Escola Tècnica Superior d'Enginyeria ETSE-UV

MÁSTER UNIVERSITARIO EN BIOINFORMÁTICA

VNIVERSITAT DE VALÈNCIA

TRABAJO DE FIN DE MÁSTER

METAANÁLISIS FUNCIONAL DE ESTUDIOS DE METILACIÓN DE CÁNCER DE MAMA

AUTOR/A:

ANTONIO MANUEL TRASSIERRA FRESCO

TUTORES:

FRANCISCO GARCÍA GARCÍA
MIGUEL LOZANO IBÁÑEZ

SEPTIEMBRE, 2017

VNIVERSITAT DE VALÈNCIA

Escola Tècnica Superior d'Enginyeria ETSE-UV

MÁSTER UNIVERSITARIO EN BIOINFORMÁTICA

TRABAJO DE FIN DE MÁSTER

METAANÁLISIS FUNCIONAL DE ESTUDIOS DE METILACIÓN DE CÁNCER DE MAMA

AUTOR/A:

ANTONIO MANUEL TRASSIERRA FRESCO

TUTORES:

FRANCISO GARCÍA GARCÍA

MIGUEL LOZANO IBÁÑEZ

TRIBUNAL:

PRESIDENTE/A:

VOCAL 1:

VOCAL 2:

FECHA DE DEFENSA:

CALIFICACIÓN:

Índice de contenido

Abstract

Cáncer es el nombre que se da a un conjunto de enfermedades en las que se observa un descontrolado proceso de división en las células del organismo. En todos los tipos de cáncer (desde los que forman tumores sólidos a los que están en la sangre, llamados leucemias), algunas de las células del organismo empiezan a dividirse continuamente y se diseminan a los tejidos de alrededor, invadiéndolos, en un proceso denominado como metástasis.

El cáncer de mama es el tipo de cáncer con mayor incidencia y mortalidad entre la población femenina a nivel mundial. Dentro del cáncer de mama encontramos varios subtipos con diferentes pronósticos. En este estudio, nos centramos en los subtipos de cáncer de mama basal y dos subtipos suyos (lum-a y lum-b), her-2, brca, carcinoma ductal invasivo (idc) y carcinoma lobular invasivo (ilb).

El principal objetivo de este trabajo es proporcionar una mejor caracterización funcional de esta enfermedad a través de metodologías de análisis novedosas, como el metaanálisis funcional. Para ello evaluamos un conjunto de estudios de metilación en cáncer de mama, obteniendo un marco funcional común a los diversos subtipos de esta patología de interés.

La información del nivel de metilación se resumirá a nivel de gen mediante dos aproximaciones basadas en el promedio y el valor absoluto. Posteriormente realizaremos un análisis de enriquecimiento funcional para cada grupo de genes seguido del correspondiente metaanálisis.

Mediante el uso de esta metodología obtenemos una gran cantidad de funciones comunes que se encuentran alteradas entre los diferentes subtipos de esta enfermedad, como son diversas rutas de señalización como los receptores tirosin-quinasa dependientes de colágeno. Gracias a esta batería de resultados podemos comprobar cómo en estas células se desarrolla la actividad del cáncer. Esto implica una mayor activación del sistema inmune y el reconocimiento de varios *hallmarks* del cáncer a través de distintos procesos biológicos y rutas afectadas.

Todas estas funciones se relacionan con distintos grupos de genes, los cuales posteriormente, podrán ser usados como biomarcadores tumorales para futuros ensayos clínicos en busca de potenciales tratamientos contra la enfermedad.

Introducción

1. El cáncer

El término cáncer hace referencia a un conjunto de enfermedades que pueden afectar a cualquier parte del organismo, con una proliferación masiva y descontrolada de células anormales. Las células cancerosas pueden incluso invadir partes adyacentes del cuerpo o propagarse a otros órganos provocando en última instancia tumores. Esto último recibe el nombre de metástasis, las cuales son la principal causa de la muerte por cáncer ("OMS | Cáncer," 2017).

Estas alteraciones en las células son resultado de interacciones entre los factores genéticos del paciente y tres categorías de agentes externos:

- Carcinógenos físicos, entre los que se encuentran las radiaciones ultravioletas e ionizantes.
- Carcinógenos químicos, como el amianto, los componentes del humo del tabaco, las aflatoxinas (contaminantes de alimentos), etc.
- Carcinógenos biológicos, como algunos virus, bacterias y parásitos.

Además de estos causantes, se debe de tener en cuenta que el cáncer es una enfermedad que tiene mayor incidencia a edades avanzadas. El motivo de este fenómeno se cree que es la acumulación de sucesivos errores en las células tras innumerables ciclos de replicación.

Los tumores a su vez pueden ser de dos tipos, benignos o malignos:

- Tumor benigno. Es un tumor que no se extiende a otras partes del cuerpo, es decir, no produce metástasis. Los tumores benignos pueden ser extirpados y tras esto no se suele observar recurrencia o reincidencia del cáncer. Un tumor benigno puede transformarse en uno maligno.
- Tumor maligno. Es un tumor que puede extenderse a otras partes del cuerpo dada su gran capacidad invasiva (metástasis). Aunque sea extirpado, puede reaparecer de nuevo. Los tumores malignos son causantes de la muerte del enfermo si no son tratados e incluso si el enfermo recibiera un tratamiento óptimo, no tiene asegurada la desaparición y cura del cáncer (Mareel & Leroy, 2003). Sin embargo, en la actualidad, el índice de curación de algunos tipos de cáncer se encuentra en torno al 95% de los casos, llegando a ser este índice mayor que el de algunas enfermedades infecciosas y desórdenes metabólicos.

El proceso de metástasis ocurre, por tanto, en aquellas líneas celulares tumorales que son más fuertes desde un punto de vista darwiniano. Las células cancerosas deben sobrevivir en un entorno hostil como es el cuerpo del individuo en donde se han originado, por tanto, estas células precisan de una gran heterogeneidad genética que viene dada por la inestabilidad genómica inherente a las células cancerosas y que en última instancia aumenta la capacidad de estas líneas celulares de adquirir competencia metastásica.

La integridad del ADN en las líneas celulares cancerosas puede verse comprometida de varias maneras como son fallos en los "*checkpoints*" del ciclo celular, aberraciones en los genes encargados de la reparación del ADN, etc. Sin ir más lejos, la mitad de los cánceres en seres humanos sufren la pérdida del gen supresor de tumores p53, el cual codifica para

una proteína interna que responde a los daños del ADN eliminando la célula que los contenga (Massagué, 2008).

La figura 1 ilustra una representación gráfica de la capacidad de metástasis que tienen los tumores malignos.

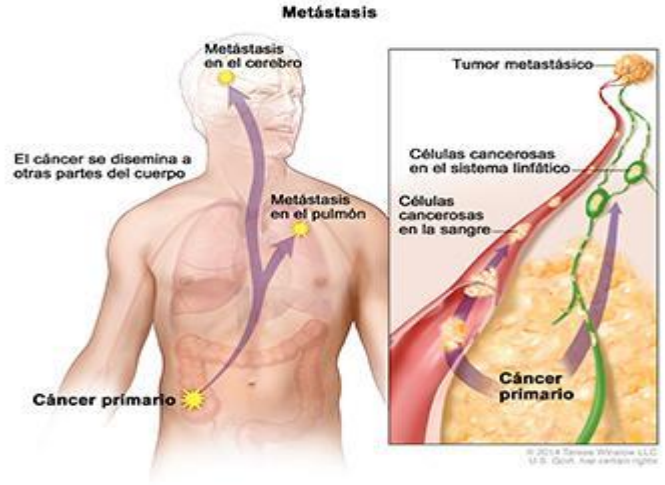

Figura 1. Mecanismo de metástasis. Las células cancerosas son transportadas por la sangre y la linfa hasta nuevos lugares donde producirán un nuevo tumor de origen metastásico.

1.1 La creciente incidencia del cáncer

Actualmente el cáncer tiene un gran impacto en la sociedad. Es la enfermedad con el mayor índice de mortalidad en muchos países, sobre todo en aquellos que poseen un mayor índice de desarrollo humano (HDI, Human Development Index). Adicionalmente, el número de nuevos casos que se diagnostican cada año es mayor con una clara tendencia ascendente según diversos estudios (Bray, Jemal, Grey, Ferlay, & Forman, 2012).

Los tipos de cáncer con mayor incidencia y mortalidad anuales son el de pulmón, el de mama y el de colón. Tanto el cáncer de pulmón como el de colon son comunes a ambos sexos, pero el cáncer de mama en su mayoría solo en mujeres.

El cáncer de mama es el principal tipo de cáncer diagnosticado en las mujeres en regiones con una alta HDI, siendo la cuarta en regiones pobres con una baja HDI. Este tipo de cáncer originó sólo en el año 2008 el diagnóstico de 1.400.000 nuevos casos de mujeres con esta enfermedad. Sin embargo, su mortalidad no es tan elevada como la incidencia con la que se presenta, debido en gran parte a diversos tipos de tratamientos que existen hoy en día, donde encontramos la cirugía, el tratamiento hormonal y la radioterapia entre otros.

En la figura 2 se representan los distintos niveles de incidencia de los tipos de cáncer más comunes, así como sus índices de incidencia acumulativo por edad, índice de mortalidad e índice de mortalidad acumulativo por edad. Además, en la figura 3 se presenta un mapa que muestra la distribución de la incidencia de los tipos de cáncer a nivel mundial en las mujeres, pudiendo observar cómo el cáncer de mama tiene una incidencia prácticamente global.

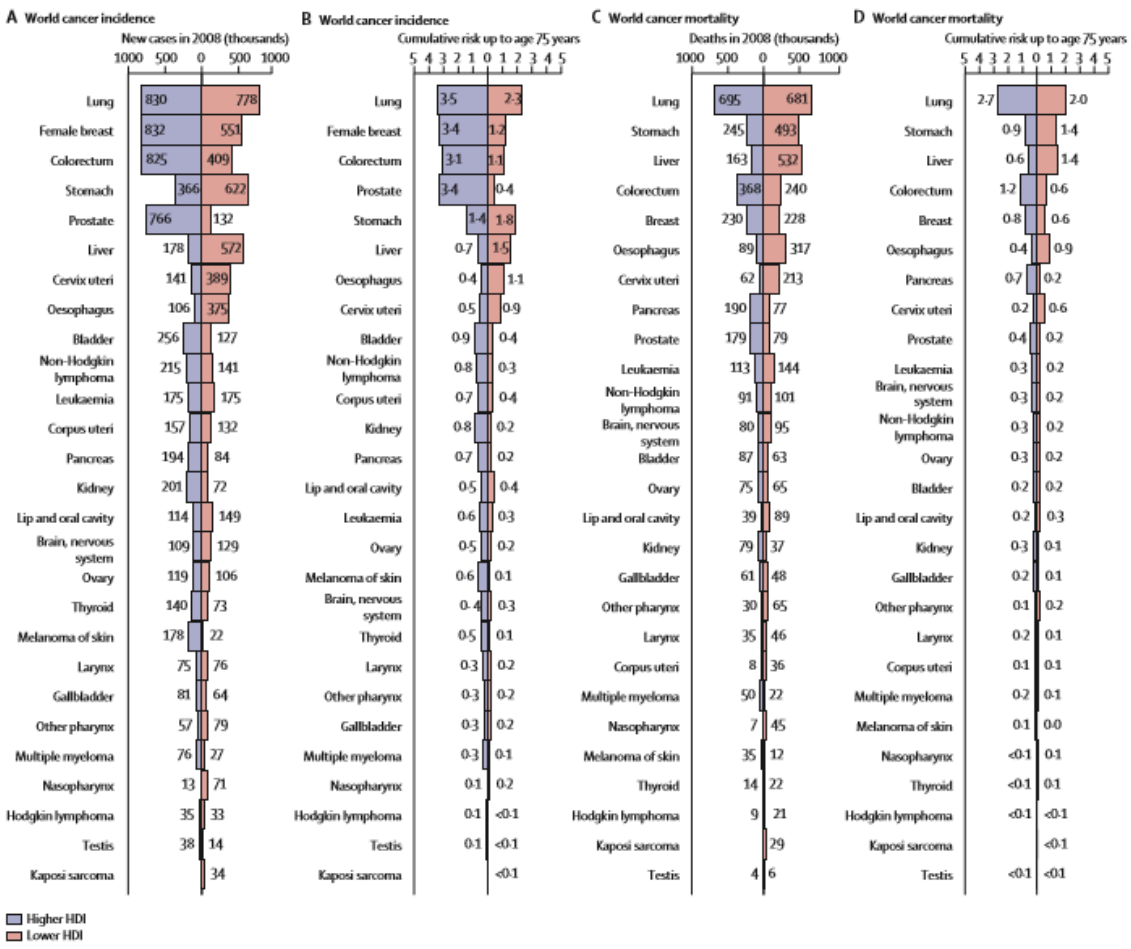

Figura 2. Distribución de la incidencia y mortalidad de diversos tipos de cáncer en 2008.

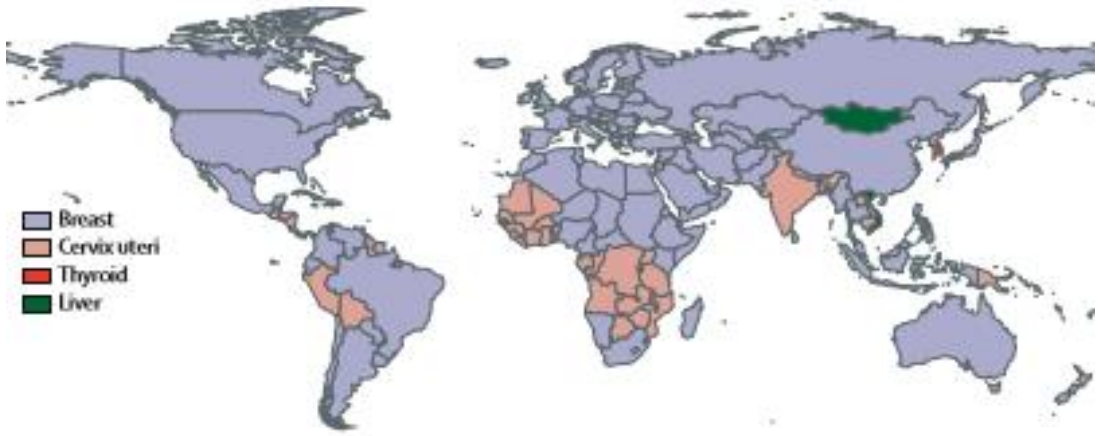

Figura 3. Distribución geográfica de la incidencia de los principales tipos de cáncer que afectan a mujeres. Como se puede comprobar, el cáncer de mama está presente en la mayoría de las áreas geográficas.

1.2 El cáncer de mama

El cáncer de mama es un tipo de cáncer que, como se ha citado anteriormente, afecta a mujeres en su mayoría. Este cáncer provoca un desorden en el ciclo celular de las células pertenecientes a los tejidos de las glándulas mamarias desencadenando una continuidad anormal del ciclo celular. Además, no presenta un patrón de herencia claro, dado que más del 90% de los casos de cáncer de mama han sido esporádicos.

En la actualidad se han realizado diversos hallazgos sobre el cáncer de mama, viéndose entre otras cosas, que no es una enfermedad exclusiva del ser humano. Y es que podemos encontrar este cáncer en otros mamíferos del reino animal, como es el caso del ratón doméstico o *Mus domesticus*. Otros hallazgos realizados sobre este tipo de cáncer es que algunos factores que lo provocan, generan un incremento de mutaciones en el ADN y un reparo disminuido del mismo (mutaciones en BRCA1 y BRCA2, dos genes que producen proteínas supresoras de tumores).

Hasta el día de hoy, son varios los subtipos de cáncer de mama que se conocen. En el presente estudio se recogen varios de ellos como son:

- El carcinoma ductal invasivo. Se da en el 90% de los casos, las células que producen este cáncer son las encargadas de tapizar los conductos que durante la lactancia llevan la leche desde los acinos glandulares (donde ésta se produce) hasta los conductos galactóforos, situados detrás de la areola y el pezón, donde se acumula para salir al exterior (Lugones Botell, Miguel ; Ramírez Bermúdez, 2009).

- El carcinoma lobular invasivo. Este es el segundo tipo de cáncer de mama más común. Recibe su nombre por tener capacidad invasiva (propagarse a tejidos adyacentes al que se ha generado) y por haberse originado en los lobulillos productores de leche los cuales vacían su contenido en los conductos que llevan la leche al pezón.

- El cáncer de mama HER-2 positivo. Esta variedad de cáncer de mama presenta un genotipo negativo para los receptores de estrógenos, por lo que conlleva un mayor riesgo al no poder paliar este tipo de cáncer con algunos tratamientos hormonales. Sin embargo en los últimos años se han introducido nuevos fármacos como el trastruzumab, el cual es un anticuerpo monoclonal que se une a los receptores de HER-2 inhibiendo la señal de proliferación que manda este factor de crecimiento (Cmi, 2007).

- El cáncer de mama de tipo basal, también conocido este último como triple negativo. Este tipo de cáncer mamario es incapaz de responder a tratamientos hormonales o con anticuerpos debido a que presenta negatividad para los receptores de estrógenos, los progesterónicos y los receptores de HER-2. Sin embargo, tiene una buena respuesta frente a tratamientos quimioterapéuticos.

- Dos subtipos de cáncer de mama luminal, llamados lum-A y lum-B que fueron descubiertos en un reciente estudio (Stefansson et al., 2015). Los cánceres de mama de tipo luminal presentan una alta expresión de receptores de estrógenos por lo que la terapia hormonal funciona bastante bien en ellos. En consecuencia, es el

subtipo de cáncer de mama con mejor pronóstico. El tipo lum-A se relaciona con un patrón de metilación heterogéneo a lo largo del genoma, en comparación con lum-B que presenta un patrón de metilación mayor que el que tienen individuos sanos en islas CpG.

Hoy en día, existen muchas maneras para estudiar el cáncer y diversos tipos de enfermedades, siendo la más común los estudios de ARN-Seq. No obstante, hay otro tipo de metodología que envuelve a una parte quizás más desconocida de nuestro ADN, el epigenoma. A través de estudios epigenéticos (donde destacan sobre todo los estudios de metilación) podemos llegar a la comprensión de los mecanismos que están actuando en diversos tipos de enfermedades como es el cáncer de mama.

2. La epigenética, la otra cara de la genética

2.1 ¿Qué es la epigenética?

El surgimiento del término epigenética hace cerca de un siglo, cambió la forma de trabajar de investigadores y otros profesionales, los cuales buscaban pistas sobre la alteración de los genes siguiendo los métodos convencionales, aunque con la idea de que había algo más, algo que podía provocar cambios en los genes más allá de la propia secuencia génica. Epigenética significa, literalmente, además de los cambios en la secuencia génica, aunque ha ido evolucionando hasta incluir cualquier proceso que altera la actividad genética sin alterar la secuencia del gen. Las modificaciones epigenéticas pueden ser transmitidas a células hijas o ser revertidas.

Hoy en día, una gran variedad de enfermedades, trastornos e indicadores de salud tienen una evidencia de regulación epigenética, esto incluye, entre otros, a casi todos los tipos de cáncer, disfunción cognitiva y enfermedades respiratorias, cardiovasculares, autoinmunes y neurocomportamentales.

Algunos de los vectores que se conocen o se sospecha que puedan estar tras estos procesos incluyen metales pesados, pesticidas, hormonas, radioactividad, virus, el humo del tabaco, etc. En los últimos años los estudios basados en el epigenoma han empezado a sonar con fuerza, como uno de los mecanismos para la comprensión de aquellos cambios que no podemos apreciar simplemente observando la secuencia de ADN, además, estos mecanismos son un conocimiento necesario para muchos aspectos de la genética moderna, como las células madre, la clonación, el envejecimiento, la conservación de las especies, la evolución, la biología sintética y la agricultura (Weinhold, 2006).

2.2 Mecanismos epigenéticos

Entonces, ¿cuáles son los mecanismos bajo los cuales actúa la epigenética?

En la actualidad se conocen múltiples mecanismos, entre ellos cabe destacar los de metilación, acetilación, fosforilación, ubiquitinación y sumoilación (un tipo de modificación covalente de proteínas llevada a cabo en las células de diversos organismos mediante pequeñas proteínas llamadas SUMO, *small ubiquitin-like modifier*). Otros tipos de mecanismos epigenéticos están siendo actualmente estudiados y debatidos entre la comunidad científica. El ritmo de crecimiento de estudios de metilación es cada vez mayor gracias en parte a las nuevas tecnologías que nos acontecen hoy en día.

2.2.1 La metilación del ADN

Uno de los mecanismos epigenéticos mejor conocidos en la actualidad y en el que se centran los estudios de este trabajo es el mecanismo de metilación del ADN, el cual se descubrió por primera vez en una muestra de cáncer humano en el año 1983. En esencia, este mecanismo consiste en la incorporación o substracción de grupos metilo (CH_3) preferentemente en sitios donde varias bases de citosina aparecen consecutivamente. Cuando este grupo metilo se une a las citosinas hablamos de 5-metilcitosinas. Usualmente estas citosinas metiladas siempre se encuentran adyacentes a un nucleótido de guanina, lo que llamamos sitio CpG (la p hace referencia a que los nucleótidos están enlazados por un fosfato).

Este proceso está mediado por enzimas ADN metiltransferasas, las cuales forman una familia de enzimas cuyos deberes varían desde actuar *de novo* colocando el patrón de metilación inicial en la secuencia de ADN o en cambio, ejecutar labores de mantenimiento, copiando la metilación existente en una hebra de ADN a su nueva hermana tras el proceso de replicación del ADN (Phillips, 2008). En la **figura 4** se muestra el cambio que se induce en la posición 5 de la citosina, la cual pasa a ser 5-metilcitosina.

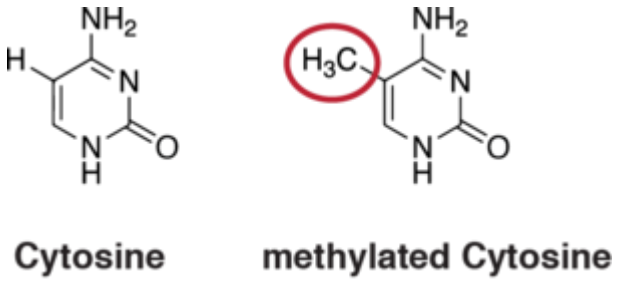

Figura 4. La modificación en la posición 5' de la citosina con un grupo metilo la transforma en 5-metilcitosina.

En mamíferos, la metilación puede ser observada de manera esporádica y escasa a través de varios sitios CpG distribuidos por todo el genoma o bien, el caso contrario, en islas CpG (las cuales son regiones de ADN donde solo encontramos estos nucleótidos y que forman el 40% de los promotores de los genes de mamíferos) y en ciertos tramos de aproximadamente 1 kilobase de longitud donde también podemos encontrar sitios CpG. La metilación de regiones inadecuadas puede desencadenar una respuesta inapropiada como el silenciamiento de determinados tipos de genes, como los genes supresores de tumores en las células cancerosas.

Los patrones de metilación varían entre individuos y especies, por ejemplo, en invertebrados, a diferencia de lo que se ha comentado previamente en mamíferos, el patrón de metilación es de tipo mosaico. Pero ambos tienen una cosa en común y es que las regiones metiladas se encuentran sobre todo cerca de los promotores de los genes lo que da una idea del papel regulador que la metilación debe tener en el funcionamiento de estos.

2.2.2 Metilación del ADN y su rol en la expresión génica

Durante muchos años se ha extendido el pensamiento de que la metilación del ADN juega un papel crucial en la expresión génica y que, de alguna manera, bloqueaba la unión de los factores de transcripción a las zonas metiladas. Sin embargo, actualmente el rol exacto de la metilación en la expresión génica es algo incierto dado que existen algunas variaciones.

No obstante, hay cierto consenso en que es esencial para la diferenciación celular y el desarrollo embrionario. Para probar esto, se han realizado muchos estudios a lo largo de los años, y han demostrado que existe una diferencia en los patrones de metilación entre las distintas células de un mismo individuo, además de que a medida que un promotor se encuentra más metilado, menor es la transcripción de su gen o incluso puede que no se transcriba.

De esta forma, individuos de la misma especie suelen compartir un patrón de metilación similar, con algunas variaciones, pero sí que existe un patrón de metilación distinto entre los tipos celulares que forman nuestro organismo.

2.2.3 Metilación del ADN y su asociación con enfermedades

Dado el rol que tiene la metilación en la expresión y estabilidad génica, parece claro que fallos o errores en este proceso pueden dar lugar a consecuencias catastróficas para la célula, las cuales pueden desembocar en enfermedades.

Actualmente hay muchos estudios e investigaciones centradas en este tipo de enfermedades relacionadas con los patrones de metilación. Por lo que los resultados de éstas serán de un gran valor para la posterior puesta en marcha de un posible tratamiento o prevención.

Añadir que, los mayores esfuerzos se están invirtiendo en la investigación del cáncer y los genes supresores de tumores. Estos genes supresores de tumores están silenciados normalmente en el cáncer debido a la hipermetilación que sufren en sus regiones promotoras. Esto choca con la hipometilación general que presentan las líneas celulares cancerosas, donde gracias a esto y a la hipermetilación de zonas que tienen que ver con regulación del ciclo celular y apoptosis consiguen aumentar en conjunto la propagación de la metástasis cancerosa. En la **figura 5** podemos ver un esquema de la metilación del ADN que sirve de ejemplo para visualizar su relación con el cáncer.

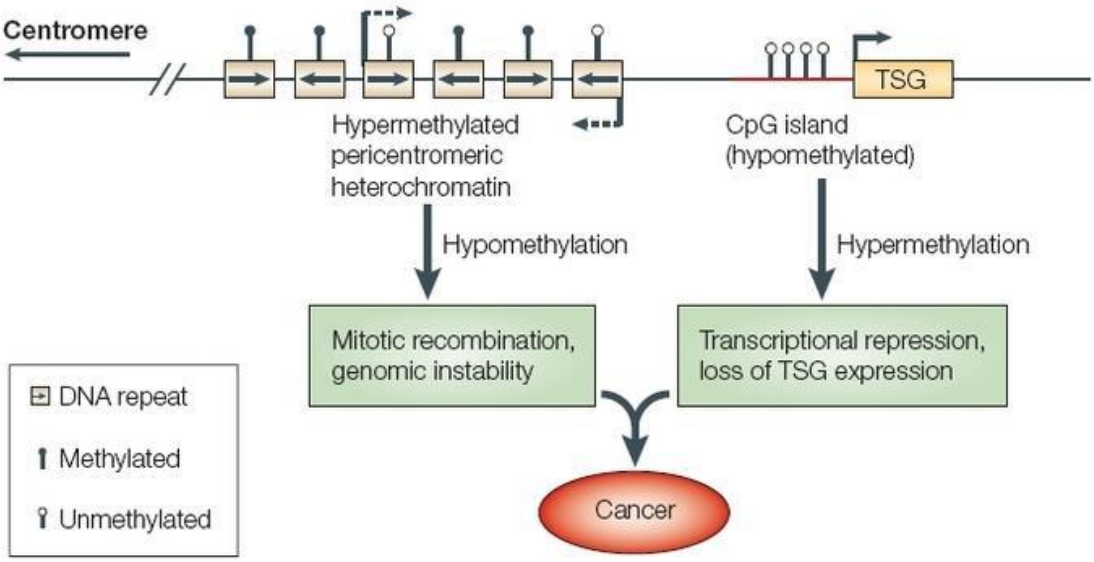

Figura 5. La imagen representa una región genómica de una célula normal. Ésta contiene bastantes regiones repetidas de heterocromatina pericentomérica hipermetilada y un gen supresor de tumores activo (TSG) transcripcionalmente que está asociado con una isla CpG (marcada en rojo) que se encuentra hipometilada. En las células cancerosas, las regiones ricas en repeticiones de heterocromatina se encuentran hipometiladas, lo que contribuye a la inestabilidad genómica (un rasgo común en células tumorales) provocando un incremento de los procesos de recombinación mitótica. De esta manera, también se producen eventos de metilación "de novo" en células cancerosas que pueden silenciar a genes supresores de tumores como el de la imagen. De esta manera, si la isla CpG asociada a TSG se hipermetilara, este gen se silenciaría provocando que no se transcribiera. Estos cambios en la metilación son bastante conocidos por ser eventos tempranos en génesis de los tumores. © 2005 Nature Publishing Group Robertson, K. DNA methylation and human disease. Nature Reviews Genetics 6, 598.

Objetivos e hipótesis iniciales

El presente trabajo consiste en un metaanálisis funcional de diversos estudios de cáncer de mama que comparten un mismo diseño experimental, esto es, control/sano vs caso/enfermedad.

A priori, cabría esperar que encontráramos patrones de metilación claros en las muestras correspondientes a los diversos tipos de cáncer de mama que apuntasen a un silenciamiento de genes supresores de tumores, control del ciclo celular, etc. Esto se interpretaría como una diferencia de metilación significativa entre los dos grupos comparados.

El objetivo principal de este trabajo es la realización de una caracterización de las funciones y mecanismos moleculares de esta enfermedad. Para ello, tendremos que descubrir si hay afectadas funciones comunes entre los distintos estudios evaluados, así como su identificación y descripción del papel que tienen en el cáncer de mama. Además, este abordaje nos permite ir "hacia atrás" y detectar biomarcadores que estén implicados en las funciones con una común y significativa sobrerrepresentación en todos los estudios (bien en los pacientes con la enfermedad o en los controles).

Para la consecución de este objetivo, se evaluarán diferentes abordajes para resumir la información del nivel de metilación a información a nivel de gen y el impacto que éstos producen en los resultados funcionales generados.

Posteriormente se elaborará un análisis de enriquecimiento funcional por grupos de genes para cada una de las comparaciones realizadas cuyo resultado servirá como entrada de datos al metaanálisis funcional.

Materiales y métodos

El esquema de trabajo que seguimos en el presente estudio se encuentra recogido de manera esquemática en la figura 6.

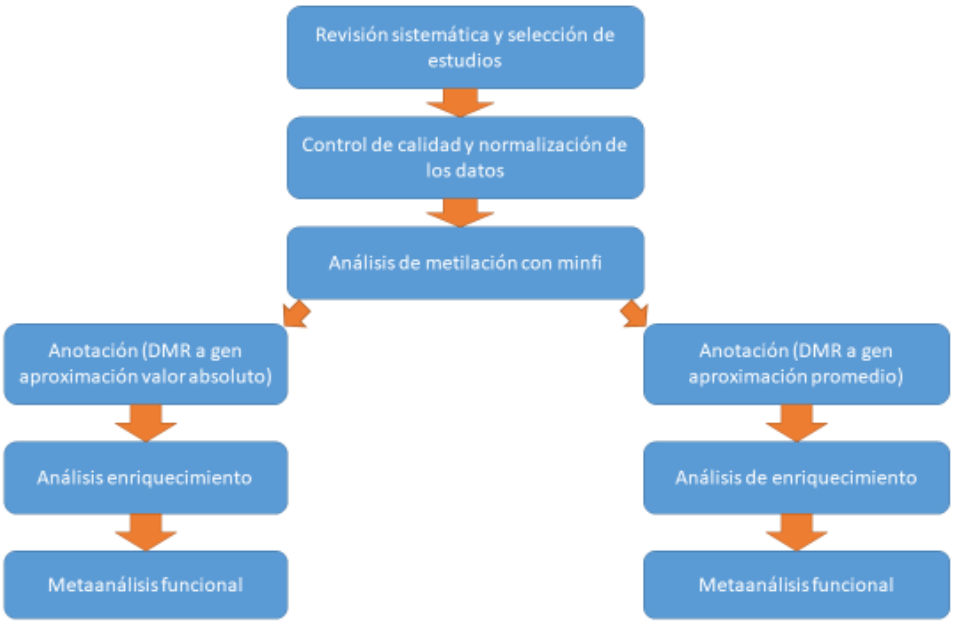

Figura 6. Línea de trabajo que seguimos en el presente estudio para el análisis de datos de metilación funcional y su posterior metaanálisis.

A continuación se desarrollan cada uno de los pasos del esquema de la figura 6.

1. Revisión sistemática y selección de estudios de metilación

La primera parte del presente estudio consiste en una revisión sistemática de estudios de metilación que se encuentran en el repositorio GEO (Gene Expression Omnibus) del NCBI (National Center for Biotechnology Information). Es un repositorio público que fue generado en el año 2000, el cual almacena datos de distintas tecnologías de alto rendimiento que han sido proporcionados por investigadores para su difusión.

La validez de un buen metaanálisis depende en gran medida de la identificación y selección de los estudios originales. Los criterios de inclusión y exclusión de los estudios, un diseño muestral adecuado y la valoración de la calidad de los estudios son elementos necesarios para obtener unos resultados robustos e interpretables en el metaanálisis. Por lo tanto, es necesario que estos estudios posean diseños experimentales similares, así que buscamos aquellos que tengan un diseño “caso vs control”. En caso de que el estudio tenga múltiples comparaciones separaremos éstas en varias comparaciones entre los distintos tipos de tumores frente a los controles (es decir, estudio 1: tumor tipo 1 vs controles, tumor tipo 2 vs controles, etc).

El filtrado que se llevó a cabo para los estudios consta de varias partes, siendo la primera de ellas la selección del objeto de la investigación del cual trataban estos. La elección fue recopilar estudios sobre cáncer de mama, una enfermedad desgraciadamente cada vez más presente en la actualidad.

El siguiente criterio de filtrado fue la selección de la plataforma donde se llevó a cabo los experimentos recogidos en los estudios. Para ello tras una búsqueda de las distintas plataformas que ofrecen soporte para estudios de metilación, decidimos optar por el chip 450K de metilación para Homo Sapiens de Illumina (Illumina HumanMethylation450 BeadChip). Esta plataforma fue seleccionada por ofrecer una gran resolución (cerca de 450000 sitios de metilación conocidos en el genoma) a un coste muy asequible y por su amplia utilización para este tipo de estudios (Kurdyukov & Bullock, 2016; Sun, Cunningham, Slager, & Kocher, 2015).

El conocimiento del funcionamiento de esta tecnología es vital para saber qué datos son los que estamos obteniendo y sobre los que estamos trabajando. Esta información es la que se detalla en el siguiente apartado.

1.1 Infinium HumanMethylation450 BeadChip Kit

La tecnología elegida en la selección de los estudios fue el chip 450k de Illumina, el cual nos proporciona una medida cuantitativa de la metilación que tiene lugar a nivel de posición en un sitio CpG, pudiendo estudiar más de 485000 sitios de metilación en cada una de las muestras que componen el experimento (cada sonda se corresponde con una de estas posiciones). El chip puede analizar hasta 12 muestras por array y en dos canales diferentes (rojo y verde). Además, cuenta con un protocolo de trabajo propio.

Estos kits, son una combinación de una completa y exhaustiva cobertura seleccionada por expertos contratados por Illumina que incluyen distintos elementos como: el 99% de las secuencias de referencia de los genes con una media de 17 sitios CpG por cada región génica distribuidos a lo largo del promotor, el 95% de las islas CpG (con una cobertura adicional en las “costas” de las islas y las regiones que las flanquean), sitios de metilación CpG que se encuentran fuera de las islas CpG y sitios con una metilación diferencial identificada en estudios de tejidos tumorales vs normales (con múltiples tipos de cáncer), entre otros. Estas características potencian la idoneidad de este kit para estudiar el epigenoma. En la figura 7 podemos apreciar el aspecto que tienen estos chips.

Figura 7. Pareja de chips 450K de Illumina.

Este chip consigue detectar la metilación de las bases de citosina en el genoma gracias al tratamiento de este previamente con bisulfito. De esta manera las citosinas que no están

metiladas, al ser tratadas con bisulfito pasarán a ser uracilo mientras que las citosinas que si se encuentran metiladas permanecerán sin cambios.

Gracias a esta reacción química y con la ayuda de dos tipos de sondas diseñadas específicamente para este chip podemos obtener los valores de intensidad de señal de las muestras.

Las dos sondas son la "bead U" y la "bead M", haciendo referencia a sitios que no estén metilados y a sitios que sí lo estén respectivamente (unmethylated y methylated).

El funcionamiento lo podemos ver de manera esquemática en la figura 8.

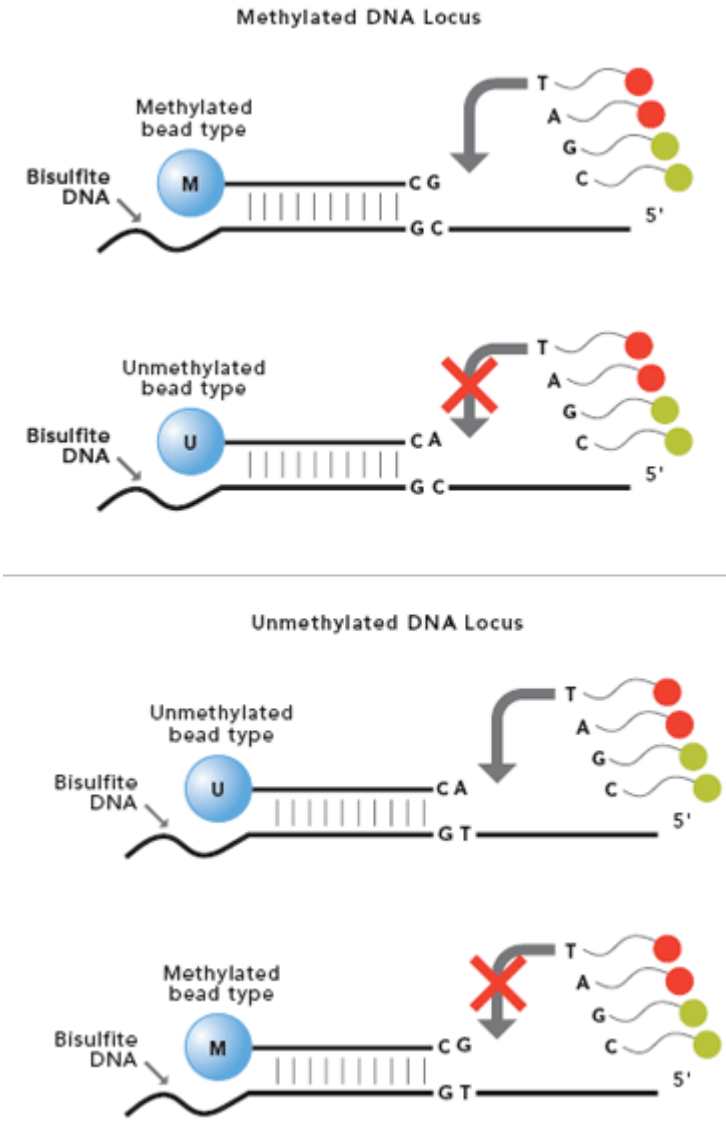

Figura 8. Funcionamiento del chip 450K.

Como se puede observar en la figura 8, los locus metilados hibridan con las sondas de tipo M ya que sus citosinas permanecen intactas, sin embargo, los locus que no presentan metilación no pueden hibridar con las sondas tipo M debido a una incompatibilidad en las bases, en cambio, sí que pueden hacerlo con los locus tipo U.

Posteriormente gracias a la introducción de ddNTPs (didesoxinucleótidos, nucleótidos que carecen de un grupo 3-hidroxilo) marcados se produce la fluorescencia. De esta manera podemos obtener el nivel de metilación de un locus gracias al ratio de la intensidad de señal metilada frente al ratio de la intensidad de señal no metilada.

Dependiendo del diseño de la sonda que usemos, la señal será detectada en diferentes colores:

- Para el diseño de **tipo I,** las dos señales serán medidas en el mismo color. Una sonda se hará cargo de la señal metilada y otra de la no metilada.
- En el diseño de **tipo II**, solo se usa una sonda, de manera que la señal metilada es medida por la intensidad verde y la señal no metilada es medida por la intensidad roja.

En la figura 9 se esquematizan estos diseños.

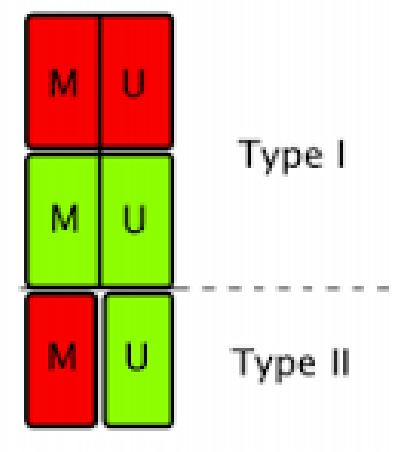

Figura 9. Diseño de las sondas que utiliza el chip 450k de Illumina, en el diseño de tipo I las señales son medidas en el mismo color mientras que en el tipo II la señal metilada es medida por la intensidad roja y la no metilada por la verde.

1.2 Selección de estudios

Los criterios iniciales de selección fueron el tipo de cáncer y la plataforma tecnológica empleada. A continuación ajustamos algunos pequeños filtros en el buscador de GEO. La "query" utilizada para la obtención de los estudios fue las que ilustra la figura 10.

Search	Add to builder	Query	Items found	Time
#7	Add	Search (((breast cancer) AND Homo Sapiens[Organism]) AND GPL13534) AND methylation Filters: Series	49	11:28:45

Figura 10. Parámetros para el filtrado de los diversos estudios de cáncer de mama que cumplen con los criterios que estamos buscando. En la búsqueda se obtienen 49 resultados correspondientes a la columna Items found.

A través de esta búsqueda se obtuvieron 49 estudios candidatos. La selección se refinó mediante un filtrado manual que incluía la revisión de la descripción de cada estudio, comprobando que presentaran el mismo diseño experimental y que además tuviesen toda

la información que necesitamos para llevar a cabo el estudio (identificación de las muestras, de las señales de intensidad metilada y no metilada, etc.).

2. Procesamiento y análisis primario de los datos

Para el análisis de los datos del Illumina HumanMethylation450 BeadChip (al cual llamaré chip 450k para abreviar) existen diversos paquetes en R que nos proporcionan una visión esquematizada del flujo de trabajo que necesitaremos realizar para trabajar con datos de metilación. Para llevar a cabo este análisis se escogió el paquete *minfi* disponible en Bioconductor en https://bioconductor.org/packages/release/bioc/html/minfi.html.

Este paquete incluye funcionalidades para preprocesar, tomar medidas del control de calidad de las muestras e identificar sitios de metilación diferencial.

2.1 Descarga de los datos sin procesar procedentes de GEO

Tras la selección de los estudios participantes en el proyecto, se descargaron los datos sin procesar del repositorio GEO utilizando el software R (Development Core Team, 2009).

Los ficheros con los que espera trabajar *minfi* para realizar el análisis de expresión diferencial tienen la extensión ".IDAT". Son archivos con un formato propio de la casa comercial del chip 450k, es decir, Illumina.

Desafortunadamente, en GEO, es muy extraño conseguir encontrar que los responsables de los estudios suban a la plataforma este tipo de archivo. En su lugar nos solemos encontrar con los datos sin procesar en formato ".txt" o ".csv". Esto supuso algunos ajustes en la metodología de trabajo con el paquete *minfi*, el cual espera recibir los datos sin procesar en formato ".IDAT", para convertirlos y normalizarlos en una matriz donde se representan las distintas intensidades de señal tanto metilada como no metilada.

Al no poder obtener los datos en forma de ".IDAT", procesamos y normalizamos los datos antes de proporcionárselos a *minfi*. En los apartados que siguen, se describen los ajustes realizados.

Tras la obtención de los datos sin procesar, se realizó la evaluación de la calidad de los mismos.

2.2 Control de calidad de los datos

Minfi nos ofrece un control de calidad de las muestras que componen nuestros experimentos. Esta evaluación se ejerce sobre los datos sin procesar y consiste en estudiar el logaritmo de la media de las señales tanto del canal metilado como las del canal no metilado. Las muestras que tengan una buena calidad tenderán a agruparse en clústeres con un logaritmo de su media muy similar. El punto de corte que determina si la calidad de una muestra es buena o mala varía en función de cada estudio.

En la figura 11 queda reflejado cómo debieran verse los logaritmos de las medias de intensidad de las señales de muestras con una buena calidad.

	mMed	uMed
	<numeric>	<numeric>
5723646052_R02C02	11.69566	11.82058
5723646052_R04C01	11.99046	11.95274
5723646052_R05C02	11.55603	12.05393
5723646053_R04C02	12.06609	12.09276
5723646053_R05C02	12.23332	12.08448
5723646053_R06C02	11.36851	11.60594

Figura 11. Ejemplo de datos sin procesar con buena calidad.

Además de esto, se realizaron análisis de clustering y análisis de componentes principales (PCA) para terminar de valorar la calidad y reducir sesgos producidos por alguna comparación que aportase mucha variabilidad.

Si observamos que la calidad de los datos sin procesar es buena, procederemos a la normalización de los mismos.

2.3 Normalización de los datos

La normalización de los datos es una parte importante de cualquier análisis. El objetivo de la normalización es identificar y eliminar variaciones sistemáticas (llamadas comúnmente "ruido") conservando la señal biológica. Mediante la normalización se pretende asegurar que las diferencias en intensidad realmente reflejan la expresión diferencial de los genes (o la metilación diferencial de ciertas regiones en el genoma en nuestro caso) y que no haya sesgos artificiales debidos a factores técnicos.

El paquete *minfi* dispone de varias herramientas para normalizar nuestros datos, pero lamentablemente para esto sólo admite datos en formato ".IDAT" por lo que tuvimos que normalizar los datos de una manera externa a la proporcionada por el paquete. Para ello normalizamos los datos sin procesar a través del método de normalización por cuantiles, método muy similar al que ofrecen ellos y que denominan normalización funcional, el cual es el propio método de normalización por cuantiles con algunas consideraciones más.

La normalización por cuantiles es un método de normalización muy extendido en Bioinformática que asume que todos los arrays de nuestro experimento tienen la misma distribución. Consiste en hacer que dos o más distribuciones sean idénticas en propiedades estadísticas, de manera que buscamos una escala común entre ellas.

En la figura 12 podemos ver una explicación de cómo se lleva a cabo.

Figura 12. Normalización por cuantiles La normalización por cuantiles se lleva a cabo de la siguiente manera: Primero, ordenamos las columnas de la matriz de intensidad (X) de menor señal a mayor señal, pasamos a tener Xsort. Segundo, calculamos las medias de cada una de las filas y sustituimos los valores de la matriz ordenada por las medias de cada una de las filas. Por último, restablecemos el orden original que tenían en la primera matriz (X).

Una vez que tenemos los datos normalizados volvemos a ejecutar sobre estos un análisis exploratorio para evaluar la calidad de los mismos (PCA y clúster). Posteriormente podemos prepararlos para el análisis de metilación diferencial con *minfi*.

2.4 Análisis de metilación diferencial

El análisis de metilación diferencial llevado a cabo con *minfi* requiere dos matrices, una donde tendremos nuestros datos normalizados y la siguiente donde especificaremos las muestras que son controles y las que son casos.

Es conveniente conocer cierto tipo de terminología que se usa en el paquete *minfi* y cuando se trabaja con datos de metilación:

- **Valores Beta**. El valor beta hace referencia a un estadístico que se obtiene tras la siguiente operación: $\beta = \mathrm{M}/(\mathrm{M} + \mathrm{U} + 100)$ donde M es el nivel de intensidad de la señal metilada para una sonda dada y U el nivel de señal no metilada para la misma sonda. 100 hace referencia a una constante que viene marcada por defecto en el paquete.

  Este estadístico es importante ya que es con el que se realizarán las subsecuentes operaciones para determinar si un sitio (sonda) se encuentra diferencialmente metilado entre las dos condiciones que se comparan.

- **Valores M**. El valor M es otro estadístico que se suele usar. Se obtiene mediante la operación: $Mval = \log(M/U)$ donde M hace referencia al nivel de intensidad de la señal metilada para una sonda dada y U al nivel de intensidad de señal no metilada para la misma sonda. En este estudio se ha trabajado únicamente con los valores Beta.

- **DMP**. Son las siglas de *Differentially Methylated Position*. Hace referencia a una única posición genómica que se encuentra diferencialmente metilada entre las dos condiciones que se están comparando.

- ***DMR***. Son las siglas de *Differentially Methylated Region*. Se refiere a cuando varias DMP consecutivas están diferencialmente metiladas en la misma dirección (esto es, cuando estas posiciones están metiladas significativamente o no metiladas significativamente) entre los dos grupos que se comparan.

Lo interesante en este caso es obtener todas las *DMR* que se encuentran en cada uno de los estudios. Para ello usamos una de las funciones que se incluyen en el paquete *minfi*, concretamente la función *bumphunter* (Jaffe et al., 2012)

2.4.1 Configurando y usando bumphunter

La función de bumphunter que se encuentra en *minfi* es una versión del algoritmo de "*bump hunting"* creado por el grupo de Jaffe Andrew en 2012. El algoritmo trata de buscar regiones diferencialmente metiladas a lo largo del genoma entre las dos comparaciones con las que se ejecuta. Para ello bumphunter define clústeres de sondas. Estos clústeres son simplemente grupos de sondas cuyas dos posiciones genómicas entre cada una no es mayor que una cierta distancia que se define y que en la función recibe el nombre de **mapGap**.

A grandes rasgos, el algoritmo calcula primero el estadístico de t para cada localización genómica. Entonces, define como región candidata para estar metilada diferencialmente un clúster de sondas para el cual todos los estadísticos de t exceden un umbral predefinido. Para comprobar la significancia de las regiones candidatas, el algoritmo emplea un test de permutaciones, estas permutaciones son representadas por el parámetro **B.** Normalmente se llevan a cabo 1000 permutaciones para un número de regiones candidatas razonable el cual suele ser menos de unas 30000. El motivo tras esto es por el gran tiempo de computación que requiere el esquema de permutaciones. Los autores del paquete recomiendan, en caso de que nuestro número de regiones candidatas sea mayor, aumentar el punto de corte por el cual se considera que la señal no es simplemente ruido, llamado **cutoff** y que por defecto está predefinido como 0.2.

Un ejemplo de la salida que nos proporciona el algoritmo podemos verla en la **figura 13.**

	chr	start	end	value	area	cluster	indexStart	indexEnd
15861	chr8	145103393	145107199	0.3767581	6.404887	194325	238277	238293
4810	chr13	113425756	113428172	0.4257673	5.960743	57562	337302	337315
4064	chr12	54446019	54447349	0.3278039	5.900470	46543	311839	311856
17813	chr10	11206772	11208339	-0.4148710	5.393322	21466	251989	252001
18360	chr10	130844121	130844899	-0.5869309	5.282378	29724	269921	269929
4054	chr12	54409207	54409770	0.4370705	5.244846	46529	311742	311753

	L	clusterL	p.value	fwer	p.valueArea	fwerArea
	17	17	0	0	0	0
	14	24	0	0	0	0
	18	28	0	0	0	0
	13	15	0	0	0	0
	9	9	0	0	0	0
	12	42	0	0	0	0

Figura 13. Primeras seis filas del objeto que compone la salida de la función bumphunter implementada en el paquete minfi.

Como podemos observar en la figura 13, la salida de la función nos muestra por cada fila una *DMR* con un valor de confianza asociado (columnas de p.valores y fwer, también conocido como *family wise error*). Además, nos muestran el intervalo genómico en el que se encuentran aportando información tanto del cromosoma como del inicio y el fin de la *DMR*. Los otros valores son algo más accesorios (como la longitud del clúster, el identificador del mismo, etc.) salvo la columna que recibe el nombre de **value**. Esta columna indica la diferencia media de metilación en la *DMR* entre las dos comparaciones. En nuestro caso, este valor refleja la comparación entre caso versus control de manera que valores positivos reflejan un mayor nivel de metilación en esa región en las muestras de tipo tumoral y un valor negativo refleja un nivel de metilación mayor en las muestras de tipo control.

En el caso de nuestro estudio, debido a que necesitábamos todas las regiones que se encontraban en el chip, ejecutamos el algoritmo con un punto de corte o **cutoff** igual a 0. De esta manera obtenemos cerca de 260000 regiones candidatas por cada estudio. Al establecer como 0 el valor de cutoff, estamos aumentando en gran medida el número de posibles *DMR* que tenemos. Los autores del paquete aconsejan un cutoff de 0.2 debido a que después de ejecutar el esquema de permutaciones la mayoría de las *DMR* candidatas no serán significativas. Sin embargo, debido al abordaje que realizaremos en el estudio, nos son necesarias todas las regiones que se encuentran en el chip. Esto es debido a que, para llevar a cabo el análisis de enriquecimiento posterior, necesitamos toda la información posible sin importar su grado de significación para no perder información y además los datos recogidos en la columna *value*, ordenados por un ranking de mayor a menor nivel de metilación. Dado que el número de regiones candidatas que le indicaremos al programa era enorme, nos vimos obligados a reducir drásticamente el número de permutaciones que se llevaba a cabo a 50 (los autores recomiendan 1000 con un punto de corte igual a 0.2). Esto hacía que cada uno de los estudios tuviese un tiempo de computación para obtener el ranking de metilación de unos 3 días completos ejecutando el análisis en paralelo en una máquina con 4 núcleos.

Una vez que tenemos todas las regiones y su valor correspondiente de metilación diferencial, el siguiente paso fue anotar estas regiones al gen al que apuntan.

2.5 Anotación (de *DMR* a gen)

La anotación de las regiones se llevó a cabo con el paquete *bumphunter* de R, recogido en http://bioconductor.org/packages/release/bioc/html/bumphunter.html. La versión del genoma usada fue la hg19 para *Homo Sapiens* recogida en el *UCSC Genome Browser*.

Gracias a su función *matchGenes*, a través del número del cromosoma y la posición de inicio y fin de la región, obtenemos el gen al que apunta dicha *DMR*.

Debemos de tener en cuenta que un gen puede ser diana de varias *DMR*, es decir, el nivel de metilación último de un gen, en más de una ocasión estará condicionada por dos o más *DMR*. Para resumir los niveles de metilación diferencial de cada gen en una única medida que incluya la información procedente de sus *DMR* utilizamos dos abordajes que desembocaron en diferentes resultados:

1. Promedio de todos los valores de metilación que apuntaban a un mismo gen.

2. Selección del valor de metilación diferencial más alto en valor absoluto de todos los valores que apuntaban a un mismo gen.

Una vez acabada la anotación obtuvimos una matriz con los identificadores de los genes y el valor de metilación diferencial asociado a cada uno de ellos (esto es, la columna llamada *value*). Esta matriz es la que usamos como datos de entrada para el análisis de enriquecimiento.

3. Enriquecimiento funcional

Hasta ahora, siguiendo la parte anterior del análisis, tenemos varias listas de genes ordenadas según su nivel de metilación diferencial.

Por otra parte, disponemos de una gran cantidad de información biológica recogida en bases de datos que nos pueden proporcionar una visión interesante de los resultados anteriores.

Es decir, podemos caracterizar la información que tenemos a nivel de gen a una información a nivel de función biológica (anotación funcional), esta asociación la conseguimos en el presente estudio gracias a las bases de datos de Gene Ontology y KEGG.

- Gene Ontology. Contiene anotaciones de procesos biológicos, componentes celulares y funciones moleculares asociadas a grupos de genes (Ashburner et al., 2011).

- KEGG. Presenta redes de interacciones moleculares en forma de rutas (Kanehisa & Goto, 2000).

En este estudio realizamos un abordaje que requiere de la información biológica de estas bases de datos, esto es, un análisis para el enriquecimiento de grupos de genes basado en modelos de regresión logística (GSA, Gene Set Analysis) (Montaner, 2010; Montaner, Minguez, Al-shahrour, & Dopazo, 2009).

3.1 Métodos de análisis de grupos de genes: GSA

El método para el abordaje funcional que se utilizó en este estudio fue el GSA.

Los análisis de grupos de genes (GSA) son capaces de modelizar con éxito la importancia del gen más débil, reforzando la interpretación funcional de los datos genómicos.

La implementación de este tipo de análisis en el presente estudio se llevó a cabo para no perder un ápice de información, como ya se ha mencionado antes, ni la del gen más débil. Es debido a esto que usamos el GSA sobre otros tipos de análisis como podrían ser los de sobrerrepresentación.

Los análisis de sobrerrepresentación (AS), consisten en realizar la interpretación funcional a nivel de una lista de genes que están filtrados por algún criterio específico. El criterio normalmente suele ser el p valor ajustado, en nuestro caso el FWER.

Una vez que tenemos el grupo de genes que tiene una diferencia de expresión significativa entre las comparaciones realizadas, se aplican sobre estos genes un test de asociación para cada función biológica o ruta de señalización. Básicamente lo que se realiza en este punto

es una comparación entre dos listas de genes para discernir qué roles biológicos son los que se encuentran más sobrerrepresentados en nuestro grupo de genes de interés.

Los AS son una de las herramientas más utilizadas actualmente pero presentan serias desventajas que se basan en la pérdida de información causada por la cota que supone el uso de un grupo de genes previamente seleccionado (aquellos con un p valor ajustado significativo, por ejemplo menor a 0.05) y el tratamiento igualitario que se les aplica (Dopazo, 2009). Estas limitaciones son superadas por abordaje con el procedimiento de GSA para el enriquecimiento funcional.

En la figuras 14 y 15 se presentan de manera gráfica la principal diferencia entre los métodos de análisis de grupos de genes (GSA) frente a los métodos de análisis de sobrerrepresentación.

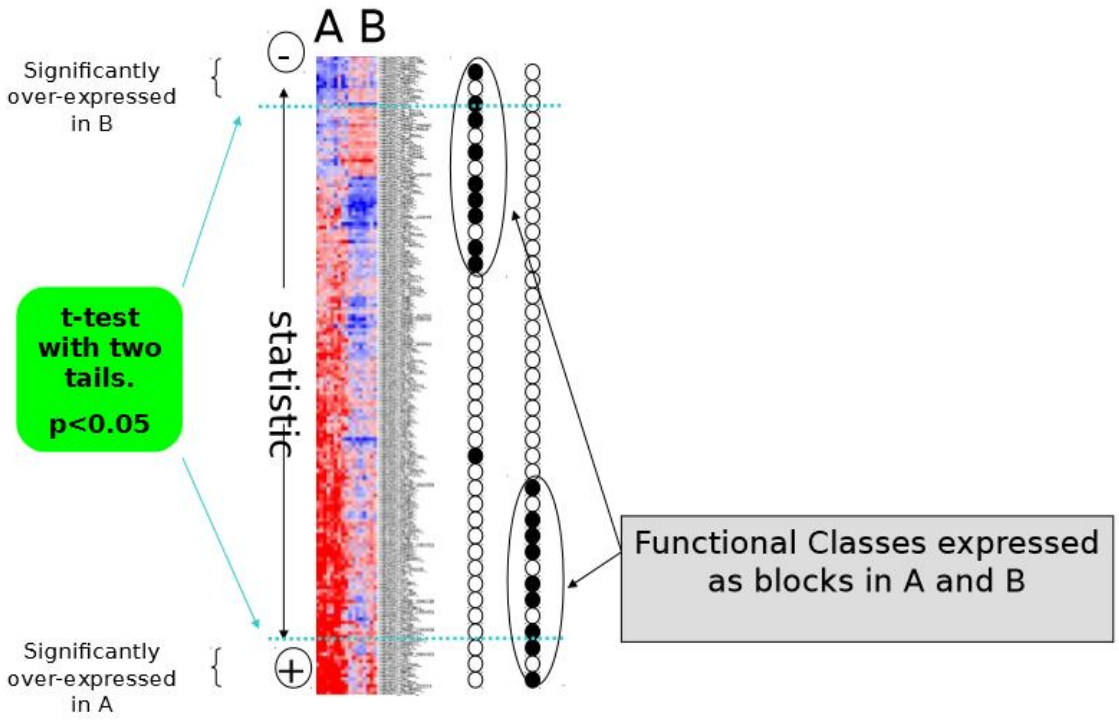

Figura 14. Representación gráfica de un análisis GSA. Como se puede observar, gracias a este tipo de aproximación podemos obtener genes que no están significativamente expresados entre las dos comparaciones A y B pero sí que participan en los clases funcionales diferencialmente expresadas.

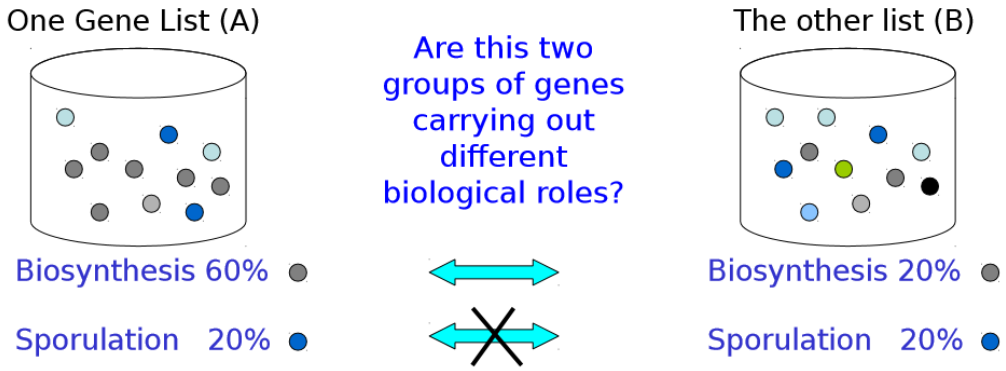

Figura 15. Ejemplo ilustrativo del funcionamiento de los análisis de sobrerrepresentación donde compara entre dos grupos de genes para comprobar si hay alguna función biológica sobrerrepresentada en alguno de ellos.

Como podemos ver en la figura 14, gracias al abordaje con GSA somos capaces de obtener más genes con un nivel de significación menor pero que forman parte de las clases funcionales que están diferencialmente expresadas.

Realizamos un GSA para cada una de las comparaciones. Adicionalmente, se separaron los términos GO en cada una de sus respectivas ontologías, es decir, procesos biológicos, funciones moleculares y componentes celulares independientemente del nivel de significación (FWER) asociado a cada gen. Esto es debido a que, como se ha comentado anteriormente, con esta aproximación queremos ver hasta el papel del gen más débil en el experimento.

De la misma manera, se realizó un GSA con los genes de cada una de las comparaciones con las rutas KEGG.

Posteriormente, cuando obtuvimos los resultados de cada uno de los diferentes GSA, usamos estos resultados como datos de entrada para el metaanálisis funcional. No sólo nos quedamos con las funciones significativas de cada análisis GSA, sino que nos quedamos con todas. Esto supone varias ventajas cuando integramos la información con el metaanálisis, las cuales van desde una mejor captura e integración de las funciones significativas y sobrerrepresentadas en cada grupo experimental hasta una mayor potencia estadística (gracias a que trabajamos con todo el conjunto de las muestras, lo que supone un tamaño muestral mucho mayor que si trabajásemos con un estudio individual).

4. Metaanálisis funcional de los datos

4.1 Introducción al metaanálisis

La aparición de tecnologías de alto rendimiento como los microarrays y la secuenciación masiva han generado una gran cantidad de datos genómicos. Para albergar tal cantidad de datos, la comunidad científica ha creado diversos repositorios que permiten la organización sistemática y el acceso libre a los datos. Entre ellos cabe destacar: **Gene Expression**

Omnibus (GEO, https://www.ncbi.nlm.nih.gov/geo/), **ArrayExpress** (https://www.ebi.ac.uk/arrayexpress/) y **Sequence Read Archive** (SRA, https://www.ncbi.nlm.nih.gov/sra). El uso de estos conjuntos de datos proporciona una importante fuente de información genómica, ofreciendo datos adicionales para análisis y evaluación de metodologías.

Si bien es cierto que el coste de las tecnologías de alto rendimiento ha ido decreciendo, la mayoría de los estudios que se realizan todavía tienen como restricción para el número de muestras este problema. Esto desemboca en que los análisis que se realizan en estos estudios con tan poca cantidad de muestras tienden a carecer de poder de detección. Este inconveniente puede ser salvado con el metaanálisis.

Las revisiones sistemáticas y los metaanálisis se han consolidado como herramientas metodológicas que ofrecen información con elevado nivel de calidad y rigor científico (García-García, 2016).

A través de una revisión sistemática de distintos estudios basados en cáncer de mama podemos integrar toda la información empírica disponible sobre esta enfermedad con un metaanálisis. El metaanálisis es una metodología estadística que permite combinar resultados sobre los efectos procedentes de diferentes estudios individuales que han sido identificados y seleccionados críticamente a partir de una revisión sistemática.

Este tipo de procedimiento se empezó a usar en ciencias sociales y psicología arrojando buenos resultados. Posteriormente se empezó a aplicar cada vez más en publicaciones médicas (Haidich, 2010) hasta llegar a diversas áreas científicas como la Genómica.

El metaanálisis nos permite obtener una medida combinada del efecto de interés con una mayor precisión que la ofrecida por los estudios individuales que han sido seleccionados previamente en la revisión sistemática, y por lo tanto ofrece una mayor potencia estadística cuantificando la variabilidad de los estudios individuales. Sin embargo, una inadecuada selección de los métodos estadísticos e interpretación de resultados puede suponer una limitación en su uso e interpretación.

4.2 Metaanálisis funcional

La mayoría de los métodos de metaanálisis de datos genómicos se centran en un nivel de gen o variante. Sin embargo, el uso de procedimientos de metaanálisis a nivel de función amplía el número de escenarios donde interpretar los resultados de estudios genómicos procedentes de diferentes tecnologías, proporcionando una mejor comprensión biológica y clínica.

El metaanálisis funcional empleado en este estudio sigue la misma metodología que el propuesto por (García-García, 2016), el cual consta de varias partes que se resumen a continuación.

4.2.1 Configuración y exploración de matrices de entrada

A partir de los resultados del enriquecimiento funcional con los modelos logísticos (GSA) obtuvimos una medida del efecto entre enfermos y controles que se representa mediante

logaritmos de los odds ratios (razón de ventajas) entre ambas condiciones. Para cada una de las funciones que se evaluaron en las bases de datos de interés, disponemos de la medida del efecto y su varianza. Ambos elementos son necesarios como entrada para el metaanálisis:

- Matriz de logaritmos de los *odds* ratios
- Matriz de varianzas de los logaritmos de los *odds* ratios.

La exploración de estas matrices permite un mejor conocimiento de los datos así como la detección de valores atípicos y una cuantificación de los valores perdidos procedentes de la combinación de las distintas comparaciones.

4.2.2 Análisis de heterogeneidad y combinación de efectos

En este paso, para cada una de las funciones de las bases de datos seleccionadas, se realiza un metaanálisis que combina el efecto medido de todas las comparaciones analizadas. La combinación y presentación de los resultados se lleva a cabo utilizando diversas técnicas estadísticas cuya elección depende del tipo de medida de resultado/efecto que se ha utilizado y de la valoración del grado de heterogeneidad de las comparaciones.

En un metaanálisis, primero se evalúa la heterogeneidad de los estudios y en función de la misma se decide el modelo de estimación de la variabilidad del efecto. Sin embargo, al trabajar con miles de metaanálisis simultáneamente (tantos como funciones), consideramos la información a priori de la que disponemos para cada una de las comparaciones para la selección de la técnica de combinación del efecto y tras su ejecución computacional, valoramos los resultados del estudio de heterogeneidad para confirmar la idoneidad del procedimiento seleccionado.

En la obtención de la medida resumen, se ponderan los resultados obtenidos de las comparaciones entre un tipo de cáncer frente a individuos sanos por la inversa de su varianza. De este modo, las comparaciones que presenten mucha variabilidad tendrán menos peso sobre el efecto combinado. La heterogeneidad entre las comparaciones puede ser medida por un modelo de efectos aleatorios o bien no ser incluida, utilizando en este caso un modelo de efectos fijos.

La diferencia entre los modelos de efectos fijos y los modelos aleatorios es que en los modelos de efectos fijos, la heterogeneidad entre las comparaciones no está contemplada, de manera que todas las comparaciones estiman el mismo efecto y las diferencias observadas se deben únicamente al azar. Sin embargo, el modelo de efectos aleatorios incorpora la posible heterogeneidad de los efectos entre los distintos estudios. En este caso, la ponderación en la medida de la combinación del efecto incluye tanto la variabilidad entre-comparaciones como la variabilidad intra-comparación. En un contexto genómico, la heterogeneidad presente entre las distintas comparaciones suele ser debida al empleo de distintas plataformas de tecnologías de alto rendimiento, a un distinto tamaño muestral y a diferentes tipos de contrastes entre grupos experimentales. Es por esto, que un modelo de efectos aleatorios se ajusta mejor a las características de las comparaciones recogidas en este estudio.

Para el metaánalisis de cada término funcional se usaron las funciones incluidas en el paquete *metafor* (Viechtbauer, 2010) de R que incorpora la implementación de varios modelos de efectos aleatorios así como la del modelo de efectos fijos. Los modelos de efectos aleatorios que se usaron fueron:

- DL (DerSimonian & Laird, 2014).

- HE (Hedges, Gurevitch, & Curtis, 1999).

- HS (Hunter & Schmidt, 2014).

Nos centramos sobre todo en el modelo de efectos aleatorios DL para la combinación de las medidas del efecto por ser más adecuado en el contexto descrito.

4.2.3 Análisis de sensibilidad y evaluación de sesgos

Tras la realización del metaanálisis, la influencia de cada una de las comparaciones es evaluada mediante un análisis de sensibilidad. Para ello, en la metodología se reproduce el metaanálisis excluyendo la comparación que se quiera revisar. Si los resultados son similares a los obtenidos con el total de comparaciones, se estará garantizando la robustez del metaanálisis.

La revisión de posibles sesgos es importante para garantizar una adecuada interpretación de los resultados. Para ello se utilizaron gráficos de embudo, los cuales nos informan de la presencia de heterogeneidad entre las comparaciones y además de la presencia de un posible sesgo de publicación.

Además, la presencia de algunas comparaciones cuya variabilidad y magnitud son muy diferentes del resto pueden producir una fuerte influencia en el metaanálisis. La exclusión de esta comparación en el análisis permite considerar los cambios en el modelo ajustado y valorar su influencia. Para comprobar esto utilizamos varias medidas de diagnóstico de datos influyentes en representadas en forma de gráficos (residuos estandarizados, las distancias de Cook, covarianza de ratios, etc.) para cada función.

4.2.4 Representación e interpretación de resultados

La metodología incluida en el metaanálisis funcional proporciona resultados globales sobre el conjunto de las anotaciones y resultados específicos a nivel de función, ofreciendo una evaluación de la integración de los estudios en el modelo.

Entre los resultados globales del metaanálisis se incluyen diversas tablas y gráficas, siendo el primer resultado global de la metodología utilizada una tabla resumen que informa sobre el desarrollo y obtención de resultados significativos tras la combinación de la información de las comparaciones iniciales. Gracias a la representación gráfica se favorece una rápida interpretación de los resultados del metaanálisis.

Resultados

1. Revisión sistemática y selección de estudios de metilación

En la revisión sistemática de los estudios de metilación de cáncer de mama realizados con el chip 450k de Illumina y con el diseño experimental caso-control, se seleccionaron 49 estudios de interés. En una segunda fase, se comprobaron los criterios de inclusión y exclusión, además de las características específicas de los estudios, refinándose la búsqueda inicial. A continuación se detalla el proceso de evaluación de algunos de estos estudios:

- **GSE78751.** En este estudio se evalúa un cáncer de mama triple negativo. Comparan muestras de tejido mamario canceroso provenientes de 23 pacientes contra 11 muestras de tejido adyacente sanos de los mismos. Y contra 12 muestras metastásicas en los nodos linfáticos. Todas las muestras se validan contra una cohorte de 70 muestras de tejido sano. Muestras totales usadas en el estudio: 116. Este estudio se seleccionó puesto que si eliminamos las muestras de metástasis en los nodos linfáticos podemos obtener una comparación de tejido sano frente a tejido canceroso mamario.
- **GSE71626.** En este estudio los autores comparan muestras tumorales de cáncer de mama, próstata e hígado invasivos contra otras variantes menos invasivas de los mismos. Muestras totales usadas en el estudio: 17. Este estudio no fue seleccionado debido a que carece de los grupos experimentales que buscamos los cuales son controles (sanos) frente a casos (cáncer de mama).
- **GSE49143.** En este estudio se caracteriza el metiloma de la línea de células cancerosas NCI60 en un panel el cual incluye células de cáncer de mama. Muestras totales usadas en el estudio: 60. Se descartó por no incluir la comparación entre los grupos experimentales que buscábamos.
- **GSE72277.** En este estudio se comparan 3 tipos de muestras de cáncer de mama. 47 correspondientes a una variante mutada de BRCA1 conocida, 65 muestras correspondientes a la variante de BRCA1 cuya mutación es la común en el cáncer de mama y 38 muestras tumorales donde BRCA1 no era el causante del cáncer de mama al no estar mutado. De entre todas estas muestras, escogen un 60 para llevar a cabo el estudio. Muestras totales usadas en el estudio: 60. Este estudio no se utilizó para la comparación puesto que no cumplía con el criterio que buscábamos para los grupos experimentales. Es cierto que trataban con muestras donde BRCA1 no estaba mutado, pero estas muestras seguían siendo tumorales y por lo tanto este estudio no cumplía la condición control frente a tumor que buscábamos.
- **GSE69118.** Este estudio trata de metilación en el endocrino por lo que se descartó inmediatamente.
- **GSE38548.** Aquí se compararon células somáticas híbridas correspondientes a dos tipos de cáncer de mama: basal y luminal, resistentes a puromicina y a G418. Se generaron controles de fusión entre la misma línea celular (homofusiones) y heterofusiones de diferentes células tumorales basales fueron generadas también como controles. Muestras totales usadas en el estudio: 27 híbridos. Este estudio se

descartó porque no cumplía con la comparación entre grupos experimentales que buscábamos.

- **GSE37754.** Los autores de este estudio buscaban asociación entre metabolitos que se encuentran en células tumorales comparando 27 muestras de tejido tumoral frente a 19 muestras de tejido adyacente no tumoral. Muestras totales usadas en el estudio: 46. No se seleccionó porque tenía más de un tipo de control con diferente comportamiento frente a metabolitos.
- **GSE61744.** Comparan los perfiles de metilación en una misma línea celular de cáncer de mama bajo inhibición a PJ-34 y otras donde no lo inhiben. Se cree que este gen tiene gran importancia en el proceso de metilación del ADN. Muestras totales usadas en el estudio: 2. No se seleccionó por no cumplir las comparaciones que requeríamos en los grupos experimentales y por el número tan bajo de muestras.
- **GSE45958.** En el estudio se comparan distintas células de cáncer de mama. Se exponen a 0, 2 y 6G de radiación y se observa cómo cambia el patrón de metilación tras 1, 2, 4, 8, 24, 48 y 72h desde que se sometieron a radiación. Se usan como controles 10 muestras tumorales a las que no se las irradia frente a 14 muestras tumorales irradiadas. Muestras totales usadas: 24. No se seleccionó el experimento porque no ofrecía los grupos experimentales que requeríamos para poder hacer las comparaciones.
- **GSE51032.** Este estudio se trata de una cohorte que fue producida en la Fundación de Genética Humana en Turín, Italia. En el estudio se comparan 845 muestras, 188 hombres y 657 mujeres a los cuales después de estudiar su metiloma se les hizo un seguimiento a lo largo varios años. Este estudio no se seleccionó porque los datos no proporcionaban información sobre que muestras eran tumorales y cuales sanas. Esto sumado al gran número de muestras que incluía además de no estar publicado hicieron que no lo seleccionásemos.
- **GSE66695.** Este estudio compara 40 muestras sanas frente a 80 muestras tumorales de cáncer de mama. En un principio, este estudio era uno de los que fueron elegidos pero al descargarnos los datos y trabajar con la matriz de intensidades de señales descubrimos que los nombres que les daban a las muestran en la matriz no concordaban con los que recibían en GEO. Intentamos ponernos en contacto con la responsable del estudio pero sin éxito por lo que finalmente el estudio fue descartado.
- **GSE52865.** En este estudio se comparan 40 muestras de tejido tumoral de cáncer de mama frente a 17 tejidos de mama sanos. Entre las 40 muestras tumorales encontramos diferentes subtipos de cáncer de mama, como el basal, dos tipos de luminal y HER2. Muestras totales del experimento: 57. Este estudio se seleccionó para llevar a cabo el metaanálisis debido a que si separábamos los subtipos de cáncer de mama obtendríamos 4 comparaciones adicionales. Es decir, a partir de este estudio se realizaron las comparaciones de control frente a cáncer de mama basal, control frente a los dos subtipos de cáncer de mama luminal y control frente a cáncer de mama de tipo HER2.

- **GSE59901.** Este experimento incluía muestras procedentes de varios subtipos y líneas celulares de cáncer de mama. Puesto que las muestras procedentes de líneas celulares eran una para cada línea celular las descartamos quedando entonces únicamente los subtipos de cáncer de mama. Estos estaban formados por 8 muestras tumorales de cáncer de mama provocado por la mutación de BRCA1, 10 muestras tumorales de cáncer de mama de tipo IDC (carcinoma ductal invasivo) y otras 10 muestras tumorales de cáncer de mama de tipo ILB (carcinoma lobulal invasivo). Estas muestras fueron comparadas con 4 muestras controles de tejido mamario sano. Muestras totales: 36 (32 tras deshacernos de las líneas celulares). Este fue el último estudio seleccionado por cumplir con nuestra preferencia de grupos experimentales. A partir de él se sacaron 3 comparaciones más.

Tras esta revisión, se seleccionaron 3 estudios (GSE78751, GSE52865 y GSE59901), obteniendo 8 comparaciones con las que realizar nuestros estudios de metilación y el posterior metaanálisis funcional. Podríamos haber añadido más estudios para obtener más comparaciones, pero dado que ya teníamos un buen número de ellas y que el tiempo de procesado de datos de metilación es bastante alto, decidimos quedarnos con las 8 que teníamos.

2. Procesamiento y análisis primario de los datos

2.1 Descarga de los datos sin procesar procedentes de GEO

Tras la descarga de los datos, tuvimos que procesar los mismos para ajustarlos a nuestro paquete de análisis de metilación además de separar cada estudio en varias comparaciones anteriormente mencionadas. En la tabla 1 podemos ver cómo se separaron los estudios de manera esquemática.

ESTUDIO	COMPARACIÓN 1	COMPARACIÓN 2	COMPARACIÓN 3	COMPARACIÓN 4
gse52865 contiene controles sanos frente a cáncer de mama de tipo basal, her2, luma y lumb	Controles frente a cáncer de mama de tipo basal.	Controles frente a cáncer de mama de tipo her-2.	Controles frente a cáncer de mama de tipo lum-a.	Controles frente a cáncer de mama de tipo lumb.
	COMPARACIÓN 5	**COMPARACIÓN 6**	**COMPARACIÓN 7**	
gse59901 contiene controles sanos frente a cáncer de mama de tipo brc, idc e ilb	Controles frente a cáncer de mama de tipo brc.	Controles frente a cáncer de mama de tipo idc.	Controles frente a cáncer de mama de tipo ilb.	
	COMPARACIÓN 8			
gse78751 contiene controles sanos frente a cáncer de mama de tipo idc y metástasis en nodos linfáticos	Controles frente a cáncer de mama de tipo idc.	Se descarta la comparación de nodos linfáticos al no ser cáncer un tipo de cáncer de mama.		

Tabla 1. Comparaciones de grupos experimentales de interés incluidas en los estudios seleccionados.

En la tabla 2 se muestran las once primeras filas y cinco primeras columnas de la matriz de intensidades de señal sin procesar para la primera comparación (control versus cáncer de mama de tipo basal).

SONDA	CONTROL 1 M	CONTROL 1 U	CONTROL 2 M	CONTROL 2 U
cg00000029	526.79	3073.79	580.91	2408.70
cg00000108	5131.96	434.13	5035.42	493.55
cg00000109	1615.51	478.95	1900.49	683.20
cg00000165	560.64	1960.65	849.67	2051.94
cg00000236	2537.56	363.57	2898.73	588.37
cg00000289	660.25	434.96	1227.76	595.37
cg00000292	5604.48	3656.51	4835.94	3645.30
cg00000321	1561.87	4767.99	1458.12	3894.02
cg00000363	1628.28	6720.34	1195.21	5855.03
cg00000622	27.39	13229.85	112.70	9675.99

Tabla 2. Muestra de la manipulación de los datos que sirven de entrada para el paquete minfi. En la tabla se pueden observar las 11 primeras filas así como las 5 primeras columnas. M y U hacen referencia a los distintos canales de metilación y no metilación respectivamente. Datos sin normalizar.

A continuación, se procedió a un análisis exploratorio y control de calidad de los datos sin procesar.

2.2 Control de calidad de los datos

El control de calidad de los datos sin procesar se llevó a cabo mediante el uso de herramientas incluidas en el paquete *minfi,* así como por métodos alternativos como análisis clúster y PCA. En la figura 16 podemos ver el control de calidad para los datos sin procesar de la primera comparación que se realizó (control versus cáncer de mama tipo basal).

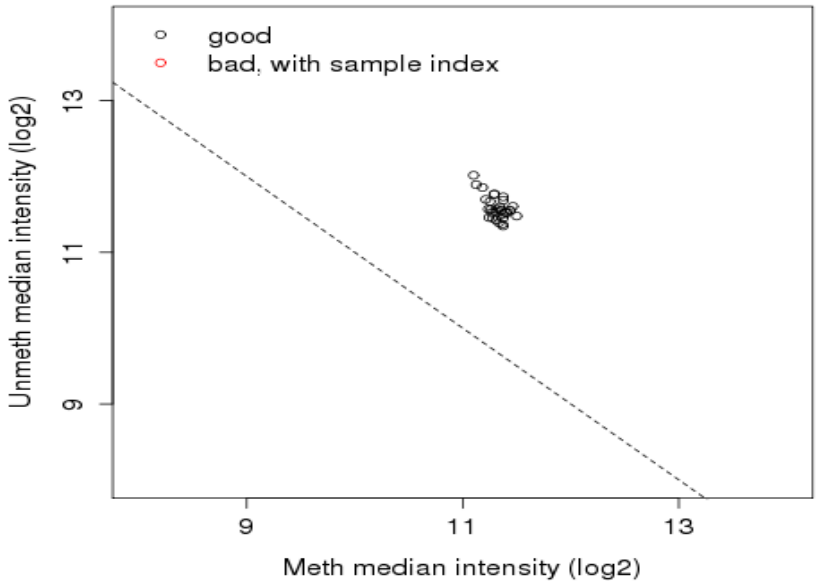

Figura 16. Control de calidad de los datos sin procesar para la comparación control vs cáncer de mama tipo basal. Las 7 comparaciones restantes ofrecieron un resultado análogo a la aplicación de este control de calidad, todas ellas lo pasaron.

Los creadores del paquete aclaran que para datos de metilación, se considera que estos tienen una buena calidad cuando el logaritmo de la media tanto de la señal de metilación como de no metilación de cada muestra es similar.

Otra medida de control de calidad que se utilizó para los estudios fue el análisis de componentes principales, como el que se muestra en la figura 17 para la primera comparación control vs cáncer de mama de tipo basal.

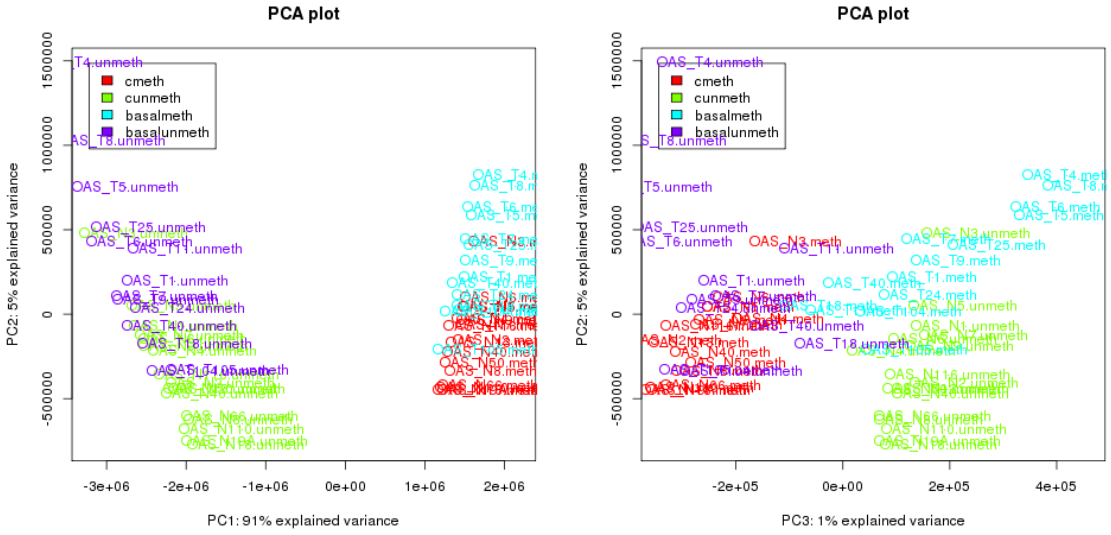

Figura 17. Análisis de componentes principales para la comparación control vs cáncer de mama tipo basal. En rojo y verde las intensidades de señal metilada y no metilada de los controles respectivamente. En azul y violeta las intensidades de señal metilada y no metilada de las muestras tumorales. Las 7 comparaciones restantes ofrecieron un resultado análogo a la aplicación de este control de calidad.

En el análisis de componentes principales, si miramos la ilustración donde se explica el 91% de la varianza, se observa cómo las muestras tienden a distribuirse entre ellas según si están o no metiladas antes que por pertenecer a uno de los distintos grupos experimentales (control y caso).

Por último, para comprobar este nivel de agrupación, se procedió a la realización de un análisis clúster como se puede visualizar en la figura 18 para la comparación anterior.

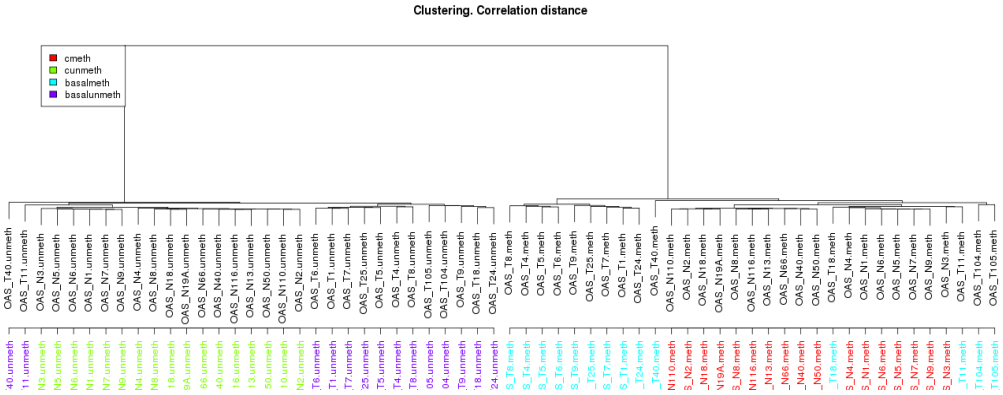

Figura 18. Análisis clúster para la comparación control vs cáncer de mama tipo basal. En rojo y verde las intensidades de señal metilada y no metilada de los controles respectivamente. En azul y violeta las intensidades de señal metilada y no metilada de las muestras tumorales. Las 7 comparaciones restantes ofrecieron un resultado análogo a la aplicación de este control de calidad.

El análisis clúster confirmó que las muestras se agrupan en base a si eran las provenientes de la señal metilada o la no metilada antes que a ser de distintos grupos experimentales.

Tras comprobar la calidad de las muestras sin procesar se procedió a su respectiva normalización.

2.3 Normalización de los datos

La normalización de los datos se llevó a cabo por el método de los cuantiles, método análogo al que recomiendan los creadores del paquete *minfi* para datos tipo control vs caso.

Tras la normalización, los datos para la primera comparación (el resto son análogas) quedaron como se muestra en la tabla 3.

SONDA	CONTROL 1 M	CONTROL 1 U	CONTROL 2 M	CONTROL 2 U
cg00000029	683.49	2858.57	772.35	2530.29
cg00000108	6076.86	430.43	6785.92	413.57
cg00000109	1851.77	474.66	1971.89	605.40
cg00000165	719.96	1924.65	1045.25	2144.74
cg00000236	2930.78	361.23	3301.73	507.72
cg00000289	824.93	431.31	1352.53	514.88
cg00000292	6592.05	3321.79	6462.37	3866.44
cg00000321	1791.99	4191.39	1539.92	4138.21
cg00000363	1866.96	5800.07	1327.09	6512.99
cg00000622	93.54	11760.81	166.095	11505.57

Tabla 3. Primeras filas y columnas tras la normalización por cuantiles para la primera comparación control vs cáncer de mama tipo basal. M y U hacen referencia a los distintos canales de metilación y no metilación respectivamente.

Podemos obtener una visión completa cómo cambia la distribución de los datos tras la normalización con las figuras 19 y 20 para la comparación control vs cáncer de mama de tipo basal.

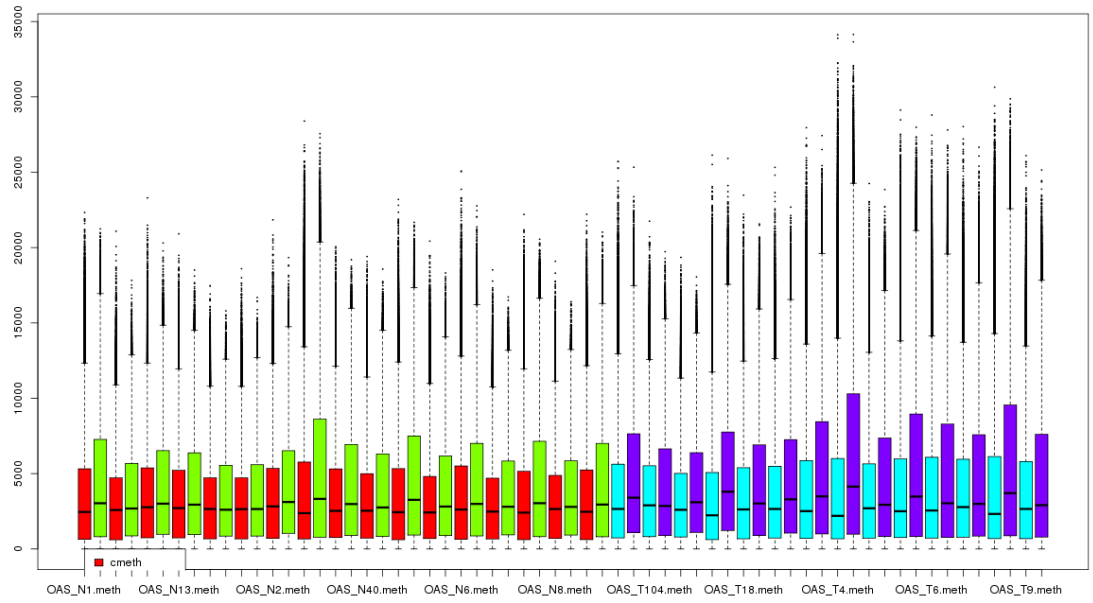

Figura 19. Distribución de las intensidades de señal de los datos sin normalizar para la comparación control vs cáncer de mama de tipo basal. En rojo y verde las señales metiladas y no metiladas de los controles respectivamente. En azul y violeta las señales metiladas y no metiladas del cáncer de mama de tipo basal.

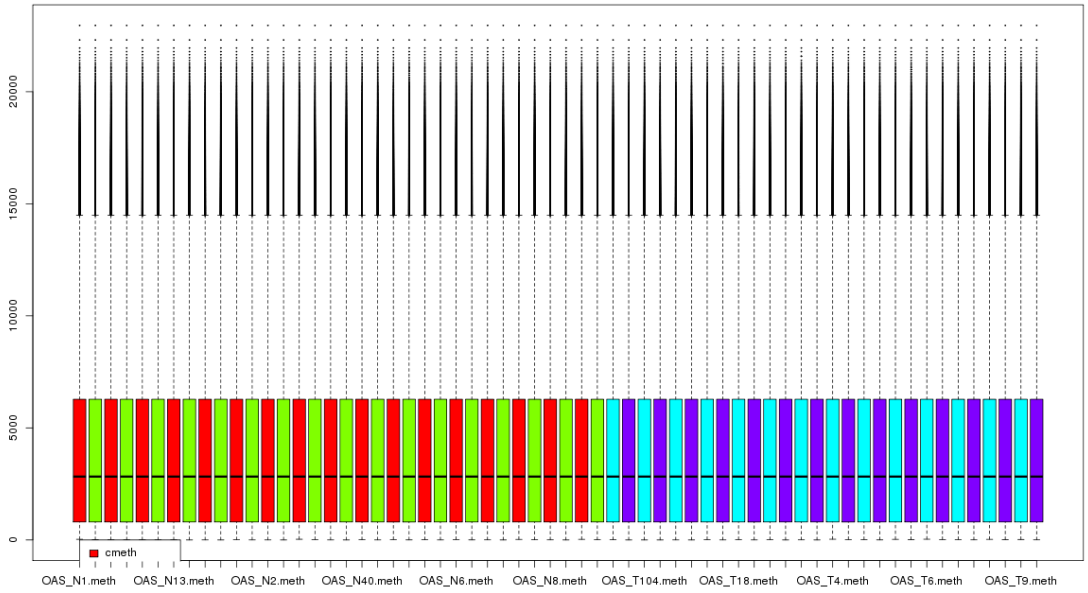

Figura 20. Distribución de las intensidades de señal de los datos normalizados para la comparación control vs cáncer de mama de tipo basal. En rojo y verde las señales metiladas y no metiladas de los controles respectivamente. En azul y violeta las señales metiladas y no metiladas del cáncer de mama de tipo basal. Gracias a la normalización conseguimos una distribución homogénea de los niveles de intensidad de señal para cada una de las muestras.

Una vez tuvimos los datos normalizados, el siguiente paso fue comprobar la calidad de estos mediante la aplicación de nuevo de los análisis clúster y PCA.

Procedemos primero con el análisis de componentes principales para la primera comparación, se muestra en la figura 21.

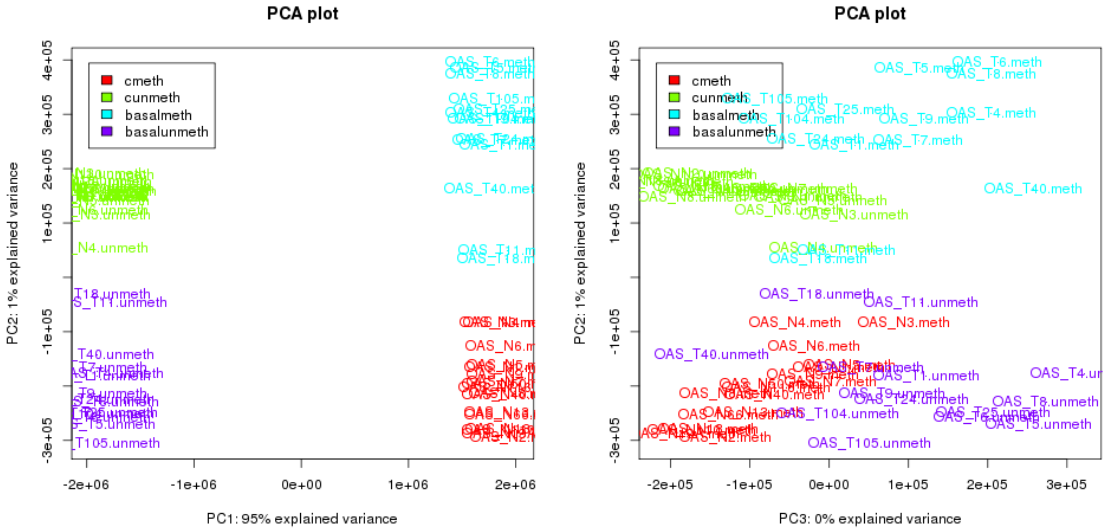

Figura 21. Análisis de componentes principales para los datos normalizados en la primera comparación control vs cáncer de mama tipo basal. En rojo y verde se muestran las intensidades de señal metilada y no metilada de los controles mientras que en azul y violeta lo hacen para los tumores, respectivamente.

Gracias al análisis de componentes principales podemos observar que las muestras se separan en 4 grupos. Dos grandes grupos que se separan entre sí dependiendo de si la

señales son metiladas o no metiladas (95% de la varianza explicada) y otros dos subgrupos que nos separan los grupos experimentales (1% de la varianza explicada).

Después del análisis de componentes principales podemos observar el análisis clúster de correlación en la figura 22 donde observamos que todo sigue igual a cuando los datos no estaban normalizados.

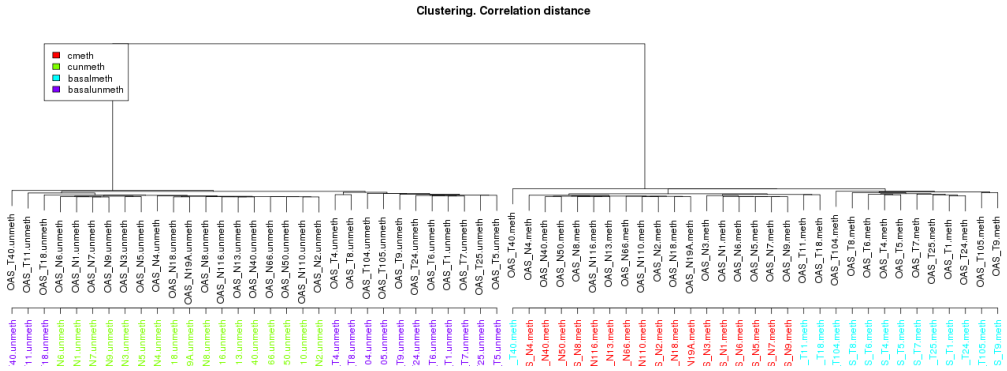

Figura 22. Análisis clúster de correlación para la comparación control vs cáncer de mama tipo basal. No hubo cambios respecto a su versión sin normalizar.

En las demás comparaciones se siguió un patrón similar menos en la perteneciente al estudio GSE78751 donde se comparaba cáncer de mama de tipo idc frente a controles sanos. Podemos ver su análisis de componentes principales en la figura 23.

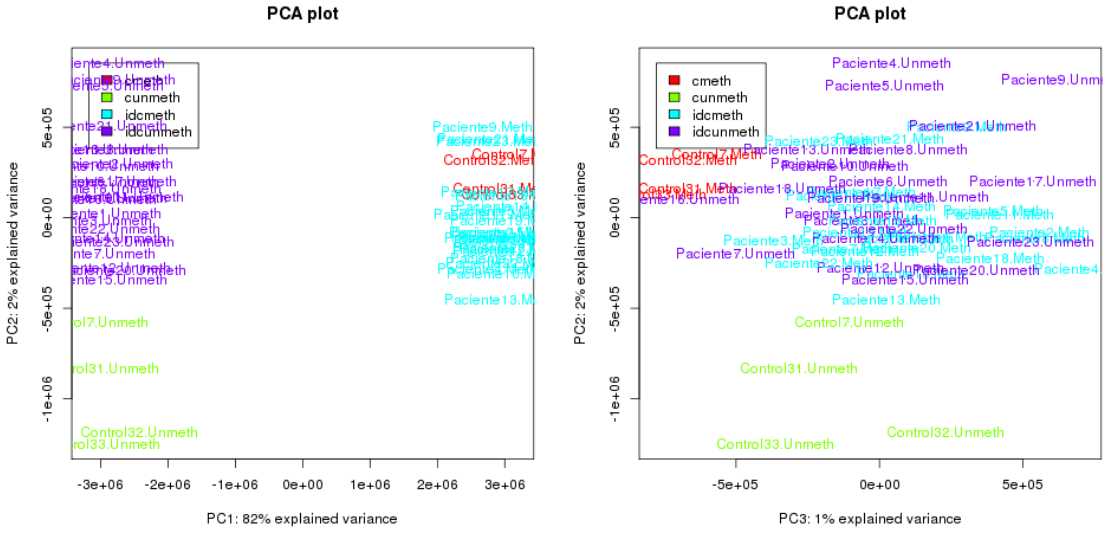

Figura 23. Análisis de componentes principales para la comparación de controles sanos frente a cáncer de mama de tipo invasivo ductal (idc).

Como podemos observar en su análisis de componentes principales, esta comparación no terminaba de tener las señales procedentes de los controles metilados totalmente separadas de las procedentes de los casos metilados. Fue esta la razón que nos hizo pensar que añadiría más variabilidad al metaanálisis y por consiguiente se descartó esta comparación quedándonos únicamente con las 7 restantes, las cuales sí que mantenían una apreciable diferenciación de cada uno de los grupos. Tras la normalización y

comprobación última de la calidad de las muestras abordamos el análisis de metilación diferencial.

2.4 Análisis de metilación diferencial

En este análisis llevado a cabo con el paquete de análisis de datos de metilación *minfi* usamos como entrada las matrices de intensidad de señal normalizadas y que pasaron nuestro filtro de control de calidad. Tenemos 7 de ellas pertenecientes a los estudios GSE52865 y GSE59901.

Tras el análisis de metilación diferencial *minfi* nos devolvió un objeto en forma de matriz donde tenemos las regiones diferencialmente metiladas asociadas a su posición genómica y nivel de significación, entre otros datos. La salida que obtenemos tras el análisis de metilación diferencial para la comparación entre controles sanos y cáncer de mama de tipo basal se muestra en la tabla 4 de manera simplificada.

chr	***start***	***End***	***value***	***cluster***	***L***	***clusterL***	***p.value***	***fwer***
chr6	32119616	32121758	0.1754458	163827	48	87	0	0
chr12	115131142	115135700	-0.1535845	50318	53	74	0	0
chr6	32063459	32066582	-0.1345020	163807	60	73	0	0
chr6	33287199	33289280	0.1712754	164140	42	181	0	0
chr6	29520752	29521803	0.1876998	163140	38	40	0	0

Tabla 4. Matriz resultante tras el análisis de metilación diferencial llevado a cabo por minfi para la comparación entre controles sanos y cáncer de mama de tipo basal. La matriz se encuentra simplificada ya que, además de que la información que se ha eliminado es accesoria, de esta manera ofrece una mejor visión de las primeras filas de la tabla.

La manera de interpretar la salida de *minfi* para la tabla 4 en la primera fila es la siguiente: hemos obtenido una región diferencialmente metilada en el cromosoma 6, empieza en la posición 32119616 y continúa hasta la posición 32121758. La diferencia media de metilación entre los grupos experimentales es 0.1754 siendo las muestras pertenecientes a los casos las que se encuentran por tanto más metiladas (proviene de la comparación señal tumor – señal control). Esta *DMR* (región diferencialmente metilada) pertenece al clúster número 163827 (un clúster en *minfi* son varias sondas juntas que apuntan a una misma dirección de metilación) que tiene una longitud de 48 sitios metilados. El p valor asociado a esta *DMR* es 0, así como su FWER.

Tras la obtención de las regiones diferencialmente metiladas el siguiente paso consistió en anotarlas como genes.

2.5 Anotación (de *DMR* a gen)

Tras anotar las regiones diferencialmente metiladas como genes, construimos una matriz con los genes y el valor que hace referencia a la diferencia media de metilación entre los grupos experimentales.

Varias *DMR* pueden apuntar a un mismo gen pero necesitamos la información de metilación diferencial resumida en un único valor. Para esto utilizamos dos aproximaciones que fueron:

- Determinación del promedio de las medidas de metilación en todas las *DMR* asociadas a un mismo gen.
- Selección de la medida de metilación con valor absoluto mayor, entre las *DMR* de un mismo gen.

El resultado de aplicar este criterio para la comparación entre controles sanos frente a las muestras de cáncer de mama de tipo basal, queda reflejado en la tabla 5 para la aproximación realizando el promedio y en la tabla 6 para la aproximación tomando la *DMR* que tuviese mayor valor absoluto. En ambas se muestran las 6 primeras filas con la diferencia media de metilación ordenada según un ranking de mayor a menor (mayor más positivo, menor más negativo) que nos será necesario para llevar a cabo el análisis de grupos de genes (GSA).

Genes	Value
HOTAIRM1	0.3032731
GPR33	0.2834010
ZNF436-AS1	0.2508752
PFN3	0.2495328
LINC00189	0.2153137

Tabla 5. Resultado de la anotación de genes para la comparación entre controles y cáncer de mama de tipo basal con el método de aproximación del promedio.

Genes	Value
SEPW1	0.4627702
MAML3	0.4369506
SEL1L	0.4284124
GUSBP2	0.4102202
SEC14L1	0.4060176

Tabla 6. Resultado de la anotación de genes para la comparación entre controles y cáncer de mama de tipo basal con el método de aproximación del valor absoluto.

Una vez tenemos todas las *DMR* resumidas a nivel de gen realizamos el análisis de enriquecimiento funcional.

3. Enriquecimiento funcional

3.1 Métodos de análisis de grupos de genes: GSA

El análisis de enriquecimiento funcional se llevó a cabo con GSA. Para ello utilizamos anotación procedente tanto de Gene Ontology como de KEGG Pathways. La caracterización funcional se realizó para cada una de las 7 comparaciones bajo las distintas anotaciones. Además, los términos GO se desglosaron según su tipo de ontología por lo que finalmente ejecutamos 21 GSA para términos GO y 7 para rutas de KEGG.

Cabe recordar que este procedimiento se hizo por duplicado, ya que comparábamos el resultado de la aproximación de resumir las *DMR* a nivel de gen por el promedio de la intensidad de las señales frente a resumirlas por el valor absoluto más alto.

3.1.1 Aproximación valor absoluto

Los resultados de los análisis GSA para los términos GO con las *DMR* resumidas a nivel de gen por el valor absoluto se muestran en las tablas 7, 8 y 9 correspondiendo cada una a los procesos biológicos, funciones moleculares y componentes celulares respectivamente.

COMPARACIÓN	OVER	UNDER	SIG.OVER	SIG.UNDER
C VS BASAL	4415	3601	25	28
C VS HER2	3807	4209	14	58
C VS LUMA	4919	3097	31	14
C VS LUMB	4330	3686	34	16
C VS BRC	4628	3388	44	25
C VS IDC	4740	3276	45	11
C VS ILB	4811	3205	24	9

Tabla 7. GSA con términos GO de procesos biológicos. En la columna "Comparación", C hace referencia a controles sanos contra los diferentes subtipos de cáncer de mama. Over y Under hacen referencia al número de funciones que están más representadas en tumores y controles, respectivamente. Sig.over y Sig.under hace referencia al número de funciones que están más representadas en tumores y controles significativamente de forma respectiva (esto lo averiguamos gracias al signo del logaritmo del odds ratio/LOR. Si el signo es positivo lo consideraremos over y si es negativo, under).

COMPARACIÓN	OVER	UNDER	SIG.OVER	SIG.UNDER
C VS BASAL	1001	642	2	2
C VS HER2	884	757	0	1
C VS LUMA	1018	623	3	8
C VS LUMB	914	727	3	2
C VS BRC	1014	627	4	15
C VS IDC	941	700	7	5
C VS ILB	910	731	7	6

Tabla 8. GSA con términos GO de funciones moleculares. En la columna "Comparación", C hace referencia a controles sanos contra los diferentes subtipos de cáncer de mama. Over y Under hacen referencia al número de funciones que están más representadas en tumores y controles, respectivamente. Sig.over y Sig.under hace referencia al número de funciones que están más representadas en tumores y controles significativamente de forma respectiva.

COMPARACIÓN	OVER	UNDER	SIG.OVER	SIG.UNDER
C VS BASAL	608	347	1	0
C VS HER2	575	380	1	0
C VS LUMA	653	302	19	5
C VS LUMB	607	348	0	0
C VS BRC	643	312	11	1
C VS IDC	620	335	7	0
C VS ILB	590	365	0	0

Tabla 9. GSA con términos GO de componentes celulares. En la columna "Comparación", C hace referencia a controles sanos contra los diferentes subtipos de cáncer de mama. Over y Under hacen referencia al número de funciones que están más representadas en tumores y controles, respectivamente. Sig.over y Sig.under hace referencia al número de funciones que están más representadas en tumores y controles significativamente de forma respectiva.

Los GSA resultantes de cada una de las diferentes comparaciones con la anotación de KEGG con la aproximación del valor absoluto están resumidos en la tabla 10.

COMPARACIÓN	OVER	UNDER	SIG.OVER	SIG.UNDER
C VS BASAL	215	87	0	1
C VS HER2	128	174	0	1
C VS LUMA	178	124	1	3
C VS LUMB	217	85	0	1
C VS BRC	197	105	1	1
C VS IDC	172	130	0	1
C VS ILB	178	124	0	1

Tabla 10. GSA con rutas de KEGG. En la columna "Comparación", C hace referencia a controles sanos contra los diferentes subtipos de cáncer de mama. Over y Under hacen referencia al número de funciones que están más representadas en tumores y controles, respectivamente. Sig.over y Sig.under hace referencia al número de funciones que están más representadas en tumores y controles significativamente de forma respectiva.

3.1.2 Aproximación promedio

Los resultados de los análisis GSA para los términos GO con las *DMR* resumidas a nivel de gen mediante el promedio se muestran en las tablas 11, 12 y 13 correspondiendo cada una a los procesos biológicos, funciones moleculares y componentes celulares respectivamente.

COMPARACIÓN	OVER	UNDER	SIG.OVER	SIG.UNDER
C VS BASAL	4256	3760	55	183
C VS HER2	4428	3588	21	101
C VS LUMA	5023	2993	103	45
C VS LUMB	4972	3044	47	23
C VS BRC	4645	3371	197	125
C VS IDC	5239	2777	167	39
C VS ILB	5506	2510	156	16

Tabla 11. GSA con términos GO de procesos biológicos. En la columna "Comparación", C hace referencia a controles sanos contra los diferentes subtipos de cáncer de mama. Over y Under hacen referencia al número de funciones que están más representadas en tumores y controles, respectivamente. Sig.over y Sig.under hace referencia al número de funciones que están más representadas en tumores y controles significativamente de forma respectiva.

COMPARACIÓN	OVER	UNDER	SIG.OVER	SIG.UNDER
C VS BASAL	970	673	8	21
C VS HER2	959	682	3	9
C VS LUMA	1062	579	31	36
C VS LUMB	1008	633	20	12
C VS BRC	1031	610	49	40
C VS IDC	1031	610	39	26
C VS ILB	1008	633	27	9

Tabla 12. GSA con términos GO de funciones moleculares. En la columna "Comparación", C hace referencia a controles sanos contra los diferentes subtipos de cáncer de mama. Over y Under hacen referencia al número de funciones que están más representadas en tumores y controles, respectivamente. Sig.over y Sig.under hace referencia al número de funciones que están más representadas en tumores y controles significativamente de forma respectiva.

COMPARACIÓN	OVER	UNDER	SIG.OVER	SIG.UNDER
C VS BASAL	638	317	18	14
C VS HER2	671	284	2	3
C VS LUMA	725	230	49	4
C VS LUMB	703	252	22	4
C VS BRC	702	253	70	12
C VS IDC	725	230	37	7
C VS ILB	708	247	26	2

Tabla 13. GSA con términos GO de componentes celulares. En la columna "Comparación", C hace referencia a controles sanos contra los diferentes subtipos de cáncer de mama. Over y Under hacen referencia al número de funciones que están más representadas en tumores y controels, respectivamente. Sig.over y Sig.under hace referencia al número de funciones que están más representadas en tumores y controles significativamente de forma respectiva.

Y para la anotación procedente de KEGG, la tabla 14 muestra el resumen de los GSA para cada una de las comparaciones.

COMPARACIÓN	OVER	UNDER	SIG.OVER	SIG.UNDER
C VS BASAL	201	101	3	3
C VS HER2	194	108	0	1
C VS LUMA	198	104	3	4
C VS LUMB	196	106	1	2
C VS BRC	221	81	7	3
C VS IDC	198	104	8	2
C VS ILB	203	99	6	1

Tabla 14. GSA con rutas de KEGG. En la columna "Comparación", C hace referencia a controles sanos contra los diferentes subtipos de cáncer de mama. Over y Under hacen referencia al número de funciones que están más representadas en tumores y controles, respectivamente. Sig.over y Sig.under hace referencia al número de funciones que están más representadas en tumores y controles significativamente de forma respectiva.

Como podemos observar, obtuvimos un mayor número de funciones significativas con la aproximación del promedio. Por lo que esperábamos que el metaanálisis con este abordaje arrojaría un mayor número de funciones significativas entre todos los estudios.

Una vez que obtuvimos los resultados de cada uno de los distintos GSA ejecutamos el último paso de este estudio, el metaanálisis funcional.

4. Metaanálisis funcional

Gracias a la inclusión de todas las muestras y funciones que obtuvimos con los diferentes análisis GSA, el metaanálisis funcional cobró una potencia estadística aún mayor.

Se realizaron dos tipos de metaanálisis en función del abordaje con el que obtuvimos los datos de los GSA, es decir, un metaanálisis para los GSA obtenidos mediante nuestra aproximación por valores absolutos más altos y otra para el promedio.

A su vez, para cada una de las bases de datos empleadas (procesos biológicos, funciones moleculares y componentes celulares de la Gene Ontology, y rutas de KEGG) se ejecutó el metaanálisis en cada uno de los dos escenarios descritos.

4.1 Aproximación valor absoluto

Los resultados para los metaanálisis realizados con la aproximación del valor absoluto se recogen a continuación.

4.1.1 Metaanálisis para los procesos biológicos

El resumen de las funciones significativas (con un p valor ajustado menor o igual a 0.05) para cada uno de los distintos modelos evaluados en el metaanálisis para los procesos biológicos se muestra en la tabla 15.

MODELO	OVER	UNDER	SIG.OVER	SIG.UNDER	SIG.LOR.OVER	SIG.LOR.UNDER
DL	4468	3357	1895	1068	240	133
HE	4647	3357	1892	1069	241	135
HS	4648	3357	1933	1107	243	135
FE	4645	3359	2033	1152	246	135

Tabla 15. Resultados del metaanálisis para los procesos biológicos por la aproximación del valor absoluto. Over y Under hacen referencia al número de funciones que están más metiladas en tumores y controels, respectivamente. Sig.Over y Sig.Under muestran las funciones significativas para un p valor ajustado de 0.05 en tumores y controles. Sig.Lor.Over y Sig.Lor.Under hacen referencia al número de funciones que están más metiladas en tumores y controles de manera significativa y que además tienen un LOR menor que -0.5 o mayor que 0.5. De esta manera nos quedamos con las funciones que tienen al menos un patrón de metilación de 1.41 veces diferente entre los grupos experimentales.

Como indicamos anteriormente, tomamos como referencia los resultados del modelo DL (DerSimonian) por recoger la posible variabilidad interestudios y ajustarse mejor que los demás a nuestras necesidades.

Se detectó un gran número de funciones significativas (hecho esperable en estudios de cáncer). Para centrarnos en aquellas funciones con mayor sobrerrepresentación en alguno de los grupos experimentales, incorporamos un filtro sobre el LOR (el logaritmo del odds ratio). De esta manera nos quedamos con aquellas funciones que tengan un LOR menor de -0.5 y mayor de 0.5. En la tabla 16 podemos ver algunas de estas funciones con más detalle.

ID	NOMBRE	LOWER_BOUND	SUMMARY_LOR	UPPER_BOUND	PVALOR	P.AJUST	QE	QEP	SE
GO:0007221	positive regulation of transcription of Notch receptor target	1.045	1.387	1.729	0	0	6.319	0.388	0.174
GO:1905007	positive regulation of epithelial to mesenchymal transition involved in endocardial cushion formation	0.914	1.362	1.809	0	0	17.723	0.007	0.228
GO:0010519	negative regulation of phospholipase activity	0.686	0.975	1.265	0	0	1.185	0.978	0.148
GO:0060573	cell fate specification involved in pattern specification	0.568	0.836	1.103	0	0	2.451	0.874	0.137
GO:0043652	engulfment of apoptotic cell	0.509	0.772	1.034	0	0	2.775	0.837	0.134
GO:0060581	cell fate commitment involved in pattern specification	0.446	0.66	0.873	0	0	2.224	0.898	0.109
GO:0051006	positive regulation of lipoprotein lipase activity	0.32	0.565	0.809	0	0	5.814	0.444	0.125
GO:0007501	mesodermal cell fate specification	0.186	0.561	0.937	0.003	0.013	21.704	0.001	0.191
GO:0030011	maintenance of cell polarity	0.334	0.546	0.759	0	0	5.4	0.494	0.109
GO:0038169	somatostatin receptor signaling pathway	0.17	0.54	0.91	0.004	0.015	8.868	0.181	0.189
GO:0038170	somatostatin signaling pathway	0.17	0.54	0.91	0.004	0.015	8.868	0.181	0.189
GO:0051893	regulation of focal adhesion assembly	0.333	0.537	0.741	0	0	20.323	0.002	0.104
GO:0090109	regulation of cell-substrate junction assembly	0.333	0.537	0.741	0	0	20.323	0.002	0.104
GO:0061469	regulation of type B pancreatic cell proliferation	0.285	0.529	0.772	0	0	2.299	0.89	0.124
GO:0043654	recognition of apoptotic cell	0.152	0.525	0.898	0.006	0.02	12.227	0.057	0.19
GO:0072602	interleukin-4 secretion	0.147	0.504	0.862	0.006	0.02	2.296	0.891	0.182
GO:0046633	alpha-beta T cell proliferation	-0.657	-0.505	-0.353	0	0	6.076	0.415	0.077
GO:0002726	positive regulation of T cell cytokine production	-0.711	-0.51	-0.309	0	0	3.802	0.703	0.103
GO:0002523	leukocyte migration involved in inflammatory response	-0.866	-0.532	-0.199	0.002	0.008	15.866	0.014	0.17
GO:0038063	collagen-activated tyrosine kinase receptor signaling pathway	-0.926	-0.533	-0.141	0.008	0.025	12.119	0.059	0.2
GO:0044546	NLRP3 inflammasome complex assembly	-0.778	-0.535	-0.292	0	0	1.601	0.952	0.124
GO:2001171	positive regulation of ATP biosynthetic process	-0.893	-0.576	-0.258	0	0.002	1.62	0.951	0.162
GO:0061737	leukotriene signaling pathway	-0.909	-0.592	-0.275	0	0.001	1.255	0.974	0.162
GO:0045953	negative regulation of natural killer cell mediated cytotoxicity	-0.837	-0.601	-0.365	0	0	1.444	0.963	0.12
GO:0090594	inflammatory response to wounding	-1.069	-0.618	-0.167	0.007	0.024	12.445	0.053	0.23
GO:0002309	T cell proliferation involved in immune response	-1.04	-0.72	-0.401	0	0	3	0.809	0.163
GO:2000318	positive regulation of T-helper 17 type immune response	-1.193	-0.931	-0.669	0	0	3.339	0.765	0.134
GO:0002001	renin secretion into blood stream	-1.275	-0.986	-0.697	0	0	5.528	0.478	0.147

Tabla 16. Procesos biológicos significativos comunes a todas las comparaciones estudiadas.

- **QE y QEp** representan el estadístico de contraste y el valor p respectivamente de la prueba de Der Simonian y Laird, utilizada para detectar la presencia de heterogeneidad entre los estudios. La hipótesis nula apunta a la presencia de homogeneidad de los estudios con un intervalo de confianza al 95%. Como podemos observar en la tabla 16 se detecta la presencia de heterogeneidad y se confirma la adecuación de un modelo de efectos aleatorios que incorpore esta variabilidad.

- **LOR** es la estimación del efecto combinado de todos los estudios. Es el logaritmo del odds ratio. El signo negativo indica mayor presencia de genes con niveles de metilación altos en la segunda clase experimental (tumores) respecto a la primera (controles sanos). La magnitud de este indicador cuantifica la sobrerrepresentación de la función en un grupo frente al otro.

- La estimación del efecto estudiado se acompaña de su intervalo de confianza al 95% construido con la variabilidad estimada en el modelo seleccionado (**SE**). **Lower** y **upper bound** además, hacen referencia al intervalo de confianza descrito. La no inclusión del valor 0 en el intervalo, confirmaría la significatividad del LOR.

- El **p valor** nos informa del nivel de significación de una función dada y el **p valor ajustado** es el nivel de significación corregido mediante el método de Benjamini y Hochberg (Benjamini & Hochberg, 1995).

Además, podemos profundizar aún más en una función específica. Por ejemplo para la función GO:0043654, el reconocimiento de una célula apoptótica, podemos ver el efecto aportado de cada estudio en la estimación del efecto global que tiene esta función en la figura 24.

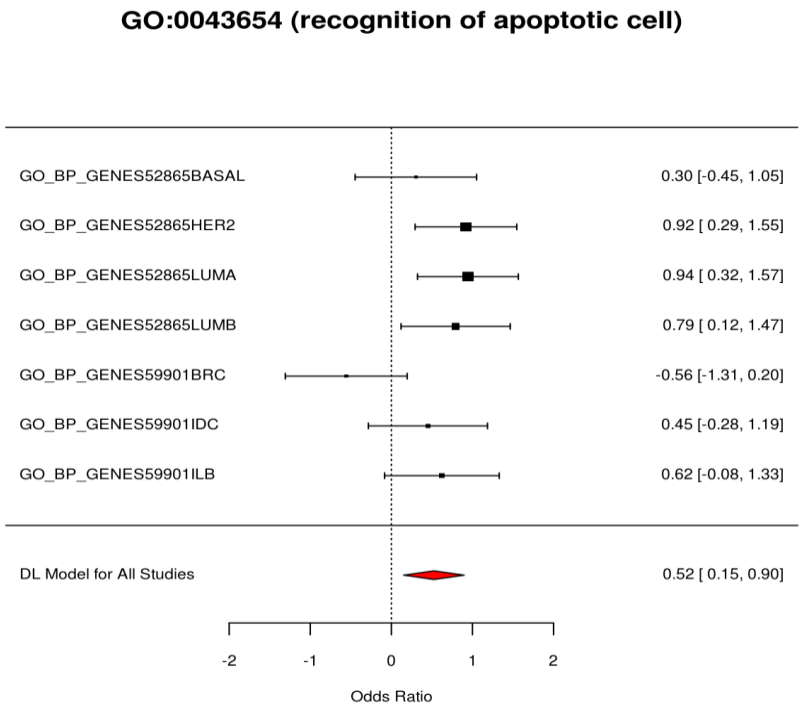

Figura 24. Efecto aportado de cada uno de los estudios/comparaciones al efecto global de la función GO:0043654.

En la figura 24:

- A la izquierda de la figura se enumeran los estudios incluidos en el metaanálisis.

- En la parte derecha se incluye la estimación de la medida resumen individual de cada estudio y su intervalo de confianza.

- En el centro de la figura se visualiza la medida del efecto (en forma de cuadrado negro) cuyo tamaño es inversamente proporcional a la variabilidad presente en el estudio/comparación en concreto. El cuadro está ubicado dentro de un segmento que representa los extremos de su intervalo de confianza. Si el intervalo de confianza de la estimación del efecto no incluye el valor 0, entonces el nivel de sobrerrepresentación será significativo. En caso contrario (inclusión del valor 0 en el intervalo de confianza), el nivel de sobrerrepresentación determinado no será significativo.

- En la parte inferior, se representa el resultado global del metaanálisis en forma de un rombo de color rojo. Su posición con respecto a la línea del efecto nulo también nos informa sobre la significación estadística global (en este caso 0.52) mientras que su anchura nos informa de su precisión (intervalo de confianza). El valor de 0.52 apunta a un mayor nivel de metilación diferencial en las muestras de cáncer de mama respecto a las muestras de tejido sano.

- También en la parte inferior derecha tenemos el valor de significación de los intervalos de confianza.

La presencia de algunos estudios cuya magnitud y variabilidad son muy diferentes del resto pueden producir una fuerte influencia en el metaanálisis. En la figura 25 podemos ver un análisis de estudios influyentes para la misma función, GO:0043654.

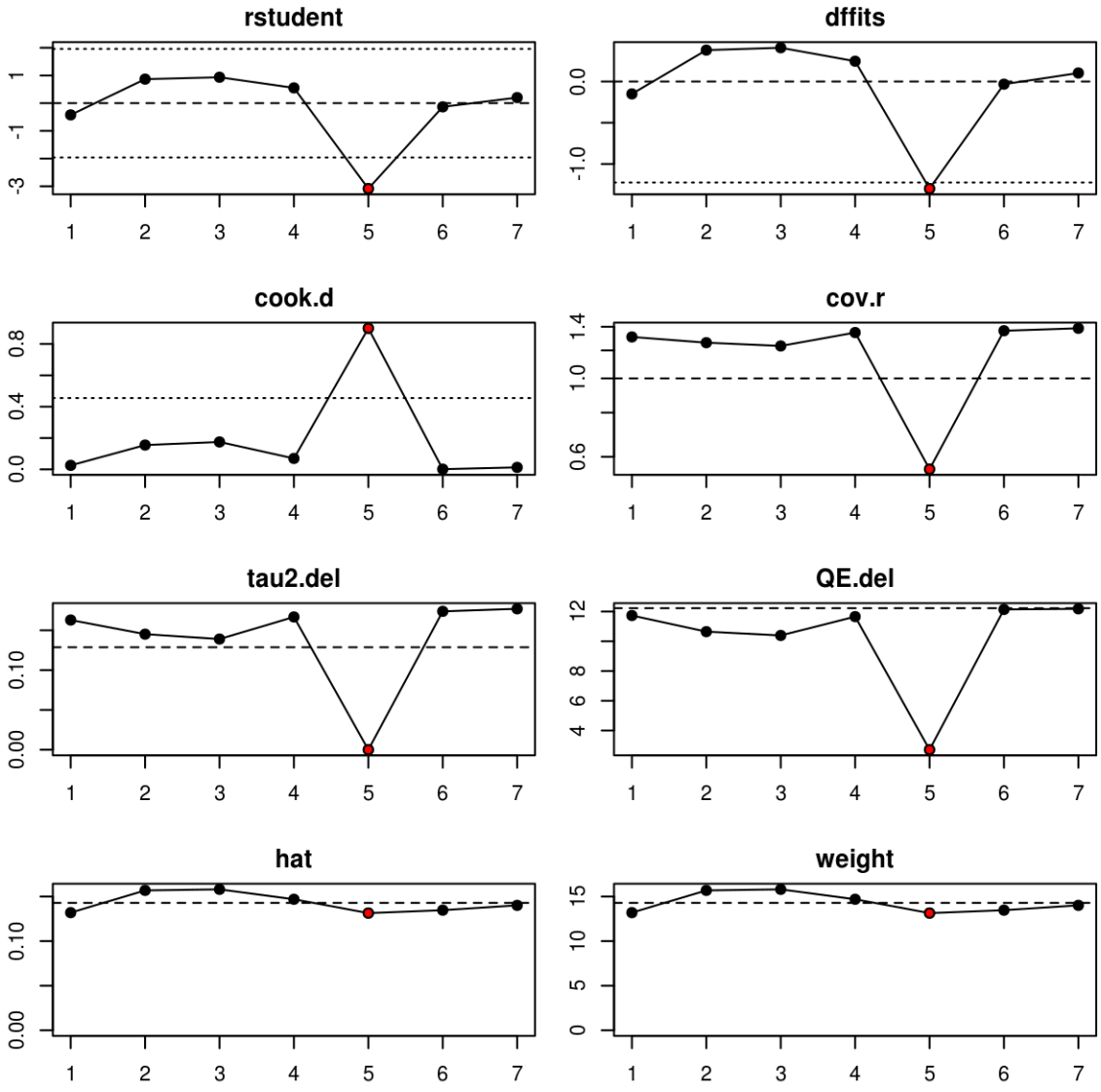

Figura 25. Análisis de estudios o comparaciones influyentes para la función GO:0043654. Entre las medidas de diagnóstico de datos influyentes se recogen: residuos estandarizados, distandias de Cook, covarianza de ratios, etc.

El diagnóstico de los datos influyentes para la función evaluada, detecta un comportamiento diferente para el estudio o comparación 5, tal como queda reflejado en el descenso de QE y el incremento en la distancia de Cook. La valoración de esta comparación en el resto de funciones, proporcionará información para mantenerla o excluirla del estudio global.

Otro tipo de representación gráfica para extraer información acerca de una función en específico en el metaanálisis son los gráficos de embudo. Mediante este tipo de gráficos se evalúa la variabilidad de los distintos estudios, así como los sesgos. En este tipo de representaciones se muestra la magnitud del efecto medido (eje X) mediante a una medida de precisión (eje Y), como la desviación estándar o el inverso de la varianza. Cada punto

representa un estudio/comparación primario/a y las conclusiones sobre el gráfico se extraen tras el análisis de la nube de puntos (Sterne et al., 2001). La figura 26 presenta la relación entre el efecto estudiado y los distintos indicadores de la variabilidad para la función GO:0043654.

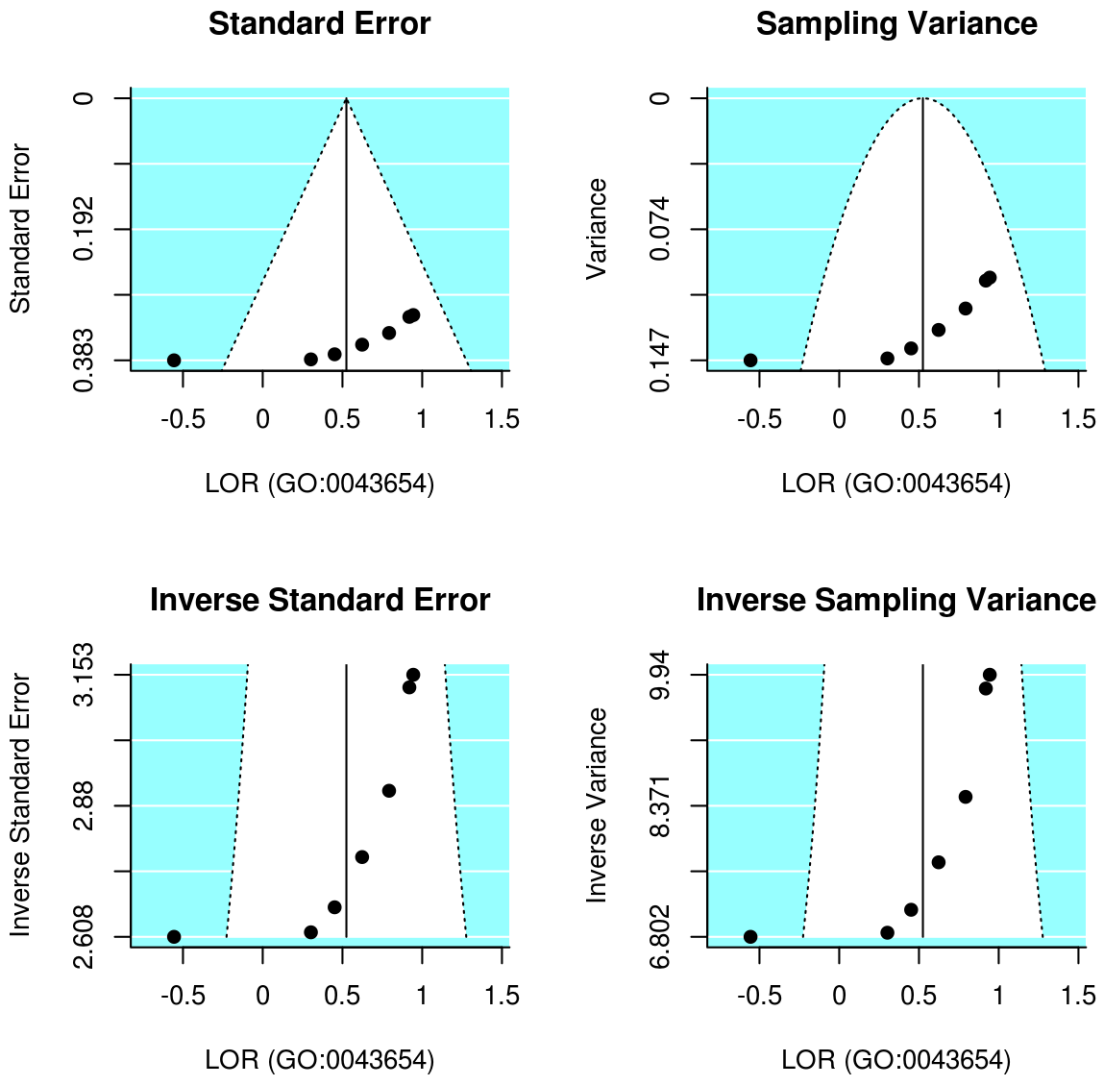

Figura 26. Relación entre el efecto estudiado y los distintos indicadores de la variabilidad para la función GO:0043654. El eje X representa los valores del logaritmo del odds ratio y el eje Y: el error estándar, la varianza en el muestreo, y sus respectivos inversos (medidas de precisión).

Como podemos apreciar tras analizar los distintos gráficos que componen la figura 26 todos los estudios menos uno (cáncer tipo brc) quedan dentro de la región de confianza. Para determinar si mantenemos o excluimos este estudio deberíamos revisar su comportamiento en el resto de funciones. Podemos observar leves cambios en la distribución de la variabilidad de los dos primeros gráficos (error estándar o varianza) entre los estudios que crece a medida que lo hace el LOR. Por último, vemos como los estudios con LOR por encima de 0.5 muestran un ligero incremento de error estándar y variabilidad que queda al descubierto cuando observamos los dos gráficos restantes que recogen los valores inversos del error estándar o varianza.

4.1.2 Metaanálisis para las funciones moleculares

El resumen de las funciones significativas (con un p valor ajustado menor o igual a 0.05) y con LOR menor a -0.5 y superior a 0.5 para cada uno de los distintos modelos evaluados en el metaanálisis para las funciones moleculares se recoge en la tabla 17.

MODELO	OVER	UNDER	SIG.OVER	SIG.UNDER	SIG.LOR.OVER	SIG.LOR.UNDER
DL	988	653	363	199	52	41
HE	988	653	364	201	52	41
HS	988	653	375	205	52	42
FE	990	651	385	213	52	42

Tabla 17. Resultados del metaanálisis para las funciones moleculares por la aproximación del valor absoluto.

Encontramos una descripción más detallada de algunas de las funciones moleculares significativas y con LOR menor que -0.5 y mayor que 0.5 para el modelo Der Simonian en la tabla 18.

4.1.3 Metaanálisis para los componentes celulares

El resumen de las funciones significativas (con un p valor ajustado menor o igual a 0.05) y con LOR menor a -0.5 y superior a 0.5 para cada uno de los distintos modelos evaluados en el metaanálisis para las funciones moleculares se recoge en la tabla 19.

MODELO	OVER	UNDER	SIG.OVER	SIG.UNDER	SIG.LOR.OVER	SIG.LOR.UNDER
DL	643	311	257	68	30	10
HE	643	311	257	68	30	10
HS	643	311	263	69	30	10
FE	643	311	272	72	30	10

Tabla 19. Resultados del metaanálisis para los componentes celulares por la aproximación del valor absoluto.

La descripción detallada de algunos de los términos representados de manera significativa y con un LOR menor a -0.5 y superior a 0.5 para el modelo de Der Simonian se encuentra en la tabla 20.

4.1.4 Metaanálisis para las rutas KEGG

La tabla 21 recoge el número de rutas KEGG con un patrón de metilación diferencial significativa y común entre las distintas comparaciones que captaron los distintos modelos evaluados en el metaanálisis. Se acompaña también con el filtro basado en el LOR.

MODELO	OVER	UNDER	SIG.OVER	SIG.UNDER	SIG.LOR.OVER	SIG.LOR.UNDER
DL	191	110	70	17	0	1
HE	191	110	70	17	0	1
HS	191	110	70	19	0	1
FE	191	110	73	19	0	1

Tabla 21. . Resultados del metaanálisis para las rutas KEGG por la aproximación del valor absoluto.

La tabla 22 contiene la única ruta con un nivel de metilación diferencial común y significativa (además del filtro con el LOR) de todas las comparaciones realizadas para el metaanálisis de los términos KEGG.

ID	NOMBRE	LOWER_BOUND	SUMMARY_LOR	UPPER_BOUND	PVALOR	P.AJUST	QE	QEP	SE
GO:0001727	lipid kinase activity	0.563	0.888	1.212	0	0	1.47	0.961	0.166
GO:0019966	interleukin-1 binding	0.314	0.595	0.876	0	0	7.541	0.274	0.143
GO:0030368	interleukin-17 receptor activity	0.26	0.564	0.868	0	0.002	2.546	0.863	0.155
GO:0042609	CD4 receptor binding	0.212	0.563	0.914	0.002	0.008	3.062	0.801	0.179
GO:0004716	receptor signaling protein tyrosine kinase activity	-0.868	-0.521	-0.174	0.003	0.013	10.338	0.111	0.177
GO:0004955	prostaglandin receptor activity	-0.932	-0.649	-0.366	0	0	8.577	0.199	0.144
GO:0035529	NADH pyrophosphatase activity	-1.023	-0.678	-0.333	0	0.001	1.508	0.959	0.176
GO:0004957	prostaglandin E receptor activity	-1.335	-0.902	-0.468	0	0	9.558	0.145	0.221

Tabla 18. Funciones moleculares significativas comunes a todas las comparaciones estudiadas para el modelo Der Simonian. LOR filtrado menor que -0.5 y mayor que 0.5.

ID	NOMBRE	LOWER_BOUND	SUMMARY_LOR	UPPER_BOUND	PVALOR	P.AJUST	QE	QEP	SE
GO:0005915	zonula adherens	0.203	0.653	1.103	0.004	0.017	17.781	0.007	0.229
GO:0016342	catenin complex	0.242	0.517	0.792	0	0.002	1.419	0.965	0.14
GO:0042101	T cell receptor complex	-0.773	-0.589	-0.405	0	0	5.486	0.483	0.094
GO:0061702	inflammasome complex	-0.829	-0.616	-0.403	0	0	1.448	0.963	0.109
GO:0042105	alpha-beta T cell receptor complex	-0.922	-0.62	-0.317	0	0.001	4.529	0.605	0.154
GO:0097169	AIM2 inflammasome complex	-0.979	-0.678	-0.376	0	0	1.421	0.965	0.154

Tabla 20.Componentes celulares significativos comunes a todas las comparaciones estudiadas para el modelo Der Simonian. LOR filtrado menor que -0.5 y mayor que 0.5.

ID	NOMBRE	LOWER_BOUND	SUMMARY_LOR	UPPER_BOUND	PVALOR	P.AJUST	QE	QEP	SE
HSA00780	Biotin metabolism	-1.064	-0.75	-0.436	0	0	6.277	0.393	0.16

Tabla 22. Ruta KEGG con un nivel de metilación diferencial significativo y común a todas las comparaciones estudiadas para el modelo Der Simonian. LOR filtrado menor que -0.5 y mayor que 0.5.

4.2 Aproximación con el promedio

Los resultados de los diferentes metaanálisis para el segundo método de resumen de la señal de metilación, el promedio, se detallan en adelante. Dada la enorme cantidad de datos que obtuvimos (más que en el metaanálisis con los valores absolutos), se filtran también estos resultados por el LOR. Filtramos para obtener aquellas funciones con un LOR menor a -0.5 o mayor a 0.5, lo que quiere decir que nos quedamos con las funciones que al menos están 1.41 veces (aproximadamente, $2^{0.5} = 1.414$) más metiladas en alguno de los dos grupos experimentales.

4.2.1 Metaanálisis para los procesos biológicos

La tabla 23 muestra un resumen con los procesos biológicos con una metilación diferencial común y significativa (con un p valor ajustado menor o igual a 0.05) para cada uno de los modelos evaluados.

MODELO	OVER	UNDER	SIG.OVER	SIG.UNDER	SIG.LOR.OVER	SIG.LOR.UNDER
DL	4939	3069	2798	1428	314	225
HE	4939	3069	2798	1432	317	227
HS	4939	3069	2804	1440	317	226
FE	4939	3069	2817	1447	316	227

Tabla 23. Procesos biológicos resultantes tras el metaanálisis con la aproximación del promedio. Sig.Lor.Over y Sig.Lor.Under hacen referencia al filtro impuesto con el LOR.

Una descripción más detallada de algunos de los procesos significativos que identificamos por el modelo Der Simonian se encuentra en la tabla 24.

4.2.2 Metaanálisis para las funciones moleculares

La tabla 25 muestra un resumen con las funciones moleculares con una metilación diferencial común y significativa (con un p valor ajustado menor o igual a 0.05) para cada uno de los modelos evaluados.

MODELO	OVER	UNDER	SIG.OVER	SIG.UNDER	SIG.LOR.OVER	SIG.LOR.UNDER
DL	1030	611	570	301	29	26
HE	1030	611	570	299	29	25
HS	1030	611	570	301	29	26
FE	1030	611	570	301	29	28

Tabla 25. Funciones moleculares resultantes tras el metaanálisis con la aproximación del promedio. Sig.Lor.Over y Sig.Lor.Under hacen referencia al filtro impuesto con el LOR.

La descripción de algunas de estas funciones la encontramos de manera más detallada en la tabla 26.

4.2.3 Metaanálisis para los componentes celulares

La tabla 27 muestra un resumen con componentes celulares con una metilación diferencial y significativa (con un p valor ajustado menor o igual a 0.05) para cada uno de los modelos evaluados.

MODELO	OVER	UNDER	SIG.OVER	SIG.UNDER	SIG.LOR.OVER	SIG.LOR.UNDER
DL	713	242	458	100	67	24
HE	713	242	458	100	67	24
HS	713	242	458	103	67	24
FE	713	242	458	103	67	24

Tabla 27. Componentes celulares resultantes tras el metaanálisis con la aproximación del promedio. Sig.Lor.Over y Sig.Lor.Under hacen referencia al filtro impuesto con el LOR.

En la tabla 28 podemos observar algunos componentes celulares significativos que identificamos con el modelo de Der Simonian.

4.2.4 Metaanálisis para las rutas KEGG

La tabla 29 muestra un resumen con los procesos biológicos con una metilación diferencial y significativa (con un p valor ajustado menor o igual a 0.05) para cada uno de los modelos evaluados.

MODELO	OVER	UNDER	SIG.OVER	SIG.UNDER	SIG.LOR.OVER	SIG.LOR.UNDER
DL	212	90	121	42	2	1
HE	212	90	121	42	2	1
HS	212	90	122	42	2	1
FE	212	90	123	43	2	1

Tabla 29. Rutas KEGG resultantes tras el metaanálisis con la aproximación del promedio. Sig.Lor.Over y Sig.Lor.Under hacen referencia al filtro impuesto con el LOR.

Las rutas provenientes de KEGG significativas encontradas en el metaanálisis con la aproximación del promedio en el modelo Der Simonian aparecen con detalle en la tabla 30.

Como podemos comprobar, finalmente el número de funciones que detectamos con el metaanálisis funcional con la aproximación del promedio mayor es mayor que con la aproximación del valor absoluto más alto. Normalmente, esperaríamos que todo vaya en la misma línea: más funciones significativas en un estudio individual, más funciones en el metaanálisis.

En el análisis de este tipo de datos ómicos, al trabajar con promedios, “suavizamos la señal” y, dentro del mismo estudio, esto repercute en un mayor nivel de significación funcional. De ahí que obtengamos una mayor batería de resultados. Por otra parte, trabajando con valores absolutos, se mantiene una señal inicial “sin suavizar” que captura un menor número de funciones por estudio.

ID	NOMBRE	LOWER_BOUND	SUMMARY_LOR	UPPER_BOUND	PVALOR	P.AJUST	QE	QEP	SE
GO:0032055	negative regulation of translation in response to stress	0.805	1.018	1.231	0	0	1.049	0.984	0.109
GO:0042661	regulation of mesodermal cell fate specification	0.607	0.862	1.118	0	0	5.326	0.503	0.13
GO:0009996	negative regulation of cell fate specification	0.561	0.859	1.157	0	0	6.85	0.335	0.152
GO:0033129	positive regulation of histone phosphorylation	0.526	0.793	1.06	0	0	1.947	0.925	0.136
GO:0042659	regulation of cell fate specification	0.575	0.77	0.965	0	0	4.508	0.608	0.1
GO:2000049	positive regulation of cell-cell adhesion mediated by cadherin	0.468	0.755	1.041	0	0	0.398	0.999	0.146
GO:2000323	negative regulation of glucocorticoid receptor signaling pathway	0.444	0.751	1.058	0	0	0.376	0.999	0.157
GO:0010519	negative regulation of phospholipase activity	0.448	0.739	1.031	0	0	1.281	0.973	0.149
GO:0009912	auditory receptor cell fate commitment	0.4	0.69	0.981	0	0	2.273	0.893	0.148
GO:0001768	establishment of T cell polarity	0.401	0.669	0.937	0	0	1.491	0.96	0.137
GO:0035912	dorsal aorta morphogenesis	0.396	0.66	0.923	0	0	0.133	1	0.134
GO:0036072	direct ossification	0.356	0.659	0.963	0	0	2.272	0.893	0.155
GO:0072592	oxygen metabolic process	0.329	0.656	0.983	0	0	1.219	0.976	0.167
GO:1902231	positive regulation of intrinsic apoptotic signaling pathway in response to DNA damage	0.38	0.641	0.902	0	0	0.455	0.998	0.133
GO:1902237	positive regulation of endoplasmic reticulum stress-induced intrinsic apoptotic signaling pathway	0.409	0.628	0.847	0	0	0.508	0.998	0.112
GO:0038161	prolactin signaling pathway	0.254	0.581	0.909	0	0.002	1.367	0.968	0.167
GO:0032633	interleukin-4 production	-0.657	-0.529	-0.4	0	0	1.756	0.941	0.066
GO:0002836	positive regulation of response to tumor cell	-0.797	-0.534	-0.271	0	0	0.47	0.998	0.134
GO:0002839	positive regulation of immune response to tumor cell	-0.797	-0.534	-0.271	0	0	0.47	0.998	0.134
GO:0015732	prostaglandin transport	-0.718	-0.534	-0.351	0	0	2.163	0.904	0.094
GO:0002418	immune response to tumor cell	-0.786	-0.574	-0.361	0	0	1.381	0.967	0.108
GO:0002347	response to tumor cell	-0.785	-0.589	-0.393	0	0	1.204	0.977	0.1
GO:2000319	regulation of T-helper 17 cell differentiation	-0.877	-0.604	-0.331	0	0	1.63	0.95	0.139
GO:0002001	renin secretion into blood stream	-1.3	-1.022	-0.743	0	0	1.289	0.972	0.142
GO:2001185	regulation of CD8-positive, alpha-beta T cell activation	-1.292	-1.051	-0.811	0	0	3.655	0.723	0.123
GO:2000564	regulation of CD8-positive, alpha-beta T cell proliferation	-1.439	-1.144	-0.85	0	0	4.06	0.669	0.15
GO:0002309	T cell proliferation involved in immune response	-1.465	-1.181	-0.898	0	0	0.572	0.997	0.145

Tabla 24. Procesos biológicos significativos comunes a todas las comparaciones estudiadas para el modelo Der Simonian. LOR filtrado menor que -0.5 y mayor que 0.

ID	NOMBRE	LOWER_BOUND	SUMMARY_LOR	UPPER_BOUND	PVALOR	P.AJUST	QE	QEP	SE
GO:0001055	RNA polymerase II activity	0.89	1.095	1.3	0	0	1.889	0.93	0.105
GO:0001727	lipid kinase activity	0.669	0.981	1.293	0	0	2.163	0.904	0.159
GO:0008440	inositol-1,4,5-trisphosphate 3-kinase activity	0.286	0.531	0.776	0	0	1.045	0.984	0.125

Tabla 26. Funciones moleculares significativas comunes a todas las comparaciones estudiadas para el modelo Der Simonian. LOR filtrado menor que -0.5 y mayor que 0.5.

ID	NOMBRE	LOWER_BOUND	SUMMARY_LOR	UPPER_BOUND	PVALOR	P.AJUST	QE	QEP	SE
GO:1990393	3M complex	0.719	1.032	1.345	0	0	1.183	0.978	0.16
GO:0072357	PTW/PP1 phosphatase complex	0.399	0.676	0.953	0	0	0.74	0.994	0.141
GO:0061702	inflammasome complex	-0.848	-0.635	-0.421	0	0	0.991	0.986	0.109
GO:0042101	T cell receptor complex	-0.894	-0.7	-0.506	0	0	7.108	0.311	0.099
GO:0097169	AIM2 inflammasome complex	-1.133	-0.824	-0.515	0	0	2.744	0.84	0.158
GO:0042105	alpha-beta T cell receptor complex	-1.302	-1.001	-0.7	0	0	6.899	0.33	0.154

Tabla 28.Componentes celulares significativos comunes a todas las comparaciones estudiadas para el modelo Der Simonian. LOR filtrado menor que -0.5 y mayor que 0.5.

ID	NOMBRE	LOWER_BOUND	SUMMARY_LOR	UPPER_BOUND	PVALOR	P.AJUST	QE	QEP	SE
HSA01220	Degradation of aromatic compounds	0.279	0.547	0.815	0	0	1.04	0.984	0.137
HSA04950	Maturity onset diabetes of the Young	0.369	0.508	0.647	0	0	7.984	0.239	0.071
HSA04740	Olfactory transduction	-1.489	-1.216	-0.942	0	0	142.996	0	0.139

Tabla 30. Rutas KEGG con un nivel de metilación diferencial significativo y común a todas las comparaciones estudiadas para el modelo Der Simonian. LOR filtrado menor que -0.5 y mayor que 0.5.

Como se puede apreciar, conseguimos un gran número de resultados, ya que se efectuaron diferentes abordajes en distintos escenarios y para cada uno de ellos, disponemos de una batería de resultados gráficos y numéricos por función. En el presente trabajo se muestran solo algunos de ellos, a nivel informativo, pero todos están disponibles para su revisión detallada (consultar apartado Anexos). Resaltar además que estos resultados se muestran en diferentes niveles: desde lo más general (tablas con las funciones obtenidas para cada método de metaanálisis) a lo más específico (gráficas e indicadores por función).

Discusión de los resultados

De los dos abordajes realizados para resumir la información de las *DMR* a nivel de gen, el que determina el promedio proporciona un mayor número de resultados. Esto no es de extrañar, al contrario, era algo esperado ya que al promediar, como se comentaba anteriormente, aumentamos la significación de las funciones de los estudios individuales (también obtuvimos un mayor número de funciones significativas en los GSA pertenecientes al promedio). Sin embargo, en un contexto biológico, promediar regiones dispares (aunque apunten al mismo gen) puede no ser del todo correcto. Con esto nos referimos a promediar por ejemplo una *DMR* que apunta a la región promotora de un gen con alguna otra que esté más distal y cuya dirección (si se encuentra más metilada en controles o casos) puede ser contraria a la primera, de manera que estaríamos suavizando y enmascarando la señal.

Los mecanismos que hacen que un gen se exprese o no, dependiendo de las *DMR* que apuntan a él, parecen cobrar más sentido cuando resumimos la información de metilación con la aproximación del valor absoluto más alto. En esta aproximación tenemos una única señal sin suavizar, donde además, obtenemos resultados más concluyentes y robustos después de una completa revisión de los mismos. Con esto último nos referimos a resultados esperados en cáncer y que confirman la adecuación de la aproximación. Con el promedio obtenemos una mayor cantidad de resultados, pero entre ellos tenemos bastantes que no coinciden con lo esperado en la literatura e incluso son contrarios a la misma como discutiremos a continuación.

1. Metaanálisis para los procesos biológicos

Empecemos analizando el metaanálisis para los procesos biológicos con la aproximación del valor absoluto, cuyos resultados se muestran en la tabla 31 de manera simplificada para la interpretación funcional de los mismos.

ID	NOMBRE	SUMMARY_LOR
GO:0007221	positive regulation of transcription of Notch receptor target	1.387
GO:1905007	positive regulation of epithelial to mesenchymal transition involved in endocardial cushion formation	1.362
GO:0010519	negative regulation of phospholipase activity	0.975
GO:0060573	cell fate specification involved in pattern specification	0.836
GO:0043652	engulfment of apoptotic cell	0.772
GO:0060581	cell fate commitment involved in pattern specification	0.66
GO:0051006	positive regulation of lipoprotein lipase activity	0.565
GO:0007501	mesodermal cell fate specification	0.561
GO:0030011	maintenance of cell polarity	0.546
GO:0038169	somatostatin receptor signaling pathway	0.54
GO:0038170	somatostatin signaling pathway	0.54
GO:0051893	regulation of focal adhesion assembly	0.537
GO:0090109	regulation of cell-substrate junction assembly	0.537
GO:0061469	regulation of type B pancreatic cell proliferation	0.529
GO:0043654	recognition of apoptotic cell	0.525
GO:0072602	interleukin-4 secretion	0.504
GO:0046633	alpha-beta T cell proliferation	-0.505
GO:0002726	positive regulation of T cell cytokine production	-0.51
GO:0002523	leukocyte migration involved in inflammatory response	-0.532
GO:0038063	collagen-activated tyrosine kinase receptor signaling pathway	-0.533
GO:0044546	NLRP3 inflammasome complex assembly	-0.535
GO:2001171	positive regulation of ATP biosynthetic process	-0.576
GO:0061737	leukotriene signaling pathway	-0.592
GO:0045953	negative regulation of natural killer cell mediated cytotoxicity	-0.601
GO:0090594	inflammatory response to wounding	-0.618
GO:0002309	T cell proliferation involved in immune response	-0.72
GO:2000318	positive regulation of T-helper 17 type immune response	-0.931
GO:0002001	renin secretion into blood stream	-0.986

Tabla 31. Términos GO referentes a procesos biológicos para la aproximación con el valor absoluto, se muestran únicamente el ID, nombre y el logaritmo del odds ratio.

Cuando trabajamos con datos de cáncer es frecuente que aparezcan un gran número de funciones alteradas, incluso algunas de ellas sin una aparente conexión con el cáncer. En nuestro caso en particular, hemos seleccionado para la discusión aquellas funciones que guardan una relación clara y directa y otras donde deberemos inferirla.

La manera de interpretar los resultados de metilación es a través del LOR. La forma en la cual hacerlo se indica a continuación para el GO:0043654 (reconocimiento de una célula apoptótica):

- Esta función es significativa y posee un LOR de 0.525, lo que quiere decir que se encuentra metilada diferencialmente en las muestras correspondientes a los subtipos de cáncer de mama. Cuando hablamos de metilación en epigenómica, normalmente se asocia un alto nivel de metilación con un bajo nivel de expresión de los genes asociados. En este caso, la interpretación de este resultado significaría que un grupo de genes cuya función es reconocer una célula apoptótica (posiblemente para degradarla y eliminarla) se encuentran con un perfil de expresión bajo en las muestras cancerosas con respecto a las muestras sanas. Que esta función se encuentre con un nivel de metilación alto en las muestras cancerosas va en consonancia con la literatura y con lo que esperábamos. Uno de los *hallmarks* del cáncer es su evasión a la muerte celular programada (Hanahan & Weinberg, 2011) (Poon, Lucas, Rossi, & Ravichandran, 2014). Normalmente la apoptosis celular se lleva a cabo en respuesta al estrés fisiológico, siendo en las células cancerosas la inestabilidad del ADN y las continuas divisiones. En este caso, la consecución de la evasión a la muerte celular programada estaría siendo posible gracias al silenciamiento de una batería de genes encargados de reconocer células apoptóticas.

Teniendo presente estas cuestiones, pasamos a analizar y discutir los resultados del metaanálisis para los términos GO referentes a procesos biológicos.

Empezando por la función con mayor nivel de metilación diferencial en las muestras cancerosas, la denominada con el término GO:0007221 que hace referencia a la regulación positiva de la transcripción del receptor de la proteína Notch. Como hemos mencionado anteriormente, nos encontramos con una función con un fuerte nivel de metilación en las células cancerosas, lo que deriva, a priori, en un bajo nivel de expresión de los genes encargados de llevarla a cabo. Según la literatura (Dotto, 2008), la ruta de señalización de Notch lleva a cabo un papel importante en la supresión de tumores de distintos carcinomas en la piel. Según concluyen, es posible además que durante esta cascada de señalización p53 (otra importante proteína supresora de tumores) se una a los receptores de Notch para que lleven a cabo su actividad. Con esto en mente, no es de extrañar que esta maquinaria supresora de tumores se encuentre con un perfil de expresión bajo (silenciado mediante metilación) en las muestras pertenecientes a cáncer de mama, lo que derivaría en una mayor proliferación de las células cancerosas.

El término GO:1905007 no presenta una conexión directa con el cáncer de mama. Nos encontramos frente a un silenciamiento de los genes encargados de regular positivamente la transición de tejido epitelial a mesenquimal en la formación de los cojines endocárdicos. En un principio, no esperamos esta función dentro de las células de cáncer de mama. Además, al pasar de tejido epitelial a mesenquimal las células ganan capacidad migratoria, invasiva y de resistencia frente la apoptosis (Kalluri & Weinberg, 2009). Estas capacidades recuerdan a las que poseen las células cancerosas por lo que un silenciamiento de estos genes que impidan estas capacidades no es esperado en la literatura. Aun así, al tratarse de una función envuelta en la formación de tejido endocárdico no guarda una relación directa con nuestras células tumorales de cáncer de mama.

Los genes encargados de la regulación negativa de la actividad de la fosfolipasa, recogidos en el término GO:0010519, tienen un patrón de metilación alto en las muestras tumorales (LOR positivo), lo que conlleva un nivel de expresión bajo. En otras palabras, los genes que se encargan de amortiguar la actividad de la fosfolipasa están siendo silenciados en las muestras cancerosas. Este es otro resultado que confirma la adecuación de nuestro método de estudio de metilación, ya que las fosfolipasas son importantes mediadoras en las señales intracelulares e intercelulares. Éstas generan mediadores lipídicos que pueden actuar promoviendo la génesis de tumores (Park et al., 2012). Estos mecanismos de promoción de tumores incluyen la proliferación, migración, invasión y angiogénesis.

En los términos GO:0060573 y GO:0060581, relativos a la expresión celular que las células tienen marcadas como destino desde que se forman, el alto grado de metilación en las células cancerosas respecto a las normales (control) indica que la expresión de los genes que regulan estas funciones está silenciada. En el cáncer, muchas de las células se encuentran indiferenciadas y con formas anormales, por lo que no es raro encontrarnos con que estas funciones se encuentren silenciadas. Esto permite a las células cancerosas permanecer indiferenciadas o escasamente diferenciadas, de manera que proliferan de una manera más agresiva e irregular (National Cancer Institute, 2013).

La actividad de la lipoproteína lipasa (LPL) recogida bajo el término GO:0051006 tiene un perfil de metilación alto en las muestras cancerosas respecto a las controles sanas. Es decir, tenemos una baja expresión de los genes que regulan la actividad de esta proteína. La lipoproteína lipasa hidroliza triglicéridos en ácidos grasos libres que pueden ser utilizados para el crecimiento de las células. Ciertos estudios (Kuemmerle et al., 2011) afirman que la presencia de LPL es crucial en diversos tipos de cáncer, donde el aporte de ácidos grasos libres por su parte a partir de la toma de triglicéridos en el torrente sanguíneo promueve el crecimiento de este tipo de células. Además, afirman que en la línea de células cancerosas HeLa (la línea cancerosa más antigua que se conserva), el "*knockdown*" del gen que codifica esta proteína tiene como resultado una inhibición del crecimiento de las mismas. Otros en cambio (Kinlaw, Quinn, Wells, Roser-Jones, & Moncur, 2006), más en la línea de nuestros resultados encontraron que las células de cáncer de mama no expresan LPL. De esta manera, las células de cáncer de mama sólo tendrían acceso a los lípidos circundantes que proporcionase su ambiente local. Detallan como además, esto explica el motivo por el que diversos cánceres de mama no sobreviven en áreas menos ricas en ácidos grasos como los nódulos linfáticos en ausencia de la expresión de Spot14, una proteína nuclear que suple la falta de LPL en el cáncer de mama (media la lipogénesis a través de progestina).

Uno de los *hallmarks* del cáncer aparece bajo el término GO:0030011. La función que refiere este término es la del mantenimiento de la polaridad de la célula. Como podemos observar, se encuentran con un nivel de metilación mayor en las muestras tumorales (LOR de 0.546) lo que indica que los genes responsables del mantenimiento de esta polaridad se expresan menos en este tipo de células que en las sanas. Estamos dentro de lo que esperamos, puesto que el mantenimiento de la polaridad de la célula es necesaria para la homeostasis en mamíferos, pero en nuestras células tumorales ésta no se mantiene. Varias proteínas cruciales para el mantenimiento de la polaridad de la célula son conocidas como

protooncogenes (genes cuyos productos promueven el crecimiento y división de las células los cuales mutados y desprovistos de regulación, son causantes de cáncer) o genes supresores de tumores. La desregulación del mantenimiento de la polaridad de la célula es motivo de cáncer en sus células hijas debido a divisiones anormales de la misma que no repartan adecuadamente el material genético (Lee & Vasioukhin, 2008).

La vía de señalización de la somatostatina aparece con los términos GO:0038169 y GO:0038170. Los receptores de somatostina inician una ruta de señalización que desemboca en un mayor nivel de apoptosis y menor proliferación. Según nuestros resultados, esta vía se encuentra más apagada (menos expresada) en los individuos con cáncer de mama. Esto parece en un principio lógico, rutas que culminan con una señal apoptótica contraria a la proliferación suelen estar apagadas en varios tipos de cáncer. Sin embargo, según varios estudios (Evers, Parekh, Townsend, Thompson, & Thompson, 1991; Kharmate, Rajput, Lin, & Kumar, 2013) los receptores de somatostatina se encuentran en gran cantidad en células de mama cancerosas. Esto parece indicar que en nuestros estudios, algún punto de la ruta de señalización de la somatostatina podría estar afectada. Según la literatura citada anteriormente, es posible que la somatostatina active a genes supresores de tumores como PTEN y p53, por lo que parece que puede que sea este punto de la vía de señalización de la somatostatina el que se encuentre con un nivel bajo de expresión en nuestras muestras. Aun así, tal es la potencia del tratamiento con somatostatina que se han abierto líneas de investigación para combatir el cáncer a través de sus análogos. Esta es una ruta de señalización que valdría la pena estudiar con más esfuerzo para descubrir qué punto de la ruta pudiese estar afectado.

Podemos comprobar que funciones como la regulación del conjunto de adherencia focal, así como la regulación del ensamblaje de unión célula-substrato, presentan un nivel de metilación mayor en las células tumorales (GO:0051893 y GO:0090109, respectivamente). Esto podría entorpecer la unión de las células cancerosas a la matriz extracelular de manera que quedaran más libres y tuviesen mayor capacidad migratoria e invasiva.

En la tónica de componentes de la matriz extracelular encontramos también la función representada por el término GO:0038063 que corresponde a la vía de señalización del receptor tirosina-quinasa activado por colágeno. En esta función podemos observar un mayor nivel de metilación en las muestras controles respecto a las cancerosas. Esto implica que pudiera ser que la vía de señalización estuviese menos expresada en individuos sanos que en enfermos. Esta hipótesis cobra sentido al consultar la literatura (Fu et al., 2013). Según estos investigadores, ésta vía culmina con la activación de otras vías de señalización como son PI3K/Akt y Ras/ERK. Como comprobamos, esto culmina en la activación cascadas que controlan la mayoría de los *hallmarks* de cáncer: supervivencia celular, ciclo, celular, metabolismo, movilidad e inestabilidad genómica (Fruman & Rommel, 2014). De esta manera parece que la sobreactivación de esta función en las muestras de cáncer de mama tiene una asociación con la proliferación y otras características del cáncer.

La secreción de la renina en sangre (GO:0002001) parece tener un mayor nivel de expresión en las muestras tumorales. El sistema renina angiotensina parece estar afectado en el cáncer de mama, sobre todo los receptores para la angiotensina 1 y 2 (Vinson, Barker, &

Puddefoot, 2012). De esta manera, parece que la angiotensina II es capaz de actuar vía receptor de angiotensina tipo I, provocando una proliferación de las células tumorales y un aumento de la angiogénesis (formación de vasos sanguíneos, que aumenta la invasividad del tumor).

Por último, terminamos discutiendo las funciones que obtenemos en este metaanálisis que están relacionadas con el sistema inmunitario.

Primero pasamos a comentar la única con LOR positivo, hablamos del término GO:0072602 referente a la secreción de interleucina 4. Parece que la interleucina 4 puede promover la apoptosis en células cancerosas de mama (Nagai & Toi, 2000). En nuestro caso en las células cancerosas parece que los genes causantes de su secreción se encuentran silenciados. Esto podría ser un modo en el que las células cancerosas de mama podrían protegerse del efecto devastador que tiene la interleucina 4 sobre ellas.

El resto de funciones que tienen que ver con el sistema inmunitario tienen un LOR negativo, lo que quiere decir que están más presentes en las muestras correspondientes a cáncer de mama. A priori esto podría ser esperable, dado que es una respuesta defensiva de nuestro organismo contra algo extraño (las células cancerosas causantes de la enfermedad). Entre estas funciones destacamos algunas como las pertenecientes a los términos GO:0002309 y GO:2000318 que hacen referencia a la proliferación de linfocitos T como respuesta inmune y la regulación positiva de las células Th17 (*T-helper 17*) en la respuesta inmune. Aunque podríamos pensar que una respuesta defensiva de nuestro organismo siempre es algo favorable, hay veces en que puede actuar impulsando la proliferación, invasión y diseminación del cáncer. Hay estudios en los que se sugiere que la respuesta inflamatoria encontrada en algunos tipos de cáncer puede ser debida a una inflamación crónica, resultando en un ambiente rico en células inmunitarias que promueven la angiogénesis y la proliferación celular (Disis, 2010). Las propiedades potenciadores del crecimiento de este tipo de respuesta inflamatoria se han comparado con la inflamación consistente con la cicatrización de las heridas (Coussens & Werb, 2002).

Como podemos observar para el metaanálisis con los procesos biológicos, muchas funciones que encontramos en la aproximación del valor absoluto se encuentran también en la del promedio. Esto podemos comprobarlo en la tabla 32, donde mostramos los procesos biológicos pertenecientes al metaanálisis con la aproximación del promedio.

ID	NOMBRE	SUMMARY_LOR
GO:0032055	negative regulation of translation in response to stress	1.018
GO:0042661	regulation of mesodermal cell fate specification	0.862
GO:0009996	negative regulation of cell fate specification	0.859
GO:0033129	positive regulation of histone phosphorylation	0.793
GO:0042659	regulation of cell fate specification	0.77
GO:2000049	positive regulation of cell-cell adhesion mediated by cadherin	0.755
GO:2000323	negative regulation of glucocorticoid receptor signaling pathway	0.751
GO:0010519	negative regulation of phospholipase activity	0.739
GO:0009912	auditory receptor cell fate commitment	0.69
GO:0001768	establishment of T cell polarity	0.669
GO:0035912	dorsal aorta morphogenesis	0.66
GO:0036072	direct ossification	0.659
GO:0072592	oxygen metabolic process	0.656
GO:1902231	positive regulation of intrinsic apoptotic signaling pathway in response to DNA damage	0.641
GO:1902237	positive regulation of endoplasmic reticulum stress-induced intrinsic apoptotic signaling pathway	0.628
GO:0038161	prolactin signaling pathway	0.581
GO:0032633	interleukin-4 production	-0.529
GO:0002836	positive regulation of response to tumor cell	-0.534
GO:0002839	positive regulation of immune response to tumor cell	-0.534
GO:0015732	prostaglandin transport	-0.534
GO:0002418	immune response to tumor cell	-0.574
GO:0002347	response to tumor cell	-0.589
GO:2000319	regulation of T-helper 17 cell differentiation	-0.604
GO:0002001	renin secretion into blood stream	-1.022
GO:2001185	regulation of CD8-positive, alpha-beta T cell activation	-1.051
GO:2000564	regulation of CD8-positive, alpha-beta T cell proliferation	-1.144
GO:0002309	T cell proliferation involved in immune response	-1.181

Tabla 32. Términos GO referentes a procesos biológicos para la aproximación con el promedio, se muestran únicamente el ID, nombre y el logaritmo del odds ratio. En color verde aquellos términos comunes o similares encontrados en el metaanálisis con la aproximación del promedio con respecto a los seleccionados en el valor absoluto y en color naranja se muestran aquellos distintos.

Como podemos comprobar en la tabla 32, hay un alto nivel de coincidencia entre las funciones significativas de ambas aproximaciones. Sin embargo, en el abordaje que utiliza el promedio, aparecen un gran número de procesos no relacionados con las células del cáncer de mama. De esta manera podemos observar procesos biológicos como la morfogénesis de la aorta dorsal, el proceso metabólico del oxígeno o la osificación directa.

Es cierto que entre las funciones que no son comunes encontramos algunas como la regulación negativa de la vía de señalización de receptores de glucocorticoides, cuyo término se corresponde con GO:2000323. Esto quiere decir que la regulación negativa de esta vía se encuentra silenciada en las muestras tumorales, estando por tanto esta vía activada. Hay estudios que indican que esta vía de señalización produce una señal de supervivencia únicamente en células tumorales pertenecientes a cáncer de mama mientras

que en otros tipos de cáncer la señal es de apoptosis celular (Moutsatsou & Papavassiliou, 2008).

Otra función que encontramos como no común y es interesante comentar es la ruta de señalización de la prolactina cuyo término es GO:0038161. Podemos observar como tiene un nivel de expresión menor en las muestras tumorales ya que su nivel de metilación es mayor. Según la información que encontramos en la literatura (Sethi, Chanukya, & Nagesh, 2012), altos niveles de prolactina se asocian con un aumento de la carcinogénesis en cáncer de mama y en cáncer colorrectal. De manera que el resultado del metaanálisis para este término parece que muestra el sentido contrario y no coincidente con la información recogida en la literatura existente sobre este tema. Es por este motivo, en el que obtenemos resultados contrarios a los esperados y por los comentados anteriormente por los que decidimos utilizar la aproximación del valor absoluto como una fuente más fiable de información.

2. Metaanálisis para las funciones moleculares y componentes celulares

A continuación se muestran unas tablas resúmenes de los metaanálisis para las funciones moleculares y los componentes para la aproximación con el valor absoluto. No se muestran con el promedio por, como hemos dicho anteriormente y se puede comprobar en las tablas 26 y 28, los resultados de esta aproximación no parecen ser fieles representantes de lo que deberíamos encontrar en un principio.

En las tablas 33 y 34 podemos encontrar de manera resumida los resultados destacados de los metaanálisis de las funciones moleculares y componentes celulares para la aproximación del valor absoluto.

ID	NOMBRE	SUMMARY_LOR
GO:0001727	lipid kinase activity	0.888
GO:0019966	interleukin-1 binding	0.595
GO:0030368	interleukin-17 receptor activity	0.564
GO:0042609	CD4 receptor binding	0.563
GO:0004716	receptor signaling protein tyrosine kinase activity	-0.521
GO:0004955	prostaglandin receptor activity	-0.649
GO:0035529	NADH pyrophosphatase activity	-0.678
GO:0004957	prostaglandin E receptor activity	-0.902

Tabla 33. Términos GO referentes a funciones moleculares para la aproximación con el valor absoluto. Se muestran únicamente el ID, el nombre y el logaritmo del odds ratio.

ID	NOMBRE	SUMMARY_LOR
GO:0005915	zonula adherens	0.653
GO:0016342	catenin complex	0.517
GO:0042101	T cell receptor complex	-0.589
GO:0061702	inflammasome complex	-0.616
GO:0042105	alpha-beta T cell receptor complex	-0.62
GO:0097169	AIM2 inflammasome complex	-0.678

Tabla 34. Términos GO referentes a componentes celulares para la aproximación con el valor absoluto. Se muestran únicamente el ID, el nombre y el logaritmo del odds ratio.

Como podemos observar en las tablas 33 y 34, los términos GO hacen referencia a funciones moleculares y componentes celulares que guardan relación con los procesos biológicos descritos anteriormente. Por ejemplo, fijándonos en la tabla 33 podemos observar un nivel alto de metilación de funciones moleculares que tienen que ver con distintos tipos de interleucinas, tal y como obteníamos anteriormente con los procesos biológicos.

Además, obtenemos otros términos nuevos que no habíamos visto anteriormente. Llama la atención el bajo nivel de expresión de la actividad de la quinasa lipídica en las células cancerosas. Aunque al ser un término tan general, no podemos saber a qué tipo de quinasa se refiere, podría ser PIK3, en cuyo caso no iría en consonancia con otros estudios o cualquier otro tipo.

Es interesante el silenciamiento de los genes que intervienen en la unión del receptor CD4 de los linfocitos T, de esta manera las células cancerosas escaparían al efecto antitumoral que media este tipo de receptores presentes en la superficie de linfocitos T, monocitos y macrófagos (Kim & Cantor, 2014).

Por último en lo referente a los resultados del metaanálisis de las funciones moleculares recogidas en la tabla 33 resaltamos dos términos referentes a la actividad que llevan a cabo los receptores de prostaglandinas. Estos términos son GO:0004955 y GO:0004957, los cuales presentan un mayor nivel de metilación en las muestras controles, lo que quiere decir que los genes que se encargan de llevar a cabo la actividad de estos receptores tienen un nivel de expresión menor en los individuos sanos analizados. Dicho de otra manera, los genes que codifican para la actividad de estos receptores se encuentran muy expresados en las muestras cancerosas. Las prostaglandinas son un conjunto de sustancias de carácter lipídico derivadas del ácido araquidónico. Median en funciones homeostáticas como los mecanismos de patogenicidad, incluyendo la respuesta inflamatoria (promoviéndola). Normalmente las prostaglandinas facilitan la progresión tumoral en células cancerosas debido a una sobreexpresión de la enzima COX-2 que cataliza prostaglandinas como la PGE2 (Ricciotti & FitzGerald, 2011). Este resultado tiene sentido biológico dentro del marco en el que estamos trabajando (mayor expresión en células tumorales) lo que reafirma la adecuación del método de análisis.

En la tabla 34 referente al metaanálisis de los componentes celulares, nos encontramos con términos relacionados con lo visto anteriormente como se indicaba con antelación. Cabe destacar el inflamasoma, una estructura del sistema inmune que puede tanto acabar con la actividad de las células tumorales como promoverla según el tipo de epitelio en el que se encuentre. En el cáncer de mama en concreto, parece tener una actividad que favorece la promoción del tumor (Kolb, Liu, Janowski, Sutterwala, & Zhang, 2014).

3. Metaanálisis para rutas KEGG

Finalmente llega el turno del último metaanálisis funcional, el referente a las rutas KEGG para la aproximación del valor absoluto. Con el nivel de filtrado que hemos aplicado a los términos GO, obtenemos una ruta significativa y con un LOR superior a 0.5 e inferior a -0.5, la cual se muestra en la tabla 35.

ID	NOMBRE	SUMMARY_LOR
HSA00780	Biotin metabolism	-0.75

Tabla 35. Única ruta KEGG obtenida con el filtro del LOR aplicado a la aproximación del valor absoluto. Se muestran únicamente su ID, nombre y el logaritmo del odds ratio.

Esta ruta, que supera el filtro impuesto en el LOR es la correspondiente con el metabolismo de la biotina, el cual se muestra con mayor detalle en la figura 27.

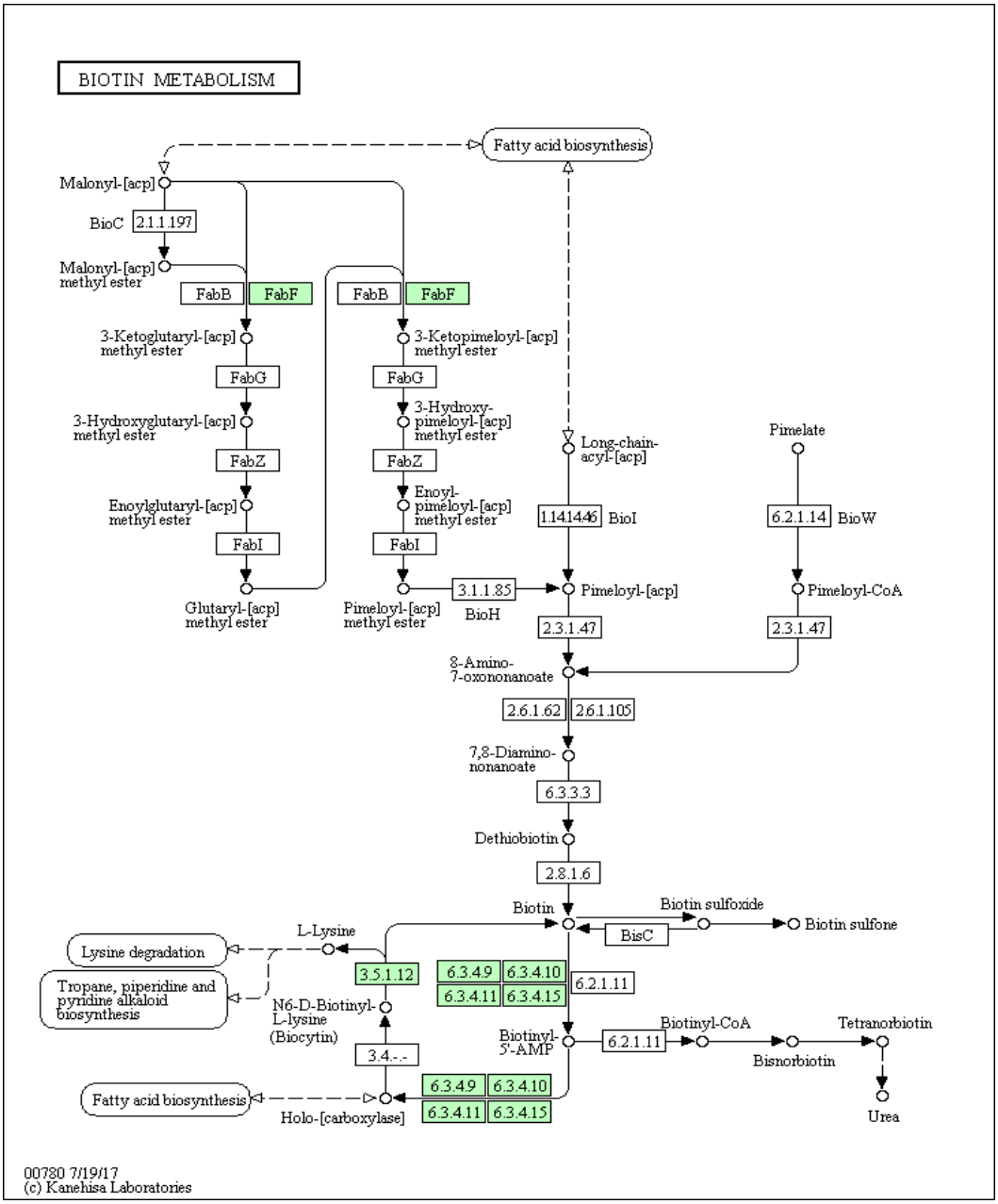

Figura 27. Metabolismo de la biotina obtenido desde KEGG Pathways.

Al comprobar la dirección de la metilación vemos que esta ruta se encuentra más metilada en las muestras de tipo control, lo que implica un menor nivel de expresión de los genes que la componen. Esto quiere decir que los genes de la ruta o, la propia ruta tiene un mayor nivel de expresión en las muestras correspondientes a cáncer de mama. La biotina, en consonancia con la literatura, es una vitamina comúnmente llamada vitamina B7 la cual podría ser incorporada de manera más agresiva en las células de cáncer de mama (en comparación con células mamarias normales) para mantener un alto estado proliferativo de las mismas (Vadlapudi, Vadlapatla, Pal, & Mitra, 2013).

Aunque no dispongan de un LOR que supere el punto de corte establecido anteriormente, en el metaanálisis funcional de las rutas KEGG con el valor absoluto encontramos rutas significativas directamente relacionadas con el cáncer. Entre ellas, cabe destacar las pertenecientes a los términos hsa05215 y hsa05217, los cuales representan las vías de señalización del cáncer de próstata y de las células de carcinoma basal. Ambas funciones tienen un nivel de metilación mayor en las muestras cancerosas respecto a las controles LOR de 0.08 y 0.189), lo que indica que varios genes pertenecientes a esta ruta tienen un nivel de expresión bajo en las muestras tumorales. Si prestamos atención a las figuras 28 y 29, correspondientes cada una los términos hsa05215 y hsa05217 respectivamente, podemos suponer que los genes con un perfil de expresión bajo en las muestras tumorales serían los supresores de tumores. Nos referimos a genes como PTEN, p53 y PTCH1 los cuales aparecen en estas vías de señalización como inhibidores del crecimiento celular. De esta forma, podría ser que algunos de los genes implicados en estas vías de señalización lo estuviese también en el cáncer de mama.

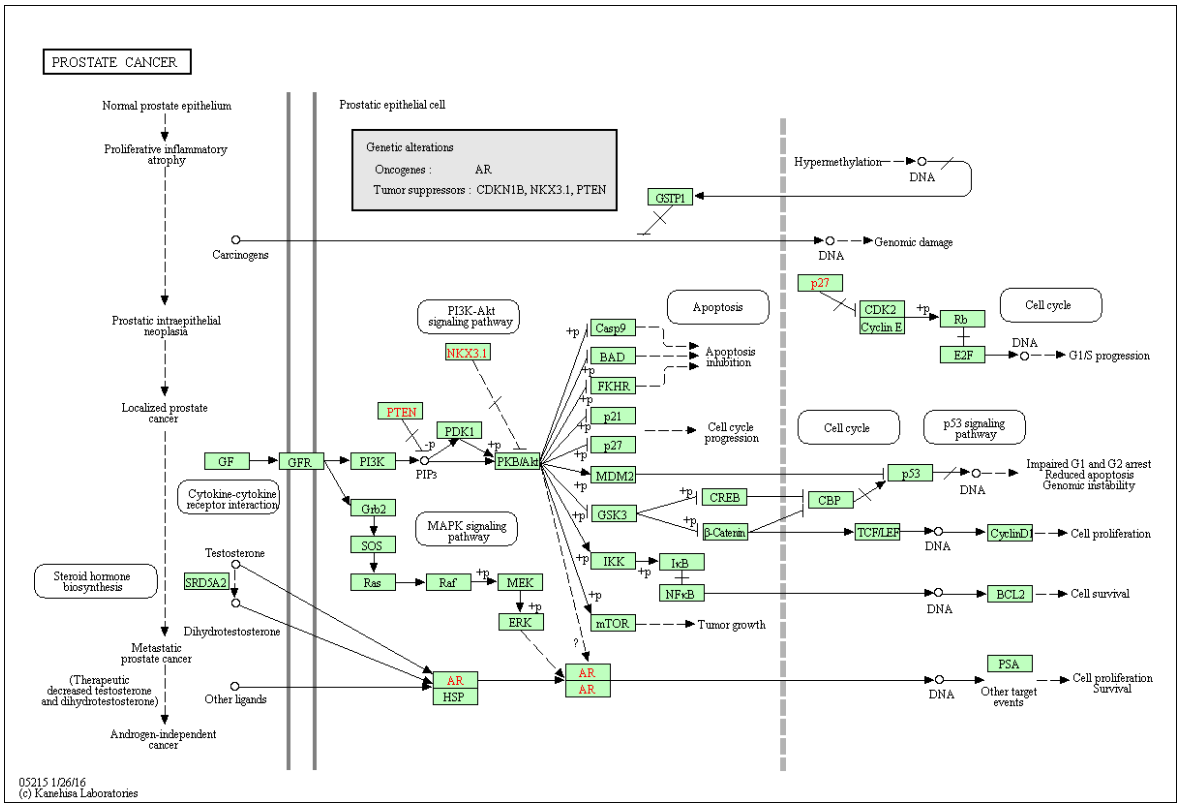

Figura 28. Vía de señalización alterada en el cáncer de próstata obtenida de KEGG Pathways.

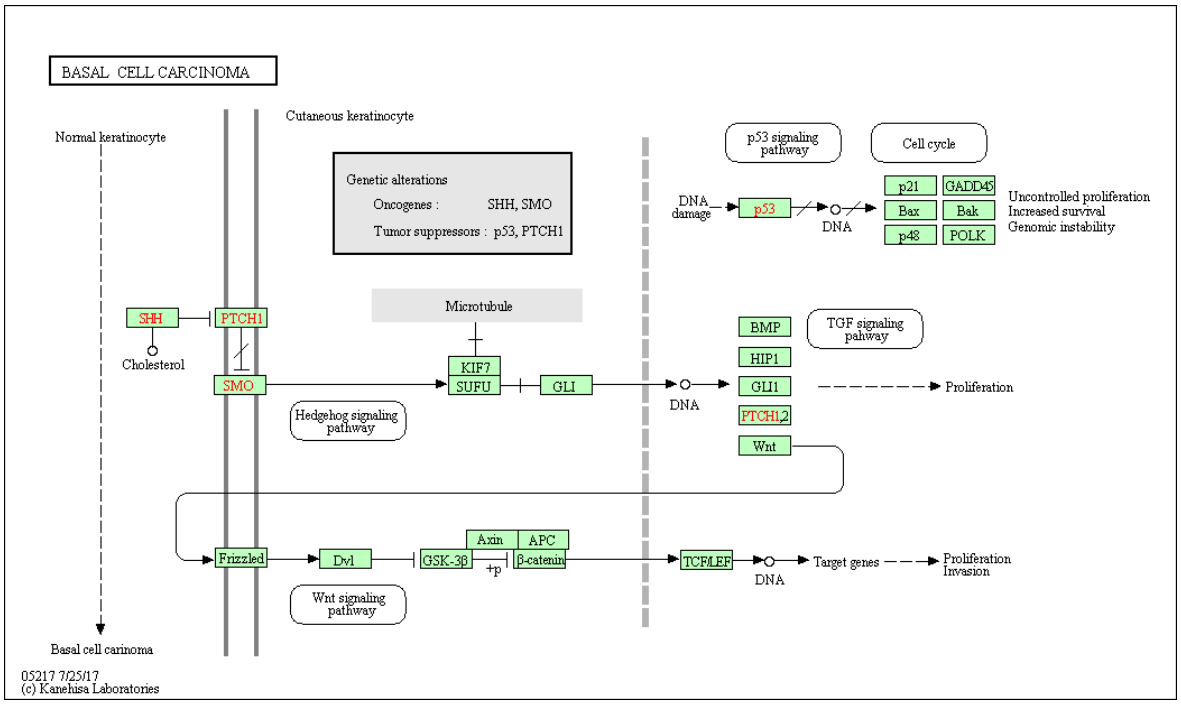

Figura 29. Vía de señalización alterada en las células de carcinoma basal obtenida de KEGG Pathways.

Los resultados para los distintos metaanálisis comentados anteriormente aportan una mejor comprensión de los mecanismos moleculares del cáncer de mama y suponen por una parte, la corroboración de algunas funcionalidades ya descritas en la literatura y por otra parte, la detección de nuevas funciones que posibilitan nuevos abordajes e hipótesis de trabajo en futuros estudios.

Conclusiones y perspectivas futuras

El cáncer es una de las causas principales de muerte en la sociedad moderna, aun así, no comprendemos totalmente todos sus mecanismos y medios de actuación. Es importante no sólo centrarse en estudios de transcriptómica, sino mirar hacia otras ómicas y aumentar la cantidad de herramientas y conocimientos que nos proporcionen un mayor entendimiento de esta enfermedad. La epigenética es una de estas ómicas que necesita un mayor estudio y esfuerzo por parte de la comunidad científica para discernir los mecanismos “ocultos” que subyacen a la regulación genética.

Gracias a los métodos de metaanálisis aplicados en datos ómicos, podemos ser capaces de detectar un mayor número de funciones que se traducirán posteriormente en genes de interés que actúen como marcadores tumorales. De esta manera, con estos marcadores tumorales podrán llevarse a cabo nuevos ensayos clínicos, donde busquemos de una forma más focalizada soluciones contra esta enfermedad.

Una de las líneas futuras de trabajo es la revisión de los resultados obtenidos con expertos oncólogos, que nos puedan ayudar a estudiar y visualizar con detalle el medio de actuación y marco de cada una de las funciones que obtenemos en el metaanálisis. Adicionalmente, sería adecuado evaluar otras soluciones complementarias que resuman la señal de metilación a nivel de gen, para comprobar cómo afecta a los resultados obtenidos. Por último, sería conveniente una mejora de las estrategias computacionales para optimizar el tiempo de ejecución de los análisis bioinformáticos.

Bibliografía

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., … Sherlock, G. (2011). Gene Ontology : tool for the unification of biology, *25*(1), 25–29. https://doi.org/10.1038/75556.Gene

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*. https://doi.org/10.2307/2346101

Bray, F., Jemal, A., Grey, N., Ferlay, J., & Forman, D. (2012). Global cancer transitions according to the Human Development Index (2008-2030): A population-based study. *The Lancet Oncology*, *13*(8), 790–801. https://doi.org/10.1016/S1470-2045(12)70211-5

Cmi, D. (2007). Subtipos clínicos y genéticos de cáncer de mama : individualización del tratamiento, *24*, 569–570.

Coussens, L. M., & Werb, Z. (2002). Inflammation and cancer. *Nature*, *420*(6917), 860–7. https://doi.org/10.1038/nature01322

DerSimonian, R., & Laird, N. (2014). *Meta-analysis in clinical trials*. *Controlled Clinical Trials* (Vol. 7). https://doi.org/10.1016/0197-2456(86)90046-2

Development Core Team. (2009). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org. Retrieved from https://www.r-project.org/

Disis, M. L. (2010). Immune regulation of cancer. *Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology*, *28*(29), 4531–8. https://doi.org/10.1200/JCO.2009.27.2146

Dopazo, J. (2009). Formulating and testing hypotheses in functional genomics. https://doi.org/10.1016/j.artmed.2008.08.003

Dotto, G. P. (2008). Notch tumor suppressor function. *Oncogene*, *27*(38), 5115–5123. https://doi.org/10.1038/onc.2008.225.Notch

Evers, B. M., Parekh, D., Townsend, C. M., Thompson, J. C., & Thompson, J. C. (1991). Somatostatin and analogues in the treatment of cancer. A review. *Annals of Surgery*, *213*(3), 190–8. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/1671812

Fruman, D. A., & Rommel, C. (2014). PI3K and cancer: lessons, challenges and opportunities. *Nature Reviews. Drug Discovery*, *13*(2), 140–56. https://doi.org/10.1038/nrd4204

Fu, H.-L., Valiathan, R. R., Arkwright, R., Sohail, A., Mihai, C., Kumarasiri, M., … Fridman, R. (2013). Discoidin domain receptors: unique receptor tyrosine kinases in collagen-mediated signaling. *The Journal of Biological Chemistry*, *288*(11), 7430–7. https://doi.org/10.1074/jbc.R112.444158

García-García, F. (2016). *Métodos de análisis de enriquecimiento funcional en estudios genómicos*. Universidad de Valencia.

Haidich, A. (2010). Meta-analysis in medical research, *14*(Suppl 1), 29–37.

Hanahan, D., & Weinberg, R. A. (2011). Hallmarks of cancer: The next generation. *Cell*,

144(5), 646–674. https://doi.org/10.1016/j.cell.2011.02.013

Hedges, L. V., Gurevitch, J., & Curtis, P. S. (1999). THE META-ANALYSIS OF RESPONSE RATIOS IN EXPERIMENTAL ECOLOGY. *Ecology*, *80*(4), 1150–1156. https://doi.org/10.1890/0012-9658(1999)080[1150:TMAORR]2.0.CO;2

Hunter, J. E., & Schmidt, F. L. (2014). *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings*. Retrieved from https://www.gwern.net/docs/statistics/2004-hunterschmidt-methodsofmetaanalysis.pdf

Jaffe, A. E., Murakami, P., Lee, H., Leek, J. T., Fallin, M. D., Feinberg, A. P., & Irizarry, R. A. (2012). Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. *International Journal of Epidemiology*, *41*(1), 200–209. https://doi.org/10.1093/ije/dyr238

Kalluri, R., & Weinberg, R. a. (2009). Review series The basics of epithelial-mesenchymal transition. *Journal of Clinical Investigation*, *119*(6), 1420–1428. https://doi.org/10.1172/JCI39104.1420

Kanehisa, M., & Goto, S. (2000). KEGG : Kyoto Encyclopedia of Genes and Genomes, *28*(1), 27–30.

Kharmate, G., Rajput, P. S., Lin, Y.-C., & Kumar, U. (2013). Inhibition of tumor promoting signals by activation of SSTR2 and opioid receptors in human breast cancer cells. *Cancer Cell International*, *13*(1), 93. https://doi.org/10.1186/1475-2867-13-93

Kim, H.-J., & Cantor, H. (2014). CD4 T-cell Subsets and Tumor Immunity: The Helpful and the Not-so-Helpful. *Cancer Immunology Research*, *2*(2), 91–98. https://doi.org/10.1158/2326-6066.CIR-13-0216

Kinlaw, W. B., Quinn, J. L., Wells, W. A., Roser-Jones, C., & Moncur, J. T. (2006). Spot 14: A marker of aggressive breast cancer and a potential therapeutic target. *Endocrinology*, *147*(9), 4048–4055. https://doi.org/10.1210/en.2006-0463

Kolb, R., Liu, G. H., Janowski, A. M., Sutterwala, F. S., & Zhang, W. (2014). Inflammasomes in cancer: A double-edged sword. *Protein and Cell*, *5*(1), 12–20. https://doi.org/10.1007/s13238-013-0001-4

Kuemmerle, N. B., Rysman, E., Lombardo, P. S., Flanagan, A. J., Lipe, B. C., Wells, W. A., … Kinlaw, W. B. (2011). Lipoprotein lipase links dietary fat to solid tumor cell proliferation. *Mol Cancer Ther*, *10*(3), 427–436. https://doi.org/1535-7163.MCT-10-0802 [pii]\r10.1158/1535-7163.MCT-10-0802

Kurdyukov, S., & Bullock, M. (2016). DNA Methylation Analysis: Choosing the Right Method. *Biology*, *5*(1), 3. https://doi.org/10.3390/biology5010003

Lee, M., & Vasioukhin, V. (2008). Cell polarity and cancer – cell and tissue polarity as a non-canonical tumor suppressor. *Journal of Cell Science*, *121*(8). Retrieved from http://jcs.biologists.org/content/121/8/1141

Lugones Botell, Miguel ; Ramírez Bermúdez, M. (2009). Aspectos históricos y culturales sobre el cáncer de mama. *Revista Cubana Medicina General Integral*, *25*(3), 160–166.

Mareel, M., & Leroy, A. (2003). Clinical, cellular, and molecular aspects of cancer invasion.

Physiological Reviews, *83*, 337–376. https://doi.org/10.1152/physrev.00024.2002

Massagué, J. (2008). Hacia Una Comprensión Del Cáncer, 203–215.

Montaner, D. (2010). Multidimensional Gene Set Analysis of Genomic Data, *5*(4). https://doi.org/10.1371/journal.pone.0010348

Montaner, D., Minguez, P., Al-shahrour, F., & Dopazo, J. (2009). Gene set internal coherence in the context of functional profiling, *13*, 1–13. https://doi.org/10.1186/1471-2164-10-197

Moutsatsou, P., & Papavassiliou, A. G. (2008). The glucocorticoid receptor signalling in breast cancer. *Journal of Cellular and Molecular Medicine*, *12*(1), 145–63. https://doi.org/10.1111/j.1582-4934.2007.00177.x

Nagai, S., & Toi, M. (2000). Interleukin-4 and breast cancer. *Breast Cancer (Tokyo, Japan)*, *7*(3), 181–6. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11029795

National Cancer Institute. (2013). Grado de un tumor - National Cancer Institute. Retrieved September 2, 2017, from https://www.cancer.gov/espanol/cancer/diagnostico-estadificacion/pronostico/hoja-informativa-grado-tumor

OMS | Cáncer. (2017). *WHO*. Retrieved from http://www.who.int/mediacentre/factsheets/fs297/es/

Park, J. B., Lee, C. S., Jang, J.-H., Ghim, J., Kim, Y.-J., You, S., … Ryu, S. H. (2012). Phospholipase signalling networks in cancer. *Nature Reviews Cancer*, *12*(11), 782–792. https://doi.org/10.1038/nrc3379

Phillips, T. (2008). The Role of Methylation in Gene Expression | Learn Science at Scitable. *Nature Education*, *1*(1), 116. Retrieved from https://www.nature.com/scitable/topicpage/the-role-of-methylation-in-gene-expression-1070

Poon, I. K. H., Lucas, C. D., Rossi, A. G., & Ravichandran, K. S. (2014). Apoptotic cell clearance: basic biology and therapeutic potential. *Nature Reviews. Immunology*, *14*(3), 166–80. https://doi.org/10.1038/nri3607

Ricciotti, E., & FitzGerald, G. A. (2011). Prostaglandins and inflammation. *Arteriosclerosis, Thrombosis, and Vascular Biology*, *31*(5), 986–1000. https://doi.org/10.1161/ATVBAHA.110.207449

Sethi, B. K., Chanukya, G. V, & Nagesh, V. S. (2012). Prolactin and cancer: Has the orphan finally found a home? *Indian Journal of Endocrinology and Metabolism*, *16*(Suppl 2), S195-8. https://doi.org/10.4103/2230-8210.104038

Stefansson, O. A., Moran, S., Gomez, A., Sayols, S., Arribas-Jorba, C., Sandoval, J., … Esteller, M. (2015). A DNA methylation-based definition of biologically distinct breast cancer subtypes. *Molecular Oncology*, *9*(3), 555–568. https://doi.org/10.1016/j.molonc.2014.10.012

Sterne, J. A., Egger, M., on, A. C. E. for the C. G., Trials, I., Lau, J., Ohrvik, J., & Furangen, A. (2001). Funnel plots for detecting bias in meta-analysis: guidelines on choice of axis. *Journal of Clinical Epidemiology*, *54*(10), 1046–55. https://doi.org/10.1016/S0895-

4356(01)00377-8

Sun, Z., Cunningham, J., Slager, S., & Kocher, J. (2015). Base resolution methylome profiling : considerations in platform selection , data preprocessing and analysis, *7*, 813–828.

Vadlapudi, A. D., Vadlapatla, R. K., Pal, D., & Mitra, A. K. (2013). Biotin uptake by T47D breast cancer cells: Functional and molecular evidence of sodium-dependent multivitamin transporter (SMVT). *International Journal of Pharmaceutics*, *441*(1–2), 535–543. https://doi.org/10.1016/j.ijpharm.2012.10.047

Viechtbauer, W. (2010). Conducting Meta-Analyses in R with the metafor Package. *Journal of Statistical Software*, *36*(3), 1–48. https://doi.org/10.1103/PhysRevB.91.121108

Vinson, G. P., Barker, S., & Puddefoot, J. R. (2012). The renin-angiotensin system in the breast and breast cancer. *Endocrine Related Cancer*, *19*(1), R1–R19. https://doi.org/10.1530/ERC-11-0335

Weinhold, B. (2006). Epigenetics: the science of change. *Environmental Health Perspectives.*, *114*(3), 160–167. https://doi.org/10.1289/ehp.114-a160

Anexos

1. Código en R utilizado para desarrollar el estudio

El código empleado para desarrollar el estudio se encuentra en https://github.com/atrassierra/tfm.

El código facilitado permite la reproducibilidad de los análisis realizados.

2. Material suplementario

Todos los resultados gráficos y numéricos están disponibles en https://drive.google.com/file/d/0B0ovsM0Kot4qR3ZUUUFqN04yejg/view?usp=sharing.