Escuela Técnica Superior de Ingeniería

Departamento de Informática

Programa de doctorado en Tecnologías de la Información, Comunicaciones y Computación

TESIS DOCTORAL

SOLUCIONES BIOINFORMÁTICAS PARA EL ANÁLISIS DE DATOS ÓMICOS, DESCUBRIMIENTO DE CONOCIMIENTO Y DIAGNÓSTICO GENÉTICO EN *SPARUS AURATA* Y OTROS ORGANISMOS BIOLÓGICOS

Autora

Beatriz Soriano Salvador

Directores

Carlos Llorens Candela

Jaume Pérez Sánchez

Vicente Arnau Llombart

Enero 2023

Esta tesis ha sido financiada por el Ministerio de Ciencia e Innovación a través de la ayuda “DI-17-09134” para contratos para la formación de doctores en empresas (Doctorados Industriales)

AGRADECIMIENTOS

En el curso de este proyecto de Tesis, he tenido la suerte de contar con el apoyo de diversas personas. A todas ellas quiero agradecer su ayuda durante los años de trabajo en esta tesis doctoral y, sin las cuales, no podría haber finalizado.

En primer lugar, a mis codirectores de tesis, los doctores Carlos Llorens, Jaume Pérez y Vicente Arnau, por la confianza brindada y la guía constante que ha hecho posible el desarrollo de este trabajo. A Carlos quería agradecer, especialmente, la paciencia que ha tenido en muchas ocasiones, su gran ayuda y el haberme ofrecido la oportunidad de llevarla a cabo.

Esta tesis ha sido posible gracias a la inestimable colaboración de los siguientes equipos de investigación que han contribuido con datos biológicos para realizar los análisis que se presentan en este trabajo y que han servido para diseñar, poner a punto y validar las herramientas bioinformáticas generadas en esta tesis: Dra. Celia Perales y Dra. Maria Eugenia Soria de la Fundación Jiménez Díaz, Dr. Alfonso Navas del Museo Nacional de Ciencias Naturales de CSIC, Dra. Ana Oleaga y Dr. Ricardo Pérez-Sánchez del Instituto de Recursos Naturales y Agrobiología de Salamanca, y Dr. José Vicente Bagán del Hospital General Universitario de Valencia.

Este trabajo tampoco habría sido posible sin mis compañeros de Biotechvana, a los que les agradezco su ayuda en todos los campos y por generar un ambiente de trabajo inmejorable.

Quiero agradecerles a mis padres sus consejos, su cariño, su dedicación y todos los esfuerzos hechos para que haya podido llegar a realizar este trabajo. A ellos y también a mi hermana, Helena, les agradezco su apoyo incondicional en absolutamente todo lo que he podido necesitar y el arrojar luz cuando, en algunas ocasiones, lo veía todo oscuro.

A mi pareja, mejor amigo y futuro marido, Will. Por estar siempre a mi lado en los momentos más complicados, por tu amor, tu comprensión, tu paciencia y la ayuda que me ofreces día a día. No creo que hubiese podido llevar a cabo esta tesis sin ti.

Mi último agradecimiento es para el Ministerio de Ciencia e Innovación, por su financiación a la tesis doctoral que me ha permitido finalizar la formación de investigadores en empresas (Doctorados Industriales, DI-17-09134). Con la petición de que el presupuesto destinado a este fin se vea incrementado y pueda llegar al máximo de estudiantes.

RESUMEN

Con el incremento de datos generados mediante el uso de las tecnologías de secuenciación, es necesario el diseño de protocolos y herramientas que permitan el análisis y la integración de los mismos con el objetivo de entender y comprender los sistemas biológicos que forman parte de cada estudio en particular. Estas herramientas, además, es preferible que sean intuitivas y de fácil manejo para los usuarios, de manera que puedan ser utilizadas por cualquier investigador y no solamente por aquellos que sean expertos o tengan un conocimiento más avanzado en el campo de la bioinformática.

En esta tesis se presentan una serie de herramientas destinadas al análisis de datos procedentes de secuenciación para dar soporte a los estudios llevados a cabo en colaboración con distintas instituciones como son el Instituto de Acuicultura Torre de la Sal (IATS-CSIC), el Instituto de Recursos Naturales y Agrobiología de Salamanca (IRNASA), la Fundación Jiménez Díaz, el Museo Nacional de Ciencias Naturales de CSIC y el Hospital General Universitario de Valencia.

En primer lugar, se desarrolló e implementó un flujo de trabajo para el ensamblaje de *novo* y anotación de genomas eucariotas ricos en duplicaciones, el cual se incluyó en la herramienta DeNovoSeq. Este protocolo fue testado ensamblando de *novo* y anotando el genoma de *Sparus aurata* (dorada) a partir de datos de secuenciación aportados por el IATS-CSIC. Como resultado, este protocolo no solo permitió la obtención de un borrador de alta calidad del genoma de la dorada y con un tamaño (1,24 Gb) más próximo al esperado según el análisis de k-mer realizado, sino que permitió establecer una hipótesis sobre el origen de la mayor parte de las expansiones sufridas por el genoma de esta especie, sugiriendo que estas derivan de las actividades de los elementos genéticos móviles y de la respuesta inmunitaria como procesos para la adaptabilidad de la especie.

En segundo lugar, se rediseñó y adaptó un *pipeline*, llamado VQS-haplotyper, a partir de un *pipeline* creado por Mercedes Guerrero-Murillo y Josep Gregori i Font, y basado en el paquete de R llamada QSutils, para la identificación y cuantificación de cuasiespecies en muestras procedentes de pacientes infectados por un virus concreto. En este caso, se utilizaron muestras de pacientes infectados por el virus SARS-CoV-2 proporcionadas por la Fundación Jiménez Díaz. La modificación del *pipeline* original nos permitió obtener los cambios de nucleótidos y deleciones que caracterizaban los haplotipos presentes en las muestras para una abundancia relativa mínima de 0,5% y 0,1%, obteniendo un total de 105 y 1.173 mutaciones y/o deleciones, respectivamente. De esta manera, VQS-haplotyper es capaz de detectar pequeños cambios en la secuencia de un virus que pueden influir en las características del mismo.

En tercer lugar, se desarrollaron e implementaron dos protocolos para el análisis de datos RNA-seq, incluyendo el análisis de enriquecimiento de términos GO y rutas metabólicas, uno a partir de datos procedentes de secuenciación de *novo*, es decir, sin genoma de referencia disponible, y otro a partir de datos procedentes de

resecuenciación, es decir, con genoma de referencia disponible. Estos dos protocolos se implementaron para dar soporte a diversos estudios utilizando muestras de las siguientes especies: *Ornithodoros erraticus* y *Ornithorodos moubata*, proporcionadas por el IRNASA; *Anisakis pegreffii*, *Anisakis simplex s.s.* y sus híbridos, proporcionadas por el Museo Nacional de Ciencias Naturales de CSIC; *Homo sapiens*, proporcionadas por el Hospital General Universitario de Valencia. Complementariamente al protocolo para el análisis RNA-seq sin genoma de referencia disponible, ha sido necesario desarrollar un protocolo para el ensamblaje de *novo* de transcriptomas consenso contra el que, posteriormente, se mapeen las lecturas. Ambas implementaciones han permitido conocer las diferencias entre distintas condiciones de estudio o entre distintas especies. Con ello, es posible establecer potenciales antígenos que puedan ser diana para posibles terapias o vacunas, dilucidar posibles relaciones y diferencias entre distintas especies o descubrir biomarcadores de, por ejemplo, cáncer.

Por último, en colaboración con el IATS-CSIC, desarrollamos SAMBA (*Structure-Learning of Aquaculture Microbiomes Using a Bayesian-Network Approach*), una implementación informática de un modelo de red bayesiano para investigar cómo se relacionan entre sí los pan-microbiomas de los peces y todas las demás variables de un sistema acuícola concreto. SAMBA se basa en un modelo entrenable de red bayesiana que aprende la estructura de red de un sistema de acuicultura utilizando información de distintas variables bióticas y abióticas de importancia en la piscicultura, con especial atención a los datos microbianos proporcionados por la secuenciación de amplicones 16S. SAMBA acepta variables tanto cualitativas como cuantitativas y trata de forma convincente las diferencias en la composición microbiana derivadas de la variación técnica o biológica entre microbiomas de distintos especímenes. Para ello, SAMBA contiene una variedad de herramientas para preanalizar los datos y elegir una distribución para construir y entrenar el modelo de red bayesiana. Una vez creado y validado el modelo, el usuario puede interrogarlo y obtener información sobre el sistema modelizado en dos modos diferentes: “Report” y “Prediction”. En el modo “Report”, SAMBA informa de cómo el pan-microbioma y todas las demás variables que intervienen en el sistema de acuicultura modelizado se influyen mutuamente y cuáles son las probabilidades de cada relación. En el modo “Prediction”, la aplicación predice cómo cambiarían probablemente la diversidad y el perfil funcional del pan-microbioma en función de cualquier cambio realizado en otras variables. Finalmente, SAMBA implementa un completo editor gráfico de redes que permite navegar, editar y exportar los resultados. El funcionamiento de SAMBA ha sido testado y validado utilizando estándares de comunidades microbianas y comunidades de microbiota intestinal de doradas de piscifactoría (*Sparus aurata*) procedentes de diferentes ensayos de alimentación, arrojando un valor de precisión en todos los casos superior al 0,62.

En definitiva, esta tesis ha contribuido no solamente con el desarrollo de protocolos y herramientas que permiten y facilitan el análisis y la integración de diferentes datos NGS, sino que, además, ha contribuido con nuevo conocimiento biológico en diversos campos de estudio.

ABSTRACT

With the increase in data generated through the use of sequencing technologies, there is a need to design protocols and tools that allow the analysis and integration of these data in order to understand and comprehend the biological systems that are part of each particular study. These tools, moreover, should preferably be intuitive and user-friendly, so that they can be employed by any researcher and not only by those that are experts or have a more advanced knowledge in the field of bioinformatics.

This thesis presents a series of tools for the analysis of sequencing data so as to provide support for the studies carried out in collaboration with different institutions such as Instituto de Acuicultura Torre de la Sal (IATS-CSIC), Instituto de Recursos Naturales y Agrobiología de Salamanca (IRNASA), Fundación Jiménez Díaz, Museo Nacional de Ciencias Naturales de CSIC and Hospital General Universitario de Valencia.

Firstly, a workflow for de novo assembly and annotation of duplication-rich eukaryotic genomes was developed and implemented, and was included in the DeNovoSeq tool. This protocol was tested by de novo assembly and annotation of the *Sparus aurata* (gilthead sea bream) genome from sequencing data provided by the IATS-CSIC. As a result, this protocol not only allowed us to obtain a high quality draft of the gilthead sea bream genome with a size (1.24 Gb) closer to the expected size according to the k-mer analysis performed, but also allowed us to establish a hypothesis about the origin of most of the expansions suffered by the genome of this species. This suggests that they derive from the activities of the mobile genetic elements and the immune response as processes for the adaptability of the species.

Secondly, a pipeline, called VQS-haplotyper, was adapted and redesigned from a pipeline created by Mercedes Guerrero-Murillo and Josep Gregori i Font, and based on the R package called QSutils, to identify and quantify quasispecies in samples from patients infected by a specific virus. In this case, samples from patients infected by the SARS-CoV-2 virus provided by the Fundación Jiménez Díaz were used. The modification of the original pipeline allowed us to obtain nucleotide changes and deletions that characterized the haplotypes present in the samples for a minimum relative abundance of 0.5% and 0.1%, obtaining a total of 105 and 1,173 mutations and/or deletions, respectively. In this way, VQS-haplotyper is able to detect small changes in the sequence of a virus that can influence its characteristics.

Thirdly, two protocols were developed and implemented to analyze RNA-seq data, including the enrichment analysis of GO terms and metabolic pathways, one from de novo sequencing data, i.e. with no reference genome available, and the other from resequencing data, i.e. with reference genome available. These two protocols were implemented to provide support for several studies using samples from the following species: *Ornithodoros erraticus* and *Ornithodors moubata*, provided by the IRNASA; *Anisakis pegreffii*, *Anisakis simplex s.s.* and their hybrids, provided by the Museo Nacional de Ciencias Naturales de CSIC and *Homo sapiens*, provided by the Hospital

General Universitario de Valencia. Complementary to the protocol for RNA-seq analysis with no reference genome available, it has been necessary to develop a protocol for the de novo assembly of consensus transcriptomes against which, subsequently, the reads are mapped. Both implementations have provided insight into the differences between distinct study conditions or between distinct species. With this, it is viable to establish potential antigens that could be targets for possible therapies or vaccines, to elucidate possible relationships and differences between different species or to discover biomarkers of, for example, cancer.

Finally, in collaboration with the IATS-CSIC, we developed SAMBA (Structure-Learning of Aquaculture Microbiomes Using a Bayesian-Network Approach), a computer implementation of a Bayesian network model to investigate how fish pan-microbiomes and all other variables in a given aquaculture system are related to each other. SAMBA is powered by a Bayesian network trainable model that learns the network structure of an aquaculture system using information from distinct biotic and abiotic variables of importance in fish farming, with special focus on microbial data provided from 16S amplicon sequencing. SAMBA accepts both qualitative and quantitative variables and convincingly deals with the differences in microbial composition derived by the technical or biological variation among microbiomes of distinct specimens. To this end, SAMBA is implemented with a variety of tools to pre-analyze the data and choose a distribution to build and train the Bayesian network model. Once the model has been created and validated, the user can interrogate the model and obtain information about the modelled system in two different modes: Report and Prediction. Using the Report mode SAMBA reports how the pan-microbiome and all other variables involved in the modelled aquaculture system influence each other and what the conditional probabilities of each relation are. Under the Prediction mode, the application predicts how the diversity and functional profile of the pan-microbiome would likely change depending on any alteration made on other variables. Finally, SAMBA implements a comprehensive graphical network editor allowing the user to navigate, edit and export outcomes. The performance of SAMBA has been tested and validated using microbial community standards and gut microbiota communities of farmed gilthead sea bream (*Sparus aurata*) from different feeding trials, giving an accuracy value in all cases higher than 0.62.

In conclusion, this thesis has not only contributed to the development of protocols and tools that allow us to facilitate the analysis and integration of different NGS data, but has also contributed with new biological knowledge in various fields of study.

TABLA DE CONTENIDOS

1. INTRODUCCIÓN

1.1. Sobre las técnicas de secuenciación masiva como fuentes de información biológica en disciplinas como la genómica y la transcriptómica

En la actual era post-genómica, llamada así en referencia a la consecución del primer borrador del genoma humano, existen dos disciplinas básicas de secuenciación; una orientada a generar nuevo conocimiento ómico a partir de aquellas especies modelo cuyo genoma o transcriptoma no ha sido secuenciado todavía y otra orientada a volver a secuenciar especies ya previamente secuenciadas con el fin de caracterizar diferencias entre distintos grupos a estudio de la misma especie (pueden ser variedades, razas, poblaciones, casos a estudio, etc.). La primera disciplina es popularmente conocida como secuenciación de *novo*. En tanto que la otra disciplina se conoce como resecuenciación. En estas líneas, las actuales tecnologías de secuenciación masiva (referidas como NGS por el acrónimo en inglés de *Next Generation Sequencing*) permiten secuenciar genomas y transcriptomas con mayor eficiencia y profundidad que las tecnologías tradicionales de secuenciación por electroforesis capilar de Sanger, que es considerada como de primera generación.

Las tecnologías NGS han abierto una nueva era en la genómica y la biología molecular al proporcionar mayor rendimiento de datos con un costo menor y niveles de profundidad (*high throughput*) nunca vistos anteriormente, lo cual permite la investigación del genoma a escala poblacional. Existen tres principales ventajas en el uso de tecnologías NGS frente a las tecnologías de primera generación. Primero, los métodos NGS no requieren un procedimiento de clonación, permitiendo preparar librerías para la secuenciación en un sistema libre de células. Segundo, las tecnologías NGS procesan millones de reacciones de secuenciación en paralelo al mismo tiempo. Tercero, la detección de bases se realiza de forma cíclica y en paralelo (Park and Kim, 2016).

La reducción de los costes de secuenciación NGS ha facilitado la translación de esta tecnología más allá de la academia, llegando progresivamente a campos como la agricultura, la acuicultura, la biotecnología y la biomedicina (OPATHY Consortium and Gabaldon, 2019).

Actualmente, hay una cierta oferta de metodologías NGS, todas ellas en continuo desarrollo en pro de aumentar el rendimiento, la longitud de las lecturas y su precisión. Las tecnologías actuales van desde la secuenciación por síntesis, implementada por Illumina, hasta la secuenciación basada en nanoporos, que permite la obtención de lecturas largas (hasta 1 Mb), pero con menos precisión y rendimiento.

La aplicabilidad de las tecnologías NGS, entre otras, permite las siguientes aproximaciones ómicas (Park and Kim, 2016):

- Reconstruir genomas pertenecientes a especies que se caracterizan por vez primera y para las que aún no se tiene una secuencia previa de referencia.

- Caracterizar *small* RNAs presentes en el genoma de diferentes especies.

- Caracterizar, mediante resecuenciación, la diversidad genética de un organismo para el que existe un genoma de referencia, obteniendo variaciones de una única posición nucleotídica (SNPs, acrónimo en inglés de *Single Nucleotide Polymorphism*), variaciones estructurales, variaciones en el número de copia, etc. Para obtener este tipo de datos, se puede realizar secuenciación de DNA, secuenciación de RNA o secuenciación del epigenoma.

- Caracterizar los patrones de expresión dados resultantes de la secuenciación de RNA de una muestra dada. Esto permite examinar el *splicing* del RNA (en el caso de eucariotas) y realizar análisis comparativos entre diferentes casos a estudio mediante expresión y enriquecimiento diferencial.

- Examinar y comparar los patrones de metilación del DNA entre distintas muestras a partir de estudios del epigenoma y de sistemas regulatorios del genoma. A este fin, hay tres tecnologías NGS disponibles: ChIP-Seq, que permite el análisis de las interacciones proteína-DNA; la secuenciación bisulfito, que permite detectar únicamente las citosinas metiladas mediante su transformación en uracilos; y MeDIP-Seq, consistente en aislar fragmentos de DNA metilados utilizando un anticuerpo contra la 5-metilcitosina (5mC), obteniendo los fragmentos mediante inmunoprecipitación.

- Estudiar las comunidades microbianas de muestras de distintas procedencias.

- Identificar nuevos patógenos.

En esta tesis, se ha usado datos procedentes de diferentes experimentos de ómica, tanto de *novo* como de resecuenciación (incluyendo transcriptómica, genómica y metagenómica), obtenidos a partir de diferentes organismos para desarrollar nuevos protocolos y herramientas bioinformáticas.

El primero de ellos es *Sparus aurata* (dorada). La dorada es un pez costero marino templado que pertenece a la familia *Sparidae*, orden Perciformes (Pérez-Sánchez et al., 2019). La dorada es una especie hermafrodita protándrica, ya que madura como macho durante su primer y segundo año, pero la mayoría de los individuos cambian a hembras entre el segundo y el cuarto año de vida (Zohar et al., 1978). La plasticidad ecológica de la dorada se evidencia por el amplio rango de temperatura (5 – 33°C) y salinidad (3—70%) que es capaz de soportar (Kleszczyńska et al., 2006; Ortega, 2008). Estas características euritermales y eurihalinas, en combinación con una notable resistencia a los factores estresantes de la acuicultura, hacen de esta especie un pez único con una gran plasticidad para la cría y los entornos desafiantes (Pérez-Sánchez et al., 2019). Debido a estas características, esta especie de pez es económicamente importante y altamente cultivada en el área del Mediterráneo, con una producción anual de más de 262.000 toneladas métricas en 2020, concentrada en Turquía, Grecia, Egipto y España (FAO, 2022), y siendo la tercera especie más cultivada en Europa (APROMAR, 2020). La dorada habita naturalmente en el Mediterráneo y Atlántico oriental, desde las Islas Británicas y el Estrecho de Gibraltar hasta Cabo Verde y las Islas Canarias, lo que respalda estudios previos sobre estructura genética que muestran una fuerte subdivisión

genética entre las poblaciones atlánticas y mediterráneas (Alarcón et al., 2004; De Innocentiis et al., 2004). Curiosamente, también se han encontrado fuertes subdivisiones a corta distancia a lo largo de las costas de Túnez (Ben Slimen et al., 2004) o entre la costa francesa y la argelina (Chaoui et al., 2009). Sin embargo, el flujo de genes ocurre sin restricciones a lo largo de la costa de Italia, en ausencia de barreras físicas y ecológicas entre los mares Adriático y Mediterráneo (Franchini et al., 2012). El hecho de que esta especie tenga una gran resistencia a factores estresantes y una gran plasticidad que le permite adaptarse a diferentes temperaturas y salinidades, junto con el hecho de que se trata de una especie altamente cultivada y de gran importancia económica, hace de la dorada un organismo interesante para el estudio y la implementación de protocolos y herramientas que permitan la mejora de su producción. A partir de esta especie, se ha obtenido datos de genómica, para el desarrollo de un protocolo de ensamblaje de *novo* y su posterior predicción y anotación de genes y moviloma, y datos de metagenómica, para el desarrollo de una herramienta que establezca asociaciones entre los taxones y variables experimentales dadas.

Otras dos especies utilizadas en esta tesis han sido *Ornithodoros erraticus* y *Ornithodoros moubata*, dos especies de garrapatas, las cuales se utilizaron para el desarrollo de un protocolo de RNA-seq de *novo* a partir de datos de transcriptómica, al tratarse de dos especies para las que todavía no hay un genoma de referencia disponible. Las garrapatas son ectoparásitos hematófagos de importancia médica y veterinaria en todo el mundo porque causan lesiones directas a sus huéspedes y transmiten una gran variedad de patógenos que afectan a los animales salvajes y domésticos, a las mascotas y a los humanos, causando importantes pérdidas económicas a nivel mundial (Pérez-Sánchez et al., 2021). Ambas especies pertenecen a la familia de los argásidos, una de las tres familias de garrapatas, que se caracteriza por una duración corta en cuanto a alimentación de sangre lo que hace que, junto con su especialización en microhábitats protegidos, su papel en la salud humana y animal sea, generalmente, ignorado (Hoogstraal, 1985). Sin embargo, estas especies pueden causar toxicosis, parálisis, irritación, alergias y desangrado (Vial, 2009).

También para el desarrollo de este protocolo de RNA-seq de *novo* a partir de la secuenciación de transcritos, se utilizan diversas especies de Anisakis (*Anisakis simplex s.s.*, *Anisakis pegreffii* e híbridos de ambas). Anisakis es un género de nematodos que parasita mamíferos marinos, peces, moluscos y crustáceos. Estos nematodos completan su ciclo vital en el estómago de los cetáceos y, con menos frecuencia, de los pinnípedos, que se infectan después de devorar huéspedes que albergan las larvas infectivas (Van Thiel et al., 1960). Los humanos también pueden ser infectados al consumir pescado crudo o poco cocinado, o carne de cefalópodos, pudiendo causar anisakiasis cuando las larvas penetran en el tracto gastrointestinal, la cual se caracteriza por manifestaciones gastrointestinales agudas de epigastralgia, náuseas, dolor abdominal y diarrea (Llorens et al., 2018). La exposición de los humanos a especies de Anisakis puede provocar también reacciones alérgicas como angiodema, urticaria y anafilaxis (Audicana and Kennedy, 2008), incluso aunque el pescado haya sido congelado o debidamente cocinado. La anisakiasis se perfila como un importante problema epidemiológico. Se han informado de más de 20.000 casos de anisakiasis en todo el mundo desde 1960 (Shweiki et al., 2014), con una mayor incidencia en áreas como Japón, los Países Bajos, Francia,

España, Alemania y California, donde tradicionalmente se come pescado crudo o es cada vez más habitual.

Para el diseño y desarrollo de un protocolo RNA-seq a partir de datos de resecuenciación, se utilizaron muestras humanas procedentes de pacientes con leucoplasia verrucosa proliferativa (PVL) y de pacientes sanos con el objetivo de investigar la biología molecular de esta enfermedad y determinar por qué esta enfermedad presenta una alta tasa de transformación maligna a un carcinoma oral de células escamosas (OSCC) (Llorens et al., 2021).

El último organismo utilizado ha sido el betacoronavius SARS-CoV-2 para el desarrollo de un protocolo de detección de haplotipos a partir de datos procedentes de secuenciación. El virus SARS-CoV-2 surgió en la población humana en 2019, y es el agente causante de la actual enfermedad pandémica COVID-19 (Huang et al., 2020). La detección de haplotipos o variantes en este virus es importante para la detección de patrones mutacionales que tengan influencia en características como la transmisibilidad y para la detección de mutantes de escape viral que limiten la eficacia de los agentes inmunitarios y antivirales (Martínez-González et al., 2022).

1.2. Sobre la bioinformática orientada al análisis de datos NGS

Debido a la universalización del NGS en la investigación biológica, su uso se ha ido expandiendo en los distintos laboratorios de investigación, tanto empresariales como académicos. Esta expansión se ha acompañado de un aumento significativo en el desarrollo de técnicas, protocolos y herramientas bioinformáticas necesarias para extraer e interpretar la información biológicamente significativa de las muestras generadas a través de las tecnologías NGS (Holzinger et al., 2014). Los protocolos bioinformáticos varían, en primer lugar, en función de si el proyecto es de *novo* o de si es un proyecto de resecuenciación. Por ejemplo, si hablamos de un estudio RNA-seq, consistente en el análisis de expresión diferencial entre diferentes grupos a estudio, en un proyecto de resecuenciación se utiliza el genoma de referencia para mapear las lecturas y, posteriormente realizar el análisis de expresión diferencial. Sin embargo, en el caso de un proyecto de *novo*, es necesario hacer un ensamblaje del genoma o del transcriptoma que sirvan de referencia para hacer este análisis.

Entre las distintas aplicaciones bioinformáticas que han surgido orientadas al análisis ómico basado en datos NGS (genómica, transcriptómica, metagenómica, etc.) encontramos lo siguiente:

- Ensambladores que permitan la reconstrucción de *novo* de genomas y transcriptomas cuya secuencia se caracteriza por vez primera sin información de referencia previa. Estos ensambladores se encargan de integrar y juntar las lecturas generadas por las tecnologías NGS como si fuera un rompecabezas al alinear regiones con superposición para construir la secuencia del genoma o del transcriptoma (Park and Kim, 2016).

- Herramientas para la identificación y predicción de *small* RNAs en genomas mediante procesos de anotación por homología o de elucidación de su estructura secundaria y/o terciaria.

- Software para el análisis de variantes mediante un proceso denominado llamada de variantes, que permiten obtener todo tipo de variantes presentes en las secuencias y compararlas entre distintas condiciones experimentales, entre individuos de una población o entre distintas cepas de un mismo organismo. Una vez conocidas las variantes, existen herramientas para predecir sus efectos moleculares en los genes y proteínas, anotando las variantes que previamente se han descrito en la bibliografía.

- Protocolos para el análisis diferencial de genes entre dos condiciones dadas en una célula, tejido u organismo. Esto es una aproximación importante para descifrar la fisiología molecular de la célula. Este tipo de protocolos suelen incluir cuatro pasos: a) análisis de calidad y preprocesado de las muestras; b) mapeo de las lecturas contra el genoma de referencia; c) cuantificación de los niveles de expresión de genes y transcritos; d) identificación de genes y transcritos específicos que están diferencialmente expresados entre condiciones; e) identificación de funciones y rutas metabólicas enriquecidas entre condiciones.

- Herramientas para el análisis de datos de metilación diferencial. En este caso, las herramientas para este tipo de análisis suelen estar implementadas en el lenguaje de programación R (R Core Team, 2022). Dependiendo de la tecnología de secuenciación utilizada, la herramienta a utilizar para el análisis de los datos es diferente. Sin embargo, todas estas herramientas permiten la detección de regiones diferencialmente metiladas entre condiciones siguiendo, en general, los siguientes pasos: a) análisis de calidad y preprocesado de las muestras; b) mapeo de las lecturas contra el genoma de referencia; c) cuantificación de los niveles de metilación para las regiones establecidas; d) identificación de las regiones que están diferencialmente metiladas entre condiciones.

- Protocolos y herramientas que permiten el análisis de las comunidades microbianas englobando la secuenciación de genomas, identificación de secuencias codificantes de proteínas, comparación de genomas para la identificación de las funciones de un gen, la inferencia de fenotipos a partir de los genotipos, etc. Estas herramientas pueden dividirse en dos grupos generales: i) técnicas para la búsqueda y alineamiento de secuencias que compare un nuevo genoma contra un conjunto de genes conocidos, de manera que se anota la estructura y función de estos genes; ii) técnicas como la minería de datos, análisis estadísticos, redes neuronales, redes bayesianas, algoritmos genéticos y técnicas de comparación de gráficos para identificar patrones comunes, características y funciones de alto nivel (Bansai, 2005).

En conclusión, el desarrollo de las distintas técnicas bioinformáticas actuales han permitido incrementar el uso del NGS en la investigación biológica y clínica (Park and Kim, 2016). Sin embargo, la bioinformática (y la biología computacional) no solamente se enfrenta al volumen de datos, sino también a la creciente necesidad de análisis y

modelos integrativos que permitan interpretar la complejidad de los mismos bajo una visión de conjunto (Holzinger et al., 2014). Todos estos avances han permitido realizar mejoras sustanciales en las técnicas de procesado, ensamblaje/mapeo y anotación de datos generados *de novo* mediante herramientas y bases de datos que permitan desarrollar nuevo conocimiento en la forma de herramientas genómicas y computacionales (anotaciones de perfiles, secuencias, etc.). Dichos protocolos, no obstante, son complejos de usar y requieren de expertos bioinformáticos e innumerables horas dedicadas, sobre todo si hablamos de genomas eucariotas. Además, dichos protocolos consisten en flujos de trabajo o *pipelines* que, normalmente y debido a su complejidad, implican la necesidad de conocimientos de comandos avanzados en entornos Unix. Un *pipeline* consiste en un conjunto de elementos de procesado de datos conectados en serie, donde el resultado de uno de los elementos es introducido en el siguiente.

Debido a la dificultad que conlleva ejecutar y trabajar con *pipelines* y flujos de trabajo utilizando comandos, durante los últimos años han aparecido soluciones que permiten realizar los análisis usando interfaces ofreciendo así al usuario herramientas amistosas para la realización o ejecución de diversos análisis. Estas herramientas se ejecutan en el ordenador del usuario y son sencillas de utilizar cuando se trata de administrar tareas específicas, pero no tienen la versatilidad y variedad de opciones y parámetros que ofrecen los protocolos que se ejecutan por comandos. Una excelente solución a este dilema es el uso de estos protocolos que se ejecutan mediante comandos a través de una interfaz gráfica (Hafez et al., 2022). Con este objetivo nacieron iniciativas como el proyecto Galaxy (Afgan et al., 2018) o el GPRO Suite (Futami et al., 2011), un proyecto bioinformático de Biotechvana que proporciona soluciones personalizadas de interfaz gráfica de usuario (GUI) para el análisis de datos ómicos en servidores remotos (la nube) o en el ordenador del usuario. El proyecto GPRO será marco de implementación de algunos de los desarrollos realizados en esta tesis.

1.3. Nuevas aproximaciones y enfoques para el análisis de datos NGS a partir de la biología integrativa y de sistemas e inteligencia artificial

Hoy en día, la bioinformática tiene un papel esencial en el desciframiento e integración de datos ómicos generados por tecnologías de alto rendimiento (NGS incluido). La evolución de esta disciplina se dirige ahora hacia áreas emergentes como la genómica integradora y traslacional y, en última instancia, hacia la medicina personalizada (Molidor et al., 2003). Esto es, los genes y los productos génicos constituyen sistemas moleculares complejos interconectados y que interaccionan de múltiples maneras. La comprensión de estos sistemas, sus interacciones y sus propiedades requiere información multidisciplinar que es contribuida a partir de varios campos, como la genómica, la proteómica, la metabolómica o los perfiles sistemáticos de fenotipos a nivel de células y organismos, entre otros (Molidor et al., 2003). Este hecho pone de manifiesto la importancia de la biología integrativa y de sistemas como herramienta de investigación para entender o comprender los sistemas biológicos. Sin embargo, este objetivo se ve dificultado por nuestra heterogeneidad a la hora de almacenar, procesar y entender el conocimiento biológico (Zhang et al., 2011a).

En las últimas décadas, se ha experimentado un importante incremento en la emergencia de bases de datos online para la gestión de dicho conocimiento. Véase, la IUBMB Enzyme Nomenclature (http://www.chem.qmul.ac.uk/iubmb/enzyme/), ENZYME de Expasy (Gasteiger et al., 2003), BRENDA (Placzek et al., 2017), KEGG (Kotera et al., 2012), METACyc (Caspi et al., 2016), así como las orientadas al diagnóstico genético como COSMIC en Cáncer (Forbes et al. 2015), y OMIM en desórdenes genéticos (https://www.omim.org). La existencia de una amplia variedad y diversidad de bases de datos implica que hay también maneras diferentes de representar datos similares, lo que dificulta su integración y procesamiento con el fin de obtener una visión en conjunto de dichos datos. Cuando se trata de investigación biológica, es crucial crear, adoptar e implementar estándares biológicos, ya que sin ellos es prácticamente imposible conseguir la integración de datos (Kher et al., 2010; Mathew et al., 2007). Por estándar, se entiende como un término o estructura que representa una entidad biológica, es decir, que representa cualquier tipo de unidad de información biológica (Lapatas et al., 2015). Los estándares facilitan la reutilización de datos, así como su intercambio. Además, ayudan a superar las dificultades relacionadas con la interoperabilidad entre diferentes formatos de datos, arquitecturas, nomenclaturas e infraestructuras. La ausencia de estándares implica una pérdida sustancial de productividad y menos datos disponibles para los investigadores. Con un conjunto único de estándares comunes, es posible crear herramientas para integrar datos mediante diferentes aproximaciones (Lapatas et al., 2015).

Estas aproximaciones suelen está dirigidas a datos ya conocidos y presentes en diferentes bases de datos. Pero, cuando se realiza un experimento *de novo* o estamos trabajando con datos no presentes en las bases de datos, no es posible realizar la integración de datos utilizando estas aproximaciones. Es por ello por lo que surge la necesidad de construir modelos que integren las diferentes capas de datos ómicos que puedan generarse en los estudios biológicos. A partir de este punto surge la biología integrativa y de sistemas como un área de investigación multidisciplinar que pretende describir el funcionamiento de un organismo mediante la construcción y análisis de diversas capas ómicas que permitan desgranar la complejidad de dichos organismos entendida como un todo. Estas capas están ligadas unas a otras e ilustran las interacciones que ocurren dentro de la célula, dentro de los órganos y, a su vez, en los propios individuos (Hanafi et al., 2019). Está ampliamente aceptado que el entendimiento integral de un sistema biológico solo puede provenir de un análisis en conjunto de todas estas capas ómicas (Joyce and Palsson, 2006; Gómez-Cabrero et al., 2014).

Un punto de partida de este análisis es utilizar el concepto matemático de redes para representar capas ómicas. Una red o grafo consta de nodos (o vértices) y enlaces/asociaciones (o aristas). En las redes biológicas, los nodos suelen representar entidades biológicas discretas a nivel molecular (por ejemplo, genes, proteínas, metabolitos, fármacos, taxones, etc.) o fenotípico (por ejemplo, enfermedades), mientras que los enlaces o asociaciones representan relaciones físicas, funcionales o químicas entre pares de entidades (Vidal et al., 2011). En las últimas dos décadas, las redes han sido una de las herramientas matemáticas más ampliamente utilizadas para el modelado y análisis de datos ómicos (Aittokallio and Schwikowski, 2006). Particularmente, estas herramientas pueden aplicarse al estudio de redes de interacción

proteína-proteína (PPI), redes de interacción génica (GI), redes de interacción de metabolitos (MI), redes de interacción microbiana y redes de coexpresión génica (Co-Ex), extrayendo información biológica valiosa a partir de los diferentes tipos de maquinaria molecular que se encuentran dentro de la célula. Sin embargo, una comprensión más completa de un sistema biológico se puede conseguir mediante un análisis conjunto e integrativo de todos estos tipos de redes (Gligorijević and Pržulj, 2015). En función del tipo de datos que se quiera integrar, los métodos de integración de redes pueden dividirse en homogéneos o heterogéneos. La integración homogénea lidia con la integración de redes que están formadas por el mismo tipo de nodos (por ejemplo, proteínas o genes), pero diferentes tipos de enlaces entre estos nodos. Sin embargo, la mayoría de los datos biológicos son heterogéneos, consistiendo en varios tipos de entidades biológicas y varios tipos de relaciones. Estos datos pueden representarse como colecciones de redes interrelacionadas con varios tipos de nodos y enlaces. Por tanto, la integración heterogénea lidia con la minería colectiva de estas redes y con la construcción de un modelo unificado (Gligorijević and Pržulj, 2015).

Las estrategias para la integración de datos se dividen en las siguientes categorías:

- Temprana o completa: combina diferentes conjuntos de datos en uno solo para la construcción del modelo. Este tipo de aproximación normalmente requiere una transformación de cada conjunto de datos en una representación común, lo que deriva, en algunos casos, en pérdida de información (Lanckriet et al., 2004; Žitnik and Župan, 2015).

- Tardía o decisión: crea un modelo para cada conjunto de datos y luego los combina en un modelo unificado. La construcción de modelos a partir de cada conjunto de datos por separado ignora sus relaciones mutuas, lo que a menudo resulta en una reducción del rendimiento del modelo final (Gevaert et al., 2006; Žitnik and Župan, 2015).

- Intermedia o parcial: combina datos a través de la inferencia de un modelo conjunto. A menudo se ha preferido esta estrategia debido a su superior precisión predictiva, tal y como reportan muchos estudios (Lanckriet et al., 2004; Gevaert et al., 2006; van Vilet et al., 2012; Žitnik and Župan, 2015; Pavlidis et al., 2002), pero hay algunos estudios que reportan una superioridad de las estrategias temprana y tardía frente a la intermedia (Ozen et al., 2009). Esta estrategia no requiere la transformación de ningún dato, lo que deriva en que no haya pérdida de información.

La mayoría de los métodos basados en redes utilizan maneras simples de integrar diferentes tipos de datos y crear una representación o modelo integrado de un conjunto de redes. Por ejemplo, en una red de integración homogénea, una manera de construir una red integrada es mediante la fusión de enlaces de todas las redes que contengan el mismo conjunto de nodos. Esto se consigue mediante una simple suma de las matrices de adyacencia que representan cada una de las redes individuales. Sin embargo, este enfoque pasa por alto los problemas de compatibilidad entre redes individuales en la construcción de la red integrada (Gligorijević and Pržulj, 2015). Otros enfoques que intentan superar esta desventaja crean una suma ponderada de matrices de adyacencia

para construir la matriz de adyacencia de la red integrada (Mostafavi et al., 2008; Mostafavi and Morris, 2012; Chen et al., 2013). Los coeficientes de ponderación se obtienen resolviendo un problema de regresión lineal, que asigna pesos más bajos a las redes "menos importantes". Sin embargo, dicha ponderación depende del problema, es decir, la estructura de la red integrada resultante depende del problema biológico que se encuentra en estudio (Gligorijević and Pržulj, 2015).

En una red de integración heterogénea, la mayoría de los estudios han integrado redes que contienen diferentes tipos de nodos y enlaces aplicando métodos simples de proyección (Sun et al., 2014; Davis and Chawla, 2011; Goh et al., 2007). Es decir, proyectan capas de un red en otra que es de interés. No obstante, este método de proyección a menudo da como resultado pérdida de información. Por ejemplo, al proyectar una red de interacción gen-gen sobre una red de asociación enfermedad-enfermedad, la información sobre las conexiones génicas se pierde junto con toda la estructura de la red de interacción gen-gen. Por lo tanto, mediante el uso de estos métodos, no es posible analizar patrones de conectividad de enfermedades y genes simultáneamente (Gligorijević and Pržulj, 2015).

Métodos más sofisticados capaces de analizar patrones de conectividad de varias redes simultáneamente utilizan procesos de difusión. Estos métodos exploran simultáneamente la estructura y las relaciones mutuas de cada red al mismo tiempo, y, basándose en toda esta información, crean una red integrada (Gligorijević and Pržulj, 2015). Aunque estos métodos están principalmente diseñados para un par de redes interrelacionadas, algunos estudios extienden su uso para el manejo de más redes (Huang et al., 2013). A pesar de la posibilidad de manejar más redes, con la inclusión de múltiples redes, crece el número iteraciones necesarias para conseguir la difusión de información y, por lo tanto, aumenta el tiempo de ejecución del algoritmo, lo que hace que la escalabilidad de estos métodos sea limitada (Gligorijević and Pržulj, 2015).

Otro método basado en redes consiste en el uso de redes bayesianas, las cuales son modelos gráficos probabilísticos que combinan conceptos como la probabilidad y la teoría de grafos para representar y modelar las relaciones causales entre variables aleatorias que describen los datos (Sachs et al., 2005). Las redes bayesianas se representan mediante grafos acíclicos directos, donde los nodos representan las variables aleatorias y los enlaces directos representan las asociaciones y las probabilidades condicionales entre pares de variables. Por ejemplo, la distribución de probabilidad condicional (DPC) entre la variables X y la variable Y ($p(X|Y)$) representa la probabilidad de X dado el valor de Y. Las DPC pueden modelar dependencias condicionales entre variables discretas y continuas, o una combinación de ambas (Gligorijević and Pržulj, 2015). Las redes bayesianas se aplican en muchas tareas en el campo de la biología de sistemas, incluyendo la inferencia de relaciones en comunidades microbianas (Sazal et al., 2020), el modelado de rutas de señalización de proteínas (Sachs et al., 2005), la predicción de funciones génicas (Troyanskaya et al., 2003) y la inferencia de redes celulares (Friedman, 2004).

Recientemente, las redes bayesianas se han utilizado como un método adecuado para la integración y el modelado de varios tipos de datos biológicos. Uno de los mayores desafíos en biología de sistemas es un problema de inferencia de redes a partir de

fuentes de datos dispares (Rider et al., 2013). A pesar de ello, este tipo de redes juega un papel muy importante en la descripción y predicción de comportamientos complejos de un sistema respaldado por la evidencia de una variedad de datos biológicos diferentes (Schadt et al., 2009). Es decir, las redes bayesianas son un buen marco para la integración de redes biológicas debido a su capacidad para capturar la dependencia condicional entre variables de datos a través de la construcción de DPC. A pesar de este hecho, este tipo de método tiene ciertas desventajas consistentes en el tiempo computacional requerido para la construcción del modelo de red, dependiente principalmente del número de nodos que componen una red, en la imposibilidad de representar bucles, que son muy importantes en muchas redes biológicas, y en la captura de los enlaces o asociaciones más importantes, perdiendo parte de la información (Gligorijević and Pržulj, 2015).

Un último método también basado en redes consiste aquellos métodos basados en kernel. Estos métodos pertenecen a la clase de métodos estadísticos *machine learning*, utilizados para el análisis de patrones de datos, como puede ser tareas de aprendizaje, *clustering*, clasificación, regresión, correlación, selección de características, etc. (Schölkopf et al., 2004). La capacidad de estos métodos basados en Kernel para el manejo de datos estructurados complejos los coloca en una posición ideal para la integración de datos heterogéneos (Daemen et al., 2009). Los métodos que utilizan matrices de datos de kernel incluyen: máquinas de vectores de soporte (SVM) (Hearst et al., 1998), análisis de componentes principales (PCA) (Jolliffe, 2005) y análisis de correlación canónica (CCA) (Hardoon et al., 2004).

Las matrices ernel, que representan la similaridad entre todos los pares de puntos de datos, permiten representar diferentes tipos de datos (Borgwardt, 2011), aunque elegir el método kernel más adecuado no es sencillo. En la mayoría de ocasiones, en vez de construir una única matriz kernel, generalmente se construyen múltiples matrices utilizando diferentes medidas de similaridad (Gönen and Alpaydin, 2011). Estas matrices se combinan linealmente en una sola, la cual se utiliza para los posteriores análisis. Esta aproximación se llama *multiple kernel learning* (MKL) y se ha demostrado que consigue un mejor rendimiento que los métodos que utilizan una única matriz kernel, especialmente en el análisis de datos genómicos (Wang et al., 2014).

Los métodos basados en kernel se propusieron en un primer momento como una técnica para la integración de datos en el artículo de 2004 publicado por Lanckriet et al. En este artículo, se utiliza el método SVM para la clasificación de proteínas en proteínas de membrana o en proteínas ribosomales. También se ha demostrado el poder de los métodos basados en kernel para la integración de datos moleculares, estructurales y fenotípicos para la reutilización de fármacos (Napolitano et al., 2013) y para la integración de datos clínicos con datos genómicos (Daemen et al., 2007), tal y como ocurre con los métodos basados en redes bayesianas.

A partir de estos ejemplos, se puede ver que la integración de datos mediante el uso de métodos basados en kernel tiene varias ventajas. La principal es que pueden integrar una amplia variedad de tipos de datos. Además, una combinación lineal de diversas matrices kernel proporciona una forma selectiva de contabilizar conjuntos de datos al asignar pesos más bajos a conjuntos de datos menos informativos y más ruidosos. Por

el contrario, una gran desventaja es que los conjuntos de datos heterogéneos deben transformarse en un espacio de características comunes para integrarse correctamente, lo que puede conducir a la pérdida de información (Gligorijević and Pržulj, 2015).

En conclusión, la aplicación de cualquiera de estos métodos descrito anteriormente para la integración de datos puede dar una visión más global y en conjunto de datos procedentes de diferentes ómicas que no están suficientemente representadas en las bases de datos y/o que proceden de organismos para los que no se tiene información previa de referencia, como es el caso de *Sparus aurata* (o dorada), conduciendo al descubrimiento de conocimiento.

Dentro de este concepto de integración de datos para el descubrimiento de conocimiento se enmarca SAMBA (de las siglas en inglés *Structure-Learning of Aquaculture Microbiomes Using a Bayesian-Network Approach*) (Soriano et al., 2022), una aplicación desarrollada durante este trabajo en R que permite establecer asociaciones entre variables experimentales y taxones, y predecir el posible valor de estos taxones en unas condiciones dadas mediante el uso de redes bayesianas y utilizando datos metagenómicos procedentes de *Sparus aurata*.

2. OBJETIVOS

El objetivo general para este doctorado industrial ha consistido en escalar la plataforma bioinformática de Biotechvana para la computación de lado servidor (principalmente basada en la GPRO suite) con nuevos protocolos, flujos de trabajo y *pipelines* bioinformáticos implementados mediante interfaces gráficas de usuario (GUIs, del acrónimo en inglés de *Guide User Interface*) o por la línea de comandos. Todo ello orientado a facilitar la gestión y el análisis de datos ómicos (genoma, transcriptoma, metagenoma, secretoma, nutrigenoma, etc) y optimizar el descubrimiento del conocimiento y el diagnóstico genético. A este fin, se han usado distintos organismos modelo (con y sin referencia genómica previa) como fuente de datos biológicos. Principalmente y mayoritariamente, hemos usado la dorada (*Sparus aurata*) como organismo modelo de especial interés en nutrogenómica y cuyo genoma ha sido reconstruido en esta tesis en colaboración con el Instituto de Acuicultura Torre de la Sal (IATS) de CSIC. Dada la universalidad de las herramientas aquí generadas hemos tenido también la oportunidad de utilizar otros organismos como caso a estudio en el desarrollo de estas herramientas. En particular, hemos contado con datos de transcriptoma humano gracias a la colaboración del Hospital General Universitario de Valencia, con datos de transcriptoma de diferentes especies de anisakis y sus híbridos, gracias a la colaboración con el Museo Nacional de Ciencias Naturales de CSIC, con datos de secretoma de dos especies de garrapata (*Ornithodoros erraticus* y *Ornithodoros moubata*) gracias a la colaboración con el Instituto de Recursos Naturales y Agrobiología de Salamanca (IRNASA) de CSIC, y, finalmente, con datos del virus SARS-CoV2 gracias a la colaboración con la Fundación Jiménez Díaz. Para cumplir este ambicioso objetivo principal, la investigación realizada en esta tesis se ha estructurado en objetivos específicos de dos tipos: objetivos específicos técnicos y objetivos específicos biológicos. En cuanto a los objetivos específicos técnicos:

1. Diseñar, implementar y validar un flujo de trabajo (*pipeline*) para el ensamblaje de *novo*, predicción génica y anotación de genomas eucariotas usando como caso a estudio el genoma de *Sparus aurata*.

2. Diseñar, implementar y validar un flujo de trabajo (*pipeline*) para el ensamblaje de *novo* y anotación de transcriptomas eucariotas usando como casos a estudios secretomas salivares de *Ornithodoros erraticus* y *Ornithodoros moubata,* y los transcriptomas procedentes de *Anisakis simplex s.s.*, *Anisakis pegreffii* y sus híbridos.

3. Diseñar, implementar y validar un flujo de trabajo (*pipeline*) para el mapeo sobre referencia genómica y/o transcriptómica con análisis de expresión y enriquecimiento diferencial de transcriptomas con y sin fichero de anotación GTF/GFF, y en distintas condiciones (control vs caso a estudio) usando como casos a estudios datos RNA-seq de anisakis*, Ornithodoros erraticus* y *Ornithodoros moubata,* y *Homo sapiens*.

4. Diseñar un *pipeline* para la caracterización de elementos genéticos móviles en genomas eucariotas utilizando *Sparus aurata* como organismo modelo.

5. Definir y diseñar las funciones y requerimientos de interfaces para la implementación software de los protocolos de genómica y transcriptómica aquí diseñados en distintas aplicaciones del GPRO suite.

6. Diseñar, implementar y validar una aplicación de inteligencia artificial basado en una red bayesiana orientada a inferir y predecir cómo los pan-microbiomas de los peces y otras variables, implicadas en la dinámica de un sistema acuícola dado (en este caso, *Sparus aurata*), están relacionados y se influyen mutuamente, a partir de datos de abundancia procedentes de amplicones 16S y metadatos experimentales.

7. Poner a punto un protocolo para la detección de haplotipos en muestras, procedentes de secuenciación de virus aislados a partir de pacientes infectados con SARS-CoV2.

En cuanto a los objetivos específicos biológicos:

1. Caracterizar el genoma de *Sparus aurata* mediante su ensamblado, predicción de genes y anotación de los mismos.

2. Caracterización del moviloma de *Sparus aurata*, incluyendo retroelementos no-LTR, retroelementos LTR, transposones de DNA y retroelementos YR.

3. Caracterización de los perfiles de mutaciones presentes en los haplotipos de SARS-CoV2.

4. Obtención de genes diferencialmente expresados en los estudios realizados para la validación de los protocolos de RNA-seq.

5. Obtención de términos *Gene Ontology* (GO) y de rutas metabólicas diferencialmente enriquecidas en los estudios realizados para la validación de los protocolos RNA-seq

3. ENSAMBLAJE *DE NOVO* DEL GENOMA DE *SPARUS AURATA*, PREDICCIÓN DE GENES Y SU ANOTACIÓN

3.1. Contexto

La metodología de ensamblaje de *novo* de un genoma es el proceso bioinformático por el cual se reconstruye un genoma biológico por primera vez y sin ninguna información o referencia previa a partir de un conjunto de lecturas procedentes de secuenciación (Miller et al., 2010; Nagarajan and Pop, 2013). El objetivo de este proceso es determinar la secuencia del genoma a estudio usando fragmentos de secuencia muestreados aleatoriamente (Chaisson et al., 2015). Una reconstrucción precisa es crucial, ya que tanto la continuidad como la precisión del ensamblaje pueden afectar a los resultados obtenidos mediante análisis posteriores realizados a partir de este ensamblaje (Denton et al., 2014). En condiciones ideales, es decir, una cobertura uniformemente alta y un genoma con pocas secuencias repetitivas, se puede determinar un ensamblaje con el enfoque más simple consistente en fusionar lecturas con superposición máxima. Sin embargo, este método es demasiado simplista para ensamblar genomas con organizaciones complejas de manera precisa. Además, la cobertura de secuencia es casi nunca uniforme, y está constituido por secuencias codificantes para genes pero también por repeticiones de diferente longitud, número de copias y altamente divergentes, lo que complican el proceso de reconstrucción del genoma a estudio (Chaisson et al., 2015). Para abordar este problema en el caso del genoma de la dorada, que es particularmente grande y rico en repeticiones, se propuso el uso de una estrategia híbrida para el ensamblaje *de novo* de genomas complejos consistente en el uso combinado de lecturas cortas, generadas mediante tecnología de secuenciación Illumina *pair end* + *mate pair*, y de lecturas largas generadas por la tecnología de secuenciación de tercera generación (TGS, del inglés *Third Generation Sequencing*) PacBio. Aquí cabe considerar que las lecturas procedentes de TGS tienen una tasa de error más elevada que aquellas procedentes de NGS; sin embargo, debido a su longitud, permiten la resolución de regiones de tamaño pequeño y medio de repeticiones que son problemáticas cuando solamente se utilizan lecturas cortas (Haghshenas et al., 2020). Por otro lado, la secuenciación *mate pair* implica la generación de librerías *paired-end* con insertos más largos, facilitando el mapeo de pequeñas regiones repetitivas. Su uso junto con librerías *pair end*, cuyo tamaño de inserto es menor, proporciona una buena combinación de longitudes de lectura para una cobertura de secuenciación máxima en todo el genoma.

El uso de una estrategia híbrida es muy popular por las siguientes razones (Haghshenas et al., 2020):

1) Las lecturas cortas tienen una elevada precisión y los secuenciadores de Illumina las pueden generar con un alto rendimiento a un coste más bajo.

2) Muchos conjuntos de datos de lecturas cortas están disponibles públicamente para muchos genomas.

3) Para algunas tareas como la llamada de variantes, las lecturas cortas proveen una mejor resolución debido a su mayor precisión.

4) A diferencia de los ensamblajes de PacBio, cuya precisión aumenta con la profundidad de la cobertura gracias a su modelo de error aleatorio imparcial (Myers, 2014), la construcción de genomas de referencia de calidad únicamente a partir de lecturas procedentes de Nanopore sigue siendo un desafío debido a los sesgos en la llamada de bases, incluso con una alta cobertura.

Tras la obtención de la primera versión de un ensamblaje de novo de genoma, normalmente es necesario cubrir los *gaps* o huecos generados durante el ensamblaje, sobre todo si se trata de un genoma complejo. Al proceso de cubrir *gaps* se le denomina *gap filling* que consiste en la reconstrucción de fragmentos de secuencia indefinidos o no resueltos entre dos regiones contiguas del mismo *contig o scaffold* de un ensamblaje separadas por uno o más *gaps* de longitud dada. En este caso, las secuencias flanqueantes del *gap* son conocidas y sirven de anclaje para resolver el *gap* añadiendo las lecturas procedentes de secuenciación Illumina. Estos *gaps* se producen al ser regiones difíciles de ensamblar, ya sea porque se trata de una región con baja cobertura o porque contiene secuencias repetitivas (Walve, 2017). Tras este proceso de *gap filling*, es recomendable la realización de un paso de *scaffolding* (la obtención de secuencias más largas resultante a partir de la unión de dos o más *contigs*) mediante el uso de lecturas largas procedentes de tecnologías TGS y/o lecturas *pair end*.

Posteriormente a la reconstrucción del genoma mediante este proceso de ensamblaje, es necesaria la predicción de genes, tanto codificantes de proteína como no codificantes, y su anotación funcional. La predicción de genes codificantes de proteína permite obtener la estructura exón-intrón de estos posibles genes que puedan estar presentes en genomas eucariotas. Esta predicción se realiza mediante el uso de software específicos que permiten la identificación de estas estructuras por búsquedas de similaridad o por métodos *ab initio* de predicción. El primer tipo es un proceso simple consistente en encontrar similitudes en secuencias de genes entre ESTs (del inglés, *Expressed Sequence Tags*), proteínas u otros genomas para el genoma de interés a partir de alineamientos locales o globales. Esta aproximación está basada en la suposición de que las regiones funcionales del genoma (los exones) están más conservadas que otras regiones no funcionales. Una vez que existe similitud entre una cierta región genómica y una EST, DNA o proteína, se puede utilizar esta información para inferir la estructura o función del gen de esa región. La principal limitación de este tipo de enfoque es que solamente alrededor de la mitad de los genes que se están descubriendo tienen una homología significativa con los genes en las bases de datos (Wang et al., 2004). El segundo tipo de aproximación consiste en utilizar la estructura de los genes como una plantilla para la detección de genes, lo que se conoce como predicción *ab initio*. La predicción de genes *ab initio* se basa en la información procedente de dos tipos de secuencia: sensores de señal y sensores de contenido. Los sensores de señal hacen referencia a motivos de secuencia corta, como sitios de *splicing*, puntos de ramificación, tramos de polipirimidina, codones de inicio y codones de paro. La detección de exones

debe basarse en los sensores de contenido, que se refieren a los patrones de uso de codones que son exclusivos de una especie, y permiten distinguir las secuencias codificantes de las secuencias no codificantes circundantes mediante algoritmos de detección estadística (Wang et al., 2004). En estas líneas, los software más exitosos están basados en el algoritmo llamado *Hidden Markov Model* (HMM) (Eddy, 1996), el cual ha sido utilizado durante décadas en el reconocimiento de patrones (Guigó et al., 2000; Rabiner and Juang, 1986). Los HMMs se han generalizado para permitir que un estado en el modelo genere más de un símbolo, dando lugar a los llamados *Generalized* HMM (GHMM), los cuales proporcionan un marco intuitivo para representar y reconocer genes con sus diversas características funcionales (Kulp et al., 1996). Una vez obtenidas las secuencias de los genes, estos se anotan por similaridad a genes conocidos en las bases de datos de conocimiento, normalmente anotados a partir de genomas de especies cercanas, describiendo su función, lo que va a permitir conocer y comprender mejor el organismo a estudio.

La detección y anotación de RNAs no codificantes puede realizarse mediante métodos de predicción de *novo* o por búsquedas de similaridad. A pesar de que la predicción de *novo* de RNAs no codificantes es un foco principal de investigación, hoy en día no hay un método suficientemente maduro para identificar de forma fiable los genes no codificantes en un genoma (Griffiths-Jones, 2007). Normalmente, la predicción de RNAs no codificantes se ha basado, en gran medida, en el potencial de una secuencia para adoptar una estructura secundaria. Sin embargo, se ha demostrado que las estructuras predichas de genes de RNA no son significativamente más estables que aquellas que forman secuencias aleatorias de RNA, al menos para hacer una distinción fiable a nivel de genoma (Rivas and Eddy, 2000). Por lo tanto, una estructura de RNA estable identificada no indica importancia funcional.

Los enfoques más recientes explotan la idea de que las estructuras de RNA funcionalmente significativas deberían estar conservadas en especies cercanas. Simulaciones computacionales han demostrado que un pequeño número de mutaciones probablemente cambie significativamente la estructura secundaria, dando así una función implícita a una estructura conservada (Huynen et al., 1997). Por lo tanto, predecir la estructura secundaria consenso a través de un alineamiento de secuencias de múltiples especies es mucho más útil que hacerlo sobre una sola secuencia. Existen diversos softwares dedicados a intentar proporcionar una probabilidad de que el alineamiento de secuencias adopte una estructura de RNA conservada (Griffiths-Jones, 2007). Algunos de estos software son RNAz (Washietl et al., 2005) y EvoFold (Pedersen et al., 2006). Si se comparan los resultados de estos dos software utilizando el genoma humano, EvoFold predice aproximadamente 10.000 transcritos de RNA con estructura secundaria (Pedersen et al., 2006), mientras que RNAz predice más de 35.000 (Washietl et al., 2005). Además, en un estudio de 2007 de Washielt *et al*., los autores mostraron que la cantidad de predicciones solapantes realizadas con ambos software era menos del 10% y estimaron tasas de *False Discovery Rate* del 50% al 70%, haciendo que la cantidad de predicciones que se pueden validar sea indeterminada. Es por ello por lo que cuando hablamos de anotaciones fiables de RNAs no codificantes, se tiene que recurrir a la búsqueda de homologías de familias de RNA conocidas utilizando, para ello, software como *Basic Local Aligment Search Tool* (BLAST) (Altschul et al., 1990) o BLAST-

like Alignment Tool (BLAT) (Kent, 2002), que permiten anotar secuencias de RNA no codificante mediante su similaridad con estas familias de RNA conocidas.

La reconstrucción de un genoma también hace posible el estudio de su moviloma. Se entiende por moviloma como el conjunto de elementos genéticos móviles (MGE, del inglés *Mobile Genetic Element*) que están presentes en una célula (Siefert, 2009), incluyendo transposones, intrones, repeticiones de baja complejidad, y, también considerados por algunos autores, RNAs no codificantes y genes quiméricos. En eucariotas, los transposones son los MGEs más abundantes en el genoma. Los transposones son secuencias de DNA que se pueden mover de una localización a otra del genoma. Estos elementos fueron identificados por primera vez hace más de 70 años por la genetista Barbara McClintock (McClintock, 1950). Aunque inicialmente los biólogos fueron escépticos sobre este descubrimiento, durante las siguientes décadas se hizo evidente que los transposones se encuentran en casi todos los organismos (tanto procariotas como eucariotas) y, por lo general, en grandes cantidades (Pray, 2008). Debido a su capacidad de poder moverse en el genoma, pueden actuar como agentes mutagénicos, generar reordenamientos en el DNA, influir en la expresión génica y, en definitiva, desempeñar un papel importante en la evolución del genoma del huésped. Es por ello que la identificación, caracterización y análisis de los transposones y su dinámica es importante para una mejor comprensión de la estructura y evolución tanto de los genomas como del propio DNA móvil (Holmes, 2002; Kidwell and Lisch, 1997). Sin embargo, el interés científico sobre los transposones radica no solo en los estudios evolutivos derivados (ya que se consideran marcadores fósiles), sino también en su papel en los genomas del huésped, su utilidad como herramientas biotecnológicas incluso para aplicaciones médicas genéticas y otras posibles implicaciones futuras (Levin and Moran, 2011; Nesmelova and Hackett, 2010).

Los transposones pueden dividirse en dos clases principales en función de su mecanismo de transposición, y cada clase se puede subdividir según el mecanismo de integración cromosómica (Bourque et al., 2018). Los transposones de clase I, también conocidos como retrotransposones, se movilizan a través de un mecanismo de "copiar y pegar" mediante el cual un intermediario de RNA se transcribe inversamente en una copia de cDNA que se integra en otra parte del genoma (Boeke et al., 1985). Para los retrotransposones de repetición terminal larga (LTR), la integración se produce por medio de una reacción de escisión y transferencia de cadena catalizada por una integrasa muy parecida a la que poseen los retrovirus (Brown et al., 1987). Para los retrotransposones que no son LTR, que incluyen elementos nucleares intercalados largos y cortos (LINE y SINE, respectivamente), la integración cromosómica se acopla a la transcripción inversa a través de un proceso denominado transcripción inversa cebada por el objetivo (Luan et al., 1993). Dentro de esta clase también se encuentran los retroelementos YR (del inglés, *Tyrosine Recombinase*), que se caracterizan por la falta de la proteasa y por la sustitución de la integrasa por una recombinasa de tirosina; y los retrotransposones Penelope (PLEs), los cuales se caracterizan por tener una organización estructural peculiar y por tener la habilidad de retener intrones durante la transposición. Estas características sugieren que los PLEs constituyen una clase antigua de retroelementos (Arkhipova, 2006; Schostak et al., 2008). Los elementos de clase II, también conocidos como transposones de DNA, se movilizan a través de un

intermediario de DNA, ya sea directamente a través de un mecanismo de "cortar y pegar" (Greenblatt and Brink, 1963; Rubin et al., 1982) o, en el caso de Helitrons, un mecanismo "*peel-and-paste*" de replicación que implica un intermedio de DNA circular (Grabundzija et al., 2016). Cada subclase de transposones se divide en subgrupos o superfamilias que normalmente se encuentran en una amplia gama de organismos, pero comparten una organización genética común y un origen monofilético. Por ejemplo y con acuerdo con el Comité Internacional de Taxonomía de Virus (ICTV), los elementos Ty3/gypsy (Llorens et al., 2020) y Ty1/copia (Llorens et al., 2021) son dos géneros de retrotransposones LTR que se encuentran prácticamente en todos los grupos principales de eucariotas (Malik and Eickbush, 2001), mientras que los elementos Bel/Pao (Soriano et al., 2021) solamente se encuentran en genomas de metazoos. Del mismo modo, Tc1/mariner, hAT (hobo-Ac-Tam3) y MULE (elementos similares a mutadores) son tres superfamilias de transposones de DNA que están muy extendidas en el árbol eucariota (Feschotte and Pritham, 2007).

3.2. Material y métodos

3.2.1. Secuenciación de DNA/RNA para la obtención de librerías del genoma

El material de DNA genómico se utilizó para la preparación de dos librerías TrueSeq Illumina estándar (Illumina, Inc.) con un tamaño medio de 360 y 747 pares de bases, respectivamente. Se utilizó el sistema Illumina NextSeq500 como plataforma de secuenciación en un formato *paired-end* de 2x150 para generar aproximadamente 600 millones de lecturas. Además, se implementaron dos estrategias diferentes para ayudar a la realización del ensamblaje *de novo* posteriormente: 1) se usó el Nextera Mate-Pair Preparation Kit (Illumina, Inc.) para crear dos librerías *mate-pair* (MP), con un tamaño de inserto entre 5 y 8 kilobases, mediante la plataforma Illumina NextSeq500 a una profundidad de 11 Gb en formato 2x75 MP, y 2) el DNA genómico se envió a Macrogen (Seúl, Corea del Sur) para la construcción de 12 librerías de células en tiempo real de molécula única (SMRT), con tamaño de inserto de hasta 50 kilobases, utilizando PacBio RS II (Pacific Biosciences) como sistema de secuenciación. Además, se construyeron ocho librerías de RNA-seq mediante el protocolo de preparación Illumina TrueSeq RNA-seq. La secuenciación de librerías indexadas se realizó en Illumina Hiseq v3, lo que resultó en aproximadamente 11 a 17 millones de lecturas por muestra (1x75 lecturas *single-end*) de seis muestras de músculo esquelético y 22 a 27 millones de pares de lectura (2x150 lecturas *paired-end*) de dos muestras de intestino agrupadas.

3.2.2. Ensamblaje de *novo* del genoma de *Sparus aurata*

Las librerías de células SMRT se preprocesaron utilizando la función de recorte del ensamblador CANU (Koren et al., 2017). Las librerías Illumina *paired-end* se procesaron con FastQC 0.11.7 (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/), para analizar la calidad de las lecturas y luego con CUTADAPT v1.16 (Martin, 2011) y Prinseq 0.20.4 (Schmieder y Edwards, 2011) para eliminar adaptadores y lecturas o fragmentos de las mismas que no cumplen los filtros de tamaño y calidad. El análisis de calidad y el

preprocesado de las librerías de Illumina MP se realizó con FastQC y Platanus (Kajitani et al., 2014). Estos protocolos para el preprocesado de las librerías se ejecutaron utilizando la herramienta DeNovoSeq perteneciente al GPRO Suite (Futami et al., 2011).

Para la estimación del posible tamaño del genoma se utilizó la herramienta Jellyfish (Marçais and Kingsford, 2011), la cual permite hacer esta estimación mediante el cálculo de la distribución de conteo de *k-mers* en el conjunto de librerías *paired-end* procedentes de Illumina. Las librerías MP y *paired-end* de Illumina se utilizaron como *input* en el ensamblador SOAPdenovo2 v2.04-r241 (Luo et al., 2012) para realizar el ensamblaje hibrido *paired-end* más *mate pair* del genoma de la dorada. Con el objetivo de testar diferentes valores de *k-mer*, se realizaron diferentes ensamblajes, obteniendo que el *k-mer* 63 fue el mejor a utilizar al tener las mejores métricas. A continuación, para mejorar la secuencia consenso y cerrar *gaps*, se realizaron dos rondas de la siguiente estrategia combinada: 1) Eliminación de duplicados con la herramienta Dedupe del paquete BBTools (sourceforge.net/projects/bbmap/); 2) *Gap filling* usando las lecturas corregidas procedentes de PacBio en PBJelly (English et al., 2012); 3) *Gap filling* utilizando las librerías *paired-end* y MP con la herramienta GapCloser perteneciente a SOAPdenovo; 4) Reensamblaje híbrido utilizando las lecturas corregidas SMRT junto con las lecturas *pair-end* y MP procedentes de Illumina con Opera 2.0.6 (Gao et al., 2011); 5) Reensamblaje guiado por transcriptoma usando como referencia el transcriptoma de dorada (Calduch-Giner et al., 2013) con el software L_RNA_scaffolder (Xue et al., 2013).

3.2.3. Predicción *ab initio* de genes codificantes de proteínas y su anotación

Para la predicción de genes codificantes de proteínas se montó un protocolo de trabajo basado en el software AUGUSTUS 3.3 (Stanke et al., 2008). Primero, se llevó a cabo una ronda inicial de entrenamiento + predicción consistente, en primer lugar, en la obtención de los parámetros más adecuados para la predicción de genes *ab initio* a partir del entrenamiento de los mismos usando trece especies de peces (*Astyanax mexicanus*, *Danio rerio*, *Gadus morhua*, *Gasterosteus aculeatus*, *Latimeria chalumnae*, *Lepisosteus oculatus*, *Oreochromis niloticus*, *Oryzias latipes*, *Petromyzon marinus*, *Poecilia formosa*, *Takifugu rubripes*, *Tetraodon nigroviridis* y *Xiphophorus maculatus*) disponibles en la base de datos de Ensembl *release* 87 (Cunningham et al., 2015). Mediante el uso de los proteomas de las especies anteriormente citadas y el ensamblaje híbrido previamente descrito, se crearon un total de 13 *training sets*, uno por cada especie, utilizando para ello la herramienta Scipio 1.4 de AUGUSTUS (Keller et al., 2008). A continuación, se realizó un entrenamiento para cada especie utilizando el *training set* creado en el *script* autoAugTrain de AUGUSTUS para después realizar trece predicciones de genes distintas con el *script* autoAugPred, una por cada especie utilizada. Posteriormente, se creó un archivo de anotación único obtenido a partir de la combinación de los trece archivos de anotación resultantes de las predicciones utilizando, para ello, la herramienta Cuffmerge (Trapnell et al., 2012), la cual permite la eliminación de secuencias duplicadas. A través de este fichero de anotación único, se obtuvo las secuencias nucleotídicas de los transcritos y las correspondientes secuencias de proteínas traduciendo los transcritos mediante el *script* OrfPredictor (Min et al., 2005).

Las proteínas obtenidas se utilizaron en Scipio 1.4 para generar un nuevo *training set* para llevar a cabo la segunda ronda de entrenamiento y de predicción de genes. Este nuevo entrenamiento, utilizando de nuevo el *script* autoAugTrain, permitió establecer los parámetros más adecuados para la posterior re-predicción de genes codificantes de proteína mediante el *script* autoAugPred. Una vez establecidos los parámetros, se procedió a la re-predicción utilizando el siguiente material publicado con anterioridad en forma de pistas (*hints*):

- Transcriptoma de dorada (Calduch-Giner et al., 2013). Este material ha permitido la creación de pistas a partir de ESTs. Para ello, se usó BLAT (Kent, 2002) para el alineamiento de estos ESTs contra el genoma de la dorada. El archivo PSL generado se filtró mediante el *script* pslCDnaFilter y se obtuvo las pistas en formato GFF mediante el *script* blat2hints.pl. Ambos *scripts* pertenecen al paquete de AUGUSTUS.

- Datos de RNA-seq procedentes de músculo e intestino (Piazzon et al., 2019), junto a aquellos datos de hígado, branquias y bazo, obtenidos del SRA archive del NCBI (Bioproject ID: PRJNA507368). Los ficheros BAM procedentes de estos RNA-seq se filtraron utilizando el *script* filterBam de AUGUSTUS. El BAM filtrado se reordenó utilizando SAMtools (Li et al., 2009). A continuación, se procedió a obtener las pistas referentes a intrones y exones. Para los primeros, se transformó la información de los BAM a pistas mediante el *script* de AUGUSTUS llamado bam2hints. Para los segundos, se transformaron los BAM ordenados mediante SAMtools en formato wig gracias a la ejecución del *script* bam2wig, un formato que permite la visualización de la densidad. Por último, se consiguió las pistas de exones a partir de los ficheros en formato wig utilizando el *script* denominado wig2hints. De esta manera, se consiguió dos ficheros GFF por archivo BAM, el primero con las pistas de los intrones y el segundo conteniendo las pistas de los exones.

Las proteínas traducidas para la obtención del *training set* se utilizaron también para la obtención de pistas para la posterior predicción de genes. Estas proteínas se mapearon contra el genoma de la dorada utilizando Exonerate (Slater and Birney, 2005), ejecutado a través del *script* de AUGUSTUS llamado startAlign. El fichero resultante se utilizó como *input* en el *script* align2hints para obtener un archivo GFF.

Todos los ficheros GFF conteniendo las pistas se fusionaron en un único fichero y se utilizó para la predicción de genes mediante AUGUSTUS. Como en la primera ronda de predicción, los transcritos se obtuvieron mediante el software Gffread (Trapnell et al., 2012), extrayéndolos a partir del fichero de anotación GFF generado por AUGUSTUS y del genoma de la dorada. Una vez extraídos, se tradujeron a proteína utilizando, de nuevo, el *script* OrfPredictor. Los transcritos resultantes se anotaron utilizando la opción BLASTX del paquete BLAST (Altschul et al., 1990), realizando una búsqueda contra las bases de datos Swissprot/Uniprot, NR del NCBI y el transcriptoma de dorada del IATS-CSIC usando un valor de corte del e-valor de 10^{-5}.

3.2.4. Detección y anotación de los genes no codificantes de proteína

Para la detección y anotación de genes no codificantes de proteína en el genoma de *Sparus aurata*, se construyó una base de datos no redundante de *small* y *long non coding* RNAs mediante la extracción de este tipo de secuencias de Ensembl *release* 87 (Cunningham et al., 2015) presentes en los genomas de las trece especies de peces utilizadas para la predicción y anotación de genes codificantes de proteína. Se creó una base de datos adicional que contenía tRNAs de *Danio rerio*, *Gasterosteus aculeatus*, *Oryzias latipes*, *Petromyzon marinus*, *Takifugu rubripes* y *Tetraodon nigroviridis* procedentes de UCSC (http://gtrnadb2009.ucsc.edu). A continuación, se realizó una búsqueda por similaridad mediante BLAT (Kent, 2002) para detectar y anotar los *non coding* RNAs que pudieran estar presentes en el genoma de la dorada. Se obtuvo un fichero resultante por cada especie que fueron convertidos a formato GFF a través del *script* blat2gff.pl (Gupta, 2013), se ordenaron por posición y *scaffold* mediante la función *sort* del paquete de utilidades BEDTools (Quinlan and Hall, 2010) y se eliminaron secuencias duplicadas utilizando la función *merge* del mismo paquete. Por último, se realizó un procedimiento final de curado basado en fusionar entradas del GFF y que pertenecieran al mismo *scaffold*. Este procedimiento consistió en los siguientes pasos:

1. Fusionar entradas procedentes de la misma secuencia de referencia *parent* y que fueran consecutivas en 5-10 nucleótidos. Para ello, se programó un *script* llamado "1_join_by_parent.py".

2. Fusionar entradas procedentes de la misma secuencia de referencia *target* y con la misma posición de inicio. Para ello, se programó el *script* "2_join_by_target_and_pos.py".

3. Fusionar entradas que fueran sobrelapantes y pertenecieran al mismo biotipo. Para ello se programó el *script* "3_join_overlapping.py".

4. Fusionar entradas que estuvieran soportadas por un único transcrito real procedente del transcriptoma de *Sparus aurata* (Calduch-Giner et al., 2013) mediante el *script* "4_join_gff_by_psl.py".

Tras el curado, se eliminaron las secuencias cortas contenidas en secuencias más largas mediante el uso del *script* "5_remove_seqs.py".

Todos los *scripts* citados anteriormente se programaron en Python y se encuentran disponibles dentro de la carpeta llamada ncRNAs en el repositorio denominado "GPRO scripts" en la cuenta de Github perteneciente a Biotechvana (https://github.com/biotechvana/GPRO-scripts).

3.2.5. Identificación y detección del moviloma de *Sparus aurata*

Tal y como se ha indicado anteriormente, el moviloma de un genoma está compuesto por intrones, RNAs no codificantes, transposones, repeticiones de baja complejidad y genes quiméricos.

En el caso del genoma de la dorada, los intrones se identificaron a partir de la predicción *de novo* de genes codificantes de proteína, mientras que la detección y anotación de RNAs no codificantes se ha explicado en el apartado "3.2.4. Detección y anotación de los genes no codificantes de proteína".

Para identificar y anotar los MGEs restantes, se utilizó, en primer lugar, RepeatModeler 1.0.11 (Smit and Hubley, 2015) para la caracterización *de novo* de los elementos móviles y repeticiones de la dorada. El uso de RepeatModeler consistió en los siguientes pasos: 1) creación de una base de datos del genoma de dorada para que adopte el formato necesario para utilizarse en este programa y 2) identificación y caracterización *de novo* de familias de elementos que puedan estar presentes en el genoma. En segundo lugar, se utilizó el software RepeatMasker 4.0.7 (Smit et al., 2015) para identificar repeticiones simples, repeticiones de baja complejidad y repeticiones intercaladas en el genoma de *Sparus aurata*. Para ello, se usó la base de datos Repbase 22.09 (Bao et al., 2015), GyDB 2.0 (Llorens et al., 2011) y las familias de repeticiones identificadas de *novo* mediante RepeatModeler como librerías. A continuación, los software LTR finder (Xu and Wang,2007) y Einverted de EMBOSS (Rice et al., 2000) se utilizaron también para caracterizar las LTRs y repeticiones invertidas, respectivamente.

Por último, para la anotación de MGEs, se utilizó el programa BLASTX de BLAST para anotar las secuencias obtenidas a partir de RepeatMasker contra las bases de datos Repbase y GyDB utilizando un valor umbral de e-valor de 10^{-5}. Todas las anotaciones correspondientes con genes codificantes de proteína asociadas a MGEs (genes quiméricos), obtenidas mediante el uso de la función Intersect del paquete BEDTools, se extrajeron de la anotación de genes codificantes realizada anteriormente y se utilizaron como *queries* en una búsqueda BLAST contra las bases de datos Repbase 22.09 y GyDB 2.0. Todos los resultados fueron curados mediante la fusión de características superpuestas con la misma anotación o separadas por menos de 100 nucleótidos.

Para facilitar el posterior uso de los protocolos de ensamblaje, predicción de genes y anotación descritos en esta sección de métodos, estos protocolos han sido implementados como flujos de trabajo o *pipelines* gestionados mediante interfaces gráficas (GUIs) en la aplicación DeNovoSeq del GPRO Suite que se describe en el apartado "Implementación de *pipelines* en la herramienta DeNovoSeq del GPRO Suite" presentado a continuación en los resultados.

3.3. Resultados

3.3.1. Ensamblaje de *novo* del genoma de la dorada

El genoma de la dorada se ensambló utilizando una estrategia híbrida mediante el uso de Illumina *paired-end* más *mate pair* y PacBio como plataformas de secuenciación. En la Figura 3.1, se muestran las etapas y los pasos seguidos desde la secuenciación hasta la anotación de los genes predichos.

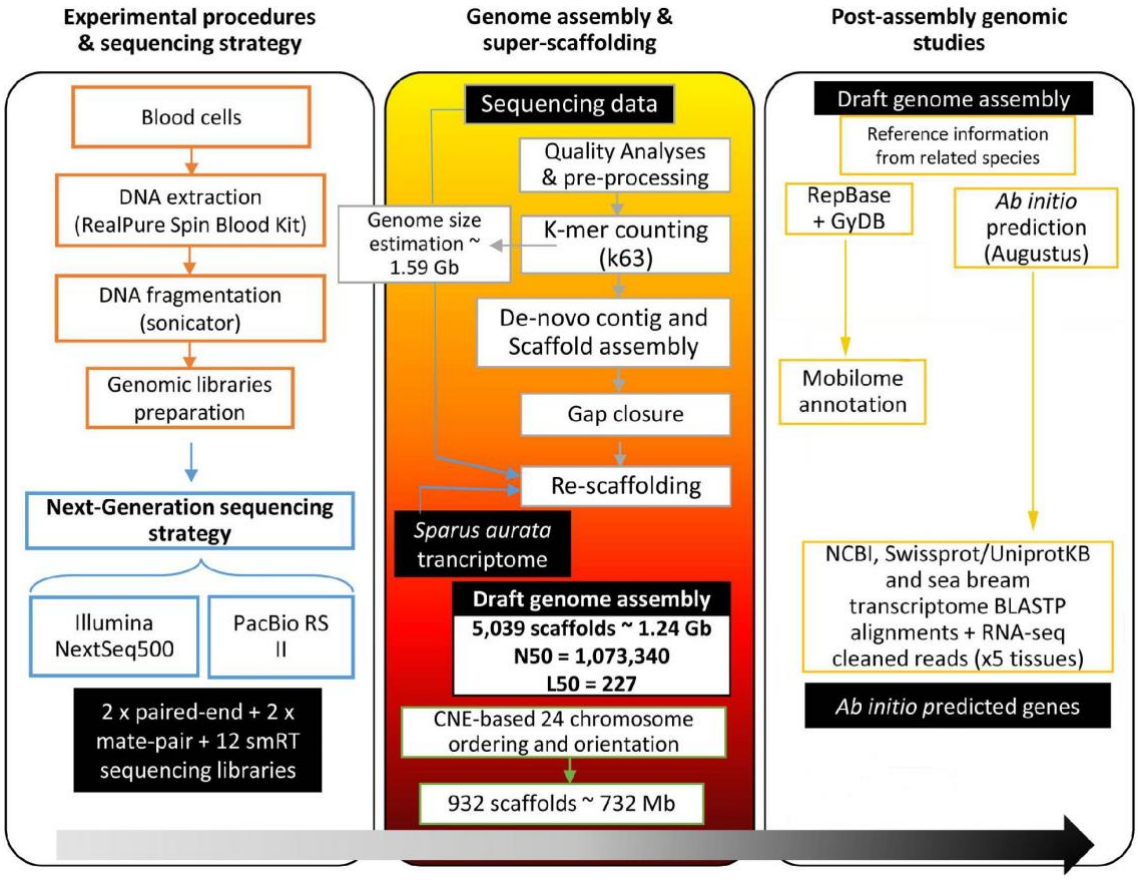

Figura 3.1. Flujo de trabajo del ensamblaje del genoma de dorada. Los recuadros negros con letra blanca indican los recursos genómicos generados de acuerdo con los siguientes pasos: procedimiento experimental y secuenciación, ensamblaje del genoma, predicción y anotación de genes codificantes de proteína.

La anteriormente citada estrategia permitió obtener un borrador del genoma de la dorada de 1,24 GB aproximadamente (Tabla 3.1). Este ensamblaje resultó en un total de 5.039 *scaffolds* con un N50 de 1,07 Mb y un L50 de 227 *scaffolds*. El porcentaje de *contigs* que se ensamblaron en *scaffolds* fue de 99,2%, con un tamaño promedio de *scaffold* de 247,38 kb y con un contenido medio de GC de 39,82%.

Tabla 3.1. Estadísticas resumidas de la anotación de genes en el genoma de la dorada.

	Ensamblaje	***Super-scaffolding***
Tamaño del genoma (en Mb)	1.246.531.774	732.670.891
Número de *scaffolds*	5.039	932
Rango de tamaño (Mín-Máx)	765-16.075163	1.868-12.047.293
Número de regiones codificantes (CDS) predichas	55.423	30.455
Tamaño medio de las CDS (bp)	10.134	11.756
Descripciones únicas	21.275	16.046
Tamaño medio de genes (bp)	10.134	11.756
Numero de exones codificantes	364.433	208.299
Número de intrones	306.674	178.167
Tamaño medio de los exones codificantes	184,18	173,75
Tamaño medio de los intrones	1.751	1.806
Bases totales asociadas a intrones (Mb)	598	358
Densidad de genes	0,048	0,042
Tasa de duplicación basada en anotaciones (CDS/Descripciones únicas)	2,43	1,90
Tamaño medio de proteínas	375	396
Exones/transcrito (excluyendo genes con un solo exón)	5,95	6,70
Intrones/transcrito (excluyendo genes con un solo exón)	5,14	5,84

El *super-scaffolding*, realizado por otro miembro del equipo de investigación, obtenido a partir del uso de 7700 CNEs (del inglés, *Conserved Non-coding Elements*) derivados del mapa de ligamiento genético del primer *paper* del genoma de la dorada (Pauletto et al., 2018), dio como resultado la ordenación y orientación del 57,8% de la longitud total del ensamblaje (~732 Mb) en 24 *super-scaffolds*. Para más detalles sobre las estadísticas del ensamblaje y del *super-scaffolding*, véase la Tabla 3.1.

Dado que el tamaño del genoma (Tabla 3.1) resultó ser largamente superior al obtenido por otro grupo de investigación (Pauletto et al., 2018) en otra investigación paralela e

independiente a nuestro trabajo (760 Mb), procedimos a evaluar la veracidad de nuestros resultados mediante el análisis de *k-mers* utilizando las lecturas *paired-end*. Como se puede observar en la Figura 3.2A, el análisis de *k-mers* mostró una frecuencia de longitud de lectura de 63-mer con un tamaño de genoma estimado de aproximadamente 1,59 Gb (pico principal), incluidos 543 Mb de *k-mers* repetidos (pico de repetición), lo que indica que hay una gran cantidad de secuencias repetidas en el genoma. Como hemos indicado, la longitud total de nuestro genoma borrador fue de aproximadamente 1,24 Gb, lo que representa el 78% del tamaño total estimado del genoma. Según esto, la cobertura de ensamblaje promedio fue de 67,8x, y el 90% del genoma ensamblado total se incluyó en los 1.613 *scaffolds* más grandes, tal y como se muestra en la Figura 3.2B. Esto quiere decir que nuestro borrador del genoma de la dorada resultó ser más completo que el borrador obtenido por el grupo de Pauletto et al., que fue ensamblado sin considerar las repeticiones.

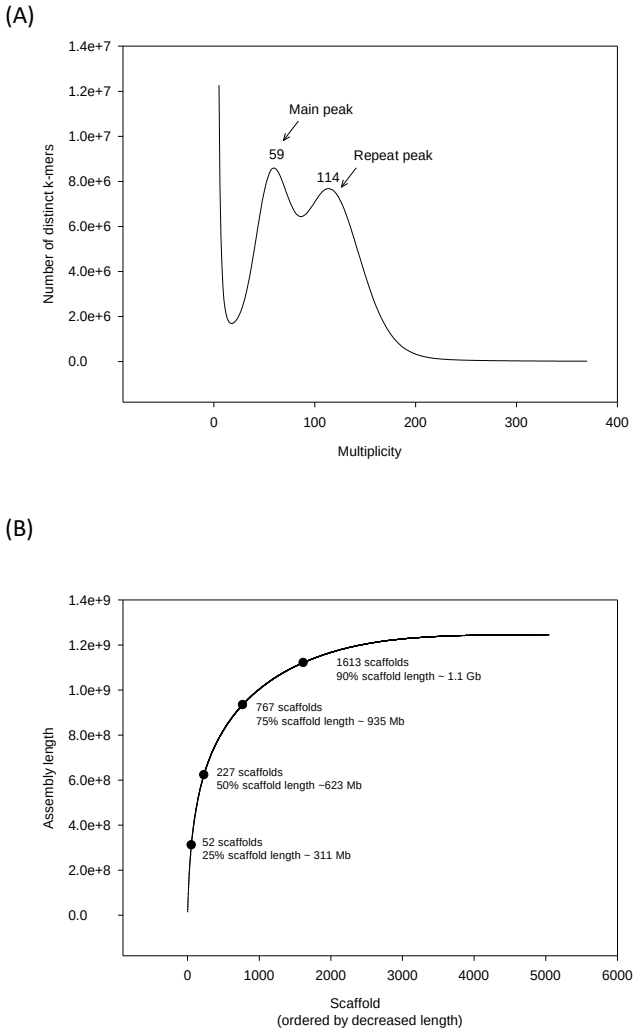

Figura 3.2. Estimación de tamaño del genoma basado en *k-mers* y distribución de *scaffolds*. (A) Histograma de frecuencia de 63-mer para el ensamblaje de dorada para la estimación del tamaño del genoma. (B) Longitud acumulada de los *scaffolds* ensamblados ajustados a la longitud del *scaffold*. Los puntos resaltados destacan el número de *scaffolds* comprimidos por debajo del 25, 50, 75 y 90% de la longitud total del *scaffold*.

3.3.2. Anotación del genoma de la dorada

En el proceso de anotación se realizó una primera predicción *ab initio* de genes codificantes de proteína utilizando AUGUSTUS v3.3 (Stalke et al., 2008). Para respaldar el establecimiento del modelo utilizado, se procesaron para generar un atlas de expresión génica en tejidos un total de ocho librerías de RNA-seq (6 de músculo esquelético y 2 de intestino) en combinación con librerías adicionales de hígado (4), bazo (3) y branquias (3). Las lecturas secuenciadas se mapearon frente a las predicciones *ab initio*, infiriéndose un total de 55.423 genes codificantes de proteína según el análisis transcriptómico de RNA-seq y la homología frente a UniProtKB (The UniProt Consortium, 2021), NR (O'Leary et al., 2016) o la base de datos del transcriptoma de dorada del IATS-CSIC (Calduch-Giner et al., 2013). Este procedimiento generó un total de 21.275 descripciones de genes únicos con 9.250 genes de una sola copia. Nótese que hasta el 90% de los genes únicos corresponden con regiones de los 1.613 *scaffolds* más grandes (Figura 3.3B). La longitud media de este tipo de genes es de 10.134 pb con tamaños medios de exón e intrón de 184 y 1.751 pb, respectivamente. Esto produce una longitud de proteína promedio de 375 aminoácidos.

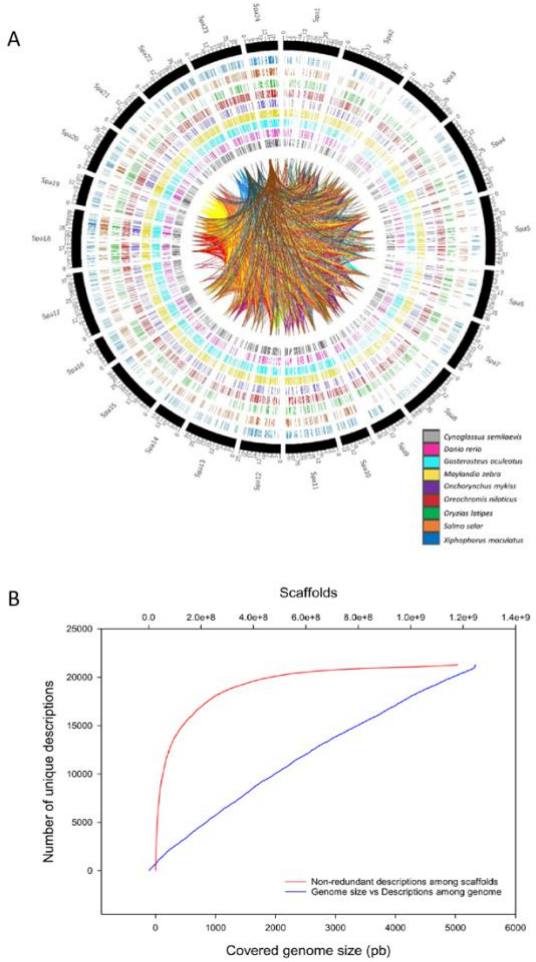

Figura 3.3. Descripciones únicas de genes, distribución y características genéticas. (A) Diagrama circos que representa las relaciones de homología entre la dorada y los genes de otras especies de peces. (B) Distribución acumulativa de anotaciones de genes no redundantes entre *scaffolds* ordenados por longitud.

Se estudiaron las relaciones de homología entre los genes contenidos en los *super-scaffolds* de dorada y los genes secuenciados en otras especies (Figura 3.3A), así como sus relaciones sinténicas obtenidas por otro compañero del equipo. De los 30.455 genes presentes en el *super-scaffolding* de la dorada, 25.806 (84,73%) tenían ortólogos en al menos una de las especies analizadas, siendo tilapia del Nilo, (*O. niloticus*, 20.562), zebra mbuna (*M. zebra*, 19.717), platy *(X. maculatus*, 15.093) y pez espinoso (*G. aculeatus*, 14.612), las especies que comparten más genes ortólogos con la dorada, mientras que los números más bajos de ortólogos se obtuvieron en la trucha arcoiris (*O. mykiss*, 8.866) y el pez cebra (*D. rerio*, 4.288) (Figura 3.3A). El nivel de conservación de la ortología refleja la proximidad filogenética entre las especies comparadas.

La anotación funcional de genes de dorada utilizando ontologías génicas (GO) dio como resultado un perfil funcional de la dorada (Figura 3.4) consistente en un conjunto diverso de categorías funcionales asignadas a 43.221 genes (Componente Celular, 41.423; Función Molecular, 38.505; Proceso Biológico, 38.588). Las 12 categorías principales de cada ontología para las descripciones de proteínas no redundantes se muestran en la Figura 3.4A. Los términos GO del Componente Celular tuvieron el recuento de genes más alto con términos GO de *cytoplasm* (GO:0005737; 20.689), *plasma membrane* (GO:0005886; 16.138) e *integral to membrane* (GO:0016021; 12.436). Los términos GO de Función Molecular más abundantes comprendían *metal ion binding* (GO:0043167; 9.210), *DNA binding* (GO:0003677; 7.041) y *ATP binding* (GO:0005524; 6.518). Los términos GO de procesos biológicos más representados fueron *transcription DNA-dependent* (GO:0006351; 6.222), *signal transduction* (GO:0007165; 3.851) y *multicellular organismal development* (GO:0007275; 2.908). Cuando se testó el enriquecimiento de términos GO entre genes quiméricos/compuestos, los 3.648 genes duplicados con 108 anotaciones de proteínas no redundantes generaron 184 procesos biológicos enriquecidos (p-valor ajustado < 0,05). Estos genes cubren diferentes términos GO relacionados con el sistema inmunitario (26%), el ciclo celular (16%), la iniciación de la traducción (11%), la respuesta a la actividad (11%), la transducción de señales (6%), el proceso de desarrollo (5%) y crecimiento (2%), entre otros (Figura 3.4B). La relación entre las categorías funcionales se ilustra mediante un diagrama de Venn, que muestra 87 descripciones de genes no redundantes de las cinco categorías funcionales principales (Figura 3.4C). Este procedimiento destacó que la alta representación del sistema inmunitario en genes quiméricos/compuestos se debió principalmente a una amplia superposición de términos GO inmunitarios con otras categorías funcionales enriquecidas. Curiosamente, se encontró intersecciones principales entre el proceso del sistema inmunitario, el ciclo celular y la transducción de señales, que comprenden 15 términos GO enriquecidos y 15 descripciones de genes únicos, correspondientes a diferentes isoformas de la proteína NLRC3 y los dominios NATCH, LRR y PYD que contienen la proteína 12.

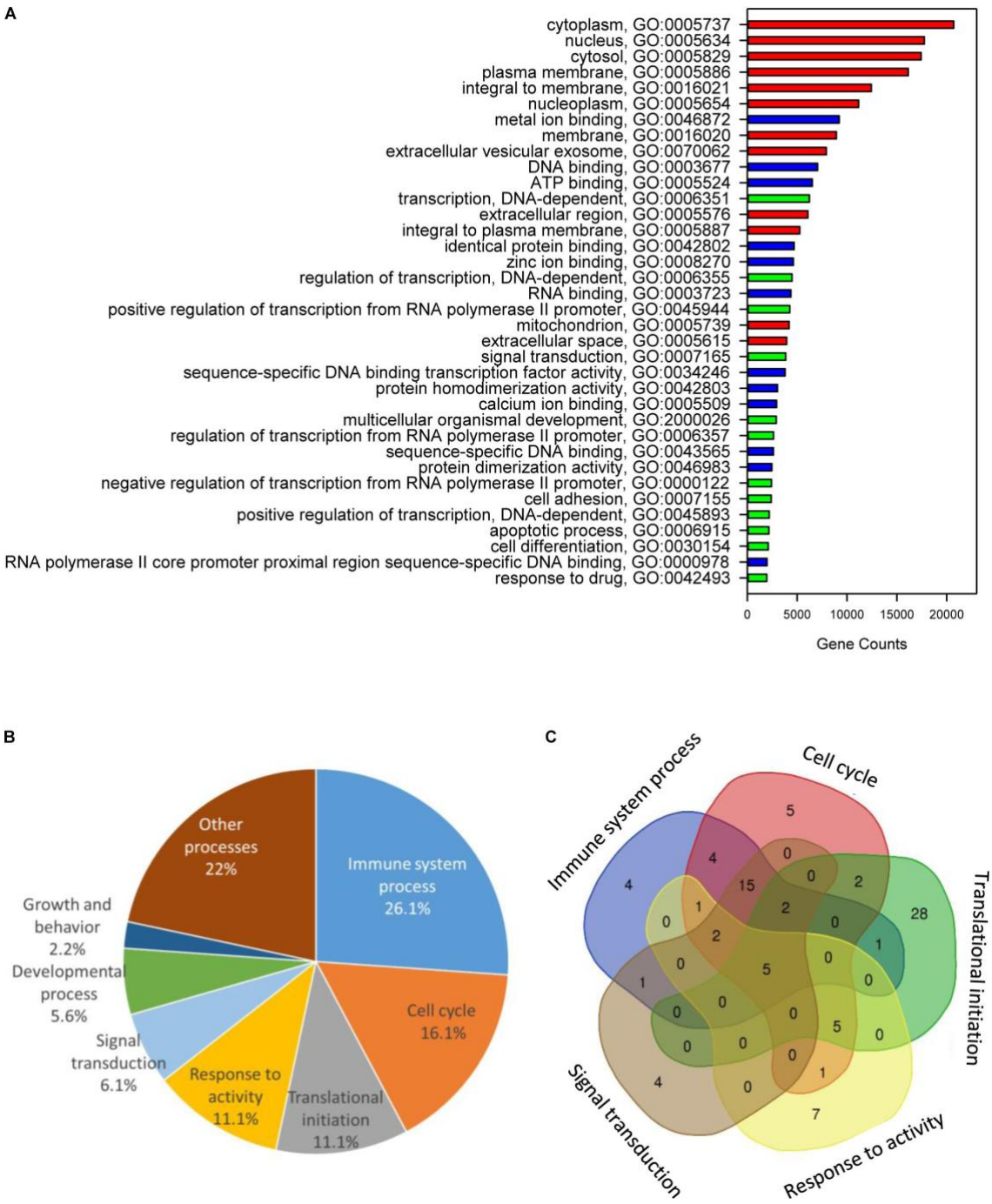

Figura 3.4. Anotación funcional de genes quiméricos y enriquecimiento de ontologías génicas. (A) Análisis de anotación funcional de ontologías de genes (GO) sobre todo el modelo de genes que muestra los principales procesos biológicos (rojo), funciones moleculares (azul) y componentes celulares (verde) para los genes que se encuentran en el genoma de la dorada. (B) Diagrama circular que representa el porcentaje de categorías funcionales de términos GO enriquecidas con procesos biológicos. (C) Diagrama de Venn que representa la superposición de las descripciones de genes únicos entre las principales categorías funcionales.

3.3.3. El moviloma del genoma de la dorada

El moviloma de la dorada tiene un tamaño de 944 Mb representando con ello el 75% del tamaño completo del genoma de este organismo (1,24 Gb). El 60% de este moviloma (599 Mb) está constituido por intrones, mientras que el resto de MGEs están ampliamente distribuidos a lo largo de los distintos *scaffolds*. Véase la Tabla 3.2.

Tabla 3.2. Moviloma de la dorada

Clase	Tipo	Grupos MGE	Tamaño (bp) en el genoma	Porcentaje de moviloma	Porcentaje de genoma
Clase I	Retroelementos con LTRs	4	27.226.719	2,88	2,18
	Retroelementos sin LTRs	14	27.743.197	2,94	2,23
	Retroelementos YR	1	222.290	0,02	0,02
Clase II	Transposones de DNA	27	99.635.002	10,55	7,99
Intrones	Intrones	1	598.945.346	63,44	48,05
Genes quiméricos	Genes relacionados con retroelementos sin LTRs	3	7.390.894	0,78	0,59
Genes multicopia	Genes relacionados con retroelementos con LTRs	4	729.596	0,08	0,06
	Genes relacionados con transposones de DNA	10	5.867.940	0,62	0,47
	Desconocido	2	4.107.023	0,44	0,33
	Genes de ncRNA	1	53.849	0,01	0,004
	Genes relacionados con retroelementos YR	1	23.583	0,002	0,002
	Genes con repeticiones	1	1.630	0,0002	0,0001
	Genes de tipo viral	1	214.839	0,023	0,017
	Genes peptidasa Clan AA	11	47.283	0,005	0,004
	Genes Scan/Krab	1	2.801	0,0003	0,0002
Genes ncRNA	Long ncRNAs	10	10.750.971	1,14	0,86
	Small ncRNAs	11	1.036.782	0,11	0,083
DNA repetitivo	Repeticiones de *novo*	2500	159.623.699	16,91	12,81
	Repeticiones conocidas	5	475.972	0,05	0,04
Tamaño total del moviloma			944.099.416	100	75,7
Fracción excluyendo intrones			345.154.070	36,5	27,6

Las repeticiones de baja complejidad abarcaron 160,5 Mb (16,91%), con aproximadamente 160 Mb correspondientes a 2500 familias de repeticiones clasificadas como específicas *de novo* en la dorada. Los 0,5 Mb restantes correspondieron a repeticiones conocidas (repeticiones invertidas y/o en tándem, así como satélites y microsatélites) también presentes en otros genomas de peces. Los MGEs de clase I (5,84%) comprendían 27,2 Mb de retroelementos LTR (Ty3/Gypsy, BEL/Pao, Ty1/Copia y Retroviridae-like), 27,8 Mb de retroelementos no LTR (distribuidos en 14 familias, principalmente LINE, del inglés *Long Interspersed Nuclear Element*, y SINE del inglés *Short Interspersed Nuclear Element*), y 0,2 Mb de retrotransposones DIRS similares a YR. Los MGEs de clase II (10,55%) incluían 99,6 Mb divididos en 27 grupos de transposones de DNA (principalmente elementos hAT, Tc1/Marines, PIF/Harbinger y PiggyBac). La última fracción del moviloma correspondió con RNA no codificante (1,25%) y genes quiméricos/compuestos (1,96%). En la Tabla 3.3, se muestra una lista completa de genes de RNA no codificante (ncRNA), que incluye tanto *long* (11 Mb constituidos por 10 grupos, principalmente lincRNA, pseudogenes y transcritos procesados) como *short* (1 Mb dividido en 11 grupos, principalmente microRNA, tRNA y snoRNA) ncRNA.

Tabla 3.3. RNAs no codificantes predichos y anotados en el genoma de la dorada.

	Tipo de ncRNA	**Subtipos de ncRNA**	**ncRNAs con nombre de gen**	**ncRNAs anotados**
lncRNA	Processed_pseudogene	16	13	100
	Processed_transcript	45	151	151
	Pseudogene	178	3	942
	Antisense	37	136	136
	IG_V_pseudogene	1	1	1
	TR_V_gene	5	15	15
	lincRNA	1.089	134	5.528
	Sense_intronic	5	35	35
	Misc_RNA	17	21	22
	Unprocessed_pseudogene	25	45	45
	Subtotal lncRNAs	1.418	554	6.975
sncRNA	miRNA	806	1.166	5.668
	rRNA	69	92	104
	snoRNA	221	288	496
	snRNA	144	183	237
	sRNA	4	6	6
	tRNA	329	0	2.272
	Subtotal sncRNAs	1.573	1.735	8.783
	Total ncRNAs (lncRNAs and sncRNAs)	**2.991**	**2.289**	**15.758**

Respecto a los genes quiméricos/compuestos, es decir, aquellos genes con exones de origen móvil, estos se dividieron en 10 grupos: aquellos con similaridad significativa a retroelementos sin LTRs (7 Mb), aquellos con similaridad significativa a retroelementos con LTRs (0.7 Mb), aquellos con similaridad significativa a transposón de DNA (5,8 Mb), aquellos con similaridad significativa a genes no codificantes (0,053 Mb), repeticiones (0,001 Mb), aquellos con similaridad significativa a virus (0,2 Mb), aquellos con similaridad significativa a retroelementos YR (0,02 Mb), así como peptidasas del clan AA (0,047 Mb), genes Scan/Krab (0,008) y genes desconocidos (4 Mb). La representación Krona de los subniveles en los que se divide el moviloma se puede ver en la Figura 3.5.

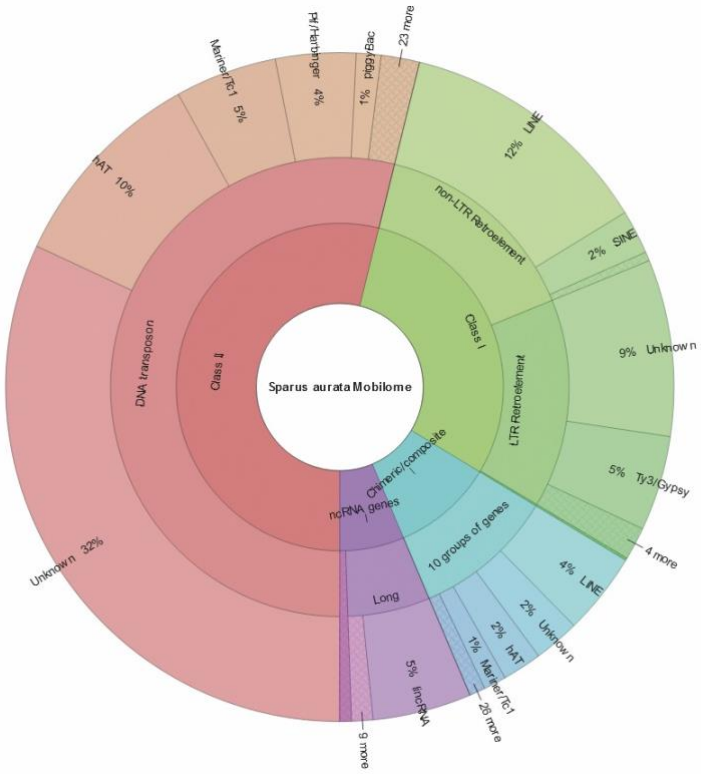

Figura 3.5. Representación Krona de los subniveles del moviloma de la dorada.

3.3.4. Implementación de *pipelines* en la herramienta DeNovoSeq del GPRO Suite

La herramienta DeNovoSeq del GPRO Suite permite nuevos enfoques o aproximaciones de *novo*, proporcionando al usuario un acceso a los flujos de trabajo para el ensamblaje y la anotación de nuevos genomas y/o transcriptomas sin secuencia de referencia previa mediante una interfaz gráfica del usuario (GUI). Centrándonos en el flujo de trabajo seguido para la obtención del ensamblaje y anotación del genoma, actualmente DeNovoSeq presenta el siguiente *pipeline*:

Preprocesado -> Ensamblaje -> Predicción de genes -> Anotación

Respecto al preprocesado, se ha implementado herramientas para el análisis de calidad de las muestras como para el filtrado de las mismas. En el primer caso, FastQC (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/) (Figura 3.6) es la herramienta

utilizada para obtener un reporte de calidad que indique qué filtros aplicar para mejorar la calidad de las muestras. Los diferentes software implementados para aplicar estos filtros son CUTADAPT (Martin, 2011), el cual elimina los adaptadores que puedan estar presentes en las muestras, y PRINSEQ (Schmieder y Edwards, 2011), que recorta, filtra o elimina lecturas que no cumplen con los criterios establecidos, ya sea de tamaño, calidad o cantidad de indeterminaciones, por ejemplo, utilizando diferentes parámetros implementados en la interfaz de esta herramienta.

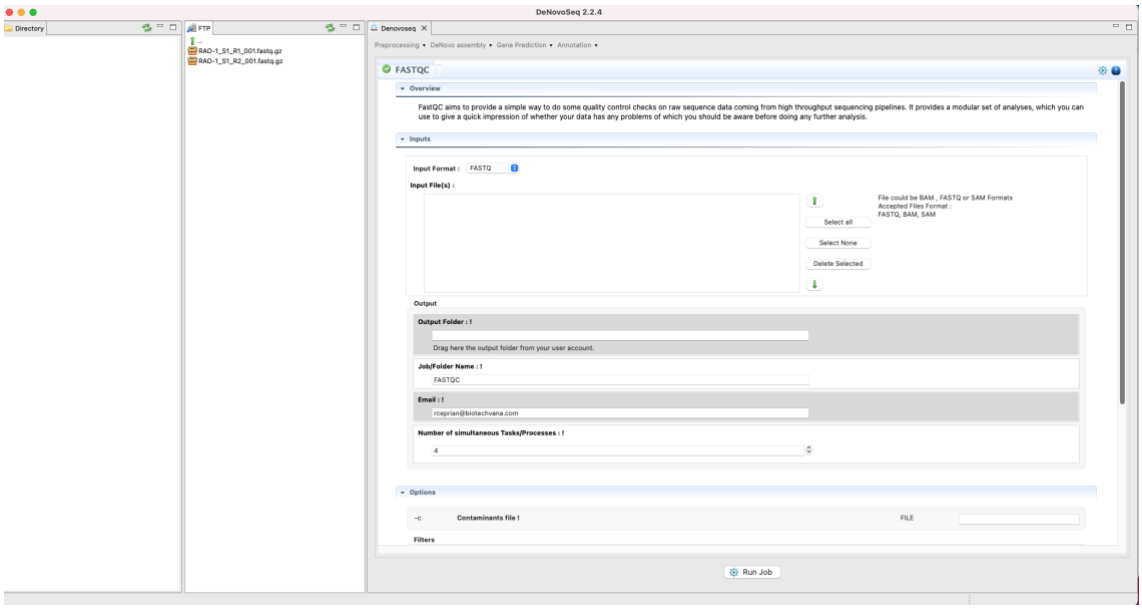

Figura 3.6. Interfaz implementada en DeNovoSeq para ejecutar FastQC.

Tras el preprocesado, le sigue el ensamblaje del genoma. Dentro de la aplicación DeNovoSeq, este paso del flujo de trabajo está dividido en el ensamblaje propiamente dicho, donde se implementa la herramienta SOAPdenovo2 (Luo et al., 2012), el *gap filling*, donde se puede utilizar GapCloser, perteneciente a SOAPdenovo, para completar *gaps* que hayan podido surgir durante el proceso de ensamblado, y el *scaffolding*, que usa Opera (Gao et al., 2011) para unir *contigs* y formar secuencias más largas mediante el uso de librerías *paired-end* y *mate pair* de lecturas cortas, y lecturas largas procedentes de TGS. Todos estos software están implementados de forma que es posible modificar los parámetros de cada uno de ellos y adecuarlos a las necesidades de cada genoma a estudio.

Una vez obtenido el ensamblaje, el siguiente paso es la predicción de genes con AUGUSTUS (Stanke et al., 2008). Al realizarse esta predicción de genes sobre una especie para la que no se tiene una referencia previa, es necesario seguir el siguiente flujo de trabajo implementado en DeNovoSeq:

1. Entrenamiento (Figura 3.7). Establece los parámetros más adecuados para la posterior predicción de genes en una especie concreta. Para ello, se debe partir del propio ensamblaje y de un *training set*, el cual puede ser un conjunto de proteínas en formato fasta, un fichero GenBank con las estructuras de los genes o un archivo GFF que incluya los genes. Por último, es necesario especificar un

nombre para la especie con la que se esté trabajando, en este caso, fue *Sparus aurata*.

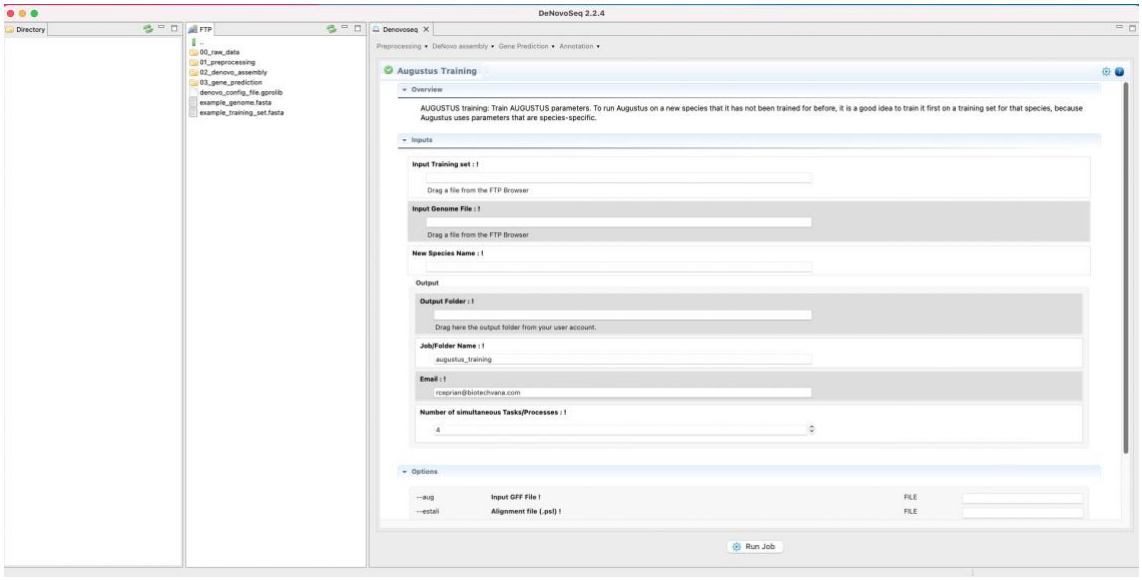

Figura 3.7. Interfaz de AUGUSTUS para el entrenamiento de una especie.

2. Predicción (Figura 3.8). Realiza la predicción de genes a partir de los parámetros establecidos durante el entrenamiento. Como parámetros necesarios se encuentran el ensamblaje y el nombre de la especie con la que se esté trabajando para tener en cuenta los parámetros establecidos durante el entrenamiento.

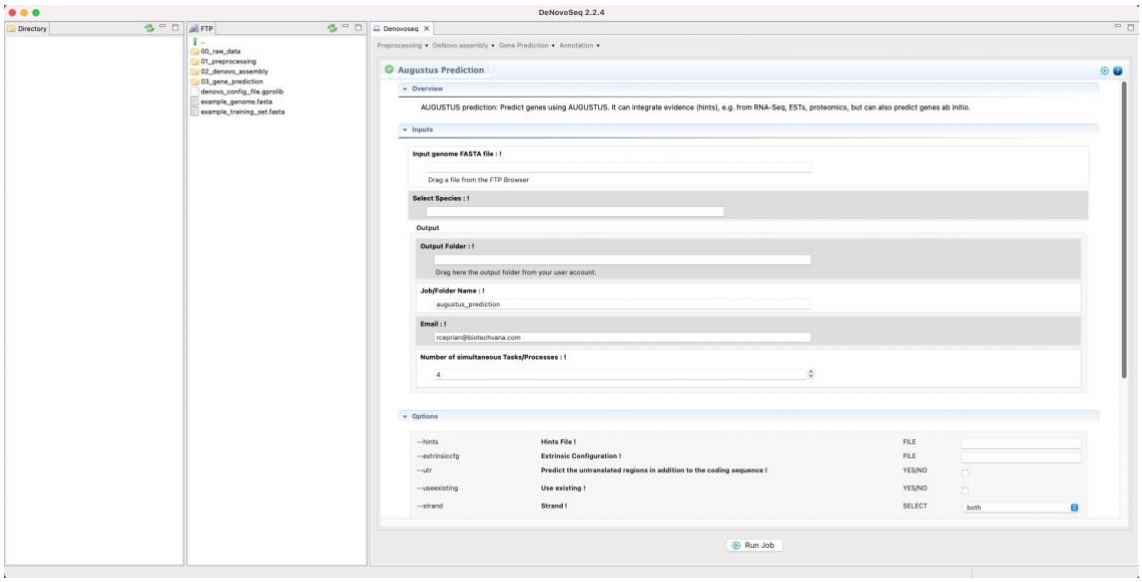

Figura 3.8. Interfaz para la predicción de genes con AUGUSTUS.

En la predicción de genes es posible incorporar, de forma opcional, pistas o *hints* sobre la estructura génica procedentes de otras fuentes como datos de RNA-seq, proteínas o ESTs (del inglés, *Expressed Sequence Tag*). En el caso de datos procedentes de RNA-seq, el protocolo comienza a partir de las librerías ya filtradas, de manera que se mapean contra el genoma utilizando BLAT (Kent, 2002), GSNAP (Wu et al., 2016) o TopHat

(Trapnell et al., 2012) como mapeadores. El resultado es la obtención de dos ficheros GFF, el primero conteniendo las coordenadas de los exones y el segundo conteniendo las coordenadas de los intrones. Si se obtienen pistas a partir de proteínas, éstas se alinean contra el genoma mediante el software Exonerate (Slater and Birney, 2005) obteniendo, finalmente, un archivo GFF con las posiciones del genoma donde se encuentran estas proteínas. Por último, para la obtención de pistas a partir de datos de ESTs, se utiliza BLAT para alinear las ESTs contra el genoma, obteniendo un fichero PSL que se procesa para conseguir un fichero GFF con las regiones del genoma contra las que se han alineado estos ESTs. Como se ha indicado anteriormente, estas pistas son opcionales, ya que AUGUSTUS permite realizar la predicción de genes *ab initio*, pero éstas pretenden mejorar las predicciones utilizando esta información previa.

El último paso del flujo de trabajo implementado en DeNovoSeq es la anotación de los genes predichos mediante BLAST (Altschul et al., 1990), comparando las secuencias predichas contra una base de datos de genes o proteínas cuya función se conoce. Para ello, se ha creado un subflujo de trabajo consistente en:

1. Formatear un fichero fasta de nucleótidos o proteínas específicos pertenecientes, por ejemplo, a especies cercanas de la especie de interés, o importar una base de datos precompilada de referencia como puede ser la NT o NR del NCBI. Ambas opciones permiten obtener una base de datos en formato adecuado para poder realizar la anotación con BLAST.

2. Ejecutar la comparación con BLAST entre los genes predichos y la base de datos que se va a utilizar como referencia, utilizando la base de datos como *subject* y los genes predichos como *query*. Dentro de esta interfaz se debe especificar el programa a utilizar de BLAST en función de si queremos realizar la comparación utilizando una base de datos de nucleótidos (programa BLASTN) o de proteínas (programa TBLASTN).

3. El resultado de BLAST es una serie de ficheros XML que se procesan en este subflujo de trabajo para obtener finalmente un fichero CSV con la anotación final de los genes predichos del genoma.

3.4. Discusión

Los constantes avances en las tecnologías de secuenciación y la reducción de los costes han mejorado substancialmente la capacidad de generar secuencias genómicas de alta calidad (Metzker, 2010). La lista de genomas en la base de datos del NCBI (www.ncbi.nlm.nih.gov/genome/browse) incluye 340 genomas de peces de 248 especies diferentes, con más de 30 correspondientes a especies de peces de especial relevancia dada su importancia económica o su papel importante como especies modelo de investigación. Durante el desarrollo de esta tesis, se ha generado y puesto a disposición del público un borrador de alta calidad del genoma de la dorada como un esfuerzo para generar nuevas herramientas genómicas para un pez modelo especialmente en la acuicultura en todo el área del Mediterráneo. La estrategia de secuenciación, que combina lecturas *paired-end* + *mate pair* + *Smart cells* ha dado como

resultado uno de los borradores de genomas de peces más completos en términos de cantidad de *scaffolds* por tamaño ensamblado (5.039 *scaffolds* en un ensamblaje de 1,24 Gb). Los intentos anteriores en peces estrechamente relacionados dieron como resultado genomas de referencia muy fragmentados debido a que los genomas de peces suelen ser ricos en repeticiones y al uso de protocolos de ensamblaje basados únicamente en estrategias de secuenciación de lectura corta. Por ejemplo, los genomas públicos de la lubina europea (*Dicentrarchus labrax*, 680 Mb), el pez globo verde manchado (*Tetraodon nigroviridis*, 342 Mb) o el molly amazónico (*Poecilia Formosa*, 830 Mb) se dividen en 46.509, 27.918 y 25.474 *scaffolds*, respectivamente (Jaillon et al., 2004; Tine et al., 2014; Warren et al., 2018). Asimismo, y como hemos indicado, el primer borrador del genoma de la dorada incluía 55.202 *scaffolds* en un ensamblaje de 760 Mb (Pauletto et al., 2018), mientras que el segundo borrador del genoma de la dorada presente en el NCBI (*Bioproject accession* PRJEB31901) comprende 833 Mb aproximadamente, ambos por debajo del ensamblaje realizado durante el desarrollo de este trabajo, el cual es más próximo al verdadero tamaño del genoma de la dorada que el generado por el grupo de Pauletto et al., tal y como revelan nuestros análisis. Este hecho produjo una mayor cantidad de descripciones anotadas de genes únicos al comparar este genoma ensamblado con los dos anteriores (21.275 frente a 13.835 y 19.631).

Ciertamente, los peces comprenden el grupo más grande y diverso de vertebrados, con un tamaño de genomas secuenciados que oscila entre 342 Mb en *Tetraodon nigroviridis* y 2,9 Gb en *Salmo salar* (Yuan et al., 2018). El genoma ensamblado desenmascarado de la dorada realizado en esta tesis es, por lo tanto, de tamaño intermedio (1,24 Gb), aunque se espera que el genoma completo sea alrededor de 350 Mb más largo. De hecho, el ensamblaje actual contiene más de 5.000 descripciones de genes únicas que no están presentes en el súper *scaffolding* basado en el primer borrador del genoma (Pauletto et al., 2018). Las estimaciones del tamaño del genoma de la dorada en función de la citometría de flujo de los glóbulos rojos arrojaron un tamaño del genoma más pequeño (aproximadamente 930 Mb) (Peruzzi et al., 2005). Sin embargo, la precisión de la técnica es limitada debido a las altas fuentes de variación intraensayo (hasta 10%) e interensayo (20-26%) (Pedersen, 1971; Gregory, 2005). Ciertamente, las diferencias en los estándares de tamaño del genoma interno/externo, la preparación de muestras, las estrategias de tinción o la deriva estocástica de los instrumentos pueden dar lugar a diferencias significativas en dichas estimaciones del tamaño del genoma (Doležel et al., 1998). En consecuencia, los métodos computacionales (por ejemplo, k-recuentos de frecuencia de mer usado en esta tesis) están surgiendo como enfoques más fiables para las estimaciones del tamaño del genoma (Sun et al., 2018). Además, otro resultado importante obtenido a partir del análisis de *k-mers* fue un segundo pico pronunciado que es indicativo de una gran cantidad de secuencias repetidas. La mayoría de las predicciones de genes reportadas mostraron un grado suficiente de divergencia, lo que apoya la idea de que se tratan de verdaderas expansiones de genes y no duplicaciones segmentales, ya que los resultados del análisis de redundancia basado en genes transcritos activamente mostraron solamente 1,01% de estas duplicaciones erróneas, indicando que el genoma tiene una muy baja proporción de errores en el ensamblaje. El análisis de sintenia también apoyó la duplicación de genes, que dificulta el establecimiento de bloques de sintenia entre especies, probablemente como resultado

de la sobrerrepresentación de expansiones génicas durante la evolución del linaje de la dorada.

Para obtener información sobre la evolución del genoma de la dorada y estudiar en más detalle el origen de estos altos niveles de duplicación genómica, se realizó un análisis de filoma por parte de un compañero del equipo de investigación y disponible en www.phylomedb.org (phylome ID 714). Su estudio confirmó lo que mostraba el análisis de sintenia y mostró una media de 2.024 copias para los 55.423 genes transcritos activamente, en al menos uno de los tejidos analizados como representación de los tejidos metabólica e inmunológicamente relevantes. Este número de transcritos regulados por tejidos con un alto porcentaje de duplicaciones ofrece la posibilidad de una mayor plasticidad adaptativa en un entorno evolutivo desafiante. Es importante destacar que el enriquecimiento funcional de los genes duplicados específicos del linaje de la dorada, obtenidos a partir del filoma, evidenció una mayor presencia de integración de DNA, transposición y producción de inmunoglobulinas. Este hallazgo sugiere que la mayor parte de las expansiones sufridas por el genoma de la dorada se derivan de las actividades de los MGEs y de la respuesta inmunitaria como procesos clave en la adaptabilidad de la especie.

El moviloma caracterizado destacó una representación abundante de MGEs, así como de una serie de genes quiméricos que aparentemente evolucionaron a partir de la co-domesticación y/o cooptación de MGEs. De hecho, la cooptación es un mecanismo recurrente que ha contribuido a las innovaciones en varios niveles de señalización celular y expresión génica varias veces durante la evolución de los vertebrados (Arkhipova et al., 2012). La fuente más importante de cooptación de genes en este genoma de dorada fue los retrotransposones LINE y los transposones de DNA Tc1/Mariner, que han sido descritos ampliamente en modelos de mamíferos como ejemplos de domesticación de elementos transponibles (Jangam et al., 2017). Entre estos genes quiméricos (Tabla suplementaria S3.1), surgió un número relevante de receptores similares a NOD (NLR), que incluyen proteínas que contienen NACHT, LRR y PYD (NLRP), y dominios CARD de receptores similares a NOD (NLRC). Estos receptores son sensores innatos involucrados en el monitoreo intracelular para detectar patógenos que han escapado a la vigilancia extracelular y endosomal, lo que indica la necesidad de los peces de detectar amenazas en un ambiente rico en patógenos (Pérez-Sánchez et al., 2019).

En resumen, una estrategia de secuenciación combinada de lecturas cortas y largas produjo un borrador de alta calidad del genoma de la dorada. La alta cobertura y profundidad de este ensamblaje dan como resultado un recurso valioso para las próximas aplicaciones basadas en NGS (como RNA-seq o Methyl-seq), análisis de metatranscriptoma, *loci* de rasgos cuantitativos (QTL) y estudios de organización espacial de genes realizados para mejorar los rasgos de este pez altamente cultivado. Por último, el análisis de ensamblaje del genoma de la dorada sugiere que los elementos transponibles son probablemente la causa principal del tamaño agrandado del genoma, ya que son los responsables de desencadenar reordenamientos genómicos, sustituciones, deleciones e inserciones (Kidwell, 2002), lo que conlleva el aumento del tamaño y la complejidad del genoma, además de nuevas combinaciones de genes que

dan lugar a funciones biológicas modificadas o nuevas (Lynch and Conery, 2000), siendo el genoma de la dorada un excelente ejemplo de ello.

3.5. Publicaciones *peer-review* relacionadas con este capítulo en esta tesis

A) Pérez-Sánchez J, Naya-Català F, Soriano B, Piazzon MC, Hafez A, Gabaldón T, Llorens C, Sitjà-Bobadilla A, Calduch-Giner JA. 2019. Genome Sequencing and Transcriptome Analysis Reveal Recent Species-Specific Gene Duplications in the Plastic Gilthead Sea Bream (Sparus aurata). Frontiers in Marine Science, 6:760. https://doi.org/10.3389/fmars.2019.00760.

Contribución de la autora de esta tesis a este trabajo: Ejecución del protocolo para el ensamblaje de *novo* del genoma de la dorada, su posterior anotación y co-redacción del artículo.

B) Llorens C, Soriano B, Krupovic M, ICTV Report Consortium. 2020. ICTV Virus Taxonomy Profile: Metaviridae. The Journal of General Virology, 101(11):1131-1132. https://doi.org/ 10.1099/jgv.0.001509.

Contribución de la autora de esta tesis a este trabajo: Ejecución de los alineamientos y construcción de los árboles filogenéticos para ver las relaciones entre las diferentes familias dentro del género Metaviridae y co-redacción del artículo.

C) Llorens C, Soriano B, Krupovic M, ICTV Report Consortium. 2021. ICTV Virus Taxonomy Profile: Pseudoviridae. The Journal of General Virology, 102(3):001563. https://doi.org/ 10.1099/jgv.0.001563.

Contribución de la autora de esta tesis a este trabajo: Ejecución de los alineamientos y construcción de los árboles filogenéticos para ver las relaciones entre las diferentes familias dentro del género Pseudoviridae y co-redacción del artículo.

D) Soriano B, Kuprovic M, Llorens C, ICTV Report Consortium. 2021. ICTV Virus Taxonomy Profile: Belpaoviridae 2021. Journal of General Virology, 102(11):001688. https://doi.org/ 10.1099/jgv.0.001688.

Contribución de la autora de esta tesis a este trabajo: Ejecución de los alineamientos y construcción de los árboles filogenéticos para ver las relaciones entre las diferentes familias dentro del género Belpaoviridae y co-redacción del artículo.

4. PIPELINE PARA LA DETECCIÓN Y CUANTIFICACIÓN DE CUASIESPECIES DEL VIRUS SARS-CoV-2

4.1. Contexto

La población de virus de RNA en un huésped no consiste en un único haplotipo consenso, sino en un conjunto de secuencias relacionadas denominadas cuasiespecies. Una cuasiespecie describe un tipo de estructura poblacional en la que colecciones de genomas estrechamente relacionados están sometidas a un proceso continuo de variación genética, competencia y selección. La cuasiespecie ha cobrado gran importancia en virología porque ofrece una interpretación de la amplia plasticidad, tanto genética como fenotípica, que presentan muchos virus, en particular los de RNA para adaptarse a entornos cambiantes (Domingo, 1999). El concepto de cuasiespecies fue definido por primera vez, a partir de estudios teóricos en 1977, por Eigen y Schuster (Eigen and Schuster, 1977), quienes definieron el término cuasiespecie como una determinada distribución de especies macromoleculares con secuencias estrechamente interrelacionadas, dominadas por una o varias copias maestras (degeneradas), contando estas con una mayor capacidad replicativa.

En biología, una especie es una clase de individuos caracterizados por un determinado comportamiento fenotípico. A nivel de genotipo, los individuos de una especie concreta pueden diferir, pero todas las especies están representadas por cadenas de DNA de estructura muy uniforme, aunque lo que las distingue individualmente es la propia secuencia de sus nucleótidos. Al tratar con estas moléculas, que son unidades replicativas, solamente se utilizan estas diferencias para definir las especies. Las diferencias son, por supuesto, expresadas también por diferentes características fenotípicas, como las tasas de replicación, los tiempos de vida, la tasa de error, etc. (Eigen and Schuster, 1977). De hecho, diversos resultados experimentales han demostrado que los genomas minoritarios de un espectro mutante pueden incluir mutaciones que confieran resistencia a inhibidores antivirales, anticuerpos neutralizantes o células T citotóxicas, pueden alterar la capacidad de inducir interferón o de responder al mismo, virulencia o estabilidad de partículas (Domingo et al., 2012; Agol and Gmyl, 2018; Lauring and Andino, 2010; Moreno et al., 2017; García-Arriaza et al., 2006; Briones and Domingo, 2008; Chumakov et al., 1991; Holland, 1992; Perales, 2020).

Este tipo de efectos fenotípicos se han ido observando en diversos estudios (Karamitros et al., 2020; Khateeb et al., 2022; Al Khatib et al., 2020) sobre el virus del SARS-CoV-2, surgido en Wuhan (China) a finales del año 2019 y responsable de la pandemia COVID-19 que se originó ese mismo año. Es por ello por lo que el estudio de los espectros mutantes de este virus es esencial para entender su patogénesis viral y la respuesta a las presiones selectivas, por lo que es necesario caracterizarlos mediante la secuenciación de muestras procedentes de pacientes infectados por el virus (Martínez-

González et al., 2022). La necesidad de analizar bioinformáticamente estas muestras de pacientes infectados por el virus SARS-CoV-2 es lo que ha motivado en esta tesis el rediseño y adaptación de un *pipeline* creado por Mercedes Guerrero-Murillo y Josep Gregori i Font, y basado en el paquete de R llamado QSutils (Guerrero-Murillo and Gregori, 2020), utilizado previamente para la detección de cuasispecies en el virus de la Hepatitis C (HCV).

4.2. Material y métodos

4.2.1. Implementación y flujo de trabajo

Para la detección de cuasiespecies en el virus del SARS-CoV-2 a partir de datos NGS, hemos adaptado el paquete QSutils (Guerrero-Murillo and Gregori, 2020) para su uso en los datos de secuenciación del virus del SARS-CoV-2, dando como resultado el *pipeline* llamado VQS-haplotyper (disponible en https://github.com/biotechvana/VQS-haplotyper), programado en R (R Core Team, 2022) y que consta del flujo de trabajo mostrado en la Figura 4.1.

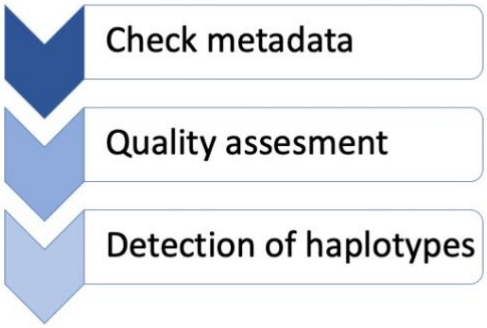

Figura 4.1. Flujo de trabajo seguido por el *pipeline* VQS-haplotyper.

La ejecución de dicho *pipeline* requiere de una serie de carpetas y ficheros necesarios para que el funcionamiento del *pipeline* sea correcto. Estas carpetas y ficheros se encuentran descritos en el repositorio VQS-haplotyper de Biotechvana alojado en la URL de Github anteriormente citada. Además, es también necesario el uso de las librerías de R llamadas Biostrings (Pagès et al., 2022), ShortRead (Morgan et al., 2009), data.table (Dowle and Srinivasan, 2021), stringr (Wickham, 2022), RColorBrewer (Neuwirth, 2022) y optparse (Davis, 2022), y del software FLASH (Magoc and Salzberg, 2011) de Illumina, proporcionado también en el repositorio.

4.2.2. Comprobación de los metadatos disponibles

Para realizar la comprobación de los metadatos disponibles, es necesario la ejecución del *script* llamado "00_CheckMetadata-RAVs-v1.17.R". Este *script* verifica toda la información relativa a las descripciones de las muestras (disponible en el fichero samples.csv) y a las descripciones de los *primers* (disponible en el fichero primers.csv), que estén presentes las secuencias de referencia por cada amplicón (dentro del fichero llamado AmpliconRefSeqs.fna) y que estén disponibles los ficheros FASTQ correspondientes a las muestras sobre las que se va a realizar la detección de cuasiespecies.

4.2.3. Análisis de calidad de las muestras y su filtrado

Para el análisis de calidad de las muestras, se ejecutó el *script* llamado "01_MiSeq_RAV_QA_Pipeline-v2.2.R", el cual se modificó añadiéndole la posibilidad de personalizar los parámetros que, antes del rediseño realizado durante este trabajo, venían por defecto. De esta manera, el análisis de calidad y el filtrado se pueden personalizar en función del virus del que provengan las muestras a analizar.

Los parámetros añadidos son el *overlap* mínimo entre las lecturas *forward* y *reverse* en FLASH (min.ov), el *overlap* máximo entre las lecturas *forward* y *reverse* en FLASH (max.ov), la fracción de *mismatches* admitidos en el *overlap* en FLASH (err.lv), y el máximo porcentaje de bases que tengan una calidad por debajo de 30 en valores PHRED por lectura (ThrQ30). Estos parámetros se añadieron utilizando la función make_option() del paquete de R optparse, especificando el parámetro, el valor por defecto, el tipo de dato que va a contener y una descripción que sirve como ayuda para entender el funcionamiento de cada uno de ellos.

Este *script* para el análisis de calidad de las muestras consta del siguiente flujo de trabajo:

6. Extensión de lecturas mediante FLASH de Illumina. Para ello, se usó el *script* "R1R2_to_FLASH_pl.R", el cual lee los ficheros FASTQ utilizados como *input* y extiende las lecturas mediante FLASH utilizando una serie de parámetros como el *overlap* mínimo (min.ov), el *overlap* máximo (max.ov) y el porcentaje máximo de *mismatches* admitidos en el *overlap* de secuencias (err.lv).

6. Análisis de calidad por posición en los ficheros FASTQ *forward* y *reverse* originales y en los ficheros FASTQ con las lecturas extendidas procedentes de FLASH usando el *script* "PoolQCbyPos-v2.3.R". Este *script* genera un informe PDF por *pool* con la calidad por posición en valores de PHRED y la distribución del tamaño de las lecturas. La calidad por posición es obtenida a partir de la función quality() del paquete de R Biostrings.

6. Análisis de calidad por lectura en los ficheros FASTQ procedentes de FLASH ejecutando el *script* "PoolQCbyRead-v1.2.R", en el que se utiliza de nuevo la función quality() del paquete de R Biostrings para obtener la calidad para cada una de las bases de las lecturas y para, posteriormente, generar un informe

donde se muestre la frecuencia de bases por lectura que tengan una calidad por debajo de 30 en valores PHRED.

6. Análisis de la distribución del tamaño de las lecturas en los ficheros FASTQ procedentes de FLASH ejecutando el *script* "LenPeaksByPool-v2.R", generando un único fichero PDF incluyendo esta información para cada uno de los *pools*.

6. Uso del *script* "FiltByQ30.R" para el filtrado de secuencias de los ficheros FASTQ procedentes de FLASH, eliminando aquellas cuyo porcentaje de bases con calidades por debajo de 30 en valores PHRED sea superior al indicado en el parámetro ThrQ30. Para ello, se obtuvo la calidad, mediante la función quality() del paquete Biostrings, de cada una de las bases. Posteriormente, se calcula el porcentaje de bases respecto del total con una calidad inferior a 30, de manera que todas aquellas secuencias con un porcentaje superior al indicado fueron eliminadas. En este paso se produce un PDF con un par de histogramas, uno por número de secuencias y otro por porcentaje de secuencias, que indican el número o porcentaje de lecturas que quedan para los posteriores análisis.

6. Por último, se obtuvo de nuevo la calidad de cada una de las bases de las secuencias filtradas de los ficheros FASTQ obtenidos con FLASH, utilizando, para ello, el *script* llamado "PoolFiltQCByPos-v2.2.R". Este *script* también utiliza la función quality() para obtener la calidad de cada una de las bases que componen las lecturas y realiza un reporte por *pool* similar al que realiza el *script* "PoolQCbyPos-v2.3.R".

Los *scripts* de este *pipeline* para el análisis de calidad y el posterior filtrado de las lecturas se modificaron para aceptar diferentes delimitadores en los ficheros samples.csv y primers.csv usando la función fread() del paquete de R llamado data.table.

4.2.4. Detección y cuantificación de cuasiespecies en muestras de virus procedentes de secuenciación de amplicones

La detección y cuantificación de cuasiespecies se realizó mediante la ejecución del *script* llamado "02_VQS-haplotyper-v1.06.R", el cual detecta las cuasiespecies presentes en las muestras procedentes de FLASH tras aplicar filtros de abundancia, similaridad, tamaño o número de indeterminaciones, entre otros. El *pipeline* se modificó respecto al original permitiendo introducir estos filtros en forma de parámetros, los cuales permiten adaptarlo a las características del organismo con el que se esté trabajando. Los parámetros que se han incluido en el *pipeline* son el número máximo de *mismatches* en un adaptador específico (pmm.mx), número mínimo de lecturas por secuencia (min.reads), número máximo de indeterminaciones que sea admisible (max.Ns), máximo número tolerado de diferencias entre las secuencias y su referencia (max.diffs), máximo número tolerado de *gaps* (max.gaps), tipo de referencia (*generic* o *consesus*) que se utiliza para filtrar (ref.type), método para el cálculo de las frecuencias (*Sum* o *Intersect*) (method), mínimo número de lecturas por las que filtrar los haplotipos antes del paso de intersección (min.rd), mínimo porcentaje de abundancia para filtrar los haplotipos antes del paso de intersección (a.cut), mínimo porcentaje de abundancia

para guardar haplotipos poco representados (ni.thr), mínimo porcentaje para el primer filtro de abundancia (var.thr), mínimo porcentaje para el segundo filtro de abundancia (ab.thr) y valor entre 0 y 1 para multiplicar el tamaño de las secuencias y seleccionar el tamaño mínimo tras la eliminación de adaptadores (min.size). Todos estos parámetros poseen valores por defecto, por lo que puede ejecutarse el *pipeline* con estos valores por defecto o modificando uno o los que sean necesarios.

Este *pipeline* ejecuta diversos *scripts* para llevar a cabo la detección de cuasiespecies. Aquellos que necesitan obtener la información de los ficheros de entrada samples.csv y primers.csv se modificaron para aceptar estos ficheros con diferentes delimitadores utilizando la función fread() del paquete de R data.table. Cada *script* que compone el *pipeline* tiene un propósito y se ejecutan en el siguiente orden:

- RAV_FF_PrimersSplit_pl-v1.7R: *script* para la eliminación de adaptadores de las lecturas de cada muestra, resultando en dos ficheros FASTA por amplicón y muestra, siendo uno *forward* si se ha encontrado en la lectura el adaptador *forward*, y siendo el otro *reverse* si se ha encontrado en la lectura el adaptador *reverse*, conteniendo los haplotipos encontrados junto con sus frecuencias (porcentaje de lecturas que soportan un haplotipo respecto del total de lecturas) y el número de lecturas que soportan un haplotipo. En este *script* se hizo dos cambios respecto al original. El primero consiste en modificar la forma en la que se realiza la búsqueda de los adaptadores en los ficheros de entrada para eliminarlos de las lecturas. Originalmente, se buscaba directamente el identificador del adaptador para extraer su secuencia para que pudiera ser eliminada. Sin embargo, esto generaba problemas cuando el nombre de un adaptador estaba contenido en otro, ya que se seleccionaban de forma simultánea y generaba un error a la hora de eliminarlos. Como solución, se utilizó la expresión regular "\b + nombre del adaptador + \b" para hacer la búsqueda por palabras completas y no contenidas en otras. El segundo cambio consiste en la introducción del parámetro min.size, que multiplica el tamaño que tendría la secuencia de referencia para seleccionar un tamaño mínimo de lectura. Esto permite la detección de deleciones en los pasos posteriores del *pipeline*.

- SeqFreqTable-MPFC-v4.5.R: *script* que alinea los haplotipos contra la secuencia de referencia del amplicón correspondiente, eliminando aquellos haplotipos que no cumplen los filtros de número máximo de gaps (max.gaps), indeterminaciones (max.Ns) o diferencias con respecto a la secuencia de referencia (max.diffs), así como aquellos que no están soportados por el número mínimo de lecturas establecido (min.reads). Este es el paso donde se realiza la llamada de variantes, detectando inserciones, deleciones y polimorfismos de un solo nucleótido (SNPs) sobre los haplotipos que han superado los filtros nombrados. La ejecución de este *script* resulta en un fichero fasta *forward* y otro *reverse* por amplicón y muestra, conteniendo los haplotipos alineados y filtrados junto con sus frecuencias y número de lecturas que lo soportan. El principal cambio en este *script* es el permitir la detección de deleciones e inserciones en los haplotipos. Esto se consigue eliminando la función en el código que corregía los *gaps* usando la secuencia de referencia y que evitaba la detección de indels y modificando parte de la función FilterCorrectHaplos() que llamaba a la antigua

función que corregía los *gaps*. Además, se corrigió la función CorrNs() para que no produjese errores a la hora de corregir las indeterminaciones a partir de la secuencia de referencia.

- nt-IntersectFW&RVHaplos-v5.3.R: *script* que intersecciona los haplotipos *forward* y *reverse* de un amplicón y muestra, eliminando aquellos que se encuentran duplicados y volviendo a calcular las frecuencias de cada uno de ellos teniendo en cuenta las lecturas de ambos ficheros. El resultado es un único fichero de haplotipos no redundantes por amplicón y muestra conteniendo la frecuencia y el número de lecturas que soportan cada haplotipo, siempre y cuando se haya encontrado este haplotipo en el fichero *forward* y en el fichero *reverse*. Como este *script* genera un informe que reporta los puntos de mutación teniendo en cuenta todos los haplotipos por amplicón y muestra, fue necesario modificarlo para que leyese las secuencias de referencia y se indicase en cada punto de mutación cuál es el nucleótido de referencia.

- AbFilterConsHaplos-050-v2.2.R: *script* que genera un fichero fasta por amplicón y muestra con aquellos haplotipos que tengan una abundancia igual o superior a la especificada en el parámetro var.thr, a partir del fichero de haplotipos no redundantes, recalculando las frecuencias de cada haplotipo teniendo en cuenta el total de lecturas de todos los haplotipos que han superado el filtro de abundancia. Como en el *script* anterior, también se reportan los puntos de mutación de aquellos haplotipos por amplicón y muestra que han pasado el filtro de abundancia, por lo que este *script* también se modificó para leer las secuencias de referencia y añadir el nucleótido de referencia para cada posición reportada. Además, se añadió la posibilidad de comparar los haplotipos contra la secuencia de referencia para buscar nucleótidos que sean divergentes respecto a ésta cuando solamente un haplotipo por muestra y amplicón haya superado el filtro de abundancia.

- MinRdConsHaplos-010-v2.2.R: *script* que genera un segundo fichero fasta por amplicón y muestra con aquellos haplotipos que tengan una abundancia igual, utilizando una abundancia más baja que en el *script* anterior, o superior a la especificada en el parámetro ab.thr, a partir del fichero de haplotipos no redundantes, recalculando las frecuencias de cada haplotipo teniendo en cuenta el total de lecturas de todos aquellos haplotipos que han superado el filtro de abundancia. También se modificó para leer las secuencias de referencia y reportar la base de referencia en los puntos de mutación encontrados en los haplotipos que han pasado el segundo filtro de abundancia por amplicón y muestra. Como en el *script* anterior, se añadió la posibilidad de reportar nucleótidos divergentes respecto a la referencia cuando solamente un haplotipo ha pasado el filtro de abundancia establecido.

- RareHplProfile-v2.3.R: *script* que genera un informe sobre el porcentaje medio del total que representan los haplotipos, teniendo en cuenta todos los amplicones, con una frecuencia de 0,01%, 0,1% y 1%. Originalmente, este *script* estaba programado para tener en cuenta únicamente tres amplicones, por lo que si se tenía más no se tenían en cuenta para el análisis. Por este motivo, se modificó para aceptar más de tres amplicones mediante el uso de un *for loop*

sobre los datos de haplotipos de baja frecuencia de todos los amplicones y obtener, de esta manera, valores promedios.

- NtFastasSummary-v1.2.R: *script* que genera un reporte sobre los cambios que hay entre los haplotipos encontrados en las muestras a partir de los ficheros generados en el *script* llamado "AbFilterConsHaplos-050-v2.1.R".
- Nt01FastasSummary-v1.2.R: *script* que genera un reporte sobre los cambios que hay entre los haplotipos encontrados en las muestras a partir de los ficheros generados en el *script* llamado "MinRdConsHaplos-010-v2.1.R".

Por último, muchos de estos *scripts* utilizan funciones procedentes del *script* llamado "global.v4.6.R". En este *script* se modificaron dos funciones. La primera fue la función llamada PosTbl.w(), la cual genera una tabla que reporta la frecuencia de un nucleótido en una posición dada de una secuencia. Se añadió a esta función el que reporte los *gaps* encontrados, es decir, las deleciones. La segunda fue la función llamada SummaryMuts.w(), la cual produce una matriz de mutaciones poblacionales dadas las frecuencias de cada secuencia. Esta función se modificó para añadir a esta matriz la información sobre el nucleótido de referencia en las posiciones donde se han encontrado mutaciones en los haplotipos. Además, se creó una nueva función, llamada mutations(), para reportar aquellas posiciones que sean divergentes con respecto a la referencia cuando solamente un haplotipo haya superado el filtro de abundancia.

4.3. Resultados

Hemos rediseñado y modificado un *pipeline* preexistente, basado en el paquete de R QSutils (Guerrero-Murillo and Gregori, 2020), para que sea capaz de detectar mutaciones puntuales e indels en los haplotipos contenidos en muestras procedentes de secuenciación de virus de RNA. Para comprobar el correcto funcionamiento de VQS-haplotyper, tanto del análisis de calidad como de la detección de cuasiespecies, se utilizaron muestras de 30 pacientes infectados por el virus SARS-CoV-2, cuyo identificador BioProject del SRA Archive es PRJEB48766, procedentes de la secuenciación de seis amplicones (dos de la región Spike y cuatro de la región Nsp12): SpkA1 (22874-23266), SpkA2 (23261-23644), Nsp12A1 (14536-14919), Nsp12A2 (14911-15297), Nsp12A3 (15289-15693) y Nsp12A4 (15670-16053). Las 30 muestras proceden de 30 pacientes admitidos en el hospital Fundación Jiménez Díaz entre el 3 y el 29 de abril de 2020, es decir, durante la primera ola de COVID-19 en España.

4.3.1. Resultados del *pipeline* para el análisis de calidad de las muestras de SARS-CoV-2

El *pipeline* de calidad tiene como principal objetivo conocer la calidad de la muestras y filtrar aquellas secuencias que no cumplen con los criterios de calidad establecidos. Este *pipeline* se ejecutó con las opciones por defecto, siendo estas min.ov 20, max.ov 300, err.lv 0,1 y thrQ30 0,05, y usando el siguiente comando para ejecutarlo:

Rscript 01_MiSeq_RAV_QA_Pipeline-v2.2.R

La ejecución de este *pipeline* produce una serie de gráficas e informes que permiten saber si las muestras utilizadas, en este caso las procedentes de pacientes infectados por el virus SARS-CoV-2, son válidas para la detección de cuasiespecies.

Como se ha visto anteriormente, el primer paso de este *pipeline* es la extensión de las lecturas mediante FLASH, obteniendo como resultado lo que se muestra en la Figura 4.2.

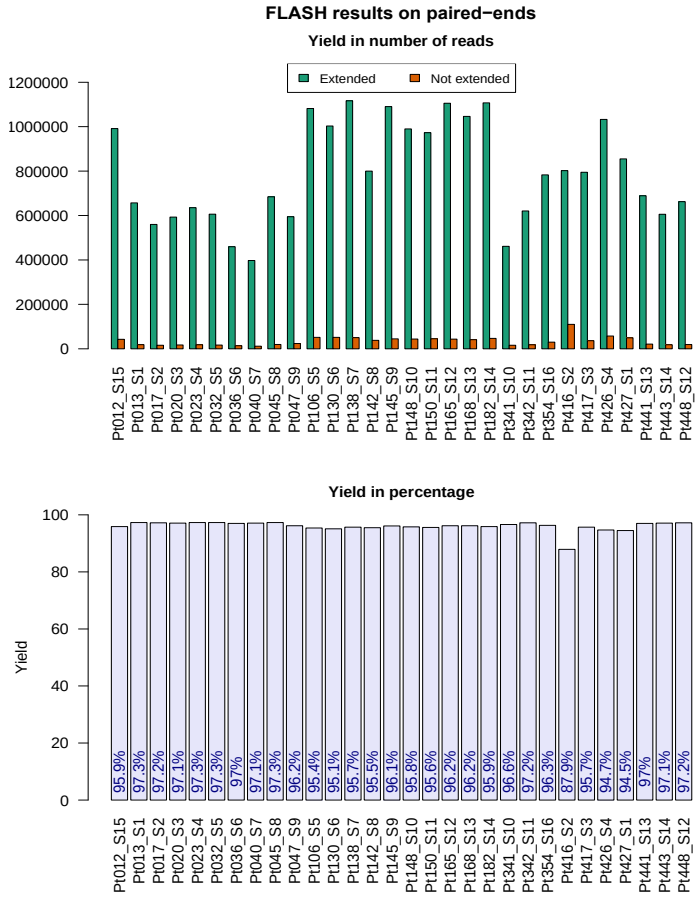

Figura 4.2. Gráficas resultantes tras la extensión de lecturas con FLASH. A) Gráfica que muestra el número de lecturas extendidas y no extendidas por muestra. B) Gráfica que muestra el rendimiento de la extensión de lecturas por muestra.

Como puede observarse en la Figura 4.2, el porcentaje de lecturas que resultaron aptas para continuar con el análisis de calidad estuvo, en todos los casos, por encima del 85%, por lo que todas las muestras se quedaron con un número suficiente de lecturas para continuar con el análisis de calidad, demostrando que la secuenciación de las muestras funcionó correctamente.

Tras la extensión de lecturas, se obtuvo la calidad por posición, la calidad media de cada lectura y la distribución del tamaño de las lecturas para cada muestra. En la Figura 4.3,

se muestra el resultado del perfil de calidad por posición en las lecturas para la muestra Pt012 a modo de ejemplo. En el caso de las lecturas pertenecientes a los ficheros *forward* (R1) y *reverse* (R2) se observó que, en todas las muestras, la calidad disminuye en el tramo final de las mismas (Figura 4.3A). Esto es un efecto común y conocido en la secuenciación mediante Illumina, ya que con el paso del tiempo se acumulan errores en la secuenciación que hacen que se produzca este efecto al final de las lecturas. Sin embargo, este efecto mejoró al fusionar las lecturas *forward* y *reverse* usando FLASH (Figura 4.3B).

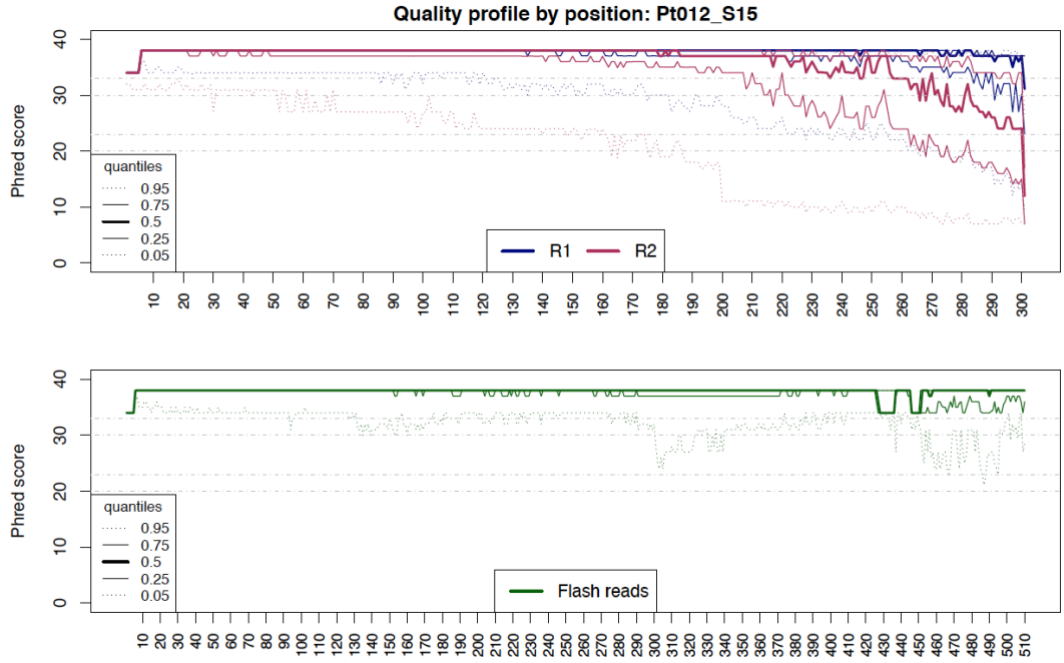

Figura 4.3. Perfiles de calidad por posición por tipos de fichero para la muestra Pt012. A) Perfiles de calidad por posición para los ficheros *forward* (R1) y *reverse* (R2). B) Perfil de calidad por posición del fichero con las lecturas extendidas.

Por otro lado, la distribución de tamaños de las lecturas permitió conocer el número de lecturas, tras la extensión de las mismas, que se quedarían con un tamaño adecuado para los posteriores análisis. Este tamaño adecuado vendría dado por el tamaño de las secuencias de referencia utilizadas para cada amplicón. Para las secuencias de referencia utilizadas en el estudio con muestras de SARS-CoV-2, el tamaño de las secuencias de referencia varía entre 384 pb y 405 pb.

En el caso de la muestra Pt012, tal y como se puede observar en la Figura 4.4, la gran mayoría de las lecturas tienen un tamaño por encima de 400 pb, indicando que la gran mayoría de ellas son lecturas válidas para los posteriores análisis en cuanto a tamaño se refiere. Esto se cumple para cada una de las muestras pertenecientes al estudio realizado.

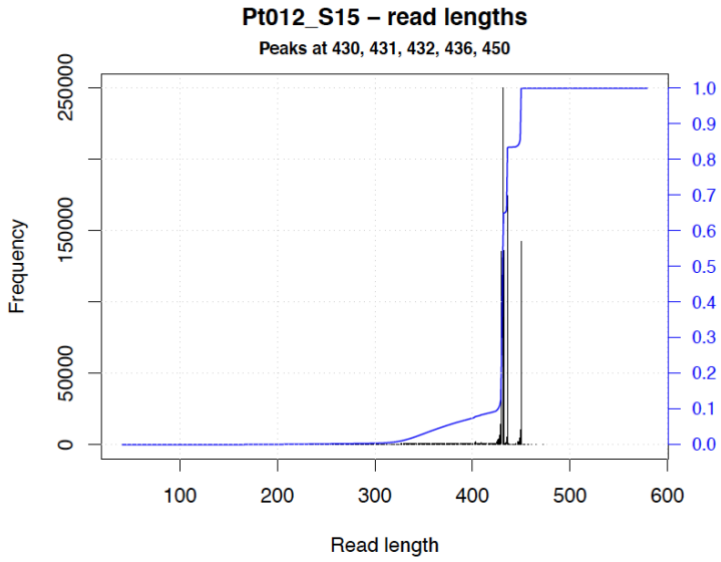

Figura 4.4. Gráfica relacionada con la distribución de lecturas y sus tamaños en la muestra Pt012 tras la extensión de lecturas.

La calidad no solamente se calculó por posición, sino que el *pipeline* generó la calidad por lectura en cada muestra, indicando el número de bases cuya calidad está por debajo de 30 en valores PHRED y la fracción de bases por lectura con una calidad por debajo de la indicada (Figura 4.5).

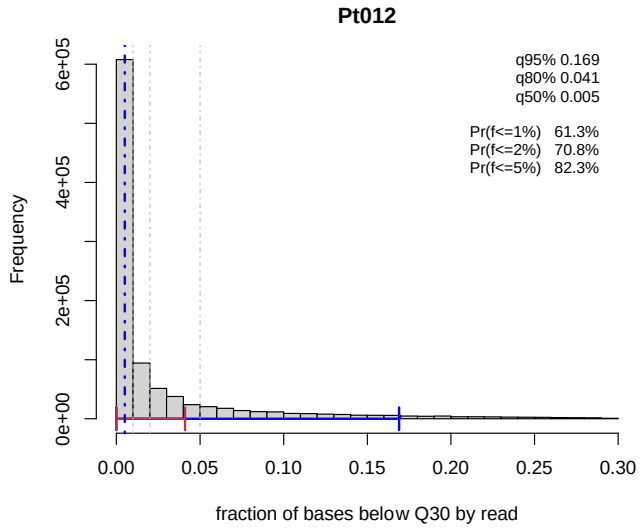

Figura 4.5. Gráfica que muestra la frecuencia del número de bases que tiene una calidad por debajo de 30 en valores PHRED para la muestra Pt012.

Tras estos análisis de calidad, se procedió a eliminar aquellas lecturas cuyo porcentaje de bases con calidad por debajo de 30 en valores PHRED se encuentre por encima del umbral establecido por defecto (5%), generando una gráfica como la que se muestra en la Figura 4.6, la cual indica el porcentaje de lecturas por muestra que se utilizaron finalmente para la detección de cuasiespecies. En todas las muestras, el porcentaje de lecturas, respecto del que se tenía inicialmente, con un porcentaje de bases con calidad

inferior a 30 en valores PHRED menor del 5%, siempre estaba por encima del 75%, tal y como se muestra en la Figura 4.6.

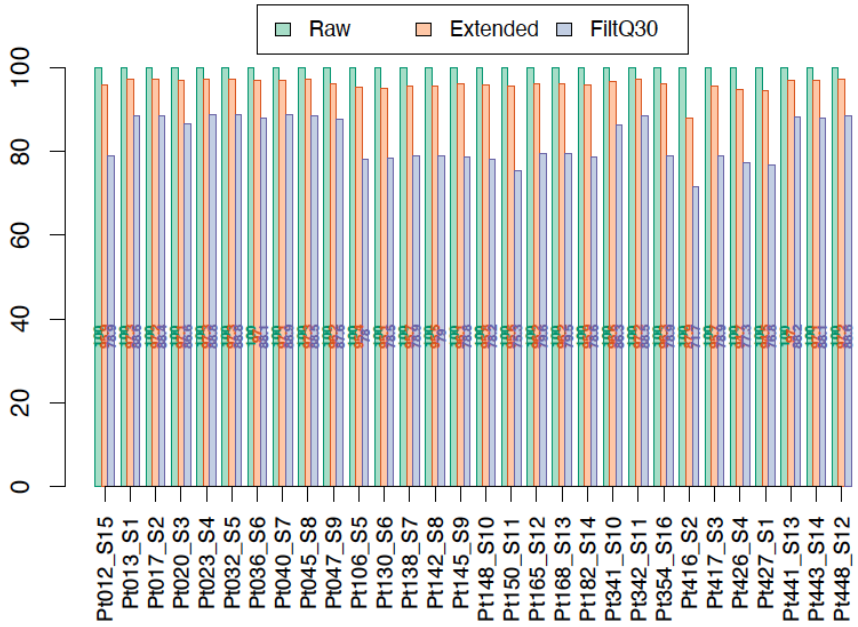

Figura 4.6. Gráfica generada tras el filtrado por calidad mostrando, para cada muestra, el porcentaje de lecturas crudas, el porcentaje de lecturas extendidas y el porcentaje de lecturas restantes tras el filtrado de calidad.

4.3.2. Resultados del *pipeline* de detección y cuantificación de cuasiespecies en las muestras de SARS-CoV-2

Tras el análisis y el posterior filtrado de calidad de las muestras de SARS-CoV-2 correspondientes al identificador BioProject PRJEB48766, se ejecutó el *pipeline* para la detección de haplotipos en estas mismas muestras utilizando los valores por defecto en los parámetros del *pipeline*, excepto para el número máximo de gaps (--max_gaps), para el cual se estableció un valor de 90, permitiendo la detección de deleciones más grandes:

```
Rscript 02_VQS-haplotyper-v1.06.R --max_gaps 90
```

Al utilizar los parámetros por defecto, se obtienen resultados para dos filtros de abundancia en los haplotipos obtenidos: 0,5% y 0,1%. En el caso de la abundancia al 0,5%, el número medio de lecturas finales que forman parte de los haplotipos por amplicón y muestra es 61.106, con un valor mínimo y máximo por amplicón y muestra entre 72 y 121.367. En cuanto al filtro de abundancia al 0,1%, el número medio de lecturas finales que forman parte de los haplotipos por amplicón y muestra es 76.210, siendo el valor mínimo de 72 y el valor máximo de 149.295. Véase la Tabla suplementaria S4.1 para más información sobre el número de lecturas por muestra, amplicón y abundancia.

Para proporcionar una imagen general de la divergencia del SARS-CoV-2 y la heterogeneidad de su espectro mutantes a una frecuencia mínima de 0,5%, se construyeron dos *heatmaps* que representan la frecuencia de cada mutación, incluyendo mutaciones puntuales y deleciones al no haberse detectado inserciones, en las regiones codificantes Nsp12 (Figura 4.7) y S (Figura 4.8), en relación con la secuencia genómica del aislado Wuhan-Hu-1 identificado con el siguiente *accession* del NCBI NC_045512.2.

Considerando todos los pacientes, el número de posiciones que incluían una variación (ya sea mutación puntual o una deleción) fue alrededor de 2 veces mayor en la región codificante S (51 con una modificación genómica de 774 posiciones analizadas) que en la región codificantes Nsp12 (polimerasa) (54 posiciones modificadas de 1.521 analizadas). Además de las mutaciones minoritarias en cada espectro mutante, también estaban presentes un total de seis mutaciones dominantes diferentes (14805, 15906, 23042, 23127, 23380 y 23403) en relación a la secuencia de referencia (aquellas con frecuencias entre el 90% y el 100%), las cuales se identifican como “Divergencia” en la Figura 4.7 y en la Figura 4.8. De esta manera, el 96,3% de las mutaciones en la región Nsp12 se encontraron con unas frecuencias que oscilaban entre el 0,5% y el 15%, mientras que sólo el 3,7% correspondía a mutaciones de divergencia. En el caso de la región S, el 92,2% de las mutaciones en esta región tenían una frecuencia entre el 0,5% y el 30%, por lo que solamente el 7,8% eran mutaciones de divergencia. Curiosamente, y teniendo en cuenta los datos de ambas regiones analizadas, 66 de las 105 mutaciones puntuales (62,9%) se detectaron con frecuencias inferiores al 2% dentro de los espectros mutantes.

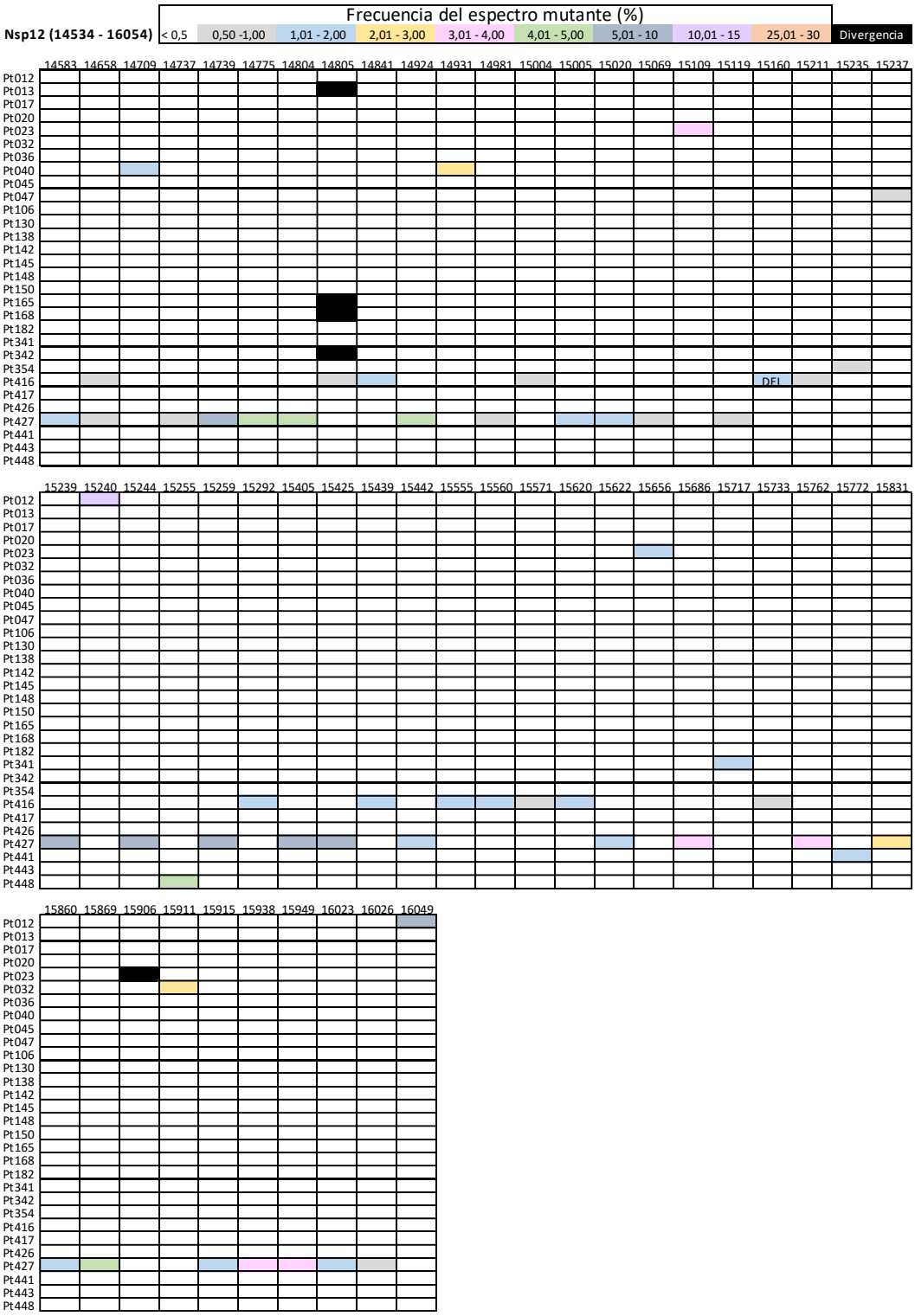

Figura 4.7. *Heatmap* mostrando las mutaciones puntuales y deleciones de la región Nsp12 de los 30 pacientes infectados por el virus SARS-CoV-2 cuya frecuencia se encuentra por encima del 0,5%.

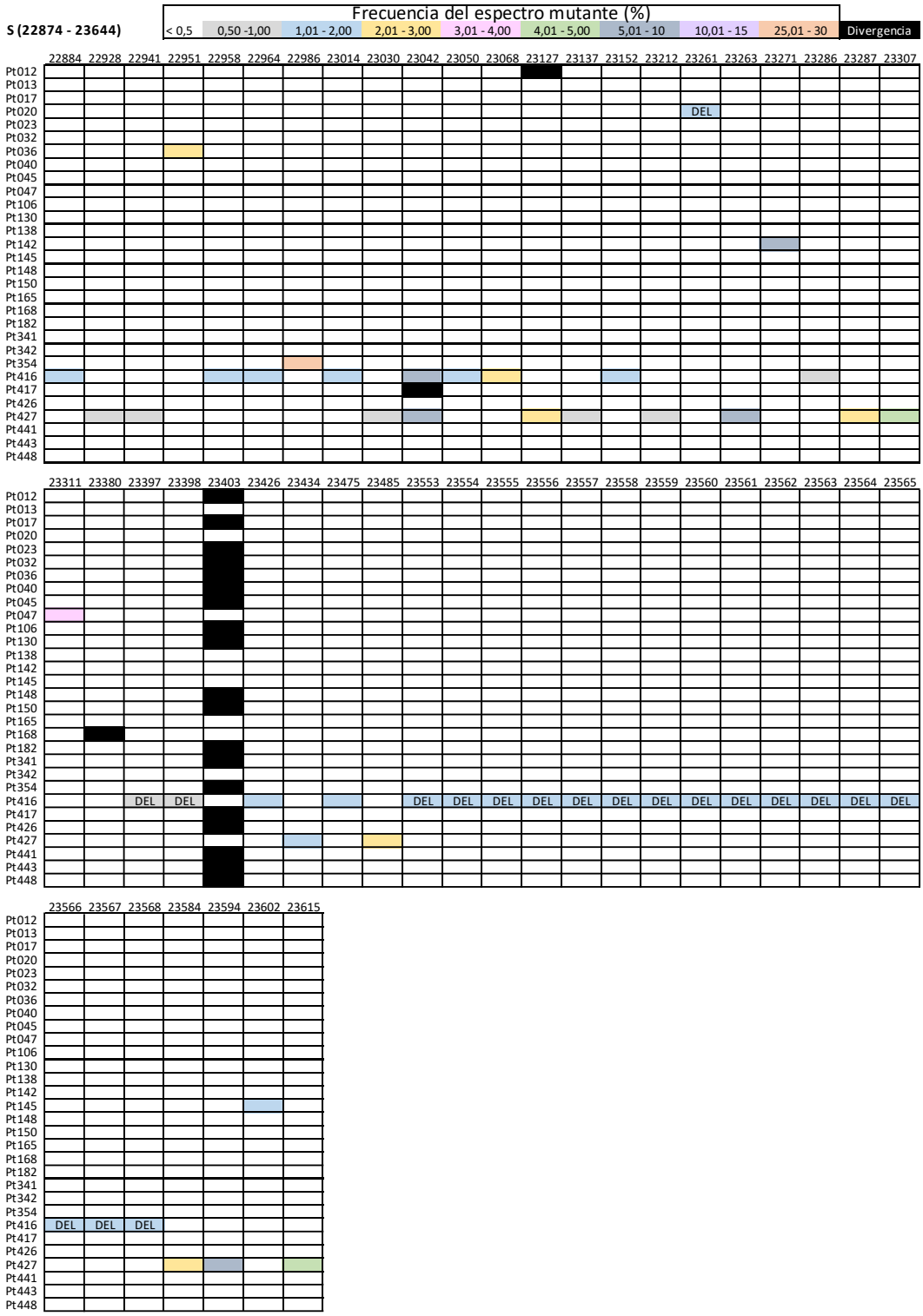

Figura 4.8. *Heatmap* mostrando las mutaciones puntuales y deleciones de la región S de los 30 pacientes infectados por el virus SARS-CoV-2 cuya frecuencia se encuentra por encima del 0,5%.

Estos mismos análisis se llevaron a cabo con un filtro de abundancia mínima de los haplotipos del 0,1% para profundizar más en la composición del espectro mutante. Se construyeron dos *heatmaps*, uno para la región codificante Nsp12 (Figura 4.9) y otro

para la región codificante S (Figura 4.10), mostrando las frecuencias de los puntos de mutación y deleciones encontradas por encima del umbral de abundancia establecido, también en relación a la secuencia del aislado Wuhan-Hu-1, la cual se utilizó como referencia.

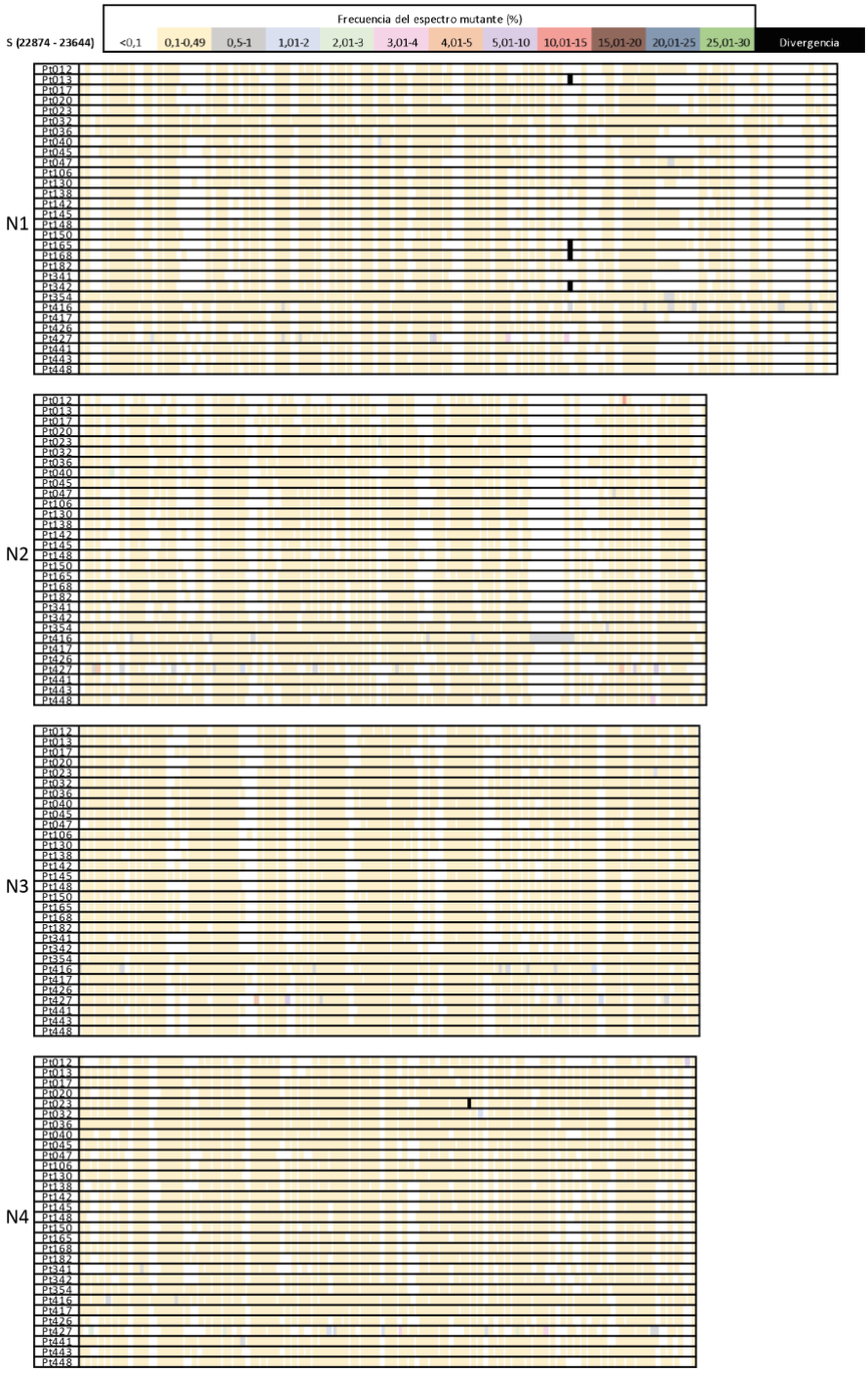

Figura 4.9. *Heatmap* mostrando las mutaciones puntuales y deleciones de la región Nsp12 de los 30 pacientes infectados por el virus SARS-CoV-2 cuya frecuencia se encuentra por encima del 0,1%. Para más información sobre las mutaciones y deleciones detectadas en la región Nsp12, véase la Tabla suplementaria S4.2.

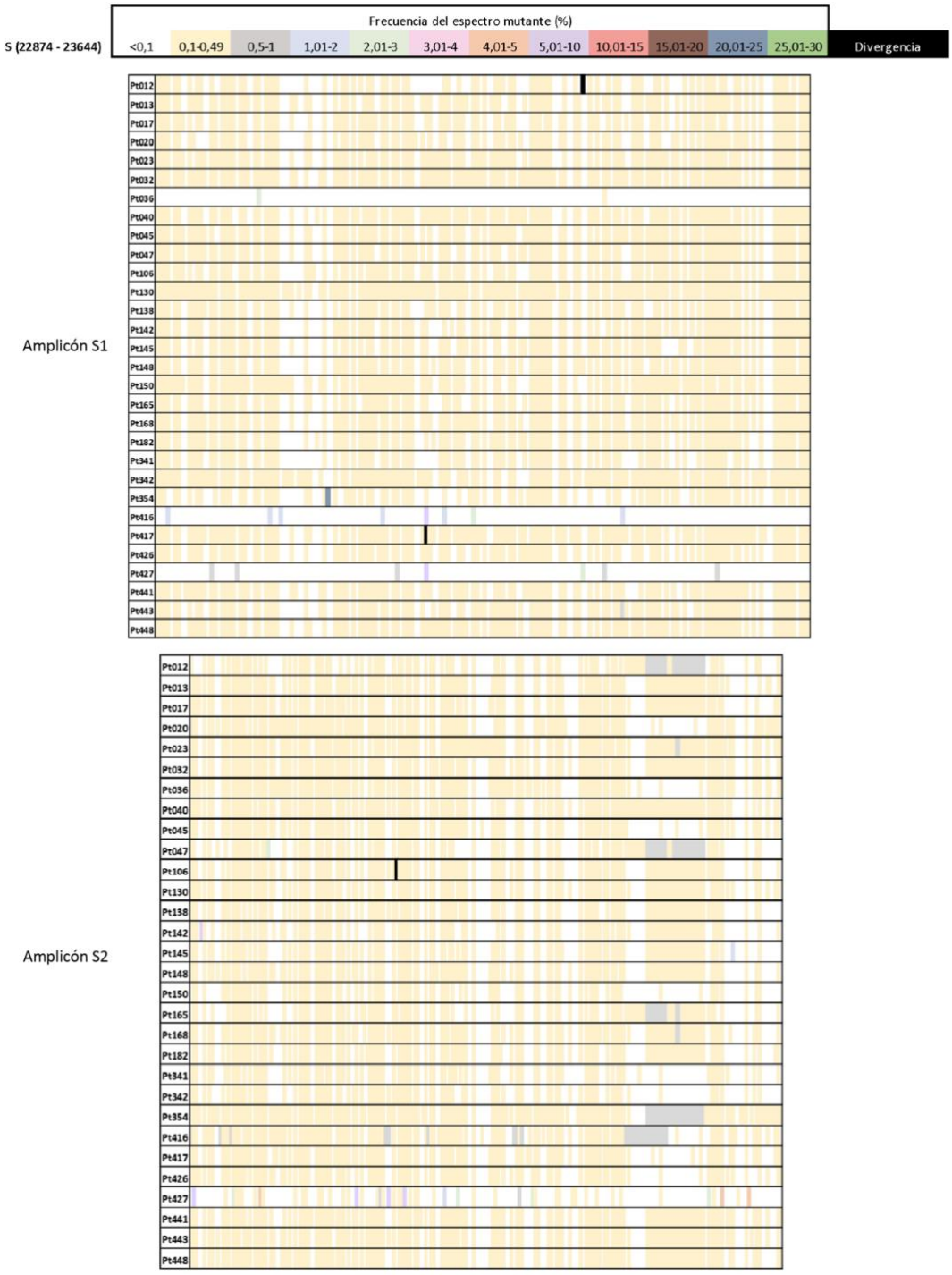

Figura 4.10. *Heatmap* mostrando las mutaciones puntuales y deleciones de la región S de los 30 pacientes infectados por el virus SARS-CoV-2 cuya frecuencia se encuentra por encima del 0,1%. Para más información sobre las mutaciones y deleciones detectadas en la región S, véase la Tabla suplementaria S4.2.

El porcentaje de posiciones con una mutación es de 49,64% (755 posiciones de un total de 1521) para la región codificante Nsp12 y de 51,55% (399 posiciones de un total de 774) para la región codificante S. También, en este caso, se detectaron las mismas mutaciones dominantes que cuando el filtro de abundancia se situaba en el 0,5% de frecuencia de los haplotipos, clasificándose también como “Divergencia” en la Figura 4.9 y en la Figura 4.10. En este caso, es decir, utilizando un mínimo de abundancia del 0,1%,

el 88,34% de las mutaciones en la región Nsp12 se encontraron con unas frecuencias, en al menos alguna de las muestras, entre 0,1% y 0,5%, el 11,39% de las mutaciones de esta región se encontraron con unas frecuencias entre el 0,5% y el 15% y solamente un 0,27% de las mutaciones se clasificaban como divergentes. En la región codificante S, el 81,70% de las mutaciones tenían una frecuencia que oscilaba entre el 0,1% y el 0,5%, el 17,30% de las mismas tenían una frecuencia entre el 0,5% y el 15% y únicamente el 1% se correspondía con mutaciones divergentes respecto a la secuencia de referencia. Cabe destacar que todas las mutaciones detectadas con el corte de abundancia al 0,5% son detectadas con el corte de abundancia del 0,1% y siempre con una frecuencia superior al 0,5%.

Por último, al reducir el valor de la frecuencia mínima de los haplotipos del 0,5% al 0,1%, se produjo un aumento significativo del número de mutaciones diferentes, el número total de mutaciones anotadas y el número de deleciones en los espectros mutantes del SARS-CoV-2, siendo este aumento menor en el caso de las deleciones. Este hecho se debe al mayor número de haplotipos diferentes que han aparecido tras el corte de frecuencia del 0,1%, lo que indica una gran sobreabundancia de mutaciones y deleciones de baja frecuencia en los espectros mutantes del SARS-CoV-2.

4.3.3. Implementación del *pipeline* VQS-haplotyper en la herramienta STATools del GPRO Suite

La herramienta STATools del GPRO Suite implementa una serie de *scripts* programados principalmente en R, aunque también en otros lenguajes de programación, para el análisis estadístico y representación gráfica de datos biológicos. Se decidió incluir el *pipeline* VQS-haplotyper en esta herramienta al estar compuesto por un flujo de trabajo completamente nuevo y estar programado en R. Centrándonos en este flujo de trabajo para la obtención de las frecuencias de los haplotipos identificados a partir de datos de secuenciación de amplicones, STATools presenta el siguiente *pipeline*:

Comprobación de metadatos -> *Pipeline* para el análisis de calidad -> VQS-haplotyper

Respecto a la comprobación de metadatos, se implementó en STATools el *script* llamado "00_CheckMetadata-RAVs-v1.17.R", el cual comprueba que todos los ficheros *input*, como son el fichero conteniendo los metadatos de las muestras, el fichero conteniendo los metadatos de los adaptadores utilizados en la secuenciación, el fichero FASTA conteniendo las secuencias de referencia para cada amplicón y una carpeta conteniendo los ficheros FASTQ correspondientes a las muestras, tienen el formato correcto. La interfaz creada a este fin es la que se puede ver en la Figura 4.11.

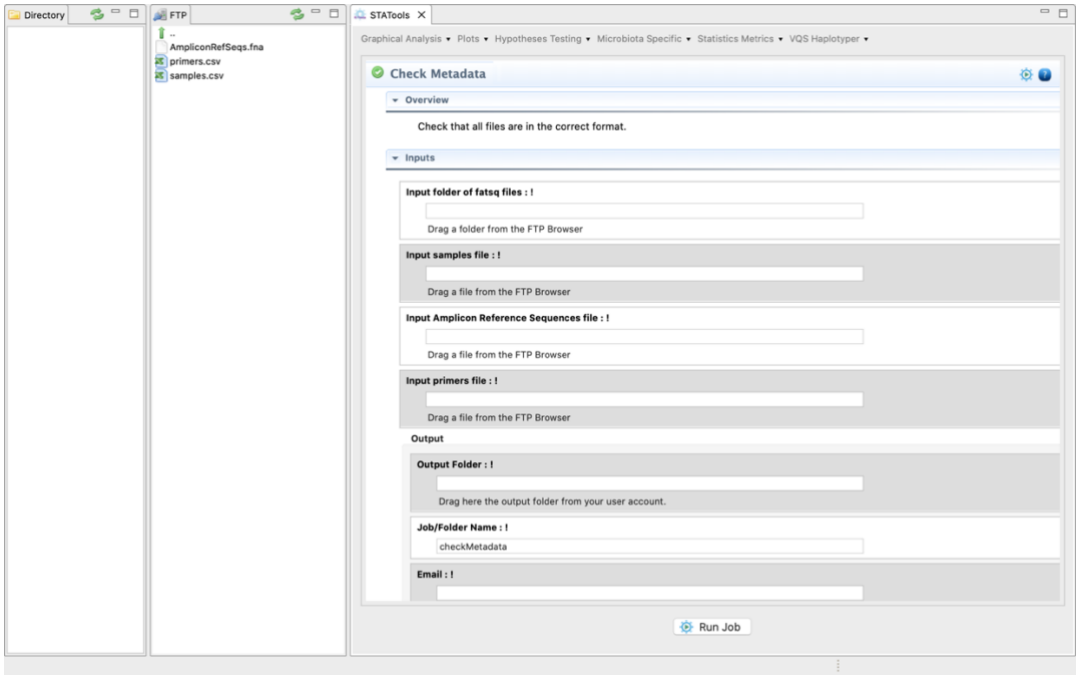

Figura 4.11. Interfaz implementada en STATools para la comprobación de metadatos.

Tras la comprobación de los metadatos, se realiza el análisis de calidad de las lecturas contenidas en los ficheros FASTQ y su filtrado por calidad. Dentro de la aplicación STATools, se implementó el *script* "01_MiSeq_RAV_QA_Pipeline-v2.2.R" junto con sus parámetros para poder adaptarlo a las muestras analizadas, y diferentes apartados en los que se puede introducir los ficheros *input* necesarios, que son los mismos que se introducen en el paso para la comprobación de metadatos. La interfaz es la que puede verse en la Figura 4.12.

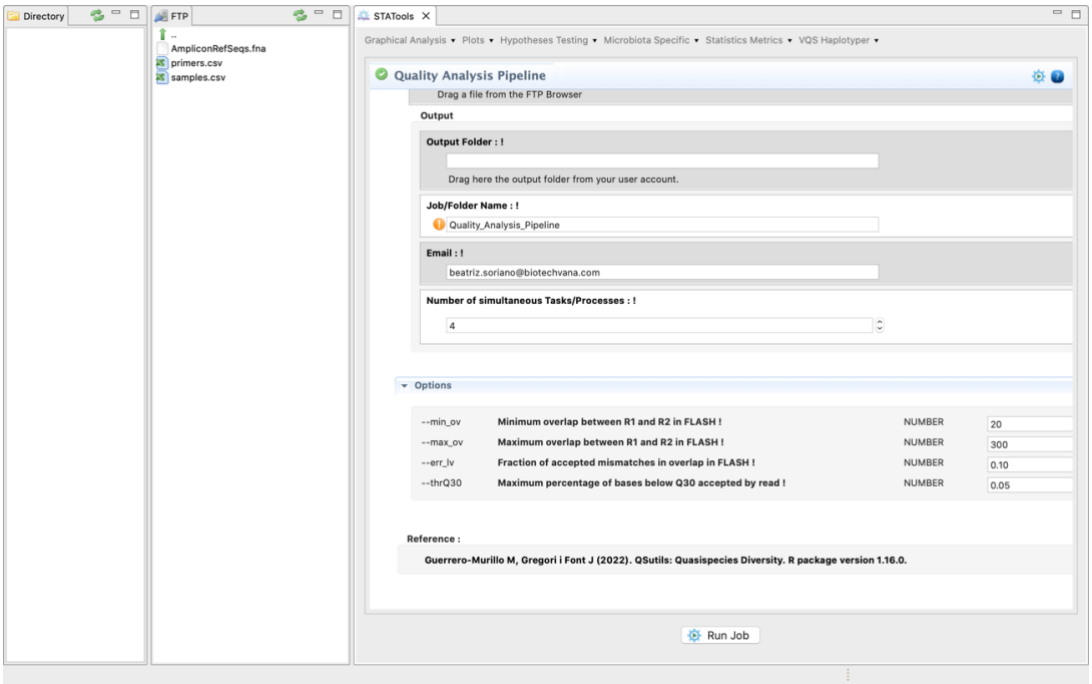

Figura 4.12. Interfaz implementada en STATools para el análisis de calidad y filtrado de lecturas en las muestras.

Una vez realizado el análisis de calidad y el posterior filtrado de las lecturas, el último paso es la ejecución del *pipeline* VQS-haplotyper para la detección y cuantificación de haplotipos presentes en las muestras procedentes de secuenciación. Para ejecutarlo, se creó una interfaz en STATools para el *script* "02_VQS-haplotyper-v1.06.R", donde se introduce como *input* el fichero con los metadatos de las muestras, el fichero de metadatos de los adaptadores utilizados para la secuenciación, el fichero FASTA conteniendo las secuencias de referencia de los amplicones y la carpeta creada tras el análisis de calidad. Esta interfaz tiene, además, la posibilidad de modificar los parámetros que el usuario considere oportunos para adaptar el *pipeline* a sus muestras, tal y como se muestra en la Figura 4.13.

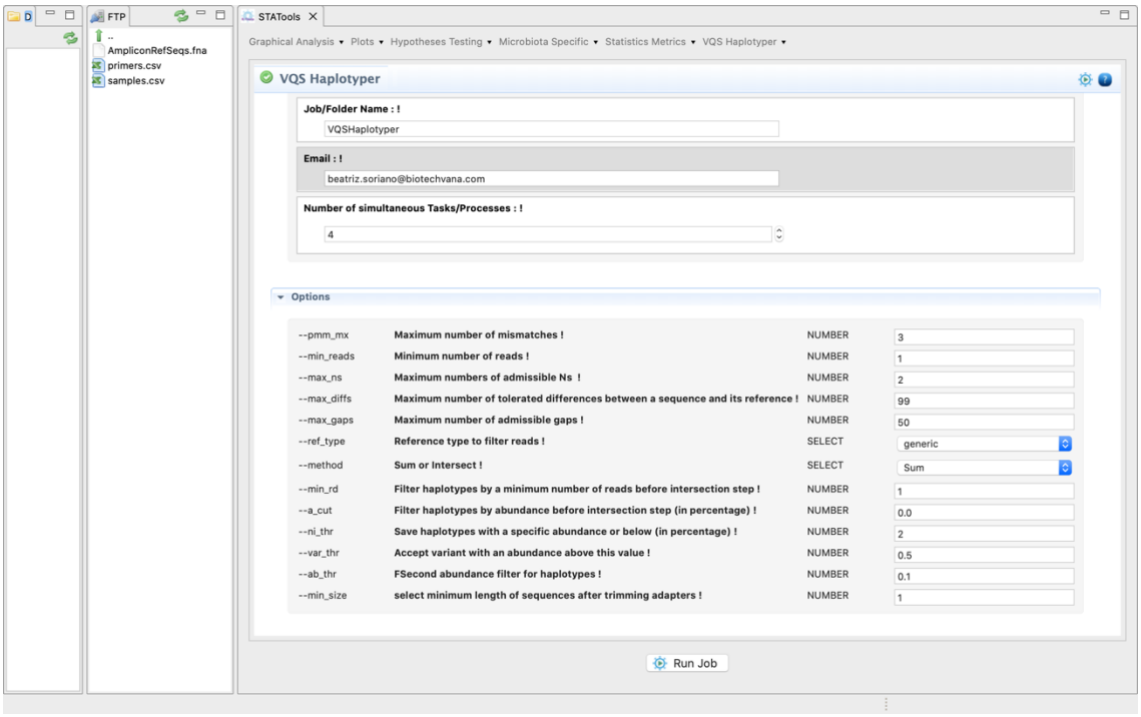

Figura 4.13. Interfaz implementada en STATools para la ejecución del *pipeline* VQS-haplotyper.

4.3.4. Resultados de la comparación entre VQS-haplotyper y SeekDeep

Los resultados procedentes del *pipeline* VQS-haplotyper se compararon con aquellos que se obtuvieron con SeekDeep (Hathaway et al., 2018), un *pipeline* pensado para su uso en datos de secuenciación de amplicones dirigidos a la estimación de la frecuencia de los haplotipos de múltiples muestras de una población.

Utilizando un filtro de abundancia de los haplotipos del 0,5%, en la región Nsp12, SeekDeep detectó 19 mutaciones puntuales y 3 deleciones que no fueron detectadas por VQS-haplotyper. Por el contrario, VQS-haplotyper detectó solamente una mutación puntual y una deleción que no aparecieron en los resultados obtenidos con SeekDeep. Respecto a la región S, SeekDeep detectó 6 mutaciones puntuales y 2 deleciones no presentes en los resultados del *pipeline* VQS-haplotyper, mientras que VQS-haplotyper identificó 12 mutaciones puntuales y una deleción que no fueron detectadas por SeekDeep. Estas diferencias se encuentran reflejadas en la Tabla 4.1, donde se indican

las posiciones donde se han detectado cambios en unas posiciones dadas con un *pipeline* pero no con el otro.

Tabla 4.1. Posiciones del genoma del SARS-CoV-2 conteniendo las mutaciones puntuales y deleciones únicamente identificadas por VQS-haplotyper o por SeekDeep al 0,5% de abundancia.

Región	**Cambios únicos**	**VQS-haplotyper**	**SeekDeep**
Nsp12	Mutaciones puntuales	15235	14554, 14558, 14597, 14665, 14859, 14923, 14936, 14946, 15135, 15189, 15323, 15357, 15393, 15583, 15661, 15700, 15735, 15756, 15868
	Deleciones	15160	14851-14858, 15203-15207, 15209-15215
S	Mutaciones puntuales	22884, 22928, 22941, 22951, 22958, 22964, 23014, 23030, 23050, 23068, 23137, 23212	23280, 23283, 23374, 23409, 23443, 23544
	Deleciones	23261	23433–23483, 23569–23582

En el caso del filtro de abundancia de haplotipos al 0,1%, las mutaciones puntuales identificadas por SeekDeep pero que no fueron detectadas con VQS-haplotyper a una abundancia del 0,5%, sí aparecieron en los resultados utilizando este filtro de abundancia más bajo, incluso la mayoría con una abundancia superior al 0,5% en alguna de las muestras. Esto es debido al mayor número de haplotipos que han pasado el corte de abundancia al 0,1%. En el caso contrario, es decir, las mutaciones puntuales detectadas por VQS-haplotyper pero no con SeekDeep a una abundancia del 0,5%, todas las mutaciones puntuales son identificadas por SeekDeep a excepción de dos en las posiciones 15235 y 22941. Respecto a las deleciones, las dos encontradas por VQS-haplotyper no se consiguieron identificar con SeekDeep con el filtro de abundancia al 0,1%. Sin embargo, VQS-haplotyper consiguió detectar las deleciones 14851-14858, 15203-15207, 15209-15213, 23450-23458 y 23569-23582, pero no las deleciones entre las posiciones 15214-15215, 23433-23449 y 23459-23483.

La comparación de los resultados obtenidos por VQS-haplotyper y SeekDeep para un mínimo de abundancia del 0,1% pueden verse en la Tabla 4.2. En esta tabla se muestra que, en la región Nsp12, VQS-haplotyper identificó 155 mutaciones puntuales y 4 deleciones no detectadas con SeekDeep, mientras que este último detectó solamente 3 mutaciones puntuales no identificadas con VQS-haplotyper. Sin embargo, detectó 7 deleciones, algunas de las cuales con un tamaño de hasta 51 nucleótidos, que no estuvieron presentes en los resultados de VQS-haplotyper. En la región S, 80 mutaciones puntuales y 2 deleciones fueron identificadas por VQS-haplotyper y no por SeekDeep, y 4 mutaciones puntuales y 4 deleciones fueron únicas de este último *pipeline*.

Tabla 4.2. Posiciones del genoma del SARS-CoV-2 conteniendo las mutaciones puntuales y deleciones únicamente identificadas por VQS-haplotyper o por SeekDeep al 0,1% de abundancia.

Región	**Cambios únicos**	**VQS-haplotyper**	**SeekDeep**
Nsp12	Mutaciones puntuales	14546, 14578, 14587, 14603, 14620, 14634, 14640, 14650, 14651, 14683, 14689, 14696, 14698, 14699, 14706, 14708, 14715, 14716, 14748, 14752, 14753, 14779, 14781, 14787, 14788, 14793, 14795, 14799, 14803, 14806, 14807, 14819, 14821, 14865, 14880, 14881, 14895, 14907, 14909, 14910, 14913, 14915, 14927, 14933, 14934, 14940, 14943, 14966, 14970, 14971, 14972, 14977, 14991, 15002, 15008, 15021, 15023, 15028, 15039, 15041, 15042, 15054, 15056, 15067, 15076, 15091, 15097, 15101, 15123, 15130, 15132, 15133, 15144, 15170, 15174, 15185, 15186, 15200, 15201, 15220, 15224, 15225, 15235, 15249, 15261, 15264, 15283, 15284, 15293, 15295, 15304, 15316, 15318, 15326, 15336, 15345, 15361, 15362, 15367, 15379, 15398, 15400, 15402, 15411, 15427, 15428, 15459, 15485, 15501, 15512, 15514, 15549, 15553, 15568, 15573, 15577, 15609, 15621, 15624, 15636, 15639, 15675, 15677, 15693, 15702, 15710, 15720, 15722, 15742, 15745, 15758, 15765, 15775, 15803, 15813, 15832, 15856, 15872, 15874, 15875, 15876, 15889, 15893, 15923, 15927, 15932, 15950, 15959, 15967, 15975, 15990, 15993, 15995, 16032, 16040	14920, 15669, 15712
	Deleciones	14849-14850, 14917-14918, 15160, 15201-15202	14576-14627, 14861-14866, 14906, 14911-14912, 15022, 15024-15040, 15214-15215
S	Mutaciones puntuales	22891, 22919, 22925, 22931, 22941, 22947, 22969, 22970, 22972, 22974, 22979, 22980, 22987, 23002, 23013, 23024, 23031, 23038, 23040, 23044, 23045, 23060, 23064, 23071, 23075, 23083, 23085, 23091, 23092, 23094, 23113, 23115, 23119, 23120, 23128, 23143, 23171, 23183, 23192, 23221, 23227, 23231, 23234, 23237, 23238, 23239, 23277, 23316, 23323, 23329, 23353, 23355, 23369, 23377, 23382, 23397, 23422, 23424, 23431, 23432, 23450, 23458, 23473, 23484, 23491, 23494, 23510, 23534, 23541, 23553, 23558, 23570, 23578, 23579, 23583, 23596, 23613, 23617, 23618, 23627	22873, 23061, 23267, 23645
	Deleciones	23148, 23261	23422-23448, 23459-23483, 23541-23552, 23583-23584

A la vista de los resultados expuestos, podemos concluir que VQS-haplotyper tiene un mejor rendimiento a la hora de identificar mutaciones puntuales, mientras que SeekDeep parece obtener mejores resultados en la detección de deleciones. Este último punto pretende ser mejorado en VQS-haplotyper en futuras versiones del *pipeline*.

4.4. Discusión

Con el incremento del uso de tecnologías de secuenciación a lo largo del tiempo, la capacidad de discernir bioinformáticamente la diversidad de secuencias presentes en una muestra vírica es fundamental para responder diferentes cuestiones biológicas relacionadas, sobre todo, con la capacidad de adaptación del virus a ambientes cambiantes. Teniendo esto en cuenta, se diseñó e implementó VQS-haplotyper a partir de la versión del año 2020 del paquete QSutils, una herramienta capaz de obtener el perfil de haplotipos víricos presentes en una muestra dada incluso cuando la frecuencia de estos es baja, evitando falsos positivos pudiendo filtrar aquellos haplotipos que estén soportados por un número bajo de lecturas, teniendo asociada una frecuencia muy baja. Además, permite descartar *outliers* al eliminar secuencias cuya calidad es baja, las cuales cuentan con un número de bases por debajo de 30 en valores PHRED por encima de un valor establecido por el usuario. Previamente al rediseño del *pipeline*, el uso del paquete QSutils solamente permitía detectar cambios de nucleótido, ya que corregía las deleciones e inserciones en base a la secuencia de referencia, y únicamente permitía trabajar con secuencias que tuviesen todas el mismo tamaño para cada amplicón. VQS-haplotyper suple estas carencias al ser capaz de manejar diferentes tamaños de secuencia al realizar alineamientos globales, lo que permite detectar y cuantificar inserciones y deleciones en los haplotipos. Con VQS-haplotyper, el usuario puede establecer unos valores específicos a las variables del *pipeline*, adaptándolos a sus muestras concretas, gracias a la implementación de parámetros introducidos a este fin. Estos parámetros permiten seleccionar el filtro por el que se eliminan secuencias de baja calidad, número de indeterminaciones, diferencias o *gaps* por los que excluir haplotipos, o filtros de abundancia, quedándonos con aquellos haplotipos que estén por encima de los valores seleccionados, entre otros. Otras modificaciones importantes implementadas en el *pipeline* permiten la detección y cuantificación de mutaciones divergentes respecto a la secuencia de referencia (mutaciones con una frecuencia entre el 90%-100%). Además, la introducción del nucleótido de referencia y el número y frecuencia de los *gaps* en los informes generados, los cuales contienen las mutaciones observadas en cada amplicón y muestra, facilitan la interpretación de los datos.

En conclusión, VQS-haplotyper permite expandir el uso de las tecnologías de secuenciación de amplicones para la detección de haplotipos, incluso de aquellos que tienen una frecuencia baja, lo que es crucial para identificar y cuantificar pequeños cambios que puedan afectar a las características de un virus como puede ser el SARS-CoV-2.

4.5. Publicaciones *peer-review* relacionadas con este capítulo en esta tesis

A. Martínez-González B, Soria ME, Vázquez-Sirvent L, Ferrer-Orta C, Lobo-Vega R, Mínguez P, de la Fuente L, Llorens C, Soriano B, Ramos R, Cortón M, López-Rodríguez R, García-Crespo C, Gallego I, de Ávila AI, Gómez J, Enjuanes L, Salar-Vidal L, Esteban J, Fernández-Roblas R, Gadea I, Ayuso C, Ruiz-Hornillos J, Verdaguer N, Domingo E, Perales C. 2022. SARS-CoV-2 Point Mutation and Deletion Spectra and Their Association with Different Disease Outcomes. Microbiology Spectrum, 10(2):e0022122.

Contribución de la autora de esta tesis a este trabajo: Rediseño y puesta a punto del protocolo y ejecución de todos los análisis bioinformáticos.

5. DISEÑO DE PROTOCOLOS PARA ESTUDIOS DE EXPRESIÓN DIFERENCIAL Y TRANSCRIPTÓMICA COMPARATIVA USANDO DATOS DE RNA-SEQ CON Y SIN GENOMA DE REFERENCIA

5.1. Contexto

La transcriptómica consiste en el estudio del transcriptoma, término que hace referencia al conjunto de transcritos de RNA que se producen a partir de un genoma y que están presentes en una célula, tejido u organismo en un momento dado, secuenciados mediante métodos de alto rendimiento. El transcriptoma es dinámico, de manera que tiende a responder a, por ejemplo, cambios en el ambiente, condiciones experimentales o a diferentes tratamientos (Sripathi et al., 2021; Lowe et al., 2017). Comprender el transcriptoma es esencial para interpretar los elementos funcionales que determinan la biología molecular de células, tejidos e individuos, siendo particularmente relevante para entender el desarrollo y las enfermedades (Wang et al., 2009).

Los objetivos clave de la transcriptómica son: catalogar todas las especies de transcritos, incluidos RNA mensajero y no codificantes (incluyendo de pequeño y gran tamaño); determinar la estructura transcripcional de los genes, en cuanto a sus sitios de inicio, los extremos 5' y 3', los patrones de *splicing* (en caso de eucariotas) y otras modificaciones postranscripcionales; y todavía más importante, cuantificar los niveles de expresión cambiantes de cada transcrito en diferentes condiciones (Wang et al., 2009). Es por todo ello que la transcriptómica ofrece conocimientos importantes sobre la estructura, expresión y regulación de genes y ha sido estudiada en muchos organismos a lo largo del tiempo (Jain, 2012; Casamassimi et al., 2017; Lowe et al., 2017).

Actualmente, la principal metodología para el estudio del transcriptoma es el RNA-seq, una tecnología que permite la secuenciación de los transcritos de RNA para, a continuación, cuantificar sus abundancias en unas muestras dadas, identificando aquellos genes que se encuentran activos en un momento determinado, y determinando diferencias de abundancia en los transcritos de las distintas muestras (Lowe et al., 2017). Las ventajas de esta tecnología frente al uso de microarrays, la primera tecnología utilizada para el estudio del transcriptoma, son el alto nivel de reproducibilidad de los datos, la identificación y cuantificación de la expresión de transcritos e isoformas desconocidas, el bajo ruido de fondo que produce la metodología y su alta resolución (Costa-Silva et al., 2017; Wang et al., 2009).

Debido a la necesidad de estudiar el transcriptoma en diferentes especies en distintas condiciones o tiempos, en este capítulo de la tesis, diseñamos e implementamos dos

protocolos RNA-seq: un protocolo de *novo* para el análisis de expresión diferencial de transcritos procedentes de muestras de especies cuyo genoma no se ha secuenciado todavía; y un segundo protocolo para el análisis de expresión diferencial a partir de muestras de resecuenciación. Para el desarrollo del protocolo de *novo* se utilizaron datos de dos especies de anisakis (*Anisakis simplex s.s.* y *Anisakis pegreffii*) y sus híbridos (Llorens et al., 2018) analizados gracias a una colaboración con el doctor Alfonso Navas del Museo Nacional de Ciencias Naturales de CSIC, y de dos especies de garrapatas, *Ornithodoros erraticus* y *Ornithodoros moubata* (Pérez-Sánchez et al., 2021; Oleaga et al., 2021), analizados gracias a la colaboración con los doctores Ana Oleaga y Ricardo Pérez-Sánchez del Instituto de Recursos Naturales y Agrobiología de Salamanca; mientras que para el desarrollo del protocolo de RNA-seq a partir de datos de resecuenciación se utilizaron muestras humanas procedentes de biopsias orales de pacientes sanos y pacientes con lesiones leucoplásicas verrugosas (Llorens et al., 2021). Esto gracias a la colaboración del profesor doctor José Vicente Bagán del Hospital General Universitario de Valencia.

5.2. Material y métodos

Para el diseño de todos los posibles flujos de trabajo (*workflow* en inglés) y/o pasos de *pipelines* e implementarlos en nuestro protocolo de transcriptómica comparativa, se contemplaron las siguientes metodologías y estrategias.

5.2.1. Análisis de calidad y preprocesado

El primer paso en nuestro protocolo de RNA-seq contempla el análisis de calidad y preprocesado de las muestras. Para el análisis de calidad se implementó la herramienta FastQC (Andrews, 2010), generando un reporte de calidad donde se especifica estadísticos básicos de las muestras, calidad de las muestras por posición en las lecturas, tamaño de las lecturas, porcentaje de GC, presencia de indeterminaciones (Ns) o presencia de adaptadores, entre otros. Tras el análisis de calidad, se procedió a implementar herramientas de preprocesado de las muestras en función de los resultados del reporte de calidad. En el caso de que las muestras tuvieran adaptadores, estos se eliminaron mediante el uso de Cutadapt (Martin, 2011) especificando la secuencia de estos. Una vez eliminados los adaptadores, se preprocesaron las muestras por calidad, tamaño y porcentaje de indeterminaciones en la secuencia, entre otros parámetros, mediante el uso de las herramientas PRINSEQ (Schmieder and Edwards, 2011), FASTX-toolkit (Hannon 2016) y Trimmomatic (Bolger 2014). En el caso de las muestras procedentes de especies para las que no se tiene un genoma de referencia, como las especies de garrapatas y las especies de anisakis, se implementó, la herramienta FASTQCollapser para eliminar las lecturas duplicadas, y la herramienta FASTQIntersect para eliminar aquellas lecturas que no estén en el fichero *forward* y sí en *reverse*, o viceversa. El uso de ambas herramientas hizo que el posterior ensamblaje

de *novo* del transcriptoma fuera más rápido y consumiera menos recursos computacionales.

5.2.2. Estrategia para el análisis de transcriptoma sin genoma de referencia

Esta estrategia, que hemos denominado "Mapping & Counting" (es decir mapeo y conteo), se implementó para abordar el análisis de expresión diferencial de transcriptomas secuenciados de *novo* y que no disponen de un genoma de referencia previo. Este es el caso de las dos especies de garrapatas y de las dos especies de anisakis y sus híbridos donde, en ambos casos, era necesario construir un transcriptoma de *novo* para cada una de las muestras mediante el uso de Oases (Schulz et al., 2012) y utilizando un rango de *kmers* de 21 a 31 en el caso de las especies de anisakis y sus híbridos, de 57 a 67 en *Ornithodoros erraticus* y de 87 a 97 en el caso de *Ornithodoros moubata*. A continuación, para cada especie se realizó un transcriptoma consenso utilizando Minimus2 del paquete Amos (Amos Package) y eliminando redundancias a una similaridad del 0,95 utilizando CD-HIT (Fu et al., 2012). Para anotar los diferentes transcriptomas consenso se usó la implementación del paquete NCBI-BLAST (Altschul et al., 1990) realizada en la herramienta DeNovoSeq (ya citada en el capítulo 1 de esta tesis). Para las especies de anisakis concretamente se usó las aplicaciones software BLASTX y BLASTn del paquete NCBI-BLAST usando las bases de datos no redundantes de proteínas (NR) y nucleótidos (NT) del NCBI (O'Leary et al., 2016), así como también la base de datos COG (del inglés, *Eukaryotic Orthologous Groups*) (Tatusov et al., 2003) como bases de datos *subject* en la anotación de los transcritos. A la postre, para los transcriptomas de las dos especies de garrapatas se utilizó la base de datos no redundante (NR) del NCBI restringida a artrópodos, la base de datos Uniprot (The UniProt Consortium, 2021) y el conjunto de proteínas procedentes del genoma de otra especie de garrapata llamada *Ixodes scapularis*. Los términos Gene Ontology (GO) (Gene Ontology Consortium, 2015), los términos InterPro (Finn et al., 2017) y los códigos de enzima se asignaron a las anotaciones en base a los *accessions* del GeneBank, para las especies de anisakis, y en base a los *accessions* de Uniprot, en el caso de las especies de garrapata, utilizando la herramienta Worksheet del GPRO Suite (Futami et al., 2011). Por último, se anotó las rutas metabólicas asociadas a los términos GO y códigos de enzima anotados a partir de la información contenida en la base de datos KEGG (en inglés, *Kyoto Encyclopedia of Genes and Genomes*) (Kanehisa and Goto, 2000) también utilizando la herramienta Worksheet del GPRO Suite.

Una vez reconstruido el transcriptoma consenso se procedió a establecer un paso de mapeo basado en los mapeadores Bowtie2 (Langmead and Salzberg, 2012), BWA (Li & Durbin 2009) e HISAT (Kin et al 2015). De entre estos, se seleccionó Bowtie2 para mapear las lecturas preprocesadas en los estudios de anisakis y garrapata contra los respectivos transcriptomas consenso generados previamente. En el caso de garrapata, el rendimiento de mapeo medio en todas las muestras pertenecientes a *Ornithodoros erraticus* fue superior a 97,4%, mientras que el de *Ornithodoros moubata* estuvo por encima del 98% en todas las muestras. En el caso de anisakis, el rendimiento de mapeo medio estuvo por encima del 97% en todas las muestras.

El siguiente paso consistió en cuantificar los patrones de expresión en base a las lecturas mapeadas contra los transcritos de los transcriptomas consenso obtenidos en cada estudio. A este fin, se implementaron dos herramientas en nuestro protocolo, Corset (Davidson y Oshlack 2014) y HtSeq (Anders et al. 2015), que permiten ambas procesar los ficheros BAM resultantes del paso de mapeo y agrupar los transcritos en clústeres y asignar el número de cuentas que han mapeado contra cada uno de ellos, generando un fichero de conteniendo las cuentas por clúster y un fichero donde se muestra la correspondencia entre los clústeres y los transcritos. De entre estas dos herramientas, se seleccionó Corset para cuantificar los patrones de expresión mediante conteo.

El último paso de esta estrategia consiste en el análisis de expresión diferencial. A este fin se implementaron dos herramientas de R, DESeq (Love et al. 2014) y EdgeR (Robinson et al, 2010). Para los estudios citados se seleccionó finalmente EdgeR. Con ello, el fichero de cuentas por clúster se utilizó como *input* para EdgeR, con el objetivo de realizar los análisis de expresión diferencial en cada estudio. Más específicamente, en el estudio sobre anisakis, las comparaciones realizadas fueron tres: híbridos vs *A. pegreffii*, híbridos vs *A. simplex s.s.* y *A. pegreffii* vs *A. simplex s.s*. En el estudio de Anisakis se usaron dos réplicas técnicas por cada especie a comparar y se consideró como significativos aquellos transcritos diferencialmente expresados cuyo valor FDR estuviera por debajo de 0,05 tras la corrección del p-valor mediante el método de Benjamini-Hochberg (Benjamini and Hochberg, 1995) aplicada por EdgeR. En los dos estudios sobre garrapatas, las comparaciones fueron las siguientes: 7 días tras alimentación vs no alimentación, 14 días tras alimentación vs no alimentación y 14 días tras alimentación vs 7 días tras alimentación. En los estudios de garrapatas se usaron también dos réplicas biológicas por condición y se consideraron transcritos diferencialmente expresados aquellos que no solo tuvieran un valor de FDR (del inglés, *False Discovery Rate*), corregido también por el método de Benjamini-Hochberg, por debajo de 0,05 sino que también tuviesen un log2 *fold change* (logFC) por encima de 1 o por debajo de -1.

5.2.3. Estrategia para el análisis de transcriptoma con genoma de referencia

La segunda estrategia en nuestro protocolo, que hemos denominado “Tophat/Hisat & Cufflinks”, se implementó para abordar el análisis de expresión diferencial de transcriptomas secuenciados con genoma de referencia previo y acompañado por ficheros GTF/GFF, que son ficheros con anotaciones sobre dicho genoma. Esto incluye información sobre las coordenadas de los genes y su estructura intrón-exón (en caso de eucariotas) y otros elementos reguladores, *non-coding* RNAs y elementos móviles. Así como también, en algunos casos, descripciones funcionales de cada gen y anotaciones de GOs y rutas metabólicas para cada gen. Como hemos indicado previamente, esta estrategia se usó para analizar datos RNA-seq resecuenciados a partir de muestras humanas procedentes de biopsias orales de pacientes sanos y pacientes con lesiones leucoplásicas verrucosas para poder determinar el perfil de expresión de los pacientes con esta patología. Más concretamente, el estudio está basado en la comparación de dos grupos, uno de 10 pacientes con Leucoplasia Verrucosa Proliferativa (PVL) y otro de 5 pacientes sanos como grupo control. Cabe añadir que esta segunda implementación

del protocolo de transcriptómica comparativa aquí diseñado se usó también en distintos estudios de transcriptoma de la dorada usando el genoma generado en el marco de esta tesis (Capítulo 1) bajo distintas condiciones experimentales. No se presentan aquí estos resultados de transcriptómica comparativa en dorada porque finalmente se han enmarcado dentro de la tesis doctoral de un compañero de CSIC perteneciente al grupo de investigación de Nutrigenómica del Centro de Torre de la Sal (en Castellón) dirigido por el profesor Jaume Pérez, también co-director de esta tesis.

El primer paso de este protocolo consistió en implementar herramientas de mapeo que permitan trabajar con ficheros GTF/GFF tanto de procariotas como eucariotas, siendo estos últimos casos de extrema complejidad dado la necesidad de gestionar los sitios de *splicing* para la reconstrucción de los transcritos. A este fin, se implementaron los mapeadores Tophat2 (Kim et al., 2013), STAR (Doubin et al. 2013) e Hisat (previamente implementado en el anterior protocolo pero que permite usar genomas de referencia y GTFs/GFFs). Finalmente, para el estudio de humanos usamos el mapeador Tophat2 para mapear las librerías FASTQ contra el genoma de *Homo sapiens* GRCh38.95, Ensembl *release* 96 (Cunningham et al., 2019). El rendimiento de mapeo de las muestras en este estudio tuvo un valor entre el 86% y el 96%. Los siguientes pasos de esta estrategia consistían en reconstruir un transcriptoma usando el fichero GTF/GFF como anotación guía para luego realizar un estudio de expresión diferencial entre el grupo control y los casos de leucoplasia. Con este propósito, seleccionamos el paquete Cufflinks (Trapnell et al., 2012) para obtener las herramientas apropiadas para realizar todos los pasos. Más concretamente, la herramienta Cufflinks fue usada para la reconstrucción de los transcriptomas de cada una de las muestras. En tanto que la herramienta Cuffdiff se usó para realizar el análisis de expresión diferencial entre el grupo de leucoplasia y el grupo control usando 10 réplicas biológicas en el primer grupo y 5 réplicas biológicas en el segundo. Nótese también que Cufflinks permite el análisis de expresión diferencial para genes, transcritos, promotores, sitios de *splicings*, etc. El análisis se realizó a nivel genes. Los genes diferencialmente expresados que resultaron significativos fueron aquellos cuyo valor de FDR (del inglés, *False Discovery Rate*) se encontró por debajo de 0,05. Los análisis estadísticos y la visualización de los datos posteriores se realizaron utilizando CummeRbund (Goff et al., 2019). Finalmente, la anotación de descripciones funcionales, términos GO, códigos de enzima y rutas metabólica de los genes diferencialmente expresados se realizó a partir de la asignación correspondiente al *accession* de Ensembl de cada gen utilizando la herramienta BioMart de Ensembl (Smedley et al., 2009; Zhang et al., 2011b) en combinación con la aplicación Worksheet de GPRO.

5.2.4. Análisis de enriquecimiento de Gene Ontology (GO) y de rutas metabólicas

Bajo las dos estrategias, los análisis de enriquecimiento de GO y de rutas metabólicas de los transcritos diferencialmente expresados se realizaron utilizando el paquete de R GOseq (Young et al., 2010). Los análisis de enriquecimiento de las rutas metabólicas de los transcritos diferencialmente expresados se realizaron utilizando la base de datos KEGG. Para ello, se utilizaron los códigos de enzima (EC) asociados a los términos GO enriquecidos para descargar los mapas KEGG (Kotera et al., 2012) y recuperar la información de las rutas implicadas. Los términos GO y las rutas metabólicas enriquecidos que mostraban p-valores ajustados por debajo de 0,05 (corrección FDR) en la distribución Wallenius se consideraron significativos.

5.3. Resultados

5.3.1. Implementación de protocolos

En este capítulo hemos diseñado un protocolo de transcriptómica comparativa basado en herramientas de ejecución por línea de comandos (en inglés, *Command Line Interface* software o su acrónimo CLI software) más comunes en el tópico. Todo ello para analizar datos RNA-seq obtenidos tanto a partir de experimentos de resecuenciación como de secuenciación de *novo*. El flujo de pasos implementados en este protocolo puede verse en la Figura 5.1. Este protocolo se basa en los siguientes pasos: “Análisis de calidad y preprocesado”, donde se proporcionan distintas herramientas para el análisis de calidad y el preprocesado de las librerías FASTQ; “Mapeo”, que ofrece herramientas para mapear las lecturas de las librerías FASTQ contra las secuencias de referencia (genoma o transcriptoma); “Ensamblaje de transcriptoma y/o Cuantificación” para ensamblar y cuantificar los patrones de expresión del transcriptoma de las muestras del caso a estudio mediante el procesamiento de los ficheros BAM obtenidos en el paso de mapeo; “Expresión diferencial” para la comparación de los distintos grupos/condiciones en una comparación dada; “Enriquecimiento diferencial” para evaluar el enriquecimiento diferencial de los términos *Gene Ontology* (GO) y/o de las rutas metabólicas. Se permite dos rutas posibles en el flujo de trabajo. La primera sigue el protocolo “Tophat/Hisat & Cufflinks” (Trapnell et al., 2012) donde los mapeadores de *splicing* como Tophat2 (Kim et al., 2013) se han combinado con el paquete Cufflinks (Trapnell et al., 2012; Goff et al., 2019) para realizar el mapeo de *splicing* y los análisis de expresión diferencial en estudios de RNA-seq de resecuenciación. En este protocolo se utilizan como referencia secuencias genómicas acompañadas de archivos GTF/GFF.

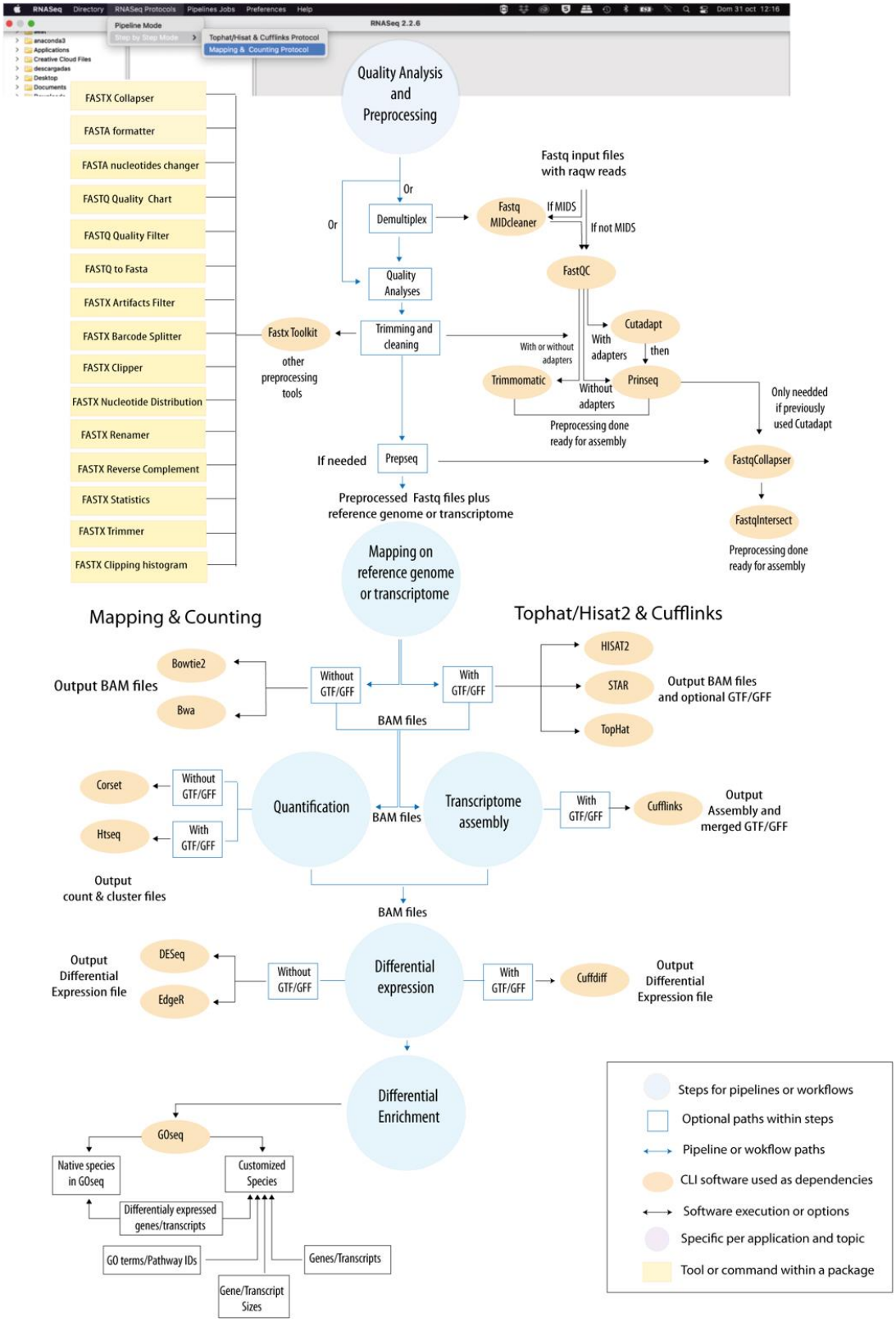

Figura 5.1. Protocolo RNA-seq. Pasos computacionales para los análisis de expresión diferencial y de enriquecimiento. Este protocolo se basa en los siguientes pasos: Análisis de calidad y preprocesado, mapeo, cuantificación, ensamblaje de transcriptoma, expresión y enriquecimiento diferenciales. La figura muestra un resumen de todas las herramientas disponibles para cada paso. Se implementan dos rutas posibles dentro de este protocolo alternativos: "*Mapping & Counting*" y "Tophat/Hisat2 & Cufflinks".

La segunda ruta del protocolo se denomina "*Mapping & Counting*" y está basada en los mapeadores de DNA/RNA, como Bowtie2 (Langmead and Salzberg, 2012), combinados con herramientas para la cuantificación del transcriptoma, como Corset (Davidson and Oshlack, 2014), para realizar el análisis de expresión diferencial, por ejemplo, con EdgeR (Robinson et al., 2010). Este flujo de trabajo se suele utilizar en estudios de RNA-seq basados en referencias de secuencias sin disponibilidad de archivos GTF/GFF como transcriptomas ensamblados de *novo*. El paso final en ambos flujos de trabajo es el análisis de enriquecimiento diferencial de categorías GO y/o rutas metabólicas utilizando GOseq (Young et al., 2010).

En pro de testar el correcto funcionamiento de estos protocolos, se utilizaron muestras de dos especies de anisakis y los híbridos de ambas (identificador BioProject PRJNA316841; BioSamples SAMN4592605 (*A. pegreffii*), SAMN04592630 (híbridos de *A. pegreffii* y de *A. simplex s.s.*) y SAMN04592599 (*A. simplex s.s.*)), muestras humanas (identificador BioProject PRJNA592439; BioSamples SAMN13426702 hasta SAMN13426716), muestras procedentes de *Ornithodoros erraticus* (identificador BioProject PRJNA666995; BioSamples SAMN16339901 hasta SAMN16339906) y muestras de *Ornithodoros moubata* (identificador BioProject PRJNA667315; BioSamples SAMN16365573 hasta SAMN16365578).

5.3.2. Resultados del protocolo de RNA-seq de *novo* aplicado a las muestras de anisakis

Las lecturas resultantes del preprocesado de las muestras (ver métodos) se ensamblaron para obtener tres transcriptomas de *novo* contribuidos por el doctor Navas, uno para *Anisakis simplex s.s*., otro para *Anisakis pegreffii* y el último para taxones híbridos formados a partir de estas dos especies. Los tres transcriptomas fueron similares en cuanto a extensión y complejidad. El ensamblaje de *A. simplex s.s.* fue más largo y tuvo un número más elevado de transcritos predichos, aparentemente debido al número más alto de isoformas encontradas (30.366 más que el transcriptoma de *A. pegreffi y* 45.059 más que el ensamblaje de las especies híbridas), aunque esto se pudo deber, probablemente, a problemas técnicos de secuenciación, a niveles elevados de *splicing* alternativo o a una elevada heterocigosidad en el transcriptoma de *A. simplex s.s.*, ya que el número de unigenes es, en cambio, bastante equilibrado (4.657 unigenes más que el ensamblaje de *A. pegreffii* y 6.989 más que el ensamblaje híbrido). A partir de estos transcriptomas, se generó un transcriptoma consenso cuyas métricas pueden verse en la Tabla 5.1, junto con las métricas de los transcriptomas por especie y el transcriptoma de los híbridos.

Los transcriptomas obtenidos fueron anotados utilizando búsquedas BLAST realizadas contra la base de datos NR, NT y COG del NCBI, considerando un e-valor mínimo de 10^{-5}. Los términos GO y las rutas metabólicas se anotaron, a continuación, a partir de los *accessions* de las proteínas que se anotaron mediante BLAST a partir de la base de datos NR. Todas las anotaciones se encuentran resumidas en la Tabla 5.1. Las mejores coincidencias detectadas por especie en el BLAST realizado contra la base de datos NR

fueron proporcionadas por proteínas de otros nematodos como *Ascaris suum*, *Brugia malayi* y *Caernorhabditis elegans* (Figura 5.2A). Aunque el transcriptoma de *Anisakis simplex s.s.* tenía el mayor número de transcritos, muchos de ellos no tuvieron homología en las búsquedas BLAST (Figura 5.2B).

Tabla 5.1. Métricas para los transcriptomas ensamblados de *novo*.

Resumen	***A. simplex s.s.***	***A. pegreffii***	**Híbrido**	**Consenso**
Tamaño total del transcriptoma (bp)	88.007.524	68.071.234	50.568.936	67.459.080
Unigenes (Loci)	36.645	31.988	29.656	-
Transcritos (isoformas)	121.907	91.541	76.848	75.380
Transcrito más largo (bp)	10.774	11.724	10.798	16.240
Transcrito más corto (bp)	100	100	100	107
% transcritos > 1 Kb	20,3	21,7	17,9	28,7
N50	973	1.026	885	1.276
L50	25.878	19.118	16.598	15.385
%A	30,87	30,25	30,11	30,41
%C	19,60	20,03	20,36	18,64
%G	19,64	19,98	20,49	20,60
%T	29,89	29,74	29,04	29,87
%Ns	0	0	0	0

Como se muestra en la Tabla 5.2, si consideramos solamente los transcritos anotados con las bases de datos NR o NT del NCBI, los tres transcriptomas resultan en un número similar de transcritos anotados (isoformas ensambladas) y unigenes (loci a los que se asigna un conjunto de transcritos). Si consideramos los identificadores de genes (GI) para aproximar el número de genes expresados, podemos afirmar que los transcriptomas reconstruidos de *A. simplex s.s.*, *A. pegreffii* y los híbridos se anotaron en función de 12.785, 12.489 y 12.324 genes potenciales, respectivamente. Se encontraron un total de 18.641 GIs no redundantes al tener en cuenta las anotaciones de los tres transcriptomas en conjunto, las cuales se aproximan al pan-transcriptoma anotado caracterizado por estas especies de anisakis. Del total de 18.641 GIs totales no redundantes, los tres transcriptomas comparten 7.535 GIs que proporcionan

información funcional adicional en forma de 3.260 anotaciones COG, 2.810 anotaciones de dominios InterPro, 1.185 términos GO y 267 códigos de enzima (Figura 5.2C). Como también se muestra en esta figura, el transcriptoma reconstruido de *A. simplex s.s.* comparte 1.463 GIs con *A. pegreffi* y 1.264 GIs con los híbridos.

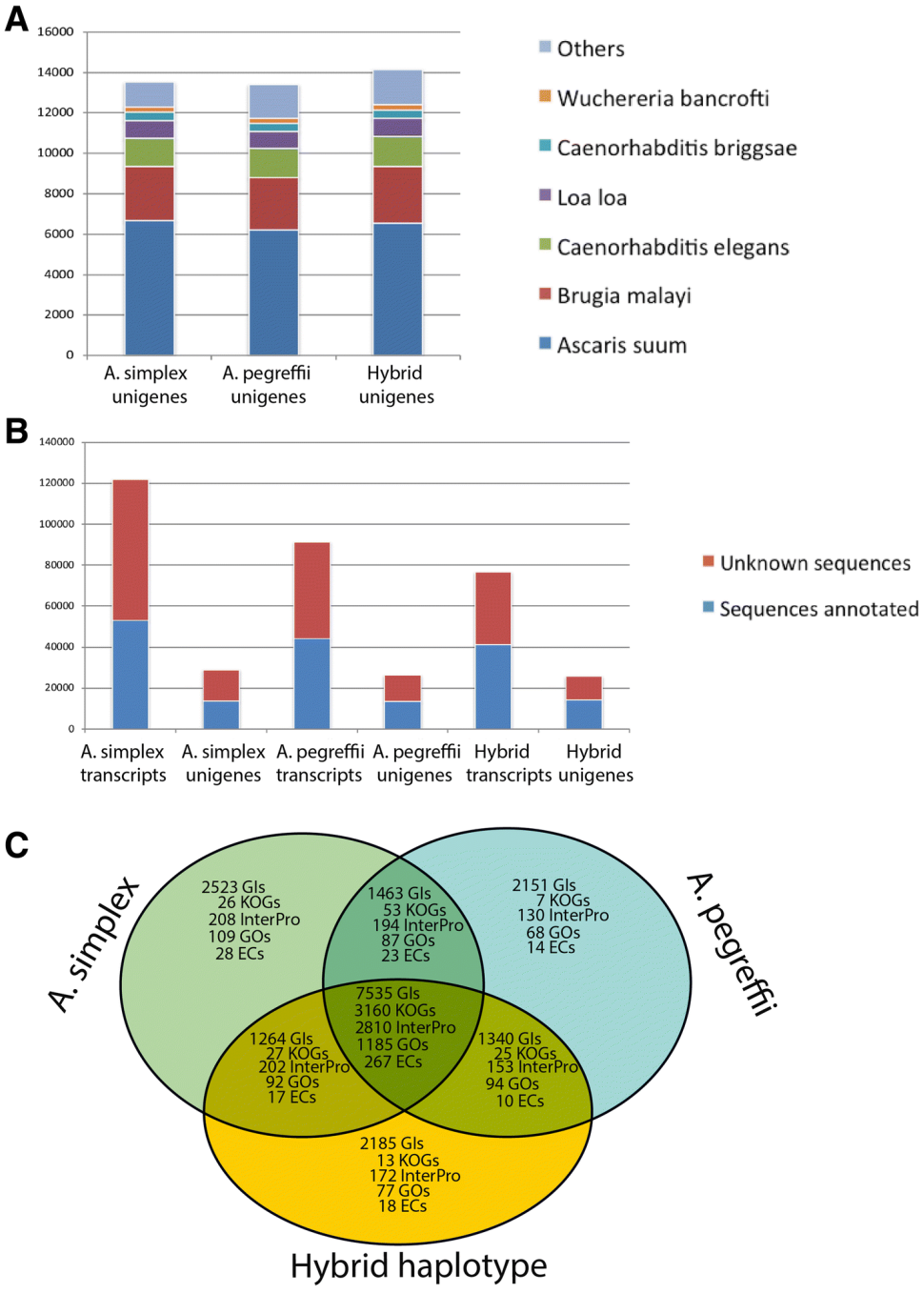

Figura 5.2. Anotaciones de los transcriptomas por especie. A) BLAST top coincidencias por especie utilizando la NR como base de datos de referencia para la anotación de unigenes. B) Distribución de transcritos y unigenes anotados y no anotados mediante BLAST utilizando la NR como base de datos de referencia. C) Diagrama de Venn mostrando las anotaciones no redundantes únicas y compartidas por transcriptoma.

A su vez, los híbridos comparten 1.340 con *A. pegreffii*. Por lo tanto, esto sugiere que, basándose en las anotaciones de los tres transcriptomas, el híbrido comparte al menos 8.875 GIs con *A. pegreffii* y 8.799 GIs con *A. simplex s.s*. En otras palabras, al menos el 48% y el 47,6% de las anotaciones del transcriptoma del híbrido tienen homólogos en *A. pegreffii* y *A. simplex s.s.*, respectivamente. *A. simplex s.s.* comparte 8.998 GIs con *A. pegreffii*, lo que significa que al menos el 48,7% de los unigenes anotados en *A. simplex s.s.* tienen homólogos en *A. pegreffi.*

Tabla 5.2. Resumen de anotaciones por transcriptoma de anisakis.

Secuencias con anotaciones	***A. simplex s.s.***		***A. pegreffii***		**Híbridos**	
	Transcritos	**Unigenes**	**Transcritos**	**Unigenes**	**Transcritos**	**Unigenes**
Identificadores de genes (GI) NR/NT	62.483	14.057	60.816	14.393	53.741	14.183
Clústeres COG	38.717	9.999	29.977	9.444	28.114	9.928
Dominios InterPro	17.093	5.938	13.514	5.122	12.884	5.125
Términos GO	14.701	5.165	11.773	4.485	11.101	4.501
Códigos de enzima	5.416	1.995	4.237	1.671	3.913	1.617
Proteínas predichas	49.755	14.384	41.324	13.506	38.782	13.834

El transcriptoma resultante de la fusión de los transcriptomas de las distintas especies y sus híbridos consistió en un total de 75.380 secuencias consenso, las cuales se utilizaron para realizar el mapeo y comparar los niveles de expresión de los transcritos de las especies de anisakis y sus híbridos. De estos 75.380 transcritos consensos, 32.999 no tuvieron anotación y se correspondieron con secuencias desconocidas expresadas en al menos 2 de los taxones, por lo que podrían tratarse de proteínas desconocidas, potenciales RNAs no codificantes y/o, incluso, elementos móviles. Los 42.381 transcritos consensos restantes correspondieron a secuencias anotadas en base a 12.511 proteínas diferentes, lo que significa que cualquier interpretación biológica de la expresión diferencial y el enriquecimiento basado en el transcriptoma consenso se basa en los patrones de expresión de 12.511 genes codificantes. Tras el mapeo contra el transcriptoma consenso, las 75.380 secuencias consenso se agruparon en 74.751 clústeres, corrigiendo el sesgo potencial en las isoformas, tal y como se detectó en el transcriptoma de *Anisakis simplex s.s*. El BCV (del inglés, *Biological Coefficient of Variation*) y la dispersión media de las muestras utilizadas se infirieron y evaluaron para valorar la idoneidad de las librerías FASTQ, resultando en valores bastante adecuados (BCV = 0,4359 y dispersión = 0,19005) para un estudio de expresión diferencial. A esta misma conclusión se llegó al trazar el BCV contra el logaritmo medio de las cuentas por

millón (CPM) de las lecturas mapeadas (Figura 5.3A). Adicionalmente, se representó una gráfica MDS (del inglés, *Multidimensional Scaling*) basada en el logFC de las diferencias entre los distintos taxones, la cual mostró que mientras las muestras utilizadas como réplicas se acercaban unas a otras no presentando diferencias entre las réplicas a excepción de los híbridos que eran más heterogéneos, las muestras de diferentes taxa estaban muy separadas, lo que indica que hay diferencias importantes entre las especies de anisakis y sus híbridos (Figura 5.3B). Consistente con estos hechos, de los 74.751 clústeres, 8.239 resultaron diferencialmente expresados en el análisis realizado entre los híbridos y *A. pegreffii*, 23.549 estuvieron diferencialmente expresados en la comparación realizada entre los híbridos y *A. simplex s.s.*, y 24.813 resultaron diferencialmente expresados al comparar *A. pegreffii* y *A. simplex s.s.*, siempre utilizando un valor FDR menor a 0,05 para considerar estos transcritos diferencialmente expresados como significativos (Figura 5.3C). Estos resultados en los análisis de expresión diferencial sugieren que los patrones de expresión de *A. pegreffii* tienen un mayor peso en los híbridos. Los resultados completos de los análisis de expresión diferencial se pueden ver en la Tabla suplementaria S5.1, proporcionada con el material electrónico suplementario adjunto a esta tesis.

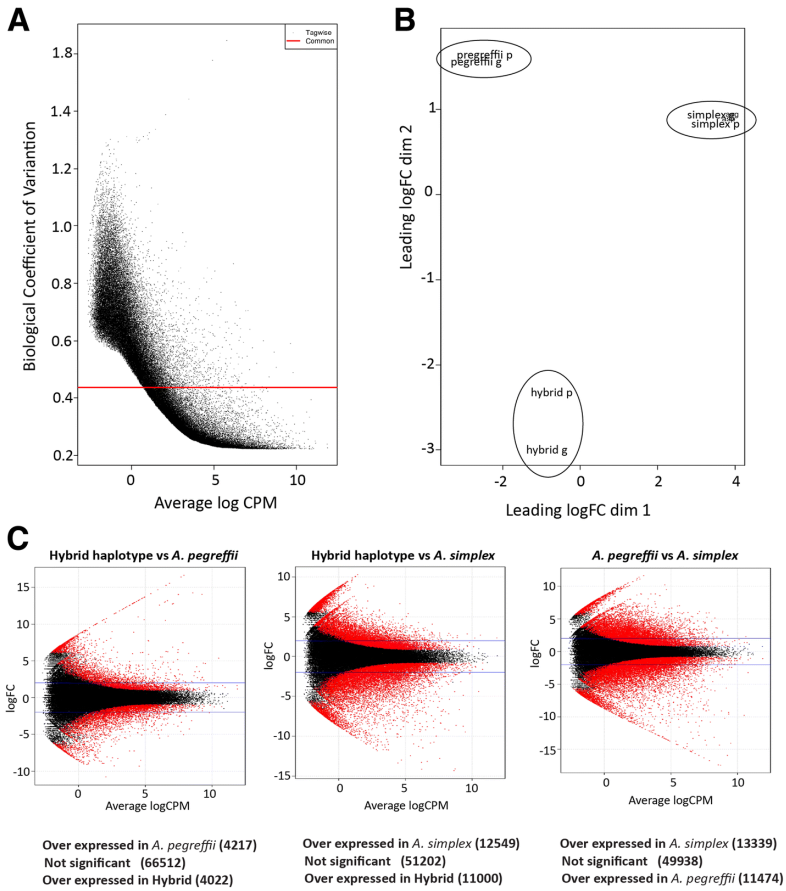

Figura 5.3. Patrones de expresión diferencial en los transcriptomas. A) BCV frente al logaritmo medio de las cuentas por millón (CPM). B) Gráfica MDS basada en el logFC de las diferencias entre los taxones y las réplicas secuenciadas por taxón. C) Gráficos MA (uno por comparación realizada) representando el logFC frente al log CPM medio por cada clúster en cada par de muestras comparadas. Los clústeres con un valor FDR menor de 0,05 se representan en rojo. Cada gráfico MA va acompañado de un resumen de los resultados de cada comparación de expresión diferencial.

Los 272 códigos de enzima identificados como comunes entre los tres transcriptomas se vincularon a 53 rutas metabólicas relacionadas con los requerimientos nutricionales y la biosíntesis, el metabolismo energético, el metabolismo de xenobióticos, el procesamiento de información ambiental, la transducción de señales y el procesamiento de información genética. Diez de las 53 rutas se detectaron como diferencialmente enriquecidas en al menos una de las comparaciones realizadas con un valor FDR menor de 0,05 (Tabla suplementaria S5.2).

Por último, también se realizaron análisis de enriquecimiento de GOs basados en los 1.211 términos GOs compartidos por los 3 transcriptomas, los cuales revelaron un enriquecimiento diferencial de 91 GOs con un FDR menor a 0,05 en al menos una de las tres comparaciones (Tabla suplementaria S5.2). Cuarenta de estos 91 GOs diferencialmente enriquecidos correspondieron a GOs de procesos biológicos, 37 a GOs de procesos moleculares y 14 a GOs de componentes celulares (Figura 5.4).

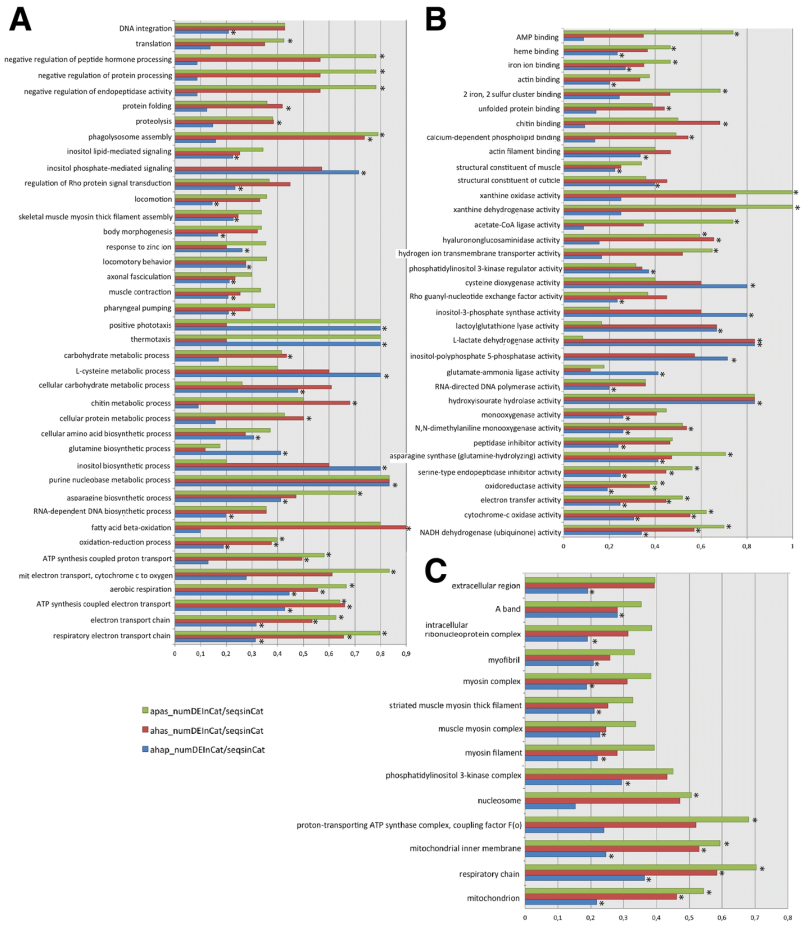

Figura 5.4. Enriquecimiento diferencial de ontologías GO. A) Gráfica de barras basada en la relación entre el número de transcritos diferencialmente expresados y los transcritos ensayados con anotaciones GOs de procesos biológicos detectados como enriquecidos diferencialmente en al menos 1 de las 3 comparaciones realizadas. Las barras coloreadas en azul hacen referencia a la comparación entre los híbridos y *A. pegreffii*; las barras en color rojo hacen referencia a la comparación entre los híbridos y *A. simplex s.s.*; y las barras verdes hacen referencia a la comparación entre *A. pegreffii* y *A. simplex s.s*. Los GOs que fueron significativos con un FDR menor a 0,05 están resaltados con un asterisco. B) El mismo gráfico de barras basado en los GOs de función molecular. C) El mismo gráfico de barras basado en los GOs de componentes celulares.

Trece de estos 91 fueron significativos en todas las comparaciones y corresponden a funciones, procesos y sub-localizaciones asociadas con el metabolismo energético, incluyendo principalmente oxidasas, oxidorreductasas, deshidrogenasas y otras enzimas y proteínas de unión con roles asignados a procesos redox y portadores de electrones. El enriquecimiento del resto de los GOs mostrados en la Figura 5.4 fue significativo en uno o dos de las tres comparaciones realizadas e hicieron referencia a funciones moleculares, procesos biológicos y componentes celulares relacionados con las rutas metabólicas enriquecidas a las que nos hemos referido previamente, pero también con la transposición, el transporte, el crecimiento y el desarrollo, la locomoción, y con la región extracelular. En términos generales y para casi todos los términos GO evaluados, la relación entre los transcritos expresados diferencialmente y los ensayados fue, de nuevo, menor si se comparan los híbridos con *A. pegreffii* que cuando se comparan los híbridos con *A. simplex s.s.* o cuando se compara este último con *A. pegreffii*. A su vez, la comparación entre los híbridos y *A. pegreffii* presentó un enriquecimiento mucho mayor que las demás: de los 91 GOs enriquecidos significativamente, 67 términos fueron significativos en la comparación entre los híbridos y *A. pegreffii*, mientras que 27 GOs fueron significativos cuando se comparan los híbridos con *A. simplex s.s.* y 34 cuando se compara este último con *A. pegreffii*.

5.3.3. Resultados del protocolo de RNA-seq de *novo* en muestras de *Ornithodoros erraticus*

El objetivo de este estudio era ensamblar de *novo* y anotar el transcriptoma a partir de 6 muestras de la especie *Ornithodoros erraticus* contribuidas por los doctores Ana Oleaga y Ricardo Pérez-Sánchez para luego realizar el análisis de expresión diferencial entre tres condiciones fisiológicas diferentes: sin alimentación (0 días), 7 días tras alimentación (7 días) y 14 días tras alimentación (14 días). En primer lugar, se realizó un transcriptoma por muestra, obteniendo 82.838 y 88.590 transcritos para las muestras de la condición sin alimentación, 61.373 y 53.779 transcritos para las muestras de la condición 7 días tras alimentación, y 73.236 y 77.721 transcritos para las muestras de la condición 14 días tras alimentación. En todos los casos se tuvo en cuenta las lecturas con al menos 100 nucleótidos. Los transcriptomas consenso para cada condición resultaron en 106.223, 75.491 y 93.846 transcritos para las condiciones sin alimentación, 7 días tras alimentación y 14 tras alimentación, respectivamente. Adicionalmente, un transcriptoma consenso se obtuvo a partir de la fusión de los transcriptomas ensamblados para cada una de las muestras, resultando en 103.041 transcritos con las siguientes métricas: N50 de 1.884 bp, 1.194 bp de tamaño medio de transcrito y siendo 26.653 bp el tamaño del transcrito más largo. Para más información sobre las métricas completas de estos ensamblajes, véase la Tabla suplementaria S5.3. También véase la Tabla 5.3 para más información sobre las métricas básicas de los transcriptomas por condición y del transcriptoma consenso. Los 103.042 transcritos del transcriptoma consenso se filtraron en base a: filtro de redundancia, resultando en 97.343 clústeres; filtro en base al nivel de expresión, de manera que se eliminaron aquellos transcritos con un valor RPKM (del inglés, *Read Per Kilobase per Million reads*) menor a 1 ya que lo más probable es que sean artefactos del proceso de ensamblaje o

que representen expresión de fondo (de Castro et al., 2017), de manera que quedaron 28.061 transcritos; filtro por tamaño de ORF (del inglés, *Open Reading Frame*) predicha por transcrito, eliminando transcritos con un tamaño de ORF < 240 pb desde el codón de inicio hasta el codón de paro, resultando en 22.007 transcritos finales (Tabla 5.3).

Tabla 5.3. Ensamblaje del transcriptoma de la glándula salival de *O. erraticus* y anotación.

	Resumen	0 días	7 días	14 días	Consenso
Métricas de los ensamblajes de los transcriptomas	**Tamaño del transcriptoma**	130.120.686	93.400.441	118.099.223	123.061.068
	Número de transcritos	106.223	75.491	93.846	103.041
	Transcrito más largo (bp)	17.265	17.479	26.653	26.653
	Transcrito más corto (bp)	100	100	100	100
	% de transcritos > 1 kb	41,32	40,83	41,52	39,97
	Tamaño medio de transcrito (bp)	1.224	1.237	1.258	1.194
	Mediana del tamaño de los transcritos (bp)	815	797	817	780
	N50 (bp)	1.875	1.979	1.978	1.884
	L50	19.869	13.592	16.813	18.414
Estadísticas de anotación del transcriptoma	**Clústeres**	—	—	—	97.343
	Transcritos > 1 RPKM	—	—	—	28.061
	ORF completa > 240 bp	—	—	—	22.007
	Anotado en al menos una base de datos	—	—	—	18.961 (86%)
	Códigos de enzima asignados	—	—	—	3.608 (16,4%)
	Términos GO asignados	—	—	—	9.249 (42%)
	Rutas KEGG asignadas	—	—	—	3.725 (16,9%)
	Identificadores InterPro asignados	—	—	—	9.116 (41,7%)
	Número de proteínas no redundantes	—	—	—	9.355

Las búsquedas por BLAST de los 22.007 transcritos contra las bases de datos NR del NCBI, Uniprot y las proteínas del genoma de la especie *Ixodes scapularis* dio como resultado la anotación de 18.961 (86,16%) secuencias con e-valor menor a 10^{-5}, de los cuales 14.219 (75%) mostraron una similaridad de secuencia por encima del 60%. Véase la Tabla suplementaria S5.4 para ver la anotación completa. Las 3.046 (13,84%) secuencias restantes no mostraron una homología significativa contra ninguna de las bases de datos de referencia. Estas secuencias podrían representar proteínas todavía desconocidas, secuencias codificantes mal ensambladas sin importancia biológica o incluso RNAs largos no codificantes que pueden ser difíciles de distinguir de las ORFs mal ensambladas (de Castro et al., 2017; Dinger et al., 2008).

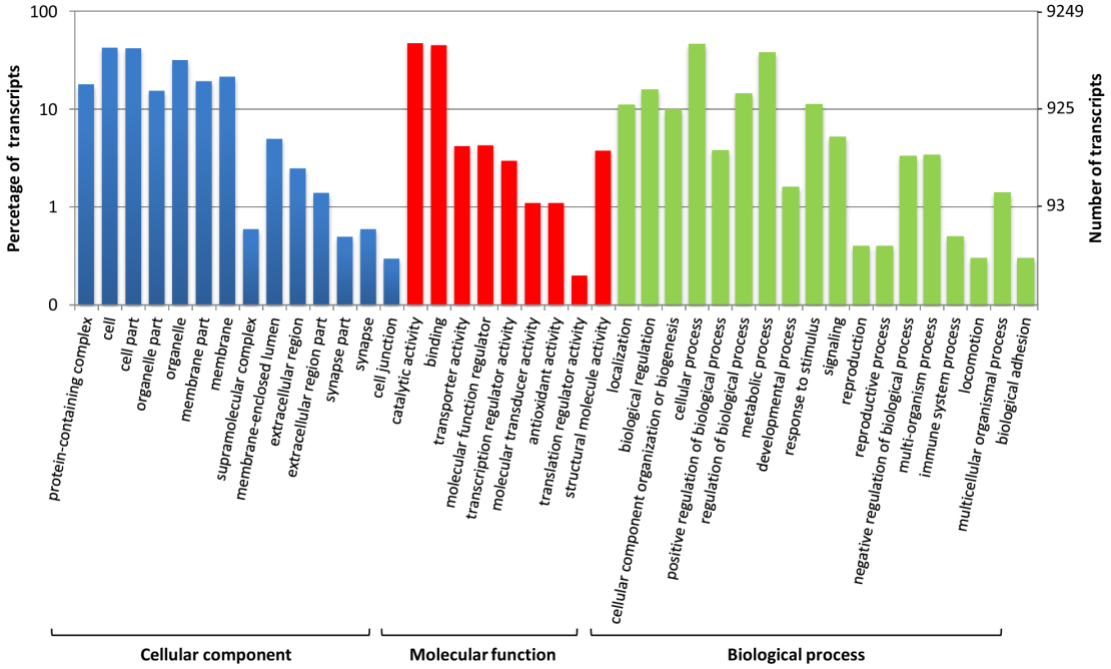

Figura 5.5. Distribución de ontologías de genes del transcriptoma de *Ornithodoros erraticus*. Los términos GO de nivel 2 de componentes celulares se representan en azul, de funciones moleculares en rojo y de procesos biológicos en verde. Estos incluyen 18.719 componentes celulares, 10.222 funciones moleculares y 15.476 procesos biológicos. Las barras representan el porcentaje y el número de transcritos anotados en cada categoría.

Las 18.961 secuencias anotadas correspondieron con 9.355 proteínas predichas no redundantes con *accessions* únicos y fueron funcionalmente caracterizadas utilizando las bases de datos GO y KEGG mediante su asociación a las anotaciones Uniprot. Se asignaron términos GO a un total de 9.249 transcritos, incluyendo 18.719 componentes celulares, 10.222 funciones moleculares y 15.476 procesos biológicos. En la Figura 5.5 se representa los transcritos clasificados según el componente celular, la función molecular y el proceso biológico, utilizando términos GO de nivel 2. Los componentes celulares se clasificaron en 14 categorías, de las cuales las 7 más abundantemente representadas fueron célula (n = 3.969), parte de la célula (n = 3.907), orgánulo (n = 2.939), membrana (n = 1.994), parte de membrana (n = 1.795), complejo que contiene proteínas (n = 1.667) y parte del orgánulo (n = 1.435). La clasificación por función molecular dio lugar a 9 categorías. Las más abundantemente representadas eran las de actividad catalítica (n = 4.412) y de unión (n = 4.187). Las categorías restantes estaban

significativamente menos representadas e incluían función molecular reguladora (n = 397), actividad transportadora (n = 386), molécula estructural (n = 349), regulación de la transcripción (n = 274), actividad antioxidante (n = 101), transductor molecular (n = 98) y regulación de la traducción (n = 18). La clasificación de los procesos biológicos dio lugar a 18 categorías. Las 8 más abundantes fueron: procesos celulares (n = 4.281), proceso metabólico (n = 3.518), regulación biológica (n = 1.458), regulación del proceso biológico (n = 1.335), respuesta a estímulo (n = 1.037), localización (n = 1.019), organización o biogénesis de componentes celulares (n = 916) y señalización (n = 477).

El análisis de los transcritos anotados en la base de datos KEGG permitió identificar las rutas metabólicas activas en las glándulas salivales. De este modo, se asignaron 3.725 secuencias a 627 enzimas, 100 rutas y 13 clases de rutas (Tabla suplementaria S5.5). Las 30 rutas más representadas se clasificaron en 7 clases de rutas e incluyeron 2.687 (72,12%) de las 3.725 enzimas (Figura 5.6). Estas enzimas pertenecen en su mayoría a rutas implicadas en el metabolismo de los lípidos (1.180), los carbohidratos (416), los aminoácidos (382) y los nucleótidos (335).

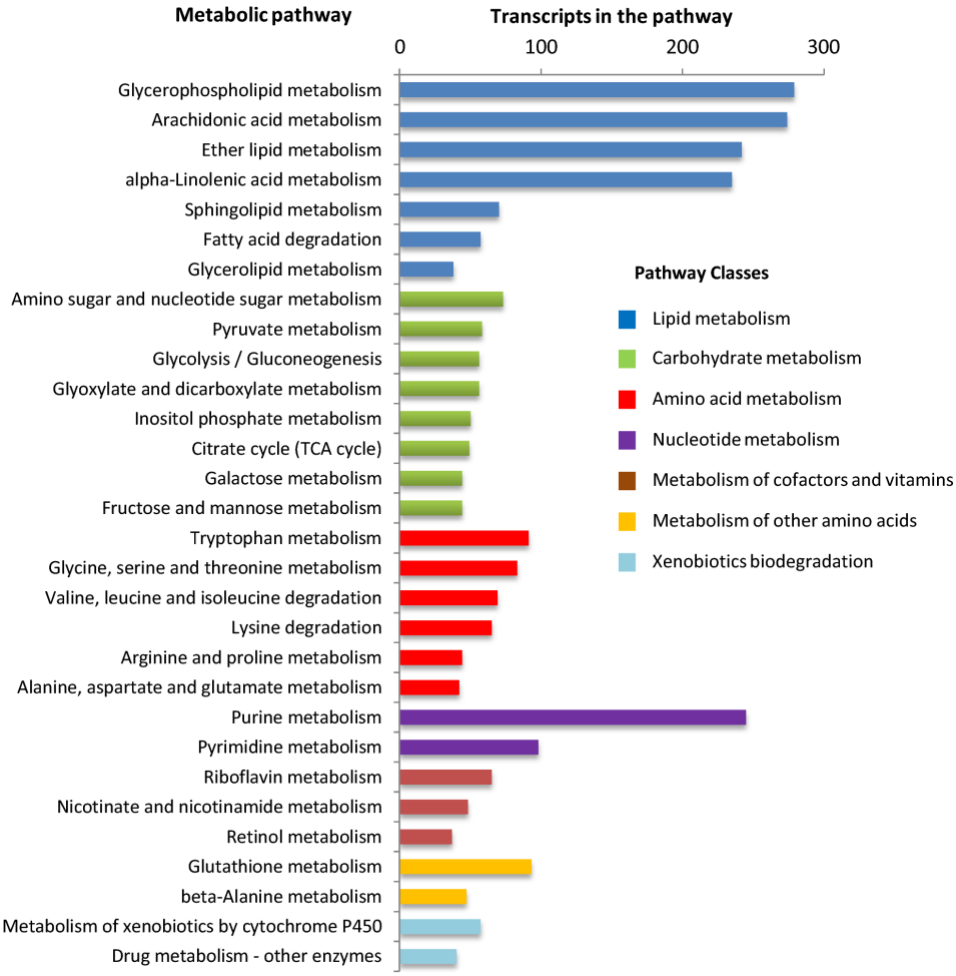

Figura 5.6. Gráfica representando las 30 rutas metabólicas KEGG más abundantes identificadas en el transcriptoma de *O. erraticus*, incluyendo 2.687 transcritos agrupados en 7 clases de rutas.

Tras la caracterización del transcriptoma consenso, se realizaron los análisis de expresión diferencial comparando los niveles de expresión de los genes entre estos tres estados fisiológicos: 7 días tras alimentación vs no alimentación (7 vs 0), 14 días tras alimentación vs 7 días tras alimentación (14 vs 7) y 14 días tras alimentación vs sin alimentación (14 vs 0). Los resultados completos de estos análisis de expresión diferencial pueden verse en la Tabla suplementaria S5.6. La mayor expresión diferencial se observó a los 7 días después de la alimentación (7 vs 0) con 3.170 transcritos diferencialmente expresados, de los cuales 2.088 estaban regulados al alza (log2FC > 1 y FDR < 0,05) y 1.082 estaban regulados a la baja (log2FC < -1 y FDR < 0,05) (Figura 5.7A). Entre 7 y 14 días tras la alimentación (14 vs 7), solamente hubo ligeras variaciones, ya que solo se detectaron 122 transcritos expresados diferencialmente (Figura 5.7B). La comparación entre la condición basal y los 14 días después de la alimentación (14 vs 0) mostró 1.482 transcritos expresados diferencialmente, de los cuales 1.234 fueron regulados al alza.

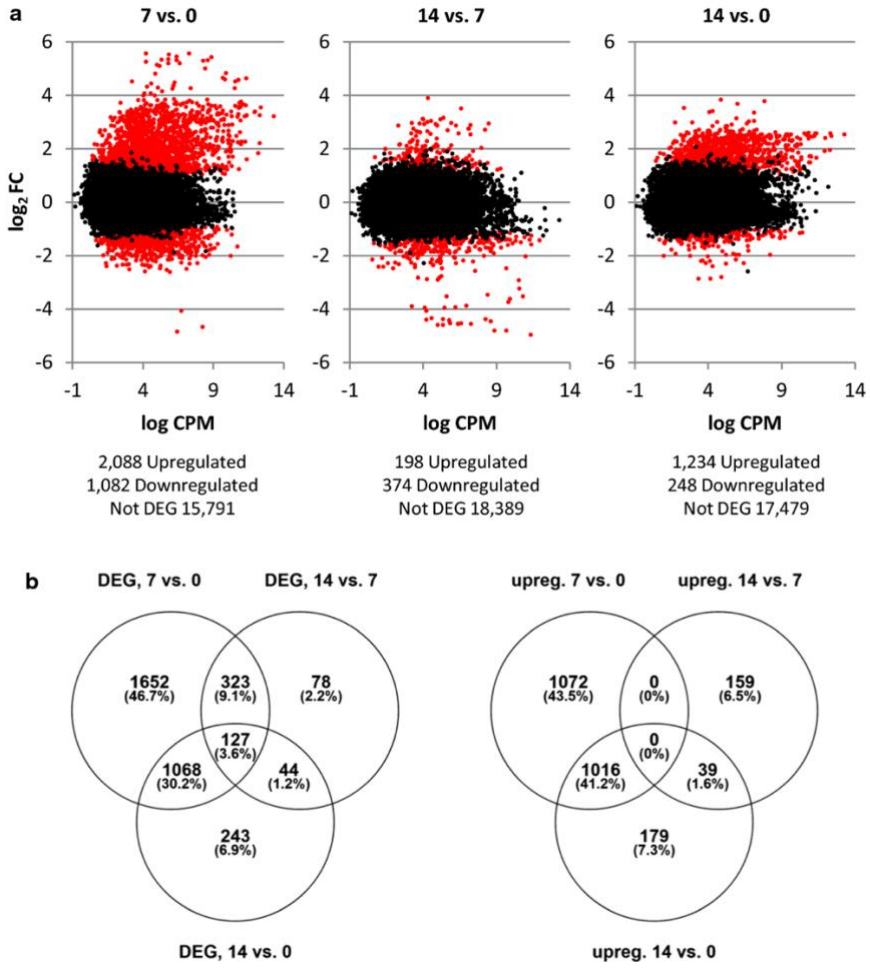

Figura 5.7. Patrones de expresión diferencial del transcriptoma de *O. erraticus*. A) Volcano *plots* representando el log2 *Fold Change* (log2FC) contra el logaritmo de las cuentas por millón para cada transcrito en cada una de las comparaciones. Transcritos con un FDR menor de 0,05 y un log2FC mayor o igual a 1 o menor o igual a -1 se consideraron como diferencialmente expresados y se representan en rojo. Transcritos fuera de estos valores no estuvieron diferencialmente expresados y se representan en negro. B) Diagramas de Venn mostrando el número de genes diferencialmente expresados (a la izquierda) y los diferencialmente regulados al alza (a la derecha) en cada comparación.

Cabe destacar que el 82,3% de los transcritos que estaban diferencialmente expresados a los 14 días después de la alimentación (1.016) ya estaban expresados diferencialmente a los 7 días después de la alimentación. Estos resultados indican que la mayor parte de la expresión génica diferencialmente regulada en las glándulas salivales se produjo en los primeros 7 días tras la alimentación, siendo menos importante a partir del día 7.

Los resultados de los análisis de enriquecimiento de GOs pueden verse en la Tabla suplementaria S5.7, que recopila los términos GO significativamente sobrerrepresentados asignados a los genes diferencialmente expresados. Además, la Figura 5.8 muestra los 20 principales procesos biológicos, funciones moleculares y componentes celulares significativamente sobrerrepresentados en las tres comparaciones.

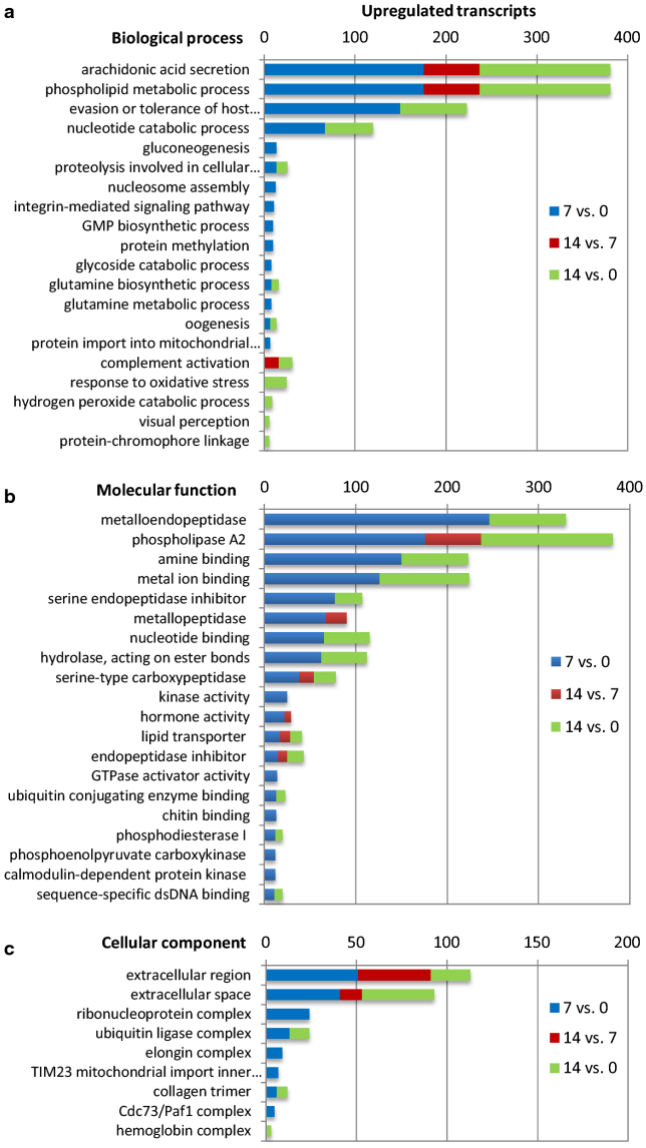

Figura 5.8. Enriquecimiento de términos GO en el transcriptoma de *O. erraticus*. 20 términos GO más significativamente sobrerrepresentados de procesos biológicos (a), función molecular (b) y componente celular (c) mostrando el número de genes diferencialmente expresados en las comparaciones 7 vs 0 (azul), 14 vs 7 (rojo) y 14 vs 0 (verde).

Hasta 98 GOs se encontraron como significativamente sobrerrepresentados (FDR < 0,05) en al menos una de las comparaciones: 36 procesos biológicos, 53 funciones moleculares y 9 componentes celulares. Para la categoría proceso biológico, el análisis de enriquecimiento mostró que 20, 11 y 20 términos GO estaban significativamente sobrerrepresentados en las comparaciones 7 vs 0, 14 vs 7 y 14 vs 0, respectivamente. La Figura 5.8a muestra que, entre los 20 procesos biológicos sobrerrepresentados, las categorías con un mayor número de secuencias reguladas al alza corresponden a proteínas involucradas en la secreción de ácido araquidónico, los procesos metabólicos de los fosfolípidos, la evasión o tolerancia de la respuesta de defensa del huésped y los procesos catabólicos de los nucleótidos. Para la categoría función molecular, el análisis de enriquecimiento mostró que 37, 10 y 34 términos GO estaban significativamente sobrerrepresentados en las comparaciones 7 vs 0, 14 vs 7 y 14 vs 0, respectivamente. La Figura 5.8b muestra que, a los 7 y 14 días tras la alimentación, las funciones moleculares asignadas al mayor número de secuencias fueron la actividad de la metaloendopeptidasa, la actividad de la fosfolipasa A2 (PLA2), la unión a aminas, la unión a iones metálicos y la actividad inhibidora de la endopeptidasa de tipo serina. Todas estas ontologías están relacionadas con grupos funcionales y familias de proteínas altamente reguladas a los 7 y 14 días de la alimentación. Entre ellas se encuentran proteínas con actividad PLA2, 5'-nucleotidasas/apiasas, lipocalinas, metalopeptidasas e inhibidores de proteasas. En cuanto a los componentes celulares, solamente 8, 2 y 5 GOs están significativamente sobrerrepresentados en las comparaciones 7 vs 0, 14 vs 7 y 14 vs 0, respectivamente. Los términos GO más sobrerrepresentados correspondieron a secuencias asignadas a compartimentos extracelulares (Figura 5.8c).

Por último, el análisis de enriquecimiento de rutas metabólicas en los genes diferencialmente expresados reveló que 9 rutas biológicas y 5 tipos de rutas se encontraban significativamente sobrerrepresentadas (FDR < 0,05) en al menos una de las tres comparaciones, y que seis, cuatro y ocho rutas biológicas se encontraban significativamente sobrerrepresentadas en las comparaciones 7 vs 0, 14 vs 7 y 14 vs 0, respectivamente. Véase la Tabla suplementaria S5.8 para más detalle sobre los análisis de enriquecimiento de rutas metabólicas. Este patrón de enriquecimiento va en paralelo con los patrones observados para el enriquecimiento de GOs y la regulación diferencial de genes, ya que la mayoría de las rutas están enriquecidas en las comparaciones 7 vs 0 y/o 14 vs 0, teniendo 733 y 625 secuencias, respectivamente. Tres rutas de los tipos metabolismo de carbohidratos y metabolismo de los aminoácidos aparecieron como enriquecidas solamente en la comparación 14 vs 0 (37 secuencias), mientras que la única ruta que apareció como enriquecida en la comparación 7 vs 0 fue del tipo biodegradación y metabolismo de xenobióticos (15 secuencias). En conjunto, las rutas sobrerrepresentadas están relacionadas con el metabolismo de lípidos, los aminoácidos, los carbohidratos, la energía y los xenobióticos, siendo las rutas del metabolismo de lípidos las que contienen el mayor número de secuencias sobreexpresadas (709 y 580 para las comparaciones 7 vs 0 y 14 vs 0, respectivamente (Tabla 5.4). Esta observación concordó con el elevado número y nivel de expresión de los transcritos sobreexpresados anotados como enzimas con actividad PLA2 (Figura 5.9), que participan en varias rutas metabólicas de los lípidos, como el metabolismo de los glicerofosfolípidos, el ácido araquidónico, los lípidos de éter y el ácido alfa-linolénico (Figura 5.6).

Tabla 5.4. Rutas metabólicas diferencialmente enriquecidas en *O. erraticus* en cada comparación.

Tipo de ruta	Identificador de la ruta	Ruta	Secuencias en ruta	Número de secuencias diferencialmente expresadas en las rutas		
				7 vs 0	14 vs 7	14 vs 0
Metabolismo de lípidos	map00590	Metabolismo del	274	181	61	148
	map00564	Metabolismo de los	279	176	61	144
	map00565	Metabolismo de los	242	176	61	144
	map00592	Metabolismo del	235	176	61	144
Metabolismo de carbohidratos	map00630	Metabolismo del glioxilato y del dicarboxilato	52	-	-	17
Metabolismo de aminoácidos	map00250	Metabolismo de la alanina, aspartato y glutamato	40	-	-	11
	map00350	Metabolismo de la tirosina	34	-	-	9
Metabolismo de la energía	map00910	Metabolismo del nitrógeno	17	9	-	8
Metabolismo y biodegradación de xenobióticos	map00983	Metabolismo de fármacos y otras enzimas	40	15	-	-

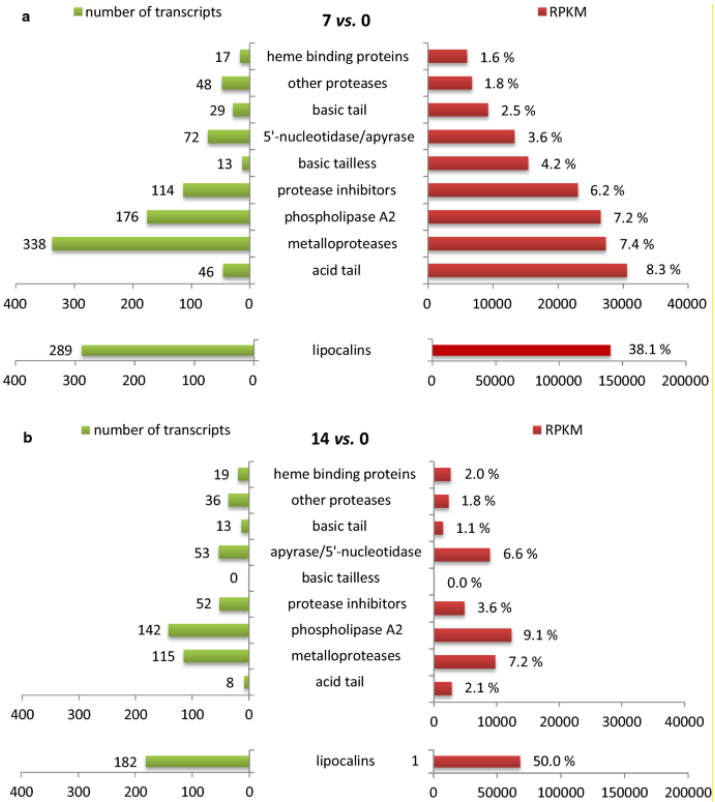

Figura 5.9. Número de transcritos anotados regulados al alza (barras verdes) y el nivel de expresión en RPKM (barras rojas) para cada proteína/familia en las comparaciones entre a) 7 días tras alimentación vs sin alimentación (7 vs 0) y b) 14 días tras alimentación vs sin alimentación (14 vs 0). Los porcentajes mostrados al final de cada barra roja representan la relación entre el nivel de expresión de cada grupo/familia y la expresión total en RPKM de todo el transcriptoma anotado como regulado al alza.

5.3.4. Resultados del protocolo RNA-seq de *novo* en muestras de *Ornithodoros moubata*

El objetivo de este estudio era ensamblar de *novo* y anotar el transcriptoma a partir de 6 muestras de la especie *Ornithodoros moubata* contribuidas por los doctores Ana Oleaga y Ricardo Pérez-Sánchez, para luego realizar el análisis de expresión diferencial entre tres condiciones fisiológicas diferentes: sin alimentación (0 días), 7 días tras alimentación (7 días) y 14 días tras alimentación (14 días). En primer lugar, se realizó un transcriptoma por muestra, obteniendo 70.133 y 84.920 *contigs* para las réplicas biológicas de la condición sin alimentación, 51.108 y 42.603 para la condición 7 días tras alimentación y 51.245 y 56.631 para la condición 14 días tras alimentación. Solamente las lecturas con más de 100 nucleótidos se utilizaron para realizar los ensamblajes. Véase la Tabla suplementaria S5.9 para más detalles sobre los ensamblajes obtenidos por muestra. A continuación, se realizó un transcriptoma consenso por estado fisiológico obteniendo 60.283, 41.911 y 39.563 transcritos para las condiciones sin alimentación, 7 días tras la alimentación y 14 días tras la alimentación, respectivamente. Para facilitar los siguientes análisis comparativos, se obtuvo un transcriptoma consenso a partir de la fusión de los 6 ensamblajes de *novo* realizados previamente (uno por muestra). El transcriptoma consenso estuvo formado por 80.684 transcritos con las siguientes métricas: un valor de N50 de 2.041 nucleótidos, un tamaño medio de transcrito de 1.397 nucleótidos y un valor de transcrito más largo de 17.994 nucleótidos. Los 80.684 transcritos se filtraron por redundancia resultando en 76.194 clústeres, de los cuales 54.845 tuvieron una ORF predicha con un tamaño mayor de 240 nucleótidos y fueron seleccionados para su anotación funcional, caracterización, expresión diferencial y análisis de enriquecimiento. Véase la Tabla 5.5 para más información sobre las métricas básicas de los transcriptomas por condición y del transcriptoma consenso.

Las búsquedas BLAST de los 54.845 transcritos contra las bases de datos NR del NCBI, Uniprot y el genoma de *Ixodes scapularis* permitieron la anotación de 41.011 (74,78%) secuencias utilizando un e-valor menor a 10^{-5}, de las cuales 30.171 (73,6%) mostraron una similaridad de secuencia por encima del 60%. Las 13.834 secuencias restantes (25,22%) no mostraron una homología significativa a ninguna secuencia presente en estas bases de datos. Como en el caso de *Ornithodoros erraticus*, este conjunto de secuencias podría representar proteínas todavía desconocidas, pero también secuencias potencialmente mal ensambladas sin significancia biológica o RNAs no codificantes largos. Para las 41.011 ORFs anotadas, se encontraron un total de 16.760 *accessions* no redundantes. Para más información sobre la anotación de estas secuencias, véase la Tabla suplementaria S5.10.

Los transcritos anotados fueron funcionalmente caracterizados utilizando las bases de datos Gene Ontology (GO) y KEGG mediante su asociación a las anotaciones Uniprot. En primer lugar, se clasificaron las secuencias respecto a su función molecular y proceso biológico en la base de datos GO. A casi la mitad de los transcritos (18.096 de 41.011) se les asignaron términos GO que incluían 24.600 procesos biológicos y 25.926 funciones moleculares. La Figura 5.10 representa los genes clasificados en relación con su función molecular y proceso biológico utilizando términos GO de nivel 2. En las gráficas de la Figura 5.10 solamente se incluyen categorías soportados por más de 400 secuencias.

Tabla 5.5. Métricas del ensamblaje del transcriptoma de la glándula salival de *O. moubata*.

Resumen	0 días	7 días	14 días	Consenso
Tamaño del transcriptoma	84.620.350	55.354.425	56.561.536	112.734.597
Número de transcritos	60.283	41.911	39.563	80.684
Transcrito más largo (bp)	15.640	17.994	15.564	17.994
Transcrito más corto (bp)	101	101	100	101
% de transcritos > 1 kb	51,94	47,03	52,10	50,33
Tamaño medio de transcrito (bp)	1.403	1.320	1.429	1.397
N50 (bp)	1.965	1.985	2.016	2.041
L50	13.225	8.442	8.536	16.802
%A	26,2	25,95	25,89	26,24
%C	23,91	24,14	24,12	23,84
%G	23,39	23,57	23,65	23,39
%T	26,43	26,27	26,26	26,42
%N	0,07	0,07	0,09	0,12
Clústeres	__	__	__	76.194
Clústeres con ORFs ≥ 240 nt	__	__	__	54.845
Transcritos anotados	__	__	__	41.011
Número no redundantes de *accessions*	__	__	__	16.760

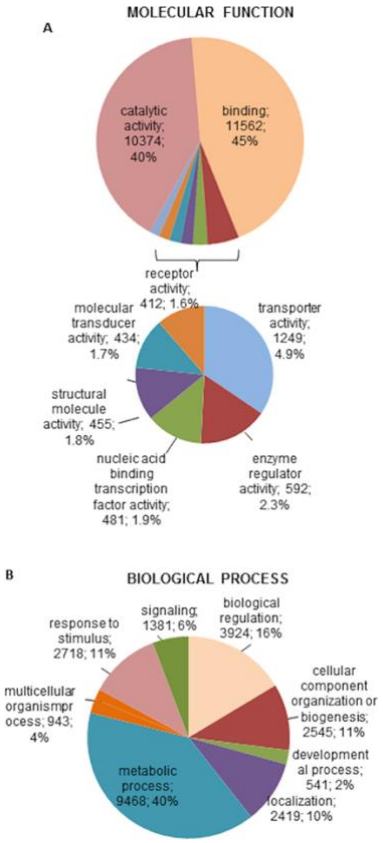

Figura 5.10. Clasificación de los transcritos anotados en relación con la (A) función molecular y (B) proceso biológico. Solamente se incluyen en los gráficos las categorías soportadas por más de 400 secuencias. Para cada categoría, se indica el número y porcentaje de secuencias.

Las categorías de funciones moleculares más abundantemente representadas fueron catalíticas (n = 10.374) y actividad de unión (n = 11.562), las cuales en conjunto se encuentran en el 85% de los transcritos. Las categorías significativamente menos representadas fueron actividad transportadora (n = 1.249), regulador de enzima (n = 592), factor de transcripción de unión a ácidos nucleicos (n = 481), estructural (n = 455), transductor molecular (n = 434) y actividad receptora (n = 412) (Figura 5.10A). La clasificación por procesos biológicos dio lugar a ocho categorías: la más abundantemente representada fue procesos metabólicos (n = 9.468) seguida de regulación biológica (n = 3.924), respuesta a estímulo (n = 2.718), organización o biogénesis de componente celular (n = 2.545), localización (n = 2.419), señalización (n = 1.381), procesos de organismos multicelulares (n = 943) y procesos de desarrollo (n = 541) (Figura 5.10B).

Para identificar las rutas metabólicas activas en las glándulas salivales de *O. moubata*, las 41.011 secuencias anotadas se analizaron en la base de datos KEGG. Hasta 5.977 secuencias se incluyeron en 103 rutas metabólicas agrupadas en 13 tipos generales (Tabla suplementaria S5.11). En la Figura 5.11 se muestran las 30 rutas más representadas, las cuales se agrupan en 9 tipos e incluyen 4.285 secuencias enzimáticas. Estas enzimas están mayoritariamente involucradas en rutas metabólicas de lípidos (1,252), nucleótidos (863), carbohidratos (739), aminoácidos (669) y biodegradación de xenobióticos (280).

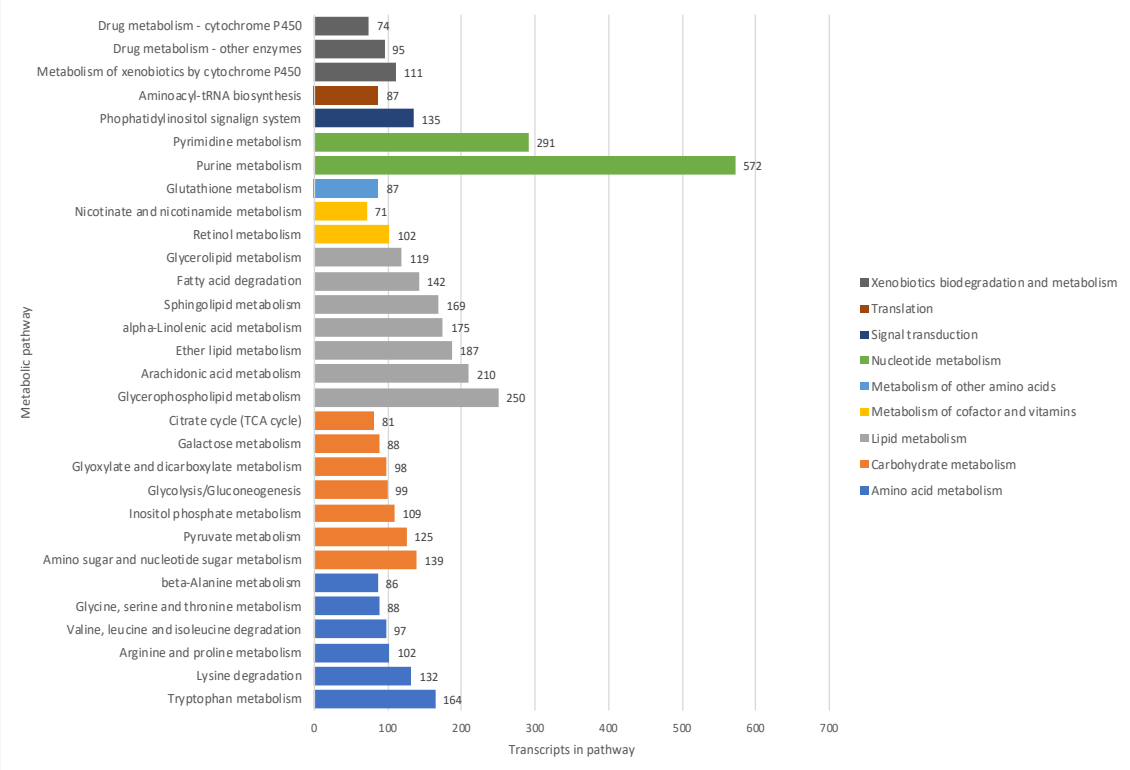

Figura 5.11. Gráfica representando las 30 rutas metabólicas KEGG más abundantes identificadas en el transcriptoma de *O. moubata*, incluyendo 4.285 transcritos agrupados en 9 clases de rutas.

Después de la caracterización funcional del transcriptoma, se caracterizó e identificó los genes diferencialmente expresados en *O. moubata* en tres condiciones: sin alimentación (SG0 o condición basal), 7 días tras la alimentación (SG7) y 14 días tras la alimentación (SG14). A continuación, se realizaron tres análisis de expresión diferencial para comparar los niveles de expresión entre estos tres estados fisiológicos: SG7 vs SG0, SG14 vs SG7 y SG14 vs SG0. Los resultados completos de los análisis de expresión diferencial pueden verse en la Tabla suplementaria S5.12.

A los 7 días después de la alimentación, la maquinaria de transcripción de genes parece totalmente activa, ya que, en este punto de tiempo, se observa la mayor expresión diferencial con respecto a la condición basal (SG7 vs SG0). En total, se identificaron 8.579 transcritos diferencialmente expresados, de los cuales 6.287 estaban regulados al alza (log2FC > 1 y FDR < 0,05). Entre los 7 y 14 días tras la alimentación (SG14 vs SG7), solamente hubo ligeros cambios, ya que solo se detectaron 203 transcritos expresados de forma diferencial. En consecuencia, la expresión génica diferencial entre la condición basal y 14 días después de la alimentación (SG14 vs SG0) refleja una situación similar a la observada entre la condición basal y 7 días después de la alimentación. A los 14 días después de la alimentación, se detectaron hasta 5.606 transcritos diferencialmente expresados, de los cuales 4.906 estaban regulados al alza (Tabla suplementaria S5.12, Figura 5.12).

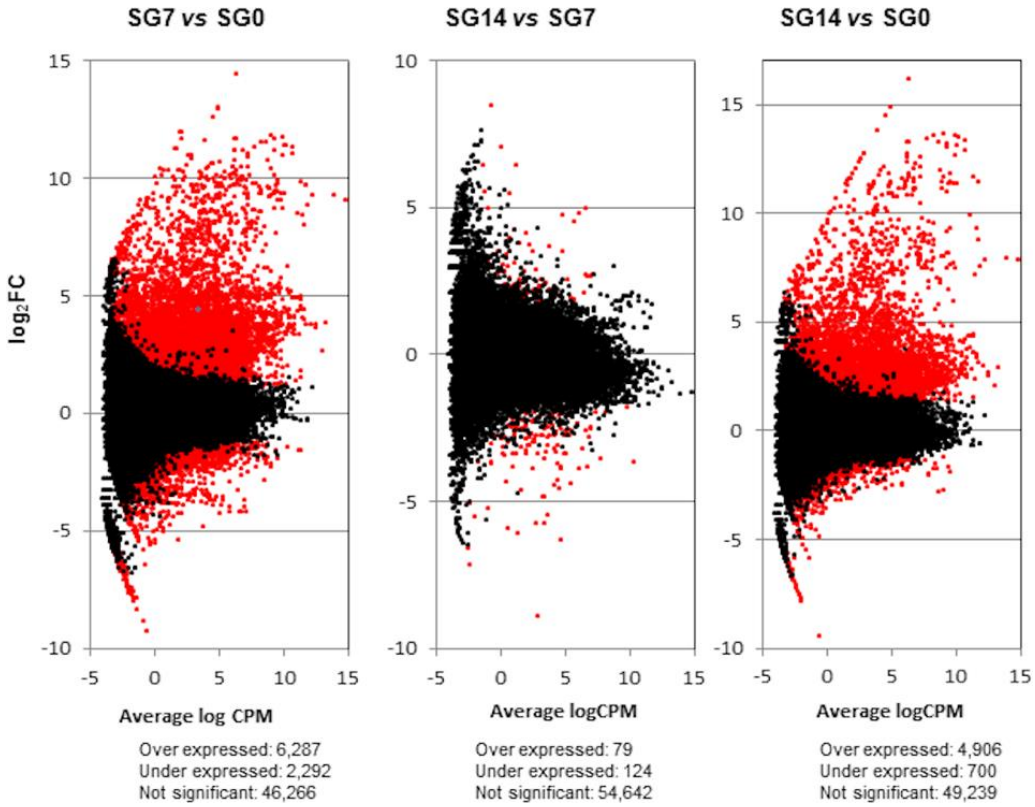

Figura 5.12. Volcano *plots* (uno por cada análisis de expresión diferencial) que representan el log *Fold Change* (log2FC) frente a la media del logaritmo de las cuentas por millón (CPM) por cada transcrito en cada par de muestras comparadas. Los transcritos diferencialmente expresados con FDR < 0,05 y log2FC ≥ 1 o ≤ -1 se representan en rojo. SG0 indica sin alimentación; SG7 y SG14 indican 7 y 14 días tras la alimentación, respectivamente.

El diagrama de Venn mostrado en la Figura 5.13 representa el número de transcritos diferencialmente expresados en cada comparación. El 86% de los transcritos que se encontraron como diferencialmente expresados a los 14 días tras la alimentación (SG14 vs SG0), se habían encontrado también diferencialmente expresados a los 7 días tras la alimentación (SG7 vs SG0). Estos resultados indican que la mayor expresión se produce en los primeros 7 días después de la alimentación.

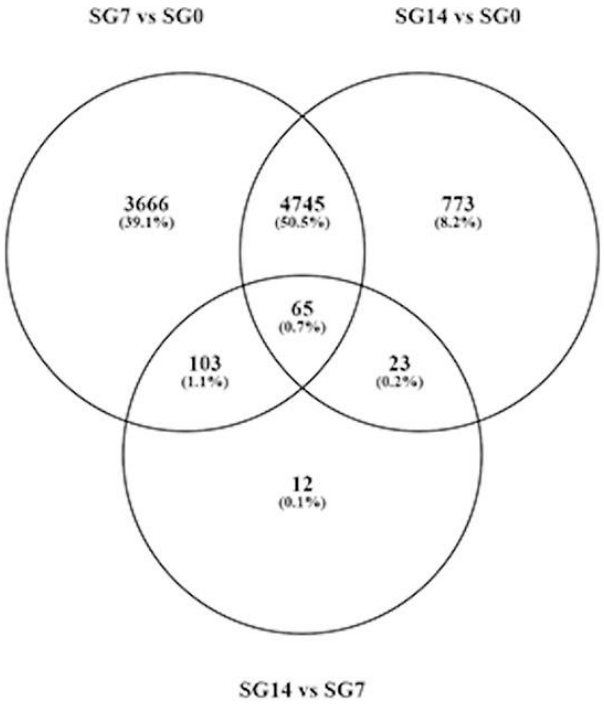

Figura 5.13. Diagrama de Venn mostrando el número de genes diferencialmente expresados en cada comparación realizada, siendo SG0 sin alimentación, SG7 7 días tras la alimentación y SG14 14 días tras la alimentación.

Una vez identificados los transcritos diferencialmente expresados, se procedió a asignarles términos GO significativamente enriquecidos, los cuales se encuentran recopilados en la Tabla suplementaria S5.13. Adicionalmente, las Figuras 5.14 y 5.15 muestran los 20 GOs más significativamente sobrerrepresentados de procesos biológicos y funciones moleculares, respectivamente.

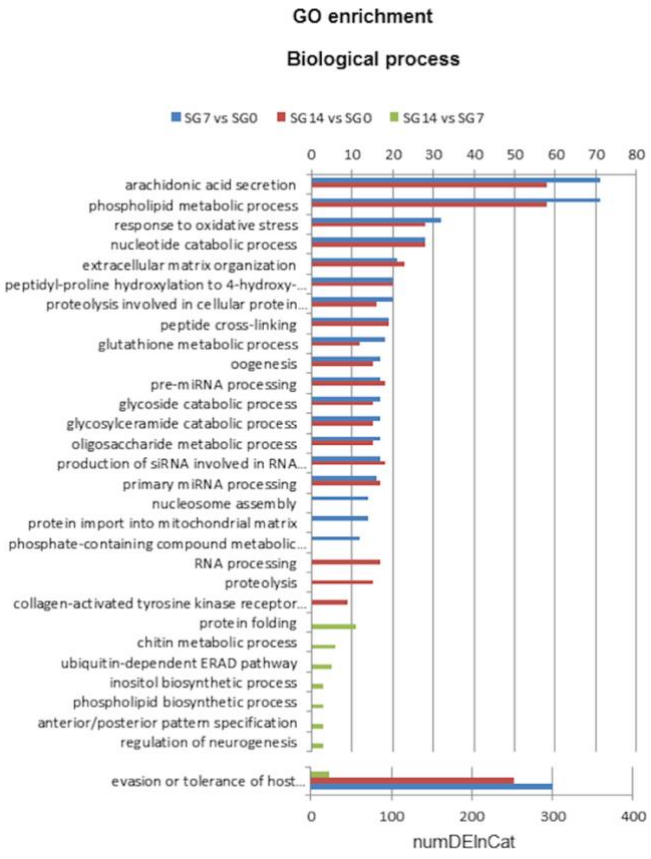

Figura 5.14. 20 términos GO de la categoría proceso biológico más diferencialmente enriquecidos en cada comparación.

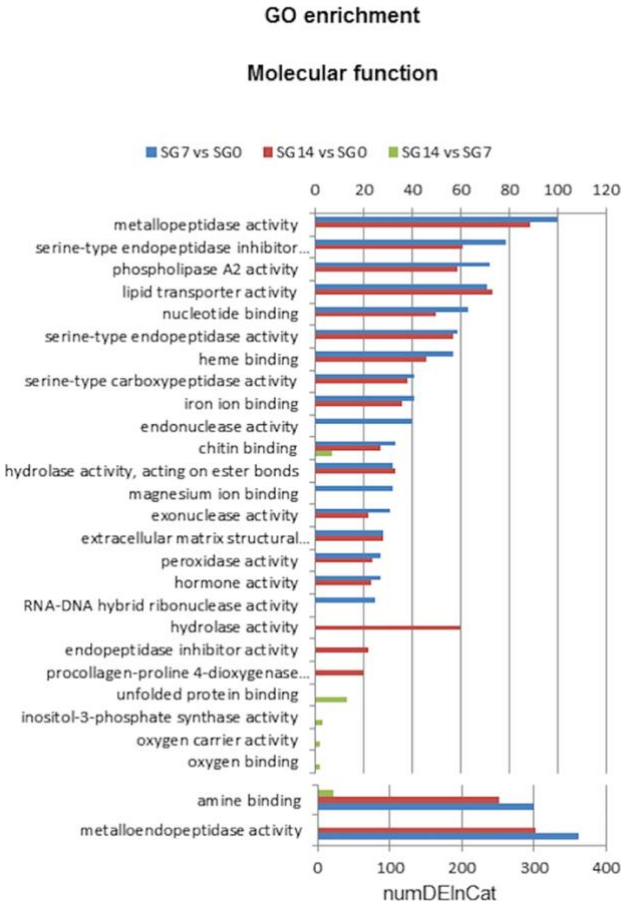

Figura 5.15. 20 términos GO de la categoría función molecular más diferencialmente enriquecidos en cada comparación.

Los análisis de enriquecimiento de GOs relevaron 144 GOs diferencialmente enriquecidos (FDR < 0,05) en al menos una de las tres comparaciones realizadas, siendo 53 correspondientes a proceso biológico, 78 a función molecular y 13 a componentes celulares. En el caso de la categoría proceso biológico, se encontró que 39, 29 y 8 términos GO estaban significativamente enriquecidos en las comparaciones SG7 vs SG0, SG14 vs SG7 y SG14 vs SG0, respectivamente (Tabla suplementaria S5.13). La Figura 5.14 muestra los 20 términos GO de la categoría proceso biológico más sobrerrepresentados en cada comparación. Las categorías con el mayor número de secuencias reguladas al alza corresponden a proteínas implicadas en la evasión o tolerancia de la respuesta de defensa del huésped, la secreción de ácido araquidónico, los procesos metabólicos de los fosfolípidos, la respuesta al estrés oxidativo y los procesos catabólicos de los nucleótidos. Estas categorías reflejan varias familias de proteínas cuyos genes están abundantemente regulados a los 7 y 14 días después de la alimentación como la familia de la lipocalinas, las proteínas con actividad fosfolipasa, las apilasas y las proteínas asociadas a las respuestas al estrés.

Para los términos GO de función molecular, el análisis de enriquecimiento reveló que 63, 54 y 6 términos GO estaban significativamente enriquecidos en las comparaciones SG7 vs SG0, SG14 vs SG0 y SG14 vs SG7, respectivamente (Tabla suplementaria S5.13). La Figura 5.15 muestra que, a los 7 y 14 días después de la alimentación, las funciones moleculares asignadas a un mayor número de secuencias fueron la actividad metalopéptida y la unión a aminas, seguidas de varias actividades peptidasas, el inhibidor de peptidasas, la actividad fosfolipasa A2 y el transportador de lípidos.

Los términos GO de compartimentos celulares más enriquecidos en SG7 y SG14 son secuencias asignadas a compartimentos extracelulares (Tabla suplementaria S5.13).

Por último, se obtuvo las rutas metabólicas significativamente enriquecidas asociadas a los transcritos diferencialmente expresados, las cuales se muestra en la Tabla suplementaria S5.14 y la Tabla 5.5. Estos análisis revelaron que 17 rutas metabólicas, agrupadas en 8 tipos, estaban diferencialmente enriquecidas (FDR < 0,05) en al menos una de las tres comparaciones, y que 13, 10 y 1 rutas metabólicas estaban significativamente enriquecidas en las comparaciones SG7 vs SG0, SG14 vs SG0 y SG14 vs SG7. Este patrón de enriquecimiento coincide con el patrón de enriquecimiento observado en los términos GO y la expresión diferencial de genes. Es decir, la mayoría de las rutas estuvieron enriquecidas en la comparación SG7 vs SG0, para la que acumulan el mayor número de secuencias (278). Además, la mayoría de estas rutas también estuvieron enriquecidas en la comparación SG14 vs SG0, aunque acumulando un menor número de secuencias (181). También se observó que tres rutas de los tipos biosíntesis y metabolismo de glicanos y degradación de xenobióticos estaban enriquecidas solamente en la comparación SG14 vs SG0 (21 secuencias) y que el metabolismo de fosfatos de inositol era la única ruta enriquecida en la comparación SG14 vs SG7 (3 secuencias).

En conjunto, las rutas enriquecidas implicaban el metabolismo de los aminoácidos, los carbohidratos, los lípidos, la energía, los glicanos, los cofactores y las vitaminas, y

xenobióticos, siendo el metabolismo de los lípidos y de los aminoácidos las rutas con mayor número de secuencias sobreexpresadas (Tabla 5.5).

Tabla 5.5. Rutas metabólicas diferencialmente enriquecidas en cada comparación.

Tipo de ruta	Identificador de la ruta	Ruta	Secuencias en ruta	Número de secuencias diferencialmente expresadas		
				SG7 vs SG0	SG14 vs SG7	SG14 vs SG0
Metabolismo de aminoácidos	map00330	Metabolismo de arginina y prolina	83	40	-	32
	map00300	Biosíntesis de lisina	12	7	-	-
	map00940	Biosíntesis de fenilpropanoides	28	21	-	18
	map00350	Metabolismo de tirosina	22	10	-	12
Metabolismo de carbohidratos	map00520	Metabolismo del amino azúcar y	82	22	-	-
	map00562	Metabolismo del inositol fosfato	8	-	3	-
	map00040	Interconversiones de pentosa y	6	4	-	-
Metabolismo de la energía	map00910	Metabolismo del nitrógeno	11	6	-	-
	map00190	Fosforilación oxidativa	43	20	-	-
Biosíntesis y metabolismo del glicano	map00603	Biosíntesis de glicoesfingolípidos	7	-	-	4
	map00512	Biosíntesis de mucina tipo O-	10	-	-	5
Metabolismo de lípidos	map00565	Metabolismo de lípidos del éter	180	74	-	59
	map00061	Biosíntesis de ácidos grasos	34	22	-	23
	map00140	Biosíntesis de hormonas	22	14	-	8
Metabolismo de cofactores y vitaminas	map00670	*One carbon pool by folate*	9	6	-	-
Metabolismo de otros aminoácidos	map00480	Metabolismo del glutatión	81	32	-	17
Metabolismo y biodegradación de xenobióticos	map00980	Metabolismo de xenobióticos por el citocromo P450	41	-	-	12

5.3.5. Resultados del protocolo de RNA-seq de resecuenciación en muestras humanas

En el caso de las muestras de humano contribuidas por el profesor José Vicente Bagán, tras el análisis de calidad y el preprocesado de las mismas, se mapearon directamente contra el genoma de referencia. Los ficheros BAM resultantes se utilizaron para realizar el análisis de expresión diferencial entre el grupo PVL y el grupo control. Los resultados completos de este análisis pueden verse en la Tabla suplementaria S5.15. En total, se identificaron 140 genes diferencialmente expresados con un valor de FDR menor a 0,05. De estos 140 genes diferencialmente expresados (Figura 5.16A), 40 se encontraron regulados a la baja y 100 regulados al alza en el grupo PVL frente al grupo control. Como se muestra en la Figura 5.16B y C, la dispersión y la mediana de los valores de expresión fueron similares en ambos grupos, lo que apoyó la viabilidad de la comparación de ambos grupos en un análisis de expresión diferencial. La Figura 5.16D revela una razonable sobredispersión individual entre las muestras debido a su variabilidad intrínseca.

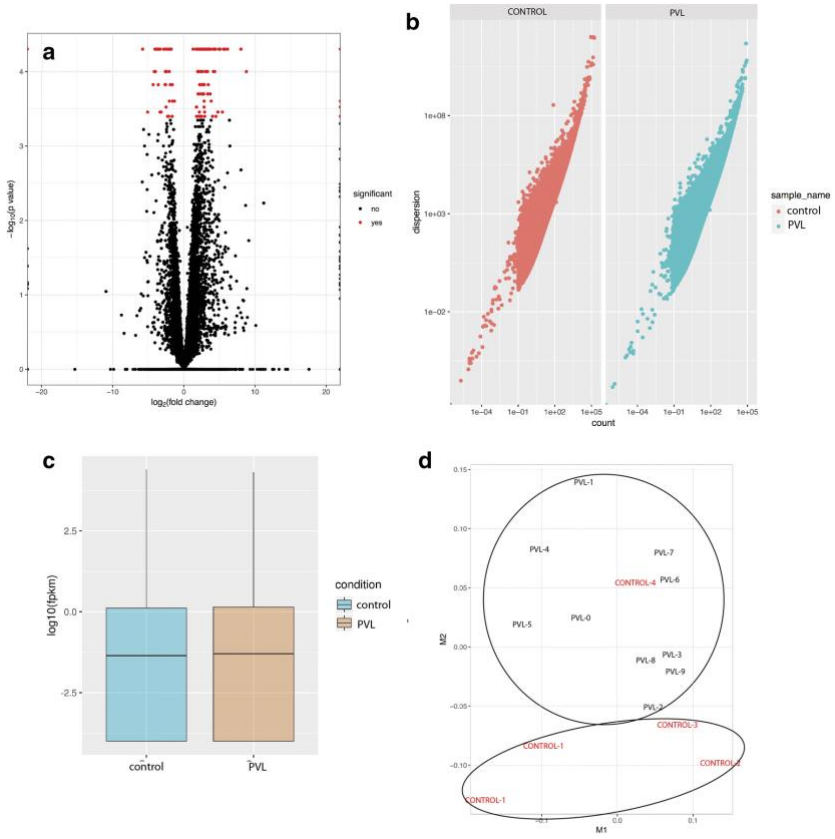

Figura 5.16. 140 genes se encontraron diferencialmente expresados entre los grupos PVL y control. A) Volcano *plot* representando el logaritmo negativo de los p-valores ajustados frente a los valores log2 *Fold Change* obtenidos en el análisis de expresión diferencial. Los puntos rojos resaltan los 140 genes diferencialmente expresados bajo un FDR < 0,05. B) Gráfico de dispersión que representa la desviación del umbral en las librerías de secuenciación utilizadas en cada grupo frente a los recuentos de fragmentos por kilobase de transcrito por millón de lecturas mapeadas (FPKM). C) Gráfico de caja de la expresión de los FPKM por grupo en escala logarítmica. D) Gráfico de escalamiento multidimensional (MDS) que muestra las relaciones de agrupamiento y las fuentes de variabilidad de las librerías del grupo control y PVL.

Obsérvese que las muestras mostraron una buena separación por la condición patológica en este análisis, revelando algunas diferencias de variación biológica, y, por lo tanto, algunas diferencias a nivel de expresión diferencial entre las muestras PVL y las muestras control, excepto el control-4 que mostró una mayor variación intrínseca que otras muestras control. El control-4 se trata de un individuo sano cuya desviación respecto a los otros controles se podría deber a muchas razones (genética, epigenética, ambiental). A pesar de ello, esta muestra se mantuvo en el análisis de expresión diferencial porque cualquier sesgo que pudiera aportar al estudio fue corregido por el resto de las muestras control.

Por último, se investigaron las diferencias funcionales en la expresión génica, determinadas por el enriquecimiento diferencial de las categorías de GO, realizando una comparación de GOseq a nivel de todo el transcriptoma. Este análisis identificó incrementos estadísticamente significativos en la expresión de 44 GOs (Tabla suplementaria S5.16). Seis de los 44 GOs enriquecidos corresponden a componentes celulares relacionados con diferentes sublocalizaciones como la membrana plasmática del espacio extracelular, el complejo del componente del complemento C1, y la sinapsis inmunológica. Otros 26 de los 44 GOs enriquecidos correspondían a procesos biológicos que estaban relacionados principalmente con procesos del sistema inmunitario, así como con la quimiotaxis, la liberación de iones de calcio secuestrados, la motilidad y la adhesión celular, la morfogénesis, la homeostasis de la formación y la diferenciación, y la regulación positiva de la transcripción. El resto del enriquecimiento de GOs observado en el grupo PVL pertenecía a 12 GOs de funciones moleculares, incluyendo genes que codifican para factores de transcripción y proteínas con funciones de unión, especialmente para la actina y la unión del citoesqueleto. En la Figura 5.17 se muestra una gráfica con el número de secuencias asignadas a cada GO significativamente enriquecido cuyo número de secuencias asignadas sea igual o mayor a 100.

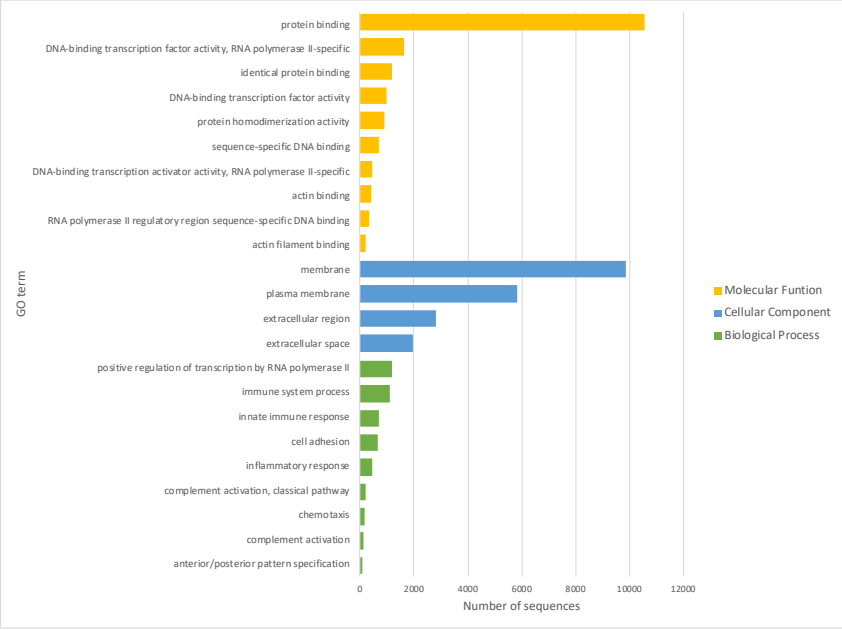

Figura 5.17. Términos GO diferencialmente enriquecidos en la comparación PVL vs control soportados por al menos 100 secuencias.

5.3.6. Implementación de los protocolos RNA-seq en la aplicación RNASeq del GPRO Suite

Los protocolos descritos en este capítulo se implementaron en la herramienta RNASeq del GPRO Suite (Hafez et al., 2022). Esta aplicación está dedicada al manejo, gestión y ejecución de estos protocolos y flujos de trabajo para el análisis de expresión diferencial y de enriquecimiento a partir de datos de secuenciación. Los flujos de trabajo implementados en esta herramienta son los mostrados anteriormente en la Figura 5.1.

Para facilitar a todo tipo de usuario el uso de todas estas herramientas que normalmente se ejecutan mediante línea de comandos (requiriendo conocimientos más avanzados en informática y sobre el entorno Linux), se creó una interfaz gráfica para cada una de las herramientas que componen los protocolos, de manera que la gestión de las mismas se realiza a través de servidores web. Además, se creó un gestor *pipeline* donde el usuario escoge el protocolo de interés y los parámetros más convenientes para sus muestras, de manera que cada una de las herramientas del protocolo de se ejecutan de forma automática. Esta interfaz gráfica es la que se puede observar en la Figura 5.18.

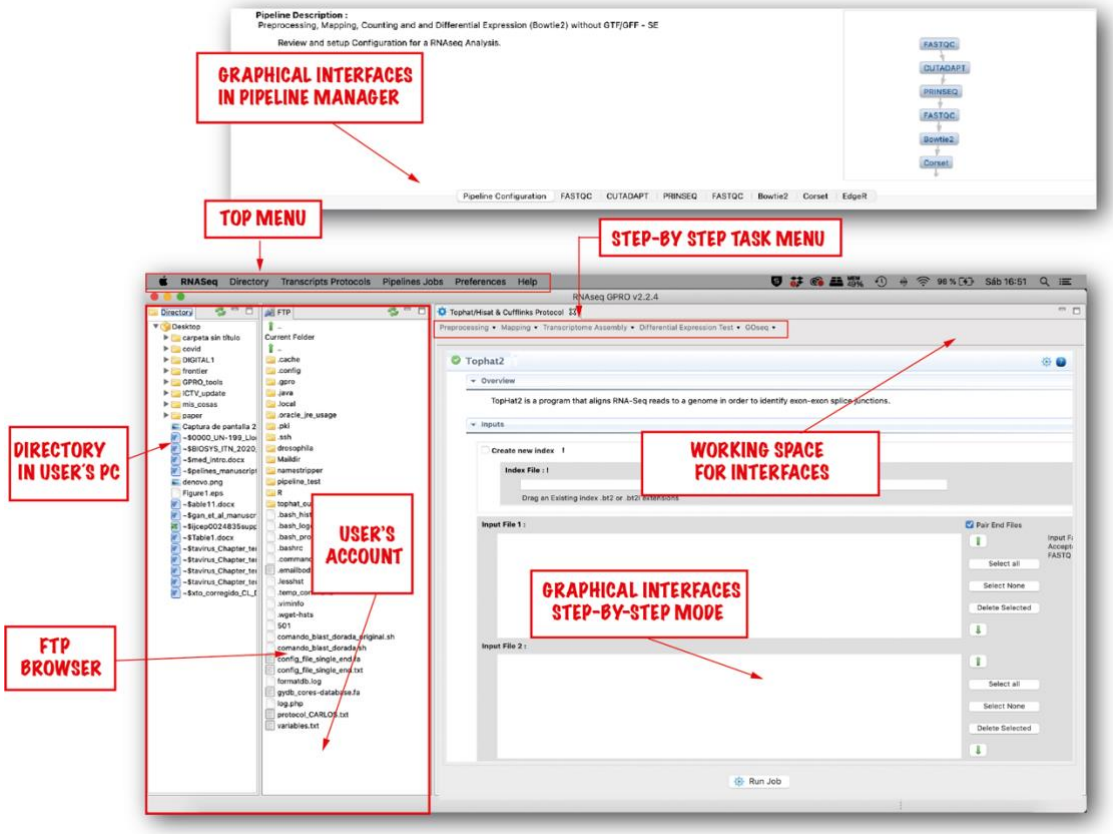

Figura 5.18. Interfaz de usuario de RNASeq. Una vez que se ha vinculado la aplicación a un servidor, el usuario debe seguir los siguientes pasos. (1) Transferir los archivos *input* desde el directorio del PC del usuario al servidor utilizando el navegador FTP (del inglés, *File Transfer Protocol*). (2) Seleccionar el modo de uso, el cual puede ser paso a paso o *pipeline*. (3) Arrastrar los ficheros *input* desde el servidor a los campos de entrada de la interfaz seleccionada. (4) Escoger una carpeta *output*. (5) Establecer las opciones y los parámetros de interés. (6) Ejecutar el análisis.

Por último, la herramienta Oases, utilizada para el ensamblaje de *novo* de los transcriptomas de anisakis y ambas especies de garrapata, se implementó en la herramienta DeNovoSeq del GPRO Suite para facilitar su uso a todo tipo de usuarios. La interfaz creada para esta herramienta es la que se puede ver en la Figura 5.19.

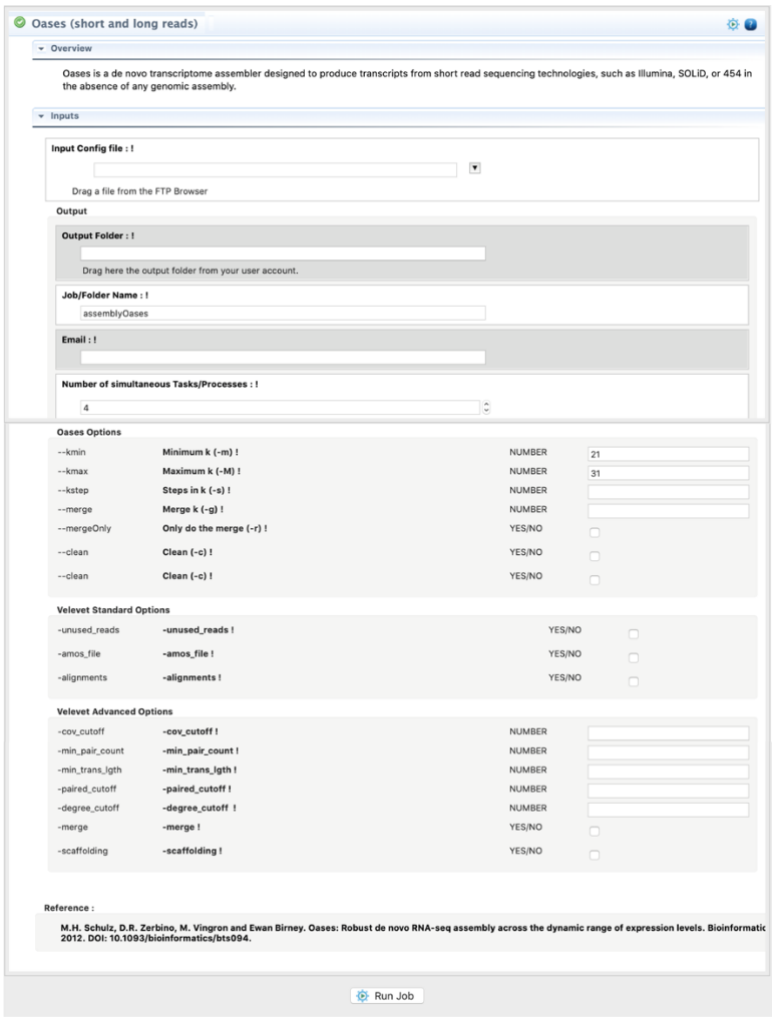

Figura 5.19. Interfaz creada para la ejecución de la herramienta Oases en la aplicación DeNovoSeq del GPRO Suite.

5.4. Discusión

Con la bajada en los costes de secuenciación que se ha ido produciendo en los últimos años, el análisis RNA-seq ha pasado rápidamente a convertirse en una tecnología más asequible y ha ganado popularidad como herramienta en investigación, permitiendo estudiar los cambios en un transcriptoma en respuesta a cambios en el ambiente, condiciones experimentales o tratamientos, entre otros. Con la necesidad de estudiar los cambios en los transcriptomas sin referencia genómica disponible hemos usado los transcriptomas de *Ornithodoros moubata* y *Ornithodoros erraticus* en tres estados fisiológicos diferentes, así como dos especies de anisakis y sus híbridos. En paralelo hemos implementado un protocolo para estudiar transcriptomas con referencia genómica conocida donde aquí hemos utilizado como caso a estudio los patrones de expresión diferencial que se producen entre muestras humanas de pacientes sanos y de pacientes con leucoplasia verrucosa proliferativa. Con todo esto hemos implementado un protocolo de análisis RNA-seq con dos rutas estratégicas, una para analizar transcriptomas de *novo* y otro para analizar datos de resecuenciación con genoma conocido.

Ambas implementaciones nos han permitido establecer las diferencias en cuanto expresión diferencial de genes entre las distintas condiciones de estudio o entre diferentes especies, como es el caso de las muestras de anisakis, dando como resultado aquellos genes que se encontraban diferencialmente expresados en una especie o condición. Con ello, es posible dilucidar potenciales antígenos que puedan ser posibles dianas para posibles terapias o diseños vacunales como es el caso de ambas especies de garrapatas; dilucidar las posibles relaciones y diferencias entre distintas especies, como es el caso de las especies de anisakis y sus híbridos; o descubrir biomarcadores de cáncer, como es el caso del estudio con muestras humanas.

En último lugar, cabe decir que las herramientas implementadas en ambos protocolos se ejecutan normalmente mediante línea de comandos en un entorno Linux, dificultando que esta tecnología llegue a investigadores biológicos para que estos puedan trabajar con estos protocolos de manera autónoma. Con el fin de que cualquier investigador fuera capaz de realizar análisis RNA-seq, ambas estrategias se han implementado en la aplicación RNASeq del GPRO Suite, pudiéndose ejecutar tanto en modo *pipeline*, de manera automática, o en modo “paso a paso”, donde el investigador ejecuta cada herramienta de forma independiente una tras otra. Esta aplicación proporciona una interfaz amistosa para cada herramienta y a través de la cual se pueden ajustar los distintos protocolos a las muestras con las que se esté trabajando al poder personalizar los parámetros de cada una de las herramientas.

5.5. Publicaciones *peer-review* relacionadas con este capítulo en esta tesis

A. Hafez A, Soriano B, Elsayed AA, Futami R, Ceprián R, Ramos-Ruiz R, Martínez G, Roig FJ, Torres-Font MA, Naya-Català F, Calduch-Giner JA, Trilla-Fuertes L, Gámez-Pozo A, Arnau V, Sempere JM, Pérez-Sánchez J, Gabaldón T, Llorens C. 2022. Client applications and Server Side docker for management of RNASeq and/or VariantSeq workflows and pipelines of the GPRO Suite. Accepted in Genes journal. Preprint available in arXiv. https://doi.org/10.48550/arXiv.2202.07473

Contribución de la autora de esta tesis a este trabajo: Co-primera autora contribuyendo en el diseño e implementación del protocolo, programación de los *scripts* en R en Python de los *pipelines* y flujos de trabajo de la herramienta, validación y testado de funciones, co-redacción del artículo.

B. Pérez-Sánchez R, Carnero-Morán Á, Soriano B, Llorens C, Oleaga A. 2021. RNA-seq analysis and gene expression dynamics in the salivary glands of the argasid tick Ornithodoros erraticus along the trophogonic cycle. Parasites Vectors 14, 170. https://doi.org/10.1186/s13071-021-04671-z

Contribución de la autora de esta tesis a este trabajo: Diseño y puesta a punto del protocolo, ejecución de todos los análisis bioinformáticos y co-redacción del artículo.

C. Oleaga A, Soriano B, Llorens C, Pérez-Sánchez R. 2021. Sialotranscriptomics of the argasid tick Ornithodoros moubata along the trophogonic cycle. PLOS Neglected Tropical Diseases 15(2): e0009105.https://doi.org/10.1371/journal.pntd.0009105.

Contribución de la autora de esta tesis a este trabajo: Diseño y puesta a punto del protocolo, ejecución de todos los análisis bioinformáticos y co-redacción del artículo.

D. Llorens C, Soriano B, Trilla-Fuertes L, Bagan L, Ramos-Ruis R, Gamez-Pozo A, Peña C, Bagan JV. 2021. Immune expression profile identification in a group of proliferative verrucous leukoplakia patients: a pre-cancer niche for oral squamous cell carcinoma development. Clin Oral Invest.https://doi.org/10.1007/s00784-020-03575-z

Contribución de la autora de esta tesis a este trabajo: Diseño y puesta a punto del protocolo, ejecución de todos los análisis bioinformáticos y co-redacción del artículo.

E. Llorens C, Arcos SC, Robertson L, Ramos R, Futami R, Soriano B, Ciordia S, Careche M, González-Muñoz M, Jiménez-Ruiz Y, Carballeda-Sangiao N, Moneo I, Albar JP, Blaxter M and Navas A. 2018.Functional insights into the infective larval stage of Anisakis simplex s.s., Anisakis pegreffii and their hybrids based on gene expression patterns. BMC Genomics, 19:592. https://doi.org/10.1186/s12864-018-4970-9

Contribución de la autora de esta tesis a este trabajo: ejecución de las anotaciones de los transcriptomas consenso.

6. SAMBA, UNA APLICACIÓN BASADA EN REDES BAYESIANAS PARA LA PREDICCIÓN DE CAMBIOS EN LA COMPOSICIÓN Y FUNCIÓN DE LA MICROBIOTA EN ACUICULTURA

6.1. Contexto

Las comunidades bacterianas intestinales de los peces están sometidas a grandes fluctuaciones que reflejan el complejo diálogo entre el hospedador y las comunidades microbianas residentes. Tales interacciones son impulsadas por un número de factores intrínsecos (por ejemplo, el genotipo, el estado fisiológico, la patobiología) y ambientales(por ejemplo, el medio ambiente, el estilo de vida y la dieta) que tienen la capacidad de modificar la composición y diversidad de la microbiota intestinal, así como su función y actividad metabólica (Egerton et al., 2018). De hecho, la investigación actual en doradas de piscifactoría (*Sparus aurata*) resaltó el uso de la microbiota intestinal como un criterio fiable para evaluar el éxito de la cría selectiva para producir peces más sanos y resistentes con cambios en la composición de la dieta y la disponibilidad limitada de alimentos convencionales (Piazzon et al., 2020; Naya-Català et al., 2022a; Naya-Català et al., 2022b). Sin embargo, nuestra comprensión sobre la dinámica de las interrelaciones entre los microbiomas de peces y sus hospedadores todavía se encuentra en su infancia debido a los múltiples factores bióticos y abióticos que intervienen en este tipo de dinámica (Faust, 2021; Liu et al., 2021). Ciertamente, actualmente hay un reconocido interés por las herramientas de biología de sistemas como las redes bayesianas (BN) y los sistemas de aprendizaje de estructuras (Scutari, 2009; Michiels et al., 2021) que pueden utilizarse de forma convincente para modelizar sistemas biológicos complejos. En el campo de la microbiología, las redes bayesianas han demostrado ser una herramienta poderosa para inferir relaciones direccionales dentro de las comunidades microbianas (Sazal et al., 2020; Sazal et al., 2021), y para analizar redes funcionales en datos metagenómicos (Hobbs et al., 2016). El estado del arte actual en redes bayesianas incluye varias herramientas con visualizaciones interactivas, como shinyBN (Chen et al., 2019a) o BayesianLab (Conrady and Jouffe, 2015), que solamente son capaces de trabajar con variables discretas y conjuntos de datos pequeños. La reciente herramienta llamada BayesSuites (Michiels et al., 2021) supera estas dificultades, siendo capaz de gestionar variables continuas y grandes conjuntos de datos centrados en el campo de la neurociencia, aunque no es capaz de realizar inferencia sobre variables discretas para establecer distribuciones de probabilidad condicional. En la investigación sobre acuicultura, las redes bayesianas se han utilizado para integrar mejor el conocimiento sobre la gestión sostenible basada en los ecosistemas (Yuniarti et al., 2021), pero la estructuras de las interrelaciones y

dependencias entre los múltiples factores bióticos y abióticos distintos que intervienen en la dinámica de un determinado sistema acuícola sigue siendo en su mayor parte desconocida. El objetivo de este capítulo es presentar SAMBA (del inglés, *Structure-Learning of Aquaculture Microbiomes using a Bayesian-Network Approach*), la implementación software de un modelo bayesiano que permite investigar y/o predecir cómo los pan-microbiomas de los peces y otras variables, implicadas en la dinámica de un sistema acuícola dado, están relacionados y se influyen mutuamente.

6.2. Material y métodos

6.2.1. Implementación y disponibilidad

SAMBA es una solución web para el análisis de redes. El componente *backend* es un flujo de trabajo implementado en R y Python que utiliza dependencias de software de terceros, mientras que el componente *frontend* es el entorno de interfaces gráficas de usuario implementado para gestionar SAMBA. El ejecutable, código fuente de SAMBA y los conjuntos de datos para poder testar la aplicación están disponibles en el repositorio llamado SAMBA de la cuenta de Github de Biotechvana (https://github.com/biotechvana/SAMBA/).

6.2.2. Base algorítmica, diseño de SAMBA y su implementación

SAMBA se ha implementado como una aplicación web interactiva utilizando R, Python y el paquete de R llamado shiny (Chang et al., 2021), y está estructurada en cinco módulos:

1) Input Network
2) Network Reports
3) Prediction
4) Network Viewer
5) Results

Toda la implementación integra funciones adicionales del paquete Future (Bengtsson, 2021) para ejecutar los pasos de aprendizaje, entrenamiento y predicción en segundo plano, lo que permite que otras tareas de SAMBA funcionen mientras se ejecutan dichos pasos.

6.2.2.1. Input Network

Este primer módulo crea el modelo de red bayesiana (modelo BN) a partir de los datos disponibles utilizando la opción "*Learning and training*". El modelo es una BN híbrida creada usando el paquete bnlearn de R (Scutari, 2010) e implementado para inferir las dependencias condicionales entre taxones y variables experimentales, elucidando relaciones direccionales entre taxón-taxón, variable experimental-variable experimental, y variable experimental-taxón. Como *input*, este componente acepta archivos planos que contiene datos de variables experimentales y cuentas crudas de taxones a partir de datos 16S. Se puede aplicar un filtro opcional para eliminar taxones que tengan un número bajo de cuentas crudas en un porcentaje específico de muestras considerando todas las muestras (opción "Total") o las muestras de un grupo (opción "Group"), especificando, en este caso, una variable experimental de interés. Este filtro, denominado filtro de prevalencia, puede aplicarse antes o después de calcular el modelo BN completo. Además, se puede utilizar una *blacklist* y/o *whitelist* como *input*s opcionales para forzar a la red a considerar relaciones específicas (en el caso de la *whitelist*) o a descartarlas (en el caso de la *blacklist*) de acuerdo con un conocimiento previo sobre cómo debería ser la estructura de la red.

El modelo se construye siguiendo cuatro pasos:

I) Normalización: la BN acepta el *input* de datos discretos y continuos para las variables experimentales, aunque permite la discretización de variables continuas bajo tres métodos: "quantile", "interval" o "Hartemink" (Hartemink, 2001). El último método ("Hartemink") intenta maximizar la información mutua entre variables. Si no se discretizan las variables continuas, se realiza un test de Shapiro (Shapiro and Wilk, 1965) para saber si estas variables siguen una distribución normal y, si no es el caso, se realizar automáticamente una transformación logarítmica para estas variables. Las cuentas crudas de un taxón en una muestra se normalizan aplicando la siguiente fórmula:

$$NC_{ij} = \frac{X_{ij} * \sum_{j=1}^{n} X_{ij}}{\left(\sum_{i=1}^{t} X_{ij}\right)}$$

Donde NCij es el recuento normalizado de un taxón (i) en una muestra concreta (j), Xij son las cuentas crudas de un taxón concreto (i) en una muestra concreta (j), n es el número de muestras del conjunto de datos y t es el número de taxones del conjunto de datos.

II) Filtrado: paso opcional consistente en eliminar aquellos taxones cuyas cuentas crudas están por debajo de un límite en un porcentaje de muestras (filtro de prevalencia). Este filtro tiene dos opciones: A) "Total", donde se consideran todas las muestras para eliminar los taxones; y B) "Group", donde se agrupan las muestras por diferentes estados de una variable experimental indicada por el usuario. Para cada estado, debe introducirse un porcentaje y, a continuación, se elimina un taxón si no cumple todas las condiciones

definidas. Este filtro puede aplicarse antes o después de construir el modelo BN.

III) Construcción y aprendizaje de la estructura de la red: para la distribución logarítmica normal, el recuento normalizado (NC) se transforma logarítmicamente mediante la siguiente fórmula:

$$(\ln (NC + 1))$$

A continuación, los datos normalizados con transformación logarítmica se fusionan con los datos de las variables experimentales en un único objeto de R. Por el contrario, para la distribución binomial negativa cero-inflada (ZINB), los recuentos normalizados se combinan con los datos de las variables experimentales sin transformación logarítmica. El conjunto de datos completo se utiliza para construir y entrenar el modelo utilizando las funciones hc() y bn.fit() del paquete bnlearn, respectivamente. La función hc() de bnlearn aprende la estructura de una red bayesiana mediante una *hill-climbing greedy search*. Nótese que en la literatura científica se han propuesto tres enfoques para el aprendizaje de estructuras: "*constraint-based*", "*score-based*" y "*hybrid*" (Scutari et al., 2019). Siguiendo a Scutari et al., SAMBA utiliza un algoritmo de aprendizaje *score-based* (*hill-climbing search*) para el aprendizaje de estructuras por dos razones: en primer lugar, este tipo de algoritmos suelen ser más rápidos que los otros dos tipos de algoritmos; en segundo lugar, este tipo de algoritmos son, en general, más precisos tanto para tamaños de muestra pequeños como grandes. El *hill-climbing search* (Selman and Gomes, 2006) explora el espacio del DAG mediante adición, eliminación o inversión de un único link. La implementación optimizada utiliza la captura de puntuaciones, la descomposición de puntuaciones y la equivalencia de puntuaciones para reducir el número de pruebas duplicadas (Scutari, 2010). La *hill-climbing search* proporciona tres funciones de puntuación para asignar un índice a nuestro modelo de red:

a. *The Akaike Information Criterion* (AIC)

b. *The Bayesian Information Criterion* (BIC)

c. *The multinomial log-likelihood* (loglik)

Estas puntuaciones estiman la calidad de un modelo. El método de puntuación se ajusta para el método hc(), de modo que generalmente se prefiere una puntuación más alta. Para la distribución ZINB, la puntuación se calcula sobre el modelo de regresión ajustado.

La estructura aprendida también dependerá de si el usuario proporciona o no una *whitelist* y/o *blacklist*. Estas listas se utilizan en la función hc() del paquete bnlearn.

IV) Ajuste de los parámetros del modelo: una vez construida la estructura de la BN, se utiliza el conjunto de datos completo para ajustar los parámetros de la red utilizando la función bn.fit() del paquete bnlearn o un método de ajuste mixto utilizando bn.fit() y zeroinfl() del paquete pscl (Zeileis et al., 2008). La función bn.fit() ajusta, asigna o sustituye los parámetros de una BN basándose en su estructura, mientras que zeroinfl() ajusta un modelo binomial negativo inflado a cero de un taxón dado utilizando la estructura de la red. Si se aplica el filtro de prevalencia, se utiliza un conjunto de datos filtrado que contiene los nodos restantes de la red para ajustar los parámetros. Una vez ajustado, se calcula la fuerza de cada asociación o link utilizando la función arc.strength() del paquete bnlearn para eliminar aquellos enlaces que tengan un valor de fuerza superior a los umbrales introducidos por el usuario en la interfaz. Esta fuerza se calcula utilizando el valor BIC y el criterio de *Mutual Information* (MI). Si un enlace tiene un valor de fuerza superior a los umbrales marcados pero se encuentra incluido en la *whitelist*, esta relación se conserva. Los resultados de este modelo BN son:

   - Archivo RData que contiene las relaciones BN modelada entre variables y taxones.
   - Dos archivos de texto que contienen los valores de fuerza de cada conexión utilizando los criterios BIC y MI.
   - Archivo CSV que contiene los recuentos de taxones normalizados.
   - Archivo CSV que contiene los recuentos de taxones normalizados en escala logarítmica.
   - Fichero log.
   - Si se ha realizado un filtrado, un archivo CSV que incluye los recuentos normalizados en escala logarítmica de los taxones que han superado las condiciones introducidas, y un archivo CSV que contiene el nombre de los taxones que han sido eliminados de la red.

La función Future() del paquete future (Bengtsson, 2021) se utiliza durante los pasos mencionados para permitir al usuario seguir utilizando el resto de la aplicación mientras se calcula un modelo en segundo plano, ya que este proceso puede llevar mucho tiempo dependiendo de los datos *input*.

Además, este módulo permite al usuario cargar el modelo BN creado para posteriores análisis en SAMBA utilizando la opción “Load network” y comparar dos condiciones experimentales diferentes (o evidencias) utilizando la opción “Evidence/Control”. En este último modo, el usuario puede seleccionar una o dos evidencias para reportar estadísticas resumidas sobre los valores de cada taxón dentro de cada evidencia. La tabla incluye, para cada taxón, el recuento medio, la desviación estándar y los rangos de los cuartiles. En esta interfaz también se visualiza un histograma y un gráfico de densidad como representación visual de la distribución del taxón dada cada evidencia.

6.2.2.2. Network Reports

En el módulo "Network Reports" se informa de la CPT (del inglés, *Conditional Probability Table*) de cada nodo de la red. Se puede seleccionar una CPT específica mediante un menú desplegable, la cual se mostrará en la sección llamada "Conditional probability table". El usuario puede descargar la CPT actual o todas las CPTs en un archivo zip. Una CPT muestra el tipo de relaciones entre diferentes taxones en diferentes condiciones experimentales.

Además, en este módulo, el usuario también puede investigar la relación parcial de cada nodo con la opción "DAG", visualizando el *Markov blanket* de cada nodo e informando de cualquier implicación comprobable sobre la independencia condicional.

6.2.2.3. Prediction

El tercer módulo de SAMBA asigna valores a los taxones e infiere el metagenoma dadas una condiciones experimentales específicas (estados o valores) definidas por el usuario. La predicción de los valores de los taxones consiste en predecir los recuentos normalizados de cada taxón dada una combinación de estados de variables experimentales dada como una evidencia.

Para ejecutar la predicción de valores de este módulo, se requiere un fichero RData conteniendo el modelo BN, el cual debe cargarse en la opción "Load network" disponible en el módulo llamado "Input Network", y una evidencia, es decir, una combinación de estados de las variables experimentales discretas.

Para la ejecución de la inferencia del metagenoma, es necesario un fichero de texto que contenga los recuentos de cada taxón y muestra, y un archivo fasta que contenga las secuencias correspondientes a cada taxón.

La predicción de valores de los taxones se realiza a partir del muestreo de las redes dadas las evidencias especificadas por el usuario. Una vez obtenidas las muestras de la red, se calcula la siguiente tabla resumen:

a. El valor medio o esperado de las cuentas normalizadas.

b. La desviación estándar de las cuentas previstas en todas las muestras.

c. Rangos de cuartiles.

d. Una desviación estándar alrededor del rango medio (media $\pm$ la desviación estándar) y el valor correspondiente de la densidad de probabilidad del rango. La densidad de probabilidad del rango se calcula a partir de la densidad de probabilidad de las muestras generadas.

e. Se infiere un resumen similar opcional sobre los datos *input* originales dadas las mismas pruebas, incluyendo el promedio de las cuentas normalizadas para cada taxón para las condiciones experimentales seleccionadas utilizando la función

mean() implementada en R y la desviación estándar para cada taxón para las condiciones experimentales seleccionadas utilizando la función sd() también implementada en R.

En los diferentes modos que utilizan la distribución Log-normal, las frecuencias normalizadas en escala logarítmica se obtienen mediante la función cpdist() de bnlearn. Esta función devuelve un conjunto de datos que contiene las muestras generadas a partir de la distribución condicional de los nodos en función de los estados seleccionados para cada variables experimental en una combinación y el peso de cada valor.

En el modo de distribución ZINB, se implementa un método de muestreo personalizada para muestrear a partir de los modelos ZINB ajustados de cada taxón de la red. En primer lugar, se extrae de la red una ruta de muestreo basada en la jerarquía de nodos. A continuación, la ruta de muestreo se utiliza para muestrear cada nodo de la ruta en orden mediante el muestreo aleatorio a partir de la distribución ZINB utilizando los coeficientes del modelo ajustado. Para tener en cuenta la propagación de la incertidumbre a través de la ruta de muestreo, la media ejecutada del modelo se indica junto a las medias de las muestras.

Se implementó un paso de corrección opcional para corregir la predicción final utilizando los datos *input* como conocimiento a priori para inferir unas muestras posteriores corregidas. Utilizando la regla de Bayes, la densidad de probabilidad de las muestras generadas se multiplica por la densidad de probabilidad de los datos *input* dada una evidencia para obtener nuevas muestras posteriores.

A continuación, se calcula una media ponderada del valor predicho normalizado (NPV) para las muestras generadas mediante la siguiente función:

$$\mathrm{NPV} = e^{\left(\frac{\sum_{i=1}^{n} P_i \times W_i}{\sum_{i=1}^{n} W_i}\right)} - 1$$

Donde n es el número de muestras generadas a partir de la red, i es una muestra generada específica, P es el valor de la muestra generada y W son los pesos de una muestra generada.

Respecto a la inferencia del metagenoma utilizando PICRUSt2 (Douglas et al., 2020), ésta consta de cuatro pasos:

1. Colocar las secuencias en el árbol de referencia. Este árbol de referencia se basa en 20.000 secuencias 16S de genomas de la base de datos Integrated Microbial Genomes (Chen et al., 2019b). Este paso consiste en:
    - Alinear las secuencias de los taxones con un alineamientos de secuencias múltiples de secuencias 16S de referencia con HMMER (Potter et al., 2018).
    - Encontrar las posiciones más probables de los taxones en el árbol de referencia con EPA-NG (Barbera et al., 2019) o SEPP (Janssen et al., 2018).

- Fichero *output* (tree.out) conteniendo el árbol con las posiciones más probables para cada taxón con GAPPA (Czech et al., 2020).

2. Predicción *hidden-state* de familias de genes, prediciendo el número de copias y enzimas de familias de genes para cada taxón. Se generan dos ficheros gz como *output*: 16S_predicted_and_nsti.tsv.gz y EC_predicted.tsv.gz.

3. Generación de predicciones del metagenoma. Los perfiles funcionales del metagenoma por muestra se generan basándose en las funciones predichas para cada secuencia de estudio. La tabla de abundancia de secuencias especificada será normalizada por el número predicho de copias de genes marcadores. La predicción genera cinco archivos *output*:

    - pred_metagenome_unstat.tsv.gz: abundancias globales de los códigos de enzima por muestra.
    - pred_metagenome_unstrat_descrip.tsv.gz: descripciones funcionales añadidas al archivo pred_metagenome_unstrat.tsv.gz.
    - pred_metagenome_contrib.tsv.gz: tabla estratificada que desglosa cómo contribuyen los taxones a las abundancia de las familias de genes en cada muestra.
    - seqtab_norm.tsv.gz: tabla de abundancia de taxones normalizada por el número de copias 16S predicho.
    - weighted_nsti.tsv.gz: valor NSTI medio por muestra teniendo en cuenta la abundancia relativa de los taxones. Este archivo puede ser útil para identificar muestras atípicas en el conjunto de datos.

4. Inferencia a nivel de ruta, que infiere la presencia y abundancia de rutas basándose en la abundancia de familias de genes en una muestra. En el directorio que se genera, llamado “pathways_out”, se crean varios archivos *output*:

    - path_abun_contrib.tsv.gz: tabla estratificada que desglosa cómo contribuyen los taxones a la abundancia de rutas en cada muestra.
    - path_abun_unstrat.tsv.gz: abundancias globales de las rutas por muestra.
    - path_abun_unstrat_descrip.tsv.gz: descripciones funcionales añadidas al archivo path_abun_unstrat.tsv.gz.

6.2.2.4. Network Viewer

El cuarto módulo de SAMBA, desarrollado en su gran mayoría por una compañera del equipo, representa el grafo del modelo BN permitiendo al usuario modificar/editar/personalizar el color, la forma, la etiqueta o el tamaño, entre otros, de nodos y links. Además, este módulo permite crear grupos de nodos para visualizarlos en lugar de la estructura completa, o para trabajar con un subgrafo específico. El fichero RData que contiene el modelo BN debe utilizarse como *input* en la opción "Load network" del primer módulo.

El gráfico se construye utilizando varias funciones: la función subgraph() del paquete bnlearn se utiliza para trazar el grafo; las funciones visIgraph() y renderVisNetwork() del paquete visNetwork (Almende et al., 2019) se utilizan para proporcionar una visualización interactiva; y la función strength.viewer() del paquete bnviewer (Fernandes, 2020), que muestra la fuerza de las relaciones probabilísticas expresadas por los enlaces de una red bayesiana y utiliza el promedio del modelo para construir un grafo que contenga solamente los enlaces significativos.

Este módulo contiene una tabla llamada "Node info", que muestra el CPT de un nodo concreto, y otra tabla llamada "Edge info" que muestra la fuerza de una relación específica. También se muestran dos tablas que incluyen información sobre el conjunto completo de nodos y relaciones mediante las funciones datatable() y dataTableProxy() del paquete DT (Yihui, 2022). Estas funciones permiten tanto navegar y filtrar la información, como interactuar directamente con el grafo resaltando, seleccionando o editando funciones, para lo cual se ha utilizado funciones del paquete VisNetwork, como visOptions(), visRedraw(), visSetSelection(), visUpdateNodes() y visUpdateEdges(), y código JavaScript introducido mediante la función runjs() del paquete shinyjs (Attali, 2021) y la función JS() del paquete htmlwidgets (Vaidyanathan et al., 2022). Sobre esta base 1) se han incorporado funciones de edición, filtrado y resaltado que pueden encontrarse en la barra lateral izquierda, y 2) también se ha introducido la posibilidad de descomponer el grafo en componentes o subgrafos utilizando la función decompose() del paquete igraph (Csardi and Nepusz, 2006).

El gráfico resultante puede descargarse como archivo HTML, PNG, JPEG o PDF. Además, se puede hacer una captura de pantalla para exportar una imagen de la red actual utilizando el paquete shinyscreenshot (Attali et al., 2021).

6.2.2.5. Results

El quinto módulo de SAMBA permite al usuario descargar un fichero zip conteniendo los archivos resultantes obtenidos a partir de la opción "Learning and training" del primer módulo y un archivo zip resultante de la predicción del metagenoma, realizada a partir de la pestaña "Prediction", a través de un menú desplegable.

6.2.3. Guía del usuario

SAMBA proporciona el siguiente flujo de trabajo para saber cómo se relacionan entre sí las variables y los taxones de un conjunto de datos determinado. El primer paso es crear el modelo BN a partir del conjunto de datos del usuario seleccionando la pestaña "Input Network" y la opción "Learning and training". Como se muestra en la Figura 6.1, se necesitan dos archivos CSV, uno con los datos de las variables experimentales y otro con las cuentas crudas de cada taxón por muestra. Una vez cargados ambos archivos, las variables experimentales y los datos de las cuentas crudas pueden revisarse en la pestañas "Experimental Variables Review" y "Count Data Review", respectivamente.

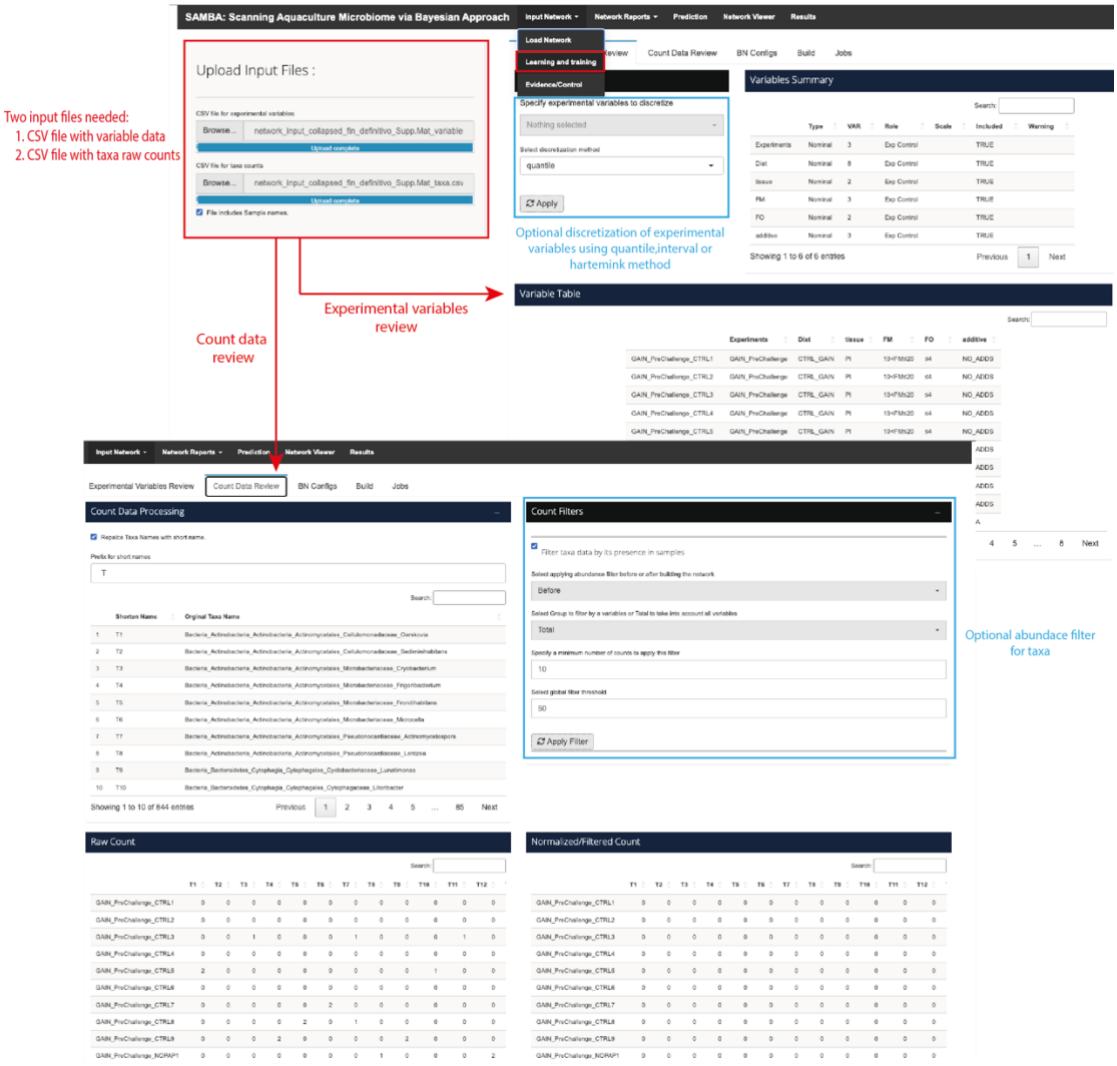

Figura 6.1. Opción "Learning and training". Arriba, interfaz para la pestaña "Experimental Variables Review", que incluye un resumen y una tabla sobre las variables, y una opción para discretizar una o más variables si es necesario. Abajo, interfaz para la pestaña "Count Data Review", que incluye una opción para cambiar los nombres largos de los taxones por otros más cortos, una opción para filtrar los taxones por su abundancia en las muestras, y dos tablas que muestran las cuentas crudas y normalizadas/filtradas por taxón y muestra.

En la pestaña “Experimental Variables Review”, el usuario puede visualizar los datos de las variables experimentales. Además, las variables experimentales continuas pueden discretizarse seleccionando una o más de ellas mediante un menú desplegable que contiene los nombres de las variables, y utilizando uno de los siguiente métodos: “quantile”, “interval” o “hartemink”. Hay que tener en cuenta que la discretización de las variables afectará a las relaciones que se establecerán durante los pasos de aprendizaje y entrenamiento de la construcción del modelo BN.

En la pestaña “Count Data Review”, el usuario puede, por un lado, visualizar los datos de cuentas crudas y normalizadas, y, por otro, cambiar los nombres de los taxones por otros más cortos y filtrar o eliminar aquellos taxones por sus abundancias en las muestras.

Tras la revisión de las variables experimentales y de las cuentas, el usuario puede modificar o adaptar las opciones de configuración para la construcción del modelo BN. Estas opciones incluyen: “Network score” (AIC, BIC, loglik o BIC-ZINB), “Taxa distribution” (Log-Normal o ZINB), "Mutual Information threshold", "Bayesian Information Criterion (BIC) threshold", "Remove Experimental Variables interactions/relations from the network model structure", "Blacklist" and "Whitelist" (Figura 6.2). La opción “Remove Experimental Variables interactions/relations from the network model structure” permite al usuario predecir los valores de los taxones en un conjunto de condiciones no estudiadas anteriormente en un experimento. Para más información sobre el proceso de construcción de la red bayesiana, consulte la sección “6.2.2.1. Input Network” de este capítulo.

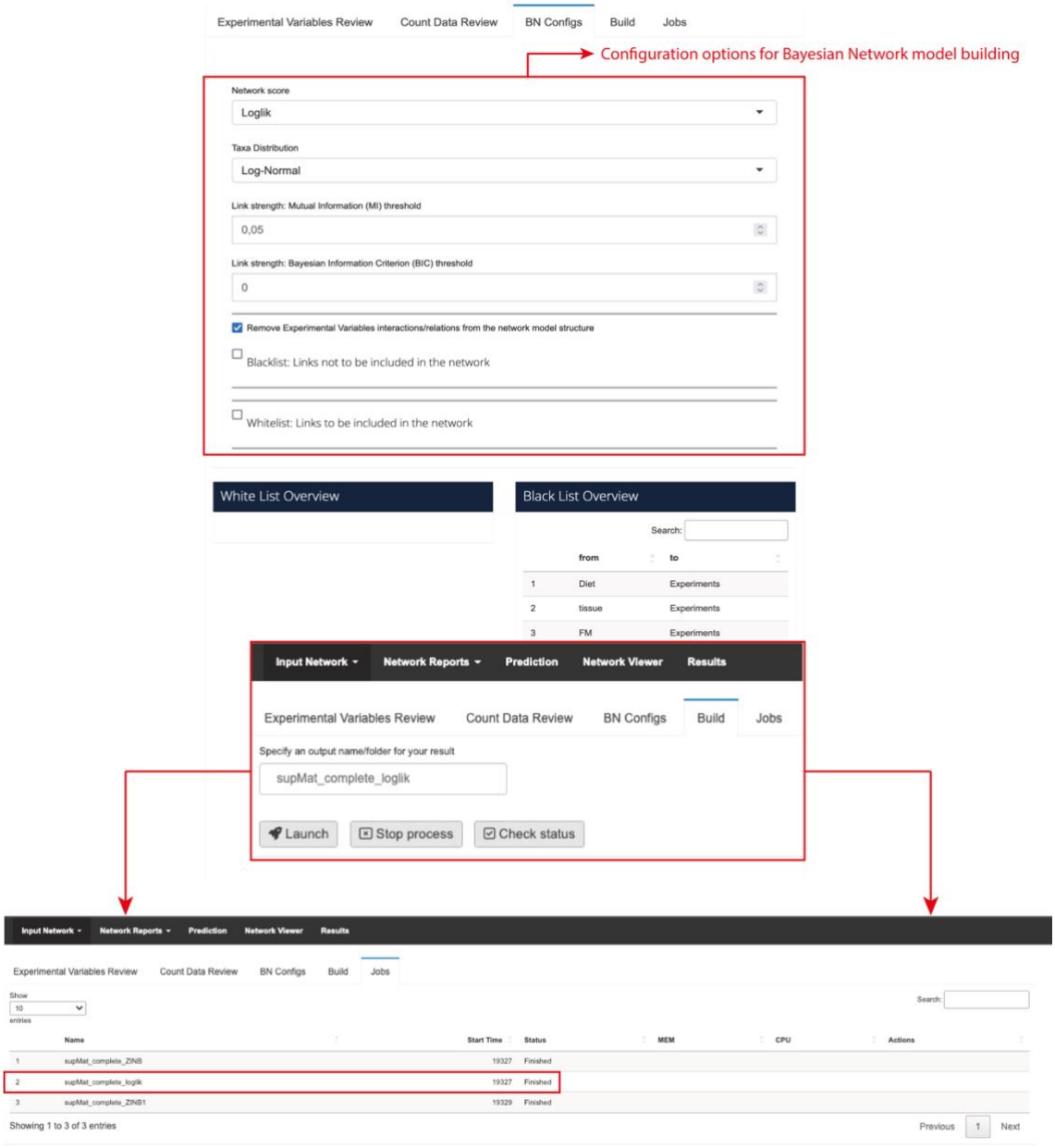

Figura 6.2. Arriba, interfaz para configurar la construcción del modelo de red bayesiana, incluyendo la puntuación o *score* de la red, *taxa distribution*, los umbrales de fuerza de enlace (*Mutual Information* y BIC), *blacklist*, *whitelist* y la opción "Remove Experimental Variables interactions/relations from the network model structure". También se muestra una tabla para la *whitelist* y otro para la *blacklist*. En el centro, interfaz para construir el modelo BN, Abajo, interfaz para el seguimiento de los trabajos lanzados actualmente o en el pasado.

Una vez finalizada la construcción del modelo de red bayesiana, el usuario puede descargar el archivo zip resultante a través de la pestaña "Results" (Figura 6.3).

Figura 6.3. Interfaz para descargar el resultado de la construcción del modelo BN.

A continuación, para seguir el análisis en SAMBA, el usuario debe cargar el fichero RData generado en "Load Network", dentro de la opción "Input Network" (Figura 6.4).

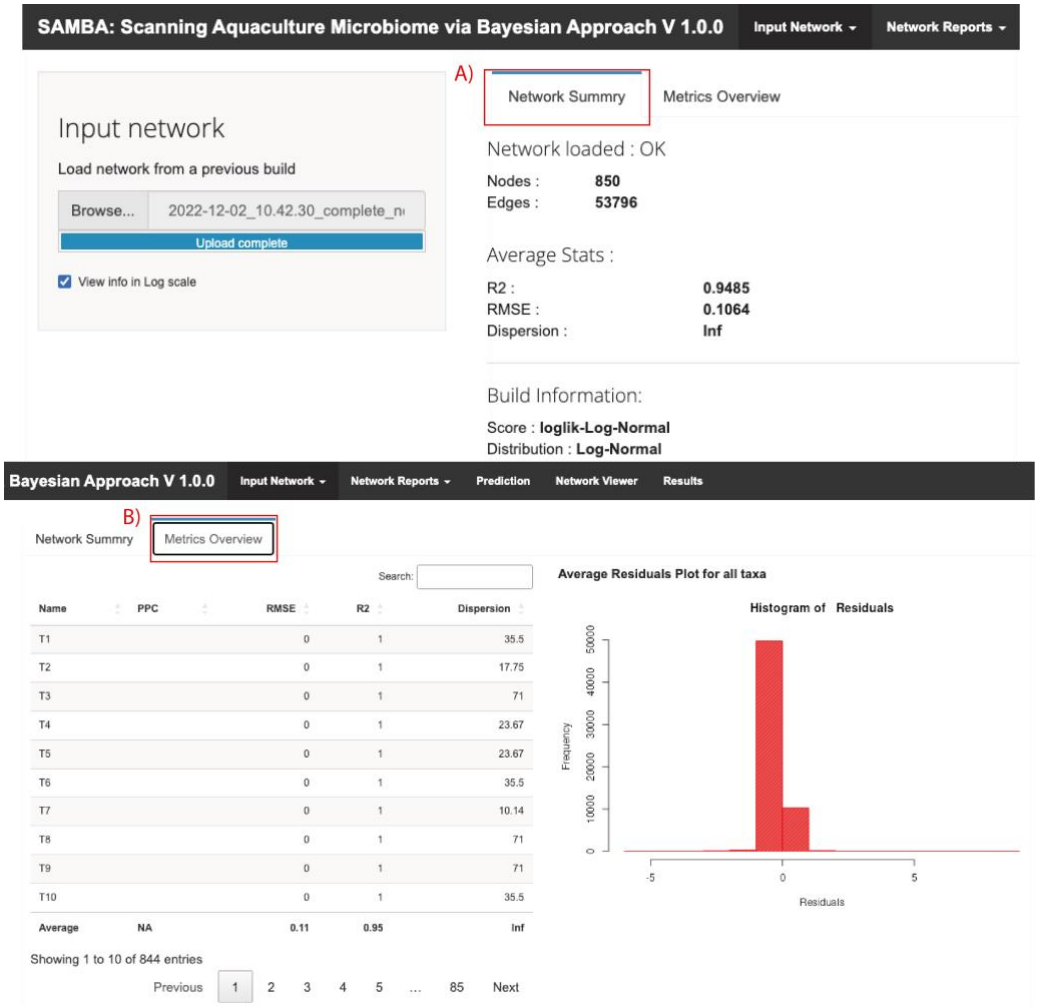

Figura 6.4. Interfaz para cargar el modelo BN generado en SAMBA. A) Pestaña "Network Summary" que incluye información básica sobre la red. B) Pestaña "Metrics Overview" que incluye estadísticas sobre los taxones y un histograma de los residuos medios de todos los taxones.

Esta interfaz ofrece información básica sobre la red en la pestaña "Network Summary", incluido el número de nodos y links, estadísticas medias e información sobre la construcción del modelo. Además, la pestaña "Metrics Overview" muestra una tabla con información sobre PPC, error cuadrático medio (RMSE), R2 y dispersión de los taxones; y un histograma que muestra los residuos medios de todos los taxones.

Una vez cargado el archivo RData, el usuario puede visualizar y comparar cómo se distribuyen los datos en diferentes condiciones o evidencias utilizando la opción "Evidence/Control" dentro de la pestaña "Input Network" (Figura 6.5). Aquí, haciendo clic en el nombre de un taxón, se muestra la siguiente información:

1) una tabla para una combinación que incluye el número de muestras que coinciden con la evidencia introducida, el valor medio, la desviación estándar y los cuartiles para cada taxón.

2) Un histograma para una combinación que muestra las frecuencias de las cuentas normalizadas.

3) Un gráfico de densidad que muestra la estimación de la densidad del kernel, es decir, una versión suavizada del histograma.

Figura 6.5. Interfaz "Evidence/Control". A) Configuración global de la comparación. En este ejemplo, las condiciones EGG.HYDRO y CTRL_EWH (referencia) se comparan con EGG.HYDRO y LOWFMFO (*target*). B) Diagrama de densidad de la comparación. La referencia aparece en rojo y el *target* en azul. C) Tablas con información sobre los taxones en las condiciones de referencia y *target*. D) Histograma de frecuencias de cuentas normalizadas del taxón T770 en las condiciones de referencia (rojo) y *target* (azul).

Los informes sobre la red pueden obtenerse después de cargar el archivo RData y seleccionando una opción ("CPTs" o "DAG") en la pestaña "Network Reports". En la opción "CPTs", como se muestra en la Figura 6.6, para taxones o variables experimentales continuos, SAMBA muestra sus CPTs incluyendo el tipo de relaciones (negativas o positivas) con otros taxones o variables en diferentes configuraciones discretas de los nodos padres. También se muestra un histograma de los residuos y un gráfico Q-Q normal.

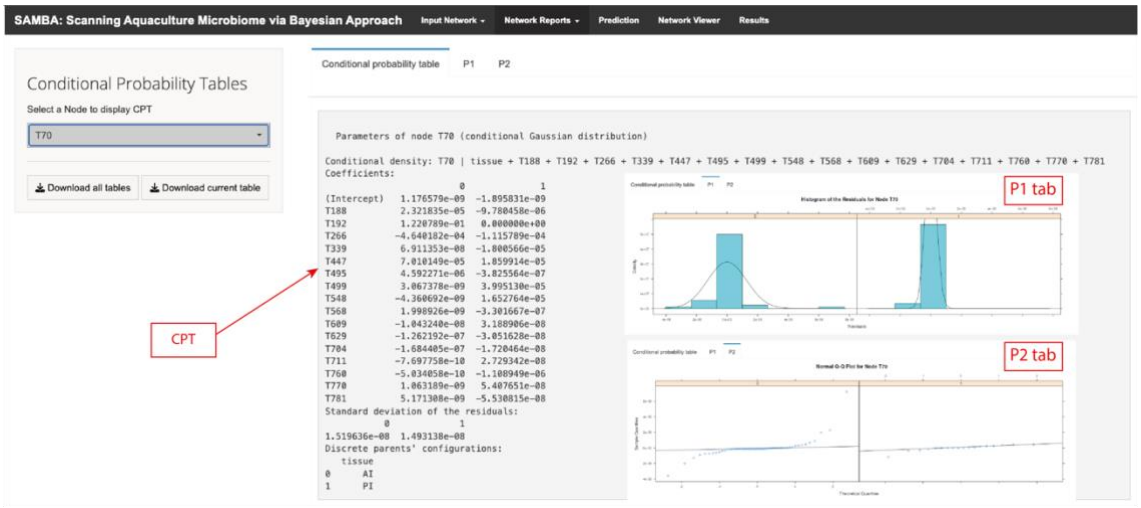

Figura 6.6. Figura mostrando la CPT, el histograma de residuos y el gráfico Q-Q normal para el taxón T70.

Para una variable discreta, SAMBA muestra su CPT incluyendo la probabilidad de cada estado dadas otras variables discretas y un gráfico de probabilidades condicionales para el nodo (Figura 6.7).

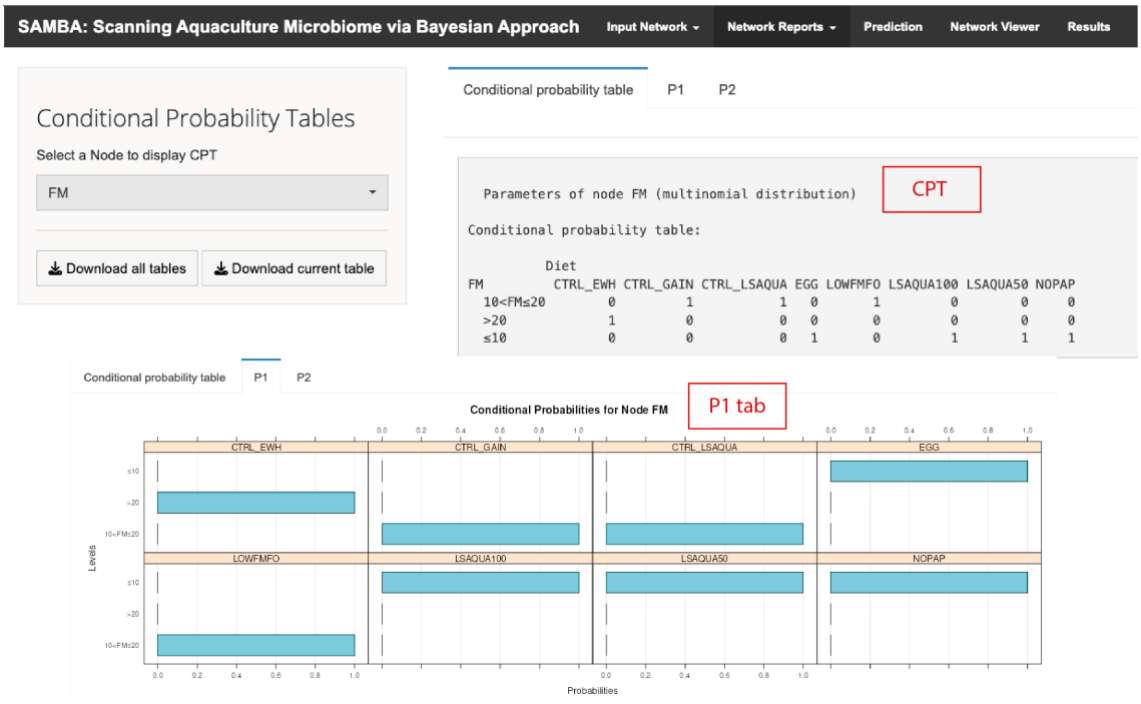

Figura 6.7. Figura mostrando la CPT y el gráfico de probabilidades condicionales para el nodo FM.

En la opción “DAG”, se muestran las relaciones directas y/o indirectas de un determinado taxón, tal y como se muestra en la Figura 6.8. También se muestra la dependencia condicional con variables experimentales y las implicaciones comprobables sobre la independencia condicional.

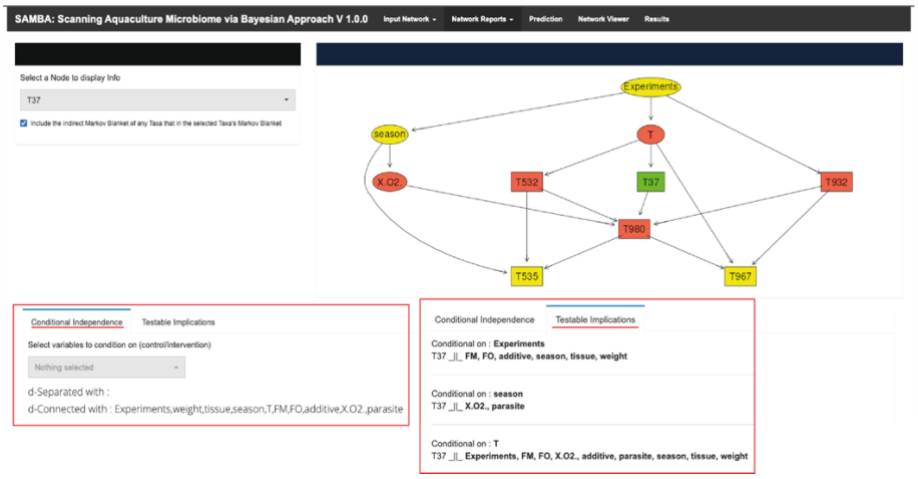

Figura 6.8. Interfaz para la opción “DAG” para el nodo T37 (en verde). Los nodos con una relación indirecta están coloreados en amarillos y los nodos con una relación directa están coloreados en rojo. Los nodos que corresponden a taxones se representan con una forma cuadrada mientras que los nodos que corresponden con variables experimentales se representan con una forma ovalada. También se muestran la independencia condicional y las implicaciones comprobables del nodo T37.

La pestaña “Prediction” permite al usuario obtener valores predichos para uno o más taxones dadas unas condiciones experimentales específicas dentro de la opción “Predicted value”, e inferir el metagenoma dentro de la opción “Infer metagenome”. Para predecir los valores de los taxones, el usuario debe seleccionar los nombres de los taxones en un menú desplegable y en qué condiciones experimentales quiere hacer la predicción de valores dentro de la opción “Predicted value” (Figura 6.9). La predicción puede realizarse teniendo en cuenta los datos input originales como conocimiento previo si el usuario selecciona esta opción (“Use Input Data as strong prior for network prediction"). Además, se puede añadir un resumen de los datos originales a la tabla resultante. Para más información sobre el proceso de predicción y sus resultados, consulte la sección “6.2.2.3. Prediction” de este capítulo.

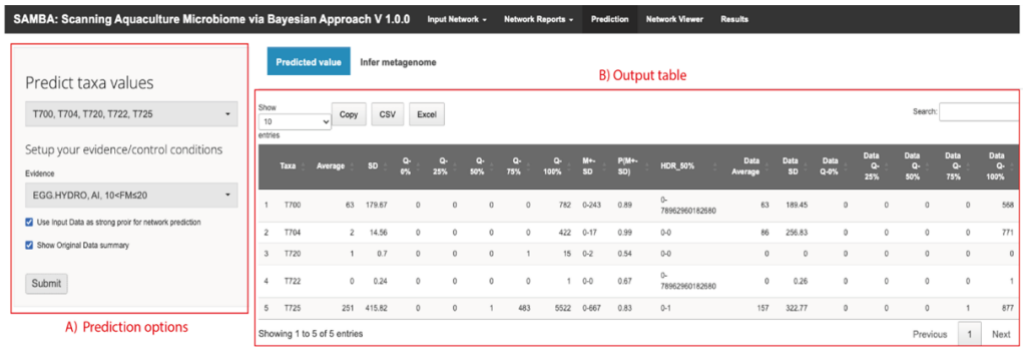

Figura 6.9. Interfaz para la predicción de valores de los taxones. A) Opciones de predicción para seleccionar taxones y condiciones experimentales (evidencia). B) Tabla resultante que contiene los datos de la predicción y el resumen de los datos originales.

La inferencia del metagenoma puede realizarse en la opción “Infer metagenome” cargando un archivo de texto que incluya las cuentas por taxón y muestra, y un archivo fasta que incluya las secuencias de referencia de los taxones. Los resultados de la predicción del metagenoma pueden descargarse desde la pestaña “Results”.

El último paso consiste en visualizar y editar la red. Para ello, el usuario debe ir a la pestaña “Network Viewer” y pulsar el botón “Refresh” para visualizar la estructura de la red. Como se muestra en la Figura 6.10, la pestaña “Network Viewer” está compuesta por una sección que incluye (A) ajustes globales, (B) un panel gráfico, (C) un menú “Edit”, “Select” y “Filter”, (D) un gestor de datos para nodos y links, y (E) una tabla de información de nodos y links.

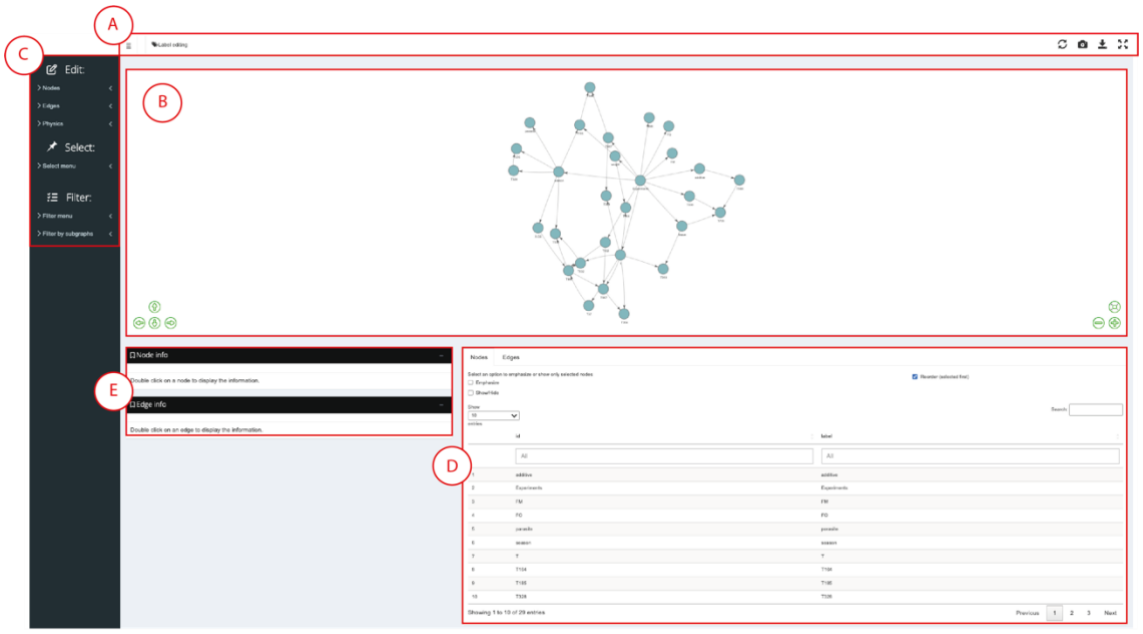

Figura 6.10. Interfaz “Network Viewer”. A) Ajustes globales, que incluyen las opciones "Display the edition sidebar", "Label editing", "Refresh Network", "Take a screenshot", "Save network" and "Full-screen mode". B) Panel gráfico, que incluye opciones para cambiar la posición de la red, reajustar el tamaño y la posición, y ampliar o reducir el tamaño. C) Menú que contiene opciones para modificar la forma, tamaño, color, sombra y etiqueta de nodos y links (menú “Edit”); para seleccionar un subconjunto de nodos relacionados con el nodo que el usuario seleccionó a partir de la dirección de los link y su grado de distancia (menú “Select”); y para crear grupos de nodos seleccionándolo en el gráfico o introduciendo los nombres de los nodos, y para resaltar o solamente mostrar un subgrafo concreto (menú “Filter”). D) Gestor de datos que incluye información sobre nodos y links en formato tabla. E) Información de nodos y link que muestra la CPT del nodo y la fuerza del link, respectivamente.

El panel gráfico permite al usuario visualizar la red y cambiar su posición, ampliar o reducir su tamaño y volver a su posición original. Además, los nodos del gráfico pueden moverse a conveniencia del usuario haciendo clic sobre ellos y arrastrándolos. El usuario puede modificar y personalizar el color, el tamaño, las características de las etiquetas, la forma y la sombra de los nodos y links haciendo clic en los nodos o links de interés y cambiando estas características en el menú “Edit”. Este menú tiene también una opción llamada “Fisics”, que incluye funciones para fijar los ejes X y/o Y, de modo que el panel gráfico quede fijo, o para dar un efecto rebote a la red. Los nodos no solo se pueden

seleccionar manualmente, sino que también pueden seleccionarse mediante el menú “Select”. Si se elige un nodo, el usuario puede seleccionar, a través de este menú, un grupo de nodos relacionados que sean padres y/o hijos (opción “Direction”) de este nodo dado un cierto grado de distancia (en links) entre ellos. Otra opción para crear un conjunto de nodos es seleccionarlos en la tabla de información sobre los nodos. Cuando se selecciona un conjunto de nodos, el usuario puede crear un grupo utilizando el menú “Filter” para destacarlo u ocultar nodos que no estén presentes en este conjunto, de forma que el usuario pueda trabajar con los nodos de interés. Una alternativa para crear un grupo en este menú es escribir los nombres de los nodos en la opción “Input a list of nodes" o seleccionar un subgrafo concreto de la red utilizando la opción “Filter by subgraphs". Cuando se selecciona un conjunto o grupo de nodos, se pueden aplicar modificaciones de características de nodos y links en todos ellos al mismo tiempo utilizando el menú “Edit”. Una vez editado y/o filtrado el grafo, el usuario puede guardarlo en varios formatos, como HTML, PDF, PNG o JPEG.

Por último, en este módulo, el usuario dispone de información adicional sobre la CPT de un nodo seleccionado (sección “Node info”), la fuerza de un enlace seleccionado (sección “Edge info”), una tabla que contiene las etiquetas e identificadores de los nodos (pestaña “Nodes”), y una tabla que incluye las relaciones establecidas entre los nodos, sus identificadores y sus respectivas fuerzas (pestaña “Edges”).

6.3. Resultados

6.3.1. Visión general

SAMBA es una aplicación web que lee e indexa datos de distintas variable bióticas y abióticas que intervienen en la dinámica de un determinado sistema de acuicultura para aprender la estructura de red de ese sistema y crear y entrenar un modelo de red bayesiana basado en esa información. Las variables típicas de importancia en los sistemas de acuicultura para construir un modelo de red bayesiano con SAMBA son el sexo, la edad, el tamaño, los antecedentes genéticos, el tejido, las frecuencias microbianas (anotaciones taxonómicas a partir de la secuenciación del amplicón 16S), la estación del año, la composición de la dieta, el pH, la temperatura, etc. A continuación, SAMBA infiere las probabilidades condicionales de todas las posibles interrelaciones de dependencia detectadas entre estas variables indexadas en el modelo de red bayesiana, con especial atención a los distintos taxones que constituyen el pan-microbioma de los peces del sistema modelado, a partir del cual SAMBA también permite inferir un perfil del metagenoma. Una vez creado, entrenado y validado el modelo de red bayesiana, SAMBA permite al usuario inferir y/o extraer información del modelo de red como informe o predicción. Además, SAMBA implementa un completo editor gráfico de redes que proporciona múltiples herramientas distintas para visualizar, editar, personalizar y exportar redes en distintos formatos.

Como se muestra en la Figura 6.11, esta GUI está estructurada en cinco módulos que proporcionan implementaciones de interfaz a los distintos módulos que componen SAMBA:

1) Input Network. Módulo que representa el primer paso del flujo de trabajo de SAMBA. Se compone de tres secciones diferentes: "Load Network", donde se puede cargar un modelo BN para realizar análisis posteriores en SAMBA; "Evidence/Control", donde SAMBA expone una tabla resumen para cada taxón y un gráfico que muestra la distribución de valores en una condición o dos condiciones diferentes si el usuario selecciona la opción de compararlas; y "Learning and training", que lee los datos *input* y construye un modelo BN utilizando el paquete de R llamado bnlearn (Scutari, 2010) con el algoritmo Hill-Climbing de aprendizaje basado en puntuaciones (Selman and Gomes, 2006).

2) Network Reports. Una vez creado y entrenado el modelo BN, este módulo utiliza los paquetes bnlearn y dagitty (Textor et al., 2016) para obtener información clave sobre la causalidad potencial y las relaciones entre los taxones microbianos y todas las demás variables implicadas en la dinámica de las especies acuícolas modeladas. En particular, el usuario puede obtener un informe con métricas y estadísticas con información de la estructura de red del sistema modelado, incluyendo tablas de probabilidad condicional (CPT), valores de probabilidad condicional (CPV), y un gráfico de residuos y Q-Q para cada nodo.

3) Prediction. Este módulo utiliza los paquetes bnlearn y pscl (Zeileis et al., 2008) para inferir predicciones sobre los cambios más probables en la diversidad y las frecuencias del pan-microbioma en función de los cambios de otras variables. Este módulo también permite inferir el perfil metagenómico funcional asociable al pan-microbioma modelado utilizando PICRUSt2 (Douglas et al., 2020). Además, el módulo permite al usuario cambiar la predicción de acuerdo con la experiencia experimental en el sistema modelado, permitiendo así calibrar las predicciones de la red hasta que converjan con la observación del mundo real.

4) Network Viewer. Este módulo es la implementación de un completo visor gráfico de redes basado en distintas funciones de los siguientes paquetes de R: igraph (Csardi and Nepusz, 2006), visNetwork (Almende et al., 2019), bnviewer (Fernandes, 2020), shinythemes (Chang, 2021), shinyBS (Bailey, 2022), tibble (Muller and Wickham, 2022), colourpicker (Attali, 2022), shinydashboard (Chang and Borges, 2021), shinydashboardPlus (Granjon, 2021), shinyscreenshot (Attali et al., 2021), shinyjs (Attali, 2021), htmlwidgets (Vaidyanathan et al., 2022), stringr (Wickham, 2022), purrr (Henry and Wickham, 2021) y DT (Yihui, 2022). Este editor de grafos proporciona múltiples herramientas distintas para visualizar, editar, personalizar y exportar grafos en distintos modos y formatos, facilitando la navegación e interpretación del grafo (o un subgrafo) resultante del modelo BN.

5) Results. Este último módulo es la sección donde SAMBA almacena los resultados y desde donde el usuario puede descargar los ficheros resultantes generados desde la interfaz "Learning and training" y un archivo zip resultante de la predicción del metagenoma realizada a partir de la pestaña "Prediction".

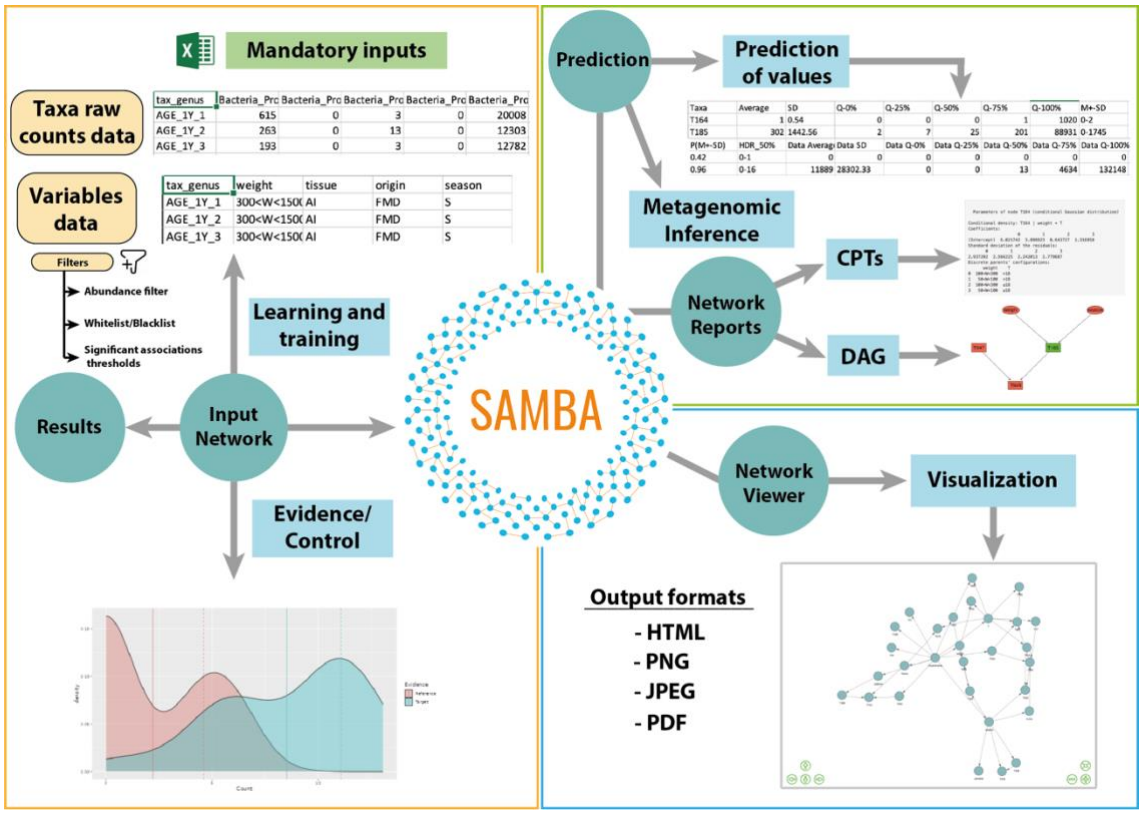

Figura 6.11. SAMBA se implementa en 5 módulos: 1) Input Network. Se puede crear un modelo BN a partir de los datos *input* del usuario utilizando la función "Learning and training". Además, la distribución de valores para una condición o una combinación de condiciones se muestra en la interfaz "Evidence/Control". 2) Prediction. En este módulo se puede realizar la predicción de los valores de los taxones y la inferencia del metagenoma. 3) Network Reports. Módulo para obtener Tablas de Probabilidad Condicional (CPT) y/o Grafos Acíclicos Directos (DAG) para un nodo. 4) Network Viewer. Módulo para visualizar, editar, personalizar y exportar grafos. 5) Results. Módulo para descargar un archivo zip que contiene los resultados de la interfaz "Learning and training".

6.3.2. Validación de caso de estudio

A lo largo del capítulo, se ha expuesto la capacidad de SAMBA para transformar conjuntos de datos *input* de abundancia de taxones en modelos de asociación que contienen conexiones causales putativas entre taxones. Considerando todo este conjunto de relaciones en el modelo de asociación, SAMBA calcula un valor predicho para cada taxón en unas condiciones específicas dadas por las variables experimentales. Sin embargo, este cálculo puede verse influido por la variabilidad técnica y biológica de los conjuntos de datos taxonómicos. Además, los valores predichos calculados deben preservar los valores absolutos y relativos de abundancia de taxones presentes en los datos *input*. Por lo tanto, es necesario validar la correlación entre los valores predichos

y los observados en los archivos *input*. En esta sección de este capítulo, se muestran los resultados de esta correlación para garantizar el buen rendimiento de SAMBA. Para ello, se crearon dos conjuntos de datos de prueba utilizando datos semisintéticos y empíricos (https://github.com/biotechvana/SAMBA). Los conjuntos de datos de prueba incluían dos archivos que contenían las variables experimentales (es decir, dieta, tejido, aditivo, etc.) y las cuentas crudas de cada taxón para cada muestra. Para más información sobre cómo construir los modelos BN con SAMBA, consulte la sección "6.2.3. Guía del usuario" de este texto.

6.3.2.1. Datos semisintéticos: comunidad Mock

El primer intento de validar SAMBA se realizó con un conjunto de datos semisintéticos que minimizaba el efecto de la variabilidad biológica. Se secuenciaron ocho réplicas de ZymoBIOMICS™ Microbial Community Standard II (Zymo Research Corp., CA, Estados Unidos) con la plataforma Oxford Nanopore MinION. Esta comunidad sintética contenía 7 bacterias cuyas cuentas celulares se conocían y distribuían en una escala logarítmica, oscilando entre el 95,9% (*Listeria monocytogenes*) y el 0,0001% (*Staphylococcus aureus*) al secuenciar para cada una de ellas el gen 16S rRNA. Se creó una segunda tabla *input* para este conjunto de datos que contenía las variables artificiales para cada réplica. La etiqueta EXTMETH hacía referencia al tipo de método de extracción utilizado y contenía las condiciones EXT1 y EXT2, para dos métodos diferentes. Las diferentes condiciones de PCR durante el procedimiento se etiquetaron con la variables experimental PCR, que contenía las condiciones PCR1 y PCR2 (Tabla 6.1).

Tabla 6.1. Resumen de todas las combinaciones potenciales y no solapantes de variables experimentales en los datos semisintéticos.

Réplicas	EXTMETH	PCR
Mock Rep. 1	EXT1	PCR1
Mock Rep. 2	EXT1	PCR1
Mock Rep. 3	EXT1	PCR2
Mock Rep. 4	EXT1	PCR2
Mock Rep. 5	EXT2	PCR1
Mock Rep. 6	EXT2	PCR1
Mock Rep. 7	EXT2	PCR2
Mock Rep. 8	EXT2	PCR2

Como era de esperar, no se observó variabilidad biológica entre las muestras, y la abundancia relativa de cada taxón en la comunidad simulada se mantuvo en las ocho réplicas tras la secuenciación y la asignación taxonómica, independientemente del método de extracción y de las condiciones de PCR (Figura 6.12A). Tras la predicción con SAMBA, todos los nodos participaron en al menos una asociación en la red resultante (Figura 6.12B). En el modelo se observa como la variables EXPMETH se asocia con todos los taxones, mientras que el efecto de la variable PCR solamente afecta a las asociaciones con *B. subtilis* y *S. enterica*. Para la validación del rendimiento predictivo de SAMBA para este modelo, se obtuvo los valores predichos para cada taxón bajo todas las combinaciones no solapantes de condiciones y variables experimentales. Los ocho valores resultantes se correlacionaron linealmente con el valor medio observado en las condiciones correspondientes. La correlación de datos por pares fue estadísticamente significativa (P-valor < 0,05) y las regresiones mostraron altos valores R2 ajustados, que oscilaron entre 0,6 y 0,89 para las cuatro bacterias más abundantes de la comunidad (Figura 6.12C-F).

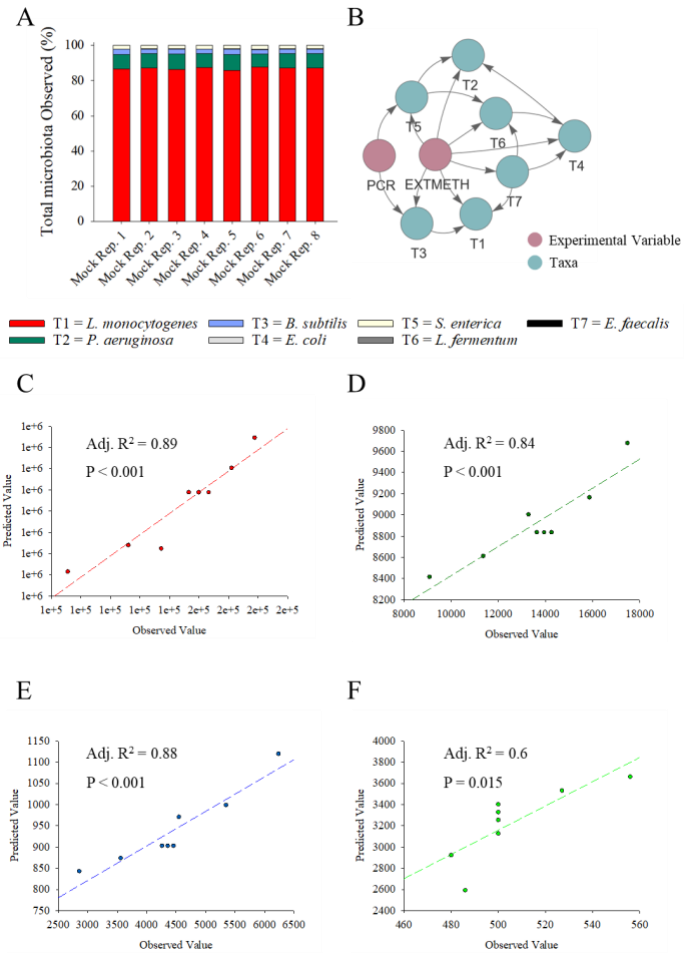

Figura 6.12. (A) Diagramas de barras que muestran la abundancia relativa de las bacterias que componen la comunidad simulada en las ocho réplicas. (B) Captura de pantalla del modelo de asociación creado por SAMBA. El resto de los paneles de la figura muestran los resultados de la correlación entre los valores observados presentes en el conjunto de datos *input* y los valores predichos por la herramienta para (C) *L. monocytogenes*, (D) *P. aeruginosa*, (E) *S. enterica* y (F) *B. subtilis*. Las regresiones mostraron la correlación entre los valores de una combinación específica y no superpuesta de estados de variables experimentales.

6.3.2.2. Datos empíricos: conjunto de datos reales

El conjunto de datos reales se construyó utilizando datos de tres ensayos de alimentación publicados para analizar la población de microbiota autóctona de secciones anteriores y posteriores del intestino de 72 peces seleccionados al azar. Brevemente, estos ensayos de alimentación revelaron cómo variaba la microbiota intestinal con (i) proteínas bacterianas unicelulares y proteínas animales procesadas como principal fuente de proteína dietética (LSAQUA; PRJNA713764; Solé-Jimenez et al., 2021), (ii) suplementación de dietas vegetales con hidrolizados de clara de huevo con actividad antioxidante y antiobesogénica en ratas obesas (EGGHYDRO; PRJNA705868; Naya-Català et al., 2021a), y (iii) nuevas formulaciones de piensos suplementadas con un promotor de salud intestinal (GAIN_PRE; PRJNA750446; Piazzon et al., 2022) (Tabla 6.2). Para este primer intento de validación de SAMBA con datos reales, todos los ensayos de alimentación se llevaron a cabo en paralelo (primavera-verano de 2020) en la misma infraestructura de investigación (Instituto de Acuicultura Torre de la Sal, Castellón, España) en condiciones naturales de luz y temperatura a nuestra latitud (40°5'N; 0°10'E), utilizando peces hermanos del mismo lote de criadero (Avramar, Burriana, España) para minimizar la variabilidad ambiental y genética del metagenoma de los peces, que presenta una gran variabilidad individual dentro de los ensayos ganaderos y entre otros.

Tabla 6.2. Resumen de los experimentos con dorada incluidos en el conjunto de datos reales. Todos los experimentos mencionados consistieron en la amplificación de las regiones hipervariables V3-V4 del gen 16S rRNA para evaluar los cambios en la composición de la microbiota adherida a la mucosa intestinal.

Experimentos	Pez	Estímulo	Duración (días)	Referencia
LSAQUA	27	LSAqua SusPro (proteínas animales procesadas + proteínas unicelulares bacterianas) como sustituto de la harina de pescado.	56	Solé-Jiménez et al., 2021
EGGHYDRO	27	Dieta baja en harina de pescado/aceite de pescado con/sin suplementación con un hidrolizado bioactivo de clara de huevo.	56	Naya-Català et al., 2021a
GAIN_PRE	18	Dietas sin harina de pescado con/sin un aditivo comercial para piensos beneficioso para la salud (SANACORE® GM).	34	Piazzon et al., 2022

Para construir el archivo de variables experimentales en el procedimiento de validación, definimos cada ensayo con varias etiquetas (véase la Tabla 6.3). La etiqueta DIET se refería a la composición general de los alimentos para peces, indicando el nivel de inclusión en la dieta de harina y aceite de pescado. Concretamente, la etiqueta FM comprendía tres niveles porcentuales de harina de pescado (FM≤10; 10<FM≤20; FM>20), mientras que la etiqueta FO sólo se refería a dos niveles porcentuales (FO≤4; 4<FO<12). La etiqueta TISSUE definía la porción intestinal objetivo (AI = Anterior; PI = Posterior). Por último, la etiqueta ADDITIVE indicaba la existencia de un aditivo o de un sustitutivo proteico comercial (NO_ADDS = sin aditivos; SANA = con SANACORE®GM; EWH = con hidrolizado de clara de huevo; LSAQUA = LSAqua SusPro®).

Tabla 6.3. Resumen de todas las combinaciones posibles y no solapantes de variables experimentales.

Experimentos	DIET	FM	FO	TISSUE	ADDITIVE
GAIN_PRE	CTRL1	10<FM≤20	≤4	PI	NO_ADDS
GAIN_PRE	NOPAP	≤10	4<FO<12	PI	SANA
EGGHYDRO	CTRL2	>20	4<FO<12	AI	NO_ADDS
EGGHYDRO	EGG	≤10	≤4	AI	EWH
EGGHYDRO	LOWFMFO	10<FM≤20	≤4	AI	NO_ADDS
LSAQUA	CTRL3	10<FM≤20	4<FO<12	AI	NO_ADDS
LSAQUA	LSAQUA50	≤10	4<FO<12	AI	LSAQUA
LSAQUA	LSAQUA100	≤10	4<FO<12	AI	LSAQUA

En total, se recuperaron 844 géneros de bacterias en al menos una muestra de las tres pruebas de alimentación, pero la mayoría de los taxones solamente estaban presentes en un pequeño subconjunto de muestras y la matriz de datos resultante suele tener muchos datos cuyo valor es cero. En consecuencia, 537 taxones (63,6% del total) estaban presentes en menos de 10 muestras, y la mayoría de los taxones con muchos ceros coincidentes en las muestras pueden ser objeto de una regresión significativa, aunque puede deberse a una mala interpretación de los taxones poco abundantes. En consecuencia, la eliminación de datos cero también debe considerarse parte del procedimiento normal. Para ello, creamos y entrenamos tres modelos BN diferentes (Tabla 6.4) utilizando el *score* loglik y la distribución Log-Normal de taxones. El modelo TOTAL incluyó todos los taxones detectados (844) en todas las muestras analizadas, disminuyendo los taxones alcanzados inicialmente tras la aplicación del filtro de prevalencia implementado en SAMBA. De este modo, se englobaron un total de 170 taxones en el modelo TF25, que comprendía los taxones presentes en al menos el 25% de las muestras, y 48 taxones en el modelo TF50, que incluía únicamente los taxones presentes en al menos el 50% de las muestras.

Tabla 6.4. Resumen de los modelos BN resultantes construidos con el conjunto de datos empíricos de 844 géneros y 6 variables experimentales.

	TOTAL	**TF25**	**TF50**
Taxones	844	170	48
Var. Exp.	6	6	6
Nodos	850	176	54
Relaciones*	13.876	1.546	145
Taxones en relaciones*	844	170	48
Var. Exp. en relaciones*	6	6	6
Nodos en relaciones*	850	176	54
% Nodos en relaciones*	100	100	100

*Relaciones significativas (MI < 0.05 y BIC SCORE < 0)

Independientemente del modelo construido, la jerarquía de las variables experimentales era la misma (Figura 6.13). La variables EXPERIMENTS estaba en la parte superior de la jerarquía, e influía en las variables DIET y TISSUE. Los tres ensayos de alimentación diferían precisamente en estos dos elementos, por lo que se esperaban estos resultados. La etiqueta DIET controla las variables FM, FO y ADDITIVES, que son básicamente los componentes de los piensos.

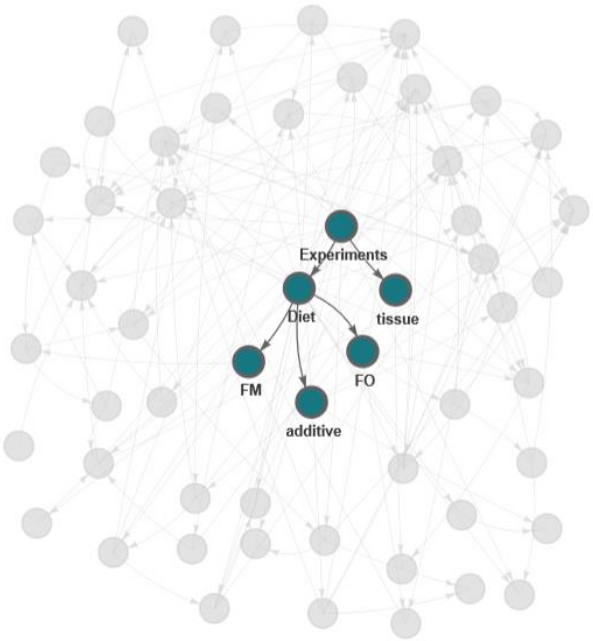

Figura 6.13. Captura de pantalla del modelo TF50 construido con SAMBA que muestra las relaciones significativas calculadas entre las variables experimentales. La misma jerarquía se calculó también en los modelo TOTAL y TF25.

Además de los resultados anteriores, todos los taxones asignados estaban implicados en al menos una relación significativa (MI < 0,05 y BIC SCORE < 0). Así, el número de relaciones significativas osciló entre 13.786 en el modelo TOTAL, 1.546 en el modelo TF25 y 145 en el modelo TF50. Este hallazgo es indicativo de que un gran número de taxones que solamente están presentes en una pocas muestras están incluidos en el conjunto de relaciones. Sin embargo, teniendo en cuenta que los taxones que faltan no son los mismos en cada condición, el primer paso en un proceso de filtrado suave y fiable es eliminar todos los taxones que tenían un valor predicho de 0 y un valor observado medio menor de 1 en todos los ensayos considerados por separado. A continuación, calculamos la regresión lineal entre los valores observados y los predichos, y nos quedamos con sus parámetros clave (p-valor y R2 ajustado) para obtener los taxones más informativos. Por último, nos centramos en los valores predichos y los dividimos en Verdaderos Positivos (VP) y Falsos Positivos (FP). Se consideró que un valor predicho era un VP cuando caía dentro del rango definido por la media ± la desviación estándar de la observación real en una condición experimental dada. Un FP se especifica cuando la predicción está fuera de ese rango o cuando el valor predicho es 0 y el valor observado es mayor que 0. La precisión de las mediciones se calculó entonces como el índice entre VP y la suma de VP y FP. Repetimos estos pasos para los tres modelos BN y los resultados se presentan en la Tabla 6.5.

Curiosamente, la regresión lineal de la información observada y la predicha mostró una fuerte significación estadística en los modelos TF25 y TF50 (p-valor < 0,001) en las tres pruebas de alimentación, y los parámetros R2 ajustados oscilaron entre 0,56-0,98 y 0,51-0,99, respectivamente. En comparación, el modelo TOTAL no superó el umbral estadístico en la condición GAIN_PRE. Ciertamente, el número de FP detectados fuera del rango de los datos observacionales fue mínimo o nulo en todos los modelos. Sin embargo, existe un número notable de taxones que se predicen como 0, pero que tienen un valor superior a 0 en los datos observacionales (15-76 en TOTAL, 31-34 en el modelo TF25 y 8-13 en TF50). La existencia de estos casos produce una ligera disminución del parámetro de precisión, aunque se mantiene dentro de valores aceptables por encima de 0,62. Según estos resultados, si el usuario introduce todos los datos, el modelo resultante puede inferir una mayor cantidad de taxones informativos y fiables, pero la significatividad de la correlación entre los valores predichos y los observados no está asegurada para algunas de las condiciones experimentales. SAMBA supera esta situación con la implementación de filtros de prevalencia que disminuyen el número de falsos positivos, certificando la correlación entre valores predichos y observados.

Tabla 6.5. Resumen de los resultados de la correlación entre los valores observados y previstos de los datos empíricos utilizando las condiciones de la variable EXPERIMENTS. Los valores P en negrita indican significación estadística (p-valor < 0,05).

Modelo	Condición	Nodos (≠ 0)	%Nodos (≠ 0)	R^2 Adj.[1]	p-valor[2]	VP[3]	FP[4]	FP_0[5]	Precisión[4]
TOTAL	LSAQUA	320	37,91	0,55	**0,011**	304	1	15	0,95
	EGGHYDRO	200	23,70	0,98	**<0,001**	123	1	76	0,62
	GAIN_PRE	234	27,73	0,41	0,060	194	0	40	0,83
TF25	LSAQUA	101	59,41	0,56	**<0,001**	67	0	34	0,66
	EGGHYDRO	89	52,35	0,98	**<0,001**	57	1	31	0,64
	GAIN_PRE	93	54,71	0,70	**<0,001**	60	0	33	0,65
TF50	LSAQUA	44	91,67	0,51	**<0,001**	30	1	13	0,68
	EGGHYDRO	41	85,42	0,99	**<0,001**	31	0	10	0,76
	GAIN_PRE	39	81,25	0,75	**<0,001**	31	0	8	0,79

[1]Valor del parámetro R2 ajustado tras la regresión lineal de los valores predichos y observados. [2]P-valor de la regresión lineal. [3]Verdaderos Positivos (VP). [4]Falsos Positivos (FP) predichos fuera de rango (Media ± Desviación estándar de los datos observacionales). [5]FP medido como 0 en la predicción con un valor mayor de 0 en los datos observacionales. [6]Precisión = VP / (VP + FP).

6.4. Discusión

La modelización de relaciones complejas entre los componentes físicos y biológicos de los sistemas acuícolas es uno de los retos futuros en el ámbito de la producción de alimentos en el contexto del cambio climático y el crecimiento de la población humana (Abberton et al., 2016; Poore and Nemecek, 2018). En este capítulo, hemos presentado SAMBA, la implementación software de un modelo BN con múltiples aplicabilidades prácticas. SAMBA puede ayudar a revelar las relaciones y dependencias dentro de un pan-microbioma, identificando taxones que constituyen el núcleo del pan-microbioma y determinando su orden de colonización. Esta información puede ser de especial interés en piscicultura, ya que la dinámica y las jerarquías de los pan-microbiomas de los peces de un determinado sistema acuícola pueden utilizarse para estudiar los efectos de las fórmulas dietéticas, los estímulos ambientales o los factores de trazabilidad, contribuyendo todos ellos a mejorar la sostenibilidad de la industria acuícola (Terova et al., 2022). En general, SAMBA es muy intuitivo y fácil de usar gracias a su interfaz gráfica de usuario, y puede ser considerado como un proyecto en continuo progreso, comprometiéndonos a implementar nuevas funciones y herramientas en futuras

versiones de SAMBA, como por ejemplo, nuevos métodos de normalización, nuevas distribuciones de datos y/o nuevas funciones para detectar el tamaño mínimo de la muestra de entrenamiento para un experimento dado, lo que puede ayudar a aumentar el poder de predicción de las redes bayesianas (ver, por ejemplo, Hu et al., 2021).

Durante su validación, los valores predichos por SAMBA teniendo en cuenta todas las relaciones establecidas en los diferentes conjuntos de datos mostraron un alto grado de similitud con los valores observados tras los procedimientos de secuenciación de amplicones y asignación taxonómica. En ausencia de una alta variabilidad biológica y de datos distintos de cero, SAMBA mostró una correlación casi completa con los conjuntos de datos sintéticos observacionales, lo que hace que esta herramienta sea muy adecuada para datos menos dispersos o con un número reducido de datos perdidos. Cuando se utilizan conjuntos de datos metagenómicos reales, el uso de los filtros de prevalencia implementados en SAMBA ayuda al usuario a reducir el impacto negativo de la dispersión y variabilidad de los taxones entre las condiciones experimentales.

En resumen, SAMBA permite inferir relaciones potenciales dentro de las entidades que constituyen una comunidad microbiana, ayudando a entender cómo estas comunidades establecen dependencias causales entre ellas en las barreras mucosas de los peces. Por último, cabe destacar que, aunque hemos probado y validado SAMBA utilizando datos de peces, la aplicación se puede utilizar para modelar cualquier tipo de relaciones de red microbioma-hospedador utilizando como caso de estudio cualquier otro organismo complejo, incluso los seres humanos.

6.5. Publicaciones peer-review relacionadas con este capítulo en esta tesis

La única publicación relacionada con este capítulo de tesis es la correspondiente a la aplicación SAMBA, donde la autora de esta tesis es primera autora al contribuir con el diseño, la algorítmica y la implementación de SAMBA, así como con la redacción del correspondiente artículo:

Soriano B, Hafez A, Naya-Català F, Moldovan RA, Toxqui-Rodríguez S, Piazzon MC, Arnau V, Llorens C, Pérez-Sánchez J. 2022. SAMBA: Structure-Learning of Aquaculture Microbiomes Using a Bayesian-Network Approach. Submitted to Genes journal. Preprint available in biorxiv: https://doi.org/10.1101/2022.12.30.522281.

7. DISCUSIÓN GENERAL

Los avances en la secuenciación de nueva generación (NGS) han cambiado la forma en que los investigadores realizan análisis comparativos basados en datos de secuenciación y resecuenciación. La implementación de estas aproximaciones en los procedimientos rutinarios de laboratorio sigue siendo un reto, ya que requieren la ejecución secuencial de protocolos complejos y variables para extraer y procesar la información biológicamente relevante a partir de los datos de secuenciación en bruto. Estos protocolos suelen denominarse *pipelines* y/o flujos de trabajo, y normalmente se llevan a cabo utilizando software ejecutado por línea de comandos (CLI). La ventaja de los *pipelines* basados en herramientas CLI es que pueden personalizarse para objetivos específicos y utilizar la amplia gama de software de libre acceso producido por la comunidad científica. Esto es particularmente útil en las aproximaciones NGS, donde los requisitos de cada *pipeline* difieren dependiendo de los datos a analizar (Conesa et al., 2016; Geraci et al., 2020; Minei et al., 2018; Jung et al., 2020; Tritt et al., 2012). Por ejemplo, los *pipelines* RNA-seq varían en función de la disponibilidad de archivos GTF/GFF (formato de archivo que proporciona información sobre las características de los genes de una secuencia de referencia) y de la secuencia de referencia (puede ser un genoma, un transcriptoma, un panel de genes, etc.). Otra ventaja de los protocolos basados en herramientas CLI es que funcionan tanto en ordenadores personales (PC) como en servidores remotos. Esto permite la gestión y el análisis simultáneo de múltiples muestras, una práctica típica en los enfoques NGS. Las desventajas de los *pipelines* basados en herramientas CLI es que su implementación y uso solamente puede lograrse en entornos Linux/Unix y requiere conocimientos informáticos avanzados para instalar software de terceros, escribir *scripts* y ejecutar procesos con la línea de comandos. En otras palabras, estos protocolos están restringidos a bioinformáticos experimentados.

En los últimos años se han desarrollado distintas soluciones de interfaz gráfica de usuario (GUI) para proporcionar a los usuarios finales herramientas fáciles de usar para analizar de forma autónoma sus datos NGS. Muchas de ellas son plataformas comerciales distribuidas bajo licencias de pago, mientras que otras son herramientas gratuitas o de acceso público. Las plataformas comerciales que ofrecen funciones para el análisis de datos NGS suelen ser plataformas cruzadas de escritorio que se ejecutan en cualquier PC y sistema operativo (SO), pero también pueden ser plataformas web proveedoras de servicios en la nube. Algunos usuarios consideran que estas plataformas comerciales merecen la pena porque: I) presentan herramientas propietarias específicas; II) su gestión solo requiere conocimientos informáticos a nivel usuario; III) se implementan bajo marcos Java o C++ muy eficientes y seguros. Ejemplos de soluciones comerciales populares con herramientas para el análisis de datos NGS son Qiagen (CLC) OmicSoft (https://digitalinsights.qiagen.com), Geneious (http://www.geneious.com), Partek (https://www.partek.com/partek-genomics-suite), y OmicsBox (https://www.biobam.com/omicsbox). A pesar de sus ventajas, los paquetes comerciales siguen estando limitados en comparación con los *pipelines* basados en herramientas CLI que se pueden actualizar y mejorar más fácilmente con nuevas funciones y herramientas que los paquetes comerciales. Por esta razón, existe

una tendencia creciente entre las empresas bioinformáticas a habilitar *plugins* para software CLI de terceros en sus plataformas.

Por otro lado, la mayoría de las soluciones GUI gratuitas o disponibles públicamente son plataformas impulsadas por software CLI. Estas plataformas suelen implementarse como soluciones del lado del cliente y del lado del servidor, donde el componente del lado del cliente es la aplicación de escritorio y/o web que permite ejecutar las herramientas CLI a través de la GUI, y el componente del lado del servidor es la plataforma que aloja las herramientas CLI y que proporciona los entornos Linux y R que permiten ejecutarlas. Dos ejemplos representativos de soluciones públicas ya ensambladas del lado del cliente y del lado del servidor son Unipro Ugene (Okonechnikov et al., 2012) y Chipster (Kallio et al., 2011). Ugene es una plataforma cruzada desarrollada en C++ con distintas herramientas para el análisis de biología molecular, incluyendo *pipelines* y componentes basados en las herramientas CLI más comunes para distintas ómicas (Golosova et al., 2014). Chipster es otra plataforma que proporciona entornos GUI para una colección de herramientas CLI para el análisis de datos de ómica distintiva. Chipster se desarrolló originalmente como una aplicación Java de escritorio, pero actualmente está disponible como servidor web. Chipster también puede instalarse en el PC del usuario aunque solamente en entornos Linux al tratarse de una aplicación cliente acoplada a un paquete que contiene todas las dependencias del lado del servidor, principalmente software CLI que solo funciona en SO Linux. Ugene y Chipster son soluciones de código abierto que permiten a sus usuarios implementar nuevas herramientas y algunos niveles de personalización para configurar flujos de trabajo específicos. Estas plataformas son, sin duda, un recurso valioso para los usuarios con conocimientos bioinformáticos avanzados, pero podrían resultar complejas de manejar y, por tanto, un reto para otros usuarios cuyos conocimientos y experiencia sean más biológicos.

También pueden implementarse de *novo* soluciones GUI del lado del cliente y de lado del servidor basados en herramientas CLI. Estos enfoques personalizados suelen implementarse como instancias web utilizando Python o R porque son lenguajes interpretados más concisos y fáciles de manejar que Java o C++. En el caso de los desarrolladores de R, la estrategia más común para este tipo de implementaciones es combinar RStudio (http://www.rstudio.com), un entorno integrado para la programación de R, con shiny (Chang et al., 2021), un paquete para implementar soluciones de servidor web basadas en herramientas de R. En cuanto a Python, los desarrolladores suelen utilizar frameworks como Bioconda (Gruning et al., 2018) y Galaxy (The Galaxy Community, 2022), una plataforma de módulos web precompilados que actúan como GUIs para herramientas CLI específicas. Los módulos Galaxy pueden combinarse de diferentes maneras para construir flujos de trabajo personalizados para el análisis ómico, siendo el repertorio de herramientas soportados por el proyecto Galaxy muy extenso. Sin embargo, la implementación de una solución con Galaxy sigue siendo enrevesada, ya que la instalación y configuración de combinaciones específicas de módulos Galaxy es una tarea compleja que requiere conocimientos avanzados de bioinformática y una amplia experiencia en sistemas informáticos. Además, Galaxy sólo funciona en entornos Linux, principalmente porque las herramientas CLI utilizadas como dependencias del lado del servidor son herramientas Linux. Sin embargo, los usuarios avanzados interesados en ejecutar un enfoque Galaxy bajo otros sistemas operativos

como Windows o MacOS tienen la opción de crear contenedores Docker (Merkel, 2014) para desplegar allí el componente del lado del servidor de Galaxy.

Es por todo ello que el objetivo principal de esta tesis ha sido diseñar y desarrollar herramientas y protocolos que permitan y faciliten el análisis de datos procedentes de secuenciación a cualquier investigador, independientemente de la experiencia que tenga en el campo de la bioinformática, para posteriormente implementarlas dentro de las distintas herramientas que componen el GPRO Suite de Biotechvana, un proyecto bioinformático que proporciona soluciones personalizadas GUI para el análisis de datos ómicos en servidores remoto o en el PC del usuario sea cual sea su sistema operativo.

Esta tesis ha sido desarrollada bajo financiación otorgada por el Ministerio de Ciencia e Innovación para el desarrollo de este Doctorado Industrial en colaboración entre diversos centros de investigación (incluyendo el Instituto de Acuicultura Torre de la Sal (IATS-CSIC), la Fundación Jiménez Díaz, el Museo Nacional de Ciencias Naturales de CSIC, el Instituto de Recursos Naturales y Agrobiología de Salamanca (IRNASA) y el Hospital General Universitario de Valencia), y Biotechvana, empresa dedicada al campo de la bioinformática. Como resultado de estas interacciones entre Biotechvana y los diferentes centros de investigación, los protocolos y herramientas generados durante esta tesis se han aplicado a diversos estudios, enfatizando su utilidad práctica dentro de la investigación con datos NGS. Esto incluye *pipelines* y herramientas para el ensamblaje de *novo* de genomas (Capítulo 1) y transcriptomas eucariotas (Capítulo 3), para el análisis RNA-seq con y sin genoma de referencia (Capítulo 3), para la identificación y cuantificación de haplotipos y variantes en muestras víricas (Capítulo 2), y para el establecimiento de relaciones entre la microbiota y diversos factores bióticos y abióticos mediante el uso de un modelo de red bayesiana (Capítulo 4). En el contexto de la investigación, todos los *pipelines* y herramientas han sido desarrollados con el objetivo de ayudar a los investigadores a entender el contexto biológico en cada uno de los estudios que se han llevado a cabo, sirviendo estos estudios, a su vez, para la validación de estos protocolos. En el contexto industrial, todos los desarrollos, excepto el presentado en el Capítulo 4, han sido incorporados a diversas aplicaciones (DeNovoSeq, RNAseq y STATools) que forman parte del producto denominado GPRO Suite perteneciente a Biotechvana, haciendo que este suite de aplicaciones aumente su valor de mercado con la incorporación de los *pipelines* desarrollados durante la realización de esta tesis. A continuación, ofreceré un análisis general de todas las soluciones desarrolladas en la tesis presentada.

7.1. *Pipeline* para el ensamblaje de *novo* y anotación de genomas eucariotas con un alto porcentaje de repeticiones en su secuencia

Las técnicas para el ensamblaje de *novo* de un genoma en la investigación genómica es una tarea fundamental que permite la creación de genomas de referencia de alta calidad para muchas especies, facilitando su uso posterior en otros estudios de investigación. Cuando hablamos de genomas eucariotas, el principal reto se da a la hora de ensamblar regiones repetitivas, ya que estas regiones, que se encuentran dispersas por la secuencia del genoma, se agrupan algorítmicamente en unas pocas secuencias debido a su similitud durante el ensamblaje. Este es el caso de *Sparus aurata* (la dorada), cuyo análisis k-mer mostró un segundo pico relativo al porcentaje de repeticiones en su secuencia del genoma. La elección de esta especie se debió a que se trata una especie de relevancia dentro de la colaboración con el IATS-CSIC dada su importancia económica, al ser la tercera especie más cultivada de Europa, y dada su capacidad para adaptarse a diferentes temperaturas y salinidades.

Con el objetivo de conseguir un genoma de alta calidad que pudiese utilizarse en estudios posteriores por parte del IATS-CSIC, se desarrolló un protocolo para la obtención de un genoma de referencia para la dorada a partir de la secuenciación combinada de lecturas cortas y largas, incluyendo su predicción y anotación de genes (codificantes y no codificantes) y moviloma, tal y como se expone en el Capítulo 1 de esta tesis. El rendimiento del ensamblaje del genoma siguiendo la estrategia propuesta, es decir, utilizando una combinación de lecturas cortas y lecturas largas durante el ensamblaje, fue realmente bueno, obteniendo finalmente un genoma con tamaño total de 1,24 Gb. El análisis k-mer no solo mostró que el genoma de la dorada era rico en repeticiones, sino que también indicó que el tamaño estimado para este genoma es de 1,59 Gb. Esto indica que el genoma obtenido a través de esta estrategia representaría el 78% del tamaño esperado para este genoma. Los 0,35 Gb faltantes pueden no encontrarse en nuestro genoma por dos motivos: 1) no se ha conseguido secuenciar ese conjunto de secuencias que faltan; 2) estas secuencias no han podido ser ensambladas con nuestro protocolo. A pesar de no haber conseguido ensamblar el genoma al completo, el tamaño conseguido es muy superior al obtenido en los dos ensamblajes anteriores realizados para esta especie (Pauletto et al. en 2018, 760 Mb; *Bioproject accession* PRJEB31901, 833 Mb), lo que indica que nuestro *pipeline* ha sido capaz de lidiar y ensamblar parte de las regiones ricas en repeticiones presentes en este genoma.

Con el propósito de conocer si estas duplicaciones del genoma de la dorada eran verdaderas, se realizó un análisis de redundancia basado en genes transcritos activamente, un análisis de sintenia y un análisis de filoma. Los tres análisis concluyeron que la gran mayoría de las duplicaciones eran reales, detectando únicamente un 1,01% como duplicaciones erróneas. De hecho, el análisis de filoma sugirió que esta gran cantidad de duplicaciones derivaron de las actividades de los elementos genéticos móviles y de la respuesta inmunitaria como procesos clave en la adaptabilidad de la especie, al observar un enriquecimiento funcional en actividades de integración de DNA, transposición y producción de inmunoglobulinas. La caracterización del moviloma, no solamente sugirió que las duplicaciones y el aumento del complejidad del genoma de la dorada es debido a los elementos genéticos móviles, sino que también, por la caracterización de los genes quiméricos, destacó la presencia de un número elevado de

receptores relacionados con el monitoreo intracelular para detectar patógenos que han escapado a la vigilancia extracelular y endosomal. Esto indica que la dorada, y en general los peces, han ido evolucionando para mejorar su capacidad para detectar amenazas en un ambiente rico en patógenos.

Por último, hacer énfasis en que el genoma de la dorada resultante del protocolo diseñado durante esta tesis ha servido como genoma de referencia en múltiples estudios de investigación (Naya-Català et al., 2021b; Aslam et al., 2020; Picard-Sánchez et al., 2020; Colás-Ruiz et al., 2022; Naya-Català et al., 2021c; Naya-Català et al., 2021d; Serna-Duque et al., 2022a; Serna-Duque et al., 2022b; Naya-Català et al., 2022b; Szczygiel et al., 2021; Riera-Ferrer et al., 2022; Naya-Català et al., 2022a).

7.2. *Pipeline* para la identificación, caracterización y cuantificación de cuasiespecies en muestras víricas

Debido a la pandemia del COVID-19 surgió la colaboración las doctoras Celia Perales y Maria Eugenia Soria de la Fundación Jiménez Díaz con el propósito de rediseñar y adaptar un *pipeline* creado previamente por Mercedes Guerrero-Murillo y Josep Gregori i Font, y basado en el paquete de R llamado QSutils (Guerrero-Murillo and Gregori, 2020), para el análisis bioinformático de muestras de pacientes infectados por el virus SARS-CoV-2 con el objetivo de detectar diferentes variantes del virus. El rediseño de este *pipeline* se propuso por dos razones: 1) el *pipeline* había sido diseñado y testado exclusivamente para muestras de pacientes infectados por Hepatitis C; 2) el *pipeline* no permitía la detección de inserciones ni deleciones al no ser capaz de manejar diferentes tamaños de secuencias y por corregir los haplotipos detectados utilizando la secuencia de referencia. El rediseño de este *pipeline* para solventar estas limitaciones dio lugar a la herramienta VQS-haplotyper, la cual se implementó dentro de la herramienta STATools del GPRO Suite y a la que se le introdujeron diferentes parámetros para su adaptación al análisis de muestras procedentes de pacientes infectados por un virus dado.

Como se ha expuesto en el Capítulo 2, a partir de muestras de 30 pacientes infectados por este virus, el uso de este *pipeline* permitió detectar y cuantificar un total de 51 mutaciones en la región S y 54 mutaciones en la región Nsp12, incluyendo deleciones, al aplicarse un filtro de abundancia del 0,5% a los haplotipos. Cuando se aplicó un filtro de abundancia del 0,1%, el número de mutaciones en la región S creció hasta las 399 y en la región Nsp12 hasta las 755, también incluyendo las deleciones.

Con el objetivo de validar el desempeño de VQS-haplotyper, los resultados obtenidos se compararon con aquellos conseguidos a través de la ejecución de la herramienta SeekDeep. De forma general, a una abundancia del 0,5%, el rendimiento de VQS-haplotyper es ligeramente inferior al de SeekDeep en cuanto a detección de cambios de nucleótido y de deleciones. Esto es debido, posiblemente, a la forma en la que cada una de las herramientas computa la abundancia de cada haplotipo detectado. Sin embargo, cuando se establece una abundancia mínima de 0,1%, el rendimiento de VQS-haplotyper es notablemente superior al de SeekDeep, detectando 155 mutaciones puntuales en la región Nsp12 y 80 en la región S adicionales a las detectadas con

SeekDeep. Cabe destacar que las mutaciones puntuales, excepto dos, que detectaba SeekDeep y no VQS-haplotyper con una abundancia del 0,5%, sí se detectan cuando el límite de abundancia baja a 0,1%, lo que apoya la idea de que este hecho se deba a una diferencia en la forma de calcular las abundancias relativas de cada haplotipo. Por el contrario, SeekDeep seguía teniendo un mejor rendimiento a la hora de detectar deleciones, identificando 11 deleciones adicionales a las identificadas con VQS-haplotyper, algunas de las cuales llegando a tener un tamaño de hasta 51 nucleótidos. Es por ello que se propone la mejora de este aspecto para las versiones futuras de este *pipeline*.

7.3. Protocolos para el análisis RNA-seq con y sin genoma de referencia

El análisis RNA-seq ha ido ganando popularidad con el tiempo al tratarse de una tecnología cada vez más asequible y permitir estudiar los cambios que se producen en la expresión de los genes entre diferentes condiciones, tratamientos o, incluso, especies. Debido a la colaboración con el IRNASA, el Museo Nacional de Ciencias Naturales de CSIC y el Hospital General Universitario de Valencia surgió la oportunidad de diseñar e implementar dos protocolos para el análisis RNA-seq y de enriquecimiento diferencial dentro de la herramienta RNASeq del GPRO Suite: el primero para datos de secuenciación de *novo*, es decir, datos de especies cuyo genoma de referencia no está disponible; y el segundo para datos procedentes de resecuenciación, es decir, de especies cuyo genoma de referencia sí está disponible. Para el primer protocolo se utilizaron, por una parte, datos procedentes de la secuenciación de dos especies de garrapatas (*Ornithodoros erraticus* y *Ornithodoros moubata*) proporcionados por el IRNASA y, por otra parte, datos de dos especies de anisakis (*Anisakis simplex s.s.* y *Anisakis pegreffii*) y sus híbridos proporcionados por el Museo Nacional de Ciencias Naturales de CSIC. Para el desarrollo del protocolo para datos de especies con genoma de referencia disponible, se utilizaron datos de secuenciación de humanos proporcionados por el Hospital General Universitario de Valencia.

En el protocolo para datos de secuenciación de *novo* fue necesario incluir un *pipeline* para el ensamblaje de *novo* y anotación de transcriptomas consenso contra los que mapear las lecturas durante el análisis RNA-seq. Este protocolo resultó ser muy eficiente al obtener unos rendimientos de mapeo medios superiores al 97,4%, 98% y 97% en todas las muestras de *O. erraticus*, *O. moubata* y anisakis, respectivamente. Estos rendimientos de mapeo tan elevados sugirieron que los transcriptomas consenso obtenidos con este *pipeline* fueron realmente completos, incluyendo la mayoría de las lecturas secuenciadas.

En el estudio con especies de anisakis y sus híbridos, se compararon las diferencias a nivel de expresión de transcritos entre ellos. Estas comparaciones dieron lugar a que 8.239 resultaran diferencialmente expresados en la comparación entre los híbridos y *A. pegreffii*, 23.549 en la comparación entre los híbridos y *A. simplex s.s.*, y 24.813 resultaron diferencialmente expresados al comparar *A. pegreffii* y *A. simplex s.s.* Estos resultados sugirieron que los patrones de expresión de los híbridos eran más similares a los obtenidos en *A. pegreffii*, de manera que esta especie tendría un mayor peso en los híbridos que *A. simplex s.s*. Este hecho también fue apoyado cuando se analizó el

enriquecimiento de términos GO, de manera que para casi todos los GOs analizados, la relación entre transcritos diferencialmente expresados y transcritos ensayados era menor en esta comparación que en las otras dos comparaciones adicionales.

En los estudios con *O. erraticus* y *O. moubata*, se realizaron tres comparaciones: 7 días tras alimentación vs sin alimentación, 14 días tras alimentación vs sin alimentación y 14 días tras alimentación vs 7 días tras alimentación. En ambos estudios, el mayor número de transcritos diferencialmente expresados se dieron a los 7 días de la alimentación, seguido de la condición 14 tras alimentación y habiendo muy pocas diferencias entre ambos tiempos tras la alimentación de la garrapata. Estos resultados indican que la mayor parte de la expresión génica diferencialmente regulada en las glándulas salivales se produjo en los primeros 7 días tras la alimentación, siendo menos importante a partir del día 7. De hecho, cuando se analizó el enriquecimiento de términos GO de la ontología función molecular de estas comparaciones en ambas especies, se evidenció que la mayoría de GOs enriquecidos estaban relacionados con el proceso de alimentación sanguínea de las garrapatas, incluyendo actividad PLA2, actividad metalopeptidasa o actividad de unión a aminas, entre otros. El análisis de enriquecimiento de rutas metabólicas en ambas especies de garrapata también apoyó este hecho, de manera que las rutas metabólicas enriquecidas a los 7 y 14 días tras la alimentación estaban relacionadas con la actividad PLA2, es decir, estaban relacionadas con la alimentación de las garrapatas.

En cuanto al estudio llevado a cabo con muestras humanas para comparar los patrones de expresión de genes de pacientes sanos y de pacientes con leucoplasia verrucosa proliferativa, se encontraron 140 genes diferencialmente expresados especialmente enriquecidos en términos GO de la ontología de proceso biológico relacionados con la modulación de la vigilancia inmunitaria y la morfogénesis de tejidos y órganos. La principal hipótesis que apoya este hecho es que, al ser los desórdenes de PVL bastante resistentes a los tratamientos y al presentar una alta recurrencia y tasa de evolución a cáncer oral (Lozovskaia and Iukova, 1979), estas lesiones modifican la inmunovigilancia como uno de los primeros pasos de la iniciación tumoral.

En resumen, ambos protocolos implementados en la herramienta RNASeq del GPRO Suite han sido validados y testados con múltiples estudios transcriptómicos utilizando diferentes secuencias de referencia y contextos experimentales (Pérez-Sánchez R et al., 2021; Oleaga et al., 2021; Llorens et al., 2021; Llorens et al., 2018), permitiendo conocer las diferencias de expresión entre las diferentes condiciones o especies, como es el caso de anisakis, obteniendo qué genes/transcritos se encontraban diferencialmente expresados entre ellas, siendo los resultados obtenidos concordantes con los esperado. Con ello, cabe la posibilidad de obtener potenciales antígenos diana para terapias o vacunas, dilucidar las relaciones entre diferentes especies, de detectar posibles biomarcadores de cáncer o de desarrollar nuevas terapias basadas en enfoques inmunoterapéuticos.

7.4. SAMBA, un modelo de red bayesiana para establecer relaciones e influencias entre taxones y variables experimentales en un determina sistema

En la actualidad, existen diferentes soluciones basadas en redes bayesianas con visualizaciones interactivas, tales como shinyBN (Chen at al., 2019), BayesianLab (Conrady and Jouffe, 2015) o BayesSuite (Michiels et al., 2021). Sin embargo, en el caso de las dos primeras, solamente deja trabajar con variables discretas y con conjuntos pequeños de datos, mientras que la tercera sí supera estas restricciones, pero está enfocada al campo de la neurociencia y no es capaz de realizar inferencias sobre variables discretas. Debido a estas limitaciones, se decidió, en colaboración con el IATS-CSIC, desarrollar e implementar la herramienta SAMBA (del inglés, *Structure-Learning of Aquaculture Microbiomes using a Bayesian-Network Approach*), la implementación software de un modelo BN, con el objetivo de investigar y/o predecir cómo los diferentes taxones y variables experimentales de un sistema dado se relacionan entre sí y se influyen mutuamente, identificando qué taxones forman el núcleo del pan-microbioma y determinando su orden de colonización. Esto es de especial importancia dentro del campo de la piscicultura, ya que puede utilizarse para conocer o estudiar los efectos de una determinada dieta, de estímulos ambientales o de factores de trazabilidad en el pan-microbiomas de los peces de un determinado sistema acuícola, contribuyendo todos ellos a mejorar la sostenibilidad de la industria acuícola. Con el fin de facilitar el uso de SAMBA por parte de cualquier usuario, se diseñó una interfaz gráfica de usuario con el paquete shiny de R, dividida en cinco módulos diferentes, haciendo que SAMBA sea una herramienta muy intuitiva y de fácil manejo. Gracias a la implementación de estos cinco módulos que componen SAMBA, esta herramienta no solamente es capaz de establecer relaciones e influencias entre las variables, sino que, por ejemplo, también permite explorar y comparar la distribución de valores de cada taxón dadas unas condiciones experimentales concretas, lo que ayuda a comprender mejor el modelo de red bayesiana generado por SAMBA.

Para conocer si el rendimiento de SAMBA era adecuado, se llevó a cabo la validación de esta herramienta comparando los valores predichos por ella contra las observaciones reales en dos conjuntos de datos, uno constituido por datos semisintéticos y otro formado por datos reales procedentes de tres experimentos diferentes (Solé-Jimenez et al., 2021; Naya-Català et al., 2021a; Piazzon et al., 2022). En los datos semisintéticos, la correlación entre ambos fue casi perfecta, haciendo que esta herramienta sea ideal cuando hay una variabilidad biológica baja o cuando hay un número reducido de datos perdidos (en este caso, valores cero). Cuando se utilizaron conjuntos de datos metagenómicos reales, se realizó la validación utilizando los datos completos (TOTAL) y utilizando el filtro de prevalencia al 25% (TF25) y al 50% (TF50). Los resultados mostraron que la precisión fue inferior que en los datos semisintéticos, aunque en todos los casos (modelo TOTAL, modelo TF25 y modelo TF50) se encontró por encima del 0,62, un valor aceptable. Con los resultados obtenidos, se deduce que, si se utilizan todos los datos, el modelo resultante puede ser más informativo, pero puede verse afectada la significatividad de la correlación entre los valores predichos y los valores observados en algunas condiciones experimentales, ya que una de las condiciones resultó no ser significativa en el modelo TOTAL. Es por ello por lo que es importante la aplicación del filtro de prevalencia implementado en SAMBA, ya que ayuda a reducir el impacto

negativo de la dispersión y variabilidad de los taxones entre las condiciones experimentales.

Cabe destacar que esta herramienta, a pesar de que solamente se ha testado con los datos metagenómicos procedentes de diferentes estudios realizados con la dorada, podría utilizarse con otras especies al tratarse de un modelo de red bayesiana que solamente tiene en cuenta los datos *input* que introduce el usuario en la herramienta. Sin embargo, en el futuro son necesarios más estudios con especies diferentes para comprobar esta hipótesis.

8. CONCLUSIONES GENERALES

1. A nivel bioinformatico, esta tesis ha contribuido a la generación de herramientas para diferentes análisis ómicos cuya implementación mediante interfaces gráficas de usuario permite que puedan ser utilizadas por cualquier usuario de manera autónoma, incluso aunque no disponga de conocimientos bioinformáticos. Con todo esto hemos podido diseñar e implementar:

    a. Un protocolo específico para el análisis de *novo* de genomas eucariotas y de transcriptomas eucariotas que se ha implementado en la aplicación DeNovoSeq del GPRO suite

    b. Un protocolo específico para el análisis RNA-seq con distintas rutas analíticas tanto para datos de *novo* como para datos de resecuenciación, que se ha implementado en la aplicación RNAseq del GPRO suite.

    c. Un protocolo *pipeline* específico denominado VQS-haplotyper para el análisis de variantes virales de muestras de SARS-CoV-2 adaptado de otro ya existente para el virus de la hepatitis C. Este *pipeline* se ha implementado en la aplicación STATools del GPRO suite.

    d. Un modelo de tipo red bayesiana cuya implementación software ha sido denominada SAMBA. Este modelo de red bayesiana permite inferir y modelar las relaciones potenciales entre los taxones de una comunidad microbiana y los factores bióticos y abióticos que intervienen en la dinámica de un determinado acuicultivo. Todo ello para descifrar la red de influencias e interacciones en estas variables unas a otras dentro del sistema. Esta red bayesiana es un modelo de inteligencia artificial que una vez entrenado permite predecir como las distintas variables pueden cambiar de estado dependiendo de la dinámica de cambios en otras variables.

2. Las herramientas bioinformáticas creadas en el marco del proyecto GPRO suite, han sido compiladas tanto en versión de escritorio como en versión *cloud*. Todo ello para permitir el uso de estas herramientas tanto en PC, bajo cualquier sistema operativo, como para su uso como herramientas de computación en la nube instaladas en servidores remotos.

3. A nivel biológico, esta tesis ha contribuido a generar nuevo conocimiento biológico reconstruyendo, anotando y/o comparando el siguiente material ómico:

    a. Un genoma de alta calidad del organismo modelo *Sparus aurata* (la dorada) reconstruido de *novo* y anotado en el marco de esta tesis. Esto incluye los genes codificantes y no codificantes de este organismo, así como también su moviloma (conjunto de elementos móviles) que constituye el 75% del tamaño del genoma. Este dato es altamente relevante porque se demuestra que el genoma de la dorada es mucho más grande de lo que originalmente se pensaba y todo apunta a que el moviloma ha sido uno de los mecanismos genómicos que ha facilitado la expansión de dicho genoma a lo largo de la

evolución. Cabe igualmente decir que la disponibilidad de este genoma ha permitido a su vez desplegar un número importante de experimentos de transcriptómica comparativa en otras tesis, donde se ha usado el genoma de la dorada como secuencia de referencia.

b. Los sialotranscriptomas (transcriptoma salivar) de dos especies de garrapatas, *Ornithodoros moubata* y *Ornithodoros erraticus*, caracterizados por vez primera bajo distintas condiciones dietéticas. Los patrones de expresión de cada especie fueron comparados bajo distintas condiciones dietéticas. En *O. moubata,* los grupos de proteínas funcionales más abundantemente sobrerrepresentados en el caso a estudio fueron las lipocalinas, las proteasas (especialmente las metaloproteasas), los inhibidores de proteasas, incluida la familia Kunitz/BPTI, las proteínas con actividad fosfolipasa A2, las proteínas de cola ácida, las proteínas de cola básica, las vitelogeninas, la familia 7DB y las proteínas implicadas en la inmunidad y defensa de las garrapatas. En *O. erraticus*, las familias de proteínas más abundantemente sobrerrepresentadas en el caso a estudio fueron las lipocalinas, las proteínas de cola ácida y básica, las proteasas (en particular las metaloproteasas), los inhibidores de proteasas, las fosfolipasas A2 secretadas, las 5'-nucleotidasas/apiasas y las proteínas similares a la vitelogenina de unión a hemo. Todas ellas están funcionalmente relacionadas con la ingestión de sangre y la regulación de las respuestas defensivas del hospedador, por lo que pueden ser interesantes antígenos protectores candidatos para vacunas.

c. Se reconstruyó de *novo* los transcriptomas de dos especies de Anisakis (*Anisakis simplex s.s.* y *Anisakis pegreffii*) y sus híbridos. Todo ello para elucidar las diferencias entre los distintos patrones de expresión de *A. simplex s.s., A. pegreffii* y en el estadío L3. En concreto, se observa fuertes efectos del origen de los progenitores, indicando que los híbridos son entidades biológicas intermedias entre sus especies parentales y, por tanto, de destacado interés en el estudio de la especiación en nematodos. Se demuestra también que ambas especies y sus híbridos comparten más genes alergénicos de lo que se creía.

d. Se caracterizó y comparó los patrones de expresión de una cohorte de pacientes humanos con leucoplasia verrucosa proliferativa respecto a un grupo de pacientes de sanos con el fin de explorar las causas que motivan la alta tasa de transformación maligna a cáncer de estas lesiones. Concretamente, lo que encontramos es que los patrones de expresión génica de los pacientes sanos y no sanos diferían en 140 genes cuya desregulación tiene un impacto funcional en el funcionamiento normal del sistema inmunitario. Este perfil de expresión inmunitaria proporciona una hipótesis plausible para explicar la transformación en carcinoma oral de células escamosas observada en 6 de los 10 casos ensayados.

e. La implementación de VQS-haplotyper ha permitido la caracterización y cuantificación de haplotipos circulando en muestras de pacientes infectados por el virus SARS-CoV-2 a una abundancia mínima del 0,05% y del 0,01%.

Concretamente hemos sido capaces de detectar un total de 105 posiciones con mutaciones al 0,05% y 1.154 posiciones mutadas al 0,01%, teniendo en cuenta las dos regiones analizadas (S y Nsp12).

f. La red bayesiana aquí denominada como SAMBA ha resultado ser una herramienta capaz de elucidar las relaciones potenciales entre los componentes de una comunidad microbiana y los factores bióticos y abióticos en acuicultivos de dorada. Todo ello con una alta correlación ello pese a la alta dispersión que se observa entre las frecuencias de los distintos taxones microbianos entre las muestras 16S obtenidas incluso de un mismo experimento. Todo ello gracias a las distintas herramientas implementadas como el filtro de prevalencia, el cual asegura que disminuyen el número de falsos positivos, certificando la correlación entre valores predichos por SAMBA y valores observados, con múltiples aplicabilidades prácticas. Con todo ello, SAMBA es una excelente herramienta que puede ayudar a revelar las relaciones y dependencias dentro de un pan-microbioma, identificando los taxones que constituyen el núcleo del pan-microbioma y determinando su orden de colonización. Esta información puede ser de especial interés en piscicultura, ya que la dinámica y las jerarquías de los pan-microbiomas de los peces de un determinado sistema acuícola pueden utilizarse para estudiar los efectos de las fórmulas dietéticas, los estímulos ambientales o los factores de trazabilidad, contribuyendo todos ellos a mejorar la sostenibilidad de la industria acuícola.

9. BIBLIOGRAFÍA

Abberton M, Batley J, Bentley A, Bryant J, Cai H, Cockram J, de Oliveira AC, Cseke LJ, Dempewolf H, De Pace C, Edwards D, Gepts P, Greenland A, Hall AE, Henry R, Hori K, Howe GT, Hughes S, Humphreys M, Lightfoot D, Marshall A, Mayes S, Nguyen HT, Ogbonnaya FC, Ortiz R, Paterson AH, Tuberosa R, Valliyodan B, Varshney RK, Yano M. 2016. Global agricultural intensification during climate change: a role for genomics. Plant Biotechnology Journal, 14:1095–1098.

Afgan E, Baker D, Batut B, van den Beek M, Bouvier D, Cech M, Chilton J, Clements D, Coraor N, Grüning BA, Guerler A, Hillman-Jackson J, Hiltemann S, Jalili V, Rasche H, Soranzo N, Goecks J, Taylor J, Nekrutenko A, Blankenberg D. 2018. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic Acids Research, 46:W537-W544.

Agol VI, Gmyl AP. 2018. Emergency Services of Viral RNAs: Repair and Remodeling. Microbiology and Molecular Biology Reviews, 82:e00067–17.

Aittokallio T, Schwikowski B. 2006. Graph-based methods for analysing networks in cell biology. Brief Bioinformatics, 7:243–255.

Al Khatib HA, Benslimane FM, Elbashir IE, Coyle PV, Al Maslamani MA, Al-Khal A, Al Thani AA, Yassine HM. 2020. Within-Host Diversity of SARS-CoV-2 in COVID-19 Patients With Variable Disease Severities. Frontiers in Cellular and Infection Microbiology, 10:575613.

Alarcón JA, Magoulas A, Georgakopoulos T, Zouros E, Alvarez MC. 2004. Genetic comparison of wild and cultivated European populations of the gilthead sea bream (Sparus aurata). Aquaculture, 230:65–80.

Almende BV, Benoit T, Titouan R. 2019. visNetwork: Network Visualization using 'vis.js' Library. R package version 2.0.9. https://CRAN.R-project.org/package=visNetwork.

Altschul SF, Gish W, Miller W, Myers EW and Lipman DJ. 1990. Basic local alignment search tool. JMolBiol 215:403-410.

Amos Package [http://sourcegorge.net/projects/amos]

Andrews S. 2010. FastQC: A Quality Control Tool High Throughput Sequence Data [Online]. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

APROMAR. 2020. Aquaculture in Spain 2021 [WWW Document]. URL. http://www.apromar.es/content/informes-anuales accessed 07.06.22.

Arkhipova IR, Batzer MA, Brosius J, Feschotter C, Moran JV, Schmitz J, Jurka J. 2012. Genomic impact of eukaryotic transposable elements. Mobile DNA, 3:19.

Arkhipova IR. 2006. Distribution and phylogeny of Penelope-like elements in eukaryotes. Systematic Biology, 55:875-85.

Aslam ML, Carraro R, Sonesson AK, Meuwissen T, Tsigenopoulos CS, Rigos G, Bargelloni L, Tzokas K. 2020. Genetic Variation, GWAS and Accuracy of Prediction for Host Resistance to Sparicotyle chrysophrii in Farmed Gilthead Sea Bream (Sparus aurata). Frontiers in Genetics, 11:594770.

Attali D. 2022. colourpicker: A Colour Picker Tool for Shiny and for Selecting Colours in Plots. R package version 1.2.0. https://CRAN.R-project.org/package=colourpicker.

Attali D. 2021. shinyjs: Easily Improve the User Experience of Your Shiny Apps in Seconds. R package version 2.1.0. https://cran.r-project.org/web/packages/shinyjs/index.html.

Attali D, von Hertzen N, Grey E. 2021. shinyscreenshot: Capture Screenshots of Entire Pages or Parts of Pages in 'Shiny'. R package version 0.2.0. https://cran.r-project.org/web/packages/shinyscreenshot/index.html.

Audicana MT, Kennedy MW. 2008. Anisakis simplex: from Obscure Infectious Worm to Inducer of Immune Hypersensitivity. Clinical Microbiology Reviews, 21:360–79.

Bailey E. 2022. shinyBS: Twitter Bootstrap Components for Shiny. R package version 0.61.1. https://CRAN.R-project.org/package=shinyBS.

Bansai AK. 2005. Bioinformatics in microbial biotechnology – a mini review. Microbial Cell Factories, 4:19.

Bao W, Kojima KK, Kohany O. 2015. Repbase Update, a database of repetitive elements in eurakyotic genomes. Mobile DNA, 6:11.

Barbera P, Kozlov AM, Czech L, Morel B, Darriba D, Flouri T, Stamatakis A. 2019. EPA-ng: Massively Parallel Evolutionary Placement of Genetic Sequences. Systematic Biology, 68:365-369.

Ben Slimen H, Guerbej H, Ben Othmen A, Ould Brahim I, Blel H, Chatti N. 2004. Genetic differentiation between populations of gilthead sea bream (Sparus aurata) along the tunisian coast. Cybium, 28:45–50.

Bengtsson H. 2021. A Unifying Framework for Parallel and Distributes Processing in R using Futures. The R Journal, 13:273-291.

Benjamini Y, Hochberg Y. 1995. Controlling the false Discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B, 57:289-300.

Boeke JD, Garfinkel DJ, Styles CA, Fink GR. 1985. Ty elements transpose through an RNA intermediate. Cell, 40:491-500.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics, 30:2114-2120.

Borgwardt K. 2011. Kernel methods in bioinformatics. In Handbook of statistical bioinformatics (eds Lu HH-S, Schlkopf B, Zhao H), Springer Handbooks of Computational Statistics, pp. 317–334. Berlin, Germany: Springer.

Bourque G, Burns KH, Gehring M, Gorbunova V, Seluanov A, Hammell M, Imbeault M, Izsvák Z, Levin HL, Macfarlan TS, Mager DL, Feschotte C. 2018. Ten things you should know about transposable elements. Genome Biology, 19:199.

Briones C, Domingo E. 2008. Minority report: hidden memory genomes in HIV-1 quasispecies and possible clinical implications. AIDS Reviews, 10:93-109.

Brown PO, Bowerman B, Varmus HE, Bishop JM. 1987. Correct integration of retroviral DNA in vitro. Cell, 49:347-56.

Calduch-Giner JA, Bermejo-Nogales A, Benedito-Palos L, Estensoro I, Ballester-Lozano G, Sitjà-Bobadilla A, Pérez-Sánchez J. 2013. Deep sequencing for de novo construction of a marine fish (*Sparus aurata*) transcriptome database with a large coverage of protein-coding transcripts. BMC Genomics, 14:178.

Casamassimi A, Federico A, Rienzo M, Esposito S, Ciccodicola A. 2017. Transcriptome profiling in human diseases: new advances and perspectives. International Journal of Molecular Sciences, 18:E1652.

Caspi R, Billington R, Ferrer L, Foerster H, Fulcher CA, Keseler IM, Kothari A, Krummenacker M, Latendresse M, Mueller LA, Ong Q, Paley S, Subhraveti P, Weaver DS, Karp PD. 2016. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Research, 44:D471-480.

Chaisson MJP, Wilson RK, Eichler EE. 2015. Genetic variation and de novo assembly of human genomes. Nature Reviews Genetics, 16:627-640.

Chang W, Cheng J, Allaire JJ, Sievert C, Schloerke B, Xie Y, Allen J, McPherson J, Dipert A, Borger B. 2021. shiny: web application framework for r. R package version 1.6.0. https://CRAN.R-project.org/package=shiny.

Chang W. 2021. shinythemes: shinythemes: Themes for Shiny. R package version 1.2.0. https://CRAN.R-project.org/package=shinythemes.

Chang W, Borges B. 2021. shinydashboard: Create Dashboards with 'Shiny'. R package version 0.7.2. https://cran.r-project.org/web/packages/shinydashboard/index.html.

Chaoui L, Kara MH, Quignard JP, Faure E, Bonhomme F. 2009. Strong genetic differentiation of the gilthead sea bream Sparus aurata (L., 1758) between the two western banks of the Mediterranean. Comptes Rendus Biologies, 332:329–335.

Chen J, Zhang R, Dong X, Lin L, Zhu Y, He J, Christiani DC, Wei Y, Chen F. 2019a. shinyBN: an online application for interactive Bayesian network inference and visualization. BMC Bioinformatics, 20:711.

Chen IA, Chu K, Palaniappan K, Pillay M, Ratner A, Huang J, Huntemann M, Varghese N, White JR, Seshadri R, Smirnova T, Kirton E, Jungbluth SP, Woyke T, Eloe-Fadrosh EA, Ivanova NN, Kyrpides NC. 2019b. IMG/M v.5.0: an integrated data management and comparative analysis system for microbial genomes and microbiomes. Nucleic Acids Research, 47:D666-D677.

Chen Y, Hao J, Jiang W, He T, Zhang X, Jiang T, Jiang R. 2013. Identifying potential cancer driver genes by genomic data integration. Scientific Reports, 3:3538.

Chumakov KM, Powers LB, Noonan KE, Roninson IB, Levenbook IS. 1991. Correlation between amount of virus with altered nucleotide sequence and the monkey test for acceptability of oral poliovirus vaccine. PNAS, 88:199–203.

Colás-Ruiz NR, Ramirez G, Courant F, Gomez E, Hampel M, Lara-Martín PA. 2022. Multi-omic approach to evaluate the response of gilt-head sea bream (Sparus aurata) exposed to the UV filter sulisobenzone. Science of The Total Environment, 803:150080.

Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, Szcześniak MW, Gaffney DJ, Elo LL, Zhang X, Mortazavi A. 2016. A survey of best practices for RNA-seq data analysis. Genome Biology, 17:13.

Conrady S, Jouffe L. 2015. Bayesian networks and BayesiaLab: a practical introduction for researchers (Vol. 9). Franklin: Bayesia USA.

Costa-Silva J, Domingues D, Lopes FM. 2017. RNA-seq differential expression analysis: An extended review and a software tool. PLoS One, 12:e0190152.

Csardi G, Nepusz T. 2006. The igraph software package for complex network research. InterJournal, Complex Systems:1695.

Cunningham F, Achuthan P, Akanni W, Allen J, Amode MR, Armean IM, Bennett R, Bhai J, Billis K, Boddu S, Cummins C, Davidson C, Dodiya KJ, Gall A, Girón CG, Gil L, Grego T, Haggerty L, Haskell E, Hourlier T, Izuogu OG, Janacek SH, Juettemann T, Kay M, Laird MR, Lavidas I, Liu Z, Loveland JE, Marugán JC, Maurel T, McMahon AC, Moore B, Morales J, Mudge JM, Nuhn M, Ogeh D, Parker A, Parton A, Patricio M, Abdul Salam AI, Schmitt BM, Schuilenburg H, Sheppard D, Sparrow H, Stapleton E, Szuba M, Taylor K, Threadgold G, Thormann A, Vullo A, Walts B, Winterbottom A, Zadissa A, Chakiachvili M, Frankish A, Hunt SE, Kostadima M, Langridge N, Martin FJ, Muffato M, Perry E, Ruffier M, Staines DM, Trevanion SJ, Aken BL, Yates AD, Zerbino DR, Flicek P. 2019. Ensembl 2019. Nucleic Acids Research, 47:D745-D751.

Cunningham F, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S, Gil L, Girón CG, Gordon L, Hourlier T, Hunt SE, Janacek SH, Johnson N, Juettemann T, Kähäri AK, Keenan S, Martin FJ, Maurel T, McLaren W, Murphy DN, Nag R, Overduin B, Parker A, Patricio M, Perry E, Pignatelli M, Riat HS, Sheppard D, Taylor K, Thormann A, Vullo A, Wilder SP, Zadissa A, Aken BL, Birney E, Harrow J, Kinsella R, Muffato M, Ruffier M, Searle SM, Spudich G, Trevanion SJ, Yates A, Zerbino DR, Flicek P. 2015. Ensembl 2015. Nucleic Acids Research, 43:D662-669.

Czech L, Barbera P, Stamatakis A. 2020. Genesis and Gappa: processing, analyzing and visualizing phylogenetic (placement) data. Bioinformatics, 36:3263-3265.

Daemen A, Gevaert O, Ojeda F, Debucquoy A, Suykens JAK, Sempoux C, Machiels J-P, Haustermans K, De Moor B. 2009. A kernel-based integration of genome-wide data for clinical decision support. Genome Medicine, 1:39.

Daemen A, Gevaert O, De Moor B. 2007. Integration of clinical and microarray data with kernel methods. In Engineering in medicine and biology society, 2007. EMBS 2007. 29th Annual Int. Conf. of the IEEE, pp. 5411–5415. Piscataway, NJ:IEEE.

David TL. 2022. optparse: Command Line Option Parser. R package version 1.7.3.

Davidson NM, Oshlack A. 2014. Corset: enabling differential gene expression analysis for de novo assembled transcriptomes. Genome Biology, 15:410.

Davis DA, Chawla NV. 2011. Exploring and exploiting disease interactions from multirelational gene and phenotype networks. PLoS ONE, 6:e22670.

De Castro MH, de Klerk D, Pienaar R, Rees DJG, Mans BJ. 2017. Sialotranscritomics of Rhipicephalus zambeziensis reveals intricate expression profiles of secretory proteins and suggest tight temporal transcriptional regulation during blood-feeding. Parasite Vectors, 10:384.

De Innocentiis S, Lesti A, Livi S, Rossi AR, Crosetti D, Sola L. 2004. Microsatellite markers reveal population structure in gilthead sea bream Sparus aurata from Atlantic Ocean and Mediterranean Sea. Fisheries Sci., 70:852-859.

Denton JF, Lugo-Martínez J, Tucker AE, Schrider DR, Warren WC, Hahn MW. 2014. Extensive error in the number of genes inferred from draft genome assemblies. PLoS Computational Biology, 10:e1003998.

Dinger ME, Pang KC, Merce TR, Mattick JS. 2008. Differentiating protein-coding and noncoding RNA: challenges and ambiguities. PLoS Computational Biology, 4:e1000176.

Doležel J, Greilhuber J, Lucretti S, Meister A, Lysák MA, Nardi L, Obermayer R. 1998. Plant Genome Size Estimation by Flow Cytometry: Inter-laboratory Comparison. Annals of Botany, 82:17-26.

Domingo E. 1999. Quasispecies. Encyclopedia of Virology, 1999:1431-1436.

Domingo E, Sheldon J, Perales C. 2012. Viral quasispecies evolution. Microbiology and Molecular Biology Reviews, 76:159-216.

Douglas GM, Maffei VJ, Zaneveld JR, Yurgel SN, Brown JR, Taylor CM, Huttenhower C, Langille MGI. 2020. PICRUSt2 for prediction of metagenome functions. Nature Biotechnology, 38:685-688.

Dowle M, Srinivasan A. 2021. data.table: Extension of 'data.frame'. R package version 1.14.2.

Eddy SR. 1996. Hidden Markov models. Current Opinion in Structural Biology, 6:361-365.

Egerton S, Culloty S, Whooley J, Stanton C, Ross RP. 2018. The Gut Microbiota of Marine Fish. Frontiers in Microbiology, 9:873.

Eigen M, Schuster P. 1977. A principle of natural self-organization. Naturwissenschaften, 64:541-565.

English AC, Richards S, Han Y, Wang M, Vee V, Qu J, Qin X, Muzny DM, Reid JG, Worley KC, Gibbs RA. 2012. Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. PLoS ONE, 7:e47768.

FAO. 2022. FishStatJ – software for fishery statistical time series. Available at: http://www.fao.org/fishery/statistics/software/fishstatj.

Faust K. 2021. Open challenges for microbial network construction and analysis. The ISME Journal, 15:3111-3118.

Fernandes R. 2020. bnviewer: Bayesian Networks Interactive Visualization and Explainable Artificial Intelligence. R package version 0.1.6. https://CRAN.R-project.org/package=bnviewer.

Feschotte C, Pritham EJ. 2007. DNA Transposons and the Evolution of Eukaryotic Genomes. Annual Review of Genetics, 41:331-368.

Finn RD, Attwood TK, Babbitt PC, Bateman A, Bork P, Bridge AJ, Chang HY, Dosztányi Z, El-Gebali S, Fraser M, Gough J, Haft D, Holliday GL, Huang H, Huang X, Letunic I, Lopez R, Lu S, Marchler-Bauer A, Mi H, Mistry J, Natale DA, Necci M, Nuka G, Orengo CA, Park Y, Pesseat S, Piovesan D, Potter SC, Rawlings ND, Redaschi N, Richardson L, Rivoire C, Sangrador-Vegas A, Sigrist C, Sillitoe I, Smithers B, Squizzato S, Sutton G, Thanki N, Thomas PD, Tosatto SC, Wu CH, Xenarios I, Yeh LS, Young SY, Mitchell AL. 2017. InterPro in 2017-beyond protein family and domain annotations. Nucleic Acids Research, 45:D190-D199.

Forbes SA, Beare D, Gunasekaran P, *et al*. 2015. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Research, 43:D805-811.

Franchini P, Sola L, Crosetti D, Milana V, Rossi AR. 2012. Low levels of population genetic structure in the gilthead sea bream, Sparus aurata, along the coast of Italy. ICES Journal of Marine Science, 69:41-50.

Friedman N. 2004. Inferring cellular networks using probabilistic graphical models. Science, 303(5659):799-805.

Fu L, Niu B, Zhu Z, Wu S, Li W. 2012. CD-HIT: accelerated for clustering the next generation sequencing data. Bioinformatics, 28:3150-3152.

Futami R, Muñoz-Pomer A, Viu JM, Dominguez-Escribá L, Covelli L, Bernet GP, Sempere JM, Moya A, Llorens C. 2011. GPRO: the professional tool for management, functional analysis and annotation of omic sequences and databases. Biotechvana Bioinformatics: 2011-SOFT3.

Gao S, Sung WK, Nagarajan, N. 2011. Opera: reconstructing optimal genomic scaffolds with high-throughput paired-end sequences. Journal of Computational Biology, 18:1681-1691.

García-Arriaza J, Ojosnegros S, Dávila M, Domingo E, Escarmis C. 2006. Dynamics of mutation and recombination in a replicating population of complementing, defective viral genomes. Journal of Molecular Biology, 360:558-572.

Gasteiger E, Gattiker A, Hoogland C, Duvaud S, Ivanyi I, Appel RD, Bairoch A. 2003. ExPASy: The proteomics server for in-depth protein knowledge and analysis. Nucleic Acids Research, 31:3784-7.

Gene Ontology Consortium. 2015. Gene Ontology Consortium: going forward. Nucleic Acids Research, 43:D1049-56.

Geneious, Dotmatics. Available online: http://www.geneious.com.

Geraci F, Saha I, Bianchini M. 2020. Editorial: RNA-Seq Analysis: Methods, Applications and Challenges. Frontiers in Genetics, 11:220.

Gevaert O, Smet FD, Timmerman D, Moreau Y, Moor BD. 2006. Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks. Bioinformatics, 22:e184-e190.

Gligorijević V, Pržulj N. 2015. Methods for biological data integration: perspectives and challenges. Journal of the Royal Society Interface, 12:20150571.

Goff L, Trapnell C, Kelley D. 2019. CummeRbund: analysis, exploration, manipulation, and visualization of Cufflinks high-throughput sequencing data. R package version: 2.26.0.

Goh K-I, Cusick ME, Valle D, Childs B, Vidal M, Barabsi A-L. 2007. The human disease network. PNAS, 104:8685-8690.

Golosova O, Henderson R, Vaskin Y, Gabrielian A, Grekhov G, Nagarajan V, Oler AJ, Quiñones M, Hurt D, Fursov M, Huyen Y. 2014. Unipro UGENE NGS pipelines and components for variant calling, RNA-seq and ChIP-seq data analyses. PeerJ, 2:e644.

Gómez-Cabrero D, Abugessaisa I, Maier D, Teschendorff A, Merkenschlager M, Gisel A, Ballestar E, Bongcam-Rudloff E, Conesa A, Tegnér J. 2014. Data integration in the era of omics: current and future challenges. BMC Systems Biology, 8:I1.

Gönen M, Alpaydın E. 2011. Multiple kernel learning algorithms. Journal of Machine Learning Research, 12:2211-2268.

Grabundzija I, Messing SA, Thomas J, Cosby RL, Bilic I, Miskey C, Gogol-Döring A, Kapitonov V, Diem T, Dalda A, Jurka J, Pritham EJ, Dyda F, Izsvák Z, Ivics Z. 2016. A Helitron transposon reconstructed from bats reveals a novel mechanism of genome shuffling in eukaryotes. Nature Communications, 7:10716.

Granjon D. 2021. shinydashboardPlus: Add More 'AdminLTE2' Components to 'shinydashboard'. R package version 2.0.3. https://cran.r-project.org/web/packages/shinydashboardPlus/index.html.

Greenblatt IM, Brink RA. 1963. Transpositions of modulator in maize into divided and undivided chromosome segments. Nature, 197:412-3.

Gregory TR. 2005. Animal Genome Size Database. Available at: http://www.genomesize.com/faq.php

Griffiths-Jones S. 2007. Annotating noncoding RNA genes. Annual Reviews of Genomics and Human Genetics, 8:279-98.

Gruning B, Dale R, Sjodin A, Chapman BA, Rowe J, Tomkins-Tinch CH, Valieris R, Koster J, Bioconda T. 2018. Bioconda: sustainable and comprehensive software distribution for the life sciences. Nature methods, 15:475-476.

Guerrero-Murillo M and Gregori i Font J. 2020. QSutils: Quasispecies Diversity. R package version 1.8.0.

Guigó R, Agarwal P, Abril JF, Burset M, Fickett JW. 2000. An Assessment of Gene Prediction Accuracy in Large DNA Sequences. Genome Research, 10(10):1631-1642.

Gupta, V. 2013. blat2gff.pl. Perl script, https://github.com/vikas0633/perl/blob/master/blat2gff.pl.

Hafez A, Soriano B, Elsayed AA, Futami R, Ceprián R, Ramos-Ruiz R, Martínez G, Roig FJ, Torres-Font MA, Naya-Català F, Calduch-Giner JA, Trilla-Fuertes L, Gámez-Pozo A, Arnau V, Sempere JM, Pérez-Sánchez J, Gabaldón T, Llorens C. 2022. Client applications and Server Side docker for management of RNASeq and/or VariantSeq workflows and pipelines of the GPRO Suite. Accepted in Genes journal. Preprint available in arXiv. https://doi.org/10.48550/arXiv.2202.07473.

Haghshenas E, Asghari H, Stoye J, Chauve C, Hach F. 2020. HASLR: Fast Hybrid Assembly of Long Reads. iScience, 23:101389.

Hanafi H, Rafii F, Hassani BDR, Kbir MA. 2019. Integration Methods for Biological Data Sources. In: Ben Ahmed, M., Boudhir, A., Younes, A. (eds) Innovations in Smart Cities Applications Edition 2. SCA 2018. Lecture Notes in Intelligent Transportation and Infrastructure. Springer, Cham.

Hannon lab. 2016. FASTX-Toolkit: FASTQ/a short-reads pre-processing tools. Available online: http://hannonlab.cshl.edu/fastx_toolkit.

Hardoon D, Szedmak S, Shawe-Taylor J. 2004. Canonical correlation analysis: an overview with application to learning methods. Neural Computation, 16:2639-2664.

Hartemink AJ. 2001. Principled computational methods for the validation discovery of genetic regulatory networks. https://dspace.mit.edu/handle/1721.1/8699.

Hathaway NJ, Parobek CM, Juliano JJ, Bailey JA. 2018. SeekDeep: single-base resolution de novo clustering for amplicon deep sequencing. Nucleic Acids Research, 46:e21.

Hearst MA, Dumais ST, Osman E, Platt J, Scholkopf B. 1998. Support vector machines. IEEE Intelligent Systems and their Applications, 13:18-28.

Henry L, Wickham H. 2021. purrr: Functional Programming Tools. R package version 0.3.5. https://cran.r-project.org/web/packages/purrr/.

Hobbs ET, Pereira T, O'Neill PK, Erill I. 2016. A Bayesian inference method for the analysis of transcriptional regulatory networks in metagenomic data. Algorithms for Molecular Biology, 11:19.

Holland JJ, editor. 1992. Genetic diversity of RNA viruses Current Topics in Microbiology and Immunology. Berlin: Springer-Verlag.

Holmes I. 2002. Transcendent elements: whole-genome transposon screens and open evolutionary questions. Genome Research, 12:1152-5.

Holzinger A, Dehmer M, Jurisica I. 2014. Knowledge Discovery and interactive Data Mining in Bioinformatics – State-of-the-Art, future challenges and research directions. BMC Bioinformatics 15:I1.

Hoogstraal H. 1985. Argasid and Nuttalliellid ticks as parasites and vectors. Advances in Parasitology, 24:135-23.

Hu J, Zou W, Wang J, Pang L. 2021. Minimum training sample size requirements for achieving high prediction accuracy with the BN model: A case study regarding seismic liquefaction. Expert Systems with Applications, 185:115702.

Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, Zhang L, Fan G, Xu J, Gu X, Cheng Z, Yu T, Xia J, Wei Y, Wu W, Xie X, Yin W, Li H, Liu M, Xiao Y, Gao H, Guo L, Xie J, Wang G, Jiang R, Gao Z, Jin Q, Wang J, Cao B. 2020. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet, 395:497-506.

Huang Y-F, Yeh H-Y, Soo V-W. 2013. Inferring drug–disease associations from integration of chemical, genomic and phenotype data using network propagation. BMC Medical Genomics, 6:1-14.

Huynen M, Gutell R, Konings D. 1997. Assessing the reliability of RNA folding usinf statistical mechanics. Journal of Molecular Biology, 267:1104-12.

Jaillon O, Aury JM, Brunet F, Petit JL, Stange-Thomann N, Mauceli E, Bouneau L, Fischer C, Ozouf-Costaz C, Bernot A, Nicaud S, Jaffe D, Fisher S, Lutfalla G, Dossat C, Segurens B, Dasilva C, Salanoubat M, Levy M, Boudet N, Castellano S, Anthouard V, Jubin C, Castelli V, Katinka M, Vacherie B, Biémont C, Skalli Z, Cattolico L, Poulain J, De Berardinis V, Cruaud C, Duprat S, Brottier P, Coutanceau JP, Gouzy J, Parra G, Lardier G, Chapple C, McKernan KJ, McEwan P, Bosak S, Kellis M, Volff JN, Guigó R, Zody MC, Mesirov J, Lindblad-Toh K, Birren B, Nusbaum C, Kahn D, Robinson-Rechavi M, Laudet V, Schachter V, Quétier F, Saurin W, Scarpelli C, Wincker P, Lander ES, Weissenbach J, Roest Crollius H. 2004. Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. Nature, 431:946-57.

Jain M. 2012. Next-generation sequencing technologies for gene expression profiling in plants. Briefings in Functional Genomics, 11:63-70.

Jangam D, Feschotte C, Betrán E. 2017. Transposable element domestication as an adaptation to evolutionary conflicts. Trends in Genetics, 33:817-831.

Janssen S, McDonald D, Gonzalez A, Navas-Molina JA, Jiang L, Xu ZZ, Winker K, Kado DM, Orwoll E, Manary M, Mirarab S, Knight R. 2018. Phylogenetic Placement of Exact Amplicon Sequences Improves Associations with Clinical Information. mSystems, 3:e00021-18.

Jolliffe I. 2005. Principal component analysis. New York, NY: Wiley Online Library.

Joyce AR, Palsson BØ. 2006. The model organism as a system: integrating omics data sets. Nature Reviews Molecular Cell Biology, 7:198-210.

Jung H, Ventura T, Chung JS, Kim WJ, Nam BH, Kong HJ, Kim YO, Jeon MS, Eyun S. 2020. Twelve quick steps for genome assembly and annotation in the classroom. PLOS Computational Biology, 16:e1008325.

Kajitani R, Toshimoto K, Noguchi H, Toyoda A, Ogura Y, Okuno M, Yabana M, Harada M, Nagayasu E, Maruyama H, Kohara Y, Fujiyama A, Hayashi T, Itoh T. 2014. Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. Genome Research, 24:1384-1395.

Kallio MA, Tuimala JT, Hupponen T, Klemelä P, Gentile M, Scheinin I, Koski M, Käki J, Korpelainen EI. 2011. Chipster: user-friendly analysis software for microarray and other high-throughput data. BMC Genomics, 12:507.

Kanehisa M, Goto S. 2000. KEGG: kyoto enciclopedia of genes and genomes. Nucleic Acids Research, 28:27-30.

Karamitros T, Papadopoulou G, Bousali M, Mexias A, Tsiodras S, Mentis A. 2020. SARS-CoV-2 exhibits intra-host genomic plasticity and low-frequency polymorphic quasispecies. Journal of Clinical Virology, 131:104585.

Keller O, Odronitz F, Stanke M, Kolimar M, Waack S. 2008. Scipio: Using protein sequences to determine the precise exon/intron structures of genes and their orthologs in closely related species. BMC Bioinformatics, 9:278.

Kent WJ. 2002. BLAT --the BLAST-like alignment tool. Genome Research, 12(4):656-664.

Khateeb D, Gabrieli T, Sofer B, Hattar A, Cordela S, Chaouat A, Spivak I, Lejbkowicz I, Almog R, Mandelboim M, Bar-On Y. 2022. SARS-CoV-2 variants with reduced infectivity and varied sensitivity to the BNT162b2 vaccine are developed during the course of infection. PLoS Pathogens, 18:e1010242.

Kher S, Dickerson J, Rawat N. 2010. Biological pathway data integration trends, techniques, issues and challenges: A survey. In: Nature and Biologically Inspired Computing (NaBIC), Second World Congress On. IEEE:177-82.

Kidwell MG. 2002. Transposable elements and the evolution of genome size in eukaryotes. Genetica, 115:49-63.

Kidwell MG, Lisch D. 1997. Transposable elements as sources of variation in animals and plants. PNAS, 94:7704-11.

Kim D, Langmead B, Salzberg SL. 2015. HISAT: a fast spliced aligner with low memory requirements. Nature methods, 12:357-360.

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biology, 14:R36.

Kleszczyńska A, Vargas-Chacoff L, Gozdowska M, Kalamarz H, Martínez-Rodríguez G, Mancera JM, Kulczykowska E. 2006. Arginine vasotocin, isotocin and melatonin responses following acclimation of gilthead sea bream (Sparus aurata) to different environmental salinities. Comp. Biochem. Physiol. A, 145:268-273.

Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Research, 27:722-736.

Kotera M, Hirakawa M, Tokimatsu T, Goto S, Kanehisa M. 2012. The KEGG databases and tools facilitating omics analysis: latest developments involving human diseases and pharmaceuticals. Methods MolBiol, 802:19-39.

Kulp D, Haussler D, Reese MG, Eeckman FH. 1996. A generalized hidden Markov model for the recognition of human genes in DNA. Proceedings International Conference on Intelligent Systems for Molecular Biology, 4:134-42.

Lanckriet GRG, De Bie T, Cristianini N, Jordan MI, Noble WS. 2004. A statistical framework for genomic data fusion. Bioinformatics, 20:2626-2635.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nature methods, 9:357:359.

Lapatas V, Stefanidakis M, Jiménez RC, Via A, Schneider MV. 2015. Data integration in biological research: an overview. Journal of Biological Research (Thessalon), 22:9.

Lauring AS, Andino R. 2010. Quasispecies theory and the behavior of RNA viruses. PLoS Pathogens, 6:e1001005.

Levin HL, Moran JV. 2011. Dynamic interactions between transposable elements and their hosts. Nature Reviews Genetics, 12:615-27.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics, 25:2078-2079.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics, 25:1754-1760.

Liu Z, Ma A, Mathé E, Merling M, Ma Q, Liu B. 2021. Network analyses in microbiome based on high-throughput multi-omics data. Briefings in Bioinformatics, 22:1639-1655.

Llorens C, Soriano B, Krupovic M, ICTV Report Consortium. 2021. ICTV Virus Taxonomy Profile: Pseudoviridae. The Journal of General Virology, 102:001563.

Llorens C, Soriano B, Trilla-Fuertes L, Bagan L, Ramos-Ruis R, Gamez-Pozo A, Peña C, Bagan JV. 2021. Immune expression profile identification in a group of proliferative verrucous leukoplakia patients: a pre-cancer niche for oral squamous cell carcinoma development. Clinical Oral Investigations, 25:2645-2657.

Llorens C, Soriano B, Krupovic M, ICTV Report Consortium. 2020. ICTV Virus Taxonomy Profile: Metaviridae. The Journal of General Virology, 101:1131-1132.

Llorens C, Arcos SC, Robertson L, Ramos R, Futami R, Soriano B, Ciordia S, Careche M, González-Muñoz M, Jiménez-Ruiz Y, Carballeda-Sangiao N, Moneo I, Albar JP, Blaxter M and Navas A. 2018.Functional insights into the infective larval stage of Anisakis simplex s.s., Anisakis pegreffii and their hybrids based on gene expression patterns. BMC Genomics, 19:592.

Llorens C, Futami R, Covelli L, Domínguez-Escribà L, Viu JM, Tamarit D, Aguilar-Rodríguez J, Vicente-Ripolles M, Fuster G, Bernet GP, Maumus F, Muñoz-Pomer A, Sempere JM, Latorre A, Moya A. 2011. The Gypsy Database (GyDB) of mobile genetic elements: release 2.0. Nucleic Acids Research, 39:D70-4.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology, 15:550.

Lowe R, Shirley N, Bleackley M, Dolan S, Shafee T. 2017. Transcriptomics technologies. PLoS Computational Biology, 13:e1005457.

Lozovskaia GS, Iukova GS. 1979. Kropotovo Biological Station of the N.K. Kol'tsov Institute of Developmental Biology of the Academy of Sciences of the USSR. Ontogenez, 10:524-528.

Luan DD, Korman MH, Jakubczak JL, Eickbush TH. 1993. Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. Cell, 72:595-605.

Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y, Tang J, Wu G, Zhang H, Shi Y, Liu Y, Yu C, Wang B, Lu Y, Han C, Cheung DW, Yiu S-M, Peng S, Xiaoqian Z, Liu G, Liao X, Li Y, Yang H, Wang J, Lam T-W, Wang J. 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience, 1:18.

Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. Science, 290:1151-1155.

Magoc T, Salzberg S. 2011. FLASH: Fast length adjustment of short reads to improve genome assemblies. Bioinformatics, 27:2957-2963.

Malik HS, Eickbush TH. 2001. Phylogenetic analysis of ribonuclease H domains suggests a late, chimeric origin of LTR retrotransposable elements and retroviruses. Genome Research, 11:1187-97.

Marçais G, Kingsford C. 2011. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics, 27:764-770.

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet journal, 17:10-12.

Martínez-González B, Soria ME, Vázquez-Sirvent L, Ferrer-Orta C, Lobo-Vega R, Mínguez P, de la Fuente L, Llorens C, Soriano B, Ramos R, Cortón M, López-Rodríguez R, García-Crespo C, Gallego I, de Ávila AI, Gómez J, Enjuanes L, Salar-Vidal L, Esteban J, Fernández-Roblas R, Gadea I, Ayuso C, Ruiz-Hornillos J, Verdaguer N, Domingo E, Perales C. 2022. SARS-CoV-2 Point Mutation and Deletion Spectra and Their Association with Different Disease Outcomes. Microbiology Spectrum, 10:e0022122.

Mathew JP, Taylor BS, Bader GD, Pyarajan S, Antoniotti M, Chinnaiyan AM, et al. 2007. From bytes to bedside: Data integration and computational biology for translational cancer research. PLoS Comput Biol, 3:12.

McClintock B. 1950. The origin and behavior of mutable loci in maize. PNAS, 36:344-355.

Mell P, Grance T. 2011. The NIST definition of cloud computing. NIST Special Publication, 800:7.

Merkel, D. 2014. Docker: lightweight Linux containers for consistent development and deployment. Linux Journal, 2014:2.

Metzker, ML. 2010. Sequencing technologies – the next generation. Nature Reviews Genetics, 11:31-46.

Michiels M, Larrañaga P, Bielza C. 2021. BayeSuites: An open web framework for massive Bayesian networks focused on neuroscience. Neurocomputing, 428:166-181.

Miller JR, Koren S, Sutton G. 2010. Assembly algorithms for next-generation sequencing data. Genomics, 95:315-327.

Min XJ, Butler G, Storms R, Tsang A. 2005. OrfPredictor: predicting protein-coding regions in EST-derived sequences. Nucleic Acid Research, 33:W677-W680.

Minio R, Hoshina R, Ogura A. 2018. De novo assembly of middle-sized genome using MinION and Illumina sequencers. BMC Genomics, 19:700.

Molidor R, Sturn A, Maurer M, Trajanoski Z. 2003. New trends in bioinformatics: from genome sequence to personalized medicine. Experimental Gerontology, 38:1031-1036.

Moreno E, Gallego I, Gregori J, Lucía-Sanz A, Soria ME, Castro V, Beach NM, Manrubia S, Quer J, Esteban JI, Rice CM, Gómez J, Gastaminza P, Domingo E, Perales C. 2017. Internal Disequilibria and Phenotypic Diversification during Replication of Hepatitis C Virus in a Noncoevolving Cellular Environment. Journal of Virology, 91:e02505-16.

Morgan M, Anders S, Lawrence M, Aboyoun P, Pagès H, Gentleman R. 2009. ShortRead: a Bioconductor package for input, quality assessment and exploration of high-throughput sequence data. Bioinformatics, 25:2607-2608.

Mostafavi S, Morris Q. 2012. Combining many interaction networks to predict gene function and analyze gene lists. Proteomics, 12:1687–1696.

Mostafavi S, Ray D, Warde-Farley D, Grouios C, Morris Q. 2008. GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. Genome Biology, 9:S4.

Muller K, Wickham H. 2022. tibble: Simple Data Frames. R package version 3.1.8. https://CRAN.R-project.org/package=tibble.

Myers G. 2014. Efficient local alignment discovery amongst noisy long reads. Brown D, Morgenstern B (Eds.), International Workshop on Algorithms in Bioinformatics, pp. 52-67. Berlin, Germany: Springer.

Nagarajan N, Pop M. 2013. Sequence assembly demystified. Nature Reviews Genetics, 14:157-167.

Napolitano F, Zhao Y, Moreira VM, Tagliaferri R, Kere J, D'Amato M, Greco D. 2013. Drug repositioning: a machine learning approach through data integration. Journal of Cheminformatics, 5:30.

Naya-Català F, Piazzon MC, Calduch-Giner JA, Sitjà-Bobadilla A, Pérez-Sánchez J. 2022a. Diet and host genetics drive the bacterial and fungal intestinal metatranscriptome of gilthead sea bream. Frontiers in Microbiology, 13:883738.

Naya-Català F, Piazzon MC, Torrecillas S, Toxqui-Rodríguez S, Calduch-Giner JÀ, Fontanillas R, Sitjà-Bobadilla A, Montero D, Pérez-Sánchez J. (2022b). Genetics and Nutrition Drive the Gut Microbiota Succession and Host-Transcriptome Interactions through the Gilthead Sea Bream (*Sparus aurata*) Production Cycle. Biology, 12:1744.

Naya-Català F, Wiggers G, Piazzon MC, López-Martínez MI, Estensoro I, Calduch-Giner JA, Martínez-Cuesta MC, Requena T, Sitjà-Bobadilla A, Miguel M, Pérez-Sánchez J. 2021a. Modulation of gilthead sea bream gut microbiota by a bioactive egg white hydrolysate: Interactions between bacteria and host lipid metabolism. Frontiers in Marine Science, 8:698484.

Naya-Català F, do Vale Pereira G, Piazzon MC, Fernandes AM, Calduch-Giner JA, Sitjà-Bobadilla A, Conceição LEC, Pérez-Sánchez J. 2021b. Cross-Talk Between Intestinal Microbiota and Host Gene Expression in Gilthead Sea Bream (Sparus aurata) Juveniles: Insights in Fish Feeds for Increased Circularity and Resource Utilization. Frontiers in Physiology, 12:748265.

Naya-Català F, Martos-Sitcha JA, de las Heras V, Simó-Mirabet P, Calduch-Giner JA, Pérez-Sánchez J. 2021c. Targeting the Mild-Hypoxia Driving Force for Metabolic and Muscle Transcriptional Reprogramming of Gilthead Sea Bream (Sparus aurata) Juveniles. Biology, 10:416.

Naya-Català F, Simó-Mirabet P, Calduch-Giner JA, Pérez-Sánchez J. 2021d. Transcriptomic profiling of Gh/Igf system reveals a prompted tissue-specific

differentiation and novel hypoxia responsive genes in gilthead sea bream. Scientific Reports, 11:16466.

Nesmelova IV, Hackett PB. 2010. DDE transposases: Structural similarity and diversity. Advanced Drug Delivery Reviews, 62:1187-95.

Neuwirth, E. 2022. RColorBrewer: ColorBrewer Palettes. R package version 1.1-3.

O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, Astashyn A, Badretdin A, Bao Y, Blinkova O, Brover V, Chetvernin V, Choi J, Cox E, Ermolaeva O, Farrell CM, Goldfarb T, Gupta T, Haft D, Hatcher E, Hlavina W, Joardar VS, Kodali VK, Li W, Maglott D, Masterson P, McGarvey KM, Murphy MR, O'Neill K, Pujar S, Rangwala SH, Rausch D, Riddick LD, Schoch C, Shkeda A, Storz SS, Sun H, Thibaud-Nissen F, Tolstoy I, Tully RE, Vatsan AR, Wallin C, Webb D, Wu W, Landrum MJ, Kimchi A, Tatusova T, DiCuccio M, Kitts P, Murphy TD, Pruitt KD. 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Research, 44:D733-45.

Okonechnikov K, Golosova O, Fursov M. 2012. UGENE team. Unipro UGENE: a unified bioinformatics toolkit. Bioinformatics, 28:1166-7.

Oleaga A, Soriano B, Llorens C, Pérez-Sánchez R. 2021. Sialotranscriptomics of the argasid tick Ornithodoros moubata along the trophogonic cycle. PLoS Neglected Tropical Diseases, 15:e0009105.

OmicsBox, Biobam SL. Available online: https://www.biobam.com/omicsbox.

OPATHY Consortium (including Hafez A, and Llorens C), Gabaldon T. 2019. Recent trends in molecular diagnostics of yeast infections: from PCR to NGS. FEMS microbiology reviews, 43:517–547.

Ortega A. 2008. Cultivo de dorada (Sparus aurata). In: Cuadernos de acuicultura (Espinosa de los Monteros, J ed. Fundación Observatorio Español de Acuicultura: Madrid.

Ozen A, Gonen M, Alpaydin E, Haliloglu T. 2009. Machine learning integration for predicting the effect of single amino acid substitutions on protein stability. BMC Structural Biology, 9:66.

Park ST, Kim J. 2016. Trends in Next-Generation Sequencing and a New Era for Whole Genome Sequencing. International Neurourology Journal, 20:S76-83.

Partek Genomic Suite Version 7, Partek Inc. Available online: https://www.partek.com/partek-genomics-suite.

Pauletto M, Manousaki T, Ferraresso S, Babbucci M, Tsakogiannis A, Louro B, Vitulo N, Quoc VH, Carraro R, Bertotto D, Franch R, Maroso F, Aslam ML, Sonesson AK, Simionati B, Malacrida G, Cestaro A, Caberlotto S, Sarropoulou E, Mylonas CC, Power DM, Patarnello T, Canario AVM, Tsigenopoulos C, Bargelloni L. 2018. Genomic analysis of Sparus aurata reveals the evolutionary dynamics of sex-biased genes in a sequential hermaphrodite fish. Communications Biology, 1:119.

Pavlidis P, Cai J, Weston J, Noble WS. 2002. Learning gene functional classifications from multiple data types. Journal of Computational Biology, 9:401–411.

Pedersen JS, Bejerano G, Siepel A, Rosenbloom K, Lindblad-Toh K, Lander ES, Kent J, Miller W, Haussler D. 2006. Identification and Classification of Conserved RNA Secondary Structure in the Human Genome. PLoS Computational Biology, 2:e33.

Perales C. 2020. Quasispecies dynamics and clinical significance of hepatitis C virus (HCV) antiviral resistance. International Journal of Antimicrobial Agents, 56:105562.

Pérez-Sánchez R, Carnero-Morán A, Soriano B, Llorens C, Oleaga A. 2021. RNA-seq analysis and gene expression dynamics in the salivary glands of the argasid tick Ornithodoros erraticus along the trophogonic cycle. Parasites & Vectors, 14:170.

Pérez-Sánchez J, Naya-Català F, Soriano B, Piazzon MC, Hafez A, Gabaldón T, Llorens C, Sitjà-Bobadilla A, Calduch-Giner JA. 2019. Genome Sequencing and Transcriptome Analysis Reveal Recent Species-Specific Gene Duplications in the Plastic Gilthead Sea Bream (Sparus aurata). Frontiers in Marine Science, 6:760.

Peruzzi S, Chatain B, Menu B. 2005. Flow cytometric determination of genome size in European sea bass (Dicentrarchus labrax), gilthead seabream (Sparus aurata), thinlip mullet (Liza ramada) and European eel (Anguilla Anguilla). Aquatic Living Resources, 18:77-81.

Piazzon MC, Naya-Català F, Perera E, Palenzuela O, Sitjà-Bobadilla A, Pérez-Sánchez J. 2020. Genetic selection for growth drives differences in intestinal microbiota composition and parasite disease resistance in gilthead sea bream. Microbiome, 8:168.

Piazzon MC, Naya-Català F, Pereira GV, Estensoro I, Del Pozo R, Calduch-Giner JA, Nuez-Ortín WG, Palenzuela O, Sitjà-Bobadilla A, Diaz J, Conceição LEC, Pérez-Sánchez J. 2022. A novel fish meal-free diet formulation supports proper growth and does not impair intestinal parasite susceptibility in gilthead sea bream (Sparus aurata) with a reshape of gut microbiota and tissue-specific gene expression patterns. Aquaculture, 558:738362.

Picard-Sánchez A, Estensoro I, Perdiguero P, del Pozo R, Tafalla C, Piazzon MC, Sitjà-Bobadilla A. 2020. Passive Immunization Delays Disease Outcome in Gilthead Sea Bream Infected With Enteromyxum leei (Myxozoa), Despite the Moderate Changes in IgM and IgT Repertoire. Frontiers in Immunology, 11:581361.

Placzek S, Schomburg I, Chang A, Jeske L, Ulbrich M, Tillack J, Schomburg D. 2017. BRENDA in 2017: new perspectives and new tools in BRENDA. Nucleic Acids Research, 45:D380-D388.

Poore J, Nemecek T. 2018. Reducing Food's Environmental Impacts Through Producers and Consumers. Science, 360:987–992.

Potter SC, Luciani A, Eddy SR, Park Y, Lopez R, Finn RD. 2018. HMMER web server: 2018 update. Nucleic Acids Research, 46:W200-W204.

Pray L. 2008. Transposons: The jumping genes. Nature Education, 1:204.

QIAGEN CLC Genomics Workbench. Available online: https://digitalinsights.qiagen.com.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics, 26:841-842.

R Core Team. 2022. R: A Language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Rabiner LR, Juang BH. 1986. An introduction to hidden Markov models. IEEE ASSP Magazine, 3:4-16.

Rider AK, Chawla NV, Emrich SJ. 2013. A survey of current integrative network algorithms for systems biology, pp. 479–495. Amsterdam, The Netherlands: Springer.

Riera-Ferrer E, Piazzon MC, del Pozo R, Palenzuela O, Estensoro I, Sitjà-Bobadilla A. 2022. A bloody interaction: plasma proteomics reveals gilthead sea bream (Sparus aurata) impairment caused by Sparicotyle chrysophrii. Parasites & Vectors, 15:322.

Rivas E, Eddy SR. 2000. Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. Bioinformatics, 16:583-605.

Robinson MD, McCarthy DJ, Smyth GK. 2010. EdgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics, 26:139-140.

RStudio Team. RStudio: Integrated Development for R. RStudio, PBC, Boston, MA. Available online: http://www.rstudio.com.

Rubin GM, Kidwell MG, Bingham PM. 1982. The molecular basis of P-M hybrid dysgenesis: the nature of induced mutations. Cell, 29:987-94.

Sachs K, Perez O, Pe'er D, Lauffenburger DA, Nolan GP. 2005. Causal protein signaling networks derived from multiparameter single-cell data. Science, 308:523-529.

Sazal M, Stebliankin V, Mathee K, Yoo C, Narasimhan G. 2021. Causal effects in microbiomes using interventional calculus. Scientific Reports, 11:5724.

Sazal M, Mathee K, Ruiz-Perez D, Cickovski T, Narasimhan G. 2020. Inferring directional relationships in microbial communities using signed Bayesian networks. BMC Genomics, 21:663.

Schadt E, Friend S, Shaywitz D. 2009. A network view of disease and compound screening. Nature Reviews Drug Discovery, 8:286–295.

Schmieder R, Edwards R. 2011. Quality control and preprocessing of metagenomic datasets. Bioinformatics, 27:863-864.

Schölkopf B, Tsuda K, Vert J-P. 2004. Kernel methods in computational biology. Cambridge, MA: MIT Press.

Schostak N, Pyatkov K, Zelentsova E, Arkhipova I, Shagin D, Shagina I, Mudrik E, Blintsov A, Clark I, Finnegan DJ, Evgen'ev M. 2008. Molecular dissection of Penelope transposable element regulatory machinery. Nucleic Acids Research, 36:2522-9.

Schulz MH, Zerbino DR, Vingron M, Birney E. Oases: robust de novo RNA-seq assembly across dynamic range of expression levels. Bioinformatics, 28:1086-1092.

Scutari M, Graafland CE, Gutiérrez JM. 2019. Who learns better Bayesian network structures: Accuracy and speed of structure learning algorithms. International Journal of Approximate Reasoning, 115:235-253.

Scutari, M. 2010. Learning Bayesian Networks with the bnlearn R Package. Journal of Statistical Software, 35:1-22.

Scutari, M. 2009. Structure variability in Bayesian networks. Available in: https://doi.org/10.48550/arXiv.0909.1685.

Selman B, Gomes CP. 2006. Hill-climbing Search. In, Nadel, L. (Ed), *Encyclopedia of Cognitive Science*. Wiley, New York.

Serna-Duque JA, Cuesta A, Esteban MA. 2022a. Massive gene expansion of hepcidin, a host defense peptide, in gilthead seabream (Sparus aurata). Fish & Shellfish Immunology, 124:563-5671.

Serna-Duque JA, Cuesta A, Sánchez-Ferrer A, Esteban MA. 2022b. Two duplicated piscidin genes from gilthead seabream (Sparus aurata) with different roles in vitro and in vivo. Fish & Shellfish Immunology, 127:730-739.

Shapiro SS, Wilk MB. 1965. An analysis of variance test for normality (complete samples). Biometrika, 52:591-611.

Shweiki E, Rittenhouse DW, Ochoa JE, Punja VP, Zubair MH, Baliff JP. 2014. Acute Small-Bowel Obstruction From Intestinal Anisakiasis After the Ingestion of Raw Clams; Documenting a New Method of Marine-to-Human Parasitic Transmission. Open Forum Infectious Diseases, 1:ofu087.

Siefert JL. 2009. Defining the mobilome. Methods in Molecular Biology, 532:13-17.

Slater GS, Birney E. 2005. Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics, 6:31.

Smedley D, Haider S, Ballester B, Holland R, London D, Thorisson G, Kasprzyk A. 2009. BioMart – biological queries made easy. BMC Genomics, 10:22.

Smit AFA, Hubley R. 2015. RepeatModeler Open-1.0. http://www.repeatmasker.org.

Smit AFA, Hubley R, Green P. 2015. RepeatMasker Open-4.0. http://www.repeatmasker.org.

Solé-Jiménez P, Naya-Català F, Piazzon MC, Estensoro I, Calduch-Giner JA, Sitjà-Bobadilla A, Van Mullem D, Pérez-Sánchez J. 2021. Reshaping of Gut Microbiota in Gilthead Sea

Bream Fed Microbial and Processed Animal Proteins as the Main Dietary Protein Source. Frontiers in Marine Science, 8:705041.

Soriano B, Hafez A, Naya-Català F, Moldovan RA, Toxqui-Rodríguez S, Piazzon MC, Arnau V, Llorens C, Pérez-Sánchez J. 2022. SAMBA: Structure-Learning of Aquaculture Microbiomes Using a Bayesian-Network Approach. Submitted to Genes journal. Preprint available in biorxiv: https://doi.org/10.1101/2022.12.30.522281.

Soriano B, Kuprovic M, Llorens C. 2021. ICTV Virus Taxonomy Profile: Belpaoviridae 2021. Journal of General Virology, 102:001688.

Sripathi VR, Anche VC, Gossett ZB, Walker LT. 2021. Recent applications of RNA sequencing in food and agriculture. In (Ed.), Applications of RNA-seq in Biology and Medicine. IntechOpen.

Stanke M, Diekhans M, Baertsch R, Haussler D. 2008. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics 24:637-644.

Sun H, Ding J, Piednoël M, Schneeberger K. 2018. findGSE: estimating genome size variantion within human and Arabidopsis using k-mer frequencies. Bioinformatics, 34:550-557.

Sun K, Buchan N, Larminie C, Pržulj N. 2014. The integrated disease network. Integrative Biology, 6:1069–1079.

Szczygiel J, Kamińska-Gibas T, Petit J, Jurecka P, Wiegertjes G, Irnazarow I. 2021. Re-evaluation of common carp (Cyprinus carpio L.) housekeeping genes for gene expression studies – considering duplicated genes. Fish & Shellfish Immunology, 115:58-69.

Tatusov RL, Federova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA. 2003. The COG database: an updated version includes eukaryotes. BMC Bioinformatics, 4:41.

Terova G, Naya-Català F, Rimoldi S, Piazzon MC, Torrecillas S, Toxqui MS, Fontanillas R, Calduch-Giner J, Hostins B, Sitjà-Bobadilla A, Montero D, Pérez-Sánchez J. 2022. Highlights from gut microbiota survey in farmed fish - European sea bass and gilthead sea bream case studies. Aquaculture Europe, 47:5–10.

Textor J, van der Zander B, Gilthorpe MS, Liskiewicz M, Ellison GT. 2016. Robust causal inference using directed acyclic graphs: the R package 'dagitty'. International Journal of Epidemiology, 45:1887-1894.

The Galaxy Community. 2022. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update. Nucleic Acids Research, 50:W345–W351.

The UniProt Consortium. 2021.UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Research, 49:D480-D489.

Tine M, Kuhl H, Gagnaire PA, Louro B, Desmarais E, Martins RS, Hecht J, Knaust F, Belkhir K, Klages S, Dieterich R, Stueber K, Piferrer F, Guinand B, Bierne N, Volckaert FA, Bargelloni L, Power DM, Bonhomme F, Canario AV, Reinhardt R. 2014. European sea bass genome and its variation provide insights into adaptation to euryhalinity and speciation. Nature Communications, 5:5770.

Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L. 2012. Differential Gene and Transcript Expression Analysis of RNA-seq Experiments with TopHat and Cufflinks. Nature Protocols, 7:562-578.

Tritt A, Eisen JA, Facciotti MT, Darling AE. 2012. An integrated pipeline for de novo assembly of microbial genomes. PLoS One, 7:e42304.

Troyanskaya OG, Dolinski K, Owen AB, Altman RB, Botstein D. 2003. A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae). PNAS; 100:8348–8353.

Vaidyanathan R, Xie Y, Allaire JJ, Cheng J, Sievert C, Russell K. 2022. htmlwidgets: HTML Widgets for R. R package version 1.6.0. https://cran.r-project.org/web/packages/htmlwidgets/index.html.

Van Thiel PH, Kuipers FC, Roskman RT. 1960. A nematode parasitic to herring causing acute abdominal syndromes in man. Tropical and geographical medicine, 12:97-113.

van Vliet MH, Horlings HM, van de Vijver MJ, Reinders MJ, Wessels LF. 2012. Integration of clinical and gene expression data has a synergetic effect on predicting breast cancer outcome. PLoS ONE, 7:e40358.

Vial L. 2009. Biological and ecological characteristics of soft ticks (Ixodida: Argasidae) and their impact for predicting tick and associated disease distribution. Parasite, 16:191-202.

Vidal M, Cusick ME, Barabási A-L. 2011. Interactome networks and human disease. Cell, 144:986–998.

Walve R, Salmela L, Mäkinen V. 2017. Variant genotyping with gap filling. PLoS ONE, 12:e0184608.

Wang X, Xing EP, Schaid DJ. 2014. Kernel methods for large-scale genomic data analysis. Brief. Bioinformatics, 16:183–192.

Wang Z, Gerstein M, Snyder M. 2009. RNA-seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics, 10:57-63.

Wang Z, Chen Y, Li Y. 2004. A Brief Review of Computational Gene Prediction Methods. Genomics Proteomics Bioinformatics, 2:216-221.

Warren WC, García-Pérez R, Xu S, Lampert KP, Chalopin D, Stöck M, Loewe L, Lu Y, Kuderna L, Minx P, Montague MJ, Tomlinson C, Hillier LW, Murphy DN, Wang J, Wang Z, Garcia CM, Thomas GCW, Volff JN, Farias F, Aken B, Walter RB, Pruitt KD, Marques-Bonet

T, Hahn MW, Kneitz S, Lynch M, Schartl M. 2018. Clonal polymorphism and high heterozygosity in the celibate genome of the Amazon molly. Nature Ecology and Evolution, 2:669-679.

Washietl S, Pedersen JS, Korbel JO, Stocsits C, Gruber AR, Hackermüller J, Hertel J, Lindemeyer M, Reiche K, Tanzer A, Ucla C, Wyss C, Antonarakis SE, Denoeud F, Lagarde J, Drenkow J, Kapranov P, Gingeras TR, Guigó R, Snyder M, Gerstein MB, Reymond A, Hofacker IL, Stadler PF. 2007. Structured RNAs in the ENCODE selected regions of the human genome. Genome Research, 17:852-864.

Washietl S, Hofacker IL, Stadler PF. 2005. Fast and reliable prediction of noncoding RNAs. PNAS, 102:2454-2459.

Wickham, H. 2022. stringr: Simple, Consistent Wrappers for Common String Operations. R package version 1.5.0. https://cran.r-project.org/web/packages/stringr/index.html.

Wu TD, Reeder J, Lawrence M, Becker G, Brauer MJ. 2016. GMAP and GSNAP for Genomic Sequence Alignment: Enhancements to Speed, Accuracy, and Functionality. Methods in Molecular Biology, 1418:283-334.

Xue W, Li JT, Zhu YP, Hou GY, Kong XF, Kuang YY, Sun XW. 2013. L-RNA_scaffolder: scaffolding genomes with transcripts. BMC Genomics, 14:604.

Yihui X, Cheng J, Tan X. 2022. DT: A Wrapper of the JavaScript Library 'DataTables'. R package version 0.26 https://cran.r-project.org/web/packages/DT/index.html.

Young MD, Wakefield MJ, Smyth GK, Oshlack A. Gene ontology analysis for RNA-seq: accounting for selection bias. Genome Biology, 2010:R14.

Yuan Z, Liu S, Zhou T, Tian C, Bao L, Dunham R, Liu Z. 2018. Comparative genome analysis of 52 fish species suggests differential associations of repetitive elements with their living aquatic environments. BMC Genomics, 19:141.

Yuniarti I, Glenk K, McVittie A, Nomosatryo S, Triwisesa E, Suryono T, Santoso AB, Ridwansyah I. 2021. An application of Bayesian Belief Networks to assess management scenarios for aquaculture in a complex tropical lake system in Indonesia. PLoS One, 16:e0250365.

Zeileis A, Kleiber C, Jackman S. 2008. Regression Models for Count Data in R. Journal of Statistical Software, 27:1-25.

Zhang Z, Bajic VB, Yu J, Cheung K, Townsend JP. 2011a. Data Integration in Bioinformatics: Current Efforts and Challenges. In (Ed.), Bioinformatics – Trend and Methodologies. IntechOpen.

Zhang J, Haider S, Baran J, Cros A, Guberman JM, Hsu J, Liang Y, Yao L, Kasprzyk A. 2011b. BioMart: a data federation framework for large collaborative projects. Database (Oxford), 2011:bar038.

Žitnik M, Župan B. 2015. Data fusion by matrix factorization. IEEE Transactions on Pattern Analysis and Machine Intelligence, 37:41–53.

Zohar Y, Abraham M, Gordin H. 1978. The gonadal cycle of the captivity-reared hermaphroditic teleost Sparus aurata (L.) during the first two years of life. Ann. Biol. Anim. Biochem. Biophys., 18:877-882.

APÉNDICE A: MATERIAL SUPLEMENTARIO

Capítulo 3. Ensamblaje de novo del genoma de *Sparus aurata*, predicción de genes y su anotación.

- Tabla suplementaria S3.1.
  https://biotechvana.com/files/supplementary_material/Tabla_S3.1.xlsx

Capítulo 4. *Pipeline* para la detección y cuantificación de cuasiespecies del virus SARS-CoV-2.

- Tabla suplementaria S4.1.
  https://biotechvana.com/files/supplementary_material/Tabla_S4.1.xlsx
- Tabla suplementaria S4.2.
  https://biotechvana.com/files/supplementary_material/Tabla_S4.2.xlsx

Capítulo 5. Diseño de protocolos para estudios de expresión diferencial y transcriptómica comparativa usando datos de RNA-seq con y sin genoma de referencia.

- Tabla suplementaria S5.1.
  https://biotechvana.com/files/supplementary_material/Tabla_S5.1.xlsx
- Tabla suplementaria S5.2.
  https://biotechvana.com/files/supplementary_material/Tabla_S5.2.xlsx
- Tabla suplementaria S5.3.
  https://biotechvana.com/files/supplementary_material/Tabla_S5.3.xlsx
- Tabla suplementaria S5.4.
  https://biotechvana.com/files/supplementary_material/Tabla_S5.4.xlsx
- Tabla suplementaria S5.5.
  https://biotechvana.com/files/supplementary_material/Tabla_S5.5.xlsx
- Tabla suplementaria S5.6.
  https://biotechvana.com/files/supplementary_material/Tabla_S5.6.xlsx
- Tabla suplementaria S5.7.
  https://biotechvana.com/files/supplementary_material/Tabla_S5.7.xlsx
- Tabla suplementaria S5.8.
  https://biotechvana.com/files/supplementary_material/Tabla_S5.8.xlsx
- Tabla suplementaria S5.9.
  https://biotechvana.com/files/supplementary_material/Tabla_S5.9.xlsx
- Tabla suplementaria S5.10.
  https://biotechvana.com/files/supplementary_material/Tabla_S5.10.xlsx

- Tabla suplementaria S5.11.
  https://biotechvana.com/files/supplementary_material/Tabla_S5.11.xlsx
- Tabla suplementaria S5.12.
  https://biotechvana.com/files/supplementary_material/Tabla_S5.12.xlsx
- Tabla suplementaria S5.13.
  https://biotechvana.com/files/supplementary_material/Tabla_S5.13.xlsx
- Tabla suplementaria S5.14.
  https://biotechvana.com/files/supplementary_material/Tabla_S5.14.xlsx
- Tabla suplementaria S5.15.
  https://biotechvana.com/files/supplementary_material/Tabla_S5.15.xlsx
- Tabla suplementaria S5.16.
  https://biotechvana.com/files/supplementary_material/Tabla_S5.16.xlsx

APÉNDICE B: RECURSOS ONLINE

- **RNASeq**
  - https://gpro.biotechvana.com/download/RNAseq
  - https://gpro.biotechvana.com/tool/RNAseq/manual
- **DeNovoSeq**
  - https://gpro.biotechvana.com/download/DeNovoSeq
  - https://gpro.biotechvana.com/tool/DeNovoSeq/manual/
- **STATools**
  - https://gpro.biotechvana.com/download/STATools
  - https://gpro.biotechvana.com/tool/STATools/manual/
- **VQS-haplotyper**
  - https://github.com/biotechvana/VQS-haplotyper
- **SAMBA**
  - https://github.com/biotechvana/SAMBA
- **GPRO scripts**
  - https://github.com/biotechvana/GPRO-scripts

APÉNDICE C: PATENTES Y PROPIEDAD INTELECTUAL

Título: GPRO Suite. **Autores**: AHMED HAFEZ, BEATRIZ SORIANO, CARLOS LLORENS. **Solicitud NO**: V-959-20. **NÚMERO DE ENTRADA**: 09/2023/9. **Fecha de prioridad**: 23-11-2020. **Entidad**: BIOTECH VANA S.L. **País**: España. **Compañía exploradora IT**: BIOTECH VANA S.L (software).