Prediction model of response to a certain treatment in childhood obesity

Pol Castellano Escuder
Máster universitario en Bioinformática y bioestadística
Área del trabajo: Estadística y Bioinformática

Nombre Consultor/a: Esteban Vegas Lozano
Nombre Profesor/a externo: Josep Jimenez Chillarón

24 de Mayo de 2017

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Prediction model of response to a certain treatment in childhood obesity*
Nombre del autor:	*Pol Castellano Escuder*
Nombre del consultor/a:	*Esteban Vegas Lozano*
Nombre del PRA:	*Josep Jimenez Chillarón*
Fecha de entrega (mm/aaaa):	*05/2017*
Titulación:	*Máster universitario en Bioinformática y bioestadística*
Área del Trabajo Final:	*Estadística y Bioinformática*
Idioma del trabajo:	*Español*
Palabras clave	*Modelo, Predicción, Metilación*

Resumen del Trabajo (máximo 250 palabras): *Con la finalidad, contexto de aplicación, metodología, resultados i conclusiones del trabajo.*

En este estudio disponemos de una cohorte muy concreta de niños y niñas prepuberales, con obesidad severa. Todos los sujetos fueron sometidos a un tratamiento de seis meses donde solo la mitad, aproximadamente, respondió de la forma esperada. Entonces, creamos una clasificación de estos sujetos en función de si respondían o no al tratamiento.

Posteriormente, se analizó la metilación del genoma de cada uno de los individuos, encontrando diferencias epigenéticas significativas entre el patrón de los dos grupos.

Aquí proponemos un modelo de predicción basado en el panel epigenético de cada fenotipo (responden o no lo hacen) con finalidad diagnóstico en la práctica clínica. El modelo se ha realizado a partir de una regresión limma con validación externa "Leave One Out". Hemos creado tantos modelos de regresión como individuos tenemos en la muestra, dejando cada vez un individuo fuera del modelo. Finalmente, usamos los predictores intersección entre todos los modelos para crear nuestro modelo de predicción final.

La finalidad principal del modelo es poder determinar con antelación si un individuo va a responder o no al tratamiento, pudiendo usar otros métodos alternativos en caso de predicción negativa. De este modo podemos agilizar la mejora del individuo ahorrando tiempo al equipo médico y sobretodo, al sujeto

y a su familia.

Abstract (in English, 250 words or less):

Here we work with a very specific cohort of prepubertal children, with severe obesity. Subjects were under a six-month treatment where only about half responded as expected. Then, we created a classification of these subjects depending on whether or not they responded to treatment.

Subsequently, we analyzed genome methylation of each subject, finding significant epigenetic differences between two groups.

Here we propose a prediction model based on the epigenetic panel of each phenotype (high responder or low responder) for diagnostic purposes in clinical practice. The model was made from a limma regression with external validation "Leave One Out". We have created as many regression models as samples we have, leaving an individual out of the model each time. Finally, we use the overlap predictors between all models to create our final prediction model.

The purpose of the model is to be able to determine in advance if an individual is going to respond or not to the treatment, being able to use other alternative methods in case of negative prediction. So we can speed up the improvement of the individual, saving time to the medical team and, above all, the subject and his family.

Agradecimientos

Este trabajo no hubiera sido posible sin la ayuda del grupo de endocrinología del Hospital Sant Joan de Déu, tanto dentro como fuera del laboratorio, y a mi director, el Dr. Josep Jiménez Chillarón, por darme la oportunidad de poder formar parte de su grupo.

También me gustaría agradecer a mi tutor interno, el Dr. Esteban Vegas Lozano, la supervisión que me ha dado en todo momento del trabajo.

Finalmente, dar las gracias también al grupo del Dr. Alexandre Perera, de la Universitat Politècnica de Catalunya, por la ayuda en la realización de este proyecto. En especial, dar las gracias a María Maqueda, por su gran implicación en el proyecto y su soporte en todo momento.

Índice

Lista de figuras

1. Introducción

1.1 Contexto y justificación del Trabajo

Desde el grupo de investigación "Endocrinología Metabólica" (dentro del área "Enfermedades pediátricas con inicio en etapas iniciales de la vida") del Instituto de Investigación Sanitaria del Hospital de San Joan de Déu, reclutamos una cohorte clínica de 26 niños/as (17 niños y 9 niñas) prepuberales con obesidad severa que entraron en un programa nutricional para reducir peso. Se tuvo en cuenta la aparición prematura de la pubertad de las niñas respecto a los niños, teniendo en cuenta solo los sujetos que cumplieran estrictamente las condiciones propias de la fase prepuberal.

Antes de empezar el "tratamiento" (tiempo 0) se tomaron los datos de todos los sujetos (edad, sexo, peso, altura, hemoglobina, etc.). El programa consistía en un período de 6 meses bajo una dieta y un ejercicio moderado (programa llevado a cabo por los médicos del hospital). A los 6 meses se volvieron a tomar los datos de todos los sujetos.

Calculamos el zBMI de cada sujeto antes y después del tratamiento. Esta medida consiste en el BMI (Body Mass Index), calculado a partir de la altura y peso, corregido también por edad y sexo [1,2]. Observamos que la diferencia de zBMI (DzBMI) antes y después del programa no era homogénea en todos los sujetos, obteniendo un grupo que respondía mejor al tratamiento (incremento de la disminución del peso) y otro grupo que respondía peor. Calificamos a los individuos que respondían bien como HR (high responders) y a los que respondían mal como LR (low responders). Se obtuvieron 13 HR (9 niños y 4 niñas) y 13 LR (8 niños y 5 niñas).

El criterio para discriminar los dos grupos fue escogido por los jefes del grupo de investigación y consistió en un valor de DzBMI de -0.3. Valores más pequeños indicaban que el sujeto era HR y valores superiores indicaban que el individuo era LR.

Posteriormente, el grupo decidió caracterizar el patrón global de metilación de ADN en muestras de sangre de los mismos niños y niñas. Se usó un array de metilación (Illumina EPIC 850K arrays®) con 850.000 sondas (resolución de nucleótidos). Algunas regiones presentaron una metilación diferencialmente significativa (DMR) entre los individuos que respondieron al tratamiento y los que no lo hicieron [3].

Sorprendentemente, algunas de estas regiones estaban asociadas a genes que regulan proliferación y diferenciación celular, lo que sugirió que las vías implicadas en la estabilidad de la adiposidad infantil afectan la función de los pre-adipocitos [4].

Ante los buenos resultados del análisis preliminar, se decidió analizar en profundidad los datos del array e intentar desarrollar un modelo capaz de discriminar si un individuo respondería o no al tratamiento basándonos en algunos biomarcadores del panel epigenético obtenido por el array.

La importancia del potencial modelo de predicción recae en saber con antelación si un niño o niña va a responder al tratamiento explicado anteriormente o no, apostando por otros tratamientos y pudiendo agilizar la mejora del sujeto dado el segundo caso.

En este trabajo usaremos herramientas bioinformáticas para determinar, de las regiones identificadas, el número mínimo necesario capaz de discriminar eficientemente los individuos que no responderán a una intervención de los que si lo harán. El objetivo final es desarrollar un panel epigenético con capacidad pronóstico que se pueda aplicar en la práctica clínica (modelo de predicción).

Posteriormente se tratará de identificar nuevas vías metabólicas implicadas en el desarrollo y establecimiento de la obesidad infantil. Este

apartado se realizará a partir de la anotación en bases de datos (GO, KEGG...) de dichas regiones diferencialmente metiladas, con la finalidad de caracterizar posibles rutas, enzimas y procesos biológicos implicados en nuestro modelo de obesidad.

En cuanto a la justificación del trabajo, desde las prácticas de grado (Biología) me empecé a interesar por el metabolismo de la obesidad y la diabetes. Es por este motivo que busqué unas prácticas extracurriculares dónde trabajaran en dicho campo.

Cuando el grupo decidió usar un array de metilación, dada la gran cantidad de datos que esto conlleva, pensé que el análisis bioinformático y el desarrollo del panel epigenético (uso de técnicas y modelos estadísticos) podía ser una buena opción como trabajo final del máster.

1.2 Objetivos del Trabajo

El principal objetivo consiste en desarrollar un panel epigenético con capacidad pronóstico para aplicarlo a la práctica clínica diaria. A partir de las regiones diferencialmente metiladas que hemos identificado, determinaremos un panel epigenético que incluirá el número mínimo necesario de marcadores que permitan discriminar de manera eficaz los individuos que responderán de los que no responderán a un tratamiento. En conclusión, realizar un modelo de predicción.

El objetivo específico del anterior consistiría en definir el número mínimo de CpG*, aproximadamente entre 1 y 10, que nos permita hacer distinción eficazmente de los individuos que responden al tratamiento para la obesidad y los que no responden.

Como segundo objetivo, nos hemos marcado identificar nuevas vías metabólicas implicadas en el desarrollo y establecimiento de la obesidad infantil estableciendo una relación entre la metilación* y la expresión de genes. A partir de esta relación contrastaremos la expresión en múltiples bases de datos, buscando así todas las vías implicadas relacionadas descritas, así como las enzimas y otros procesos biológicos implicados generando una red de procesos implicados y relacionados directa e indirectamente.

* *Las palabras marcadas con un asterisco aparecen definidas en el glosario*

1.3 Enfoque y método seguido

En el estudio tenemos una matriz de 26 muestras (sujetos) por el número de CpG significativos. Sabemos que cada individuo tiene asociada una etiqueta según el grupo (HR o LR) y una DzBMI.

Desde el inicio, consideramos óptima la obtención de los CpG significativos a partir del paquete de R *"limma",* frecuentemente utilizado en el análisis de microarrays debido a su alta efectividad y calidad en los resultados así como una gran variedad de funciones para el análisis [5]. Para la obtención de los CpG significativos usaremos las etiquetas HR/LR como regresoras (variable respuesta).

Para escoger nuestros CpG se tendrá en cuenta un p-valor con FDR < 0.05 y un "effect size" (cambio entre las medias del valor de *beta* de cada variable entre los dos grupos) mayor del 10%. Tomamos como referencia un cambio del 10% entre las *betas* (puede parecer muy poco restrictivo) debido a los resultados de otros artículos de metilación donde aún son menos restrictivos (effect size > 5%). [6,7]

Una vez obtenidos los CpG, para el desarrollo del panel epigenético* con valor pronóstico podemos utilizar modelos predictivos basados en modelos lineales, como la regresión PLS (Partial least squares regression), métodos de análisis discriminante como PLS-DA (Partial least squares discriminant analysis), métodos para la selección de variables o predictores óptimos como las técnicas de SPLS (Sparse Partial Least Squares), técnicas *clustering* o PCA (Principal Component Analysis), entre otras.

También podemos usar métodos de machine learning basados en *"training-test"* (validación cruzada) como el K-Fold.

> *"La validación cruzada es una técnica utilizada para evaluar los resultados de un análisis estadístico y* ***garantizar que son independientes de la partición entre datos de entrenamiento y prueba.*** *Consiste en repetir y calcular la media aritmética obtenida de las medidas de evaluación sobre diferentes particiones.* ***Se utiliza en entornos donde el objetivo principal es la predicción*** *y se quiere estimar cómo de preciso es un modelo que se llevará a cabo a la práctica. Es una técnica muy utilizada en proyectos de inteligencia artificial para validar modelos generados."* [8]

Sin embargo, consideramos que tenemos una muestra muy pequeña (n=26) para aplicar los métodos de validación cruzada. Por otro lado, consideramos que hacer la selección de predictores con alguno de los métodos anteriores como PLS, PLS-DA o PCA, era "hacer trampa" ya que estábamos condicionando el modelo de predicción a unos CpG escogidos a partir de toda nuestra muestra, sin validación externa.

Por este motivo decidimos que la mejor opción en nuestro caso era crear 26 modelos de regresión con limma, usando en cada modelo 25 de nuestras muestras y descartando una diferente cada vez, usando todos los CpG significativos en los 26 modelos. Lo que haríamos de este modo es crear un LOO (Leave One Out) externo como validación para nuestro modelo final.

El paso posterior sería escoger los "mejores" CpG de los coincidentes en los 26 modelos como predictores de nuestro modelo.

No obstante, nuestros individuos se dividen en dos grupos (HR/LR) en función del DzBMI usando el cutoff -0.3 para dividir los dos grupos. Por lo que nos planteamos también utilizar la diferencia de BMI corregida (DzBMI) como variable regresora en vez de la etiqueta HR/LR (que se determina a partir del DzBMI). Ésto se debe a que el regresor DzBMI tiene mas grados de libertad pero no posee cutoff, ya que es una variable continua cuantitativa, considerándola un mejor regresor para la creación de los modelos.

Decidimos escoger para cada modelo (x26) sus CpG significativos en función de los valores *m* y del DzBMI, de modo que cada modelo tiene unos CpG significativos escogidos a partir de 25 muestras, dejando una como LOO. Cuando hagamos el *overlap* más adelante éste si será "lícito", ya que el modelo no procederá de todas las muestras a la vez y por tanto, "no estará condicionado".

De todos modos se realizará una selección de variables con mayor capacidad predictiva a partir del método SPLS para compararlas con nuestros predictores obtenidos a partir de los 26 modelos con limma. Para los dos modelos SPLS se tendrán en cuenta los CpG significativos obtenidos a partir de las etiquetas HR/LR y se usará la variable DzBMI como regresora.

Finalmente, para la realización del segundo objetivo utilizaremos el paquete de R SIGORA. Este paquete nos permite enriquecer genes en las bases de datos KEGG y REACTOME para el genoma humano usando el mismo pipeline. [9]

Primero tendremos que anotar los genes que coinciden con las posiciones de nuestros CpG significativos a partir de la tabla de anotaciones que nos proporciona la casa comercial del array (https://support.illumina.com/array/array_kits/infinium- methylationepic-beadchip-kit/downloads.html).

Una vez anotados y separados los genes, obtendremos el valor Entrez de cada gen (único para cada uno) usando el paquete de R "org.Hs.eg.db ". Enriqueceremos los valores de Entrez con el paquete SIGORA en las dos bases de datos mencionadas anteriormente usando un p-valor de 0.01 y la corrección de Bonferroni.

Por último analizaremos los resultados y haremos una discusión desde el punto de vista biológico para explicar el epigenoma de nuestros fenotipos.

1.4 Planificación del Trabajo

1.4.1 Tareas:

Aquí numeramos las principales tareas a realizar durante el trabajo. Éstas constituirán el cuerpo del trabajo, sin embargo, contemplamos la posibilidad de incluir nuevos apartados si es preciso para la comprensión del trabajo.

1. Hallar los CpG significativos entre los dos grupos (entre 788373 CpG totales) mediante los respectivos test y filtros (limma, efect size...) [10].
2. Una vez obtenidos los CpG significativos construiremos 26 modelos (26 limmas "con validación LOO") y haremos la evaluación de dichos modelos.
3. Seleccionaremos el modelo óptimo o en cualquier caso escogeremos los CpG overlap en los 26 modelos para la creación de un modelo consenso.
4. Enriquecimiento de los CpG anteriores en bases de datos [11].
5. Explicación biológica del epigenoma en relación al fenotipo (la explicación será a partir de expresión/regulación gènica).
6. Redacción de esta memoria y preparación de la presentación.

1.4.2 Calendario:

En el siguiente diagrama de Grantt vemos el calendario para la realización del trabajo. En morado marcamos la planificación real y en amarillo la posible demora de las tareas, ambas por semanas.

Durante la primera semana obtendremos los CpG diferencialmente metilados entre los dos grupos. En principio no esperamos ningún tiempo de demora en este punto.

Las siguientes dos semanas (más unos tres días para una posible demora) nos servirán para la realización de los 26 modelos de regresión lineal (excluyendo una muestra diferente en cada modelo).

En la cuarta semana decidiremos el modelo óptimo (probablemente un modelo consenso explicativo) y empezaremos el enriquecimiento de los genes situados donde se encuentran nuestros CpG significativos. Esta última tarea puede alargarse unos dos o tres días más de lo previsto.

Por último, la quinta y sexta semana las dedicaremos a realizar una discusión para los resultados del enriquecimiento obtenidos y para finalizarlo en el caso que no lo esté ya.

Todo el tiempo restante a partir del paso anterior lo dedicaremos a la redacción de la memoria y a la preparación de la presentación.

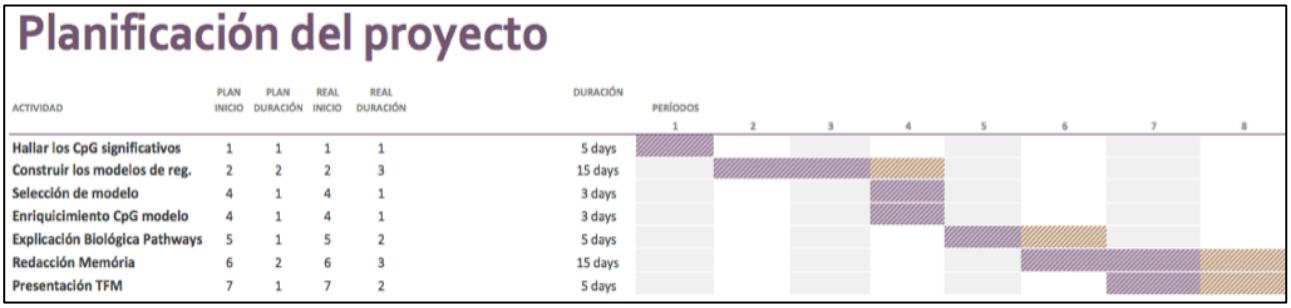

Figura 1. Diagrama de Grantt con la temporalización del proyecto

1.4.3 Hitos:

Será muy importante cumplir el calendario sobretodo las primeras 3-4 semanas. Es decir, no retrasar la obtención de los CpG significativos y la creación de los modelos.

Entonces para la PEC 2 o "fase de trabajo 1" entregaría los modelos de regresión finalizados.

Para la PEC 3 o "fase de trabajo 2" terminaría de escoger el modelo óptimo (en el caso que lo haya) si no lo he terminado para la PEC 2 y realizaría la parte de enriquecimiento y discusión.

1.4.4 Análisis de riesgos:

Personalmente creo que el único "riesgo" es la validación del modelo con arrays de metilación de una segunda cohorte de niños. Esto se debe

a que la segunda cohorte no habrá terminado el tratamiento para la fecha de entrega del TFM.

La intención es la obtención de un modelo, que provisionalmente solo será explicativo de nuestra cohorte (exclusivamente). Quedaremos pendientes de su validación para cohortes externas.

1.5 Breve sumario de productos obtenidos

1. Tabla con todos los CpG significativos en función del valor m y la etiqueta HR/LR.

   No los vamos a usar para crear los modelos de predicción pero si para el enriquecimiento. Solo usamos la etiqueta DzBMI para hacer la regresión y crear el modelo. Para el enriquecimiento y la discusión usamos el valor marcado por los investigadores.

   Obtenemos 214 CpG diferencialmente metilados entre los dos grupos para los 26 individuos. El producto se ha obtenido con un modelo lineal limma y las restricciones p-valor con FDR < 0.05 y effect size mayor al 10%.

2. Posibles modelos consenso a partir de los 26 modelos LOO con limma.

   Consta de un informe detallado del proceso de la obtención de los 26 modelos de regresión así como propone 4 modelos consenso finales. Incluye las curvas ROC y los valores AUC para los modelos, así como otras aportaciones interesantes para la posterior selección del modelo más óptimo.

3. Modelos SPLS para la obtención de los predictores óptimos.

   Realizados con dos paquetes de R diferentes (mixOmics y spls). Obtenemos unos CpG óptimos para predecir que compararemos con los obtenidos a partir de nuestros 26 modelos.

4. Análisis de enriquecimiento.

   Consiste en el proceso de anotación de los CpG diferencialmente metilados (según etiquetas HR/LR) con los genes que coinciden con sus posiciones y del enriquecimiento de estos genes correspondientes a los CpG significativos.

1.6 Breve descripción de los otros capítulos de la memoria

Estudio del array

Para entender e interpretar bien los resultados hace falta estudiar el soporte usado para el experimento (tipo de array y sus características).

En este apartado comentaremos como se calculan e interpretan los valores para analizar que nos devuelve el array.

CpG diferencialmente metilados

Consiste en la obtención de los CpG diferencialmente metilados con el criterio HR y LR según un cutoff de -0.3 DzBMI.

Es un apartado importante ya que estos CpG son los que se van a usar para enriquecer y hacer la discusión.

Obtención de 26 modelos de regresión con limma (validación LOO)

Este apartado contiene el código R junto con la explicación del proceso de obtención de los 26 modelos de regresión.

Modelo consenso

Directamente vinculado con el apartado anterior aquí se proponen cuatro ejemplos de modelo consenso calculados a partir de los anteriores.

Seleccionamos el modelo consenso óptimo, que consistirá en nuestro modelo de predicción final.

SPLS

Este apartado contiene dos modelos SPLS para nuestros CpG significativos (según HR/LR) usando el DzBMI como regresor.

Un modelo se ha realizado con el paquete "mixOmics" y el otro con el paquete "spls". Comparamos los predictores óptimos obtenidos entre los dos modelos anteriores y nuestro modelo consenso.

Enriquecimiento

Apartado que consiste en el proceso de anotación de los CpG diferencialmente metilados (según etiquetas HR/LR) con los genes que coinciden con sus posiciones y del enriquecimiento de estos genes.

El enriquecimiento se realiza en dos bases de datos (KEGG y REACTOME) mediante el paquete de R Sigora.

Discusión

En el último apartado tratamos de dar una explicación al epigenoma de nuestros individuos según el fenotipo.

Nos basamos en los resultados del apartado anterior para construir una hipótesis sobre nuestro modelo de obesidad infantil.

2. Resto de capítulos

2.1 Estudio del array

Trabajamos con el kit de Illumina "Infinium MethylationEPIC Kit". El *output* del array de metilación nos devuelve unos valores de *beta* y unos valores de *m*.

Para trabajar con estos valores necesitamos saber cómo se calculan, cómo se comportan, la relación entre ellos, sus características, etc. Esto nos servirá para elegir con criterio un valor u otro para hacer según que análisis.

Esta parte del trabajo (más teórica, pero imprescindible) no es muy extensa, sin embargo, nos ha servido para usar el valor *m* en la selección de CpG significativos, ya que este valor es homocedástico.

A la vez que usaremos los valores de *beta* para el filtro posterior (effect size). Este segundo filtro a partir de *beta* esta respaldado por muchos artículos de metilación [6,7].

Definición del valor *m* i del valor *beta* [12]:

Tal como se define en el artículo "*Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis*", cito textualmente:

"El valor beta es la relación entre la intensidad de la sonda metilada y la intensidad global (suma de las intensidades de la sonda metilada y no metilada). Siguiendo la notación utilizada por el ensayo de metilación de Illumina, el valor Beta para un sitio CpG interrogado se define como:

$$Beta_i = \frac{\max(y_{i,methy},0)}{\max(y_{i,unmethy},0)+\max(y_{i,methy},0)+\alpha}$$

Donde "yi, methy" y "yi, unmethy" son las intensidades medidas por la i-parte de la sonda metilada y no metilada, respectivamente. Para evitar valores negativos después del ajuste de fondo, cualquier valor negativo se restablecerá a 0. Illumina recomienda agregar un valor constante α (por defecto, α = 100) al denominador para regularizar el valor Beta cuando las intensidades de la sonda metilada y no metilada son bajas. El valor Beta da como resultado un número entre 0 y 1, o 0 y 100%. En condiciones ideales, un valor de cero indica que todas las copias del sitio CpG en la muestra estaban completamente no metiladas (no se midieron moléculas metiladas) y un valor de uno indica que se metiló cada copia del sitio. Si asumimos que las intensidades de la sonda siguen una distribución gamma, entonces el valor Beta sigue una distribución Beta.

El valor M se calcula como el log con base 2 de la relación de las intensidades de la sonda metilada frente a la sonda no metilada:

$$M_i = \log_2 \left(\frac{\max(y_{i,methy}, 0) + \alpha}{\max(y_{i,unmethy}, 0) + \alpha} \right)$$

Puede haber grandes cambios debido a pequeños errores en la estimación de la intensidad, porque para los valores de intensidad muy pequeños (especialmente entre 0 y 1), los pequeños cambios en la intensidad de la sonda a ser metilada y no metilada llevan a grandes cambios en la Valor-M."

En nuestros datos podemos ver como la mayor parte de los CpG se encuentran en los rangos 0-0.2 (hipometilados) y 0.8-1 (hipermetilados) (**Figura 15 - Anexo**).

Relación entre valor *m* y valor *beta*:

Aunque el incremento del valor *m* y del valor *beta* indiquen más metilación en la sonda y viceversa, los dos valores no tienen una relación lineal. A continuación, la ecuación y el gráfico que relacionan los dos valores:

$$Beta_i = \frac{2^{M_i}}{2^{M_i} + 1}; M_i = \log_2 \left(\frac{Beta_i}{1 - Beta_i} \right)$$

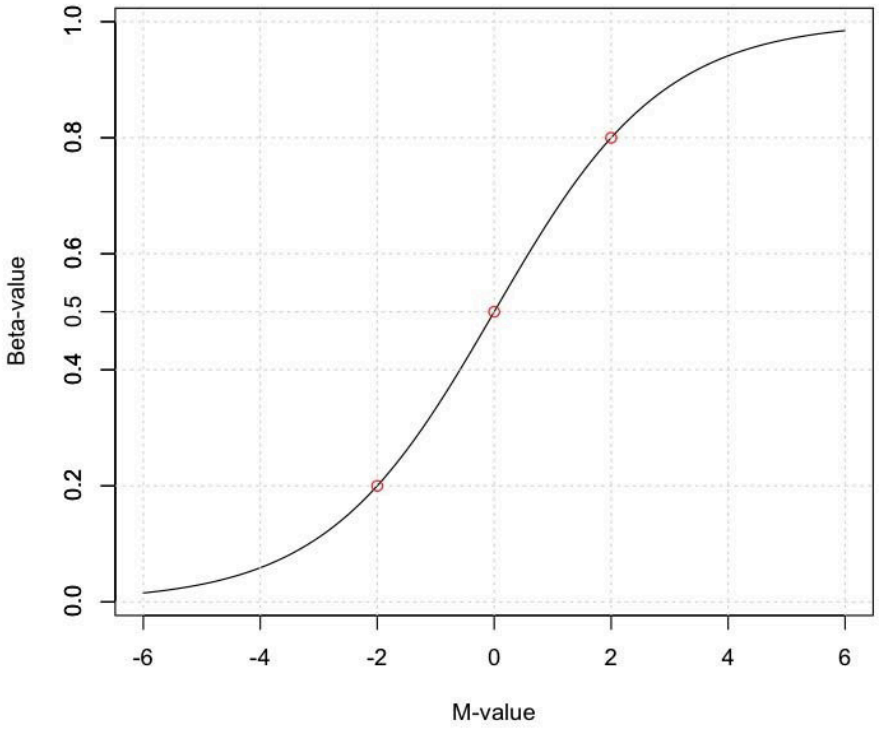

Figura 2. Gráfico de la relación entre los valores m y *beta*. [12]

La distribución de la desviación estándar a través de diferentes niveles de metilación:

La heterocedasticidad de los datos puede ser un problema a la hora de analizar datos de alto rendimiento como los arrays.

Un modo de probar la homocedasticidad de los datos es visualizando las relaciones entre la media y la desviación estándar.

Citamos de nuevo el mismo artículo:

"La siguiente figura muestra las relaciones de desviación media y estándar del valor Beta y el valor M. Los puntos rojos representan la desviación estándar mediana dentro de una ventana local. La desviación estándar del valor Beta se comprime en gran medida en los rangos bajos (entre 0 y 0,2) y altos (entre 0,8 y 1).

Esto significa que el valor Beta tiene una heterocedasticidad significativa en el rango de metilación bajo y alto. El problema de la heteroscedasticidad se resuelve eficazmente después de transformar el valor Beta en el valor M."

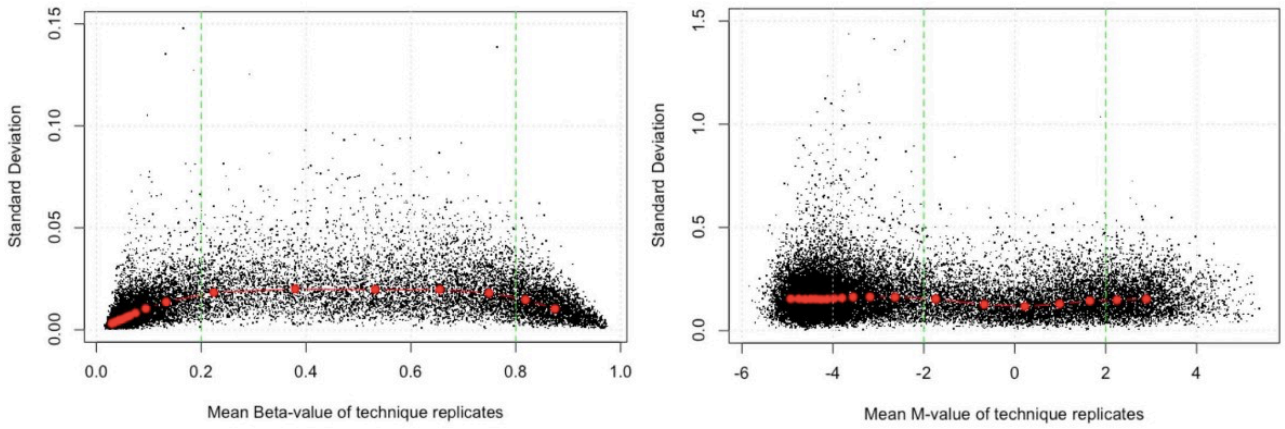

Figura 3. Gráficos de los valores medios y desviación estándar de los valores *m* y *beta*. [12]

En base a los resultados descritos en la bibliografía [12] y a otros artículos de metilación [6,7], utilizaremos los valores de m para seleccionar los CpG diferencialmente metilados (debido a la homocedasticidad de este valor) y los valores de *beta* para calcular el "effect size".

2.2 CpG diferencialmente metilados

El principal objetivo de este apartado es realizar el análisis de metilación diferencial de CpG entre los grupos HR y LR (usaremos las etiquetas HR/LR como regresor).

Antes de empezar a trabajar con los datos, la empresa Making Genetics se encargó de realizar el filtraje de los datos crudos del array y la normalización. Por este motivo nosotros trabajamos a partir de 788.373 CpG y no a partir de 850.000 CpG (número total de sondas del array). Se descartaron las sondas según los siguientes criterios:

- Filtrado de las sondas con un p-valor de detección superior a 0.01.
- Filtrado de las sondas con un *beadcount* inferior a 3 en al menos el 5% de las muestras.
- Filtrado de las sondas asociadas SNPs.
- Filtrado de las sondas que se alinean a múltiples ubicaciones.
- Filtrado de las sondas del cromosoma X o Y.

Ahora ya podemos empezar con el análisis. Para el test se utilizó el paquete limma sobre los valores de m de las diferentes muestras (26 muestras, HR/LR). Entonces solo necesitamos:

- Los M-value dispuestos en una matriz con los valores dispuestos en filas correspondientes a CpGs y columnas a muestras.
- Los fenodatos dispuestos en un dataframe con las muestras dispuestas en filas y en columnas la información de las variables que vayamos a utilizar (sea la etiqueta de HR/LR, edad ó genero por ejemplo). A continuación:

Sample_Name	Sample_Group	Sample_Sex	Sample_Age
MO13_0M	LR	M	9,01
MO14_0M	LR	M	8,2
MO18_0M	LR	F	9,38
MO23_0M	HR	M	10,11
MO27_0M	LR	F	9,11
MO31_0M	LR	F	7,12
MO36_0M	LR	M	10,35
MO39_0M	LR	M	10,85
MO42_0M	LR	F	8,74
MO48_0M	LR	F	8,6
MO52_0M	LR	M	9,46
MO53_0M	LR	M	7
SA10_0M	LR	M	10,72
SA12_0M	LR	M	9
MO28_0M	HR	F	7,39
MO29_0M	HR	M	9,58
MO30_0M	HR	F	8,47
MO32_0M	HR	M	8,63
MO34_0M	HR	F	8,97
MO41_0M	HR	M	8,22
MO44_0M	HR	M	8,46
MO45_0M	HR	M	8,48
MO47_0M	HR	F	9,07
SA11_0M	HR	M	10
SA32_0M	HR	M	6,09
SA37_0M	HR	M	9,98

Tabla 1. Fenodatos de todas las muestras del estudio.

Una vez importados y ordenados los datos, procedemos con el análisis limma:

```
require(limma)
initialmodel <- model.matrix( ~ Sample_Group + Sample_Sex, data)
model <- lmFit(mvalHRLR, initialmodel)
modelstats <- eBayes(model)
```

A parte de la etiqueta principal (HR/LR) incluimos en el modelo las variables "sex" y "age".

Los resultados muestran como en base a la variable "age" no se prioriza ningún CpG por lo tanto esta variable es susceptible de ser eliminada del modelo.

Y ahora obtenemos prioritarios (FDR 5%) para la etiqueta HR y LR:

```
#Cuántos hay con adj pval <0.05 (FDR)
num_rank <- sum(p.adjust(modelstats$p.value[,2],
method = "fdr")<0.05)
```

```
#Hacemos una selección de los que tienen un adj pval
<0.05
rankedCpG_HRLR <-topTable(modelstats,
number=num_rank ,coef=2, adjust="fdr",p.value=0.05)
```

Importamos los valores de *beta* para hacer el cálculo del effect size, aplicando el segundo filtro.

No necesitamos todos los CpGs, por lo que reducimos dicho dataframe a solo aquellos CpGs que han sido priorizados por limma (FDR).

```
CpGs_betas_limma <- CpGs_betas[which (rownames
(CpGs_betas) %in% rownames (rankedCpG_HRLR)),]
```

Calculamos la media por grupo (HR, LR):

```
LRs <- data$Sample_Name[which(data$Sample_Group %in%
"LR")]

CpGs_betas_LR <- CpGs_betas_limma[,which (colnames
(CpGs_betas_limma) %in% LRs)]

CpGs_betas_HR <- CpGs_betas_limma[,-(which(colnames
(CpGs_betas_limma) %in% LRs))]

#Media de los valores de Beta para el grupo de LR y
HR respectivamente

LR_avg <- apply(CpGs_betas_LR,1,function(x) mean(x))

HR_avg <- apply(CpGs_betas_HR,1,function(x) mean(x))

#Effect size para cada CpG. Expresado en %

Effect_size <- (LR_avg-HR_avg)*100
```

Juntamos ambas informaciones para finalmente hacer la selección de CpGs diferencialmente metilados.

```
CpGs_HRvsLR <- data.frame (Cg_ID = rownames
(CpGs_betas_limma) , HR_avg = HR_avg, LR_avg =
LR_avg, Effect_size = Effect_size)

#Y ahora antes de unir los resultados de limma,
reordenamos los resultados de limma para hacer una
unión directa

rankedCpG_HRLR2 <- rankedCpG_HRLR[match (rownames
(CpGs_HRvsLR),rownames(rankedCpG_HRLR)),]

#Finalmente unimos resultados

CpGs_HRvsLR <- cbind(CpGs_HRvsLR, rankedCpG_HRLR2
[,3:6])
```

Los CpG ahí listados ya tienen un adj p-val < 0.05 por FDR. Para acabar de tener los CpG priorizados, ahora hay que filtrar estos por effect size >= 10% (valor absoluto).

```
CpGs_HRvsLR_FINAL <- CpGs_HRvsLR[abs(Effect_size) >=
10 ,]
```

Por lo tanto, ahora resultan **214 CpGs con FDR < 0.05 y effect_size >=10%.**

Vamos a basar el análisis de enriquecimiento en estos CpGs diferencialmente metilados (dos filtros).

La tabla 2 nos muestra a continuación los primeros 43 de estos 214 CpG diferencialmente metilados.

	Cg_ID	HR_avg	LR_avg	Effect_size	t	P.Value	adj.P.Val	B
cg04088940	cg04088940	0.6625531	0.52090715	-14.16459	5.131557	1.883723e-05	0.01627479	2.860740349
cg17210938	cg17210938	0.3372966	0.21818368	-11.91129	4.967643	2.955040e-05	0.01724634	2.454520431
cg23958373	cg23958373	0.8283046	0.71031289	-11.79917	3.897080	5.462003e-04	0.03414555	-0.179396869
cg15022308	cg15022308	0.5139409	0.61575981	10.18189	-5.656598	4.472207e-06	0.01519741	4.154516252
cg22308949	cg22308949	0.4523825	0.34348455	-10.88980	3.798393	7.108782e-04	0.03729783	-0.416660838
cg03847932	cg03847932	0.7870617	0.67872754	-10.83342	4.818313	4.453451e-05	0.01822040	2.084172618
cg05834845	cg05834845	0.7335269	0.61331754	-12.02093	3.484361	1.626388e-03	0.04990465	-1.159653747
cg18254123	cg18254123	0.3355945	0.15371285	-18.18817	4.042504	3.695286e-04	0.03040264	0.172870089
cg11731671	cg11731671	0.4043339	0.56152517	15.71913	-3.948099	4.763768e-04	0.03266898	-0.056146557
cg14157435	cg14157435	0.5378353	0.31226583	-22.55695	5.834871	2.752573e-06	0.01519741	4.589603969
cg01821018	cg01821018	0.6577446	0.50326328	-15.44813	4.140081	2.838990e-04	0.02822094	0.410743471
cg07234876	cg07234876	0.8544814	0.72433084	-13.01506	3.885767	5.629923e-04	0.03441979	-0.206673899
cg23066280	cg23066280	0.5423558	0.43939292	-10.29629	4.157136	2.710871e-04	0.02794878	0.452428703
cg11949518	cg11949518	0.4992519	0.37060367	-12.86482	4.445239	1.237807e-04	0.02290502	1.160533783
cg17658113	cg17658113	0.7456859	0.61214968	-13.35363	4.679054	6.526349e-05	0.01978922	1.738928305
cg23541304	cg23541304	0.4331690	0.63748973	20.43207	-3.639858	1.082006e-03	0.04322830	-0.794252897
cg17611936	cg17611936	0.6560047	0.54240049	-11.36042	3.713735	8.901154e-04	0.04034816	-0.618880468
cg10510935	cg10510935	0.4077341	0.28026607	-12.74680	4.212525	2.332865e-04	0.02679429	0.588016513
cg00977403	cg00977403	0.4289942	0.53064459	10.16504	-6.267658	8.552119e-07	0.01487481	5.633274160
cg16702083	cg16702083	0.3242596	0.21431652	-10.99431	4.131310	2.907185e-04	0.02842548	0.389317858
cg06379361	cg06379361	0.1485377	0.26778829	11.92506	-3.561462	1.329606e-03	0.04643341	-0.979139403
cg12923728	cg12923728	0.5715093	0.71182406	14.03147	-5.241415	1.393279e-05	0.01592417	3.132583337
cg13803195	cg13803195	0.7863446	0.64147345	-14.48711	4.232653	2.208810e-04	0.02645094	0.637362069
cg11189868	cg11189868	0.7205789	0.62052750	-10.00514	4.346682	1.619701e-04	0.02439917	0.917571374
cg05630272	cg05630272	0.4201654	0.54507983	12.49145	-5.439723	8.089933e-06	0.01534281	3.621975935
cg05333146	cg05333146	0.7628294	0.66089617	-10.19332	4.387061	1.450853e-04	0.02372564	1.017036449
cg26839010	cg26839010	0.6703617	0.51246638	-15.78954	4.527573	9.883151e-05	0.02165698	1.363940168
cg12150299	cg12150299	0.5165918	0.41417851	-10.24133	5.014757	2.596276e-05	0.01668719	2.571334217
cg00036352	cg00036352	0.3973827	0.22693003	-17.04526	4.559471	9.056923e-05	0.02130141	1.442831549
cg18768136	cg18768136	0.5139501	0.65425956	14.03094	-4.474574	1.142463e-04	0.02252451	1.232964156
cg04259027	cg04259027	0.4280642	0.60176827	17.37040	-3.763868	7.792573e-04	0.03853056	-0.499285314
cg11657665	cg11657665	0.6849288	0.50926370	-17.56651	4.548662	9.328899e-05	0.02136951	1.416094449
cg13981132	cg13981132	0.2195538	0.32168892	10.21351	-4.393927	1.423932e-04	0.02361159	1.033959196
cg03498762	cg03498762	0.7633543	0.65413743	-10.92169	3.844422	6.287752e-04	0.03570436	-0.306193076
cg14648920	cg14648920	0.3952282	0.52859442	13.33663	-4.434913	1.273214e-04	0.02299583	1.135048921
cg02218090	cg02218090	0.4442194	0.31595923	-12.82602	3.699376	9.246097e-04	0.04091190	-0.653050829
cg11121987	cg11121987	0.3671089	0.47094160	10.38327	-4.295474	1.862057e-04	0.02523181	0.791603629
cg04738746	cg04738746	0.6065185	0.48255131	-12.39672	4.193971	2.453298e-04	0.02717060	0.542564985
cg05387464	cg05387464	0.4834284	0.34616089	-13.72675	5.282867	1.243498e-05	0.01592417	3.235037436
Showing 1 to 43 of 214 entries								

Tabla 2. Tabla de los primeros 43 de los 214 CpG diferencialmente metilados teniendo en cuenta un FDR< 0.05 y un effect size >= 10%.

Para finalizar el apartado, incluimos algunos gráficos de los datos crudos (pasado el filtraje y normalización) con el fin de compararlos con los mismos gráficos para los 214 CpG significativos.

En primer lugar vemos el manhattan plot de los 788.373 CpG (**Figura 4**). Este gráfico nos permite tener una visión general de nuestro array. Observamos una gran cantidad de puntos (CpG) por encima de la línea horizontal (que indica un FDR de 0.05), lo que indica que todos estos CpG serán candidatos para estar entre nuestros 214 CpG priorizados (**Tabla 2**). Se denominan "puntos o CpGs sugestivos".

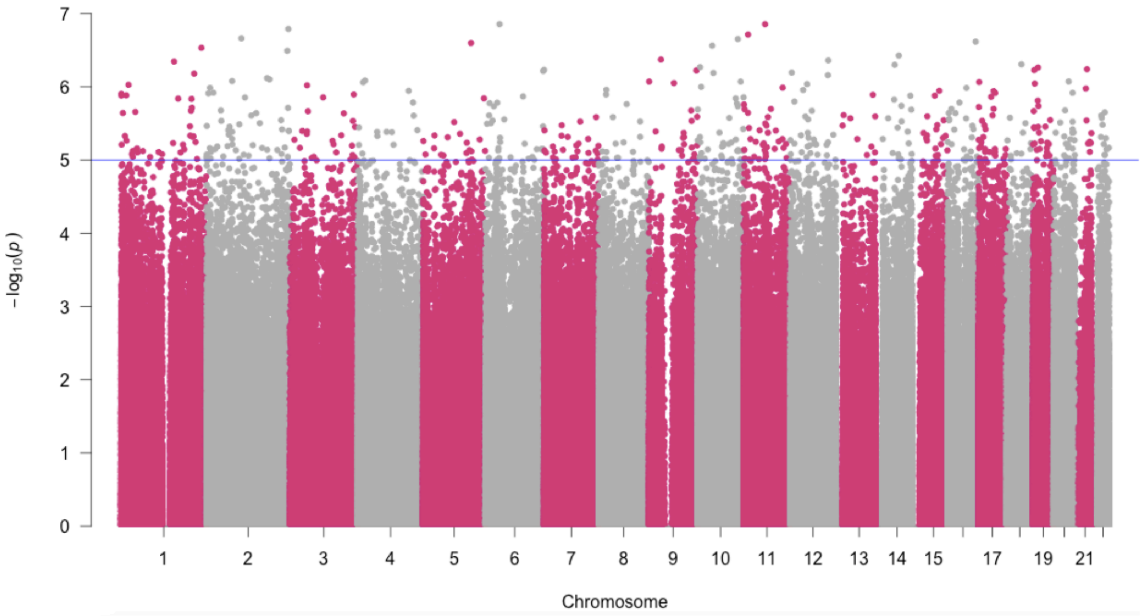

Figura 4. Manhattan plot de los datos después del filtraje. La línea azul indica un FDR de 0.05.

A primera vista no vemos más CpG significativos en ningún cromosoma en concreto. Hay una mayor cantidad en el cromosoma 1 o 2 pero muy probablemente se deba a la mayor longitud de estos cromosomas respecto a los otros.

Por otro lado tenemos el volcano plot, también de todos los puntos (CpG) del array después del filtraje (**Figura 5**). Sería una forma parecida al manhattan plot de tener una visión global de los resultados. Pero a diferencia del manhattan, que solo incluye el p-valor, el volcano plot también incluye el effect size, por lo que los dos ejes nos sirven para

discriminar nuestros CpG significativos en función de los dos filtros. Los puntos rojos equivaldrían aproximadamente a los 214 CpG de la Tabla 2.

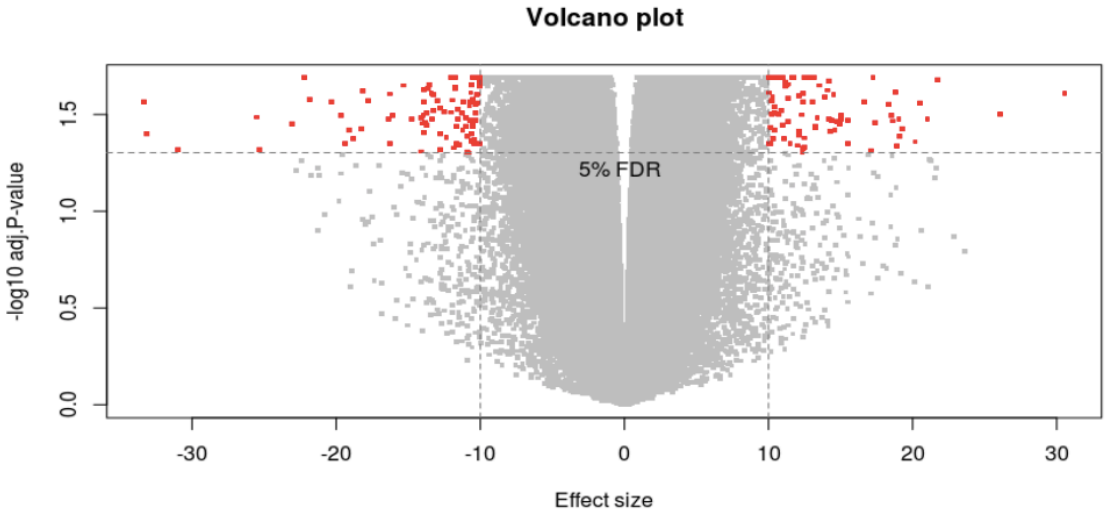

Figura 5. Volcano plot de los datos después del filtraje. El eje vertical nos indica un FDR de 0.05 y el eje horizontal un effect size del 10%.

Vemos en el dendrograma con los 788.373 CpG (**Figura 6**) que los dos grupos no segregan prácticamente. Sin embargo, cuando graficamos el dendrograma solo para los 214 CpG significativos (**Figura 7**) vemos una segregación casi total, donde solo un individuo con la etiqueta "LR" queda mal clasificado.

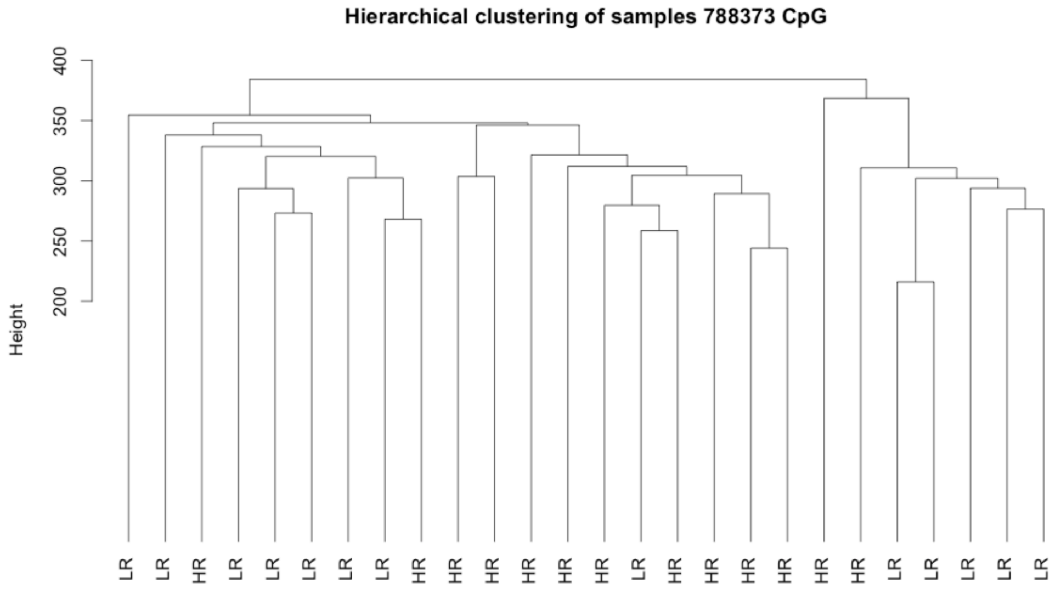

Figura 6. Dendrograma de los 26 individuos por los 788373 CpG

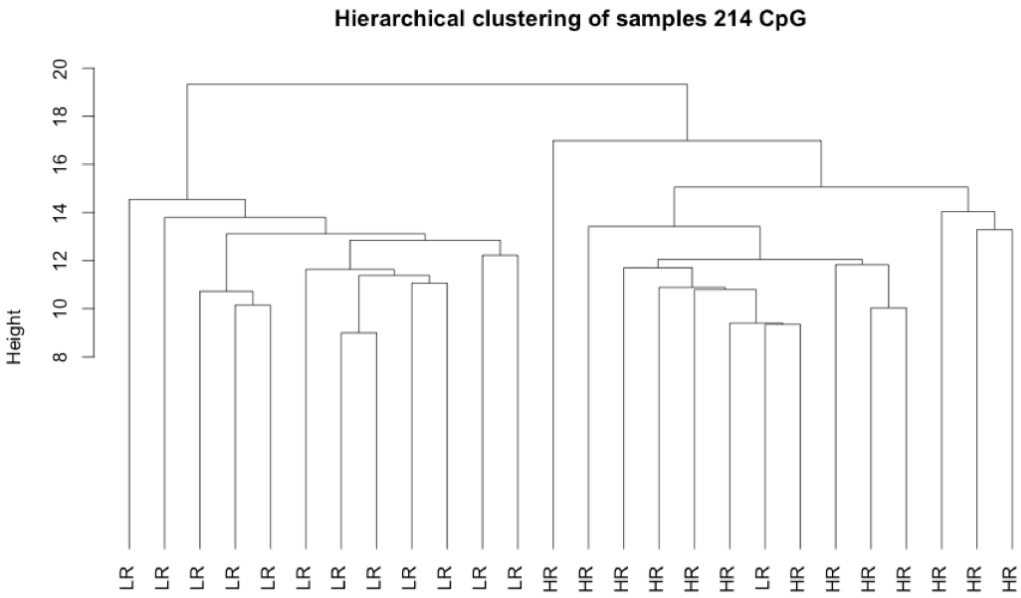

Figura 7. Dendrograma de los 26 individuos por los 214 CpG diferencialmente metilados

Por último calculamos los modelos PCA de los datos completos (788373 CpG) y de los CpG significativos. Hacemos las gráficas de la proporción de la varianza del modelo explicada para cada uno.

Para el primer caso (**Figura 8**) vemos como el primer componente explica solo el 25% de la varianza del modelo aproximadamente, mientras que el segundo y el tercero explican aproximadamente el 9% y el 6%.

Por otro lado, en el segundo modelo (**Figura 9**) vemos como el primer componente ya explica por él solo el 50% de la varianza, mientras que el segundo y el tercero explican aproximadamente el 7% y el 6%.

Concluimos que el modelo mejora claramente seleccionando solo nuestros 214 CpG diferencialmente metilados y por tanto, parecen ser buenos marcadores para discriminar los dos grupos.

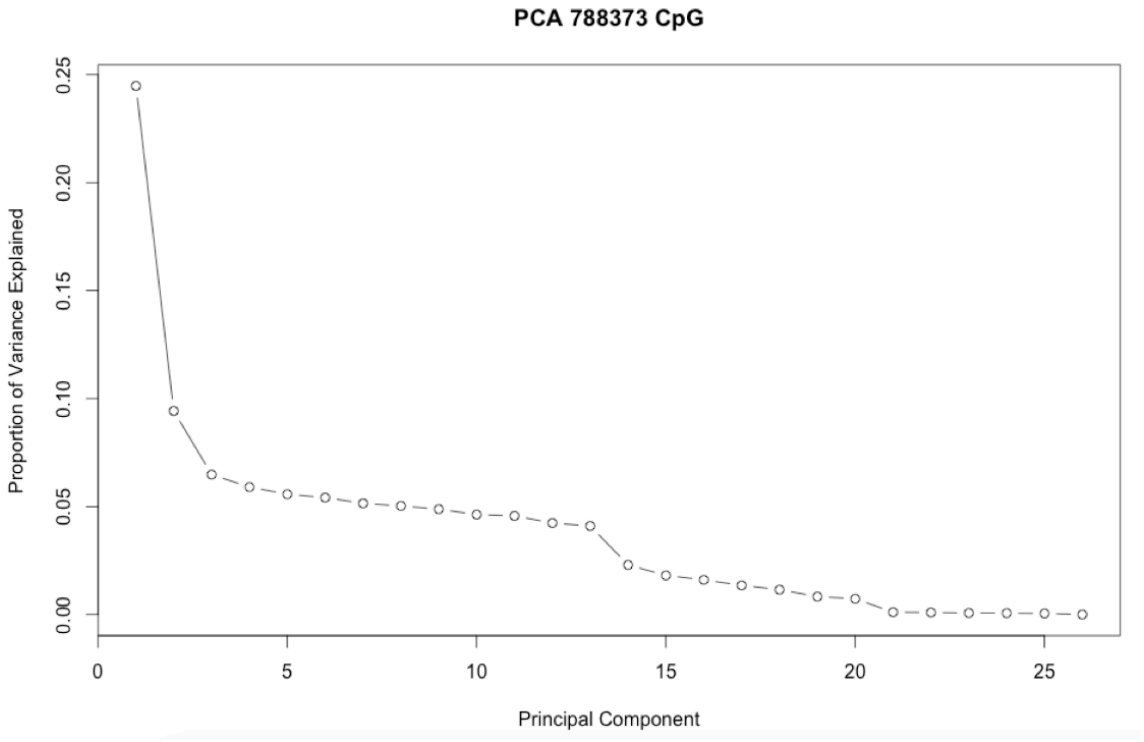

Figura 8. Principal Component Analysis. Proporción de la varianza explicada por los componentes en los 788373 CpG.

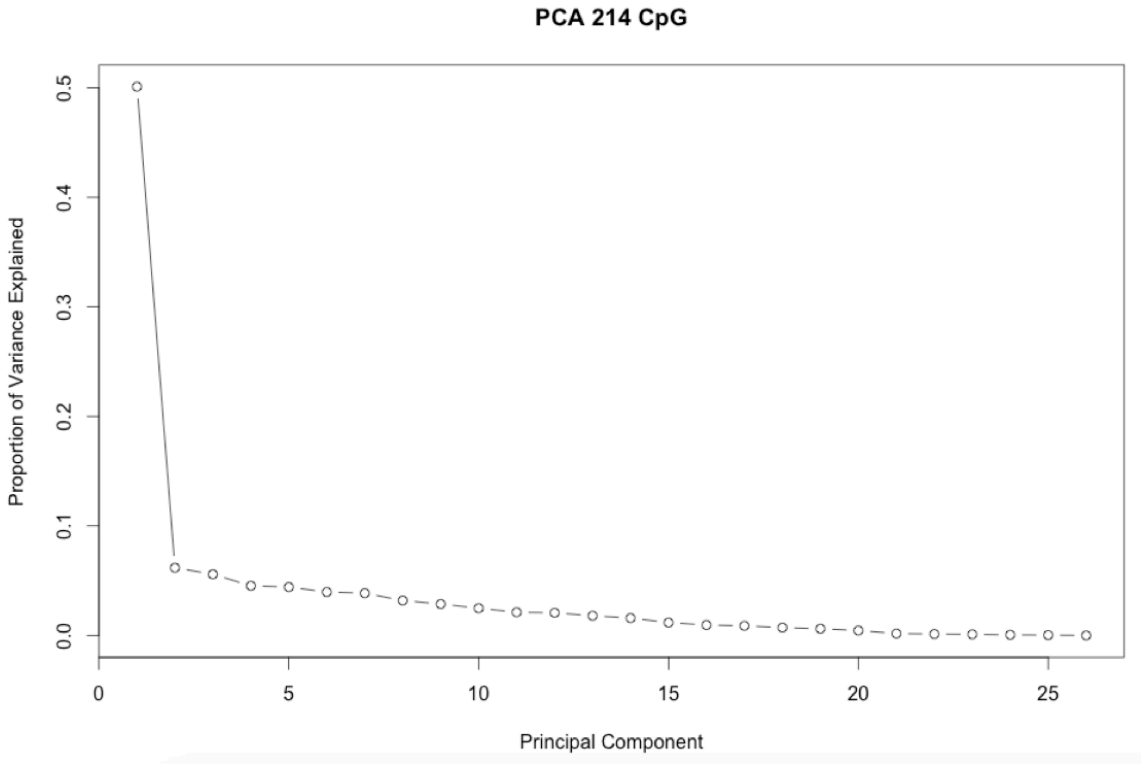

Figura 9. Principal Component Analysis. Proporción de la varianza explicada por los componentes en los 214 CpG significativos.

2.3 Obtención de 26 modelos de regresión con limma (validación LOO)

Como hemos comentado en apartados anteriores, para la creación de los modelos vamos a utilizar los CpG diferencialmente metilados obtenidos a partir de las etiquetas DzBMI, según el criterio seguido por el grupo. La decisión se tomó debido al restringente cutoff de las etiquetas HR/LR para hacer la regresión (y posterior predición) y al mayor número de grados de libertad que aportaban las etiquetas DzBMI. Además, la prueba con las nuevas etiquetas nos serviría para considerar un posible cambio del cutoff (para discriminar los grupos) inicial en función de los futuros resultados.

Este apartado consiste en el diseño de un algoritmo que nos permita realizar 26 modelos de regresión limma (excluyendo una muestra diferente en cada uno) a partir de los CpG significativos obtenidos a partir del valor m y la etiqueta de DzBMI mediante un proceso de iteración con código R. Además se añadirá la opción de realizar los modelos a partir de las etiquetas HR/LR y la opción de utilizar también los valores de m para calcular el *effect size*.

Estos 26 modelos nos van a proporcionar las 26 listas de CpG significativos que usaremos para generar un modelo consenso final, en función del *overlap* de éstas mismas.

Para llevar a cabo el proceso vamos a necesitar:

- Los m-values (mvalHRLR) completos (788373 CpG) de las 26 muestras que utilizaremos para limma y para priorizar (p-valor y si se decide, effect size).
- Los *beta*-values (CpGs_*beta*s) completos (788373 CpG) de las 26 muestras que utilizaremos para priorizar (effect size).
- Los metadatos que todas las 26 muestras: HR/LR, género, edad y DzBMI.

Esta función creará tantos listados de CpG priorizados como muestras encuentre. En cada listado se habrán considerado *n* muestras -1 con el objetivo de hacer un LOO externo a limma.

Hay varios parámetros de entrada a esta función dependiendo de qué se quiera hacer, básicamente:

- Input DzBMI: si TRUE el predictor a utilizar en la regresión limma será DzBMI, si FALSE es la etiqueta HR/LR (como en el apartado anterior).
- Input param_cutoff: selecciona los valores *beta* o los valores m para realizar el segundo filtrado de CpGs después de limma (effect size).
- Input value_cutoff: indicamos qué valor (en valor absoluto) de cutoff para el effect size (de *beta* o de m, según especificado anteriormente) aplicamos. (En el apartado anterior usábamos 10%, página 21).
  NOTA: Este último sin porcentajes.

Vamos a ponerle el nombre *"lists_x_model_tests"* a nuestra función. A continuación adjuntamos y explicamos el algoritmo diseñado:

```
#Definimos los parámetros de la función y cargamos
el paquete limma

require(limma)
lists_x_model_tests <- function(mvalHRLR,CpGs_betas,
data,sample_to_exclude, DzBMI=TRUE, param_cutoff =
"beta", value_cutoff= 0.1)
{
```

```
#Escogemos el conjunto de 25 muestras que queremos
analizar y creamos una matriz de datos a partir de
la matriz entera - Extrayendo "sample_to_exclude"

datasel <- data[-(which(data$Sample_Name %in%
sample_to_exclude)),]

mvalsel <-  mvalHRLR[,-(which(colnames(mvalHRLR)
%in% sample_to_exclude))]

betasel <- CpGs_betas[,-(which(colnames (CpGs_betas)
%in% sample_to_exclude))]

#Aplicamos limma en base a la etiqueta HR/LR (por
defecto) o por lo contrario, en base al valor de
DzBMI - Fase para la primera priorización

if(DzBMI == FALSE) {initialmodel <- model.matrix( ~
Sample_Group + Sample_Sex, datasel)}
if(DzBMI == TRUE) {initialmodel <- model.matrix( ~
dzBMI + Sample_Sex, datasel)}

#Hacemos el modelo con limma en función de los
valores de m de las 25 muestras escogidas "mvalsel"

model <- lmFit(mvalsel, initialmodel)
modelstats <- eBayes (model)

#Listado de priorización de limma (por variable
HR/LR o DzBMI)

num_rank <- sum(p.adjust(modelstats$p.value[,2],
method = "fdr")<0.05)
```

```
rankedCpG_HRLR <- topTable(modelstats, number=
num_rank,coef=2, adjust="fdr",p.value=0.05)

#Selección de CpGs de limma con valores de beta y
mval - Guardamos ambos y luego filtraremos por el
elegido

CpGs_beta_limma <- betasel[which(rownames(betasel)
%in% rownames(rankedCpG_HRLR)),]

CpGs_mval_limma <- mvalsel[which(rownames(mvalsel)
%in% rownames(rankedCpG_HRLR)),]

#Calculamos la media por grupo (HR, LR) para el
effect size cutoff. Será lo mismo si utilizamos
DzBMI (cutoff de -0.3) que fue lo que se utilizó
para las etiquetas. Lo hacemos x4 (HR beta, LR beta,
HR m, LR m)

LRs <- datasel$Sample_Name [which
(datasel$Sample_Group %in% "LR")]

CpGs_beta_LR <- CpGs_beta_limma[,which(colnames
(CpGs_beta_limma) %in% LRs)]

CpGs_beta_HR <- CpGs_beta_limma[,-(which(colnames
(CpGs_beta_limma) %in% LRs))]

CpGs_mval_LR <- CpGs_mval_limma[,which (colnames
(CpGs_mval_limma) %in% LRs)]

CpGs_mval_HR <- CpGs_mval_limma[,-(which(colnames
(CpGs_mval_limma) %in% LRs))]
```

```
#Media de los valores de Bet/Mval según especificado
para el grupo de LR y HR respectivamente

LR_beta_avg <- apply(CpGs_beta_LR, 1, function(x)
mean(x))
HR_beta_avg <- apply(CpGs_beta_HR, 1, function(x)
mean(x))
LR_mval_avg <- apply(CpGs_mval_LR, 1, function(x)
mean(x))
HR_mval_avg <- apply(CpGs_mval_HR, 1, function(x)
mean(x))

#Effect size (deltaBeta o deltaM) para cada CpG. NO
expresado en %!!!

Effect_size_beta <- (LR_beta_avg-HR_beta_avg)
Effect_size_mval <- (LR_mval_avg-HR_mval_avg)

#Fusión resultados limma y effect size

CpGs_HRvsLR <- data.frame(Cg_ID =
rownames(CpGs_beta_limma), HR_beta_avg =
HR_beta_avg, LR_beta_avg = LR_beta_avg,
Effect_size_beta = Effect_size_beta, HR_mval_avg =
HR_mval_avg, LR_mval_avg = LR_mval_avg,
Effect_size_mval = Effect_size_mval)

#Reordenamos los resultados de limma para hacer una
unión directa

rankedCpG_HRLR2 <-rankedCpG_HRLR[match(rownames
(CpGs_HRvsLR),rownames(rankedCpG_HRLR)),]
```

```
#Unimos los resultados

CpGs_HRvsLR <- cbind(CpGs_HRvsLR, rankedCpG_HRLR2
[,3:6])

#Selección de CpGs - Filtrado por effect size - sus
valores de beta o mval según input (Fase para la
segunda priorización)

if(param_cutoff == "beta") {CpGs_HRvsLR_FINAL <-
CpGs_HRvsLR[abs(Effect_size_beta) >=value_cutoff,]}

if(param_cutoff == "mval") {CpGs_HRvsLR_FINAL <-
CpGs_HRvsLR[abs(Effect_size_mval) >=value_cutoff,]}

CpGs_HRvsLR_FINAL$Cg_ID <- as.character
(CpGs_HRvsLR_FINAL$Cg_ID)

return(CpGs_HRvsLR_FINAL)
}
```

Hasta este punto hemos definido la función. Ahora hay que evaluar su viabilidad. Probamos el algoritmo:

```
#Creamos un objeto con la lista de todos los sujetos
ya que tendremos que excluir uno cada vez en un
proceso de iteración

toexclude <- data$Sample_Name
```

El objeto "*toexclude*" es el equivalente a la primera columna (*Sample_Name*) de la **Tabla 1.**

```
#Probamos nuestra función con los parámetros
siguientes
#Análisis de metilación diferencial en base a la
etiqueta HR/LR - Effect size Beta cutoff 0.1

lists_of_CpGs <- lapply(toexclude, function(x)
lists_x_model_tests(mvalHRLR,CpGs_betas,data,x,DzBMI
= FALSE, param_cutoff = "beta",value_cutoff = 0.1))

#A cada elemento de la lista, podemos darle el
nombre de la muestra que hemos excluido
especificando delante "wo" (without) para aclarar
que es sin esa muestra

names(lists_of_CpGs) <- paste("wo",data$Sample_Name,
sep="_")
unlist(lapply(lists_of_CpGs, function(x) dim(x)[1]))
```

```
> unlist(lapply(lists_of_CpGs, function(x) dim(x)[1]))
wo_M053_0M wo_M014_0M wo_M048_0M wo_M018_0M wo_SA12_0M wo_M013_0M wo_M052_0M wo_M031_0M
       217        140        269        368        178        206        139        201
wo_M039_0M wo_M027_0M wo_M023_0M wo_M042_0M wo_SA10_0M wo_M036_0M wo_M045_0M wo_SA32_0M
        69        168        267        144        233        209        119        131
wo_M034_0M wo_M032_0M wo_M047_0M wo_SA11_0M wo_M044_0M wo_M030_0M wo_M028_0M wo_M029_0M
       302        222        186        225        205        139        214        175
wo_M041_0M wo_SA37_0M
        87        290
```

Tabla 3. Producto del algoritmo para la creación de los 26 modelos. Obtenemos un listado de 26 modelos con el número de CpG significativos que hay en cada uno. Cada uno lleva el nombre de la muestra no incluida en el modelo.

Obtenemos los 26 modelos esperados para los parámetros que definimos (**Tabla 3**). Comprobamos que el algoritmo funciona.

El siguiente paso consiste en repetir el análisis para diferentes parámetros que consideremos lógicos (más restricción, menos restricción...), escoger el óptimo (según criterio del grupo) y determinar el modelo consenso para predecir.

NOTA: Para obtener la intersección de CpG entre los 26 modelos (**Tabla 4**) usaremos el código:

```
Reduce(intersect,lapply(lists_of_CpGs, function(x)
x$Cg_ID))
```

```
> Reduce(intersect,lapply(lists_of_CpGs, function(x) x$Cg_ID))
 [1] "cg04088940" "cg17210938" "cg14157435" "cg11949518" "cg17658113" "cg12923728" "cg05630272"
 [8] "cg26839010" "cg00036352" "cg11657665" "cg05387464" "cg08240913" "cg11601967" "cg25377862"
[15] "cg01565529" "cg08394602" "cg04863005" "cg05509609" "cg18872420" "cg10776230" "cg01229327"
[22] "cg21043213" "cg15398841" "cg00846098" "cg23881939" "cg02530824" "cg10126788" "cg03963391"
[29] "cg26173906" "cg24989447" "cg15953550" "cg24892374" "cg16007266" "cg03857535" "cg09127314"
[36] "cg09203111" "cg06565913" "cg09439754" "cg16072462" "cg07911738" "cg14920334" "cg16080552"
[43] "cg05348875" "cg20992785" "cg02602925" "cg10807027" "cg24790788" "cg16322792" "cg06834998"
[50] "cg00835812" "cg09365459" "cg15288329" "cg23743554" "cg04057818" "cg05440482"
```

Tabla 4. Producto de la intersección de CpG entre de los 26 modelos.

Creemos óptima la utilización del regresor DzBMI por los motivos explicados anteriormente así como el uso del valor de m para el primer filtro. Vamos a modificar solo los parámetros para el *effect size* (segundo filtro). Todas las opciones propuestas parten del siguiente principio:

$$mval\_CpG_i \sim \beta_0 + \beta_1 DzBMI + \beta_2 Gender$$

y obtener un listado de CpGs priorizados al aplicar test estadístico para testear la hipótesis alternativa $\beta 1 \neq 0$. Se aplica FDR 5%.

Como estamos aplicando un LOO externo, obtenemos n listados de CpGs donde para cada uno se ha extraído una de las muestras. El procedimiento es el mismo para todos los casos. La diferencia radica en el siguiente paso: segundo filtro para la selección final de CpGs en base al valor de *beta* o al valor de M.

Proponemos 4 opciones distintas para crear los 26 modelos:

1. **Cutoff $|\Delta\beta| \geq 0.1$** (10%).

   Según la bibliografía [6,7] vemos que el valor de *beta* es el valor estándar para aplicar el *effect size* y creemos que un 10% ya és suficientemente restrictivo para un análisis de metilación. Se

adjuntan los resultados a continuación:

wo_M053_0M	wo_M014_0M	wo_M048_0M	wo_M018_0M	wo_SA12_0M	wo_M013_0M	wo_M052_0M	wo_M031_0M
200	133	241	258	165	204	136	184
wo_M039_0M	wo_M027_0M	wo_M023_0M	wo_M042_0M	wo_SA10_0M	wo_M036_0M	wo_M045_0M	wo_SA32_0M
96	187	216	127	216	145	44	166
wo_M034_0M	wo_M032_0M	wo_M047_0M	wo_SA11_0M	wo_M044_0M	wo_M030_0M	wo_M028_0M	wo_M029_0M
268	201	187	262	178	141	192	151
wo_M041_0M	wo_SA37_0M						
12	221						

Tabla 5. Producto del algoritmo para la creación de los 26 modelos. Listado de 26 modelos con el número de CpG significativos que hay en cada uno. Cada uno lleva el nombre de la muestra no incluida en el modelo. Parámetros = Cutoff |ΔB| ≥ 0.1.

2. **Cutoff |Δβ| ≥ 0.15** (15%).

   Incrementamos la restricción hasta un cutoff del 15%, dejando fuera todos los CpG significativos con un *effect size* del 10 al 15% (no incluido). Seguimos trabajando con los valores de *beta*. Se adjuntan los resultados a continuación:

wo_M053_0M	wo_M014_0M	wo_M048_0M	wo_M018_0M	wo_SA12_0M	wo_M013_0M	wo_M052_0M	wo_M031_0M
43	23	59	50	34	53	27	41
wo_M039_0M	wo_M027_0M	wo_M023_0M	wo_M042_0M	wo_SA10_0M	wo_M036_0M	wo_M045_0M	wo_SA32_0M
23	43	46	25	49	31	12	38
wo_M034_0M	wo_M032_0M	wo_M047_0M	wo_SA11_0M	wo_M044_0M	wo_M030_0M	wo_M028_0M	wo_M029_0M
59	53	38	70	42	33	36	32
wo_M041_0M	wo_SA37_0M						
3	47						

Tabla 6. Producto del algoritmo para la creación de los 26 modelos. Listado de 26 modelos con el número de CpG significativos que hay en cada uno. Cada uno lleva el nombre de la muestra no incluida en el modelo. Parámetros = Cutoff |ΔB| ≥ 0.15.

3. **Cutoff |Δβ| ≥ 0.2** (20%).

   Incrementamos la restricción hasta un cutoff del 20%, dejando fuera todos los CpG significativos con un *effect size* del 10 al 20% (no incluido). Seguimos trabajando con los valores de *beta*. Se adjuntan los resultados a continuación:

wo_M053_0M	wo_M014_0M	wo_M048_0M	wo_M018_0M	wo_SA12_0M	wo_M013_0M	wo_M052_0M	wo_M031_0M
15	9	20	19	12	14	9	11
wo_M039_0M	wo_M027_0M	wo_M023_0M	wo_M042_0M	wo_SA10_0M	wo_M036_0M	wo_M045_0M	wo_SA32_0M
8	13	15	9	16	9	5	7
wo_M034_0M	wo_M032_0M	wo_M047_0M	wo_SA11_0M	wo_M044_0M	wo_M030_0M	wo_M028_0M	wo_M029_0M
19	24	11	23	15	10	12	10
wo_M041_0M	wo_SA37_0M						
2	10						

Tabla 7. Producto del algoritmo para la creación de los 26 modelos. Listado de 26 modelos con el número de CpG significativos que hay en cada uno. Cada uno lleva el nombre de la muestra no incluida en el modelo. Parámetros = Cutoff |ΔB| ≥ 0.20.

4. **Cutoff |ΔM| ≥ 1**.

   En este punto pasamos a utilizar el valor m. No está descrito el uso del valor m para calcular el *effect size*, sin embargo, conocemos que este valor tiene un dominio [-6,6]. Usaremos un valor de 1 como cutoff. Se adjuntan los resultados a continuación:

wo_M053_0M	wo_M014_0M	wo_M048_0M	wo_M018_0M	wo_SA12_0M	wo_M013_0M	wo_M052_0M	wo_M031_0M
79	49	100	106	65	90	49	74
wo_M039_0M	wo_M027_0M	wo_M023_0M	wo_M042_0M	wo_SA10_0M	wo_M036_0M	wo_M045_0M	wo_SA32_0M
37	81	89	44	93	59	16	75
wo_M034_0M	wo_M032_0M	wo_M047_0M	wo_SA11_0M	wo_M044_0M	wo_M030_0M	wo_M028_0M	wo_M029_0M
104	97	71	121	70	61	76	59
wo_M041_0M	wo_SA37_0M						
6	90						

Tabla 8. Producto del algoritmo para la creación de los 26 modelos. Listado de 26 modelos con el número de CpG significativos que hay en cada uno. Cada uno lleva el nombre de la muestra no incluida en el modelo. Parámetros = Cutoff |ΔM| ≥ 1.

2.4 Modelo consenso

En este apartado trataremos de obtener el modelo de predicción final. Nos basaremos en las cuatro opciones propuestas al final del apartado anterior.

Ahora que no hemos tenido en cuenta todas las muestras para hacer el modelo ya estamos en condiciones de aplicar una regresión PLS (Partial Least Squares) para determinar el modelo final.

Si construimos un modelo PLSR utilizando la intersección de las listas del apartado anterior y utilizamos los valores m para los CpG involucrados de la siguiente forma:

$$DzBMI_{excluded\_sample} \sim \beta_0 + \beta_1 CpG_1 + \beta_2 CpG_2 + \cdots + \beta_k CpG_k$$

donde *k* sería el número de CpGs priorizados para cada caso y realizáramos la predicción de la muestra excluida en cada caso obtendríamos cuatro modelos de predicción distintos de los que tendríamos que escoger el óptimo. Vamos a proceder.

- Modelo 1 basado en el cutoff de *beta* del 0.1:

En primer lugar anotamos los CpG que forman la intersección a partir del archivo de anotación que nos proporciona la casa comercial del array (Illumina).

Generamos una tabla con la anotación (**Tabla 9**) en la que mostramos el ID del CpG, el cromosoma donde se encuentra, la posición, el nombre del gen en que se sitúa (si cae en algún gen) y la posición respecto al gen.

	CpGID	CHR	MAPINFO	UCSC_RefGene_Name	UCSC_RefGene_Group
167	cg04057818	2	67487963	LOC102800447	Body
355	cg14157435	2	206628692	NRP2;NRP2;NRP2;NRP2;NRP2	Body;Body;Body;Body;Body
744	cg05387464	2	9956256		
1612	cg00036352	8	144636448	GSDMD	5'UTR
2299	cg16007266	16	57050314	NLRC5	TSS1500
3019	cg08240913	10	117969024	GFRA1;GFRA1;GFRA1	Body;Body;Body
5589	cg23743554	11	65321226	LTBP3;LTBP3;LTBP3	Body;Body;Body
6282	cg18872420	14	78023429	SPTLC2	Body

Tabla 9. Producto de la anotación de los CpG que forman la intersección para los 26 modelos creados a partir del criterio del cutoff |ΔB| ≥ 0.1.

Adjuntamos el código de R para calcular el modelo consenso a partir de estos CpG intersección:

```
#Creamos un objeto con todos los DzBMI
dzb <- data$dzBMI

#El objeto "consensus_betavalue01" contiene los CpG
intersección de los 26 modelos
CpGs_selected <- consensus_betavalue01

#Selecionamos los valores de m de los CpG
intersección
mval_selected <- mvalHRLR[which(rownames(mvalHRLR)
%in% CpGs_selected),]

#Juntamos los CpG con "+" para crear el modelo
selection <- paste(CpGs_selected,collapse='+')

#Creamos la equación del modelo
f_selected <- formula(paste("dzb ~ ",selection))

#Creamos el modelo PLS con validación LOO y
jakknife*
modpls_selected <- plsr(f_selected,
data=as.data.frame(t(mval_selected)), validation =
"LOO",jackknife=TRUE)
```

```
#Realizamos un test jackknife del mismo paquete pls

jack.test(modpls_selected, ncomp=1)
```

NOTA: La función "jack.test" realiza pruebas t aproximadas de los coeficientes de regresión basados en estimaciones de varianza jackknife. [13]

Para el modelo 1 obtenemos la siguiente curva ROC*:

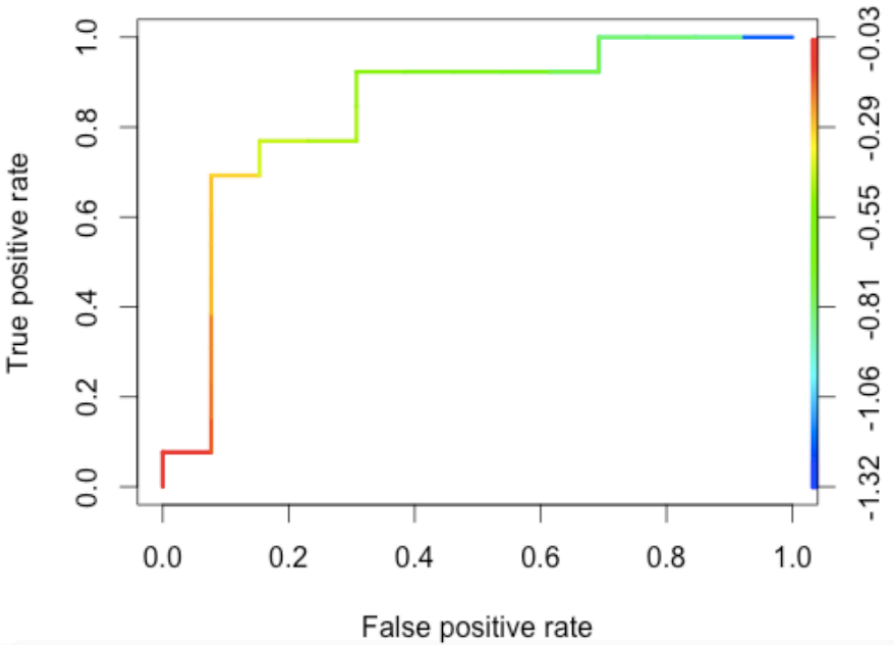

Figura 10. Curva ROC para el modelo de 8 CpG obtenidos con cutoff |ΔB| ≥ 0.1.

En el ámbito sanitario, las curvas ROC también se denominan **curvas de rendimiento diagnóstico**. [14]

El AUC* para este modelo (**Figura 10**) es del 84% y el cutoff DzBMI óptimo se encuentra en -0.426 (para este valor sensibilidad* del 77% y especificidad* del 85%).

Aquí mostramos los resultados del modelo explicativo PLSR para un componente (que presenta el menor valor de PRESS).

```
dzb ~ cg14157435 + cg00036352 + cg05387464 + cg08240913 + cg18872420 +
    cg16007266 + cg23743554 + cg04057818
```

```
> modpls_selected$validation$PRESS
     1 comps  2 comps 3 comps  4 comps  5 comps  6 comps 7 comps  8 comps
dzb 1.939827 2.263937 2.69263 2.816453 2.815648 2.874687 2.88551 2.888362
> jack.test(modpls_selected, ncomp=1)
Response dzb (1 comps):
             Estimate Std. Error Df t value  Pr(>|t|)
cg14157435 -0.0798113  0.0083020 25 -9.6135 7.062e-10 ***
cg00036352 -0.0771572  0.0178811 25 -4.3150 0.0002199 ***
cg05387464 -0.0487775  0.0066348 25 -7.3517 1.058e-07 ***
cg08240913 -0.0495441  0.0096373 25 -5.1409 2.590e-05 ***
cg18872420 -0.0676075  0.0189271 25 -3.5720 0.0014732 **
cg16007266 -0.0841298  0.0149264 25 -5.6363 7.257e-06 ***
cg23743554 -0.0583814  0.0108829 25 -5.3645 1.456e-05 ***
cg04057818  0.0817991  0.0163676 25  4.9976 3.750e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Modelo 2 basado en el cutoff de *beta* del 0.15:

Identificamos los CpG intersección de la **Tabla 6** y generamos la tabla de anotaciones (**Tabla 10**) en la que mostramos el ID del CpG, el cromosoma donde se encuentra, la posición, el nombre del gen en que se sitúa (si cae en algún gen) y la posición respecto al gen.

	CpGID	CHR	MAPINFO	UCSC_RefGene_Name	UCSC_RefGene_Group
355	cg14157435	2	206628692	NRP2;NRP2;NRP2;NRP2;NRP2	Body;Body;Body;Body;Body
2299	cg16007266	16	57050314	NLRC5	TSS1500

Tabla 10. Producto de la anotación de los CpG que forman la intersección para los 26 modelos creados a partir del criterio del cutoff |ΔB| ≥ 0.15.

Para el siguiente modelo PLS construido con la intersección (2 CpG):

```
dzb ~ cg14157435 + cg16007266
```

Obtenemos la siguiente curva ROC (**Figura 11**):

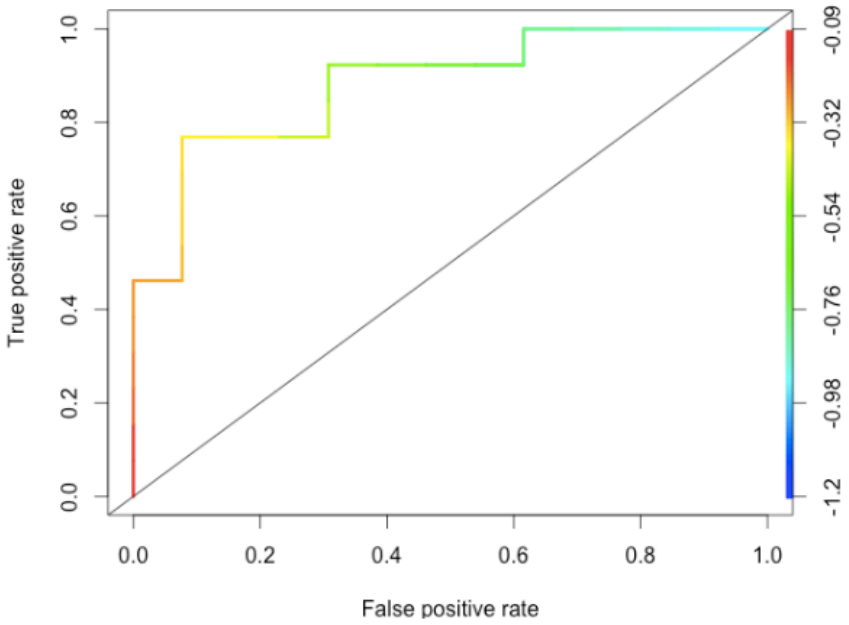

Figura 11. Curva ROC para el modelo de 2 CpG obtenidos con cutoff |ΔB| ≥ 0.15.

El AUC del modelo (**Figura 11**) es del 88% y el cutoff DzBMI óptimo se encuentra en -0.358 (para este valor sensibilidad del 77% y especificidad del 92%).

Y ahora el modelo explicativo PLSR con este listado reducido (para ncomp=1 que presenta el menor valor de PRESS, los valores son altos comparados con el anterior).

```
> modpls_selected$validation$PRESS
     1 comps  2 comps
dzb 2.984965 3.209845
> jack.test(modpls_selected, ncomp=1)
Response dzb (1 comps):
            Estimate Std. Error Df t value  Pr(>|t|)
cg14157435 -0.180143   0.032638 25 -5.5194 9.785e-06 ***
cg16007266 -0.189890   0.040083 25 -4.7375 7.358e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Modelo 3 basado en el cutoff de *beta* del 0.20:

La intersección de CpG para la realización de este modelo (**Tabla 7**) coincide exactamente con los mismos CpG que el modelo anterior (**Tabla 10**).

Al obtener la misma intersección hemos de realizar un modelo a partir de los mismos CpG, por lo que obtenemos exactamente los mismos resultados que en el **modelo 2**.

La restricción para el *effect size* de |ΔB| ≥ 0.15 y |ΔB| ≥ 0.20 son equivalentes desde el punto de vista de la creación de un modelo consenso.

- Modelo 4 basado en el cutoff de m ≥ 1:

Identificamos los CpG intersección de la **Tabla 8** y generamos la tabla de anotaciones (**Tabla 11**) en la que mostramos el ID del CpG, el cromosoma donde se encuentra, la posición, el nombre del gen en que se sitúa (si cae en algún gen) y la posición respecto al gen.

	CpGID	CHR	MAPINFO	UCSC_RefGene_Name	UCSC_RefGene_Group
167	cg04057818	2	67487963	LOC102800447	Body
355	cg14157435	2	206628692	NRP2;NRP2;NRP2;NRP2;NRP2	Body;Body;Body;Body;Body
1612	cg00036352	8	144636448	GSDMD	5'UTR
2299	cg16007266	16	57050314	NLRC5	TSS1500

Tabla 11. Producto de la anotación de los CpG que forman la intersección para los 26 modelos creados a partir del criterio del cutoff |ΔM| ≥ 1.

Para el siguiente modelo PLS construido con la intersección (4 CpG):

```
dzb ~ cg14157435 + cg00036352 + cg16007266 + cg04057818
```

Obtenemos la siguiente curva ROC (**Figura 12**):

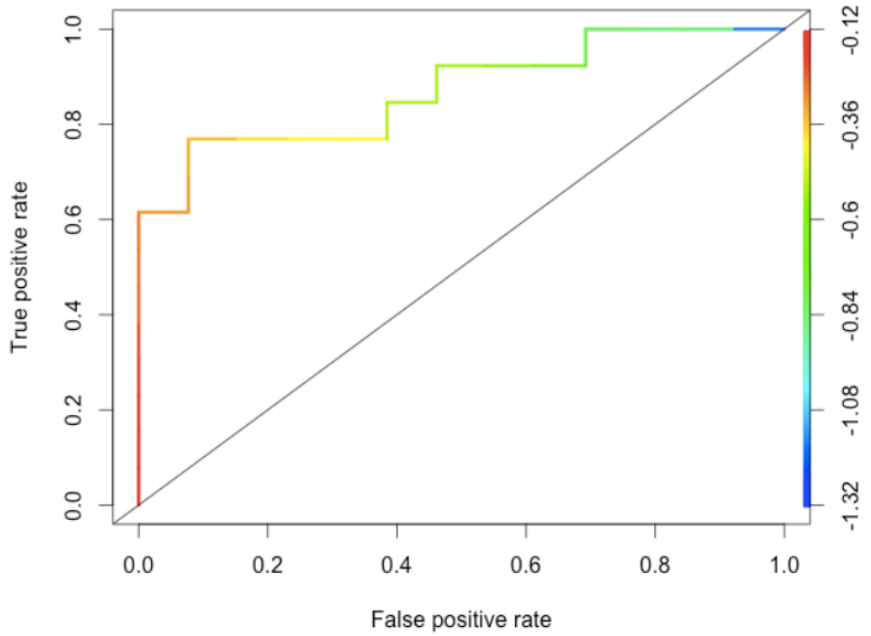

Figura 12. Curva ROC para el modelo de 4 CpG obtenidos con cutoff |ΔM| ≥ 1.

Donde el AUC (**Figura 12**) es del 87% y el cutoff DzBMI óptimo se encuentra en -0.354 (para este valor sensibilidad del 77% y especificidad del 92%).

Y para finalizar, el modelo explicativo PLSR con este listado reducido (para ncomp=1, que tiene el menor valor de PRESS).

```
> modpls_selected$validation$PRESS
     1 comps  2 comps  3 comps  4 comps
dzb 2.099961 2.316893 2.418547 2.443645
> jack.test(modpls_selected, ncomp=1)
Response dzb (1 comps):
            Estimate Std. Error Df t value  Pr(>|t|)
cg14157435 -0.114174   0.014945 25 -7.6395 5.389e-08 ***
cg00036352 -0.110378   0.023087 25 -4.7810 6.572e-05 ***
cg16007266 -0.120352   0.023414 25 -5.1401 2.595e-05 ***
cg04057818  0.117018   0.022191 25  5.2733 1.841e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Una vez obtenidos los 4 modelos (el primero con 8 predictores, el segundo y tercero con 2 y el cuarto con 4) vemos que todos los

predictores obtenidos en los modelos 2, 3 y 4 son subgrupos de los 8 predictores del modelo 1.
Si nos fijamos en el valor de AUC (que nos indica "como de bueno es" nuestro modelo de predicción) vemos que en los tres modelos (ya que el 2 y el 3 son el mismo) este valor se sitúa en torno al 85% (100% = modelo de predicción "perfecto").

Los valores de sensibilidad de los tres modelos coinciden en el 77% mientras que la especificidad tiene valores entre 85% (el más bajo) hasta 92% (los otros dos).

Aunque el valor de AUC para el primer modelo sea el más bajo, con un 84% (aún así, considerado un buen modelo de predicción) y la especificidad sea del 85% (también menor respecto a los otros dos modelos), consideramos óptimo el primer modelo ya que creemos que solo 2 o 4 predictores (para todo el genoma) son muy pocos, dada la variabilidad en el epigenoma durante etapas de desarrollo. Así mismo, consideramos óptimo un número de 8 predictores CpG para la fabricación de un kit - diagnóstico en términos efectividad - precio (más de 10-20 predictores encarecerían mucho el producto para la mejora de efectividad que supondrían).

En conclusión, nuestro modelo de predicción final para predecir el valor del DzBMI en función de los valores de m de los 8 CpG escogidos, en sujetos prepuberales con obesidad, sigue la siguiente equación:

$$DzBMI \approx (-0.0798113m1) + (-0.0771572m2) + (-0.0487775m3) + (-0.0495441m4) + (-0.0676075m5) + (-0.0841298m6) + (-0.0583814m7) + (0.0817991m8)$$

Este modelo de predicción tiene un valor AUC del 84%, una sensibilidad del 77% y una especificidad del 85% para el cutoff de DzBMI de -0.426, que nos servirá para discriminar los individuos en HR o LR, es decir, para realizar el diagnóstico.

2.5 SPLS

Este breve apartado contiene dos modelos SPLS para nuestros 214 CpG significativos (según etiqueta HR/LR) usando el DzBMI como regresor.

Un modelo se ha realizado con el paquete "mixOmics" y el otro con el paquete "spls". Comparamos los predictores óptimos obtenidos con los dos modelos SPLS entre ellos y entre los predictores de nuestro modelo consenso.

- Modelo SPLS con el paquete de R "mixOmics"

Adjuntamos el código R para la obtención de los predictores óptimos con mixOmics:

```
#Cargamos el paquete
library(mixOmics)

#Probamos el modelo inicialmente con 10 comp.
ncomp = 10

#spls con 10 componentes
res.spls<-spls(beta_sig,dzBMI,ncomp = ncomp, keepX =
c(rep(10, ncomp)), mode = 'regression')

#Mfold validation (x10 folds)
tune.spls <- perf(res.spls, validation = 'Mfold',
folds = 10, criterion = 'all', progressBar = TRUE)
```

Obtenemos un valor de PRESS mínimo para 5 componentes (**Figura 13**) y graficamos los *loadings* (variables) (**Figura 14**).

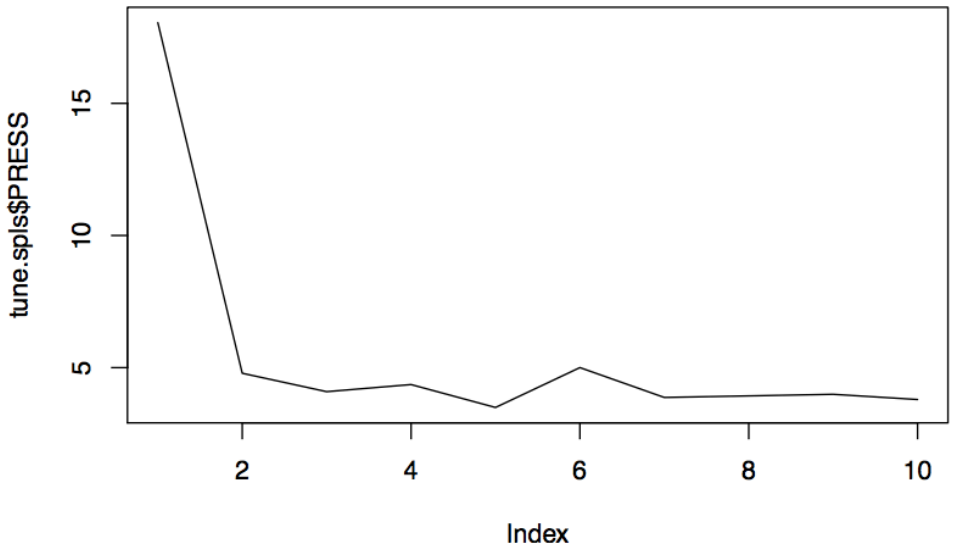

Figura 13. Curva del valor PRESS para el modelo SPLS realizado con el paquete de R "spls". Vemos un valor mínimo en 5 componentes.

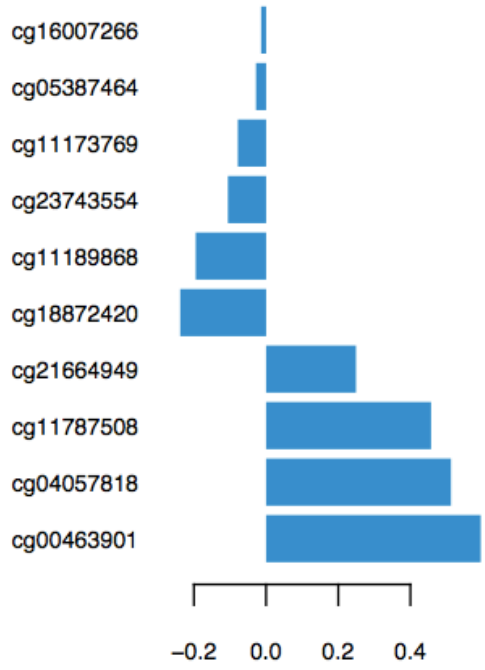

Figura 14. Loadings para el modelo SPLS realizado con el paquete de R "spls". Vemos los 10 mejores predictores.

Vemos como 5 de las variables que aparecen en el gráfico “loadings” coinciden con el modelo de 8 CpG realizado con limma. Las variables coincidentes son: ***cg05387464, cg18872420, cg16007266, cg23743554*** *y* ***cg04057818.***

- Modelo SPLS con el paquete de R "spls"

Realizamos un nuevo modelo SPLS utilizando el paquete de R "spls". Adjuntamos el código a continuación:

```
#Cargamos el paquete
library(spls)

#Realizamos el modelo usando 5 componentes ya que
hemos visto en el moselo mixOmics que corresponden
al menor valor de PRESS
f<-spls(beta_sig,dzBMI,K=5,eta=0.95);f
```

```
  SPLS chose 12 variables among 214 variables


  Selected variables:
  cg05401945 cg23118773 cg12119625 cg18872420
  cg21664949 cg16613240 cg00463901 cg03047376
  cg11787508 cg10062460 cg04057818 cg09357926
```

Vemos como el paquete spls nos selecciona 12 predictores de los 214 CpG significativos. Ahora, 2 de las 12 variables seleccionadas coinciden con nuestro modelo de 8 CpG realizado con limma. Las variables coincidentes son: ***cg18872420*** *y* ***cg04057818.***

Observamos como el overlap entre los 2 modelos SPLS es de 5 predictores: ***cg21664949***, ***cg18872420***, ***cg00463901***, ***cg11787508*** *y* ***cg04057818.***

Consideramos que estos resultados obtenidos respaldan en gran parte nuestro modelo de regresión limma con 8 CpG predictores.

2.6 Enriquecimiento

Este apartado contiene la parte de significación biológica del trabajo. A continuación llevaremos a cabo la anotación de nuestros 214 CpG diferencialmente metilados (según la etiqueta HR/LR) y el posterior enriquecimiento de los genes que corresponden con las posiciones de estos CpG.

Para el enriquecimiento de genes hemos utilizado el paquete de R SIGORA [9]. Este paquete nos permite enriquecer genes en las bases de datos KEGG y REACTOME para el genoma humano usando el mismo pipeline.

Primero hemos anotado los genes que coinciden con las posiciones de nuestros CpG significativos a partir de la tabla de anotaciones que nos proporciona la casa comercial del array (https://support.illumina.com/array/array_kits/infinium- methylationepic-beadchip-kit/downloads.html).

Obtenemos una tabla de anotación (**Tabla 12**) para nuestros 214 CpG. La tabla contiene el ID del CpG, el cromosoma, la posición, el nombre del gen en que se sitúa (si cae en algún gen), la referencia del gen y la posición respecto al gen:

IlmnID	CHR	MAPINFO	UCSC_RefGene_Name	UCSC_RefGene_Accession	UCSC_RefGene_Group
cg14157435	2	206628692	NRP2;NRP2;NRP2;NRP2	NM_201279;NM_018534;N	Body;Body;Body;Body;I
cg05387464	2	9956256			
cg09127314	1	152161683			
cg12372706	2	107592370			
cg03963391	16	34808479			
cg06758833	16	73575220			
cg11732055	10	91482145	KIF20B;KIF20B	NM_016195;NM_00128425	Body;Body
cg08693938	14	73439289	ZFYVE1;ZFYVE1;ZFYVE1;	NM_178441;NM_00128173	Body;Body;Body;Body
cg22166084	17	3602036			
cg09764652	14	55712439			
cg16081325	12	106517443	NUAK1	NM_014840	Body
cg20163335	20	61309524			
cg06538684	12	12511223	LOH12CR2;LOH12CR1	NR_024061;NM_058169	TSS1500;Body
cg07684019	5	139174529	PSD2	NM_032289	TSS1500
cg21664949	10	118015682	GFRA1;GFRA1;GFRA1	NM_145793;NM_005264;N	Body;Body;Body
cg12271079	10	117997745	GFRA1;GFRA1;GFRA1	NM_001145453;NM_00526	Body;Body;Body
cg12426470	17	1104627			
cg04109556	13	19185909			
cg15022308	10	43447071			
cg07114310	7	51538921			
cg18056749	20	55836268	BMP7	NM_001719	Body
cg04057818	2	67487963	LOC102800447	NR_110564	Body
cg27336068	10	1336103	ADARB2	NM_018702	Body
cg26173906	19	14376389			
cg06565913	16	1584452	IFT140;TMEM204	NM_014714;NM_024600	Body;1stExon
cg00846098	21	43198791			
cg14920334	12	4829827	GALNT8;GALNT8	NM_017417;NM_017417	1stExon;5'UTR
cg05630272	10	1517218	ADARB2	NM_018702	Body
cg02602925	6	171019460			
cg16322792	1	120165303	ZNF697	NM_001080470	3'UTR
cg09548084	6	8436218	SLC35B3;SLC35B3;SLC3	NM_015948;NM_00114254	TSS1500;TSS1500;TSS1
cg14210765	18	73628830			
cg24989447	22	31730238	PIK3IP1-AS1;PATZ1;PAT	NR_110542;NM_014323;NM	TSS1500;Body;Body;Bo
cg11121987	11	132940088	OPCML	NM_001012393	Body
cg20992785	6	168533507			
cg01669927	12	85152040			
cg08394602	19	22123885			
cg10807027	2	206628773	NRP2;NRP2;NRP2;NRP2	NM_201267;NM_018534;N	Body;Body;Body;Body;I
cg04738746	10	78780052	KCNMA1;KCNMA1;KCN	NM_002247;NM_00116135	Body;Body;Body;Body;I

Tabla 12. Tabla de anotación para nuestros 214 CpG significativos.

Una vez anotados y separados los genes, hemos obtenido el valor Entrez* de cada gen (único para cada uno) usando el paquete de R "org.Hs.eg.db ". Posteriormente enriquecemos los valores de Entrez (equivalentes a los genes) con el paquete SIGORA en las dos bases de datos mencionadas anteriormente usando un p-valor de 0.05 FDR. Adjuntamos el código de enriquecimiento:

```
#Separamos los genes en las columnas
"UCSC_RefGene_Name" y "UCSC_RefGene_Accession"
d <- cbind.data.frame(GeneSymbol =
unlist(strsplit(as.character(data$UCSC_RefGene_Name)
,";")),
                      GeneAcc =
unlist(strsplit(as.character(data$UCSC_RefGene_Acces
sion),";")), stringsAsFactors = FALSE)
```

```
#Anotamos el valor Entrez (único para cada gen) a
partir del "UCSC_RefGene_Accession" y el paquete
"org.Hs.eg.db"
require(org.Hs.eg.db)

# Eliminamos los genes que no existen en la base de
datos del paquete
d2 <- d[d$GeneAcc %in% ls(org.Hs.egACCNUM2EG),]

#Obtenemos el valor Entrez a partir de GeneAcc
Entrez <- unlist(mget(d2$GeneAcc,
org.Hs.egACCNUM2EG))
d2 <- cbind(d2,Entrez,stringsAsFactors =FALSE)

### Enriquecimiento de genes - SIGORA test method
#KEGG PATHWAY database

library(sigora)
sig_KEGGresults <- sigora(GPSrepo = kegH, level = 2,
markers = TRUE, queryList = unique(d2$Entrez),
saveFile=NULL, weighting.method = "invhm")

#REACTOME database

sig_REACTOMEresults<-sigora(GPSrepo = reaH, level =
4, markers = TRUE, queryList = unique(d2$Entrez),
saveFile=NULL, weighting.method = "invhm")
```

Para nuestra sorpresa, explorando la base de datos KEGG no obtenemos ningún *pathway* sobrerrepresentado (**Tabla 13**).

Por otro lado, en la base de datos REACTOME solo encontramos un *pathway* significativo (**Tabla 14**): "***Endosomal/Vacuolar pathway***".

description	pvalues	Bonferroni	successes
Basal cell carcinoma	0.0002388	0.06686	3.00
Natural killer cell mediated cytotoxicity	0.0031480	0.88140	3.72
Cell adhesion molecules (CAMs)	0.0090980	1.00000	3.62
cGMP-PKG signaling pathway	0.0217200	1.00000	2.17

Tabla 13. Resultado del enriquecimiento en la base de datos KEGG. No obtenemos ningún pathway sobrerrepresentado.

description	pvalues	Bonferroni	successes
Endosomal/Vacuolar pathway	1.785e-27	1.571e-24	15.00
RMTs methylate histone arginines	8.958e-04	7.883e-01	2.00
Defective B4GALT7 causes EDS, progeroid type	3.636e-03	1.000e+00	1.00

Tabla 14. Resultado del enriquecimiento en la base de datos REACTOME. Obtenemos solo un pathway sobrerrepresentado, "*Endosomal/Vacuolar pathway*".

Nos encontramos delante de una vía metabólica muy genérica, implicada en muchos procesos biológicos, por lo que nos va a dificultar notablemente hacer una hipótesis explicativa del fenotipo en torno a este pathway.

En este punto finaliza la parte bioinformática y estadística del trabajo. A continuación vamos a añadir un breve apartado que pretende aportar una significación biológica a los resultados anteriores.

2.7 Discusión

Respecto a los resultados previstos del enriquecimiento de genes obtenemos un resultado un poco diferente al esperado. En un caso óptimo, al enriquecer los genes que coinciden con las posiciones de los CpG diferencialmente metilados, esperaríamos obtener un buen número de *pathways* relacionados con el fenotipo del estudio, que en nuestro caso, es la obesidad infantil. Lo ideal sería encontrar *pathways* relacionados con la síntesis de lípidos, distribución de ácidos grasos, metabolismo de la glucosa, gestión de las reservas, etc.

A diferencia de eso, para la base de datos KEGG no hemos obtenido ningún *pathway* sobrerrepresentado y para la base de datos REACTOME solo hemos obtenido uno significativo (en los dos análisis se ha aplicado la corrección de Bonferroni).

Es complicado dar una explicación biológica del fenotipo en base a nuestro resultado. El *pathway* que obtenemos es extremadamente genérico y no tiene una vinculación clara con nuestro fenotipo.

Probablemente obtenemos este resultado porque hemos sido muy restrictivos a la hora de escoger los CpG significativos, por lo que nuestros genes a enriquecer tienen una *n* demasiado pequeña. Nuestra mejor opción sería disminuir la restricción del segundo filtro (*effect size*) al 5% o incluso eliminarlo, teniendo en cuenta solo el p-valor < 0.05 (FDR).

De ésta manera obtendríamos un listado de genes a enriquecer mucho mayor y probablemente, muchos más *pathways* sobrerrepresentados.

Por otro lado, tenemos otra explicación alternativa a la de tener una lista demasiado reducida de genes. Al tratar con datos epigenéticos, se nos plantea la segunda explicación.

Podemos poner en duda el efecto de la metilación sobre el gen en que se encuentra (no sobre el fenotipo). De este modo explicaríamos que la

metilación diferencial en nuestro modelo no está vinculada a la expresión de los genes en que se encuentra y en consecuencia tampoco a los *pathways* relacionados con dichos genes.

La metilación se encuentra en una "dimensión" distinta a la expresión, es decir, no siempre está relacionada. Hay que tener en cuenta que la metilación en un punto del genoma puede afectar a otros puntos enormemente distales, incluso en otros cromosomas (vía *enhancers*, por ejemplo).

Es un hecho que la metilación afecta a la expresión de genes, lo que no esta claro en muchas ocasiones es "cómo?". Y aún menos cuando hablamos de CpG sites aislados en el genoma y no de una agregación de nucleótidos metilados (Differential Methylated Region).

Estamos convencidos que la metilación es relevante en nuestro fenotipo. Sin embargo, hace falta profundizar en el estudio para averiguar como está contribuyendo, ya que probablemente el efecto de la metilación sobre nuestro fenotipo no sea directa, es decir, no sea mediante expresión de genes y consecuentemente, *pathways* alterados.

Por ejemplo, podría contribuir impidiendo o dificultando la unión de algún factor de transcripción que afecte la expresión de genes que se encuentran a kb de distancia (por tanto no presentes en nuestra anotación), podría afectar dificultando la unión de alguna histona al DNA, alterando algún patrón conformacional, afectando algún *enhancer*, etc.

Viabilidad del producto:

Hemos de tener muy claro que el modelo consenso de predicción obtenido consiste (de momento) en un modelo explicativo exclusivamente de nuestra muestra hasta su validación en una cohorte externa.

Hay que tener en cuenta la enorme variabilidad en el epigenoma durante las fases de desarrollo (como es nuestro caso), por lo que no podemos asegurar el éxito de nuestro modelo en otras cohortes.

Aún así, la validación externa es el siguiente paso. En el peor de los casos, si el modelo no predijera bien en otras cohortes de niños/as obesos/as, calificaríamos nuestro modelo como un "modelo de predicción exclusivo para clasificar los individuos de nuestra cohorte".

Por otro lado, si la validación se confirmara (lógicamente este es el objetivo esperado) tendríamos un "modelo de predicción del DzBMI para niños y niñas prepuberales con obesidad severa" (modelo general).

3. Conclusiones

3.1 Lecciones aprendidas

Este proyecto de fin de master me ha servido para mejorar enormemente mis conocimientos de estadística (principalmente) y de bioinformática (parte de anotaciones y enriquecimiento de genes).

Creo que ha sido muy importante también para mejorar mi aprendizaje autónomo, dándome la oportunidad de aportar ideas, llevarlas a cabo y hablar con expertos e incluso trabajar con ellos.

Hago una valoración muy positiva de este trabajo ya que ha sido una gran oportunidad y una gran experiencia dentro de un ámbito relativamente nuevo para mi, como es la biocomputación.

Me he sentido muy motivado a lo largo de todo el proyecto por la idea de trabajar en un producto final con una posible aplicación real en la práctica clínica.

Finalmente, puntualizar de nuevo que somos conscientes de que está pendiente la validación del modelo en una cohorte externa, y que hasta el momento de su validación, lo habremos de considerar un modelo de predicción explicativo exclusivo de nuestra muestra.

3.2 Cumplimiento de los objetivos

Creemos haber logrado el objetivo principal del proyecto, que consistía en el desarrollo del modelo de predicción. Conscientes de que disponíamos de una muestra más bien pequeña para lograr el objetivo marcado, creemos que hemos hecho un buen trabajo con los medios de los que disponíamos y hacemos una valoración muy positiva de ello.

A parte de la valoración positiva que hacemos para nuestro modelo en

particular, también estamos muy satisfechos con la metodología usada para lograr el objetivo y creemos que puede ser un buen método para el desarrollo de modelos de predicción en muestras relativamente pequeñas.

Sin embargo, por lo que respecta al segundo objetivo no podemos decir lo mismo. No hemos podido explicar nuestro fenotipo a partir de nuestro patón epigenético con nuestro análisis de enriquecimiento.

Hemos obtenido muchos menos *pathways* de los esperados, y los que hemos obtenido (solo uno) no se relaciona (al menos directamente) con nuestro fenotipo.

3.3 Seguimiento y planificación

A lo largo de todo el trabajo no ha habido ninguna desviación en la temporalización. Tanto en la PEC 2 como en la PEC 3 se han entregado todos los productos marcados en la planificación además de las actividades no planificadas realizadas.

Hemos mantenido el cronograma durante todo el proyecto ya que lo hemos podido seguir correctamente.

3.3.1 Actividades no planificadas

En primer lugar, para poder trabajar con los valores de m y de *beta* que nos devuelve el array necesitábamos saber cómo se calculan, cómo se comportan, la relación entre ellos, sus características, etc.

Aunque no habíamos incluido esta parte (más teórica, pero imprescindible) en la planificación, fue crucial para poder entender los datos y escoger con criterio qué valor utilizar.

En segundo lugar, sabemos que nuestros individuos se dividen en dos

grupos (HR/LR) en función del DzBMI.

Para realizar la parte de los modelos (exclusivamente), nos planteamos utilizar la diferencia de BMI corregida (DzBMI) como regresor en vez de la etiqueta HR/LR (que se determina a partir del DzBMI). Esto se debe a que el regresor DzBMI es una variable continua que tiene más información que la variable HR/LR, que corresponde a la discretización en dos estados de DzBMI.

Por último, también hemos incluido dos análisis "spls" para nuestros CpG significativos con la finalidad de hacer una selección de variables externa a nuestro modelo limma LOO y posteriormente comparar si hay intersección de estas variables seleccionadas con las variables que escogimos para hacer el modelo.

3.4 Líneas de trabajo futuro

1. Validación de nuestro modelo en la misma cohorte extendida y en cohortes externas.

En el caso de una validación positiva pasaríamos a considerar nuestro modelo como un modelo general. En caso contrario, éste quedaría como un modelo de predicción explicativo exclusivo de nuestra muestra.

2. Análisis de enriquecimiento con más genes mediante una bajada de restricción en el segundo filtro (*effect size*).

Al disminuir la restricción del segundo filtro o incluso eliminarlo, obtenemos una lista mucho más extensa de CpG, y en consecuencia, de genes.

Si enriquecemos estos nuevos genes es muy probable que obtengamos resultados mucho más buenos que en el

enriquecimiento realizado. De este modo, obtendremos una visión más amplia y nos será más fácil razonar diferentes hipótesis que expliquen nuestros fenotipos.

3. Posible validación de los resultados del punto anterior en el laboratorio (mediante PCR, Western Blot…).

Dependiendo del resultado anterior, en el laboratorio podríamos hacer alguna validación de expresión de algunos genes, interacciones proteína-proteína, etc.

4. Glosario

- **CpG:** Regiones del DNA donde una citosina es seguida por una guanina en la secuencia lineal de bases a lo largo de su dirección 5 '→ 3'.

- **Metilación:** Proceso por el cual se añaden grupos metilo al DNA.

- **Epigenética:** Estudio de los factores que, sin corresponderse a elementos de la genética clásica, básicamente los genes, juegan un papel muy importante en la genética moderna interaccionando con estos primeros. Pueden ser elementos como la metilación o acetilación.

- **Curva ROC:** Representación gráfica de la sensibilidad frente a la especificidad para un sistema clasificador binario según se varía el umbral de discriminación.

- **AUC:** Área bajo la curva ROC.

- **Sensibilidad:** Capacidad de un estimador (modelo) para dar como casos positivos los casos realmente positivos. Se denominan verdaderos positivos.

- **Especificidad:** Capacidad de nuestro estimador (modelo) para dar como casos negativos los casos realmente negativos. Se denominan verdaderos negativos.

- **Jackknife:** Técnica de muestreo especialmente útil para estimar la varianza y el sesgo. El estimador Jackknife de un parámetro se encuentra dejando sistemáticamente cada observación a partir de un conjunto de datos y el cálculo de la estimación y luego encontrar el promedio de estos cálculos.

- **Entrez value:** Valor único para cada gen. Permite acceder a la base de datos del National Center for Biotechnology Information (NCBI).

5. Bibliografía

(1) Law C, Cole T, Cummins S, et al. *A pragmatic evaluation of a family-based intervention for childhood overweight and obesity.* Public Health Research, No. 2.5. NIHR Journals Library , Southampton (UK) , 2014. [Apendix 10]

(2) Law C, Cole T, Cummins S, et al. *A pragmatic evaluation of a family-based intervention for childhood overweight and obesity.* Public Health Research, No. 2.5. NIHR Journals Library , Southampton (UK) , 2014. [Chapter 3]

(3) Butcher, L. M., & Beck, S. (2015). Probe Lasso: a novel method to rope in differentially methylated regions with 450K DNA methylation data. *Methods*, *72*, 21-28.

(4) Katzmarzyk, P. T., Barlow, S., Bouchard, C., Catalano, P. M., Hsia, D. S., Inge, T. H., ... & Spruijt-Metz, D. (2014). An evolving scientific basis for the prevention and treatment of pediatric obesity. *International journal of obesity*, *38*(7), 887-905.

(5) 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.

(6) Zhang, X., Justice, A. C., Hu, Y., Wang, Z., Zhao, H., Wang, G., ... & Xu, K. (2016). Epigenome-wide differential DNA methylation between HIV-infected and uninfected individuals. *Epigenetics*, *11*(10), 750-760.

(7) Joubert, B. R., Håberg, S. E., Nilsen, R. M., Wang, X., Vollset, S. E., Murphy, S. K., ... & Ueland, P. M. (2012). 450K epigenome-wide scan identifies differential DNA methylation in newborns related to maternal smoking during pregnancy.

(8) https://es.wikipedia.org/wiki/Validación_cruzada [16/5/2017]

(9) https://cran.r-project.org/web/packages/sigora/sigora.pdf [10/04/2017]

(10) Jin, X., Liu, X., Li, X., & Guan, Y. (2016). Integrated Analysis of DNA Methylation and mRNA Expression Profiles Data to Identify Key Genes in Lung Adenocarcinoma. *BioMed Research International*, *2016*.

(11) Ung, M. H., Varn, F. S., Lou, S., & Cheng, C. (2015). Regulators associated with clinical outcomes revealed by DNA methylation data in breast cancer. *PLoS Comput Biol*, *11*(5), e1004269.

(12) Du, P., Zhang, X., Huang, C.C. et al. *Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis*. BMC Bioinformat. 2010;11:587.

(13) https://cran.r-project.org/web/packages/pls/pls.pdf [16/03/2017]

(14) https://es.wikipedia.org/wiki/Curva_ROC#Curvas_ROC_para_pruebas_diagn.C3.B3sticas [16/03/2017]

6. Anexos

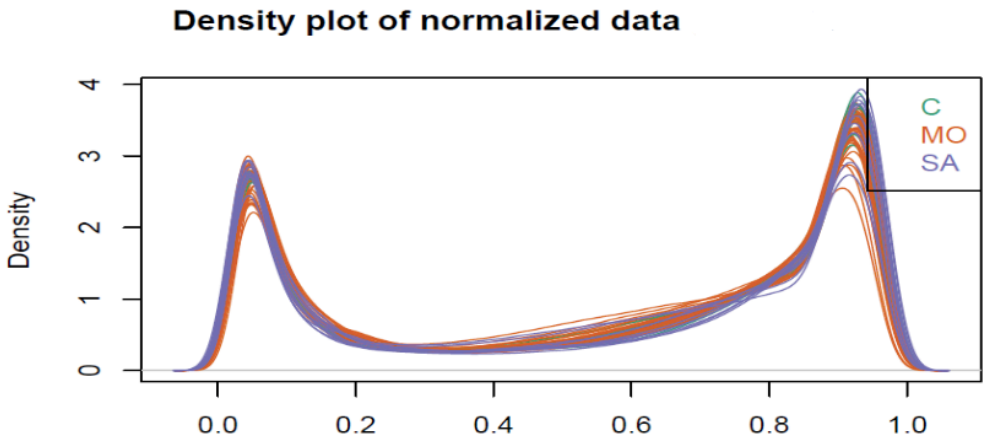

Figura 15. Gráfico de densidad de los datos normalizados. Figura proporcionada por la empresa encargada del filtraje y normalización Making Genetics. En el eje horizontal encontramos los valores de *beta* y en el vertical la densidad. Podemos ver la distribución de rangos de nuestros datos.