UNIVERSIDAD COMPLUTENSE DE MADRID
FACULTAD DE FARMACIA
Departamento de Microbiología y Parasitología

TESIS DOCTORAL

Bioinformática funcional y su aplicación en genómica, proteogenómica y reposicionamiento de fármacos

MEMORIA PARA OPTAR AL GRADO DE DOCTOR

PRESENTADA POR

Mónica Franch Sarto

Director

Alberto Pascual Montano

Madrid
Ed. electrónica 2019

UNIVERSIDAD COMPLUTENSE DE MADRID

FACULTAD DE FARMACIA

DEPARTAMENTO DE MICROBIOLOGÍA Y PARASITOLOGÍA

TESIS DOCTORAL

BIOINFORMÁTICA FUNCIONAL Y SU APLICACIÓN EN GENÓMICA, PROTEOGENÓMICA Y REPOSICIONAMIENTO DE FÁRMACOS

MEMORIA PARA OPTAR AL GRADO DE DOCTORA

PRESENTADA POR:

Mònica Franch Sarto

DIRECTOR:

Alberto Pascual Montano

AGRADECIMIENTOS

Este apartado de la tesis se escribe con la satisfacción de haber concluido un apartado de la vida. Un apartado que no ha sido fácil en algunos momentos pero que ha resultado ser muy gratificante.

Me gustaría agradecer a todas las personas que lo han hecho posible. Para empezar a mis padres que siempre me ha apoyado en todas mis decisiones y siempre siempre han confiado en mí, ¡gracias a ellos estoy aquí! Al igual que mi hermana que es mi otra mitad, y sin ella la vida no sería tan bonita y divertida. Us estimo molt! Por supuesto, también quería darle las gracias als meus avis. Que aparte de su apoyo y su cariño incondicional están todos aprendiendo ciencia para entender a lo que me dedico. ¡Y ya son todos unos expertos! I al meu avi Josep que encara que ja no estigui aquí ser que estaria molt orgullós de mi.

Durante estos años he conocido a mucha gente que me ha acompañado en este recorrido. Por empezar me gustaría darle las gracias a Alberto mi director, porque es un profesional excelente y una mejor persona. Durante el tiempo que he trabajado con él me ha enseñado lo que es la pasión por el trabajo y su ejemplo y motivación han conseguido que hoy este aquí. Por supuesto también quería darle las gracias a Blanca por siempre estar a mi lado, su cariño diario y todos los buenísimos consejos que me ha dado. Habéis sido una familia para mi durante todo este tiempo, mi familia cuba-catalana.

También quería mencionar a los compañeros con los que empecé esta etapa de mi vida Dannys, Rubén, Tabas y Joao que, aunque luego cada uno haya seguido su camino han seguido a mi lado celebrando cada uno de los logros conseguidos.

Asimismo, quería agradecer a todo el laboratorio del proteómica por haber sido tan buenos compañeros en este último periodo de la tesis, Adán, Miguel, Antonio, Gema, Rosana, Alberto, Manuel, Sergio, Lola, Virginia, Mari Carmen y en especial a Fernando por permitirme formar parte de este gran grupo y volver a ilusionarme con mi trabajo y enseñarme tanto de proteómica, como también lo han hecho Concha Gil y Lucia Monteoliva.

Y en este apartado tampoco podía faltar la familia que se elige, mis amigas y amigos de Mollet y Madrid. ¡No podría haber elegido mejor!

Y como siempre lo mejor para el final, gracias Javi por toda tu paciencia apoyo y cariño. Sabes que te quiero mucho y sin ti no lo hubiera logrado.

TABLA DE CONTENIDO

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

1. INTRODUCCIÓN

1.1. Bioinformática funcional

En los últimos años, el uso de datos ómicos en biomedicina está creciendo muy rápidamente. Para explorar en detalle esta avalancha de nuevos datos producidos por las tecnologías ómicas, es indispensable la aplicación de técnicas avanzadas de análisis y cálculo computacional que permitan extraer la información biológica disponible en ellos, convirtiendo estos datos biológicos en información valiosa. Esta Tesis Doctoral se titula "Bioinformática funcional y su aplicación en Genómica, Proteogenómica y reposicionamiento de fármacos", para su realización se han desarrollado y aplicado técnicas Bioinformáticas para analizar datos ómicos y situarlos dentro de un contexto biológico para convertir firmas de genes en mecanismos biológicos o fenotipos. Los estudios realizados durante esta Tesis giran en torno a la Bioinformática funcional y por ende al Dogma Central de la Biología Molecular que trata los tres procesos fundamentales llevados a cabo por los ácidos nucleicos: replicación, transcripción y traducción (Figura 1). En la introducción de esta tesis se pretende introducir las ómicas más comunes, las tecnologías utilizadas y los problemas biológicos relevantes para la tesis, definiendo los problemas biológicos concretos en cada uno de las aportaciones de la tesis.

Los ácidos nucleicos, y el ácido desoxirribonucleico (ADN) en particular, almacenan y transmiten la información necesaria para el funcionamiento de los seres vivos. La principal característica que define la estructura del ADN es su secuencia nucleotídica formada por adenina (A), citosina (C), guanina (G) y timina (T). Estas cuando se combinan forman unidades funcionales llamadas genes. Cada gen proporciona las instrucciones para formar un producto funcional, es decir, una molécula necesaria para desempeñar un trabajo en la célula. En muchos casos, el producto funcional es una proteína y se forma a partir de la transcripción previa del ADN a ácido ribonucleico (*ARN*). Pero no todos los genes codifican proteínas, algunos proporcionan moléculas de ARN funcionales, como los *ncRNA* (ARN no codificante, del inglés *non-coding RNA*) o *lncRNA* (del inglés, *long non-coding RNA*) que desempeñan papeles en la traducción.

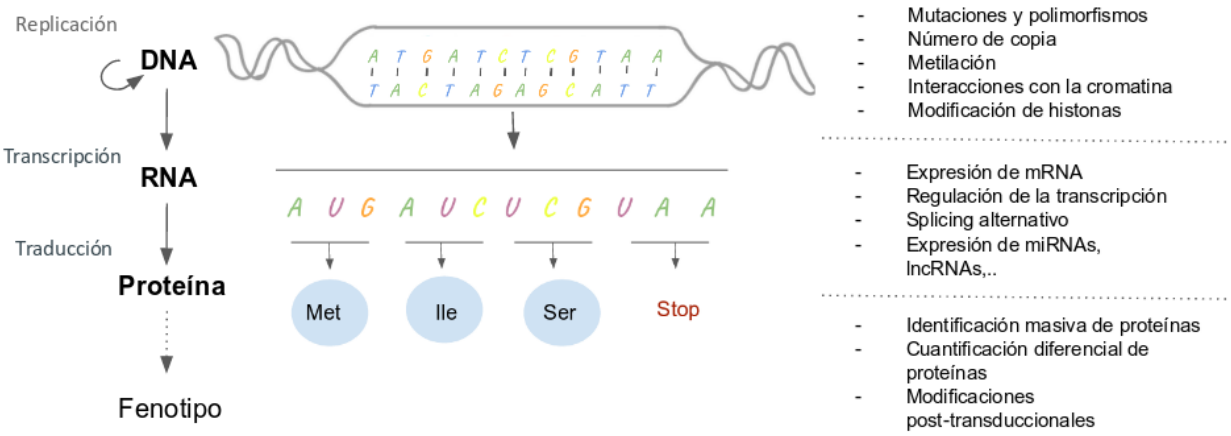


Figura 1. Esquema del dogma central de la biología y los principales estudios ómicos a nivel de ADN, RNA y proteína.

En el primer paso de la expresión génica, conocida como transcripción, la información del ADN se utiliza como molde para crear moléculas de ARN. El ARN se sintetiza utilizando una de las cadenas de ADN como plantilla y tiene la misma estructura química, excepto que la timina se reemplaza por uracilo (U). Algunas moléculas de ARN pueden ser el producto final en sí mismas (ARN funcionales) y algunas, a su vez, pueden usarse como una plantilla para la creación de las proteínas en un proceso llamado traducción. Las proteínas están compuestas por secuencias de aminoácidos, cada uno de los cuales está codificado por un triplete de nucleótidos de ARN. En la naturaleza son 20 los aminoácidos más comunes que se concatenan mediante enlaces peptídicos. Estos aminoácidos están formados por un átomo central de carbono, conocido como carbono alfa, al que enlazan: un grupo amino (NH2), un grupo carboxilo (COOH), un átomo de hidrógeno y una cadena denominada lateral que caracteriza a cada aminoácido.

A pesar de que prácticamente todas las células de nuestro cuerpo contienen la misma información genética, no todas las células son iguales. Por ejemplo, una célula de la piel es distinta a una del hueso, debido a que no todas las células utilizan todo el ADN que tienen disponible. Normalmente se considera que sólo están activos aquellos genes que se están transcribiendo a ARNm (ARN mensajero) para producir proteínas o ARNs funcionales, aunque la cantidad de ARNm puede no correlacionar exactamente con la cantidad de proteína. Los ARNm presentes en una célula o las proteínas que se están expresando pueden ser un buen indicativo de los procesos que se están llevando a cabo en ellas [1].

Por ejemplo, en el genoma humano existen 19.901 genes codificantes de proteínas, pero al incluir otras formas de ARN como *lncRNA*, *ncRNA* y pseudogenes, el número total de genes asciende a 58.381 (estadísticas de Gencode de noviembre de 2017, https://www.gencodegenes.org/). A partir de estos genes han sido predichas 19.656 proteínas de las cuales 2.186 aún no han sido identificadas (estadísticas de *HUPO* de mayo de 2018, https://www.hupo.org/).

Debido a que el ADN es la base molecular de todo sistema vivo, pequeños fallos o variaciones en su secuencia (mutaciones) pueden provocar errores vitales. Existen diferentes tipos de mutaciones, aquellas que afectan al producto génico y las que no. Cuando estas mutaciones se extienden a un grupo suficientemente grande de la población se consideran polimorfismos.

1.2. Ciencias ómicas

El término ómico se refiere al estudio global de los sistemas celulares en un nivel concreto. Las principales ciencias ómicas desarrolladas en los últimos años son la Genómica, la Transcriptómica, la Proteómica y la Metabolómica. Refiriéndose la Genómica al estudio de los genes del ADN; la Transcriptómica al estudio de

transcritos o ARN mensajero; la Proteómica al estudio de proteínas y la Metabolómica al estudio de los metabolitos. Estas disciplinas dependen del análisis de un gran volumen de datos, y por lo tanto se valen de la Bioinformática y de técnicas rápidas y automatizadas de alto rendimiento (*high-throughout techniques*) [2].

1.2.1. Genómica

La Genómica es el campo de la Genética que intenta comprender el contenido, la organización, la función y la evolución de la información molecular del ADN albergada en el genoma completo. Conocer dicha secuencia, nos permite poder identificar los genes contenidos y estudiar las funciones de los mismos de forma detallada.

Con la secuenciación del genoma humano se inició una nueva era en la investigación biomédica. Antes, se estudiaban los genes uno por uno, su localización cromosómica, su función y su asociación con enfermedades, y actualmente en la era Genómica, estudiamos de forma masiva el genoma completo observando los cambios que se generan bajo diferentes condiciones o circunstancias. Esta nueva era se inició con la finalización del Proyecto Genoma Humano, que empezó en 1990 y su objetivo fue la secuenciación y ensamblaje completo del genoma humano. Fue en febrero de 2001 cuando las empresas *HUGO* y *Celera* lograron secuenciar el genoma y publicaron los primeros borradores que cubrían el 90% del genoma. El 10% restante correspondía a heterocromatina y se secuenció en 2006.

La Genómica se divide en dos ramas principales: la Genómica estructural orientada a la caracterización y determinación de la conformación tridimensional de las proteínas y la Genómica funcional disciplina que se orienta hacia la recolección sistemática de información sobre las funciones de los genes. Esta última, emplea técnicas de análisis masivo para el estudio de genes, proteínas y metabolitos. Se podría decir que llena el hueco existente entre el conocimiento de las secuencias de un gen y su función [3]. En términos generales, la Genómica funcional trata de explicar el origen de un fenotipo determinado a partir de los cambios generados en cualquiera de los niveles moleculares antes mencionados. De esta manera, además de estudiar la secuencia del ADN en sí misma, se puede ramificar en otras muchas aproximaciones ómicas.

1.2.2. Transcriptómica

La Transcriptómica estudia y compara transcriptomas, es decir, los conjuntos de ARN mensajeros o tránscritos presentes en una célula, tejido u organismo. Comprender el transcriptoma es esencial para interpretar los elementos funcionales del genoma y revelar los constituyentes moleculares de las células y los tejidos, y también para comprender el desarrollo y las enfermedades [4]. Los objetivos de la Transcriptómica son: catalogar todas las especies de transcritos; incluyendo ARNm, *ncRNA* y *snRNA*, para determinar la estructura

transcripcional de los genes como los sitios de inicio, los extremos 5' y 3', patrones de *splicing* y modificaciones post-transcripcionales; y además cuantificar los niveles cambiantes de expresión en diferentes condiciones [5].

Varias tecnologías se han desarrollado para deducir y cuantificar el transcriptoma. Desde métodos basados en la hibridación como los microchips de ADN (*microarrays*) o métodos basados en secuencias, hasta llegar a la tecnología actual, la secuenciación masiva (*NGS, Next Generation Sequencing*).

1.2.2.1. De los microarrays a la secuenciación masiva

Un *microarray* o chip de ADN consiste en una superficie sólida a la que se une una colección de fragmentos de ADN llamadas sondas. Las superficies empleadas para fijar el ADN son muy variables y pueden ser de vidrio, plástico e incluso de silicona. Su funcionamiento consiste básicamente en marcar con moléculas fluorescentes las cadenas complementarias de los fragmentos de ADN o ARN de la muestra a analizar para que hibriden con las sondas, de esta forma, para cada secuencia se obtiene un valor de fluorescencia que representa la cantidad de ADN/ARN con esa secuencia que hay en la muestra estudiada. Los *microarrays* suelen utilizarse para identificar aquellos genes que tienen una expresión diferente en distintas condiciones. Por ejemplo, para detectar que genes producen o están involucrados en ciertas enfermedades, se comparan los niveles de expresión de los genes entre células sanas y células que están desarrollando alguna enfermedad.

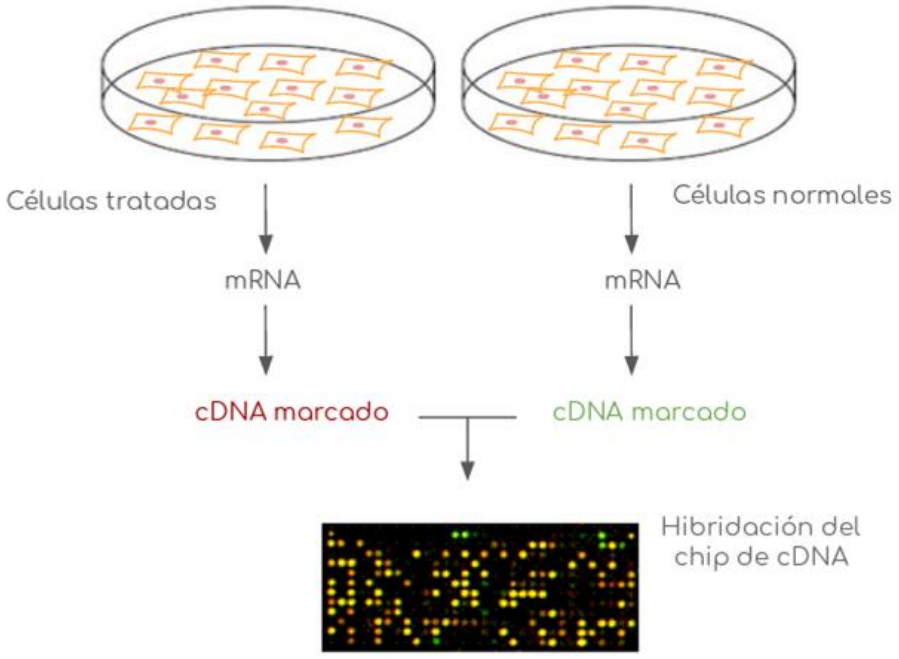


Figura 2. Diagrama de un experimento típico de chip de dos canales. Cada una de las muestras es marcada con un fluróforo de diferente color (rojo y verde) que se hibridan en un mismo chip de ADN.

Existen diferentes tipos de *microarrays*, como los *microarrays* de dos canales, los chips de oligonucleótidos de ADN y los chips de ADN para genotipado. En la Figura 2 se muestra un diagrama de un experimento típico de chip de dos canales. En este tipo de chips se marcan con un fluoróforo diferente las dos muestras biológicas que se quieren estudiar, éstas se mezclan e hibridan sobre un mismo chip de ADN que a continuación se escanea para visualizar los resultados. Esta metodología se suele utilizar para detectar genes que se activan o se reprimen en distintas condiciones. Los enfoques basados en la hibridación son métodos de alto rendimiento y de relativamente bajo coste. Sin embargo, estos métodos tienen varias limitaciones como tener que confiar en la secuencia existente del genoma y obtener un rango de detección limitado debido al ruido de fondo y a la saturación de la señal. Además, la comparación de niveles de expresión en diferentes experimentos es a menudo difícil y la normalización puede ser complicada [5] .

En contraste con los métodos de *microarrays*, los enfoques basados en secuencias determinan directamente la secuencia de ADNc (ADN complementario). La secuenciación por Sanger fue de las primeras en utilizarse. En esta técnica se aísla el fragmento de ADN a estudiar, se amplifica mediante *PCR* y se secuencia. Sin embargo, tiene limitaciones es una técnica cara, lenta, sólo permite secuenciar fragmentos cortos de ADN y no cuantifica.

Este escenario cambió cuando la farmacéutica Roche desarrolló una nueva técnica llamada piro-secuenciación, para secuenciar millones de fragmentos de ADN de una forma muchísimo más rápida y económica que el método Sanger. Al poco tiempo, otras compañías comenzaron a desarrollar técnicas similares con la misma idea: aislar contenido genómico, trocearlo en fragmentos pequeños y secuenciarlos todos a la vez. Posteriormente mediante algoritmos computacionales, juntar todos los fragmentos en el orden adecuado para obtener el genoma original. La diferencia entre todas estas técnicas es el principio químico que usan para secuenciar el ADN. Estas técnicas se llaman Técnicas de Secuenciación masiva (en inglés *NGS*, *Next-Generation Sequencing*) y algunos ejemplos de las plataformas de secuenciación más utilizadas son: *Illumina* (*Illumina*), *454* (*Roche*), *Ion Torrent* (*Life Technologies*), *Solid* (*Applied Biosystems*), *PacBio* (*Pacific Biosciences*).

Con el desarrollo de estas nuevas tecnologías de secuenciación de alto rendimiento se originó un nuevo método para mapear y cuantificar los transcriptomas llamado *RNA-seq*. Éste método tiene claras ventajas frente a otras aproximaciones existentes y ha revolucionado la forma de estudiar el transcriptoma. A diferencia de los *microarrays*, que se basan únicamente en los transcritos conocidos contenidos en el chip, la técnica *RNA-seq* ofrece una visión sin procedentes del transcriptoma de un determinado tejido o tipo de célula permitiendo la realización de diferentes análisis que no se podían realizar anteriormente.

Como se representa en la Figura 3 para realizar un análisis de *RNA-seq* primero se convierte la población de moléculas de ARN (total o fraccionada) a analizar en fragmentos de ADNc y a cada molécula, se les une adaptadores con secuencia única en uno, o en ambos extremos. Cada molécula es secuenciada para obtener secuencias cortas desde un extremo (*single-end sequencing*) o desde ambos extremos (*paired-end sequencing*).

Estas lecturas suelen tener un tamaño de unas 30-500 bp dependiendo de la plataforma utilizada para secuenciar. En principio cualquier tecnología de secuenciación masiva puede ser utilizada para *RNA-seq*. Los datos adquiridos de la secuenciación consisten en una lista de secuencias cortas junto con la calidad asociada a cada lectura. Las lecturas obtenidas son alineadas al genoma de referencia con el fin de construir un mapa del transcriptoma sobre el genoma, y así poder cuantificar los niveles de transcripción para saber si un gen se está expresando, confirmar o revisar los extremos 5' o 3' anotados en la referencia, mapear los límites de los exones/intrones, etc [6].

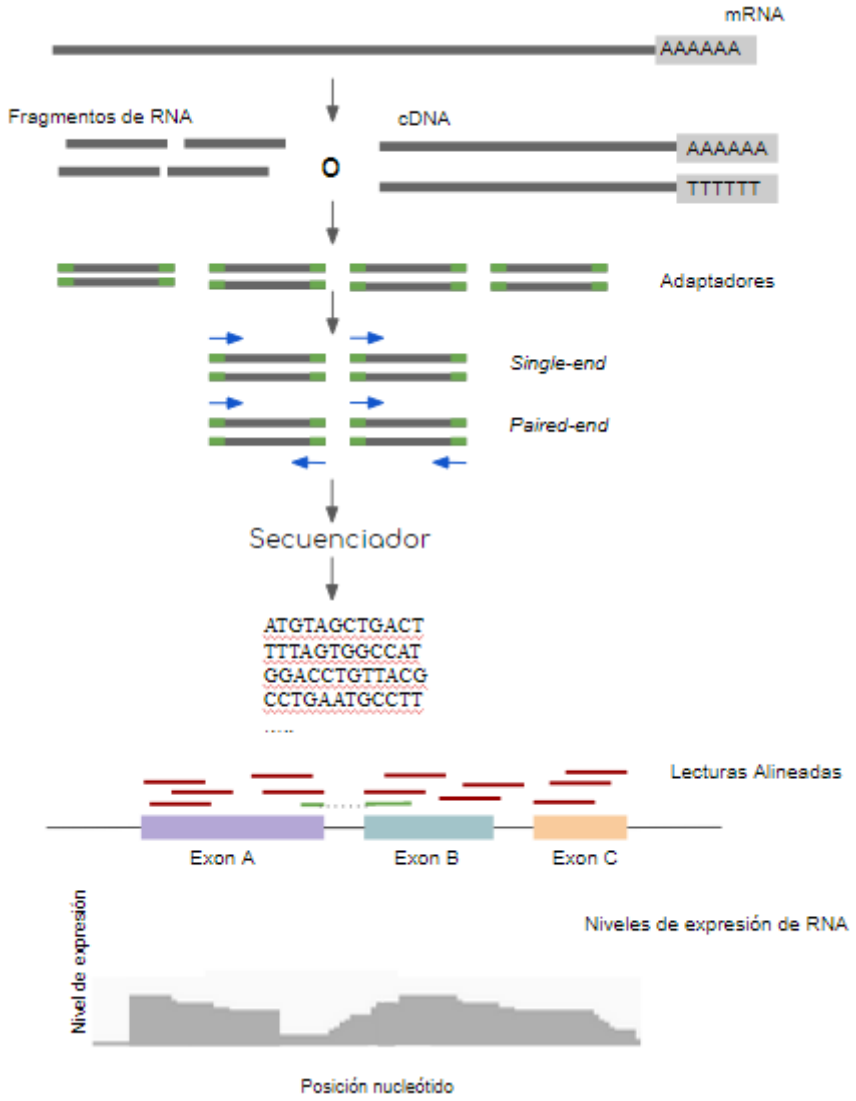


Figura 3. Representación de un experimento típico de RNA-seq. Los ARNm se convierten a ADNc o se fragmentan en pequeños ARNs. A continuación, se añaden adaptadores (verdes) a los fragmentos y se secuencian según si se quieren lecturas *paired-end* o *single-end*. Del secuenciador se obtienen lecturas con las secuencias que son alineadas al genoma de referencia. Finalmente se cuantifican las lecturas alineadas en el genoma.

Esta técnica obtiene unos resultados con una resolución mucho mayor que con las metodologías descritas anteriormente [6,7]. Permite el análisis de expresión de los genes, el estudio de variantes de *splicing*, la expresión de alelos específicos, y el descubrimiento de variantes de transcritos raras y/o nuevas [8]. Además,

también se puede realizar análisis de transcriptoma en organismos en los que no existe un genoma de referencia (secuenciación del transcriptoma *de novo*).

1.2.2.2. Análisis de datos de secuenciación masiva

En la mayoría de estudios *RNA-seq*, los análisis de datos constan de los siguientes pasos clave representados en la Figura 4 [9,10] : control de calidad y preprocesamiento de las lecturas, alineamiento de las lecturas, cuantificación de las lecturas alineadas a los genes, normalización de los datos e identificación de los genes con expresión diferencial , e interpretación biológica y análisis de enriquecimiento funcional [11,12].

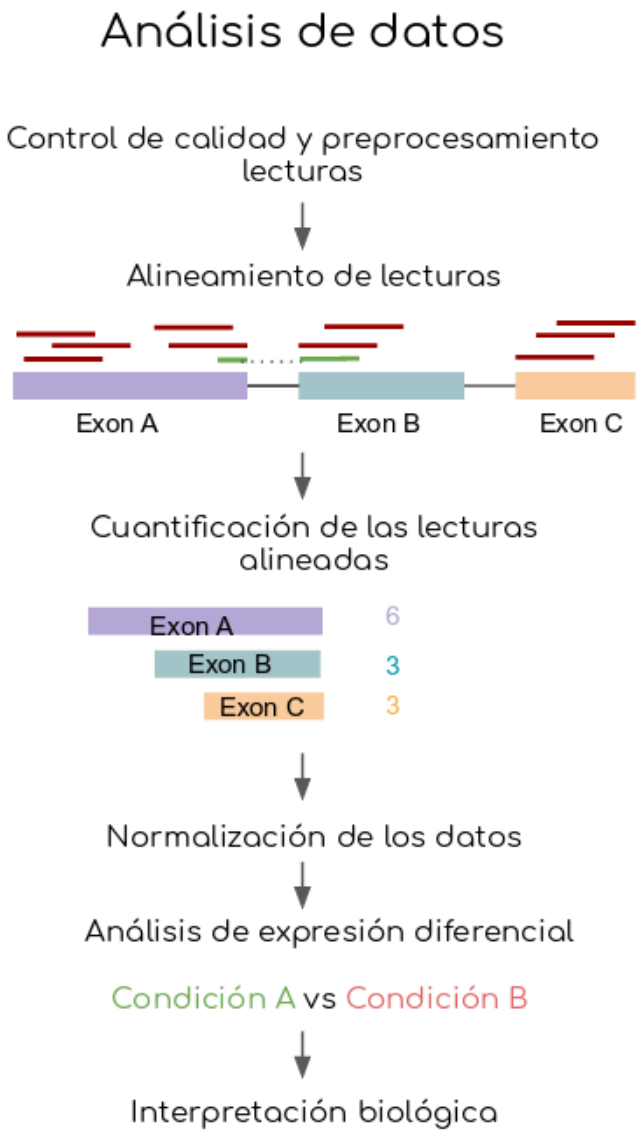


Figura 4. Pasos claves en el análisis de datos de *RNA-seq*. El primer paso consiste en comprobar la calidad de las lecturas para proceder al alineamiento de éstas al genoma o transcriptoma de referencia. A continuación, se cuentan las lecturas alineadas en regiones de interés para posteriormente ser normalizadas y realizar un análisis de expresión diferencial. Finalmente, con los datos resultantes, se lleva a cabo la interpretación biológica.

1.2.2.2.1. Control de calidad y preprocesamiento

El primer paso en todo análisis de *RNA-seq* es comprobar que la calidad de las lecturas obtenidas en la secuenciación es correcta. Analizar datos de mala calidad puede afectar al análisis y a la interpretación de los datos, dando lugar a resultados inexactos. La baja calidad puede ser debida a problemas en la preparación de las librerías y/o problemas en la propia secuenciación. Las lecturas pueden contener artefactos de *PCR*, secuencias con adaptadores, sesgos específicos de secuencias, contaminantes, etc.

Existen algunas herramientas computacionales como *PRINSEQ* [13] o *FASTQC* [14] entre otros, para visualizar y evaluar la calidad de los datos. Cuando se observa una mala calidad, existen programas como *Cutadapt* [15], *FASTX-Toolkit* [16] y *Trimmomatic* [17] que permiten procesar y eliminar aquellas lecturas no aptas para el análisis.

1.2.2.2.2. Alineamiento

El segundo paso consiste en alinear las lecturas generadas por los secuenciadores a un genoma o transcriptoma de referencia para descubrir que posiciones han sido transcritas. Alinear al genoma de referencia y no al transcriptoma tiene la ventaja de poder encontrar nuevos genes e isoformas, pero requiere de la habilidad del analista de datos y de la precisión del programa para poder cortar las lecturas. Esto es debido a que las bibliotecas de *RNA-seq* han sido construidas a partir de transcritos de ARN por lo que las secuencias intrónicas no están presentes en las lecturas y esto es un problema en las secuencias que abarcan exon-exon (Figura 5). En estos casos, cuando se alinea al genoma de referencia se utiliza un método que suele consistir en dos pasos: en el primero alinean todas las lecturas posibles, que alinearán en los exones, y en el segundo paso se rompen las lecturas que no han sido alineadas previamente y se trataran de alinear de forma independiente [4]. En los últimos años se han desarrollado programas como *TopHat2* [18], *GSNAP* [19] , *QPALMA* [20], *STAR* [21] y *SOAPSplice* [22] que permiten romper estas lecturas, pero existen otros como *Bowtie* [23], *BWA* [24] y *SOAP* [25] que están especializados en alinear lecturas cortas.

Otra característica a tener en cuenta al alinear son las lecturas múltiples, es decir, aquellas lecturas que alinean en varias regiones del genoma o transcriptoma. Estas lecturas múltiples son muy comunes en genomas largos y complejos, pudiendo constituir del 10 al 40% del genoma [26]. Existen tres estrategias para solucionar el problema. La primera consiste en descartar todas estas lecturas, pero conlleva una pérdida de información importante produciendo un sesgo en la cuantificación de expresión de los genes. La segunda estrategia consiste en asignar una posición aleatoriamente asumiendo que todas las posiciones tienen la misma probabilidad de que el ARN se haya generado en cada una de las posiciones, lo que muchas veces no es válido. Y la última estrategia, y más utilizada, consiste en reportar todos los alineamientos con un número máximo

especificado por el usuario, como por ejemplo reportar 10 alineamientos. El problema principal de esta aproximación es que el umbral de corte es totalmente arbitrario [11].

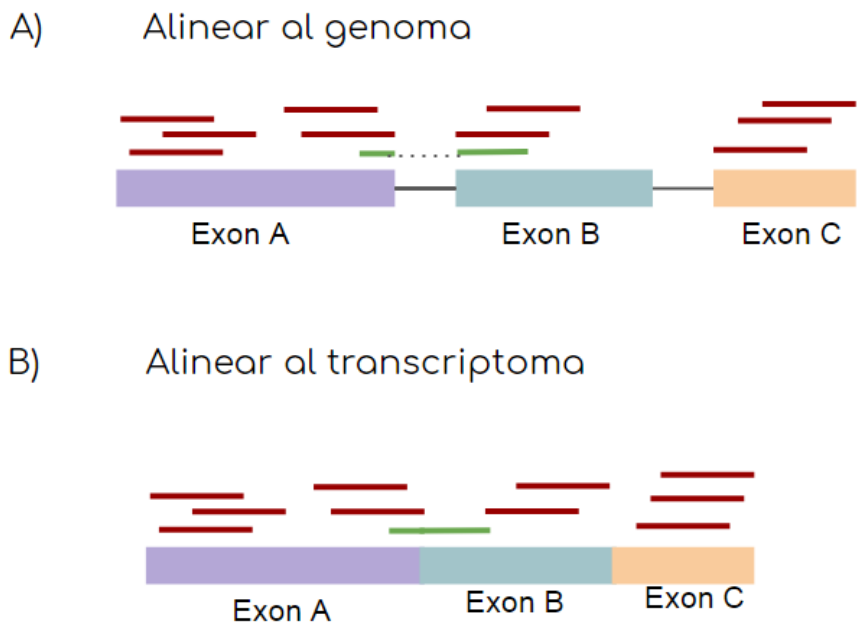


Figura 5. Alineamiento a un genoma (A) y a un transcriptoma (B) de referencia. El alineamiento a un genoma a veces requiere que las lecturas se rompan para poder alinear entre intrones, en cambio en el alineamiento a un transcriptoma no.

1.2.2.2.3. Cuantificación

Independientemente del método utilizado para el alineamiento, el siguiente paso es estimar la expresión, contando cuantas lecturas han sido asignadas a cada uno de los genes. Muchos métodos han sido desarrollados para inferir la abundancia de genes e isoformas como *RSEM* [27], *Cufflinks* [4], *IsoEM* [28], *FeatureCounts* [29], y *HTSeq* [30]. Estos algoritmos se suelen dividir en dos categorías: las aproximaciones basadas en los transcritos y las aproximaciones que se basan en las uniones de exones.

Las aproximaciones basadas en los transcritos tratan de distribuir las lecturas entre distintas isoformas de un gen. Sin embargo, existe una alta superposición genómica entre las distintas lecturas y resulta difícil estimar la expresión de isoformas individuales (ver Figura 6) [31]. En cambio, los métodos basados en uniones de exones son mucho más simples porque todas las lecturas alineadas en los exones se contemplan como las propias del gen. Una lectura se asigna al gen siempre que tenga suficiente superposición con cualquiera de sus exones. En comparación con las isoformas, las lecturas pueden asignarse a genes con mucha más confianza. Por lo tanto, el método de recuento basado en exón de unión es comúnmente utilizado en *RNA-seq*,

aunque en los recuentos a nivel de genes es complicado distinguir isoformas [32].

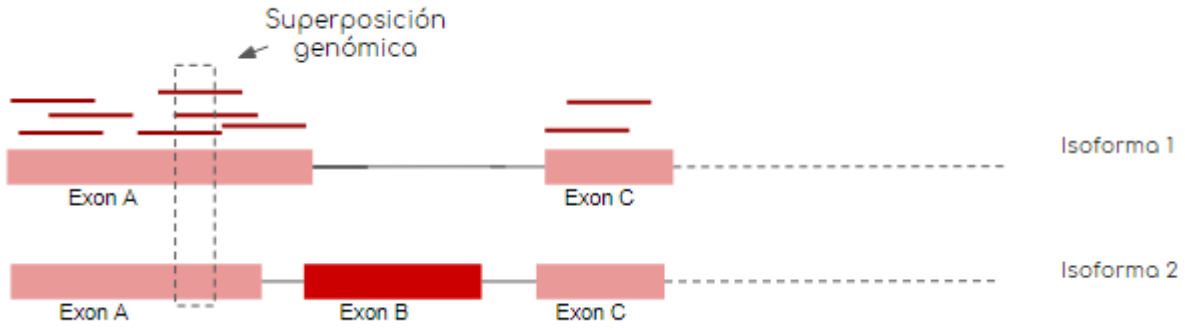


Figura 6. Superposición genómica. En la figura se representa la dificultad de estimar la expresión de isoformas por la superposición genómica.

1.2.2.2.4. Normalización y análisis de expresión diferencial

Una vez se obtienen las estimaciones de expresión, es necesario garantizar, mediante la normalización que los niveles de expresión entre las muestras son comparables. Para ello, es importante tener en consideración varias características como la longitud del transcrito y la profundidad de la secuenciación de la muestra. La mayoría de análisis de *RNA-seq* consisten en observar la expresión diferencial de los genes y con ese objetivo se han desarrollado varios programas que normalizan y calculan las diferencias entre las muestras como *DESeq* [33], *EdgeR* [34][35], *GENE-Counter* [36]y *Cuffdiff2* [37]. Sin embargo, hoy en día no existe un consenso sobre qué aproximación es la más acertada [11]. Estos programas generan una lista de genes con valores estadísticos que representan la diferencia de expresión de los genes entre las distintas condiciones.

El siguiente y último paso del análisis es la interpretación biológica de estos genes diferencialmente expresados y el análisis de enriquecimiento funcional que se detalla en el apartado 1.2.6. de la introducción.

1.2.2.3. Análisis de variantes

Los análisis de *RNA-seq* también permiten identificar variantes. Las variantes consisten en cambios de nucleótidos en la secuencia de ADN (Figura 7). Estas pueden ser un polimorfismo de nucleótido único (del inglés, *Single Nucleotide Polimorfism* o *SNP*) o una inserción o eliminación (también llamado *INDEL*) de nucleótidos de la secuencia que suelen conllevar un desplazamiento del marco de lectura de la proteína.

Los polimorfismos de nucleótido único o *SNP* se pueden clasificar en diferentes tipos: variaciones sinónimas, aquellas que el cambio de nucleótido no genera un cambio en el aminoácido y las variaciones no sinónimas que son aquellas que sí generan un cambio en el aminoácido. El impacto del cambio puede variar según el

triplete que se forme, pudiendo llegar a truncar la proteína si se introduce un codón *STOP* en su secuencia. Identificar estas variaciones genéticas es crucial para revelar la relación entre genotipo y fenotipo además pueden dar información importante acerca de ciertas enfermedades.

Existen varias herramientas para identificar estas variaciones, pero el primer paso como en cualquier experimento de *RNA-seq* consiste en alinear las lecturas al genoma de referencia. Cuando se quieren encontrar variantes, se debe especificar en los parámetros del alineador que acepte alineamientos no perfectos, es decir, que permita que las lecturas alineen en regiones en que la secuencia de la lectura no sea exactamente igual a la de referencia.

A partir de este punto existen herramientas como *SAMtools* con *mpileup* [38] o *Gatk* [39] que buscan aquellas posiciones donde existe una cantidad suficiente de lecturas en las cuales el nucleótido de la lectura no corresponde con el de referencia. A cada posible mutación se le asigna una puntuación que contemple el número de lecturas que respaldan la evidencia, la calidad del alineamiento, etc. Estas mutaciones se suelen representar en ficheros con formato *VCF*.

Estas mutaciones suelen filtrarse por calidad y luego ser anotadas (asignadas a un gen) con programas como *SnpEff* [40] o *VEP* [41] que permiten asociar información biológica a la anotación, como por ejemplo describir el gen al que está afectando la mutación e inferir cual podría ser el impacto de la mutación en la proteína.

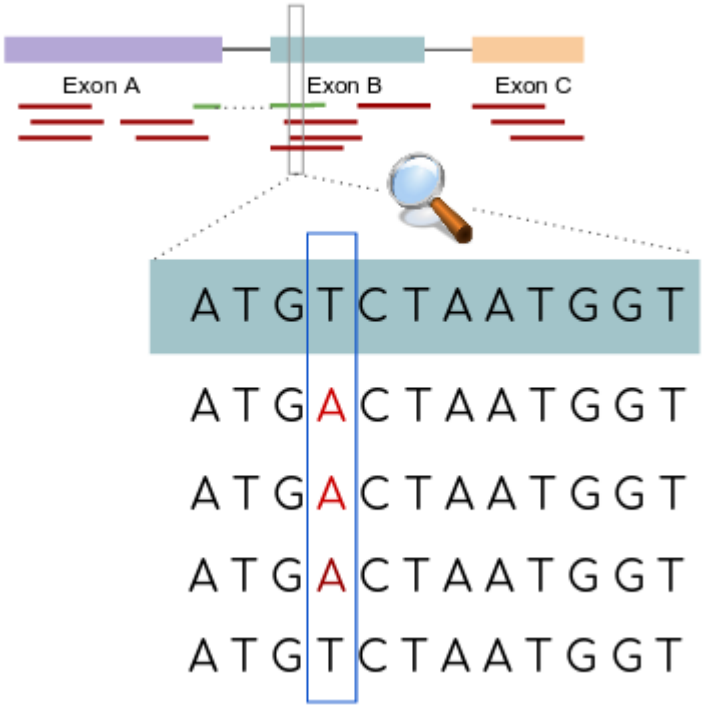


Figura 7. Variaciones en las lecturas alineadas al genoma de referencia. Las 4 lecturas que han alineado en la región del genoma representada contienen variaciones con respecto el genoma de referencia. En esta posición el genoma de referencia contiene Timidina mientras que 3 de las 4 lecturas alineadas corresponden a Adenina.

1.2.3. Proteómica

La Proteómica se define como el conjunto de técnicas o tecnologías utilizadas para la obtención de información funcional de las proteínas, y tiene por objetivo el análisis, identificación y caracterización del proteoma celular. El concepto proteoma fue acuñado en 1994 por Marc Wilkins fusionando las palabras proteína y genoma. Si el genoma es la dotación génica de una célula u organismo, el proteoma es entendido como el conjunto total de proteínas expresadas por los genes de una célula, tejido u organismo. Sin embargo, mientras que el genoma es básicamente el mismo en todas las células de un organismo, el proteoma es mucho más variable, siendo diferente en los distintos tipos celulares. Además, la cantidad de material genético en la célula no tiene por qué correlacionar con el rango de concentración de la proteína. Las secuencias génicas pueden sufrir *splicing* alternativo y generar diferentes variantes proteicas a partir de un gen o modificarse después de la traducción dando otras variantes por lo que el proteoma siempre será más grande que el número de genes [42] (Figura 8).

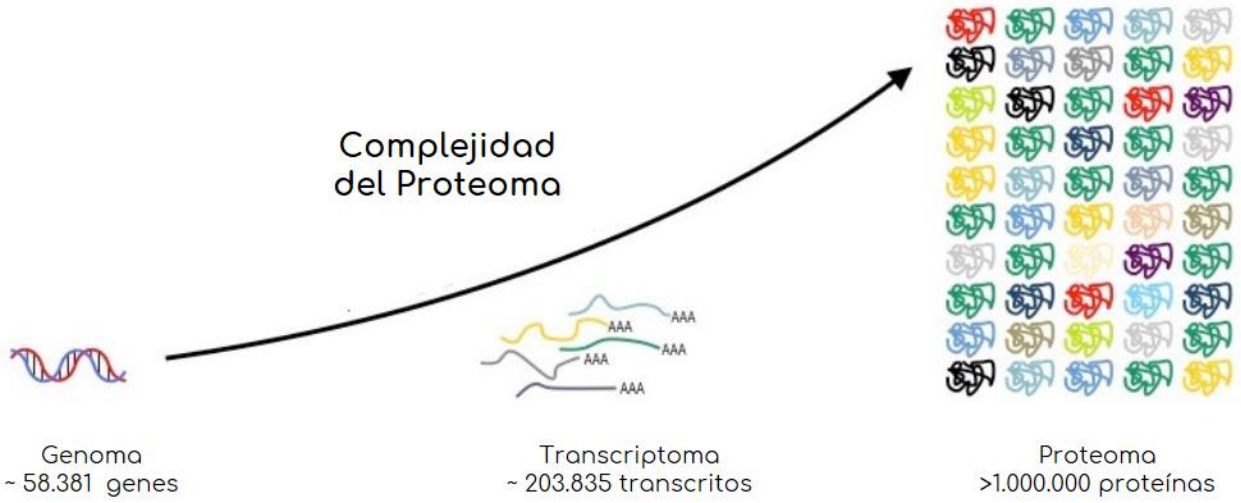


Figura 8. Representación de la complejidad del proteoma respecto al genoma. El genoma humano contiene aproximadamente 58.000 genes mientras que el transcriptoma alrededor de 200.000 trascritos. Las cadenas polipeptídicas pueden ser procesadas de varias formas diferentes generando diferentes proteínas por cada transcrito. Figura modificada de http://ib.bioninja.com.au/standard-level/topic-2-molecular-biology/24-proteins/proteome.html

El desarrollo de la Proteómica ha experimentado un enorme avance en los últimos años, gracias fundamentalmente al desarrollo tecnológico. Hoy en día, el estudio del proteoma incluye técnicas muy sofisticadas basadas en métodos electroforéticos y cromatográficos, espectrometría de masas y tecnología bioinformática. La espectrometría de masas, en particular, se ha convertido en una herramienta indispensable para la Proteómica.

1.2.3.1. Identificación de proteínas por Espectrometría de masas

La espectrometría de masas es una herramienta analítica muy poderosa que se basa en la obtención de iones a partir de moléculas en fase gaseosa que se separaran de acuerdo con su relación masa/carga (*m/z*). Estos iones se detectan por medio de un dispositivo adecuado obteniendo un espectro de masas que representa la abundancia relativa de los distintos iones en función de su *m/z*. Los espectros de masas pueden suministrar información sobre la estructura de especies moleculares complejas, las relaciones isotópicas de los átomos en las muestras, y la composición cualitativa y cuantitativa de analitos en muestras complejas [43,44].

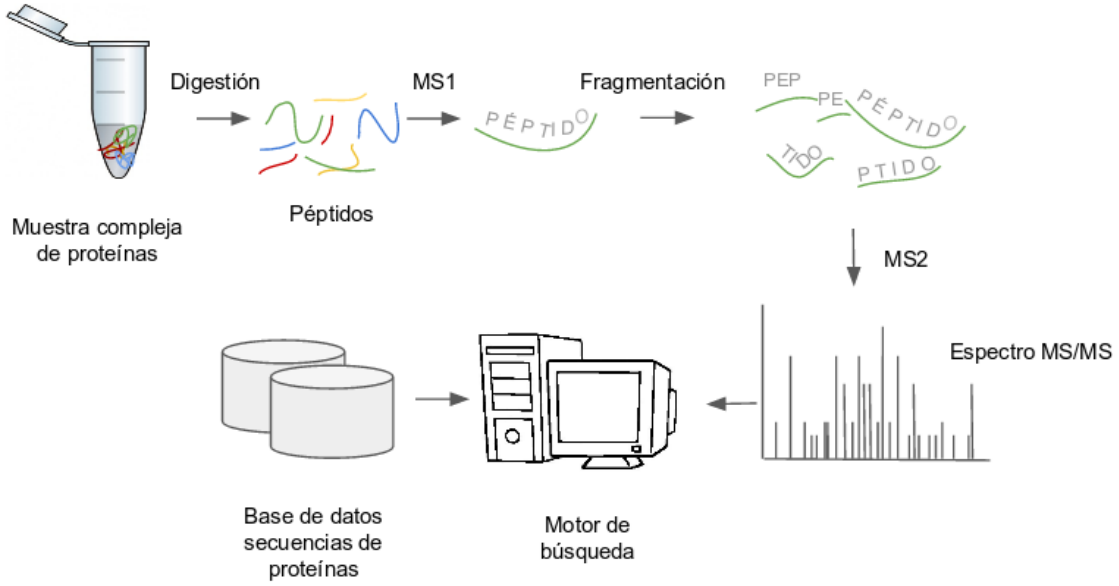


Figura 9. Flujo de trabajo típico para la identificación y caracterización de proteínas utilizando MS/MS. Una muestra compleja de proteínas es digerida con una enzima generando péptidos. Estos péptidos son introducidos al espectrómetro de masas. Los péptidos son seleccionados uno a uno (MS1) e inducidos a fragmentación (MS2), posiblemente por colisión y a continuación se captura el espectro MS/MS. El espectro de fragmentación es comparado con la base de datos para obtener la secuencia del péptido.

La Espectrometría de Masas es esencial en el contexto de la Proteómica actual debido a su alta capacidad de análisis, su sensibilidad y su precisión en la determinación de masas moleculares proteicas. En la Figura 9 se muestra el típico flujo de trabajo para la identificación y caracterización de proteínas por espectrometría de masas en tándem. Partiendo de una muestra proteica, que puede ser una proteína o varias, una enzima (generalmente tripsina) corta las proteínas en péptidos. A continuación, la muestra es introducida al espectrómetro, y para ello es necesario regular el flujo de entrada de la muestra que se suele hacer mediante la cromatografía líquida de alta resolución (*HPLC*). Una vez los péptidos están en el espectrómetro, son seleccionados uno a uno para ser detectados (huella peptídica) o para ser inducidos a fragmentación y capturar

sus espectros de *MS/MS* (espectrometría de masas en tándem). A continuación, con un motor de búsqueda se determina que secuencia peptídica de la base de datos corresponde mejor al espectro generado. Esto es posible porque previamente, estas bases de datos de proteínas son digeridas *in silico* utilizando la conocida especificidad de la enzima que ha sido utilizada para la digestión de la muestra, y son calculadas las masas de los péptidos. Si la masa calculada de un péptido coincide con la del péptido observado, las masas de los iones esperados se calculan y se comparan con los valores experimentales, asignando ese péptido al espectro.

1.2.3.1.1. Espectrómetro de masas

Esencialmente, el espectrómetro de masas debe ser capaz de desempeñar cuatro funciones:

1) Vaporizar substancias de volatilidades muy diferentes.
2) Originar iones a partir de moléculas neutras en fase gaseosa.
3) Separar los iones en función de su relación masa/carga.
4) Detectar los iones formados y registrar la información adecuadamente.

Los espectrómetros de masas utilizados en el análisis de proteínas o péptidos deben estar formados por las siguientes partes:

- **Sistema de introducción de muestras**

Existen diferentes métodos para la introducción de muestras y la elección depende de la naturaleza de las mismas. Los sistemas cromatográficos son los más populares actualmente. Este acoplamiento al espectrómetro consiste generalmente en una columna capilar (comúnmente de fase reversa) de caudal controlado que permite que los analitos de una muestra compleja sean separados e introducidos gradualmente en el equipo para ser detectados. Existen sistemas como ESI (del inglés, *Electroespray ionization*), en el que el sistema de entrada y la fuente de iones forman parte de un único componente.

- **Fuente de iones**

La fuente de iones es la parte del Espectrómetro de Masas que es capaz de ionizar la muestra, originando iones a partir de las moléculas neutras mediante la ganancia o pérdida de un electrón o protón. Las técnicas de ionización más utilizadas en Proteómica son *ESI* y *MALDI* (del inglés, *Matrix-Assisted Laser Desorption/Ionization*) aunque existen también otros métodos un poco menos utilizados

  - La técnica *MALDI* [45] consiste en embeber la muestra en una matriz líquida, que posteriormente se seca para producir una mezcla cristalizada. Esta se irradia con un láser pulsado que provoca que la matriz absorba esta energía y la convierta en energía de excitación ocurriendo la transferencia de iones al analito, generando una nube gaseosa. Las colisiones de la muestra con las

moléculas de gas contenidas en el equipo ionizan la muestra provocando la aceleración de los iones hacia la cámara de vacío del espectrómetro de masas.

- En *ESI* [46] los analitos se encuentran en fase liquida disueltos en un solvente orgánico volátil. Este analito cuando deja a tras la columna cromatográfica, atraviesa un capilar de metal cargado. Este último a su vez, provoca una diferencia de potencial entre el capilar y la fuente de iones. Este fenómeno trae como consecuencia la formación de un aerosol cuando el líquido sale del capilar. Conforme el solvente se evapora, las moléculas de analito se aproximan, se repelen y cuando la repulsión de las cargas positivas vence la tensión superficial, estallan las gotas (Explosión de Coulomb) quedando en suspensión y siendo así introducidos en un sistema de vacío hacia el espectrómetro.

- **Analizador, para la separación de iones.**

A la salida de la fuente de iones, se tiene una mezcla de diversos iones que deben ser separados para ser detectados de forma individual. El analizador de masas es la parte del instrumento en la que los iones se separan en base su relación *m/z* y suele consistir en una cámara de vacío donde se aplica algún tipo de campo eléctrico o magnético para poder ver las diferencias entre los iones al moverse a través de la cámara. Los analizadores más utilizados en Proteómica son: los analizadores de tipo TOF, los Cuadropolos, orbitrap y las Trampas iónicas [43,44]. Existen más tipos y además los analizadores pueden combinarse formando por ejemplo un *QTOF* (*Quadrupole-Time Of Flight*) que consiste en un analizador de tipo cuadrópolo acoplado a un analizador de tiempo de vuelo o un *Qtrap* (*Quadrupole-Ion Trap*) que es un cuadrópolo unido a una trampa iónica.

- **Detector**

El detector convierte el haz de iones que incide sobre él en una señal eléctrica medible para poder ser procesada y almacenada. El detector más empleado es el multiplicador de electrones que consiste en varias placas sobre las cuales se aplica una diferencia de potencial. Cuando un ion incide sobre su superficie se produce una descarga de electrones que inciden en otra placa y así sucesivamente de forma que se produce una reacción en cadena consiguiendo una amplificación de señal.

1.2.3.1.2. Técnicas Proteómicas para la identificación de proteínas

Existen dos métodos principales para la identificación de proteínas por espectrometría de masas. El método basado en la huella peptídica y los basados en la secuencia peptídica [47]. El método de huella peptídica consiste en digerir una muestra con una enzima proteolítica y adquirir los espectros MS generados por las masas de los péptidos. Al realizarse con una enzima con un patrón de corte específico estos cortes son conocidos y por lo tanto el espectro que genera puede ser utilizado para identificar la proteína, comparando las

masas de los espectros teóricos con aquellos obtenidos en el experimento. Sin embargo, a pesar de esta especificidad, una enorme variedad de proteínas implica una mayor variedad de péptidos generados a partir de ellas que pueden tener unas masas muy similares. Por ese motivo se requiere que la proteína se encuentre previamente aislada. Esta técnica se lleva a cabo generalmente por espectrometría de masas MALDI-TOF [48].

Por otro lado, existe la espectrometría de masa en tándem o *MS/MS* que se basa en la secuencia peptídica y permite trabajar con muestras complejas [47]. En la espectrometría de masas en tándem o *MS/MS*, los péptidos una vez ionizados y en el interior del analizador, son sometidos a una fragmentación adicional. Los péptidos se fragmentan generando iones más pequeños lo que hace que el patrón de fragmentación sea más específico de la secuencia original como se representa en la Figura 10. En este proceso, primero se escanean todos los péptidos ionizados introducidos al espectrómetro y se registran los espectros *MS1* (con sus valores *m/z* y las intensidades de cada ion). A continuación, se aíslan estos iones, denominados iones precursores, y son fragmentados en péptidos más pequeños (o iones fragmento). El espectro *MS2* que se adquiere (también llamado espectro de fragmentación) registra los valores *m/z* e intensidades de los fragmentos de cada uno de los péptidos precursores aislados y fragmentados. Este espectro de fragmentación, contiene la información necesaria para deducir la secuencia de aminoácidos del péptido que lo origina.

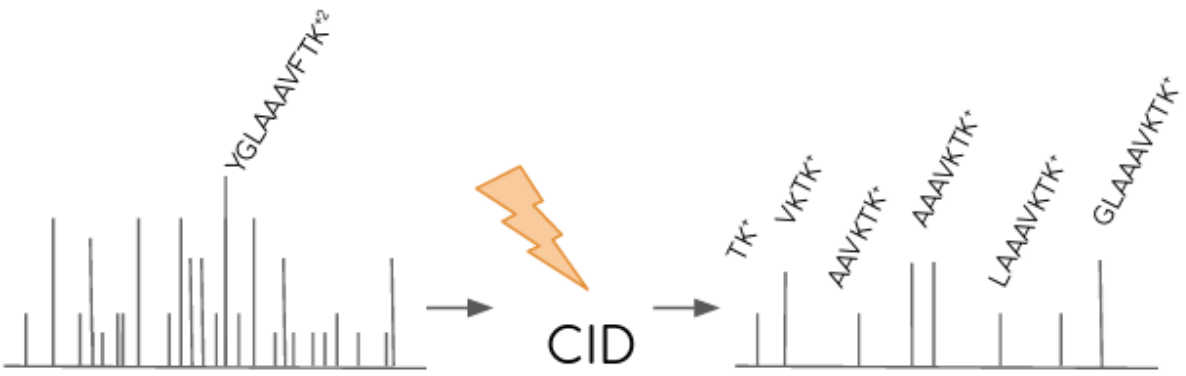


Figura 10. Fragmentación del péptido precursor. El péptido precursor (generalmente el de mayor intensidad) es seleccionado en el espectro *MS1* para ser fragmentado y generar el espectro *MS2* o *MS/MS*.

Esto aumenta la resolución del análisis porque permite distinguir péptidos intactos que tienen masas muy similares pero cuyos patrones de fragmentación *MS/MS* son diferentes, permitiendo así, incrementar el rendimiento del experimento en muestras con mezclas de proteínas complejas.

La fragmentación de péptidos se realiza por diferentes metodologías como, por ejemplo: CID (*collision induced dissociation*), ETD (*electron transfer dissociation*) y HCD (*Higher-energy collision dissociation*), entre otras. La fragmentación tipo CID es uno de los métodos más utilizado en proteómica. En él se selecciona un determinado precursor (ion parental) que será fragmentado por colisiones con un gas inerte (Nitrógeno), obteniendo así los iones fragmentos [44].

1.2.3.1.3. Bases de datos y motores de búsqueda

El método más utilizado para la asignación de secuencia peptídica a un espectro *MS/MS* es mediante la utilización de bases de datos con motores de búsqueda. Las bases de datos consisten en las secuencias de las proteínas del organismo y estas pueden obtenerse de diferentes bases de datos como del *NCBI* (https://www.ncbi.nlm.nih.gov/) o *Uniprot* (http://www.uniprot.org/) que son los repositorios más utilizados, pero también pueden obtenerse de bases de datos específicas del organismo. Todas estas bases de datos son fáciles de obtener y de descargar. Las bases de datos suelen tener diferencias entre ellas, pudiendo ser más redundantes o variar en la calidad de la anotación. La decisión al utilizar una u otra depende del experimento y del organismo con el que se trabaja. En el caso de búsquedas de polimorfismos, se suelen utilizar bases de datos grandes como *Entrez Protein (*https://www.ncbi.nlm.nih.gov/protein/*)* que pueden contener variaciones, sin embargo, las búsquedas en bases de datos grandes pueden llevar a un aumento de falsos positivos [49]. Actualmente se está utilizando la proteogenómica para encontrar variaciones en proteínas, hablaremos de este campo en el apartado 1.2.4 de la introducción.

Los motores de búsqueda deben cumplir dos funciones: Primero obtener espectros o patrones teóricos para cada una de las secuencias depositadas en la base de datos, y segundo, proporcionar un método para comparar cada uno de los patrones teóricos con el espectro real obtenido en el espectrómetro. En la

Tabla 1 se citan algunos de los motores de búsqueda más utilizados [50,51] en los que algunos son de uso gratuito y otros requieren de una licencia, además de si son sistemas óptimos para computación de alto rendimiento (del inglés *High performance computing*, *HPC*).

Tabla 1. Motores de Búsqueda

Motor de búsqueda	***URL***	**Licencia/Gratuitos**	**Óptimo para sistema *HPC***
MASCOT	http://www.matrixscience.com/	Licencia	Si
Sequest	https://www.thermofisher.com	Licencia	Si
OMSSA	http://pub-chem.ncbi.nlm.nih.gov/omssa	Gratuito	Si
X!Tandem	http://www.thegpm.org/tandem/	Gratuito	Si
FindPept	https://web.expasy.org/findpept/	Gratuito	No
Comet	http://comet-ms.source-forge.net/	Gratuito	Si

1.2.3.2. Análisis de datos proteómicos

Actualmente la Proteómica *Shotgun* es muy utilizada y se emplea principalmente para identificar proteínas en mezclas complejas, combinando la cromatografía líquida de alta resolución (del inglés, *High Performance Liquid Chromatography*, *HPLC*) con espectrometría de masas en tándem. Los datos obtenidos de un experimento típico de Proteómica *shutgun* consisten en un registro de las relaciones *m/z* y las intensidades de todos los iones de los fragmentos resultantes generados a partir del ion precursor aislado. Estos datos requieren de un análisis computacional para poder extraer de los espectros la información necesaria para identificar la secuencia de aminoácidos del péptido y así saber que proteínas contiene la muestra (Figura 11).

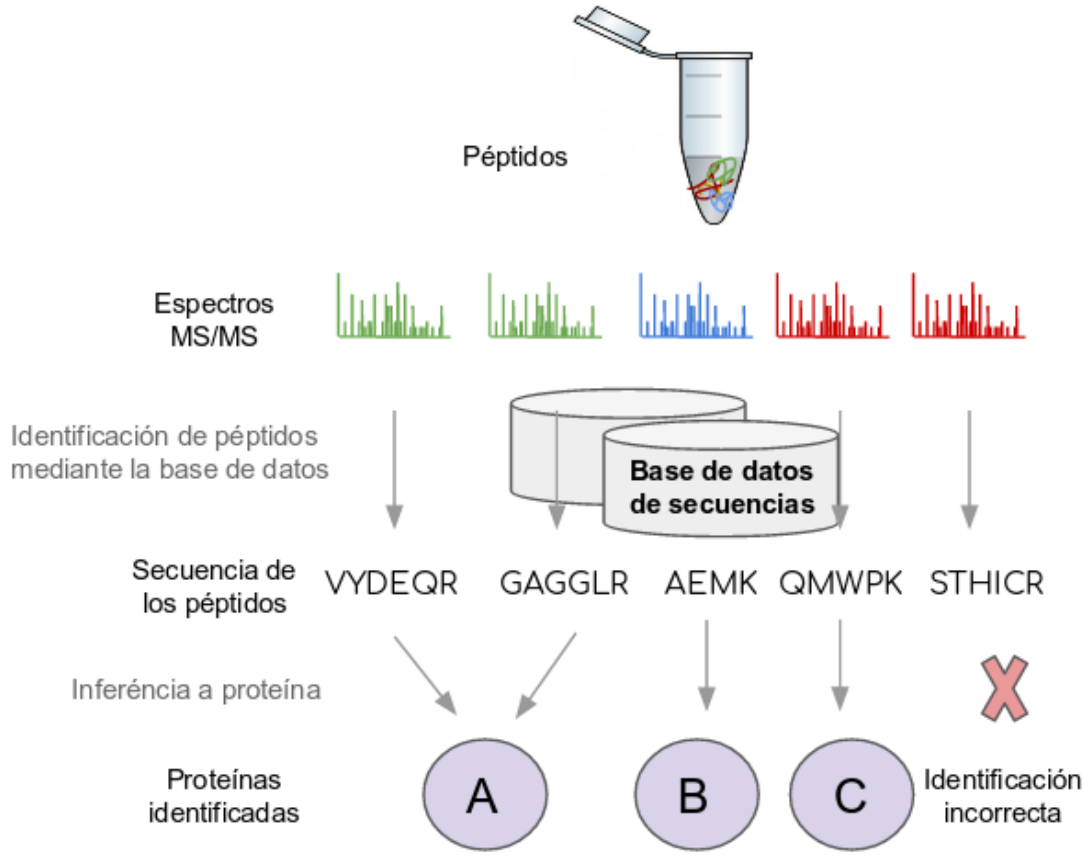


Figura 11. Proceso general para la identificación de proteínas. La mezcla de péptidos es analizada por el espectrómetro de masas obteniendo los espectros *MS/MS* de cada uno de ellos. A partir de estos espectros y los motores de búsqueda se asignan los péptidos que serán validados estadísticamente y aquellas identificaciones incorrectas se descartarán del análisis. Las secuencias obtenidas de los péptidos son utilizadas para hacer la inferencia a proteína.

1.2.3.2.1. Formato de los ficheros

Existen muchos formatos de ficheros para representar los datos obtenidos del espectrómetro de masas. Estos podrían dividirse en dos grandes grupos: Los que contienen los datos crudos que suelen ser dependientes de la instrumentación utilizada, y los formatos que contienen los datos procesados, que han sido creados

para tener y poder extraer los datos de una forma más sencilla. Ejemplos de datos del primer grupo son los de formato *dat* (*Bruker*), raw (Thermo Xcalibur), wiff (ABI/Sciex), etc. que están asociados al software utilizado [50]. Los archivos con formato *mzXML* y *mzML* son independientes del origen o software utilizado en su procesamiento y se crearon para intentar estandarizar todos los datos proteómicos [52]. Es necesario conocer el formato de los datos y como estos están organizados para proceder con el análisis. Afortunadamente, hoy en día existen programas como *Msconverter* [53] para poder cambiar de un formato a otro.

1.2.3.2.2. Asignación péptido-espectro

Como hemos comentado anteriormente existen muchas herramientas conocidas como motores de búsqueda para la asignación péptido-espectro (PSM). Todas actúan de forma similar: toman un espectro experimental de *MS/MS* como entrada y lo comparan con patrones de fragmentación teóricos construidos a partir de la base de datos.

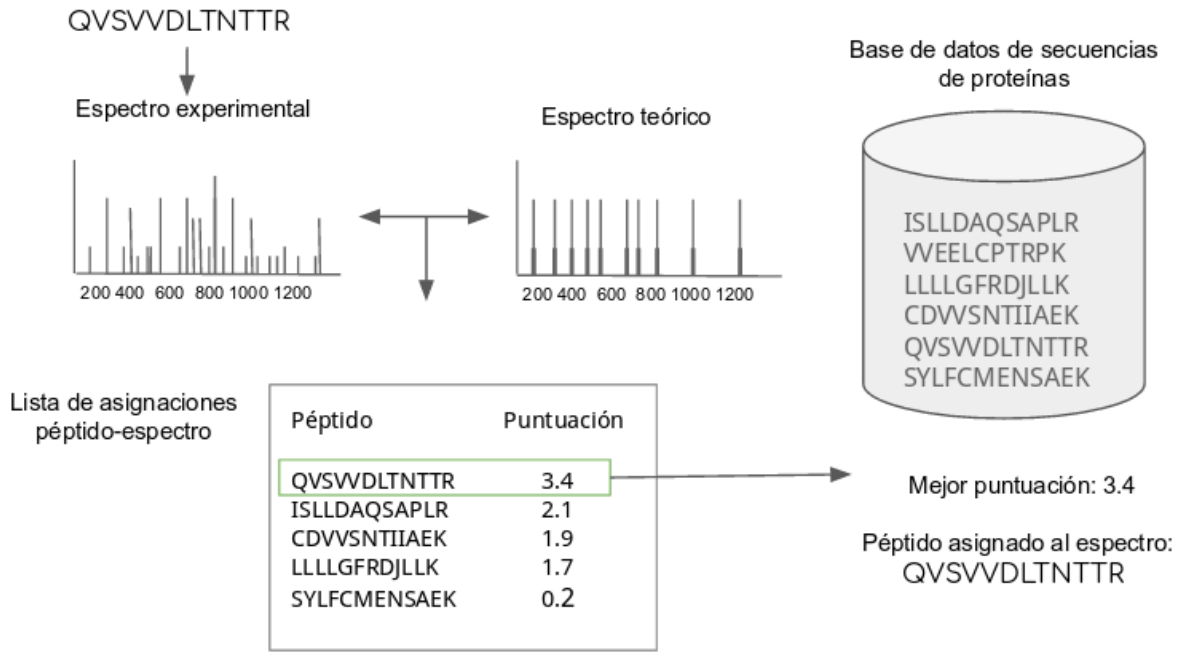


Figura 12. Esquema gráfico de cómo actúan los motores de búsqueda. El espectro experimental adquirido es comparado con los espectros teóricos generados a partir de la base de datos. El motor de búsqueda asigna una puntuación según la similitud entre los espectros, y los péptidos candidatos resultantes son ordenados según su puntuación. El péptido con mayor puntuación es seleccionado para los siguientes análisis.

La búsqueda puede realizarse bajo ciertos criterios especificados por el usuario como la tolerancia de masa, la enzima proteolítica utilizada y los tipos de modificaciones post-traduccionales. Entonces los patrones de

fragmentación son calculados teniendo en cuenta estas especificaciones resultando en una lista de coincidencias (secuencias peptídicas) clasificadas de acuerdo al sistema de puntuación propio de cada motor de búsqueda. Este sistema de puntuación representa el grado de similitud entre el espectro experimental y el espectro de fragmentación teórico. El resultado con mejor puntuación será el péptido más probable correspondiente al espectro analizado (Figura 12).

1.2.3.2.3. Validación estadística por FDR

Como en otros experimentos de alto rendimiento el resultado típico de un experimento de Proteómica es una lista de inferencias, generalmente en forma de asignación espectro-péptido (*PSM*). Las puntuaciones asignadas a cada inferencia son un reflejo de la confianza de la asignación. Por lo general únicamente nos interesan las que tienen mejor puntuación, pero para ello necesitamos establecer un umbral de corte [54]. Una forma intuitiva de establecer el umbral es utilizando el *FDR* (*del inglés, False Discovery Rate*) o Tasa de Falsos Positivos, que es la fracción esperada de descubrimientos para los cuales la hipótesis nula es verdadera. La hipótesis nula indica que el péptido (o proteína) está identificado incorrectamente, y la hipótesis alternativa lo contrario, que la asignación es correcta [54].

En el contexto de la Proteómica, la *FDR* es la estimación de falsos positivos presentes en los resultados. Existen diferentes estrategias para estimarlo, pero la más utilizada es la utilización de bases de datos señuelo concatenadas a la base de datos real para realizar las búsquedas como proponen [55]. Esta base de datos contiene el mismo número de secuencias reales (*T, target*) y secuencias ficticias (*D, decoy*) que muchas veces consisten en las secuencias reales del revés o desordenadas. Este método asume que el número de secuencias señuelo que pasen el umbral establecido es el número de asignaciones falsas en los resultados bajo este umbral, teniendo en cuenta que, al concatenar las dos bases de datos el tamaño de ésta es el doble.

$$FDR = \frac{2\,x\,D}{(T + D)}$$

Calcular la *FDR* a nivel de *PSM*, péptido o proteína no es lo mismo. Aunque el objetivo principal de un experimento típico de *shoutgun* es evaluar el nivel de proteína en la muestra, las hipótesis de estos experimentos son testadas a nivel de espectro. El cálculo de *FDR* a nivel de proteína es más complicado por la variabilidad en la abundancia de las proteínas y la distribución no aleatoria de los péptidos en las proteínas, ya que no existe correspondencia uno a uno entre péptidos y proteínas. Otra forma para establecer el umbral de *FDR* es realizando dos búsquedas independientes consecutivas, una sobre la base de datos con las secuencias reales y otra con las secuencias señuelo. En estos casos la estimación del *FDR* resulta conservadora ya que al compararse toda la población de espectros con las secuencias señuelo se asignan espectros que podrían asignarse correctamente a una secuencia real, por lo que se pierde competencia en la asignación y se produciría una sobreestimación de *PSM* incorrectos [56].

1.2.3.2.4. Inferencia de proteínas a partir de péptidos

En la mayoría de estudios, el investigador está interesado en la identificación de proteínas en lugar de péptidos. Estos péptidos deben agruparse de acuerdo con la proteína a la que corresponden, y la confianza estadística debe recalcularse para ser a nivel de proteína. Sin embargo, existen varias dificultades que complican este proceso de ensamblar péptidos a proteínas.

Uno de los problemas a los que nos enfrentamos se representa en la Figura 13 donde se muestra qué péptidos identificados correctamente tienden a agruparse a un pequeño número de proteínas, pero existen péptidos identificados incorrectamente que se asignan a nuevas proteínas que no pertenecen a la muestra. Por lo tanto, el porcentaje de error en proteínas sería significativamente mayor que en péptidos [56].

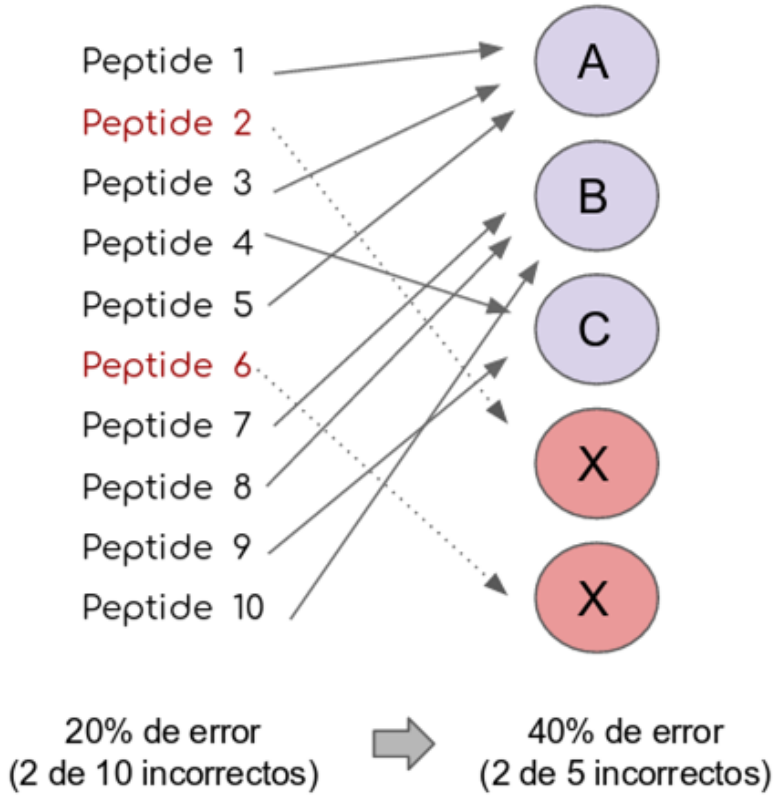


Figura 13. Inferencia de péptido a proteína. La figura representa el agrupamiento de péptidos a proteínas. De los 10 péptidos identificados dos identificaciones son incorrectas (20% de error). Estas dos identificaciones corresponden a dos proteínas distintas incrementando el error a un 40% a nivel de proteína.

Otra de las dificultades de la inferencia es cuando un péptido corresponde a más de una proteína de la base de datos, por lo que no podemos saber a qué proteína deberíamos asignar el péptido presente en la muestra. Estos casos a menudo resultan ser proteínas homólogas, variantes de *splicing,* o entradas redundantes de la

base de datos y son especialmente problemáticos cuando se trabaja con una muestra compleja de proteínas [56,57].

La mayoría de motores de búsqueda permiten al usuario ver los péptidos inferidos a las proteínas. Sin embargo, en estudios a gran escala muchas veces se trabaja con múltiples datos adquiridos y procesados en distintos tiempos y es complicado estudiar todos los casos. Actualmente existen herramientas computacionales para combinar las asignaciones de diferentes experimentos y obtener una lista única de proteínas identificadas. La herramienta más popular para ello es *ProteinProphet* [56] que tiene en cuenta las dificultades mencionadas anteriormente. Este programa a partir de las identificaciones de péptidos y sus probabilidades, calcula la probabilidad de que una proteína esté presente en una muestra combinando conjuntamente las probabilidades obtenidas de la identificación de los péptidos. Para ello, aquellos péptidos que son la única evidencia de que una proteína existe, son penalizados, pero no son excluidos, mientras que aquellos que se encuentran en proteínas en las que se han agrupado más péptidos se recompensan. Los péptidos compartidos se distribuyen entre todas sus proteínas, y se obtiene una lista mínima de proteínas que pueden explicar todos los péptidos observados. Además, *ProteinProphet* también colapsa las entradas redundantes de la base de datos en una única entrada, representando la identificación como un grupo de proteínas. Finalmente, el programa produce una lista de proteínas con sus péptidos identificados y sus probabilidades entre otra información.

1.2.4. Proteogenómica

La Proteómica S*hotgun,* descrita en el capítulo anterior, requiere de comparar los espectros *MS/MS* contra los espectros teóricos generados a partir de la base de datos de secuencias de proteínas. Esta estrategia asume que todos los productos génicos del genoma son conocidos y están contenidos en la base de datos utilizada. Sin embargo, esta premisa no es completamente cierta y conlleva a que existan péptidos que no puedan ser identificados, como péptidos que contengan mutaciones, representen nuevos *loci* codificantes de proteína o sean resultado de *splicing* o de polimorfismos no sinónimos [58] (Figura 14).

Existen algunas estrategias para identificar estos péptidos, pero la aproximación alternativa que cada vez está siendo más utilizada es la Proteogenómica. La Proteogenómica es un nuevo campo en el que se combinan la Genómica y la Proteómica. El término fue introducido por primera vez en 2004 y consiste en la generación de bases de datos de secuencias de proteínas o péptidos personalizadas con secuencias de interés para ser utilizadas por los motores de búsqueda.

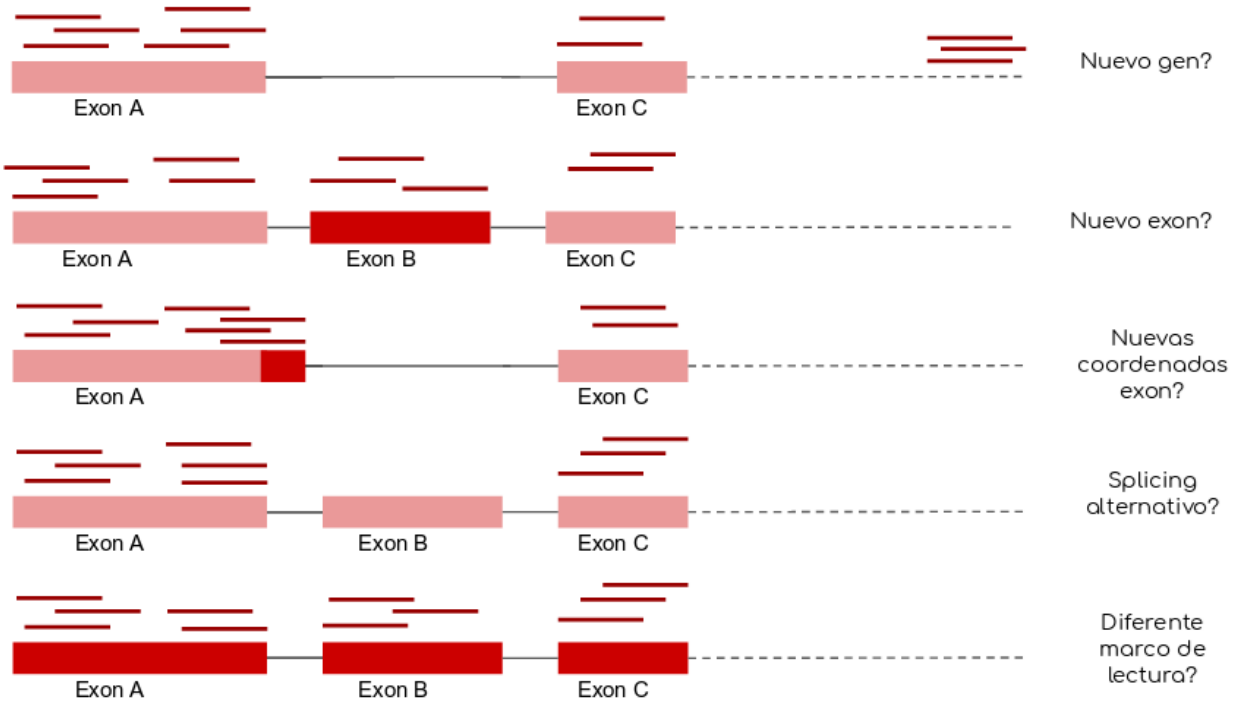


Figura 14. Posibles elementos a identificar con Proteogenómica. En rojo se representan los nuevos elementos que podrían identificarse si se contemplan las lecturas alineadas al genoma. Figura modificada de [58]

Estas bases de datos se pueden crear utilizando diferentes estrategias como: generar secuencias utilizando los 6 marcos de lectura de la secuencia genómica, predecir genes con metodologías *Ab initio*, utilizar un marcador de secuencia expresada o *EST (*del inglés, *expressed sequence tag) ,* utilizar transcritos de ARN anotados y traducirlos utilizando 3 marcos de lectura, incluyendo variantes albergadas en repositorios públicos como *dnSNP* (https://www.ncbi.nlm.nih.gov/SNP/) o a partir de datos de secuenciación del ADN o ARNm gracias al empleo de técnicas como el *RNA-seq* [59].

Muchas de estas metodologías propuestas incrementan el tamaño de la base de datos, lo que puede conllevar a un aumento del número de falsos positivos además de un incremento del tiempo computacional [49]. Este aumento del tamaño puede requerir de variaciones en el flujo de trabajo como tener que dividir la base de datos y realizar varias búsquedas independientes [49]. Por lo tanto, la base de datos resultante ideal debe ser un equilibrio en el que se incluyan entradas interesantes, pero tratando de no incrementar mucho el tamaño de la base de datos.

1.2.4.1. Herramientas para Proteogenómica

En los últimos años se han desarrollado varias herramientas que permiten la generación de estas bases de datos, algunas de ellas se detallan en la Tabla 2.

Las herramientas que se detallan en la tabla utilizan diferentes metodologías para la generación de las bases de datos. Según el objetivo del experimento y la anotación existente del organismo es más interesante la utilización de una o de otra. Sin embargo, muchas de ellas requieren de conocimiento bioinformático para poder ser utilizadas o estar trabajando con algún organismo que esté bien anotado en repositorios de referencia como *Ensembl* o *Refseq*. Por ejemplo, *CustomProDB* [60] es una de las herramientas más conocidas en proteogenómica y consiste en un paquete de *R* que permite generar bases de datos que incorporan variaciones o nuevas zonas de unión procedentes de datos de *RNA-seq* permitiendo obtener una base de datos específica del experimento. Sin embargo, ésta requiere de conocimiento del lenguaje de programación R para su utilización complicando el análisis.

Tabla 2. Herramientas desarrolladas para proteogenómica.

Herramientas	URL
CustomProDB	http://bioconductor.org/packages/release/bioc/html/customProDB.html
SpliceDB	http://www.softberry.com/spldb/SpliceDB.html
MSProGene	http://sourceforge.net/projects/msprogene/
The Proteogenomic Mapping Tool	http://www.agbase.msstate.edu/tools/pgm/
SpliceVista	https://github.com/yafeng/SpliceVista
Peppy	http://geneffects.com/peppy
PGTools	http://qcmg.org/bioinformatics/PGTools
QUILTS	http://openslice.fenyolab.org/cgi-bin/quilts_cgi_v2.0.pl
PGA	https://bioconductor.org/packages/release/bioc/html/PGA.html
ProteoAnnotator	http://www.proteoannotator.org/

1.2.5. Extracción de información biológica para la integración de datos

Los experimentos masivos de datos han posibilitado el estudio de la expresión de miles de genes y proteínas simultáneamente, esclareciendo así no solo su función individual, sino también las complejas interacciones que ocurren en el interior de una célula.

Tabla 3. Bases de datos biológicas

Tipo de Base de Datos	Base de datos	URL
Principales BD de secuencias de nucleótidos	NCBI (EEUU)	https://www.ncbi.nlm.nih.gov/
	EMBL (Europa)	https://www.ebi.ac.uk/
	DDBJ (Japón)	https://www.ddbj.nig.ac.jp
Algunas bases de datos de genomas de organismos concretos	Flybase (Drosophila)	http://flybase.org/
	SGD (Levadura)	https://www.yeastgenome.org/
	TAIR (Arabidopsis)	https://www.arabidopsis.org/
	ENSEMBL (Hombre, ratón y otros)	https://www.ensembl.org/index.html
Principales BD de proteínas	Uniprot	http://www.uniprot.org/
	Swiss-prot	http://www.ebi.ac.uk/swissprot/
	PDB	https://www.rcsb.org/
	SCOP	http://scop.mrc-lmb.cam.ac.uk/scop/
Bibliografía	Pubmed	https://www.ncbi.nlm.nih.gov/pubmed/
Rutas metabólicas	KEGG	http://www.kegg.jp/
Enfermedades genéticas humanas	OMIM	https://www.omim.org/

Las diversas bases de datos públicas son fuentes de información muy potentes que pueden enriquecer los resultados obtenidos por este tipo de análisis. Sin embargo, debido a la proliferación en los últimos años de estas fuentes de datos y al estar implicado no un reducido grupo de genes sino miles de ellos, es imposible acudir a ellas manualmente. Para ello, es necesario un estudio del tipo de datos disponible y el desarrollo de técnicas de recuperación, almacenamiento e integración automáticas que permitan extraer dicha información de las bases de datos e integrarlos a los resultados experimentales para obtener unos resultados de mayor calidad.

Las bases de datos biológicas contienen información de diferente índole abarcando los diferentes campos de la biología. Algunos recursos como Ensembl (www.ensembl.org) y NCBI (https://www.ncbi.nlm.nih.gov/) reúnen mucha información como secuencias de nucleótidos, proteínas, estructura de proteínas, genomas, expresión genética, bibliografía, taxonomía, metabolismo, factores de transcripción, etc. Pero además de estas también existen otras bases de datos más específicas que son muy utilizadas. Algunas de ellas se detallan en la Tabla 3.

1.2.5.1. Las bases de datos Ensembl, NCBI y Uniprot

La enorme cantidad de información biológica que se está acumulando en los últimos años se encuentra en diferentes bases de datos públicas. Incorporar esta información en los análisis es interesante para obtener unos resultados de mayor calidad.

La información básica sobre la que se asientan casi todas las herramientas y recursos en Bioinformática son las secuencias de nucleótidos y proteínas. Una vez es secuenciado el genoma, la información es procesada y anotada, empezando por la predicción de genes, proceso en el que se identifica qué fragmentos del genoma corresponden a genes. Con esta información ya es posible conocer las proteínas codificadas y diversa información estructural o funcional asociada.

Existen muchas bases de datos de diferentes tipos, sin embargo, en un ámbito más general los dos modelos de información biológica más usados son los soportados por *NCBI* en USA y por el *EBI* europeo. Estas bases de datos están conectadas a diversos recursos permitiendo al usuario acceder a toda la información disponible de un organismo de una forma clara y sencilla. Esta abarca desde la secuencia de nucleótidos del genoma hasta las variaciones encontradas o las secuencias de las proteínas de un organismo. Casi todas las bases de datos importantes contienen bases de datos curadas y de construcción automática. Las de construcción automática su crecimiento es muy rápido pero su contenido no es tan preciso; en cambio las bases de datos curadas crecen lentamente, pero ofrecen información fiable.

Centrándonos en las bases de datos genómicas o de nucleótidos del *NCBI*, *Genebank* (https://www.ncbi.nlm.nih.gov/genbank/) es la base de datos de construcción automática, mientras que *Refseq* (*The Reference sequence Database*, https://www.ncbi.nlm.nih.gov/refseq/) es su base de datos revisada y libre de redundancias. Ambas contienen información del genoma, tránscritos y proteínas en más de 100.000 organismos en *Genbank* y en 75,218 organismos en *RefSeq*. Cada registro de estas bases de datos está conectada a otras bases de datos pudiendo integrar información de hasta 40 bases de datos distintas. El acceso a esta compleja base de datos es posible gracias al sistema desarrollado *Entrez* [61] que mantiene y proporciona el acceso a toda esta compleja red de información.

Por otro lado, está *Ensembl* que es un proyecto conjunto entre EMBL- EBI y el instituto Sanger. Éste está constituido por *Ensembl* (*https://www.ensembl.org*) y *Ensembl Genomes* (http://ensemblgenomes.org/). *Ensembl* contiene 112 organismos vertebrados, mientras que *ENSEMBL Genomes* se divide en: *ENSEMBL Fungi*, *ENSEMBL Bacteria*, *ENSEMBL Metazoa* , *ENSEMBL Plants* y *ENSEMBL Protists* albergando a 45161 organismos [62]. La base de datos revisada y curada de *Ensembl* se denomina *Vega* (*Vertebrate Genome Annotation*) [63].

Respecto al proteoma, se puede decir que *Uniprot* (http://www.uniprot.org/) [64], mantenido también por *EBI* entre otros, es el repositorio de información por excelencia. Esta base de datos está dividida en dos: *Swiss-Prot* que consiste en la base de datos curada y revisada manualmente en la que no se encuentran redundancias y *TrEMBL* que contiene entradas analizadas computacionalmente y anotadas automáticamente. *TrEMBL* fue creada porque el incremento de datos generados por los proyectos de secuenciación era tal, que era imposible poder revisar todas las entradas al tiempo que se iban generando.

A cada gen, transcrito y proteína se les asigna un identificador único en cada base de datos. Ese índice de identificadores es el esqueleto que sostiene toda la información contenida en *Ensembl*, *Uniprot* y *NCBI*.

1.2.5.2. Bases de datos funcionales

Gracias a las diferentes técnicas experimentales utilizadas hoy en día, conocemos mucho mejor el papel en el que están involucrados los genes o proteínas en los diferentes organismos. Del mismo modo que con las bases de datos genómicas y proteómicas, han aparecido numerosos proyectos para tratar de cubrir diferentes aspectos de la biología. Existen muchas bases de datos diferentes que abarcan numerosos campos de la biología como interacciones de proteínas, regulación genética, información estructural, información funcional, etc. Las bases de datos funcionales más populares se podrían decir que son *Gene Ontology* y *Kegg*, que permiten anotar los genes con información biológica además de poder realizar análisis de enriquecimiento funcional.

1.2.5.2.1. Gene Ontology

Gene Ontology o *GO* (http://www.geneontology.org/) es un importante proyecto bioinformático que provee un vocabulario controlado, estructurado, bien definido y común que describe el papel de los genes y sus productos en diversos organismos [65–67].

El proyecto *Gene Ontology* consiste en tres ontologías independientes y accesibles públicamente. Cada una representa un concepto clave en la Biología Molecular y éstas son [65]:

- Función Molecular: Describen las funciones que llevan a cabo los productos génicos.

- Proceso Biológico: Se trata de los grandes procesos biológicos, como la mitosis o el metabolismo de las purinas, que son llevados a cabo por conjuntos ordenados de funciones moleculares.
- Componente Celular: Estructuras subcelulares, localizaciones y complejos macromoleculares.

Cada una de estas ontologías está formada por una serie de anotaciones con identificador alfanumérico, nombre común y definición (por ejemplo, *catabolic process (GO:0009056)).* A estas anotaciones se asignan los genes o proteínas correspondientes, permitiendo asociar a cada producto génico un contexto molecular, celular y de proceso biológico.

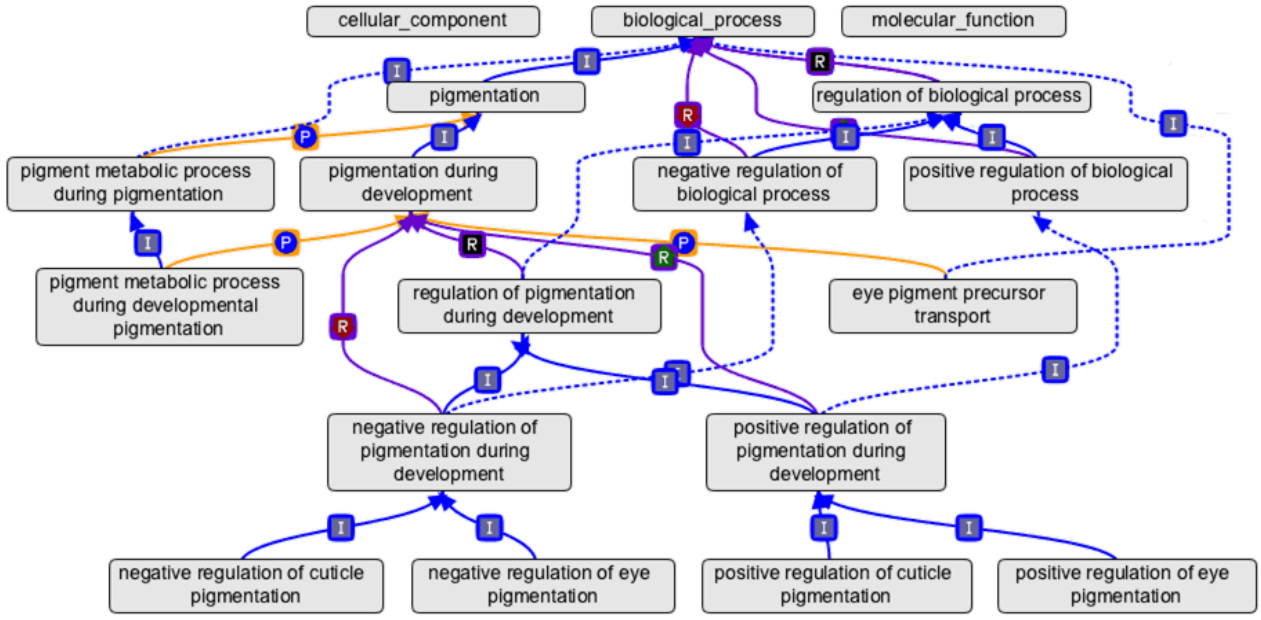


Figura 15. Estructura de Gene Ontology. La figura muestra algunos niveles de las anotaciones de los procesos biológicos. Imagen tomada de http://geneontology.org/page/ontology-structure

La estructura de GO se puede describir utilizando los términos de un grafo, dónde cada término GO es un nodo y las relaciones entre los nodos son las aristas. Se podría considerar una estructura jerárquica, donde los términos "*hijos*" son más específicos que los términos "*padres*", sin embargo, a diferencia de una estructura jerárquica estricta, un término puede tener más de un término "padre". Por ejemplo, en la Figura 15 se muestra que en procesos biológicos el término "*negative regulation of pigmentation during development*" tiene dos padres "*pigmentation during development*" y "*regulation of pigmentation during development*". Esto es debido a que dentro de un grupo pueden existir subtipos.

1.2.5.2.2. KEGG

Kyoto Encyclopedia of Genes and Genomes o *KEGG* (http://www.kegg.jp/) es una base de datos que se define como la representación computacional del sistema biológico [68]. Esta base de datos almacena conocimiento para comprender la información funcional de alto nivel de las células y organismos que forman parte del complejo genómico y molecular. *KEGG* se divide en cuatro bloques: *KEGG GENES* que alberga información de los genes y proteínas, *KEGG LIGAND* que contiene información de sustancias exógenas y endógenas, K*EGG PATHWAY* que consiste en diagramas moleculares de redes de interacción y reacción y *KEGG BRITE* que relaciona los diferentes bloques citados. *KEGG PATHWAY* es el módulo más utilizado y consiste en una colección de mapas dibujados manualmente representando el conocimiento experimental sobre el metabolismo y otras funciones de la célula y el organismo como procesos celulares, enfermedades humanas, desarrollo de fármacos, etc. Cada mapa contiene una red de interacciones y reacciones moleculares que vinculan los genes a sus productos génicos (principalmente proteínas), permitiendo conocer qué genes o proteínas participan en estos procesos (Figura 16)

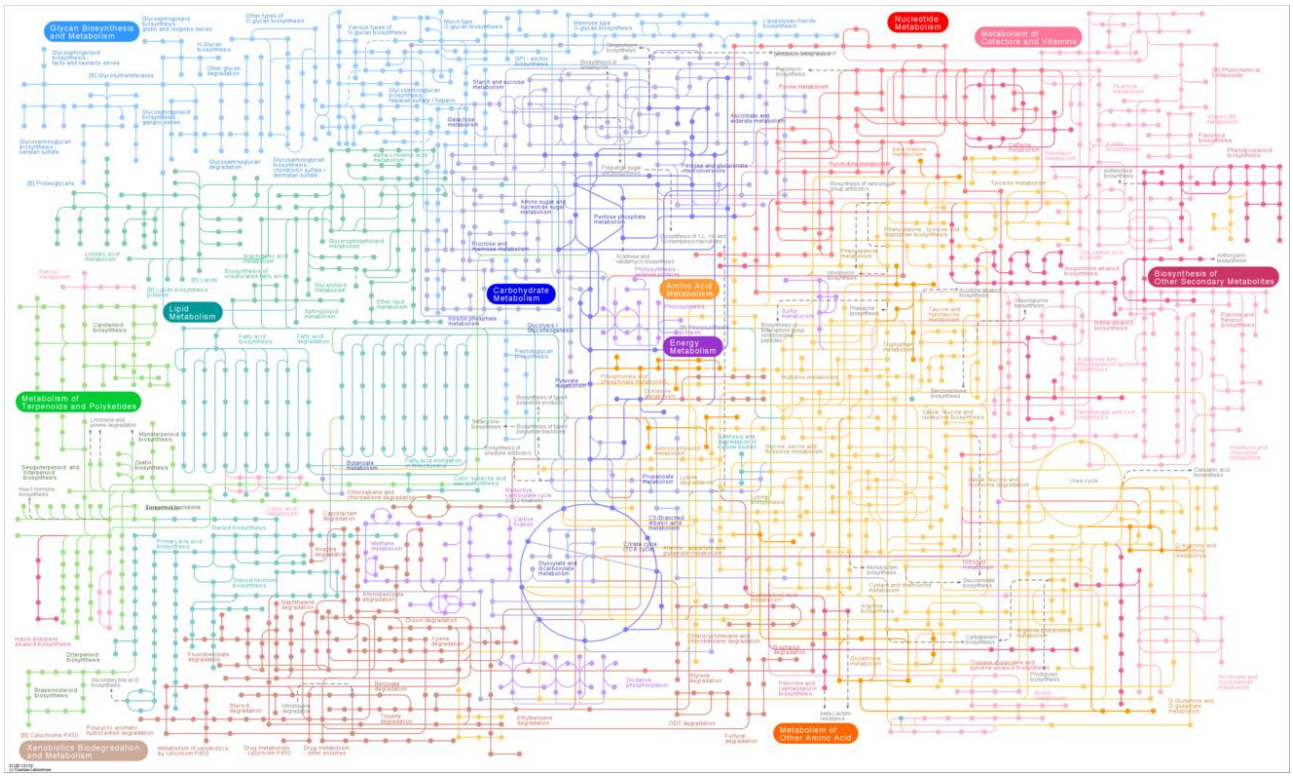


Figura 16. Diagrama de la ruta metabólica de Kegg

1.2.5.3. Repositorios de datos biológicos

Además de bases de datos de información biológica, también existen bases de datos que albergan los datos del experimento. Estos repositorios de datos son muy útiles porque permiten compartir y utilizar diferentes experimentos de la comunidad científica.

1.2.5.3.1. Repositorios de datos biológicos en Genómica

Los repositorios de datos biológicos genómicos son bases de datos que contienen experimentos de *micro-arrays*, de secuenciación masiva y otras formas de experimentos de alto rendimiento introducidos por la comunidad científica. Aproximadamente el 90% de los experimentos son datos de expresión génica que abarcan diferentes campos de la Biología.

GEO [69,70] y *ArrayExpress* [71] son los repositorios de datos biológicos más populares en la actualidad. De hecho, existe información duplicada en ambas bases de datos, ya que *ArrayExpress* importa sistemáticamente parte de los datos publicados en *GEO*.

Los datos almacenados en estos repositorios representan la investigación de la comunidad científica porque muchas revistas requieren que los datos del estudio estén disponibles públicamente. Esto permite que los datos puedan ser evaluados y/o reanalizados para ser utilizados en otros proyectos. Estos repositorios albergan diferentes tipos de datos que se generan en los análisis: datos crudos, procesados, metadatos, etc. Además de albergar los datos, también ofrecen herramientas como *GEO2R* o *ArrayExpress Bioconductor package* para facilitar la descarga, el análisis o la visualización de estos datos.

Es importante conocer la estructura de los datos para poder acceder a ellos y saber qué información contienen. Por ejemplo, en *GEO* existen principalmente dos tipos de datos: los que se almacenan en el formato enviado por los autores, y, por otro lado, conjuntos de datos procesados por expertos de *GEO*. Los datos enviados por los autores consisten en tres registros (Figura 17):

- **Plataforma (*GPL*)**

GPL es el registro que describe la plataforma utilizada, como, por ejemplo, qué *microarray* o secuenciador se ha empleado para el análisis, además de en el caso de los *microarrays*, una plantilla con las sondas o los genes utilizados para el mismo.

- **Muestras (*GSM*)**

Las muestras o *GSM* describe el protocolo experimental utilizado en la muestra y contiene los datos de la propia muestra. A cada una se le asigna un número de acceso *GEO* que empieza por *GSM* seguido de número y este debe tener asociado un *GPL y* puede estar relacionado con más de un *GSE*.

- **Series (*GSE*)**

Las series o *GSE* son agrupaciones de muestras que están relacionadas entre sí en un experimento. Estas van asociadas a una descripción completa del estudio, y como en los otros casos se le asigna *GSE* seguido de un número.

Figura 17. Esquema de la organización de los datos en GEO.

Dado que esta información suele ser muy heterogénea, en algunos casos *GEO* construye nuevos conjuntos de datos llamados *GDS.*

- **Conjuntos de datos (*GDS*)**

Los conjuntos de datos o *GDS* son datos que han sido extraídos, reanalizados y procesados según la descripción del experimento por expertos de *GEO*. Para ello, seleccionan un conjunto de muestras biológicas con una plataforma en común y reanalizan los datos creando un nuevo conjunto de datos con un formato homogéneo. A estos datos también se puede acceder utilizando el identificador único *GDS* seguido de un número.

En el caso de *Array Express* la organización de datos es similar a la de *GEO*. De una forma parecida, en esta base de datos la organización va en torno a los experimentos, describiendo los distintos ensayos que pertenecen al estudio.

1.2.5.3.2. Repositorios de datos biológicos en Proteómica

Como sucede con los datos genómicos, los resultados generados en Proteómica también son almacenados en repositorios públicos para ser compartidos con la comunidad científica. El repositorio oficial de datos biológicos en Proteómica es el consorcio *ProteomeXchange* (*PX*, http://www.proteomexchange.org/). Este consorcio se formó en 2006 pero no fue hasta el 2011 cuando se empezó a estandarizar la subida y difusión de los datos de espectrometría de masas a nivel mundial [72].

PX esté compuesto por los repositorios *PRIDE, PeptideAtlas, PASSEL, MassIVE* y *jPOST.* Éstos se formaron con distintos objetivos, por ejemplo, *PRIDE* representa la información igual que fue analizada por el investigador, mientras que *PeptideAtlas* reprocesa los datos con un flujo de trabajo común (*TTP*, *Trans-Proteomic Pipline*) para mostrar uniformidad en todos los resultados [73]. Este consorcio ha logrado unificar las diferentes bases de datos alojando los datos en un portal común.

1.2.6. Análisis de enriquecimiento funcional

En muchas ocasiones los resultados de experimentos realizados con tecnologías ómicas consisten en largas listas de genes o proteínas de interés. La interpretación de estos datos puede ser una tarea compleja que se puede facilitar con la realización de análisis funcionales. El análisis funcional de genes hace referencia a la anotación y análisis estadístico de listas de genes mediante métodos computacionales que identifican anotaciones funcionales con las que los genes están significativamente relacionados.

Estos tipos de análisis son muy utilizados y como consecuencia existe una gran variedad de herramientas que según [74] se clasifican en: análisis de enriquecimiento singular, análisis de enriquecimiento de conjuntos de genes y análisis de enriquecimiento integrativo y modular.

1.2.6.1. Análisis de enriquecimiento singular

El análisis de enriquecimiento singular fue el primer tipo de análisis que se desarrolló y se basa en comparar la frecuencia esperada de un término de anotación con la frecuencia esperada por azar. Para ello se aplica un test estadístico que suele ser χ^2 o el test de Fisher mediante una distribución binomial o hipergeométrica para determinar qué anotaciones están significativamente enriquecidas en la lista de entrada con respecto a la lista de referencia, que normalmente suelen ser todos los genes del genoma. El p-valor calculado representa la significancia de la asociación entre el término y la lista, aquellos términos individuales con, por ejemplo, un p-valor menor de 0.05 se consideran enriquecidos. En este contexto se suelen utilizar anotaciones de diferentes fuentes, unas de las más populares son *Gene Ontology* y *KEGG* [75]. Hoy en día existen muchas herramientas que hacen este tipo de análisis, un ejemplo son *GeneCodis* [75–77], *David* [78], *Panther* [79].

1.2.6.2. Análisis de enriquecimiento integrativo y modular

El análisis de enriquecimiento singular considera cada término independientemente, sin embargo, no todos los términos son independientes. Por ejemplo, los términos *GO* están relacionados entre sí, lo que puede llevar a la redundancia o creer que un concepto biológico está más enriquecido por obtener más términos. De hecho, esto puede suceder porque la fuente de anotación tenga procesos más anotados que otros, ya que estas fuentes son actualizadas constantemente.

Existen herramientas que consideran estas relaciones entre los términos como son *Ontologizer* [80], *David* [78] y *GeneCodis* [77] mejorando la sensibilidad y especificidad del análisis. Estos análisis se llaman análisis de enriquecimiento integrativo y modular y consisten en incorporar un segundo análisis al singular teniendo en cuenta las dependencias entre genes o proteínas inferidas de redes biológicas, grafos de ontologías o combinaciones de diversos tipos de anotaciones.

1.2.6.3. Análisis de enriquecimiento de conjuntos de genes

El análisis de enriquecimiento de conjuntos de genes incorpora los valores de expresión, valores de *Fold Change* o los p-valores de todos los genes para hacer un test de análisis de significación estadística. En este análisis se evalúa toda la lista de genes ordenados según su correlación con el fenotipo (datos de perfil molecular) en el contexto de la anotación, y no únicamente aquellos diferencialmente expresados o más interesantes para el estudio.

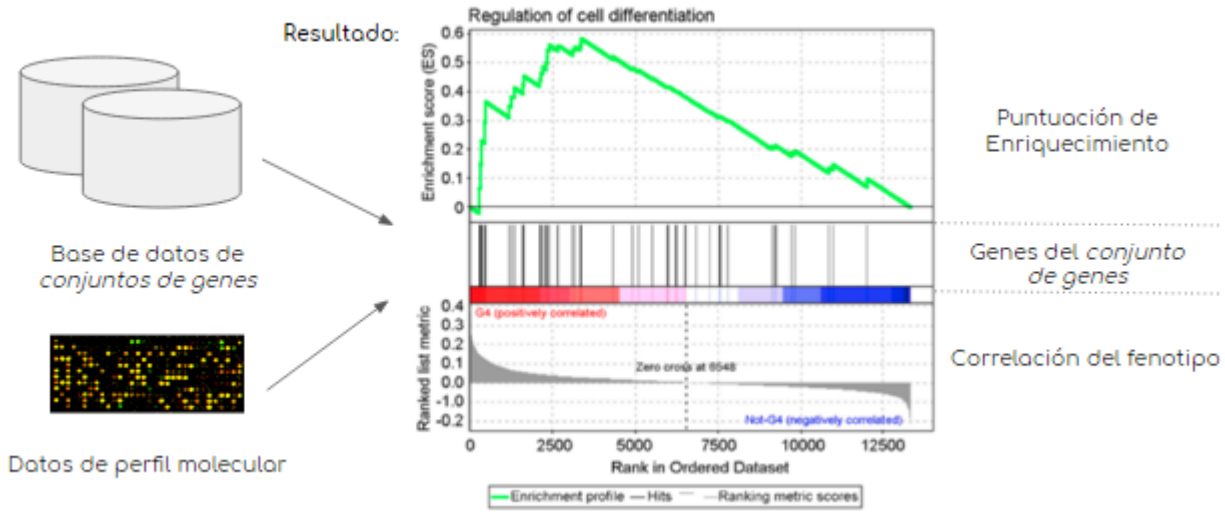


Figura 18. Análisis de enriquecimiento de conjuntos de genes en GSEA.

El análisis de enriquecimiento de conjuntos de genes o *GSEA*, evalúa los datos del análisis a nivel de conjuntos de genes. Los conjuntos de genes son definidos basándose en el conocimiento biológico, como, por ejemplo, información publicada sobre vías metabólicas. El objetivo de *GSEA* es determinar si los miembros de un conjunto de genes tienden a aparecer en los extremos de la lista (parte superior o inferior), y así determinar si el conjunto de genes se correlaciona con uno de los fenotipos [81]. La herramienta más popular es *GSEA* del *Broad Institute* [81], pero existen otras como *GeneTrail* [82], *FatiScan* [83], etc. (Figura 18)

1.2.7. Conversión de identificadores de genes o proteínas

Como hemos estado comentando, el conocimiento biológico se encuentra actualmente esparcido en bases de datos diferentes e independientes. Cada una de estas bases de datos pueden contener identificadores de genes propios. Para poder integrar los datos de diferentes fuentes, o bien poder utilizar herramientas desarrolladas para la realización de diferentes análisis, es necesario convertir los identificadores a uno común y/o a uno reconocible para la herramienta que se vaya a utilizar.

Un ejemplo en el que es necesaria la conversión de identificadores es en los *microarrays*, para cuyo análisis se deben traducir los identificadores de las sondas previamente a un identificador conocido como *Official Gene Name*, *Genbank*, *RefSeq*, *UniProt*, etc. Este proceso es lento y tedioso, y aún más importante, suele llevar a una traducción incompleta o inexacta y en consecuencia perder información. Por ello se han creado herramientas que permiten esta conversión como *DAVID Gene ID Conversion Tool* [84] o *Biomart* [85] que permiten convertir identificadores de diferentes organismos de forma sencilla.

2. OBJETIVOS

En la introducción de este trabajo se presentan múltiples técnicas experimentales utilizadas en las ciencias ómicas. Es fácil suponer que la integración de todas ellas representa el escenario ideal para poder entender la complejidad biológica de los seres vivos. Sin embargo, en términos prácticos esta integración no es nada simple y solo es posible mediante la integración de distintas disciplinas científicas. La bioinformática intenta ser un punto central en este contexto y en particular esta complejidad en la integración ha constituido una de las motivaciones centrales de este trabajo.

Por lo tanto, el objetivo general de esta tesis doctoral es el desarrollo y aplicación de métodos bioinformáticos para el análisis de datos biológicos procedentes de diversas plataformas, así como su integración y aplicación para obtener una visión global de los genes, proteínas y procesos biológicos alterados. Estos métodos se aplicarán tanto a datos procedentes de diferentes laboratorios para responder a preguntas científicas concretas como a datos existentes en repositorios públicos para crear herramientas tales como las destinadas al reposicionamiento de fármacos.

Estos objetivos se traducen de una manera detallada en:

- Elaboración de diferentes flujos de análisis automáticos o semi-automáticos para estudiar datos biológicos procedentes de diferentes técnicas ómicas, tales como datos de secuenciación masiva (*RNA-Seq)* o espectrometría de masas; utilizando datos de múltiples laboratorios o abordando datos procedentes de repositorios públicos comparando varias muestras de forma simultánea.

- Desarrollo de una metodología para identificar el interactoma de las células madre hematopoyéticas HSC mediante la integración de datos biológicos procedentes de diversas fuentes de información.

- Desarrollo de una herramienta bioinformática para el reposicionamiento de fármacos a través de la comparación de perfiles de expresión de diferentes bases de datos públicas. Asimismo, se desarrollarán herramientas complementarias e interfaces para poder visualizar y comparar este tipo de experimentos de una forma sencilla e intuitiva.

- Elaboración de una herramienta para la generación de bases de datos Proteogenómicas que incorporen variaciones y nuevas uniones a las bases de datos de referencia utilizadas para realizar las búsquedas de *MS/MS* e identificar nuevas proteínas.

Cada uno de estos objetivos será desarrollado en la sesión de aportaciones principales. Para facilitar su lectura y coherencia, estos serán presentados siguiendo una estructura completa de introducción, metodología, resultados y discusión. Cabe destacar que casi todos los desarrollos de los objetivos de esta tesis llevan asociados resultados biológicos significativos. Estos se describen en detalle en cada uno de los apartados de aportaciones.

3. APORTACIONES PRINCIPALES

3.1. Metodología bioinformática para la identificación de reguladores involucrados en la expansión y migración de células madre hematopoyéticas.

Hoy en día se está constatando que las células de un organismo multicelular interactúan con su entorno celular inmediato adaptando el entorno a sus necesidades. Esta afirmación es cierta para las células madre y el nicho en el que residen, como en el caso de las *HSC* (del inglés, *hematopoietic stem cells*) [86,87]. Esto también se ha visto en el desarrollo de tumores, donde las células del tumor influyen en el nicho cambiándolo para que favorezca el mantenimiento y la progresión del tumor, y así no apoyar el mantenimiento celular normal. A pesar de que existan trasplantes de células madre todavía no se conoce mucho sobre cómo éstas interactúan [88–90] y esta información es necesaria para mejorar la efectividad de los fármacos y encontrar nuevos tratamientos.

Este proyecto se llevó a cabo conjuntamente con el *Stem Cell Institute de la Katholike Universitet de Leuven* y propone analizar los niveles de ARN celular para identificar los perfiles de expresión del transcriptoma en las células *HSC* y el nicho, y a partir de los mismos derivar las posibles redes de interacción y señalización. Este análisis interactómico a partir del transcriptoma se extendió a las distintas fases del desarrollo. En concreto, esta propuesta se basa en el desarrollo de métodos bioinformáticos para detectar y evaluar las posibles relaciones entre las *HSC* y los nichos durante el desarrollo en el hígado fetal.

3.1.1. Células madre

El concepto de células madre ha atraído un gran interés debido a una serie de acontecimientos durante la última década, como el desarrollo de medios para mantener y expandir las células madre embrionarias humanas, o el descubrimiento de pequeñas poblaciones de células madre en varios órganos del organismo desarrollado que tienen la capacidad de regenerar todas las diferentes células del tejido [91].

Las células madre se definen por su habilidad de auto-renovarse, que significa dejar como mínimo una copia idéntica de ellas mismas después de cada división celular, y de diferenciarse en todas las células maduras del órgano o tejido al que pertenecen. El concepto de células madre fue propuesto por Till y McCulloch en el año 1960 cuando demostraron la existencia de células madre multipotentes mientras estudiaban la regeneración del sistema sanguíneo *in vivo* [92].

Las células madre se clasifican por la diversidad de su progenie en: totipotentes (las que tienen habilidad para dividirse y producir todas las células diferenciadas de un organismo), pluripotentes (capaces de formar todos los tejidos del cuerpo), multipotentes (capaces de crear los múltiples linajes celulares que constituyen un órgano o tejido), oligopotentes (aquellas que su progenie únicamente pertenecen a dos o más linajes dentro de un tejido) y unipotentes (que sólo contribuyen a un linaje) [93] (Ver Figura 19).

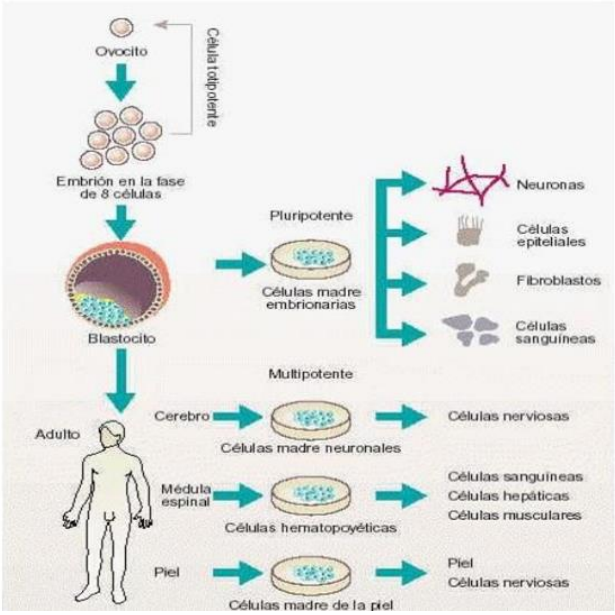


Figura 19. Tipos de células madre. Las células madre totipotentes tienen la capacidad de desarrollar un embrión completo; las pluripotentes pueden dar origen a cualquier tipo celular y las multipotentes dan origen a los distintos tipos celulares del órgano o tejido del que proceden. Figura tomada de http://www.dfarmacia.com/farma/

Las células que forman el cigoto son las células totipotentes y tienen el potencial de convertirse en cualquier tipo de célula. Días después de la fecundación, las células totipotentes se especializan originando las células madre pluripotentes o también llamadas células embrionarias que dan origen a todos los tejidos del cuerpo. En el organismo adulto encontramos el resto de tipos de células madre. Existen algunos tipos celulares como las células madre mesenquimales y hematopoyéticas capaces de diferenciarse en más de un tipo celular, o las células madre específicas de tejido (unipotentes) que tienen la función de renovarse o regenerarse si se produce algún daño tisular.

3.1.1.1. Células madre hematopoyéticas

El sistema sanguíneo es utilizado como paradigma para entender las células madre de tejido, su biología, y su participación en el envejecimiento, las enfermedades y la oncogénesis.

Las *HSC* se encuentran en la médula ósea, la sangre periférica y la sangre del cordón umbilical. Son células multipotentes capaces de producir precursores sanguíneos para cada linaje, que se diferenciarán y producirán células sanguíneas maduras, incluyendo glóbulos rojos, megacariocitos, células mieloides (monocito, macrófago y neutrófilo) y linfocitos [94]. Estas células sanguíneas maduras suelen tener una vida corta, por lo tanto, se requiere de las células hematopoyéticas para reponer las células progenitoras de los diferentes linajes sanguíneos [95]. Las células madre hematopoyéticas se dividen en dos grupos en función del tiempo que mantienen la capacidad de auto-renovación. Células capaces de auto-renovarse indefinidamente (*long term subset*, *LT*) o células con esta capacidad durante aproximadamente 8 semanas (*short term subset*, *ST*).

Las células con una capacidad limitada de auto-renovación, se diferencian en progenitores multipotentes. Estos tienen la capacidad de diferenciarse en múltiples pero limitados tipos celulares para poder diferenciarse en diferentes progenitores, como los progenitores mieloide o linfoide, y obtener finalmente una progenie diferenciada funcionalmente a través de un proceso de maduración irreversible [87,91].

3.1.1.2. Desarrollo de las *HSC*

El conjunto inicial de las *HSC* se forma durante la embriogénesis, en un procedimiento complejo que envuelve diferentes espacios anatómicos (el saco vitelino, la región de la aorta-gónada-mesonefros, la placenta y el hígado fetal), colonizando finalmente la médula ósea al nacer (Ver Figura 20) [95].

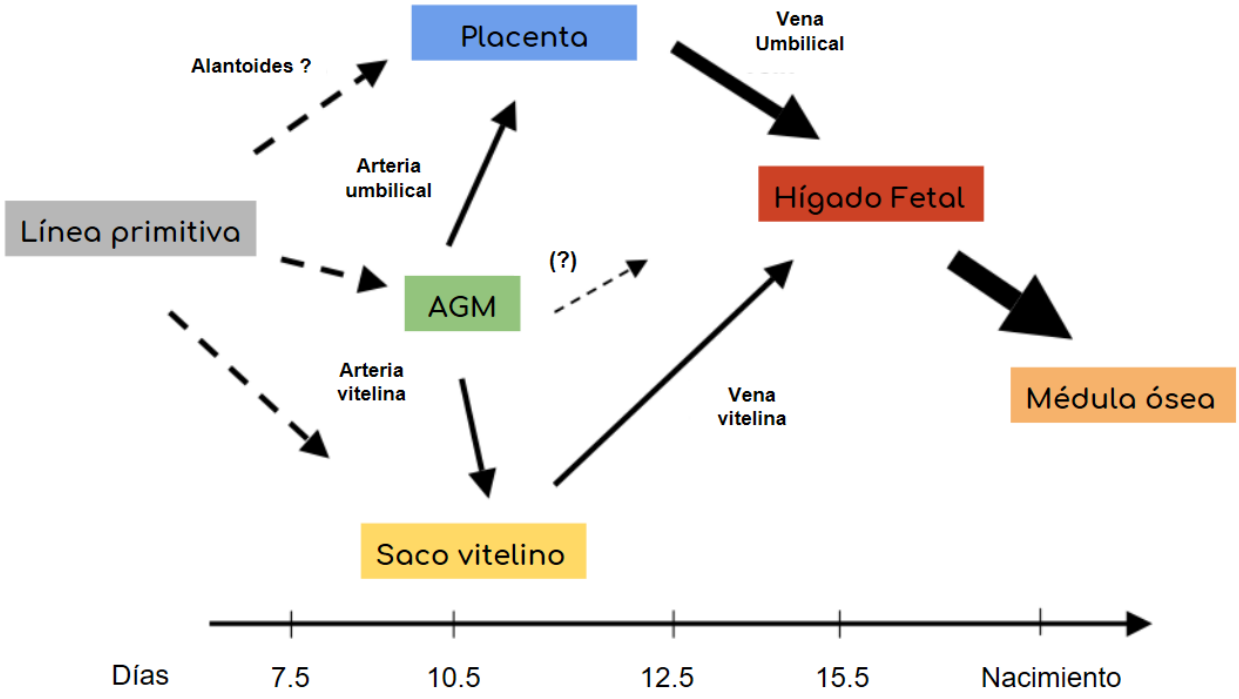


Figura 20. Rutas migratorias y circulatorias que conectan los sitios hematopoyéticos fetales en ratón. De la línea primitiva migran al saco vitelino (amarillo), a la región de la aorta-gónada-mesonefros (verde) y a la placenta (azul). Se cree que la especificación de las HSC se produce durante la migración. Alrededor del estadio E11-12.5 el hígado fetal (rojo) es colonizado, y después se establecen en la médula ósea (naranja) (Figura tomada y modificada de [95]).

Durante este proceso, las *HSC* que se están desarrollando migran a través de distintos nichos embrionarios los cuales les proveen señales para que establezcan y conserven su habilidad de autorrenovarse y diferenciarse [87,96]. Los nichos son lugares exclusivos de los tejidos que protegen y regulan las células madre a través de mecanismos biofísicos, bioquímicos y celulares [90]. Es concebible que el micro-ambiente, especialmente en el hígado fetal (*Feltal Liver, FL*), se modifique durante el desarrollo con el fin de satisfacer las necesidades que se requieren para la diferenciación en distintos linajes y la expansión de las *HSC* [97].

El hígado fetal es el órgano hematopoyético primario y el sitio más importante para que las células se expandan y se diferencien. La primera etapa del hígado fetal empieza en el estadio E11-12.5, cuando el hígado es colonizado por las *HSC*s que han migrado desde el saco vitelino, la región aorta-gónada-mesonefros y la placenta a partir de la vena umbilical y la vena vitelina [98].

Durante la embriogénesis, las *HSC* no solo consiguen obtener una progenie madura, sino que también se expanden simétricamente para crear un conjunto adecuado para la vida postnatal. La expansión de las *HSC* en el hígado fetal consigue el punto más alto, un máximo de alrededor 1000 *HSC*, en el estadio E15.5-E16.5, después, poco a poco pierde su capacidad de proliferación [95].

Finalmente, antes del nacimiento, la localización de la hematopoyesis cambia del hígado fetal a la médula ósea alrededor del estadio 18.5 donde se establece durante la época adulta y entra en el estadio G0 del ciclo celular [95,99]. La elección de la residencia no es casual, se establecen en la médula ósea ya que los alrededores del hueso proporcionan una protección física a agresiones externas, y los elementos celulares de la médula ósea nutren y mantienen las *HSC*. La migración de las *HSC* a la médula ósea (*Bone marrow*, *BM*) es comúnmente conocido como *homing* [100].

3.1.1.3. Aplicaciones médicas de las *HSC*

La capacidad única de las *HSC* de auto-renovarse y restaurar el sistema linfohematopoyético ha sido utilizada en varias aplicaciones médicas. Al principio las *HSC* se utilizaron únicamente para reconstruir médula ósea dañada pero los tratamientos han evolucionado no centrándose únicamente en tratamientos de enfermedades de la sangre, y actualmente son utilizadas para trastornos de todos los órganos del cuerpo humano [101].

Este gran interés en el campo de las células madre es debido a que su gran potencial terapéutico puede tener importantes aplicaciones en la ingeniería de tejidos, medicina regenerativa, terapia celular y terapia génica [102]. El concepto de terapia celular es el de reparar, sustituir o complementar células enfermas por células sanas. Actualmente existe la forma de suministrar células madre que puedan diferenciarse en nuevas células o tejidos sanos para aliviar o incluso curar una amplia gama de condiciones difíciles de tratar [103]. Se suelen tratar varias enfermedades con trasplantes de *HSC* como las que se muestran en la Tabla 4. Sin embargo, hay mucho por investigar sobre las características específicas de las células madre, como los mecanismos involucrados en la diferenciación y reparación para poder probar la eficacia y seguridad de nuevas terapias basadas en las células madre.

Tabla 4. Enfermedades comúnmente tratadas con transplantes de *HSC.* (Información obtenida de [104,105])

Transplante	Tipo	Enfermedad
Transplante Autólogo	Cáncer	Mieloma multiple Linfoma no Hodgkin Leucemia mieloide aguda Neuroblastoma Cancer de ovarios Tumores de células germinales
	Otras enfermedades	Transtornos autoimunes Amiloidosis
Transplante Alogénico	Cáncer	Leucemia mieloide aguda Leucemia linfoblastica aguda Leucemia mieloide cronica Linfoma no Hodgkin Enfermedad Hodgkin Síndromes mielodisplásicos Trastornos mieloproliferativos Leucemia linfocítica crónica Mieloma multiple
	Otras enfermedades	Anemia aplástica Anemia de Fanconi Anemia de Diamond-Blackfan Talasemia Anemia falciforme Inmunodeficiencia combinada severa Sindrome de Wiskott-Aldrich Error congénito del metabolismo

3.1.2. Metodología bioinformática

Como parte de este trabajo desarrollamos una metodología bioinformática para detectar y evaluar las posibles relaciones entre las *HSC* y los nichos durante el desarrollo en el hígado fetal. La primera etapa de la metodología consistió en analizar los transcriptomas correspondientes a los diferentes tiempos para determinar y cuantificar los niveles de expresión, mientras que la segunda etapa se basó en predecir las posibles interacciones y conocer los reguladores intrínsecos involucrados, para saber si estas relaciones juegan un rol determinante en la expansión o migración en las *HSC*.

Estas relaciones se estudiaron basándonos en el esquema dela Figura 21. Primero nos centramos en aquellos elementos que se están expresando, y con esa información estudiamos si existen evidencias de que las proteínas secretadas del nicho interaccionen con algún receptor de las *HSC* del hígado fetal. A continuación,

observamos qué elementos están involucrados en la interacción y qué efecto conllevan para evaluar su relevancia en el estudio.

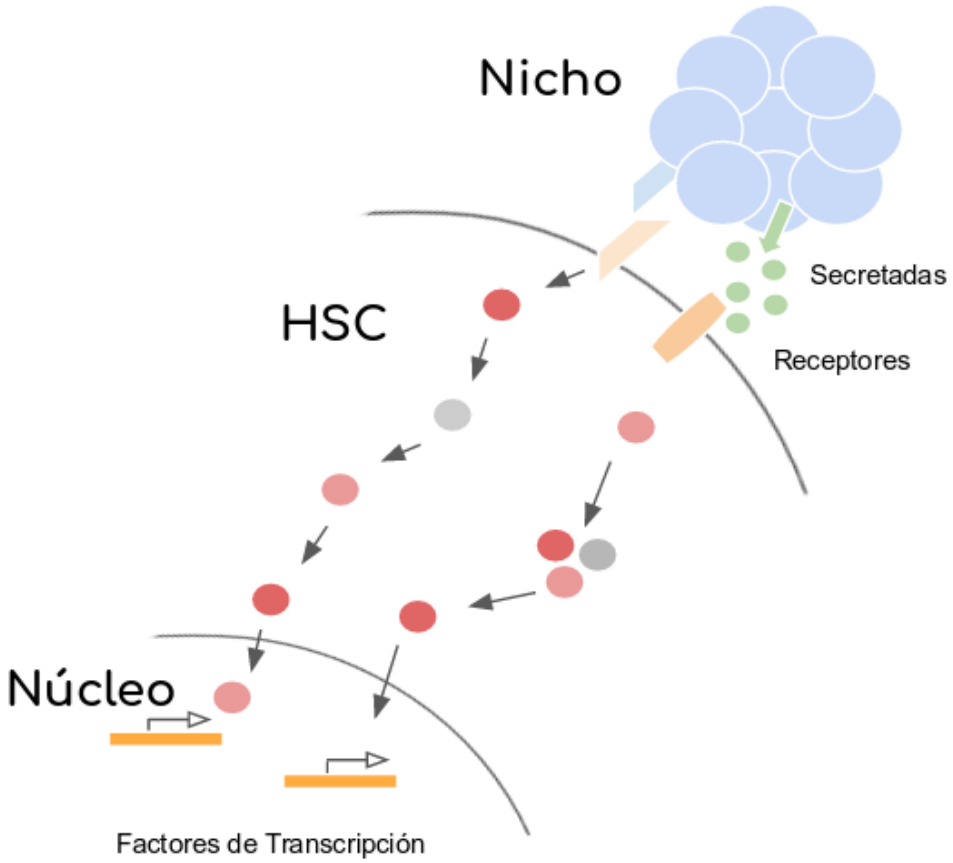


Figura 21. Figura esquemática del proceso a estudiar. Las proteínas secretadas por el nicho interaccionan con receptores de las células hematopoyéticas activando una cascada de señalización que involucran factores de transcripción que participan en rutas relacionadas con la expansión y migración de las *HSC*.

Este proceso fue realizado en las distintas fases del desarrollo. Se obtuvieron muestras del hígado fetal y de los nichos en diferentes días del desarrollo embrionario (12.5, 14.5, 16.5 y 18,5). Las relaciones se establecieron según la Figura 22. Estudiando posibles relaciones entre el nicho FL12.5 con *HSC* FL 12.5 y *HSC* FL14.5, el nicho FL14.5 con HSC FL 14.5 y *HSC* FL16.5 y el nicho FL16.5 con *HSC* FL 16.5 y *HSC* FL18.5. Estas etapas fueron llevadas a cabo mediante el desarrollo de metodologías bioinformáticas existentes que detallaremos en los capítulos siguientes y que fueron codificadas en el lenguaje *Perl* para este proyecto.

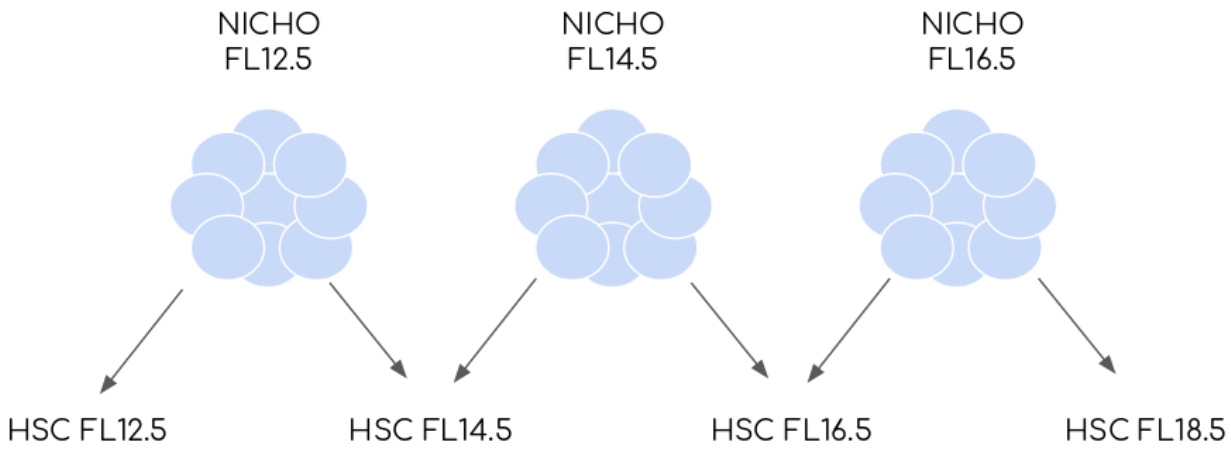


Figura 22. Relaciones entre HSC del hígado fetal y los nichos a estudiar. El nicho FL12.5 interactúa con FL12.5 y FL14.5; el nicho FL14.5 interactúa con FL14.5 y FL16.5 y por último el nicho FL16.5 interactúa con FL16.5 y FL18.5.

3.1.2.1. Muestras

Se obtuvieron 11 muestras de diferentes estadios y localizaciones (nicho o *HSC*) de ratón (*Mus Musculus*) por microdisección con láser. Esta última es una técnica que permite aislar células de interés de regiones microscópicas en tejidos celulares.

Ocho de las once muestras pertenecen a las *HSC* en distintos estadios del desarrollo embrionario en el hígado fetal (E12.5, E14.5, E16.5 y E18.5) con una réplica para cada muestra. Cada estadio se refiere a los días de desarrollo del feto. Para el estudio de la relación de los microambientes con el desarrollo de las *HSC*, se obtuvieron muestras de los nichos en los estadios E12.5, E14.5 y E16.5 (Tabla 5).

Tabla 5. Muestras utilizadas para el estudio

Muestra	Días de desarrollo embrionario	Número de replicas
Hígado Fetal HSC	E12.5	2
Hígado Fetal HSC	E14.5	2
Hígado Fetal HSC	E16.5	2
Hígado Fetal HSC	E18.5	2
HSC nicho	E12.5	1
HSC nicho	E14.5	1
HSC nicho	E18.5	1

Estas muestras se secuenciaron mediante la tecnología *RNA-seq* que permite revelar la presencia y cantidad de ARN en una muestra biológica en un momento dado y así estudiar los genes que se están expresando. El análisis de *RNA-seq* consiste en diversos pasos, empezando por la preparación de una librería específica para la secuenciación, hasta el posterior análisis bioinformático que detallamos en los siguientes puntos.

3.1.2.2. Análisis de datos

Utilizando el secuenciador *Illumina HiSeq2000* se generaron lecturas *paired-end* (secuenciadas por ambos extremos*)* y *strand specific* (que tienen en cuenta la orientación de las lecturas). Los ficheros crudos obtenidos del secuenciador (ficheros *FASTQ*) contienen la secuencia biológica junto a las calidades asociadas a cada nucleótido de dicha secuencia.

3.1.2.2.1. Calidad de las secuencias

El primer paso a realizar en todo análisis de secuenciación masiva es comprobar la calidad de las secuencias, y determinar si es necesario reducir el ruido que contienen las lecturas debido a la secuenciación o a alguna contaminación en la preparación de la muestra. Es posible comprobar la calidad porque los sistemas de secuenciación estiman la probabilidad del que los nucleótidos identificados sean erróneos. Esta estimación se calcula utilizando la escala *Phred* que se define:

$$Phred = -10\log(prob\ error)$$

Y es específica de cada tecnología y la calcula el software del equipo. La empresa BGI (*Beijing Genomics Institute*) que es la que secuenció las muestras, previamente a la entrega de los resultados realiza unos pasos de filtrado que consisten en: eliminar las lecturas con adaptadores, eliminar las lecturas en que las bases no identificadas son superiores al 5% y eliminar las lecturas con más de un 30% de la secuencia con bases de baja calidad (una base con baja calidad es aquella que la calidad de secuenciación no es superior a 10 en escala *Phred*). Aunque las lecturas fueron filtradas por la compañía de secuenciación, comprobamos con la herramienta *FASTQC* [14] que la calidad fuese la deseada para poder proceder al alineamiento con el menor ruido en las lecturas.

3.1.2.2.2. Alineamiento de las lecturas

El siguiente paso fue mapear las lecturas contra el genoma de referencia mm9 utilizando la herramienta *TopHat* [4,18]. *TopHat* es un alineador que se basa en el método de la transformada de *Burrows-Wheeler*. Para el alineamiento se empleó el genoma de referencia mm9 de *UCSC (University of California at Santa Cruz)* y las opciones:

- *-r*: Para especificar la distancia media interior esperada entre los dos fragmentos de *paired-end*.

- *-p:* Número de *cores* utilizados por el programa
- *--no-discordant*: reporta únicamente aquellos fragmentos que tengan su *paired-end*.
- *--no-mixed*: reporta aquellos alineamientos en los que los dos fragmentos de *paired-end* han podido ser mapeados.

Con el fichero con formato *BAM* obtenido del alineamiento, se utilizaron herramientas de la suite *SAMtools* [38] para ordenar e indexar el archivo. A continuación, para confirmar que el análisis es correcto se visualizó con *IGV* (*Integrative Genomics Viewer*). Esta es una herramienta de visualización que permite la exploración de grandes ficheros de datos genómicos de distintos tipos y de diferentes especies en tiempo real [106].

3.1.2.2.3. Cuantificación

Para medir los niveles de expresión en las distintas muestras del estudio se utilizó el paquete *Genomic Features* [107] del lenguaje de programación *R* que permite al usuario descargar fácilmente las localizaciones genómicas de los transcritos, los exones y *CDS* (*Coding Sequence*) de un organismo dado, ya sea de la base de datos de *UCSC* o *de Ensembl*. Utilizando este paquete calculamos el número de lecturas alineadas en cada gen (a nivel de exones).

3.1.2.2.4. Normalización

Los análisis de *RNA-seq* proporcionan aproximaciones cuantitativas de la abundancia de los transcritos, en forma de conteo. Sin embargo, el conteo tiene que ser normalizado para eliminar los sesgos técnicos procedentes de la secuenciación, en particular, la longitud de las especies de ARN y la profundidad en la secuenciación de la muestra. Para ello, se procedió a normalizar los datos a valores *RPKM* (*Reads per KB per million reads*). *RPKM* es una medida de cuantificación de la expresión génica de datos de *RNA-seq* que se basa en normalizar el total de la longitud de la lectura y el número de lecturas secuenciadas.

$$RPKM = \frac{Número\ de\ lecturas\ alineadas \times 1\ millón\ de\ lecturas}{(longitud\ del\ transcrito \times número\ de\ lecturas\ total)}$$

Con la media de los datos normalizados realizamos un análisis de los componentes principales (*PCA, Principal Component Analysis*). Esta técnica es utilizada para reducir la dimensionalidad de conjuntos de datos; además de poder observar y comprobar las diferencias entre las muestras. Con este análisis observamos en la Figura 23 que las muestras FL6.5 y FL18.5 son las únicas que se agrupan por su similitud.

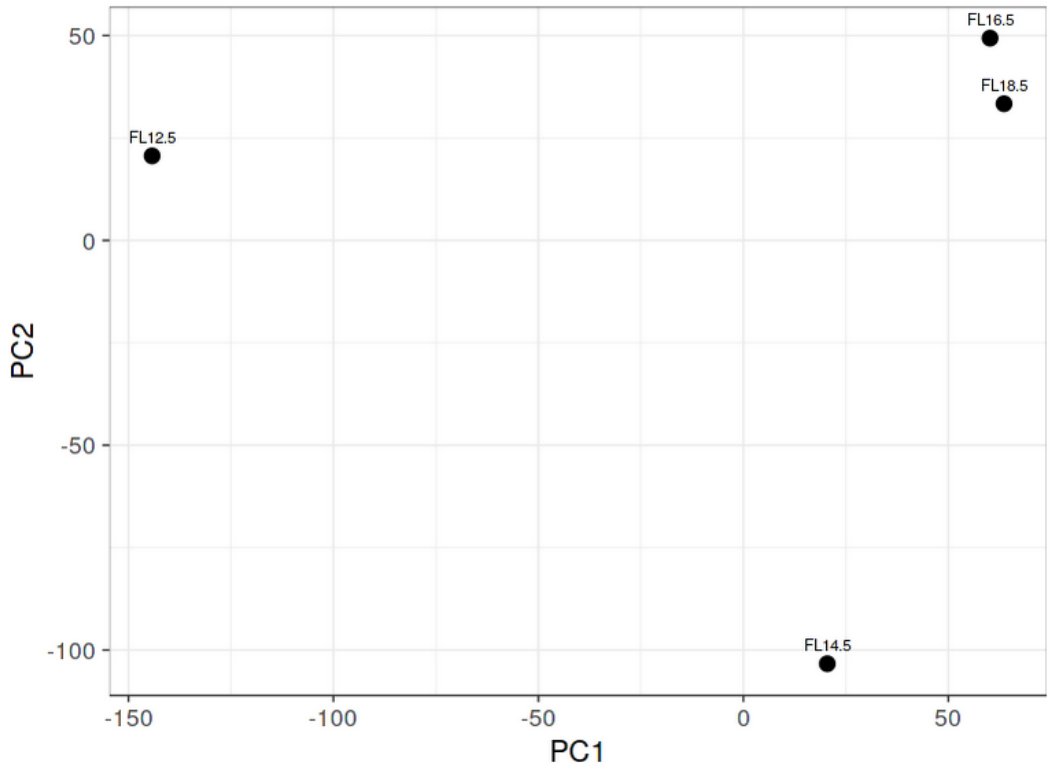


Figura 23. PCA de los datos de expresión normalizados por RPKM de las muestras *HSC*.

3.1.2.2.5. Umbral de activación

Al estudiar la expresión de los genes en distintos momentos de la embriogénesis, y no tratarse de analizar genes diferencialmente expresados, requerimos de un umbral de estimación de expresión real para cada tiempo. De aquí en adelante llamaremos a éste término umbral de activación. El método utilizado para seleccionar el umbral fue el descrito en [108].

Este método se basa en la estimación de la Tasa de Falsos Positivos y Tasa de Falsos Negativos (*FNR*, *False Negative Rate*) en diferentes niveles de expresión. El *FDR* se refiere a la cantidad de lecturas correspondientes a los genes (exones) que realmente corresponden a ruido en la secuenciación, esto se obtiene estimando el ruido de las regiones intergénicas. El *FNR* se refiere a las lecturas que no se han considerado en el conteo debido a errores en la secuenciación, fallos en el alineamiento, u otros factores.

Para seguir éste método, a partir del fichero obtenido del análisis de *RNA-seq* que contiene la media de la cuantificación de los genes en RPKM, se generó una gráfica de densidad agrupando los genes en distintos niveles de expresión con el fin de mostrar el porcentaje de genes que tienen un determinado RPKM. La gráfica también se creó para las regiones intergénicas. Según el estudio de [109], las regiones intergénicas no deberían pertenecer a intrones o a zonas de 10 kb flanqueando los genes, por lo tanto, estas no han sido incluidas en el conteo de las regiones intergénicas.

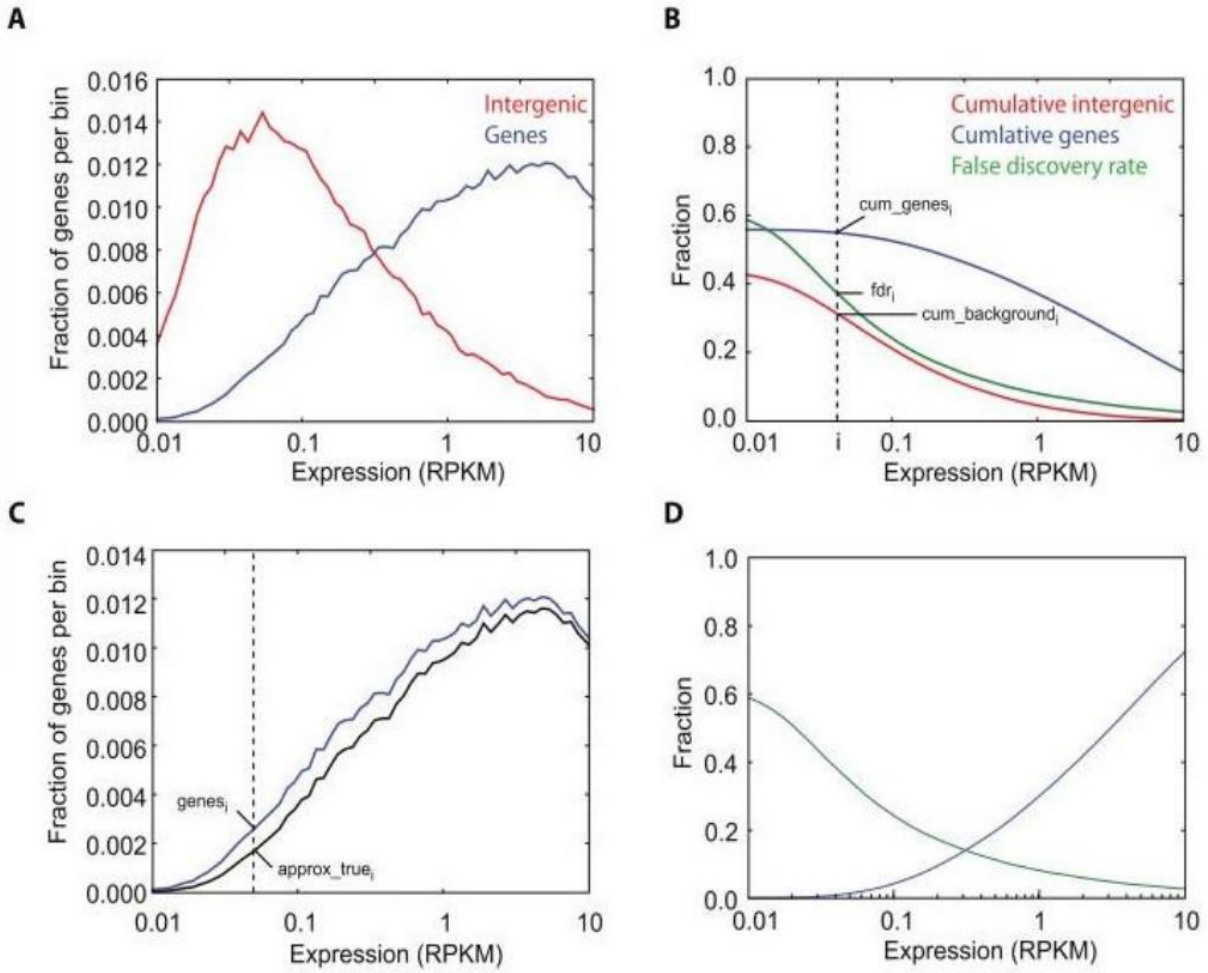


Figura 24. Estimación del FDR y el FNR en los diferentes niveles de expresión. Figura tomada de [108]

Como en el trabajo [108], el autor representa en la Figura 24, gráficas correspondientes a la cantidad acumulada de genes y regiones intergénicas expresadas por encima de ciertos niveles de expresión (Figura 24B). Para cada posición fue calculado el *FDR* de la siguiente manera:

$$FDRi = \frac{\frac{cum_{interg_i}}{cum_{genes_i}} \times \left(1 - cum_{genes_i}\right)}{1 - cum_{interg_i}}$$

Que es utilizado para estimar el verdadero número de genes expresados en cada región (Figura 24C):

$$genes\_est_i = cum\_genes_i \times FDR_i$$

A partir de esta estimación de genes se calculó el FNR mediante:

$FNRi = 1 - \frac{\sum_{j=i}^{n} genes\_estj}{\sum_{j=1}^{n} genes\_estj}$ representándose en la Figura 24D.

Finalmente, en la Figura 24D se representa la *FDR* y *FNR* para diferentes valores de expresión. El valor en el que ambas se intersectan se seleccionó como umbral de activación, siendo este el punto de equilibrio entre ambos valores. Este procedimiento fue repetido para cada uno de los tiempos obteniendo los datos representados en la Tabla 6 que son los umbrales de activación para cada una de las muestras y el número de genes asignados a cada grupo (genes no detectados, genes no expresados y genes expresados).

Tabla 6. Valor del umbral de activación en cada estadio y número de genes no detectados, no expresados y expresados al establecer el umbral.

Muestra	Umbral de activación	Genes no detectados	Genes no expresados	Genes expresados
Hígado Fetal HSC E12.5	0.465668	7850	2161	13274
Hígado Fetal HSC E14.5	0.5598743	7427	2733	13125
Hígado Fetal HSC E16.5	0.5992199	7732	2309	13244
Hígado Fetal HSC E18.5	0.576851	7729	2406	13150
HSC nicho E12.5	0.5708743	8530	1822	12933
HSC nicho E14.5	0.4870799	9038	1598	12649
HSC nicho E18.5	0.4621148	9580	1762	13943

3.1.2.3. Bases de datos de anotaciones utilizadas

Se llevó a cabo una búsqueda exhaustiva de receptores, factores de transcripción y proteínas secretadas en diferentes bases de datos, con el objetivo de asociar estas anotaciones a los genes de los datos de *RNA-seq* que superan el umbral de activación. Además, se obtuvo información de interacciones y rutas metabólicas para la posterior integración e interpretación de los datos.

Esta información fue adquirida de diferentes fuentes de información, como *Uniprot* (*http://www.uniprot.org/*) y *Gene Ontology* (*http://www.geneontology.org/*). Ambos son proyectos bioinformáticos relevantes que contienen información funcional de diversos organismos, de bases de datos específicas de localización o función.

Uniprot es una base de datos de acceso libre de secuencias de proteínas e información funcional. Muchas de sus entradas derivan de proyectos de secuenciación del genoma y además contiene una gran cantidad de información sobre la función biológica de las proteínas derivada de literatura científica [64]. Para la anotación se utilizó la base de datos *Swiss-Prot* que tiene como objetivo proporcionar información fiable con un alto

nivel de anotación y un nivel mínimo de redundancia. De esta base de datos extrajimos aquellas proteínas anotadas como receptores, factores de transcripción y proteínas secretadas en *Mus musculus*.

Por otro lado, *Gene Ontology* es un proyecto que proporciona un vocabulario controlado de términos para describir procesos biológicos, funciones moleculares, y componentes celulares o localizaciones de productos génicos [67]. Estos términos *GO* se usan como atributos de los productos génicos por parte de las distintas bases de datos, facilitando así búsquedas estandarizadas. De estos datos se obtuvieron aquellos genes que se encuentran anotados como *Receptor complex (GO:0043235)* y/o *Receptor activity (GO:0004872)* para los receptores, y como *Extracellular space (GO:0005615)* para las proteínas secretadas. Aparte de estas dos grandes fuentes de información también se utilizaron bases de datos específicas de función o localización celulares detalladas a continuación.

3.1.2.3.1. Bases de datos de receptores

Las bases de datos de receptores consultadas fueron: *International Union of Basic and Clinical Pharmacology* (*IUPHAR*,http://www.guidetopharmacology.org/) [110] y *Human Plasma Membrane Receptome* (*HPMR*, http://www.receptome.org/HPMR/). *IUPHAR* es una asociación sin ánimo de lucro con el objetivo de proporcionar acceso a un gran repositorio de datos de receptores, canales iónicos y medicamentos, mientras que *HPMR* contiene familias de receptores de humanos que participan en rutas de señalización. Esta última base de datos se creó con el objetivo de estudiar la evolución de los receptores [111].

3.1.2.3.2. Bases de datos de proteínas secretadas

Secreted Protein Database (*SPD*,http://spd.cbi.pku.edu.cn) es una base de datos de proteínas secretadas en ratón, rata y humano que contiene entradas de las bases de datos S*wiss-Prot*, *TrEMBL*, *RefSeq*, *Ensembl* y *CBI-gene*, además de proteínas secretadas predichas [112].

3.1.2.3.3. Bases de datos de interacciones

Para detectar qué proteínas secretadas del nicho podrían interactuar con receptores expresados en células *HSC*, se recopilaron las interacciones conocidas en humano y ratón de las bases de datos *INTACT* del *EMBL-EBI (*https://www.ebi.ac.uk/intact/) [113] y *Molecular interaction database* [114] (*MINT*, http://mint.bio.uniroma2.it/mint/) . Estas son dos bases de datos que contienen interacciones proteína-proteína manualmente revisadas de datos procedentes de la literatura científica y datos experimentalmente validados.

3.1.2.3.4. Bases de datos de factores de transcripción

Con el fin de saber qué rutas están involucradas en la migración y expansión de las *HSC*, se realizó una búsqueda exhaustiva de los *targets* de los factores de transcripción expresados en los distintos tiempos. Las

bases de datos utilizadas para la obtención de factores de transcripción y sus targets fueron *Pazar, TRANSFAC, Integrated Transcription Factor Plataform (ITFP)* e *Ingenuity*.

Transfac es una base de datos que requiere licencia y contiene información revisada por expertos de factores de transcripción de organismos eucariotas. La base de datos incluye sus sitios de unión y los perfiles de unión a ADN, además de sus *targets* entre otra información [115]. También se utilizaron repositorios públicos como *PAZAR* [116] y *ITFP* [117], que además de información experimentalmente validada utiliza predicciones en organismos modelo. Finalmente utilizamos la herramienta "*Upstream analysis*" de *Ingenuity,* que permite obtener los factores de transcripción de los datos de *RNA-seq* estudiados.

3.1.2.3.5. Ingenuity para la obtención de rutas metabólicas

Ingenuity Pathways Analysis (IPA) (www.ingenuity.com) es una herramienta web que contiene interacciones biológicas y químicas, además de anotaciones funcionales que han sido manualmente revisadas, entre otra información. El software *IPA* contiene distintas herramientas para realizar diversos análisis. En este proyecto se utilizó el software para interpretar los datos de expresión de *Mus musculus* en el contexto de procesos biológicos y rutas metabólicas. Para ello, se introdujeron los datos de expresión obtenidos del análisis de *RNA-seq* en cada tiempo del hígado fetal y se seleccionó el tipo de análisis "*Core analysis*". Éste análisis realiza una evaluación rápida de las vías de señalización y metabólicas de las redes moleculares y de los procesos biológicos que están significativamente más alterados por los datos de interés introducidos. Estos datos pueden ser exportados, pero también se puede obtener información detallada acerca de la ruta y de los genes involucrados; o también obtener un diagrama interactivo de la ruta.

3.1.2.4. Integración de datos

El diagrama del proceso con los *scripts* utilizados para obtener las rutas de mayor interés para el estudio, se representa en la Figura 25. El primer paso fue integrar la información obtenida de las diferentes bases de datos para identificar aquellos receptores, proteínas secretadas y factores de transcripción expresados en los datos de *RNA-seq.* A continuación, se filtraron las rutas de *IPA* utilizando la información biológica obtenida para obtener únicamente aquellas interesantes para el estudio, quedándonos con las rutas con un p-valor mayor a 0.01 y con un factor de transcripción y receptor (con interacción con proteína expresada conocida) expresados.

3.1.2.4.1. Integración de anotaciones

Los identificadores de los datos obtenidos de las distintas bases de datos fueron unificados a uno común para poder comparar, integrar resultados y eliminar las redundancias propias de la utilización de distintas fuentes de información. Se utilizó como identificador *ENSEMBL GENE ID* de *Mus musculus*. Esta conversión fue llevada

a cabo utilizando la herramienta *DICT (David Gene ID Conversion Tool)* que convierte los distintos identificadores al identificador seleccionado, y además informa sobre cuáles son los identificadores de ese gen en otras especies [84] . Este último punto es importante, porque las bases de datos específicas utilizadas no siempre contienen datos de ratón y en aquellos casos en que la base de datos contiene datos experimentalmente validados, se recogió la información de humano y se anotó como ratón utilizando esta herramienta.

Juntados todos los datos, convertidos a un identificador común y eliminadas las redundancias se obtuvieron 2.488 receptores, 1.758 proteínas secretadas, y 2.626 factores de transcripción. Este procedimiento fue llevado a cabo mediante el desarrollo y utilización de los *scripts* en *Perl* (Bases_datos.pl y Ensemble_gene_Id.pl).

3.1.2.4.2. Selección de los genes expresados

A partir de los datos obtenidos del análisis de *RNA-seq* y mediante una serie de *scripts* se seleccionaron aquellos genes de los datos de expresión en el hígado fetal que se encontraban expresados según el umbral de activación propuesto para cada uno de los tiempos. Estos se agruparon según su correspondencia a receptores o factores de transcripción, obteniendo así los receptores y factores de transcripción expresados en un determinado estadio de tiempo.

El mismo análisis se realizó para los genes de los datos de expresión de *RNA-seq* pertenecientes al nicho, en este caso se seleccionaron aquellos genes que correspondían a proteínas secretadas.

Las proteínas expresadas resultantes en cada estadio se agruparon según: nicho y nicho-1. Es decir, el conjunto de secretadas para el estadio FL12.5 se constituyó únicamente por secretadas expresadas en el nicho FL12.5, las del estadio FL14.5 por secretadas expresadas en los nichos FL12.5 y FL14.5, las del estadio FL16.5 por las de los nichos FL14.5 y FL16.5, y las del estadio FL18.5 por las del nicho FL16.5.

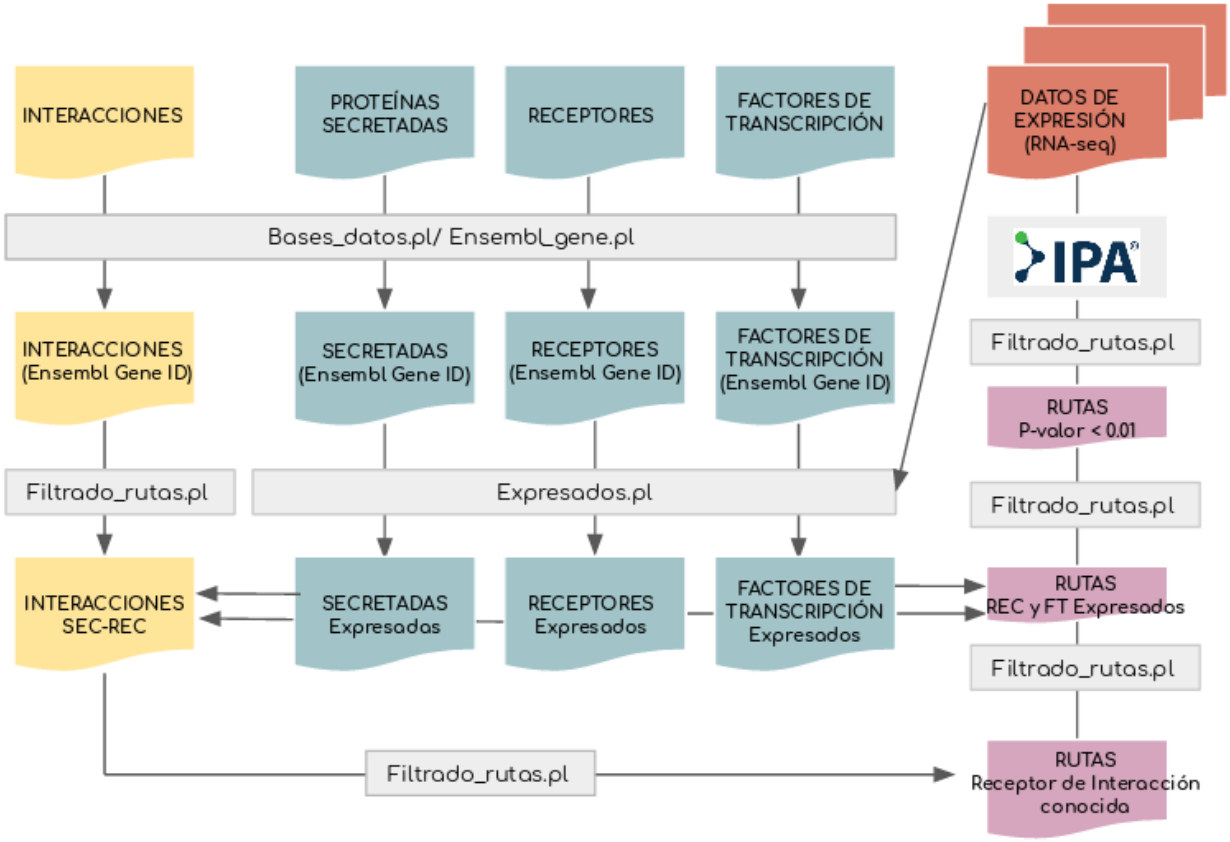


Figura 25. Diagrama del proceso realizado para obtener las rutas interesantes para el estudio. Las proteínas obtenidas en las distintas bases de datos fueron convertidas a un identificador común (*Ensemble Gene ID*) mediante los scripts Bases_datos.pl y Ensembl_gene.pl. Seguidamente con el script Activos.pl se seleccionaron aquellos factores de transcripción, receptores y proteínas secretadas que superaran el umbral de expresión y que aparecían en los datos de expresión. Los datos de expresión del análisis de *RNA-seq* se analizaron con *IPA* y mediante el script Filtrado_rutas.pl se filtraron las rutas para obtener aquellas con un p-valor mayor a 0.05, aquella que contenían un receptor y un factor de transcripción activado y finalmente aquellas que contenían un receptor con interacción con proteína secretada conocida.

3.1.2.4.3. Análisis de las rutas

Con el propósito de encontrar información sobre cuáles son las rutas que podrían estar involucradas en cada estadio del análisis, se utilizó *Ingenuity Pathways Analysis (IPA)*. Este permite a través de un conjunto de datos (en este caso, los datos de expresión de *HSC* en FL) mostrar cuales son las rutas más significativas. Al introducir los datos de expresión de los diferentes tiempos resultantes del análisis de *RNA-seq* (FL12.5, FL14.5, FL16.5 y FL18.5), se obtuvieron 436, 267, 449 y 441 rutas respectivamente (Tabla 7).

Cada archivo contiene el nombre de la ruta, el -log (p-valor) asociado a la ruta que es calculado mediante el test de *Fisher*, el ratio (moléculas activadas en cada ruta versus el número total de moléculas de la ruta) y los genes que se encuentran en la ruta.

Las rutas obtenidas fueron estudiadas y filtradas con el fin de obtener únicamente aquellas relevantes para el estudio y su posterior validación. Para empezar, se seleccionaron únicamente aquellas rutas

significativamente afectadas por los datos de expresión (–log mayor que 1.3 (p-valor <= 0.05)). A continuación, nos centramos en aquellas rutas que contenían un receptor y un factor de transcripción expresado en cada uno de los tiempos, lo que redujo bastante las rutas a estudiar. Y, por último, dado que el objetivo del análisis es obtener las rutas desencadenadas por las interacciones de las *HSC* con sus microambientes (nichos), el siguiente filtrado se focalizó en obtener aquellas rutas que entre sus genes contenían un receptor que pudiera interaccionar con una proteína secretada del nicho. Para este filtrado utilizamos las 7.476 interacciones anotadas en *INTACT* y *MINT*, y estudiamos las rutas que contenían un receptor expresado del que se conoce su interacción con una proteína secretada que también se encuentre expresada en el tiempo estudiado. Todos estos filtros se aplicaron con el *script* en *Perl* Filtrado_rutas.pl.

Tabla 7. Número de rutas en cada proceso de filtrado.

Número de rutas en cada tiempo	FL12.5	FL14.5	FL16.5	FL18.5
Resultados IPA	436	267	449	441
P-valor <=0.05	262	70	264	259
Receptor y Factor de transcripción	187	40	181	176
Interacción (Receptor)	64	14	65	54
Puntuación 3	37	11	26	28

Al finalizar los distintos análisis se obtuvieron 64, 14, 65 y 54 rutas en los estadios FL12.5, FL14.5, FL16.5 y FL18.5 respectivamente como muestra la Tabla 3. El estadio FL14.5 fue el que obtuvo un número menor de rutas, pero también partió de menos rutas que en los otros casos: 267 respecto 436, 449 y 441.

Para concluir se estudiaron cada una de las rutas resultantes con el software *IPA*, que permite visualizar la ruta distinguiendo qué moléculas de los datos introducidos se expresan en una cierta condición y cuáles no. Dividimos las listas en 4 grupos asignando los números 0, 1, 2 y 3, dónde 3 representa la ruta más interesante en el contexto que estamos estudiando y el 0 la menos relacionada con estos procesos, pudiéndose incluso eliminar del análisis. La selección manual se basó en el tipo de receptor, la función de la ruta y la expresión observada en las distintas moléculas. Para que la ruta obtuviese un 3, esta debería estar constituida por un receptor extracelular, que estuviese involucrada en algún proceso de migración y/o desarrollo y, además, que las moléculas que forman parte de la ruta se encontraran expresadas. En la Figura 26 se observan las puntuaciones que han recibido las rutas, mostrando que en todos los estadios la mayoría de rutas han recibido un 2 o un 3 y ninguna un 0.

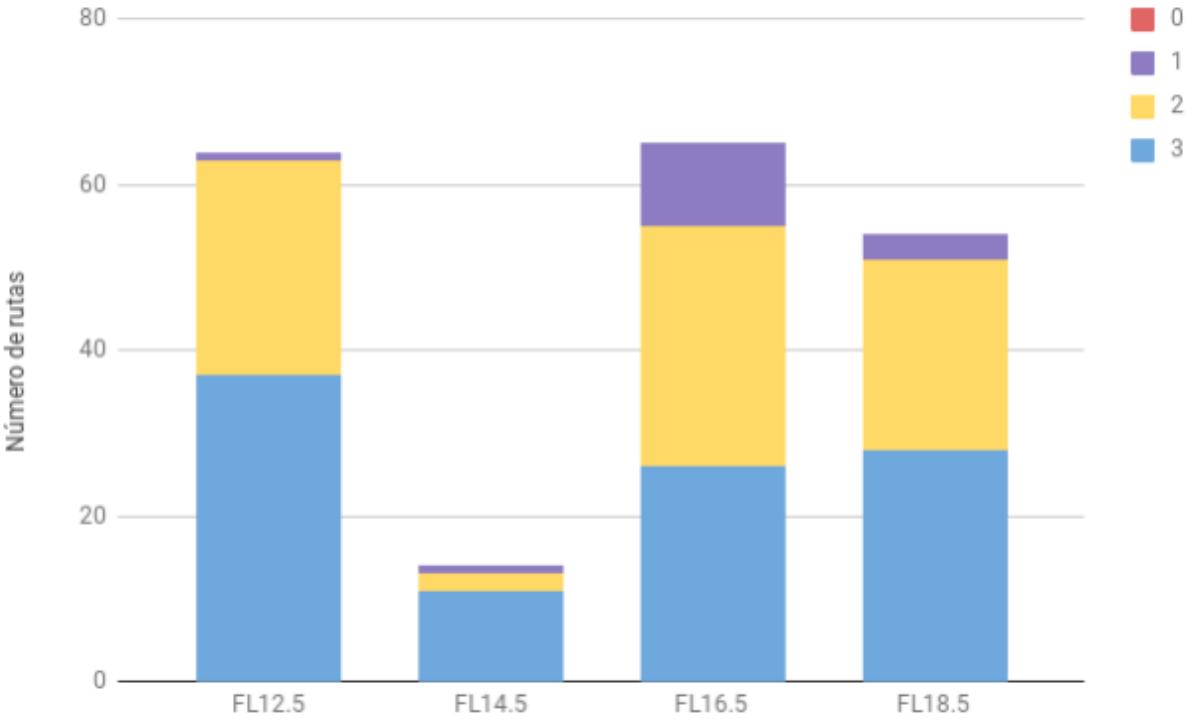


Figura 26.Número de rutas obtenidas al final del filtrado en cada estudio y sus puntuaciones asignadas.

3.1.3. Resultados

Los resultados obtenidos al emplear esta metodología son una lista de rutas potencialmente interesantes para explicar los procesos que se dan durante la expansión y migración de las *HSC* en el hígado fetal. También una lista de proteínas secretadas y receptores que podrían mostrar las diferencias entre los tiempos y ayudar a explicar los procesos.

3.1.3.1. Rutas obtenidas

De todas las rutas resultantes, 58 obtuvieron la puntuación 3. El tiempo que más rutas consiguió fue FL12.5 con 37 rutas seguido por FL18.5, FL16.5 y FL14.5 con 28,26 y 11 respectivamente. A continuación, en la Tabla 8, se muestran las 58 rutas que han obtenido la máxima puntuación.

Tabla 8. Rutas finales obtenidas en cada estadio.

RUTA	FL12.5	FL14.5	FL16.5	FL18.5
SIGNALING BY RHO FAMILY GTPASES	x			x
ILK SIGNALING	x		x	x
FACTORS PROMOTING CARDIOGENESIS IN VERTEBRATES	x			
FAK SIGNALING	x			x
TEC KINASE SIGNALING	x		x	x
ERK/MAPK SIGNALING	x		x	x
NF-κB ACTIVATION BY VIRUSES	x			x
MACROPINOCYTOSIS SIGNALING	x			
COLORECTAL CANCER METASTASIS SIGNALING	x			x
INTEGRIN SIGNALING	x		x	x
GLIOMA SIGNALING	x		x	x
PROTEIN KINASE A SIGNALING	x		x	
PI3K/AKT SIGNALING	x		x	x
RHOGDI SIGNALING	x		x	x
HUNTINGTON'S DISEASE SIGNALING	x			x
ROLE OF PKR IN INTERFERON INDUCTION AND ANTIVIRAL RESPONSE	x			
SEMAPHORIN SIGNALING IN NEURONS	x			
PAK SIGNALING	x		x	
PHOSPHOLIPASE C SIGNALING	x		x	x
IL-6 SIGNALING	x		x	x
GERM CELL-SERTOLI CELL JUNCTION SIGNALING	x		x	x
RAC SIGNALING	x		x	x
WNT/β-CATENIN SIGNALING	x		x	x
SERTOLI CELL-SERTOLI CELL JUNCTION SIGNALING	x			x
AGRIN INTERACTIONS AT NEUROMUSCULAR JUNCTION	x			
REGULATION OF EIF4 AND P70S6K SIGNALING	x			x
AXONAL GUIDANCE SIGNALING	x			
CDC42 SIGNALING	x		x	x
PTEN SIGNALING	x		x	x
VEGF FAMILY LIGAND-RECEPTOR INTERACTIONS	x			
ACTIN CYTOSKELETON SIGNALING	x		x	x
PAXILLIN SIGNALING	x		x	x

VEGF SIGNALING	x			
ACTIN NUCLEATION BY ARP-WASP COMPLEX	x			x
EPHRIN RECEPTOR SIGNALING	x			
LEUKOCYTE EXTRAVASATION	x			
NITRIC OXIDE SIGNALING IN THE CARDIOVASCULAR SYSTEM	x			
ROLE OF MACROPHAGES, FIBROBLASTS AND ENDOTHELIAL CELLS IN RHEUMATOID ARTHRITIS		x		
ROLE OF OSTEOBLASTS, OSTEOCLASTS AND CHONDROCYTES IN RHEUMATOID ARTHRITIS		x		
LXR/RXR ACTIVATION		x		
ROLE OF PATTERN RECOGNITION RECEPTORS IN RECOGNITION OF BACTERIA AND VIRUSES		x		
ALTERED T CELL AND B CELL SIGNALING IN RHEUMATOID ARTHRITIS		x		
HEPATIC FIBROSIS / HEPATIC STELLATE CELL ACTIVATION		x		
REGULATION OF THE EPITHELIAL-MESENCHYMAL TRANSITION PATHWAY		x		
LPS/IL-1 MEDIATED INHIBITION OF RXR FUNCTION		x		
IL-10 SIGNALING		x		
MIF-MEDIATED GLUCOCORTICOID REGULATION		x		
HEPATIC CHOLESTASIS		x		
IL-9 SIGNALING			x	
IL-15 SIGNALING			x	
IL-2 SIGNALING			x	
PPAR SIGNALING			x	
ROLE OF JAK1 AND JAK3 IN γC CYTOKINE SIGNALING			x	
IL-4 SIGNALING			x	
P38 MAPK SIGNALING			x	x
NEUREGULIN SIGNALING			x	x
REGULATION OF CELLULAR MECHANICS BY CALPAIN PROTEASE				x
TNFR1 SIGNALING				x

Como se observa en la tabla, los tiempos que más comparten resultados son el FL16.5 y FL18.5, aunque el tiempo FL12.5 también comparte alguna ruta. En cambio, las rutas obtenidas para el tiempo FL14.5 son exclusivas para este estadio.

Además de las rutas, resultó interesante saber qué proteínas secretadas y receptores están involucrados en la obtención de las rutas finales para poder estudiar si es importante validar alguno de estos elementos. La Tabla 9 muestra los receptores y proteínas secretadas involucradas en las rutas finales. En estos datos se sigue observando una clara diferencia entre el tiempo FL14.5 y el resto de tiempos.

Tabla 9.Receptores y proteínas secretadas involucrados en la obtención de las rutas.

RECEPTORES	FL12.5	FL14.5	FL16.5	FL18.5
ITGB1	x		x	x
MET		x		
IL1RAP		x		
TRAF2		x		
IL1R1		x		
TLR4		x		
MYH9	x		x	x
KDR	x	x		
IGF2R	x		x	x
ITGB2	x		x	x
LRP6	x		x	x
LRP5	x		x	x
TNFRSF1A	x		x	x
EBI3	x		x	
DLG4	x			x

SECRETADAS	FL12.5	FL14.5	FL16.5	FL18.5
CD40		x		
IL1RL1		x		
IL15RA		x		
IL1B		x		
LY96		x		
CSF1	x		x	x
VEGFA	x	x		
CTSD	x		x	x

ITGAM	x		x	x
MESDC2	x		x	x
FADD	x		x	x
FBLN2	x		x	x
LRP8	x			x
NPTN	x			x
TXN1	x			x
GPI1	x			x
CLU	x			x
PDIA3	x			x
SOD1	x			x
IL12A	x		x	
LYZ2	x			x
KDR		x		
IL1R1		x		
VEGFB	x		x	x

Estudiando los datos, se observan algunos genes ya descritos en procesos de células hematopoyéticas como Csf1 o KDR, que tienen un rol esencial en la regulación de la proliferación y diferenciación de las HSC [87,118] . La interacción que da lugar al mayor número de rutas es la formada por el receptor ITGB1 y la secretada VEGFB.

Para estudiar las diferencias entre nichos, y así ver las diferentes señales que transmiten según el tiempo, se representaron las proteínas secretadas expresadas para cada tiempo en un diagrama de Venn (Figura 27). Como se ha descrito anteriormente las proteínas secretadas están agrupadas según nicho y nicho-1.

En la representación se muestra que la mayoría de proteínas son comunes. 601 proteínas se expresan en todos los estadios, 101 en tres estadios (70 en FL12.5, FL14.5 y FL16.5; y 31 en FL14.5, FL16.5 y FL18.5) y 168 en dos estadios de las cuales 47 corresponden a FL14.5 y FL16.5; 61 a FL12.5 y FL14.5 y 60 a FL16.5 y FL18.5. Estas 168 proteínas serían las específicas de tiempo.

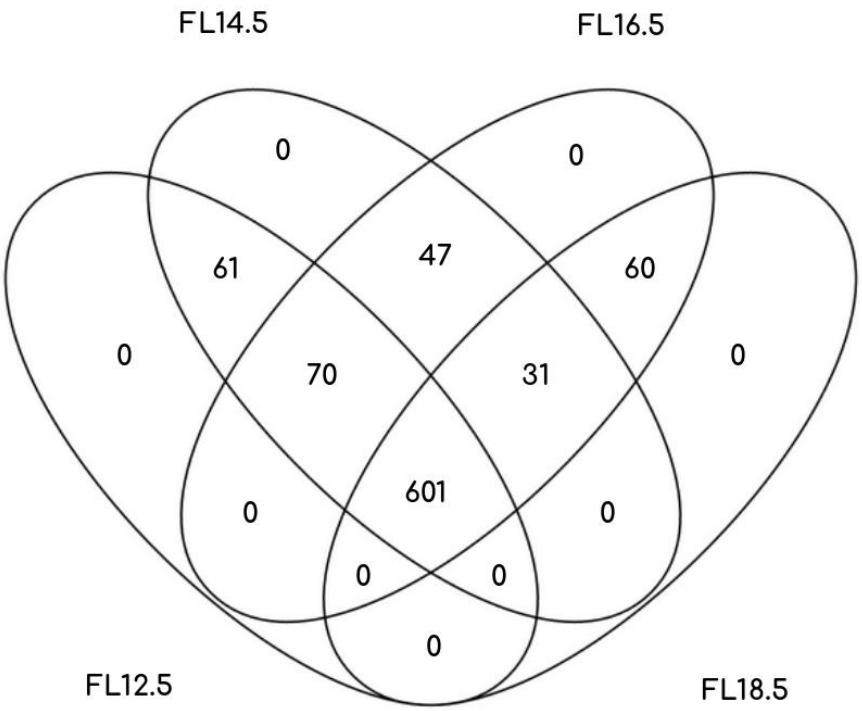


Figura 27. Proteínas secretadas expresadas agrupadas según nicho y nicho -1.

3.1.3.2. Interpretación de los resultados obtenidos

El hecho de que la mayoría de las rutas se hayan clasificado en los grupos 2 y 3, indica que el método obtiene rutas potencialmente significativas para el estudio. Aunque la información acerca de las señales de los nichos y como afectan es escasa hoy en día [100], se estudió en mayor profundidad las rutas obtenidas y se observó que los datos resultantes contienen moléculas y rutas importantes descritas en otros estudios.

En el estudio de [118], el autor comenta que la secreción de efectores solubles como *angiopoietin -like3 (Angptl4)* y *insulin-like growth factor-2 (IGF2)* promueven la expansión, la diferenciación y la maduración de *HSC*. Esto apoyaría los resultados obtenidos en nuestro estudio, ya que ambas proteínas se encuentran activas en los nichos para los tres estadios. Otro caso es la quimiocina *CXCL12* que se encuentra activa en todos los estadios y se cree que es necesaria para el proceso de *homing* según [100] y la regulación de la migración en las *HSC* [94]. También, estos mismos autores comentan que la molécula *VEGF*, que en nuestro estudio está expresada en todos los estadios y que forma parte de la ruta *VEGF SIGNALING* obtenida en el estadio FL12.5, podría inducir a la formación de tejido hematopoyético, además de a la proliferación celular.

Por otro lado, nos fijamos en aquellas proteínas secretadas que difieren entre estadios para ver si existen diferencias entre los tiempos. Nos centramos en aquellas proteínas específicas de estadio. De las 168 proteínas secretadas obtenidas en únicamente dos tiempos, solo *LY96, IL1B y CD40* influyeron en la obtención

de rutas. *LY96* y *IL1B* se encontraban expresadas en el nicho 12,5 mientras que *CD40* en el nicho 14,5. Además de las 11 rutas resultantes del estadio FL14.5, 10 de ellas han sido obtenidas por la interacción de una de estas 3 proteínas secretadas con un receptor y ninguna de estas rutas se encuentra en las otras fases del desarrollo. No se ha encontrado información acerca de su importancia, en el caso de que exista, con la migración y expansión de las *HSC*, pero *IL1B* y *CD40* participan en la proliferación celular, mientras que *LY96* en respuesta inmune.

Cuando observamos la tabla con las rutas obtenidas vemos que el tiempo FL12.5 es el que ha obtenido más rutas, 37, seguido por FL18.5, FL16.5 y FL14.5 con 28, 26 y 11 respectivamente. Las muestras FL16.5 y FL18.5 son las que más tienen en común y eso podría ser porque en esas últimas fases del desarrollo ya las células sufren menos cambios. Que estos dos tiempos contengan expresiones de genes similares, concuerda con el análisis de PCA resultante de las muestras antes de ser analizadas donde mostraban que las muestras más similares eran las de los tiempos FL16.5 y FL18.5.

Según varios estudios [94][96][98], los procesos involucrados en la generación y migración de las HSC son desconocidos. Hasta el momento se sabe que NOTCH1 se requiere para la producción de HSC en AGM (aorta-gónada-mesonefros), participando en el proceso otras rutas como CoupTF-II, CDX-HOX y la ruta de activación de proaglandinas. Por otro lado, *WNT/ WNT/β-CATENIN* es conocida por su importancia en la autorrenovación, proliferación y diferenciación de las células madre adultas [90,119], y la familia VEGF que se requiere para la proliferación de HSC adultas y para atraer células mieloides a sitios específicos.

De los procesos comentados en los resultados obtenidos encontramos la ruta *WNT/ WNT/β-CATENIN* en todos los tiempos menos en el 14.5 y también la familia VEGF. Lo que sugiere que podrían no únicamente ser esenciales en células adultas. Sin embargo, ninguna de las rutas requeridas para la producción de HSC en AGM se ha encontrado con nuestra metodología, que significaría que cuando llegan al hígado fetal no solo cambia la localización, sino que los procesos conocidos que participan en AGM, ya no estarían activos.

Por otro lado, destacamos que en muchas de las rutas obtenidas aparecen interleucinas (*IL*). Las Interleucinas son sintetizadas principalmente por los leucocitos, aunque en algún caso también pueden intervenir células endoteliales o del estroma del timo o de la médula ósea, por lo que sería interesante ver por qué en el nicho FL16.5 se activan más rutas que involucran interleucinas que en FL18.5. También, los resultados del tiempo FL14.5 fueron distintos al resto de estadios: no solo consiguieron menos rutas, sino que, en los análisis de rutas, de anotaciones y el gráfico PCA reflejan una clara diferencia de expresión en este tiempo con respecto al resto.

Es importante destacar que este estudio está aún en desarrollo y la evaluación exhaustiva de todas las rutas y sus implicaciones en las señalizaciones de *homing* y migración está aún en progreso, actualmente los datos resultantes los tiene el grupo con el que colaboramos para estudiarlos en detalle y validar experimentalmente en pez cebra algunos datos como ya realizaron en el anterior trabajo [120] en el que se analizaron muestras en el estadio FL14.4 y cuando las HSC alcanzan la médula ósea mediante un análisis de expresión diferencial.

3.1.3.3. Estrategia bioinformática

La estrategia bioinformática llevada a cabo para obtener una lista de las posibles rutas y moléculas involucradas en la expansión y migración de las *HSC* en diferentes estadios de la embriogénesis, ha consistido en la creación de una serie de *scripts* en lenguaje *Perl* y la utilización de algunas aplicaciones bioinformáticas existentes. En esta metodología se han comparado diferentes estadios del desarrollo de las HSC utilizando un umbral de activación y no realizando un análisis de expresión diferencial como en el anterior trabajo realizado [120]. Con esta metodología y utilizando los datos procedentes de *RNA-seq* se han obtenido rutas metabólicas y moléculas a estudiar que podrían estar involucradas en la expansión y migración de las *HSC*.

En esta estrategia se han tenido en cuenta algunos parámetros como el p-valor, el umbral de activación y las moléculas obtenidas de las bases de datos. El cambio de cualquiera de estos parámetros modifica significativamente los resultados obtenidos, especialmente la introducción o eliminación de elementos en las bases de datos.

Como las rutas obtenidas serán analizadas por un laboratorio experimental, nos interesa no introducir más rutas de las que se creen necesarias. Por ese motivo se seleccionó un p-valor de 0.05, para que únicamente se trabajara con aquellas rutas que *IPA* consideraba significativamente afectada por la expresión de los genes. En cambio, la búsqueda de anotaciones para identificar los datos de *RNA-seq* fue crítica para el análisis. Los resultados varían mucho si las proteínas de los datos de expresión no están anotadas o si la anotación no es correcta, especialmente en el caso de las proteínas secretadas y los receptores.

Al analizar los datos obtenidos de otras bases de datos como *Swiss prot* se observó, que muchas moléculas anotadas como *secreted protein* correspondían a receptores y no a proteínas secretadas como se esperaba. Por lo tanto, las proteínas se obtuvieron especialmente de bases de datos especializadas y/o se buscaron por términos ajustados de *Gene Ontology* y posteriormente, fueron revisadas manualmente.

Debido a que la mayor parte de datos de *RNA-seq* se analizan realizando un análisis de expresión diferencial comparando dos condiciones, la estimación del umbral de activación para procesos fisiológicos normales fue complejo. Este trabajo requiere un método que calcule la expresión en cada tiempo, debido a que el estudio se lleva a cabo en diferentes estadios de desarrollo. Para ello se utilizó el método descrito por [108]. Aunque los umbrales de activación parezcan valores muy bajos y similares entre ellos, al realizar el análisis, se han excluido muchas moléculas no expresadas y se han observado diferencias entre los diferentes tiempos, aunque entendemos que este enfoque no es tan restrictivo como un análisis diferencial.

3.1.4. Conclusiones

El conocimiento actual sobre las *HSC* ha permitido el tratamiento de varias enfermedades, sin embargo, el conocimiento de los mecanismos moleculares de los nichos para regular las *HSC* está aún por descubrir.

Existen evidencias que sugieren que las *HSC* se encuentran distribuidas en nichos y reciben señales moleculares que las proveen de los requerimientos funcionales necesarios. Entender cómo estos nichos mantienen las *HSC*, no sólo hará avanzar nuestra comprensión de la Biología, sino también nos puede proporcionar las ideas necesarias para el diseño de futuras estrategias para su utilización en el ámbito clínico. En base a estas evidencias, se ha creado una estrategia bioinformática que permite, a partir de datos de *RNA-seq* y basándose en un contexto biológico concreto, obtener una lista de rutas que podrían recibir señales del nicho y estar involucradas en procesos de migración y expansión de las *HSC* para que el laboratorio experimental las estudie y valide. Los resultados obtenidos corroboran estas evidencias mostrando diferencias interesantes y significativas entre los diferentes tiempos.

La estrategia llevada a cabo en este trabajo permite obtener aquellas rutas con unas características determinadas. Además, la metodología desarrollada es flexible y permite modificar los parámetros al igual que incluir nuevas anotaciones que se vayan descubriendo. No obstante, sabemos que es necesario el posterior estudio experimental de las rutas, para corroborar los resultados obtenidos *in silico.*

Desde el punto de vista bioinformático, esta aportación de la Tesis ha permitido la aplicación práctica de distintas materias bioinformáticas de actualidad. Entre otras, el análisis de datos de secuenciación masiva para estudios de transcriptómica, análisis funcional, análisis interactómico, bases de datos de anotaciones, interacciones, dianas de factores de transcripción entre otras. Así mismo el uso de distintos paquetes de software, tanto académicos como comerciales, ha sido muy enriquecedor en un proyecto de estas dimensiones. Por último, todas las tareas de programación específica desarrolladas han permitido dar a este proyecto un carácter global desde el punto de vista bioinformático.

3.2. NFFINDER: Una herramienta bioinformática para el reposicionamiento de fármacos mediante experimentos transcriptómicos

La neurofibromatosis es un conjunto de enfermedades raras provocadas por un trastorno genético del sistema nervioso que se caracteriza por el desarrollo de múltiples tumores benignos en los nervios del cuerpo y la piel. Como sucede en un alto porcentaje de enfermedades raras y genéticas, no existe en la actualidad un tratamiento específico para la enfermedad. El diseño de nuevos fármacos implica un proceso que lleva muchos años de investigación, grandes inversiones de dinero y muy pocas nuevas drogas son capaces de pasar a las fases clínicas de la investigación.

En este contexto y conjuntamente con la fundación *Children Tumor Fundation* se desarrolló *NFFINDER*, una herramienta para la creación de hipótesis de nuevos tratamientos a todo tipo de enfermedades mediante el reposicionamiento de fármacos. El reposicionamiento de fármacos (del inglés, *drug repurposing* o *repositioning*) es una estrategia mediante la cual se identifican nuevos compuestos o drogas a partir de su capacidad para tratar enfermedades distintas de aquellas para las que fueron diseñadas originalmente. La estrategia utilizada en esta herramienta se basa en la identificación de genotipos similares u opuestos (signaturas moleculares), que den pistas sobre otras patologías o procesos similares para los cuales sí existen drogas efectivas, utilizando una estrategia similar a la de *MARQ* [121] y *CMap* [122]. Además, para completar el trabajo se desarrollaron aproximaciones novedosas como incluir datos de *RNA-seq* e información regulatoria de miARNs, además de poner de manifiesto a los autores de los trabajos para promover colaboraciones en la investigación.

3.2.1. Neurofibromatosis

La neurofibromatosis es una enfermedad autosómica dominante causada en humanos por la deficiencia de uno de los genes de neurofibromina, *NF1* o *NF2*. La enfermedad se caracteriza por el desarrollo de tumores en el sistema nervioso. Los tumores suelen ser benignos y los síntomas pueden ser leves. Sin embargo, las complicaciones de la neurofibromatosis pueden conllevar a la pérdida de audición, deterioro del aprendizaje, problemas cardíacos y de los vasos sanguíneos (cardiovasculares), pérdida de la visión y dolor intenso.

Existen tres tipos de neurofibromatosis: La neurofibromatosis tipo 1 (*NF1*) es el tipo más común de neurofibromatosis y ocurre en uno de cada 3000 nacimientos a nivel mundial. Es causada por mutaciones en el gen *NF1* que produce la proteína neurofibromina que participa en la regulación del crecimiento de las células y también actúa como gen supresor de tumor. La mutación del gen *NF1* conlleva a un crecimiento celular sin control [123]. Aunque puede surgir de forma esporádica, un 10% de los pacientes con NF1 suele desarrollar *MPNST* (del inglés, *Malignant peripheral nerve sheath tumor*s), que es una manifestación más severa de la enfermedad y se caracteriza por la formación de sarcomas raros localizados principalmente en el tronco y extremidades que pueden llegar a formar metástasis [124,125].

La neurofibromatosis tipo 2 (*NF2*) es menos frecuente, ocurre a una de cada 25.000 personas y suele manifestarse con tumores benignos en los oídos, aunque pueden crecer en otros nervios del cuerpo. Es causada por la mutación en el gen *NF2* que se encuentra en el cromosoma 22 y produce la proteína merlina [126]. Por último, la *Schwannomatosis* es el tipo más raro de neurofibromatosis, afectando a una de cada 40.000 personas y se caracteriza por la formación de tumores en los nervios del cuero cabelludo, espinales y periféricos.

Hoy en día no existe un tratamiento efectivo para la neurofibromatosis. Los tratamientos existentes intentan maximizar el crecimiento y desarrollo saludables y controlar las complicaciones tan pronto como surjan. Cuando la neurofibromatosis causa tumores grandes o tumores que presionan un nervio, la cirugía puede aliviar estos síntomas. Algunas personas pueden beneficiarse con otras terapias, como la radiocirugía estereotáctica o medicamentos para controlar el dolor.

3.2.2. Reposicionamiento de drogas en perfiles transcripcionales

El proceso de desarrollo de un nuevo fármaco suele ser un proceso largo en el que se invierte mucho dinero y muchos recursos. El proceso se suele dividir en cuatro fases y se requiere de una inversión media de US$1,78 millones y un promedio de 13.5 años [127] . Aun así, únicamente un 10% de los medicamentos investigados supera las distintas fases. La industria farmacéutica almacena estos medicamentos rechazados y normalmente utiliza esta información en otro tipo de estrategias como es el reposicionamiento de fármacos.

El reposicionamiento de fármacos es básicamente la utilización de compuestos ya conocidos, en enfermedades para las cuales no habían sido diseñados inicialmente. Esta estrategia permite reducir costes y tiempo invertido para llegar hasta la aprobación de un medicamento.

Existen diferentes casos exitosos de reposicionamiento de fármacos, el más conocido es el del Sildenafilo (Viagra), que inicialmente fue estudiada como terapia contra la angina de pecho, pero los científicos de *Pfizer* descubrieron que era mucho más eficiente para tratar la disfunción eréctil. También existen otros ejemplos que han llegado a ser rentables como el *Bupropión*, originalmente utilizado para la depresión fue reutilizado para dejar de fumar y la *Talidomida* que se creó para las náuseas matutinas y actualmente se utiliza para el mieloma múltiple [128].

El proceso de reposicionamiento está cada vez más en auge al existir una gran cantidad de información, procesos, dianas, compuestos estudiados, etc. hasta fases muy avanzadas, que pueden ser valorados en el contexto de nuevas enfermedades y que podrían eventualmente recortar significativamente el tiempo y dinero invertidos en el desarrollo de fármacos. Existen varias estrategias efectivas para el proceso de reposicionamiento: pueden centrarse en la droga, en la diana o en la misma enfermedad. Ejemplos de estas aproximaciones hay muchos, como la búsqueda de las características estructurales de las moléculas ya aprobadas para ciertas indicaciones o enfermedades, o el uso de biología de sistemas para evaluar de qué manera ciertas dianas están posicionadas con respecto a otras dianas exitosas. De la misma manera, el uso

de técnicas ómicas de alto rendimiento en las enfermedades también es utilizado para identificar fenotipos similares u opuestos que den pistas sobre otras patologías o procesos similares para los cuales si existan drogas efectivas.

3.2.3. La herramienta *NFFinder*

Es en el último punto donde se enmarca esta aportación de la tesis, que, aprovechando toda la experiencia, estudios y resultados acumulados durante los últimos años en diferentes grupos de investigación diseñamos un sistema computacional para el reposicionamiento de fármacos haciendo uso de técnicas ómicas.
A partir de un fenotipo de interés (por ejemplo, una enfermedad rara) *NFFinder* busca y evalúa los datos existentes para encontrar condiciones (enfermedades, drogas, líneas celulares tratadas con compuestos, etc.) que producen fenotipos similares o antagónicos susceptibles de estar relacionados con el proceso buscado. De esta manera es posible la identificación de drogas, compuestos y enfermedades relacionadas que servirían de candidatos para ser reposicionados.

3.2.3.1. Construcción de la base de datos de NFFinder

NFFinder utiliza perfiles de expresión para comparar los fenotipos. El perfil de expresión es la medida de actividad de miles de genes simultáneamente, creando una imagen global de la función celular para un fenotipo concreto. Los perfiles se obtienen de realizar el estudio de análisis diferencial en experimentos de transcriptómica como el análisis de chips de ADN o los experimentos de *RNA-seq*. Para este proyecto se han utilizado diferentes fuentes de información para la obtención de datos como *DrugMatrix* [129], *GEO* y *Connectivity Map (CMap)*[122].
La base de datos *GEO (Gene Expression Omnibus)* del *NCBI (National Center for Biotechnology Information)* [69] es un repositorio público de datos experimentales de *microarrays*, secuenciación masiva y otras formas de datos de genómica funcional que contienen más de 2.000.000 de muestras. Estos datos son introducidos por la comunidad científica con el objetivo de facilitar una evaluación independiente de los resultados o poder volver a analizar los datos. Los datos utilizados para este estudio son aquellos que han sido revisados manualmente e introducidos en los llamados *GEO DataSets (GDS)*. Estos datos contienen información adicional sobre los factores presentes en el experimento, originalmente introducidos para poder realizar tareas de comparación y agrupamiento dentro de estos datos. Los factores pueden representar cualquier tipo de agrupamiento de muestras, desde tiempos a tipo de células o tejidos, compuestos utilizados, etc., y son los que se utilizan para comparar las muestras y poder crear perfiles de expresión.
Con el lenguaje *R* se creó un flujo para analizar los 3.254 *GDS* con los que obtuvimos 16.432 firmas moleculares o perfiles de expresión. Estos datos fueron descargados con el paquete de *R GEOquery* [70]. Dentro de cada experimento se analizaron las palabras clave para encontrar muestras significativas como los controles y así poder realizar comparaciones del resto de muestras contra esta. En el caso de no haber control,

se compararon todas contra todas, lo que llevó a comparaciones sin sentido que el usuario tendrá que descartar cuando aparezcan en los resultados. El proceso de generación de perfiles consistió en el análisis diferencial entre las muestras utilizando el paquete *Limma* [130]. Después se ordenaron los genes por *t-valor*, o *fold-change* en el caso de no existir suficientes réplicas para hacer un análisis estadístico, creando el perfil de expresión.

Drug Matrix y *Cmap* son bases de datos especializadas que contienen datos de expresión de muestras tratadas con fármacos u otros compuestos químicos. La idea principal de estos proyectos es evaluar diferentes fármacos en diferentes condiciones y tejidos. Aparte de *GEO*, utilizamos estas bases de datos de las que obtuvimos 6100 firmas moleculares de *Cmap* y 5288 de *DrugMatrix*, alcanzando un total de 27820 firmas moleculares o perfiles de expresión.

3.2.3.1.1. Etiquetado de los términos relacionados con fármacos y enfermedades

Las firmas se etiquetaron según si el experimento del que se obtuvo la firma trataba de estudiar un fármaco o analizar una enfermedad. En el caso de *Cmap* y *DrugMatrix* el etiquetado fue sencillo ya que son bases de datos que contienen experimentos de compuestos específicos en diferentes líneas celulares o tejidos de ratas.

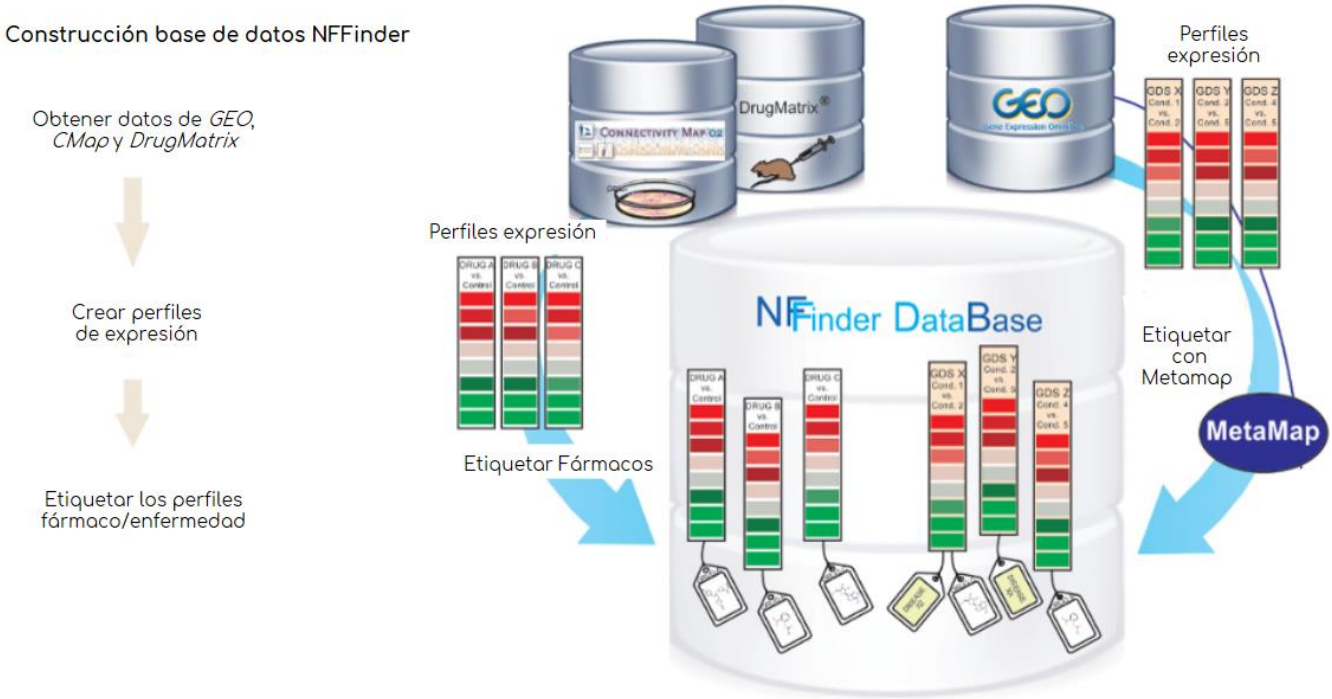


Figura 28. Esquema de la construcción de la base de datos de NFFinder [130].

En cambio, las firmas extraídas de *GEO* fueron un poco más complicadas de conseguir porque los experimentos contienen diferentes condiciones experimentales y clases, y es difícil discernir qué información etiqueta correctamente el experimento. Este problema lo solventamos utilizando *Metamap*, una herramienta

de análisis de lenguaje natural especializada en texto biológico [131]. A partir del texto de las descripciones del experimento de cada *GEO DataSet* se extrajeron los términos. Estos términos se filtraron manualmente eliminando aquellos que no tenían sentido o no eran relevantes para el estudio, quedándonos con los relacionados con fármacos y enfermedades. Esta lista final se utilizó para etiquetar cada perfil de expresión obtenido de *GEO*. El flujo de la creación de la base de datos se muestra en la Figura 28.

3.2.3.1.2. miARN *a* ARNm

NFFinder también acepta una lista de miARNs como datos de entrada. Con la utilización de las bases de datos *miRWalk* [132], *miRecords* [133], TarBase [134] y *miRTarBase* [135] recopilamos las interacciones experimentalmente validadas *miARN-ARNm* y estas interacciones se tradujeron al formato común *miRBase*, para asociar miARN a ARNm [136].

Como la mayoría de interacciones miARN-ARNm conocidas son de inhibición, de aquellos miARNs introducidos como sobreexpresados inferimos la lista de genes inhibidos y al revés, de los miARNs introducidos como inhibidos inferimos la lista de genes sobreexpresados, para proceder con el análisis descrito anteriormente como si de genes se trataran.

3.2.3.1.3. Identificación de expertos

Con el objetivo de promover las colaboraciones y la investigación de la Neurofibromatosis se relacionaron los perfiles de expresión con los autores de los experimentos. Esta información se sacó del campo "*citation"* de los *GEO Datasets* y se anotaron en cada uno de los perfiles, pudiendo así relacionar fenotipo con investigador. En los resultados, se asignó una puntuación a los expertos según el número de veces que aparecían sus estudios en los resultados.

3.2.3.2. Comparación de perfiles

NFFinder es una herramienta web para poder crear hipótesis para el reposicionamiento de fármacos basándose en comparaciones de perfiles de expresión. A partir de genes diferencialmente expresados de un experimento, la herramienta permite realizar consultas a la base de datos realizando comparaciones de firmas de expresión con un método similar al utilizado en la herramienta *MARQ* [121]. Este método utiliza una estadística de rango para asignar una puntuación y un p-valor a las dos listas de genes introducidas: genes sobreexpresados e inhibidos. Cada una de estas listas recibe una puntuación al compararse con un perfil de expresión, premiando a los genes de estas listas que se encuentran también diferencialmente expresados de una forma significativa en el perfil comparado.

La puntuación se calcula utilizando una estadística ponderada como la de *Kolmogorov-Smirnov*. Una puntuación positiva indica que la firma de expresión de la base de datos comparte una proporción significativa de genes sobreexpresados e inhibidos con los introducidos por el usuario, es decir que la mayoría

de los genes sobreexpresados e inhibidos introducidos también se encuentra así en la firma. En cambio, una puntuación negativa significa la situación opuesta: que los genes sobreexpresados se encuentran inhibidos en la firma, y los inhibidos se encuentran sobreexpresados.

Una vez calculadas todas las puntuaciones entre la entrada y las firmas, las puntuaciones se escalan entre 0 y 100, siendo 100 la puntuación que correspondería a la firma más similar en el caso de una comparación directa, y la más opuesta en el caso de una comparación inversa. La significación estadística de cada puntuación se calculó para cada firma utilizando un test de permutaciones aleatorias, creando listas de genes del mismo número que la de entrada. Como miles de firmas son comparadas con los datos de entrada y esto puede dar falsos positivos, el p-valor fue corregido por *FDR*.

3.2.3.3. Plataforma de integración y visualización de los datos

La interfaz de entrada de la herramienta es la que se muestra en la Figura 29. *NFFinder* está configurada para introducir los datos necesarios para el análisis de una forma sencilla y clara. Existen dos campos donde poner los genes o miARNs diferencialmente expresados y un campo para especificar qué tipo de datos son; también se debe seleccionar en que base de datos se quieren realizar las búsquedas y si se quiere obtener una búsqueda directa o inversa.

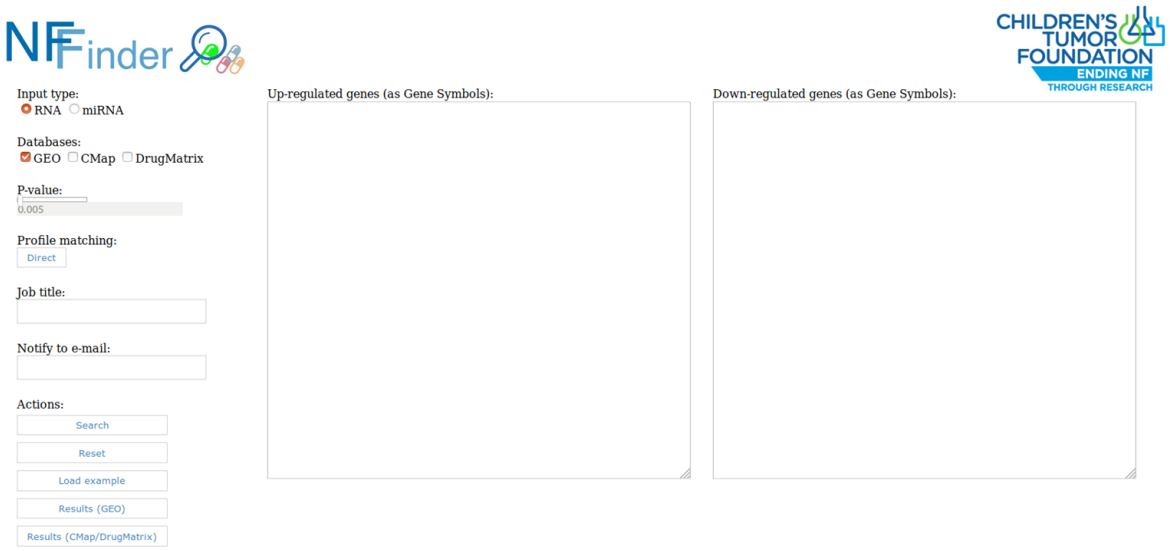


Figura 29. Pantalla inicio herramienta NFFinder.

Para la implementación de la herramienta y la visualización de resultados se utilizó la plataforma *Spotfire*. *Spotfire* es una plataforma de análisis que ayuda a explorar los datos y conectarlos con una mayor rapidez y de una forma sencilla e interactiva. Con el objetivo de facilitar la visualización de los datos para generar hipótesis que permitan avanzar en el tratamiento de la neurofibromatosis, o de otras enfermedades sin tratar, se generaron distintas visualizaciones para la mayor comprensión de los datos (ver Figura 30).

Estas visualizaciones consisten en: (A) una lista con los resultados obtenidos detallados y ordenados según la puntuación obtenida, (B) Un resumen de los resultados, (C) Los experimentos obtenidos etiquetados como drogas o enfermedades (D) un *heatmap* en los que se representa la expresión de genes de interés (E) relaciones enfermedad-droga que se han encontrado en los resultados y (F) gráfico con los autores de los datos de los resultados. Las visualizaciones se encuentran conectadas entre sí, eso significa que, si se selecciona algún experimento, los datos seleccionados se marcan en las diferentes visualizaciones.

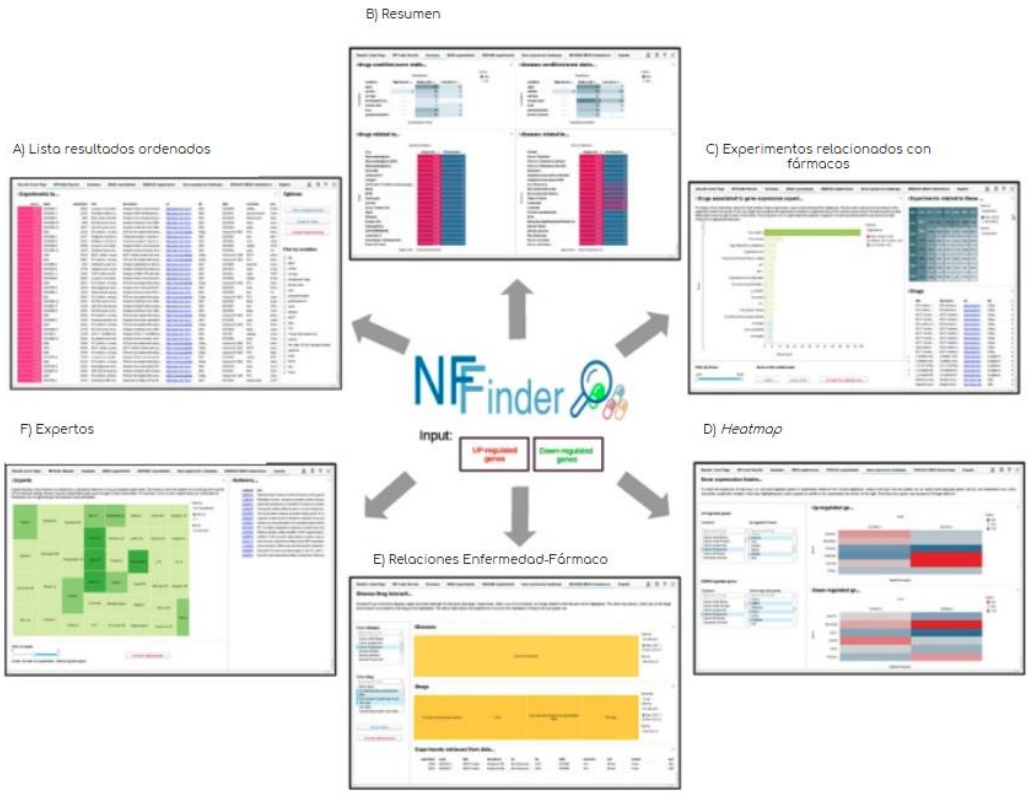


Figura 30. Resultados de salida generados por NFFinder con Spotfire.

3.2.4. Resultados

3.2.4.1. Caso de uso: Reposicionamiento de fármacos en Neurofibromatosis

Con el objetivo de encontrar algún fármaco que pueda ayudar en la búsqueda del tratamiento de la neurofibromatosis y poder ilustrar la funcionalidad de la aplicación, utilizamos un estudio [137] en el que se utilizan *microarrays* para comparar células *MPNST* con células *Schwann* normales. Los genes diferencialmente expresados del estudio se analizaron con *NFFinder* para tratar de encontrar alguna droga que pudieran revertir el fenotipo introducido. Con este propósito seleccionamos en NFFinder las bases de datos *Cmap* y *DrugMatrix* y buscamos perfiles inversos.

Como resultado se obtuvieron 775 perfiles con un total de 391 compuestos. Nos centramos en los 30 fármacos con mayor puntuación junto con los 10 más abundantes en los resultados, formando un conjunto de 32 compuestos. Un 56% de los fármacos obtenidos están relacionados con tratamientos contra el cáncer, un 12% con desórdenes neurológicos, un 12% problemas de la piel y un 3% con neoplasias benignas. El 40 % del total de fármacos relacionados con tratamientos contra el cáncer se utilizan para tratar diferentes tumores malignos del sistema nervioso, como por ejemplo la *Tricostatina A (TSA)* , que es uno de los mejor puntuados (23 veces entre los 30 mejores puntuaciones), y que muestra efectividad en otros tipos de tumores como el cáncer de mama o el carcinoma de células escamosas gracias a su capacidad de detener la proliferación celular y desencadenar la apoptosis [138,139].

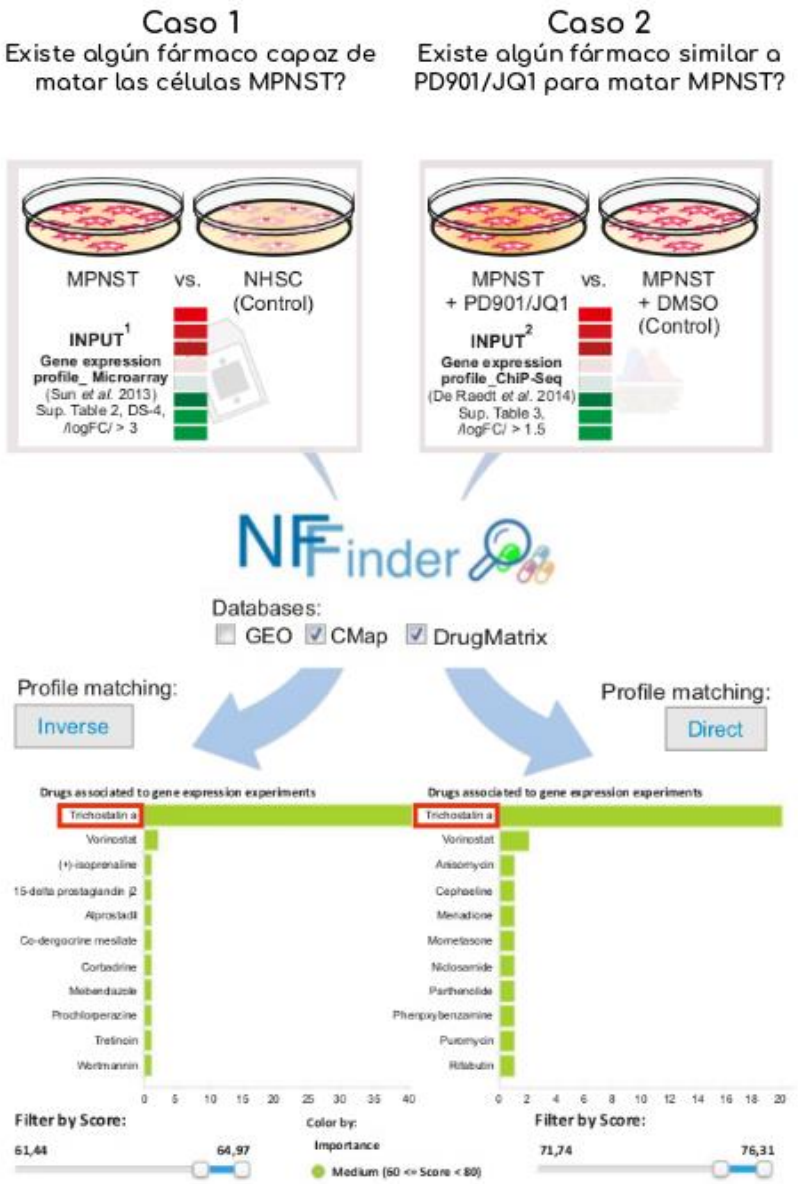


Figura 31. Descripción de los dos casos de uso ilustrados. Caso 1 datos de entrada de [137] analizando MPNST vs Control y Caso 2 datos de [140] con MPNST PD901/JQ1 vs MPNST DMSO. Visualización en NFFinder de los resultados cuando se ha realizado una búsqueda directa (Caso 1) e inversa (Caso2) a *Cmap* y *DrugMatrix*.

.

A continuación, se utilizaron los mismos datos para buscar enfermedades similares (búsqueda directa en la base de datos *GEO*). De los resultados obtenidos, descartamos entre un 20-30% de los resultados porque las comparaciones realizadas no tenían sentido. Del resto de resultados un 55% correspondían a cáncer además de obtenerse experimentos de *MPNST* en el que también se estudiaba la expresión de tumores.

Se llevó a cabo un segundo caso con datos del trabajo de [140] que comparaban células *MPNST* control (tratadas con *DMSO)* con células tratadas con *PD-901/JQ1*. El compuesto PD-901 actúa como inhibidor de la ruta *RAS-MEK* contrarrestando la deficiencia de la neurofibromina mientras que *JQ1* compensa las deficiencias de *SUZ12* o *DEE,* componentes del complejo *PCR2*. Se cree que esta combinación de medicamentos podría utilizarse en los tratamientos de *MPNST* porque mutaciones en *SUZ12* generalmente están relacionadas con ablaciones de *NF* y la deficiencia doble de *NF1* y *SUZ12* coopera para desarrollar *MPNST,* entre otros tumores. En este caso queríamos ver si existe algún tratamiento parecido a la combinación de estos fármacos que pudiera estar en el mercado para utilizarse en el tratamiento de la Neurofibromatosis. Para ello utilizamos *NFFinder* introduciendo los datos de expresión y seleccionando las bases de datos *DrugMatrix* y *Cmap* y realizamos una búsqueda directa.

También obtuvimos como mejor resultado *Trichostatin A* (*TSA)* ocupando 19 posiciones de los 30 mejores resultados. Estos datos nos reafirman que *TSA* podría ser un fármaco interesante para probar su efecto en *MPNST.*

3.2.4.2. Mejoras de NFFinder sobre otras herramientas

Para ver las mejoras de *NFFinder* respecto a otras herramientas parecidas como *Connectivity map (Cmap)* y *Combinatorial Drug Assembler (CDA)*, utilizamos un conjunto de genes diferencialmente expresados de cáncer gástrico del estudio publicado por [141]. La herramienta MARQ no ha sido evaluada porque se encuentra desactualizada y NFFinder se podría considerar una versión mejorada de esta porque utiliza más bases de datos (no únicamente GEO).

Connectivity map, además de albergar datos de expresión de cultivos de células humanas tratadas con diferentes compuestos, también permite al usuario realizar consultas de perfiles con un algoritmo parecido al utilizado por *GSEA* [81].

CDA utiliza también los datos de *CMap*. Esta herramienta parte de dos listas de genes, sobreexpresados e inhibidos, que son procesadas para posteriormente realizar un análisis de enriquecimiento de rutas de señalización. La misma realiza un análisis mucho más complejo, teniendo en cuenta las similitudes entre rutas metabólicas [142] . Pero esta forma de realizar las comparaciones depende del conocimiento acerca de estas rutas, además de las asociaciones entre perfiles y fármacos o enfermedades, por lo que el grado de error puede ser mayor, o incluso puede que se estén obviando mecanismos de señalización no conocidos hasta el momento, ya que depende de las anotaciones existentes.

Los resultados obtenidos se muestran en la Tabla 10. Los tres métodos identifican los 4 componentes *LY-294002*, *Trichostatin A*, *Tanespimycin*, *Vorinostat* con un p-valor menor a 0.005. Debido a que *CDA* hace una asociación a rutas metabólicas, los compuestos *Resveratrol* y *Trifluoperazine* no son estadísticamente significativos.

El estudio de [141] demostró mediante un análisis de expresión génica que *Vorinostat* es una nueva droga terapéutica válida para el tratamiento del cáncer gástrico. Aunque este estudio prueba que las tres herramientas son adecuadas para trabajar con hipótesis, *NFFinde*r exhibe mejoras claras respecto las otras. En primer lugar, la base de datos con la que compara perfiles es mucho más grande, incluye las de estas dos herramientas (*Cmap y CDA*) además de *DrugMatrix* y *GEO*. La inclusión de la base de datos GEO permite no solo realizar búsquedas de drogas, sino que también reorientar la pregunta buscando enfermedades con un perfil parecido para el que sí exista una droga, o buscar drogas parecidas que produzcan menos toxicidad y sí puedan ser utilizadas para tratar la enfermedad.

Estos dos sistemas requieren de un registro previo para realizar la consulta mientras que *NFFinder* no. Además de proporcionar información complementaria sobre los perfiles, como expertos con los que colaborar o relaciones enfermedad-droga, *NFFinder* muestra los resultados de un forma más clara y ordenada.

Tabla 10. Comparativa resultados con Cmap, CDA y NFFinder.

Fármaco	Orden	CMap	CDA	NFFinder
		p-valor	p-valor	p-valor
LY-294002	1	0.000	0.000	0.001
TRICHOSTATIN A	2	0.000	0.000	0.001
RESVERATOL	3	0.00016	0.4018	0.001
TRIFLUOPERAZINE	4	0.00058	0.2475	0.001
TANESPIMYCIN	5	0.0008	0.000	0.001
VORINOSTAT	6	0.00098	0.000	0.001

3.2.5. Conclusiones

Esta aportación de la tesis presenta *NFFINDER*, una herramienta para crear hipótesis de reposicionamiento de fármacos, en la que, mediante un fenotipo de interés, busca y evalúa datos existentes para encontrar

condiciones que producen fenotipos similares o antagónicos susceptibles de estar relacionados con el proceso buscado.

Es la primera herramienta que reúne las tres bases de datos *GEO, Cmap* y *DrugMatrix* reuniendo un total de 27.820 firmas diferentes: 16.432 firmas procedentes de *GEO DataSets*, 5.288 de *DrugMatrix* y 6.100 de *CMap*. Estas firmas han sido asociadas con información de compuestos, fármacos y enfermedades para poder establecer conexiones entre firmas, y permitir una visualización de los resultados más completa. Además, se ha dado importancia a los investigadores de los laboratorios que producen este tipo de datos para establecer colaboraciones entre investigadores y que así sea más fácil encontrar un tratamiento.

En concreto al estudiar el fenotipo de *MPNST* de Neurofibromatosis, *Trichostatin A* apareció como un candidato interesante a estudiar. Este resultado se consiguió al realizar una búsqueda para revertir el perfil de células *MPNST*, y al realizar una búsqueda para encontrar fenotipos parecidos a células tratadas con compuestos que se podrían utilizar para enfermedad.

3.2.6. Participación en el proyecto

El desarrollo de esta herramienta ha sido la suma del trabajo de diferentes autores. Personalmente en este proyecto he llevado a cabo la búsqueda y análisis de las bases de datos transcriptómicas existentes y la evaluación para su integración a NFFinder y así obtener más perfiles transcripcionales con los que realizar las búsquedas. Por otro lado, en un contexto más biológico, he estudiado las bases moleculares de la neurofibromatosis a través de experimentos transcriptomicos para evaluar los resultados obtenidos con NFFinder para la enfermedad y de esta forma ver también la funcionalidad de la herramienta. Por último, he participado en el diseño e implementación de las diferentes visualizaciones de los resultados en el software *Spotfire.*

3.3. Análisis proteogenómico en *Candida albicans*

C. albicans es un hongo conocido por provocar la infección Candidiasis, y que habitualmente se encuentra como comensal en la microbiota humana de los tractos intestinal y urinario. Muchos estudios de proteómica se han realizado con el objetivo de buscar biomarcadores para su diagnóstico o pronóstico, y entender los procesos biológicos en los que interviene. Actualmente, un 66% del proteoma predicho ha sido identificado [143].

Es conocido que cada individuo contiene cientos de variaciones de nucleótidos no sinónimas (*nsSNV*, *nonsynonymus single nuclotide variants*) en su genoma, llevando a polimorfismos en los aminoácidos que codifican las proteínas. Es importante detectar y cuantificar estas variaciones a nivel de proteína para estudiar la regulación post-traduccional, la expresión alélica y otros procesos biológicos importantes. Sin embargo, la mayoría de estos casos no son detectados en estudios proteómicos estándar debido a que las bases de datos que utilizan los motores de búsqueda no contienen esta información.

En este trabajo, realizado conjuntamente con el grupo de investigación de la Dra. Concha Gil de la Universidad Complutense de Madrid, nos propusimos mejorar las búsquedas por espectrometría de masas en *C. albicans* utilizando métodos de Proteogenómica. Para llevar a cabo nuestro objetivo, hemos generado una base de datos personalizada con variaciones (*SNV*, *single nucleotide variations* e *INDELS*, *insertions and deletions*) y nuevas zonas de unión de exones a partir de información genómica y transcriptómica utilizando datos de repositorios públicos. La misma ha sido utilizada en búsquedas por espectrometría de masas, con el propósito de identificar nuevos péptidos en *C. albicans*.

3.3.1. El organismo *C. albicans*

C. albicans es un microorganismo eucariota y diploide. También es polifórmico, debido a su capacidad de adquirir al menos tres morfologías diferentes: levadura (forma redonda u ovoide), hifa y pseudohifa (Figura 32Figura 32). Las dos últimas son formas elongadas consideradas filamentosas. La forma de levadura facilita la diseminación del hongo en el torrente sanguíneo, mientras que la hifa es responsable de la penetración e invasión de los tejidos [144].

C. albicans es considerado un microorganismo patógeno oportunista, porque, aunque está presente en la microbiota gastrointestinal y el tracto urinario, tiene la capacidad de invadir otros tejidos y causar enfermedad cuando se crea un desequilibrio con respecto al resto de microorganismos con los que convive en nuestro organismo. Estas infecciones se conocen como Candidiasis y suelen surgir cuando el sistema inmune del hospedador se encuentra debilitado, permitiendo a C. *albicans* proliferar y propagarse, provocando enfermedades de gravedad variable. Las infecciones más benignas se caracterizan por un sobrecrecimiento local en la piel y en las superficies mucosas, observándose cambios en la microbiota normal y dando lugar a infecciones superficiales irritantes. Las infecciones más graves son aquellas producidas cuando el hongo

consigue acceder al torrente sanguíneo, que puede ocurrir en individuos con deficiencias graves en la inmunidad celular (por ejemplo, en pacientes de SIDA), y pueden llegar a ser letales.

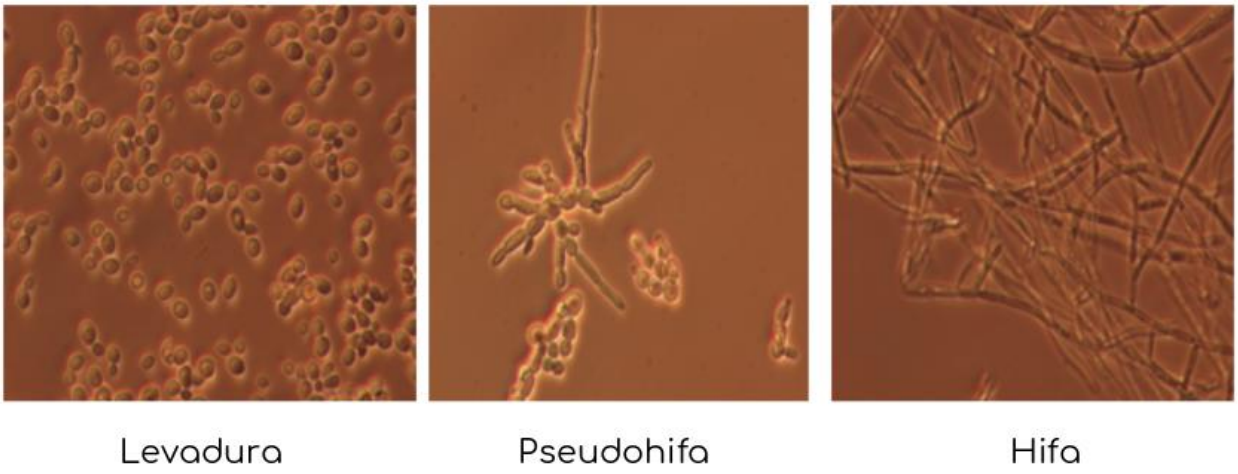


Figura 32. Tres formas morfológicas de *C. albicans*. En forma de levadura (1) dónde las células son pequeñas y forma redonda. La forma pseudohifa (2) las células van obteniendo una forma más alargada y por ultimo hifa (3) corresponden a células alargadas. Imagen tomada de [145]

El genoma diploide de *Candida* se encuentra anotado en la base de datos C*andida Genome Database (CGD)* que es actualmente el repositorio de datos más completo de las especies del género *Candida*. El genoma de *C. albicans SC5314* (A22) se presenta como un ensamblaje a nivel de cromosoma y existen las secuencias de los alelos A y B para cada cromosoma. Este ensamblaje surgió porque permite análisis más sensibles y de especificidad alélica [145]. Según *CGD*, el genoma contiene 6.198 *ORFs* (*Open reading frame*, Marco abierto de lectura) (30 Mayo, 2018) de los cuales 1.678 están anotados como *ORFs* verificados en los que existe evidencia experimental de un producto génico, y 4.368 como no caracterizados lo que indica que no existe evidencia suficiente para afirmar que existe un producto génico [146] .

Se han llevado a cabo un elevado número de experimentos proteómicos desde diferentes puntos de vista y múltiples aproximaciones, pero estos datos confirman que aún, la parte del proteoma predicho, sigue sin ser conocido o no está bien anotado. Una buena caracterización del microoganismo sería de gran valor porque permitiría incrementar el conocimiento de proteínas involucradas en los mecanismos de virulencia e infección y esto, ayudaría a diseñar nuevas estrategias para el diagnóstico y el tratamiento de la Candidiasis.

3.3.2. Creación de una base de datos proteogenómica para *C. albicans*

3.3.2.1. Obtención de datos de *RNA-seq*

El valor añadido de este trabajo reside en la utilización de varios experimentos de expresión génica de diferentes laboratorios. Utilizar y combinar estos datos permitió crear una base de datos con la que realizar

búsquedas de espectrometría de masas de manera más global, es decir, menos específica de experimento y así poder ser utilizada por toda la comunidad científica de investigadores de *C. albicans.*

Tabla 11. Lista de experimentos de RNA-seq utilizados en el estudio. (*Referencia)

Experimento*	Número De muestras	Año del estudio	Plataforma	Identificador Muestras
GSE73409 [148]	5	2015	Illumina HiSeq 2000	SRR2513862, SRR2513863, SRR2513864, SRR2513865, SRR2513866
GSE71902 [149]	2	2015	Illumina HiSeq 2500	SRR2153488, SRR2153489
GSE49310 [150]	4	2013	Illumina HiSeq 2000	SRR944219, SRR944229, SRR944234, SRR944238
GSE52236 [151]	1	2013	Illumina Genome Analyzer IIx	SRR1027797, SRR1027798
GSE38426 [152]	6	2012	Illumina HiSeq 2000	SRR503465, SRR503466, SRR503467, SRR503471, SRR503472, SRR503473
GSE41749 [153]	6	2012	Illumina Genome Analyzer IIx, Illumina HiSeq 2000	SRR604748, SRR604749, SRR771361, SRR771362, SRR771363, SRR771364
GSE35233	3	2012	Illumina Genome Analyzer IIx	SRR400661, SRR400662, SRR400663
GSE68477 [154]	3	2015	Illumina HiSeq 2500	SRR2005826, SRR2005827, SRR2005828
GSE56174 [155]	6	2014	Illumina HiSeq 2000	SRR1204813, SRR1204814, SRR1204815, SRR1204816, SRR1204817, SRR1204818
GSE56091 [156]	12	2014	Illumina HiSeq 2000	SRR424346, SRR424348, SRR424571, SRR424574, SRR419386, SRR419388, SRR420197, SRR420192, SRR420196, SRR420198, SRR420200, SRR423934
GSE53073 [157]	4	2013	Illumina Genome Analyzer IIx	SRR1044400, SRR1044401, SRR1044402, SRR1044403

Para la obtención de datos de secuenciación masiva se utilizó *GEO*, un repositorio de datos de expresión génica introducidos por la comunidad científica [69,147]. *GEO* contiene datos de secuenciación masiva para analizar expresión y regulación génica, epigenómica y otros aspectos de la genómica funcional en los que se utilizan métodos como *RNA-seq*, *ChIP-seq*, *miRNA-seq, RIP-seq, HiC-seq, methyl-seq,* etc. Este recurso contiene datos crudos y datos procesados. Estos se pueden descargar gratuitamente en diferentes formatos.

La búsqueda se realizó utilizando los términos “C. albicans AND RNA-seq” y los resultados obtenidos se revisaron, procesando todos los componentes del estudio, para finalmente quedarnos con 53 muestras control de 11 experimentos diferentes [148–157](Tabla 11). Éstas muestras se descargaron con las herramientas del paquete sra-toolkit [158] que permiten acceder y descargar los datos en diferentes formatos como ABI SOLiD native, fasta, fastq, sff, sam o Illumina native. Técnicamente, se utilizó la herramienta fastq-dump para obtener los datos en formato fastq y poder proceder a comprobar su calidad.

3.3.2.2. Análisis de datos para la obtención de variantes

Los datos de expresión génica obtenidos del repositorio público *GEO* fueron procesados con el objetivo de identificar diferentes isoformas, variaciones y/o nuevas anotaciones para ser utilizadas en las bases de datos de proteómica. Esta nueva base de datos fue utilizada por el motor de búsqueda para combinar con los espectros de *C. albicans*. El análisis de datos contiene diferentes pasos que incluyen: la evaluación de la calidad de las secuencias, el mapeo, el ensamblaje y la identificación de variaciones o nuevas uniones.

3.3.2.2.1. Calidad de las secuencias

Después de descargar los ficheros, la calidad de las secuencias se comprobó utilizando el programa *FASTQC* [14], que permite visualizar una serie de parámetros relativos a la calidad de las lecturas secuenciadas para saber si es necesario reducir la longitud de las lecturas de secuenciación o eliminar artefactos de la secuenciación.

Los ficheros con formato *Fastq* contienen las secuencias de las lecturas y sus correspondientes valores de calidad de secuenciación representados en escala de *Phred*.

$$Phred\ score = -10\log(prob\ error)$$

El programa FASTQC está dividido en diferentes módulos que analizan la calidad de la secuenciación de cada nucleótido individual de las lecturas, identifican secuencias sobrerrepresentadas que podrían corresponder a adaptadores de la librería de secuenciación o secuencias contaminantes, el contenido en GC, la distribución del tamaño de las secuencias, entre otros aspectos.

Es muy importante que las lecturas sean de buena calidad para que alineen al genoma de referencia. En el caso de que las lecturas no sean de buena calidad, existen herramientas que permiten disminuir el ruido o eliminar las secuencias de mala calidad. Para este análisis se utilizó la herramienta *fastq_quality_trimmer* de la colección *FASTX-ToolKit* [16] con las opciones *Q33 -l 50 -t 28*. Estos parámetros cortaron los nucleótidos de una calidad inferior a 28 y eliminaron las secuencias de una longitud menor a 50 nucleótidos, utilizando la codificación de calidad Phred33.

Utilizando estos parámetros se eliminaron las muestras S*RR1044400, SRR1044401, SRR1044402, SRR1044403* del experimento *GSE53073* y *SRR1027798* del experimento *GSE52236*, obteniendo un conjunto final de 48 muestras de 10 diferentes experimentos.

3.3.2.2.2. Procesamiento de RNA-seq

Después de comprobar la calidad de las secuencias y eliminar el ruido, las lecturas resultantes fueron alineadas con el programa *TopHat* [4,18] utilizando como referencia el genoma de *C. albicans SC5314* (A22) de *Candida Genome Database* [146] e indicando como longitud mínima y máxima de los intrones 20 y 800, respectivamente. Los archivos *BAM* resultantes fueron ordenados e indexados con el paquete de herramientas *SAMtools* [38] y visualizados con el programa *IGV* [106].

3.3.2.2.3. Detección de SNV e INDELS

Para obtener las variaciones se utilizó la herramienta *mpileup* de *SAMtools* y a continuación se filtraron aquellas variaciones que tenían una calidad inferior a 10 con *vcfutils.pl* [159]. La calidad de las variaciones también se representa utilizando la escala *Phred* y un valor de 10 indica que 1 de cada 10 variaciones puede ser errónea. Las variaciones resultantes, tanto puntuales como inserciones o deleciones, fueron anotadas con el programa *Snpeff* [40]. El flujo utilizado se muestra en la Figura 33.

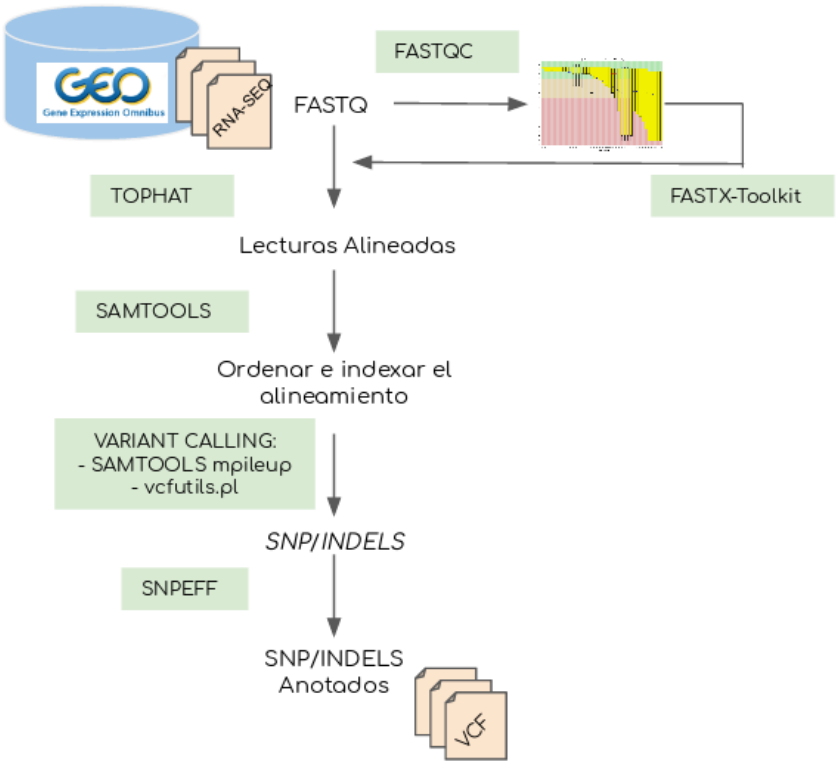


Figura 33. Flujo de análisis transcriptómico para obtener SNV e INDELS.

Existen variaciones codificantes y no codificantes, según en qué región del genoma se encuentren. Las mutaciones codificantes se pueden clasificar en tres tipos: sinónimas, que no provocan un cambio en el

aminoácido codificado y por lo tanto no varía la función de la proteína; las mutaciones denominadas "sin sentido", que son en las que se ha añadido un codón de terminación antes del final de la proteína, truncándola y provocando que pierda su función o que no se exprese; y las mutaciones "no sinónimas" que son aquellas donde el aminoácido varía y es por tanto más difícil de determinar qué efecto conlleva.

A partir de esta información seleccionamos las variaciones de tipo no sinónimas y sin sentido; y generamos péptidos de 80 aminoácidos alrededor de la mutación como en trabajos anteriores [49,160]. Seguidamente se filtraron estas mutaciones, manteniendo únicamente aquellas reportadas en más de un experimento de *RNA-seq*. A cada una de las entradas del fichero resultante *FASTA* se les añadió una cabecera con el nombre de la proteína a la que pertenece el péptido y los identificadores de los experimentos en los que detectaron las variaciones.

3.3.2.2.4. Identificación de nuevas uniones entre exones

Para encontrar nuevas zonas codificantes como nuevos genes, nuevas isoformas de un gen, distintas coordenadas de comienzo y final de regiones codificantes, etc. el genoma de *C. albicans* fue alineado nuevamente con *Tophat*, esta vez utilizando el fichero de anotación de *C. albicans* de CGD. El alineamiento se realizó dos veces, una de ellas utilizando la opción *no-novel-junctions* y otra sin activar esta opción, obteniendo dos archivos de resultados: uno con las nuevas uniones y otro con únicamente aquello que alinea al genoma de referencia.

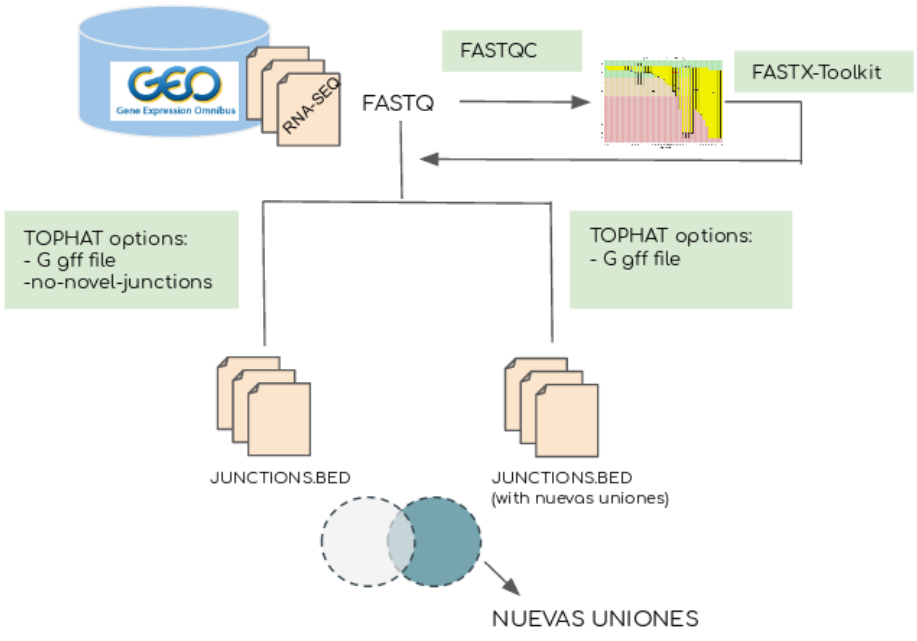


Figura 34. Flujo de análisis transcriptómico para obtener nuevas uniones

Posteriormente se compararon ambos archivos *junctions.bed* para extraer las coordenadas de las nuevas uniones, seleccionando únicamente aquellas que se encontraban en más de un experimento y contenían más de 5 lecturas de RNA-seq (Figura 34), siguiendo la estrategia descrita en los estudios [21,60]. Estas

coordenadas se alargaron 66 nucleótidos por ambos lados y las secuencias genómicas correspondientes fueron traducidas a secuencias de péptidos. La traducción se realizó utilizando los 3 marcos de lectura, teniendo en cuenta si se encontraban en la cadena negativa o positiva, de manera similar a la utilizada en [161].

Para reducir el tamaño de la base de datos y tratar de no introducir falsos positivos [49], aquellas secuencias que contenían más de 5 codones de terminación en la secuencia también se eliminaron. El identificador de cada entrada se compuso por las coordenadas de las nuevas uniones, el marco de lectura utilizado y los experimentos en los que se encontraron estas nuevas anotaciones

3.3.2.3. Base de datos con variaciones y nuevas uniones

La base de datos utilizada para realizar las búsquedas fue el resultado de anexar las proteínas de referencia, *SNV/INDELS* y las nuevas uniones, todo en formato *FASTA* (Tabla 12). Según [162] es preferible que las bases de datos se unan para utilizar únicamente una y no hacer dos análisis con dos bases de datos distintas; por ejemplo, una con las proteínas de referencia y otra con las variaciones. El estudio demuestra que cuando se utilizan dos bases de datos se pierde la competitividad entre proteínas ya que un espectro bien identificado con un péptido, podría tener una mejor identificación con la otra base de datos y esto solo puede resolverse utilizando una única referencia.

Tabla 12. Composición de la base de datos proteogenómica para *C. albicans*.

Base de datos	Número de entradas
Proteínas de referencia	12.421
SNV + INDELS	868
Nuevas Uniones	462

Cada entrada de la base de datos contiene una cabecera en formato FASTA con un identificador para saber a qué conjunto de datos corresponde, así como los datos necesarios para poder hacer un seguimiento en el caso de ser necesario.

Juntando los *SNV* encontrados y las nuevas uniones el tamaño de la base de datos incrementó únicamente un 10,7%, por lo que no debería aumentar el número de falsos positivos [163–165]. Esto es gracias a los filtrados previamente descritos, ya que el incremento de las bases de datos es uno de los problemas más importantes a los que hay que hacer frente en Proteogenómica.

3.3.2.3.1. Caracterización de la base de datos

Esta base de datos proteogenómica se generó a partir de la identificación de 48.486 mutaciones de las que 16.428, un 34%, correspondían a mutaciones sin sentido y no sinónimas. De todas estas variaciones solo se seleccionaron aquellas que se encontraban en más de una muestra, reduciendo el número de variantes a 868. Este filtrado se realizó porque nuestro principal interés era crear una base de datos que reuniera variaciones y que pudiera ser de utilidad para toda la comunidad científica que estudia *C. albicans* y de esta forma no detectaran únicamente variaciones cuando hicieran proteómica dirigida o se generaran experimentos específicos para proteogenómica (mismas muestras de *RNA-seq* y de *MS/MS*) que eleva significativamente el coste del experimento.

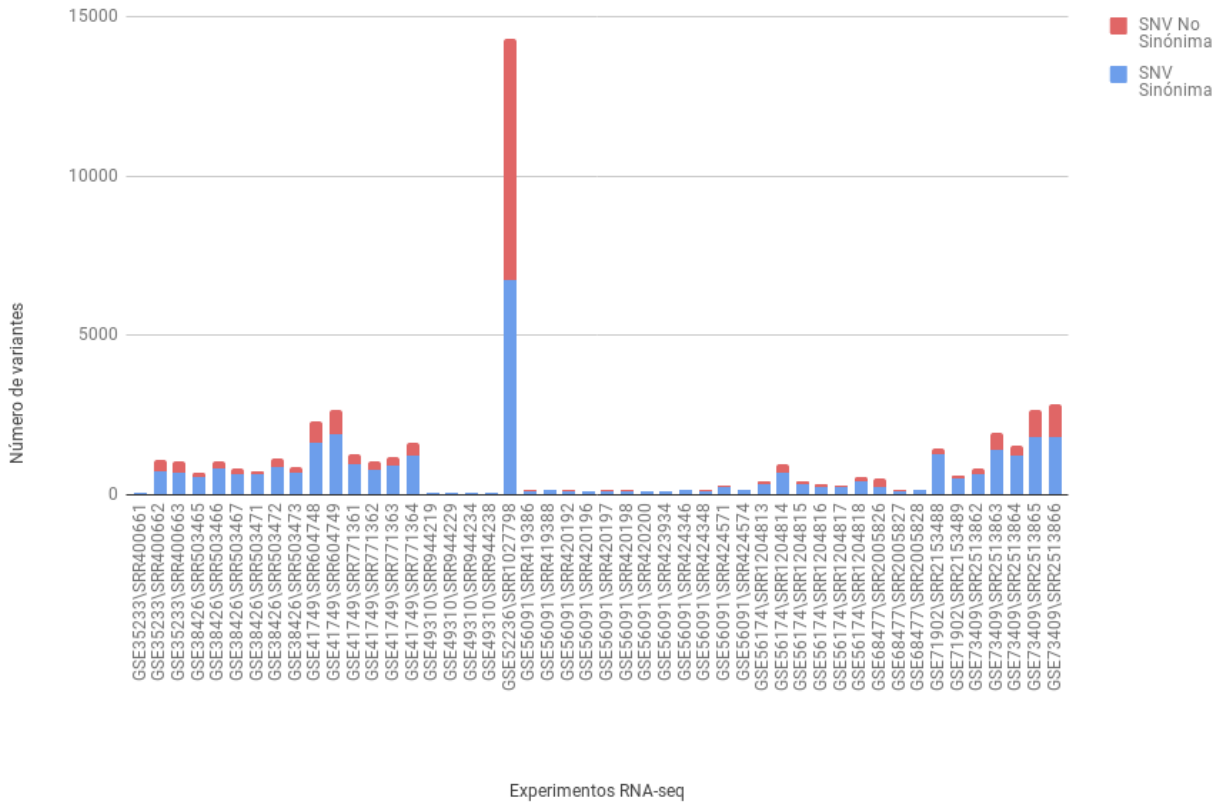


Figura 35. Representación del número de mutaciones encontradas en las diferentes muestras.

En la Figura 35 se representan todas las variantes que se encontraron en cada una de las muestras de los experimentos utilizados en el estudio. La muestra *SRR1027798* del experimento *GSE52236* destaca porque obtiene un número muy elevado de variantes comparado con los otros experimentos. El estudio en detalle de los datos no arrojó ninguna conclusión firme sobre la variabilidad de esta muestra, únicamente se vio que el número de lecturas secuenciadas totales superaba la media del resto de muestras, unas 280.000.000 lecturas, mientras que la media obtenida en los otros experimentos fue de 31.705.034 lecturas. Como se tuvo en cuenta el número de lecturas alineadas para detectar una mutación, esta alta cobertura podría explicar el

aumento, aunque pensamos que podría existir otra fuente de variabilidad. Este experimento *(GSE52236)* está compuesto únicamente por dos muestras: una en la que se obtuvo este elevado número de mutaciones y la otra que se descartó porque no superaba los filtros de calidad establecidos, debido a que, al eliminar los artefactos debidos a la secuenciación, las secuencias resultaron ser más cortas de 50 nucleótidos. Para descartar las posibles variaciones debidas a las diferencias entre experimentos, solo se incluyeron a la base de datos aquellas mutaciones que aparecían en más de un experimento, por lo que todas las variaciones de la muestra *SRR1027798* no se ven reflejadas en la base de datos final. Estos resultados se muestran en la Figura 36, donde se representa en qué experimentos se encuentran las 868 variaciones que se utilizaron para generar la base de datos de variantes.

Es de esperar que los experimentos con muchas muestras fueran los que acumularan más mutaciones cuando se aplicaron los filtros descritos, porque existe mucha variabilidad entre estudios.

Esta variabilidad puede ser debida a, por ejemplo, las diferencias en la preparación de las muestras, en los protocolos utilizados para la generación de la biblioteca de secuenciación, los efectos de lotes, etc. así como la variabilidad propia de la muestra biológica [166].

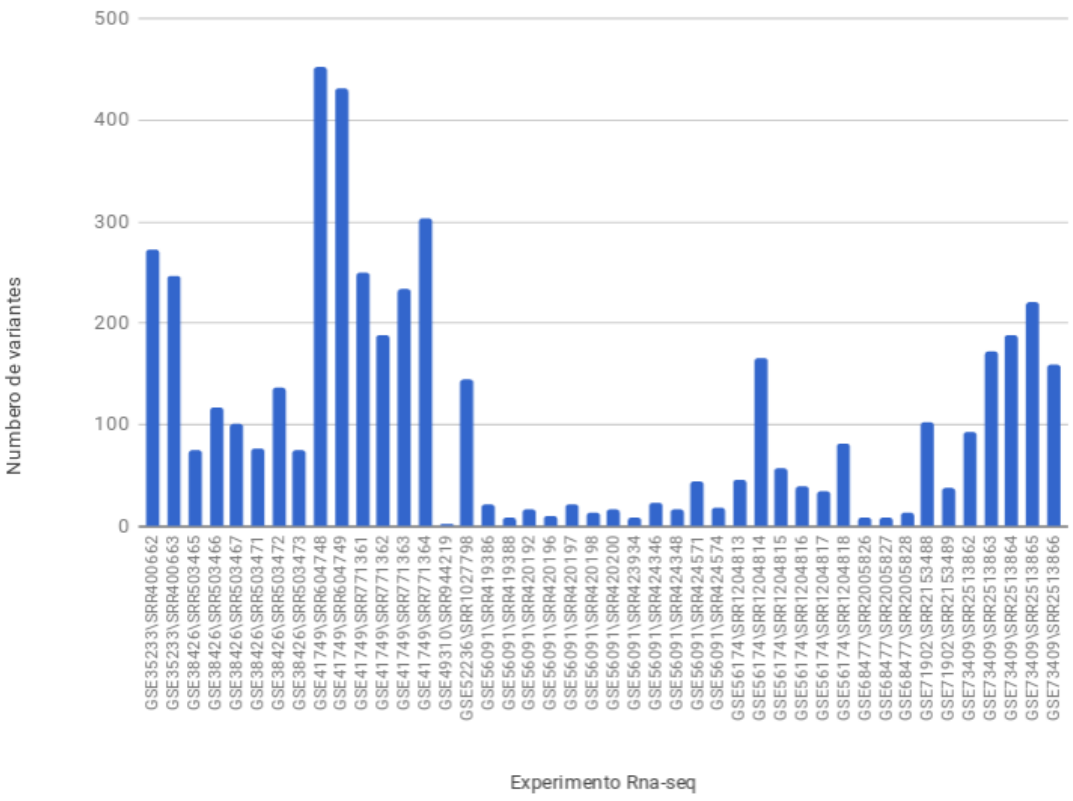


Figura 36. Representación de las variaciones filtradas en las diferentes muestras.

3.3.3. Identificación de péptidos

3.3.3.1. Experimentos de MS/MS

Los espectros utilizados se obtuvieron de *PeptideAtlas*, que es un repositorio de datos crudos de diferentes experimentos de *MS/MS* (espectrometría de masas en tándem), que además contiene los péptidos y proteínas identificadas al analizar los datos con el flujo *Trans Proteomics Pipeline (TTP*).

Tabla 13. Lista de experimentos de MS/MS utilizados en el estudio. (*Referencia)

Experimento*	Instrumento	Número de ficheros
Calb_ves_secretome [169]	LTQ XL Orbitrap	12
Calb_surfome [168]	LTQ Orbitrap Velos	14
Calb_offGel [143]	LTQ XL Orbitrap	24
Calb_subcel_fract [143]	LTQ XL Orbitrap	95
SILAC_phos_OrbitrapVelos_1 [168]	Orbitrap Velos	3
SILAC_phos_OrbitrapVelos_2 [168]	Orbitrap Velos	3
SILAC_phos_OrbitrapVelos_3 [168]	Orbitrap Velos	3
SILAC_phos_OrbitrapVelos_4 [168]	Orbitrap Velos	3
SILAC_phos_OrbitrapXL_1A_TiO2 [168]	Orbitrap XL	5
SILAC_phos_OrbitrapXL_1A [168]	Orbitrap XL	3
SILAC_phos_OrbitrapXL_1B [168]	Orbitrap XL	6
SILAC_phos_OrbitrapXL_1B_TiO2 [168]	Orbitrap XL	6
SILAC_phos_OrbitrapXL_2 [168]	Orbitrap XL	6
SIILAC_phos_OrbitrapXL_3 [168]	Orbitrap XL	6
SILAC_phos_OrbitrapXL_4 [168]	Orbitrap XL	5
Calb_acidic_subproteome [170]	LTQ	3
Calb_memb [170]	LTQ	11
Hyphal_extract_OrbitrapVelos [168]	Orbitrap Velos	4
Yeast_extract_OrbitrapVelos [168]	Orbitrap Velos	4
Calb_extract_3TOF [168]	Triple TOF	3

Este flujo emplea secuencialmente los programas *PeptideProphet, ProteinProphet* y *iProphet* [167] y recopila información de diferentes organismos como humano, ratón y diferentes levaduras [167]. Para el desarrollo de este proyecto se utilizaron los 20 experimentos [143,168–170] existentes de *C. albicans* formados por 219 ficheros mgf (Tabla 13) que se convirtieron al formato común *.MzXML* con el programa *msconverter* [53].

3.3.3.2. Asignación espectro-péptido

Para la identificación de péptidos se utilizó el motor de búsqueda *X!Tandem*, un software que permite combinar espectros de masas en tándem con secuencias de péptidos para la identificación de proteínas [171]. La estrategia general consiste en realizar una digestión teórica a las secuencias de la base de datos para que el motor de búsqueda recorra estas secuencias para cada espectro observado, seleccionando aquellos péptidos con valores de masa y carga (*m/z*) similares al del ion precursor en el espectro empírico utilizando las mismas especificaciones. Este motor de búsqueda se utilizó para analizar los 219 ficheros mgf de experimentos de C*. albicans* que fueron analizados con distinta instrumentación y especificaciones (Tabla 14).

Los resultados obtenidos se evaluaron aplicando el test estadístico denominado FDR [172]. La idea del *FDR* se basa en determinar la proporción de *PSM* (Asignación Péptido-Espectro, del inglés *Peptide to Spectrum Match*) incorrectos que se aceptan en todo el conjunto del experimento, y para ello utilizamos una base de datos señuelo o *decoy* concatenada a la base de datos original. Las bases de datos señuelo consisten en un grupo de secuencias generadas artificialmente (revirtiendo las secuencias de la base de datos) y que por lo tanto no corresponden a ninguna proteína conocida. El umbral de corte utilizado para el *FDR* fue del 0.01 y se cortó a nivel de espectro y de péptido considerando únicamente aquellos péptidos con una longitud superior a 8 aminoácidos. Por lo tanto, aceptamos 1 *PSM* incorrecto cada 100 *PSM* y volvimos a filtrar los resultados aceptando un 1 péptido incorrecto cada 100 péptidos. No filtramos a nivel de proteína porque el interés residía en la identificación de péptidos con la variación.

Tabla 14. Parámetros utilizados en las búsqueda con X!Tandem.

Parámetros	SILAC phos OrbitrapVelos/ SILAC phos Orbitrap XL	Hyphal extract OrbitrapVelos/ Yeast extract OrbitrapVelos	Calb acidic subproteome/ Calb memb	Calb extract 3TOF	Calb ves secretome/ Calb subcel fract/ Calb surfome / Calb offGel
Tolerancia en la masa del precursor monoisotópico	30 ppm	30ppm	3.1amu	50 ppm	30 ppm
Tolerancia en la masa de los iones fragmento	0.4 amu	0.4 amu	0.4 amu	0.4 amu	0.4 amu
Número de terminales trípticos	1	1	1	1	1
Número de cortes permitidos	2	2	2	2	2
Modificaciones	M Oxidation (variable); STY Phospho (variable); LR SILAC (variable); n 42.0106 (variable); C Carbamidomethyl (fixed)	M Oxidation (variable); n 42.0106 (variable); C Carbamidomethyl (fixed)	M Oxidation (variable); n 42.0106 (variable); C Carbamidomethyl (fixed)	M Oxidation (variable); n 42.0106 (variable); C Carbamidomethyl (fixed)	M Oxidation (variable); n 42.0106 (variable); C Carbamidomethyl (fixed)

3.3.4. Resultados

3.3.4.1. Péptidos identificados y validaciones

Al analizar los espectros con *X!Tandem* con la nueva base de datos identificamos nuevos péptidos prototípicos (péptidos únicos en la base de datos) que contienen la mutación. Identificamos 13 péptidos que corresponden a 11 proteínas diferentes. Por otro lado, las nuevas uniones únicamente se identificaron cuando bajamos el umbral de *FDR* al 5%. Todos los espectros fueron analizados visualmente mediante el motor de búsqueda *Mascot*, y ninguno de los espectros correspondientes a las nuevas uniones obtuvo una buena asignación al péptido. Tampoco obtuvo un buen espectro de fragmentación el péptido *YRWPKLHVVQSTK* correspondiente a la proteína *C6_00020W_B*. A continuación, se lanzaron las búsquedas con los motores de

búsqueda *PEAKS* [173] y *Mascot* para ver que puntuación asignaba estos motores a los péptidos identificados (Tabla 15) y también poder evaluar la calidad del espectro asignado de manera visual.

Tabla 15. **Péptidos prototípicos identificados con FDR 1%.** En rojo están marcados los aminoácidos mutados y en azul los aminoácidos de referencia. A continuación, se muestran las puntuaciones obtenidas en los diferentes motores de búsqueda y por último los resultados de las validaciones. Si/ No representa si el resultado obtenido confirma la mutación o no. --- indica que no se han obtenido resultados.

Descripciones		**Puntuaciones motores de búsqueda**			**Validaciones**		
Péptidos identificados	**Proteínas**	**X!Tandem**	**MASCOT**	**PEAKS**	**Espectro Visual**	**SANGER**	**Validación Espectro**
ENHVDDSLIEDAEK → (D)	C1_11200wp_a	63	87	94.31	Si	Si	---
AQYTFGSPVSEK→ (L)	C4_03190wp_b	30.6	12	--	Si	---	---
LNLANTVFSTGPIITLFYFGNGLNSGQHLCDLEISLPRGLIPR→ (S)	C5_01190wp_a	32.1	--	--	---	---	---
YRWPKLHVVQSTK → (K)	C6_00020wp_b	29.7	10	--	No	---	---
VQSFLFENHSEACTAK → (L)	GLK1	72.3	69	124.73	Si	Si	Si
KVQSFLFENHSEACTAK → (L)	GLK1	55.4	50	70.33	SI	Si	Si
GNGNGNESGGGDDDDKEEDDDDDEITEPSTSTASGDKK → (N)	MAP2	69.6	13	--	Si	---	---
GNGNGNESGGGDDDDKEEDDDDDEITEPSTSTASGDK → (N)	MAP2	66.2	60	23.13	Si	---	---
NSLITNKPGSISEAANEISQENNNNNNNK → (N)	NUP60	69.4	43	71.49	Si	---	---
KYEPNLDGPYQVQEVLGK →	ORF298	28.8	--	--	---	---	---
SNIGSGSGRGSGASGSSSGGGASINGNSIFGRSGYDDEEDDDEEK → (D)	PTC2	39.2	--	--	---	---	---
NSGGGSGGGGSQTTPQFIK → (G)	RBK1	67	89	94.58	Si	Si	Si
QQQEEQQAQQSEK → (Q)	TBF1	45.7	61	81.97	Si	---	---

Las proteínas que obtuvieron una mejor puntuación fueron *GLK1*, *MAP2*, *NUP60*, *RBK1* y *C1_11200W_A*. Con el objetivo de obtener más evidencias que confirmasen estas mutaciones encontradas llevamos a cabo dos validaciones:

1. Se validaron los espectros mediante péptidos sintéticos. Para ello se sintetizaron los péptidos con el objetivo de comparar los espectros de fragmentación obtenidos de los péptidos sintéticos y de los péptidos candidatos, y ver si los íones carcaterísticos de una de las series "y" o "b" son coincidentes entre sí. Por las características de los péptidos y la metodología del método, no todos pudieron ser sintetizados, únicamente *GLK1*, *RBK1 y C1_11200W_A* (Anexo II).
2. Se secuenció la región del ADN que contenía la mutación. Para ello se amplificaron las regiones que contenían las mutaciones mediante *PCR* (*Polimerase Chain reaction*) ajustando las condiciones en cada región, para a continuación secuenciarlas con el método de secuenciación SANGER. Los ficheros de secuenciación fueron visualizados con el software *Chromas* (Anexo III).

Los resultados obtenidos se resumen en la Tabla 15**.** Las proteínas *GLK1* y *RBK1* pudieron ser validadas por los dos métodos y se confirmaron las mutaciones. Por otro lado, el péptido de la proteína *C1_11200W_A* no se pudo sintetizar, pero los resultados obtenidos en los motores de búsqueda y con el método Sanger fueron satisfactorios debidos a las altas puntuaciones obtenidas, además de encontrar presencia de los dos nucleótidos (mutados y referencia) aproximadamente en la misma cantidad en los resultados de secuenciación, no pudiéndose discernir cuál de las dos formas es la más común.
Para concluir, los péptidos correspondientes a las proteínas *NUP60* y *MAP2* no pudieron ser validados debido a su longitud y a la composición nucleica de la región que dificultó la secuenciación, pero las puntuaciones asignadas por los motores de búsqueda además de los espectros de fragmentación obtenidos indican que la mutación podría existir.

3.3.4.2. Comparación con *PeptideAtlas*

Estos resultados se compararon con los resultados de *PeptideAtlas* para ver las diferencias obtenidas al reanalizar los datos con la nueva base de datos. *PeptideAtlas* además de albergar los ficheros crudos de *MS/MS* también contiene información sobre las proteínas y péptidos que se identificaron cuando se analizaron estos mismos datos con la base de datos de referencia. En la última versión (febrero 2015) se encontraban identificadas 4225 proteínas obtenidas por 71175 péptidos. En nuestro análisis únicamente hablamos de péptidos sin hacer inferencia a proteína porque el objetivo fue identificar las mutaciones.

En la Tabla 16 se observa la información que contenía *PeptideAtlas* de las proteínas de las que se identificaron los péptidos, además del número de experimentos de *RNA-seq* en los que se detectaron las mutaciones. Un campo de la tabla informa de si la proteína estaba considerada como identificada en *PeptideAtlas* y el otro campo refleja si el péptido que hemos identificado se había identificado anteriormente sin la mutación.

Tabla 16. Número de experimentos de *RNA-seq* en los que se detectaron las mutaciones e información de *PeptideAtlas*.

Proteína	Nº experimentos *RNA-seq*	Péptido identificado?	*Canonical Protein*?
C1_11200wp_a	10	SI	SI
C4_03190wp_b	27	NO	NO
C5_01190wp_a	2	NO	SI
C6_00020wp_b	2	NO	NO
GLK1	7	SI	SI
MAP2	9	SI	NO
NUP60	7	SI	SI
ORF298	10	NO	NO
PTC2	9	NO	SI
RBK1	23	NO	NO
TBF1	6	SI	SI

Los resultados los dividimos en dos grupos de acuerdo a la hipótesis que nos planteamos. La hipótesis se basó en la identificación del péptido, si este había sido identificado o no. Si la identificación era positiva podría ser que se estuviera identificando una variación respecto al aminoácido de referencia, *SAAV* (del inglés, *single amino acid variation*), y si no, nos planteamos que podría ser debido a que la secuencia estuviera mal anotada en la base de datos de referencia.

Dos de los casos más interesantes dentro del grupo de secuencias no identificadas fueron las proteínas *RBK1* y *C4_03190W_B* que tampoco contenían suficientes evidencias para que *PeptideAtlas* confirmara la identificación de la proteína. En estos casos las mutaciones se encontraban en muchos experimentos de *RNA-seq*, en 23 y 27 experimentos respectivamente, además también se realizó un análisis de secuencias con *BLAST* [174] y tenían 100% homología con otras cepas de *C. albicans.* Estos datos sugirieron que podría ser una mala anotación del genoma. Sin embargo, al visualizar los cromatogramas de *Sanger* para RBK1 se identificaron los dos nucleótidos, por lo que descartamos la mala anotación del genoma. En el caso de RBK1 se hizo la inferencia a proteína con la nueva base de datos y resultó en la identificación de la proteína al incorporar este péptido como identificado.

Siguiendo con aquellos casos en los que el péptido no habían sido identificado, la mutación de la proteína *ORF298* se detectó en 10 experimentos de *RNA-seq* y sorprendentemente no se había identificado ningún otro péptido de la proteína. En este caso las mutaciones encontradas correspondían a huecos en la anotación

debido a que la base de datos oficial contenía *X* en estas posiciones, por lo que supondría que no se identificó debido a una mala anotación.

Al centrarnos en aquellos péptidos en los que sí se había identificado el péptido sin la mutación, vimos que las mutaciones fueron identificadas en un rango de 6 a 10 experimentos de *RNA-seq*. *MAP2* y *GLK1*, además fueron identificadas por dos péptidos y, en dos y en tres ficheros de espectros respectivamente. Siguiendo la hipótesis descrita anteriormente, creemos que podríamos haber identificado *SAAVs* en estas proteínas. Especialmente en los casos de *GLK1 y C1_11200W_A* en los que se confirmaron las evidencias en las validaciones.

3.3.4.3. Herramienta para la creación de bases de datos para proteogenómica

Uno de los problemas principales del campo de la Proteogenómica, es que al ser un área de creación reciente no existen muchos programas para crear estas bases de datos, sobre todo para utilizar datos de especies menos estudiadas como es *C. albicans*. El programa más conocido es *CustomProDB* [60] que permite generar bases de datos de *SNV, INDELS* y nuevas uniones a exones a partir de los datos de *RNA-seq*. También existen otros programas como *PGA* [175] o *Quilts* [176]. Sin embargo, ninguno de ellos contiene el genoma de *C. albicans,* por lo que no se utilizaron en este estudio.

Para solucionar este problema, se generó un *script* que permite crear estas bases de datos a partir de un fichero FASTA, un fichero de anotaciones y uno con las variaciones encontradas. Un valor añadido del programa es que no solo crea la base de datos con mutaciones, sino que también genera la base de datos de proteínas de referencia a partir de estos ficheros. Esto es importante porque existen discrepancias entre secuencias para una misma región del genoma en las bases de datos públicas y estas discrepancias hacen difícil evaluar la incidencia y las variaciones encontradas en las proteínas [177]. Por lo tanto, solventamos este problema utilizando los mismos ficheros para alinear, buscar variantes y crear la base de datos, tanto de las proteínas de referencia como de las nuevas variantes.

El *script* utiliza como argumentos la secuencia de nucleótidos del organismo (*FASTA*), el fichero de anotaciones (*GTF* o *GFF3*) y el fichero con las mutaciones (*VCF*). También existen algunos parámetros importantes como son el tipo de secuencias finales a obtener (proteínas o péptidos mutados), y en el caso que sean péptidos, de qué longitud. Además, también se puede decidir con qué código *IUPAC* se traducen las secuencias genómicas a proteínas. Esta característica se introdujo porque el organismo *C. albicans* traduce el triplete CUG a Serina y no a Leucina como en otros organismos[178].

El fichero de salida contiene las proteínas de referencia según el fichero de anotación y los péptidos o proteínas con SNV /INDELS en formato FASTA para ser utilizados como base de datos.

3.3.5. Conclusiones

En este estudio se generó una base de datos mediante proteogenómica para *C. albicans* que permitió la identificación de diversas variantes proteicas. Se creó una base de datos completa, robusta y no específica de experimento para que pudiera ser utilizada en otras búsquedas de MS/MS de C. albicans. La generación de la base de datos fue a través del desarrollo de una herramienta en Python, que permite obtener estas bases de datos para cualquier organismo de una forma sencilla y sin necesidad de tener conocimientos de programación.

Por lo que sabemos hasta el momento, este es el primer trabajo proteogenómico en *C. albicans* y abarca todos los pasos de esta aproximación: desde la generación de la base de datos hasta la validación de los resultados, pasando por el desarrollo de una herramienta para generar estas bases de datos. Los resultados obtenidos reflejaron una mejora en las identificaciones permitiendo refinar las bases de datos de referencia. Además, los resultados fueron confirmados mediante validaciones con péptidos sintéticos y secuencias de *Sanger*, que confirmaron el potencial de la utilización de la proteogenómica.

4. DISCUSIÓN

En esta tesis doctoral se han realizado una serie de trabajos que han abordado diferentes aspectos de la bioinformática funcional, con el principal objetivo de dotar a la comunidad científica de herramientas y metodologías útiles para transformar los datos procedentes de experimentos ómicos en información útil, que explique las características del mismo y permita la generación de nuevas hipótesis. La investigación se ha desarrollado en forma de aportaciones principales, y en todos nuestros trabajos se han procesado los datos para extraer la información, y a partir de estos se han elaborado metodologías o herramientas necesarias para responder a una pregunta científica concreta. Cada una de las aportaciones contiene su propia discusión para facilitar su lectura y hemos presentado esta sección como una discusión general de la tesis en la que se resumen los diferentes trabajos presentados.

En esta Tesis se han generado dos flujos de trabajo para el análisis de datos procedentes de experimentos de *RNA-seq* y Proteómica *shotgun*. El análisis de datos de *RNA-seq* incluye el control de calidad, el alineamiento, cuantificación, y según el objetivo a alcanzar se completó el flujo de trabajo con filtrado de expresión de genes, búsqueda de mutaciones, nuevas uniones entre exones o generando perfiles de expresión. El proceso de análisis de los datos procedentes de Proteómica *shotgun* incluye la conversión de ficheros a un formato común, la utilización de motores de búsqueda y la validación estadística por *FDR.*

Según los diferentes contextos biológicos se han desarrollado diferentes estrategias metodológicas para su aplicación a los datos procesados, y así convertir estos datos biológicos en información valiosa. Todas las aportaciones han partido de datos procedentes de experimentos de *RNA-seq*.
La secuenciación de ARN o RNA-Seq es un método común para analizar la expresión génica y descubrir nuevas especies de ARN. Existen varios métodos para el análisis general de expresión génica y varias aplicaciones específicas [180], como la detección de isoformas y fusión de genes, perfiles de expresión génica, secuenciación dirigida, etc. Para cada una de las aplicaciones no existe un único flujo de trabajo ni un programa óptimo para analizar los datos, y cada punto del análisis (diseño experimental, control de calidad, alineamiento de las lecturas, cuantificación, visualización, etc.) puede ser discutido y modificado para obtener unos mejores resultados según el objetivo de la investigación y el organismo a estudiar [181,182]. Aunque el análisis más común de *RNA-seq* consiste en realizar un análisis de expresión diferencial en el que se compara la expresión de los genes en dos muestras diferentes [184]. Actualmente las tecnologías de secuenciación continúan avanzando, y se siguen desarrollado nuevos métodos para el análisis e integración de datos, ejemplo de ello son los trabajos presentados en esta tesis doctoral, y se espera que surjan nuevos métodos de RNA-Seq en el futuro [183].
Nuestra inicial contribución basó su fundamento en detectar y evaluar las posibles relaciones ente las *HSC* y los nichos durante el desarrollo de las *HSC* en el hígado fetal. Para ello centramos el flujo de análisis de datos de *RNA-seq* en filtrar los genes según su expresión estableciendo umbrales mediante el método de [108] que se basa en contar las lecturas intergénicas y contarlas como ruido. La elección del método fue un punto

importante del trabajo, el objetivo era estudiar lo que sucedía en cada uno de los estadios del desarrollo embrionario y no compararlos entre ellos, por lo que queríamos evitar la expresión diferencial que es uno de los métodos más conocidos e utilizados [182,184]. Para establecer el umbral de corte, previamente normalizamos las lecturas por RPKM. Utilizamos este método y no FPKM, TPM o CPM para imitar la estrategia que siguieron los autores del estudio. Sin embargo, la utilización del método FPKM hubiera sido más adecuada ya que se trataba de muestras paired-end y este en lugar de contar lecturas cuenta fragmentos, por lo que tiene en cuenta que dos lecturas puedan alinear en un fragmento y no cuenta las lecturas dos veces. Por otro lado, TPM es un método similar a RPKM y FPKM que cada vez se está empleando más, que al normalizar tiene en consideración la longitud del gen y la profundidad de la secuenciación, y también se hubiera podido utilizar en el análisis. Sin embargo, el problema principal del análisis residió en tener únicamente dos réplicas de cada muestra, siendo tres el mínimo de réplicas que se suelen utilizar en estos análisis, aunque según [188] se debería utilizar un mínimo de 6 réplicas por muestra.

Aunque los umbrales obtenidos eran bastante similares, las diferencias en los datos fueron significativas (Tabla 7). Otro de los puntos interesantes fue la creación de la estrategia bioinformática para la integración de datos biológicos procedentes de diferentes bases de datos como de interacciones, anotaciones y factores de transcripción entre otras. En esta metodología se trató de contextualizar la información obtenida simulando el posible funcionamiento celular y así obtener una lista de posibles procesos interesantes a estudiar, y validar aquellos que podrían estar relacionados con desarrollo de las *HSC*. La metodología es dependiente de la información contenida en las bases de datos, así que la incorporación o eliminación de elementos en éstas variarían los resultados. Es la primera vez que una metodología de estas características es desarrollada, principalmente porque se basa fundamentalmente en el diseño experimental del experimento con el objetivo de resolver una pregunta concreta en el contexto de las HSC. Sin embargo, esta metodología se podría utilizar para resolver otras preguntas científicas en otro contexto e organismos.

Los resultados obtenidos corroboraron el funcionamiento de la estrategia utilizada. La mayoría de las rutas resultantes fueron potencialmente significativas para el estudio, se clasificaron en los grupos 2 y 3, e incluso algunas las rutas ya se habían descrito como que participaban en el desarrollo de las HSC; como la ruta *VEGF SIGNALING* detectada que induce la formación de tejido hematopoyético y la proliferación celular [94] o la ruta *WNT/ WNT B-Catenin* que es conocida por su participación en la autorrenovación, proliferación y diferenciación de las células madre adultas [90,119]. También observamos proteínas secretadas como *Angptl4 y IGF-2* que promueven la expansión, diferenciación y la maduración de las HSC [118] o como *CXCL12* que se expresa en todos los estadios y se cree que es necesaria para el proceso de *homing* y la regulación y migración de las *HSC* [94,100]. Además, gracias al establecimiento de umbral de expresión pudimos observar y analizar aquellas diferencias significativas existentes entre los distintos tiempos, algunos de ellas ya descritos en otros estudios [90,100,119]. En vista del potencial de la tecnología nos propusimos realizar otro tipo de análisis con también datos de RNA-seq pero desde una perspectiva distinta, generando perfiles de expresión que es

la medida de la actividad de miles de genes simultáneamente para crear una imagen global de la función celular o fenotipo.

Bajo esta idea, llevamos a cabo nuestra segunda aportación, en la que a partir de estos perfiles generamos una herramienta web para la generación de hipótesis para el reposicionamiento de fármacos. Para ello se analizaron datos transcriptómicos, no únicamente de *RNA-seq*, también de chips de ADN obtenidos de diferentes repositorios públicos como *GEO, DrugMatrix* y *CMap*; reuniendo un total de 27.820 perfiles o firmas diferentes [179]. *NFFinder* es la primera herramienta que reúne tres bases de datos distintas y un número tan elevado de perfiles génicos asociados a información de: compuestos, fármacos, enfermedades y expertos científicos que permite establecer relaciones entre perfiles. NFFinder requiere de una lista de genes procedentes de datos de expresión génica, donde los experimentos de las distintas bases de datos son ordenados por su parecido con los datos de entrada mediante un algoritmo de reconocimiento de patrones. Llevando a cabo este procedimiento se puede a partir de un fenotipo de interés buscar y evaluar los datos existentes para encontrar condiciones que producen fenotipos similares o antagónicos susceptibles de estar relacionados con el proceso buscado, y de esta manera generar hipótesis en el reposicionamiento de fármacos.

Se realizó un caso de uso con NFFinder para ver la funcionalidad de la herramienta, éste se basó encontrar algún fármaco que pudiera usarse en el tratamiento de la neurofibromatosis. Para ello se utilizaron genes diferencialmente expresados al comparar células MPNST con células Swann normales del estudio [137], y se buscó un perfil inverso seleccionando las bases de datos CMap y DrugMatrix. Los resultados obtenidos mostraron que la Tricostatina A (TSA) podría ser un candidato interesante para revertir el fenotipo inducido, según [138,139] este fármaco se ha utilizado en otros tipos de tumores por su capacidad de detener la proliferación celular y desencadenar la apoptosis, por lo que podría ser interesante. Además de ver su funcionalidad, comparamos NFFinder con otras herramientas parecidas como Cmap [122] y CDA [142]. CDA utiliza también los datos de CMap, y también parte de dos listas de genes, pero a diferencia de NFFinder, realiza un enriquecimiento de rutas de señalización, por lo que puede aumentar el grado de error en los resultados porque depende del conocimiento de las rutas y las anotaciones existentes, además de las asociaciones entre perfiles, fármacos o enfermedades. Para comparar las distintas herramientas utilizamos genes diferencialmente expresados de cáncer gástrico obtenidos del estudio [141] que muestra que Vorinostat es una droga terapéutica válida para el tratamiento de cáncer gástrico. Los resultados obtenidos en las tres herramientas (Tabla 10) reflejan las mejoras de NFFinder versus a las otras dos. En primer lugar, la base de datos con la que se comparan los perfiles es más grande, incluyendo los perfiles de las otras dos herramientas. Además, NFFinder no solo permite realizar búsqueda de drogas, sino que también reorientar la pregunta para buscar enfermedades con un perfil parecido para el que, si existe una droga, o para buscar una droga parecida que produzca menos toxicidad y que pueda ser utilizada para tratar la enfermedad. La utilización de NFFinder no

se ciñe a un organismo concreto, se pueden realizar búsquedas para cualquier organismo, pero el resultado mostrará aquellos perfiles parecidos de los organismos que se encuentren en la base de datos.

La utilización de repositorios públicos ha sido un recurso utilizado en este trabajo de investigación por el gran valor añadido que ofrecen [185,186]. Los repositorios permiten reanalizar una gran cantidad de datos analizados en diferentes plataformas y diversas condiciones. En esta tesis doctoral, estas fuentes de información han permitido el desarrollo de una herramienta con una gran cantidad de datos experimentales como *NFFinder* y la realización de un trabajo más específico en el que hemos utilizado todos los datos generados hasta el momento de *C. albicans* teniendo en cuenta la variabilidad entre diferentes laboratorios como en la última aportación de esta tesis. La elección de los repositorios utilizados no ha sido al azar, actualmente existen diferentes repositorios genómicos como *ArrayExpress*, *dgGAP*, *Japanese Genotype-phenotype Archive*, etc. estos fueron estudiados para ver qué datos contenían y en que formato se almacenaban. En ambos estudios observamos que GEO era la base de datos con más entradas, e incluso vimos muchos datos ya contenidos en GEO en otras bases de datos por lo que se generaban redundancias y nos decantamos por únicamente utilizar GEO que es el repositorio genómico más completo [69,70]. También se utilizaron otros repositorios para complementar los diferentes objetivos propuestos; para *NFFinder* también se utilizó información de los repositorios *Drugmatrix* y *Cmap* que contienen perfiles de expresión de células tratadas con fármacos, o como en la última aportación que obtuvimos los datos de un repositorio de datos de proteómica llamado *PeptideAtlas*, que además de albergar los ficheros crudos de *MS/MS* también contiene información sobre los péptidos y proteínas que se identificaron al analizar los datos.

En esta última aportación se generó una base de datos mediante Proteogenómica para *C. albicans* que permitió la identificación de diversas variantes proteicas. Se creó una base de datos completa y robusta que contenía *SNV, INDELS* y nuevas uniones mediante la utilización de experimentos de *RNA-seq* del repositorio *GEO*. Esta fue utilizada para analizar todos los experimentos del repositorio *Peptide Atlas* de *C. albicans* en los que pudimos identificar nuevos péptidos. Para la generación y creación de esta base de datos desarrollamos un programa sencillo que puede ser utilizado sin necesidad de tener conocimientos de programación. Actualmente existen programas para la generación de estas bases de datos como *CustomProDB*, *Quilts* [176] o *PGA* [175], sin embargo, ninguno de ellos contiene el genoma actualizado de *C. albicans*, por lo que no se pudieron utilizar en el estudio. La mayor ventaja del programa que desarrollamos es la opción de poder utilizar cualquier genoma secuenciado obtenido de cualquier base de datos, ya que en muchos organismos los genomas más actualizados no siempre se encuentran en las bases de datos más conocidas, sino en bases de datos especializadas de organismo como ha sido el caso de *C. albicans* que lo hemos obtenido de *CGD* [146].

Para la utilización del programa es necesario un fichero FASTA, el fichero de anotación y el fichero .vcf con las variaciones. Existen muchas herramientas para detectar *SNV* pero las más utilizadas son *GATK* y *mpileup* de *SAMtools*. Existen varios estudios que comparan las herramientas y algunas de las diferencias entre estas

dos herramientas son: como preprocesan el alineamiento, *GATK* elimina las lecturas con baja calidad mientras que *SAMtools* las utiliza todas; el modelo de probabilidad de genotipo utilizado es distinto, el modelo de *SAMtools* deriva de *BAQ* mientras que el de *GATK* de *Dindel*, que principalmente difieren en como asumen los errores de secuencia; *SAMtools* utiliza filtros establecidos mientras que *GATK* aprende los filtros de los datos, por lo que utilizar *GATK* cuando los datos son de humano suele ser recomendable porque se tienen suficientes datos para entrenar el modelo, etc. En este estudio utilizamos *SAMtools* por dos razones principales, el tamaño del organismo y por cómo se preprocesa el alineamiento porque queríamos controlar el filtrado de lecturas.

Otro punto a discutir fue el tamaño de la base de datos para realizar las búsquedas de *MS/MS*, varios estudios [49,163-165] indican que un tamaño muy grande da lugar a un aumento de los falsos positivos. Para reducir el tamaño se filtraron algunas entradas, porque según [162] es preferible utilizar una única base de datos y no dividirla en dos y realizar dos búsquedas porque se pierde competividad entre las identificaciones. Esto es debido a que un espectro bien identificado con un péptido podría tener una mejor identificación con la otra base de datos. El filtrado consistió en incorporar variaciones no sinónimas y aquellas que se encontraron en más de una muestra reduciendo el tamaño de 48.486 a 1.330 entradas, incrementando únicamente la base de datos un 10,7% con respecto a la base de datos de referencia (Tabla 12).

Mediante esta nueva base de datos se identificaron con un FDR del 1%, 13 nuevos péptidos con *SNV* y ninguna nueva unión a exón. El hecho de no haber identificado ninguna nueva unión creemos que es porque C. albicans no es un organismo con muchos intrones [187], además de porque cortamos con un umbral de FDR muy restrictivo. Las identificaciones fueron comparadas con las identificaciones de *PeptideAtlas* y dividimos los resultados en dos grupos de acuerdo a la hipótesis que nos planteamos. La hipótesis se basó en si el péptido ya había sido identificado sin la mutación o no en *PeptideAtlas*. Si la identificación era positiva podría ser que se estuviera identificando *SAAVs*, y si no, nos planteamos que podría ser debido a que la secuencia estuviera mal anotada en la base de datos de referencia. En el grupo de los péptidos no identificados sin la mutación, encontramos las proteínas C4_03190wp_a, C5_01190wp_a, C6_00020wp_b, ORF298, PTC2 y RBK1. Entre ellas nos llamaron la atención las proteínas C4_03190wp_a y RBK1 porque las mutaciones fueron identificadas en muchos experimentos de RNA-seq, 23 y 27 experimentos respectivamente, y la secuencia identificada tenía un 100% de homología con otras cepas de *Candida* al hacer un *BLAST*. Sin embargo, al visualizar el cromatograma de Sanger de RBK1 se identificaban los dos nucleótidos, por lo que descartamos la mala anotación del genoma.

En el grupo de péptidos sí identificados encontramos las proteínas C1_11200wp_a, GLK1, MAP2, NUP60 y TBF1. Entre ellas, GLK1 y C1_11200wp_a pudimos confirmar las evidencias encontradas en las validaciones. Sin embargo, aunque las otras variaciones no pudieron validarse, al analizar los espectros visualmente obtuvieron una buena asignación al péptido por lo que creemos que son interesantes a tener en consideración.

Este estudio se realizó con *C. albicans* y se generó una base de datos más completa para las búsquedas de MS/MS. El método empleado y la herramienta desarrollada, llena el vacío que existía en el campo de la protegenómica en microorganismos y permite hacer proteogenómica en cualquier organismo para así poder identificar nuevas proteínas y /o mejorar la anotación de los genomas existentes.

No únicamente esta última aportación es aplicable al mundo de los microorganismos, cada una de los trabajos realizados también, y aunque en esta tesis se han tratado de resolver diferentes casos concretos, la idea general ha sido generar flujos de trabajo que puedan ser utilizados en otros organismos menos estudiados. Es importante destacar la diferencia de información existente en las bases de datos de organismos superiores con respecto la información de microorganismos y la necesidad que hay de generar más datos y más herramientas bioinformáticas para estrechar la brecha existente.

5. CONCLUSIONES

En esta tesis doctoral se han abordado diferentes aspectos de la bioinformática funcional. Para ello se han desarrollado una serie de metodologías y herramientas que constituyen un marco de trabajo para la investigación con datos ómicos permitiendo obtener información de interés a partir de datos biológicos. Si bien es cierto que estos trabajos representan problemas distintos y en contextos experimentales diferentes, es también importante destacar que la solución a los mismos tiene un hilo conductor común, que son las técnicas experimentales, las herramientas bioinformáticas y las metodologías de integración. Una de las enseñanzas de este trabajo es que no existe una única aproximación que pueda ser aplicada a todos los problemas y necesidades que se presentan en los laboratorios de biología molecular, siendo necesario el desarrollar y aplicar las existentes a los problemas específicos, además de generar metodologías nuevas que le integren. En este sentido se ha enmarcado esta tesis doctoral, permitiéndonos obtener resultados relevantes y alcanzar las siguientes conclusiones:

1. Se han desarrollado flujos de trabajo para el análisis de experimentos de *RNA-seq* que incluyen el alineamiento, la cuantificación, la normalización y expresión diferencial o filtrado de expresión por umbral de activación, búsqueda de variantes y nuevas uniones entre exones.
2. Se ha elaborado una metodología para la integración de datos biológicos a datos de *RNA-seq* que permite estudiar y evaluar el interactoma obteniendo una visión global y específica de su funcionalidad.
3. Se han procesado y analizado los experimentos de las bases de datos *GEO*, *Connectivity Map* y *Drug Matrix* para generar perfiles de expresión y asociarlos a fármacos, compuestos, enfermedades y/o expertos.
4. Se ha desarrollado una herramienta web para el reposicionamiento de drogas que permite comparar perfiles de expresión génica para encontrar condiciones que producen fenotipos similares o antagónicos susceptibles de estar relacionados con el proceso buscado.
5. Se ha elaborado un programa para la generación de bases de datos Proteogenómicos que incorpore variaciones y nuevas uniones, además de la base de datos de referencia, en cualquier organismo de una forma sencilla y sin necesidad de tener conocimientos de programación. Para esto, se ha desarrollado un flujo de trabajo para el análisis de experimentos procedentes de espectrometría de masas que permite la conversión de ficheros a un formato común, la utilización del motor de búsqueda *X!Tandem* y la validación estadística por *FDR*.
6. Se ha creado una base de datos Proteogenómica para el organismo *C. albicans* más completa para ser utilizada para la comunidad científica y que ha permitido la identificación de nuevos péptidos.

REFERENCIAS

1. Koussounadis, A., Langdon, S.P., Um, I.H., Harrison, D.J., and Smith, V.A. (2015). Relationship between differentially expressed mRNA and mRNA-protein correlations in a xenograft model system. Sci. Rep. *5*, 1–9.

2. Joyce, A.R., and Palsson, B.Ø. (2006). The model organism as a system: integrating “omics” data sets. Nat. Rev. Mol. Cell Biol. *7*, 198–210.

3. Hieter, P., and Boguski, M. (1997). Functional Genomics : It ’ s All How You Read It. October *278*, 601–602.

4. Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., and Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat. Protoc. *7*, 562–578.

5. Wang, Z., Gerstein, M., and Snyder, M. (2010). RNA-Seq : a revolutionary tool for transcriptomics. Nat. Rev. Genet. *10*, 57–63.

6. Nagalakshmi, U., Waern, K., and Snyder, M. (2010). RNA-Seq: a method for comprehensive transcriptome analysis. Curr. Protoc. Mol. Biol. *Chapter 4*, Unit 4.11.1-13.

7. Zhu, M., Dahmen, J.L., Stacey, G., and Cheng, J. (2013). Predicting gene regulatory networks of soybean nodulation from RNA-Seq transcriptome data. BMC Bioinformatics *14*, 278.

8. Marot, G., Castel, D., Estelle, J., Guernec, G., Jagla, B., Servant, N., Jouneau, L., Laloe, D., Gall, C. Le, and Schae, B. (2012). A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. Brief. Bioinform. *14*, 671–83.

9. Capobianco, E. (2014). RNA-Seq Data: A Complexity Journey. Comput. Struct. Biotechnol. J. *11*, 123–130.

10. Garber, M., Grabherr, M.G., Guttman, M., and Trapnell, C. (2011). Computational methods for transcriptome annotation and quantification using RNA-seq. Nat. Methods *8*, 469–77.

11. Zhao, S., Baohong Zhang, Y.Z., William Gordon, Sarah Du, T.P., and Schack, M.V. and D. von (2012). Bioinformatics for RNA-Seq Data Analysis. In Bioinformatics - Updated Features and Applications, I. Y. Abdurakhmonov, ed. (United States of America: INTECH), p. 27.

12. Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M.W., Gaffney, D.J., Elo, L.L., Zhang, X., *et al.* (2016). A survey of best practices for RNA-seq data analysis. Genome Biol. *17*, 1–19.

13. Schmieder, R., and Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. Bioinformatics *27*, 863–864.

14. Andrews Simon (2015). FastQC: a quality control tool for high throughput sequence data.

15. Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal *17*, 10.

16. Gordon, A., and Hannon, G. (2010). FASTX-Toolkit. FASTQ/A short-reads pre-processing tools. Available at: http://hannonlab.cshl.edu/fastx_toolkit/.

17. Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120.

18. Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S.L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. *14*, R36.

19. Wu, T.D., and Nacu, S. (2010). Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics *26*, 873–881.

20. De Bona, F., Ossowski, S., Schneeberger, K., and Rätsch, G. (2008). Optimal spliced alignments of short sequence reads. Bioinformatics *24*, 174–180.

21. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: Ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.

22. Huang, S., Zhang, J., Li, R., Zhang, W., He, Z., Lam, T.W., Peng, Z., and Yiu, S.M. (2011). SOAPsplice: Genome-wide ab initio detection of splice junctions from RNA-Seq data. Front. Genet. *2*, 1–12.

23. Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. *10*.

24. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760.

25. Li, R., Li, Y., Kristiansen, K., and Wang, J. (2008). SOAP: Short oligonucleotide alignment program. Bioinformatics *24*, 713–714.

26. Costa, V., Angelini, C., De Feis, I., and Ciccodicola, A. (2010). Uncovering the complexity of transcriptomes with RNA-Seq. J. Biomed. Biotechnol. *2010*, 19.

27. Li, B., and Dewey, C.N. (2011). RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics *12*.

28. Nicolae, M., Mangul, S., Măndoiu, I., and Zelikovsky, A. (2010). Estimation of alternative splicing isoform frequencies from RNA-seq data. Algorithms Mol. Biol. *9*, 202–214.

29. Liao, Y., Smyth, G.K., and Shi, W. (2014). FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics *30*, 923–930.

30. Anders, S., Pyl, P.T., and Huber, W. (2015). HTSeq-A Python framework to work with high-throughput sequencing data. Bioinformatics *31*, 166–169.

31. Zhao, S., Xi, L., and Zhang, B. (2015). Union exon based approach for RNA-seq gene quantification: To be or not to be? PLoS One *10*, 1–21.

32. Kanitz, A., Gypas, F., Gruber, A.J., Gruber, A.R., Martin, G., and Zavolan, M. (2015). Comparative assessment of methods for the computational inference of transcript isoform abundance from RNA-seq data. Genome Biol. *16*, 1–26.

33. Simon Anders, and Wolfgang Huber (2010). Differential expression analysis for sequence count data. Genome Biol. *11*, R106.

34. Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2009). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139–140.

35. Robinson, M.D., and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. *11*, R25.

36. Cumbie, J.S., Kimbrel, J.A., Di, Y., Schafer, D.W., Wilhelm, L.J., Fox, S.E., Sullivan, C.M., Curzon, A.D., Carrington, J.C., Mockler, T.C., *et al.* (2011). GENE-counter: A computational pipeline for the analysis of RNA-seq data for gene expression differences. PLoS One *6*.

37. Trapnell, C., Hendrickson, D.G., Sauvageau, M., Goff, L., Rinn, J.L., and Pachter, L. (2013). Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat. Biotechnol. *31*, 46–53.

38. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

39. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., *et al.* (2009). The Genome Analysis Toolkit: A MapReduce framework for

analyzing next-generation DNA sequencing data. Genome Res. *20*, 254–260.

40. Cingolani P, Platts A, Wang LL, et al. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w 1118 ; iso-2; iso-3. Fly (Austin). *6*, 80–92.

41. McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R.S., Thormann, A., Flicek, P., and Cunningham, F. (2016). The Ensembl Variant Effect Predictor. Genome Biol. *17*, 1–14.

42. Merrill, S. a., and Mazza, A.-M. (2006). Reaping the Benefits of Genomic and Proteomic Research: Intellectual Property Rights, Innovation and Public Health (Washington, DC: The National Academies Press).

43. Girolamo, F., Lante, I., Muraca, M., and Putignani, L. (2013). The Role of Mass Spectrometry in the "Omics" Era. Curr. Org. Chem. *17*, 2891–2905.

44. Glish, G.L., and Vachet, R.W. (2003). The basics of mass spectrometry in the twenty-first century. Nat. Rev. Drug Discov. *2*, 140–150.

45. Yalcin, E.B., and de la Monte, S.M. (2015). Review of Matrix-Assisted Laser Desorption Ionization-Imaging Mass Spectrometry for Lipid Biochemical Histopathology. J. Histochem. Cytochem. *63*, 762–771.

46. Ho, C.S., Lam, C.W.K., Chan, M.H.M., Cheung, R.C.K., Law, L.K., Lit, L.C.W., Ng, K.F., Suen, M.W.M., and Tai, H.L. (2003). Electrospray ionisation mass spectrometry: principles and clinical applications. Clin. Biochem. *24*, 3–12.

47. Yates, J.R. 3rd (1998). Database searching using mass spectrometry data. Electrophoresis *19*, 893–900.

48. Gevaert, K., Vandekerckhove, J., Sciences, H., and Vandekerckhove, J. (2000). Protein identification methods in proteomics. Electrophoresis *21*, 1145–54.

49. Sheynkman, G.M., Shortreed, M.R., Frey, B.L., Scalf, M., and Smith, L.M. (2014). Large-scale mass spectrometric detection of variant peptides resulting from nonsynonymous nucleotide differences. J. Proteome Res. *13*, 228–240.

50. Nesvizhskii, A.I. (2007). Mass Spectrometry Data Analysis in Proteomics First. R. Matthiesen, ed. (Odense, Denmark: Humana Press).

51. Quandt, A., Espona, L., Balasko, A., Weisser, H., Brusniak, M.Y., Kunszt, P., Aebersold, R., and Malmström, L. (2014). Using synthetic peptides to benchmark peptide identification software and search parameters for MS/MS data analysis. EuPA Open Proteomics *5*, 21–31.

52. Deutsch, E.W. (2010). Mass Spectrometer Output File Format mzML. Curr. Protoc. Protein Sci. *604*, 319–331.

53. Comai, L., and Katz, J.E. (2017). Data Conversion with ProteoWizard msConvert. In Proteomics Methods and Protocols, L. Comai, J. E. Katz, and P. Mallick, eds. (Los Angeles: Humana Press), p. 375.

54. The, M., Tasnim, A., and Käll, L. (2016). How to talk about protein-level false discovery rates in shotgun proteomics. Proteomics *16*, 2461–2469.

55. Elias, J.E., and Gygi, S.P. (2007). Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nat. Methods *4*, 207–214.

56. Nesvizhskii, A.I., Keller, A., Kolker, E., and Aebersold, R. (2003). A statistical model for identifying proteins by tandem mass spectrometry. Anal. Chem. *75*, 4646–4658.

57. Rappsilber, J., and Mann, M. (2002). What does it mean to identify a protein in proteomics? Trends Biochem. Sci. *27*, 74–78.

58. Walley, J.W., and Briggs, S.P. (2015). Dual use of peptide mass spectra: Protein atlas and genome annotation. Curr. Plant Biol. *2*, 21–24.

59. Nesvizhskii, A.I. (2014). Proteogenomics: Concepts, applications and computational strategies. Nat. Methods *11*, 1114–1125.

60. Wang, X., Zhang, B., and Wren, J. (2013). CustomProDB: An R package to generate customized protein databases from RNA-Seq data for proteomics search. Bioinformatics *29*, 3235–3237.

61. Coordinators, N.R. (2015). Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. *43*, D6–D17.

62. Kersey, P.J., Allen, J.E., Allot, A., Barba, M., Boddu, S., Bolt, B.J., Carvalho-Silva, D., Christensen, M., Davis, P., Grabmueller, C., *et al.* (2018). Ensembl Genomes 2018: An integrated omics infrastructure for non-vertebrate species. Nucleic Acids Res. *46*, D802–D808.

63. Wilming, L.G., Gilbert, J.G.R., Howe, K., Trevanion, S., Hubbard, T., and Harrow, J.L. (2008). The vertebrate genome annotation (Vega) database. Nucleic Acids Res. *36*, 753–760.

64. Apweiler, R. (2004). UniProt: the Universal Protein knowledgebase. Nucleic Acids Res. *32*, 115D–119.

65. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., *et al.* (2011). Gene Ontology : tool for the unification of biology. *25*, 25–29.

66. Harris, M.A., Deegan, J.I., Ireland, A., Lomax, J., Ashburner, M., Tweedie, S., Carbon, S., Lewis, S., Mungall, C., Day-Richter, J., *et al.* (2008). The Gene Ontology project in 2008. Nucleic Acids Res. *36*, 440–444.

67. Blake, J.A., Christie, K.R., Dolan, M.E., Drabkin, H.J., Hill, D.P., Ni, L., Sitnikov, D., Burgess, S., Buza, T., Gresham, C., *et al.* (2015). Gene ontology consortium: Going forward. Nucleic Acids Res. *43*, D1049–D1056.

68. Kanehisa, M. (2006). From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res. *34*, D354–D357.

69. Barrett, T., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., Holko, M., *et al.* (2013). NCBI GEO: archive for functional genomics data sets--update. Nucleic Acids Res. *41*, D991-5.

70. Sean, D., and Meltzer, P.S. (2007). GEOquery: A bridge between the Gene Expression Omnibus (GEO) and BioConductor. Bioinformatics *23*, 1846–1847.

71. Brazma, A., Parkinson, H., Sarkans, U., Shojatalab, M., Vilo, J., Abeygunawardena, N., Holloway, E., Kapushesky, M., Kemmeren, P., Lara, G.G., *et al.* (2003). ArrayExpress - A public repository for microarray gene expression data at the EBI. Nucleic Acids Res. *31*, 68–71.

72. Deutsch, E.W., Csordas, A., Sun, Z., Jarnuczak, A., Perez-Riverol, Y., Ternent, T., Campbell, D.S., Bernal-Llinares, M., Okuda, S., Kawano, S., *et al.* (2017). The ProteomeXchange consortium in 2017: Supporting the cultural change in proteomics public data deposition. Nucleic Acids Res. *45*, D1100–D1106.

73. Vizcaíno, J.A., Deutsch, E.W., Wang, R., Csordas, A., Reisinger, F., Ríos, D., Dianes, J.A., Sun, Z., Farrah, T., Bandeira, N., *et al.* (2014). ProteomeXchange provides globally coordinated proteomics data submission and dissemination. Nat. Biotechnol. *32*, 223–226.

74. Huang, D.W., Sherman, B.T., and Lempicki, R.A. (2009). Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. *37*, 1–13.

75. Nogales-Cadenas, R., Carmona-Saez, P., Vazquez, M., Vicente, C., Yang, X., Tirado, F., Carazo, J.M., and Pascual-Montano, A. (2009). GeneCodis: interpreting gene lists through enrichment analysis and

integration of diverse biological information. Nucleic Acids Res. *37*, W317-22.

76. Carmona-Saez, P., Chagoyen, M., Tirado, F., Carazo, J.M., and Pascual-Montano, A. (2007). GENECODIS: a web-based tool for finding significant concurrent annotations in gene lists. Genome Biol. *8*, R3.

77. Tabas-Madrid, D., Nogales-Cadenas, R., and Pascual-Montano, A. (2012). GeneCodis3: A non-redundant and modular enrichment analysis tool for functional genomics. Nucleic Acids Res. *40*, 478–483.

78. Dennis, G., Sherman, B.T., Hosack, D.A., Yang, J., Gao, W., Lane, H., and Lempicki, R.A. (2003). DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol. *4*, R60.

79. Thomas, P.D., Campbell, M.J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., Diemer, K., Muruganujan, A., and Narechania, A. (2003). PANTHER: A library of protein families and subfamilies indexed by function. Genome Res. *13*, 2129–2141.

80. Bauer, S., Grossmann, S., Vingron, M., and Robinson, P.N. (2008). Ontologizer 2.0 - A multifunctional tool for GO term enrichment analysis and data exploration. Bioinformatics *24*, 1650–1651.

81. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., *et al.* (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. U. S. A. *102*, 15545–50.

82. Backes, C., Keller, A., Kuentzer, J., Kneissl, B., Comtesse, N., Elnakady, Y.A., Müller, R., Meese, E., and Lenhof, H.P. (2007). GeneTrail-advanced gene set enrichment analysis. Nucleic Acids Res. *35*, 186–192.

83. Al-Shahrour, F., Arbiza, L., Dopazo, H., Huerta-Cepas, J., M??nguez, P., Montaner, D., and Dopazo, J. (2007). From genes to functional classes in the study of biological systems. BMC Bioinformatics *8*, 1–17.

84. Huang, D.W., Sherman, B.T., Stephens, R., Baseler, M.W., Lane, H.C., and Lempicki, R. a (2008). DAVID gene ID conversion tool. Bioinformation *2*, 428–30.

85. Guberman, J.M., Ai, J., Arnaiz, O., Baran, J., Blake, A., Baldock, R., Chelala, C., Croft, D., Cros, A., Cutts, R.J., *et al.* (2011). BioMart Central Portal: An open database network for the biological community. Database *2011*.

86. Calvi, L., Adams, G., Weibrecht, K., Weber, J., Olson, D., Knight, M., Martin, R., Schipani, E., Divieti, P., Bringhurst, F., *et al.* (2003). Osteoblastic cells regulate the haematopoietic stem cell niche. Nature *425*, 841–846.

87. Seita, Jun; Weissman, I.L. (2010). Hematopoietic Stem Cell: Self-renewal versus Differentiation. Wiley Interdiscip Rev Syst Biol Med *2*, 640–653.

88. Zhang, J., Niu, C., Ye, L., Huang, H., XiHe, W.-G.T., Jason Ross, J.H., Johnson, T., Feng, J.Q., Harris, S., M., W.L., *et al.* (2003). Identification of the haematopoietic stem cell niche and control of the niche size Jiwang. Nature *425*, 836–41.

89. Kiel, M.J., Yilmaz, Ö.H., Iwashita, T., Yilmaz, O.H., Terhorst, C., and Morrison, S.J. (2005). SLAM family receptors distinguish hematopoietic stem and progenitor cells and reveal endothelial niches for stem cells. Cell *121*, 1109–1121.

90. Redondo, P.A., Pavlou, M., Loizidou, M., and Cheema, U. (2017). Elements of the niche for adult stem cell expansion. J. Tissue Eng. *8*, 1–18.

91. Greaves, M. (2007). Stem and progenitor cell involvement in acute lymphoblastic leukemia Anders Castor Lund strategic center for stem cell biology and cell therapy and.

92. Maehle, A.-H. (2014). Ambiguous Cells: The Emergence of the Stem Cell Concept in the Nineteenth and Twentieth Centuries. Notes Rec R Soc L. *65*.

93. Perez Pomares, J.M. CELL GENITORS AND STEM CELLS. THE CONCEPT OF STEM CELL NICHE (Malaga).

94. Orkin, S.H., and Zon, L.I. (2008). Hematopoiesis: an evolving paradigm for stem cell biology. Cell *132*, 631–44.

95. Mikkola, H.K. a, and Orkin, S.H. (2006). The journey of developing hematopoietic stem cells. Development *133*, 3733–44.

96. Hackney, J. a, Charbord, P., Brunk, B.P., Stoeckert, C.J., Lemischka, I.R., and Moore, K. a (2002). A molecular profile of a hematopoietic stem cell niche. Proc. Natl. Acad. Sci. U. S. A. *99*, 13061–6.

97. Emambokus, N.R., and Frampton, J. (2003). The glycoprotein IIb molecule is expressed on early murine hematopoietic progenitors and regulates their numbers in sites of hematopoiesis. Immunity *19*, 33–45.

98. McKinney-Freeman, S., Cahan, P., Li, H., Lacadie, S. a, Huang, H.-T., Curran, M., Loewer, S., Naveiras, O., Kathrein, K.L., Konantz, M., *et al.* (2012). The transcriptional landscape of hematopoietic stem cell ontogeny. Cell Stem Cell *11*, 701–14.

99. Gekas, C., Dieterlen-Lièvre, F., Orkin, S.H., and Mikkola, H.K. a (2005). The placenta is a niche for hematopoietic stem cells. Dev. Cell *8*, 365–75.

100. Hidalgo, A. (2008). Hematopoietic stem cell homing: The long, winding and adhesive road to the bone marow. Inmunologia *27*, 22–35.

101. Felfly, H., and Haddad, G. (2014). Hematopoietic stem cells: potential new applications for translational medicine. J Stem Cells *9*, 163–97.

102. Liras, A. (2010). Future research and therapeutic applications of human stem cells: General, regulatory, and bioethical aspects. J. Transl. Med. *8*, 131.

103. Copelan, E. a (2006). Hematopoietic stem-cell transplantation. N. Engl. J. Med. *354*, 1813–26.

104. Copelan, E.A. (2006). Hematopoietic Stem-Cell Transplantation. N. Engl. J. Med. *354*, 1813–1826.

105. Hatzimichael, E., and Tuthill, M. (2010). Hematopoietic stem cell transplantation. Stem Cells Cloning Adv. Appl. *3*, 105–117.

106. Thorvaldsdóttir, H., Robinson, J.T., and Mesirov, J.P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief. Bioinform. *14*, 178–92.

107. Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M.T., and Carey, V.J. (2013). Software for Computing and Annotating Genomic Ranges. PLoS Comput. Biol. *9*, 1–10.

108. Ramsköld, D., Wang, E.T., Burge, C.B., and Sandberg, R. (2009). An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. PLoS Comput. Biol. *5*, e1000598.

109. Flegel, C., Manteniotis, S., Osthold, S., Hatt, H., and Gisselmann, G. (2013). Expression profile of ectopic olfactory receptors determined by deep sequencing. PLoS One *8*, e55368.

110. Harding, S.D., Sharman, J.L., Faccenda, E., Southan, C., Pawson, A.J., Ireland, S., Gray, A.J.G., Bruce, L., Alexander, S.P.H., Anderton, S., *et al.* (2017). The IUPHAR/BPS Guide to PHARMACOLOGY in 2018: updates and expansion to encompass the new guide to IMMUNOPHARMACOLOGY. Nucleic Acids Res. *46*, 1091–1106.

111. Ben-Shlomo, I., Yu Hsu, S., Rauch, R., Kowalski, H.W., and Hsueh, A.J.W. (2003). Signaling Receptome: A Genomic and Evolutionary Perspective of Plasma Membrane Receptors Involved in Signal Transduction. Sci. Signal. *2003*, RE9.

112. Chen, Y., Zhang, Y., Yin, Y., Gao, G., Li, S., Jiang, Y., Gu, X., and Luo, J. (2005). SPD--a web-based secreted protein database. Nucleic Acids Res. *33*, D169-73.

113. Kerrien, S., Aranda, B., Breuza, L., Bridge, A., Broackes-Carter, F., Chen, C., Duesbury, M., Dumousseau, M., Feuermann, M., Hinz, U., *et al.* (2012). The IntAct molecular interaction database in 2012. Nucleic Acids Res. *40*, D841-6.

114. Chatr-aryamontri, A., Ceol, A., Palazzi, L.M., Nardelli, G., Schneider, M.V., Castagnoli, L., and Cesareni, G. (2007). MINT: the Molecular INTeraction database. Nucleic Acids Res. *35*, D572-4.

115. Matys, V. (2006). TRANSFAC(R) and its module TRANSCompel(R): transcriptional gene regulation in eukaryotes. Nucleic Acids Res. *34*, D108–D110.

116. Portales-Casamar, E., Arenillas, D., Lim, J., Swanson, M.I., Jiang, S., McCallum, A., Kirov, S., and Wasserman, W.W. (2009). The PAZAR database of gene regulatory information coupled to the ORCA toolkit for the study of regulatory sequences. Nucleic Acids Res. *37*, D54-60.

117. Zheng, G., Tu, K., Yang, Q., Xiong, Y., Wei, C., Xie, L., Zhu, Y., and Li, Y. (2008). ITFP: An integrated platform of mammalian transcription factors. Bioinformatics *24*, 2416–2417.

118. Coskun, S., and Hirschi, K.K. (2010). Establishment and regulation of the HSC niche: Roles of osteoblastic and vascular compartments. Birth Defects Res. C. Embryo Today *90*, 229–42.

119. Ferraro, F., and Celso, C. Lo (2010). ADULT STEM CELLS AND THEIR NICHES. Adv Exp Med Biol *695*, 155–168.

120. Manesia, J.K., Franch, M., Tabas-Madrid, D., Nogales-Cadenas, R., Vanwelden, T., Van Den Bosch, E., Xu, Z., Pascual-Montano, A., Khurana, S., and Verfaillie, C.M. (2017). Distinct Molecular Signature of Murine Fetal Liver and Adult Hematopoietic Stem Cells Identify Novel Regulators of Hematopoietic Stem Cell Function. Stem Cells Dev. *26*, 573–584.

121. Vazquez, M., Nogales-Cadenas, R., Arroyo, J., Botías, P., García, R., Carazo, J.M., Tirado, F., Pascual-Montano, A., and Carmona-Saez, P. (2010). MARQ: an online tool to mine GEO for experiments with similar or opposite gene expression signatures. Nucleic Acids Res. *38*, W228-32.

122. Lamb, J., Crawford, E.D., Peck, D., Modell, J.W., Blat, I.C., Wrobel, M.J., Lerner, J., Brunet, J., Subramanian, A., Kenneth, N., *et al.* (2012). The Connectivity MAP: Using Gene-Expression Signatures to Connect Small Molecules, Genes, and Disease. Science *313*, 1929–1936.

123. Hong Shen Ming., Harper S Peter., U.M. (1996). Molecular genetics of neurofibromatosis type 1 (NF1). J. Med. Genet. *33*, 2–17.

124. Friedrich, R.E., Hartmann, M., and Mautner, V.F. (2007). Malignant Peripheral Nerve Sheath Tumors (MPNST) in NF1-affected children. Anticancer Res. *27*, 1957–1960.

125. Rueda-Arenas, E., Pinilla-Orejarena, A., García-Corzo, J.R., and Lozano-Ortiz, D. (2016). Tumor maligno de la vaina del nervio periférico retroperitoneal en un niño preescolar. Bol. Med. Hosp. Infant. Mex. *73*, 188–195.

126. Korf, B., Messiaen, L., Horvitz, R.A., Heiberger, Y.A., and Knight, P.B. (2016). ¿Acaba de recibir un diagnóstico de schwannomatosis? Información básica (Children's Tumor Foundation (New York)).

127. Li, Y.Y., and Jones, S.J.M. (2012). Drug repositioning for personalized medicine. Genome Med. *4*, 27.

128. Baker, N.C., Ekins, S., Williams, A.J., and Tropsha, A. (2018). A bibliometric review of drug repurposing. Drug Discov. Today *23*, 661–672.

129. Ganter, B., Tugendreich, S., Pearson, C.I., Ayanoglu, E., Baumhueter, S., Bostian, K.A., Brady, L., Browne, L.J., Calvin, J.T., Day, G.J., *et al.* (2005). Development of a large-scale chemogenomics database to improve drug candidate selection and to understand mechanisms of chemical toxicity and

action. J. Biotechnol. *119*, 219–244.

130. Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. *43*, e47.

131. Aronson, A.R., and Lang, F.-M. (2010). An overview of MetaMap: historical perspective and recent advances. J. Am. Med. Inform. Assoc. *17*, 229–36.

132. Dweep, H., Sticht, C., Pandey, P., and Gretz, N. (2011). MiRWalk - Database: Prediction of possible miRNA binding sites by " walking" the genes of three genomes. J. Biomed. Inform. *44*, 839–847.

133. Xiao, F., Zuo, Z., Cai, G., Kang, S., Gao, X., and Li, T. (2009). miRecords: an integrated resource for microRNA-target interactions. Nucleic Acids Res. *37*, D105-10.

134. Sethupathy, P., Corda, B., and Hatzigeorgiou, A.G. (2006). TarBase: A comprehensive database of experimentally supported animal microRNA targets. RNA *12*, 192–197.

135. Chou, C.-H., Shrestha, S., Yang, C.-D., Chang, N.-W., Lin, Y.-L., Liao, K.-W., Huang, W.-C., Sun, T.-H., Tu, S.-J., Lee, W.-H., *et al.* (2017). miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions. Nucleic Acids Res. *46*, 296–302.

136. Kozomara, A., and Griffiths-Jones, S. (2014). miRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic Acids Res. *42*, D68-73.

137. Sun, D., Haddad, R., Kraniak, J.M., Horne, S.D., and Tainsky, M. a (2013). RAS/MEK-independent gene expression reveals BMP2-related malignant phenotypes in the Nf1-deficient MPNST. Mol. Cancer Res. *11*, 616–27.

138. Li, H., Soriano, M., Cordewener, J., Muino, J.M., Riksen, T., Fukuoka, H., Angenent, G.C., and Boutilier, K. (2014). The Histone Deacetylase Inhibitor Trichostatin A Promotes Totipotency in the Male Gametophyte. Plant Cell *26*, 195–209.

139. Wang, C., Chen, Y., Gao, J., Lyu, G., Su, J., Zhang, Q., Ji, X., Yan, J.-Z., Qiu, Q., Zhang, Y., *et al.* (2015). Trichostatin A Is a Histone Deacetylase Inhibitor with Potent Antitumor Activity against Breast Cancer in Vivo 1. Evidence-Based Complement. Altern. Med. *344*, 2014.

140. De Raedt, T., Beert, E., Pasmant, E., Luscan, A., Brems, H., Ortonne, N., Helin, K., Hornick, J.L., Mautner, V., Kehrer-Sawatzki, H., *et al.* (2014). PRC2 loss amplifies Ras-driven transcription and confers sensitivity to BRD4-based therapies. Nature *514*, 247–251.

141. Claerhout, S., Lim, J.Y., Choi, W., Park, Y.-Y., Kim, K., Kim, S.-B., Lee, J.-S., Mills, G.B., and Cho, J.Y. (2011). Gene expression signature analysis identifies vorinostat as a candidate therapy for gastric cancer. PLoS One *6*, e24662.

142. Lee, J.-H., Kim, D.G., Bae, T.J., Rho, K., Kim, J.-T., Lee, J.-J., Jang, Y., Kim, B.C., Park, K.M., and Kim, S. (2012). CDA: combinatorial drug discovery using transcriptional response modules. PLoS One *7*, e42573.

143. Vialas, V., Sun, Z., Reales-Calderón, J.A., Hernáez, M.L., Casas, V., Carrascal, M., Abián, J., Monteoliva, L., Deutsch, E.W., L., R.M., *et al.* (2015). A comprehensive Candida albicans PeptideAtlas build enables deep proteome coverage. *344*, 1173–1178.

144. De la Calle Rodriguez N, Santa Velez C, C.C.N. (2012). Factores de virulencia para la infeccion de tejidos queratinizados por Candida albicans y hongos dermatofitos. Rev CES Med *26*, 43–55.

145. Muzzey, D., Schwartz, K., Weissman, J.S., and Sherlock, G. (2013). Assembly of a phased diploid Candida albicans genome facilitates allele-specific measurements and provides a simple model for repeat and indel structure. Genome Biol. *14*, R97.

146. Skrzypek, M.S., Binkley, J., Binkley, G., Miyasato, S.R., Simison, M., and Sherlock, G. (2017). The Candida Genome Database (CGD): Incorporation of Assembly 22, systematic identifiers and visualization of high throughput sequencing data. Nucleic Acids Res. *45*, D592–D596.

147. Ron, E., Domrachev, M., and Lash, A.E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. *30*, 207–210.

148. Tscherner, M., Zwolanek, F., Jenull, S., Sedlazeck, F.J., Petryshyn, A., Frohner, I.E., Mavrianos, J., Chauhan, N., von Haeseler, A., and Kuchler, K. (2015). The Candida albicans Histone Acetyltransferase Hat1 Regulates Stress Resistance and Virulence via Distinct Chromatin Assembly Pathways. PLoS Pathog. *11*, 1–32.

149. Lane, S., Di Lena, P., Tormanen, K., Baldi, P., and Liua, H. (2015). Function and regulation of Cph2 in Candida albicans. Eukaryot. Cell *14*, 1114–1126.

150. Cottier, F., Tan, A.S.M., Chen, J., Lum, J., Zolezzi, F., Poidinger, M., and Pavelka, N. (2015). The Transcriptional Stress Response of Candida albicans to Weak Organic Acids. Genes|Genomes|Genetics *5*, 497–505.

151. Muzzey, D., Sherlock, G., and Weissman, J.S. (2014). Extensive and coordinated control of allele-specific expression by both transcription and translation in Candida albicans. Genome Res. *24*, 963–973.

152. Hnisz, D., Bardet, A.F., Nobile, C.J., Petryshyn, A., Glaser, W., Schöck, U., Stark, A., and Kuchler, K. (2012). A Histone Deacetylase Adjusts Transcription Kinetics at Coding Sequences during Candida albicans Morphogenesis. PLoS Genet. *8*.

153. Grumaz, C., Lorenz, S., Stevens, P., Lindemann, E., Schöck, U., Retey, J., Rupp, S., and Sohn, K. (2013). Species and condition specific adaptation of the transcriptional landscapes in Candida albicans and Candida dubliniensis. BMC Genomics *14*.

154. Warren, J.J., Blanchette, D., Dawson, D. V, Teresa, A., Phipps, K.R., Starr, D., and Drake, D.R. (2017). Transcriptional landscape of trans-kingdom communication between Candida albicans and Streptococcus gordonii. Mol Oral Microbiol *31*, 136–161.

155. Wartenberg, A., Linde, J., Martin, R., Schreiner, M., Horn, F., Jacobsen, I.D., Jenull, S., Wolf, T., Kuchler, K., Guthke, R., *et al.* (2014). Microevolution of Candida albicans in Macrophages Restores Filamentation in a Nonfilamentous Mutant. PLoS Genet. *10*.

156. Liu, Y., Shetty, A.C., Schwartz, J.A., Bradford, L.L., Xu, W., Phan, Q.T., Kumari, P., Mahurkar, A., Mitchell, A.P., Ravel, J., *et al.* (2015). New signaling pathways govern the host response to C. albicans infection in various niches. Genome Res. *125*, 679–689.

157. Wells, M.L., Washington, O.L., Hicks, S.N., Nobile, C.J., Hartooni, N., Wilson, G.M., Zucconi, B.E., Huang, W., Li, L., Fargo, D.C., *et al.* (2016). Post-transcriptional regulation of transcript abundance by a conserved member of the tristetraprolin family in Candida albicans. Mol Microbiol. *95*, 1036–1053.

158. Kodama, Y., Shumway, M., and Leinonen, R. (2012). The sequence read archive: Explosive growth of sequencing data. Nucleic Acids Res. *40*, 2011–2013.

159. Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., *et al.* (2011). The variant call format and VCFtools. Bioinformatics *27*, 2156–2158.

160. Tabas-Madrid, D., Alves-Cruzeiro, J., Segura, V., Guruceaga, E., Vialas, V., Prieto, G., García, C., Corrales, F.J., Albar, J.P., and Pascual-Montano, A. (2015). Proteogenomics dashboard for the human proteome project. J. Proteome Res. *14*, 3738–3749.

161. Sheynkman, G.M., Shortreed, M.R., Frey, B.L., and Smith, L.M. (2013). Discovery and Mass Spectrometric Analysis of Novel Splice-junction Peptides Using RNA-Seq. Mol. Cell. Proteomics *12*,

2341–2353.

162. Wang, X., and Zhang, B. (2014). Integrating Genomic, Transcriptomic, and Interactome Data to Improve Peptide and Protein Identi fi cation in Shotgun Proteomics. J. Proteome Res. *13*, 2715–2723.

163. Bunger, M.K., Cargile, B.J., Sevinsky, J.R., Deyanova, E., Yates, N.A., Hendrickson, R.C., and Stephenson, J.L. (2007). Detection and validation of non-synonymous coding SNPs from orthogonal analysis of shotgun proteomics data. J. Proteome Res. *6*, 2331–2340.

164. Li, J., Su, Z., Ma, Z.-Q., Slebos, R.J.C., Halvey, P., Tabb, D.L., Liebler, D.C., Pao, W., and Zhang, B. (2011). A Bioinformatics Workflow for Variant Peptide Detection in Shotgun Proteomics. Mol. Cell. Proteomics *10*, M110.006536.

165. McAfee, A., and Foster, L.J. (2017). Proteogenomics: Recycling Public Data to Improve Genome Annotations 1st ed. (Vancouver: Elsevier Inc.) Available at: http://dx.doi.org/10.1016/bs.mie.2016.09.020.

166. Rau, A., Marot, G., and Jaffrézic, F. (2014). Differential meta-analysis of RNA-seq data from multiple studies. BMC Bioinformatics *15*, 91.

167. Desiere, F. (2006). The PeptideAtlas project. Nucleic Acids Res. *34*, D655–D658.

168. Gil-Bona, Ana, Marcela Parra-Giraldo, C., Hernáez, M.L., Reales-Calderón, J.A., Solis, N. V, Filler, S.G., Monteoliva, L., and Gil, C. (2016). Candida albicans cell shaving uncovers new proteins involved in cell wall integrity, yeast to hypha transition, stress response and host-pathogen interaction. Proteomics, J *127*, 340–351.

169. Gil-Bona, A., Llama-Palacios, A., Parra, C.M., Vivanco, F., Nombela, C., Monteoliva, L., and Gil, C. (2015). Proteomics unravels extracellular vesicles as carriers of classical cytoplasmic proteins in Candida albicans. J. Proteome Res. *14*, 142–153.

170. Monteoliva, L., Martinez-Lopez, R., Pitarch, A., Hernaez, M.L., Serna, A., Nombela, C., Albar, J.P., and Gil, C. (2011). Quantitative proteome and acidic subproteome profiling of Candida albicans yeast-to-hypha transition. J. Proteome Res. *10*, 502–517.

171. Muth, T., Vaudel, M., Barsnes, H., Martens, L., and Sickmann, A. (2010). XTandem Parser: An open-source library to parse and analyse X!Tandem MS/MS search results. Proteomics *10*, 1522–1524.

172. Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. B *57*, 289–300.

173. Zhang, J., Xin, L., Shan, B., Chen, W., Xie, M., Yuen, D., Zhang, W., Zhang, Z., Lajoie, G.A., and Ma, B. (2012). PEAKS DB: De Novo Sequencing Assisted Database Search for Sensitive and Accurate Peptide Identification. Mol. Cell. Proteomics *11*, M111.010587.

174. Johnson, M., Zaretskaya, I., Raytselis, Y., Merezhuk, Y., McGinnis, S., and Madden, T.L. (2008). NCBI BLAST: a better web interface. Nucleic Acids Res. *36*, 5–9.

175. Wen, B., Xu, S., Zhou, R., Zhang, B., Wang, X., Liu, X., Xu, X., and Liu, S. (2016). PGA: an R/Bioconductor package for identification of novel peptides using a customized database derived from RNA-Seq. BMC Bioinformatics *17*, 244.

176. Ruggles, K. V., Tang, Z., Wang, X., Grover, H., Askenazi, M., Teubl, J., Cao, S., McLellan, M.D., Clauser, K.R., Tabb, D.L., *et al.* (2016). An Analysis of the Sensitivity of Proteogenomic Mapping of Somatic Mutations and Novel Splicing Events in Cancer. Mol. Cell. Proteomics *15*, 1060–1071.

177. Nielsen, R., Paul, J.S., Albrechtsen, A., and Song, Y.S. (2011). Genotype and SNP calling from next-generation sequencing data. Nat. Rev. Genet. *12*, 443–451.

178. Miranda, I., Silva-dias, A., Rocha, R., Teixeira-Santos, R., Coelho, C., Gonçalves, T., Santos, M.A.S., Pina-

Vaz, C., Solis, Norma V, Filler, S.G., *et al.* (2013). Candida albicans CUG Mistranslation Is a Mechanism To Create Cell Surface Variation. MBio *4*, 1–9.

179. Setoain, J., Franch, M., Martínez, M., Tabas-Madrid, D., Sorzano, C.O.S., Bakker, A., Gonzalez-Couto, E., Elvira, J., and Pascual-Montano, A. (2015). NFFinder: An online bioinformatics tool for searching similar transcriptomics experiments in the context of drug repositioning. Nucleic Acids Res. *43*, W193–W199.

180. Hrdlickova, R., Toloue, M., & Tian, B. (2017). RNA-Seq methods for transcriptome analysis. Wiley Interdisciplinary Reviews. RNA, 8(1).

181. Han, Yixing et al. "Advanced Applications of RNA Sequencing and Challenges." Bioinformatics and Biology Insights 9.Suppl 1 (2015): 29–46.

182. Costa-Silva, Juliana, Douglas Domingues, and Fabricio Martins Lopes. "RNA-Seq Differential Expression Analysis: An Extended Review and a Software Tool." Ed. Zhi Wei. PLoS ONE 12.12 (2017)

183. Conesa, Ana et al. "A Survey of Best Practices for RNA-Seq Data Analysis." Genome Biology 17 (2016): 13.

184. Khang, Tsung Fei, and Ching Yee Lau. "Getting the Most out of RNA-Seq Data Analysis." Ed. Jaume Bacardit. PeerJ 3 (2015): e1360.

185. Piwowar, Heather A., and Todd J. Vision. "Data Reuse and the Open Data Citation Advantage." Ed. Xiaolei Huang. PeerJ 1 (2013): e175. PMC. Web. 10 July 2018.

186. Rung, Johan and Brazma Alvis. "Reuse of public genome-wide gene expression data." Nature Reviews Genetics (2012):14.

187. Bruno, Vincent M. et al. "Comprehensive Annotation of the Transcriptome of the Human Fungal Pathogen Candida Albicans Using RNA-Seq." Genome Research 20.10 (2010): 1451–1458.

188. Schurch, N. J., Schofield, P., Gierliński, M., Cole, C., Sherstnev, A., Singh, V., … Barton, G. J. (2016). How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? *Rna*, *22*(10), 1641.

ABREVIATURAS

ADN: Ácido desoxi-ribonucleico

ARN: Ácido ribonucleico

ARNm: Ácido ribonucleico mensajero

miARN: micro Ácido ribonucleico

PCR: Reacción en cadena de polimerasa

ADNc: DNA complementario

CDS: secuencia codificante

EST: Marcador de secuencia expresada

GO: Gene Ontology

GSEA: Enriquecimiento de conjuntos de genes

HSC: Células madre hematopoyéticas

FL: Hígado fetal

BM: Médula ósea

RPKM: Reads per KB per million reads

PCA: Análisis de components principals

FDR: Tasa de falsos positivos

FNR: Tasa de Falsos Negativos

NF: Neurofibromatosis

INDELS: Inserciones o deleción

SAAVs: Variaciones de amino ácido único

SNP: Polimorfismo de nucleótido único

SNV: Variaciones de nucleótido único

TTP: Trans-Proteomic pipline

ORFs: marco de lectura abierto

m/z: masa/carga

PSM: Asignación Péptido-Espectro

RESUMEN

El término ómico se refiere al estudio global de los sistemas celulares en un nivel concreto. Las principales ciencias ómicas desarrolladas en los últimos años son la Genómica, la Transcriptómica, la Proteómica y la Metabolómica. Estas disciplinas se basan en el análisis de un gran volumen de datos, y para ello se valen de la Bioinformática y de técnicas rápidas y automatizadas de alto rendimiento.

La Genómica es el campo de la Genética que intenta comprender el contenido, la organización, la función y la evolución de la información molecular del ADN albergada en el genoma completo. Conocer dicha secuencia, nos permite poder identificar los genes contenidos y estudiar las funciones de los mismos de forma detallada. Gracias a los avances en genómica apareció la transcriptómica que es el campo en el que se estudia y se comparan los conjuntos de ARN mensajeros o tránscritos presentes en una célula, tejido u organismo. Varias tecnologías se han desarrollado para deducir y cuantificar el transcriptoma, actualmente la tecnología más utilizada es la secuenciación masiva de ARN (*RNA-seq*). En la que se parte de una población de *ARNm* y se obtienen las lecturas de las secuencias. El análisis computacional parte de estas lecturas y consiste en un control de calidad y procesamiento de las lecturas, el alineamiento de éstas al genoma de referencia, la cuantificación de las lecturas alineadas en los genes y la normalización de los datos e identificación de los genes mediante expresión diferencial.

La proteómica en cambio se define como el conjunto de técnicas o tecnologías utilizadas para la obtención de información funcional de las proteínas, y tiene por objetivo el análisis, identificación y caracterización del proteoma celular. En la proteómica *shotgun,* el primer paso del experimento consiste en la digestión de las proteínas de la muestra mediante una enzima, generalmente tripsina. Los péptidos obtenidos son separados por cromatografía líquida e ionizados para entrar en el espectro de masas dónde son separados en función de la proporción entre su masa y su carga (*m/z*). En la espectrometría de masas en tándem (MS/MS), los péptidos con mayor intensidad son seleccionados para ser fragmentados generando espectros MS/MS, colecciones de valores *m/z* y de intensidad para cada precursor y sus fragmentos. Obtenidos estos datos, comienza el análisis computacional. Mediante los motores de búsqueda, los espectros adquiridos son enfrentados contra una base de datos de secuencias de proteínas. A cada par espectro-péptido se les asigna una puntuación que mide el grado de similitud. Estas puntuaciones son evaluadas con test estadísticos como *el Test de Falsos Positivos, FDR* mediante la utilización de secuencias generadas artificialmente llamadas señuelo.

Combinando estas omicas se encuentra la Proteogenómica que tiene como objetivo identificar y descubrir nuevos péptidos que, con las bases de datos utilizadas habitualmente por los motores de búsqueda, no se identificarían. Esta aproximación consiste en la generación de bases de datos de secuencias de proteínas o péptidos con secuencias de interés obtenidas de análisis transcriptómicos.

Estas técnicas generan un gran crecimiento de la cantidad de datos biológicos de los que se puede extraer mucha información. Un aspecto importante de estas técnicas ómicas es la diseminación de los resultados experimentales. Actualmente existen repositorios públicos *online* que albergan los datos generados por estas tecnologías que permiten compartir, utilizar y reanalizar estos experimentos generados por la comunidad científica. En el campo de la genómica se encuentra *GEO* y *ArrayExpress* que contienen datos de experimentos de *microarrays*, de secuenciación masiva y otras formas de experimentos de alto rendimiento. En proteómica los repositorios *PRIDE* y *PeptideAtlas* tienen un papel importante.

Objetivos

El objetivo general de esta Tesis doctoral es el desarrollo y aplicación de métodos bioinformáticos para el análisis de datos biológicos procedentes de diversas plataformas, así como su integración y aplicación para obtener una visión global de los genes, proteínas y procesos biológicos alterados. Estos métodos se aplicaron tanto a datos procedentes de diferentes laboratorios para responder a preguntas científicas concretas como a datos existentes en repositorios públicos para crear herramientas tales como las destinadas al reposicionamiento de fármacos.

Resultados

En la primera aportación de esta Tesis se ha desarrollado una metodología bioinformática para el análisis de datos transcriptómicos e integración de datos biológicos. Primero se ha establecido un flujo de análisis de experimentos de *RNA-seq* que incluye alineamiento, cuantificación, normalización y filtrado de expresión por umbral de activación. A continuación, se integraron los resultados obtenidos con los datos biológicos procedentes de diferentes bases de datos, para así evaluar y estudiar el interactoma obteniendo una visión global y específica de su funcionalidad. Mediante la utilización de esta metodología se obtuvo una lista de rutas metabólicas y genes de interés para su posterior validación experimental.

En la segunda aportación de la Tesis se ha desarrollado una herramienta web para el reposicionamiento de drogas que permite comparar perfiles de expresión génica para encontrar condiciones que producen fenotipos similares o antagónicos susceptibles de estar relacionados con el proceso buscado. Para ello se han procesado y analizado los experimentos transcriptómicos de las bases de datos *GEO*, *Connectivity Map* y *Drug Matrix* para generar perfiles de expresión y asociarlos a fármacos, compuestos, enfermedades y/o expertos.

En la tercera y última aportación se ha realizado un estudio proteogenómico para la identificación de SNV, INDELS y nuevas uniones en *C. albicans*. Con este objetivo, se ha desarrollado una herramienta para generar bases de datos con información procedente de experimentos de *RNA-seq* para realizar las búsquedas por espectrometría de masas y así identificar estas variaciones. El resultado de la búsqueda ha consistido en 13 nuevos péptidos correspondientes a 11 proteínas diferentes que han sido validados experimentalmente.

Conclusiones

En este trabajo se han abordado diferentes aspectos de la bioinformática funcional. Para ello se han desarrollado una serie de metodologías y herramientas que constituyen un marco de trabajo para la investigación con datos ómicos permitiendo obtener información de interés a partir de datos biológicos. Si bien es cierto que estos trabajos representan problemas distintos y en contextos experimentales diferentes, es también importante destacar que la solución a los mismos tiene un hilo conductor común, que son las técnicas experimentales, las herramientas bioinformáticas y las metodologías de integración. Una de la enseñanza de este trabajo es que no existe una única aproximación a todos los problemas que encontramos en cualquier laboratorio de biología molecular y es necesario desarrollar y aplicar las metodologías existentes a los problemas específicos, además de generar metodologías nuevas que le integren.

SUMMARY

Introduction

The terms *omics* refers to the collective characterization and quantification studies perform on a specific level of the cellular system. In the last few years, the main omics sciences developed are Genomics, Transcriptomics, Proteomics and Metabolomics. All these branches are based on the analysis of large volumes of data that require the help of bioinformatics tools and fast and automated high performance techniques.

Genomics is a part of the genetic field that analyses content, organization, function and evolution contained in the entire genome. Indeed, the knowledge of DNA sequence is crucial to identify and codify the function of the genes. The successful and the advent of new technologies in the genomic field contribute to the evolution of Transciptomics focused on the classification and quantification of set of transcripts in a cell at a specific developmental stage or physiological condition.

A deep understanding of the trascriptome is essential to interpret the functional elements of the genome and to reveal the molecular constituents of cells and tissues in order to shed light on cellular development and associated diseases. Although several technologies have been developed to investigate and quantify the transcriptome, currently the most widely used technology is *RNA-seq*. A typical *RNA-seq* experiment enable the determination of the fragment sequence of a population of RNA. These results are further processed in the data analysis steps: (1) quality check and preprocessing of raw sequence reads, (2) mapping reads to a reference genome or transcriptome, (3) counting reads mapped to individual genes or transcripts, (4) normalization and identification of differential expression.

Proteomics is a set of techniques or technologies that aims to obtain functional information of proteins in order to characterize the cellular proteome. Specifically, *Shotgun* proteomic experiments usually consists of a digestion of sample proteins into different peptides, often using a proteolytic enzyme such as *trypsin*. These peptides are then separated by liquid chromatography before entering into the spectrometer, to be selected one at a time using the first stage of mass analysis. Subsequently, each isolated peptide is fragmented, possibly by collision, and the second stage of mass analysis is used to capture an MS/MS spectrum. Finally, each MS/MS spectrum is analysed by a software that determines which peptide sequence gives the best match in a database of proteins. To assess confidence in the peptide identification, the scores are evaluated using False Discovery Rate (FDR) by using a decoy sequences.

The field that combines genomics and proteomics is called Proteogenomics. It aims to identify and discover new peptides that would not be identified with a reference database leading to generate new database that contains proteins or peptides with sequence of interest that are obtained from transcriptomic data.

Objectives

The main objective of the thesis was directed to develop and apply bioinformatics methods in the analysis of biological data from different platforms and to integrated them for a thorough understanding of genes, proteins and altered biological processes. To validate this approach, these methods were applied to data from public repositories and to experimental data from different laboratories to answer to specific scientific questions.

Results

In the first contribution of this thesis we developed a bioinformatics methodology for the analysis of transcriptomic data and its integration of biological data. An automatic flow for processing *RNA-seq* data has been established. This analysis includes alignment, quantification, normalization, and expression filtered by a threshold. Then, the results obtained were combined with biological data taken from different databases in order to evaluate and study the interactome, in a global and local sight. The use of this methodology results in a list of metabolic pathways and genes of interest.

In the second contribution we developed a web tool for drug repurposing named NFFinder. This tool allows to compare gene expression profiles to find conditions that produce similar or antagonistic phenotypes that can revert or be related with the studied disease. To create Nffinder, we processed and analysed transcriptomics profiles from GEO, DrugMatrix and CMap generating expression profiles and associating them with drugs compounds, diseases and experts.

In the third and last contribution, we carried on a proteogenomics study for the identification of SNV, INDELS and Novel Junctions in C. albicans. Thus, we developed a tool to generate a database with information from transcriptomic experiments. This database was used in mass spectrometry searches identifying 13 new peptides corresponding to 11 different proteins.

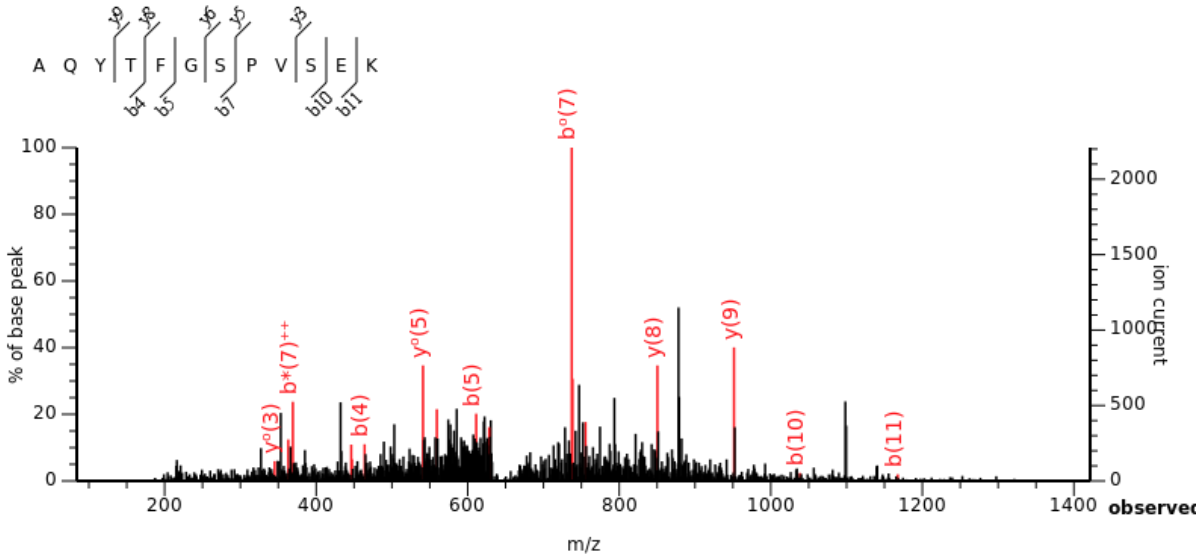
Conclusions

In this work different aspects of functional bioinformatics have been addressed. We developed different methodologies and tools for *omics* data to obtain information of interest from biological data. This work faces different scientific problems in different experimental contexts, displaying that the only possible solution has a common line: the experimental techniques, the bioinformatics tools and integration methodologies.

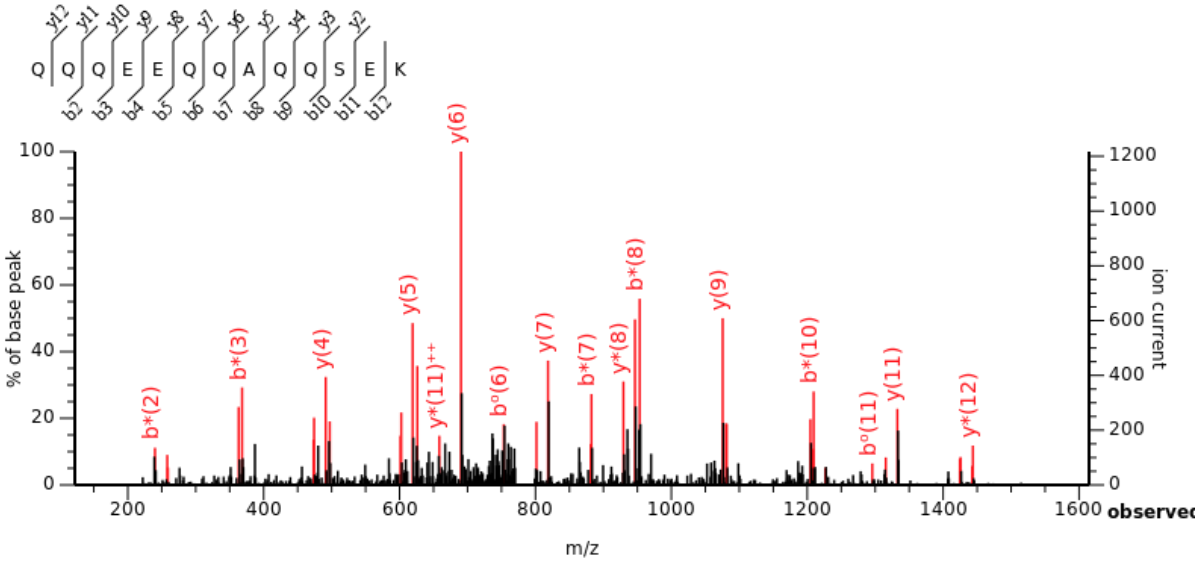
ANEXOS

Anexo I: Espectros MS2 de los péptidos candidatos obtenidos con MASCOT. En este anexo se encuentran los espectros obtenidos de aquellos péptidos que no han sido sintetizados.

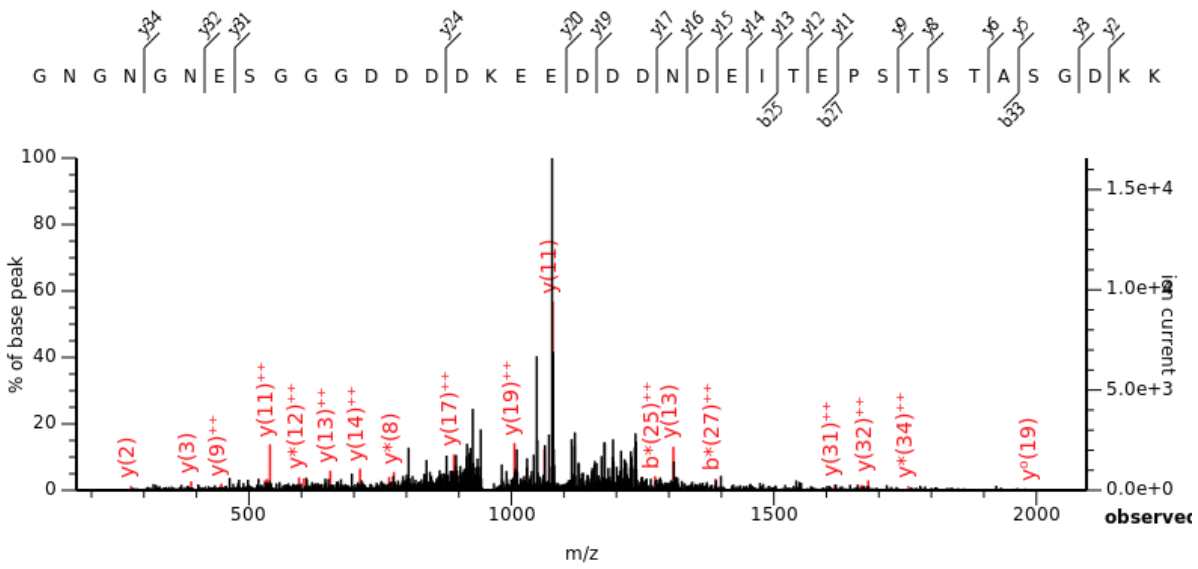
C4_03190W_B AQYTFGSPVSEK



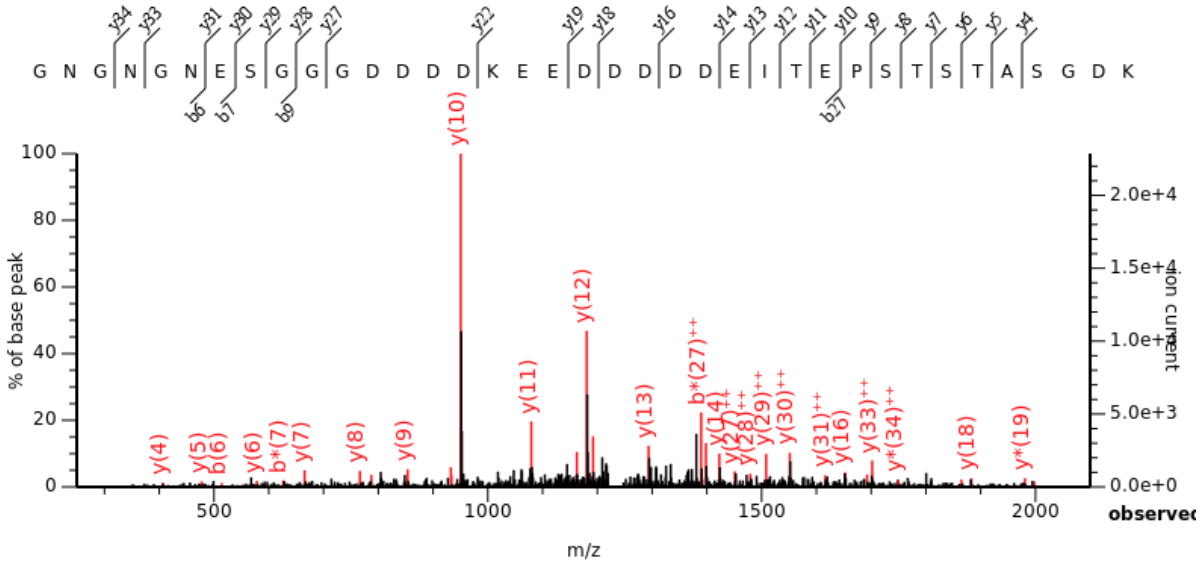
TBF1 QQQEEQQAQQSEK



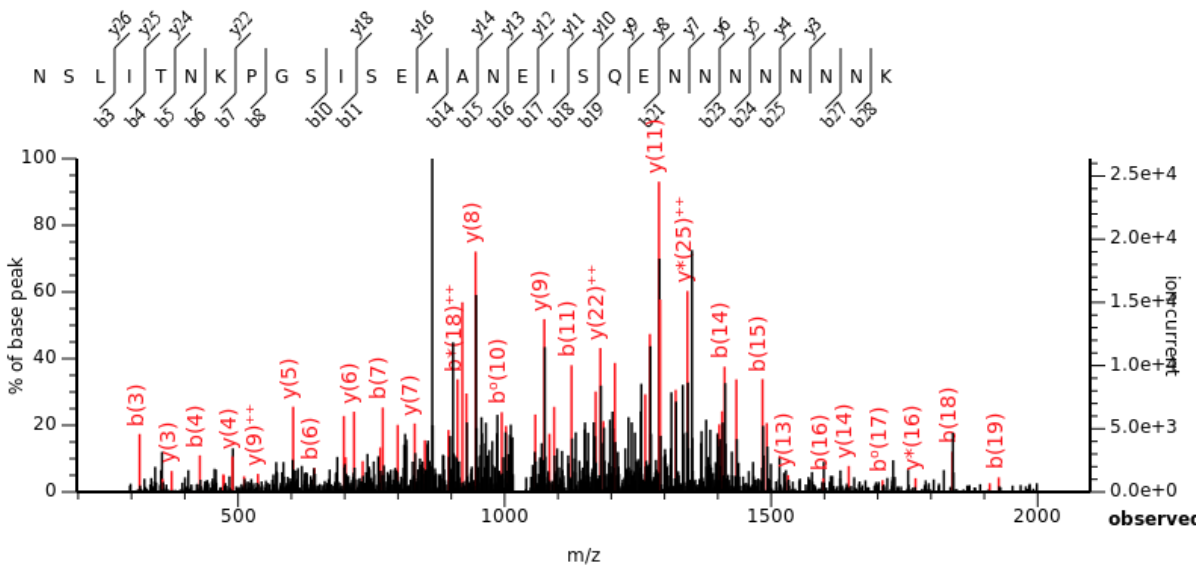
MAP2 GNGNGNESGGGDDDDKEEDDDDDEITEPSTSTASGDKK



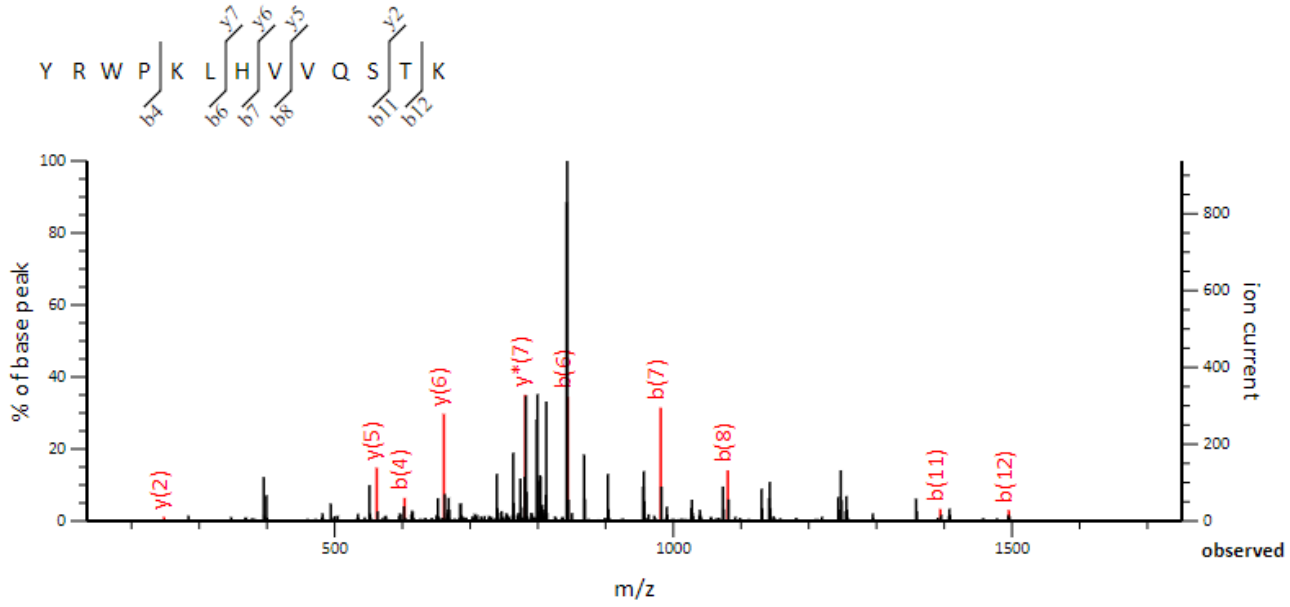
MAP2 GNGNGNESGGGDDDDKEEDDDDDEITEPSTSTASGDK



NUP60 NSLITNKPGSISEAANEISQENNNNNNNK



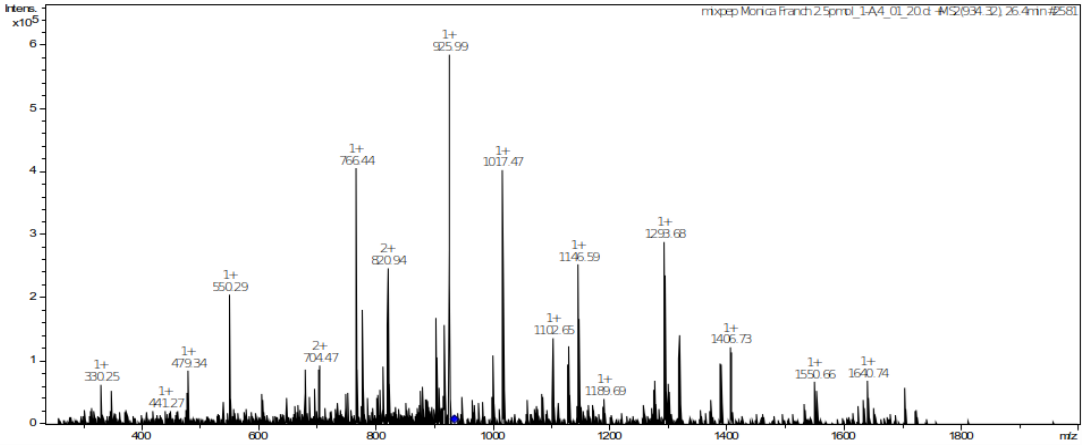
C6_00020W_B YRWPKLHVVQSTK



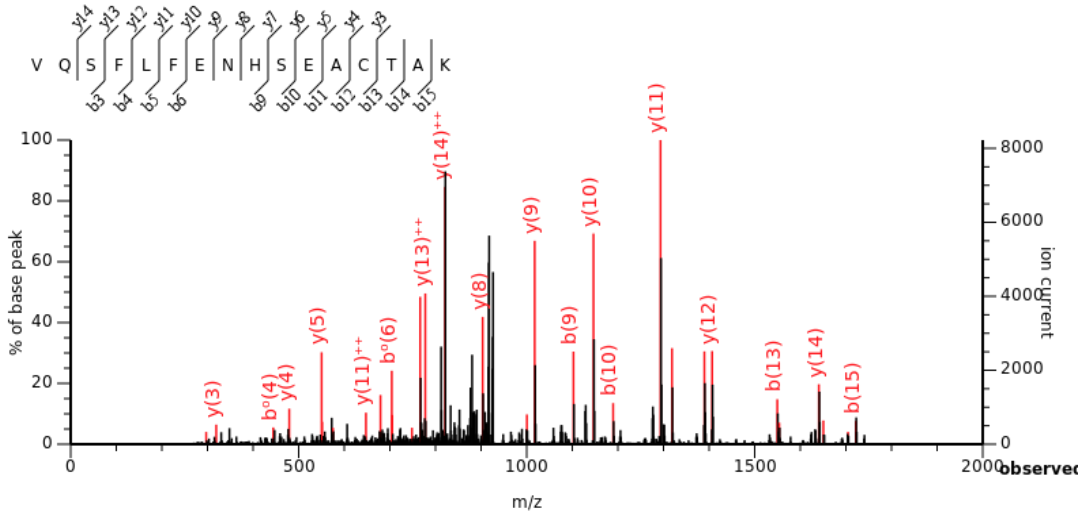
Anexo II: Espectros MS2 de los péptidos sintéticos, de los péptidos candidatos y tabla de anotación de iones. Cada espectro MS2 candidato se compara con el espectro de fragmentación del péptido sintético correspondiente. Se incluye la tabla de anotación de iones para las comparaciones.

GLK1 VQSFLFENHSEACTAK

Espectro Péptido Candidato



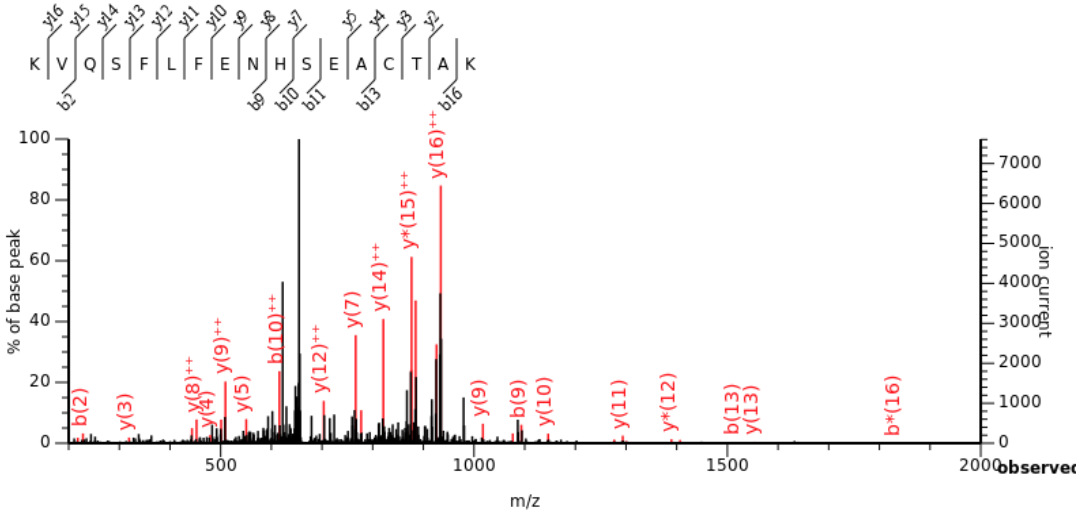
Espectro Péptido Sintético



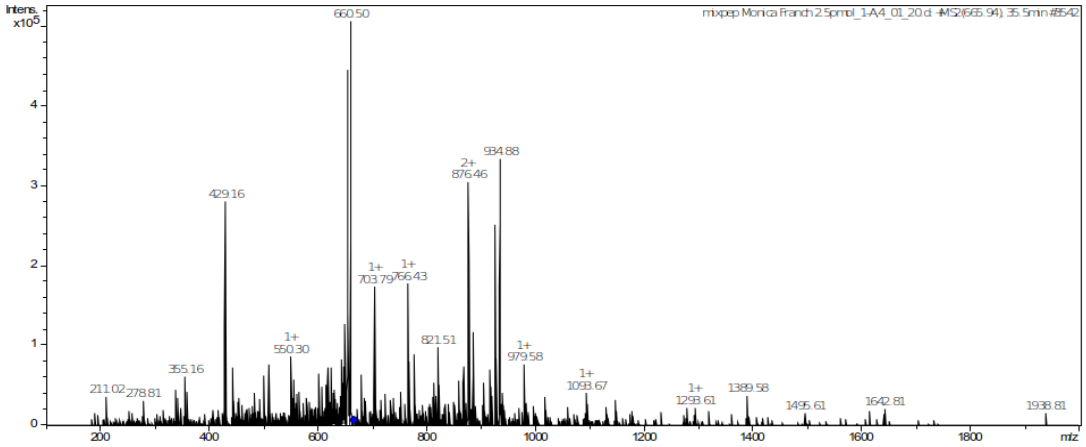
b	b+2				y	y+2
---	---	1	V	16	---	---
228.1343	---	2	Q	15	1768.7964	884.9019
315.1663	---	3	S	14	1640.7379	820.8726
462.2347	---	4	F	13	1553.7058	777.3566
575.3188	---	5	L	12	1406.6374	703.8224
722.3872	---	6	F	11	1293.5534	647.2803
851.4298	---	7	E	10	1146.4849	573.7461
965.4727	---	8	N	9	1017.4424	509.2248
1102.5316	551.7694	9	H	8	903.3994	452.2034
1189.5636	595.2855	10	S	7	766.3405	383.6739
1318.6062	659.8068	11	E	6	679.3085	340.1579
1389.6434	695.3253	12	A	5	550.2659	275.6366
1549.6745	775.3409	13	C(57.022)	4	479.2288	240.1180
1650.7222	825.8647	14	T	3	319.1976	160.1024
1721.7593	861.3833	15	A	2	218.1499	109.5786
---	---	16	K	1	147.1128	74.0600

GLK1 KVQSFLFENHSEACTAK

Espectro Péptido Candidato



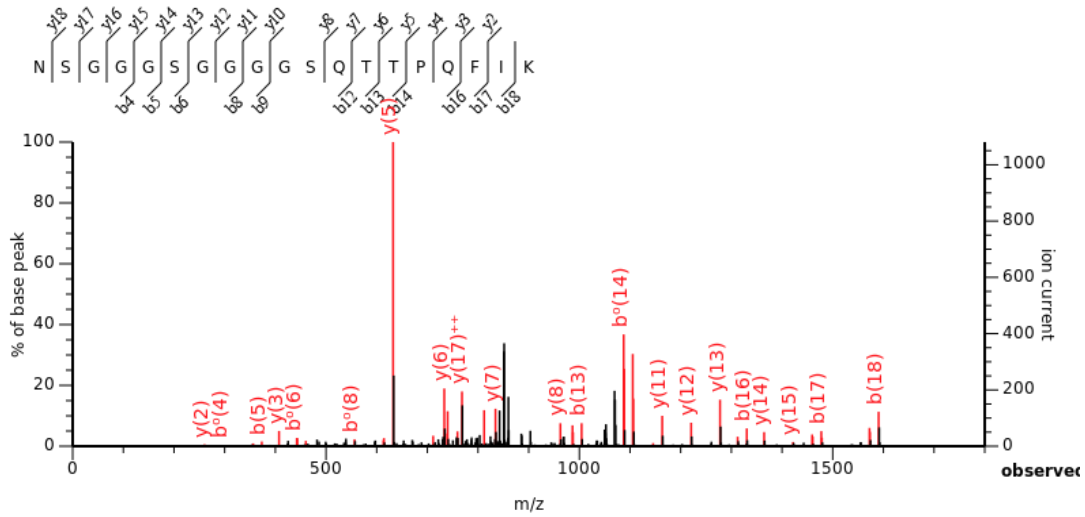
Espectro Péptido Sintético



b	b^{+2}				y	y^{+2}
---	---	1	K	17	---	---
228.1707	114.5890	2	V	16	1867.8649	934.4361
356.2292	178.6183	3	Q	15	1768.7964	884.9019
443.2613	222.1343	4	S	14	1640.7379	820.8726
590.3297	295.6685	5	F	13	1553.7058	777.3566
703.4137	352.2105	6	L	12	1406.6374	703.8224
850.4822	425.7447	7	F	11	1293.5534	647.2803
979.5247	490.2660	8	E	10	1146.4849	573.7461
1093.5677	547.2875	9	N	9	1017.4424	509.2248
1230.6266	615.8169	10	H	8	903.3994	452.2034
1317.6586	659.3329	11	S	7	766.3405	383.6739
1446.7012	723.8542	12	E	6	679.3085	340.1579
1517.7383	759.3728	13	A	5	550.2659	275.6366
1677.7695	839.3884	14	C(57.022)	4	479.2288	240.1180
1778.8172	889.9122	15	T	3	319.1976	160.1024
1849.8543	925.4308	16	A	2	218.1499	109.5786
---	---	17	K	1	147.1128	74.0600

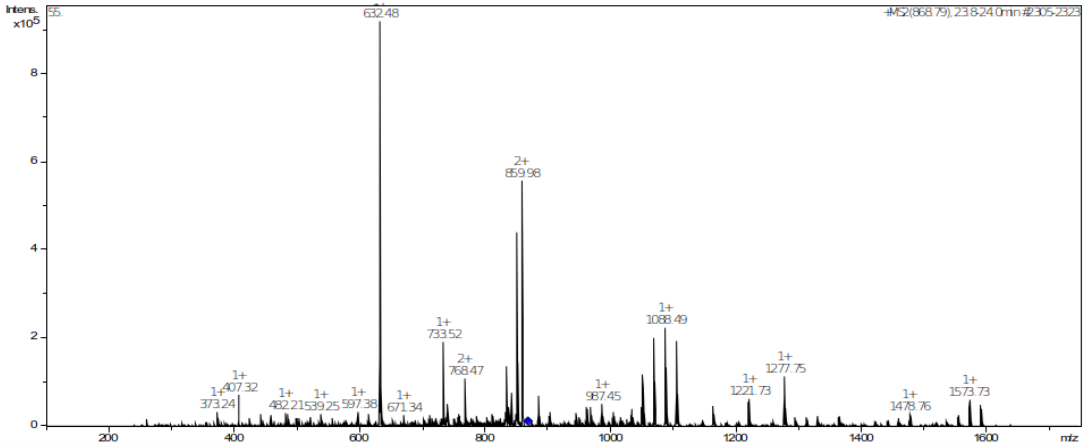
RBK1 NSGGGSGGGGSQTTPQFIK

Espectro Péptido Candidato



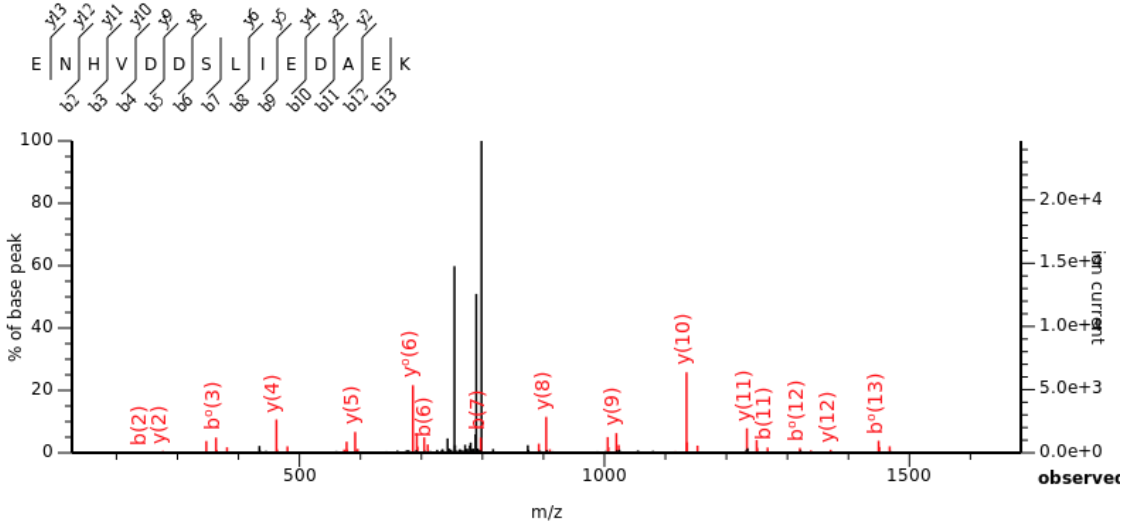
b			y	y^{+2}
---	1	N 19	---	---
202.0822	2	S 18	1622.7769	811.8921
259.1037	3	G 17	1535.7449	768.3761
316.1252	4	G 16	1478.7234	739.8653
373.1466	5	G 15	1421.7019	711.3546
460.1787	6	S 14	1364.6805	682.8439
517.2001	7	G 13	1277.6484	639.3279
574.2216	8	G 12	1220.6270	610.8171
631.2430	9	G 11	1163.6055	582.3064
688.2645	10	G 10	1106.5840	553.7957
775.2965	11	S 9	1049.5626	525.2849
903.3551	12	Q 8	962.5306	481.7689
1004.4028	13	T 7	834.4720	417.7396
1105.4505	14	T 6	733.4243	367.2158
1202.5032	15	P 5	632.3766	316.6919
1330.5618	16	Q 4	535.3239	268.1656
1477.6302	17	F 3	407.2653	204.1363
1590.7143	18	I 2	260.1969	130.6021
---	19	K 1	147.1128	74.0600

Espectro Péptido Sintético



C1_11200W ENHVDDSLIEDAEK

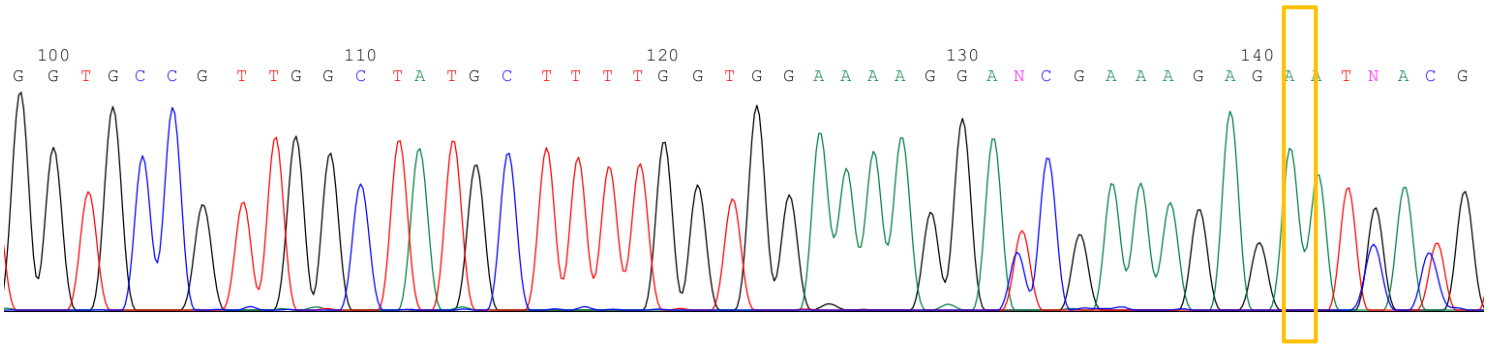
Espectro Péptido Candidato



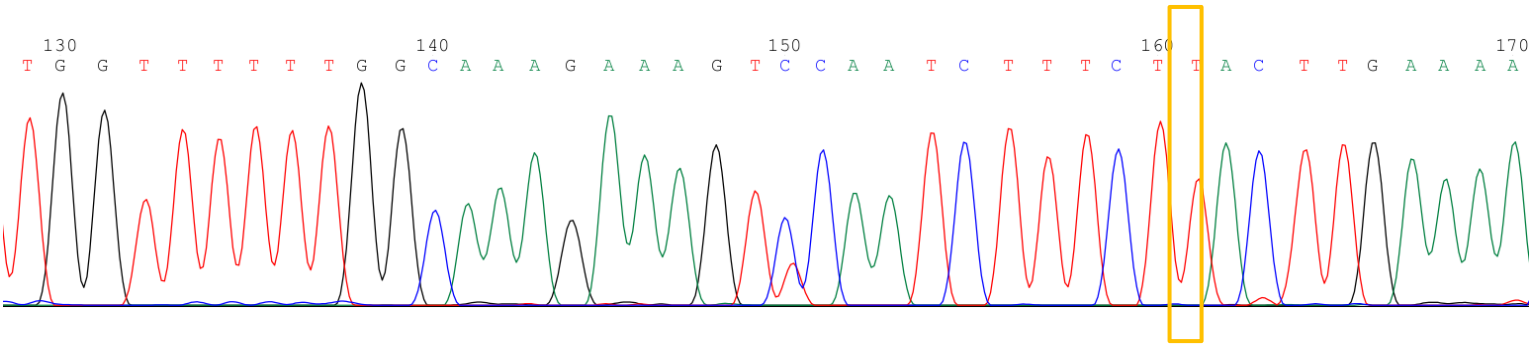
b	b+2			y	y+2
---	---	1	E 14	---	---
244.0928	---	2	N 13	1484.6863	742.8468
381.1517	191.0795	3	H 12	1370.6434	685.8253
480.2201	240.6137	4	V 11	1233.5845	617.2959
595.2471	298.1272	5	D 10	1134.5161	567.7617
710.2740	355.6406	6	D 9	1019.4891	510.2482
797.3060	399.1567	7	S 8	904.4622	452.7347
910.3901	455.6987	8	L 7	817.4302	409.2187
1023.4742	512.2407	9	I 6	704.3461	352.6767
1152.5168	576.7620	10	E 5	591.2620	296.1347
1267.5437	634.2755	11	D 4	462.2195	231.6134
1338.5808	669.7940	12	A 3	347.1925	174.0999
1467.6234	734.3153	13	E 2	276.1554	138.5813
---	---	14	K 1	147.1128	74.0600

Anexo III: Resultados de la secuenciación Sanger. En amarillo está indicada la posición de la mutación. (Referencia → Mutación)

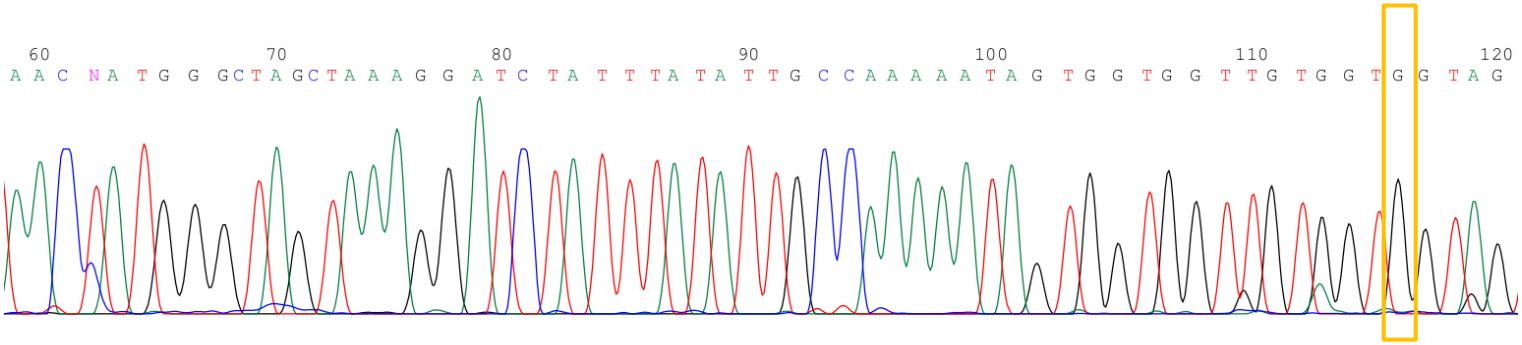
C1_11200W_A G → C



GLK1 C → T



RBK1 G → A



Anexo IV: Publicaciones relacionadas con la tesis doctoral

Manesia, J.K., Franch, M., Tabas-Madrid, D., Nogales-Cadenas, R., Vanwelden, T., Van Den Bosch, E., Xu, Z., Pascual-Montano, A., Khurana, S., and Verfaillie, C.M. (2017). Distinct Molecular Signature of Murine Fetal Liver and Adult Hematopoietic Stem Cells Identify Novel Regulators of Hematopoietic Stem Cell Function. Stem Cells Dev. *26*, 573–584.

Setoain, J., Franch, M., Martínez, M., Tabas-Madrid, D., Sorzano, C.O.S., Bakker, A., Gonzalez-Couto, E., Elvira, J., and Pascual-Montano, A. (2015). NFFinder: An online bioinformatics tool for searching similar transcriptomics experiments in the context of drug repositioning. Nucleic Acids Res. *43*, W193–W199.

Franch, M., Monteoliva, L., Amador-Garcia, A., Ramos-Fernández A., Corrales FJ.,Pascual-Montano A., Gil C. Proteogenomics: Combining Next Generation Sequencing and Mass Spectrometry in *C. albicans* . (En Revisión)