Métodos de enriquecimiento funcional para la evaluación de la significación biológica en análisis bioinformáticos

Grado en Estadística - Facultad de Ciencias

Universidad de Valladolid

Autor: Jorge García
Tutores: Itziar Fernández Martínez
Yolanda Larriba González

Curso 2023/2024

Resumen

El análisis de enriquecimiento funcional es una técnica fundamental en las ciencias ómicas que permite la interpretación biológica de datos de alto rendimiento. Este análisis se utiliza para identificar conjuntos de genes, proteínas u otras entidades biológicas que están significativamente representados en un experimento, proporcionando así una comprensión más profunda de los procesos biológicos subyacentes. En este trabajo se exploran las bases de datos más comúnmente utilizadas para el análisis de enriquecimiento funcional, así como los métodos estadísticos empleados para evaluar la significación de los resultados obtenidos. Además, se discute la visualización de estos resultados y se presentará una comparación práctica entre dos métodos populares de enriquecimiento funcional: el análisis de sobre-representación (SEA) y el análisis de enriquecimiento de conjuntos génicos (GSEA). Esta comparación se realizará utilizando herramientas automatizadas para evaluar la eficacia y precisión de los métodos en diferentes escenarios experimentales.

Palabras clave: Análisis de Enriquecimiento Funcional, Ciencias Ómicas Bases de Datos Biológicas, SEA (Análisis de Sobre-Representación), GSEA (Gene Set Enrichment Analysis)

Abstract

Functional enrichment analysis is a fundamental technique in omics sciences that enables the biological interpretation of high-throughput data. This analysis is used to identify sets of genes, proteins, or other biological entities that are significantly represented in an experiment, thereby providing a deeper understanding of the underlying biological processes. This work explores the most commonly used databases for functional enrichment analysis, as well as the statistical methods employed to evaluate the significance of the obtained results. Additionally, the visualization of these results are discussed, and a practical comparison between two popular approaches to functional enrichment: over-representation analysis (SEA) and gene set enrichment analysis (GSEA) is presented. This comparison is conducted using automated tools to assess the effectiveness and accuracy of the methods in different experimental scenarios.

Keywords: Functional Enrichment Analysis, Omics Sciences, Biological Databases, SEA, GSEA

Dedicatoria

Me gustaría agradecer en este espacio a todas las personas que han hecho posible que a día de hoy pueda haber terminando este trabajo y con ello mis estudios en InDat.

Gracias a los profesores que me han ido guiando desde mi etapa en el colegio Nuestra Señora del Carmen hasta el último curso del grado universitario aquí en la UVA, muy especialmente a mis tutoras Itziar Fernández y Yolanda Larriba que me han aconsejado y facilitado el desarrollo de este documento enormemente.

Gracias sobretodo a mis amigos, mis tíos, mis abuelos, mis padres y mi hermano que han confiado siempre en mí, que me dan fuerzas cada día y por quienes he llegado hasta aquí.

Para todos ellos y para todos los que tengáis la curiosidad de leer este trabajo, os mando mi abrazo más sincero.

Índice general

Índice de Figuras

Índice de Tablas

Capítulo 1 - Introducción

En la era de las ciencias ómicas, la capacidad de generar grandes volúmenes de datos biológicos ha superado significativamente la capacidad de analizarlos e interpretarlos de manera efectiva. El análisis de enriquecimiento funcional se ha convertido en una herramienta esencial para dar sentido a estos datos, permitiendo a los investigadores identificar patrones biológicos significativos y derivar conclusiones útiles a partir de ellos. Este análisis se centra en determinar si ciertos conjuntos de genes, proteínas u otras entidades biológicas están sobrerrepresentados en listas de interés, comparándolos con lo que se esperaría por azar.

El análisis de enriquecimiento funcional es especialmente relevante en el contexto de las ciencias ómicas, como la genómica, la proteómica y la transcriptómica, los resultados a menudo se presentan como listas de genes o proteínas que muestran diferencias significativas en su expresión entre distintas condiciones experimentales. Sin embargo, la simple identificación de estos elementos no es suficiente para comprender su relevancia biológica. Aquí es donde entra en juego el análisis de enriquecimiento funcional, ayudando a identificar qué funciones biológicas, procesos o rutas metabólicas están predominantemente asociados con los cambios observados [1, 2].

En este documento además se hará un repaso de las bases de datos biológicas que contienen información detallada sobre funciones génicas, rutas metabólicas, interacciones proteína-proteína y otras relaciones biológicas que son cruciales para interpretar los datos de manera coherente. Algunas de las bases de datos más utilizadas en el análisis de enriquecimiento funcional Gene Ontology (GO) [3], Kyoto Encyclopedia of Genes and Genomes (KEGG) [4] y Reactome [5] entre otras.

Además, se explorarán las diversas opciones de visualización de los resultados del análisis de enriquecimiento funcional que permite a los investigadores interpretar los datos de manera intuitiva y efectiva. Existen diversas herramientas y métodos para visualizar estos resultados, incluyendo gráficos de barras, gráficos de red y gráficos de enriquecimiento, que ayudan a representar las relaciones complejas entre los términos enriquecidos y los genes o proteínas de interés [6, 7, 8].

Por otro lado, desde el punto de vista estadístico se describirán en detalle varios de los métodos para realizar el análisis de enriquecimiento funcional, siendo los enfoques más comunes el análisis de sobre-representación (SEA, del inglés *Simple Enrichment Analysis)*, que consiste en evaluar la sobrerrepresentación de genes de interés en conjuntos predefinidos, y el análisis de conjuntos génicos (GSEA, del inglés *Gene Set Enrichment Analysis*) que, con el mismo objetivo, requiere de una variable cuantitativa que permite ordenar la lista de genes según su relación con el fenómeno de interés. Cada método tiene sus ventajas y limitaciones, y la elección del método adecuado puede depender de la naturaleza de los datos y de las preguntas de investigación específicas.

Este trabajo se propone además describir el funcionamiento de herramientas que automatizan el proceso de enriquecimiento funcional como DAVID [6], GSEA [7] y ClusterProfiler [8] poniendo en práctica esta última través de una serie de experimentos simulados y análisis comparativos que permitan determinar la capacidad de SEA y GSEA para detectar conjuntos de productos génicos que deberían estar enriquecidos en escenarios controlados.

En lo que sigue, el trabajo se estructura en un documento que cuenta con cuatro capítulos en los que se desarrolla el contenido del trabajo, un apéndice con la recopilación de todos los acrónimos empleados y una bibliografía. Concretamente, los capítulos que conforman el contenido principal de la memoria son: 'Capítulo 2 - Análisis de enriquecimiento funcional' en el que se definen las ciencias ómicas, la necesidad del enriquecimiento funcional, las bases de anotación de términos, las herramientas de enriquecimiento y las opciones de visualización, 'Capítulo 3 - Fundamentos estadísticos de los métodos evaluados' donde se describe en detalle los métodos empleados en las dos principales aproximaciones del enriquecimiento funcional (SEA y GSEA), 'Capítulo 4 - Resultados y discusión' que detalla el proceso de experimentación y puesta a prueba de los métodos descritos y por último 'Capítulo 5 - Conclusiones' el cual recoge las principales ideas y resultados obtenidos a lo largo del desarrollo del documento.

En el desarrollo de este trabajo han resultado útiles los conocimientos adquiridos en las asignaturas de Fundamentos de Programación, Algoritmos y Computación, Análisis de Datos Categóricos, Inferencia Estadística I, Inferencia Estadística II, Computación Estadística.

Capítulo 2 - Análisis de enriquecimiento funcional

2.1. Ciencias ómicas

En los últimos años, con el rápido desarrollo de nuevas tecnologías llamadas de alto rendimiento, han surgido lo que hoy se conocen como ciencias ómicas. El término ómico, acuñado en los años 80, se refiere a la recogida y análisis de grandes cantidades de datos que representan la estructura y función de un sistema biológico a un nivel determinado[9]. La característica distintiva de las tecnologías ómicas cosiste en su capacidad holística, esto significa que, en lugar de estudiar una sola parte o aspecto de un sistema biológico, estas tecnologías permiten examinar todos los componentes y cómo interactúan entre sí a nivel global. Esta visión holística permite obtener una comprensión más completa y detallada del funcionamiento biológico de tal forma que hasta ahora, nos hemos beneficiado de estas ciencias para la creación de biomarcadores que se asocian o predicen un proceso biológico, ya sea normal o relacionado con una enfermedad [10, 11]. Dentro de la ómica se pueden diferenciar cuatro grandes grupos: la genómica, la transcriptómica, la proteómica y la metabolómica.

La genómica considera todo el conjunto del DNA (del inglés *Deoxyribonucleic Acid*) completo de un organismo, incluidos todos sus genes, conocido como el "genoma". Con la llegada de la NGS (del inglés *Next Generation Sequencig*), la obtención de datos a escala genómica se ha vuelto mucho más fácil, ampliando nuestra capacidad para analizar y entender genomas completos y disminuyendo la brecha entre genotipo, la composición genética de un organismo, y fenotipo, la manifestación observable de esa genética, como las características físicas y el comportamiento.

La transcriptómica estudia el transcriptoma, es decir, el total del RNA (del inglés *Ribonucleic Acid*), que procede de la transcripción de DNA, expresado por una célula o tejido. Aquí la tecnología que dio inicio al campo fueron los microarrays de expresión génica, pero fue la reciente llegada, al igual que en la genómica, de las NGS lo que ha provocado una revolu-

ción permitiendo secuenciar millones o incluso miles de millones de fragmentos de RNA en paralelo de forma rápida, eficaz y no excesivamente costosa. En la actualidad, una de las principales aplicaciones de la transcriptómica es el análisis de la expresión de genes implicados en diferentes tipos de cánceres.

La proteómica se encarga del análisis de los conjuntos de proteínas en una célula u organismo procedentes de la traducción de mRNA (del inglés *messenger RNA*). El análisis proteómico ofrece información de las proteínas que están presentes en cierto momento lo que facilita descubrir nuevos biomarcadores e interacciones entre distintas proteinas.

La última gran rama de la ómica que cabe resaltar es la metabolómica que consiste en el perfilado global de los metabolitos presentes en una muestra biológica. La metabolómica permite detectar una amplia variedad de moléculas como péptidos, aminoácidos, carbohidratos, vitaminas, entre otros [11].

2.2. Análisis de datos de alta dimensión

Dentro del contexto de las ciencias ómicas, surgen datos de alta dimensión, caracterizados por contener una gran cantidad de variables en relación al tamaño muestral. El análisis de este tipo datos requiere de técnicas estadísticas y computacionales avanzadas para manejar, procesar y extraer información útil de conjuntos de datos extremadamente grandes y complejos.

En muchas ocasiones, los análisis planteados se realizan marginalmente, variable a variable. El resultado suele consistir en largas listas de genes, proteínas o metabolitos en las que se trata de identificar si ciertos grupos, con funciones biológicas comunes o que participan en las mismas vías metabólicas, están sobrerrepresentados en dichas listas comparado con lo que se esperaría por azar. Esto ayuda a comprender mejor las funciones biológicas, procesos celulares y rutas metabólicas más relevantes en el contexto del estudio, proporcionando una visión más clara sobre los mecanismos biológicos subyacentes. Cuando estos conjuntos conocidos aparecen sobrerrepresentados en las listas de interés, se dice que están enriquecidos [12] [1]

Por otro lado, aunque hasta el momento se ha hecho referencia a conjuntos de genes, es necesario precisar el término vía biológica. Las vías biológicas son mecanismos de los sistemas biológicos para reaccionar a estímulos tanto internos como externos, que pueden ir desde comenzar la coagulación de una herida hasta guiar el desarrollo de un óvulo fecundado. Estas vías están compuestas por una secuencia de pasos mediada por enzimas y proteínas que interactúan entre sí y que pueden involucrar múltiples componentes celulares, como metabolitos,

[1] Para referirse a los distintos productos génicos (genes, proteínas metabolitos etc) y para facilitar la nomenclatura de aquí en adelante se hará referencia a ellos como genes y a sus medidas obtenidas a partir de las tecnologías de alto rendimiento como niveles de expresión.

proteínas, o ácidos nucleicos en tre otros [13]. Es por ello que conocer cuáles de estas vías están desreguladas en cierta condición, es decir que están formadas por componentes cuyos niveles de expresión son significativamente distintos a los que presentarían en condiciones normales, es de mucha más utilidad que descubrir qué genes lo están. De este modo surge la necesidad de poder acceder a la información de qué vías biológicas existen y qué genes las componen.

2.3. Bases de datos biológicas

Las bases de datos de anotación de términos biológicos se crean para proporcionar un marco estructurado y estandarizado para describir los atributos y relaciones de entidades biológicas. Estas bases de datos permiten a los científicos comunicar y compartir datos de manera más efectiva, asegurando que los términos utilizados para describir genes, proteínas y otros elementos biológicos sean consistentes y comprensibles a nivel global.

Gracias a estas bases de datos es posible determinar qué serie de vías biológicas se van a estudiar en un análisis de enriquecimiento funcional y qué genes las componen. Sin embargo no existe una única base de datos que reúna todo el conocimiento si no que existen múltiples alternativas distintas con datos recogidos y verificados por distintos equipos. Cómo es evidente algunas bases de datos gozan de mayor prestigio y conocimiento recopilado que otras y es por ello que a continuación se detalla el funcionamiento de tres de las más importantes utilizadas en este trabajo.

2.3.1. Gene Ontology

La base de datos GO (del inglés *The Gene Ontology*) es una ontología creada en 1998 desarrollada por el *Gene Ontology Consortium*, un grupo de biólogos de diferentes instituciones que reconocieron la necesidad de una terminología estándar y organizada para describir los atributos de los genes y sus productos en diferentes organismos. Su principal objetivo es proporcionar un vocabulario estandarizado y una estructura jerárquica para describir las funciones de los genes y productos génicos en todos los organismos vivos.

Los datos contenidos en GO son el resultado de un proceso continuo de curación y revisión de información de literatura científica y bases de datos experimentales. Se basan en evidencia experimental, como estudios de expresión génica, mutagénesis, estudios de función génica y análisis de interacciones proteína-proteína. Además, se utilizan criterios estrictos para garantizar la calidad y precisión de la información proporcionada. Sin embargo, es importante tener en cuenta que la información en GO está en constante evolución a medida que se descubren nuevos conocimientos en el campo de la biología molecular y celular.

GO se organiza en tres subontologías principales que cubren diferentes aspectos de la función biológica de los genes y productos génicos:

- Componente Celular (CC, del inglés *Cellular Component*): Describe la localización subcelular donde un gen o producto génico ejerce su función, como el núcleo, la membrana plasmática o el citoplasma.
- Proceso Biológico (BP, del inglés *Biological Process*): Describe las actividades y procesos celulares que son llevados a cabo por un gen o producto génico, como la replicación del DNA, la señalización celular o el metabolismo de carbohidratos.
- Función Molecular (MF, del inglés *Molecular Function*): Describe las funciones específicas que un gen o producto génico realiza a nivel molecular, como la actividad enzimática, la unión a ligandos o el transporte de moléculas.

Cada término anotado en GO cuenta con distintos elementos:

- Identificador único y nombre: nombre como por ejemplo 'mitocondria'o 'transporte transmembrana de glucosa', un ID único de 7 dígitos precedido por el prefijo 'GO:' como podrían ser GO:0005739 o GO:1904659.
- Aspecto: Especifica a cuál de las tres subontologías (CC, BP, MF) pertenece el término.
- Definición: Descripción textual de lo que representa el término y referencias.
- Relación: Indica cómo cierto término se relaciona con otros en la ontología. Todos los términos, menos los términos raíz de cada subontología, tienen como relación 'is a', que significa que son una subclase de otro término de la ontología. Las principales y más comunes relaciones usadas en GO son:
    - 'is a': Es la relación que forma la estructura básica de GO. SI se dice que A 'is a' B entonces es que A es un subtipo del nodo B.
    - 'part of': Esta relación es útil para representar relaciones parte-todo de tal forma que si A 'part of' B, A es un componente integral de B, lo que significa que A no puede existir de manera independiente en el contexto biológico definido. A siempre estará presente como una parte dentro de la estructura o proceso más amplio que es B.
    - 'has part': Es la relación inversa a "parto of". Cuando decimos que B 'has part A', significa que B incluye a A como uno de sus componentes. B está compuesto por A junto con otros posibles componentes.
    - 'regulates': Cuando se dice que A 'regulates' B, significa que A ejerce algún tipo de influencia sobre B. Esta influencia puede ser directa o indirecta, y puede afectar la actividad, frecuencia, tasa o extensión de B. Esta relación se utiliza específicamente

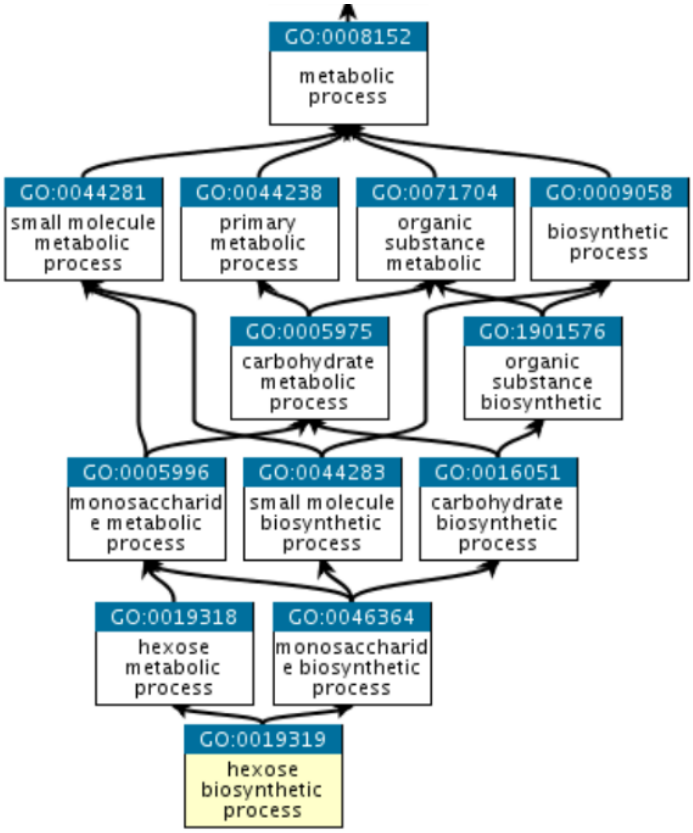


Figura 2.1: Grafo de términos GO - Imagen tomada de [3]

para significar 'necesariamente' regula: si A y B están presentes, A siempre regula B, pero B puede no estar siempre regulado por A.

Todas estas relaciones y términos dan lugar a una estructura de grafo acíclico no dirigido como se puede observar en la Figura 2.1 en la que los distintos términos se relacionan de forma jerárquica partiendo desde conceptos más generales en la parte superior del árbol a conceptos de más bajo nivel en los últimos nodos, aunque cabe señalar que cualquier término puede tener más de un "padre". Es interesante destacar que más que un único grafo podrían distinguirse 3 distintos que están relacionados, uno para cada aspecto (CC, BP, MF) de tal forma que cada uno tiene su nodo raíz y no hay relaciones 'is a' entre términos de distintos aspectos (aunque los grafos se comuniquen con el resto de relaciones existentes).

Sin embargo uno de los aspectos más importantes de GO, especialmente en el caso del análisis de enriquecimiento funcional, es la existencia de las anotaciones GO. Las anotaciones GO se crean asociando un gen a uno o varios de los términos GO descritos, por lo que gracias a ellas es posible determinar cómo funciona cada gen a nivel molecular (según sus anotaciones con términos MF) , en qué parte de la célula funciona (según sus anotaciones con términos CC) y qué procesos biológicos (vías, programas) ayuda a llevar a cabo (según sus anotaciones con términos PB).

Cada anotación GO cuenta con al menos 4 elementos que son: el identificador del producto génico, el identificador del término GO, una referencia y una evidencia.

Todas las anotaciones GO están respaldadas en última instancia por la literatura científica, ya sea directa o indirectamente. En GO, la evidencia de apoyo se presenta en forma de códigos de evidencia GO y una referencia publicada o una descripción de la metodología utilizada para crear la anotación. Los códigos de evidencia GO describen el tipo de evidencia que puede ser de uno entre los siguientes tipos:

- Evidencia experimental: Indica que hay evidencia experimental directa que apoya la anotación del gen en el término indicado.
- Anotaciones filogenéticas: Las anotaciones de base filogenética se derivan de un modelo explícito de ganancia y pérdida de función génica en ramas específicas de un árbol filogenético. Cada anotación inferida puede rastrearse hasta las anotaciones experimentales directas que sirvieron de base para esa afirmación.
- Análisis computacional: significa que la información sobre la función o característica de un gen ha sido inferida a través de herramientas y algoritmos bioinformáticos en lugar de métodos experimentales directos.
- Declaración del autor: Los códigos de declaración del autor indican que la anotación se ha realizado a partir de una declaración realizada por el autor o autores en la referencia citada.
- Declaración de 'curator': Indican una anotación hecha en base a juicios que no corresponden a ninguno de los otros códigos.
- Anotación electrónica: Las anotaciones electrónicas no se revisan manualmente (aunque el propio método suele someterse a diversas evaluaciones de calidad). Estas anotaciones se basan en última instancia en la homología y/u otra información experimental o de secuencias, pero generalmente no pueden rastrearse hasta una fuente experimental.

Además de toda esta información sobre la fiabilidad de las anotaciones el *Gene Ontology Consortium* implementa una serie de consultas automáticas para comprobar la calidad de las anotaciones aportadas a GO [14].

Con la información contenida en las anotaciones de esta ontología es posible obtener para cada proceso biológico los genes que presentan anotaciones con dicho término logrando así definir el conjunto de genes para cada vía biológica.

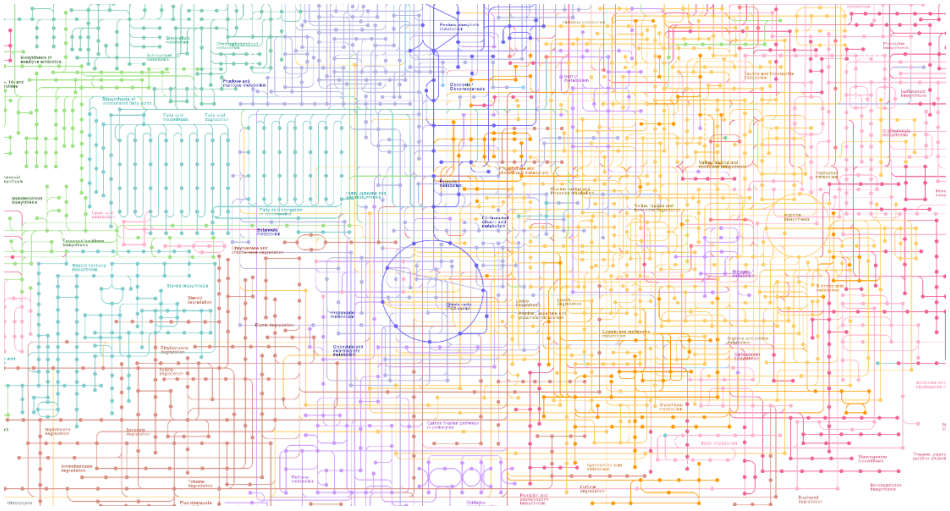


Figura 2.2: Gráfico de rutas metabólicas de KEGG Pathway - Imagen tomada directamente de KEGG [4]

2.3.2. KEGG

La base de datos KEGG (del inglés *Kyoto Encyclopedia of Genes and Genomes*) es un repositorio integral que facilita el análisis detallado de las funciones de los genes y cómo se relacionan con funciones biológicas más complejas. Fue creada en 1995 el Instituto de Investigación Química de la Universidad de Kioto y es mantenida por Minoru Kanehisa y su equipo.

KEGG proporciona herramientas y bases de datos que ayudan a interpretar datos genéticos. Esto incluye información sobre los genes, los procesos biológicos en los que participan, y cómo estos genes y procesos interactúan en sistemas celulares completos. En otras palabras, KEGG permite a los científicos comprender mejor el papel de los genes dentro del contexto más amplio de las funciones y procesos biológicos, ayudando en el estudio de la biología y la medicina. KEGG cuenta con dos bases de datos fundamentales.

Por un lado, la base de datos 'Pathway' es una colección de diagramas gráficos (mapas de vías) para las vías bioquímicas que son en su mayoría vías metabólicas. Cada vía o red se presenta en forma de diagrama interactivo con una forma similar a lo que se puede observar en la Figura 2.2, donde los componentes moleculares se representan gráficamente y se vinculan con información detallada sobre su función y regulación. Las vías metabólicas y las redes de señalización se organizan en una jerarquía de categorías y subcategorías, que facilita la navegación y la búsqueda de información.

Para cada componente de un gráfico de vía en la base de datos 'Pathway' de KEGG, se proporciona información detallada sobre su función y regulación. Esto puede incluir:

- Nombre del componente: Identificación del componente molecular, como genes, pro-

teínas, metabolitos u otras moléculas biológicas.

- Función: Descripción de la función biológica o actividad molecular del componente.
- Regulación: Información sobre cómo se regula la actividad del componente, incluidos los factores que activan o inhiben su función.
- Interacciones**:** Relaciones con otros componentes en la misma vía o red biológica, incluidas las interacciones físicas, las vías metabólicas relacionadas y cualquier retroalimentación o regulación cruzada.
- Referencias**:** Enlaces a la literatura científica que respalda la información proporcionada sobre el componente, como estudios experimentales, revisiones o bases de datos externas.

Por otro lado la base de datos de 'Genes' cuenta con información de 24 genomas completos y 12 parciales. Cada entrada en la base de datos representa un gen individual y sus campos más importantes son:

- Identificadores de Genes y nombre: Un identificador único para cada gen, generalmente un código alfanumérico así como el nombre del gen.
- Organismo: Nombre del organismo KEGG al que pertenece el gen.
- KO: Los números KO (KEGG Orthology) se asignan mediante análisis de homología, que busca genes y proteínas similares en diferentes organismos. Si un gen o proteína es homólogo a un miembro conocido de una clase funcional específica, se le asigna el mismo número KO.
- 'Pathway': Vías biológica a en las que participa el gen, asignadas a través de su número KO. Consiste en una lista de identificadores de entradas de la base de datos '*Pathway*' de KEGG.
- Secuencias de DNA y RNA: Las secuencias nucleotídicas del gen, incluyendo intrones y exones.
- Secuencias de Proteínas: Secuencias de aminoácidos si el gen codifica una proteína.
- Otras Bases de Datos: Links a bases de datos externas.

Toda esta información se obtienen de una variedad de fuentes públicas y privadas, incluyendo: bases de datos de secuencias de genes (GenBank, RefSeq, EMBL/DDJB), bases de datos de literatura biomédica (PubMed, MEDLINE, JNLPBA), bases de datos de interacciones moleculares (BIND, IntAct, Reactome) y datos experimentales de genómica, proteómica, metabolómica y transcriptómica. El lector puede ampliar esta información en [15, 16, 17].

2.3.3. Molecular Signatures Database

La base de datos *Molecular Signatures Database* (MSigDB) es una base de datos crucial para el análisis de enriquecimiento de conjuntos génicos. Fue desarrollada por el *Broad Institute*, una colaboración entre el Instituto Tecnológico de Massachusetts y la Universidad de Harvard, siendo publicada por primera vez en un artículo seminal en 2005. MSigDB ha sido mantenida y mejorada continuamente por investigadores de este incluyendo a destacados investigadores en el campo de la bioinformática y la biología computacional. La validación de los conjuntos de genes incluidos en MSigDB es rigurosa. Los conjuntos deben estar respaldados por publicaciones revisadas por pares y, en muchos casos, provienen de colaboraciones con investigadores externos que han identificado y caracterizado estos conjuntos de genes en estudios experimentales.

MSigDB contiene decenas de miles de anotaciones de conjuntos de genes y está organizada en varias colecciones de estos pertenecientes a la categoría de 'Human' o 'Mouse', teniendo cada colección un propósito específico:

- H 'Hallmark Gene Sets': Conjuntos de genes que representan procesos biológicos fundamentales y están bien definidos.
- C1 'Positional Gene Sets': Conjuntos de genes agrupados por su localización cromosómica.
- C2 'Curated Gene Sets': Conjuntos de genes basados en publicaciones científicas y bases de datos biológicas, incluyendo conjuntos de vías de señalización y regulaciones.
- C3: 'Regulatory Target Gene Sets': Conjuntos de genes que incluyen secuencias diana de factores de transcripción y microRNA.
- C4: 'Computational Gene Sets': Conjuntos de genes derivados de análisis computacionales de datos de expresión génica.
- C5 'GO Gene Sets': Conjuntos de genes categorizados según términos de Gene Ontology.
- C6 'Oncogenic Signatures': Conjuntos de genes asociados con firmas oncogénicas.
- C7 'Immunologic Signatures': Conjuntos de genes relacionados con firmas inmunológicas.
- MH 'Mouse-ortholog hallmark gene sets': Versiones de conjuntos de genes en la colección Hallmarks de MSigDB mapeados a sus ortólogos en ratón.
- M1 'Positional gene sets': Conjuntos de genes correspondientes a bandas citogenéticas cromosómicas de ratón.

- M2 'Curated gene sets': Conjuntos de genes de bases de datos de vías en línea, publicaciones en PubMed y conocimiento de expertos en el dominio.
- M3 'Regulatory target gene sets': Conjuntos de genes basados en predicciones de dianas génicas para secuencias de microRNA y sitios de unión predichos de factores de transcripción.
- M5 'Ontology gene sets':Consisten en genes anotados por el mismo término de ontología.
- M8 'Cell type signature gene sets': Conjuntos de genes curados a partir de marcadores de grupos identificados en estudios de secuenciación de células individuales de tejidos de ratón.

Cada conjunto de genes en MSigDB es cuidadosamente anotado y vinculado a la literatura científica, lo que facilita su uso en diversos análisis de enriquecimiento y estudios comparativos [18, 19].

2.4. Metodologías de enriquecimiento funcional

Las bases de datos descritas, entre otras, intervienen en una primera etapa del enriquecimiento funcional. Existen múltiples herramientas (PANTHER [20], DAVID [6], Enrichr [21], GSEA [7] o `ClusterProfiler` [8] en la que se centra este documento) que se encargan de realizar el mapeado sistemático de la ingente cantidad de genes inicial procedente de las tecnologías de alto rendimiento a los términos de cierta anotación biológica (términos GO, KEGG entre otros) para posteriormente examinar el enriquecimiento de los genes para cada uno de los términos de anotación [22].

Estas herramientas cumplen con la función de transformar las listas de genes y sus niveles de expresión producidos por los experimentos de alto rendimiento en listas de procesos biológicos, funciones moleculares o vías biológicas que se encuentran enriquecidas junto con medidas de su significación estadística, en un proceso similar al que se representa en la Figura 2.3. Para obtener dichas listas estas herramientas emplean alguno de los múltiples métodos estadísticos existentes que se pueden dividir en realidad en tres grandes familias que se detallan a continuación.

2.4.1. 1º Generación - *Simple Enrichment Analysis*

La primera generación de métodos desarrollados fueron lo que se conoce como SEA (del inglés *Simple Enrichment Analysis*) u ORA (del inglés *Over Representation Analysis*). Estos enfoques surgieron en un primer momento tras el desarrollo de las tecnologías de alto rendimiento y junto con la aparición de GO. Para emplear estos métodos es necesario proporcionar una

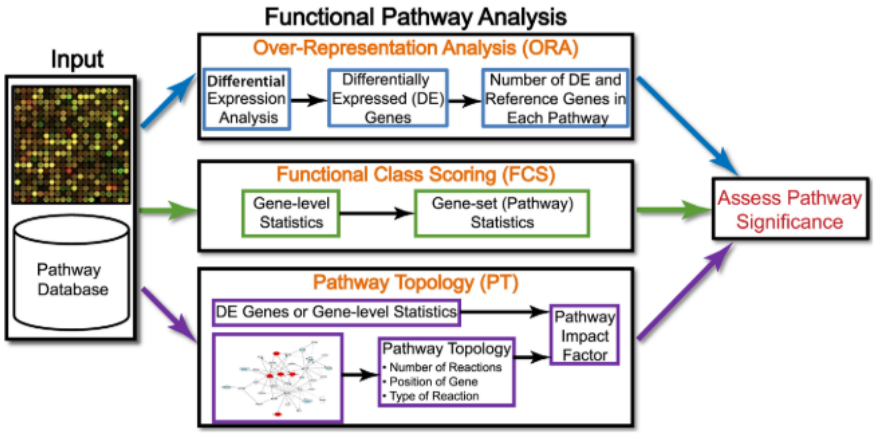


Figura 2.3: Proceso de análisis de enriquecimiento funcional - Imagen tomada de [23]

lista de genes 'interesantes' para el investigador que se pueden obtener, por ejemplo, teniendo en cuenta el número de veces, o *fold change* (FC), que aumenta o disminuye la medida de cada gen en la condición estudiada respecto a condiciones normales de tal forma que para cada vía biológica se cuenta el número de genes anotados en ella que pertenecen a la lista proporcionada.

A continuación, se lleva a cabo el cálculo de un p-valor que determine la probabilidad de que el número de genes de interés en cada vía se haya dado de forma aleatoria en vez de estar motivado por la naturaleza del proceso, cálculo que puede realizarse con métodos estadísticos clásicos como Chi-Cuadrado, el test exacto de Fisher o un test de enriquecimiento basado en la distribución hipergeométrica.

Estos métodos cuentan con varias limitaciones que hacen preferibles a las nuevas generaciones de algoritmos. En primer lugar estos métodos emplean una discretización creando una lista de genes de interés sin tener en cuenta el valor numérico de los niveles de expresión de los experimentos de los que provienen desechando así información que podría ser de gran valor.

Esta discretización tiene también como inconveniente añadido la selección del límite de significación estadística α que permite hacer la selección de genes de interés. La variación de este valor provoca inevitablemente la variación del conjunto de genes DE y en consecuencia la variación del conjunto de vías biológicas detectadas como enriquecidas. Por lo tanto existe en este proceso un componente de subjetividad que hay que tener en cuenta.

Además de esto, al tratar a todos los genes por igual, en el SEA se asume que los genes son independientes entre sí, pero precisamente el análisis de enriquecimiento funcional se basa en la premisa de que existen reacciones coordinadas en los genes, por lo que ignorar la correlación entre estos puede llevar a resultados sesgados. De igual forma en SEA no se tiene en cuenta la relación entre vías biológicas lo que constituye una pérdida de información adicional [23, 22].

2.4.2. 2º Generación - *Gene Set Enrichment Analysis*

La segunda generación de métodos de enriquecimiento funcional es la conocida como GSEA (del inglés *Gene Set Enrichment Analysis*) o FCS (del inglés *Functional Class Scoring*). En contraposición con los métodos de primera generación, no es necesario proporcionar una lista de genes preseleccionados previamente. Concretamente, la principal diferencia con estos métodos es que los genes se ordenan según alguna medida continua de la cantidad de relación con el fenómeno que se está estudiando, eliminando la limitación de la discretización mencionada en los métodos de primera generación. Al aplicar estos métodos cambios sutiles en colecciones de genes que antes no superaban el límite de selección para ser genes diferencialmente expresados (DE, del inglés *Differentially Expressed*) pueden ser ahora tenidos en cuenta.

La idea de los algoritmos GSEA pasa por ordenar en primer lugar los genes según cierto estadístico para después emplear la posición en el *ranking* de los genes anotados en cada vía con el objetivo de determinar si la distribución de dichos genes en la lista es significativamente distinta a lo esperable por el azar, es decir si los *rankings* del conjunto de genes de dicha vía se sitúan en posiciones demasiado altas o demasiado bajas.

Este paso se lleva a cabo mediante algún estadístico como el test de Kolmogorov Smirnov o el test de Wilcoxon dando lugar a lo que se conoce como *Maximum Enrichment Score*. Finalmente, para obtener el p-valor simplemente se realiza un test de permutaciones cambiando múltiples veces las etiquetas de los genes en las distintas medidas y recalculando en cada iteración el estadístico anterior. De esta forma es posible obtener una distribución de dicho estadístico y obtener un p-valor del estadístico medido.

Sin embargo, la limitación de no tener en cuenta las relaciones entre vías biológicas prevalece en los métodos GSEA lo que puede conllevar a errores debido al hecho de que cada gen puede pertenecer a distintas vías de modo que cuando una vía está verdaderamente enriquecida otras con varios de sus genes anotados podrían ser también identificadas como enriquecidas sin estarlo [22, 23, 24].

2.4.3. 3º Generación - *Modular Enrichment Analysis*

La última generación de métodos de enriquecimiento funcional se conoce como MEA (del inglés *Modular Enrichment Analysis*) o PT (del inglés *Pathway Topology*). MEA introduce la consideración de la estructura y relaciones de la red de anotaciones biológicas. Esto significa que no solo se analizan los términos aislados, sino también sus interrelaciones y cómo interactúan dentro de redes más grandes de procesos biológicos.

MEA utiliza las interrelaciones entre los términos de anotación biológica, como las relaciones

jerárquicas y de dependencia entre los términos GO, para mejorar la especificidad y sensibilidad del descubrimiento de términos enriquecidos. Basándose en la similitud de expresión genética, las jerarquías y las interacciones entre términos recogidas en las bases de datos como GO, los términos de anotación y genes se agrupan en redes o módulos funcionales que representan procesos biológicos completos para los cuales se calcula el enriquecimiento.

Al considerar las relaciones entre términos, MEA puede proporcionar una visión más precisa y relevante biológicamente de los resultados de enriquecimiento. Además este enfoque ayuda a minimizar las redundancias en los términos de anotación, enfocándose en módulos o redes biológicas completas en lugar de términos aislados.

Sin embargo la limitación evidente de los métodos MEA es la posibilidad de la existencia de términos 'huérfanos' que podrían quedarse fuera del análisis al no tener relaciones fuertes con otros términos [22, 23, 25].

Aunque este enfoque es interesante y merece la pena hacerle mención, por simplicidad y limitaciones de tiempo y espacio, en este trabajo no se evalúa ningún procedimiento perteneciente a esta familia.

2.5. Herramientas de enriquecimiento funcional

Las herramientas de enriquecimiento funcional realizan automáticamente tareas como la asignación de anotaciones funcionales a los genes accediendo a las bases de datos correspondientes, la identificación de conjuntos de genes relevantes mediante los métodos estadísticos descritos, el cálculo de p-valores de enriquecimiento y la visualización de resultados de tal forma que eliminan la necesidad de realizar estas tareas manualmente, ahorrando tiempo y esfuerzo.

Existen multitud de herramientas que pueden utilizarse para llevar a cabo el enriquecimiento funcional como pueden ser PANTHER [20], DAVID [6], Enrichr [21], GSEA [7] y en la que se centrará este documento que es `ClusterProfiler` [8]. Muchas de ellas están disponibles a través de Internet de forma totalmente libre, pero para la realización de este trabajo se ha elegido `ClusterProfiler` por ser muy completa y fácil de integrar con el resto de código R que se emplea.

2.5.1. ClusterProfiler

`ClusterProfiler` [8] es una herramienta de análisis funcional ampliamente utilizada en bioinformática para el enriquecimiento y la visualización de datos genómicos. Fue creada por Guangchuang Yu y se publicó por primera vez en 2012. La herramienta es mantenida activamente por Guangchuang Yu y el equipo de desarrolladores de Bioconductor [26], una

plataforma de software para el análisis de datos genómicos en R. Esta herramienta emplea como principales bases de anotación GO, KEGG y MSigDb entre otros.

Para poder hacer uso de esta plataforma en R es necesario asegurarse de tener Bioconductor instalado para poder después cargar el paquete de `ClusterProfiler`.

Uso del paquete y generación de resultados

Una vez cargada la librería ya es posible acceder a las distintas funciones que esta ofrece. `ClusterProfiler` es posible o bien realizar un análisis de enriquecimiento funcional con una base de datos de conjuntos de genes propia para lo cual es posible usar las funciones `enrich()` (para SEA) y `GSEA() (para GSEA)` o bien emplear funciones que hagan uso de las bases de anotación mencionadas en cuyo caso se emplearán las funciones `enrichGO()`, `gseGO()`, `enrichKEGG()`, `gseKEGG()`, `enrichDO()` y `gseDO()`. Estas funciones permiten realizar el análisis de enriquecimiento funcional a partir de una lista de genes y una base de anotación permitiendo además cambiar parámetros como el tamaño mínimo o máximo de los conjuntos de genes a tener en cuenta, el límite del p-valor a partir del cuál se determina que cierto conjunto está enriquecido o el valor del parámetro de peso p empleado en GSEA [24].

Todas las funciones mencionadas dan como resultado un objeto de clase 'enrichResult' o 'gseaResult' con información detallada de qué términos están enriquecidos y cuáles son sus genes. Para poder obtener información de estos resultados basta con utilizar la función `summary()` sobre estos objetos o emplear el paquete de visualización `enrichplot` como se detalla en la sección 2.6

Algoritmos empleados

A diferencia de otras herramientas de enriquecimiento populares como DAVID que solo ofrece la posibilidad de realizar el análisis de enriquecimiento funcional empleando métodos SEA o la herramienta GSEA que solo ofrece la opción de aplicar métodos GSEA, `ClusterProfiler` ofrece ambos enfoques. En el caso de SEA el software implementa el test exacto de Fisher mientras que en el caso de GSEA el algoritmo es el presentado por Subramanian [24] de forma que en las funciones de GSEA se puede establecer el valor de cierto parámetro p como el parámetro 'exponent' para poder cambiar entre las distintas variantes existentes.

2.6. Visualización de los resultados

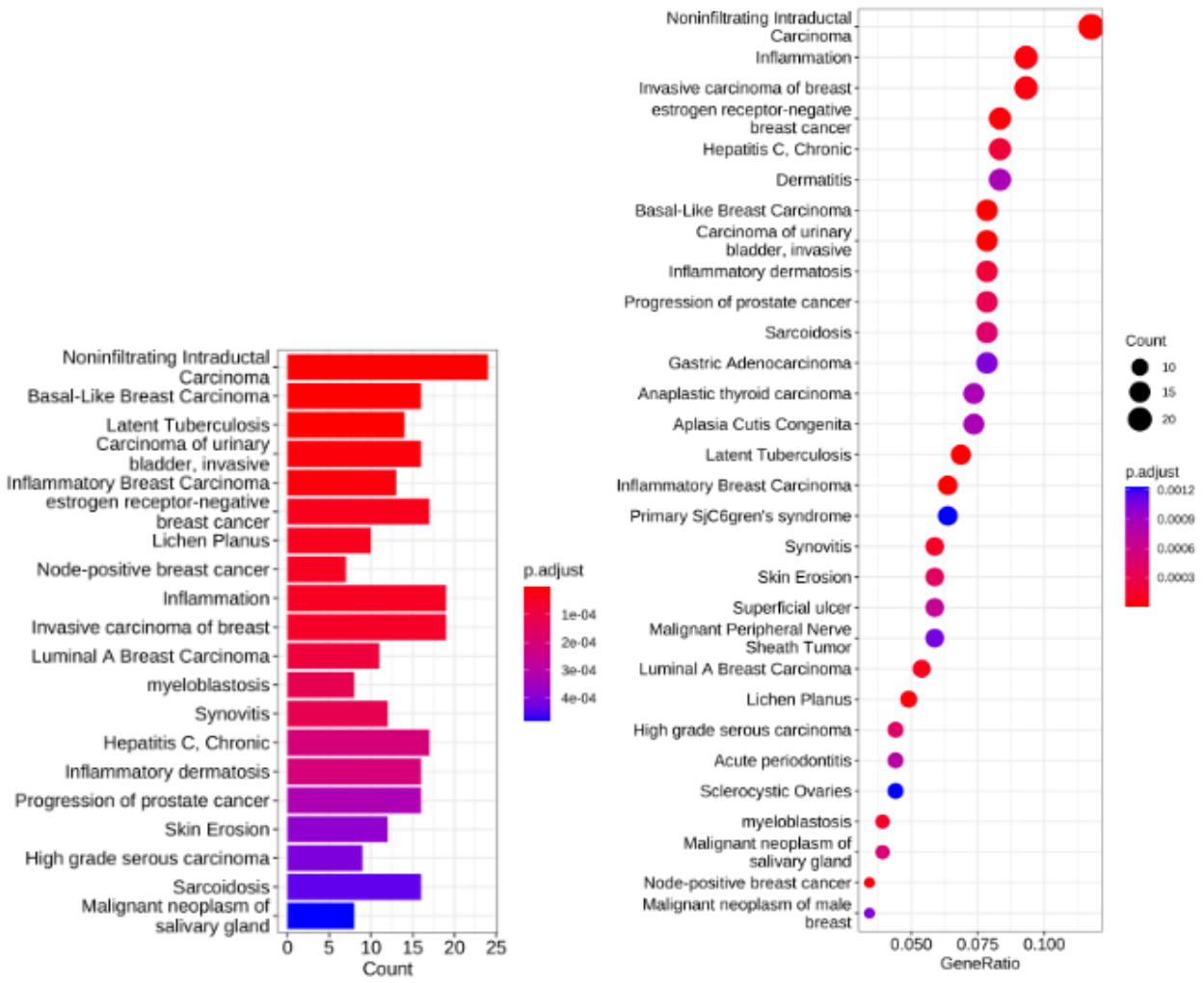
Una vez realizado el análisis de enriquecimiento funcional es natural buscar una forma visual y efectiva de mostrar los resultados obtenidos. Para ello existen múltiples alternativas

pero en este documento se va a centrar el foco en las opciones descritas en [8, 27] que ofrece el paquete `enrichplot` de R que se utiliza para visualizar el resultado del análisis de enriquecimiento SEA y GSEA de algunas de las herramientas antes mencionadas, en concreto `ClusterProfiler` [8]. Todas las imágenes presentadas en este apartado proceden de la documentación de `ClusterProfiler` [28]

2.6.1. Diagrama de barras y diagrama de puntos

Dos plots sencillos para obtener una idea de qué vías están más enriquecidas son los plots de barras o de puntos en los que se muestran las vías cuya prueba estadística (SEA o GSEA) muestra un p-valor menor o mayor representado de acuerdo con la tonalidad de las barras. Además de esta información también se representa el conteo de genes de cada vía o el ratio, es decir, el número de genes en el conjunto de entrada que pertenecen al término de ontología, dividido por el número esperado de genes en el conjunto de referencia que pertenecerían al mismo término .

Estas dos últimas magnitudes se pueden representar con la longitud de las barras así como con la distancia y tamaño de los puntos como puede observarse en las Figuras 2.4a y 2.4b.



(a) Diagrama de barras de `enrichplot`

(b) Diagrama de puntos de `enrichplot`

Figura 2.4: Estos plots están disponibles con las funciones `dotplot()` y `barplot()` de `enrichplot` [29].

2.6.2. Red Gen-Concepto

Además de conocer cuáles son las vías detectadas como enriquecidas muchas veces es también necesario conocer qué genes las componen. La solución a esta necesidad es la creación de una red en la que los nodos pequeños representan genes de los que salen una o más líneas que los conectan con nodos más grandes los cuales representan las distintas vías biológicas tal y como se muestra en la Figura 2.5.

Además de esta estructura básica se puede añadir color indicando el FC de cada gen o cambiar el tamaño de los nodos de las vías para representar cuántos genes tienen asociados.

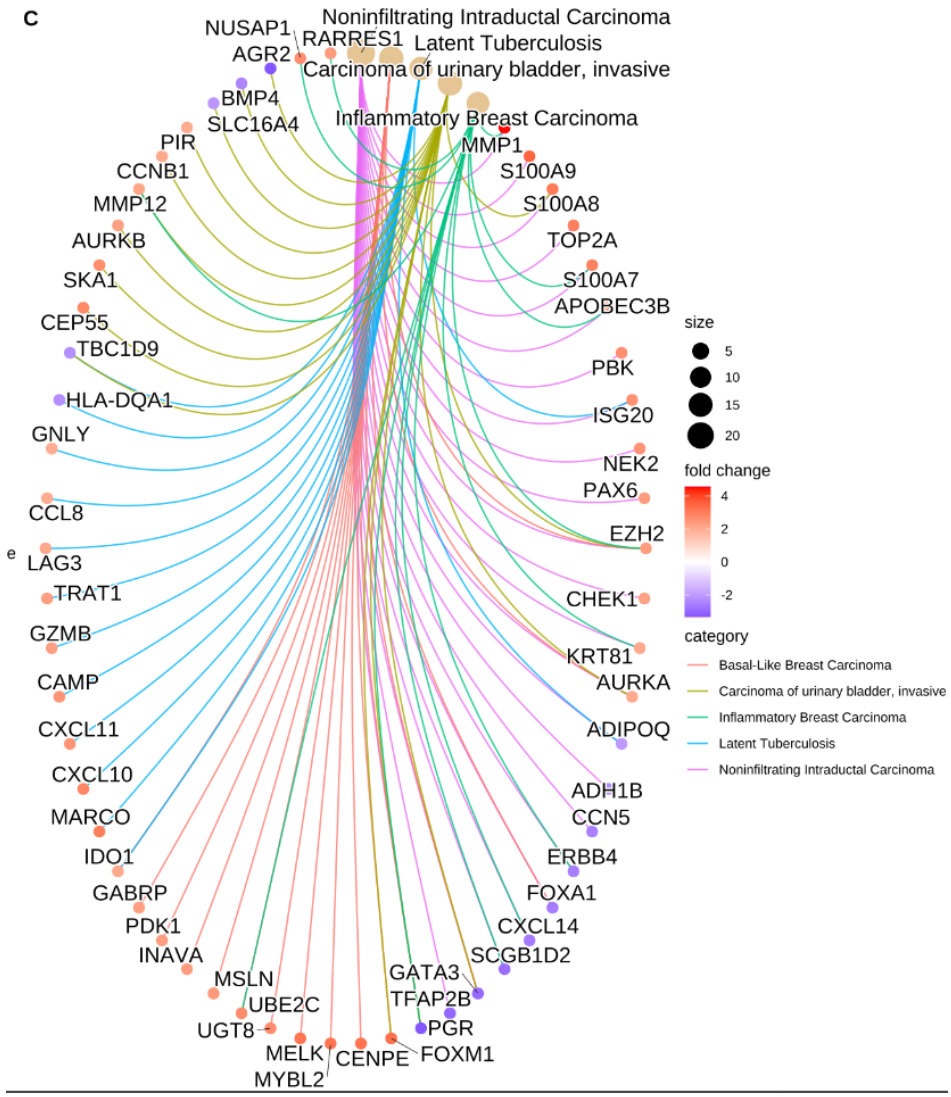


Figura 2.5: Red Gen-Concepto creada con la función `cnetplot()` de `enrichplot` [29].

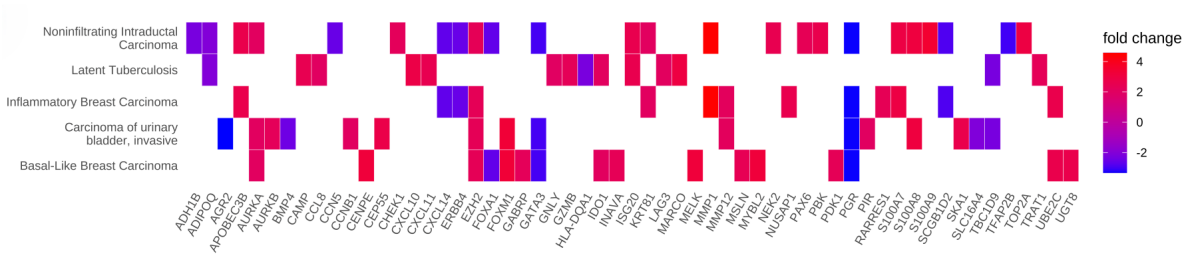


Figura 2.6: Mapa de calor creado con la función `heatplot()` de `enrichplot` [29].

2.6.3. Clasificación funcional con mapa de calor

El problema de gráficos como el gráfico de red mostrado anteriormente es que es fácil que al aumentar el número de vías biológicas a tener en cuenta aumente del mismo modo la complejidad de interpretación.

Por ello una alternativa es el uso de un mapa de calor en el que cada columna represente un gen y cada fila una vía determinada de tal manera que cada celda coloreada significa que el gen de su columna está anotado en la vía enriquecida que representa su fila. Además de esto se puede añadir también la información del FC a través del color.

2.6.4. Árbol

Otra cuestión de interés al obtener la lista de vías enriquecidas se trata de lograr agruparlas de forma que se puedan establecer distintos grupos de vías similares. Una forma de hacer esto es aplicando el *clustering* jerárquico.

Sin entrar en detalles sobre el funcionamiento de los algoritmos de *clustering* la idea consiste en calcular la distancia entre cada par de vías, utilizando por ejemplo el índice de Jaccard que mide la proporción de genes compartidos entre vías, de tal forma que en cada iteración el algoritmo busca los dos *clusters* más similares entre sí y los fusiona hasta quedarse con el número de *clusters* deseado.

Tanto el número de *clusters* final como el criterio para determinar la similitud entre *clusters* (centroide, media, mediana, completo) pueden especificarse.

De este modo el resultado final es un gráfico como el que se muestra en la Figura 2.7 en el que se agrupa con un color distinto las vías que comparten más genes entre sí.

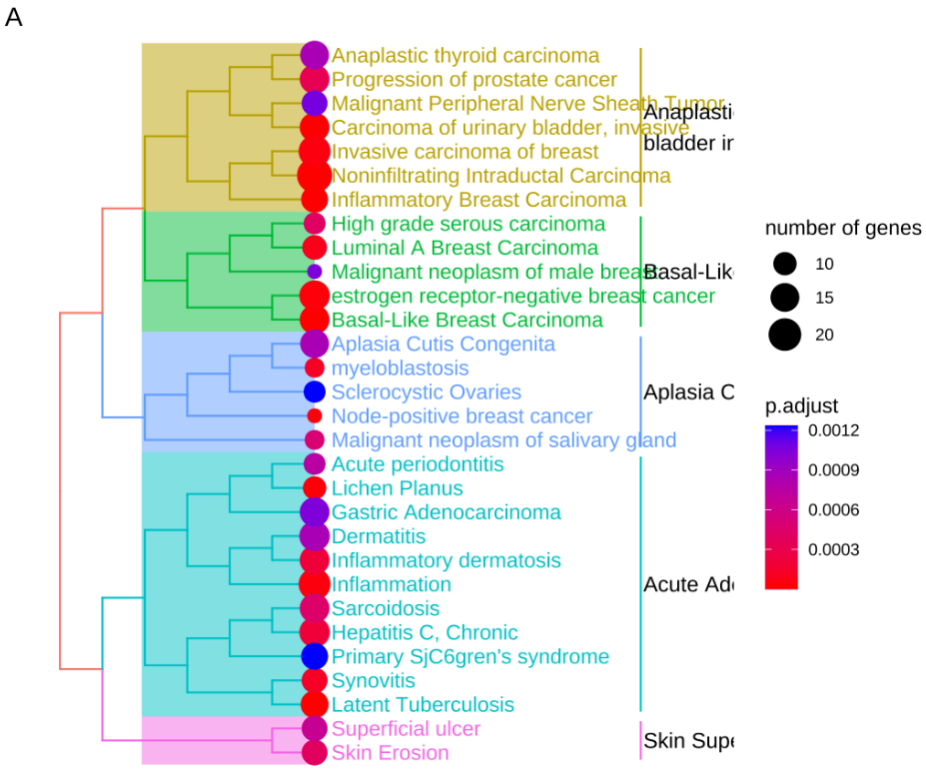


Figura 2.7: Árbol de *clustering* jerárquico creado con la función `treeplot()` de `enrichplot` [29].

2.6.5. Mapa de enriquecimiento

Con un propósito similar al del árbol descrito anteriormente, el mapa de enriquecimiento muestra las vías enriquecidas de tal forma que las más cercanas y con más uniones entre sí simbolizan vías con más genes en común tal y como se muestra en la Figura 2.8. Además es posible crear grupos de tal forma que los nodos de distintos *clusters* se representan con colores diferentes.

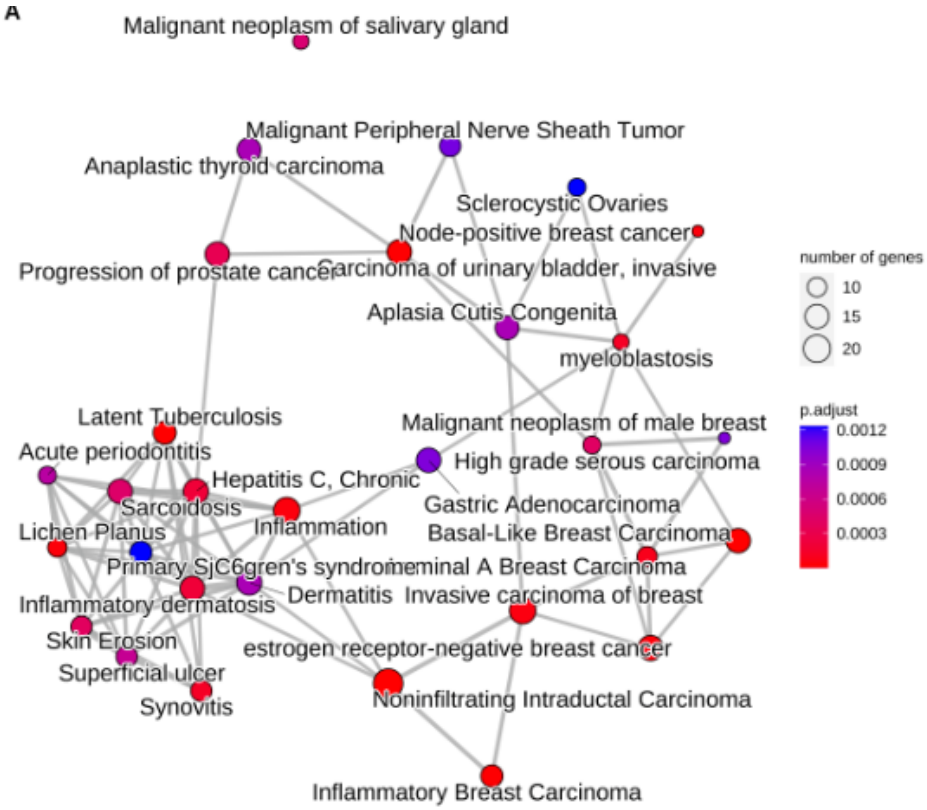


Figura 2.8: Mapa de enriquecimiento creado con la función `emaplot()`. Además existe la opción de crear *clusters* previamente con la función `compareCluster()`. Ambas funciones pertenecen al paquete `enrichplot` [29].

2.6.6. Gráfico de línea de cresta

Este gráfico se emplea para visualizar la distribución de las medidas de los genes más importantes de las vías biológicas detectadas como enriquecidas con curvas dispuestas en tantas filas como vías a comparar reflejando cada una de ellas la densidad de cada distribución como se observa en la Figura 2.9.

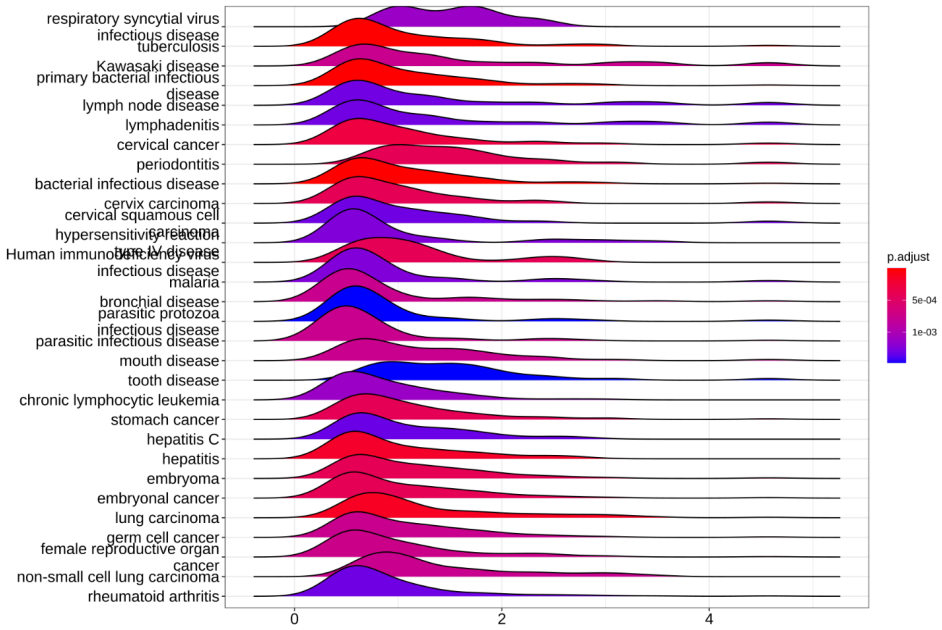


Figura 2.9: Gráfico de línea de cresta creado con la función `ridgeplot()` de `enrichplot` [29].

2.6.7. *Running score* y lista preordenada de GSEA

Otro plot de gran utilidad en el análisis de los resultados de GSEA son los plots de *running score*. En estos gráficos para cada vía biológica se crea una curva que representa la puntuación de enriquecimiento acumulativa a medida que se avanza por la lista ordenada de genes de forma que en el eje horizontal del gráfico la izquierda indica posiciones más altas de la lista y la derecha posiciones más bajas. Esta curva crece cuando el gen de la lista ordenada pertenece al conjunto de genes anotados en la vía y baja cuando no lo están.

Además de esta curva, el gráfico muestra marcas verticales en la parte inferior para indicar las posiciones de la lista en las que han aparecido los genes de la vía.

La posición en el eje X donde la curva alcanza su valor máximo (o mínimo) indica el punto de mayor enriquecimiento. Este valor máximo es el *Maximum Enrichment Score* ya mencionado. De esta forma en un supuesto caso de medidas de expresión de genes en el que la lista se ordene por FC de modo que en la parte superior se encuentren los genes sobreexpresados y en la parte inferior se encuentren los genes subexpresados, para cierta vía biológica el hecho de que el valor máximo de la curva se encuentre a la izquierda (principio de la lista) o derecha (final de la lista) supondrá conocer si la vía está enriquecida por sobreexpresión o infraexpresión respectivamente.

Por último la superposición de curvas con colores tal y como se muestra en 2.10 permite comparar distintas vías.

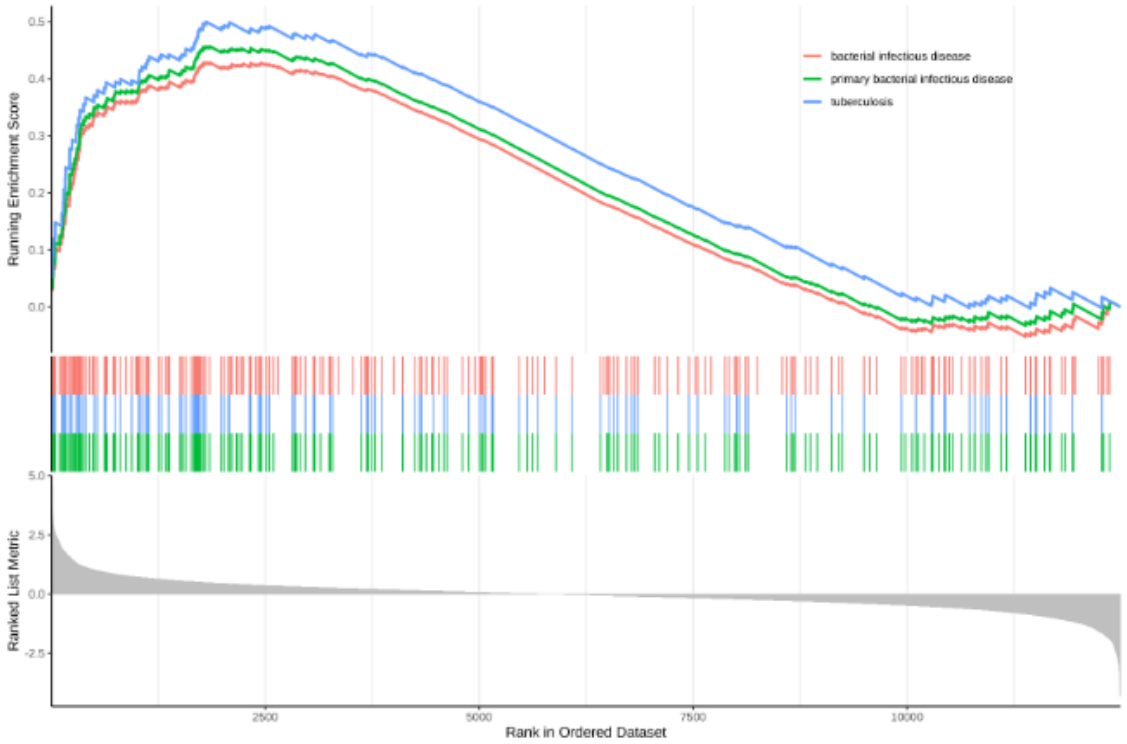


Figura 2.10: Plot de GSEA creado con la librería `gseaplot()` de `enrichplot` [29].

Capítulo 3 - Fundamentos estadísticos de los métodos evaluados

Una vez explicado el contexto de las ciencias ómicas y las tecnologías de alto rendimiento, la necesidad del enriquecimiento funcional y sus distintas metodologías, las ontologías disponibles y las opciones de visualización de los resultados obtenidos es importante describir las técnicas estadísticas que se emplean en las distintas herramientas de enriquecimiento funcional puesto que hoy en día estas aplicaciones gozan de una gran popularidad pero son a menudo utilizadas por los profesionales a modo de 'caja negra', es decir, sin un conocimiento profundo de los cálculos subyacentes [12].

En la sección 2.4 se ha presentado una visión general de los distintos enfoques metodológicos que actualmente se utilizan para llevar a cabo un enriquecimiento funcional. Este capítulo se centra en los fundamentos estadísticos que se han utilizado en este trabajo.

3.1. Single Enrichment Analysis (SEA)

El SEA u ORA parte de una situación en la que se cuenta con un listado de genes seleccionados de entre todos los evaluados en el experimento según cierto criterio. Este criterio suele ser establecido combinando dos conceptos: la significación estadística y la relevancia práctica. Para cuantificar el primero se suelen utilizar p-valores asociados a contrastes sobre la diferencia de expresión de cada gen en las distintas condiciones de estudio, y para el segundo una medida habitual es el FC.

A partir de la lista de N genes evaluados, y dado el grupo de m_i genes anotados en la categoría i-ésima de la base de datos que se va a utilizar como fuente de conocimiento biológico, denotado por C_i, es posible construir la tabla de contingencia representada en la Tabla 3.1. En esta tabla, n es el número de genes DE en el conjunto de N genes evaluados y k el número de genes DE anotados en la categoría C_i.

	$\in C_i$	$\notin C_i$	
DE	k	n - k	n
$\overline{DE}$	m_i-k	N + k - n - m_i	N -n
	m_i	N-m_i	N

Tabla 3.1: Tabla de contingencia

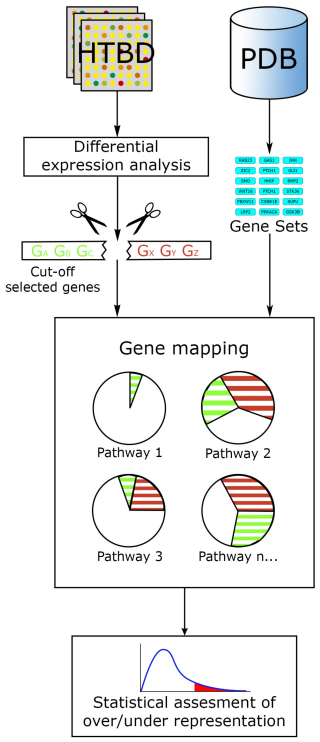


Figura 3.1: Esquema del Single Enrichment Analysis - Imagen tomada de [30]

Para determinar si el número de genes expresados diferencialmente en su conjunto de genes C_i es mayor de lo que se esperaría simplemente por azar es posible utilizar estadísticos de contraste basados en varias distribuciones clásicas de la estadística.

3.1.1. Test exacto de Fisher

Una solución para comprobar si existe asociación entre el conjunto de genes DE y el conjunto de genes anotados en la categoria i es utilizar el test exacto de Fisher asociado a la tabla de contingencia 2x2 representada en la Tabla 3.1. Este test se utiliza en la práctica cuando se quiere comprobar si existe dependencia entre dos variables cualitativas con dos niveles cada una de ellas.

Para poder utilizar este test es necesario que se cumplan ciertas condiciones:

- Las observaciones de la muestra deben ser independientes unas de otras y contribuir únicamente en una celda de la tabla
- Las frecuencias marginales deben ser fijas

El test de Fisher emplea lo que se conoce como un modelo de urna de forma que el conjunto de genes podría ser comparado con el número de N 'bolas' blancas y negras que se encuentran en una urna. El test consiste en hacer un muestreo sin reemplazamiento de m_i bolas (los genes pertenecientes a la vía biológica Ci) y determinar si la proporción de n bolas blancas (genes DE) entre las N totales es significativamente mayor de lo que se esperaría por azar [31].

En estas condiciones, la probabilidad de obtener exactamente k genes DE anotados en C_i será:

$$\Pr(X = k) = \frac{\binom{m_i}{k}\binom{N-m_i}{n-k}}{\binom{N}{n}}$$

donde

- $\binom{m_i}{k}$ es el número de formas de elegir k genes de entre los m_i genes anotados en C_i
- $\binom{N-m_i}{n-k}$ es el número de formas de elegir $n-k$ genes DE que no están anotados en C_i de entre los $N-m_i$ disponibles
- $\binom{N}{n}$ es el número de formas de elegir una muestra de n genes DE de la población total de N genes.

La probabilidad de observar k genes o más será:

$$\Pr(X \geq k) = \sum_{x=k}^{\min(m_i,n)} \frac{\binom{m_i}{x}\binom{N-m_i}{n-x}}{\binom{N}{n}}$$

el p-valor que cuantifica como es de probable haber observado k (o más) genes DE anotados en C_i asumiendo que no hay relación entre estar anotado en esa categoría y estar DE.

Uno de los principales problema de este test tiene que ver con que al aumentar el número de observaciones el número de posibles combinaciones en la tabla de contingencia crece de manera exponencial. Por esta razón cuando el número de observaciones es mayor puede ser más recomendable utilizar otra aproximación [32].

Otro aspecto a tener en cuenta es desde luego el tamaño del conjunto total de genes estudiados. La significación estadística no depende únicamente del número de genes DE ni, ni del número de genes en C_i, si no que también es dependiente de N, el número de genes totales que se evalúan. Por esta razón si el número de genes estudiado es extremadamente alto o bajo la significación estadística puede alejarse de la significación biológica.

3.1.2. Test χ^2

Como se ha mencionado con anterioridad, el test exacto de Fisher suele utilizarse cuando el número de observaciones es pequeño debido al coste computacional que implicaría un estudio de miles de genes. En estos casos, o en aquellos en los que se evaluen más de dos condiciones experimentales, será más conveniente utilizar en su lugar una aproximación como lo es el test χ^2 de Pearson.

A la hora de aplicar este test es necesario tener en cuenta de nuevo ciertas condiciones:

- Las observaciones de la muestra deben ser independientes unas de otras y contribuir únicamente en una celda de la tabla
- Las frecuencias esperadas en cada celda de la tabla de contingencia deben ser mayores o iguales a 5, de lo contrario la aproximación no será buena

El test χ^2 de Pearson parte de la tabla contingencia ilustrada en la Tabla 3.1 y se basa en las diferencias entre los valores observados y esperados.

$$\chi^2 = \sum_{r,c} \frac{(O_{rc} - E_{rc})^2}{E_{rc}}$$

donde:

- O_{rc} es el valor observado en la celda de la fila r y columna c.
- E_{rc} es el valor esperado en la celda de la fila r y columna c, que se calcula como el producto de las marginales.

En las condiciones anteriores, se puede asumir que este estadístico test seguirá una distribución chi-cuadrado con $(r-1)(c-1)$ grados de libertad [32].

3.2. Gene Set Enrichment Analysis (GSEA)

Como respuesta a los inconvenientes de SEA mencionados en la sección 2.4 surgieron a partir de las propuestas de Mootha en 2003 [33] y Subramanian et al. en 2005 [24] los métodos de 'Functional Class Scoring' (FCS) o 'Gene Set Enrichment Analysis' (GSEA).

Estos métodos se caracterizan por ejecutarse en tres pasos bien diferenciados [23].

Cálculo del éstadístico a nivel de gen

El primer paso consiste en el cálculo de un estadístico para cada gen utilizando los niveles de expresión medidos.

Este punto de partida es el mismo que en los métodos SEA, con la diferencia de que, en este caso en lugar de obtener una lista binaria que indique si el gen está sobrerrepresentado o no, se dispone de una medida cuantitativa del grado de asociación entre el gen y la condición evaluada. Esta medida permite ordenar la lista de genes y construir diferentes estadísticos. La elección de este estadístico es muy variada. Algunos de los recomendados por proporcionar los mejores resultados son los que proporcionan mejores resultados son el Test Moderado de Welch, la Mínima Diferencia Significativa, el *Signal To Noise Ratio* y el Test de BaumgartnerWeiss-Schindler [34].

Cálculo del estadístico a nivel de vía biológica

Una vez calculado el estadístico a nivel de gen se cuenta con una lista L de todos los genes del estudio ordenados por dicho valor. A partir de esta lista se calcula un estadístico a nivel de vía biológica, conocido como Enrichment Score (ES), que servirá para determinar como de sobrerepresentada está cierta vía en los extremos de la lista. El estadístico a elegir en este paso puede ser tan sencillo como una media o una mediana de los estadísticos a nivel de gen correspondientes a los genes que componen cierta vía biológica, o algo más complejos como el estadístico de Kolmogorov Smirnov, la suma de rangos de Wilcoxon, o el estadístico Maxmean [23].

El método más extendido y el que se implementa en paquetes como `clusterProfiler` es la versión GSEA de Subramanian et al.[24] que se basa en el estadístico de Kolmogorov-Smirnov.

Kolmogorov-Smirnov El estadístico de Kolmogorov-Smirnov de una muestra se emplea como técnica no paramétrica utilizada para evaluar si una muestra de datos sigue una distribución específica de tal forma que esta distribución de referencia puede ser continua o discreta, y no se hacen suposiciones sobre sus parámetros. Para realizar la comparación se obtiene primero la 'función de distribución empírica'. Dada una muestra $X_1, X_2, \ldots, X_n$ la función de distribución empírica $F_n(x)$ se define como

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq x)$$

donde n es el tamaño de la muestra, e $I(X_i \leq x)$ es la función idicador que toma el valor 1 si $X_i \leq x$ y 0 en caso contrario.

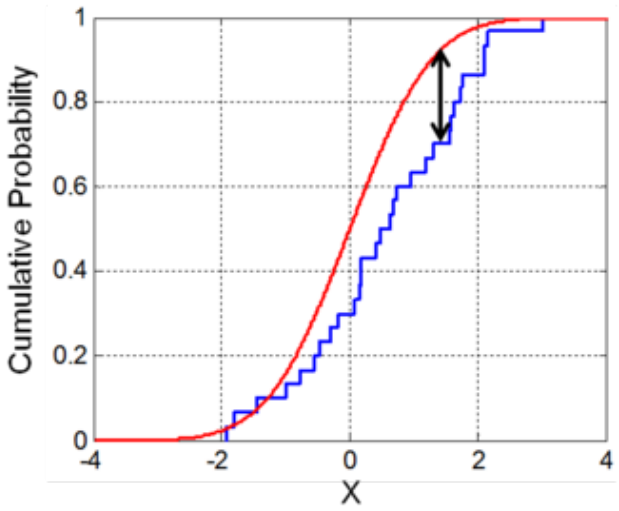


Figura 3.2: Representación gráfica del estadístico de Kolmogorov-Smirnov

El estadístico de Kolmogorov-Smirnov como la mayor desviación de la función de distribución empírica respecto a una función de distribución de referencia $F_0(x)$: $D = max(|F_n(x) - F_0(x)|)$ [35]. El estadístico puede entenderse fácilmente con la ilustración presentada en la Figura 3.2.

Enrichment Score Tomando N el número total de genes estudiados y L la lista ordenada de genes según el estadístico a nivel de gen, se define C_i como el conjunto de genes anotados en la vía biológica i-ésima, y N_{C_i} como el número de genes en C_i. Además se toma r_k como el estadístico a nivel de gen para cada gen g_k perteneciente a dicho conjunto.

Con esta notación es además posible definir la suma N_{R_i} de los estadísticos a nivel de gen de los genes en C_i, elevados a un parámetro de peso p.

$$N_{R_i} = \sum_{g_k \in C_i} |r_k|^p.$$

El ES se calcula en función de la diferencia entre las probabilidades acumuladas P_{hit} y P_{miss}, definidas como:

$$P_{\text{hit}}(C_i, j) = \sum_{\substack{g_k \in C_i \\ k \leq j}} \frac{|r_k|^p}{N_{R_i}}, \quad \text{donde } N_{R_i} = \sum_{g_k \in C_i} |r_k|^p,$$

$$P_{\text{miss}}(C_i, j) = \sum_{\substack{g_k \notin C_i \\ k \leq j}} \frac{1}{N - N_{C_i}},$$

donde j es una posición en la lista ordenada L.

Por tanto, el proceso consiste en recorrer la lista ordenada L construyendo las dos funciones de distribución que se comparan utilizando una modificación de la distancia de Kolmogorov-Smirnov. La función P_{hit} que incrementa su magnitud por cada gen g_k del conjunto C_i una cantidad que depende de la cantidad de asociación de ese gen con la condición estudiada, r_k. La función P_{miss} incrementa su magnitud por cada gen que no está en el conjunto C_i. [24].

En la versión original en la que se basa este algoritmo propuesta por Mootha [33] el estadístico empleado era simplemente el de Kolmogorov Smirnov por lo que el valor del incremento aplicado era siempre el mismo para todos los genes. Sin embargo esta aproximación contribuía a que fuera posible detectar como enriquecidos valores situados en zonas intermedias de la lista por lo que se cambió el incremento de tal forma que este dependiera de $|r_k|^p$, es decir el valor absoluto del estadístico a nivel de gen elevado a cierto parámetro p que permite variar entre el estadístico de Kolmogorov Smirnov (p=0) y esta propuesta de Subramanian et al. (típicamente p=1) [36]

Determinación de la significación estadística del ES

Una vez calculado el ES es natural preguntarse si su valor es lo suficientemente alto como para determinar que la vía biológica asociada a dicho conjunto de genes está enriquecida. Al ser Kolmogorov Smirnov un test no paramétrico la primera opción para establecer la significación estadística se trata de un test de permutaciones [37].

Para determinar la significación sería deseable conocer como se distribuirían los valores del ES. El test de permutaciones consiste en cambiar de posición de forma aleatoria, o bien las etiquetas que muestran el fenotipo o condición de cada muestra, o bien realizando una permutación de la etiqueta con los identificadores de genes como se indica en la documentación de implementaciones como la de `clusterProfiler` [8].

Poniendo el foco en este segundo enfoque el proceso consiste en tomar la lista ordenada L de genes con sus distintos estadísticos a nivel de gen y aplicar la permutación un número elevado de veces P (en torno a las 1000 [36]) calculando en cada una de ellas el ES tal y como se ha descrito anteriormente. Este proceso dará como resultado la distribución nula de este estadístico y con ella es posible determinar cómo de probable es haber observado por azar el valor observado del ES (ES_0), en otras palabras será posible establecer el p-valor (p_val) del enriquecimiento del término como:

$$p\_val = \frac{\sum_{b=1}^{P} I(ES_b \geq ES_0)}{P}$$

cuando $ES_0 \geq 0$

$$p\_val = \frac{\sum_{b=1}^{P} I(ES_b \leq ES_0)}{P}$$

cuando $ES_0 < 0$

Leading-Edge Subset

Dentro de cada conjunto de genes analizado no todos los miembros tienen por qué participar en el proceso biológico por lo que a menudo es útil extraer los miembros centrales de conjuntos de genes con altas puntuaciones que contribuyen al ES.

En GSEA este conjunto de genes son aquellos que aparecen en la lista ordenada L antes de alcanzarse el máximo del estadístico acumulativo descrito anteriormente. En la Figura 2.10 estos genes se corresponderían con los que aparecen antes del pico que muestra el gráfico.

Este subconjunto de genes pueden definirse como el núcleo que representa la señal de enriquecimiento y pueden ser útiles para agrupar conjuntos de genes que participan en los mismos procesos biológicos [24].

3.3. Corrección de comparaciones múltiples

Dado que el número de categorías a contrastar es bastante elevado, es necesario establecer algún tipo de corrección para controlar el problema de comparaciones múltiples, es decir la aparición de categorías falsamente enriquecidas simplemente por azar.

Control de FDR

La *False Discovery Rate* (FDR) es una medida utilizada para corregir el riesgo de obtener resultados falsos positivos en múltiples contrastes de hipótesis. Específicamente, la FDR es la proporción de resultados significativos que son falsos positivos entre todos los resultados significativos obtenidos, es decir es la proporción de descubrimientos que son incorrectos entre aquellos que se consideran "significativos".

En el contexto del enriquecimiento funcional, para cada vía biológica, se calcula un p-valor que sirve para evaluar la hipótesis de en la lista de genes de interés, la correspondiente vía biológica está sobrerepresentada. Entre los p-valores significativos, existirán falsos positivos, es decir vías biológicas consideradas sobrerepresentadas erróneamente. Uno de los métodos más comunes para controlar el FDR es el procedimiento de Benjamini-Hochberg (BH) [38]. Este método ajusta los p-valores para que la proporción de falsos positivos se mantenga bajo un umbral prefijado, conocido como q-valor. En concreto, el procedimiento de BH es como sigue:

1. Se ordenan los p-valores de todas las pruebas en orden ascendente. Suponiendo que hay M pruebas, $p_1, p_2, ..., p_M$ son los p-valores ordenados
2. Se asigna un rango a cada p-valor

3. Se calcula el p-valor ajustado para cada p-valor ordenado:

$$p_i^{adj} = \text{mín}(1, \min_{j \leq i}(\frac{M * p_j}{j}))$$

donde p_j es el p-valor original sin ajustar.

Estimación de FDR en GSEA

Para estimar el FDR de cada conjunto de genes en GSEA primero se normalizan los ES de cada conjunto de genes dividiéndolos por la media de los ES de las permutaciones para ese conjunto. Esto produce una puntuación de enriquecimiento normalizada, *Normalized Enrichment Score* (NES), que permite la comparación entre diferentes conjuntos de genes.

El FDR para cada conjunto de genes se calcula utilizando un test de permutaciones. Se crean muchas listas de genes permutadas para generar una distribución nula de NES y el FDR se calcula como la proporción de NES en la distribución nula que son al menos tan extrema como el NES observado, ajustada por la proporción de NES observados que tienen valores tan o más extremos.

Por lo tanto el procedimiento sería de la forma:

1. Para un conjunto de genes C_i, se define su puntuación de enriquecimiento observada NES_i.
2. Con todos los NES de todas las permutaciones de todos los conjuntos se obtiene una distribución nula NES_{nulo}
3. Se cuenta el número de permutaciones donde la puntuación de enriquecimiento de un conjunto de genes en la distribución nula es mayor o igual a NES_i.
4. Se ajusta este conteo por el número total de permutaciones y por la fracción de NES observados que son al menos tan extremos como NES_i.

$$\text{FDR}(i) = \frac{\sum I(NES_{nulo} \geq NES_i)}{\sum I(NES_{observados} \geq NES_i)}$$

Capítulo 4 - Resultados y discusión

En este capítulo se comparan los resultados obtenidos utilizando el test de Fisher y dos versiones del GSEA propuesto por Subramanian et al. [24]. La comparación se lleva a cabo en diferentes escenarios simulados, en los que se induce una desregularización controlada de cierto conjunto de genes, así como en un conjunto de datos de expresión en el que se tiene cierta información sobre qué vías biológicas deberían estar enriquecidas.

4.1. Métodos a comparar

Como se ha mencionado a lo largo de todo el documento los dos grandes enfoques de análisis de enriquecimiento funcional son tanto SEA como GSEA. Es por ello que se ha considerado conveniente utilizar como herramienta para llevar a cabo el enriquecimiento `ClusterProfiler`, que da soporte a ambos.

En representación de los métodos SEA se empleará el método exacto de Fisher (al cual se hará referencia como SEA de aquí en adelante), implementado en la función `enricher()` de `ClusterProfiler`, y como métodos GSEA se aplicará el propuesto por Subramanian et al. [24] con dos valores del parámetro p: 1 y 1.5 (GSEA_P1 y GSEA_P1.5) para comprobar si existen diferencias notables al modificar este parámetro, aunque en ciertos apartados la comparación se limitará a SEA y GSEA_P1.

4.2. Datos simulados

Se simulan siete escenarios, todos ellos basados en unos niveles de expresión base, generados a partir de la distribución normal estándar. Siguiendo el esquema planteado en [39], se generan 1000 genes y 50 muestras, 25 de ellas control y las otras 25 tratamiento. Además, se consideran 50 conjuntos de genes, con 20 genes cada uno de ellos. Las diferencias entre los diferentes escenarios de simulación se refieren a la cantidad de expresión diferencial entre grupos de

tratamiento y el número de genes diferencialmente expresados dentro de un mismo conjunto de genes, tal y como se describe a continuación:

- Escenario 1: Con el objetivo de evaluar la capacidad de detección de los métodos SEA y GSEA frente a un cambio significativo en todos los genes del conjunto, se aumenta 0.5 unidades en la clase tratamiento de los 20 genes del primer conjunto considerado.
- Escenario 2: Para evaluar la sensibilidad de los métodos en la detección de cambios sutiles en todos los genes del conjunto, se aumenta 0.2 unidades en la clase tratamiento de los 20 genes del primer conjunto.
- Escenario 3: Con el propósito de evaluar la capacidad de SEA y GSEA para detectar enriquecimientos parciales cuando solo una parte del conjunto de genes está afectada, se aumenta 0.4 unidades en los primeros 10 genes y 0.2 unidades en los siguientes 10 genes del primer conjunto considerado en la clase tratamiento.
- Escenario 4: Para comparar la sensibilidad de los métodos utilizados frente a cambios opuestos dentro del mismo conjunto de genes, se aumenta 0.5 unidades en los primeros 10 genes y se disminuye 0.5 unidades en los siguientes 10 genes en la clase tratamiento del primer conjunto considerado.
- Escenario 5: Con el objetivo de evaluar cómo SEA y GSEA manejan efectos opuestos dentro del mismo conjunto de genes, lo que podría complicar la interpretación del enriquecimiento, se aumenta 0.6 unidades en los primeros 5 genes en la clase tratamiento del primer conjunto.
- Escenario 6: Para comparar la capacidad de SEA y GSEA para detectar enriquecimientos cuando solo una fracción de los genes del conjunto está afectada, se aumenta 0.4 unidades en 10 genes aleatorios en la clase tratamiento del primer conjunto considerado.
- Escenario 7: Con la intención de evaluar la robustez de los métodos empleados frente a cambios aleatorios y heterogéneos, se aumenta aleatoriamente con una media de 0.5 unidades en los primeros 10 genes y 0.3 unidades en los siguientes 10 genes en la clase tratamiento del primer conjunto.

En todos los escenarios, el conjunto de interés es el conjunto 1, siendo este el único conjunto que los métodos de enriquecimiento funcional deberían identificar como significativamente enriquecido.

Para comprender la diferencia en la distribución de los valores de expresión de las clases control y tratamiento del primer conjunto de genes con respecto al resto se han creado los gráficos de densidad que se muestran el la Figura 4.1 en la que se puede observar como los escenarios 1 y 7 son los que presentan diferencias de distribución más claras mientras que en el resto de casos los cambios son más sutiles.

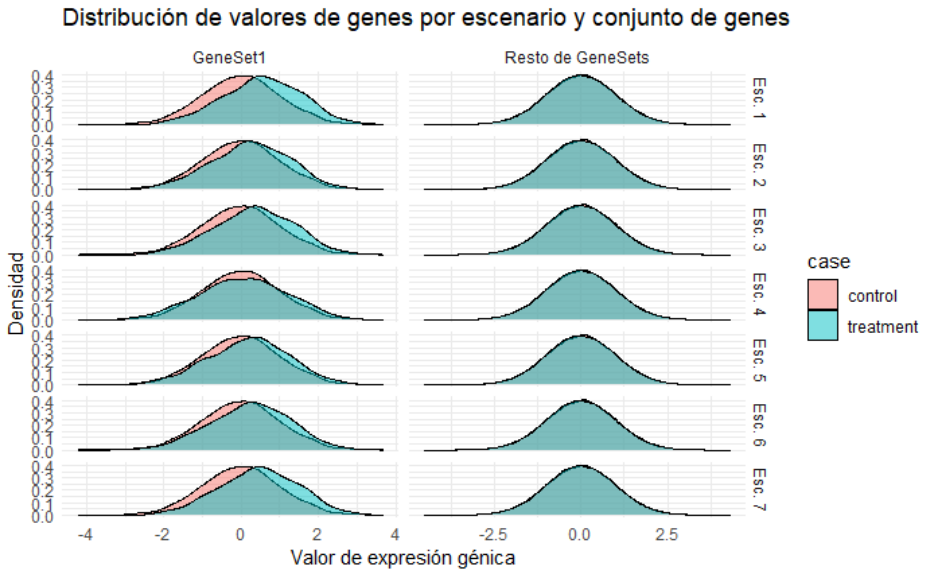


Figura 4.1: Comparativa entre el conjunto de genes 1 y los 49 conjuntos restantes con respecto a la diferencia de distribución de los valores de expresión entre control y tratamiento

4.2.1. Análisis de expresión diferencial

Para llevar a cabo el análisis de expresión diferencial se utilizó el paquete limma disponible en Bionductor [26] que puede consultarse en el Anexo A.

Se identificaron los genes DE filtrando los genes basándose en los criterios de p-valor (p-valor ¡0.05) y en la magnitud del cambio en el logFC de tal forma que se estableció como valor límite 0.8 (logFC 0.8) para todos los escenarios, lo que significa que se considera una expresión diferencial clínicalmente relevante valores de 1.74 o má veces en el tratamiento que en el control.

4.2.2. Análisis de enriquecimiento funcional

Para el análisis de enriquecimiento mediante el método SEA, se utilizó la función `enricher()` del paquete `ClusterProfiler` con la lista de genes DE y para el análisis de enriquecimiento mediante GSEA se preparó un *ranking* de genes basado en el logFC con todos los genes. Este *ranking* se utilizó como entrada para la función `GSEA()` de `ClusterProfiler`.

Los resultados obtenidos se resumen en la Tabla 4.1 y la Figura 4.2 donde se pueden apreciar los diferentes -log10(p-valores) que los tres métodos han obtenido al realizar el análisis de enriquecimiento sobre el primer conjunto, es decir, el único conjunto que debería detectarse como enriquecido. A continuación, se discuten los resultados en los diferentes escenarios.

- **Escenario 1:** En este escenario, donde el cambio en todos los genes del conjunto fue pronunciado, todos los métodos detectaron el enriquecimiento con p-valores extremadamente bajos. Esto indica una alta sensibilidad de todos los métodos ante cambios significativos en todos los genes del conjunto.
- **Escenario 2:** Aquí, con un aumento más sutil pero igualmente uniforme en todos los genes del conjunto de genes 1 en la clase tratamiento, SEA no logró detectar enriquecimiento significativo, mientras que tanto GSEA_P1 como GSEA_P1.5 sí lo hicieron. Esto sugiere que los métodos GSEA son más sensibles para detectar cambios más pequeños en los conjuntos de genes.
- **Escenario 3:** En este escenario, con un aumento mixto de 0.4 y 0.2 en la clase tratamiento del conjunto de genes 1, SEA detectó enriquecimiento significativo, al igual que GSEA_P1 y GSEA_P1.5. Todos los métodos mostraron capacidad para detectar enriquecimientos parciales aunque los p-valores en GSEA son claramente menores.
- **Escenario 4:** Para este escenario, donde hubo efectos opuestos dentro del mismo conjunto de genes, SEA detectó enriquecimiento significativo, mientras que GSEA_P1 y GSEA_P1.5 también lo hicieron. Esto indica que todos los métodos fueron sensibles a los cambios opuestos dentro del mismo conjunto de genes.
- **Escenario 5:** En este caso, con un aumento de 0.6 en el tratamiento solo para un cuarto de los genes del conjunto 1, SEA detectó enriquecimiento mientras que GSEA no lo hizo. Esto sugiere que GSEA ofrece un mejor funcionamiento con cambios que afecten a la mayoría de los genes del conjunto mientras que SEA se beneficia de que estas cambios aunque fueran en un pequeño número de genes fueran muy pronunciados y se representaran en la lista de genes DE.
- **Escenario 6:** Aquí, con un aumento de 0.4 unidades en 10 genes aleatorios del conjunto de genes 1 en la clase tratamiento, SEA no detectó enriquecimiento significativo (p-valor = NA), mientras que GSEA_P1 y GSEA_P1.5 sí lo hicieron. Esto indica que los métodos GSEA pueden ser más efectivos para detectar enriquecimientos cuando hay mayor equilibrio entre el número de genes enriquecidos y la intensidad de la desregulación de estos.
- **Escenario 7:** En este último escenario, con un aumento aleatorio con media 0.5 para los primeros 10 genes y media 0.3 para los siguientes 10 genes del conjunto de genes 1 en la clase tratamiento, todos los métodos detectaron enriquecimiento significativos. Esto sugiere que los métodos son robustos frente a cambios aleatorios y heterogéneos en el conjunto de genes.

En resumen, los resultados indican que los métodos GSEA_P1 y GSEA_P1.5 son más sensibles y efectivos que SEA en la mayoría de los escenarios evaluados, especialmente cuando

los cambios son sutiles y el criterio del FC limita la lista de genes DE, aunque sorprendentemente al producirse cambios intensos pero focalizados en unos pocos genes del conjunto SEA funciona mejor.

Escenario	**SEA_log_pvalue**	**GSEA_p1_log_pvalue**	**GSEA_p1.5_log_pvalue**
Escenario 1	8.726026	8.301030	7.393444
Escenario 2	NA	1.441223	2.015649
Escenario 3	1.698970	3.386043	3.867662
Escenario 4	6.931886	8.301030	7.160837
Escenario 5	1.698970	NA	NA
Escenario 6	NA	1.878503	1.531750
Escenario 7	1.698970	4.727566	2.997633

Tabla 4.1: Tabla con los -log10(p-valores) obtenidos en el análisis de enriquecimiento funcional para el primer conjunto ordenados por método y escenario. Los valores NA se corresponden con p-valores superiores al umbral de significación estadística fijado en 0.05

Por otra parte es importante mencionar que se puede asegurar la especificidad de los métodos en todos los escenarios puesto que ninguno detectó como enriquecido ningún conjunto diferente al primero.

4.3. Conjunto P53

Este conjunto de datos incluye niveles de expresión de 8655 genes en 33 muestras con una mutación en el gen P53 y 17 muestras sin dicha mutación, que se han obtenido utilizando microarrays Affymetrix U95. La matriz de expresión está disponible en el paquete GSAR de Bioconductor [40]. Se espera encontrar un enriquecimiento en los conjuntos de genes relacionados con las vías asociadas a las mutaciones en P53. El gen P53 es un supresor de tumores que juega un papel crucial en el proceso de señalización de la apoptosis (muerte celular programada). En particular, la proteína P53 actúa como un factor de transcripción que, en condiciones normales, inhibe el crecimiento celular y estimula la muerte celular cuando se detecta estrés celular.

4.3.1. Análisis de expresión diferencial

Antes de hacer el enriquecimiento funcional se llevó a cabo el análisis de expresión diferencial que se puede consultar en el Anexo A.

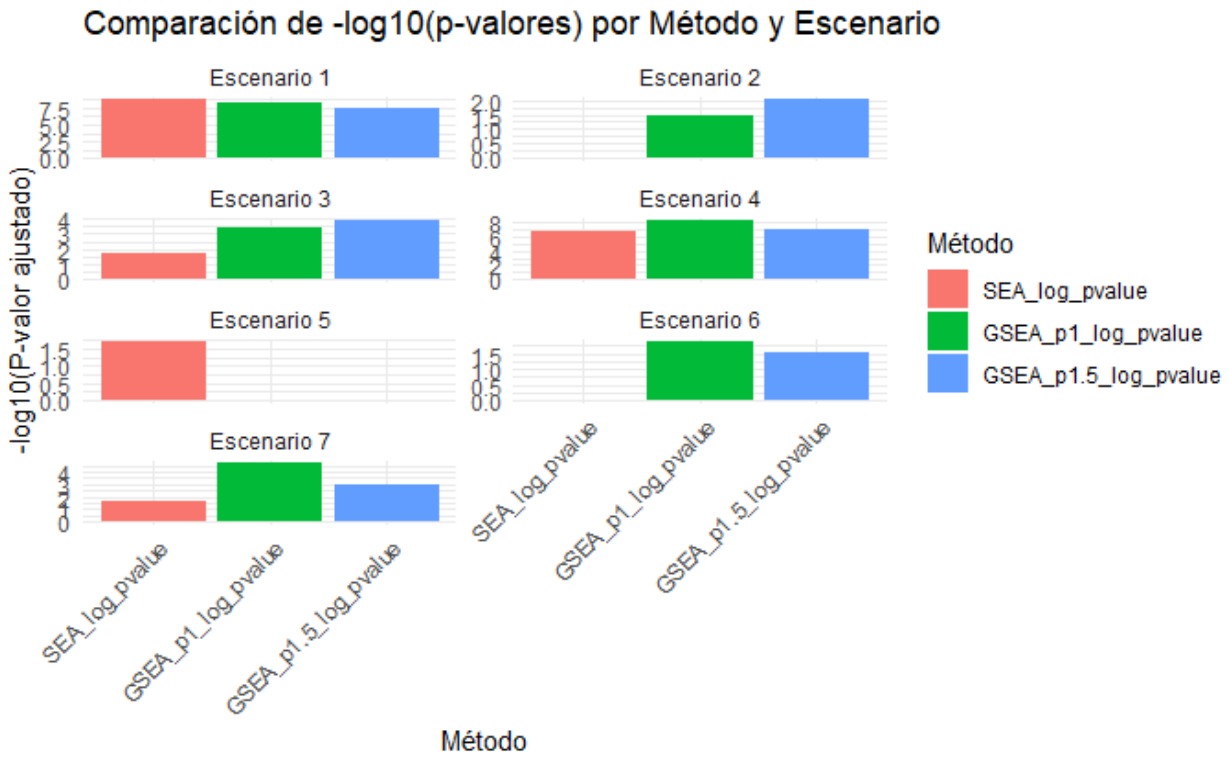


Figura 4.2: -log10(p-valores) obtenidos por cada método en el análisis de enriquecimiento del primer conjunto de genes para los 7 escenarios.

Se utilizó el paquete limma disponible en Bioconductor [26]. El análisis consiste en el ajute de un modelo lineal gen a gen y, a partir de dicho ajuste, utilizar el estadístico *empirical Bayes moderated t-test* [41] para contrastar la igualdad de expresión entre las dos clases de muestras: con y sin mutación en el gen P53. Posteriormente, se identificaron los genes DE entre las condiciones de tratamiento y control. Este paso se llevó a cabo filtrando los genes basándose en el criterio del logFC de tal forma que se estableció como valor límite 1 ($logFC > 1$), lo que significa que hay una expresión diferencial del doble o más en el grupo de muestras con mutación.

4.3.2. Análisis de enriquecimiento funcional

Para el análisis de enriquecimiento mediante el método SEA, se utilizó la función `enricher()` del paquete `clusterProfiler` con la lista de genes DE.

Por otro lado, para el análisis de enriquecimiento mediante GSEA se preparó un *ranking* de genes basado en el logFC de los genes DE. Este *ranking* se utilizó como entrada para la función `GSEA()` de `ClusterProfiler`.

Es necesario añadir que como base de datos de anotación para llevar a cabo el análisis GSEA se utilizó MSigDB [19] y en concreto las firmas genéticas anotadas para la especie

'Homo sapiens' de la categoría C6 correspondiente a conjuntos de genes asociados con firmas oncogénicas.

Comparacion de métodos de enriquecimiento con MSigBD Los resultados obtenidos al realizar el análisis de enriquecimiento funcional empleando la base de datos MSigDB se resumen en la Tabla 4.2 donde se pueden apreciar los diferentes p-valores obtenidos con los tres métodos y la Figura 4.3 que muestra el solapamiento entre los conjuntos de genes detectados por los tres métodos.

Como se hace evidente al observar los resultados los tres métodos son capaces de detectar los conjuntos de genes relacionados con el gen P53, en concreto estos conjuntos son P53_DN.V1_DN y P53_DN.V1_UP.

Sin embargo la diferencia notable entre los métodos GSEA y SEA tiene relación con la especificidad de los resultados puesto que SEA identifica como enriquecidos 10 conjuntos adicionales que tienen relación con otros oncogenes.

Se comprueba por tanto que los resultados obtenidos con GSEA contienen menos ruido y gozan de una mayor especificidad que puede ser importante en muchos aplicaciones de estos métodos. Por otra parte se observa como los resultados de GSEA en este caso no varían al cambiar el parámetro p.

Term	**SEA_pvalue**	**GSEA_p1_pvalue**	**GSEA_p1.5_pvalue**
P53_DN.V1_DN	9.234830e-51	1e-10	1e-10
P53_DN.V1_UP	6.873306e-37	1e-10	1e-10
LEF1_UP.V1_UP	6.673369e-04	NA	NA
BCAT.100_UP.V1_UP	1.109631e-03	NA	NA
RPS14_DN.V1_UP	2.284235e-03	NA	NA
RAF_UP.V1_DN	2.284235e-03	NA	NA
BMI1_DN.V1_UP	2.317701e-03	NA	NA
E2F1_UP.V1_DN	8.872327e-03	NA	NA
MEL18_DN.V1_UP	1.035082e-02	NA	NA
PGF_UP.V1_DN	3.765311e-02	NA	NA
KRAS.DF.V1_UP	3.765311e-02	NA	NA
HOXA9_DN.V1_UP	3.765311e-02	NA	NA

Tabla 4.2: Tabla con el p-valor ajustado de los conjuntos detectados como enriquecidos por los tres métodos al aplicarlos sobre los datos del conjunto P53. Los valores NA se corresponden con p-valores superiores al umbral de significación estadística fijado en 0.05.

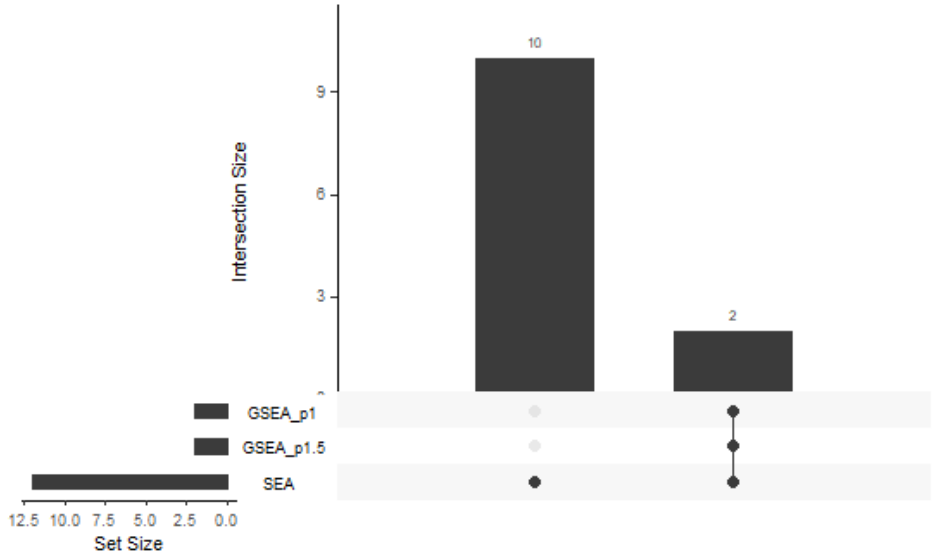


Figura 4.3: Solapamiento de los conjuntos de genes detectados como enriquecidos por los tres métodos en el caso del conjunto P53

Análisis de robustez y especificidad Para evaluar la robustez y especificidad de GSEA y SEA se ha planteado un experimento en el que se ha seleccionado de forma aleatoria 1000 genes diferencialmente expresados (DE) de entre los 8655 incluidos en el conjunto P53. Se ha llevado a cabo el enriquecimiento funcional utilizando tanto SEA como GSEA en 500 selecciones distintas. Para cada una de ellas, se ha evaluado la presencia o no de los conjuntos relacionados con P53 entre los conjuntos enriquecidos significativamente, así como el número de conjuntos distintos a ellos que también son finalmente enriquecidos.

Los resultados obtenidos pueden observarse en la Tabla 4.3 en la que se muestra el número de veces que cada método ha detectado cada conjunto. Estos datos muestran como SEA no es capaz de detectar los conjuntos de interés de forma sistemática a lo largo de las 500 simulaciones y que por el contrario identifica como significativos muchos otros conjuntos, mientras que GSEA muestra una gran consistencia a la hora de detectar los conjuntos de interés y manteniendo una gran especificidad.

Estos datos evidencia la dependencia de los métodos SEA del punto de corte con el que se elige la lista de genes DE. En contraposición, GSEA se muestra muy robusto ante los cambios de genes en la lista lo que supone una ventaja enorme que permite aplicar el método en escenarios en los que no es posible lograr una gran precisión al definir los genes de interés.

Term	**Method**	**Count**
P53_DN.V1_DN	GSEA	499
P53_DN.V1_DN	SEA	104
P53_DN.V1_UP	GSEA	498
P53_DN.V1_UP	SEA	93
otros	GSEA	777
otros	SEA	92685

Tabla 4.3: Número de veces que cada método ha detectado diferentes conjuntos a lo largo de 500 simulaciones de genes DE aleatorios.

Capítulo 5 - Conclusiones

En este trabajo se ha explorado dos aproximaciones diferentes para llevar a cabo lo que se conoce en el entorno de la bioinformática como enriquecimiento funcional. Este tipo de análisis consiste en identificar y evaluar la representación de categorías funcionales o rutas biológicas dentro de un conjunto de genes, comparado con lo que se esperaría por azar, con el objetivo principal de descubrir funciones biológicas, procesos celulares y rutas metabólicas que estén sobrerrepresentadas en el conjunto de datos analizados, lo que puede proporcionar información valiosa sobre los mecanismos biológicos subyacentes. La primera aproximación explorada ha sido el método de sobrerrepresentación (SEA), que se basa en contrastar la frecuencia de genes de una categoría funcional específica en el conjunto de datos de interés con la frecuencia de esos mismos genes en un conjunto de referencia. La segunda aproximación estudiada ha sido el análisis de conjuntos de genes (GSEA), que clasifica todos los genes del conjunto de datos de interés según una medida de expresión o de asociación y luego evalúa si los genes pertenecientes a una categoría funcional específica tienden a aparecer en las partes más extremas de la lista ordenada.

Se ha evaluado su funcionamiento tanto en datos reales como simulados, en los que se tiene cierto conocimiento respecto a los conjuntos que deberían detectarse como enriquecidos. Los resultados han sido bastante favorables para el método GSEA, algo que cabía esperar, puesto que este tipo de metodología proporciona una visión más completa incorporando una medida cuantitativa de la asociación y no solo una lista de genes de interés. Esto permite detectar enriquecimientos sutiles y distribuidos a lo largo de toda la lista de genes, lo que puede ser especialmente útil en estudios donde los cambios en la expresión génica son más graduales y no necesariamente dicotómicos. Además, GSEA puede ser más robusto frente a variaciones en los datos y errores de medición, dado que utiliza un enfoque basado en la clasificación de genes en lugar de depender exclusivamente de umbrales arbitrarios de significación estadística. Se ha podido comprobar que GSEA es, por lo general, más robusto y específico, que obtiene p-valores más bajos para los conjuntos que deberían estar enriquecidos y que es capaz de detectar cambios uniformes pero muy sutiles que a SEA se le escapan por su fuerte dependencia del análisis de expresión diferencial. Aunque sin duda sorprende el

caso del escenario 5 de los datos simulados, en el que un aumento importante del nivel de expresión en el tratamiento, pero solamente en un cuarto de los genes, fue detectado por SEA y no por GSEA.

Un parámetro a fijar en el método GSEA es el peso de la medida de asociación, que se ha denotado en este trabajo como p. Se han evaluado dos valores de dicho parámetros y se ha podido comprobar que, por lo menos bajo la condición de igualdad de tamaño de los conjuntos de genes a detectar, no ofrece resultados marcadamente diferentes, al menos con los valores evaluados en este trabajo.

Aunque en este trabajo la elección del parámetro p no se ha mostrado crítica, en la práctica debería ser considerada cuidadosamente en función de la naturaleza de los datos y los objetivos específicos del análisis. Cabe esperar que, cuando p tome un valor mayor, el análisis se focalice más en los genes situados en los extremos de la lista ordenada, dando mayor peso a aquellos con asociaciones más fuertes. Es posible que no se hayan considerado valores de p suficientemente extremos en este estudio, lo que constituye una posible limitación del mismo. En estudios futuros, evaluar una gama más amplia de valores para este parámetro podría proporcionar una comprensión más profunda de su influencia en los resultados del análisis de enriquecimiento funcional. Ajustar p de manera adecuada puede mejorar la sensibilidad del análisis para detectar señales biológicas relevantes, especialmente en situaciones donde se espera que los cambios en la expresión génica sean más pronunciados en un subconjunto de genes

En cualquier caso parece lógico ante los resultados obtenidos la recomendación de la elección de GSEA como método de enriquecimiento funcional frente a la alternativa que ofrece SEA.

Por último destacar que, durante la realización de este trabajo se ha podido comprobar que llevar a cabo análisis de datos con información bioinformática presenta desafíos significativos debido a la diversidad y heterogeneidad de las fuentes de conocimiento e información involucradas. La bioinformática integra información de múltiples disciplinas como la genética, la biología molecular, la bioquímica, la estadística y la informática, cada una con sus propias metodologías, formatos y terminologías. Esta multiplicidad de fuentes de conocimiento dificulta la unificación y el procesamiento de los datos, ya que a menudo carecen de estandarización. En el contexto concreto de este trabajo, la falta de homogeneidad se manifiesta en dos aspectos fundamentalmente: en primer lugar en la variedad de bases de datos, y en segundo las diferentes herramientas de análisis de enriquecimiento disponibles.

También es importante destacar que la metodología utilizada, tanto en GSEA como en SEA, se basa en métodos clásicos de la estadística como el test de Fisher y el test de Kolmogorov-Smirnov, que proporcionan una base sólida para evaluar la relevancia biológica de las categorías funcionales enriquecidas. La aplicación de estas pruebas clásicas en el contexto de la bioinformática pone de manifiesto la utilidad y la adaptabilidad de las herramientas es-

tadísticas tradicionales para abordar problemas complejos en la biología moderna e inspira su aplicación en investigaciones biomédicas reales.

Acrónimos

- BP: Biological Process
- CC: Cellular Component
- DE: Diferencialmente Expresado
- DNA: Deoxyribonucleic Acid
- ES: Enrichment Score
- FC: Fold-Change
- FCS: Functional Class Scoring
- FDR: False Discovery Rate
- GO: The Gene Ontology
- GSEA: Gene Set Enrichment Analysis
- GSEA_P1: Gene Set Enrichment Analysis con parámetro de peso p igual a 1
- GSEA_P1.5: Gene Set Enrichment Analysis con parámetro de peso p igual a 1.5
- KEGG: Kyoto Encyclopedia of Genes and Genomes
- MEA: Modular Enrichment Analysis
- MF: Molecular Function
- MSigDB: Molecular Signatures Database
- NES: Normalized Enrichment Score
- NGS: Next Genereation Sequencing
- ORA: Over Representation Analysis
- PT: Pathway Topology

- RNA: Ribonucleic Acid
- SEA: Singular Enrichment Analysis

Bibliografía

[1] Alex Emmons. «Pathways and Gene Sets: What is Functional Enrichment Analysis?» En: *Bioinformatics Training and Education Program (BTEP)* (2024). Accessed: 2024-07-15. URL: https://bioinformatics.ccr.cancer.gov/btep/pathways-and-gene-sets-what-is-functional-enrichment-analysis/.

[2] Konrad J. Karczewski et al. «The mutational constraint spectrum quantified from variation in 141,456 humans». En: *Nature* 581.7809 (2022). Accessed: 2024-07-15, págs. 434-443. DOI: 10.1038/s41586-020-2308-7.

[3] URL: https://geneontology.org/docs/ontology-documentation/.

[4] KEGG: Kyoto Encyclopedia of Genes and Genomes. *KEGG: Kyoto Encyclopedia of Genes and Genomes*. Accessed: 2024-06-16. 2024. URL: https://www.genome.jp/kegg/.

[5] Marc Gillespie et al. «The reactome pathway knowledgebase 2022». En: *Nucleic acids research* 50.D1 (2022), págs. D687-D692.

[6] Da Wei Huang, Brad T Sherman y Richard A Lempicki. «Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources». En: *Nature Protocols* 4.1 (2009), págs. 44-57. ISSN: 1750-2799. DOI: 10.1038/nprot.2008.211.

[7] Broad Institute. «GSEA User Guide». En: *GSEA User Guide* (). Accessed: 2024-07-1. URL: https://www.gsea-msigdb.org/gsea/doc/GSEAUserGuideFrame.html.

[8] Tianzhi Wu et al. «clusterProfiler 4.0: A universal enrichment tool for interpreting omics data». En: *The innovation* 2.3 (2021).

[9] Xiaofeng Dai y Li Shen. «Advances and Trends in Omics Technology Development». En: *Frontiers in Medicine* 9 (2022). ISSN: 2296-858X. DOI: 10.3389/fmed.2022.911861.

[10] Mario Vailati-Riboni, Valentino Palombo y Juan J. Loor. «What Are Omics Sciences?» En: *Periparturient Diseases of Dairy Cows: A Systems Biology Approach*. Ed. por Burim N. Ametaj. Cham: Springer International Publishing, 2017, págs. 1-7. ISBN: 978-3-319-43033-1. DOI: 10.1007/978-3-319-43033-1_1.

[11] Maria Eugenia Frigolet Vázquez Vela y Ruth Gutiérrez-Aguilar. «Ciencias "ómicas",¿ cómo ayudan a las ciencias de la salud?» En: *Revista Digital Universitaria* 18.7 (2017).

[12] Kangmei Zhao y Seung Yon Rhee. «Interpreting omics data with pathway enrichment analysis». En: *Trends in Genetics* 39.4 (2023), págs. 308-319. ISSN: 0168-9525. DOI: https://doi.org/10.1016/j.tig.2023.01.003.

[13] URL: https://www.genome.gov/es/about-genomics/fact-sheets/Vias-Biologicas.

[14] The Gene Ontology Consortium et al. «The Gene Ontology knowledgebase in 2023». En: *Genetics* 224.1 (mar. de 2023), iyad031. ISSN: 1943-2631. DOI: 10.1093/genetics/iyad031.

[15] Minoru Kanehisa et al. «KEGG as a reference resource for gene and protein annotation». En: *Nucleic acids research* 44.D1 (2016), págs. D457-D462.

[16] M. Kanehisa y S. Goto. «KEGG: Kyoto Encyclopedia of Genes and Genomes». En: *Nucleic Acids Research* 28.1 (ene. de 2000), págs. 27-30. ISSN: 0305-1048. DOI: 10.1093/nar/28.1.27.

[17] Hiroyuki Ogata et al. «KEGG: Kyoto Encyclopedia of Genes and Genomes». En: *Nucleic Acids Research* 27.1 (ene. de 1999), págs. 29-34. ISSN: 0305-1048. DOI: 10.1093/nar/27.1.29.

[18] Anthony Castanza et al. «The Molecular Signatures Database Revisited: Extending Support for Mouse Data». En: (oct. de 2022). DOI: 10.1101/2022.10.24.513539.

[19] URL: https://www.gsea-msigdb.org/gsea/msigdb.

[20] Paul D. Thomas et al. «PANTHER: Making genome-scale phylogenetics accessible to all». En: *Protein Science* 31.1 (2022), págs. 8-22. DOI: https://doi.org/10.1002/pro.4218.

[21] Waseem Jawaid. *enrichR: Provides an R Interface to 'Enrichr'*. Ver. 3.2. 22 de mar. de 2023. URL: https://CRAN.R-project.org/package=enrichR.

[22] Da Wei Huang, Brad T Sherman y Richard A Lempicki. «Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists». En: *Nucleic acids research* 37.1 (2009), págs. 1-13.

[23] Purvesh Khatri, Marina Sirota y Atul J Butte. «Ten years of pathway analysis: current approaches and outstanding challenges». En: *PLoS computational biology* 8.2 (2012), e1002375.

[24] Aravind Subramanian et al. «Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles». En: *Proceedings of the National Academy of Sciences* 102.43 (2005), págs. 15545-15550.

[25] Maysson Al-Haj Ibrahim et al. «A topology-based score for pathway enrichment». En: *Journal of Computational Biology* 19.5 (2012), págs. 563-573.

[26] Robert C. Gentleman et al. «Bioconductor: Open software for bioinformatics». En: *Genome Biology* 5.1 (2004), R80. URL: https://www.bioconductor.org/packages/release/bioc/html/biodbNcbi.html.

[27] Guangchuang Yu y Qing-Yu He. «ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization». En: *Molecular BioSystems* 12.2 (2016), págs. 477-479.

[28] Guangchuang Yu. *Introduction — Biomedical Knowledge Mining using GOSemSim and clusterProfiler — yulab-smu.top*. https://yulab-smu.top/biomedical-knowledge-mining-book/index.html. [Accessed 09-06-2024].

[29] G. Yu. *enrichplot: Visualization of Functional Enrichment Result*. Ver. 1.24.0. R package. 2024. URL: https://yulab-smu.top/biomedical-knowledge-mining-book/.

[30] Miguel A García-Campos, Jesús Espinal-Enríquez y Enrique Hernández-Lemus. «Pathway analysis: state of the art». En: *Frontiers in physiology* 6 (2015), págs. 1-2.

[31] Robert Castelo M.C. Erick Cuevas Fernández. *Mini curso junio 2021: Análisis de enriquecimiento funcional de conjuntos de genes en R*. https://github.com/comunidadbioinfo/minicurso_junio_2021. 01-04-2024. 2021.

[32] Jun. de 2024. URL: https://en.wikipedia.org/wiki/Fisher%27s_exact_test.

[33] Vamsi K Mootha et al. «PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes». En: *Nature genetics* 34.3 (2003), págs. 267-273.

[34] Joanna Zyla et al. «Ranking metrics in gene set enrichment analysis: do they matter?» En: *BMC Bioinformatics* 18.1 (2017), pág. 256. ISSN: 1471-2105. DOI: 10.1186/s12859-017-1674-0.

[35] Frank J Massey Jr. «The Kolmogorov-Smirnov test for goodness of fit». En: *Journal of the American statistical Association* 46.253 (1951), págs. 68-78.

[36] Jing Shi y Michael G Walker. «Gene set enrichment analysis (GSEA) for interpreting gene expression profiles». En: *Current Bioinformatics* 2.2 (2007), págs. 133-137.

[37] P Good. «A practical guide to resampling methods for testing hypotheses». En: *(No Title)* (1994).

[38] Jüri Reimand et al. «Pathway enrichment analysis and visualization of omics data using g: Profiler, GSEA, Cytoscape and EnrichmentMap». En: *Nature protocols* 14.2 (2019), págs. 482-517.

[39] Luca Abatangelo et al. «Comparative study of gene set enrichment methods». En: *BMC Bioinformatics* 10.1 (sep. de 2009), pág. 275. ISSN: 1471-2105. DOI: 10.1186/1471-2105-10-275. URL: https://doi.org/10.1186/1471-2105-10-275.

[40] Yasir Rahmatallah et al. «GSAR: Bioconductor package for gene set analysis in R». En: *BMC Bioinformatics* 18 (2017), pág. 61.

[41] G. K. Smyth. «Linear models and empirical Bayes methods for assessing differential expression in microarray experiments». En: *Statistical Applications in Genetics and Molecular Biology* 3.1 (2004), Article 3. URL: `http://www.statsci.org/smyth/pubs/ebayes.pdf`.