

Trabajo Epigenómica: Active promoter (H3K4me3,+H3K27Ac)

Autores:

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

Objetivo

En este trabajo el objetivo es analizar y estudiar el estado 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.

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 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.

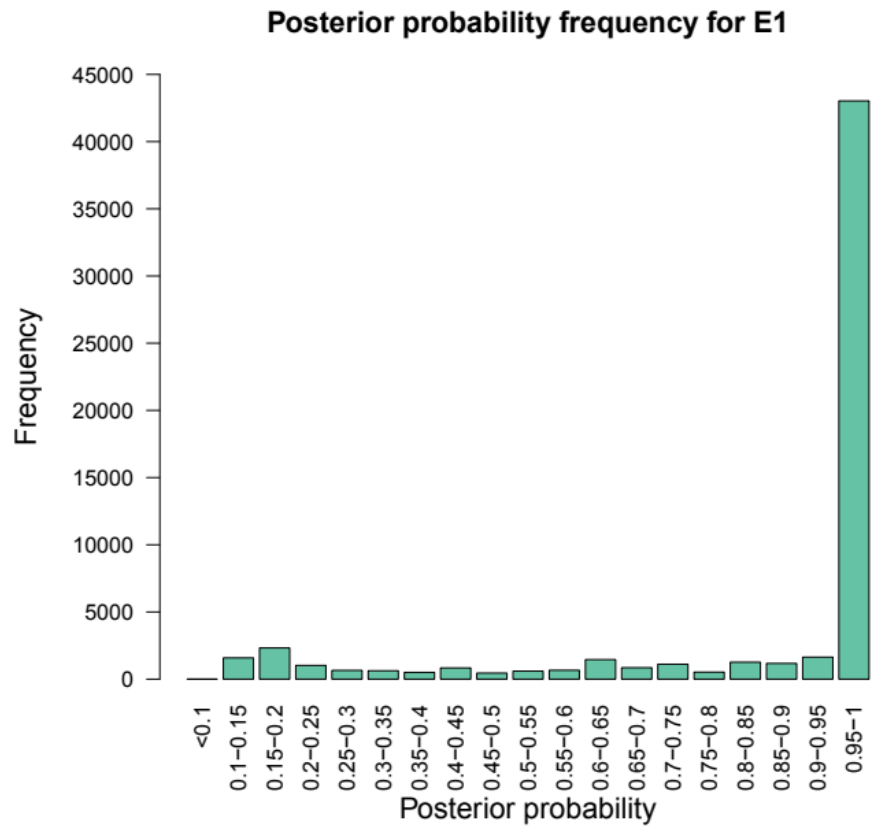


Figura 1- Distribución de los segmentos del estado 1 en función de la probabilidad posterior

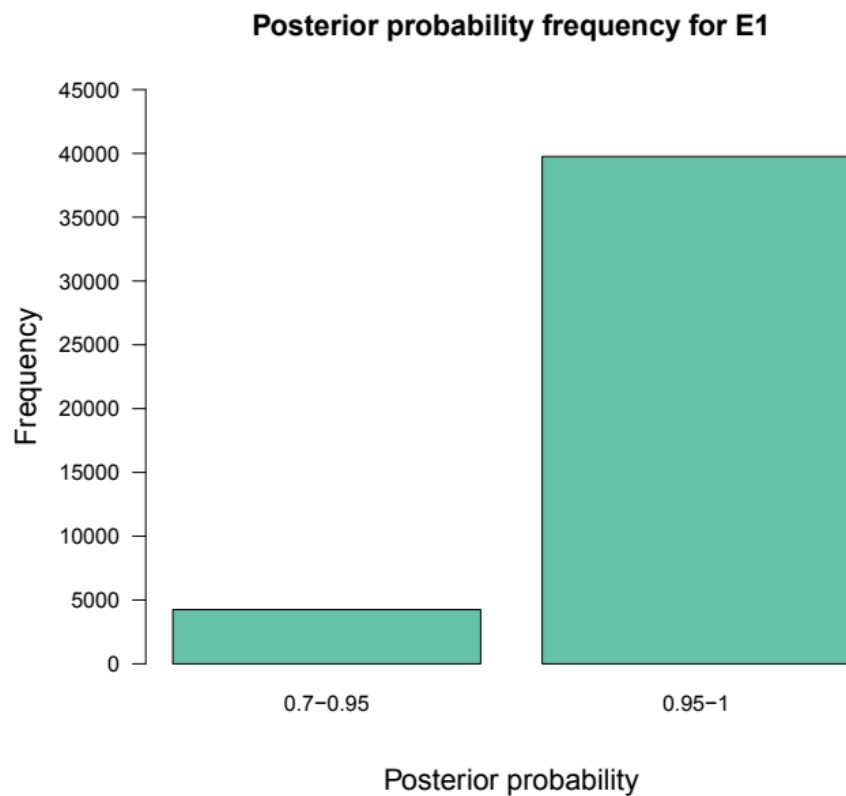


Figura 2: Segmentos seleccionados para el análisis (44008).

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:

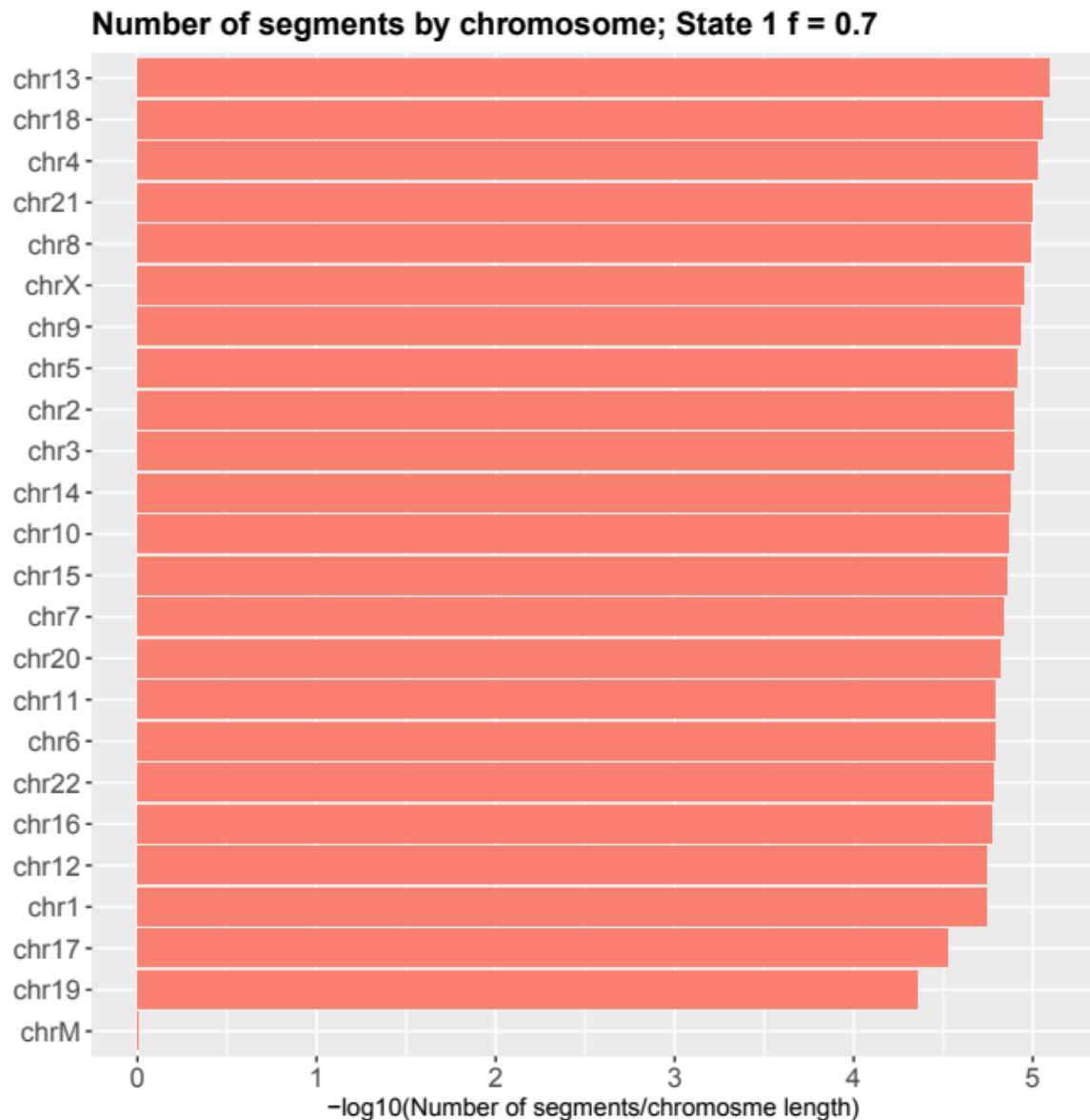


Figura 2: Número de segmentos por cromosoma tras normalizar por el tamaño de cada cromosoma. Resultado obtenido tras la intersección de los segmentos solapantes en ambas réplicas biológicas de monocito.

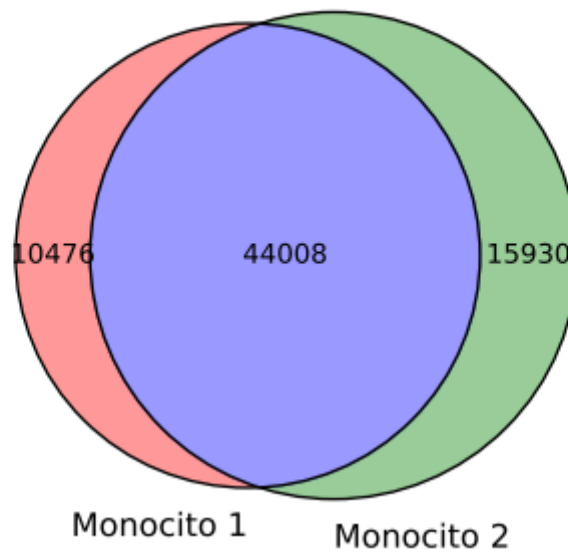
Venn diagram: State 1 $f=0.7$ 

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: Anotar los segmentos. Como mínimo, se deberá dar el porcentaje de segmentos que solapan con *protein-coding genes* en estado.

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
Enhancers_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_promoters	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 al 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.

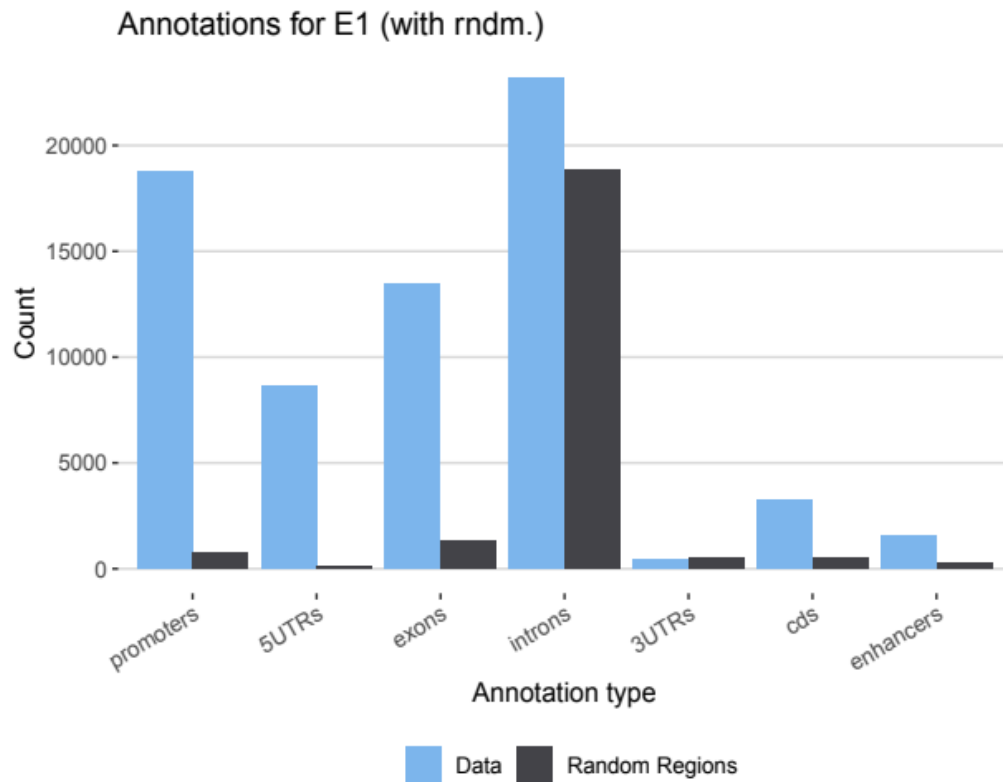


Figura 4: Comparación de la anotación en nuestros datos con datos generados aleatoriamente

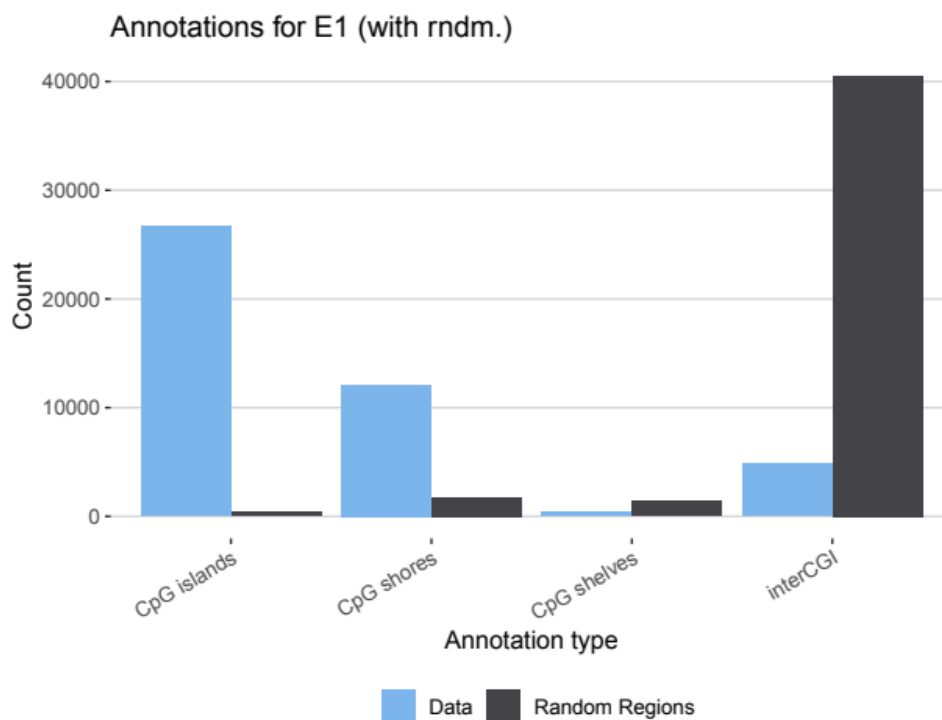


Figura 5: Comparación de la anotación de islas CpG de nuestros datos con datos generados aleatoriamente.

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.

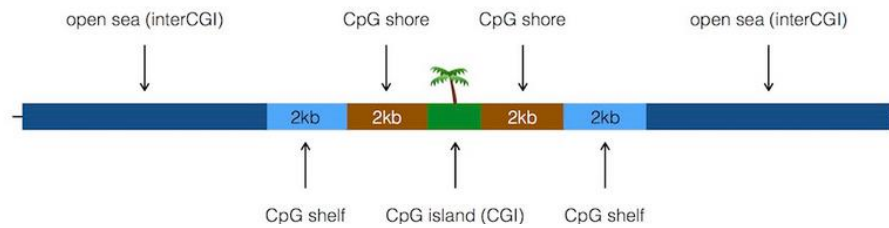
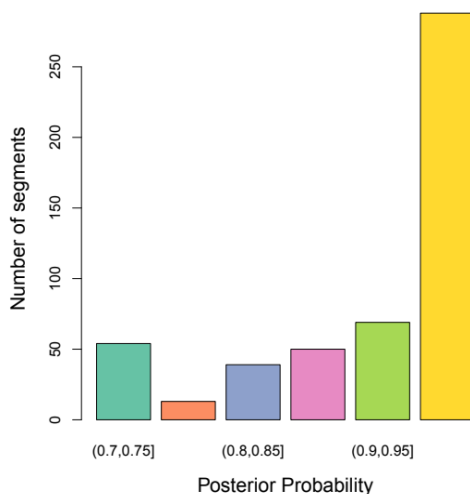


Figura 6: Esquema de la anotación de islas CpG.

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 6, 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. De nuevo, los resultados señalan que el E1 se relaciona con la transcripción, promoviéndola.

Igualmente, se estudió la distribución de la anotación en función de la probabilidad posterior.

Distribution of CDS annotation among E1 posterior probability



Distribution of promoter annotation among E1 posterior probability

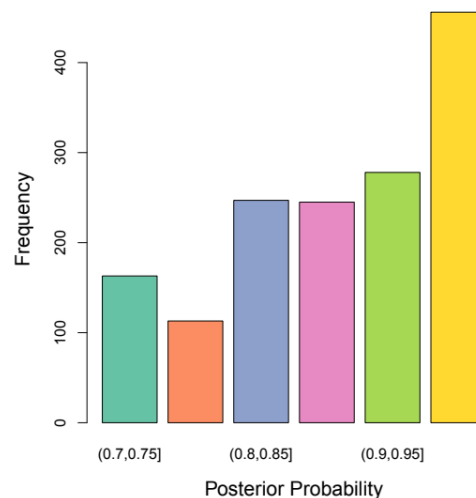


Figura 7: Distribución de la probabilidad posterior de anotaciones CDS y promotores.

Por ejemplo, en las imágenes anteriores observamos que la mayoría de segmentos anotados proceden de la región con una probabilidad posterior de 0.95 o superior. Esto apoya la seguridad de las anotaciones, puesto que los segmentos que más se asocian con el E1 son aquellos que más relacionados se encuentran con estos términos. Igualmente, se observa que recuperar los segmentos con una probabilidad posterior superior o igual a 0.7 permite ampliar el número de anotaciones de estos términos.

Asimismo, con el objetivo de obtener mayor información y completar los resultados extraídos del análisis con “annotatr”, 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. A continuación, se muestran los resultados de la anotación de los segmentos de E1 seleccionados empleando los parámetros de solapamiento por defecto en GREAT (región basal más extensión de 5 kbs aguas arriba, 1 kb aguas abajo y 1000 kbs para regiones distales).

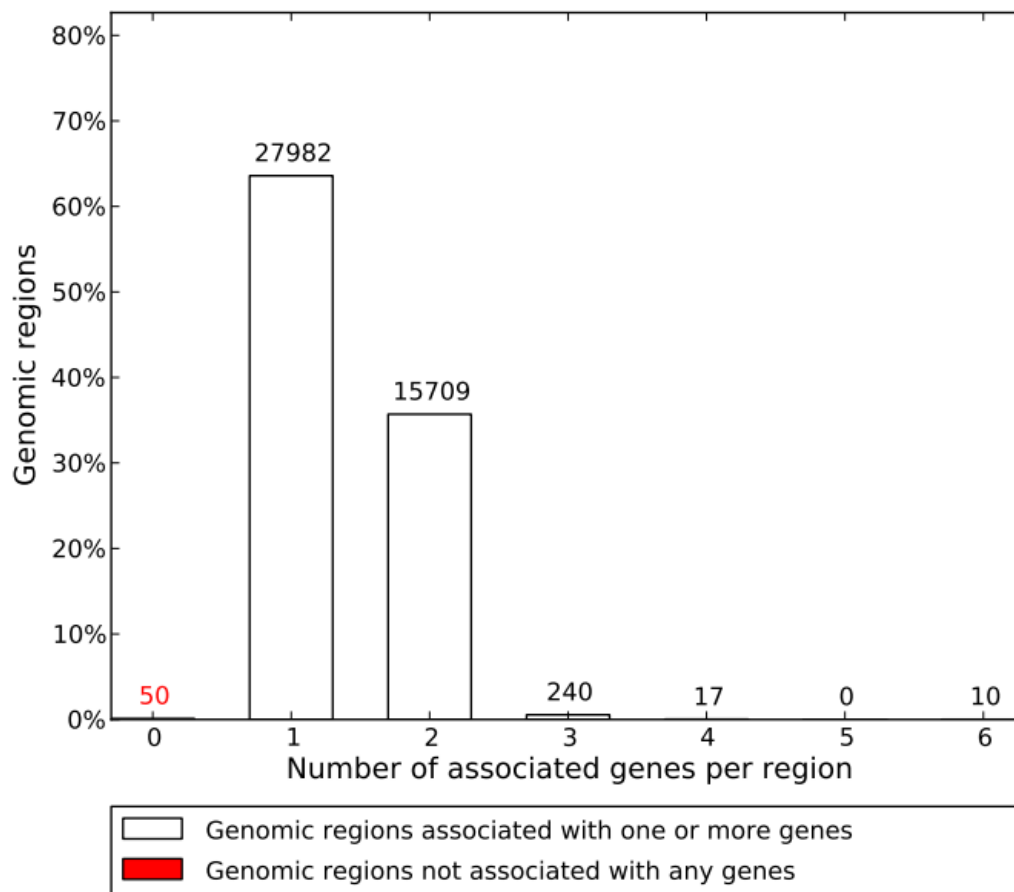


Figura 6: Diagrama de barras que muestra el número de segmentos de E1 asociados a genes

Como se observa en la Figura 3, tan solo 50 (0.1%) de los 44008 segmentos del E1 no se encuentran relacionados con ningún gen. Con el paquete “annotatr” se estudia con mayor detalle con qué tipo de elemento se asocian los 43958 segmentos restantes, que solapan con 10931 genes de los 18,549 genes presentes en GREAT.

En la Figura 6, 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.

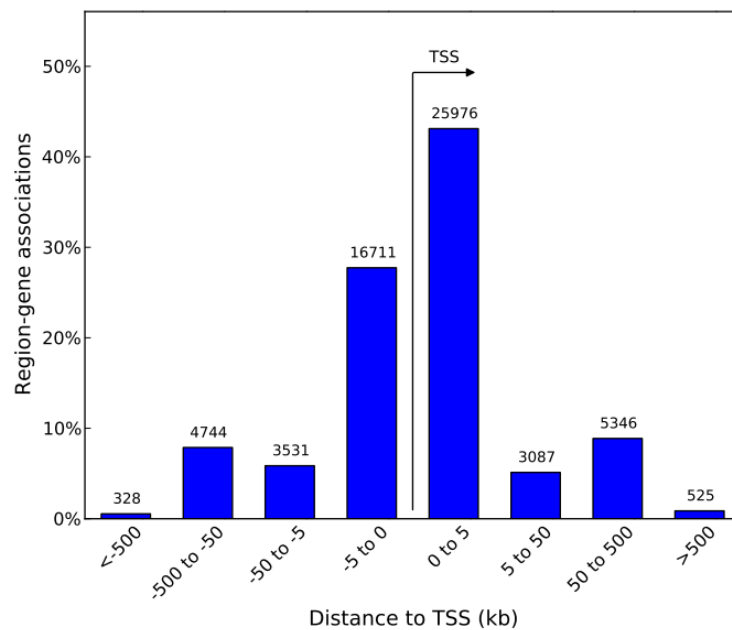


Figura 7: Distancia en kilobases de los segmentos de E1 anotados con respecto al sitio de inicio de la transcripción (*Transcription Start Site*, TSS).

Igualmente, los resultados de las anotaciones GO validan que los segmentos de E1 estudiados pertenecen a células del sistema inmune relacionadas con la defensa frente a patógenos, y que se encuentran relacionados con la transcripción y la regulación epigenética de la cromatina. Por ejemplo, en la Figura 5 se observa que, entre los términos GO sobre procesos biológicos asociados a nuestros segmentos de interés, hay 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...*). En el paso 4 de este trabajo, se observarán algunos genes relacionados con estos términos GO en el *UCSC genome browser*.

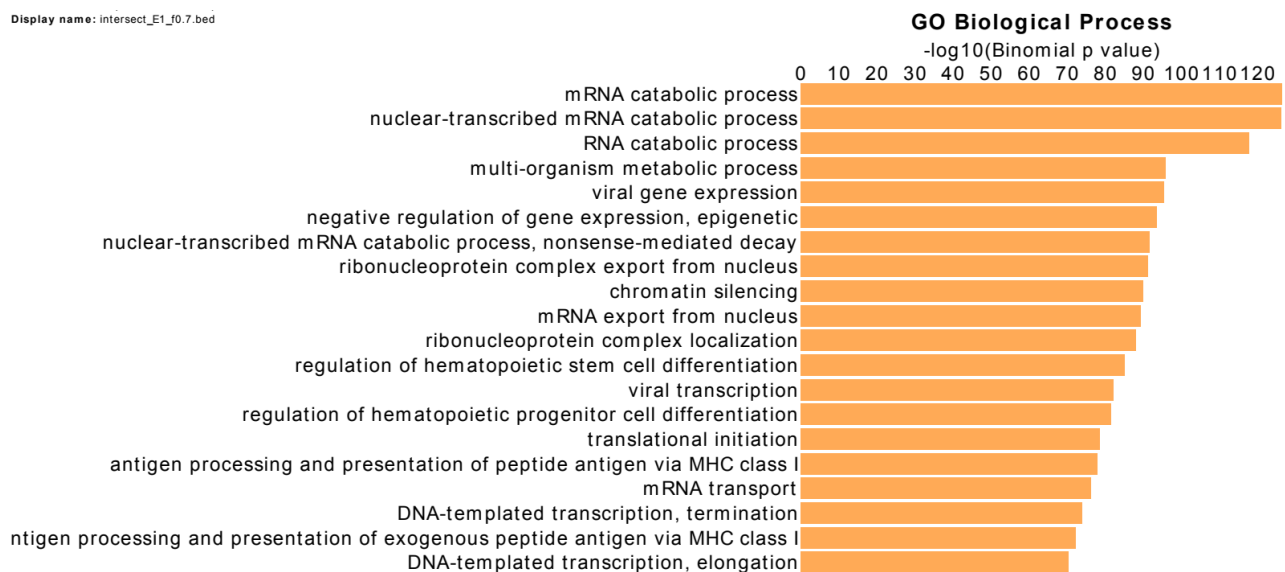


Figura 8: Top 20 anotaciones GO de procesos biológicos relacionados con los segmentos de E1.

Job ID: 20200313-public-4.0.4-o7kzqk
Display name: intersect_E1_f0.7.bed

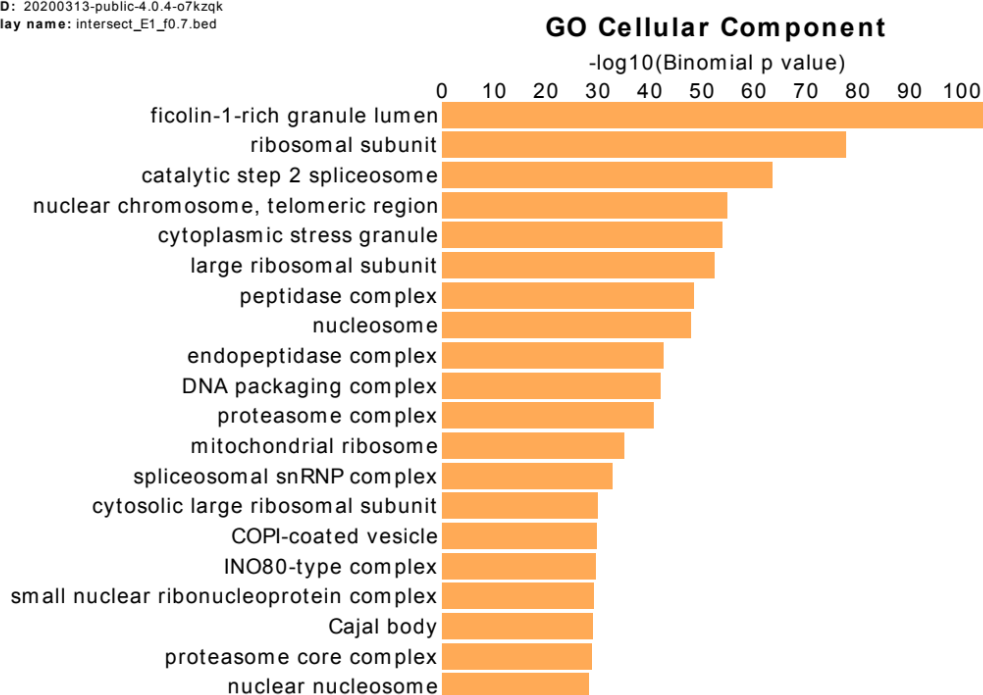


Figura 9: Top 20 términos GO de elementos celulares relacionados con los segmentos de E1.

Job ID: 20200313-public-4.0.4-o7kzqk
Display name: intersect_E1_f0.7.bed

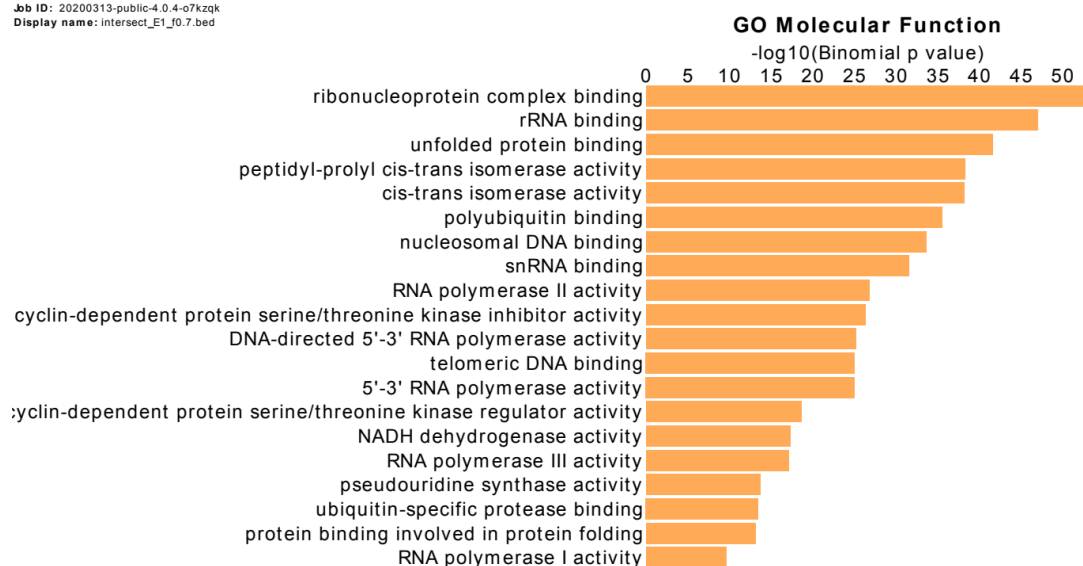


Figura 10: Top 20 términos GO de función molecular relacionados con los segmentos E1.

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 3 análisis de enriquecimiento frente a distintas bases de datos: anotación de términos GO, para la validación de los resultados de GREAT; anotación de rutas KEGG (<https://www.genome.jp/kegg/>), para conocer los procesos en los que están implicados los genes anotados; y, finalmente, 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 último caso, dado que el estado de cromatina que estamos estudiando está relacionado con la 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.

Respecto a la anotación con términos GO, los resultados obtenidos con “clusterProfiler” son muy parecidos a los presentados anteriormente, lo que nos permite validarlos.

En cuanto a los resultados del enriquecimiento con rutas KEGG, se puede observar en la Figura 11 cómo claramente la mayoría de las entradas obtenidas son rutas relacionadas con el sistema inmune, presentando además un p-valor ajustado muy bajo.

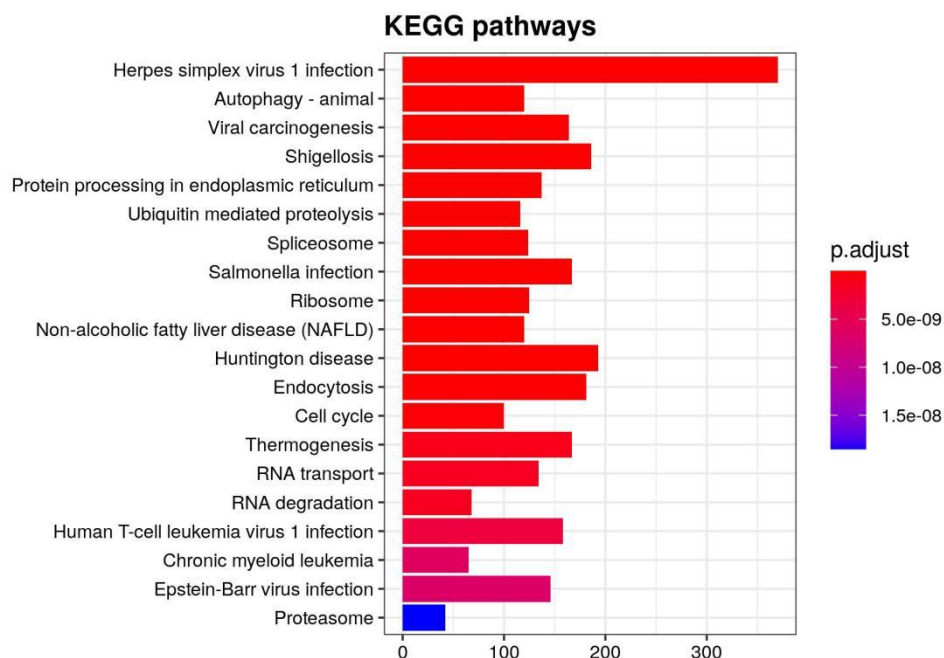


Figura 11: Top 20 rutas KEGG relacionadas con los genes relacionados con los segmentos E1.

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.

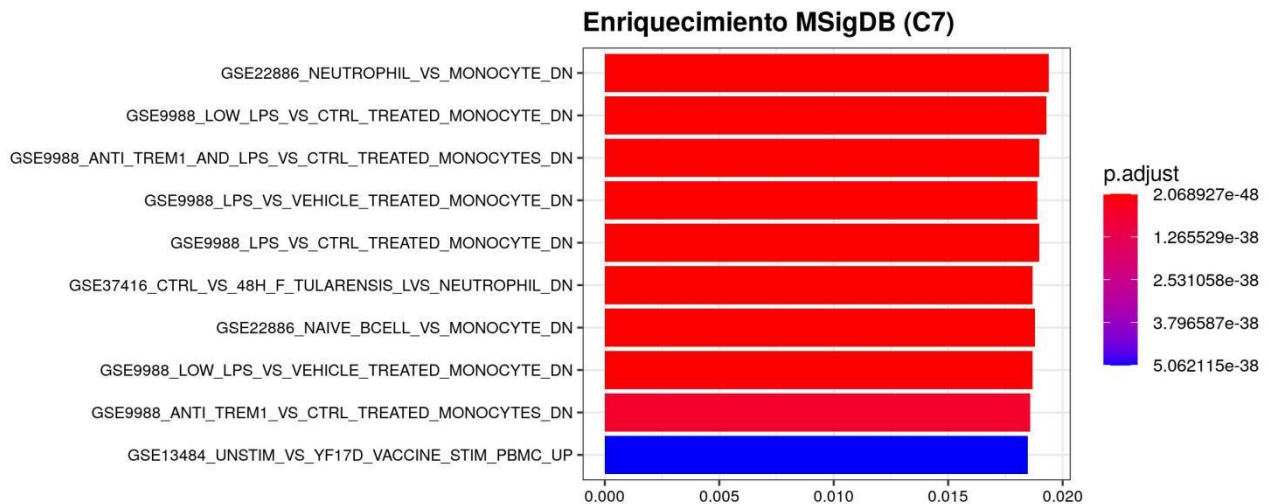


Figura 12: Top 20 entradas resultado de enriquecimiento contra MSigDB (colección C7: células inmunes).

Paso 3: Descargar los picos de DNase I en monocitos de ENCODE y 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 nucleótidos, 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.

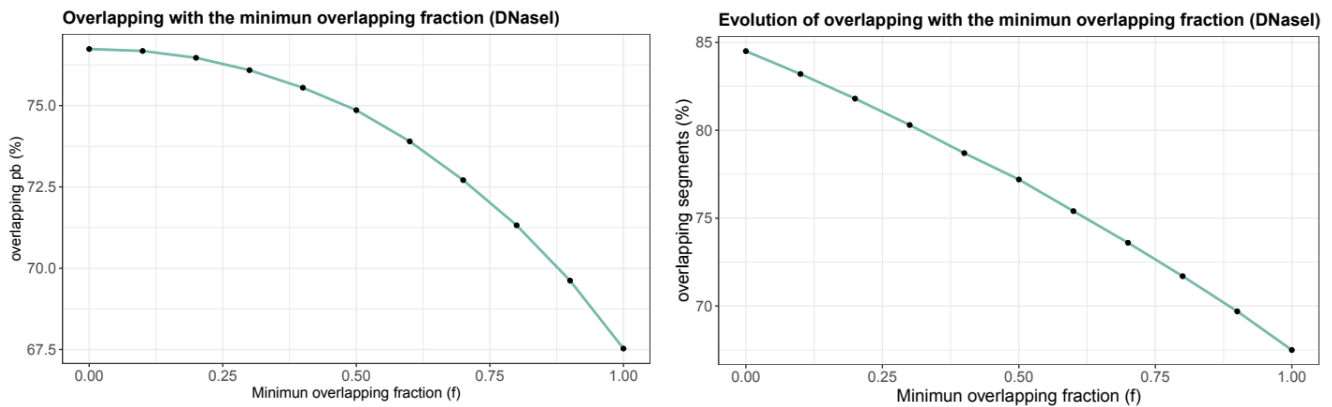


Figura 13: Evolución del número de regiones solapantes con respecto a la fracción mínima solapante.

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.

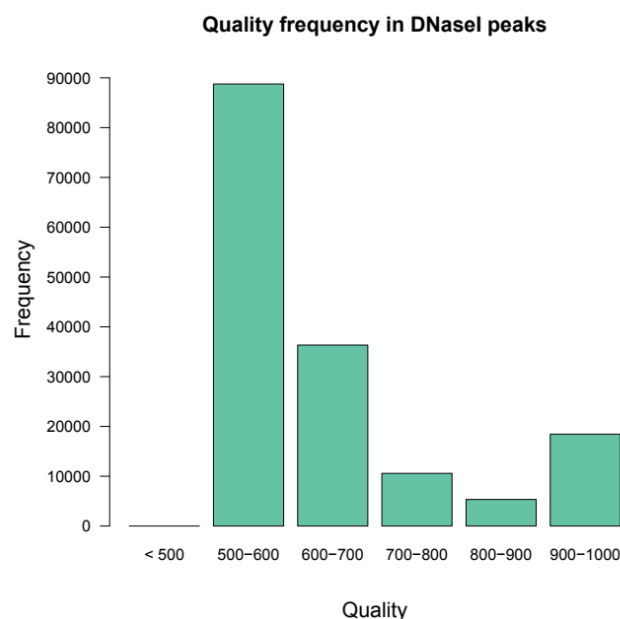


Figura 14: Histograma de la calidad de las regiones accesibles a la DNase I en monocitos CD14+ procedentes de ENCODE.

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.

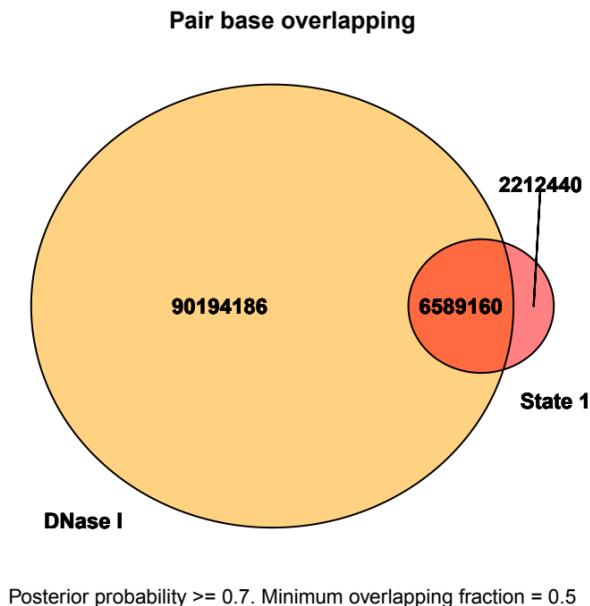


Figura 15: Diagrama de Venn con el número de pares de bases solapantes entre la DNase I y el E1.

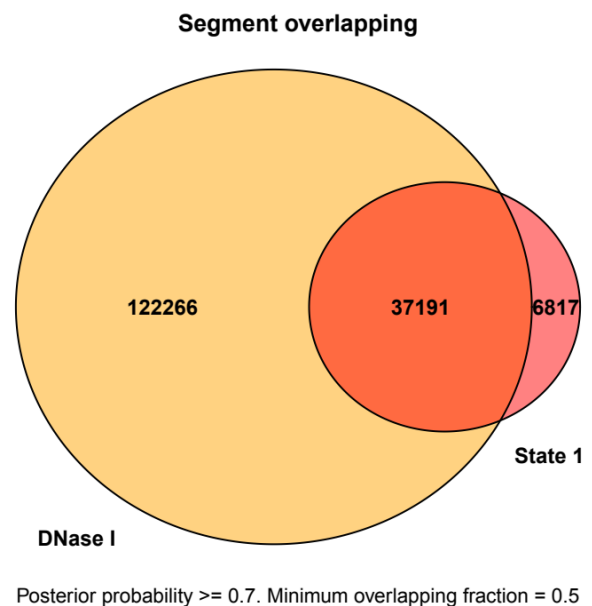


Figura 16: Diagrama de Venn con el número de segmentos solapantes entre la DNase I y el E1.

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

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, de los cuales se seleccionaron algunos 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/macrófagos 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).

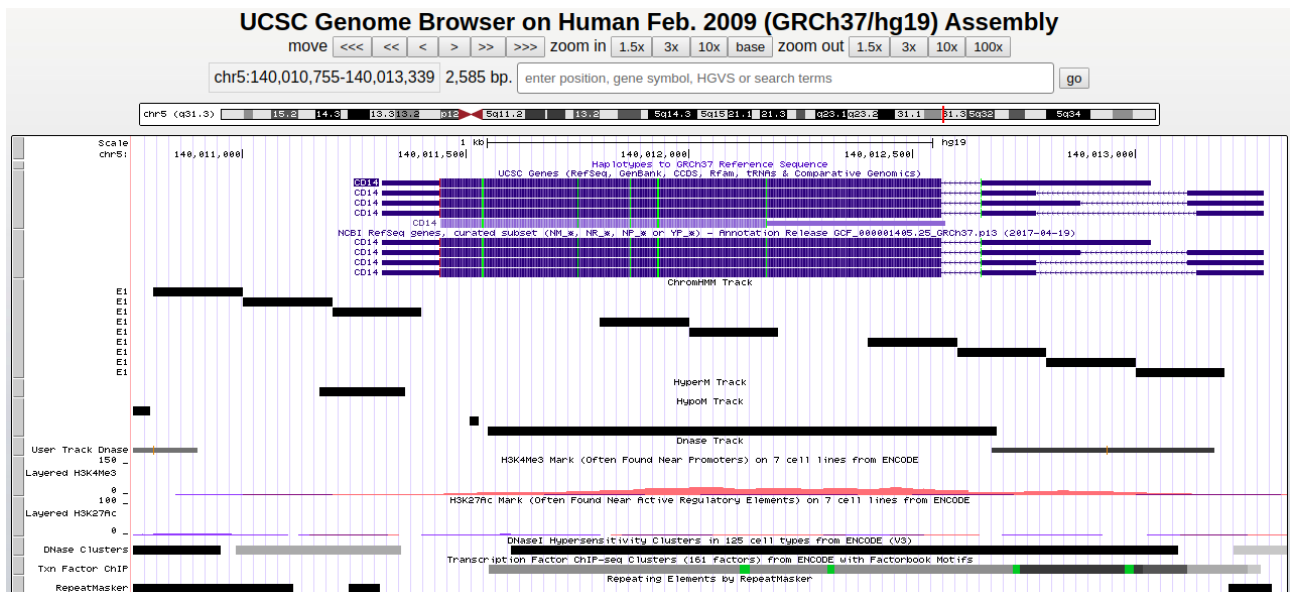


Figura 17: Visualización del gen CD14 en el navegador genómico de la UCSC.

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.

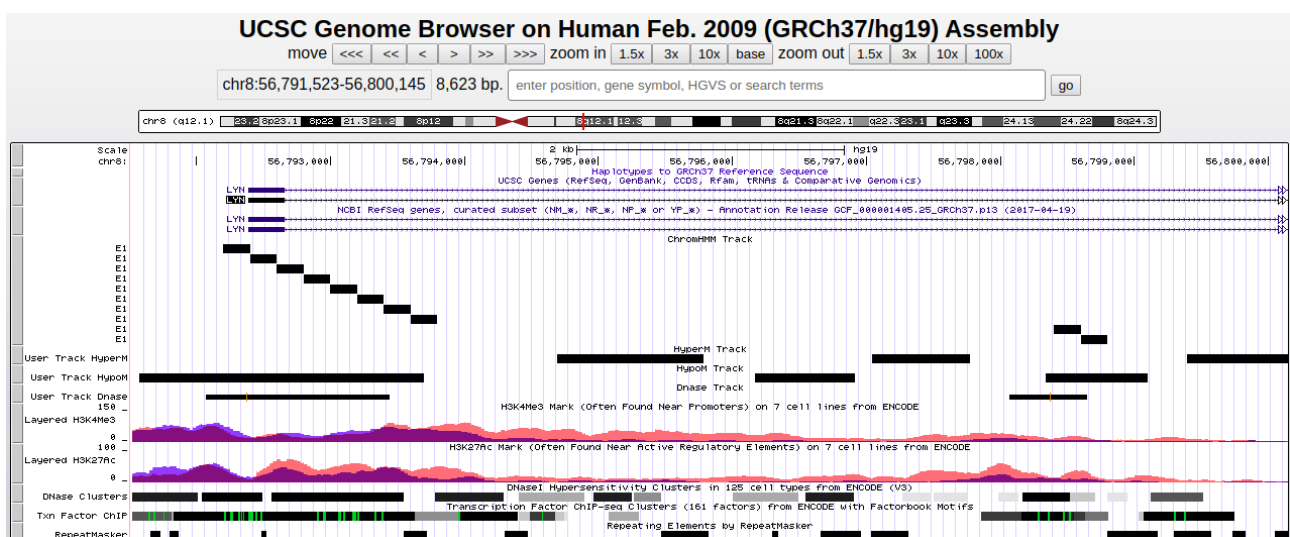


Figura 18: Visualización del gen LYN en el navegador genómico de la UCSC.

El gen EIF2A codifica para un factor de transcripción iniciador de la transcripción mediante la formación de los complejos de preiniciación 80S sensibles a puomicina y la síntesis de poly(U) a

bajas concentraciones de magnesio celular. 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. Sin embargo, este caso incluye el hecho de que dicha región con las características descritas presenta al gen SERP1 en dirección contraria (unos cuantos cientos de pbs aguas arriba del TSS de EIF2A), el cual se relaciona con la síntesis de proteínas. Esto sugiere la posibilidad de que se trate de un promotor bidireccional. 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.

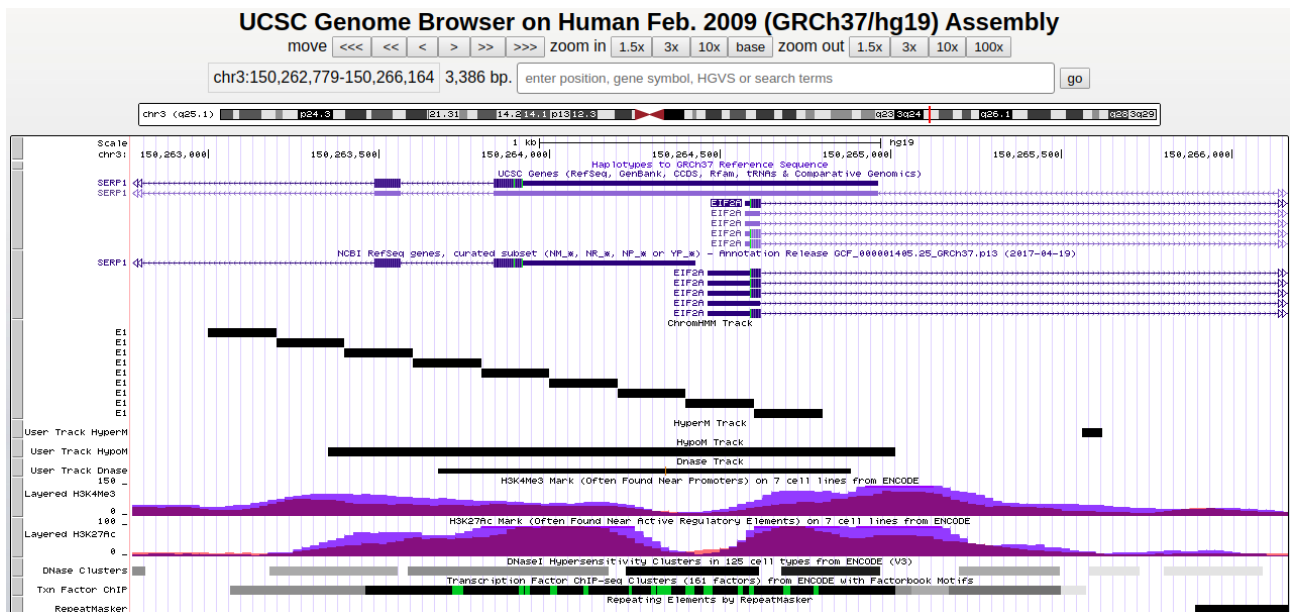


Figura 19: Visualización del gen EIF2A en el navegador genómico de la UCSC.

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.

En esta sección se muestran los pasos seguidos para la búsqueda de motivos enriquecidos en los segmentos. Al igual que en el resto del documento, calcularemos el enriquecimiento que se han detectado los dos ficheros y que tienen asociado una probabilidad posterior de estado 1 mayor que 0.7. El objetivo 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), cuyo funcionamiento se describe a continuación:

1. Descubrimiento Ab-initio de motivos, que identifica nuevos patrones en las regiones de ChIP-seq que pueden deberse a sitios de unión de TF.
2. Búsqueda de motivos asociados con sitios de unión de TF conocidos.
3. Visualización de motivos para mostrar las ubicaciones relativas y las fuerzas de unión de los TF en las regiones identificadas.
4. Análisis de la afinidad de unión al ADN de cada región de entrada para el TF correspondiente.
5. Comparación de motivos ab-initio con los motivos de unión de TF conocidos.

MEME-ChIP es, por tanto, una herramienta potente que proporciona varias herramientas 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
```

```
GCGCGGGGCTGCTCGGGCCGGGAAGCCGCGGAGTCGCGTGAGCACCGCCCGCCGGGCCCTGTGCCCCGCTTTCGGTCAGGCCTC  
CTGGGCCCCGTGCGCAGTCCGGGCCGGCGGGGAACGCGGCTCCGGGCGTGTGGCGGGCGGAGAGGAAGCGTTTGTGGCGCGGC  
ACGTCGTGTGCTAGCCCGGGAGCGGCGGGCAG
```

```
>chr10:182000-182200
```

```
GGCTGGCCCCGAGGCCCGGGGCGGAGATTCGTGCGTCCCCGGGGCTGTGAGCGACCCCGGCAGCGAGGCCCGCCCTCAACAG  
GTCCTCGCGGACCTCCGGGACAGTCTGTGGGGTCCGCCGCCCTCTCCACCCAGAGCTCCGGGGAGAAGCTGCTGAGGACGCG  
GCTGGGACGAGGGGGGGCGCCGGGACCCGGA
```

```
>chr10:976000-976200
```

```
CGCCAGGCAGGTTGGGGAAGTATGGAAATTGCACACTGCACACACCACCTGACCACCAAGAAAAAGAGGACACCACTCTTCCCCA  
CTGTGGTCTTCACAACTACAACCTCAGCACCTATTTCAAGTGCAATTTAAATATTCAAACAACCTGAACTAAGAGAAAAAACTTCGAA  
AGGGAAAATAACTTGTCTGTCCCCT
```

```
>chr10:976200-976400
```


```
TCAATGTTACCTTGGAGTTGCTCTCATCTCTAGAAAAAGGCCTCACGCATACTGTTTATACTTACAACTCCACTGCCCTGTCAAGGA  
AAACCTTCCTGTTTTCTCAGTTATCCAAAACCTCAAACCTCAATTTACTATTAAGACAAGTTATTGGGCATCCACTAGAGAATCTCAA  
ACGAAAGCCAGATAGGAGGTA
```

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.

Data Submission Form

Perform motif discovery, motif enrichment analysis and clustering on large nucleotide datasets.

Select the motif discovery and enrichment mode ?

☒ Classic mode ☐ Discriminative mode ☐ Differential Enrichment mode 

Select the sequence alphabet

Use sequences with a standard alphabet or specify a custom alphabet. ?

☒ DNA, RNA or Protein ☐ Custom No file chosen

Input the primary sequences

Enter the (equal-length) nucleotide sequences to be analyzed. ?

intersect_E1_10.7.fasta ?

Input the motifs

Select, upload or enter a set of known motifs. ?

?

?

Input job details

(Optional) Enter your email address. ?

(Optional) Enter a job description. ?

Note: if the combined form inputs exceed 80MB the job will be rejected.

Figura 20: Configuración utilizada en MEME-ChIP.

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. 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.

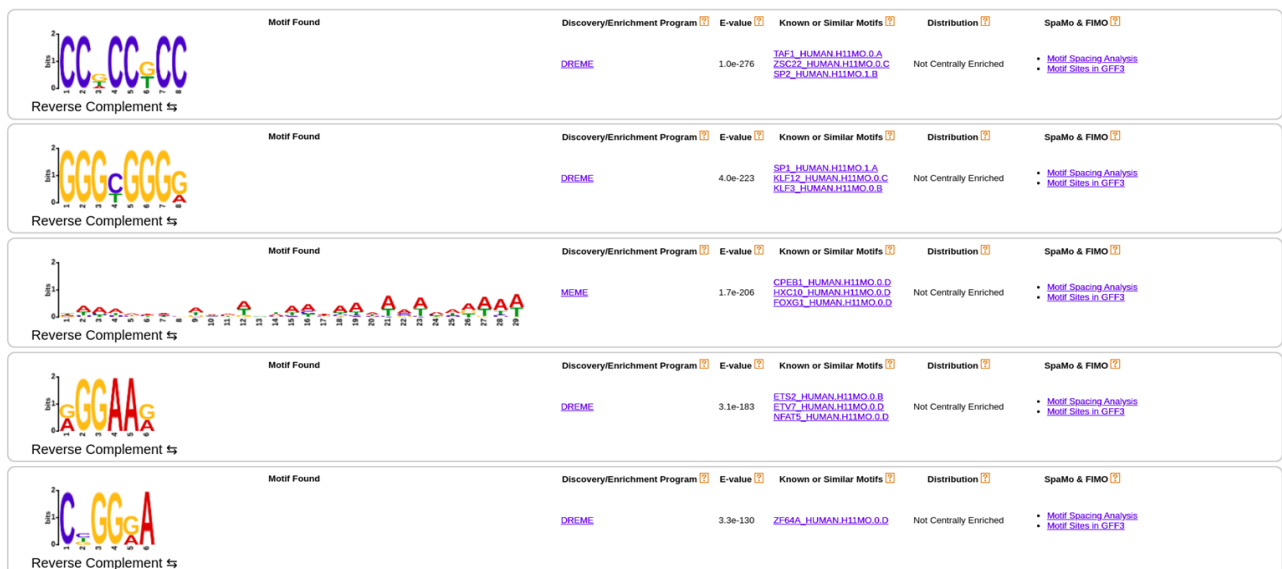


Figura 21: Motivos de unión *de novo* encontrados por MEME, Dreme y CentriMo.

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.

Rank	Motif	P-value	log P-value	% of Targets	% of Background	STD(Bg STD)	Best Match/Details
1		1e-662	-1.525e+03	62.43%	49.05%	55.8bp (71.4bp)	Elk4(ETS)/Hela-Elk4-ChIP-Seq(GSE31477)/Homer(0.807) More Information Similar Motifs Found
2		1e-312	-7.196e+02	52.32%	43.14%	55.5bp (68.6bp)	YY2/MA0748.2/Jaspar(0.802) More Information Similar Motifs Found
3		1e-225	-5.192e+02	69.30%	61.81%	55.2bp (68.6bp)	YY2/MA0748.2/Jaspar(0.659) More Information Similar Motifs Found
4		1e-218	-5.028e+02	35.68%	28.58%	56.2bp (70.0bp)	ZBED1/MA0749.1/Jaspar(0.720) More Information Similar Motifs Found
5		1e-202	-4.651e+02	38.72%	31.71%	55.4bp (67.1bp)	ETV6/MA0645.1/Jaspar(0.795) More Information Similar Motifs Found
6		1e-181	-4.190e+02	37.85%	31.23%	55.3bp (68.1bp)	PB0179.1_Sp100_2/Jaspar(0.650) More Information Similar Motifs Found
7		1e-163	-3.764e+02	37.49%	31.22%	55.9bp (72.1bp)	PB0192.1_Tefap2e_2/Jaspar(0.811) More Information Similar Motifs Found
8		1e-109	-2.518e+02	12.63%	9.32%	55.8bp (68.0bp)	NFYA/MA0060.3/Jaspar(0.853) More Information Similar Motifs Found

Figura 22: Top 8 motivos enriquecidos mediante HOMER.

Como podemos observar en la Figura 22, 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 (); 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.

Regiones hiper e hipometiladas

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. Las regiones hipometiladas, en concreto las relacionadas con promotores e islas CpG (elementos enriquecidos en nuestros segmentos de E1) se asocian con una regulación positiva de la transcripción. En conclusión, estos resultados apoyan a los obtenidos anteriormente en la afirmación de que el E1 y sus marcas de histonas corresponden a marcas transcripcionalmente activas.

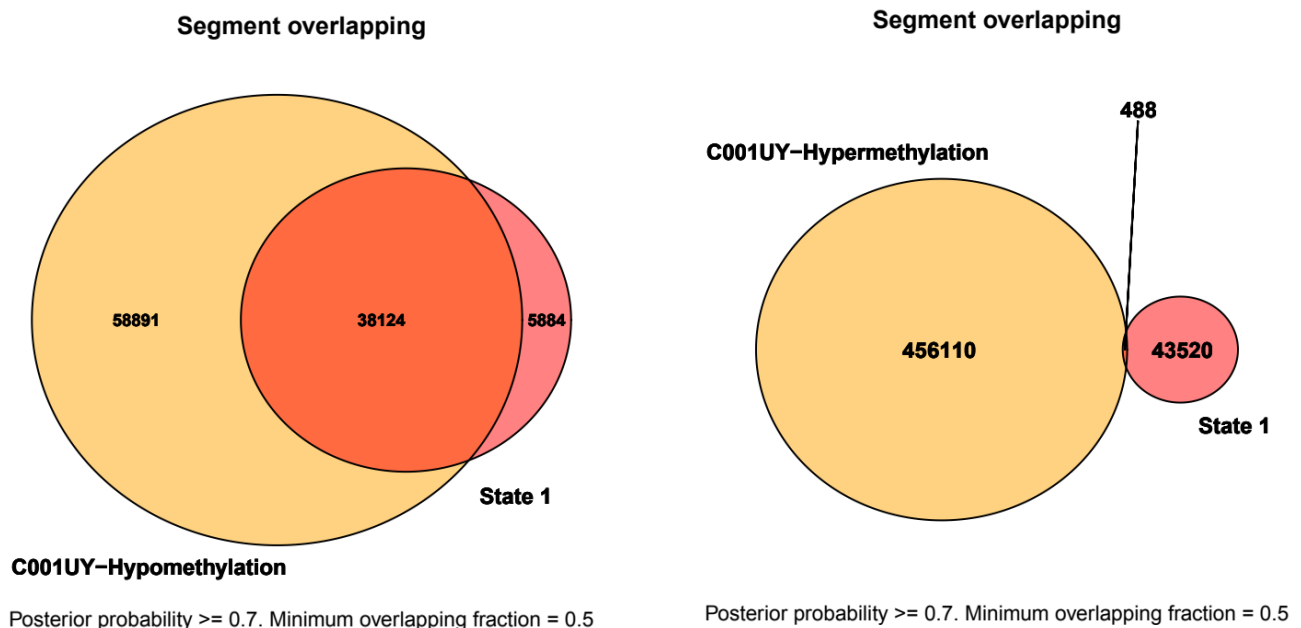


Figura 22: Solapamiento entre los segmento E1 y regiones hipo- e hipermetiladas.

Las regiones hipometiladas, en concreto las relacionadas con promotores e islas CpG (elementos enriquecidos en nuestros segmentos de E1), se asocian con una regulación positiva de la transcripción. En conclusión, estos resultados apoyan a los obtenidos anteriormente en la afirmación de que el E1 y sus marcas de histonas corresponden a marcas transcripcionalmente activas.

La metilación de citosinas del DNA y de histonas es un mecanismo muy importante en la regulación de la transcripción. Las regiones promotoras e islas CpG (situadas éstas últimas mayormente 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. 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.

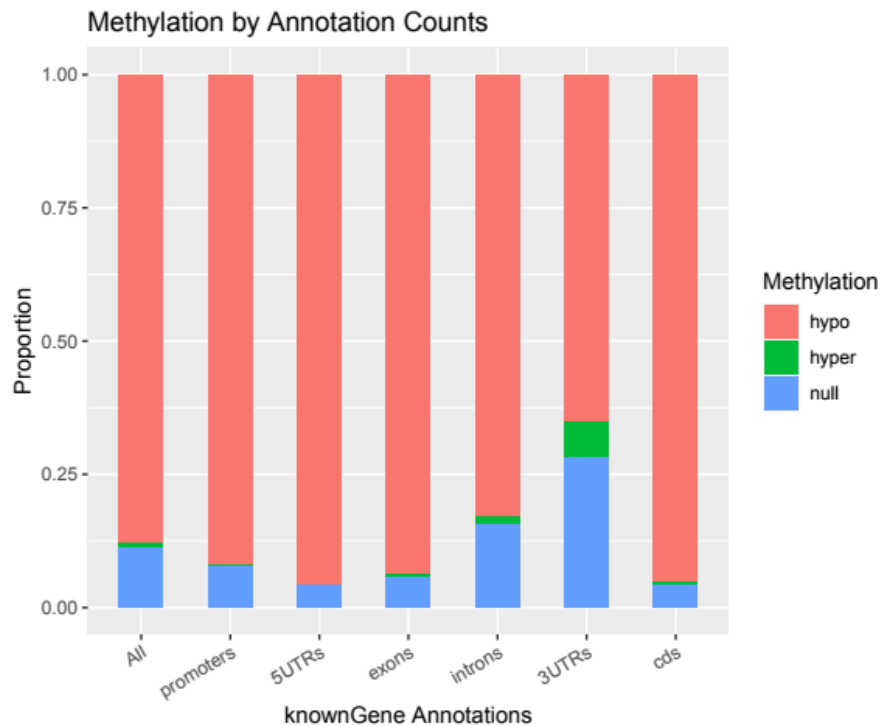


Figura 23: Proporción de segmentos de E1 que corresponden a segmentos que solapan con regiones hipometiladas, hipermetiladas o ninguna de ellas, en función de las diferentes anotaciones generadas.

En la Figura 23, comprobamos en primer lugar que los segmentos solapantes con las regiones hipometiladas representan la mayor proporción de segmentos, seguidos por los que no solapan con regiones hipo o hipermetiladas ('null') y, en último lugar, los segmentos que solapan con regiones hipermetiladas. Cabe destacar también, que la proporción de regiones hipometiladas es superior a la suma de segmentos hipermetilados y 'null'. Entre las distintas anotaciones, destacamos el hecho de que la anotación de promotores, 5'UTRs, CDS y exones, todas ellas relacionadas con una activación de la transcripción, presentan la mayor proporción de segmentos solapantes con regiones hipometiladas y sin apenas segmentos solapantes con regiones hipermetiladas. Por el contrario, los intrones y las regiones 3'UTRs muestran una menor proporción de segmentos hipometilados (aunque siguen siendo la proporción superior al 50%) y un mayor contenido en regiones hipermetiladas. El efecto de la metilación fuera de los promotores es hoy día un debate en la comunidad científica, pero el hecho de que los intrones y las regiones 3'UTRs muestren un patrón de metilación algo diferente nos indica que las marcas del estado E1 se asocian con mayor claridad en las zonas 5' donde se lleva a cabo la transcripción, situándose así en mayor medida en promotores e islas CpG, como se comprobó anteriormente.

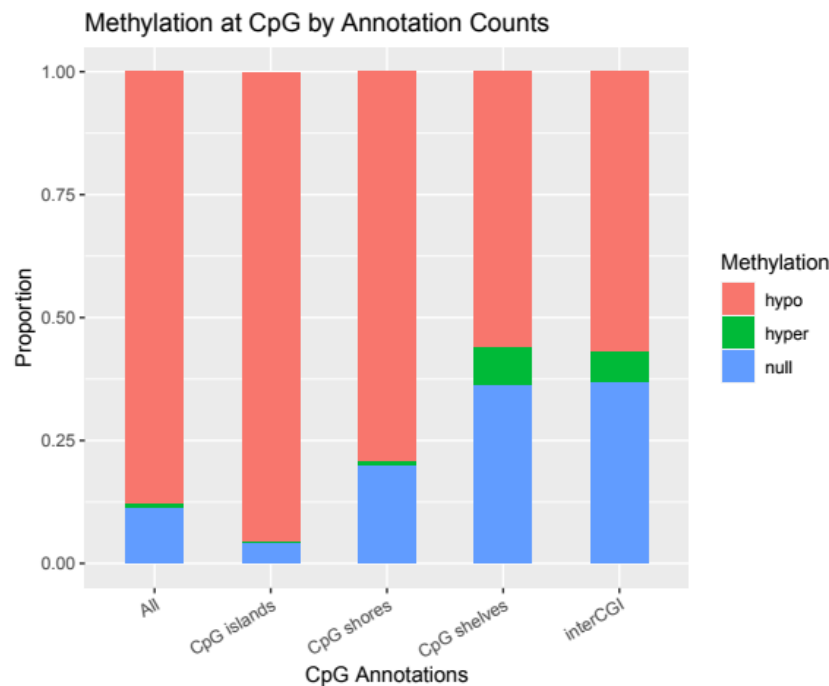


Figura 24: - Proporción de segmentos de E1 que corresponden a segmentos que solapan con regiones hipometiladas, hipermetiladas o ninguna de ellas, en función de las diferentes anotaciones generadas para las islas CpG.

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.

Para facilitar la visualización de los resultados, 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. En estas figuras, comprobamos 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 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.

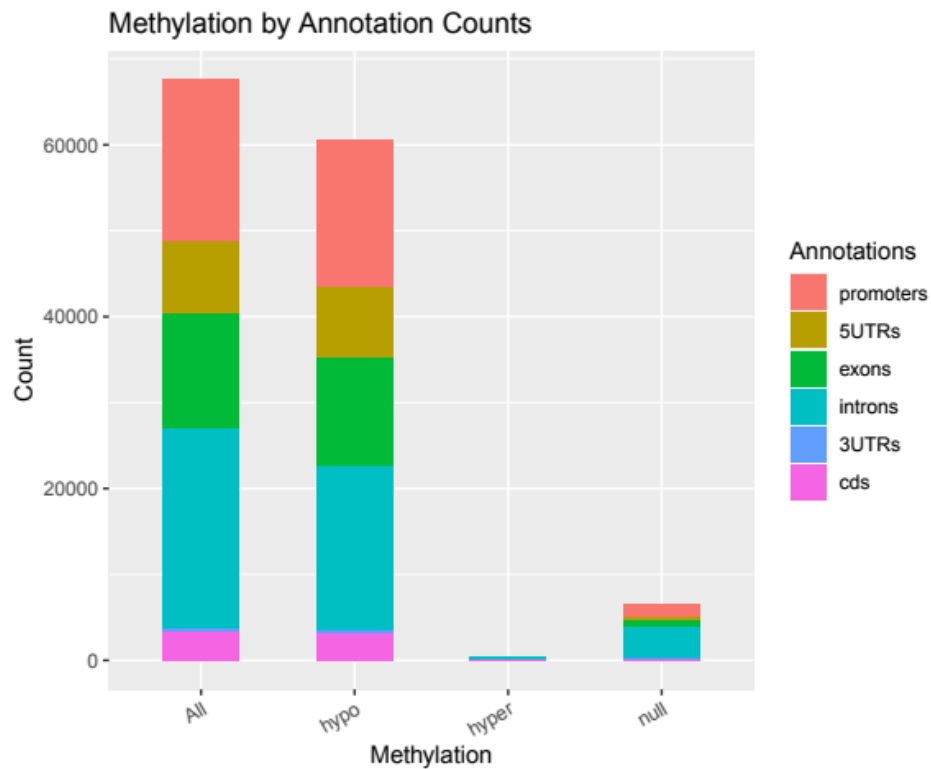


Figura 25: Distribución de las anotaciones en función de los distintos niveles de metilación de los segmentos del estado E1.

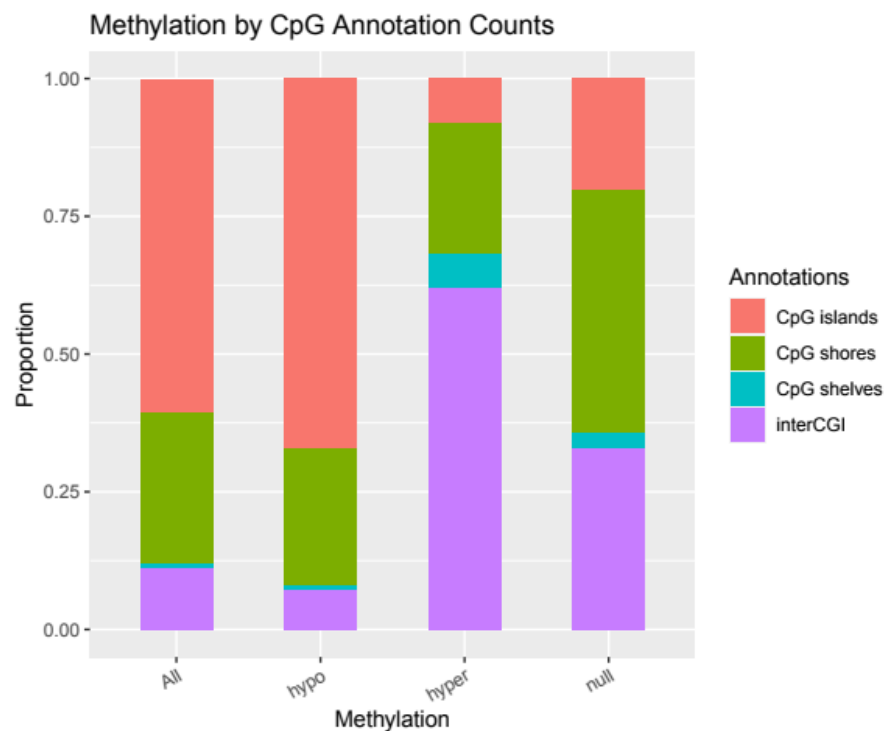


Figura 26: - Distribución de las anotaciones de las islas CpG en función de los distintos niveles de metilación de los segmentos del estado E1.

Material adicional: 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:

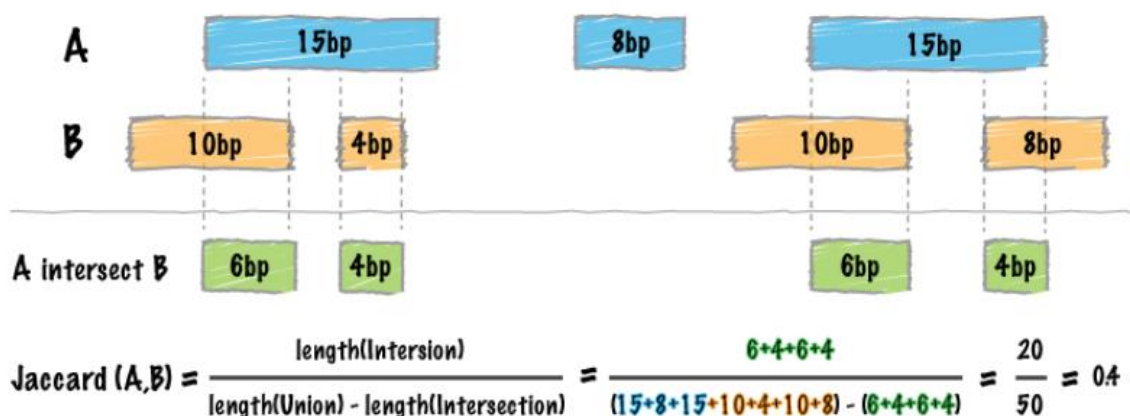


Figura 26: - Representación del cálculo del índice de Jaccard. Fuente: Documentación bedtools (<https://bedtools.readthedocs.io/en/latest/content/tools/jaccard.html>)

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

Solapamiento	Intersección (pb)	Unión (pb)	Nº. intersecciones	Índice de Jaccard
Monocito 1 vs Monocito 2	1206600	20014600	13203	0.60286
Segmentos E1 vs DNase I	6754621	98830325	11233	0.0683456
Segmentos E1 vs Hipometilación	7602983	46129738	10633	0.164817
Segmentos E1 vs Hipermetilación	104934	234298306	480	4e-5

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 menor

índice 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.

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 transcripción (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.
- Las islas CpG, situándose a su vez estos segmentos en regiones hipometiladas.
- 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”...)

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.

Bibliografía

- Alberts, B., 2016. *Introducción A La Biología Celular*. Buenos Aires [etc.]: Médica Panamericana.
- 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.
- 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., Churpek, J.E., Keel, S.B., Walsh, T., Lee, M.K., Loeb, K.R., Gulsuner, S., Pritchard, C.C., Sanchez-Bonilla, M., Delrow, J.J., Basom, R.S., Forouhar, M., Gyurkocza, B., Schwartz, B.S., Neistadt, B., Marquez, R., Mariani, C.J., Coats, S.A. 2015. Germline ETV6 mutations in familial thrombocytopenia and hematologic malignancy. *Nature Genetics*, 47, pp.189.185.