Trabajo Epigenómica: Estado 1 (H3K4me3, H3K27Ac)

# Autores:

Sara Dorado Alfaro, Diego Mañanes Cayero, Alejandro Martín Muñoz, Álvaro Huertas García

**Contenido**

[Autores: 1](#_Toc37235849)

[Objetivo 2](#_Toc37235850)

[Antecedentes 2](#_Toc37235851)

[Análisis 3](#_Toc37235852)

[Paso 1: Obtener los segmentos que tengan el mismo estado en los dos replicados de monocitos. 3](#_Toc37235853)

[Paso 2: Paso 2: Anotar los segmentos 4](#_Toc37235854)

[Paso 3: Calcular el porcentaje de solapamiento entre DNaseI-peaks y vuestros segmentos de trabajo. 9](#_Toc37235856)

[Paso 4: Visualizar una región del genoma en el *UCSC browser*. 10](#_Toc37235857)

[Paso 5: Búsqueda de motivos enriquecidos. 13](#_Toc37235859)

[Paso 6: Regiones hiper e hipometiladas 15](#_Toc37235861)

[Anexo: Material adicional 18](#_Toc37235863)

[Anexo: Análisis estadístico de los solapamientos 21](#_Toc37235864)

[Bibliografía 22](#_Toc37235865)

# Objetivo

En este trabajo el objetivo es analizar y estudiar el estado 1 de cromatina asignado a dos modificaciones de la histona 3 (H3), la trimetilación en la lisina 4 (H3K4me3) y la acetilación de la lisina 27 (H3K27Ac).

# Antecedentes

En este trabajo, el material de partida son los segmentos de cromatina asignados a uno de los 11 estados calculados por ChromHMM, como se realizó en la práctica del día 27 de febrero en el aula. El software ChromHMM emplea los modelos ocultos de Markov (“Hidden Markov Models, HMM) para calcular distintos estados de cromatina, cada uno de ellos caracterizado por la combinación de distintas marcas epigenéticas (Ernst and Kellis, 2017). En nuestro caso, el estado de estudio es el estado 1, en adelante E1, que se caracteriza por la combinación de la trimetilación en la lisina 4 (H3K4me3) y la acetilación de la lisina 27 (H3K27ac) de la histona 3.

Las regiones del DNA con un nivel de metilación bajo (24%) se encuentran enriquecidas en la modificación de histonas H3K4me3. Del mismo modo, las modificaciones H3K4me3 y H3K27Ac, junto a otras modificaciones como H3K4me1, H3K9Ac, son consideradas marcas de cromatina activas, correlacionadas de forma negativa con la hipermetilación del DNA (Sharifi-Zarchi et al., 2017).

En lo que se refiere a la modificación H3K27Ac, se localiza en los enhancers promoviendo la activación de la transcripción, siendo interesante destacar que esta modificación de histona se asocia con la presencia de super enhancers (clústeres de enhancers que inducen una fuerte respuesta transcipional) (Benetatos and Vartholomatos, 2018).

Con respecto a la modificación H3K4me3, ésta se asocia a regiones promotoras con bajos niveles de metilación. Algunos autores (Sharifi-Zarchi et al., 2017) indican la importancia de esta modificación de histonas a la hora de diferenciar regiones en el genoma. Según estos autores, sólo uno de los tres estados de metilación (metilación de DNA, H3K4me1, H3K4me3) es alto en una región genómica. Si el estado es DNA metilado la región se considera inactiva. Si se encuentra enriquecida en la marca H3K4me1, la región es un enhancer y si es la marca H3K4me3, la región es promotora.

En conclusión, las modificaciones de histonas H3K4me3 y H3K27Ac se asocian con el inicio de la transcripción, con regiones genómicas con bajos niveles de metilación de DNA y las regiones promotoras y enhancers, respectivamente.

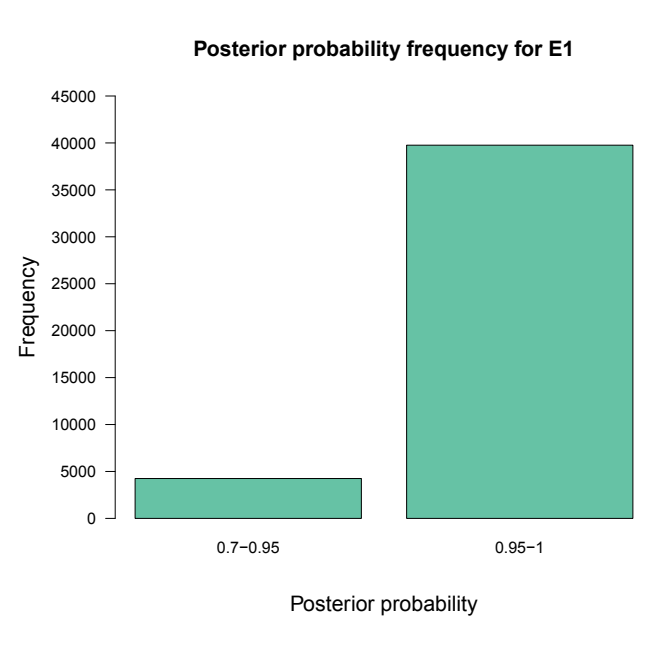
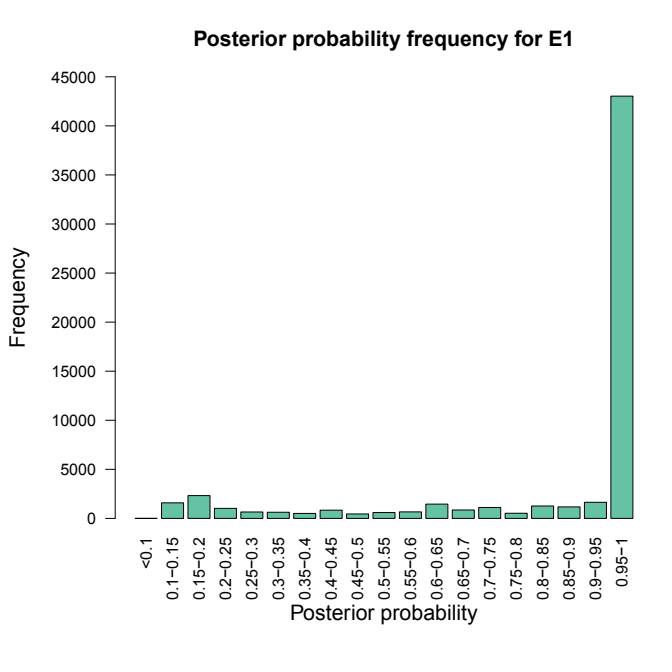
# Análisis

## Paso 1: Obtener los segmentos que tengan el mismo estado en los dos replicados de monocitos.

La cromatina empleada en este trabajo procede de dos réplicas biológicas de monocitos CD14+ CD16- de humano. El primer paso de nuestro trabajo es generar los archivos de partida del estudio. Para ello, se procede a calcular el número de segmentos de 200 pbs que solapan en ambos replicados biológicos. Este paso es fundamental para asegurar que los estados asignados a cada segmento son correctos. Igualmente, para mayor seguridad, se emplean los archivos de la carpeta “POSTERIOR” para generar el archivo de la intersección de segmentos entre las dos réplicas.

Entre los archivos de la carpeta “POSTERIOR” hay un documento por cromosoma, en el que cada línea corresponde a un segmento de 200 pbs y cada columna indica la probabilidad posterior de dicho de segmento a pertenecer a cada uno de los 11 estados. De este modo, se estableció como umbral de selección de segmentos el valor de probabilidad posterior 0.7, extrayéndose únicamente los segmentos que igualaran o superaran ese umbral para E1 en ambas réplicas biológicas.

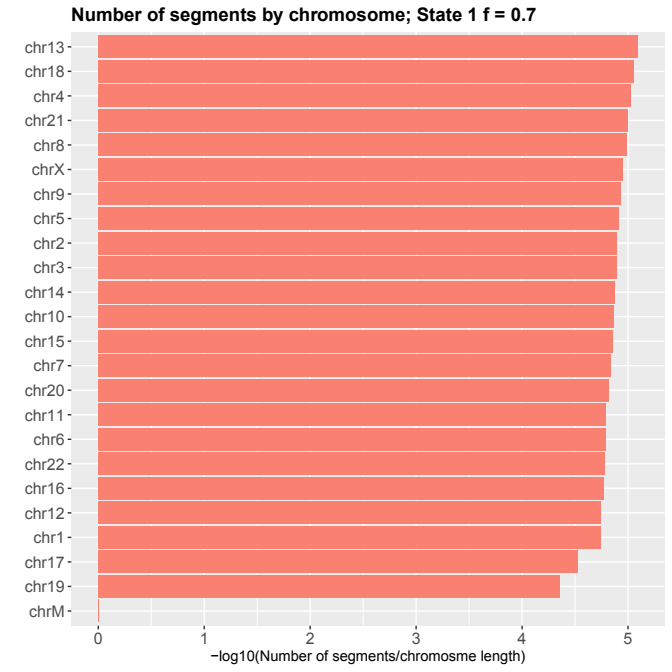
Se emplea código en bash y awk (archivo “filtrador.sh”) para hacer este paso y juntar todos los segmentos de todos los cromosomas. Al final del trabajo se muestra el enlace al proyecto de github (<https://github.com/Huertas97/Transcriptomics>) donde está disponible todo el código empleado.

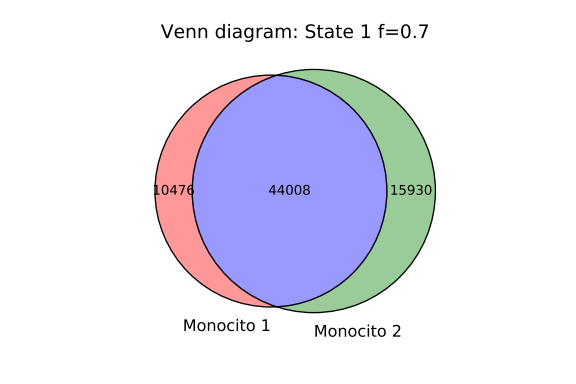
**

**Figura 1-** Distribución de los segmentos del estado 1 en función de la probabilidad posterior

**Figura 2:** Segmentos seleccionados para el análisis (44008).

A continuación, se extraen los segmentos solapantes de los ficheros generados en el paso anterior para cada réplica de monocito. Para ello se emplea el comando “- intersect” de la herramienta “bedtools v2.29.1” sin ninguna opción adicional, puesto que los segmentos de ambas réplicas biológicas tienen 200 pbs de longitud y sólo pueden coincidir en su totalidad. A continuación, se muestran algunos resultados de este primer paso:

 **Figura 2:** Número de segmentos por cromosoma tras normalizar por el tamaño de cada cromosoma. Resultado obtenido tras la intersección de los segmentos solapantes en ambas réplicas biológicas de monocito.

 **Figura 3:** Diagrama de Venn que muestra el número de segmentos solapantes entre ambas réplicas biológicas, tras la intersección de los segmentos con una probabilidad superior o igual a 0.7 de ser asignados al E1.

## Paso 2: Paso 2: Anotar los segmentos

Una vez se han extraído los segmentos comunes entre ambas réplicas biológicas de monocito con una probabilidad de ser asignados al E1 mayor o igual a 0.7, se procede a caracterizar estos segmentos mediante su anotación. Se emplea el paquete de R “annotatr” versión 3.10 (https://www.bioconductor.org/packages/release/bioc/html/annotatr.html).

Además, con el objetivo de obtener información funcional sobre los genes anotados con los paquetes anteriores, posteriormente se procede a llevar a cabo un enriquecimiento funcional con el siguiente software:

* GREAT: Genomic Regions Enrichment of Annotations Tool (http://great.stanford.edu/public/html/splash.php).
* clusterProfiler (https://bioconductor.org/packages/release/bioc/html/clusterProfiler.html), versión 3.14.3.

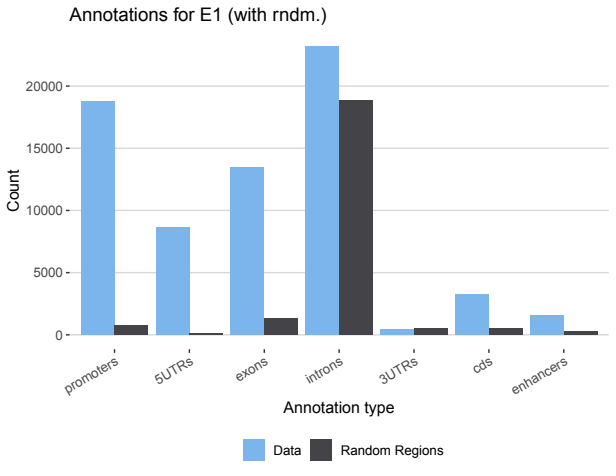
El cálculo del número de segmentos que solapan con genes codificantes para proteínas se realizó mediante la anotación de los segmentos que pertenecían a CDS mediante el paquete “annotatr”. Se emplea este paquete porque dispone de la función “summarize\_annotations”, que muestra un resumen de las anotaciones presentes en los datos obtenidos tras la búsqueda (Tabla 1).

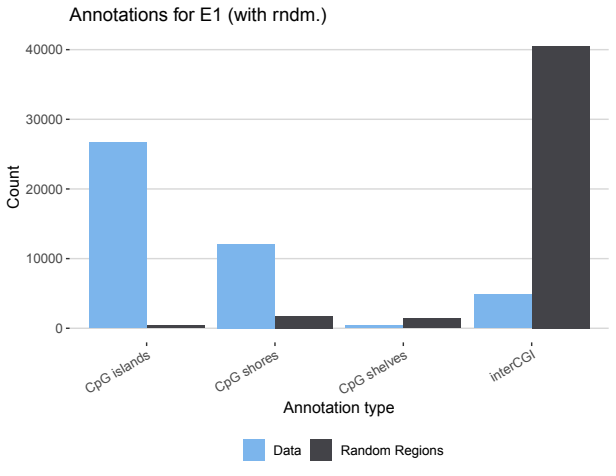
|  |  |  |
| --- | --- | --- |
| Tipo de anotación | Número de segmentos anotados | Porcentaje de anotación |
| CpG\_inter | 4863 | 11 |
| CpG\_island | 26661 | 61 |
| CpG\_shelves | 402 | 1 |
| CpG\_shores | 12128 | 28 |
| Enchancers\_fantom | 1533 | 3 |
| Genes\_1to5kb | 8907 | 20 |
| Genes\_3UTRs | 451 | 1 |
| Genes\_5UTRs | 8615 | 20 |
| Genes\_cds | 3265 | 7 |
| Genes\_exons | 13450 | 31 |
| Genes\_intergenic | 1877 | 4 |
| Genes\_introns | 23180 | 53 |
| Genes\_promters | 18801 | 43 |

**Tabla 1:** Resumen de las anotaciones presentes tras la anotación de segmentos con “annotatr”.

Con las anotaciones obtenidas, se comprueba que un 7% de los segmentos se asocian con regiones codificantes para proteínas (CDS). Es importante señalar que se estableció como mínimo un solapamiento de 100 pbs entre los segmentos de 200 pbs y las coordenadas del genoma anotado para asignar la anotación al segmento. Se estableció este valor porque los exones tienen un tamaño de alrededor de 120 pbs y los intrones un tamaño de 2 kbs en regiones genómicas que contienen un 30-40% de GC y una longitud media de 500 pbs en las regiones con más de 50% de GC (Alberts, 2016). A pesar de que las regiones 5’-UTR y 3’-UTR pueden tener desde 60-80 pbs a 4 kbs (Chatterjee and Pal, 2009) y podrían no anotarse si tuvieran menos de 100 pbs, se prioriza la anotación de exones, intrones, CDS y promotores frente a estas regiones (Figura 4).

A partir de los resultados obtenidos de la anotación, se comprueba que las 3 anotaciones más abundantes son islas CpG, intrones y promotores. Para comprobar si esta anotación está enriquecida en nuestros segmentos con respecto el genoma, el paquete “annotatr” presenta la función *randomize\_regions* que, a partir de un set de regiones y el genoma, devuelve un nuevo set de regiones del mismo tamaño distribuidas aleatoriamente en el genoma, las cuales fueron posteriormente anotadas con las mismas condiciones que los segmentos de estudio. En la Figura 5 se puede comprobar cómo las anotaciones de promotor, exones, CDS y regiones 5’UTR se encuentran claramente enriquecidas en nuestros datos. Igualmente, se ve algo de enriquecimiento en los enhancers, en menor medida en los intrones y nada en las regiones 3’UTR. Con respecto a los intrones, vemos que la anotación por azar es muy elevada, ya que se corresponden al 50% del genoma del ser humano (Lamolle and Musto, 2018). Este hecho hace que sea más probable encontrar un elevado número de anotaciones en ellos.

**Figura 4:** Comparación de la anotación en nuestros datos con datos generados aleatoriamente

 **Figura 5:** Comparación de la anotación de islas CpG de nuestros datos con datos generados aleatoriamente.

En conclusión, estas anotaciones demuestran que los segmentos del E1 están relacionados con la transcripción, situándose en promotores, enhancers, regiones 5’UTR y regiones codificantes. Asimismo, las islas CpG son regiones del genoma con un alto contenido en GC susceptibles de ser metiladas para regular epigenéticamente el estado de condensación de la cromatina. Estos elementos genómicos se sitúan especialmente en los promotores. En la Figura 5, se observa claramente un enriquecimiento de nuestros segmentos en las islas CpG y las regiones colindantes a ellas. La anotación “interCGI” hace referencia al resto de anotaciones diferentes a islas CpG, y se puede comprobar que es menor en nuestros datos.

Igualmente, en caso de que se tenga interés, en el apartado “Anexo: Material adicional” se muestra la distribución de la anotación en función de la probabilidad posterior. Esta distribución de las anotaciones apoya la selección de la probabilidad posterior 0.7 como umbral.

Asimismo, con el objetivo de completar los resultados extraídos del análisis con “annotatr”, se emplea la herramienta web GREAT v4.0.4 permite anotar los segmentos con los términos GO de localización celular, función molecular, proceso biológico y fenotipos humanos. GREAT se encarga de realizar la anotación de los segmentos mediante cálculos estadísticos generados por la asociación de regiones genómicas (segmentos) con genes cercanos. La asociación tiene dos pasos: en primer lugar, se asocia cada gen a un dominio regulador y, posteriormente, cada segmento es asociado con los genes que pertenecen al dominio regulador con el que solapa. Para el análisis se emplearon los parámetros por defecto: región basal más extensión de 5 kbs aguas arriba, 1 kb aguas abajo y 1000 kbs para regiones distales Las figuras obtenidas con esta herramienta se encuentran en github (<https://github.com/Huertas97/Transcriptomics>), pero para facilitar su accesibilidad se han añadido al “Anexo: Material adicional”.

Entre los resultados obtenidos destaca que tan solo 50 (0.1%) de los 44008 segmentos del E1 no se encuentran relacionados con ningún gen. Además, se observa que la mayoría de los segmentos anotados se sitúan cerca o muy cerca del sitio de inicio de la transcripción (TSS). A pesar de que la longitud y la secuencia de los promotores humanos es variable, los elementos más importantes (denominados en inglés como “cores”) se sitúan en un rango cercano al TSS, ~100 pbs aguas arriba y ~100 pbs aguas abajo (Landolin et al., 2010). En consecuencia, los resultados obtenidos indican que nuestras marcas de interés (H3K4me3 + H3K27Ac) se encuentran relacionadas con los promotores.

Otro resultado importante obtenido en GREAT son las notaciones de los términos GO. Entre los términos GO asociados a los segmentos del estado E1 encontramos términos relacionados con el procesamiento del RNA y la transcripción (*mRNA catabolic process, RNA catabolic process, elongation, negative regulation of gene expression epigenetic*…) al igual que términos relacionados con el sistema inmune (*regulation of hematopoietic stem cell differentiation, regulation of hematopoietic progenitor cell differentiation*…).

Además de los términos GO obtenidos mediante GREAT, se utilizó el paquete de R “clusterProfiler” para llevar a cabo la anotación funcional de los genes anotados con “annotatr” tanto para contrastar los resultados obtenidos con GREAT, como para recabar nueva información de las regiones asociadas a E1. Para ello, se han llevado a cabo la anotación de rutas KEGG (https://www.genome.jp/kegg/), y un análisis de enriquecimiento frente a la base de datos Molecular Signatures Database (MSigDB, https://www.gsea-msigdb.org/gsea/msigdb/index.jsp).

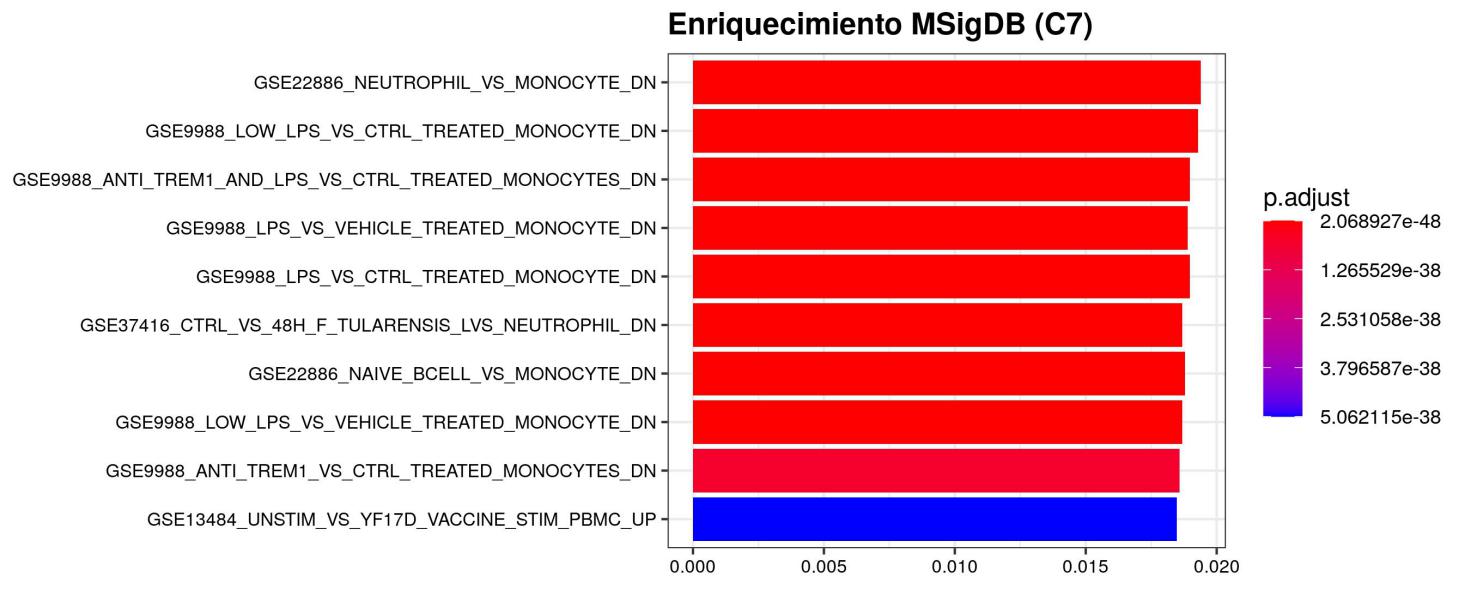
En este últumo caso, dado que el estado de cromatina que estamos estudiando está relacionado con ĺa activación de la expresión genética, el objetivo es comprobar si efectivamente los genes anotados se relacionan con el tipo celular de estudio, monocitos. Respecto a los parámetros de significación utilizados, en todos los casos se lleva a cabo corrección por testeo múltiple de los p-valores mediante el método de Benjamini-Hochberg (BH) y se establece un umbral de significancia igual a 0.05.

En cuanto a los resultados del enriquecimiento con rutas KEGG, se puede observar en la Figura 6 cómo claramente la mayoría de las entradas obtenidas son rutas relacionadas con el sistema inmune, presentando además un p-valor ajustado muy bajo.

### unnamed-chunk-4-2

**Figura 6:** Top 20 rutas KEGG relacionadas con los genes relacionados con los segmentos E1.

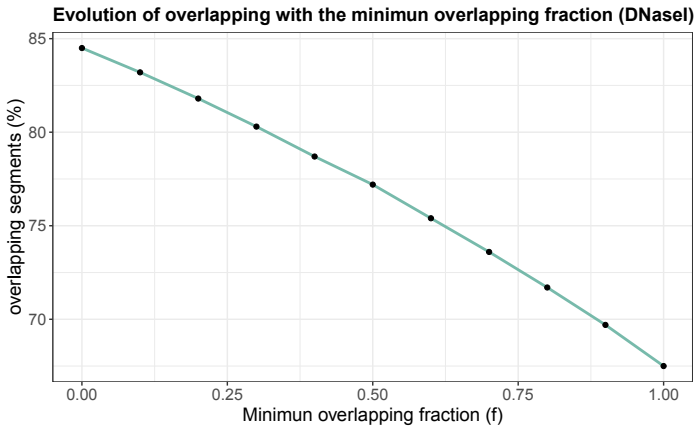
Finalmente, se llevó a cabo el análisis de enriquecimiento contra MSigDB. Esta base de datos consiste en diferentes sets de genes relacionados con algún área en particular, como por ejemplo perfiles de tejidos oncológicos, etc. En este caso, se utilizó la colección dedicada a células del sistema inmune (C7) para, como se ha comentado anteriormente, validar el tipo celular con el que estamos trabajando. Como podemos comprobar en la Figura 12, efectivamente la mayoría de entradas obtenidas proceden de experimentos relacionados con monocitos. Este hecho nos permite confirmar que, los genes de las regiones anotadas con E1 son relativos al perfil de los monocitos.



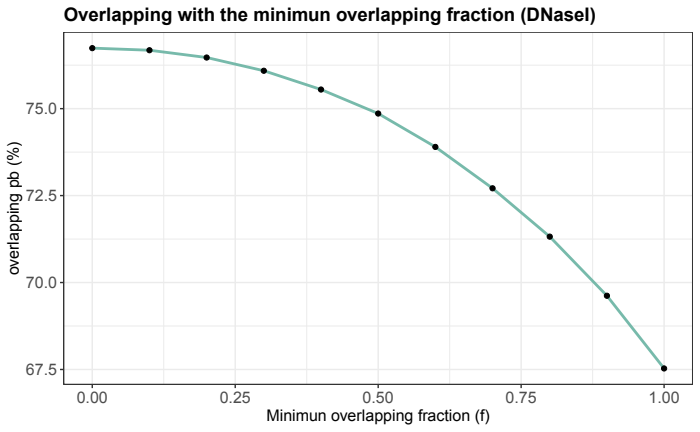
**Figura 7:** Top 20 entradas resultado de enriquecimiento contra MSigDB (colección C7: células inmunes).

## Paso 3: Calcular el porcentaje de solapamiento entre DNaseI-peaks y vuestros segmentos de trabajo.

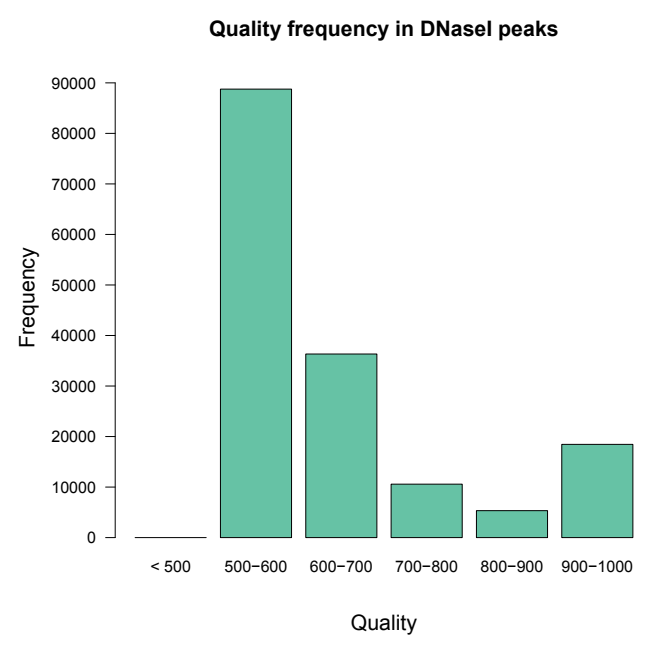
En primer lugar, es importante indicar que la DNaseI es una endonucleasa capaz de romper el enlace fosfodiéster entre dos nucléotidos, tanto en DNA monocatenario como bicatenario en las regiones accesibles de la cromatina. Esta enzima se emplea en la técnica DNase I-seq para identificar regiones hipersensibles a DNase I (“DNase I Hypersensitive Site”, DHS) a lo largo del genoma. Las regiones genómicas donde actúa la Dnase I son consideradas marcadores de regiones reguladoras de DNA, regiones de inicio de transcripción, enhancers y silenciadores. En otras palabras, las regiones genómicas secuenciadas en un experimento de Dnase I-seq corresponderían a las regiones genómicas accesibles, donde la maquinaria de transcripción podría llevar a cabo su función (Sullivan et al., 2015).

En nuestro caso, se emplean segmentos de DNase I de monocitos CD14+ de la versión del genoma hg19 procedentes de ENCODE. El solapamiento entre nuestros segmentos y los de la DNase I permite conocer qué segmentos de E1 se encuentran accesibles a la maquinaria de transcripción. El solapamiento se calcula de forma equivalente a la realizada en el paso 1. Sin embargo, dado que la longitud de los segmentos procedentes de la DNase I no son uniformes (media de 606 pbs) y son superiores a los 200 pbs de los segmentos del E1, se selecciona como archivo de referencia el archivo con los segmentos de E1. De este modo, se procede a buscar aquellos segmentos de E1 que solapan con la DNase I y no al revés. Esto, además, permite aplicar una fracción mínima de solapamiento de 100 pbs, de modo que solo se seleccionen aquellos segmentos de E1 que al menos solapan en 100 pbs con segmentos de DNAse I. Este valor se estableció a partir del estudio de la variación del número de segmentos y pares de bases solapantes a medida que aumenta la fracción solapante.

**Figura 8:** Evolución del número de regiones solapantes con respecto a la fracción mínima solapante.

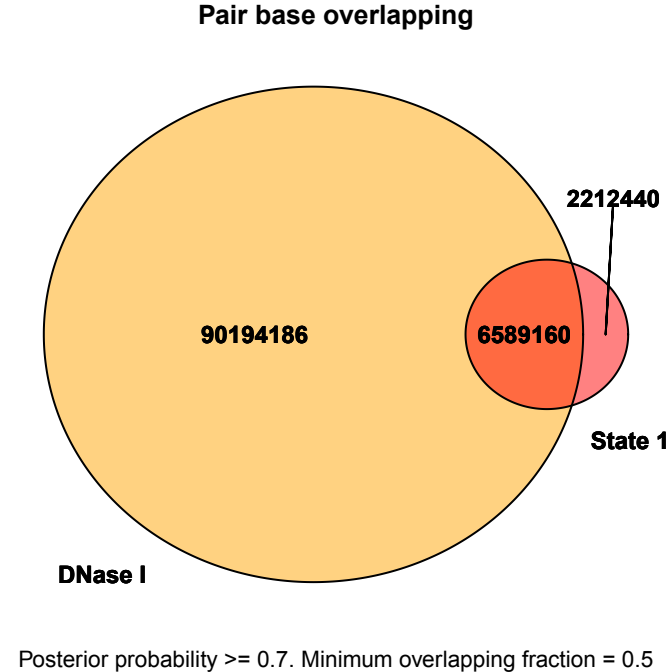
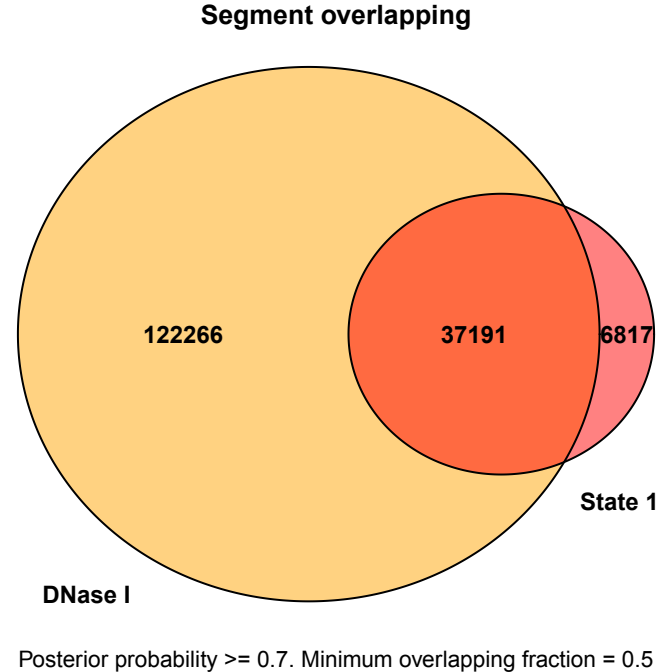
En la Figura 13 se comprueba que el número de regiones solapantes disminuye a medida que se es más restrictivo con la fracción mínima solapante. Cabe destacar que el cambio del número de segmentos solapantes es lineal, mientras que el número de pares de bases solapantes disminuye de forma curvilínea, siendo la caída suave al principio y brusca al final. Esto puede explicarse por el hecho de que, a medida que aumenta la fracción mínima solapante, el número de segmentos que dejan de solapar disminuye de forma progresiva y equivalente, pero el número de pares de bases que se pierden con esos segmentos es cada vez mayor, haciendo más brusca la caída. Con estos resultados, se considera que es adecuado establecer como fracción mínima de solapamiento 0.5, lo que equivale a restringir el solapamiento a 100 pbs como mínimo.

Igualmente, antes de proceder a calcular el solapamiento, se procedió a estudiar la calidad de los picos presentes en el archivo procedente de ENCODE. En la Figura 14 se comprueba que todos los picos presentan una calidad superior a 500, encontrándose la mayoría en el rango 500-600. De este modo, ningún pico es eliminado, pues la calidad media se sitúa en el rango 100-1000 recomendado por ENCODE.



**Figura 9:** Histograma de la calidad de las regiones accesibles a la DNase I en monocitos CD14+ procedentes de ENCODE.

Una vez se han explorado los datos descargados de ENCODE, se procede a calcular el porcentaje de solapamiento entre nuestros segmentos y los picos de DNase I. El solapamiento en función de los segmentos es del 85%, mientras que en función de los pares de base es del 75% (Figuras 15 y 16). En cualquiera de los casos, queda claro que nuestros segmentos se encuentran de forma abundante en regiones accesibles de la cromatina, apoyando la hipótesis de que el E1 corresponde a regiones transcripcionalmente activas.



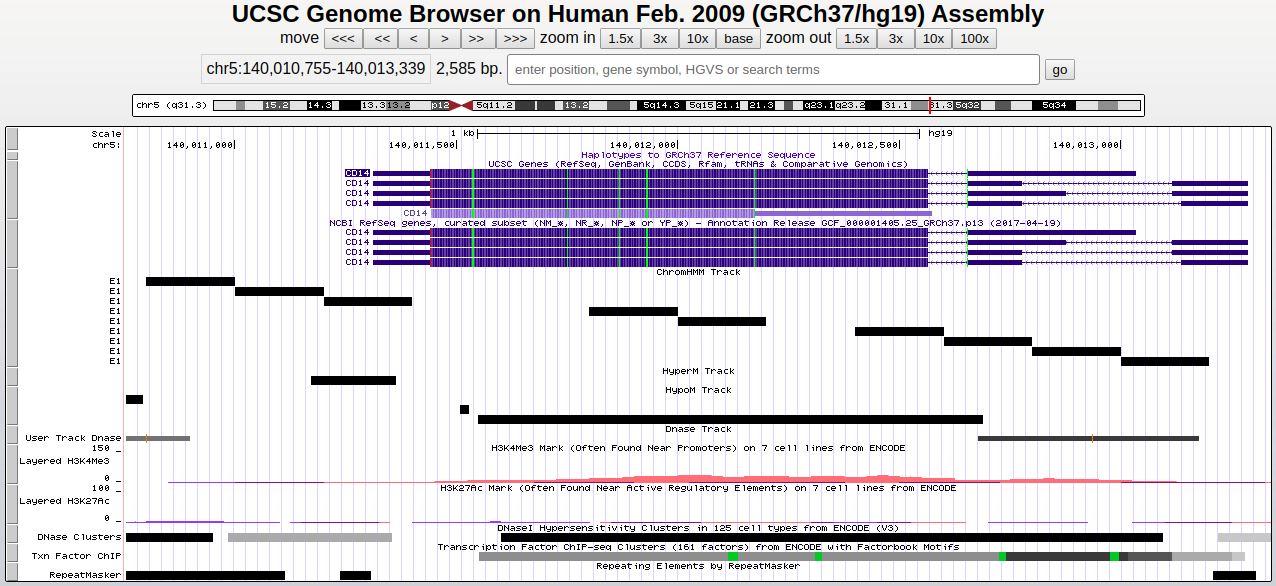
## Paso 4: Visualizar una región del genoma en el *UCSC browser*.

**Figura 11:** Diagrama de Venn con el número de segmentos solapantes entre la DNase I y el E1.

**Figura 10:** Diagrama de Venn con el número de pares de bases solapantes entre la DNase I y el E1.

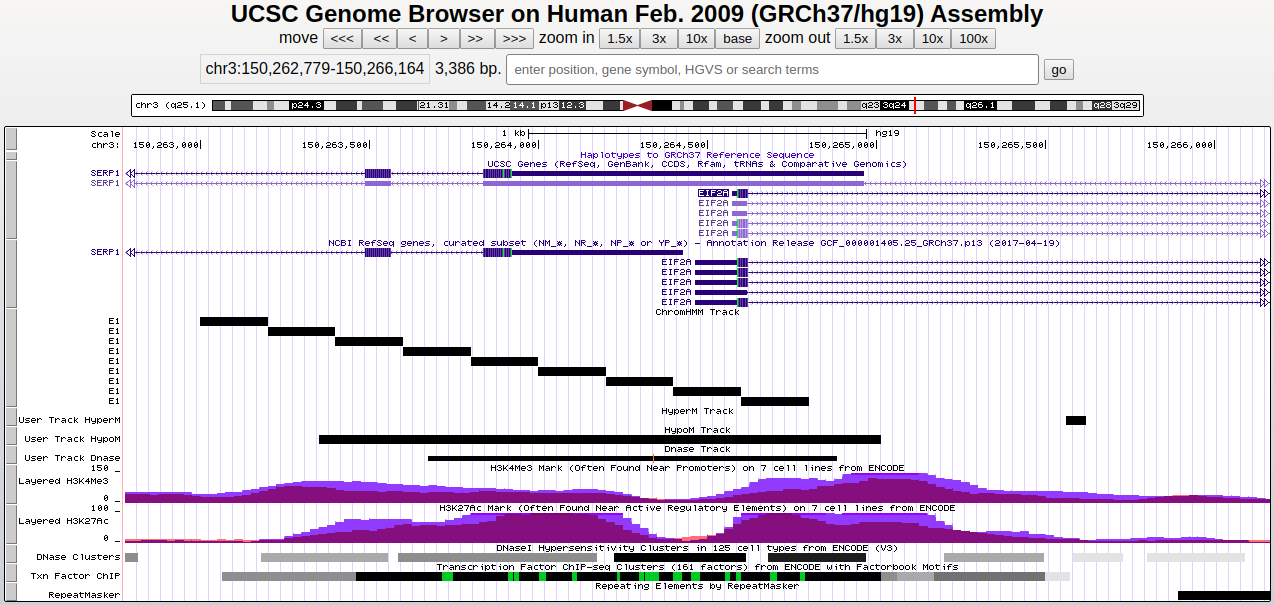
Este paso es muy importante para visualizar y contextualizar los resultados obtenidos. Para ello, en el navegador genómico de la UCSC se suben: el fichero bed con la intersección de segmentos con el E1 como más probable entre ambos monocitos, el fichero de marcas de hiper e hipometilación y el fichero con los picos de la DNase I de células sanguíneas de ENCODE. Además, se eligió mostrar las modificaciones de histonas de interés solo en los tipos celulares disponibles asociados al sistema inmune (GM12878 y K562), al cual pertenecen las células de estudio. A continuación, se visualizaron una serie de genes relacionados con los términos GO mostrados anteriormente y cuya función evidencia la participación en el proceso de transcripción génica, epigenómica y diferenciación celular de las células monocíticas: CD14, LYN y EIF2A.

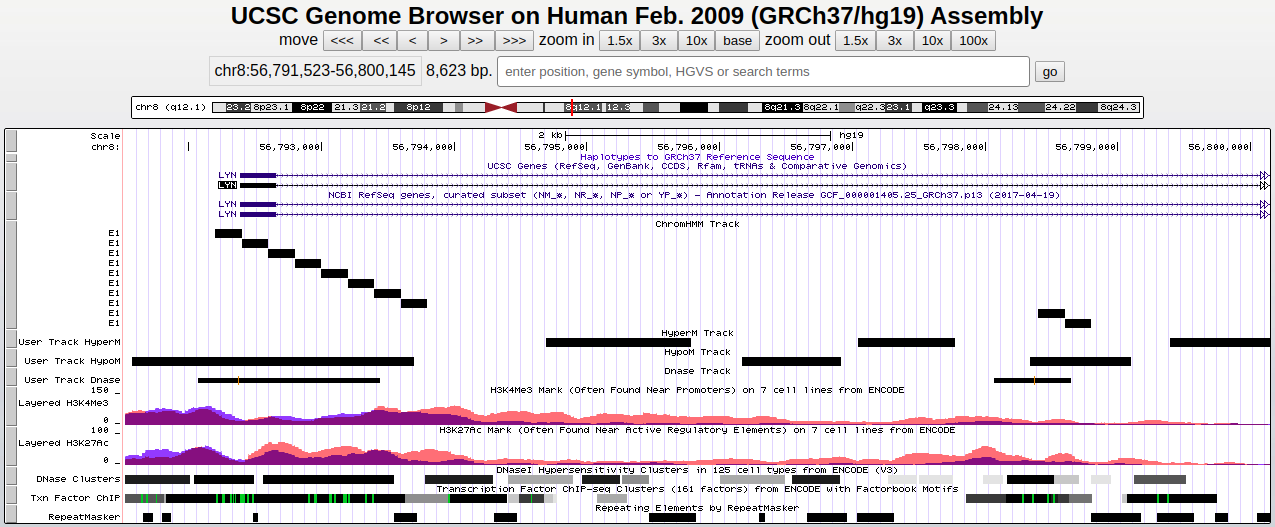
El gen CD14 codifica para una proteína que se localiza de forma específica en la superficie de monocitos/macrógafos participando en el reconocimiento de oligosacáridos procedentes de patógenos. En la Figura 17 se puede observar cómo los segmentos de E1 se sitúan predominantemente en el inicio del gen (parte derecha de la imagen), coincidiendo con la teórica posición del promotor, aunque también con algunos exones e intrones. Además, dichos segmentos de E1 en el inicio solapan con regiones hipometiladas (datos procedentes del paciente C001UY de BLUEPRINT) y regiones accesibles (track de la DNase I). Por último, hay que destacar que estos segmentos de E1 se acompañan de la existencia de la marca H3K4me3 y, en mucha menor medida, pero presente, de la H3K27Ac (marcas de promotor activo).

 **Figura 12:** Visualización del gen CD14 en el navegador genómico de la UCSC.

**Figura 17:** Visualización del gen CD14 en el navegador genómico de la UCSC.

El gen LYN codifica para una tirosina quinasa involucrada en la degranulación celular y la diferenciación hematopoyética. Es considerado un proto-oncogen por su participación en el desarrollo celular y su desregulación se asocia a enfermedades como coreocantocitosis o sarcoma. En la Figura 18 se puede comprobar que los segmentos de E1 solapan de una forma más evidente con la región inicial del gen, siendo ésta la ubicación teórica del promotor del mismo, y se acompañan de considerables picos tanto de H3K4me3 como de H3K27Ac. También se repite el solapamiento con regiones de hipometilación y accesibles.

El gen EIF2A codifica para un factor de transcripción iniciador de la transcripción. En la Figura 19 se observa que la situación es muy parecida a la del gen anterior, con una presencia evidente de los segmentos de E1 y marcas epigenéticas de interés, de hipometilación y de accesibilidad en la zona teórica del promotor. Además, este promotor podría ser bidireccional, dado que en sentido contrario se encuentra el gen SERP1, relacionado con la síntesis de proteínas. De hecho, en las anotaciones presentes en la base de datos GeneCards se puede observar que los promotores asociados a SERP1 tienen también como diana a EIF2A. Adicionalmente, la existencia de este promotor bidireccional podría suponer un sistema de regulación de la expresión de ambos genes.

 **Figura 13:** Visualización del gen LYN en el navegador genómico de la UCSC.

### Figura 14: Visualización del gen EIF2A en el navegador genómico de la UCSC.

En conclusión, de las visualizaciones realizadas llama la atención que los segmentos de E1 correspondientes a estos genes se ubican en el extremo 5’ de los mismos, en los cuales también se observa hipometilación (asociada a expresión), solapamiento con segmentos de DNase I (región accesible) y presencia de las marcas epigenéticas H3K4me3 y H3K27ac (asociadas a promotores activos). Esto implica que los genes observados, de expresión típica en monocitos, muestran sus promotores activos y podrían transcribirse.

## Paso 5: Búsqueda de motivos enriquecidos.

El objetivo de esta sección es obtener información sobre motivos de unión conocidos de diversos factores de transcripción (TF) que aparecen enriquecidos en nuestros segmentos, aportando más información al análisis bioinformático (McLeay, y Bailey, 2010). La herramienta utilizada es MEME-ChIP (Machanick y Bailey, 2011. MEME-ChIP es una herramienta potente tanto para el análisis e identificación como la visualización de motivos de unión de TF.

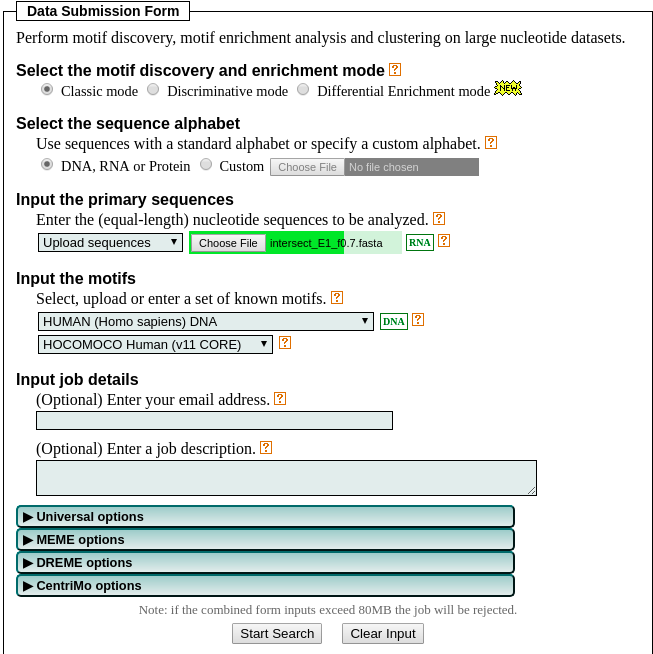
No obstante, para usar esta herramienta es necesario conocer las secuencias asociadas a los segmentos de interés. Es por ello necesario un paso previo para extraer las secuencias a partir de las coordenadas cromosómicas que contiene el fichero *.bed*. Esto se puede hacer con *bedtools* de la siguiente manera:

bedtools getfasta -fi hg19.fa -bed $file\_in -fo $file\_out

donde hg19.fa es el fasta con el genoma de referencia, file\_in es el fichero *.bed* con las coordenadas cromosómicas de las que queremos obtener las secuencias y file\_out es un nuevo fichero fasta con las secuencias de los segmentos. En nuestro caso, se ha utilizado el genoma de referencia disponible en la web del USCS, que puede descargarse en <http://hgdownload.soe.ucsc.edu/goldenPath/hg19/bigZips/>. Así, obtenemos un fichero fasta con los segmentos a estudiar:

>chr10:181800-182000

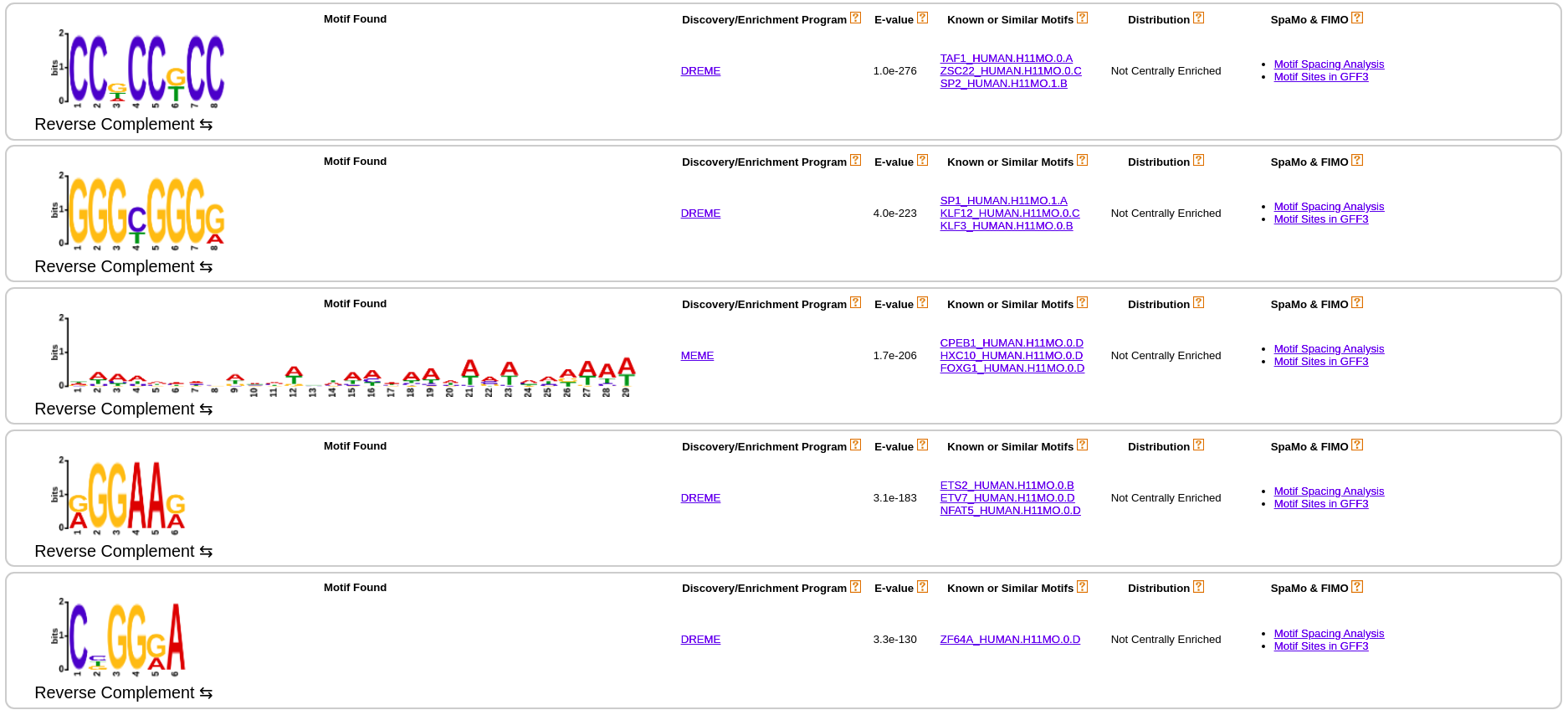
GCGCGGGGCCTGCTCGGGCCGGGAAGCCGCGGAGTCGCGTGAGCACCGCCCGCCGGGCCCTGTGCCCCGCTTTCGGTCAGGCCTCCTGGGCCCGTGCGCAGTCCGGGCCGGCGGGGAACGCGGCTCCGGGCGTGTGGCGGGCGGAGAGGAAGCGTTTGTCGGCGCGGCACGTCGTGTGCTAGCCCGGGAGCGGCGGGCAG

Evidentemente, todas las secuencias son segmentos de 200 pares de bases. A continuación, hacemos el análisis con MEME-ChIP con la siguiente configuración mostrada en la Figura 20.

Para la búsqueda se utiliza la base de datos de HOCOMOCO, que contiene motivos de unión de factores de transcripción conocidos para *Homo sapiens*. La salida de MEME-ChIP es un HTML con la información del análisis.

Los motivos significativos encontrados *de novo* por MEME, DREME y CentriMo son agrupados por similitud y ordenados por E-valor. En la Figura 21 se muestran los cinco más significativos.

**Figura 15:** Configuración utilizada en MEME-ChIP

Como podemos observar, los dos primeros presentan un alto contenido en C y Gs, lo cual tiene sentido biológico por el hecho de que un alto porcentaje de regiones cromosómicas que recogen nuestros segmentos son islas CpG, áreas relacionadas con la expresión genética.

**Figura 16:** Motivos de unión de novo encontrados por MEME, Dreme y CentriMo.

### Con el fin de tener más información, se llevó a cabo un análisis de enriquecimiento de motivos con una segunda herramienta: HOMER. Al igual que en el caso anterior, se lleva a cabo sobre el mismo set de secuencias correspondiente a los E1 con una probabilidad posterior mayor de 0.7. Además, los parámetros utilizados fueron los establecidos por defecto.

**Figura 17:** Top 8 motivos enriquecidos mediante HOMER.

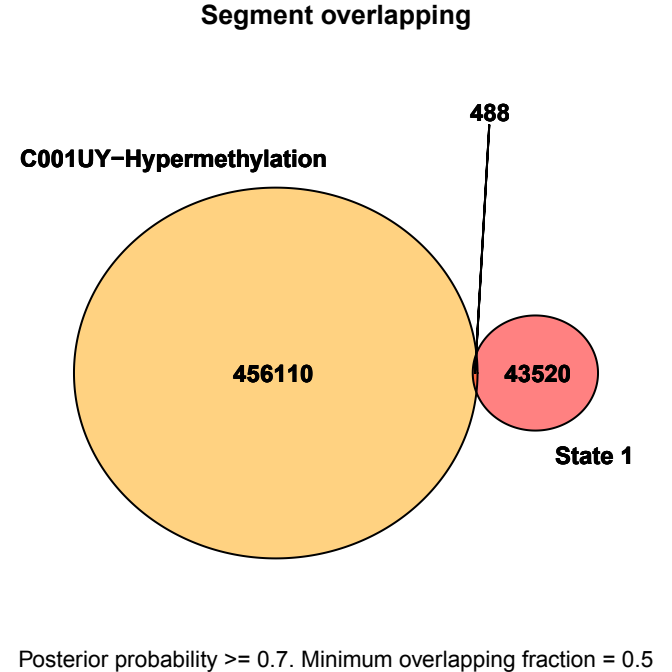
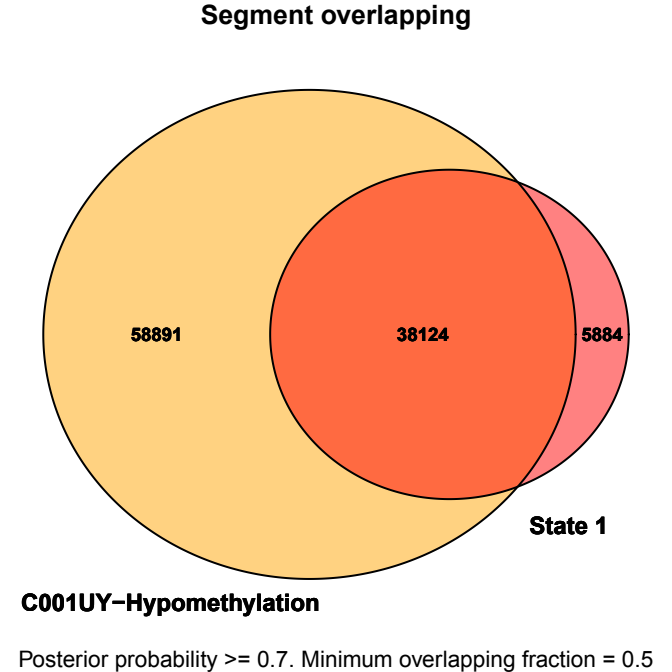
Como podemos observar en la Figura 17, en la, el set de bases nucleótidos encontrados en estos motivos es más amplio, pareciendo ser en general más ricos en citosinas y timinas. Algunos de los factores de transcripción (TF) asociados a estos motivos son ELK4, YY2, para el cual se encuentran dos entradas; ZBED1, ETV6 o NFYA, entre otros. Por dar algunos ejemplos, podríamos destacar ETV6, TF con actividad activadora específico de la RNA-polimerasa II que juega un papel relevante en los fenómenos de hematopoyesis y transformación maligna (Zhang et al., 2015); o NFYA que activa el core que controla los ciclos circadianos, ARNTL/BMAL1.

Cabe destacar que algunos de los motivos encontrados mediante MEME-CHiP son validados mediante HOMER, como es el caso de ETV6, el cual está relacionado de hecho con las células inmunes.

### Paso 6: Regiones hiper e hipometiladas

La metilación de citosinas del DNA y de histonas es un mecanismo muy importante en la regulación de la trasncripción. Las regiones promotoras e islas CpG (situadas éstas últimas mayor mente en los promotores) hipermetiladas se asocian con una represión de la transcripción, mientras que la hipometilación de las mismas se asocia con la activación de la transcripción.

En la Figura 22, se observa claramente como los segmentos del estado 1 se relacionan en mayor medida con regiones hipometiladas que con regiones hipermetiladas, 87% y 1%, respectivamente.

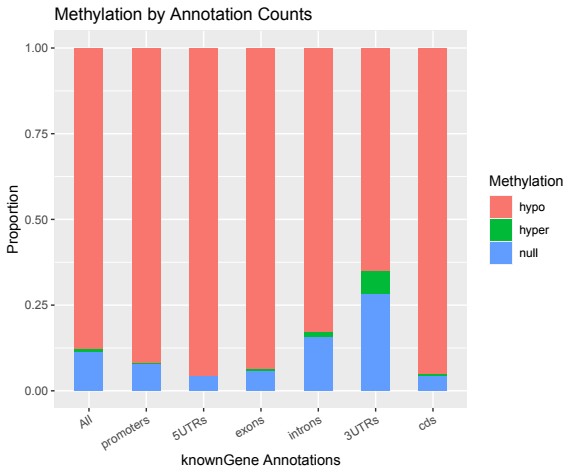
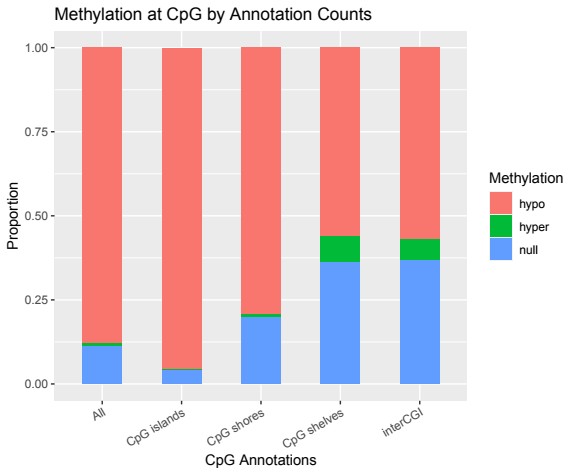


**Figura 18:** Solapamiento entre los segmento E1 y regiones hipo- e hipermetiladas.

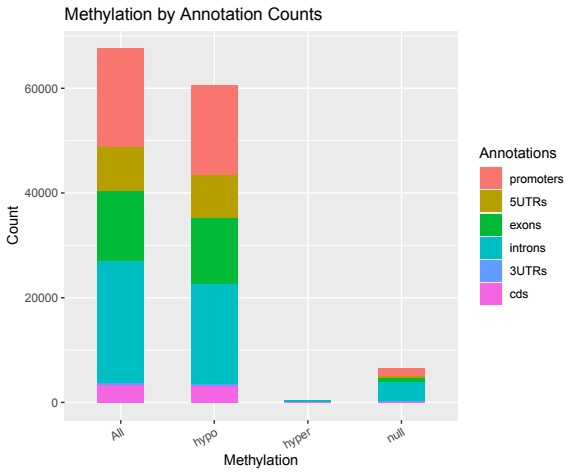
Las regiones hipometiladas, en concreto las relacionadas con promotores e islas CpG se asocian con una regulación positiva de la transcripción. Con los resultados anteriores del porcentaje de solapamiento de los segmentos correspondientes al estado E1 con las regiones hiper e hipometiladas del paciente C001YU, podemos concluir que las marcas H3K4me3+H3K27Ac del estado E1 se relacionan con la activación de la transcripción y con genes transcripcionalmente activos. Para dar más evidencias que corroboren esta afirmación, se analiza a continuación la distribución de las anotaciones generadas por el paquete de R “annotatr” en función del estado de metilación.

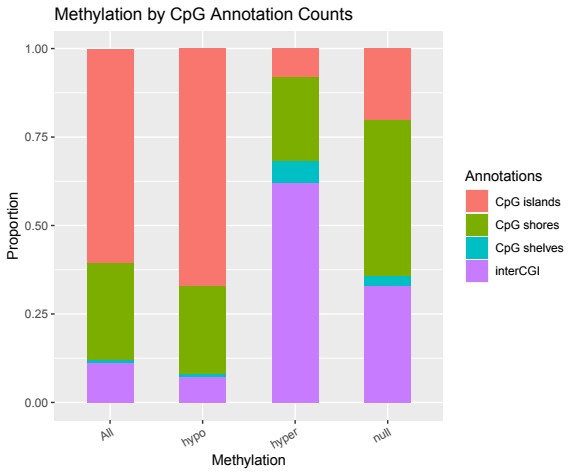
Para realizar el análisis de la distribución de las anotaciones en función del estado de metilación, se asigna con la etiqueta “hypo”, “hyper” o “null” a los segmentos de E1 según si el segmento solapa con las regiones cromosómicas del archivo bed del paciente C001YU de hipometilación, hipermetilación o ninguno de ellos, respectivamente, empleando “bedtools v2.29.1” con un umbral de 100 pb de solapamiento.

En la Figura 24, también se puede comprobar que los segmentos solapantes con regiones hipometiladas son los más abundantes dentro de los segmentos asociados a islas CpG, siendo las regiones hipermetiladas minoritarias. Es muy interesante observar como las islas CpG tienen una gran proporción de segmentos hipometilados, y a penas segmentos ‘null’ o hipermetilados. Por lo tanto, se apoya la hipótesis de que los segmentos del estado E1 se asocian con la activación de la transcripción. Igualmente, se observa que a medida que nos alejamos de las propias islas CpG la proporción de hipometilación disminuye, demostrando que la distribución de la hipometilación no es un artefacto y que se sitúa de forma específica en regiones muy importantes donde la metilación ejerce su papel regulador.



**Figura 19:** Proporción de segmentos de E1 que corresponden a segmentos que solapan con regiones hipometiladas, hipermetiladas o ninguna de ellas, en función de las diferentes anotaciones generadas.





**Figura 20:** Distribución de las anotaciones en función de los distintos niveles de metilación de los segmentos del estado E1.

También se muestran en las Figura 25 y Figura 26 cual es la proporción de cada anotación en cada uno de los diferentes niveles de metilación. Se comprueba de nuevo que son los intrones y los promotores las anotaciones más abundantes en nuestros segmentos, pero como se observó anteriormente en la Figura 5 cuando se comparaba la anotación con unos segmentos generados aleatoriamente a partir del genoma, los intrones anotados no mostraban una gran diferencia con la anotación aleatoria, mientras que sí lo hacían el resto de las anotaciones, en especial los promotores. Del mismo modo, las regiones hipometiladas muestran esta distribución enriquecida en promotores, 5’UTRs, exones y CDS, mientras que la región hipermetilada es insignificante. Con respecto a las islas CpG, podemos observar que se asocian a un estado de hipometilación, disminuyendo en gran medida su presencia en ‘null’ y, en especial, en hipermetilación.Conclusiones

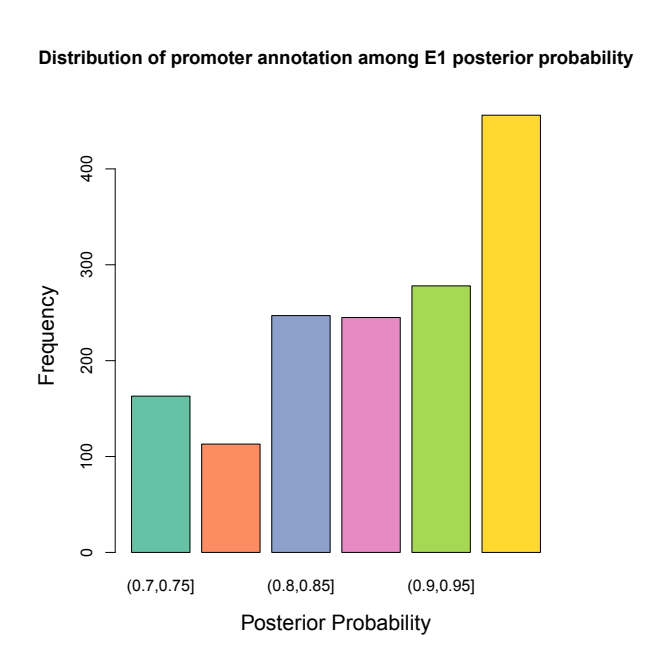
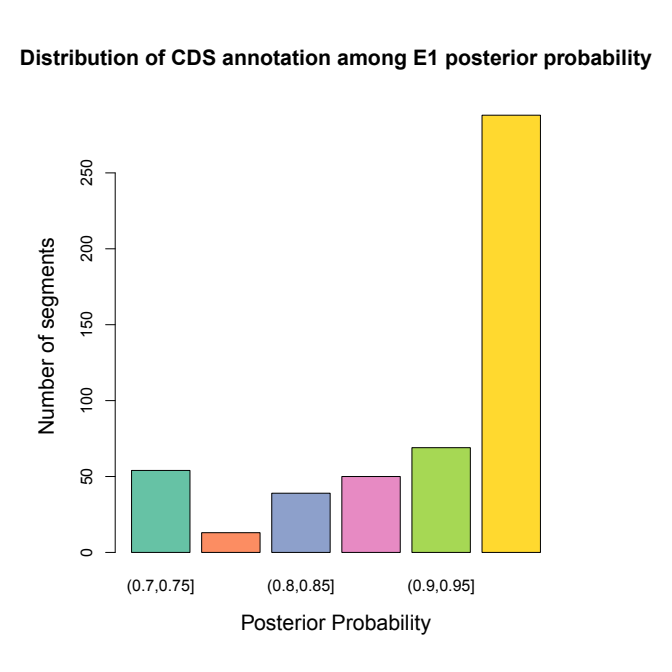
En el presente trabajo se ha mostrado como las marcas H3K4me3+H3K27Ac del estado E1 se asocian con:

* Elementos necesarios para iniciar el proceso de transcripción: promotores, enhancers y regiones 5’UTRs.
* Regiones de la cromatina espacialmente accesibles por la maquinaria de transcipción ( 85% de solapamiento con regiones cromosómicas obtenidas por la acción de DNase I)
* El estado de hipometilación, solapando al menos en 100 pb un 87% los segmentos del estado E1 con regiones hipometiladas del genoma de monocitos.
* Anotación funcional de términos GO relacionada con el sistema inmune del que forman parte los monocitos (“regulation of hematopoietic progenitor cell differentiation”, “regulation of hematopoietic stem cell differentiation”...) y con la maquinaria de transcripción y regulación epigenética (“mRNA catabolic process”, “RNA pol II”, “elongation”... )
* Las islas CpG, situándose a su vez estos segmentos en regiones hipometiladas.

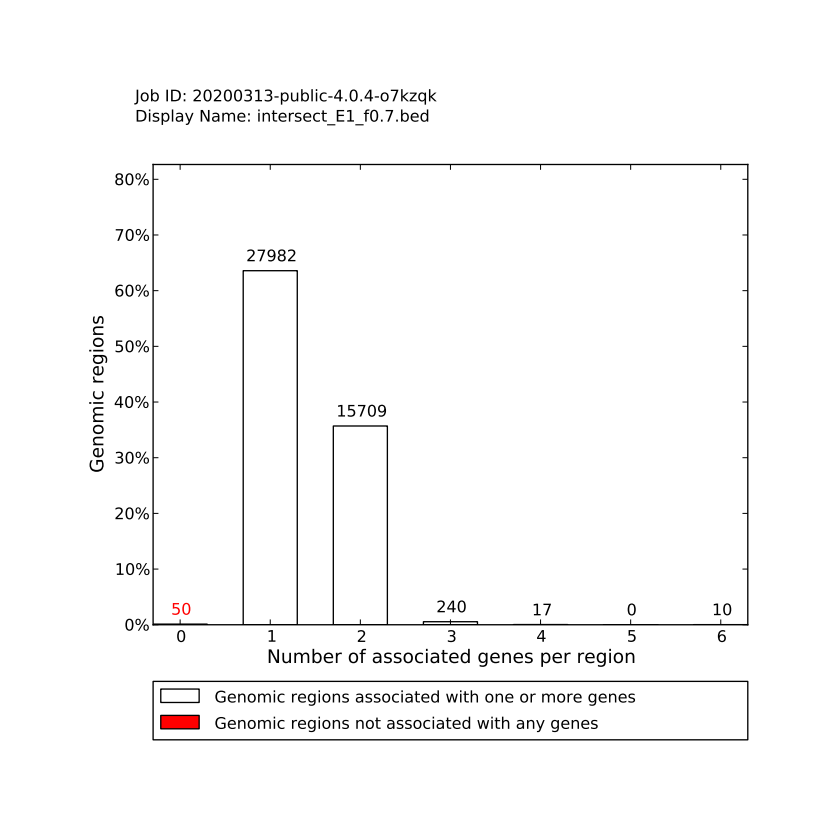
Todos estos resultados se comprobaron también mediante su visualización en el UCSC browser donde se seleccionaron genes (HLA-A, EIF2A y CD14) relacionados con el proceso de transcripción y actividad biológica de monocitos.

Finalmente, el análisis de enriquecimiento de motivos de unión de factores de transcripción (FT) con MEME-CHiP y HOMER, permite asociar los segmentos del estado 1 con FT ricos en citosinas y timinas, estando éstos relacionados con los términos GO analizados previamente.

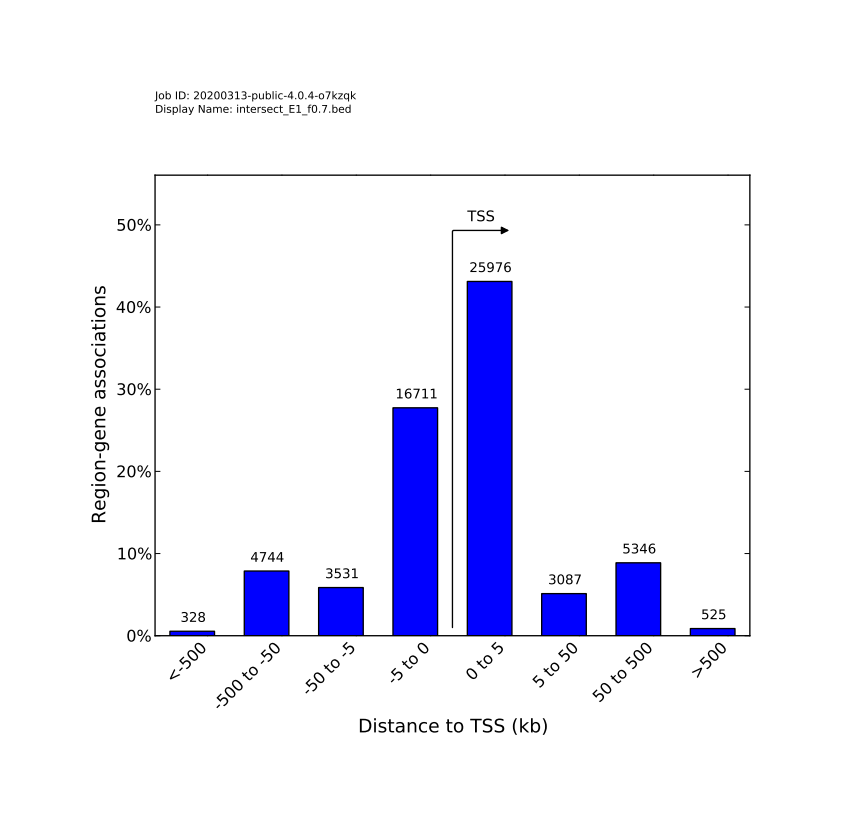
# Anexo: Material adicional

Es importante señalar que estas imágenes están disponibles en el repositorio de github (<https://github.com/Huertas97/Transcriptomics>). No obstante, para facilitar su accesibilidad se han incorporado en este Anexo.

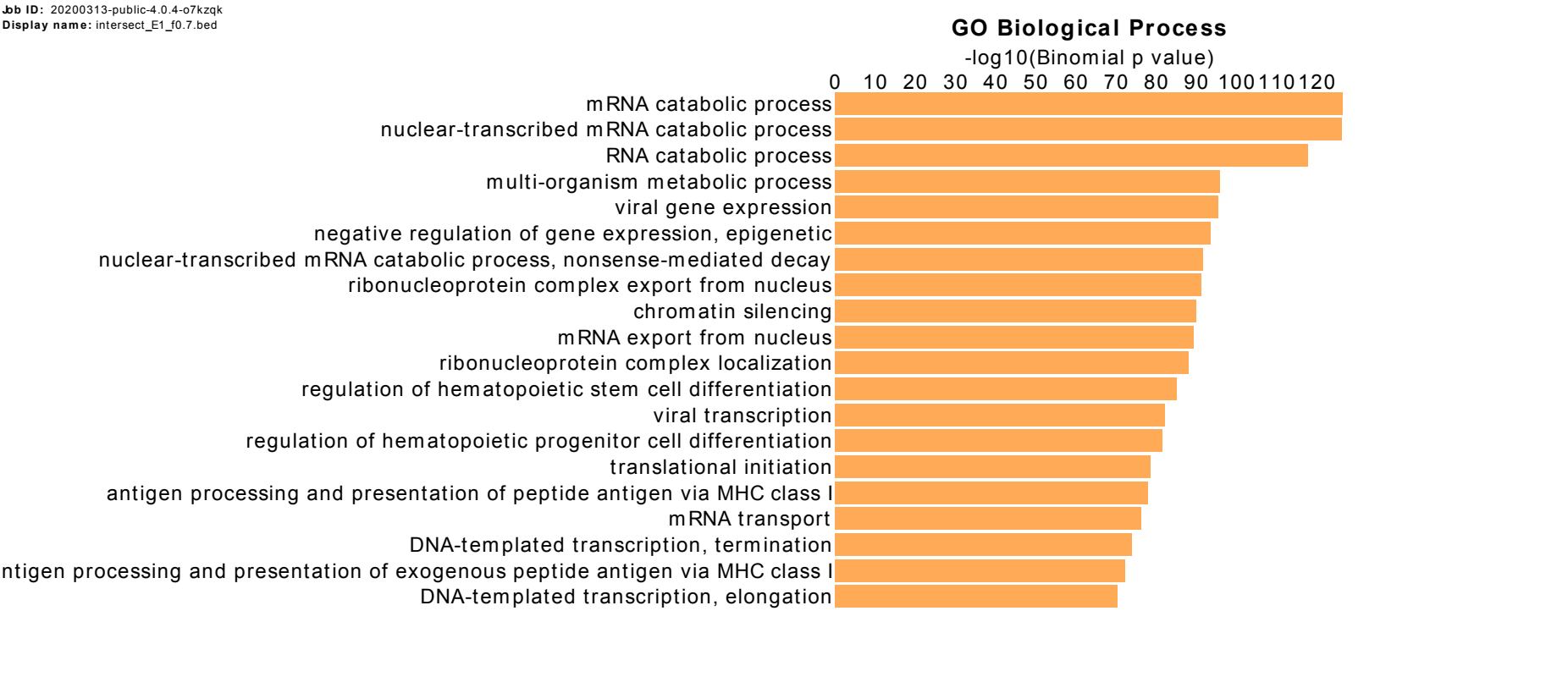
**Figura Anexo 1:** Distribución de la probabilidad posterior de anotaciones CDS y promotores.



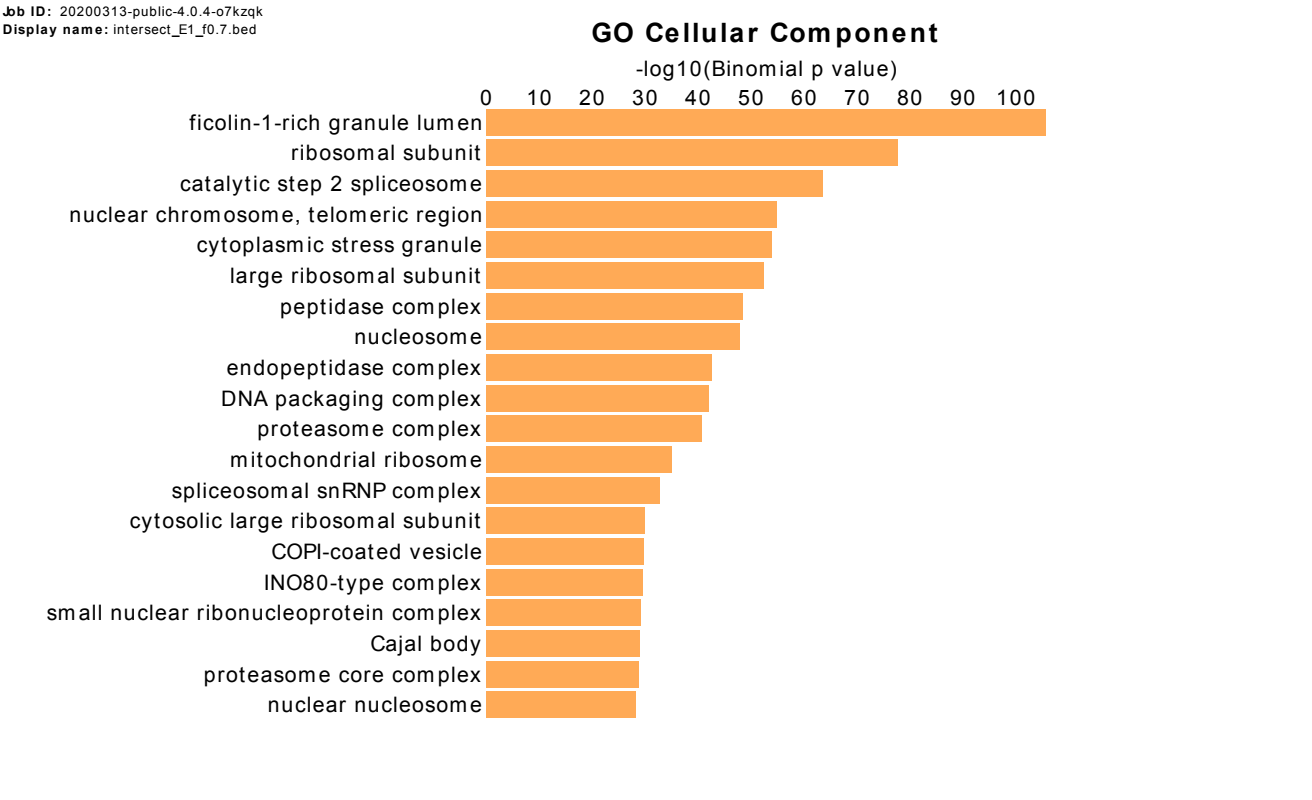
**Figura Anexo 2:** Diagrama de barras procedente de GREAT que muestra el número de segmentos de E1 asociados a genes



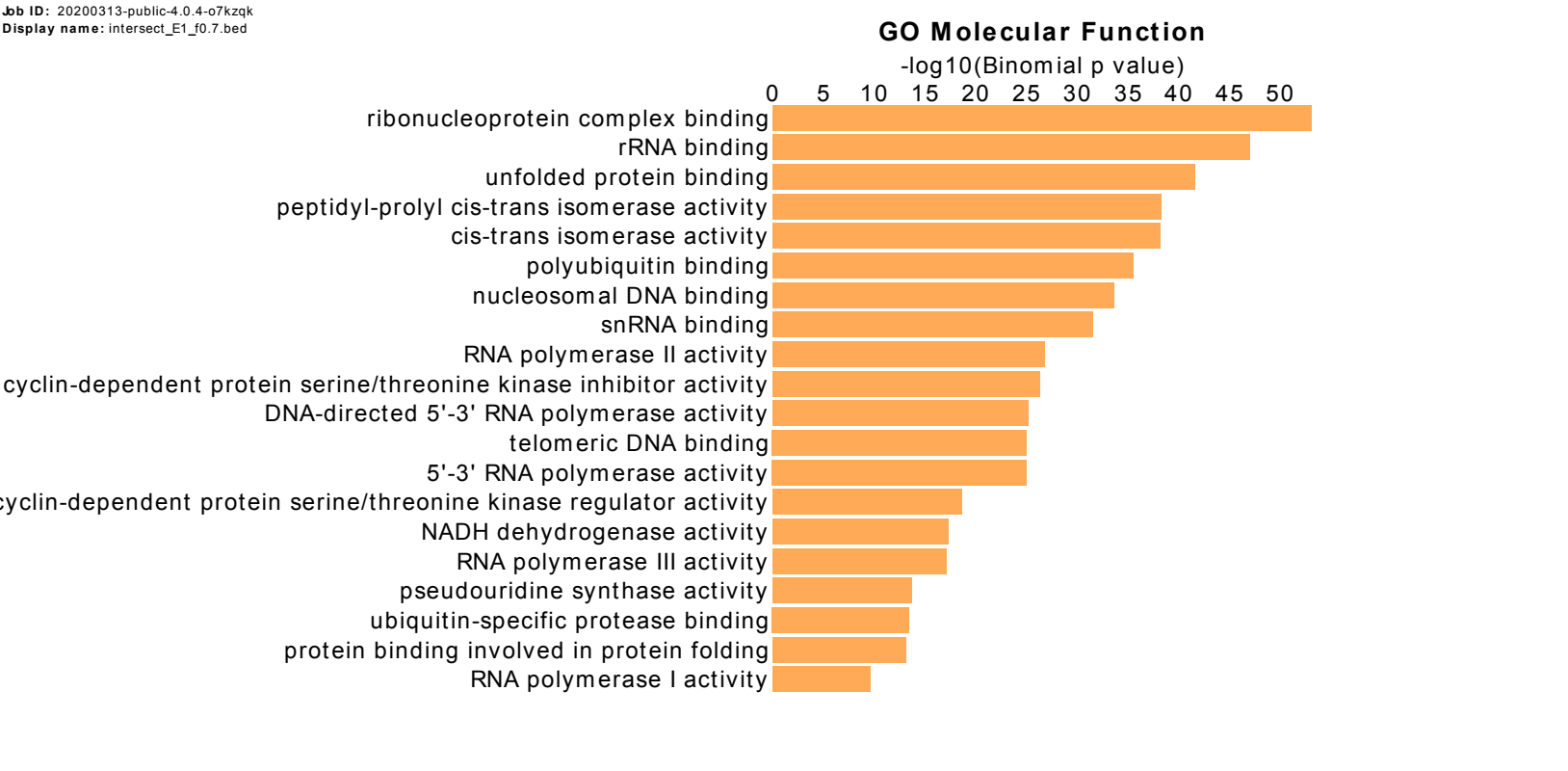
**Figura Anexo 3:** Distancia en kilobases de los segmentos de E1 anotados con respecto al sitio de inicio de la transcripción (Transcription Start Site, TSS).



**Figura Anexo 4:** Top 20 anotaciones GO de procesos biológicos relacionados con los segmentos de E1 obtenidas con GREAT.

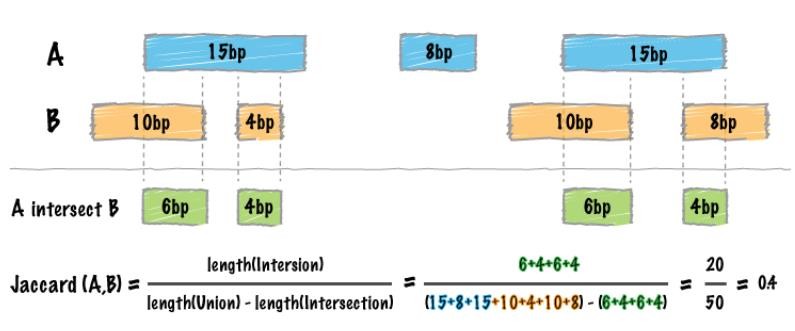


**Figura Anexo 5:** Top 20 términos GO de elementos celulares relacionados con los segmentos de E1 obtenidas con GREAT.

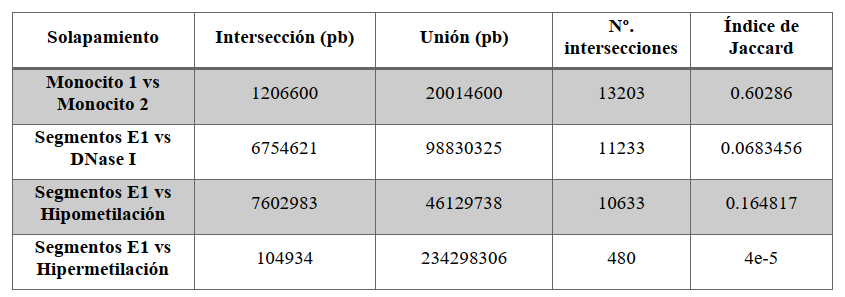


**Figura Anexo 5:** Top 20 términos GO de función molecular relacionados con los segmentos E1 obtenidos con GREAT.

# Anexo: Análisis estadístico de los solapamientos

Adicionalmente al cálculo del porcentaje de segmentos que solapan con la herramienta intersect “bedtools v.2.229.1”, se analiza la similitud de los dos sets de regiones cromosómicas mediante el índice de Jaccard. Este método estadístico mide el ratio entre el número de pares de la intersección y el número de pares de bases de la unión de dos sets de regiones cromosómicas, como se muestra a continuación:

**Figura 26:** Representación del cálculo del índice de Jaccard. Fuente: Documentación bedtools ([https://bedtools.readthedocs.io/en/latest/content/tools/jaccard.html)](https://bedtools.readthedocs.io/en/latest/content/tools/jaccard.html)

En consecuencia, los cálculos de la intersección de los apartados anteriores se acompañan con su correspondiente índice de Jaccard:

En el caso del solapamiento entre ambas réplicas biológicas comprobamos que el índice de Jaccard es elevado (0.60286), lo que demuestra la similitud entre ambas réplicas.

No obstante, en el resto de los solapamientos se observa que el índice de Jaccard es menor. Esto se debe a la diferencia en el número de pares de bases entre los segmentos del estado 1 y el resto de los archivos. Por ejemplo, en el caso del solapamiento con DNase I se observa que el tamaño de la unión es, aproximadamente, el doble de la unión con la Hipometilación, lo que conlleva a un menoíndice de Jaccard a pesar de que, como se demostró anteriormente en el paso 3, los segmentos de E1 solapan en un 85% con la DNase I. Sin embargo, sí queda claro que el solapamiento entre los segmentos del estado 1 y la Hipermetilación es muy pequeño, dado que el índice de Jaccard es mucho menor que el correspondiente al resto de solapamientos

# Bibliografía

Benetatos, L. and Vartholomatos, G., 2018. Enhancer DNA methylation in acute myeloid leukemia and myelodysplastic syndromes. *Cellular and Molecular Life Sciences*, 75(11), pp.1999-2009.

Cavalcante, R.G., Sartor, M.A. 2017. annotatr: genomic regions in context. Bioinformatics, 33(15), pp.2381-2383.

Chatterjee, S. and Pal, J., 2009. Role of 5′- and 3′-untranslated regions of mRNAs in human diseases. *Biology of the Cell*, 101(5), pp.251-262.

Ernst, J. and Kellis, M., 2017. Chromatin-state discovery and genome annotation with ChromHMM. *Nature Protocols*, 12(12), pp.2478-2492.

Lamolle, G. and Musto, H., 2018. *Genoma Humano. Aspectos Estructurales*.

Landolin, J., Johnson, D., Trinklein, N., Aldred, S., Medina, C., Shulha, H., Weng, Z. and Myers, R., 2010. Sequence features that drive human promoter function and tissue specificity. *Genome Research*, 20(7), pp.890-898.

McLean, C., Bristor, D., Hiller, M. *et al.* 2010. GREAT improves functional interpretation of *cis*-regulatory regions. *Nature Biotechnology*, 28**,**pp.495–501.

McLeay, R.C., Bailey, T.L. 2010. Motif Enrichment Analysis: a unified framework and an evaluation on ChIP data. *BMC Bioinformatics*, 11, pp165.

Philip Machanick, Timothy L. Bailey. 2011. MEME-ChIP: motif analysis of large DNA datasets, *Bioinformatics*, 27(12), pp.1696–1697.

Sharifi-Zarchi, A., Gerovska, D., Adachi, K., Totonchi, M., Pezeshk, H., Taft, R., Schöler, H., Chitsaz, H., Sadeghi, M., Baharvand, H. and Araúzo-Bravo, M., 2017. DNA methylation regulates discrimination of enhancers from promoters through a H3K4me1-H3K4me3 seesaw mechanism. *BMC Genomics*, 18(1).

Sullivan, A., Bubb, K., Sandstrom, R., Stamatoyannopoulos, J. and Queitsch, C., 2015. DNase I hypersensitivity mapping, genomic footprinting, and transcription factor networks in plants. *Current Plant Biology*, 3-4, pp.40-47.

Yu, G., Wang, L., Hanm Y., He, Q. 2012. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS: A Journal of Integrative Biology, 16(5), pp.284-287.

[Zhang, M.Y.](https://www.uniprot.org/uniprot/?query=author:%22Zhang+M.Y.%22&sort=score), [Churpek, J.E.](https://www.uniprot.org/uniprot/?query=author:%22Churpek+J.E.%22&sort=score), [Keel, S.B.](https://www.uniprot.org/uniprot/?query=author:%22Keel+S.B.%22&sort=score), [Walsh, T.](https://www.uniprot.org/uniprot/?query=author:%22Walsh+T.%22&sort=score), [Lee, M.K.](https://www.uniprot.org/uniprot/?query=author:%22Lee+M.K.%22&sort=score), [Loeb, K.R.](https://www.uniprot.org/uniprot/?query=author:%22Loeb+K.R.%22&sort=score), [Gulsuner, S.](https://www.uniprot.org/uniprot/?query=author:%22Gulsuner+S.%22&sort=score), [Pritchard, C.C.](https://www.uniprot.org/uniprot/?query=author:%22Pritchard+C.C.%22&sort=score), [Sanchez-Bonilla, M.](https://www.uniprot.org/uniprot/?query=author:%22Sanchez-Bonilla+M.%22&sort=score), [Delrow, J.J.](https://www.uniprot.org/uniprot/?query=author:%22Delrow+J.J.%22&sort=score), [Basom, R.S.](https://www.uniprot.org/uniprot/?query=author:%22Basom+R.S.%22&sort=score), [Forouhar, M.](https://www.uniprot.org/uniprot/?query=author:%22Forouhar+M.%22&sort=score), [Gyurkocza, B.](https://www.uniprot.org/uniprot/?query=author:%22Gyurkocza+B.%22&sort=score), [Schwartz, B.S.](https://www.uniprot.org/uniprot/?query=author:%22Schwartz+B.S.%22&sort=score), [Neistadt, B.](https://www.uniprot.org/uniprot/?query=author:%22Neistadt+B.%22&sort=score), [Marquez, R.](https://www.uniprot.org/uniprot/?query=author:%22Marquez+R.%22&sort=score), [Mariani, C.J.](https://www.uniprot.org/uniprot/?query=author:%22Mariani+C.J.%22&sort=score), [Coats, S.A.](https://www.uniprot.org/uniprot/?query=author:%22Coats+S.A.%22&sort=score) 2015. Germline ETV6 mutations in familial thrombocytopenia and hematologic malignancy. *Nature Genetics*, 47, pp.189.185.