

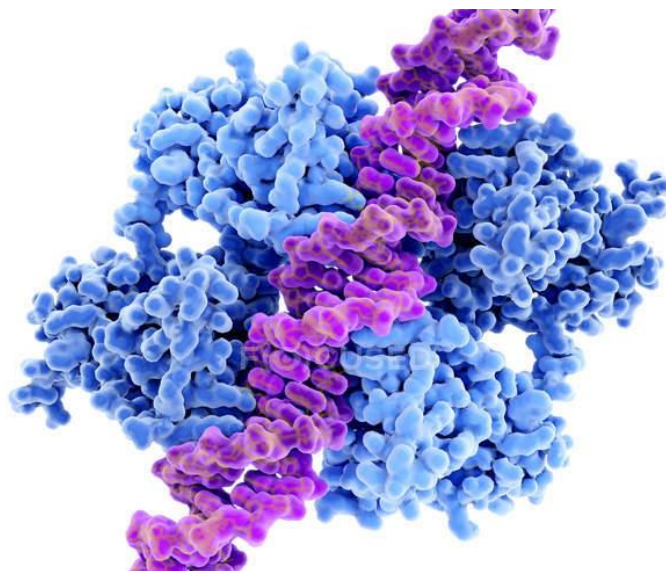


TRABAJO FIN DE GRADO

Grado en Biotecnología

Facultad de Biología – Universidad de Murcia

Generación de modelos para el análisis de experimentos ChIP-Seq sobre las proteínas BCLAF1 y CTCF



Autor: Sergio Vela Moreno

Tutor: Jesualdo Tomás Fernández Breis

Curso 2020-2021



D./Dña.Sergio Vela Moreno....., con DNI nº:..... 77598208A.....,
estudiante del Gradode Biotecnología..... de la Facultad de Biología de la
Universidad de Murcia, **DECLARO:**

Que el Trabajo de Fin de Grado que presento para su exposición y defensa
titulado

... Generación de modelos para el análisis de experimentos ChIP-Seq sobre las
proteínas BCLAF1 y CTCF...

y cuyo/s tutor/es son

D./Dña.Jesualdo Tomás Fernández Breis.....

D./Dña.

.....

**es original y que todas las fuentes utilizadas para su realización han
sido debidamente citadas en el mismo.**

Murcia, a ...13....deJunio.... de 2021.

Firma



*Declaración de originalidad del Trabajo Fin de
Grado*

Listado de abreviaturas:

BGN: Background Gene

BGM: Background Genome

ChIP-Seq: Chromatine Immunoprecipitation followed by Sequencing

ChSqM: ChIP-Seq Model

FM: Functional Model

GMM: Genome Model

GNM: Gene Model

HBGM: HumanBackground Genome

HBGN: Human Background Gene

HGMM: Human Genome Model

HGNM: Human Gene Model

HPM: Human Pathway Model

NGS: Next Generation Sequencing

TES: Transcription End Site

TSS: Transcription Start Site

WGS: Whole Genome Sequencing

ChSqM: ChIP-Seq Model

Índice

1. Resumen.....	
1.1. Resumen en castellano	
1.2. Abstract	
1. Introducción:	1
1.1. Descripción de un experimento ChIP-Seq:.....	2
2. Objetivo:.....	4
2.1. Objetivos específicos:.....	4
3. Materiales y métodos:	5
3.1. Información de las proteínas:	6
3.2. Información de las líneas celulares:	7
3.3. Modelos de conocimiento y probabilísticos:	8
3.3.1. Modelos de conocimiento:	8
3.3.2. Modelos probabilísticos:	13
3.4. Marco de análisis:.....	17
3.4.1. Nivel de región:	17
3.4.2. Nivel de gen:.....	18
3.4.3. Nivel de ruta:	19
4. Resultados:	20
4.1. BCLAF1:.....	20
4.1.1. Nivel de región:	20
4.1.2. Nivel de gen:.....	24
4.1.3. Nivel de ruta:	26
4.2. CTCF:.....	28
4.2.1. Nivel de región:	28
4.2.2. Nivel de gen:.....	32
4.2.3. Nivel de ruta:	34
5. Discusión de resultados:.....	36
5.1. BCLAF1:.....	36
5.2. CTCF:.....	37
6. Conclusiones:	39
6.1. BCLAF1:.....	39
6.2. CTCF:.....	39
7. Bibliografía:	40

1. Resumen

1.1. Resumen en castellano

El estudio sobre la regulación de la transcripción es complejo y está en continua evolución, lo que con el paso del tiempo ha llevado al desarrollo de nuevas técnicas experimentales que permiten obtener distintos datos a partir de muestras biológicas. Actualmente las tecnologías más utilizadas para el estudio de proteínas específicas a nivel del reguloma son las tecnologías de secuenciación de próxima generación (NGS por sus siglas en inglés). Esto es debido a su alta sensibilidad, rendimiento y a su bajo coste económico y de tiempo. Estas tecnologías son capaces de generar una gran cantidad de datos en un solo experimento, por lo que ha surgido la necesidad de desarrollar herramientas de software capaces de analizar los amplios volúmenes de datos que proporcionan. En este estudio se emplean datos obtenidos mediante una técnica NGS, la Inmunoprecipitación de cromatina seguida de secuenciación (ChIP-Seq), que ha permitido caracterizar proteínas de unión al DNA usando *Homo sapiens* como el organismo modelo. Se ha llevado a cabo el análisis del comportamiento de dos proteínas distintas, cada una en una línea celular específica: BCLAF1, en la línea celular k-562, y CTCF, en la línea celular astrocito obtenida a partir de distintas secciones de tejido. Para analizar dichos datos se han generado distintos modelos, tanto de conocimiento como probabilísticos, que permiten analizar el comportamiento de la proteína a tres niveles: nivel de región, de gen y funcional. Gracias a estos modelos ha sido posible encontrar patrones específicos de unión para cada proteína en sus respectivas líneas celulares, así como determinar qué rutas metabólicas se encuentran más afectadas por las proteínas estudiadas en cada una de las líneas celulares.

1.2. Abstract

The study on the regulation of transcription is complex and in continuous evolution, which over time has led to the development of new experimental techniques, allowing obtainance of different data from biological samples. Currently the most widely used technologies for the study of specific proteins at the regulome level are next generation sequencing techniques (NGS). This is due to its high sensitivity, performance and its low economic and time cost. These technologies are capable of generating a large amount of data in a single experiment, which involves the need to develop software tools capable of analyzing such large volumes of data provided. This study uses data obtained by an NGS technique, chromatin immunoprecipitation followed by sequencing (ChIP-Seq), which has made it possible to characterize DNA-binding proteins using *Homo sapiens* as the model organism. Analysis of the behavior of two different proteins has been carried out, each one in a specific cell line: BCLAF1, in the k-562 cell line, and CTCF, in the astrocyte cell line obtained from different tissue sections. To analyze these data, different models have been generated, both knowledge and probabilistic, which allow analyzing the behavior of the protein at three levels: region, gene and functional level. Thanks to these models, it has been possible to find specific binding patterns for each protein in their respective cell lines, as well as to determine which metabolic pathways are most affected by the proteins studied in each of the cell line.

1. Introducción:

Conocer las interacciones proteína-DNA a lo largo del genoma y los diferentes mecanismos epigenéticos es necesario para poder comprender las vías de regulación de los genes que llevan a cabo los distintos procesos biológicos de un organismo^{1,2}. La regulación de la transcripción es compleja y está influenciada por diversos factores como la modificación del DNA y el estado de empaquetamiento de la cromatina, los cuales son procesos dinámicos, al contrario de lo que se creía en un principio^{1,3}, por lo que surge la necesidad de nuevos avances tecnológicos en el campo de la genética y de la genómica, tanto estructural como funcional, para caracterizar genomas y epigenomas en distintos tipos de células y condiciones, los cuales pueden ayudar a mejorar la comprensión de procesos de desarrollo celular y de diversas enfermedades^{1,4}.

Un conjunto de estas tecnologías empleadas para la obtención de datos a partir de muestras biológicas son las tecnologías de secuenciación de moléculas de DNA o de cDNA, obtenidas a partir de la transcripción reversa de moléculas de RNA^{5,6}. Este tipo de tecnología se ha desarrollado mucho con el paso del tiempo, y se clasifica en distintas generaciones:

- **1ª generación:** esta generación incluye el método Sanger (1977), el cual permite secuenciar un fragmento de DNA de unos 800 nucleótidos en cada experimento. Este método sigue siendo muy usado hoy en día, y fue el utilizado en 1979 para secuenciar el primer genoma completo en la historia^{5,7}.
- **2ª generación:** esta generación también se conoce como “Secuenciación de próxima generación” (NGS), y permite secuenciar simultáneamente millones de fragmentos pequeños de DNA, que tienen una longitud entre 50 y 300 nucleótidos^{5,7}.
- **3ª generación:** esta tecnología todavía se encuentra en desarrollo, pero intenta secuenciar un único fragmento de DNA con una longitud entre 1 y 100 kb por experimento^{5,7}.

Las tecnologías NGS son actualmente las más utilizadas, ya que han sido capaces de aumentar considerablemente su sensibilidad y rendimiento a la vez que han reducido su coste económico y de tiempo. Estas tecnologías son capaces de generar una gran cantidad de datos en un solo experimento, consecuentemente, no sólo ha sido necesario un gran avance a nivel de laboratorio sino también a nivel informático, desarrollando herramientas de software capaces de analizar dichos datos^{5,6}.

La herramienta principal que se utiliza para experimentos de 2ª generación es la inmunoprecipitación de cromatina (ChIP), una técnica empleada para identificar las regiones de DNA por las que una proteína específica presenta afinidad para unirse^{1,8}.

Gracias a la hibridación con micromatrices, los fragmentos de DNA que resultan de la inmunoprecipitación de cromatina pueden ser identificados. Esta técnica se conoce como ChIP-chip, y proporciona una visión a escala del genoma de las interacciones proteína-DNA^{1,9,10}.

Los grandes avances tecnológicos en secuenciación de próxima generación (NGS) permiten secuenciar decenas o cientos de millones de fragmentos cortos de DNA en un solo experimento. Esto ha permitido su aplicación en distintas áreas, siendo una de las más importantes la secuenciación a nivel de genoma completo (WGS por sus siglas en inglés)^{1,11,12}.

Otra de las principales aplicaciones de NGS es la técnica de inmunoprecipitación de cromatina seguida de secuenciación (ChIP-Seq), la cual permite que los fragmentos de interés de DNA se secuencien directamente, sin ser necesaria su hibridación a una micromatriz. En comparación con la técnica ChIP-chip, ChIP-Seq proporciona datos significativamente más precisos, requiere de menos instrumental y permite la secuenciación de fragmentos de DNA más pequeños (hasta aproximadamente 35 kb)^{1,13}.

1.1. Descripción de un experimento ChIP-Seq:

La secuenciación de Inmunoprecipitación de Cromatina es utilizada para la identificación de regiones de DNA a las que se une una determinada proteína. Esta técnica se compone de dos fases: una primera fase en la que se lleva a cabo la inmunoprecipitación de los fragmentos de cromatina (ChIP), y una segunda fase en la que se lleva a cabo la secuenciación de los fragmentos obtenidos (Seq). Gracias a ello, esta clase de experimento proporciona una representación a escala genómica de las interacciones entre el DNA y una proteína en concreto^{5,14}, para un tejido o una línea celular específicos. Sus etapas principales (Figura 1) son: i. Administración de la célula, ii. Entrecruzamiento, iii. Lisis celular, iv. Corte de la cromatina, v. Inmunoprecipitación de la cromatina, vi. Entrecruzamiento reverso, vii. Preparación de la librería, viii. Secuenciación, ix. Control de calidad, x. Mapeo, xi. Control de calidad, xii. Definición de picos, xiii. Análisis funcional, xiv. Identificación de motivos⁵.

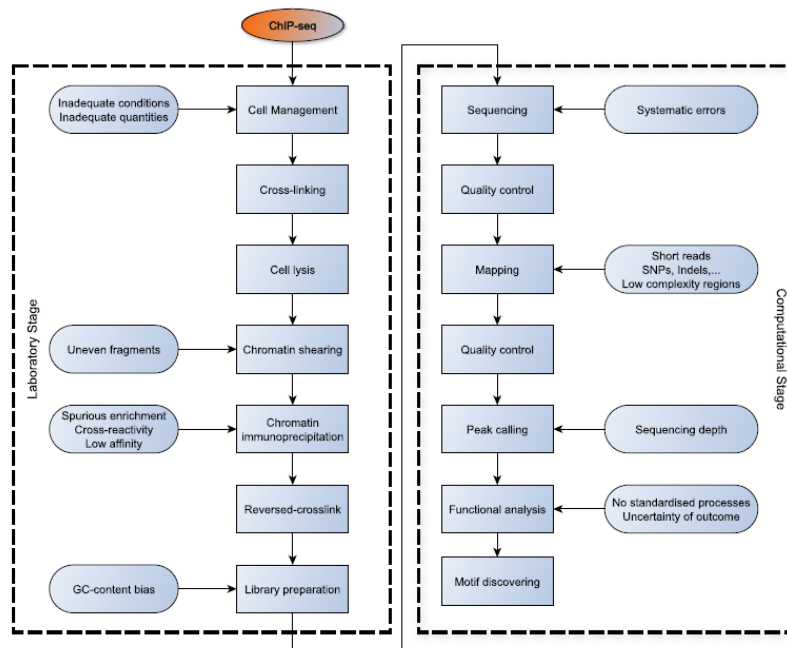


Figura 1. Etapas generales de un experimento ChIP-Seq. En esta figura se representan las distintas fases de un experimento ChIP-Seq (formas rectangulares) y los problemas principales que derivan de las mismas (formas ovaladas)⁵

Las tecnologías NGS generan gran cantidad de datos por experimento. Consecuentemente, el avance a nivel experimental está acompañado por avances a nivel informático, que incluyen herramientas de software para el análisis de dichos datos generados^{1,4,5}. Un ejemplo de estas herramientas son los modelos de conocimiento y probabilísticos que se describen y generan⁵, y que se usarán en dos contextos distintos para realizar un análisis a nivel del reguloma empleando *Homo sapiens* como organismo modelo: por un lado se analizan tres estudios que emplean la proteína BCLAF1 en la línea celular k-562, y por otro lado se analizan tres experimentos que usan la proteína CTCF en la línea celular astrocito. Indicar que se ha llevado a cabo un proceso de selección de proteínas usando la base de datos ReMap2020¹⁵, que es una base de datos que integra datos de reguladores transcripcionales procedentes de experimentos de unión proteína-DNA en dos organismos: *Homo sapiens* y *Arabidopsis thaliana*. Las proteínas que se han seleccionado son BCLAF1, ya que tiene múltiples experimentos en una misma línea celular (k-562), y CTCF, ya que presenta experimentos en una misma línea celular (Astrocito) obtenida de 3 secciones de tejido distintas: cerebro, cerebelo y médula espinal.

2. Objetivo:

El objetivo general de este trabajo es analizar experimentos ChIP-Seq asociados a dos proteínas y líneas celulares en concreto: BCLAF1 en la línea celular k-562 y CTCF en la línea celular astrocito, para lo que se generaran distintos modelos de conocimiento y probabilísticos.

2.1. Objetivos específicos:

A continuación se enuncian los objetivos específicos para cada una de las dos proteínas seleccionadas que se emplean en este estudio.

Para la proteína BCLAF1, se analizan experimentos en la línea celular k-562, con el fin determinar qué regiones, genes y rutas se encuentran más afectadas por la acción de la proteína a partir de los modelos que se generan y encontrar patrones de unión específicos en dicha línea celular. También se busca analizar cuál de ellos resulta más próximo al comportamiento real de la proteína.

Para la proteína CTCF, se analizan experimentos en la línea celular astrocito obtenida de distintas secciones de tejido, y se pretende determinar qué regiones, genes y rutas se encuentran más afectadas por la acción de la proteína a partir de los modelos que se generan y encontrar patrones de unión específicos en dicha línea celular. También se busca analizar cuál de los modelos resulta más próximo al comportamiento real de la proteína y cómo varía el comportamiento de la misma de unión según la sección de tejido en la que se realiza el experimento.

3. Materiales y métodos:

En este apartado se describen los pasos llevados a cabo para el análisis de estos experimentos (Figura 2):

- i) Obtención de datos necesarios a partir de diversos repositorios
- ii) Generación de los modelos de conocimiento a partir de dichos datos
- iii) Obtención de los modelos probabilísticos (a partir de los modelos de conocimiento)
- iv) Análisis a distintos niveles
- v) Obtención del perfil estandarizado de la proteína.

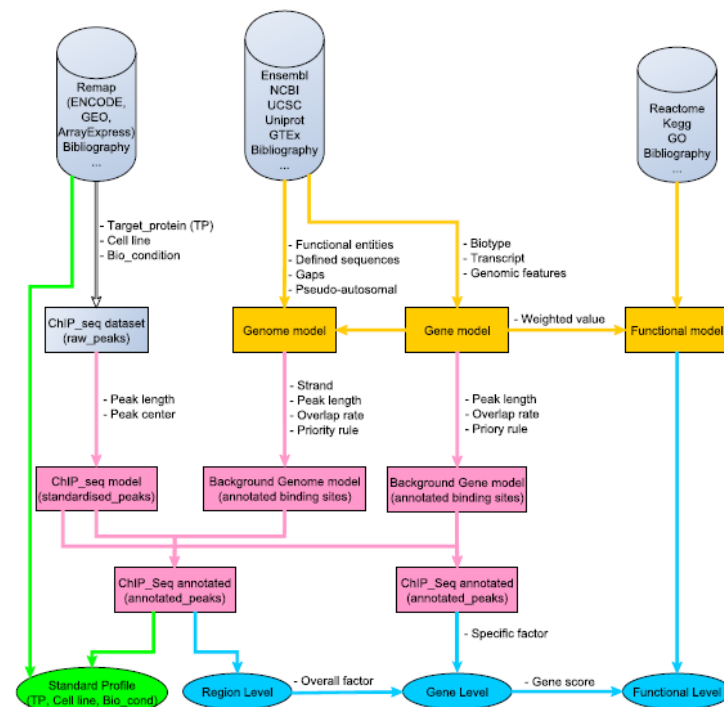


Figura 2. Vista general del procedimiento de la metodología seguida para el análisis a distintos niveles realizado en este estudio⁵. En gris aparecen indicados distintos repositorios bioinformáticos, en amarillo los modelos de conocimiento, en rosa los modelos probabilísticos, en azul los niveles de los que se compone el marco de análisis que se sigue, y en verde el perfil estandarizado, específico de proteína, línea celular y condición biológica.

3.1. Información de las proteínas:

A continuación se incluye información sobre ambas proteínas empleadas en este estudio, BCLAF1 (Tabla1) y CTCF (Tabla 2).

- **BCALF1:**

Gen	BCLAF1 ¹⁶
Descripción	Factor de transcripción 1 asociado a BCL2 ¹⁶
Localización	Cromosoma 6: 136,256,627-136,289,851 cadena rezagada ¹⁶
Número de transcritos	24 ¹⁶
ID del transcrito canónico	ENS00000531224.6 ¹⁶
Código en UniProt	Q9NYF8 ¹⁷
Función	Esta proteína se trata de un represor transcripcional que interacciona con proteínas de la familia BCL2 y cuya sobreexpresión provoca apoptosis ¹⁸

Tabla 1. Características generales de la proteína BCLAF1. En esta tabla se incluyen las principales características de la proteína BCLAF1, obtenidas de distintas bases de datos, tales como su gen, localización y función.

- **CTCF:**

Gen	CTCF ¹⁶
Descripción	Factor de unión a CCCTC ¹⁶
Localización	Cromosoma 16: 67,562,467-67,639,177 cadena adelantada ¹⁶
Número de transcritos	17 ¹⁶
ID del transcrito canónico	ENS00000264010.10 ¹⁶
Código en UniProt	P49711 ¹⁷
Función	Esta proteína, que presenta motivos de dedo de zinc (ZF), interviene en la regulación transcripcional al unirse a insulators y evitar la interacción de enhancers y silenciadores con la región promotora. Entre sus funciones destaca la de actuar como supresora de tumores y participar en la regulación de mecanismos epigenéticos ¹⁸

Tabla 2. Características generales de la proteína CTCF. En esta tabla se incluyen las principales características de la proteína CTCF, obtenidas de distintas bases de datos, tales como su gen, localización y función

3.2. Información de las líneas celulares:

A continuación se incluye información sobre ambas líneas celulares empleadas en este estudio (k-562 en el caso de la proteína BCLAF1, y astrocito en el caso de la proteína CTCF).

- **K-562:**

Línea celular obtenida en 1970 a partir del derrame pleural de una mujer de 53 años que padecía leucemia mieloide crónica (CML) en fase blástica (proporción de células sanguíneas inmaduras (blastocitos) mayor al 30% de las células sanguíneas o de médula ósea)¹⁵.

- **Astrocito:**

Línea celular con forma estrellada que se clasifica como un tipo de células grandes de la neuroglia (macroglia) en el sistema nervioso central. Se trata de

las células de la neuroglia más grandes y abundantes en el cerebro y en la espina dorsal¹⁵.

3.3. Modelos de conocimiento y probabilísticos:

En este apartado se describirán los dos tipos de modelos, de conocimiento y probabilísticos, que se han generado para el estudio de las proteínas anteriormente mencionadas.

3.3.1. Modelos de conocimiento:

Los modelos de conocimiento son una representación de las distintas entidades genómicas y la relación entre ellas, la cual no solo muestra características generales del genoma del organismo estudiado, sino también características específicas de cada línea celular o tejido del organismo. Estas características se agrupan en:

- Estructurales y organizativas (*Genome Model*),
- Unidades de información (*Gene Model*)
- Funciones y procesos biológicos en el organismo (*Functional Model*)⁵.

3.3.1.1. Human Genome Model (HGMM):

El *Genome Model* (GMM) representa el conocimiento de interés sobre el genoma del organismo en estudio. Este modelo es el primero que se genera, y alberga las regiones del genoma con secuencia de nucleótidos conocida, permitiendo el mapeado de las lecturas del experimento en ellas. Puede ser específico de la línea celular o tejido, usándose para la proteína BCLAF1 la línea celular k-562 y para la proteína CTCF la línea celular astrocito⁵.

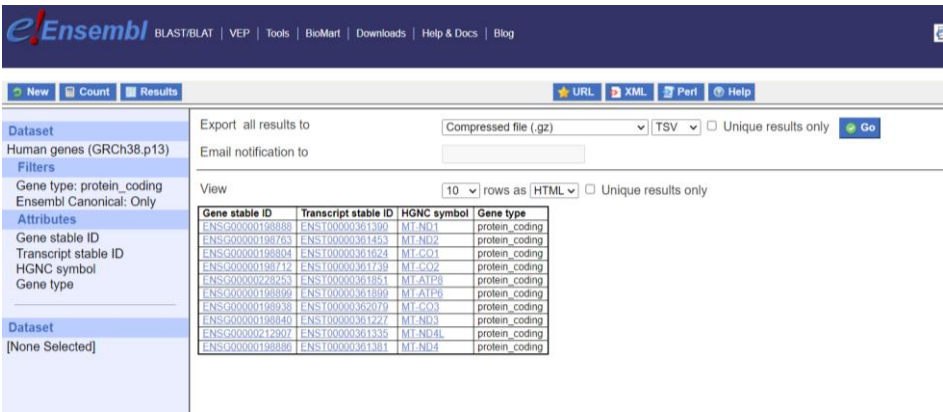
En este trabajo se analizan conjuntos de datos de experimentos ChIP-Seq que tienen como base el genoma humano. Para llevar a cabo la generación del *Human Genome Model* (HGMM), se tomó en un principio el Genome Reference Consortium Human Build 38 patch release 13 (GRCh38.p13) como base sobre la que poder ubicar mediante un sistema de coordenadas cromosómicas las distintas entidades funcionales. El modelo generado no incluye ni los huecos, regiones sin secuencia conocida, ni las regiones pseudo-autosomales del cromosoma Y, que no pueden ser analizadas mediante los algoritmos del programa ya que sólo son aplicables al cromosoma X. Además se trata de una representación haploide del genoma humano

ya que los algoritmos empleados para obtener los conjuntos de datos de los experimentos ChIP-Seq estudiados no pueden discernir si un pico corresponde a uno u otro cromosoma homólogo o a ambos⁵.

La otra finalidad de este modelo es delimitar la posición que ocupa cada entidad funcional en el genoma. Una entidad funcional es toda aquella entidad genómica funcionalmente autosuficiente, y son responsables de la dimensionalidad del modelo. Las entidades genómicas pueden ser regiones con al menos un producto funcional, como en el caso de un transcrito codificante de proteína, o puede ser la propia secuencia de la entidad la que realice una función específica (secuencia funcional), como en el caso de los *enhancers*. Las entidades funcionales están constituidas por características genómicas, las cuales son regiones genómicas que realizan una función concreta sobre la entidad funcional a la que pertenecen. Si una región del GMM no incluye ninguna entidad funcional, se le atribuye la clase intergénica, que indica que no tiene función o que no se conoce todavía⁵.

Las entidades funcionales que se han incluido en la generación de los modelos para las dos líneas celulares de este estudio han sido los genes codificantes de proteínas y *enhancers* específicos de línea celular.

Los datos para el primer tipo de entidad incluidos han sido obtenidos mediante el uso de la herramienta *Biomart* de *Ensembl*¹⁶, seleccionando como parámetros de búsqueda los genes que codifican proteínas y usando los datos correspondientes al transcrito que la propia base de datos considera como el transcrito canónico. Los datos que se obtuvieron fueron el *Gene_id*, *Transcript_id*, *Hgnc symbol* y *Gene type* (*Biotipo*) (Figura 3).



Gene stable ID	Transcript stable ID	HGNC symbol	Gene type
ENSG00000198888	ENST00000361290	MT-ND1	protein_coding
ENSG00000198763	ENST00000361453	MT-ND2	protein_coding
ENSG00000198804	ENST00000361624	MT-CO1	protein_coding
ENSG00000198712	ENST00000361739	MT-CO2	protein_coding
ENSG00000228243	ENST00000361951	MT-ATP8	protein_coding
ENSG00000198899	ENST00000361999	MT-ATP9	protein_coding
ENSG00000198938	ENST00000362079	MT-CO3	protein_coding
ENSG00000198840	ENST00000361227	MT-ND3	protein_coding
ENSG00000212907	ENST00000361335	MT-ND4L	protein_coding
ENSG00000198896	ENST00000361381	MT-ND4	protein_coding

Figura 3. Herramienta *Biomart* de la base de datos *Ensembl*¹⁶, empleada para la obtención de los genes codificantes de proteínas incluidos en la generación del *Human Genome Model*. En los parámetros de búsqueda se han seleccionado los genes que codifican proteínas y los datos correspondientes al transcrito que la propia base de datos considera como el transcrito canónico. Los datos que se obtienen son *Gene_id*, *Transcript_id*, *Hgnc symbol* y *Gene type* (*Biotipo*).

Para la línea celular K-562 el espacio de muestra considerado es el formado por los cromosomas {1..22,X} con 19195 genes, mientras que, para la línea celular astrocito el espacio de muestra considerado es el formado por los cromosomas {1..22,X,Y} con 19241 genes. Las diferencias en el espacio de muestra se deben al género de la fuente de cada línea celular.

Los ficheros que contienen *enhancers* han sido obtenidos de la base de datos *EnhancerAtlas 2.0*¹⁹, y son específicos de para cada una de las dos líneas celulares. Para la línea celular k-562 se ha incluido un total de 43753 *enhancers*, y para la línea celular astrocito se incluyen 44489 *enhancers*. Una vez obtenidos ha sido necesario utilizar la herramienta *Assembly Converter* de *Ensembl*¹⁶ para poder convertir dichos datos a la versión 38 del genoma humano de referencia humano, ya que los datos que proporciona esta base se encuentran mapeados en *Genome Reference Consortium Human Build 37 (GRCh37)*. Como resultado se obtiene un archivo en formato *BED* que contiene las coordenadas de dichos *enhancers* en el genoma.

Finalmente, como se dispone de los modelos generados que toman como base la versión *Genome Reference Consortium Human Build 38 patch release 12 (GRCh38.p12)*, se comprueba que apenas existe variación entre los modelos que usan un parche u otro como genoma de referencia (al mantenerse en ambos casos la versión 38 significa que no se han añadido o eliminado genes en dicha actualización), por lo que se decide usar los modelos de dicha versión en lugar de los que derivan de la versión actual ya que se han empleado en otros estudios dando buenos resultados⁵.

3.3.1.2. Human Gene Model (HGNM):

Un *Gene Model (GNM)* representa los tipos de entidades funcionales que contienen un producto que realice función. Un GNM consta de:

- 1) **Biotipo del GNM**, que para este modelo es genes codificantes de proteínas⁵
- 2) **Cadena de DNA en la que se localiza el GNM**, que puede ser la cadena adelantada, rezagada o ambas. En este modelo se emplean ambas⁵
- 3) **Transcrito del gen**. En este caso se incluye el transcrito canónico de cada gen de acuerdo a los criterios de la base de datos *Ensembl*¹⁶. Se compone de uno o varios exones y, en caso de que el gen sea codificante de proteína, de una secuencia codificante (CDS). Además posee un sitio de inicio de la transcripción (TSS) y un sitio de fin de la transcripción (TES)⁵
- 4) **Características genómicas**, que son regiones funcionales que corresponden a una entidad funcional concreta. Una característica genómica posee una clase

específica, un punto de referencia en el GNM que sirve para localizarla en el genoma, y una extensión en el genoma a partir de dicho punto⁵.

En este modelo se definen las características genómicas presentes en el *Human Gene Model* (HGNM), indicando su localización en el genoma y su extensión. Las 7 clases de características genómicas de las que se propone que consta un HGNM son:

- i. Promotor: se trata de la región núcleo que se encuentra justo alrededor del TSS^{5,20,21}
- ii. Regiones proximales: son las regiones reguladoras secundarias, localizadas a los lados de la región promotora⁵
- iii. Regiones distales: esta es la región que suele albergar los sitios de unión con menor afinidad para la proteína estudiada, o bien, regiones represoras, que se encuentran aguas arriba del gen, al lado de la región proximal^{5,22}
- iv. Escisión: es la región de procesamiento del mRNA maduro y rodea al TES^{5,23,24}
- v. Empalme: Representa los sitios dador y aceptor, que corresponden a los motivos reconocidos por la maquinaria encargada de llevar a cabo el proceso de splicing o espliceosoma^{5,25}
- vi. Exón: consiste en todas las regiones que formarán parte del mRNA maduro^{5,26}
- vii. Intrón: se trata de cada una de las regiones transcritas que no forman parte del mRNA maduro^{5,26}.

5) **Entidades funcionales**, que son secuencias funcionales como enhancer o insulador⁵. El conjunto de entidades funcionales junto a sus características genómicas, punto de referencia e intervalos se muestra en la tabla 3⁵.

Functional Entities	Genomic Features	Ref. Position	Start (bp)	End (bp)
Protein - Coding Gene	distal	Transcription Start Site	-4500	-2001
	proximal	Transcription Start Site	-2000	-301
			+151	+1000
	promoter	Transcription Start Site	-300	+150
	splice	Donor: End of exon	-10	+10
		Acceptor: Start of exon	-60	+10
	cleavage	Transcription End Site	-75	+250
	exon	Start of exon	0	-
		End of exon	-	0
intron (n)		End of exon (n)	+1	-
		Start of exon (n+1)	-	-1
Enhancer	enhancer	-	average 600 nt.	

Tabla 3. Localización y regiones que abarcan las características genómicas pertenecientes a las dos entidades funcionales: genes que codifican proteínas y enhancers⁵. Se toma como base la cadena adelantada. La longitud media de la clase enhancer varía según la línea celular empleada, pero en ambos casos se emplea TSS como el punto de referencia.

Tras comprobar las estadísticas para distintas distancias entre el TSS y los enhancers considerados ligados al gen respectivo para el caso de ambas proteínas, se determina que la mejor distancia a usar en el caso de la línea celular k-562 es ± 100000 bp, que genera una mediana de 11004 sitios de unión por gen del TSS y en el caso de la línea celular astrocito es ± 50000 bp del TSS, con una mediana de 11920 sitios de unión por gen.

3.3.1.3. Functional Model: Modelo de ruta humana (HPM):

Un *Functional Model* (FM) representa como una relación lineal de elementos no repetidos²⁷ todo recurso jerárquico funcional o entidad relacional que asocie una función biológica a entidades funcionales del organismo que se estudia, como por ejemplo una anotación funcional o una ruta metabólica⁵. Una anotación funcional es un código asociado a cualquier proceso biológico, función molecular o componente celular, como los términos GOs²⁸. Una ruta metabólica es una sucesión ordenada de reacciones bioquímicas en las que uno o más compuestos químicos, conocidos como sustratos, pasan de un estado inicial a un estado final, dando lugar a uno o varios productos. Este tipo de entidades se pueden encontrar en repositorios como *Reactome*²⁹.

Un FM que corresponde a una anotación funcional se compone de los GNMs de los genes asociados a dicha anotación funcional. Los elementos de un FM para una ruta metabólica son los GNMs de los genes que pueden tanto intervenir directamente en las reacciones de dicha ruta o participar en su regulación, aunque también pueden ser otras rutas metabólicas que, por ejemplo, afecten a la concentración de sustrato o cofactores implicados, inhibiendo a la ruta⁵.

A cada elemento del FM se le adjudica un número real al que se denomina factor de relevancia, que indica la contribución de dicho elemento en el FM respectivo. Este factor de contribución puede estimarse por distintos procedimientos⁵, y para este estudio tiene un valor de 1 en ambos casos.

En el nivel funcional de este estudio se utilizan como FMs el conjunto de las rutas metabólicas en humanos, que se pueden obtener en la base de datos *Reactome* v.73²⁹, y constituyen el denominado modelo de ruta humana (HPM). A partir de *Reactome* se obtienen dos ficheros de texto, uno que localiza cada ruta en el genoma, y contiene los nombres de los genes que se asocian a ellas, y otro que establece una jerarquía entre HPMs⁵. Siguiendo dicha jerarquía establecida por *Reactome*, las HPMs

se agrupan en tres niveles jerárquicos, que van desde los HPMS más genéricas a las menos. En el primer nivel se incluyen los 26 HPMS más genéricos, mientras que el segundo nivel incluye 143 HPMS y el tercero 1634 HPMS.

3.3.1.4. ChIP-Seq Dataset:

Se trata del conjunto de regiones de DNA que han resultado enriquecidas en un experimento ChIP-Seq, y se obtienen al someter las lecturas de un experimento ChIP-Seq a un algoritmo *peak calling*. Estas regiones son los denominados picos en bruto, y son específicos de proteína, línea celular y de las condiciones biológicas a las que se somete el experimento^{5,15}. Son el conjunto de datos que se quieren analizar.

Un pico en bruto está definido por cromosoma, inicio, fin, cadena de DNA a la que pertenece el pico en bruto (Puede ser la cadena adelantada (+), rezagada (-) o no estar definida (.)), y calidad.

Los picos de los experimentos de ambas proteínas estudiadas en este trabajo se han obtenido de la base de datos *ReMap2020*¹⁵:

- Para la proteína BCLAF1 se han utilizado los picos correspondientes a los experimentos ENCSR000BKH (5859 picos), ENCSR492LTS (6904 picos) y ENCSR792IYC (4292 picos), realizados en la línea celular k-562¹⁵.
- Para la proteína CTCF se han utilizado los picos correspondientes a los experimentos ENCSR000AOO (42499 picos), ENCSR000DSU (58906 picos) y ENCSR000DSZ (53412 picos), en la línea celular astrocito¹⁵.

3.3.2. Modelos probabilísticos:

Los modelos probabilísticos son un cambio en la disposición del conocimiento en los modelos anteriores, necesario para su procesamiento matemático. Dichos modelos se generan a:

- Nivel del genoma completo (*Background Genome*)
- Nivel de unidad de información (*Background Gene*).
- Entre estos modelos también se incluyen los que contienen los datos obtenidos de un experimento ChIP-seq, los cuales también son reorganizados para su procesamiento matemático (*ChIP-seq Model*)⁵.

3.3.2.1 Human Background Genome (HBGM):

El *Background genome* (BGM) es la totalidad de sitios de unión anotados que puede contener un GMM definido según unos parámetros. Dichos sitios de unión anotados son regiones del genoma que presentan la misma probabilidad de que la proteína estudiada se una a ellos, y se caracterizan por el cromosoma en que se encuentran, sus coordenadas de inicio y fin, su cadena y la clase de característica genómica a la que pertenecen, o la clase intergénica si no se conoce su función, de entre todas las que conforman el GMM del que se obtiene dicho BGM⁵.

Para configurar un BGM son necesarios:

- 1) **Tamaño de muestra:** proporción del GMM que forma parte del BGM (puede tratarse del GMM completo como en los casos de este estudio).
- 2) **Longitud de pico:** es la longitud que todos los sitios de unión anotados comparten.
- 3) **Proporción de solapamiento:** porcentaje que determina el menor número de nucleótidos pertenecientes a un sitio de unión anotado que debe superponerse con la región de una característica genómica específica para considerar que ese sitio pertenece a dicha clase.
- 4) **Regla de prioridad:** en caso de que para un sitio de unión anotado se superpongan dos o más características genómicas, determina cuál de las clases se le debe adjudicar. De este modo se logra que cada sitio de unión anotado pertenezca a una única clase.
- 5) **Cadena:** indica en qué cadena se encuentran las entidades funcionales usadas para generar el BGM. Su valor puede ser cadena adelantada, rezagada o ambas. Al evaluar cómo se comporta una proteína en estudio se podría determinar si es específica de cadena.

Los valores para la longitud de pico y la proporción de solapamiento se ajustan según la precisión de los experimentos ChIP-seq estudiados⁵.

En este caso el BGM estudiado es el que se corresponde con la población de sitios de unión anotados de un individuo humano (HBGM). En dicho HBGM se consideran las 8 características genómicas definidas en el *Human Genome Model* y la clase intergénica, que son tratadas como 9 clases independientes entre sí, por lo que se trata de un modelo multinomial de 9 dimensiones (BM:9dm) (Figura 4).

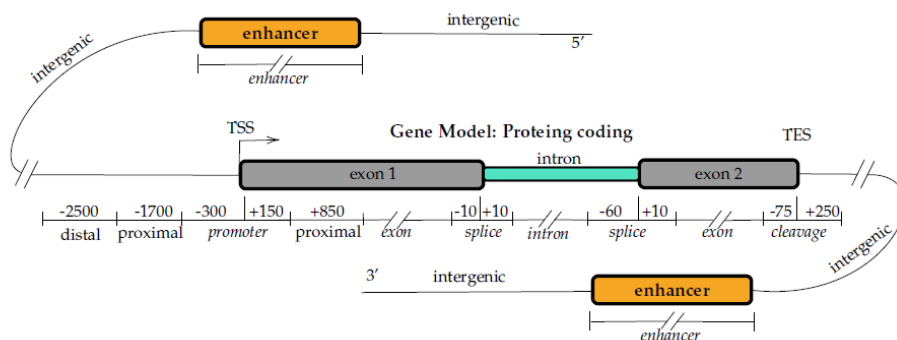


Figura 4. Representación de la colocación y longitud (pares de bases) de las distintas características genómicas en la cadena adelantada de una molécula de DNA⁵. Se trata de un HBGM de nueve dimensiones ya que aparecen representadas 9 características genómicas distintas. Se compone de siete características genómicas que usan TSS y TES y las coordenadas inicio y fin de sus exones como puntos de referencia, La clase enhancer por su parte usa como puntos de referencia sus propias coordenadas inicio y fin. La parte de la molécula que no contiene ninguna característica genómica se considera de la clase intergénica.

Para cada una de las dos proteínas en estudio se ha utilizado un único HBGM, correspondiente a ambas cadenas, que tiene los parámetros mostrados en las tablas 4 y 5:

HBGM ID	Cadena	Regla de prioridad	Longitud de pico (pb)	Proporción de solapamiento (%)	Tamaño de la muestra
BM:9dm	Ambas	promotora>proximal>distal> enhancer>escisión>empalme>exón> intrón>intergénica	31	20	{1..22, X}

Tabla 4. Representación de los parámetros usados para generar el HBGM de la proteína BCLAF1 en la línea celular k-562

HBGM ID	Cadena	Regla de prioridad	Longitud de pico (pb)	Proporción de solapamiento (%)	Tamaño de la muestra
BM:9dm	Ambas	promotora>proximal>distal> enhancer>escisión>empalme>exón> intrón>intergénica	31	20	{1..22, X,Y}

Tabla 5. Representación de los parámetros usados para generar el HBGM de la proteína CTCF en la línea celular astrocito

3.3.2.2. Human Background Gene (HBGN):

El Background Gene (BGN) se trata de todos los sitios de unión anotados presentes en un GNM que sigue unos parámetros concretos. Dichos sitios de unión anotados presentan la misma probabilidad de que la proteína estudiada se una a ellos, y a cada sitio de unión anotado en un BGN se asigna una única clase de las distintas características genómicas descritas en el GNM del que deriva dicho BGN. Las características genómicas que forman el BGN incluyen las regiones funcionales relacionadas con el GNM del que deriva así como las que pertenecen a las entidades funcionales asociadas a dicho GNM. Puesto que la clase intergénica no se encuentra definida en el GNM del que derivan, los BGNs tampoco incluyen dicha clase⁵.

Los atributos necesarios para configurar un BGN son:

- 1) **Transcrito:** puede ser canónico o específico de línea celular. En este caso se usa el canónico
- 2) **Longitud de pico:** es la longitud que todos los sitios de unión anotados comparten.
- 3) **Proporción de solapamiento:** porcentaje que determina el menor número de nucleótidos pertenecientes a un sitio de unión anotado que debe superponerse con la región de una característica genómica específica para considerar que ese sitio pertenece a dicha clase.
- 4) **Regla de prioridad:** en caso de que para un sitio de unión anotado se superpongan dos o más características genómicas, determina cuál de las clases se le debe adjudicar, logrando que cada sitio de unión anotado pertenezca a una única clase.

En este caso el BGN estudiado es el que se corresponde con la población de sitios de unión anotados de un individuo humano (HBGN). En dicho HBGN se consideran tanto las características genómicas definidas en el *Human Gene Model* (HGNM), como las características genómicas de clase *enhancer*.

Los picos de unión anotados que conforman un BGN pertenecen al BGM que se obtiene con los mismos valores para los parámetros longitud de pico, proporción de solapamiento, regla de prioridad y cadena que dicho BGN⁵. Para ambos casos, el resto de parámetros empleados es el mismo que para sus respectivos HBGM, es decir, una longitud de pico de 31 pares de bases, una proporción de solapamiento del 20% y la regla de prioridad promotora > proximal > distal > *enhancer* > escisión > empalme > exón > intrón.

3.3.2.3. ChIP-Seq Model (ChSqM):

En cada línea celular se lleva a cabo un proceso de estandarización de pico a partir de sus respectivos *ChIP-Seq Datasets*, en el que se pretende delimitar la región que contiene el sitio de unión dentro de un pico en bruto⁵. Para ello primero se lleva a cabo una selección de picos, en la que los picos en bruto que se considera que comparten un mismo sitio de unión son agrupados, obteniéndose así un único pico a partir de ellos, el cual es denominado pico estandarizado⁵. A continuación se calcula el centro del pico estandarizado, que resulta de calcular la media aritmética de los centros o los *summit* (coordenada dentro de un determinado pico en la que existe mayor probabilidad de que se encuentre el centro del sitio de unión de la proteína) de los picos en bruto que lo conforman. En este trabajo se ha usado la media aritmética de los *summit* para calcular el centro⁵. Por último se determina la longitud del pico, cuyo valor coincide con el del BGM con el que se analizará el ChSqM y es de 31 pb para los ChSqM derivados de ambas proteínas⁵.

Los atributos de un pico estandarizado son: **i)** Cromosoma, **ii)** Inicio, **iii)** Fin, **iv)** Cadena, **v)** Centro del pico y **vi)** Longitud del pico.

3.4. Marco de análisis:

En este apartado se describe la metodología que se sigue para el estudio a nivel de la distribución de las clases de sitios de unión anotados a las que se une la proteína (Nivel de región) se muestran sus preferencias por unas clases u otras, a nivel de cada uno de los GNMs pertenecientes al BGM utilizado (Nivel de gen) se obtiene en cuál de ellos la proteína presenta un comportamiento más significativamente alejado de lo esperado, y a nivel de cada uno de los recursos funcionales modelados (Modelo funcional: nivel de ruta) se obtiene el efecto acumulado de la proteína en base a los elementos que los conforman⁵.

3.4.1. Nivel de región:

Los picos estandarizados generados en el ChSqM se someten a un proceso de anotación sobre un HBGM, que actúa como la población de referencia y consta de sitios de unión anotados, generando así su respectivo ChIP-Seq anotado. Tras ello, se realiza el recuento de cuántos picos anotados corresponden a cada característica genómica presente en el HBGM, a partir del cual, usando un script en el lenguaje de programación R v3.6.2, se puede obtener su perfil característico, en el que se incluyen datos estadísticos para realizar análisis más exhaustivos⁵.

El marco de análisis a nivel de región de un perfil característico concreto permite analizar si la proteína a la que dicho perfil pertenece se une a los diferentes sitios de unión por azar o si se une de manera predefinida a regiones correspondientes a una determinada clase genómica. Para ello se lleva a cabo un *exact test of goodness of fit Chi-squared*⁵.

El valor de Z-score es dependiente de la cantidad de veces observadas que la proteína se une a sitios de unión categorizados para cada característica genómica y a partir de él se puede calcular su puntuación de preferencia. Se calcula mediante la fórmula⁵:

$$Z_{score} = \frac{X_i - \bar{X}_i}{s_i}$$

A su vez permite comparar si dos ChIP-Seq anotados corresponden a un mismo HBGM sobre el que se han anotado. Para esto se realiza un *exact test of homogeneity of fit Chi-squared*, que permite cuantificar la similitud general de comportamiento, las diferencias de comportamiento acorde a cada característica genómica, y evaluar de forma cualitativa cómo evoluciona dicho comportamiento⁵.

3.4.2. Nivel de gen:

El objetivo de este nivel es asignar una puntuación de gen a cada uno de los GNMs en función de su respectivo BGN utilizado en la generación de un BGM, sobre el cual se ha llevado a cabo el proceso de anotación de los picos de un ChSqM. La puntuación de gen indica la influencia que tiene la proteína estudiada sobre el respectivo un GNM. Las fases para calcular dicha puntuación de gen, usando al ser humano como organismo modelo son descritas en la Tesis de Ginés Almagro⁵:

Todos aquellos GNMs, para un ChSqM concreto, que presentan una puntuación de gen mayor a 0, forman el ranking característico. Dicho ranking muestra a qué GNMs afecta más la proteína en estudio, tanto a nivel general (puntuación de gen) como de las características genómicas (puntuación de característica). El ranking obtenido se puede someter a un proceso de filtrado según el q-value de sus GNMs, generando el ranking seleccionado⁵. Para los tres experimentos que usan BCLAF1 se han seleccionado 5855, 1358 y 523 GNMs respectivamente, Para los experimentos en

cerebro, médula espinal y cerebelo que usan CTCF se han seleccionado 3037, 4882 y 4380 GNMs respectivamente.

El análisis a este nivel permite comparar diferencias de HGNMs entre dos ChSqMs anotados sobre un mismo HBGM. Para ello se agrupan los ChSqMs y se seleccionan todos los vectores gen que presentan significancia en al menos uno de los rankings característicos. A continuación se procede a la estandarización de dichos vectores, tras lo que es posible obtener el vector gen diferencia, al llevar a cabo la resta de los vectores estandarizados que representan el HGNM en cada uno de los ChSqMs, cuyo módulo es el atributo “*dist*”. La representación de las distintas características genómicas y su atributo “*dist*” se conoce como *ránking diferencial*. Cuanto mayor sea el atributo *dist*, más afectado estará por la proteína un vector gen que el otro. Se puede realizar una clasificación de HGNMs según su *ranking diferencial* de cada par de ChSqM mediante la elaboración de una gráfica *Heatmap*⁵. Una vez obtenido el *ranking diferencial* asociado a un par de ChSqMs, se realiza el test estadístico basado en rankings *Wilcoxon signed-rank test*, usando la función *wilcox.test* del paquete “*stats*” de R v.3.6.1, que permite realizar una comparación cuantitativa⁵.

3.4.3. Nivel de ruta:

En esta etapa se identifica cómo afecta la proteína a nivel funcional o biológico en el organismo, realizado en este estudio a nivel de ruta metabólica. Para ello se han dividido las distintas rutas en tres niveles jerárquicos. Se pretende asignar una puntuación de función a cada FM, la cual indica cuán influenciado está dicho FM por la proteína estudiada. A cada FM se le asigna su correspondiente vector funcional a partir de su vector gen, procedente del *ranking* característico obtenido en el nivel de gen, a partir del cual se puede calcular su puntuación de función. En este nivel se considera un vector funcional como el resultado de combinar todos los vectores gen que pertenecen a su FM, y que tienen como coeficiente los factores de relevancia adjudicados dentro de la ruta (1 para todos los vectores considerados en este caso). La puntuación de ruta se obtiene a partir del módulo del vector ruta en los tres niveles jerárquicos en que se encuentran divididas⁵. Al igual que para el nivel de gen, es posible realizar una comparación cuantitativa por pares de ChSqMs a partir de sus *rankings diferenciales*, que se obtienen una vez los vectores se han estandarizado. Una vez obtenidos los *rankings diferenciales* asociados a cada par de ChSqMs en los tres niveles, se realiza el test estadístico basado en *rankings Wilcoxon signed-rank test*, al igual que se ha descrito en el apartado anterior, pero esta vez se realiza para los tres niveles jerárquicos en que se organizan las rutas⁵.

4. Resultados:

Aquí se muestran los resultados obtenidos del estudio realizado para ambas proteínas, indicando características específicas de las mismas a tres niveles a los tres niveles mencionados en el apartado 3.3.

4.1. BCLAF1:

En esta sección se analiza el comportamiento de la proteína BCLAF1 en la línea celular K-562. Para ello se han utilizado los modelos ChIP-Seq obtenidos de tres experimentos distintos, que son ENCSR000BKH, ENCSR492LTS y ENCSR792IYC, cuya condición es que usasen la misma proteína y línea celular¹⁵.

El objetivo de este estudio es determinar si las condiciones entre los tres experimentos que utilizan el *wild type* de la misma proteína en la misma línea celular son similares y así poder encontrar patrones específicos para dicha proteína.

4.1.1. Nivel de región:

Se sigue lo descrito en el apartado 3.3.1. En las figuras 5, 6 y 7 se pueden observar respectivamente representados los valores para las frecuencias relativas de cada una de las características genómicas que componen el HBGM de BM: 9dm empleado en los tres experimentos anteriormente mencionados.

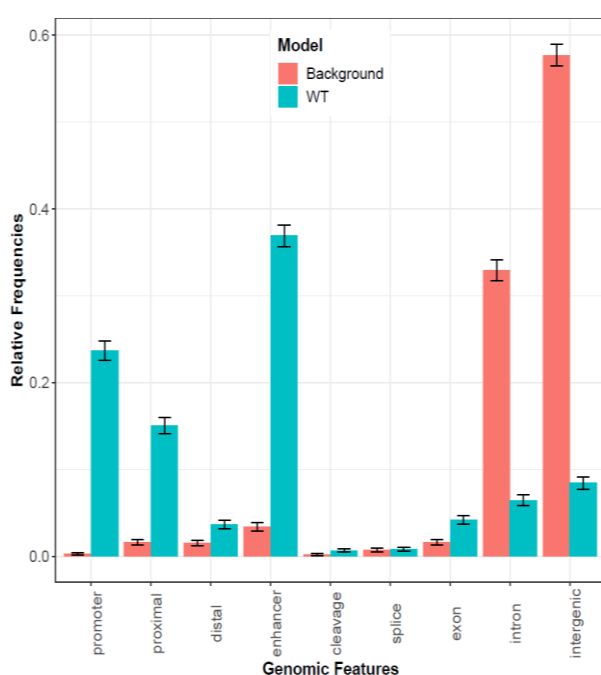


Figura 5. Representación de las frecuencias relativas a cada característica genómica para el HBGM de 9dm obtenido a partir de los datos para el experimento ENCSR000BKH. Los valores representados en rojo corresponden a los valores esperados de acuerdo a los cálculos derivados del HBGN, mientras que los valores representados en azul se corresponden con los observados en cada experimento.

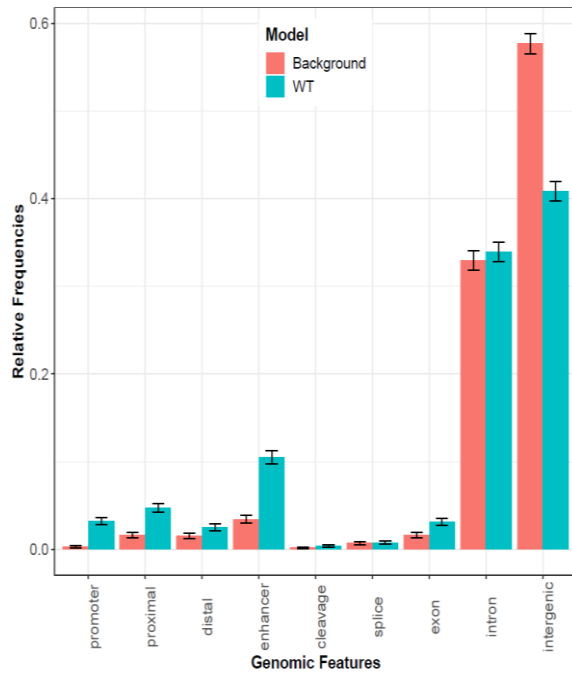


Figura 6. Representación de las frecuencias relativas a cada característica genómica para el HBGM de 9dm obtenido a partir de los datos del experimento ENCSR492LTS. Los valores representados en rojo corresponden a los valores esperados de acuerdo a los cálculos derivados del HBGN, mientras que los valores representados en azul se corresponden con los observados en cada experimento.

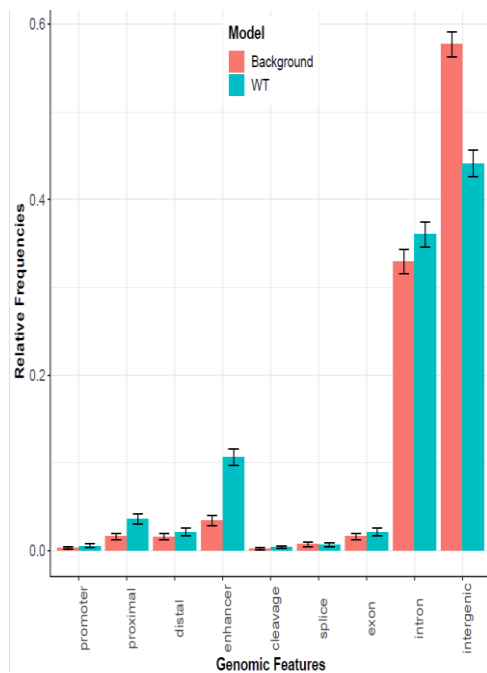


Figura 7. Representación de las frecuencias relativas a cada característica genómica para el HBGM de 9dm obtenido a partir de los datos del experimento ENCSR792IYC. Los valores representados en rojo corresponden a los valores esperados de acuerdo a los cálculos derivados del HBGN, mientras que los valores representados en azul se corresponden con los observados en cada experimento.

Para el primer experimento el 92% de los picos anotados pertenecen a una característica genómica que pertenece a una entidad funcional mientras que el 8% restante pertenece a la clase intergénica, que no tiene función conocida. Para el segundo experimento el 60% corresponde alguna característica genómica con entidad funcional. En el caso del tercer experimento, tan solo el 56% de los picos tienen función conocida. Para los tres experimentos, en contra de lo esperado por azar, la

proteína se une menos a regiones intergénicas y más a regiones de función conocida, destacando la clase *enhancer* y promotora para el primer experimento.

Para saber si la proteína sigue un patrón aleatorio o no a nivel de todo el genoma, se lleva a cabo *un exact test of goodness of fit Chi-squared*. Debido a que para los ChIP-Seq anotados de los tres experimentos los *exact tests of goodness of fit Chi-squared* devuelven un *p-value* aproximadamente 0, se confirma que el comportamiento de la proteína BCLAF1 en la línea celular K-562 no es aleatorio, sino que presenta predisposición a unirse a regiones correspondientes a una determinada clase genómica en un grado diferente a lo que cabría esperar por azar.

El perfil de preferencia de BCLAF1 por una determinada clase de característica genómica resulta contradictorio ya que, si se analizan los Z-score de los tres experimentos, no es la misma clase genómica la que se encuentra mayormente representada. De este modo para el primer experimento mencionado el valor de Z-score más alto es el de la clase promotor, con un valor de 328, y el segundo valor de Z-score más elevado es el de la clase *enhancer*, con un valor de 141. Destaca además que la clase intergénica tiene un valor de -76, por lo que se observa que la proteína se une a menos regiones de las cuales no se conoce función de lo que se esperaba acorde al HBGM empleado. Del mismo modo, el valor de Z-score para la clase genómica intrón es negativo, es decir, se observa que la proteína se une menos a regiones de dicha clase genómica (Figura 8). De acuerdo a los resultados del segundo experimento, el Z-score más elevado es el de la clase promotora, que tiene un valor de 44, y el segundo Z-score más alto es el de la clase *enhancer*, con un valor 32, por lo que, tanto este experimento como el primero concuerdan en que la proteína BCLAF1 tienen preferencia por unirse a regiones de la clase genómica promotora y *enhancer*. Al igual que pasaba en el caso anterior, el valor de Z-score para la clase intergénica es negativo, en este caso -28, lo que significa que se conoce la función de más regiones de las que cabía esperar. Sin embargo, en el tercer experimento, el valor de Z-score más alto obtenido corresponde a la clase *enhancer*, con un valor de 26, mientras que el segundo Z-score con valor más elevado es el de la clase proximal, con un valor de 10. Para este experimento, al igual que pasaba con los anteriores, el valor de Z-score para la clase intergénica es negativo, siendo en este caso de -18.

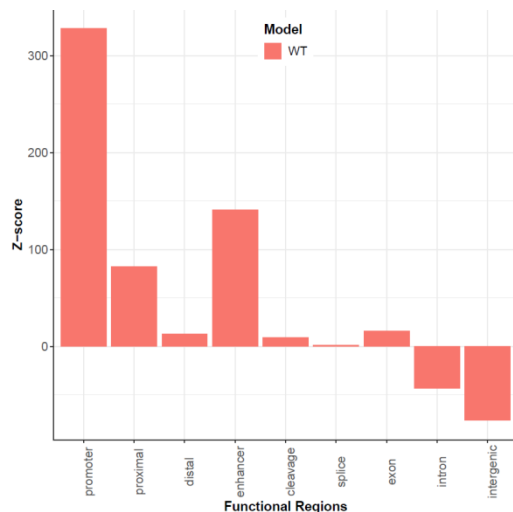


Figura 8. Representación de los valores de Z-score para cada clase genómica del experimento ENCSR000BKH. Los valores positivos corresponden a las regiones con características genómicas a las que la proteína se une más de lo esperado, destacando la clase promotora y *enhancer*. Los valores negativos pertenecen a las regiones a las que la proteína se une menos de lo esperado, que son regiones de la clase intergénica e intrón.

Al comparar los tres ChIP-Seq anotados de los tres experimentos mediante *an exact test of homogeneity of fit Chi-squared*, el comportamiento más similar de BCLAF1 se da entre los ChSqMs correspondientes a ENCSR492LTS y a ENCSR792IYC. De entre las otras dos comparaciones, se puede determinar que los dos ChSqMs que más difieren entre sí son los correspondientes a ENCSR000BKH y ENCSR492LTS (Tabla 6).

	WT_ENCSR000BKH - WT_ENCSR492LTS			WT_ENCSR000BKH - WT_ENCSR792IYC			WT_ENCSR492LTS - WT_ENCSR792IYC		
	Estadística	Grado de libertad	p-value	Estadística	Grado de libertad	p-value	Estadística	Grado de libertad	p-value
Human background genome									
BM:9dm	6041.04	8	0	5568.02	8	0	127.62	8	8.79×10^{-24}

Tabla 6. Valores obtenidos a partir del *exact test of homogeneity of fit Chi-squared* llevado a cabo para los distintos pares de los tres experimentos que emplean la proteína BCLAF1 en la línea celular k-562.

Analizando las características genómicas, se puede determinar que las mayores diferencias entre los ChSqMs de ENCSR000BKH y de ENCSR492LTS se deben a la clase intergénica e intrón, seguido de la clase *enhancer* y promotora.

Las clases genómicas que mayoritariamente contribuyen a las diferencias entre los ChSqMs de ENCSR000BKH y de ENCSR792IYC son la clase intergénica e intrón, y posteriormente la clase promotora y *enhancer*.

Por su parte, las diferencias entre los ChSqMs de ENCSR492LTS y de ENCSR792IYC son consecuencia en su mayor parte de la clase promotora (Figura 9).

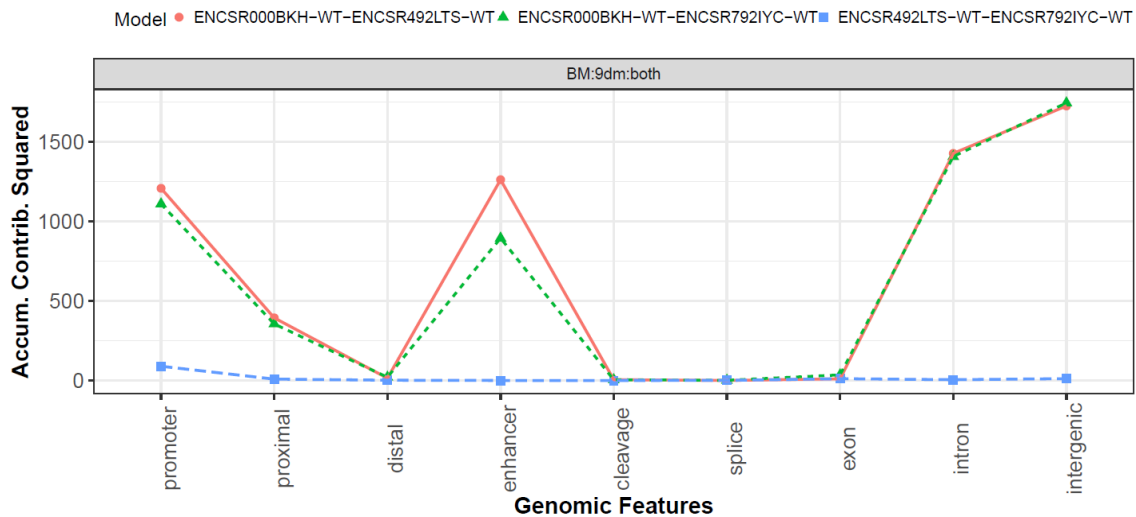


Figura 9. Representación gráfica de las *accumulated contribution squared* por cada característica genómica en los distintos pares de los tres experimentos que emplean la proteína BCLAF1 en la línea celular k-562.

La representación de las frecuencias relativas para los tres ChSqMs respecto a la variación de cada característica genómica determina, como se ha mencionado anteriormente, que los modelos derivados de los datos obtenidos para el experimento 2 y para el experimento 3 son los más similares (Figura 10).

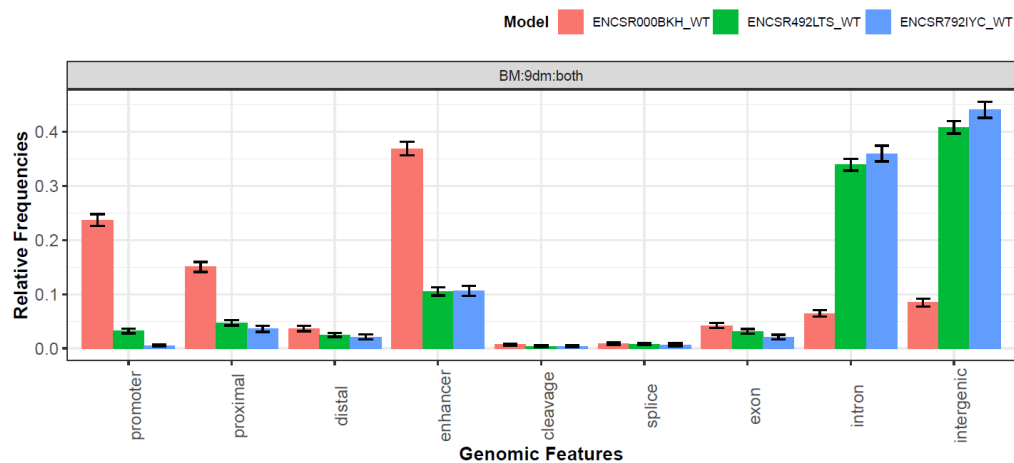


Figura 10. Gráfico de barras de las *joint relative frequencies* y sus intervalos de confianza 95% por cada característica genómica para los tres ChSqMs derivados de los experimentos ENCSR000BKH, ENCSR492LTS y ENCSR792IYC

4.1.2. Nivel de gen:

Para realizar este análisis se han seguido los pasos descritos en el apartado 3.3.2. En este caso el ranking seleccionado se corresponde con el ranking característico, y se han obtenido tres rankings característicos en total, uno para cada ChSqM derivado de los tres experimentos anteriormente mencionados. Para cada ranking característico los

HGNMs que poseen mayor puntuación de gen, es decir, más afectados por la proteína son "HIST1H4I" para el experimento 1, cuyo valor se debe principalmente a la clase *enhancer*, "ZMYND8" para los experimentos 2 y 3, cuyo valor se debe al igual que en el primero a la clase *enhancer*. Tras obtener el ranking diferencial, se puede realizar una clasificación de HGNMs de cada par de ChSqM mediante la elaboración de una gráfica *Heatmap* (Figura 11). Una vez obtenido el ranking diferencial asociado a cada par de ChSqMs, se ha realizado una comparación cuantitativa empleando el test estadístico basado en rankings *Wilcoxon signed-rank test* (Figura 12).

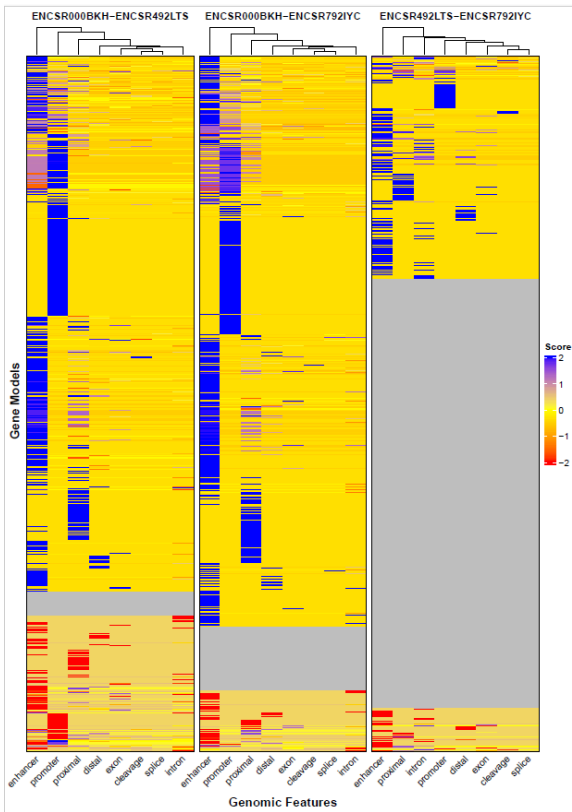


Figura 11. Gráfica *Heatmap* obtenida para las tres posibles combinaciones de pares de ChSqMs derivados de los tres experimentos que usan BCLAF1 como proteína y k-562 como línea celular. Los HGNMs con una puntuación de gen igual a cero son irrelevantes para un par puesto que no se encuentran afectados en ninguno de los ChSqMs (color gris). Los HGNMs se ordenan de mayor a menor atributo "dist", siendo los HGNMs de la parte superior los que se encuentran más afectados en el primer componente del par de ChSqMs, los de la parte central los que se encuentran igual de afectados en ambos ChSqMs, y los de la parte inferior los que se afectan más en el segundo ChSqM.

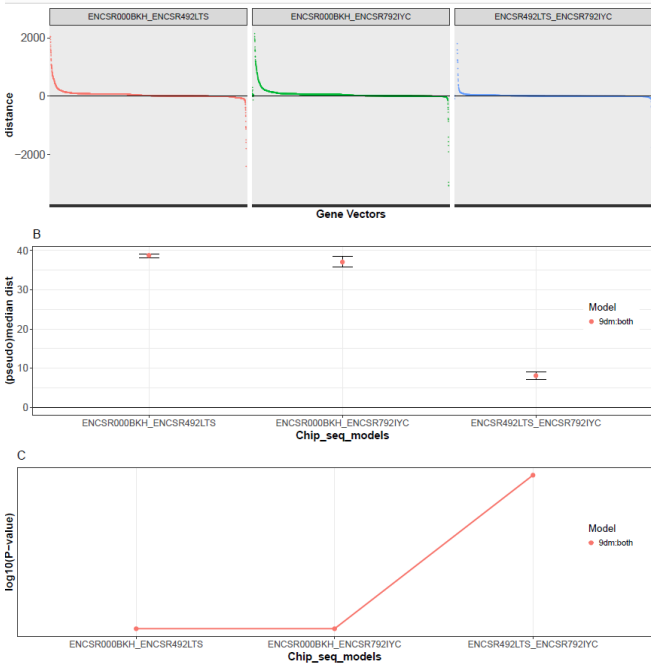


Figura 12. Representación de los resultados del test *Wilcoxon* a nivel de gen para los pares de ChSqMs obtenidos a partir de los tres experimentos de la proteína BCLAF1 en la línea celular k-562. A) Gráfica de los ranking diferenciales para los distintos pares derivados de los tres experimentos analizados. B) (Pseudo)mediana de la distancia con su intervalo de fiabilidad del 95% para cada par de ChSqMs. C) Representación de los p-valores obtenidos al realizar el *Wilcoxon Signed-rank test* en los rankings diferenciales de A), expresados con signo negativo en escala logarítmica.

Un p-value bajo, como en el caso de estos experimentos, indica que es muy probable que exista un efecto global entre dicho par, siendo el par de experimentos 2-3 el de menor probabilidad ya que tiene un p-valor alto relativo al de los otros dos pares (el p-valor del par 2-3 es $2.0648 \cdot e^{-146}$ en un rango de $[1.25 \cdot e^{-302}, 1]$, mientras que el de los otros dos pares es 0 en un rango de valores $[0, 1]$). La (pseudo)mediana indica que globalmente los HGNMs a los que afecta más la proteína se localizan en el experimento 1, ya que en ambos pares de ChSqMs a los que pertenece, 1-2 y 1-3, el mayor efecto ocurre a favor de él. Entre los ChSqMs que conforman el par 2-3, existe un efecto a favor del ChSqM derivado del experimento 2, pero es considerablemente menor que el efecto sobre los otros dos pares. Los perfiles de los rankings diferenciales indican que dichos efectos se deben sobre todo a la acumulación de vectores gen diferencia con bajo valor dist, aunque en ambos extremos aparecen HGNMs con alto valor absoluto de dist. Además, los vectores gen diferencia que representan los HGNMs del par 2-3 se disponen de forma más simétrica.

4.1.3. Nivel de ruta:

En esta etapa se siguen los pasos descritos en el apartado 3.3.3 para analizar a nivel de ruta la proteína BCLAF1. Como este estudio se realiza a nivel de ruta metabólica, los vectores obtenidos a partir de los rankings característicos de los HPMs se denominan vectores ruta. El ranking seleccionado, al igual que el ranking característico, coincide con el del nivel de gen y, al igual que en dicho nivel, coincide con la totalidad del ranking característico. La mayor puntuación de ruta, es decir, la ruta más afectada por la proteína, en el nivel jerárquico 1 corresponde a la ruta “Signal Transduction:R-HSA” para el primer experimento pero a “Metabolism:R-HSA” para los otros dos, a nivel 2 corresponde a “RNA Polymerase II Transcription:R-HSA” para los tres experimentos, y a nivel 3 corresponde a “Generic Transcription Pathway:R-HSA” para los tres. Una vez obtenidos los rankings diferenciales asociados a cada par de ChSqMs en los tres niveles (9 en total, 3 por cada uno de los tres niveles jerárquicos en que se han organizado las rutas), se realiza el test estadístico basado en rankings Wilcoxon signed-rank test en los tres niveles jerárquicos en que se organizan las rutas para llevar a cabo una comparación cuantitativa (Figura 13).

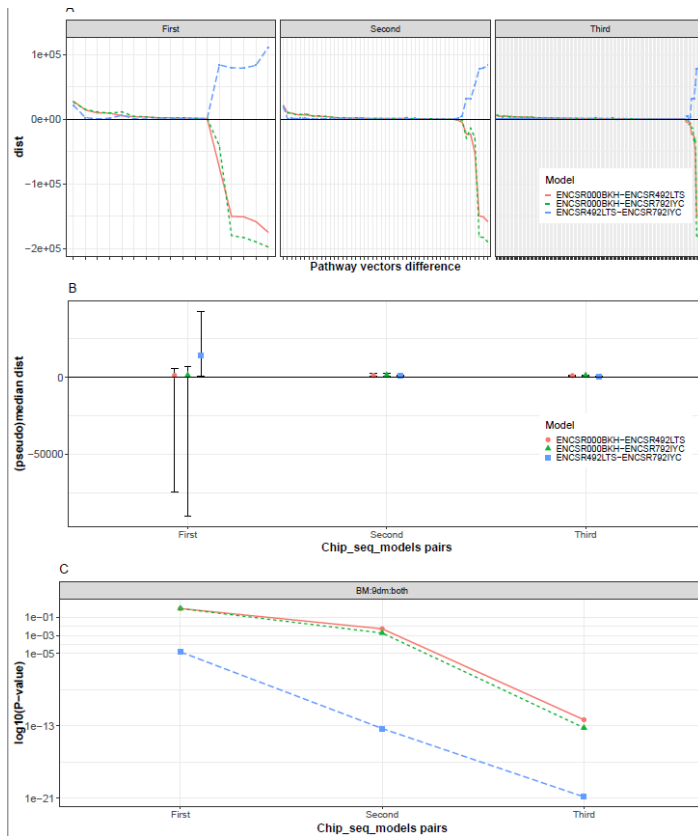


Figura 13. Representación de los resultados del test Wilcoxon a nivel de ruta para los pares de ChSqMs obtenidos a partir de los tres xperimentos de la proteína BCLAF1 en la línea celular k-562.

A) Gráfica de los ranking diferenciales para los distintos pares derivados de los tres experimentos analizados, representados para los tres niveles jerárquicos de ruta. B) (Pseudo)mediana de la distancia con su intervalo de fiabilidad del 95% para cada par de ChSqMs en los tres niveles jerárquicos de ruta. C) Representación de los p-values obtenidos al realizar el Wilcoxon Signed-rank test en los rankings diferenciales de A), expresados con signo negativo en escala logarítmica.

En el primer nivel jerárquico de ruta, el p-value para el par de experimentos 2-3 es de $1.53 \cdot 10^{-5}$ mientras el de los otros dos pares de experimentos es de 0.96 en un intervalo de $[1.53 \cdot 10^{-5}, 1]$ para los tres experimentos. En el segundo nivel jerárquico de ruta, para el par 1-2 es de 0.005, para el par 1-3 es de 0.002, y para el par 2-3 es de $4.97 \cdot 10^{-14}$, en un intervalo de $[3.55 \cdot 10^{-15}, 1]$ para los tres experimentos. En el tercer nivel jerárquico, el p-value para el par 1-2 es de $4.66 \cdot 10^{-13}$ en un intervalo $[8.55 \cdot 10^{-50}, 1]$, para el par 1-3 es de $5.98 \cdot 10^{-14}$ en un intervalo $[1.71 \cdot 10^{-49}, 1]$ y para el par 2-3 es de $1.36 \cdot 10^{-21}$ en un intervalo $[2.19 \cdot 10^{-47}, 1]$. El par 2-3, constituido por ENCSR492LTS y ENCSR792IYC, presenta la mayor probabilidad en los tres niveles. La (pseudo)mediana indica que en general los HPMS más afectados de los tres ChSqMs estudiados son los del primer experimento (ENCSR000BKH) porque el mayor efecto ocurre a su favor. Los HPMS menos afectados de los tres corresponden al tercer experimento (ENCSR792IYC). Acorde a los rankings diferenciales representados, los efectos se deben a valores elevados de dist que tienen signo negativo en los tres grados jerárquicos, y se dan entre ENCSR000BKH y los otros dos ChSqMs. Existe un reparto asimétrico respecto a $dist=0$, pero conforme aumenta el grado jerárquico dicho reparto se vuelve más simétrico para los tres pares. El par que tiene la distribución más simétrica en los tres grados jerárquicos es el par ENCSR492LTS-ENCSR792IYC.

4.2. CTCF:

En esta sección se analiza el comportamiento de la proteína CTCF en la línea celular astrocito. Para ello se han utilizado los modelos ChIP-Seq obtenidos de tres experimentos distintos, que son ENCSR000AOO, ENCSR000DSU y ENCSR000DSZ, que usan dicha proteína y línea celular extraída de secciones de distintos tejidos, que son cerebro, médula espinal y cerebelo respectivamente.

El objetivo de este estudio es determinar si los resultados para la misma proteína en la misma línea celular se ven afectados según la sección de tejido de la que se extraen, y poder encontrar patrones específicos en tejido para la proteína.

4.2.1. Nivel de región:

El procedimiento a seguir es el mismo que el descrito en el apartado 3.3.1. En las figuras 14, 15 y 16 se pueden observar representados los valores para las frecuencias relativas de cada una de las características genómicas que componen el HBGM de BM:9dm empleado en los tres experimentos anteriormente mencionados.

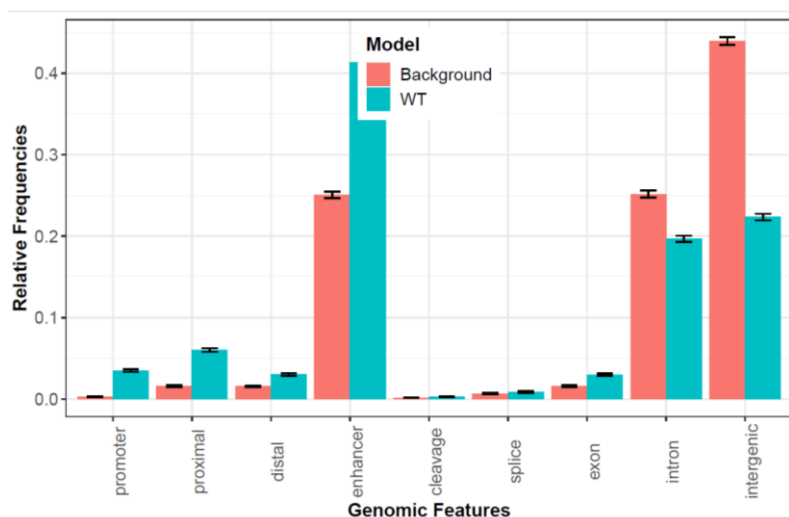


Figura 14. Representación de las frecuencias relativas a cada característica genómica para el HBGM de 9dm obtenido a partir de los datos del experimento ENCSR000AOO. Los valores representados en rojo corresponden a los valores esperados de acuerdo a los cálculos derivados del HBGN, mientras que los valores representados en azul se corresponden con los observados en cada experimento.

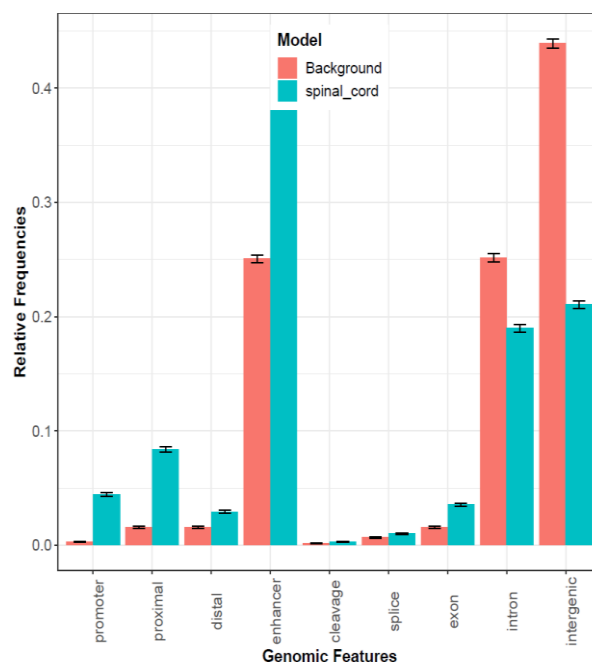


Figura 15. Representación de las frecuencias relativas a cada característica genómica para el HBGM de 9dm obtenido a partir de los datos del experimento ENCSR000DSU. Los valores representados en rojo corresponden a los valores esperados de acuerdo a los cálculos derivados del HBGN, mientras que los valores representados en azul se corresponden con los observados en cada experimento.

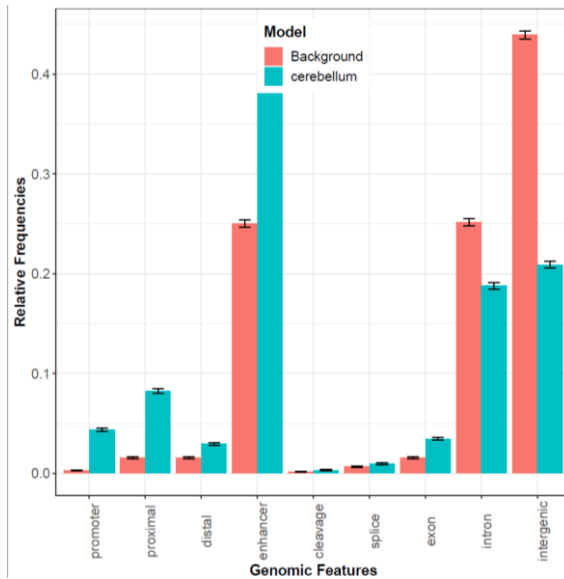


Figura 16. Representación de las frecuencias relativas a cada característica genómica para el HBGM de 9dm obtenido a partir de los datos del experimento ENCSR000DSZ Los valores representados en rojo corresponden a los valores esperados de acuerdo a los cálculos derivados del HBGN, mientras que los valores representados en azul se corresponden con los observados en cada experimento.

Para el primer experimento el 78% de los picos anotados pertenecen a una característica genómica que pertenece a una entidad funcional mientras que el 22% restante pertenece a la clase intergénica, que no tiene función conocida. Para el segundo experimento el 79% corresponde alguna característica genómica con entidad funcional. De igual forma, en el caso del tercer experimento, el 79% de los picos tienen función conocida.

En el caso de esta proteína, para los ChIP-Seq anotados de los tres experimentos los exact tests of goodness of fit Chi-squared devuelven un p-value aproximadamente 0, confirmando que el comportamiento de la proteína CTCF en la línea celular astrocito no es aleatorio, sino que presenta predisposición a unirse a regiones correspondientes a una determinada clase genómica en contra de lo que cabría esperar por azar.

El perfil de preferencia de CTCF por una determinada clase de característica genómica resulta, en general, es concordante ya que para los Z-score de los tres experimentos se trata de la misma clase genómica la que se encuentra mayormente representada. De este modo, para el experimento que emplea cerebro como tejido, el valor de Z-score más alto es el de la clase promotora, con un valor de 121, y el segundo valor de Z-score más elevado es el de la clase enhancer (77), seguido muy de cerca por la clase proximal (73). Destaca además que la clase intergénica tiene un valor de -90, por lo que se observa que la proteína se une a menos regiones de las cuales no se conoce función de lo que se esperaba acorde al HBGM empleado. Del mismo modo, el valor de Z-score para la clase genómica intrón es negativo (-26), es decir, se observa que la proteína se une menos a regiones de dicha clase genómica (Figura 17). De acuerdo a los resultados del experimento que emplea médula espinal,

el Z-score más elevado es el de la clase promotora, que tiene un valor de 185, y el segundo Z-score más alto, a diferencia que en el caso anterior, es el de la clase proximal, con un valor de 132. El tercer valor más elevado es el de la clase *enhancer* (80). Por lo que, según este experimento, la proteína CTCF tiene preferencia por unirse a regiones de la clase genómica promotora, proximal y *enhancer*. Al igual que pasaba en el caso anterior, el valor de Z-score para la clase intergénica es negativo, en este caso -112, lo que significa que se conoce la función de más regiones de las que cabía esperar, y el valor de Z-score para la clase intrón también es negativo (-35). El perfil de Z-score para el experimento que usa cerebelo como tejido es muy similar al del segundo experimento, siendo el valor de Z-score más alto correspondiente a la clase promotora, con un valor de 173, el segundo valor de Z-score más elevado el de la clase proximal (123), y el tercero el de la clase *enhancer* (80). Para este experimento, al igual que pasaba con los anteriores, el valor de Z-score para las clases intergénica e intrón es negativo, siendo en este caso de -107 y -34 respectivamente.



Figura 17. Representación de los valores de Z-score para cada clase genómica del experimento que emplea cerebro. Los valores positivos corresponden a las regiones con características genómicas a las que la proteína se une más de lo esperado, destacando la clase promotora y *enhancer*. Los valores negativos pertenecen a las regiones a las que la proteína se une menos de lo esperado, que son regiones de la clase intergénica e intrón

Al comparar los tres ChIP-Seq anotados de los tres experimentos se determina que el comportamiento más similar de CTCF se da entre el par de ChSqMs correspondientes a médula espinal y cerebelo. De entre las otras dos comparaciones, se puede determinar que los dos ChSqMs que más difieren entre sí son los correspondientes a cerebro y médula espinal (Tabla 7).

	Cerebro - médula espinal			Cerebro-Cerebelo			Médula espinal- Cerebelo		
	Estadística	Grado de libertad	p-value	Estadística	Grado de libertad	p-value	Estadística	Grado de libertad	p-value
Human background genome									
BM:9dm	360.15	8	6.17*e ⁻⁷³	300.23	8	3.67*e ⁻⁶⁰	8.61	8	0.38

Tabla 7. Valores obtenidos a partir del exact test of homogeneity of fit Chi-squared llevado a cabo para los distintos pares de los tres experimentos que emplean la proteína CTCF en la línea celular astrocito.

Analizando las características genómicas, se puede determinar que las mayores diferencias entre los ChSqMs del experimento que usa cerebro y el que emplea médula espinal se deben a la clase proximal y promotora principalmente, seguido de la clase enhancer. Las clases genómicas que mayoritariamente contribuyen a las diferencias entre los ChSqMs de cerebro y de cerebelo son la clase proximal, promotora e intergénica. Por su parte, las diferencias entre los ChSqMs de cerebelo y de médula espinal son consecuencia en su mayor parte de la clase enhancer, empalme e intrón (Figura 18).

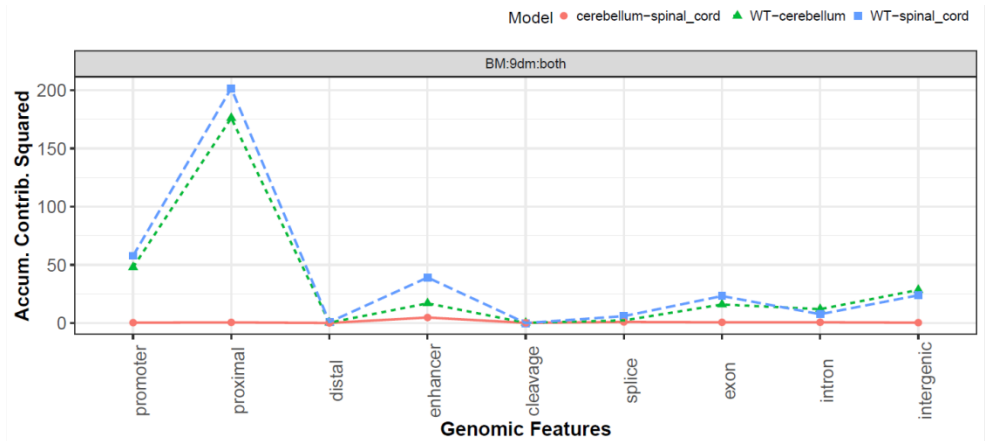


Figura 18. Representación gráfica de las accumulated contribution squared por cada característica genómica en los distintos pares de los tres experimentos que emplean la proteína CTCF en la línea celular astrocito.

La representación de las frecuencias relativas para los tres ChSqMs respecto a la variación de cada característica genómica determina, como se ha mencionado anteriormente, que los modelos derivados de los datos obtenidos para el experimento en cerebelo y para el experimento en médula espinal son los más similares, por lo tanto, y a pesar de que la proteína CTCF sigue un modelo muy similar en los tres tejidos usados para estos experimentos, el comportamiento de unión de dicha proteína resulta más parecido en células obtenidas de cerebelo y médula espinal. En general, CTCF tiene preferencia por regiones pertenecientes a características genómicas de la clase enhancer, intergénica e intrón (Figura 19).

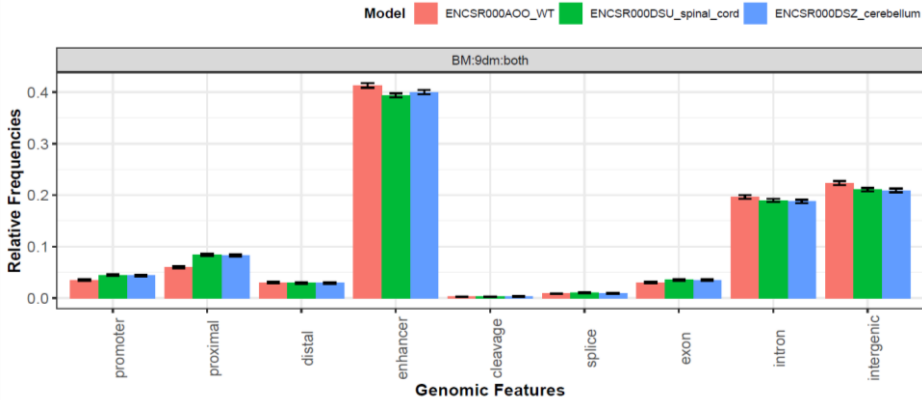
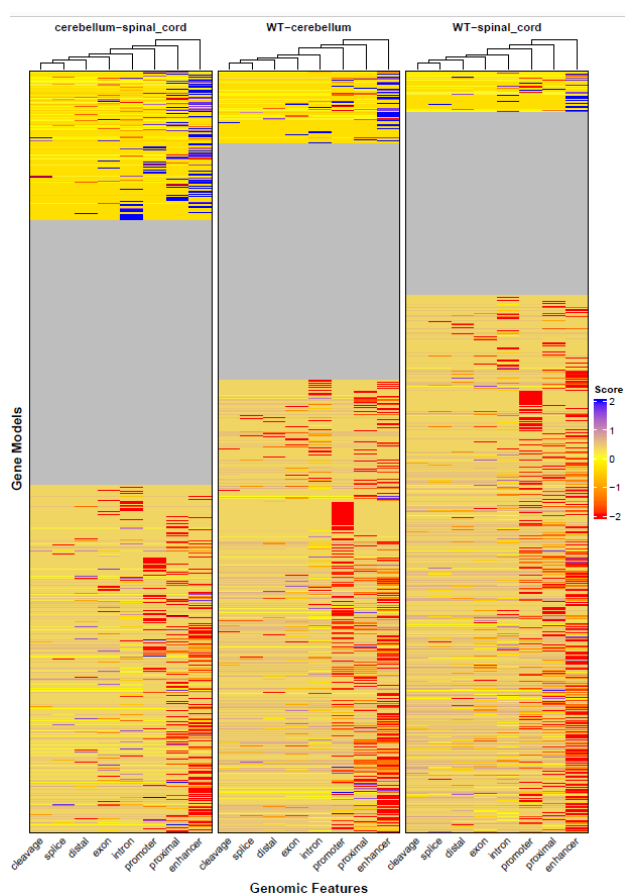


Figura 19. Gráfico de barras de las joint relative frequencies y sus intervalos de confianza 95% por cada característica genómica para los tres ChSqMs derivados de los experimentos ENCSR000A00, ENCSR000DSU y ENCSR000DSZ.

4.2.2. Nivel de gen:

En este caso se sigue el mismo procedimiento que el descrito en el apartado 3.3.2. Nivel de gen. Al igual que en dicho apartado, el ranking seleccionado se corresponde con el ranking característico, y se han obtenido tres rankings característicos en total, uno para cada ChSqM derivado de los tres experimentos mencionados que usan CTCF como proteína y astrocito como línea celular¹⁵. Para cada ranking característico los HGNMs que poseen mayor puntuación de gen, es decir, más afectados por la proteína son "MGST2" para cerebro y médula espinal, cuyo valor se debe principalmente a la clase *enhancer* en ambos, y "ZMIZ1" para cerebelo, cuyo valor al igual que en los otros dos casos se debe principalmente a la clase *enhancer*.



Como se ha mencionado, el análisis a este nivel permite comparar diferencias de HGNMs entre dos ChSqMs anotados sobre un mismo HBGM, representando una clasificación de HGNMs según su ranking diferencial de cada par de ChSqMs mediante una gráfica Heatmap (Figura 20).

Figura 20. Gráfica Heatmap obtenida para las tres posibles combinaciones de pares de ChSqMs derivados de los tres experimentos que usan CTCF como proteína y astrocito como línea celular. Los HGNMs con una puntuación de gen igual a cero son irrelevantes para un par puesto que no se encuentran afectados en ninguno de los ChSqMs (color gris). Los HGNMs se ordenan de mayor a menor atributo "dist", siendo los HGNMs de la parte superior los que se encuentran más afectados en el primer componente del par de ChSqMs, los de la parte central los que se encuentran igual de afectados en ambos ChSqMs, y los de la parte inferior los que se afectan más en el segundo ChSqM.

Una vez obtenido el ranking diferencial asociado a estos pares de ChSqMs, se realiza el test estadístico basado en rankings Wilcoxon signed-rank test, permitiendo realizar una comparación cuantitativa (Figura 21).

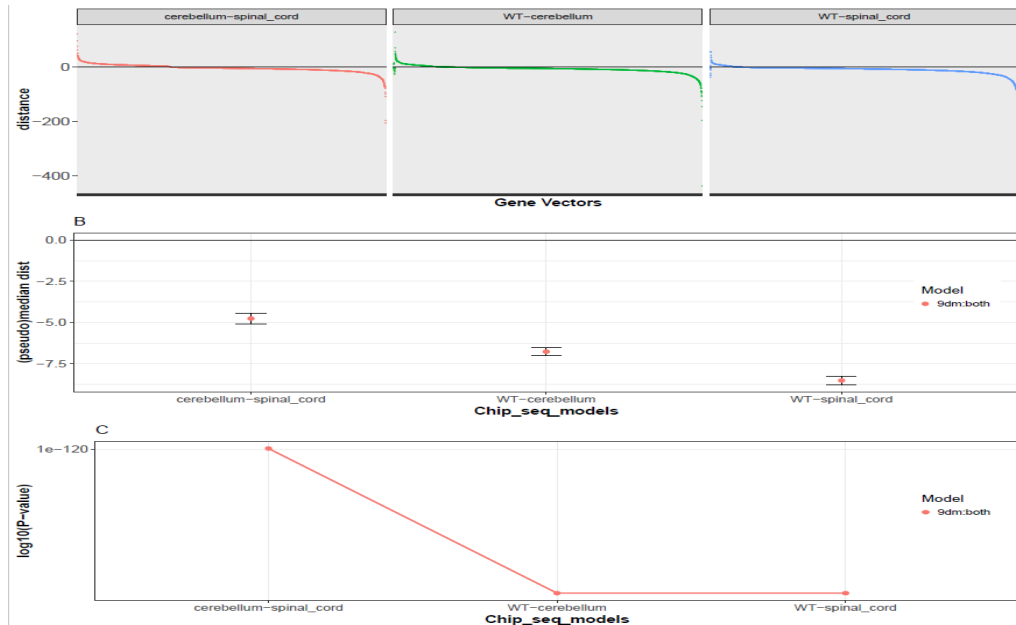


Figura 21. Representación de los resultados del test Wilcoxon a nivel de gen para los pares de ChSqMs obtenidos a partir de los tres experimentos de la proteína CTCF en la línea celular astrocito. A) Gráfica de los rankings diferenciales para los distintos pares derivados de los tres tejidos analizados. B) (Pseudo)mediana de la distancia con su intervalo de fiabilidad del 95% para cada par de ChSqMs. C) Representación de los p-values obtenidos al realizar el Wilcoxon Signed-rank test en los rankings diferenciales de A), expresados con signo negativo en escala logarítmica.

Un p-value bajo indica que es muy probable que exista un efecto global entre dicho par, como ocurre en los tres experimentos, siendo el par cerebelo-médula espinal el de menor probabilidad ya que tiene un p-valor alto relativo al de los otros dos pares (el p-valor del par cerebelo-médula espinal es $9.3267 \cdot 10^{-120}$ en un rango de $[0,1]$, mientras que el de los otros dos pares es 0 en un rango de valores $[0,1]$). La (pseudo)mediana indica que globalmente los HGNMs a los que más afecta la proteína se localizan en el modelo de médula espinal, ya que en ambos pares de ChSqMs a los que pertenece, cerebro- médula espinal y cerebelo-médula espinal, el mayor efecto ocurre a favor de médula espinal. Entre los ChSqMs que conforman el par cerebro-cerebelo, existe un efecto a favor del ChSqM derivado del experimento que emplea cerebelo como tejido. Los perfiles de los rankings diferenciales indican que dichos efectos se deben sobre todo a la acumulación de vectores gen diferencia con bajo valor dist, aunque en ambos extremos aparecen HGNMs con alto valor absoluto de dist. Además, los vectores gen diferencia que representan los HGNMs del par cerebelo-médula espinal se disponen de forma más simétrica.

4.2.3. Nivel de ruta:

En esta etapa se identifica cómo afecta la proteína a nivel funcional o biológico en el organismo, realizado en este estudio a nivel de ruta metabólica según las distintas secciones de tejido empleadas, para lo cual se siguen los pasos descritos en el apartado 3.3.3, esta vez con los experimentos correspondientes a la proteína CTCF.

El ranking seleccionado, al igual que el ranking característico, coincide con el del nivel de gen (Apartado 4.2.2) y, al igual que en dicho nivel, coincide con la totalidad del ranking característico. La mayor puntuación de ruta, es decir, la ruta más afectada por la proteína, en el nivel jerárquico 1 corresponde a la ruta “Signal Transduction:R-HSA” para los tres tejidos, a nivel 2 corresponde a “Post-translational protein modification:R-HSA” para los tres tejidos y a nivel 3 corresponde a “Generic Transcription Pathway:R-HSA” para los tres. Se realiza una comparación cuantitativa por pares de ChSqMs a partir de sus rankings diferenciales. En total se obtienen 9 rankings diferenciales, 3 por cada nivel jerárquico de los tres en que se han organizado las rutas. Una vez obtenidos los rankings diferenciales asociados a cada par de ChSqMs en los tres niveles, se realiza el test estadístico basado en rankings Wilcoxon signed-rank test, esta vez, para los tres niveles jerárquicos en que se organizan las rutas (Figura 22).

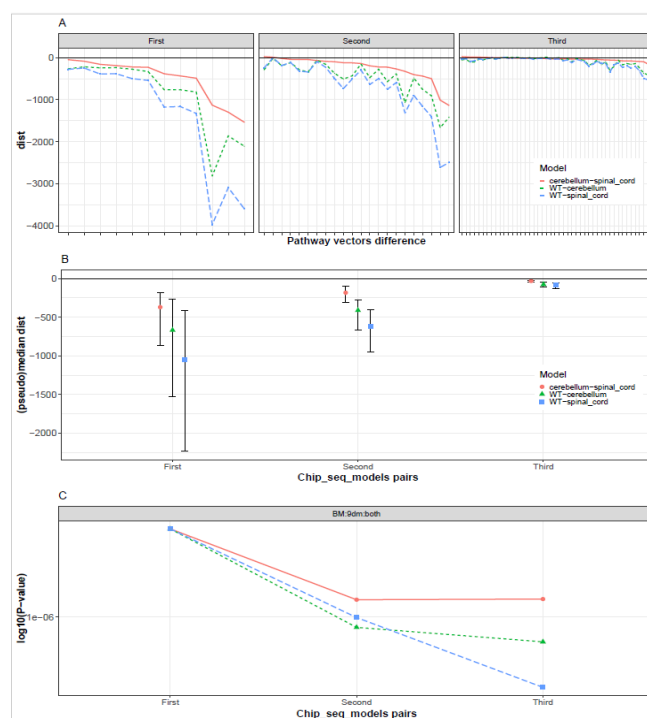


Figura 22. Representación de los resultados del test Wilcoxon a nivel de ruta para los pares de ChSqMs obtenidos a partir de los tres experimentos de la proteína CTCF en la línea celular astrocyto. A) Gráfica de los ranking diferenciales para los distintos pares derivados de los tres experimentos analizados, representados para los tres niveles jerárquicos de ruta. B) (Pseudo)mediana de la distancia con su intervalo de fiabilidad del 95% para cada par de ChSqMs en los tres niveles jerárquicos de ruta. C) Representación de los p-values obtenidos al realizar el Wilcoxon Signed-rank test en los rankings diferenciales de A), expresados con signo negativo en escala logarítmica.

En el primer nivel jerárquico de ruta, el p-value es de 0.00048 en un intervalo de $[0.00048, 1]$ para los tres experimentos. En el segundo nivel jerárquico de ruta, para el par cerebro-médula espinal es de 9.54×10^{-7} en un intervalo de $[9.54 \times 10^{-7}, 1]$ mientras que para el par cerebro- cerebelo es de 4.77×10^{-7} en un intervalo $[4.77 \times 10^{-7}, 1]$ y para el par cerebelo-médula espinal es de 3.34×10^{-6} para un intervalo $[4.77 \times 10^{-7}, 1]$. En el tercer nivel jerárquico, el p-value para el par cerebro-médula espinal es de 7.02×10^{-9} en un intervalo $[5.68 \times 10^{-14}, 1]$, para el par cerebro-cerebelo es de 1.73×10^{-7} en un intervalo $[4.55 \times 10^{-13}, 1]$ y para el par médula espinal-cerebelo es de 3.5×10^{-6} en un intervalo $[4.55 \times 10^{-13}, 1]$. Al estandarizar y comparar en cada nivel se observa que el par cerebro-cerebelo presenta la mayor probabilidad para el segundo nivel, mientras que en el tercer nivel jerárquico es el par cerebro-médula espinal. La (pseudo)mediana indica que en general los HPMS más afectados de los tres ChSqMs estudiados son los del experimento que emplea médula espinal porque el mayor efecto ocurre a su favor. Los HPMS menos afectados de los tres corresponden al experimento que usa cerebro como tejido. Acorde a los rankings diferenciales representados, los efectos se deben a valores elevados de dist que tienen signo negativo en los dos primeros grados jerárquicos, y se dan entre el ChSqM del experimento que emplea cerebro y los otros dos ChSqMs. Para el tercer grado estos valores de dist son más pequeños pero más numerosos. Existe un reparto asimétrico respecto a $\text{dist}=0$, pero conforme aumenta el grado jerárquico dicho reparto se vuelve considerablemente más simétrico para los tres pares. El par que tiene la distribución más simétrica en los tres grados jerárquicos es el par cerebelo-médula espinal.

5. Discusión de resultados:

En este apartado se exponen las consecuencias biológicas derivadas de los resultados obtenidos, representados en el apartado 4.

5.1. BCLAF1:

Los porcentajes amplios, dentro del rango entre [56%, 92%], de los picos anotados que se unen a regiones con característica genómica no intergénica (figuras 4, 5 y 6), indican que el HBGM utilizado incluye un porcentaje variado de regiones de interés funcional para BCLAF1 según los experimentos analizados, siendo el experimento ENCSR000BKH el que más porcentaje recoge.

Al analizar cómo se comporta BCLAF1 desde la perspectiva de la clase de características genómicas a las que se une, en ninguno de los tres ChSqMs se une por azar, sin embargo, se observan diferencias a este nivel, presentando preferencia clara por las clases promotora y *enhancer* en el primer ChSqM y por las clases intergénica e intrón en los otros dos. El comportamiento más parecido a este nivel se da entre los experimentos 2 y 3.

Si se analiza el comportamiento de BCLAF1 desde la perspectiva de los HGNMs con mayor gene score afectados para cada uno de los ChSqMs de este estudio, de forma cualitativa se aprecia que la clase de característica genómica que más influye en la puntuación de los genes seleccionados es la clase promotora seguida de la clase *enhancer* excepto para el tercer experimento, en el que la clase de característica genómica que más influye es la clase *enhancer* seguida de proximal. Al fijarse en los gráficos *heatmap*, que representan los rankings de diferencias, se observa que aparece una mayor similitud entre el par de los experimentos ENCSR000BKH-ENCSR492LTS. Al cuantificar los rankings diferenciales mediante su p-value, se obtiene que los HGNMs del experimento 1 son los que se diferencian más respecto a los de los otros dos experimentos, que aunque menos también se diferencian entre sí.

Realizando el análisis del comportamiento de BCLAF1 desde la perspectiva de los HPMS afectados para cada uno de los ChSqMs de este estudio, de forma cualitativa, se aprecia que para dichos ChSqMs el orden de prevalencia de la clase de característica genómica que más influye en la puntuación de ruta es prácticamente el mismo en los tres grados jerárquicos, destacando la clase *enhancer* y la clase promotora. Hay que destacar que en los tres ChSqMs de este estudio, los HPMS más genéricos (nivel uno), que se encuentran más afectados son relativas al metabolismo

de proteínas y de RNA, metabolismo y transcripción. En general los HPMS más afectados de los tres ChSqMs estudiados son los del primer experimento (ENCSR000BKH).

Analizado el comportamiento de la proteína BCLAF1 en los tres niveles se determina que a nivel de región es diferente respecto a las clases genómicas por las que tiene preferencia, sobre todo entre el experimento 1 y los otros dos. Al estudiar el comportamiento de la proteína respecto a los HGNMs afectados en el nivel de gen no se observan diferencias tan significativas. No obstante, en el *Pathway level* sí vuelven a haber diferencias significativas al comparar los HPMS afectados en los tres ChSqMs, siendo el del experimento 1 el más afectado.

La proteína BCLAF1 codifica un represor transcripcional cuya sobreexpresión induce apoptosis, por lo que lo esperado es que tenga un modelo similar al descrito siguiendo los datos obtenidos en el experimento ENCSR000BKH, que se une mayoritariamente a regiones de las clases genómicas *enhancer* y promotora. Además, para uno de estos dos experimentos (ENCSR729IYC), los autores han indicado que el experimento no muestra datos fiables debido a que se emplea un anticuerpo poco específico, por lo que a falta de realizar más experimentos se considera que el perfil del modelo 1 puede ser el más parecido al de la proteína real.

5.2. CTCF:

Los porcentajes amplios, dentro del rango entre [78%,79%], de los picos anotados que se unen a regiones con característica genómica no intergénica (figuras 12, 13 y 14), indica que el HBGM utilizado incluye un porcentaje alto de regiones de interés funcional para CTCF según los experimentos analizados, siendo los experimentos que emplean médula espinal y cerebelo los que más porcentaje recogen.

Al analizar cómo se comporta CTCF desde la perspectiva de la clase de características genómicas a las que se une, en ninguno de los tres ChSqMs se une por azar, presentando preferencia clara por las clases promotora y *enhancer*, en el modelo de cerebro, mientras que en el modelo de médula ósea y cerebelo el orden de las clases por las que la proteína presenta mayor preferencia es primero la clase promotora, seguido de la proximal y en tercer lugar la clase *enhancer*. El comportamiento más parecido a este nivel se da entre las secciones de médula espinal y cerebelo.

Si se analiza el comportamiento de CTCF desde la perspectiva de los HGNMs con mayor gene score afectados para cada uno de los ChSqMs de este estudio, de forma cualitativa se aprecia que la clase de característica genómica que más influye en la puntuación de gen es la clase promotora seguida de la clase proximal, excepto para el experimento que emplea cerebro como tejido, en el que la segunda clase de característica genómica que más influye es la clase *enhancer*. Al fijarse en los gráficos *heatmap*, que representan los rankings de diferencias, se observa que aparece una mayor similitud entre el par de los experimentos cerebelo-médula espinal. Al cuantificar los rankings diferenciales mediante su p-value, se obtiene que los HGNMs del experimento de médula espinal son los que más se diferencian respecto a los otros dos experimentos (especialmente el par médula espinal-cerebelo), que aunque menos también se diferencian entre sí.

Realizando el análisis del comportamiento de CTCF desde la perspectiva de los HPMS afectados para cada uno de los ChSqMs de este estudio, de forma cualitativa, se aprecia que para dichos ChSqMs el orden de prevalencia de la clase de característica genómica que más influye en la puntuación de ruta es prácticamente el mismo en los tres grados jerárquicos, destacando la clase promotora y la clase *enhancer*. Hay que destacar que en los tres ChSqMs de este estudio, los HPMS más genéricos (nivel uno), que se encuentran más afectados son relativas a la transducción de señales, metabolismo de proteínas, metabolismo y transcripción. En general los HPMS más afectados de los tres ChSqMs estudiados son los que derivan del experimento con médula espinal. (ENCSR000DSU).

Analizado el comportamiento de la proteína CTCF en los tres niveles se determina que a nivel de región en general coincide respecto a las clases genómicas por las que tiene preferencia, siendo los experimentos en médula espinal y cerebelo los más similares. Al estudiar el comportamiento de la proteína respecto a los HGNMs afectados en el nivel de gen no se observan diferencias tan significativas. No obstante, en el *Pathway level* sí vuelven a haber diferencias significativas al comparar los HPMS afectados en los tres ChSqMs, siendo de nuevo el par médula espinal-cerebelo el más similar.

6. Conclusiones:

6.1. BCLAF1:

A partir de los estudios analizados para la proteína BCLAF1, y teniendo en cuenta que algunos de los resultados mostrados son erróneos debido al fallo a nivel experimental, se puede determinar que:

- En la línea celular k-562, la proteína presenta patrones específicos de unión.
- Tiene preferencia por las regiones de característica genómica *enhancer* y promotora.
- Las rutas a las que más afecta, a nivel general, se relacionan con la transducción de señales, el metabolismo y la expresión génica.
- Se ha demostrado que no siempre se obtienen los resultados esperados partiendo de los mismos elementos (proteína, línea celular, condición biológica).
- Son necesarios más estudios en esta línea celular para poder obtener resultados más precisos.

6.2. CTCF:

A partir de los estudios analizados para la proteína CTCF se puede determinar que:

- En la línea celular astrocito, la proteína presenta patrones específicos de unión.
- Tiene preferencia por las regiones de característica genómica promotora en los tres tejidos empleados.
- En cerebro se observan patrones específicos de unión distintos de los observados en cerebelo y médula espinal.
- En cerebro la segunda característica genómica por la que CTCF muestra preferencia es la clase *enhancer*, en médula espinal y cerebelo es la clase proximal.
- Las rutas a las que más afecta, a nivel general, se relacionan con la transducción de señales, el metabolismo y la transcripción.

7. Bibliografía:

1. Park, P. Applications of next-generation sequencing: ChIP-seq: advantages and challenges of a maturing technology. *Nat. Rev. Genet.* **10**, 669 (2009).
2. Farnham, P. J. Insights from genomic profiling of transcription factors. *Nat. Rev. Genet.* **10**, 605–616 (2009).
3. Berger, S. L. The complex language of chromatin regulation during transcription. *Nature* **447**, 407–412 (2007).
4. Bernstein, B. E., Meissner, A. & Lander, E. S. The mammalian epigenome. *Cell* **128**, 669–681 (2007).
5. Almagro-Hernández, G.. Modelos de análisis semántico de información y conocimiento genético y genómico para el estudio de enfermedades genéticas y cáncer. (2020).
6. Pettersson, E., Lundberg, J. & Ahmadian, A. Generations of sequencing technologies. *Genomics* **93**, 105–111 (2009).
7. Shendure, J. & Ji, H. Next-generation DNA sequencing. *Nat. Biotechnol.* **26**, 1135–1145 (2008).
8. Solomon, M. J., Larsen, P. L. & Varshavsky, A. Mapping protein-DNA interactions in vivo with formaldehyde: evidence that histone H4 is retained on a highly transcribed gene. *Cell* **53**, 937–947 (1988).
9. Blat, Y. & Kleckner, N. Cohesins bind to preferential sites along yeast chromosome III, with differential regulation along arms versus the centric region. *Cell* **98**, 249–259 (1999).
10. Ren, B. *et al.* Genome-wide location and function of DNA binding proteins. *Science* **290**, 2306–2309 (2000).
11. Hillier, L. W. *et al.* Whole-genome sequencing and variant discovery in *C. elegans*. *Nat. Methods* **5**, 183–188 (2008).
12. Ley, T. J. *et al.* DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* **456**, 66–72 (2008).
13. Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316**, 1497–1502 (2007).
14. Nakato, R. & Shirahige, K. Recent advances in ChIP-seq analysis: from quality management to whole-genome annotation. *Brief. Bioinform.* **18**, 279–290 (2017).
15. Chèneby, J. *et al.* ReMap 2020: A database of regulatory regions from an integrative analysis of Human and Arabidopsis DNA-binding sequencing experiments. *Nucleic Acids Res.* **48**, D180–D188 (2020).
16. Howe, K. L. *et al.* Ensembl 2021. *Nucleic Acids Res.* **49**, D884–D891 (2021).
17. Bateman, A. *et al.* UniProt: The universal protein knowledgebase in 2021. *Nucleic Acids Res.* **49**, D480–D489 (2021).
18. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733–D745 (2016).
19. Gao, T. & Qian, J. EnhancerAtlas 2.0: An updated resource with enhancer annotation in 586 tissue/cell types across nine species. *Nucleic Acids Res.* **48**, D58–D64 (2020).
20. Suzuki, Y. *et al.* Identification and characterization of the potential promoter regions of 1031 kinds of human genes. *Genome Res.* **11**, 677–684 (2001).
21. Juven-Gershon, T. & Kadonaga, J. T. Regulation of gene expression via the core promoter and the basal transcriptional machinery. *Dev. Biol.* **339**, 225–229 (2010).
22. Venters, B. J. & Pugh, B. F. How eukaryotic genes are transcribed. *Crit. Rev. Biochem. Mol. Biol.* **44**, 117–141 (2009).
23. Richard, P. & Manley, J. L. Transcription termination by nuclear RNA polymerases. *Genes Dev.* **23**, 1247–1269 (2009).
24. Proudfoot, N. J. Ending the message: Poly(A) signals then and now. *Genes Dev.* **25**, 1770–1782 (2011).
25. Stephens, R. M. & Schneider, T. D. Features of spliceosome evolution and function inferred from an analysis of the information at human splice sites. *J. Mol. Biol.* **228**, 1124–1136 (1992).
26. Zerbino, D. R. *et al.* Ensembl 2018. *Nucleic Acids Res.* **46**, D754–D761 (2018).
27. Almagro-Hernández, G., García-Sánchez, F., de la Morena-Barrio, M. E., Corral, J. & Fernández-Breis, J. T. Angel: Towards a Multi-level Method for the Analysis of Variants in Individual Genomes BT - Bioinformatics and Biomedical Engineering. in (eds. Ortuño, F. & Rojas, I.) 47–58 (Springer International Publishing, 2016).
28. Finke, W., Rachimow, C. & Pfützner, B. Untersuchungen zu Wasserdargebot und Wasserverfügbarkeit im Ballungsraum Berlin im Rahmen des Verbundprojektes GLOWA Elbe. *Hydrol. und Wasserbewirtschaftung* **48**, 2–11 (2004).
29. Jassal, B. *et al.* The reactome pathway knowledgebase. *Nucleic Acids Res.* **48**, D498–D503 (2020).