Trabajo Epigenómica

# Autores:

* Sara Dorado Alfaro
* Diego Mañanes
* 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

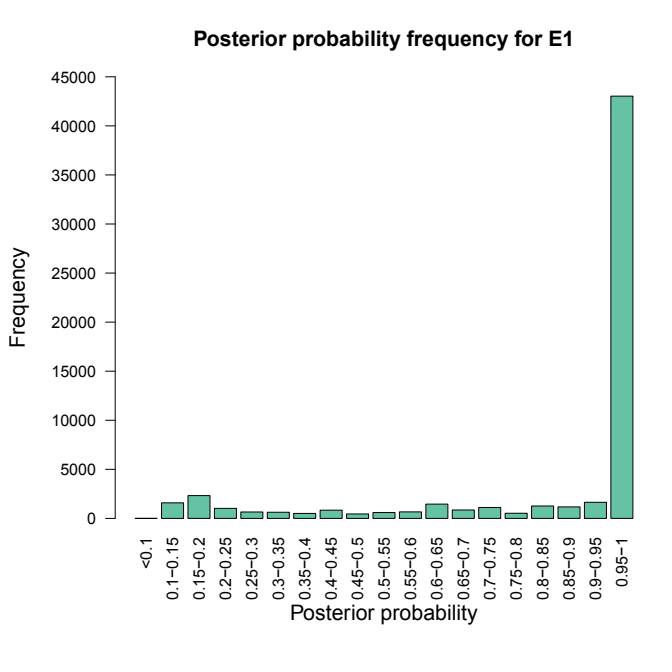
En este trabajo, el material de partida son los segmentos de cromatina asignados a uno de los 11 estados calculados por ChromHMM, como se realizó en la práctica del día 27 de febrero en el aula. El software ChromHMM emplea los modelos ocultos de Markov (“Hidden Markov Models, HMM) para calcular distintos 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.

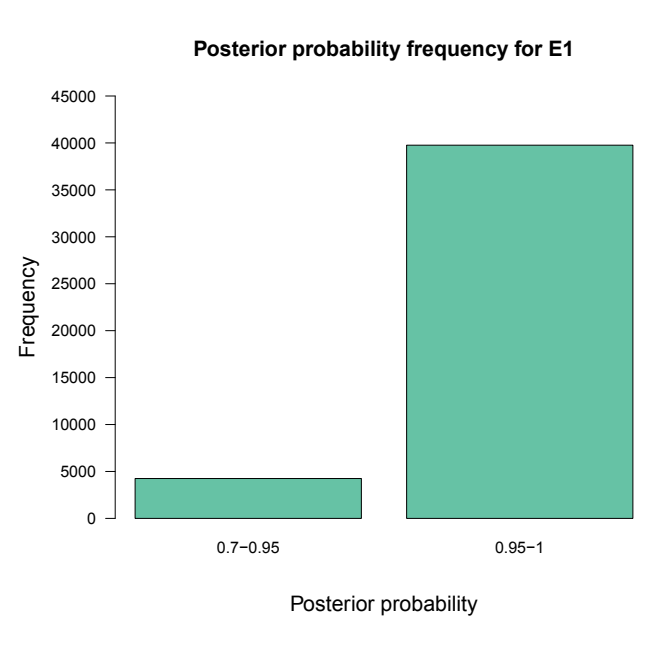
# Análisis

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

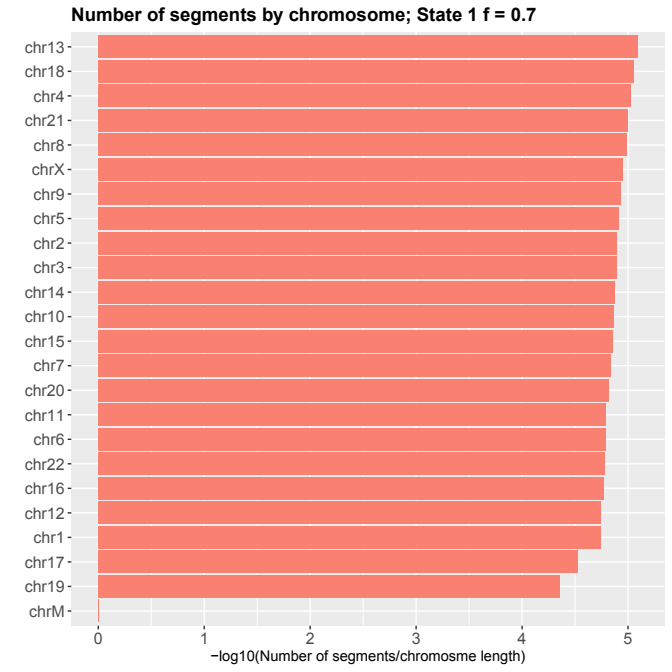
La cromatina empleada en este trabajo procede de dos réplicas biológicas de monocitos CD14+ CD16- de humano. El primer paso de nuestro trabajo es generar los archivos de partida del estudio. Para ello, se procede a calcular el número de segmentos de 200 pbs que solapan en ambos replicados biológicos. 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.

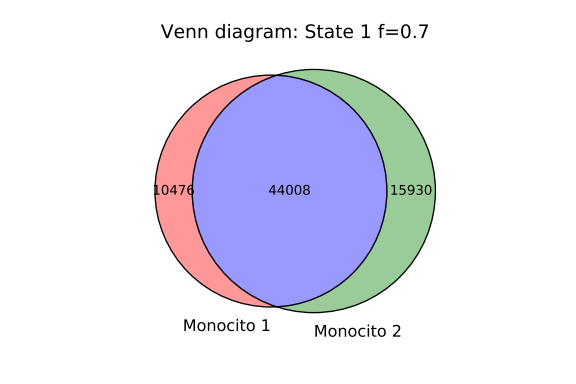
Entre los archivos de la carpeta “POSTERIOR” hay un documento por cromosoma, en el que cada línea corresponde a un segmento de 200 pbs y cada columna indica la probabilidad posterior de dicho de segmento a pertenecer a cada uno de los 11 estados. De este modo, se estableció como umbral de selección de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

  
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 3- 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 4: 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 emplean los siguientes paquetes de R:

* annotatr (<https://www.bioconductor.org/packages/release/bioc/html/annotatr.html>)
* ChIPseeker (<https://bioconductor.org/packages/release/bioc/html/ChIPseeker.html>)

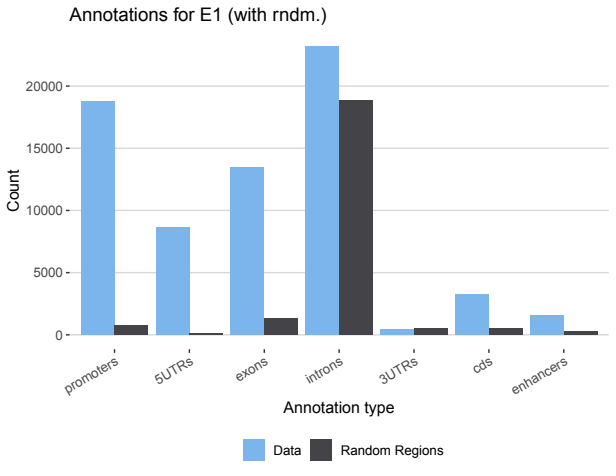
Igualmente, se emplean herramientas de anotación y caracterización funcional disponibles en la web:

* GREAT: Genomic Regions Enrichment of Annotations Tool (<http://great.stanford.edu/public/html/splash.php>)
* PANTHER Classification Analysis (<http://www.pantherdb.org/>)

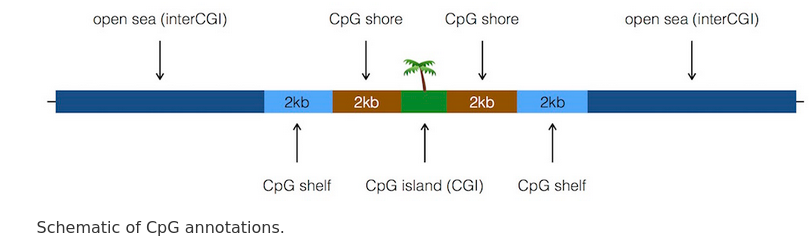
El cálculo del número de segmentos que solapan con genes codificantes para proteínas se realizó mediante la anotación de los segmentos que pertenecían a CDS mediante el paquete “annotatr”. Se emplea este paquete porque dispone de la función “summarize\_annotations”, que muestra un resumen de las anotaciones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



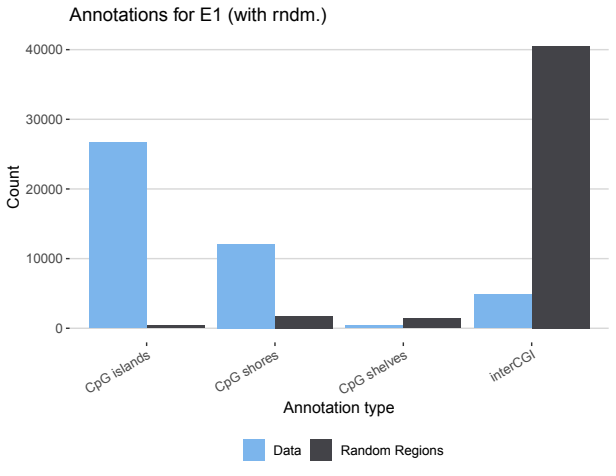
Con las anotaciones obtenidas, se comprueba que un 7% de los segmentos se asocian con regiones codificantes para proteínas (CDS). Es importante señalar que se estableció como mínimo un solapamiento de 100 pbs entre los segmentos de 200 pbs y las coordenadas del genoma anotado para asignar la anotación 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 5- Comparación de la anotación en nuestros datos con datos generados aleatoriamente

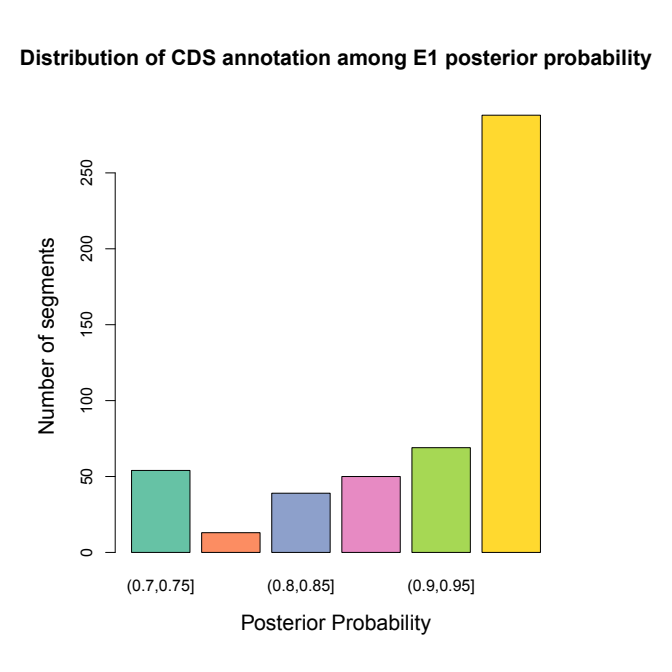
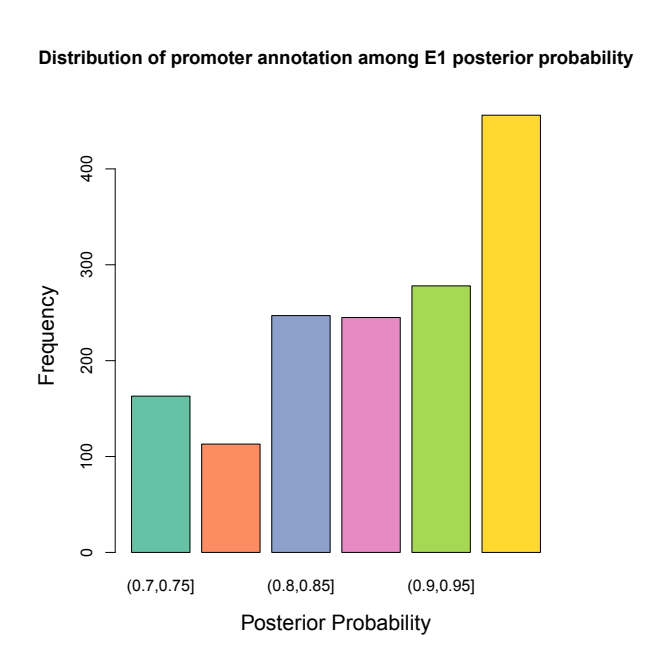
Entre los resultados obtenidos de la anotación, se comprueba que las 3 anotaciones más abundantes son las islas CpG, intrones y promotores. Para comprobar si esta anotación está enriquecida en nuestros segmentos con respecto el genoma, se generaron fragmentos aleatorios del mismo tamaño que los de estudio y se anotaron. 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 presencia por azar de intrones en la anotación es muy elevada, además de que corresponden al 50% del genoma del ser humano (Lamolle and Musto, 2018), por lo que tiene sentido que su anotación sea abundante. 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 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 nuestro datos. De nuevo, los resultados señalan que el E1 se relaciona con la transcripción, promoviéndola.

  
Figura 6- Comparación de la anotación de islas CpG de nuestros datos con datos generados aleatoriamente

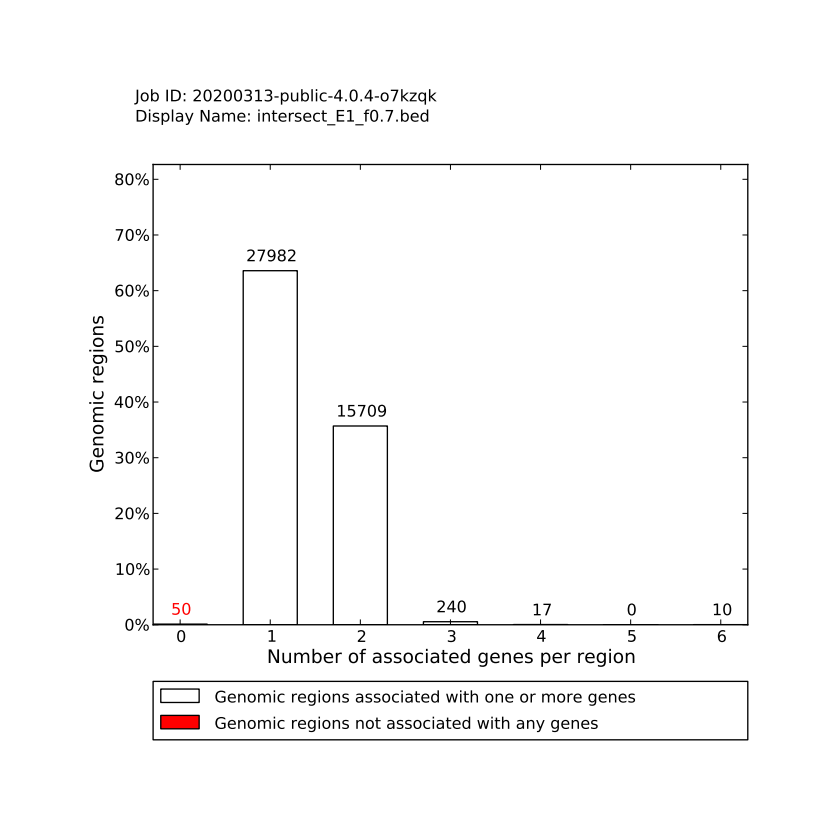
Igualmente, se estudió la distribución de la anotación en función de la probabilidad posterior.



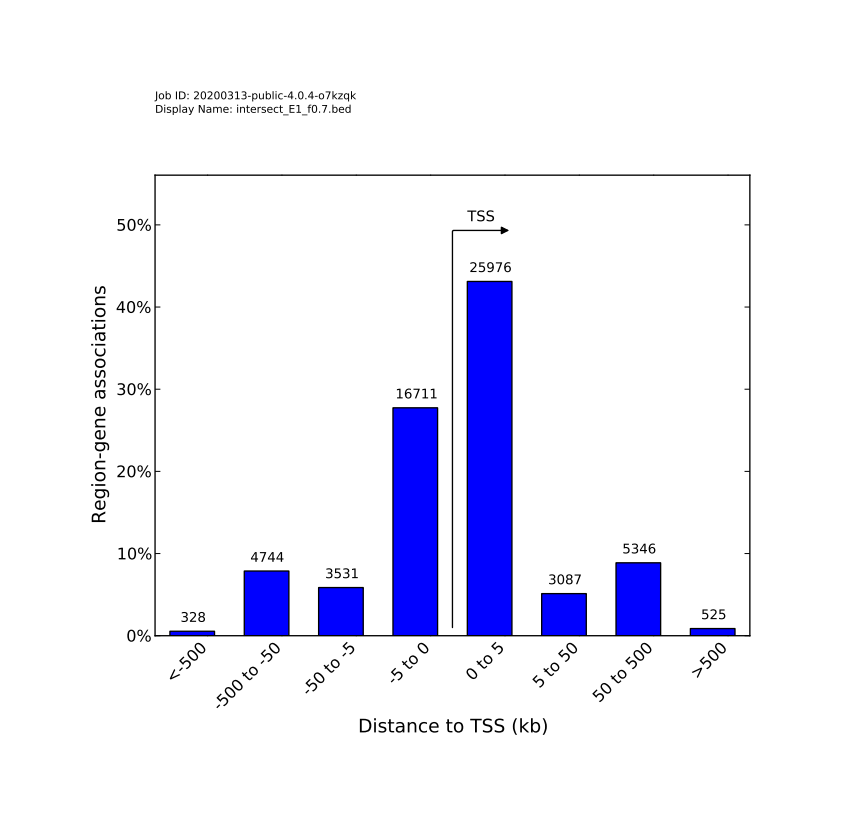
Por ejemplo, en las imágenes anteriores observamos que la mayoría de segmentos anotados proceden de la región con una probabilidad posterior de 0.95 o superior. Esto apoya la seguridad de las anotaciones, puesto que los segmentos que más se asocian con el E1 son aquellos que más relacionados se encuentran 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, 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):

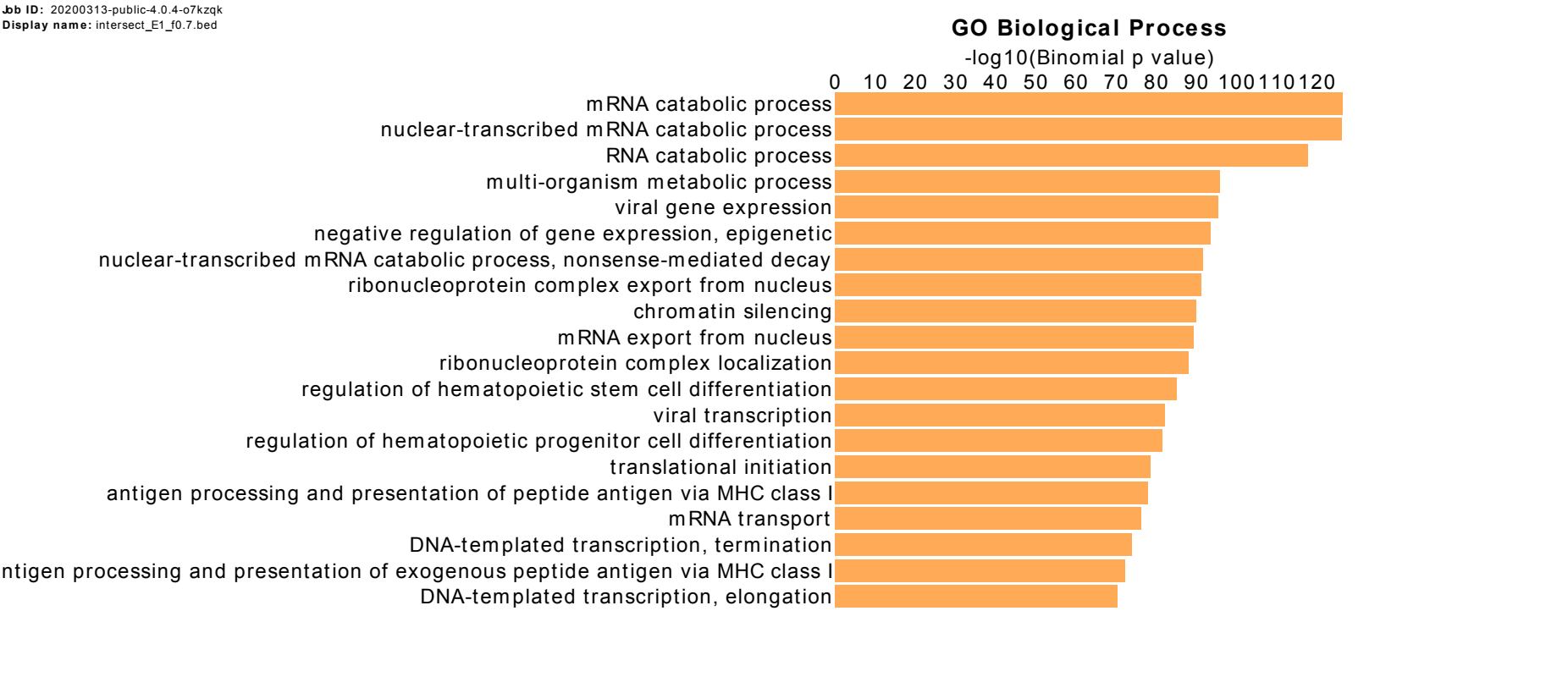
  
Figura 7: Diagrama de barras que muestra el número de segmentos de E1 asociados a genes

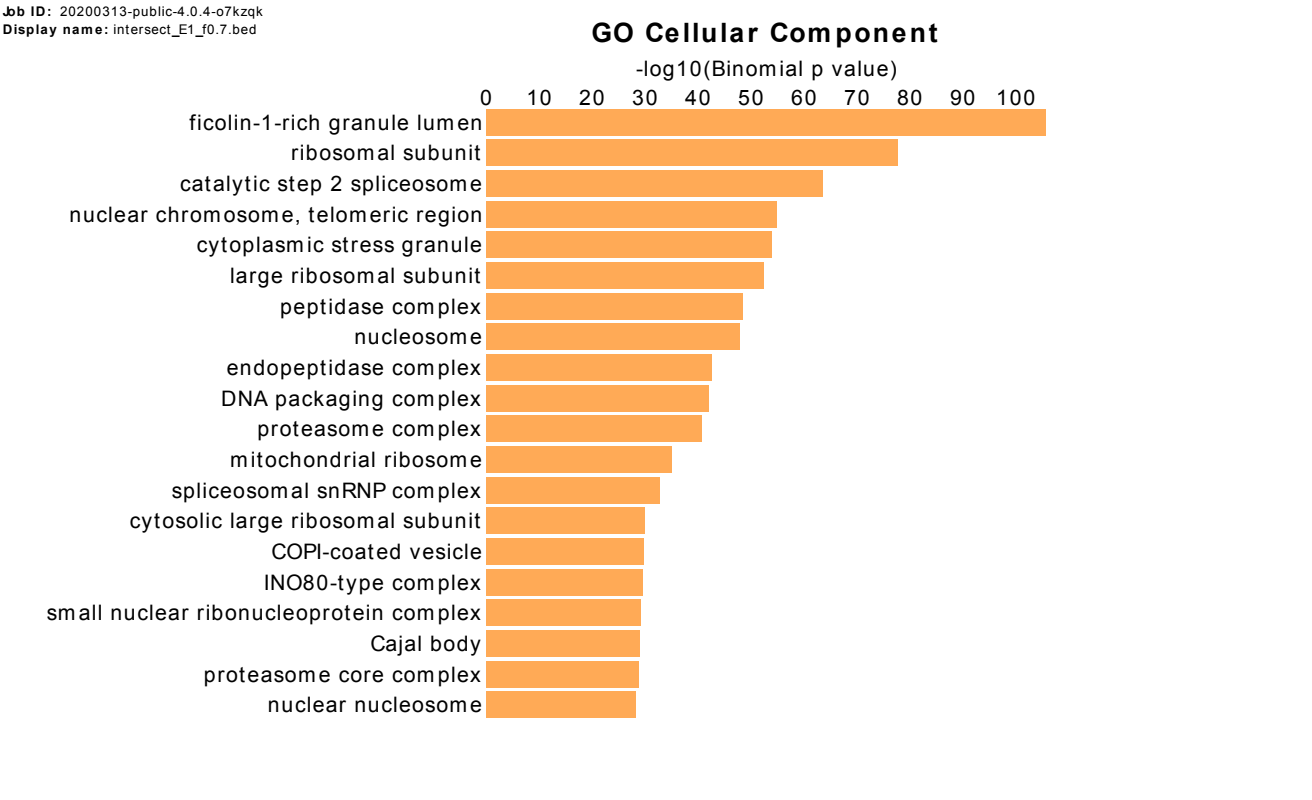
Como se observa en la Figura 3, tan solo 50 (0.1%) de los 44008 segmentos del E1 no se encuentran relacionados con ningún gen. Con el paquete “annotatr” se estudia con mayor detalle con qué tipo de elemento se asocian los 43958 segmentos restantes, que solapan con 10931 genes de los 18,549 genes presentes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

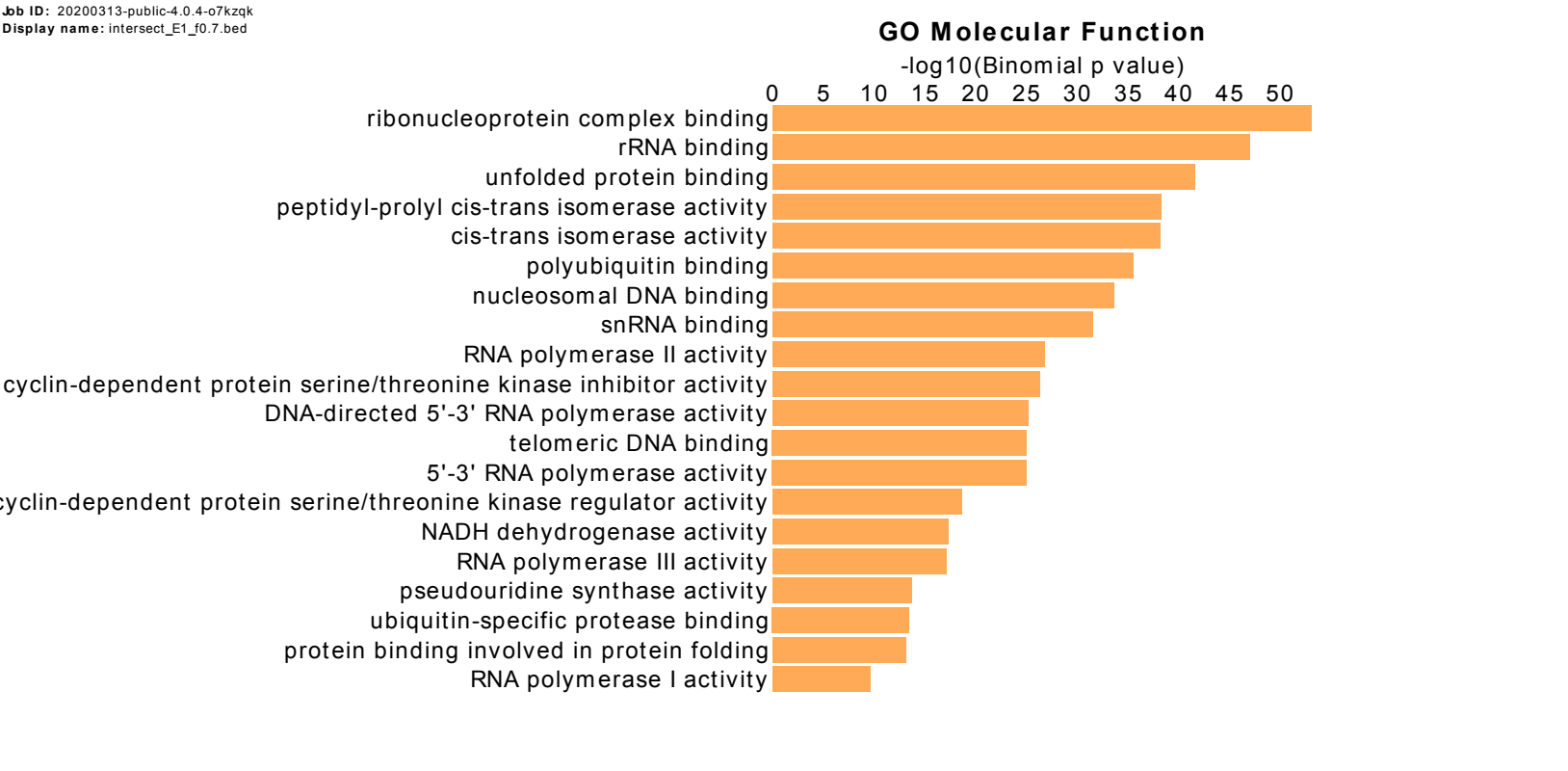
  
Figura 8: Distancia en kilobases de los segmentos de E1 anotados con respecto al sitio de inicio de la transcripción (Transcription Start Site, TSS).

En la figura anterior (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.

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 9: Top 20 anotaciones GO de procesos biológicos relacionados con los segmentos de E1

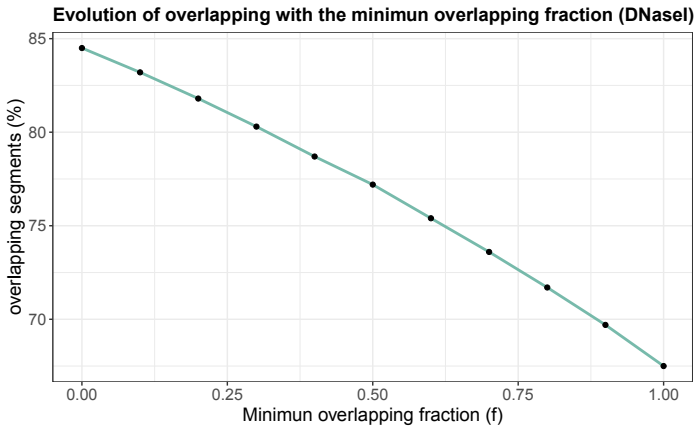
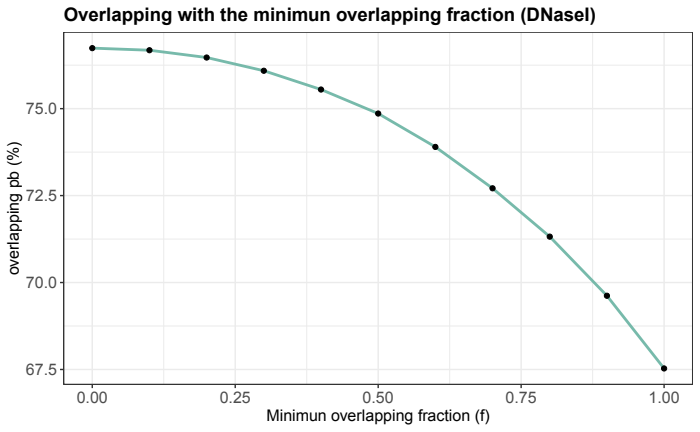
  
Figura 10: Top 20 términos GO de elementos celulares relacionados con los segmentos de E1

  
Figura 11: Top 20 términos GO de función molecular relacionados con los segmentos E1.

### 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.

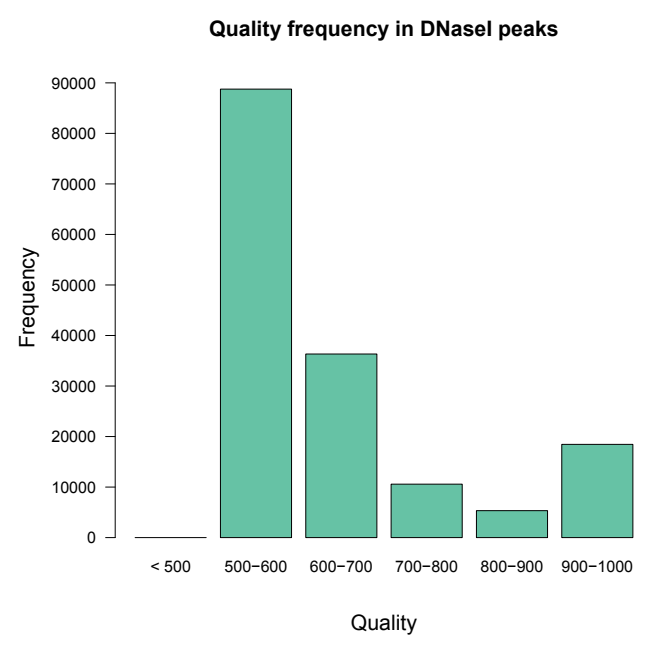
En primer lugar, es importante indicar que la DNaseI es una endonucleasa capaz de romper el enlace fosfodiéster entre dos nucléotidos, tanto en DNA monocatenario como bicatenario en las regiones accesibles de la cromatina. Esta enzima se emplea en la técnica DNase I-seq para identificar regiones hipersensibles 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.

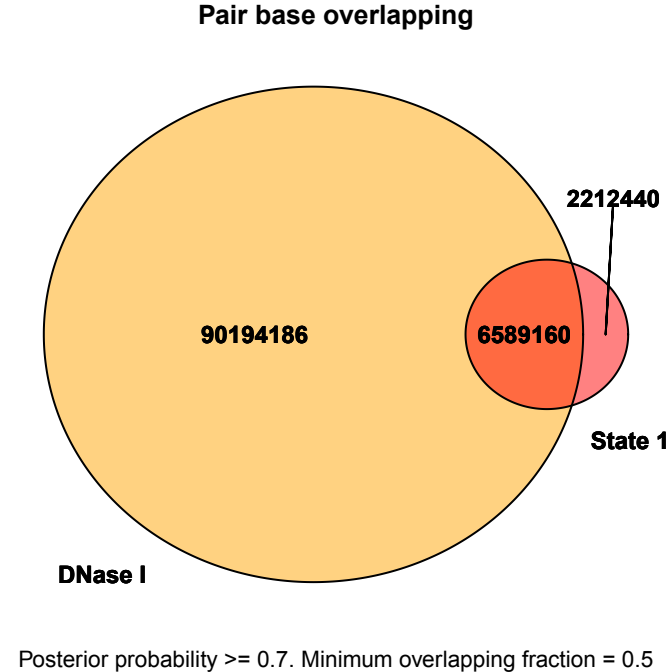


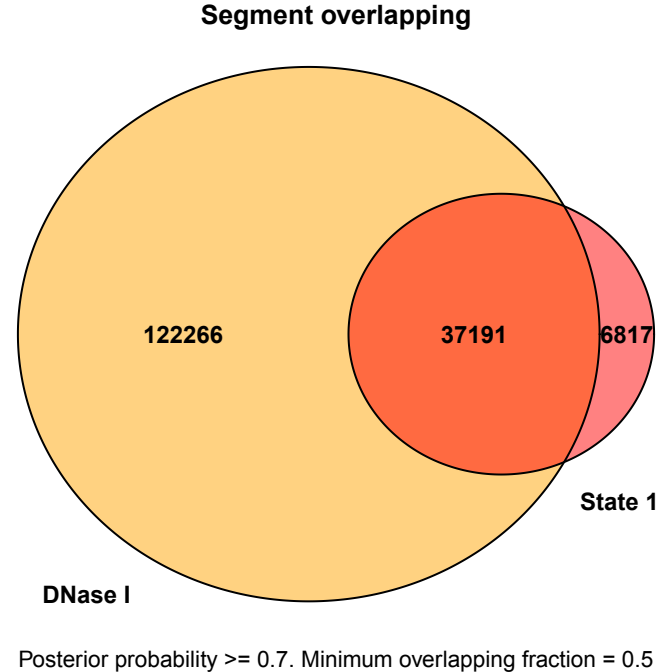
En las dos imágenes anteriores se comprueba que el número de regiones solapantes disminuye a medida que se es más restrictivo con la fracción mínima solapante. Cabe destacar que el cambio del número de segmentos solapantes es lineal, mientras que el número de pares de bases solapantes disminuye de forma 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 12 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 12- Histrograma 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%. 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 13- Diagrama de Venn con el número de pares de bases solapantes entre la DNase I y el E1

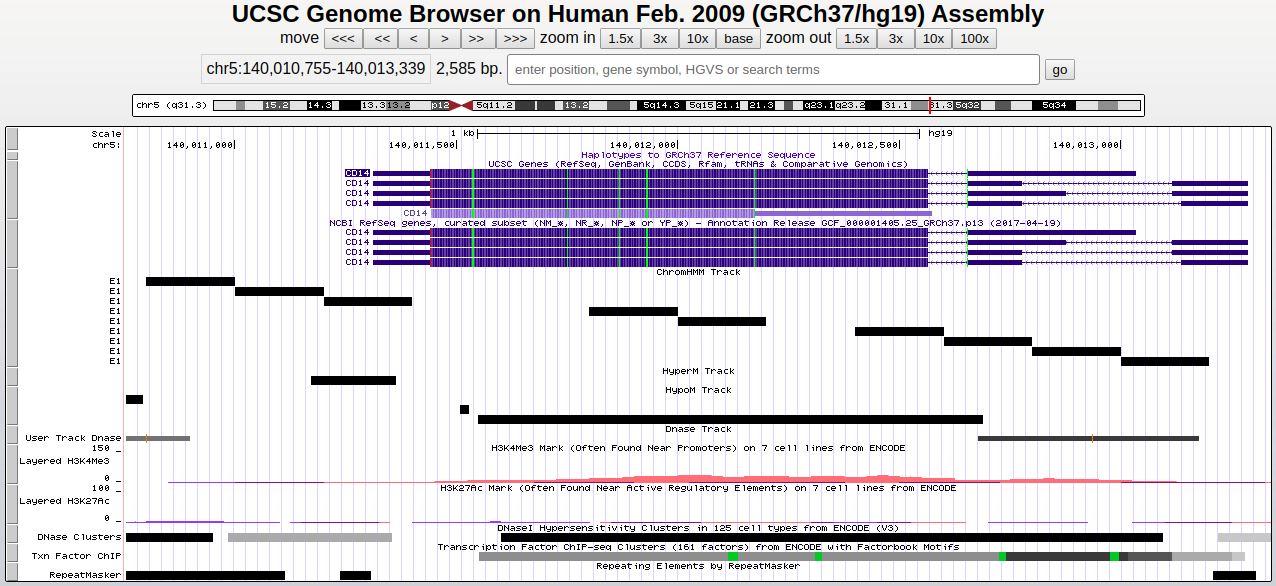
  
Figura 14- Diagrama de Venn con el número de segmentos solapantes entre la DNase I y el E1

### Paso 4) Visualizar una región del genoma en el UCSC browser

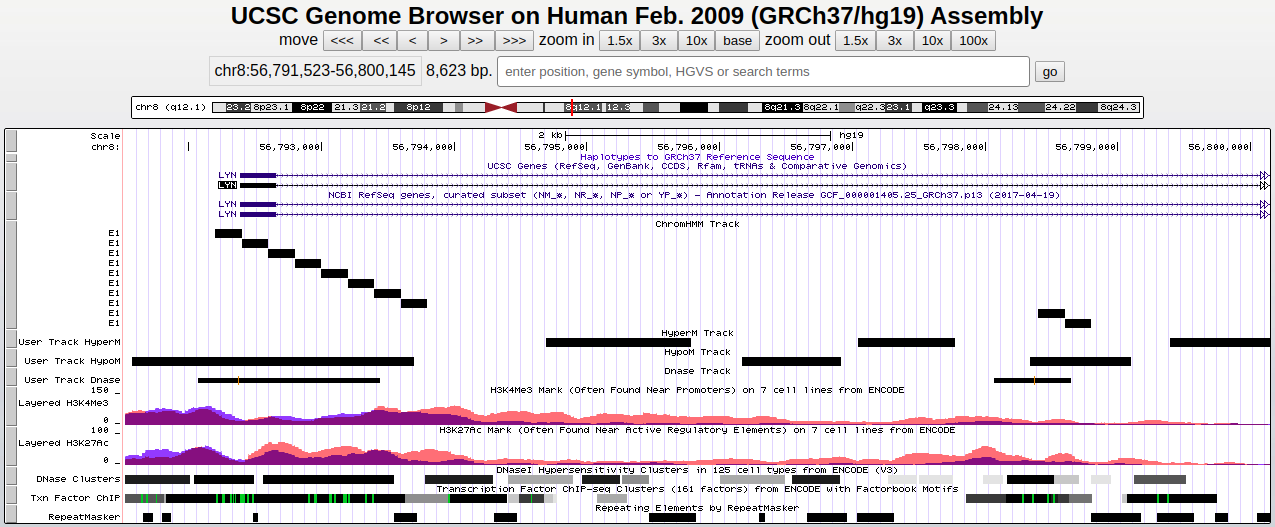
Este paso es muy importante para visualizar y contextualizar los resultados obtenidos. Para ello, en el navegador genómico de la UCSC se suben: el fichero bed con la intersección de segmentos con el E1 como más probable entre ambos monocitos, el fichero de marcas de hiper e hipometilación y el fichero 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/macrógafos participando en el reconocimiento de oligosacáridos procedentes de patógenos. En la Figura 15 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 15- 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 16 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 16 - 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 puromicina y la síntesis de poly(U) a bajas concentraciones de magnesio celular. En la Figura 17 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.

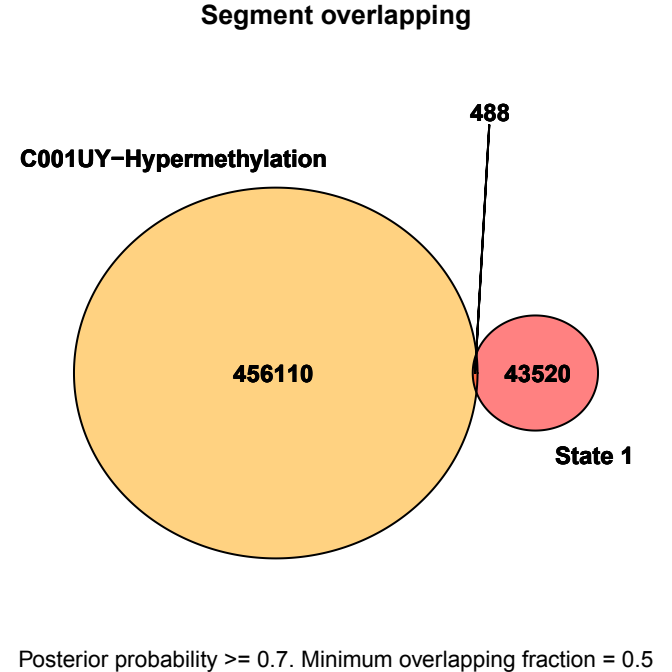
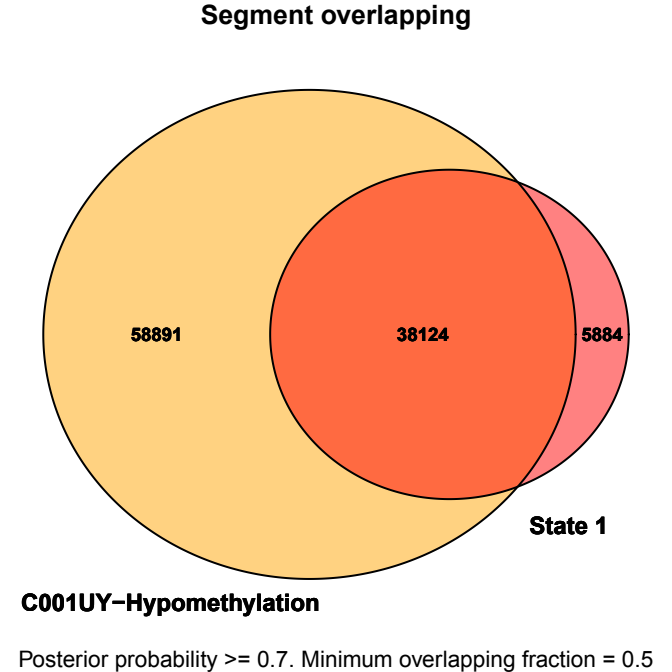
### EIF2A

Figura 17 - 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.

### Regiones hiper e hipometiladas

En las dos imágenes a continuación, 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.



Bibliografía

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

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.

Alberts, B., 2016. *Introducción A La Biología Celular*. Buenos Aires [etc.]: Médica Panamericana.

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.

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

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.