Relación entre las alteraciones génicas asociadas a los mecanismos epigenéticos y los perfiles de metilación aberrante hallados en el ADN canceroso global

Lorena Ponce Ruiz

Máster en Bioinformática y Bioestadística
Epigenómica y Cáncer

Supervisora:
Izaskun Mallona González

05 de junio de 2019

FICHA FINAL DEL TRABAJO

Título del trabajo:	Relación entre las alteraciones génicas asociadas a los mecanismos epigenéticos y los perfiles de metilación aberrante hallados en el ADN canceroso global
Nombre del autor:	Lorena Ponce Ruiz
Nombre del consultor/a:	Izaskun Mallona González
Nombre del PRA:	Ferran Prados Carrasco
Fecha de entrega:	05 de junio de 2019
Titulación:	Máster en Bioinformática y Bioestadística
Área del trabajo:	Área 3, Subárea 6: Epigenómica y cáncer
Idioma del trabajo:	Español
Palabras clave:	Cáncer, epigenética, mutaciones somáticas, variación en el número de copias génicas, aprendizaje automático
Key words:	Cancer, epigenetics, somatic mutations, copy-number variation, Machine Learning

Resumen:

La relevancia de las alteraciones genéticas y epigenéticas en la formación de tumores, su desarrollo y metástasis es bien conocida en la actualidad. A pesar de ello, en la mayoría de los casos no hay evidencias asociativas entre los perfiles de metilación aberrante y el estado genético de las células cancerosas. El objetivo del presente trabajo es esclarecer dicha relación mediante la búsqueda de patrones en el número de copias y en las mutaciones de genes candidatos que puedan ser predictivos de perfiles de metilación aberrante o viceversa. Para tal objeto, se seleccionarán sendas bases de datos con información acerca de las mutaciones somáticas, la variación en el número de copias génicas y los perfiles de metilación presentes en tejido normal y tejido canceroso de 4 cohortes distintas, las cuales se corresponden con 4 tipos de cáncer: pulmón, próstata, mama y colon; por sus siglas en TCGA: LUAD, PRAD, BRCA y COAD, respectivamente. El método de análisis está enmarcado en el proceso conocido como minería de datos o *datamining*, el cual comprende una serie de técnicas relacionadas con la estadística, la inteligencia artificial, la visualización de datos y otros muchos campos que permiten encontrar patrones en grandes bases de datos como las que se trabajarán en el presente trabajo.

Abstract:

The relevance of genetic and epigenetic alterations in the formation of tumors, their development and metastasis are well known at present. Nonetheless, in most cases there is no associative evidence between aberrant methylation profiles and the genetic status of cancer cells. The aim of this paper is to clarify this relationship by looking for patterns in the number of copies and mutations of candidate genes that may be predictive of aberrant methylation profiles or vice versa. For this purpose, databases will be selected with information about somatic mutations, variation in the number of gene copies and methylation profiles present in normal tissue and cancerous tissue from 4 different cohorts, which correspond to 4 types of cancer: lung, prostate, breast and colon; by its initials in TCGA: LUAD, PRAD, BRCA and COAD, respectively. The method of analysis is framed in the process known as datamining, which includes a series of techniques related to statistics, artificial intelligence, data visualization and many other fields that allow to find patterns in large databases as those that will be worked on in the present work.

Índice

A. PLAN DE PROYECTO

1. CONTEXTO Y JUSTIFICACIÓN DEL PROYECTO

De acuerdo con la Organización Mundial de la Salud (OMS) el cáncer es la segunda causa de muerte en el mundo, donde casi una de cada seis defunciones a nivel global es debida a esta enfermedad. Los principales problemas asociados a esta elevada tasa de mortalidad están vinculados con problemas en la detección, diagnóstico y tratamiento del cáncer. Debido a todo ello, las investigaciones para comprender completamente sus mecanismos y descubrir nuevas formas de prevenir y tratar esta enfermedad son fundamentales.

Desde el descubrimiento en los 80s del primer oncogen (Barbacid, 1987), cada vez se han encontrado más evidencias vinculando las alteraciones genéticas y epigenéticas con una expresión génica anormal en los procesos de inicio, desarrollo y expansión tumoral. Un ejemplo de la relación entre los procesos genéticos y epigenéticos es evidente en la línea celular de cáncer de colon HCT116, en la cual un alelo de MLH1 y CDKN2A está mutado genéticamente, mientras que el otro está silenciado por un proceso de metilación del ADN (Baylin y Ohm, 2006). Esta falta de expresión en los genes MLH1 y CDKN2A provoca desajustes en el proceso de reparación del ADN y en la regulación del ciclo celular, que desemboca en la aparición de células cancerosas.

El cáncer engloba más de 200 enfermedades distintas con diversos factores de riesgo y epidemiologías, por lo que mejorar nuestro conocimiento sobre la importancia de los cambios epigenéticos y genéticos que lo favorecen, puede abrir las puertas al desarrollo de estrategias de pronóstico y terapéuticas más efectivas. Algunos tratamientos como la 5-azacitadina (Tran et al. 2011) y la decitabina (Plimack et al. 2007) ya han demostrado su eficacia; mientras que otras terapias epigenéticas presentan un gran potencial (Arrowsmith et al. 2012; Plass et al. 2013). Así mismo la aplicación de la terapia epigenética mejora la respuesta a otros tratamientos contra el cáncer, como la quimioterapia (Juergens et al., 2011). A pesar de ello, en la mayoría de los casos faltan evidencias concluyentes que asocien los procesos de alteraciones genéticas con los perfiles epigenéticos aberrantes específicos de cada tipo de tumor (Florean et al. 2011), denotando la importancia de su estudio.

2. OBJETIVOS DEL PROYECTO

1) El perfil de metilación en el ADN canceroso se caracteriza por presentar patrones de hipometilación en las regiones intergénicas, relacionadas con la activación de oncogenes (Sandoval y Esteller, 2012), y un marcado patrón de hipermetilación en las regiones promotoras que favorece el silenciamiento de los genes supresores de tumores (Esteller, 2007). Una de las principales preguntas es si estos patrones de metilación son específicos para un determinado tipo de cáncer o comparables entre distintos tipos de cáncer, por ello el primer objetivo del presente trabajo es **hallar los diferentes perfiles de metilación aberrante y las regiones que presentan dicha "huella" de metilación en las distintas cohortes de estudio y compararlos entre sí**. El descubrimiento de unos perfiles específicos claros en cada una de las cohortes podría favorecer el pronóstico temprano de estos tipos de tumores.

2) Hay un gran número de alteraciones genéticas que se repiten a lo largo de todos los tipos de cáncer (Baylin y Ohm, 2006), y que están implicadas en los genes reguladores de los mecanismos epigenéticos, como: DMNT3, IDH1/2 o H3.3 (Plass et al. 2013; Shen y Laird, 2013). El segundo objetivo del trabajo es **estudiar la relación entre las alteraciones genéticas en los mecanismos de regulación epigenética, considerando tanto las mutaciones somáticas como la variación en el número de copias (CNV), y los perfiles de metilación aberrante presentes en el ADN global de muestras de tejido canceroso**. Aunque hay un gran número de evidencias vinculando los cambios tanto genéticos como epigenéticos con la oncogénesis, su relación aún no está clara. Así mismo, los principales estudios sobre el papel de las alteraciones genéticas en el desarrollo del cáncer se centran sobre todo en el efecto de las mutaciones, pasando desapercibidas otras variaciones como las CNVs. La medida en la que las CNVs contribuyen a la enfermedad humana todavía no se conoce, a pesar de ello, algunos estudios han demostrado que ciertos tipos de cáncer presentan un patrón específico y aberrante en el número de copias que posee (Ni et al., 2013).

3) El tercer objetivo, aunque de menor importancia con respecto a los anteriores, es **estudiar la relación entre la expresión génica y el estado de metilación** de las cohortes de estudio.

4) El último objetivo, el cual es dependiente de la disponibilidad de tiempo para su realización es: el **desarrollo de una herramienta web empleando Shiny y RStudio** que permita estudiar el pronóstico y diagnóstico del paciente a partir de los datos de respuesta a tratamiento y medidas de supervivencia de TCGA (*total* y *disease-free*) e implementar el pipeline desarrollado en el proceso de minería de datos.

2.1. ENFOQUE Y MÉTODOS SEGUIDOS

Para llevar a cabo el presente trabajo se han empleado bases de datos con información acerca de las mutaciones somáticas, el número de copias, los perfiles de metilación y la expresión génica presentes en tejido normal y tejido canceroso de 4 cohortes distintas, las cuales se corresponden con 4 tipos de cáncer: pulmón (LUAD), próstata (PRAD), mama (BRCA) y colon (COAD), entre paréntesis se especifica el nombre de proyecto **Atlas del Genoma del Cáncer** (TCGA) seleccionado, disponibles en el Consorcio Internacional de los Genomas del Cáncer (ICGC).

En el análisis de los datos se ha usado la minería de datos o *datamining*, la cual comprende una serie de técnicas relacionadas con la estadística, la inteligencia artificial, la visualización de datos y otros muchos campos que permiten encontrar patrones en grandes bases de datos como las que se han trabajado en el presente trabajo. Más concretamente, para hallar el perfil de metilación significativo en cada una de las cohortes que pueda ser explicado por un patrón alterado en los genes candidatos, se ha llevado a cabo una primera clasificación no supervisada de las regiones de metilación de las muestras tumorales que presentaban un nivel de significación diferencial con respecto a las muestras de tejido sano, empleando para ello el algoritmo *K-means*. Seguidamente, y tras el aislamiento de las mutaciones y CNVs presentes en los genes implicados en los mecanismos de regulación epigenética, se llevó a cabo un algoritmo supervisado *Random Forest* y un *Support Vector Machines*, para comprobar si los patrones de metilación aberrante encontrados en las muestras tumorales de las distintas cohortes pueden ser explicados por las alteraciones genéticas presentes en las mutaciones y CNVs seleccionadas.

2.2. PLANIFICACIÓN DEL PROYECTO

Tareas

A continuación, se exponen las tareas realizadas a lo largo del proyecto:

1) **Selección y descarga de las bases de datos**, entre las que se incluyen los datos de las mutaciones somáticas, los CNVs, los datos de metilación y de expresión génica de las cuatro cohortes de estudio procedentes del consorcio TCGA.
2) **Análisis de las propiedades de los datos y preprocesamiento.** Este apartado engloba la previsualización de los datos, detección de datos anómalos y preparación de los datos para los análisis posteriores.
3) **Análisis de los datos de metilación** obtenidos a partir del consorcio internacional TCGA y procedentes de la tecnología *Illumina Infinium HumanMethylation450* (450K) *BeadChip*. Este se divide a su vez en:
    a) Búsqueda de los perfiles de metilación diferencial entre las muestras de tejidos normal y canceroso de las diferentes cohortes.
    b) Comparación de los perfiles de metilación hallados en los diferentes tipos de cáncer de estudio para comprobar la especificidad de los mismos.

4) **Estudio de la relación del estado de expresión génica y el estado de metilación de las cohortes**, que permita esclarecer la relación entre el estado de metilación (hipometilación o hipermetilación) y la activación o silenciamiento génico.
5) **Estudio de la relación entre el marco mutacional y el número de copias con respecto a los perfiles de metilación hallados.** Para ello se extraen aquellas mutaciones y número de copias presentes en los genes implicados en los mecanismos de alteración epigenética, seguidos de un análisis de *Machine Learning* que aclare esta relación.
6) **Testeo** de los resultados obtenidos del apartado 5.
7) **Desarrollo de la herramienta web empleando Shiny y Rstudio.** Finalmente, y debido a las limitaciones explicadas en el apartado siguiente, *Análisis de Riesgos*, este objetivo no se pudo realizar *per se*. A pesar de ello, y dado que el desarrollo completo del trabajo se llevó a cabo con R, sí se ha podido realizar una interfaz estática que podría cumplir (al menos parcialmente) este objetivo.

Temporización

En el siguiente apartado se muestra la planificación llevada finalmente a cabo, inclusive los entregables que faltan por entregar (Figura 1):

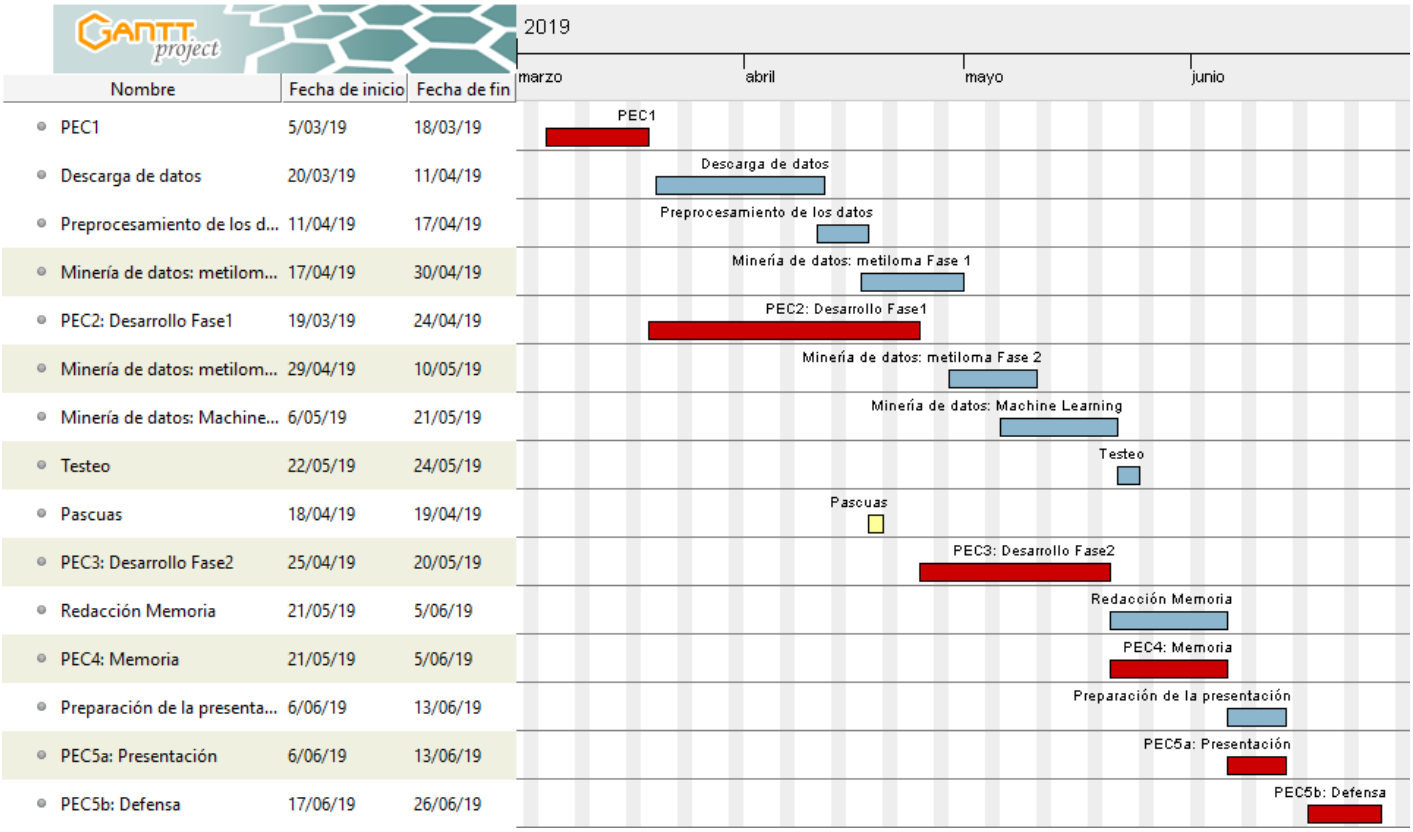


Figura 1: Proyecto Gantt con la temporización del proyecto, coincidiendo con las tareas descritas en el apartado anterior *Tareas*.

Análisis de riesgos

Hay tres grandes puntos que han limitado cumplir de manera exitosa con el total de los objetivos planeados:

1) **La obtención de los datos necesarios para el desarrollo del trabajo.**

Las bases de datos seleccionadas para el trabajo proceden de un repositorio público, el consocio internacional TCGA, que se nutre de los resultados obtenidos por una red de investigaciones conformada por instituciones de todo el mundo. Al realizar un trabajo concreto a partir de un repositorio público uno de los principales problemas, de cara a la selección de datos, es la disponibilidad de los datos necesarios para desarrollar el trabajo. Este trabajo en concreto es dependiente de la disponibilidad de muestras con datos apareados procedentes de tejido normal y canceroso para poder establecer los perfiles de metilación diferenciales y significativos de las células cancerosas. Por otra parte, para eliminar factores de error se ha decidido emplear las muestras procedentes de los mismos pacientes en todas las bases de datos de descarga (metilación, expresión génica, mutaciones y CNV), lo cual ha supuesto un sesgo para el desarrollo de algunas tareas como el estudio de la relación entre las alteraciones genéticas y los perfiles de metilación. Además, durante la descarga de los datos hubo varios problemas referentes a las librerías, como ciertas incompatibilidades con el sistema operativo, que dificultaron a su vez el proceso de descarga.

2) **La falta de capacidad del ordenador portátil empleado para desarrollar el trabajo.**

El ordenador tiene unas características limitadas, que entorpece el trabajo con bases de datos de gran tamaño como es el caso del presente trabajo; de manera que en algunos casos la memoria se satura impidiendo el procesamiento y en otros casos se precisa de mucho tiempo de computación. Para evitar este problema se ha reducido la base de datos, filtrando por ejemplo en base a la significación de los datos.

A continuación, en la Tabla 1 se especifican las características del ordenador empleado para la codificación del trabajo.

Tabla 1: Características del ordenador empleado durante el desarrollo del proyecto.

Características del procesador:	i7-3632QM
Características de la RAM:	8GB de DDR3 a 1600MHz
Versión del Sistema Operativo:	Windows 10

3) **La falta de tiempo disponible por causas laborales**

Finalmente, mis condiciones laborales particulares dificultan el dedicarle el tiempo deseado al desarrollo del proyecto. Este es un proyecto ambicioso y para su completa realización se precisaría de más tiempo disponible, especialmente para poder hacer frente a los problemas que van surgiendo a lo largo del desarrollo del proyecto y van redirigiendo la metodología a emplear.

3. RESUMEN DE LOS RESULTADOS A ENTREGAR

A continuación, se enumeran los entregables asociados al desarrollo del proyecto:

- ✓ La planificación del proyecto
- ✓ La memoria del proyecto
- ✓ El código en R disponible en el repositorio GitHub
- ✓ Una presentación virtual con su correspondiente vídeo
- ✓ Autoevaluación del proyecto

Además de todas las PECs entregadas con anterioridad a la tutora del proyecto.

4. BREVE DESCRIPCIÓN DE LOS PRÓXIMOS CAPÍTULOS

En los próximos capítulos se incluye la memoria del proyecto con los apartados típicos de un artículo científico: *introducción*, donde se aporta una visión global del efecto de las alteraciones genéticas y epigenéticas en la formación de una células cancerosas, así como sus técnicas de estudio; *métodos*, donde se describen todos los procedimientos empleados para desarrollar este proyecto; *resultados y discusión*, apartado en el cual se aportan los resultados obtenidos con una interpretación de los mismos; y, finalmente, las *conclusiones* extrapolables de los resultados del proyecto.

B. MEMORIA

1. INTRODUCCIÓN

1.1. BREVE INTRODUCCIÓN A LA EPIGENÓMICA

La epigenómica engloba todos aquellos cambios heredables que regulan la expresión génica sin afectar a la secuencia misma de ADN (Jones y Baylin, 2007). Estos cambios son dependientes de otros factores externos como la edad (Toyota et al., 1999) o la exposición a ciertos factores ambientales. Así mismo, las modificaciones epigenéticas desempeñan un papel fundamental en el desarrollo de un gran número de enfermedades humanas tales como el cáncer, enfermedades inflamatorias, enfermedades metabólicas y enfermedades neurodegenerativas.

El mecanismo de modificación epigenética más ampliamente estudiado es la metilación del ADN (Jones, 2012). La metilación es un proceso reversible en el cual las enzimas ADN-metiltransferasas (DNMTs) catalizan la conversión de una citosina en una 5-metilcitosina mediante la adición de un grupo metilo (CH_3·). La mayor parte de las 5-metilcitosinas presentes en el genoma de los mamíferos se localiza en los dinucleótidos 5'-CpG-3' (Riggs y Jones, 1983). El genoma humano haploide contiene ~28 millones de regiones CpG tanto en estado metilado como no metilado (Stevens et al., 2013). Dentro del marco epigenético hay tres tipos más de modificaciones del ADN menos conocidas: 5-hidroximetilcitosina, 5-formilcitosina y 5-carboxilcitosina (Wu y Zhang, 2011).

Otro de los mecanismos epigenéticos mejor conocidos es la modificación química de las histonas. Al menos 11 tipos de modificaciones post-traduccionales (PTMs) se han documentado sobre 60 residuos de aminoácidos diferentes en las histonas, incluyendo la acetilación, la fosforilación, la metilación, la ubiquitinación, la sumoylation, la propionilación, la butirilización, la citrulinación, la formilación, la isomerización de la prolina y la ADP ribosilación (Martin y Zhang, 2007; Ruthenburg et al., 2007).

Las modificaciones epigenéticas en el ADN y en las histonas regulan y remodelan la estructura de la cromatina afectando a su estabilidad, y, por ende, al funcionamiento de los procesos nucleares como la transcripción, la reparación del ADN y la replicación (Lomvardas y Thanos, 2002; Sanders et al., 2004; Laird et al., 2010; Jones et al., 2012). Así mismo, todos los cambios epigenéticos están vinculados entre sí, un ejemplo de ello son los resultados en los estudios realizados por Nan et al. 1998 y Jones et al. 1998, en los que demostraron que la metilación de una citosina podía atraer proteínas de unión a ADN metiladas e histonas deacetilasas (HAT) a islas CpG metiladas durante la compactación de la cromatina, provocando el silenciamiento de los genes.

1.2. EPIGENOMA Y CÁNCER

Se denomina tumor, o neoplasia, a la masa de tejido originado por una división anormal y descontrolada de las células eucarióticas. Cuando estas células invaden y destruyen los tejidos circundantes, el tumor es considerado maligno, y es llamado cáncer. La transformación de una célula sana en una célula tumoral es un proceso complejo que tradicionalmente se había asociado a alteraciones puramente genéticas, favorecidas por otros factores como son las infecciones víricas. A pesar de ello, los nuevos avances en las técnicas de mapeado de metilomas completos conforman un nuevo escenario donde todos los aspectos del cáncer, desde la formación tumoral hasta la progresión y metástasis están favorecidos por alteraciones en el genoma y en el epigenoma (Hanahan y Weinberg, 2011).

Los perfiles de metilación aberrante en las células cancerosas se caracterizan por un marcado estado de hipometilación global en las regiones intergénicas con perfiles de hipermetilación aberrante en las regiones promotoras de los genes, frecuentemente asociadas a regiones ricas en dinucleótidos CpG, denominadas islas CpG (CGI) (Lengauer, C. et al., 1997; Baylin et al., 1998). Cada vez en un número más creciente de genes se identifican estas regiones hipermetiladas asociadas a un estado canceroso de la célula (Jones y Baylin, 2007). Mientras que los perfiles de hipermetilación se asocian con replicación retardada, cromatina condensada e inhibición del inicio de la transcripción (Antequera et al., 1990; Delgado et al., 1998; Jones y Laird, 1998; Baylin y Jones, 2011), los patrones de hipometilación están vinculados con la inestabilidad genómica, contribuyendo a la transformación celular (Kullis et al., 2010).

Un ejemplo de cómo el estado de metilación de un gen puede favorecer el desarrollo tumoral sería el caso de la hipermetilación del promotor de un segundo gen de reparación del ADN que codifica la enzima glutatión S transferasa (GSTPI, también conocida como GST3), este patrón se ha detectado en el 80 - 90% de los pacientes con cáncer de próstata y aparece con frecuencia en el cáncer de mama y otros tumores (Esteller, 1998; Cains et al., 2001). Este cambio epigenético tiene el potencial de predisponer al daño por radicales de oxígeno que conduciría a mutaciones en las adeninas durante la progresión del tumor.

Estudios recientes han demostrado que el proceso de formación y progresión de tumores se debe a la activación progresiva de los genes dominantes del crecimiento (oncogenes) y la inactivación de los genes supresores de tumores (Herman et al., 1994; Weinstein, 2002; Grady y Carethers, 2008), lo que confiere una ventaja en el crecimiento a las células cancerosas. En las primeras etapas del tumor, estas alteraciones son inducidas por trastornos epigenéticos y genéticos (Feinberg et al. 2006) y diseminadas por expansión clonal (Beerenwinkel et al. 2007).

Ciertos tipos de tumores, e incluso subtipos de tumores, presentan un paisaje concreto de metilación aberrante, sugiriendo que la metilación de subconjuntos específicos de genes contribuye al desarrollo de tumores específicos (Herman et al., 1994; Costello et al., 2000). Un ejemplo de ello serían los patrones de metilación del gen BRCA1 hallados en el cáncer de ovario y de mama (Esteller et al., 2000).

Metiloma y mutaciones

La relación entre el estado de metilación global del ADN canceroso y las alteraciones génicas que presenta es, posiblemente, una de las preguntas más frecuentes en este campo, ¿son las alteraciones epigenéticas las causas o las consecuencias de la inactividad génica? La clave para dilucidar esta cuestión puede pasar por resolver los mecanismos por los cuales se inician las alteraciones epigenéticas en el cáncer para las cuales Costello et al., 2001 propone dos opciones. La primera de ella es la pérdida de los factores de protección de las regiones metiladas, particularmente de las CGI; y la segunda de ella es que la metilación es un proceso activo que provoca un inapropiado silenciamiento génico en ciertas ocasiones. Ambas teorías están respaldadas por diversos estudios científicos: Graff et al., 1995; Macleod y Szyf; 1995; Millar et al., 2000; Rhee et al., 2000; entre otros.

A pesar de que esta cuestión sigue hoy en día sin tener una respuesta clara, se ha demostrado que ciertos patrones de metilación están vinculados con *hotspots* mutacionales; revelando la sinergia entre ambos procesos. Un ejemplo de ello sería el caso del gen supresor de tumores p53, el cual se encuentra alterado en aproximadamente un 50% de todos los tumores humanos (Greenblatt et al., 1994; Harris, 1996). Este gen contiene 23 CGI metiladas en su región intergénica, representando un 8% de la secuencia génica total. A pesar de ello, el 33% de las mutaciones halladas en esta región se localizan en las CGI metiladas (Laird y Jaenisch, 1996).

Por otro lado, es recurrente la presencia de mutaciones en los genes reguladores de los mecanismos epigenéticos, predisponiendo la metilación del ADN o la modificación de las histonas. Ejemplo de ello, es el caso del gen DNMT3a cuya mutación aparece en el 25% de los pacientes de leucemia mieloide (AML) (Ley et al., 2010). Este gen junto al gen DNMT3b se encargan de metilar el ADN *de novo* durante la embriogénesis (Okano et al., 1999).

Dos estudios llevados a cabo en el Broad Institute del Massachutsets Institute of Technology y la Universidad de Harvard concuerdan en sus resultados al hallar que las mutaciones somáticas en tres genes: DNMT3A, TET2 y ASXL1 (todos ellos relacionados con los mecanismos de modificación epigenética) en las células hematopoyéticas están asociados con el desarrollo de cáncer hematológico, debido a la acumulación de estas mutaciones por hematopoyesis clonal (Genovese et al., 2014; Jaiswal et al., 2014). Todo ello indica que la detección temprana del perfil genético y epigenético de las células potencialmente cancerosas pueden mejorar la prognosis temprana del cáncer.

Todo parece indicar, pues, que a pesar de no conocer la causalidad de las alteraciones producidas en el genoma y epigenoma, ambas están relacionadas y actúan conjuntamente en los procesos de formación y desarrollo del cáncer.

Metiloma y CNVs

Dentro del marco de las alteraciones genéticas, la variabilidad en el número de copias génicas es uno de los factores de variabilidad menos estudiados. A pesar de ello, algunos estudios han demostrado que ciertos tipos de cáncer, como el cáncer de próstata o de pulmón, presentan un patrón específico y aberrante de CNVs (Holcomb et al., 2009; Cancer Genome Atlas N., 2012). La relación entre las CNVs y los perfiles de metilación, debido a la dificultad de su estudio, no se ha esclarecido; pero en ciertos tipos de cánceres como el carcinoma hepatocelular (HCC) se ha demostrado que la correlación entre el número de copias y las modificaciones epigenéticas permiten la identificación de varios subtipos de HCC de manera exitosa. Además, el mismo estudio obtuvo que los subtipos de HCC con mayor frecuencia de CNV aberrante, presentaban a su vez un perfil de metilación aberrante más marcado (Woo et al., 2017). Remarcando así, la importancia de estudiar otros tipos de alteraciones genéticas más allá de las mutaciones.

Perspectivas biológicas y clínicas

El rol de la epigenética en la oncogénesis está cada vez más claro, especialmente en los estadios tempranos de la formación tumoral. Este es, pues, un buen momento para reorientar los métodos de detección y tratamiento del cáncer. La terapia epigenética ya ha sido efectiva en el caso de fármacos como la azacitidina (Tran et al., 2011). y la decitabina (Plimack et al. 2007). Ambos medicamentos se emplean para bloquear los agentes de hipometilación en el tratamiento de tumores sólidos en el síndrome mielodisplásico (MDS). Así mismo, otros tratamientos relacionados con la inhibición de las DNMTs, como RG101 (Brueckner et al., 2005), está obteniendo resultados prometedores en estudios *in vitro*. Otra aproximación sería la inhibición de las histonas deacetilasas, como el ácido hidroxámico suberoilanilida (SAHA) (Bolden et al., 2006) o la Romidepsina, que acaba de ser aprobado por la U.S. Food and Drug Administration (FDA) para el tratamiento del linfoma cutáneo de células T (Plass et al., 2013).

El gran potencial de la terapia epigenética radica en que son cambios acumulables y detectables en las etapas tempranas del desarrollo del cáncer; y que, a diferencia de los cambios genéticos, las alteraciones epigenéticas son eventos reversibles. A pesar de ello, este tipo de terapia todavía presenta algunos retos como la inespecificidad del tratamiento, pero en algunos estudios se ha demostrado que los inhibidores de la metilación del ADN únicamente actúan en las células en división y parece que los medicamentos activan preferentemente aquellos genes que se han silenciado de manera anormal en las células cancerosas (Karpf et al., 1999; Liang et al., 2002).

1.3. ESTUDIO DE LAS BASES EPIGENÉTICAS DEL CÁNCER

El estudio del epigenoma es un campo relativamente nuevo debido principalmente a las limitaciones técnicas. El avance en las técnicas de secuenciación, especialmente el desarrollo de la secuenciación masiva (*Next-generation sequencing*, NGS), aplicada al mapeo de la cromatina y la metilación del ADN ha permitido una revolución en la forma de entender la formación y el desarrollo del cáncer.

La metilación del ADN actualmente se estudia empleando el tratamiento de las muestras con bisulfito (también conocida como Methyl-Seq). En este proceso los residuos de citosina no metilados se convierten en uracilos por un tratamiento con bisulfito sódico previo a la secuenciación, sin modificar las citosinas metiladas (Wang et al., 1980). Hay un gran número de técnicas y protocolos que combinan el tratamiento con bisulfito junto con la secuenciación, algunos de ellos son la secuenciación de bisulfito de genoma completo (WGBS), la secuenciación de bisulfito dirigida, la secuenciación de bisulfito de representación reducida (RRBS) y, más recientemente, también la secuenciación de bisulfito de una sola célula (scBS) (Masser et al., 2018). El desarrollo de estas técnicas ha permitido mejorar la especificidad y cobertura génica de los análisis, y con ello el número y la calidad de los resultados. La manera de almacenar los datos obtenidos a partir de NGS, así como las estrategias para analizarlos e interpretarlos, supone un nuevo reto al cual la minería de datos (*datamining*) y los grandes repositorios, tales como el Atlas del Genoma del Cáncer (TCGA), la Enciclopedia de los Elementos del ADN (ENCODE) o *Therapeutically Applicable Research to Generate Effective Treatments* (TARGET), pretenden hacer frente.

1.4. APRENDIZAJE AUTOMÁTICO

La minería de datos abarca todas aquellas técnicas que permiten explorar *in silico* las grandes bases de datos (*Big Data*) para encontrar patrones o tendencias que expliquen el comportamiento de los datos. Entre las técnicas que comprende la minería de datos se incluyen tanto estudios estadísticos como el conocido aprendizaje automático, o más conocido por su nombre en inglés *Machine Learning* (ML). El ML es una de las técnicas más populares y mejor establecidas para la clasificación de los datos procedentes de *microarrays*, que está desbancando a los tests estadísticos clásicos por su capacidad para identificar patrones en grandes bases de datos. Los algoritmos de ML se clasifican en supervisados, como *Support Vector Machines* o *Random Forest*, y no supervisados, como *k-means*. La principal diferencia entre ambos tipos de algoritmos radica en que en el primero de ellos se conoce el grupo de pertenencia o "etiqueta" de los datos, mientras que en el segundo no.

k-means

K-means es un método de agrupamiento o *clustering* no supervisado. A diferencia de los métodos de clasificación, el *clustering* no crea un modelo a partir de los datos, sino que crea nuevos datos a partir de la información que se infiere de la relación entre los mismos datos no etiquetados. El algoritmo *k-means* asigna cada n observación a un clúster k previamente definido. Para ello todas las características predictoras se sitúan en un espacio multidimensional, y la pertenencia a un clúster u otro reside en cuán cerca se sitúe cada observación al centroide del clúster. El objetivo de este algoritmo es agrupar las observaciones

minimizando las diferencias dentro de un clúster y aumentando las diferencias entre los clústeres. Dada la sencillez de este modelo es muy empleado en casos de agrupación no supervisada.

Support Vector Machines

Las máquinas de soporte vectorial (*Support Vector Machines*, SVM) son un conjunto de algoritmos de aprendizaje supervisado empleados tanto en problemas de clasificación como de regresión. Los algoritmos de SVM representan los puntos de los datos en el espacio mediante un hiperplano de separación. El **vector soporte** se define como el vector entre los puntos, de cada clase, más cercanos al hiperplano. Los algoritmos SVM pertenecen a la familia de los clasificadores lineales. A pesar de ello, en algunos casos los datos no son linealmente separables, para compensar este hecho se introduce el parámetro C y se diseña la función *kernel*, como la sigmoide o la gaussiana (Lantz, 2015).

Estos algoritmos son ampliamente utilizados debido a su elevada precisión y facilidad de manejo, y se han empleado en muchos estudios vinculados a los datos de expresión obtenidos con la tecnología de *microarray*, por ejemplo, en la identificación de enfermedades genéticas como el cáncer (Statnikov et al., 2005).

Entre las bondades de este conjunto de algoritmos destaca que no se ve afectado en gran medida por el "ruido" y no es propenso a sobreajustes. A pesar de ello, alguna de las debilidades de estos algoritmos es que no produce un modelo, lo que limita la comprensión de la relación entre las características y las clases, puede ser lento de entrenar y encontrar el mejor modelo requiere de probar diferentes combinaciones de parámetros y *kernels*.

Random Forest

Los bosques aleatorios (*Random Forest*, RF) se construyen usando el proceso conocido como partición recursiva, en el cual los datos se dividen repetidamente en subconjuntos más pequeños, y así sucesivamente. La partición se detiene cuando los datos dentro de un subconjunto son lo suficientemente homogéneos, o se ha cumplido el criterio para finalizar la división. Aquel atributo que permite separar mejor los datos respecto a la clasificación final se conoce como **Ganancia de Información**. El método para escoger el nodo a partir del cual surgirán las ramas más homogéneas y, por tanto, proporciona una Ganancia de información mayor, se calcula a partir de la **Entropía de Shannon** o grado de incertidumbre de una muestra:

$$Entropía\ (S) = \sum_{i=1}^{C} - p_i log_2(p_i)$$

Donde S es el conjunto de datos, C el número de clasificaciones a usar y p_i la proporción de datos que hay en la clasificación i en la muestra (Lantz, 2015).

En los últimos años, el algoritmo RF ha ganado popularidad entre la comunidad de bioinformáticos y es cada vez más utilizado en la clasificación de datos de *microarrays* y otros datos moleculares (Diaz-Uriarte y Alvarez de Andres, 2006). La razón es que (1) es aplicable cuando hay más predictores que observaciones, (2) no se ve afectado por el "ruido", (3)

incorpora las interacciones entre predictores, (4) se basa en la teoría del aprendizaje conjunto que permite al algoritmo aprender con precisión funciones de clasificación simples y complejas, (5) es aplicable para tareas de clasificación tanto binarias como de categoría múltiple, y (6) no precisa de parámetros muy ajustados (Breiman, 2001). A pesar de ello, el modelo no es fácilmente interpretable.

2. MÉTODOS

2.1. SELECCIÓN DE LA BASE DE DATOS Y PREPROCESAMIENTO

En este estudio se han considerado las bases de datos procedentes de cuatro cohortes distintas: cáncer de mama (*Breast Invasive Carcinoma*, BRCA), cáncer de colon (*Colon Adenocarcinoma*, COAD), cáncer de próstata (*Prostate Adenocarcinoma*, PRAD) y cáncer de pulmón (*Lung Adenocarcinoma*, LUAD). Estos cuatro tipos de cáncer han sido seleccionados en base a su elevada tasa de incidencia y mortalidad (Organización Mundial de la Salud, 2018), además estos comparten entre sí que son adenocarcinomas.

El Atlas del Genoma del Cáncer

El **Atlas del Genoma del Cáncer** (TCGA, por sus siglas en inglés) es un proyecto iniciado en 2006 con el objetivo de caracterizar los genomas cancerosos de distintos tipos de tumores e integrar los catálogos completos de datos genómicos, trascriptómicos, epigenómicos y proteómicos. Instituciones de todo el mundo han unido sus fuerzas para lograr este objetivo, siendo superadas sus expectativas en 2013, cuando ya contaban con más de 7000 casos analizados procedentes de 27 tipos de tumores distintos. Actualmente el Consorcio Internacional del Genoma del Cáncer (ICGC) incorpora los datos de TCGA.

El proyecto, dentro del marco epigenético, se centra en el proceso de metilación del ADN, dado que este es, sin duda, la modificación epigenética mejor caracterizada (Jones, 2012) y la única aproximación clínica al epigenoma en la actualidad. Los datos de metilación seleccionados proceden de la plataforma *Illumina Infinium HumanMethylation450* alineados con el genoma de referencia *hg19*. Para la descarga se empleó la librería de R *TGCABiolinks*, que permite el acceso al repositorio TCGA. A partir de la tecnología de *Illumina* se obtienen los valores beta de metilación, los cuales están comprendidos en un rango entre el 0 y el 1, donde 0 se corresponde con las sondas totalmente desmetiladas y el 1 con las sondas totalmente metiladas.

La preparación de la base de datos de metilación para su análisis consistió en la eliminación de las sondas correspondientes a rs (ensayo SNP), que se emplean en la medida de la variación genómica y no en el nivel de metilación, pudiendo afectar a los resultados sobre el estado de metilación global (Wang et al., 2012). Seguidamente se eliminaron las sondas correspondientes a los cromosomas X e Y, para eliminar artefactos potenciales que se originan de la presencia de una proporción diferente de hombres y mujeres (Marabita et al., 2013), y finalmente los valores faltantes (NA).

Illumina 450k BeadChip

La tecnología de *Infinium HumanMethylation450 array* permite la obtención del estado de metilación de más de 450k sitios CpG localizados a lo largo de todo el genoma, cubriendo el 99% de los genes de la base de datos *RefSeq*. Para lograrlo, esta tecnología emplea dos tipos de ensayos químicos diferentes: *Infinium I* e *Infinium II*. *Infinium I* usa dos sondas (una para el alelo metilado y otra para el alelo no metilado), utilizando un único color como señal de detección, mientras que *Infinium II* emplea una sonda que identifica e hibrida con el ADN metilado y no metilado indistintamente, y que diferencia empleando dos canales de color (Wang et al., 2018). Posteriormente, tras un proceso de preprocesamiendo de los datos, el estado de metilación se determina mediante el valor beta (Dedeurwaeder et al., 2013):

$$\beta = \frac{M}{M + U + \alpha}$$

donde M se corresponde con la intensidad de metilación, U con la intensidad de desmetilación en cada una de las regiones CpG y α generalmente es 100.

Los datos de expresión génica seleccionados se correspondían con los datos normalizados de la secuenciación de ARN disponibles en el repositorio TCGA a partir de la plataforma *Illumina HiSeq*, los cuales están alineados con el genoma de referencia *hg19*. El proceso de normalización del número de lecturas se lleva a cabo por el método FPKM (Fragmentos esperados por Kilobase de transcripción por Millón de fragmentos secuenciados) y su variante, el Cuartil Superior FPKM:

$$FPKM = \frac{RC_g * 10^9}{RC_{pc} * L}; \qquad FPKM - UQ = \frac{RC_g * 10^9}{RC_{g75} * L}$$

donde RC_g es el número de lecturas mapeadas en un gen, y RC_{g75} es su percentil 75; RC_{pc} el número de lecturas asignadas a todos los genes codificantes de proteínas y L el tamaño del gen en pares de bases (Bullard et al., 2010; Li y Dewey, 2011). Ambos métodos son análogos a RPKM (Mortazavi et al., 2008). Al igual que en la descarga de los datos de metilación se empleó la librería de R *TCGABiolinks*.

Para el estudio de las alteraciones genéticas se seleccionaron las bases de datos de las mutaciones somáticas y la variación en el número de copias (CNVs) presentes en las muestras tumorales de cada una de las cohortes de estudio. La base de datos de las mutaciones somáticas se corresponde con mutaciones no sinónimas, y para su descarga se emplearon las librerías *TCGABiolinks* y *cuatedTCGAData*, atendiendo a cuál de las dos contuviera el mayor número de muestras coincidentes con el resto de las bases de datos. La diferencia en las muestras disponibles en cada una de las librerías reside en que la primera obtiene los datos de TCGA a partir del portal del cáncer Genomic Data Commons (GDC) y la segunda a partir del repositorio Broad Genome Data Analysis Center (GDAC) Firehose. Por el contrario, para la descarga de las CNVs únicamente se empleó *curatedTCGAData*. La base de datos con las CNVs disponía de los datos procedentes del algoritmo GISTIC. Este algoritmo trata de identificar las regiones de amplificación o eliminación significativamente alteradas en conjuntos de pacientes, indicando el nivel de número de copias por gen: "-2" deleción total, "-1" deleción parcial, "0" no hay cambios, "1" amplificación parcial, y "2" amplificación total (cBioPortal).

En las bases de datos correspondientes a los datos de metilación y expresión génica las muestras seleccionadas se corresponden con muestras apareadas de tejido normal y tumoral, siendo identificadas a partir de su código de barras o *barcodes* de TCGA. Además, las muestras seleccionadas en todas las bases de datos (metilación, CNVs, mutaciones y expresión génica) debían ser coincidentes entre sí, es decir, debían pertenecer al mismo paciente; evitando así añadir el error asociado a la variación intrínseca del sujeto. Esto ha conducido a cierta disparidad en el tamaño muestral de cada una de las bases de datos. En la *Tabla 2* se muestra el tamaño muestral de cada una de las bases de datos seleccionadas para cada una de las cohortes de estudio:

Tabla 2: Número de muestras incluidas en cada una de las bases de datos descargadas. El número de muestras de tejido tumoral y normal son coincidentes.

Cohorte	Base de datos	Tumoral	Normal	TOTAL
COAD	Metilación	30	30	**60**
	Expresión	32	16	**48**
	Mutación	27		**27**
	CNV	29		**29**
BRCA	Metilación	50	50	**100**
	Expresión	58	41	**99**
	Mutación	49		**49**
	CNV	49		**49**
LUAD	Metilación	28	28	**56**
	Expresión	40	56	**58**
	Mutación	20		**20**
	CNV	28		**28**
PRAD	Metilación	50	50	**100**
	Expresión	50	35	**85**
	Mutación	43		**43**
	CNV	50		**50**

Previo a su análisis las bases de datos se analizaron gráficamente para comprobar la presencia de artefactos técnicos que pudieran afectar a los resultados de los análisis y precisaran un proceso de filtrado.

2.2. ANÁLISIS DEL METILOMA

El análisis del metiloma se subdivide a su vez atendiendo a los distintos objetivos de estudio.

1) **Búsqueda de los perfiles de metilación diferencial entre las muestras de tejidos normal y canceroso para cada una de las cohortes de estudio**

En primer lugar, se llevó a cabo una exploración de las diferencias entre el estado de metilación global del ADN de las muestras pertenecientes a tejido normal y tumoral de cada una de las cohortes por separado. Para tal fin se compararon gráfica y estadísticamente las medias de cada paciente en cada uno de los tejidos, empleando para ello la función *TCGAvisualize_meanMethylation* de la librería *TCGABiolinks*.

En segundo lugar, se procedió a estudiar las regiones o sondas diferencialmente metiladas en cada tipo de muestras, para ello R dispone de varios paquetes con funciones implementadas para tal fin. En el presente trabajo se compararon los resultados obtenidos en tres librerías: *TCGABiolinks*, *ABC.RAP* y *limma*; las cuales emplean diferentes aproximaciones estadísticas.

TCGABiolinks presenta una función, *TCGAanlayze_DMR*, que calcula las regiones diferencialmente metiladas (DMR) entre dos grupos a partir de los valores beta obtenidos directamente del repositorio TCGA. Para ello y, en primer lugar, calcula la media de metilación de cada uno de los sitios CpG en las muestras tumorales y normales respectivamente. Posteriormente la expresión diferencial de ambos grupos se analiza empleando el test de Wilcoxon ajustado por el método de Benjamini y Hochberg (Benjamini y Hochberg, 1995). Los parámetros de corte especificados fueron de una diferencia mínima entre los valores beta absolutos de 0.25 y un p.valor ajustado menor o igual a 0.01 (TCGABiolinks Help Documents-a).

La siguiente función empleada fue *ttest_data* de la librería *ABC.RAP*, la cual emplea el análisis t-test para varianzas desiguales (ABC.RAP vignettes).

Finalmente, la última librería utilizada fue *limma*, la cual está implementado para el análisis de *microarrays*. El procedimiento de uso de esta librería comienza por el desarrollo de una matriz de diseño y de contrastes para realizar el análisis mediante modelos lineales, que engloba el análisis de la varianza y la regresión. En el presente trabajo se diseñó una matriz identificando las muestras tumorales y normales de cada una de las cohortes, y la matriz de contrastes se correspondía con las comparaciones de cada muestra normal con su correspondiente tumoral para cada una de las cohortes. Seguidamente las pruebas de significación comprenden un t-test con la varianza regularizada mediante el empleo de modelos bayesianos empíricos. Finalmente, a fin de controlar el número de falsos positivos (FDR), los p.valores se ajustan empleando el método de Benjamini y Hochberg (Ritchie et al., 2015), al igual que la función de *TCGABiolinks*.

En el análisis llevado a cabo con *ABC.RAP* y *limma* se emplearon los valores M de metilación. Estos, de acuerdo con Du et al. 2010, presentan mejores propiedades estadísticas que los valores beta, por el contrario, los valores beta son más fáciles de interpretar y se suelen emplear en el análisis gráfico, debido a que sus valores están comprendidos entre 0 y 1 y los valores M entre 5 y -5. A continuación se muestra la fórmula para su cálculo a partir de los valores beta:

$$M = log_2\left(\frac{\beta}{1-\beta}\right)$$

2) Comparación del perfil de metilación aberrante global entre las diferentes cohortes de estudio

Para estudiar la relación o cercanía entre los perfiles globales de metilación en las distintas cohortes de estudio se emplearon tanto técnicas estadísticas como de *Machine Learning.* En primer lugar, se llevó a cabo un gráfico de proximidades (*Multi-Dimensional Scaling*, MDS) con las medias de las muestras tumorales y normales de cada una de las cohortes. Los gráficos MDS son un método de agrupamiento no supervisado que permite visualizar la relación de cercanía o distancia entre las variables. Este método se basa en el análisis de las componentes principales. A continuación, se repitió el proceso, pero en esta ocasión con la distancia euclídea de todas las sondas presentes en cada una de las muestras tumorales de las distintas cohortes.

Seguidamente se llevó a cabo un análisis jerárquico con 2 métodos distintos, los cuales emplean diferentes algoritmos para calcular la distancia entre las muestras. El primero de ellos es el método de *amalgamiento completo* (*complete linkage*), también conocido como estrategia de distancia máxima o similitud mínima, en el cual la distancia o similitud entre dos clústeres viene dada por sus elementos más dispares. El segundo se corresponde con el método Ward, este es un procedimiento jerárquico donde en cada etapa se unen los dos clústeres para los cuales se tenga el menor incremento en el valor total de la suma de los cuadrados de las diferencias de cada individuo al centroide del clúster.

El siguiente proceso para comparar el estado de metilación de las muestras tumorales de las distintas cohortes se llevó a cabo mediante un análisis con *limma*, de manera similar al explicado en el apartado anterior, pero en esta ocasión se compararon las muestras tumorales de todas las cohortes entre sí ajustando para ello la matriz de contrastes.

Finalmente, la especificidad del estado de metilación de cada una de las sondas se estudió mediante el algoritmo de agrupamiento no supervisado, *k-means* (explicado en el apartado 1.4). Este método se aplicó con la función *kmeans* de la librería *stats*, especificando el valor de k o número de clústeres en 4, uno por cada cohorte, y empleando el algoritmo Hartigan -Wong, ya que presenta mejores resultados que el resto de métodos disponibles (kmeans RDocumentation).

En todos los análisis anteriormente descritos se emplearon los valores M de metilación debido a sus mejores propiedades estadísticas.

2.3. ANÁLISIS DE LA RELACIÓN ENTRE EL ESTADO DE EXPRESIÓN GÉNICA Y EL ESTADO DE METILACIÓN GLOBAL

La relación entre el estado de expresión génica y el estado de metilación es un objetivo menor del proyecto, para llevarlo a cabo se emplearon las funciones implementadas en la librería *TCGABiolinks*. Dado que los datos se descargaron previamente normalizados, el primer paso fue obtener los genes diferencialmente expresados (DEGs) en cada una de las cohortes con la función *TCGAanalyze_DEA* (*Differential Expression Analysis*, DEA). Esta función llama a su vez a un conjunto de funciones de la librería *edgeR*, especializada en el estudio de la expresión génica diferencial. Los parámetros de corte especificados para la selección de los DEGs fueron de una valor de logFoldChange en valores absolutos mayor o igual a 1 y un p.valor menor o igual a 0.01 (TCGABiolinks Help Documents-b). Tras obtener los genes diferencialmente expresados y los sitios CpG diferencialmente metilados (descrito en el apartado 1), la relación entre ambos se estableció mediante un gráfico *starburst* (*TCGAvisualize_starburst*). Este gráfico representa el log_{10} del p.valor corregido para los datos de metilación en el eje X y para los datos de expresión génica en el eje Y. Además, todos aquellos genes que cumplían ciertos parámetros como una diferencia mínima entre los valores beta absolutos de 0.25 y un valor de logFoldChange en valores absolutos mayor o igual a 1 se remarcaban. El proceso se repitió dos veces, la primera considerando aquellos genes que presentaban un p.valor ajustado tanto para la expresión génica como para la diferencia del estado de metilación menor o igual a 10^{-2} y la segunda para un p.valor ajustado menor o igual a 10^{-5}.

2.4. ANÁLISIS DE LA RELACIÓN ENTRE EL PERFIL DE METILACIÓN ABERRANTE Y LAS ALTERACIONES GENÉTICAS

El último objetivo del presente proyecto es esclarecer la relación entre las alteraciones genéticas presentes en los mecanismos de regulación del epigenoma y los perfiles de metilación aberrante hallados en cada una de las cohortes de estudio. Este estudio se llevó a cabo únicamente con las bases de datos correspondientes a BRCA y PRAD, debido a que los *barcodes* indicativos de los pacientes debían coincidir en todas las bases de datos, y el tamaño muestral final coincidente en las bases de datos procedentes de COAD y LUAD era menor de 30, tamaño insuficiente para llevar a cabo una clasificación de ML significativa.

El primer paso para llevar a cabo este análisis fue la selección de las mutaciones y número de copias que se correspondían con los genes candidatos, extraídos de la literatura (You y Jones, 2012; Plass et al., 2013; Feinberg et al., 2016). Seguidamente, los perfiles de metilación aberrante en cada una de las cohortes se extrajeron de un análisis *k-means* empleando los valores beta de las regiones significativas obtenidas con *limma*. El número de clústeres idóneos (k) para cada cohorte se calculó con tres métodos: el método de la silueta, el método del codo (*elbow method*) y un análisis de discrepancias (*gap analysis*), disponibles con la función *fviz_nbclust* de la librería *factoextra* (fviz_nbclust RDocumentation). Finalmente, el número de k idóneos se seleccionó en base a cuál de los métodos aportaba una distribución de datos más homogénea, evitando así que uno de los perfiles tenga mucho más peso que los otros y balancee la clasificación. El último paso fue el uso de dos algoritmos *Machine Learning*:

Random Forest y *Suppot Vector Machines* (explicados en el apartado 1.4), para la clasificación de los perfiles de metilación de acuerdo con las alteraciones genéticas de los genes candidatos, comprobando así si existe un patrón en las alteraciones genéticas que explique dichos perfiles. Para ello se emplearon la función *ksvm* de la librería *kernlab* en el desarrollo del algoritmo SVM y la función *randomForest*, de la librería que lleva el mismo nombre, para el desarrollo del algoritmo RF. En el desarrollo del algoritmo SVM se emplearon diferentes modelos: ANOVA, el método lineal y el método Gaussiano. El análisis de ML se repitió incluyendo únicamente los perfiles de metilación y las CNVs.

Tabla 3: Extracto de la base de datos empleadas en la clasificación ML, donde se incluyen el perfil de metilación obtenido tras el agrupamiento *k-means* y el nivel de número de copias por gen y el número de mutaciones somáticas de los genes implicados en los mecanismos de modificación epigenética para cada una de las muestras coincidentes en las tres bases de datos.

Muestra	cnv.TP53	mut.TP53	Perfil.Metilación
TCGA-A7-A0D9-01	-1	15	2
TCGA-AC-A2FM-01	-2	2	3

Todos los procesos descritos en esta sección se llevaron a cabo con R. Para más información acerca de la versión accedan al Anexo 3: Versión de R.

3. RESULTADOS Y DISCUSIÓN

Búsqueda del perfil de metilación aberrante en las muestras tumorales de las distintas cohortes de estudio

El primer análisis llevado a cabo fue la comparación entre las muestras de tejido tumoral y normal de los distintos tipos de cáncer por separado. El perfil de metilación medio de cada muestra se representa en el diagrama de cajas (Figura 2), en el cual se incluye el p.valor de la comparación entre las medias de metilación de las muestras tumorales y normales para cada tipo de cáncer respectivamente. Una primera visualización refleja una mayor disparidad entre las medias de metilación de las muestras tumorales, mientras que la varianza de las muestras normales es mucho menor. A pesar de ello, la gran mayoría de las muestras presentan un valor beta medio en torno a 0.45 y 0.5. La media general para de cada tipo de muestras (normales y tumorales) son relativamente similares en todos las cohortes, exceptuando PRAD, correspondiente al cáncer de próstata, el cual es el único que presenta una diferencia significativa entre las medias de las muestras normales y tumorales con un p.valor igual a 2.5e-07 (inferior a 0.001). Estos resultados son similares a los observados al comparar la frecuencia de los valores beta en cada tipo de muestra (Figura 1 Anexo 2), donde se observa que apenas hay diferencias en el estado de metilación global, la única excepción es el cáncer de próstata que presenta un aumento de las frecuencias de los valores beta superiores a 0.75. Todo ello parece indicar que este tipo de cáncer presenta un marcado perfil de hipermetilación favoreciendo el proceso tumoral.

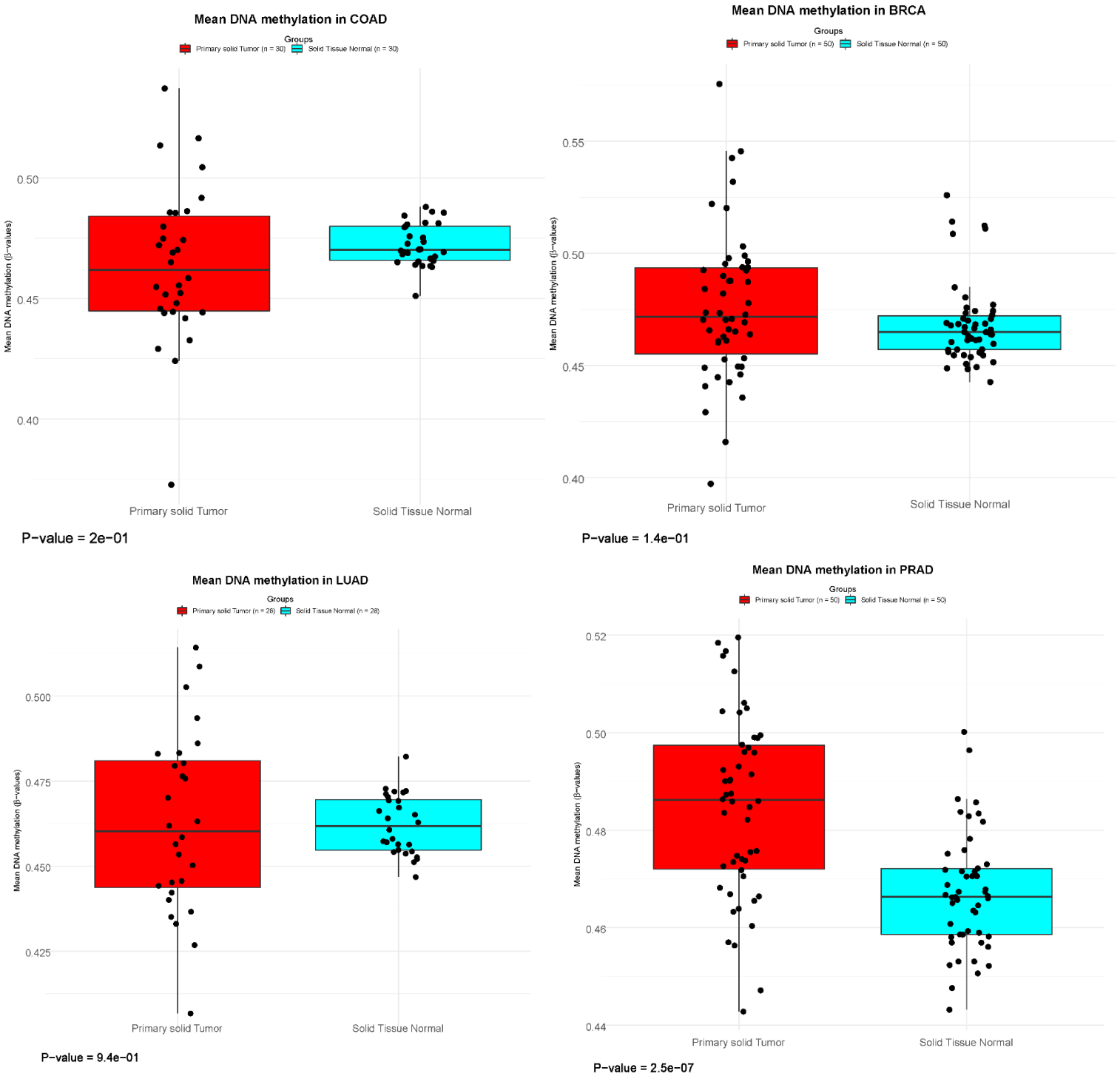


Figura 2: Diagrama de cajas con la distribución de las medias de metilación en las muestras tumorales (rojo) y normales (azul), calculadas a partir de los valores β, en los distintos tipos de cáncer de estudio. De arriba abajo y de izquierda a derecha se representan: cáncer de colon (COAD), de mama (BRCA), de pulmón (LUAD) y de próstata (PRAD).

Para comprender mejor las diferencias entre los estados de metilación de las muestras tumorales y normales de los distintos tipos de cáncer estudiados se llevó a cabo un análisis de las regiones o sondas diferencialmente metiladas con 3 librerías de R distintas. A continuación, en la Tabla 4 se muestran el número total de regiones diferencialmente metiladas entre muestras normales y tumorales obtenidas con cada una de las funciones de R empleadas para tal fin.

Como se puede apreciar el número de sondas significativamente diferentes entre las muestras tumorales y normales obtenidas con *ABC.RAP* y *limma* son similares, mientras que los resultados con *TCGABiolinks* se corresponden aproximadamente con la quinta parte del número obtenido con el resto de las funciones. Las principales diferencias entre la función de *TCGABiolinks* y las funciones del resto de librerías es que ésta emplea los valores beta de metilación, y no los M, y el análisis estadístico se lleva a cabo con un test estadístico no paramétrico como es el test de Wilcoxon, mientras que las otras funciones emplean un t-test corregido. Por esta razón, y debido a que presenta un sistema de corrección de falsos positivos con el método Benjamini y Hochberg, en los siguientes análisis se emplearon los sitios diferencialmente metilados obtenidos con *limma*.

Comparando el número de regiones significativas en cada una de las cohortes se aprecia que BRCA posee el mayor número de sitios diferencialmente metilados tanto con *ABC.RAP* como con *limma*. Estos resultados varían ligeramente con los obtenidos en otros estudios, que típicamente han relacionado el cáncer de mama con un perfil poco metilado, en comparación de otros tipos de cáncer como el de colon (Witte et al., 2014; Costello et al., 2001). A pesar de ello, a raíz de una investigación obtenida a partir de los datos de TCGA se ha obtenido que al menos uno de los cuatro subtipos de cáncer de mama presenta un marcado patrón de hipermetilación (Cancer Genoma Atlas N., 2012).

Por otro lado, COAD y PRAD, ostentan la segunda y tercera posición atendiendo a la librería *ABC.RAP* o *limma*, A pesar de ello, el número de regiones significativas es muy parecida en ambas cohortes. Finalmente, LUAD, presenta el menor número de regiones diferencialmente metiladas en ambos paquetes, pero al igual que ocurre con PRAD, hay mayores diferencias entre las funciones empleadas. Estas diferencias no se aprecian de manera tan drástica al comparar los resultados significativos de *limma* a partir del p.valor y no del p.valor ajustado como están incluidos en la Tabla 3; en este caso el número de regiones significativas es de 173590 para BRCA, 139458 para COAD, 130722 para PRAD y 102236 para LUAD; resultados mucho más parecidos a los obtenidos con *ABC.RAP*.

Tabla 4: Número de regiones significativamente metiladas o desmetiladas (p.valor <= 0.01) en cada una de las cohortes de estudio y comparando con los tres paquetes de R empleados en el análisis. Los resultados mostrados de *limma* se corresponden con los obtenidos empleando el p.valor ajustado.

	COAD	BRCA	PRAD	LUAD
TCGABiolinks	24380	15079	9513	3534
ABC.RAP	135671	163337	146647	123808
limma	124593	162265	117682	76239

A continuación, en la Tabla 5, se muestran el número de regiones hipometiladas e hipermetiladas del total de regiones significativas obtenidas con *limma*. En esta tabla se puede apreciar que prácticamente en todos los tipos de tumores el número de regiones hipermetiladas en el tejido tumoral es mayor que las hipometiladas, este patrón es especialmente claro en PRAD y BRCA, donde hay una gran diferencia en el número total de regiones hipometiladas e hipermetiladas respectivamente (Figura 3b y 3c). Estos resultados contrastan en gran medida con lo establecido en un gran número de revisiones, en las que destacan que el ADN global canceroso se caracteriza por un estado de hipometilación general y presenta, en menor número, unas regiones hipermetiladas asociadas a las CGI (Esteller et al. 2007).

COAD, por otro lado, presenta un mayor número de sondas hipometiladas. Otros estudios llevados a cabo con muestras de cáncer de colon han encontrado resultados parecidos con un 44% de regiones hipometiladas frente a los 59% obtenidas en el presente proyecto, y un 56% de regiones hipermetiladas frente a un 41% (Irizarri et al., 2009). Teniendo en cuenta la variabilidad presente entre sujetos, la cual está especialmente influenciada por la edad (Toyota et al., 1999); en ambos trabajos se encuentran distribuciones más o menos homogéneas de sitios hipo e hipermetiladas. Finalmente, LUAD es la cohorte que presenta una menor diferenciación entre el número total de sondas hipermetiladas e hipometiladas.

Tabla 5: Número de regiones diferencialmente hipometiladas e hipermetiladas en cada una de las cohortes de estudio. Los resultados se corresponden con las regiones diferencialmente metiladas que presentan un p.valor ajustado menor o igual a 0.01.

	COAD	BRCA	PRAD	LUAD
Hipometiladas	73710	69801	38211	37489
Hipermetiladas	50883	92462	79471	38750
TOTAL SIGNIFICATIVAS	**124593**	**162265**	**117682**	**76239**

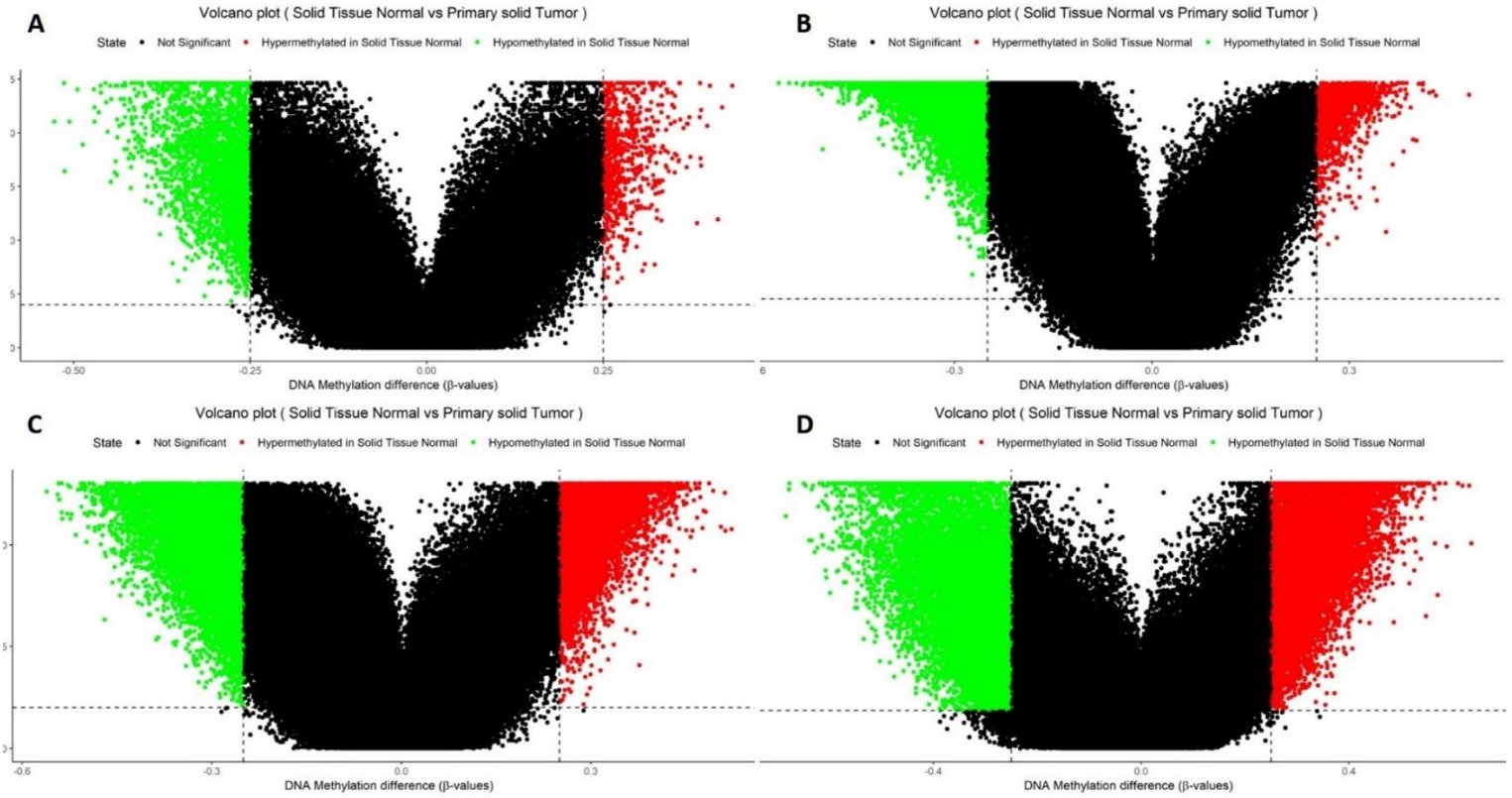


Figura 3: Volcano plots en los que se indican las regiones diferencialmente metiladas en las distintas cohortes de estudio: pulmón (A), próstata (B), mama (C) y colon (D). En verde se muestran las regiones hipermetiladas en las muestras de tejido tumoral y en rojo las hipometiladas en las muestras tumorales con respecto a las normales.

Comparación entre el perfil de metilación global de las distintas cohortes de estudio

Seguidamente se procede a analizar la relación entre el estado de metilación de las cuatro cohortes de estudio. La primera inspección de la relación entre tipos de cáncer se llevó a cabo mediante un gráfico de proximidades con las medias de los valores M de las muestras tanto tumorales como normales de los cuatro tipos de cohortes (Figura 4). En este gráfico se puede apreciar que la primera componente principal (eje X) está más relacionada con el tipo de tumor, dividiendo prácticamente por completo a las muestras tanto tumorales como normales de COAD y LUAD a la derecha, y PRAD y BRCA a la izquierda. Por otro lado, la segunda componente principal (eje Y) parece explicar la diferencia entre el perfil de metilación tumoral y normal. Los resultados son coincidentes al estudiar la distancia euclídea de las muestras tumorales (Figura 5). En la Figura 5 se muestra, además, como los puntos correspondientes a las muestras de una cohorte forma una nube de puntos que no se entremezcla con la de las otras cohortes. Todo ello parece indicar que a nivel global cada tipo de cáncer presenta un patrón de metilación específico y único.

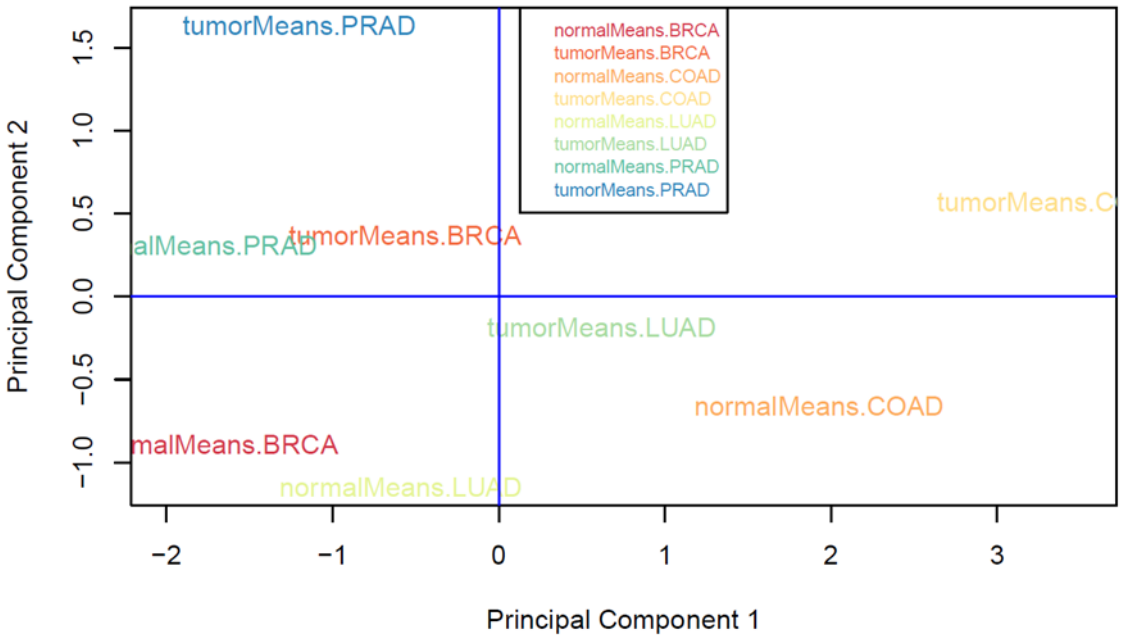


Figura 4: Análisis de proximidades o MDS de las muestras tumorales y normales de las cuatro cohortes de estudio.

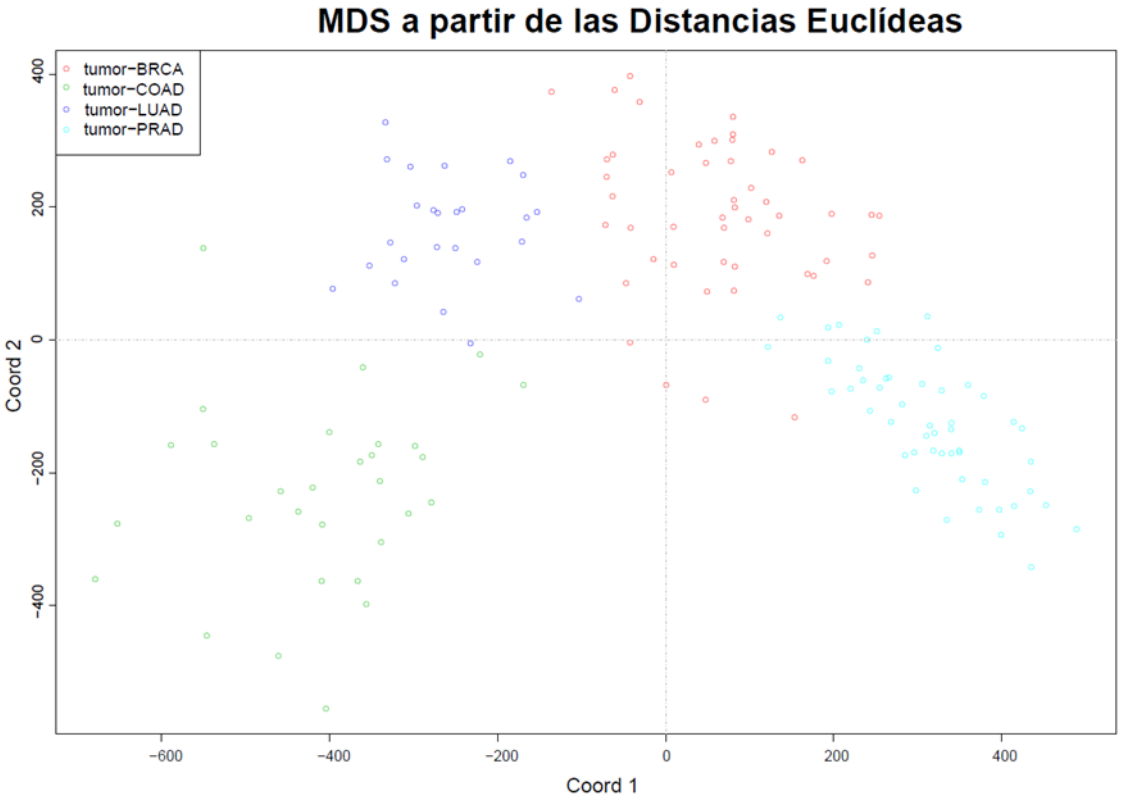


Figura 5: Gráfico de proximidades a partir de la distancia euclídea de los sitios CpG presentes en las muestras tumorales de las cohortes de estudio: BRCA en rojo, COAD en azul oscuro, LUAD en verde y PRAD en azul claro.

Otros análisis como el análisis jerárquico con el método Ward (Figura 6) y el análisis de agrupamiento con *k-means* (Tabla 6) consiguen agrupar todas las muestras tumorales de acuerdo con el tipo de cohorte al cual corresponden (exceptuando 1 muestra agrupada erróneamente en el caso de *k-means*). Estos datos corroboran la especificidad de un perfil global de metilación aberrante para cada tipo de tumor atendiendo al tejido en el cual se desarrolla. En otro estudio empleando un método de clasificación jerárquica no supervisada a partir de datos de metilación procedentes del repositorio TCGA ha conseguido a su vez una clasificación prácticamente completa de las muestras de tejido tumoral de cáncer de colon y pulmón (Hao et al., 2017), confirmando la especificidad de los patrones globales de metilación para cada tipo de tumor. Comentar que en el caso de la agrupación jerárquica con el método *complete linkage* no se obtuvieron los resultados tan precisos como los obtenidos con el método Ward (Figura 2 Anexo 2).

Además, en el dendograma se puede apreciar (Figura 6) como BRCA y PRAD están más relacionadas entre sí y lo mismo ocurre con LUAD y COAD. Estos resultados coinciden con los obtenidos en los análisis de proximidades. Una posible explicación de ello puede deberse a que BRCA y PRAD presentan un estado de hipermetilación más marcado, mientras que COAD Y LUAD poseen un estado de metilación intermedio.

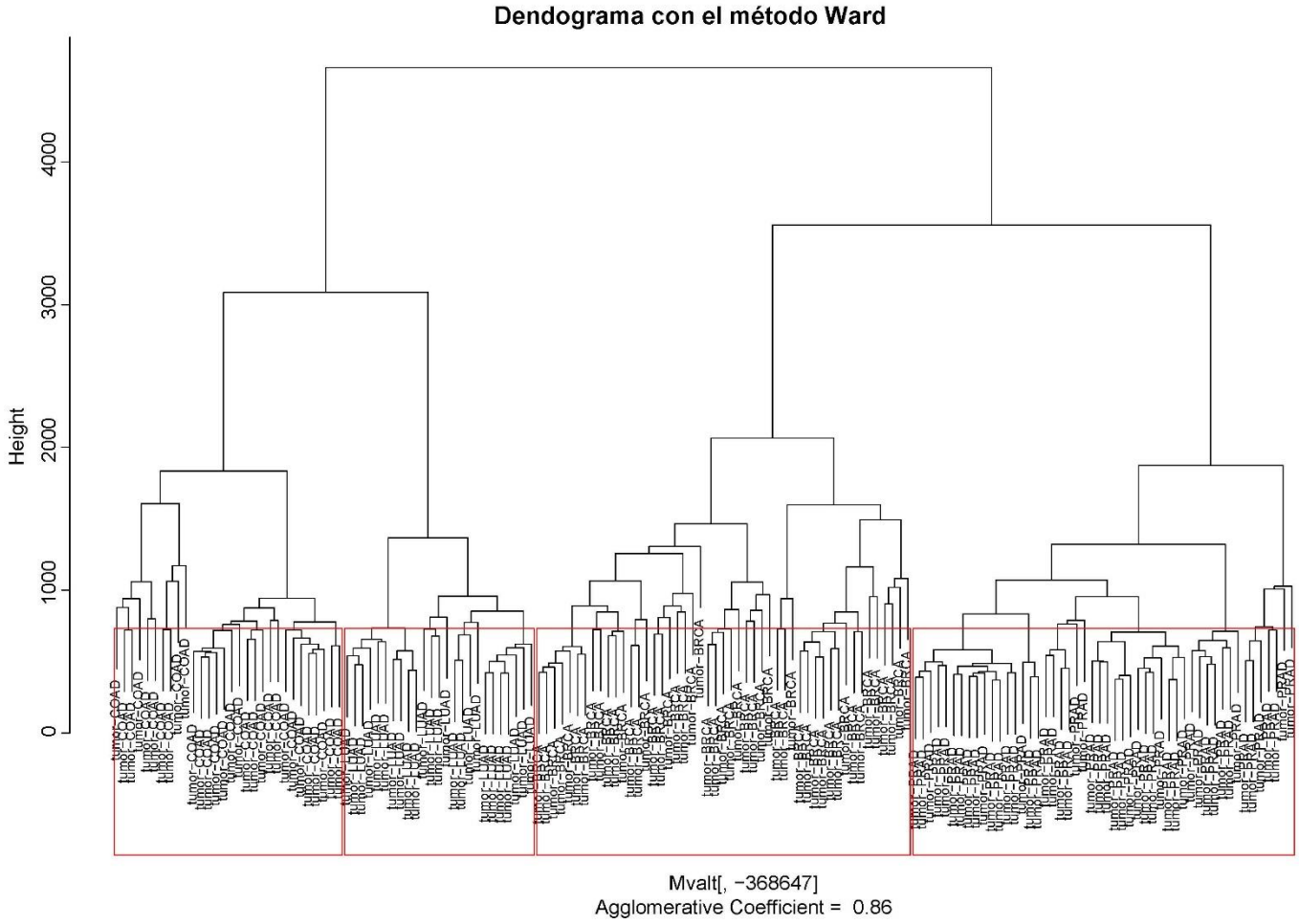


Figura 6: Dendograma empleando el método de Ward. En el dendograma se han añadido 4 clústeres que coinciden con cada una de las cohortes de estudio. De izquierda a derecha se representan las muestras de COAD, LUAD, BRCA y PRAD.

Tabla 6: Resultados obtenidos a partir de la clasificación con *k-means*, en esta tabla re representa el número de muestras clasificadas en cada clúster y la cohorte de origen a la cual corresponde.

Clústeres	BRCA	COAD	LUAD	PRAD
1	48	0	0	0
2	0	30	0	0
3	0	0	0	50
4	1	0	25	0

Los resultados obtenidos a partir de la matriz de contrastes de *limma* comparando las muestras tumorales de las distintas cohortes de estudio entre sí se refleja en la Tabla 7. En esta tabla se muestra el número de regiones coincidentes y aquellas que presentan un perfil de metilación diferencial entre sí. Como se aprecia, únicamente comparando BRCA con COAD y PRAD respectivamente se obtiene un mayor número de regiones similares, mientras que el resto de las comparaciones presentan mayores sitios CpG diferencialmente metilados ente sí, originando mayor distancia entre los perfiles de metilación global. Estos resultados concuerdan con los obtenidos en el resto de los análisis, en los cuales el perfil de metilación de PRAD es más similar a BRCA y este, a su vez de COAD. Esta relación se observa de manera más clara al comparar las siluetas resultado de la agrupación de las muestras con *k-means* (Figura 3 Anexo 2), donde PRAD y BRCA se solapan casi por completo, y COAD y BRCA se solapan ligeramente.

Tabla 7: Número de regiones significativas y no significativas comparando las muestras tumorales de las cuatro cohortes de estudio con *limma*. En azul se indican el valor con mayor número de regiones.

	COAD-BRCA	COAD-LUAD	COAD-PRAD	BRCA-LUAD	BRCA-PRAD	PRAD-LUAD
N. de sondas significativas	176767	259084	194051	263077	177859	289482
N. de sondas no significativas	193150	110833	175866	106840	192058	80435

A partir de los resultados obtenidos en el estudio diferencial del estado de metilación entre las muestras tumorales y normales, así como los resultados procedentes de la comparación entre las cuatro cohortes de estudio, es posible afirmar que cada tipo de tumor presenta un patrón de metilación global característico que lo diferencia del resto. A pesar de ello, al comparar dichos perfiles entre las muestras tumorales y normales apenas hay diferencias significativas, exceptuando en el cáncer de próstata. Todo ello remarca la importancia de enfocar el estudio del epigenoma del cáncer en las regiones diferencialmente metiladas, y más concretamente estudiar su grado de metilación a lo largo del proceso de desarrollo tumoral. Este tipo de estudio permitirá esclarecer las regiones pronóstico de cada tipo de tumor y pueden resultar claves en el diagnóstico temprano.

Relación entre el estado de expresión génica y el estado de metilación global

A continuación, se muestra la tabla (Tabla 8) con los resultados obtenidos al comparar la expresión génica y la metilación en el ADN canceroso. En esta se aprecia que, a excepción de LUAD, hay mayor número de genes infraexpresados que sobreexpresados. Estos resultados están secundados por un gran número de estudios, en los que se vincula el desarrollo tumoral con un silenciamiento de los genes, especialmente de los genes supresores de tumores (Herman y Baylin, 2003; Egger et al., 2004; Feinberg y Ticko, 2004; Esteller, 2005). Por otro lado, al contrario de lo que muchos artículos remarcan, tanto los genes infraexpresados como los sobreexpresados se ven influenciados en gran medida por los procesos de hipermetilación. Para entender este proceso en mayor medida sería preciso conocer qué región génica se encuentra hipermetilada y cómo afecta al proceso de transcripción. En general, se ha documentado cómo la hipermetilación en las CGI reprime el proceso de transcripción al imposibilitar al complejo ARN polimerasa el acceso al ADN debido a la presencia de una proteína de unión a una metilcitosina (Baylin y Herman, 2000). De una manera u otra, las regiones que presentan un estado de metilación anormal, tanto hipermetilado como hipometilado, influyen en el acceso de los factores de transcripción al ADN dificultando o favoreciendo el proceso de transcripción. La representación gráfica de estos datos está disponible en la Figura 7 del Anexo 2.

Tabla 8: Número de genes infraexpresados y sobreexpresados y su relación con el estado de metilación que presenta, hipometilado o hipermetilado en las muestras tumorales de cada una de las cohortes de estudio. Los siguientes resultados se obtuvieron con los siguientes parámetros de corte: diferencia mínima de metilación, en valores absolutos, mayor o igual a 0.25; presentar un valor de logFoldChange mayor o igual a 1; y un valor FDR de expresión y metilación menor o igual a 10^{-2} y 10^{-5}, respectivamente.

	COAD		BRCA		PRAD		LUAD	
	10^{-2}	10^{-5}	10^{-2}	10^{-5}	10^{-2}	10^{-5}	10^{-2}	10^{-5}
Infra + Hiper	394	142	225	179	195	161	30	19
Infra + Hipo	181	84	88	70	29	20	2	1
Sobre + Hiper	158	55	161	104	122	78	75	30
Sobre + Hipo	142	68	97	53	16	5	22	14

Relación entre el perfil de metilación aberrante y las alteraciones génicas

Finalmente se comentará la relación entre el perfil de metilación aberrante y las alteraciones génicas en el cáncer de mama y próstata. Para la clasificación de los perfiles de metilación, en primer lugar, se calcularon el número de k óptimos, que en todas las cohortes variaban entre 2 y 4. El hecho de que ambas cohortes presenten un número de clústeres mayores de 1, indica que todos los tipos de tumores pueden ser clasificados en subtipos de acuerdo con el perfil de metilación que presenta. Seguidamente se llevó a cabo el algoritmo *k-means*, en la Tabla 9 se refleja el tamaño muestral presente en cada clúster para cada una de las cohortes empleadas en este análisis (BRCA y PRAD). De estas agrupaciones se seleccionó aquella que presentaba un tamaño muestral más homogéneo en los distintos clústeres, siendo k igual a 4 (indicado en azul, Tabla 9) tanto en PRAD como en BRCA. En el caso de BRCA otro estudio identifica la presencia de cuatro subtipos principales en esta cohorte (Cancer Genoma Atlas N., 2012). A pesar de tratar de seleccionar la distribución más homogénea, en ambos tipos de cáncer hay un perfil más dominante que el resto.

Tabla 9: Tamaño muestral de las bases de datos de cada una de las cohortes atendiendo a los tres valores de k óptimos. En azul se representa el número de clúster seleccionado para la clasificación de los perfiles de metilación en cada tipo de tumor.

k	BRCA	PRAD
2	14,35	7,43
3	11,27,11	7,12,31
4	**13,19,11,6**	**7,16,18,9**

Tras la clasificación de los perfiles de metilación con *k-means* y el aislamiento del número de copias y las mutaciones relacionadas con los genes implicados en los mecanismos epigenéticos (para más información ver Anexo1: Genes candidatos), se procedió a la clasificación ML con los algoritmos *Support Vector Machines* y *Random Forest*. El primer algoritmo aplicado fue SVM, en las Tablas 10 y 11 se muestran los parámetros obtenidos al estudiar la relación de los perfiles de metilación tanto con las mutaciones somáticas como con las CNVs, y únicamente con las CNVs, respectivamente. Seguidamente los valores obtenidos tras aplicar el algoritmo RF se muestran en la Tabla 12.

Tabla 10: Parámetros obtenidos al procesar los datos procedentes del cáncer de mama (BRCA) y próstata (PRAD) con un algoritmo *Support Vector Machines* con un modelo lineal, gaussiano y un ANOVA. En este modelo se tienen en cuenta tanto la contribución de las mutaciones somáticas como la variación en el número de copias.

MODELO	Accuracy	Kappa	Accuracy Lower	Accuracy Upper
BRCA				
Lineal	0.400	0.172	0.163	0.677
ANOVA	0.533	0.314	0.266	0.787
Gaussiano	0.333	0.000	0.118	0.616
PRAD				
Lineal	0.357	-0.016	0.128	0.649
ANOVA	0.429	0.097	0.177	0.711
Gaussiano	0.643	0.340	0.351	0.872

Tabla 11: Parámetros obtenidos al procesar los datos procedentes del cáncer de mama (BRCA) y próstata (PRAD) con un algoritmo *Support Vector Machines* con un modelo lineal, gaussiano y un ANOVA. En este modelo se tienen en cuenta únicamente la variación en el número de copias.

MODELO	**Accuracy**	**Kappa**	**Accuracy Lower**	**Accuracy Upper**
BRCA				
Lineal	0.333	0.080	0.118	0.616
ANOVA	0.533	0.314	0.266	0.787
Gaussiano	0.533	0.000	0.118	0.616
PRAD				
Lineal	0.357	0.016	0.128	0.649
ANOVA	0.357	0.000	0.128	0.649
Gaussiano	0.643	0.340	0.351	0.872

Tabla 12: Parámetros obtenidos al procesar los datos procedentes del cáncer de mama (BRCA) y próstata (PRAD) con un algoritmo *Random Forest*.

DATOS	**Accuracy**	**Kappa**	**Accuracy Lower**	**Accuracy Upper**	**P.valor**
BRCA mutaciones + CNV	0.400	0.151	0.163	0.677	0.382
BRCA CNV solo	0.533	0.356	0.269	0.787	0.088
PRAD mutaciones + CNV	0.357	0.000	0.128	0.647	0.910
PRAD CNV solo	0.357	-0.016	0.128	0.649	0.910

Como es posible apreciar en las tablas anteriores, en ambos algoritmos e incluyendo o no las mutaciones somáticas en la clasificación, los mejores valores de precisión (*accuracy*) obtenidos son de 0.643 y 0.533; indicando que alrededor del 50 y el 65% de los datos estarían bien agrupados en la clasificación. Así mismo el mejor valor del estadístico kappa sería 0.356 al clasificar las muestras de BRCA únicamente con las CNVs empleando RF y 0.340 al aplicar el modelo Gaussiano de SVM en PRAD, considerando tanto las mutaciones como las CNVs, o solo las CNVs. De acuerdo con Lantz et al., 2015, estos valores del estadístico kappa se corresponderían con una clasificación justa. Esto quiere decir que en los modelos aplicados se puede apreciar una tendencia, pero no es suficiente para afirmar una relación clara entre los perfiles de metilación y las alteraciones genéticas.

En base a los resultados obtenidos, no es posible afirmar qué algoritmo es mejor para clasificar los datos procedentes de *microarrays*, y más concretamente para estudiar las diferencias entre el estado genético y epigenético de las muestras; ya que para BRCA la mejor clasificación se obtiene con RF y para PRAD con SVM. A pesar de ello, la selección del modelo en SVM es clave

para la optimización del algoritmo, como es posible apreciar en las Tablas 10 y 11, donde el modelo lineal presenta los peores resultados. Por otro lado, la eliminación de las mutaciones somáticas en el marco general del análisis no mejora los resultados, pero en ambas cohortes los mejores estadísticos se obtienen al considerar únicamente las CNVs. En la Figura 7 se representa el grado de importancia de las primeras 30 variables, como se puede observar en estas posiciones únicamente se sitúan las CNVs; además, al eliminar las mutaciones (Fig.7b) el grado de importancia aumenta ligeramente, con un máximo de 0.8 sobre 1.

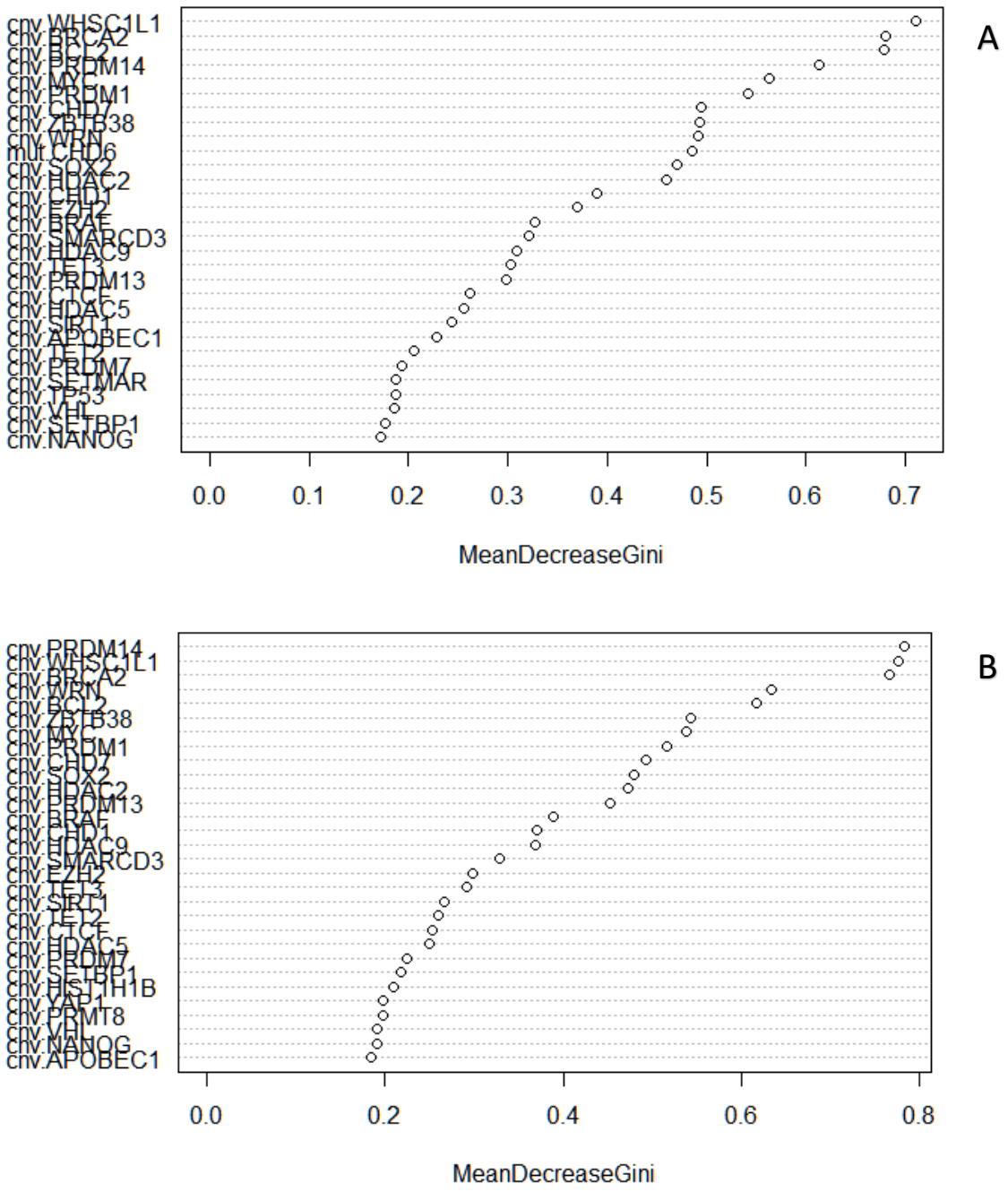


Figura 7: Importancia de las 30 primeras variables analizadas en el algoritmo *Random Forest*, obtenidas del análisis de las muestras tumorales de cáncer de próstata considerando las mutaciones somáticas y las CNVs (A) y únicamente las (CNVs).

A continuación, se enumeran algunas de las razones por las cuáles se han obtenido unos valores tan bajos de precisión y estadísticos kappa:

- El número de muestras coincidentes entre las bases de estudio, aun considerando únicamente las bases de datos con mayor tamaño muestral, sigue siendo muy bajo. La base de datos final empleada en los algoritmos para BRA contiene 46 muestras y la de PRAD posee únicamente 43. Esta limitación es clave en la fase de entrenamiento del modelo.
- Algunos de los grupos de clasificación obtenidos con *k-means* estaban poco representados y al dividir la base de datos para su entrenamiento y posterior testeo, su baja representación ha podido afectar a su correcta clasificación.
- La mayoría de las mutaciones somáticas se presentan con una frecuencia muy baja (Wood, 2007), por lo que para poder representar el efecto de las mutaciones en el marco epigenético se precisaría de una base de datos con mayor tamaño muestral.

Por todo ello, el principal problema en el desarrollo de los algoritmos ML fue la falta de un tamaño muestral representativo. Se descartan otras posibles alteraciones como los artefactos técnicos, debido a la selección estricta de las sondas empleadas, seleccionadas por su significatividad. Además, en el estudio llevado a cabo por Feber et al., 2014 se demuestra que las CNVs no alteran la señal de los chips *Infinium*.

A pesar de lo anteriormente expuesto, los resultados obtenidos indican, que tras depurar el modelo y aumentar el tamaño muestral, el número de copias anormal y las mutaciones somáticas presentes en los genes candidatos podrían explicar en gran parte los perfiles de metilación hallados. Abriendo las puertas a la mejor comprensión de la relación entre las alteraciones genéticas y epigenéticas en el desarrollo del cáncer.

4. CONCLUSIONES

En base a los resultados obtenidos se concluye que:

- Cada tipo de tejido presenta un perfil de metilación concreto y diferencialmente excluyente del hallado en otros tipos de tejidos.

- Los perfiles de metilación específicos de cada tipo de tejido no son equitativos al estado anormal de la célula. Por el contrario, las grandes diferencias entre el estado tumoral y normal se deben a las alteraciones producidas en regiones concretas del genoma, las cuales favorecen el proceso de formación y desarrollo tumoral.

- El estado de metilación diferencial y los genes que lo presentan en las muestras tumorales permite, incluso, la subclasificación de los tipos de tumores. Es por ello por lo que la identificación de estas regiones y conocer su evolución a lo largo del proceso tumoral puede ser un elemento clave para la detección temprana del cáncer y la búsqueda de dianas específicas en las terapias epigenéticas.

- La expresión anormal de los genes está influenciada tanto por los procesos de metilación como de desmetilación. Esto puede deberse a las regiones específicas que presentan el estado de metilación alterado, por ejemplo, en el cuerpo o en la región promotora del gen.

- El estudio de las alteraciones génicas, tanto las mutaciones somáticas como la variación en el número de las copias génicas, presentes en los genes reguladores de los procesos de metilación presentan un gran potencial para esclarecer la relación entre los procesos genéticos y epigenéticos implicados en el desarrollo del cáncer. En especial la variación del número de copias, del cual se dispone de poca bibliografía y, de acuerdo con los resultados, presenta una relación más marcada con los perfiles de metilación que las mutaciones somáticas estudiadas.

- El ML es una herramienta muy eficaz en la búsqueda de patrones en grandes bases de datos. Previo a su uso es conveniente la optimización de los parámetros, modelos y datos a estudiar para conseguir mejores resultados en la clasificación.

A continuación se incluye el link de acceso al repositorio GitHub con el código en R.

5. ABREVIACIONES

BRCA: Cáncer de mama según TCGA (*Breast Invasive Carcinoma*)

CGI: Islas CpG

CNV: Variación en el número de copias (*Copy-Number Variation*)

COAD: Cáncer de colon según TCGA (*Colon Adenocarcinoma*)

DEA: Análisis de expresión diferencial (*Differential Expression Analysis*)

DMR: Regiones diferencialmente metiladas (*Differential Methylated Regions*)

DNMTs: Enzimas ADN-metiltransferasas

ENCODE: Enciclopedia de los Elementos del ADN (*Encyclopedia of DNA Elements*)

FDR: Número de falsos positivos (*False Discovery Rate*)

FPKM: Fragmentos esperados por Kilobase de transcripción por Millón de fragmentos secuenciados (*Fragments Per Kilobase Million*)

GDC: *Genomic Data Commons*

GEGs: Genes diferencialmente expresados (*Differential Expression Genes*)

HAT: Histonas deacetilasas

HCC: Carcinoma hepatocelular

ICGC: Consorcio Internacional de los Genomas del Cáncer (*International Cancer Genome Consortium*)

LUAD: Cáncer de pulmón según TCGA (*Lung Adenocarcinoma*)

MDS: Análisis de proximidades (*Multi-Dimensional Scaling*)

MDS: Síndrome mielodisplásico

ML: Aprendizaje automático (*Machine Learning*)

NA: Valores faltantes

NGS: Secuenciación masiva (*Next-Generation Sequencing*)

PRAD: Cáncer de próstata según TCGA (*Prostate Adenocarcinoma*)

PTMs: Modificaciones post-traduccionales (*Post-Traductional Modification*)

RF: Bosques aleatorios (*Random Forest*)

RRBS: Secuenciación de bisulfito de representación reducida (*Reduced-Representation Bisulfite Sequencing*)

ScBS: Secuenciación de bisulfito de una sola célula (single-cell reduced-representation Bisulfite Sequencing)

SVM: Máquinas de soporte vectorial (*Support Vector Machines*)

TARGET*: Therapeutically Applicable Research to Generate Effective Treatments*

TCGA: Atlas del Genoma del Cáncer (*The Cancer Genome Atlas*)

WGBS: Secuenciación de bisulfito de genoma completo (*Whole-Genome Bisulfite Sequencing*)

6. GLOSARIO

- **Acetilación**: Adición de un grupo acetilo desde una molécula de acetil-CoA a un resido de lisina presente en las histonas. Este proceso está catalizado por la enzima histona acetiltransferasas (HAT).
- **Adenocarcinoma**: Cáncer que empieza en las células glandulares (secretoras). Las células glandulares se encuentran en el tejido que reviste ciertos órganos internos; producen y liberan sustancias en el cuerpo, como el moco, los jugos digestivos u otros líquidos (definición obtenida por el Instituto Nacional del Cáncer, NIH).
- **ADP ribosilación**: Trasferencia enzimática de una molécula de una ADP-ribosa a una proteína aceptora a partir de un dinucleótido de nicotinamida y adenina (NAD^+ en su forma oxidada). Esta reacción es catalizada por la ADP-ribosiltransferasas.
- **Barcodes de TCGA**: El código de barras TCGA es el identificador principal de los datos de muestras biológicas dentro del proyecto TCGA, y contiene información sobre el tipo de muestra (tumoral o normal), el tipo de tejido, el centro de estudio, características asociadas a su identificación analítica (vial, porción, etc) y el nombre del proyecto (Para más información acceda a: TCGA Barcode).
- **Citrulinación**: Conversión del residuo arginina a citrulina proceso catalizado por la enzima peptidil arginina deiminasa (PAD).
- **Fosforilación**: Adición de un grupo fosfato procedente del ATP al grupo de oxhidrilo de una cadena lateral del aminoácido del objetivo.
- **Isomerización**: Transformación de una molécula en un isómero diferente, las dos conformaciones posibles son cis y trans.
- **Oncogénesis (o carcinogénesis)**: Proceso por el cual una célula normal se convierte en una célula cancerosa.
- **Sumoylation**: Adición covalente de un miembro de la familia SUMO (*small ubiquitin-like modifier*) a un residuo de lisina de una proteína diana (Wilkinson & Henley, 2010).
- **Ubiquitinación**: Adición de una o varias moléculas de ubiquitina, una proteína pequeña, de manera covalente a proteínas diana en sus residuos de lisina.

7. BIBLIOGRAFÍA

[1] Antequera, F., Boyes, J., and Bird, A. (1990). High levels of de novo methylation and altered chromatin structure at CpG islands in cell lines. *Cell*, *62*(3), 503-514.

[2] Arrowsmith C.H., Bountra C., Fish P.V., Lee K. and Schapira M. (2012). Epigenetic protein families: a new frontier for drug discovery. Nature reviews *Drug discovery*, (5), 384

[3] Barbacid, M. (1987). Ras genes. *Annual review of biochemistry*, *56*(1), 779-827.).

[4] Baylin, S. B., and Herman, J. G. (2000). DNA hypermethylation in tumorigenesis: epigenetics joins genetics. *Trends in genetics*, *16*(4), 168-174.

[5] Baylin, S. B., and Jones, P. A. (2011). A decade of exploring the cancer epigenome—biological and translational implications. *Nature Reviews Cancer*, 11(10), 726.

[6] Baylin, S. B., and Ohm, J. E. (2006). Epigenetic gene silencing in cancer–a mechanism for early oncogenic pathway addiction?. *Nature Reviews Cancer*, *6*(2), 107.

[7] Baylin, S. B., Herman, J. G., Graff, J. R., Vertino, P. M., and Issa, J. P. (1997). Alterations in DNA methylation: a fundamental aspect of neoplasia. In *Advances in cancer research* (Vol. 72, pp. 141-196). Academic Press.

[8] Baylln, S. B., Herman, J. G., Graff, J. R., Vertino, P. M., and Issa, J. P. (1997). Alterations in DNA methylation: a fundamental aspect of neoplasia. In *Advances in cancer research* (Vol. 72, pp. 141-196). Academic Press.

[9] Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, *57*(1), 289-300.

[10] Bolden, J. E., Peart, M. J., and Johnstone, R. W. (2006). Anticancer activities of histone deacetylase inhibitors. *Nature reviews Drug discovery*, *5*(9), 769.

[11] Brecqueville M., Cervera N., Gelsi-Boyer V., Murati A. Adelaide J., Chaffanet M., Rey J., Vey N., Mozziconacci M.J. and Birnbaum D. (2011). Rare mutations in DNMT3A in myeloproliferative neoplasms and myelodysplastic syndromes. *Blood Cancer Journal*, *1*(5), e18.

[12] Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

[13] Bullard, J. H., Purdom, E., Hansen, K. D., and Dudoit, S. (2010). Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC bioinformatics, 11*(1), 94.

[14] Cairns, P., Esteller, M., Herman, J. G., Schoenberg, M., Jeronimo, C., Sanchez-Cespedes, M., et al. (2001). Molecular detection of prostate cancer in urine by GSTP1 hypermethylation. *Clinical Cancer Research*, *7*(9), 2727-2730.

[15] Cancer Genome Atlas Network. (2012). Comprehensive molecular portraits of human breast tumours. *Nature*, *490*(7418), 61.

[16] Costello, J. F., Frühwald, M. C., Smiraglia, D. J., Rush, L. J., Robertson, G. P., Gao, X., et al. (2000). Aberrant CpG-island methylation has non-random and tumour-type–specific patterns. *Nature genetics*, *24*(2), 132.

[17] Dedeurwaerder, S., Defrance, M., Bizet, M., Calonne, E., Bontempi, G., and Fuks, F. (2013). A comprehensive overview of Infinium HumanMethylation450 data processing. *Briefings in bioinformatics*, *15*(6), 929-941.

[18] Delgado, S., Gómez, M., Bird, A., and Antequera, F. (1998). Initiation of DNA replication at CpG islands in mammalian chromosomes. *The EMBO journal*, *17*(8), 2426-2435.

[19]Díaz-Uriarte, R., and De Andres, S. A. (2006). Gene selection and classification of microarray data using random forest. *BMC bioinformatics*, *7*(1), 3.

[20]Du P, Zhang X, Huang CC, et al. (2010) Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics*, *11*(1): 587.

[21]Egger, G., Liang, G., Aparicio, A., and Jones, P. A. (2004). Epigenetics in human disease and prospects for epigenetic therapy. *Nature*, *429*(6990), 457.

[22]Esteller, M. (2005). Aberrant DNA methylation as a cancer-inducing mechanism. *Annual Review of Pharmacology and Toxicology, 45*, 629-656.

[23]Esteller, M. (2007). Cancer epigenomics: DNA methylomes and histone-modification maps. *Nature reviews genetics*, *8*(4), 286.

[24]Esteller, M., Corn, P. G., Urena, J. M., Gabrielson, E., Baylin, S. B., and Herman, J. G. (1998). Inactivation of glutathione S-transferase P1 gene by promoter hypermethylation in human neoplasia. *Cancer research*, *58*(20), 4515-4518.

[25]Esteller, M., Silva, J. M., Dominguez, G., Bonilla, F., Matias-Guiu, X., Lerma, E., et al. (2000). Promoter hypermethylation and BRCA1 inactivation in sporadic breast and ovarian tumors. *JNCI: Journal of the National Cancer Institute*, *92*(7), 564-569.

[26]Feber, A., Guilhamon, P., Lechner, M., Fenton, T., Wilson, G. A., Thirlwell, C., et al. (2014). Using high-density DNA methylation arrays to profile copy number alterations. *Genome biology*, *15*(2), R30.

[27]Feinberg, A. P., and Tycko, B. (2004). The history of cancer epigenetics. *Nature Reviews Cancer*, **4**(2), 143.

[28]Feinberg, A. P., Koldobskiy, M. A., and Göndör, A. (2016). Epigenetic modulators, modifiers and mediators in cancer aetiology and progression. *Nature Reviews Genetics*, *17*(5), 284.

[29]Feinberg, A. P., Ohlsson, R., and Henikoff, S. (2006). The epigenetic progenitor origin of human cancer. *Nature reviews genetics*, *7*(1), 21.

[30] Florean, C., Schnekenburger, M., Grandjenette, C., Dicato, M. and Diederich, M. (2011). Epigenomics of leukemia: from mechanisms to therapeutic applications. *Epigenomics*, *3*(5), 581–609.

[31]Genovese, G., Kähler, A. K., Handsaker, R. E., Lindberg, J., Rose, S. A., Bakhoum, S. F., et al. (2014). Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. *New England Journal of Medicine*, *371*(26), 2477-2487.

[32]Grady, W. M., and Carethers, J. M. (2008). Genomic and epigenetic instability in colorectal cancer pathogenesis. *Gastroenterology*, *135*(4), 1079-1099.

[33]Graff, J. R., Herman, J. G., Lapidus, R. G., Chopra, H., Xu, R., Jarrard, D. F., et al. (1995). E-cadherin expression is silenced by DNA hypermethylation in human breast and prostate carcinomas. *Cancer research*, *55*(22), 5195-5199.

[34]Greenblatt, M. S., Bennett, W. P., Hollstein, M., and Harris, C. C. (1994). Mutations in the p53 tumor suppressor gene: clues to cancer etiology and molecular pathogenesis. *Cancer research*, *54*(18), 4855-4878.

[35]Hanahan, D., and Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *Cell*, *144*(5), 646-674.

[36]Hao, X., Luo, H., Krawczyk, M., Wei, W., Wang, W., Wang, J., et al. (2017). DNA methylation markers for diagnosis and prognosis of common cancers. *Proceedings of the National Academy of Sciences*, *114*(28), 7414-7419.

[37]Harris, C. C. (1996). Structure and function of the p53 tumor suppressor gene: clues for rational cancer therapeutic strategies. *JNCI: Journal of the National Cancer Institute*, *88*(20), 1442-1455.

[38]Herman, J. G., and Baylin, S. B. (2003). Gene silencing in cancer in association with promoter hypermethylation. *New England Journal of Medicine*, *349*(21), 2042-2054.

[39]Herman, J. G., Latif, F., Weng, Y., Lerman, M. I., Zbar, B., Liu, S., et al. (1994). Silencing of the VHL tumor-suppressor gene by DNA methylation in renal carcinoma. *Proceedings of the National Academy of Sciences*, *91*(21), 9700-9704.

[40] Holcomb, I. N., Young, J. M., Coleman, I. M., Salari, K., Grove, D. I., Hsu, et al. (2009). Comparative analyses of chromosome alterations in soft-tissue metastases within and across patients with castration-resistant prostate cancer. *Cancer research*, *69*(19), 7793-7802.

[41]Irizarry, R. A., Ladd-Acosta, C., Wen, B., Wu, Z., Montano, C., Onyango, P., et al. (2009). The human colon cancer methylome shows similar hypo-and hypermethylation at conserved tissue-specific CpG island shores. *Nature genetics*, *41*(2), 178.

[42]Jaiswal, S., Fontanillas, P., Flannick, J., Manning, A., Grauman, P. V., Mar, B. G., ... and Higgins, J. M. (2014). Age-related clonal hematopoiesis associated with adverse outcomes. *New England Journal of Medicine*, *371*(26), 2488-2498.

[43]Jones, P. A. (2012). Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature Reviews Genetics*, *13*(7), 484.

[44]Jones, P. A., and Baylin, S. B. (2007). The epigenomics of cancer. *Cell*, *128*(4), 683-692.

[45]Jones, P. A., and Laird, P. W. (1999). Cancer-epigenetics comes of age. *Nature genetics*, *21*(2), 163.

[46]Jones, P. L., Veenstra, G. C. J., Wade, P. A., Vermaak, D., Kass, S. U., Landsberger, N., et al. (1998). Methylated DNA and MeCP2 recruit histone deacetylase to repress transcription. *Nature genetics*, *19*(2), 187.

[47]Juergens, R. A., Wrangle, J., Vendetti, F. P., Murphy, S. C., Zhao, M., Coleman, B., et al. (2011). Combination epigenetic therapy has efficacy in patients with refractory advanced non–small cell lung cancer. *Cancer discovery*, *1*(7), 598-607.

[48]Karpf, A. R., Peterson, P. W., Rawlins, J. T., Dalley, B. K., Yang, Q., Albertsen, H., and Jones, D. A. (1999). Inhibition of DNA methyltransferase stimulates the expression of signal transducer and activator of transcription 1, 2, and 3 genes in colon tumor cells. *Proceedings of the National Academy of Sciences*, *96*(24), 14007-14012.

[49]Kulis, M. et al. (2010) DNA methylation and cancer. *Advances in genetics*, *70*, 27-56.

[50]Laird, P. W. (2010). Principles and challenges of genome-wide DNA methylation analysis. *Nature Reviews Genetics*, *11*(3), 191.

[51]Laird, P. W., and Jaenisch, R. (1996). The role of DNA methylation in cancer genetics and epigenetics. *Annual review of genetics*, *30*(1), 441-464.

[52]Lantz, Brett. 2015. Machine learning with R. Packt Publishing Ltd.

[53]Lengauer, C., Kinzler, K. W., and Vogelstein, B. (1997). DNA methylation and genetic instability in colorectal cancer cells. *Proceedings of the National Academy of Sciences*, *94*(6), 2545-2550.

[54]Ley, T.J., Ding, L., Walter, M.J., McLellan, M.D., Lamprecht, T., Larson, D.E., Kandoth, C., Payton, J.E., Baty, J., Welch, J., et al. (2010). DNMT3A mutations in acute myeloid leukemia. *New England Journal of Medicine*, *363*(25), 2424-2433.

[55]Liang, G., Gonzales, F. A., Jones, P. A., Orntoft, T. F., and Thykjaer, T. (2002). Analysis of gene induction in human fibroblasts and bladder cancer cells exposed to the methylation inhibitor 5-aza-2'-deoxycytidine. *Cancer research*, *62*(4), 961-966.

[56]Lomvardas, S., and Thanos, D. (2002). Modifying gene expression programs by altering core promoter chromatin architecture. *Cell*, *110*(2), 261-271.

[57]MacLeod, A. R., and Szyf, M. (1995). Expression of antisense to DNA methyltransferase mRNA induces DNA demethylation and inhibits tumorigenesis. *Journal of Biological Chemistry*, *270*(14), 8037-8043.

[58]Marabita, F., Almgren, M., Lindholm, M. E., Ruhrmann, S., Fagerström-Billai, F., Jagodic, M., et al. (2013). An evaluation of analysis pipelines for DNA methylation profiling using the Illumina HumanMethylation450 BeadChip platform. *Epigenetics*, *8*(3), 333-346.

[59]Martin, C., and Zhang, Y. (2007). Mechanisms of epigenetic inheritance. *Current opinion in cell biology*, *19*(3), 266-272.

[60]Masser, D. R., Hadad, N., Porter, H., Stout, M. B., Unnikrishnan, A., Stanford, D. R., and Freeman, W. M. (2018). Analysis of DNA modifications in aging research. *Geroscience*, *40*(1), 11-29.

[61]Methylation of histone H4 lysine 20 controls recruitment of Crb2 to sites of DNA damage. *Cell*, *119*(5), 603-614.

[62]Millar, D. S., Paul, C. L., Molloy, P. L., and Clark, S. J. (2000). A Distinct Sequence (ATAAA) n Separates Methylated and Unmethylated Domains at the 5'-End of theGSTP1 CpG Island. *Journal of Biological Chemistry*, *275*(32), 24893-24899.

[63]Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods*, *5*(7), 621.

[64]Nan, X., Ng, H. H., Johnson, C. A., Laherty, C. D., Turner, B. M., Eisenman, R. N., and Bird, A. (1998). Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex. *Nature*, *393*(6683), 386.

[65]Okano, M., Bell, D. W., Haber, D. A., and Li, E. (1999). DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell*, *99*(3), 247-257.

[66]Plass, C., Pfister, S. M., Lindroth, A. M., Bogatyrova, O., Claus, R., and Lichter, P. (2013). Mutations in regulators of the epigenome and their connections to global chromatin patterns in cancer. *Nature reviews genetics*, *14*(11), 765.

[67]Rhee, I., Jair, K. W., Yen, R. W. C., Lengauer, C., Herman, J. G., Kinzler, K. W., ... and Schuebel, K. E. (2000). CpG methylation is maintained in human cancer cells lacking DNMT1. *Nature*, *404*(6781), 1003.

[68]Riggs, A. D., and Jones, P. A. (1983). 5-methylcytosine, gene regulation, and cancer. In *Advances in cancer research* (Vol. 40, pp. 1-30). Academic Press.

[69]Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., and Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic acids research*, *43*(7), e47-e47.

[70]Ruthenburg, A. J., Li, H., Patel, D. J., and Allis, C. D. (2007). Multivalent engagement of chromatin modifications by linked binding modules. *Nature reviews Molecular cell biology*, *8*(12), 983.

[71]Sandoval, J., and Esteller, M. (2012). Cancer epigenomics: beyond genomics. *Current opinion in genetics and development*, *22*(1), 50-55.

[72]Statnikov, A., Aliferis, C. F., Tsamardinos, I., Hardin, D., and Levy, S. (2004). A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. *Bioinformatics*, *21*(5), 631-643.

[73]Stevens, M., Cheng, J. B., Li, D., Xie, M., Hong, C., Maire, C. L., et al. (2013). Estimating absolute methylation levels at single-CpG resolution from methylation enrichment and restriction enzyme sequencing methods. *Genome research*, *23*(9), 1541-1553.

[74]Toyota, M., Ahuja, N., Ohe-Toyota, M., Herman, J. G., Baylin, S. B., and Issa, J. P. J. (1999). CpG island methylator phenotype in colorectal cancer. *Proceedings of the National Academy of Sciences*, *96*(15), 8681-8686.

[75]Tran, H. T. T., Kim, H. N., Lee, I. K., Kim, Y. K., Ahn, J. S., Yang, D. H., ... and Kim, H. J. (2011). DNA methylation changes following 5-azacitidine treatment in patients with myelodysplastic syndrome. *Journal of Korean medical science*, *26*(2), 207-213.

[76]Wang, D., Yan, L., Hu, Q., Sucheston, L. E., Higgins, M. J., Ambrosone, C. B., et al. (2012). IMA: an R package for high-throughput analysis of Illumina's 450K Infinium methylation data. *Bioinformatics*, *28*(5), 729-730.

[77]Wang, R. Y. H., Gehrke, C. W., and Ehrlich, M. (1980). Comparison of bisulfite modification of 5-methyldeoxycytidine and deoxycytidine residues. *Nucleic Acids Research*, *8*(20), 4777-4790.

[78]Wang, Z., Wu, X., and Wang, Y. (2018). A framework for analyzing DNA methylation data from Illumina Infinium HumanMethylation450 BeadChip. *BMC bioinformatics*, *19*(5), 115.

[79]Weinstein, I. B. (2002). Addiction to oncogenes--the Achilles heal of cancer. *Science*, *297*(5578), 63-64.

[80]Wilkinson, K. A., and Henley, J. M. (2010). Mechanisms, regulation and consequences of protein SUMOylation. *Biochemical Journal*, *428*(2), 133-145.

[81]Woo, H. G., Choi, J. H., Yoon, S., Jee, B. A., Cho, E. J., Lee, J. H., et al. (2017). Integrative analysis of genomic and epigenomic regulation of the transcriptome in liver cancer. *Nature communications*, *8*(1), 839.

[82]Wood, L. D., Parsons, D. W., Jones, S., Lin, J., Sjöblom, T., Leary, R. J., et al. (2007). The genomic landscapes of human breast and colorectal cancers. *Science*, 318(5853), 1108-1113.

[83]Wu, H., and Zhang, Y. (2011). Mechanisms and functions of Tet protein-mediated 5-methylcytosine oxidation. *Genes and development*, *25*(23), 2436-2452.

[84]You, J. S., and Jones, P. A. (2012). Cancer genetics and epigenetics: two sides of the same coin?. *Cancer cell*, *22*(1), 9-20.

Webs empleadas en la bibliografía

- ABC.RAP vignettes: https://cran.r-project.org/web/packages/ABC.RAP/vignettes/ABC.RAP.html
- Bread GDAC Firehose: http://gdac.broadinstitute.org/
- cBioPorta FAQ : https://www.cbioportal.org/faq#what-is-gistic-what-is-rae
- Definición de adenocarcinoma: https://www.cancer.gov/espanol/publicaciones/diccionario/def/adenocarcinoma
- Definición de epigenómica: https://www.cancer.gov/espanol/publicaciones/diccionario/def/epigenomica
- Fviz_nclust RDocumentation: https://www.rdocumentation.org/packages/factoextra/versions/1.0.5/topics/fviz_nbclust
- GDC Data Portal: https://portal.gdc.cancer.gov/
- Kmeans RDocumentation: https://www.rdocumentation.org/packages/stats/versions/3.6.0/topics/kmeans
- Normalización FPKM: https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/Expression_mRNA_Pipeline/#fpkm
- OMS: https://www.who.int/es/news-room/fact-sheets/detail/cancer
- RefSeq: https://www.ncbi.nlm.nih.gov/refseq/
- TCGA Barcode: https://docs.gdc.cancer.gov/Encyclopedia/pages/TCGA_Barcode/
- TCGA: https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga
- TCGABiolinks Help Documents: https://www.bioconductor.org/packages/devel/bioc/vignettes/TCGAbiolinks/inst/doc/analysis.html#tcgaanalyze_dmr:_differentially_methylated_regions_analysis

8. ANEXO 1: GENES CANDIDATOS

GEN	FUNCIÓN	COHORTE	GEN	FUNCIÓN	COHORTE
DNMT1	ADN metiltransferasa	COAD,BRCA,LUAD	MGMT	ADN metiltransferasa	
DNMT3A	ADN metiltransferasa		KRAS	Proto-oncogen	
DNMT3B	ADN metiltransferasa	LUAD,BRCA	APC	Regulador de la vía de señalización WNT	COAD
SMARCA2	Relacionado con las proteínas SWI/SNF		TP53	Codifica una proteína supresora de tumores	BRCA
SMARCA4	Relacionado con las proteínas SWI/SNF	LUAD	STAT1	Acetil transferasa	
SMARCD3	Relacionado con las proteínas SWI/SNF		STAT3	Acetil transferasa	
ARID1A	Subunidad de BAF	COAD,LUAD	YAP1	Regulador transcripcional	
ARD1B	Subunidad de BAF	BRCA	CTCF	Factor de unión CCCTC	
ARID2	Subunidad de PBAF		BRCA1	Asociado a reparadores del ADN	BRCA
CHD5	Helicasa		BRCA2	Asociado a reparadores del ADN	
CHD1	Helicasa		VHL	Codifica una proteína supresora de tumores	COAD
CHD3	Helicasa	COAD,PRAD,LUAD	BRAF	Proto-oncogen, quinasa serina/treonina	
CHD4	Helicasa	COAD,PRAD,LUAD	EHMT1	Histona lisina metiltransferasa	
CHD6	Helicasa	COAD,PRAD,LUAD	EHMT2	Histona lisina metiltransferasa	
CHD7	Helicasa	COAD,PRAD,LUAD	CBX5	Cromobox 5	
CHD8	Helicasa	COAD,PRAD,LUAD	UHRF1	Ubiquitina ligasa	
TET1	5’metilcitosina hidrolasa		OCT4	Transportador de catión/carnitina	
TET2	5’metilcitosina hidrolasa		SOX2	SRY-box 2	
TET3	5’metilcitosina hidrolasa		NANOG	Factor de transcripción NANOG	
MBD1	Dominio de unión a metil-CpG	COAD	KLF4	Factor de transcripción del dedo de zinc	COAD, BRCA
MDB4	Dominio de unión a metil-CpG	COAD,LUAD,BRCA	GSTP1	Glutatión S-transferasa	COAD
HDAC2	Histona deacetilasa	COAD	MLH1	Homólogo 1 MutL	COAD
HDAC4	Histona deacetilasa	BRCA	GAT4		COAD
HDAC9	Histona deacetilasa	PRAD	GAT5		BRCA, LUAD
MLL	Histona metiltransferasa	COAD,LUAD,BRCA	SRBC		COAD
MLL3	Histona metiltransferasa	LUAD	WRN	Helicasa	
MLL4	Histona metiltransferasa	LUAD	RIZ1	Codifica una proteína supresora de tumores	
SETD1A	Histona lisina metiltransferasa	BRCA	BRD4	Dominio-bromo	BRCA,COAD
EZH2	Histona metiltransferasa	BRCA,PRAD, COAD,LUAD	CDK2N2A	Quinasa dependiente de ciclina	
NSD1	Receptor nuclear de unión a SET	COAD	MYC	Proto-oncogen	

GEN	FUNCIÓN	COHORTE	GEN	FUNCIÓN	COHORTE
NSD2	Receptor nuclear de unión a SET	COAD	SETP9	Proteína sin función clara	COAD
SETD2	Histona lisina metiltransferasa		BCL2	Regulador de apoptosis	
SETD3	Hisitidina metiltransferasa		TFAP2E	Factor de transcripción AP-2 α	
SETD7	Histona lisina metiltransferasa		TTF2	Factor de transcripción	
SETMAR	Dominio SET		PRMT6	Arginina metiltransferasa	BRCA
EP300	Codifica una proteína de unión p300		PRMT8	Arginina metiltransferasa	BRCA
ZBTB38	Factor de transcripción del dedo de zinc		H3F3A	Familia de las histonas 3A	
ZMYM4	Dedo zinc tipo MYM		H3F3B	Familia de las histonas 3B	
ZMYM6	Dedo zinc tipo MYM		HIST1H3B	Familia de las histonas miembro b	
ZMYM8	Dedo zinc tipo MYM		HIST1H1B	Familia de las histonas miembro b	
IDH1	Isocitrato deshidrogenasa		NSD1	Receptor nuclear de unión SET	
IDH2	Isocitrato deshidrogenasa		APOBEC1	Enzima de edición de ARNm	
CDKN2A	Quinasa dependiente de ciclina		PRDM1	Dominio SET	
PRDM2	Dominio SET		KDM2A	Lisina demetilasa	
PRDM4	Dominio SET		KDM2B	Lisina demetilasa	
PRDM6	Dominio SET		KDM3A	Lisina demetilasa	
PRDM7	Dominio SET		KDM3B	Lisina demetilasa	
PRDM8	Dominio SET		KDM4B	Lisina demetilasa	
PRDM11	Dominio SET		KDM6A	Lisina demetilasa	
PRDM12	Dominio SET		KDM6B	Lisina demetilasa	
PRDM13	Dominio SET		SMYD2	Dominio SET y MYND	
PRDM14	Dominio SET		SMYD3	Dominio SET y MYND	
PRDM15	Dominio SET		AID	5'citidina deaminasa	
PRDM16	Dominio SET		ASXL	Potenciador del grupo trithorax y polycomb	
SETBP1	Codifica una proteína de unión SET		PRMT1	Arginina metiltransferasa	
WHSC1	Histona lisina metiltransferasa		PRMT5	Arginina metiltransferasa	
WHSC1L1	Histona lisina metiltransferasa		LSD1	Dedo zinc	
HDAC2	Histona deacetiltransferasa		UTX	Histona demetilasa	
HDAC5	Histona deacetiltransferasa	BRCA,COAD,PRAD	BRG1	ATPasa de BAF	LUAD
HDAC7A	Histona deacetiltransferasa	BRCA,COAD,PRAD	SIRT1	Histona deacetiltransferasa	BRCA,COAD, PRAD
BRGM	ATPasa de BAF	PRAD			

Tabla 1: Listado de genes seleccionados por su relación con los mecanismos epigenéticos, cuyas mutaciones y CNVs han sido aisladas para su estudio. En el caso de que se haya establecido una relación entre las alteraciones en dicho gen y algunas cohortes de estudio aparece indicado en la columna "Cohorte" (You y Jones, 2012; Plass et al., 2013; Feinberg et al., 2016)

9. ANEXO 2: FIGURAS ADICIONALES

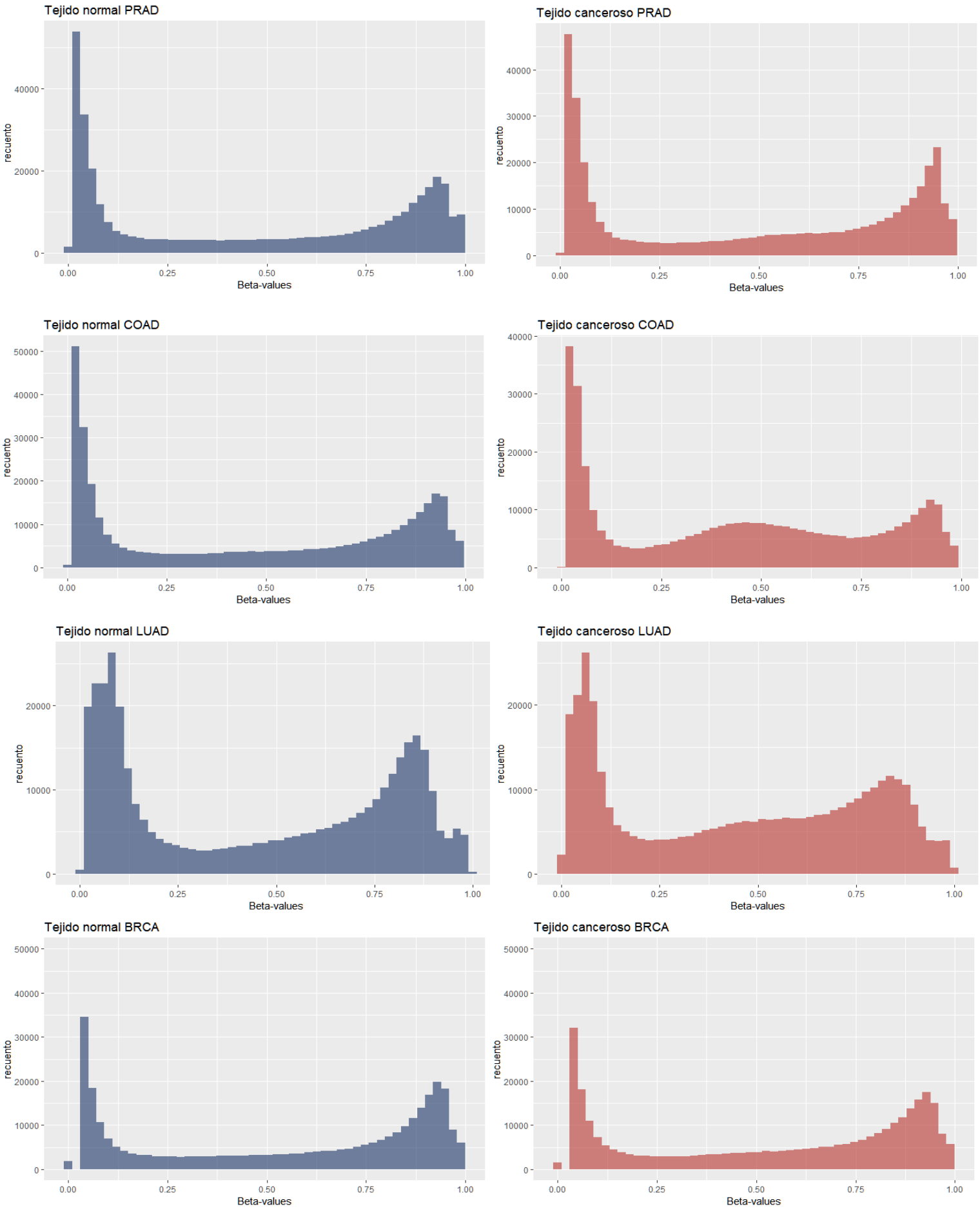


Figura 1: Frecuencias de los valores beta en las muestras normales (azul) y tumorales (rojo) para cada una de las cohortes de estudio, de arriba abajo: PRAD, COAD, LUAD y BRCA.

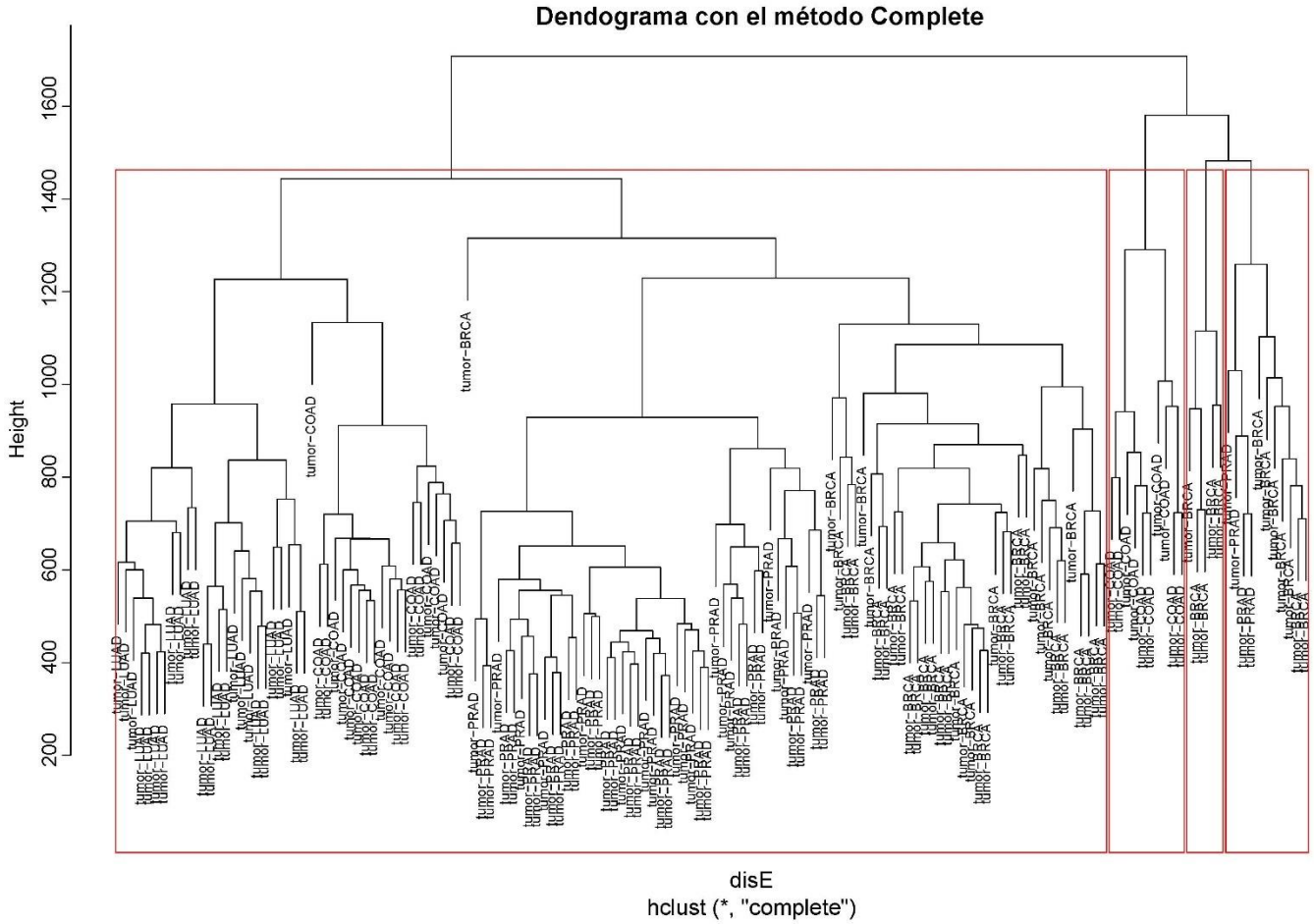


Figura 2: Dendograma empleando el método *complete*. En el dendograma se han añadido 4 clústeres que coinciden con cada una de las cohortes de estudio.

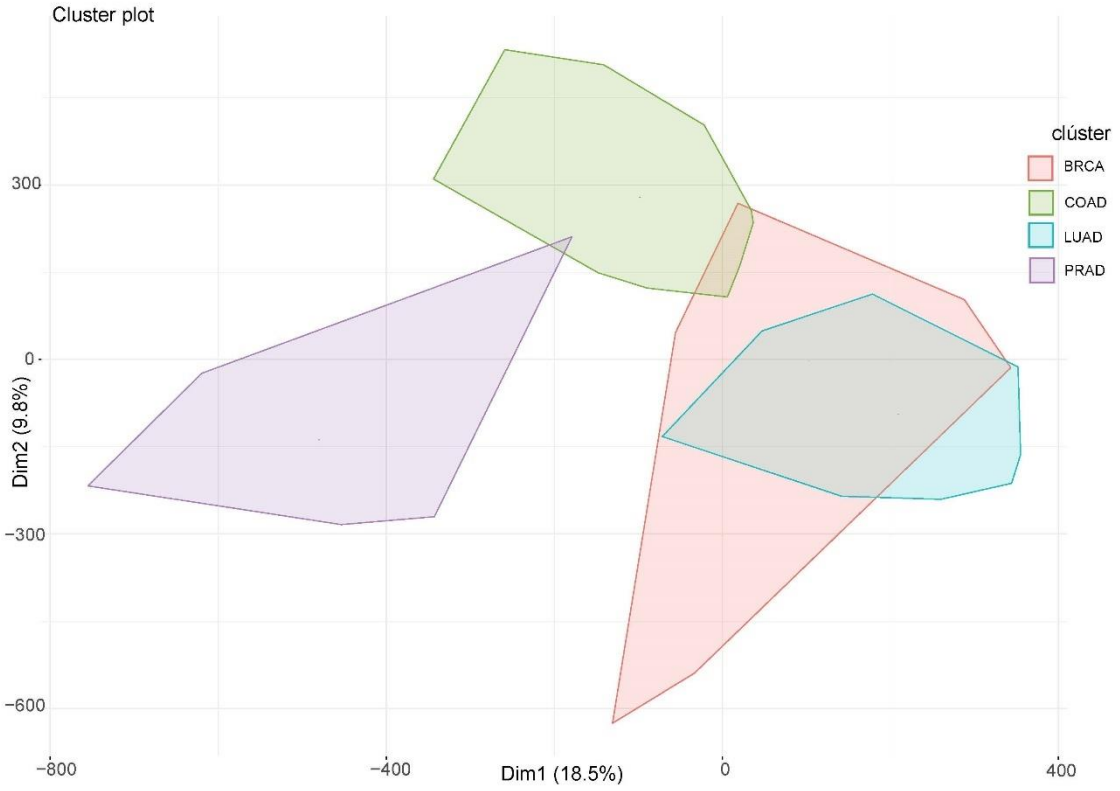


Figura 3: Silueta de las muestras agrupadas en cada una de las cohortes empleando el método de agrupación no supervisado k-means. En rosa se representa las muestras clasificadas como BRCA, en verde como COAD, en azul como UAD y en morado como PRAD.

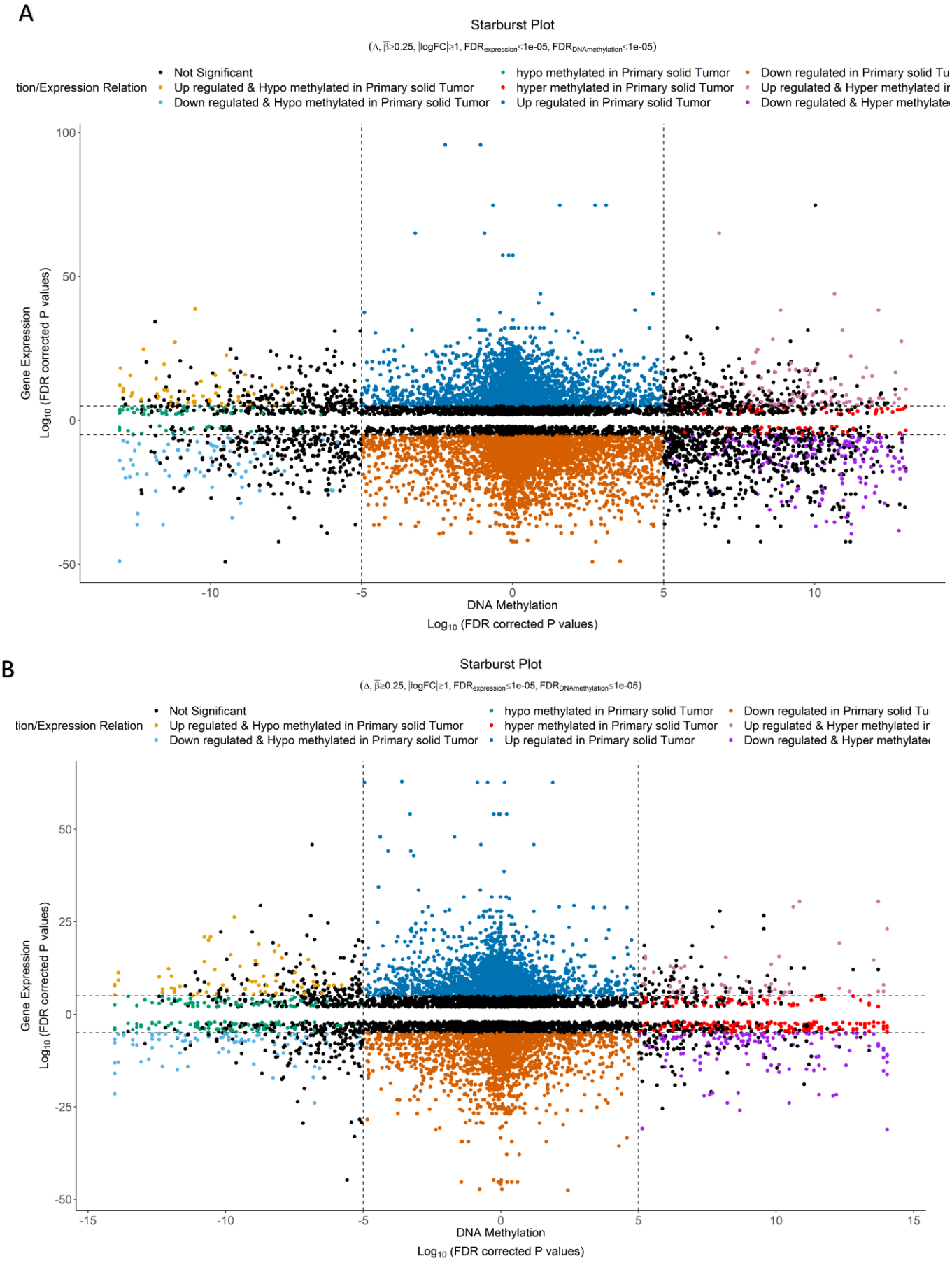
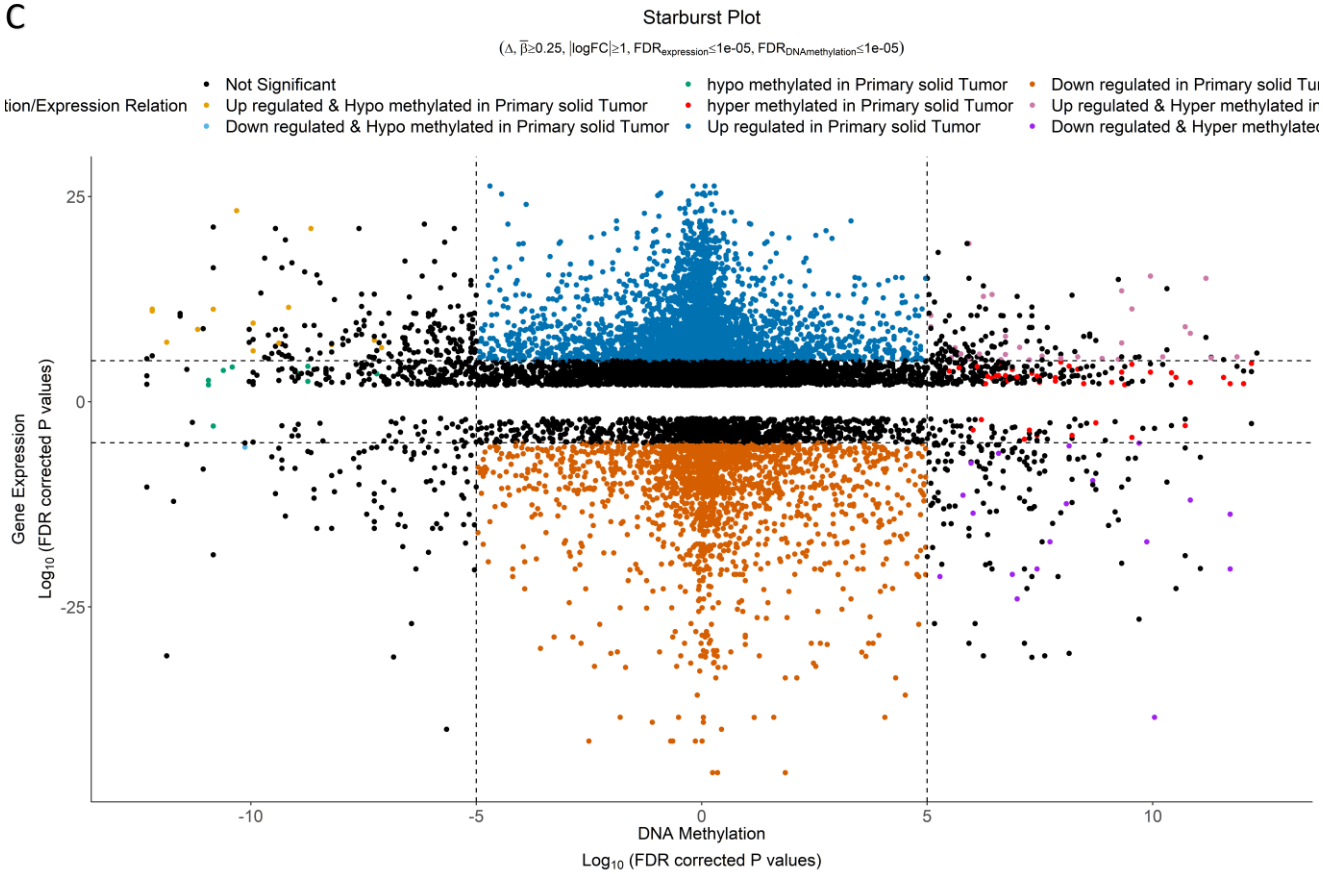
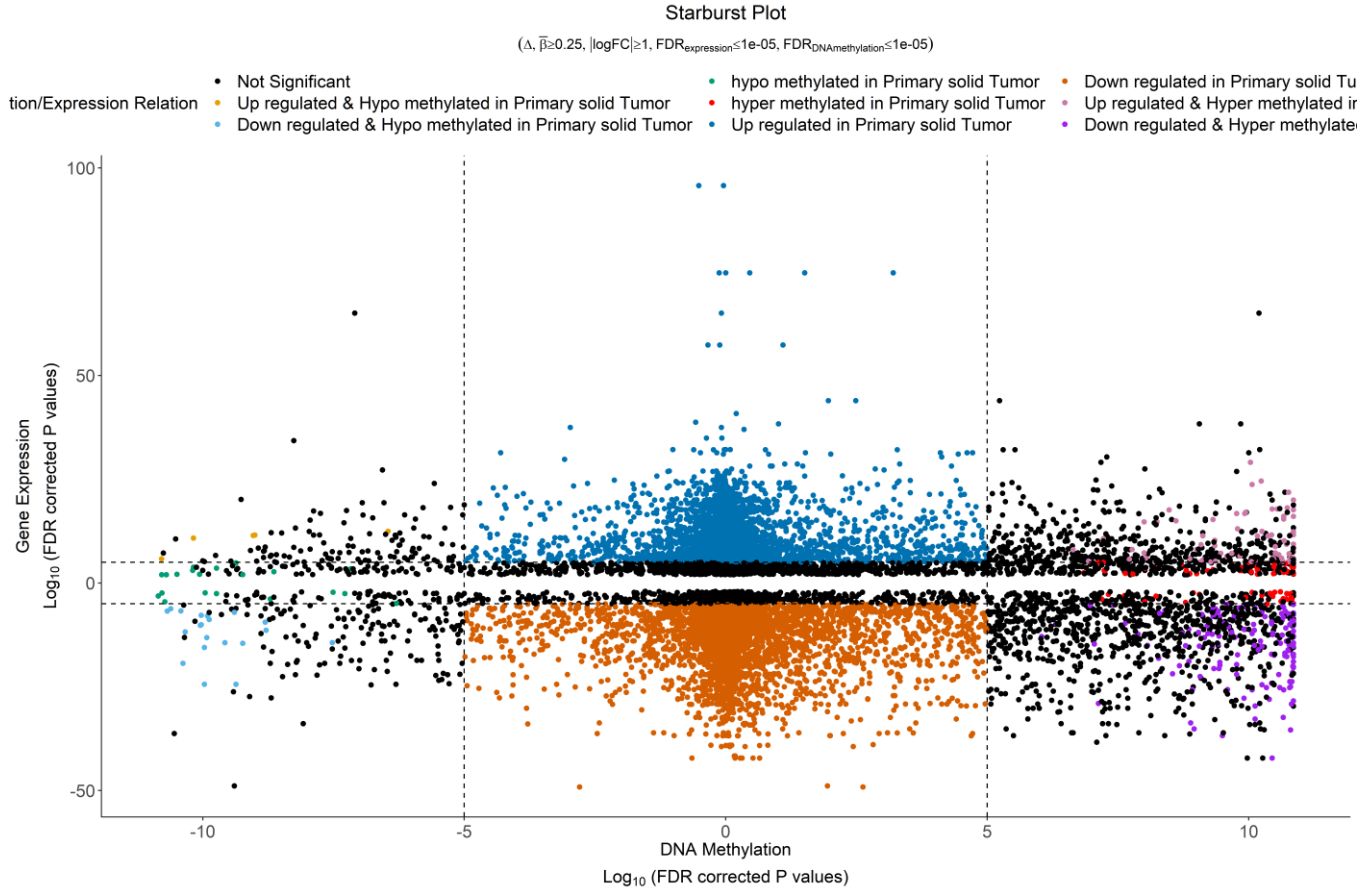


Figura 7: Starburst plot comparando los datos de metilación (eje X) y expresión génica (Y) de las cuatro cohortes de estudio: mama (A), colon (B), pulmón (C) y próstata (D); C y D se muestran en la página siguiente. En morado (derecha-abajo) se muestran los genes silenciados y que presenta regiones hipermetiladas en las muestras tumorales de manera significativa, en violeta (derecha-arriba) representa los genes activados y se corresponde con regiones hipermetiladas, en azul claro (izquierda-abajo) se muestran los genes silenciados correspondientes a regiones hipometiladas, y en amarillo (izquierda-arriba) aparecen los genes activados y se corresponde con regiones hipometilados. Los resultados se corresponden con aquellos que presentan una significación menor o igual a 10-5.

C


Starburst Plot
(Δ, β̄≥0.25, |logFC|≥1, FDR_expression≤1e-05, FDR_DNAmethylation≤1e-05)
tion/Expression Relation
Not Significant
Up regulated & Hypo methylated in Primary solid Tumor
Down regulated & Hypo methylated in Primary solid Tumor
hypo methylated in Primary solid Tumor
hyper methylated in Primary solid Tumor
Up regulated in Primary solid Tumor
Down regulated in Primary solid Tu
Up regulated & Hyper methylated in
Down regulated & Hyper methylated
25
0
-25
Gene Expression
Log10 (FDR corrected P values)
-10
-5
0
5
10
DNA Methylation
Log10 (FDR corrected P values)

D


Starburst Plot
(Δ, β̄≥0.25, |logFC|≥1, FDR_expression≤1e-05, FDR_DNAmethylation≤1e-05)
tion/Expression Relation
Not Significant
Up regulated & Hypo methylated in Primary solid Tumor
Down regulated & Hypo methylated in Primary solid Tumor
hypo methylated in Primary solid Tumor
hyper methylated in Primary solid Tumor
Up regulated in Primary solid Tumor
Down regulated in Primary solid Tu
Up regulated & Hyper methylated in
Down regulated & Hyper methylated
100
50
0
-50
Gene Expression
Log10 (FDR corrected P values)
-10
-5
0
5
10
DNA Methylation
Log10 (FDR corrected P values)

10. ANEXO 3: VERSIÓN DE R

```
> sessionInfo()
R version 3.5.1 (2018-07-02)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows 7 x64 (build 7600)

Matrix products: default

locale:
[1] LC_COLLATE=Spanish_Spain.1252  LC_CTYPE=Spanish_Spain.1252    LC_MONETARY=Spanish_Spain.1252
[4] LC_NUMERIC=C                   LC_TIME=Spanish_Spain.1252

attached base packages:
[1] parallel  stats4    stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
 [1] RaggedExperiment_1.6.0      TCGAbiolinks_2.10.5         curatedTCGAData_1.4.3
 [4] MultiAssayExperiment_1.8.3  SummarizedExperiment_1.12.0 DelayedArray_0.8.0
 [7] BiocParallel_1.16.6         matrixStats_0.54.0          Biobase_2.42.0
[10] GenomicRanges_1.34.0        GenomeInfoDb_1.18.2         randomForest_4.6-14
[13] caret_6.0-84                ggplot2_3.1.1               lattice_0.20-35
[16] kernlab_0.9-27              dplyr_0.8.1                 IRanges_2.16.0
[19] S4Vectors_0.20.1            BiocGenerics_0.28.0         knitr_1.22

loaded via a namespace (and not attached):
  [1] R.utils_2.8.0                tidyselect_0.2.5             RSQLite_2.1.1
  [4] AnnotationDbi_1.44.0         grid_3.5.1                   DESeq_1.34.1
  [7] munsell_0.5.0                codetools_0.2-15             preprocessCore_1.44.0
 [10] withr_2.1.2                  colorspace_1.4-1             highr_0.8
 [13] rstudioapi_0.10              GenomeInfoDbData_1.2.0       KMsurv_0.1-5
 [16] hwriter_1.3.2                bit64_0.9-7                  downloader_0.4
 [19] generics_0.0.2               ipred_0.9-9                  xfun_0.7
 [22] ggthemes_4.2.0               EDASeq_2.16.3                R6_2.4.0
 [25] doParallel_1.0.14            locfit_1.5-9.1               bitops_1.0-6
 [28] assertthat_0.2.1             promises_1.0.1               scales_1.0.0
 [31] nnet_7.3-12                  gtable_0.3.0                 sva_3.30.1
 [34] wheatmap_0.1.0               sesameData_1.0.0             timeDate_3043.102
 [37] rlang_0.3.4                  genefilter_1.64.0            cmprsk_2.2-7
```

```
 [40] GlobalOptions_0.1.0          splines_3.5.1                rtracklayer_1.42.2
 [43] lazyeval_0.2.2               ModelMetrics_1.2.2           selectr_0.4-1
 [46] broom_0.5.2                  BiocManager_1.30.4           yaml_2.2.0
 [49] reshape2_1.4.3               GenomicFeatures_1.34.8       backports_1.1.4
 [52] httpuv_1.5.1                 tools_3.5.1                  lava_1.6.5
 [55] RColorBrewer_1.1-2           DNAcopy_1.56.0               Rcpp_1.0.1
 [58] plyr_1.8.4                   progress_1.2.2               zlibbioc_1.28.0
 [61] purrr_0.3.2                  RCurl_1.95-4.12              prettyunits_1.0.2
 [64] ggpubr_0.2                   rpart_4.1-13                 GetoptLong_0.1.7
 [67] zoo_1.8-5                    ggrepel_0.8.1                cluster_2.0.7-1
 [70] magrittr_1.5                 data.table_1.12.2            circlize_0.4.6
 [73] survminer_0.4.3              aroma.light_3.12.0           hms_0.4.2
 [76] mime_0.6                     evaluate_0.13                xtable_1.8-4
 [79] XML_3.98-1.19                gridExtra_2.3                shape_1.4.4
 [82] compiler_3.5.1               biomaRt_2.38.0               tibble_2.1.1
 [85] crayon_1.3.4                 R.oo_1.22.0                  htmltools_0.3.6
 [88] mgcv_1.8-24                  later_0.8.0                  tidyr_0.8.3
 [91] geneplotter_1.60.0           lubridate_1.7.4              DBI_1.0.0
 [94] ExperimentHub_1.8.0          matlab_1.0.2                 ComplexHeatmap_1.20.0
 [97] MASS_7.3-50                  ShortRead_1.40.0             Matrix_1.2-14
[100] readr_1.3.1                  cli_1.1.0                    R.methodsS3_1.7.1
[103] gower_0.2.1                  km.ci_0.5-2                  pkgconfig_2.0.2
[106] sesame_1.0.0                 GenomicAlignments_1.18.1     recipes_0.1.5
[109] xml2_1.2.0                   foreach_1.4.4                annotate_1.60.1
[112] XVector_0.22.0               prodlim_2018.04.18           rvest_0.3.4
[115] stringr_1.4.0                digest_0.6.18                ConsensusClusterPlus_1.46.0
[118] Biostrings_2.50.2            rmarkdown_1.12               survMisc_0.5.5
[121] edgeR_3.24.3                 curl_3.3                     shiny_1.3.2
[124] Rsamtools_1.34.1             rjson_0.2.20                 nlme_3.1-137
[127] jsonlite_1.6                 limma_3.38.3                 pillar_1.4.0
[130] httr_1.4.0                   survival_2.42-3              interactiveDisplayBase_1.20.0
[133] glue_1.3.1                   iterators_1.0.10             bit_1.1-14
[136] class_7.3-14                 stringi_1.4.3                blob_1.1.1
[139] AnnotationHub_2.14.5         latticeExtra_0.6-28          memoise_1.1.0
[142] e1071_1.7-1
> RStudio.Version()

$version
[1] ‘1.2.1511’
```