Análisis e integración de información de datos biológicos mediante análisis funcional

por

Juan Cruz Rodriguez

Presentado ante la Facultad de Matemática, Astronomía, Física y Computación

para obtener el grado de

Doctor en Ciencias de la Computación

de la

UNIVERSIDAD NACIONAL DE CÓRDOBA

Septiembre 2019

FAMAF - UNC

Director: Elmer Andrés Fernández

Va Castro, va Castro, va Castro, sacó el centro pasado, Riggio, ...

Aquella felicidad vuelve a atravesar mi cuerpo al defender esta tesis.

Agradecimientos

Llega el tan ansiado día de defender mi tesis doctoral, luego de más de 5 años de investigación sin duda hay miles de agradecimientos que mencionar. Antes que nada, me siento bendecido y afortunado de haber nacido en la República Argentina. Le agradezco a mi país no solo por haberme brindado educación gratuita de excelencia, si no que además la beca con la cual pude preocuparme exclusivamente en mi investigación durante estos años de estudios doctorales.

En segundo lugar, creo que el mayor agradecimiento se lo debo al Dr. Elmer A. Fernández, mi director, quien me inició, guió y formó en el universo de la investigación científica. Con todas las concordancias y discusiones que haya pasado con él, todo aprendizaje resulta positivo para la vida. Asimismo agradezco a mis compañeros de trabajo de la UCC, sus comentarios y sugerencias fueron un pilar fundamental para la calidad de este trabajo de tesis -sin contar el gran aporte de amistad y cerveceadas-.

Obviamente, este trabajo no hubiera sido posible sin momentos de esparcimiento de quien escribe. Aquí es donde debo principalmente agradecer a toda mi familia: má, pá, sin saber mucho de ciencia o investigación, siempre me apoyaron en lo que me hiciera feliz. A mi abuelito, que cada vez que me preguntaba sobre la facu, me daba ánimos a seguir más y más. Por otra parte mis grupos de amigos, los del barrio, los de la facu, los de la cancha, grupos bien distintos, pero que siempre brindaron toneladas de cariño ~~y asados y escabio~~. Y si a esparcimiento me refiero, no puedo no incluir al Instituto Atlético Central Córdoba, fuente de gran parte de mis alegrías.

Finalmente, imposible cerrar esta sección sin agradecer al Whisky, no, no soy

alcohólico, me refiero a mi amigo de cuatro patas. El ser más cariñoso que conozco, el que mayor parte de tiempo estuvo presente durante mi doctorado -ahora mismo está acá mirándome con cara de ir a pasear-. Esas pausas de 15 minutos que me exigía, fueron fundamentales para despejar el cerebro y remontar con mejores ideas.

Gracias, gracias a todos, les dedico este trabajo, y esta parte de mi vida.

Resumen

El análisis funcional refiere a un conjunto de técnicas que tienen como fin detectar aquellas funciones o procesos que se encuentran desregulados en un experimento biológico. Con el continuo avance en las tecnologías de obtención de expresión de muestras biológicas, la cantidad de bases de datos de libre disponibilidad aumenta constantemente. Las técnicas de análisis funcional se basan en el estudio de un único experimento, en la era del *Big Data* resulta natural notar la necesidad de explotar esta gran cantidad de bases de datos para su integración, y así, generar nuevas fuentes de información.

Esta tesis propone, como objetivo principal, brindar una metodología que permita integrar grandes cantidades de bases de datos de expresión biológica. Integrando información de diversas poblaciones, fenotipos, enfermedades, entre otros, se podrá detectar patrones que caractericen cada grupo. Como primer instancia de tesis, se realizó una comparación exhaustiva de diversas alternativas para llevar a cabo el análisis funcional. Con tantas alternativas existentes, que siguen diversos supuestos e ideas, esta evaluación nos llevó a la creación del pipeline de *Análisis Funcional Integrador*: `IFA` (del inglés *Integrative Functional Analysis*). El `IFA` realiza su análisis tomando alternativas que otorgaron los mejores resultados desde un punto de vista biológico y estadístico.

Para cumplir con el objetivo principal de esta tesis, presentamos la herramienta `MIGSA` (del inglés *Massive and Integrative Gene Set Analysis*). Gracias a esta herramienta, es ahora posible llevar a cabo un análisis funcional masivo e integrador de

grandes cantidades de bases de datos biológicas que provienen tanto de distintas poblaciones como de distintas fuentes biológicas (genes, proteínas, etc.). Además, `MIGSA` proveé diversas herramientas que permiten explorar y visualizar fácilmente los resultados, y de esta manera, validar y generar nuevas hipótesis de estudio. La utilidad de nuestra herramienta fue comprobada ya que permitió, para sub-grupos de cáncer de mama -con pronósticos bien distintivos-, detectar genes y procesos biológicos que los caracterizan. `MIGSA` representa una herramienta que permite detectar efectivamente aspectos biológicos que podrían ser blancos de drogas, y así contrarrestar la condición bajo estudio.

Clasificación (ACM CCS 2012):

- *Applied computing ~> Life and medical sciences ~> Computational biology*
- *Applied computing ~> Life and medical sciences ~> Bioinformatics*

Palabras claves: *Minería de datos - Ciencia de datos - Integración de información - Bioinformática*

Índice general

Prefacio

Actualmente enfermedades como el cáncer, han sido abordadas mediante el análisis de genes individuales. Si bien a través de los años se han desarrollado terapias específicas contra oncogenes determinados, que han disminuido su mortalidad a corto plazo, la enfermedad encuentra caminos alternativos para volver a manifestarse con diferente intensidad y a diferentes tiempos. Por otro lado, hay evidencias que diversas firmas moleculares con muy pocos genes en común son capaces de estratificar, de manera similar, la misma población en términos del desarrollo de la enfermedad. Si bien los genes detectados o presentes en las diversas firmas moleculares no coinciden, sí lo hacen las vías de acción implicadas en el desarrollo de la enfermedad. Por esta razón, los estudios genómicos ya no se enfocan en la identificación de listas minimalistas de genes, sino más bien en la aplicación de complejas metodologías bioinformáticas que relacionan información existente, tanto experimental como de bases de datos, para identificar cuáles son las funciones biológicas, moleculares y lugares donde éstos están actuando. Este conjunto de metodologías que permite identificar esas funcionalidades se conoce como Análisis Funcional (AF). La identificación de estas funciones biológicas, moleculares y metabólicas que están activas o deprimidas en un determinado contexto patológico o experimental, no solo permite un abordaje sistémico de la misma, sino que es fundamental para el descubrimiento de información y verificación de hipótesis.

El AF permite identificar, estadísticamente, los mecanismos biológicos desregula-

dos en un experimento. Los métodos de AF se basan en la evaluación, no de genes individuales, sino de grupos de genes, bajo el supuesto de que su acción coordinada impacta un mismo término biológico. Esta tarea se lleva a cabo consultando grandes bases de datos donde grupos de genes están asociados a cada mecanismo biológico. Las estrategias más comunmente utilizadas para el AF son el Análisis de Sobre-Representación y Puntuación Funcional de Clase (ASR y PFC, respectivamente). La principal diferencia entre ambos enfoques es que el primero utiliza una lista de genes de interés como entrada, que suele ser la lista de genes diferencialmente expresados, mientras que los métodos de PFC utilizan todos los genes disponibles en el experimento así como sus valores de expresión.

Tanto para el ASR como PFC se han desarrollado varios algoritmos con sus propios supuestos y parámetros de entrada, para los que cada autor pretende demostrar o enfatizar la superioridad de su algoritmo sobre los demás. Sin embargo, todas las comparaciones disponibles se basan en la evaluación en términos de adecuación de los supuestos de la distribución, estimación del p-valor, eficiencia computacional, entre otros, en lugar de evaluarlos desde un punto de vista de la información biológica obtenida. Por lo tanto, seleccionar el algoritmo apropiado y el ajuste de sus parámetros no es una decisión trivial para los investigadores. Más aún, no queda claro si un método es completamente superior al resto o si los resultados de cada algoritmo son independientes, complementarios, o igualmente útiles.

Por otra parte, el surgimiento y rápido avance de las tecnologías de obtención de niveles de expresión biológica, han llevado a la disponibilidad de miles de experimentos en repositorios públicos, no solo con información a nivel de genes si no que también a otros niveles moleculares como proteínas o transcriptos. La disponibilidad de estas grandes fuentes de información crearon oportunidades sin precedentes para estudiar enfermedades humanas. Integrando información funcional de diversos repositorios como de distintas fuentes moleculares es posible llegar a una caracterización

de grupos de sujetos de interés. Atacando aspectos funcionales activos por uno u otro grupo de sujetos bajo estudio se logra el desarrollo de terapias personalizadas. Es por ello que resulta fundamental poder realizar una comparación y caracterización de multiples fuentes de datos a nivel funcional.

La presente tesis proporciona una metodología integradora que permite, desde el punto de vista funcional, comparar correctamente grandes cantidades de bases de datos de experimentos provenientes tanto de diversas fuentes ómicas como de distintos grupos de estudio. El primer desafío consistió en evaluar las ventajas y desventajas de los algoritmos existentes de AF. El segundo desafío fue adaptar los datos provenientes de diversas fuentes ómicas al *pipeline* convencional de AF. Finalmente se desarrolló una herramienta, `MIGSA`, que permite una evaluación integradora de grandes colecciones de bases de datos biológicas.

La organización del documento de tesis es como sigue:

- **Capítulo 1:** introduce al lector en el concepto del **análisis funcional** y las distintas metodologías para llevarlo a cabo. Se presenta el concepto de **ontologías** biológicas, y la información presente en ellas.

- **Capítulo 2:** expone las diversas **fuentes de datos** biológicas que resultan de interés para el presente trabajo de tesis. Como así también los diversos repositorios de bases de datos públicas utilizadas.

- **Capítulo 3:** muestra los aportes realizados en este trabajo de tesis en el contexto del **análisis funcional** integrador. Los aportes están dirigidos a la comparación, desde el punto de vista biológico, de diversos algoritmos junto a sus diferentes parámetros.

- **Capítulo 4:** presenta la **herramienta desarrollada** que permite llevar a cabo un **análisis funcional masivo** e integrador de múltiples bases de datos. Nuestra herramienta proporciona métodos de exploración y visualización que

permiten responder preguntas, como así también desarrollar nuevas hipótesis.

- **Capítulo 5:** exhibe **desafíos que surgieron** durante el desarrollo del presente trabajo, y **como fueron afrontados**. Estos desafíos no estaban directamente relacionados con los objetivos bajo estudio, pero aportaron gratamente al resultado final de la tesis.

- **Capítulo 6:** muestra las **conclusiones y trabajos futuros** producto de la presente tesis. Se destacan los diferentes aportes realizados al estado del arte, así como también las posibles líneas que se pueden continuar a partir de lo realizado a lo largo del doctorado.

Capítulo 1

Análisis funcional

El desarrollo de las tecnologías de obtención de niveles de expresión biológica, dio lugar a grandes avances en la biología, permitiendo así pasar del análisis individual de genes o proteínas a obtener información de todo el genoma y el proteoma. Con esta información, principalmente se lleva a cabo lo que se conoce como análisis de expresión diferencial, el cual permite obtener listas de genes ó proteínas que se encuentran expresadas diferencialmente en distintas condiciones de estudio, por ejemplo, tejido tumoral vs. tejido normal.

Sin embargo, resulta de un gran interés biológico evaluar la interacción de todo el conjunto de genes sobre diversos mecanismos o funciones biológicas. Esta tarea se lleva a cabo consultando grandes bases de datos, conocidas como ontologías, las cuales almacenan información biológica funcional a nivel de genes. Cada ontología codifica dicha información en conjuntos de genes, cada uno, asociado a una función, localización, enfermedad particular. Cada conjunto de genes tiene un nombre o descripción del evento, y los genes específicos que se sabe interactúan para llevar a cabo ese evento.

Una vez determinada la información ontológica a utilizar, se aplican técnicas estadísticas para evaluar si la relación que se observa en el experimento es un evento

azaroso o no, al compararlo con un comportamiento de referencia (Rivals, Personnaz, Taing, & Potier, 2006). De esta manera se obtiene para cada conjunto de genes, una probabilidad que determina si dicha función biológica está desregulada según la condición experimental bajo estudio. Este conjunto de métodos estadísticos permiten realizar el Análisis Funcional (AF) partiendo de la matriz, con niveles de expresión de cada gen (filas) para cada muestra (columnas), y la ontología a evaluar.

1.1. Ontologías

En biología, las ontologías son grandes bases de datos de anotación, las cuales mediante vocabulario controlado permiten almacenar, de una manera estructurada, la información conocida. Por ejemplo, las ontologías utilizadas para el AF contienen información conocida sobre grupos de genes, que al interactuar desencadenan algún concepto o término biológico.

1.1.1. Gene Ontology

Existen grandes cantidades de ontologías, cada una con hasta decenas de miles de mecanismos biológicos descriptos. En la presente tesis se analizaron las categorías de una de las ontologías de mayor difusión en la comunidad científica: Gene Ontology (GO) (Ashburner et al., 2000). En esta ontología, la información se encuentra estructurada en tres categorías:

- **Funciones moleculares (FM)** describen actividades que ocurren a nivel molecular. Sin especificar dónde, cuándo, o en qué contexto, tiene lugar la acción. Las FM pueden ser tan generales como “actividad catalítica”, “actividad del transportador” o “binding”; o específicas como “actividad de la adenilatociclasa”.

- **Procesos biológicos (PB)** refieren a una serie de eventos realizados por uno

o más conjuntos ordenados de FM. Un proceso biológico se lleva a cabo mediante un conjunto particular de FM, de una manera altamente regulada y en una secuencia temporal particular. Ejemplos de términos de PB generales son "apoptosis" o "transducción de señales". Ejemplos más específicos son "proceso metabólico de la pirimidina" o "transporte de alfa-glucósidos".

- **Componentes celulares (CC)** describen un componente de una célula, con la condición de que forme parte de un objeto más grande; puede ser una estructura anatómica (por ejemplo, un núcleo o retículo endoplásmico rugoso) o un grupo de productos génicos (por ejemplo, un ribosoma, un proteasoma o un dímero proteico). Ejemplos de CC son "parte citoplasmática de la membrana plasmática", "mitocondria" o "ribosoma". A diferencia de las otras dos categorías de GO, los conceptos de CC no se refieren a procesos sino a la anatomía celular.

En resumen, en GO se tiene la información de qué genes participan en un PB o FM, y en que CC actúan. Un ejemplo de término de GO es "apoptosis", el cual pertenece a la categoría PB, y se conforma por los genes:

```
conj_genes["apoptosis (GO:0006915)"]
```

```
$`apoptosis (GO:0006915)`
 [1] "ZBTB16"   "DNAJB13"  "SFRP5"    "RRAGC"    "IAPP"     "ELMO1"
 [7] "BAX"      "PDCD4"    "TNFSF9"   "PDCD2"    "PDCD1"    "PTPN6"
[13] "RAF1"     "CYFIP2"   "PPP1R15A" "GPR65"    "AHR"      "TNF"
[19] "CIB1"     "FOXO3"    "KIAA1967" "E2F1"     "AKT1"     "CSE1L"
[25] "NLRP1"    "PHLDA2"   "FIS1"     "SIAH1"    "YARS"     "SEMA6A"
[31] "DAXX"     "GML"      "GADD45B"  "GADD45A"  "LY86"     "..."
```

Vale la pena aclarar que un gen particular puede participar en más de un término y categoría particular. Cada una de las tres categorías de GO está estructurada como

un grafo acíclico dirigido (árbol), donde cada término tiene relaciones definidas con uno o más términos del mismo dominio. Cada nodo del grafo tiene asociados los genes que participan en dicho término. Cada grafo se encuentra organizado de una manera jerárquica, donde a mayor profundidad, se representa un concepto biológico más específico, y por ende, disminuye la cantidad de genes en cada término. En este sentido, un gen que se encuentra anotado en un nodo dado, también se encuentra anotado en sus nodos ancestros. En la Figura 1.1 se puede observar el subárbol con el PB "apoptosis" y sus términos ancestros.

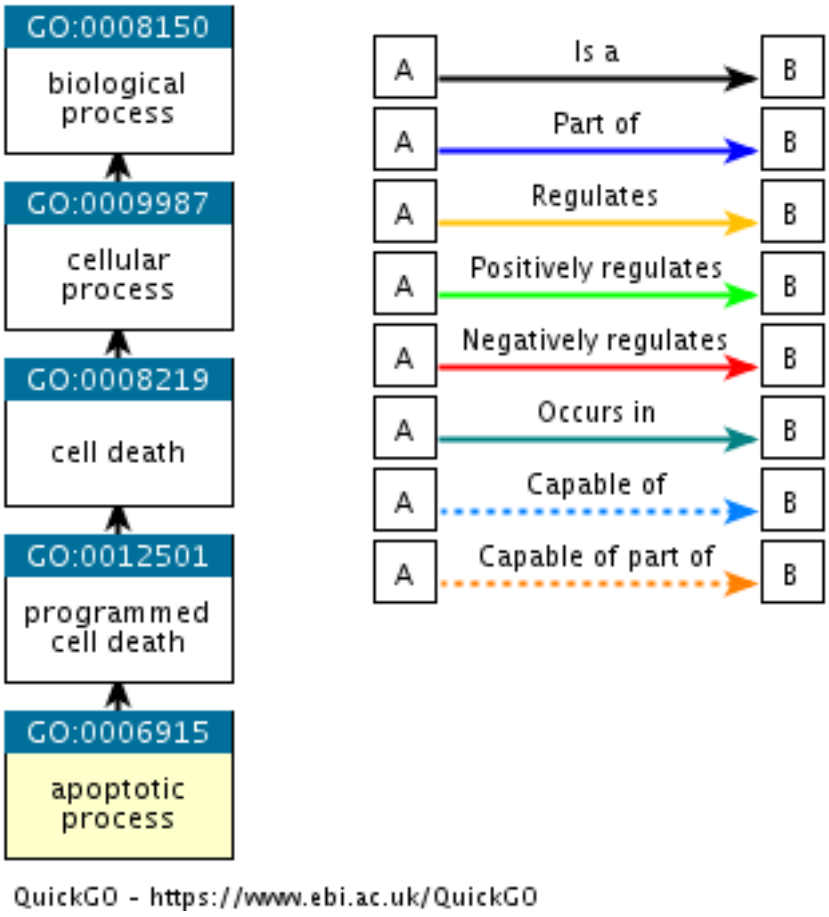


Figura 1.1: Subárbol de Gene Ontology, donde se presenta el término "apoptosis" (apoptotic process) y sus términos ancestros. Imagen extraída de `https://www.ebi.ac.uk/QuickGO/term/GO:0006915`.

El vocabulario de GO está diseñado para ser agnóstico a las especies, e incluye términos aplicables a procariotas, eucariotas, organismos unicelulares y multicelula-

res. Las revisiones continuas de la ontología están a cargo de un equipo de editores de ontología con amplia experiencia en biología y representación de conocimientos computacionales.

1.1.2. Otras ontologías

Si bien en la presente tesis se analizó la ontología GO, existen cientos de otras ontologías que pueden ser indagadas mediante AF. Entre estas otras ontologías existentes, debido a su gran popularidad en la comunidad científica, vale la pena mencionar:

KEGG

La ontología provista por el KEGG (Kanehisa & Goto, 2000) contiene, para cada término o vía metabólica, además de los genes que lo influencian, un diagrama visual que representa el conocimiento experimental sobre el mismo y varias otras funciones que interactúan. Cada diagrama contiene una red de interacciones y reacciones moleculares y está diseñado para vincular los genes del genoma con los productos génicos (principalmente proteínas) de la vía. Esto ha permitido examinar qué vías y funciones asociadas, es probable que estén codificadas en el genoma.

Ejemplos de términos presentes en esta ontología son “apoptosis” o “ciclo celular” (Figura 1.2). En estos diagramas se identifican tres tipos de elementos: cajas rectangulares para representar productos de genes, flechas para el flujo de las reacciones, y cajas con bordes redondeados para vincular a otras vías involucradas en el proceso. A su vez, esta base de datos se complementa con un conjunto de tablas de grupos ortólogos, donde se encuentra la información de subvías conservadas, que son especialmente útiles para la predicción de funciones de genes.

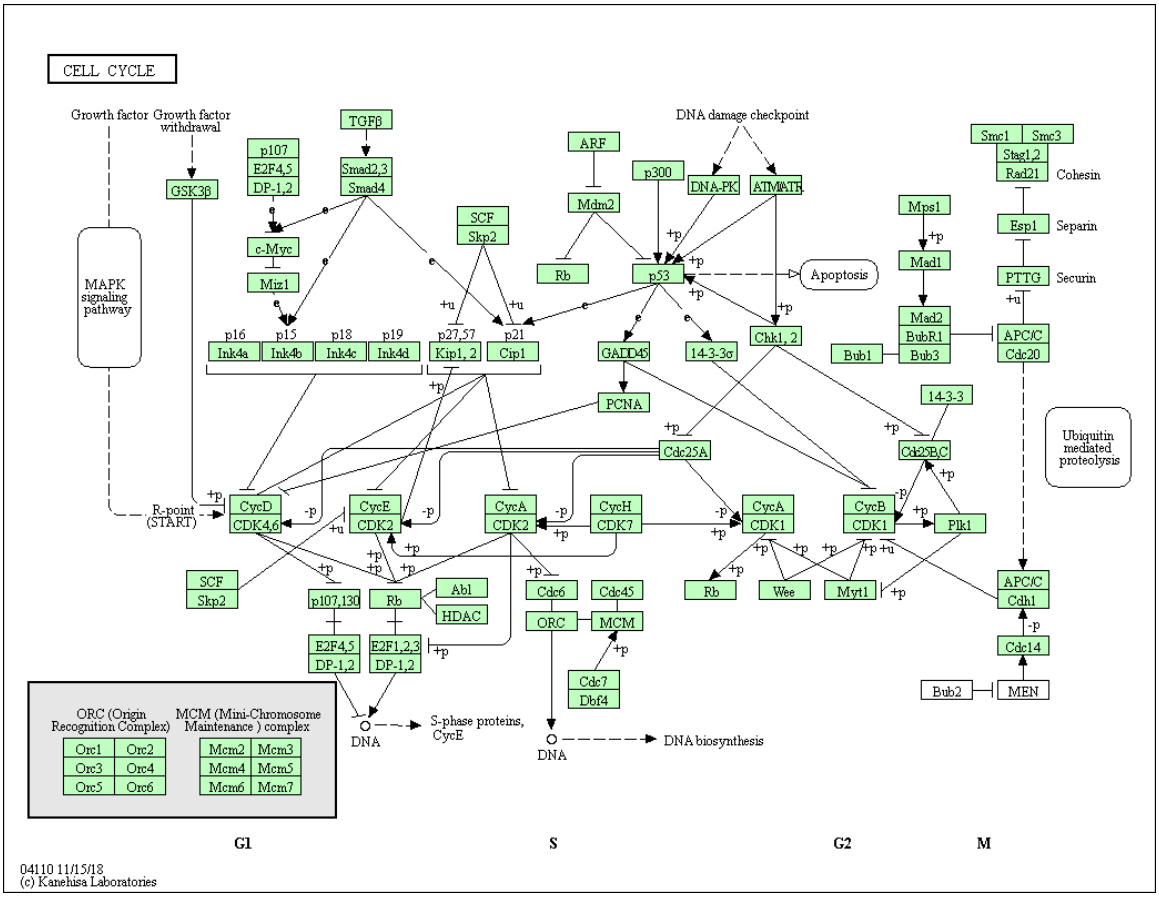


Figura 1.2: Diagrama de KEGG del "ciclo celular" (cell cycle). Imagen extraída de `www.genome.jp/kegg-bin/show_pathway?hsa04110`.

Según los desarrolladores, es una "representación informática" del sistema biológico, formada por bloques de construcción y diagramas de cableado, más específicamente, bloques de construcción de genes o proteínas, bloques de construcción de químicos de pequeñas moléculas y reacciones, y diagramas de cableado de interacción molecular y redes de reacción.

BioCarta

BioCarta (Nishimura, 2001) fue creado por un grupo de científicos e ingenieros que reconocieron las limitaciones y obsolescencia de la literatura científica de la época. Es muy similar a la ontología del KEGG en el sentido que cada conjunto de genes viene acompañado de un diagrama de las interacciones. Sin embargo, en esta ontología, los

diagramas presentan un mayor detalle gráfico como se puede observar en la Figura 1.3, donde se presenta el diagrama del término "corticosteroides y cardioprotección".

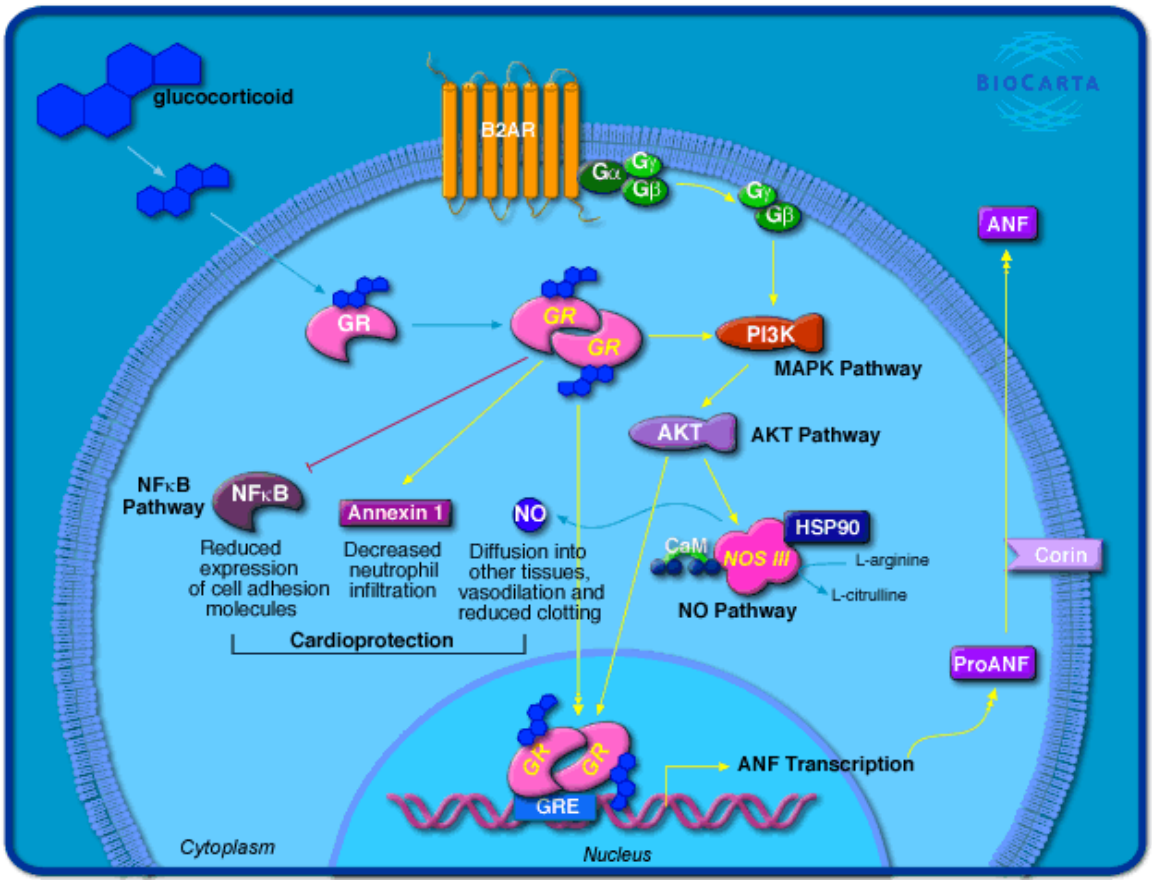


Figura 1.3: Diagrama de BioCarta del término "corticosteroides y cardioprotección". Imagen extraída de `https://david.ncifcrf.gov/biocarta.jsp?path=h_gcrPathway$Corticosteroids+and+cardioprotection&termId=30000124`.

La principal diferencia de BioCarta con KEGG, es que BioCarta fue diseñado para proporcionar un foro para la difusión de información, así como para promover la colaboración entre investigadores, educadores y estudiantes. BioCarta alienta a los científicos a contribuir con contenidos relacionados con sus áreas específicas de especialización. De esta manera, se espera que la contribución de un gran número de especialistas en ciencias de la vida en muchos campos diferentes permita la construcción de modelos detallados y precisos de las interacciones de los productos génicos que superen los esfuerzos de un solo grupo.

Reactome

Reactome: una base de datos de reacciones, vías y procesos biológicos. Reactome (Joshi-Tope et al., 2005) está desarrollado por biólogos expertos, en colaboración con un equipo editorial, todos ellos biólogos con nivel de doctorado. El contenido tiene referencias cruzadas a muchas bases de datos bioinformáticas. El objetivo principal de Reactome es representar visualmente las vías biológicas con todo detalle mecanicista, a la vez que se ponen a disposición los datos de la fuente en un formato computacionalmente accesible.

La unidad central del modelo de datos de Reactome es la reacción. Las entidades (genes, proteínas, complejos y pequeñas moléculas) que participan en las reacciones forman una red de interacciones biológicas y se agrupan en vías. Ejemplos de vías biológicas en Reactome incluyen la "función inmune innata y adquirida", la "regulación transcripcional", y como se detalla en la Figura 1.4, la "apoptosis".

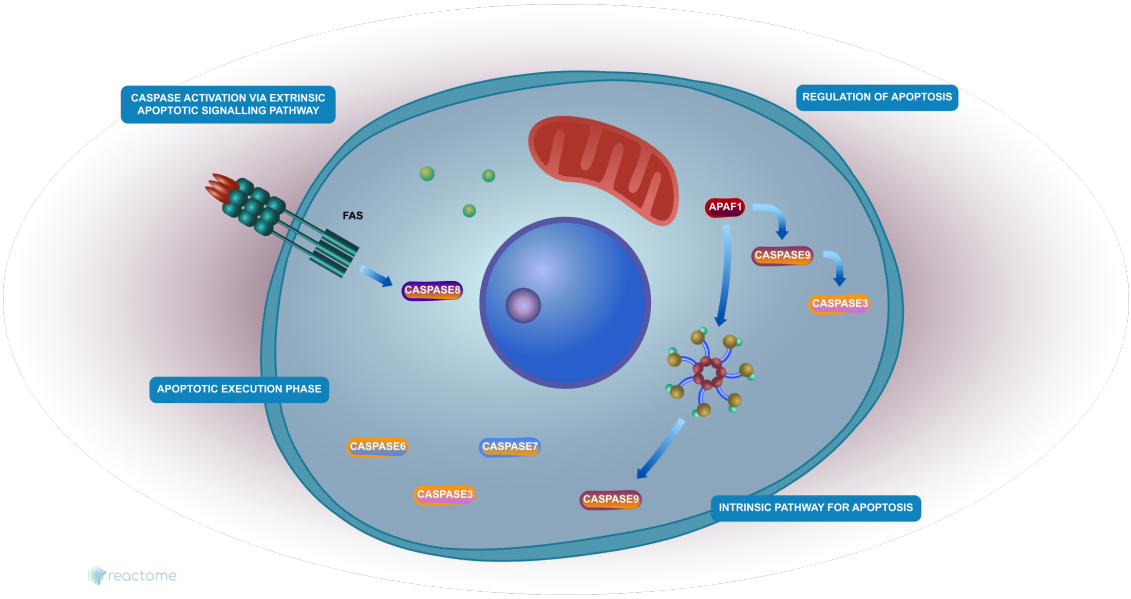


Figura 1.4: Diagrama de Reactome del término "apoptosis". Imagen extraída de `https://www.reactome.org/content/detail/R-HSA-109581`.

Las vías representadas en Reactome son específicas para cada especie, y cada

paso de la vía está respaldado por citas bibliográficas que contienen una verificación experimental del proceso representado. Si no existe una verificación experimental, las vías pueden contener pasos inferidos, pero sólo si un biólogo experto, nombrado como *autor de la vía*, y un segundo biólogo, nombrado como *revisor*, están de acuerdo en que esto es una inferencia válida a hacer. Las vías humanas se utilizan para generar computacionalmente, mediante un proceso basado en la ortología, las vías derivadas de otros organismos.

MSigDB

MSigDB (Liberzon et al., 2011) es la ontología desarrollada por el Broad Institute (`https://www.broadinstitute.org/`), esta ontología es un compilado de ciertos conjuntos de genes provenientes de otras ontologías, entre ellas las descriptas previamente. MSigDB agrupa todos sus conjuntos de genes dentro de ocho categorías:

- **Conjuntos de genes distintivos**: resumen y representan estados o procesos biológicos específicos bien definidos. Estos conjuntos de genes fueron generados por una metodología computacional basada en la identificación de superposiciones entre conjuntos de genes en MSigDB y la retención de genes que muestran detalles de expresión coordinada.
- **Conjuntos de genes posicionales**: correspondientes a cada cromosoma humano y a cada banda citogenética que tenga al menos un gen.
- **Conjuntos de genes curados**: obtenidos y curados a partir de diversas fuentes, como bases de datos de ontologías (KEGG, BioCarta y Reactome), literatura biomédica y el conocimiento de expertos en la materia.
- **Conjuntos de genes del motivo**: representan objetivos potenciales de regulación por factores de transcripción o microARNs. Los conjuntos consisten en genes agrupados por contener motivos de secuencia corta en común, pertenecientes a regiones codificadoras no proteicas. Los motivos representan elementos

de regulación cis conocidos o probables en promotores y 3'-UTRs.

- **Conjuntos de genes computacionales**: definidos por la minería de grandes colecciones de datos orientados al cáncer.
- **Conjuntos de genes de GO**: contienen conjuntos de genes seleccionados de GO.
- **Firmas oncogénicas**: representan firmas de vías celulares que a menudo están desreguladas en el cáncer. La mayoría de las firmas se generaron directamente a partir de datos de genes que implicaban la perturbación de genes cancerígenos conocidos.
- **Firmas inmunológicas**: representan estados celulares y perturbaciones dentro del sistema inmunológico. Las firmas fueron generadas por la curación manual de estudios publicados en inmunología humana y de ratones.

1.2. Metodologías de Análisis Funcional (AF)

El análisis funcional refiere a técnicas que permiten, dado un experimento, evaluar el impacto de la interacción de grupos de genes sobre características biológicas. Partiendo de una matriz de expresión de genes con muestras pertenecientes a una de dos condiciones, los algoritmos de AF buscan detectar aquellos términos biológicos que se encuentran desregulados entre ambas condiciones. Una vez determinada la base de datos ontológica de interés, el investigador debe seleccionar cuál técnica de AF utilizar.

Se diferencian dos categorías de algoritmos para llevar a cabo el AF (D. W. Huang et al., 2008a). La principal diferencia entre estas categorías es la estrategia que llevan a cabo para el análisis:

1.2.1. Análisis de Sobre-Representación (ASR)

Esta metodología requiere definir la lista de genes de interés o **candidatos**, generalmente aquellos que se encuentran diferencialmente expresados en el experimento, es decir, un vector con los nombres de los genes. Adicionalmente requiere de una segunda lista de genes, que utiliza para especificar lo que es esperable como comportamiento de **referencia** del modelo biológico, generalmente todos los genes presentes en el experimento. Luego, para cada término, realiza de forma independiente un test de hipótesis comparando las proporciones observadas sobre los candidatos con respecto a la referencia. De este modo, para cada término se obtiene un p-valor asociado que denota si existe evidencia de que las proporciones son diferentes.

En la Figura 1.5 se puede apreciar un diagrama de lo que sería el proceso completo para llevar a cabo un ASR. De este diagrama se desprenden varios aspectos que resultan variables. Diversos autores han propuesto diferentes test de hipótesis (Falcon & Gentleman, 2006; Fang & Gough, 2014; Fresno & Fernández, 2013; D. W. Huang et al., 2008b), y principalmente de allí es que surgen grandes cantidades de alternativas de ASR. Más aún, la elección de la lista de referencia no queda del todo clara, y se desprenden dos sugerencias con gran validez estadística: utilizar todos los genes del experimento (como en la Figura 1.5), o utilizar todo el genoma de la especie en cuestión (Fresno et al., 2012).

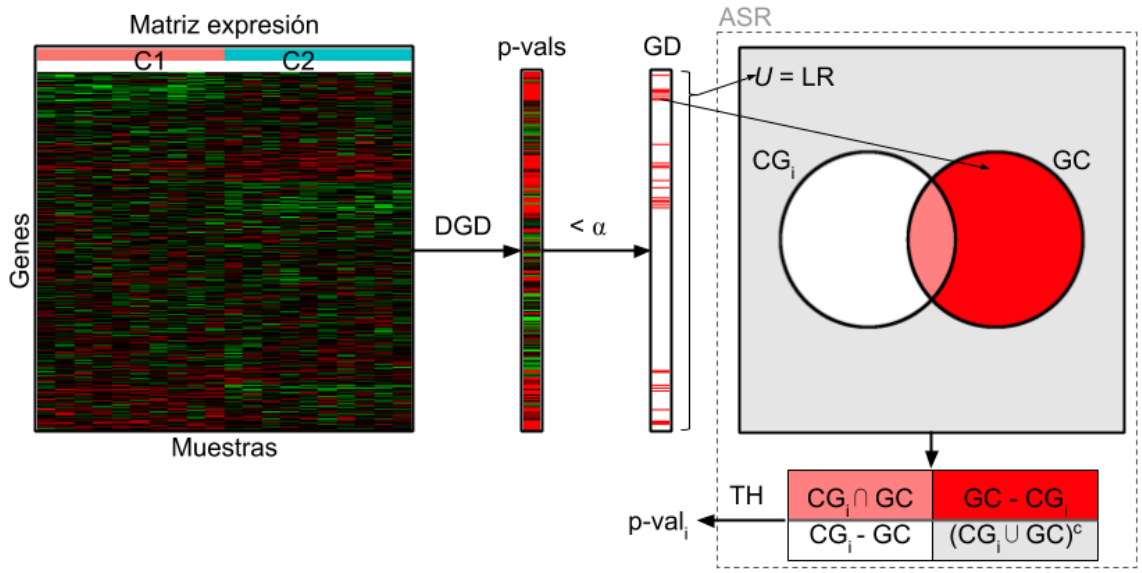


Figura 1.5: Diagrama clásico del proceso completo de Análisis de Sobre-Representación (ASR). Este proceso parte de una Matriz de expresión, con Genes en filas y Muestras en columnas, y la etiqueta de condición de cada muestra (C1 y C2). Mediante un algoritmo de Detección de Genes Diferenciales (DGD) se obtiene para cada gen un p-valor asociado (p-vals). Aplicando un nivel de corte α a estos p-valores ($< \alpha$), se obtiene cuáles son Genes Diferenciales (GD) y cuáles no. Aquí es donde efectivamente comienza el ASR, los algoritmos de ASR requieren de dos listas: la de genes candidatos (usualmente los GD), y la Lista de Referencia (LR; usualmente todos los genes detectados en el experimento). A partir de estas listas de genes, para cada Conjunto de Genes (CG_i) se genera una tabla de contingencia, a la cual se le aplica un Test de Hipótesis (TH) del cual resulta un p-valor asociado al CG_i ($p\text{-val}_i$).

1.2.2. Puntuación Funcional de Clase (PFC)

Esta metodología construye un ordenamiento de los genes a partir de la totalidad del perfil de expresión. El objetivo es determinar si todos los miembros de una característica biológica de interés, están distribuidos aleatoriamente o no (en algún extremo) a lo largo del ordenamiento generado (Subramanian et al., 2005). Para ello, los genes se organizan mediante algún criterio de ordenamiento que refleje la diferencia entre ambas condiciones bajo estudio, por ejemplo, para cada gen obtener la diferencia de medias entre ambas condiciones. Luego, para un término biológico en particular, se

recorre este ordenamiento para calcular el máximo del enriquecimiento inducido, lo cual se conoce como **Enrichment Score** (ES) del término. Para ello, a cada gen se le aplica una función de coste que aumenta (o disminuye) proporcionalmente a la correlación de su nivel de expresión con el fenotipo de las condiciones, cada vez que encuentra un gen que pertenece (o no) a la lista de miembros de la categoría de interés. Este ES se compara con la distribución nula generada por las permutaciones en las etiquetas de las condiciones, con el fin de evaluar si el ordenamiento original es esperable o generado por azar. En caso de que difiera del azar, existen diferentes criterios para definir, para el experimento dado, cuáles son los genes más influyentes en el término. Normalmente se utilizan los genes más cercanos al máximo del ES, o los genes pertenecientes al segmento más pequeño entre el coste máximo y el principio o el final de la lista ordenada.

En la Figura 1.6 se observa el diagrama de lo que sería un proceso completo de PFC. De este diagrama, diversos aspectos difieren dependiendo del algoritmo de PFC a utilizar. Por ejemplo, ciertos algoritmos sugieren hacer permutación en las etiquetas de los genes, otros en las etiquetas de las muestras, y finalmente otros, en las etiquetas de genes del ordenamiento generado. Más aún, la principal diferencia entre uno y otro algoritmo de PFC reside en la función de ES, si bien la idea por detrás resulta similar, las ponderaciones de cada gen varían. Finalmente, vale la pena mencionar que ciertos algoritmos fijan diferentes filtros de los conjuntos de genes a analizar ya que suponen que fuera de esos límites el algoritmo pierde potencia estadística.

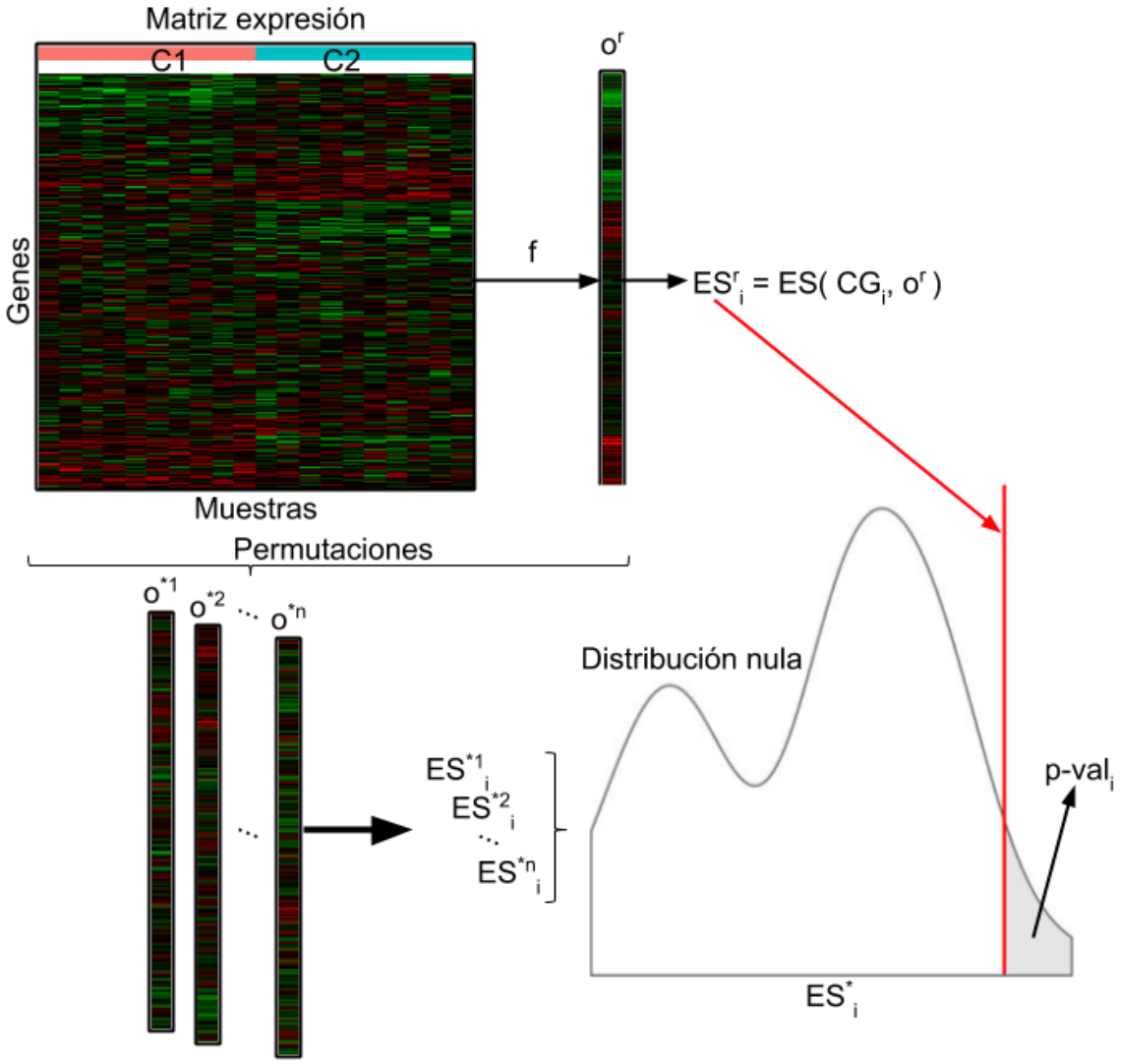


Figura 1.6: Diagrama clásico del proceso completo de Puntuación Funcional de Clase (PFC). Este proceso parte de una Matriz de expresión, con Genes en filas y Muestras en columnas, y la etiqueta de de condición de cada muestra (C1 y C2). Mediante alguna función particular (f) se obtiene un ordenamiento real (o^r) para los genes, es decir, para cada gen un valor numérico que le asigna una posición de orden. Luego, mediante una función llamada Enrichment Score (ES), partiendo de un conjunto de genes particular (CG_i) y el ordenamiento, se obtiene un valor de enriquecimiento real para el conjunto de genes (ES_i^r). Esta función ES es diferente dependiendo del algoritmo de PFC utilizado. Luego, mediante permutaciones sobre las condiciones, las etiquetas de los genes, ó del ordenamiento, se generan nuevos vectores de ordenamiento (o_1^*, o_2^*, ..., o_n^*). Con cada uno de estos ordenamientos permutados, se genera un valor de enriquecimiento permutado para el conjunto de genes (ES_i^{*1}, ES_i^{*2}, ..., ES_i^{*n}). Dado que se cuenta con n muestras permutadas de valores de enriquecimiento, se procede a realizar una distribución nula, con la cual se compara contra el valor de enriquecimiento real ES_i^r, y por ende se obtiene un p-valor asociado al conjunto de genes dado ($p\text{-val}_i$).

1.2.3. Análisis de Enriquecimiento Modular

Vale la pena mencionar que en la literatura se suele encontrar una tercer alternativa de AF, llamada Análisis de Enriquecimiento Modular. Esta metodología requiere, adicionalmente, que los CG presenten alguna relación entre ellos. En la presente tesis no se tiene en cuenta esta metodología ya que en el común de las bases de datos de CG no existe relación entre sus términos, lo cual resulta en un gran limitante para el análisis.

1.3. Comentarios finales

Tanto el ASR como PFC se utilizan para saber si un término se enriquece (o no) en la comparación de las condiciones estudiadas. Sin embargo, la formulación del problema es diferente para cada caso. Para realizar el análisis a través del ASR es necesario proporcionar dos listas, una de ellas de referencia y la otra de los genes candidatos. Este último suele estar formado por aquellos genes que se han identificado como expresados de forma diferencial entre dos condiciones experimentales (por ejemplo, tumor vs. normal) para un umbral definido. En PFC, se utiliza una matriz de expresión única que contiene todos los genes detectados por el experimento y, a continuación, se utiliza el criterio de ordenación propuesto para medir el enriquecimiento. En este sentido, PFC en comparación al ASR, no utiliza un umbral para definir la lista de candidatos, y por ende PFC utiliza toda la información presente en el experimento.

Si bien, existen principalmente dos metodologías para realizar el AF, tanto para ASR como para PFC, existen grandes cantidades de diferentes algoritmos desarrollados, cada uno con sus propios supuestos, test, y parametrizaciones. En este sentido, no fue evaluado qué beneficios trae un algoritmo frente al resto. Por consiguiente, para el investigador, la elección del algoritmo de AF a utilizar termina siendo más una

cuestión de elección aleatoria, que una elección adaptada al experimento en cuestión, de la hipótesis bajo estudio, o del enfoque de la investigación.

Capítulo 2

Fuentes de datos biológicas

La célula es la unidad estructural, funcional y biológica básica de todos los organismos vivos conocidos. Los organismos pueden clasificarse como unicelulares (compuestos de una sola célula, incluidas las bacterias) ó multicelulares (incluidas las plantas y los animales). Existen dos tipos de células, las eucariotas, que contienen núcleo celular, y las procariotas, que no lo contienen.

Las células eucariotas están compuestas por diversos orgánulos como la membrana, el citoplasma y el núcleo (Figura 2.1). La membrana envuelve y protege a la célula, y regula lo que entra y sale (selectivamente permeable). Dentro de la membrana, el citoplasma ocupa la mayor parte del volumen de la célula, y la separa del núcleo celular. El más prominente de los orgánulos es el núcleo celular, el cual aloja el material genético de la célula. El material genético se presenta como Acido DesoxirriboNucleico (ADN), el cual está organizado en una o más moléculas, llamadas cromosomas.

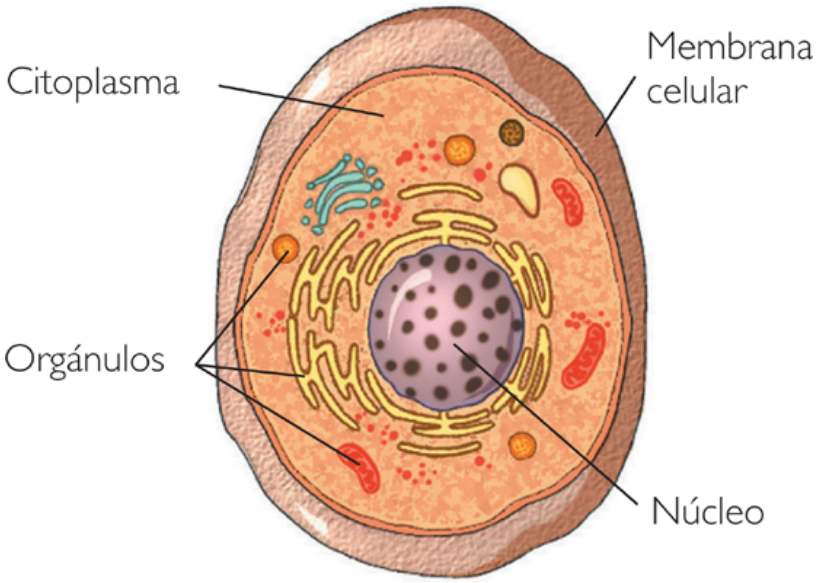


Figura 2.1: Diagrama simplificado de una célula eucariota. Imagen extraída de `https://biologiarubenurjc.wordpress.com/2012/03/19/membrana-nucleo-y-citoplasma/`.

La información biológica contenida en un organismo está codificada en su secuencia de ADN. El ácido desoxirribonucleico esta organizado en dos cadenas que se enrollan una alrededor de la otra para formar una doble hélice (Figura 2.2) que lleva las instrucciones genéticas utilizadas en el crecimiento, desarrollo, funcionamiento y reproducción de todos los organismos conocidos. El ADN esta conformado por unidades más pequeñas conocidas como nucleótidos. Cada nucleótido está compuesto por un azúcar llamado desoxirribosa, un grupo de fosfatos, y por una de cuatro bases nitrogenadas que son la citosina (C), guanina (G), adenina (A) ó timina (T). Los nucleótidos están unidos entre sí en una cadena por enlaces covalentes entre el azúcar de un nucleótido y el fosfato del siguiente. Las bases nitrogenadas de las dos cadenas de nucleótidos separadas se unen, según las reglas de emparejamiento de bases (A con T y C con G). Ambas cadenas de ADN almacenan la misma información biológica. Las regiones relevantes del ADN se encuentran localizadas en los cromosomas y se denominan genes. Si bien la cadena de ADN contiene millones de nucleótidos, solo un pequeño porcentaje de ella codifica proteínas (alrededor del 2 % para los humanos).

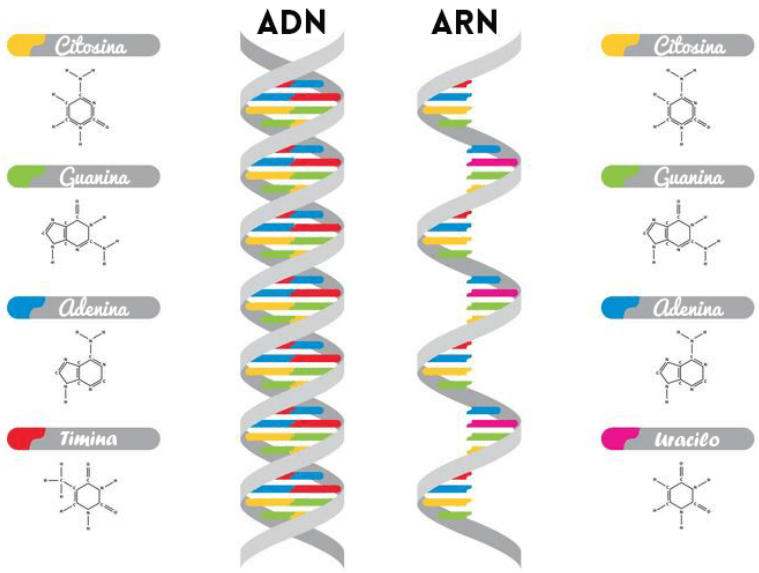


Figura 2.2: Diagrama de la estructura del ADN y ARN. Imagen extraída de `https://diferencias-entre.org/diferencias-entre-adn-y-arn/`.

Las células utilizan el ADN para el almacenamiento de información a largo plazo. Por otra parte, para las demás tareas en que sea necesaria la información genética, las células utilizan el Ácido RiboNucleico (ARN). El ARN se obtiene a partir del ADN para llevar a cabo tareas celulares como la síntesis de proteínas, las cuales son cadenas de aminoácidos que tienen funcionalidades básicas tanto para el metabolismo como para la fisiología celular y, en consecuencia, del organismo. La decodificación del material genético comienza dentro del núcleo celular, donde las hebras de ARN se crean utilizando el ADN como plantilla en un proceso llamado transcripción. Al igual que el ADN, el ARN se ensambla como una cadena de nucleótidos, pero a diferencia del ADN, se encuentra como una única hebra (Figura 2.2), donde las bases timina son reemplazadas por uracilo (U).

Los organismos celulares utilizan ARN mensajero (ARNm) para transmitir información genética. Orgánulos llamados ribosomas procesan el ARNm tomando cada combinación de tres nucleótidos para codificar cada uno de los 20 aminoácidos posibles. Posteriormente, los ribosomas generan la cadena de aminoácidos decodificados

y así conforman la proteína codificada.

De los resultados de estos procesos celulares, existen diversos aspectos biológicos de interés científico: el genoma, el proteoma, el metaboloma, entre otros, los cuales se conocen como las diversas fuentes ómicas. A partir de una muestra biológica, para cada fuente ómica, existen tecnologías capaces de medir sus niveles de expresión. Es decir, para una misma muestra es posible obtener niveles de expresión tanto de genes, proteínas, etc.

En la presente tesis nos centraremos únicamente en aquellas ómicas que permitan obtener sus niveles de expresión en forma de matriz. Por ejemplo, en genómica, al secuenciar m muestras, es posible obtener una matriz G_{g*m}, con expresión obtenida para g genes, donde $G[i, j]$ será un valor numérico representando el nivel de expresión del *i-ésimo* gen para la *j-ésima* muestra.

2.1. Tecnologías de obtención de expresión biológica

La presente sección tiene como objetivo mencionar brevemente aspectos pertinentes sobre las tecnologías mediante las cuales se obtienen los niveles de expresión de las ómicas analizadas en la tesis.

2.1.1. Microarreglos de ADN

Un microarreglo es una superficie sólida donde pequeños fragmentos de ADN (sondas) son dispuestos en forma de matriz bidimensional. Cada celda contiene secuencias de ADN correspondientes a **genes**, ligados químicamente en cada celda. Para medir la expresión génica se extrae ARNm de muestras biológicas, este ARNm es luego copiado (transcripción reversa) obteniendo como resultado ADN complementario o ADNc. A este último se lo amplifica incorporando moléculas fluorescentes en las ré-

plicas. Las copias fluorescentes del ADNc se vuelcan luego, sobre el microarreglo. De esta manera, las secuencias marcadas con moléculas fluorescentes se hibridizan (se "pegan") a su cadena complementaria presente en el microarreglo. Luego, el microarreglo se escanea excitando las celdas con un láser y midiendo la intensidad de luz emitida por las moléculas fluorescentes. El resultado de escanear un microarreglo es una imagen por cada microarreglo. La intensidad medida en cada celda es, en principio, proporcional a la cantidad de ARNm, específico para esa celda, presente en la muestra biológica (Fernández, Alvarez, Podhajcer, & Stolovitzky, 2007).

Las imágenes resultantes del escaneo son procesadas por programas informáticos que identifican las celdas del microarreglo y miden la intensidad de luz registrada. Como resultado del procesamiento de la imagen se obtiene una serie de datos por cada celda perteneciente a cada microarreglo. En particular se obtienen los valores de intensidad de la celda, de la intensidad que rodea a la celda (intensidad de fondo), algunos índices que aportan información sobre las características de la celda (por ejemplo, el área y el perímetro de la misma) y la distribución de las intensidades dentro de cada celda. Toda esta información se utiliza para determinar la calidad del escaneo (Fresno et al., 2014). Una vez eliminadas aquellas celdas defectuosas o que presentan niveles de calidad de señal inadecuados, se normalizan los valores, comúnmente aplicando logaritmo. Finalmente se obtiene como resultado un archivo con la intensidad o el nivel de expresión de cada celda o gen.

Luego de procesar los archivos resultantes de microarreglos, y repetir el procedimiento para varios sujetos, se llega a una matriz de expresión con genes en filas y sujetos en columnas. Cada valor de la matriz representa la intensidad o nivel de expresión de un gen para una muestra. Luego de una correcta normalización de esta matriz, se llega a una del estilo a la que se muestra a continuación; sub-matriz de 6×5 de datos reales provenientes de microarreglos de ADN:

	A2-A0CM-01A	A2-A0D0-01A	A2-A0D1-01A	A2-A0D2-01A	A2-A0EQ-01A
ZBTB16	-0.12125000	0.00425	-0.98975	3.31300000	0.5387500
DNAJB13	-0.48200000	-0.53250	-0.60450	-0.61750000	-0.2670000
SFRP5	-0.02033333	0.22900	0.28440	-0.03216667	-0.3128333
RRAGC	1.23525000	0.68625	0.89750	1.71275000	0.9832500
IAPP	0.11500000	0.06250	1.26450	0.43500000	1.0490000
ELMO1	0.19487500	-0.35275	-0.66825	0.36437500	0.6860000

Dado que los valores de expresión provienen de niveles de intensidad de luz, los datos son valores continuos. Por ende, es de esperar que la distribución de los genes, para cada sujeto, se asemeje a una distribución *Normal*, como se observa en la Figura 2.3.

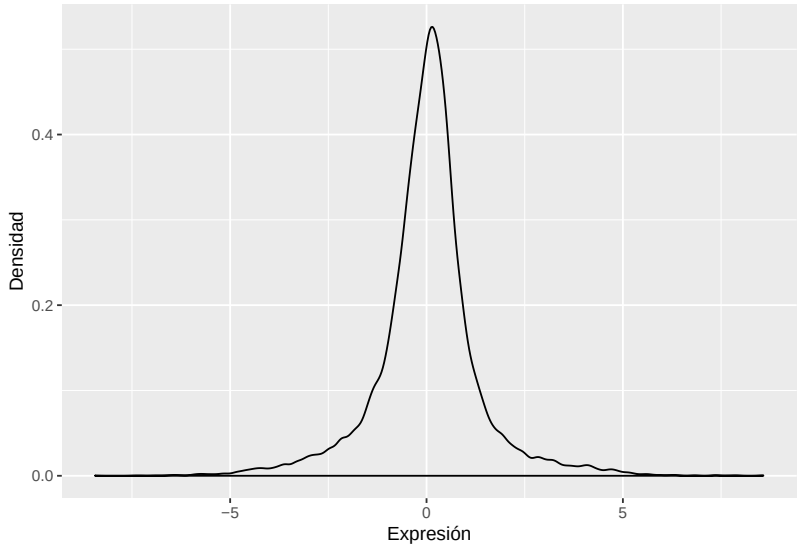


Figura 2.3: Densidad de los valores de expresión de los genes de un sujeto, para datos obtenidos mediante Microarreglos de ADN.

2.1.2. iTRAQ

El método iTRAQ se basa en el marcado químico, con etiquetas de masa variable, de las aminas de los péptidos de las digestiones de **proteínas** presentes en una muestra biológica. Actualmente hay dos reactivos utilizados principalmente, que pueden usarse para marcar todos los péptidos de diferentes muestras. Estas muestras luego

se agrupan y generalmente se fraccionan mediante cromatografía líquida, y se analizan mediante espectrometría de masas en tándem. Luego se realiza una búsqueda en la base de datos utilizando los datos de fragmentación para identificar los péptidos marcados y, por lo tanto, las proteínas correspondientes. La fragmentación de la etiqueta adjunta genera un ión indicador de baja masa molecular que se puede usar para cuantificar relativamente los péptidos y las proteínas a partir de las cuales se originaron.

A nivel peptídico, las señales de los iones indicadores de cada espectro permiten calcular la abundancia relativa (ratio) de los péptidos identificados por este espectro. Las proporciones combinadas de los péptidos de una proteína representan la cuantificación relativa de esa proteína. De esta manera se obtiene una matriz resultante de proteínas $\times$ sujetos, con valores del ratio de la expresión de una proteína para un sujeto dado. Si bien la matriz esta indexada a nivel de proteínas, resulta de mayor interés estudiarla a nivel de genes, por ello, consultando bases de datos de anotación, se traduce cada proteína al gen que la produce. Luego de una normalización adecuada se obtiene una matriz como la que se presenta:

	A2-A0CM-01A	A2-A0D0-01A	A2-A0D1-01A	A2-A0D2-01A	A2-A0EQ-01A
RRAGC	0.17794796	0.27751403	-0.13956920	0.098823426	-0.1745463
ELMO1	0.39888804	0.30589111	-0.09178467	-0.126851824	0.7252130
BAX	0.35139262	0.08414816	0.23821506	-0.118018092	-0.1782356
PDCD4	0.04627689	0.02277544	0.65416083	-0.353588366	-0.3033616
PDCD2	-0.19497549	0.55829865	-0.15272404	-0.024984301	0.6028903
PTPN6	0.17697246	0.19056327	-0.38121552	-0.002048376	0.8563106

Dado que los valores de expresión provienen de niveles de señales, los datos son valores continuos. Por ende, es de esperar que la distribución se asemeje a una distribución *Normal*, como se observa en la Figura 2.4. Vale la pena aclarar que ya que esta matriz sigue una distribución similar a la obtenida mendiante Microarreglos de ADN,

es común que se utilicen los mismos métodos de análisis para ambas fuentes de datos.

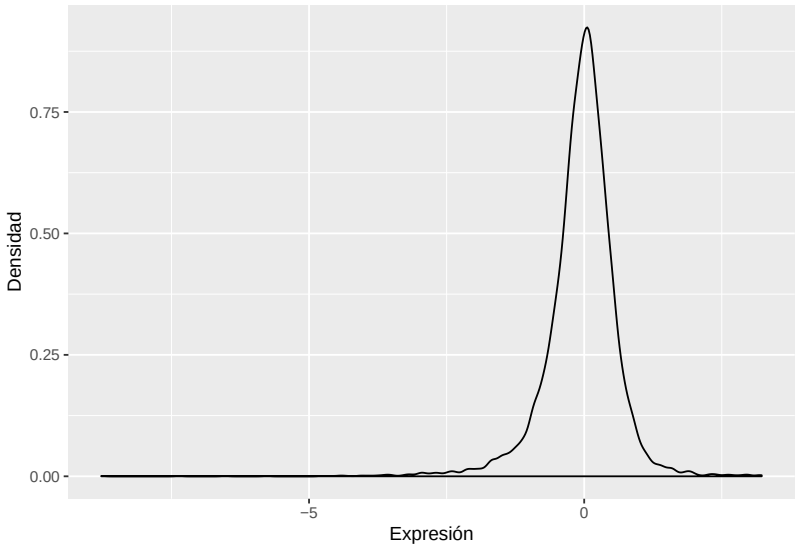


Figura 2.4: Densidad de los valores de expresión de los genes de un sujeto, para datos obtenidos mediante iTRAQ.

2.1.3. Secuenciación de ARN

Este tipo de tecnologías se basa en poder obtener para una muestra biológica, las secuencias de **nucleótidos** detectadas. En este sentido se obtienen millones de cadenas de nucleótidos, que al mapearlos (unirlos) permiten obtener la cantidad de veces que subcadenas del genoma aparecen en la muestra. Al poseer información a nivel de nucleótidos, surgen ventajas con respecto a los microarreglos de ADN, por ejemplo, poder detectar mutaciones de nucleótidos, e incluso estudiar la muestra a niveles más detallados que genes (transcriptos, exones, intrones, etc.).

Una vez obtenido el ARN de la muestra biológica, se retrotranscribe para obtener ADN complementario (ADNc) a estas cadenas. El proceso siguiente es la fragmentación en donde, mediante cortes en secciones aleatorias, se llevan estas grandes cadenas de ADNc a fragmentos de entre 300 y 1000 nucleótidos. Cada uno de estos fragmentos es posteriormente clonado varias veces, produciendo millones de copias de cada fragmento. Aquí es cuando comienza el proceso de secuenciación propiamente dicho. En

el caso de la secuenciación por síntesis, por ejemplo de las plataformas de *Illumina*, los fragmentos de ADNc se ponen en un pool que contiene nucleótidos individuales marcados con fluoróforos, cada una de las cuatro bases con un color diferente. Al entrar en contacto el ADNc con estos nucleótidos libres, se incorporan como hebra complementaria del ADNc, de esta manera se logra incorporar un color a cada una de las bases de los fragmentos de ADNc. Luego, cada fragmento pasa por un sistema que permite leer el color incorporado a cada nucleótido. Es así que traduciendo los colores, se logra detectar los nucleótidos que componen cada fragmento. Al finalizar el proceso de secuenciación de ARN, se cuenta con un archivo con millones de lecturas, una por cada fragmento. Cada lectura se representa con 4 líneas en el archivo, como la que se muestra a continuación:

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

La primer línea es un identificador de la secuencia, la segunda es la que contiene la secuencia de nucleótidos leídos, la tercera es un campo opcional, y la cuarta es la calidad de lectura de cada nucleótido.

Una vez que se cuenta con el archivo resultante de la secuenciación, comienza el proceso bioinformático en sí. El primer paso consiste en la reconstrucción de los transcriptos de las lecturas para determinar de qué genes proceden. Este paso consiste en mapear las lecturas con el genoma ó transcriptoma respectivo. En términos sencillos, si se piensa en las lecturas como piezas de un rompecabezas, el genoma ó transcriptoma es la imagen que se obtendrá uniendo correctamente las piezas. Para el proceso de mapeado, una alternativa es realizar un ensamblaje de novo para inferir las secuencias de las transcripciones sin usar más información que la contenida en las lecturas. Por otro lado, el conocimiento a priori del genoma o transcriptoma de la es-

pecie en estudio se puede utilizar como una referencia que facilitará la reconstrucción, lo cual se conoce como mapeado con referencia. Al final de la etapa de mapeado, se cuenta con información referente a qué gen o transcripto pertenece cada fragmento leído.

Al momento inicial, cuando se obtiene el ARN de la muestra biológica, un gen más activo se encontrará más expresado que el resto, y por consiguiente, el gen presentará mayor cantidad de fragmentos asociados. Para obtener el nivel de expresión de cada gen ó transcripto particular, simplemente se puede tomar la cantidad de fragmentos que fueron mapeados al mismo. De este modo, mediante secuenciación de ARN se permite obtener una matriz de genes × sujetos, con un conteo para cada gen y sujeto, como se puede apreciar a continuación:

	A2-A0CM-01A	A2-A0D0-01A	A2-A0D1-01A	A2-A0D2-01A	A2-A0EQ-01A
ZBTB16	89	31	126	449	81
DNAJB13	11	11	44	31	43
SFRP5	1	2	0	0	0
RRAGC	2797	1453	1680	4573	2631
IAPP	162	78	167	151	195
ELMO1	1740	677	540	1758	4204

Dado que los valores de expresión se desprenden de la cantidad de fragmentos mapeados, los datos son valores de conteos (no continuos). Por ende, es de esperar que la distribución de los genes, para cada sujeto, se asemeje a una distribución de *Poisson* o $BinomialNegativa$, como se observa en la Figura 2.5.

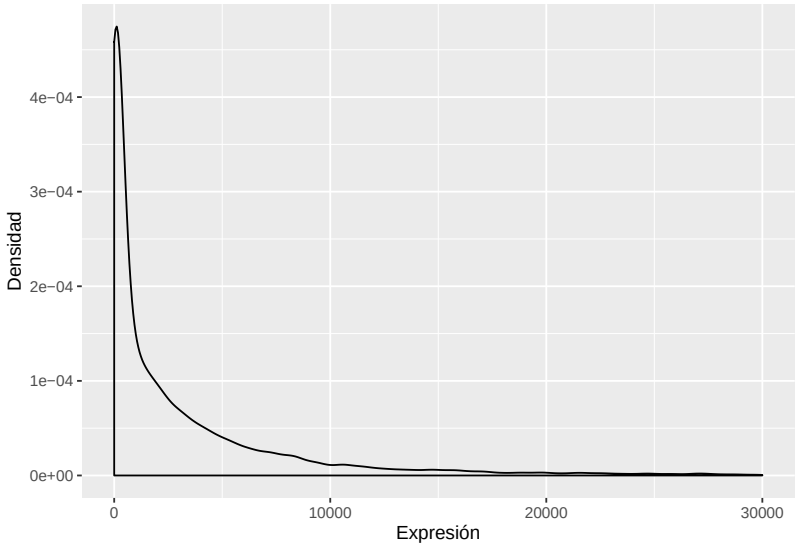


Figura 2.5: Densidad de los valores de expresión de los genes de un sujeto, para datos obtenidos mediante Secuenciación de ARN.

La era de la transcriptómica tuvo su máximo desarrollo con la evolución de las tecnologías de secuenciación de alto rendimiento. La exploración de todo el transcriptoma simultáneamente y a profundidades sin precedentes ha sido posible a partir de estas tecnologías. Esta terminología se refiere a alto rendimiento, en el sentido del paralelismo en la secuenciación, lo que permite investigar millones de fragmentos de ADN en una sola ejecución. Su rápida expansión se justifica por su versatilidad, que ha permitido el estudio de experimentos complejos a escalas hasta ahora inalcanzables e incluso en organismos nunca antes estudiados.

Con el desarrollo de este tipo de tecnología no solo se logró obtener información de expresión a nivel de **genes**, sino que al tener las cadenas de nucleótidos, es posible extraer información de expresión a niveles biológicos menores que genes, como ser **transcriptos**, **exones**, **isoformas**, etc.

2.2. Repositorios de datos de expresión

Para los diversos análisis realizados en la presente tesis, nos centramos principalmente en datos de cáncer de mama. Sin embargo, también se utilizaron repositorios con datos de cáncer de próstata. Un detalle de las bases de datos de libre acceso utilizadas se puede observar en la Tabla 2.1. A lo largo de este trabajo se utilizaron un total de 25 matrices de expresión de cáncer de mama y cuatro de cáncer de próstata, provenientes de microarreglos de ADN.

Adicionalmente se utilizaron datos de cáncer de mama provenientes del proyecto "el atlas del genoma del cáncer" (TCGA; del inglés *The Cancer Genome Atlas*). El proyecto TCGA provee, para una misma muestra, niveles de expresión provenientes de diversas fuentes ómicas, de este proyecto se utilizaron matrices de expresión de microarreglos de ADN, de proteínas medidas mediante iTRAQ, y de genes medidos mediante secuenciación de ARN. Vale la pena aclarar que no necesariamente para toda muestra perteneciente al TCGA se cuenta con datos provenientes de las tres tecnologías, en este sentido, se cuenta con 97 sujetos con muestras de las tres tecnologías.

Tabla 2.1: Bases de datos utilizadas en la presente tesis. Nombre de la base de datos; tipo de cáncer que presentan los sujetos; tecnología de obtención de información biológica utilizada; cantidad de genes presentes; cantidad de sujetos; referencias a la fuente de datos.

Nombre	Cáncer	Tecnología	#Genes	#Muestras	Referencias
Camcap	Próstata	Microarreglos	18.718	199	(Ross-Adams et al., 2015)
Grasso			17.289	122	(Grasso et al., 2012)
Taylor			17.950	179	(Taylor et al., 2010)
Varambally			17.043	19	(Varambally et al., 2005)
Mainz	Mama		13.091	200	(Schmidt et al., 2008)
Nki			13.120	337	(Van't Veer et al., 2002)

Nombre	Cáncer	Tecnología	#Genes	#Muestras	Referencias
Transbig			13.091	198	(Chin et al., 2006)
Unt			18.528	133	(Sotiriou et al., 2006)
Upp			18.528	251	(Miller et al., 2005)
Vdx			13.091	344	(Minn et al., 2007)
					(Wang et al., 2005)
Cal			13.091	118	(Chin et al., 2006)
Dfhcc			20.365	115	(Li et al., 2010)
Dfhcc2			20.365	84	(Silver et al., 2010)
Dfhcc3			20.365	40	(Richardson et al., 2006)
Duke2			20.389	160	(Bonnefoi et al., 2007)
Emc2			20.365	204	(Bos et al., 2009)
Eortc10994			13.091	49	(Farmer et al., 2005)
Expo			20.365	353	(Bittner, 2005)
Hlp			19.985	53	(Natrajan et al., 2010)
Irb			20.365	129	(Lu et al., 2008)
Lund2			12.288	105	(Saal et al., 2007)
Maqc2			13.091	230	(Shi et al., 2006)
Mccc			19.949	75	(Waddell et al., 2010)
Mda4			13.091	129	(Liedtke et al., 2008)
					(Hess et al., 2006)
Msk			13.091	99	(Minn et al., 2005)
Nccs			13.091	183	(Yu et al., 2008)
Pnc			20.365	92	(Dedeurwaerder et al., 2011)
Stk			18.528	159	(Pawitan et al., 2005)
Unc4			17.779	305	(Prat et al., 2010)

Nombre	Cáncer	Tecnología	#Genes	#Muestras	Referencias
TCGA			16.207	547	(Weinstein et al., 2013)
TCGA		ARN	19.948	547	(Weinstein et al., 2013)
TCGA		iTRAQ	10.625	105	(Weinstein et al., 2013)

2.3. Condiciones experimentales

Como se mencionó en la Sección 1.2, para llevar a cabo el AF, es necesario contar con dos condiciones de interés a contrastar. Es decir, dos condiciones para las cuales encontrar aquellos mecanismos biológicos que las diferencian.

2.3.1. Cáncer de mama

Para cáncer de mama, Perou et al. (Parker et al., 2009) desarrolló un clasificador - PAM50 - que, a partir de datos de microarreglos de ADN, asigna a cada sujeto en uno de 5 subtipos: Luminal A, Luminal B, Her2, Basal ó Normal. Dicha clasificación se basa en los niveles de expresión detectados, para cada sujeto, en 50 genes específicos (Parker et al., 2009). Contrastar de a pares estos grupos resulta de un alto interés biológico ya que se sabe que cada grupo es diferente al resto en aspectos como tiempo de sobrevida, reacción a distintas drogas, entre otros. Y por ende, los términos biológicos característicos de cada grupo PAM50 son los que determinan su comportamiento. Para cada sujeto se obtuvo su clasificación PAM50 mediante el paquete de R `genefu` (Gendoo et al., 2015).

2.3.2. Cáncer de próstata

En el caso del cáncer de próstata, ya que contamos con solo cuatro bases de datos, y una de ellas con solo 19 sujetos, no se contrastaron subtipos de la enfermedad. Para

este tipo de cáncer se contrastaron aquellas muestras provenientes de tejido tumoral contra provenientes de tejido normal. De aquí se desea detectar aquellos términos biológicos que caractericen el desarrollo de un tumor maligno en comparación a uno benigno.

2.4. Comentarios finales

Gracias al rápido avance de las tecnologías de obtención de expresión biológica disminuyeron notablemente sus costos, y por ende, el aumento de proyectos internacionales con mayor número de muestras y nivel de detalle. La disponibilidad libre de estas fuentes de información biológica crearon oportunidades sin precedentes para estudiar enfermedades humanas. Habiendo grandes cantidades de bases de datos biológicas de diversas poblaciones, como de distintas tecnologías ómicas, la integración de información resulta en una herramienta clave para el estudio de enfermedades (Cleveland, 2001). Sin embargo, este tipo de estudios se viene realizando a nivel de poblaciones individuales.

Resulta fundamental llevar este tipo de análisis a la comparación de diversas poblaciones. Integrando información funcional de diversos repositorios como de distintas fuentes moleculares es posible llegar a una caracterización de cada grupo de interés. Atacando aspectos funcionales activos por uno u otro grupo bajo estudio se logra el desarrollo de terapias personalizadas. Es por ello que resulta fundamental poder realizar una comparación y caracterización de múltiples fuentes de datos y poblaciones a nivel funcional.

Capítulo 3

Análisis Funcional Integrador

3.1. Motivación

La complejidad y heterogeneidad de ciertas enfermedades, como el cancer, demuestran que el análisis de los genes Diferencialmente Expresados (DE) no resulta suficiente para descrifrar el fenómeno biológico subyacente (Reis-Filho & Pusztai, 2011). Por el contrario, resulta solo el punto de partida de un proceso de exploración en el que se buscan patrones utilizando diversas fuentes de información (Goeman & Bühlmann, 2007), un proceso conocido como Análisis Funcional (AF), el cuál fue descripto en el Capítulo 1. Existen principalmente dos enfoques para llevar a cabo el AF: el Análisis de Sobre-Representación (ASR), y la Puntuación Funcional de Clase (PFC) (Manoli et al., 2006; Pavlidis, Qin, Arango, Mann, & Sibille, 2004). Una de las principales críticas al ASR es que requiere de una lista de genes candidatos definida por el usuario, generalmente estableciendo un umbral de corte de los genes DE (Goeman & Bühlmann, 2007; Khatri, Sirota, & Butte, 2012; Manoli et al., 2006; Pavlidis et al., 2004; Tian et al., 2005). Es por ello que los métodos de PFC emergen como una alternativa que supera esa limitación utilizando no solo todos los genes presentes en el experimento, sino que también sus niveles de expresión, donde los genes

son ponderados de acuerdo a alguna métrica relacionada con el fenotipo analizado (Subramanian et al., 2005).

Se han propuesto varios algoritmos tanto de ASR como de PFC (Khatri et al., 2012), cada uno con supuestos y parámetros de entrada diferentes, los cuales pueden conducir a resultados muy diferentes. Adicionalmente, algunas ontologías, como Gene Ontology (Ashburner et al., 2000), organizan sus conjuntos de genes en alguna estructura particular que permite considerar estrategias de penalización adicionales para el AF. Por lo tanto, la selección del algoritmo apropiado y sus parámetros no resulta una decisión trivial para el investigador, y es una problemática que no ha sido abordada analíticamente. Por otra parte, no está claro qué se obtiene con cada método desde un punto de vista de recuperación de la información, si los resultados son independientes del método y sus parámetros o si los métodos son complementarios o igualmente útiles.

Manoli et al. (Manoli et al., 2006) realizó una comparación de ambos enfoques: ASR y PFC. Sin embargo, dicha comparación fue teniendo en cuenta solo 20 conjuntos de genes enriquecidos de los 227 que evaluó, y solo tres bases de datos con 160 sujetos en total. Por otra parte, Pavlidis et al. (Pavlidis et al., 2004) también comparó ambos enfoques, en ese trabajo se utilizaron 41 muestras apareadas de cerebro, y consideró sólo 10 de los conjuntos de genes de los 965 que analizó. Sin embargo, tanto Manoli como Pavlidis obtuvieron resultados inesperados de PFC, ya que evaluaron sólo un algoritmo y con una única parametrización posible, la cual no es comúnmente recomendada por la literatura para el algoritmo utilizado. Es por ello que, para lograr diseñar un mejor enfoque del AF, resulta fundamental un análisis exhaustivo que tenga en cuenta una variedad de algoritmos, de parametrizaciones, una gran cantidad de conjuntos de genes, así como más bases de datos y sujetos.

3.2. Validación de resultados

Uno de los principales inconvenientes en la comparación de métodos de AF es la falta de *gold standards* o de un conjunto de datos de referencia. Como lo establece Khatri y colaboradores (Khatri et al., 2012); para esta situación, el uso de conjuntos de datos biológicos reales es preferible a datos simulados, ya que estos últimos carecen de factores biológicos importantes (Khatri et al., 2012). Para superar este problema, en el presente trabajo proponemos el uso de varios experimentos para evaluar los mismos (y muy contrastantes) fenotipos de cáncer, asumiendo que deberían exhibir perfiles funcionales similares a través de los experimentos. Nuestra hipótesis es que en un meta-análisis que contrasta dos fenotipos con diferencias conocidas, los patrones de enriquecimiento funcional deben encontrarse en consenso compartidos entre todos los conjuntos de datos, independientemente del método utilizado. Por otra parte, las diferencias entre cohortes podrían considerarse como particularidades biológicas que pueden explorarse más a fondo. Por ejemplo, aunque se encontró poca o ninguna superposición de genes entre varias firmas moleculares en diferentes cohortes de pacientes con el mismo fenotipo (Ein-Dor, Zuk, & Domany, 2006), se han reportado funcionalidades comunes en términos de funciones biológicas (Reis-Filho & Pusztai, 2011). Por lo tanto, los resultados del AF deberían ser similares, mostrando un alto consenso entre los conjuntos de datos a pesar de los genes expresados diferencialmente en cada caso.

Adicionalmente, para determinar si los resultados de enriquecimiento están realmente relacionados con la condición bajo estudio, se requerirá una validación manual de la literatura, lo cual suele ser una tarea tediosa para el investigador, por lo que a modo de validación automática, resulta de extrema utilidad la Base de Datos de Toxicogenómica Comparada (BDTC). Utilizando la BDTC (Davis et al., 2014), es posible consultar para un término biológico dado, si el mismo está relacionado o no con una determinada enfermedad. La BDTC se utilizó para, programáticamente, consultar y

verificar si los términos enriquecidos por cada método están relacionados o no con la enfermedad bajo estudio: “cáncer de mama”.

3.3. Datos de entrada

3.3.1. Matrices de expresión

Para el presente trabajo, se analizaron seis bases de datos de cáncer de mama, seleccionadas debido a su fácil acceso ya que se encuentran presentes en el repositorio R **Bioconductor**. Estas bases de datos fueron medidas mediante la tecnología de microarreglos de ADN y son: Mainz, Nki, Transbig, Unt, Upp y Vdx; ver la Tabla 2.1. Para cada una de las bases de datos, se obtuvo el subtipo intrínseco de cáncer de mama de cada sujeto mediante el método PAM50 (Parker et al., 2009) utilizando la librería de R `genefu` (Gendoo et al., 2015) y siendo procesados como sugiere Sørlie et al. (Sørlie et al., 2010). Se mantuvieron únicamente aquellos sujetos clasificados como tipo Basal o Luminal A, obteniendo un total de 741 sujetos, ya que es sabido que ambos subtipos son muy contrastantes. Al comparar sobrevida de sujetos de ambos subtipos, Luminal A suele presentar un mejor pronóstico que Basal (Dai et al., 2015; Ein-Dor, Kela, Getz, Givol, & Domany, 2005; Parker et al., 2009). Por ende, se espera identificar muchos genes que se comporten contrariamente, es decir, desregulados entre ambas condiciones, y que impacten en varios términos (enriquecidos) que se detecten en común a través de las bases de datos. Para cada una de las seis matrices de expresión, se mantuvieron solo aquellos genes que contaban con identificador de genes *Entrez id* válido, y valores de expresión detectados en al menos el 50 % de las muestras de cada subtipo.

3.3.2. Conjuntos de genes

Como Conjuntos de Genes (CG) se utilizaron las tres categorías de GO. Para descargar los CG se utilizó la versión *v3.0.0* del paquete R `org.Hs.eg.db` (Carlson, Falcon, Pages, & Li, 2013), obteniendo un total de 19.693 CG.

3.4. Algoritmos y parámetros comparados

Para determinar si un CG se encontraba enriquecido por un método dado, se utilizaron los criterios por defecto o recomendados de cada método.

3.4.1. Análisis de Sobre-Representación

Dentro de los algoritmos de ASR, una de las herramientas más utilizadas es la plataforma web DAVID (Huang, Sherman, & Lempicki, 2009) -del inglés *Database for Annotation, Visualization and Integrated Discovery*-. La lista de genes DE de cada base de datos fue enviada a la plataforma Web DAVID (en adelante llamado **WD**) a través del paquete R `RDAVIDWebService` (Fresno & Fernández, 2013). Uno de los principales inconvenientes de la plataforma DAVID es que no permite evaluar CG provistos por el usuario, y al momento de realizar este análisis, la versión estable (*v6.7*) contaba con una base de conocimiento actualizada por última vez en el año 2010. Para superar este inconveniente, desarrollamos **RD**, una versión en lenguaje R del algoritmo de DAVID, que nos permite analizar cualquier CG de interés. Además, a diferencia de DAVID y RDAVIDWebService, RD no requiere de comunicación a través de internet. El tercer algoritmo de ASR evaluado es **GOstats** (Falcon & Gentleman, 2007), que fue diseñado específicamente para realizar análisis de enriquecimiento de GO. GOstats penaliza el enriquecimiento de los CG (penalización *elim*) teniendo en cuenta la estructura de grafos de GO. Finalmente, también se evaluó el método **dEnricher** (Fang & Gough, 2014). Este último método posee una representación

interna de GO, y ofrece otro método de penalización adicional al provisto por GOstats (penalización *lea*). Además dEnricher permite seleccionar el test estadístico a utilizar: *Hipergeométrico*, *Binomial* ó *Fisher*.

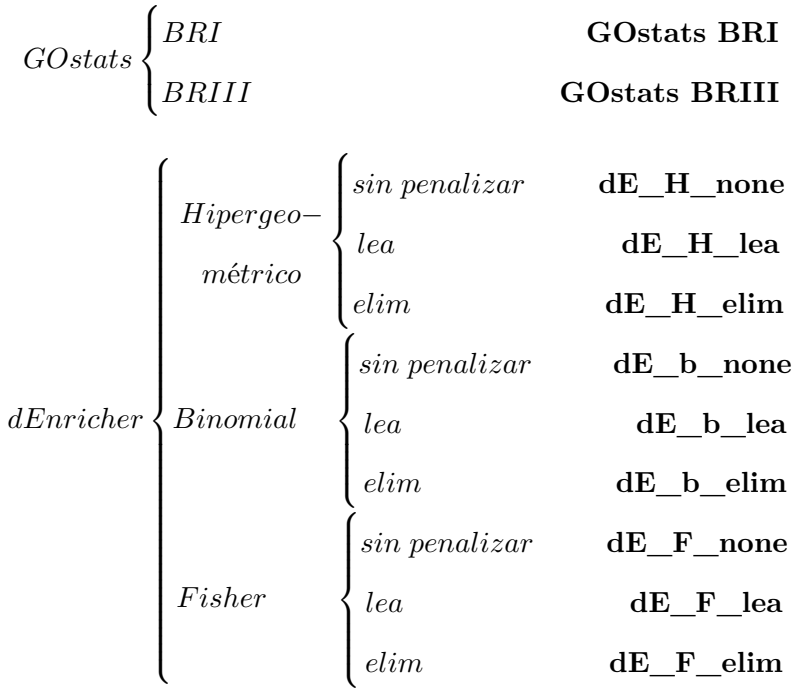
Como se mencionó en la Sección 1.2, los métodos de ASR requieren, además de los CG y la lista de genes candidatos, la lista de genes de referencia (BR; del inglés *Background Reference*). Utilizar diferentes listas de referencia pueden dar lugar a resultados diferentes del método de ASR (Fresno et al., 2012; Rivals, Personnaz, Taing, & Potier, 2007). Por ello, para cada opción de ASR propuesta, a excepción de dEnricher que no lo permite, se evaluaron listas de referencia propuestas por Fresno et al. (Fresno et al., 2012). Las listas de referencia evaluadas fueron: el genoma completo (BRI) y aquellos genes detectados en el experimento (BRIII).

Para cada base de datos, como lista de genes candidatos, se utilizaron los genes DE del experimento. Para obtener los genes DE de cada experimento se utilizó la función `treat` (McCarthy & Smyth, 2009) del paquete R `limma` (Berkeley, 2004), esta función tiene en cuenta aquellos genes que presentan un valor absoluto de Fold-Change mayor a un valor de corte llamado *treatLfc*. Los genes DE fueron aquellos que obtuvieron un p-valor ajustado por FDR $\leq 0{,}01$. Para obtener resultados comparables entre los experimentos, para cada base de datos, se seleccionó un valor *treatLfc* que obtuviera una cantidad de genes DE de alrededor el 5 % de la longitud de su BRIII.

En resumen, los métodos y parámetros analizados de ASR, junto con el identificador a utilizar (en negrita), se listan a continuación:

$$DAVID \begin{cases} BRI & \textbf{WD BRI} \\ BRIII & \textbf{WD BRIII} \end{cases}$$

$$RD \begin{cases} BRI & \textbf{RD BRI} \\ BRIII & \textbf{RD BRIII} \end{cases}$$



3.4.2. Puntuación Funcional de Clase

El método *Gene Set Enrichment Analysis* propuesto por Subramanian et al. (Subramanian et al., 2005) fue de los primeros algoritmos de PFC. El método Subramanian (SM; del inglés de *Subramanian Method*) se incluyó entre los analizados, se utilizó su implementación en lenguaje Java (*v2-2.2.1.1*), ya que la implementación R se encuentra deprecada. El SM puede ser alimentado con un vector de genes prerankeados (**SMpr**) obtenidos a través de alguna métrica adecuada, ó con la matriz de expresión de genes. Para esta última alternativa, la significancia del estadístico *Enrichment Score* (ES) se estima a través de una estrategia de permutaciones sobre las etiquetas de los genes (**SMgp**) o de las condiciones a las que pertenece cada muestra (**SMpp**), generando rankeos permutados de los genes. Para determinar a cada conjunto de genes un p-valor asociado, el SM calcula los ESs aplicando un estadístico de *Kolmogorov-Smirnov* ponderado (Subramanian et al., 2005). Esta ponderación está determinada por un peso w que, por defecto, se establece en 1, pero que tam-

bién puede sustituirse por 0 ó 2. Sin embargo, el efecto del peso w no queda claro ya que los autores sugieren diferentes parametrizaciones de acuerdo con el tipo de dato de entrada (matriz de expresión ó genes pre-rankeados). En el presente trabajo se evaluaron las combinaciones de las diferentes estrategias de permutación con las diferentes alternativas de pesos w. Para el caso de SMpr, ya que debe contarse con un vector de rankeo de los genes, se aplicó la función `eBayes` de la librería R `limma`, obteniendo así para cada gen, el estadístico t y p-valor asociados. Se analizaron tres métodos de rankeo: t, $1-p\text{-}valor$, y $-log(p\text{-}valor)$. Los métodos de PFC suelen filtrar conjuntos de genes previo a realizar el análisis. Este filtrado suele darse dependiendo de la cantidad de genes que componen cada CG. Al evaluar el SM se utilizaron los parámetros de filtrado de CG por defecto, que analiza CG que incluyen entre 15 y 500 genes.

Otra alternativa de PFC evaluada fue la librería R `mGSZ`, la cual se basa en la función de *Z-score* para el cálculo del ES, y en estimación asintótica del p-valor de cada CG, mediante permutaciones de muestras e (implícitamente) de genes (Mishra, Törönen, Leino, & Holm, 2014). El método mGSZ como entrada requiere la matriz de expresión, y rankea los genes mediante el estadístico t moderado (Berkeley, 2004), obtenido utilizando la función `eBayes` de la librería R `limma`. El método mGSZ limita el tamaño de los CG, por defecto, usando un mínimo de 5 genes (sin filtrar por máximo). Adicionalmente, se analizó filtrando CG entre 15 y 500, como en el SM, para obtener una comparación más robusta.

En resumen, los métodos y parámetros analizados de PFC, junto con el identifi-

cador a utilizar (en negrita), se listan a continuación:

$$
\begin{array}{l}
\left\{\begin{array}{l}
\left\{\begin{array}{ll}
w=0 & \textbf{SMgp0} \\
w=1 & \textbf{SMgp1} \\
w=2 & \textbf{SMgp2}
\end{array}\right. \quad \textit{Permutación de genes} \\
\left\{\begin{array}{ll}
w=0 & \textbf{SMpp0} \\
w=1 & \textbf{SMpp1} \\
w=2 & \textbf{SMpp2}
\end{array}\right. \quad \textit{Permutación de condición} \\
\textit{Subra-manian} \left\{\begin{array}{l}
\left\{\begin{array}{ll}
w=0 & \textbf{SMpr tScore 0} \\
w=1 & \textbf{SMpr tScore 1} \\
w=2 & \textbf{SMpr tScore 2}
\end{array}\right. \quad t \\
\textit{Pre-rank} \left\{\begin{array}{l}
\left\{\begin{array}{ll}
w=0 & \textbf{SMpr -log(p) 0} \\
w=1 & \textbf{SMpr -log(p) 1} \\
w=2 & \textbf{SMpr -log(p) 2}
\end{array}\right. \quad -log\ p-valor \\
\left\{\begin{array}{ll}
w=0 & \textbf{SMpr 1-p 0} \\
w=1 & \textbf{SMpr 1-p 1} \\
w=2 & \textbf{SMpr 1-p 2}
\end{array}\right. \quad 1-\ p-valor
\end{array}\right.
\end{array}\right.
\end{array}
$$

$$
mGSZ \left\{\begin{array}{ll}
CG\ filtrado[15,500) & \textbf{mGSZ[15,500)} \\
CG\ filtrado[5,\infty) & \textbf{mGSZ}[\mathbf{5},\infty)
\end{array}\right.
$$

3.5. Resultados

3.5.1. Genes diferencialmente expresados por base de datos

El número de genes DE para cada base de datos analizada se presenta en la Tabla 3.1. Aunque para cada matriz de expresión se contrastaron los mismos subtipos de cáncer de mama, se observa muy poca superposición entre pares de experimentos. Por ejemplo, los datasets Unt y Nki sólo presentan 383 genes DE en común, siendo

que Unt tiene 1.059 genes DE en total (sólo el 36 % se superponen). En general, se encontró que sólo el 12 % (195 genes) de los genes DE en común entre todas los bases de datos (la intersección) se superponen entre la unión (1.678 genes) de los genes DE de todos los datasets evaluados.

Tabla 3.1: Genes diferencialmente expresados entre bases de datos. Primera columna: Los Nombres de las bases de datos, con el umbral de valor absoluto de Fold-Change entre paréntesis. Diagonal principal: Número de genes Diferencialmente Expresados (DE) para cada base de datos (FDR $\leq$ 0,01) y porcentaje del total de genes en el experimento entre paréntesis. Triangular superior: Número de genes DE intersectados por cada par de bases de datos. Nótese que la intersección global de genes DE es sólo 195 de una unión total de 1.678 genes DE.

Nombre	Vdx	Nki	Transbig	Upp	Unt	Mainz
Vdx (0,75)	611 (4,7)	292	465	430	425	412
Nki (0,2)		568 (4,3)	310	374	383	286
Transbig (0,6)			628 (4,8)	448	461	433
Upp (0,3)				932 (5)	632	428
Unt (0,25)					1.059 (5,7)	437
Mainz (0,45)						605 (4,6)
	Intersección:	195		Unión:	1.678	

3.5.2. Análisis de estabilidad de enriquecimiento

Para evaluar la estabilidad de detección de CG enriquecidos a través de las diversas bases de datos, para cada combinación de método/parámetros, se generó un boxplot con el número de CG enriquecidos. La integración de resultados de diversas bases de datos nos permite proporcionar validación inter-estudios tal y como se indica en el análisis de Edelman et al. (Edelman et al., 2006).

Como se puede observar en los boxplot de la Figura 3.1, se aprecia una alta

variabilidad tanto entre los métodos como para las diversas parametrizaciones de un mismo método. Teniendo en cuenta solo aquellos datasets que presentaron almenos un término enriquecido, se encontró una mediana de 284 términos enriquecidos. El SM parece ser muy sensible a diferentes parametrizaciones, así como a la forma en que se le proveen los datos, es decir, a través de la matriz de expresión de genes o mediante una lista pre-rankeada. Curiosamente, para SMpr cualquier valor del factor de ponderación w devolvió casi cero términos enriquecidos. Además, para SMgp y SMpp, la selección de w podría producir enriquecimientos muy diferentes, que van desde cero términos en SMpp0 para Nki o 59 términos en SMgp2 para Vdx hasta valores extremos como 1.019 en SMpp2 para Nki o 474 términos en SMgp0 para Vdx. En particular, el método SMpp presenta comportamientos muy diferentes dependiendo del valor w. Por ejemplo, no se encontró enriquecimiento para $w = 0$, gran variabilidad resultó con $w = 1$ con un Rango InterCuartil (RIC) de 257,57, y se obtuvieron resultados concordantes con $w = 2$, es decir, pequeña dispersión sobre bases de datos (con un RIC de 83,25). Sin embargo, SMpp2 mostró un número extremo de enriquecimientos para una base de datos (1.019 para Nki) y un número muy bajo para otro (101 para Mainz), resultando en dos valores atípicos. Esto podría plantear un problema cuando se analiza un único dataset. Para el SMgp, el enriquecimiento es bastante estable en todos los datasets, con RICs de 105, 91,25 y 98,5 para SMgp0, SMgp1 y SMgp2, respectivamente, pero se obtuvo un número decreciente de enriquecimientos a medida que w aumentaba de 0 a 2, lo que arroja una mediana de 611,5, 293 y 121, respectivamente. Para $w = 1$ el número de términos enriquecidos es similar a la mediana general (284 términos enriquecidos) de los diferentes métodos, lo que sugiere que el SMgp1 se podría considerar una configuración más apropiada.

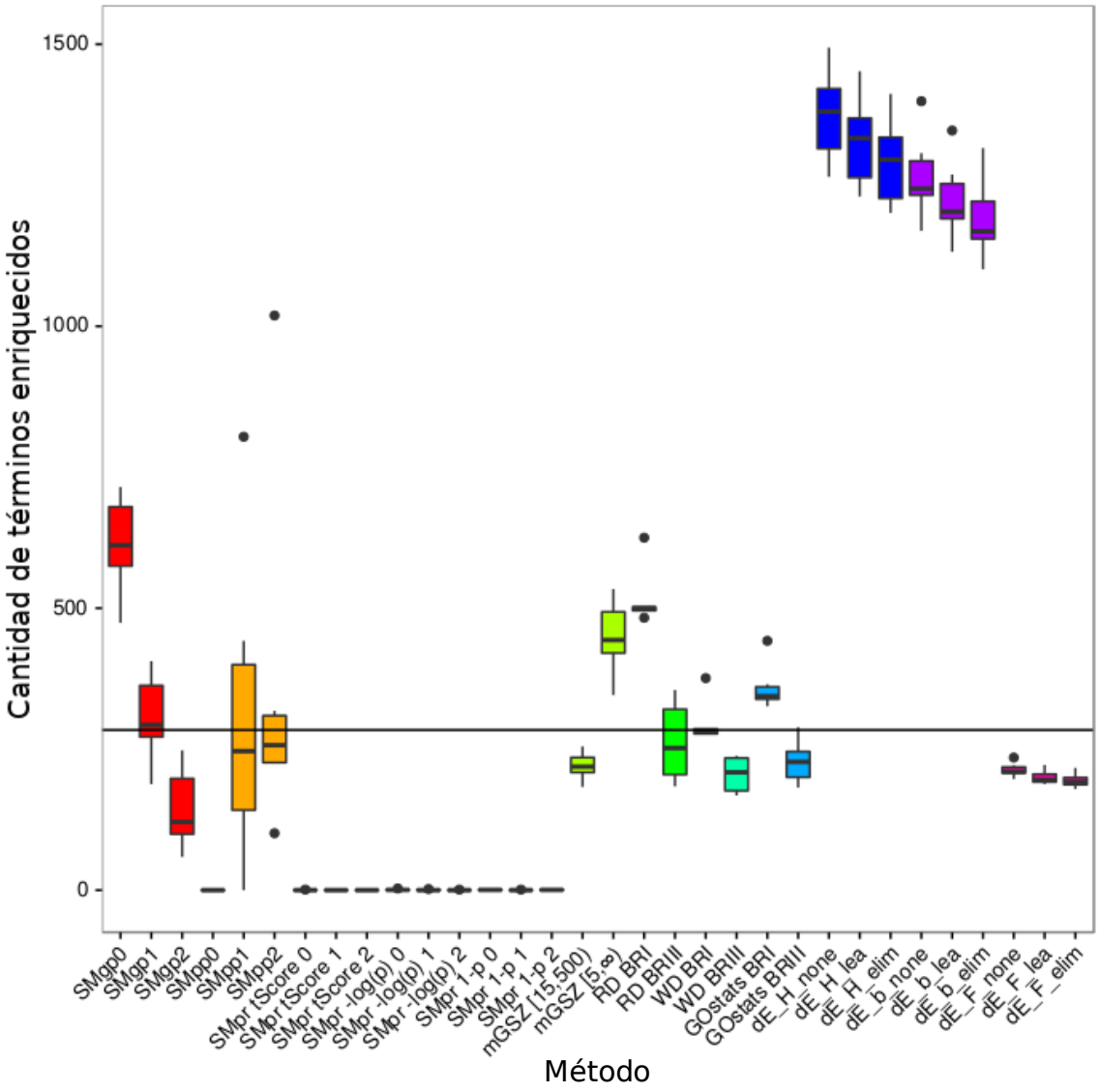


Figura 3.1: Boxplot del número de términos enriquecidos por cada método para las diferentes bases de datos. Las siglas de los métodos se describen en la Sección 3.4. Notar que las alternativas de SMpr enriquecen casi ningún conjunto de genes, mientras que la mediana de términos enriquecidos es de 284 para los métodos restantes (línea negra horizontal). El método SMpp1 obtuvo el número más variable de términos enriquecidos. A excepción de dEnricher con prueba de Fisher, todas las demás combinaciones de método/parámetro de dEnricher devolvieron valores extremos.

En el caso de mGSZ, los resultados mostraron una estabilidad adecuada entre los datasets (RICs de 26,5 para mGSZ$[15, 500)$ y 73 para mGSZ$[5, \infty)$), produciendo un

número muy similar de términos enriquecidos entre las bases de datos. Este método resultó sensible al tamaño de los CG analizados. Cuando el filtro de CG a analizar se estableció entre $[15, 500)$, se logró un número bastante conservador de términos enriquecidos, es decir, número menor de CG enriquecidos. Estos resultados fueron más estables en comparación con el valor obtenido con los límites de filtro de CG de $[5, \infty)$. Vale la pena mencionar que los términos enriquecidos obtenidos con filtro $[15, 500)$ fueron contenidos en su mayoría (93 % en promedio) por el método con el filtro de CG por defecto. Estos términos enriquecidos adicionales generalmente contenían un menor número de genes, es decir, términos más específicos que pueden ser mucho más útiles para descifrar el fenómeno biológico bajo estudio. Basándonos en este concepto, en adelante, hemos seguido las recomendaciones del autor de mGSZ en cuanto al filtrado de conjuntos de genes de $[5, \infty)$.

A excepción de las diversas parametrizaciones de dEnricher, los resultados de las alternativas de ASR fueron bastante similares entre sí, produciendo enriquecimientos bastante estables en todos los datasets: RICs de 6,25 para RD BRI; 115,75 RD BRIII; 8,75 WD BRI; 58 WD BRIII; 21,75 GOstats BRI; y 45,5 GOstats BRIII. La configuración RD BRIII mostró una mayor variabilidad que su contraparte de WD BRIII, probablemente porque se analizó un mayor número de CG (76 % más CG en promedio). Para cada alternativa de ASR, los enriquecimientos obtenidos con BRIII estaban en general contenidos en los obtenidos utilizando BRI (99 % de los términos en promedio para RD y WD; y 86 % para GOstats) en concordancia con la observación de Fresno et al. (Fresno et al., 2012).

En el caso de dEnricher, las pruebas Hipergeométrica y Binomial arrojaron un número extremo de términos enriquecidos en comparación con la mediana general de los demás métodos, obteniendo más de 1.168 términos enriquecidos, independientemente del algoritmo de penalización aplicado. Por otra parte, la prueba de Fisher arrojó resultados bastante estables, en comparación con los otros métodos de ASR,

con valores medios de 210, 195,5 y 191,5 para dE_F_none, dE_F_lea y dE_F_elim, respectivamente; con RICs inferiores a 13,75. Los términos enriquecidos obtenidos con los algoritmos dE_F_lea y dE_F_elim fueron 100 % contenidos en los obtenidos con dE_F_none; más aún, el 86 % de estos términos enriquecidos adicionales encontrados por dE_F_none estaban relacionados con el cáncer de mama cuando fueron consultados en la BDTC. Para los siguientes análisis se descartaron las combinaciones de parámetros de dEnricher, excepto dE_F_none.

Sólo aquellos métodos y configuraciones que devolvieron un número concordante de términos enriquecidos entre datasets, y alrededor de la mediana general de los métodos (284 términos enriquecidos) fueron considerados para los siguientes análisis, es decir, SMgp1, SMpp2, mGSZ$[5, \infty)$, RD BRI, RD BRIII, WD BRI, WD BRIII, GOstats BRI, GOstats BRIII y dE_F_none.

3.5.3. Análisis de profundidad en Gene Ontology

Gene Ontology está organizado en tres grafos acíclicos dirigidos (árboles: "funciones moleculares", "procesos biológicos" y "componentes celulares"), donde un nodo hijo representa un término más específico biológicamente que su padre. Para explorar la especificidad biológica sobre la estructura de GO, se calculó el número mínimo de ramas entre el nodo y la raíz (la profundidad) de cada término enriquecido. Luego, se calculó una tabla de frecuencias del número de términos enriquecidos por profundidad (agrupando los resultados para cada base de datos).

El porcentaje de términos enriquecidos agrupados por profundidad para cada método se muestra en la Figura 3.2. Todos los métodos tienden a explorar profundidades en su mayoría entre tres y seis. Los métodos mGSZ, dE_F_none, SMpp2 y GOstats BRIII enriquecieron el mayor número de términos específicos, es decir, nodos más profundos u hojas, con profundidades > 6: 13 %, 12,8 %, 10,7 % y 9,8 % del enriquecimiento total, respectivamente. Además, WD y RD proporcionaron mayor

enriquecimiento de términos generales, es decir, en profundidades < 3, nodos más cercanos al nodo raíz de cada árbol de GO: 13,7 % para RD BRI, 11,9 % para WD BRI, 9,7 % para WD BRIII y 8,8 % para RD BRIII. Dentro de los métodos de ASR, proporcionalmente se enriquecen más términos cerca de la raíz cuando se usa BRI que BRIII.

Para cada método de ASR, excepto dEnricher, BRI enriqueció un número mayor de términos que BRIII. Sin embargo, como se mencionó anteriormente, tanto en WD, RD - como en GOstats - la mayoría de los términos enriquecidos por BRIII también fueron enriquecidos por BRI. Como se discutió en Fresno et al., usar BRIII, estadísticamente, tiene más sentido que usar BRI; por lo tanto, su uso se sugiere y se usa de aquí en adelante.

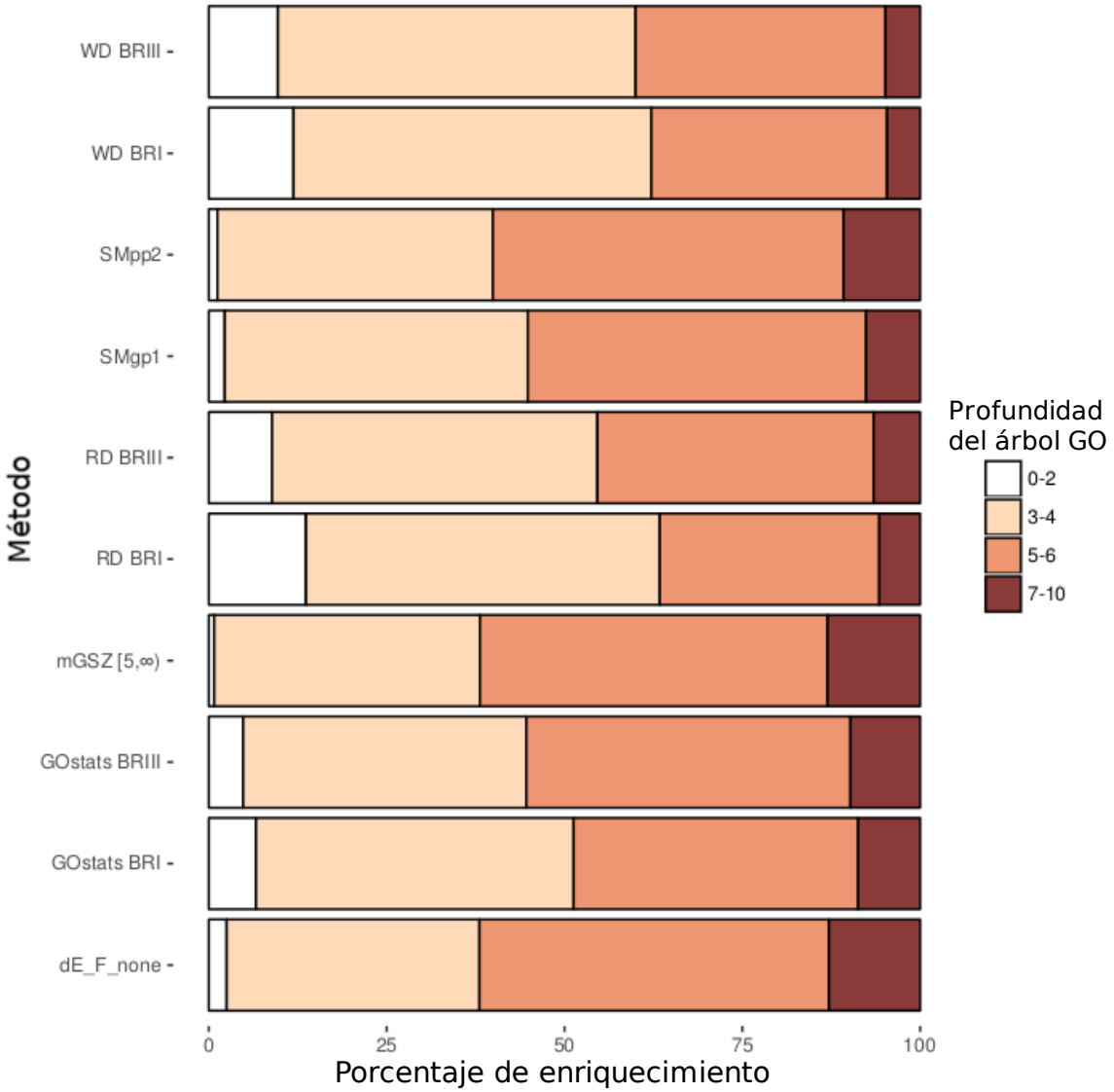


Figura 3.2: Profundidad de enriquecimiento de Gene Ontology (GO) para cada método. Los colores más oscuros representan términos más profundos de la jerarquía del árbol GO. Note que todos los métodos tienden a enriquecer las profundidades en su mayoría entre tres y seis. Los métodos WD y RD enriquecieron los términos más superficiales de la estructura de árbol de GO. Por otro lado, mGSZ, dE_F_none, SMpp2 y GOstats BRIII enriquecieron los términos más profundos.

3.5.4. Análisis de consenso

Para evaluar la hipótesis de que los fenotipos comparados deberían presentar perfiles de enriquecimiento similares en todas las bases de datos, se construyó una matriz

de enriquecimiento $E = e_{mdt}$. En esta matriz, cada fila contiene un término GO y cada columna contiene una combinación de método/parámetros por base de datos. Donde cada celda e_{mdt} de la matriz se definió como sigue:

$$e_{mdt} = \begin{cases} 1 & si\ m\ \textbf{enriqueció}\ t\ en\ la\ base\ de\ datos\ d \\ 0 & si\ m\ \textbf{no enriqueció}\ t\ en\ la\ base\ de\ datos\ d \\ NA & si\ m\ \textbf{no analizó}\ t\ en\ la\ base\ de\ datos\ d \end{cases}$$

donde $m = \{1, \ldots, M\}$ combinación de método/parámetros; $t = \{1, \ldots, T\}$ un conjunto de genes; $d = \{1, \ldots, D\}$ una base de datos.

Aquellos términos que no fueron enriquecidos en ningún conjunto de datos fueron eliminados. Usando la librería R `vegan`, se aplicó un agrupamiento jerárquico a las filas y columnas de E a través de la distancia Jaccard y enlace promedio para agrupar automáticamente perfiles de enriquecimiento similares. En la Figura 3.3 se observa el *heatmap* resultante de la matriz E. El dendrograma superior muestra que los conjuntos de datos analizados tienden a agruparse según el método aplicado, excepto para SMpp2, GOstats y RD, que presentan un conjunto de datos disperso. También puede observarse una diferenciación entre los resultados obtenidos por ASR y PFC, es decir, los métodos de PFC forman un cluster separado, mientras que los métodos de ASR se dividen en dos sub-clusters principales: el de RD/WD y el de GOstats/dE_F_none. Como era esperado, puede verse un subconjunto de términos enriquecidos por casi todos los métodos a través de los conjuntos de datos (filas etiquetadas como **A**). El 64 % de términos enriquecidos en al menos el 80 % de los conjuntos de datos para cada método también fueron enriquecidos por mGSZ en la misma proporción. Esto sugiere que, hasta cierto punto, todos los métodos tienden a proporcionar la misma información. Sin embargo, cada método también proporciona términos exclusivamente enriquecidos (filas etiquetadas como **E**). La comparación de los métodos de PFC y ASR muestra que RD y WD tienden a enriquecer algunos términos que no se en-

riquecen por ningún otro método; lo mismo sucede con dE_F_none y con mGSZ, lo que sugiere que PFC y ASR se complementan entre sí. Los métodos dE_F_none, GOstats y RD, así como mGSZ, analizan más términos que los demás métodos (un número menor de $e_{mdt} = NA$). Además, el 63 % de los términos enriquecidos en al menos el 80 % de los conjuntos de datos por los SM también fueron enriquecidos por mGSZ en la misma proporción, mientras que el 47 % de los términos enriquecidos por GOstats, RD y WD fueron enriquecidos por dE_F_none, lo que sugiere que mGSZ y dE_F_none pueden ser utilizados como métodos de referencia para PFC y ASR, respectivamente.

Dado que todos los experimentos comparan los mismos fenotipos de cáncer de mama, se esperaba encontrar resultados concordantes entre los conjuntos de datos para cada combinación de método/parámetros. Así, el número de términos enriquecidos en casi todos los conjuntos de datos (Frecuencia de Consenso en Enriquecimiento; FCE), así como los términos que no se enriquecieron en casi todos los conjuntos de datos (Frecuencia de Consenso en No Enriquecimiento; $FCNE$), se utilizaron como indicador de la estabilidad del método. Basado en este supuesto, definimos las métricas de comparación listadas en la Tabla 3.2. La concordancia entre los conjuntos de datos para cada método mostró que mGSZ supera con un 45 % de términos enriquecidos concordantes (FCE), seguido de dE_F_none, RD y WD con un 39 %, SMgp1 con un 36 %, SMpp2 con un 30 %, y GOstats con un 29 %. La concordancia entre los términos no enriquecidos ($FCNE$) obtuvo 91 % para dE_F_none, RD y GOstats, 89 % para WD, 85 % para mGSZ, 83 % para SMgp1, y 82 % para SMpp2. Por lo tanto, todos los métodos parecen tener un alto consenso para los términos no enriquecidos y un bajo consenso para los enriquecidos. Ambos conceptos son importantes a la hora de enfrentarse al AF, ya que no debe perderse ningún término biológicamente significativo, ni debe haber términos enriquecidos incorrectamente. De aquí se desprende que el análisis de PFC utilizando mGSZ demostró ser el método más consensuado,

mientras que dE_F_none y RD para la contraparte de ASR.

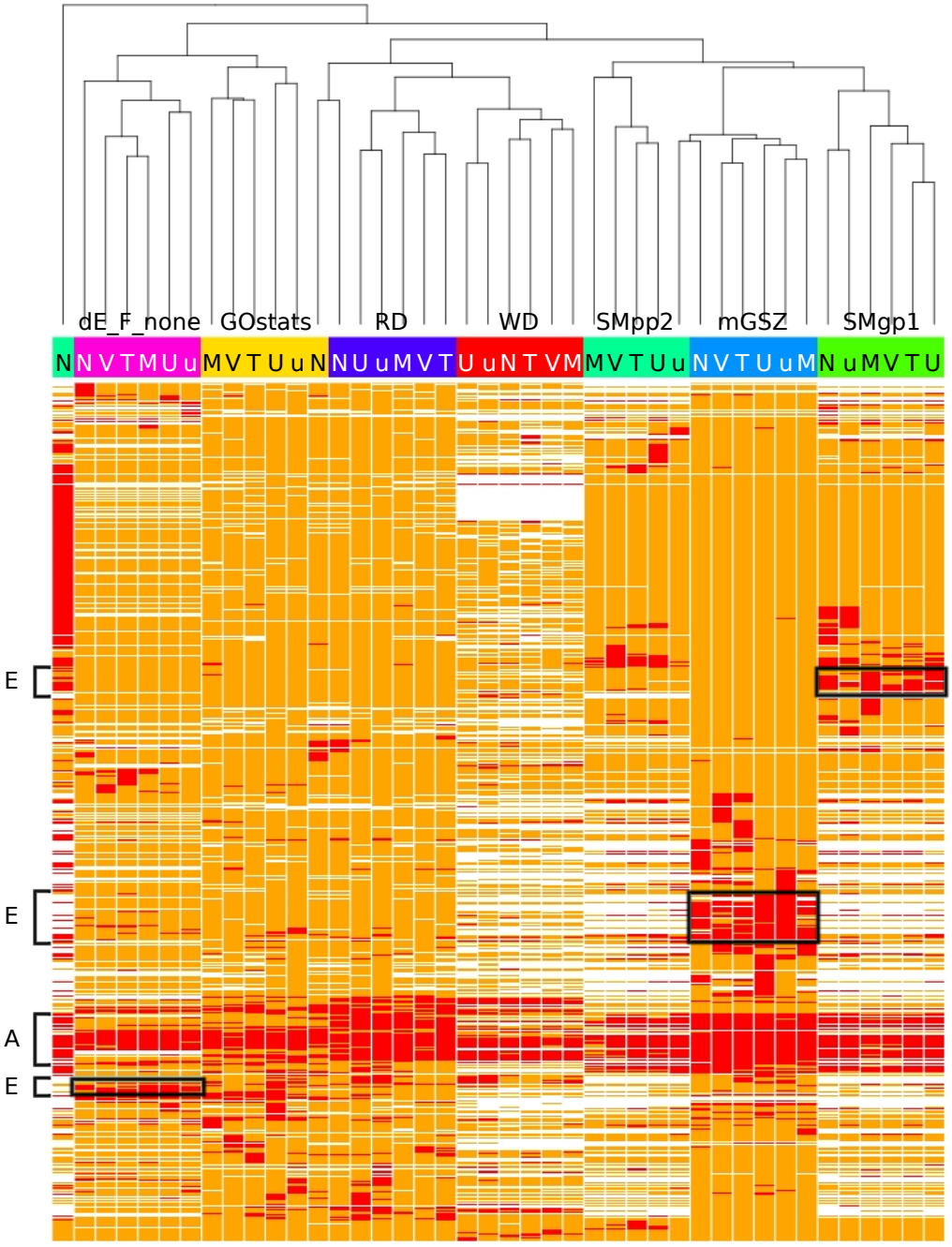


Figura 3.3: Heatmap de enriquecimiento. En columnas, cada combinación de método/parámetros por base de datos; en filas, esos conjuntos de genes (términos) enriquecidos en al menos una base de datos. Notar que los métodos de PFC y ASR se agrupan por separado en el dendrograma. Celdas rojas indican enriquecimiento, naranjas indican que no hay enriquecimiento, y blancas muestran términos que no fueron analizados. Se observan subconjuntos de términos que resultaron enriquecidos por casi todos los métodos analizados (**A**), y subconjuntos de términos enriquecidos exclusivamente por una sola combinación de método/parámetros (para todas las bases de datos; **E**). El color de la etiqueta de cada columna representa el algoritmo utilizado, y la letra representa la letra inicial del conjunto de datos: V Vdx; N Nki; T Transbig; U Upp; u Unt; M Mainz.

Tabla 3.2: Métricas de comparación. $d = \{1, \ldots, D\}$ base de datos seleccionada; $t = \{1, \ldots T\}$ conjunto de genes seleccionado; m: combinación de método/parámetros utilizada; $t_e = 0{,}8$ y $t_n = 0{,}2$ umbrales para el enriquecimiento y el no enriquecimiento, respectivamente; I función indicadora.

Nombre	Definición
Frecuencia de enriquecimiento inter-método	$F_{mt} = \frac{1}{D} \sum_{d=1}^{D} e_{mdt}$
Consenso en Enriquecimiento	$CE_m = \sum_{t=1}^{T} I(F_{mt} > t_e)$
Consenso en No Enriquecimiento	$CNE_m = \sum_{t=1}^{T} I(F_{mt} < t_n)$
No Consenso	$NC_m = \sum_{t=1}^{T} I(t_n \leq F_{mt} \leq t_e)$
Frecuencia de CE	$FCE_m = \frac{CE_m}{CE_m + NC_m}$
Frecuencia de CNE	$FCNE_m = \frac{CNE_m}{CNE_m + NC_m}$

Dado que RD fue capaz de analizar más CG que WD, que puede ser utilizado con CG provistos por el usuario, y no depende de una conexión a Internet, se prefirió sobre WD. Por lo tanto, este último fue excluido de los análisis siguientes.

3.5.5. Enriquecimiento exclusivo y relevancia de términos

Para determinar si los métodos pueden considerarse complementarios o no desde la perspectiva de la información obtenida, se analizaron los términos exclusivamente enriquecidos (TEE) para cada método, es decir, los términos enriquecidos por el método m en el 80 % de los datasets, pero en menos del 20 % de los demás $m' \neq m$, es decir:

$TEE_m = \{t | F_{mt} > t_e \wedge \forall m' \neq m : F_{m't} < t_n\}$ donde t es cualquier conjunto de genes; m alguna combinación de método/parámetros; $t_e = 0{,}8$ y $t_n = 0{,}2$ umbrales para el enriquecimiento y el no enriquecimiento, respectivamente.

Para evaluar la asociación de los términos exclusivamente enriquecidos con el

fenotipo bajo estudio, se evaluó la relación de cada término perteneciente a cada TEE_m utilizando la base de datos dla BDTC (Davis et al., 2014). Utilizando la BDTC, para cada término se pudo determinar su condición patológica en relación con el concepto de "cáncer de mama".

Considerando los términos enriquecidos exclusivamente por cada método (TEE_m), encontramos que mGSZ y dE_F_none fueron los que obtuvieron más TEE_m para PFC y ASR, respectivamente (ver Tabla 3.3). Adicionalmente, mGSZ proporcionó más términos relacionados con el cáncer de mama de acuerdo con la BDTC, así como términos mucho más específicos (profundidad > 6). En el caso del ASR, dE_F_none obtuvo más TEE_m, también relacionados con el cáncer de mama.

Tabla 3.3: Número de términos enriquecidos exclusivos. El número de términos enriquecidos relacionados con el cáncer de mama según la Base de Datos de Toxigenómica Comparativa se presenta entre paréntesis. Notar que mGSZ y dE_F_none enriquecen el mayor número de términos para PFC y ASR, respectivamente.

	Profundidad en los arboles de GO				
Método	0-2	3-4	5-6	7-10	Total
SMpp2			2 (0)	1 (1)	3 (1)
SMgp1	2 (1)	15 (9)	7 (4)		24 (14)
mGSZ		26 (13)	27 (12)	8 (3)	61 (28)
dE_F_none	1 (1)	4 (3)	5 (3)	3 (1)	13 (8)
RD	4 (0)	2 (1)			6 (1)
GOstats		1 (0)			1 (0)

3.5.6. Análisis Funcional Integrador

Como resultado de este estudio exhaustivo (Rodriguez, González, Fresno, & Fernández, 2015; Rodriguez et al., 2016a, 2016b; Rodriguez, Prato, Llera, & Fernández, 2017), en la Figura 3.4 se presenta el pipeline del análisis funcional integrador (IFA; del inglés *Integrative Functional Analysis*). El IFA proporciona como una herramienta de software, cuyo código R se encuentra en el repositorio `https://github.com/jcrodriguez1989/IFA`. Para utilizar la función principal del IFA, el usuario debe pro-

porcionar la matriz de expresión, y la especificación de los fenotipos de cada sujeto. En caso que no se proporcionen los CG, IFA utilizará los CG de GO más actualizados provistos por la librería R `org.Hs.eg.db`. La lista de genes DE y su rankeo son calculados por la herramienta IFA mediante un modelo lineal utilizando la librería de R `limma`, con los cuales se llevan a cabo los análisis mGSZ y de dE_F_none. De este modo, IFA proporciona un enfoque sencillo, unificado y global para el AF.

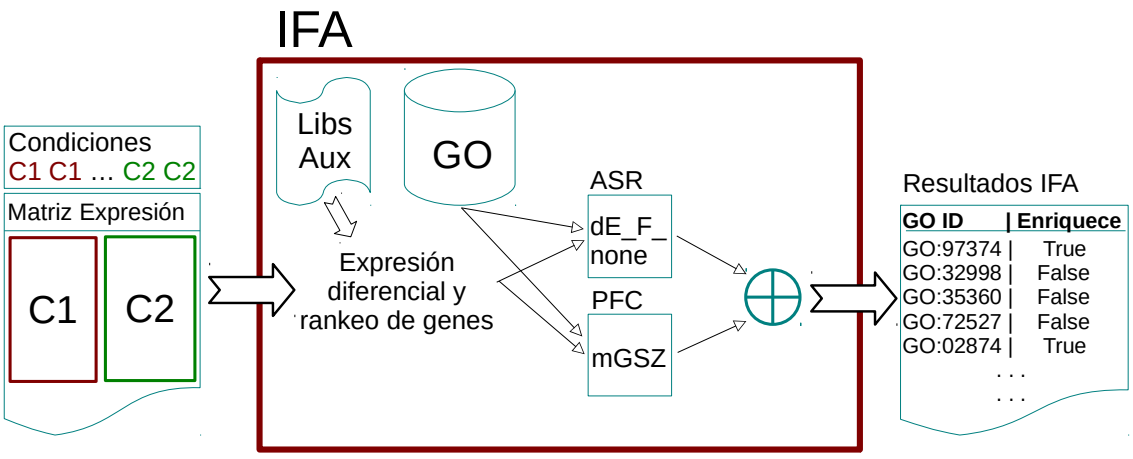


Figura 3.4: Flujo de trabajo del Análisis Funcional Integrador (IFA). El usuario proporciona la matriz de expresión y las correspondientes etiquetas de fenotipo como entrada. El IFA utiliza librerías auxiliares de R para obtener los genes expresados diferencialmente, rankearlos y realizar análisis de PFC y ASR. Finalmente, los resultados de enriquecimiento obtenidos por el IFA integran tanto los resultados del ASR como los de la PFC.

3.5.7. IFA sobre TCGA

Para demostrar la utilidad del pipeline del IFA, se obtuvo el dataset proveniente de microarreglos de ADN de cáncer de mama del reconocido proyecto TCGA. En este dataset, 86 sujetos resultaron clasificados como tipo Basal y 198 como Luminal A. Se utilizó un valor $treatLfc = 1$ de manera de obtener alrededor del 5 % de los genes DE. Los resultados del IFA para este conjunto de datos se utilizaron como caso de prueba y, por lo tanto, se evaluaron y compararon con los conjuntos de datos

analizados previamente (Tabla 3.1). Como en la Sección 3.5.4, la matriz de consenso de resultados del IFA se presentó mediante un *heatmap* que se puede observar en la Figura 3.5, donde 812 términos resultaron enriquecidos para el conjunto de datos de TCGA. El treinta y tres por ciento de los términos (270) se encontraron enriquecidos también en todos los demás conjuntos de datos (marca **A** en la Figura 3.5), el 43 % (352 términos) se encontraron enriquecidos en más del 80 % de los otros conjuntos de datos, y el 15 % (123 términos) se encontraron enriquecidos exclusivamente en TCGA (marca $\mathbf{E}^*$ en la Figura 3.5). En esta matriz de consenso, 445 términos resultaron enriquecidos en al menos el 80 % de todos los conjuntos de datos, de los cuales 232 (52 %) estaban relacionados con el cáncer de mama según la BDTC. En particular, la proporción media de términos exclusivamente enriquecidos presentes en la BDTC en cada conjunto de datos fue del 42 %, con Nki primero con 170 términos exclusivos (84 en la BDTC), y con Mainz último con 58 términos (30 en la BDTC).

Los términos relacionados con la hormona y el receptor de estrógeno, la transición G1/S del ciclo celular mitótico, la replicación del ADN, la organización del huso mitótico, el desenrollado dual del ADN, la actividad de la histona cinasa, la actividad de la helicasa de hibridación, entre otros, se encontraron enriquecidos, en común, en todos los conjuntos de datos (marca **A** en la Figura 3.5), lo que respalda las diferencias de proliferación entre los subtipos de cáncer de mama de tipo Basal y Luminal A. Además, términos como la actividad de la proteína de señalización del receptor tirosina cinasa, la diferenciación de células madre y otros relacionados con la diferenciación celular, sólo se encontraron en los conjuntos de datos de TCGA (marca $\mathbf{E}^*$ en la Figura 3.5).

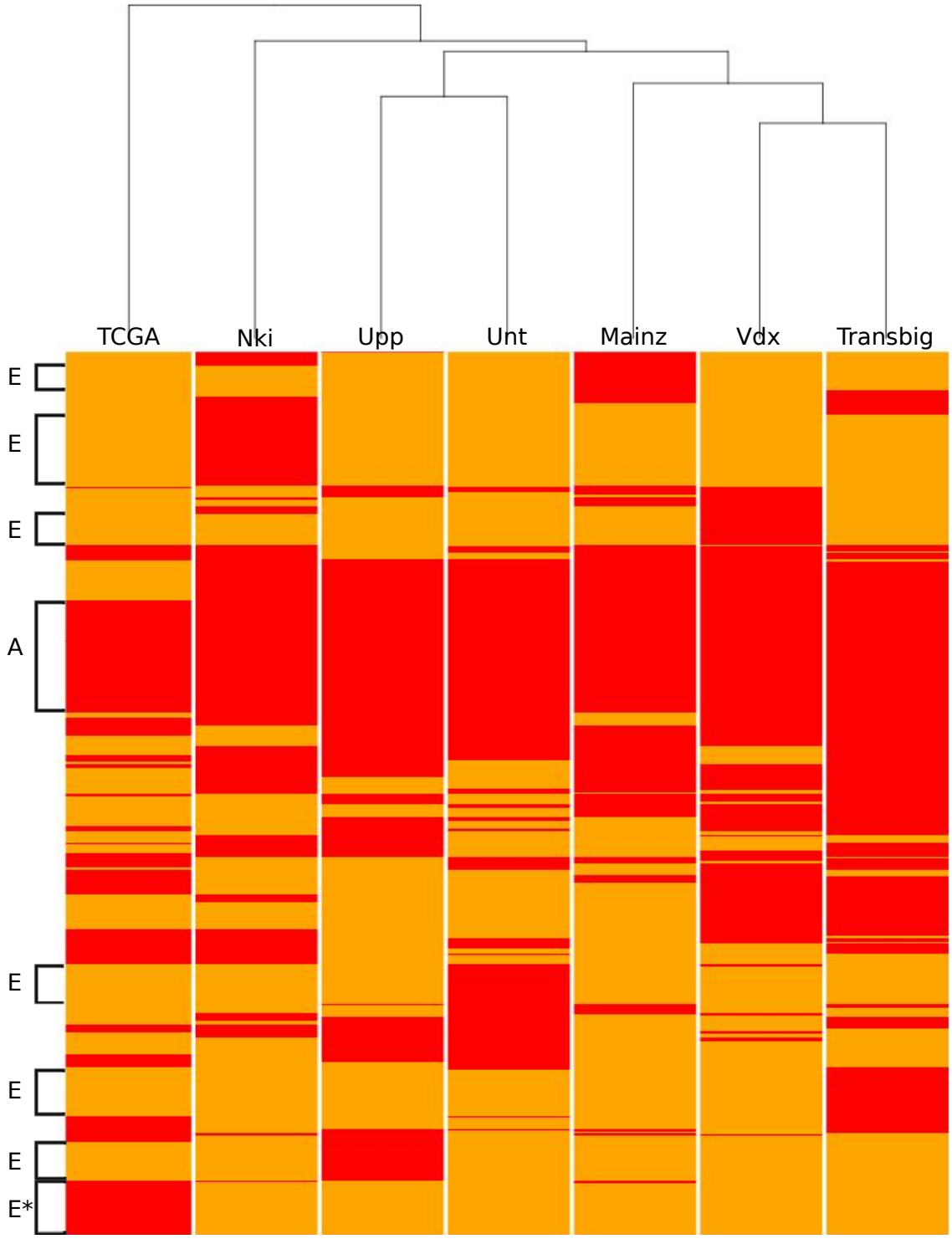


Figura 3.5: Heatmap de enriquecimiento del Análisis Funcional Integrador en cáncer de mama (incluyendo TCGA). Los datasets se ubican en columnas; y los conjuntos de genes (términos), enriquecidos en al menos un dataset, se presentan en filas. Celdas rojas indican enriquecimiento, y las anaranjadas indican que no hay enriquecimiento. Se observan subconjuntos concordantes de términos enriquecidos entre cada conjunto de datos (**A**) y subconjuntos de términos enriquecidos exclusivamente en un solo conjunto de datos (**E**).

3.5.8. IFA sobre datasets de cáncer de próstata

Con el fin de verificar que los resultados del IFA no dependen de la patología analizada, sino que puede extenderse a otros escenarios, el IFA se probó sobre cuatro datasets de cáncer de próstata disponibles en el repositorio R **Bioconductor**. En total, se obtuvieron 519 sujetos de las bases de datos: 74 sujetos con cáncer de próstata benigno y 125 con tumor para Camcap; 29 benignos y 150 tumor para Taylor; 6 vs. 13 para Varambally; y 28 vs. 94 para Grasso. Para obtener alrededor del 5 % de los genes DE, se utilizaron valores de $treatLfc$ de 0,2, 0,15, 0,2, y 0,45 para Camcap, Taylor, Varambally y Grasso, respectivamente. Para el análisis de Varambally, los p-valores de los genes no fueron ajustados, ya que se no se obtuvieron genes DE bajo ningún valor de $treatLfc$ con un valor de corte de FDR fijado en 0,01.

La matriz de consenso resultante se muestra en la Figura 3.5, en la que 163 términos fueron enriquecidos en al menos el 80 % de los conjuntos de datos, de los cuales 99 (61 %) están relacionados con el cáncer de próstata según la BDTC. En particular, la proporción media de términos exclusivamente enriquecidos presentes en la BDTC en cada conjunto de datos fue del 44 %, con Taylor primero con 448 términos exclusivos (194 en la BDTC) y Varambally último con 212 términos (98 en la BDTC).

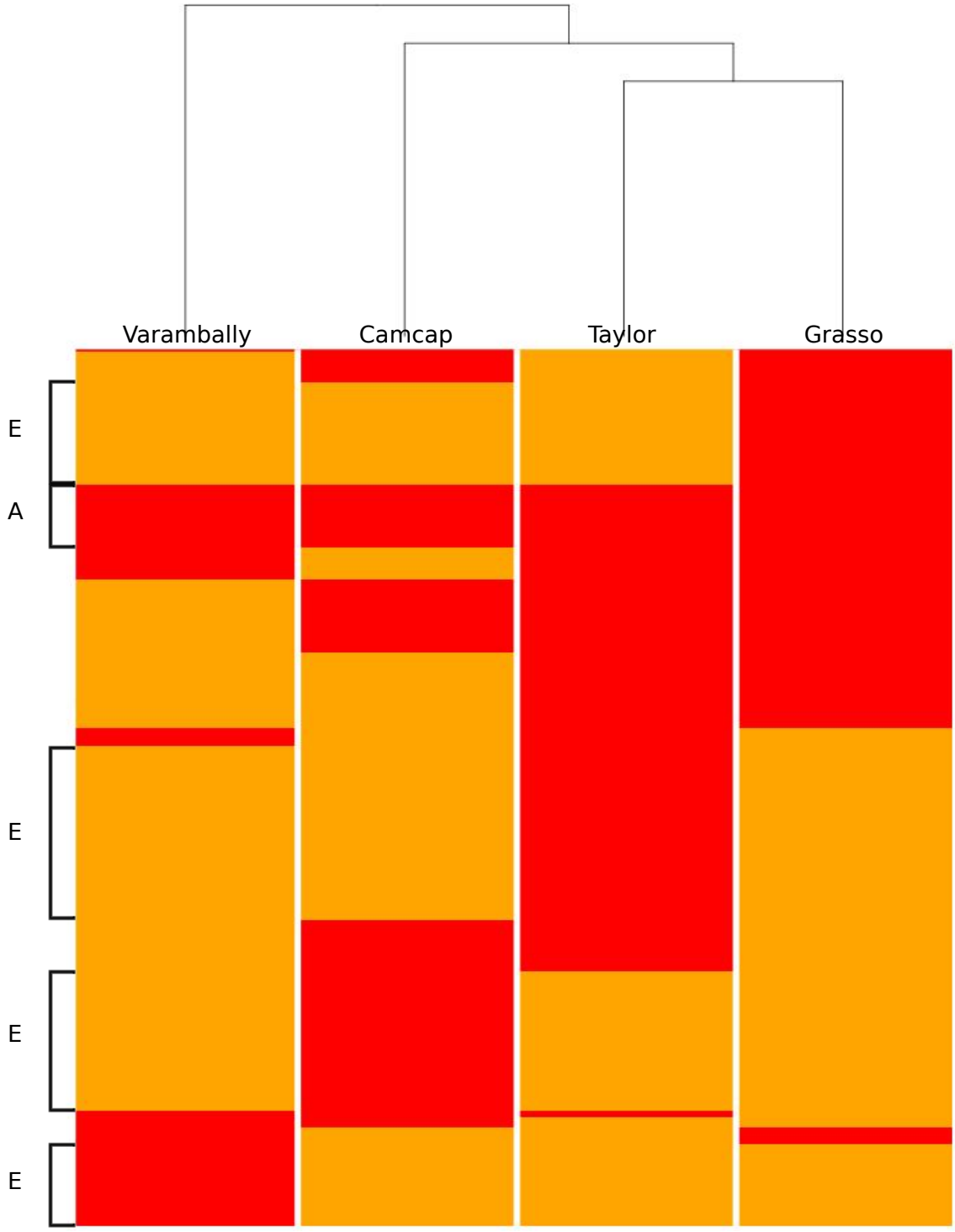


Figura 3.6: Heatmap de enriquecimiento del Análisis Funcional Integrador en cáncer de próstata. Los datasets se ubican en columnas; y los conjuntos de genes (términos), enriquecidos en al menos un dataset, se presentan en filas. Celdas rojas indican enriquecimiento, y las anaranjadas indican que no hay enriquecimiento. Se observan subconjuntos concordantes de términos enriquecidos entre cada conjunto de datos (**A**) y subconjuntos de términos enriquecidos exclusivamente en un solo conjunto de datos (**E**).

3.6. Conclusiones

3.6.1. Comparación de métodos

En este capítulo se pudo mostrar que los resultados del AF pueden varían dependiendo del método y los parámetros utilizados. Esto podría influir negativamente en la interpretación biológica si no se aborda adecuadamente. Por ejemplo, el método de PFC Subramanian mostró una alta sensibilidad a diferentes configuraciones de parámetros y datos de entrada. Los métodos SMpp0 y SMpr, utilizando el valor de corte estadístico recomendado, casi no devolvieron términos enriquecidos para los conjuntos de datos analizados. Estos resultados fueron bastante inesperados porque la naturaleza de los subtipos de cáncer de mama considerados tienen mecanismos biológicos subyacentes muy contrastantes y se han reportado resultados de supervivencia altamente opuestos (Parker et al., 2009). Sin embargo, cuando el usuario sólo tiene una lista ordenada de genes, no hay otra alternativa que la versión pre-rankeada (SMpr) para realizar PFC. En este caso, sugerimos utilizar el rankeo por $1 - p\text{-}valor$ con $w = 1$, y verificar con diversos valores de corte que devuelvan términos enriquecidos. Cuando se alimentó el SM con la matriz de expresión y sus etiquetas de fenotipo, se encontró que tanto las permutaciones de fenotipo como de genes eran muy sensibles al valor de ponderación w elegido, lo que producía diferentes cantidades de términos enriquecidos, así como diferentes niveles de variabilidad de enriquecimiento entre los conjuntos de datos. La permutación del fenotipo con $w = 2$ (SMpp2) parece proporcionar resultados estables, pero la aparición de conjuntos de datos con un número extremo de términos enriquecidos desalienta su uso. Además, cuando $w = 1$ los resultados fueron muy inestables (RIC alto). La estrategia de permutación de genes con $w = 1$ y $w = 0$ (SMgp1 y SMgp0) proporcionó resultados muy estables, pero estas parametrizaciones enriquecieron casi el doble de términos que todos los demás métodos analizados. Cuando se usó $w = 2$ (SMgp2), se obtuvo un número bajo de

términos enriquecidos. En desacuerdo con los autores del SM, recomendamos SMgp sobre SMpp. Sin embargo, estamos de acuerdo con su recomendación sobre el valor de ponderación $w = 1$, es decir, recomendamos SMgp1 sobre cualquier otra configuración de SM.

El método mGSZ, a comparación del SM, fue más estable entre los diversos conjuntos de datos, produciendo un alto consenso entre datasets (alto FCE) y proporcionando el mayor número de términos informativos y exclusivamente enriquecidos (TEE). Se observó que, cuando no se utilizaba límite superior de tamaño de CG (mGSZ$[5, \infty)$), se enriquecieron términos específicos e informativos adicionales a con el límite [15, 500). Por lo tanto, fomentamos el uso de mGSZ$[5, \infty)$. Otra ventaja de mGSZ sobre el SM es que el primero tiene una implementación R actualizada, mientras que el segundo requiere de un entorno Java.

En cuanto a las metodologías de ASR, aunque el uso de BRI produce resultados estables y contiene los términos enriquecidos utilizando BRIII, este último es más apropiado desde un punto de vista estadístico (Fresno et al., 2012). Además, se demostró que BRIII a diferencia de BRI no presenta un número extremo de términos enriquecidos, pero tiene un rango más variable de términos enriquecidos sobre los conjuntos de datos utilizados. La versión R implementada de DAVID (RD) fue desarrollada para funcionar lo más similar posible a la plataforma web de DAVID (WD), con la ventaja de utilizar una base de datos de anotaciones GO actualizada. Aún más, cualquier CG deseado de interés puede ser analizado. Además, RD no requiere una conexión a Internet ni una cuenta registrada en el sitio web de DAVID. En el caso de GOstats, se demostró que es demasiado variable cuando se prueba el enriquecimiento a través de bases de datos (bajo FCE). Esto podría volverse problemático cuando se analiza un solo conjunto de datos, es decir, daría una visión biológica muy limitada del experimento que se está analizando. Para las diversas parametrizaciones de dEnricher, se obtuvieron valores extremos de enriquecimiento cuando se utilizó

test *Hipergeométrico* o *Binomial*; sin embargo, cuando se utilizó la prueba de *Fisher*, se obtuvieron resultados concordantes. Además, al no aplicar ningún algoritmo de penalización (dE_F_none) se obtuvieron términos adicionales relacionados con la enfermedad bajo estudio. El dE_F_none resultó ser el método más estable entre los conjuntos de datos para las alternativas de ASR (el más alto FCE) y superó a sus competidores en cuanto al número de términos informativos enriquecidos exclusivamente (TEE_m).

En resumen, concluimos que si los parámetros se establecen correctamente, el AF obtiene información biológica consensuada y significativa a pesar del bajo nivel de genes DE en común entre diversas bases de datos. Se demostró que tanto el ASR como la PFC proporcionan resultados complementarios que pueden integrarse para obtener una visión biológica más amplia. Además, su integración nos permite abarcar una mayor profundidad de la estructura GO, una característica deseable a la hora de contrastar condiciones experimentales (Fresno et al., 2012). En consecuencia, proponemos utilizar el pipeline del IFA, que realiza análisis simultáneos de ASR y PFC a través de dE_F_none y mGSZ, que resultaron ser los métodos más efectivos respectivamente. Ambos enfoques se basan en el mismo modelo lineal a través de la conocida librería R `limma` (Berkeley, 2004). Por lo tanto, proporciona un marco completo y unificado que utiliza únicamente la matriz de expresión, el diseño experimental y (si no se dispone de CG) la base de datos GO de `org.Hs.eg.db` (Carlson et al., 2013). Aunque este trabajo se basó en los conjuntos de genes GO, cualquier otra base de conocimiento de conjuntos de genes podría ser aplicada para el IFA.

3.6.2. Aplicación del IFA

La aplicación del IFA para estudiar los términos regulados diferencialmente entre los subtipos de cáncer de mama Luminal A y Basal resultó en la detección de varios términos altamente relacionados con el cáncer de mama. Por ejemplo, IFA permi-

tió detectar términos relacionados con las vías de señalización de los receptores de hormonas y estrógenos, los cuales están fuertemente relacionados con los subtipos de cáncer de mama analizados, ya que los sujetos Luminal A son dependientes de estrógeno, mientras que los sujetos Basales no lo son (Bastien et al., 2012; Dai et al., 2015; Goldhirsch et al., 2013). A partir del IFA, los resultados concordantes revelaron términos asociados con el proceso de desenrollado del ADN, un evento asociado con el inicio de la síntesis de ADN y relacionado con la facilitación de la actividad de las helicasas. Se ha reportado que la helicasa BACH1/FANCJ está mutada en el cáncer de mama de inicio temprano, especialmente relacionado con el gen hereditario del cáncer de mama *BRCA1* (Cantor et al., 2001). La mayoría de los cánceres de mama relacionados con el gen *BRCA1* son triplemente negativos y Basales (Atchley et al., 2008), por lo tanto son esperadas las diferencias en los genes que regulan el proceso de desenrollado del ADN entre Luminal A y Basal. Además, el IFA reveló un grupo de términos que sólo estaban regulados de forma diferencial en el conjunto de datos de TCGA. De ellos, la actividad de la histona metiltransferasa en H3-K9 fue uno de los términos más profundos encontrados en la estructura de árbol de GO. La metilación de la histona H3-K9 se ha correlacionado con la formación de heterocromatina y la represión transcripcional, que puede regular la expresión del receptor de estrógeno (Sharma et al., 2005). Además, el término de actividad de la tirosina cinasa está involucrada en la terapia de elusión en cánceres de mama triplemente negativos - Basales - (Scaltriti, Elkabets, & Baselga, 2016). La evaluación de la capacidad de generalización del IFA en los conjuntos de datos sobre el cáncer de próstata también arrojó resultados concordantes e informativos. Además, como era de esperar y visto en el caso del cáncer de mama, se enriquecieron términos presentes en consenso entre los conjuntos de datos, así como términos específicos enriquecidos para cada uno de ellos.

Estos hallazgos apoyan la utilidad de nuestra propuesta desde la perspectiva de

la minería de datos biológicos. Además, el marco de análisis propuesto, IFA, supera las limitaciones de las bases de conocimiento que presentan los métodos, minimiza los parámetros a definir por el usuario, facilita el AF, permite la comparación de diferentes cohortes de pacientes, como se muestra con los resultados de TCGA y cáncer de próstata, y se proporciona gratuitamente como código abierto R en GitHub.

Capítulo 4

Análisis masivo e integrador de conjuntos de genes

4.1. Motivación

El surgimiento y rápido avance de las tecnologías de secuenciación de alto rendimiento han llevado a la disponibilidad de miles de experimentos en repositorios públicos como son el Gene Expression Omnibus (Barrett et al., 2012) y Array Express (Kolesnikov et al., 2014). Más aún, se han llevado a cabo varios proyectos de gran escala, centrados en la comprensión integrada de la complejidad de las enfermedades humanas, como The Cancer Genome Atlas (TCGA) (Weinstein et al., 2013) y el US-LACRN (Llera et al., 2015) para el estudio del cáncer. Comúnmente, los conjuntos de datos disponibles incluyen diferentes tipos de datos provenientes de múltiples tecnologías ómicas para los mismos sujetos, proporcionando información a diferentes niveles moleculares y creando oportunidades sin precedentes para estudiar enfermedades humanas. La integración de estos datos genéticos, genómicos y proteómicos puede conducir a una mejor identificación de grupos homogéneos de sujetos, así como de patrones biológicos comunes y distintivos entre los grupos. La caracterización de

grupos de sujetos que presentan una misma condición patológica, contribuye al desarrollo de estrategias diagnósticas y de tratamiento más adecuadas (Chen & Snyder, 2013; Kedaigle & Fraenkel, 2018). En este sentido, el auge de la medicina personalizada y la disponibilidad de repositorios de datos moleculares de alto rendimiento han aumentado la necesidad de herramientas adecuadas que permitan a los investigadores traslacionales gestionar y explorar estas fuentes de datos (Canuel, Rance, Avillach, Degoulet, & Burgun, 2014; Fernandez & Casares, 2018).

La integración y el análisis de grandes conjuntos de datos provenientes de múltiples fuentes ómicas se ha convertido en una tarea cada vez más desafiante. Los primeros esfuerzos para identificar patrones distintivos relacionados con fenotipos o contrastes (comparaciones entre pares de fenotipos) específicos se han basado en el análisis de expresión diferencial. Estos esfuerzos condujeron a la identificación de diferentes listas de genes, proteínas u otras características genómicas, con una *ligera superposición* entre los conjuntos de datos (Creighton et al., 2018; Hoadley et al., 2014; Korkola et al., 2007; Meng, Kuster, Culhane, & Gholami, 2014; Metzger Filho, Ignatiadis, & Sotiriou, 2011; Rodriguez et al., 2016a; Verhaak et al., 2010; Weinstein et al., 2013). Por el contrario, cuando la estrategia de análisis se basaba en el enriquecimiento funcional, se identificaron *grandes cantidades* de vías biológicas, funciones moleculares y procesos biológicos en común entre diferentes conjuntos de datos (Creighton et al., 2018; Meng et al., 2014; Rodriguez et al., 2016a). Estos resultados denotan que, a pesar de las posibles diferencias en las características ómicas probadas en varios ensayos, la biología subyacente presenta características similares, lo que lleva a resultados fenotípicos o pronósticos similares. Por lo tanto, el análisis funcional de Conjuntos de Genes (CG) resulta en un mejor enfoque para la integración de múltiples conjuntos de datos.

El principal inconveniente de los análisis de expresión diferencial o de enriquecimiento funcional es la falta de integración de datos de múltiples ómicas. Esto ha

motivado al surgimiento de un campo de investigación, conocido como *Big Omics Analysis*, en el que la integración de tipos de datos dispares se ha vuelto cada vez más importante para capturar la heterogeneidad de los procesos biológicos (Fernandez & Casares, 2018; Liu, Shen, & Pan, 2016). Los enfoques recientes se basan en aprendizaje automático (Shen, Olshen, & Ladanyi, 2009), análisis de redes (Wang et al., 2014), o fusión de patrones (Shi et al., 2017). Estas herramientas se basan principalmente en el análisis de expresión o patrones de coexpresión de un único conjunto de datos, lo que supone una gran limitación, ya que se debe contar con datos de cada ómica para cada sujeto. Además, ninguna de estas herramientas es capaz de analizar el comportamiento funcional de la enfermedad o fenotipo; y mucho menos analizar multiples cohortes (poblaciones). Si bien una herramienta ampliamente utilizada llamada PARADIGM (Vaske et al., 2010) ha sido desarrollada para analizar los aspectos funcionales de múltiples fuentes ómicas, sólo puede proporcionar información a nivel de sujetos individuales, y por lo tanto, omite el análisis de grupos de interés, fenotipos, conjuntos de datos o tecnologías ómicas.

El análisis de grandes fuentes de datos ómicas resulta particularmente útil para el estudio de enfermedades heterogéneas, como el Cáncer de Mama (CM) (Korkola et al., 2007; Kwa, Makris, & Esteva, 2017; Reis-Filho & Pusztai, 2011). En este sentido, Perou et al. (Perou et al., 2000) definió cuatro subtipos intrínsecos de CM -introducidos en la Sección 2.3.1- que presentan diferentes perfiles. Estos subtipos han sido ampliamente evaluados a través del análisis independiente de diferentes fuentes de datos incluyendo genes, expresión de microARN, variación del número de copias, y proteínas (Network & others, 2012). Aunque estos análisis condujeron a la validación del esquema de clasificación de Perou et al., también han dado lugar a más preguntas sobre los propios subtipos. Por ejemplo, se ha descubierto que los sujetos con CM están agrupados en aún más grupos de subtipos (Curtis et al., 2012). Más aún, se han identificado nuevos grupos de CM dentro de los subtipos previamente definidos (Aure

et al., 2017). Asimismo, algunos sujetos con CM no pudieron ser asignados a ningún subtipo existente (Fresno et al., 2016). Por lo tanto, el éxito de la medicina traslacional en CM aún se ve limitado por la capacidad de las herramientas de análisis integradoras y masivas para descifrar los procesos biológicos aún desconocidos que subyacen a cada uno de estos subtipos de CM. Para lograr este éxito, resulta fundamental una herramienta que permita al investigador comparar bases de datos tanto de diversas fuentes ómicas como de diversas poblaciones o incluso enfermedades.

4.2. Adaptación del IFA a otras fuentes de datos

Como se mencionó en la Sección 3.5.6, el pipeline del IFA lleva a cabo sus análisis utilizando los métodos dE_F_none y mGSZ. Para llevar a cabo dichos análisis, la lista de genes Diferencialmente Expresados (DE) y el rankeo son calculados mediante un modelo lineal obtenido por la función `eBayes` de la librería `limma` de R. La obtención de los genes DE se lleva a cabo como un paso anterior a utilizar dE_F_none, sin embargo, para el caso de mGSZ, el rankeo de los genes se realiza internamente por dicha librería.

El ajuste del modelo lineal de `eBayes` asume que los datos de entrada siguen una distribución *Normal*, lo cual, como se vió en la Sección 2.1, se cumple para datos provenientes de microarreglos de ADN y de iTRAQ. Sin embargo, este supuesto no se cumple para datos de Secuenciación de ARN, donde vimos que, al ser datos de conteos, se asemejan a una distribución de *Poisson* o *Binomial Negativa*.

Dado el interés de desarrollar una herramienta que permita comparar, desde un punto de vista funcional, una cantidad masiva de bases de datos, tanto de diversas poblaciones como de una variedad de fuentes ómicas. Es por esto que resultó necesario adaptar el pipeline previo del IFA de modo que permita llevar a cabo el análisis a partir de datos de conteos. En particular, se debió inspeccionar el paquete R `mGSZ` y

realizarle las modificaciones pertinentes sobre la metodología de rankeo de genes.

El IFA utiliza una misma herramienta (y modelo estadístico), tanto para el ASR como para el análisis de PFC, para lograr un análisis unificado. La adaptación del IFA a datos de conteos se realizó manteniendo esta idea de unificación. En este sentido, tanto para la obtención de los genes DE como para rankearlos, se implementó una modificación del pipeline IFA que dependiendo del tipo de datos de entrada utiliza la función `eBayes` (para datos *Normales*) ó `voom` (para conteos). La función `voom` también se encuentra desarrollada en la librería `limma`, y se desarrolló principalmente con el fin de detectar genes DE para datos de conteos, la idea de esta propuesta es poder modelar los datos como si fueran *Normales* mediante mínimos cuadrados generalizados (Law, Chen, Shi, & Smyth, 2014). Para ello se realiza una transformación sobre los datos de conteos de manera de poder corregir la relación media-varianza de los datos, a partir de un ajuste *loess*. De esta manera, a partir de la matriz de conteos se obtiene una matriz de pesos asociada a la corrección necesaria y por consiguiente poder ajustar el modelo lineal mediante mínimos cuadrados generalizados. Como resultado de esta reimplementación, se obtuvo una alternativa del IFA que, de un modo transparente al usuario, permite llevar a cabo el análisis a partir tanto de datos *Normales* como de conteos.

4.3. Herramienta desarrollada

En el presente capítulo se introduce la librería R desarrollada: MIGSA -del inglés *Massive and Integrative Gene Set Analysis*-. MIGSA tiene como objetivo identificar la presencia de patrones funcionales comunes o distintivos de diversos fenotipos. Para el análisis, los términos biológicos, clínicos, o definidos por el usuario (de interés específico) se definen mediante CG. MIGSA permite a los usuarios realizar un análisis comparativo integrador de los datasets combinando diferentes experimentos de

diferentes plataformas ómicas. Nuestra librería supera la pérdida de información producida por la fusión de diferentes fuentes de datos, explota la complementariedad de los diferentes niveles ómicos y permite realizar consultas directas y supervisadas basadas en cada fenotipo.

Como se explica esquemáticamente en la Figura 4.1, MIGSA toma como entrada una lista de matrices de expresión ($L = \{M_d\}; d \in \{1, \ldots, D\}$). Cada matriz contiene datos del d-ésimo dataset, con S_d sujetos medidos a través de cualquier tecnología ómica (por ejemplo, genes, proteínas, etc.) que pueda ser anotada con identificadores *Entrez*. Para el d-ésimo conjunto de datos, MIGSA requiere la identificación del fenotipo de cada sujeto $F_d = [f_{d1}, ..., f_{dS_d}]$, con $f_{di} \in \{a, b, ...\}$ ($i \in \{1, ..., S_d\}$), para identificar cada experimento $E_d^{a,b}$ que resulta de la matriz de expresión M_d y el contraste de fenotipo *a vs. b*. Luego, en paralelo, MIGSA aplica individualmente a cada experimento el Análisis Funcional Integrador (IFA) descrito en el Cápitulo 3. Vale la pena mencionar que el IFA aplicado por MIGSA presenta dos novedades respecto a su publicación original (Rodriguez et al., 2016a): ahora permite analizar datos de conteos (como los de secuenciación de ARN), microarreglos de ADN, y proteómicos de iTRAQ; e identifica los genes específicos que contribuyeron al enriquecimiento de cada CG.

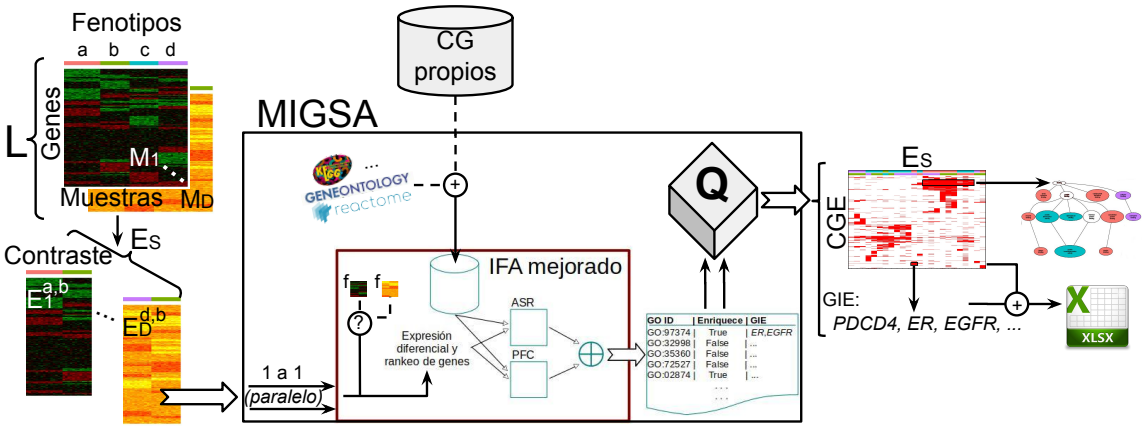


Figura 4.1: Flujo de trabajo del Análisis Masivo e Integrador de Conjuntos de Genes (MIGSA). De una lista de matrices de expresión $L = \{M_1, ..., M_D\}$, cada matriz con la identificación del fenotipo (por ejemplo a, b, c, d) para cada muestra, se realiza la lista de experimentos (Es), donde cada experimento compara un par de fenotipos de una matriz (por ejemplo $E_1^{a,b}$ contrasta los fenotipos a y b para la matriz de expresión M_1). MIGSA toma como entrada esta lista de Es y aplica, individualmente a cada experimento, una versión mejorada del IFA previamente presentado, de manera paralela. Esta versión mejorada del IFA toma una matriz de expresión y, de acuerdo con su tipo de datos, utiliza una función f diferente para obtener tanto la expresión diferencial como el rankeo de los genes, y luego realiza ambos Análisis de Sobre-Representación (ASR) y de Puntuación Funcional de Clase (FCS). Para este paso, se puede utilizar una colección generada por el usuario de conjuntos de genes ó seleccionar entre más de 130 colecciones proporcionadas por MIGSA. Los resultados de cada experimento se almacenan en un cubo de datos Q. Como salida, MIGSA proporciona diferentes alternativas de exploración y visualización, permitiendo identificar fácilmente los Conjuntos de Genes Enriquecidos (CGE) para cada experimento y, para cada conjunto de genes, los Genes Importantes para su Enriquecimiento (GIE).

Los resultados de MIGSA se resumen en un cubo de datos tridimensional, $Q^{Es,CGE,GIE}$, con tres ejes o niveles de abstracción: Experimentos (Es; es decir, contrastes de fenotipos), Conjuntos de Genes Enriquecidos (CGE; CG estadísticamente desregulados para un umbral de significancia), y la lista de Genes Importantes para el Enriquecimiento (GIE; es decir, para cada CG, los genes que fueron los principales responsables de su desregulación en un experimento específico). Así, para

el experimento $E_d^{a,b}$, MIGSA proporciona la lista de conjuntos de genes enriquecidos $CGE_d^{a,b}$, y la lista de genes importantes de enriquecimiento $GIE_d^{a,b}$ (ver Figura 4.1). Nótese que en el cubo Q los ejes son: $Es = \cup_d \cup_{a,b} E_d^{a,b}$, $CGE = \cup_d \cup_{a,b} CGE_d^{a,b}$, y $GIE = \cup_d \cup_{a,b} GIE_d^{a,b}$.

Un gen G será considerado un GIE para un determinado conjunto de genes CG y experimento $E_d^{a,b}$, si $G \in CG$, y al menos uno de:

- CG fue enriquecido por el análisis de sobre-representación, y el gen G se encontró diferencialmente expresado en $E_d^{a,b}$.
- CG fue enriquecido por puntuación funcional de clase, y el gen G conformó el *leading-edge* (Subramanian et al., 2005) de CG, es decir, la posición de G, en el rankeo, se encuentra a la izquierda/derecha de la gráfica positiva/negativa del *Enrichment Score* de CG.

MIGSA proporciona varias herramientas gráficas, que permiten a los usuarios explorar y visualizar fácilmente los resultados de sus datos. Por ejemplo, el primer eje del cubo MIGSA ($Q^{CGE,Es}$) puede representarse fácilmente mediante un *heatmap*. En este caso, los CG se distribuyen en filas y los experimentos en columnas. En particular, las filas y columnas se encuentran ubicadas de acuerdo a un ordenamiento dado por la distancia Jaccard y enlace promedio. El *heatmap* de MIGSA resume eficazmente los resultados de múltiples IFA y ofrece a los usuarios la capacidad de visualizar patrones de enriquecimiento funcional relacionados con contrastes, fenotipos o incluso con fuentes de datos específicas. El *heatmap* incluye un dendrograma superior que muestra los grupos formados por los resultados de todos los experimentos. Este dendrograma permite una rápida identificación de clusters y sub-clusters de experimentos, y su asociación con los fenotipos estudiados.

4.4. Validación de la herramienta

Para demostrar la utilidad de MIGSA como herramienta de descubrimiento, se utilizó con el fin de lograr una caracterización funcional de cada subtipo molecular PAM50 de CM. Se aplicó MIGSA para analizar una gran cantidad de bases de datos de CM donde se evaluaron todas las combinaciones posibles de pares de subtipos. La eficiencia de MIGSA se vería reflajada en encontrar genes y patrones funcionales que caractericen los diversos fenotipos analizados. En este sentido, dado que se analizaron experimentos pertenecientes a una misma enfermedad, será esperable encontrar patrones en común compartidos entre todos los contrastes. Pero también, y quizás de mayor interés biológico, resultará fundamental detectar patrones específicos que caractericen cada contraste, y más particularmente, cada subtipo.

4.5. Datos de entrada

4.5.1. Matrices de expresión

Se analizaron veinticinco conjuntos de datos de expresión génica de cáncer de mama basados en microarreglos de ADN (4.145 sujetos) previamente reportados por Haibe-Kains et al. (Haibe-Kains et al., 2012) e identificados como *microarreglos comerciales* por Fresno et al. (Fresno et al., 2016), estos conjuntos de datos son los de CM presentes en la Tabla 2.1, excepto por los de TCGA. Para cada sujeto de los datasets de Haibe-Kains (HKds), se tomó la clasificación PAM50 basada en el algoritmo `pbcmc`, según se detalla en Fresno et al. El algoritmo `pbcmc` devuelve un nivel de confianza de que un sujeto efectivamente pertenezca a un subtipo o no, de este modo se evita analizar sujetos que posiblemente estén mal asignados por `genefu`, y que por ende, provoquen un sesgo en los resultados. Se conservaron aquellos sujetos clasificados efectivamente como Basal, Luminal A, Luminal B y Her2+, mientras

que los sujetos clasificados como Normal o No Asignados fueron descartados, lo que resultó en un total de 2.756 sujetos analizados.

Además de los HKds, se analizó una población de 97 sujetos que cuentan tanto con datos transcriptómicos (secuenciación de ARN y microarreglos de ADN) y proteómicos (iTRAQ) de TCGA; se seleccionaron los mismos sujetos para cada una de las bases de datos para evitar posibles sesgos. La clasificación PAM50 para estos sujetos se obtuvo a través de la librería `pbcmc` utilizando los datos de microarreglos y la opción de estandarización de escala. Mediante `pbcmc` se obtuvieron 25, 25, 12, 20, 0, y 15 sujetos identificados como Basal, Her2+, Luminal A, Luminal B, Normal, y No Asignados, respectivamente.

Para cada conjunto de datos, se promediaron los genes con múltiples sondas o detecciones. Para cada contraste de dos subtipos, sólo se incluyeron los genes detectados en al menos el 70 % de los sujetos de cada subtipo. Además, para los datos del secuenciación de ARN, sólo se incluyeron los genes con más de 10 conteos en promedio por subtipo. Para un conjunto de datos dado y para cada subtipo dado, si había menos de ocho sujetos presentes, entonces estos sujetos fueron descartados para evitar tener una población no representativa, y en consecuencia no se analizaron los contrastes con ese subtipo para ese conjunto de datos.

Luego de los filtros mencionados, se obtuvieron 118 experimentos que contrastan los subtipos de CM para los HKds y 18 para TCGA. Para el análisis de sobrerepresentación del IFA de MIGSA, se eligieron parámetros de selección de expresión diferencial que incluyan entre el 4 y 6 % de los genes como diferencialmente expresados.

4.5.2. Conjuntos de genes

MIGSA permite realizar el IFA analizando más de 130 bases de datos disponibles de CG conocidos (Kuleshov et al., 2016), como así también utilizar cualquier colección

de CG especificada por el usuario. En este trabajo se obtuvieron (`org.Hs.eg.db` *v3.5.0*) y analizaron las tres categorías de CG de GO (Consortium, 2016), alcanzando 15,796, 1,902, y 4,604 CG de las ontologías de Procesos Biológicos (PB), Componentes Celulares (CC), y Funciones Moleculares (FM), respectivamente. Un CG se consideró CGE si obtuvo un p-valor $\leq 0,01$ por el IFA.

4.6. Estrategias para el análisis de eficiencia

Como primer instancia, se aplicó MIGSA sobre los experimentos de los HKds. Con el cubo resultante de MIGSA, se combinaron los resultados para definir los conceptos:

- **Términos Enriquecidos en Consenso por Contraste** (**TECC**): Para cada contraste, esos CG (y GIE) enriquecidos en al menos $K\,\%$ de los experimentos del contraste (donde K es un nivel definido por el usuario, aquí usamos $K = 50$), es decir, $\boldsymbol{TECC}^{a,b} = \{CG : \sum_{j=1}^{D} I(CG \in CGE_j^{a,b}) \geq \frac{K}{100} N^{a,b}\}$ donde I es la función indicadora, D es el número de datasets (aquí $D = 25$ HKds), y $N^{a,b}$ el es el número de experimentos que contrastan el subtipo a con b.
- **Perfil Transcriptómico Funcional** (**PTF**): Para cada subtipo, aquellos **TECC** (y sus respectivos GIE) presentes en común al contrastar un subtipo contra cada uno de los restantes, es decir, $\boldsymbol{PTF}_a = \cap_{b \neq a} \boldsymbol{TECC}^{a,b}$.

Luego, se analizaron los **TECC** para obtener un perfil funcional para cada contraste de los subtipos de CM. Adicionalmente, se obtuvieron los **PTF** basados en los experimentos de los HKds y se exploraron conjuntamente con los GIE correspondientes, para obtener una caracterización funcional de cada subtipo de CM.

Posteriormente, para demostrar la capacidad del paquete MIGSA de integrar múltiples conjuntos de datos ómicos, se aplicó MIGSA a los conjuntos de datos de TCGA de CM, y los resultados se compararon con los **TECC** definidos con los experimentos de los HKds. Para evaluar la concordancia entre los resultados de MIGSA sobre

los conjuntos de datos de transcriptómica y proteómica del TCGA, se analizaron los CGE (y los GIE) para definir:

- **Términos Enriquecidos por Transcriptómica** (**TET**): Para cada subtipo, aquellos CG enriquecidos consistentemente en experimentos transcriptómicos. Para identificar los **TET**, primero, para cada contraste, se obtuvo la intersección de CGE tanto en los datos de secuenciación de ARN como en los de microarreglos de TCGA. Luego, para cada subtipo, el conjunto de términos presentes en los tres contrastes del subtipo dado se denominó Términos Enriquecidos por Transcriptómica.
- **Términos Enriquecidos por Proteómica** (**TEP**): Para cada subtipo, aquellos CG enriquecidos consistentemente en experimentos proteómicos. Para obtener los **TEP**, primero, se identificaron los CGE encontrados en los experimentos de iTRAQ del TCGA para cada contraste. Luego, se seleccionó el conjunto de términos presentes en todos los contrastes de un subtipo dado.

Finalmente, para cada subtipo, se evaluó la concordancia en la detección de CGE a través de diferentes plataformas y conjuntos de datos ómicos comparando los **TET** y **TEP** del TCGA, con los **PTF** encontrados analizando los experimentos de los HKds. Utilizando MIGSA, se llevo a cabo una inspección de los árboles de GO para encontrar intersecciones entre **PTF**, **TET** y **TEP** para las categorías de PB, FM y CC de cada subtipo.

4.7. Resultados

4.7.1. MIGSA sobre datasets de microarreglos

Se analizaron, con MIGSA, los 118 experimentos de HKds. En la Figura 4.2 se observa el *heatmap* que ilustra el primer eje del cubo de resultados ($Q^{CGE,Es}$), donde

el dendrograma superior revela los clusters formados por los resultados de todos los experimentos que contrastan los subtipos de CM. Al analizar los dendrogramas se nota la conformación de clusters generados principalmente por los subtipos contrastados (no se encontró efecto de lote o plataforma). En particular, el cluster más grande contiene todos los experimentos que contrastan el subtipo Luminal A (marca LA en la Figura 4.2; 63 experimentos/columnas). Dentro de este cluster, se identificaron dos sub-cluster distintos. Uno de estos sub-clusters consiste en 23 de un total de 25 (92 %) experimentos evaluando Luminal A contra Basal (marca LA-Ba). El otro sub-cluster incluyó 18 de 20 (90 %) experimentos contrastando los dos subtipos Luminales (marca LA-LB). Los experimentos que comparan Luminal A contra Her2+ (LA-H) fueron agrupados principalmente en dos clusters que contienen seis y nueve de 18 (83 %) experimentos respectivamente.

El segundo gran cluster divisable en la Figura 4.2 incluye todos los experimentos que contrastan los subtipos Basal con Luminal B ó Her2+ (marca Ba). En este cluster, se identificaron dos sub-clusters, uno con 15 de 20 (75 %) experimentos contrastando Basal con Luminal B (marca Ba-LB), y otro con 12 de 18 (67 %) experimentos contrastando Basal con Her2+ (marca Ba-H). Además de estos dos grandes sub-clusters, se diferencia uno con experimentos que contrastan Luminal B con Her2+ (marca LB-H), con 14 de 17 (82 %) experimentos. En particular, este último cluster tiene la menor cantidad de términos enriquecidos, lo que sugiere una gran similitud entre estos dos subtipos.

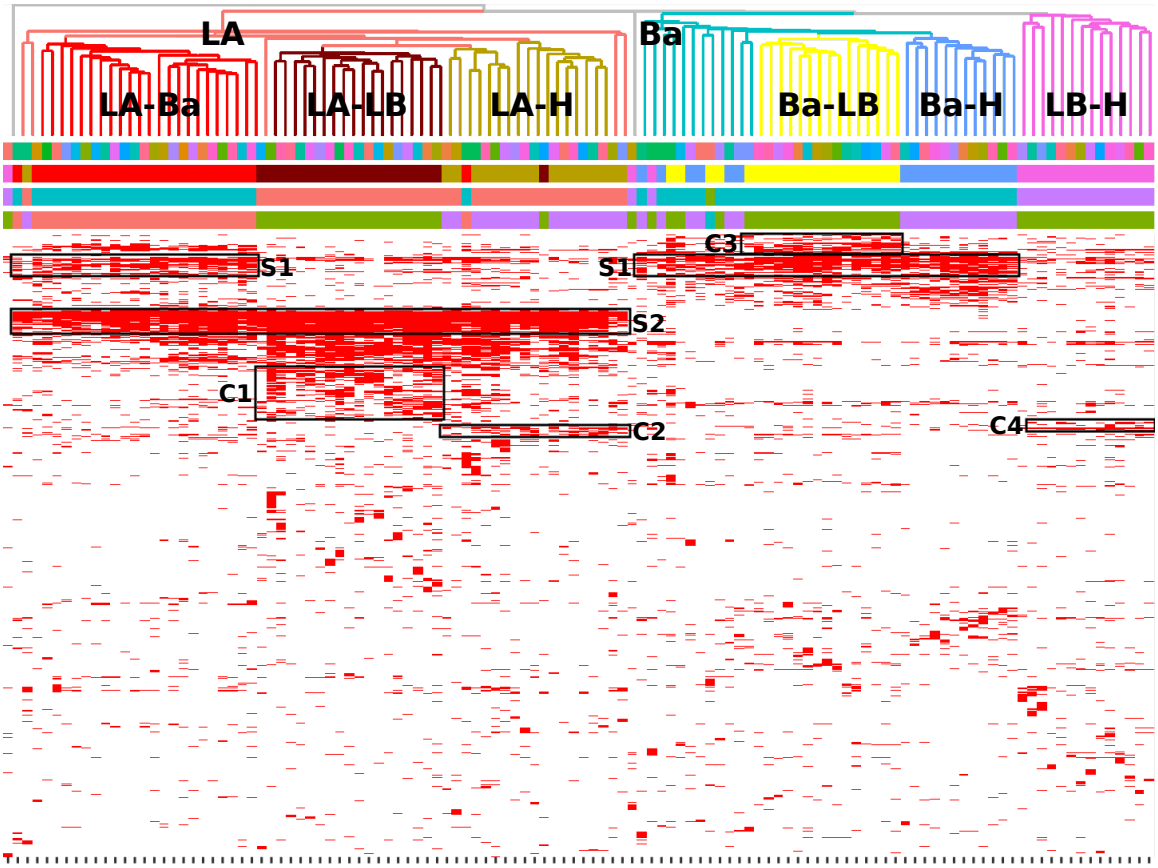


Figura 4.2: Conjuntos de genes enriquecidos por análisis funcional para cada experimento. Heatmap: conjuntos de genes en filas, experimentos en columnas; ordenados por distancia Jaccard y enlace promedio. Conjuntos de genes enriquecidos en rojo. Las casillas S1 y S2 contienen conjuntos de genes enriquecidos en común para subtipos, y las casillas C1, C2, C3 y C4 para contrastes (comparación entre pares de subtipos). Barras de columnas (de arriba abajo): datasets (cada color representa uno de los 25 datasets analizados); contraste; subtipo1 y subtipo2, salmón es Luminal A (LA), turquesa es Basal (Ba), púrpura es Her2+ (H), y el verde es Luminal B (LB). Clusters de dendrogramas: Se demarcan para los subtipos Basal (Ba) y Luminal A (LA), y uno por cada contraste.

El *heatmap* de la Figura 4.2 también reveló patrones de enriquecimiento comunes específicos de subtipos para todos los subtipos contrastados. Por ejemplo, las cajas S1 muestran CGE que se encuentran comúnmente en los contrastes con el subtipo Basal, mientras que la caja S2 indica patrones de enriquecimiento de Luminal A. Además, también se observan patrones de enriquecimiento específicos de contrastes, es decir,

conjuntos de términos que denotan diferencias entre dos subtipos determinados. Por ejemplo, la caja C1 contiene aquellos CG que diferencian Luminal A de Luminal B, la caja C2 para Luminal A vs. Her2+, C3 para Basal vs. Luminal B, y C4 para Luminal B vs. Her2+.

Las capacidades exploratorias de MIGSA se utilizaron para sub-dividir el *heatmap* mostrado en la Figura 4.2 en *sub-heatmaps* específicos de cada subtipo (es decir, Basal vs. todos los demás, Her2+ vs. todos los demás, Luminal B vs. todos los demás, y Luminal A vs. todos los demás). Estos *sub-heatmap* proporcionaron una visión detallada de los patrones de enriquecimiento de cada subtipo, y pueden observarse en (Rodriguez, Merino, Llera, & Fernández, 2019). En particular, los dendrogramas de estas figuras mostraron una mejor agrupación por contrastes, es decir, clusters para cada contraste que agruparon más experimentos que en la Figura 4.2.

El número de **TECC** y de términos en **PTF** encontrados se presenta en la Tabla 4.1. Como puede observarse, el número más bajo de **TECC** (184) se encuentra para el contraste Luminal B vs. Her2+, mientras que el contraste Luminal A vs. Luminal B muestra el valor más alto, con 983 CG enriquecidos. En términos de los **PTF**, Luminal B y Her2+ fueron los que presentaron el menor número de términos, mientras que Luminal A alcanzó el mayor número de CG en su **PTF**, seguido por el subtipo Basal.

Cabe señalar que los **PTF** no son mutuamente excluyentes, lo cuál era de esperar ya que todos los fenotipos contrastados son subtipos de la misma enfermedad. Veintisiete CG se encontraron presentes, en común, en los **PTF** de los cuatro subtipos de CM, mientras que 506 CG fueron únicos (exclusivos) para un solo subtipo. El número de CG exclusivos observados para cada **PTF** fue de 103 para Basal, uno para Luminal B, siete para Her2+ y 395 para Luminal A. Además, se encontraron 33 CG en común entre dos **PTF** diferentes: diez en $\boldsymbol{PTF}_{Basal} \cap \boldsymbol{PTF}_{Her2+}$, ocho en $\boldsymbol{PTF}_{Basal} \cap \boldsymbol{PTF}_{LuminalB}$, siete en $\boldsymbol{PTF}_{Basal} \cap \boldsymbol{PTF}_{LuminalA}$, y ocho en $\boldsymbol{PTF}_{LuminalB} \cap \boldsymbol{PTF}_{LuminalA}$.

Tabla 4.1: Conjuntos de genes en consenso. Número de Términos Enriquecidos en Consenso por Contraste (**TECC**), definidos como aquellos conjuntos de genes enriquecidos en al menos el 50 % de los experimentos evaluados para cada contraste (fila vs. columna). La columna "Intersección" indica el número de conjuntos de genes encontrados en todos los **TECC** de cada subtipo dado, lo cual define el Perfil Transcriptómico Funcional (**PTF**) de cada subtipo de cáncer de mama.

Contraste	Basal	Luminal B	Her2+	Luminal A	Intersección
Basal	-	488	437	768	155
Luminal B		-	184	983	44
Her2+			-	707	44
Luminal A				-	437

4.7.2. Exploración de los PTF de los subtipos de cáncer de mama

La exploración de los **PTF** reveló varios CG de GO previamente relacionados con los subtipos de CM. La visualización basada en árboles de la estructura GO, proporcionada por MIGSA, se utilizó para visualizar el **PTF** de cada subtipo en las categorías PB, CC y FM de GO. Los árboles de GO de los **PTF** de cada subtipo pueden encontrarse en (Rodriguez et al., 2019), donde cada término fue codificado por colores de acuerdo al **PTF** al que pertenece. La visualización en árbol nos permitió explorar más a fondo las relaciones entre los **PTF** de cada subtipo. La exploración de los árboles reveló que los términos no estaban dispersos al azar, sino que se encontraban conformando ramas a diversas profundidades, dominadas principalmente por subtipos particulares. Además, los CG presentes en común en los **PTF** de todos los subtipos fueron encontrados ubicados en nodos poco profundos que involucran términos generales y comúnmente encontrados, tales como citoplasma, desarrollo de sistemas, muerte y proliferación celular.

El análisis del árbol de PB reveló que las ramas que involucran términos relacionados con la regulación de la vía de señalización de los receptores de estrógeno sólo aparecieron en el **PTF** del subtipo Basal. Esto está motivado por el hecho de que, genes implicados en estos CG, incluyendo *ESR1*, se han encontrado previamente sobre-expresados en los subtipos Luminales y en algunos tumores Her2+ (Network & others, 2012), y sub-expresados en los sujetos Basal (Valentin, Da Silva, Privat, Alaoui-Jamali, & Bignon, 2012). Por esta razon, estos términos no se encontraron entre los **TECC** de los contrastes que involucran a los subtipos Luminales y HER2+ y, en consecuencia, no estaban presentes en el **PTF** de estos subtipos. Además, los términos asociados con el filamento de actina, la transición epitelio-mesénquima y el estrés del retículo endoplásmico aparecieron en el árbol de PB como **PTF** de Basal. Estas ramas son procesos bien conocidos relacionados con los tumores Basales, que los diferencian de otros subtipos (Guen et al., 2017). Los genes *MLPH* y *FOXC1* se encontraron entre los GIE presentes en estas ramas, y fueron consistentemente sub y sobre-expresados, respectivamente, en todos los experimentos en los que se contrastó Basal. Cabe mencionar que estos dos genes están involucrados en una firma molecular recientemente desarrollada para cánceres de mama triple negativos, altamente relacionados con tumores Basales (Santuario-Facio et al., 2017). Además, los GIE presentes en términos relacionados con la diferenciación celular, como *GATA3* y *FOXA1*, se encontraron desregulados en el subtipo Basal para todos los contrastes evaluados, en concordancia con los hallazgos del "Grupo del Atlas del Genoma del Cáncer" (Network & others, 2012). A su vez, *ERBB4*, un oncogén drogable (un gen cuya expresión puede afectarse con drogas), se encontró desregulado en el 95 % de los contrastes con Basal evaluados, lo que sugiere que los sujetos Basales no se verían afectados por los fármacos que ataquen el gen *ERBB4*.

El subtipo Luminal A resultó ser muy diferente de los demás en lo que se refiere a los procesos de regulación del ciclo celular y replicación del ADN, en donde

los términos como la transición G1/S del ciclo celular y la unión del origen de la replicación del ADN aparecieron como su **PTF**. Es bien sabido que los procesos de progresión a través del ciclo celular y la replicación del ADN están profundamente involucrados en la proliferación (Hanahan & Weinberg, 2000), que es la principal diferencia entre Luminal A y los subtipos más agresivos como Luminal B, Her2+, y Basal (Prat & Perou, 2011; Yersal & Barutca, 2014). En particular, estos términos muestran la principal diferencia entre los dos subtipos Luminales, ya que los tumores Luminal B expresan clásicamente niveles más altos de Ki67 (Cheang et al., 2009). Adicionalmente, se encontró que los GIE que provocan la proliferación celular, tales como *AURKA*, *PCNA*, *CHEK2*, y *RAD51*, estaban sub-expresados en Luminal A con respecto a los otros subtipos. Mientras que, genes como *ERBB2*, *GRB7*, y *FGFR4* fueron encontrados sobre-expresados en todos los experimentos en los que Her2+ fue contrastado, en concordancia con lo informado por el "Grupo del Atlas del Genoma del Cáncer" (Network & others, 2012).

4.7.3. Integración de resultados de MIGSA de TCGA y Haibe-Kains

Los resultados de MIGSA para cada tecnología de expresión del TCGA se combinaron con los resultados de los HKds. La Figura 4.3 muestra el *heatmap* del eje $Q^{CGE,Es}$ del cubo de resultados de MIGSA. El análisis del dendrograma superior mostró que los contrastes evaluados a partir de los datos del TCGA se asemejan a los respectivos resultados de HKds.

Como en la Figura 4.2, al analizar la Figura 4.3 se encontró un gran cluster con 11 de 12 experimentos contrastando Luminal A (marca LA), y otro cluster incluyendo los ocho experimentos contrastando el subtipo Basal con Luminal B ó Her2+ (marca Ba). Curiosamente, ambos clusters identificados contienen un cluster más específico, que agrupa únicamente experimentos transcriptómicos (marcas LA^T y Ba^T). Además,

para cada contraste, se encontró un cluster que agrupaba solamente datos transcriptómicos, es decir, de microarreglos y secuenciación de ARN del TCGA, junto con los **TECC** de HKds (marcas T). Este resultado se encuentra en acuerdo con lo reportado por el "Grupo del Atlas del Genoma del Cáncer" (Network & others, 2012) para un subgrupo de sujetos Luminal B y Her2+. En concordancia con los resultados del "Grupo del Atlas del Genoma del Cáncer" para arrays de proteínas en fase inversa (Network & others, 2012), los experimentos proteómicos que contrastan Basal con Luminal B ó Her2+ se agruparon cerca de sus contrapartes transcriptómicas. Por otro lado, para los demás contrastes, la expresión de proteínas parece proporcionar información funcional complementaria a su contraparte transcriptómica.

Para cada subtipo, se obtuvieron los **TET** y los **TEP**, y los correspondientes GIE, y se analizaron en conjunto con los correspondientes **PTF** encontrados previamente. En la Tabla 4.2, para cada subtipo, se presenta el número de términos que se encuentran exclusivamente en **PTF**, **TET** y **TEP**, y los que se encuentran exclusivamente en cada intersección. Analizando estos los resultados, encontramos que la concordancia entre los **PTF** y su contraparte del TCGA es notable. Por ejemplo, de los 680 términos que pertenecen a los **PTF** (suma de celdas de la fila Total para columnas que involucran los **PTF**), el 38 % también se encontró en **TEP** o **TET** de los datos del TCGA. En particular, los subtipos Basal y Luminal A fueron los únicos que exhibieron términos en común entre transcriptómica (**PTF** y **TET**) y proteómica (**TEP**). También se encontró que casi el 80 % de esos términos en los **TET** también se encontraron en los **PTF**, evidenciando la alta consistencia de MIGSA para analizar los conjuntos de datos de transcriptómica de los HKds y TCGA. Por el contrario, y para todos los subtipos de CM, esos términos enriquecidos con datos proteómicos no se encontraron frecuentemente entre los **TET**. Este hallazgo sugiere que los **TEP** proporcionan información complementaria a la contraparte transcriptómica, como se reportó previamente (Meng et al., 2014). Sin embargo, será necesaria una investiga-

ción más profunda que incluya un mayor número de datasets proteómicos para poder definir un Perfil **Proteómico** Funcional.

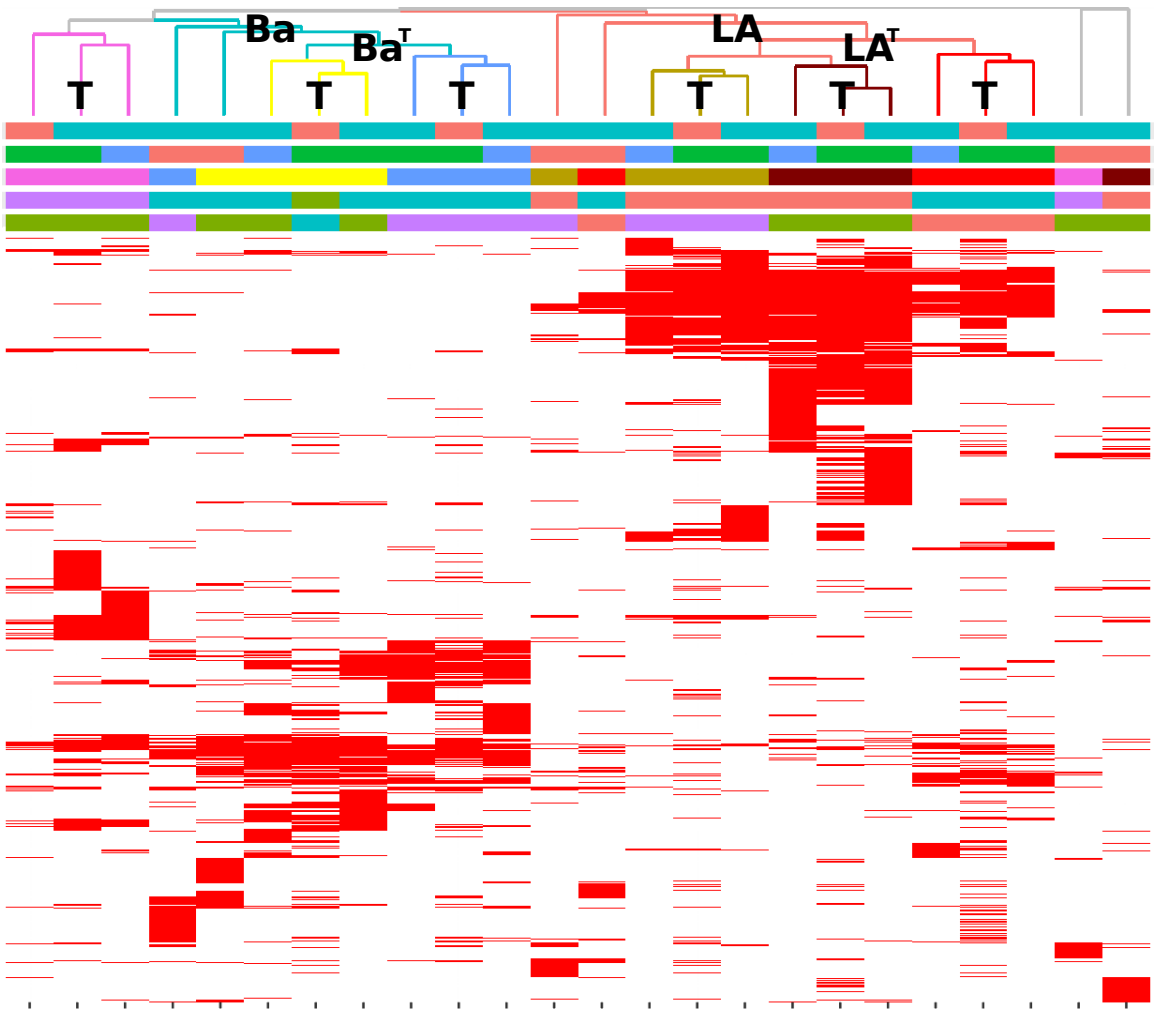


Figura 4.3: Conjuntos de genes enriquecidos por análisis funcional para TCGA. Heatmap: conjuntos de genes en filas, experimentos en columnas; ordenados por distancia Jaccard y enlace promedio. Conjuntos de genes enriquecidos en rojo. Barras de columnas (de arriba abajo): salmón es "términos enriquecidos por consenso para los conjuntos de datos de Haibe-Kains", en turquesa datos de TCGA; azul datos de secuenciación de ARN, verde de microarreglos, en salmón de iTRAQ; contraste (comparación entre pares de subtipos); subtipo1 y subtipo2, salmón es Luminal A, turquesa es Basal, púrpura es Her2+, y el verde es Luminal B. Clusters de dendrogramas: Se demarcan para los subtipos Basal (Ba) y Luminal A (LA), para experimentos transcriptómicos de Basal (Ba^T) y de Luminal A (LA^T), y para experimentos transcriptómicos específicos por contraste (T).

Tabla 4.2: Número de términos enriquecidos compartidos entre plataformas y múltiples ómicas para cada subtipo.

Subtipo	**PTF^{E}**	**TEP^{E}**	**TET^{E}**	**(PTF ∩ TEP)^{E}**	**(PTF ∩ TET)^{E}**	**(TEP ∩ TET)^{E}**	**PTF∩ TEP∩ TET**	**Total**
Basal	109	46	21	12	19	0	15	222
Luminal B	37	7	25	0	7	0	0	76
Her2+	42	20	1	0	2	0	0	65
Luminal A	234	2	15	3	186	0	14	454
Total	422	75	62	15	214	0	29	817

PTF: Perfil Transcriptómico Funcional obtenido para cada subtipo a partir de 25 conjuntos de datos basados en microarreglos. **TEP**: Términos Enriquecidos por Proteómica a partir de datos del TCGA. **TET**: Términos Enriquecidos por Transcriptómica a partir de datos del TCGA. $(X)^{E}$ denota los términos en X pero no en las intersecciones de X con los otros conjuntos. $(X \cap Y)^{E}$ denota los términos en X y en Y pero no en las intersecciones de X ó Y con los otros conjuntos.

4.7.4. Exploración de los resultados de TCGA en conjunto con los PTF

Para cada subtipo, se analizaron los árboles de GO incluyendo los términos enriquecidos por cada categoría presente en las columnas de la Tabla 4.2. Como se señaló anteriormente, los términos enriquecidos únicamente en datos proteómicos se encontraron principalmente en los subtipos Basal y Her2+. Por otra parte, para Luminal A y Luminal B los árboles están dominados principalmente por ramas enriquecidas en transcriptómica.

Un hallazgo interesante al analizar los **TEP** fue la identificación del gen *INPP4B* como un GIE al contrastar los subtipos Basal y Her2+ con los Luminales. En particular, la sobre-expresión de *INPP4B* detectada en los tumores Luminales es consistente con su perfil de expresión reportado por el "Grupo del Atlas del Genoma del Cáncer" (Network & others, 2012) para datos de arrays de proteínas en fase inversa.

Para los sujetos Basales, los GIE tales como *DHRS3*, *LDHB*, *HSD17B8*, *CBR4*, *CYB5A*, y *PHGDH* fueron encontrados en ramas relacionadas con la actividad de la oxidoreductasa. De estos, los GIE de deshidrogenasa, *DHRS3*, *LDHB*, y el *PHGDH*

fueron encontrados sobre-expresados en experimentos proteómicos en los cuales el subtipo Basal fue contrastado. Apoyando este hallazgo, la sobre-expresión de la proteína *PHGDH* ha sido previamente relacionada con tumores Basales (Gromova et al., 2015). Además, este gen se ha asociado con la proliferación celular, la migración y los procesos de invasión, y recientemente se ha identificado como un biomarcador inmunohistoquímico de cáncer (Song, Feng, Lu, Lin, & Dong, 2018).

El análisis de los árboles de Luminal B incluye una rama de **TEP** relacionada con la secreción hormonal. En esta rama, se encontraron genes tales como *GATA3*, *IRS1*, *BAD*, *SYTL4*, *NDUFAF2*, *GPLD1*, *EGFR*, *FGB*, *FGG*, y *LYN*. En particular, para el contraste proteómico Luminal B vs. Her2+, *EGFR* se encontró des-regulado para Luminal B, mientras que *GATA3* estaba sobre-expresado, tal como se informa para arrays de proteínas en fase inversa por el "Grupo del Atlas del Genoma del Cáncer" (Network & others, 2012). Otra rama, relacionada con los ribosomas, fue encontrada para Luminal B, en la cual varios genes ribosómicos mitocondriales fueron encontrados como GIE: *MRPL18*, *MRPL28*, *MRPL37*, *MRPL45*, *MRPL48*, *MRPL50*, *MRPL53*, y *MRPL54*; todos ellos fueron encontrados sobre-expresados en Luminal B contrastándolo con Luminal A.

Como se muestra en la Tabla 4.2, 29 términos previamente identificados en los **PTF** de los subtipos de CM también fueron detectados en los respectivos **TEP** y **TET**. Entre estos, se encontraron términos que diferencian el subtipo Basal, como el término de FM de unión al distroglicano; y el de PB de regulación positiva de la vía de señalización del receptor de estrógeno intracelular. Para el subtipo Luminal A, se detectó una rama de CC relacionada con el cromosoma condensado, y términos de PB asociados a ciclos celulares mitóticos y meióticos. Veintidós GIE fueron encontrados contribuyendo al enriquecimiento de términos para **PTF**, **TEP** y **TET**. De ellos, 11 fueron identificados para el subtipo Basal (por ejemplo, *AR*, *GATA3*, *FOXA1*, *NFIB*, *AGR2*, *CA12*, y *DSC2*) y 11 para los tumores Luminal A (por ejemplo, *TOP2A*,

CCNB1, *CDC20*, *UBE2C*, *TTK*, y *KIF2C*). Cabe destacar que todos estos genes se comportaron de acuerdo con las observaciones realizadas por el "Grupo del Atlas del Genoma del Cáncer" (Network & others, 2012).

4.8. Conclusiones

En este capítulo se presentó el paquete R `MIGSA` para el Análisis Masivo e Integrador de Conjuntos de Genes. Esta herramienta realiza un análisis funcional completo/integrador, masivamente sobre varias matrices de expresión de diferentes repositorios y fuentes ómicas. `MIGSA` facilita la búsqueda de patrones funcionales -que son denotados por conjuntos de genes y que explican fenotipos particulares- por medio de herramientas exploratorias y gráficas. MIGSA también permite a los usuarios indagar en conjuntos de genes particulares para buscar los genes específicos responsables del enriquecimiento funcional. Las capacidades analíticas y exploratorias de `MIGSA` como herramienta de análisis funcional masivo, en conjunto con el hecho de que a nuestro entender, es la única herramienta existente que permite la comparación de múltiples experimentos desde un punto de vista funcional, la convierten en un enfoque novedoso.

La utilidad de `MIGSA` quedó demostrada debido a la caracterización funcional de los subtipos de Cáncer de Mama (CM) definidos por Perou et al. La aplicación de `MIGSA` sobre 118 experimentos de microarreglos de CM resultó en la identificación de Perfiles Transcriptómicos Funcionales (**PTF**) y Genes Importantes de Enriquecimiento (GIE) para cada subtipo. La fiabilidad de nuestros hallazgos fue confirmada usando resultados previamente reportados en la literatura. La identificación de GIE con un comportamiento de expresión consistente a través de conjuntos de datos resulta útil para el descubrimiento de genes provocadores de enfermedades, y por ende, buenos blancos a atacar por drogas.

La capacidad de `MIGSA` de integrar múltiples conjuntos de datos ómicos se demostró aplicándolo a los datasets de CM del TCGA compuestos tanto de datos transcriptómicos (secuenciación de ARN y microarreglos) como proteómicos (iTRAQ). Los resultados obtenidos se utilizaron para identificar, para cada subtipo, el conjunto de Términos Enriquecidos por Transcriptómica (**TET**) y de Términos Enriquecidos por Proteómica (**TEP**). Al integrar y comparar estos resultados con los **PTF** encontrados previamente, se descubrió una alta consistencia de patrones funcionales entre los términos enriquecidos en datos de TCGA y en los **PTF**. Dado que `MIGSA` es una herramienta tan flexible, el procedimiento presentado puede ser replicado tanto en otros subtipos de CM como en otras enfermedades con el fin de explorar perfiles funcionales. El paquete `MIGSA` se encuentra libremente disponible -bajo una licencia GPLv2- en el repositorio Bioconductor `https://bioconductor.org/packages/MIGSA/`, y cuenta con más de 3.000 descargas según las estadísticas del repositorio `https://bioconductor.org/packages/stats/bioc/MIGSA/` desde su primera versión, en Marzo de 2017.

Capítulo 5

Desafíos que surgieron durante el trabajo de tesis

5.1. Eficiencia computacional del algoritmo `mGSZ`

Como se vio en el Capítulo 3, el método `mGSZ` resultó ser la mejor alternativa para llevar a cabo el análisis funcional del tipo *Puntuación Funcional de Clase*. La función principal de `mGSZ`, con sus parámetros por defecto -que sugerimos como resultado del Capítulo 3-, realiza un máximo de 200 permutaciones y filtra solo aquellos conjuntos de genes que presentan menos de cinco genes. En este sentido, llevar a cabo el análisis funcional de `mGSZ` de un experimento ordinario, por ejemplo, Basal vs. Luminal A de datos de microarreglos del TCGA sobre Gene Ontology, en una computadora estándar, puede tomar alrededor de 2,46 horas, como mostramos a continuación.

Esta demora del algoritmo, puede resultar despreciable si se desea analizar un único experimento. Sin embargo, como objetivo global del presente trabajo de tesis, se propuso una herramienta de exploración, desde un punto de vista funcional, de grandes cantidades de experimentos. Si extrapolamos linealmente la demora del análisis funcional de un experimento a 118 experimentos (el número de experimentos

analizados en el Capítulo 4), estaríamos hablando de una espera de 290,28 horas, es decir, 12,09 días. Dada esta situación, resultó fundamental afrontar este desafío, mediante la optimización del algoritmo original de `mGSZ`, de manera de disminuir el tiempo de ejecución.

5.1.1. Optimización del algoritmo `mGSZ`

Para disminuir el tiempo de ejecución de `mGSZ`, en primer instancia se debió analizar el código del paquete `mGSZ` de manera de encontrar porciones de código críticas. Mediante los paquetes `profvis` y `microbenchmark` se detectaron tanto aquellas porciones de código de más lenta ejecución, como aquellas que más frecuentemente se ejecutaban. Como resultado de este análisis, se obtuvieron unos pocos fragmentos de código que fueron **objetivos críticos** a optimizar.

La optimización de estos objetivos críticos se llevo a cabo aplicando diversas estrategias de optimización generales (Cooper & Torczon, 2011; Wolfe, 1996) como también algunas específicas del lenguaje R (Burns, 2011; Gillespie & Lovelace, 2016; Wickham, 2014). Esta versión optimizada de la función `mGSZ` se implementó en el paquete `MIGSA`, y puede utilizarse, independientemente de `MIGSA`, llamando a la función `MIGSAmGSZ`.

Con el fin de evaluar las mejoras en tiempo de ejecución de `MIGSAmGSZ` con respecto a `mGSZ`, se analizaron ambas funciones dándoles como entrada los mismos input: la matriz de expresión de microarreglos de ADN de cáncer de mama del TCGA (16.207 genes $\times$ 237 sujetos), donde se contrastaron los subtipos Basal vs. Luminal A, analizando los conjuntos de genes del Gene Ontology y KEGG (20.425 conjuntos de genes). El análisis se llevó a cabo utilizando un procesador Intel(R) Xeon(R) E5-2620 v3 @ 2.40GHz (24 núcleos), y 128GB de memoria RAM.

Dicha comparación se puede replicar mediante el siguiente código:

```r
library("BiocParallel")
library("mGSZ")
library("MIGSA")
library("MIGSAdata")

# Cargamos la matriz de expresión del TCGA
data(tcgaMAdata)
subtipos <- tcgaMAdata$subtypes
expresion <- tcgaMAdata$geneExpr

dim(expresion) # #genes y #sujetos
```

```
[1] 16207   237
```

```r
table(subtipos) # #sujetos de cada subtipo
```

```
subtipos
Basal  LumA
   95   142
```

```r
# Cargamos los conjuntos de genes de KEGG y Gene Ontology
# utilizando funciones de MIGSA
conj_genes <- list(
  KEGG = downloadEnrichrGeneSets("KEGG_2015")[[1]],
  BP = loadGo("BP"),
  CC = loadGo("CC"),
  MF = loadGo("MF")
)
conj_genes <- do.call(c, lapply(conj_genes, MIGSA:::asList))
```

```r
# Ejecutamos mGSZ original y medimos su tiempo de ejecución
set.seed(8818)
mGSZ_tiempo <- system.time({
  mGSZ_res <- mGSZ(expresion, conj_genes, subtipos)$mGSZ
})
```

```r
# Ejecutamos MIGSAmGSZ y medimos su tiempo de ejecución
register(SerialParam()) # linea que aclararemos a continuación
set.seed(8818)
MIGSAmGSZ_tiempo <- system.time({
  MIGSAmGSZ_res <- MIGSAmGSZ(expresion, conj_genes, subtipos)
})
```

```r
# Vemos el tiempo total ("elapsed") y calculamos Speedup
mGSZ_tiempo <- mGSZ_tiempo[["elapsed"]]
MIGSAmGSZ_tiempo <- MIGSAmGSZ_tiempo[["elapsed"]]
c(
  mGSZ = mGSZ_tiempo / 60 / 60,            # en horas
  MIGSAmGSZ = MIGSAmGSZ_tiempo / 60 / 60,  # en horas
  Speedup = mGSZ_tiempo / MIGSAmGSZ_tiempo
)
```

```
    mGSZ MIGSAmGSZ   Speedup
2.461000  1.553534  1.584130
```

Como se observa, la versión optimizada, `MIGSAmGSZ`, presenta un *speedup* de 1,6X sobre el algoritmo original -de demorar 2,46 horas, la nueva versión pasó a demorar 1,55 horas-.

Por otra parte, la implementación original de mGSZ se ejecuta, exclusivamente, de modo secuencial. En la actualidad, siendo tan fácil el acceso a computadores multiprocesador, el algoritmo mGSZ deja de lado una alternativa muy favorable en cuanto a la eficiencia en el tiempo de ejecución.

Como segunda fase de la reimplementación de mGSZ, detectamos cuáles porciones de código de MIGSAmGSZ podrían ejecutarse en paralelo y otorgarían un *speedup* considerable. Utilizando el paquete BiocParallel, se paralelizaron aquellas rutinas que consideramos pertinentes, y se brindó una sencilla interfaz de paralelismo al usuario. El análisis de la ganancia obtenida, gracias al paralelismo de MIGSAmGSZ, se puede replicar mediante el siguiente código:

```
# Evaluamos MIGSAmGSZ con 1, 2, 4, 8, 10, 12 y 14 núcleos
nucleos <- c(1, 2, 4, 8, 10, 12, 14)
resultados <- lapply(nucleos, function(act_nucl) {
  # Mediante las siguientes 4 líneas de código se le indica
  # la cantidad de núcleos en las que debe ejecutarse MIGSAmGSZ.
  # El parámetro 'workers' indica la cantidad de núcleos a utilizar.
  register(MulticoreParam(
    workers = act_nucl, threshold = "DEBUG",
    progressbar = TRUE
  ))

  set.seed(8818)
  MIGSAmGSZ_tiempo <- system.time({
    MIGSAmGSZ_res <- MIGSAmGSZ(expresion, conj_genes, subtipos)
  })

  return(list(tiempo = MIGSAmGSZ_tiempo, res = MIGSAmGSZ_res))
```

```
})
```

```
# Para cada ejecución conservamos el tiempo total ("elapsed")
# y calculamos Speedup y Eficiencia sobre mGSZ
tiempos_segs <- sapply(resultados, function(act_res) {
  act_res$tiempo[["elapsed"]]
})
names(tiempos_segs) <- nucleos
metricas <- rbind(
  "Demora (mins)" = tiempos_segs / 60, # en minutos
  Speedup = mGSZ_tiempo / tiempos_segs,
  Eficiencia = (mGSZ_tiempo / tiempos_segs) / nucleos
)
round(metricas, 2)
```

```
                1     2     4     8    10    12    14
Demora (mins) 93.21 46.50 24.98 15.63 13.67 14.79 28.43
Speedup        1.58  3.18  5.91  9.45 10.81  9.98  5.19
Eficiencia     1.58  1.59  1.48  1.18  1.08  0.83  0.37
```

Como puede observarse en la tabla superior, sin importar el número de núcleos en los que se haya corrido `MIGSAmGSZ`, su rendimiento fue superior al de `mGSZ`. Ejecutándose en un núcleo muestra un *speedup* de 1,6X, alcanzando un máximo de 10,8X con diez núcleos. Gracias a la optimización desarrollada, es posible obtener en 14 minutos los mismos resultados que se obtenían en 2,46 horas de ejecución con el `mGSZ` original. Extrapolando a 118 experimentos, pasaríamos de una demora de 12,09 días, a tan solo 27,53 horas.

Como se mencionó previamente, la función optimizada y paralelizada `MIGSAmGSZ`

se encuentra incluída en el paquete `MIGSA`, y se utiliza siguiendo los mismos parámtros que la versión original `mGSZ`.

Capítulo 6

Conclusiones y trabajo futuro

En esta tesis se presentó la problemática que presenta el Análisis Funcional (AF) frente a grandes cantidades de bases de datos provenientes de diversas fuentes ómicas. Dicha problemática surge debido a la falta de actualización de las metodologías de AF existentes, en contraste con el rápido avance de las tecnologías de obtención de datos de muestras biológicas. Con el surgimiento de tecnologías para obtener datos biológicos de nuevas fuentes, las técnicas de AF quedan obsoletas u otorgan resultados incorrectos. Por otra parte, al disminuir los costos de obtención de datos, la disponibilidad de grandes cantidades de bases de datos es desaprovechada por las metodologías de AF.

En el Capítulo 1 se introdujo al lector en el concepto del análisis funcional. Para ello, en la Sección 1.1, se presentó la noción de *Ontología* en el contexto de la biología. Se detallaron varios grupos conocidos de ontologías, como KEGG, BioCarta, Reactome, MSigDB, y Gene Ontology, donde se mostró cómo se estructura su información en conceptos, categorías, términos o vías metabólicas. Luego, en la Sección 1.2, se presentaron las distintas metodologías de AF existentes. Se explicó, a grandes rasgos, la idea subyacente a las metodologías de *Análisis de Sobre-Representación* (ASR) y de *Puntuación Funcional de Clase* (PFC) detallando sus ventajas y limitaciones.

El Capítulo 2 presentó diversas fuentes de datos biológicas que la tecnología permite cuantificar y que son objeto de análisis. En la Sección 2.1 se profundizó en las tecnologías de Microarreglos de ADN, iTRAQ, y Secuenciación de ARN, ya que permiten obtener matrices de expresión, las cuáles son utilizadas como input del AF. Para cada tecnología se especificó la forma en la que obtienen los valores de expresión, y el tipo de dato que devuelven. Luego, en la Sección 2.2, se comentó sobre los diversos repositorios de libre acceso a bases de datos de expresión, la cantidad de sujetos que presenta cada uno y las ómicas que analizan. Finalmente, se mencionaron las condiciones experimentales que resultaban de interés en estos repositorios, para ser inspeccionadas bajo AF.

En el Capítulo 3 se llevó a cabo un comparación exhaustiva de diversos algoritmos pertenecientes a las metodologías tanto de ASR como de PFC, junto con combinaciones de sus parámetros. Se mostró que los resultados del AF pueden variar notablemente dependiendo del método y los parámetros utilizados. Lo cual puede influir negativamente en la interpretación biológica si no se aborda adecuadamente. De este capítulo concluimos que tanto el ASR como la PFC proporcionan resultados complementarios que pueden integrarse para obtener una visión biológica más amplia del experimento en estudio. Por ende, presentamos un pipeline de Análisis Funcional Integrador, `IFA`, que realiza análisis simultáneos de ASR y PFC, proporcionando un marco completo y unificado de AF. Finalmente, se evaluó desde el punto de vista de la biología, los resultados que obtiene este pipeline `IFA`, lo cual demostró su capacidad de detección de funciones biológicas adecuadas para el fenómeno bajo estudio, además de presentar resultados concordantes entre diversos experimentos.

En el Capítulo 4 se presentó el paquete R desarrollado para la presente tesis, `MIGSA`. Esta herramienta permite realizar un AF completo, integrador, y masivo sobre grandes cantidades de bases de datos. `MIGSA` posibilita detectar genes y patrones funcionales biológicos, por medio de herramientas exploratorias y gráficas, que ca-

racterizan poblaciones, fenotipos, o grupos de interés. La utilidad de `MIGSA` quedó demostrada debido a la caracterización funcional de los subtipos de cáncer de mama, donde se logró definir un perfil funcional para cada subtipo. `MIGSA` efectivamente permitió integrar múltiples bases de datos, y múltiples fuentes de datos (ómicas).

Finalmente, en el Capítulo 5 se introdujeron aquellos desafíos que surgieron durante el transcurso de la investigación de la presente tesis, pero que no estaban relacionados directamente con el objetivo bajo estudio. Al notar la lentitud del algoritmo de PFC utilizado por el pipeline `IFA` y el paquete `MIGSA`, el desafío principal consistió en disminuir su tiempo de ejecución. En este sentido, luego de un extenso trabajo de optimización y paralelización del algoritmo, se logró un notable *speedup* de 10,8X en el mejor de los casos.

Como propuesta de **trabajo a futuro**, se desprenden dos líneas de estudio a contemplar. Por una parte, como se mencionó en la Sección 4.2, se debió adaptar el algoritmo `mGSZ` para que resultara adecuado para analizar datos provenientes de secuenciación de ARN. Para esta adaptación se propuso utilizar la metodología `voom`, la cuál modela este tipo de datos mediante mínimos cuadrados generalizados (Law et al., 2014). Sin embargo, resulta necesario llevar a cabo una comparación y evaluación extensiva de otras propuestas de manejo de datos de secuenciación de ARN frecuentemente mencionadas por la comunidad científica, entre ellas, `edgeR` (Robinson, McCarthy, & Smyth, 2010) y `DESeq2` (Love, Huber, & Anders, 2014).

Por otra parte, el paquete `MIGSA` presenta varios aspectos que pueden ser extendidos, y de los cuales creemos que continuar su desarrollo resultará en notables beneficios para la comunidad científica. Por una parte, con el continuo avance de las tecnologías de obtención de datos ómicos, resulta fundamental actualizar constantemente la herramienta `MIGSA` de manera que permita el análisis de tipos de datos con diversos supuestos estadísticos subyacentes, como ejemplo de esto, un desafío pendiente es lograr analizar datos de transcriptos de genes. Como segundo objetivo,

resultaría interesante poder extender `MIGSA` al análisis de nuevos diseños experimentales. En la actualidad, `MIGSA` permite utilizar cualquier diseño experimental, sin embargo no se ha evaluado la correctitud estadística de utilizar otros diseños -aparte de caso vs. control- en este tipo de AF. Como tercer objetivo a futuro, proponemos idear e implementar nuevas técnicas de exploración y visualización de los resultados de `MIGSA`, esto permitirá nuevas estrategias para inspeccionar resultados y generar nuevas hipótesis. Finalmente, creemos que extender `MIGSA` de manera que incluya una interfaz gráfica de usuario, aumentará el público de la herramienta, incorporando aquellos usuarios no tan acostumbrados al ambiente de la programación. Mediante la librería de R `shiny`, crear interfaces que corren código R resulta sencillo (Rodriguez, Vargas, & Fernández, 2018) y extremadamente útil.

Bibliografía

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., ... others. (2000). Gene ontology: Tool for the unification of biology. *Nature Genetics*, *25*(1), 25.

Atchley, D. P., Albarracin, C. T., Lopez, A., Valero, V., Amos, C. I., Gonzalez-Angulo, A. M., ... Arun, B. K. (2008). Clinical and pathologic characteristics of patients with brca-positive and brca-negative breast cancer. *Journal of Clinical Oncology*, *26*(26), 4282–4288.

Aure, M. R., Vitelli, V., Jernström, S., Kumar, S., Krohn, M., Due, E. U., ... others. (2017). Integrative clustering reveals a novel split in the luminal a subtype of breast cancer with impact on outcome. *Breast Cancer Research*, *19*(1), 44.

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., ... others. (2012). NCBI geo: Archive for functional genomics data sets—update. *Nucleic Acids Research*, *41*(D1), D991–D995.

Bastien, R. R., Rodrı'guez-Lescure, Á., Ebbert, M. T., Prat, A., Munárriz, B., Rowe, L., ... others. (2012). PAM50 breast cancer subtyping by rt-qPCR and concordance with standard clinical molecular markers. *BMC Medical Genomics*, *5*(1), 1.

Berkeley, C. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *E-Book Available at Http://Www. Bepress. Com/Sagmb/Vol3/Iss1/Art3.[PubMed]*.

Bittner, M. (2005). Expression project for oncology (expo). *National Center for Biotechnology Information. Available at: Http://Www. Ncbi. Nlm. Nih. Gov/Geo.*

Bonnefoi, H., Potti, A., Delorenzi, M., Mauriac, L., Campone, M., Tubiana-Hulin, M., ... others. (2007). RETRACTED: Validation of gene signatures that predict the response of breast cancer to neoadjuvant chemotherapy: A substudy of the eortc 10994/big 00-01 clinical trial. Elsevier.

Bos, P. D., Zhang, X. H.-F., Nadal, C., Shu, W., Gomis, R. R., Nguyen, D. X., ... others. (2009). Genes that mediate breast cancer metastasis to the brain. *Nature*, *459*(7249), 1005.

Burns, P. (2011). *The r inferno*. Lulu. com.

Cantor, S. B., Bell, D. W., Ganesan, S., Kass, E. M., Drapkin, R., Grossman, S., ... others. (2001). BACH1, a novel helicase-like protein, interacts directly with brca1 and contributes to its dna repair function. *Cell*, *105*(1), 149–160.

Canuel, V., Rance, B., Avillach, P., Degoulet, P., & Burgun, A. (2014). Translational research platforms integrating clinical and omics data: A review of publicly available solutions. *Briefings in Bioinformatics*, *16*(2), 280–290.

Carlson, M., Falcon, S., Pages, H., & Li, N. (2013). Org. Hs. Eg. Db: Genome wide annotation for human. R package version.

Cheang, M. C., Chia, S. K., Voduc, D., Gao, D., Leung, S., Snider, J., ... others. (2009). Ki67 index, her2 status, and prognosis of patients with luminal b breast cancer. *JNCI: Journal of the National Cancer Institute*, *101*(10), 736–750.

Chen, R., & Snyder, M. (2013). Promise of personalized omics to precision medicine. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, *5*(1), 73–82.

Chin, K., DeVries, S., Fridlyand, J., Spellman, P. T., Roydasgupta, R., Kuo, W.-L., ... others. (2006). Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. *Cancer Cell*, *10*(6), 529–541.

Cleveland, W. S. (2001). Data science: An action plan for expanding the technical

areas of the field of statistics. *International Statistical Review*, *69*(1), 21–26.

Consortium, G. O. (2016). Expansion of the gene ontology knowledgebase and resources. *Nucleic Acids Research*, *45*(D1), D331–D338.

Cooper, K., & Torczon, L. (2011). *Engineering a compiler*. Elsevier.

Creighton, C. J., Chen, F., Zhang, Y., Gibbons, D. L., Deneen, B., Kwiatkowski, D., & Ittmann, M. (2018). Pan-cancer molecular classes transcending tumor lineage across 32 cancer types, multiple data platforms, and over 10,000 cases. *Clinical Cancer Research*, clincanres–3378.

Curtis, C., Shah, S. P., Chin, S.-F., Turashvili, G., Rueda, O. M., Dunning, M. J., ... others. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, *486*(7403), 346.

Dai, X., Li, T., Bai, Z., Yang, Y., Liu, X., Zhan, J., & Shi, B. (2015). Breast cancer intrinsic subtype classification, clinical use and future trends. *American Journal of Cancer Research*, *5*(10), 2929.

Davis, A. P., Grondin, C. J., Lennon-Hopkins, K., Saraceni-Richards, C., Sciaky, D., King, B. L., ... Mattingly, C. J. (2014). The comparative toxicogenomics database's 10th year anniversary: Update 2015. *Nucleic Acids Research*, gku935.

Dedeurwaerder, S., Desmedt, C., Calonne, E., Singhal, S. K., Haibe-Kains, B., Defrance, M., ... others. (2011). DNA methylation profiling reveals a predominant immune component in breast cancers. *EMBO Molecular Medicine*, *3*(12), 726–741.

Edelman, E., Porrello, A., Guinney, J., Balakumaran, B., Bild, A., Febbo, P. G., & Mukherjee, S. (2006). Analysis of sample set enrichment scores: Assaying the enrichment of sets of genes for individual samples in genome-wide expression profiles. *Bioinformatics*, *22*(14), e108–e116.

Ein-Dor, L., Kela, I., Getz, G., Givol, D., & Domany, E. (2005). Outcome signature genes in breast cancer: Is there a unique set? *Bioinformatics*, *21*(2), 171–178.

Ein-Dor, L., Zuk, O., & Domany, E. (2006). Thousands of samples are needed

to generate a robust gene list for predicting outcome in cancer. *Proceedings of the National Academy of Sciences*, *103*(15), 5923–5928.

Falcon, S., & Gentleman, R. (2006). Using gostats to test gene lists for go term association. *Bioinformatics*, *23*(2), 257–258.

Falcon, S., & Gentleman, R. (2007). Using gostats to test gene lists for go term association. *Bioinformatics*, *23*(2), 257–258.

Fang, H., & Gough, J. (2014). Thednet'approach promotes emerging research on cancer patient survival. *Genome Medicine*, *6*(8), 64.

Farmer, P., Bonnefoi, H., Becette, V., Tubiana-Hulin, M., Fumoleau, P., Larsimont, D., ... others. (2005). Identification of molecular apocrine breast tumours by microarray analysis. *Breast Cancer Research*, *7*(2), P2–11.

Fernandez, E. A., & Casares, F. M. (2018). Current challenges for big omics data analytics and precision medicine. *Med Sci Tech*, *59*, 1–3.

Fernández, E., Alvarez, M., Podhajcer, O., & Stolovitzky, G. (2007). Genomica funcional: En busca de la función de los genes. *National Academy of Science, Argentine*, *13*, 63–76.

Fresno, C., & Fernández, E. A. (2013). RDAVIDWebService: A versatile r interface to david. *Bioinformatics*, *29*(21), 2810–2811.

Fresno, C., González, G. A., Merino, G. A., Flesia, A. G., Podhajcer, O. L., Llera, A. S., & Fernández, E. A. (2016). A novel non-parametric method for uncertainty evaluation of correlation-based molecular signatures: Its application on pam50 algorithm. *Bioinformatics*, *33*(5), 693–700.

Fresno, C., González, G. A., Merino, G. A., Rodriguez, J. C., Balzarini, M. G., & Fernández, E. A. (2014). Control de calidad de microarreglos de adn mediante descomposición anova-pca/pls. *XIX Reunión Científica Del Grupo Argentino de Biometría*.

Fresno, C., Llera, A. S., Girotti, M. R., Valacco, M. P., López, J. A., Podhajcer,

O. L., ... Fernández, E. A. (2012). The multi-reference contrast method: Facilitating set enrichment analysis. *Computers in Biology and Medicine*, *42*(2), 188–194.

Gendoo, D. M., Ratanasirigulchai, N., Schröder, M. S., Paré, L., Parker, J. S., Prat, A., & Haibe-Kains, B. (2015). Genefu: An r/bioconductor package for computation of gene expression-based signatures in breast cancer. *Bioinformatics*, *32*(7), 1097–1099.

Gillespie, C., & Lovelace, R. (2016). *Efficient r programming: A practical guide to smarter programming*. .º'Reilly Media, Inc.".

Goeman, J. J., & Bühlmann, P. (2007). Analyzing gene expression data in terms of gene sets: Methodological issues. *Bioinformatics*, *23*(8), 980–987.

Goldhirsch, A., Winer, E. P., Coates, A., Gelber, R., Piccart-Gebhart, M., Thürlimann, B., ... others. (2013). Personalizing the treatment of women with early breast cancer: Highlights of the st gallen international expert consensus on the primary therapy of early breast cancer 2013. *Annals of Oncology*, *24*(9), 2206–2223.

Grasso, C. S., Wu, Y.-M., Robinson, D. R., Cao, X., Dhanasekaran, S. M., Khan, A. P., ... others. (2012). The mutational landscape of lethal castration-resistant prostate cancer. *Nature*, *487*(7406), 239.

Gromova, I., Gromov, P., Honma, N., Kumar, S., Rimm, D., Talman, M.-L. M., ... Moreira, J. (2015). High level phgdh expression in breast is predominantly associated with keratin 5-positive cell lineage independently of malignancy. *Molecular Oncology*, *9*(8), 1636–1654.

Guen, V. J., Chavarria, T. E., Kröger, C., Ye, X., Weinberg, R. A., & Lees, J. A. (2017). EMT programs promote basal mammary stem cell and tumor-initiating cell stemness by inducing primary ciliogenesis and hedgehog signaling. *Proceedings of the National Academy of Sciences*, 201711534.

Haibe-Kains, B., Desmedt, C., Loi, S., Culhane, A. C., Bontempi, G., Quackenbush, J., & Sotiriou, C. (2012). A three-gene model to robustly identify breast cancer molecular subtypes. *Journal of the National Cancer Institute*, *104*(4), 311–325.

Hanahan, D., & Weinberg, R. A. (2000). The hallmarks of cancer. *Cell*, *100*(1), 57–70.

Hess, K. R., Anderson, K., Symmans, W. F., Valero, V., Ibrahim, N., Mejia, J. A., ... others. (2006). Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer. *Journal of Clinical Oncology*, *24*(26), 4236–4244.

Hoadley, K. A., Yau, C., Wolf, D. M., Cherniack, A. D., Tamborero, D., Ng, S., ... others. (2014). Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell*, *158*(4), 929–944.

Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2008a). Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, *37*(1), 1–13.

Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2008b). Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nature Protocols*, *4*(1), 44.

Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nature Protocols*, *4*(1), 44–57.

Joshi-Tope, G., Gillespie, M., Vastrik, I., D'Eustachio, P., Schmidt, E., Bono, B. de, ... others. (2005). Reactome: A knowledgebase of biological pathways. *Nucleic Acids Research*, *33*(suppl_1), D428–D432.

Kanehisa, M., & Goto, S. (2000). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, *28*(1), 27–30.

Kedaigle, A., & Fraenkel, E. (2018). Turning omics data into therapeutic insights. *Current Opinion in Pharmacology*, *42*, 95–101.

Khatri, P., Sirota, M., & Butte, A. J. (2012). Ten years of pathway analysis: Current approaches and outstanding challenges. *PLoS Computational Biology*, *8*(2),

e1002375.

Kolesnikov, N., Hastings, E., Keays, M., Melnichuk, O., Tang, Y. A., Williams, E., ... others. (2014). ArrayExpress update—simplifying data submissions. *Nucleic Acids Research*, *43*(D1), D1113–D1116.

Korkola, J. E., Blaveri, E., DeVries, S., Moore, D. H., Hwang, E. S., Chen, Y.-Y., ... Waldman, F. M. (2007). Identification of a robust gene signature that predicts breast cancer outcome in independent data sets. *BMC Cancer*, *7*(1), 61.

Kuleshov, M. V., Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., ... others. (2016). Enrichr: A comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Research*, *44*(W1), W90–W97.

Kwa, M., Makris, A., & Esteva, F. J. (2017). Clinical utility of gene-expression signatures in early stage breast cancer. *Nature Reviews Clinical Oncology*, *14*(10), 595.

Law, C. W., Chen, Y., Shi, W., & Smyth, G. K. (2014). Voom: Precision weights unlock linear model analysis tools for rna-seq read counts. *Genome Biology*, *15*(2), R29.

Li, Q., Eklund, A. C., Juul, N., Haibe-Kains, B., Workman, C. T., Richardson, A. L., ... Swanton, C. (2010). Minimising immunohistochemical false negative er classification using a complementary 23 gene expression signature of er status. *PloS One*, *5*(12), e15031.

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P., & Mesirov, J. P. (2011). Molecular signatures database (msigdb) 3.0. *Bioinformatics*, *27*(12), 1739–1740.

Liedtke, C., Mazouni, C., Hess, K. R., André, F., Tordai, A., Mejia, J. A., ... others. (2008). Response to neoadjuvant therapy and long-term survival in patients with triple-negative breast cancer. *Journal of Clinical Oncology*, *26*(8), 1275–1281.

Liu, B., Shen, X., & Pan, W. (2016). Integrative and regularized principal compo-

nent analysis of multiple sources of data. *Statistics in Medicine*, *35*(13), 2235–2250.

Llera, A. S., Podhajcer, O. L., Breitenbach, M., Santini, L., Muller, B., Daneri Navarro, A., ... others. (2015). Translational cancer research comes of age in latin america. In *American sci transl med* (Vol. 7, p. 319). American Association for the Advancement of Science.

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biology*, *15*(12), 550.

Lu, X., Lu, X., Wang, Z. C., Iglehart, J. D., Zhang, X., & Richardson, A. L. (2008). Predicting features of breast cancer with gene expression patterns. *Breast Cancer Research and Treatment*, *108*(2), 191.

Manoli, T., Gretz, N., Gröne, H.-J., Kenzelmann, M., Eils, R., & Brors, B. (2006). Group testing for pathway analysis improves comparability of different microarray datasets. *Bioinformatics*, *22*(20), 2500–2506.

McCarthy, D. J., & Smyth, G. K. (2009). Testing significance relative to a fold-change threshold is a treat. *Bioinformatics*, *25*(6), 765–771.

Meng, C., Kuster, B., Culhane, A. C., & Gholami, A. M. (2014). A multivariate approach to the integration of multi-omics datasets. *BMC Bioinformatics*, *15*(1), 162.

Metzger Filho, O., Ignatiadis, M., & Sotiriou, C. (2011). Genomic grade index: An important tool for assessing breast cancer tumor grade and prognosis. *Critical Reviews in Oncology/Hematology*, *77*(1), 20–29.

Miller, L. D., Smeds, J., George, J., Vega, V. B., Vergara, L., Ploner, A., ... others. (2005). An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proceedings of the National Academy of Sciences*, *102*(38), 13550–13555.

Minn, A. J., Gupta, G. P., Padua, D., Bos, P., Nguyen, D. X., Nuyten, D., ... others. (2007). Lung metastasis genes couple breast tumor size and metastatic spread. *Proceedings of the National Academy of Sciences*, *104*(16), 6740–6745.

Minn, A. J., Gupta, G. P., Siegel, P. M., Bos, P. D., Shu, W., Giri, D. D., ... Massagué, J. (2005). Genes that mediate breast cancer metastasis to lung. *Nature*, *436*(7050), 518.

Mishra, P., Törönen, P., Leino, Y., & Holm, L. (2014). Gene set analysis: Limitations in popular existing methods and proposed improvements. *Bioinformatics*, *30*(19), 2747–2756.

Natrajan, R., Weigelt, B., Mackay, A., Geyer, F. C., Grigoriadis, A., Tan, D. S., ... others. (2010). An integrative genomic and transcriptomic analysis reveals molecular pathways and networks regulated by copy number aberrations in basal-like, her2 and luminal cancers. *Breast Cancer Research and Treatment*, *121*(3), 575–589.

Network, C. G. A., & others. (2012). Comprehensive molecular portraits of human breast tumours. *Nature*, *490*(7418), 61.

Nishimura, D. (2001). BioCarta. *Biotech Software & Internet Report: The Computer Software Journal for Scient*, *2*(3), 117–120.

Parker, J. S., Mullins, M., Cheang, M. C., Leung, S., Voduc, D., Vickery, T., ... others. (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of Clinical Oncology*, *27*(8), 1160.

Pavlidis, P., Qin, J., Arango, V., Mann, J. J., & Sibille, E. (2004). Using the gene ontology for microarray data mining: A comparison of methods and application to age effects in human prefrontal cortex. *Neurochemical Research*, *29*(6), 1213–1222.

Pawitan, Y., Bjöhle, J., Amler, L., Borg, A.-L., Egyhazi, S., Hall, P., ... others. (2005). Gene expression profiling spares early breast cancer patients from adjuvant therapy: Derived and validated in two population-based cohorts. *Breast Cancer Research*, *7*(6), R953.

Perou, C. M., Sørlie, T., Eisen, M. B., Van De Rijn, M., Jeffrey, S. S., Rees, C. A., ... others. (2000). Molecular portraits of human breast tumours. *Nature*, *406*(6797), 747.

Prat, A., Parker, J. S., Karginova, O., Fan, C., Livasy, C., Herschkowitz, J. I., ... Perou, C. M. (2010). Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. *Breast Cancer Research*, *12*(5), R68.

Prat, A., & Perou, C. M. (2011). Deconstructing the molecular portraits of breast cancer. *Molecular Oncology*, *5*(1), 5–23.

Reis-Filho, J. S., & Pusztai, L. (2011). Gene expression profiling in breast cancer: Classification, prognostication, and prediction. *The Lancet*, *378*(9805), 1812–1823.

Richardson, A. L., Wang, Z. C., De Nicolo, A., Lu, X., Brown, M., Miron, A., ... Ganesan, S. (2006). X chromosomal abnormalities in basal-like human breast cancer. *Cancer Cell*, *9*(2), 121–132.

Rivals, I., Personnaz, L., Taing, L., & Potier, M.-C. (2006). Enrichment or depletion of a go category within a class of genes: Which test? *Bioinformatics*, *23*(4), 401–407.

Rivals, I., Personnaz, L., Taing, L., & Potier, M.-C. (2007). Enrichment or depletion of a go category within a class of genes: Which test? *Bioinformatics*, *23*(4), 401–407.

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). EdgeR: A bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, *26*(1), 139–140.

Rodriguez, J. C., González, G. A., Fresno, C., & Fernández, E. A. (2015). Integrative functional analysis improves information retrieval in breast cancer. *Iberoamerican Congress on Pattern Recognition*.

Rodriguez, J. C., González, G. A., Fresno, C., Llera, A. S., & Fernández, E. A. (2016a). Improving information retrieval in functional analysis. *Computers in Biology and Medicine*, *79*, 10–20.

Rodriguez, J. C., Merino, G. A., Llera, A. S., & Fernández, E. A. (2019). Massive integrative gene set analysis enables functional characterization of breast cancer

subtypes. *Journal of Biomedical Informatics*, 103157.

Rodriguez, J. C., Merino, G. A., Prato, L., Llera, A. S., & Fernández, E. A. (2016b). The impact of rna-seq differential expression algorithms on over-representation analysis of gene sets. *ISCB Latin America 2016*.

Rodriguez, J. C., Prato, L., Llera, A. S., & Fernández, E. A. (2017). Effects of rna-seq genes analysis on the over-representation analysis of gene sets. *XXI Congreso Argentino de Bioingeniería - SABI 2017*.

Rodriguez, J. C., Vargas, C., & Fernández, E. A. (2018). ShinyWYSIWYG: A shiny what you see is what you get editor. *Latinamerican Conference About the Use of R in R&D 2018*.

Ross-Adams, H., Lamb, A., Dunning, M., Halim, S., Lindberg, J., Massie, C., ... others. (2015). Integration of copy number and transcriptomics provides risk stratification in prostate cancer: A discovery and validation cohort study. *EBioMedicine*, *2*(9), 1133–1144.

Saal, L. H., Johansson, P., Holm, K., Gruvberger-Saal, S. K., She, Q.-B., Maurer, M., ... others. (2007). Poor prognosis in carcinoma is associated with a gene expression signature of aberrant pten tumor suppressor pathway activity. *Proceedings of the National Academy of Sciences*, *104*(18), 7564–7569.

Santuario-Facio, S. K., Cardona-Huerta, S., Perez-Paramo, Y. X., Trevino, V., Hernandez-Cabrera, F., Rojas-Martinez, A., ... others. (2017). A new gene expression signature for triple-negative breast cancer using frozen fresh tissue before neoadjuvant chemotherapy. *Molecular Medicine*, *23*, 101.

Scaltriti, M., Elkabets, M., & Baselga, J. (2016). Molecular pathways: AXL, a membrane receptor mediator of resistance to therapy. *Clinical Cancer Research*, clincanres–1458.

Schmidt, M., Böhm, D., Von Törne, C., Steiner, E., Puhl, A., Pilch, H., ... Gehrmann, M. (2008). The humoral immune system has a key prognostic impact in node-

negative breast cancer. *Cancer Research*, *68*(13), 5405–5413.

Sharma, D., Blum, J., Yang, X., Beaulieu, N., Macleod, A. R., & Davidson, N. E. (2005). Release of methyl cpg binding proteins and histone deacetylase 1 from the estrogen receptor α (er) promoter upon reactivation in er-negative human breast cancer cells. *Molecular Endocrinology*, *19*(7), 1740–1751.

Shen, R., Olshen, A. B., & Ladanyi, M. (2009). Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*, *25*(22), 2906–2912.

Shi, L., Reid, L. H., Jones, W. D., Shippy, R., Warrington, J. A., Baker, S. C., ... others. (2006). The microarray quality control (maqc) project shows inter-and intraplatform reproducibility of gene expression measurements. *Nature Biotechnology*, *24*(9), 1151.

Shi, Q., Zhang, C., Peng, M., Yu, X., Zeng, T., Liu, J., & Chen, L. (2017). Pattern fusion analysis by adaptive alignment of multiple heterogeneous omics data. *Bioinformatics*, *33*(17), 2706–2714.

Silver, D. P., Richardson, A. L., Eklund, A. C., Wang, Z. C., Szallasi, Z., Li, Q., ... others. (2010). Efficacy of neoadjuvant cisplatin in triple-negative breast cancer. *Journal of Clinical Oncology*, *28*(7), 1145.

Song, Z., Feng, C., Lu, Y., Lin, Y., & Dong, C. (2018). PHGDH is an independent prognosis marker and contributes cell proliferation, migration and invasion in human pancreatic cancer. *Gene*, *642*, 43–50.

Sotiriou, C., Wirapati, P., Loi, S., Harris, A., Fox, S., Smeds, J., ... others. (2006). Gene expression profiling in breast cancer: Understanding the molecular basis of histologic grade to improve prognosis. *Journal of the National Cancer Institute*, *98*(4), 262–272.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., ... others. (2005). Gene set enrichment analysis: A knowledge-based approach

for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, *102*(43), 15545–15550.

Sørlie, T., Borgan, E., Myhre, S., Vollan, H. K., Russnes, H., Zhao, X., ... Rødland, E. (2010). The importance of gene-centring microarray data. *The Lancet Oncology*, *11*(8), 719–720.

Taylor, B. S., Schultz, N., Hieronymus, H., Gopalan, A., Xiao, Y., Carver, B. S., ... others. (2010). Integrative genomic profiling of human prostate cancer. *Cancer Cell*, *18*(1), 11–22.

Tian, L., Greenberg, S. A., Kong, S. W., Altschuler, J., Kohane, I. S., & Park, P. J. (2005). Discovering statistically significant pathways in expression profiling studies. *Proceedings of the National Academy of Sciences*, *102*(38), 13544–13549.

Valentin, M. D., Da Silva, S. D., Privat, M., Alaoui-Jamali, M., & Bignon, Y.-J. (2012). Molecular insights on basal-like breast cancer. *Breast Cancer Research and Treatment*, *134*(1), 21–30.

Van't Veer, L. J., Dai, H., Van De Vijver, M. J., He, Y. D., Hart, A. A., Mao, M., ... others. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, *415*(6871), 530.

Varambally, S., Yu, J., Laxman, B., Rhodes, D. R., Mehra, R., Tomlins, S. A., ... others. (2005). Integrative genomic and proteomic analysis of prostate cancer reveals signatures of metastatic progression. *Cancer Cell*, *8*(5), 393–406.

Vaske, C. J., Benz, S. C., Sanborn, J. Z., Earl, D., Szeto, C., Zhu, J., ... Stuart, J. M. (2010). Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using paradigm. *Bioinformatics*, *26*(12), i237–i245.

Verhaak, R. G., Hoadley, K. A., Purdom, E., Wang, V., Qi, Y., Wilkerson, M. D., ... others. (2010). Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in pdgfra, idh1, egfr, and nf1. *Cancer Cell*, *17*(1), 98–110.

Waddell, N., Arnold, J., Cocciardi, S., Da Silva, L., Marsh, A., Riley, J., ... others. (2010). Subtypes of familial breast tumours revealed by expression and copy number profiling. *Breast Cancer Research and Treatment*, *123*(3), 661–677.

Wang, B., Mezlini, A. M., Demir, F., Fiume, M., Tu, Z., Brudno, M., ... Goldenberg, A. (2014). Similarity network fusion for aggregating data types on a genomic scale. *Nature Methods*, *11*(3), 333.

Wang, Y., Klijn, J. G., Zhang, Y., Sieuwerts, A. M., Look, M. P., Yang, F., ... others. (2005). Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *The Lancet*, *365*(9460), 671–679.

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., ... others. (2013). The cancer genome atlas pan-cancer analysis project. *Nature Genetics*, *45*(10), 1113.

Wickham, H. (2014). *Advanced r*. Chapman; Hall/CRC.

Wolfe, M. J. (1996). *High performance compilers for parallel computing*. Addison-Wesley.

Yersal, O., & Barutca, S. (2014). Biological subtypes of breast cancer: Prognostic and therapeutic implications. *World Journal of Clinical Oncology*, *5*(3), 412.

Yu, K., Ganesan, K., Tan, L. K., Laban, M., Wu, J., Zhao, X. D., ... others. (2008). A precisely regulated gene expression cassette potently modulates metastasis and survival in multiple solid cancers. *PLoS Genetics*, *4*(7), e1000129.