VNIVERSITAT DE VALÈNCIA

Programa de doctorado Estadística y Optimización

Métodos estadísticos en el estudio de la variabilidad genómica en cáncer

Tesis doctoral de

José Carbonell Caballero

Dirigida por

David Conesa

Antonio López-Quílez

Joaquín Dopazo

Realizada en la

Facultad de Ciencias Matemáticas

Departamento de Estadística e Investigación Operativa

Julio de 2021

Agradecimientos

Bueno, pues parece que después de unos cuantos años peleando, por fin llegamos al final.

Terminar esta tesis no solo ha supuesto completar el trabajo más duro que he realizado en mi vida, también significa cerrar un ciclo personal que me ha proporcionado experiencias muy valiosas para el futuro. Atrás quedan todas las tardes, fines de semana y horas que he tenido que robar al reloj para poder avanzar un poquito cada día. Ahora queda disfrutar y, sobre todo, asimilar todo el trabajo realizado.

En primer lugar, me gustaría agradecer a mis directores todo el apoyo recibido en esta tesis. Gracias David y Antonio por ayudarme a completar esta tarea tan difícil, vuestros consejos y comentarios han sido una referencia para mi en este camino. También, quiero agradecer a Ximo todo su apoyo, los 10 años que compartimos en el CIPF fueron una etapa muy enriquecedora para mi, aprendí muchísimo y descubrí que la bioinformática es una de las formas más increibles de aproximarse a la ciencia.

De esta etapa también me gustaría destacar al grupo de personas con las que tuve la suerte de coincidir. Gracias Quino, Matías, Cankut, Alex, Marta, Alicia, Dan, Pacolo, Susi, Paco y las dos Marianas, os agradezco mucho las risas, los chistes malos, los videos frikis en *youtube*, las paellas en el Perelló y, en especial, toda la cantidad de comida que regularmente traíamos y compartíamos cada día en el trabajo, sin todas esas calorías físicas y emocionales no podría ser la persona que soy ahora.

Por supuesto, no quiero olvidar a otras ilustres compañeras de etapas anteriores

como Luz, Rosa, Patricia Sebastián, Sonia o Patri Diaz. Y como no, agradecer a mis tres irónicos mosqueteros Jorge, Roberto y Quique todo el soporte moral en estos últimos años, es dificil encontrar personas con las que discutir de naranjas y física cuantica en la misma conversación y eso es algo que siempre he valorado muchísimo. Espero que en breve lo podamos celebrar en persona y definamos por fin un plan para hacernos ricos.

También quiero mencionar a personas especiales como Mariam, Pepemo o Stefan, además de haber compartido grandes momentos (y montones de tapas en *Malagueños*), también les debo haber descubierto el mundo de la investigación y haber sembrado en mi el gusanillo de hacer una tesis doctoral.

Por supuesto, no podría olvidarme de mis *Pim-Pam-eras* amigas Eva y Sonia, que durante estos últimos años han tenido que aguantar mil veces las preocupaciones y angustias de una tesis que parecía no acabar nunca, os debo una invitación de las buenas.

Además, me gustaría agradecer a Adonis su apoyo en muchos momentos de la tesis. También a Guille y Silvina, cuya pasta frola y alfajores hicieron que muchas tardes llegara a casa con la energía suficiente para seguir trabajando un ratito más.

Por supuesto, y muy en especial, me gustaría agradecer a Vanessa todo el apoyo y comprensión que tuvo conmigo en los momentos más complicados de esta etapa, gracias por escuchar sin reproches, creo que una parte de mi siempre estará en deuda contigo.

Por último, y más importante, quiero dedicarle esta tesis a mi familia, y agradecerle mucho que hayan absorbido parte de mis preocupaciones durante estos años. De verdad, prometo no volver a hacer otra tesis.

"Al andar se hace camino
Y al volver la vista atrás
Se ve la senda que nunca
Se ha de volver a pisar
Caminante no hay camino
sino estelas en la mar"

Antonio Machado

A mi familia

UNIVERSITAT DE VALÈNCIA

Resumen

Facultad de Ciencias Matemáticas
Departamento de Estadística e Investigación Operativa
Programa de Doctorado en Estadística y Optimización

El estudio del cáncer representa una de las áreas de investigación más importantes en el ámbito de las enfermedades humanas. A pesar de los enormes recursos invertidos en la investigación biomédica en los últimos años, todavía representa una de las patologías con mayor índice de mortalidad. El cáncer describe a un conjunto enfermedades cuyo factor común reside en el crecimiento incontrolado de un grupo de células a las que denominamos tumor. Se trata de patologías con una base genética muy heterogénea, incluso cuando se compara a individuos con el mismo tipo de cáncer. Esta circunstancia convierte en un desafío a la búsqueda de marcadores comunes que puedan ser empleados posteriormente en el diseño de terapias de uso general.

El proceso de tumorigénesis se inicia tras la acumulación de una serie de alteraciones somáticas en componentes esenciales de las células, lo que se traduce en la sobre-estimulación de funciones como el crecimiento y la proliferación celular, mientras se inhiben otros mecanismos de control destinados a eliminar aquellas células del tejido que muestran un comportamiento errático. En los tumores, el motor básico de cambio lo constituye la aparición de nuevas alteraciones somáticas, siendo en ocasiones su generación de naturaleza casi aleatoria. Como resultado, es habitual observar en tumores reales un abanico variado de configuraciones genéticas, produciendo la aparición de diferentes subpoblaciones celulares, llamadas clones, que evolucionan, compiten y cooperan en un proceso análogo a la selección natural de Darwin. Esta circunstancia ha hecho de la caracterización de mutaciones somáticas un elemento fundamental en el estudio de la biología molecular del cáncer.

En este contexto, la biología computacional (BC) constituye una de las herramientas más valiosas. De forma destacada, la BC ha contribuido con el desarrollo de una generación de herramientas computacionales destinadas a la detección robusta de alteraciones somáticas en los genomas de pacientes con cáncer. Asimismo, a través del modelado estadístico, la BC ha permitido el establecimiento de patrones específicos de expresión génica, dando lugar a una caracterización funcional exhaustiva del conjunto de genes recurrentes que muestran un papel relevante en diversos tipos de cáncer. Además, la BC ha proporcionado una enorme variedad de métodos estadísticos destinados a la cuantificación y análisis de las rutas moleculares, las cuales describen cómo se producen las interacciones entre proteínas para llevar a cabo las funciones esenciales en la célula. Esta aproximación ha permitido abordar el estudio del cáncer mediante la biología de sistemas, una disciplina que describe a las enfermedades como alteraciones específicas en partes de un sistema, más allá de los cambios producidos en sus elementos individuales. De forma significativa, este abordaje ha permitido entender por qué alteraciones en proteínas diferentes provocan cambios similares en las rutas moleculares, de gran relevancia en el estudio de la variabilidad genómica del cáncer.

Al margen de la variabilidad biológica observada, los métodos de análisis desarrollados han necesitado modelar en mayor o menor medida el grado de incertidumbre presente en los protocolos de adquisión de datos, como la secuenciación genómica. Con independencia del método de adquisición, todas las modalidades cuentan con un conjunto de fuentes de ruido de diversa índole, con capacidad para alterar de forma significativa la fiabilidad de la cuantificación, dando lugar a numerosos falsos positivos en los análisis posteriores. En el contexto del análisis molecular del cáncer, el modelado del ruido tendrá gran relevancia, especialmente en el ámbito del genotipado somático. En este caso, debido a las características intrínsecas del problema, un conjunto significativo de mutaciones reales estarán respaldadas por cambios de poca magnitud, comparables a lo que esperaríamos por la presencia de un artefacto de ruido.

Uno de los enfoques estadísticos más interesantes a la hora de modelar la

variabilidad lo representa el uso de métodos de detección de componentes latentes. Estas metodologías permiten describir a las observaciones como una combinación lineal de un conjunto finito, y típicamente pequeño, de componentes esenciales. Para esta tarea, existen diferentes aproximaciones, que tratan de realizar el proceso de optimización a través de criterios muy distintos, como el de la maximización de la varianza (*PCA*), el de la independencia estadística entre componente (*ICA*) o el de la combinación de elementos positivos para construir el conjunto de observaciones (*NMF*).

El presente trabajo tiene por objetivo el estudio de la heterogeneidad genómica en el cáncer. Para ello, se describe un protocolo general de análisis que comienza con la estimación de mutaciones somáticas, para después evaluar su efecto sobre las distintas rutas de señalización. En este caso, se hará uso de un modelo jerárquico de factorización que permitirá determinar simultáneamente las componentes latentes, tanto a nivel de gen, como a nivel de ruta molecular. Dichas componentes se corresponderán con las diferentes estrategias de supervivencia implementadas en los tumores reales, lo que proporcionará una visión global de la heterogeneidad genómica desde diferentes niveles de abstracción, incluyendo mutaciones, genes, rutas y, finalmente, las características biológicas compartidas por todos los cánceres, conocidas como *hallmarks* del cáncer.

El Capítulo 1 comienza con una introducción general al contexto de la tesis. En primer lugar, se realizará una descripción general sobre los *hallmarks* del cáncer, que tendrán un impacto en el diseño de los modelos estadísticos que serán planteados en los capítulos siguientes. Después, se describirán las contribuciones más importantes de la BC, haciendo especial hincapié en las técnicas de cuantificación de rutas moleculares, de gran relevancia en esta tesis. Por último, se aportará una descripción general sobre las técnicas de estimación de componentes latentes más comunes, describiendo sus ventajas e inconvenientes, en función del contexto de estudio.

En el Capítulo 2 se describe la implementación de un protocolo computacional que tiene por objetivo la predicción robusta de mutaciones somáticas en un grupo de pacientes con cáncer. Para ello, tendrá especial relevancia el modelado del ruido

y su aplicación a la hora de estimar el grado de error esperado en cada región genómica. El modelado del ruido permitirá la obtención de una versión corregida de los datos, y en consecuencia, la obtención de un genotipo más robusto. Este enfoque será especialmente relevante a la hora de recuperar mutaciones somáticas que, debido al grado de heterogeneidad celular y de contaminación normal, estén soportadas por un número muy reducido de lecturas. Este capítulo se compone de tres apartados principales. En primer lugar, se describe la selección de un conjunto de estimadores estadísticos destinado a cuantificar de forma precisa el nivel de ruido desde diferentes puntos de vista. Los estimadores serán después evaluados a lo largo de diferentes experimentos planteados en varios organismos modelo, con el fin de evaluar su capacidad a la hora de detectar regiones genómicas que por su naturaleza muestran un alto grado de susceptibilidad al ruido. El segundo apartado se centra en la implementación del modelo de genotipado somático. Para ello, en primer lugar se describe la construcción del modelo de ruido a partir de los estimadores propuestos y su aplicación a la hora de corregir los datos de entrada. Después, se describe el modelo estadístico de genotipado, que de forma característica tendrá en cuenta la hipotética presencia de diversas subpoblaciones celulares, cuya presencia tendría un efecto directo en la distribución de frecuencias alélicas observada. En tercer lugar, se describe la implementación de una herramienta de simulación de tumores destinada a la evaluación exhaustiva de la herramienta de genotipado somático propuesta y su comparación frente a otras herramientas bien establecidas en el campo. Esta herramienta simula la evolución de un linaje de células tumorales durante un periodo de ciclos definido. En este caso, la incorporación progresiva de mutaciones somáticas condicionará el grado de adaptación al medio de las células afectadas, simulando así las diferentes presiones selectivas observadas en tumores reales.

El Capítulo 3 tiene por objetivo el estudio de la heterogeneidad genómica observada entre pacientes, bajo una perspectiva de biología de sistemas. Para ello, se describe un modelo jerárquico de factorización que permite la obtención simultánea de un conjunto de componentes latentes a nivel de gen y sus correspondientes componentes a nivel de ruta molecular. El protocolo comienza con la integración

de las mutaciones somáticas y los valores de expresión con el fin de caracterizar la actividad de cada uno de los genes implicados. A continuación, se describe la metodología empleada para la cuantificación de las rutas moleculares, centrada en el desarrollo de una versión adaptada de la herramienta *Hipathia* para su uso posterior en el proceso de optimización. Después, se describe el diseño del modelo jerárquico de factorización, y su aplicación a un conjunto de funciones moleculares que muestran una alteración significativa al comparar a pacientes frente a los individuos normales. Finalmente, se analiza cómo se distribuyen los pesos de las componentes latentes obtenidas por el modelo jerárquico a lo largo de los individuos, determinando así si éstas se asocian preferencialmente a características fenotípicas de interés como el subtipo de cáncer.

Para acabar, en el capítulo 4 se ofrece una serie de conclusiones generales de la tesis, haciendo hincapié en las ventajas de un protocolo de análisis como el aquí descrito y su aplicación en casos reales. Asimismo, se describirán algunas líneas de trabajo futuras que permitirán seguir desarrollando las distintas metodologías propuestas.

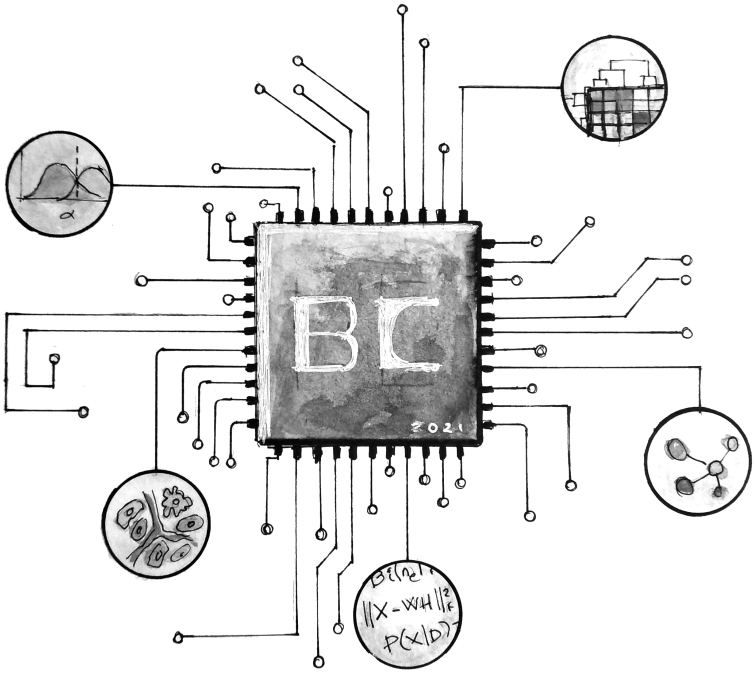
Índice general

Índice de figuras

Índice de tablas

CAPÍTULO 1

Visión general del contexto



En este capítulo se realizará una introducción general al contexto de la tesis. En particular, se realizará una descripción de las características comunes a cualquier cáncer, denominadas hallmarks, que condicionarán el diseño de los modelos estadísticos a desarrollar. Además, se describirán las contribuciones más importantes de la biología computacional en el estudio del cáncer, haciendo especial hincapié en los métodos de cuantificación de rutas moleculares. Por último, se realizará un recorrido general por los métodos más comunes de estimación de componentes latentes, de gran importancia en la última parte de la tesis.

1.1. Biología del cáncer

El término cáncer describe a un conjunto de enfermedades complejas de origen genético caracterizadas por el crecimiento incontrolado de un grupo de células a las que llamamos tumor (Vogelstein y cols., 2013). Representa una de las mayores lacras de este siglo y a pesar de los enormes esfuerzos realizados en investigación en las últimas décadas, todavía es una de las enfermedades con mayor grado de mortalidad y con una prevalencia incrementada por factores demográficos como el envejecimiento de la población (Torre, Siegel, Ward, y Jemal, 2016; Vineis y Wild, 2014).

Existen numerosas razones para justificar la carencia de tratamientos efectivos que permitan curar o cronificar la enfermedad con el objetivo de no representar una amenaza contra la supervivencia de los pacientes. Sin duda, entre ellas destaca la enorme complejidad y el alto grado de heterogeneidad observado en pacientes que cuentan con el mismo tipo de tumor (Burrell, McGranahan, Bartek, y Swanton, 2013; Dagogo-Jack y Shaw, 2018; McGranahan y Swanton, 2017), siendo a menudo complicado encontrar marcadores comunes que permitan diseñar tratamientos efectivos de uso general.

En mayor o menor medida, cada tumor cuenta con un conjunto mínimo de alteraciones somáticas en proteínas clave, necesario para iniciar un comportamiento

aberrante que supone el inicio del cáncer. Asimismo, la adquisición de alteraciones nuevas durante el ciclo de vida del tumor será también esencial para permitir su progreso y desarrollo sostenido dentro del tejido. De forma general, estas alteraciones desembocan en la sobreactivación de procesos esenciales, como la proliferación celular y la inhibición de funciones de control como la apoptosis. Esta secuencia de eventos se enfrenta directamente a los mecanismos de control presentes en todas las células normales, elementos que han evolucionado durante toda nuestra historia con el fin de preservar la supervivencia del organismo en detrimento de la supervivencia de la célula (Venkatesan, Birkbak, y Swanton, 2017).

En un porcentaje amplio de casos, se trata de enfermedades asociadas a la exposición continuada a un agente nocivo, como el tabaco o la radiación solar (Narayanan, Saladi, y Fox, 2010; Torre y cols., 2016). También, elementos relacionados con el estilo de vida, como la dieta o el ejercicio físico, influirán en la predisposición a desarrollar la enfermedad (Turner y Lloyd, 2017; Uzunlulu, Telci-Caklili, y Oguz, 2016). Asimismo, a nivel molecular, otros mecanismos como la acumulación de mutaciones somáticas debido a errores esporádicos en la maquinaria de replicación (Alexandrov y Stratton, 2014; Martincorena y Campbell, 2015) o mutaciones germinales de carácter hereditario en genes esenciales, como los genes *BRCA1/BRCA2* en cáncer de mama (Wang y cols., 2012), tendrán un papel relevante en la aparición y desarrollo del cáncer (Sheikh y cols., 2015).

La investigación realizada en el estudio del cáncer en las últimas décadas ha permitido constatar que, con independencia del tejido de origen, todos los cánceres muestran algunas pautas comunes, como el crecimiento incontrolado o la evasión al sistema inmunitario (Vogelstein y cols., 2013). Estas observaciones han motivado en los últimos años un esfuerzo notable con el fin de encontrar y caracterizar propiedades inherentes a cualquier tumor, que potencialmente podrían aportar una visión más mecanística de la enfermedad y en consecuencia, la posibilidad de definir tratamientos más efectivos y transversales a diferentes tipos de cáncer.

Con esta idea, Hanahan y Weinberg propusieron en el año 2000 lo que se conoce como *hallmarks* (Hanahan y Weinberg, 2000) del cáncer (Figura 1.1), extendidos por

los mismos autores (Hanahan y Weinberg, 2011) en 2011. Los *hallmarks* pretenden ser una descripción general de las habilidades necesarias en cada tumor para poder llevar su desarrollo fuera del control del tejido circundante. A continuación, se describe cada uno de ellos:

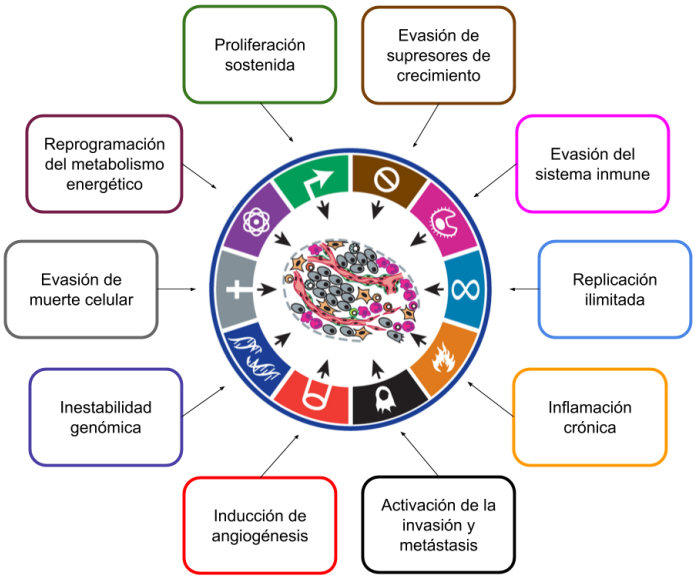


Figura 1.1: *Diagrama representativo de los hallmarks del cáncer. (Figura adaptada del artículo original de Hanahan y Weinberg).*

1) *Proliferación sostenida*

   Las células tumorales muestran una profileración crónica y exhacerbada, mostrando alteraciones importantes en los mecanismos destinados a regular las distintas fases del ciclo celular. Se trata de una de las características más importantes dentro del tumor, ya que le permite compensar la acción del sistema inmunitario y evadir las señales de control externo enfocadas a limitar su proliferación.

2) *Evasión de supresores de crecimiento*

   Las células tumorales muestran alteraciones específicas cuyo efecto provoca la evasión y bloqueo de los programas externos destinados a controlar y limitar el crecimiento celular. En este contexto, también es necesario la inhibición de

determinadas proteínas supresoras de tumor que en caso de estrés o daño celular actúan bloqueando tanto el crecimiento, como la profileración celular.

3) *Replicación ilimitada*

Las células tumorales muestran habitualmente una capacidad de replicación ilimitada. Se debe principalmente a la regulación del acortamiento de los telómeros cromosómicos que en células normales suele llevar a limitar el número máximo de divisiones celulares, llevando a las células a un estado latente de no proliferación, conocido como senescencia, o a la activación de los programas de muerte celular controlada. La vía habitual para este *hallmark* suele implicar alteraciones en la proteína telomerasa.

4) *Evasión de muerte celular*

La muerte celular controlada representa un mecanismo evolutivo presente en células normales destinado a eliminar del tejido aquellas células que muestran alteraciones importantes. Las células tumorales han de ser capaces de bloquear e inhibir estas señales internas con el fin de poder seguir acumulando alteraciones genómicas durante su desarrollo.

5) *Inducción de angiogénesis*

El crecimiento y la proliferación celular sobreestimulada requieren la disposición de nutrientes y recursos necesarios, así como la posibilidad de disponer de diferentes vías para excretar compuestos de desecho. Para ello, las células tumorales promueven la creación de nueva vascularización que habitualmente implica la aparición de nuevas células endoteliales y su ensamblaje en forma de tubos. Además, este mecanismo representa un paso esencial en la propagación y metástasis del tumor a otros órganos.

6) *Evasión del sistema inmunitario*

Las células tumorales adquieren mecanismos que les permiten mermar total o parcialmente el efecto del sistema inmunitario. De forma característica, se ha observado que la infiltración de células del sistema inmunitario a menudo resulta ventajosa para el tumor, ya que permite promover la reparación del tejido y favorecer otros mecanismos como la angiogénesis.

7) *Inflamación crónica*

La inflamación crónica resulta una característica observada de forma recurrente en

los tumores, ya que ésta promueve de forma indirecta la reparación del tejido y la creación de nueva vascularización, lo que favorece el desarrollo del tumor.

8) *Reprogramación del metabolismo energético*
La alteración del metabolismo celular es también una característica ampliamente observada en tumores. Al igual que en el caso de la angiogénesis, las células tumorales necesitan exhacerbar su metabolismo con el fin de incrementar los niveles de energía necesarios para mantener su crecimiento y proliferación. Para llevar a cabo este *hallmark*, las células tumorales necesitan reprogramar su metabolismo, que se caracteriza por un incremento considerable en el consumo de glucosa y su procesamiento fuera de las mitocondrias, conocido como efecto *Warburg*.

9) *Inestabilidad genómica*
La inestabilidad genómica representa una de las características más apreciables de las células tumorales. Dicho mecanismo promueve la aparición de nuevas alteraciones cromosómicas y con ello, la capacidad de disponer rápidamente de nuevas habilidades adquiridas que permitan a las células tumorales ser más competitivas en el tejido.

10) *Activación de la invasión y metástasis*
La metástasis supone el último paso en la colonización de diferentes tejidos por parte del tumor. Para ello, las células tumorales deben promover la degradación de la matriz extracelular, con el fin de ganar movilidad, llegar al sistema vascular y propagarse a otros tejidos.

La forma en la que se implementan y mantienen los *hallmarks* en cada tumor resulta de gran complejidad, ya que el enorme grado de heterogeneidad observada, tanto entre pacientes, como a nivel intratumoral, sugiere la existencia de multitud de posibles vías de ejecución (Sanchez-Vega y cols., 2018), probablemente esculpidas de forma particular por la presión selectiva ejercida por el tejido en cada tumor. Entender como funciona este proceso y qué restricciones del tejido condicionan las posibles soluciones viables puede tener gran impacto en el campo de la investigación molecular en cáncer, haciendo del estudio de la heterogeneidad genómica una de las vías de investigación más importantes.

Con este fin, es necesario contar con la aplicación de métodos estadísticos y computacionales que proporcionen formas eficientes de medir la actividad tumoral

y de dilucidar las distintas pautas esenciales para el desarrollo de un tumor en cada paciente. Para ello, es necesario también integrar el conocimiento biológico previo contenido en las bases de datos, y manejado en forma de anotaciones y redes de interacción. Por último, es importante contar con una visión más mecanística de la enfermedad, donde las células se describan mediante un sistema jerárquico con diferentes niveles de abstracción que pueda ser interrogado con el fin de entender como las alteraciones somáticas presentes en proteínas clave influyen en el desarrollo de procesos celulares más elevados que implementan y sostienen a los *hallmarks* del cáncer.

1.2. Biología computacional y de sistemas

La biología computacional (BC) es un campo multidisciplinar que integra diferentes áreas afines, como la computación, la estadística o las matemáticas, con el propósito de estudiar y resolver problemas biológicos. Gracias a los avances tecnológicos, se ha convertido en una disciplina estratégica, que ha ganado un enorme peso incluso en ámbitos más experimentales. Se trata de una herramienta fundamental en el estudio de características celulares como la actividad génica, la integridad del genoma o la dinámica de la cromatina, entre otras.

Las nuevas tecnologías aplicadas al estudio de la biología, en especial las técnicas de secuenciación masiva, han transformado el estudio de la biología molecular, proporcionando grandes cantidades de datos en cada experimento. Esto ha requerido el desarrollo de herramientas computacionales que permitan el procesamiento primario de los datos crudos y su posterior análisis por medio de modelos estadísticos que extraen e infieren nuevo conocimiento, especialmente en sistemas biológicos donde las entidades de interés (como las proteínas) muestran relaciones de carácter complejo.

A nivel general, la BC ha sido pieza fundamental en el avance de numerosas áreas de la biología molecular. Una de las más importantes ha sido la identificación

de genes codificantes en secuencias genómicas crudas mediante el uso de modelos de Markov (Do y Choi, 2006), y el estudio de la homología de secuencias, que ha permitido la asignación de funciones biológicas a nuevos genes identificados en función de su homología con genes ya descritos (Lipman y Pearson, 1985). Asimismo, destaca la genómica comparativa (Hardison, 2003; Mullikin, 2014), donde el modelado estadístico ha sentado las bases para el estudio del árbol evolutivo de la vida. También, la predicción de la estructura secundaria y terciaria de las proteínas (Wardah, Khan, Sharma, y Rashid, 2019) ha permitido la predicción y estudio de las interacciones entre proteínas dentro de la célula. Por otro lado, el ensamblaje y reconstrucción de genomas (Giani, Gallo, Gianfranceschi, y Formenti, 2020; Wee y cols., 2019) ha permitido determinar características genotípicas relacionadas con rasgos fenotípicos en plantas y animales. Además, en el terreno de la biomedicina, la BC ha permitido un progreso notable en áreas como el reposicionamiento de drogas (Ezzat, Wu, Li, y Kwoh, 2019; Karaman y Sippl, 2019), que busca la aplicación de nuevas dianas terapeúticas a drogas ya existentes, o la medicina personalizada (Ashley, 2016; Davis, Kumbale, Zhang, y Voit, 2019), donde las alteraciones específicas de cada paciente son cuantificadas con el fin de diseñar terapias personalizadas de mayor eficacia que las terapias de uso general.

Asimismo, la BC ha sido pieza clave en la extracción de nuevo conocimiento a partir de importantes consorcios internacionales destinados a entender la genética humana y su impacto sobre la regulación génica. Destacan el proyecto *1000 genomes* (Abecasis, 2012) (proyecto pionero que permitió por primera vez secuenciar de forma masiva a diferentes poblaciones mundiales), el proyecto *HapMap* (The International HapMap Consortium, 2003) (destinado a la creación de un mapa de haplotipos humanos) y el proyecto *ENCODE* (Dunham y cols., 2012) (cuyo objetivo ha sido caracterizar la actividad funcional y transcripcional de regiones genómicas no contenidas en zonas codificantes).

En el contexto del cáncer, la BC ha permitido importantes avances a la hora de determinar alteraciones genómicas de diferente naturaleza, entre ellas, aberraciones cromosómicas (Li y cols., 2020; Yi y Ju, 2018), variaciones en el número copias

en determinadas regiones genómicas (Kanagal-Shamanna y cols., 2018), pequeñas inserciones y deleciones (Chen y cols., 2016), así como mutaciones somáticas puntuales (Cibulskis y cols., 2013; Goya y cols., 2010; Koboldt y cols., 2012; Roth y cols., 2012; Saunders y cols., 2012), siendo esta última, probablemente la vía más importante de estudio en el cáncer. Además, también se han aplicado numerosos enfoques computacionales en el análisis de rutas moleculares (Kuenzi e Ideker, 2020), el análisis filogenético de las diferentes subpoblaciones de clones existentes en un tumor (Eaton, Wang, y Schwartz, 2018; Watkins y Schwarz, 2018) y algunas aproximaciones interesantes que implican la integración de diferentes tecnologías ómicas (Chakraborty, Hosen, Ahmed, y Shekhar, 2018; Finotello y Eduati, 2018).

También, en el estudio del cáncer han surgido algunos consorcios muy relevantes que han permitido un aumento considerable del tamaño muestral, y con ello, la posibilidad de confirmar de forma fiable algunas alteraciones relevantes de baja frecuencia. Entre ellos destaca el consorcio americano de cáncer *TCGA* (Tomczak, Czerwinska, y Wiznerowicz, 2015) y su extensión a nivel internacional con el proyecto *ICGC* (Hudson y cols., 2010) que recoge más de 20000 individuos en 33 tipos de cáncer, contando además con diferentes ómicas, como la transcriptómica, la genómica o la proteómica. En este caso, la BC ha tenido un papel relevante en el análisis estadístico de las diferentes cohortes de individuos (Tomczak y cols., 2015), así como de algunas aproximaciones más globales, conocidas como análisis de *pan-cancer* integrando diferentes tipos de cáncer con el objetivo de encontrar patrones comunes (Alexandrov y cols., 2020; Gerstung y cols., 2020; Rheinbay y cols., 2020).

1.2.1. Interpretando el sistema

La base genética de muchas enfermedades resulta de gran complejidad. Una de las causas principales reside en los patrones de interacción de las proteínas en el interior de la célula y de cómo éstos orquestan las diferentes funciones esenciales a realizar. Bajo esta perspectiva, y más allá de las alteraciones específicas en cada gen o individuo, resulta vital analizar el impacto de las alteraciones con un punto

de vista más global, que permita evaluar su efecto en la célula, describiendo a ésta como un sistema complejo compuesto de diferentes elementos interconectados. Este concepto describe los fundamentos de la biología de sistemas (Tavassoly, Goldfarb, y Iyengar, 2018), que tiene por objetivo el estudio de las interacciones de sus elementos más básicos a la hora de realizar las diferentes funciones esenciales en la célula.

El uso de la BC en el ámbito de la biología de sistemas ha permitido grandes avances. Uno de los más relevantes ha sido la creación de repositorios y bases de datos donde acumular todo el conocimiento biológico generado hasta el momento, con el propósito de comprender la función de los genes en el contexto de las funciones celulares. Entre los repositorios más destacados figuran *Gene Ontology* (Ashburner y cols., 2000) (que define una ontología de términos biológicos para describir las distintas funciones celulares), *KEGG* (Kanehisa y Goto, 2000) (que proporciona una descripción gráfica de las rutas moleculares más importantes) o *Reactome* (Fabregat y cols., 2018) (que proporciona un entorno gráfico y computacional para interrogar las interacciones proteicas en el contexto de las rutas). Asimismo, otros repositorios generales como *Ensembl* (Zerbino y cols., 2018) o *Uniprot* (The UniProt Consortium, 2017) han sido de enorme importancia a la hora de aglutinar y federar la información para su acceso por cualquier miembro de la comunidad científica. Esta infraestructura ha permitido el desarrollo de numerosas herramientas computacionales para la cuantificación de los procesos celulares y su entendimiento en el contexto de las enfermedades.

1.2.2. Cuantificación de rutas moleculares

El análisis de expresión diferencial representa la vía más empleada en la BC para la comparación a nivel molecular de dos o más fenotipos (Byron, Van Keuren-Jensen, Engelthaler, Carpten, y Craig, 2016; Oshlack, Robinson, y Young, 2010). A pesar de que permite determinar con facilidad aquellos genes que muestran un comportamiento muy distinto entre las condiciones de estudio, la información proporcionada a nivel de gen suele ser a menudo compleja y descontextualizada,

especialmente cuando se busca una explicación más mecanística a los procesos internos que producen una determinada enfermedad. En este contexto, es habitualmente necesario profundizar en aquellos procesos biológicos en los que intervienen los genes que han resultado ser significativos, con el fin de poder realizar una interpretación más coherente de los resultados en un análisis real.

Con este fin, la BC ha proporcionado diferentes metodologías para recoger los estadísticos obtenidos a nivel de gen, y proyectarlos sobre las rutas moleculares en las que a su vez participan (Al-Shahrour y cols., 2007; Hidalgo y cols., 2016), con el propósito de realizar inferencia estadística sobre ellas. Para ello, es necesario recoger e integrar dicha información biológica con el fin de definir un vocabulario acotado de términos biológicos que constituirán las unidades básicas de análisis en este tipo de metodologías. En la práctica, los términos pueden constituir cualquier tipo de etiqueta biológica que pueda ser anotada a nivel de gen, aunque mayoritariamente serán funciones biológicas o rutas moleculares en las que los genes intervienen.

Los primeros métodos de cuantificación de términos fueron conocidos como métodos de sobrerepresentación funcional (del inglés *ORA, overrepresentation analysis*). En este tipo de metodologías se parte de un conjunto de genes seleccionados, obtenidos generalmente a partir de un análisis previo de expresión diferencial, y se determina qué términos aparecen anotados con una frecuencia mayor de lo esperado al compararla frente a un conjunto de genes de referencia formado habitualmente por el resto de genes del genoma. De forma intuitiva, si el porcentaje de genes anotados en la lista de genes de interés es significativamente superior al de la lista de referencia, diremos que se trata de una función o término enriquecido. Este planteamiento ha llevado a la aplicación natural de test estadísticos de comparación de proporciones, como el test exacto de Fisher (Fisher, 1925), o el test hipergeométrico (Harkness, 1965), el cual, tras la construcción de una tabla de contingencia, nos permite estimar el correspondiente p-valor como:

$$P(k_i) = 1 - \sum \frac{\binom{M}{n}\binom{N-M}{n-m}}{\binom{M}{n}}, \tag{1.1}$$

donde N es el número total de genes estudiado, n el número de genes seleccionados, M el total de genes anotados con el término y m el número de genes seleccionados

y anotados con el término.

Este enfoque representa la aproximación más usada en este tipo de estudios ya que permite la detección de funciones biológicas alteradas entre condiciones de estudio empleando únicamente un conjunto de genes de interés, cuya selección tiene lugar a partir de procedimientos de índole muy diverso, como el mencionado análisis de expresión diferencial, la presencia de mutaciones o cualquier otra característica que pueda ser medida mediante las técnicas de secuenciación actuales. Sin embargo, también muestra algunas limitaciones claras. En primer lugar, la lista de genes seleccionados suele implicar la aplicación de umbrales fijos de significancia estadística, donde algunos genes relevantes dentro del término podrían quedar por debajo del umbral debido a que muestran un cambio menos pronunciado. Además, todos los genes tienen el mismo peso dentro del análisis, con independencia de si han mostrado un valor más extremo en la comparación.

Parte de estas carencias fueron resueltas por los métodos de enriquecimiento funcional orientados a grupos de genes (del inglés *GSEA, Gene Set Enrichment Analysis*), los cuales representan la evolución natural de los métodos *ORA*. En este caso, la inferencia se realiza a partir de una lista o *ranking* ordenado de los genes, cuyo valor suele ser el del estadístico obtenido en la expresión diferencial. Esta aproximación permite evitar el uso de un umbral estricto de significancia y facilita que cada gen aporte un peso específico dentro del modelo (generalmente el de su estadístico), planteando el procedimiento como un análisis que trata de determinar si el conjunto de genes anotados a un determinado término biológico muestra una tendencia clara a agruparse hacia uno de los dos extremos de la lista. El método más usado de *GSEA* (Mootha y cols., 2003; Subramanian y cols., 2005) calcula un estadístico de enriquecimiento a partir de un recorrido acumulado a lo largo del *ranking*. Asimismo, otras soluciones posteriores emplearon aplicaciones iterativas del test exacto de Fisher (Al-Shahrour y cols., 2007), o la aplicación de modelos de regresión logística (Montaner y Dopazo, 2010).

A pesar de que los métodos *GSEA* corrigen algunas de las limitaciones de los métodos *ORA*, tienen en común el hecho de no tener en cuenta la topología

subyacente cuando los términos a analizar son rutas moleculares y se dispone de una red que nos permite entender cómo se produce la interacción entre las distintas proteínas que la forman. En este caso, los métodos *ORA* y *GSEA* tratan a las rutas como bloques uniformes donde todos los genes tienen un mismo rol. Esta limitación es particularmente delicada por el hecho de que a menudo, algunos de los genes incluidos en las rutas no tienen la función de promover su ejecución, sino de inhibir o regular su funcionamiento (Figura 1.2).

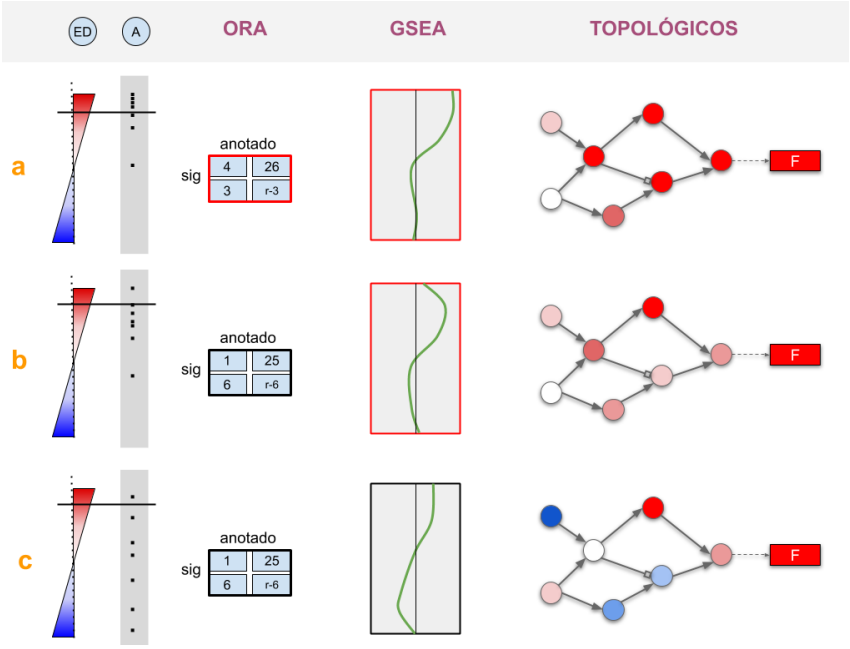


Figura 1.2: *Diagrama representativo de los diferentes métodos de cuantificación de rutas moleculares en 3 escenarios distintos: a) los genes pertenecientes a la ruta se agrupan de forma preferente clara en la parte alta del ranking; b) los genes pertenecientes a la ruta muestran una tendencia clara a agruparse en la parte alta del ranking, con algunos genes por debajo del umbral de significancia; c) los genes anotados a la ruta muestran una distribución uniforme a lo largo del ranking. En este caso, solo los métodos topológicos son capaces de detectar la alteración presente en la ruta.*

Para dar respuesta a estas limitaciones, en la última década han aparecido herramientas para modelizar y posteriormente analizar estadísticamente la topología de las rutas moleculares. Estos métodos generalmente abordan el modelado

matemático de las rutas mediante el uso de grafos, donde los nodos son los genes y las aristas sus interacciones (Amadoz y Hidalgo, 2018). Para ello, la información topológica es recogida a partir de bases de datos específicas como *KEGG* (Kanehisa y Goto, 2000) o *Reactome* (Fabregat y cols., 2018) que describen diagramas canónicos de las rutas moleculares más conocidas.

Uno de los métodos más representativos dentro de esta categoría ha sido la herramienta *SPIA* (Tarca y cols., 2009), la cual no solo modeliza los cambios en la expresión génica, sino también la naturaleza de las interacciones entre las proteínas incluidas y su posición dentro de la ruta molecular. Otra de las aproximaciones más interesantes fue propuesta por los autores de *CliPPER* (Martini, Sales, Massa, Chiogna, y Romualdi, 2013), donde la topología se modela haciendo uso de modelos gráficos. En este caso se realiza un análisis de partes específicas del grafo con el fin de poder detectar puntos calientes, potencialmente alterados, al comparar entre condiciones. También *DEGraph* (Jacob, Neuvial, y Dudoit, 2012) aportó un punto de vista diferente al problema mediante el diseño de una compleja metodología en la que los valores de expresión son analizados en el contexto de la topología a partir de un análisis basado en la transformada de Fourier. Después *DEAP* (Haynes, Higdon, Stanberry, Collins, y Kolker, 2013), con una aproximación más sencilla, permitía la obtención de un valor diferencial en cada ruta, a partir de la acumulación recursiva de los cambios de expresión observados en cada uno de los posibles caminos contenidos en el grafo. Poco después, *subSPIA* (Li, Shen, Shang, y Liu, 2015) aportó un nuevo método en el que se realiza un análisis en el contexto de la topología conectando de una forma óptima todos los genes diferencialmente expresados en el grafo, con el fin de detectar partes alteradas de la red.

Por último, *Hipathia* (Hidalgo y cols., 2016), uno de los métodos más recientes, se centra en el análisis de las rutas de señalización (Chacón-Solano y cols., 2019; Esteban-Medina, Peña-Chilet, Loucera, y Dopazo, 2019; Petkevicius y cols., 2019; Peña-Chilet y cols., 2019). En concreto, *Hipathia* descompone cada ruta en un conjunto de posibles subgrafos o caminos que conectan a las proteínas de entrada (típicamente receptores de membrana) con las proteínas de salida (típicamente

proteínas efectoras), lo que le permite modelar apropiadamente las diferentes funciones que puedan ser producidas dentro de la misma ruta. Para llevar a cabo la cuantificación, en cada camino *Hipathia* simula la propagación de una señal virtual que comienza en los nodos iniciales y se difunde hasta el nodo de salida. En este caso, el flujo de salida resultante, el cual constituye el valor de actividad de la cascada, dependerá del valor de expresión de cada uno de los genes que la señal debe atravesar. Esta aproximación tiene un marcado carácter biológico, ya que permite emular el flujo de información que ocurre en una cascada de señalización real cuando un ligando del medio extracelular se une a un receptor de membrana y desemboca en una cascada de interacciones que tiene por objetivo la activación de una proteína efectora. Entre las ventajas más importantes de *Hipathia* destaca el hecho de poder estimar un valor de actividad para cada individuo y no simplemente un valor diferencial en la comparación, como aportan la mayor parte de métodos. Asimismo, también modela de forma natural el rol de algunas proteínas que actúan como inhibidores dentro de la ruta, ya que el cómputo del flujo de salida en cada nodo de la red tiene en cuenta la naturaleza de las interacciones.

1.2.3. Modelado de la variabilidad

Con independencia de su naturaleza, los datos reales se componen de una mezcla de diferentes fuentes de variabilidad. Al realizar un estudio es habitual observar que las componentes más importantes se relacionan con los rasgos fenotípicos de los individuos seleccionados. Sin embargo, en mayor o menor medida, todos los datos cuantitativos cuentan con una serie de fuentes de variabilidad no deseadas. En la mayor parte de casos, se debe a aspectos relacionados con el protocolo de adquisición de datos, como fluctuaciones aleatorias en los aparatos de medida. Por otro lado, diseños experimentales que muestran un incorrecto balanceo entre grupos de algunas características fenotípicas relevantes (como la edad) pueden aportar fuentes de variabilidad no deseadas a los datos. En este contexto, la estadística juega un papel esencial, ya que de forma natural nos permite determinar cuando una diferencia observada entre grupos es realmente consistente.

En la práctica, cuado se realiza un estudio es habitual aglutinar todas las fuentes de variabilidad técnica en un término conocido como efecto *batch* (Goh, Wang, y Wong, 2017). Este término suele describir al conjunto de fuentes de variabilidad producidas durante la adquisición de los datos, debido a parámetros técnicos como el uso de diferentes aparatos de medida en cada muestra, la adquisición en días distintos o la realización del proceso medida por personal diferente. Estas fuentes de variabilidad han de ser necesariamente tenidas en cuenta con el fin de ofrecer resultados robustos en la fase de análisis, requiriendo habitualmente la aplicación de técnicas de normalización específicas para mitigar sus efectos (Johnson, Li, y Rabinovic, 2007).

A nivel general, se considera ruido a cualquier fuente de variabilidad no deseada. En ocasiones el ruido puede ser de carácter más aleatorio, pudiendo ser promediado mediante la simple repetición por medio de réplicas, pero en otras ocasiones, puede ser de carácter más sistemático, produciendo errores con un patrón más complejo, a menudo relacionado con la escala de los datos. En el contexto particular de la biología, los datos también cuentan con multitud de sesgos y fuentes de variabilidad no deseadas. Estas fuentes se relacionan principalmente con diferencias en la construcción de librerías genómicas, en la amplificación *PCR*, o en la secuenciación del material genético.

En el contexto del análisis de datos de secuenciación masiva, el cual comprende la mayor parte de estudios computacionales actuales en biología molecular, las lecturas obtenidas cuentan con errores producidos por el secuenciador al digitalizar los nucleótidos que las forman. Además, errores en algunas partes del protocolo de análisis primario, como el mapeo de las lecturas al genoma de referencia, aportan ciento de grado de inconsistencia que favorece la aparición de interpretaciones erróneas del perfil molecular de algunas regiones genómicas. Debido a esto, cualquier herramienta computacional destinada a inferir características genómicas a partir de lecturas de secuenciación masiva debe tener en cuenta algunos estimadores de ruido, como las calidades de secuenciación o las calidades de mapeo. Un contexto especialmente sensible a estos efectos lo constituye la predicción de mutaciones

somáticas en cáncer (Cibulskis y cols., 2013; Goya y cols., 2010; Koboldt y cols., 2012; Roth y cols., 2012; Saunders y cols., 2012), donde las alteraciones encontradas suelen estar soportadas por un número de lecturas muy reducido, habitualmente en el rango de lo esperado por ruido.

Estimación de componentes latentes

Una de las aproximaciones más interesantes en el modelado de la variabilidad ha sido el uso de técnicas de estimación de componentes latentes (ECL) (Kossenkov y Ochs, 2010; Stein-O'Brien y cols., 2018). El objetivo de este enfoque consiste en la obtención de una serie de patrones latentes en los datos que potencialmente constituyen los bloques básicos a partir de los cuales se forman las observaciones, obtenidas generalmente como una combinación lineal de éstos.

La ECL, a menudo conocida en el ámbito del tratamiento de señales como separación ciega de fuentes (del inglés *BSS*, *Blind Source Separation*), se ejemplificó en 1953 a partir del problema conocido como *Cocktail Party* (Cherry, 1953). Este ejemplo describe una reunión social donde diferentes personas conversan simultáneamente. Después, gracias a la existencia de diferentes grabaciones (u observaciones) del mismo evento, las técnicas de *BSS* son capaces de separar las voces individuales de cada una de las personas presentes en forma de componentes independientes.

La ECL ha tenido numerosas aplicaciones en el campo de la biomedicina, en particular en el procesamiento de datos de encefalografía (Cashero y Anderson, 2011; Sun, Liu, y Beadle, 2005), electromiografía (Petersen, Buchner, Eger, y Rostalski, 2017), o técnicas funcionales de imagen por resonancia (Ding, Lee, y Lee, 2013; Esposito y cols., 2005), entre otras.

Cuando se realiza un análisis real, la ECL trabaja bajo la hipótesis de que los individuos pertenecientes a los mismos grupos compartirán características genómicas comunes, que a la postre estarán representadas mediante componentes específicas. En este caso, es habitual construir un modelo donde el número de componentes

empleado suele ser muy pequeño en comparación con el número de individuos de la muestra o el número de variables. Por esta razón, es habitual también referirse a los métodos de ECL como métodos de reducción de la dimensionalidad, ya que son capaces de obtener una versión aproximada de los datos de entrada a partir de un conjunto mucho más reducido de variables.

En el contexto de la BC, la ECL es conocida habitualmente como factorización de matrices ya que, de forma habitual, los datos suelen representarse mediante matrices bidimensionales, donde las filas representan a las distintas variables de interés (como los genes) y las columnas a las distintas observaciones. Sus aplicaciones han sido principalmente relevantes en el *clustering* (Türkmen, 2015), el análisis de interacciones (Hofree, Shen, Carter, Gross, e Ideker, 2013), la deconvolución de patrones mutacionales (Alexandrov, Nik-Zainal, Wedge, Campbell, y Stratton, 2013; Bayati y cols., 2020) o la deconvolución de los distintos tipos celulares presentes en muestras tisulares (Gaujoux y Seoighe, 2012; Repsilber y cols., 2010).

Durante las últimas décadas se han propuesto numerosas estrategias para la extracción de componentes latentes a partir de un conjunto de señales de entrada. Las estrategias difieren principalmente en los requisitos o restricciones aplicadas tanto a las componentes como a las matrices de mezcla, siendo en general, un proceso de carácter iterativo.

Probablemente, el método de factorización más empleado sea *PCA* (*Principal Component Analysis*) (Hotelling, 1933). *PCA* trata de descomponer la matriz de entrada en un conjunto de componentes principales que representan direcciones de máxima variabilidad en el espacio multidimensional formado por las variables. Por su construcción, las componentes de *PCA* son ortogonales entre sí, y se ordenan de forma natural en función de la fracción de varianza explicada que capturan de la varianza total del modelo.

Existen diferentes aproximaciones para abordar el *PCA*, como el análisis de valores y vectores propios, análisis de factores o la factorización *SVD* (Eckart C, 1936; Hestenes, 1958) (*Singular Value Decomposition*). Esta última representa una

de las alternativas más populares ya que dispone de una solución analítica. *SVD* descompone la matrix de entrada X en 3 matrices:

$$X \approx USQ^T, \tag{1.2}$$

donde U y Q son ortogonales ($U^TU = 1$ y $Q^TQ = 1$) y representan los vectores singulares de la matriz. Por su parte S es una matriz diagonal proporcional a la desviación estándar asociada a los r vectores singulares. En una visión general vemos que:

$$F = US = USQ^TQ = XQ, \tag{1.3}$$

donde F representa la matriz de componentes principales y Q la matriz de mezcla.

La extensión natural de *PCA* es habitualmente representada por *ICA (Independent Component Analysis*). *ICA*, propuesto en 1984 por Jeanny Hérault y Bernard Ans y popularizado por Pierre Comon en 1994 (Comon, 1994), realiza una descomposición de la matriz de entrada en componentes latentes que muestran un alto grado de independencia estadística, lo que en la práctica elimina la restricción de ortogonalidad impuesta por *PCA* en el espacio multidimensional formado por las variables de entrada. *ICA* trata de maximizar la independencia estadística, lo que en la práctica equivale a minimizar la normalidad de la proyección de los datos sobre las componentes encontradas, empleando para ello momentos de orden elevado (como la kurtosis), así como otras métricas como la entropía negativa o la función tangente.

En la práctica, la tendencia normal en los datos es eliminada en un paso previo llamado *whitening* donde todos los momentos de orden 2 son iguales a 1. De esta forma, las componentes pueden ser no-ortogonales.

ICA define el proceso de factorización como:

$$X = AS, \tag{1.4}$$

donde X representa la matriz de entrada, A la matriz de mezcla y S el conjunto de señales latentes.

En el caso general (como el *Cocktail Party*) se asume que el número de compo-

nentes es igual al número de señales, entonces la matriz de mezcla se convierte en una matriz cuadrada de $n \times n$ elementos, de forma que:

$$Y = WX, \tag{1.5}$$

donde Y se corresponde con la estimación de las componentes latentes independientes y W se define como:

$$W = DPA^{-1}, \tag{1.6}$$

siendo D una matriz diagonal y no singular, y P una matriz de permutación.

Uno de los algoritmos más usados en este contexto es *FastICA* (Hyvärinen, 1999, 2013), el cual maximiza la entropía negativa, ya que ésta es una medida ideal para estimar la no-normalidad de los datos:

$$J(y) = H(y_G) - H(y), \tag{1.7}$$

donde y_G es un vector aleatorio distribuido de forma normal con la misma matriz de covarianzas que el vector y. En este caso, podemos representar la información mútua como:

$$I(y) = J(y) - \sum_i J(y_i). \tag{1.8}$$

Otro de los métodos más habituales en la ECL es *NMF* (Non negative matrix factorization). *NMF*, propuesta inicialmente por Paatero y Tapper como *positive factorization* (Paatero y Tapper, 1994) y reintroducida por Lee y Seung en 1999 (Lee y Seung, 1999), realiza una descomposición de las señales de entrada a partir de una combinación lineal de las componentes latentes, donde tanto la matriz de entrada, como las matrices de componentes y mezcla, han de ser necesariamente positivas. Esta restricción tiene un gran impacto sobre la factorización, ya que describe a las observaciones como una suma ponderada de partes más elementales. Esto supone una gran diferencia con otros métodos como *PCA* o *ICA*, ya que estos muestran valores negativos que han de ser cancelados al combinar diferentes componentes. Además, la restricción de positividad tiende a producir matrices de mezcla más dispersas, lo que en algunos contextos puede ser muy adecuado.

La *NMF* se representa habitualmente de la siguiente forma:

$$X \approx WH, \tag{1.9}$$

donde X indica la matriz de datos de entrada, W la matriz de componentes latentes y H la matriz de mezcla. *NMF* plantea la factorización a partir de la minimización de la siguiente expresión:

$$f = \|X - WH\|_f^2, \tag{1.10}$$

donde $\|.\|_f^2$ se corresponde con la norma de Frobenius.

A nivel general, todos los métodos de factorización proporcionan soluciones coherentes, siendo el contexto de aplicación el que debe definir la idoneidad de uno u otro método. En la práctica, las restricciones impuestas por los métodos los hace disponer de características deseables distintas.

A diferencia de *ICA* y *NMF*, *PCA* permite establecer un orden natural en las componentes encontradas, ya que éstas se ordenan en función de su varianza explicada en el modelo. Esto lo hace un método adecuado cuando el objetivo es realizar una reducción de la dimensionalidad de los datos. Además, *PCA* resulta un método ideal a la hora de evaluar la calidad de los datos de entrada y la posibilidad de encontrar diferentes efectos de *batch*, ya que nos permite determinar si la característica fenotípica principal que agrupa a las muestras bajo estudio se alinea preferencialmente con la primera componente, o por el contrario, existen otras fuentes de variabilidad no deseadas que capturan un porcentaje no despreciable de la variabilidad presente en los datos. Sin embargo, los requisitos de ortogonalidad hacen de *PCA* un método poco adecuado cuando los datos no muestran una dispersión normal sobre las direcciones máximas de variabilidad (Figura 1.3).

Por otro lado, a pesar de que tanto *ICA* como *NMF* no muestran un orden natural en las componentes, debido a su forma de factorizar permiten encontrar diferentes componentes asociadas a fuentes específicas de variabilidad, aunque éstas no capturen un porcentaje importante de la variabilidad total.

Por su parte, *NMF* es el método que proporciona una mejor interpretabilidad

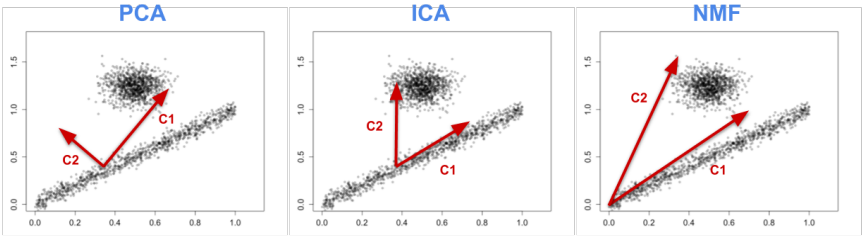


Figura 1.3: *Diferentes enfoques a la hora de factorizar un conjunto de datos con dos subpoblaciones latentes. En el caso de PCA, sus restricciones de ortogonalidad producen una estimación poco adecuada de las componentes, siendo ICA y NMF, más adecuadas para este contexto. A diferencia de ICA, las restricciones de no negatividad de NMF producen que sus componentes estimadas comiencen en el origen de coordenadas.*

de las componentes obtenidas, ya que, el hecho de no disponer de valores negativos que puedan ser cancelados entre componentes, le proporciona una interpretación más directa de los valores obtenidos en cada componente encontrada. En este caso, *NMF* encuentra de forma más natural las diferentes partes que componen el conjunto total de datos.

La Figura 1.4 representa un ejemplo ilustrativo de como *PCA*, *ICA* y *NMF* realizan el proceso de factorización, obteniendo unas matrices con características muy diferentes. En este caso se dispone de un corpus de imágenes faciales empleado para realizar la factorización, mostrando el vector de mezcla de una de las observaciones a título descriptivo. En primer lugar, se observa como *NMF* proporciona un conjunto componentes latentes que describen partes elementales de la caras, permitiendo una interpretación sencilla de la muestra de ejemplo como una suma ponderada de estas componentes. Por su parte, *PCA* proporciona unas componentes que no describen partes específicas, sino aspectos globales de las caras, haciendo muy complicada su interpretación. Por último, se aprecia como el enfoque de *ICA* también proporciona componentes con un carácter más elemental que *PCA*, aunque, también con valores positivos y negativos.

Además de los métodos aquí descritos, hay que resaltar que existen multitud de aproximaciones en la literatura, en especial, aproximaciones más probabilísticas

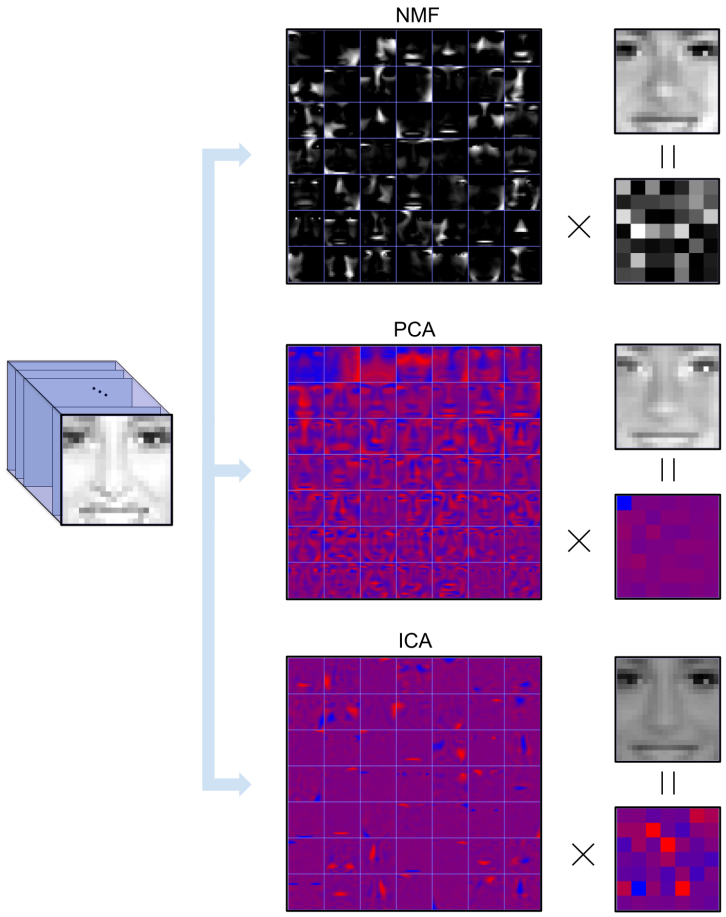


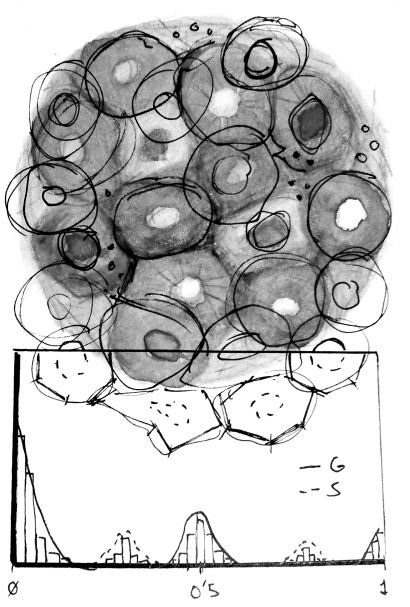
Figura 1.4: *Diferentes métodos de factorización aplicados a un corpus de imágenes faciales (http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html). En la imagen se muestra el conjunto de componentes obtenido con cada método, y el vector de pesos obtenido para la imagen de test.*

de *PCA* (Tipping y Bishop, 1999), *ICA* (Poppe y cols., 2013) y *NMF* (Bayar, Bouaynaya, y Shterenberg, 2014). Asimismo, cabe destacar que las metodologías descritas representan el caso bidimensional, aunque existen soluciones particulares que permiten extender la factorización a un número mayor de dimensiones. Es el caso de *Non negative Tensor factorization* (Kim, He, y Park, 2014), que proporciona una extensión natural a *NMF* mediante el uso de tensores. Estas aproximaciones son especialmente útiles cuando lo que se desea es realizar un proceso de factorización que integre diferentes niveles de información, como la integración de diferentes ómicas en BC.

Por último, también es importante destacar la existencia de numerosos trabajos donde las aproximaciones clásicas se modifican y extienden con el fin de poder añadir términos adicionales en las funciones de coste que permitan alcanzar soluciones más específicas. Es el ejemplo particular de la incorporación de términos de regularización para las matrices de componentes y mezcla, que contribuyan a la obtención de soluciones más dispersas. Por otro lado, la aplicación de modelos de factorización combinados que incluyan a más de una matriz representa también una expansión interesante. De hecho, se trata de la aproximación seguida en esta tesis, a través de la definición de un modelo de factorización jerárquico que analiza simultáneamente la actividad de genes y rutas moleculares.

CAPÍTULO 2

Alteraciones somáticas y variabilidad intra-celular en el cáncer



En este capítulo se presentarán una serie de metodologías estadísticas diseñadas para la obtención de un modelo de genotipado somático robusto, un proceso esencial dentro del análisis rutinario de muestras de cáncer. Para ello, en primer lugar se describirán algunas características intrínsicas que han sido previamente observadas en los tumores. Estas propiedades, tendrán un gran impacto en el diseño de los modelos estadísticos, ya que permiten explicar el enorme grado de heterogeneidad genómica observada en pacientes reales. Además, tendrá especial relevancia el modelado estadístico de la variabilidad técnica, ya que, debido a las limitaciones del contexto, una buena parte de las mutaciones somáticas encontradas serán compatibles con el efecto producido por un artefacto de ruido.

2.1. Heterogeneidad intra-celular en el cáncer

La heterogeneidad intra-celular constituye una de las características principales de cualquier tumor (Burrell y cols., 2013; Dagogo-Jack y Shaw, 2018), y son las condiciones de inicio y las particularidades de su entorno las que determinarán su evolución y grado de variabilidad. A pesar de que la secuencia de alteraciones genómicas que origina el comienzo de un cáncer suele ser desconocida, habitualmente implica la mutación de genes esenciales para el funcionamiento de la célula. Esta circunstancia desemboca progresivamente en una cascada de eventos que lleva a las células tumorales a sobre-estimular funciones relacionadas con el crecimiento y la replicación celular, y a inhibir procesos destinados al control, como la apostosis (Hanahan y Weinberg, 2011; Martincorena y Campbell, 2015).

En este contexto, la naturaleza del proceso de mutagénesis y la presión ejercida por el tejido normal facilita la aparición de diferentes subgrupos o comunidades dentro del tumor, a las que llamamos *clones*. Los clones muestran características comunes, ya que generalmente derivan de un progenitor común, pero también algunas características propias que adquieren durante su desarrollo. Esta descripción, muy próxima a la ecología, explica con gran facilidad una parte importante de la heterogeneidad observada, siendo las subpoblaciones de clones, y sus interacciones,

las responsables de construir la identidad global del tumor.

2.1.1. Expansión e interacción entre clones

La visión clásica del cáncer sitúa su inicio en una primera célula sana que, a partir de uno o varios eventos, adquiere las alteraciones específicas necesarias para arrancar el inicio de la enfermedad. Esta célula progenitora representa el primer ancestro de un linaje celular que irá mutando progresivamente y, en consecuencia, adquiriendo las habilidades necesarias para adaptarse al medio, colonizar el tejido circundante y, ocasionalmente, realizar metástasis a otros órganos. El linaje compartirá mutaciones somáticas, y por tanto, habilidades adquiridas, principalmente aquellas presentes en la célula progenitora que dio lugar al linaje.

En la masa tumoral, los cambios en el ADN suponen el motor principal de su desarrollo, ya que le permiten implementar constantemente nuevas estrategias de adaptación. Dichos cambios provienen de dos fuentes principales (Torre y cols., 2016; Vineis y Wild, 2014): eventos externos que producen daño en el ADN, como la luz ultravioleta o agentes contaminantes, y eventos internos, como fallos en la replicación cromosómica. De forma complementaria, ambos procesos irán generalmente acompañados de la inhibición de aquellas proteínas encargadas de detectar y corregir alteraciones en el genoma, lo que potencialmente perpetuará de forma indefinida las aberraciones aparecidas dentro del tumor.

Durante el desarrollo tumoral, cada célula cancerosa continuará proliferando y acumulando nuevas alteraciones, llevando al linaje a una situación de divergencia continua. En este contexto, el conjunto de mutaciones presentes en cada célula condicionará en gran medida su grado de adaptación, y por tanto, su capacidad de supervivencia frente a otras células del tejido. Este proceso lleva de forma natural a un escenario de competición celular, y a la observación de un fenómeno de presión selectiva que se describe de forma análoga al proceso de evolución de las especies descrito por Darwin (Darwin, 1859). En concreto, aquellas mutaciones que proporcionen una mejor adaptación al entorno incrementarán la supervivencia

de las células portadoras, produciendo, a su vez, que dichas mutaciones se fijen en la población celular. En contraposición, aquellas mutaciones perjudiciales acabarán mermando la supervivencia de las células afectadas, eliminándolas progresivamente del tejido.

Además del cambio fenotípico producido por el efecto de las mutaciones, es importante resaltar que la presión selectiva no será necesariamente uniforme a lo largo de toda la masa tumoral, ya que el acceso a recursos esenciales, como nutrientes y fuentes de oxígeno, no es completamente homogéneo en el tejido. Asimismo, también la interacción con el tejido sano adyacente condiciona el grado de presión selectiva al que es sometido cada célula tumoral, en especial, la exposición a células del sistema inmunitario (Gonzalez, Hagerling, y Werb, 2018). En este sentido, la expansión progresiva de la masa tumoral provoca que las células tumorales situadas en la periferia del tumor (de mayor contacto con el tejido sano) estén sometidas a una presión selectiva mayor. Esta circunstancia provoca que a menudo la propia la tasa mutacional pueda ser una característica bajo presión selectiva, ya que, aun a riesgo de producir alteraciones con efectos muy negativos en la célula portadora, éstas permitirán acelerar de forma significativa la búsqueda de nuevas configuraciones con mayor ventaja selectiva.

Con independencia del mecanismo implementado, la adaptación a nuevos entornos tendrá un papel esencial en la expansión del tumor. Destaca la transición de células epiteliales a mesenquimales multipotentes (Edge, 2016) (*EMT, epithelial-mesenchymal transition*) donde las células pierden su adhesión a otras células del tejido, ampliando así sus capacidades migratorias, y renovando su capacidad para diferenciarse en otro tipo celular. De igual forma, la metástasis (Lambert, Pattabiraman, y Weinberg, 2016) representa el caso más extremo de este proceso, ya que las células tumorales colonizadoras tendrán que ser capaces de adaptarse a un entorno muy distinto al de origen.

Diferentes modelos evolutivos

La mayor parte de mutaciones adquiridas durante el desarrollo del tumor son de carácter neutral, ya que en la práctica no modifican de forma sutancial la capacidad de adaptación al entorno de las células portadoras. Sin embargo, durante el proceso, y de forma cíclica, algunas células tumorales sufrirán alteraciones puntuales que aumentarán de forma considerable sus capacidades. Estas mutaciones, llamadas habitualmente *drivers*, llevarán a la descendencia directa de la célula mutada a una mejor adaptación al entorno, y por consiguiente, a una expansión progresiva en el tejido, desplazando no solo a las células sanas, sino al resto de células tumorales que no sean portadoras de dicho cambio. Este proceso, representado en la Figura 2.1, lleva a la creación de un nuevo clon dominante (Axelrod, Axelrod, y Pienta, 2006) y, de forma colateral, al consiguiente cuello de botella en el tumor, donde las mutaciones pertenecientes al resto de clones desaparecerán del tejido, fijando únicamente las del linaje dominante. Este proceso puede ocurrir de forma cíclica y producirse de dos maneras distintas: de forma abrupta, mediante la aparición de una o varias alteraciones cromosómicas simultáneas (Koltsova y cols., 2019; Marcozzi, Pellestor, y Kloosterman, 2018), o de forma gradual, donde la adquisición progresiva de cambios puntuales incrementará paulatinamente la capacidad de adaptación del linaje hasta el punto de expansión.

A pesar de que el modelo basado en un único clon dominante (Nowell, 1976) ha sido ampliamente usado para explicar la dinámica interna de un tumor, en la actualidad comienzan a proponerse otras alternativas (Axelrod y cols., 2006; Greaves y Maley, 2012) que permiten explicar de forma más sencilla el enorme grado de heterogeneidad intra-celular observado en pacientes reales, cuando estos son sometidos a diferentes biopsias a lo largo del espacio ocupado por el tumor (Gerlinger y cols., 2012). Más allá del gradiente de variabilidad observado en cualquier cáncer, cada vez está más aceptada la idea de que una coexistencia de varios clones dominantes pueda no solo ser algo viable, sino incluso potenciador para el desarrollo y propagación del tumor. Este punto de vista, muy alejado de la hipótesis monoclonal, cambia la percepción del tumor como un único conjunto de

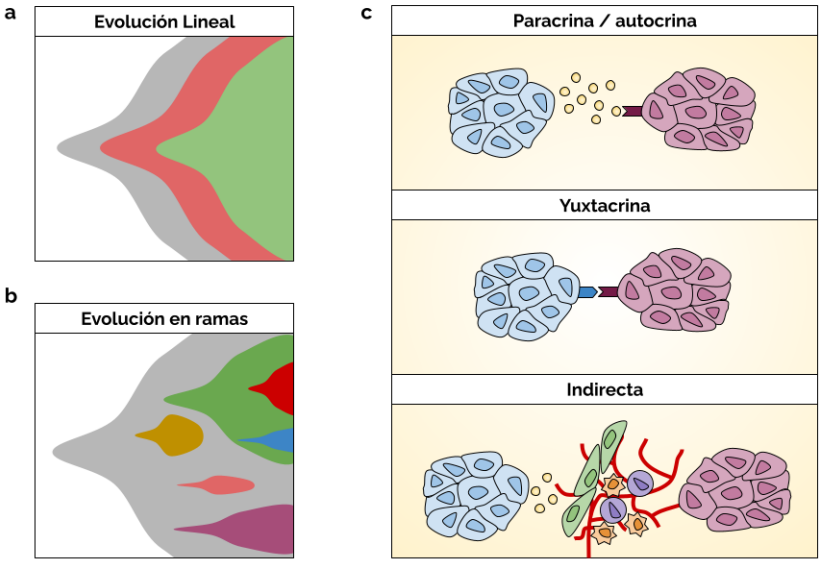


Figura 2.1: *Diferentes modelos evolutivos y mecanismos de comunicación entre clones en el seno de un tumor. a) Modelo de evolución lineal donde el clon dominante surge del mismo linaje. b) Modelo de evolución en ramas donde diferentes clones dominantes pueden coexistir en el mismo espacio. c) Mecanismos de comunicación entre clones.*

células estrechamente relacionadas y plantea una visión más general describiendo a una comunidad compuesta por diferentes subpoblaciones celulares, donde la identidad del tumor ya no corresponde mayoritariamente a un único clon dominante, sino a la propia mezcla de subpoblaciones. En este sentido, algunos autores han planteado puntos de vista muy próximos a la ecología (Maley y cols., 2017), con el fin de explicar que dinámicas internas estarían regulando la presencia de los diferentes clones, sus interacciones y, en última instancia, sus proporciones dentro de la masa tumoral.

Cuando varios clones coexisten en el mismo territorio celular se produce de forma natural un escenario de competición a nivel clonal. Aquel clon con mejores capacidades tendrá mayor probabilidad de expandirse por el tejido, mermando así la presencia de otros clones. Además, los clones más viables no solo tendrán mayor presencia debido a capacidades intrínsecas, como la velocidad de proliferación o su resistencia al efecto del sistema inmunitario, a menudo, algunos clones serán capaces de secretar al medio moléculas con un efecto citotóxico sobre el resto (Vivarelli, Wagstaff, y Piddini, 2012). Por otro lado, al margen del escenario clásico de competición, es importante resaltar que la existencia de varios clones dominantes en el mismo tumor a menudo describe una potencial co-dependencia entre ellos. Esta situación podría producir que la eliminación completa de un clon específico por parte del sistema inmunitario tuviera también un efecto claramente negativo sobre el resto. De esta forma, se describe un modelo evolutivo de cooperación entre clones (Greaves y Maley, 2012; Maley y cols., 2017), que no necesariamente elimina la competición natural por los recursos.

La interacción entre los diferentes clones puede producirse mediante diferentes mecanismos (Tabassum y Polyak, 2015), representados en la Figura 2.1c. La forma más sencilla de comunicación se produce mediante la señalización yuxtacrina, donde las células interaccionan por contacto físico directo, facilitando el trasvase de ligandos generados por una célula emisora hacia otra célula receptora. Asimismo, las células también interaccionan mediante la señalización paracrina, donde los productos secretados al medio extracelular por parte de un clon son también

recibidos por el resto de células existentes en el microentorno más local. Además, la interacción entre clones también puede producirse de forma más indirecta, donde los cambios producidos por una célula en el medio extracelular afectan de forma directa al resto de células. Un ejemplo claro lo constituye la secreción de factores angiogénicos, cuyo efecto provoca la formación de un nuevo sistema vascular. También la secreción de proteasas por parte de un clon permitirá a todas las células tumorales expandirse con mayor facilidad por el tejido sano.

Por último, cabe resaltar que la importancia en la cooperación clonal se ha validado muy bien en modelos animales, donde la implantación de un único clon dominante en roedores (*xenografts*) lleva a una merma considerable de su capacidad de expansión en el nuevo huesped, muy diferente a la observada después de implantar el tumor completo (Cleary, Leonard, Gestl, y Gunther, 2014).

2.1.2. Mutaciones en el estudio del cáncer

La búsqueda de alteraciones somáticas en los genomas de pacientes con cáncer representa la vía más importante de estudio dentro de este conjunto de enfermedades. Las alteraciones genómicas forman parte de la tumorigénesis y su localización exacta a nivel cromosómico determinará el impacto final sobre genes, rutas moleculares, y en última instancia, el fenotipo de la célula.

El abanico de posibles alteraciones somáticas se divide en diferentes categorías. A grandes rasgos podríamos hablar de cambios a gran escala, donde incluiríamos deleciones, inserciones y translocaciones de regiones genómicas en un rango mayor a 1kb, y cambios a pequeña escala, donde incluiríamos reordenamientos más pequeños, incluyendo también mutaciones puntuales, las cuales son el objeto principal de estudio en este capítulo de la tesis.

Gracias al aumento significativo en el tamaño muestral proporcionado por grandes consorcios de cáncer como el *TCGA* (Tomczak y cols., 2015) o *ICGC* (Hudson y cols., 2010) ha sido posible determinar de forma fiable un conjunto de genes alterados en cáncer de forma recurrente. Dichos genes, conocidos como *drivers*,

se distribuyen en dos categorías principales: oncogenes, cuyas alteraciones somáticas les permiten orquestar los programas transcripcionales necesarios para el desarrollo del cáncer, y los supresores de tumor, los cuales se encargan habitualmente de velar por la integridad del ADN y disparar los programas de muerte celular controlada en caso necesario, habitualmente deshabilitados en las células tumorales.

La caracterización del rol de un potencial gen *driver* suele realizarse a partir de ensayos experimentales muy concretos, en ocasiones complementados por la enorme cantidad de información biológica contenida en repositorios públicos (Ashburner y cols., 2000; Fabregat y cols., 2018; Kanehisa y Goto, 2000). Al margen de la vía experimental, el rol de un determinado gen puede ser evaluado mediante el análisis de la distribución de mutaciones sobre su región codificante (Vogelstein y cols., 2013). En este caso, una distribución relativamente uniforme se relaciona con un perfil supresor de tumor, ya que la proteína podría ser truncada por medio de alteraciones somáticas en múltiples sitios de su secuencia de nucleótidos. De forma contraria, los oncogenes muestran una distribución mucho más concentrada en sitios específicos de su secuencia, relacionados con los cambios estructurales que le proporcionan a la proteína su ganancia de función (Figura 2.2).

La identificación y caracterización de genes *driver* ha sido una de las tareas más importantes en el estudio del cáncer. Los resultados obtenidos a nivel experimental han sido habitualmente almacenados en multitud de bases de datos con el fin de compartir dicho conocimiento con el resto de la comunidad científica, resultando de gran relevancia para grupos de investigación con un carácter más computacional. Dentro de estas bases de datos destaca *COSMIC* (Tate y cols., 2019), la cual se encarga de mantener un censo actualizado de genes *driver*, que contiene además información sobre mutaciones somáticas. Asimismo, los proyectos *TCGA* (Tomczak y cols., 2015) e *ICGC* (Hudson y cols., 2010) han desarrollado interfaces *web* de acceso a sus datos, ofreciendo información acerca de la frecuencia mutacional en distintos tipos de cáncer.

De igual manera, el diagnóstico ha sido beneficiado por la caracterización de mutaciones frecuentes. Destacan algunas alteraciones, conocidas como biomarca-

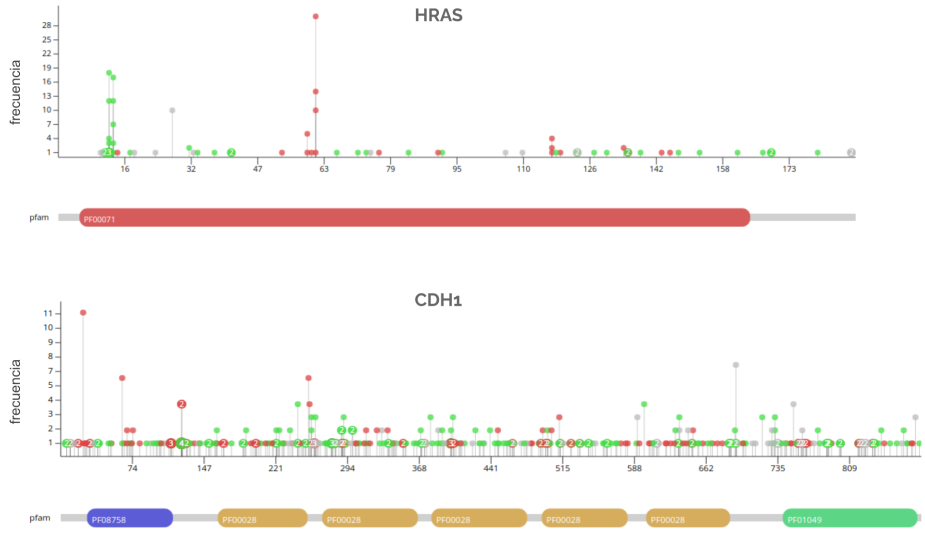


Figura 2.2: *Distribución de mutaciones a lo largo de la región codificante del oncogen HRAS y el supresor de tumor CDH1, obtenida del repositorio del proyecto ICGC. La figura muestra una acumulación excesiva de individuos mutados en dos posiciones específicas del oncogen HRAS correspondientes a las posiciones de ganancia de función. Por su parte, el supresor de tumor CDH1 muestra una distribución de individuos mutados más uniforme. Nótese la diferencia de escala entre ambos genes.*

dores, descritas en los genes *PML* y *RARA* en leucemia promielocítica aguda , mutaciones en *BRCA1* y *BRCA2* en cáncer de mama y ovario, VHL en cáncer de riñón, EGFR en cáncer de hígado y carcinoma pulmonar, *RAS* y *BRAF* en cáncer de tiroides (Nikiforov y Nikiforova, 2011) y *BRAF* en melanoma, entre otros.

2.1.3. Metodologías actuales para el genotipado somático

La detección de mutaciones somáticas representa una de las tareas más complejas dentro del análisis estadístico de un cáncer. La heterogeneidad intrínseca de los tumores produce que a menudo muchas de las alteraciones somáticas más relevantes estén presentes en una proporción celular baja de la muestra tumoral. La contaminación de tejido normal, por su parte, diluye aún más la muestra, lo que en la práctica se traduce en que las mutaciones somáticas candidatas son observadas en un porcentaje muy reducido de las lecturas que cubren una determinada región genómica de interés. Esta circunstancia describe el problema general, marcando el diseño de todo el conjunto de herramientas bioinformáticas desarrolladas durante los últimos años para poder dar solución a este problema. En concreto, la tarea más delicada se centra en determinar cuando un numero reducido de lecturas con el mismo cambio describe una mutación somática real, o por el contrario, solo cambios aleatorios o sistemáticos producidos por diferentes fuentes potenciales de error durante las distintas fases del protocolo, como la amplificación *PCR*, el proceso de secuenciación o el protocolo primario de análisis. Ofrecer una solución fiable al problema resulta de gran importancia, ya que, el conjunto de mutaciones somáticas encontradas en la muestra tumoral representa en muchos casos el punto de partida para un análisis más completo y sistemático que abarca no solo los genes afectados, sino aquellos procesos celulares o rutas moleculares que podrían estar alterados en un determinado cáncer.

La primera aproximación a la detección de mutaciones somáticas fue abordada mediante predictores de variantes germinales (Li, 2011; McKenna y cols., 2010). Este abordaje permitía detectar de forma fiable mutaciones somáticas presentes en un porcentaje celular significativo de la muestra, y por tanto, en un número

razonable de lecturas. Por el contrario, erraba a la hora de detectar de forma fiable mutaciones más diluidas ya que en este caso, la proporción de lecturas observada habitualmente se aleja del 50 % esperado para una mutación heterocigota.

Posteriormente, surgieron métodos específicamente diseñados para detectar mutaciones somáticas. Se han empleado enfoques estadísticos muy distintos, pero en general, la muestra tumoral suele ir pareada con una muestra de tejido sano circundante, que permite filtrar falsos positivos y determinar si una mutación somática es en realidad germinal. Además, en todos los casos se emplean filtros preestablecidos sobre algunos indicadores de ruido con el fin de eliminar de forma preventiva posiciones genómicas poco fiables.

Los primeros métodos específicamente diseñados para el genotipado somático se basaron en la aplicación de test de proporciones, como el test exacto de *Fisher* (Fisher, 1925). A nivel general, las herramientas hacen uso de los conteos de lecturas para estimar la verosimilitud de la hipótesis somática, frente a la hipótesis de ruido, o la existencia de una variante germinal diluida por efectos del muestreo aleatorio. *Varscan* (Koboldt y cols., 2012) fue una de las primeras herramientas basadas en esta idea, seguida posteriormente por *Mutascope* (Yost y cols., 2013), especialmente diseñada para modelos con amplicones, y *Shimer* (Hansen, Gartner, Mei, Samuels, y Mullikin, 2013), la primera en incorporar de forma eficiente una corrección de test múltiple.

También, los enfoques bayesianos tuvieron gran importancia en este ámbito. Dicha aproximación permitió diseñar modelos más generales y flexibles con capacidad para incorporar parámetros adicionales dentro del modelo que permitan estimar de forma más fiable la probabilidad somática frente a otras fuentes de variabilidad. En este contexto *SNVmix* (Goya y cols., 2010), y su extension *JointSNVMix* (Roth y cols., 2012), hicieron uso de modelos gráficos para plantear el modelo, implementando un optimizador basado en el algoritmo *EM* (del inglés *Expectation-Maximization*) (Dempster y Rubin, 1977) para inferir la frecuencia somática dominante dentro del tumor y emplear ésta durante el genotipado somático. También en el terreno bayesiano, *SomaticSnipper* (Larson y cols., 2012) planteó

un predictor somático que estima de forma conjunta la probabilidad a posteriori de los genotipos tumor y normal, determinando así cuando una mutación es realmente somática y no germinal. Posteriormente, *Strelka* (Saunders y cols., 2012) aportó un interesante modelo donde la muestra tumoral se modela como una mixtura de tejido normal, tumoral y ruido, integrando algunas fuentes típicas de error como el sesgo de hebra (*strand bias*), ofreciendo así excelentes resultados, especialmente en muestras secuenciadas mediante secuenciadores de tipo *Illumina*. El grado de contaminación normal también fue explícitamente incorporado dentro del modelado somático por parte de *Virmid* (Kim y cols., 2013), la cual realiza una estimación fiable del grado de contaminación normal para después corregir dicho sesgo en las frecuencias encontradas. Después *MuTect* (Cibulskis y cols., 2013) planteó un modelo donde la hipótesis somática, estimada teniendo en cuenta todo el espectro de variabilidad alélica dentro del tumor, se compara frente a la hipótesis de ruido, siendo especialmente sensible para mutaciones de baja frecuencia alélica. En este caso, la estimación del genotipo se hace empleando umbrales basados en experimentos previos, e incorporando el concepto de *panel de individuos sanos*, lo que permite el genotipado de muestras tumorales que no tengan explícitamente una muestra normal pareada.

También, se han diseñado otros enfoques más específicos para incorporar información poblacional sobre la estructura de haplotipos (Usuyama y cols., 2014), o cuando se dispone de múltiples biopsias del mismo paciente (Josephidou, Lynch, y Tavaré, 2015). También, han surgido diferentes herramientas de consenso en el campo. Concretamente, Goode y colaboradores (2013) combinaron *MuTect*, *JointSNVMix2* y *SomaticSniper* ofreciendo mejores resultados que las herramientas empleadas por separado.

El aprendizaje automático (*Machine learning*) también ha sido empleado en el pasado. En concreto, Ding y colaboradores (2012) entrenaron diferentes clasificadores con el fin de detectar mutaciones reales en posiciones candidatas. A pesar de lo interesante de su enfoque, la carencia de un modelo estadístico de genotipado ha limitado su uso posterior. Asimismo, las últimas técnicas de aprendizaje profundo

(*Deep Learning*) han realizado aportaciones en este contexto, ofreciendo interesantes resultados tanto en el genotipado somático (Mohammad y cols., 2019), como en el refinado posterior (Ainscough y cols., 2018). Por último, las últimas versiones de *Mutect2* y *Strelka2* han ofrecido mejoras con respecto a sus versiones anteriores, incorporando técnicas de ensamblado local.

2.2. Objetivos del capítulo

El presente capítulo de la tesis se centra en la detección robusta de las mutaciones somáticas presentes en una muestra tumoral. Para ello, tendrá especial relevancia el modelado estadístico de los artefactos de ruido presentes en las muestras a analizar y de cómo se emplea éste para distinguir de forma fiable mutaciones somáticas reales frente a cambios observados de naturaleza artefactual.

En un primer paso, se describe el análisis preliminar realizado para determinar si el conjunto de indicadores de ruido habitualmente calculado en estudios de genotipado es capaz de predecir con fiabilidad aquellas regiones del genoma que muestran una mayor acumulación de errores y por tanto una mayor susceptibilidad al ruido. En este apartado, los indicadores de ruido se calcularán sobre un conjunto de muestras de secuenciación recogidas para 4 organismos distintos, donde la comparación entre las diferentes versiones de sus genomas permitirá determinar con anterioridad donde se sitúan las regiones problemáticas a detectar.

Después de demostrar la utilidad de los indicadores de ruido, se describe el protocolo de análisis propuesto para la detección de mutaciones somáticas. En este caso, el modelado de los indicadores de ruido tendrá un papel esencial a la hora de predecir el número de cambios artefactuales esperados para cada base del genoma, lo cual permitirá de forma robusta determinar cuando una mutación somática es fiable a pesar de estar presente en pocas lecturas.

Por último, se describe la metodología seguida para simular de forma realista un conjunto de muestras tumorales, y sus correspondientes parejas de tejido normal.

Este punto tiene gran relevancia a la hora de evaluar de forma precisa el modelo estadístico propuesto para la predicción de mutaciones somáticas, y su comparación frente al resto de herramientas bien establecidas en el campo.

2.3. Descripción de la metodología de análisis

2.3.1. Protocolo de análisis primario

Las técnicas de ultrasecuenciación han supuesto importantes avances en el estudio de las enfermedades desde un punto de vista molecular. Con independencia de la técnica, la mayor parte de protocolos se basan en la secuenciación de fragmentos de ADN o ARN con el fin de poder identificar características particulares, como la expresión génica o la presencia de mutaciones en el genoma.

Existen variaciones en el protocolo dependiendo del tipo de secuenciación, pero a grandes rasgos, el protocolo de análisis primario se compone de los siguientes pasos:

1. *Control de calidad*:
   Supone el primer contacto con la muestra secuenciada y sirve para evaluar si la muestra puede ser empleada para análisis posteriores. En particular, se evalúan las calidades de secuenciación a nivel de nucleótido, especialmente en función de su posición en la lectura, la composición general en frecuencias nucleotídicas, el grado de complejidad de la librería, la distribución de longitudes o la presencia de lecturas duplicadas, entre otras.

2. *Eliminación de adaptadores y filtrado de secuencias de baja calidad*:
   En esta fase se eliminan los restos de adaptadores incluidos por el secuenciador durante el proceso de digitalización. Además, en ocasiones, se recortan las colas de las lecturas, generalmente de peor calidad, y en caso necesario, se elimina de forma directa las lecturas con una calidad media muy baja.

3. *Mapeo de lecturas*:

   El conjunto de lecturas de buena calidad se alinea contra el genoma de referencia, con el fin de poder reconstruir el genoma o transcriptoma de la muestra.

4. *Filtrado del mapeo*:

   Filtrado de lecturas con baja calidad de mapeo. Se realiza en base a las calidades definidas por el *software* de mapeo, indicando lecturas mal alineadas o lecturas con mapeos múltiples. Asimismo, se eliminan duplicados producidos durante la amplificación *PCR*.

5. *Análisis específico*:

   Describe cualquier tipo de análisis estadístico posterior realizado a partir de la muestra mapeada. Destacan los análisis de expresión génica, la búsqueda de alteraciones genómicas o la detección de zonas de regulación por parte de factores de transcripción.

2.3.2. Evaluación preliminar de los indicadores de ruido

Cuando se realiza el análisis de una región genómica particular es importante poder determinar el grado de fiabilidad observado en las lecturas que cubren la región, ya que éstas constituyen la materia prima para cualquier tipo de análisis posterior. Con este fin, es habitual disponer de un conjunto de estadísticos descriptivos que reflejan de forma sencilla el grado de incertidumbre producido durante el alineamiento. Los estadísticos se obtienen directamente de las lecturas que cubren la región y permiten realizar tanto una evaluación general de la región, como una evaluación más detallada de cada nucleótido cubierto, de especial relevancia en contextos como la búsqueda de mutaciones somáticas.

Existen diferentes indicadores que permiten evaluar la muestra desde diferentes puntos de vista. Uno de los indicadores más relevantes lo constituye la calidad de secuenciación. Se trata de un vector de valores proporcionado por el propio aparato de secuenciación para reflejar el grado de incertidumbre obtenido a la hora de

emitir cada nucleótido de la secuenciación. Se trata de valores enteros y codificados en formato *Phred* (Ewing y Green, 1998; Ewing, Hillier, Wendl, y Green, 1998) con el fin de ser almacenados de forma muy eficiente. A partir del valor de calidad es posible obtener la probabilidad de error de la siguiente forma:

$$P = 10^{-Q/10}, \tag{2.1}$$

donde P es la probabilidad de error y Q la calidad de secuenciación. Este indicador es particularmente importante en la predicción de mutaciones, ya que permite ponderar de forma natural cada lectura en el cómputo de cada genotipo.

Otro de los indicadores más habituales lo constituye la calidad de mapeo, también codificada en formato *Phred*. A diferencia de la calidad de secuenciación (definida con resolución de nucleótido), éste se trata de un valor proporcionado para cada lectura que el *software* de mapeo emplea para reflejar el grado de incertidumbre obtenido al tratar de alinear las lecturas. Mientras que valores bajos describen con mucha probabilidad lecturas mapeadas en sitios incorrectos, valores altos describen un probabilidad muy baja de error.

Además de los mencionados, existen otros indicadores que miden de forma más indirecta el grado de error en el mapeo. Por ejemplo, algunos indicadores informan sobre el número de nucleótidos recortados por el mapeador en los bordes de las lecturas, el número de pequeñas inserciones o deleciones (*indels*) presentes, o el número de alineamientos no primarios en la región de interés. En la Tabla 2.1 se muestra una relación completa de los indicadores de ruido empleados en este apartado de la tesis, donde además de las calidades de secuenciación y mapeo descritas, se incluyen algunos indicadores más específicos calculados *ex profeso*.

Descripción del protocolo

El objetivo de esta parte de la tesis consiste en determinar si los indicadores de ruido más habituales son predictores fiables a la hora de estimar el grado de error en una región genómica determinada. Para ello, se ha diseñado un protocolo destinado a la construcción de un nuevo estadístico que combina los distintos indicadores de

Acrónimo	Descripción	Tipo	Ámbito
BEP	Probabilidad media de error por base	Todas	QC
MEP	Probablidad media de error de mapeo	Todas	QC
BE MWZ	Estadístico Z de Mann-Whitney al comparar el error medio por base	Variantes	QC
ME MWZ	Estadístico Z de Mann-Whitney al comparar el error medio de mapeo	Variantes	QC
RP MWZ	Estadístico Z de Mann-Whitney al comparar la posición relativa del cambio en la lectura	Variantes	QC
CE MWZ	Estadístico Z de Mann-Whitney al comparar el número elementos en el código *CIGAR*	Variantes	QC
SOR	Razón obtenida al dividir el sesgo de hebra de lecturas con variante y referencia	Variantes	QC
AAF	Frecuencia alélica del alelo alternativo principal	Todas	VA
OAAF	Frecuencia alélica de los alelos alternativos secuendarios	Todas	VA
ND	Diversidad nucleotídica	Todas	VA
H	Heterocigosidad alélica	Todas	VA
PI	Parsimonia informativa	Todas	VA
PP	Proporción de lecturas mapeadas a una distancia correcta de su pareja	Todas	QC
MUF	Proporción de lecturas cuya pareja no fue mapeada	Todas	QC
IF	Frecuencia de *Indels*	Todas	QC
CF	Frecuencia de *clippping* en las lecturas	Todas	QC

Tabla 2.1: Relación de los distintos indicadores de ruido empleados en la evaluación de ensamblajes. Los estadísticos se agrupan principalmente en indicadores de control de calidad (QC) o de descripción de variabilidad alélica (VA)

ruido para determinar de forma conjunta el grado de error esperado en cada región (Carbonell-Caballero y cols., 2017). Después, el modelo resultante será evaluado en diferentes contextos donde el grado de error es conocido a priori.

Una de las estrategias más interesantes para evaluar el modelo de ruido se centra en determinar su capacidad para detectar regiones genómicas mal ensambladas en diferentes genomas de referencia (incluido el genoma humano), ya que después del correspondiente mapeo, estas zonas suelen mostrar indicadores de calidad bajos. En este contexto, la mayor parte de herramientas disponibles evalúan la fiabilidad

de un ensamblaje a nivel global. Algunas herramientas recientes han extendido esta aproximación con el fin de obtener una cuantificación específica para cada región del genoma. El ejemplo más característico es la herramienta *REAPR* (Hunt y cols., 2013), diseñada para evaluar un ensamblaje a partir del mapeo obtenido por una muestra de secuenciación del mismo organismo. En este caso, la muestra es secuenciada en el modo *paired-end* con el objetivo de encontrar variaciones estructurales incompatibles con el genoma. También destacan otras herramientas como *misFinder* (Zhu y cols., 2015) o *QUAST* (Gurevich, Saveliev, Vyahhi, y Tesler, 2013), las cuales emplean un abordaje parecido a *REAPR*, pero ayudado por los genomas de especies cercanas. Asimismo, otras herramientas han surgido en el contexto de genomas bacterianos (Walker y cols., 2014) o en el contexto de la metagenómica (Mikheenko, Saveliev, y Gurevich, 2016).

La nueva herramienta propuesta se describe de forma general en la Figura 2.3. La primera parte del protocolo consiste en la construcción del conjunto de distribuciones empíricas correspondientes a cada indicador de ruido incluido en el modelo. Para ello, como paso previo, se realiza una parcelación del genoma en ventanas de un tamaño definido, donde se promedia el valor de cada parámetro.

El valor de cada indicador x en una ventana W se calcula dependiendo de su naturaleza. En particular existen 2 tipos de indicadores: (i) aquellos definidos en cada nucleótido de la ventana y (ii) aquellos definidos únicamente sobre ciertas bases, como la mutaciones somáticas. Para el primer caso, el indicador x se calcula como:

$$x = \sum_{i \in W} \frac{r_i}{l}, \tag{2.2}$$

siendo

$$r_i = \sum_{j \in S} \frac{y_{ij}}{s}. \tag{2.3}$$

En este caso, l se corresponde al tamaño de la ventana, y r_i a la suma calculada en cada posición relativa i de la ventana, siendo y_{ij} el valor del indicador para la muestra j.

Cuando el valor de la ventana solo está definido sobre determinados nucleótidos

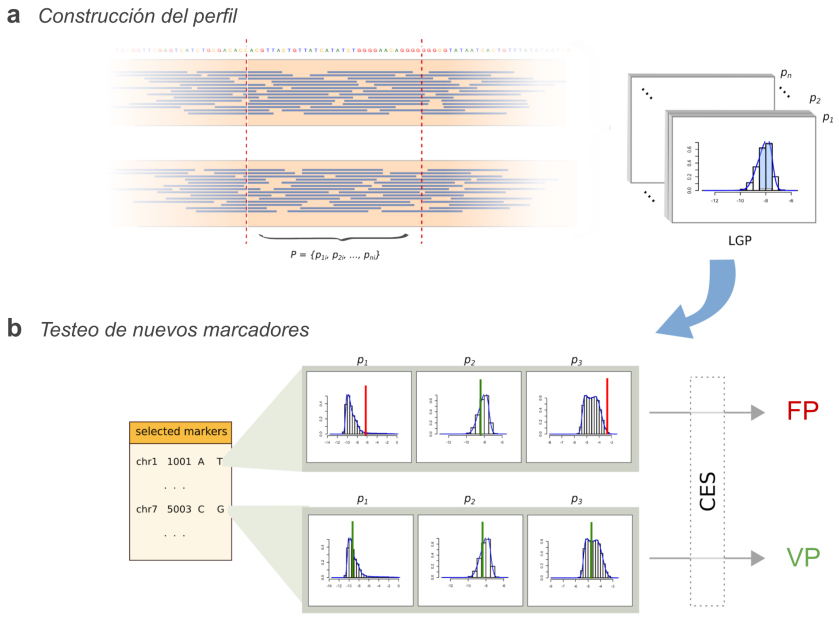


Figura 2.3: *Vista general del protocolo de evaluación de genomas. a) El genoma se parcela en ventanas de tamaño definido, donde las lecturas mapeadas son usadas para el cómputo de los indicadores. El conjunto total de valores obtenido para cada indicador define su distribución empírica. b) Los indicadores de ruido son calculados para las ventanas que contienen a las nuevas posiciones a evaluar. Después los valores de los indicadores son contrastados sobre sus distribuciones empíricas con el fin de determinar si sus valores se alejan de lo esperado.*

o muestras, el cálculo es similar, pero restringido a dichas posiciones:

$$x = \sum_{i \in V} \frac{r_i}{v}, \tag{2.4}$$

siendo

$$r_i = \sum_{j \in S_v} \frac{y_{ij}}{s_v}, \tag{2.5}$$

V el conjunto de nucleótidos evaluables en la ventana y s_v el conjunto de muestras evaluables, mientras que v y s_v son sus respectivos tamaños.

En un caso general, todos los nucleótidos de la ventana estarían ponderados de la misma forma a la hora de resumir el valor de cada indicador en la ventana. Sin embargo, en ciertas ocasiones puede ser necesario ponderar de forma especial una posición central en la ventana (como por ejemplo un nucleótido mutado). En estos casos, las ventanas se centran en cada nucleótido de interés, asignando mayor relevancia al centro de la ventana:

$$x = \sum_{i \in W} r_i w_i, \tag{2.6}$$

siendo

$$\sum_{i \in W} w_i = 1, \tag{2.7}$$

y w_i el peso asignado a cada posición relativa i dentro de la ventana. Si la ponderación se realiza de forma lineal, w_i se calcula de la siguiente forma:

$$w_i = \frac{W_c - |i - Wc|}{\sum_{k \in W} Wc - |k - W_c|}, \tag{2.8}$$

donde W_c se corresponde con el centro de la ventana.

Una vez realizado el resumen de los indicadores en cada ventana del genoma, se procede a construir una distribución empírica para cada indicador a partir del conjunto de valores obtenido. Este conjunto de distribuciones permitirá determinar el rango de variabilidad natural de cada indicador a lo largo de todo el genoma, y por tanto, determinar fácilmente cuando un valor extremo indica una región con una susceptibilidad anómala.

A la hora de evaluar una determinada ventana (o región alrededor de una base

en el modo ponderado), simplemente habrá que contrastar el valor resumen de cada indicador con su correspondiente distribución empírica almacenada. Después, el conjunto de p-valores empíricos obtenido por los indicadores es a su vez combinado con el fin de obtener un único valor resúmen, al que llamamos CES (del inglés *Combined Estimator Score*). Este valor, se obtiene aplicando el método de combinación de p-valores de *Fisher*, en particular:

$$\text{CES} = 1 - \Phi^{-1}\left(-2\sum_{q=1}^{m} log(p_q)\right), \tag{2.9}$$

donde Φ^{-1} se corresponde con la función de distribución (densidad) acumulada de una χ^2 con $2m$ grados de libertad.

Experimentos propuestos

Con el fin de evaluar la fiabilidad del estimador combinado, se plantean diferentes experimentos. En particular, se evaluarán diferentes ensamblajes pertenecientes a organismos reales que contienen errores en sus secuencias, para después determinar si el estimador combinado es capaz de detectar dichas regiones incorrectamente ensambladas.

Para establecer el grado de integridad de cada región genómica, el ensamblaje a evaluar se compara con una versión mejorada del propio genoma, correspondiente en la mayor parte de casos, a una versión del genoma obtenida posteriormente. En este caso, el grado de discrepancia obtenido en cada región genómica entre ambas versiones define el grado de error esperado. Para ello, se define un estadístico de similitud basado en el protocolo de alineamiento *BLAST* (Altschul, Gish, Miller, Myers, y Lipman, 1990). En concreto, el ensamblaje a evaluar se divide en ventanas de tamaño fijo (el mismo tamaño empleado en la construcción del modelo), para después mapear cada porción frente a la versión más nueva del genoma. En esta caso, el estadístico de similitud s se define como:

$$s = \sqrt{b_1 - b_2}, \tag{2.10}$$

donde b_1 y b_2 se corresponden con el estadístico de alineamiento obtenido con

BLAST para el primer y segundo *hit* respectivamente. De esta forma, valores altos del estadístico de similitud se corresponderán con regiones con un alineamiento de gran identidad y con un *hit* único, mientras que valores bajos se corresponderán con alineamientos pobres y mapeos de carácter múltiple

Una vez realizado el resumen de los indicadores en cada ventana del genoma, se procede a construir una distribución empírica para cada indicador a partir del conjunto de valores obtenido. Este conjunto de distribuciones permitirá determinar el rango de variabilidad natural de cada indicador a lo largo de todo el genoma, y por tanto, determinar fácilmente cuando un valor extremo indica una región con una susceptibilidad anómala.

Una vez realizado el resumen de los indicadores en cada ventana del genoma, se procede a construir una distribución empírica para cada indicador a partir del conjunto de valores obtenido. Este conjunto de distribuciones permitirá determinar el rango de variabilidad natural de cada indicador a lo largo de todo el genoma, y por tanto, determinar fácilmente cuando un valor extremo indica una región con una susceptibilidad anómala.

A la hora de evaluar una determinada ventana (o región alrededor de una base en el modo ponderado), simplemente habrá que contrastar el valor resumen de cada indicador con su correspondiente distribución empírica almacenada. Después, el conjunto de p-valores empíricos obtenido por los indicadores es a su vez combinado con el fin de obtener un único valor resúmen, al que llamamos CES (del inglés *Combined Estimator Score*). Este valor, se obtiene aplicando el método de combinación de p-valores de *Fisher*, en particular:

$$\text{CES} = 1 - \Phi^{-1}\left(-2\sum_{q=1}^{m} log(p_q)\right), \tag{2.11}$$

donde Φ^{-1} se corresponde con la funció

A la hora de evaluar una determinada ventana (o región alrededor de una base en el modo ponderado), simplemente habrá que contrastar el valor resumen de cada indicador con su correspondiente distribución empírica almacenada. Después,

el conjunto de p-valores empíricos obtenido por los indicadores es a su vez combinado con el fin de obtener un único valor resúmen, al que llamamos CES (del inglés *Combined Estimator Score*). Este valor, se obtiene aplicando el método de combinación de p-valores de *Fisher*, en particular:

$$\text{CES} = 1 - \Phi^{-1}\left(-2\sum_{q=1}^{m} log(p_q)\right), \tag{2.12}$$

donde Φ^{-1} se corresponde con la funció.

Para demostrar el funcionamiento del estimador propuesto (2.12) se plantea como primer experimento una evaluación del genoma humano. Esta elección se basa en que representa el organismo modelo más estudiado en biología molecular. Desde su primer borrador (Lander y cols., 2001), y hasta su primer ensamblaje estable (International Human Genome Sequencing Consortium, 2004) ha sufrido decenas de actualizaciones, a medida que las técnicas de secuenciación mejoraban. Actualmente, es considerado un genoma muy estable con pocas actualizaciones. Bajo este escenario, se plantean dos experimentos distintos. En el primero, un conjunto de regiones parcheadas en el genoma v37 (GRCh37, GCA000001405.1) fueron comparadas frente a un conjunto de regiones aleatorias, representando la variabilidad natural del genoma humano en términos de susceptibilidad al ruido. Para ello, 50 muestras humanas fueron descargadas del proyecto *1000 genomes* (Abecasis, 2012) y empleadas para construir el set de distribuciones empíricas ($l = 200$). Después, se obtuvieron los estimadores combinados para el conjunto total de regiones (parcheadas y aleatorias) y comparados para evaluar sus diferencias.

Como segundo experimento, se evaluó el protocolo en un contexto de genotipado. Concretamente, 30 exomas aleatoriamente seleccionados del proyecto *1000 genomes* (Abecasis, 2012) fueron descargados y genotipados para después ser comparados contra un *microarray* de genotipado sobre las mismas regiones. En este caso, el estimador combinado se comparó frente al número de discrepancias obtenidas al comparar el genotipado obtenido con técnicas de secuenciación frente al obtenido mediante el *microarray* de genotipado. Además, en este caso, el conjunto de distribuciones empíricas fue construido con otro set de individuos, pudiendo así

evaluar el grado de generalización obtenido por el protocolo a la hora de evaluar un conjunto nuevo de muestras.

Además, con el fin de poder realizar una evaluación más exhaustiva del protocolo, se plantean los siguientes experimentos sobre otros organismos modelo, cuyo genoma no está tan consolidado como el genoma humano. En concreto, se evaluaron los genomas de los siguientes organismos:

- *Arabidopsis Thaliana (Ath)*

  Representa uno de los organismos modelo más estudiados en biología de plantas. La última versión del genoma contiene alrededor de 136Mb y puede considerarse bastante estable. En este caso, el objetivo es comparar las dos últimas versiones de su ensamblaje, donde aquellas regiones del genoma anterior que hayan sido reensambladas en el nuevo genoma constituirán las regiones con artefactos a detectar. En la práctica, para cada ventana evaluada en el modelo estadístico se obtendrá el valor del estimador combinado y éste se comparará frente al estadístico de similitud basado en *BLAST*, que refleja el nivel de cambios sufrido entre genomas.

- *Saccharomyces cerevisiae (Sce)*

  Se trata de una especie de levadura cuyo genoma mide 12Mb. Es también un organismo eucariota modelo, siendo ampliamente empleado en biología molecular, de gran relevancia a la hora de evaluar interacciones entre genes esenciales. Para este experimento, se descargaron 79 muestras de levadura desde el repositorio *SRA* (Leinonen, Sugawara, y Shumway, 2011) (*SRP091984*) y se mapearon contra un ensamblaje realizado *ad-hoc* empleando las lecturas de una de las muestras (*SRR4446970*), generando así un genoma con numerosos errores. Después, se obtuvieron las distribuciones empíricas de cada indicador parcelando el genoma en ventanas de tamaño $l = 100$ nucléotidos. Por último, se comparó el estimador combinado para cada ventana con el estadístico de similitud basado en *BLAST*, obtenido al comparar el ensamblaje creado *ad-hoc*, con el actual genoma de referencia (*GCF_000146045.2 NCBI*).

- *Aeromonas hydrophilia (Ahy)*

Se trata de una bacteria heterotrópica presente en numerosos entornos humanos, incluyendo fuentes de agua dulce. Es resistente a numerosos antibióticos y causas algunas enfermedades humanas. Su genoma mide aproximadamente 5Mb y fue incluido en el proyecto *GAGE-B* (Magoc y cols., 2013) donde diferentes genomas bacterianos fueron ensamblados por diferentes herramientas bajo evaluación. Para esta comparación, se comparó el genoma ensamblado mediante la herramienta *Abyss* con el genoma de referencia oficial de *Ahy* (*NC_008570 NCBI*). Al igual que en el experimento anterior, el estimador de error combinado se comparó con el estadístico de similitud basado en *BLAST*.

2.3.3. Predicción de mutaciones somáticas

Una vez probada la capacidad de los distintos indicadores a la hora determinar la susceptibilidad al ruido en cada región genómica, el objetivo de este apartado de la tesis consiste en diseñar e implementar un predictor de mutaciones de somáticas que incorpore dichos indicadores de ruido, con el fin de reducir el número de falsos positivos. El modelo de genotipado somático desarrollado en este apartado de la tesis se presenta como una herramienta bioinformática bajo el nombre *SOM-hi*.

Modelado del ruido técnico

El protocolo de análisis comienza con la construcción de un modelo estadístico que tiene por objetivo la predicción del nivel de ruido esperado en cada posición del genoma. Para ello, el modelo toma como variables de entrada al conjunto de indicadores de ruido calculados en cada nucleótido, y devuelve como variable respuesta el número de cambios artefactuales (n_e) esperados para el conjunto de lecturas que cubren dicha posición genómica.

El entrenamiento del modelo se realiza mediante un predictor de tipo *Random Forest*, que permite modelar de forma muy eficiente el número de errores esperado.

En este caso, el predictor emplea una pequeña porción aleatoria de la muestra normal (típicamente 5Kb-10kb), de la que se excluyen aquellas posiciones en las que existe una probabilidad alta de mutación germinal. El objetivo en este caso consiste en aprender la estructura del ruido presente en la muestra empleando posiciones en las que no existe ninguna mutación y por tanto, solo cambios artefactuales. Teniendo en cuenta la tasa habitual de mutaciones germinales (10^{-3}) se puede considerar que la muestra de entrenamiento tendrá inicialmente muy pocas variantes germinales.

Una vez entrenado el modelo, éste se aplicará a cada posición genómica que se desee interrogar. Para cada posición se dispone de dos posible alelos, siendo a el alelo de referencia, y b un potencial alelo alternativo, donde $a, b \in \{A, T, C, G\}$. En cada posición se observan n lecturas, donde n_a y n_b describen el número de lecturas dando soporte al alelo de referencia y un posible alelo alternativo, respectivamente, siendo $n = n_a + n_b$ el conjunto total de lecturas disponibles.

La aplicación del modelo de ruido a cada posición candidata proporciona una estimación de los n_e posibles cambios artefactuales. Este dato nos permite recalcular de forma sencilla los conteos observados para cada alelo, obteniendo unos valores corregidos, donde los n_e estimados serán substituidos preferencialmente por alelos de referencia. En concreto:

$$\hat{n_a} = n_a + n_e, \tag{2.13}$$

$$\hat{n_b} = n_b - n_e. \tag{2.14}$$

Genotipado germinal

El genotipado germinal se realiza en primer lugar sobre la muestra normal. Su inferencia permite determinar si una determinada mutación presente en el tumor es de naturaleza somática o germinal. De forma intuitiva, si también la muestra normal contiene el mismo alelo alternativo, definiremos la variante como germinal, siendo necesariamente somática si la muestra normal no contiene dicho alelo en una proporción de lecturas razonable.

El genotipado germinal comienza con la definición de la variable discreta θ, definida como el genotipo existente en una determinada posición genómica a evaluar. Asumiendo dos copias por cada cromosoma, $G = \{aa, ab, bb\}$ describe el conjunto de posibles genotipos válidos para θ. La inferencia sobre θ se realiza a partir del conjunto de lecturas cubriendo dicha posición, en concreto n_a y n_b. A nivel práctico, obtendremos el valor más probable de θ a partir de la estimación de la probabilidad a posterior de cada genotipo posible (G). Para ello, definimos la probabilidad a posterior de θ a partir del teorema de Bayes:

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}. \tag{2.15}$$

La probabilidad a priori de θ se define mediante la tasa de cambio alélico por base (d) observada en estudios previos. Con un valor típico de 10^{-3}, correspondiente a una variante heterocigota cada 1000bp (*base pairs*), d se relaciona con θ de la siguiente forma:

$$P(\theta) = \begin{cases} 1 - d - d^2 & \text{si } \theta = aa \\ d & \text{si } \theta = ab \\ d^2 & \text{si } \theta = bb. \end{cases}$$

Para construir la función de verosimilitud de cada genotipo, definimos π como la probabilidad de encontrar un alelo alternativo en una lectura dada, estando su valor condicionado a cada genotipo posible de la siguiente forma:

$$\pi = \begin{cases} \epsilon & \text{si } \theta = aa \\ 0{,}5 & \text{si } \theta = ab \\ 1 - \epsilon & \text{si } \theta = bb, \end{cases}$$

siendo ϵ la tasa de error esperada (típicamente 10^{-6}).

En este caso, emplearemos un modelo binomial con parámetro π, cuyos fracasos y éxitos se corresponden con los conteos obtenidos para las lecturas con alelo de referencia (n_a) y alternativo (n_b) respectivamente. De esta forma, la probabilidad

asociada a cada genotipo posible se define como

$$P(n_b, n|\theta) = \begin{cases} Bi(n_b|n, \pi_{aa}) & \text{si } \theta = aa \\ Bi(n_b|n, \pi_{ab}) & \text{si } \theta = ab \\ Bi(n_b|n, \pi_{bb}) & \text{si } \theta = bb. \end{cases}$$

Por último, obtenemos la probabilidad a posteriori para cada genotipo como:

$$\forall i \in aa, ab, bb \rightarrow P(\theta = i|D) = \frac{P(\theta = i|D)P(\theta = i)}{\sum_{j \in \{aa,ab,bb\}} P(\theta = j|D)P(\theta = j)}, \tag{2.16}$$

donde el genotipo con mayor probabilidad a posteriori será elegido como el genotipo presente en cada posición.

Genotipado somático

El conjunto de posiciones genómicas que muestran al menos una lectura con cambio después de la corrección por el modelo de ruido representa el conjunto de posiciones candidatas. La distribución de frecuencias alélicas para este conjunto muestra claramente una composición heterogénea formada por posiciones referencia con cambios residuales no corregidos por el modelo de ruido, posiciones con mutaciones germinales que muestran un cierto grado de variación debido al muestreo aleatorio, posiciones con mutaciones germinales homocigotas con cambios no corregidos por el modelo, y finalmente, posiciones con cambios somáticos, algunos de los cuales mostrarán frecuencias cercanas al nivel del ruido. Con esta visión del problema, el genotipado somático de la muestra tumoral comienza con el ajuste de una mixtura de distribuciones normales sobre la distribución de frecuencias alélicas obtenida para el conjunto de posiciones candidatas. La mixtura se realiza empleando el paquete *MClust* (Scrucca, Fop, Murphy, y Raftery, 2016) de *R* (R Core Team, 2019) especificando un rango variable de posibles distribuciones, desde un mínimo de 4 (que cubriría de forma general a los 4 genotipos posibles), hasta un máximo definido por el usuario, donde se contempla la existencia de varias subpoblaciones de clones que puedan describir diferentes frecuencias alélicas dominantes. En este caso, el número óptimo de distribuciones será seleccionado empleado el criterio

BIC (Schwarz, 1978).

Después de ajustar la mixtura de distribuciones, aquella componente con la media más baja será considerada como la distribución que representa a la variabilidad residual en las posiciones con alelos iguales a la referencia, siendo el resto de distribuciones empleadas para designar a posiciones con mutaciones germinales y somáticas. Además, las proporciones obtenidas para cada distribución de la mixtura durante el ajuste nos permite definir las probabilidades a priori para los distintos genotipos sin necesidad de recurrir a estimaciones externas que podrían no representar correctamente a la muestra tumoral estudiada. En este caso, se define:

$$P(\theta_t) = \begin{cases} \rho_1(1-\phi/m)+\phi/m & \text{si } \theta_t = aa \\ \sum_{i=2}^{k} \rho_k(1-\phi/m) & \text{si } \theta_t \neq aa, \end{cases}$$

donde ρ_i se corresponde con la proporción de la distribución i-ésima en la mixtura, k al número óptimo de distribuciones empleado, ϕ al número de posiciones no candidatas ($n_b = 0$) y m el número total de posiciones interrogadas.

Por otro lado, la probabilidad a posteriori del genotipo referencia se define como:

$$P(\theta_t = aa|D) \propto Bi(n_b|n, \pi_1)P(\theta_t = aa), \tag{2.17}$$

donde π_1 se corresponde con la media de la distribución asignada al genotipo referencia, equivalente al parámetro π_{aa} en el genotipado germinal. Asimismo, la probabilidad a posteriori para la hipótesis de mutación se define como

$$P(\theta_t \neq aa|D) \propto P(\theta_t \neq aa) \sum_{i=2}^{k} Bi(n_b|n, \pi_i)\rho_i', \tag{2.18}$$

donde ρ_i' se corresponde con la proporción relativa de la distribución i-ésima de forma que $\sum_{i=2}^{k} \rho_i' = 1$.

Finalmente, el genotipo final se define en función de las probabilidades a posteriori de θ en el tumor y del genotipo normal obtenido durante el genotipado

somático.

$$\theta_t = \begin{cases} aa & \text{si } P(\theta_t = aa|D) > P(\theta_t \neq aa|D) \\ ab & \text{si } P(\theta_t = aa|D) < P(\theta_t \neq aa|D) \text{ y } \theta_n = ab \\ bb & \text{si } P(\theta_t = aa|D) < P(\theta_t \neq aa|D) \text{ y } \theta_n = bb \\ s & \text{si } P(\theta_t = aa|D) < P(\theta_t \neq aa|D) \text{ y } \theta_n = aa, \end{cases}$$

donde s representa al genotipo somático.

Por último, después del genotipado somático, cada posición candidata será evaluada para detectar la posible presencia de artefactos inusualmente extremos. Estas posiciones, caracterizadas por mostrar valores extremos en algunos indicadores de ruido, no serán bien capturadas por el modelo, produciendo así falsos positivos en el resultado. Para corregir este efecto, y de forma análoga al apartado anterior, se estimará el p-valor empírico de cada indicador de ruido, para después ser combinados mediante el método de *Fisher*. En este caso, si el p-valor combinado resulta ser significativo, se considerará que existen evidencias suficientes para eliminar dicha posición del resultado.

2.3.4. Evaluación del modelo de genotipado somático

Tal y como se ha descrito, la predicción de mutaciones somáticas constituye un proceso complejo, principalmente debido a efectos adversos como la heterogeneidad intra-celular o las condiciones de ruido y contaminación normal que afectan a cada muestra. Estos factores constituyen una fuente importante de falsos positivos para todas aquellas herramientas de análisis que trabajan a partir del mapeo realizado después de la secuenciación.

Debido a esto, es necesario contar con métodos de validación que permitan cuantificar de forma exhaustiva la sensibilidad y especificidad de las herramientas usadas, en particular, las de genotipado somático, especialmente en aquellos casos donde la frecuencia alélica de las mutaciones esta comprendida en el rango de lo esperado por ruido.

En un contexto general, las validaciones experimentales constituyen la alternativa más fiable. Sin embargo, en el campo del genotipado somático esta aproximación resulta complicada, ya que la propia heterogeneidad genómica dificulta también el diseño de los adaptadores empleados por las técnicas de validación experimentales. Este hecho explica la casi total inexistencia de estudios experimentales que incluyan la validación de un número suficientemente amplio de mutaciones genómicas, para ser empleado posteriormente como referencia en el diseño de nuevas herramientas computacionales. Una de las principales excepciones lo constituye el consorcio *SEQC* (Fang y cols., 2019). Se trata de un proyecto muy longevo destinado a evaluar como el uso de diferentes tecnologías de secuenciación condicionan la aparición de falsos positivos en los protocolos de análisis bioinformático. En este caso, el consorcio proporciona una pareja de muestras (tumor y normal) secuenciada mediante diferentes tecnologías, y diferentes centros de secuenciación, proporcionando así un conjunto de datos muy fiable.

Debido a esto, en este contexto resulta interesante recurrir al diseño de métodos de simulación que permita recrear las condiciones que conforman a una muestra tumoral real. Existen pocas herramientas diseñadas para esta tarea. Cabe destacar *tHapMix* (Ivakhno y cols., 2017) y *OncoSimulR* (Diaz-Uriarte, 2017), que aunque de gran utilidad, no permiten definir con total control los diferentes parámetros clave en la simulación, como el número de ciclos de la simulación, el tamaño de la población, así como otros parámetros que definen la probabilidad de mutación en cada ciclo. Asimismo, existen otros trabajos interesantes centrados en el estudio de la evolución de un tumor desde un punto de vista más físico (Ghaffarizadeh, Heiland, Friedman, Mumenthaler, y Macklin, 2016), aunque sin la capacidad de generar muestras de secuenciación que puedan ser empleadas por las herramientas de análisis posterior.

Simulación de linajes tumorales

En este apartado se describe el diseño de una herramienta de simulación de tumores que se ha desarrollado para validar y comparar las herramientas

de genotipado somático. El objetivo general de esta herramienta consiste en la simulación de una población de células tumorales y sus células normales adyacentes, con el fin de producir como resultado un conjunto de mutaciones somáticas cuya distribución de frecuencias alélicas resulte similar a la de un tumor real. Para ello, durante un número de ciclos definido por el usuario, cada célula incluida en la población celular irá progresivamente mutando, dividiéndose y desapareciendo en función de los parámetros de presión selectiva surgidos a partir de las mutaciones somáticas introducidas durante el proceso. En este caso, cada mutación contribuirá al *fitness* de las células portadoras, definido como la capacidad de adaptación al medio. Al final del proceso, la población celular final será empleada para generar una muestra de secuenciación a la que se le incorporará cierta cantidad de ruido y contaminación normal, con el fin de dotarla de realismo. Este diseño permite disponer tanto de la secuencia de nucleótidos del genoma de cada célula incluida en la simulación, como del conjunto total como una muestra agregada, directamente comparable a una muestra de secuenciación real.

La simulación parte de un conjunto finito de células normales, donde cada célula consta de 2 haplotipos (o copias cromosómicas) sobre los que se irán acumulando las mutaciones generadas durante cada paso de la simulación. Seguidamente, a este conjunto de células se le añaden 1 o más células tumorales progenitoras, que constituyen los ancestros de todas las células tumorales simuladas durante el proceso. Después de inicializar la población celular, y a partir de la tasa de mutaciones por base definida por el usuario (μ), se simula el conjunto de mutaciones germinales que estará presente en todas las células de la población. En este conjunto de mutaciones se define $\textit{fitness} = 0$, describiendo así su carácter neutral. Después se añaden 1 o más mutaciones somáticas a los progenitores tumorales, las cuales, con $\textit{fitness} \gg 0$, aportarán la primera ventaja selectiva que las células tumorales muestran frente a las normales.

Para cada ciclo, la herramienta evalúa el destino de cada célula de la población. Para ello, en primer lugar se evalúa la aparición de nuevas variantes somáticas haciendo uso de los parámetros π_s^T y π_s^N, que definen la probabilidad de aparición

de una o más mutaciones somáticas en un ciclo dado para las células tumorales y normales respectivamente. Si finalmente se incorporan mutaciones somáticas, en un segundo paso se decide el número de mutaciones a introducir en la célula. Para ello, se define la probabilidad de cada cantidad como:

$$P(n_s = x) = \frac{\pi_s^x}{\sum_{i=1}^{m} \pi_s^i}, \tag{2.19}$$

donde n_s se corresponde con el número de mutaciones somáticas y m con el número máximo de mutaciones permitido.

Después de generar el conjunto de mutaciones, es necesario definir el *fitness* asociado a cada mutación generada. En particular, el *fitness* se distribuye según una gaussiana centrada en 0 y con varianza σ_T^2 si la célula es tumoral, o centrada en 0 y con varianza σ_N^2 si la célula es normal. En este caso, $\sigma_T \gg \sigma_N$ lo que permite reflejar tanto el aumento de *fitness* con respecto a las células normales, como la posibilidad de originar mutaciones inviables en las células tumorales.

Después de la fase de mutación, se recalcula el *fitness* de cada célula en función del *fitness* aportado por cada una de sus mutaciones:

$$f_c = \sum_{i}^{v} f_i + r. \tag{2.20}$$

En este caso, el *fitness* celular es empleado para determinar qué conjunto de células muestra una mejor adaptación al medio. De esta forma, aquellas células con mejor *fitness* podrán dividirse en la siguiente iteración, siendo el resto eliminadas de la población. Tal y como se aprecia, el *fitness* de cada célula es modificado mediante el término r que se distribuye como $N(0, \sigma_R)$. Esta perturbación aleatoria permite reflejar de forma sencilla las variaciones en la presión selectiva debido a características dinámicas del medio como la disponibilidad de nutrientes a lo largo de tejido. En la práctica, este término permite que algunas células con menor *fitness* puedan sobrevivir en la población, evitando así una expansión excesiva y poco realista de los clones dominantes. La Tabla 2.2 se muestra una relación de los parámetros de la herramienta, junto a los valores típicos empleados en las simulaciones de esta tesis.

Al final de la simulación, la herramienta permite disponer de una población de

Parámetro	Descripción	Valor por defecto
N	Número de iteraciones empleadas en la simulación	1000
Tamaño de la población	Número de células empleadas en la simulación	100
Número de progenitores	Número de progenitores tumorales	1
μ	Tasa de mutación germinal por nucleótido	0.001
π_s^N	Probabilidad de aparición de 1 o más mutaciones somáticas por unidad de tiempo en células normales	0.01
π_s^T	Probabilidad de aparición de 1 o más mutaciones somáticas por unidad de tiempo en células tumorales	0.75
Mutaciones iniciales	Número de mutaciones somáticas iniciales en el progenitor somático con *fitness* muy positivo	1
Tasa homocigótica	Tasa de mutaciones homocigóticas	0.1
TiTv	Tasa de transiciones/transversiones	2.3
σ_N	Desviación estándar en la distribución de *fitness* normal	0.1
σ_T	Desviación estándar en la distribución de *fitness* tumoral	2
σ_R	Desviación estándar aleatoria *fitness* tumoral	2

Tabla 2.2: Parámetros definidos en la simulación de linajes tumorales.

células que han experimentado un proceso de presión selectiva similar al que podemos encontrar en las células tumorales que componen una muestra real. Después, se generan los haplotipos correspondientes a cada célula de la población, empleando para la generación de lecturas un patrón de calidades similar al que muestran los aparatos de secuenciación reales. Para ello, se hará uso de la herramienta *ART* (Huang, Li, Myers, y Marth, 2012), diseñada para tal propósito.

Así pues, se plantean dos conjuntos de simulaciones destinados a evaluar la sensibilidad y especificidad de la herramienta de genotipado en función del grado de contaminación normal y del nivel de ruido aleatorio respectivamente. En concreto, se han generado simulaciones con 2 esquemas distintos:

- Contaminación normal: de 0 a 90 %, a intervalos de 10 %.
- Niveles de ruido: de 0 a 9 niveles, a intervalos de 1.

2.4. Resultados

2.4.1. Evaluación de los indicadores de error

Los indicadores de error fueron evaluados mediante el protocolo propuesto. En concreto, se construyeron los modelos de error correspondientes y se evaluó la eficacia del estimador de error combinado en el contexto de cada experimento.

En primer lugar se analizaron las regiones parcheadas del genoma humano. Tal y como se aprecia en la Figura 2.4a, se obtuvieron importantes diferencias en la distribución del estimador combinado para el conjunto de regiones aleatorias y el conjunto de regiones parcheadas. De igual forma, en el experimento de variantes, el estimador combinado mostró una clara tendencia creciente a medida que el número de diferencias entre las dos tecnologías de genotipado empleadas aumentaba (Figura 2.4b). Este punto además fue corroborado por la mayoría de indicadores individuales.

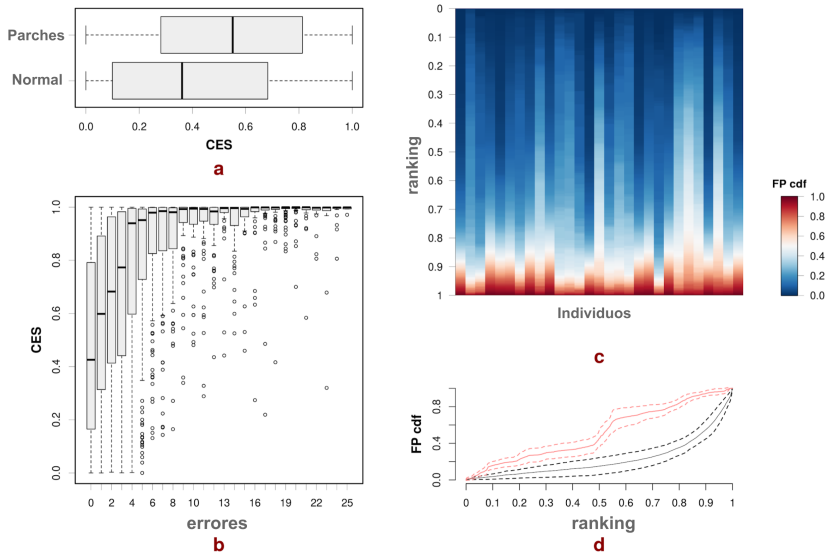


Figura 2.4: *Evaluación del estimador combinado en diferentes escenarios; a) distribución de valores para el conjunto de regiones aleatorias y parcheadas; b) distribución de valores para diferentes rangos de error obtenidos al comparar el genotipado por ultrasecuenciación frente al genotipado por microarray; c) distribución acumulada de falsos positivos en el ranking obtenido al ordenar las posiciones genómicas de 20 muestras en función del estimador combinado; d) perfil de densidad acumulada de falsos positivos en el ranking definido por el estimador combinado (negro) y por la herramienta REAPR (rojo).*

Después, se evaluó si el estimador combinado es capaz de ofrecer a priori una manera natural de ordenar las posiciones genómicas genotipadas en función de su probabilidad estimada de error. En este caso, tal y como se aprecia en la Figura 2.4c, el orden proporcionado por el modelo estuvo estrechamente relacionado con la densidad de falsos positivos encontrada, mostrando una acumulación clara de falsos positivos al final del *ranking*. La Figura 2.4d muestra una vista resumida del *ranking*, donde se compara con el mismo resultado proporcionado por la herramienta *REAPR*. En este caso se aprecia como *REAPR* muestra una distribución de falsos positivos más uniforme a lo largo del *ranking*.

Asimismo, se analizaron los modelos obtenidos para los otros tres organismos modelo. En concreto, la evaluación del ensamblaje de *Ath* mostró gran similitud entre las dos versiones (*TAIR8* y *TAIR10*) de genoma evaluadas (Figura 2.5a), lo cual confirma el reducido número de actualizaciones aplicadas al genoma en los últimos años.

De forma similar, los indicadores para el ensamblaje *de novo* de *Sce* también mostraron un alto grado de similitud. Con 2337 *scaffolds* y un tamaño de 11,669,271bp (95 % del genoma original, N50=61,488bp), el ensamblaje muestra una distribución de valores de similitud centrada en valores altos (Figura 2.5b), lo que sugiere que el ensamblaje construido únicamente con lecturas cortas produce un genoma razonable. De forma contraria, la evaluación del ensamblaje *de novo* descargado para *Ahy* mostró una gran parte de la densidad de distribución sobre valores bajos, lo que sugiere que el ensamblaje obtenido tendría mucho margen de mejora.

La distribución de valores para los indicadores individuales de ruido mostraron también grandes diferencias entre los genomas representados por un ensamblaje aceptable (*Ath* y *Sce*) y el genoma de *Ahy* (Figura 2.6). Notablemente, el indicador *MEP* y los indicadores basados en el test de *Mann-Whitney* muestran una clara tendencia descendiente cuando el estadístico de similitud también decrece. También, los indicadores de variabilidad alélica (*AF*, *ND*, *H* y *PI*) mostraron una tendencia similar en todos los casos, especialmente acusada en el genoma de *Ahy*, demostrando

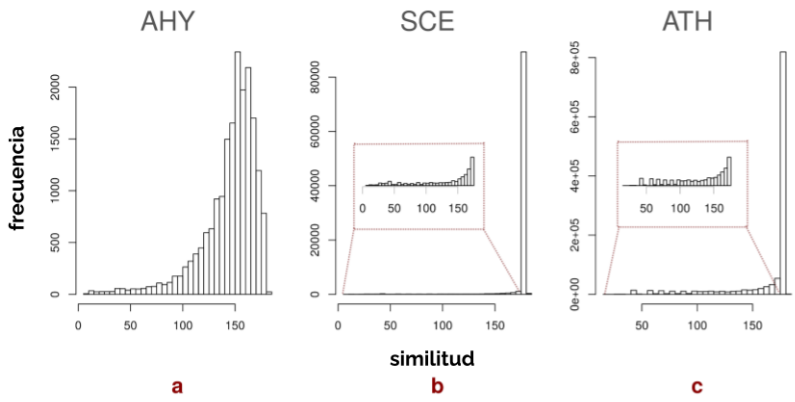


Figura 2.5: *Distribución del estadístico de similitud en los tres organismos modelo analizados.*

una relación robusta entre la susceptibilidad al ruido y el número de diferencias encontradas en la región de interés con respecto al genoma de referencia. En especial, se aprecia una tendencia clara para los indicadores individuales cuando se observa un valor de similitud por debajo de 120, lo que pone de manifiesto la utilidad de los indicadores cuando se aprecian diferencias sutiles en el ensamblaje. Por otro lado, los indicadores *CP*, *PP*, *MUF* e *IF* son especialmente sensibles en *Ahy* y *Sce*, pero con menor potencia estadística en *Ath*, reflejando de nuevo las diferencias entre ensamblajes de calidad aceptable y los ensamblados basados en lecturas cortas.

Respecto al estimador combinado de error, se aprecia gran concordancia con los valores de similitud obtenidos (Figura 2.7). En particular, se observa como valores próximos a 1 para el estimador combinado coinciden con valores de similitud por debajo de 120, mostrando de forma clara errores en el ensamblaje de los genomas.

El conjunto total de ventanas cuyo estimador combinado mostró valores estadísticamente significativos en *Ath* agrupó un total de 2,187,900bp (1.8 % del genoma) con señales claras de artefactos. Asimismo, el estimador mostró en *Sce* 321,800bp (2.8 % del genoma) y 434,900bp (8.7 % del genoma) en *Ahy* para su posterior revisión. En este caso, la distribución de valores para el estimador combinado (Figura 2.7d) mostró diferencias claras para los tres tipos de parches (inserciones,

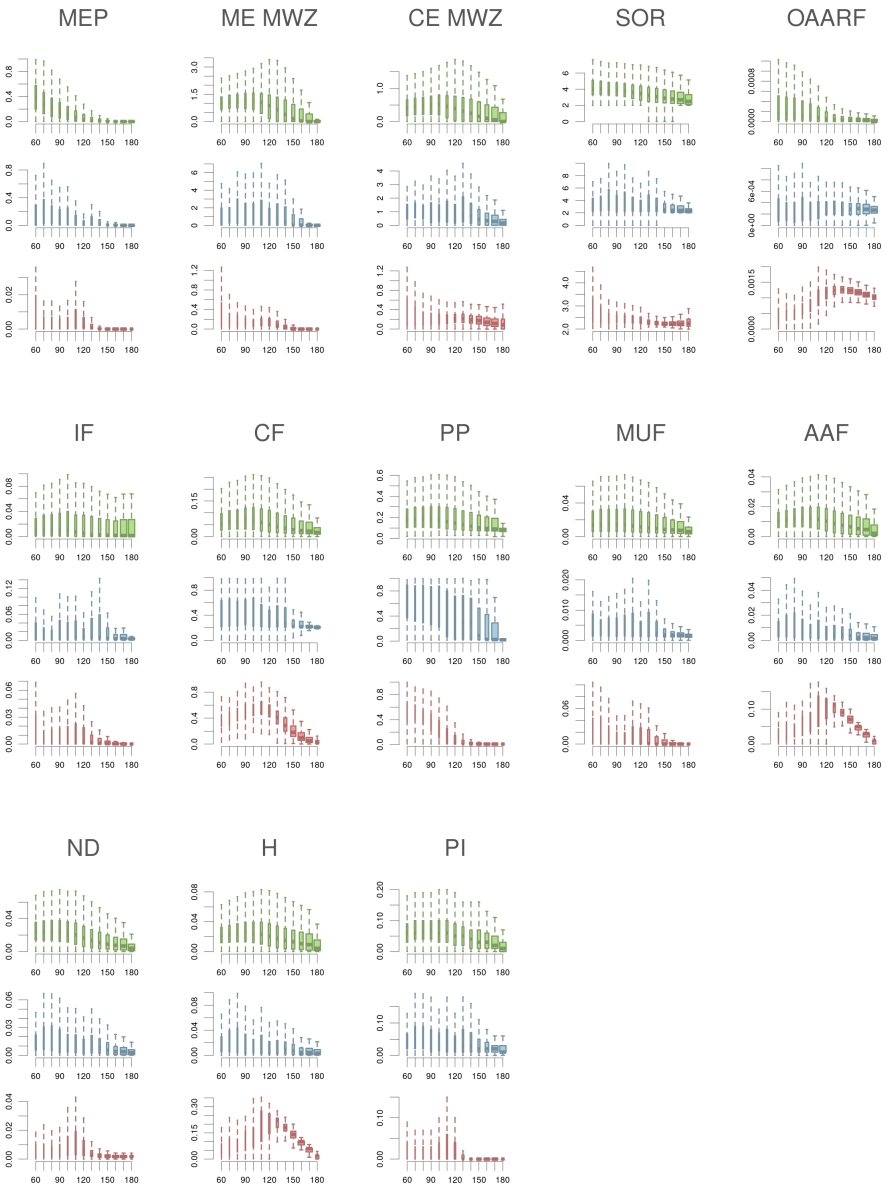


Figura 2.6: *Distribución de los indicadores de ruido individuales en los tres organimos analizados. De arriba a abajo Ahy, Sce y Ath.*

deleciones y modificaciones), confirmado por casi todos los indicadores individuales.

También los valores estimados por *REAPR* mostraron una buena coincidencia con los valores de similitud (Figuras 2.7e-g), especialmente para aquellas regiones con grandes errores de ensamblaje. Sin embargo, *REAPR* mostró menor sensibilidad que el estimador combinado para aquellas regiones con errores parciales de ensamblaje, diferencias que también se observaron para las regiones parcheadas (Figura 2.7h).

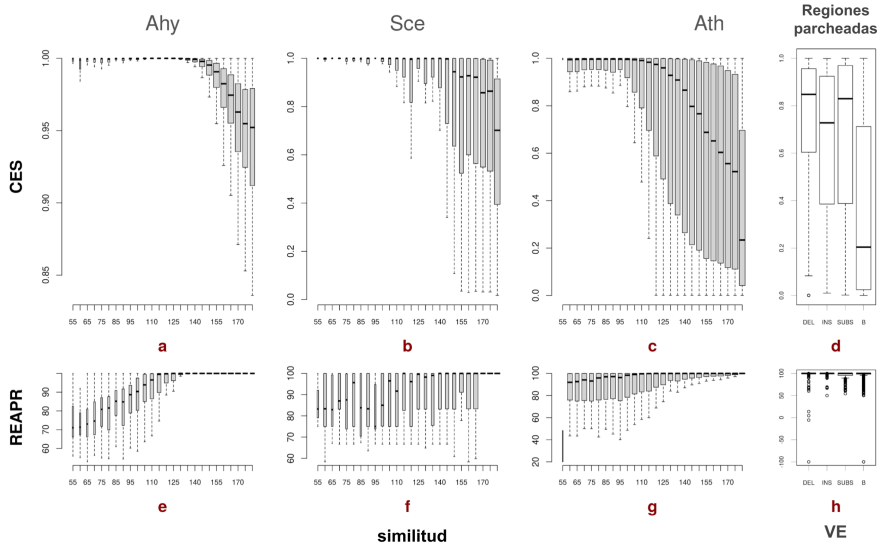


Figura 2.7: *Distribución del estimador combinado en función de diferentes escenarios. a, b y c) distribución del estimador combinado en función del estadístico de similitud obtenido al comparar las diferentes regiones de los genomas de Ahy, Sce y Ath, respectivamente. d) distribución del estimador combinado para diferentes categorías de regiones parcheadas en el genoma de Ath. e, f y g) distribución del estadístico de REAPR en función del estadístico de similitud. h) distribución del estadístico de REAPR para diferentes categorías de regiones parcheadas en el genoma de Ath.*

2.4.2. Predicción de mutaciones somáticas

Resultados con simulaciones

La herramienta de simulación de tumores se empleó para generar 10 simulaciones distintas, representando a 10 muestras tumorales reales generadas con parámetros similares. En cada simulación se emplearon 100 iteraciones y una población de 200 células, con un único progenitor tumoral. Se empleó una tasa de mutación germinal $d = 10^{-3}$ y una tasa de mutación somática asociada al tejido normal de $d_s = 10^{-6}$. En cada iteración se definió una probabilidad de mutación somática de $d_s^T = 0{,}75$ con un máximo de 3 mutaciones por célula. El *fitness* asociado a las mutaciones germinales fue igual a cero. En cambio, para las mutaciones somáticas tumorales, se empleó un *fitness* distribuido como $N(\mu = 0, \sigma = 2)$. Asimismo, para cada célula se empleó una desviación del *fitness* distribuida como $N(\mu = 0, \sigma = 0{,}25)$.

El conjunto total de las simulaciones mostró resultados parecidos. Al final de las 100 iteraciones, se obtuvo una media de 1233.8 mutaciones somáticas ($\sigma = 136{,}67$), de las cuales 859.4 ($\sigma = 156{,}01$) tuvieron una frecuencia alélica igual o superior a 0.01, lo que representaría en un caso ideal al menos 1 lectura mutada en una simulación de la muestra a 100× de cobertura.

El conjunto de frecuencias alélicas asociadas a las mutaciones somáticas mostró una distribución principalmente centrada en valores bajos (Figura 2.8a). Este patrón se corresponde con un gran número de mutaciones somáticas presentes en un porcentaje muy reducido de las células de la población. El resultado encaja en la distribución del número de mutaciones aparecidas y mantenidas en cada unidad de tiempo (Figura 2.8b). En este caso, se observa como la mayor parte de mutaciones mantenidas al final de la simulación surgen a tiempos muy próximos al final, lo que podría describir la expansión del último clon dominante. Se observó también como la población tumoral colonizó rápidamente el tejido (Figura 2.8c), consiguiendo un fracción total de células tumorales alrededor de la iteración 10. Asimismo, los valores de *fitness* muestran una distribución bastante simétrica (Figura 2.8d),

con una media centrada en 0 y una desviación estándar de 1.88, lo que describe mutaciones con potencial para mejorar las capacidades de las células, pero también con mutaciones que suponen una clara merma en su capacidad de adaptación.

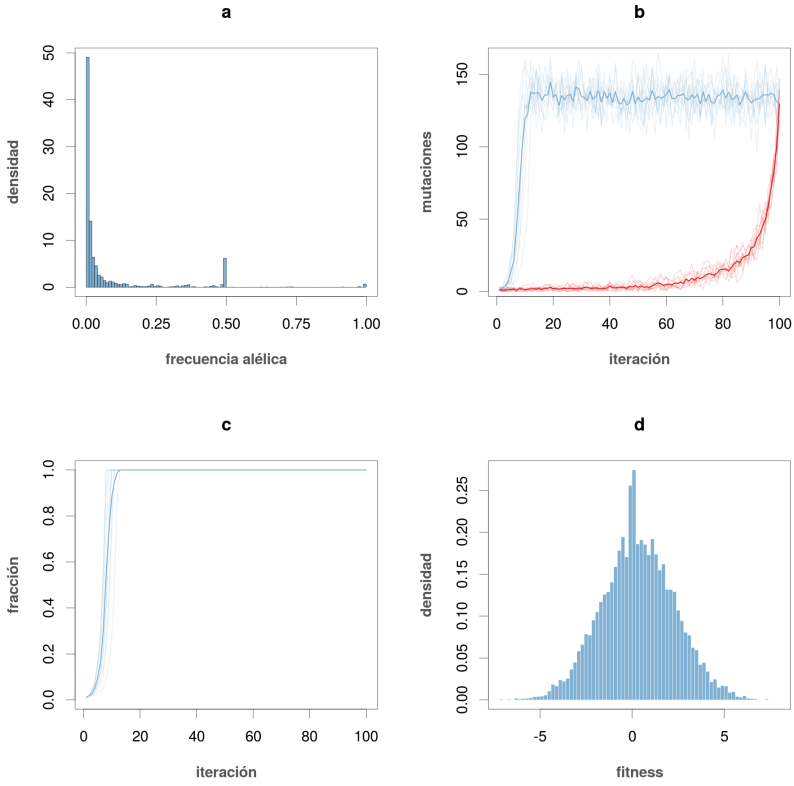


Figura 2.8: *Descripción de los parámetros obtenidos por las simulaciones; a) frecuencia alélica observada en la población de células al finalizar la simulación; b) número de mutaciones presentes en la población en cada iteración y la proporción de mutaciones conservada al final de la simulación (rojo); c) proporción de células tumorales en cada iteración; d) distribución del valor de fitness obtenido al final de la simulación.*

De cada serie simulada se generaron las lecturas correspondientes (en formato *fastq*) para las parejas normal-tumor a lo largo de los 10 niveles de ruido planteados y los 10 niveles de contaminación normal, haciendo un total de 20 parejas normal-tumor para cada serie y 400 muestras simuladas. El total de muestras simuladas se

empleó para evaluar y comparar la herramienta propuesta (SOM-hi) frente a dos de las herramientas más fiables en el campo (*MuTect2* y *Strelka2*).

En primer lugar se evaluó la sensibilidad y especificidad de las herramientas bajo diferentes niveles de ruido. En este caso, los resultados mostraron en todos los casos una disminución significativa en el rendimiento de las herramientas. En concreto, se aprecia un descenso claro en el número de mutaciones reales encontradas (verdaderos positivos) para todas las herramientas (Figura 2.9). Este descenso fue especialmente claro para *Strelka2*, donde a niveles de ruido superiores a 2 el descenso fue muy acusado, perdiendo prácticamente la totalidad de mutaciones para niveles superiores a 6. En el caso de *MuTect2* y *SOM-hi*, ambas herramientas mostraron una gestión mucho más eficiente de niveles altos de ruido. En concreto, ambas herramientas encontraron un porcentaje amplio de las mutaciones detectadas para niveles bajos de ruido, siendo *SOM-hi* ligeramente superior en casi todos los niveles de ruido evaluados.

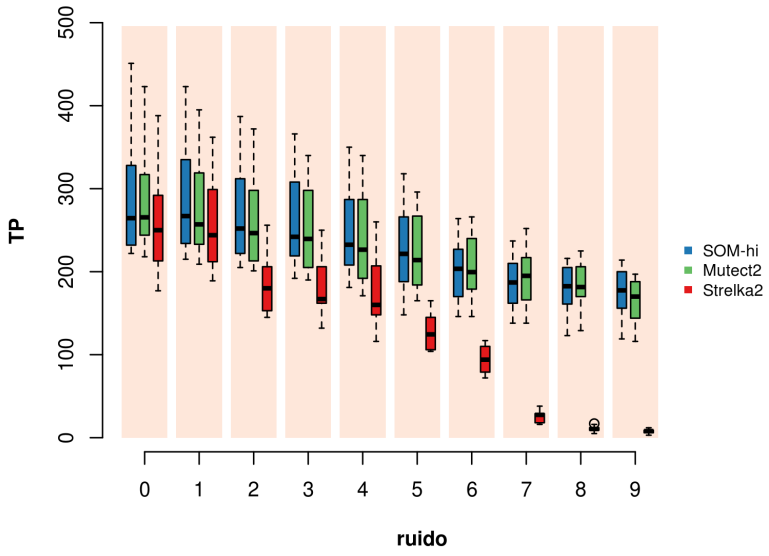


Figura 2.9: *Número de verdaderos positivos (TP) para diferentes niveles de ruido.*

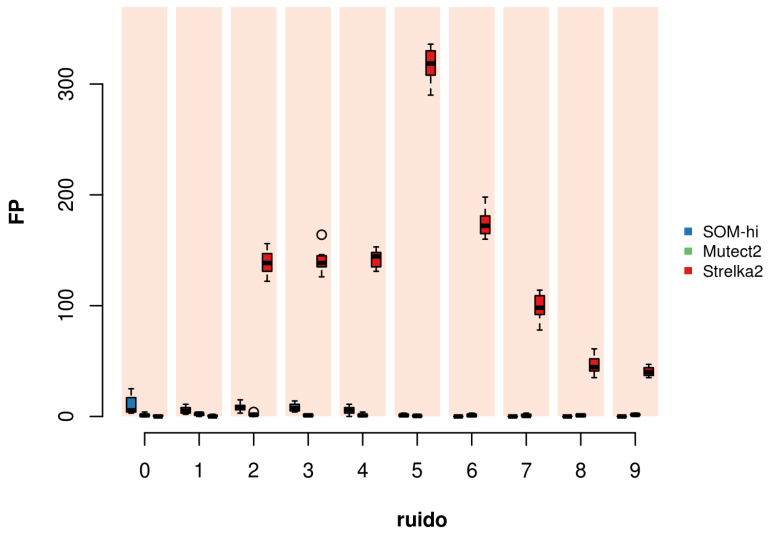


Figura 2.10: *Número de falsos positivos (FP) para diferentes niveles de ruido.*

También en el caso de falsos positivos (Figura 2.10) tanto *SOM-hi* como *Mutect2* mantuvieron un número muy reducido, siendo *SOM-hi* ligeramente peor. *Strelka2* por su parte mostró un gran número de falsos positivos para la mayoría de los niveles, mostrando un patrón más errático y no tan progresivo como en el caso de los verdaderos positivos. Ambos resultados se confirman al observar el área bajo la curva (AUC), medida que permite evaluar conjuntamente verdaderos y falsos positivos. En este caso *SOM-hi* mostró un comportamiento muy superior a *Strelka2* y ligeramente superior a *Mutect2* en casi todos los casos (Figura 2.11).

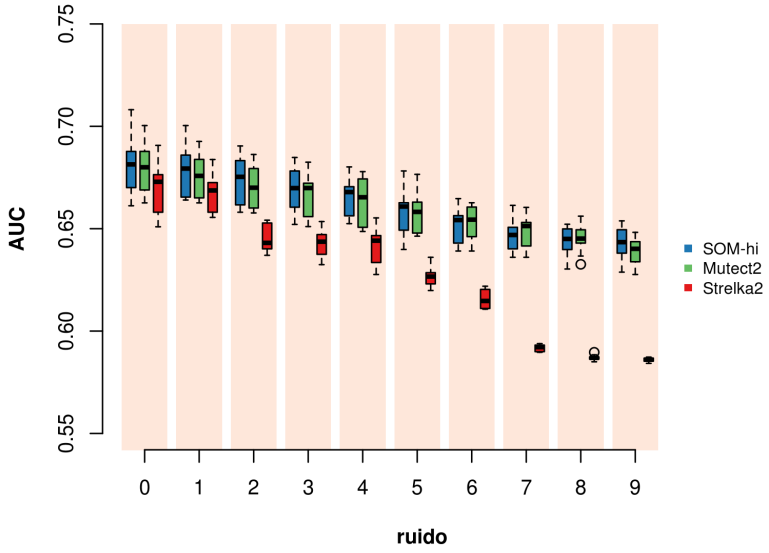


Figura 2.11: *Área bajo la curva obtenida para cada herramienta en los distintos niveles de ruido evaluados.*

Por su parte, el incremento de los niveles de contaminación normal también produjo una merma significativa en el rendimiento de las herramientas (Figura 2.12), aunque con diferencias menores entre ellas. En concreto, tanto *SOM-hi* como *Mutect2* recuperaron un número de mutaciones reales superior al de *Strelka2*. En el caso de falsos positivos (Figura 2.13), todas las herramientas obtuvieron un número

muy reducido, siendo *SOM-hi* ligeramente peor que el resto. A pesar de esto, en el cómputo global, *SOM-hi* mostró un comportamiento superior en el AUC (Figura 2.14).

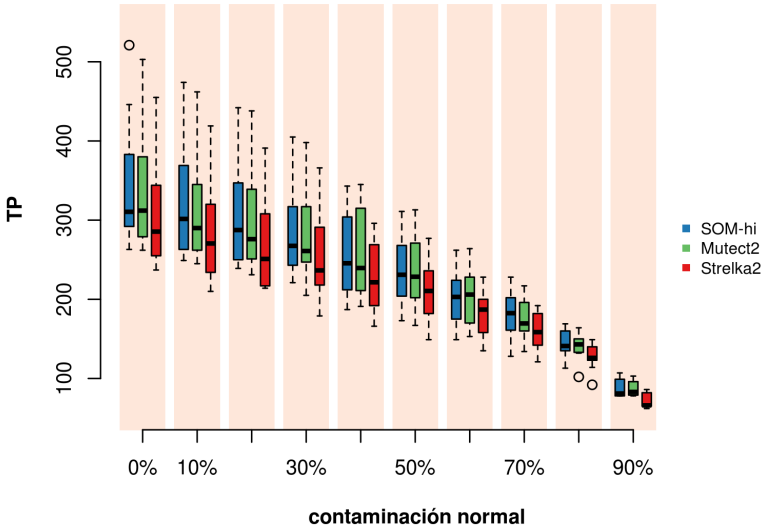


Figura 2.12: *Número de verdaderos positivos (TP) para diferentes niveles de contaminación normal.*

Además, se evaluó el rendimiento de la herramienta propuesta en función del tamaño muestral elegido durante el entrenamiento. La Figura 2.15 muestra la evolución del número de falsos y verdaderos positivos a lo largo del rango de tamaños muestrales evaluados. Tal y como se aprecia, se obtuvo un número de verdaderos positivos razonable incluso para tamaños muestrales muy reducidos, lo que describe un entrenamiento aceptable incluso para porciones muy pequeñas de la muestra normal. Sin embargo, de forma complementaria, se obtuvo un número de falsos positivos muy elevado para los tamaños muestrales pequeños, reduciéndose drásticamente a partir de tamaños superiores a 1Kbp. Por otro lado, se obtuvo una reducción inesperada del número de verdaderos positivos para tamaños muestrales

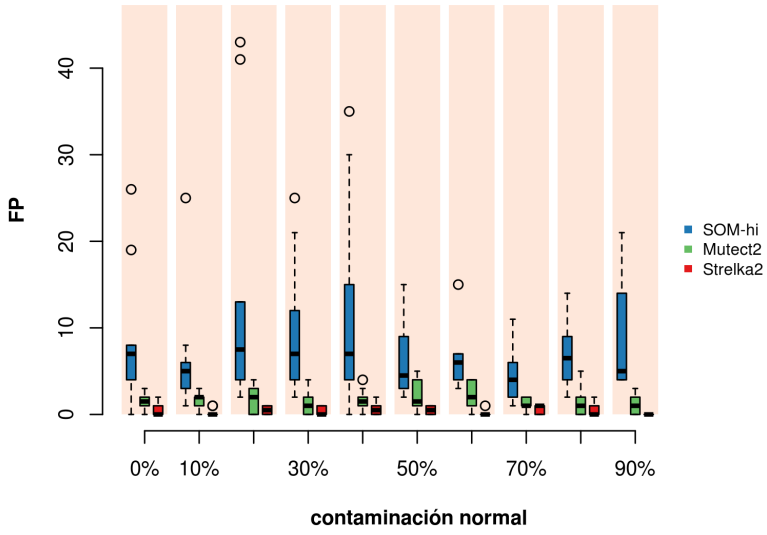


Figura 2.13: *Número de falsos positivos (FP) para diferentes niveles de contaminación normal.*

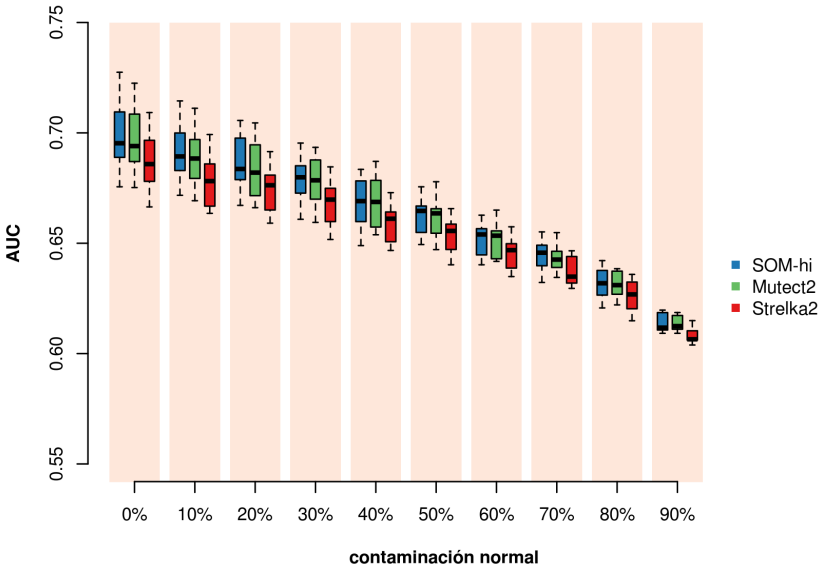


Figura 2.14: *Área bajo la curva obtenida para cada herramienta en los distintos niveles de contaminación normal evaluados.*

grandes ($> 50Kbp$), indicando probablemente un sobreentrenamiento producido por la inclusión de variantes germinales reales en los sets de entrenamiento.

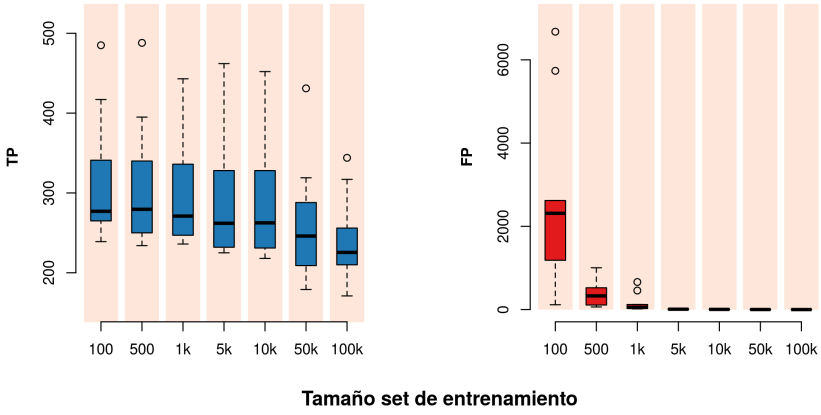


Figura 2.15: *Número de verdaderos positivos (TP) y falsos positivos (FP) obtenido con la herramienta SOM-hi para diferentes tamaño muestrales.*

Resultados en muestras reales

La metodología propuesta (*SOM-hi*) y las dos herramientas bajo comparación (*Strelka* y *Mutect2*) fueron aplicadas a tres experimentos procedentes del proyecto *SEQC*. Los resultados obtenidos se muestran en la Tabla 2.3. En este caso, el número de mutaciones somáticas a detectar fue de 78,375bp, con un número de posiciones sin mutación de 146,015,391bp.

Tal y como se aprecia, los resultados obtenidos varían de forma considerable entre las réplicas correspondientes a los 3 aparatos de secuenciación empleados. Mientras que para la réplica 3 tanto *Mutect2* como *Strelka* proporcionan un número de FP y VP más adecuado que *SOM-hi*, para las réplicas 1 y 2, *SOM-hi* proporciona un número de FP mucho más acotado que el de las otras dos herramientas, donde el porcentaje de FP en *Mutect2* y *Strelka* es muy superior al proporcionado por *SOM-hi*.

	secuenciador	VP	FP	VN	FN	Precisión	Exhaustividad
Strelka	HiSeq 1500	1777	15442	145999949	76598	0.1032000	0.0226730
	HiSeq 4000	1852	28337	145987054	76523	0.0613468	0.0236300
	HiSeq 2500	2089	4019	146011372	76286	0.3420105	0.0266539
Mutect2	HiSeq 1500	1339	11743	146003648	77036	0.1023544	0.0170845
	HiSeq 4000	1327	22189	145993202	77048	0.0564297	0.0169314
	HiSeq 2500	1833	4678	146010713	76542	0.2815236	0.0233876
SOM-hi	HiSeq 1500	1625	3915	146011476	76750	0.2933213	0.0207337
	HiSeq 4000	1599	8560	146006831	76776	0.1573974	0.0204019
	HiSeq 2500	1795	5881	146009510	76580	0.2338458	0.0229027

Tabla 2.3: Resultados obtenidos mediante la herramientas *Strelka*, *Mutect2* y *SOM-hi* para las muestras del proyecto *SEQC*. Se describe el número de verdaderos positivos (VP), falsos positivos (FP), verdaderos negativos (VN), falsos negativos (FN), así como los indicadores de precisión ($\frac{VP}{VP+FP}$) y exhaustividad ($\frac{VP}{VP+FN}$).

2.5. Conclusiones

La variabilidad intracelular constituye una de las características más comunes dentro de cualquier tumor. Se trata de una importante área de investigación que cada vez está adquiriendo mayor relevancia en el estudio de la enfermedad, no solo a la hora de comprender las características moleculares más generales del cáncer, sino también a la hora de comprender como la heterogeneidad celular condiciona la efectividad de las estrategias terapéuticas y de cómo la composición clonal y la interacción entre sus distintas subpoblaciones influye en el desarrollo y expansión del tumor.

En este capítulo de la tesis, se han abordado los efectos de la heterogeneidad celular a la hora de realizar un genotipado somático robusto, el cual constituye en muchos casos el punto de partida para un análisis bioinformático más global que permita evaluar la integridad de los genes y su funcionamiento en el contexto de las rutas moleculares. Para poder desarrollar esta tarea, ha sido necesario realizar un modelado explícito de las distintas fuentes de variabilidad contenidas en los datos. En concreto, se ha hecho uso de un conjunto de indicadores de error obtenidos a partir de mapeos convencionales que han permitido predecir y estimar con gran

fiabilidad el grado de error en cada región genómica evaluada.

En la primera parte de este capítulo se ha planteado un modelo estadístico que integra los distintos indicadores de ruido con el fin de detectar regiones genómicas incorrectamente ensambladas. A lo largo de diferentes experimentos realizados en varios organismos, se ha podido demostrar la capacidad predictiva de los indicadores a la hora de detectar regiones genómicas que posteriormente han estado sujetas a correcciones en versiones genómica más nuevas o de mayor calidad. En concreto, los indicadores han detectado regiones parcheadas y posiciones susceptibles a errores de genotipado en el genoma humano, diferencias claras entre las últimas versiones del genoma de *Arabidopsis Thaliana* y regiones genómicas incorrectamente construidas en los ensamblajes obtenidos *ad-hoc* en *Saccharomyces cerevisiae* y *Aeromonas hydrophilia*. Estos resultados demuestran la idoneidad del enfoque y sugieren un uso similar en diferentes ámbitos como el genotipado somático abordado posteriormente.

En este apartado se ha planteado también la definición de un estimador combinado construido a partir de los p-valores empíricos obtenidos por cada indicador de ruido individual. Este abordaje ha permitido disponer de la sensibilidad necesaria para detectar regiones incorrectamente ensambladas en el genoma humano, el cual constituye probablemente el mejor ensamblaje disponible entre los organismos modelo. Además, el estimador combinado ha sido capaz de detectar regiones con errores de genotipado mediante un modelo entrenado con un conjunto de muestras diferente al evaluado, lo que permite plantear la construcción de modelos de error de uso genérico a partir de poblaciones amplias de datos. Por último, la herramienta propuesta obtuvo mejores resultados que *REAPR*, la cual constituye la herramienta *gold standard* en el ámbito de la evaluación de ensamblajes.

Una vez validada la capacidad predictiva de los indicadores de ruido, se ha empleado una aproximación similar en el diseño de un modelo de genotipado somático robusto. En concreto, se ha construido un modelo que permite estimar el número de cambios artefactuales esperado en cada posición genómica debido a las diferentes fuentes de ruido latentes. El modelo ha permitido analizar de forma robusta posiciones genómicas complejas con mutaciones soportadas por muy

pocas lecturas. En este caso, la estimación proporcionada por el modelo permite determinar qué fracción de los cambios observados es de naturaleza artefactual, lo que en la práctica se traduce en una tasa de falsos positivos muy baja.

La herramienta de genotipado somático propuesta (*SOM-hi*) ha demostrado ser competitiva frente al resto de herramientas bien establecidas en el campo (*Strelka* y *Mutect*), con una extensa evaluación en diferentes niveles de ruido y distintos grados de contaminación normal. Uno de los puntos interesantes de la herramienta lo constituye el aprendizaje específico de la estructura del ruido presente en cada muestra. Este enfoque permite conseguir el máximo rendimiento ya que cada observación, en función del protocolo de laboratorio y del tipo de secuenciación empleado, dispondrá de unos artefactos de ruido específicos. Esta aproximación ha permitido conseguir resultados aceptables incluso cuando los niveles de ruido han estado situados en rangos muy por encima de lo normal. En este contexto, ha sido importante la selección de posiciones genómicas en la muestra normal para entrenar el modelo de ruido. Para ello, se han evaluado diferentes tamaños poblacionales para determinar el tamaño óptimo de entrenamiento en el modelo. En los experimentos realizados se demuestra como un conjunto de solo 5000 posiciones genómicas resulta suficiente para obtener un modelo de ruido fiable, lo que constituye un porcentaje muy reducido con respecto al tamaño genómico total, incluso cuando se realiza un genotipado solo de regiones codificantes.

También en rangos de contaminación normal altos, el protocolo propuesto ha funcionado con gran solvencia, demostrando así su sensibilidad en la detección de mutaciones somáticas de baja frecuencia. Esta característica resulta de gran relevancia ya que, tal y como se ha visto en las simulaciones realizadas, la mayor parte de mutaciones somáticas presentes en una muestra tumoral se observan en una frecuencia poblacional muy por debajo de lo que una secuenciación estándar podría capturar.

En este contexto, se ha demostrado la eficacia y flexibilidad obtenida al emplear un clasificador jerárquico como *Random Forest*. A pesar de renunciar a comprender la relación entre las variables predictivas (indicadores de ruido) y la variable

respuesta (número de cambios artefactuales), tal y como podría obtenerse mediante el uso de modelos clásicos como los *GLM*, el clasificador ha demostrado ser de gran utilidad a la hora de modelar indicadores de ruido de diferentes naturalezas. Este punto abre la posibilidad futura del uso de clasificadores basados en *Deep Learning* que permitan mejorar el rendimiento del genotipado. También, el enfoque planteado sugiere una mejora futura del modelo de error solo con la ampliación paulatina de nuevos indicadores de ruido, ya que estos permiten observar y cuantificar desde diferentes puntos de vista las distintas fuentes de error latentes.

Por último, una de las aportaciones más interesantes de este capítulo ha sido el desarrollo de una herramienta de simulación de linajes tumorales, y de su adaptación para generar muestras de secuenciación que incorporen patrones de error realistas. En este contexto, es de gran importancia el desarrollo de herramientas de simulación que permiten explorar diferentes escenarios evolutivos, con el fin de avanzar en la comprensión de las dinámicas de interacción entre los diferentes clones con el tejido normal adyacente. Bajo esta línea, en este trabajo se ha desarrollado una herramienta de simulación que ha permitido explorar la estructura interna de un linaje tumoral a través de diferentes series, mostrando puntos relevantes a la hora de entender como se distribuye la frecuencia alélica de las variantes somáticas en función de la población celular resultante después del tiempo de simulación.

Como trabajo futuro queda la simulación de linajes con diferentes progenitores y con poblaciones celulares más grandes y estrategias que permitan simular la coexistencia de diferentes clones, no solo mediante en un escenario de competición, sino mediante la implementación de co-dependencias entre clones. En este sentido, resultará especialmente interesante desarrollar un modelo espacial que permita situar a las células en un entorno tridimensional donde se pueda estimar de forma más precisa la presión selectiva en cada célula en función de la proximidad a recursos esenciales, o dependiendo de la cercanía de otras células del tejido.

CAPÍTULO 3

Modelado de la variabilidad genómica en las rutas moleculares

En este capítulo de la tesis se abordará el estudio de la heterogeneidad genómica entre pacientes con el mismo tipo de cáncer. Para ello, se propodrá un modelo jerárquico de factorización que permite la obtención simultánea de un conjunto de componentes latentes a nivel de gen y sus correspondientes componentes a nivel de ruta molecular. Finalmente, se describirán los análisis aplicados al resultado proporcionado por el modelo jerárquico, con el fin de determinar la relevancia de las componentes encontradas en el contexto de la enfermedad.

3.1. Contexto biológico de las rutas moleculares

Las rutas moleculares describen como se producen las distintas interacciones entre proteínas dentro de la célula para llevar a cabo las tareas esenciales durante todo su ciclo de vida. Dichas tareas representan programas transcripcionales implementados y seleccionados durante toda la evolución, siendo en algunos casos mecanismos muy antiguos presentes en los primeros organismos multicelulares (Herman y cols., 2018). Las rutas no solo definen procesos internos de mantenimiento y reparación, además, describen cómo es la respuesta celular a los distintos estímulos externos a los que las células deben atender para cumplir su función en el tejido (Figura 3.1).

El mecanismo de respuesta a un estímulo está condicionado a características particulares como el tipo celular. Además, las condiciones del medio extracelular en el que la célula está inmersa modificarán de forma significativa su comportamiento. En concreto, la matriz extracelular y el conjunto de moleculas que circulan por el medio representan estímulos valiosos que informan a la célula sobre las condiciones del tejido. Asimismo, el contacto mecánico directo con otras células estimulará determinadas respuestas, que en ocasiones, propagarán a células vecinas.

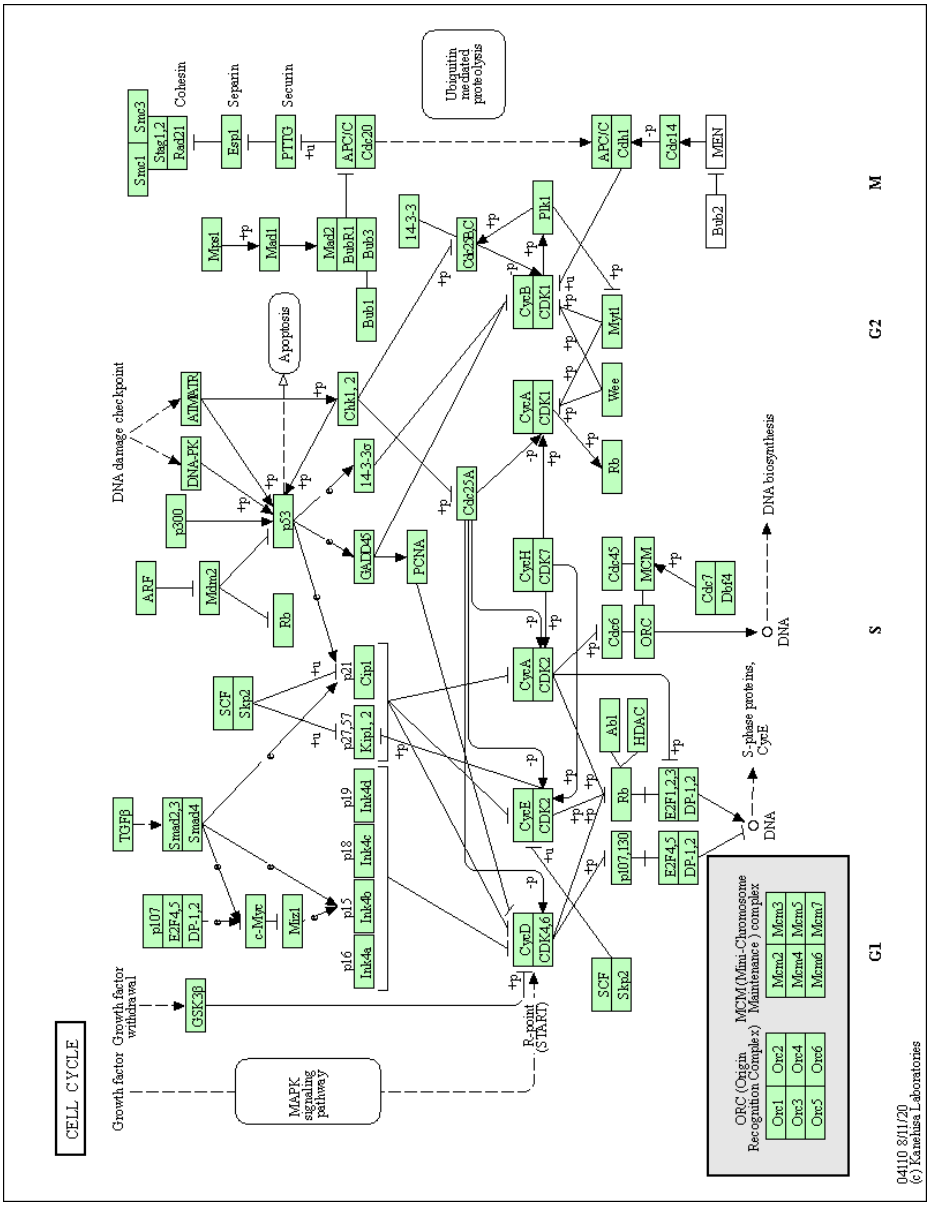


Figura 3.1: *Ruta molecular del ciclo celular recogida en el repositorio KEGG (https://www.genome.jp/pathway/hsa04110).*

3.1.1. Interacciones entre proteínas

En el contexto celular las proteínas representan los actores principales en el funcionamiento y ejecución de las rutas moleculares, por medio de sus interacciones y de su relación con otras moléculas más elementales del medio, como metabolitos u hormonas. A nivel estructural las proteínas están formadas por una secuencia de aminoácidos codificada a partir de un proceso conocido de forma clásica como el *dogma central de la biología* (Crick, 1970). Este proceso comprende (i) la transcripción de la región codificante de un determinado gen a su correspondiente ARN mensajero, (ii) la translocación de éste al citoplasma, y (iii) su posterior traducción a la secuencia de aminoácidos correspondiente por parte de los ribosomas. Como característica esencial, las proteínas adoptarán progresivamente durante su traducción una conformación tridimensional específica que determinará sus potenciales interacciones con otros elementos de la célula, y con ello, su función.

Las interacciones entre proteínas se pueden producir mediante diferentes mecanismos. Uno de los más habituales consiste en la formación de complejos proteicos formados por dos o más proteínas, donde la conformación tridimensional adquirida por el conjunto permite disponer de interacciones (o funciones) nuevas que no realizan sus proteínas integrantes por separado. Este tipo de agregación representa un tipo de interacción muy estable, frente a otras interacciones más transitorias.

Por otro lado, algunas proteínas con actividad enzimática son capaces de modificar la estructura tridimensional de otras proteínas mediante la incorporación o eliminación de grupos químicos en sitios específico de su estructura. Estas transformaciones se conocen como modificaciones postraduccionales (MPT) (Uversky, 2013), ya que describen específicamente cambios aparecidos en las proteínas después de su traducción por parte de los ribosomas. Las transformaciones incluyen (de)fosforilaciones (destinadas a activar o desactivar otras proteínas mediante la incorporación o eliminación de un grupo fosfato), (de)metilaciones (destinadas a regular las proteínas que componen los nucleosomas), o ubiquitinaciones (destinadas a marcar una proteína para su degradación posterior mediante la incorporación

de una o más moleculas de ubiquitina), entre otras. El cambio a nivel bioquímico puede provocar la modificación de la estructura tridimensional, o simplemente cambiar sus propiedades termodinámicas, lo que dota a la proteína de la capacidad de interaccionar con otras proteínas nuevas, siendo a menudo un mecanismo que permite pasar de una forma inactiva a otra activa de la proteína. Las MPT son esenciales en la regulación de las rutas moleculares, ya que permiten modular de forma dinámica la función de las proteínas en la célula. Las MPT generalmente se producen en sitios específicos de las proteínas, mostrando a menudo más de un sitio distinto de modificación, lo que permite disponer de un abanico de funciones amplio para determinadas proteínas esenciales en la célula (Kastenhuber y Lowe, 2017).

En este contexto destacan las proteínas quinasas, las cuales se encargar de fosforilar a otras proteínas a través de la incorporación de un grupo fosfato, tomado de las moléculas de ATP (adenosín tri-fosfato) presentes en el citoplasma, y generando como subproducto ADP (adenosín di-fosfato). Esta transformación puede potencialmente activar o desactivar una proteína. De forma complementaria, hay que destacar a las proteínas fosfatasas, las cuales se encargar de realizar el proceso inverso, eliminando un grupo fosfato de su proteína diana.

Un contexto particular donde las MPT son de especial relevancia lo representan las rutas de señalización. Estas describen la respuesta celular a los distintos ligandos (como hormonas) presentes en el medio extracelular. Se trata de un tipo de respuesta muy rápida en la célula, en contraposición a la respuesta transcripcional que suele tardar horas. El proceso consiste en el disparo de una cascada de señalización formada por una sucesión de interacciones entre diferentes proteínas o complejos. La cascada comienza habitualmente en la membrana nuclear, atraviesa el citoplasma celular, y acaba activando una proteína efectora concreta, generalmente a través de una MPT específica. Después, la proteína efectora se translocará al núcleo de la célula, donde, con la ayuda de otras proteínas, se unirá de forma específica al ADN de la región promotora de una serie de genes a los cuales promoverá su expresión. Estos genes, a su vez, serán los encargados de dar respuesta a medio

plazo al estímulo que originó la cascada de señalización.

3.1.2. Representación matemática de las rutas moleculares

Los grafos constituyen el planteamiento matemático más empleado para representar de manera formal las interacciones entre proteínas. En concreto, un grafo G se define como:

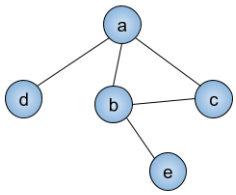
$$G = (V, I),$$

donde V representa a los k vértices e I al vector de interacciones, compuesto por n pares de vértices donde:

$$I = \{p_1,p_2, ..., p_n\}$$
$$p_i = x_i, y_i$$
$$i \in [1, n]$$
$$x_i, y_i \in V.$$

En el contexto de las rutas moleculares, los vértices generalmente representan a las proteínas, y las aristas, sus interacciones, pudiendo ocasionalmente ponderar o etiquetar las aristas en función del tipo de interacción observada entre las parejas de proteínas. De manera general, los grafos pueden contener interacciones no dirigidas (o sin dirección preferente), como las interacciones físicas entre parejas de proteínas dentro de un complejo proteico, o dirigidas (con dirección preferente), como las MPT en rutas de señalización, donde una proteína con actividad enzimática modifica a otra proteína que actúa como sustrato.

Además de la representación típica mediante una lista de vértices e interacciones, una de las formas más comunes de representar a un grafo es la creación de una matriz de adyacencia $k \times k$ (Figura 3.2). La matriz refleja si existe una interacción entre la proteína situada en la fila i y la proteína situada en la columna j. De forma característica, la matriz de adyacencia es simétrica cuando las interacciones son no dirigidas, y no simétrica, cuando el grafo es dirigido.



	a	b	c	d	e
a	0	1	1	1	0
b	1	0	1	0	1
c	1	1	0	0	0
d	1	0	0	0	0
e	0	1	0	0	0

Figura 3.2: *Grafo no dirigido compuesto por 5 nodos, junto a su matriz de adyacencia correspondiente.*

Los grafos han sido ampliamente utilizados en biología computacional para representar relaciones de distinta naturaleza entre proteínas o genes. Las interacciones pueden representar relaciones directas, como las interacciones físicas ya mencionadas, o también representar relaciones más complejas, como patrones de coexpresión, o asociaciones indirectas, como la pertenencia al grupo de genes que producen una misma enfermedad. Las redes biológicas generalmente representan un solo tipo de interacción, pero ocasionalmente pueden también mostrar interacciones de diferente naturaleza, así como integrar diferentes tipos de nodos, formando redes más heterogéneas. Esta aproximación permite disponer de esquemas más flexibles, con capacidad para representar de forma conjunta diferentes tipos o niveles de información, lo que facilita la identificación de relaciones particulares entre genes que difícilmente podrían ser detectadas a partir de redes homogéneas independientes. Por supuesto, estas aproximaciones requerirán también de enfoques estadísticos específicos, que tengan en cuenta tanto la naturaleza del nodo, como de la interacción a analizar.

La aplicación de la teoría de grafos en el estudio de las redes biológicas (Koutrouli, Karatzas, Paez-Espino, y Pavlopoulos, 2020) permite extraer y cuantificar de forma sencilla parámetros numéricos que caracterizan el rol o la importancia de cada proteína dentro de la red, pudiendo en consecuencia, abordar su análisis posterior mediante métodos de inferencia estadística de diferente índole. Dentro del conjunto de parámetros que se pueden medir para cada nodo de la red destaca el número de

interacciones, como medida general de su relevancia en la red. En este contexto, una de las observaciones más típicas en las redes biológicas es que suelen mostrar un perfil conocido como *scale free*, donde la distribución del número de interacciones por nodo sigue típicamente una curva exponencial. Este patrón refleja que la mayor parte de nodos muestran un número muy reducido de interacciones, siendo un grupo pequeño de nodos los que acumulan la mayor parte de la estructura de la red. En el caso real, estos nodos representan a las proteínas más esenciales dentro de la célula, siendo recurrentemente mutadas en enfermedades como el cáncer.

Además, existen otros parámetros que ayudan a caracterizar el peso de un nodo en la red. Se trata de parámetros que miden la centralidad del nodo. En concreto, es posible calcular la centralidad de cercanía (longitud media entre el nodo de interés y cualquier otro nodo en la red), la centralidad de intermediación (el número de caminos entre cualquier par de nodos que atraviesa el nodo de interés), así como otras medidas más sofisticadas como la centralidad de valor propio (medida a partir de la contribución del nodo sobre el primer vector propio de la matriz de adyacencia).

Asimismo, también es importante señalar que es posible cuantificar parámetros que caracterizan a una red biológica en sentido global. Algunos de estos parámetros consisten en promedios obtenidos a partir de los parámetros calculados a nivel de nodo, como la anchura del grafo (longitud media del camino entre cada par de nodos), o la centralidad media, pero también otros como el número de subredes o comunidades, que solo tienen sentido a nivel de red. Además, también es posible caracterizar partes específicas de la red a través de parámetros como el coeficiente de agrupamiento, que mide el nivel de interconexión entre un conjunto de nodos y su similitud con un grafo con conexión total. Estas aproximaciones resultan interesantes, ya que permiten identificar subredes latentes, o vecindarios con características relevantes para explicar una enfermedad.

3.1.3. Visión jerárquica de la heterogeneidad genómica en el cáncer

Tal y como se ha descrito en el capítulo anterior, existe un alto grado de heterogeneidad genómica observada, tanto a nivel intratumoral, como entre pacientes con el mismo tipo de cáncer. La enorme variedad de mecanismos surgidos y seleccionados en el seno del tumor establece el primer nivel de heterogeneidad, donde aquellos genes más relevantes para su desarrollo podrían mostrar un abanico de diferentes alteraciones genómicas, con un efecto parecido sobre su función.

De forma complementaria, también el grado de heterogeneidad observado a nivel de gen resulta muy elevado, siendo en ocasiones difícil encontrar genes recurrentemente mutados en un porcentaje significativo de pacientes con el mismo tipo de cáncer. Esta característica fundamental sugiere necesariamente que la alteración selectiva de conjuntos diferentes de genes puede conducir a fenotipos celulares con características muy similares, algo también habitual en otras enfermedades multigénicas además del cáncer. En este caso, la estructura natural de las rutas moleculares explica el segundo nivel de heterogeneidad, ya que, la alteración de puntos distintos de la misma red tiene potencialmente la capacidad de inhibir o sobreactivar de forma significativa su funcionamiento. Ejemplos característicos son: la alteración de diferentes proteínas pertenecientes al mismo complejo (lo que conduce a la eliminación del complejo y con ello su función en la célula), el bloqueo de diferentes proteínas esenciales dentro de la ruta (lo que lleva al bloqueo de su actividad), o la alteración de diferentes proteínas inhibidoras dentro de la red (lo que lleva a la sobreactivación de la ruta).

Por último, también a un nivel jerárquico más alto es posible explicar parte de la hetegeneidad genómica observada. Tal y como se ha descrito, los *hallmarks* del cáncer describen de forma general aquellas características comunes a cualquier tumor. A nivel biológico, los *hallmarks* se descomponen a su vez en multitud de rutas moleculares distintas, implicadas en procesos esenciales como el metabolismo, la apoptosis o la proliferación celular. Este punto establece el tercer nivel de

heterogeneidad, donde la alteración combinada de diferentes rutas moleculares conducirá a la activación de un mismo *hallmark*.

De esta forma, desde la perspectiva de la biología de sistemas, se define una visión jerárquica de la heterogeneidad genómica en el cáncer, desde las mutaciones somáticas, hasta las rutas moleculares que regulan en los *hallmarks* del cáncer.

3.2. Objetivos del capítulo

En este capítulo de la tesis se ofrece una visión más sistémica de las alteraciones observadas en los pacientes, y de cómo éstas afectan a las rutas de señalización. El objetivo principal en este caso consiste en el desarrollo y aplicación de un modelo de factorización que permita la detección de patrones latentes, que posteriormente se relacionarán con las distintas estrategias implementadas en los tumores para desarrollar los *hallmarks* del cáncer.

Para llevar a cabo esta tarea emplearemos métodos de estimación de componentes latentes. En concreto, se describe la implementación de un modelo jerárquico de factorización basado en la técnica *Non-negative Matrix Factorization* (*NMF*). El modelo planteado nos permitirá factorizar de forma simultánea las matrices que describen la actividad de los tumores, tanto a nivel de gen, como a nivel de ruta molecular, teniendo como principal requisito que las componentes latentes encontradas en ambos niveles sean compatibles entre sí.

A nivel práctico, el modelo jerárquico no describe una descomposición global de las matrices de actividad. De hecho, se aplica de forma sucesiva a un conjunto de matrices que describen la actividad de los genes y las rutas implicadas en aquellas funciones biológicas alteradas en los pacientes con cáncer. Este diseño estratificado no solo permite que las factorizaciones sean más abordables, sino que además proporciona una descripción más comprensible y detallada sobre la estructura interna de cada función biológica.

Por último, una vez obtenidas las componentes latentes, éstas se analizan para determinar si se relacionan con parámetros clínicos como el subtipo o la supervivencia de los individuos. Asimismo, los elementos más relevantes dentro de cada componente encontrada son analizados con el objetivo de determinar su carácter oncogénico y su potencial uso como biomarcador en la enfermedad.

3.3. Descripción de la metodología de análisis

El protocolo de análisis que se propone comienza con la estimación de la actividad de los genes en el conjunto de muestras seleccionadas para el análisis. Para ello, se define la matriz $X_g \in \mathbb{R}^{m \times n}$, compuesta por m genes y n individuos, obtenida mediante la combinación de la matriz X_g^e, que describe el nivel de expresión de cada gen en cada individuo, y la matriz de afectación X_g^v, que describe el efecto de las mutaciones somáticas sobre la estructura de los genes. Después, dicha matriz se introduce en la herramienta *Hipathia*, obteniendo así la matriz X_p que cuantifica la actividad de cada ruta de señalización en los mismos individuos.

Las matrices X_g y X_p describen la actividad de los tumores desde dos niveles de abstracción distintos, y constituyen las matrices de entrada para el método de extracción de componentes latentes. A continuación, se describen todos los apartados necesarios para la generación del modelo jerárquico y su aplicación en un análisis real.

3.3.1. Integración de mutaciones somáticas y expresión

El conjunto de mutaciones somáticas presentes en el genoma de un individuo tiene un efecto directo sobre el patrón de actividad de sus genes. Aunque en la mayoría de casos las mutaciones tienen un efecto débil sobre el fenotipo celular, un conjunto específico de mutaciones somáticas será mantenido gracias la presión selectiva, produciendo cambios importantes en el comportamiento del tumor. Dichas mutaciones somáticas tienen mayoritariamente un efecto inhibidor, alterando o

truncando la estructura de las proteínas afectadas. Sin embargo, en algunos casos, las mutaciones se producen en posiciones específicas, dotando a las proteínas mutadas de una ganancia de función específica que resultará relevante a la hora de orquestar los procesos necesarios para llevar a cabo los *hallmarks* del cáncer.

La extracción de mutaciones somáticas se realiza habitualmente mediante una herramienta de predicción similar a la descrita en el capítulo anterior. Su aplicación proporciona una lista de mutaciones somáticas observadas en el genoma de cada muestra, acompañadas de información contextual de gran importancia, como su posición genómica o el tipo de cambio observado. Esta información permite determinar qué genes han sido potencialmente afectados, pudiendo en este punto, si se requiere, realizar un análisis adicional destinado a determinar la presencia de mutaciones o genes alterados de forma recurrente en un conjunto de individuos con el mismo tipo de cáncer.

En este protocolo, las mutaciones somáticas encontradas se almacenan en formato *VCF* (Danecek y cols., 2011), que constituye el estándar en el área. A continuación, se determina el efecto de las mutaciones sobre los genes afectados, determinando en cada caso si el cambio de nucleótido observado a nivel genómico proporciona a su vez un cambio importante sobre la secuencia de aminoácidos de la proteína correspondiente. Para ello, se hace uso de la herramienta *SnpEff* (Cingolani y cols., 2012), la cual toma como entrada el fichero *VCF* y genera una versión anotada del mismo, añadiendo para cada mutación el tipo de efecto causado en cada una de las isoformas de los genes potencialmente afectados. En particular, *SnpEff* resume el efecto de cada mutación en función de 4 categorías generales (*HIGH*, *MODERATE*, *LOW* y *MODIFIER*) que describen diferentes grado de afectación (Tabla 3.1).

Después de caracterizar el conjunto de mutaciones es necesario agregar dicha información a nivel de gen con el propósito de generar la matriz X_g^v que refleja el nivel de afectación de cada gen en cada individuo. En particular, para un gen i y un individuo j, se propone agregar las mutaciones encontradas de la siguiente

Efecto	Descripción	Peso (ω)
HIGH	Cambios estructurales importantes, como la presencia de un codón de parada prematuro, o un cambio en la pauta de lectura	0.99
MODERATE	Mutaciones que no producen un cambio global en la estructura de la proteína, pero sí producen cambio o pérdida de aminoácidos concretos	0.75
LOW	Mutaciones en la secuencia codificante que no producen un cambio de aminoácido en la proteína	0.1
MODIFIER	Mutaciones que no afectan a la región codificante del gen	0.05

Tabla 3.1: Categorías proporcionadas por la herramienta *SnpEff* para la caracterización del efecto producido por una mutación dada en un gen particular.

forma:

$$X^{v}_{g,i,j} = \left[1 - \sum_{v}^{V^{-}_{i,j}} \omega(\psi(v))\right] \left[1 + \sum_{v}^{V^{+}_{i,j}} \omega(\psi(v))\right], \tag{3.1}$$

donde $V^{-}_{i,j}$ y $V^{+}_{i,j}$ se corresponden con el conjunto de mutaciones que producen una pérdida y ganancia de función, respectivamente; ψ con el efecto más importante producido por la mutación v en cualquiera de las isoformas del gen; y ω a su peso asociado (Tabla 3.1).

Para determinar si una mutación concreta pertenece al grupo de mutaciones somáticas con ganancia de función se plantea un contraste de hipótesis, cuya hipótesis nula (H_0) asume que la mutación muestra un número de individuos compatible con lo observado para el resto de mutaciones en el mismo gen, indicando la hipotesis alternativa (H_A) que la mutación está en un número de individuos mayor de lo esperado. En la práctica, para considerar una ganancia de función el nivel de significancia obtenido al aplicar el contraste debe estar por debajo de 0.05, definiendo el p-valor como:

$$p = 1 - \Phi^{-1}(\eta), \tag{3.2}$$

donde η se corresponde con el número de individuos afectados por la mutación evaluada, y Φ^{-1} con la función de distribución (densidad) acumulada de la distribución que describe el número de individuos afectados en el conjunto total de mutaciones observadas en el mismo gen.

Por último, una vez agregadas las mutaciones, se procede a combinar la matriz resultante X_g^v con la matriz de expresión X_g^e, generando así la matriz definitiva de actividad génica:

$$X_g = X_g^e \odot X_g^v, \tag{3.3}$$

donde $\odot$ se corresponde con el producto escalar entre matrices.

3.3.2. Cuantificación de rutas moleculares de señalización

Las rutas moleculares describen como se produce la interacción entre las distintas proteínas del medio para llevar a cabo las funciones biológicas necesarias durante el ciclo de vida de las células. Dado que su actividad depende directamente de las proteínas, en las últimas dos décadas han surgido gran multitud de métodos estadísticos y computacionales destinados a inferir su actividad a partir de la expresión de los genes que las integran (Haynes y cols., 2013; Jacob y cols., 2012; Martini y cols., 2013; Tarca y cols., 2009).

Para cuantificar la actividad de las rutas moleculares empleamos en este caso la herramienta *Hipathia* (Hidalgo y cols., 2016), anteriormente mencionada. *Hipathia* permite el modelado y cuantificación de un conjunto de rutas de señalización descritas en la base de datos *KEGG* (Kanehisa y Goto, 2000). Estas rutas cubren la mayor parte de funciones esenciales en la célula. Además, se trata de funciones que, por su relevancia, aparecen recurrentemente mutadas en el cáncer, por lo que su modelado estadístico resulta de gran utilidad a la hora explorar y entender a un nivel más sistémico los cambios producidos en un tumor.

Hipathia modela las distintas rutas de señalización mediante el uso de grafos dirigidos, donde los nodos representan proteínas individuales o complejos, y las aristas sus interacciones, describiendo típicamente la acción de un nodo enzimático que activa o inhibe a otra proteína. Como paso previo, *Hipathia* disecciona cada ruta de señalización en uno o más subgrafos que representan cascadas de señalización individuales, cuyo objetivo general es responder a un estímulo específico, como una señal del medio extracelular, o un suceso interno como la detección de daño en el

ADN. En ambos casos, el resultado al final de la cascada de interacciones suele ser la activación de una proteína efectora destinada a traslocarse al interior del núcleo y promover la expresión de un conjunto de genes que responderán posteriormente al estímulo detectado.

En la práctica *Hipathia* es una función que transforma la matriz de actividad génica en una matriz de actividad de rutas. En otras palabras, es una función que cumple que:

$$\hbar : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}^{p \times n} \tag{3.4}$$

$$X_p = \hbar(X_g), \tag{3.5}$$

siendo m el número de genes, p el número total de cascadas de señalización y n el número de individuos.

Para cuantificar el nivel de actividad de una cascada de señalización, *Hipathia* simula la propagación de una señal virtual que recorre la red desde los nodos iniciales hasta el nodo final del subgrafo. En este caso, la señal virtual trata de reproducir el flujo de información real que se produce cuando un ligando del medio extracelular es capturado por un receptor de membrana, disparando una cascada de interacciones que culmina con la activación de la proteína efectora. Para tener en cuenta la actividad de los genes, *Hipathia* modula el flujo que atraviesa cada nodo de la red en función de su nivel de actividad. Además, dado que cada proteína puede ser activada o inhibida por varias proteínas o complejos, el flujo de salida de un nodo particular j, se define a partir de la siguiente ecuación:

$$s_j^{(t)} = v_j \left[1 - \prod_{a \in A} (1 - s_a^{(t-1)}) \right] \left[\prod_{i \in I} (1 - s_i^{(t-1)}) \right], \tag{3.6}$$

donde $s_j^{(t)}$ se corresponde con el flujo de salida del nodo j en la iteración t; v_j es su valor de actividad; y, finalmente, $s_a^{(t-1)}$ y $s_i^{(t-1)}$ representan, respectivamente, el valor de flujo de salida calculado en la iteración anterior en cada uno de sus A activadores e I inhibidores. Esta ecuación se aplica a cada nodo a medida que la señal virtual se propaga por el subgrafo, siendo la cantidad de flujo a la salida del nodo final el valor que directamente describe la actividad de la cascada. Además,

dado que existe la posibilidad de contar con bucles de retroalimentación dentro de las rutas, el proceso se mantiene de forma iterativa hasta que el valor de flujo de salida permanece de forma estable.

Generación de un conjunto de ecuaciones equivalentes a *Hipathia*

Hipathia emplea un proceso iterativo para estimar la actividad de cada cascada de señalización. Este enfoque le permite modelar de forma eficiente los bucles presentes en el grafo, pero complica su integración en procesos más generales de optimización, como la factorización de matrices. En concreto, al no calcular de forma analítica el valor de flujo de cada cascada, no se dispone de un conjunto de ecuaciones que represente dicho proceso y que permita el cálculo de derivadas parciales en un proceso de optimización clásico, como los basados en un gradiente descendente.

Para solventar esta limitación, se diseñó una versión modificada de *Hipathia* (Figura 3.3) con el objetivo de proporcionar una ecuación característica para cada cascada de señalización, que fuera equivalente al proceso iterativo de *Hipathia* durante un número de ciclos previamente definido. Como resultado, redefinimos la función de *Hipathia* como $\hbar = [J_1, J_2, ..., J_p]$, donde J_i se corresponde con la ecuación asociada a la i-ésima cascada de señalización.

Para obtener el conjunto de ecuaciones se realizó una ejecución controlada de *Hipathia*, tomando como entrada una matriz de actividad generada de forma aleatoria. En este caso, además del cálculo de la señal, en cada nodo visitado durante el proceso iterativo se emitió a la salida su ecuación correspondiente, donde se describe su valor de flujo de salida en cada iteración realizada en función del valor de flujo de salida de sus inhibidores y activadores particulares en el paso anterior. Este proceso generó como resultado un conjunto de ecuaciones que definen el valor de flujo de salida en cada nodo en función del tiempo. Después, se tomó la última ecuación emitida para cada nodo final, y de forma recursiva, se fueron substituyendo los términos definidos a tiempo t por los términos que incluyen a

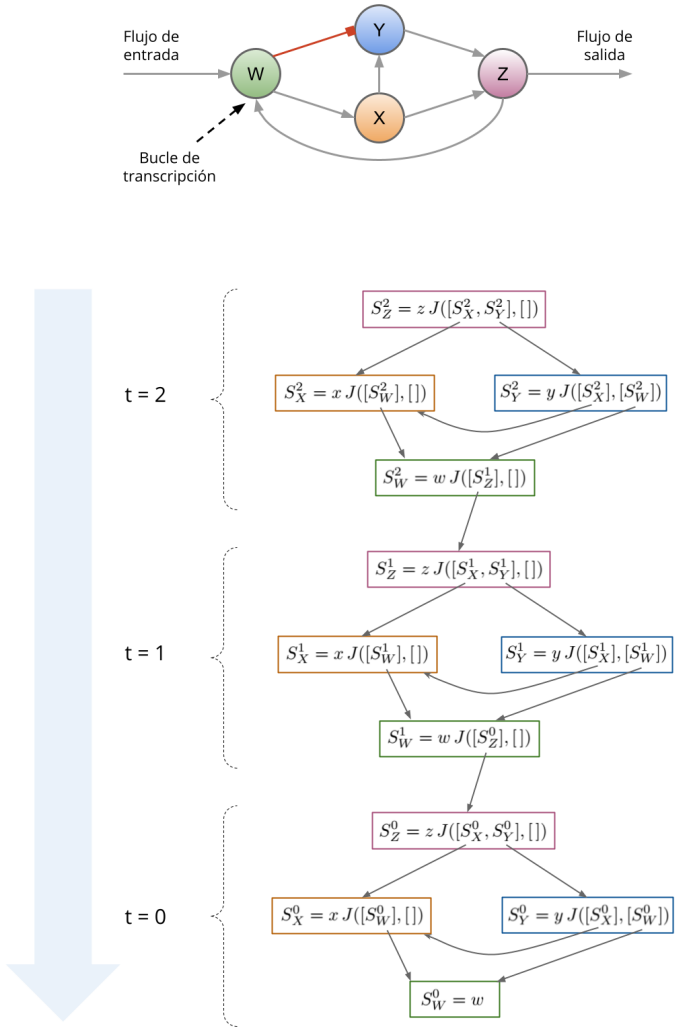


Figura 3.3: *Obtención de una ecuación equivalente al algoritmo de Hipathia durante 3 ciclos para una red de ejemplo. Los términos que constituyen la ecuación del flujo de salida del nodo Z a tiempo $t = 2$ (S_z^2) se substiyen de forma recursiva hasta llegar a los valores recogidos a tiempo $t = 0$. Las variables $\{w, x, y, z\}$ se corresponden con los valores de actividad de cada nodo. $J(A, I)$ se corresponde con la función característica de Hipathia, siendo A e I los vectores de flujos de entrada a un nodo que activan e inhiben respectivamente.*

sus reguladores a tiempo $t-1$. Al finalizar el proceso recursivo, se obtuvo una expresión para cada nodo final que describe su valor de flujo de salida en función del valor de los nodos de la red a tiempo $t=0$, que se corresponde con el valor de actividad proporcionado por la matriz de entrada. Finalmente, el conjunto de ecuaciones final fue validado frente al valor original de la herramienta.

3.3.3. Detección de funciones biológicas alteradas

La cuantificación de las rutas moleculares nos permite estimar la actividad de las células tumorales desde un punto de vista más sistémico. Esta aproximación permite en la práctica disponer de una visión más global del cáncer, y por consiguiente, menos reduccionista, como la ofrecida por enfoques clásicos donde los genes son los principales protagonistas. En este caso, el objetivo consiste en estimar las funciones biológicas alteradas en los tumores a través del patrón de actividad observado en las cascadas de señalización que las regulan, cuantificadas gracias a la aplicación de *Hipathia*.

Para poder definir las diferentes funciones biológicas de interés es necesario recurrir a una base de datos como *Gene Ontology* (*GO*), la cual describe una ontología de términos o funciones biológicas en las que las proteínas participan. Para poder cuantificar dichas funciones, en primer lugar, es necesario anotar cada una de las cascadas de señalización a los términos de *GO* en los que intervienen. En este caso, se sigue el protocolo propuesto en la publicación original de *Hipathia*, donde las anotaciones disponibles a nivel de gen o proteína son proyectadas sobre las rutas de señalización en las que participan. En concreto, cada cascada de señalización hereda los términos de *GO* anotados a su proteína efectora, ya que su objetivo final es activar dicha proteína, y por tanto, sus funciones asociadas. Hay que resaltar que cada proteína puede estar involucrada en diferentes funciones biológicas, y por tanto, también sus cascadas. De la misma forma, una determinada función biológica puede estar regulada por diferentes rutas de señalización, representando así diferentes escenarios donde la célula necesita iniciar la misma respuesta en función a diferentes estímulos.

Esta aproximación no solo permite fácilmente determinar aquellas funciones biológicas en las que una determinada cascada está involucrada, sino también, obtener un grupo de cascadas que regulan la misma función biológica. En la práctica, la actividad de este grupo de cascadas se representa mediante la submatriz X_p, que se acompaña de la correspondiente submatriz de genes X_g, definiendo así la entrada al modelo jerárquico de factorización que se describe en el apartado siguiente.

Para poder determinar las funciones alteradas en los individuos enfermos, en primer lugar se construye la matriz X_f, correspondiente a la matriz de actividad de funciones, análoga a las matrices X_p y X_g. En este caso, el vector de actividad para cada función biológica se obtiene promediando la actividad de cada uno de las cascadas que la regulan (X_p), en concreto:

$$X_{f,i,j} = \sum_{l \in \Lambda} \frac{X_{p,l,j}}{\lambda}, \tag{3.7}$$

donde $X_{f,i,j}$ se corresponde con el valor de actividad de la función biológica i en el individuo j, calculado como el promedio del valor de las Λ cascadas de señalización que la regulan, siendo λ el número de cascadas.

Después, para cada función biológica representada en la matriz X_f, se aplica un test no paramétrico de comparación de medias (Mann y Whitney, 1947), con el fin de definir si dicho término muestra diferencias significativas al comparar los valores asociados a los individuos normales frente a los tumorales. En caso positivo, el término es considerado como una potencial función biológica alterada y por tanto, incluido en el proceso posterior de factorización.

En un contexto real, el número de funciones biológicas alteradas en un tumor es bastante elevado, lo que introduce una carga computacional muy importante en los análisis posteriores. Para reducir este efecto, se propone utilizar la estructura de la ontología representada en *GO* para poder aliviar el número de resultados significativos, por medio de la agrupación de términos que aparecen muy cerca en la ontología, y que, por tanto, describen a los mismos procesos biológicos desde puntos de vista ligeramente distintos. Esta aproximación se realiza a partir del cálculo de distancias semánticas entre los términos que han resultado ser inicialmente

significativos, teniendo como objetivo la obtención de grupos de términos, que posteriormente se reduce a solo un representante, limitando así el número de términos significativos en el análisis, y por tanto, el número de factorizaciones a realizar.

Para realizar esta tarea, se hizo uso del paquete de *R GOSemSim* (Yu y cols., 2010), que realiza el cálculo de distancias semánticas entre los términos significativos. Después, para cada término evaluado, se determina qué otros terminos están a un distancia semántica muy cercana y si alguno de ellos ha mostrado un p-valor más pequeño en el análisis inicial. En caso positivo, el término es propuesto para su eliminación, dejando como representante provisional al término con p-valor más pequeño al que se asocia. De esta forma, al final del proceso, se obtiene una lista muy reducida de términos, conservando la misma cantidad de información biológica.

3.3.4. Modelo jerárquico de factorización

Tal y como se ha descrito, el modelo jerárquico de factorización planteado en esta tesis tiene por objetivo la extracción de un conjunto de componentes latentes que describan las diferentes estrategias implementadas en los tumores de pacientes reales para desarrollar los *hallmarks* del cáncer.

Con esta visión, se plantea un modelo jerárquico *gen*→*ruta*→*función*, donde el proceso de factorización se realiza de forma simultánea para la matriz de actividad de rutas (X_p) y su correspondiente matriz de actividad de genes (X_g) involucradas en la ejecución de una determinada función biológica. En este caso, la función objetivo de la optimización incorpora una serie de términos adicionales que permiten llegar a soluciones donde ambos conjuntos de componentes son compatibles, lo que permite describir de forma adecuada los distintos niveles de heterogeneidad observados en las muestras reales.

A continuación, se describen los diferentes pasos para implementar dicho modelo.

Estimación de componentes latentes

Las matrices de actividad X_g y X_p representan el punto de partida del modelo jerárquico. De forma análoga al abordaje clásico de *NMF*, el proceso de factorización se define para genes y rutas de la siguiente forma:

$$X_g \approx W_g H_g \tag{3.8}$$

$$X_p \approx W_p H_p, \tag{3.9}$$

donde $W_g \in \mathbb{R}^{m_g \times k_g}$ y $W_p \in \mathbb{R}^{m_p \times k_p}$ son las matrices de componentes, $H_g \in \mathbb{R}^{k_g \times n}$ y $H_p \in \mathbb{R}^{k_p \times n}$ las matrices de mezcla, m_g y m_p el número de genes y cascadas, y k_g y k_p representan número de componentes empleado pare genes y rutas, respectivamente.

La función objetivo de *NMF* aplicada de forma independiente a cada matriz de actividad trata de minimizar la siguiente expresión:

$$f = \|X - WH\|_f^2 , \tag{3.10}$$

donde $\|\cdot\|_f^2$ se corresponde con la norma de *Frobenius*, definida con el fin de evaluar la distancia entre la matriz de entrada X y la solución propuesta en cada iteración.

El binomio de esta expresión (3.10) se descompone de la siguiente forma:

$$f = \frac{1}{2}(XX^T - 2WHX^T + WHH^TWT), \tag{3.11}$$

de la cual se derivan las derivadas parciales correspondientes a W y H:

$$\frac{\partial f}{\partial W} = -XH^T + WHH^T \tag{3.12}$$

$$\frac{\partial f}{\partial H} = -W^TX + W^TWH. \tag{3.13}$$

La optimización puede resolverse mediante un esquema clásico de gradiente descendente. Sin embargo, Lee y Seung propusieron (1999) un método de optimización basado en actualizaciones multiplicativas para ambas matrices, lo que permite una convergencia más rápida y evita la necesidad de truncar valores negativos

aparecidos en las matrices al aplicar el gradiente durante la actualización. Durante la optimización las matrices del modelo se actualizan de la siguiente forma:

$$W = W \odot \frac{XH^T}{WHH^T} \tag{3.14}$$

$$H = H \odot \frac{W^T X}{W^T WH}. \tag{3.15}$$

La aplicación del método de Lee y Seung permite factorizar ambas matrices de actividad ofreciendo una solución razonable. Sin embargo, este abordaje no tiene en cuenta la necesidad de encontrar dos conjuntos de componentes que sean compatibles entre sí para asegurar la consistencia del modelo jerárquico. Para resolver esta limitación, añadimos el siguiente término a la función objetivo:

$$\left\| \hbar(W_g) - W_p S^T \right\|_f^2. \tag{3.16}$$

Este término permite evaluar la diferencia entre ambos conjuntos de componentes. En concreto, se compara el conjunto de componentes a nivel de ruta (W_p) con el conjunto de componentes obtenido al aplicar la función de *Hipathia* sobre las componentes a nivel de gen ($\hbar(W_g)$). Dado que ambas matrices suelen disponer de un número diferente de componentes, para poder realizar dicha comparación es necesario contar con la matriz auxiliar $S \in \mathbb{R}^{k_g \times k_p}$, que se define como una matriz binaria.

S concentra la esencia del modelo jerárquico, ya que describe qué componentes a nivel de gen se asocian con la misma componente a nivel de ruta. Debido a su estructura binaria, y con el fin de producir una convergencia suave, S se aproxima durante la optimización mediante una expresión sigmoidal de la siguiente forma:

$$S = \frac{1}{1 + e^{-Y}}, \tag{3.17}$$

donde Y, con las mismas dimensiones que S, permite pasar del rango [-Inf, +Inf] al rango [0, 1].

Además, dentro del modelo es importante penalizar soluciones de S que lleven a más de un valor distinto de cero por fila, lo que llevaría a la asignación de algunas

componentes a nivel gen a más de una componente a nivel de ruta. Para ello, se añade la siguiente expresión:

$$\left\| SO_{k_p} - O_{k_g} \right\|_f^2, \tag{3.18}$$

donde O_{k_p} y O_{k_g} se corresponden con vectores columna con valores igual a 1 de tamaño k_p y k_p respectivamente. En este caso, la expresión favorece que la suma de valores por fila sea igual a 1 para S, lo cual combinado con la expresión anterior, permite acelerar la convergencia hacia una solución idónea para S.

Por otro lado, es importante que la matriz S refleje una correspondencia balanceada entre componentes a nivel de gen y ruta, evitando soluciones que contengan componentes a nivel de ruta con un número muy elevado de componentes asociadas a nivel de gen. Para ello, se añade el siguiente término:

$$\left\| S^T O_{k_g} - \frac{k_g}{k_p} O_{k_p} \right\|_f^2, \tag{3.19}$$

donde $\frac{k_g}{k_p}$ representa el número esperado de componentes a nivel de gen asociados a la misma componente a nivel de ruta, en un caso ideal.

De forma análoga al término que facilita la compatibilidad entre ambos juegos de componentes, se añade un término adicional que penaliza soluciones donde las matrices de mezcla a ambos niveles no sean coherentes. En concreto:

$$\left\| S^T H_g - H_p \right\|_f^2. \tag{3.20}$$

Finalmente, la función de error completa que incluye los pesos asociados a cada término queda definida de la siguiente forma:

$$\begin{aligned} f =& \alpha \left\| X_g - W_g H_g \right\|_f^2 + \beta \left\| X_p - W_p H_p \right\|_f^2 + \\ &+ \gamma_1 \left\| \hbar(W_g) - W_p S^T \right\|_f^2 + \gamma_2 \left\| S^T H_g - H_p \right\|_f^2 + \\ &+ \rho_1 \left\| SO_{kp} - O_{k_g} \right\|_f^2 + \rho_2 \left\| S^T O_{k_g} - \frac{k_g}{k_p} O_{k_p} \right\|_f^2. \end{aligned}$$

De esta ecuación, se derivan las siguientes derivadas parciales para cada término que nos permiten actualizar cada una de las matrices del modelo durante la

optimización:

$$
\begin{aligned}
W_g^{(t+1)} &= W_g^{(t)} \odot \frac{\alpha X_g H_g^T + \gamma_1 \dfrac{\partial \hbar(W_g)}{\partial W_g} W_p S^T}{\alpha W_g H_g H_g^T + \gamma_1 \dfrac{\partial \hbar(W_g)}{\partial W_g} \hbar(W_g)} \\
H_g^{(t+1)} &= H_g^{(t)} \odot \frac{\alpha W_g^T X_g + \gamma_2 S H_p}{\alpha W_g^T W_g H_g + \gamma_2 S S^T H_g} \\
W_p^{(t+1)} &= W_p^{(t)} \odot \frac{\beta X_p H_p^T + \gamma_1 \hbar(W_g) S}{\beta W_p H_p H_p^T + \gamma_1 W_p S^T S} \\
H_p^{(t+1)} &= H_p^{(t)} \odot \frac{\beta W_p^T X_p + \gamma_2 S^T H_g}{\beta W_p^T W_p H_p + \gamma_2 H_p} \\
Y^{(t+1)} &= Y^{(t)} - \eta \odot [\\
&\quad - \left(\gamma_1 \hbar(W_g)^T W_p + \gamma_2 H_g H_p^T + \rho_1 O_{kg} O_{kp}^T + \rho_2 \frac{k_g}{k_p} O_{kg} O_{kp}^T \right) \\
&\quad + \left(\gamma_1 S W_p^T W_p + \gamma_2 H_g H_g^T S + \rho_1 S O_{kp} O_{kp}^T + \rho_2 O_{kg} O_{kg}^T S \right) \\
&\quad] \odot \frac{\partial S}{\partial Y},
\end{aligned}
$$

donde

$$
\frac{\partial S}{\partial Y} = S \odot (1 - S), \tag{3.21}
$$

siendo η el factor de aprendizaje empleado en la convergencia de la matriz Y.

En el modelo $\frac{\partial \hbar(X)}{\partial X} \in \mathbb{R}^{m \times p}$ nos permiten determinar cual es la contribución de cada uno de los m genes en las p cascadas para una matriz X. El valor de dicha expresión para un gen g y una cascada c se define como:

$$
\left[\frac{\partial \hbar(X)}{\partial X} \right]_{gc} = \sum_{i=1}^{n} \frac{\partial J_c}{\partial g} (X_{Qi}), \tag{3.22}
$$

donde i se corresponde con cada una de las observaciones incluidas en la matriz X, y J_c con la ecuación correspondiente a la cascada c en la función *Hipathia* a partir de los Q genes que la integran.

Por último, tal y como se aprecia, cada término de la ecuación incluye un peso que ha de ser ajustado antes de realizar la optimización. Para ello, se usa el paquete de *R DEoptim* (Ardia, Boudt, Carl, Mullen, y Peterson, 2011), el cual, por

mediación de un algoritmo genético, realiza una búsqueda en el espacio de error. En este caso, con el objetivo de emplear una carga computacional asumible, cada ejecución del modelo jerárquico dentro del algoritmo genético empleará únicamente 100 iteraciones, con un total de 50 pasos dentro de la búsqueda.

Estimación del numero óptimo de componentes latentes

El parámetro característico de cualquier factorización matricial lo representa el número de componentes latentes a emplear en el proceso. Se trata de un parámetro habitualmente desconocido y su adecuada selección resulta fundamental ya que, el uso de un valor por debajo del idóneo llevaría a una factorización subóptima, y un número demasiado elevado a una fragmentación excesiva de las componentes originales.

Una de las aproximaciones más habituales al problema consiste en la realización de sucesivas ejecuciones, empleando un rango de valores candidatos suficientemente amplio, para después determinar cual es el número de componentes que ofrece una factorización más adecuada. Esta aproximación, aunque robusta, muestra algunas limitaciones importantes. La primera es a nivel computacional, ya que dependiendo de la amplitud del rango, podría ser necesario realizar un gran número de factorizaciones, teniendo que escoger necesariamente un número máximo de iteraciones bastante reducido para no disponer de una carga computacional inasumible. Por otro lado, resulta totalmente necesario emplear métricas que ponderen la complejidad del modelo frente a la bondad del ajuste, ya que, de forma natural, un número elevado de componentes permitirá realizar un ajuste más preciso, teniendo como límite un número equivalente al número de observaciones empleado.

La estimación del número de componentes ha sido objeto de estudio en multitud de trabajos. Alguna de las aproximaciones más empleadas utilizan los conocidos criterios *AIC* (Akaike, 1974) o *BIC* (Schwarz, 1978) que tratan de ponderar el error obtenido frente a la complejidad del modelo en modelos de regresión. También

destaca el coeficiente de correlación cofenético (Saraçli, Doğan, y Doğan, 2013), que evalua la concordancia entre la distancia euclídea de las muestras y la distancia obtenida a partir de un *clustering* jerárquico realizado con la matriz de mezcla. El método de la silueta (Rousseeuw, 1987) también ha sido utilizado para este propósito, ya que permite estimar la distancia de cada observación a otros grupos adyacentes de muestras incluidos en el *clustering* jerárquico, lo que proporciona de forma implícita una forma de evaluar la consistencia obtenida mediante las matrices del modelo.

El modelo jerárquico planteado en esta tesis permite aplicar de forma natural una restricción importante en la estimación del número de componentes, ya que, la pareja de valores óptima para genes (k_g) y rutas (k_p) debería ser aquella que maximizara la relación entre las componentes encontradas en ambos niveles. En la práctica, sin embargo, esta aproximación resulta complicada ya que, al tratarse de una factorización simultánea, el proceso de selección necesitaría $|K_p| \times |K_g|$ ejecuciones, donde K_p y K_g se corresponden con el conjunto de valores posibles para rutas y genes, respectivamente.

Para poder reducir la carga computacional, se plantea una fase inicial compuesta por una secuencia de sucesivas factorizaciones realizadas de forma independiente para genes y rutas, empleando para ello el esquema de factorización sugerido por Lee y Seung (1999). El objetivo en este caso es acotar la búsqueda a un intervalo reducido que potencialmente contenga al valor óptimo para k_p y k_g, para después aplicar el modelo jerárquico únicamente sobre este subconjunto de valores. Esta solución limita el numero de factorizaciones a $|K_p^0| + |K_g^0| + |K_p| \times |K_g|$, donde K_g^0 y K_p^0 se corresponden con los intervalos iniciales, y K_g y K_p se corresponden con los intervalo acotados.

La selección del intervalo se basa en la evolución de la bondad ajuste, a medida que incrementamos el número de componentes en la factorización. En particular, se trata de seleccionar un punto en la curva de ajuste a partir del cual el error

cuadrático cae por debajo de un cierto umbral, definido de la siguiente forma:

$$k^0 = \underset{k}{\operatorname{argmin}} \sqrt{k^2 + wy^2}, \tag{3.23}$$

donde k^0 representa al valor seleccionado de todos los posibles $k \in K$, y es el error en el ajuste y w es una constante de normalización (típicamente con un valor de 3).

Además, gracias a que la bondad del ajuste muestra típicamente una curva exponencial (Figura 3.4), en la práctica es posible estimar su valor sin necesidad de probar todos los posibles valores de k dentro del rango definido. En concreto, el valor de error se ajusta mediante la minización de la siguiente expresión:

$$\left\| \epsilon_k - e^{-zk'} \right\|^2, \tag{3.24}$$

donde $k' = k - min(k)$, ϵ_k se corresponde con el error (normalizado entre 0 y 1) obtenido al emplear un número k de componentes, y z es el parámetro a optimizar.

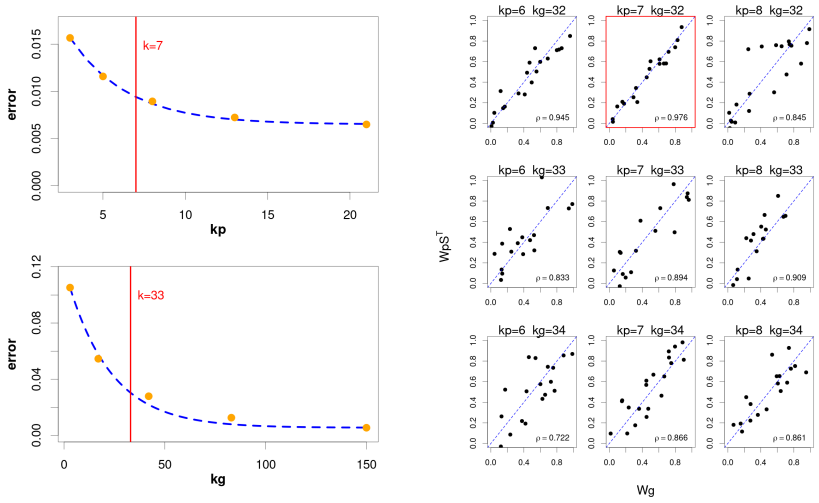


Figura 3.4: *Método empleado para estimar el número óptimo de componentes en un caso de ejemplo. A la izquierda se describe el ajuste exponencial de la curva de error para rutas (arriba) y genes (abajo), tomando como puntos de referencia a 5 factorizaciones realizadas. A la derecha se describe el resultado obtenido por el modelo jerárquico para las 9 alternativas seleccionadas, marcando en rojo la solución seleccionada.*

El ajuste exponencial nos permite reducir el número de ejecuciones a solo dos

valores candidatos situados en los extremos del rango, y a unos pocos (típicamente 3) distribuidos de forma uniforme por todo el intervalo. El rango final a explorar se define por el intervalo $K_p = [k_p^0 - d, k_p^0 + d]$ y $K_g = [k_g^0 - d, k_g^0 + d]$, para rutas y genes respectivamente, donde d, con un valor típico de 1, define la amplitud del rango a explorar por el modelo jerárquico.

Por último, en el rango seleccionado se aplica el modelo de factorización jerárquico, seleccionando como valor óptimo para k_p y k_g aquel con proporcione una mayor coherencia entre las componentes Wg y Wp correspondientes. En concreto, se selecciona el valor de k_g y k_p que minimice la siguiente expresión:

$$(k_p, k_g) = \underset{K_p, K_g}{\operatorname{argmin}} \left\| \hbar(W_g^{(x)}) - W_p^{(y)} S^{(xy)T} \right\|_f^2 \tag{3.25}$$

$$\forall (x, y), x \in K_g, y \in K_p. \tag{3.26}$$

3.3.5. Análisis del modelo jerárquico de factorización

La ejecución del modelo jerárquico de factorización proporciona una descripción detallada de cómo se produce la regulación de una determinada función biológica en los individuos bajo estudio. La base de esta descripción está constituida por el conjunto de componentes encontradas durante el proceso, y su caracterización posterior permite determinar elementos importantes como el caracter oncogénico de las componentes, o la relación de sus genes más relevantes con estudios previos realizados en la enfermedad. Además, la distribución de los pesos de las componentes a lo largo de los individuos permitirá determinar si éstas se relacionan con características clínicas relevantes.

A continuación se describen los diferentes pasos a realizar para el análisis de las matrices optimizadas por el modelo.

Bondad del ajuste

El primer paso en la evaluación del modelo de factorización consiste en determinar si la solución encontrada, compuesta por el conjunto de componentes latentes y sus correspondientes matrices de mezcla, constituye una aproximación adecuada a las matrices originales de entrada X_g y X_p. Esta evaluación se lleva a cabo de forma independiente para genes y rutas, mediante la aplicación de un modelo de regresión lineal simple que toma como variable independiente a la matriz original y como variable dependiente a la matriz estimada.

Asimismo, resulta fundamental evaluar si las componentes latentes obtenidas a nivel de gen se agrupan correctamente y proporcionan soluciones similares a sus componentes asociadas a nivel de ruta, describiendo así una convergencia adecuada del modelo jerárquico de factorización. Al igual que antes, la relación entre ambos conjuntos de componentes se evalúa mediante un modelo de regresión simple entre la matriz $\hbar(W_g)$ y su correspondiente matriz $W_p S^T$. Además, también se visualiza la matriz de correlación entre ambos términos, pudiendo así determinar si componentes no asociadas en ambos niveles también muestran cierto grado de correlación, lo que podría indicar una convergencia a una solución poco adecuada. De forma equivalente, se evalua también la similitud entre las matrices de mezcla obtenidas en ambos niveles. En este caso, se determina si el peso de cada componente a lo largo de los individuos a nivel de ruta (H_p) resulta equivalente al peso acumulado de sus componentes asociadas a nivel de gen ($S^T H_g$).

Por último, se analiza la matriz S que describe la asociación entre las componentes encontradas a ambos niveles. En particular, la matriz S permite determinar si algunas componentes a nivel ruta muestran una asociación con un número elevado de componentes a nivel de gen, lo que indicaría una estrategia a nivel de ruta altamente heterogénea. Además, la visualización de la matriz S nos permite determinar si al final de la optimización alguna de las componentes a nivel de ruta no muestra asociación con ninguna componente a nivel de gen, lo que indicaría una convergencia inadecuada del modelo.

Caracterización de las componentes

Cada una de las componentes encontradas representa una combinación particular de valores de los elementos más básicos que las forman. Mientras que las componentes a nivel de ruta describen valores específicos en las cascadas de señalización anotadas a la función biológica estudiada, las componentes a nivel de gen, describen valores específicos en los genes o complejos que las forman. A nivel molecular, las combinaciones de valores representan estratégias observadas en los pacientes a la hora de regular la función biológica bajo estudio, y a un nivel más sencillo, describen qué elementos más básicos han de ser sobreactivados o inhibidos simultáneamente para producir el efecto deseado.

Aunque a nivel biológico hay que considerar que el conjunto total de genes y cascadas incluidos en la factorización está involucrado en la regulación de la función biológica, no significa que necesariamente todos los elementos esten alterados en alguno de los subgrupos de individuos del estudio. En este sentido, es lógico considerar que una parte importante de los elementos básicos no mostrarán un valor particularmente extremo en ninguna de las componentes, por lo que podrían ser considerados como elementos poco relevantes a la hora de explicar los mecanismos alterados en la enfermedad.

Siguiendo este planteamiento, el primer paso en la caracterización se basa en determinar para cada componente cual es su conjunto de elementos básicos más relevante, considerando al resto poco informativos. Hay que resaltar que un elemento podría ser muy relevante en una componente en particular y ser poco informativo en el resto, lo que indicaría que se trata de una característica particular y esencial de dicha componente. Para determinar si un elemento básico es relevante en alguna de las componentes, se toma su distribución de valores a lo largo de todas ellas y se evalua si en alguna de las componentes muestra un valor extremo, considerando como extremo un valor más allá de 1.5 veces el rango intercuartílico. Esta aproximación lleva a considerar como no relevantes en el conjunto total de componentes a aquellos elementos cuya distribución no muestre ningún valor

extremo.

Además, en el caso de las componentes a nivel de gen, es posible realizar un análisis adicional. En este caso, se trata de evaluar si un determinado gen, aun mostrando una variación pequeña entre las componentes, podría producir un gran efecto en el espacio de las rutas. Este concepto se sustenta en la estructura topológica de las rutas moleculares, donde algunas proteínas o complejos con un papel central podrían ser nodos cuya variación resulta ser más importante. De forma complementaria, algunos genes que muestran gran variación en el espacio de los genes, podrían no propagar dicho cambio en el espacio de las rutas, describiendo así a genes cuya variación resulta poco relevante para la regulación de la función biológica.

Para evaluar si en una componente en particular un determinado gen es relevante en el espacio de las rutas, es necesario emplear la función de *Hipathia*. En concreto, la función se aplica de forma repetida para la misma componente, fijando el valor de todos los genes, a excepción del gen a evaluar, el cual toma de forma secuencial el valor obtenido a lo largo de todas las componentes. En este caso, se considera que el gen es relevante si el valor de alguna de las cascadas de la componente original constituye un valor extremo al compararla frente al resto de repeticiones.

Caracterización oncogénica de las componentes

Además del análisis anterior, el modelo de factorización permite determinar el carácter oncogénico de cada una de las componentes encontradas a nivel de gen. Para ello, se determina si algunos de los genes considerados relevantes en cada componente han sido previamente descritos como genes de cáncer, comparando además sus niveles de actividad frente a los de otros genes previamente no descritos.

Para definir el rol de cada gen, hacemos uso del censo de genes disponible en el repositorio *COSMIC* (Tate y cols., 2019). En él, se recogen los genes implicados en gran multitud de estudios de cáncer, haciendo énfasis en aquellos que aparecen alterados de forma recurrente, especialmente entre diferentes tipos de cáncer.

COSMIC distribuye los genes de cáncer en 2 categorías principales: oncogenes y supresores de tumor. Además, aporta una tercera categoría que describe a aquellos genes cuyo comportamiento es ambivalente en función del contexto celular.

Para llevar a cabo este análisis se evalua el valor de actividad de los genes relevantes en cada componente en función de la categoría a la que pertenezcan, considerando las categorías aportadas por *COSMIC*, más una categoría adicional para agrupar a aquellos genes que no tienen un rol oncogénico conocido.

Análisis de subtipos

Las matrices de mezcla obtenidas durante la factorización describen el peso que cada una de las componentes encontradas tiene a lo largo de los individuos. Esta información resulta de gran importancia a la hora de caracterizar a los individuos del estudio, ya que aquellos que tiendan a estar representados por las mismas componentes constituyen de forma natural grupos de sujetos con un perfil molecular parecido.

Siguiendo esta idea, dentro de este subapartado se realiza un *clustering* jerárquico a partir de la matriz de mezcla, aplicándose de forma independiente para genes y rutas. En este caso, se trata de determinar si las agrupaciones observadas en los individuos se corresponden con los subtipos conocidos en la enfermedad, lo cual sugeriría que las componentes encontradas por el modelo describen estrategias específicas de cada subtipo.

De forma complementaria al análisis de *clustering*, se obtiene una matriz de prevalencias, que describe el porcentaje de individuos de cada subtipo con una contribución significativa de cada componente. Este punto permite determinar qué componentes tienen un peso mayor en los invididuos sanos, y cuales representan mejor a uno o varios subtipos de la enfermedad, siendo a nivel práctico un indicador adicional sobre el carácter oncogénico de las componentes.

Para determinar la prevalencia de la componente k en el subgrupo ω, empleare-

mos la siguiente expresión:

$$\Gamma_k^{\omega} = \frac{\sum_i^{|\omega|} \tau(H_{k,i})}{|\omega|}, \tag{3.27}$$

donde $H_{k,i}$ se corresponde con el valor de mezcla para cada individuo i en la componente k, y τ a la función de pertenencia.

Dicha función de pertenencia se define para cada $H_{k,i}$ a partir del ajuste de dos distribuciones Beta (η_k^0 y η_k^1) al vector de valores de mezcla de cada componente ($H_{k,:}$). Las dos distribuciones representan al conjunto de individuos con una contribución residual de la componente (η_k^0) y al conjunto de individuos con un peso significativo (η_k^1), cuya función de pertenencia se define como:

$$\tau(H_{k,i}) = \begin{cases} 1 & \text{si } Be(H_{k,i}|\alpha_1, \beta_1)P(\eta_k^1) > Be(H_{k,i}|\alpha_0, \beta_0)P(\eta_k^0) \\ 0 & \text{en caso contrario,} \end{cases}$$

donde α_0 y β_0 se corresponden los parámetros característicos de la distribución η_k^0; y α_1 y β_1 son los parámetros característicos de la distribución η_k^1.

Análisis de supervivencia

Uno de los puntos más interesantes del análisis consiste en determinar si las componentes encontradas muestran algún tipo de relación significativa con variables clínicas adicionales disponibles para los individuos del estudio.

En este caso, se hace especial hincapié en la variable clínica que describe la supervivencia de los individuos, ya que constituye la característica fenotípica de mayor relevancia a nivel clínico en la enfermedad. En particular, se trata de determinar si aquellos individuos que muestran un peso significativo en una determinada componente, muestran también una probablidad de supervivencia distinta, lo que sugeriría que la sobreactivación o inhibición de sus elementos básicos constituye un mecanismo molecular que conduce a una forma más agresiva de la enfermedad.

Para llevar a cabo este análisis en cada componente, se estratifica a los individuos

en función del peso obtenido durante la factorización, dividiendo así a los sujetos en grupos de individuos con una contribución importante de la componente, frente a individuos con una contribución baja o nula. Después, se obtiene una curva de *Kaplan-Meier* con el objetivo de evaluar gráficamente la supervivencia en función de la estratificación obtenida. Asimismo, se realiza un contraste de hipótesis para determinar si las curvas pertenecientes a cada grupo muestran diferencias significiativas en la supervivenca. El análisis se realiza mediante el paquete *Survival* (Therneau, 2015) de *R*.

Análisis de sinergias entre componentes

Además de los análisis previamente descritos, las matrices de mezcla nos permiten determinar la existencia de sinergias entre componentes, correspondientes a grupos de componentes independientes con una distribución parecida a lo largo de los individuos. Una sinergia positiva entre dos o más componentes describe a un grupo de estrategias que ocurren de forma simultánea en los mismos individuos. Este hecho sugiere que dichas componentes forman parte de una estrategia de mayor escala, en la que probablemente es necesario alterar simultáneamente todos los elementos relevantes que aporta cada componente del grupo.

De forma complementaria, la presencia de sinergias negativas nos indicaría la existencia de estrategias excluyentes que no podrían convivir en los mismos individuos. Este concepto es conocido también como exclusividad y permite obtener información muy valiosa acerca de la secuencia de eventos producida en los pacientes para desarrollar el tumor.

Para llevar a cabo este análisis, se hace uso de las prevalencias obtenidas en el apartado anterior, evaluando el nivel de sinergia observado entre cada pareja posible de componentes. Para ello, se aplica un test de χ^2 con el objetivo de determinar si la aparición simultánea de dos componentes concretas en los mismos individuos ocurre con una frecuencia mayor de lo esperado en función de las frecuencias individuales de cada componente.

Además, en el caso de las sinergias positivas, es posible construir una red de sinergias, donde los nodos son las componentes y las aristas las sinergias significativas obtenidas mediante el test χ^2. Después, con el objetivo de poder determinar la presencia de grupos de 2 o más componentes conectadas, aplicamos un algoritmo de detección de comunidades a la red de sinergias obtenida.

3.3.6. Meta-análisis de funciones

La aplicación del modelo jerárquico de factorización a cada una de las funciones biológicas alteradas proporciona información muy valiosa acerca de como se implementan los *hallmarks* del cáncer en los distintos individuos enfermos. Tal y como se ha descrito, a partir del resultado proporcionado por cada factorización se realiza un *clustering* jerárquico destinado a describir como se distribuyen los individuos en el espacio de las componentes y determinar si estos se agrupan en función de su subtipo específico. En caso positivo, indicaría que las componentes encontradas en el modelo recogen de forma precisa las diferentes estrategias que los subtipos implementan a la hora de regular la función biológica estudiada.

En este contexto, a pesar de que los diferentes subtipos describen a nivel global perfiles moleculares muy distintos, no necesariamente implica la existencia de diferencias significativas en todas las funciones biológicas a estudiar. Hay que tener en cuenta que dichas funciones son seleccionadas por mostrar diferencias significativas entre individuos sanos y enfermos en su conjunto, sin tener en cuenta los distintos subtipos existentes. Este planteamiento sugiere que (i) en cada factorización podemos obtener agrupaciones distintas en función de si uno o varios subtipos muestran una regulación diferente al resto, y (ii) que para conseguir un agrupamiento general de los individuos es necesario integrar el resultado de todas las factorizaciones realizadas, ya que estas describen por separado partes específicas del mismo sistema.

Con este objetivo, se propone realizar un meta-análisis destinado a la obtención de una matriz global de distancias (D) que acumule las distancias obtenidas entre

los individuos en cada una de las funciones biológicas estudiadas. Dicha matriz se obtiene de forma independiente para genes y rutas. La distancia $D_{i,j}$ entre las muestras i y j se define mediante la siguiente expresión:

$$D_{i,j} = \sum_{q \in Q} d^q_{i,j} w^q, \tag{3.28}$$

donde $d^q_{i,j}$ representa la distancia euclidea obtenida de la matriz de mezcla en cada función q del conjunto total de funciones Q, y w^q al peso asociado a dicha función, calculado a partir del estadístico obtenido inicialmente al comparar a los individuos sanos frente a los enfermos.

Por último, a partir de las matrices globales de distancia obtenidas de forma independiente para genes y rutas realizamos el correspondiente *clustering* jerárquico, con el fin de determinar su similitud con los subtipos reales.

3.3.7. Experimentos

El modelo jerárquico de factorización se ha evaluado mediante dos estrategias complementarias. En primer lugar se evaluó su rendimiento mediante un conjunto de simulaciones destinadas a determinar si el modelo descrito es capaz de (i) estimar inicialmente el número de componentes esperado y (ii) recuperar de forma fiable las componentes incluidas en la simulación. En este caso, la evaluación tiene en cuenta tanto los conjuntos de componentes a nivel de gen y ruta, como su correcta relación jerárquica representada por la matriz S.

En segundo lugar, se aplicó el modelo jerárquico de factorización, y su posterior análisis, a un conjunto de muestras de cáncer de mama, con el fin de ejemplificar su uso en un contexto real.

A continuación se describe con detalle estos dos escenarios.

Simulaciones

Las simulaciones obtenidas en este apartado tratan de reflejar una estructura jerárquica de componentes similar a la esperada en un conjunto de muestras reales, lo que permite evaluar el modelo de factorización propuesto en un escenario realista.

El primer paso a realizar consiste en la simulación de las componentes latentes a nivel de ruta. Para ello, en primer lugar, es necesario escoger el conjunto de m_p cascadas que regulan la función biológica ficticia a simular, seleccionadas a partir del conjunto total de rutas disponibles en *Hipathia*. Después, se recogen los m_g genes incluidos en dichas rutas, obteniendo así el conjunto final de genes y cascadas que forman parte de la factorización.

Con el fin de simular la matriz $W_p \in \mathbb{R}^{m_p \times k_p}$, correspondiente a las k_p componentes latentes a nivel de ruta, se comienza con la inicialización de una matriz aleatoria a nivel de gen ($W_g^0 \in \mathbb{R}^{m_g \times k_p}$) construida como $W_g^0 \sim Beta(10, 10)$. Después, la función de *Hipathia* es aplicada a dicha matriz, proporcionando así el conjunto de componentes latentes a nivel de ruta empleado en la simulación:

$$W_p = \hbar(W_g^0). \tag{3.29}$$

Después, se procede a definir la matriz S, que describe al modelo jerárquico, y que en la práctica define el número de componentes a nivel de gen asociadas a cada una de las componentes a nivel de ruta. A continuación, se procede a inicializar la matriz de componentes a nivel de gen, también mediante una distribución Beta: $W_g \sim Beta(10, 10)$. Con el fin de poder obtener componentes a nivel de gen compatibles con la componente a nivel de ruta a la que se asocian, es necesario aplicar una modelo de factoración con el siguiente criterio de optimización:

$$\min \left\| \hbar(W_g) - W_p S^T \right\|_f^2 . \tag{3.30}$$

Ya que en este punto, tanto W_p como S son conocidas, únicamente es necesario optimizar el valor de W_g, empleando para ello el siguiente esquema de gradiente

descendente:

$$W_g = W_g - \eta \left[\frac{\partial \hbar(W_g)}{\partial W_g} \hbar(W_g) - \frac{\partial \hbar(W_g)}{\partial W_g} W_p S^T \right], \tag{3.31}$$

donde $\eta = 0{,}0001$ se corresponde con el factor de aprendizaje y $\frac{\hbar(W_g)}{\partial W_g}$ con la derivada parcial entre genes y rutas definida en la descripción del modelo jerárquico.

Una vez definidas las matrices de componentes (W_g y W_p) y su relación jerárquica (S), se procede a definir las correspondientes matrices de mezcla (H_p y H_g). Para ello, es tambien necesario aplicar un modelo de factorización, minimizando en este caso la siguiente expresión:

$$\min \|\hbar(W_g H_g) - W_p H_p\|_f^2 . \tag{3.32}$$

Después H_p y H_g se optimizan de la siguiente forma:

$$H_g = H_g - \alpha \left[W_g^T \frac{\partial \hbar(W_g H_g)}{\partial W_g H_g} \hbar(W_g H_g) - W_g^T \frac{\partial \hbar(W_g H_g)}{\partial W_g H_g} W_p H_p \right] \tag{3.33}$$

$$H_p = H_p - \alpha \left[W_p^T W_p H_p - W_p^T \hbar(W_g H_g) \right] . \tag{3.34}$$

Finalmente, las matrices obtenidas (W_g, W_p, H_g, H_p y S) son almacenadas con el fin de poder ser utilizadas posteriormente durante la evaluación del modelo jerárquico.

Análisis con datos reales

El modelo jerárquico de factorización ha sido también evaluado mediante datos correspondientes a pacientes reales. En este caso, se ha hecho uso de la cohorte de pacientes de cáncer de mama incluida en el consorcio internacional de cáncer *ICGC*. Para ello, se descargaron los datos de expresión génica desde el portal oficial del proyecto (https://daco.icgc.org/), seleccionando específicamente la última versión (*release* 28). En este caso, los datos descargados aparecen en formato *MAF*, donde cada linea representa el conteo de lecturas de secuenciación obtenidas para un gen particular, en un paciente concreto. Con el fin de poder disponer de los datos en un formato más práctico, se diseñó un *script* en *R* para convertir el formato de

entrada en una matriz, cuyas filas representan a los genes, y cuyas columnas a los individuos, obteniendo así la matriz de conteos brutos. Después, dicha matriz fue normalizada mediante paquete de *R DeSeq2* (Love, Huber, y Anders, 2014).

Asimismo, se descargaron los datos correspondientes a las mutaciones somáticas de los mismos pacientes. En este caso, a partir del fichero descargado, también en formato *MAF*, se generó el correspondiente archivo en formato *VCF*, a partir del cual se obtuve la matriz de afectación (X_g^v) a partir del protocolo descrito anteriormente.

Por último, la información clínica de los pacientes seleccionados fue descargada del repositorio *GDAC* (https://gdac.broadinstitute.org/) del *Broad Institute*.

3.4. Resultados

El modelo jerárquico de factorización ha sido evaluado mediante dos aproximaciones distintas. En primer lugar, las simulaciones descritas en el apartado anterior han permitido cuantificar la capacidad del modelo a la hora de recuperar las componentes previamente introducidas. Asimismo, se evaluó el protocolo empleado para estimar el número óptimo de componentes. Por último, el modelo jerárquico ha sido aplicado a una cohorte de pacientes con cáncer de mama, con el fin de ilustrar su uso en un escenario real. A continuación, se describen los resultados en ambas categorías

3.4.1. Análisis de simulaciones

La metodología desarrollada ha sido evaluada con un total de 100 simulaciones, diseñadas para reproducir la estructura jerárquica que genes y rutas moleculares muestran en el funcionamiento de las células.

En primer lugar, se evaluó el protocolo empleado para estimar el número óptimo de componentes a emplear durante la factorización. En este caso, el protocolo

fue además comparado con dos métodos clásicos de selección: el coeficiente de correlación cofenético y el método de la silueta. La Figura 3.5 muestra los resultados obtenidos en dicha comparación.

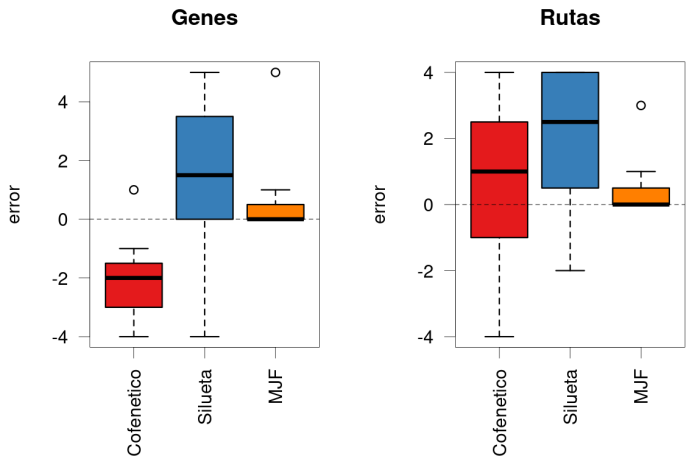


Figura 3.5: *Distribución de error obtenido en la estimación del número óptimo de componentes para el coeficiente de correlación cofenético, el método de la silueta y el modelo jerárquico de factorización (MJF), para genes y rutas, respectivamente.*

Tal y como se observa, la estimación ofrecida por el protocolo proporcionó un error mucho más reducido que el ofrecido por los otros dos métodos clásicos, reproduciendo el mismo patrón tanto para genes, como para rutas. Además, de forma notable, el protocolo propuesto ofreció en una buena parte de casos el número exacto de componentes esperado.

Por otro lado, las simulaciones fueron empleadas para el evaluar la capacidad del modelo a la hora de recuperar las matrices de componentes y mezcla introducidas. La Figura 3.6 muestra los valores de correlación obtenidos entre las matrices originales y las matrices estimadas obtenidas por el método original propuesto por Lee y Seung, el método *NNLS* (Kim y Park, 2008), y el modelo jerárquico propuesto.

En este caso, los resultados mostraron para todos los métodos una correlación cercana a 1 entre las matrices originales de actividad y sus correspondientes ma-

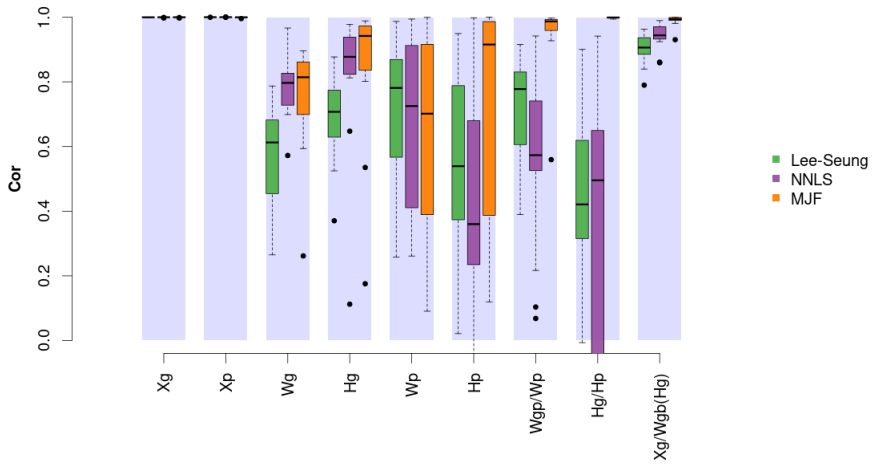


Figura 3.6: *Distribución de valores de correlación obtenidos entre las matrices simuladas y las obtenidas por el modelo de factorización originalmente propuesto por Lee y Seung, por el método basado en mínimos cuadrados no negativos (NNLS) y por el modelo jerárquico de factorización (MJF).*

trices estimadas, tanto a nivel de gen, como a nivel de ruta. Asimismo, el modelo jerárquico obtuvo las mejores estimaciones para las matrices de componentes y mezcla, salvo en el caso de las componentes a nivel de ruta, donde el método proporcionó correlaciones ligeramente peores. De forma característica, las correlaciones obtenidas entre las matrices de componentes a nivel de ruta y sus correspondientes componentes asociadas a nivel de gen, fueron notablemente superiores en el modelo jerárquico. De forma similar, las matrices de mezcla también mostraron correlaciones muy superiores, proporcionando en este caso valores próximos a 1.

3.4.2. Análisis con datos reales

El conjunto de muestras seleccionadas en este análisis se corresponde con el grupo de pacientes cuyo progreso en la enfermedad se situa en el estadio II. Esta selección tiene por objetivo reducir el número de componentes latentes a aquellas que describan el subtipo como única variable de interés, sin añadir otras caraterísticas clínicas que podrían estar incrementando la heterogeneidad de la muestra. Como requisito adicional, únicamente se tuvieron en cuenta aquellos individuos que simultáneamente dispusieron de secuenciación transcriptómica y genómica, con el fin de poder integrar correctamente el nivel de expresión y el efecto de las mutaciones somáticas en todos los individuos del estudio.

La Tabla 3.2 muestra la distribución de subtipos a lo largo de la cohorte de individuos seleccionados. En este caso, se aprecia una distribución asimétrica en el tamaño de los subtipos, muy similar a la distribución general incluyendo todos los estadios.

Selección de términos

A partir del conjunto de anotaciones proyectadas sobre las rutas de señalización se seleccionaron un total de 6381 términos biológicos recogidos por *Gene Ontology* (*GO*), en la categoría de *proceso biológico*. Para cada uno de los términos, se extrajo

Subtipo	Total	Frecuencia
Basal	60	0.1863
Her2	31	0.0963
LumA	107	0.3323
LumB	65	0.2019
Normal	59	0.1832

Tabla 3.2: Número total de individuos por subtipo en la cohorte seleccionada

el conjunto de genes incluidos en las rutas que participan en su regulación, y a partir de ellos, la submatriz de actividad génica (X_g). Después, se empleó la función de *Hipathia* con el fin de obtener la correspondiente matriz de actividad a nivel de ruta (X_p), utilizada, a su vez, para generar el vector de valores que mide la actividad del término en el conjunto total de individuos (X_f).

Después, con el fin de poder aplicar de forma coherente el modelo de factorización propuesto, el conjunto de términos seleccionados (6381) fue reducido a un total 1435 términos biológicos cuya matriz de actividad a nivel de ruta mostró un número de cascadas de señalización en el rango [10, 40].

A continuación, el análisis comparativo proporcionó un total de 1134 funciones biológicas con diferencias significativas en su valor de actividad, al comparar la población de individuos tumorales frente a la población de individuos normales. Por último, el análisis de similitud semántica permitió reducir el número de términos significativos a un total de 53 términos no redundantes, que describen funciones biológicas independientes, alteradas de forma característica en los individuos con cáncer de mama.

Aplicación del modelo jerárquico

Convergencia del modelo

El modelo jerárquico de factorización fue aplicado al grupo de términos signifi-

cativos no redundantes seleccionado en la fase anterior, obteniendo como resultado el conjunto de componentes latentes, tanto a nivel de gen, como a nivel de ruta, que describe las características intrínsecas de cada individuo y subtipo de la cohorte.

A nivel global, el proceso de optimización mostró una convergencia adecuada en todos los términos biológicos analizados. En particular, la bondad del ajuste obtenida entre las matrices de entrada (X_g y X_p) y las obtenidas por los modelos mostró un valor medio de varianza explicada de 0.9577 y 0.9910 para genes (Figura 3.7a) y rutas (Figura 3.7b), respectivamente.

Asimismo, el modelo jerárquico proporcionó para la totalidad de los términos biológicos analizados un grado de similitud muy alto entre las componentes a nivel ruta y sus correspondientes componentes asociadas a nivel de gen ($W_p S^T \approx \hbar(W_g)$), obteniendo en promedio una varianza explicada de 0.9933 (Figura 3.7c). De manera muy similar, las matrices de mezcla obtenidas en ambos niveles ($H_p \approx H_g S^T$) obtuvieron un valor de 0.9989 (Figura 3.7d). Este punto demuestra la idoneidad del modelo a la hora de encontrar soluciones coherentes entre genes y rutas.

De forma complementaria, la estructura de la matriz S también proporcionó algunas claves interesantes acerca de la convergencia del modelo. En particular, el número observado de componentes a nivel gen asociadas a cada componente a nivel de ruta, fue en promedio 1.73 veces mayor de lo esperado (Figura 3.7e), lo que indica que algunas componenes a nivel de ruta muestran una heterogeneidad mayor de lo esperado.

Agrupaciones de individuos en función de sus componentes

A partir de las matrices de mezcla obtenidas en cada una de las factorizaciones se realizó un *clustering* jerárquico con el objetivo de evaluar cómo se agrupan los distintos individuos en cada función biológica analizada. En la mayoría de casos, los resultados obtenidos mostraron una clara tendencia de los individuos a agruparse en función del subtipo al que pertenecen, especialmente notable para los subtipos *Normal* y *Basal*.

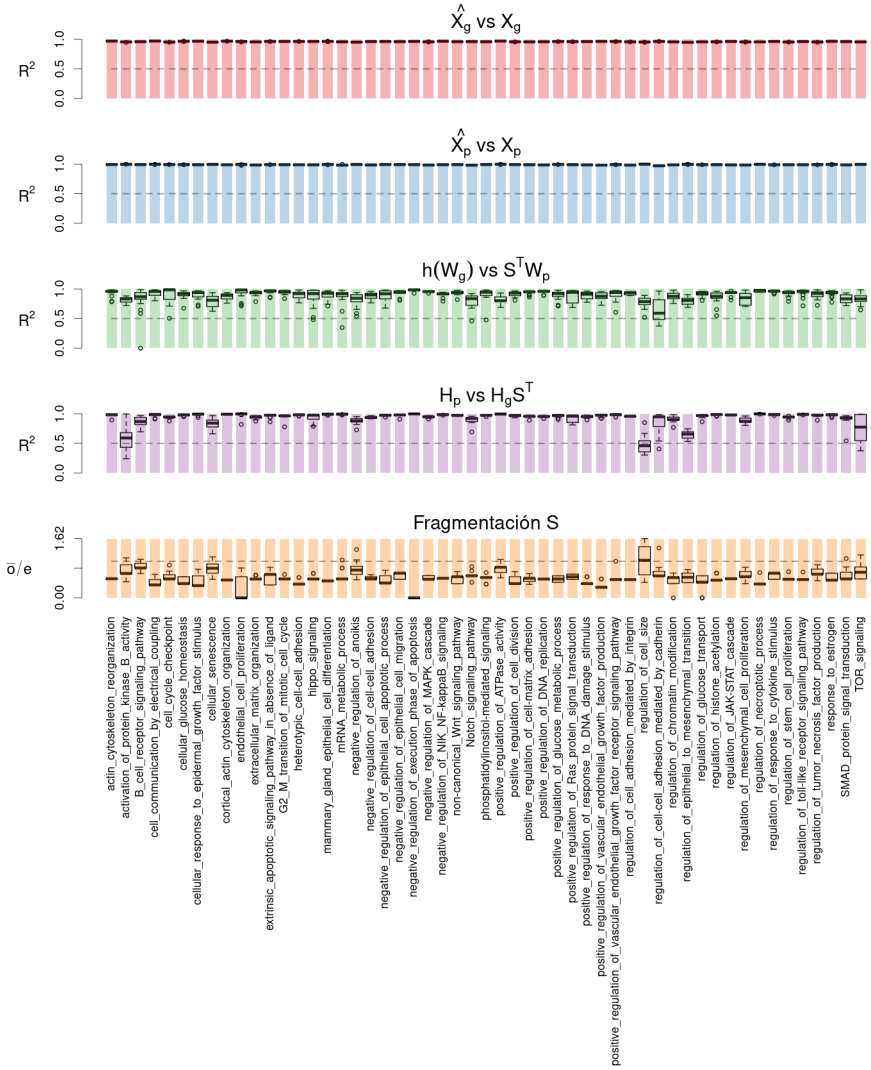


Figura 3.7: *Resultados generales obtenidos para los 53 términos biológicos factorizados. La figura muestra la bondad del ajuste (R^2) obtenida al comparar: (i) las matrices originales frente a las obtenidas por el modelo, (ii) ambos conjuntos de componentes ($W_p S^T \approx \hbar(W_g)$), (iii) sus pesos a lo largo de los individuos ($H_p \approx H_g S^T$), (iv) así como el nivel de fragmentación obtenido para la matriz S al comparar el número esperado de componentes a nivel gen asociadas a la misma componente a nivel de ruta frente al observado.*

Asimismo, la tendencia general se confirmó al evaluar el porcentaje de verdaderos positivos obtenido al comparar el subtipo real de los individuos con las agrupaciones obtenidas mediante un clasificador *Kmeans*, tomando como entrada a las matrices de mezcla, tanto a nivel de gen (Figura 3.8), como a nivel de ruta (Figura 3.9).

Además, los resultados generales permiten confirmar que la precisión obtenida por las matrices de mezcla a la hora de recuperar el subtipo original de los individuos ofrece un rango de valores comparable al obtenido al realizar el mismo proceso con las matrices originales de entrada (Figura 3.10). En particular, en el caso de la matriz de mezcla a nivel de gen, los resultados ofrecen una mejora en los subtipos *Basal* y *Her2*, siendo por el contrario, notablemente peores al clasificar al subtipo *Luminal B*. En el caso de la matriz de mezcla a nivel de ruta, los resultados son notablemente mejores en todos los casos, a excepción, nuevamente, del subtipo *Luminal B*, contando en este caso con menores diferencias con las clasificaciones obtenidas mediante las matrices originales de entrada. Estos resultados demuestran que pese a que el modelo proporciona una reducción considerable en la dimensionalidad de los datos, el conjunto de componentes latentes captura con éxito las características intrínsecas de cada subtipo, y las particularidades que los diferencian del resto.

Prevalencia en los subtipos

Un análisis pormenorizado de los pesos obtenidos en la matriz de mezcla para cada componente encontrada permite determinar si éstas se asocian preferencialmente a uno o más subgrupos de individuos. Esta aproximación tiene especial relevancia a la hora de determinar el rol de cada componente en la enfermedad, ya que el carácter oncogénico de una componente podrá manifestarse mediante un peso significativamente mayor en el grupo de muestras tumorales frente a los individuos normales.

Los resultados obtenidos en la caracterización de los pesos muestran cómo de manera general algunas de las componentes encontradas se asocian específicamente al grupo de sujetos sanos (Figura 3.11), tanto para las componentes a nivel de gen, como a nivel de ruta. De manera similar, se observaron componentes asociadas a

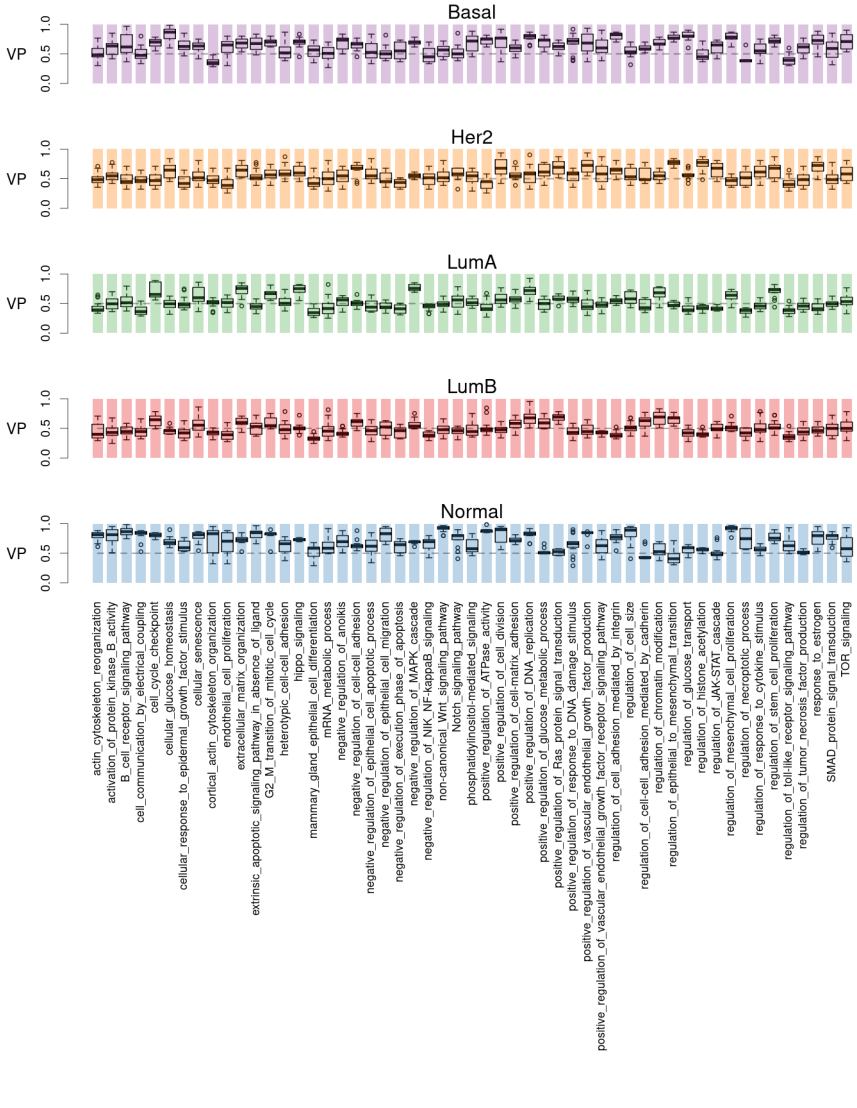


Figura 3.8: *Porcentaje de verdaderos positivos (VP) obtenido al clasificar a los individuos en función de la matriz de mezcla a nivel de gen a lo largo de los 53 términos seleccionados.*

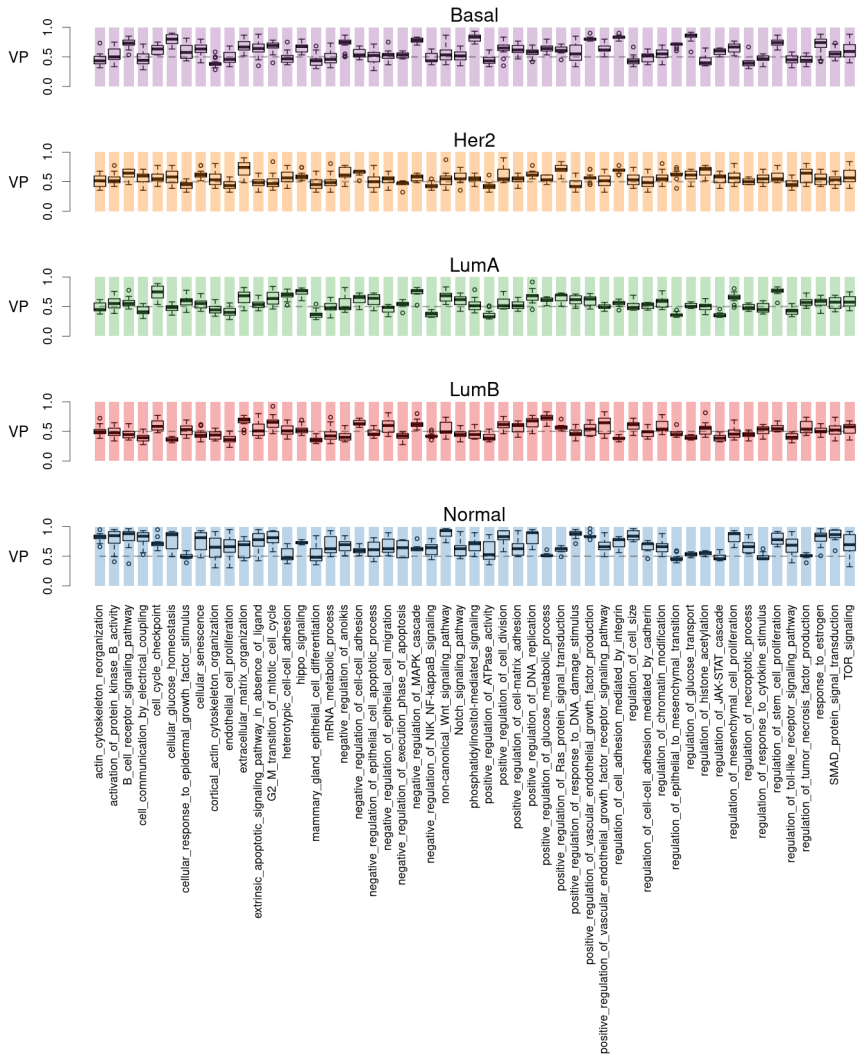


Figura 3.9: *Porcentaje de verdaderos positivos (VP) obtenido al clasificar a los individuos en función de la matriz de mezcla a nivel de ruta a lo largo de los 53 términos seleccionados.*

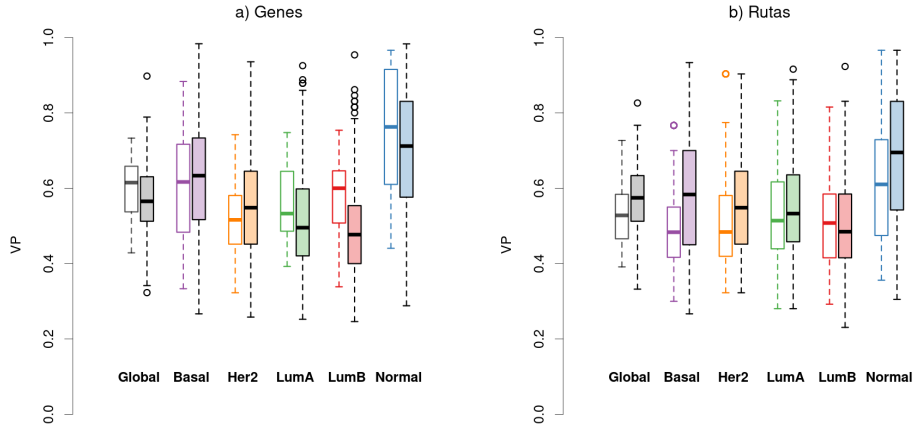


Figura 3.10: *Porcentaje de verdaderos positivos (VP) obtenido al clasificar cada uno de los subtipos mediante las matrices originales de entrada y las matrices de mezcla obtenidas por los modelos propuestos, tanto a nivel de gen, como a nivel de ruta.*

uno o más subtipos de cáncer, teniendo en algunos casos un peso marginal para el conjunto de muestras sanas. De forma complementaria, se observaron componentes con un peso relativamente uniforme a lo largo de todos los individuos, lo que indicaría una asociación a características propias del tejido, común a todos los subgrupos.

Con el objetivo de caracterizar de forma exhaustiva la asociación entre las componentes encontradas y los subtipos presentes en la cohorte, se aplicó un test de comparación de medias entre cada par de subtipos, añadiendo una comparación adicional entre las muestras normales y las muestras tumorales en su conjunto. Dichas comparaciones permitieron derivar una etiqueta que describe la asociación de cada componente a cada subgrupo particular. En este caso, cuando una componente mostró un peso medio significativamente mayor en el grupo de muestras tumorales frente al grupo de muestras normales, dicha componente fue etiquetada como *Tumoral*, siendo etiquetada como *Normal* cuando las diferencias significativas se produjeron en sentido contrario. Además, cuando el peso medio de un subtipo

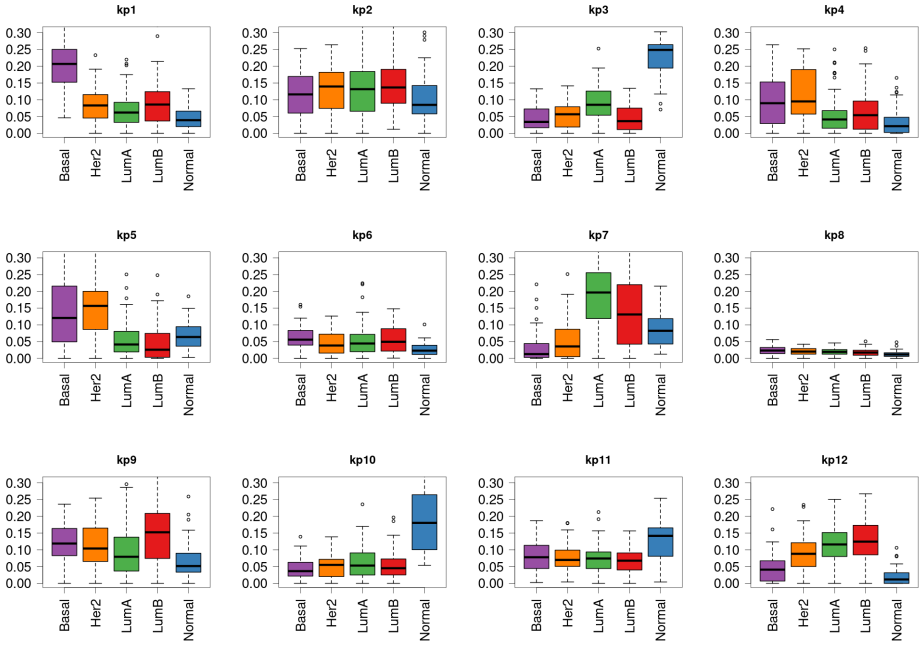


Figura 3.11: *Distribución de valores de la matriz de mezcla a nivel de ruta para la función biológica "cellular glucose homeostasis". Los valores de la matriz aparecen estratificados por subtipo y componente. Tal y como se aprecia, las diferentes componentes muestran asociaciones específicas a uno o más subgrupos de individuos.*

fue significativamente mayor a otro subtipo de la muestra, así como al grupo de muestras normales, dicha componente fue también etiquetada con el subtipo correspondiente.

En este caso, la comparación permitió determinar el número de componentes asociadas a cada subtipo de la cohorte (Figura 3.12) en el conjunto total de funciones. De forma característica, se aprecia que en todas las funciones biológicas analizadas a nivel de gen, todos los subgrupos mostraron al menos una componente asociada. En el caso de las factorizaciones a nivel de ruta, el resultado es similar, a excepción de los términos *mRNA metabolic process* y *endothelial cell migration*, que no mostraron ninguna componente asociada para los subtipos *Basal* y *Her2*, respectivamente. Asimismo, se observó que la mayor parte de componentes encontradas se asociaron al grupo de muestras tumorales, lo que describe una mayor heterogeneidad intrínseca, frente al grupo de individuos normales. Además, se observó también que en la mayoría de funciones biológicas, existen componentes de carácter general sin asociación específica a ningún subgrupo.

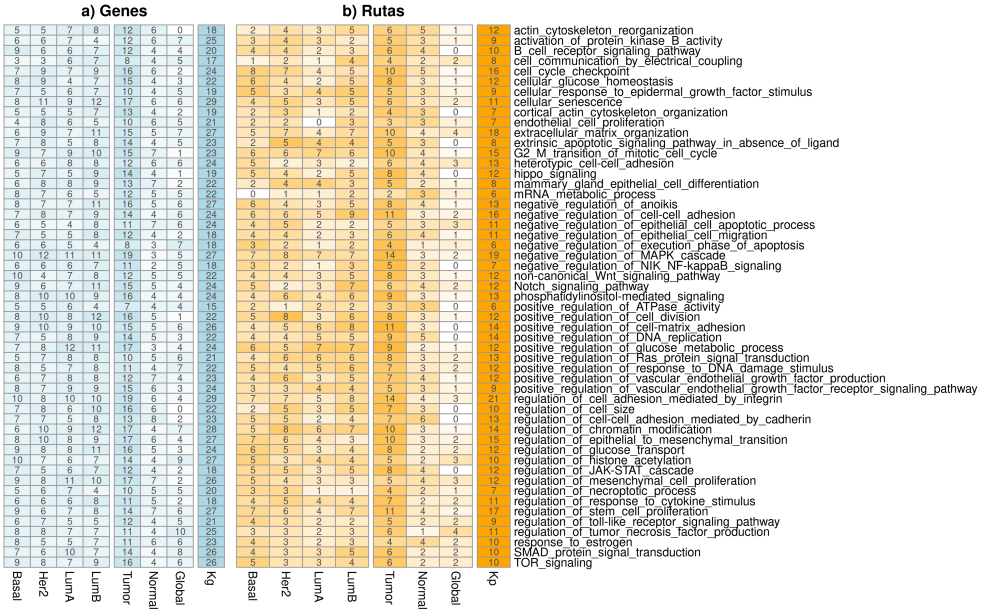


Figura 3.12: *Número de componentes asociadas a cada subtipo en el conjunto de funciones biológicas analizadas, tanto a nivel de ruta, como a nivel de gen.*

Por otro lado, el análisis de prevalencias permitió determinar para cada función biológica el porcentaje de individuos de cada subtipo con una contribución significativa de las componentes encontradas (Figura 3.13). En este caso, los resultados mostraron que el porcentaje asociado a cada subtipo describe una relación coherente con la etiqueta asociada a cada componente.

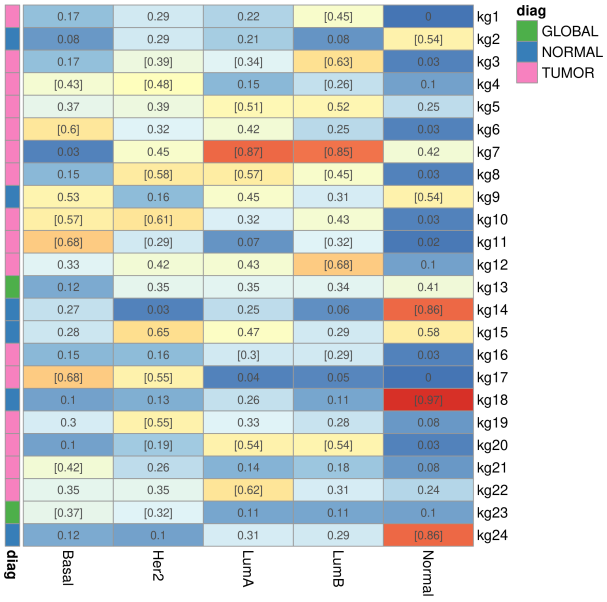


Figura 3.13: *Matriz de prevalencias obtenida para la función biológica "cell cycle checkpoint". La matriz muestra la fracción de individuos por subtipo con una contribución significativa en cada componente. A la izquierda, aparece la caracterización de cada componente en función de las prevalencias obtenidas.*

Por último, con el objetivo de evaluar el peso de cada componente en el modelo, se emplearon dos estrategias. En primer lugar se empleó un predictor de tipo *Random Forest* para determinar si el peso de cada componente tuvo relevancia a la hora de predecir el subtipo. Asimismo, se acumuló el peso de cada componente a lo largo de los individuos, con el objetivo de evaluar el peso global de cada componente en el modelo. Los resultados obtenidos confirmaron que la mayoría de componentes encontradas tuvieron tanto un peso relevante en la predicción del subtipo de (Figura 3.14), como un peso acumulado no despreciable en la factorización. De forma complementaria, se comprobó la existencia de un porcentaje pequeño de

componentes con bajo poder discriminatorio y con un peso acumulado muy bajo, probablemente encargadas de modelar características técnicas ajenas al problema biológico.

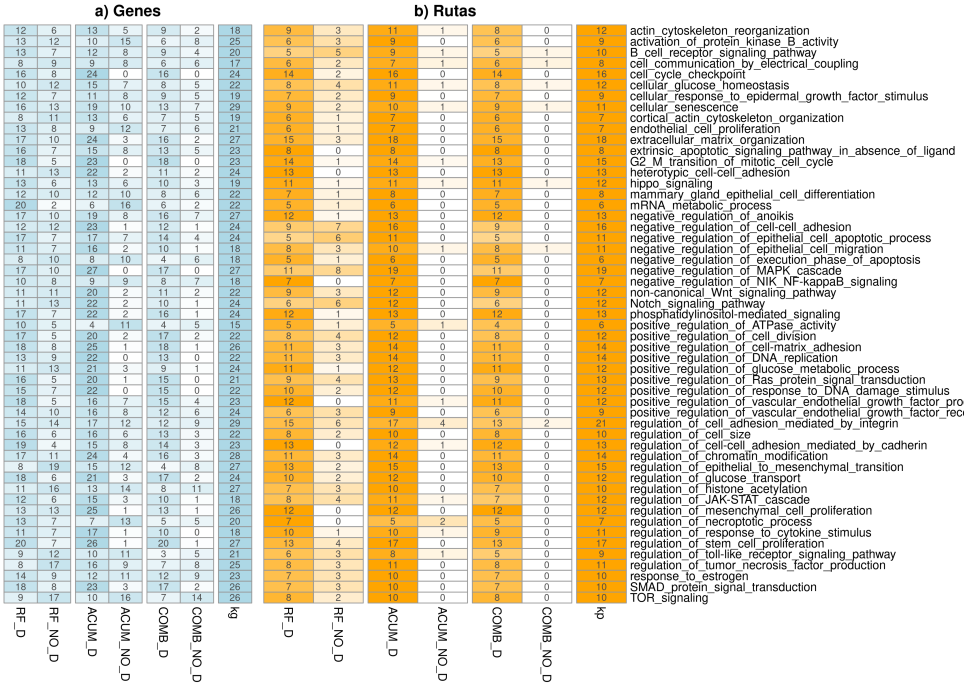


Figura 3.14: *Estimación de componentes relevantes obtenida en cada una de las funciones biológicas, tanto a nivel de ruta, como a nivel de gen. RF_D/RF_NO_D: Número de componentes relevantes (y no relevantes) a la hora de predecir el subtipo con un clasificador de tipo Random Forest. ACUM_D/ACUM_NO_D: Número de componentes relevantes (y no relevantes) obtenido al acumular el peso de cada componente a lo largo de los individuos. COMB_D/COMB_NO_D: Número de componentes relevantes (y no relevantes) combinando ambas estrategias.*

Análisis de supervivencia

Además del análisis de subtipos, los pesos obtenidos en las matrices de mezcla para cada componente permitieron determinar su relación con la probabilidad de supervivencia de los individuos enfermos. Para ello, el peso de cada componente fue estratificado en 3 niveles (alto, normal y bajo) y después empleado para realizar el correspondiente análisis de supervivencia. Los resultados de este análisis

permitieron comprobar la existencia de algunas componentes cuyo peso se asoció significativamente a la probabilidad de supervivencia (Figura 3.15). En concreto, se encontraron un total de 34 componentes significativas a nivel de gen, ofreciendo una media de 0.64 componentes significativas por factorización. En el caso de las componentes a nivel de ruta, se encontraron un total de 28 componentes significativas, con una media de 0.52 componentes por factorización. De forma característica, las componentes significativas encontradas, tanto a nivel de gen, como a nivel de ruta, se concentraron en menos de la mitad de las funciones biológicas, lo que podría poner de manifiesto la existencia de rutas biológicas con mayor relevancia en la supervivencia de los individuos.

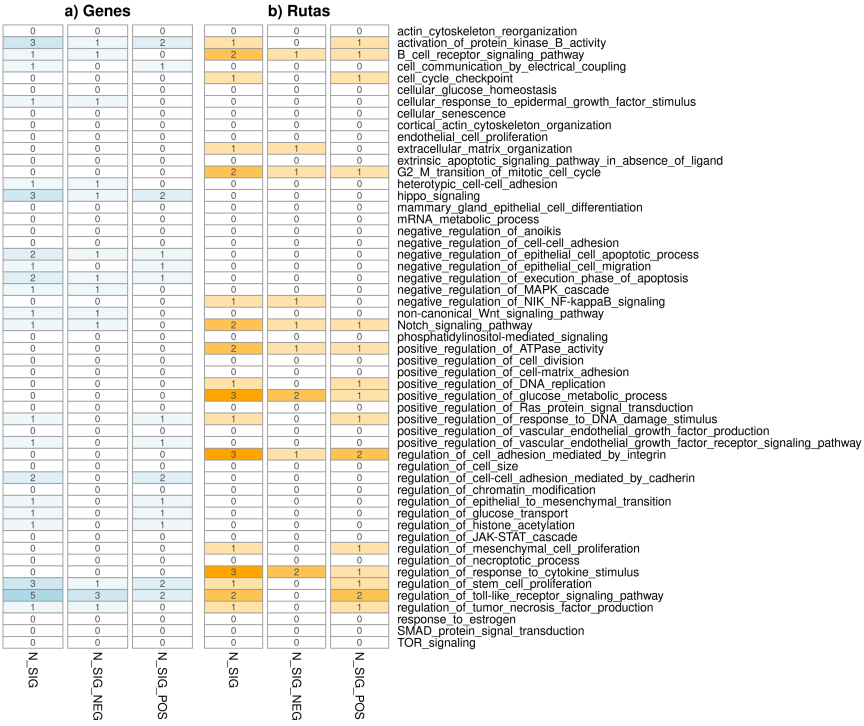


Figura 3.15: *Número de componentes con una relación significativa con la supervivencia de los individuos para el conjunto de funciones biológicas analizadas, tanto a nivel de gen, como a nivel de ruta.*

Este hecho refleja la influencia de cada función biológica en la supervivencia de los individuos. Además, de forma característica, se encontraron tanto componentes

cuya contribución se asoció a una mayor mortalidad, como componentes que se asociaron a individuos con mayor probabilidad de supervivencia, lo que describe a trayectorias menos agresivas en la enfermedad.

Sinergia entre componentes

El análisis de sinergias permitió determinar la existencia de multitud de parejas de componentes con un solapamiento significativo en los mismos individuos. Este patrón, observado tanto a nivel de gen, como a nivel de ruta, se repite además de forma habitual en la práctica totalidad de funciones biológicas analizadas (Figura 3.16).

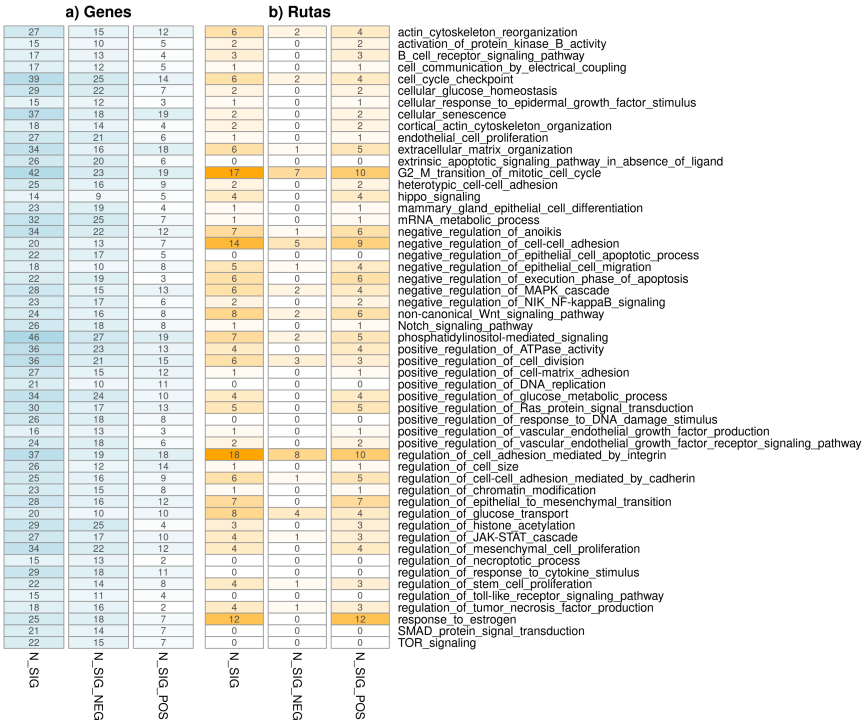


Figura 3.16: *Sinergias significativas obtenidas entre parejas de componentes para el conjunto total de funciones biológicas analizadas, tanto a nivel de gen, como a nivel de ruta. Se describe el número total de sinergias (N_SIG), así como el número de sinergias positivas (N_SIG_POS) y negativas (N_SIG_NEG).*

Asimismo, se hizo uso de las redes de sinergia para representar la existencia

de grupos más amplios de componentes con una participación conjunta (Figura 3.17). Este resultado permite comprender como las distintas componentes latentes de una función biológica se combinan para construir las observaciones, y regular las funciones biológicas en las que participan.

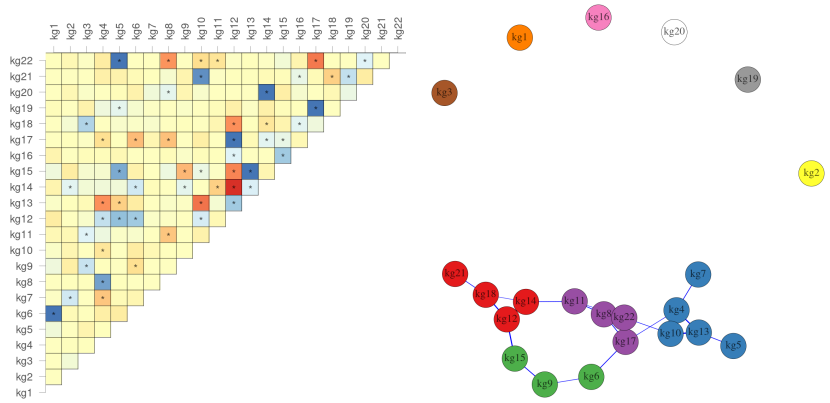


Figura 3.17: *Matriz de sinergias obtenidas para la función biológica "cellular senescence" (las sinergias significativas aparecen resaltadas con el símbolo *). Además, se describe la red de sinergias positivas, marcando en diferentes colores las distintas agrupaciones de componentes. Asimismo, aparecen componentes individuales que no han mostrado sinergias significativas con otras componentes.*

De forma complementaria, se observó un gran número de sinergias de tipo negativo, tanto a nivel de gen, como a nivel de ruta. En este caso, se trata de parejas de componentes que rara vez ocurren en los mismos individuos, mostrando por tanto un caracter excluyente. Este resultado permite comprobar la existencia de caminos distintos para la enfermedad, cuya descripción representa un gran valor a nivel clínico.

Nodos relevantes en las componentes

El análisis realizado en cada función biológica permitió determinar qué genes fueron los más relevantes dentro de cada componente encontrada. En primer lugar, el análisis permitió determinar que en promedio aproximadamente un 40 % de los genes fueron relevantes en al menos una componente del modelo (Figura 3.18). De

forma complementearia, aproximadamente un 10 % de genes fueron considerados relevantes por mostrar valores muy pequeños, compatibles con una inhibición selectiva.

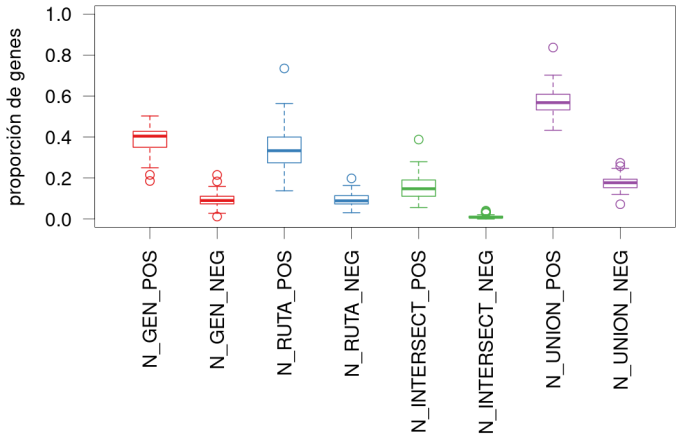


Figura 3.18: *Proporción de genes relevantes en el conjunto total de componentes encontradas a nivel de gen. Se describe el número de genes relevantes de carácter positivo obtenido en el espacio de los genes (N_GEN_POS), en el espacio de las rutas (N_RUTA_POS) y combinando ambos criterios mediante la intersección (N_INTERSECT_POS) y la unión (N_UNION_POS). Asimismo, se recojen los mismos valores para los genes relevantes de carácter negativo.*

Los resultados obtenidos mostraron que en promedio un 30 % de genes fueron considerados relevantes a través de su evaluación en el espacio de las rutas, al proporcionar valores particularmente altos en al menos una cascada de señalización. De forma complementaria, aproximadamente un 10 % de genes mostraron valores particularmente bajos en el espacio de las rutas, compatibles con la inhibición de una o más cascadas de señalización

De forma característica, los resultados obtenidos mostraron un solapamiento

parcial entre ambos criterios empleados para determinar la relevancia de los genes, visible además a lo largo de todas las factorizaciones (Figura 3.19). Este resultado confirmó la existencia de genes cuya variabilidad no produjo el efecto esperado en el espacio de las rutas.

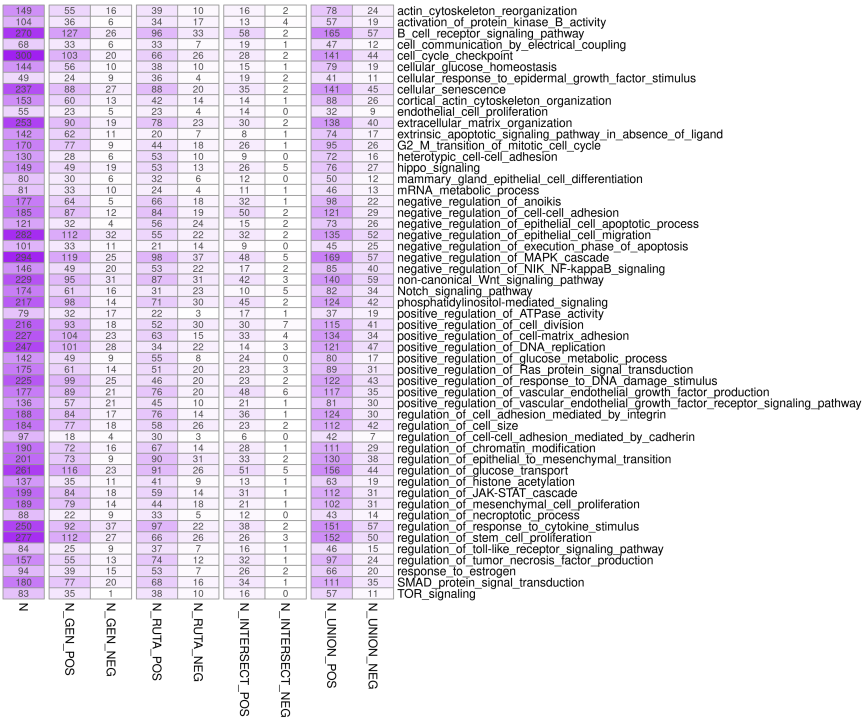


Figura 3.19: *Número de genes relevantes obtenidos en cada una de las funciones biológicas analizadas.*

Con el objetivo de evaluar la importancia de los genes seleccionados durante el análisis, se analizó el grado de solapamiento entre el conjunto de genes relevantes y varios conjuntos de genes con implicaciones previas en la enfermedad (Figura 3.20). En particular, se evaluó mediante un test exacto de *Fisher* el solapamiento con los conjuntos de genes correspondientes a los clasificadores PAM50, MammaPrint y Oncotype. Además, se incluyeron los genes correspondientes al censo de genes de cáncer del repositorio COSMIC, en sus tres categorias existentes (*oncogenes*, *tumor supressors*, y *fusion genes*). En este caso, los resultados mostraron un solapamiento significativo con aquellos genes considerados relevantes por ofrecer valor extremos en

el espacio de las rutas. Sin embargo, los genes considerados relevantes por mostrar valores extremos en el espacio de los genes, no proporcionaron el solapamiento significativo esperado. Este resultado sugiere que el uso del efecto de la variación en el espacio de las rutas podría ser un criterio más fiable a la hora de seleccionar los genes más relevantes dentro de una componente.

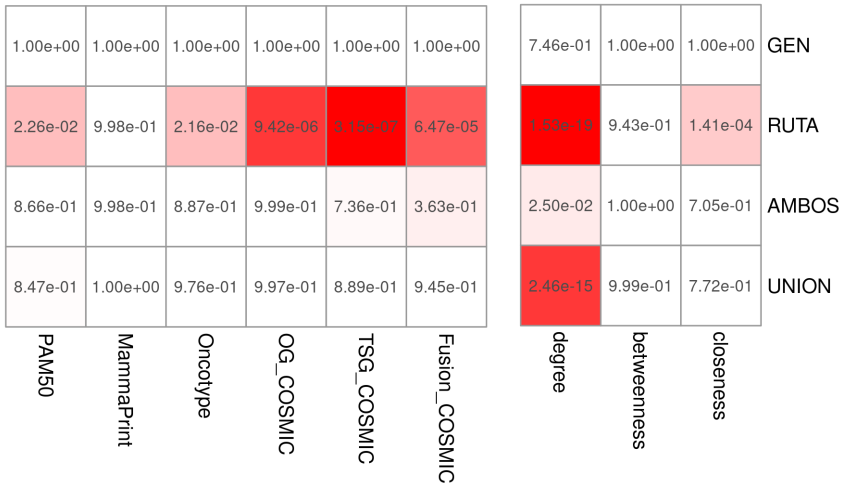


Figura 3.20: *Solapamiento obtenido entre diferentes conjuntos de genes con implicaciones previas en la enfermedad y los genes relevantes obtenidos el espacio de las rutas, el espacio de los genes y combinando ambos criterios. Asimismo, se incluye la comparación entre genes relevantes y no relevantes, realizada a partir de tres parámetros de red.*

Este último resultado fue también confirmado a la hora de evaluar tres parámetros típicos que miden la importancia de un nodo dentro de una red. En este caso, únicamente se encontraron diferencias significativas entre los genes relevantes en el espacio de las rutas y los no relevantes, al comparar los valores obtenidos para los parámetros *degree* y *closeness* mediante un test de comparación de medias no paramétrico (*Mann Whitney*) de una cola.

Análisis de genes driver

De forma complementaria al análisis anterior, se evaluó el nivel de actividad

de los genes relevantes en cada componente encontrada en función de su pertenencia a las categorías descritas en *COSMIC*, añadiendo una categoría adicional correspondiente a aquellos que no han sido previamente descritos en el cáncer. En este caso, se observaron patrones muy distintos entre las componentes (Figura 3.21), describiendo perfiles donde los oncogenes muestran una actividad mayor, o de forma contraria, donde éstos tienen un nivel de actividad similar al resto de genes.

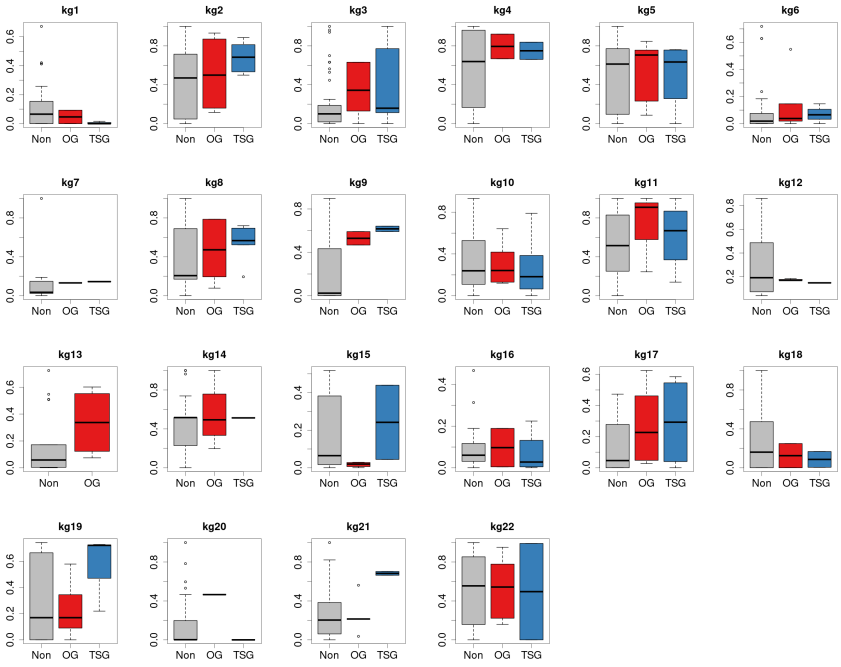


Figura 3.21: *Distribución de valores de actividad de los genes relevantes para la función biológica "positive regulation of cell division". Los valores aparecen estratificados por componente y rol conocido.*

Además, se evaluó el nivel de actividad de cada tipo de gen en función de la etiqueta previamente definida para cada componente, describiendo si éstas tienen mayoritariamente un peso más importante en las muestras normales, en las tumorales, o de forma global en todo el conjunto (Figura 3.22). En este caso, se observó como los genes no descritos previamente en el cáncer tiene un nivel de

actividad mayor en las componentes asociadas a las muestras normales, a diferencia de los oncogenes cuya actividad se incrementa en las componentes asociadas a las muestras tumorales. Además, los supresores de tumor, aun teniendo un rango de actividad más bajo, también mostraron un mayor nivel de actividad en las componentes de tipo normal.

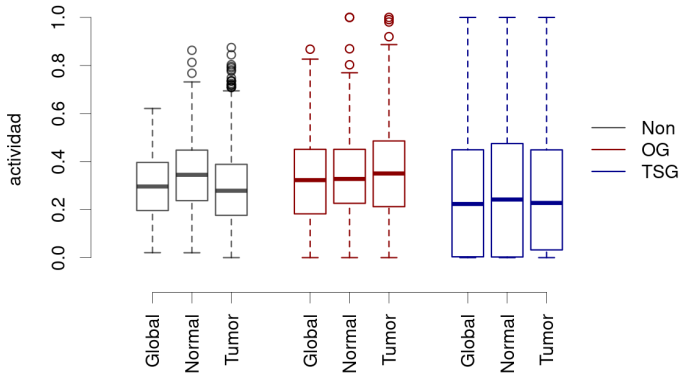


Figura 3.22: *Distribución de la actividad de los genes en función del tipo de componente y del tipo de descripción obtenida para los genes en el repositorio COSMIC.*

Meta-análisis

El análisis realizado en cada una de las factorizaciones ha permitido describir las estrategias que se encuentran de forma latente en los pacientes para desarrollar los *hallmarks* de la enfermedad. Aunque esta aproximación resulta muy valiosa a la hora de entender cómo se regula cada una de las funciones biológicas alteradas, no proporciona una visión global sobre las diferencias entre los distintos subtipos moleculares presentes en el conjunto de individuos, ya que, tal y como se ha descrito, para una función particular no es estrictamente necesario observar un comportamiento distinto entre todos los subgrupos.

Con este objetivo, se realizó un meta-análisis destinado a cuantificar de forma global las diferencias o distancias entre cada uno de los individuos, en el contexto

del subgrupo al que pertenecen. Para ello, se combinaron las matrices de distancia obtenidas en cada una de las factorizaciones, con el fin de construir una matriz consenso que describe la distancia global entre los individuos. En este caso, se emplearon dos métricas distintas a la hora de medir la distancia entre cada par de individuos: distancia euclidea y correlación lineal. La Figura 3.23 muestra los resultados obtenidos a partir de cada una de las matrices obtenidas, incluyendo también como referencia al resultado obtenido con las matrices de distancia calculadas a partir las matrices originales de entrada (X_g y X_p).

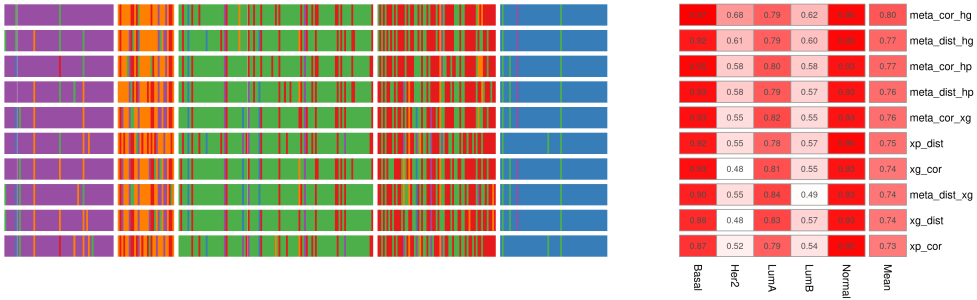


Figura 3.23: *Predicción del subtipo de los individuos obtenida mediante las matrices de distancia euclídea y correlación lineal, las matrices de entrada globales (xg_dist, xg_cor, xp_dist, xp_cor) y la combinación de las matrices de mezcla de cada una de las funciones biológicas analizadas (meta_dist_hg, meta_cor_hg, meta_dist_hp, meta_cor_hp). A la izquierda se muestran los individuos agrupados en función de su subtipo original y coloreados en función del subtipo obtenido en la predicción. A la derecha, se describe el porcentaje de verdaderos positivos obtenido en cada alternativa. Las diferentes estrategias aparecen ordenadas en función de su capacidad de predicción.*

De forma característica, los agrupamientos basados en el consenso de las matrices de mezcla proporcionaron una clasificación más precisa en comparación con las matrices originales de entrada. Este resultado, permite confirmar que la integración de los modelos jerárquicos obtenidos en cada función biológica proporciona un entendimiento más preciso sobre los distintos subtipos moleculares presentes en la muestra.

Descripción del sistema

La aplicación del modelo jerárquico ha proporcionado una vista detallada sobre la composición interna de cada individuo a la hora de regular aquellas funciones biológicas alteradas en las muestras tumorales. En particular, el modelo jerárquico ha descrito como las componentes más relevantes a nivel de gen se combinan para construir las componentes a nivel de ruta, y cómo éstas, a su vez, regulan cada función biológica bajo estudio.

Esta aproximación permite en la práctica realizar un estudio detallado sobre la variabilidad biológica observada en un grupo de pacientes a la hora de regular las funciones biológicas que han mostrado un grado de alteración importante. En particular, la aproximación proporciona a nivel práctico una vista comprensible sobre el grado de heterogeneidad observado en un determinado subgrupo de individuos, describiendo de forma clara las diferentes soluciones encontradas para el grupo de sujetos.

La Figura 3.24 muestra de forma gráfica el resultado obtenido tras la aplicación del modelo jerárquico. En este caso, la gráfica describe el subconjunto de componentes a nivel de ruta (W_p) con una contribución relevante en el subtipo seleccionado, acompañadas de la matriz de mezcla (H_p) binarizada, con el fin de determinar de forma más clara la relevancia de cada componente en cada individuo analizado. De forma complementaria, se incluyen las componentes asociadas a nivel de gen (W_g) que describen las posibles estrategias observadas para producir una misma respuesta a nivel de ruta, acompañadas también de su correspondiente matriz de mezcla (H_g) binarizada.

Además, el modelo jerárquico describe las diferentes combinaciones de las componentes latentes observadas en los individuos seleccionados, proporcionando también su frecuencia de aparición, con el fin de determinar cuales son las combinaciones más relevantes dentro del subtipo y cuales representan casos más aislados. De forma implícita, el modelo jerárquico describe cuales son los genes más relevantes en cada componente seleccionada, indicando además si su actividad estimula o

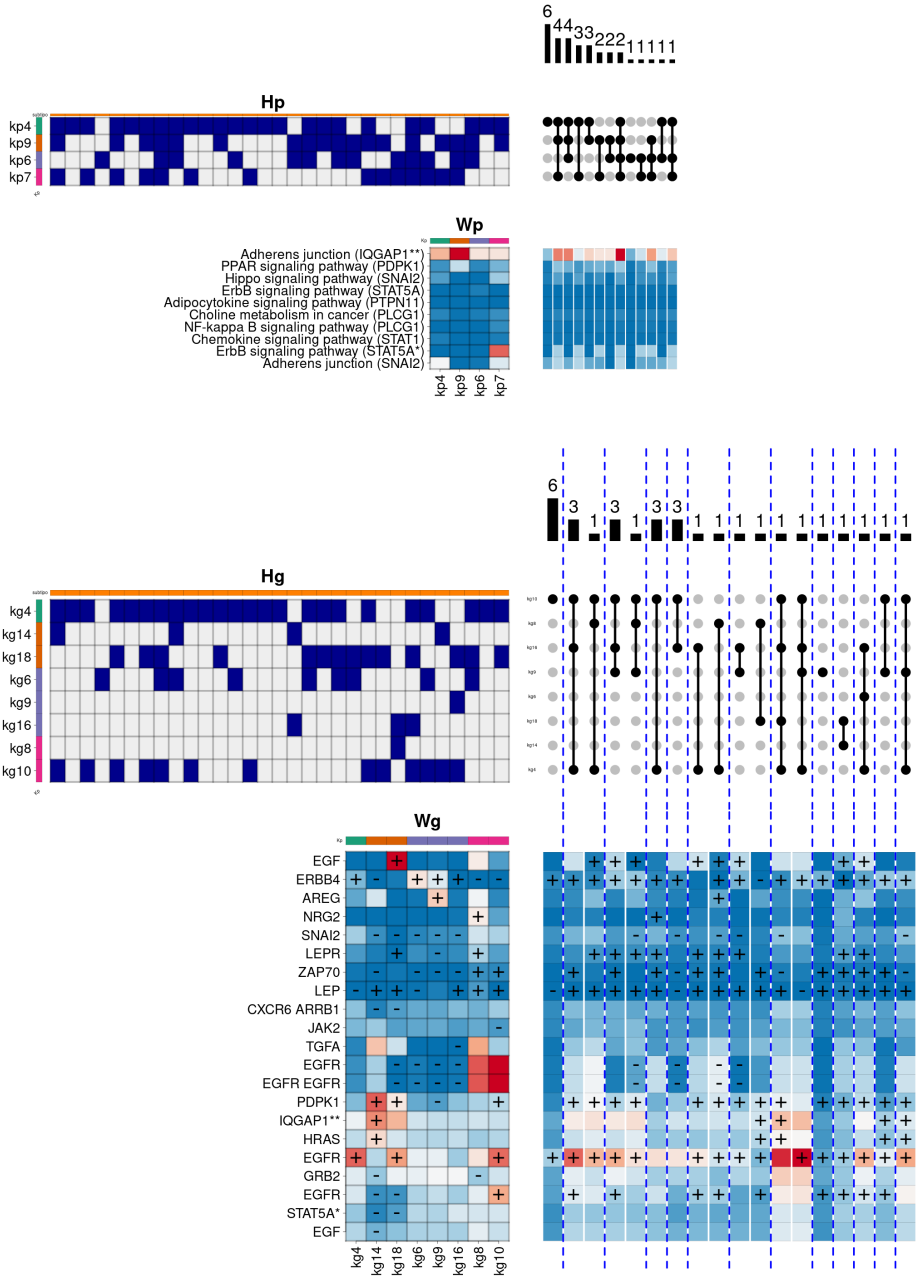


Figura 3.24: *Representación gráfica del modelo jerárquico obtenida en la función biológica "cellular response to epidermal growth factor stimulus", aplicada al subtipo Her2.*

inhibe la actividad de la componente en el espacio de las rutas. Además, dado que las soluciones se construyen combinando las componentes, es posible propagar la relevancia de cada gen en cada componente para determinar a su vez, la relevancia de cada gen en cada solución encontrada, lo que proporciona una vista aún clara de las diferentes estrategias.

La figura describe también la composición interna del subtipo *Her2* en la función biológica *cellular response to epidermal growth factor stimulus*. Este subtipo se caracteriza por tener una actividad muy reducida en los receptores hormonales de estrógenos y progesterona, y por mostrar un alto grado de sobreactivación en el gen *ERBB2*. Este gen pertenece a la familia de receptores de membrana del factor de crecimiento epidérmico, encargado de regular, entre otras cosas, la proliferación celular, contribuyendo de forma relevante a que el subtipo *Her2* muestre una mayor mortalidad que otros subtipos. La combinación de componentes a nivel de ruta más frecuente consiste en el uso exclusivo de la componente *kp4* (coloreada en verde), que se compone principalmente de la sobreactivación de dos cascadas de señalización procedentes de la ruta *Adherens junction*. En el espacio de los genes, la componente *kp4* se asocia con una única solución representada por la componente a nivel de gen *kg4*, que muestra como nodos relevantes los genes *ERBB4* y *EGFR*, pertenecientes también a la familia de receptores del factor de crecimiento epidérmico. Asimismo, se observan dos combinaciones secundarias que añaden a la componente anterior la componente *kp9* (coloreada en naranja), encargada de contribuir también a la sobreactivación de las dos cascadas de la ruta *Adherens junction* y de activar parcialmente la ruta *PPAR signaling pathway*. La componente *kp9* se representa en el espacio de los genes mediante dos soluciones distintas, correspondientes a las componentes *kg14*, que sobreactiva simultáneamente a los genes *PDPK1* (*Phosphoinositide-dependent kinase 1*, involucrado en la respuesta a factores de crecimiento e insulina), *IQGAP1* (*Phosphoinositide-dependent kinase 1*, involucrado en la regulación del ciclo celular) y *HRAS* (*HRas Proto-Oncogene, GTPase*, perteneciente a la familia *RAS* de oncogenes), y la componente *kg18*, que sobreactiva nuevamente a los genes *PDPK1* y *EGFR*, y añade al propio factor de crecimiento epidérmico *EGF* (*epidermal growth factor*). Estas dos componentes

adicionales, completan la combinación mediante la incorporación de las componentes a nivel de ruta *kp7* (coloreada en rosa), encargada de incrementar la actividad de la ruta *ErbB signaling pathway*, y la componente *kp6* (coloreada en magenta) encargada de sobreactivar de nuevo una de las dos cascadas de la ruta *Adherens junction*. Dichas componentes se representan en el espacio de los genes mediante las componentes *kg6/8/16* y *kg8/10* respectivamente, contribuyendo a la sobreactivación de otros genes relevantes como *AREG* (*Amphiregulin*, también perteneciente a la familia de receptores del factor de crecimiento epidérmico) o *NREG2* (*Neuregulin 2*, involucrado en el crecimiento y diferenciación de las células epiteliales). De esta forma, la composición proporciona un mapa detallado sobre los elementos más importantes a nivel de gen y a nivel de ruta, y de cómo éstos se relacionan para construir las combinaciones más frecuentes dentro del subtipo.

La misma aproximación ha sido aplicada para entender la heterogeneidad observada en los pacientes pertenecientes al subtipo *Basal*. Los tumores de tipo *Basal* se caracterizan por ser habitualmente negativos tanto en los receptores de hormona, como en el receptor *ERBB2*. Esta circunstancia impide el uso de fármacos habituales en el tratamiento de la enfermedad, mostrando además un mayor crecimiento tumoral, mayor agresividad y una mortalidad superior al resto de subtipos.

Una de las vias de señalización más alteradas en el subtipo *Basal* lo constituye la ruta *Notch signaling pathway*. Se trata de una ruta muy conservada en mamíferos, con un rol esencial en el desarrollo embrionario y la regulación del destino celular en las glándulas mamarias a lo largo de diferentes estadios de su desarrollo. La Figura 3.25 muestra la composición interna del subtipo *Basal* en dicha función biológica.

La vista ofrecida por el modelo jeráquico ofrece una descripción mas variada que la obtenida en el análisis anterior con el subtipo *Her2*, seguramente producido por el mayor tamaño muestral del grupo *Basal*. En este caso, se observan 3 combinaciones principales representando a la mayor parte de individuos, donde la activación de las rutas *Wnt* y *Notch* aparece con gran intensidad. En este caso, dichas combinaciones

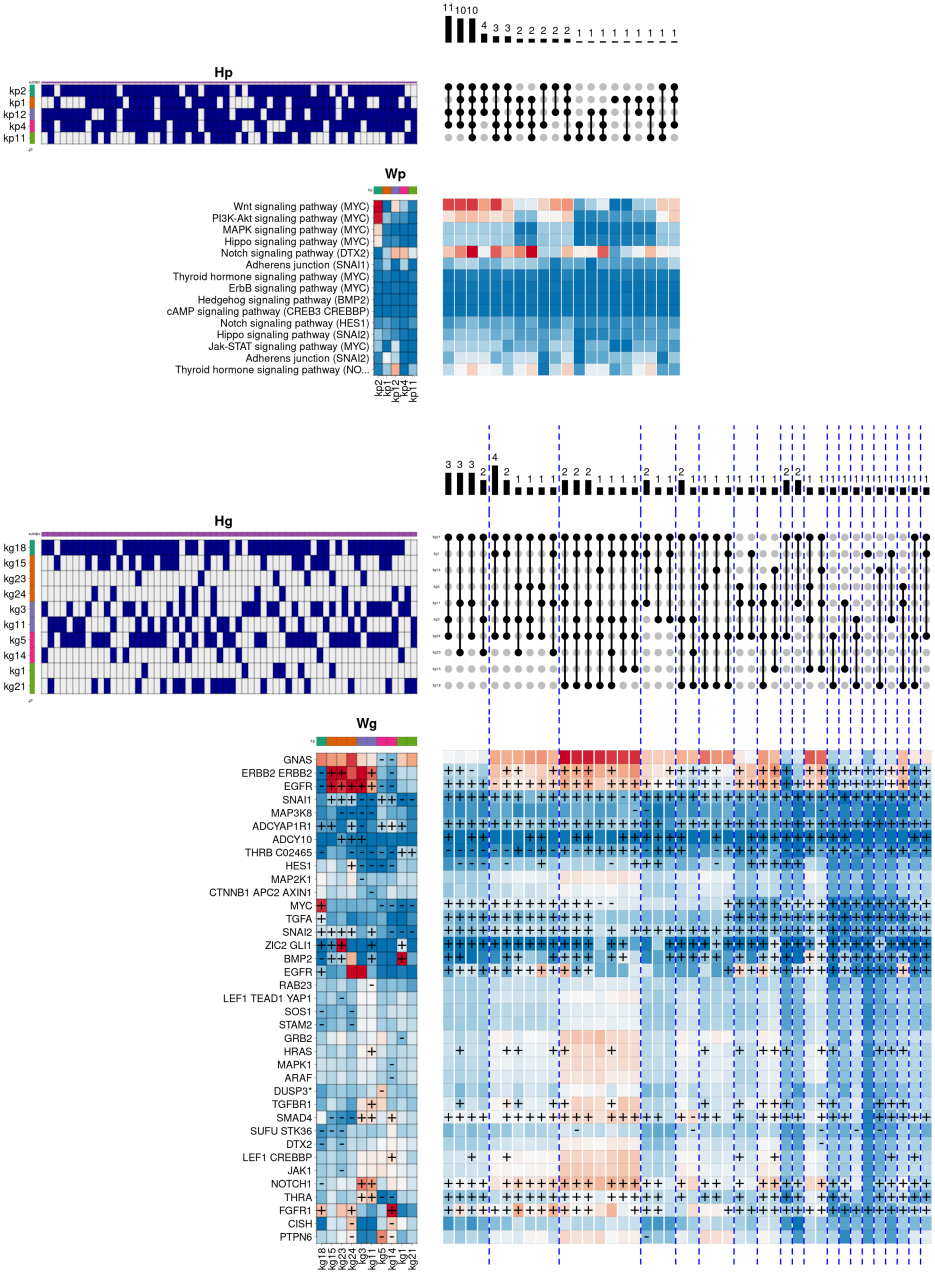


Figura 3.25: *Representación gráfica del modelo jerárquico obtenida en la función biológica "Notch signalling", aplicada al subtipo Basal.*

incluyen de manera común a las componentes *kp2* (verde oscuro), *kp12* (magenta) y *kp4* (rosa). La primera de las componentes (*kp2*) describe la sobreactivación de las cascadas de señalización procedentes de las rutas de señalización *Wnt*, *PI3K-Akt*, *MAPK* e *Hippo*, habitualmente alteradas en diferentes tipos de cáncer. Esta componente se asocia en el espacio de los genes con la componente *kg18*, la cual sobreactiva de forma simultánea a los genes *MYC* (*MYC Proto-Oncogene*, oncogen implicado en la regulación del ciclo celular y la apoptosis), *TGFA* (*Transforming Growth Factor Alpha*, un ligando natural del receptor del factor de crecimiento epidérmico), *SNAI2* (*Snail Family Transcriptional Repressor 2*, un represor natural del complejo e-caderina, habitualmente inhibido en carcinomas), *EGFR* (receptor del factor de crecimiento epidérmico) y *FGFR1* (*Fibroblast Growth Factor Receptor 1*, previamente involucrado en el crecimiento tumoral y la metástasis). Por otro lado, la componente *kp12* contribuye a la combinación activando la ruta de señalización *Notch*, de gran relevancia en el subtipo, y la ruta de señalización que describe la recepción de la hormona tiroides. Esta componente se asocia en el espacio de los genes con las componentes *kg3/11*. Dichas componentes, además de sobreactivar los receptores *EGFR* y *ERBB2*, contribuyen a la composición sobreactivando los genes *TGFBR1* (*Transforming Growth Factor Beta Receptor 1*, otro de los receptores del factor de crecimiento transformante beta), *SMAD4* (*SMAD Family Member 4*, activadas por el factor de crecimiento transformante beta), *THRA* (*Thyroid Hormone Receptor Alpha*, receptor nuclear de la hormona de tiroides) y, de forma notable, el gen *NOTCH1* (*Notch Receptor 1*, implicado en la regulación del destino celular, diferenciación y proliferación). De forma complementaria, la componente *kp4* contribuye reforzando la activación de la ruta *Notch* e incluyendo la ruta *Adherens junction*. Esta componente se asocia en el espacio de los genes con las componentes *kg5/14*, las cuales, además de reforzar la activación de los genes *SMAD4* y *FGFR1*, contribuyen sobreactivando el gen *SNAI1* (*Snail Family Transcriptional Repressor 1*, también represor del complejo e-caderina, e implicado en la transición epitelial-mesenquimal).

Además de las 3 componentes principales, las dos soluciones secundarias implican a las componentes *kp1* (naranja) y *kp11* (verde claro). En este caso, la componente

kp1 contribuye a reforzar la activación de las rutas *PI3K-Akt*, *Notch* y *Adherence junction*. En particular, esta componente se asocia en el espacio de los genes con las componentes *kg15/23/24*, las cuales sobreactivan de forma clara los receptores *ERBB2* y *EGFR* previamente descritos. Además, también sobreactivan los genes *SNAI1*, *SNAI2* y *BMP2* (*Bone Morphogenetic Protein 2*, un ligando natural del factor de crecimiento transformante beta, implicado en la regulación de la proliferación celular, la diferenciación y la respuesta inmunitaria). Por último, la componente *kp11*, la cual permite reforzar de nuevo la recepción de la hormona de tiroides y la activación de la ruta *Notch*, se asocia con las componentes a nivel de gen *kg1/21*, encargadas de sobreactivar los genes *THRB* (*Thyroid Hormone Receptor Beta*, también receptor nuclear de la hormona de tiroides), *BMP2* y *ZIC2* (*Zic Family Member 2*, represor natural del receptor de dopamina).

A diferencia de los subtipos *Basal* y *Her2*, los subtipos *Luminal A* y *B* se caracterizan por ser positivos en los marcadores correspondientes a los receptor de hormonas, especialmente el de estrógenos. Se trata de los subtipos más habituales en el conjunto de pacientes con cáncer de mama, y muestran un mejor pronóstico que el resto, en parte debido a la variedad de fármacos disponibles para modular el efecto de los receptores de hormona.

La Figura 3.26 muestra el resultado obtenido por el modelo jerárquico en los dos subtipos luminales para la función biológica *Response to estrogen*. Además de la información mostrada anteriormente, en este caso también el gráfico describe las componentes que son específicas del subtipo *Luminal A* (en verde), las que son específicas del subtipo *Luminal B* (en rojo) y las que son comunes a ambos grupos (en negro). De forma complementaria, también el nombre de los genes en la matriz de componentes muestra el color del subtipo al que se asocia. En este caso, un determinado gen es específico de un subgrupo si únicamente se muestra como gen relevante en aquellas componentes que son también específicas de dicho subgrupo. Esta descripción permite fácilmente determinar las diferencias y similitudes entre los dos subtipos, mostrando sus genes específicos más relevantes, las componentes en las que participan y combinaciones que forman.

En este caso, se observa que ambos subtipos comparten gran parte de las componentes. Sin embargo, en el caso de las combinaciones, se aprecia que mayoritariamente son específicas de un subtipo *Luminal A*, probablemente debido a disponer de un mayor tamaño muestral.

En el gráfico se observa como la combinación más frecuente a nivel de ruta engloba a ambos subtipos, empleando a las componentes *kp3* y *kp10*. La primera de las dos componentes muestra una activación razonable para las cascadas relacionadas con la respuesta celular a la hormona tiroides y las rutas *Adherens junction* y *Tight junction*, relacionadas con la interacción física que se produce entre células adyacentes. En este caso, la componente *kp3* se relaciona en el espacio de los genes con las componentes *kg5/11/19*. Dichas componentes tienen como genes relevantes a los receptores del factor de crecimiento epidérmico *ERBB4* y *AREG*, y de forma notable, al gen *ESR1* que codifica la isoforma principal del receptor de estrógenos. La segunda componente a nivel de ruta (*kp10*) contribuye a la activación de las mismas rutas que la componente *kp3*, aunque añadiendo en este caso también la ruta *ErbB signalling pathway*, implicada en la regulación de diversas funciones biológicas esenciales. Las componentes a nivel de gen a las que se asocia son *kg8/20/21* y contribuyen nuevamente a la activación de los genes *ERBB4* y *ESR1*. Además, una de las componentes (*kg8*) activa de forma específica a los genes *EREG* (*Epiregulin*, un ligando específico de los receptores del factor de crecimiento epidérmico) y *GPER1* (*G Protein-Coupled Estrogen Receptor 1*, cuya proteína se localiza en el retículo endoplasmático y se une preferencialemente a los estrógenos). Asimismo, también los genes *HBEGF* (*Heparin Binding EGF Like Growth Factor*, un parálogo importante del gen *AREG*) y *MMP2* (*Metallopeptidase*, perteneciente a una familia enzimas encargadas de degradar algunos componentes de la matriz extracelular) son activados en el mismo grupo.

La segunda combinación más frecuente a nivel de ruta se asocia específicamente al subtipo *Luminal A* y añade la componente *kp4* a las dos anteriores, activando de forma notable una cascada distinta de la ruta *Adherens junction* y otra via adicional de la respuesta a la hormona tiroides. En este caso, la componente se

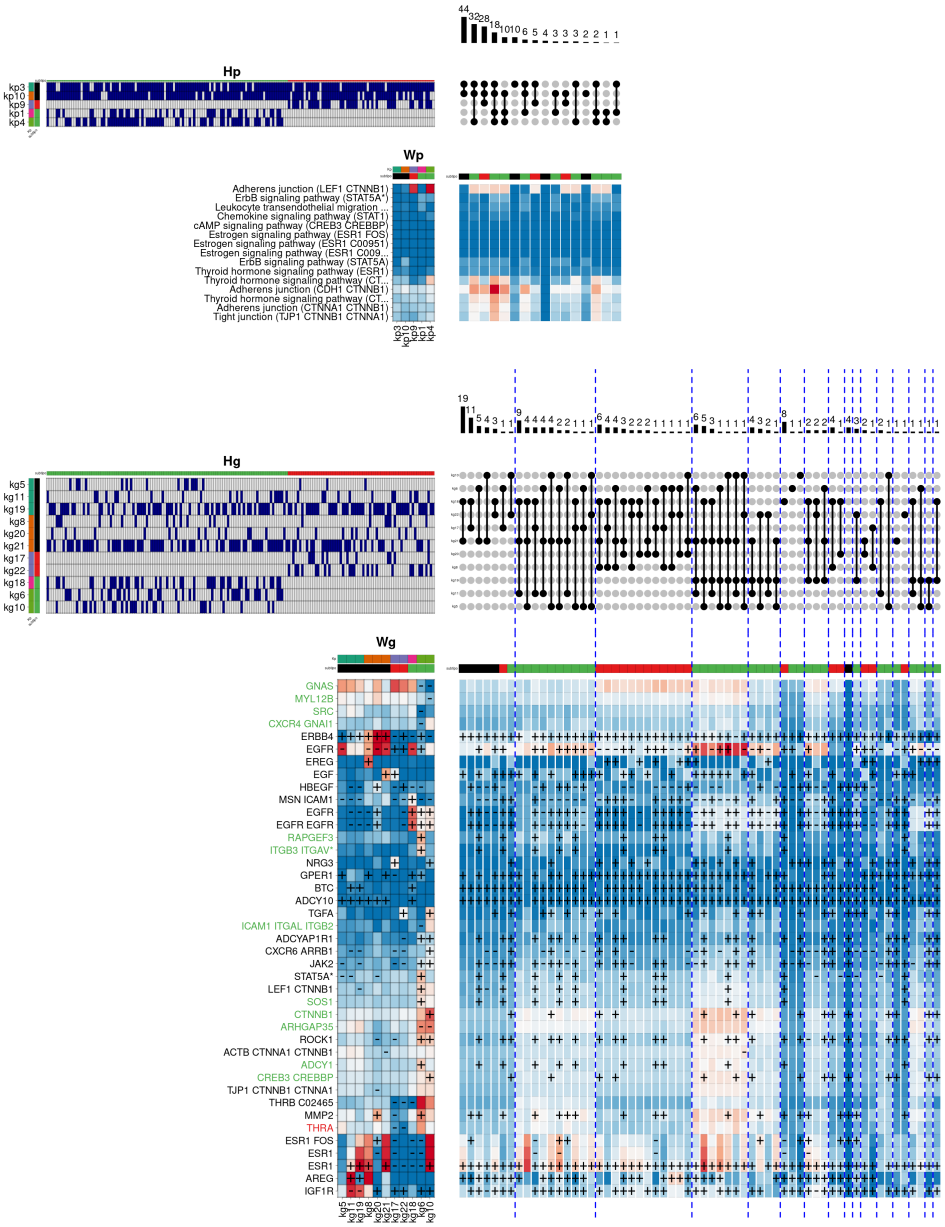


Figura 3.26: *Representación gráfica del modelo jerárquico obtenida en la función biológica "response to estrogen", aplicada a los subtipos Luminal A y Luminal B.*

asocia a las componentes a nivel de gen *kg10/6/10*. Estas componentes activan de forma simultánea a gran cantidad de genes, especialmente la componente *kg6*, que activa a los genes *EGFR*, *RAPGEF3* (*Rap Guanine Nucleotide Exchange Factor 3*, involucrado en angiogénesis), *ITGB3* (*Integrin Subunit Beta 3*), *GPER1*, *STAT5A* (*Signal Transducer And Activator Of Transcription 5A*, que regula la expresión de las proteínas de la leche durante la lactancia), *LEF1* (*Lymphoid Enhancer Binding Factor 1*), *SOS1* (*SOS Ras/Rac Guanine Nucleotide Exchange Factor 1*), *ROCK1* (*Rho Associated Coiled-Coil Containing Protein Kinase 1*), *ADCY1* (*Adenylate Cyclase 1*) y *MMP2*. También la componente *kg10* activa de forma simultánea a un número elevado de genes, algunos compartidos con la componente *kg6*, pero activando de forma específica a los genes *NRG3* (*Neuregulin 3*, un ligando específico del receptor *ERBB4*), *TGFA*, *CTNNB1* (*Catenin Beta 1*) y *CREB3* (*CAMP Responsive Element Binding Protein 3*, involucrado en la regulación del ciclo celular), entre otros.

La tercera combinación es en este caso específica del subtipo *Luminal B*. En concreto, la combinación agrega a la componente *kp9*, que activa de forma notable la via anterior de la ruta *Adherens junction*, además de activar de forma razonable otras cascadas activadas por las componentes anteriores. La componente *kp9* se asocia específicamente en el espacio de los genes a las componentes *kg17/22* que muestran un patrón bastante heterogéneo, mientras que la compnente *kg17* activa de forma clara a los genes *EGF* y *NRG3*, la componente *kg22* activa a *TGFA* y de forma débil a *HBEGF*.

Una vista global del modelo jerárquico determina una cantidad importante de genes asociados al subtipo *Luminal A* (en verde) debido a que muestran gran actividad en las componentes que son específicamente relevantes para los individuos del subtipo. Esta visión sugiere que los dos subtipos luminales, pese a compartir el estado de los receptores de hormona, muestran evidentes diferencias en cuanto a como regulan la respuesta a estrógenos.

Otra de las diferencias importantes entre ambos subtipos luminales lo constituye el mayor nivel de profileración celular observado en los tumores de tipo *Luminal*

B, produciendo un aumento significativo de la mortalidad frente a los individuos del subtipo *Luminal A*. Con el fin de determinar diferencias importantes entre ambos subtipos en este proceso biológico, la Figura 3.27 describe la regulación de la función *G2 M transition of mitotic cell cycle*, que describe una parte esencial del ciclo celular.

En particular, la Figura 3.27 muestra dos combinaciones a nivel de ruta que concentran a la mayoría de individuos. Ambas combinaciones hacen uso de la componente *kp12* (verde) encargada de activar parcialmente a las rutas de señalización *FoxO*, *Hippo*, *AMPK*, *Chemokine* y *Progesterone mediated oocyte maturation*, así como de la sobreactivación de la ruta *PI3K-Akt*. Esta componente se asocia en el espacio de los genes a las componentes *kg3/10*. En este caso, la componente *kg10* se asocia a ambos subtipos, teniendo como genes relevantes a *PRKCA* (*Protein Kinase C Alpha*, involucrado en la adhesión, diferenciación y proliferación celular) *BDNF*, y *SPDYA* (*Speedy/RINGO Cell Cycle Regulator Family Member A*, involucrado en la transición de fase del ciclo celular). La primera combinación resulta ser específica del subtipo *Luminal A* y, además de incluir a la componente *kp12*, añade las componentes *kp10* (naranja) y *kp2* (magenta), encargadas de activar con gran intensidad las rutas de señalización *FoxO* y *ErbB*, y de activar parcialmente a las rutas *MAPK* y *TGF-beta*. Dichas componentes se asocian en el espacio de los genes con las componentes *kg13/20* y *kg15/19* respectivamente. En este caso, dichas componentes activan de forma característica a los inhibidores de las ciclinas *CDKN1A*, *CDKN1B* y *CDKN2B*, además de activar a un bloque de genes formado por *NGFR*, *NGFRAP1*, *EDRNB* (*Endothelin Receptor Type B*), *PRKCA*, *EDF1* y *NTF4*.

Por su parte, la segunda combinación es específica del subtipo *Luminal B* y, además de la componente global *kp12*, incorpora a las componentes *kp6* (rosa) y *kp4* (verde claro), que activan de forma significativa a la ruta *PI3K-Akt* y la ruta de activación de la proteína *P53*. Dichas componentes se asocian en el espacio de los genes con las componentes *kg16* y *kg4/18* respectivamente. En este caso, la componente *kg16*, que muestra una prevalencia mayoritaria en la combinación, se

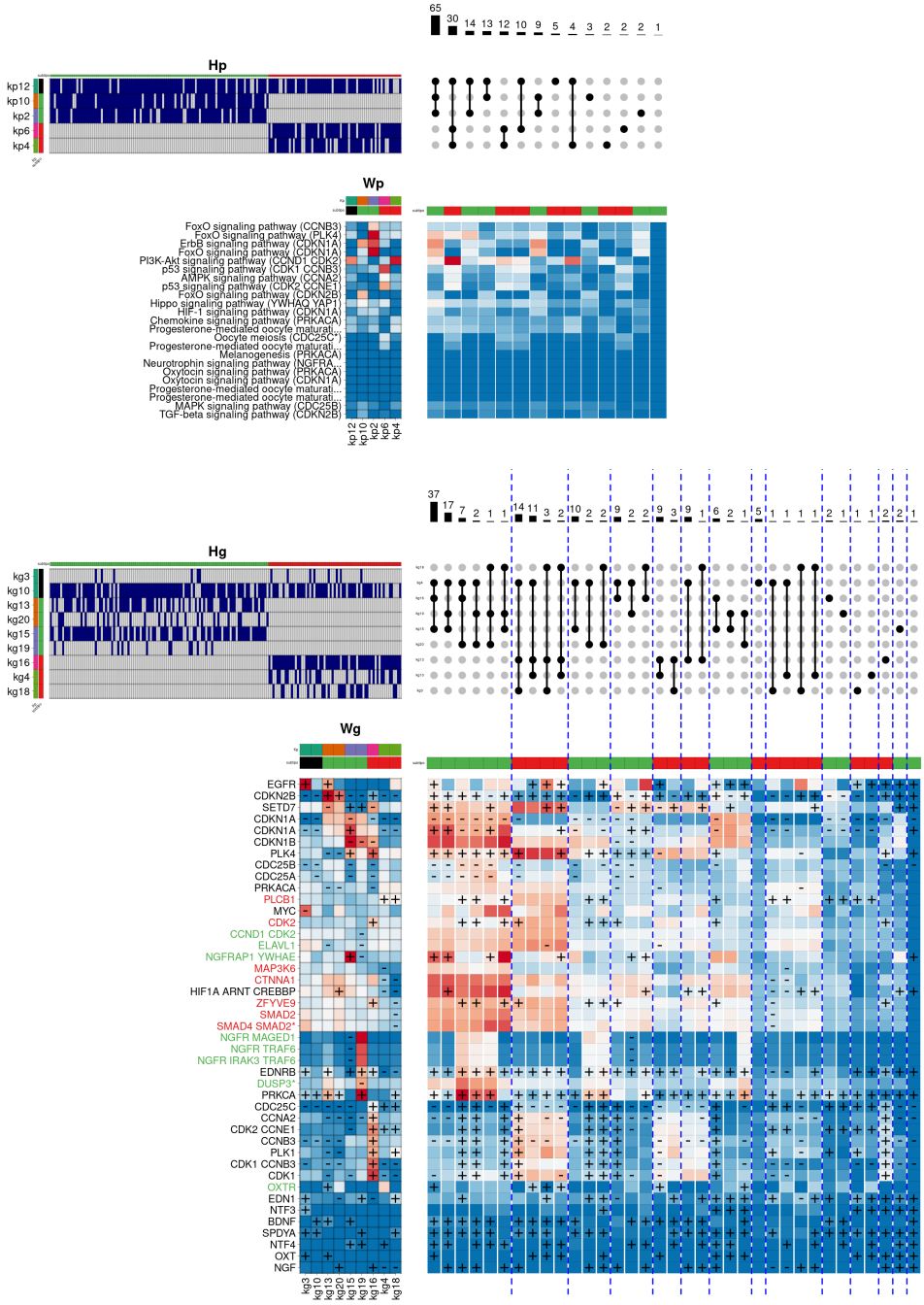


Figura 3.27: *Representación gráfica del modelo jerárquico obtenida en la función biológica "G2 M transition of mitotic cell cycle", aplicada a los subtipos Luminal A y Luminal B.*

encarga de sobreactivar de forma característica un conjunto de genes directamente relacionados con la regulación positiva del ciclo celular. En particular, la componente activa a los genes *CDC25C* (*Cell Division Cycle 25C*), *CCNA2* (*Cyclin A2*), *CDK1* (*Cyclin Dependent Kinase 1*), *CDK2* (*Cyclin Dependent Kinase 2*), *CCNB3* (*Cyclin B3*) y *PLK1* (*Polo Like Kinase 1*, altamente activo durante la mitosis).

Tal y como se observa, las diferencias encontradas entre ambas combinaciones permiten explicar las diferencias entre ambos subtipos luminales. En concreto, mientras que el subtipo *Luminal A* activa de forma clara algunos inhibidores del ciclo celular, las componentes asociadas al subtipo *Luminal B* sobreactivan de forma simultánea un bloque de genes que promueven dicha función.

Siguiendo esta misma línea, es posible aplicar la misma aproximación para estudiar de forma global a todos los subtipos presentes en el estudio, describiendo no solo la heterogeneidad presente en los individuos para una función biológica dada, sino también los elementos comunes entre subtipos. La Figura 3.28 describe el resultado del modelo jerárquico a lo largo de todos los subtipos incluidos en el estudio para la función biológica *mammary gland epithelial cell differentiation*. En este caso, la mayoría de componentes a nivel de ruta contribuye tanto a las muestras tumorales como a las normales. Únicamente se observaron dos componentes específicas de tumor. En concreto, la componente *kp8* se asocia de forma específica a los individuos del subtipo *Basal*. Esta componente activa de forma específica a las rutas de señalización *TNF* (*Tumor Necrosis Factor*), involucrada en numerosos procesos esenciales en la célula, y en especial, a la ruta *Fc epsilon RI*, involucrada en la respuesta inflamatoria.

Asimismo, la componente *kp1* se asocia específicamente a los individuos del subtipo *Luminal B*. En este caso, la componente activa de forma parcial la ruta de señalización *Hippo*, encargada de regular el tamaño de los órganos en la mayoría de animales, la ruta *Tigh Junction*, encargada de establecer una barrera permeable de difusión entre células adyacentes, y en especial la ruta *Jak-STAT*, que describe una cascada de eventos altamente pleiotrópica que se involucra en multitud de procesos biológicos esenciales como el ciclo celular, el metabolismo de lípidos, la

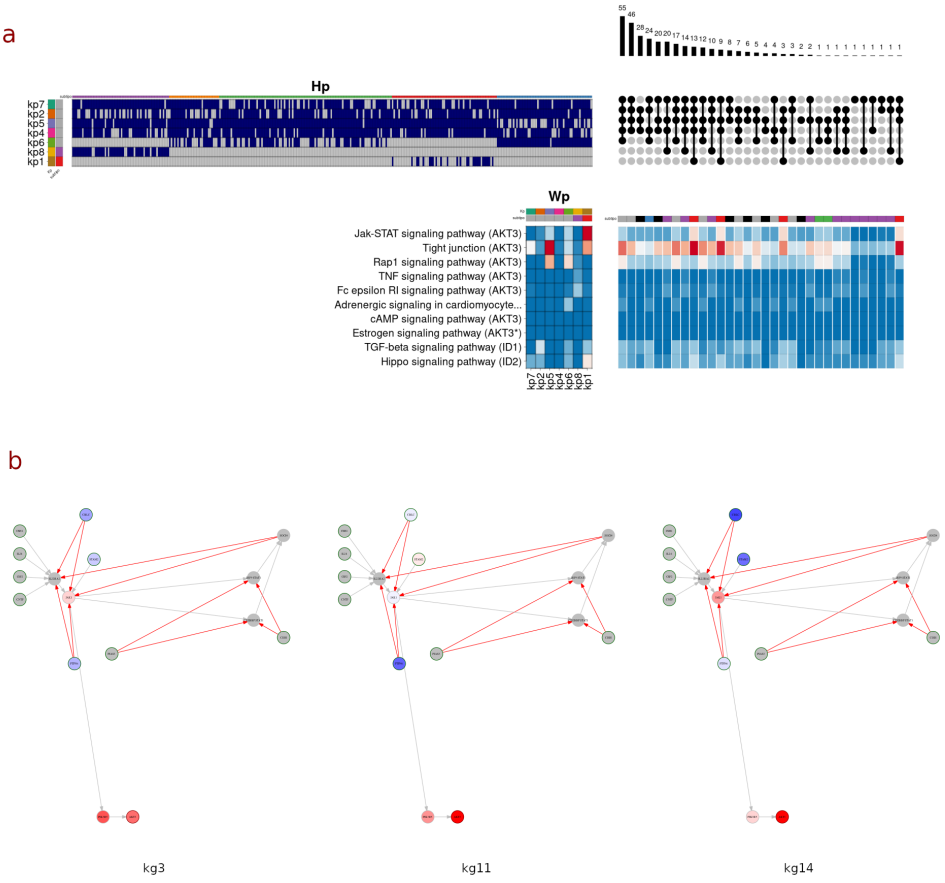


Figura 3.28: *Representación gráfica del modelo jerárquico a nivel de ruta obtenida en la función biológica "mammary gland epithelial cell differentitation", aplicada a todos los subtipos. a) Matrices obtenidas a nivel de ruta. b) Actividad de los nodos para la ruta Jak-STAT en las componentes kg3/11/14.*

diferenciación o la apoptosis. En este caso, y a modo de ejemplo, la Figura 3.28b muestra el patrón de activación de dicha cascada a cargo de las tres componentes a nivel de gen (*kg3/11/14*) asociadas a la componente *kp1*. Esta descripción permite entender el grado de heterogeneidad observado en los pacientes pertenecientes al subtipo *Luminal B*, ya que, cada componente representa un patrón de activación distinto a la hora de producir el mismo flujo de salida en la cascada. En un análisis más detallado, las tres soluciones se centran en activar de forma significativa los dos últimos nodos de la cascada (*PI3KR5* y *AKT3*) y de reducir el nivel de activación de algunos los nodos que actúan como inhibidores de *JAK1*, el cual, en caso de reducir del todo su actividad, produciría un cuello de botella que bloquearía la actividad de la cascada.

Siguiendo el mismo planteamiento, es posible estudiar otras funciones biológicas alteradas de forma general en todos los cánceres. La Figura 3.29 muestra el resultado del modelo jerárquico para la función biológica *Positive regulation of response to DNA damage stimulus*. Esta función engloba diferentes mecanismos involucrados en la respuesta celular al daño en el genoma, que en células normales suele llevar a activar procesos de reparación o, en caso grave, a la apoptosis celular con el fin de preservar la integridad del tejido.

En particular, la Figura 3.29 permite identificar una variedad más amplia en la especificidad de cada componente a los distintos subtipos, apareciendo dos componentes (*kp6* y *kp1*) asociadas a varios subtipos tumorales, dos componentes (*kp10* y *kp5*) asociadas a los individuos del subtipo *Basal*, una componente (*kp12*) asociada al subtipo *Her2* y una componente asociada específicamente a los individuos normales. Esta última componente activa de forma muy clara a las rutas pleiotrópicas *Jak-STAT* y *PI3K-Akt*, y también a la ruta de respuesta a la hormona tiroides, seguramente específica del tejido mamario. Asimismo, la componente *kp5*, asociada de forma específica al subtipo *Basal*, activa de forma notoria a la ruta *P53*. Esta ruta orquesta de manera principal la respuesta al daño en el ADN alrededor de la proteína *P53*, la cual constituye el gen más mutado en el conjunto de cánceres contenido en el repositorio del consorcio *ICGC*.

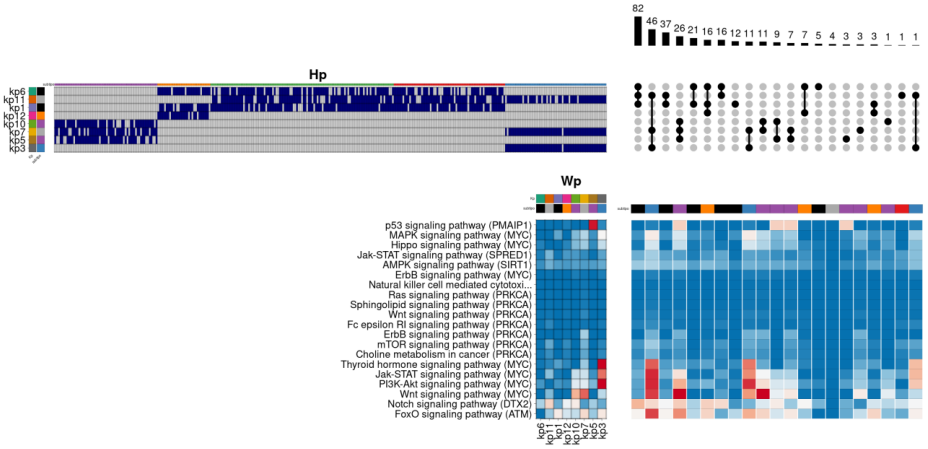


Figura 3.29: *Representación gráfica del modelo jerárquico a nivel de ruta obtenida en la función biológica "positive regulation of response to DNA damage stimulus", aplicada a todos los subtipos.*

Otro de los *hallmarks* más importantes en el cáncer lo define la relación entre las células tumorales y su microentorno más inmediato, donde las células tumorales son capaces de modificar el comportamiento de otras células normales en su propio beneficio. Para ello, la relación entre las células tumorales y la matrix extracelular juega un papel importante. La Figura 3.30 describe la configuración de componentes y combinaciones encontrada por el modelo jerárquico para el término *Positive regulation of cell-matrix adhesion*. Se trata de una función biológica con una enorme variedad de soluciones, apoyadas en un conjunto de componentes que muestran un nivel de especificidad por los diferentes subtipos bastante alto. Se aprecian multitud de componentes asociadas a más de un subtipo tumoral (*kg4/22/25/10/17/3/21/11/18/5/9*). Asimismo, también se observaron componentes específicas para los subtipos *Luminal B* (*kg6/17*), *Her2* (*kg8/13*), *Basal* (*kg20/24*), y *Normal* (*kg5/9*). Estos dos últimos grupos de componentes destacan sobre el resto, ya que activan en bloque un gran número de genes. En particular, las componentes del subtipo *Basal* activan los reguladores de la adhesión celular *FGR* (*FGR Proto-Oncogene, Src Family Tyrosine Kinase*) y *LYN* (*LYN Proto-Oncogene, Src Family Tyrosine Kinase*), así como un conjunto genes de carácter pleiotrópico,

en concreto *DAPP1* (*Protein Tyrosine Phosphatase Receptor Type C*), *RELB* (*RELB Proto-Oncogene, NF-KB Subunit*), *PTPRC* (*Protein Tyrosine Phosphatase Receptor Type C*), *PTK2B* (*Protein Tyrosine Kinase 2 Beta*), entre otros.

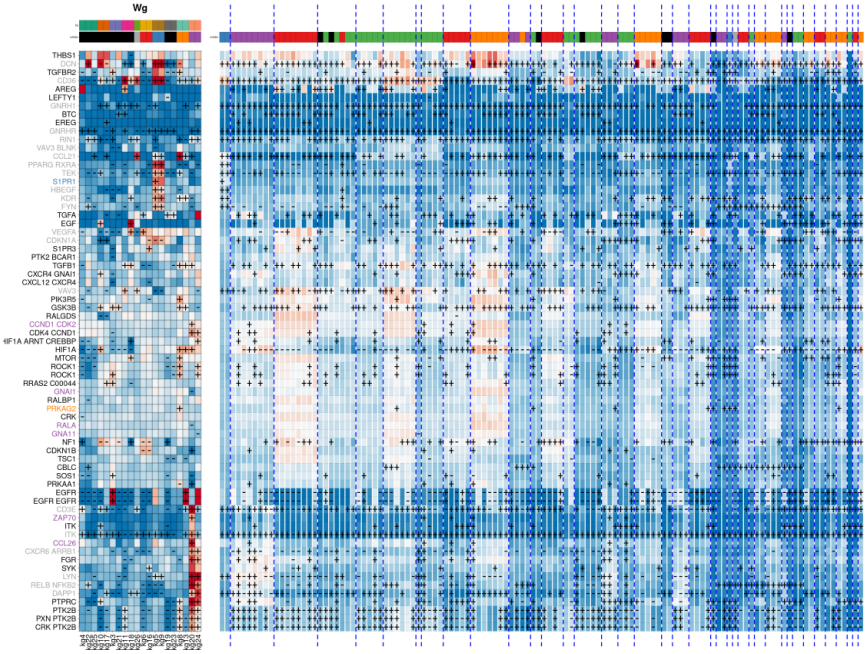


Figura 3.30: *Representación gráfica del conjunto de componentes a nivel de gen y sus correspondientes combinaciones obtenida en la función biológica "positive regulation of cell-matrix adhesion", aplicada a todos los subtipos.*

Por el contrario, las componentes asociadas al subrupo *Normal* activan de forma clara a un conjunto distinto de genes, involucrado en la regulación basal de la adhesión entre las células y la matriz extracelular. En concreto, se activan los genes *CCL21* (*C-C Motif Chemokine Ligand 21*), *PPARG* (*Peroxisome Proliferator Activated Receptor Gamma*), *TEK* (*TEK Receptor Tyrosine Kinase*, regulador de la adhesion focal entre células), *S1PR1* (*Sphingosine-1-Phosphate Receptor 1*), *HBEGF* (*Heparin Binding EGF Like Growth Factor)*, *KDR* (*Kinase Insert Domain Receptor*, factor de crecimiento para las células endoteliales) y *FYN* (*FYN Proto-Oncogene, Src Family Tyrosine Kinase*).

3.5. Conclusiones

Los análisis realizados han permitido modelar a nivel estadístico la composición interna de los pacientes, describiendo de una manera detallada un conjunto de soluciones o estrategias a nivel celular que los individuos implementan a la hora de regular una determinada función biológica alterada en la enfermedad. Este punto de vista ha proporcionado una descripción cuantitativa sobre las diferencias y similitudes que los diferentes subtipos presentan, proporcionando claves muy valiosas a la hora de comprender el grado de heterogeneidad dentro de cada subtipo.

La aproximación al problema se ha llevado a cabo desde la perspectiva de la biología de sistemas, abordando el modelado estadístico de la función celular tanto a nivel de gen, como a nivel de ruta molecular. Para ello, el modelo desarrollado ha partido de un conjunto de ecuaciones derivadas de la herramienta *Hipathia*, que contienen de forma implícita la estructura y topología de las redes de señalización. En particular, las ecuaciones describen el peso que cada gen tiene en aquellas rutas en las que participa, permitiendo así conectar los dos niveles de información incluidos en el modelo.

El modelo estadístico desarrollado se basa en la técnica de factorización de matrices *NMF*. Esta técnica, que impone como restricción la positividad de los valores de las matrices del modelo, describe las observaciones como una suma positiva de partes más elementales. Esta característica facilita la interpretación directa de cada una de las componentes encontradas en la factorización, ya que, a diferencia de otras técnicas como *PCA* o *ICA*, no existen valores negativos que deban ser previamente cancelados mediante la combinación de varias componentes. En este caso, se ha desarrollado un modelo estadístico que factoriza de forma simultánea las matrices de entrada correspondientes a la actividad de los genes y las rutas en los individuos del estudio. Para ello, durante la optimización, el modelo impone una serie de restricciones que permiten conectar de forma jerárquica ambos conjuntos de componentes, produciendo que sean compatibles en el espacio de las rutas y que, por tanto, proporcionen soluciones coherentes a nivel biológico.

La aplicación del modelo jerárquico de factorización ha proporcionado resultados satisfactorios. La primera parte de su evaluación se ha llevado a cabo mediante el uso de un conjunto de simulaciones, especialmente diseñadas para reproducir la estructura jerárquica que las células muestran en su funcionamiento. En este caso, las simulaciones realizadas permitieron evaluar de forma precisa la capacidad del modelo a la hora de encontrar un conjunto de componentes latentes previamente introducidas, mostrando un rendimiento superior al de un abordaje no jerárquico realizado mediante algunas de las aproximaciones más clásicas de la técnica *NMF*. Asimismo, las simulaciones han permitido evaluar el grado de error esperado por el método a la hora de estimar el número óptimo de componentes a emplear en el modelo, ofreciendo una precisión muy superior a la obtenida por otras aproximaciones clásicas como el coeficiente de correlación cofenético o el método de la silueta.

De forma complementaria, la aplicación del modelo jerárquico al conjunto de pacientes reales, también produjo resultados interesantes. En primer lugar, la evaluación de los distintos modelos obtenidos para cada una de las funciones biológicas alteradas mostró un grado de convergencia muy adecuado, obteniendo una bondad del ajuste cercada a 1 con respecto a la matrices originales de entrada. Además, tanto las matrices de componentes, como las matrices de mezcla, mostraron en general un grado de correlación muy alto al comparar las soluciones encontradas a nivel de gen con las soluciones encontradas a nivel de ruta, lo que valida de forma implícita el planteamiento seguido a la hora de definir la estructura jerárquica del modelo.

Los resultados obtenidos mostraron también de forma recurrente una asociación preferencial de las componentes encontradas a alguno de los subtipos incluidos en la muestra, demostrando así la capacidad del modelo para capturar la estructura interna de la cohorte de individuos. Esta asociación permitió además caracterizar a nivel oncogénico cada una de las componentes encontradas, en función de si mostraron un perfil asociado a las muestras tumorales, un perfil más general asociado al tejido de estudio, o una asociación específica a las muestras sanas. Asimismo, los

resultados mostraron la asociación significativa entre algunas de las componentes encontradas y la probabilidad de supervivencia de los individuos, capturando así el efecto producido por la acción combinada de varios genes relevantes en la evolución de la enfermedad.

Los resultados obtenidos mostraron también de forma recurrente la presencia de sinergias significativas de carácter positivo entre parejas de componentes, lo que sugiere la existencia de alteraciones específicas en partes distintas de la célula que han de ser necesariamente combinadas para producir un fenotipo determinado. Esta aproximación fue además representada mediante el uso de redes, proporcionando así una descripción más global sobre la composición interna de los individuos. Además, de forma notable, se observaron sinergias negativas, lo que sugiere la existencia de trayectorias biológicas excluyentes, donde necesariamente algunas alteraciones observadas (representadas mediante componentes) estarían condicionadas a la existencia previa de otras alteraciones.

Por otro lado, el análisis de los resultados obtenidos por cada modelo jerárquico permitió determinar cuales fueron los genes más relevantes dentro de cada función biológica estudiada, determinando además su peso dentro de cada componente. Los genes clasificados como relevantes dentro del modelo mostraron un solapamiento significativo con algunos conjuntos de genes esenciales, como el censo de genes de cáncer recopilado por *COSMIC*, o los conjuntos de genes incluidos en los predictores *MammaPrint*, *Oncotype* o *PAM50*. Además, los solapamientos obtenidos demuestran que el efecto producido por la variación de un gen en el espacio de las rutas constituye el método más preciso para evaluar su relevancia en una determinada componente o función biológica. Este punto confirma la necesidad de contar con un sistema que integre la estructura topológica de las rutas moleculares.

Por su parte, el meta-análisis de funciones proporcionó un mejor agrupamiento de los subtipos al combinar las matrices de mezcla, frente al obtenido mediante las matrices originales de entrada. Este punto resulta de gran importancia ya que el proceso de factorización proporciona de forma implícita una reducción en la dimensionalidad de los datos, demostrando de nuevo la capacidad del modelo

a la hora de capturar satisfactoriamente un conjunto reducido de componentes esenciales que describen de forma adecuada a los individuos de la cohorte. Además, la aproximación seguida confirma que una combinación adecuada de los resultados obtenidos a lo largo de todas las funciones biológicas constituye una representación más precisa de la identidad de los diferentes subtipos que la proporcionada por cada modelo de forma independiente.

La parte más esencial de los resultados ha sido la interpretación gráfica obtenida por el modelo jerárquico, la cual ha permitido representar aquellas funciones biológicas que potencialmente podrían explicar las diferencias entre los subtipos. En este caso, la representación gráfica no solo permitió determinar cuales fueron las principales componentes a nivel de ruta para cada subtipo, sino también las combinaciones de componentes más frecuentes en los individuos, lo que constituye una representación directa de la heterogeneidad genómica dentro de un grupo de muestras. Además, el modelo jerárquico permitió determinar la existencia de componentes y combinaciones asociadas a más de un subtipo en la enfermedad, lo que describe potencialmente características comunes que permitirían el diseño de fármacos más generales. En este sentido, el modelo jerárquico contribuyó también a identificar de forma clara las componentes a nivel de gen más frecuentes dentro de las combinaciones observadas. Esta representación podría permitir el diseño de terapias combinadas que tuvieran por objetivo el conjunto de genes que simultáneamente se activan dentro de una misma componente a nivel de gen, considerada esencial dentro de uno o varios subtipos.

A nivel biológico, los modelos jerárquicos permitieron entender la estructura interna del subtipo *Her2* en la regulación del factor de crecimiento epidérmico, la estructura interna del subtipo *Basal* en la ruta de señalización *Notch* y las diferencias entre los subtipos *Luminal A* y *Luminal B* en la regulación de la respuesta a estrógenos o la regulación del ciclo celular. Asimismo, se evaluaron algunas funciones biológicas importantes en la enfermedad como la diferenciación de las células epiteliales mamarias, u otros procesos relevantes en el cáncer como la respuesta al daño en el ADN, o la interación entre las células y la matriz extracelular.

Se trata de resultados que confirman los hallazgos previos en la enfermedad y que validan los resultados obtenidos en el resto de funciones biológicas alteradas, cuya estructura resulta menos conocida en el estudio del cáncer.

Uno de los puntos más importantes del protocolo propuesto, ha sido la integración de los valores de expresión y las mutaciones somáticas observadas en los individuos. En este caso, el valor de actividad de cada gen se obtuvo mediante la combinación del efecto de sus mutaciones somáticas (representando las posibles pérdidas o ganancias de función), y del valor de expresión génica medido en cada individuo (que de forma natural integra el resultado de diferentes capas de regulación cromosómica, como los cambios epigenéticos, epigenómicos o alteraciones en el número de copias en el ADN). Se trata sin duda de un punto esencial a tener en cuenta en el futuro, ya que todavía existen algunos tipos de mutaciones somáticas cuyo efecto en la estructura de las proteínas sigue estar bien modelado.

Una de las limitaciones principales dentro del protocolo propuesto lo constituye el tiempo de cómputo empleado para optimizar las matrices del modelo, superior al de otras aproximaciones más sencillas. En este caso, el modelo incluye un número elevado de términos y matrices que facilitan la obtención de soluciones muy específicas a nivel biológico, pero que añaden una carga computacional importante al proceso de factorización. Además, una parte esencial de la optimización lo constituye el cálculo de las derivadas parciales de los nodos en cada una las rutas, proceso que añade una carga todavía más pesada al proceso. En este sentido, la búsqueda de métodos de inicialización podrían contribuir a aliviar de forma considerable el tiempo de cómputo, conservando las virtudes del modelo.

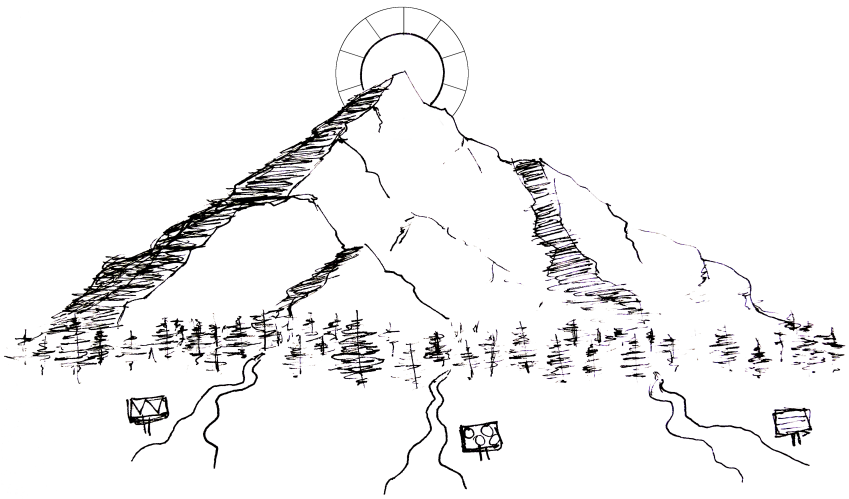
Otro de los puntos delicados lo constituye el ajuste de los pesos de cada término dentro del modelo de factorización. En este caso, los parámetros fueron optimizados empleando técnicas de optimización global, donde el número de iteraciones empleado en cada ejecución del modelo fue necesariamente ajustado a un número muy reducido, con el fin de no producir una carga computacional inasumible. En este sentido, será interesante explorar en el futuro otras alternativas de optimización global que pudieran producir una convergencia más rápida, así como la deducción

de reglas heurísticas que pudiesen acotar de forma natural el rango de los pesos en el modelo.

Como trabajo futuro, sin duda destaca la aplicación del modelo jerárquico a otro tipo de rutas moleculares, como las rutas metabólicas. En este caso, el diseño del modelo permitiría integrar cualquier tipo de ruta molecular, siempre que estuviera representada mediante un conjunto de ecuaciones que describan el peso de cada gen en cada parte más elemental del sistema. Asimismo, el uso de otras ontologías de anotación diferentes a *GO* podría permitir el análisis de otro tipo de entidades más abstractas como los *hallmarks* del cáncer. Por último, como trabajo principal quedaría la búsqueda de alternativas automáticas que permitan analizar la enorme cantidad de resultados proporcionada por el protocolo propuesto en un análisis real.

CAPÍTULO 4

Conclusiones generales y líneas de futuro



El análisis de la heterogeneidad genómica representa una de las líneas de investigación más importantes en el estudio del cáncer. El motivo es que constituye uno de los mayores obstáculos a la hora de comprender la secuencia de alteraciones genómicas necesarias para iniciar el proceso de tumorigénesis, condicionando a nivel práctico el diseño de tratamientos genéricos que frenen la enfermedad e impidan la recaída de los pacientes.

A continuación, y en forma de listado, se resumen las conclusiones principales de este trabajo:

- En esta tesis se describe un protocolo de análisis global destinado al estudio de la heterogeneidad genómica en un conjunto de individuos afectados por una enfermedad como el cáncer de mama. El protocolo parte de la estimación de las mutaciones somáticas en los individuos del estudio, para después, en una segunda fase, proyectar su efecto sobre la estructura de los genes y las rutas moleculares en las que participan.

- El trabajo realizado se ha abordado desde el punto de vista de la biología de sistemas, ofreciendo una descripción detallada de aquellas funciones biológicas alteradas que conducen al desarrollo de los *hallmarks* del cáncer.

- Los modelos construidos a partir del uso de técnicas de factorización de matrices han permitido la obtención de una serie de componentes latentes que describen de forma directa las diferentes estrategias genómicas que los tumores de cada subtipo implementan para poder llevar a cabo la regulación de las funciones biológicas alteradas.

- Además de la heterogeneidad biológica, en esta tesis ha sido fundamental el modelado de la variabilidad técnica, el cual ha sido tratado desde diferentes niveles. En particular, el modelado de los estadísticos de ruido ha sido una pieza clave a la hora de proporcionar una predicción más fiable sobre las posibles mutaciones somáticas presentes. Asimismo, el uso de técnicas de factorización aplicadas en la segunda parte de la tesis ha contribuido a amortiguar gran parte de la variabilidad de origen técnico.

- Por otro lado, cabe destacar la importancia que en esta tesis ha tenido el diseño y uso de la simulación como herramienta para evaluar de forma precisa la fiabilidad de los modelos estadísticos planteados, especialmente en contextos como el genotipado somático, donde las validaciones experimentales resultan escasas. En este caso, a lo largo de la tesis se han planteado simulaciones cuidadosamente diseñadas con el fin de representar de forma realista la estructura interna que las muestras reales contienen, pudiendo así proporcionar resultados fiables durante la fase de construcción de los modelos.

- Los modelos estadísticos propuestos en esta tesis han permitido el estudio de la heterogeneidad genómica desde diferentes puntos de vista y diferentes niveles de información a nivel biológico. En particular, el abordaje seguido en la primera parte de la tesis ha permitido modelar la variabilidad intratumoral con el propósito de ofrecer una predicción más robusta sobre las mutaciones somáticas presentes en el genoma de un individuo. De forma complementaria, en la segunda parte de la tesis, el modelo jerárquico ha proporcionado una vista detallada de la heterogeneidad genómica entre los individuos que pertenecen a un mismo subtipo.

- Por último, en este trabajo se han descrito las alteraciones producidas por mutaciones somáticas y su efecto en la actividad de los genes, las rutas moleculares en las que participan, y las funciones biológicas que regulan, lo que constituye una visión muy detallada y estructurada sobre la variabilidad genómica de una enfermedad tan compleja como el cáncer.

Para finalizar, a continuación se describen las principales líneas de trabajo futuro, ya descritas a lo largo de la tesis:

- Se prevé la simulación de linajes tumorales con diferentes progenitores somáticos, incluyendo poblaciones celulares mucho más grandes.

- Otra línea es el diseño de estrategias que faciliten la coexistencia de diferentes clones en la misma simulación tumoral. En este caso, los clones no sólo

interactuarían mediante un escenario de competición, sino también mediante la implementación de co-dependencias entre ellos.

- Además, se plantea el desarrollo de un modelo espacial en el simulador de tumores. Esta opción permitirá situar a las células en un entorno tridimensional donde se pueda estimar de forma más precisa la presión selectiva en cada célula en función de la proximidad a recursos esenciales, o dependiendo de la cercanía de otras células del tejido.

- Otra posibilidad abierta es la aplicación del modelo jerárquico a otro tipo de rutas moleculares, como las rutas metabólicas. En este caso, el diseño del modelo jerárquico permite integrar cualquier tipo de ruta molecular, siempre que esté representada mediante un conjunto de ecuaciones que describan el peso de cada gen en cada parte del sistema.

- También a destacar la posibilidad de emplear otras ontologías de anotación diferentes a *GO* para estratificar las cascadas de señalización en el modelo jerárquico. Esta opción podría permitir el análisis de otro tipo de entidades más abstractas, como los *hallmarks* del cáncer.

- Por último, quedaría la búsqueda de alternativas automáticas que permitan analizar la enorme cantidad de resultados proporcionada por el protocolo propuesto en un análisis real, incluyendo la generación de informes y la busqueda automática de marcadores relevantes para la enfermedad.

Bibliografía

Abecasis, G. R. e. a. (2012, Nov). An integrated map of genetic variation from 1,092 human genomes. *Nature*, *491*(7422), 56–65.

Ainscough, B. J., Barnell, E. K., Ronning, P., Campbell, K. M., Wagner, A. H., Fehniger, T. A., ... Griffith, O. L. (2018, 12). A deep learning approach to automate refinement of somatic variant calling from cancer sequencing data. *Nat. Genet.*, *50*(12), 1735–1743.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716-723. doi: 10.1109/TAC.1974.1100705

Alexandrov, L. B., Kim, J., Haradhvala, N. J., Huang, M. N., Tian Ng, A. W., Wu, Y., ... Yu, W. (2020, 02). The repertoire of mutational signatures in human cancer. *Nature*, *578*(7793), 94–101.

Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J., y Stratton, M. R. (2013, Jan). Deciphering signatures of mutational processes operative in human cancer. *Cell Rep*, *3*(1), 246–259.

Alexandrov, L. B., y Stratton, M. R. (2014, Feb). Mutational signatures: the patterns of somatic mutations hidden in cancer genomes. *Curr. Opin. Genet. Dev.*, *24*, 52–60.

Al-Shahrour, F., Arbiza, L., Dopazo, H., Huerta-Cepas, J., M?nguez, P., Montaner, D., y Dopazo, J. (2007, Apr). From genes to functional classes in the study of biological systems. *BMC Bioinformatics*, *8*, 114.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., y Lipman, D. J. (1990, Oct). Basic local alignment search tool. *J. Mol. Biol.*, *215*(3), 403–410.

Amadoz, A., y Hidalgo, M. R. (2018). A comparison of mechanistic signaling pathway activity analysis methods. *Briefings in bioinformatics*(June), 1–14. doi: 10.1093/bib/bby040

Ardia, D., Boudt, K., Carl, P., Mullen, K. M., y Peterson, B. G. (2011). Differential Evolution with DEoptim: An application to non-convex portfolio optimization. *The R Journal*, *3*(1), 27–34. Descargado de https://journal.r-project.org/archive/2011-1/RJournal_2011-1_Ardia~et~al.pdf

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., ... Sherlock, G. (2000, May). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, *25*(1), 25–29.

Ashley, E. A. (2016, 08). Towards precision medicine. *Nat. Rev. Genet.*, *17*(9), 507–522.

Axelrod, R., Axelrod, D. E., y Pienta, K. J. (2006). Evolution of cooperation among tumor cells. *Proceedings of the National Academy of Sciences*, *103*(36), 13474–13479. doi: 10.1073/pnas.0606053103

Bayar, B., Bouaynaya, N., y Shterenberg, R. (2014, Feb). Probabilistic non-negative matrix factorization: theory and application to microarray data analysis. *J Bioinform Comput Biol*, *12*(1),

1450001.

Bayati, M., Rabiee, H. R., Mehrbod, M., Vafaee, F., Ebrahimi, D., Forrest, A. R. R., y Alinejad-Rokny, H. (2020, Jan). CANCERSIGN: a user-friendly and robust tool for identification and classification of mutational signatures and patterns in cancer genomes. *Sci Rep*, *10*(1), 1286.

Burrell, R. A., McGranahan, N., Bartek, J., y Swanton, C. (2013, Sep). The causes and consequences of genetic heterogeneity in cancer evolution. *Nature*, *501*(7467), 338–345.

Byron, S. A., Van Keuren-Jensen, K. R., Engelthaler, D. M., Carpten, J. D., y Craig, D. W. (2016, May). Translating RNA sequencing into clinical diagnostics: opportunities and challenges. *Nat. Rev. Genet.*, *17*(5), 257–271.

Carbonell-Caballero, J., Amadoz, A., Alonso, R., Hidalgo, M. R., Çubuk, C., Conesa, D., ... Dopazo, J. (2017, Nov). Reference genome assessment from a population scale perspective: an accurate profile of variability and noise. *Bioinformatics*, *33*(22), 3511–3517.

Cashero, Z., y Anderson, C. (2011). Comparison of EEG blind source separation techniques to improve the classification of P300 trials. *Conf Proc IEEE Eng Med Biol Soc*, *2011*, 7183–7186.

Chacón-Solano, E., León, C., Díaz, F., García-García, F., García, M., Escámez, M., ... del Río, M. (2019). Fibroblasts activation and abnormal extracellular matrix remodelling as common hallmarks in three cancer-prone genodermatoses. *British Journal of Dermatology*, 0–3. Descargado de http://doi.wiley.com/10.1111/bjd.17698 doi: 10.1111/bjd.17698

Chakraborty, S., Hosen, M. I., Ahmed, M., y Shekhar, H. U. (2018). Onco-Multi-OMICS Approach: A New Frontier in Cancer Research. *Biomed Res Int*, *2018*, 9836256.

Chen, X., Schulz-Trieglaff, O., Shaw, R., Barnes, B., Schlesinger, F., Killberg, M., ... Saunders, C. T. (2016, 04). Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics*, *32*(8), 1220–1222.

Cherry, E. C. (1953). *Coktail Party Effect Cherry 1953.pdf* (Vol. 25) (n.º 5).

Cibulskis, K., Lawrence, M. S., Carter, S. L., Sivachenko, A., Jaffe, D., Sougnez, C., ... Getz, G. (2013, mar). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature biotechnology*, *31*(3), 213–9. Descargado de http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3833702{&}tool=pmcentrez{&}rendertype=abstract doi: 10.1038/nbt.2514

Cingolani, P., Platts, A., Coon, M., Nguyen, T., Wang, L., Land, S., ... Ruden, D. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, snpeff: Snps in the genome of drosophila melanogaster strain w1118; iso-2; iso-3. *Fly*, *6*(2), 80-92.

Cleary, A. S., Leonard, T. L., Gestl, S. A., y Gunther, E. J. (2014). Tumour cell heterogeneity maintained by cooperating subclones in Wnt-driven mammary cancers. *Nature*, *508*(1), 113–117. Descargado de http://dx.doi.org/10.1038/nature13187 doi: 10.1038/nature13187

Comon, P. (1994). Independent component analysis, A new concept? *Signal Processing*, *36*(3), 287–314. doi: 10.1016/j.sigpro.2015.03.006

Crick, F. (1970, Aug). Central dogma of molecular biology. *Nature*, *227*(5258), 561–563.

Dagogo-Jack, I., y Shaw, A. T. (2018, 02). Tumour heterogeneity and resistance to cancer therapies. *Nat Rev Clin Oncol*, *15*(2), 81–94.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., ... Wang, J. (2011, Aug). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158.

Darwin, C. (1859). *On the origin of species by means of natural selection.* London: Murray. (or the Preservation of Favored Races in the Struggle for Life)

Davis, J. D., Kumbale, C. M., Zhang, Q., y Voit, E. O. (2019, 08). Dynamical systems approaches to personalized medicine. *Curr. Opin. Biotechnol.*, *58*, 168–174.

Dempster, L., y Rubin. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*, *39*(1), 1–38. doi: 10.1115/1.3424485

Diaz-Uriarte, R. (2017). OncoSimulR: Genetic simulation with arbitrary epistasis and mutator genes in asexual populations. *Bioinformatics*, *33*(12), 1898–1899. doi: 10.1093/bioinformatics/btx077

Ding, J., Bashashati, A., Roth, A., Oloumi, A., Tse, K., Zeng, T., ... Shah, S. P. (2012, jan). Feature-based classifiers for somatic mutation detection in tumour-normal paired sequencing data. *Bioinformatics (Oxford, England)*, *28*(2), 167–75. Descargado de http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3259434{&}tool=pmcentrez{&}rendertype=abstract doi: 10.1093/bioinformatics/btr629

Ding, X., Lee, J. H., y Lee, S. W. (2013, Apr). Performance evaluation of nonnegative matrix factorization algorithms to estimate task-related neuronal activities from fMRI data. *Magn Reson Imaging*, *31*(3), 466–476.

Do, J. H., y Choi, D. K. (2006, Apr). Computational approaches to gene prediction. *J. Microbiol.*, *44*(2), 137–144.

Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., Davis, C. A., Doyle, F., ... Birney, E. (2012, Sep). An integrated encyclopedia of DNA elements in the human genome. *Nature*, *489*(7414), 57–74.

Eaton, J., Wang, J., y Schwartz, R. (2018, 07). Deconvolution and phylogeny inference of structural variations in tumor genomic samples. *Bioinformatics*, *34*(13), i357-i365.

Eckart C, Y. G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika*, *1*(3), 211–218. Descargado de papers2://publication/uuid/DE3742B5-B0D3-401B-99DD-B066C04BD9DE

Edge, L. (2016). Emt: 2016. , 21–45. doi: 10.1016/j.cell.2016.06.028

Esposito, F., Scarabino, T., Hyvarinen, A., Himberg, J., Formisano, E., Comani, S., ... Di Salle, F. (2005). Independent component analysis of fMRI group studies by self-organizing clustering. *NeuroImage*, *25*(1), 193–205. doi: 10.1016/j.neuroimage.2004.10.042

Esteban-Medina, M., Peña-Chilet, M., Loucera, C., y Dopazo, J. (2019, Jul). Exploring the druggable space around the Fanconi anemia pathway using machine learning and mechanistic models. *BMC Bioinformatics*, *20*(1), 370.

Ewing, B., y Green, P. (1998, Mar). Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.*, *8*(3), 186–194.

Ewing, B., Hillier, L., Wendl, M. C., y Green, P. (1998, Mar). Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.*, *8*(3), 175–185.

Ezzat, A., Wu, M., Li, X. L., y Kwoh, C. K. (2019, 07). Computational prediction of drug-target interactions using chemogenomic approaches: an empirical survey. *Brief. Bioinformatics*, *20*(4), 1337–1357.

Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., ... D'Eustachio, P. (2018, 01). The Reactome Pathway Knowledgebase. *Nucleic Acids Res.*, *46*(D1), D649-D655.

Fang, L. T., Zhu, B., Zhao, Y., Chen, W., Yang, Z., Kerrigan, L., ... Consortium, T. S. W. G. o. S.-I. (2019). Establishing reference samples for detection of somatic mutations and germline variants with NGS technologies. *bioRxiv*, 625624. Descargado de https://www.biorxiv.org/content/10.1101/625624v1 doi: 10.1101/625624

Finotello, F., y Eduati, F. (2018). Multi-Omics Profiling of the Tumor Microenvironment: Paving the Way to Precision Immuno-Oncology. *Front Oncol*, *8*, 430.

Fisher, R. (1925). Statistical Methods for Research Workers. *Edinburgh*.

Gaujoux, R., y Seoighe, C. (2012, Jul). Semi-supervised Nonnegative Matrix Factorization for gene expression deconvolution: a case study. *Infect. Genet. Evol.*, *12*(5), 913–921.

Gerlinger, M., Rowan, A. J., Horswell, S., Larkin, J., Endesfelder, D., Gronroos, E., ... Swanton, C. (2012). Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *New England Journal of Medicine*, *366*(10), 883-892. Descargado de https://doi.org/10.1056/NEJMoa1113205 (PMID: 22397650) doi: 10.1056/NEJMoa1113205

Gerstung, M., Jolly, C., Leshchiner, I., Dentro, S. C., Gonzalez, S., Rosebrock, D., ... Van Loo, P. (2020, 02). The evolutionary history of 2,658 cancers. *Nature*, *578*(7793), 122–128.

Ghaffarizadeh, A., Heiland, R., Friedman, S., Mumenthaler, S., y Macklin, P. (2016). *PhysiCell: an Open Source Physics-Based Cell Simulator for 3-D Multicellular Systems*. doi: 10.1101/088773

Giani, A. M., Gallo, G. R., Gianfranceschi, L., y Formenti, G. (2020). Long walk to genomics: History and current approaches to genome sequencing and assembly. *Comput Struct Biotechnol J*, *18*, 9–19.

Goh, W. W. B., Wang, W., y Wong, L. (2017, 06). Why Batch Effects Matter in Omics Data, and How to Avoid Them. *Trends Biotechnol.*, *35*(6), 498–507.

Gonzalez, H., Hagerling, C., y Werb, Z. (2018). Roles of the immune system in cancer: From tumor initiation to metastatic progression. *Genes and Development*, *32*(19-20), 1267–1284. doi: 10.1101/GAD.314617.118

Goode, D. L., Hunter, S. M., Doyle, M. A., Ma, T., Rowley, S. M., Choong, D., ... Campbell, I. G. (2013, jan). A simple consensus approach improves somatic mutation prediction accuracy. *Genome medicine*, *5*(9), 90. Descargado de http://genomemedicine.com/content/5/9/90 doi: 10.1186/gm494

Goya, R., Sun, M. G. F., Morin, R. D., Leung, G., Ha, G., Wiegand, K. C., ... Shah, S. P. (2010, mar). SNVMix: predicting single nucleotide variants from next-generation sequencing of tumors. *Bioinformatics (Oxford, England)*, *26*(6), 730–6. Descargado de http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2832826{&}tool=pmcentrez{&}rendertype=abstract doi: 10.1093/bioinformatics/btq040

Greaves, M., y Maley, C. C. (2012, jan). Clonal evolution in cancer. *Nature*, *481*(7381), 306–13. Descargado de http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3367003{&}tool=pmcentrez{&}rendertype=abstract doi: 10.1038/nature10762

Gurevich, A., Saveliev, V., Vyahhi, N., y Tesler, G. (2013, apr). QUAST: quality assessment tool for genome assemblies. *Bioinformatics (Oxford, England)*, *29*(8), 1072–5. Descargado de http://www.ncbi.nlm.nih.gov/pubmed/23422339

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

Hanahan, D., y Weinberg, R. A. (2011, Mar). Hallmarks of cancer: the next generation. *Cell*, *144*(5), 646–674.

Hansen, N. F., Gartner, J. J., Mei, L., Samuels, Y., y Mullikin, J. C. (2013, jun). Shimmer: detection of genetic alterations in tumors using next-generation sequence data. *Bioinformatics (Oxford, England)*, *29*(12), 1498–503. Descargado de http://bioinformatics.oxfordjournals.org/content/29/12/1498.long doi: 10.1093/bioinformatics/btt183

Hardison, R. C. (2003, Nov). Comparative genomics. *PLoS Biol.*, *1*(2), E58.

Harkness, W. L. (1965, 06). Properties of the extended hypergeometric distribution. *Ann. Math. Statist.*, *36*(3), 938–945. doi: 10.1214/aoms/1177700066

Haynes, W. A., Higdon, R., Stanberry, L., Collins, D., y Kolker, E. (2013). Differential expression analysis for pathways. *PLoS Comput. Biol.*, *9*(3), e1002967.

Herman, P. E., Papatheodorou, A., Bryant, S. A., Waterbury, C. K. M., Herdy, J. R., Arcese, A. A., ... Bloom, O. (2018, 01). Highly conserved molecular pathways, including Wnt signaling, promote functional recovery from spinal cord injury in lampreys. *Sci Rep*, *8*(1), 742.

Hestenes, M. R. (1958). Inversion of matrices by biorthogonalization and related results. *J. Soc. Indust. Appl. Math.*, *6*(I), 91–92.

Hidalgo, M. R., Cubuk, C., Amadoz, A., Salavert, F., Carbonell-Caballero, J., y Dopazo, J. (2016). High throughput estimation of functional cell activities reveals disease mechanisms and predicts relevant clinical outcomes. *Oncotarget*. doi: 10.18632/oncotarget.14107

Hofree, M., Shen, J. P., Carter, H., Gross, A., y Ideker, T. (2013, nov). Network-based stratification of tumor mutations. *Nature methods*, *10*(11), 1108–15. Descargado de http://dx.doi.org/10.1038/nmeth.2651 doi: 10.1038/nmeth.2651

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, *24*(6), 417–441. doi: 10.1037/h0071325

Huang, W., Li, L., Myers, J. R., y Marth, G. T. (2012, feb). ART: a next-generation sequencing read simulator. *Bioinformatics (Oxford, England)*, *28*(4), 593–4. Descargado de http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3278762{&}tool=pmcentrez{&}rendertype=abstract doi: 10.1093/bioinformatics/btr708

Hudson, T. J., Anderson, W., Aretz, A., Barker, A. D., Bell, C., Bernabé, R. R., ... Wainwright, B. J. (2010). International network of cancer genome projects. *Nature*, *464*(7291), 993–998. doi: 10.1038/nature08987

Hunt, M., Kikuchi, T., Sanders, M., Newbold, C., Berriman, M., y Otto, T. D. (2013, May). REAPR: a universal tool for genome assembly evaluation. *Genome Biol.*, *14*(5), R47.

Hyvärinen, A. (1999). Fast and robust fixed-point algorithms for independent component analysis. *IEEE Trans Neural Netw*, *10*(3), 626–634.

Hyvärinen, A. (2013, Feb). Independent component analysis: recent advances. *Philos Trans A Math Phys Eng Sci*, *371*(1984), 20110534.

International Human Genome Sequencing Consortium. (2004). Finishing the euchromatic sequence of the human genome. *Nature*, *431*(7011), 931–45. Descargado de http://www.ncbi.nlm.nih.gov/pubmed/15496913 doi: 10.1038/nature03001

Ivakhno, S., Colombo, C., Tanner, S., Tedder, P., Berri, S., y Cox, A. J. (2017). THapMix: Simulating tumour samples through haplotype mixtures. *Bioinformatics*, *33*(2), 280–282. doi: 10.1093/

bioinformatics/btw589

Jacob, L., Neuvial, P., y Dudoit, S. (2012). More power via graph-structured tests for differential expression of gene networks. *Annals of Applied Statistics*, *6*(2), 561–600. doi: 10.1214/11-AOAS528

Johnson, W. E., Li, C., y Rabinovic, A. (2007, Jan). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, *8*(1), 118–127.

Josephidou, M., Lynch, A. G., y Tavaré, S. (2015). multiSNV: a probabilistic approach for improving detection of somatic point mutations from multiple related tumour samples. *Nucleic acids research*, *43*(9), e61. Descargado de http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4482059{&}tool=pmcentrez{&}rendertype=abstract doi: 10.1093/nar/gkv135

Kanagal-Shamanna, R., Hodge, J. C., Tucker, T., Shetty, S., Yenamandra, A., Dixon-McIver, A., ... Fang, M. (2018, 12). Assessing copy number aberrations and copy neutral loss of heterozygosity across the genome as best practice: An evidence based review of clinical utility from the cancer genomics consortium (CGC) working group for myelodysplastic syndrome, myelodysplastic/myeloproliferative and myeloproliferative neoplasms. *Cancer Genet*, *228-229*, 197–217.

Kanehisa, M., y Goto, S. (2000, Jan). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, *28*(1), 27–30.

Karaman, B., y Sippl, W. (2019). Computational Drug Repurposing: Current Trends. *Curr. Med. Chem.*, *26*(28), 5389–5409.

Kastenhuber, E. R., y Lowe, S. W. (2017). Putting p53 in Context. *Cell*, *170*(6), 1062–1078. Descargado de http://dx.doi.org/10.1016/j.cell.2017.08.028 doi: 10.1016/j.cell.2017.08.028

Kim, H., y Park, H. (2008).
, *30*(2), 713–730.

Kim, J., He, Y., y Park, H. (2014). *Algorithms for nonnegative matrix and tensor factorizations: A unified view based on block coordinate descent framework* (Vol. 58) (n.º 2). doi: 10.1007/s10898-013-0035-4

Kim, S., Jeong, K., Bhutani, K., Lee, J. H., Patel, A., Scott, E., ... Bafna, V. (2013). Virmid: accurate detection of somatic mutations with sample impurity inference. *Genome biology*, *14*(8), R90. Descargado de http://www.ncbi.nlm.nih.gov/pubmed/23987214 doi: 10.1186/gb-2013-14-8-r90

Koboldt, D. C., Zhang, Q., Larson, D. E., Shen, D., McLellan, M. D., Lin, L., ... Wilson, R. K. (2012, mar). VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome research*, *22*(3), 568–76. Descargado de http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3290792{&}tool=pmcentrez{&}rendertype=abstract doi: 10.1101/gr.129684.111

Koltsova, A. S., Pendina, A. A., Efimova, O. A., Chiryaeva, O. G., Kuznetzova, T. V., y Baranov, V. S. (2019). On the Complexity of Mechanisms and Consequences of Chromothripsis: An Update. *Front Genet*, *10*, 393.

Kossenkov, A. V., y Ochs, M. F. (2010). Matrix factorisation methods applied in microarray data analysis. *International Journal of Data Mining and Bioinformatics*, *4*(1), 72–90. doi: 10.1504/IJDMB.2010.030968

Koutrouli, M., Karatzas, E., Paez-Espino, D., y Pavlopoulos, G. A. (2020). A Guide to Conquer the Biological Network Era Using Graph Theory. *Front Bioeng Biotechnol*, *8*, 34.

Kuenzi, B. M., y Ideker, T. (2020, 04). A census of pathway maps in cancer systems biology. *Nat. Rev. Cancer*, *20*(4), 233–246.

Lambert, A. W., Pattabiraman, D. R., y Weinberg, R. A. (2016). Emerging Biological Principles of Metastasis. *Cell*, *168*(4), 670–691. Descargado de http://dx.doi.org/10.1016/j.cell.2016.11.037 doi: 10.1016/j.cell.2016.11.037

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., ... International Human Genome Sequencing, C. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860–921. Descargado de http://www.ncbi.nlm.nih.gov/pubmed/11237011$\backslash$backslash$nhttp://www.nature.com/nature/journal/v409/n6822/pdf/409860a0.pdf doi: 10.1038/35057062

Larson, D. E., Harris, C. C., Chen, K., Koboldt, D. C., Abbott, T. E., Dooling, D. J., ... Ding, L. (2012, feb). SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics (Oxford, England)*, *28*(3), 311–7. Descargado de http://bioinformatics.oxfordjournals.org/content/28/3/311.long doi: 10.1093/bioinformatics/btr665

Lee, D. D., y Seung, H. S. (1999, Oct). Learning the parts of objects by non-negative matrix factorization. *Nature*, *401*(6755), 788–791.

Leinonen, R., Sugawara, H., y Shumway, M. (2011). The sequence read archive. *Nucleic Acids Research*, *39*(SUPPL. 1), 2010–2012.

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, *27*(21), 2987–2993. doi: 10.1093/bioinformatics/btr509

Li, X., Shen, L., Shang, X., y Liu, W. (2015). Subpathway Analysis based on Signaling-Pathway Impact Analysis of Signaling Pathway. *PLoS ONE*, *10*(7), e0132813.

Li, Y., Roberts, N. D., Wala, J. A., Shapira, O., Schumacher, S. E., Kumar, K., ... Zhang, C. Z. (2020, 02). Patterns of somatic structural variation in human cancer genomes. *Nature*, *578*(7793), 112–121.

Lipman, D. J., y Pearson, W. R. (1985, Mar). Rapid and sensitive protein similarity searches. *Science*, *227*(4693), 1435–1441.

Love, M. I., Huber, W., y Anders, S. (2014). Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biology*, *15*, 550. doi: 10.1186/s13059-014-0550-8

Magoc, T., Pabinger, S., Canzar, S., Liu, X., Su, Q., Puiu, D., ... Salzberg, S. L. (2013). GAGE-B: An evaluation of genome assemblers for bacterial organisms. *Bioinformatics*, *29*(14), 1718–1725.

Maley, C. C., Aktipis, A., Graham, T. A., Sottoriva, A., Boddy, A. M., Janiszewska, M., ... Shibata, D. (2017). Classifying the evolutionary and ecological features of neoplasms. *Nature Reviews Cancer*, *17*(10), 605–619. Descargado de http://dx.doi.org/10.1038/nrc.2017.69 doi: 10.1038/nrc.2017.69

Mann, H. B., y Whitney, D. R. (1947). On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, *18*(1), 50 – 60. Descargado de https://doi.org/10.1214/aoms/1177730491 doi: 10.1214/aoms/1177730491

Marcozzi, A., Pellestor, F., y Kloosterman, W. P. (2018). The Genomic Characteristics and Origin of Chromothripsis. *Methods Mol. Biol.*, *1769*, 3–19.

Martincorena, I., y Campbell, P. J. (2015, Sep). Somatic mutation in cancer and normal cells. *Science*,

349(6255), 1483–1489.

Martini, P., Sales, G., Massa, M. S., Chiogna, M., y Romualdi, C. (2013, Jan). Along signal paths: an empirical gene set approach exploiting pathway topology. *Nucleic Acids Res.*, *41*(1), e19.

McGranahan, N., y Swanton, C. (2017, 02). Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future. *Cell*, *168*(4), 613–628.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., ... DePristo, M. A. (2010, Sep). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.*, *20*(9), 1297–1303.

Mikheenko, A., Saveliev, V., y Gurevich, A. (2016). MetaQUAST: Evaluation of metagenome assemblies. *Bioinformatics*, *32*(7), 1088–1090.

Mohammad, S., Sahraeian, E., Liu, R., Lau, B., Podesta, K., Mohiyuddin, M., y Lam, H. Y. K. (2019). somatic mutation detection. *Nature Communications*, 1–10. Descargado de http://dx.doi.org/10.1038/s41467-019-09027-x doi: 10.1038/s41467-019-09027-x

Montaner, D., y Dopazo, J. (2010, Apr). Multidimensional gene set analysis of genomic data. *PLoS ONE*, *5*(4), e10348.

Mootha, V. K., Lindgren, C. M., Eriksson, K. F., Subramanian, A., Sihag, S., Lehar, J., ... Groop, L. C. (2003, Jul). PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.*, *34*(3), 267–273.

Mullikin, J. C. (2014, Sep). The evolution of comparative genomics. *Mol Genet Genomic Med*, *2*(5), 363–368.

Narayanan, D. L., Saladi, R. N., y Fox, J. L. (2010, Sep). Ultraviolet radiation and skin cancer. *Int. J. Dermatol.*, *49*(9), 978–986.

Nikiforov, Y. E., y Nikiforova, M. N. (2011). Molecular genetics and diagnosis of thyroid cancer. *Nature Reviews Endocrinology*, *7*(10), 569–580. Descargado de http://dx.doi.org/10.1038/nrendo.2011.142 doi: 10.1038/nrendo.2011.142

Nowell, P. C. (1976). The Clonal Evolution of Tumor Cell Populations. *Science*, *194*(4260), 23–28.

Oshlack, A., Robinson, M. D., y Young, M. D. (2010). From RNA-seq reads to differential expression results. *Genome Biol.*, *11*(12), 220.

Paatero, P., y Tapper, U. (1994). Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, *5*(2), 111–126.

Petersen, E., Buchner, H., Eger, M., y Rostalski, P. (2017, Apr). Convolutive blind source separation of surface EMG measurements of the respiratory muscles. *Biomed Tech (Berl)*, *62*(2), 171–181.

Petkevicius, K., Virtue, S., Bidault, G., Jenkins, B., Çubuk, C., Morgantini, C., ... Vidal-Puig, A. (2019, 08). Accelerated phosphatidylcholine turnover in macrophages promotes adipose tissue inflammation in obesity. *Elife*, *8*.

Peña-Chilet, M., Esteban-Medina, M., Falco, M. M., Rian, K., Hidalgo, M. R., Loucera, C., y Dopazo, J. (2019, 12). Using mechanistic models for the clinical interpretation of complex genomic variation. *Sci Rep*, *9*(1), 18937.

Poppe, A. B., Wisner, K., Atluri, G., Lim, K. O., Kumar, V., y Macdonald, A. W. (2013, Sep). Toward a neurometric foundation for probabilistic independent component analysis of fMRI data. *Cogn Affect Behav Neurosci*, *13*(3), 641–659.

R Core Team. (2019). R: A language and environment for statistical computing [Manual de software

informático]. Vienna, Austria. Descargado de https://www.R-project.org/

Repsilber, D., Kern, S., Telaar, A., Walzl, G., Black, G. F., Selbig, J., ... Jacobsen, M. (2010, Jan). Biomarker discovery in heterogeneous tissue samples -taking the in-silico deconfounding approach. *BMC Bioinformatics*, *11*, 27.

Rheinbay, E., Nielsen, M. M., Abascal, F., Wala, J. A., Shapira, O., Tiao, G., ... Zhang, C. Z. (2020, 02). Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. *Nature*, *578*(7793), 102–111.

Roth, A., Ding, J., Morin, R., Crisan, A., Ha, G., Giuliany, R., ... Shah, S. P. (2012, apr). JointSNVMix: a probabilistic model for accurate detection of somatic mutations in normal/tumour paired next-generation sequencing data. *Bioinformatics (Oxford, England)*, *28*(7), 907–13. Descargado de http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3315723{&}tool=pmcentrez{&}rendertype=abstract doi: 10.1093/bioinformatics/bts053

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53-65. Descargado de https://www.sciencedirect.com/science/article/pii/0377042787901257 doi: https://doi.org/10.1016/0377-0427(87)90125-7

Sanchez-Vega, F., Mina, M., Armenia, J., Chatila, W. K., Luna, A., La, K. C., ... Mariamidze, A. (2018, 04). Oncogenic Signaling Pathways in The Cancer Genome Atlas. *Cell*, *173*(2), 321–337.

Saraçli, S., Doğan, N., y Doğan, I. (2013). Comparison of hierarchical cluster analysis methods by cophenetic correlation. *Journal of Inequalities and Applications*, *2013*, 1–8. doi: 10.1186/1029-242X-2013-203

Saunders, C. T., Wong, W. S. W., Swamy, S., Becq, J., Murray, L. J., y Cheetham, R. K. (2012, jul). Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics (Oxford, England)*, *28*(14), 1811–7. Descargado de http://www.ncbi.nlm.nih.gov/pubmed/22581179 doi: 10.1093/bioinformatics/bts271

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, *6*(2), 461 – 464. Descargado de https://doi.org/10.1214/aos/1176344136 doi: 10.1214/aos/1176344136

Scrucca, L., Fop, M., Murphy, T. B., y Raftery, A. E. (2016). mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal*, *8*(1), 205–233. Descargado de https://journal.r-project.org/archive/2016-1/scrucca-fop-murphy-etal.pdf

Sheikh, A., Hussain, S. A., Ghori, Q., Naeem, N., Fazil, A., Giri, S., ... Al Tamimi, D. M. (2015). The spectrum of genetic mutations in breast cancer. *Asian Pac. J. Cancer Prev.*, *16*(6), 2177–2185.

Stein-O'Brien, G. L., Arora, R., Culhane, A. C., Favorov, A. V., Garmire, L. X., Greene, C. S., ... Fertig, E. J. (2018). Enter the Matrix: Factorization Uncovers Knowledge from Omics. *Trends in Genetics*, *34*(10), 790–805. Descargado de https://doi.org/10.1016/j.tig.2018.07.003 doi: 10.1016/j.tig.2018.07.003

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., ... Mesirov, J. P. (2005, Oct). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.*, *102*(43), 15545–15550.

Sun, L., Liu, Y., y Beadle, P. J. (2005). Independent component analysis of EEG signals. *Proceedings of the 2005 IEEE International Workshop on VLSI Design and Video Technology, IWVDVT 2005*, 293–296. doi: 10.1109/iwvdvt.2005.1504590

Tabassum, D. P., y Polyak, K. (2015). Tumorigenesis: it takes a village. *Nature Publishing Group*, *15*(8), 473–483. Descargado de http://dx.doi.org/10.1038/nrc3971 doi: 10.1038/nrc3971

Tarca, A. L., Draghici, S., Khatri, P., Hassan, S. S., Mittal, P., Kim, J. S., ... Romero, R. (2009, Jan). A novel signaling pathway impact analysis. *Bioinformatics*, *25*(1), 75–82.

Tate, J. G., Bamford, S., Jubb, H. C., Sondka, Z., Beare, D. M., Bindal, N., ... Forbes, S. A. (2019). COSMIC: The Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Research*, *47*(D1), D941–D947. doi: 10.1093/nar/gky1015

Tavassoly, I., Goldfarb, J., e Iyengar, R. (2018, 10). Systems biology primer: the basic methods and approaches. *Essays Biochem.*, *62*(4), 487–500.

The International HapMap Consortium. (2003, Dec). The International HapMap Project. *Nature*, *426*(6968), 789–796.

The UniProt Consortium. (2017, 01). UniProt: the universal protein knowledgebase. *Nucleic Acids Res.*, *45*(D1), D158-D169.

Therneau, T. M. (2015). A package for survival analysis in s [Manual de software informático]. Descargado de https://CRAN.R-project.org/package=survival (version 2.38)

Tipping, M. E., y Bishop, C. (1999, January). Probabilistic principal component analysis. *Journal of the Royal Statistical Society, Series B*, *21*(3), 611-622. Descargado de https://www.microsoft.com/en-us/research/publication/probabilistic-principal-component-analysis/ (Available from http://www.ncrg.aston.ac.uk/Papers/index.html)

Tomczak, K., Czerwinska, P., y Wiznerowicz, M. (2015). The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp Oncol (Pozn)*, *19*(1A), 68–77.

Torre, L. A., Siegel, R. L., Ward, E. M., y Jemal, A. (2016, Jan). Global Cancer Incidence and Mortality Rates and Trends–An Update. *Cancer Epidemiol. Biomarkers Prev.*, *25*(1), 16–27.

Türkmen, A. C. (2015). A Review of Nonnegative Matrix Factorization Methods for Clustering. , 1–23. Descargado de http://arxiv.org/abs/1507.03194

Turner, N. D., y Lloyd, S. K. (2017, Apr). Association between red meat consumption and colon cancer: A systematic review of experimental results. *Exp. Biol. Med. (Maywood)*, *242*(8), 813–839.

Usuyama, N., Shiraishi, Y., Sato, Y., Kume, H., Homma, Y., Ogawa, S., ... Imoto, S. (2014, aug). HapMuC: somatic mutation calling using heterozygous germ line variants near candidate mutations. *Bioinformatics (Oxford, England)*, btu537–. Descargado de http://bioinformatics.oxfordjournals.org/content/early/2014/09/03/bioinformatics.btu537.long doi: 10.1093/bioinformatics/btu537

Uversky, V. (2013). Posttranslational modification. En S. Maloy y K. Hughes (Eds.), *Brenner's encyclopedia of genetics (second edition)* (Second Edition ed., p. 425-430). San Diego: Academic Press. Descargado de https://www.sciencedirect.com/science/article/pii/B9780123749840012031 doi: https://doi.org/10.1016/B978-0-12-374984-0.01203-1

Uzunlulu, M., Telci-Caklili, O., y Oguz, A. (2016). Association between Metabolic Syndrome and Cancer. *Ann. Nutr. Metab.*, *68*(3), 173–179.

Venkatesan, S., Birkbak, N. J., y Swanton, C. (2017, 02). Constraints in cancer evolution. *Biochem. Soc. Trans.*, *45*(1), 1–13.

Vineis, P., y Wild, C. P. (2014, Feb). Global cancer patterns: causes and prevention. *Lancet*, *383*(9916), 549–557.

Vivarelli, S., Wagstaff, L., y Piddini, E. (2012, jan). Cell wars: regulation of cell survival and proliferation by cell competition. *Essays in biochemistry*, *53*, 69–82. Descargado de http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3707360{&}tool=pmcentrez{&}rendertype=abstract doi: 10.1042/bse0530069

Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz, L. a., y Kinzler, K. W. (2013, mar). Cancer genome landscapes. *Science (New York, N.Y.)*, *339*(6127), 1546–58. Descargado de http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3749880{&}tool=pmcentrez{&}rendertype=abstract doi: 10.1126/science.1235122

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., ... Earl, A. M. (2014). Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE*, *9*(11).

Wang, F., Fang, Q., Ge, Z., Yu, N., Xu, S., y Fan, X. (2012, Mar). Common BRCA1 and BRCA2 mutations in breast cancer families: a meta-analysis from systematic review. *Mol. Biol. Rep.*, *39*(3), 2109–2118.

Wardah, W., Khan, M. G. M., Sharma, A., y Rashid, M. A. (2019, Aug). Protein secondary structure prediction using neural networks and deep learning: A review. *Comput Biol Chem*, *81*, 1–8.

Watkins, T. B. K., y Schwarz, R. F. (2018, 04). Phylogenetic Quantification of Intratumor Heterogeneity. *Cold Spring Harb Perspect Med*, *8*(4).

Wee, Y., Bhyan, S. B., Liu, Y., Lu, J., Li, X., y Zhao, M. (2019, 02). The bioinformatics tools for the genome assembly and analysis based on third-generation sequencing. *Brief Funct Genomics*, *18*(1), 1–12.

Yi, K., y Ju, Y. S. (2018, 08). Patterns and mechanisms of structural variations in human cancer. *Exp. Mol. Med.*, *50*(8), 98.

Yost, S. E., Alakus, H., Matsui, H., Schwab, R. B., Jepsen, K., Frazer, K. A., y Harismendy, O. (2013, aug). Mutascope: sensitive detection of somatic mutations from deep amplicon sequencing. *Bioinformatics (Oxford, England)*, *29*(15), 1908–9. Descargado de http://bioinformatics.oxfordjournals.org/content/29/15/1908.long doi: 10.1093/bioinformatics/btt305

Yu, G., Li, F., Qin, Y., Bo, X., Wu, Y., y Wang, S. (2010). Gosemsim: an r package for measuring semantic similarity among go terms and gene products. *Bioinformatics*, *26*(7), 976-978. doi: 10.1093/bioinformatics/btq064

Zerbino, D. R., Achuthan, P., Akanni, W., Amode, M. R., Barrell, D., Bhai, J., ... Flicek, P. (2018, 01). Ensembl 2018. *Nucleic Acids Res.*, *46*(D1), D754-D761.

Zhu, X., Leung, H. C. M., Wang, R., Chin, F. Y. L., Yiu, S. M., Quan, G., ... Wang, Y. (2015). misFinder: identify mis-assemblies in an unbiased manner using reference and paired-end reads. *BMC Bioinformatics*, *16*(1), 386. Descargado de http://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-015-0818-3