Comparative genomics to unravel adaptive mechanisms in *Saccharomyces*

PhD Thesis

Laura Gutiérrez Macías

January, 2021

Supervisors :

Dr. Eladio Barrio Esparducer

Dr. Christina Toft

VNIVERSITAT DE VALÈNCIA

Dpto. de Bioquímica y Biología Molecular

Doctorado en Biomedicina y Biotecnología

VNIVERSITAT DE VALÈNCIA **Departament de Genètica**

I2SysBio INSTITUTE FOR INTEGRATIVE SYSTEMS BIOLOGY

El Dr. Eladio Barrio Esparducer, Catedrático de Genética de la Universitat de València y la Dra. Christina Toft, Investigadora SEJI del Instituto de Biología Integrativa de Sistemas (I2SysBio), Universitat de València.

CERTIFICAN

Que Doña Laura Gutiérrez Macías, Graduada en Biotecnología por la Universidad de León, ha realizado bajo su dirección el trabajo titulado: "Comparative genomics to unravel adaptive mechanisms in *Saccharomyces*", que presenta para optar al grado de Doctor en el programa de Biomedicina y Biotecnología por la Universitat de València. Asimismo, certifican haber dirigido y supervisado tanto los distintos aspectos del trabajo como su redacción.

Y para que conste a los efectos oportunos, en cumplimiento de la legislación vigente, firman el presente certificado en

Valencia, a 14 de enero de 2021

Los directores:

Fdo. Eladio Barrio

Fdo. Christina Toft

This PhD Thesis work was supported by AGL2015-67504-C3-3-R and RTI2018-093744-B-C32 grants from the Spanish Government and European Union ERDF-FEDER.

Contents

Resumen amplio en español **1**

Introduction **15**
Yeasts and the Fungi kingdom 15
Saccharomyces yeasts 17
Ecology and geography of *Saccharomyces* species 19
Saccharomyces in fermentations 22
New trends in winemaking 23
Molecular mechanisms involved in generating genomic diversity in *Saccharomyces* 25
Single Nucleotide Polimorphisms (SNPs) 26
Gene duplication 27
Gross chromosomal rearrangements (GCRs) 30
Reticulated evolution in *Saccharomyces* 30
Comparative genomics as a powerful tool to identify molecular mechanisms involved in adaptation 32
Whole-genome sequencing 32
Phylogenomics 34
Detecting natural selection in genomic data 34

Objectives **36**

1 Comparative genomics between *Saccharomyces kudriavzevii* and *Saccharomyces cerevisiae* applied to identify mechanisms involved in adaptation **39**
1.1 Introduction 39
1.2 Materials and methods 43
1.2.1 Assembly and annotation 43
1.2.2 Orthology and alignment 44
1.2.3 Signatures of positive selection 44

1.2.4 Testing constant rate of evolution 46
1.2.5 Functional divergence 47
1.2.6 Duplicated genes 48
1.2.7 Gene onthology and pathway enrichment analyses 49
1.3 Results 50
1.3.1 Differential adaptive evolution between *S. cerevisiae* and *S. kudriavzevii* 50
1.3.2 Evidence of adaptive evolution in genes related with known physiological differences between the two *Saccharomyces* species 55
1.3.3 Adaptive evolution in genes for which no previous physiological data is available 56
1.4 Discussion 58

2 GWideCodeML: a Python package for testing evolutionary hypotheses at the genome-wide level and its application in detecting signatures of positive selection in *Saccharomyces uvarum* 61
2.1 Introduction 61
2.2 GWideCodeML development 64
2.2.1 Input files 64
2.2.2 GWideCodeML workflow 65
2.2.3 Nested models implemented in the package 65
2.2.4 Additional options and modules 67
2.2.5 Output files 68
2.2.6 GWideCodeML testing 68
2.2.7 GWideCodeML compared to other software 69
2.2.8 GWideCodeML availability 70
2.3 Case study: GWideCodeML applied to detect signatures of positive selection in *S. uvarum.* 71
2.3.1 Genome dataset 71
2.3.2 Species tree phylogeny 71
2.3.3 Evolutionary hypothesis testing using GwideCodeML 74
2.3.4 Duplicates 74
2.3.5 Gene Ontology and Pathway Enrichment Analyses 74
2.4 Results 75
2.4.1 GWideCodeML increases both the number of analysed genes and the statistical power of the analysis 75
2.4.2 Cell-wall and chemical homeostasis related genes showed signatures of positive selection in the *S. uvarum* clade 76
2.4.3 Ribosome and glucose fermentation genes have multiple codon positions under positive selection 76

2.4.4 Enrichment of ohnologues under positive selection 77
2.5 Discussion 80

3 Convergent adaptation of *Saccharomyces uvarum* to sulphite, an antimicrobial preservative widely-used in human-driven fermentations 83
3.1 Introduction 83
3.2 Materials and Methods 87
3.2.1 Yeast strains, media, and fermentations. 87
3.2.2 Edited strains construction. 87
3.2.3 Genome sequencing, assembly, and annotation. 88
3.2.4 Phylogenetic analyses. 88
3.2.5 Analyses of the origin of the shared chromosomal rearrangement among BMV58, CECT12600, and NPCC1417 strains. 89
3.2.6 Southern blot analysis. 89
3.2.7 Gene Expression Determination. 90
3.2.8 Sulphite tolerance assay. 91
3.3 Results 92
3.3.1 Two new recombination events in the *SSU1* promoter of *S. uvarum* strains 92
3.3.2 Strains carrying the chromosomal rearrangements in the *SSU1* promoter are more tolerant to sulphite 94
3.3.3 Phylogenetic reconstruction and the origin of the *SSU1*-promoter recombination. 97
3.4 Discussion 102

4 High-quality new assemblies of *Saccharomyes* genomes provide insights into their evolutionary dynamics 106
4.1 Introduction 106
4.2 Materials and methods 109
4.2.1 Genome sequencing and assembly 109
4.2.2 Annotation 109
4.2.3 Pangenome analysis 111
4.2.4 Subtelomeric regions 112
4.2.5 Subtelomeric protein-coding gene families 112
4.2.6 Phylogeny reconstruction 113
4.3 Results 114
4.3.1 Highly accurate annotations of *Saccharomyces* long-read based genome assemblies 114
4.3.2 Early-divergent species *S. eubayanus* and *S. uvarum* show striking differences in subtelomeric lengths 114

4.3.3 Expansion of subtelomeric protein-coding gene families is responsible for a species pattern differentiation 120
4.3.4 Core genes have roles in basic cellular maintenance functions while species-specific genes have a wide range of functions. . . 121
4.3.5 Thiamine biosynthesis and maltose metabolism genes show variations in their genome positions. 125
4.4 Discussion 127

General Discussion **130**

Conclusions **135**

Bibliography **138**

Appendix **156**

List of Figures

1 *Saccharomyces* life cycle. 16
2 *Saccharomyces* taxonomic rearrangements 18
3 *Saccharomyces* cladogram and estimated nucleotide pairwise distances 18
4 Species phylogeny of the pre- and post-WGD species 29
5 *Saccharomyces* whole-genome sequencing projects pie chart 33

1.1 Species-based comparative genomics approach. 51
1.2 Functional divergence along *S. kudriavzevii* genome. 53
1.3 Functional divergence among a subset of metabolic pathways. 54

2.1 GWideCodeML workflow. 66
2.2 Species tree 73
2.3 Glucose fermentation genes under positive selection 78

3.1 New *SSU1* promoter variants found in *S. uvarum*. 93
3.2 Confirmation of the presence of VII^{XVI} recombination in the *S. uvarum* strains 94
3.3 Relative *SSU1* expression and growth in *S. uvarum* strains during fermentation. 96
3.4 Relative *SSU1* expression in *S. uvarum* wild type and edited strains grown in a fermentation experiment. 98
3.5 Phylogenetic analysis of the *S. uvarum* sequenced genomes. 100
3.6 Determination of a selective sweep in the NPCC1417 genome. 101

4.1 Annotation pipeline 110
4.2 Subtelomeric lengths in *Saccharomyces* species 117
4.3 Subtelomeric gene densities in *Saccharomyces* species 118
4.4 Hierarchical clustering dendrogram 119
4.5 Phylogeny of the *PGU*-coding gene family 122
4.6 *Saccharomyces* pangenome 124
4.7 *Saccharomyces* phylogeny. 125

List of Tables

1 Ecology and biogeography of *Saccharomyces* species 20

1.1 List of strains and sources of the genomic sequences used in Chapter 1. 45
1.2 Number of genes with a positive result in positive selection, functional divergence and Tajima's relative rate test analyses. 55

2.1 Overlapping features between GWideCodeML and other bioinformatics tools 70
2.2 List of strains and sources of the genomic sequences used in Chapter 2 72
2.3 GWideCodeML testing results 75
2.4 Number of genes under positive selection. 79

4.1 Number of annotated genes in *Saccharomyces* assemblies 115
4.2 Largest subtelomeric protein-coding gene families 120

Resumen amplio en español

Las levaduras son organismos unicelulares pertenecientes al reino Fungi. En particular, las levaduras del subfilo Saccharomycotina comprenden en torno a dos tercios de todas las levaduras descritas hasta el momento. Estas levaduras, en su estado haploide y diploide, se reproducen asexualmente tras la mitosis por gemación de una célula idéntica. Esta forma de reproducción es la más frecuente aunque también puden reproducirse sexualmente por la hibridación de dos esporas.

La diversidad genética de las levaduras es comparable a la diversidad existente entre animales y plantas. Hay más de 1000 especies descritas dentro del subfilo Saccharomycotina siendo el organismo *Saccharomyces cerevisiae* el más estudiado. *S. cerevisiae* no sólo es un organismo modelo utilizado en diversos estudios de investigación básica sino que también es una levadura clave en la industria biotecnológica. Es el organismo predominante en la mayoría de procesos como fermentaciones de bebidas alcohólicas o elaboración de pan.

El género *Saccharomyces* se compone de ocho especies, incluyendo *S. cerevisiae*. Las levaduras de este género poseen genomas muy compactos distribuidos en 16 cromososomas compuestos por unas 12 megabases de genoma nuclear y unas 80 kilobases de genoma mitocondrial. *S. cerevisiae* fue el primer organismo eucariota en tener su secuencia genómica completa y a día de hoy es,

probablemente, el genoma eucariota mejor secuenciado y anotado. La divergencia nucleotídica existente dentro del género *Saccharomyces* es muy alta; las dos especies más lejanas tienen una divergencia comparable a la existente entre humanos y pájaros. Estas caractarísticas de interés hacen que *Saccharomyces* haya sido propuesto como modelo para estudios evolutivos. Sin embargo, cuando se trata de las especies no-*S. cerevisiae*, hay mucha menos información disponible acerca de su ecología y muchos menos genomas secuenciados disponibles.

Las especies de *Saccharomyces* se han aislado en ambientes naturales de diferentes localizaciones de la Tierra, asociadas sobre todo con árboles del orden Fagales como hayas y robles. *S. uvarum* y *S. cerevisiae* son las únicas especies del género aisladas en ambientes controlados por el ser humano como son las fermentaciones. *S. kudriavzevii* fue aislada por primera vez en Japón a partir de lodo y hojas en descomposición del suelo. Después se aisló en Europa en la corteza de robles de Portugal, España y Francia. *S. uvarum*, como hemos mencionado, se ha aislado de ambientes fermentativos como vino y sidra aunque también está ampliamente extendida por ambientes naturales de regiones con temperaturas más bajas como La Patagonia o Europa del Este. Existe un linage de *S. uvarum* compuesto de cepas aisladas de Nueva Zelanda cuya divergencia nucleotídica respecto a otras *S. uvarum* oscila en torno al 5% con lo cual se considera que está en proceso de especiación y pudiendo llegar a ser especies diferentes. Entre los factores que más influyen en el crecimiento óptimo de las distintas especies, la temperatura es el más determinante, responsable de separar las especies de *Saccharomyces* en dos grupos: termotolerantes y criotolerantes. Gracias a estas diferencias, se cree que las especies pueden vivir en simpatría sin competir entre ellas, como por ejemplo *S. cerevisiae* (termotolerante) y *S. kudriavzevii* (criotolerante), cepas de estas especies se han aislado del mismo roble.

Los procesos de producción de bebidas fermentadas han sido muy importantes

en la historia de la humanidad y datan del 7000 antes de Cristo. La fermentación alcohólica es un proceso anaerobio en el que las levaduras transforman los azúcares en etanol y dióxido de carbono como principales productos, además de otros compuestos de alto valor para lo industria como son los compuestos aromáticos. Durante la fermentación, las levaduras tienen que lidiar con algunos factores de estrés como la hiperosmolaridad, estrés oxidativo, concentración de etanol, temperatura y adición de compuestos antimicrobianos como el sulfito.

En los últimos años, las nuevas demandas de la industria vínica han aumentado el interés de los enólogos por levaduras alternativas a la especie *Saccharomyces* que puedan utilizarse para enfrentarse a los nuevos retos del sector debido, entre otros factores, al cambio climático. Se buscan también levaduras que tengan temperaturas de crecimiento óptimas más bajas, que produzcan menor etanol y que tengan un buen perfil aromático. Las levaduras criotolerantes del género *Saccharomyces*, como *S. kudriavzevii* y *S. uvarum*, son especies de interés que pueden ser de gran utilidad en la industria enológica ya que cumplen esos requerimientos. *S. kudriavzevii* es una especie cuya temperatura óptima de crecimiento está en torno a unos 23ºC, más baja que la de *S. cerevisiae* que está alrededor de los 30ºC. Las fermentaciones vínicas a bajas temperaturas son de gran interés debido a que facilitan la retención de aromas deseables. Otro dato interesante es que *S. kudriavzevii* dirige el flujo de carbono sobre todo hacia la producción de glicerol lo que hace también que produzca menos etanol. Aunque *S. kudriavzevii* tiene características de interés, no llega a mejorar el rendimiento fermentativo que tiene *S. cerevisiae*.

S. uvarum es la única especie del género, aparte de *S. cerevisiae*, que se ha aislado de ambientes fermentativos controlados por el ser humano. Se ha aislado en fermentaciones de sidra de manzana y en fermentaciones vínicas a baja temperatura (12ºC) mejorando incluso el rendimiento de *S. cerevisiae* a la misma temperatura. Estudios previos demuestran que la estrategia para adaptarse con éxito a bajas

temperaturas podría ser un incremento de la actividad de la ruta del shikimato.

El estudio de aislados de las distintas especies del género *Saccharomyces* ha revelado una gran diversidad fenotípica correspondiente con una gran diversidad genómica. Los mecanismos moleculares responsables de generar diversidad genómica son los principales responsables de la biodiversidad existente en este género de levaduras. A continuación se comentan brevemente los mecanismos moleculares más importantes. Variaciones de un sólo nucleótido (SNPs): son variaciones de un sólo nucleótido a la largo de la secuencia de ADN, donde hay varias alternativas entre diferentes individuos de poblaciones o especies diferentes. También pueden ocurrir pequeñas inserciones o deleciones en cuyo caso se conocen como indels. Este tipo de variaciones, si están en regiones codificantes, pueden causar una alteración en la estructura de la proteína, función o intereacciones proteína-proteína. En el caso de encontrarse en regiones reguladoras, pueden causar variación en la expresión de los genes. Duplicación génica: es un mecanismo molecular clave en la generación de nuevas funciones génicas. La teoría dice que la mayoría de genes duplicados retornan a estado de copia única relativamente pronto tras la duplicación. Esto ocurre porque al haber dos copias, las restricciones selectivas se relajan en una de ellas lo que causa que se acumulen mutaciones deletéreas hasta que finalmente ocurre la pseudogeneización. La principal causa responsable de la retención de un gen es que el incremento en la dosis génica esté favorecido selectivamente. Otras veces, puede ocurrir un proceso de neo-funcionalización del gen tras su duplicación, adquiriendo una nueva función mediante divergencia funcional. También puede ocurrir la sub-funcionalización, cuando ambas copias se reparten la función del gen original. Los principales mecanismos por los cuales puede ocurrir una duplicación génica son una duplicación del genoma completo por aloploidización, duplicaciones génicas a pequeña escala, frecuentes sobre todo en regiones subteloméricas y por último, aneuploidías o duplicaciones de cromosomas completos. Reordenaciones cromosómicas: este fenómeno está mediado por recombinación no-homóloga. Puede

tener un papel en los mecanismos de adaptación tempranos ya que pueden causar alteraciones en la expresión de un gen o de varios genes. Evolución reticulada: se conoce también como herencia no-vertical. Ocurre cuando hay un intercambio de material genético entre diferentes linajes. En levaduras los fenómenos de hibridación interespecífica son muy frecuentes ya que las barreras pre-cigóticas son muy leves. Existen híbridos entre casi todas las especies del género, siendo *S. pastorianus* probablemente el más conocido. Este híbrido es el responsable de las fermentaciones de cerveza tipo lager y es un híbrido entre las especies *S. eubayanus* y *S. cerevisiae*. También se han encontrado híbridos de *S. cerevisiae* y *S. kudriavzevii* responsables de fermentaciones vínicas europeas y de *S. cerevisiae* y *S. uvarum* participando en fermentaciones a baja temperatura de sidra y vino. En ocasiones, una hibridación puede ir seguida de retrocruzamiento con el parental y, tras varias rondas, dar lugar a la aparición de introgresiones, que son pequeñas regiones en el genoma de una especie que provienen de otra especie. Finalmente, también son conocidos los eventos de transferencia horizontal que son mecanismos a través de los cuales una especie adquiere genes de otra especie que está más alejada y que no pertenece directamente a su líena ancestral. Todos estos fenómenos de intercambio genético se ha visto que pueden ser importantes en los mecanismos de adaptación de levaduras a ambientes, sobre todo a ambientes fermentativos controlados por el ser humano.

El desarrollo de las técnicas de secuenciación, su mejora y su abaratamiento ha permitido que haya habido un gran aumento de los genomas secuenciados en los últimos años. Como ya se ha comentado, *S. cerevisiae* fue el primer eucariota en tener su genoma secuenciado en el año 1996. A día de hoy contamos con miles de genomas secuenciados de distintas cepas, poblaciones y especies del género *Saccharomyces*. De todos los proyectos de secuenciación de ADN publicados hasta la fecha, más del 90% corresponden a *S. cerevisiae*. Hay un desequilibrio significativo en los esfuerzos de secuenciación dedicados a las especies. Esa falta de datos dificulta los ensayos de genómica comparada aplicados al género al completo. La calidad de

los genomas de referencia es un factor a tener en cuenta. *S. cerevisiae*, es el eucariota con el genoma mejor secuenciado y anotado. La combinación de técnicas de segunda generación (lecturas cortas de hasta 300 pb), junto con las recientes tecnologías de tercera generación (lecturas largas, más de 1000 pb), permiten secuenciar genomas con una alta calidad y resolver regiones que anteriormente no se podía usando sólo las secuencias de lectura corta. Esas regiones, son regiones repetitivas como las regiones subteloméricas, y son regiones de gran importancia en mechanismos adaptativos ya que contienen muchas familias de genes duplicados. A pesar de estos avances, aún hay varios genomas del género que permacen sin actualizar, dificultando el estudio de esas regiones en concreto. La falta de genomas de referencia también compromete el éxito de los proyectos de genómica comparada que implican al género al completo.

En el primer capítulo de la presente tesis doctoral, nos centramos en las diferencias genómicas entre *S. cerevisiae* y *S. kudriavzevii*. Para ello, en primera instancia, se secuenciaron dos cepas de *S. kudriavzevii* aisladas de corteza de Roble de la zona de Ciudad Real (CR85) y Castellón (CA111). Esas secuencias genómicas se ensamblaron y se añadieron a las secuencias ya publicadas de *S. kudriavzevii* IFO 1802 (cepa tipo) y ZP591, aislada de Portugal. Se decidió anotar todas las cepas con un método propio. Por un lado, transferimos la anotación de un genoma bien anotado como es el de *S. cerevisiae* S288C a nuestros ensamblajes. Por otro lado, utilizamos un método de secuenciación *de novo*. Con un programa propio, mezclamos ambas anotaciones y curamos manualmente los genes. De esta manera, se evitamos que haya genes parálogos que se anoten erróneamente. Una vez anotadas con éxito nuestras secuencias, añadimos secuencias anotadas de *S. cerevisiae* tanto de ambientes naturales como domesticados. Finalmente escogimos, como especie *outgroup*, la especie *Torulaspora delbrueckii*. Utilizamos las secuencias alineadas de genes ortólogos entre todas estas cepas y especies para realizar un test de selección positiva de tipo “rama-sitio”. En este tipo de test asumimos que puede haber una presión selectiva en una rama concreta (*foreground*) diferente a las

otras ramas del árbol y además asumimos que una secuencia codificante no tiene porqué estar sometida a la misma presión selectiva a lo largo de toda la secuencia. Para hacer este análisis, primero fijamos como rama objetivo, la rama que da lugar a las cepas de *S. kudriavzevii* y después repetimos el análisis fijando la rama de las cepas de *S. cerevisiae*. También aplicamos un test de tasas (test de Tajima) donde vemos aquellos genes cuyas tasas evolutivas están aceleradas en una rama o en otra. Finalmente, utilizamos las secuencias génicas traducidas a secuencias aminoacídicas para identificar divergencia funcional entre las dos especies de *Saccharomyces*. A partir de los análisis de divergencia funcional observamos que las rutas metabólicas de respuesta al estrés osmótico y oxidativo junto con la ruta de metabolismo de esfingolípidos fueron las que contuvieron un mayor número de genes con divergencia funcional entre ambas especies. En cuanto a los análisis de selección positiva, observamos tres genes de interés con posibles regiones de su secuencia sometidas a selección positiva. Estos genes fueron el gen *FBA1* que codifica una aldolasa que actúa en la glucólisis, justo donde las triosas se dirigen bien al final de la glucólisis o bien a la síntesis de glicerol. También observamos selección positiva en *ARO4*, un gen de la ruta de síntesis de aminoácidos aromáticos, y *DAL3* un gen que participa en la degradación de alantoína. El gen *DAL4* pertenece al clúster de genes de alantoína los cuales se encontraron todos bajo divergencia funcional. En cuanto al análisis del test de tasas, observamos que cuatro genes de la ruta de riboflavina mostabran señales de tener sus tasas aceleradas en la rama de *S. kudriavzevii*. Teniendo en cuenta los dos análisis de selección positiva y aceleración de tasas, observamos que tres genes arrojaron resultados positivos para ambos test. Los genes *FBA1*, *RQC2*, perteneciente a un complejo ribosomal, y el gen *ZIP1* que codifica una proteína con un dominio tropomiosina que no contienen las cepas de *S. cerevisiae*. El gen *FBA1*, es un gen esencial de la glucólisis que interviene en la síntesis de glicerol. Las diferencias metabólicas estudiadas por otros autores entre *S. cerevisiae* y *S. kudiravzevii* señalan que una de las diferencias más notables es el desvío en la

glucólisis hacia la producción glicerol en *S. kudriavzevii*. Este mecanismo podría ser un mecanismo adaptativo dado el papel que tiene el glicerol como osmorregulador a baja temperatura. El gen *DAL3* y todo el clúster también es un resultado interesante al intervenir en la degradación de alantoína. La alantoína se ha encontrado en ambientes muy parecidos a los de *S. kudriavzevii*, en exudados de árboles y además se ha visto que tiene un efecto importante en la *fitness* de las levaduras.

Dada la variedad de resultados, decidimos centrarnos únicamente en el análisis de selección positiva, ya que nos parecía el más estricto de todos y el que nos podía dar resultados más concretos. También éramos conscientes de la importancia de añadir más secuencias si queríamos tener éxito en nuestros análisis. Cuántas más secuencias, mayor poder estadístico tiene el análisis y los resultados serán más fiables. Para poder añadir más secuencias de manera automática y utilizar el programa de selección positiva que, a priori, está diseñado para analizar gen a gen, lo cual dificulta mucho la tarea, decidimos crear un programa. En este programa, además, implementamos nuevos modelos de selección para probar con nuestros datos. A parte del modelo de “rama-sitio”, implementamos el modelo de ramas y el modelo de sitios. Una vez desarrollado con éxito nuestro programa, lo aplicamos en la detección de selección positiva en la especie *S. uvarum* utilizando los tres modelos. En primer lugar, en el modelo de “rama-sitio”, obtuvimos 89 genes bajo selección positiva, un cosiderable aumento con respecto al análisis realizado previamente. De estos genes, observamos que doce de ellos estaban relacionados con pared celular. También observamos una alta proporción de genes relacionados con homeostasis. Otros resultados a destacar fueron los genes que codificaban para: una alcohol deshydrogenasa (*ADH4*), el factor transcripcional activador de los genes de la ruta de aminoácidos aromáticos (*ARO80*), la permeasa general de aminoácidos (*GAP1*) y una transferasa de manosas (*MNN2*). Decidimos clasificar los genes en función de si eran parálogos o no y observamos un enriquecimiento en genes onólogos (parálogos cuyo origen es la duplicación del genoma completo). Es decir, que una alta proporción

de los genes observados bajo selección positiva en *S. uvarum* eran parálogos. En cuanto al análisis de sitios, también obtuvimos unos resultados muy interesantes con 49 genes bajo selección positiva. En esos 49 genes había un enriquecimiento en términos GO (ontología de genes) en los componentes de ribosoma y pared celular. También encontramos cinco genes que codifican enzimas de la ruta de fermentación de la glucosa. Esos genes eran *CDC19, ENO2, FBA1, GPM1* y *PDC1* entre los que observamos de nuevo a la aldolasa *FBA1*. Cuando ponemos los resultados en común, además de lo mencionado, obervamos varios genes que codifican proteínas de la familia de las manoproteínas. Esta familia de proteínas se ha visto que tiene un papel fundamental en la respuesta adaptativa de las levaduras ante un choque térmico por frío. Además ya se había visto que estos genes eran los que mayores divergencias presentaban entre varias especies de *Saccharomyces*, con lo cual podrían haber sido objeto de la acción de la selección positiva, fijando cambios que mejorasen su respuesta a estrés por frío.

En el tercer capítulo, analizamos los genomas de *S. uvarum* en busca de señales de domesticación. Esta especie, es la única del género, junto con *S. cerevisiae*, que ha sido aislada de ambientes controlados por el ser humano como fermentaciones de vino y sidra de manzana. A bajas temperaturas, *S. uvarum* tiene una capacidad fermentativa más alta que *S. cerevisiae.* El estudio de señales de domesticación ya se había realizado previamente con algunos genomas de *S. uvarum* donde se observaron introgresiones de *S. eubayanus* en cepas de sidra. El análisis de los genomas de nuestra colección de *S. uvarum* reveló dos posibles translocaciones en cepas distintas en la región promotora del gen *SSU1*, el gen que codifica para la bomba de sulfito celular, que se encarga de extraer el sulfito del interior de la célula. El sulfito es un compuesto que se utiliza comúnmente en la industria enológica y se usa como antimicrobiano para evitar la aparición de microorganismos no deseados. Este compuesto, también es tóxico para las levaduras encargadas de realizar la fermentación, con lo cual, la obtención de levaduras tolerantes a sulfito es

una característica deseada para el sector enológico. Un incremento en la expresión de este gen da lugar a levaduras más tolerantes, algo que se ha demostrado en cepas vínicas de *S. cerevisiae*. En estas cepas, se han descrito varios eventos de recombinación independientes en el promotor de *SSU1* que han dado lugar a una mayor tolerancia a este compuesto. Realizamos ensayos por goteo para evaluar la tolerancia a etanol de nuestras cepas de *S. uvarum* y observamos una correlación entre las cepas que poseían la recombinación demostrando una mayor tolerancia. Se realizaron análisis de expresión génica del gen *SSU1* observando un incremento de la expresión de ese gen en las cepas que poseían la recombinación en fermentaciones llevadas a cabo en presencia y ausencia de sulfito en el medio, concluyendo que el gen se encontraba constitutivamente activo en las cepas recombinadas. Para demostrar que la alta expresión en ese gen estaba causada por la región promotora recombinada, se realizaron mutantes donde se insertaron las dos versiones del promotor en cepas *Wild Type* (WT). Se llevaron a cabo fermentaciones de los mutantes en presencia y ausencia de sulfito y se midió la expresión de *SSU1* observando un aumento significativo de la expresión con respecto a la cepa WT. Sin emabrgo, el aumento no llegó a los niveles de las cepas originalmente recombinantes con lo cual creemos que debe haber otros elementos de la región promotora aguas abajo del gen que pueden tener que ver con el aumento de expresión. Finalmente, analizamos los genomas para intentar explicar el posible origen de la recombinación en dintistintas cepas localizadas en diferentes partes de la filogenia. Concluimos que hubo un origen único en el ancestro común de todas esas cepas que después se mantuvo en algunas cepas domesticadas debido a la presión selectiva. Nuestra hipótesis se respalda tras observar una zona de barrido selectivo en la región del promotor de una de las cepas recombinadas y mosaico. En este trabajo ejemplificamos un caso de convergencia adaptativa a distintos niveles taxonómicos ya que demostramos que recombinaciones independientes en la región promotora de *SSU1* dan lugar al mismo efecto en dos especies tan alejadas como son *S. cerevisiae* y *S. uvarum*. Además, este estudio pone

en el punto de mira al gen *SSU1* como diana de selección para diferentes estudios.

En el último capítulo de esta tesis, utilizamos tecnologías de secuenciación de tercera generación, o tecnologías de secuenciación de lectura larga, para obtener ensamblajes de alta calidad de las especies *S. mikatae*, *S. kudriavzevii* y *S. uvarum*. Esta tecnología ya se había usado con otras especies del género *Saccharomyces* para obtener genomas de referencia y hacer análisis de población, sin embargo, nunca se había utilizado para secuenciar estas tres especies. Utilizamos además, secuenciaciones de Illumina, *mate-pair* y *paired-end* para corregir los ensamblajes ya que esta tecnologías tienen la desventaja de tener una alta tasa de error a nivel de nucleótidos individuales. Una vez obtuvimos los ensamblajes, utilizamos nuestro propio procedimiento de anotación, el mismo que utilizamos en el capítulo 1. Tras anotar esos genomas y los genomas publicados de otras especies del género, obtuvimos el pan-genoma de *Saccharomyces* utilizando criterios de homología y sintenia. Observamos que 4950 formaban parte del core del pan-genoma. El análisis de enriquecimiento funcional en términos GO reveló que había enriquecimiento en funciones del ciclo celular como genes involucrados en mitosis y meiosis. En definitiva, los genes del core estaban relacionados con genes básicos del mantenimiento celular. Por otro lado, también estudiamos los genes específicos de cada especie de *Saccharomyces*. Esos genes se correspondían con funciones más relacionadas con adaptación al nicho específico de cada especie. Por ejemplo, en *S. uvarum* y *S. eubayanus*, observamos dos genes relacionados con respiración celular *YAT1* y *CYB2*. Si bien es cierto que se sabe de sobra que *S. cerevisiae* es un microorganismo que principalemente fermenta glucosa, poco se sabe de las fuentes de carbono no-fermentables y de su relación con las especies de *Saccharomyces*. Podría ser que estas especies tuvieran capacidad de utilizar otros tipos de fuentes de carbono con mayor eficiencia que *S. cerevisiae*. También observamos, en esas mismas especies, genes homólogos a una galactosa permeasa y una transferasa de manosas. En *S. kudriavzevii* encontramos un gen que codifica una ceramida sintasa, un gen que

en crecimiento en quimiostato, se ha visto sobre expresado en *S. cerevisiae* ante un choque térmico por baja temperatura. Dada la alta calidad de los ensamblajes, decidimos estudiar a fondo las regiones subteloméricas. Para ello, lo primero que hicimos fue definir esas regiones subteloméricas. Muchos investigadores fijan una longitud concreta en todos los cromosomas de aproximadamente 35 kb desde el telómero para definir qué es la región subtelomérica. Sin embargo, nosotros utilizamos otro criterio, utilizado previamente, en el que los genes hacen de límites subteloméricos. Es decir, definimos el límite en el primer gen conservado desde el telómero entre todas las especies y cepas analizadas. Una vez definidos estos límites, calculamos los valores de longitud en pares de bases y los valores de densidad génica medidos en número de genes anotados en esa región. Estos valores los calculamos para todos los genomas de nuestro estudio y para todos los subtelómeros. Haciendo un análisis de clustering de estos valores observamos un patrón que nos separa las especies en dos grandes grupos, uno formado por *S. cerevisiae* y *S. paradoxus* y el otro compuesto por el resto de especies. Además de estudiar los subtelómeros usando estos datos, también identificamos familias de proteínas en regiones subteloméricas y su posible expansión o duplicación en las diferentes especies del género *Saccharomyces*. El análisis de familias subteloméricas nos reveló aquellas familias compuestas por un mayor número de genes entre los distintos subtelómeros de las especies. Entre esas familias encontramos la familia de las seripauperinas, diferentes transportadores de membrana como los transportadores de hexosa y de amino ácidos y también genes relacionados con el catabolismo de carbohidratos. Analizando particularmente aquellas familias compuestas por un mayor número de copias en unas especies concretas observamos grandes diferencias responsables de un patrón de especiación. Por ejemplo, una familia de los genes codificantes de endopoligalacturonasas se encontraba duplicada en *S. jurei* y *S. mikatae* con una copia más y en *S. uvarum* y *S. eubayanus* con hasta tres copias más totalmente divergentes del gen conservado en todas las *Saccharomyces*. Estos

genes son muy importantes para las levaduras que fermentan sustratos de plantas. Dada la ecología de las especies, podrían ser responsables de la fermentación de estos sustratos en las especies del género *Saccharomyces*. En *S. kudriavzevii* observamos la duplicación de una familia de manitol deshidrogenasas. Este resultado es interesante ya que se han caracterizado levaduras altamente productoras de manitol en el género *Candida* y estas levaduras están aisladas de lodo y tierra, un ambiente parecido a donde se aislaron las primeras *S. kudriavzevii* encontradas en suelo y hojas en descomposicion.

A partir de esta tesis doctoral se han podido extraer las conclusiones que paso a detallar a continuación.

En el capítulo 1 concluimos que la mitad de las proteínas analizadas entre *S. cerevisiae* y *S. kudriavzevii* tenían evidencias de divergencia funcional según nuestro método. Una alta proporción de ellas están relacionadas con estrés oxidativo y síntesis de esfingolípidos. Los genes del metabolismo de riboflavinas muestras señales de aceleración de tasas evolutivas en la especie de *S. kudriavzevii*.

En el capítulo 2 observamos que el incremento de secuencias tanto en la rama de interés como en los *outgroups* incrementa el número de genes detectados bajo selección positiva y aumenta el poder estadístico del método de análisis. Los genes bajo selección positiva en *S. uvarum* están enriquecidos en onólogos. Varios genes codificantes de manoproteínas de la pared celular mostraron evidencias de selección positiva según diferentes modelos evolutivos. El gen *FBA1*, codificante de una aldolasa de la glucólisis, mostró evidencias de selección positiva en las ramas de *S. uvarum* y *S. kudriavzevii* en test evolutivos independientes. Este gen podría tener un papel crucial en las especies criotolerantes del género que son altamente productoras de glicerol.

En el capítulo 3 descubrimos que cepas de *S. uvarum* aisladas de ambientes

fermentativos muestran señales de domesticación en su genoma. Las reordenaciones cromosómicas identificadas en las cepas de *S. uvarum* causan la sobreexpresión del gen *SSU1*, el gen responsable de la detoxificación de sulfito en levaduras. Las reordenaciones cromosómicas confieren una ventaja selectiva a las cepas de *S. uvarum* que crecen en fermentaciones donde el sulfito se usa como aditivo antimicrobiano. Un único origen de la recombinación tipo VIItXVI ocurrió en el ancestro común de las especies holárticas. Esta recombinación se mantuvo en algunas poblaciones domesticadas por presión selectiva. La región del gen *SSU1* es una diana de evolución causando una convergencia evolutiva entre *S. cerevisaie* y *S. uvarum*.

En el capítulo 4 concluimos que el core del pangenoma de *Saccharomyces* contiene genes codificantes de proteínas relacionadas con ciclo celular y homeostasis. Los genes específicos poseen una gran variedad de funcionalidades. Las regiones subteloméricas tienen una tremenda variabilidad en términos de longitud y densidad génica. A pesar de esta variabilidad, algunos patrones se han podido observar responsables de la separación entre especies termotolerantes y criotolerantes. Las familias de proteínas más abundantes localizadas en regiones subteloméricas tienen funciones específicas de adaptación al ambiente. Se ha podido observar la expansión o duplicación de familias de genes subteloméricos siguiendo un patrón de especies. *S. uvarum* muestra un incremento de las endo-poligalacturonasas mientras que *S. kudriavzevii* muestra un incremento de la familia de las manitol deshydrogenasas.

Introduction

Yeasts and the Fungi kingdom

Fungi is one of the most diverse kingdoms of living organisms. It comprehends different organisms such as mushrooms, lichens, moulds, smuts, rusts and yeasts. These organisms have had a crucial role in human-life history as they have been involved in essential contributions to the fields of medicine, research or human industry. Yeasts stand out among the Fungi kingdom as a fascinating group of unicellular organisms that lack fruiting bodies, can divide either by fission or budding and can display different sexual states (Kurtzman and Sugiyama, 2015). They do not conform a monophyletic group in the phylogenetic tree, and it has been demonstrated that they have convergently emerged multiple times from distinct lineages of Ascomycota or Basidiomycota (Nagy et al., 2014). The diverse heterotrophic metabolisms exhibited in yeasts have allowed them to successfully colonize every continent and every major aquatic and terrestrial biome (Kurtzman et al., 2011; Opulente et al., 2018).

Budding yeasts, from the subphylum Saccharomycotina, gather almost two-thirds of all known yeasts. Diploid budding yeast cells usually reproduce asexually, after mitosis and buds off genetically identical cells. This reproduction mode is

frequent when yeasts are growing in environments under favourable conditions with no nutrient limitations. When lacking nutrients such as nitrogen, diploid cells can undergo sporulation after meiosis, producing a tetrad containing four haploid spores. These dormant spores are resistant to adverse environmental conditions, and they can become metabolically active haploid gametes when returning to favourable conditions. Spores can display two mating types (a/α) codified by the MAT locus located at chromosome III. Spores can mate and return to a diploid state (sexual reproduction) when they find another spore with the opposite mating-type (Figure 1).

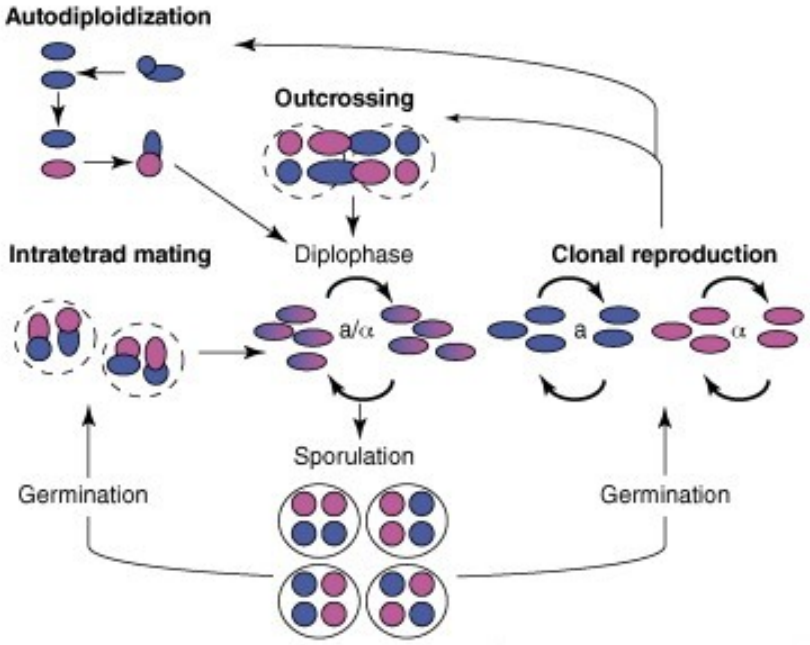


Figure 1: *Saccharomyces* life cycle. Adapted from Replansky et al. (2008)

The budding yeast genetic diversity is very high and comparable to the diversity found between animals and plants (Shen et al., 2018). More than 1,000 species have been described (Hittinger et al., 2015; Kurtzman et al., 2011) within Saccharomycotina. One of these species is the model organism *Saccharomyces cerevisiae* which was the first eukaryotic organism to have its whole genome sequenced (Goffeau et al., 1996).

Saccharomyces yeasts

During the last decades, the *Saccharomyces* group of yeasts has been defined according to different criteria based on their physiological properties (e.g. the ability to grow and ferment compounds), reproduction (the biological species concept based on interespecies sterility and intraspecies fertility) (Naumov, 1987), or DNA-DNA reassociation studies (Vaughan Martini, 1989)). These definitions gave rise to the description of three species: *S. cerevisiae, S. paradoxus* and *S. bayanus* formerly known as *Saccharomyces sensu stricto* group (Figure 2). The increase in the geographic locations sampled and the improvement of the isolation sampling techniques together with the development of the Next Generation Sequencing (NGS) technologies have led to the current definition of eight *Saccharomyces* species (Dujon and Louis, 2017) (Figure 2). The rise of NGS technologies facilitated the species' whole-genome sequencing being a turning point for the species classification. The whole-genome sequencing project's feasibility allowed the redefinition of the *Saccharomyces* genus based on phylogenetic criteria rather than reproductive isolation.

The *Saccharomyces* genomes contain 16 chromosomes with around 12 Mb of the nuclear genome and 70-90 Kb of mitochondrial DNA (mtDNA). Yeast genomes are characterized by their low number of introns and their loss of active transposons (Dujon, 2006). There are approximately 6,000 genes in the *Saccharomyces* genomes, with a high proportion of paralogues, coming from the whole-genome duplication event, also known as ohnologues (Wolfe, 2000), and genes arising from small-scale duplications. Over the evolutionary time of yeasts, extensive duplication has occurred leading to the expansion of specific gene families such as the sugar-utilization gene family of *Saccharomyces* genus with potential roles on niche-specific adaptation. *Saccharomyces* genus shows a sequence divergence which is much higher than

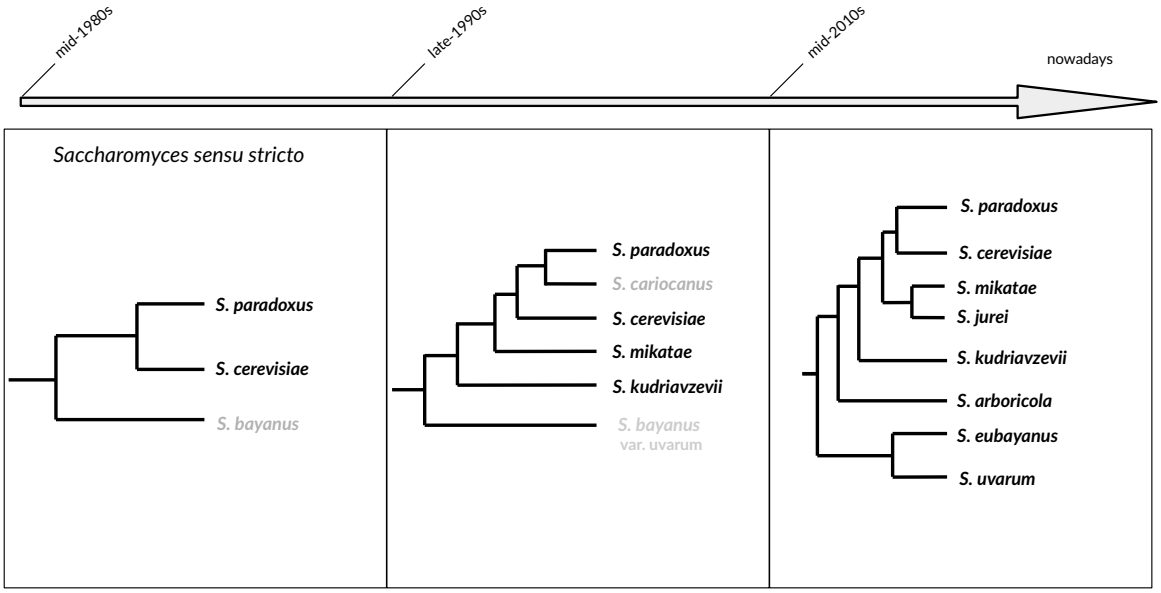
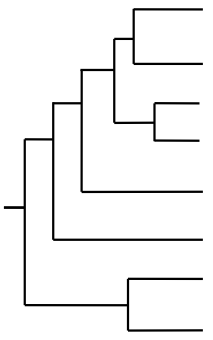


Figure 2: *Saccharomyces* taxonomic rearrangements. The picture shows the main changes in the *Saccharomyces* taxonomy over the years. By the mid-1980s to mid-1990s, the use of DNA–DNA reassociation (Vaughan Martini, 1989; Vaughan Martini and Kurtzman, 1985)) and the biological species definition (Naumov, 1987) allowed the consolidation of the *Saccharomyces* yeasts into three species. By the late 1990s, the use of the biological species definition, along with electrophoretic karyotyping and presence/absence of specific repeated sequences, resulted in the discovery of three new species (*S. cariocanus, S. mikatae*, and *S. kudriavzevii*) and the clarification of *S. bayanus var. uvarum*. In recent years, whole genome sequencing together with genetic analysis has resulted in the current view of the group.

anticipated by the similarities described between the species (Figure 3). The nucleotide identity found between *S. cerevisiae* and the early-divergent clade (*S. eubayanus*–*S. uvarum*) is comparable with values of identity found between humans and birds (Dujon, 2006). This high sequence divergence is commonly observed between yeasts species being other genera like *Lachancea* or *Eremothecium* even more diverged than *Saccharomyces*.



	S. eubayanus	S. uvarum	S. arboricola	S. kudriavzevii	S. mikatae	S. jurei	S. paradoxus	S. cerevisiae
S. eubayanus	0	0.07	0.21	0.21	0.23	0.23	0.22	0.24
S. uvarum	0.07	0	0.2	0.2	0.23	0.23	0.22	0.23
S. arboricola	0.21	0.2	0	0.19	0.21	0.22	0.2	0.22
S. kudriavzevii	0.21	0.2	0.19	0	0.2	0.2	0.19	0.21
S. mikatae	0.23	0.23	0.21	0.2	0	0.07	0.16	0.18
S. jurei	0.23	0.23	0.22	0.2	0.07	0	0.16	0.18
S. paradoxus	0.22	0.22	0.2	0.19	0.16	0.16	0	0.11
S. cerevisiae	0.24	0.23	0.22	0.21	0.18	0.18	0.11	0

Figure 3: *Saccharomyces* cladogram (left) and estimated nucleotide pairwise distances (right).Pairwise distances between species were calculated from the alignment of 4759 shared orthologous genes using the ”dist.dna” function from the *ape* R package (Paradis and Schliep, 2019) under the ”K89” model (Kimura, 1981).

Ecology and geography of *Saccharomyces* species

Saccharomyces genus has been proposed as a model clade for ecology and evolution research studies. However, when non-*S. cerevisiae* species are studied, there is currently an important lack of information about the ecological niches. The available information suggests that optimal growth conditions depend on the combination of different parameters like pH, radiation, temperature, insect, plan hosts and metabolism. Understanding *Saccharomyces* ecology is fundamental to shed light on the history of this genus and its evolutionary dynamics. *S. cerevisiae* is one of the most studied eukaryotes because of its industrial relevance and its use as a model organism in many research studies, resulting in its genome sequence being the best annotated among eukaryotes (Cherry et al., 2012; Goffeau et al., 1996). It has been traditionally associated with human-related environments such as fermented beverages production or bread-making but has also been found in a wide range of natural environments worldwide (Peter et al., 2018). The rest of the species of the genus have been found in natural environments, mainly isolated from tree barks, soil and insects' gut (Table 1). *S. kudriavzevii* was first isolated from soil and decayed leaves in Japan (Naumov et al., 2000a) and later European populations were discovered in oak barks (Erny et al., 2012; Lopes et al., 2010; Sampaio and Gonçalves, 2008).

Besides *S. cerevisiae*, *S. uvarum* is the only non-hybrid *Saccharomyces* species traditionally found in anthropic environments such as wine and cider-making (Almeida et al., 2014; Rodríguez et al., 2014). *S. uvarum* and *S. eubayanus* were grouped as a single species (*S. bayanus*) until a few years ago with both of them having been isolated from regions of Patagonia (Libkind et al., 2011). *S. eubayanus* is the parent together with *S. cerevisiae* of the *S. pastorianus* hybrid responsible for the lager beer fermentation and it has been isolated from America, Asia and New Zealand (Bing et al., 2014; Gayevskiy and Goddard, 2016; Libkind et al., 2011; Peris et al., 2016, 2014;

Table 1: Ecology and biogeography of *Saccharomyces* species.

Species	Geographic locations	Ecology
S. cerevisiae	Asia, Europe, North and South America	Wild and domesticated strains. Wild strains isolated from Fagaceae. Domesticated species associated with wine, beer, sake fermentation, dairy products, etc., Clinical isolates
S. paradoxus	Eurasia and North America. European populations found in New Zealand	Wild strains associated with bark and soil of Fagaceae (*Quercus* spp.)
S. jurei	Europe (France)	Wild strains associated with Fagaceae (*Quercus* spp.)
S. mikatae	Asia (China and Japan)	Wild strains associated with Fagaceae (*Quercus* spp.)
S. kudriavzevii	Asia (Japan and Taiwan) and Southern Europe (Portugal, Spain, France)	Wild strains associated with Fagaceae (*Quercus* spp.), soil and decayed leaves.
S. arboricola	Asia (China and Taiwan) and Australasia (New Zealand)	Wild strains associated with Fagaceae (*Quercus, Nothofagus, Juglans* and *Cyclobalanopsis*)
S. uvarum	South and North America, Eurasia and Australasia	Mostly wild strains associated with Fagaceae (*Nothofagus, Araucaria*) and domesticated strains isolated from wine and cider fermentations
S. eubayanus	South and North America and East Asia	Wild strains associated with Fagaceae (mainly *Nothofagus, Araucaria* and *Quercus rubra*) and *Pinaceae* (*Cedrus* sp., *Pinus taeda*)

Rodríguez et al., 2014). There is no evidence of *S. eubayanus* isolated in Europe, yet researchers suspect that it probably remains undiscovered (Peris et al., 2014). Although *Saccharomyces* yeasts are present in a wide range of environments, the most frequent hosts are oak trees in the Northern Hemisphere and beeches in the Southern Hemisphere. The reason why *Saccharomyces* species (excluding *S. cerevisiae*) have not been found in other locations might be due to sampling biases introduced during enrichment culturing. Besides, main sampling efforts have been focused only on particular regions of our planet, especially Europe, North America and Japan. In the last years, increased sampling efforts done by different research groups have been proved to be efficiently leading to new *S. uvarum* and *S. eubayanus* isolations in Argentina, Chile and China (Almeida et al., 2014; Bing et al., 2014; Peris et al., 2014; Rodríguez et al., 2014), and the discovery of a brand new *Saccharomyces* species in France, *S. jurei* (Naseeb et al., 2017). Therefore, it is likely that future sampling may reveal undiscovered *Saccharomyces* populations or species. Environments near oak trees (forest soil and exudates of these trees) seem to be the optimal niches for most wild *Saccharomyces* species (Boynton and Greig, 2014). This association of wild *Saccharomyces* yeasts is unexpected regarding the ability of *S. cerevisiae* to grow on high-sugar substrates. *S. cerevisiae* has a preference for producing alcohol via fermentation even when there is oxygen available, which is known as the Crabtree effect (Pronk et al., 1996). This phenomenon is believed to be an adaptive mechanism for growing in high-sugar environments as Crabtree-positive yeasts can outcompete Crabtree-negative microorganisms (Piškur et al., 2006). Nevertheless, wild *Saccharomyces* are rarely found on fruit or in other high-sugar environments, but mostly associated with tree barks. A possible explanation might be that these yeasts are contaminants of a nearby sugar-rich substrate. For example, *S. eubayanus* and *S. uvarum* have been isolated from *Cyttaria* stromata in *Nothofagus* bark, a very high sugar environment optimal for yeast communities (Čadež et al., 2019; Libkind et al., 2011). Sometimes, samples have been collected from tree exudates which are

also sugar-rich environments (Charron et al., 2014; Naumov et al., 1998). Another cause might be that yeasts have been adapted to grow in small amounts of sugar, and they may exhibit a mechanism different from the Crabtree effect to succeed in the environment. This assumption is widely accepted for non-*S. cerevisiae* species. Temperature is one of the most important factors that have been determinant in the *Saccharomyces* genus evolution (Gonçalves et al., 2011; Salvadó et al., 2011). Species with different optimal growth optimal temperatures have been found living in sympatry. For example, *S. paradoxus* (thermotolerant) and *S. uvarum* (cryotolerant) were isolated from the same oak barks (Sampaio and Gonçalves, 2008) as well as *S. cerevisiae* (thermotolerant) and *S. kudriavzevii* (cryotolerant) (Lopes et al., 2010; Sampaio and Gonçalves, 2008). In all these reported cases of sympatry between *Saccharomyces* species, temperature preferences were the strongest phenotypic discontinuity. Therefore, one plausible hypothesis is that the adaptation to different temperature niches allows their coexistence without competition exclusion.

Saccharomyces in fermentations

Fermented beverages have played key roles in the development of human culture and technology. Evidence for the production of fermented beverages dates these biotechnological processes back to 7000 B.C. (McGovern et al., 2004), in the ancient Chinese culture. Molecular DNA analysis of samples from Egyptian wine jars allowed identifying *S. cerevisiae* as the microorganism responsible for the wine fermentation since, at least, 3150 B.C. (Cavalieri et al., 2003). By the year 500 B.C., the wine was already spread out in the Mediterranean civilizations. Winemaking was also introduced into America during the colonization in the XVI century, and in South Africa in the XVII century (Pretorius, 2000). Alcoholic fermentation is the anaerobic transformation of sugars into carbon dioxide and ethanol by yeasts. Ethanol and carbon dioxide are the main products of the alcoholic fermentations but not

unique as several flavour compounds of a great value for the winemaking industry are also produced. During alcoholic fermentation, yeasts are subject to several stress conditions (Carrasco et al., 2001). The most important stresses are temperature, oxidative stress, hyperosmolarity, nitrogen starvation, ethanol concentration, and preservative compounds like sulphites. Temperature is one of the key parameters influencing alcoholic fermentations.

New trends in winemaking

The use of low temperatures (12 – 15 ºC) in fermentation is a common practice in the wine industry as it increases the retention of flavour volatile compounds (aromas) (Molina et al., 2007). However, decreasing temperatures in fermentations carried out by *S. cerevisiae* strains might cause fermentation problems. When *S. cerevisiae* ferments at low temperatures, its lag phase increases and growth rate decreases leading to a higher risk of halted or sluggish fermentation (Blateyron and Sablayrolles, 2001). Another important challenge in the winemaking industry is the climate change responsible for the modifications in the composition and properties of the grape must (Borneman et al., 2013). The increase of temperature due to climate change results in an earlier grape maturation which causes an imbalance between the sugar content and the phenolic maturity in the grapes. A direct consequence of this is an increment of the ethanol content and a higher stringency of the final of wines. To avoid these risks, wineries demand yeasts with lower ethanol yields and higher glycerol and mannoprotein productions as these compounds balance the wine astringency (White et al., 2006). New *Saccharomyces* yeasts have attracted the winemaking industry's attention as a poor exploited resource of yeast biodiversity. *Saccharomyces* species and their interspecific hybrids can solve the problems of the wine industry associated with climate change (Pérez-Torrado et al., 2018). Our group focuses on *S. uvarum* and *S. kudriavzevii* species as alternative yeast starters to accomplish the wine industry's

latest challenges (Gamero et al., 2013; Minebois et al., 2020; Peris et al., 2016; Stribny et al., 2016a; Tronchoni et al., 2012). In the next paragraphs, I will describe the fermentation characteristics that have turned both species into outstanding candidates for the wine industry.

S. kudriavzevii is a cryotolerant species that has not yet been found in anthropogenic environments but its hybrids with *S. cerevisiae* have been found in wine and beer fermentation at low temperatures (Erny et al., 2012; Gallone et al., 2019; Gangl et al., 2009; González et al., 2008, 2007; Langdon et al., 2019). Its optimal growth temperature is around 23 ºC (Salvadó et al., 2011) which is lower than the optimal growth of *S. cerevisiae* (>30ºC). This is an advantage to retain aromas in wine fermentations. However, the fermentative performance of *S. kudriavzevii* is not at the same high level as *S. cerevisiae*. Metabolic studies comparing both species during wine fermentations showed how *S. kudriavzevii* mainly directs the carbon flux towards the glycerol biosynthesis under low pH, high sugar concentrations and low temperatures instead of ethanol (Arroyo-López et al., 2010a). It has also been reported that *S. kudriavzevii* produces more biomass and glycerol than *S. cerevisiae* in chemostat cultures at 12 ºC. This increased glycerol yield was correlated with a higher expression of the glycerol-3-phosphate dehydrogenase 1 (Gpd1p) (Oliveira et al., 2014). A metabolic study revealed that *S. kudriavzevii* presented an increased NAD+ synthesis in response to cold temperatures (López-Malo et al., 2013).

S. uvarum have been found in industrial fermentations carried out at low temperatures (Demuyter et al., 2004; Naumov et al., 2000b) having good fermentative capabilities, performing even better than *S. cerevisiae* under the same conditions. It has also shown to produce higher amounts of glycerol and lower amounts of acetic acid and ethanol when compared to *S. cerevisiae* (Castellari et al., 1994; Rainieri et al., 1999). Additionally, *S. uvarum* has been characterized by its capability to release desirable flavour components, particularly 2-phenylethanol and 2-phenylethyl, which

provides stronger aromatic intensity in wines (Gamero et al., 2013). Differences in gene regulations and enzymatic activities have been identified as responsible for the differences in aroma profiles between *S. uvarum* and *S. cerevisiae* (Gamero et al., 2014; Stribny et al., 2015, 2016a). It is believed, based on metabolic studies at low temperatures, that the strategy of *S. uvarum* to adapt to cold temperatures is an increase of the shikimate pathway activity (López-Malo et al., 2013)

Molecular mechanisms involved in generating genomic diversity in *Saccharomyces*

The studies of natural *Saccharomyces* isolates have revealed a tremendous phenotypic diversity explained by a corresponding genetic variation. The molecular mechanisms involved in promoting these genetic changes are the main drivers of the existing yeast biodiversity. The study of those mechanisms involved in adaptation is an important task to understand biodiversity, population structure, and the evolutionary history of *Saccharomyces* yeasts. Mutations are the ultimate source of genetic variation, the consequence of evolutionary forces such as natural selection, genetic drift and gene flow. According to the neo-Darwinian theory of evolution, natural populations contain enough variation to respond to any sort of selection. This genetic variation is the result of the occurrence of different alleles originated by mutation and homologous recombination. The gradual accumulation of minor changes in allele frequencies, due to the natural selection action, resulting in adaption, or due to fixation by genetic drift, explains evolution. Organisms have different molecular mechanisms to respond to environmental stresses and to evolve the corresponding adaptive functions. In particular, unicellular organisms can show a quick adaptation to different environmental changes in diverse niches. Comparative genomics of the same organisms adapted to grow in distinct environments allows identifying the

underlying mechanisms of adaptive evolution. Studies using budding yeasts have revealed several molecular mechanisms responsible for adaptation in both natural and experimentally evolved populations (reviewed at Dujon and Louis (2017); Marsit et al. (2017)). Adaptation can be accomplished through small-scale nucleotide changes like base insertions, deletions or substitutions. These modifications can change gene expression, protein structures and protein-protein interactions. Large-scale genome changes like chromosome duplications, translocations or rearrangements are also drivers of adaptation in yeasts. These genomic changes may alter the gene expression by modifying the genomic context or the gene dosage through copy number variations (CNV). Additionally, it is widely accepted that organisms can evolve through reticulated evolution by exchanging genetic material. In the next subsections, the main mechanisms involved in generating genetic diversity responsible for adaptive mechanisms will be addressed.

Single Nucleotide Polymorphisms (SNPs)

SNPs are single nucleotide positions in DNA at which different sequence alleles exist in individuals within the same population or species. These molecular changes in the DNA sequence correspond to the single nucleotide substitutions or small nucleotide insertion-deletions (indels). SNPs and indels arise due to errors in DNA replication or repair system. The nucleotide polymorphisms have become very useful as genetic markers to reveal the evolutionary histories of populations and to establish the phylogenetic relationships within *S. cerevisiae* populations (Gallone et al., 2018; Liti et al., 2009; Peter et al., 2018) and other *Saccharomyces* species (Almeida et al., 2014; Peris et al., 2016). SNPs placed in coding regions can mainly affect protein structure, protein function and protein-protein interactions while SNPs in regulatory regions could alter the gene expression. Examples of such polymorphism have been found in natural *S. cerevisiae* strains where point mutations in the transcriptional factors *IME1, RME1*

and *RSF1* were associated with an improved sporulation efficiency (Gerke et al., 2009). Another example of polymorphism observed at a coding sequence was reported in the *SSU1* gene involved in sulphite resistance (Aa et al., 2006).

Gene duplication

Gene duplication is a molecular mechanism of great importance as it is believed to be the primary source of new gene and function emergence (Ohno, 1970). Population genetics theory formulates that most duplicated genes return to single copies soon after duplication. The main reason is that duplicate redundancy causes the relaxation of the selective constraints in one of the two copies which can accumulate deleterious mutations leading to decay and erosion and becoming a non-functional pseudogene (Ohno, 1970; Zhang, 2003). Many retained duplicates are conserved because the increase in the gene dosage (increased transcription levels) might confer a selective advantage. For example, *SUL1*, the gene encoding a high-affinity sulfate permease, was found duplicated in *S. cerevisiae* strains evolved in a sulfate-limited environment (Payen et al., 2014). Increases in the gene dosage are not always beneficial, and sometimes it can cause stoichiometric imbalance (Papp et al., 2003) or have fitness costs (Kondrashov, 2012; Kondrashov and Kondrashov, 2006). Therefore, a considerable number of duplicates are retained through other mechanisms, probably by the acquisition of new functions (neo-functionalization) or by partitioning the ancestral function between the paralogues (sub-functionalization) (He and Zhang, 2005; Lynch and Katju, 2004). The pair of paralogues encoding two alcohol-dehydrogenases, *ADH1* and *ADH2*, is an excellent example of a neo-functionalization in yeasts. While *ADH1* maintains the ancestral function, *ADH2* appears to have diverged to reconverts ethanol to acetaldehyde. This neofunctionalization has been proposed as one of the critical mechanisms underlying the Crabtree effect (Thomson et al., 2005). A sub-functionalization example in

Saccharomyces yeasts can be found in the galactose use pathway. The *GAL1-GAL3* paralogous pair evolved from the same ancestral gene copy that encodes a protein with both transcriptional regulator and galactokinase functions. After gene duplication, each paralogous gene specialized in one of the original functions: *GAL1* encodes a galactokinase and *GAL3* a transcriptional regulator of the GAL genes (Hittinger and Carroll, 2007).

The main mechanisms involved in gene duplication are: (i) the duplication of a single gene or segment of surrounding genes, (ii) the chromosome duplication (aneuploidy), (iii) the duplication of the whole-genome. Small-scale duplications or segmental duplications (SSD) are those that involve a single gene or group of adjacent genes. Many multigene families are encoded in the yeast genomes with three or more duplicated genes indicating the occurrence of consecutive or segmental duplications. Tandem gene duplications are widespread in yeasts (Dujon et al., 2004) and especially abundant in subtelomeric regions (Brown et al., 2010). A whole-genome duplication (WGD) is a process by which the whole set of chromosomes is doubled. The sequencing of *S. cerevisiae* allowed the identification of conserved blocks of duplicated genes, indicating an ancestral duplication of the entire genome (Wolfe and Shields, 1997), also known as the WGD event. Subsequent studies revealed that this event occurred just before the separation of *Vanderwaltozyma polyspora* originating a clade of post-WGD species (Figure 4). This genome duplication was followed by extensive genome rearrangements and gene losses that shaped the post-WGD species genomes with only a minor fraction of the WGD-derived duplicates (ohnologues) retained (Gordon et al., 2009; Seoighe and Wolfe, 1998). The high levels of synteny showed by the duplicated gene blocks led to the hypothesis that the WGD event's origin was an autopolyploidization (Gordon et al., 2009). However, a later study with whole-genome sequencing data of more species demonstrated that the most likely hypothesis that explains the WGD event was an ancient hybridization and subsequent duplication (allopolyploidization) (Marcet-Houben and Gabaldón, 2015).

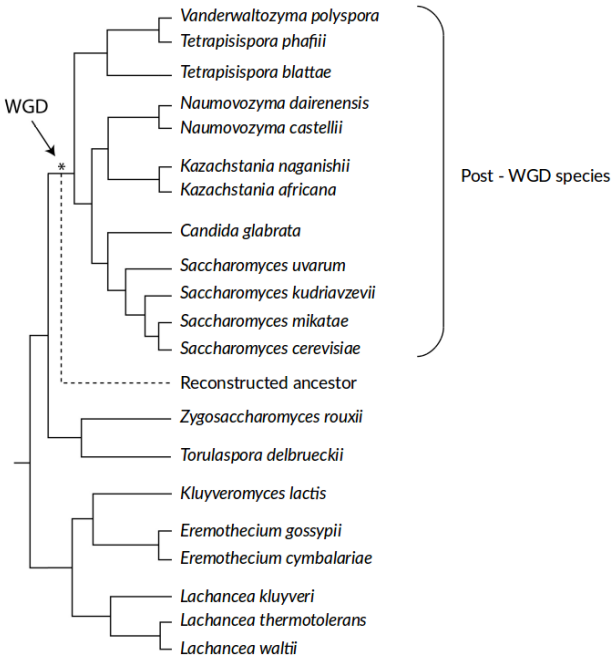


Figure 4: Species phylogeny of the pre- and post-WGD species. Phylogeny of species from the Saccharomycotina subphylum. Adapted from Byrne and Wolfe (2005)

The *S. cerevisiae* genome contains approximately 2,000 pairs of paralogues of which 554 are ohnologues derived from the WGD (Byrne and Wolfe, 2005), while the remainder originated from SSD. Many of these paralogues could have been retained because they might provide a selective advantage. In budding yeasts, they contributed to metabolic innovation by altering the regulatory networks and increasing the glycolytic fluxes (Conant and Wolfe, 2007). Finally, another frequent mechanism by which duplicates can emerge in yeast genomes is through changes in chromosome copy numbers, also known as aneuploidies. Aneuploidies are generated by chromosomal non-disjunction during mitosis or meiosis, resulting in a disproportion of gene products and disruption of their interactions. Although aneuploidies have been proven to be very harmful to many organisms, they are widespread in yeasts and can be advantageous under stressful environmental conditions (Selmecki et al., 2006, 2015). The most critical consequence of aneuploidy is the change in the gene dose, which allows adaptation to unstable environments (Bakalinsky and Snow, 1990). An example of the advantage of aneuploidies in yeast was reported in *S. cerevisiae* wine strains, where

an extra-copy of chromosome III was correlated to increased ethanol tolerance in wine strains (Morard et al., 2019).

Gross chromosomal rearrangements (GCRs)

Gross chromosomal rearrangements (GCRs) events are mediated through ectopic (non-homologous) recombination between repeated sequences such as Ty sequences or other repeated sequences (Rachidi et al., 1999). GCRs have a crucial role in the early stages of adaptation as they can cause changes in the gene expression of one or multiple genes. Chromosomal rearrangements have been observed to increase copper tolerance in *S. cerevisiae* strains isolated from Evolution Canyon (Israel) (Chang et al., 2013). Another case of chromosomal rearrangement driving adaptation was reported in *S. cerevisiae* wine strains where a chromosome translocation resulted in the overexpression of the *SSU1*, the gene responsible for the sulphite detoxification of the cell (Pérez-Ortín et al., 2002), leading to strains more tolerant to the presence of sulphites in wine fermentations.

Reticulated evolution in *Saccharomyces*

Reticulated evolution, also known as nonvertical inheritance, refers to the interchange of the total or partial genetic material between different lineages. This phenomenon is responsible for the arising of evolutionary histories that can not be represented using bifurcating trees as a phylogenetic network better represents them. One of the main molecular mechanisms causing this gene flow is the interspecific hybridization, which is very frequent due to their weak prezygotic barriers. Hybridization produces a genetic exchange between divergent lineages that can generate new traits, allowing the adaptation to novel environments through the expression of hybrid vigour (Bell and Travis, 2005). *Saccharomyces* interspecific hybrids have been

found in almost all combinations of species and they are recurrent in anthropic environments such as industrial fermentations (Langdon et al., 2019; Morard et al., 2020; Pérez-Través et al., 2014). *S. pastorianus* is probably the most outstanding hybrid example within *Saccharomyces*. This hybrid between *S. cerevisiae* and *S. eubayanus* is responsible for the lager beer fermentations (Libkind et al., 2011). Hybridization followed by backcrossing with the parental lineage can lead to genomic introgressions which are particular genomic regions of one species integrated into the genome of another species. The mechanisms by which introgressions arise in reproductively isolated species, like yeasts, have always been unclear. However, a recent study has shed light on this (Angiolo et al., 2020), by demonstrating how genomic instability in the ancestral yeast hybrid, founder of an *S. cerevisiae* alpechin lineage with *S. paradoxus* introgressions, generated homozygous genome blocks that allows fertility's restoration and caused the origin of the introgressions. Introgressions also have key roles in adaptive phenotypic traits of domesticated *Saccharomyces*. It has been reported that genomic regions corresponding to different *Saccharomyces* species, especially *S. eubayanus*, have been introgressed into European *S. uvarum* strains isolated from wine and cider fermentations (Almeida et al., 2014). These regions are enriched in genes related to nitrogen metabolism, demonstrating their potential adaptive role. Wine *S. cerevisiae* strains have also shown introgressed genomic regions of other *Saccharomyces* (Borneman et al., 2016) highlighting the importance of introgression in adapting to anthropic environments where these genomic events are more frequent than in nature. Horizontal Gene Transfer (HGT) occurs when a species acquires new genes from a distantly related species or a species that does not belong to its ancestry line. The main difference between HGT and introgression is that in the former, the sexual reproduction (hybridization) is not the only mechanism responsible for the genetic transfer. Other mechanisms like transformation, bacterial conjugation, and transduction can be involved part in the process. HGT has been reported as a mechanism that facilitates the adaptation to novel environments by the acquisition of

new biological functions (Keeling and Palmer, 2008). Three large genomic regions (A, B and C regions) have been acquired by wine *S. cerevisiae* strains from distantly related yeasts (Novo et al., 2009). In particular, region C, a region of 158 kb length, was transferred from *Torulaspora microellipsoides* to *S. cerevisiae* conferring a robust competitive advantage during wine fermentation (Marsit et al., 2015a).

Comparative genomics is a powerful tool to identify molecular mechanisms involved in adaptation

Whole-genome sequencing

Genomic studies of yeasts have increased considerably in the last years (Libkind et al., 2020). The advances in the next-generation sequencing technologies (NGS) and their decrease in costs have promoted a considerable increase in the number of published whole-genome sequencing projects. Since the sequencing of *S. cerevisiae*, lots of yeasts have been sequenced, allowing an improvement of our understanding of the evolution and ecology of yeasts. Of the more than 60,000 DNA sequencing projects that include the sequencing of any of the *Saccharomyces* species (excluding hybrids), more than 90% belong to *S. cerevisiae* sequencing projects (Figure 5). *S. paradoxus* and *S. eubayanus* are the second and third, respectively, most-sequenced *Saccharomyces* species. There is a significant imbalance in the sequencing efforts with most *Saccharomyces* genomes considerably underrepresented compromising the success of the comparative genomics analyses using the entire *Saccharomyces* genus. Comparative genomics analyses applied to a whole genus can shed light on its evolution and the molecular mechanisms underlying the adaptation of its species to different environments.

Whole-genome sequencing has been used to identify large-scale structural

Figure 5: *Saccharomyces* whole-genome sequencing projects pie chart. Total number of whole-genome sequencing (WGS) projects submitted to the SRA (Sequence Read Archive (NCBI), `https://www.ncbi.nlm.nih.gov/sra`) of *Saccharomyces* species (excluding hybrids) was collected and represented in this pie chart. Slices are color-coded according the *Saccharomyces* species sequences. Last accession to SRA database: October 2020.

variants such as inversions or reciprocal translocations which are common among yeast genomes (Yue et al., 2017). The used sequencing technology has a critical impact on the potential to identify this kind of variantions. Short-read sequencing technologies (second-generation sequencing technologies) are currently the most used due to their low costs and low error rates. However, short-length reads (up to 300 bp) make difficult, and sometimes even impossible, to assemble particular genomic regions such as repetitive subtelomeric regions and regions with structural variations between samples. The third-generation sequencing technologies like Oxford Nanopore and PacBio SMRT offer an opportunity to succeed in the assembling of these regions due to the long length of their reads (>1,000 bp). However, they should always be combined with short-read technologies as they have a higher error rate. These long read technologies have been used to provide new high-quality end-to-end genome assemblies of some *Saccharomyces* species (Brickwedde et al., 2018; Naseeb et al., 2018; Yue et al., 2017) while other genome references (*S. uvarum, S. kudriazevii* or *S. mikatae*) remain outdated.

Saccharomyces phylogenomics

Population genomics studies have become very popular to identify the different lineages within the same species and to study genomic and phenotypic variation. These approaches are beneficial when there is no gene flow between populations. *S. cerevisiae* population genomics studies have shed light on the evolutionary history of this species and how their lineages grouped in the phylogenetic tree according to their ecology and isolation source (Gallone et al., 2018; Liti et al., 2009; Peter et al., 2018). These approaches have been successfully applied also to other species of the *Saccharomyces* genus like *S. paradoxus* (Xia et al., 2017), *S. arboricola* (Gayevskiy and Goddard, 2016), *S. uvarum* (Almeida et al., 2014) and *S. eubayanus* (Nespolo et al., 2020; Peris et al., 2014). However, more in-depth analyses of the under-sequenced *Saccharomyces* species' can not be accomplished yet due to the scarcity of genomic data.

Detecting natural selection in genomic data

Natural selection's action follows a simple principle: those traits conferring increased chances of survival and reproduction will have a higher probability of being transmitted to the next generation, and tend to become more frequent in population over time (Darwin, 1859). According to the neutral theory (Kimura, 1983) postulated later, most of the emerged mutations are selectively neutral or deleterious. Advantageous mutations responsible for adaptation then constitute a small fraction of the total emerged mutations. There are various forms of natural selection, such as negative (purifying) selection, balancing selection, or positive (adaptive) selection. Positive selection is the most studied form of selection. It occurs when an allele is favoured by the action of the natural selection, increasing the frequency of the favoured allele over time, leading to its potential fixation in the population. An allele's fixation

can leave signatures locally in the genome that can be studied through statistical methods on testable hypotheses. The most widely used statistical method is the dN/dS ratio. This measure is based on the assumption that synonymous (dS) DNA mutations accumulate neutrally, and hence the selective pressure on non-synonymous (dN) can be measured by the ratio dN/dS (ω), and would therefore be equal to one in a neutral evolution scenario and lower than one if the purifying selection is acting to maintain the functional constraints of the coding gene. This ratio might be greater than one when positive selection favours adaptive amino-acid changes in the encoded protein sequence, which means, non-synonymous substitutions. Identification of coding-sequences under positive selection among different lineages can unveil adaptive mechanisms. This method applied at a genome-wide level allows identifying gene candidates responsible for adaptation among different lineages. Some studies have explored the action of natural selection in some *Saccharomyces* species (Kawahara and Imanishi, 2007), and in particular genes of *S. cerevisiae* (Becker-Kettern et al., 2016; Oppler et al., 2019). However, there is no study exploring the action of natural selection at the genome level using whole-genome sequencing data of different *Saccharomyces* species.

Objectives

Saccharomyces genus is a fascinating model for evolutionary biology due to its high genetic and phenotypic diversity. The improvement in the sampling efforts during the last decade has resulted in the isolation of *Saccharomyces* strains from a wide range of sources worldwide. *S. cerevisiae* is probably the most well known eukaryotic system and the dominant organism in most industrial fermentations. Due to the new challenges of the winemaking industry (e.g., global warming and changing customer demands), alternative *Saccharomyces* species have attracted researchers' attention during the last decade as a poorly exploited resource of biodiversity. Among them, *S. kudriavzevii* and *S. uvarum* appear as two promising candidates because of their lower optimum growth temperature, higher glycerol production and different aromatic profiles when compared to *S. cerevisiae*. However, the mechanisms involved in the adaptation of these species to their true ecological niches remain unexplored. With the advent of the NGS technologies, comparative genomics studies applied to identify adaptive mechanisms have considerably increased during the last years, especially those studies focused on *S. cerevisiae*. *Saccharomyces* species' sequencing efforts are completely unbalanced, resulting in the scarcity of sequences of non – *S. cerevisiae* species. This lack of data complicates the use of comparative genomics to understand the adaptive mechanisms acting in the different species of the genus. Taking these

previous issues into consideration, the present thesis is aimed at achieving the following objectives:

1) The analysis of the genomic differences between *S. cerevisiae* and *S. kudriavzevii* by applying a comparative genomics approach of their coding-sequences.
    a) The sequencing, assembly and annotation of new genomes of *S. kudriavzevii*.
    b) The study of functional divergence between *S. cerevisiae* – *S. kudriavzevi*i orthologous genes.
    c) The identification of genes under positive selection and having accelerated evolutionary rates.
2) The study of signatures of positive selection in *S. uvarum*.
    a) The development of a new tool to facilitate the detection of positive selection signatures at the genome-wide level.
    b) The identification of genes under positive selection in *S. uvarum*.
3) The exploration of the genomic footprints of domestication in *S. uvarum*.
    a) The identification of different recombination events in *S. uvarum* isolated from anthropic environments.
    b) The analysis of genotype-phenotype correlations in *S. uvarum*.
    c) The elucidation of the origin of the convergent adaptation in *S. uvarum* populations.
4) The comparative study of genomic differences among all *Saccharomyces* species using high-quality genome references.
    a) The sequencing, assembly and annotation of *S. uvarum, S. kudriavzevii* and *S. mikatae* using both short and long-read technologies.
    b) The study of the pangenome of the S*accharomyces* genus.

c) The analysis of subtelomeric region variability in *Saccharomyces* species and the dynamics of subtelomeric gene families.

This doctoral thesis was organized into four chapters:

Chapter 1, *Comparative genomics between S. kudriavzevii and S. cerevisiae applied to identify mechanisms involved in adaptation* corresponds to objective 1.

Chapter 2, *GWideCodeML: a Python package for testing evolutionary hypotheses at the genome-wide level and its application in detecting signatures of positive selection in S. uvarum*, addresses objective 2.

Chapter 3, *Convergent adaptation of S. uvarum to sulphite, an antimicrobial preservative widely used in human-driven fermentations* deals with objective 3.

Chapter 4, *High-quality new assemblies of Saccharomyes genomes provide insights into their evolutionary dynamics*, which aims to unravel objective 4.

CHAPTER 1

Comparative genomics between *Saccharomyces kudriavzevii* and *Saccharomyces cerevisiae* applied to identify mechanisms involved in adaptation

1.1 Introduction

How species have adapted to new environments by the action of natural selection shaping their genomes is a key question in modern biology since Charles Darwin proposed the theory of natural selection to explain the origin of adaptations. The Modern Synthesis (Neo-Darwinism), reconciling Darwin's theory of evolution and Mendelian genetics, was based on the idea that most natural populations contain

This chapter is published in Macías et al. (2019) *Frontiers in Genetics* .

enough genetic variation, generated by mutation, to respond to any sort of selection, and explained adaptation as the gradual evolution resulting from changes in the frequencies of the genetic variants acted upon by natural selection. However, with the proposal of the neutral theory of molecular evolution (Kimura, 1983), it has widely been assumed that most mutations are neutral or deleterious, depending on their functional constraints. In contrast, advantageous mutations constitute a very small fraction of the total but are responsible for adaptation. As deleterious mutations are removed by purifying selection, most genetic variation, both within-species polymorphisms and between-species divergence, is the result of the action of genetic drift, with a negligible contribution of the rare beneficial mutations fixed by positive selection (Kimura, 1983). In recent years, several authors propose conciliation between neutralism and selectionism by considering that fixed neutral mutations can become advantageous by shifts in the selective pressures, and hence, promote later evolutionary adaptation (Wagner, 2008). According to Lynch (2007)): "the non-adaptive force of random genetic drift set the stage for future paths of adaptive evolution in novel ways that would not otherwise be possible."

In the genomic era, an important challenge is to determine whether patterns of genome variation can be explained by random genetic drift or selection. However, the rapid acquisition of more and more genome sequences, together with the development and improvement of statistical methods for comparative genomics, allow us to unveil the evolutionary forces responsible for adaptation at the molecular level.

The genus *Saccharomyces* is composed of eight species (Boynton and Greig, 2014): *S. arboricola, S. cerevisiae, S. eubayanus, S. kudriavzevii, S. mikatae, S. paradoxus, S. uvarum,* and the recently described *S. jurei* (Naseeb et al., 2017). Yeasts belonging to this genus have mostly been isolated from wild environments. The exception is *S. cerevisiae* , one of the most well-studied microorganisms, which has also been found in a wide range of human-manipulated fermentative environments

such as wine, cider, sake, beer, bread, etc., as well as in traditional fermentations (Gallone et al., 2018; Liti et al., 2009; Peter et al., 2018). In a lesser extent, *S. uvarum* is also present in wine and cider fermentations from regions of cold climate, where coexists or even replaces *S. cerevisiae* (Almeida et al., 2014; Rodríguez et al., 2017). In addition, different types of interspecific *Saccharomyces* hybrids have also been isolated in fermentations from cold regions (González et al., 2006; Morales and Dujon, 2012; Pérez-Través et al., 2014). Another interesting species from this genus is *S. kudriavzevii*. This species is isolated only from wild environments, such as oak barks and decayed leaves in Asia (Naumov et al., 2000a, 2013) and Europe (Erny et al., 2012; Lopes et al., 2010; Sampaio and Gonçalves, 2008). Although *S. kudriavzevii* has never been found in fermentations, its double hybrids with *S. cerevisiae* and triple with *S. cerevisiae* and *S. uvarum* appear, and even dominate, in wine, beer and cider fermentations in regions of cold climates (Peris et al., 2018).

To understand the contribution of the *S. kudriavzevii* parent to its hybrids, several comparative physiological studies between *S. cerevisiae* and *S. kudriavzevii* have been performed (Arroyo-López et al., 2009; Belloch et al., 2008; Gangl et al., 2009; González et al., 2007). This way, these results indicate that hybrids acquired the high alcohol tolerance trait of *S. cerevisiae* (Arroyo-López et al., 2010b), and the better adaptation to grow at low temperatures of *S. kudriavzevii* (Salvadó et al., 2011). These physiological differences have been related to modifications in the components of lipid membrane of both species (Tronchoni et al., 2012), and in the production of glycerol (Arroyo-López et al., 2010a). The lower ethanol yield and the higher glycerol synthesis, together with differences in the aroma production (Stribny et al., 2015) and an optimal growth under low pH (Arroyo-López et al., 2009) indicate that *S. kudriavzevii* and its hybrids are good potential candidates for future applications in the wine industry (Alonso-del Real et al., 2017; Pérez-Torrado et al., 2018).

At the same time, different studies have been performed to unravel the genetic

basis responsible for the phenotypic differences observed between *S. cerevisae* and *S. kudriavzevii*, especially for those that study the low temperature adaptation. The analysis of the glycerol synthesis pathway showed that the higher glycerol production in *S. kudriavzevii* is due to an enhanced enzymatic activity of its glycerol-3-phosphate dehydrogenase Gpd1p (Oliveira et al., 2014). A transcriptomic study revealed that *S. kudriavzevii* exhibits a higher ability to initiate the translation of crucial genes in cold adaptation (Tronchoni et al., 2014). A systems biology study applied to both *S. cerevisiae* and *S. kudriavzevii* revealed that pathways such as lipid, oxidoreductase and vitamin metabolism were directly involved with the fitness of these species at low temperatures (Paget et al., 2014). Main phenotypic differences between *S. kudriavzevii* and *S. cerevisiae* are described but most of genes responsible for these phenotypes remain unknown.

In the present study, we applied for the first time diverse comparative approaches to study adaptive differences and functional divergence between both *Saccharomyces* species at genome-wide level. We sequenced, *de novo* assembled and annotated two new genomes of *S. kudriavzevii* strains isolated from Spanish tree barks. We used complete genome sequences of these strains as well as those from two *S. kudriavzevii* previously sequenced (Scannell et al., 2011), and four genome sequences from representative *S. cerevisiae* strains, to identify selective shifts in a set of orthologous genes in both *S. cerevisiae* and *S. kudriavzevii* leading branches. Functional divergence among orthologous proteins was also quantified leading to the identification of the most functional divergent pathways between *S. kudriavzevii* and *S. cerevisiae*.

1.2 Materials and methods

1.2.1 Assembly and annotation

Saccharomyces kudriavzevii strains CR85 and CA111 were isolated in a previous work (Lopes et al., 2010) and their genomes were sequenced in this study. These strains were sequenced by Illumina MiSeq with paired-end 300 bp reads. In addition, *S. kudriavzevii* CR85 was also sequenced using Roche 454 shotgun sequencing and paired-end reads of 8 kb.

De novo assembly of the *S. kudriavzevii* CR85 genome was carried out using MIRA v3.4.1.11 and GS de novo Assembler (Roche/454 Life Sciences, Branford, CT, United States). Manual checking and corrections of the assembly were done using Consed (Gordon et al., 1998).

Assembly of CA111 strain was performed using Velvet v1.1 (Zerbino and Birney, 2008) to determine the best k-mer value and then Sopra v1.4. (Dayarian et al., 2010) was used for *de novo* assembly. To get the scaffolds into chromosome structure, ultra-scaffolds were generated with an in-house script, which orders the contigs according to their homology to a reference genome, in our case *S. cerevisiae* S288c. A whole genome aligner MUMmer (Kurtz et al., 2004) was used to generate this information. Final assembly sizes of 11.75 and 11.89 Mb were obtained for *S. kudriavzevii* CR85 and *S. kudriavzevii* CA111, respectively (Table S1.1).

Reannotations of *S. kudriavzevii* IFO1802 and ZP591 genomes (Scannell et al., 2011) were also performed due to problems with the original annotations. Two approaches were used for the annotation of the four *S. kudriavzevii* genomes: first, a transfer of annotations from *S. cerevisiae* S288c (Goffeau et al., 1996) by using RATT tool (Otto et al., 2011), and second, a novel gene prediction with Augustus (Stanke

and Morgenstern, 2005). Finally, the annotations were manually verified by using Artemis (Rutherford et al., 2000). With this pipeline, 5664 genes in IFO1802 strain were annotated, 5575 in ZP591 strain, 5623 in CR85, and 5492 in CA111.

1.2.2 Orthology and alignment

We also used four well-annotated genome sequences from different populations of *S. cerevisiae* (Liti et al., 2009), as representative strains of this species (Table 1.1). The genome sequence of *Torulaspora delbrueckii* (Gordon et al., 2011) was used as outgroup. This species was selected because it diverged from the *Saccharomyces* genus before the Whole Genome Duplication (WGD) event (Wolfe and Shields, 1997). This was done to ensure the use of orthologous reference sequences in the analyses, which is not necessarily true if a post-WGD species is selected as outgroup due to differential loss of paralogous (ohnologous) genes (Scannell et al., 2007). Orthology among the three species was defined according to synteny information available in the Yeast Genome Order Browser (YGOB) (Byrne and Wolfe, 2005).

Alignments for all orthologous sequences were obtained using Mafft v7.221 (Katoh and Standley, 2013). A total number of 4164 orthologous genes were found in common among the three species. In some cases, as *T. delbrueckii* was a pre-WGD species, the same gene sequence was aligned against two different gene sequences of *Saccharomyces* genomes, those duplicated genes generated by the WGD event, according to the YGOB.

1.2.3 Signatures of positive selection

To identify genes being potentially under positive selection in both *S. kudriavzevii* and *S. cerevisiae* branches, we performed a comparison of the likelihood scores of selection models implemented in the branch-site CodeML software of the PAML

Table 1.1: List of strains and sources of the genomic sequences used in Chapter 1.

Strain	Origin	Source	Reference
S. kudriavzevii CR85	Ciudad Real, Spain	*Quercus ilex* bark	This study
S. kudriavzevii CA111	Castellón, Spain	*Quercus ilex* bark	This study
S. kudriavzevii ZP591	Cast. Vide, Portugal	*Quercus pyrenaica* bark	(Scannell et al., 2011)
S. kudriavzevii NBRC1802	Japan	Decayed leaf	(Scannell et al., 2007)
S. cerevisiae T73	Alicante, Spain	Wine	(Morard et al., 2019)
S. cerevisiae S288c	-	Laboratory	(Goffeau et al., 1996)
S. cerevisiae YPS128	Pennsylvania, USA	*Quercus* forest soil	(Liti et al., 2009)
S. cerevisiae Y9	Indonesia	Ragi (similar to sake)	(Liti et al., 2009)

package, version 4.5 (Yang, 2007). The branch-site test was used to detect positive selection acting at specific codons in a defined branch of a phylogenetic tree. This branch is known as the foreground and the rest as background branches. The branch-site test compares a model considering three fractions of codon sites with a null model with only two fractions of codons. In the three-fraction model, the first fraction (p_0) evolved in both foreground and background branches with a non-synonymous/synonymous substitution ratio of $\omega_0 < 1$ (purifying selection), the second (p_1) with $\omega_1 = 1$ (neutral) in both sets of branches, and the third (p_2) evolved with $\omega_0 > 1$ (positive selection) in the foreground branch but with $\omega_0 < 1$ or $\omega_1 = 1$ in the background branches. The null model considers only two fractions, one evolved with $\omega_0 < 1$ and the other with $\omega_1 = 1$ in both sets of branches. Since this is a species-based method, we first set as the foreground branch the one leading to the *S. kudriavzevii* clade. Then, we repeated the analysis by setting as foreground the *S. cerevisiae* clade branch. Both analyses were performed using *T. delbrueckii* as an outgroup species.

All genes whose Likelihood Ratio Test (LRT) χ^2 analysis, with one degree of freedom (the difference of free parameters between models), reached *p*-values lower

than 0.05 were considered significant, and those genes containing a fraction of codons with ω > 1 were selected as gene candidates to be under positive selection. In these genes, Bayesian posterior probabilities for site classes were estimated, with the Bayes Empirical Bayes (BEB) method (Yang et al., 2005), to identify amino acid sites under positive selection.

1.2.4 Testing constant rate of evolution

The molecular constant rate of evolution was tested for all orthologous genes analysed in both *S. cerevisiae* and *S. kudriavzevii* species, by using *T. delbrueckii* as outgroup. Tajima relative rate test (Tajima, 1993) was implemented by using an in-house built Python script. A singleton was defined as a change in the nucleotide sequence specific for every one of the three species included in the alignment. Number of observed singletons in each gene alignment was calculated according to the formulas:

$$m_1 = \sum_i \sum_{j \neq i} n_{ijj}$$

$$m_2 = \sum_i \sum_{j \neq i} n_{jij}$$

$$m_3 = \sum_i \sum_{j \neq i} n_{jji}$$

where *i* is the variable position in the alignment and *j* is the nucleotide that is conserved in two out of three sequences of the alignment. *m1, m2*, and *m3* are the total number of singletons in one alignment for *S. cerevisae*, *S. kudriavzevii* and the outgroup species, respectively.

Under the molecular clock hypothesis, the number of singletons in *S. cerevisiae* and *S. kudriavzevii* species are expected to be the same, therefore, the expected

singletons according to this hypothesis was calculated as:

$$E(m_1) = E(m_2) = m_1 + m_2/2$$

For every nucleotide alignment, the number of *S. cerevisae* and *S. kudriavzevii* singletons was calculated and it was compared with the number of expected changes under the molecular clock hypothesis. A χ^2 test with one degree of freedom was applied to assess whether the difference between the observed and the expected singletons was significant and, if so, the molecular clock hypothesis was rejected.

1.2.5 Functional divergence

In this work, functional divergence type I was identified. This type of functional divergence involves the change in selection constraints acting at specific amino acid sites of a protein in a specific phylogenetic clade (which will be defined as clade-of-interest) when compared to another clade. A method previously described by (Toft et al., 2009), was used to identify amino acid sites which have diverged significantly from the output sequence in a clade of interest with respect to the homologous sites in a second clade. This test was performed twice by defining as the clade-of-interest *S. kudriavzevii* or *S. cerevisiae*. Once all divergent amino acid sites were obtained, results were filtered by Grantham's scores (Grantham, 1974), to quantify the biochemical divergence between *S. cerevisiae* and *S. kudriavzevii* amino acids. Scores of 120 and higher were considered for further analyses as sites that have radically changed in *S. kudriavzevii* when compared to *S. cerevisiae* and which might have functional importance for the protein, these results were normalised by the protein length.

We also tested whether there was any genome region enriched in proteins

showing evidence of functional divergence. This task was assessed by checking if the mean of normalised functional divergence values from non-overlapping windows of ten genes fall within the 95% confidence interval resulting from generating a random distribution after sampling 10^6 times ten genes from the whole set of genes analysed.

Finally, functional divergence was also determined in terms of domain architecture. SUPERFAMILY hidden Markov models available in the SUPERFAMILY database (ver. 1.75., last accessed February 20, 2018) (Gough et al., 2001) were used for domain assignment according to the Structural Classification of Proteins (SCOP) database to get domain annotations for every *S. kudriavzevii* and *S. cerevisiae* orthologous pair of proteins using the same criteria as described in Grassi et al. (2010). Orthologous pairs with identical domain architecture, which exhibit no domain architecture functional divergence, were annotated as class A. Orthologues carrying similar domain architectures but differed in domain copy number were annotated as class B. Finally, class C contained *S. kudriavzevii-S. cerevisiae* orthologues whose domain architectures differed in the presence or absence of one or more domains.

1.2.6 Duplicated genes

A careful examination of duplicated genes was done after performing all analyses previously mentioned. We defined duplicated gene pairs as the resulting best reciprocal hits from all-against-all BLAST (Altschul et al., 1990) searches using BLASTP with an E-value cut-off of $1x10^{-5}$ and a bit score cut-off of 50. Duplicated pairs were then classified as ohnologous gene pairs, generated by the whole genome duplication event (WGDs) according to the Yeast Gene Order Browser (YGOB) list (Byrne and Wolfe, 2005). All other paralogous gene duplicates were considered as derived from small-scale duplications (SSDs).

1.2.7 Gene onthology and pathway enrichment analyses

For every analysis previously mentioned, a list of candidate genes was obtained. Gene ontology (GO) term and pathway enrichments were performed using the Gene List tool available in the *Saccharomyces* Genome Database [0], by considering the list of all 4164 aligned genes used in this study as background population. Results were filtered by a *p*-value lower than 0.05 after a Holm-Bonferroni test correction (Aickin and Gensler, 1996).

[0] https://yeastmine.yeastgenome.org/yeastmine/bag.do

1.3 Results

A species-based comparative genomics approach has been applied to investigate the genetic basis behind the main phenotypic differences already reported between *S. kudriavzevii* and *S. cerevisiae*. As representatives of *S. kudriavzevii*, we included four complete genome assemblies from strains of different origins. Those from strains IFO1802 and ZP591 were publicly available from a previous study (Scannell et al., 2011) and the other two, corresponding to strains isolated from oak bark samples taken in different locations of Spain, were sequenced, *de novo* assembled and annotated for the present study. Our assembly and annotation pipeline, that combines transfer of annotation and *de novo* gene prediction with a final accurate manual correction, allowed us to provide *S. kudriavzevii* high-quality annotation avoiding common errors such as paralogues mislabelling, coming from the sole use of automatic annotation pipelines. For this reason, IFO1802 and ZP591 genome assemblies were also re-annotated using the same pipeline. In the case of *S. cerevisiae*, we included in the analyses well-annotated genomes of four strains as representatives of the main lineages (Liti et al., 2009; Peter et al., 2018).

1.3.1 Differential adaptive evolution between *S. cerevisiae* and *S. kudriavzevii*

The presence of signatures of adaptive evolution in coding genes was tested in both species by using three different approaches: branch-site test of selection, Tajima's rate of evolution test and functional divergence test, which results are summarized in (Figure 1.1). GO and pathway enrichment analyses performed for genes showing a positive result simultaneously for more than one of the tests mentioned revealed no significant results. Using the branch-site model, we obtained 30 genes under positive selection when the branch leading to *S. kudriavzevii* was considered as foreground branch. Additionally, 32 genes were found under positive selection when *S. cerevisiae*

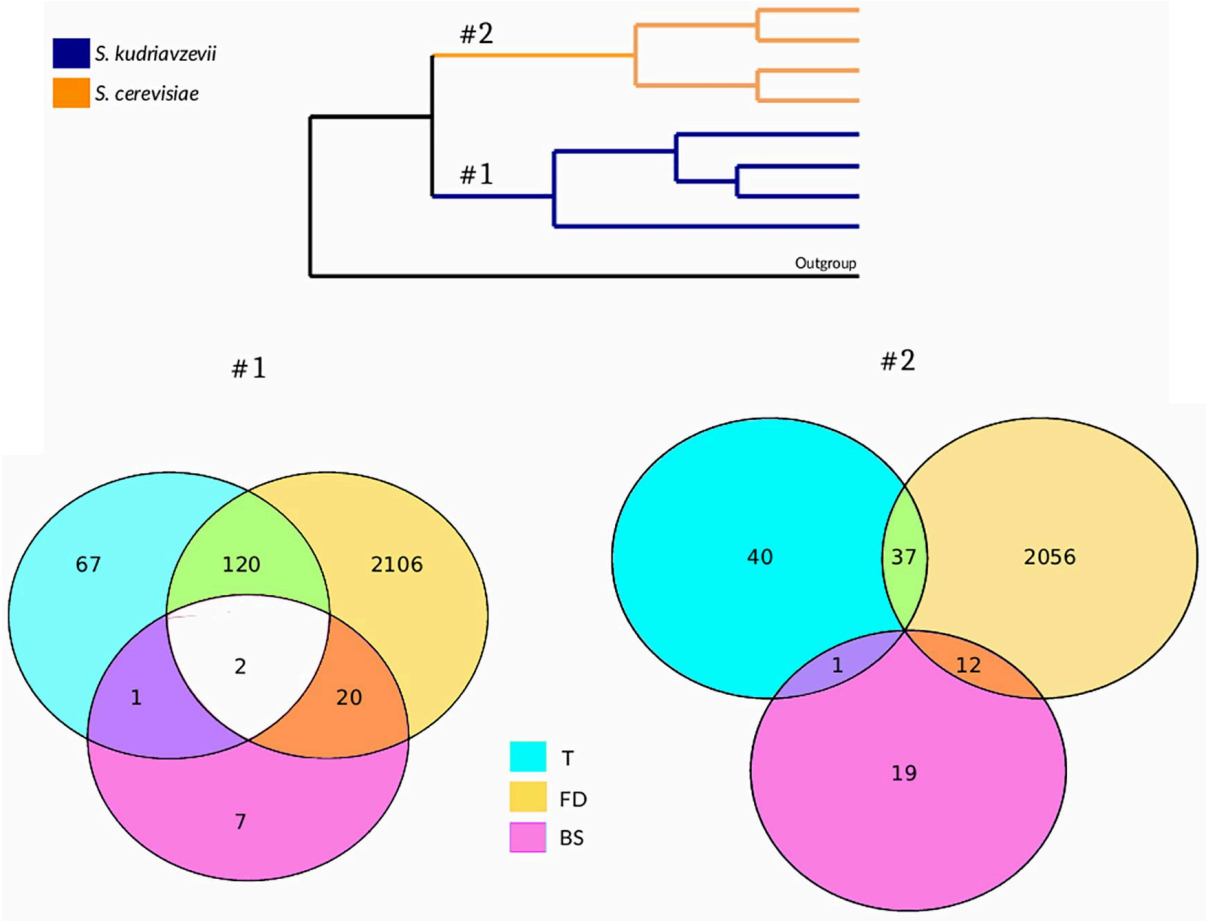


Figure 1.1: Species-based comparative genomics approach. Venn diagrams represent the number of positive genes for every statistical test performed on both *S. kudriavzevii* (#1) and *S. cerevisiae* (#2) branches of the tree to detect adaptive evolution. T, Tajima's relative rate test; FD, functional divergence test; BS, branch-site model test.

was set as the foreground branch (Table S1.2, S1.4). Neither GO nor pathway enrichment were obtained for these lists. Only two genes, *FRT2* and *RQC2*, showed evidence of positive selection on both branches.

Tajima's relative rate test was applied to detect higher rate of nucleotide substitution at specific coding sequences. Using this test, 190 genes in *S. kudriavzevii* branch and 78 genes in *S. cerevisiae* branch were obtained (Table S1.2). The difference between the numbers of genes detected on both branches was significant (Fisher's exact test: F = 2.5, *p*-value = $2.71e^{-12}$). No GO term enrichment was found for none of the two lists. No pathway enrichment was found for *S. cerevisiae* branch results whereas an enrichment of genes belonging to riboflavin pathway (*RIB2, RIB3, RIB5,* and *FMN1*) was found in *S. kudriavzevii* branch results. In the *S. kudriavzevii* branch, three genes showed an acceleration in evolutionary rates and were found to

be under positive selection: *FBA1, ZIP1,* and *RQC2*, while in the *S. cerevisiae* branch, only one gene (*STE24*) was detected in both statistical tests.

A set of proteins showing evidence of functional divergence and the specific amino acid sites that are contributing to this phenomenon was obtained for both *Saccharomyces* species (Table S1.2). Using this approach, 2248 proteins out of 4164 analysed (54%) showed evidence of functional divergence when *S. kudriavzevii* was compared to their *S. cerevisiae* orthologous proteins. On the other hand, 2105 proteins (~50%) were found to be under functional divergence when *S. cerevisiae* was set as the clade-of-interest.

To asses whether there was any region in the genome of *S. kudriavzevii* showing an enrichment in genes codifying for functionally divergent proteins, we evaluated chromosomal regions in non-overlapping windows of ten genes (Figure 1.2 and Table S1.5). One region containing ten genes in chromosomes III and IX were impoverished in proteins showing functional divergence in the *S. kudriavzevii* clade. One region in each of the chromosomes II, VII, VIII, X, XI, XIV, and XVI, two regions in chromosomes XII and XV, and five regions in chromosome IV were enriched in functionally divergent proteins.

In addition, functional divergence was evaluated in terms of protein domain architecture. Domain-based functional analysis leads us to get an additional perspective on the possible biological differences between *S. kudriavzevii* and *S. cerevisiae* orthologous proteins. SCOP domains were assigned for every *S. kudriavzevii-S. cerevisiae* orthologous pair of proteins (Table S1.2). A total number of 2544 proteins were annotated with SCOP domains for *S. cerevisiae* and 2550 for *S. kudriavzevii*. Of them, 2402 proteins were classified as class A as they showed no evidence of functional divergence in protein domain architecture. Other 54 proteins were classified as class B because they differed in copy number domain. Finally, 96 proteins were classified in class C as they carried domain architectures which differed

in the presence or absence of any of the domains of their orthologous gene. No GO or pathway enrichment were found for groups B or C. Two genes belonging to category B, *SEC7* and *SMC1*, and five genes from category C, YNL144C, *FAR1, PRP19, SMC4* and *ZIP1*, showed accelerated evolutionary rates as well.

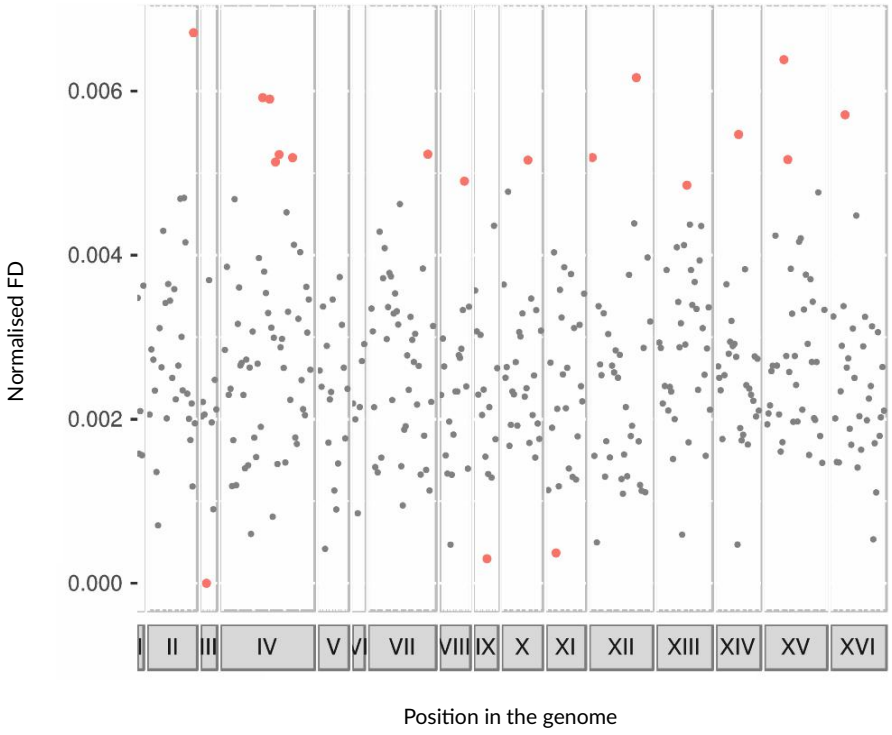


Figure 1.2: Functional divergence along *S. kudriavzevii* genome. Functional divergence along *S. kudriavzevii* genome. Enriched chromosomal regions in genes codifying for proteins with functional divergence. Points represent the mean of normalised functional divergence values of bins of ten genes and their location in every chromosome. Chromosome length is represented in x-axis and proportionally to the size. The y-axis shows functional divergence normalised values. Black points correspond to those bins that felt into the 95% confidence level of the random distribution. Red points show those bins enriched or impoverished in proteins showing evidence of functional divergence.

Contribution of genes under functional divergence to every metabolic pathway was analysed to identify those pathways more functionally divergent in *S. kudriavzevii* and *S. cerevisiae* (Figure 1.3 and Figure S1.1). Amino acid biosynthesis, glycerophospholipid metabolism, GPI-anchor biosynthesis, N-glycan biosynthesis and purine and pyrimidine metabolisms were found between the pathways containing a higher number of functionally divergent proteins in both *S. kudriavzevii* and *S. cerevisiae*.

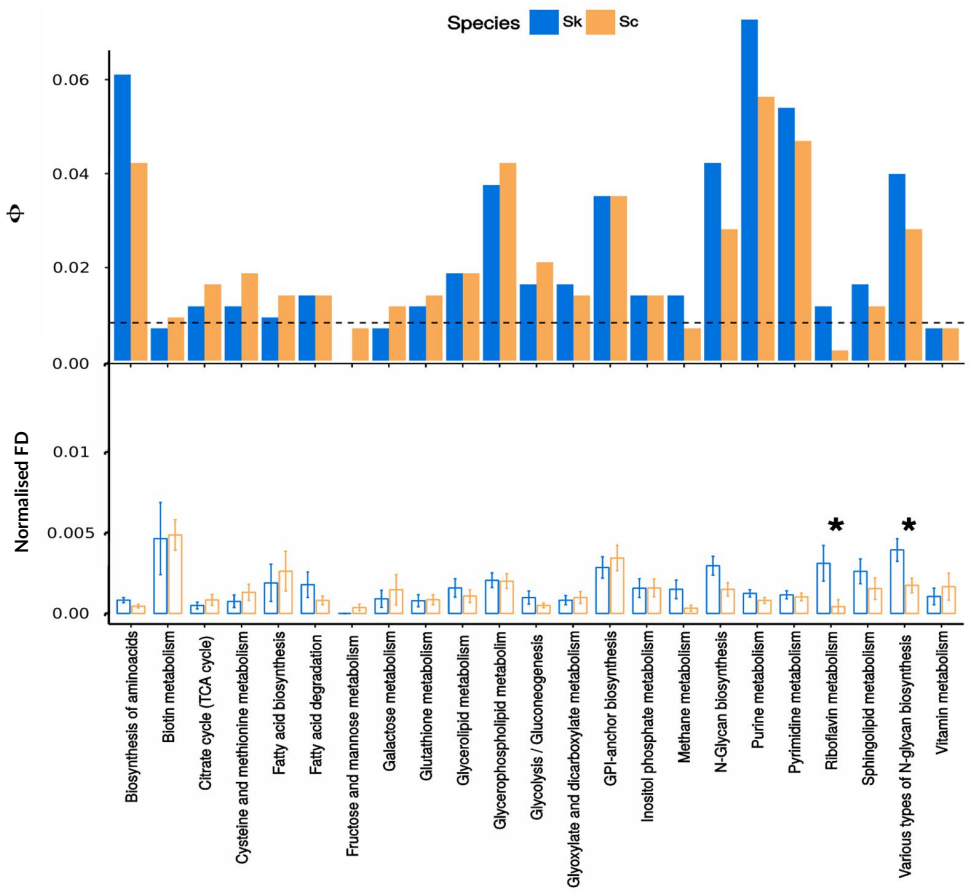


Figure 1.3: Functional divergence among a subset of metabolic pathways. (**Top panel**) Normalised contribution of genes showing evidence of functional divergence to every path. The height of the bars represents Φ, the normalised contribution of each pathway (*i*) of size (t) to the total number of genes under functional divergence when considering the whole dataset (T), calculated as $\Phi = (n_i / t)(t / T)$. Bars above the dashed line represent enriched pathways in genes under functional divergence while bars below the line show impoverished pathways. (**Bottom panel**) Normalised functional divergence values among metabolic pathways. The significance of the differences in every pathway between analysis performed with *S. kudriavzevii* (Sk) or *S. cerevisiae* (Sc) as clade-of-interest was assessed by a Wilcoxon paired signed-rank test, those significant were indicated with an asterisk.

We also assessed the significance of the differences between *S. cerevisiae* and *S. kudriavzevii* in normalised functional divergence values belonging to the different pathways (Figure S1.2). In the *S. kudriavzevii* branch, proteins related to metabolism of riboflavin and biosynthesis of various types of N-glycans showed highly normalised functional divergence values and their difference with the same values calculated for the *S. cerevisiae* branch was found to be significant (Figure 1.3., bottom panel).

Finally, functional divergence was evaluated in pairs of duplicated sequences (WGDs and SSDs) coming from gene duplication events (Table 1.2). There were not significant differences in the ratio between singletons in any of the duplicated sequences. We observed more WGDs than SSDs cases of functional divergence (*F* =

1.08), although this difference was not significant (*p*-value = 0.66).

Table 1.2: Number of genes with a positive result in positive selection, functional divergence and Tajima's relative rate test analyses. PS: number of genes positive for the branch-site analysis; MC: number of genes rejecting the molecular clock hypothesis (Tajima's relative rate test analysis); FD number of genes having functionally divergent codon sites. Results are represented per branch. Genes were classified into singletons or duplicates (WGD or SSD).

Type of gene	PS		MC		FD		Total genes analysed
	Sc	*Sk*	*Sc*	*Sk*	*Sc*	*Sk*	
Singletons	25	26	65	157	1734	1849	3439
SSD	2	1	6	19	151	169	330
WGD	5	3	7	15	221	230	395

1.3.2 Evidence of adaptive evolution in genes related with known physiological differences between the two *Saccharomyces* species

Detecting traces of positive selection could be challenging. As mentioned above, very few genes were obtained with the statistical methods applied for detecting adaptive evolution such as the branch site test and the Tajima's relative rate analysis. Contrastingly, at least half of the proteins analysed for both species showed amino acid positions that could be contributing to functional divergence. In this section, we want to highlight those genes detected in our analysis that could have a role in the phenotypic differences between *S. cerevisiae* and *S. kudriavzevii* according to physiological characterizations performed in our group. Thus, these two *Saccharomyces* species show differences in their carbon metabolism (Arroyo-López et al., 2010a; López-Malo et al., 2013; Oliveira et al., 2014). In our branch-test analysis, we detected adaptive evolution in gene *FBA1*, an essential gene encoding a fructose 1,6-bisphosphate aldolase (Schwelberger et al., 1989). This enzyme has a crucial role in the glycolysis pathway, it catalyses the conversion of a high-energy hexose, fructose 1,6-biphosphate, into two interconvertible phosphorylated trioses, glyceraldehyde-3-phosphate and dihydroxyacetone-phosphate, just at the branching point where these trioses can be directed to the end of glycolysis and ethanol fermentation or to the synthesis of glycerol. This gene also showed acceleration of

evolutionary rate in *S. kudriavzevii* branch according to the Tajima's relative rate test results.

Differences in nitrogen metabolism and aroma synthesis have also been reported (Gamero et al., 2015; Stribny et al., 2015). One of the genes under positive selection in the *S. kudriavzevii* branch is *ARO4*, which encodes a 3-deoxy-D-arabino-heptulosonate-7-phosphate synthase that catalyses the first step in aromatic amino acid biosynthesis (Künzler et al., 1992). Another gene is *DAL3*, which codifies for a ureidoglycolate lyase with a role in the third step of allantoin degradation (Yoo et al., 1985). *DAL3* belongs to the allantoin cluster (Wong and Wolfe, 2005) together with the genes *DAL1, DCG1, DAL2, DAL5, DAL7,* and *DUR1,2*. Although *DAL5* and *DAL7* were not included in the set of 4164 genes analysed because they were missing in some genomes, the rest of genes of the allantoin cluster included in the analyses encode proteins that showed signals of functional divergence (Dur1,2p, Dal1p, Dal2p, Dal3p, and Dcg1p).

The riboflavin pathway was found to be enriched for genes showing accelerated evolutionary rates in *S. kudriavzevii* branch. Riboflavin is required for the synthesis of the cofactors flavin mononucleotide (FMN) and flavin adenine dinucleotide (FAD) (Ghisla and Massey, 1989). Additionally, four genes involved in this pathway, *RIB2, RIB3, RIB5,* and *RIB7*, encoded proteins that showed functional divergence for *S. kudriavzevii*.

1.3.3 Adaptive evolution in genes for which no previous physiological data is available

Our approach also detected several genes for which no experimental data about physiological differences is available, and therefore, can be the subject of future studies. For instance, *FRT2* and *RQC2* were found under positive selection in both

S. kudriavzevii and *S. cerevisiae* branches. *FRT2*, also known as *HPH2*, encodes for a membrane protein of the endoplasmic reticulum (Burri and Lithgow, 2004), which interacts with the protein encoded by its paralogue *FRT1* (*HPH1*), duplicated after the WGD. Although their functions are not well known, both paralogues have been associated to physiological stress response as they can promote growth when there is a high concentration of Na^+ in the environment (Heath et al., 2004). *RQC2* encodes a component of the ribosome quality control (RQC) complex which takes part in the degradation of aberrant nascent proteins (Brandman et al., 2012), and also has a role in the recruitment of alanine-to-threonine- charged tRNAs (Shen et al., 2015). It has also been reported that *RQC2* is responsible for communicating translation stress signal to the heat shock transcription factor *HSF1* (Brandman et al., 2012).

Finally, *ZIP1* not only has been found under positive selection in *S. kudriavzevii* branch, it also showed accelerated evolutionary rates. In addition, amino acid positions contributing to functional divergence and different SCOP domains have been observed. *S. kudriavzevii ZIP1* encodes a Zip1p protein carrying a tropomyosin domain, which it is not present in *S. cerevisiae ZIP1*. This gene encodes a transverse filament protein that conforms the synaptonemal complex and it is required for meiotic chromosome synapsis, acting as a molecular zipper to facilitate the interaction between homologous chromosomes (Sym et al., 1993). This is correlated with the GO enrichment results of functionally divergent proteins on both *S. cerevisiae* and *S. kudriavzevii* clades, which revealed an enrichment in biological processes such as cell cycle and cellular component like cellular bud neck (Table S1.3).

1.4 Discussion

Saccharomyces kudriavzevii is a species from the *Saccharomyces* genus isolated from natural environments such as tree barks and decayed leaves (Boynton and Greig, 2014). On the contrary, *S. cerevisiae* is a species very well known, isolated from a wide range of environments and frequently related to human-driven industrial processes (Goddard and Greig, 2015). Although *S. kudriavzevii* ecological niche is still not well understood, phenotypic differences existing between *S. kudriavzevii* and *S. cerevisiae* have been addressed in previous studies (reviewed in Pérez-Torrado et al. (2018)).

In an attempt to understand the genetic basis behind the main phenotypic differences between *S. kudriavzevii* and *S. cerevisiae* we have proposed to trace the genomic changes occurred as a consequence of the adaptation of these species to the different environments.

Despite this study relied on a small set of *S. kudriavzevii* genomes, we have assessed some general insights into the genetic differences between the well-studied *S. cerevisae* and *S. kudriavzevii*. Understanding the evolutionary process of the adaptation of *S. kudriavzevii* and *S. cerevisiae* requires a pluralistic approach. This way we have applied methods to detect signatures of strong selection in coding sequences combined with differences observed at protein level such as functional divergence and accelerated rates of substitution.

The positive selection analyses revealed three genes related to metabolism that might be good candidates to explain differences between both species: *FBA1, ARO4* and *DAL3*. As mentioned, *FBA1* is involved in the synthesis of dihydroxyacetone phosphate, the precursor of the glycerol synthesis. Previous studies have shown how *S. kudriavzevii* is able to produce higher amounts of glycerol when compared to *S.*

cerevisiae (Arroyo-López et al., 2010a). Here we proposed that the positive selection observed in this gene together with the acceleration in the evolutionary rates, as shown after the performance of the Tajima's relative rate test, are signatures of adaptation and the *S. kudriavzevii* version of the *FBA1* may have an importance in the synthesis of glycerol in the cell (Oliveira et al., 2014). In addition, the reaction catalysed by this enzyme has been proposed as cold-favouring (Paget et al., 2014), so the thermal stability of this enzyme could be also an important factor to take into account to explain the patterns of adaptation observed in this gene (Gonçalves et al., 2011).

ARO4 is involved in aromatic amino acid biosynthesis. Previous works have demonstrated the differences in amino acid metabolism among closely related *Saccharomyces* species and the ability of *S. kudriavzevii* to produce different amounts of aroma compounds such us higher alcohols and acetate esters from amino acidic precursors (Stribny et al., 2015). Therefore, we proposed that the evidence of positive selection observed in this gene could be related to the phenotype already mentioned.

DAL3 is part of the allantoin gene cluster (Wong and Wolfe, 2005), and it is involved in the allantoin degradation pathway. Allantoin has been found in similar environments as those in where *S. kudriavzevii* has been isolated, like tree bark exudates, and it has been demonstrated to have an important effect on the fitness of yeast living in natural environments (Filteau et al., 2017). This nitrogen source, especially when it is limited, has been shown to cause a rapid effect in the yeast genomes due to environmental adaptation (Gresham et al., 2010). The evidence of selection acting on this gene, together with the fact that the whole cluster showed functional divergence, could explain that *S. kudriavzevii* is better adapted to natural environments in which allantoin is more frequently found rather than in human-related environments.

Differences in functional divergence values revealed that proteins belonging to metabolism of riboflavin pathway was significantly different in *S. kudriavzevii* than in

S. cerevisiae. *RIB2, RIB3, RIB5* and *FMN1* were also found to have their evolutionary rates accelerated when compared to *S. cerevisiae*. Positions contributing to protein functional divergence were also found to be related to protein structure stability. A previous systems biology study which used these two species of yeasts because of their differences in temperature growth revealed that genes related to riboflavin were potentially affected by cold temperature because vitamins might have an important role at low temperatures (Paget et al., 2014).

The analysis of functional divergence in *S. kudriavzevii* also revealed a high number of genes involved in cellular response to osmotic and oxidative stress and sphingolipid metabolic pathway. Sphingolipids play very important roles in yeasts, being involved in signal transmission, cell recognition, regulation of endocytosis, ubiquitin-dependent proteolysis, cytoskeletal dynamics, cell cycle, translation, post-translational protein modification, and heat stress response (Cowart and Obeid, 2007).

In this work, we have increased the number of *S. kudriavzevii* genomes, which allowed us to conduct comparative analyses to unveil some of the mechanisms involved in the differential adaptation of *S. cerevisiae* and *S. kudriavzevii*. We used methods making different assumptions just to validate the reliability of our results and their interpretation. The inferred cases of positive selection deserve further research, especially with the experimental testing of functional divergence.

CHAPTER 2

GWideCodeML: a Python package for testing evolutionary hypotheses at the genome-wide level and its application in detecting signatures of positive selection in *Saccharomyces uvarum*

2.1 Introduction

With the rise of the genomic era, comparative analyses are gaining interest and are becoming more feasible due to the increase of available genomes. Testing evolutionary hypotheses to measure selective pressure in coding sequences is a common approach in evolutionary biology projects. To do this, there are different bioinformatic tools and resources such as the PAML package (Yang, 2007). Within

This chapter is published in Macías et al. (2020) *G3: GENES, GENOMES, GENETICS*.

this package, *codeml* allows estimating the ratio (ω) between non-synonymous and synonymous substitutions in protein-coding sequences. The assumption that synonymous mutations accumulate naturally, implies that ω can be used as a measure of the selective pressure on the coding sequence. In a neutral evolution scenario, this ratio will remain equal to one while it will be less or greater than one under purifying and positive selection, respectively. *Codeml* can be run with different assumptions, including the most used: ω is constant along the whole coding sequence but it can vary among branches (branch models), ω remains constant among branches, but it can differ among codon sites in a coding sequence (site models) or assuming that ω varies both among branches and sites (branch-site models). The desired model, along with other parameters, is given to *codeml* through the control file, along with the coding-sequence alignment of orthologous genes and the species tree. Implementing *codeml* in an automatic workflow for a genome-wide approach has its challenges. Among these is the negative correlation between the number of genomes and the number of orthologues shared. Increasing the number of genomes to improve statistical power will reduce the number of analysed genes. However, the lack of one particular gene in any of the analysed species might not be critical for the analysis if the number of remaining orthologues is substantial. In this work, we provide a Python package, GWideCodeML, that can be used for running codeml on a set of orthologous coding sequences under site, branch, or branch-site models. This package automatically generates the files necessary to run the codeml program including pruning of the topology to match each of the alignments in cases where some taxa are missing compared to the species tree. Chapter 1 dataset was used for package testing and benchmarking increasing the number of genes showing positive selection signals while allowing the use of several outgroup species without decreasing the number of genes included in the analysis. We used this package for evolutionary hypotheses testing to find signatures of positive selection in *S. uvarum*. This species is the only non-hybrid species, excluding *S. cerevisiae*,

isolated from human-driven environments such as wine and cider fermentation at low temperatures (Almeida et al., 2014; Rodríguez et al., 2017). *S. uvarum*, together with its sister species *S. eubayanus*, occupies a basal position in the *Saccharomyces* phylogeny, being the most distantly related to *S. cerevisiae* of the genus. It has been isolated from both anthropic environments and natural environments of America, Europe and New Zealand, extensively associated with trees of the Fagales order. A population genomics study revealed three major clades of *S. uvarum*: a clade containing Holarctic strains and some South American strains, a clade containing isolated strains from South America uniquely and, finally, a clade corresponding to the Australasian population, clearly separated from the other two clades with a 4.4% of divergence (Almeida et al., 2014). Our package, GWideCodeML, was used to study the action of natural selection in *S. uvarum* species to identify genes under positive selection with potential roles in adaptation. We applied both branch and branch-site approaches in the branch leading to the domesticated *S. uvarum* strains. Finally, we used the site model to study the action of selection at different amino acid positions, and we calculated the proportion of paralogues under positive selection.

2.2 GWideCodeML development

2.2.1 Input files

GWideCodeML requires as input, a directory with codon-aligned orthologous sequences in fasta format, and a Newick Standard tree topology containing all the species included in the analysis (Figure 2.1). A common denominator between the names within the fasta file and the taxa (e.g. species and/or strain names) within the tree, is also required. It should be noted that the pipeline itself does not handle possible duplication events and the user should make sure that the fasta files contain a maximum of one sequence per species/strain. The package contains a module that can remove duplicated genes from the alignment, by only keeping the best BLAST hit against a set reference species. Besides, the users need to provide optional configuration parameters such as the model to be tested. Furthermore, it is possible to set a minimum number of species/strains, either belonging to the foreground branch and/or to the outgroups, to filter out alignments with low statistical power. Whether the users decide to test a branch, branch-site model, or choose to set a threshold, a text file with a list of branch labels information is also required as input. In the branch and branch-side model, a 0 or 1 should be added after the branch label to indicate if the branch is foreground or background, respectively. GWideCodeML package allows for multiple branch testing, which means that multiple branches can be tested as foreground branches when numbers higher than one are added after the branch labels. In these cases, the workflow will be run as many times as set in the branch labels files.

2.2.2 GWideCodeML workflow

Codon-aligned sequences are passed through the workflow if: i) only one gene per taxa is found and ii) they pass the threshold of the number of sequences, which by default is set to zero. Hereafter, sequence names are compared between the alignment file and the species tree. The tree will be pruned until it only contains the taxa present within the alignment. From this step onwards, the workflow performance will depend on the user's model (branch, site, branch-site, or custom). In the branch or branch-site models, the new tree created (pruned or original) will be used to indicate the foreground branch. Hereafter, the control files necessary to run both the alternative and null hypotheses are created, and *codeml* (included in the PAML v4.9 package) is run with each of them. This is the most time-consuming step; however, the pipeline allows parallelisation of this task when the user provides a maximum number of cores to be used by the program set by the -p optional parameter. Once *codeml* has finished, GWideCodeML parses the output files to perform a Likelihood Ratio Test (LRT) to assess the level of significance between both hypotheses. This last part is optional, allowing the users to analyse *codeml* results by themselves or letting the pipeline do it for them. When the branch or branch-site models are selected, GWideCodeML will also run FastTree (Price et al., 2010) on each of the alignments and compare the gene and species tree topologies focusing on the foreground branches. If the studied clade is not monophyletic between the two trees, the user will be informed.

2.2.3 Nested models implemented in the package

Three nested models have been implemented in GWideCodeML for their execution. The branch model allows the dN/dS ratio to vary among branches assuming this ratio remains constant among codons (Yang, 1998; Yang and Nielsen, 1998). This model is useful for detecting positive selection acting on a particular branch. It uses

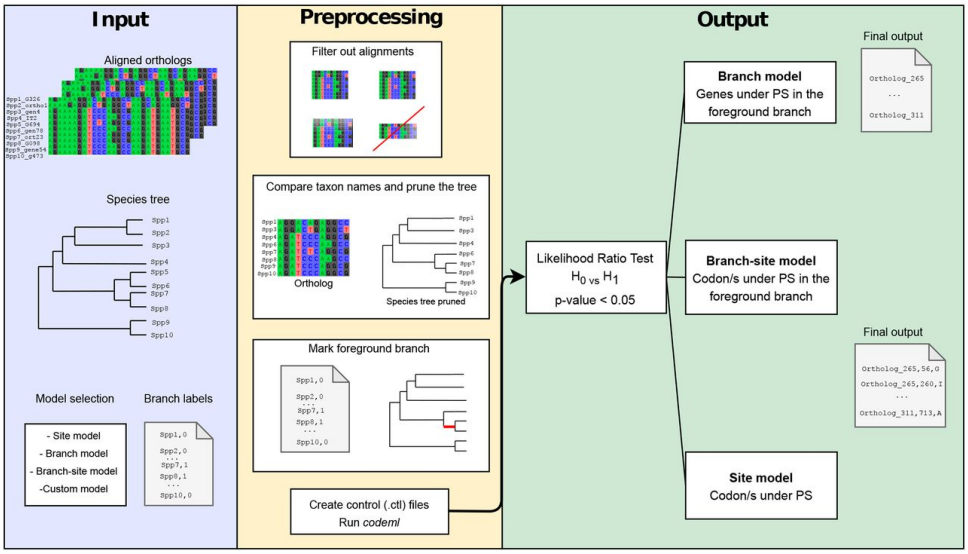


Figure 2.1: GWideCodeML workflow. Schematization of the GwideCodeML pipeline divided in the three main parts: required input, preprocessing steps necessary for creating control *codeml* files necessary for testing both null and alternative hypotheses. The last part of the pipeline shows the output file obtained after running LRTs on the *codeml* results file. This output file contains all the candidate genes of being under positive selection.

M0 (one-ratio model) and two-ratios model (Goldman and Yang, 1994) as null and alternative hypothesis testing respectively. Site models assume the dN/dS ratio might differ among codons of the coding sequence provided (Yang and Nielsen, 1998; Yang et al., 2005). To test this, models M1a or NearlyNeutral and M2a or PositiveSelection (Yang and Nielsen, 1998; Yang et al., 2005) are built-in on the workflow as null and alternative hypotheses, respectively. Finally, the branch-site model combines the two previous approaches as it allows the dN/dS ratio to vary among branches and codon sites. Model A null and model A have been used in our package to test both null and alternative hypotheses (Yang et al., 2005; Zhang et al., 2005). The three described models are nested, and therefore an LRT can be applied to assess the null hypothesis's level of significance. If the *p*-value obtained from the LRT is lower than 0.05, the null hypothesis is rejected, and this gene is further investigated. Additionally, it is possible to vary the significance level by applying Bonferroni's test correction (Miller (1981), p. 67–70) when multiple branches are tested. This method has been proven to be practical and useable (Anisimova and Yang, 2007) to correct the significance level for null hypothesis rejection depending on the number of hypotheses tested. Rejecting the null hypothesis means that the alternative hypothesis is more likely than null for the

given species topology and alignment. In addition to the LRT, it is necessary to check whether dN/dS ratio is greater than one in the foreground branch (branch or branch-site models) to determine if a gene is a candidate gene of being under positive selection. Moreover, in the cases of site and branch-site models, *codeml* offers two methods for calculating posterior probabilities for site classes to identify codons under positive selection when the LRT is significant. These methods are naïve empirical Bayes (NEB) and Bayes empirical Bayes (BEB) (Yang et al., 2005). PAML authors highly recommend ignoring NEB output and using BEB instead so the package extracts codon positions with a probability greater than 0.9 of being under positive selection according to BEB results when the LRT is significant.

2.2.4 Additional options and modules

GWideCodeml contains several additional scripts and options which can aid the user in preparing the data or get further information about the data. They include: i) a module that helps users create the alignment or alignments necessary to reconstruct a robust species phylogeny by selecting the genes containing sequences for all species included in the analysis. ii) a module that allows users to align their codon alignments with three different aligners (Mafft, Muscle or Prank). iii) The workflow proceeds with the provided species phylogeny; however, it is possible to set the parameter -gene_tree. The program will check if any of the genes within the studied clades originate from a horizontal gene transfer by comparing the species tree with the gene tree. iv) In the branch and branch-site models, the option -dnds will parse the foreground, and background omega's into a separate output file. v) a module that performs a Bonferroni's test correction when multiple branch testing is performed.

2.2.5 Output files

The package's final output is a text file with a list containing genes under positive selection depending on the analysis performed. Additionally, when the site or branch-site model is selected, gene names are accompanied by the codon positions in the alignment under positive selection. A gene may have one or more codon positions under positive selection.

2.2.6 GWideCodeML testing

The dataset of *S. cerevisiae* and *S. kudriavzevii* analysed in Chapter 1 was used for package testing and benchmarking. In this previous analysis, the dataset was conformed of nine genomes: four annotated genomes of each *Saccharomyces* species and *T. delbrueckii* as the outgroup species. We increased this dataset for testing this package, particularly the outgroup, with 18 annotated yeast genomes from The Yeast Gene Order Browser (Byrne and Wolfe, 2005). This database also contains information on the homology of different species of the Saccharomycotina subphylum. In the analysis of selection performed in Chapter 1 with nine genomes, the number of genes analysed was limited by the number of common orthologous genes shared among them: 4165 genes. In the tested dataset here, which included 26 yeast genomes, only 2753 genes were shared among all species. By setting a minimum number of three background species necessary for testing the evolutionary hypotheses on a gene and four strains in the foreground branch, our pipeline almost doubled the genes analysed to 4920 genes.

2.2.7 GWideCodeML compared to other software

Several bioinformatics resources have been developed to simplify and/or automate positive selection analyses. Some of these tools are aimed at facilitating the running of *codeml* and visualize the results in one task, meaning that the users need to run the program gene by gene (Delport et al., 2010; Steinway et al., 2010; Stern et al., 2007; Xu and Yang, 2013). There are also other bioinformatics resources created for running *codeml* in a genome-wide framework like GWideCodeML. Most of these tools have overlapping features among them and when compared to the package presented here (Table 2.1). The main benefit of our package, when compared to others, is the combination of features that allows testing site, branch, and branch-site models on filtered orthologues with a variable number of sequences among them. PosiGene (Sahm et al., 2017) is the only one of the three that can be run with a variable number of taxa in the orthologues, this is done by generating a gene tree for each run or, as GWideCodeML, prunes the provided species tree to fit each of the alignments. When looking for positive selection it is important to have a correct topology, however, both species and gene trees suffer their problems. In the case of gene trees, an incorrect topology can be generated due to long-branch attraction and natural variation between genes due to the stochastic nature of mutations (Castresana, 2007; Jeffroy et al., 2006; Rokas and Carroll, 2006). On the other hand, occurrences of horizontal gene transfer will result in a wrong species topology. GWideCodeML uses the species tree but the package will flag the genes in which the conservation of the taxa in the studied clade is not the same in the species and the gene topologies. Furthermore, PosiGene, along with POTION (Hongo et al., 2015), can perform some of the pre-processing steps, although they have been developed to run only one specific model. In contrast, LMAP (Maldonado et al., 2016) is the most flexible regarding the models offered, the three different nested models along with custom's settings, although it lacks the pre-processing steps necessary to run them on a dataset composed of orthologs with

a heterogeneous number of sequences. Other features offered by GWideCodeML, which are not included in the other three programs, are the option to test multiple branches in the same run, and the fact that provides a workflow to run multiple hypothesis testing, along with a module to analyse the output.

Table 2.1: Overlapping features between GWideCodeML and other bioinformatics tools. *1Built-in models: site model (SM), branch model (BM), branch-site model (BSM). *2PosiGene generates a new tree for each gene, where GWideCodeML prunes the provided species tree.

Feature	LMAP	POTION	PosiGene	GWideCodeML
Built-in models*1	SM, BM, BSM	SM	BSM	SM, BM, BSM
Run costume models	Yes	-	-	Yes
Easy branch labelling	Yes	-	Yes	Yes
Automatic pruning*2	-	-	Yes	Yes
Filter out low quality orthologues	-	Yes	Yes	Yes
Multithreading	Yes	Yes	Yes	Yes

2.2.8 GWideCodeML availability

The GWideCodeML package is available at GitHub (`https://github.com/lauguma/GWideCodeML`), and the GWideCodeML user guide can be found at `https://github.com/lauguma/GWideCodeML/wiki`.

2.3 Case study: GWideCodeML applied to detect signatures of positive selection in *S. uvarum*.

2.3.1 Genome dataset

Twenty-four yeasts genomes were selected for testing evolutionary hypotheses (Table 2.2). Annotation of *Saccharomyces* was carried out as described in 1.2.1. section. Genes were annotated with systematic gene names. We used the information available at the Yeast Genome Order Browser (YGOB) database (Byrne and Wolfe, 2005) to identify orthology relationships between *Saccharomyces* and non-*Saccharomyces* species. An in-house Python script was used to separate orthologous groups into different fasta files. Each gene from species that diverged before the WGD event (*L. kluyveri, K. lactis, Z. rouxii* and *T.delbrueckii*) the same gene was present twice at different fastas in case they belong to a pair of duplicated genes (ohnologues). Genes were translated into amino acid sequences, aligned with Mafft v.221 (Katoh and Standley, 2013), and back-translated into aligned codons using the Python script *fas2msa*, available in the GWideCodeML package.

2.3.2 Species tree phylogeny

A concatenated alignment was obtained from 3096 orthologous genes using the *cds2concat* script, also available in the GWideCodeML package. A maximum likelihood phylogeny was reconstructed based on the concatenated alignment by using RAxML v8.1.24 (Stamatakis, 2014), with 100 bootstrap replicates and the GTR model. The ML tree (Figure 2.2) was visualized using iTOL (Letunic and Bork, 2016). This unrooted ML tree, in the Newick standard format, was used as the species tree for subsequent analyses.

Table 2.2: List of strains and sources of the genomic sequences used in Chapter 2.

Species	Strain	Isolation source	Reference
L. kluyveri	CBS3082	*Drosophila pinicola*	(Souciet et al., 2009)
K. lactis	CLIB210	Laboratory	(Dujon et al., 2004)
Z. rouxii	CBS732	Black-grape must	(Souciet et al., 2009)
T. delbrueckii	CBS1146	Unknown	(Gordon et al., 2011)
T. blattae	CBS6284	*Blatta orientalis* (gut)	(Gordon et al., 2011)
N. castellii	CBS4309	Soil	(Gordon et al., 2011)
K. africana	CBS2517	Soil	(Gordon et al., 2011)
S. eubayanus	CBS12357	Fruiting body of *Cyttaria hariotii*	(Brickwedde et al., 2018)
S. uvarum	ZP962	*Nothofagus cunninghamii*	(Almeida et al., 2014)
S. uvarum	NPCC1290	*Araucaria araucana*	(Lairón Peris et al., n.d.)
S. uvarum	NPCC1314	Chicha	(Lairón Peris et al., n.d.)
S. uvarum	ZIM2113	Must	Chapter 3
S. uvarum	CBS7001	*Mesophylax adopersus* (gut)	(Scannell et al., 2011)
S. uvarum	BMV58	Wine	Chapter 3
S. arboricola	CBS10644	*Quercus fabri*	(Liti et al., 2013)
S. kudriavzevii	IFO1802	Decayed leaf	(Scannell et al., 2011)
S. kudriavzevii	CR85	Quercus ilex	(Macías et al., 2019)
S. mikatae	IFO1815	Soil	(Scannell et al., 2011)
S. jurei	NCYC3947	*Quercus robur*	(Naseeb et al. 2018)
S. paradoxus	N44	*Quercus mongolica*	(Yue et al. 2017)
S. cerevisiae	S288c	Laboratory	(Goffeau et al., 1996)
S. cerevisiae	T73	Wine	(Morard et al., 2019)
S. cerevisiae	YPS128	*Quercus alba*	(Yue et al., 2017)
S. cerevisiae	Y12	Sake	(Yue et al., 2017)

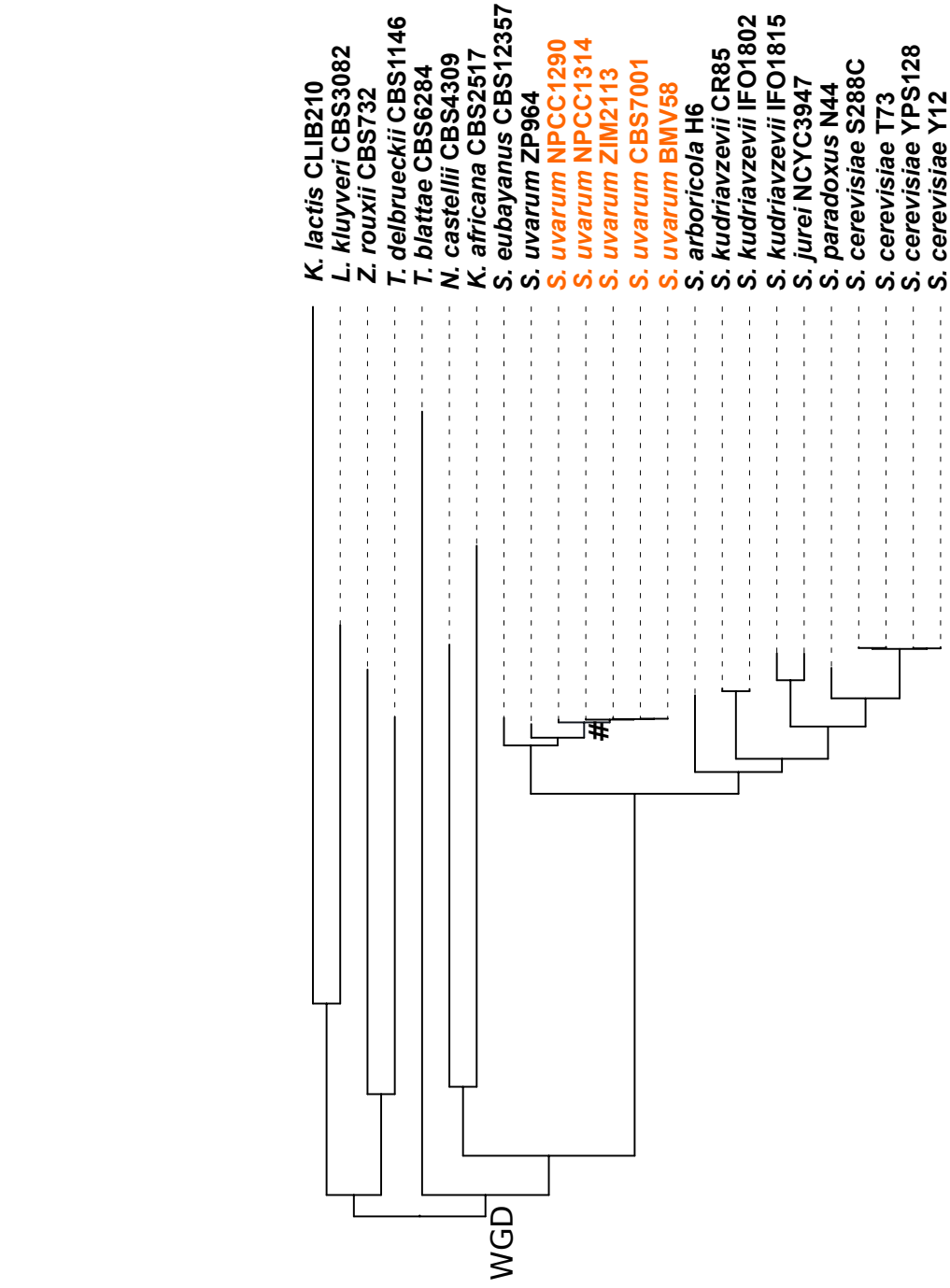


Figure 2.2: Species tree. Species tree used for running GwideCodeML in order to detect signatures of positive selection in *S. uvarum*. Tree is composed of 24 yeasts and the foreground branch is pointed out with a #. *S. uvarum* strain names belonging to the foreground branch are shown in orange. Tree was rooted for tree visualization in iTOL at the branch of the species diverged before the WGD event.

2.3.3 Evolutionary hypothesis testing using GwideCodeML

The yeast coding sequences dataset was used to find positive selection signatures using the three nested models implemented in GWideCodeML: branch, site, and branch-site. Filters were applied using the -omin and -cmin options. Genes were filtered by a minimum four outgroup sequences (-omin 4) and a minimum of three clade-of-interest (*S. uvarum*) sequences (-cmin 3). The foreground branch (clade-of-interest) was defined in the branch grouping the *S. uvarum* strains NPCC1290, NPCC1314, ZIM2113, CBS7001, and BMV58.

2.3.4 Duplicates

Genes were classified as ohnologues generated by the WGD event (WGDs), paralogues generated by small-scale duplications (SSDs), and singletons. Paralogues classification was done as described in Section 1.2.6.

2.3.5 Gene Ontology and Pathway Enrichment Analyses

GO-term and Pathway enrichments were performed as described in Section 1.2.7. The set of genes analysed after passing the filter of a minimum number of clade and outgroup species were used as the background gene set.

2.4 Results

2.4.1 GWideCodeML increases both the number of analysed genes and the statistical power of the analysis

This package not only allowed us to analyse a higher number of genes, but has also dramatically increased the statistical power of our study by including more species. The benchmarking study between the branch-site analysis performed in Chapter 1, and the GWideCodeML performance with the inclusion of more genomes in the dataset, revealed an increase in the number of positive results (Table 2.3). More specifically, we obtained 30 and 32 genes under positive selection in *S. kudriavzevii* and *S. cerevisiae* branches, respectively, in Chapter 1. However, using GWideCodeML under the branch-site model assumption, we obtained 137 and 96 in *S. kudriavzevii* and *S. cerevisiae* branches, respectively, as gene candidates to be under positive selection (Table S2.1). Increasing the number of outgroup species also affects the power to detect sites under positive selection (Goodswen et al., 2018). Besides, our package facilitates the running of additional models such as the branch and site models. The combination of the results of the three approaches tested provides more depth on how positive selection has been acting on the studied clade.

Table 2.3: GWideCodeML testing results. Number of detected genes under positive selection after running GwideCodeML twice, one for each branch, using the three built-in nested models. [*1]In site models, there is no dN/dS ratio variation among branches, therefore, it was run once. Sk: *S. kudriavzevii*; Sc: *S. cerevisiae*.

Model	Nested models (null vs alternative hypotheses)	No. genes under positive selection in Sk branch	No. genes under positive selection in Sc branch
Branch	M0 vs two-ratios	83	31
Branch-site	MA_{null} vs MA	137	96
Site	M1a vs M2a	32^{-1}	32^{-1}

2.4.2 Cell-wall and chemical homeostasis related genes showed signatures of positive selection in the *S. uvarum* clade

The branch-site model was tested on the dataset of 4495 genes using *S. uvarum* as foreground branch (excluding the more divergent Australasian *S. uvarum* ZP962 strain). We obtained 89 gene candidates having at least one codon position under positive selection (Table S2.2). No significant GO-term nor pathway enrichment was found. Further investigation on the functional annotation of the 89 genes revealed that 12 of them are related to cell-wall organization: *RCR1, GIP1, FIG2, CRH1, HSP150, DAN1, MID2, SLA2 KRE1, GAS5, TIR4* and *TIR2*. Other ten genes are related to chemical homeostasis: *FIT1, PHM8, HAL5, TARK2, MCP2, PPZ1, YCK2, BOR1, IZH4* and *FRE3*. Additional genes of interest under positive selection were the *ADH4*, encoding an alcohol dehydrogenase isoenzyme unrelated to any other yeast ADHs (De Smidt et al., 2008), *ARO80* (transcriptional activator of the aromatic catabolic genes), *GAP1* (general amino acid permease), *HXT6* (high-affinity glucose transporter), and *MNN2* (alpha-1,2-mannosyltransferase). Branch model testing on the same genes resulted in 16 gene candidates being under positive selection without GO-term or pathway enrichment. The intersection of both the branch and branch-site models results reported one common gene, *CUE4*, a gene coding for a protein of unknown function with a ubiquitin-binding domain, probably acting to facilitate the intramolecular monoubiquitination (Shih et al., 2003).

2.4.3 Ribosome and glucose fermentation genes have multiple codon positions under positive selection

The site-model was tested on our dataset, reporting 49 genes with codons under positive selection (Table S2.2). GO-term enrichment analysis showed enrichment in the biological processes of translation and ribosome assembly, and the molecular

functions of the structural constituent of the ribosome and cell-wall. Ribosome and cell-wall cellular components GO-terms were also found enriched in our site-model results. A pathway enrichment analysis was also obtained for the glucose fermentation pathway with five genes having codon positions under positive selection: *CDC19, ENO2, FBA1, GPM1* and *PDC1* (Figure 2.3). A summary of the number of positions positively selected in each gene showed that most of them have multiple sites, with the gene *SRP40*, that encodes a nucleolar serine-rich protein, showing the highest number of codons under positive selection (35). *FET5*, a multicopper oxidase, and *MDJ1*, a co-chaperone involved in protein folding in the mitochondrial matrix, also exhibited a high number of codons under positive selection, with 21 positions each.

2.4.4 Enrichment of ohnologues under positive selection

We classified genes into three categories: singletons, WGDs and SSDs to study whether they show different patterns of positive selection (Table 2.4). Under the branch-site model, the proportion of genes under positive selection was 1.6%, 1.7% and 3.6% for singletons, SSDs, and ohnologues, respectively. This high fraction of ohnologues under positive selection compared to the rest was significant (Fisher's exact test: F = 2.35, *p*-value = 0.0004). Differences between duplicates and singletons under the branch and site models were not significant. In most cases, there was only one duplicate of the pair under positive selection, but, there were particular cases in which both copies were found under positive selection. For example, the pair of SSDs, *THI20-THI21*, involved in the thiamine biosynthesis pathway, were found to be under positive selection according to the branch-site model testing. Another example was found under the site-model assumption. Both copies of the WGD ohnologous gene pairs, *RPS27A-RPS27B*, encoding components of the small ribosomal subunit, and *TIR2-TIR3*, codifying for cell-wall mannoproteins, were detected under positive selection. Interestingly, two more genes of the same family of mannoproteins, *TIR1*

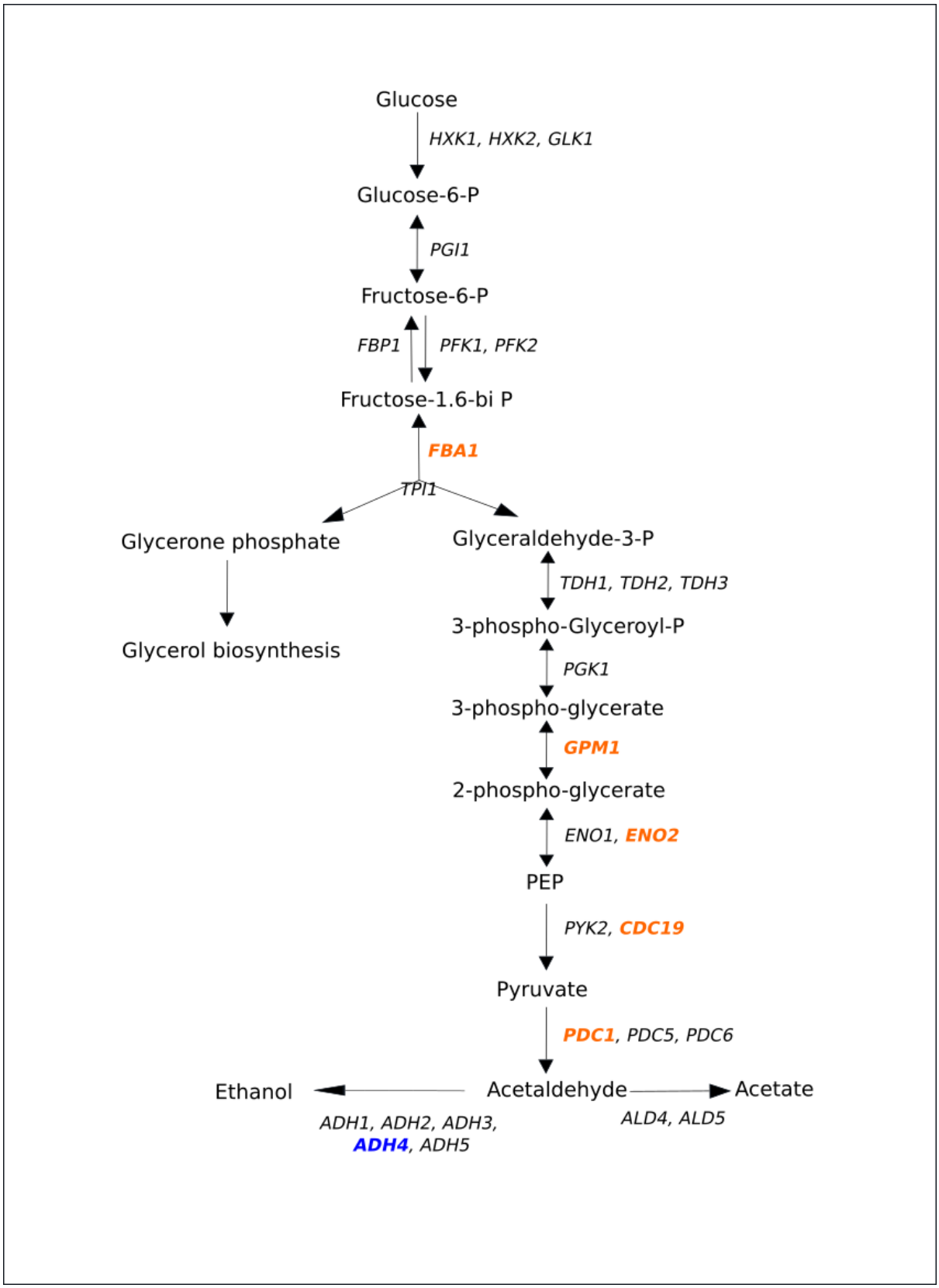


Figure 2.3: Glucose fermentation genes under positive selection. Gene names codifying for the different enzymes names are color-coded and represented next to the arrows. Orange: genes showing signatures of positive selection under the site model. Blue: gene showing signatures of positive selection under the branch-site model

and *TIR4*, also posses amino acid positions under positive selection.

Table 2.4: Number of genes under positive selection according to the model used for hypothesis testing. Genes are classified in singletons and duplicates by WGD or SSD.

Type of gene	Branch model	Branch-site model	Site model	Total no. genes analysed
Singleton	14	47	33	2943
SSD	0	12	2	727
WGD	2	30	14	825

2.5 Discussion

In this chapter, we successfully implemented a Python package to facilitate the use of codeml in genome-wide studies. This package provided us with an excellent opportunity to perform evolutionary testing with higher statistical power by adding more outgroup species without decreasing the number of genes involved in our study. Besides, the use of the site-model analysis on our dataset gave us some curious results about the evolutionary dynamics of the Saccharomycotina species yeasts. Given the good results of the benchmarking with GWideCodeML, we decided to use this package and the three implemented nested-models to detect signatures of positive selection in the *S. uvarum* clade that comprises the domesticated *S. uvarum*. The branch-site model analysis generated exciting results on the possible gene candidates responsible for the *S. uvarum* adaptation. Although no enrichment was detected, we found genes under positive selection that could play a role in the environmental adaptation like the transporters *GAP1* and *HXT6*. *GAP1*, the general amino-acid permease, is a transmembrane transporter responsible for the uptake of amino acids; it is regulated by the nitrogen source (Grenson et al., 1970). *HXT6* gene encodes a high-affinity glucose transporter commonly found mutated or amplified in experimentally evolved *S. cerevisiae* strains grown under limiting glucose concentration (Selmecki et al., 2015). Transmembrane transporter proteins have a crucial role in the adaptation of yeasts to novel environments as the main responsible proteins for nutrient uptake. Therefore, these results suggest an important role of adaptive selection favouring particular amino acid changes fixed in the *S. uvarum* clade to improve nutrient uptake and cell fitness.

The site-model was also tested, revealing enrichment of both pathway and GO-terms. The glucose fermentation pathway was found enriched with five genes with codons under positive selection (Figure 2.3). Three of them correspond to duplicates. It has been suggested that one important outcome of the WGD event, that occurred

in an ancestor of the *Saccharomyces* genus, was the increase of the glycolytic flux as a consequence of the duplication of the genes encoding hexose transporters and glycolytic enzymes (Conant and Wolfe, 2007). The identification of codons under positive selection in ohnologous genes encoding glycolytic enzymes supports this hypothesis. The enzymes of the glucose fermentation pathway showing particular positions on their amino-acid sequences under positive selection might be the main responsible for the observed increase in the glycolytic flux in the species diverging after the WGD. *FBA1* deserves special attention among all of them because it has been identified as a gene under positive selection with different selection testing approaches (Chapter 1). The enzyme codified by this gene, the fructose 1,6-bisphosphate aldolase, has a crucial role during the glycolysis, acting at the branching point where trioses are directed to the synthesis of pyruvate (complete glycolysis) or diverted to the production of glycerol. This gene could be under selection for the fine-tuning of the regulation of this glycolytic crossroad. Further research on how amino acid changes detected under positive selection are responsible for the observed phenotypic changes between *Saccharomyces* species in the glucose fermentation is required.

The family of genes encoding cell-wall mannoproteins has probably an outstanding role in the adaptive mechanisms of *S. uvarum* in particular and the Saccharomycotina species in general. The whole family of genes (*TIR1, TIR2, TIR3* and *TIR4*) were detected under positive selection under the site-model assumption. Additionally, *TIR2* and *TIR4* were found under positive selection in the *S. uvarum* species under the branch-site model. These genes encode proteins that are critical for the growth at suboptimal temperatures as they are induced under cold stress (Abe, 2007). A genome-wide study comparing gene divergences between the sympatric species *S. uvarum* and *S. cerevisiae* observed that these genes were among those showing the highest divergences between both species (Gonçalves et al., 2011). Considering that optimum growth temperature is the most prominent phenotypic difference between *Saccharomyces* species, it is probably that these genes have gone

through positive selection due to their role in temperature adaptation. Altogether, our results provide exciting clues on possible gene candidates with crucial roles in *S. uvarum* adaptive mechanisms. However, further research will be necessary to demonstrate the adaptive role of the fixed amino acid changes in the proteins detected under positive selection.

CHAPTER 3

Convergent adaptation of *Saccharomyces uvarum* to sulphite, an antimicrobial preservative widely-used in human-driven fermentations

3.1 Introduction

Organisms belonging to different lineages can evolve independently to overcome similar environmental pressures through different molecular mechanisms. This convergent evolution has been seen as evidence of the action of natural selection (Losos, 2011). In the last years, comparative genomics studies have suggested that convergent adaptations can be more frequent than expected (Martin and Orgogozo, 2013; Stern, 2013). For example, independent mutations in the gene responsible

for the vernalization have been reported between different populations of *Arabidopsis thaliana* (Shindo et al., 2005). Convergent adaptation has been observed between different species of the genus *Drosophila* with different mutations leading to the same outcome, the loss of trichomes (Sucena and Stern, 2000). Also, species of insects classified into different orders have convergently evolved to increase their tolerance to toxic compounds produced by plants (Dobler et al., 2012). These examples demonstrate that convergent adaptation has frequently occurred in nature between organisms belonging to different taxonomic levels. Evidence of convergent adaptation has also been reported in experimentally evolved populations, for example in populations of *Saccharomyces cerevisiae* evolved under glucose limitation (Kvitek and Sherlock, 2011).

The genus *Saccharomyces* is composed of eight species including the model organism *S. cerevisiae* (Dujon and Louis, 2017). There is a substantial nucleotide divergence displayed for example between *S. cerevisiae* and the species *S. uvarum* and *S. eubayanus*, comparable to the divergence found between humans and birds (Dujon, 2006). *S. cerevisiae* has traditionally been associated with food and beverage fermentations which have been traced back to 5,000 – 10,000 years ago (Cavalieri et al., 2003; McGovern et al., 2004). This domestication of *S. cerevisiae* by humans has left footprints that characterize their genome (Gallone et al., 2018; Legras et al., 2018; Peter et al., 2018). Along with *S. cerevisiae*, the species *S. uvarum* is the only natural species of the *Saccharomyces* genus that shows ecological success in human-driven fermentative environments (Fernandez-Espinar et al., 2003; Rainieri et al., 1999). *S. uvarum* coexists and even replaces *S. cerevisiae* in wine and cider fermentations performed at low temperatures, in particular at regions with oceanic or continental climate (González Flores et al., 2019; Naumov et al., 2002; Rodríguez et al., 2017). Genomic footprints of domestication, like introgressions, have also been reported in *S. uvarum* genomes (Almeida et al., 2014).

During fermentation processes, yeast cells face adverse conditions such as osmotic stress due to high sugar concentrations, high ethanol, low temperatures, low pH, and the presence of certain toxic compounds used as preservatives. One of the most common preservatives used in wine and cider fermentations is sulphite (Bauer and Pretorius, 2000). The most common molecular mechanism to deal with the presence of sulphite in the media in yeasts involves the sulphite efflux with a plasma membrane pump encoded by the gene *SSU1* (Casalone et al., 1992; Park and Bakalinsky, 2000). The strains lacking this gene showed a higher sensibility to sulphite due to the intracellular accumulation of this compound (Avram and Bakalinsky, 1997; Nadai et al., 2016). The transcription factor encoded by the *FZF1* gene has been reported to interact with the upstream promoter region of the gene *SSU1* to increase its transcription (Avram et al., 1999).

Mutations causing large-scale chromosomal rearrangements often occur in yeast populations rather than less frequent small-scale changes (Lynch et al., 2008). Even though most large-scale changes are deleterious and, therefore, quickly removed from the population, these mutations contribute to the genetic variation within the population facilitating the rapid adaptation to novel environments (Chang et al., 2013; Selmecki et al., 2009). It has been reported that specific chromosomal rearrangements in *S. cerevisiae* wine strains generate an overexpression of the *SSU1* gene that increases the tolerance to sulphite (Pérez-Ortín et al., 2002). A reciprocal translocation between chromosomes VIII and XVI replaced the promoter of the *SSU1* gene, encoding a sulphite transporter (Pérez-Ortín et al., 2002). This modification causes an increased expression of *SSU1* and, as a consequence, a greater resistance to sulphite. After this first evidence, several groups have confirmed both the presence of this rearrangement in different strains belonging to the *S. cerevisiae* wine yeast subpopulation and the advantage that sulphite resistance confers to yeasts during their competition in wine fermentation (Brion et al., 2013; Yuasa et al., 2004). The translocation VIIItXVI has been proposed not only to contribute to the ecological differentiation of wine yeasts but

also to the partial reproductive isolation between wine and wild subpopulations of *S. cerevisiae* (Clowers et al., 2015; Hou et al., 2014). Years later, another translocation event, between chromosomes XV and XVI, was described and associated with an increase in the expression of the *SSU1* gene in *S. cerevisiae* yeasts (Zimmer et al., 2014). Another molecular mechanism causing the overexpression of this gene found in *S. cerevisiae* is an inversion in chromosome XVI (García-Ríos et al., 2019).

The promoter region of the *SSU1* gene has been demonstrated to be a hotspot of evolution in *S. cerevisiae* leading to different chromosomal rearrangements with a common phenotypic outcome: an increased sulphite tolerance. This work aims to test the evidence of convergent evolution at a higher taxonomic level by using another *Saccharomyces* species isolated from human-driven environments, *S. uvarum*. In this study, several strains of *S. uvarum* isolated from a wide range of environments and geographic locations have been used to identify high sulphite tolerant strains and the underlying molecular mechanisms associated with this trait.

3.2 Materials and Methods

3.2.1 Yeast strains, media, and fermentations.

Strains were maintained and propagated in YPD media (5 g/L yeast extract, 5 g/L peptone, 20 g/L glucose). Wine fermentations were carried out in 100 mL bottles filled with 90 ml of synthetic must (100 g/L glucose, 100 g/L fructose, 6 g/L citric acid, 6 g/L malic acid, mineral salts, vitamins, anaerobic growth factors, 300 mg/L assimilable nitrogen) that simulates standard grape juice (Bely et al., 2003). Fermentations were inoculated at 5.0 × 106 cells/ml density from overnight precultures determined by measuring OD600. Bottles were closed with Muller valve caps and incubated at 25 ºC with gentle agitation. Fermentation progress was followed by daily measuring bottle weight loss. In the fermentations with MBS, after preliminary tests, a sub-lethal concentration (15 mg/l) of MBS that allow the four strains used (BMV58, CECT12600, NPCC1290, and NPCC1314) to grow was selected. All wine fermentations were performed at least in independent triplicates.

3.2.2 Edited strains construction.

To modify *SSU1* promoters in the CBS7001 strain we used the CRISPR-Cas9 technique as described by Generoso et al. (2016). The primers used are listed in Table S3.3. The plasmid pRCCN (Addgene) was used to target the *SSU1* promoter to integrate the recombinant fragments, amplified from BMV58 or BR6-2 strains. Transformations were performed following Gietz and Schiestl (2007) method. Transformants were selected in ClonNat (Sigma) YPD agar plates and verified by PCR and sanger sequencing. Finally, the positive strains were cured of the pRCCN vector.

3.2.3 Genome sequencing, assembly, and annotation.

Strains were sequenced by Illumina HiSeq 2000 with paired-end reads of 100 bp long at the Genomics section from the Central Service of Experimental Research Support (SCSIE), University of Valencia. SPAdes (Bankevich et al., 2012), with default parameters, was used for *de novo* assembly. MUMmer (Kurtz et al., 2004) was used to get the homology between the strains sequenced in this study and the reference *S. uvarum* strain CBS7001 (Scannell et al., 2011). This information was used to get scaffolds into chromosome structure (note that, in Scannell et al. (2011) annotation, chromosome X was mislabeled as chromosome XII and vice-versa). Annotation was performed as described in the 1.2.1. section.

3.2.4 Phylogenetic analyses.

Annotated genomes sequenced in this study as well as collected data from previous studies (Almeida et al., 2014; Scannell et al., 2011) were used for phylogeny reconstruction. Introgressed genes from other *Saccharomyces* species were removed from the analysis. A total number of 1265 orthologous genes were found among the 21 *S. uvarum* strains. Nucleotide sequences were translated into amino-acids and aligned with Mafft (Katoh and Standley, 2013). Aligned protein sequences were back-translated into codons. Maximum-Likelihood (ML) phylogeny reconstruction was performed for each gene using RAxML (Stamatakis, 2014) with the GTRCAT model and 100 bootstrap replicates. ML-trees were concatenated to infer a coalescence-based phylogeny using ASTRAL-III, version 5.6.3 (Zhang et al., 2018). Tree was visualized using iTOL (Letunic and Bork, 2016).

3.2.5 Analyses of the origin of the shared chromosomal rearrangement among BMV58, CECT12600, and NPCC1417 strains.

Gene sequences upstream and downstream of the *SSU1* gene were extracted to calculate genetic distances among the strains BMV58, CECT12600, and NPCC1417. Distances were calculated using the "dist.dna" function from the ape R package (Paradis and Schliep, 2019) under the "K81" model (Kimura, 1981). This method was repeated to calculate pairwise genetic distances using the BMV58 as a reference against NPCC1309 and NPCC1314 strains. An in-house python script was used to select 1,000 random windows of 20 genes within BMV58 and NPCC1417 genomes to calculated pairwise genetic distances.

3.2.6 Southern blot analysis.

We performed Southern blot analyses with karyotyping gels. Pulsed-field gel electrophoresis was performed under these conditions: 60 seconds during 12 h and 120 seconds during 14 h with an angle of 150º and a velocity of 6V/cm. The strains included were BMV58, CECT12600, NPCC1290, and NPCC1314. DNA was transferred to a nylon membrane Amersham Hybond -N+ (GE Healthcare Europe GmbH, Barcelona, Spain) according to manufactures protocol. We construct the probes using the primers listed in Table S3.3 and the PCR DIG Probe Synthesis Kit (Roche Applied Science, Mannheim, Germany). Each Southern blot analysis was done with high stringency conditions to be sure of the specificity of the probe. Hybridization was prepared with DIG Easy Hyb Granules (Roche Applied Science), following recommendations of the manufacturer for prehybridization, hybridization, and post hybridization washes. For washing, blocking, and detection of DIG-labeled probes DIG Wash and Block Buffer Set (Roche Applied Science) was used. For the detection of DIG-labeled molecules an Anti-Digoxigenin-AP, Fab fragment (1:10.000)

(Roche Applied Science), was used. Finally, CDP-Star Set (Roche Applied Science), a chemiluminescent substrate for alkaline phosphatase was used at 1:100 dilution, and images were stored after 30 min of exposition.

3.2.7 Gene Expression Determination.

For each culture, a 10–20-ml sample was taken each day of wine fermentation. The cells were quickly collected by centrifugation, washed, and frozen with liquid N2. Then, frozen cells were homogenized with a FastPrep-24 (MP Biomedicals, Santa Ana, USA) device with acid-washed glass beads (0.4 mm diameter; Sigma-Aldrich, Madrid, Spain) in LETS buffer (10 mm Tris pH 7.4, 10 mM lithium-EDTA, 100 mM lithium chloride, 1% lithium lauryl sulfate) for 30 s alternating with ice incubation six times. The phenol:chloroform method with minor modifications (Combina et al. 2014) was used to extract and purify total RNA. Then, cDNA was synthesized from the RNA and the expression of *SSU1* genes was quantified by qRT-PCR (quantitative real-time PCR). cDNA was synthesized in 13 µl using 2 µg of RNA mixed with 0.8 mM dNTP's and 80 pmol Oligo (dT). The mixture was incubated at 65 °C for 5 min and in ice for 1 min. Then, 5 mM dithiothreitol (DTT), 50 U of RNase inhibitor (Invitrogen, Waltham, USA), 1 × First-Strand Buffer (Invitrogen), and 200 U Superscript III (Invitrogen) were added in 20 µl mixture and this was incubated at 50 °C for 60 min and 15 min at 70 °C. qRT-PCR gene-specific primers (200 nM), designed from consensus sequences between the different strains, were used in 10 µl reactions, using the Light Cycler FastStart DNA MasterPLUS SYBR green (Roche Applied Science) in a LightCycler® 2.0 System (Roche Applied Science). All samples were processed for DNA concentration determination, amplification efficiency, and melting curve analysis. To obtain a standard curve, serial dilutions (10−1 to 10−5) of a mixture of all samples was used. The average of *ACT1* and *RDN18* constitutive genes was used to normalize the amount of mRNA and to safeguard repeatability, correct

interpretation, and accuracy.

3.2.8 Sulphite tolerance assay.

sulphite tolerance was tested in YEPD +TA (tartaric acid) agar plates as described by Park et. al. (1999). YEPD (2% dextrose, 2% peptone and 1% yeast extract) was supplemented with L- tartaric acid at 75 mM buffered at pH 3.5 and potassium metabisulphite (K2S2O5, MBS) was added to each plate to a final concentration of 0, 0.5, 1, 1.5, 2, 2.5, 3, 3.5, or 4 mM. Yeast precultures were grown overnight in YPD medium. Cell cultures were diluted to OD600 = 1. Then, serial 1:5 dilutions of cells were inoculated in MBS YEPD plates and incubated at 25ºC for a week.

3.3 Results

3.3.1 Two new recombination events in the *SSU1* promoter of *S. uvarum* strains

A total number of 21 *S. uvarum* genomes (Table S3.1) were assembled and examined to find structural variations in the promoter of the *SSU1* gene. Assemblies allowed us to identify two candidate chromosomal rearrangements in the promoter of this gene located at chromosome XVI (Figure 3.1A). One of them was found in the genomes of three fermentative strains (BMV58, CECT12600, and NPCC1417) and involves chromosome VII. The other rearrangement involves chromosome XI and it was found in the strain BR6-2 isolated from a fermentative environment (Almeida et al., 2014). Strains CECT12600 and BMV58 were isolated in Spain from wine fermentations, while BR6-2 and NPCC1417 were isolated from cider fermentations in France and Argentina respectively. These chromosomal rearrangements changed the genomic context in the upstream region of the *SSU1* gene (Figure 3.1B). Instead of the *NOG1* gene present in the ancestral *SSU1* promoter strains, the recombinant chromosome VIIXVI has the *BRP1* gene and the XIXVI has the *FBA1* (gene reverse strand) upstream of *SSU1*. The rearrangement observed between chromosomes VII and XVI was identified at 339 bp upstream of the *SSU1* gene start (Figure 3.1B) within a microhomology region (Figure 3.1C) similarly to the VIIXVI recombination described in *S. cerevisiae* strains. The distance between the end of this gene and the beginning of the *SSU1* gene is 422 bp and 924 bp between the starts of both genes. In the assembled genome of the *S. uvarum* BR6-2 strain, the rearrangement between chromosomes XI and XVI occurred at 393 bp upstream of the *SSU1* gene start also within a microhomology region (Figure 3.1C). Both *SSU1*-promoter recombination events described in this study occurred before the *FZF1* binding site (Figure 3.1C), a well-known *SSU1* gene transcriptional regulator, indicating that this site has been lost in these strains, as also occurred in the two recombination events described in *S.*

cerevisiae.

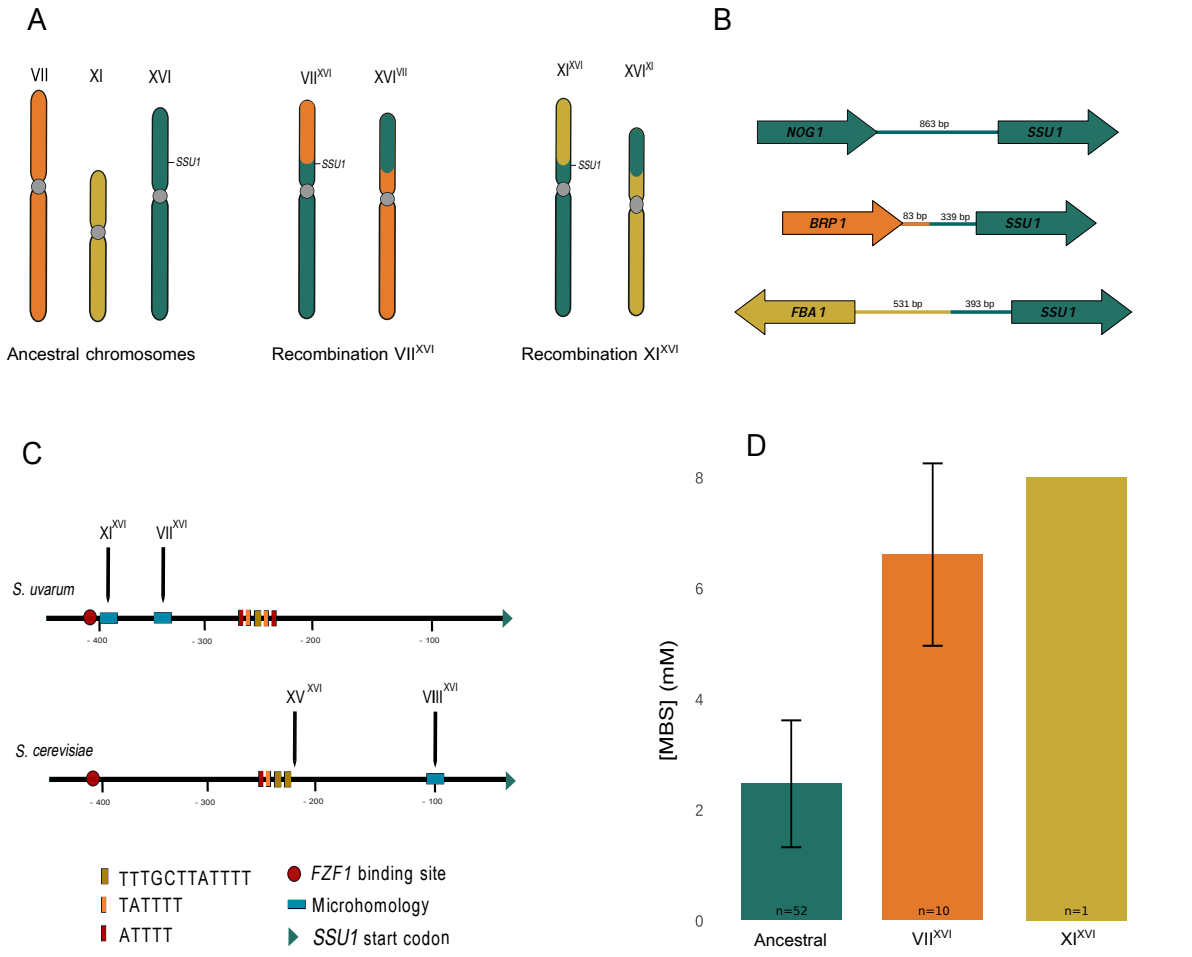


Figure 3.1: New *SSU1* promoter variants found in *S. uvarum.* Panel A. Ancestral type chromosomes; chromosomes VII and XVI after the reciprocal translocation in the *SSU1* promoter; chromosomes XI and XVI after the reciprocal translocation in the *SSU1* promoter. **Panel B.** Gene context surrounding the *SSU1* gene in the genomes with the ancestral and rearranged chromosomes. The distance between the *SSU1* gene and the previous gene is depicted in base pairs, in both the ancestral and recombinant genomes. **Panel C**. *SSU1* promoter and recombination sites described for *S. uvarum*, in this study, and *S. cerevisiae* in previous ones (Pérez-Ortín et al., 2002; Zimmer et al., 2014). *FZF1* binding site and microhomology sites are shown as well as the sites where recombinations occurred in both species reported. **Panel D.** Bar chart showing the tolerance to sulphites of the collection of *S. uvarum* strains tested by drop test assay. Ancestral strains: 52 strains without any of two rearrangements reported; VIIXVI: 10 strains with the chomosome VII and XVI rerrangement; XIXVI: one strain with the chromosome XI and XVI rearrangement. Tolerance to sulphite is measured by the maximum concentration of MBS in which cells can grow. The bars represent the mean of the maximum MBS concentration reached by each strain and the arrows represent the standard deviation.

To determine the frequency of these translocations in *S. uvarum*, we designed specific PCR tests to evaluate a collection of 61 *S. uvarum* strains obtained from different geographic locations and sources, including both natural and anthropic environments, such as wine and cider fermentations (Table S3.2). The PCR amplification allowed us to identify if any of these strains carried any of the two rearrangements identified at the *SSU1* promoter. Rearrangements between chromosomes VII and XVI were found in a total number of 10 strains while the

rearrangement involving chromosomes XI and XVI, was only identified in the BR6-2 strain. Southern blot method was used to classify the most frequent chromosomal rearrangement (VIIXVI) as a reciprocal chromosomal translocation (Figure 3.2).

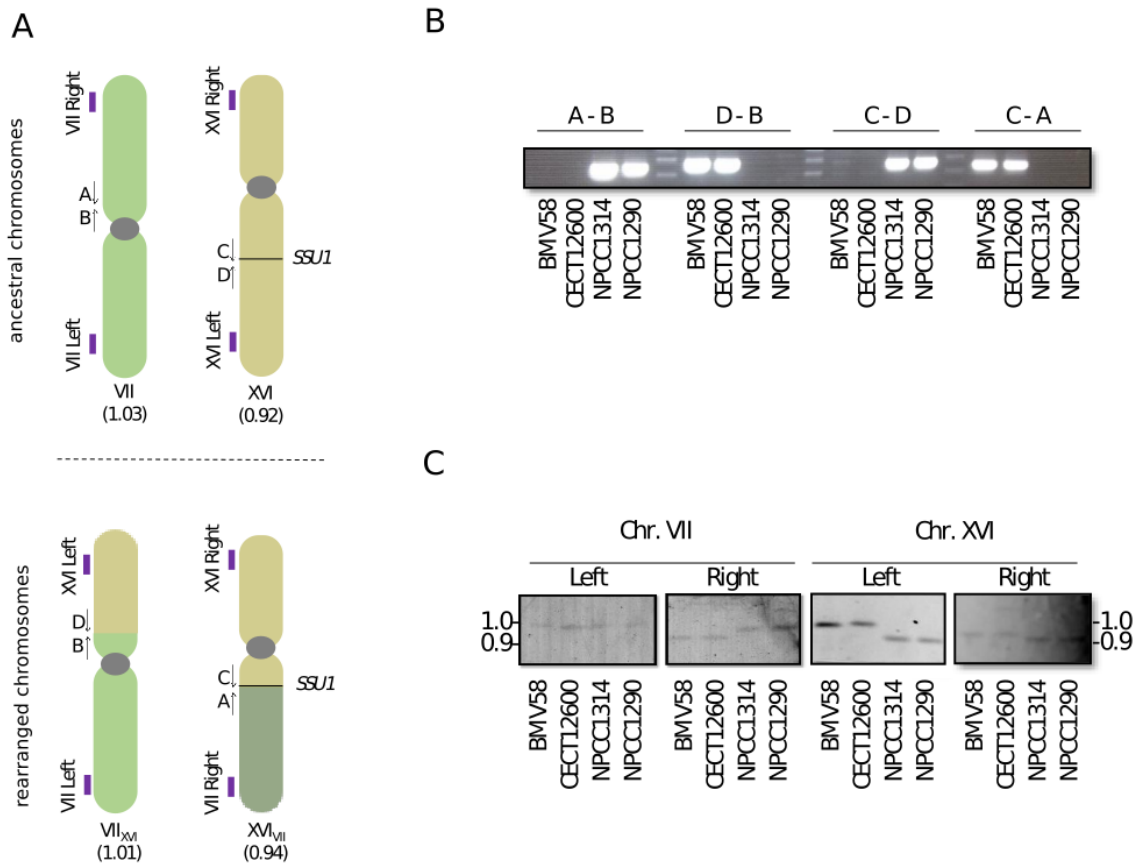


Figure 3.2: Confirmation of the presence of VII^{XVI} recombination in the *S. uvarum* strains BMV58 and CECT12600, compared to the *S. uvarum* NPCC1290 and NPCC1314 strains (ancestral chromosomes). Panel A: A schematic representation of the chromosomal location of primers (arrows) and probes (purple rectangles) used to detect ancestral (VII and XVI) and rearranged (VII^{XVI} and XVI^{VII}) chromosomes. Chromosomal size in Mbp is indicated in brackets. **Panel B:** PCR amplification used to test for the presence of ancestral chromosomes VII (primers A-B) and XVI (primers C-D) or rearranged chromosomes VII^{XVI} (primers D-B) XVI^{VII} (primers C-A). **Panel C:** Southern blots with chromosome VII and XVI left and right probes performed in genomic DNA obtained from BMV58, CECT12600, NPCC1290, and NPCC1314 *S. uvarum* strains. DNA fragment size is shown in Mbp.

3.3.2 Strains carrying the chromosomal rearrangements in the *SSU1* promoter are more tolerant to sulphite

Sulphite tolerance was evaluated by drop test assays in the 61 *S. uvarum* strains to establish a correlation between the presence of a chromosomal rearrangement and the ability to grow in high concentrations of sulphite. Sulphite tolerance was tested in

plates containing different concentrations of potassium metabisulphite (MBS) ranging from 0 to 8 mM. The results showed a significantly higher MBS resistance of the strains with any of the two described rearrangements in comparison with the strains with the ancestral type *SSU1* promoter Figure 3.1D. Only the strains carrying any of the two reported recombinations were able to grow in plates with the maximum concentration of MBS tested, while the maximum tolerable concentration of MBS of strains without recombination was 4.0 mM. This phenotypic characterization of the *S. uvarum* strains, together with the PCR amplification, allowed us to identify a clear correlation between the presence of a rearrangement in the *SSU1* promoter and the tolerance to sulphite (Figure 3.1D, Table S3.2).

To confirm that the recombination in the *SSU1* promoter was leading to an increase of the expression of this gene, qPCR studies were performed with the *S. uvarum* strains. Fermentations with and without MBS were conducted with strains carrying the most frequent recombination (VII^{XVI}). We compared the *SSU1* expression of the wine BMV58 and CECT12600 strains against the *SSU1* expression of two strains with no recombinations: the strain CBS2986 (Pérez-Través et al., 2014), isolated from wine fermentation, and the natural NPCC1290 strain isolated from an *Araucaria araucana* tree (Rodríguez et al., 2014). Relative expression of the *SSU1* gene to the strain NPCC1314 (*SSU1* promoter without recombination) was calculated (Figure 3.3). In this experiment, we observed a clear overexpression of the *SSU1* gene in the two strains carrying out the recombination VIIXVI when compared to the wild strain (NPCC1290) but also to the wine strain (CBS2986). This suggests that the *SSU1* recombination is a specific adaptation to sulphite presence rather than an adaptation to the wine environment. We also observed that the overexpression of the *SSU1* gene is not dependent on the presence of sulphite in the media. We performed a two-way analysis of variance (ANOVA) and both BMV58 and CECT12600 strains showed significantly higher expression levels than the other strains in the two conditions analysed (with and without MBS), although expression was higher with MBS

for all the strains, especially during the first two days of fermentation (Figure 3.3).

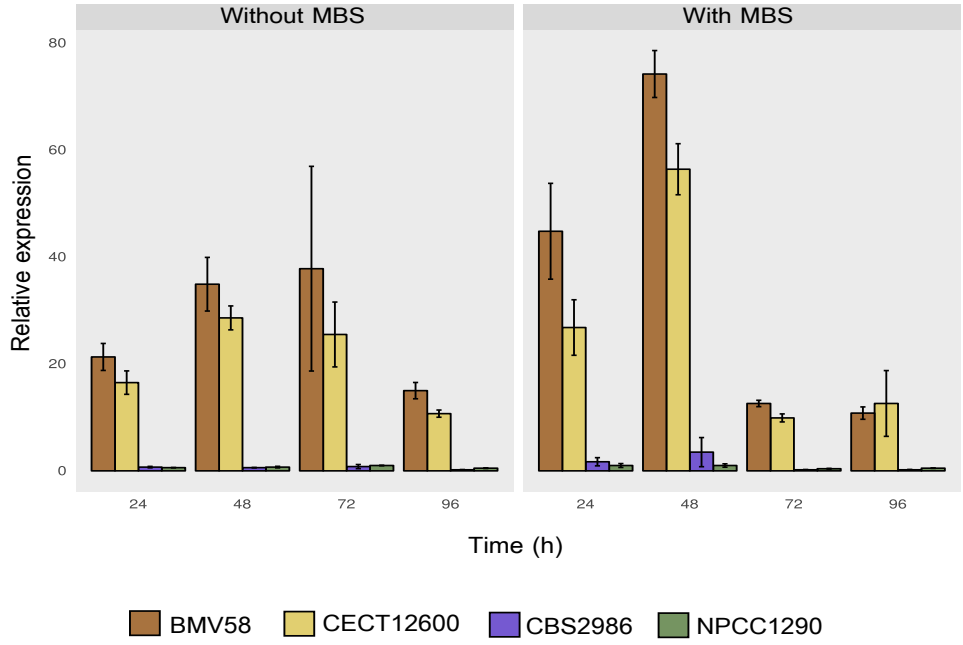


Figure 3.3: Relative *SSU1* expression and growth in *S. uvarum* strains during fermentation. Expression of the *SSU1* gene was studied during wine fermentation in synthetic must with or without sub lethal sulphite (MBS) concentration (15 mg/l) for two strains with the VIIXVI recombination (BMV58 and CECT12600) and two with the ancestral chromosomes (NPCC1290 and CBS2986). Daily samples were taken until day four and, after mRNA extraction, *SSU1* gene expression was quantified by qPCR. Two constitutive genes (*ACT1* and *RDN18*) were used to normalize qPCR data. All expression measures were relativized to the *SSU1* expression in the NPCC1314 strain (ancestral *SSU1* promoter) grown under the same fermentation conditions.

A second fermentation experiment was conducted to measure the *SSU1* expression of both BMV58 (VIIXVI) and BR6-2 (XIXVI). Besides, to demonstrate the effect of the two different recombinations in the *SSU1* gene expression, we obtained two modified versions of the *S. uvarum* type strain CBS7001, where the wild type *SSU1* promoter was substituted with the BMV58 or BR6-2 *SSU1* promoters. *SSU1* gene expression was also measured in these mutants together with the wild type CBS7001 (Figure 3.4).

First, we confirmed that both types of rearrangements generated *SSU1* overexpression compared to the wild-type strain (CBS7001). Second, we observed that, when the BR6-2 *SSU1* promoter was introduced in the CBS7001 strain,

generating the CBS7001(prBR6-2) strain, an overexpression of *SSU1* is produced in the modified strain. This overexpression was not significantly different (t-test; $p<0.05$) than that observed for the strain BR6-2, except for time point 96 h with MBS, where values were 1.6x fold higher in the strain CBS7001(prBR6-2). In the other case, when the promoter of BMV58 (VIIXVI) was introduced instead of the wild type promoter of the CBS7001 strain, generating the strain CBS7001(prBMV58), a clearer overexpression in the *SSU1* levels was observed compared with the CBS7001 strain. This phenomenon was especially prominent after the first 24 hours of fermentations with and without MBS. The overexpression of *SSU1* in the edited strain CBS7001(prBMV58) showed no significantly different values (t-test; $p<0.05$) compared to the strain BMV58 except at the 24 hours time point without MBS and at 24 h and 72 h time points with MBS when the transcriptions levels were significantly lower when compared to the BMV58 strain. These latter results suggest that, unless the new promoter of CBS7001(prBMV58) strain produces a significant overexpression of *SSU1*, other factors as the chromosomal context or other upstream/downstream elements, not transferred to CBS7001(prBMV58) could have further influenced *SSU1* expression in the BMV58 strain.

3.3.3 Phylogenetic reconstruction and the origin of the *SSU1*-promoter recombination.

A total number of 11 strains were found to have the recombinations described above. These strains were all isolated from wine or cider fermentations Table S3.2, anthropic environments where sulphite is commonly used as an antimicrobial preservative. Two of these strains were also isolated from Argentinean cider fermentation (as the strain NPCC1417). No recombination was found in the South American strains isolated from natural environments, neither in the ones isolated from chicha, a beverage performed in traditional fermentation with no sulphite addition.

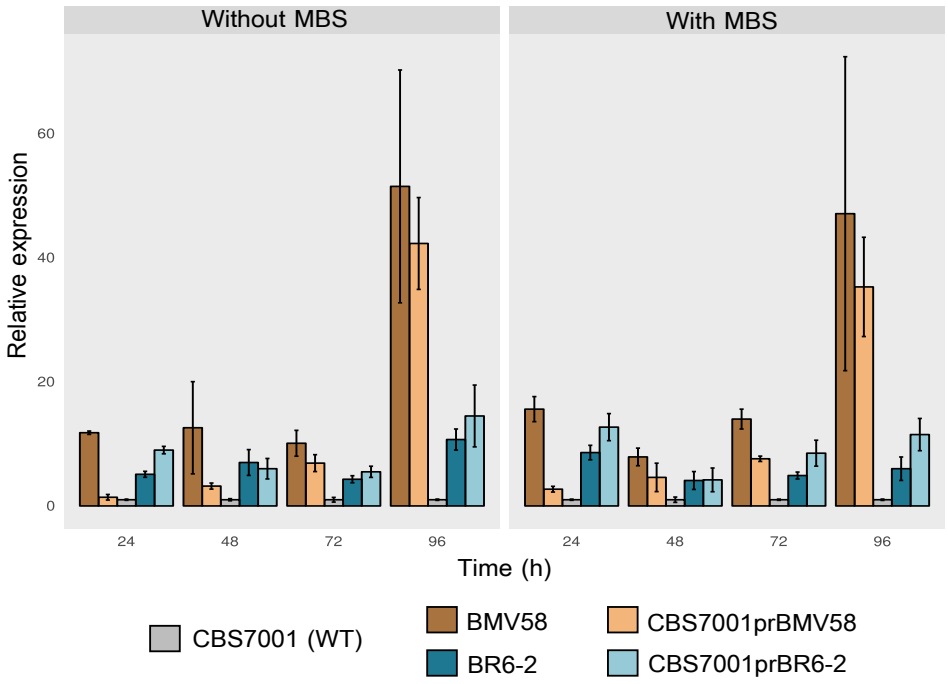


Figure 3.4: Relative *SSU1* expression in *S. uvarum* wild type and edited strains grown in a fermentation experiment. Expression of *SSU1* gene was studied during wine fermentation in synthetic must with or without sub lethal sulphite (MBS) concentration (15 mg/l) for a strain with the wild type chromosomes (CBS7001), a strain with the VIIXVI recombination (BMV58), a strain with the VIIXVI recombination (BR6-2), a modified version of the CBS7001 type strain with the BM58 *SSU1* promoter (CBS7001pBM58) and a modified version of the CBS7001 type strain with the BR6-2 *SSU1* promoter (CBS7001prBR6-2). Daily samples were taken until day four and, after mRNA extraction, *SSU1* gene expression was quantified by qPCR. Two constitutive genes (*ACT1* and *RDN18*) were used to normalize qPCR data. All expression measures were relativized to the *SSU1* expression in the CBS7011 wild type strain grown under the same fermentation conditions.

To unravel the origin of the new chromosomal translocations discovered in this study we performed a phylogenetic analysis using whole-genome sequencing data from 21 strains. The selected strains represent different origins, populations, and *SSU1* promoter versions (ancestral, VII^{XVI}, or XI^{XVI}). The phylogeny revealed that strains carrying recombinations in the *SSU1* promoter are located at different branches in the tree and they did not constitute a monophyletic group (Figure 3.5). It also revealed that the recombinant strains were not located at branches belonging to *S. uvarum* strains from Australasia or South America B populations, previously described by Almeida et al. (2014). A well-supported branch includes the strains CRUB1778 and CRUB1779, which belong to the South America B population. On the same branch is located the

NPCC1290 strain, isolated in Argentina from *Araucaria araucana* tree bark (Rodríguez et al., 2014).

As a sister group of the CRUB1778-CRUB1779-NPCC1290 cluster, different internal branches were observed. South American and European strains appear as intermixed, including those South America A and Holarctic strains described by Almeida et al. (2014). Most of these branches showed low support values, indicating that other relationships are possible. Interestingly, three Argentinean strains sequenced in this study (NPCC1417, NPCC1309, and NPCC1314) are located in a branch next to the strains classified as South America A and Holarctic. Two of these strains, NPCC1314, and NPCC1309, were isolated from traditional apple chicha fermentation, whereas the strain NPCC1417 was isolated from cider fermentation. This strain carries the chromosome VIIXVI recombination which is shared with the European strains BMV58 and CECT12600, both isolated from wine fermentations. To further investigate the origin of the chromosomal rearrangement shared between the Argentinean NPCC1417 strain and the BMV58 and CECT12600 European wine strains, we extracted the genes surrounding the *SSU1* promoter to calculate nucleotide distances. Nucleotide gene distances of these genes revealed that NPCC1417 was closer to BMV58 than to CECT12600, having more genes with nucleotide distances equal to zero. To estimate the size of the conserved region between NPCC1417 and BMV58, we extracted genes towards both sides of the rearrangement and calculate distances between both strains until these distances were higher from zero. A region of 21 genes from the *SSU1* gene towards the right side (until YPL068C gene) and 33 genes from the *SSU1* gene towards the left side (until the YGL044C gene) contained genes in NPCC1417 sharing the same coding sequences as the BMV58. The genomic region size of this selective sweep was estimated in 51,000 bp and 66,000 bp towards the right and left side, respectively, from the *SSU1* gene (Figure 3.6).

To confirm that this effect was a consequence of the action of natural selection in

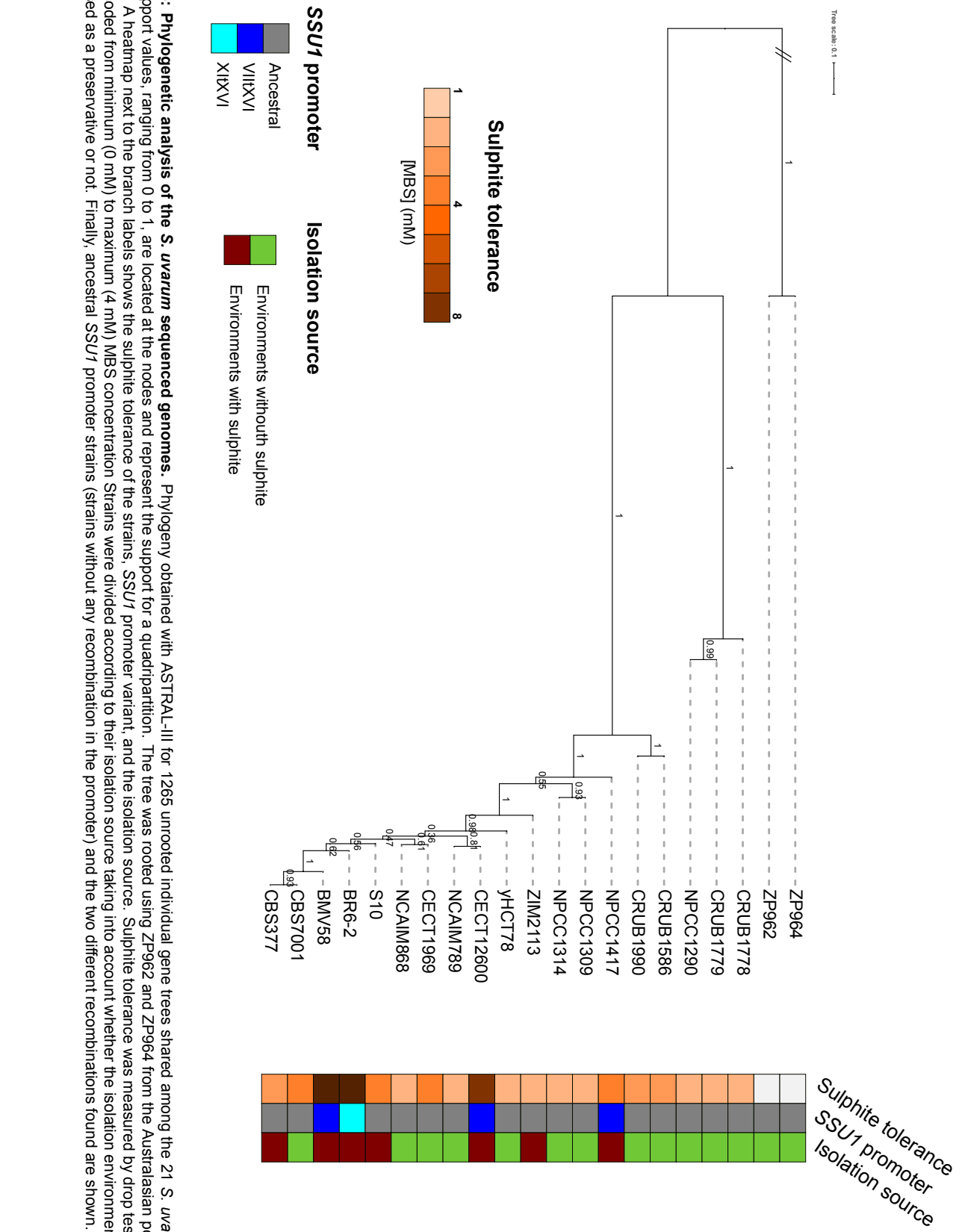


Figure 3.5: Phylogenetic analysis of the *S. uvarum* sequenced genomes. Phylogeny obtained with ASTRAL-III for 1265 unrooted individual gene trees shared among the 21 *S. uvarum* strains. Branch-support values, ranging from 0 to 1, are located at the nodes and represent the support for a quadripartition. The tree was rooted using ZP962 and ZP964 from the Australasian population as outgroups. A heatmap next to the branch labels shows the sulphite tolerance of the strains, *SSU1* promoter variant, and the isolation source. Sulphite tolerance was measured by drop test assay and it is color-coded from minimum (0 mM) to maximum (4 mM) MBS concentration Strains were divided according to their isolation source taking into account whether the isolation environment contained sulphite used as a preservative or not. Finally, ancestral *SSU1* promoter strains (strains without any recombination in the promoter) and the two different recombinations found are shown.

the *SSU1* promoter causing a selective sweep, we also calculated nucleotide genetic distances in the *SSU1* promoter of the surrounding genes of the strains NPCC1309 and NPCC1314. We selected these strains because both were isolated in the same geographic location as the NPCC1417 and they were closer in the phylogenetic tree. None of the selected genes in the NPCC1309 and NPCC1314 shared this region with NPCC1417. To study how frequent was in the NPCC1417 genome to find genomic regions containing consecutive conserved genes with the BMV58, we randomly selected 1,000 windows of twenty genes along the genome of the NPCC1417 and calculated genetic distances against the BMV58 orthologous genes. A window of 20 genes containing all its pairwise distances equal to zero resulted significantly different from the distribution created from the 1,000 randomly selected windows (p-value < 0.05; Whitney-Wilcoxon test).

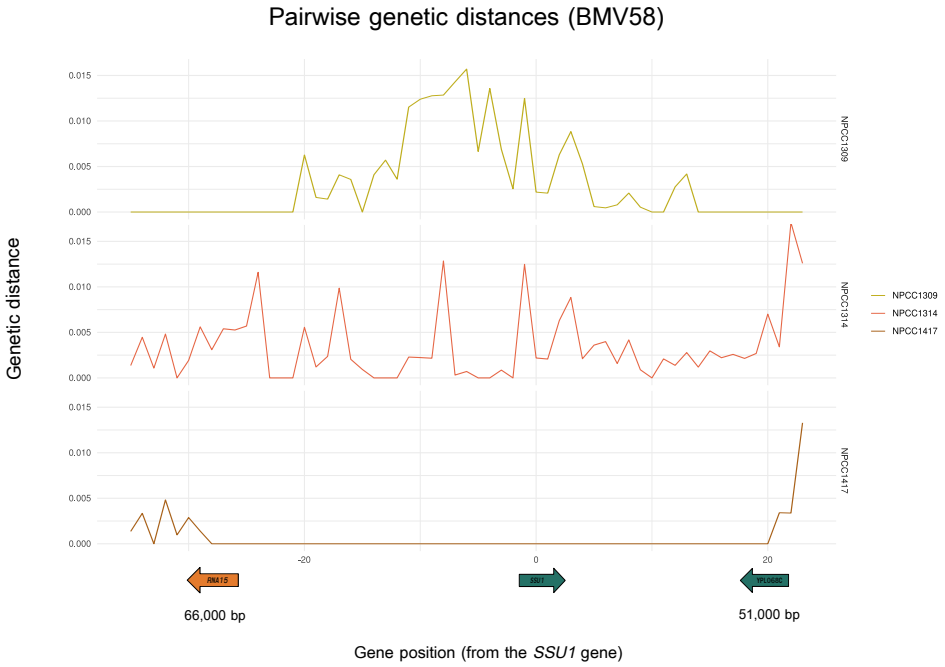


Figure 3.6: Determination of a selective sweep in the NPCC1417 genome. Pairwise genetic distances of the genes surrounding the *SSU1* promoter were calculated and represented in this study using BMV58 as reference. The x-axis represents the gene position using *SSU1* as reference (position 0). Green genes correspond to genes from the reference chromosome XVI and the orange gene corresponds to the reference chromosome VII.

3.4 Discussion

In this work, we present the case of a convergent adaptation of *S. uvarum* strains, isolated from fermentation environments, to grow in sulphite containing media, a preservative usually added in industrial processes such as wine or cider fermentation. This is the first example reported in which different chromosomal rearrangements originated by two different recombination events resulted in the overexpression of the *SSU1* gene and, therefore, an increase of the sulphite tolerance in the strains carrying the recombination.

In *S. cerevisiae*, different cases of structural variations have been described in the promoter of the *SSU1* gene. These variations include chromosomal translocations (Pérez-Ortín et al., 2002; Yuasa et al., 2004; Zimmer et al., 2014), which involve different chromosomes than those reported for *S. uvarum*, and a chromosomal inversion (García-Ríos et al., 2019). These *SSU1* promoter variants described for *S. cerevisiae* have been reported to cause the overexpression of this gene being those strains much more tolerant to the presence of sulphites in the culture media. This is the first time that a recombination in the *SSU1* promoter, providing an adaptive value, is described for another *Saccharomyces* species, different from *S. cerevisiae*.

As far as we know, our work describes the first example of a phenotypic convergence produced by independent chromosomal rearrangements in two of the most divergent Saccharomyces species, *S. cerevisiae*, and *S. uvarum* (80% of nucleotide divergence). Strains of both species exhibit rearrangements at different locations in the promoter of the *SSU1* gene that allows adaptation to tolerate high sulphite concentrations. The presence of molecular mechanisms resulting in a phenotypic convergence reinforces the enormous adaptive role that exerts the overexpression of the *SSU1* gene in industrial strains. Interestingly, the four recombination events described so far are independent, produced at different locations

of the *SSU1* promoter, and involving reciprocal translocations between chromosome XVI and different partners. Our results, including several complementary approaches, confirm the strong selection pressure that the antimicrobial effect of sulphite imposes on yeasts in human-driven fermentations, as well as remarks on the role of chromosomal rearrangements as a source of variation to promote yeast adaptations in fast-evolving environments.

The molecular mechanisms that produced the overexpression of the *SSU1* gene remains unclear. The regulation mechanism of the *SSU1* gene known until now is mediated by the five-zinc-finger transcription factor codified by the *FZF1* gene. This gene acts as a positive regulator of the *SSU1* by binding directly to its upstream promoter region (Avram et al., 1999). The Fzf1p binding sequence has been described as 5'-CTATCA-3'. This sequence is present at many sites throughout the genome but *SSU1* is the only demonstrated target. We have identified the binding sequence in the ancestral promoter *SSU1* version of strains without chromosomal rearrangements. Interestingly, both rearrangements described in this work, occurred before the *FZF1* binding site, like in *S. cerevisiae*, hence, the *SSU1* promoter region lost the Fzf1p binding site due to the chromosomal rearrangements. Our main hypothesis is that *FZF1* is not regulating the expression of the *SSU1* gene in these *S. uvarum* strains. Instead of that, this gene could be possibly constitutively active or being regulated by another of several transcription factors that have not been identified yet. We can also conclude from our experiments that the overexpression effect of the *SSU1* gene is not depending on the presence of sulphite in the media as this gene is highly expressed from the early stages of fermentation with and without sulphite.

The XI^{XVI} recombination was found in a unique European strain isolated from a cider fermentation while the VII^{XVI} recombination event is shared among European and South American strains. Previous population analyses performed on the *S. uvarum* species classify them into four differentiated populations: Australasian, South

America B, South America A, and Holarctic (Almeida et al., 2014). In a recent study (Gonzalez Flores et al., 2020), the existence of South America A population, genetically differentiated from the Holarctic population has been questioned and the authors suggest that these strains are the result of the genetic admixture of Holarctic and South America B strains. This fact, together with the high incongruence observed in our phylogenic reconstruction, leads us to think that they should not be properly considered as two different populations because they are, indeed, a mixed population. This idea is supported by the shared chromosomal rearrangement described in this study between strains isolated in Europe and Argentina. We hypothesize that these strains probably coexisted at the same location. This rearrangement was spread by sexual reproduction among different strains and it became fixed later in those strains grown in human-related environments where sulphite is used as a microbial preservative. This premise is supported by the selective sweep observed in the *SSU1* surrounding gene sequences of NPCC1417. In this region of the NPCC1417 genome, we have observed a selective sweep of approximately 50 genes where coding sequences do not show a single nucleotide variation when compared to the BMV58 corresponding orthologous sequences. We randomly sampled other genomic regions of the NPCC1417 to confirm that this was not happening in other parts of the genome confirming that this phenomenon is probably due to the strong selective pressure that sulphite exerts on yeast cells. Thus, our data suggest that VII^{XVI} recombination had a unique origin in a European strain, and then, it was inherited by these South American strains due to hybridizations between European and South American strains.

Finally, our discovery highlights the role of the *SSU1* gene promoter as a hotspot of evolution at different taxonomic levels. *S. cerevisiae* is the predominant species in sulphite containing environments as wine, cider, and other fermented beverages. However, *S. uvarum* can be also dominant in certain types of fermentation, especially those performed at lower temperatures (González Flores et al., 2019; Naumov et al., 2002; Rodríguez et al., 2017). This abundance can explain the detection of the *SSU1*

locus recombinations exactly in those species, as an adaptation to sulphite. Other species such as *Hanseniospora uvarum*, *Metschnikowia pulcherrima*, *Brettanomyces sp*. among others can be found in relatively high numbers in those environments at the beginning and even at more advanced stages of fermentations (Cousin et al., 2017; Varela, 2016). Thus, an interesting outcome of this study is that other chromosomal rearrangements in the gene responsible for the sulphite detoxification should not be discarded in other species present in those environments.

CHAPTER 4

High-quality new assemblies of *Saccharomyes* genomes provide insights into their evolutionary dynamics

4.1 Introduction

Saccharomyces cerevisiae is one of the most interesting and well-studied genetic model organisms. It was the first eukaryote of having its genome fully sequenced (Goffeau et al., 1996). Since then, several genome sequences of *S. cerevisiae* strains isolated from a wide range of environments and locations have been described (Gallone et al., 2018; Legras et al., 2018; Liti et al., 2009; Peter et al., 2018; Yue et al., 2017). The number of published sequences from whole-genome sequencing projects ranges at present 55-56 thousand. The comparative genomics analysis of a model organism with respect to its closest relatives can shed light on the

molecular mechanisms of evolution. The *Saccharomyces* genus is turning into a very attractive model for evolutionary biologists as their species possess tiny genomes of 12 Mb of nuclear genome distributed into 16 chromosomes. *Saccharomyces* genomes display less complexity than other eukaryotic genomes because of their reduced number of introns and active transposons (Dujon, 2006). Besides these genomic features, the *Saccharomyces* genus also shows a very high nucleotide divergence across its species. The amino acid identity found between the early-divergent *S. uvarum* and *S. cerevisiae* orthologous proteins is around 80%, the same as the similarity found between human and bird proteins (Dujon, 2006). *Saccharomyces* species show striking differences in their ecology. Besides *S. cerevisiae*, *S. uvarum* is the only non–hybrid species isolated from anthropogenic environments such as wine and cider fermentations (Almeida et al., 2014; Rodríguez et al., 2017). Natural habitats of wild *Saccharomyces* species are oaks and beeches (Goddard and Greig, 2015), and several species have been found coexisting in the same habitats (Charron et al., 2014; Lopes et al., 2010; Sampaio and Gonçalves, 2008). Differences in optimal growth temperatures between species have been proposed to be the reason for the coexistence of these yeasts instead of the competition (Salvadó et al., 2011). *S. eubayanus*, *S. uvarum* and *S. kudriavzevii* are considered cryotolerant species with significantly lower optimal growth temperatures than *S. cerevisia*e and *S. paradoxus*, considered as thermotolerant. Understanding how *Saccharomyces* yeasts have colonised different environments can shed light on the evolution of this genus. Large-scale structural variations or gene duplications are among the molecular mechanisms that can drive adaptation. Subtelomeric regions are repeat-rich regions proximal to the telomeres that display high mutation and recombination rates (Barton et al., 2008). These regions have been understudied when compared to non-subtelomeric genomic locations because the assembly of the subtelomeres is often hampered by the extensive sequence similarity and the presence of repetitive regions. However, the subtelomeric gene families have been demonstrated to have

critical roles in the adaptation of the organisms (Brown et al., 2010). For example, yeast gene families involved in carbohydrate utilization and biofilm formation have been identified in subtelomeric locations (Denayrolles et al., 1997; Michels and Needleman, 1984; Verstrepen and Klis, 2006). The long-read sequencing technologies (PacBio and Oxford Nanopore) have gained power and interest during the last years. The use of these tools provides an opportunity to generate continuous genome assemblies to resolve complex regions that remained unclear, such as the subtelomeres. These technologies have been proven useful for population genomic analyses of *S. cerevisiae* and *S. paradoxus* (Yue et al., 2017). They have also been used in the whole-genome sequencing of the type strains of *S. eubayanus* and *S. jurei* (Brickwedde et al., 2018; Naseeb et al., 2018). However, there is still an important lack of non-*S. cerevisiae* whole-genome sequences generated using these novel technologies. In this work, we have used a combination of short and long-read sequencing to obtain whole-genome sequences of *S. mikatae*, *S. kudriavzevii* and *S. uvarum*. We combined these data with those available from previous studies to perform a comparative genomics study of the *Saccharomyces* genus focusing on the annotation of subtelomeric regions. Our study provides new insights into the dynamics of *Saccharomyces* evolution together with new annotated genomic sequences that can serve as genome references for future research studies using *Saccharomyces* as a model genus.

4.2 Materials and methods

4.2.1 Genome sequencing and assembly

Paired-end Illumina libraries were prepared as previously described (Shen et al., 2018). Libraries were sequenced with the Illumina HiSeq 2000, HiSeq 2500 Rapid or MiSeq equipment. The quality and quantity of the finished libraries were assessed using an Agilent DNA1000 series chip assay (Agilent Technologies) and Invitrogen Qubit HS Kit (Invitrogen, Carlsbad, CA), respectively, and the library was standardized to 2 nM. Images were analysed using CASAVA version 1.8.2. PacBio sequencing was performed using the PacBio Single Molecule, Real-Time (SMRT) DNA sequencing technology (platform: PacBio RS II; chemistry: P4-C2 for the pilot phase and P6-C4 for the main phase). The raw reads were processed using the standard SMRT analysis pipeline (v2.3.0). The de novo assembly was carried out following the hierarchical genome-assembly process (HGAP) protocol. Short-reads (Illumina) were mapped against long-read based assemblies using bowtie2 (Langmead and Salzberg, 2012). Mappings were used to assess assembly correctness using Pilon (Walker et al., 2014).

4.2.2 Annotation

Annotation of corrected genome assemblies was done using a combination of two approaches (Figure 4.1). On the one hand, the curated and well-annotated genome assembly of *S. cerevisiae* S288c (Goffeau et al., 1996) was used as the reference assembly where annotated features are transferred to our assemblies using RATT (Otto et al., 2010). In this step, the yeast genomes' systematic gene names were transferred from the reference annotated assembly (S288c) to our annotations when sequence homology and synteny were conserved between them. Systematic gene names allow us to link the annotated features with the *Saccharomyces* Genome

Database (SGD) (Cherry et al., 2012) containing meaningful information about gene description and gene ontology. On the other hand, Augustus software (Stanke and Waack, 2003) was used for *de novo* gene prediction. After that, both annotations were merged. Annotated genes by RATT were conserved on those positions in which there was an overlap between both annotations. Annotated genes by Augustus were conserved in positions where RATT did not transfer any gene from the reference annotation. Artemis tool (Rutherford et al., 2000) was used to visualize annotated assemblies and to perform manual corrections.

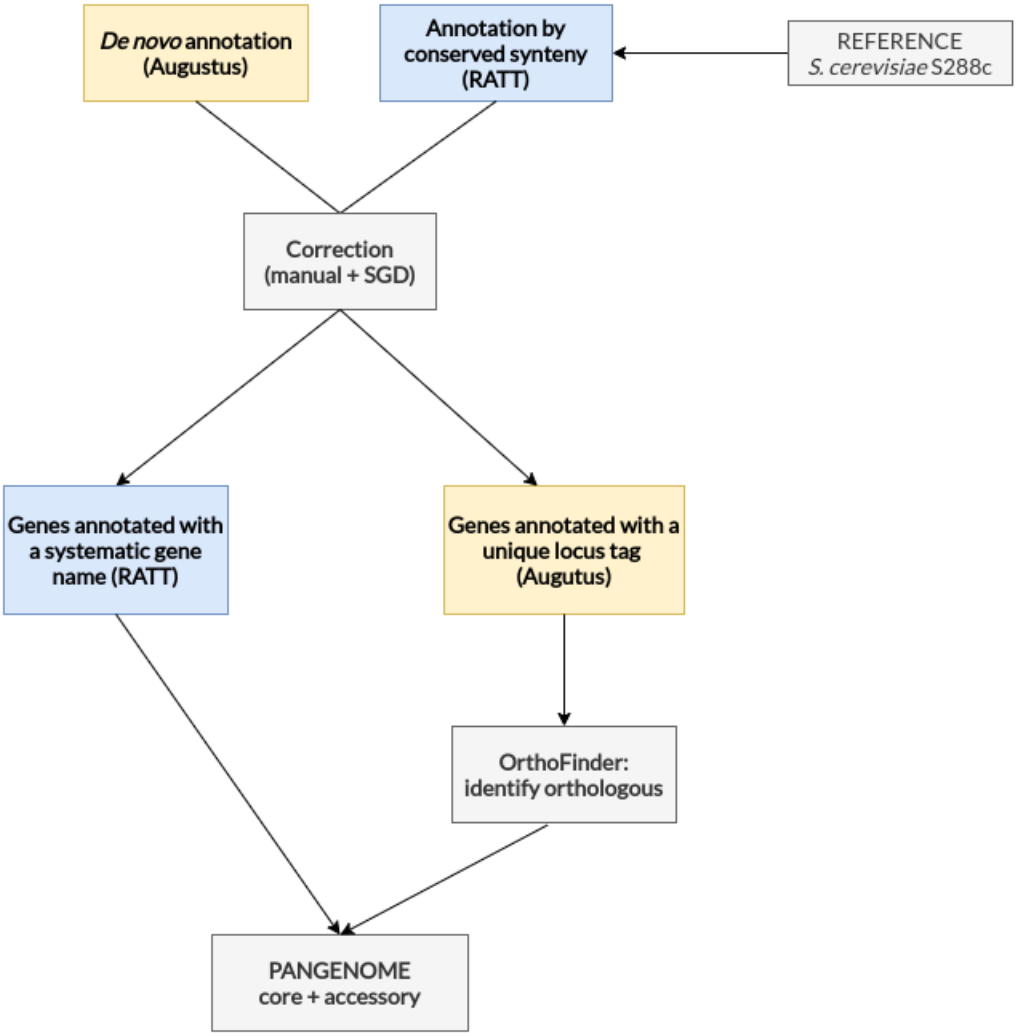


Figure 4.1: Annotation pipeline. Pipeline followed for the annotation of all the *Saccharomyces* assemblies used in this study. *Saccharomyces* assemblies used in this project and sequenced in a previous study were also annotated using this pipeline.

4.2.3 Pangenome analysis

The annotated genes were divided into two types in the final curated annotated assemblies. The first type included those annotated genes by homology and synteny conservation with respect to the reference *S. cerevisiae* S288c (RATT) named with the systematic annotation (e.g. YAL069W). These gene names were shared across all annotated assemblies, and they allowed us to identify orthologues automatically. The second type of annotated genes were annotated only by the de novo gene prediction method (Augustus). These genes were named using a locus tag which was unique for each assembly and gene. These genes are not syntenic or are absent in the reference strain (S288c) but they can be present in the other species genomes, they could also be duplicates (e.g. paralogues), or ancestral genes lost in the reference strain. To identify orthology relationships for genes not present in the reference genome, we used OrthoFinder (Emms and Kelly, 2019) with the *de novo* annotated genes. Our pangenome definition aimed to find the set of orthologous and syntenic genes common to all *Saccharomyces* species in a relation of 1-to-1. To perform this, we firstly obtained a set of common genes for each species. In the cases of *S. eubayanus, S. jurei*, and *S. mikatae*, we used all the annotated genes as we had only one strain available per species. For the rest of the species, we calculated the common set of genes shared by the strains of the same species to be used in the *Saccharomyces* pangenome definition. To identify whether our set core genes were enriched in gene ontology (GO) terms or pathway categories, we used the Gene List tool of the *Saccharomyces* Genome Database (`https://yeastmine.yeastgenome.org/yeastmine/bag.do`). Genes with a *p*-value lower than 0.05 after a Holm-Bonferroni (Aickin and Gensler, 1996) test correction were retained.

4.2.4 Subtelomeric regions

Subtelomeric regions were identified using the same criteria as Yue et al. (2017) for *S. cerevisiae* and *S. paradoxus*. In this previous work, these regions were defined based on conserved syntenic gene blocks as boundaries. In the present study, we used those boundaries and defined them again for the new set of *Saccharomyces* species. A gene was considered as a subtelomeric boundary when it is the first conserved gene from the beginning/end of the syntenic regions of each chromosome common to all the annotated *Saccharomyces* assemblies. Subtelomeric lengths and subtelomeric gene density in terms of the number of annotated genes within the subtelomeric regions were computed to unravel differences among *Saccharomyces* species. Signals of species clustering based on subtelomeric lengths and gene density values were used to perform a hierarchical clustering analysis with the Ward's method based on Euclidean distances (Ward, 1963).

4.2.5 Subtelomeric protein-coding gene families

Coding sequences of the annotated genes within subtelomeric regions were extracted to define the subtelomeric families. A database containing all the translated proteins from these coding sequences was created to identify protein-coding gene families. Blastp searches all-against-all was performed and the e-values were used to obtain a MCL clustering analysis of proteins (van Dongen, S. Graph Clustering by Flow Simulation Ph.D. thesis, University Utrecht, 2000), with an inflation parameter of 2. This step allowed us to define clusters (families) within our protein database. These protein-coding gene families were functionally annotated using OmicsBox version 1.3 (Götz et al., 2008). Differences in copy numbers within subtelomeric families across species were studied to identify the potential expansion of protein-coding gene families in particular species. A matrix containing values of the number of genes in each family

per species was created. These values were normalised according to the family sizes. This matrix was used to perform a Ward's hierarchical clustering analysis based on Euclidean distances to detect signals of species clustering.

4.2.6 Phylogeny reconstruction

Concatenated sequences of the core genes defined in the pangenome were used to obtain a phylogenetic reconstruction. Translated proteins were aligned with Mafft (Katoh and Standley, 2013) and multiple sequence alignments were back-translated into codons using an in-house Python script. Trimal (Capella-Gutiérrez et al., 2009) was used on each codon alignment for sequence trimming using the gap threshold option of 0.8. Trimmed alignments were concatenated into one single alignment. Maximum-likelihood phylogeny reconstruction was performed on the trimmed concatenated alignment using RAxML (Stamatakis, 2014) with the GTR-Γ model and 100 bootstrap replicates. The tree was rooted in the *S. eubayanus*–*S. uvarum* leading branch. The resulting tree was visualized using iTOL (Letunic and Bork, 2016).

4.3 Results

4.3.1 Highly accurate annotations of *Saccharomyces* long-read based genome assemblies

In this work, we have used long-read sequencing to obtain new end-to-end genome assemblies of five strains from three *Saccharomyces* species. Two *S. uvarum* strains: the type CBS7001^T and ZP964 from the European and Australasian populations, respectively (Almeida et al., 2014); two *S. kudriavzevii* strains: the type IFO1802^T from Japan, and the European ZP591 (Naumov et al., 2000a; Sampaio and Gonçalves, 2008); and the *S. mikatae* type strain IFO1815^T, also from Japan (Naumov et al., 2000a). Our thorough annotation pipeline was used to annotate these genome assemblies together with other *Saccharomyces* long-read-based genome assemblies obtained in previous studies (Brickwedde et al., 2018; Naseeb et al., 2018; Yue et al., 2017). With this method, we accurately identified a high number of orthologous and syntenic genes among all *Saccharomyces* species (Table 4.1). This pipeline also allowed us to avoid errors derived from the sole use of automatic annotation pipelines, such as paralog mislabeling. Our annotations will be available for the yeast scientific community and can be used as the basis of future *Saccharomyces* 'omic' studies.

4.3.2 Early-divergent species *S. eubayanus* and *S. uvarum* show striking differences in subtelomeric lengths

We defined subtelomeric gene boundaries for all *Saccharomyces* species to study the evolutionary dynamics of their subtelomeric regions. From the 32 gene boundaries previously defined by Yue et al. (2017), only 15 remain unchanged while the rest were redefined by considering the whole species set. Accordingly, the

Table 4.1: Number of annotated genes in *Saccharomyces* assemblies. A summary of the annotated genes using the pipeline described in section 4.2.2. The numbers represent the total number of annotated genes using the pipeline (3rd column) and the total number of those annotated genes that were annotated by RATT (4th column).

Species	Strain	No. total annotated genes	No. total annotated genes (RATT)
S. eubayanus	CBS12357	5682	5462
S. uvarum	CBS7001	5750	5541
S. uvarum	ZP964	5702	5484
S. kudriavzevii	IFO1802	5835	5597
S. kudriavzevii	ZP591	5764	5584
S. mikatae	NBRC1815	5692	5429
S. jurei	NCYC3947	5606	5398
S. cerevisiae	DBVPG6044	5529	5389
S. cerevisiae	DBVPG6765	5531	5430
S. cerevisiae	S288c	5989	5894
S. cerevisiae	SK1	5533	5389
S. cerevisiae	UWOPS03	5560	5410
S. cerevisiae	Y12	5535	5411
S. cerevisiae	YPS128	5530	5422
S. paradoxus	CBS432	5523	5369
S. paradoxus	N44	5519	5374
S. paradoxus	UFRJ50816	5510	5366
S. paradoxus	UWOPS91917	5472	5324
S. paradoxus	YPS138	5504	5366

new boundaries for 17 subtelomeric regions were displaced towards the centromere, increasing their sizes. Subtelomeric lengths compared across *Saccharomyces* assemblies displayed a high variability, ranging from 1 kb to almost 90 kb, with a median length of 24 kb. This variability was observed at both the species and subtelomere levels. Despite this variability, some patterns were observed on particular subtelomeres. The biggest differences were observed at the left arm of chromosomes VI, XIII, and the right arms of chromosomes VII, X, XI, and XIII (Figure 4.2). These subtelomeres showed striking length differences in the *S. uvarum* and *S. eubayanus* species compared to the rest of the species. Hierarchical clustering analysis of the subtelomeric lengths of the 32 subtelomeres confirmed our observations by dividing them into two major groups. One group included the closely related species *S. eubayanus* and *S. uvarum* and the other the rest of the species and strains sharing no clear pattern (Figure 4.4A).

Subtelomeric gene density expressed in terms of the number of annotated genes located at the subtelomeric regions was calculated. High variability was also observed at both the subtelomere and the species level. Values ranged from 0 genes to 24 genes with a median of 6 genes per subtelomere. A positive correlation between subtelomeric lengths and subtelomeric gene density was observed (Spearman correlation coefficient = 0.73, *p*-value = 4.74e-102). Subtelomeric gene densities showed high variability across subtelomeres and species level (Figure 4.3) as expected due to the correlation observed with the subtelomeric lengths. A hierarchical clustering confirmed the differentiation of the early-divergent clade of *S. uvarum* and *S. eubayanus* with respect to the rest of the *Saccharomyces* species (Figure 4.4B).

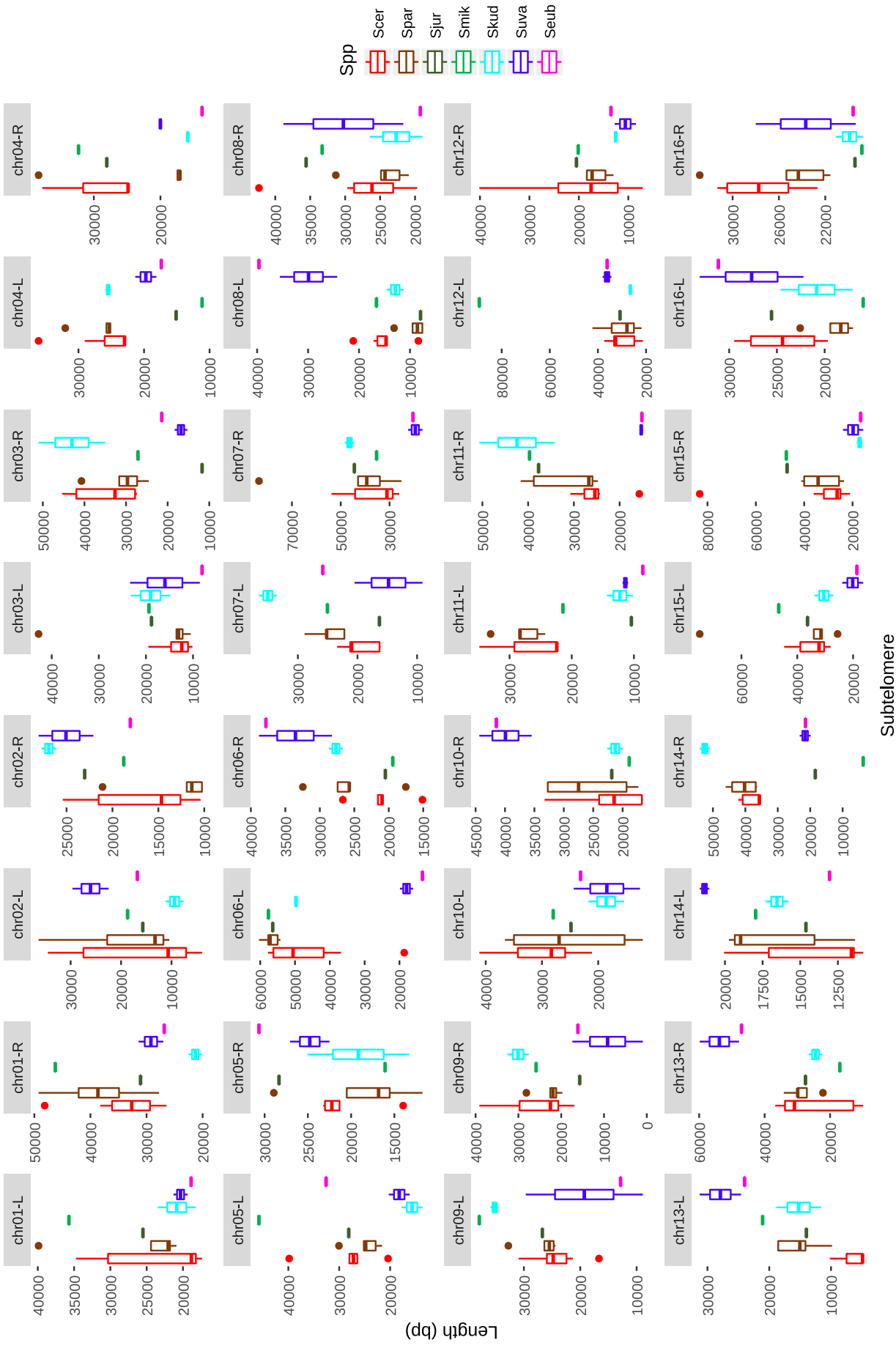


Figure 4.2: Subtelomeric lengths in *Saccharomyces* species. Subtelomeric lengths represented in base-pairs and calculated from subtelomeric gene boundaries. Number of chrosomosome is represented according to *S. cerevisiae* S288C. All large chromosomal translocation were taken into account. Subtelomere left arm (L), subtelomere right arm (R). Legend: Spp: Species; Scer: *S.cerevisiae*; Spar: *S. paradoxus*; Sjur: *S. jurei*; Smik: *S. mikatae*; Skud: *S. kudriavzevii*; Suva: *S. uvarum*; Seub: *S. eubayanus*.

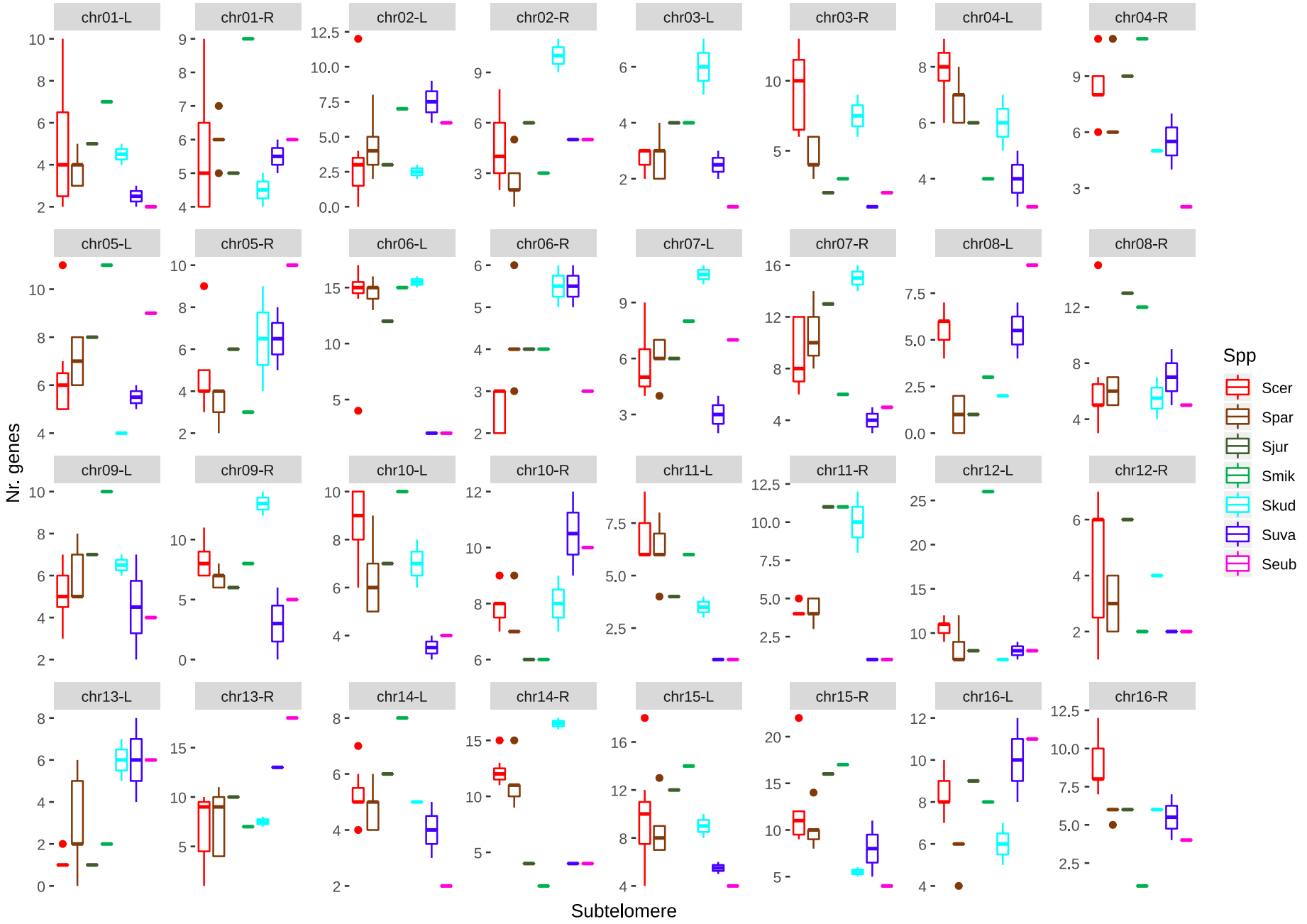


Figure 4.3: Subtelomeric gene densities in *Saccharomyces* species. Subtelomeric gene densities values represented in number of annotated ORFs in every subtelomere defined by subtelomeric gene boundaries. Number of chrosomosome is represented according to *S. cerevisiae* S288C. All large chromosomal translocation were taken into account. Subtelomere left arm (L), subtelomere right arm (R). Legend: Spp: Species; Scer: *S.cerevisiae*; Spar: *S. paradoxus*; Sjur: *S. jurei*; Smik: *S. mikatae*; Skud: *S. kudriavzevii*; Suva: *S. uvarum*; Seub: *S. eubayanus*.

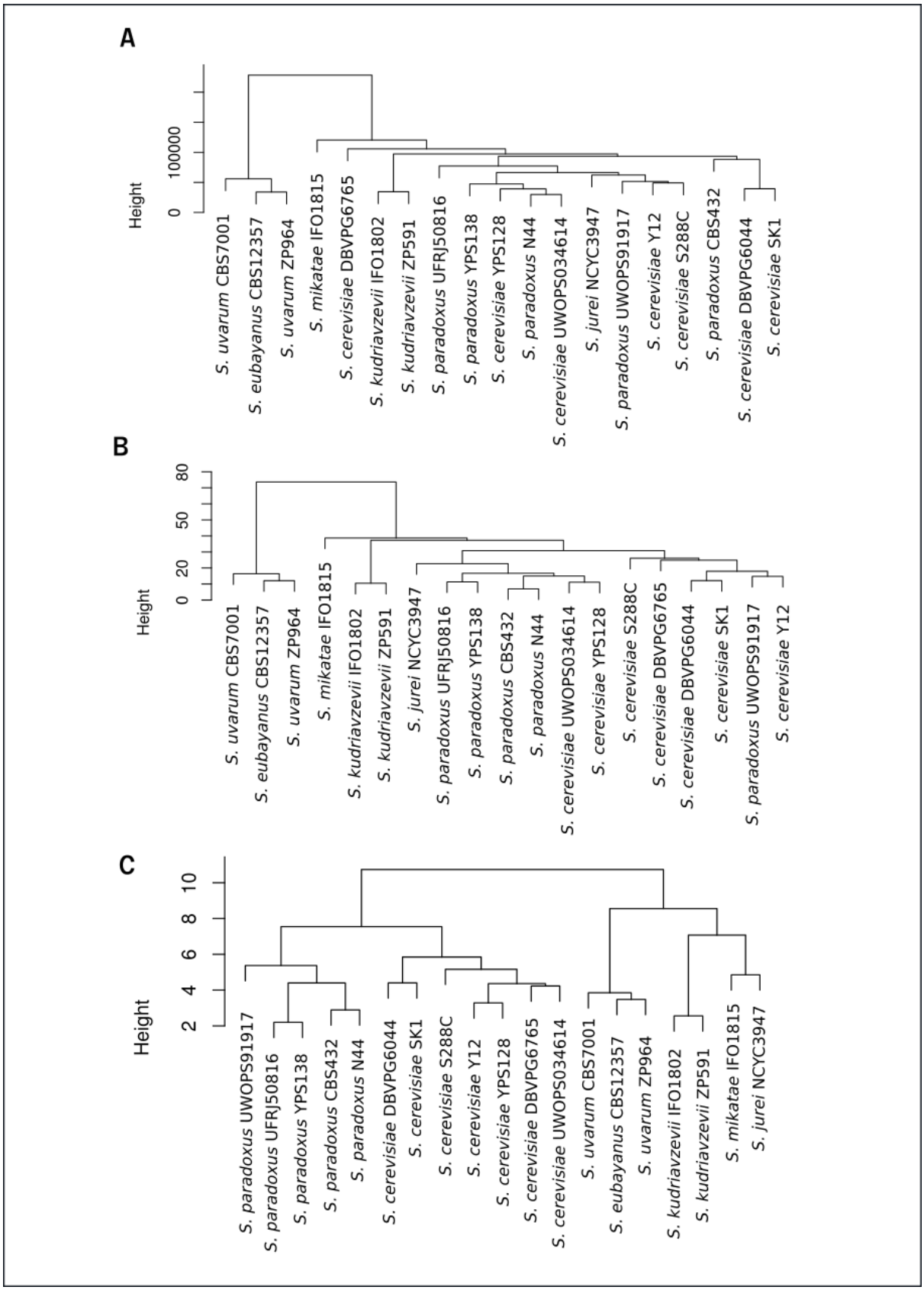


Figure 4.4: Hierarchical clustering dendrogram. Subtelomeric lengths (A). Subtelomeric gene density (B) and subtelomeric protein families copy number (C).

4.3.3 Expansion of subtelomeric protein-coding gene families is responsible for a species pattern differentiation

Subtelomeric gene families were obtained using a clustering method based on protein sequence similarities. A total of 190 clusters/families were determined of which 33 were singletons. The mean size of the clusters was 19 (singletons excluded), the same as the number of the input genomes. Subtelomeric gene families were functionally annotated, and the largest cluster consisted of 245 proteins of the seripauperin family, a multigene family located mainly in subtelomeric regions (Viswanathan et al., 1994). The ten largest subtelomeric gene families and their annotations are shown in Table 4.2, they mainly correspond to different membrane transporters, such as amino acid, hexose, and metal transporters.

Table 4.2: protein-coding gene families.

Cluster size	Functional annotation
245	Seripauperins
185	Hexose transporters
171	DUP380 subfamily of conserved genes
155	Carbohydratases
108	DNA-binding proteins; transcription factors
106	*FLO* genes
95	Aminoacid transporters
83	Cell wall; mannoprotein cell wall signaling transduction
82	Ferric transporter; gluthation exchange
81	4-amino-5-hydroxymethyl-2-methylpyrimidine phosphate synthase

Differences in copy numbers across subtelomeres of each protein family revealed species differentiation (Figure 4.4C). Two main groups were observed, one including the *S. cerevisiae* and *S. paradoxus* clade and the other the rest of species. Within these two groups, differences by species are also observed. Subtelomeric protein-encoding gene families responsible for species differentiation were further investigated. The family functionally annotated as endo-polygalacturonases showed striking copy-number differences among species. This family contains the *PGU1* gene,

a gene located at the right subtelomere of chromosome X (according to the ancestral chromosome identity based on the S288c strain). All strains of *Saccharomyces* species have the homologous and syntenic copy of the *PGU1* gene, but some species show extra copies of this endo-polygalacturonase. Thus, the closely related *S. mikatae* and *S. jurei* species possess an additional copy located at the right subtelomere of chromosome VIII. The early-divergent *S. eubayanus* and S. *uvarum* species showed also extra-copies of this gene but located at different positions. The phylogenetic tree of the protein family revealed that the extra copies of the *S. uvarum* and *S. eubayanus* strains were not monophyletic (Figure 4.5). The S-adenosyl-methionine protein families showed one extra-copy in both *S. uvarum* and *S. eubayanus* as well. Other protein families showing expansion in the early-divergent species were the methyltransferases, quinone oxidoreductases, an unknown protein family probably involved in the interaction with ribosomes, aldehyde-reductases involved in furfural detoxification, and a plasma membrane Mg^{+2} transporter. Other clades and species also exhibited differences in the expansion of subtelomeric families. In *S. kudriavzevii*, we observed the expansion of the mannitol dehydrogenase family and an integral membrane protein family. In the case of the *S. jurei–S. mikatae* clade, differences in the high-affinity cysteine-specific transporter and the nuclear transport factor 2 families were observed. Some small subtelomeric gene families are unique for particular *Saccharomyces* species. This is the case of the mannose-6-phosphate receptors, cyanamide hydratases, and enolases families, which are only found in the subtelomeres of the *S. cerevisiae* and *S. paradoxus* species.

4.3.4 Core genes have roles in basic cellular maintenance functions while species-specific genes have a wide range of functions.

Using orthology and synteny, we defined the set of core genes of the *Saccharomyces* genus. This set includes 4950 genes (Figure 4.6) in a relation

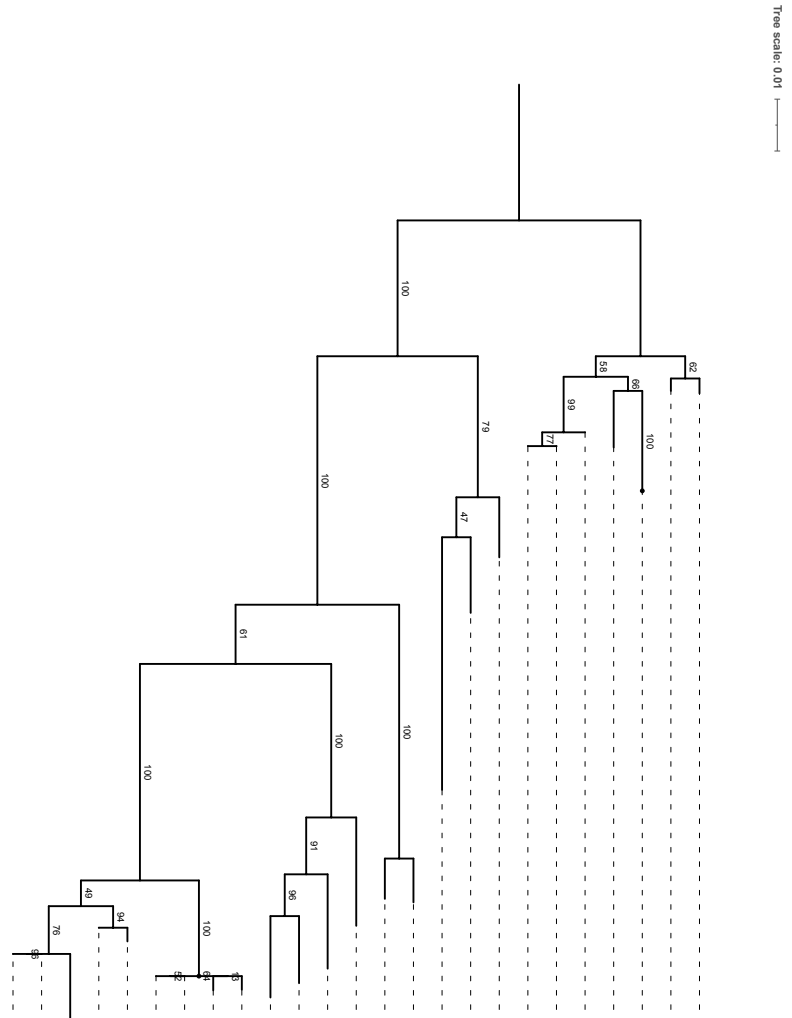


Figure 4.5: Phylogeny of the *PGU*-coding gene family. Translated proteins of the *PGU1* family were aligned with Mafft (Katoh and Standley, 2013) and back-translated into codons. Maximum-likelihood phylogeny reconstruction was obtained using RAxML (Stamatakis, 2014) with the GTR-Γ model and 100 bootstrap replicates. The tree was rooted in the *S. eubayanus*–*S. uvarum PGU* divergent genes leading branch, and visualized using iTOL (Letunic and Bork, 2016).

of one-to-one orthologous genes conserving the synteny among the 19 annotated genome assemblies. We used the 4950 genes for phylogeny reconstruction (Figure 4.7). GO-term enrichment analysis of the core genes revealed biological processes related to cell cycle, such as cellular homeostasis and mitotic and meiotic cell cycles, together with other essential functions for the cell such as Golgi vesicle transport or endocytosis.

Variable genes were further investigated focusing on *S. mikatae*, *S. kudriavzevii*, and *S. uvarum* genes, the species sequenced in this study. *S. uvarum* showed sixteen variable genes present only in this species, such as a *CYB2*-related gene located in chromosome X. *CYB2* encodes the L-lactate cytochrome-c oxidoreductase (Lodi and Guiard, 1991), a component of the mitochondria required for lactate utilization. A set of 84 variable genes are shared between the closely-related *S. uvarum* and *S. eubayanus*. Most of those genes are representatives of a series of ancestral genes that were conserved in *S. uvarum – S. eubayanus* lineage and lost after the divergence of the ancestor of the remaining species. A homologous gene to *YAT1* was found in chromosome IV. *YAT1* codifies for an outer mitochondrial carnitine acetyltransferase located at chromosome I. A *YAT1*-related gene was identified in *S. eubayanus* and the European *S. uvarum* CBS7001. Strickingly, this gene shows signals of pseudogenization in the Australasian *S. uvarum* ZP964 strain. Other genes identified as unique in this clade were homologous to *PEP1*, a transmembrane sorting receptor for vacuole hydrolases; *HXT10*, a hexose transporter; *GAL2*, a galactose permease; *PMT4*, a mannosyltransferase; *FRE1*, a ferric/cupric reductase; *ARL1*, a soluble GTPase with a role in the regulation of the membrane trafficking. *S. kudriavzevii* showed 48 species-specific genes. A gene homologous to *LAC1*, that encodes a ceramide synthase component, is located in chromosome III. *S. mikatae* shares genes with its sister species *S. jurei*, such as a *FEX1*-related gene located in chromosome IV in both species. *FEX1* encodes a fluoride transporter. Two tandem duplications of the genes *PHO3* and *ARA1*, encoding a phosphatase and an $NADP^+$-dependent arabinose

Figure 4.6: *Saccharomyces* pangenome. *Saccharomyces* pangenome is showed in the center. It was obtained by using the intersection of all the annotated genes in each *Saccharomyces* species.

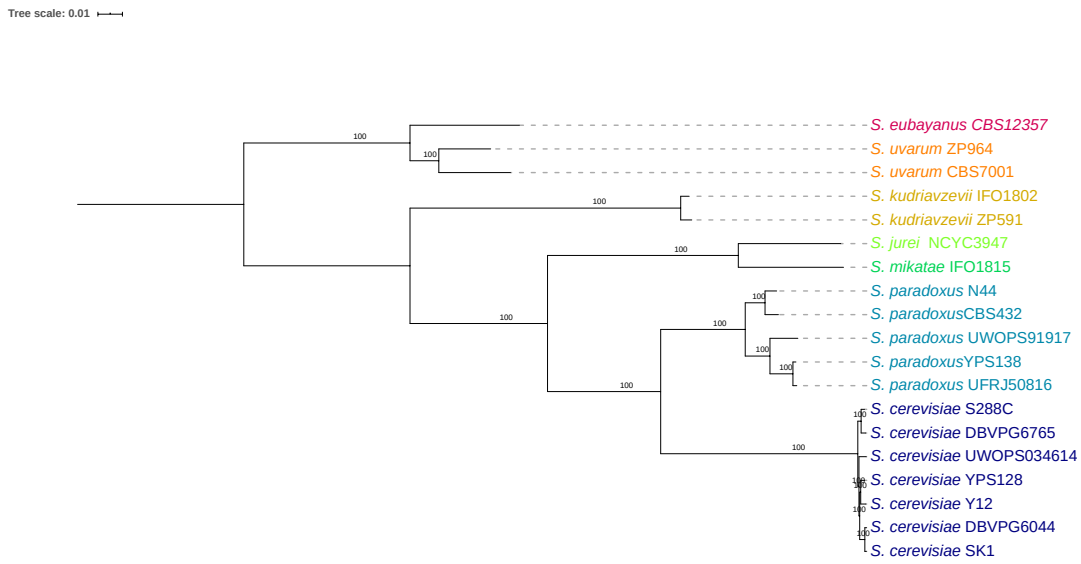


Figure 4.7: *Saccharomyces* phylogeny. Phylogeny reconstruction from the concatenated alignment of the 4950 core genes. Bootstrap support values showed range from 0 to 100.

dehydrogenase, respectively, were found in both species located in chromosome II.

4.3.5 Thiamine biosynthesis and maltose metabolism genes show variations in their genome positions.

The *S. cerevisiae* reference assembly (R.64, strain S288c) was used to study the patterns of differential synteny conservation among species. Subtelomeric regions showed the highest variations in synteny, as expected. As examples, several genes of the subtelomeric families of thiamine biosynthesis and maltose metabolism showed a different location in the subtelomeres of the species *S. eubayanus, S. uvarum, S. mikatae*, and *S. kudriavzevii*, with respect to *S. cerevisiae*. Differences in synteny conservation of single genes were also observed. *FLR1*, a plasma membrane drug transporter changed its location in *S. eubayanus, S. uvarum*, and *S. kudriavzevii*. In *S. eubayanus*, *S. uvarum* is located in their chromosome VIIItXVI while *S. kudriavzevii* exhibits two copies of this gene: one in chromosome XVI and another in chromosome II. The rest of species only maintained the copy located in chromosome II. There

are examples of genes that have undergone a process of pseudogenization during *Saccharomyces* evolution. One example is the *YAT1*-related gene in the Australasian *S. uvarum* strain, as mentioned. Another particular case is the gene *FAA3*, encoding a long-chain fatty acyl-CoA synthetase, in the sister species *S. uvarum* and *S. eubayanus*, in which a hexose transporter gene, homologous to *HXT10*, replaced the *FAA3* ortholog.

4.4 Discussion

The *Saccharomyces* genus is turning into a very attractive model for evolutionary biologists; however, when it comes to available sequences, there is an important lack of non–*S. cerevisiae* high – quality genome references. In this work, we used a combination of long-read and short-read sequencing technologies to obtain five new *Saccharomyces* end-to-end genome assemblies. Our work provides newly updated genome references of *S. kudriavzevii, S. mikatae* and *S. uvarum*. The high variability observed at the subtelomeric level when lengths and gene densities are compared, within and between species, illustrates the extraordinary instability of eukaryotic subtelomeres. Despite the variability observed, a clear pattern could be distinguished from our data that differentiates the subtelomeric evolution dynamics of the early-divergent *S.uvarum – S. eubayanus* clade from the rest of *Saccharomyces* species. These two species have been shown very similar population patterns (Almeida et al., 2014; Peris et al., 2016) with populations spread out along the same locations on Earth except for *S. eubayanus* which has not been found in Europe yet. These species are cryotolerant, showing significantly lower optimal growth temperatures, especially when compared to the *S. cerevisiae – S. paradoxus* clade. They seem to have preferences in particular environments as they have been mostly isolated from beeches (*Nothofagus*) and *Araucaria* trees in the Southern hemisphere (Almeida et al., 2014; Gonzalez Flores et al., 2020; Libkind et al., 2011), while wild *Saccharomyces* populations from the Northern hemisphere are widely associated with oaks. Our analysis allowed us to identify the largest protein families located nearby telomeric regions. These families were mainly involved in niche-specific processes like carbohydrate uptake and metabolism, stress response, and cell membrane transport. These observations support the idea of the important role that subtelomeric genes play in the rapid adaptation to novel niches. Expansion of subtelomeric protein families in particular clades have been also studied. We have focused on the possible expansion

of certain families in non-*S. cerevisiae* species. An interesting example was the gene family encoding proteins with endo-polygalacturonase activity. The gene *PGU1* is conserved at the right subtelomere of chromosome X in all *Saccharomyces* species. On one hand, *S. jurei* and *S. mikatae* showed an extra copy of this gene at the left arm of chromosome VIII which could be explained by one small-scale duplication event at the common ancestor of both species. On the other hand, *S. uvarum* and *S. eubayanus* showed expansion of this family as well. Besides the *PGU1* homolog and syntenic gene, a divergent copy of this gene was found in a different branch (Figure 4.5). The presence of this divergent copy in the common ancestor of *Saccharomyces* followed by a loss after *S. uvarum – S. eubayanus* divergence could be the most likely hypothesis. It is interesting how this divergent pectinase is conserved with two (ZP964) and three copies (CBS7001 and CBS12375). The pectinase activity is an important trait in yeasts fermenting plant substrates and the conservation of this genotype in the early-divergent clade together with the latter duplication event in *S. jurei–S. mikatae* clade shows its potential role in niche colonization as these species have been mostly isolated from tree barks. Besides, *S. uvarum–S. eubayanus* have been characterised by a high pectinolytic activity suggesting that the divergent copies of *PGU* genes could be specialised in a different role than the *PGU1* (Naumov et al., 2016). *S. kudriavzevii* showed expansion of the mannitol dehydrogenase family. Mannitol dehydrogenases are involved in mannitol biosynthesis. The mannitol is a natural acyclic polyol with an important role in stress tolerance because is accumulated at the cytosol and prevents the inactivation of metabolic processes. High-mannitol-producing yeast strains have been isolated from fermentation sludge (Song et al., 2002). *S. kudriavzevii* have been isolated from soil and decayed leaves (Naumov et al., 2000a), therefore, the expansion of this subtelomeric family suggest an important role in the adaptation of *S. kudriavzevii* to these environments. Species-specific genes might be very important for species adaptation as well. The origin of these genes could be lying on small–scale duplications. The extra copies generated by small–scale duplications

can acquire new functions (neofunctionalization) (Ohno, 1970) without being affected by purifying selection. An alternative is that gene duplication leads to asymmetric evolution of both preexisting and duplicated genes so that these functions can be optimised (subfunctionalization) (Force et al., 1999). Another plausible hypothesis to explain the presence of species-specific genes is that they could correspond to ancestral genes present in the common ancestor that were lost during the divergence of the other species. This seems to be more frequent in *S. uvarum–S. eubayanus* as they belong to the first clade to diverge and single loss events can explain the absence of this clade-specific genes in the other species. Finally, *de novo* gene birth is also an option that seems more prevalent than expected (Carvunis et al., 2012), although it is not considered in this study. We have observed interesting species-specific genes related to metabolism, such as a *CYB2* homolog in *S. uvarum*. This gene is responsible for the oxidation of lactate, a non-fermentable source, into pyruvate (Guiard, 1985). It is well known that sugars are the preferred carbon sources for *S. cerevisiae* but little is known about the assimilation of non–fermentable carbon sources in other *Saccharomyces* species. Another example is a *YAT1* homolog, which was found in *S. uvarum* and *S. eubayanus*, but in a process of pseudogenization in the Australasian *S. uvarum* ZP964 strain. *YAT1* is essential for ethanol metabolism acting as a transmembrane transporter of the activated acyl groups from the cytoplasm into the mitochondrial matrix. Further investigation will be needed to identify the function of this gene and to determine if its loss in the Australasian lineage is strain- or lineage-specific. A global view into the subtelomeric evolution dynamics together with annotation insights is provided in this study. Further research must be necessary to elucidate the role of the observed subtelomeric gene variation in the environmental adaptation of the *Saccharomyces* species.

General Discussion

The tremendous phenotypic diversity of the *Saccharomyces* species and its potential application in biotechnological processes have attracted researchers´ attention during the last decade. Understanding the main adaptive mechanisms underlying phenotypic diversity is a key question in modern biology. There are several molecular mechanisms involved in the adaptive processes of *Saccharomyces* yeasts. This doctoral thesis aimed to investigate different molecular adaptive mechanisms in *S. uvarum* and *S. kudriavzevii* species by using comparative genomics methods.

In the first chapter, we attempted to shed light on the main existing genomic differences between *S. kudriavzevii* and *S. cerevisiae*. At the time this study was carried out, our main concern was the availability of high-quality annotated genome references of *S. kudriavzevii*. Therefore, we decided to sequence two new strains and annotate both the newly and previously sequenced strains using our annotation pipeline. Despite our efforts, our study relied on a small genome dataset, mainly due to the lack of *S. kudriavzevii* sequences. However, we could assess some general insights by using different approaches. At the nucleotide level, the analysis of positive selection signatures under the branch-site model assumption revealed 30 and 32 genes under positive selection for *S. cerevisiae* and *S. kudriavzevii*, respectively. At the same time, Tajima's rate test showed 78 and 191. The intersection of the

branch-site analysis results for the two clades showed genes *FRT2* and *RQC2* as being under positive selection in both species. We observed that approximately half of the protein dataset showed evidence of functional divergence at the protein level. Interestingly, when reviewing the functional divergence by metabolic pathways, we observed that the most functionally divergent pathways between both species were 'osmotic and oxidative stress response' and 'sphingolipid metabolic pathway'. From this comprehensive analysis, we could conclude that the gene *FBA1*, encoding a fructose 1,6-biphosphate aldolase, is an excellent candidate to explore its role in metabolic differences between the two species. In specific, the nucleotide sequences analysis revealed a positive selection in the *S. kudriavzevii FBA1* gene and an acceleration of its evolutionary rates in both species. This gene acts at the glycolysis branching point where the trioses can be directed either to the ethanol fermentation or the glycerol synthesis. A higher glycerol-3-phosphate dehydrogenase activity and an increased glycerol production have been reported in *S. kudriavzevii* compared to *S. cerevisiae* (Arroyo-López et al., 2010a; Oliveira et al., 2014). The glycerol synthesis is essential in yeast metabolism because it is involved in osmoregulation (Ansell et al., 1997; Nevoigt and Stahl, 1997). This compound seems to play a critical role in the low-temperature-tolerance of cryotolerant yeasts (Izawa et al., 2004). Accelerated evolutionary rates at both branches could indicate a high divergence of this gene in both *Saccharomyces* species. Moreover, the positive selection detected in the *S. kudriavzevii* branch could clearly be attributed to the action of natural selection. Most of the riboflavin pathway encoding genes displayed accelerated evolutionary rates, and the corresponding proteins showed functional divergence. According to a systems biology approach, this pathway was demonstrated to be affected by cold-temperature (Paget et al., 2014). Functionally divergent riboflavin proteins with improved activities at low-temperatures could be possible.

Given the variability obtained using the three methods, we decided to focus on the positive selection method for our next study. Furthermore, we strived to

increase the statistical power of the testing by expanding our dataset. To achieved this, we developed a tool to facilitate the positive selection analysis using different annotated genomes and with the possibility of running other nested-models besides the branch-site assumption. GwideCodeML package was successfully implemented and applied to our previous *S. kudriavzevii* – *S. cerevisiae* analysis increasing the number of detected genes. We also used GWideCodeML to detect positive selection signatures, using the three implemented models, in the *S. uvarum* clade. In the case of the branch-site model, many genes related to the cell wall, chemical homeostasis, and those encoding the general amino acid permease *GAP1* and the hexose transporter *HXT6*, with demonstrated environmental adaptation roles, were revealed. An enrichment in ohnologues was found among the positively selected genes emphasizing gene duplication as a valuable mechanism generating functional novelties. The site-model revealed exciting results with an enrichment in genes related to translation and ribosome assembly. These findings are concordant with previous transcriptomic data suggesting translation efficiency as a critical target of adaptive evolution to face changing environments (Tronchoni et al., 2014). Additionally, the site-model results were also enriched in the glucose fermentation pathway genes where three out of the five genes were duplicates. Species diverging after the WGD showed increased glycolytic fluxes (Conant and Wolfe, 2007). The detected genes could be contributing to this effect where particular amino acid positions have been differentially fixed in the Saccharomycotina species. The different nested-models jointly pointed out cell wall mannoprotein encoding genes as positively selected genes. Finally, it is worth to mention the *FBA1* gene, in which codon positions different to those detected in Chapter 1, were under positive selection according to the site-model.

In Chapter 3, we explored the *S. uvarum* genomes to find genomic footprints of domestication. We found two different chromosomal rearrangements in the promoter region of the *SSU1* gene. The strains carrying these rearrangements showed a higher sulphite tolerance than the strains with the ancestral *SSU1* promoter version. This

evolved property is highly important from an industrial point of view as sulphite is commonly used as a preservative in wine and cider fermentations. We demonstrated that the chromosomal rearrangements were responsible for the overexpression of the *SSU1* gene. Furthermore, this effect does not seem to be constitutive and independent of the sulphite presence in the media. The origin of the most common chromosomal translocation (VII^{XVI}) was studied using additional *S. uvarum* genomic data. We concluded that the recombination event probably originated in the common ancestor of all of them. Moreover, our data confirmed that the Argentinean strains are probably admixed strains between European and South American *S. uvarum*, as suggested by previous population genomics studies (Almeida et al., 2014; Gonzalez Flores et al., 2020). Previous chromosomal rearrangements have been described in *S. cerevisiae* wine strains causing the same phenotypic outcome, an increased sulphite tolerance (García-Ríos et al., 2019; Pérez-Ortín et al., 2002; Zimmer et al., 2014). This way, our work provides an excellent example of convergent adaptation in the two S*accharomyces* species associated to anthropic environments, and our discovery highlights the *SSU1* gene as an important selection target in the adaptive evolution of domesticated yeasts.

In the last chapter of this thesis, we used long-read sequencing combined with short-read sequencing data to obtain end-to-end high-quality genome assemblies of *S. kudriavzevii, S. mikatae*, and *S. uvarum*. We combined our data with previously published data of the other species of the genus to perform a comparative genomics study of all *Saccharomyces* species. We thoroughly annotated all the assemblies to obtain a pangenome of the genus. The core genes corresponded to those genes shared among all the species in a relation of 1 to 1. Core gene functional annotation revealed cellular homeostasis and mitotic and meiotic cell cycle functions, whereas the species-specific genes were impoverished on those functions. We performed a comprehensive study of the subtelomeric regions revealing high intra- and interspecific variabilities. We could observe some patterns that differentiate species in two major

groups: one containing *S. cerevisiae* and *S. paradoxus* and the other having the rest of *Saccharomyces* species. We identified the protein families located at the subtelomeres. The largest families were mainly involved in carbohydrate metabolism, nutrient uptake, and stress response. These functions are considered niche-specific supporting the key role of subtelomeric regions and their high plasticity in the adaptive mechanisms. We identified the expansion of certain subtelomeric protein families in our species of interest. *S. uvarum*, together with its sister species *S. eubayanus*, showed the expansion of a family of endo-polygalacturonases with high pectinase activity (Naumov et al., 2016). This is an important trait in yeast fermenting plant substrates, and it could also be important for our *Saccharomyces* species according to their ecology. *S. kudriavzevii* showed the expansion of a mannitol dehydrogenases family that could be involved in their adaptation to grow in the soil. In this study, we provided a global view of the subtelomere evolutionary dynamics in the *Saccharomyces* genus. However, further research needs to be done to correlate subtelomere variation with the ecological adaptation of yeast.

Conclusions

1) Approximately half of the analysed proteins encoded in the *S.kudriavzevii* and *S. cerevisiae* genomes showed evidence of functional divergence. A high proportion of those proteins were involved in oxidative stress responses and sphingolipid metabolic pathways.
2) The riboflavin metabolism-related genes are among the most functionally divergent proteins in *S. kudriavzevii*. The encoding genes also showed accelerated evolutionary rates highlighting the differences observed between *S. kudriavzevii* and *S. cerevisiae*.
3) *FRT2* (protein of the endoplasmic reticulum membrane) and *RQC2* (component of the ribosome quality control) genes showed positive selection signatures at both *S. kudriavzevii* and *S. cerevisiae* branches.
4) The use of additional sequences to detect positive selection signatures increases the number of detected genes in a genome-wide approach.
5) The genes under positive selection in the *S. uvarum* leading branch are enriched in ohnologues.
6) Various cell wall mannoprotein-encoding genes showed positive selection signatures using different nested-models suggesting a potential adaptive role in *S. uvarum*, as well as in the rest of the species included in the analysis.

7) The *FBA1* gene, encoding the fructose 1,6-bisphosphate aldolase, was found under positive selection in independent comparative genomics studies, suggesting an important role in the glycerol production in the cryotolerant species.
8) *S. uvarum* strains isolated from human-driven fermentations showed genomic footprints of domestication.
9) The chromosomal rearrangements (VII^{XVI} and XI^{XVI}) identified at the *SSU1* promoter region cause overexpression of the *SSU1* gene that is not dependent on the sulphite presence in the media.
10) The chromosomal rearrangements confer a selective advantage to the *S. uvarum* strains growing in fermentations where sulphite is used as a preservative.
11) A single origin of the VII^{XVI} chromosomal rearrangement occurred in the common ancestor of the holarctic species. This recombination was maintained in some domesticated populations by selective pressure.
12) The *SSU1* promoter region is a hotspot of evolution, causing the *S. uvarum* and *S. cerevisiae*'s convergent adaptation to growing in sulphite containing environments.
13) The *Saccharomyces* pangenome's core genes encode proteins related to cellular homeostasis and cell cycle, whereas the species-specific genes have a wide range of molecular functions.
14) The subtelomeric regions display a tremendous variability at both lengths and gene density. Despite this variability, some patterns could be distinguished for species separation in two major groups: one containing *S. cerevisiae* and *S. paradoxus* and the other one the rest of the species.
15) The largest subtelomeric protein families play niche-specific functions such as nutrient uptake and carbohydrate metabolism.
16) The expansion of particular subtelomeric protein families following a species pattern is observed. *S. uvarum* shows an increase in the endo-polygalacturonases family, while *S. kudriavzevii* shows an increase

in the mannitol dehydrogenases.

Bibliography

Aa, E., Townsend, J. P., Adams, R. I., Nielsen, K. M. and Taylor, J. W. (2006), 'Population structure and gene evolution in *Saccharomyces cerevisiae*', *FEMS Yeast Research* **6**(5), 702–715.

Abe, F. (2007), 'Induction of DAN/TIR yeast cell wall mannoprotein genes in response to high hydrostatic pressure and low temperature', *FEBS Letters* **581**(25), 4993–4998.

Aickin, M. and Gensler, H. (1996), 'Adjusting for multiple testing when reporting research results: the Bonferroni vs Holm methods.', *American journal of public health* **86**(5), 726–8.

Almeida, P., Gonçalves, C., Teixeira, S., Libkind, D., Bontrager, M., Masneuf-Pomarède, I., Albertin, W., Durrens, P., Sherman, D. J., Marullo, P., Hittinger, C. T., Gonçalves, P. and Sampaio, J. P. (2014), 'A Gondwanan imprint on global diversity and domestication of wine and cider yeast *Saccharomyces uvarum*.', *Nature communications* **5**(May), 4044.

Alonso-del Real, J., Contreras-Ruiz, A., Castiglioni, G. L., Barrio, E. and Querol, A. (2017), 'The use of mixed populations of *Saccharomyces cerevisiae* and *S. kudriavzevii* to reduce ethanol content in wine: Limited Aeration, inoculum proportions, and sequential inoculation', *Frontiers in Microbiology* **8**, 2087.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990), 'Basic local alignment search tool', *Journal of Molecular Biology* **215**(3), 403–410.

Angiolo, M. D., Chiara, M. D., Yue, J.-x., Irizar, A., Stenberg, S., Llored, A., Barré, B., Schacherer, J., Marangoni, R. and Gilson, E. (2020), 'A yeast living fossil reveals the origin of genomic introgressions', *Nature* .

Anisimova, M. and Yang, Z. (2007), 'Multiple hypothesis testing to detect lineages under positive selection that affects only a few sites', *Molecular Biology and Evolution* **24**(5), 1219–1228.

Ansell, R., Granath, K., Hohmann, S., Thevelein, J. M. and Adler, L. (1997), 'The two isoenzymes for yeast NAD+-dependent glycerol 3-phosphate dehydrogenase encoded by *GPD1* and *GPD2* have distinct roles in osmoadaptation and redox regulation', *EMBO Journal* **16**(9), 2179–2187.

Arroyo-López, F. N., Orlić, S., Querol, A. and Barrio, E. (2009), 'Effects of temperature, pH and sugar concentration on the growth parameters of *Saccharomyces cerevisiae*, *S. kudriavzevii* and their interspecific hybrid', *International Journal of Food Microbiology* **131**(2-3), 120–127.

Arroyo-López, F. N., Pérez-Torrado, R., Querol, A. and Barrio, E. (2010a), 'Modulation of the glycerol and ethanol syntheses in the yeast *Saccharomyces kudriavzevii* differs from that exhibited by *Saccharomyces cerevisiae* and their hybrid', *Food Microbiology* **27**(5), 628–637.

Arroyo-López, F. N., Salvadó, Z., Tronchoni, J., Guillamón, J. M., Barrio, E. and Querol, A. (2010b), 'Susceptibility and resistance to ethanol in *Saccharomyces* strains isolated from wild and fermentative environments', *Yeast* **27**(12), 1005–1015.

Avram, D. and Bakalinsky, A. T. (1997), '*SSU1* encodes a plasma membrane protein with a central role in a network of proteins conferring sulfite tolerance in *Saccharomyces cerevisiae*', *Journal of Bacteriology* **179**(18), 5971–5974.

Avram, D., Leid, M. and Bakalinsky, A. T. (1999), 'Fzf1p of *Saccharomyces cerevisiae* is a positive regulator of *SSU1* transcription and its first zinc finger region is required for DNA binding', *Yeast* **15**(6), 473–480.

Bakalinsky, A. T. and Snow, R. (1990), 'The chromosomal constitution of wine strains of *Saccharomyces cerevisiae*', *Yeast* **6**(5), 367–382.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A. and Pevzner, P. A. (2012), 'SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing.', *Journal of computational biology : a journal of computational molecular cell biology* **19**(5), 455–77.

Barton, A. B., Pekosz, M. R., Kurvathi, R. S. and Kaback, D. B. (2008), 'Meiotic recombination at the ends of chromosomes in *Saccharomyces cerevisiae*', *Genetics* **179**(3), 1221–1235.

Bauer, F. and Pretorius, I. (2000), 'Yeast stress response and fermentation efficiency: how to survive the making of wine - a review', *South African Journal of Enology & Viticulture* **21**(1), 27–51.

Becker-Kettern, J., Paczia, N., Conrotte, J. F., Kay, D. P., Guignard, C., Jung, P. P. and Linster, C. L. (2016), '*Saccharomyces cerevisiae* forms D-2-hydroxyglutarate and couples its degradation to D-Lactate formation via a cytosolic transhydrogenase', *Journal of Biological Chemistry* **291**(12), 6036–6058.

Bell, M. A. and Travis, M. P. (2005), 'Hybridization, transgressive segregation, genetic covariation, and adaptive radiation'.

Belloch, C., Orlic, S., Barrio, E. and Querol, A. (2008), 'Fermentative stress adaptation of hybrids within the *Saccharomyces* sensu stricto complex', *International Journal of Food Microbiology* **122**(1-2), 188–195.

Bely, M., Rinaldi, A. and Dubourdieu, D. (2003), 'Influence of assimilable nitrogen on volatile acidity production by *Saccharomyces cerevisiae* during high sugar fermentation', *Journal of Bioscience and Bioengineering* **96**(6), 507–512.

Bing, J., Han, P. J., Liu, W. Q., Wang, Q. M. and Bai, F. Y. (2014), 'Evidence for a far east asian origin of lager beer yeast'.

Blateyron, L. and Sablayrolles, J. M. (2001), 'Stuck and slow fermentations in enology: Statistical study of causes and effectiveness of combined additions of oxygen and diammonium phosphate', *Journal of Bioscience and Bioengineering* **91**(2), 184–189.

Borneman, A. R., Forgan, A. H., Kolouchova, R., Fraser, J. A. and Schmidt, S. A. (2016), 'Whole genome comparison reveals high levels of inbreeding and strain redundancy across the spectrum of commercial wine strains of *Saccharomyces cerevisiae*', *G3: Genes, Genomes, Genetics* **6**(4), 957–971.

Borneman, A. R., Schmidt, S. A. and Pretorius, I. S. (2013), 'At the cutting-edge of grape and wine biotechnology'.

Boynton, P. J. and Greig, D. (2014), 'The ecology and evolution of non-domesticated *Saccharomyces* species', *Yeast* **31**(12), 449–462.

Brandman, O., Stewart-Ornstein, J., Wong, D., Larson, A., Williams, C. C., Li, G.-W., Zhou, S., King, D., Shen, P. S., Weibezahn, J., Dunn, J. G., Rouskin, S., Inada, T., Frost, A. and Weissman, J. S. (2012), 'A ribosome-bound quality control complex triggers degradation of nascent peptides and signals translation stress.', *Cell* **151**(5), 1042–54.

Brickwedde, A., Brouwers, N., van den Broek, M., Gallego Murillo, J. S., Fraiture, J. L., Pronk, J. T. and Daran, J. M. G. (2018), 'Structural, physiological and regulatory analysis of maltose transporter genes in *Saccharomyces eubayanus* CBS 12357T', *Frontiers in Microbiology* **9**(AUG), 1786.

Brion, C., Ambroset, C., Sanchez, I., Legras, J. L. and Blondin, B. (2013), 'Differential adaptation to multi-stressed conditions of wine fermentation revealed by variations in yeast regulatory networks', *BMC Genomics* **14**(1), 681.

Brown, C. A., Murray, A. W. and Verstrepen, K. J. (2010), 'Rapid Expansion and Functional Divergence of Subtelomeric Gene Families in Yeasts', *Current Biology* **20**(10), 895–903.

Burri, L. and Lithgow, T. (2004), 'A complete set of SNAREs in yeast.', *Traffic* **5**(1), 45–52.

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 Research* **15**(10), 1456–1461.

Čadež, N., Bellora, N., Ulloa, R., Hittinger, C. T. and Libkind, D. (2019), 'Genomic content of a novel yeast species *Hanseniaspora gamundiae* sp. Nov. From fungal stromata (*Cyttaria*) associated with a unique fermented beverage in Andean Patagonia, Argentina', *PLoS ONE* **14**(1).

Capella-Gutiérrez, S., Silla-Martínez, J. M. and Gabaldón, T. (2009), 'trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses', *Bioinformatics* **25**(15), 1972–1973.

Carrasco, P., Querol, A. and Del Olmo, M. (2001), 'Analysis of the stress resistance of commercial wine yeast strains', *Archives of Microbiology* **175**(6), 450–457.

Carvunis, A. R., Rolland, T., Wapinski, I., Calderwood, M. A., Yildirim, M. A., Simonis, N., Charloteaux, B., Hidalgo, C. A., Barbette, J., Santhanam, B., Brar, G. A., Weissman, J. S., Regev, A., Thierry-Mieg, N., Cusick, M. E. and Vidal, M. (2012), 'Proto-genes and *de novo* gene birth', *Nature* **487**(7407), 370–374.

Casalone, E., Colella, C. M., Daly, S., Gallori, E., Moriani, L. and Polsinelli, M. (1992), 'Mechanism of resistance to sulphite in *Saccharomyces cerevisiae*', *Current Genetics* **22**(6), 435–440.

Castellari, L., Ferruzzi, M., Magrini, A., Giudici, P., Passarelli, P. and Zambonelli, C. (1994), 'Unbalanced wine fermentation by cryotolerant vs. non-cryotolerant *Saccharomyces* strains', *Vitis* **33**(1), 49–52.

Castresana, J. (2007), 'Topological variation in single-gene phylogenetic trees'.

Cavalieri, D., McGovern, P. E., Hartl, D. L., Mortimer, R. and Polsinelli, M. (2003), Evidence for *S. cerevisiae* fermentation in ancient wine, *in* 'Journal of Molecular Evolution', Vol. 57, Springer, pp. S226–S232.

Chang, S. L., Lai, H. Y., Tung, S. Y. and Leu, J. Y. (2013), 'Dynamic Large-Scale Chromosomal Rearrangements Fuel Rapid Adaptation in Yeast Populations', *PLoS Genetics* **9**(1).

Charron, G., Leducq, J.-B., Bertin, C., Dubé, A. K. and Landry, C. R. (2014), 'Exploring the northern limit of the distribution of *Saccharomyces cerevisiae* and *Saccharomyces paradoxus* in North America', *FEMS Yeast Research* **14**(2), 281–288.

Cherry, J. M., Hong, E. L., Amundsen, C., Balakrishnan, R., Binkley, G., Chan, E. T., Christie, K. R., Costanzo, M. C., Dwight, S. S., Engel, S. R., Fisk, D. G., Hirschman, J. E., Hitz, B. C., Karra, K., Krieger, C. J., Miyasato, S. R., Nash, R. S., Park, J., Skrzypek, M. S., Simison, M., Weng, S. and Wong, E. D. (2012), '*Saccharomyces* Genome Database: The genomics resource of budding yeast', *Nucleic Acids Research* **40**(D1), 700–705.

Clowers, K. J., Heilberger, J., Piotrowski, J. S., Will, J. L. and Gasch, A. P. (2015), 'Ecological and genetic barriers differentiate natural populations of *Saccharomyces cerevisiae*', *Molecular Biology and Evolution* **32**(9), 2317–2327.

Conant, G. C. and Wolfe, K. H. (2007), 'Increased glycolytic flux as an outcome of whole-genome duplication in yeast', *Molecular Systems Biology* **3**(1), 129.

Cousin, F. J., Le Guellec, R., Schlusselhuber, M., Dalmasso, M., Laplace, J.-M. and Cretenet, M. (2017), 'Microorganisms in fermented apple beverages: current knowledge and future directions', *Microorganisms* **25**, 5.

Cowart, L. A. and Obeid, L. M. (2007), 'Yeast sphingolipids: Recent developments in understanding biosynthesis, regulation, and function', *Biochimica et Biophysica Acta - Molecular and Cell Biology of Lipids* **1771**(3), 421–431.

Darwin, C. (1859), *On the origin of species by means of natural selection*, London: John Murray.

Dayarian, A., Michael, T. P. and Sengupta, A. M. (2010), 'SOPRA: Scaffolding algorithm for paired reads via statistical optimization', *BMC Bioinformatics* **11**(1), 345.

De Smidt, O., Du Preez, J. C. and Albertyn, J. (2008), 'The alcohol dehydrogenases of *Saccharomyces cerevisiae*: A comprehensive review', *FEMS Yeast Research* **8**(7), 967–978.

Delport, W., Poon, A. F., Frost, S. D. and Kosakovsky Pond, S. L. (2010), 'Datamonkey 2010: A suite of phylogenetic analysis tools for evolutionary biology', *Bioinformatics* **26**(19), 2455–2457.

Demuyter, C., Lollier, M., Legras, J. L. and Le Jeune, C. (2004), 'Predominance of *Saccharomyces uvarum* during spontaneous alcoholic fermentation, for three consecutive years, in an Alsatian winery', *Journal of Applied Microbiology* **97**(6), 1140–1148.

Denayrolles, M., De Villechenon, E. P., Lonvaud-Funel, A. and Aigle, M. (1997), 'Incidence of SUC-RTM telomeric repeated genes in brewing and wild wine strains of *Saccharomyces*', *Current Genetics* **31**(6), 457–461.

Dobler, S., Dalla, S., Wagschal, V. and Agrawal, A. A. (2012), 'Community-wide convergent evolution in insect adaptation to toxic cardenolides by substitutions in the Na,K-ATPase', *Proceedings of the National Academy of Sciences of the United States of America* **109**(32), 13040–13045.

Dujon, B. (2006), 'Yeasts illustrate the molecular mechanisms of eukaryotic genome evolution', *Trends in Genetics* **22**(7), 375–387.

Dujon, B. A. and Louis, E. J. (2017), 'Genome Diversity and Evolution in the Budding Yeasts (Saccharomycotina).', *Genetics* **206**(2), 717–750.

Dujon, B., Sherman, D., Fischer, G., Durrens, P., Casaregela, S., Lafentaine, I., De Montigny, J., Marck, C., Neuvéglise, C., Talla, E., Goffard, N., Frangeul, L., Algie, M., Anthouard, V., Babour, A., Barbe, V., Barnay, S., Blanchin, S., Beckerich, J. M., Beyne, E., Bleykasten, C., Boisramé, A., Boyer, J., Cattolico, L., Confanioleri, F., De Daruvar, A., Despons, L., Fabre, E., Fairhead, C., Ferry-Dumazet, H., Groppi, A., Hantraye, F., Hennequin, C., Jauniaux, N., Joyot, P., Kachouri, R., Kerrest, A., Koszul, R., Lemaire, M., Lesur, I., Ma, L., Muller, H., Nicaud, J. M., Nikolski, M., Oztas, S., Ozier-Kalogeropoulos, O., Pellenz, S., Potter, S., Richard, G. F., Straub, M. L., Suleau, A., Swennen, D., Tekai, F., Wésolowski-Louvel, M., Westhof, E., Wirth, B., Zeniou-Meyer, M., Zivanovic, I., Bolotin-Fukuhara, M., Thierry, A., Bouchter, C., Caudron, B., Scarpelli, C., Gaillardin, C., Weissebach, J., Wincker, P. and Souciet, J. L. (2004), 'Genome evolution in yeasts', *Nature* **430**(6995), 35–44.

Emms, D. M. and Kelly, S. (2019), 'OrthoFinder: Phylogenetic orthology inference for comparative genomics', *Genome Biology* **20**(1), 238.

Erny, C., Raoult, P., Alais, A., Butterlin, G., Delobel, P., Matei-Radoi, F., Casaregola, S. and Legras, J. L. (2012), 'Ecological Success of a Group of *Saccharomyces cerevisiae*/*Saccharomyces kudriavzevii* hybrids in the Northern European wine-making environment', *Applied and Environmental Microbiology* **78**(9), 3256–3265.

Fernandez-Espinar, T. T., Barrio, E. and Querol, A. (2003), 'Analysis of the genetic variability in the species of the *Saccharomyces* sensu stricto complex', *Yeast* **20**(14), 1213–1226.

Filteau, M., Charron, G. and Landry, C. R. (2017), 'Identification of the fitness determinants of budding yeast on a natural substrate', *ISME Journal* **11**(4), 959–971.

Force, A., Lynch, M., Pickett, F. B., Amores, A., Yan, Y. L. and Postlethwait, J. (1999), 'Preservation of duplicate genes by complementary, degenerative mutations', *Genetics* **151**(4), 1531–1545.

Gallone, B., Mertens, S., Gordon, J. L., Maere, S., Verstrepen, K. J. and Steensels, J. (2018), 'Origins, evolution, domestication and diversity of *Saccharomyces* beer yeasts'.

Gallone, B., Steensels, J., Mertens, S., Dzialo, M. C., Gordon, J. L., Wauters, R., Theßeling, F. A., Bellinazzo, F., Saels, V., Herrera-Malaver, B., Prahl, T., White, C., Hutzler, M., Meußdoerffer, F., Malcorps, P., Souffriau, B., Daenen, L., Baele, G., Maere, S. and Verstrepen, K. J. (2019), 'Interspecific hybridization facilitates niche adaptation in beer yeast', *Nature Ecology and Evolution* **3**(11), 1562–1575.

Gamero, A., Belloch, C., Ibáñez, C. and Querol, A. (2014), 'Molecular analysis of the genes involved in aroma synthesis in the species *S. cerevisiae*, *S. kudriavzevii* and *S. bayanus* var. uvarum in winemaking conditions', *PLoS ONE* **9**(5), e97626.

Gamero, A., Belloch, C. and Querol, A. (2015), 'Genomic and transcriptomic analysis of aroma synthesis in two hybrids between *Saccharomyces cerevisiae* and *S. kudriavzevii* in winemaking conditions.', *Microbial cell factories* **14**(1), 128.

Gamero, A., Tronchoni, J., Querol, A. and Belloch, C. (2013), 'Production of aroma compounds by cryotolerant *Saccharomyces* species and hybrids at low and moderate fermentation temperatures', *Journal of Applied Microbiology* **114**(5), 1405–1414.

Gangl, H., Batusic, M., Tscheik, G., Tiefenbrunner, W., Hack, C. and Lopandic, K. (2009), 'Exceptional fermentation characteristics of natural hybrids from *Saccharomyces cerevisiae* and *S. kudriavzevii*', *New Biotechnology* **25**(4), 244–251.

García-Ríos, E., Nuévalos, M., Barrio, E., Puig, S. and Guillamón, J. M. (2019), 'A new chromosomal rearrangement improves the adaptation of wine yeasts to sulfite', *Environmental Microbiology* **21**(5), 1771–1781.

Gayevskiy, V. and Goddard, M. R. (2016), '*Saccharomyces eubayanus* and *Saccharomyces arboricola* reside in North Island native New Zealand forests', *Environmental Microbiology* **18**(4), 1137–1147.

Generoso, W. C., Gottardi, M., Oreb, M. and Boles, E. (2016), 'Simplified CRISPR-Cas genome editing for *Saccharomyces cerevisiae*', *Journal of Microbiological Methods* **127**, 203–205.

Gerke, J., Lorenz, K. and Cohen, B. (2009), 'Genetic interactions between transcription factors cause natural variation in yeast', *Science* **323**(5913), 498–501.

Ghisla, S. and Massey, V. (1989), 'Mechanisms of flavoprotein catalyzed reactions', *European Journal of Biochemistry* **181**(1), 1–17.

Gietz, R. D. and Schiestl, R. H. (2007), 'Frozen competent yeast cells that can be transformed with high efficiency using the LiAc/SS carrier DNA/PEG method', *Nature Protocols* **2**(1), 1–4.

Goddard, M. R. and Greig, D. (2015), '*Saccharomyces cerevisiae*: A nomadic yeast with no niche?', *FEMS Yeast Research* **15**(3), 1–6.

Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J. D., Jacq, C., Johnston, M., Louis, E. J., Mewes, H. W., Murakami, Y., Philippsen, P., Tettelin, H. and Oliver, S. G. (1996), 'Life with 6000 genes.', *Science (New York, N.Y.)* **274**(5287), 546, 563–7.

Goldman, N. and Yang, Z. (1994), 'A codon-based model of nucleotide substitution for protein-coding DNA sequences.', *Molecular Biology and Evolution* .

Gonçalves, P., Valério, E., Correia, C., de Almeida, J. M. and Sampaio, J. P. (2011), 'Evidence for divergent evolution of growth temperature preference in sympatric *Saccharomyces* species', *PLoS ONE* **6**(6).

González Flores, M., Rodríguez, M. E., Origone, A. C., Oteiza, J. M., Querol, A. and Lopes, C. A. (2019), '*Saccharomyces uvarum* isolated from patagonian ciders shows excellent fermentative performance for low temperature cidermaking', *Food Research International* **126**, 108656.

Gonzalez Flores, M., Rodríguez, M. E., Peris, D., Querol, A., Barrio, E. and Lopes, C. A. (2020), 'Human-associated migration of Holarctic *Saccharomyces uvarum* strains to Patagonia', *Fungal Ecology* **48**, 100990.

González, S. S., Barrio, E., Gafner, J. and Querol, A. (2006), 'Natural hybrids from *Saccharomyces cerevisiae*, *Saccharomyces bayanus* and *Saccharomyces kudriavzevii* in wine fermentations', *FEMS Yeast Research* **6**(8), 1221–1234.

González, S. S., Barrio, E. and Querol, A. (2008), 'Molecular characterization of new natural hybrids of *Saccharomyces cerevisiae* and *S. kudriavzevii* in brewing', *Applied and Environmental Microbiology* **74**(8), 2314–2320.

González, S. S., Gallo, L., Climent, M. D., Barrio, E. and Querol, A. (2007), 'Enological characterization of natural hybrids from *Saccharomyces cerevisiae* and *S. kudriavzevii*', *International Journal of Food Microbiology* **116**(1), 11–18.

Goodswen, S. J., Kennedy, P. J. and Ellis, J. T. (2018), 'A gene-based positive selection detection approach to identify vaccine candidates using *Toxoplasma gondii* as a test case protozoan pathogen', *Frontiers in Genetics* **9**(AUG), 332.

Gordon, D., Abajian, C. and Green, P. (1998), 'Consed: A graphical tool for sequence finishing', *Genome Research* **8**(3), 195–202.

Gordon, J. L., Armisén, D., Proux-Wéra, E., ÓhÉigeartaigh, S. S., Byrne, K. P. and Wolfe, K. H. (2011), 'Evolutionary erosion of yeast sex chromosomes by mating-type switching accidents.', *Proceedings of the National Academy of Sciences of the United States of America* **108**(50), 20024–9.

Gordon, J. L., Byrne, K. P. and Wolfe, K. H. (2009), 'Additions, losses, and rearrangements on the evolutionary route from a reconstructed ancestor to the modern *Saccharomyces cerevisiae* genome', *PLoS Genetics* **5**(5), e1000485.

Götz, S., García-Gómez, J. M., Terol, J., Williams, T. D., Nagaraj, S. H., Nueda, M. J., Robles, M., Talón, M., Dopazo, J. and Conesa, A. (2008), 'High-throughput functional annotation and data mining with the Blast2GO suite', *Nucleic Acids Research* **36**(10), 3420–3435.

Gough, J., Karplus, K., Hughey, R. and Chothia, C. (2001), 'Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure', *Journal of Molecular Biology* **313**(4), 903–919.

Grantham, R. (1974), 'Amino acid difference formula to help explain protein evolution', *Science* **185**(4154), 862–864.

Grassi, L., Fusco, D., Sellerio, A., Corà, D., Bassetti, B., Caselle, M. and Lagomarsino, M. C. (2010), 'Identity and divergence of protein domain architectures after the yeast whole-genome duplication event', *Molecular BioSystems* **6**(11), 2305–2315.

Grenson, M., Hou, C. and Crabeel, M. (1970), 'Multiplicity of the amino acid permeases in *Saccharomyces cerevisiae*. IV. Evidence for a general amino acid permease.', *Journal of bacteriology* **103**(3), 770–777.

Gresham, D., Usaite, R., Germann, S. M., Lisby, M., Botstein, D. and Regenberg, B. (2010), 'Adaptation to diverse nitrogen-limited environments by deletion or extrachromosomal element formation of the *GAP1* locus', *Proceedings of the National Academy of Sciences of the United States of America* **107**(43), 18551–18556.

Guiard, B. (1985), 'Structure, expression and regulation of a nuclear gene encoding a mitochondrial protein: the yeast L(+)-lactate cytochrome c oxidoreductase (cytochrome b2).', *The EMBO journal* **4**(12), 3265–3272.

He, X. and Zhang, J. (2005), 'Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution', *Genetics* **169**(2), 1157–1164.

Heath, V. L., Shaw, S. L., Roy, S. and Cyert, M. S. (2004), 'Hph1p and Hph2p, novel components of calcineurin-mediated stress responses in *Saccharomyces cerevisiae*', *Eukaryotic Cell* **3**(3), 695–704.

Hittinger, C. T. and Carroll, S. B. (2007), 'Gene duplication and the adaptive evolution of a classic genetic switch', *Nature* **449**(7163), 677–681.

Hittinger, C. T., Rokas, A., Bai, F. Y., Boekhout, T., Gonçalves, P., Jeffries, T. W., Kominek, J., Lachance, M. A., Libkind, D., Rosa, C. A., Sampaio, J. P. and Kurtzman, C. P. (2015), 'Genomics and the making of yeast biodiversity'.

Hongo, J. A., de Castro, G. M., Cintra, L. C., Zerlotini, A. and Lobo, F. P. (2015), 'POTION: An end-to-end pipeline for positive Darwinian selection detection in genome-scale data through phylogenetic comparison of protein-coding genes', *BMC Genomics* **16**(1), 567.

Hou, J., Friedrich, A., De Montigny, J. and Schacherer, J. (2014), 'Chromosomal rearrangements as a major mechanism in the onset of reproductive isolation in *Saccharomyces cerevisiae*', *Current Biology* **24**(10), 1153–1159.

Izawa, S., Sato, M., Yokoigawa, K. and Inoue, Y. (2004), 'Intracellular glycerol influences resistance to freeze stress in *Saccharomyces cerevisiae*: Analysis of a quadruple mutant in glycerol dehydrogenase genes and glycerol-enriched cells', *Applied Microbiology and Biotechnology* **66**(1), 108–114.

Jeffroy, O., Brinkmann, H., Delsuc, F. and Philippe, H. (2006), 'Phylogenomics: the beginning of incongruence?', *Trends in Genetics* **22**(4), 225–231.

Katoh, K. and Standley, D. M. (2013), 'MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability', *Molecular Biology and Evolution* **30**(4), 772–780.

Kawahara, Y. and Imanishi, T. (2007), 'A genome-wide survey of changes in protein evolutionary rates across four closely related species of *Saccharomyces* sensu stricto group', *BMC Evolutionary Biology* **7**(1), 9.

Keeling, P. J. and Palmer, J. D. (2008), 'Horizontal gene transfer in eukaryotic evolution'.

Kimura, M. (1981), 'Estimation of evolutionary distances between homologous nucleotide sequences', *Proceedings of the National Academy of Sciences of the United States of America* **78**(1 II), 454–458.

Kimura, M. (1983), 'The neutral theory of molecular evolution', *Cambridge: Cambridge University Press* **154**(3).

Kondrashov, F. A. (2012), 'Gene duplication as a mechanism of genomic adaptation to a changing environment'.

Kondrashov, F. A. and Kondrashov, A. S. (2006), 'Role of selection in fixation of gene duplications', *Journal of Theoretical Biology* **239**(2), 141–151.

Künzler, M., Paravicini, G., Egli, C. M., Irniger, S. and Braus, G. H. (1992), 'Cloning, primary structure and regulation of the *ARO4* gene, encoding the tyrosine-inhibited 3-deoxy-D-arabino-heptulosonate-7-phosphate synthase from *Saccharomyces cerevisiae*.', *Gene* **113**(1), 67–74.

Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C. and Salzberg, S. L. (2004), 'Versatile and open software for comparing large genomes.', *Genome biology* **5**(2), R12.

Kurtzman, C. P., Fell, J. V. and Boekhout, T. (2011), *The Yeasts, A Taxonomic Study*.

Kurtzman, C. P. and Sugiyama, J. (2015), Saccharomycotina and taphrinomycotina: The yeasts and yeastlike fungi of the ascomycota, *in* 'Saccharomycotina and taphrinomycotina: The yeasts and yeastlike fungi of the ascomycota', Springer Berlin Heidelberg, pp. 3–33.

Kvitek, D. J. and Sherlock, G. (2011), 'Reciprocal sign epistasis between frequently experimentally evolved adaptive mutations causes a rugged fitness landscape', *PLoS Genetics* **7**(4), e1002056.

Lairón Peris, M., Morard, M., Macías, L. G., González-Flores, M., Lopes, C. A. and Barrio, E. (n.d.), 'Convergent, independent introgressions in Saccharomyces uvarum strains isolated from apple juice fermentations (in preparation).', **131**.

Langdon, Q. K., Peris, D., Baker, E. P., Opulente, D. A., Nguyen, H.-V., Bond, U., Gonçalves, P., Sampaio, J. P., Libkind, D. and Hittinger, C. T. (2019), 'Fermentation innovation through complex hybridization of wild and domesticated yeasts', *Nature Ecology & Evolution* .

Langmead, B. and Salzberg, S. L. (2012), 'Fast gapped-read alignment with Bowtie 2', *Nature Methods* **9**(4), 357–359.

Legras, J. L., Galeote, V., Bigey, F., Camarasa, C., Marsit, S., Nidelet, T., Sanchez, I., Couloux, A., Guy, J., Franco-Duarte, R., Marcet-Houben, M., Gabaldon, T., Schuller, D., Sampaio, J. P. and Dequin, S. (2018), 'Adaptation of *S. cerevisiae* to fermented food environments reveals remarkable genome plasticity and the footprints of domestication', *Molecular Biology and Evolution* **35**(7), 1712–1727.

Letunic, I. and Bork, P. (2016), 'Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees', *Nucleic acids research* **44**(W1), W242–W245.

Libkind, D., Hittinger, C. T., Valeŕio, E., Goncalves, C., Dover, J., Johnston, M., Goncalves, P. and Sampaio, J. P. (2011), 'Microbe domestication and the identification of the wild genetic stock of lager-brewing yeast', *Proceedings of the National Academy of Sciences of the United States of America* **108**(35), 14539–14544.

Libkind, D., Peris, D., Cubillos, F. A., Steenwyk, J. L., Opulente, D. A., Langdon, Q. K., Rokas, A. and Hittinger, C. T. (2020), 'Into the wild: new yeast genomes from natural environments and new tools for their analysis', *FEMS yeast research* **20**(2).

Liti, G., Carter, D. M., Moses, A. M., Warringer, J., Parts, L., James, S. A., Davey, R. P., Roberts, I. N., Burt, A., Koufopanou, V., Tsai, I. J., Bergman, C. M., Bensasson, D., O'Kelly, M. J., Van Oudenaarden, A., Barton, D. B., Bailes, E., Nguyen, A. N., Jones, M., Quail, M. A., Goodhead, I., Sims, S., Smith, F., Blomberg, A., Durbin, R. and Louis, E. J. (2009), 'Population genomics of domestic and wild yeasts', *Nature* **458**(7236), 337–341.

Liti, G., Nguyen Ba, A. N., Blythe, M., Müller, C. A., Bergström, A., Cubillos, F. A., Dafhnis-Calas, F., Khoshraftar, S., Malla, S., Mehta, N., Siow, C. C., Warringer, J., Moses, A. M., Louis, E. J. and Nieduszynski, C. A. (2013), 'High quality *de novo* sequencing and assembly of the *Saccharomyces arboricolus* genome.', *BMC genomics* **14**(1), 69.

Lodi, T. and Guiard, B. (1991), 'Complex transcriptional regulation of the *Saccharomyces cerevisiae CYB2* gene encoding cytochrome b2: *CYP1(HAP1)* activator binds to the *CYB2* upstream activation site *UAS1-B2*.', *Molecular and Cellular Biology* **11**(7), 3762–3772.

Lopes, C. A., Barrio, E. and Querol, A. (2010), 'Natural hybrids of *S. cerevisiae* × *S. kudriavzevii* share alleles with European wild populations of *Saccharomyces kudriavzevii*', *FEMS Yeast Research* **10**(4), 412–421.

López-Malo, M., Querol, A. and Guillamon, J. M. (2013), 'Metabolomic Comparison of *Saccharomyces cerevisiae* and the Cryotolerant Species *S. bayanus* var. *uvarum* and *S. kudriavzevii* during Wine Fermentation at Low Temperature', *PLoS ONE* **8**(3).

Losos, J. B. (2011), 'Convergence, adaptation, and constraint', *Evolution* **65**(7), 1827–1840.

Lynch, M. (2007), The origins of genome architecture, Technical report.

Lynch, M. and Katju, V. (2004), 'The altered evolutionary trajectories of gene duplicates', *Trends in Genetics* **20**(11), 544–549.

Lynch, M., Sung, W., Morris, K., Coffey, N., Landry, C. R., Dopman, E. B., Dickinson, W. J., Okamoto, K., Kulkarni, S., Hartl, D. L. and Thomas, W. K. (2008), 'A genome-wide view of the spectrum of spontaneous mutations in yeast', *Proceedings of the National Academy of Sciences of the United States of America* **105**(27), 9272–9277.

Macías, L. G., Barrio, E. and Toft, C. (2020), 'GWideCodeML: A Python Package for Testing Evolutionary Hypotheses at the Genome-Wide Level', *G3: Genes, Genomes, Genetics* **10**, 4369 – 4372.

Macías, L. G., Morard, M., Toft, C. and Barrio, E. (2019), 'Comparative genomics between *Saccharomyces kudriavzevii* and *S. cerevisiae* applied to identify mechanisms involved in adaptation', *Frontiers in Genetics* **10**, 187.

Maldonado, E., Almeida, D., Escalona, T., Khan, I., Vasconcelos, V. and Antunes, A. (2016), 'LMAP: Lightweight Multigene Analyses in PAML', *BMC Bioinformatics* **17**(1), 354.

Marcet-Houben, M. and Gabaldón, T. (2015), 'Beyond the whole-genome duplication: Phylogenetic evidence for an ancient interspecies hybridization in the baker's yeast lineage', *PLoS Biology* **13**(8), 1–26.

Marsit, S., Leducq, J.-B., Durand, É., Marchant, A., Filteau, M. and Landry, C. R. (2017), 'Evolutionary biology through the lens of budding yeast comparative genomics', *Nature Reviews Genetics* .

Marsit, S., Mena, A., Bigey, F., Sauvage, F. X., Couloux, A., Guy, J., Legras, J. L., Barrio, E., Dequin, S. and Galeote, V. (2015a), 'Evolutionary advantage conferred by an eukaryote-to-eukaryote gene transfer event in wine yeasts', *Molecular Biology and Evolution* **32**(7), 1695–1707.

Martin, A. and Orgogozo, V. (2013), 'The loci of repeated evolution: A catalog of genetic hotspots of phenotypic variation', *Evolution* **67**(5), 1235–1250.

McGovern, P. E., Zhang, J., Tang, J., Zhang, Z., Hall, G. R., Moreau, R. A., Nuñez, A., Butrym, E. D., Richards, M. P., Wang, C. S., Cheng, G., Zhao, Z. and Wang, C. (2004), 'Fermented beverages of pre- and proto-historic China', *PNAS* **101**(51), 17593–17598.

Michels, C. A. and Needleman, R. B. (1984), 'The dispersed, repeated family of MAL loci in *Saccharomyces* spp.', *Journal of Bacteriology* **157**(3), 949–952.

Miller, R. G. (1981), *Simultaneous Statistical Inference*, Springer Series in Statistics, Springer New York, New York, NY.

Minebois, R., Pérez-Torrado, R. and Querol, A. (2020), 'Metabolome segregation of four strains of *Saccharomyces cerevisiae*, *Saccharomyces uvarum* and *Saccharomyces kudriavzevii* conducted under low temperature oenological conditions', *Environmental Microbiology* **22**(9), 3700–3721.

Molina, A. M., Swiegers, J. H., Varela, C., Pretorius, I. S. and Agosin, E. (2007), 'Influence of wine fermentation temperature on the synthesis of yeast-derived volatile aroma compounds', *Applied Microbiology and Biotechnology* **77**(3), 675–687.

Morales, L. and Dujon, B. (2012), 'Evolutionary Role of Interspecies Hybridization and Genetic Exchanges in Yeasts', *Microbiology and Molecular Biology Reviews* **76**(4), 721–739.

Morard, M., Benavent-Gil, Y., Ortiz-Tovar, G., Pérez-Través, L., Querol, A., Toft, C. and Barrio, E. (2020), 'Genome structure reveals the diversity of mating mechanisms in *Saccharomyces cerevisiae* x *Saccharomyces kudriavzevii* hybrids, and the genomic instability that promotes phenotypic diversity', *Microbial Genomics* **6**(3).

Morard, M., Macías, L. G., Adam, A. C., Lairón-Peris, M., Pérez-Torrado, R., Toft, C. and Barrio, E. (2019), 'Aneuploidy and ethanol tolerance in *Saccharomyces cerevisiae*', *Frontiers in Genetics* **10**, 82.

Nadai, C., Treu, L., Campanaro, S., Giacomini, A. and Corich, V. (2016), 'Different mechanisms of resistance modulate sulfite tolerance in wine yeasts', *Applied Microbiology and Biotechnology* **100**(2), 797–813.

Nagy, L. G., Ohm, R. A., Kovács, G. M., Floudas, D., Riley, R., Gácser, A., Sipiczki, M., Davis, J. M., Doty, S. L., De Hoog, G. S., Lang, B. F., Spatafora, J. W., Martin, F. M., Grigoriev, I. V. and Hibbett, D. S. (2014), 'Latent homology and convergent regulatory evolution underlies the repeated emergence of yeasts', *Nature Communications* **5**(1), 1–8.

Naseeb, S., Alsammar, H., Burgis, T., Donaldson, I., Knyazev, N., Knight, C. and Delneri, D. (2018), 'Whole genome sequencing, *de novo* assembly and phenotypic profiling for the new budding yeast species *Saccharomyces jurei*', *G3: Genes, Genomes, Genetics* **8**(9), 2967–2977.

Naseeb, S., James, S. A., Alsammar, H., Michaels, C. J., Gini, B., Nueno-Palop, C., Bond, C. J., McGhie, H., Roberts, I. N. and Delneri, D. (2017), '*Saccharomyces jurei* sp. Nov., isolation and genetic identification of a novel yeast species from Quercus robur', *International Journal of Systematic and Evolutionary Microbiology* **67**(6), 2046–2052.

Naumov, G. (1987), 'Genetic Basis for Classification and Identification of the Ascomycetous Yeasts', *Studies in Mycology* **30**, 469–475.

Naumov, G. I., James, S. A., Naumova, E. S., Louis, E. J. and Roberts, I. N. (2000a), 'Three new species in the *Saccharomyces* sensu stricto complex: *Saccharomyces* cariocanus, *Saccharomyces kudriavzevii* and *Saccharomyces mikatae*', *International Journal of Systematic and Evolutionary Microbiology* **50**(5), 1931–1942.

Naumov, G. I., Lee, C. F. and Naumova, E. S. (2013), 'Molecular genetic diversity of the *Saccharomyces* yeasts in Taiwan: *Saccharomyces* arboricola, *Saccharomyces cerevisiae* and *Saccharomyces kudriavzevii*', *Antonie van Leeuwenhoek, International Journal of General and Molecular Microbiology* **103**(1), 217–228.

Naumov, G. I., Masneuf, I., Naumova, E. S., Aigle, M. and Dubourdieu, D. (2000b), 'Association of *Saccharomyces* bayanus var. uvarum with some French wines: Genetic analysis of yeast populations', *Research in Microbiology* **151**(8), 683–691.

Naumov, G. I., Naumova, E. S., Antunovics, Z. and Sipiczki, M. (2002), '*Saccharomyces bayanus* var. uvarum in Tokaj wine-making of Slovakia and Hungary', *Applied Microbiology and Biotechnology* **59**(6), 727–730.

Naumov, G. I., Naumova, E. S. and Sniegowski, P. D. (1998), '*Saccharomyces paradoxus* and *Saccharomyces cerevisiae* are associated with exudates of North American oaks', *Canadian Journal of Microbiology* **44**(11), 1045–1050.

Naumov, G. I., Shalamitskiy, M. Y. and Naumova, E. S. (2016), 'New family of pectinase genes PGU1b–PGU3b of the pectinolytic yeast *Saccharomyces* bayanus var. uvarum', *Doklady Biochemistry and Biophysics* **467**(1), 89–91.

Nespolo, R. F., Villarroel, C. A., Oporto, C. I., Tapia, S. M., Vega-Macaya, F., Urbina, K., de Chiara, M., Mozzachiodi, S., Mikhalev, E., Thompson, D., Larrondo, L. F., Saenz-Agudelo, P., Liti, G. and Cubillos, F. A. (2020), 'An Out-of-Patagonia migration explains the worldwide diversity and distribution of *Saccharomyces eubayanus* lineages', *PLoS Genetics* **16**(5), e1008777.

Nevoigt, E. and Stahl, U. (1997), 'Osmoregulation and glycerol metabolism in the yeast *Saccharomyces cerevisiae*', *FEMS Microbiology Reviews* **21**(3), 231–241.

Novo, M., Bigey, F., Beyne, E., Galeote, V., Gavory, F., Mallet, S., Cambon, B., Legras, J. L., Wincker, P., Casaregola, S. and Dequin, S. (2009), 'Eukaryote-to-eukaryote gene transfer events revealed by the genome sequence of the wine yeast *Saccharomyces cerevisiae* EC1118', *Proceedings of the National Academy of Sciences of the United States of America* **106**(38), 16333–16338.

Ohno, S. (1970), *Evolution by Gene Duplication*, Springer Berlin Heidelberg.

Oliveira, B. M., Barrio, E., Querol, A. and Pérez-Torrado, R. (2014), 'Enhanced enzymatic activity of glycerol-3-phosphate dehydrogenase from the cryophilic *Saccharomyces kudriavzevii*', *PLoS ONE* **9**(1), e87290.

Oppler, Z. J., Parrish, M. E. and Murphy, H. A. (2019), Variation at an adhesin locus suggests sociality in natural populations of the yeast *Saccharomyces cerevisiae*, *in* 'Proceedings of the Royal Society B: Biological Sciences', Vol. 286, Royal Society Publishing, p. 20191948.

Opulente, D. A., Rollinson, E. J., Bernick-Roehr, C., Hulfachor, A. B., Rokas, A., Kurtzman, C. P. and Hittinger, C. T. (2018), 'Factors driving metabolic diversity in the budding yeast subphylum', *BMC Biology* **16**(1), 26.

Otto, T. D., Dillon, G. P., Degrave, W. S. and Berriman, M. (2011), 'RATT: Rapid Annotation Transfer Tool', *Nucleic Acids Research* **39**(9), 1–7.

Otto, T. D., Sanders, M., Berriman, M. and Newbold, C. (2010), 'Iterative correction of reference Nucleotides (iCORN) using second generation sequencing technology', *Bioinformatics* **26**(14), 1704–1707.

Paget, C. M., Schwartz, J. M. and Delneri, D. (2014), 'Environmental systems biology of cold-tolerant phenotype in *Saccharomyces* species adapted to grow at different temperatures', *Molecular Ecology* **23**(21), 5241–5257.

Papp, B., Pál, C. and Hurst, L. D. (2003), 'Dosage sensitivity and the evolution of gene families in yeast', *Nature* **424**(6945), 194–197.

Paradis, E. and Schliep, K. (2019), 'Ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R', *Bioinformatics* **35**(3), 526–528.

Park, H. and Bakalinsky, A. T. (2000), '*SSU1* mediates sulphite efflux in *Saccharomyces cerevisiae*', *Yeast* **16**(10), 881–888.

Payen, C., Di Rienzi, S. C., Ong, G. T., Pogachar, J. L., Sanchez, J. C., Sunshine, A. B., Raghuraman, M. K., Brewer, B. J. and Dunham, M. J. (2014), 'The dynamics of diverse segmental amplifications in populations of *Saccharomyces cerevisiae* adapting to strong selection', *G3: Genes, Genomes, Genetics* **4**(3), 399–409.

Pérez-Ortín, J. E., Querol, A., Puig, S. and Barrio, E. (2002), 'Molecular characterization of a chromosomal rearrangement involved in the adaptie evolution of yeast strains', *Genome Research* **12**(10), 1533–1539.

Pérez-Torrado, R., Barrio, E. and Querol, A. (2018), 'Alternative yeasts for winemaking: *Saccharomyces* non-cerevisiae and its hybrids', *Critical Reviews in Food Science and Nutrition* **58**(11), 1780–1790.

Pérez-Través, L., Lopes, C. A., Querol, A. and Barrio, E. (2014), 'On the complexity of the *Saccharomyces* bayanus taxon: Hybridization and potential hybrid speciation', *PLoS ONE* **9**(4).

Peris, D., Langdon, Q. K., Moriarty, R. V., Sylvester, K., Bontrager, M., Charron, G., Leducq, J. B., Landry, C. R., Libkind, D. and Hittinger, C. T. (2016), 'Complex Ancestries of Lager-Brewing Hybrids Were Shaped by Standing Variation in the Wild Yeast *Saccharomyces eubayanus*', *PLoS Genetics* **12**(7), e1006155.

Peris, D., Pérez-Torrado, R., Hittinger, C. T., Barrio, E. and Querol, A. (2018), 'On the origins and industrial applications of *Saccharomyces cerevisiae* × *Saccharomyces kudriavzevii* hybrids', *Yeast* **35**(1), 51–69.

Peris, D., Sylvester, K., Libkind, D., Gonçalves, P., Sampaio, J. P., Alexander, W. G. and Hittinger, C. T. (2014), 'Population structure and reticulate evolution of *Saccharomyces eubayanus* and its lager-brewing hybrids', *Molecular Ecology* **23**(8), 2031–2045.

Peter, J., De Chiara, M., Friedrich, A., Yue, J.-X., Pflieger, D., Bergström, A., Sigwalt, A., Barre, B., Freel, K., Llored, A., Cruaud, C., Labadie, K., Aury, J.-M., Istace, B., Lebrigand, K., Barbry, P., Engelen, S., Lemainque, A., Wincker, P., Liti, G. and Schacherer, J. (2018), 'Genome evolution across 1,011 *Saccharomyces cerevisiae* isolates', *Nature* **556**(7701), 339–344.

Piškur, J., Rozpedowska, E., Polakova, S., Merico, A. and Compagno, C. (2006), 'How did *Saccharomyces* evolve to become a good brewer?', *Trends in Genetics* **22**(4), 183–186.

Pretorius, I. S. (2000), 'Tailoring wine yeast for the new millennium: Novel approaches to the ancient art of winemaking', *Yeast* **16**(8), 675–729.

Price, M. N., Dehal, P. S. and Arkin, A. P. (2010), 'FastTree 2 - Approximately maximum-likelihood trees for large alignments', *PLoS ONE* **5**(3).

Pronk, J. T., Steensma, H. Y. and Van Dijken, J. P. (1996), 'Pyruvate metabolism in *Saccharomyces cerevisiae*', *Yeast* **12**(16), 1607–1633.

Rachidi, N., Barre, P. and Blondin, B. (1999), 'Multiple Ty-mediated chromosomal translocations lead to karyotype changes in a wine strain of *Saccharomyces cerevisiae*', *Molecular and General Genetics* **261**(4-5), 841–850.

Rainieri, S., Zambonelli, C., Hallsworth, J. E., Pulvirenti, A. and Giudici, P. (1999), '*Saccharomyces uvarum*, a distinct group within *Saccharomyces sensu stricto*', *FEMS Microbiology Letters* **177**(1), 177–185.

Replansky, T., Koufopanou, V., Greig, D. and Bell, G. (2008), 'Saccharomyces sensu stricto as a model system for evolution and ecology', *Trends in Genetics* **23**(9), 494–501.

Rodríguez, M. E., Pérez-Través, L., Sangorrín, M. P., Barrio, E. and Lopes, C. A. (2014), '*Saccharomyces eubayanus* and *Saccharomyces uvarum* associated with the fermentation of Araucaria araucana seeds in Patagonia', *FEMS Yeast Research* **14**(6), 948–965.

Rodríguez, M. E., Pérez-Través, L., Sangorrín, M. P., Barrio, E., Querol, A. and Lopes, C. A. (2017), '*Saccharomyces uvarum* is responsible for the traditional fermentation of apple chicha in Patagonia', *FEMS Yeast Research* **17**(1), fow109.

Rokas, A. and Carroll, S. B. (2006), 'Bushes in the tree of life', *Plos Biology* **4**(11), 1899–1904.

Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M. A. and Barrell, B. (2000), 'Artemis: Sequence visualization and annotation', *Bioinformatics* **16**(10), 944–945.

Sahm, A., Bens, M., Platzer, M. and Szafranski, K. (2017), 'PosiGene: Automated and easy-to-use pipeline for genome-wide detection of positively selected genes', *Nucleic Acids Research* **45**(11), e100.

Salvadó, Z., Arroyo-López, F. N., Guillamón, J. M., Salazar, G., Quero, A. and Barrio, E. (2011), 'Temperature adaptation Markedly Determines evolution within the genus *Saccharomyces*', *Applied and Environmental Microbiology* **77**(7), 2292–2302.

Sampaio, J. P. and Gonçalves, P. (2008), 'Natural populations of *Saccharomyces kudriavzevii* in Portugal are associated with Oak bark and are sympatric with *S. cerevisiae* and *S. paradoxus*', *Applied and Environmental Microbiology* **74**(7), 2144–2152.

Scannell, D. R., Frank, A. C., Conant, G. C., Byrne, K. P., Woolfit, M. and Wolfe, K. H. (2007), 'Independent sorting-out of thousands of duplicated gene pairs in two yeast species descended from a whole-genome duplication', *Proceedings of the National Academy of Sciences of the United States of America* **104**(20), 8397–8402.

Scannell, D. R., Zill, O. A., Rokas, A., Payen, C., Dunham, M. J., Eisen, M. B., Rine, J., Johnston, M. and Hittinger, C. T. (2011), 'The awesome power of yeast evolutionary genetics: New genome sequences and strain resources for the *Saccharomyces* sensu stricto genus', *G3: Genes, Genomes, Genetics* **1**(1), 11–25.

Schwelberger, H. G., Dieter, K. S. and Paltauf, F. (1989), 'Molecular cloning, primary structure and disruption of the structural gene of aldolase from *Saccharomyces cerevisiae*', *European Journal of Biochemistry* **180**(2), 301–308.

Selmecki, A., Forche, A. and Berman, J. (2006), 'Aneuploidy and isochromosome formation in drug-resistant *Candida albicans*', *Science* **313**(5785), 367–370.

Selmecki, A. M., Dulmage, K., Cowen, L. E., Anderson, J. B. and Berman, J. (2009), 'Acquisition of aneuploidy provides increased fitness during the evolution of antifungal drug resistance', *PLoS Genetics* **5**(10), e1000705.

Selmecki, A. M., Maruvka, Y. E., Richmond, P. A., Guillet, M., Shoresh, N., Sorenson, A. L., De, S., Kishony, R., Michor, F., Dowell, R. and Pellman, D. (2015), 'Polyploidy can drive rapid adaptation in yeast', *Nature* **519**(7543), 349–351.

Seoighe, C. and Wolfe, K. H. (1998), 'Extent of genomic rearrangement after genome duplication in yeast', *Proceedings of the National Academy of Sciences of the United States of America* **95**(8), 4447–4452.

Shen, P. S., Park, J., Qin, Y., Li, X., Parsawar, K., Larson, M. H., Cox, J., Cheng, Y., Lambowitz, A. M., Weissman, J. S., Brandman, O. and Frost, A. (2015), 'Rqc2p and 60S ribosomal subunits mediate mRNA-independent elongation of nascent chains', *Science* **347**(6217), 75–78.

Shen, X.-X., Opulente, D. A., Kominek, J., Zhou, X., Steenwyk, J. L., Buh, K. V., Haase, M. A. B., Wisecaver, J. H., Wang, M., Doering, D. T., Boudouris, J. T., Schneider, R. M., Langdon, Q. K., Ohkuma, M., Endoh, R., Takashima, M., Manabe, R.-I., Čadež, N., Libkind, D., Rosa, C. A., DeVirgilio, J., Hulfachor, A. B., Groenewald, M., Kurtzman, C. P., Hittinger, C. T. and Rokas, A. (2018), 'Tempo and Mode of Genome Evolution in the Budding Yeast Subphylum.', *Cell* **175**(6), 1533–1545.e20.

Shih, S. C., Prag, G., Francis, S. A., Sutanto, M. A., Hurley, J. H. and Hicke, L. (2003), 'A ubiquitin-binding motif required for intramolecular monoubiquitylation, the CUE domain', *EMBO Journal* **22**(6), 1273–1281.

Shindo, C., Aranzana, M. J., Lister, C., Baxter, C., Nicholls, C., Nordborg, M. and Dean, C. (2005), 'Role of FRIGIDA and FLOWERING LOCUS C in determining variation in flowering time of *Arabidopsis*', *Plant Physiology* **138**(2), 1163–1173.

Song, K. H., Lee, J. K., Song, J. Y., Hong, S. G., Baek, H., Kim, S. Y. and Hyun, H. H. (2002), 'Production of mannitol by a novel strain of *Candida magnoliae*', *Biotechnology Letters* **24**(1), 9–12.

Souciet, J. L., Dujon, B., Gaillardin, C., Johnston, M., Baret, P. V., Cliften, P., Sherman, D. J., Weissenbach, J., Westhof, E., Wincker, P., Jubin, C., Poulain, J., Barbe, V., Ségurens, B., Artiguenave, F., Anthouard, V., Vacherie, B., Val, M. E., Fulton, R. S., Minx, P., Wilson, R., Durrens, P., Jean, G., Marck, C., Martin, T., Nikolski, M., Rolland, T., Seret, M. L., Casarégola, S., Despons, L., Fairhead, C., Fischer, G., Lafontaine, I., Leh, V., Lemaire, M., De Montigny, J., Neuvéglise, C., Thierry, A., Blanc-Lenfle, I., Bleykasten, C., Diffels, J., Fritsch, E., Frangeul, L., Goëffon, A., Jauniaux, N., Kachouri-Lafond, R., Payen, C., Potier, S., Pribylova, L., Ozanne, C., Richard, G. F., Sacerdot, C., Straub, M. L. and Talla, E. (2009), 'Comparative genomics of protoploid Saccharomycetaceae', *Genome Research* **19**(10), 1696–1709.

Stamatakis, A. (2014), 'RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies', *Bioinformatics* **30**(9), 1312–1313.

Stanke, M. and Morgenstern, B. (2005), 'AUGUSTUS: A web server for gene prediction in eukaryotes that allows user-defined constraints', *Nucleic Acids Research* **33**(SUPPL. 2).

Stanke, M. and Waack, S. (2003), 'Gene prediction with a hidden Markov model and a new intron submodel.', *Bioinformatics (Oxford, England)* **19**, 215–225.

Steinway, S. N., Dannenfelser, R., Laucius, C. D., Hayes, J. E. and Nayak, S. (2010), 'JCoDA: A tool for detecting evolutionary selection', *BMC Bioinformatics* **11**(1), 284.

Stern, A., Doron-Faigenboim, A., Erez, E., Martz, E., Bacharach, E. and Pupko, T. (2007), 'Selecton 2007: Advanced models for detecting positive and purifying selection using a Bayesian inference approach', *Nucleic Acids Research* **35**(SUPPL.2).

Stern, D. L. (2013), 'The genetic causes of convergent evolution', *Nature Genetics* **14**(11), 751–764.

Stribny, J., Gamero, A., Pérez-Torrado, R. and Querol, A. (2015), '*Saccharomyces kudriavzevii* and *Saccharomyces uvarum* differ from *Saccharomyces cerevisiae* during the production of aroma-active higher alcohols and acetate esters using their amino acidic precursors', *International Journal of Food Microbiology* **205**, 41–46.

Stribny, J., Querol, A. and Pérez-Torrado, R. (2016a), 'Differences in enzymatic properties of the *Saccharomyces kudriavzevii* and *Saccharomyces uvarum* alcohol acetyltransferases and their impact on aroma-active compounds production', *Frontiers in Microbiology* **7**(JUN), 897.

Sucena, É. and Stern, D. L. (2000), 'Divergence of larval morphology between *Drosophila sechellia* and its sibling species caused by cis-regulatory evolution of ovo/shaven-baby', *Proceedings of the National Academy of Sciences of the United States of America* **97**(9), 4530–4534.

Sym, M., Engebrecht, J. A. and Roeder, G. S. (1993), 'ZIP1 is a synaptonemal complex protein required for meiotic chromosome synapsis', *Cell* **72**(3), 365–378.

Tajima, F. (1993), 'Simple methods for testing the molecular evolutionary clock hypothesis', *Genetics* **135**(2), 599–607.

Thomson, J. M., Gaucher, E. A., Burgan, M. F., De Kee, D. W., Li, T., Aris, J. P. and Benner, S. A. (2005), 'Resurrecting ancestral alcohol dehydrogenases from yeast', *Nature Genetics* **37**(6), 630–635.

Toft, C., Williams, T. A. and Fares, M. A. (2009), 'Genome-wide functional divergence after the symbiosis of proteobacteria with insects unraveled through a novel computational Aapproach', *PLoS Computational Biology* **5**(4), e1000344.

Tronchoni, J., Medina, V., Guillamón, J. M., Querol, A. and Pérez-Torrado, R. (2014), 'Transcriptomics of cryophilic *Saccharomyces kudriavzevii* reveals the key role of gene translation efficiency in cold stress adaptations.', *BMC genomics* **15**(1), 432.

Tronchoni, J., Rozès, N., Querol, A. and Guillamón, J. M. (2012), 'Lipid composition of wine strains of *Saccharomyces kudriavzevii* and *Saccharomyces cerevisiae* grown at low temperature', *International Journal of Food Microbiology* **155**(3), 191–198.

Varela, C. (2016), 'The impact of non-*Saccharomyces* yeasts in the production of alcoholic beverages', *Applied Microbiology and Biotechnology* **100**(23), 9861–9874.

Vaughan Martini, A. (1989), '*Saccharomyces paradoxus* comb. nov., a newly separated species of the *Saccharomyces* sensu stricto complex based upon nDNA/nDNA homologies', *Systematic and Applied Microbiology* **12**(2), 179–182.

Vaughan Martini, A. and Kurtzman, C. P. (1985), 'Deoxyribonucleic acid relatedness among species of the genus *Saccharomyces* sensu stricto', *International Journal of Systematic Bacteriology* **35**(4), 508–511.

Verstrepen, K. J. and Klis, F. M. (2006), 'Flocculation, adhesion and biofilm formation in yeasts', *Molecular Microbiology* **60**(1), 5–15.

Viswanathan, M., Muthukumar, G., Cong, Y. S. and Lenard, J. (1994), 'Seripauperins of *Saccharomyces cerevisiae*: a new multigene family encoding serine-poor relatives of serine-rich proteins', *Gene* **148**(1), 149–153.

Wagner, A. (2008), 'Neutralism and selectionism: A network-based reconciliation'.

Walker, B., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C., Zeng, Q., Wortman, J., Young, S. and Earl, A. (2014), 'Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement', *PLoS ONE* **9**(11), e112963.

Ward, J. H. (1963), 'Hierarchical Grouping to Optimize an Objective Function', *Journal of the American Statistical Association* **58**(301), 236.

White, M. A., Diffenbaugh, N. S., Jones, G. V., Pal, J. S. and Giorgi, F. (2006), 'Extreme heat reduces and shifts United States premium wine production in the 21st century', *Proceedings of the National Academy of Sciences of the United States of America* **103**(30), 11217–11222.

Wolfe, K. (2000), 'Robustness - It's not where you think it is', *Nature Genetics* **25**(1), 3–4.

Wolfe, K. H. and Shields, D. C. (1997), 'Molecular evidence for an ancient duplication of the entire yeast genome', *Nature* **387**(6634), 708–713.

Wong, S. and Wolfe, K. H. (2005), 'Birth of a metabolic gene cluster in yeast by adaptive gene relocation', *Nature Genetics* **37**(7), 777–782.

Xia, W., Nielly-Thibault, L., Charron, G., Landry, C. R., Kasimer, D., Anderson, J. B. and Kohn, L. M. (2017), 'Population genomics reveals structure at the individual, host-tree scale and persistence of genotypic variants of the undomesticated yeast *Saccharomyces paradoxus* in a natural woodland', *Molecular Ecology* **26**(4), 995–1007.

Xu, B. and Yang, Z. (2013), 'PamlX: A graphical user interface for PAML', *Molecular Biology and Evolution* **30**(12), 2723–2724.

Yang, Z. (1998), 'Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution', *Molecular Biology and Evolution* **15**(5), 568–573.

Yang, Z. (2007), 'PAML 4: Phylogenetic analysis by maximum likelihood', *Molecular Biology and Evolution* **24**(8), 1586–1591.

Yang, Z. and Nielsen, R. (1998), 'Synonymous and nonsynonymous rate variation in nuclear genes of mammals', *Journal of Molecular Evolution* **46**(4), 409–418.

Yang, Z., Wong, W. S. and Nielsen, R. (2005), 'Bayes empirical Bayes inference of amino acid sites under positive selection', *Molecular Biology and Evolution* **22**(4), 1107–1118.

Yoo, H. S., Genbauffe, F. S. and Cooper, T. G. (1985), 'Identification of the ureidoglycolate hydrolase gene in the DAL gene cluster of *Saccharomyces cerevisiae*.', *Molecular and Cellular Biology* **5**(9), 2279–2288.

Yuasa, N., Nakagawa, Y., Hayakawa, M. and Iimura, Y. (2004), 'Distribution of the sulfite resistance gene *SSU1*-R and the variation in its promoter region in wine yeasts', *Journal of Bioscience and Bioengineering* **98**(5), 394–397.

Yue, J.-X., Li, J., Aigrain, L., Hallin, J., Persson, K., Oliver, K., Bergström, A., Coupland, P., Warringer, J., Lagomarsino, M. C., Fischer, G., Durbin, R. and Liti, G. (2017), 'Contrasting evolutionary genome dynamics between domesticated and wild yeasts', *Nature Genetics* **49**(6), 913–924.

Zerbino, D. R. and Birney, E. (2008), 'Velvet: Algorithms for *de novo* short read assembly using de Bruijn graphs', *Genome Research* **18**(5), 821–829.

Zhang, C., Rabiee, M., Sayyari, E. and Mirarab, S. (2018), 'ASTRAL-III: Polynomial time species tree reconstruction from partially resolved gene trees', *BMC Bioinformatics* **19**.

Zhang, J. (2003), 'Evolution by gene duplication: An update', *Trends in Ecology and Evolution* **18**(6), 292–298.

Zhang, J., Nielsen, R. and Yang, Z. (2005), 'Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level', *Molecular Biology and Evolution* **22**(12), 2472–2479.

Zimmer, A., Durand, C., Loira, N., Durrens, P., Sherman, D. J. and Marullo, P. (2014), 'QTL dissection of lag phase in wine fermentation reveals a new translocation responsible for *Saccharomyces cerevisiae* adaptation to sulfite', *PLoS ONE* **9**(1), e86298.

The supplementary files are available online at https://drive.google.com/drive/folders/1czbMDDxClo34fzbLckus4NT2RW6E2e8H?usp=sharing. This link can also be obtained by scanning the following QR code.

Supplementary files Chapter 1

- **Table S1.1**: Assembly statistics for *S. kudriavzevii* CA111 and CR85 genome assemblies.
- **Table S1.2**: Positive selection, evolutionary rate test and functional divergence results for all genes analyzed in both *S. kudriavzevii* and *S. cerevisiae* species. PS, positive selection; FD, number of positions contributing to functional divergence in every protein; SCOP, annotations with SCOP domains as explained in Methods; NA (not available), protein domain annotation was not possible; Tajima's test, *Sk*: acceleration of evolutionary rates leading to *S. kudriavzevii* branch; *Sc*, acceleration of evolutionary rates leading to *S. cerevisiae* branch.
- **Table S1.3**: GO term enrichment for genes showing evidence of functional divergence in *S. kudriavzevii* and *S. cerevisiae* branch.
- **Table S1.4:** Amino acids sites under positive selection according to branch-site model and BEB method. Specific amino acids sites and the probability of being under positive selection were retrieved for those candidates obtained with the branch-site model.
- **Table S1.5:** Enriched/Impoverished chromosome regions in proteins with functional divergence evidence.
- **Figure S1.1:** Functional divergence among metabolic pathways. Normalized contribution of genes showing evidence of functional divergence to every path. The height of the bars represents Φ, the normalized contribution of each pathway (*i*) of size (*t*) to the total number of genes under functional divergence when considering the whole dataset (*T*), calculated as $\Phi = (n_i / t) (t / T)$. Bars above the dashed line represent enriched pathways in genes under functional divergence while bars below the line show impoverished pathways. B, biosynthesis; M, metabolism; D, degradation; aa, amino acid.
- **Figure S1.2:** *S. kudriavzevii* vs. *S. cerevisiae* differences in functional divergence among metabolic pathways. Normalized functional divergence values among metabolic pathways. The significance of the differences in every pathway between analysis performed with *S. kudriavzevii* or *S. cerevisiae* as clade-of-interest was assessed by a Wilcoxon paired signed-rank test, those significant were indicated with an '*'. B, biosynthesis; M, metabolism; D, degradation; aa, amino acid.

Supplementary files Chapter 2

- **Table S2.1:** List of genes obtained under positive selection using the GWideCodeML package under the branch-site model. We used the genomic data of the Chapter 1 to package testing. A list of genes under positive selection and their amino acid positions are shown in different sheets. The first sheet (Sc) shows the genes under positive selection when *S. cerevisiae* was set as the foreground branch, and the second sheet (Sk) when *S. kudriavzevii* was set as the foreground branch.
- **Table S2.2:** Results of GWideCodeML performance using the genomic data of the Table 2.2. Genes under positive selection under the branch-site and branch model using *S. uvarum* as the foreground branch are shown on the first and second sheet, respectively. Genes having amino acid positions under the site-model are shown on the third sheet. Amino acid positions shown correspond to the amino acid position of the reference strain (S288c) in the multiple-sequence alignment.

Supplementary files Chapter 3

- **Table S3.1:** Genome sequences used for the phylogenetic and variant calling analyses. *De novo* sequenced genomes used in this study, references and number of annotated genes.
- **Table S3.2:** Sulphite tolerance and *SSU1* promoter in a collection of *S. uvarum* strains isolated from different environments and geographic locations. Drop test assay results are represented by the number of the most diluted (from 1 to the less diluted and 6 to the most diluted) that grew in each MBS concentration tested. The type of *SSU1* promoter is represented in the last column according to the results of PCR amplification.
- **Table S3.3:** List of primers used in this study.

Supplementary files Chapter 4

- **Table S4.1:** Subtelomeres across *Saccharomyces* species. This table is divided into different sheets: Gene boundaries: subtelomere gene boundaries showed by the systematic name of the gene selected as boundary; Coordinates: subtelomere boundaries positions in the assemblies, positions were calculated from the gene start and gene end when they were acting as boundaries of the left arm and right arm subtelomeres respectively; Lengths: subtelomeric lengths (bp); Gene densities: subtelomeric gene densities showed in terms of number of annotated genes in each subtelomeric region.
- **Table S4.2:** Functional annotation of subtelomeric genes. Subtelomeric gene families resulting from the clustering method. Each row correspond to one gene. Gene name column shows which family the gene belongs (Cx) and the corresponding gene name in each assembly.
- **Table S4.3:** Normalised copy number values of each subtelomeric gene family (cluster).
- **Table S4.4:** Core genes GO enrichment. GO enrichment results of the 4950 core genes defined in pangenome. Tables: biological process, molecular function and cellular component.