribution to fitness of duplicates with and without altered transcriptomic profiles in YPD. Transcriptionally altered duplicates exhibited higher contribution to fitness than unaltered duplicates (Figure 1, C–F), discarding the possibility that altered duplicates may have less contribution to fitness than not-altered duplicates consequently being less selectively constrained to change. Both transcriptionally altered and not-altered duplicates showed a significant increase in their contribution to fitness under stress (Figure 1, C–F). However, this increase was sharper in transcriptionally altered duplicates than in unaltered duplicates.

Increased transcriptional plasticity after gene duplication

The higher transcriptional plasticity of duplicated genes in *S. cerevisiae* when compared to singletons may be the result of a biased preservation in duplicate of highly transcriptionally plastic genes. To test whether

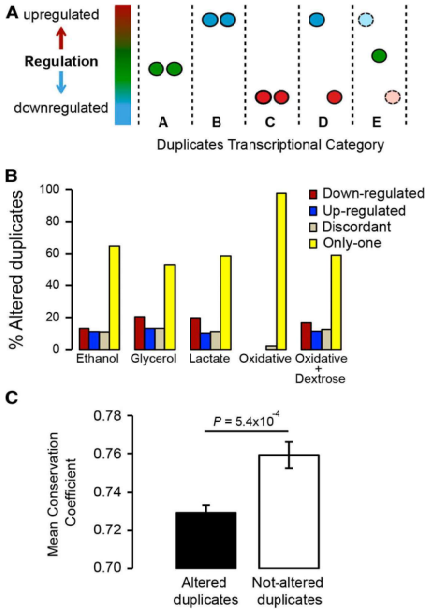


Figure 2 Duplicated genes exhibit differential patterns of transcriptional plasticity under stress. We identified five patterns of transcriptional plasticity for the duplicates of *S. cerevisiae* growing under stress conditions (A), including duplicates in which neither copy has been altered (category A), those with both copies up-regulated (category B), those with both copies down-regulated (category C), those with gene copies showing discordant transcriptional plasticities (category D), and those in which only one gene copy is altered while its sister copy is unaltered under stress (category E). Calculating the percentage of the duplicated genes belonging to each of the transcriptional categories (B), we found that the category with only one copy altered (yellow bar) is the one showing the highest percentage of the altered duplicates under stress. (C) We measured the conservation of the promoter regions for altered and not-altered duplicates and found that altered duplicates (black bar) exhibit lower conservation than not-altered duplicates under stress (white bar).

gene duplication increases transcriptional plasticity we examined the patterns of transcriptional plasticity of duplicates and singletons in the post-WGD yeast *C. glabrata*, a phylogenetically close species to *S. cerevisiae*. To this end, we asked the question of whether duplicates of *S. cerevisiae* had singleton orthologs in *C. glabrata* that were not more transcriptionally plastic than expected when compared to other *C. glabrata* singletons under stress and vice versa. We obtained RNA sequence data from a previous publication in which transcriptomic data were available under YPD conditions and under acidic stress (Linde *et al.* 2015), similar to our data on lactic acid stress. *S. cerevisiae* orthologs from *C. glabrata* were identified using syntheny information available in the pillars of YGOB (Byrne and Wolfe 2005). In total, we identified 4844 reliable *S. cerevisiae*: *C. glabrata* orthologs. Of these 4844 orthologs, 788 genes in *C. glabrata* were duplicated genes (394 pairs, File S6), of which 123 were duplicated in *C. glabrata* but not in *S. cerevisiae*. Of the 2240 duplicates of *S. cerevisiae*, we found 1019 orthologs that were singletons in *C. glabrata*. We first asked whether singletons in *C. glabrata* that are orthologs of duplicates in *S. cerevisiae* exhibit higher transcriptional plasticity than singletons in *C. glabrata* with no duplicate orthologs in *S. cerevisiae*. If this were the case, then gene duplication would have no role in transcriptional plasticity in *S. cerevisiae*. Notwithstanding that the transcriptional plasticity for a particular gene may vary among species, we found that the percentage of singletons with significant transcriptional alterations under stress in *C. glabrata* that are orthologs to *S. cerevisiae* duplicates (599 out of a total of 1019 genes, 58.7%) was not significantly higher than that of transcriptionally altered singletons in *C. glabrata* that had no duplicate orthologs in *S. cerevisiae* (1725 out of a total of 3062 singleton genes, 56.3%) (Fisher exact test: odd's ratio $F = 1.10$, $P = 0.17$). Conversely, duplicates in *C. glabrata* that were orthologs to singletons in *S. cerevisiae* exhibited a significantly higher percentage of transcriptionally altered genes under stress (82 out of a total of 123 genes, 66.7%) than singletons in *C. glabrata* (Fisher's exact test: odd's ratio $F = 1.55$, $P = 0.02$).

Differential patterns of transcriptional alterations within duplicated genes

We sought to investigate the different transcriptional profiles of pairs of duplicated genes and their contributions to the fitness of *S. cerevisiae*. We divided duplicated genes that underwent transcriptional alterations after stress into five different categories (Figure 2A and Table 2): (a) duplicates in which both of the gene copies were up-regulated under stress (called herein Up pattern); (b) duplicates with both copies down-regulated under stress (Down pattern); (c) duplicates with one copy up-regulated and one copy down-regulated under stress (Discordant pattern), (d) duplicates with one copy showing not-altered transcription under stress while its sister copy shows either up-regulation or down-regulation under stress (Only-one pattern), and (e) duplicates that remained unchanged under stress (Not-altered pattern).

In each of the stress conditions, the category "Only-one" comprised the largest number of duplicates with altered transcriptional profiles, with this category including 53–97% of the altered duplicates in the five stresses (Figure 2B). These results support the classical view of evolution by gene duplication, according to which following gene duplication one copy undergoes rapid divergence while the other copy keeps the ancestral function. Here we show that this pattern of evolution by gene duplication is also true for the regulatory evolution of duplicated genes.

Duplicate gene copies with higher transcriptional divergence under stress should show higher sequence divergence when compared to orthologous sequences from other phylogenetically related species if the basis for this transcriptional plasticity was encoded in the gene sequence. To test this hypothesis, we were able to obtain 537 reliable promoter alignments for duplicates in *S. cerevisiae* and at least four additional phylogenetically related yeast species (File S7). The main *Saccharomyces* species we compared *S. cerevisiae* to were *S. bayanus*, *S. castellii*, *S. mikatae*, *S. paradoxus*, *S. kluyveri*, and *S. kudriavzevii*. Not all species presented annotated intergenic regions but we used all those alignments that included at least four of the species. We aligned the 600 nucleotide sequence regions upstream of duplicated genes and their orthologs, as these are likely to include most if not all the regulatory elements of the genes (Ohler and Niemann 2001). We then measured the coefficient of conservation (CC) for each nucleotide site using the

Table 2 Categories of altered expression of duplicates

Stress	Number of Pairs Both Copies Concordant: Down	Number of Pairs Both Copies Concordant: Up	Number of Pairs Discordant	Number of Pairs Not-Altered	Number of Pairs One-Altered
Ethanol	89	76	74	438	677
Glycerol	159	103	102	413	777
Lactic acid	147	76	83	433	739
Oxidative	0	0	1	41	42
Oxidative + dextrose	129	87	96	448	760

entropy equation (Cover and Thomas 2006; Halabi *et al.* 2009; Ruiz-Gonzalez and Fares 2013):

$$CC = f_k^{(a)} \ln \frac{f_k^{(a)}}{q^{(a)}} + \left(1 - f_k^{(a)}\right) \ln \frac{1 - f_k^{(a)}}{1 - q^{(a)}}; \forall a \in [A, T, G, C]$$

In this equation, CC of a nucleotide (a) at position (k) in an alignment is defined as the entropy of the observed frequency of a at k ($f_k^{(a)}$) relative to the background frequency of a in all sequences of the alignment ($q^{(a)}$). Therefore, the more conserved the site the higher is its CC value. CC was averaged for each promoter and then these averages were used to compare altered duplicates (those belonging to the categories "Up," "Down," "Discordant," and "Only-one") with not-altered duplicates (those belonging to the category "Not-altered"). For each of the stress conditions we estimated the CC values for altered and not-altered duplicates. We then pulled all the data together from all stress conditions and compared the CC values of altered to that of not-altered duplicates. The CC values of duplicates with constant transcriptional profiles under stress (mean ± SE = 0.76 ± 0.005) were significantly larger than those of duplicates with altered transcriptional profiles (mean ± SE = 0.72 ± 0.01) (Figure 2C), and the difference was significant using a parametric test (t-test: $t = 3.47$, d.f. = 1140.9, $P = 5.4 \times 10^{-4}$) and a nonparametric test (Mann–Whitney U-test: $P = 0.003$), indicating that higher transcriptional plasticity of duplicates may be due to a divergence in their promoter sequences from the ancestral preduplication state.

Duplicates with different transcriptional divergence patterns exhibit different functional dependencies

To determine whether the transcriptional plasticity of duplicated genes is accompanied by a functional divergence of gene copies, we analyzed the genetic interaction network of *S. cerevisiae* and asked how many of the duplicated genes show genetic interactions between their gene copies, hence are functionally dependent on one another (Costanzo *et al.* 2010), within each of the transcriptional categories (*i.e.*, Up, Down, Discordant, Only one, and Not-altered). To this end, we used the genetic interaction map of *S. cerevisiae* as a proxy to the functions of each of the genes (Costanzo *et al.* 2010). This map contains roughly 6.5 million genetic interactions and the functional chart for 75% of the *S. cerevisiae* genes. The number of genetic interactions for a particular gene is a proxy to the number of functions it performs, as the deletion of both of the genes identified as interacting produces significantly different fitness effects than the multiplicative effect of single gene deletions (Costanzo *et al.* 2010). We identified 762,768 significant genetic interactions (*i.e.*, epistasis, ε) in *S. cerevisiae*, of which 52% were synergistic (*i.e.*, the double mutant exhibited significantly lower fitness W_{12} than the multiplicative effects of individual mutants: $\varepsilon = W_{12} - W_1W_2$; $\varepsilon < 0$) and 48% were antagonistic interactions ($\varepsilon = W_{12} - W_1W_2$; $\varepsilon > 0$). However, duplicated genes were largely biased regarding the sign epistasis, with the majority of the epistasis (89.5%) being synergistic (binomial test: $P < 2.2 \times 10^{-16}$). This pattern was also true for transcriptionally altered duplicates (89.74% synergistic epistasis). Dividing transcriptionally altered duplicates into the different categories provides similar results, with all such categories being equally enriched for duplicates with synergistic epistasis: up-regulated duplicates presented largely synergistic epistasis (varying between 86% in ethanol and 93% under oxidative stress supplemented with dextrose), and so did the only-one category (ranging between 87% of the interactions being synergistic under glycerol stress and 92.7% in ethanol stress). These percentages were of the same order in the "Down" and "Discordant" categories.

In all stress conditions the category "Down" showed the highest enrichment for those duplicates with interacting gene copies (Figure 3A). On average over all stress conditions, the genetically interacting duplicates enrichment followed the same pattern, with a distribution among the categories in the following decreasing manner: the category "Down," followed by the category "Not-altered," then the category "Only-one," then the category "Discordant," and finally the category "Up" (Figure 3A, inset box and Table 3).

The strong genetic interaction between the gene copies could be due to either each gene copy having a large fitness effect such that deleting both magnifies such an effect, or each gene copy having very low fitness effects due to genetic redundancy but deleting both significantly magnifies this effect (*i.e.*, functional compensation of a gene deletion or both gene copies are needed to perform the function because they have subfunctionalized). The category "Up" is the one with the lowest number of gene copy interactions, therefore is likely to contain very little genetic compensation, perhaps because gene copies have diverged in their function from the ancestral preduplication gene, and as such the multiplicative effect of deleting single gene copies may be as important in their contribution to fitness as the double gene deletions. The category "Discordant" shows higher levels of genetic interactions between gene copies than the category "Up," but lower levels than "Down." Since all discordant duplicates exhibit synergistic epistasis, this suggests certain functional redundancy under normal conditions for transcriptionally discordant duplicates, which also applies to the categories of "One-altered" (average percentage of synergistic epistasis among all stresses: 89.9%; binomial test: $P < 3.61 \times 10^{-7}$) and "Only-one" duplicates.

To determine whether duplicate gene copies are more dependent upon each other's functions than expected, we built sets of singleton genes for each of the duplicate sets according to their transcriptional profiles. Each of the singleton transcriptional categories was built taking random pairs of singleton genes. For example, for the "Up" category, both of the singleton genes were sampled from the set of up-regulated singleton genes under stress. We built sets of 1000 pairs and compared each of the duplicate transcriptional categories with the corresponding singleton transcriptional categories. Results show that all the categories,

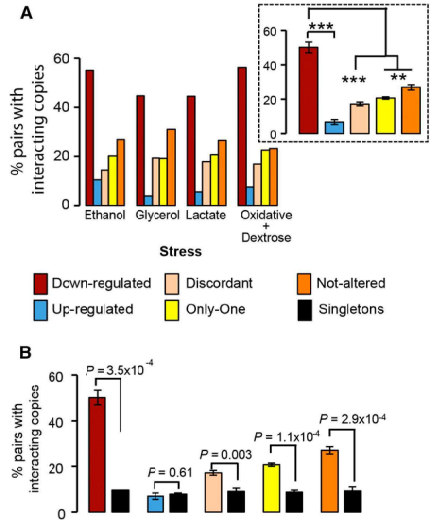


Figure 3 The genetic dependencies between gene copies of transcriptionally plastic duplicates. We measured the number of pairs within each of the transcriptional categories with evidence of genetic interactions between duplicate gene copies using the functional landscape of *S. cerevisiae* (Costanzo *et al.* 2010). (A) The percentage of pairs with interacting gene copies was very high in the category of down-regulated duplicates (red bars), very low in the category of up-regulated duplicates (blue bars), and intermediate in the other three categories under all the stress conditions examined in this study. The mean percentage of duplicates with interacting copies across the stress conditions for each transcriptional category is presented in the inset box. The category "Down" presented a larger mean percentage of duplicates whose gene copies are functionally dependent upon one another than any of the other categories. Significant differences are indicated with ** and *** when the probabilities are $P < 0.01$, $P < 0.001$, and $P < 10^{-4}$, respectively. (B) The average proportion of duplicates with interacting gene copies was compared to the proportion of genetic interactions for sets of randomly sampled pairs of singletons with altered transcription profiles under stress. Each transcriptional profile for duplicates was compared to an equivalent set of random pairs of singletons with similar transcriptional profiles. For example, up-regulated duplicates were compared to random pairs of up-regulated singletons under stress.

with the exception of the one including duplicates with both gene copies up-regulated, exhibit a significant proportion of their duplicates with interacting gene copies when compared to singletons of the same transcriptional category (Figure 3B). Therefore, up-regulated duplicates seem to exhibit evidence of independent evolution of their gene copies likely due to the finding of novel functions by each copy under stress.

Functional divergence and genetic redundancy of duplicated genes

The differences in the functional dependencies between gene copies found in the duplicate transcriptional categories hint at a different mode of evolution by gene duplication for these categories. We hypothesize, based on the patterns of genetic interactions, that the functional fate of duplicates in terms of neo- or subfunctionalization is dependent on the transcriptional category they belong to and the genetic redundancy between gene copies. Genetic redundancy has been shown to correlate with evolvability because it provides mutational robustness, which in turn increases the evolvability of genes (Draghi *et al.* 2010; Wagner 2000b, 2005).

To test whether a given transcriptional category of duplicates is more likely to have evolved neo- or subfunctionalization, we examined two parameters linked to genetic interactions: (a) the number of shared interactions between the gene copies, (b) the number of total interactions of the gene copies. Neo-functionalized duplicates involve those in which one of the gene copies has lost all ancestral functions and acquired new functions, hence is likely to have a reduced number of genetic interactions. Conversely, subfunctionalization should affect duplicates with many functions in which each copy has become specialized in a set of ancestral functions while sharing common functions with its sister gene copy, hence is likely to be overrepresented among highly interacting duplicates. Neo-functionalization should also lead to lower levels of sharing of genetic interactions between the gene copies as one of the copies has acquired novel functions that are perhaps independent from subfunctionalization. In agreement with our hypotheses, duplicates from the down-regulated category exhibited a greater number of genetic interactions (mean ± SE: 393.39 ± 14.17) than those of the up-regulated category (mean ± SE: 313.95 ± 13.95; *t*-test: $t = 3.99$, d.f. = 551.94, $P = 7.36 \times 10^{-5}$). The index of shared interactions was calculated as: $Sh_{A,B} = \frac{1}{2}\left(\frac{Sh}{N_A} + \frac{Sh}{N_B}\right)$, with $Sh_{A,B}$ referring to the mean number of shared interactions between gene copies A and B, *Sh* referring to the number of shared interactions, and *N* being the total number of interactions. Duplicates from the down-regulated category shared more interactions (mean ± SE: 0.15 ± 0.005) than those of the up-regulated category (mean ± SE: 0.12 ± 0.004; *t*-test: $t = 3.29$, d.f. = 209.27, $P = 1.1 \times 10^{-3}$). These results indicate that while up-regulated duplicates may have neo-functionalized, down-regulated duplicates have likely subfunctionalized.

Sequence divergence levels of duplicates correlate with their transcriptional profiles

To determine whether the transcriptional duplicate categories included specific functional divergence profiles between gene copies, we inferred the amino acid distances between duplicate gene copies for all transcriptional categories and stress conditions. Divergence between duplicate gene copies was calculated using Poisson-corrected distances. Under all four stresses, the duplicates of category "Down" presented the lowest distance between gene copies (Figure 4A), followed by the category "Discordant," then the category "Only-one," then the category "Up," and finally the category "Not-altered" (Figure 4A). Taking all stresses together, we found three groups of transcriptional categories according to the divergence values between gene copies of duplicates (Figure 4B). The first category is "Down": this category exhibited the lowest divergence levels between gene copies which was significantly smaller than the following group that included "Discordant" category duplicates (median divergence values for "Down": 0.05, median for "Discordant": 0.12, Wilcoxon rank test: $P < 2.2 \times 10^{-16}$). The following group included duplicates belonging to the transcriptional categories of "Only-one" (median divergence between gene copies: 0.22), "Up" (median divergence between gene copies: 0.24), and "Not-altered" (median divergence of gene copies: 0.33). The category "Discordant" exhibited significantly lower divergence between gene copies than the categories "Up," "Only-one," and "Not-altered" (Wilcoxon rank test: $P < 2.2 \times 10^{-16}$).

Importantly, the mean sequence divergence levels between the duplicate gene copies correlated negatively with the mean percentage of duplicates in which both gene copies interacted genetically (Pearson correlation: $\rho = -0.54$, $P = 0.013$; Figure 4C), indicating that the larger the divergence between the gene copies the lower is their functional dependency. This correlation became more significant when taking only those duplicates for which at least one gene copy has showed changing transcriptional patterns under stress (Pearson correlation: $\rho = -0.78$, $P = 3.4 \times 10^{-4}$). This result is in agreement with a previous study in which the differences between pairs of WGDs in which both gene copies interacted genetically and those in which gene copies did not was analysed (Musso *et al.* 2008).

These results strongly suggest that gene copies that become up-regulated (category "Up") under stress have undergone accelerated evolution and divergence from their ancestral, preduplication functions perhaps allowing the adaptation to stress conditions that are often encountered by the cell in nature.

Table 3 Number of duplicates with genetically interacting gene copies for each transcriptional category of duplicates

Stress	Number of Pairs (Total) Down-Regulated	Number of Pairs (Total) Up-Regulated	Number of Pairs (Total) Discordant	Number of Pairs (Total) Only-One	Number of Pairs (Total) Not-Altered
Ethanol	27 (49)	6 (57)	8 (55)	61 (301)	67 (249)
Glycerol	38 (85)	3 (77)	15 (77)	54 (281)	59 (190)
Lactate	36 (81)	3 (54)	10 (56)	65 (313)	55 (207)
Oxidative + dextrose	36 (64)	5 (67)	12 (71)	72 (319)	44 (189)

The origin of specific and general adaptations in S. cerevisiae

To determine whether the transcriptional plasticity of duplicates is the result of an adaptive process to face environmental perturbations, we sought to investigate whether this plasticity is stress specific (*i.e.*, the result of adaptive processes) or a general response to stress. We examined common transcriptionally altered duplicates for each of the transcriptional categories among stress conditions. We found that a substantial proportion of duplicates showed stress-specific transcriptional plasticity (Figure 5A). This pattern was the inverse in the case of transcriptionally altered singletons, with many common such singletons responding to all four stress conditions (Figure 5B). Comparison of the proportion of duplicates in each of the categories for stress response (*i.e.*, stress specific, common genes response to two, three, or four stresses) revealed more significant stress-specific transcriptional alterations in duplicates than in singletons (a mean of 34.4% of duplicates with transcriptional flexibility were stress specific compared to 26% of singletons, Fisher's exact test: odds ratio $F = 1.47$, $P = 3.3 \times 10^{-4}$), while singletons showed more common responses to all stresses than duplicates (a mean of 32.8% of singletons responded to all stress conditions compared to 23% of duplicates, Fisher's exact test: odds ratio $F = 1.64$, $P = 1.1 \times 10^{-5}$) (Figure 5C). These results reveal a fundamental difference in the transcriptional plasticity of duplicates and singletons, with evidence for the role of natural selection in duplicates' transcriptional differences as an adaptive mechanism.

Because of the fundamental difference in transcriptional plasticities between WGDs and SSDs (Figure 1B), we split the dataset for duplicates into these two groups and conducted the same comparison as above. Both the WGDs and SSDs showed very similar transcriptional flexibility patterns to the entire dataset: WGDs and SSDs had their largest transcriptional plasticity in genes that responded in a stress-specific manner (Figure 5D).

To understand the relationship between adaptation to stress and transcriptional plasticity, we analyzed how the different transcriptional alterations in duplicates may have an important role in the adaptation to oxidative stress supplemented with dextrose (Figure 6). Oxidation generates reactive oxidative molecules or species (ROS) including peroxide, superoxide, hydroxyl radicals, and single oxygen in the cell. Increasing ROS in the cell can lead to important structural damage. We found a number of important duplicates that are involved in mitochondrial respiration (*Icl1*/*Icl2*, *Shh4*/*Shh*1, and *Sdh4*/*Sdh1* duplicated genes, among others) as well as duplicated genes encoding transcriptional gluconeogenesis activators (*Cat8*/*Sip4* and *Csr2*/*Ecm21*) to be up-regulated under oxidative stress, perhaps to reduce the generation of ROS. Moreover, duplicated genes involved in NADH metabolism and oxidative processes in the glycolysis pathway (*Gpd1*/*Gpd2*, *Gpp1*/*Gpp2*) are down-regulated under stress (Figure 6). Interestingly, as previously noticed (Belli *et al.* 2004), the gene copies (*Pug1*/*Rta1*) involved in heme transport and iron ion homeostasis showed discordant expression patterns, while the gene copies (*Fit3*/*Fit1*) involved in ion transport showed transcriptional alterations only for *Fit3* (Figure 6).

The transcriptional plasticity of the duplicates belonging to the category "Only-one" was very noticeable under oxidative stress affecting functional classes required to minimize ROS, including many duplicated genes involved in ion transport (*Fit1*/*Fit3*), hexose transport (*Hxt3*/*Hxt5*), and heat stress response (*Ssa2*/*Ssa1*). Similarly, the transcriptional category "Discordant" also showed a prominent response pattern to oxidative stress, including duplicates involved in nucleotide/nucleoside metabolism (*Gnd1*/*Gnd2*, *Apa1*/*Apa2*), and in heme transport and iron ion homeostasis (*Pug1*/*Rta1*). Most of these duplicates have important roles in DNA replication and stress.

Interestingly, some of these genes are involved in many stresses but their transcriptional plasticity exhibits different patterns under different stresses. For example, the duplicates (*Gpd1*/*Gpd2* and *Gpp1*/*Gpp2*) that drive glycerol production using dextrose through the glycolysis pathway (NADH metabolism and oxidative processes) behave transcriptionally different under the different stresses. Both of the gene copies of these duplicates are down-regulated under oxidative stress supplemented with dextrose, while only one gene copy is down-regulated when the cell is subjected to stress by lactate or glycerol, and none of the gene copies showed any differential expression level with the wild type when the cell grew under ethanol stress. Similarly, the gene copies of the duplicates (*Shh4*/*Shh1* and *Sdh4*/*Sdh1*) involved in oxidation of succinate and electron transfer to ubiquinone are up-regulated under oxidative stress. However, under ethanol stress, only gene copies *Shh4* and *Sdh4* are up-regulated while their corresponding paralogs *Shh1* and *Sdh1* show wild-type expression levels (Figure 6).

DISCUSSION

In this study, we demonstrate that ancient duplicates of *S. cerevisiae* exhibit a large transcriptional plasticity when subjected to stress. This

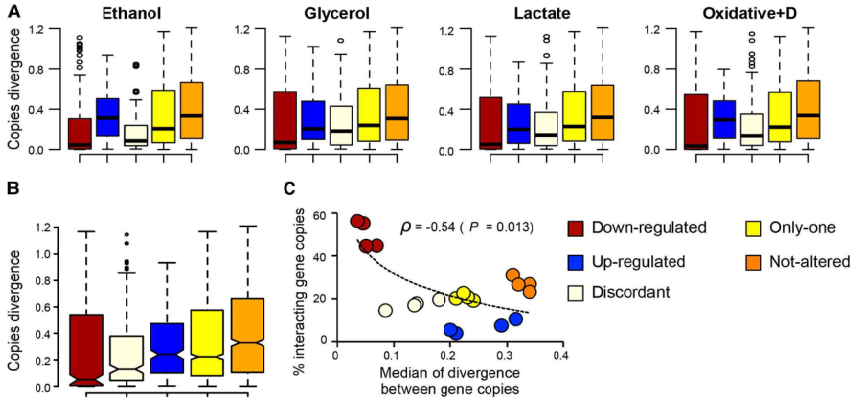


Figure 4 Functional divergence analysis of duplicates with different patterns of transcriptional plasticity. (A) Poisson-corrected amino acid distance between gene copies of duplicates for each of the transcriptional plasticity profiles ("Up," "Down," "Discordant," "Only one," and "Not-altered"). (B) Comparison of the divergence levels between gene copies for the different transcriptional profiles. (C) Correlation analysis between the percentage of of duplicates with interacting gene copies and the divergence levels between the copies (red, blue, light yellow, yellow, and orange circles refer to the duplicate class of "Down," "Up," "Discordant," "One-altered," and "Not-altered," respectively).

transcriptional plasticity may be the result of preadaptations to environmental stress. Such preadaptations may have been generated through an increase in the polymorphism of the regulatory sequence regions of duplicated genes, perhaps the same sequence changes that have led to the regulatory divergence between the gene copies of duplicates. The transcriptional divergence between gene copies has been studied in plants (Blanc and Wolfe 2004; Ha *et al.* 2007, 2009; Wang *et al.* 2012) and animals (Huminiecki and Wolfe 2004). In *S. cerevisiae*, while the genome-wide transcriptional plasticity has been reported under stress (Causton *et al.* 2001; Cormier *et al.* 2010; Ferea *et al.* 1999; Ideker *et al.* 2001; Landry *et al.* 2006; Stern *et al.* 2007), the differential patterns in this plasticity between duplicates and singletons have received little attention. Findings from these studies led authors to conclude that the transcriptional plasticity of the genes in *S. cerevisiae* is the result of a general response to a wide range of stresses, sparking the possibility that this plasticity is an emerging property resulting from a universal feature of the underlying regulatory network. In this study we show that: (i) ancient duplicates of *S. cerevisiae* exhibit a large transcriptional plasticity when subjected to stress; and (ii) the transcriptional plasticity of duplicated genes differs from that of singletons, is more complex than thought before, and is likely the result of selection for an adaptive response to specific environmental challenges.

The transcriptional changes affecting one or both of the gene copies resulting from gene duplication may be selectively advantageous in unicellular organisms because the absence of tissue-specific transcriptional subfunctionalization precludes a relief of the genetic redundancy of duplicated genes. Therefore, in unicellular eukaryotes, such as *S. cerevisiae*, the efficiency of purifying selection or positive selection must be a strong force driving the fate of duplicated genes. In agreement with this prediction, the genetic redundancy generated in *S. cerevisiae* after the WGD event that took place roughly 100 MYA was erased by purifying selection, as 92% of duplicated genes returned to single copy genes (Wolfe and Shields 1997). Despite this, the number of duplicates in *S. cerevisiae* (roughly 30% of all the genes) is higher than predicted by theory, raising the possibility that most retained genes have become functionally specialized, hence less redundant, shortly after duplication (Force *et al.* 1999; Lynch and Katju 2004; Barkman and Zhang 2009; Des Marais and Rausher 2008; He and Zhang 2005; Conant and Wolfe 2006), transcriptionally divergent (Blanc and Wolfe 2004; Ha *et al.* 2007, 2009; Wang *et al.* 2012; Francino 2005), preserved due to their higher mutational robustness (Fares 2015; Fares *et al.* 2013; Keane *et al.* 2014; Wagner 2000b, 2005), maintained owing to a selective advantage for higher gene dosage (Conant and Wolfe 2008), or kept to preserve stoichiometric balances in duplicates encoding protein complexes (Gibson and Spring 1998; Veitia 2003a,b).

In this study, the transcriptional plasticity identified in *S. cerevisiae* is likely the result of population polymorphism at the regulatory regions of duplicates (Figure 3C), which were selectively relaxed after gene duplication. This polymorphism has likely given rise to preadaptations to environments never before faced by the yeast and became fixed in the populations after facing such environmental perturbations. It is therefore likely that duplicates that show transcriptional plasticity, in particular those that become up-regulated under stress, are usually performing important functions in the cell and are hence maintained by purifying selection. However, under stress, such genes may encode new functions that provide the yeast with the ability to survive stress, a property encapsulated within the term exaptation (Gould and Vrba 1982). The question that remains is: how important an adaptive force is the transcriptional divergence against the functional divergence of duplicates in *S. cerevisiae*?

Functional divergence after gene and genome duplication has been the subject of intense scrutiny and a number of examples unequivocally correlate the origin of important gene families and functional specialization with the divergence between gene paralogs. Indeed, key globin proteins that specialized in different aspects of oxygen metabolism have originated through WGD events (Hoffmann *et al.* 2011, 2012a,b; Storz *et al.* 2011, 2013). Functional divergence has also been observed in a number of studies and has been correlated with an asymmetric increase

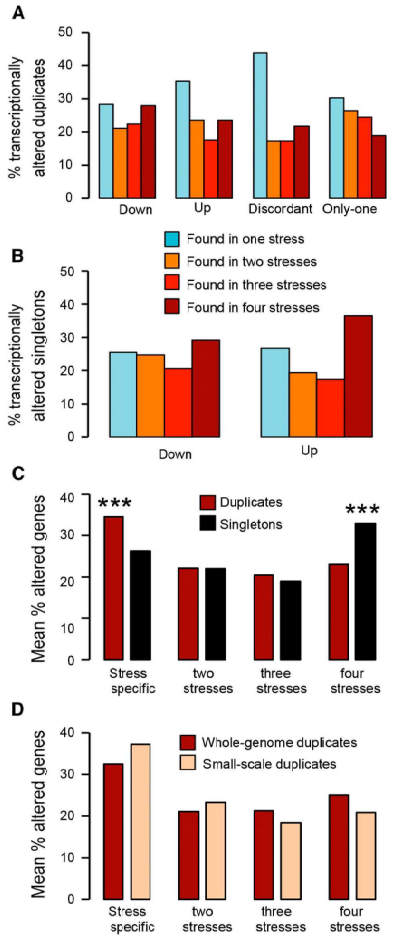


Figure 5 Transcriptional plasticity of duplicates is stress-specific. We analyzed the distribution of transcriptionally altered duplicates in the different stress conditions. (A) Proportion of the transcriptionally altered duplicates from each of the duplicates classes ("Up," "Down," "Discordant," and "Only one") that are altered in one stress only, two stresses, three stresses, or in all four stresses tested in this study. (B) Proportion of singletons that are up- or down-regulated that respond to specifically one stress only, two stresses, three stresses, or all four stresses. (C) The mean percentage of altered genes across the different transcriptional classes in duplicates and singletons that are altered under one or more type of stress. (D) The mean percentage of whole-genome duplicates and small-scale duplicates that are transcriptionally altered when *S. cerevisiae* is faced with one or more stresses. *** indicates $P < 0.001$ under a Fisher's exact test.

in the rates of sequence evolution in the duplicate gene copies (Blanc and Wolfe 2004), in good agreement with the fundamental tenet of the molecular evolution theory (Gu *et al.* 2002; Dermitzakis and Clark 2001). Expression divergence, but not transcriptional plasticity, between the copies of duplicated genes has also been demonstrated in a number of organisms.

We propose the hypothesis supporting a link between expression and functional divergence – that is, one level of divergence necessarily drives the other level. Indeed, gene expression levels largely determine the rates of evolution of the proteins they encode (Drummond *et al.* 2006; Drummond and Wilke 2008; Pal *et al.* 2001; Rocha and Danchin 2004; Wilke and Drummond 2006). The theoretical justification for this link between gene expression and its rate of evolution can be found in the misfolding–mistranslation hypothesis, according to which highly expressed genes evolve slower constrained by the need to maintain low levels of misfolded or mistranslated proteins bearing destabilizing mutations (Drummond *et al.* 2005). On the other hand, functional divergence, or acquisition of novel functions, may involve a fine-tuning of the expression of the encoding gene to perform the required function at the right rate. Whether expression divergence came before functional divergence or vice versa remains to be investigated but our data suggest a link between these two levels of divergence because different transcriptional categories exhibit different patterns of sequence evolution and divergence between gene copies (Figure 4C). We hypothesize that genetic redundancy has allowed transcriptional divergence between gene copies due to relaxed selective constraints. This has allowed divergence at the coding level driven by changes in gene expression, as gene expression is a strong determinant of sequence evolution (Drummond *et al.* 2005). Such functional divergence may have led to the acquisition of functions that enabled the adaptation to stress conditions (Figure 5C).

In this study, we determine the plasticity that each of the gene copies has at the regulatory level, the link of this plasticity with the functional dependencies among gene copies, and the role of such a link in the response to stress. Our study reveals different modes of evolution for the different transcriptional categories. Most responsive duplicates to stress present only one copy altered, following the classic view of evolution by gene duplication. The genetic dependencies and low sequence divergence between gene copies for these duplicates also reveal the mode of evolution and innovation: these duplicates exhibit the highest proportion of cases with synergistic epistasis between gene copies, which summed to the low sequence divergence between the gene copies indicates higher genetic redundancy (VanderSluis *et al.* 2010). This nontrivial pattern of evolution of novel functions is in agreement with previous predictions, according to which higher genetic redundancy allows the functional compensation between gene copies, the neutral exploration of genotypic space, and eventual finding of additional novel functions (Fares 2015; Fares *et al.* 2013; Keane *et al.* 2014; Wagner 2005). The category of up-regulated duplicates exhibits evidence of neo-functionalization based on the rapid evolution of the gene copies when compared to their ancestor and the low functional dependency of each copy on its sister copy, suggesting the acquisition of novel functions. Duplicates with both copies being down-regulated under stress present low divergence between the gene copies and significant functional dependencies among the gene copies, suggesting the subfunctionalization of the gene copies through the partition of ancestral functions. The partition of ancestral functions in these duplicates is not complete, as the gene copies share more functions than expected. This greater sharing may reflect a selective advantage for gene dosage, particularly in the ancestral state immediately postdating genome duplication (Ihmels *et al.* 2007; VanderSluis *et al.* 2010). Finally, the category in which gene copies exhibit discordant transcriptional

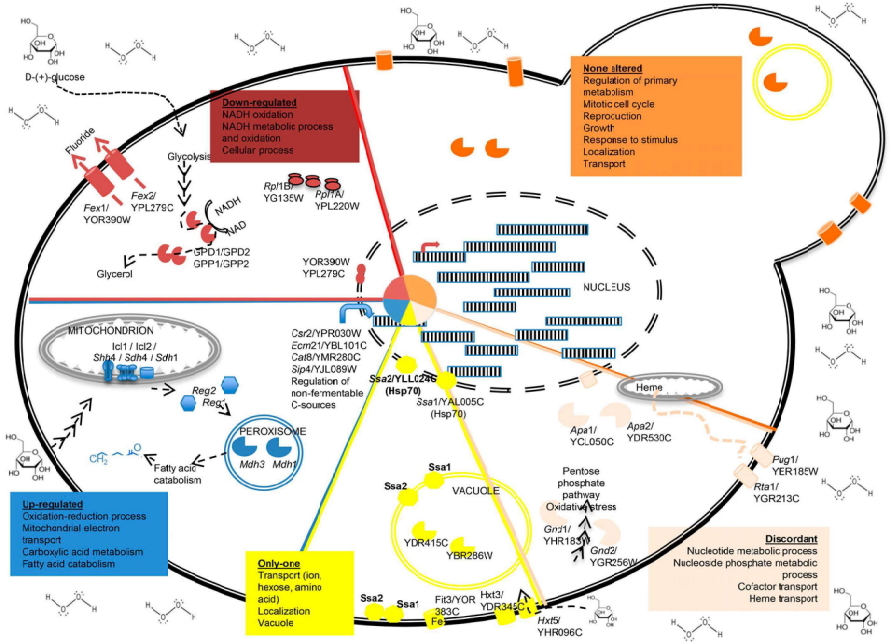


Figure 6 Schematic representation of main GO processes and selected duplicate gene pairs affected by oxidative stress in the presence of dextrose. The schematic *S. cerevisiae* cell has been partitioned in a pie shape with the proportions being in accordance to the distribution of duplicates in categories ("Up," "Down," "Discordant," "One-altered," and "Not-altered"). Color codes follow the same categorization. For those categories, "One-altered" and "Discordant," on which each gene in the pair show a differential fold-change, it has been shadowed with green if not differentially expressed, with blue if down-regulated, or with red if up-regulated. Some of the pathways affected are indicated with the corresponding genes names and ID tag, except for the most common gene names, or when not available a common name and the ID tag is provided. Curved arrows in the nucleus represent some transcription factors affected; color indicates the corresponding category. Subcellular localization of proteins has been retrieved from Saccharomyces Genome Database following Gene Ontology term analysis.

alterations under stress ("Discordant") presents a low number of genetic interactions between gene copies and low sequence divergence of the gene copies when compared to a preduplication ancestral gene, suggesting that both of the gene copies may be performing very similar functions under different conditions (Force *et al.* 1999). The category of discordant duplicates may include cases in which copies have diverged in their regulation such that one copy is active under stress and its sister copy is active under normal conditions, thereby avoiding the costly evolutionary optimization of the encoded function under two different conditions (Conant and Wolfe 2006). Remarkably, duplicates with no evidence for expression alteration under stress conditions exhibit a greater number of interactions between gene copies than expected. Because these duplicates are not affected by environmental perturbations, their enrichment for genetic interactions supports genetic buffering between the gene copies that is independent of the environment, a phenomenon important for genetic robustness (Dean *et al.* 2008; DeLuna *et al.* 2008; Fares 2015; Keane *et al.* 2014; VanderSluis *et al.* 2010).

Our results point to the expression divergence between the gene copies of duplicated genes as a strategy to yield adaptive responses to stress conditions. This divergence is the result of differential accumulation of mutations (polymorphism) in the promoters of the duplicate gene copies, such that one copy exhibits high expression polymorphism under normal environmental conditions (Keane *et al.* 2014). Since yeast generally undergo stress because of the changes in the osmotic characteristics of the medium resulting from the metabolic byproducts generated, such stress conditions provide selective advantage to those strains with polymorphic expression divergence of duplicated genes that allow response to stress. A number of findings disregard expression noise as an alternative hypothesis to the adaptive value of the observed transcriptional alterations in duplicates. First, the profiles indicating adaptations to stress (*i.e.*, the transcriptional categories including duplicates with at least one gene up-regulated) include duplicated genes that under normal conditions show as much contribution to fitness as those with no response to stress (Figure 1, C–F). Second, under stress conditions, the contribution of up-regulated duplicates to fitness is significantly higher than their contribution to fitness under normal conditions. Finally, duplicates with altered expression profiles are mostly stress-specific and affect functions that are strictly related to the interaction with the environment and signal transduction.

In conclusion, we reveal the underlying transcriptional plasticity of duplicates in *S. cerevisiae* and the potential of this plasticity to give origin to adaptations (by means of specialization of duplicates) to environmental perturbations. This study sets a new research endeavor mainly aiming at finding the as yet unexplored metabolic capabilities resulting from the evolution of duplicated genes.

ACKNOWLEDGMENTS

We thank members of the Fares laboratory for careful reading and discussion of the results in the manuscript. We are also grateful to colleagues at Trinity College for helpful discussions. This work was supported by a grant from the Spanish Ministerio de Economía y Competitividad (reference: BFU2015-66073-P) and a grant (reference: ACOMP/2015/026) from the local government Consellería de Educación Investigación, Cultura y Deporte, Generalitat Valenciana to M.A.F. C.T. was supported by a grant Juan de la Cierva from the Spanish Ministerio de Economía y Competitividad (reference: JCA-2012-14056).

LITERATURE CITED

Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang *et al.*, 1997 Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25(17): 3389–3402.

Anders, S., and W. Huber, 2010 Differential expression analysis for sequence count data. Genome Biol. 11(10): R106.

Barkman, T., and J. Zhang, 2009 Evidence for escape from adaptive conflict? Nature 462 (7274): E1; discussion E2–3.

Belli, G., M. M. Molina, J. Garcia-Martinez, J. E. Perez-Ortin, and E. Herrero, 2004 *Saccharomyces cerevisiae* glutaredoxin 5-deficient cells subjected to continuous oxidizing conditions are affected in the expression of specific sets of genes. J. Biol. Chem. 279(13): 12386–12395.

Blanc, G., and K. H. Wolfe, 2004 Functional divergence of duplicated genes formed by polyploidy during *Arabidopsis* evolution. Plant Cell 16(7): 1679–1691.

Byrne, K. P., and K. H. Wolfe, 2005 The yeast gene order browser: combining curated homology and syntenic context reveals gene fate in polyploid species. Genome Res. 15(10): 1456–1461.

Carretero-Paulet, L., and M. A. Fares, 2012 Evolutionary dynamics and functional specialization of plant paralogs formed by whole and small-scale genome duplications. Mol. Biol. Evol. 29(11): 3541–3551.

Causton, H. C., B. Ren, S. S. Koh, C. T. Harbison, E. Kanin *et al.*, 2001 Remodeling of yeast genome expression in response to environmental changes. Mol. Biol. Cell 12(2): 323–337.

Conant, G. C., and K. H. Wolfe, 2006 Functional partitioning of yeast co-expression networks after genome duplication. PLoS Biol. 4(4): e109.

Conant, G. C., and K. H. Wolfe, 2008 Turning a hobby into a job: how duplicated genes find new functions. Nat. Rev. Genet. 9(12): 938–950.

Cormier, L., R. Barbey, and L. Kuras, 2010 Transcriptional plasticity through differential assembly of a multiprotein activation complex. Nucleic Acids Res. 38(15): 4998–5014.

Costanzo, M., A. Baryshnikova, J. Bellay, Y. Kim, E. D. Spear *et al.*, 2010 The genetic landscape of a cell. Science 327(5964): 425–431.

Cover, T. M., and J. A. Thomas, 2006 *Elements of Information Theory*. Ed. 2. Wiley-Interscience, New York.

Cui, L., P. K. Wall, J. H. Leebens-Mack, B. G. Lindsay, D. E. Soltis *et al.*, 2006 Widespread genome duplications throughout the history of flowering plants. Genome Res. 16(6): 738–749.

Dean, E. J., J. C. Davis, R. W. Davis, and D. A. Petrov, 2008 Pervasive and persistent redundancy among duplicated genes in yeast. PLoS Genet. 4(7): e1000113.

DeLuna, A., K. Vetsigian, N. Shoresh, M. Hegreness, M. Colon-Gonzalez *et al.*, 2008 Exposing the fitness contribution of duplicated genes. Nat. Genet. 40(5): 676–681.

Dermitzakis, E. T., and A. G. Clark, 2001 Differential selection after duplication in mammalian developmental genes. Mol. Biol. Evol. 18(4): 557–562.

Des Marais, D. L., and M. D. Rausher, 2008 Escape from adaptive conflict after duplication in an anthocyanin pathway gene. Nature 454(7205): 762–765.

Draghi, J. A., T. L. Parsons, G. P. Wagner, and J. B. Plotkin, 2010 Mutational robustness can facilitate adaptation. Nature 463(7279): 353–355.

Drummond, D. A., and C. O. Wilke, 2008 Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell 134(2): 341–352.

Drummond, D. A., J. D. Bloom, C. Adami, C. O. Wilke, and F. H. Arnold, 2005 Why highly expressed proteins evolve slowly. Proc. Natl. Acad. Sci. USA 102(40): 14338–14343.

Drummond, D. A., A. Raval, and C. O. Wilke, 2006 A single determinant dominates the rate of yeast protein evolution. Mol. Biol. Evol. 23(2): 327–337.

Fares, M. A., 2015 The origins of mutational robustness. Trends Genet. 31(7): 373–381.

Fares, M. A., O. M. Keane, C. Toft, L. Carretero-Paulet, and G. W. Jones, 2013 The roles of whole-genome and small-scale duplications in the functional specialization of *Saccharomyces cerevisiae* genes. PLoS Genet. 9(1): e1003176.

Ferea, T. L., D. Botstein, P. O. Brown, and R. F. Rosenzweig, 1999 Systematic changes in gene expression patterns following adaptive evolution in yeast. Proc. Natl. Acad. Sci. USA 96(17): 9721–9726.

Ferris, S. D., and G. S. Whitt, 1979 Evolution of the differential regulation of duplicate genes after polyploidization. J. Mol. Evol. 12(4): 267–317.

Force, A., M. Lynch, F. B. Pickett, A. Amores, Y. L. Yan *et al.*, 1999 Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151(4): 1531–1545.

Francino, M. P., 2005 An adaptive radiation model for the origin of new gene functions. Nat. Genet. 37(6): 573–577.

Gibson, T. J., and J. Spring, 1998 Genetic redundancy in vertebrates: polyploidy and persistence of genes encoding multidomain proteins. Trends Genet. 14(2): 46–49, discussion 49–50.

Goldman, N., and Z. Yang, 1994 A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol. Biol. Evol. 11(5): 725–736.

Gould, S. J., and E. S. Vrba, 1982 Exaptation-A missing term in the science of form. Paleobiology 8(1): 12.

Gu, Z., D. Nicolae, H. H. Lu, and W. H. Li, 2002 Rapid divergence in expression between duplicate genes inferred from microarray data. Trends Genet. 18(12): 609–613.

Ha, M., W. H. Li, and Z. J. Chen, 2007 External factors accelerate expression divergence between duplicate genes. Trends Genet. 23(4): 162–166.

Ha, M., E. D. Kim, and Z. J. Chen, 2009 Duplicate genes increase expression diversity in closely related species and allopolyploids. Proc. Natl. Acad. Sci. USA 106(7): 2295–2300.

Halabi, N., O. Rivoire, S. Leibler, and R. Ranganathan, 2009 Protein sectors: evolutionary units of three-dimensional structure. Cell 138(4): 774–786.

He, X., and J. Zhang, 2005 Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. Genetics 169(2): 1157–1164.

Hoegg, S., H. Brinkmann, J. S. Taylor, and A. Meyer, 2004 Phylogenetic timing of the fish-specific genome duplication correlates with the diversification of teleost fish. J. Mol. Evol. 59(2): 190–203.

Hoffmann, F. G., J. C. Opazo, and J. F. Storz, 2011 Differential loss and retention of cytoglobin, myoglobin, and globin-E during the radiation of vertebrates. Genome Biol. Evol. 3: 588–600.

Hoffmann, F. G., J. C. Opazo, D. Hoogewijs, T. Hankeln, B. Ebner *et al.*, 2012a Evolution of the globin gene family in deuterostomes: lineage-specific patterns of diversification and attrition. Mol. Biol. Evol. 29(7): 1735–1745.

Hoffmann, F. G., J. C. Opazo, and J. F. Storz, 2012b Whole-genome duplications spurred the functional diversification of the globin gene superfamily in vertebrates. Mol. Biol. Evol. 29(1): 303–312.
Holub, E. B., 2001 The arms race is ancient history in *Arabidopsis*, the wildflower. Nat. Rev. Genet. 2(7): 516–527.
Huminiecki, L., and K. H. Wolfe, 2004 Divergence of spatial gene expression profiles following species-specific gene duplications in human and mouse. Genome Res. 14(10A): 1870–1879.
Ideker, T., V. Thorsson, J. A. Ranish, R. Christmas, J. Buhler *et al.*, 2001 Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. Science 292(5518): 929–934.
Ihmels, J., S. R. Collins, M. Schuldiner, N. J. Krogan, and J. S. Weissman, 2007 Backup without redundancy: genetic interactions reveal the cost of duplicate gene loss. Mol. Syst. Biol. 3: 86.
Innan, H., and F. Kondrashov, 2010 The evolution of gene duplications: classifying and distinguishing between models. Nat. Rev. Genet. 11(2): 97–108.
Keane, O. M., C. Toft, L. Carretero-Paulet, G. W. Jones, and M. A. Fares, 2014 Preservation of genetic and regulatory robustness in ancient gene duplicates of *Saccharomyces cerevisiae*. Genome Res. 24(11): 1830–1841.
Kim, S., M. J. Yoo, V. A. Albert, J. S. Farris, P. S. Soltis *et al.*, 2004 Phylogeny and diversification of B-function MADS-box genes in angiosperms: evolutionary and functional implications of a 260-million-year-old duplication. Am. J. Bot. 91(12): 2102–2118.
Landry, C. R., J. Oh, D. L. Hartl, and D. Cavalieri, 2006 Genome-wide scan reveals that genetic variation for transcriptional plasticity in yeast is biased towards multi-copy and dispensable genes. Gene 366(2): 343–351.
Lespinet, O., Y. I. Wolf, E. V. Koonin, and L. Aravind, 2002 The role of lineage-specific gene family expansion in the evolution of eukaryotes. Genome Res. 12(7): 1048–1059.
Li, W. H., J. Yang, and X. Gu, 2005 Expression divergence between duplicate genes. Trends Genet. 21(11): 602–607.
Linde, J., S. Duggan, M. Weber, F. Horn, P. Sieber *et al.*, 2015 Defining the transcriptomic landscape of *Candida glabrata* by RNA-Seq. Nucleic Acids Res. 43(3): 1392–1406.
Lohse, M., A. M. Bolger, A. Nagel, A. R. Fernie, J. E. Lunn *et al.*, 2012 RobiNA: a user-friendly, integrated software solution for RNA-Seq-based transcriptomics. Nucleic Acids Res. 40(Web Server issue): W622–W627.
Lynch, M., and J. S. Conery, 2000 The evolutionary fate and consequences of duplicate genes. Science 290(5494): 1151–1155.
Lynch, M., and J. S. Conery, 2003 The evolutionary demography of duplicate genes. J. Struct. Funct. Genomics 3(1–4): 35–44.
Lynch, M., and V. Katju, 2004 The altered evolutionary trajectories of gene duplicates. Trends Genet. 20(11): 544–549.
Marcet-Houben, M., and T. Gabaldon, 2015 Beyond the whole-genome duplication: phylogenetic evidence for an ancient interspecies hybridization in the Baker's yeast lineage. PLoS Biol. 13(8): e1002220.
Musso, G., M. Costanzo, M. Huangfu, A. M. Smith, J. Paw *et al.*, 2008 The extensive and condition-dependent nature of epistasis among whole-genome duplicates in yeast. Genome Res. 18(7): 1092–1099.
Ohler, U., and H. Niemann, 2001 Identification and analysis of eukaryotic promoters: recent computational approaches. Trends Genet. 17(2): 56–60.
Ohno, S., 1970 *Evolution by Gene Duplication*. Springer Verlag, New York.
Ohno, S., 1999 Gene duplication and the uniqueness of vertebrate genomes circa 1970–1999. Semin. Cell Dev. Biol. 10(5): 517–522.
Otto, S. P., and J. Whitton, 2000 Polyploid incidence and evolution. Annu. Rev. Genet. 34: 401–437.
Pal, C., B. Papp, and L. D. Hurst, 2001 Highly expressed genes in yeast evolve slowly. Genetics 158(2): 927–931.
Robinson, M. D., D. J. McCarthy, and G. K. Smyth, 2010 edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26(1): 139–140.
Rocha, E. P., and A. Danchin, 2004 An analysis of determinants of amino acids substitution rates in bacterial proteins. Mol. Biol. Evol. 21(1): 108–116.
Ruiz-Gonzalez, M. X., and M. A. Fares, 2013 Coevolution analyses illuminate the dependencies between amino acid sites in the chaperonin system GroES-L. BMC Evol. Biol. 13: 156.
Steinmetz, L. M., C. Scharfe, A. M. Deutschbauer, D. Mokranjac, Z. S. Herman *et al.*, 2002 Systematic screen for human disease genes in yeast. Nat. Genet. 31(4): 400–404.
Stern, S., T. Dror, E. Stolovicki, N. Brenner, and E. Braun, 2007 Genome-wide transcriptional plasticity underlies cellular adaptation to novel challenge. Mol. Syst. Biol. 3: 106.
Storz, J. F., J. C. Opazo, and F. G. Hoffmann, 2011 Phylogenetic diversification of the globin gene superfamily in chordates. IUBMB Life 63(5): 313–322.
Storz, J. F., J. C. Opazo, and F. G. Hoffmann, 2013 Gene duplication, genome duplication, and the functional diversification of vertebrate globins. Mol. Phylogenet. Evol. 66(2): 469–478.
Taylor, J. S., and J. Raes, 2004 Duplication and divergence: the evolution of new genes and old ideas. Annu. Rev. Genet. 38: 615–643.
Thompson, D. A., S. Roy, M. Chan, M. P. Styczynsky, J. Pfiffner *et al.*, 2013 Evolutionary principles of modular gene regulation in yeasts. eLife 2: e00603.
Tong, A. H., M. Evangelista, A. B. Parsons, H. Xu, G. D. Bader *et al.*, 2001 Systematic genetic analysis with ordered arrays of yeast deletion mutants. Science 294(5550): 2364–2368.
VanderSluis, B., J. Bellay, G. Musso, M. Costanzo, B. Papp *et al.*, 2010 Genetic interactions reveal the evolutionary trajectories of duplicate genes. Mol. Syst. Biol. 6: 429.
Veitia, R. A., 2003a Nonlinear effects in macromolecular assembly and dosage sensitivity. J. Theor. Biol. 220(1): 19–25.
Veitia, R. A., 2003b A sigmoidal transcriptional response: cooperativity, synergy and dosage effects. Biol. Rev. Camb. Philos. Soc. 78(1): 149–170.
Wagner, A., 2000a Decoupled evolution of coding region and mRNA expression patterns after gene duplication: implications for the neutralist-selectionist debate. Proc. Natl. Acad. Sci. USA 97(12): 6579–6584.
Wagner, A., 2000b Robustness against mutations in genetic networks of yeast. Nat. Genet. 24(4): 355–361.
Wagner, A., 2005 Robustness, evolvability, and neutrality. FEBS Lett. 579(8): 1772–1778.
Wang, Y., X. Wang, and A. H. Paterson, 2012 Genome and gene duplications and gene expression divergence: a view from plants. Ann. N. Y. Acad. Sci. 1256: 1–14.
Wendel, J. F., 2000 Genome evolution in polyploids. Plant Mol. Biol. 42(1): 225–249.
Wilke, C. O., and D. A. Drummond, 2006 Population genetics of translational robustness. Genetics 173(1): 473–481.
Wolfe, K. H., 2015 Origin of the yeast whole-genome duplication. PLoS Biol. 13(8): e1002221.
Wolfe, K. H., and D. C. Shields, 1997 Molecular evidence for an ancient duplication of the entire yeast genome. Nature 387(6634): 708–713.
Yang, Z., 2007 PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24(8): 1586–1591.
Zhang, Z. H., D. J. Jhaveri, V. M. Marshall, D. C. Bauer, J. Edson *et al.*, 2014 A comparative study of techniques for differential expression analysis on RNA-Seq data. PLoS One 9(8): e103207.

Communicating editor: B. J. Andrews

environmental microbiology
Environmental Microbiology (2017) **19**(3), 990–1007 doi:10.1111/1462-2920.13603

Glycerol stress in *Saccharomyces cerevisiae*: Cellular responses and evolved adaptations

Florian Mattenberger,[1†] Beatriz Sabater-Muñoz,[1,2†] John E. Hallsworth[3] and Mario A. Fares[1,2*]
[1]*Department of Abiotic Stress, Instituto de Biología Molecular y Celular de Plantas (CSIC-UPV), Valencia, Spain.*
[2]*Department of Genetic, Smurfit Institute of Genetics, University of Dublin, Trinity College, Dublin 2, Dublin, Ireland.*
[3]*Institute for Global Food Security, School of Biological Sciences, MBC, Queen's University Belfast, BT9 7BL, Northern Ireland.*

Summary

Glycerol synthesis is key to central metabolism and stress biology in *Saccharomyces cerevisiae*, yet the cellular adjustments needed to respond and adapt to glycerol stress are little understood. Here, we determined impacts of acute and chronic exposures to glycerol stress in *S. cerevisiae*. Glycerol stress can result from an increase of glycerol concentration in the medium due to the *S. cerevisiae* fermenting activity or other metabolic activities. Acute glycerol-stress led to a 50% decline in growth rate and altered transcription of more than 40% of genes. The increased genetic diversity in *S. cerevisiae* population, which had evolved in the standard nutrient medium for hundreds of generations, led to an increase in growth rate and altered transcriptome when such population was transferred to stressful media containing a high concentration of glycerol; 0.41 M (0.990 water activity). Evolution of *S. cerevisiae* populations during a 10-day period in the glycerol-containing medium led to transcriptome changes and readjustments to improve control of glycerol flux across the membrane, regulation of cell cycle, and more robust stress response; and a remarkable increase of growth rate under glycerol stress. Most of the observed regulatory changes arose in duplicated genes. These findings elucidate the physiological mechanisms, which underlie glycerol-stress response, and longer-term adaptations, in *S. cerevisiae*; they also have implications for enigmatic aspects of the ecology of this otherwise well-characterized yeast.

*For correspondence. E-mail: faresm@tcd.ie; Tel.: +353 1 8963521.
†These authors have contributed equally to this study.

Introduction

The specialist yeast *Saccharomyces cerevisiae*, known as an archetypal microbial weed in sugar-rich habitats, can also form part of microbial communities in diverse types of environments including soils, plant surfaces, and saline substrates (Botha, 2011; Cray *et al.*, 2013a; Lievens *et al.*, 2015). Whereas *S. cerevisiae* is a copiotroph, there are no reports of strains capable of biotic activity at water activities below 0.900, e.g. dried fruits, honey or sugar-saturated beet juice (Lievens *et al.*, 2015). By contrast, biomass-dense populations of this species are commonly found in sugar-rich substrates with intermediate water-activity values (0.990 to 0.920), including floral nectar, fruit juices, and the various substrates used to produce bioethanol (Cray *et al.*, 2013b; 2015; Lievens *et al.*, 2015). In its natural habitats, *S. cerevisiae* can be exposed to glycerol as a so-called byproduct of yeast and fungal metabolism (Hohmann, 2015) and in fermenting substrates, *S. cerevisiae* is known to produce and release glycerol to extracellular concentrations as high as 0.60 M (Basso *et al.*, 2008). Highly xerotolerant *S. cerevisiae* strains able to grow down to 0.880 to 0.900 water activity (see Hallsworth, 1998) must accumulate approximately 3.7 M glycerol in order to reduce the water activity of the cytosol to that of the extracellular milieu (de Lima Alves *et al.*, 2015). However, most strains can retain metabolic activity only down to 0.940 to 0.920 water activity, a value equivalent to 2.4 to 3.2 M glycerol.

Most importantly, and unlike metazoans, yeast utilizes nutrients not only as the source of energy to propel biosynthetic activity, but as the signals which control developmental, metabolic, and transcriptional activities of the cell (Broach, 2012). The robust and versatile stress biology of *S. cerevisiae* has been implicated in its ability to dominate the microbial communities within specific habitats (Cray *et al.*, 2013a); elucidating the mechanisms that underlie the dynamic responses, and

give rise to the adaptive plasticity, of yeast is imperative to underlying aspects of its molecular and cellular biology, environmental microbiology, ecology and evolution, and biotechnology. Although much of the microbial genome may be implicated in its stress biology, many fundamental aspects of the stress biology of *S. cerevisiae*, particularly the regulatory mechanisms that enable it to respond and adapt to the surrounding environment, remain only partially understood.

Recent developments in biophysical techniques have provided insights into chaotrope-induced stress mechanisms and responses, competitive ability and ecology of *S. cerevisiae* (Bhaganna *et al.*, 2010; Cray *et al.*, 2013a,b; 2015; de Lima Alves *et al.*, 2015). Whereas high extracellular glycerol concentrations can cause a transient turgor change to the *S. cerevisiae* cell, glycerol penetrates the plasma membrane within 1 to 2 min and does not therefore act as an osmotic stressor (Alemohammad and Knowles, 1974; Kiyosawa, 1991; Vilhelmsson and Miller, 2002; de Lima Alves *et al.*, 2015). Its primary mode-of-action as a stressor, therefore, is the depression of water activity and, at molar concentrations, an inhibitory level of chaotropic activity (Williams and Hallsworth, 2009; Cray *et al.*, 2013b; de Lima Alves *et al.*, 2015). Some of the most xerotolerant *S. cerevisiae* strains can grow at 3 to 4 M glycerol (Cray *et al.*, 2013a); these strains may be inhibited by both the low water-activity and high chaotropicity of the stressor (de Lima Alves *et al.*, 2015). It should be noted that the impact of chaotropic solutes on the flexibility of cellular macromolecules is dependent on the temperature as well as kosmotropic substances present (Hallsworth *et al.*, 2007; Bhaganna *et al.*, 2010; Ball and Hallsworth, 2015; Cray *et al.*, 2015; de Lima Alves *et al.*, 2015; Yakimov *et al.*, 2015). *S. cerevisiae* utilizes trehalose, which is most abundant in the cell during stationary phase, a highly kosmotropic compatible solute, to protect its macromolecular systems from the worst excesses of the combined chaotropicity of ethanol, acetaldehyde and high glycerol concentrations (Cray *et al.*, 2013a,; de Lima Alves *et al.*, 2015).

Important advances made in RNA sequencing technology have also made it possible to reveal additional, novel aspects of genetic and molecular mechanisms of stress response (see Taymaz-Nikerel *et al.*, 2016). Such developments, for instance, have enabled levels of experimental standardization and reproducibility, which elucidate biological mechanisms involved in molecular and cellular plasticity under stress and the sensitivity of a number of pathways to abiotic and biogenic stressors (Nagalakshmi *et al.*, 2008; van Dijk *et al.*, 2011; Nookaew *et al.*, 2012). Four pathways are instrumental in the effective adaptation of yeast to changes in environmental properties, including the signaling networks PKA, TORC1, Snf1, and Pho85. The PKA-signaling network controls cell growth, autophagy, glycogen synthesis, gluconeogenesis, and entry into quiescence (Taymaz-Nikerel *et al.*, 2016; Taymaz-Nikerel and Lara, 2016). TORC1, is implicated in cell growth control and stress response, with the deletion of *torc1* gene leading to increased thermotolerance and oxidative stress resistance (Cardenas *et al.*, 1999; Bjornsti and Houghton, 2004; Martin and Hall, 2005; Aramburu *et al.*, 2014). The Snf1 kinase-signaling pathway is involved in energy homeostasis and response to glucose or carbon limitations (Turcotte *et al.*, 2010; Ghillebert *et al.*, 2011; Crozet *et al.*, 2014; Emanuelle *et al.*, 2016). Under carbon-substrate limitations, there is massive reprogramming of gene expression characterized by the altered expression of genes involved in gluconeogenesis, the glyoxylate cycle, and the tricarboxylic acid cycle (Turcotte *et al.*, 2010). Finally, the Pho85 signaling pathway responds to phosphate limitations and starvation, which is central to the biosynthesis of nucleotides, phospholipids and metabolites (Huang *et al.*, 2007; Yadav *et al.*, 2016). This pathway is also involved in the response to compromised protein folding and oxide-reductive stress, and other challenges (DeRisi *et al.*, 1997; Carroll and O'Shea, 2002).

Variation in the levels of glycerol is among the most important of the challenges which require a cellular response, because glycerol is often involved in cellular homeostasis and in adjustment to changes in extracellular osmolarity (Hohmann *et al.*, 2007; Hubmann *et al.*, 2011). The levels of glycerol are maintained in *S. cerevisiae* by a fine-tuned regulation of glycerol-proton symport via the Stl1 transporter (Tulha *et al.*, 2010; Duskova *et al.*, 2015). In brief, the yeast Hot1 transcription factor binds physically to the promoter of *stl1*, and in response to osmostress, binds active stress-activated protein (Hog1), a transcription factor from the mitogen-activated protein kinase (MAPK) pathway. Once Hog1 bound to the promoter of *stl1*, this transcription factor increases the rate of transcription by recruiting the chromatin-remodeling protein Rpd3 and associating with RNA PolIII and components of the mediator complex (Alepuz *et al.*, 2003; De Nadal *et al.*, 2004). Interestingly, Hot1 transcription factor seems to bind specifically the promoter of *stl1* and is critical for its transcription (Bai *et al.*, 2015). Glycerol is also produced in *S. cerevisiae* through glycolysis and the reduction of the glycolytic intermediate dihydroxyacetone phosphate to glycerol-3-phosphate and the subsequent oxidation of NADH to NAD^+. What transcriptomic changes are involved in responding and adapting to variations in the levels of glycerol remains little understood. Most importantly, whether such transcriptomic responses can evolve sufficiently to improve cell growth under glycerol changes has not been explored.

Yeast exhibits a genome-wide transcriptomic response to stress produced by glucose limitation (Ferea *et al.*, 1999). The fact that many of the transcriptional alterations are not stress-specific (Causton *et al.*, 2001; Ideker *et al.*, 2001; Stern *et al.*, 2007; Cormier *et al.*, 2010) suggests the

possibility that these transcriptional alterations are not the result of adaptive evolutionary changes in the response to stress but is a property emerging from a universal feature underlying regulatory networks. On the other hand, the expression divergence between duplicate gene copies under standard nutrient medium and stress conditions, suggests that response to stress may be the result of adaptive changes in the regulation of duplicated genes (Blanc and Wolfe, 2004; Li *et al.*, 2005; Conant and Wolfe, 2006; Thompson *et al.*, 2013). Gene duplication is universally recognized as a source of novel functions and adaptations. This is because after the duplication of a gene, the resulting identical gene copies generate genetic redundancy and relax natural selection against one gene copy. The relaxed gene copy can explore novel genotypes and eventually access new phenotypes while its sister copy gene maintains the ancestral function (Ohno, 1970; 1999). Accordingly, gene duplication has been linked to major evolutionary leaps in plants (Otto and Whitton, 2000; Wendel, 2000; Holub, 2001; Lespinet *et al.*, 2002; Kim *et al.*, 2004; Cui *et al.*, 2006; Carretero-Paulet and Fares, 2012) and animals (Otto and Whitton, 2000; Hoegg *et al.*, 2004). Evolution of gene expression and expression plasticity through duplication is a stronger force than evolution of function, as shown in recent experimental analyses (Keane *et al.*, 2014).

To both determine whether the response to stress is adaptive or not and address the associate knowledge gaps, here we determine the transcriptomic changes in standard medium-adapted *S. cerevisiae* growing under glycerol-induced stress. The specific aims were to characterize three phenomena: (i) the transcriptomic initial response of *S. cerevisiae* growing in glycerol; (ii) impacts of the genetic background of *S. cerevisiae* population on the response to exposure to glycerol-induced stress, and (iii) transcriptomic changes underlying the adaptation of *S. cerevisiae* during exposure to glycerol-induced stress for hundreds of generations (during a 10-day period). The sudden exposure to glycerol-induced stress triggers a genome-wide transcriptomic response, and we identified these transcriptomic responses and the associated changes in cellular processes. We show that the genetic variation in the population of *S. cerevisiae* can affect the transcriptomic response to glycerol-induced stress. The study thereby reveals the fine-tuning of *S. cerevisiae* regulatory re-programming during its adaptation and growth improvement when growing in glycerol for a long period. Finally, we present evidence that transcriptionally altered duplicated genes during glycerol-induced stress are the core genes in the immediate response (i.e., short-term) and in the response to alterations in glycerol levels during the evolution of *S. cerevisiae* during a 10-day exposure to glycerol.

Results

Glycerol acts as a potent cellular stressor of S. cerevisiae

From an initial *S. cerevisiae* colony (ancestral colony) we grew a culture in our standard nutrient medium; yeast extract, peptone and dextrose medium (YPD; 0.998 water activity). This was the control population of *S. cerevisiae*, which included genetically related *S. cerevisiae* cells growing in a rich medium to which *S. cerevisiae* was adapted. Another culture was initiated from the ancestral colony in yeast extract, peptone, and 0.41 M glycerol (YPG), a stressful environment (0.990 water activity) to which *S. cerevisiae* was not adapted (Fig. 1). We measured cell density of the populations of *S. cerevisiae* in the standard medium (YPD), and stressful medium (YPG), using optical density (OD_{600nm} was determined for five biological replicates at 15-min intervals during a 92-hr period) to provide a measure of cell number and metabolic activity under these media. The *S. cerevisiae* growth curve in YPD was sigmoidal curve with a maximum growth rate of ($\mu_{max} \pm$ s.d.m.$= 0.33\ h^{-1} \pm 0.01$), as calculated by the program GrowthRates (Experimental procedures). The growth rate of *S. cerevisiae* declined in YPG (Fig. 2A), with a maximum growth rate in YPG of ($\mu_{max} \pm$ s.d.m.$= 0.16\ h^{-1} \pm 0.003$), a rate that was significantly lower than the growth rate in YPD (normal test: $z = 24.28$, $P \ll 0.001$; Fig. 2B). Importantly, the lag time in YPG (Lagt$= 21.8$ h) was approximately 10 times longer than that in YPD (Lagt$= 2.7$ h) (Fig. 2A). The difference in lagging time may be required for the regulatory reprogramming of the cell to respond to YPG. Interestingly, cell number was greater in YPG than in YPD when their respective cultures reached the stationary phase (Fig. 2A).

Transcriptome-wide, cellular stress response to glycerol

The response to glycerol stress was further characterized by comparing the transcriptome of the control population (standard medium) to that of the glycerol-stressed population (see Experimental procedures; Fig. 1). To this end, we compared the transcriptional level of each of the 5825 genes for which we obtained reliable RNA read counts between the control and the stressed populations (Experimental procedures).

Glycerol-induced stress altered significantly the expression (false discovery rate FDR < 0.05) of 2748 out of the 5842 (47.04% of the total) analyzed genes in *S. cerevisiae* (Fig. 3). Out of the 2748 differentially expressed genes, 1331 were up-regulated (i.e., $\log_2 FC > 1$) a proportion similar to that of down-regulated genes (Binomial test: $P = 0.13$). The extent of transcriptomic response was of the same order as previously reported for a number of stressful environmental conditions (Gasch *et al.*, 2000).

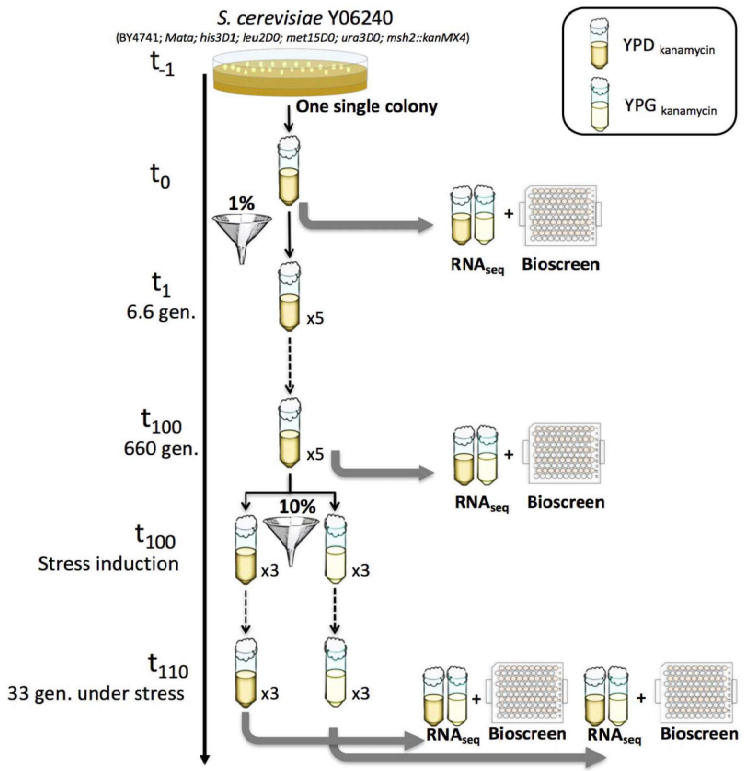


Fig. 1. Experimental procedure to test the transcriptomic response of *S. cerevisiae* populations to glycerol-induced stress.
From a single colony of *S. cerevisiae* strain YO6240, we derived a population which was used in the rest of the experiments (and named this population t_0). This population was subjected to glycerol-induced stress and the transcriptomic changes from growing in glucose (YPD) to growing in glycerol (YPG) were quantified. From the population at t_0 we evolved a population in YPD for 100 passages (approximately 660 generations of *S. cerevisiae*) by transferring a 1% dilution daily to a tube with a fresh YPD medium. Then, we subjected this population at t_{100} to glycerol-induced stress and quantified its transcriptomic response. Finally, the population at t_{100} was exposed to long-term glycerol induced stress by evolving for 10 passages (66 generations) under glycerol, with a control population growing under YPD. This evolution took place by transferring a 10% of the population to a new tube containing YPG or YPD media. All populations were grown in the presence of kanamycin to minimise the possibility of bacterial contamination.

We analyzed the distribution of de-regulated genes among the different cell process categories classified according to Gene Ontology (GO) terms using the *Saccharomyces* Genome Database (SGD: http://www.yeastgenome.org). De-regulation of genes affected fundamental processes in the cell (Supporting Information Table S1), including translation, mainly ribosomal genes and genes involved in ribosomal biogenesis, oxidation-reduction and cellular respiration, as a consequence of shifting from a fermentative (in YPD) to an oxidative growth (in YPG), genes from the mitochondrial respiratory chain complex assembly, and genes encoding proteins involved in the transport of substances across the membrane (Supporting Information Table S1). Importantly, genes involved in deriving energy by oxidation from organic compounds were also transcriptionally altered (Supporting Information

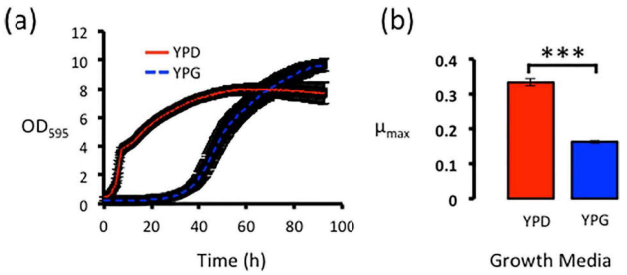


Fig. 2. Environmental changes that involve a replacement of glucose with glycerol impact negatively the ability of *S. cerevisiae* to grow.
We grew *S. cerevisiae* populations in normal conditions (YPD) or in growth media containing 3% glycerol (YPG).
A. Growth curves show that *S. cerevisiae* grows at higher rate in YPD (red sigmoidal curve) than in YPG (blue sigmoidal curve). Shaded area throughout the curve represents the standard deviations for each time point.
B. The maximum growth rate of *S. cerevisiae* in YPD (red column) was significantly greater than its growth rate in YPG (blue column) using a Wilcoxon rank test (***$P < 0.001$).

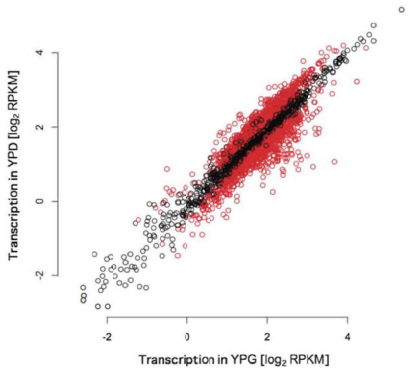


Fig. 3. Growth of *S. cerevisiae* in glycerol unfolds a genome-wide transcriptional response.
The plot represents the comparison of the expression level (measured as the logarithm of the number of reads per billion (RPKM)) of each *S. cerevisiae* gene under YPD medium (*Y*-axis) and YPG medium (*X*-axis). Expression levels of genes are represented as Reads per billion (RPKM) logarithmically transformed. Black dots in the plot represent genes with no evidence of transcriptional alterations when comparing their expression in YPD with this in YPG. Red dots represent' genes exhibiting significant transcriptional changes when comparing their expression in the two growth media and these are either up-regulated in YPG compared with YPD (red dots above the diagonal) or down-regulated (red dots below the diagonal).

Table S1). A number of environmental stress-response genes were transcriptionally altered, including genes from the pathway of positive and negative regulators of protein kinase A (PKA) such as *pkh1* and *pkh2*, *ras1*, *cdc25*, and *gpa2*. Only the gene encoding one β-subunit of the SNF1 complex, *sip2*, was transcriptionally altered. Several transcription factors were also altered, including *cat8* and *rds2*, both of which are potent activators of gluconeogenesis (Schuller, 2003; Turcotte *et al.*, 2010).

Identification of the genes that were up-regulated or down-regulated in cells grown on YPG indicated the primary physiological changes that had taken place (Fig. 4). In general, growth in YPG up-regulated genes involved in respiration, including redox and genes from the mitochondrial respiration chain assembly. It also up-regulated genes involved in deriving energy from oxidative organic compounds, TCA cycle, ATP synthesis, fatty acid oxidation, phosphorylation, and transmembrane transport, among others (Supporting Information Table S2 and Fig. 4A). In contrast, translation, represented by the biogenesis of ribosomes and maturation of ribosomal subunits, transcription, protein folding, cellular biosynthesis, glycosylation, and nucleic acid metabolism, among others were down-regulated when growing in YPG relative to YPD (Supporting Information Table S3 and Fig. 4B).

Increased genetic diversity increases the phenotypic plasticity of S. cerevisiae

Population t_0 was derived from a single colony, with this population being relatively limited in terms of genetic diversity, and thus the observed transcriptomic response may not represent a natural population. To test how the genetic variation of a population may influence the transcriptomic response when such population is challenged with an environmental change (for example, growth in YPG instead of YPD), we first increased the genetic variation in the initial population by propagating this population of *S. cerevisiae*

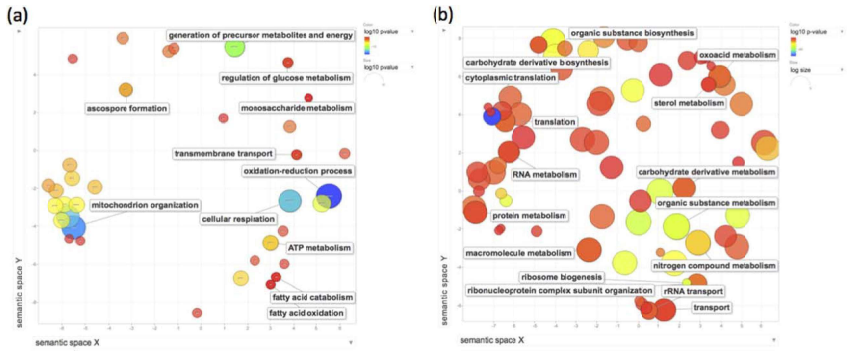


Fig. 4. Growth in YPG induces up-regulation and down-regulation of genes involved in a large number of cellular processes.
Significantly transcribed genes under glycerol stress were classified according to their gene ontology (GO) term by using Panther GO sorting system. The scatterplot shows the cluster representatives in a two-dimensional space, on which bubble color indicate the logarithm of GO term *P* value (based on the number of genes belonging to this category), whereas the bubble size indicates the frequency of the GO term in the organism database (the more general term, the bigger bubble size). Only the most relevant cluster descriptors are shown.
A. Processes enriched for up-regulated genes under YPG growth conditions.
B. Processes enriched for down-regulated genes under YPG growth conditions.

growing in YPD during 100 passages, with each passage involving the transfer of a hundredth of the population to a new flask with fresh YPD medium (Experimental procedures) (Fig. 1). In each passage, we transferred a hundredth of the population, corresponding to $\log_2 100 = 6.6$ generations of the yeast, to a new flask containing fresh YPD medium. Taking into account a mutation rate of 10^{-8} mutations/nucleotide for *S. cerevisiae* Δ*msh2* strain used in this study, on average each passage involves the fixation of one mutation per cell every day. Most of the variation accumulated during the experimental evolution of propagated population was neutral (i.e., there were no deleterious mutations fixed by genetic drift) since the maximum growth rate of the propagated populations ($\mu_{max} \pm$ s.d.m. $= 0.35 \pm 0.02$) was not significantly different from that of the original population (Normal test: $z = 1$, $P = 0.16$).

After 100 passages (hereafter t_{100}), we challenged the propagated population by growing it in YPG and analyzed its transcriptomic changes at exponential phase by comparing its transcriptomes in YPG and YPD. We then compared the transcriptomic changes at t_{100} with those at t_0. The number of transcriptionally altered genes ($N = 2429$) at t_{100} was 12% lower than that at t_0. However, only 58% of the genes transcriptionally altered at t_{100} when yeast had been cultured in YPG ($N = 1426$) were also altered at t_0, while 1003 genes that were differentially regulated in YPG compared with YPD at t_{100} were not altered at t_0 (Fig. 5A). The fact that roughly half of the transcriptomic response at t_{100} to glycerol was unaltered at t_0 is intriguing (Fig. 5A). We believe that the set of genes transcriptionally altered in both of the populations, at t_0 and t_{100} ($N = 1426$), represents the core genes of response to glycerol-induced stress, while the set of genes altered in YPG at t_0 but not at t_{100} ($N = 1321$), or those altered at t_{100} but not at t_0 ($N = 1003$) must represent transcriptomic variation that is dependent on the genetic composition of the population. Identification of the cellular processes significantly enriched for the set of altered genes in the population challenged with glycerol at t_0 and at t_{100} and the set of altered genes in only one of these populations revealed meaningful differences between the two sets of altered genes. The processes of translation, respiration, nucleotide metabolism, energy derivation by respiration, carboxylic acid metabolism, ion transmembrane transport, and ATP metabolism were enriched for the set of transcriptionally altered genes in both t_0 and t_{100} populations (Fig. 5B; Supporting Information Table S4). Processes mostly concerned with regulation of transcription and gene expression, protein export to the mitochondrion, mitochondrial translation, and transcription were significantly enriched for genes transcriptionally altered at t_0 but not t_{100} (Fig. 5C and Supporting Information Table S5). Finally, cellular processes involved in the regulation of transcription and a number of other fundamental biological processes were significantly enriched for genes altered at t_{100} but not t_0 (Supporting Information Table S6).

Out of the genes up-regulated in cell cultured in YPG (in comparison with genes' expressions in cells cultured in

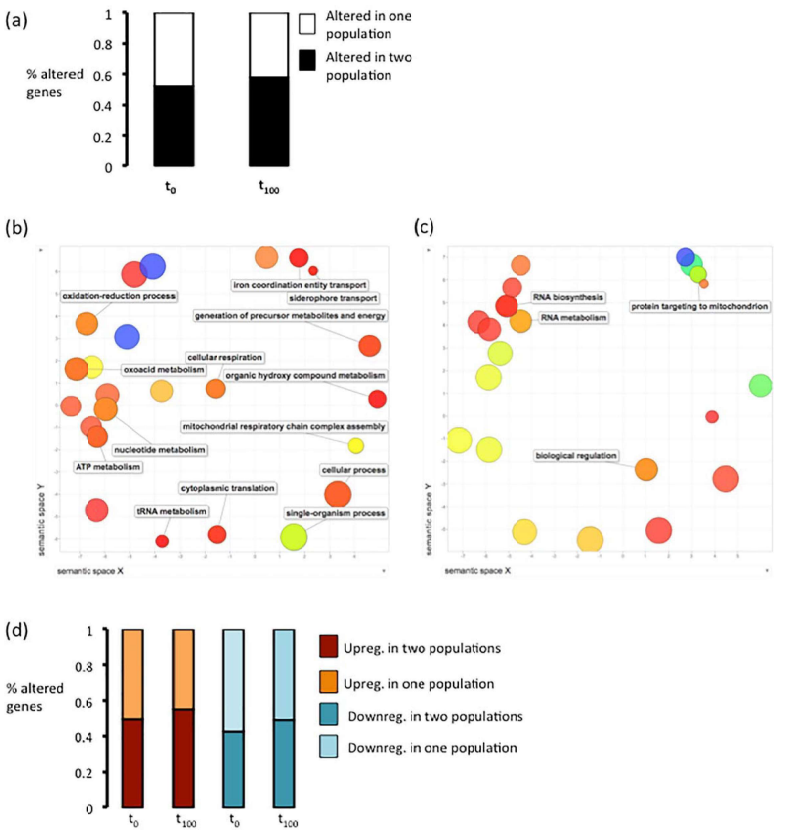


Fig. 5. Populations of *S. cerevisiae* isolated at different times of their evolution exhibit overlapping and non-overlapping transcriptional responses to glycerol.
A. Identification of genes that are transcriptionally altered in the ancestral population (t_0) and the population evolved for 100 days in YPD (t_{100}). Black areas of the columns represent the percentage of genes transcriptionally altered in that population when grown in YPG that are also altered in the other population, while white are represent genes altered in one population but not in the other.
B. Semantic clustering of cellular processes enriched for genes that were transcriptionally altered in both of the populations (at t_0 and t_{100}). The diameter of the circle represents the proportion of genes in a particular cellular process found transcriptionally altered, while the different processes are color-coded. The color of the bubbles represents the proportion of genes in a particular cellular process found transcriptionally altered (log *P* value), while the size indicates the frequency of the GO term in the organism.
C. Semantic clustering of cellular processes enriched for genes that were transcriptionally altered in t_0 but not t_{100}.
D. Proportion of genes that are up-regulated or down-regulated in one and/or both population(s).

YPD) at t_{100}, 55% ($N = 660$) were also up-regulated in cells cultured in YPG at t_0. This change is 6% greater than the down-regulated genes that were common to the t_{100} and t_0 populations (49%, $N = 602$, Fisher's exact test: $F = 1.2$, $P = 0.01$) (Fig. 5D). Only 82 genes that were up-regulated at t_{100} (i.e., approximately 6.7% of those up-regulated) were among the down-regulated genes found at t_0, pointing to a high level of consistency between

the transcriptomic data of *S. cerevisiae* isolated at t_0 and t_{100}. Therefore, in general terms, up-regulated genes at t_{100} that are transcriptionally altered at t_0 are also up-regulated at t_0, and down-regulated genes at t_{100} that are altered at t_0 are also down-regulated at t_0. The low overlap in the transcriptional profiles of the populations at t_0 and t_{100} was paralleled with differences in the growth rates of these populations in YPG. Indeed, the population evolved for 100 passages in YPD exhibited higher maximum growth rate in YPG ($\mu_{max} \pm$ s.d.m. $= 0.19\ h^{-1} \pm 0.015$) than t_0 populations ($\mu_{max} \pm$ s.d.m. $= 0.16\ h^{-1} \pm 0.003$; Normal test: $z = 3.33$, $P < 0.025$).

Transcriptomic-wide, adaptive evolution after successive generations under glycerol stress

To determine what transcriptomic changes mediate adaptive growth in glycerol, we evolved in triplicate the t_{100}-evolved population (i.e., genetically diverse population) for 10 passages separately in YPG and in YPD by transferring a tenth of the population to fresh YPG-containing flask every 24 h, with each passage allowing the transfer of $\log_2 10 = 3.33$ generations of the yeast. We used a lower dilution in the transfers in YPG than that used during the 100 passages of evolution in YPD, to increase the population size transferred and consequently the efficacy of selection in the adaptation to YPG. After 10 passages of evolution, the evolved population exhibited a significant increase in its maximum growth rate in YPG ($\mu_{max} \pm$ s.d.m. $= 0.26\ h^{-1} \pm 0.013$) compared with the populations isolated at t_{100} (Normal test: $z = 4.67$, $P < 0.01$) and t_0 (normal test: $z = 6.66$, $P < 0.001$). The growth rate of the t_{110}-population evolved in YPG was greater there than that of t_{110}-population grown in YPD ($\mu_{max} \pm$ s.d.m. $= 0.23\ h^{-1} \pm 0.01$) (Normal test: $z = 3$, $P < 0.025$).

Analysis of the transcriptomic changes in the t_{110}-population evolved in YPG in comparison with the t_{110}-population evolved in YPD revealed that adaptation to YPG involved a massive regulatory re-programming of *S. cerevisiae* with 2640 genes exhibiting differential expression in YPG compared with YPD. Out of the transcriptionally altered genes, 1274 genes (48.3% of the total) were up-regulated and 1366 were down-regulated. The list of altered genes in the t_{110}-population evolved in YPG compared with YPD with that list in the t_{100}-population yielded 1383 genes out of the 2640 as transcriptionally altered in YPG in the t_{100} and t_{110} populations (52.4%), and 1453 genes (55%) were coincidentally altered in YPG compared with YPD in t_{110} and t_0 populations. We identified core-altered genes, those that were transcriptionally altered in the t_0, t_{100}, and t_{110} populations, and thus were genes responsive to glycerol-induced stress regardless of the length of exposure of the population to glycerol. This list of genes included 902 core genes (Supporting Information Table S7) that were transcriptionally altered in YPG compared with YPD. We also identified the genes that were uniquely altered in cells cultured in YPG in each population and not in the others (Fig. 6A).

Altered core-genes and population specific genes affected different cellular processes. Processes that are involved in the transition from fermentative to oxidative metabolism, controlling osmotic stress through transporters at the environment-cell interface and energy production were enriched for core-altered genes. Examples of such genes include those encoding proteins from the mitochondrial respiratory chain complex, proteins involved in oxidative-reductive processes, cellular respiration, ion and substrates transport and ATP metabolism (Fig. 6B, Supporting Information Table S8). Cellular processes concerned with energy saving such as translation, biosynthesis processes, protein metabolism, intra-cellular transport and protein cellular localization genes were enriched for genes that were transcriptionally altered only in the t_0 population (Fig. 6C, Supporting Information Table S9). A subset of these processes were also enriched for genes transcriptionally altered in the t_{100} population (Supporting Information Table S10). Interestingly, the trafficking of proteins across the mitochondrial membrane, localization of proteins to mitochondrion, and nucleic acid metabolism were particularly enriched for altered genes in the t_{100} population (Fig. 6D). Finally, altered genes in t_{110}-population affected processes involved in stress stimuli response, cell cycle, and regulation, which were not enriched for altered genes from the t_0 and t_{100} populations (Fig. 6E, Supporting Information Table S11).

Transcriptional response to glycerol-induced stress is mainly mediated by duplicated genes

Since *S. cerevisiae* bears 2240 duplicated genes (1120 pairs, roughly 32.8% of its genome), and gene duplication is often linked to the origin of novel adaptations, we investigated whether duplicated genes have driven the transcriptomic responses to glycerol-induced stress. Of all the transcriptionally altered genes in the t_0 population, 1113 genes, corresponding to 40.5% of all transcriptionally altered genes, were duplicated genes, a proportion significantly greater than expected by chance (Binomial test: $P < 2.2 \times 10^{-16}$). Complementary to duplicated genes, 1634 of the transcriptional altered genes in this population (59.5% of all altered genes) were singletons, a proportion lower than expected by chance (binomial test: $P < 2.2 \times 10^{-16}$). Therefore, duplicated genes mostly drove transcriptional response to glycerol in the t_0 population. Duplicated genes in *S. cerevisiae* were generated through two different mechanisms, including whole genome duplication ($N = 554$ pairs, corresponding to 49.5% of all duplicates), also known as WGDs (Wolfe and Shields,

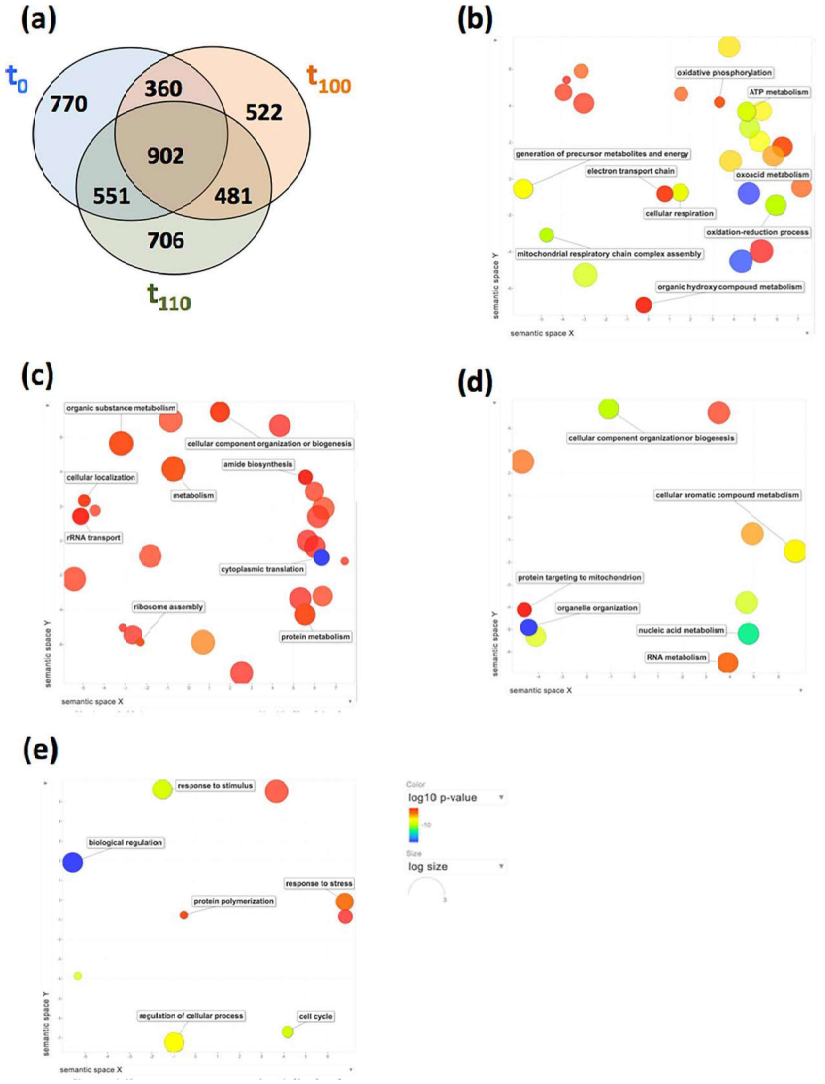


Fig. 6. Growth in glycerol induces two sets of genes, those that are common to all three populations (t_0, t_{100}, and t_{110}), also known as core genes, and those that are unique to one of the populations.
A. Venn diagram identifying core genes and unique genes transcriptionally altered in one population.
B. Semantic clustering of cellular processes enriched for genes that were transcriptionally altered in all three populations. The color of the bubbles represents the proportion of genes in a particular cellular process found transcriptionally altered (log *P* value), while the size indicates the frequency of the GO term in the organism.
C. Semantic clustering of cellular processes enriched for genes that were transcriptionally altered in the population at t_0 growing in YPG.
D. Semantic clustering of cellular processes enriched for genes that were transcriptionally altered in the population at t_{100} growing in YPG.
E. Cellular processes enriched for genes that were transcriptionally altered in the population at t_{110} growing in YPG.

1997; Marcet-Houben and Gabaldon, 2015) and small-scale duplications ($N = 566$ pairs, 50.5% of all duplicates), also known as SSDs (Fares *et al.*, 2013; Keane *et al.*, 2014). Since the evolution of duplicated genes is strongly dependent on the mechanism of duplication (Carretero-Paulet and Fares, 2012; Fares *et al.*, 2013; Keane *et al.*, 2014), we investigated whether differences in the presence of WGDs and SSDs existed in the transcriptomic responses to glycerol. We found that WGDs were significantly more represented among transcriptionally altered genes in the t_0 population cultured in YPG than expected (Binomial test: $P = 4 \times 10^{-4}$), while the opposite was true for SSDs. Similar results were observed for t_{100} and t_{110} populations (Supporting Information Table S12), although at t_{110} the population exhibited no differences between WGDs and SSDs (Binomial test: $P = 0.34$).

We identified the core-altered duplicated genes, those duplicated genes that were altered in the t_0, t_{100}, and t_{110} populations ($N = 369$, Fig. 7A). Noticeably, the proportion of core-altered duplicated genes (369/1079 = 34.2%) was significantly greater than the proportion of core-altered singletons (533/1821 = 29.3%) (Fisher's exact test: odds ratio $F = 1.26$, $P = 6.2 \times 10^{-3}$). Interestingly, a differential pattern was found between the transcriptional response of singletons and duplicates: while the majority of duplicates were down-regulated in cells cultured in YPG for all three populations, the opposite pattern was observed for singletons (Fig. 7B). Indeed, the number of up-regulated duplicates in cells cultured in YPG was lower than expected at t_0, t_{100}, and t_{110} (Fig. 7B). Conversely, singleton genes exhibited higher proportion of up-regulation than expected in all three populations (Fig. 7B).

Examination of the cellular processes enriched for altered duplicated genes identified important differences between populations adapted to YPD and those adapted to YPG. Core-altered duplicates, those that were transcriptionally responsive to glycerol in all three populations, were distributed among cellular processes concerned with transport of carbohydrates and organic substrates and respiration, including oxidation-reduction, energy derivation by oxidation of organic compounds (Fig. 7C and Supporting Information Table S12). In the t_0 population challenged with glycerol most of the processes enriched for duplicated genes that were transcriptionally altered only in the t_0 population but not in the t_{100} or t_{110} populations were concerned with translation and biosynthetic processes (Fig. 7D and Supporting Information Table S13). We found no cellular processes enriched for altered duplicated genes specifically in the t_{100} population but not in the populations corresponding to either of the other two time points. Finally, the t_{110} population adapted to YPG exhibited transcriptional alterations in a number of duplicates mostly involved in response to stress and stimuli as well as in regulation of biological processes (Fig. 7E and Supporting Information Table S14). It is worth noticing that analyses of up-regulated duplicates in the three populations also yielded very different outcomes for populations cultured in YPD (t_0 and t_{100} populations) compared with the population cultured in YPG: up-regulated duplicates in the t_0 and t_{100} populations were preferentially distributed in cellular processes of sexual reproduction, sporulation, transport, oxidative-reductive processes, and ascospore wall biogenesis, among others (Supporting Information Tables S15 and S16). In contrast to this, in the t_{110} population many catabolic processes, including energy derivation by oxidation, TCA, and lipid catabolic processes were enriched for duplicated genes that were up-regulated in cells cultured in YPG (Supporting Information Table S17).

Adaptations of cellular metabolism of experimentally evolved S. cerevisiae *populations*

How metabolically divergent is the population adapted to glycerol from those responsive to glycerol but not adapted to it? We measured the 'metabolic distance' between the t_0, t_{100}, and t_{110} populations using a simple measure of distance between cellular processes terms (see Experimental procedures). We analyzed separately the metabolic distance between pairs of the t_0, t_{100}, and t_{110} populations for core-altered genes (i.e., those genes transcriptionally altered in all three populations), up-regulated and down-regulated genes in each of the populations (Table 1). Altered genes in cells from the t_0 and t_{100} populations, were metabolically closer to each other than either was to those in the t_{110} population. Up-regulated genes that included singletons and duplicates, in the t_{100} population exhibited greater metabolic distance to the t_0 population than they did to those observed in the t_{110} population adapted to glycerol. For down-regulated genes, on the other hand, the distance between t_0 and t_{100} populations was

shorter than that of either to t_{110} population. Using only core duplicated genes, those duplicated genes transcriptionally altered in all three populations, the ancestral population showed low metabolic distance to t_{100} population, and t_{100} population was metabolically close to t_{110} population, but t_0 was distant from t_{110} population. Both the up-regulated and down-regulated sets of duplicated genes exhibited greater distance of the t_{110} population adapted to glycerol to any of the other two populations than this distance between t_0 and t_{100} populations.

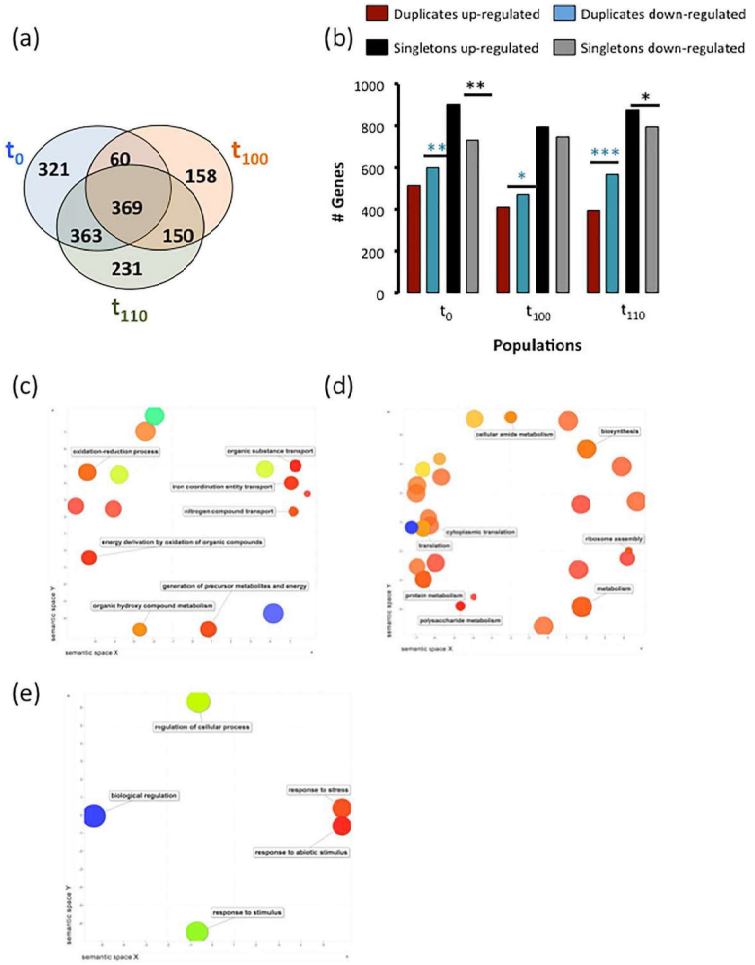


Fig. 7. Duplicated genes are more enriched than singletons for transcriptionally altered genes and show different response patterns to those of singletons under YPG growth conditions.
A. Venn diagram identifying common duplicates response in the three populations (t_0, t_{100}, and t_{110}) and duplicates uniquely responsive to glycerol in each of the populations.
B. Duplicates show more down-regulation than up-regulation, while singletons exhibit more up-regulation than down-regulation (*, **, *** refer to $P < 0.05$, $P < 0.01$, $P < 0.001$ under a binomial test, and colors identifies the most abundant pattern in the comparison between up- and down-regulated genes).
C. Semantic clustering of cellular processes enriched for duplicated genes that were transcriptionally altered in all three populations. The color of the bubbles represents the proportion of genes in a particular cellular process found transcriptionally altered (log *P* value), while the size indicates the frequency of the GO term in the organism.
D. Semantic clustering of cellular processes enriched for duplicated genes that were transcriptionally altered in in the population at t_0.
E. Semantic clustering of cellular processes enriched for duplicated genes that were transcriptionally altered in in the population at t_{110}.

Table 1. Metabolic distance between populations t_0, t_{100}, and t_{110}.

	Metabolic distance ($MD_{i,j}$)								
	Core altered			Up-regulated			Down-regulated		
Genes	t_0 vs. t_{100}	t_0 vs. t_{110}	t_{100} vs. t_{110}	t_0 vs. t_{100}	t_0 vs. t_{110} t_{110}	t_{100} vs. t_{110}	t_0 vs. t_{100}	t_0 vs. t_{110}	t_{100} vs. t_{110}
All	0.27	0.37	0.29	0.51	0.34	0.28	0.39	0.44	0.46
Duplicates	0.09	0.38	0.09	0.37	0.51	0.52	0.30	0.63	0.60

Discussion

Exposure to glycerol-induced stress triggers a genome-wide transcriptomic response

Unicellular organisms, in particular non-motile ones, must have developed a number of metabolic strategies to sense minute changes in the environment and prepare to combat environmental fluctuations. Here, we show that even small increases in the extracellular concentrations of glycerol lead to dramatic transcriptional re-programming of the yeast *S. cerevisiae*. This re-programming affects a great proportion of the genes and it extends to cellular processes affecting all the organelles in the cell. Such genome-wide re-programming has been shown before to take place under a number of stress conditions. For example, in an experiment of adaptation of *S. cerevisiae* through experimental evolution under glucose-limited conditions, authors observed a large expression re-programming affecting several hundreds of genes (Ferea *et al.*, 1999). Following these experiments, there are two kinds of transcriptomic responses, which vary with regards to the time of response, from very quick responses to very slow ones (Yosef and Regev, 2011). These two types of responses have been shown to involve a large number of genes (Taymaz-Nikerel *et al.*, 2016), in good agreement with the data presented in this study. Indeed, our results show that subjecting an initial population of *S. cerevisiae* to mild glycerol stress leads to the alteration of the expression of hundreds of genes in a very short time. Most such genes are linked to a number of pathways that include positive and negative regulators of protein kinase A and other transport pathways that are involved in buffering the effects of glycerol variation around the membrane, including genes encoding transport proteins for degradation of unfolded proteins such as SSA1, ERP2, PMT2, etc.

Two well-differentiated patterns are observed during the transcriptional alterations under stress. On the one hand, genes involved in responding directly to exposure of cells to glycerol and the activation of the respiratory metabolism of the yeast cell are up-regulated. On the other hand, genes involved in the ribosomal biogenesis or in energy-dependent processes, such as translation and transcription are down-regulated. These patterns are in agreement with the hypothesis that such a transcriptional response does not counterbalance the effects of stress on fitness but are a first step towards adapting the cell to the new perturbed environment. Accordingly, despite this large re-programming of the cell, the growth rate of the population declines under glycerol stress in the ancestral (t_0) population when compared with its growth rate in YPD. This result is concordant with the lack of correlation between the transcriptional response and fitness gain or buffering of the effects of stress on cell's growth rate (Giaever *et al.*, 2002; Giaever and Nislow, 2014). On the contrary, it seems that genes involved in environmental stress response are correlated with a decline in cell growth rate (Gasch *et al.*, 2000; Regenberg *et al.*, 2006; Castrillo *et al.*, 2007; Brauer *et al.*, 2008; Frazio *et al.*, 2008).

The genomic background of a S. cerevisiae *population impacts glycerol-stress responses*

An important question is how does the genetic background of a cell influence its response capacity to environmental perturbations? In this study, the experimental evolution of a population of *S. cerevisiae* for a greater number of generations in YPD (t_{100} population) demonstrates that the genetic composition of the population largely influences the transcriptional profile of the cells under stress. Indeed, we observed that only a fraction of the genes responding to stress in the evolved population are also responsive to stress in the ancestral non-evolved population. Since the ancestral population was founded from a single clone and, hence, lacked genetic variability, the difference in the transcriptional profile between the evolved and the ancestral population is likely the result of a change in the genotype of the population. An important result deriving from our analyses is that the population evolving in YPD, but not in the presence of glycerol, exhibits higher fitness when cultured in the presence of glycerol than its parental non-evolved population. A possible explanation of this finding is that the evolved population has explored a wide range of genotypes some of which may be adaptive to the environment containing glycerol, a phenomenon encapsulated within the term exaptation (Fares, 2015). The high dependence of the transcriptional plasticity of yeast on the genetic composition of the population supports previous suggestions that the transcriptional re-programming of the cell does not occur in a stress-specific manner (Causton *et al.*, 2001; Ideker *et al.*, 2001; Stern *et al.*, 2007; Cormier *et al.*, 2010). Instead the transcriptional response to stress may be a biological property emerging from a universal feature underlying regulatory networks. Since biological systems are known to bear distributed robustness—i.e., the product of a specific metabolic pathway could be achieved through many other alternative and unrelated pathways (Wagner, 2015), the differential transcriptional profiles of the evolved and non-evolved populations in response to glycerol may be due to this distributed robustness, which in turn can lead to exaptations.

Is glycerol-stress response a combination of adaptive responses and system-level emerging properties?

The evolution of the yeast population in the presence of glycerol (i.e., the t_{110} population) yields results supporting the regulatory potential of yeast to improve its adaptation to changing environments. We distinguish between two main responses, those generated against instantaneous environmental perturbations (i.e., quick responses) and those emerging from the selective fixation of adaptive regulatory changes that emerge when a population is subjected to constant stress. Accordingly, the quick response involves a set of genes that are de-regulated in the three populations (t_0, t_{100} and t_{110} populations). However, the adaptive response generated in the population evolved in YPG (t_{110} population) mostly includes genes involved in stress response and regulation processes. This late induction of genes involved in stress response has been previously observed in populations of yeast subjected to high temperatures (Gasch *et al.*, 2000; Causton *et al.*, 2001).

Transcriptional response to glycerol stress is driven mainly by duplicated genes

Determining whether or not there is potential to adapt to novel environments through rapid evolution of regulatory programs is an important aim in evolutionary biology, yet such an aim has remained obscure owing to the difficulty of mapping phenotypes to genotypes or assigning transcriptional changes to phenotypic variations. A clear pattern observed in this study links transcriptional variation to phenotypic adaptation to glycerol, namely that transcriptionally altered genes in the presence of glycerol are enriched for duplicated genes. The classic theory of evolution by gene duplication states that after the duplication of a gene, one of the copies explores novel genotypes freed from selection constraints because its sister copy performs the ancestral well-adapted function (Ohno, 1970; 1999; Lynch and Conery, 2000; Conant and Wolfe, 2006). This genotypic exploration affects both the functional and regulatory features of genes. We hypothesize that increasing the regulatory plasticity after gene duplication is a more likely scenario than increasing the functional plasticity, as the effects of the former are more likely to be the subject of rapid selection in changing environments while the latter requires longer evolutionary times to be selected for (Keane *et al.*, 2014; Fares, 2015). Eventually, the expression plasticity may lead to functional plasticity because the rate of evolution of a gene is strongly determined by its expression level, a link that has been observed in all organisms examined so far from viruses to mammals (Krylov *et al.*, 2003; Rocha and Danchin, 2004; Drummond *et al.*, 2005; 2006; Drummond and Wilke, 2008; Pagan *et al.*, 2012; Zhang and Yang, 2015). Testament to this is the rapid expression divergence between the copies of a duplicated gene (Blanc and Wolfe, 2004; Li *et al.*, 2005; Conant and Wolfe, 2006; Thompson *et al.*, 2013). The selective enrichment of glycerol-responsive duplicate genes raises the possibility that duplicated genes have led to major specializations in the regulatory response to changing environments, perhaps through the acquisition of novel functions or novel interactions in the cell (Fares *et al.*, 2013).

Implications for S. cerevisiae *ecology, and concluding remarks*

Our data reveal a strong link between the regulatory re-programming of the cell and the environmental change in glycerol concentration. These data uncover a rapid genome-wide de-regulation of genes involved in fundamental metabolic processes in the cell, with the up-regulation of genes that allow a rapid shift from a fermentative to a respiratory metabolism as well as genes encoding membrane transporters of solutes and ions. In contrast, down-regulated genes in cells exposed to

glycerol-induced stress affect energetic-costly processes, such as translation and transcription. We also identified a fine-tuned reprogramming of the transcriptome as an adaptive response to 10-day exposure to glycerol-induced stress. Most of the transcriptomic responses to glycerol are driven by duplicated genes, revealing an unprecedented fundamental role of these genes in the evolution of adaptive responses to environmental perturbations.

In natural habitats of *S. cerevisiae* (which may include soils, plant surfaces, animal hosts, saline environments or sugar-rich milieu), the yeast cell—like those of many microbes—is likely to experience multiple, concomitant stresses. In high-sugar substrates of >0.900 water activity, where *S. cerevisiae* can thrive, some of these stresses are self-imposed (ethanol stress, acetaldehyde stress, and organic-acid stress) and others are not (sugar-induced reduction of water activity, antimicrobials produced by competitors, sub- or supra-optimal temperatures, etc.). The interplay between such parameters, community dynamics, responses and adaptations of *S. cerevisiae*, and its ability to grow and/or remain metabolically active have been the focus on recent studies by Cray *et al.* (2013a; 2015). The findings of the current study demonstrate that glycerol, which is produced by the yeast cell as a stress protectant, can also cause collateral damage, and also shows how *S. cerevisiae* can response physiologically and can also evolutionarily adapt to this additional stress burden. A series of intriguing scientific questions remain unanswered: What is the role of duplicated genes in allowing quick switches of regulatory programs in *S. cerevisiae* subjected to environmental perturbations? What is the ecological cost of a rapid adaptation to glycerol stress? What is the impact of the regulatory re-programming of *S. cerevisiae* under stress on the evolution of protein-coding genes and the origin of novel functions? How plastic is the *S. cerevisiae* transcriptome to fluctuating environments?

Experimental procedures

Strains, culture media and culture conditions. The *S. cerevisiae* Y06240 haploid *msh2* deletion strain (BY4741; *Mata*; *his3D1*; *leud2DO*; *met15DO*; *ura3DO*; *msh2::kanMX4*) (Fares *et al.*, 2013) was used as study subject. The experimental evolution procedure is summarized in Figure 1. Briefly, from the glycerol stock, a colony was selected (t_{-1}) to start a liquid culture (t_0), from which a set of five populations were established in rich media (YPD: 2% (w/v) bacto peptone, 1% (w/v) yeast extract, 2% (w/v) dextrose; supplemented with kanamycin) and evolved through daily bottlenecks (1%) for 100 days. Populations were allowed to grow at 28°C for 24 h, each in 5 ml of media in 50 ml Corning tubes, and subjected to serial passages (1%) in a daily manner. Each 10 passages (10 days), a glycerol stock (25%) of the population was stored creating a fossil record. From passage 100 (t_{100}), populations were split in two, one half grown in normal rich media (YPD) as control populations, and the remaining half in media containing 3% (w/v) glycerol (YPG: 0.41 M glycerol, 2% (w/v) bacto peptone, 1% (w/v) yeast extract; also supplemented with kanamycin) as sole carbon source. Split populations were subjected to serial passages (at 10% dilution of original population; bottleneck of 10% of population) in a daily manner for an additional 10 days (t_{110}), as described above.

Quantitation of water activity. Water activity of culture media was quantified at 28°C, using a Novasina Humidat-IC-II water-activity machine (Novasina, Pfäffikon, Switzerland) as described by Stevenson *et al.*, in press-c. A number of precautions were taken to ensure that the volatility of glycerol did not interfere with quantification, and to minimize any other potential error to maintain a level of accuracy consistent with the sensitivity of the microbial cell (Hallsworth and Nomura, 1999; Stevenson *et al.*, 2015a,b). Calibration was carried out between each measurement of culture medium, using saturated salt solutions of known water activity (Winston and Bates, 1960). The water activity of each medium type was determined three times, and variation was within 0.001.

Determination of growth rates under YPD as well as glycerol-induced stress. Growth parameters were evaluated using the BioScreen C plate-reader system (Oy Growth Curves Ab Ldt., Hensinki, Finland) at t_0, t_{100} and t_{110}. Each time point was pre-cultured overnight at 28°C in 5 ml of the corresponding media. Precultures were used to inoculate 200 μl of fresh media (YPD or YPG) to an initial OD_{595} of 0.06 to 0.07, distributed in 100-well Honeycomb plates (Oy Growth Curves). Each time point was tested at least in triplicate in the same plate with the two media. Each experimental run was conducted with negative (blank fresh media) and positive (ancestral, t_0, lines) controls. Plates were incubated at 28°C, with continuous shaking (medium force) in the Bioscreen C. Growth was monitored for a period of 92 to 120 h taking OD_{595} measurement each 15 min. Maximum growth at exponential phase (μ_{max}) and lagging time (Lagt) were determined with GrowthRates software version 2.1 (Baty and Delignette-Muller, 2004; Hall *et al.*, 2014) (http://bellingham researchinstitute.com/software/index.html) across replicated cultures. Growth curves were constructed by plotting corrected OD versus time. OD was corrected for linearity by applying the following formula after blank subtraction to each time point: $OD_{cor} = OD_{obs} + 0.449 \times OD_{obs}^2 + 0.191 \times OD_{obs}^3$ (Warringer and Blomberg, 2003).

RNA extractions and transcriptomic analyses. The transcriptomic profiling was performed in the t_0, t_{100} and t_{110}, with three technical replicates for biological stress

condition (3% glycerol (YPG)) in comparison with the normal growth condition (YPD media). Total RNA extractions were performed with RNeasy kit (Qiagen) following manufacturer instructions. Ribosomal RNA was removed by using Ribo-Zero Gold rRNA removal yeast (Illumina) depletion kit. Stranded RNA libraries were constructed using TruSeq stranded mRNA (Illumina) from oligo-dT captured mRNAs from depleted samples. Libraries were run in NextSeq 500 (Illumina) at 75nt single read by using High Output 75 cycles kit v2.0 (Illumina). RNA libraries were sequenced at Genomic core facility at Servicio Central de Soporte a la Investigación Experimental (SCSIE) from University of Valencia, Spain.

The treatment of the RNA libraries was done following a previous study in which different methods of differential expression analyses were compared (Zhang *et al.*, 2014). Raw reads were analyzed using FastQC report and cleaned with CutAdapt as implemented in RobiNA software package v 1.2.4 (Lohse *et al.*, 2012). Low quality reads were filtered and trimmed (Phred score inferior to 20 and size less than 40 nt were discarded). Reads were then aligned with Bowtie (up to two mismatches accepted) to the reference transcriptome (PRJNA290217) from the reference S288c strain. Statistical assessment of differential gene expression was done with edgeR (Robinson *et al.*, 2010) and with DESeq (Anders and Huber, 2010) as implemented in RobiNA. A previous study compared the different expression analysis methods, concluding that edgeR and DeSeq were the best performing methods when the objective is to analyze differential expression (Zhang *et al.*, 2014). Comparison of logarithmic fold change of our expression data between edgeR and DESeq provided very strong correlation (Spearman correlation coefficient: $\rho = 0.995$, $P < 2.2 \times 10^{-16}$, Figure 1 of Supporting Information File S8). Significant expression changes were identified using a false discovery rate (FDR < 0.05). These results indicate that our quantification of expression data is robust to the method used. RNA raw reads are available from the Sequence Read Archive with accession number SRP074821.

Genes with significantly higher reads per billion (RPKM) under YPG than YPD (with a false discovery rate for the fold change of expression FDR < 0.05) were considered transcriptionally altered in YPG. Because RNA molecules can undergo degradation before being translated, we examined the correlation between the RPKMs of our transcriptomic analyses and those obtained by other groups using ribosomal profiling, a technique that measures ribosome occupancy and translation genome-wide (Albert *et al.*, 2014). Despite the large number of data available for both of the methods ($N = 4682$), we found a very strong and significant correlation between the counts of both of the methods (Spearman's correlation: $\rho = 0.77$, $P < 2.2 \times 10^{-16}$). Hence, highly transcribed genes are also highly translated and vice versa. Our data, therefore, are indicative of the levels of gene expression. RNA-seq technology is sensitive to biases in expression detection, such that often it becomes difficult distinguishing genes with very low read counts from background noise. However, it has recently been shown that RPKM metric is robust to the low expression filtering strategies (Lin *et al.*, 2016). Indeed, we could identify differentially expressed genes at read counts as low read as RPKM = 0.001 (logRPKM = −2.8).

Identification of duplicated genes. Paralogous pairs of duplicated genes were identified as the resulting best reciprocal hits from all-against-all BLAST searches using BLASTP with an *E*-value cutoff of 1E-5 and a 50 bit score(-Altschul *et al.*, 1997). Paralogs were then divided into two groups according to the mechanism of their origin: WGDs and SSDs. WGDs are those extracted from the reconciled list provided by the YGOB (Yeast Gene Order Browser, http://wolfe.gen.tcd.ie//ygob(Byrne and Wolfe, 2005)) (555 pairs of genes), and these were not subjected to subsequent SSD. All other paralogs were considered to belong to the category of SSDs (560 pairs of genes). The duplicates used in this study have been estimated to have their origin on the time point of the whole genome duplication that took place 100 MYA (Wolfe and Shields, 1997). Also, in this study we have used the SSDs that exhibit similar distribution of synonymous substitutions as those of WGDs, so roughly belonging to the same age (Fares *et al.*, 2013; Keane *et al.*, 2014).

Gene ontology—Functional categories classification and visualization. For each differential expressed gene list, gene ontology (GO) term was associated with Gene Ontology Term Finder as implemented in the Saccharomyces Genome Database (http://www.yeastgenome.org/cgi-bin/GO/goTermFinder.pl), which also include a GO enrichment analysis, with a *P* value cutoff of <0.01 (Supporting Information Tables S1 to S17). A semantic similarity score, simRel (Schlicker *et al.*, 2006) was used to summarize and remove redundant GO terms in the list, as implemented in REVIGO software with medium (0.7) allowed similarity, using *S. cerevisiae* GO term database and the *P* values from enrichment analysis (Supek *et al.*, 2011).

Software. Calculations and statistics were performed using MS Excel and R 3.2.1. (R Core team 2013), except as indicated above for differential expression, and growth parameters.

Measure of metabolic distance. To calculate the metabolic distance between the three *S. cerevisiae* populations (t_0, t_{100} and t_{110}) we compared the list of GO process terms enriched for transcriptionally altered genes between two populations (i and j) by calculating the number of shared process terms ($SP_{i,j}$) and the number of enriched terms for transcriptionally altered genes only in one of the populations but not the other (P_i and P_j). The metabolic distance

between the two populations ($MD_{i,j}$) was calculated as: $MD_{i,j} = 1 - \frac{SP_{i,j}}{Min[P_i, P_j]}$.

Here $Min[P_i, P_j]$ is the number of cellular processes enriched for transcriptionally altered genes for the population with minimum number of such processes. Following this equation, metabolic distance varies between 0, when the transcriptionally altered genes from both populations affect the same processes (i.e., $P_{i,j} = Min[P_i, P_j]$), and 1 when there is no overlap in the process terms.

Acknowledgements

This study was supported by grants from the Spanish Ministry of Economy and Competitiveness (Refs: BFU2012-36346, BFU2015-66073-P) and a grant from the local government Consellería de Educación, Investigación, Cultura y Deporte, Generalitat Valenciana (Ref: ACOMP/2012/098) to M.A.F. The authors have no conflict of interest to declare.

References

Albert, F.W., Muzzey, D., Weissman, J.S., and Kruglyak, L. (2014) Genetic influences on translation in yeast. *PLoS Genet* **10:** e1004692.

Alemohammad, M.M., and Knowles, C.J. (1974) Osmotically induced volume and turbidity changes of *Escherichia coli* due to salts, sucrose and glycerol, with particular reference to the rapid permeation of glycerol into the cell. *J Gen Microbiol* **82:** 125–142.

Alepuz, P.M., de Nadal, E., Zapater, M., Ammerer, G., and Posas, F. (2003) Osmostress-induced transcription by Hot1 depends on a Hog1-mediated recruitment of the RNA Pol II. *EMBO J* **22:** 2433–2442.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25:** 3389–3402.

Anders, S., and Huber, W. (2010) Differential expression analysis for sequence count data. *Genome Biol* **11:** R106.

Aramburu, J., Ortells, M.C., Tejedor, S., Buxade, M., and Lopez-Rodriguez, C. (2014) Transcriptional regulation of the stress response by mTOR. *Sci Signal* **7:** re2.

Bai, C., Tesker, M., and Engelberg, D. (2015) The yeast Hot1 transcription factor is critical for activating a single target gene, STL1. *Mol Biol Cell* **26:** 2357–2374.

Ball, P., and Hallsworth, J.E. (2015) Water structure and chaotropicity: their uses, abuses and biological implications. *Phys Chem Chem Phys* **17:** 8297–8305.

Basso, L.C., de Amorim, H.V., de Oliveira, A.J., and Lopes, M.L. (2008) Yeast selection for fuel ethanol production in Brazil. *FEMS Yeast Res* **8:** 1155–1163.

Baty, F., and Delignette-Muller, M.L. (2004) Estimating the bacterial lag time: which model, which precision?. *Int J Food Microbiol* **91:** 261–277.

Bhaganna, P., Volkers, R.J., Bell, A.N., Kluge, K., Timson, D.J., McGrath, J.W., *et al.* (2010) Hydrophobic substances induce water stress in microbial cells. *Microb Biotechnol* **3:** 701–716.

Bjornsti, M.A., and Houghton, P.J. (2004) The TOR pathway: A target for cancer therapy. *Nat Rev Cancer* **4:** 335–348.

Blanc, G., and Wolfe, K.H. (2004) Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution. *Plant Cell* **16:** 1679–1691.

Botha, A. (2011) The importance and ecology of yeasts in soil. *Soil Biol Biochem* **43:** 8.

Brauer, M.J., Huttenhower, C., Airoldi, E.M., Rosenstein, R., Matese, J.C., Gresham, D., *et al.* (2008) Coordination of growth rate, cell cycle, stress response, and metabolic activity in yeast. *Mol Biol Cell* **19:** 352–367.

Broach, J.R. (2012) Nutritional control of growth and development in yeast. *Genetics* **192:** 73–105.

Byrne, K.P., and Wolfe, K.H. (2005) The yeast gene order browser: Combining curated homology and syntenic context reveals gene fate in polyploid species. *Genome Res* **15:** 1456–1461.

Cardenas, M.E., Cutler, N.S., Lorenz, M.C., Di Como, C.J., and Heitman, J. (1999) The TOR signaling cascade regulates gene expression in response to nutrients. *Genes Dev* **13:** 3271–3279.

Carretero-Paulet, L., and Fares, M.A. (2012) Evolutionary dynamics and functional specialization of plant paralogs formed by whole and small-scale genome duplications. *Mol Biol Evol* **29:** 3541–3551.

Carroll, A.S., and O'shea, E.K. (2002) Pho85 and signaling environmental conditions. *Trends Biochem Sci* **27:** 87–93.

Castrillo, J.I., Zeef, L.A., Hoyle, D.C., Zhang, N., Hayes, A., Gardner, D.C., *et al.* (2007) Growth control of the eukaryote cell: A systems biology study in yeast. *J Biol* **6:** 4.

Causton, H.C., Ren, B., Koh, S.S., Harbison, C.T., Kanin, E., Jennings, E.G., *et al.* (2001) Remodeling of yeast genome expression in response to environmental changes. *Mol Biol Cell* **12:** 323–337.

Conant, G.C., and Wolfe, K.H. (2006) Functional partitioning of yeast co-expression networks after genome duplication. *PLoS Biol* **4:** e109.

Cormier, L., Barbey, R., and Kuras, L. (2010) Transcriptional plasticity through differential assembly of a multiprotein activation complex. *Nucleic Acids Res* **38:** 4998–5014.

Cray, J.A., Bell, A.N., Bhaganna, P., Mswaka, A.Y., Timson, D.J., and Hallsworth, J.E. (2013a) The biology of habitat dominance; can microbes behave as weeds? *Microb Biotechnol* **6:** 453–492.

Cray, J.A., Russell, J.T., Timson, D.J., Singhal, R.S., and Hallsworth, J.E. (2013b) A universal measure of chaotropicity and kosmotropicity. *Environ Microbiol* **15:** 287–296.

Cray, J.A., Stevenson, A., Ball, P., Bankar, S.B., Eleutherio, E.C., Ezeji, T.C., *et al.* (2015) Chaotropicity: a key factor in product tolerance of biofuel-producing microorganisms. *Curr Opin Biotechnol* **33:** 228–259.

Crozet, P., Margalha, L., Confraria, A., Rodrigues, A., Martinho, C., Adamo, M., *et al.* (2014) Mechanisms of regulation of SNF1/AMPK/SnRK1 protein kinases. *Front Plant Sci* **5:** 190.

Cui, L., Wall, P.K., Leebens-Mack, J.H., Lindsay, B.G., Soltis, D.E., Doyle, J.J., *et al.* (2006) Widespread genome duplications throughout the history of flowering plants. *Genome Res* **16:** 738–749.

de Lima Alves, F., Stevenson, A., Baxter, E., Gillion, J.L., Hejazi, F., Hayes, S., *et al.* (2015) Concomitant osmotic and chaotropicity-induced stresses in *Aspergillus wentii*: compatible solutes determine the biotic window. *Curr Genet* **61:** 457–477.

De Nadal, E., Zapater, M., Alepuz, P.M., Sumoy, L., Mas, G., and Posas, F. (2004) The MAPK Hog1 recruits Rpd3 histone deacetylase to activate osmoresponsive genes. *Nature* **427:** 370–374.

DeRisi, J.L., Iyer, V.R., and Brown, P.O. (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278:** 680–686.

Drummond, D.A., and Wilke, C.O. (2008) Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* **134:** 341–352.

Drummond, D.A., Raval, A., and Wilke, C.O. (2006) A single determinant dominates the rate of yeast protein evolution. *Mol Biol Evol* **23:** 327–337.

Drummond, D.A., Bloom, J.D., Adami, C., Wilke, C.O., and Arnold, F.H. (2005) Why highly expressed proteins evolve slowly. *Proc Natl Acad Sci USA* **102:** 14338–14343.

Duskova, M., Ferreira, C., Lucas, C., and Sychrova, H. (2015) Two glycerol uptake systems contribute to the high osmotolerance of *Zygosaccharomyces rouxii*. *Mol Microbiol* **97:** 541–559.

Emanuelle, S., Doblin, M.S., Stapleton, D.I., Bacic, A., and Gooley, P.R. (2016) Molecular insights into the enigmatic metabolic regulator, SnRK1. *Trends Plant Sci* **21:** 341–353.

Fares, M.A. (2015) The origins of mutational robustness. *Trends Genet* **31:** 373–381.

Fares, M.A., Keane, O.M., Toft, C., Carretero-Paulet, L., and Jones, G.W. (2013) The roles of whole-genome and small-scale duplications in the functional specialization of *Saccharomyces cerevisiae* genes. *PLoS Genet* **9:** e1003176.

Ferea, T.L., Botstein, D., Brown, P.O., and Rosenzweig, R.F. (1999) Systematic changes in gene expression patterns following adaptive evolution in yeast. *Proc Natl Acad Sci USA* **96:** 9721–9726.

Gasch, A.P., Spellman, P.T., Kao, C.M., Carmel-Harel, O., Eisen, M.B., Storz, G., *et al.* (2000) Genomic expression programs in the response of yeast cells to environmental changes. *Mol Biol Cell* **11:** 4241–4257.

Ghillebert, R., Swinnen, E., Wen, J., Vandesteene, L., Ramon, M., Norga, K., *et al.* (2011) The AMPK/SNF1/SnRK1 fuel gauge and energy regulator: Structure, function and regulation. *FEBS J* **278:** 3978–3990.

Giaever, G., and Nislow, C. (2014) The yeast deletion collection: A decade of functional genomics. *Genetics* **197:** 451–465.

Giaever, G., Chu, A.M., Ni, L., Connelly, C., Riles, L., Veronneau, S., *et al.* (2002) Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* **418:** 387–391.

Hall, B.G., Acar, H., Nandipati, A., and Barlow, M. (2014) Growth rates made easy. *Mol Biol Evol* **31:** 232–238.

Hallsworth, J.E. (1998) Ethanol-induced water stress in yeast. *J Ferment Bioeng* **85:** 125–137.

Hallsworth, J.E., and Nomura, Y. (1999) A simple method to determine the water activity of ethanol-containing samples. *Biotechnol Bioeng* **62:** 242–245.

Hallsworth, J.E., Yakimov, M.M., Golyshin, P.N., Gillion, J.L., D'auria, G., de Lima Alves, F., *et al.* (2007) Limits of life in $MgCl_2$-containing environments: chaotropicity defines the window. *Environ Microbiol* **9:** 801–813.

Hoegg, S., Brinkmann, H., Taylor, J.S., and Meyer, A. (2004) Phylogenetic timing of the fish-specific genome duplication correlates with the diversification of teleost fish. *J Mol Evol* **59:** 190–203.

Hohmann, S. (2015) An integrated view on a eukaryotic osmoregulation system. *Curr Genet* **61:** 373–382.

Hohmann, S., Krantz, M., and Nordlander, B. (2007) Yeast osmoregulation. *Methods Enzymol* **428:** 29–45.

Holub, E.B. (2001) The arms race is ancient history in Arabidopsis, the wildflower. *Nat Rev Genet* **2:** 516–527.

Huang, D., Friesen, H., and Andrews, B. (2007) Pho85, a multifunctional cyclin-dependent protein kinase in budding yeast. *Mol Microbiol* **66:** 303–314.

Hubmann, G., Guillouet, S., and Nevoigt, E. (2011) Gpd1 and Gpd2 fine-tuning for sustainable reduction of glycerol formation in *Saccharomyces cerevisiae*. *Appl Environ Microbiol* **77:** 5857–5867.

Ideker, T., Thorsson, V., Ranish, J.A., Christmas, R., Buhler, J., Eng, J.K., *et al.* (2001) Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* **292:** 929–934.

Keane, O.M., Toft, C., Carretero-Paulet, L., Jones, G.W., and Fares, M.A. (2014) Preservation of genetic and regulatory robustness in ancient gene duplicates of *Saccharomyces cerevisiae*. *Genome Res* **24:** 1830–1841.

Kim, S., Yoo, M.J., Albert, V.A., Farris, J.S., Soltis, P.S., and Soltis, D.E. (2004) Phylogeny and diversification of B-function MADS-box genes in angiosperms: Evolutionary and functional implications of a 260-million-year-old duplication. *Am J Bot* **91:** 2102–2118.

Kiyosawa, K. (1991) Volumetric properties of polyols (ethylene glycol, glycerol, meso-erythritol, xylitol and mannitol) in relation to their membrane permeability: Group additivity and estimation of the maximum radius of their molecules. *Biochim Biophys Acta* **1064:** 251–255.

Krylov, D.M., Wolf, Y.I., Rogozin, I.B., and Koonin, E.V. (2003) Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. *Genome Res* **13:** 2229–2235.

Lespinet, O., Wolf, Y.I., Koonin, E.V., and Aravind, L. (2002) The role of lineage-specific gene family expansion in the evolution of eukaryotes. *Genome Res* **12:** 1048–1059.

Li, W.H., Yang, J., and Gu, X. (2005) Expression divergence between duplicate genes. *Trends Genet* **21:** 602–607.

Lievens, B., Hallsworth, J.E., Pozo, M.I., Belgacem, Z.B., Stevenson, A., Willems, K.A., and Jacquemyn, H. (2015) Microbiology of sugar-rich environments: diversity, ecology and system constraints. *Environ Microbiol* **17:** 278–298.

Lin, Y., Golovnina, K., Chen, Z.X., Lee, H.N., Negron, Y.L., Sultana, H., *et al.* (2016) Comparison of normalization and differential expression analyses using RNA-Seq data from 726 individual *Drosophila melanogaster*. *BMC Genomics* **17:** 28.

Lohse, M., Bolger, A.M., Nagel, A., Fernie, A.R., Lunn, J.E., Stitt, M., and Usadel, B. (2012) RobiNA: a user-friendly, integrated software solution for RNA-Seq-based transcriptomics. *Nucleic Acids Res* **40:** W622–W627.

Lynch, M., and Conery, J.S. (2000) The evolutionary fate and consequences of duplicate genes. *Science* **290:** 1151–1155.

Marcet-Houben, M., and Gabaldon, T. (2015) Beyond the whole-genome duplication: phylogenetic evidence for an ancient interspecies hybridization in the baker's yeast lineage. *PLoS Biol* **13:** e1002220.

© 2016 Society for Applied Microbiology and John Wiley & Sons Ltd, *Environmental Microbiology*, **19,** 990–1007

Martin, D.E., and Hall, M.N. (2005) The expanding TOR signaling network. *Curr Opin Cell Biol* **17:** 158–166.
Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., and Snyder, M. (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320:** 1344–1349.
Nookaew, I., Papini, M., Pornputtapong, N., Scalcinati, G., Fagerberg, L., Uhlen, M., and Nielsen, J. (2012) A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in *Saccharomyces cerevisiae*. *Nucleic Acids Res* **40:** 10084–10097.
Ohno, S. (1970) *Evolution by Gene Duplication*. Springer Verlag.
Ohno, S. (1999) Gene duplication and the uniqueness of vertebrate genomes circa 1970-1999. *Semin Cell Dev Biol* **10:** 517–522.
Otto, S.P., and Whitton, J. (2000) Polyploid incidence and evolution. *Annu Rev Genet* **34:** 401–437.
Pagan, I., Holmes, E.C., and Simon-Loriere, E. (2012) Level of gene expression is a major determinant of protein evolution in the viral order *Mononegavirales*. *J Virol* **86:** 5253–5263.
Regenberg, B., Grotkjaer, T., Winther, O., Fausboll, A., Akesson, M., Bro, C., *et al.* (2006) Growth-rate regulated genes have profound impact on interpretation of transcriptome profiling in *Saccharomyces cerevisiae*. *Genome Biol* **7:** R107.
Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010) edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26:** 139–140.
Rocha, E.P., and Danchin, A. (2004) An analysis of determinants of amino acids substitution rates in bacterial proteins. *Mol Biol Evol* **21:** 108–116.
Schlicker, A., Domingues, F.S., Rahnenfuhrer, J., and Lengauer, T. (2006) A new measure for functional similarity of gene products based on gene ontology. *BMC Bioinformatics* **7:** 302.
Schuller, H.J. (2003) Transcriptional control of nonfermentative metabolism in the yeast *Saccharomyces cerevisiae*. *Curr Genet* **43:** 139–160.
Stern, S., Dror, T., Stolovicki, E., Brenner, N., and Braun, E. (2007) Genome-wide transcriptional plasticity underlies cellular adaptation to novel challenge. *Mol Syst Biol* **3:** 106.
Stevenson, A., Burkhardt, J., Cockell, C.S., Cray, J.A., Dijksterhuis, J., Fox-Powell, M., *et al.* (2015a) Multiplication of microbes below 0.690 water activity: Implications for terrestrial and extraterrestrial life. *Environ Microbiol* **17:** 257–277.
Stevenson, A., Cray, J.A., Williams, J.P., Santos, R., Sahay, R., Neuenkirchen, N., *et al.* (2015b) Is there a common water-activity limit for the three domains of life? *ISME J* **9:** 1333–1351.
Stevenson, A., Hamill, P.G., Dijksterhuis, J., and Hallsworth, J.E. (in press-c) Water-, pH- and temperature relations of germination for the extreme xerophiles *Xeromyces bisporus* (FRR 0025), *Aspergillus penicillioides* (JH06THJ), and *Eurotium halophilicum* (FRR 2471). *Microbial Biotechnol* DOI: 10.1111/1751-7915.12406.
Supek, F., Bosnjak, M., Skunca, N., and Smuc, T. (2011) REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS One* **6:** e21800.
Taymaz-Nikerel, H., and Lara, A.R. (2016) Editorial: Quantitative systems biology for engineering organisms and pathways. *Front Bioeng Biotechnol* **4:** 22.
Taymaz-Nikerel, H., Cankorur-Cetinkaya, A., and Kirdar, B. (2016) Genome-wide transcriptional response of *Saccharomyces cerevisiae* to stress-induced perturbations. *Front Bioeng Biotechnol* **4:** 17.
Thompson, D.A., Roy, S., Chan, M., Styczynsky, M.P., Pfiffner, J., French, C., *et al.* (2013) Evolutionary principles of modular gene regulation in yeasts. *Elife* **2:** e00603.
Tulha, J., Lima, A., Lucas, C., and Ferreira, C. (2010) *Saccharomyces cerevisiae* glycerol/H+ symporter Stl1p is essential for cold/near-freeze and freeze stress adaptation. A simple recipe with high biotechnological potential is given. *Microb Cell Fact* **9:** 82.
Turcotte, B., Liang, X.B., Robert, F., and Soontorngun, N. (2010) Transcriptional regulation of nonfermentable carbon utilization in budding yeast. *FEMS Yeast Res* **10:** 2–13.
van Dijk, E.L., Chen, C.L., D'aubenton-Carafa, Y., Gourvennec, S., Kwapisz, M., Roche, V., *et al.* (2011) XUTs are a class of Xrn1-sensitive antisense regulatory noncoding RNA in yeast. *Nature* **475:** 114–117.
Vilhelmsson, O., and Miller, K.J. (2002) Humectant permeability influences growth and compatible solute uptake by *Staphylococcus aureus* subjected to osmotic stress. *J Food Prot* **65:** 1008–1015.
Wagner, A. (2015) Causal drift, robust signaling, and complex disease. *PLoS One* **10:** e0118413.
Warringer, J., and Blomberg, A. (2003) Automated screening in environmental arrays allows analysis of quantitative phenotypic profiles in *Saccharomyces cerevisiae*. *Yeast* **20:** 53–67.
Wendel, J.F. (2000) Genome evolution in polyploids. *Plant Mol Biol* **42:** 225–249.
Williams, J.P., and Hallsworth, J.E. (2009) Limits of life in hostile environments: No barriers to biosphere function?. *Environ Microbiol* **11:** 3292–3308.
Winston, P.W., and Bates, D.H. (1960) Saturated solutions for the control of humidity in biological research. *Ecology* **41:** 6.
Wolfe, K.H., and Shields, D.C. (1997) Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* **387:** 708–713.
Yadav, K.K., Singh, N., and Rajasekharan, R. (2016) Responses to phosphate deprivation in yeast cells. *Curr Genet* **62:** 301–307.
Yakimov, M.M., La Cono, V., Spada, G.L., Bortoluzzi, G., Messina, E., Smedile, F., *et al.* (2015) Microbial community of the deep-sea brine Lake Kryos seawater-brine interface is active below the chaotropicity limit of life as revealed by recovery of mRNA. *Environ Microbiol* **17:** 364–382.
Yosef, N., and Regev, A. (2011) Impulse control: Temporal dynamics in gene transcription. *Cell* **144:** 886–896.
Zhang, J., and Yang, J.R. (2015) Determinants of the rate of protein sequence evolution. *Nat Rev Genet* **16:** 409–420.
Zhang, Z.H., Jhaveri, D.J., Marshall, V.M., Bauer, D.C., Edson, J., Narayanan, R.K., *et al.* (2014) A comparative study of techniques for differential expression analysis on RNA-Seq data. *PLoS One* **9:** e103207.

Supporting information

Additional supporting information may be found in the online version of this article at the publisher's web-site

AMERICAN SOCIETY FOR MICROBIOLOGY | mSystems®

RESEARCH ARTICLE
Ecological and Evolutionary Science

Transcriptional Rewiring, Adaptation, and the Role of Gene Duplication in the Metabolism of Ethanol of *Saccharomyces cerevisiae*

Beatriz Sabater-Muñoz,[a,b] Florian Mattenberger,[a*] Mario A. Fares,[a,b†] Christina Toft[a,c]

[a]Department of Abiotic Stress, Instituto de Biología Molecular y Celular de Plantas (IBMCP), Consejo Superior de Investigaciones Científicas-Universidad Politécnica (CSIC-UPV), Valencia, Spain
[b]Smurfit Institute of Genetics, Department of Genetics, University of Dublin, Trinity College, Dublin, Ireland
[c]Program for Systems Biology of Molecular Interactions and Regulation, Institute for Integrative Systems Biology (I²SysBio, CSIC-UV), Valencia, Spain

ABSTRACT Ethanol is the main by-product of yeast sugar fermentation that affects microbial growth parameters, being considered a dual molecule, a nutrient and a stressor. Previous works demonstrated that the budding yeast arose after an ancient hybridization process resulted in a tier of duplicated genes within its genome, many of them with implications in this ethanol "produce-accumulate-consume" strategy. The evolutionary link between ethanol production, consumption, and tolerance versus ploidy and stability of the hybrids is an ongoing debatable issue. The implication of ancestral duplicates in this metabolic rewiring, and how these duplicates differ transcriptionally, remains unsolved. Here, we study the transcriptomic adaptive signatures to ethanol as a nonfermentative carbon source to sustain clonal yeast growth by experimental evolution, emphasizing the role of duplicated genes in the adaptive process. As expected, ethanol was able to sustain growth but at a lower rate than glucose. Our results demonstrate that in asexual populations a complete transcriptomic rewiring was produced, strikingly by downregulation of duplicated genes, mainly whole-genome duplicates, whereas small-scale duplicates exhibited significant transcriptional divergence between copies. Overall, this study contributes to the understanding of evolution after gene duplication, linking transcriptional divergence with duplicates' fate in a multigene trait as ethanol tolerance.

IMPORTANCE Gene duplication events have been related with increasing biological complexity through the tree of life, but also with illnesses, including cancer. Early evolutionary theories indicated that duplicated genes could explore alternative functions due to relaxation of selective constraints in one of the copies, as the other remains as ancestral-function backup. In unicellular eukaryotes like yeasts, it has been demonstrated that the fate and persistence of duplicates depend on duplication mechanism (whole-genome or small-scale events), shaping their actual genomes. Although it has been shown that small-scale duplicates tend to innovate and whole-genome duplicates specialize in ancestral functions, the implication of duplicates' transcriptional plasticity and transcriptional divergence on environmental and metabolic responses remains largely obscure. Here, by experimental adaptive evolution, we show that *Saccharomyces cerevisiae* is able to respond to metabolic stress (ethanol as nonfermentative carbon source) due to the persistence of duplicated genes. These duplicates respond by transcriptional rewiring, depending on their transcriptional background. Our results shed light on the mechanisms that determine the role of duplicates, and on their evolvability.

KEYWORDS RNAseq, adaptive laboratory experimental evolution, clonal populations, transcriptional divergence, ethanol stress

Citation Sabater-Muñoz B, Mattenberger F, Fares MA, Toft C. 2020. Transcriptional rewiring, adaptation, and the role of gene duplication in the metabolism of ethanol of *Saccharomyces cerevisiae*. mSystems 5:e00416-20. https://doi.org/10.1128/mSystems.00416-20.

Editor Paola Flórez de Sessions, Oxford Nanopore Technologies

Address correspondence to Christina Toft, christina.toft@uv.es.

* Present address: Florian Mattenberger, Program for Pathogen Systems Biology, Institute for Integrative Systems Biology (I²SysBio, CSIC-UV), Valencia, Spain.

† Deceased.

Received 11 May 2020
Accepted 25 July 2020
Published 11 August 2020

Sensing and responding to the environment are central parts of metabolism of almost all unicellular organisms. During evolution, some budding yeasts (Saccharomycotina) faced a new source of carbon (sugars) in a new niche (nectar or fruits from the recently emerged [100 million years ago {MYA}] Angiosperms), a fact that has been postulated at the origin of fermentative metabolism with ethanol as main end product. Behind this biological innovation has been unveiled gene duplication at two scales, whole-genome (WGD) and small-scale (SSD) duplication, and genome shrinkage after it, as evolutionarily driving genomic changes (reviewed in references 1 to 3). The baker's yeast *Saccharomyces cerevisiae* is one of the most biotechnologically important species, being able to tolerate higher ethanol levels during fermentation than any other microbe (4–6).

Under sugar scarcity, yeast can switch from fermentative to respiratory metabolism using ethanol and glycerol as nonfermentative carbon sources to support growth (7, 8). This ethanol "make-accumulate-consume" strategy (Crabtree effect) has been partially linked to the yeast evolutionary origin history. However, ethanol in particular endangers the yeast metabolic activity, survival, cell morphology, growth ability, and biomass production. Ethanol also exhibits a general cell toxicity that yeasts used to control competitors' growth. This duality (nutrient and stressor) generates great concerns in the biotechnological industries (by its applications) and in the scientific community (by its molecular basis), highlighting the importance of systems biology studies (reviewed in references 9 to 12).

Experimental evolution, in particular with *Escherichia coli* and *S. cerevisiae*, has been of unprecedented relevance to unveil evolutionary pathways underlying the origin of adaptations, including as examples the adaptation of *E. coli* to citrate in the known Lenski evolution experiment (13–17) and heat stress, nutrient limitations, antibiotic treatment, or tolerance to glycerol in *S. cerevisiae* (18–24). Its use to understand the adaptation to ethanol has been addressed in only a few studies, while using ethanol as additional carbon source (4, 5, 25). Indeed, only one work revealed the genomic dynamics including point mutations, copy number variation (gene duplication), ploidy changes, and clonal interference mix in a complex evolutionary pathway that increases tolerance to ethanol (4). Nonetheless, the transcriptional rewiring occurring during this response to ethanol and its importance in comparison with the contribution of genomic changes have not been explored. Indeed, the implication in ethanol response and adaptation of duplicates, from a transcriptional perspective, have been only marginally explored recently by our group (19, 23). The interplay between duplicates and transcriptional rewiring remains unknown. It also remains elusive whether and how *S. cerevisiae* could optimize the use of ethanol as nonfermentative carbon source.

In this study, we undertake the challenge of elucidating the role of transcriptional rewiring to the response and adaptation to ethanol (as sole carbon source) in *S. cerevisiae* and revealing the link between gene duplication and ethanol usage. As already mentioned, previous studies revealed an unprecedented complexity in the genomic dynamics underlying adaptation to ethanol but, however, did not address the implication of transcriptional reprogramming of the ancestral duplicates (4, 5, 25). Here, we evolved clonal populations of *S. cerevisiae* using glucose as carbon source and challenged them to use ethanol as sole carbon source in short and long (ethanol adaptive laboratory evolution) responses. We reveal the transcriptional reprogramming basis and the interplay of this with gene duplication in the response and adaptation to ethanol.

RESULTS

Phenotypic changes of *S. cerevisiae* in response and adaptation to ethanol. At time points t_0, t_{100}, and t_{110}, we characterized growth parameters of *S. cerevisiae* populations in the standard medium (yeast extract-peptone-dextrose [YPD]) and stressful medium (yeast extract-peptone-ethanol [YPE]), using optical density measurements (Fig. 1A). The mean maximum growth rate (μ_{max}) was significantly lower at time t_0 in YPE ($\mu_{max} \pm$ standard deviation of the mean [SDm] = 0.1303 ± 0.0093 h^{-1}) than in YPD

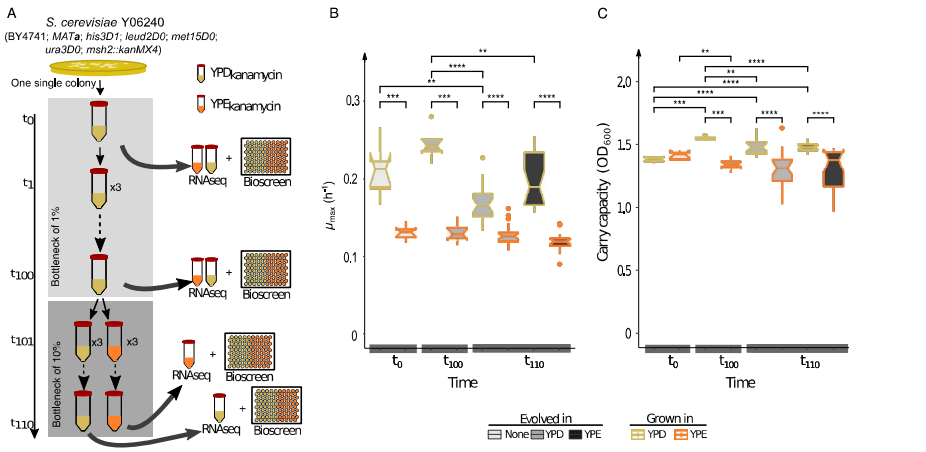


FIG 1 Experimental adaptive laboratory evolution scheme and phenotypic characterization in *Saccharomyces cerevisiae* Y06240. (A) Experimental layout. From a single *S. cerevisiae* Y06240 (*MATa; his3D1; leud2D0; met15D0; ura3D0; msh2::kanMX4*) colony, we derived a population in liquid YPD medium (called population t_0). This population was split into 3 replicates and evolved in YPD for 100 passages (approximately 660 generations) by daily transferring 1% (0.5 ml) to a new tube (50 ml) with fresh YPD medium (4.5 ml) (we called this population t_{100}). After passage 100, we started the adaptive evolution, by splitting up the evolutionary experiment into two: one continuing to evolve in YPD (with 2% glucose as carbon source) and the other replacing the glucose with 3% ethanol (medium YPE). Populations were evolved for 10 passages with a daily 10% bottleneck (approximately 33 generations; we called this population t_{110}). In the experimental scheme, the points at which phenotypic characterization and transcriptome changes (RNAseq) were carried out are indicated. (B) Phenotypic characterization was performed by characterization of population growth curves. Maximum growth rate (h^{-1}) of each population was determined at each control time point (t_0, t_{100}, and t_{110}) in their evolving medium (YPD or YPE) and in the challenge medium (YPE or YPD). (C) "Carry capacity" of each population (OD_{600}) was also determined for each population and each control time point in their evolving medium and in the challenge one. Significant differences of each growth parameter are indicated as *, **, ***, and ****, when the probabilities are $P < 0.05$, $P < 0.005$, $P < 10^{-3}$, and $P < 10^{-4}$, respectively, using a Wilcoxon rank test.

($\mu_{max} \pm$ SDm = 0.2096 ± 0.0343 h^{-1}; Wilcoxon rate test, $P = 5.8 \times 10^{-4}$). This difference between growth rates was also observed in all the evolved lines at all time points (Fig. 1B; see also Fig. S1 in the supplemental material). Diversifying the population for approximately 660 generations (t_{100}) increased the growth rate in YPD for all lines. Furthermore, the populations also increase their carry capacity after diversification (Fig. 1C; Fig. S2). However, only one of the lines increased its growth rate in YPE; the other two retained a similar growth rate as the ancestral population (Fig. S1 and S2). All lines, except one, at time t_{110} reduced the growth rate after just 33 generations. The populations evolved in YPE and when challenged to grow in YPD showed a higher growth rate than the control population evolved in YPD. This difference comes from the growth rate recovery of one of the evolved populations in YPE. The rest of the populations in t_{100} perform similarly. Overall, the evolved population in YPE reduced their growth rate in YPE compared to the evolved populations in YPD but increased their carry capacity (Fig. S1 and S2).

Up- and downregulation in response and adaptation to ethanol. Transcriptome sequencing (RNAseq) was conducted in populations t_0, t_{100}, and t_{110}, in YPD and/or challenged with ethanol (YPE) (Fig. 1A). The exposure to ethanol led to the upregulation (fold change [FC] in the expression of the genes >25%, false-discovery rate [FDR] <0.005) of 833 and 1,389 genes compared to the same population grown in YPD for t_0 and t_{100}, respectively. Of the 833 genes upregulated in t_0, 557 (66.9%) were also upregulated in t_{100} (Fig. S3A). Adaptation to ethanol stress for 10 passages led to the upregulation of 1,694 genes, of which 437 were also upregulated in t_0 and t_{100} (we call these core upregulated genes) (Fig. S3A). The exposure to ethanol led to the downregulation of 751 and 940 genes compared to the same population grown in YPD for

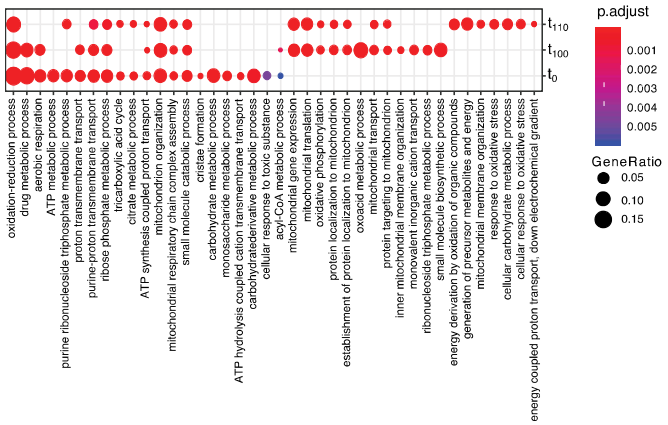


FIG 2 Biological processes enriched due to the use of 3% ethanol as the sole carbon source. Enrichment analysis of functional categories (biological process) for upregulated genes in YPE compared to YPD, at the three time points (t_0, t_{100}, and t_{110}), was performed with clusterProfiler.

t_0 and t_{100}, respectively. Of the 751 downregulated genes in t_0, 326 (43.4%) were also downregulated in t_{100} (Fig. S3B). The adaptation to ethanol stress led to the downregulation of 1,391 genes, of which 222 were also downregulated in t_0 and t_{100}.

Low overlap in the transcriptomic response and adaptation to ethanol. The number of upregulated genes among the populations t_0 and t_{100} (67% of t_0 upregulated genes are also upregulated in t_{100}) was high for the t_0 population, but t_{100} showed twice as many upregulated genes as t_0, perhaps indicating that experimental evolution in YPD for 100 passages has involved significant polymorphism in the transcriptomic reprogramming of cells in this population. Only 437 genes were core upregulated genes in all three populations. Populations t_{110}, the populations derived from t_{100} and evolved for 10 days in ethanol, showed an overlap of only 883 upregulated genes with their parental t_{100} populations despite the low number of passages separating them (Fig. S3A).

To determine whether the functions affected by the transcriptomic responses have changed among populations, we performed an analysis of Gene Ontology (GO) terms of the set of upregulated genes. Populations at t_0 and t_{100} exhibited enrichment for upregulated genes ($P < 0.01$) in similar functional categories, affecting mainly the "oxidation-reduction process," "drug metabolic process," "aerobic respiration," "proton transmembrane transport," "mitochondrion organization," "small-molecule catabolic process," "oxidoreductase activity," "cofactor binding," and "proton transmembrane transporter activity" (Fig. 2). The analysis, on the other hand, of GO term enrichment for upregulated genes in t_{110} population led to a somewhat different result. There was some overlap of enriched GO terms from t_0 and t_{110}, but more importantly, a number of GO term enrichments were specific for t_{110}. They include terms "energy derivation by oxidation of organic compounds," "response to oxidative stress," and "cellular response to oxidative stress and response to inorganic substances" (Fig. 2).

The GO term analysis for downregulated genes showed even less overlap between the three populations (Fig. S4). Population t_0 had enrichment in "cytoplasmic translation and ribosome biogenesis," which was also enriched at population t_{110}. Furthermore, population t_{110} had "ncRNA processing and methylation" enriched for downregulated genes. In contrast, population t_{100} had only a few GO terms enriched,

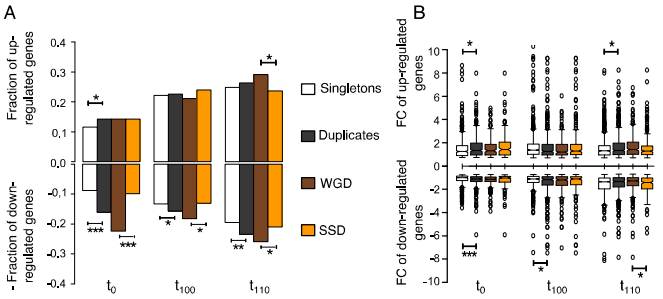


FIG 3 Genes responding transcriptionally to glucose replacement by ethanol, as carbon source, after adaptive evolution. (A) Proportion of responding genes (showing transcriptional divergence [TD]) distributed in four categories (singletons, duplicates, WGDs, and SSDs). Upregulated genes are on the positive part of the *y* axis, whereas downregulated genes are on the negative part of the axis (after being made negative for representation purposes). Fisher's exact test has been used to test if the observed fractions of TD genes are significantly different from those expected. (B) Expression difference (log fold change) in the two media, YPD and YPE. A Wilcoxon rank test has been used to test the difference in expression levels of sets of genes. Significant differences are indicated as *, **, and ***, when the probabilities are $P < 0.05$, $P < 0.005$, and $P < 10^{-3}$, respectively.

including "carbohydrate transport" and "nucleic acid phosphodiester bond hydrolysis," with none of them overlapping the other two populations.

Duplicated genes encode rapid responses and adaptations to ethanol. Since duplicated genes are involved in the origin of new functions (3, 26–30), we sought to investigate if the response to ethanol, as the sole carbon source, was mainly driven by duplicates, differentiating between WGDs (27) and SSDs (28) in our analyses.

Population t_0 exhibited 312 duplicate (14.2%) and 524 singleton (11.6%) genes out of the 833 upregulated genes. The proportion of upregulated duplicates was higher than that of singletons (Fisher's exact test: odds ratio $F = 1.22$, $P = 0.0071$) (Fig. 3A). We also observed a higher expression fold change (FC) difference in duplicates (median FC = 1.343) than in singletons (median FC = 1.244) (Wilcoxon rank test: $P = 0.0215$) (Fig. 3B). We found no difference in the response of duplicates when analyzing their origin (312 = 156 WGDs + 156 SSDs) (Fisher's exact test: odds ratio $F = 1.002$, $P = 1$). Likewise, no difference was observed in the expression fold change between WGDs (median FC = 2.92) and SSDs (median FC = 3.13) (Wilcoxon rank test: $P = 0.65$).

Population t_{100} showed no difference in the response to ethanol between upregulated duplicates (494) and singletons (1,000) (Fisher's exact test: odds ratio $F = 1.01$, $P = 0.807$). Remarkably, while the proportion of duplicates that were upregulated increased 58% after 100 passages of evolution, the proportion of singletons increased 92% in comparison with the parental population t_0 (Fig. 3A). No difference in expression fold change was observed between duplicates and singletons, nor between WGDs and SSDs.

Population t_{110} exhibited 578 upregulated duplicates and 1,116 upregulated singletons, not being significantly different (Fisher's exact test: odds ratio $F = 1.06$, $P = 0.285$). Interestingly, upregulated duplicates (median FC = 1.346) saw a higher expression fold change than singletons (median FC = 1.298) (Wilcoxon rank test: $P = 0.0288$). We found significantly more WGDs (319) responding to ethanol than SSDs (259) (Fisher's exact test: odds ratio $F = 1.23$, $P = 0.0248$), but no difference of expression fold change between the two (WGDs: median FC = 1.408; SSDs: median FC = 1.274; Wilcoxon rank test: $P = 0.0586$).

The response of duplicates to ethanol was even more apparent when looking at downregulated genes. Population t_0 showed a larger proportion of the duplicates (353) being downregulated than singletons (398) (Fisher's exact test: odds ratio $F = 1.82$, $P = 2.33 \times 10^{-14}$). Not only were there more ethanol-responding duplicates, but the

response was also higher (duplicates: median FC = −1.103; singletons: median FC = −1.004; Wilcoxon rank test: $P = 3.75 \times 10^{-4}$). Most of this response was coming from WGDs (WGDs: 245; SSDs: 108; Fisher's exact test: odds ratio $F = 2.27$, $P = 8.31 \times 10^{-12}$). No difference in the expression fold change was found between the two types of duplicates (WGDs: median FC = −1.130; SSDs: median FC = −1.019; Wilcoxon rank test: $P = 0.1804$).

Population t_{100} showed similar results as population t_0: 343 of the downregulated genes were duplicates and 597 were singletons (Fisher's exact test: odds ratio $F = 1.18$, $P = 0.024$). The expression fold change was also higher in duplicates (median FC = −1.158) than in singletons (median FC = −1.108) (Wilcoxon rank test: $P = 0.0174$). More WGDs (199) were responding to ethanol stress than SSDs (144) (Fisher exact test: odds ratio $F = 1.39$, $P = 6.28 \times 10^{-3}$), but no difference in the expression fold change was observed (WGDs: median FC = −1.186; SSDs: median FC = −1.113; Wilcoxon rank test: $P = 0.484$).

Population t_{110} had no difference in duplicated genes (514) being downregulated compared to singletons (877) (Fisher exact test: odds ratio $F = 1.20$, $P = 2.71 \times 10^{-3}$). However, duplicated genes had a higher expression fold change than singletons (duplicates: median FC = −1.332; singletons: median FC = −1.349; Wilcoxon rank test: $P = 0.338$). Interestingly, WGDs (284) were more abundant than SSDs (230) (Fisher's exact test: odds ratio $F = 1.23$, $P = 0.031$), but SSDs (median FC = −1.421) showed a higher expression fold change than WGDs (median FC = −1.271) (Wilcoxon rank test: $P = 0.0266$). The core gene of the downregulated genes consisted of more duplicates (4.3%) than singletons (2.84%) (Fisher's exact test: odds ratio $F = 2.36$, $P = 1.27 \times 10^{-4}$), with WGDs as the most affected duplicates (WGD 6.02%; SSD 2.5%: Fisher's exact test: odds ratio $F = 2.24$, $P = 1.63 \times 10^{-8}$).

Transcriptional divergence between duplicates gene copies is linked to the response and adaptation to ethanol in *S. cerevisiae*. If duplicates were linked to the response and adaptation of *S. cerevisiae* to ethanol, then we should expect the transcriptional divergence (TD) between gene copies of a duplicate to be correlated with its transcriptional patterns in ethanol. We identified those duplicated genes that exhibited a fold change expression difference between their gene copies of more than 25%. Of the 1,090 duplicated gene pairs (analysis contained both copies), 867 showed transcriptional divergence between gene copies in YPD. In the populations t_0, 274 of the 312 upregulated duplicates in ethanol belonged to duplicates with evidence of TD, a proportion greater than expected by chance (binomial test: $P = 1.846 \times 10^{-4}$) (Fig. 4A). The mean expression fold change (in logarithmic scale) of the 274 duplicates was 1.59. We compared this mean to a null distribution of means built by sampling 274 duplicates from the population of the 867 duplicates with evidence of expression divergence (Fig. 4B). The mean fold change of these duplicates was greater than expected by chance ($P = 3.0 \times 10^{-6}$). The fold change of the gene copy with the highest expression in ethanol divided by that of the least expressed gene copy is also correlated with the expression fold change of the duplicate (Pearson correlation: $r = 0.66$, $P < 2.2 \times 10^{-12}$). Importantly, among the most highly divergent and upregulated duplicates, we identified the plasma membrane H^+-ATPase (*PMA2*), translational elongation factor (*HEF3*), plasma membrane permeases (*GIT1* and *SEO1*), and a gene involved in the metabolism under respiratory conditions (*RGI2*), among others (Table S1).

In the t_{100} population 426 duplicates of the 492 showed evidence of upregulation and belonged to the set of duplicates with evidence of expression divergence, a proportion greater than expected by chance (binomial test: $P = 6.81 \times 10^{-5}$). Like in the t_0 population, upregulated duplicates exhibited greater mean expression divergence between gene copies than expected by chance (mean = 1.45, $P = 1.05 \times 10^{-3}$) (Fig. 4B). The phenotypic plasticity (expression fold change of duplicates in ethanol compared to YPD) was correlated with the expression divergence between the gene copies (Pearson correlation: $r = 0.58$, $P < 2.2 \times 10^{-16}$).

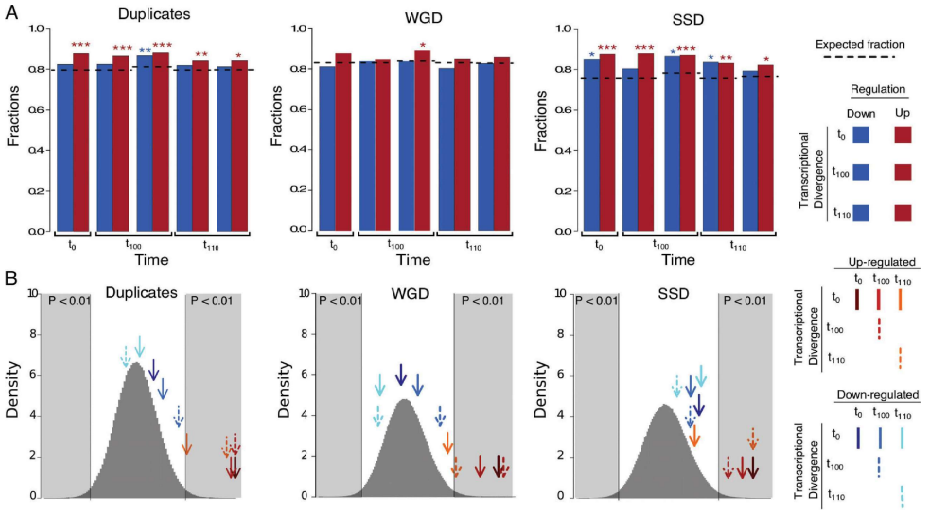


FIG 4 Proportion and mean transcriptional divergence of transcriptionally responding duplicates. (A) The proportion of TD duplicates of transcriptionally responding genes, along with the expected proportions, marked with dashed lines. Significant differences in the observed number of TDs compared to the expected number are indicated as *, **, and ***, when the probabilities are $P < 0.05$, $P < 0.005$, and $P < 10^{-3}$, respectively, using a binomial test. (B) Mean transcriptional divergence levels between duplicated genes within each of the categories (marked with arrows) are mapped onto a normal distribution build by random sampling, without replacement, of the same size from the corresponding gene pools. The gray blocks are indicating the significant part of the distribution ($P < 0.01$).

Population t_{110} also presented enrichment of upregulated duplicates for duplicates with evidence of expression divergence, with 485 out of the 576 upregulated duplicates exhibiting expression divergence (binomial test: $P = 5.22 \times 10^{-3}$). Interestingly, upregulated duplicated genes in t_{110} do not show higher transcriptional divergence than expected when calculated from the t_0 population (mean = 1.29, $P = 0.150$) but do show such when calculated from the t_{110} population (mean = 1.38, $P = 0.00120$), indicating that transcriptional divergence and upregulation are highly dependent on the current transcriptional background. Similar to the t_0 and t_{100} populations, the population of t_{110} shows correlation between phenotypic plasticity and expression divergence of duplicated genes (Pearson correlation: $r = 0.5$, $P < 2.2e{-}16$).

In contrast to this pattern for the upregulated genes, we see no correlation between TD and downregulated genes, at any of the time points (Fig. 4). The only correlation that is also present for the downregulated genes is the phenotypic plasticity in YPD and YPE (Pearson correlations: t_0, $r = 0.68$, $P < 2.2e{-}16$; t_{100}, $r = 0.67$, $P < 2.2e{-}16$; t_{110}, $r = 0.61$, $P < 2.2e{-}16$).

Understanding why up- and downregulated genes show different patterns with respect to TD, we first checked the overall TD of up- and downregulated genes (Fig. 5B). Downregulated duplicates had significantly lower TD than upregulated genes ($P = 0.00244$). The duplicated genes we are looking at are TD, which means we have one copy with a lower expression than the other (Fig. 5A). Dividing up- and downregulated genes into low and high transcriptional diverged copy (TDC), we observe, as expected, higher TDC in downregulated genes (binomial test: $P = 0.0018$) and lower TDC in upregulated genes (binomial test: $P = 1.473 \times 10^{-9}$) (Fig. 5C). All groups showed similar TD except for downregulated and low TDC, which had the lowest TD of all groups (Fig. 5C). Looking at GO enrichments of the four categories of Fig. 5C, no

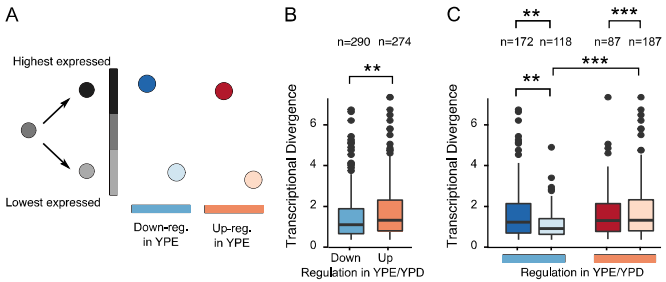


FIG 5 Transcriptional divergence of transcriptionally responding duplicates. (A) Transcriptionally divergent (TD) duplicate characterization. TD duplicates are classified according to their sign of expression (high expression shown in dark colors and low expression shown in light colors), as the responding gene can be either the gene with high expression or the gene with low expression of the duplicated pair. Blue and red indicate the downregulated and upregulated pairs in YPE, respectively. (B) Comparison of TDs of up- and downregulated genes at t_0. (C) Comparison of TD pairs at t_0, differentiating each up-expressed gene into high- and low-expression gene categories as indicated in the scheme depicted in panel A. Significant differences are indicated as *, **, and ***, when the probabilities are $P < 0.05$, $P < 0.005$, and $P < 10^{-3}$, respectively. A Wilcoxon rank test was used for testing the significance between TDs of the different categories, whereas a binomial test was used for testing the number of TDs in the different categories.

overlap is observed between the upregulated and downregulated categories (Fig. S5). Upregulated (highest transcribed copy) genes were enriched for "drug metabolic process," "energy derivation by oxidation of organic compounds," and "small molecule metabolic process," and downregulated duplicates were enriched for "cytoplasmic transition" and different ribosome processes. To determine if the behavior of a gene has influence on the response of the other duplicated copies, we further divided the groups into categories of the two copies having the same regulation profile, the two copies having different regulation profiles, or only one of the copies showing up- or down-regulation in ethanol (Fig. 6). Interestingly, we observed more duplicated copies which were both downregulated than expected ($P < 10^{-12}$), but this group also had the lowest TD. Inspecting the function of these genes, we see that a majority (48 out of the 60 genes) are ribosomal proteins. As would be expected, a lot of overlap of enriched GOs was observed between the different up- and downregulated categories. In the

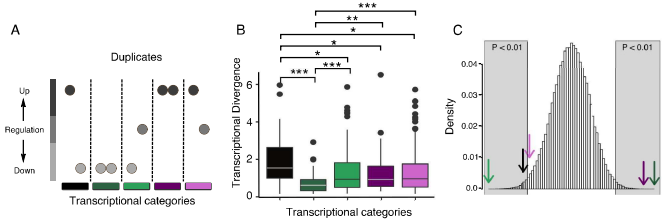


FIG 6 Characterization of TD transcriptionally responding duplicated pairs. (A) TD duplicate categorization scheme (five categories): black, one gene copy is upregulated and the other is downregulated; dark green, both duplicated genes are downregulated; light green, one duplicate is downregulated and the other is unaltered; purple, both duplicates are upregulated; violet, one duplicated gene is upregulated and the other is unaltered. (B) Comparison of TDs of the five categories described. A Wilcoxon rank test was used to determine significant differences indicated as *, **, and ***, when the probabilities are $P < 0.05$, $P < 0.005$, and $P < 10^{-3}$, respectively. (C) Mean number of genes within each of the TD categories (marked with arrows with the coloring code described for panel A), mapped onto a normal distribution build by random sampling, without replacement, of the same size from the corresponding gene pools. Gray blocks over the normal distribution indicate the significant part of the distribution ($P < 0.01$).

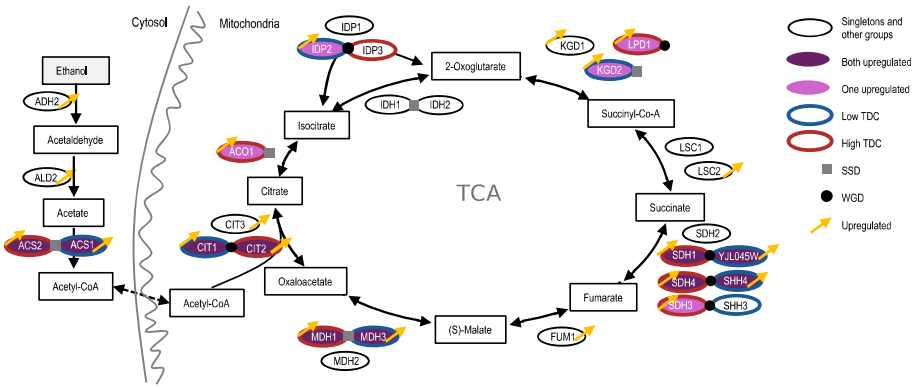


FIG 7 Pathway of nonfermentative C_2 metabolism in *S. cerevisiae* Y06240, from ethanol to TCA. Only ethanol degradation and the TCA pathway have been shown. Pathway information was taken from the KEGG pathway database. The proteins are colored after the duplicate categories set out in Fig. 5 and 6, with indication of duplicate origin (SSD or WGD) and transcription-diverged copy level (TDC, in blue or red for low or high, respectively) or upregulation (yellow arrow).

case of both duplicates being upregulated, we saw an enrichment for "carbohydrate metabolic process" and "oxidative phosphorylation." Furthermore, these categories are the only ones which observe an enrichment of a pathway, namely, "superpathway of TCA cycle and glyoxylate cycle." For the discordant duplicates, enrichment categories included "glucose 6-phosphate metabolic process," "NADP metabolic process," and "oxidoreduction coenzyme metabolic process" (Fig. S5).

Changing the carbon source from glucose to ethanol implies that the yeast goes from fermentation to aerobic respiration. Combining this with the fact that the tricarboxylic acid (TCA) cycle was enriched for upregulated duplicated pairs, we map the categorized proteins onto the two pathways (Fig. 7). At least one of the proteins involved in each of the steps was upregulated, and in most cases the duplicated pairs were upregulated (i.e., *CIT1* and *CIT2*, *MDH1* and *MDH3*, and *ACS2* and *ACS1*). Interestingly, in cases where only one of a duplicated pair was within this pathway, we saw upregulation of just one of the proteins (i.e., *ACO1*, *LDP1*, or *KGD2*), namely, the one within the TCA cycle.

Transcriptional divergence between duplicated genes plays different roles in WGDs and SSDs. If TD of duplicated genes plays the same role in WGDs and SSDs, the same patterns should be observed. It has previously been noted that WGDs are more transcriptionally divergent than SSDs (30). Using our data, we find 910 WGDs to be TD in YPD at t_0, compared to 824 SSDs, not a significant difference (Fisher's exact test: $P = 0.1389$). However, when looking at the magnitude of the TD between duplicates, we observed a significant difference between the two types, with WGDs showing a higher TD than SSDs (Wilcoxon rank sum test: $P = 0.01281$) (Fig. 5B). It is worth noting that this difference of magnitude, between WGD and SSD, disappears if we look only at TD gene copies (Wilcoxon rank sum test: $P = 0.1575$).

To determine if the TD between the gene copies of WGDs and SSDs had different influences on the response to ethanol, we looked at how many TD duplicates were up- or downregulated in YPE compared to YPD at all three time points. For the duplicates *per se*, we had seen in the section above that upregulated genes contained more TD genes than expected (Fig. 4A). When separating out the two types of duplicates, it was seen that SSDs contained more TD upregulated genes than expected at all three time points (binomial test: t_0, $P = 2.399 \times 10^{-4}$; t_{100}, $P = 7.239 \times 10^{-7}$; t_{110}, $P = 3.622 \times 10^{-3}$), as well as for downregulated genes at t_0 and t_{110} (binomial test: t_0,

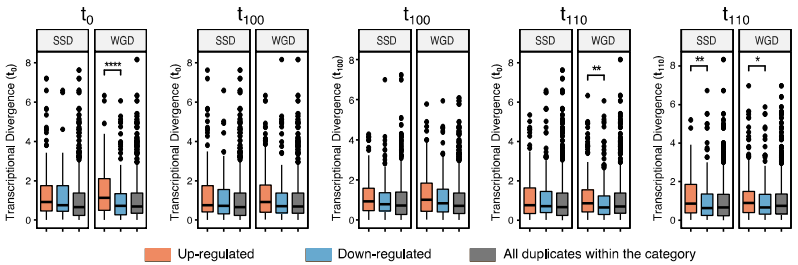


FIG 8 Distribution of transcriptional divergence per time point of the experimental evolution and per duplicate origin (SSDs and WGDs). A Wilcoxon rank test was performed to determine significant differences in the TD, indicated as *, **, ***, and ****, when the probabilities are $P < 0.05$, $P < 0.005$, and $P < 10^{-3}$, and $P < 10^{-4}$, respectively.

$P = 0.02408$; t_{110}, $P = 3.349 \times 10^{-3}$). This is the opposite pattern from what we observed in WGDs, where neither up- nor downregulated genes, at any of the time points, had more TD genes than expected (Fig. 4A). To rule out that the limit set for a duplicated gene pair to be TD was not affecting our results, we redid the analysis for the TD limit going from equal expression to a 4-fold difference (Fig. S6). In general, the pattern did not change much as the TD limit was changed, in particular at the lower TD limits.

As we saw a difference between WGDs and SSDs with respect to the quantity of TD genes that reacted to the ethanol stress, we wanted to see if there was a difference with respect to the magnitude of the TD and ethanol response. First, we compared the TDs of up- and downregulated genes. The general pattern observed was that the WGDs show a statistical difference between the magnitudes of TD of up- and downregulated genes (Wilcoxon rank sum test: t_1, $P = 6.5 \times 10^{-5}$; t_{100}, $P = 0.05$; t_{110}, $P = 5.3 \times 10^{-3}$) (Fig. 8). In contrast, SSD showed no difference between the magnitudes of the TD for up- and downregulated genes. Second, we wanted to see if the observed mean TD of differentially expressed genes was higher or lower than expected by chance. We compared the observed mean TD with the normal distribution build by random sampling of the same size from the corresponding pools (WGDs and SSDs). In this case, we observed similar patterns for both WGDs and SSDs. The TD of upregulated genes exhibits the expected mean of TD. One interesting thing for both was that the mean TD at t_{110} was significant only when calculated from t_{110} but not from t_0, indicating that the transcriptional background has an influence on the response of the duplicated genes in ethanol stress. Furthermore, neither WGDs nor SSDs showed a significantly higher mean of TD of downregulated genes than expected, at all time points (Fig. 4B).

DISCUSSION

Large transcriptional response to ethanol stress. One of the central mechanisms of unicellular organisms, and particularly nonmobile ones, is sensing and responding to changes in the environment. This is especially essential when the organism endures stress. Here, we show how changing the carbon source from glucose to ethanol leads to large transcriptional changes in the yeast *S. cerevisiae*. These changes are observed in a large percentage of the *S. cerevisiae* genes and encode a wide range of functions. Such genome-wide transcriptional changes have been shown before to take place in the response to numerous stresses, including glucose restriction (31), glycerol as the only carbon source (19), oxidative stress (32–34), environmental estrogen (35), acid tolerance (36, 37), and thermal resilience (38), among others (39). Indeed, when the initial population was switched from a glucose-containing medium to one that contained only ethanol as carbon source, hundreds of genes altered their expression. With the medium not containing glucose, the yeast is performing aerobic respiration (4, 5, 7),

and the presence of ethanol induces oxidative stress of the cells (40–45), which is reflected in our results, where we have an enrichment of "oxidation-reduction process" and "aerobic respiration" gene terms in the upregulated gene classes. Hence, two clear transcriptional patterns were observed during exposure to ethanol: (i) an upregulation of genes involved in stress response and (ii) a downregulation of ribosomal biogenesis or energy-dependent processes. Our observations agree with the suggested tradeoff cellular response (transcriptional rewiring of central metabolism) to environmental stress on yeast growth rate (36, 39, 46, 47).

The genetic background influences the transcriptional response to ethanol. The genetic background of a population influences its capability to grow and respond to stress (48–50). In this study, evolving a population of *S. cerevisiae* for a large number of generations had huge influences on the transcriptional response to ethanol, as well as increasing the growth rate in the evolved medium (YPD). We observed transcriptional changes between the evolved and the nonevolved ancestral population. This change is likely due to a genetic change in the population, as the original ancestral population had low variability as it originated from a single clone and the evolved population gained genetic variability through the evolution experiment, as described recently (51). An interesting result from our study is that the evolved population improved its fitness in the evolved medium (YPD) but showed no change in the fitness when grown in ethanol (YPE). It has previously been shown that diversification can lead to exaptation in nonevolved environments (19, 33, 51, 52). There are multiple possible reasons for us not observing this in our populations. First, increasing fitness to ethanol is hard. Many of the studies which have observed increased tolerance to ethanol see an increase of the ploidy of chromosome III (53, 54); this occurs only in diploid and polyploid *S. cerevisiae*, and we are evolving a haploid population. Second, our ancestral population might have been at a local maximum in the ethanol fitness landscape of the population. Last, despite evolving our populations for approximately 660 generations, it might not have been long enough to acquire any exaptation to ethanol, deserving further study of the mutational landscape of these populations.

Going from acute to chronic exposure of ethanol rewires the transcriptome. The evolution of the yeast population in the presence of ethanol (the t_{110} populations) uncovered the regulatory changes that occur as the population reacts to acute and chronic exposure. It has previously been suggested that reducing the growth rate can lead to increased stress tolerance by redirecting the resources (47). The chronic-exposure population (evolved population in ethanol, YPE_t_{110}) showed an enrichment of upregulation of genes involved in oxidative stress, so overall these populations were upregulating more genes involved in stress response, an indication of higher allocation of resources to stress tolerance. This agrees with the fact that we observed a lower growth rate on ethanol for the evolved population in ethanol than for the population that evolved in YPD.

Duplicated genes play an important role in the response to ethanol. The first response, of an organism to stress, is though regulatory reprogramming; hence, plasticity of the transcriptome will determine the potential for adapting to a new environment (55–57). However, this link is still not fully understood, enthralling scientists for the past 40 years and becoming of great importance recently (58), and that is predominately down to the difficulty of mapping phenotypes to genotype and assigning transcriptional changes to phenotypic variations. In this work, we clearly see a link between transcriptional variations, phenotypic response to ethanol, and gene copy number (referring here to duplicates), as the transcriptionally altered genes are enriched for duplicated genes. The classical theory behind the evolution of duplicated genes states that one gene copy is able to evolve without or with reduced selection constraints as the other gene copy is performing the ancestral function (28, 59, 60). Diversification of the gene copies happens not only at the functional level but also at the expression level (34, 61, 62). The diversification at the expression level could open up the possibility to diverge functionally, as the rate of evolution is highly linked to its expression, although recently it has been shown that low-expression transcription

factors adapt through cooperation rather than functional divergence (34, 63). It has been suggested that the whole-genome duplication even facilitated the *Saccharomyces* yeast to evolve the ability to ferment sugars under anaerobic conditions, which is not the case for other yeasts (reviewed in references 64 to 66). Here, we are forcing the yeast to use ethanol as the sole carbon source, meaning it has to perform respiration instead for fermentation, not requiring the Crabtree effect-implicated genes. In correlation with this, we observe that most of the changes of the duplicated genes were downregulation of WGD, consistent with the hypothesis that WGDs were providing the raw material for conservation of dosage-sensitive genes involved in both rewiring of rapid growth elements (ribosomal protein genes) and divergent regulation and specialization of gluconeogenesis-ethanol consumption phase versus glycolysis-ethanol production (8, 65). Taken as a whole, the rewiring of the transcriptome, and in particular the duplicated genes, indicates that the yeast cell goes into energy preservation when the carbon source is switched to ethanol.

The transcriptional background and the response to ethanol. In plants, it has been observed that duplicated genes diverge transcriptionally soon after duplication (66–69). Furthermore, a correlation between the divergence from the ancestral expression level and stress response has also been observed in plants (70). These all indicate that duplication and expression divergence are linked to adaptation and stress response (67, 71). In yeast, duplicated genes have also been shown to be transcriptionally diverged, particularly in WGD (19). In a wider study looking at transcriptional changes of duplicated genes under different stress conditions, it was observed that one of the gene copies was more transcriptionally plastic than the other (23). These all indicate that transcriptional divergence plays an important role in maintaining duplicated genes in the genome and expanding the phenotypic plasticity of the organism. Here, we observe that transcriptional divergence between gene copies is correlated with response to ethanol. In particular, responding duplicates have higher transcriptional divergence than expected. However, WGD and SSD have different parameters by which the TD influences the response to ethanol. The magnitude of the TD is important for WGD, where in contrast the number of genes with TD is important for SSD. One interesting thing that we observed in this study is the change of TD of the duplicated genes throughout our experiment and that this change was correlated with the response to ethanol, indicating that the transcriptional background is important for the actual stress response and this can change relatively quickly.

Concluding remarks. The recent advances in next-generation sequencing technologies coupled with the decrease of their prices have increased general interest in determining the role of polyploidy and transcriptional plasticity in ecological shifts or lifestyles. The switch to use ethanol as sole carbon source implied a yeast cell reprogramming to energy preservation with low growth rate but with similar biomass production due to transcriptional reprogramming of duplicates, especially those of the TCA cycle. In this work, we have unveiled that TD between duplicates and the transcriptional background affect duplicates' response to ethanol, with the magnitude of the TD being especially important for WGDs.

MATERIALS AND METHODS

Yeast culture and experimental evolution. The *Saccharomyces cerevisiae* strain Y06240 (BY4741: *MATa*; *his3D1*; *leud2D0*; *met15D0*; *ura3D0*; *msh2::kanMX4*) was used as described previously (22, 72). Briefly, a homogeneous population founded by growing a colony in a liquid culture of rich medium (YPD: 2% [wt/vol] Bacto peptone, 1% [wt/vol] yeast extract, 2% [wt/vol] dextrose; supplemented with 100 μg/ml kanamycin) (t_0) was evolved through daily bottlenecks (1%) for 100 days (t_{100}; ~660 generations), in 5 ml of YPD medium in 50-ml Corning tubes, at 28°C and 220 rpm. From passage 100 (t_{100}), the population a1 was divided into two sublines, each with three biological replicates. One subline was grown in YPD medium as control (lines Da1), whereas the second subline (lines Ea1) was grown in a medium containing 3% ethanol as the sole carbon source (YPE: 3% [vol/vol] ethanol, 2% [wt/vol] Bacto peptone, 1% [wt/vol] yeast extract; supplemented with 100 μg/ml kanamycin). The populations were evolved for another 10 passages, with a daily bottleneck of 10% of population, in 5 ml of the corresponding medium, as indicated previously. Each 10 passages, a fossil record of each line was established by preserving the entire population in 25% glycerol solution at −80°C (Fig. 1).

Growth characterization. Growth parameters for t_0, t_{100}, and t_{110} were obtained using the Bioscreen C plate-reader system (Oy Growth Curves Ab Ltd., Helsinki, Finland) as described in reference 72. Briefly, each time point was precultured overnight at 28°C, from the corresponding fossil record, and used to inoculate 200 μl of fresh medium (YPD and/or YPE) to an initial optical density at 600 nm (OD_{600}) of 0.06 to 0.07, distributed in 100-well honeycomb plates, with 6 to 7 technical replicates. The experiment was run for 78 h at 28°C with continuous shaking (high level) and taking OD_{600} measurements (brown filter) every 15 min. Each run contained at least 3 controls for each medium (uninoculated fresh medium). The data were analyzed with Growthcurver v.0.3.0 under R-studio (73).

RNA extraction and transcriptomic analysis. The RNA profiling was performed at the t_0, t_{100}, and t_{110} time points as indicated in Fig. 1A, following the same procedures as previously used (22, 72). rRNA-depleted RNA (Illumina) libraries were constructed and sequenced at the Genomic Core Facility at Servicio Central de Soporte a la Investigacion Experimental (SCSIE) from the University of Valencia, Spain. Reads (trimmed) were aligned with Bowtie2 (up to two mismatches accepted) to the reference S288c strain genome (only coding sequences [CDS]). Statistical assessment of differential gene expression was done with edgeR (74), setting false-discovery rate (FDR) at <0.005, and applying BY correction for *P* value (0.005).

Identification of duplicated genes, functional classification, and visualization. Paralogous pairs of duplicated genes were divided into two groups according to their origin mechanism: whole-genome duplicates (WGDs) or small-scale duplicates (SSDs). WGDs (555 pairs) were extracted from the reconciled YGOB list (Yeast Gene Order Browser, last accessed March 2018; http://wolfe.gen.tcd.ie//ygob [75]). SSDs (560 pairs) were identified after best reciprocal hits from all-against-all BLAST searches using BLASTP with an E value cutoff of 1E−5 and a 50-bit score (76), selecting only those that exhibit a distribution of synonymous substitutions similar to WGDs (3, 26). Differential expressed genes were further classified according to their gene ontology (GO) term as implemented in the R package clusterProfiler (77), followed by an enrichment analysis with a *P* value cutoff of <0.01 and with the *P* value being adjusted with the Benjamini and Hochberg (78) method.

Software. Unless otherwise indicated, statistics were performed using the appropriate packages in R v 3.5.1 (R Core Team [2018]).

Data availability. Raw reads are available from the Sequence Read Archive (SRA) with accession numbers PRJNA321113 (t_0 in YPD and YPE), PRJNA610243 ($a1t_{100}$ in YPD), PRJNA610541 ($a1t_{100}$ in YPE), PRJNA610474 ($Da1t_{110}$ in YPD), and PRJNA610515 ($Ea1t_{110}$ in YPE).

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

FIG S1, EPS file, 0.1 MB.

FIG S2, EPS file, 0.1 MB.

FIG S3, EPS file, 0.2 MB.

FIG S4, EPS file, 0.1 MB.

FIG S5, EPS file, 0.2 MB.

FIG S6, EPS file, 0.4 MB.

TABLE S1, CSV file, 0.1 MB.

ACKNOWLEDGMENTS

This work was supported by grants BFU2015-66073-P from the Spanish Ministry of Economy and Competitiveness (MINECO-FEDER) to M.A.F. and SEJI/2018/046 from the Generalitat Valenciana, Programa a la excelencia cientifica de investigadores juniors, to C.T. F.M. was supported by an FPI grant from the Spanish Ministry of Economy and Competitiveness (BES-2016-076677). The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

M.A.F., B.S.-M., and C.T. conceived and designed the study. B.S.-M. and F.M. performed the experiments. M.A.F., B.S.-M., and C.T. analyzed and interpreted the data and drafted the article. All authors approved the final version.

The authors have no conflict of interest to declare.

REFERENCES

1. Dittmar K, Liberles D (ed). 2010. Evolution after gene duplication. Wiley-Blackwell, Hoboken, NJ.
2. Wolfe KH, Shields DC. 1997. Molecular evidence for an ancient duplication of the entire yeast genome. Nature 387:708–713. https://doi.org/10.1038/42711.
3. Fares MA, Keane O, Toft C, Carretero-Paulet L, Jones GW. 2013. The roles of whole-genome and small-scale duplications in the functional specialization of *Saccharomyces cerevisiae* genes. PLoS Genet 9:e1003176. https://doi.org/10.1371/journal.pgen.1003176.
4. Voordeckers K, Kominek J, Das A, Espinosa-Cantú A, De Maeyer D, Arslan A, Van Pee M, van der Zande E, Meert W, Yang Y, Zhu B, Marchal K, DeLuna A, Van Noort V, Jelier R, Verstrepen KJ. 2015. Adaptation to high ethanol reveals complex evolutionary pathways. PLoS Genet 11:e1005635. https://doi.org/10.1371/journal.pgen.1005635.
5. Snoek T, Verstrepen KJ, Voordeckers K. 2016. How do yeast cells become tolerant to high ethanol concentrations? Curr Genet 62:475–480. https://doi.org/10.1007/s00294-015-0561-3.
6. Ding J, Huang X, Zhang L, Zhao N, Yang D, Zhang K. 2009. Tolerance and stress response to ethanol in the yeast *Saccharomyces cerevisiae*. Appl Microbiol Biotechnol 85:253–263. https://doi.org/10.1007/s00253-009-2223-1.
7. Gancedo JM. 1998. Yeast carbon catabolite repression. Microbiol Mol Biol Rev 62:334–361. https://doi.org/10.1128/MMBR.62.2.334-361.1998.

8. Schuller HJ. 2003. Transcriptional control of nonfermentative metabolism in the yeast *Saccharomyces cerevisiae*. Curr Genet 43:139–160. https://doi.org/10.1007/s00294-003-0381-8.
9. Haas R, Horev G, Lipkin E, Kesten I, Portnoy M, Buhnik-Rosenblau K, Soller M, Kashi Y. 2019. Mapping ethanol tolerance in budding yeast reveals high genetic variation in a wild isolate. Front Genet 10:998. https://doi.org/10.3389/fgene.2019.00998.
10. Dashko S, Zhou N, Compagno C, Piskur J. 2014. Why, when, and how did yeast evolve alcoholic fermentation. FEMS Yeast Res 14:826–832. https://doi.org/10.1111/1567-1364.12161.
11. Mullis A, Lu Z, Zhan Y, Wang T-Y, Rodriguez J, Rajeh A, Chatrath A, Lin Z. 2020. Parallel concerted evolution of ribosomal protein genes in Fungi and its adaptive significance. Mol Biol Evol 37:455–468. https://doi.org/10.1093/molbev/msz229.
12. Yang J, Bae JY, Lee YM, Kwon H, Moon H-Y, Kang HA, Yee S-B, Kim W, Choi W. 2011. Construction of *Saccharomyces cerevisiae* strains with enhanced ethanol tolerance by mutagenesis of the TAT-binding protein gene and identification of novel genes associated with ethanol tolerance. Biotechnol Bioeng 108:1776–1787. https://doi.org/10.1002/bit.23141.
13. Blount ZD, Borland CZ, Lenski RE. 2008. Historical contingency and the evolution of a key innovation in an experimental population of *Escherichia coli*. Proc Natl Acad Sci U S A 105:7899–7906. https://doi.org/10.1073/pnas.0803151105.
14. Blount ZD, Barrick JE, Davidson CJ, Lenski RE. 2012. Genomic analysis of a key innovation in an experimental *Escherichia coli* population. Nature 489:513–518. https://doi.org/10.1038/nature11514.
15. Barrick JE, Yu DS, Yoon SH, Jeong H, Oh TK, Schneider D, Lensk RE, Kim JF. 2009. Genome evolution and adaptation in a long-term experiment with *Escherichia coli*. Nature 461:1243–1247. https://doi.org/10.1038/nature08480.
16. Lenski RE, Wiser MJ, Ribeck N, Blount ZD, Nahum JR, Morris JJ, Zaman L, Turner CB, Wade BD, Maddamsetti R, Burmeister AR, Baird EJ, Bundy J, Grant NA, Card KJ, Rowles M, Weatherspoon K, Papoulis SE, Sullivan R, Clark C, Mulka JS, Hajela N. 2015. Sustained fitness gains and variability in fitness trajectories in the long-term evolution experiment with *Escherichia coli*. Proc Biol Sci 282:20152292. https://doi.org/10.1098/rspb.2015.2292.
17. Turner CB, Blount ZD, Lenski RE. 2015. Replaying evolution to test the cause of extinction of one ecotype in experimentally evolved population. PLoS One 10:e0142050. https://doi.org/10.1371/journal.pone.0142050.
18. Gresham D, Desai MM, Tucker CM, Jenq HT, Pai DA, Ward A, DeSevo CG, Botstein D, Dunham MJ. 2008. The repertoire and dynamics of evolutionary adaptations to controlled nutrient-limited environments in yeast. PLoS Genet 4:e1000303. https://doi.org/10.1371/journal.pgen.1000303.
19. Oz T, Guvenek A, Yildiz S, Karaboga E, Tamer YT, Mumcuyan N, Oza VB, Senturk GH, Cokol M, Yeh P, Toprak E. 2014. Strength of selection pressure is an important parameter contributing to the complexity of antibiotic resistance evolution. Mol Biol Evol 31:2387–2401. https://doi.org/10.1093/molbev/msu191.
20. Toprak E, Veres A, Michel JB, Chait R, Hartl DL, Kishony R. 2011. Evolutionary paths to antibiotic resistance under dynamically sustained drug selection. Nat Genet 44:101–105. https://doi.org/10.1038/ng.1034.
21. Yona AH, Manor YS, Herbst RH, Romano GH, Mitchell A, Kupiec M, Pilpel Y, Dahan O. 2012. Chromosomal duplication is a transient evolutionary solution to stress. Proc Natl Acad Sci U S A 109:21010–21015. https://doi.org/10.1073/pnas.1211150109.
22. Mattenberger F, Sabater-Munoz B, Hallsworth JE, Fares MA. 2017. Glycerol stress in *Saccharomyces cerevisiae*: cellular responses and evolved adaptations. Environ Microbiol 19:990–1007. https://doi.org/10.1111/1462-2920.13603.
23. Kaminski Strauss S, Schirman D, Jona G, Brooks AN, Kunjapur AM, Nguyen Ba AN, Flint A, Solt A, Mershin A, Dixit A, Yona AH, Csörgő B, Busby BP, Hennig BP, Pál C, Schraivogel D, Schultz D, Wernick DG, Agashe D, Levi D, Zabezhinsky D, Russ D, Sass E, Tamar E, Herz E, Levy ED, Church GM, Yelin I, Nachman I, Gerst JE, Georgeson JM, Adamala KP, Steinmetz LM, Rübsam M, Ralser M, Klutstein M, Desai MM, Walunjkar N, Yin N, Aharon Hefetz N, Jakimo N, Snitser O, Adini O, Kumar P, Soo Hoo Smith R, Zeidan R, Hazan R, Rak R, Kishony R, Johnson S, Nouriel S, Vonesch SC, Foster S, Dagan T, Wein T, Karydis T, Wannier TM, Stiles T, Olin-Sandoval V, Mueller WF, Bar-On YM, Dahan O, Pilpel Y. 2019. Evolthon: a community endeavor to evolve lab evolution. PLoS Biol 17:e3000182. https://doi.org/10.1371/journal.pbio.3000182.
24. Sandberg TE, Salazar MJ, Weng LL, Palsson BO, Feist AM. 2019. The emergence of adaptive laboratory evolution as an efficient tool for biological discovery and industrial biotechnology. Metab Eng 56:1–16. https://doi.org/10.1016/j.ymben.2019.08.004.
25. Avrahami-Moyal L, Engelberg D, Wenger JW, Sherlock G, Braun S. 2012. Turbidostat culture of *Saccharomyces cerevisiae* W303-1A under selective pressure elicited by ethanol selects for mutations in SSD1 and UTH1. FEMS Yeast Res 12:521–533. https://doi.org/10.1111/j.1567-1364.2012.00803.x.
26. Keane OM, Toft C, Carretero-Paulet L, Jones GW, Fares MA. 2014. Preservation of genetic and regulatory robustness in ancient gene duplicates of *Saccharomyces cerevisiae*. Genome Res 24:1830–1841. https://doi.org/10.1101/gr.176792.114.
27. Wolfe KH. 2015. Origin of the yeast whole-genome duplication. PLoS Biol 13:e1002221. https://doi.org/10.1371/journal.pbio.1002221.
28. Hakes L, Pinney JW, Lovell SC, Oliver SG, Robertson DL. 2007. All duplicates are not equal: the difference between small-scale and genome duplication. Genome Biol 8:R209. https://doi.org/10.1186/gb-2007-8-10-r209.
29. Ohno S. 1970. Evolution by gene duplication. Springer-Verlag, Heidelberg, Germany.
30. van Hoek MJ, Hogeweg P. 2009. Metabolic adaptation after whole genome duplication. Mol Biol Evol 26:2441–2453. https://doi.org/10.1093/molbev/msp160.
31. Guan Y, Dunham MJ, Troyanskaya OG. 2007. Functional analysis of gene duplications in *Saccharomyces cerevisiae*. Genetics 175:933–943. https://doi.org/10.1534/genetics.106.064329.
32. Ferea TL, Botstein D, Brown PO, Rosenzweig RF. 1999. Systematic changes in gene expression patterns following adaptive evolution in yeast. Proc Natl Acad Sci U S A 96:9721–9726. https://doi.org/10.1073/pnas.96.17.9721.
33. Mattenberger F, Sabater-Muñoz B, Toft C, Sablok G, Fares MA. 2017. Expression properties exhibit correlated patterns with the fate of duplicated genes, their divergence, and transcriptional plasticity in Saccharomycotina. DNA Res 24:559–570. https://doi.org/10.1093/dnares/dsx025.
34. Anand A, Chen K, Catoiu E, Sastry AV, Olson CA, Sandberg TE, Seif Y, Xu S, Szubin R, Yang L, Feist AM, Palsson BO. 2020. OxyR is a convergent target for mutations acquired during adaptation to oxidative stress-prone metabolic states. Mol Biol Evol 37:660–667. https://doi.org/10.1093/molbev/msz251.
35. Chapal M, Mintzer S, Brodsky S, Carmi M, Barkai N. 2019. Resolving noise–control conflict by gene duplication. PLoS Biol 17:e3000289. https://doi.org/10.1371/journal.pbio.3000289.
36. Bereketoglu C, Arga KY, Eraslan S, Mertoglu B. 2017. Genome reprogramming in Saccharomyces cerevisiae upon nonylphenol exposure. Physiol Genomics 49:549–566. https://doi.org/10.1152/physiolgenomics.00034.2017.
37. Mira NP, Teixeira MC, Sá-Correia I. 2010. Adaptive response and tolerance to weak acids in *Saccharomyces cerevisiae*: a genome-wide view. OMICS 14:525–540. https://doi.org/10.1089/omi.2010.0072.
38. Geng P, Zhang L, Shi GY. 2017. Omics analysis of acetic acid tolerance in Saccharomyces cerevisiae. World J Microbiol Biotechnol 33:94. https://doi.org/10.1007/s11274-017-2259-9.
39. Jarolim S, Ayer A, Pillay B, Gee AC, Phrakaysone A, Perrone GG, Breitenbach M, Dawes IW. 2013. *Saccharomyces cerevisiae* genes involved in survival of heat shock. G3 (Bethesda) 3:2321–2333. https://doi.org/10.1534/g3.113.007971.
40. Talavera D, Kershaw CJ, Costello JL, Castelli LM, Rowe W, Sims PFG, Ashe MP, Grant CM, Pavitt GD, Hubbard SJ. 2018. Archetypal transcriptional blocks underpin yeast gene regulation in response to changes in growth conditions. Sci Rep 8:7949. https://doi.org/10.1038/s41598-018-26170-5.
41. Costa V, Reis E, Quintanilha A, Moradas-Ferreira P. 1993. Acquisition of ethanol tolerance in *Saccharomyces cerevisiae*: the key role of the mitochondrial superoxide dismutase. Arch Biochem Biophys 300:608–614. https://doi.org/10.1006/abbi.1993.1084.
42. Alexandre H, Ansanay-Galeote V, Dequin S, Blondin B. 2001. Global gene expression during short-term ethanol stress in *Saccharomyces cerevisiae*. FEBS Lett 498:98–103. https://doi.org/10.1016/s0014-5793(01)02503-0.
43. van Voorst F, Houghton-Larsen J, Jønson L, Kielland-Brandt MC, Brandt A. 2006. Genome-wide identification of genes required for growth of *Saccharomyces cerevisiae* under ethanol stress. Yeast 23:351–359. https://doi.org/10.1002/yea.1359.
44. Gibson BR, Lawrence SJ, Boulton CA, Box WG, Graham NS, Linforth RST, Smart KA. 2008. The oxidative stress response of a lager brewing yeast strain during industrial propagation and fermentation. FEMS Yeast Res 8:574–585. https://doi.org/10.1111/j.1567-1364.2008.00371.x.

45. Izawa S, Inoue Y. 2009. Post-transcriptional regulation of gene expression in yeast under ethanol stress. Biotechnol Appl Biochem 53:93–99. https://doi.org/10.1042/BA20090036.
46. Teixeira MC, Mira NP, Sá-Correia I. 2011. A genome-wide perspective on the response and tolerance to food-relevant stresses in *Saccharomyces cerevisiae*. Curr Opin Biotechnol 22:150–156. https://doi.org/10.1016/j.copbio.2010.10.011.
47. López-Maury L, Marguerat S, Bähler J. 2008. Tuning gene expression to changing environments: from rapid responses to evolutionary adaptation. Nat Rev Genet 9:583–593. https://doi.org/10.1038/nrg2398.
48. Zakrzewska A, van Eikenhorst G, Burggraaff JEC, Vis DJ, Hoefsloot H, Delneri D, Oliver SG, Brul S, Smits GJ. 2011. Genome-wide analysis of yeast stress survival and tolerance acquisition to analyze the central trade-off between growth rate and cellular robustness. Mol Biol Cell 22:4435–4446. https://doi.org/10.1091/mbc.E10-08-0721.
49. Bergström A, Simpson JT, Salinas F, Barré B, Parts L, Zia A, Nguyen Ba AN, Moses AM, Louis EJ, Mustonen V, Warringer J, Durbin R, Liti G. 2014. A high-definition view of functional genetic variation from natural yeast genomes. Mol Biol Evol 31:872–888. https://doi.org/10.1093/molbev/msu037.
50. Ho C, Zhang J. 2014. The genotype-phenotype map of yeast complex traits: basic parameters and the role of natural selection. Mol Biol Evol 31:1568–1580. https://doi.org/10.1093/molbev/msu131.
51. Galardini M, Busby BP, Vieitez C, Dunham AS, Typas A, Beltrao P. 2019. The impact of the genetic background on gene deletion phenotypes in *Saccharomyces cerevisiae*. Mol Syst Biol 15:e8831. https://doi.org/10.15252/msb.20198831.
52. Nguyen Ba AN, Cvijovic I, Rojas-Echenique JI, Lawrence KR, Rego-Costa A, Liu X, Levy SF, Desai MM. 2019. High-resolution lineage tracking reveals travelling wave of adaptation in laboratory yeast. Nature 575:494–499. https://doi.org/10.1038/s41586-019-1749-3.
53. Fares MA. 2015. Survival and innovation: the role of mutational robustness in evolution. Biochimie 119:254–261. https://doi.org/10.1016/j.biochi.2014.10.019.
54. Kaboli S, Miyamoto T, Sunada K, Sasano Y, Sugiyama M, Harashima S. 2016. Improved stress resistance and ethanol production by segmental haploidization of the diploid genome in *Saccharomyces cerevisiae*. J Biosci Bioeng 121:638–644. https://doi.org/10.1016/j.jbiosc.2015.10.012.
55. Morard M, Macías LG, Adam AC, Lairón-Peris M, Pérez-Torrado R, Toft C, Barrio E. 2019. Aneuploidy and ethanol tolerance *in Saccharomyces cerevisiae*. Front Genet 10:82. https://doi.org/10.3389/fgene.2019.00082.
56. Landry CR, Oh J, Hartl DL, Cavalieri D. 2006. Genome-wide scan reveals that genetic variation for transcriptional plasticity in yeast is biased towards multi-copy and dispensable genes. Gene 366:343–351. https://doi.org/10.1016/j.gene.2005.10.042.
57. Stern S, Dror T, Stolovicki E, Brenner N, Braun E. 2007. Genome-wide transcriptional plasticity underlies cellular adaptation to novel challenge. Mol Syst Biol 3:106. https://doi.org/10.1038/msb4100147.
58. Lehner B. 2010. Conflict between noise and plasticity in yeast. PLoS Genet 6:e1001185. https://doi.org/10.1371/journal.pgen.1001185.
59. Hallin J, Landry CR. 2019. Regulation plays a multifaceted role in the retention of gene duplicates. PLoS Biol 17:e3000519. https://doi.org/10.1371/journal.pbio.3000519.
60. Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. Science 290:1151–1155. https://doi.org/10.1126/science.290.5494.1151.
61. Conant GC, Wolfe KH. 2008. Turning a hobby into a job: how duplicated genes find new functions. Nat Rev Genet 9:938–950. https://doi.org/10.1038/nrg2482.
62. Papp B, Pál C, Hurst LD. 2003. Dosage sensitivity and the evolution of gene families in yeast. Nature 424:194–197. https://doi.org/10.1038/nature01771.
63. Zhang J. 2003. Evolution by gene duplication: an update. Trends Ecol Evol 18:292–298. https://doi.org/10.1016/S0169-5347(03)00033-8.
64. Gout J-F, Kahn D, Duret L, Paramecium Post-Genomics Consortium. 2010. The relationship among gene expression, the evolution of gene dosage, and the rate of protein evolution. PLoS Genet 6:e1000944. https://doi.org/10.1371/journal.pgen.1000944.
65. Hagman A, Säll T, Compagno C, Piskur J. 2013. Yeast "make-accumulate-consume" life strategy evolved as a multi-step process that predates the whole genome duplication. PLoS One 8:e68734. https://doi.org/10.1371/journal.pone.0068734.
66. Escalera-Fanjul X, Quezada H, Riego-Ruiz L, González A. 2019. Whole-genome duplication and yeast's fruitful way of life. Trends Genet 35:42–54. https://doi.org/10.1016/j.tig.2018.09.008.
67. Blanc G, Wolfe KH. 2004. Functional divergence of duplicated genes formed by polyploidy during *Arabidopsis* evolution. Plant Cell 16:1679–1691. https://doi.org/10.1105/tpc.021410.
68. Ha M, Li WH, Chen ZJ. 2007. External factors accelerate expression divergence between duplicate genes. Trends Genet 23:162–166. https://doi.org/10.1016/j.tig.2007.02.005.
69. Ha M, Kim ED, Chen ZJ. 2009. Duplicate genes increase expression diversity in closely related species and allopolyploids. Proc Natl Acad Sci U S A 106:2295–2300. https://doi.org/10.1073/pnas.0807350106.
70. Wang Y, Wang X, Paterson AH. 2012. Genome and gene duplications and gene expression divergence: a view from plants. Ann N Y Acad Sci 1256:1–14. https://doi.org/10.1111/j.1749-6632.2011.06384.x.
71. Zou C, Lehti-Shiu M, Thomashow M, Shiu S-H. 2009. Evolution of stress-regulated gene expression in duplicate genes of *Arabidopsis thaliana*. PLoS Genet 5:e1000581. https://doi.org/10.1371/journal.pgen.1000581.
72. Mattenberger F, Sabater-Muñoz B, Toft C, Fares MA. 2017. The phenotypic plasticity of duplicated genes in *Saccharomyces cerevisiae* and the origin of adaptations. G3 (Bethesda) 7:63–75. https://doi.org/10.1534/g3.116.035329.
73. Sprouffske K, Wagner A. 2016. Growthcurver: an R package for obtaining interpretable metrics from microbial growth curves. BMC Bioinformatics 17:172. https://doi.org/10.1186/s12859-016-1016-7.
74. Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26:139–140. https://doi.org/10.1093/bioinformatics/btp616.
75. Byrne KP, Wolfe KH. 2005. The Yeast Gene Order Browser: combining curated homology and syntenic context reveals gene fate in polyploid species. Genome Res 15:1456–1461. https://doi.org/10.1101/gr.3672305.
76. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402. https://doi.org/10.1093/nar/25.17.3389.
77. Yu G, Wang L, Han Y, He Q. 2012. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS 16:284–287. https://doi.org/10.1089/omi.2011.0118.
78. Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc B 57:289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

eLife RESEARCH ARTICLE

Globally defining the effects of mutations in a picornavirus capsid

Florian Mattenberger[1], Victor Latorre[1], Omer Tirosh[2], Adi Stern[2], Ron Geller[1]*

[1]Institute for Integrative Systems Biology, I2SysBio (Universitat de València-CSIC), Paterna, Spain; [2]The Shmunis School of Biomedicine and Cancer Research, Tel-Aviv University, Tel-Aviv, Israel

Abstract The capsids of non-enveloped viruses are highly multimeric and multifunctional protein assemblies that play key roles in viral biology and pathogenesis. Despite their importance, a comprehensive understanding of how mutations affect viral fitness across different structural and functional attributes of the capsid is lacking. To address this limitation, we globally define the effects of mutations across the capsid of a human picornavirus. Using this resource, we identify structural and sequence determinants that accurately predict mutational fitness effects, refine evolutionary analyses, and define the sequence specificity of key capsid-encoded motifs. Furthermore, capitalizing on the derived sequence requirements for capsid-encoded protease cleavage sites, we implement a bioinformatic approach for identifying novel host proteins targeted by viral proteases. Our findings represent the most comprehensive investigation of mutational fitness effects in a picornavirus capsid to date and illuminate important aspects of viral biology, evolution, and host interactions.

***For correspondence:** ron.geller@uv.es

Competing interests: The authors declare that no competing interests exist.

Funding: See page 22

Received: 22 October 2020
Accepted: 11 January 2021
Published: 12 January 2021

Reviewing editor: Jan E Carette, Stanford University School of Medicine, United States

Introduction

The capsids of non-enveloped viruses are among the most complex of any viral protein. These highly multimeric structures must correctly assemble around the genome from numerous subunits, at times numbering in the hundreds, while avoiding aggregation (***Harrison, 2013***; ***Hunter, 2013***; ***Perlmutter and Hagan, 2015***). Moreover, the assembled structure must be both sufficiently stable to protect the viral genome during its transition between cells yet readily disassemble upon entry to initiate subsequent infections. For these functions to be achieved, viral capsids must encode the information for interacting with numerous cellular factors that are required to correctly fold and assemble around the genome (***Callaway et al., 2001***; ***Fields et al., 2013***; ***Geller et al., 2007***; ***Jiang et al., 2014***; ***Macejak and Sarnow, 1992***). Viral capsids also play key roles in pathogenesis, dictating host and cell tropism by encoding the determinants for binding cellular receptors (***Helenius, 2013***; ***Rossmann et al., 2002***) and mediating escape from humoral immune responses (***Cifuente and Moratorio, 2019***; ***Heise and Virgin, 2013***). As a result, viral capsids show the highest evolutionary rates among viral proteins.

The picornaviruses constitute a large group of single-stranded, positive-sense RNA viruses and include several pathogens of significant medical and economic impact (***Racaniello, 2013***). Their relative simplicity and ease of culture have made picornaviruses important models for understanding virus biology. Among the many breakthroughs achieved with these viruses was the determination of the first high-resolution structure of the capsid of an animal virus, making the picornavirus capsid the prototypical non-enveloped, icosahedral viral capsid (***Racaniello, 2013***). Picornavirus capsid genesis initiates with the co-translational release of the P1 capsid precursor protein from the viral polyprotein via the proteolytic activity of the viral encoded 2A protease (***Jiang et al., 2014***; ***Racaniello, 2013***). Subsequently, the viral encoded 3CD protease ($3CD^{pro}$) cleaves the P1 capsid precursor to liberate three capsid proteins (VP0, VP3, and VP1), generating the capsid protomer.

eLife digest A virus is made up of genetic material that is encased with a protective protein coat called the capsid. The capsid also helps the virus to infect host cells by binding to the host receptor proteins and releasing its genetic material. Inside the cell, the virus hitchhikes the infected cell's machinery to grow or replicate its own genetic material.

Viral capsids are the main target of the host's defence system, and therefore, continuously change in an attempt to escape the immune system by introducing alterations (known as mutations) into the genes encoding viral capsid proteins. Mutations occur randomly, and so while some changes to the viral capsid might confer an advantage, others may have no effect at all, or even weaken the virus.

To better understand the effect of capsid mutations on the virus' ability to infect host cells, Mattenberger et al. studied the Coxsackievirus B3, which is linked to heart problems and acute heart failure in humans. The researchers analysed around 90% of possible amino acid mutations (over 14,800 mutations) and correlated each mutation to how it influenced the virus' ability to replicate in human cells grown in the laboratory.

Based on these results, Mattenberger et al. developed a computer model to predict how a particular mutation might affect the virus. The analysis also identified specific amino acid sequences of capsid proteins that are essential for certain tasks, such as building the capsid. It also included an analysis of sequences in the capsid that allow it to be recognized by another viral protein, which cuts the capsid proteins into the right size from a larger precursor. By looking for similar sequences in human genes, the researchers identified several ones that the virus may attack and inactivate to support its own replication.

These findings may help identify potential drug targets to develop new antiviral therapies. For example, proteins of the capsid that are less likely to mutate will provide a better target as they lower the possibility of the virus to become resistant to the treatment. They also highlight new proteins in human cells that could potentially block the virus in cells.

Five protomers then assemble to form the pentamer, twelve of which assemble around the viral genome to yield the virion. Finally, in some picornaviruses, VP0 is further cleaved into two subunits, VP4 and VP2, following genomic encapsidation to generate the infectious, 240 subunit particles (***Jiang et al., 2014***; ***Racaniello, 2013***). Work over the years has identified numerous host factors that help support capsid formation (***Corbic Ramljak et al., 2018***; ***Geller et al., 2007***; ***Macejak and Sarnow, 1992***; ***Qing et al., 2014***; ***Thibaut et al., 2014***), defined antibody neutralization sites (***Cifuente and Moratorio, 2019***), and identified numerous host receptors for many members of this viral family (***Rossmann et al., 2002***).

Despite significant progress in understanding the structure and function of picornavirus capsids, a comprehensive understanding of how mutations affect viral fitness across different structural and functional attributes is lacking. To address this, we perform a comprehensive analysis of mutational fitness effects (MFE) across the complete capsid region of the human picornavirus coxsackievirus B3 (CVB3), analyzing >90% of all possible single amino acid mutations. Furthermore, using these data, we develop models to predict the effect of mutations with high accuracy from available sequence and structural information, improve evolutionary analyses of CVB3, and define the sequence preferences of several viral encoded motifs. Finally, we use the information obtained in our dataset for the sequence requirements of capsid-encoded 3CD protease cleavage sites to identify host targets of this viral protease. Overall, our data comprise the most comprehensive survey of MFE effects in a picornavirus capsid to date and provide important insights into virus biology, evolution, and interaction with the host.

Results

Deep mutational scanning of a CVB3 capsid

To generate CVB3 libraries encoding a large amount of diversity in the capsid region, we used a codon-level PCR mutagenesis method (***Bloom, 2014***). The mutagenesis protocol was performed on

the capsid precursor region P1 in triplicate to generate three independent mutagenized libraries (Mut Library 1–3; ***Figure 1A***). From these, three independent viral populations (Mut Virus 1–3) were derived by electroporation of in vitro transcribed viral RNA into HeLa-H1 cells (***Figure 1A***). High-fidelity next-generation sequencing (***Schmitt et al., 2012***) was then used to analyze the mutagenized libraries and resulting viruses, unmutagenized virus populations (WT virus 1–2), as well as controls for errors occurring during PCR (PCR) and reverse transcription (RT-PCR). High coverage was obtained for all samples (>10^6 per codon across all experimental conditions and >6.5×10^5 for the controls; ***Supplementary file 2***). Due to the high rate of single mutations within codons observed in the RT-PCR control compared to the mutagenized virus populations (***Supplementary file 2***), all single mutants were omitted from our analysis to increase the signal-to-noise ratio. While this resulted in an inability to analyze 83.4% of synonymous codons in the capsid region (1746/2094), only 2.8% of non-synonymous mutations were lost to analysis (458/16,169). Upon removing single mutations within codons, we obtained a large signal-to-noise ratio in the average mutation rate of 510× (range 449–572) and 245× (range 174–285) for the mutagenized libraries and viruses, respectively, compared to their error controls (***Figure 1B*** and ***Supplementary file 2***). On average, 0.9 (range 0.8–1.02) codon mutations were observed per genome, which was in agreement with Sanger sequencing of 59 clones (range 18–23 per library; ***Figure 1—figure supplement 1*** and ***Supplementary file 3***). As expected, the rate of stop codons, which should be invariably lethal in the CVB3 capsid, decreased significantly following growth in cells to <0.5% of that observed in the corresponding mutagenized libraries ($p<0.005$ by paired t-test on log-transformed data; ***Supplementary file 2***). No major bias was observed in the position within a codon where mutations were observed (***Figure 1—figure supplement 2***) or in the type of mutation (***Figure 1—figure supplement 2***), except for the WT virus, which had a high rate of A to G transitions in the two independent replicates analyzed. Of all 16,169 possible amino acid mutations in the capsid region (851 AA × 19 AA mutation = 16,169), a total of 14,839 amino acid mutations were commonly observed in all three mutagenized libraries, representing a 91.8% of all possible amino acid mutations in the capsid region, allowing us to globally assess the effects of the vast majority of amino acid mutations on the capsid (***Figure 1C***).

MFE across the CVB3 capsid

We next derived the MFE of each observed mutation by examining how its frequency changed relative to that of the WT sequence following growth in cells. The preferences for the different amino acids at each position (amino acid preferences [***Bloom, 2015***]) showed a high correlation between biological replicates (Spearman's $\rho > 0.83$; ***Figure 2—figure supplement 1*** and ***Supplementary file 4*** MFE). Overall, most mutations in the capsid were deleterious to growth in cell culture, with only 1.2% of mutations increasing fitness relative to the WT amino acid (***Figure 2A*** and ***Supplementary file 4***; Interactive heatmap available at https://rgellerlab.github.io/CVB3_capsid_DMS_Interactive_Heatmap/). Hotspots where mutations were tolerated were observed at several regions across the capsid (***Figure 2A***). These hotspots largely overlapped with highly variable regions in natural sequences, as measured by Shannon entropy in the enterovirus B family, indicating that lab measured MFE reflect natural evolutionary processes (***Figure 2A***, top). Indeed, a strong correlation was observed between the average MFE observed at each site and sequence variability for the enterovirus B genus (Spearman's $\rho = 0.59$, $p<10^{-16}$; ***Figure 2B***). Similarly, antibody neutralization sites overlapped with hotspots for mutations (***Figure 2A***, top), with individual mutations in antibody neutralization sites showing lower MFE ($p<10^{-16}$ by Mann–Whitney test; ***Figure 2C***). As expected, mutations were also less deleterious in loops compared to β-strands ($p<10^{-16}$ by Mann–Whitney test; ***Figure 2D***), at surface residues compared to core residues ($p<10^{-16}$ by Mann–Whitney test; ***Figure 2E***), and for mutations predicted to be destabilizing or aggregation-prone ($p<10^{-16}$ by Mann–Whitney test for both; ***Figure 2F***). Importantly, independent validation of the MFE of 10 different mutants using a sensitive qPCR-based competition assay (***Moratorio et al., 2017***) showed a strong correlation with the deep mutational scanning (DMS) results (Spearman's $\rho = 0.9$, $p<0.001$; ***Figure 2G*** and ***Supplementary file 5***). It is important to note that laboratory-measured MFE may not always reflect those in nature due to differences in the environments.

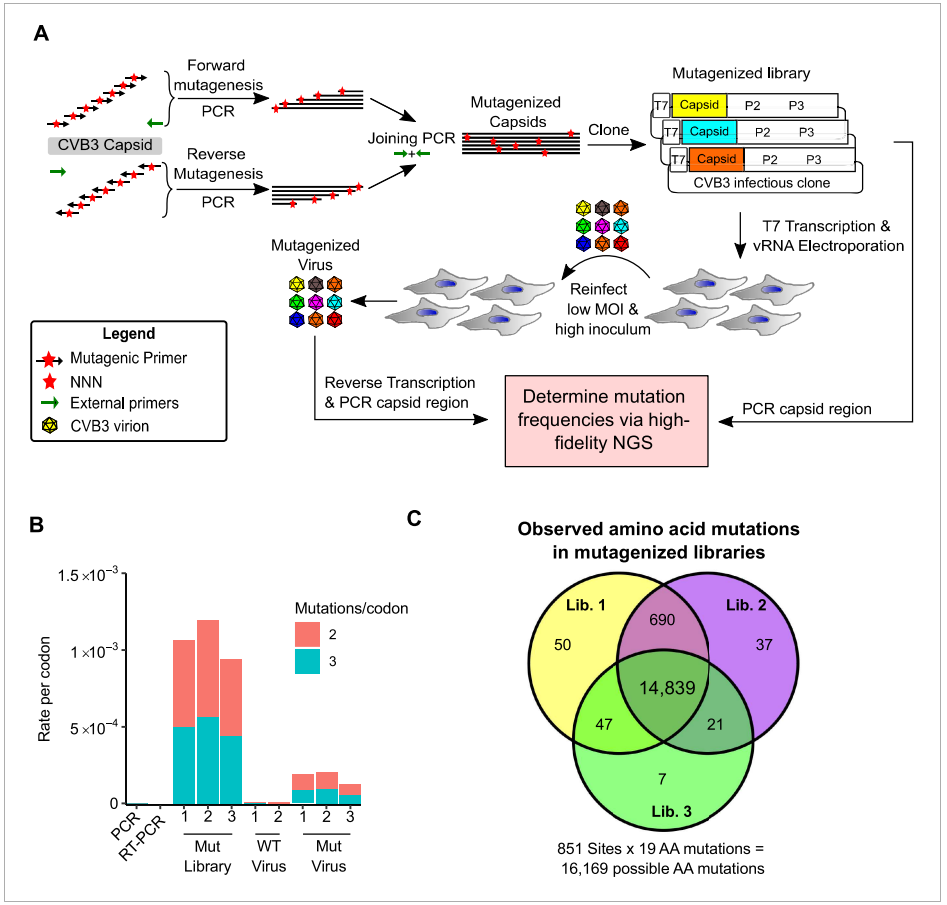


Figure 1. Deep mutational scanning (DMS) of the CVB3 capsid. (**A**) Overview of the deep mutational scanning experimental approach. A mutagenesis PCR was used to introduce all possible single amino acid mutations across the CVB3 capsid region (Mut Library 1–3). Viral genomic RNA (vRNA) produced from the mutant libraries was then electroporated into cells to generate high diversity CVB3 populations (Mut Virus 1–3). The frequency of each mutation relative to the WT amino acid was then determined in both the mutagenized libraries and the resulting virus populations via high-fidelity duplex sequencing. (**B**) The average rate of double or triple mutations per codon observed in the mutagenized libraries (Mut Library 1–3), the resulting mutagenized virus (Mut Virus 1–3), as well as controls for the error rate of the amplification and sequencing process (PCR and RT-PCR) or the WT unmutagenized virus (WT Virus 1–2). Single mutations per codon were omitted from the analysis to increase the signal-to-noise ratio. (**C**) Venn diagram showing the number of amino acid mutations observed in the mutagenized libraries. MOI: multiplicity of infection. NGS: next-generation sequencing. The online version of this article includes the following figure supplement(s) for figure 1:

Figure supplement 1. Sanger analysis of DMS libraries.
Figure supplement 2. Results of high-fidelity duplex sequencing.

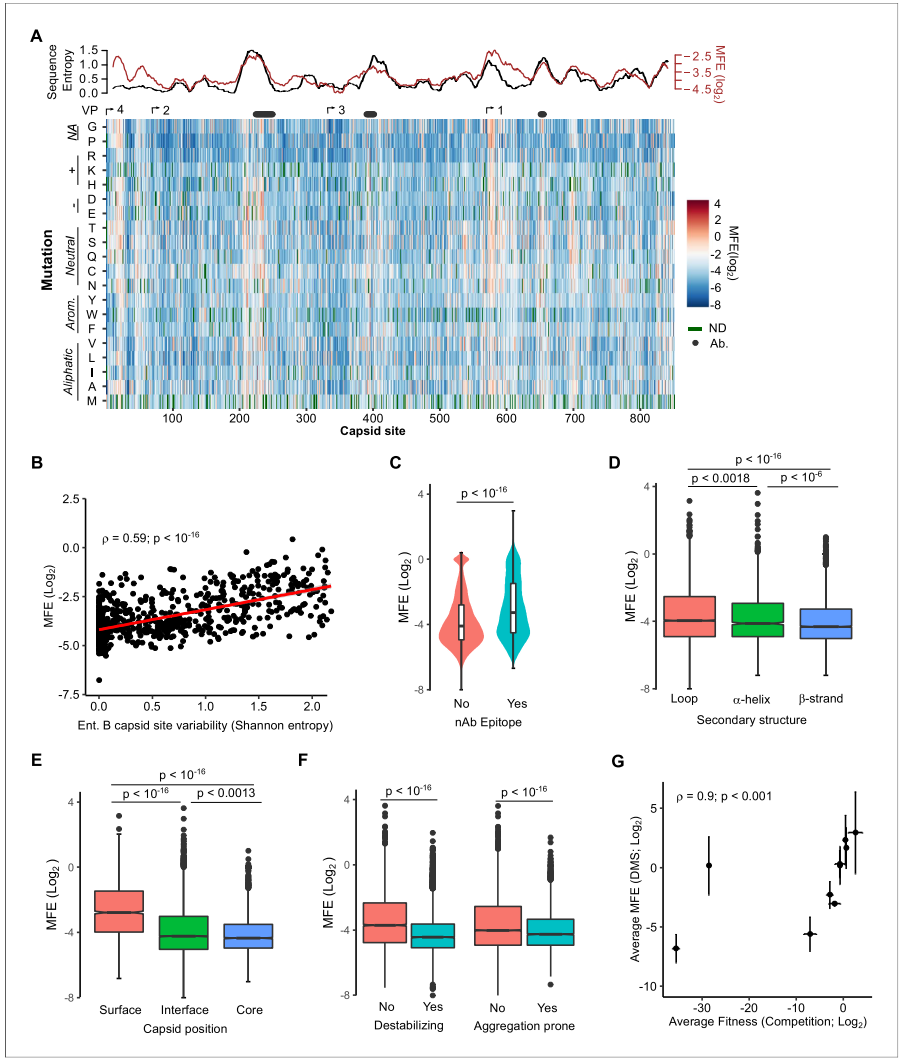


Figure 2. Mutational fitness effects (MFE) across the CVB3 capsid and their correlation with structural, evolutionary, and immunological attributes. (**A**) Overview of the MFE observed across the CVB3 capsid. Bottom: A heatmap representing the MFE of all mutations observed at each capsid site. Green indicates no data available (ND), and the positions of the mature viral proteins (VP1–4) or antibody neutralization sites (nAb) are indicated above. Top: A 21 amino acid sliding window analysis of the average sequence variation in enterovirus B genomes (Shannon entropy; black line) or a 21 amino acid

Figure 2 continued on next page

Figure 2 continued

sliding window of the average MFE observed at each capsid site (red line). (**B**) Correlation between the average MFE observed at each capsid site and variation in enterovirus B sequence alignments (Shannon entropy). (**C**) Violin plot of MFE in antibody neutralization sites versus other capsid sites. (**D–F**) Boxplots of MFE as a function of secondary structure (**D**), position in the capsid (**E**), or the predicted effect of mutations on stability or aggregation propensity (**F**). (**G**) Validation of the MFE obtained by DMS using a competition assay. For each mutant, the average and standard deviation of the MFE obtained by DMS (n = 3) is plotted against the average and standard deviation of the fitness derived using the competition assay (n = 4). A two-sided Mann–Whitney test was used for two- category comparisons.

The online version of this article includes the following figure supplement(s) for figure 2:

Figure supplement 1. Correlation of amino acid preferences observed in experimental replicates.

Prediction of MFE from available structural and sequence information

As MFE correlated with natural sequence variation and different structural features of the capsid (***Figure 2***), we next investigated if MFE could be predicted from available structural and sequence information. For this, we obtained a dataset of 52 parameters, including structural information derived from the crystal structure of the CVB3 capsid (PDB:4GB3), amino acid properties, and natural variation in available enterovirus sequences (Shannon entropy), and predicted the effects of mutation on stability and aggregation propensity using FoldX (***Schymkowitz et al., 2005***) and TANGO (***Fernandez-Escamilla et al., 2004***), respectively (***Supplementary file 6***). We then employed a random forest algorithm to identify the parameters that can best predict MFE, limiting our analysis to sites that present in the crystal structure and where mutations were observed in at least two replicates to improve accuracy (total of 9685 mutations). Overall, a model trained on 70% of the dataset was able to predict the remaining 30% of the data (2905 mutations) with high accuracy (Spearman's $\rho > 0.75$, Pearson's $r = 0.76$; $p<10^{-16}$; ***Figure 3—figure supplement 1***). Surprisingly, a random forest model trained on the top five predictors alone showed similar accuracy (Spearman's $\rho = 0.73$, Pearson's $r = 0.73$; $p<10^{-16}$; ***Figure 3B***). Excluding natural sequence variation, amino acid identity, or structural attributes reduced model predictability significantly (>20%; data not shown), suggesting

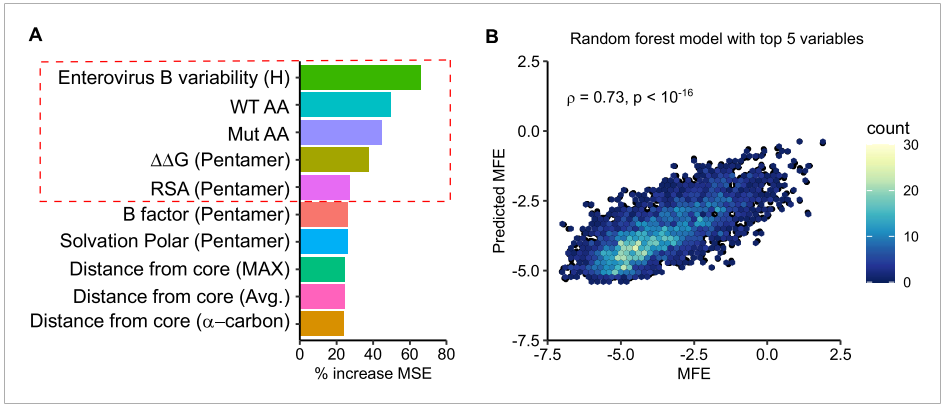


Figure 3. Prediction of MFE based on structural and sequence information. (**A**) The top 10 predictors identified in a random forest model for explaining MFE in the CVB3 capsid based on the percent of mean squared error (MSE) increase. (**B**) Hexagonal plot showing the correlation between MFE predicted using a random forest algorithm trained on the top five variables versus observed MFE. The random forest model was trained on 70% of the data and then tested on the remaining 30% (shown). RSA, relative surface area.

The online version of this article includes the following figure supplement(s) for figure 3:

Figure supplement 1. Prediction of mutational fitness effects using random forest or linear models.

a combination of evolutionary, sequence, and structural information best explains MFE. Using an alternative approach, we were able to predict the data with slightly lower accuracy using a linear model with the same five predictors ($p<10^{-16}$, Spearman's $\rho = 0.67$, Pearson's $r = 0.67$; ***Figure 3—figure supplement 1***). Together, these results suggest that the prediction of MFE in the CVB3 capsid can be achieved at relatively high accuracy based on available structural and sequence information. Due to the high conservation of capsid structure in picornaviruses, as well as the availability of numerous capsid sequences and structures, these findings are likely generalizable to related picornaviruses.

Experimentally measured MFE inform of natural evolutionary processes

We next examined if our experimentally measured MFE could improve phylogenetic models of CVB3 evolution by incorporating site-specific amino acid preferences using PhyDMS (***Hilton et al., 2017***). Indeed, significant improvement in model fit was observed (***Table 1*** PHY; $p<10^{-16}$ using a log-likelihood test compared to non-site-specific codon models), supporting the relevance of our results to understanding evolutionary processes in nature. Nevertheless, selection in nature was significantly more stringent than in the lab ($\beta = 2.18$), indicating the presence of additional selection pressures. As laboratory conditions lack selection from antibodies, we used the sum of the absolute differential selection observed at each site (***Bloom, 2017***) to examine whether known antibody neutralization sites show differential selection between the two environments (***Supplementary file 7***). Indeed, antibody neutralization sites showed significantly higher differential selection values compared to other residues ($p<10^{-6}$ by Mann–Whitney test; ***Figure 4A***). Moreover, the three sites showing the strongest overall differential selection were found in known antibody neutralization sites: positions 226 and 242 in the EF loop (residues 157 and 173 of VP2) and position 650 in the BC loop (residue 80 of VP1; ***Figure 4B–D*** and ***Supplementary file 7***). In summary, incorporation of experimentally derived amino acid preferences into phylogenetic analyses significantly improved model fit and identified residues in antibody neutralization sites that show differential selection, suggesting these may play important roles in immune evasion in vivo.

Insights into capsid-encoded motifs: myristoylation and protease cleavage

Picornavirus capsids undergo a complex assembly path to generate the infectious particle. These include myristoylation, cleavage by the viral proteases 2A and $3CD^{pro}$, as well as interaction with cellular chaperones and glutathione (***Corbic Ramljak et al., 2018***; ***Geller et al., 2007***; ***Jiang et al., 2014***; ***Qing et al., 2014***; ***Thibaut et al., 2014***; ***Figure 5A***). Having obtained a comprehensive dataset for MFE across the capsid, we next examined the sequence requirements for several of these capsid-encoded motifs. Specifically, myristoylation of the N-terminal glycine is essential for virion assembly (***Corbic Ramljak et al., 2018***). In agreement with this, the N-terminal glycine in the CVB3 capsid showed the strongest average fitness cost upon mutation in the capsid (***Figure 4—figure supplement 1*** and ***Supplementary file 4***). The remaining sites in the myristoylation motif agreed with the canonical myristoylation motif in cellular proteins (Prosite pattern PDOC00008) (***Bologna et al., 2004***), albeit with increased selectivity at three of the six positions (***Figure 4—figure supplement 1***). On the other hand, a conserved WCPRP motif in the C-terminal region of VP1 that was shown to be important for $3CD^{pro}$ cleavage of the related foot and mouth disease virus capsid (FDMV; YCPRP motif) (***Kristensen and Belsham, 2019***) was found to be intolerant to mutations

Table 1. Incorporation of DMS results in evolutionary models better describes natural CVB3 evolution compared to standard codon models.

Model	ΔAIC	Log-likelihood	Parameters	Parameter values
ExpCM	0.00	−14,580.51	6	Beta = 2.18, kappa = 7.47, omega = 0.16
Goldman-Yang M5	4187.56	−16,668.29	12	Alpha_omega = 0.30, beta_omega = 10.00, kappa = 7.15
Averaged ExpCM	4303.74	−16,732.38	6	Beta = 0.61, kappa = 7.55, omega = 0.02
Goldman-Yang M0	4371.26	−16,761.14	11	Kappa = 7.14, omega = 0.02

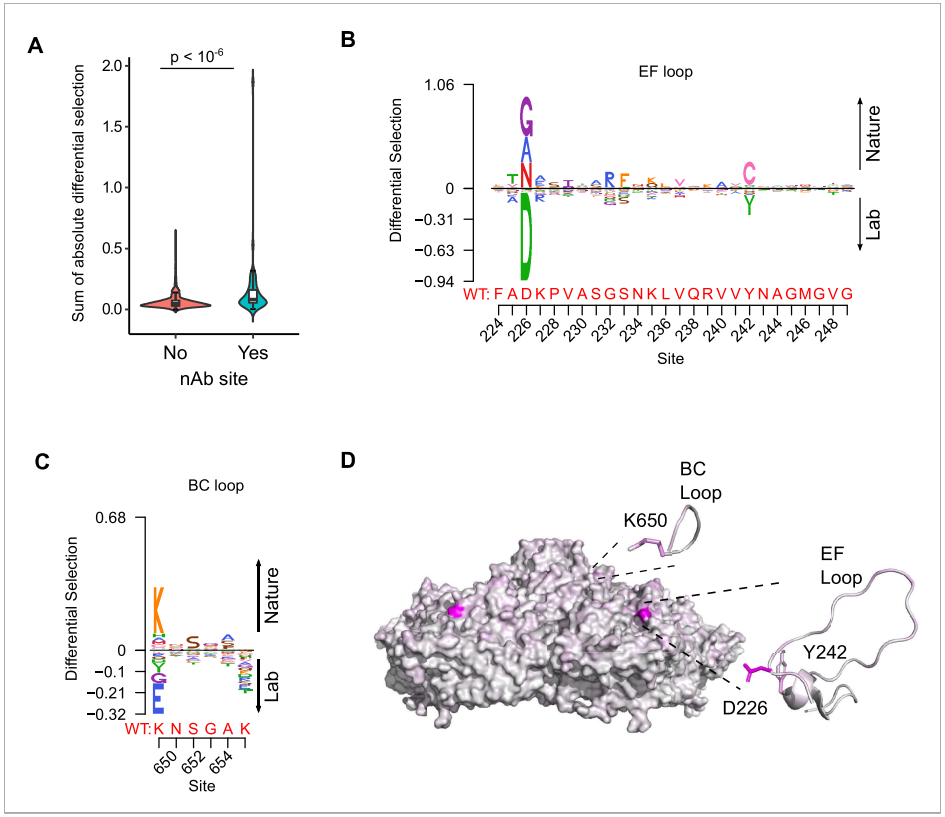


Figure 4. Antibody neutralization sites show differential selection between laboratory conditions and nature. (**A**) Violin plot showing the sum of absolute differential selection observed at capsid sites comprising antibody neutralization epitopes (nAb) versus all other capsid sites. (**B–C**) Logoplots showing the observed differential selection of sites in the EF loop or BC loop. The WT sequence is indicated in red. (**D**) The CVB3 capsid pentamer (PDB:4GB3), colored according to the amount of differential selection. The BC and EF loops are shown next to the structure together with the side chains for sites showing the highest differential selection.

The online version of this article includes the following figure supplement(s) for figure 4:

Figure supplement 1. Sequence preferences of capsid-encoded motifs.

compared to other capsid residues ($p<0.05$ versus all other positions by Mann–Whitney test; sites 815–819 in CVB3). Moreover, within this motif, the sites showing the highest average fitness cost in our DMS dataset were identical to analogous positions in FMDV that resulted in a loss of viability upon mutation to alanine (***Figure 4—figure supplement 1***; ***Kristensen and Belsham, 2019***), highlighting the conservation of this motif across different picornaviruses.

The viral 3C protease ($3C^{pro}$) cleaves the picornavirus capsid at two conserved glutamine–glycine (QG) pairs to liberate the viral capsid proteins VP0, VP3, and VP1 (***Figure 5A***). Previous work has

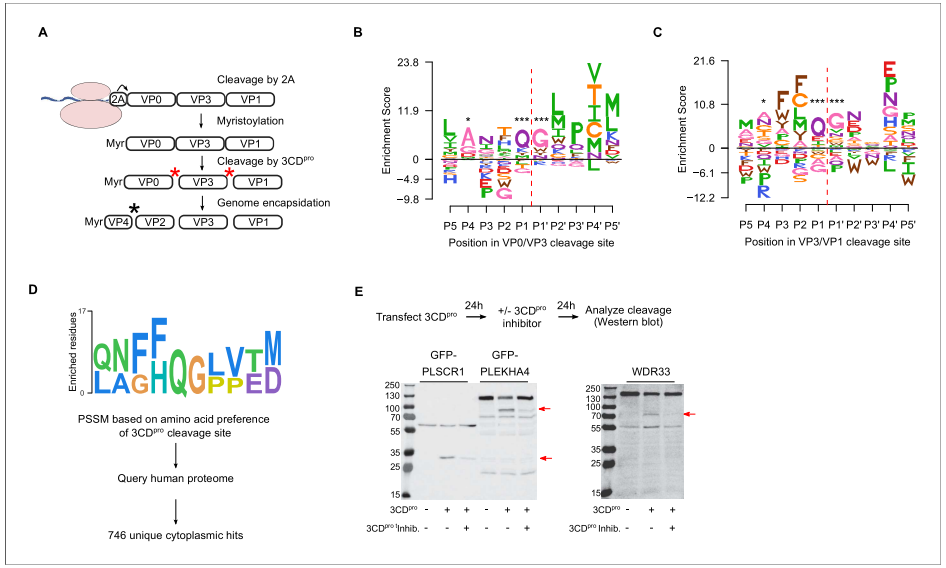


Figure 5. Sequence preference of capsid 3CD^{pro} cleavage sites and their use for the identification of novel cellular targets of the viral protease. (**A**) Overview of the CVB3 capsid maturation pathway. The CVB3 capsid precursor P1 is co-translationally cleaved by the viral 2A protease. P1 is then myristoylated and cleaved by the viral 3CD^{pro} to generate the capsid proteins VP0, VP3, and VP1. Finally, upon assembly and genome encapsidation, VP0 is further cleaved into VP4 and VP2 in a protease-independent manner to generate the mature capsid. Red and black asterisks indicated 3CD^{pro} or protease-independent cleavage events, respectively. (**B,C**) Logoplots showing amino acid preferences for the 10 amino acid regions spanning the 3CD^{pro} cleavage sites (P1–P'1) of both VP0/VP3 and VP3/VP1 in the DMS dataset. (**D**) Overview of the bioinformatic pipeline for identification of novel 3CD^{pro} cellular targets using the amino acid preferences for the capsid cleavage sites from our DMS study. A position-specific scoring matrix (PSSM) was generated based on the amino acid preferences for the 10 amino acid regions spanning the two 3CD^{pro} cleavages sites. This PSSM was then used to query the human genome for potential cellular targets, and non-cytoplasmic proteins were filtered out, yielding 746 proteins. (**E**) The cellular proteins PLSCR1, PLEKHA4, and WDR33 are cleaved by 3CD^{pro}. Western blot analysis of cells cotransfected with 3CD^{pro} and GFP-PLSCR1 or GFP-PLEKHA4 and probed with a GFP antibody or transfected with 3CD^{pro} and probed using a WDR33 antibody. When indicated, the 3CD^{pro} inhibitor rupintrivir was included to ensure cleavage was mediated by the viral protease. Red arrows indicate cleavage products of the expected size (GFP-PLSCR1 full length = 64 kDa, cleaved N-terminus = 36 kDa; GFP-PLEKHA4 full length = 118 kDa, cleaved N-terminus = 72 kDa; WDR33 full length = 146 kDa, cleaved N-terminus = 72 kDa). *p<0.05, ***p<0.001.

The online version of this article includes the following figure supplement(s) for figure 5:

Figure supplement 1. Evaluation of select hits identified as potential 3CD^{pro} target proteins.

defined the sequence specificity of several picornavirus 3C^{pro} enzymes by examining both natural sequence variation and in vitro cleavage assays using synthetic peptides (***Laitinen et al., 2016***). However, unlike other 3C^{pro}-mediated cleavage events in the viral polyprotein, the capsid is only efficiently cleaved by the precursor protein 3CD^{pro} (***Ypma-Wong et al., 1988***). To gain insights into the sequence specificity of 3CD^{pro}, we examined the amino acid preferences for a 10 amino acid region surrounding the protease cleavage site (P5–P5'). As expected based on the known specificity of the 3C protease (***Laitinen et al., 2016***), a strong preference for the presence of QG was observed at both 3CD^{pro} cleavage sites in our dataset (positions P1 and P1' in the cleavage site; ***Figure 5B,C***). Interestingly, significant correlation in amino acid preferences between the two cleavage sites was observed only at P1–P1' (Pearson's $\rho > 0.99$, $p<10^{-16}$) and P4 (Pearson's $\rho > 0.49$, $p<0.05$), as was the case in the enterovirus B alignments (Pearson's $\rho > 0.84$ and $p<10^{-6}$ for positions P4, P1, and

P'1; data not shown). Hence, the low agreement in amino acid preferences observed for most positions across the two 3CD^{pro} cleavage sites suggests cleavage is strongly dictated by positions P4, P1, and P1'.

Identification of 3CD^{pro} cellular targets based on the sequence preferences of capsid-encoded protease cleavage sites

In addition to cleaving the viral polyprotein, the picornavirus proteases cleave cellular factors to facilitate viral replication, including both antiviral factors and cellular factors that favor viral IRES-driven translation mechanism over cellular cap-dependent translation (e.g. DDX58, eIF4G, and PABP) (***Laitinen et al., 2016***; ***Sun et al., 2016***). As the canonical 3C/3CD^{pro} QG cleavage site occurs on average 1.6 times per protein in the human proteome (~33,000,000 times), we sought to examine whether the rich dataset we obtained for the amino acid preferences of the capsid 3CD^{pro} cleavage sites can be used to identify novel cellular factors that are targeted by the viral protease. Specifically, a position-specific score matrix (PSSM) was generated for the 10 amino acid regions spanning the two protease cleavage sites in the CVB3 capsid (P5–P5') based on the amino acid preferences identified in our study (***Figure 5D***). This PSSM was then used to query the human proteome for potential cleavage sites, yielding a total of 746 cytoplasmic proteins (***Figure 5D***; ***Supplementary file 8***). Eleven cellular factors that are known to be cleaved during enterovirus infection were identified using this approach, including the viral sensor Probable ATP-dependent RNA helicase DDX58 (RIG1), the immune transcription factors p65 (RELA) and interferon regulatory factor 7 (IRF7), and polyadenylate-binding protein 1 (PABPC1), an important factor in translation initiation and mRNA stability (***Supplementary file 8***; ***Jagdeo et al., 2018***; ***Laitinen et al., 2016***).

To evaluate whether our approach can identify novel cellular targets for the viral protease, we examined the ability of 3CD^{pro} to cleave eight different proteins found in the data set, focusing on those with cellular functions of potential relevance to CVB3 biology and which could be readily detected in our cell culture assay (e.g. availability of antibodies or tagged-variants, cleavage fragments of observable size, and high expression level). These included four interferon-inducible proteins (Pleckstrin homology domain containing A4, PLEKHA4; phospholipid scramblase 1, PLSCR1; NOD-like receptor family CARD domain containing 5, NLRC5; zinc finger CCCH-type containing, antiviral 1, ZC3HAV1) and four proteins involved in various cellular functions, namely apoptosis (MAGE family member D1, MAGED1), RNA processing (WD repeat domain 33, WDR33), and vesicle transport (cyclin G-associated kinase, GAK; tumor susceptibility 101, TSG101). Of these, three proteins were cleaved upon expression of the viral protease to generate fragments of the expected size (PLSCR1, PLEKHA4, and WDR33; ***Figure 5E*** and ***Supplementary file 8***). Of note, while WDR33 was predicted to harbor two potential cleavage sites, only a single cleavage event was observed. Treatment with a specific 3CD^{pro} inhibitor, rupintrivir (***Dragovich et al., 1999***), blocked the cleavage of these proteins, indicating the effect was due to the viral protease (***Figure 5D***). In contrast, five of the proteins were found to not be cleaved upon 3CD^{pro} expression, suggesting additional determinants are involved in the cleavage of host factors (***Figure 5—figure supplement 1***). Hence, our approach correctly identified 30% of the predicted cleavage sites (three of the nine different cleavage sites), indicating a strong enrichment of cellular targets of the 3CD^{pro} in the dataset.

Discussion

The picornavirus capsid is a highly complex structure that plays key roles in viral biology and pathogenesis. In the current study, we employ a comprehensive approach to define the effects of single amino acid mutations in the CVB3 capsid, measuring the effects of >90% of all possible mutations. We find that most mutations in the capsid are deleterious to growth in cell culture, with very few mutations showing higher fitness than the WT sequence (1.2% of all mutations). Similar results have been reported in other non-enveloped capsid proteins (***Acevedo et al., 2014***; ***Hartman et al., 2018***; ***Ogden et al., 2019***) as well as non-capsid viral proteins (***Ashenberg et al., 2017***; ***Bloom, 2014***; ***Doud and Bloom, 2016***; ***Du et al., 2016***; ***Haddox et al., 2016***; ***Hom et al., 2019***; ***Thyagarajan and Bloom, 2014***; ***Wu et al., 2015***). In light of these results, it is likely that the large population sizes of RNA viruses help maintain viral fitness in the face of high mutation rates and strong mutational fitness costs. It is important to note that the effect of a particular mutation on

fitness observed under laboratory conditions may not always reflect its effect in nature due to inherent differences between these two environments.

Investigation of the factors that influence MFE in the capsid revealed a strong correlation with various structural and functional attributes. These included computationally predicted effects on stability and aggregation propensity, secondary structure, and surface exposure (***Figure 2***). Surprisingly, we find that MFE can be predicted with relatively high accuracy using only five parameters: natural sequence variation, the identity of the original and mutant amino acids, the predicted effect on protein stability, and relative solvent accessibility (***Figure 3***). A recent study examined the ability of 46 different variant effect prediction tools to predict MFE from 31 different DMS datasets of both viral and non-viral proteins (***Livesey and Marsh, 2020***). Overall, viral proteins showed the lowest predictability (Spearman's correlation of <0.5). In contrast, we were able to predict MFE using a random forest model using these above-mentioned five parameters with an accuracy similar to the best prediction obtained in this analysis for any viral or non-viral protein (Pearson's r = 0.73; Spearman's ρ = 0.73; ***Figure 3B***). Interestingly, SNAP2 (***Hecht et al., 2015***), a neural network-based classifier of mutational effects that was shown to correlate well with MFE in other studies (***Gray et al., 2018***; ***Livesey and Marsh, 2020***; ***Reeb et al., 2020***), correlated poorly with our data (R^2 = −0.26). Overall, considering the relative conservation of capsid structure in picornaviruses as well as the availability of both capsid sequences and high-resolution structures for numerous members of this family, it is likely that these findings can be extrapolated to additional picornaviruses.

Incorporating site-specific amino acid preferences obtained from our DMS results into phylogenetic models was found to significantly improve model accuracy. This has been observed in DMS studies with other RNA viruses (***Bloom, 2017***; ***Doud and Bloom, 2016***; ***Haddox et al., 2018***) and indicated that our laboratory-measured MFE capture additional information that cannot be obtained from sequence analysis alone. In addition, this approach allowed us to assess which sites show differential selection patterns as a result of the distinct environments encountered in nature and the laboratory. As expected, pressure from the adaptive immune system was found to be the major difference between these environments, with residues in antibody neutralization sites showing higher differential selection compared to other sites in the capsid (***Figure 4A***). Moreover, the sites showing the highest degree of differential selection were found in known antibody neutralization sites (***Figure 4B–D***). However, why these particular residues within antibody neutralization sites show differential selection, while others do not, remain to be elucidated. It has been shown that one, or a few, sites within antibody binding regions can have strong effects on escape from antibody neutralization (***Lee et al., 2019***), potentially explaining these findings. Interestingly, while the top three sites showing differential selection were in antibody neutralization sites, the mutation showing the fourth-highest differential selection was found in the HI loop of VP1. While not classically considered an antibody epitope, this loop has been shown to interact with an antibody fragment in the picornavirus coxsackievirus A6 (***Xu et al., 2017***), is known to mediate receptor binding in different picornaviruses (***Belnap et al., 2000***; ***Xing et al., 2000***), and to interact with host cyclophilin A to facilitate uncoating (***Qing et al., 2014***). Whether these factors or others are responsible for the observed differential selection remains to be elucidated.

The CVB3 capsid encodes the information for directing myristoylation, protease cleavage, and interaction with host factors. We took advantage of our data to examine the sequence specificity and mutational tolerance of several known capsid-encoded motifs. First, we examined the amino acid preferences of the CVB3 capsid myristoylation motif. We observe a strong correlation with the canonical myristoylation pattern (Prosite pattern PDOC00008), although with greater intolerance to mutations in three of the six residues in the capsid (***Figure 4—figure supplement 1***). This is likely to stem from additional constraints imposed by capsid structure. On the other hand, we examined the amino acid preference of a conserved motif in VP1 that is required for $3CD^{pro}$-mediated cleavage of picornavirus capsids (***Kristensen and Belsham, 2019***). Our data showed a higher cost to mutation in this motif relative to other capsid positions (***Figure 4—figure supplement 1***), highlighting its importance for capsid function. Finally, we examined the sequence preferences surrounding the two $3CD^{pro}$ cleavage sites. We find a strong dependence on the cleavage site residues (positions P1 and P1′; ***Figure 5***) and to a lesser degree position P4, with large variation in the sequence preferences across the remaining positions between the two cleavage sites. Overall, our experimentally measured MFE are congruent with existing information regarding the sequence preferences of the

examined capsid motifs, yet provide in-depth insights into sequence specificity that cannot be obtained from examining natural sequence variation.

Finally, we used the amino acid preferences observed in $3CD^{pro}$ cleavage sites within the capsid to query the human genome for potential cellular targets of this protease (***Figure 5D***). Using this approach, we identify 746 cytoplasmic proteins that harbor a potential $3CD^{pro}$ target sequence, including 11 proteins previously shown to be cleaved by different picornavirus 3C proteases. We then validated our approach using eight proteins, comprising nine predicted cleavage sites. Six of the predicted cleavage sites were not affected by $3CD^{pro}$ expression (***Figure 5—figure supplement 1***). On the other hand, three proteins were observed to be specifically cleaved by the viral protease (***Figure 5E***): WD repeat domain 33 (WDR33), an important factor for polyadenylation of cellular pre-mRNAs (***Chan et al., 2014***) that has been shown to act as a restriction factor during influenza infection (***Brass et al., 2009***); the interferon-induced protein phospholipid scramblase 1 (PLSCR1), which is involved in the replication of numerous viruses, likely due to its ability to enhance the expression of certain interferon-stimulated genes (***Kodigepalli et al., 2015***); and the interferon-induced Pleckstrin homology domain containing A4 (PLEKHA4), a plasma membrane-localized signaling modulator (***Shami Shah et al., 2019***) that is currently not known to play a role in viral infection. Overall, our approach correctly predicts 30% of the identified cleavage sites. It is likely that incorporating additional selection criteria, such as accessibility of the cleavage peptide in the folded structure, can be used to further reduce false positives. Nevertheless, extrapolating our validation results to the larger dataset suggests >200 new host targets of the protease are identified, some of which could play key roles in viral biology and pathogenesis.

Materials and methods

Key resources table

Reagent type (species) or resource	Designation	Source or reference	Identifiers	Additional information
Strain, strain background (coxsackievirus B3)	pCVB3-XhoI-P1-Kpn21	10.1016/j.celrep.2019.09.014		Infectious CVB3 clone based on the Nancy strain (Taxon identifier 103903)
Strain, strain background (coxsackievirus B3)	pCVB3-XhoI-ΔP1-Kpn21	This paper		Infectious CVB3 clone without P1 region
Strain, strain background (coxsackievirus B3)	Marked reference CVB3 virus	10.1038/nmicrobiol.2017.88		Infectious CVB3 clone with silent mutations in the polymerase region used as a reference for fitness assays
Strain, strain background (*Escherichia coli*)	NZY5α	NZY Tech	MB004	Competent cells, standard cloning
Strain, strain background (*Escherichia coli*)	MegaX DH10B T1R Electrocomp cells	ThermoFisher	C6400-03	Electrocompetent cells, library cloning
Cell line (*Homo sapiens*)	HeLa-H1	ATCC	CRL-1958; RRID:CVCL_3334	Cell line for CVB3 infection and DMS library production
Cell line (*Homo sapiens*)	HEK293	ATCC	CRL-1573; RRID:CVCL_0045	Cell line used for production of CVB3 mutants and for protease cleavage
Antibody	Anti-GFP (Mouse monoclonal)	SantaCruz	Sc-9996	Western blot (1:2000)

Continued on next page

Continued

Reagent type (species) or resource	Designation	Source or reference	Identifiers	Additional information
Antibody	Anti-FLAG (Mouse monoclonal)	SantaCruz	Sc-166335	Western blot (1:2000)
Antibody	Anti-HA (Mouse monoclonal)	SantaCruz	Sc-7392	Western blot (1:2000)
Antibody	Anti-WDR33 (Mouse monoclonal)	SantaCruz	Sc-374466	Western blot (1:1000)
Antibody	Anti-TSG101 (Mouse monoclonal)	SantaCruz	Sc-136111	Western blot (1:1000)
Antibody	Anti-GAK (Mouse monoclonal)	SantaCruz	Sc-137053	Western blot (1:1000)
Antibody	Anti-MAGED1 (Mouse monoclonal)	SantaCruz	Sc-393291	Western blot (1:1000)
Recombinant DNA reagent	DMS libraries (1–3)	This paper		CVB3 infectious clone libraries with mutagenized capsid region
Recombinant DNA reagent	pUC19-HiFi-P1 (plasmid)	This paper		CVB3 capsid region used as template for DMS cloned into SalI digested pUC19 vector. Used for site-directed mutagenesis
Recombinant DNA reagent	T7 encoding plasmid (plasmid)	10.1128/jvi.02583–14	RRID:Addgene_65974	Plasmid encoding T7 polymerase for transfection
Recombinant DNA reagent	pIRES-3CDpro (plasmid)	This paper		CVB3 3 CD protease region cloned into XhoI and NotI pIRES plasmid (Clonetech)
Recombinant DNA reagent	peGFP_PLEKHA4	10.1016/j.celrep.2019.04.060		Kind gift from Dr. Jeremy Baskin GFP-PLEKHA4 expression plasmid
Recombinant DNA reagent	peGFP_PLSCR1	10.1371/journal.pone.0005006		Kind gift from Dr. Serfe Benichou GFP-PLSCR1 expression plasmid
Recombinant DNA reagent	pAcGFP-C1 WDR33	https://doi.org/10.1016/j.molcel.2018.11.036		Kind gift from Dr. Matthias Altmeyer pAcGFP-C1 WDR33 expression plasmid
Recombinant DNA reagent	FLAG-NLCR5	Addgene	RRID:Addgene_37521	NLCR5 expression plasmid
Recombinant DNA reagent	HA-ZC3HAV1	Addgene	RRID:Addgene_45907	HA-ZC3HAV1 expression plasmid
Recombinant DNA reagent	Fluc-eGFP	Addgene	RRID:Addgene_90170	Fluc-eGFP expression plasmid

Continued on next page

Continued

Reagent type (species) or resource	Designation	Source or reference	Identifiers	Additional information
Sequence-based reagent	HiFi_F	IDT	PCR primer	For generating PCR to clone libraries and sequencing: CTTTGTTGGGTTT ATACCACTTAGC TCGAGAGAGG
Sequence-based reagent	HiFi_R	IDT	PCR primer	For generating PCR to clone libraries and sequencing: CCTGTAGTTCCCCA CATACACTGCTCCG
Sequence-based reagent	DMS primers	IDT	PCR primer	Primers spanning the full coding region of the CVB3 capsid to perform codon mutagenesis. Listed in ***Supplementary file 1***.
Sequence-based reagent	2045_F	IDT	PCR primer	Primer used for Sanger sequencing. TCGAGTGTTTTTA GTCGGACG
Sequence-based reagent	2143_R	IDT	PCR primer	Primer used for Sanger sequencing. TCGAGTGTTTT TAGTCGGACG
Sequence-based reagent	3450_RT	IDT	PCR primer	Primer used for Sanger sequencing and RT-PCR. TCGAGTGTTTTT AGTCGGACG
Sequence-based reagent	qPCR_F	10.1038/nmicrobiol.2017.88	PCR primer	qPCR primer for competition assays. GATCGCATATG GTGATGATGTGA
Sequence-based reagent	qPCR_R	10.1038/nmicrobiol.2017.88	PCR primer	qPCR primer for competition assays. AGCTTCAGCGAGT AAAGATGCA
Sequence-based reagent	MGB_CVB3_wt	10.1038/nmicrobiol.2017.88	TaqManProbe	qPCR probe for competition assays. 6FAM-CGCATCGTA CCCATGG-TAMRA
Sequence-based reagent	MGB_CVB3_Ref	10.1038/nmicrobiol.2017.88	TaqManProbe	qPCR probe for competition assays. HEX-CGCTAGCTA CCCATGG-TAMRA
Sequence-based reagent	Q8D_F	IDT	PCR primer	Primer for site-directed mutagenesis: gtatcaacgGAT aagactggg
Sequence-based reagent	Q8D_R	IDT	PCR primer	Primer for site-directed mutagenesis: ttgag ctcccattttgctgt

Continued on next page

Continued

Reagent type (species) or resource	Designation	Source or reference	Identifiers	Additional information
Sequence-based reagent	K829L_F	IDT	PCR primer	Primer for site-directed mutagenesis: gagaaggcaCTA aacgtgaac
Sequence-based reagent	K829L_R	IDT	PCR primer	Primer for site-directed mutagenesis: gtattg gcagagtctaggtgg
Sequence-based reagent	K235D_F	IDT	PCR primer	Primer for site-directed mutagenesis: gggtcc aacGATttggtacag
Sequence-based reagent	K235D_R	IDT	PCR primer	Primer for site-directed mutagenesis: gga tgcgaccggtttgtccgc
Sequence-based reagent	R16G_F	IDT	PCR primer	Primer for site-directed mutagenesis: catga gaccGGActgaatgct
Sequence-based reagent	R16G_R	IDT	PCR primer	Primer for site-directed mutagenesis: tgccc cagtcttttgcgttg
Sequence-based reagent	K827G_F	IDT	PCR primer	Primer for site-directed mutagenesis: caatacgagGG Ggcaaagaac
Sequence-based reagent	K827G_R	IDT	PCR primer	Primer for site-directed mutagenesis: gcaga gtctaggtggtctagg
Sequence-based reagent	Q566M_F	IDT	PCR primer	Primer for site-directed mutagenesis: atttcgcagATGaacttttc
Sequence-based reagent	Q566M_R	IDT	PCR primer	Primer for site-directed mutagenesis: gaaaggagtgt ccttcaatag
Sequence-based reagent	T315P_F	IDT	PCR primer	Primer for site-directed mutagenesis: attacgg tcCCCatagcccca
Sequence-based reagent	T315P_R	IDT	PCR primer	Primer for site-directed mutagenesis: tgggacgtacgtggtgga
Sequence-based reagent	N395H_F	IDT	PCR primer	Primer for site-directed mutagenesis: gagaaggtcCAT tctatggaa

Continued on next page

Continued

Reagent type (species) or resource	Designation	Source or reference	Identifiers	Additional information
Sequence-based reagent	N395H_R	IDT	PCR primer	Primer for site-directed mutagenesis: tccaacatttt ggactgggac
Sequence-based reagent	T849A_F	IDT	PCR primer	Primer for site-directed mutagenesis: actacaatgGTC aatacgggc
Sequence-based reagent	T849A_R	IDT	PCR primer	Primer for site-directed mutagenesis: gatgctttgcct agtagtgg
Sequence-based reagent	K235D_F	IDT	PCR primer	Primer for site-directed mutagenesis: gggtccaacGAT ttggtacag
Sequence-based reagent	K235D_R	IDT	PCR primer	Primer for site-directed mutagenesis: ggatgcgacc ggtttgtccgc
Sequence-based reagent	3C_For	IDT	PCR primer	Primer for cloning CVB3 3 CD into pIRES: TATTCTCGAGACC ATGGGCCCTGC CTTTGAGTTCG
Sequence-based reagent	3D_Rev	IDT	PCR primer	Primer for cloning CVB3 3 CD into pIRES: TATTGCGGCCGCC TAGAAGGAGTCC AACCATTTCCT
Commercial assay or kit	NEBuilder HiFi DNA Assembly kit	NEB	E2621X	Seamless cloning
Commercial assay or kit	TranscriptAid T7 High Yield Transcription Kit	ThermoFisher Scientific	K0441	T7 in vitro transcription kit
Commercial assay or kit	*Quick*-RNA Viral kit	Zymo Research	R1035	RNA purification
Commercial assay or kit	*DNA Clean andConcentrator-5*	Zymo Research	D4013	DNA purification, gel purification
Commercial assay or kit	Luna Universal Probe One-Step RT-qPCR kit	NEB	E3006X	One-step qPCR master mix
Chemical compound, drug	Rupintivir	Tocris Biosciences	Cat. #: 6414	CVB3 3C protease inhibitor

Continued on next page

Continued

Reagent type (species) or resource	Designation	Source or reference	Identifiers	Additional information
Software, algorithm	Codon TilingPrimers	https://doi.org/10.1016/j.chom.2017.05.003		Software to design primers for mutagenesis (https://github.com/jbloomlab/CodonTilingPrimers)
Software, algorithm	Sanger Mutant Library Analysis	Dr. Jesse Bloom		Software to assess library mutagenesis by Sanger sequencing (https://github.com/jbloomlab/SangerMutantLibraryAnalysis)
Software, algorithm	Samtools	http://www.htslib.org/	version 1.5	Suite of programs for interacting with high-throughput sequencing data
Software, algorithm	Fastp	10.1093/bioinformatics/bty560		Software for NGS read trimming and QC
Software, algorithm	PicardTools, FastqToSam	https://broadinstitute.github.io/picard/	Version 2.2.4	Used to generate Bam files from Fastq files
Software, algorithm	Duplex pipeline	https://github.com/KennedyLabUW/Duplex-Sequencing; ***Kennedy et al., 2014***	Version 3.0	Analysis pipeline for duplex sequencing (UnifiedConsensus Maker.py)
Software, algorithm	VariantBam	10.1093/bioinformatics/btw111		Software to filter Bam files
Software, algorithm	BWA	https://sourceforge.net/projects/bio-bwa/files/	Version 0.7.16	Software to align NGS reads
Software, algorithm	Fgbio	http://fulcrumgenomics.github.io/fgbio/	version 1.1.0	Software used to hard-clip NGS reads
Software, algorithm	VirVarSeq	10.1093/bioinformatics/btu587	version 1.1.0	Software used to identify codons in each NGS read
Software, algorithm	Custom R scripts	This paper		Custom R scripts to process output of VirVarSeq script. Available at https://github.com/RGellerLab/CVB3_Capsid_DMS
Software, algorithm	DMS_tools2	10.1186/s12859-015-0590-4		Software to determine amino acid preferences and mutational fitness effects
Software, algorithm	TANGO	10.1038/nbt1012		Software to determine the effect of mutations on aggregation
Software, algorithm	FoldX	10.1093/nar/gki387		Software to determine the effect of mutations on stability

Continued on next page

Continued

Reagent type (species) or resource	Designation	Source or reference	Identifiers	Additional information
Software, algorithm	DSSP	http://swift.cmbi.ru.nl/gv/dssp/		Software used to obtain secondary structure and RSA within DMS_tools2
Software, algorithm	ViprDB	http://viperdb.scripps.edu/ ***Carrillo-Tripp et al., 2009***		Software used to obtain structural information on capsid sites
Software, algorithm	DECIPHER Package	10.32614/RJ-2016–025		R package for performing codon alignments
Software, algorithm	PhyDMS	doi: 10.7717/peerj.3657		For phylogenetic and differential selection analyses. https://jbloomlab.github.io/phydms/index.html
Software, algorithm	Custom R scripts	This paper		Custom R script to generate in silico peptides spanning 10AA 3 CD protease cleavage site. Available at https://github.com/RGellerLab/CVB3_Capsid_DMS
Software, algorithm	PSSMSearch	10.1093/nar/gky426		Used to generate position-specific scoring matrix and search human proteome for hits. http://slim.icr.ac.uk/pssmsearch/
Software, algorithm	Peptides R package	ISSN 2073–4859	Version 2.4.2	R package to predict molecular weight of proteins
Software, algorithm	RandomForest R package	10.1023/A:1010933404324	Version 4.6–16	R package for random forest prediction
Software, algorithm	Logolas	10.1186/s12859-018-2489-3		Package to generate logo plots in R

Viruses, cells, and plaque assays

HeLa-H1 (CRL-1958; RRID:CVCL_3334) and HEK293 (CRL-1573; RRID:CVCL_0045) cells were obtained from ATCC and were periodically validated to be free of mycoplasma. All work with CVB3 was based on the Nancy infectious clone (kind gift of Dr. Marco Vignuzzi, Institute Pasteur). Cells were cultured in culture media (Dulbecco's modified Eagle's medium [DMEM] with 10% heat-inactivated fetal bovine serum (FBS), Pen-Strep, and L-glutamine) with FBS concentrations of 2% during infection. For plaque assays, serial dilutions of the virus were used to infect confluent HeLa-H1 cells in six-well plates for 45 min, followed by overlaying the cells with a 1:1 mixture of 56˚C 1.6% agar (Arcos Organics 443570010) and 37˚C 2× DMEM with 4% FBS. Two days later, plates were fixed with formaldehyde (2% final concentration) after which the agar was removed and the cells stained with crystal violet to visualize plaques.

Deep mutational scanning

The infectious clone was modified by site-directed mutagenesis to remove an XhoI site present in the capsid region (P1) and introduce an XhoI site at position 692 as well as a Kpn2I site at position 3314, generating a pCVB3-XhoI-P1-Kpn2I clone (**Bou et al., 2019**). In addition, a pCVB3-XhoI-ΔP1-Kpn2I plasmid was generated by replacing the region between the XhoI and Kpn2I sites in pCVB3-XhoI-P1-Kpn2I with a short linker. To generate the template for DMS, the capsid region was amplified by PCR from pCVB3-XhoI-P1-Kpn2I with Phusion polymerase (Thermo Scientific) and primers HiFi-F (CTTTGTTGGGTTTATACCACTTAGCTCGAGAGAGG) and HiFi-R (CCTGTAGTTCCCCACA TACACTGCTCCG) and gel purified (Zymoclean Gel DNA Recovery Kit). Primers spanning the full coding region of the capsid region were designed using the CodonTilingPrimers software from the Bloom lab (https://github.com/jbloomlab/CodonTilingPrimers; **Dingens et al., 2017**) with the default parameters and synthesized by IDT (**Supplementary file 1**). These primers were used to perform the mutagenesis PCR on the capsid template together with the HiFi-F or HiFi-R primers in triplicate following published protocols (**Dingens et al., 2017**) with the exception that 10 rounds of mutagenesis were performed for libraries 1 and 2, while a second round of seven mutagenesis cycles was performed for library three to increase the number of mutation per clone. The products were gel purified and ligated to an XhoI and Kpn2I digested and gel purified pCVB3-XhoI-ΔP1-Kpn2I using NEBuilder HiFi DNA Assembly reaction (NEB) for 25 min. Mutagenesis efficiency was evaluated by the transformation of the assembled plasmids into NZY5α competent cells (NZY Tech), Sanger sequencing of 18–23 clones per library, and mutation analysis using the Sanger Mutant Library Analysis script (https://github.com/jbloomlab/SangerMutantLibraryAnalysis; **Bloom, 2014**). Subsequently, the assembled plasmid reactions were purified using a Zymo DNA Clean and Concentrator-5 kit (Zymo Research) and used to electroporate MegaX DH10B T1R Electrocomp cells (ThermoFisher) using a Gene Pulser XCell electroporator (Bio-Rad) according to the manufacturer's protocol. Cells were then grown overnight in a 50 mL liquid culture at 33˚C and DNA purified using the PureLink HiPure plasmid midiprep kit (Invitrogen). Transformation efficiency was estimated by plating serial dilutions of the transformation on agar plates. In total, 4.44×10^5, 1.46×10^5, and 2.19×10^5 transformants were obtained for lines 1, 2, and 3, respectively. Viral genomic RNA was then transcribed from SalI linearized, gel-purified full-length plasmids using the TranscriptAid T7 kit (Thermo Scientific), and four electroporations were performed using 4×10^6 HeLa-H1 cells in a 4 mm cuvette in 400 µL of calcium- and magnesium-free phosphate-buffered saline (PBS) using with 8 µg of RNA in a Gene Pulser XCell (Bio-Rad) set to 240 V and 950 µF. Electroporated cells were then pooled, and one-fourth was cultured for 9 hr to produce the passage 0 virus (P0). Following three freeze–thaw cycles, 2×10^6 plaque-forming units (PFU) were used to infect a 90% confluent 15 cm plate in 2.5 mL of infection media for 1 hr. Cells were then washed with PBS and incubated in 12 mL of infection media for 9 hr. Finally, cells were subjected to three freeze–thaw cycles, debris removed by centrifugation at 500 × g, and the supernatants collected to generate P1 virus stocks. All infections produced >2.38×10^6 PFU in P0 and >1.2×10^7 PFU in P1 as judged by plaque assay.

Next-generation sequencing analysis

Libraries were prepared following published protocols (**Kennedy et al., 2014**), and each library was run on a Novaseq6000 2 × 150 at a maximum of 30G per lane to reduce potential index hopping. Reads trimming was performed using fastp (**Chen et al., 2018**) (command: `–max_len1` 150 `–max_–len2` 150 –length_required 150 -x -Q -A), unsorted bam files were generated from fastq files using Picard tools FastqToSam (version 2.2.4) and merged into a single bam using the cat command of Samtools (version 1.5). The duplex pipeline was then implemented (https://github.com/KennedyLabUW/Duplex-Sequencing/UnifiedConsensusMaker.py; **Kennedy et al., 2014**). using the UnifiedConsensusMaker.py script and a minimum family size of 3, a cutoff of 0.9 for consensus calling, and an N cutoff of 0.3. The single-stranded consensus files (SSCS) were then aligned using BWA mem (version 0.7.16), sorted using Samtools, size selected to be 133 bp long using VariantBam (**Wala et al., 2016**), unaligned reads were discarded (Samtools view command with -F 4), and the resulting bam file indexed with Samtools. Subsequently, fgbio (http://fulcrumgenomics.github.io/fgbio/; version 1.1.0) was used to hard-clip 10 bp from each end and upgrade all clipping to hard-clip (-c Hard `–upgrade-clipping true –read-one-five-prime` 10 `–read-one-three-prime` 10 `–read-two-five-prime` 10 `–read-two-three-prime` 10). Variant bam was then used to keep all reads

that were between 50 and 150 bp, well-mapped, and had either no indels and less than five mutations (command –r {'':{'rules':[{'ins':[0,0],'del':[0,0],'nm':[0,4], 'mate_mapped':true,'fr':true,'length':[50,150]}]}}'). Finally, the codons in each read were identified using the VirVarSeq (**Verbist et al., 2015**) Codon_table.pl script using a minimum read quality of 20. A custom R script was then used to generate a codon counts table for each codon position by eliminating all codons containing ambiguous nucleotides and codons with a strong strand bias (StrandOddsRatio > 4), as well as all codons that are reached via a single mutation (available at https://github.com/RGellerLab/CVB3_Capsid_DMS); **Mattenberger, 2021**; copy archived at swh:1:rev:29dd205182f0886dc5bad3e6b4ddd6e786c58a75). Amino acid preferences and MFE were determined using DMStools2 (**Bloom, 2015**), with the Bayesian option and the default settings.

Structural analyses

The crystal structure PDB:4GB3 (**Yoder et al., 2012**) was used for all structural analyses. The effects of mutations on aggregation were determined using TANGO version 2.3.1 (**Fernandez-Escamilla et al., 2004**) using the default settings, and the effect on stability on the monomer and pentamer was determined using FoldX 4 (**Schymkowitz et al., 2005**) using the default settings. For the latter, the pentamer subunits were renamed to unique letters, all mutations between the reference sequence and the structure sequence were introduced using the BuildModel command, the structure was optimized using the RepairPDB command 5 or 10 times for the pentamer or monomer, respectively, and then the effects of the mutations were predicted using the BuildModel command (modified PDB files can be found at https://github.com/RGellerLab/CVB3_Capsid_DMS). Secondary structure and RSA were obtained from DSSP (http://swift.cmbi.ru.nl/gv/dssp/) using the dms_tools2.dssp function of dms_tools2, while interface, surface, and core residues as well as residue contact number, and presence in the twofold, threefold, and fivefold axes were obtained from ViprDB (http://viperdb.scripps.edu/) (**Carrillo-Tripp et al., 2009**). Distance from the center was calculated with Pymol using the Distancetoatom.py script on the monomer or pentamer. Finally, the location of antibody neutralization sites in CVB3 was obtained from an analysis of the CVB3 capsid structure in a previous publication (**Muckelbauer et al., 1995**).

Generation and evaluation of CVB3 capsid mutants

With the exception of mutant N395H (kind gift of Rafael Sanjuan) (**Bou et al., 2019**), all other mutants were generated by site-directed mutagenesis. For this, the PCR of the capsid region used as a template for DMS was phosphorylated and cloned into a SmaI digested pUC19 vector for use in the mutagenesis reactions (pUC19-HiFi-P1). For each mutant, non-overlapping primers containing the mutation in the middle of the forward primer were used to introduce the mutation with Phusion polymerase, followed by DpnI (Thermo Scientific) treatment, phosphorylation, ligation, and transformation of chemically competent bacteria. Successful mutagenesis was verified by Sanger sequencing. Subsequently, the capsid region was subcloned into pCVB3-XhoI-ΔP1-Kpn2I using XhoI and Kpn2I sites. Plasmids were then linearized with MluI, and 2 µg of plasmid was transfected into 5×10^5 HEK293 cells, together with a plasmid encoding the T7 polymerase (**Yun et al., 2015**) (Addgene 65974) using calcium phosphate. Briefly, an equal volume of 2× HBS (274 mM NaCl, 10 mM KCl, 1.4 mM Na_2HPO_4) was added dropwise to DNA containing 0.25M $CaCl_2$ while mixing, incubated 15 min at RT, and then added dropwise to cells. Following 48 hr, passage 0 (P0) virus was collected and titered by plaque assay. From this, 10^5 PFU were used to infect 90% confluent six-well HeLa-H1 cells (multiplicity of infection (MOI) 0.1) for 1 hr at 37˚C, after which the cells were washed twice with PBS and 2 mL of infection media added. Cells were then incubated until cytopathic effect (CPE) was observed. Emerging viral populations were titered by plaque assay and the capsid region sequenced to ensure no compensatory mutations or reversions arose during replication. The fitness of these mutants was then tested by direct competition with a marked reference virus using a Taqman RT-PCR method (**Moratorio et al., 2017**). Briefly, using four biological replicates, confluent HeLa-H1 cells in a 24-well plate were infected with 200 µL of a 1:1 mixture of 4×10^3 PFU (MOI 0.01) of the test and marked reference viruses for 45 min. Subsequently, the inoculum was removed, the cells were washed twice with PBS, 200 µL of infection media was added, and the cells were incubated for 24 hr at 37˚C. Finally, cells were subjected to three freeze–thaw cycles, debris removed by centrifugation at 500 × g, the supernatants collected and treated with 2 µL of RNase-Free DNaseI

(ThermoFisher) for 15 min at 37˚C, and viral RNA extracted using the *Quick*-RNA Viral Kit (Zymo Research), eluting in 20 µL. Quantification of the replication of each mutant versus the reference was performed using Luna Universal Probe One-Step RT-qPCR kit (New England BioLabs) containing 3 µL of total RNA, 0.4 µM of each qPCR primers, and 0.2 µM of each probe. The standard curve was performed using 10-fold dilutions of RNA extracted from 10^7 PFU of wild-type and reference viruses. All samples were performed with three technical replicates. The relative fitness (W) of each mutant versus the common marked reference virus was calculated using the following formula: $W = [R(t)/R(0)]^{1/t}$, where R(0) and R(t) represent the ratio of the mutant to the reference virus genomes in the initial mixture used for the infection and after 1 day (t = 1), respectively (***Carrasco et al., 2007***; ***Moratorio et al., 2017***).

Sequence variability and phylogenetic analyses

Amino acid variability was assessed using Shannon entropy. Briefly, all available, non-identical, full-genome CVB3, CVB, or enterovirus B sequences were downloaded from Virus Pathogen Resource (***Pickett et al., 2012***) (http://www.viprbrc.org) and codon-aligned using the DECIPHER package in R (available at https://github.com/RGellerLab/CVB3_Capsid_DMS). All alignment positions not present in our reference strain were removed, and a custom R script was used to calculate Shannon entropy. For phylogenetic and differential selection analyses, PhyDMS was run using the default settings on an alignment of CVB3 genomes that was processed with the phydms_prealignment module and using the average preferences from the three DMS replicates.

Identification of 3CD^{pro} cleavage sites in the human proteome

The amino acid preferences (the relative enrichment of each amino acid at each position standardized to 1) was used to generate in silico 1000 peptides spanning the 10 amino acid regions surrounding each cleavage site using a custom R script (available at https://github.com/RGellerLab/CVB3_Capsid_DMS). Specifically, for each peptide position, 100 peptides were generated that encoded each amino acid at a frequency corresponding to its preference observed in the DMS results, with the remaining positions unchanged. The resulting 1000 peptides from each cleavage site were uploaded to PSSMSearch (***Krystkowiak et al., 2018***) (http://slim.icr.ac.uk/pssmsearch/) using the default setting (psi_blast IC). Results were filtered to remove proteins indicated to be secreted, lumenal, or extracellular in the Warnings column. To test whether proteins were cleaved by the viral 3 CD protease, the corresponding region was PCR amplified from the Nancy infectious clone (primers 3C-For: TATTCTCGAGACCATGGGCCCTGCCTTTGAGTTCG and 3D-Rev: TATTGCGGCCGCCTAGAAGGAGTCCAACCATTTCCT) and cloned into the pIRES plasmid (Clonetech) using the restriction sites XhoI and NotI (pIRES-3CD^{pro}). For analysis of fusion proteins, HEK293 cells were transfected with GFP-PLEKHA4 (kind gift of Dr. Jeremy Baskin, Cornell University), GFP-PLSCR1 (kind gift of Dr. Serge Benichou, Institut Cochin), pAcGFP-WDR33 (Kind gift of Dr. Matthias Altmeyer, University of Zurich), FLAG-NLCR5 (Addgene #37521), HA-ZC3HAV1 (Addgene #45907), or the control plasmid FLuc-eGFP (Addgene #90170), together with the pIRES-3CD^{pro} plasmid using Lipofectamine 2000. Following 24 hr, proteins were collected by lysing in lysis buffer (50 mM Tris–HCl, 150 mM NaCl, 1% NP40, and protease inhibitor cocktail [Complete Mini EDTA-free, Roche]) and subjected to western blotting with the corresponding antibody (anti-GFP, Santa Cruz sc-9996; anti–FLAG, Santa Cruz sc-166335; anti-HA, Santa Cruz, sc-7392). For analysis of endogenous proteins, 3CD^{pro} was expressed for 48 hr before cell lysis, and western blotting using antibodies against WDR33 (Santa Cruz sc-374466), TSG101 (Santa Cruz sc-136111), GAK (Santa Cruz sc-137053), and MAGED1 (Santa Cruz sc-393291). When indicated, the 3C^{pro} inhibitor rupintrivir (Tocris Biosciences) was added at a concentration of 2 µM for the last 24 hr before collection. The predicted molecular weight of cleaved fragments was calculated using the mw function of the Peptides R package (version 2.4.2).

Statistical analyses

All experiments were performed with at least three biological replicates with the exception of the analysis of protein cleavage by western blotting, which was performed in duplicate. All statistical analyses were performed in R and were two tailed. For random forest prediction, the R RandomForest package (version 4.6–14) was employed using the default setting with an mtry of 10, and for the linear model, the formula lm(MFE ~ enterovirus B entropy + WT amino acid * mutant amino acid + predicted effect of mutations on stability in the pentamer + relative surface exposure) was used (available at https://github.com/RGellerLab/CVB3_Capsid_DMS). Sequence logoplots were producing using Logolas (***Dey et al., 2018***).

Data availability

Unaligned bam files have been uploaded to SRA (Bioproject PRJNA643896, SRA SRP269871, Accession SRX8663374-SRX8663384). The scripts and data required to obtain the codon count tables for all samples, to perform the random forest and linear model predictions, to generate the peptides for use with PSSMsearch, as well as the sequence alignments and modified structure files for FoldX analysis, can be found on GitHub (https://github.com/RGellerLab/CVB3_Capsid_DMS). Finally, the interactive heatmap of MFE across the capsid was generated by modifying a script from a prior publication (***Starr et al., 2020***) (available at https://github.com/jbloomlab/SARS-CoV-2-RBD_DMS/blob/master/interactive_heatmap.ipynb) and can be found on this projects' GitHub page (https://github.com/RGellerLab/CVB3_Capsid_DMS).

Acknowledgements

The authors would like to thank Dr. Javier O Cifuentes for help with the interpretation of antibody neutralization sites and Drs. Santiago Elena and Tzachi Hagai for critical reading of the manuscript. In addition, the authors would like to acknowledge the use of the Principe Felipe Research Center (CIPF) server which was co-financed by the European Union through the Operativa Program of the European Regional Development Fund (ERDF/FEDER) of the Comunitat Valenciana 2014–2020.

Additional information

Funding

Funder	Grant reference number	Author
Ministerio de Economía, Industria y Competitividad, Gobierno de España	BFU2017-86094-R	Ron Geller
Ministerio de Economía, Industria y Competitividad, Gobierno de España	RYC-2015-17517	Ron Geller
Ministerio de Economía, Industria y Competitividad, Gobierno de España	BES-2016-076677	Florian Mattenberger

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Author contributions

Florian Mattenberger, Conceptualization, Formal analysis, Investigation, Methodology, Writing - original draft, Writing - review and editing; Victor Latorre, Conceptualization, Investigation, Methodology, Writing - review and editing; Omer Tirosh, Software, Formal analysis, Methodology; Adi Stern, Software, Formal analysis, Supervision, Methodology, Writing - original draft, Writing - review and editing; Ron Geller, Conceptualization, Data curation, Formal analysis, Supervision, Funding acquisition, Investigation, Methodology, Writing - original draft, Writing - review and editing

Author ORCIDs

Florian Mattenberger https://orcid.org/0000-0002-2727-0284
Omer Tirosh https://orcid.org/0000-0001-8139-9866
Adi Stern http://orcid.org/0000-0002-2919-3542
Ron Geller https://orcid.org/0000-0002-7612-4611

Decision letter and Author response
Decision letter https://doi.org/10.7554/eLife.64256.sa1
Author response https://doi.org/10.7554/eLife.64256.sa2

Additional files

Supplementary files

- Supplementary file 1. Primers for next-generation sequencing.
- Supplementary file 2. Next-generation sequencing statistics.
- Supplementary file 3. Mutations observed by Sanger sequencing.
- Supplementary file 4. Mutational fitness effects of the mutagenized viral populations.
- Supplementary file 5. Results of qPCR validation of MFE.
- Supplementary file 6. Data used for random forest model and parameter explanation.
- Supplementary file 7. Differential selection results.
- Supplementary file 8. PSSMsearch results.
- Transparent reporting form

Data availability

Sequencing data have been uploaded to SRA (Bioproject PRJNA643896, SRA SRP269871, Accession SRX8663374-SRX8663384). All data used in the paper are either included as supplemental data and/or can be found at https://github.com/RGellerLab/CVB3_Capsid_DMS (copy archived at https://archive.softwareheritage.org/swh:1:rev:29dd205182f0886dc5bad3e6b4ddd6e786c58a75/).

The following dataset was generated:

Author(s)	Year	Dataset title	Dataset URL	Database and Identifier
Mattenberger F, Geller R	2020	Deep mutational scanning of the CVB3 capsid protein	https://www.ncbi.nlm.nih.gov/bioproject/643896	NCBI BioProject, PRJNA643896

References

Acevedo A, Brodsky L, Andino R. 2014. Mutational and fitness landscapes of an RNA virus revealed through population sequencing. *Nature* **505**:686–690. DOI: https://doi.org/10.1038/nature12861, PMID: 24284629

Ashenberg O, Padmakumar J, Doud MB, Bloom JD. 2017. Deep mutational scanning identifies sites in influenza nucleoprotein that affect viral inhibition by MxA. *PLOS Pathogens* **13**:e1006288. DOI: https://doi.org/10.1371/journal.ppat.1006288, PMID: 28346537

Belnap DM, McDermott BM, Filman DJ, Cheng N, Trus BL, Zuccola HJ, Racaniello VR, Hogle JM, Steven AC. 2000. Three-dimensional structure of poliovirus receptor bound to poliovirus. *PNAS* **97**:73–78. DOI: https://doi.org/10.1073/pnas.97.1.73, PMID: 10618373

Bloom JD. 2014. An experimentally determined evolutionary model dramatically improves phylogenetic fit. *Molecular Biology and Evolution* **31**:1956–1978. DOI: https://doi.org/10.1093/molbev/msu173, PMID: 24859245

Bloom JD. 2015. Software for the analysis and visualization of deep mutational scanning data. *BMC Bioinformatics* **16**:168. DOI: https://doi.org/10.1186/s12859-015-0590-4, PMID: 25990960

Bloom JD. 2017. Identification of positive selection in genes is greatly improved by using experimentally informed site-specific models. *Biology Direct* **12**:1. DOI: https://doi.org/10.1186/s13062-016-0172-z, PMID: 28095902

Bologna G, Yvon C, Duvaud S, Veuthey AL. 2004. N-Terminal myristoylation predictions by ensembles of neural networks. *Proteomics* **4**:1626–1632. DOI: https://doi.org/10.1002/pmic.200300783, PMID: 15174132

Bou JV, Geller R, Sanjuán R. 2019. Membrane-Associated Enteroviruses undergo intercellular transmission as pools of sibling viral genomes. *Cell Reports* **29**:714–723. DOI: https://doi.org/10.1016/j.celrep.2019.09.014, PMID: 31618638

Brass AL, Huang IC, Benita Y, John SP, Krishnan MN, Feeley EM, Ryan BJ, Weyer JL, van der Weyden L, Fikrig E, Adams DJ, Xavier RJ, Farzan M, Elledge SJ. 2009. The IFITM proteins mediate cellular resistance to influenza A H1N1 virus, west nile virus, and dengue virus. *Cell* **139**:1243–1254. DOI: https://doi.org/10.1016/j.cell.2009.12.017, PMID: 20064371

Callaway A, Giesman-Cookmeyer D, Gillock ET, Sit TL, Lommel SA. 2001. The multifunctional capsid proteins of plant RNA viruses. *Annual Review of Phytopathology* **39**:419. DOI: https://doi.org/10.1146/annurev.phyto.39.1.419, PMID: 11701872

Carrasco P, Daròs JA, Agudelo-Romero P, Elena SF. 2007. A real-time RT-PCR assay for quantifying the fitness of tobacco etch virus in competition experiments. *Journal of Virological Methods* **139**:181–188. DOI: https://doi.org/10.1016/j.jviromet.2006.09.020, PMID: 17092574

Carrillo-Tripp M, Shepherd CM, Borelli IA, Venkataraman S, Lander G, Natarajan P, Johnson JE, Brooks CL, Reddy VS. 2009. VIPERdb2: an enhanced and web API enabled relational database for structural virology. *Nucleic Acids Research* **37**:D436–D442. DOI: https://doi.org/10.1093/nar/gkn840, PMID: 18981051

Chan SL, Huppertz I, Yao C, Weng L, Moresco JJ, Yates JR, Ule J, Manley JL, Shi Y. 2014. CPSF30 and Wdr33 directly bind to AAUAAA in mammalian mRNA 3' processing. *Genes & Development* **28**:2370–2380. DOI: https://doi.org/10.1101/gad.250993.114, PMID: 25301780

Chen S, Zhou Y, Chen Y, Gu J. 2018. Fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**:i884–i890. DOI: https://doi.org/10.1093/bioinformatics/bty560, PMID: 30423086

Cifuente JO, Moratorio G. 2019. Evolutionary and structural overview of human Picornavirus capsid antibody evasion. *Frontiers in Cellular and Infection Microbiology* **9**:1–11. DOI: https://doi.org/10.3389/fcimb.2019.00283, PMID: 31482072

Corbic Ramljak I, Stanger J, Real-Hohn A, Dreier D, Wimmer L, Redlberger-Fritz M, Fischl W, Klingel K, Mihovilovic MD, Blaas D, Kowalski H. 2018. Cellular N-myristoyltransferases play a crucial Picornavirus genus-specific role in viral assembly, virion maturation, and infectivity. *PLOS Pathogens* **14**:e1007203. DOI: https://doi.org/10.1371/journal.ppat.1007203, PMID: 30080883

Dey KK, Xie D, Stephens M. 2018. A new sequence logo plot to highlight enrichment and depletion. *BMC Bioinformatics* **19**:473. DOI: https://doi.org/10.1186/s12859-018-2489-3, PMID: 30526486

Dingens AS, Haddox HK, Overbaugh J, Bloom JD. 2017. Comprehensive mapping of HIV-1 escape from a broadly neutralizing antibody. *Cell Host & Microbe* **21**:777–787. DOI: https://doi.org/10.1016/j.chom.2017.05.003, PMID: 28579254

Doud M, Bloom J. 2016. Accurate measurement of the effects of all Amino-Acid mutations on influenza hemagglutinin. *Viruses* **8**:155. DOI: https://doi.org/10.3390/v8060155

Dragovich PS, Prins TJ, Zhou R, Webber SE, Marakovits JT, Fuhrman SA, Patick AK, Matthews DA, Lee CA, Ford CE, Burke BJ, Rejto PA, Hendrickson TF, Tuntland T, Brown EL, Meador JW, Ferre RA, Harr JE, Kosa MB, Worland ST. 1999. Structure-based design, synthesis, and biological evaluation of irreversible human rhinovirus 3C protease inhibitors. 4. incorporation of P1 . *Journal of Medicinal Chemistry* **42**:1213–1224. DOI: https://doi.org/10.1021/jm9805384, PMID: 10197965

Du Y, Wu NC, Jiang L, Zhang T, Gong D, Shu S, Wu T-T, Sun R. 2016. Annotating protein functional residues by coupling High-Throughput fitness profile and Homologous-Structure analysis. *mBio* **7**:01801-16. DOI: https://doi.org/10.1128/mBio.01801-16

Fernandez-Escamilla AM, Rousseau F, Schymkowitz J, Serrano L. 2004. Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nature Biotechnology* **22**:1302–1306. DOI: https://doi.org/10.1038/nbt1012, PMID: 15361882

Fields BN, Knipe DM, Howley PM. 2013. *Fields Virology*. Philadelphia: Wolters Kluwer Health/Lippincott Williams & Wilkins.

Geller R, Vignuzzi M, Andino R, Frydman J. 2007. Evolutionary constraints on chaperone-mediated folding provide an antiviral approach refractory to development of drug resistance. *Genes & Development* **21**:195–205. DOI: https://doi.org/10.1101/gad.1505307, PMID: 17234885

Gray VE, Hause RJ, Luebeck J, Shendure J, Fowler DM. 2018. Quantitative missense variant effect prediction using Large-Scale mutagenesis data. *Cell Systems* **6**:116–124. DOI: https://doi.org/10.1016/j.cels.2017.11.003, PMID: 29226803

Haddox HK, Dingens AS, Bloom JD. 2016. Experimental estimation of the effects of all Amino-Acid mutations to HIV's Envelope Protein on Viral Replication in Cell Culture. *PLOS Pathogens* **12**:e1006114. DOI: https://doi.org/10.1371/journal.ppat.1006114, PMID: 27959955

Haddox HK, Dingens AS, Hilton SK, Overbaugh J, Bloom JD. 2018. Mapping mutational effects along the evolutionary landscape of HIV envelope. *eLife* **7**:e34420. DOI: https://doi.org/10.7554/eLife.34420, PMID: 29590010

Harrison SC. 2013. Principles of Virus Structure. In: Knipe D. M, Howley P. M (Eds). *Fields Virology*. Wolters Kluwer Health/Lippincott Williams & Wilkins. p. 52–86.

Hartman EC, Jakobson CM, Favor AH, Lobba MJ, Álvarez-Benedicto E, Francis MB, Tullman-Ercek D. 2018. Quantitative characterization of all single amino acid variants of a viral capsid-based drug delivery vehicle. *Nature Communications* **9**:1385. DOI: https://doi.org/10.1038/s41467-018-03783-y, PMID: 29643335

Hecht M, Bromberg Y, Rost B. 2015. Better prediction of functional effects for sequence variants. *BMC Genomics* **16**:S1. DOI: https://doi.org/10.1186/1471-2164-16-S8-S1

Heise MT, Virgin HW. 2013. Pathogenesis of viral infection. In: Knipe D. M, Howley P. M (Eds). *Fields Virology*. Wolters Kluwer Health/Lippincott Williams & Wilkins. p. 254–285.

Helenius A. 2013. Virus Entry and Uncoating. In: Knipe D. M, Howley P. M (Eds). *Fields Virology*. Wolters Kluwer Health/Lippincott Williams & Wilkins. p. 87–104.

Hilton SK, Doud MB, Bloom JD. 2017. Phydms: software for phylogenetic analyses informed by deep mutational scanning. *PeerJ* **5**:e3657. DOI: https://doi.org/10.7717/peerj.3657, PMID: 28785526

Hom N, Gentles L, Bloom JD, Lee KK. 2019. Deep mutational scan of the highly conserved influenza A virus M1 matrix protein reveals substantial intrinsic mutational tolerance. *Journal of Virology* **93**:1–16. DOI: https://doi.org/10.1128/JVI.00161-19

Hunter E. 2013. Virus Assembly. In: Knipe D. M, Howley P. M (Eds). *Fields Virology*. Wolters Kluwer Health/Lippincott Williams & Wilkins. p. 127–152.

Jagdeo JM, Dufour A, Klein T, Solis N, Kleifeld O, Kizhakkedathu J, Luo H, Overall CM, Jan E. 2018. N-Terminomics TAILS identifies host cell substrates of poliovirus and coxsackievirus B3 3C proteinases that modulate virus infection. *Journal of Virology* **92**:e02211-17. DOI: https://doi.org/10.1128/JVI.02211-17, PMID: 29437971

Jiang P, Liu Y, Ma HC, Paul AV, Wimmer E. 2014. Picornavirus morphogenesis. *Microbiology and Molecular Biology Reviews* **78**:418–437. DOI: https://doi.org/10.1128/MMBR.00012-14, PMID: 25184560

Kennedy SR, Schmitt MW, Fox EJ, Kohrn BF, Salk JJ, Ahn EH, Prindle MJ, Kuong KJ, Shen JC, Risques RA, Loeb LA. 2014. Detecting ultralow-frequency mutations by duplex sequencing. *Nature Protocols* **9**:2586–2606. DOI: https://doi.org/10.1038/nprot.2014.170, PMID: 25299156

Kodigepalli KM, Bowers K, Sharp A, Nanjundan M. 2015. Roles and regulation of phospholipid scramblases. *FEBS Letters* **589**:3–14. DOI: https://doi.org/10.1016/j.febslet.2014.11.036

Kristensen T, Belsham GJ. 2019. Identification of a short, highly conserved, motif required for picornavirus capsid precursor processing at distal sites. *PLOS Pathogens* **15**:e1007509. DOI: https://doi.org/10.1371/journal.ppat.1007509

Krystkowiak I, Manguy J, Davey NE. 2018. PSSMSearch: a server for modeling, visualization, proteome-wide discovery and annotation of protein motif specificity determinants. *Nucleic Acids Research* **46**:W235–W241. DOI: https://doi.org/10.1093/nar/gky426

Laitinen OH, Svedin E, Kapell S, Nurminen A, Hytönen VP, Flodström-Tullberg M. 2016. Enteroviral proteases: structure, host interactions and pathogenicity. *Reviews in Medical Virology* **26**:251–267. DOI: https://doi.org/10.1002/rmv.1883

Lee JM, Eguia R, Zost SJ, Choudhary S, Wilson PC, Bedford T, Stevens-Ayers T, Boeckh M, Hurt AC, Lakdawala SS, Hensley SE, Bloom JD. 2019. Mapping person-to-person variation in viral mutations that escape polyclonal serum targeting influenza hemagglutinin. *eLife* **8**:e49324. DOI: https://doi.org/10.7554/eLife.49324, PMID: 31452511

Livesey BJ, Marsh JA. 2020. Using deep mutational scanning to benchmark variant effect predictors and identify disease mutations. *Molecular Systems Biology* **16**:1–12. DOI: https://doi.org/10.15252/msb.20199380

Macejak DG, Sarnow P. 1992. Association of heat shock protein 70 with enterovirus capsid precursor P1 in infected human cells. *Journal of Virology* **66**:1520–1527. DOI: https://doi.org/10.1128/JVI.66.3.1520-1527.1992

Mattenberger F. 2021. CVB3_Capsid_DMS. *Software Heritage*. swh:1:rev:29dd205182f0886dc5bad3e6b4ddd6e786c58a75.https://archive.softwareheritage.org/swh:1:dir:40e0b58a90057980be80c0bbf4e39a10450c31c7;origin=https://github.com/RGellerLab/CVB3_Capsid_DMS;visit=swh:1:snp:d162eac2e7160d4ff881f6fc792188a5875294a9;anchor=swh:1:rev:29dd205182f0886dc5bad3e6b4ddd6e786c58a75/

Moratorio G, Henningsson R, Barbezange C, Carrau L, Bordería AV, Blanc H, Beaucourt S, Poirier EZ, Vallet T, Boussier J, Mounce BC, Fontes M, Vignuzzi M. 2017. Attenuation of RNA viruses by redirecting their evolution in sequence space. *Nature Microbiology* **2**:17088. DOI: https://doi.org/10.1038/nmicrobiol.2017.88

Muckelbauer JK, Kremer M, Minor I, Diana G, Dutko FJ, Groarke J, Pevear DC, Rossmann MG. 1995. The structure of coxsackievirus B3 at 3.5 å resolution. *Structure* **3**:653–667. DOI: https://doi.org/10.1016/S0969-2126(01)00201-5

Ogden PJ, Kelsic ED, Sinai S, Church GM. 2019. Comprehensive AAV capsid fitness landscape reveals a viral gene and enables machine-guided design. *Science* **366**:1139–1143. DOI: https://doi.org/10.1126/science.aaw2900

Perlmutter JD, Hagan MF. 2015. Mechanisms of virus assembly. *Annual Review of Physical Chemistry* **66**:217–239. DOI: https://doi.org/10.1146/annurev-physchem-040214-121637, PMID: 25532951

Pickett BE, Sadat EL, Zhang Y, Noronha JM, Squires RB, Hunt V, Liu M, Kumar S, Zaremba S, Gu Z, Zhou L, Larson CN, Dietrich J, Klem EB, Scheuermann RH. 2012. ViPR: an open bioinformatics database and analysis resource for virology research. *Nucleic Acids Research* **40**:D593–D598. DOI: https://doi.org/10.1093/nar/gkr859

Qing J, Wang Y, Sun Y, Huang J, Yan W, Wang J, Su D, Ni C, Li J, Rao Z, Liu L, Lou Z. 2014. Cyclophilin A associates with enterovirus-71 virus capsid and plays an essential role in viral infection as an uncoating regulator. *PLOS Pathogens* **10**:e1004422. DOI: https://doi.org/10.1371/journal.ppat.1004422, PMID: 25275585

Racaniello VR. 2013. Picornaviridae: The Viruses and Their Replication. In: Knipe M. D, Howley M. P (Eds). *Fields Virology*. Philadelphia: Wolters Kluwer Health/Lippincott Williams & Wilkins, c2013. p. 453–489.

Reeb J, Wirth T, Rost B. 2020. Variant effect predictions capture some aspects of deep mutational scanning experiments. *BMC Bioinformatics* **21**:107. DOI: https://doi.org/10.1186/s12859-020-3439-4, PMID: 32183714

Rossmann MG, He Y, Kuhn RJ. 2002. Picornavirus-receptor interactions. *Trends in Microbiology* **10**:324–331. DOI: https://doi.org/10.1016/S0966-842X(02)02383-1, PMID: 12110211

Schmitt MW, Kennedy SR, Salk JJ, Fox EJ, Hiatt JB, Loeb LA. 2012. Detection of ultra-rare mutations by next-generation sequencing. *PNAS* **109**:14508–14513. DOI: https://doi.org/10.1073/pnas.1208715109

Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, Serrano L. 2005. The FoldX web server: an online force field. *Nucleic Acids Research* **33**:W382–W388. DOI: https://doi.org/10.1093/nar/gki387

Shami Shah A, Batrouni AG, Kim D, Punyala A, Cao W, Han C, Goldberg ML, Smolka MB, Baskin JM. 2019. PLEKHA4/kramer attenuates dishevelled ubiquitination to modulate wnt and planar cell polarity signaling. *Cell Reports* **27**:2157–2170. DOI: https://doi.org/10.1016/j.celrep.2019.04.060, PMID: 31091453

Starr TN, Greaney AJ, Hilton SK, Ellis D, Crawford KHD, Dingens AS, Navarro MJ, Bowen JE, Tortorici MA, Walls AC, King NP, Veesler D, Bloom JD. 2020. Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding. *Cell* **182**:1295–1310. DOI: https://doi.org/10.1016/j.cell.2020.08.012, PMID: 32841599

Sun D, Chen S, Cheng A, Wang M. 2016. Roles of the picornaviral 3C proteinase in the viral life cycle and host cells. *Viruses* **8**:82. DOI: https://doi.org/10.3390/v8030082, PMID: 26999188

Thibaut HJ, van der Linden L, Jiang P, Thys B, Canela MD, Aguado L, Rombaut B, Wimmer E, Paul A, Pérez-Pérez MJ, van Kuppeveld FJ, Neyts J. 2014. Binding of glutathione to Enterovirus capsids is essential for virion morphogenesis. *PLOS Pathogens* **10**:e1004039. DOI: https://doi.org/10.1371/journal.ppat.1004039, PMID: 24722756

Thyagarajan B, Bloom JD. 2014. The inherent mutational tolerance and antigenic evolvability of influenza hemagglutinin. *eLife* **3**:e03300. DOI: https://doi.org/10.7554/eLife.03300

Verbist BMP, Thys K, Reumers J, Wetzels Y, Van der Borght K, Talloen W, Aerssens J, Clement L, Thas O. 2015. VirVarSeq: a low-frequency virus variant detection pipeline for Illumina sequencing using adaptive base-calling accuracy filtering. *Bioinformatics* **31**:94–101. DOI: https://doi.org/10.1093/bioinformatics/btu587

Wala J, Zhang C-Z, Meyerson M, Beroukhim R. 2016. VariantBam: filtering and profiling of next-generational sequencing data using region-specific rules. *Bioinformatics* **32**:2029–2031. DOI: https://doi.org/10.1093/bioinformatics/btw111

Wu NC, Olson CA, Du Y, Le S, Tran K, Remenyi R, Gong D, Al-Mawsawi LQ, Qi H, Wu TT, Sun R. 2015. Functional constraint profiling of a viral protein reveals discordance of evolutionary conservation and functionality. *PLOS Genetics* **11**:e1005310. DOI: https://doi.org/10.1371/journal.pgen.1005310, PMID: 26132554

Xing L, Tjarnlund K, Lindqvist B, Kaplan GG, Feigelstock D, Cheng RH, Casasnovas JM. 2000. Distinct cellular receptor interactions in poliovirus and rhinoviruses. *The EMBO Journal* **19**:1207–1216. DOI: https://doi.org/10.1093/emboj/19.6.1207

Xu L, Zheng Q, Li S, He M, Wu Y, Li Y, Zhu R, Yu H, Hong Q, Jiang J, Li Z, Li S, Zhao H, Yang L, Hou W, Wang W, Ye X, Zhang J, Baker TS, Cheng T, et al. 2017. Atomic structures of coxsackievirus A6 and its complex with a neutralizing antibody. *Nature Communications* **8**:505. DOI: https://doi.org/10.1038/s41467-017-00477-9, PMID: 28894095

Yoder JD, Cifuente JO, Pan J, Bergelson JM, Hafenstein S. 2012. The crystal structure of a coxsackievirus B3-RD variant and a refined 9-angstrom cryo-electron microscopy reconstruction of the virus complexed with decay-accelerating factor (DAF) provide a new footprint of DAF on the virus surface. *Journal of Virology* **86**:12571–12581. DOI: https://doi.org/10.1128/JVI.01592-12, PMID: 22973031

Ypma-Wong MF, Dewalt PG, Johnson VH, Lamb JG, Semler BL. 1988. Protein 3CD is the major poliovirus proteinase responsible for cleavage of the p1 capsid precursor. *Virology* **166**:265–270. DOI: https://doi.org/10.1016/0042-6822(88)90172-9

Yun T, Park A, Hill TE, Pernet O, Beaty SM, Juelich TL, Smith JK, Zhang L, Wang YE, Vigant F, Gao J, Wu P, Lee B, Freiberg AN. 2015. Efficient reverse genetics reveals genetic determinants of budding and fusogenic differences between nipah and Hendra viruses and enables real-time monitoring of viral spread in small animal models of Henipavirus infection. *Journal of Virology* **89**:1242–1253. DOI: https://doi.org/10.1128/JVI.02583-14, PMID: 25392218