Universitat Oberta de Catalunya

uoc.edu

Uso de paquetes de R/Bioconductor para análisis funcional de datos ChIP-Seq

Aarón Gallego Crespo

Máster en Bioinformática y Bioestadística

Área 4 – Subárea 1

Helena Brunel

Antoni Pérez Navarro

Fecha de entrega: 24/12/2021

Ficha del trabajo final

Título del trabajo	Uso de paquetes de R/Bioconductor para análisis funcional de datos ChIP-Seq
Nombre del autor	Aarón Gallego Crespo
Nombre del consultor	Helena Brunel
Nombre del PRA	Antoni Pérez Navarro
Fecha de entrega (mm/aaaa)	12/2021
Titulación	Máster en Bioinformática y Bioestadística
Área del trabajo final	Área 4- Subárea 1
Idioma del trabajo	Castellano
Número de créditos	15 ECTS
Palabras clave	ChIP-Seq, R/Bioconductor, pipeline, factor de transcripción, histona

Resumen

ChIP-Seq es un método de secuenciación masiva para identificar sitios de unión proteínas-ADN. Su aplicación para el estudio del perfil de unión de histonas y factores de transcripción en el genoma humano permite conocer la relevancia de estas proteínas en procesos como la diferenciación celular y la patogenia de enfermedades. El proyecto Bioconductor ofrece paquetes de software en R para el análisis de datos ChIP-Seq, suponiendo una alternativa de fácil acceso, bajo coste computacional y altamente versátil frente a otras herramientas.

En este trabajo se propuso como objetivos: 1) La búsqueda y selección de paquetes de R/Bioconductor; 2) La búsqueda y selección de conjuntos de datos ideales para aplicar los paquetes; 3) El diseño de un pipeline de análisis de datos ChIP-Seq. Este pipeline se diseñó para realizar la anotación funcional e identificación de motivos de ADN a partir de datos procedentes de tres escenarios experimentales típicos: 1) ChIP-Seq en diferentes replicas biológicas; 2) ChIP-Seq en dos condiciones diferentes; 3) ChIP-Seq en diferentes líneas celulares.

El pipeline resultante combina los paquetes ChIPSeeker y ChIPpeakAnno; rGREAT para el análisis de motivos, entre otros paquetes. Los resultados indican que: 1) ChIPpeakAnno es más flexible y específico para anotación funcional de los picos; 2) ChIPseeker ofrece mayor variedad de opciones para visualizar datos; 3) Ambos paquetes se complementan bien en sus fortalezas y debilidades, siendo más útiles juntos que separados; 4) rGREAT cumple los requerimientos básicos para el análisis de motivos, pero tiene un catálogo de funciones limitado.

Abstract

ChIP-Seq is a high-throughput sequencing method for identifying DNA-protein binding sites. Its main application for studying histones and transcription factors binding profiles in the human genome allows us to know their relevance in processes such us cell differentiation and diseases pathogenesis. The Bioconductor Project offers R software packages for ChIP-Seq data analysis, being an accessible, low computational cost and versatile alternative against other tools.

In this study, the main objective is to: 1) Search and select R/Bioconductor packages; 2) Search and select ideal datasets to apply the packages; 3) Design a ChIP-Seq data analysis pipeline. This pipeline was designed for functional annotation and DNA motifs discovery applied to data from three typical experimental

backgrounds: 1) ChIP-Seq in different biological replicates; 2) ChIP-Seq in two different experimental conditions; 3) ChIP-Seq in different cell lines.

The final pipeline combines the packages ChIPpeakAnno and ChIPseeker; rGREAT for motif analysis, among other packages. Our results indicate that: 1) ChIPpeakAnno is more flexible and specific for functional peak annotation; 2) ChIPseeker offer a bigger variety of options for data visualization; 3) Both packages complement each other well in their strengths and weaknesses, being more useful together than separately; 4) rGREAT meets the basic requirements for motif analysis, but has a limited functionalities set

Índice

Índice de figuras y tablas

1. Introducción

1.1 Contexto y justificación del Trabajo

El ChIP-Seq es un método de identificación de interacciones proteína-ADN, usualmente utilizado para histonas y/o factores de transcripción. Ambos tipos de proteínas regulan una parte importante de la expresión génica, por tanto, las aplicaciones del ChIP-Seq en el estudio de estos mecanismos reguladores permiten estudiar potenciales biomarcadores de enfermedades.

El análisis computacional de datos ChIP-Seq requiere del uso de una gran cantidad de herramientas, con diferentes requerimientos técnicos y entornos de desarrollo. Los sistemas operativos Unix/Linux son los más utilizados para el análisis de datos, seguidos de MAC OS y Windows OS. El proyecto Bioconductor ofrece una extensa variedad de paquetes de software en R, de código abierto y aplicables desde cualquiera de estos tres sistemas operativos. Están incluidos paquetes dedicados a análisis de datos ChIP-Seq, de fácil uso y relativamente bajo coste computacional. Esto pone en valor la necesidad de comparar estos paquetes para saber aprovechar al máximo sus diferentes funcionalidades en el diseño de pipelines personalizables y dinámicos, aplicables en los diferentes escenarios experimentales y datos de interés.

1.2 Objetivos

En este trabajo se ha propuesto la siguiente serie de objetivos.

Objetivos generales:

1. Búsqueda y selección de paquetes de R/Bioconductor.
2. Búsqueda y selección de conjuntos de datos.
3. Diseño de un pipeline de análisis funcional de datos ChIP-Seq

Objetivos específicos:

1.a) Revisar la bibliografía existente sobre métodos de análisis ChIP-Seq con paquetes R/Bioconductor.

1.b) Seleccionar paquetes centrados en *peak annotation* y *motif analysis*, compatibles con un entorno de desarrollo RStudio y sistema operativo Windows 10.

2.a) Revisar las bases de datos epigenéticos:

- ENCODE Portal
- Gene Expression Omnibus (GEO)
- Cistrome

2.b) Selección de datos de picos o sitios de unión de proteínas al ADN ya alineados con un genoma de referencia.

3.a) El pipeline debe ser aplicable a tres escenarios experimentales típicos en ChIP-Seq:

- ChIP-Seq de un factor de transcripción/histona en diferentes replicas biológicas de una misma línea celular.
- ChIP-Seq de un factor de transcripción/histona en una línea celular sometida a dos condiciones diferentes.
- ChIP-Seq de un factor de transcripción/histona en diferentes líneas celulares.

3.b) El pipeline debe contener los siguientes análisis:

- Estudiar la distribución de los picos a lo largo del genoma.
- Identificar el porcentaje de picos solapantes con elementos genómicos, como promotores, exones, intrones etc.
- Identificar motivos de ADN y rutas biológicas más frecuentes solapantes con los picos

1.3 Enfoque y método seguido

La estrategia seguida para llevar a cabo los objetivos planteados ha sido dividir el trabajo en tres fases. Una primera fase de búsqueda bibliográfica, de conjuntos de datos y exploración general de las diferentes funcionalidades que ofrecen los paquetes seleccionados, que se corresponde con lo descrito en la PEC 2. Una segunda fase para el diseño del pipeline aplicado a datos procedentes de tres contextos experimentales diferentes, que se corresponde con el contenido de la PEC 3. En esta fase se ha buscado aprovechar al máximo sus funcionalidades de *peak annotation* y *motif anaysis*. Finalmente, una tercera fase para la redacción de la memoria del trabajo, correspondiéndose con la PEC 4.

1.4 Planificación

A partir del enfoque general del trabajo se ha establecido el siguiente calendario con las diferentes fases del TFM:

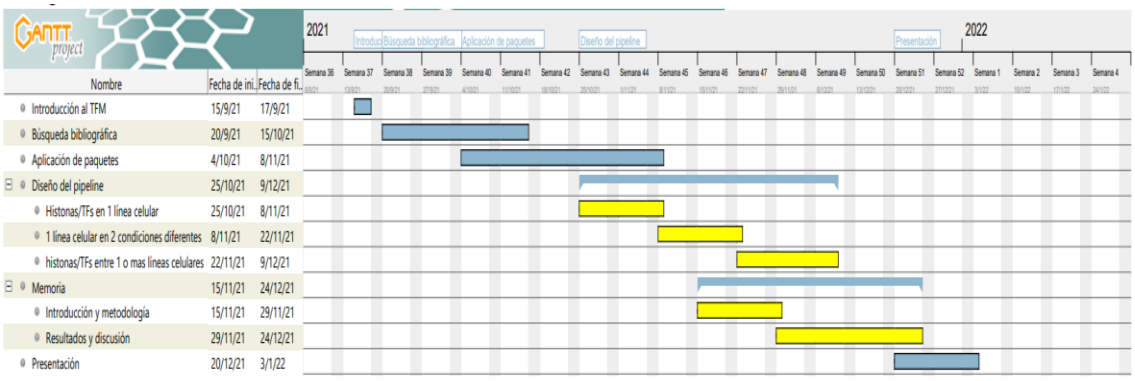


Figura 1. Calendario de planificación del TFM.

Los hitos más relevantes del trabajo han sido cuando:

1) Todos los paquetes seleccionados se han aplicado y se ha comprobado que no existen incompatibilidades de software, errores en su instalación, y se ejecutan correctamente. Este hito se cumplió el 8/11/21.

2) Se ha completado el pipeline en el que se apliquen los métodos a los datos seleccionados de manera exitosa, el 9/12/21.

3) Se ha redactado por completo la memoria del trabajo, el 24/12/21.

1.5 Breve sumario de productos obtenidos

Un pipeline escrito en R y guardado en un archivo Markdown (.Rmd), donde se realiza la anotación funcional y análisis de motivos de datos ChIP-Seq obtenidos en tres contextos experimentales diferentes.

1.6 Breve descripción de otros capítulos de la memoria

Capítulo 2: Breve descripción del estado actual de la cuestión. Qué potencial tiene el ChIP-Seq en la identificación de biomarcadores de enfermedades y qué herramientas ofrece Bioconductor para el análisis de datos ChIP-Seq.

Capítulo 3: Metodología del trabajo. Bases de datos consultadas y criterios para la selección de datos.

Capítulo 4: Resultados. Exposición de los análisis realizados mediante el pipeline, gráficos generados y tablas de resultados.

Capítulo 5: Discusión. Comentario sobre los resultados obtenidos y las limitaciones del pipeline.

Capítulo 6: Conclusiones. Recapitulación del trabajo con las conclusiones obtenidas.

Capítulo 7: Bibliografía. Bibliografía consultada en este trabajo.

Capítulo 8: Anexos. Lista de anexos a la memoria de este trabajo.

2. Estado del arte

El método ChIP-Seq consiste en la secuenciación masiva de aquellas regiones del genoma donde hay interacción proteína-ADN, obtenidas por inmunoprecipitación de la cromatina [1]. Su aplicación principalmente es la identificación de sitios de unión de proteínas en el ADN, sobre todo histonas y factores de transcripción [2]. Estas proteínas juegan un papel esencial en la expresión génica, ya sea promoviéndola o inhibiéndola, interaccionando con el ADN y con otras proteínas conformando una maquinaria celular compleja y sofisticada. Por lo tanto, ChIP-Seq permite caracterizar las redes reguladoras de la expresión génica al identificar regiones promotoras o sitios de inicio de la transcripción (TSS) y enhancers, regiones potenciadoras de la expresión génica [2, 3].

ChIP-Seq ha demostrado ser relevante para conocer cómo estos elementos determinan procesos como la identidad, desarrollo y diferenciación celular, incluso la patogenia de enfermedades [4, 5]. Entre las aplicaciones del análisis computacional de datos ChIP-Seq encontramos la predicción de la expresión génica a gran escala o la identificación de posibles biomarcadores útiles en el diagnóstico y pronóstico de enfermedades [6, 7, 8, 9]. Algunos ejemplos identificados gracias a ChIP-Seq son: 1) Alteraciones en los sitios de unión de la histona K3K27me3 y del Receptor de Estrógenos (ER) están asociadas con el desarrollo del cáncer de mama [7]; el SNP de riesgo asociado a cáncer de próstata rs339331 se sitúa en un enhancer intrónico, alterando la unión de las histonas H3K27ac y H3K4me1 [9]; 3) el SNP rs17293632, asociado a la enfermedad autoinmune de Crohn, también presenta un alelo que altera un sitio de unión para el factor de transcripción AP-1 [9].

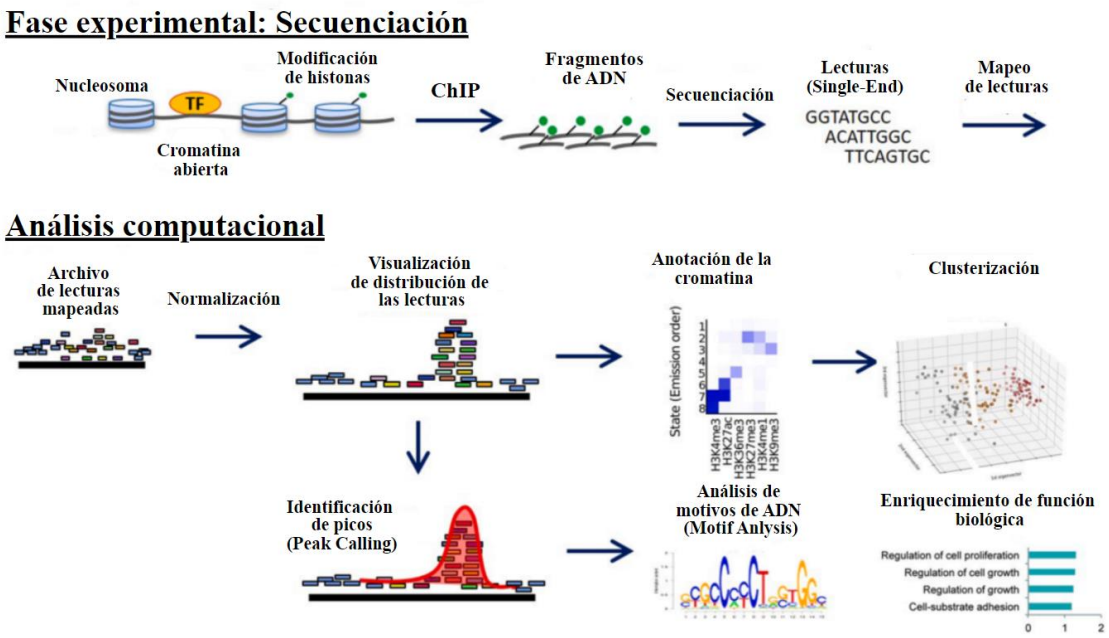


Figura 2. Workflow del método ChIP-Seq. Extraído y modificado a partir de [6].

El análisis de datos ChIP-Seq estándar se resume en el worklfow de la Figura 2. Partiendo de las lecturas obtenidas por la secuenciación, hay que alinearlas contra un genoma de referencia (*Read mapping* en la figura) e identificar donde se encuentran los picos o sitios de unión proteína-ADN (paso comúnmente conocido como *Peak calling*). Una vez realizados los controles de calidad necesarios sobre los picos identificados, se puede realizar diferentes análisis en función de las necesidades. Lo más común es realizar la anotación funcional de los picos (*Peak annotation*) y el análisis de los motivos de ADN presentes en los picos (*Motif analysis*). El primero se basa en identificar a qué distancia se encuentran los picos de los elementos genómicos, normalmente los TSS, y que porcentaje de solapamiento hay entre picos y genes. También es muy común analizar el enriquecimiento funcional de los genes solapantes para comprobar que rutas biológicas están reguladas. Por otra parte, el *motif analysis* suele consistir en la identificación *de novo* o de motivos ya conocidos, de qué factores de transcripción pueden ser diana y su distribución en el genoma.

El análisis ChIP-Seq puede conllevar un alto coste computacional, ya que normalmente se trabaja con grandes cantidades de datos, siendo

determinados pasos, como el alineamiento con el genoma humano, los que requieren de herramientas potentes. Los paquetes existentes dentro del proyecto Bioconductor suponen una alternativa de fácil acceso, código abierto, bajo coste computacional y altamente versátil para los investigadores, con la posibilidad de integrar análisis estadísticos y gráficos de datos genómicos con metadatos (ej.: bibliografía desde PubMed, anotación funcional de Entrez genes etc.) [10, 11]. Bioconductor ofrece hasta 90 paquetes dedicados a datos ChIP-Seq como, por ejemplo, ChIPQC (dedicado a control de calidad) [12] o DiffBind (dedicado al análisis diferencial de los picos) [13].

3. Metodología

En primer lugar, para la búsqueda de la bibliografía existente se ha tenido en cuenta la base de datos PubMed, del *National Center for Biotechnology Information* (NCBI) (https://pubmed.ncbi.nlm.nih.gov/), mediante la búsqueda de palabras clave como: "ChIP-Seq Analysis", "R", "Bioconductor", "workflow", "pipeline". Para la selección de paquetes también se ha consultado la web oficial del proyecto Bioconductor (https://bioconductor.org/), donde es posible acceder a los manuales de referencia, guías de usuario, al código de los paquetes, todas las versiones, entre otra documentación sobre cualquier paquete de interés.

Todos los paquetes utilizados se han instalado en su última versión y son compatibles con versiones de R ≥ 3.5.0, evitando posibles incompatibilidades que pudieran surgir entre paquetes. En este trabajo se han utilizado las versiones de R 4.1.2, BiocManager 1.30.16 y RStudio 1.3.1093. También se han tenido en cuenta paquetes de R del repositorio *Comprehensive R Archive Network* (CRAN) (https://cran.r-project.org/). Determinadas funciones de estos paquetes nos permiten acceder indirectamente también a bases de datos como: Entrez Gene (https://www.ncbi.nlm.nih.gov/gene) y UCSC (https://genome.ucsc.edu/) para anotación funcional de genes y promotores del genoma humano, y MSigDB (https://www.gsea-msigdb.org/gsea/msigdb/) para el análisis de motivos de ADN.

Por otra parte, para la búsqueda de conjuntos de datos apropiados para los siguientes análisis se han consultado las siguientes bases de datos: GEO (https://www.ncbi.nlm.nih.gov/geo/) [14, 15] y ENCODE (https://www.encodeproject.org/) [16, 17]; el principal buscador de conjuntos de datos ha sido Cistrome (http://cistrome.org/) [18, 19].

El criterio principal para seleccionar datos ChIP-Seq ha sido escoger aquellos con información de los picos de factores de transcripción o histonas

para diferentes líneas celulares humanas, siempre en formato BED, NARROWPEAK o BROADPEAK. Los dos primeros son formatos típicos de factores de transcripción, ya que sus sitios de unión suelen abarcar regiones más concretas, mientras que el tercero es un formato típico de histonas, cuyos sitios de unión suelen ser regiones más extensas. Estos datos no contienen anotación funcional y la información básica que contienen son las coordenadas genómicas (cromosoma, inicio y final de la región del sitio de unión), nombre de cada pico (generado por el software de *Peak Calling*, usualmente MACS), y un score de calidad (Figura 3).

chr1	868170	869141	MACS_peak_1	130.24
chr1	1008750	1009999	MACS_peak_2	1348.98
chr1	1014518	1016032	MACS_peak_3	233.60
chr1	1607591	1608718	MACS_peak_4	371.14
chr1	1670390	1671432	MACS_peak_5	856.53
chr1	1690019	1690917	MACS_peak_6	250.45
chr1	2179696	2180564	MACS_peak_7	76.30
chr1	2335607	2336540	MACS_peak_8	94.82
chr1	2341997	2342853	MACS_peak_9	115.69
chr1	2583334	2588147	MACS_peak_10	89.72
chr1	2619000	2621815	MACS_peak_11	90.97
chr1	2622562	2623910	MACS_peak_12	84.32
chr1	3585097	3585903	MACS_peak_13	130.71
chr1	3625250	3625992	MACS_peak_14	73.57
chr1	3773708	3774406	MACS_peak_15	76.76
chr1	6217626	6218552	MACS_peak_16	175.77
chr1	6671544	6672365	MACS_peak_17	111.33

Figura 3. Formato central de datos ChIP-Seq.

4. Resultados

4.1 Resumen de paquetes seleccionados

Los dos paquetes más importantes de entre los seleccionados son ChIPpeakAnno [versión 3.28.0] [20] y ChIPseeker [versión 1.30.2] [21], dedicados a la anotación funcional de picos. En ambos encontramos funciones que nos permiten estudiar la localización y distribución de los picos a lo largo del genoma, porcentaje de solapamientos con diferentes elementos genómicos (TSS, exones, intrones etc.) y distancia media a estos elementos.

ChIPpeakAnno incluye un mayor número de funciones, destacando las diseñadas para obtener conjuntos de datos ya precargados (ej.: *TSS.human.GRCh38,* con anotación de TSS en genoma humano) o para enriquecimiento de anotación funcional (ej.: *getEnrichedGO* y *getEnrichedPATH*). Chipseeker no incluye datos precargados, aunque sí destaca por un mayor número de paquetes dedicados a visualización de datos (ej.: *vennpie* para diagramas de Venn o *plotAnnoBar* para diagramas de barras). Ambos paquetes incluyen funciones para transformar los datos de partida a objetos tipo GRanges, paso necesario para cualquier análisis ChIP-Seq.

ChIPpeakAnno		ChIPseeker	
Clase	**Función**	**Clase**	**Función**
Anotación	*annoPeaks*	**Anotación**	*annotatePeaks*
	annotatePeakInBatch		*seq2gene*
	assignChromosomeRegion		*getBioRegion*
	bdp		*getGeneAnno*
	binOverFeature		*getGenomicAnnotation*
	genomicElementDistribution		*getNearestFeatureIndicesAndDistances*
	genomicElementUpSetR		*getPromoters*
Gráficos	*makeVennDiagram*	**Gráficos**	*covplot*
	featureAlignedDistribution		*peakHeatmap*
	featureAlignedExtendSignal		*plotPeakProf*
	featureAlignedHeatmap		*plotPeakProf2*
	featureAlignedSignal		*plotAvgProf*

	pie1		*plotAvgProf2*
	metagenePlot		*plotAnnoBar*
			plotAnnoPie
			tagHeatmap
			upsetplot
			vennpie
			vennplot
			vennplot.peakfile
Miscelánea	*condenseMatrixByColnames*	**Miscelánea**	*readPeakFile*
	convert2EntrezID		*as.GRanges*
	countPatternInSeqs		*downloadGEObedFiles*
	egOrgMap		*downloadGSMbedFiles*
	estFragmentLength		*dropAnno*
	estLibSize		*enrichAnnoOverlap*
	findEnhancers		*enrichPeakOverlap*
	findOverlappingPeaks		*getGEOgenomeVersion*
	findOverlapsOfPeaks		*getGEOInfo*
	getAllPeakSequence		*getGEOspecies*
	getAnnotation		*getSampleFiles*
	getEnrichedGO		*getTagMatrix*
	getEnrichedPATH		*mclapply*
	getGO		*overlap*
	getVennCounts		
	IDRfilter		
	mergePlusMinusPeaks		
	oligoFrequency		
	oligoSummary		
	peakPermTest		
	peaksNearBDP		
	preparePool		
	reCenterPeaks		
	summarizeOverlapsByBins		
	summarizePatternInPeaks		
	tileCount		
	tileGRanges		
	toGRanges		
	translatePattern		
	write2FASTA		
	xget		
Conjuntos de datos	*enrichedGO*		

	annotatePeaks		
	ExonPlusUtr.human.GRCh37		
	HOT.spots		
	myPeakList		
	Peaks.Ste12.Replicate1		
	Peaks.Ste12.Replicate2		
	Peaks.Ste12.Replicate3		
	TSS.human.GRCh37		
	TSS.human.GRCh38		
	TSS.human.NCBI36		
	TSS.mouse.GRCm38		
	TSS.mouse.NCBIM37		
	TSS.rat.RGSC3.4		
	TSS.rat.Rnor_5.0		
	TSS.zebrafish.Zv8		
	TSS.zebrafish.Zv9		
	wgEncodeTfbsV3		

Tabla 1. Resumen de funciones de los paquetes ChIPpeakAnno y ChIPseeker.

También se han seleccionado otros paquetes igualmente necesarios para todo el análisis:

Paquete	Versión	Descripción	Referencia
clusterProfiler	4.2.0	Visualización de enriquecimiento funcional	[22, 23]
DBI	1.1.1	Fácil acceso a bases de datos relacionales desde R	[24]
GenomicRanges	1.44.0	Fácil manipulación de datos genómicos y anotación funcional mediante objetos GRanges	[25]
org.Hs.eg.db	3.14.0	Anotación del genoma humano, basado en Gene IDs de Entrez Gene	[26]
pRoloc	1.34.0	Obtención de GO Terms de los genes	[27]
reactome.db	1.77.0	Análisis de enriquecimiento funcional vía acceso a la base de datos Reactome Pathways	[28]
ReactomePA	1.38.0		[29]
rGREAT	1.26.0	Búsqueda e identificación de motivos de ADN	[30]

TxDb.Hsapiens.UCSC.hg18.knownGene	3.2.2	Anotación del genoma humano, aportada por la UCSC (versiones NCBI36/hg18 y hg19/GRCh37 respectivamente)	[31]
TxDb.Hsapiens.UCSC.hg19.knownGene	3.2.2		[32]

Tabla 2. Resumen de paquetes R/Bioconductor seleccionados. En morado: paquetes dedicados a análisis de motivos de ADN; En verde: paquetes auxiliares necesarios para los análisis.

4.2 Resumen de datos seleccionados

Dividiendo el pipeline en tres análisis correspondientes a cada uno de los escenarios experimentales previstos en los objetivos específicos 3.a), los datos seleccionados para cada uno han sido los siguientes.

En el análisis 1 se han utilizado los ficheros en formato BED: GSM486322, GSM486323 y GSM486324 (GEO Dataset: GSE19485; disponible en https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE19485). Estos datos corresponden al perfil de unión del factor de transcripción NF-KappaB en una línea celular linfoblastoide humana (células mononucleares de sangre periférica transformadas mediante infección con el Virus Epstein-Barr) [33].

En el análisis 2 se han utilizado datos en formato BED de los ficheros GSM1003718 y GSM1003732 (GEO Dataset: GSE40867; disponible en https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE40867). Ambos ficheros corresponden al perfil de unión de la histona H3K4me3 en muestras de tumor primario de cáncer de mama ER+ sometidas a dos condiciones diferentes: alta y baja resistencia a tratamiento con inhibidores de la enzima aromatasa. Los tumores de alta resistencia al tratamiento mostraron un peor pronóstico en la enfermedad y los de baja resistencia, mejor pronóstico [34].

Finalmente, en el análisis 3 se han utilizado datos procedentes de los ficheros en formato NARROWPEAK GSM935376, GSM935310 y GSM935542

(GEO Dataset: GSE31477; disponible en https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE31477). Estos datos corresponden al perfil de unión del factor de transcripción SMC3 en las líneas celulares GM12878 (Linfocitos B transformados con Epstein-Varr), K562 (Granulocitos inmortalizados de leucemia mieloide) y HepG2 (células inmortalizadas de hepatocarcinoma) respectivamente [35].

4.3 Diseño general del pipeline

El pipeline está diseñado para ser aplicado a los tres escenarios experimentales propuestos en los objetivos generales 3.a). En los tres casos se aplica el mismo protocolo: primero, se realiza una exploración de los datos comparando el número de picos entre los diferentes conjuntos de datos; segundo, se seleccionan los picos comunes y se calculan las distancias picos-TSS; tercero, se realiza la anotación funcional de las regiones genómicas de los picos. Finalmente, se realiza un análisis de enriquecimiento de los datos y análisis de motivos de ADN. Los paquetes usados en cada etapa vienen indicados a la derecha; en morado los paquetes dedicados a análisis de anotación funcional o de análisis de motivos; en verde, los paquetes auxiliares necesarios para los análisis y el conjunto del pipeline (Figura 4).

Este pipeline está pensado para ser lo más reproducible posible para cualquier investigador que quiera ejecutarlo. Para ello se ha establecido una serie de parámetros al comienzo del documento para los archivos ChIP-Seq y las distancias TSS-picos o gen-picos que queramos usar a lo largo del análisis (Figura 5). Así, quien quiera editar estos parámetros solo tiene que hacerlo una vez al inicio sin tocar el resto del código. En este pipeline de ejemplo, todos nuestros archivos se encuentran en una carpeta *data* y trabajamos con

distancias de 0.5 kb upstream y 1 kb downstream de los TSS, y hasta 2 kb downstream de los genes.

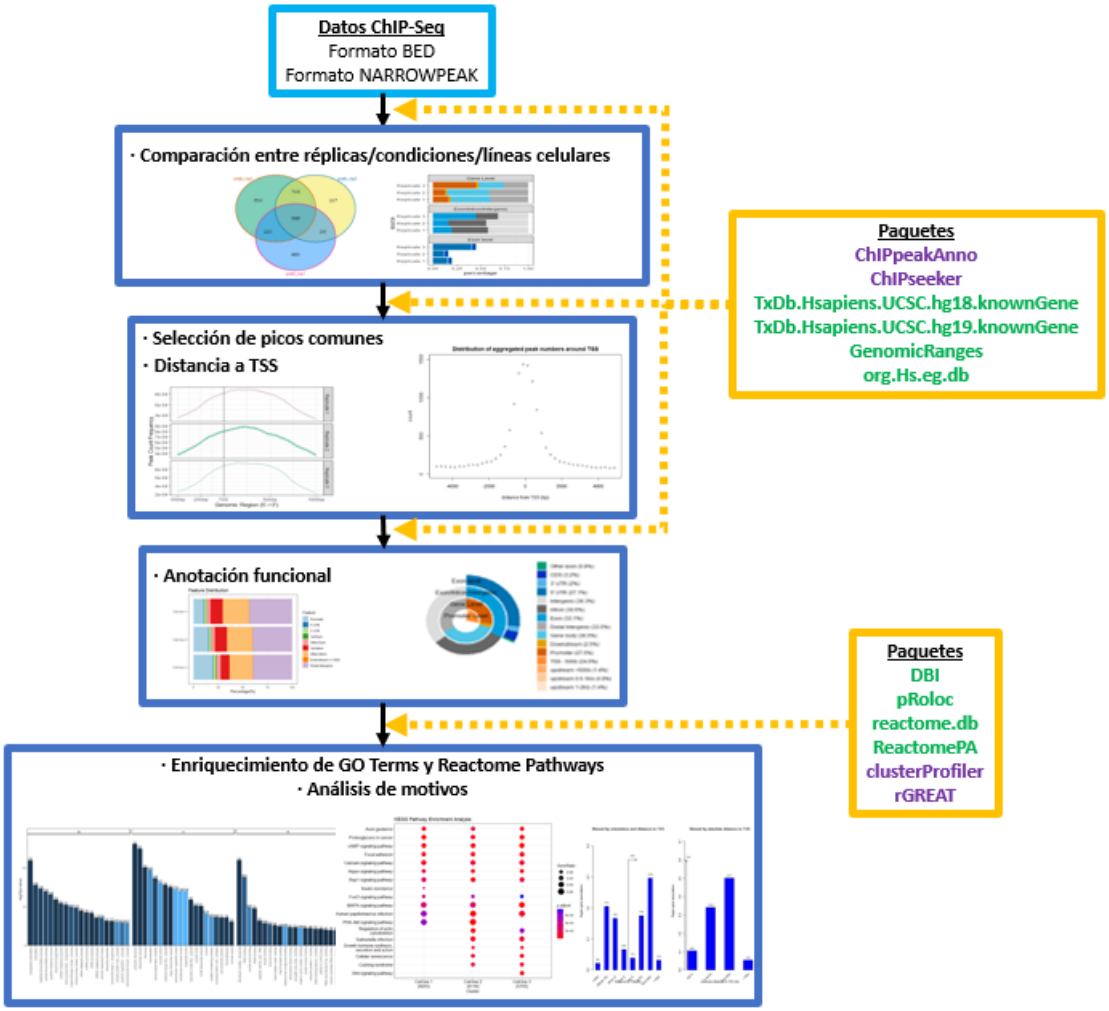


Figura 4. Resumen del pipeline de análisis de datos ChIP-Seq

```
params:
  replicate_1: data/NFKB_replicate1.bed
  replicate_2: data/NFKB_replicate2.bed
  replicate_3: data/NFKB_replicate3.bed
  condition_1: data/H3K4me3_good_outcome_4.bed
  condition_2: data/H3K4me3_poor_outcome_4.bed
  cell_line_1: data/SMC3_HepG2_hg19.narrowPeak
  cell_line_2: data/SMC3_GM12878_hg19.narrowPeak
  cell_line_3: data/SMC3_K562_hg19.narrowPeak
  prom_down: 1000
  prom_up: 500
  gene_down: 2000
```

Figura 5. Parámetros del pipeline

4.4 Análisis 1: Réplicas biológicas

Aplicando los paquetes directamente sobre los datos de partida, podemos obtener una visión global de estos. Podemos crear diagramas de Venn (mediante la función *makeVennDiagram*) y de barras (mediante *genomicElementDistribution*) fáciles de entender e igualmente informativos. En este análisis, realizado con ChIPpeakAnno, podemos comprobar las diferencias en el número de picos entre las réplicas biológicas y cuántos son compartidos (Figura 6). También, dónde se encuentran en el genoma respecto a diferentes

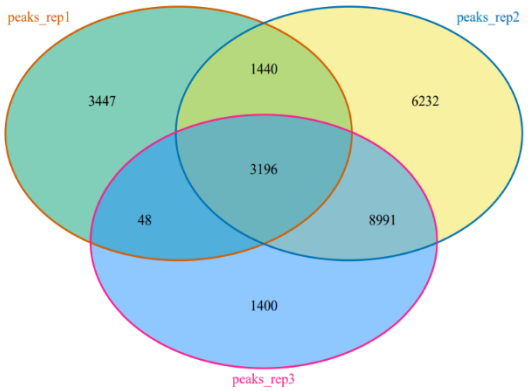


Figura 6. Diagrama de Venn entre réplicas biológicas

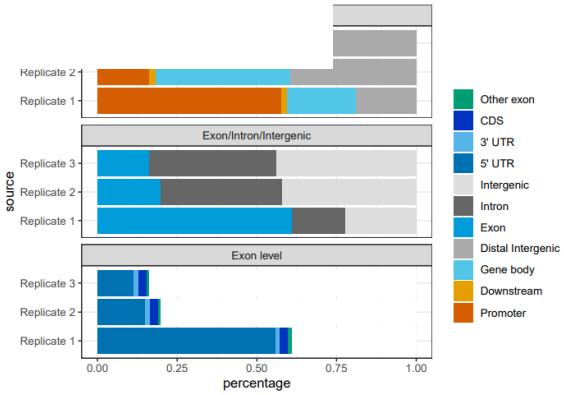
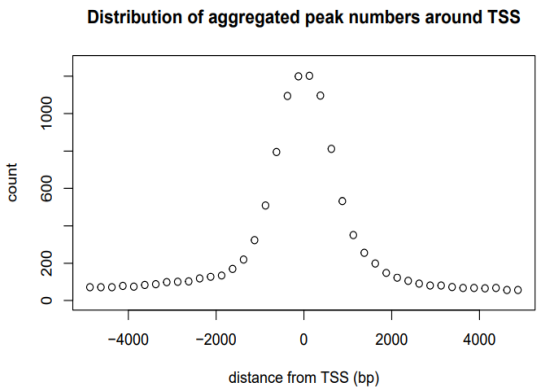


Figura 7. Solapamiento picos-elementos genómicos entre réplicas biológicas



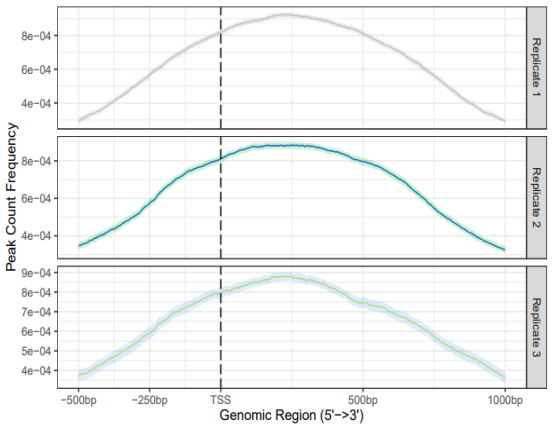


Figura 8. Distancia picos-TSS de picos comunes entre réplicas biológicas (ChIPpeakAnno, izquierda) y de cada réplica biológica (ChIPseeker, derecha)

elementos genómicos (promotores,
exones, intrones, regiones intergénicas) (Figura 7). Esto ya supone un primer paso para la anotación funcional, permitiendo comparar rápidamente las réplicas a nivel exónico, génico e intergénico. En este ejemplo, podemos destacar que la réplica 1 tiene un porcentaje de picos en promotores, exones y regiones 5'-UTR muy superior al de las otras réplicas, seguramente debido a que es la réplica con el menor número total de picos y existe un cierto sesgo. Por otra parte, un porcentaje alto de los picos totales se encuentran en regiones intergénicas, siendo esto típico debido a que los enhancers frecuentemente se encuentran en estas regiones (Figura 7).

Otro análisis preliminar típico es analizar la distancia de los picos a los TSS o promotores. ChIPpeakAnno no permite generar una figura con los gráficos de las tres réplicas, como si permite ChIPseeker. Tras seleccionar solo los picos comunes a las tres réplicas y calcular las distancias (mediante *binOverFeature* (ChIPpeakAnno) y *plotAvgProf* (ChIPseeker)) obtenemos la Figura 8 (izquierda), donde observamos que la distribución de los picos se ajusta a una curva normal entorno a los TSS. En cambio, en la figura 8 (derecha) obtenemos la distribución de todos los picos, observando que se encuentran entorno a los TSS, abarcando un área de hasta 5-7.5 kb en dirección 3' de los TSS. Esto concuerda con lo dicho previamente que un gran porcentaje de los picos situados cerca de promotores se encuentran también en exones. Además, en esta figura podemos incluir una estimación de las frecuencias al 95 % de confianza, indicado con una sombra alrededor de las líneas.

Sobre los picos comunes a las tres réplicas también podemos extraer su localización respecto a elementos genómicos (Figura 9). El gráfico tipo tarta es incluso más informativo, aportando porcentajes

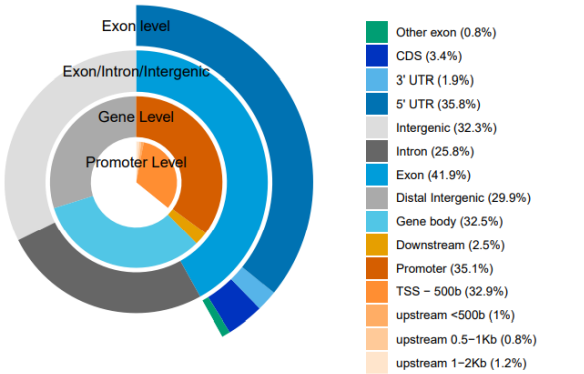


Figura 9. Solapamiento picos-elementos genómicos en picos comunes

y, si queremos, podemos
configurar una capa extra de información a nivel promotor. Podemos describir nosotros en el código de la función *genomicElementDistribution* en qué upstream de los promotores queremos estudiar el porcentaje de picos. Esto puede ser bastante útil ya que, si bien la longitud estándar de un promotor en el genoma humano es de 100-1000 pb upstream respecto a los TSS, esta longitud es muy variable, y nos puede interesar analizar con detalle la distribución concreta de los picos. Un 35.10 % de los picos comunes entre las réplicas se encuentran en promotores, y un 32.90 % en las primeras 500 pb downstream desde el TSS.

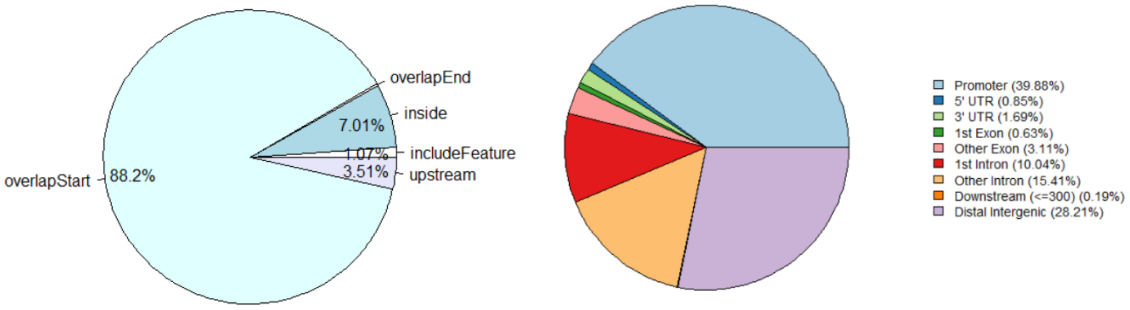


Figura 10. Anotación funcional de picos comunes (ChIPpeakAnno, izquierda; ChIPseeker, derecha)

En cuanto a la anotación funcional, con ChIPpeakAnno (mediante *annotatePeakInBatch*) podemos obtener la gráfica 10 (izquierda), la cual nos indica el porcentaje de picos solapantes con promotores de entre los picos comúnes a las tres réplicas. Nos divide los solapamientos en overlapStart (solapamiento al comienzo del promotor), overlapEnd (solapamiento al final del promotor), inside (pico totalmente dentro del promotor), includeFeature (promotor totalmente dentro del pico) y pico en posición upstream al promotor. En la figura 10 (derecha), generada con ChIPseeker (mediante *annotatePeak*), obtenemos una anotación más general ya que este paquete incluye opciones más limitadas para hacer nuestra búsqueda más selectiva. Con ChIPpeakAnno es posible buscar anotación específica de promotores bidireccionales mediante el argumento *output = "nearestBiDirectionalPromoters"* (un promotor capaz de iniciar la transcripción de dos genes diferentes), algo que no es posible en ChIPseeker.

En cuanto al análisis de enriquecimiento funcional, ChIPpeakAnno integra funciones para ello (*getEnrichedGO* y *getEnrichedPATH*, así como *enrichmentPlot* para visualización) con las que podemos generar los resultados de la Figura 11. Mediante una consulta a la base de datos Entrez Gene podemos consultar las funciones biológicas más estadísticamente significativas de los genes solapantes con los picos. Este análisis de Gene Ontology (GO) Terms nos aporta información sobre qué proceso biológico (bp), componente celular (cc) y función molecular (mf) desempeñan las proteínas codificadas por estos genes. Idénticamente, podemos extraer las rutas biológicas más significativas mediante una consulta a la base de datos Reactome Pathways.

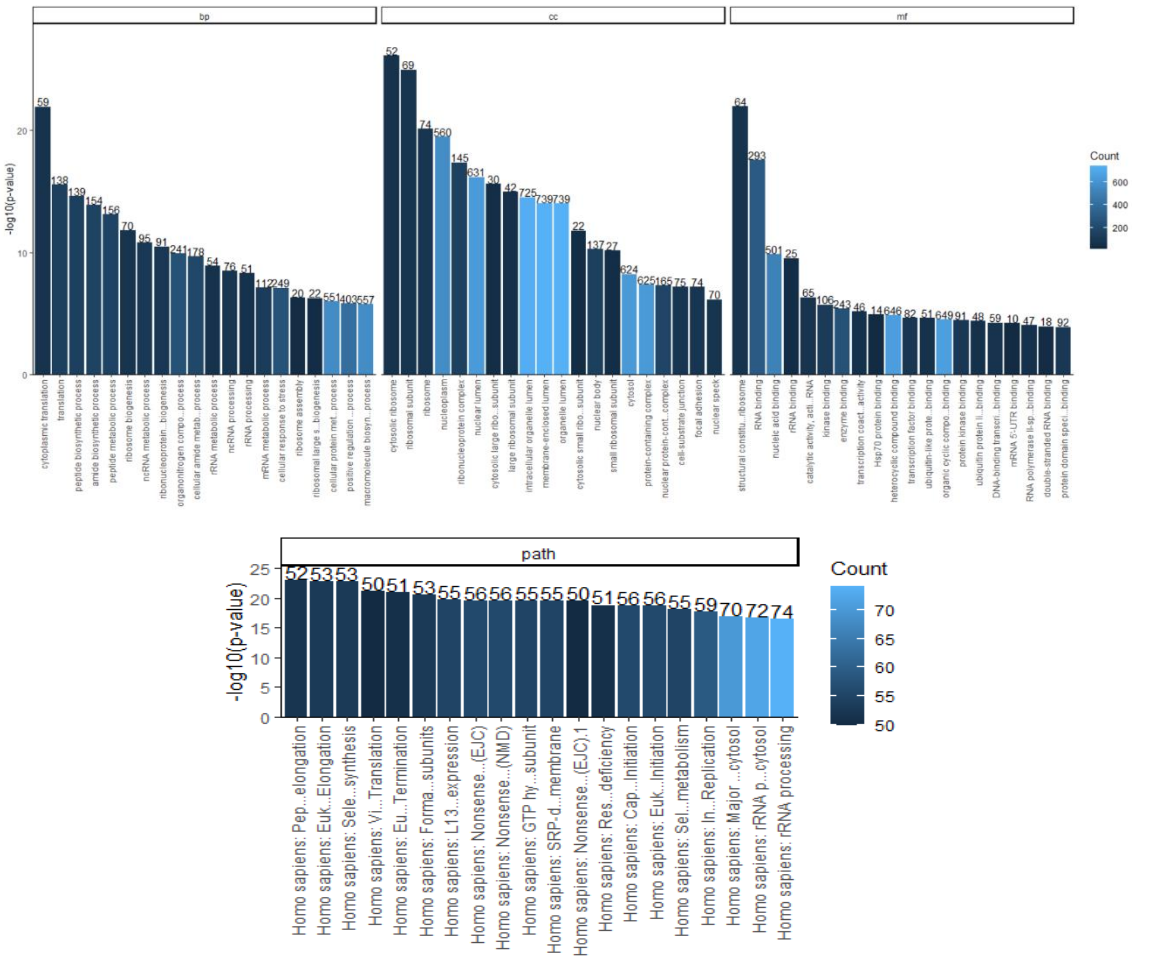


Figura 11. Enriquecimiento funcional de picos comunes (GO Terms, arriba; Reactome Pathways, abajo)

En este ejemplo, los tres procesos biológicos más significativos son *cytoplasmic translation* (GO:0002181), *translation* (GO:0006412) y *peptide biosynthetic process* (GO:0043043). Los tres componentes celulares más significativos son *cytosolic ribosome* (GO:0022626)*, ribosomal subunit* (GO:0044391) y *ribosome* (GO:0005840). Las tres funciones moleculares más significativas son *structural constituent of ribosome* (GO:0003735)*, RNA binding* (GO:0003723) y *nucleic acid binding* (GO:0003676). Por otra parte, las tres rutas biológicas más significativas son *Homo sapiens: Peptide chain elongation* (R-HSA-156902), *Homo sapiens: Eukaryotic Translation Elongation* (R-HSA-156842) *y Homo sapiens: Viral mRNA Translation* (R-HSA-192823).

En el análisis de motivos, el paquete rGREAT ofrece también un análisis vía consulta con la base de datos MSigDB. En este análisis hemos podido extraer los cuatro motivos más significativos (Tabla 3) y las dianas conocidas de factores de transcripción (Tabla 4). Finalmente, podemos conocer cuántos genes solapan por cada uno de estos motivos y a qué distancia se sitúan los motivos de los TSS (Figura 12).

Nombre	Binom Raw P-Value	Binomial Adj P-Value
Motif NNNNKGGRAANTCCCN (no known TF)	2.132133e-31	1.311262e-28
Motif GGGAMTTYCC matches NFKB RELA: v-rel reticuloendotheliosis viral oncogene homolog A, nuclear factor of kappa light polypeptide gene enhancer in B-cells 3, p65 (avian)	9.744519e-25	2.996440e-22
Motif ACWTCCK matches ETV4: ets variant gene 4 (E1A enhancer binding protein, E1AF)	7.770098e-23	1.592870e-20
Motif GGGRATTTCC matches RELA: v-rel reticuloendotheliosis viral oncogene homolog A, nuclear factor of kappa light polypeptide gene enhancer in B-cells 3, p65 (avian)	1.810137e-18	2.783086e-16

Tabla 3. Motivos de ADN más significativos en los picos comunes

Nombre	Binomial Raw P-Value	Binomial Adj P-Value
Genes bound by one of the five NF-kB subunits in U937 cells before or 1 hour after lipopolysaccharide stimulation	6.050712e-30	1.149635e-28
Targets of YY1 identified by ChIP-chip	5.728943e-15	5.442496e-14
Targets of CREB, identified by ChIP-chip in HEK293T cells in three different time points after forskolin stimulation	1.606017e-10	1.017144e-09
Targets of ETS1, identified by ChIP-chip in Jurkat T-cells	1.008537e-08	4.790551e-08

Tabla 4. Dianas de factores de transcripción conocidos más significativos en los picos comunes

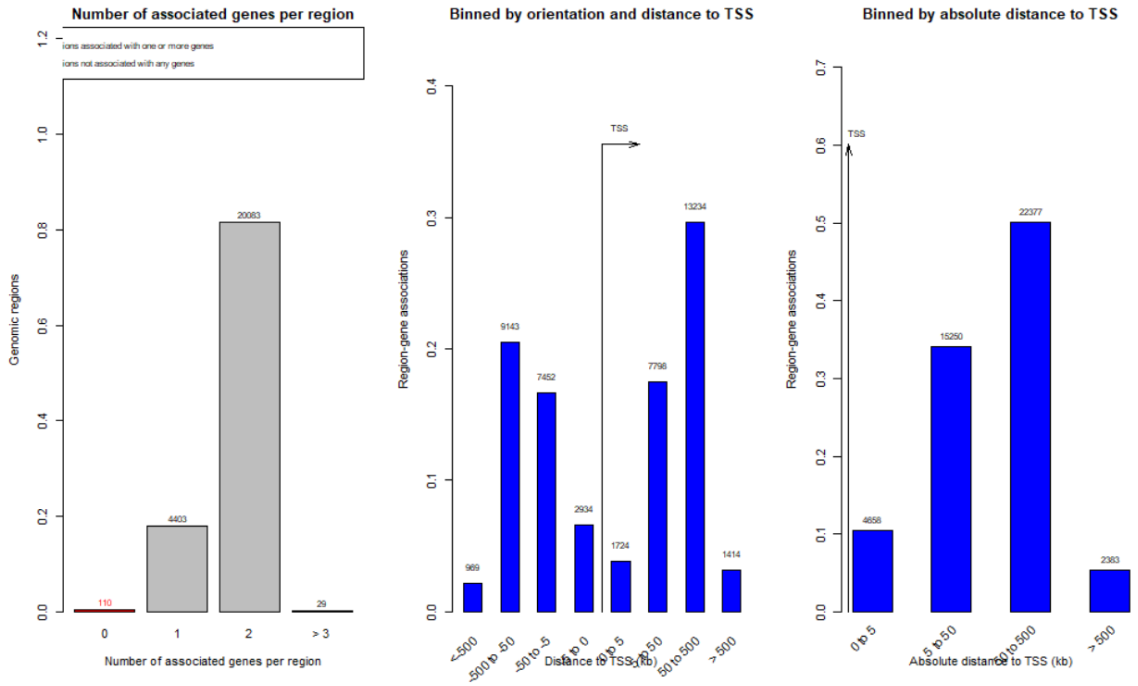


Figura 12. Genes asociados a motivos y distancias de los motivos a los TSS

4.4 Análisis 2: Condiciones diferentes

En este análisis seguimos los mismos pasos que en el anterior, la mayor parte realizada con ChIPpeakAnno con aportaciones puntuales de ChIPseeker. El diagrama de Venn (Figura 13, izquierda) indica que una gran parte de los picos (74.07 %) de la condición 2 están compartidos con los de la condición 1, que solo suponen un 40.51 % de estos. Al menos el 50 % de los picos se encuentran en promotores, siendo mayor el porcentaje en la condición 2 y de nuevo un alto porcentaje de estos picos se encuentran en exones y regiones 5'-UTR (Figura 13, derecha). El 73.10 % de los picos comunes a ambas condiciones se encuentran en promotores, y un 57.30 % se encuentran en las primeras 500 pb downstream de los TSS (Figura 15, derecha). La distribución en torno a los TSS también es bastante uniforme en los picos comunes (Figura 14, izquierda), aunque cuando observamos las dos condiciones por separado se aprecian irregularidades a ambos lados de los TSS (Figura 14, derecha), seguramente debido a aquellos picos exclusivos de cada condición.

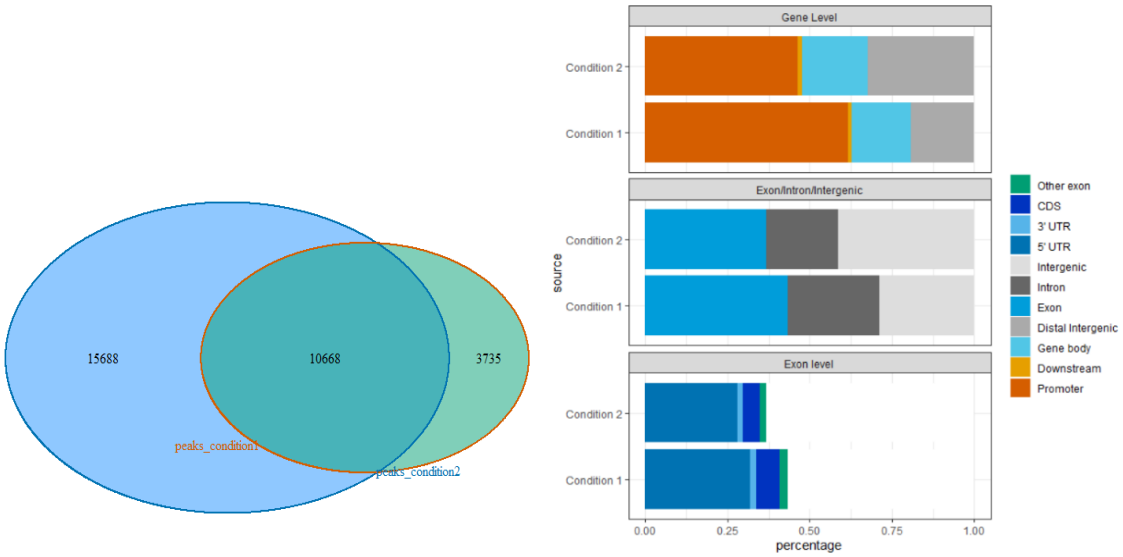


Figura 13. Diagrama de Venn de los picos (izquierda) y solapamiento con elementos genómicos en las diferentes condiciones experimentales (derecha)

En la anotación funcional (Figura 16) probamos esta vez un análisis más estricto, buscando solo solapamientos de picos con promotores bidireccionales. El 48.18 % de los picos comunes entre condiciones solapan con estos promotores, de los cuales el 61.50 % solapan al comienzo del promotor, un 35.90 % están en posición upstream y un 2.04 % están dentro del promotor.

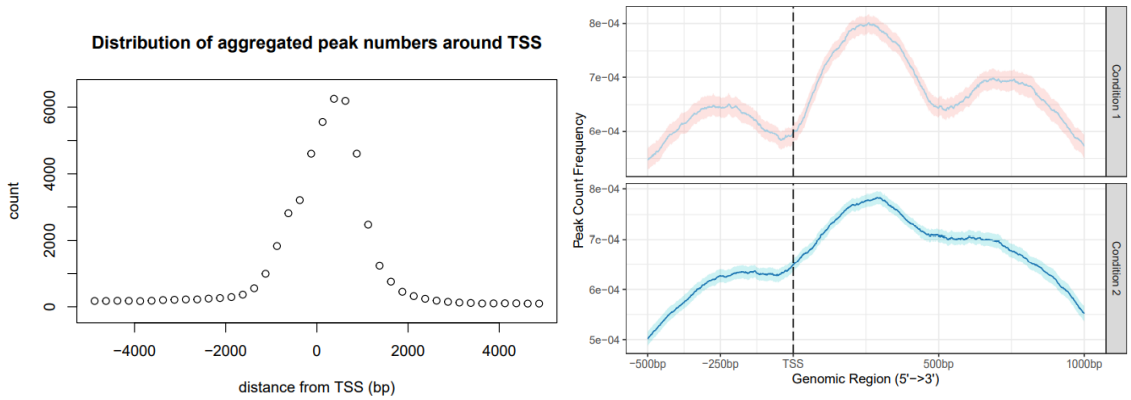


Figura 14. Distancias picos-TSS de los picos comunes (ChIPpeakAnno, izquierda) y de los picos de cada una de las condiciones experimentales (ChIPseeker, derecha)

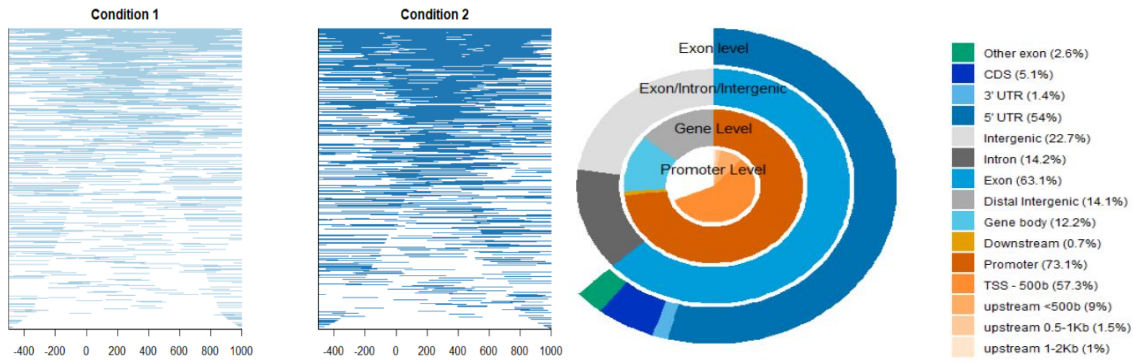


Figura 15. Heatmap de los picos distribuidos en torno a los TSS en ambas condiciones experimentales (izquierda) y solapamiento con elementos genómicos de los picos comunes (derecha)

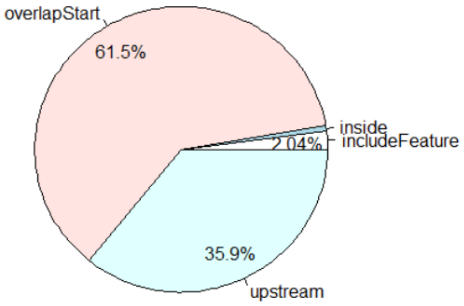


Figura 16. Anotación funcional de los picos comunes entre condiciones experimentales (ChIPpeakAnno)

El posterior enriquecimiento funcional (Figura 17) lo realizamos sobre estos picos solapantes con promotores bidireccionales. De nuevo la mejor opción que tenemos son las funciones integradas en ChIPpeakAnno. Esta vez los tres procesos biológicos más significativos son *cellular nitrogen compound metabolic process* (GO:0034641), *nucleic acid metabolic process* (GO:0090304) y *ncRNA metabolic process* (GO:0034660).

Los tres componentes celulares más significativos son *nucleoplasm* (GO:0005654), *ribonucleoprotein complex* (GO:1990904) e *intracellular organelle lumen* (GO:0070013). Las tres funciones moleculares más significativas son *RNA binding* (GO:0003723), *nucleic acid binding* (GO:0003676) *y catalytic activity, acting on RNA* (GO:0140098).

Por otra parte, las tres rutas biológicas más significativas son *Homo sapiens: Metabolism of RNA* (R-HSA-8953854), *Homo sapiens:* Translation (R-HSA-72766) *y Homo sapiens: rRNA processing* (R-HSA-72312).

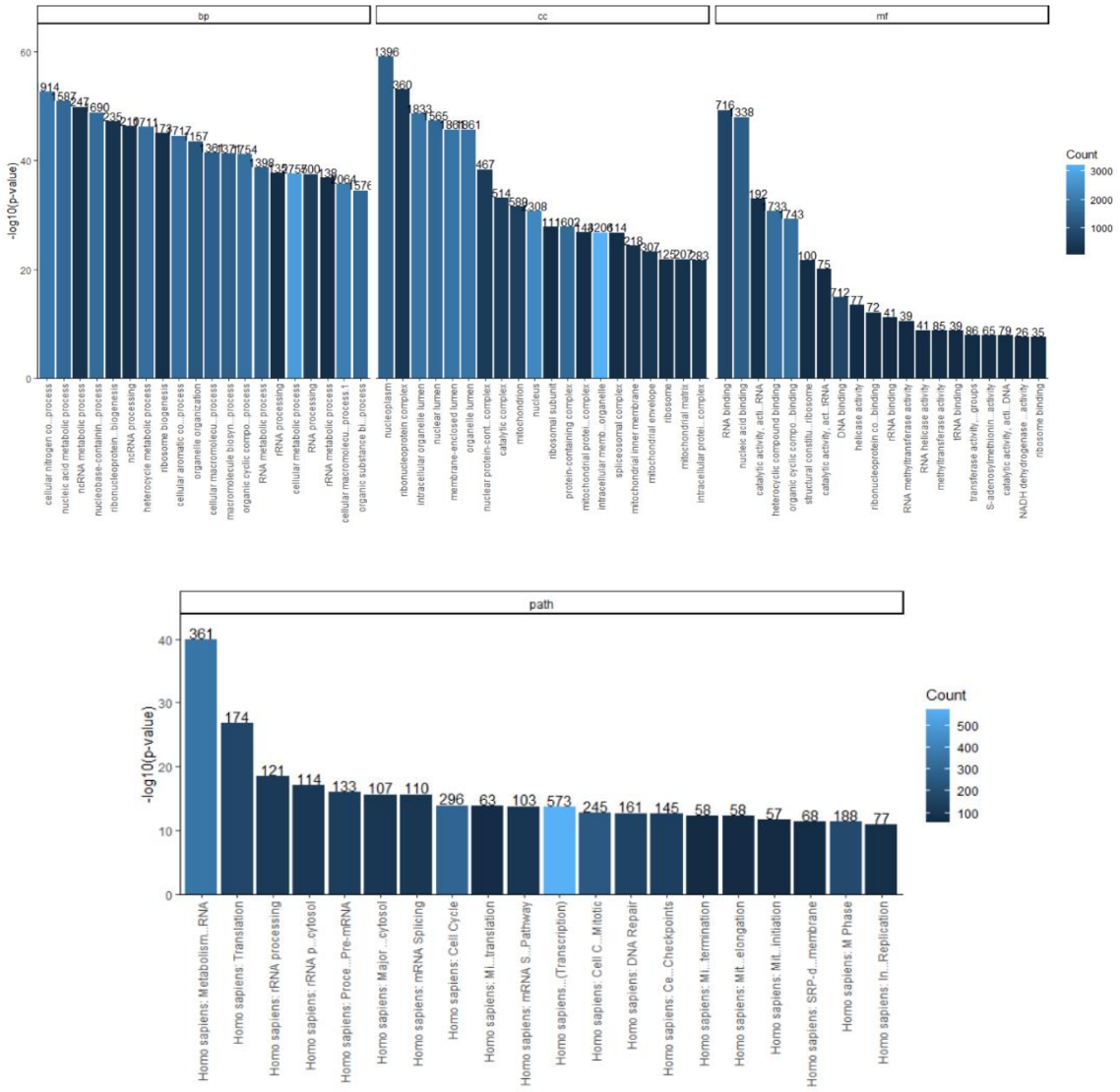


Figura 17. Enriquecimiento funcional de los picos comunes entre condiciones experimentales. Funciones biológicas más significativas, de GO Terms (arriba) y Reactome Pathways (abajo)

Finalmente, obtenemos los cuatro motivos de ADN más significativos (Tabla 5) y cuántos genes solapan por cada uno de estos motivos y a qué distancia se sitúan los motivos de los TSS (Figura 18).

Nombre	Binom Raw P-Value	Binomial Adj P-Value
Motif SCGGAAGY matches ELK1: ELK1, member of ETS oncogene family	2.362545e-53	1.452965e-50
Motif ACCGGAAGNG matches GABPB1: GA binding protein transcription factor, beta subunit 1.	1.036863e-27	3.188354e-25
Motif VCCGGAAGNGCR matches GABPA: GA binding protein	3.571133e-25	7.320823e-23

transcription factor, alpha subunit 60kDa GABPB2: GA binding protein transcription factor, beta subunit 2		
Motif RCGCANGCGY matches NRF1: nuclear respiratory factor 1	8.468444e-17	1.302023e-14

Tabla 5. Motivos de ADN más significativos en los picos comunes

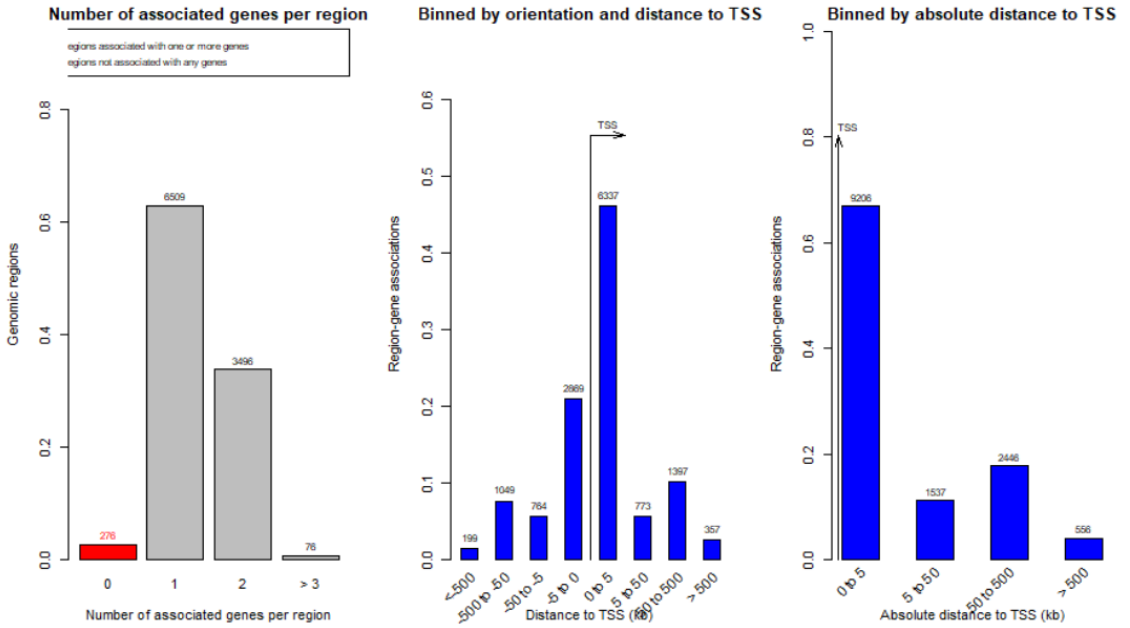


Figura 18. Genes asociados a motivos y distancias de los motivos a los TSS

4.5 Análisis 3: Líneas celulares diferentes

En este último análisis nos centramos más en ChIPseeker, con usos puntuales de ChIPpeakAnno. Comenzamos por el diagrama de Venn (creado mediante *vennplot*) (Figura 19) para comparar los picos entre las tres líneas celulares. Si bien ChIPseeker no nos ofrece la posibilidad de personalizarlo, sigue siendo una herramienta útil para

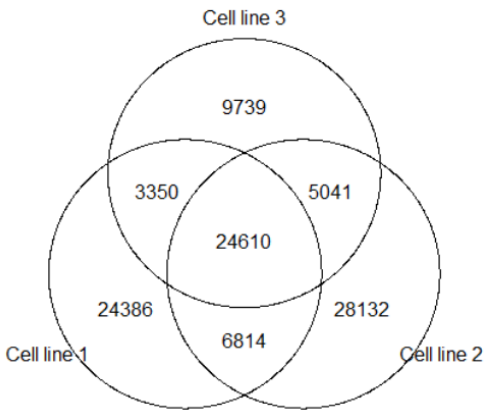


Figura 19. Diagrama de Venn entre líneas celulares

empezar el análisis. La distancia a los TSS nos indica de nuevo que los picos comunes a las tres líneas celulares se encuentran distribuidos en torno a los TSS, asemejándose a una curva normal (Figura 20, izquierda). Sin embargo, los picos en cada línea celular muestran una distribución más extensa, en las primeras 500 pb downstream de los TSS, y algo más variable según la línea celular (Figura 20, derecha). Mediante los heatmaps generados con ChIPseeker (mediante *tagHeatmap*) (Figura 21) también nos hacemos una idea de la densidad de los picos de cada línea celular en torno a los TSS.

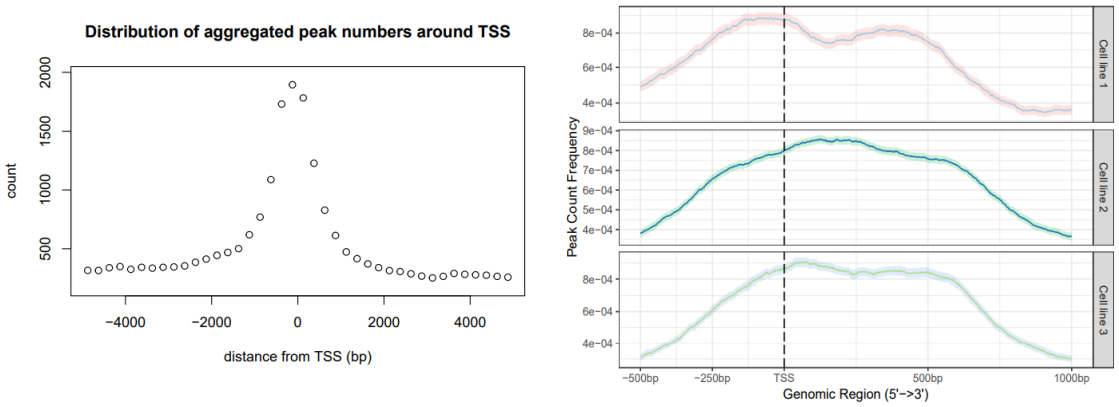


Figura 20. Distancias picos-TSS de los picos comunes entre líneas celulares (ChIPpeakAnno, izquierda) y de cada una de las líneas celulares (ChIPseeker, derecha)

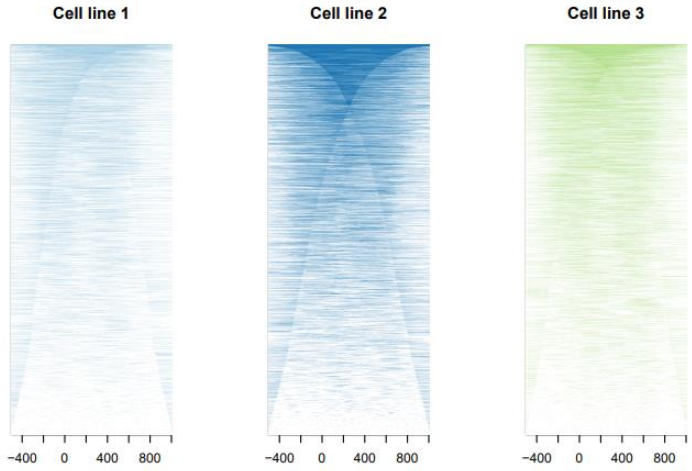


Figura 21. Densidad de los picos de cada línea celular en torno a los TSS

En la anotación funcional, si bien ChIPseeker ofrece menos posibilidades para configurar la búsqueda de anotación que ChIPpeakAnno (ej.: no incluye opción para buscar promotores bidireccionales), si incluye la opción de buscar solo anotación que esté en la misma hebra de ADN que los picos (*sameStrand = TRUE*). En este ejemplo, nos centramos en una búsqueda de anotación en ambas hebras. Comprobamos que el porcentaje de picos en promotores es relativamente bajo en las tres líneas celulares, predominando los picos situados en regiones no codificantes, ya sean regiones intergénicas o intrones (Figura 22). Esto concuerda con el hecho de que solo un 20-30 % de los picos se sitúan a una distancia de 1-3 kb de los TSS, aumentando el porcentaje conforme nos alejamos de estos. Estos diagramas de barras son vistosos y facilitan la interpretación visual de los datos, aunque solo se pueden aplicar a listas de datos ChIP-Seq, no a datos individuales. Por lo tanto, no podemos aplicarlos a los picos comunes entre las líneas celulares.

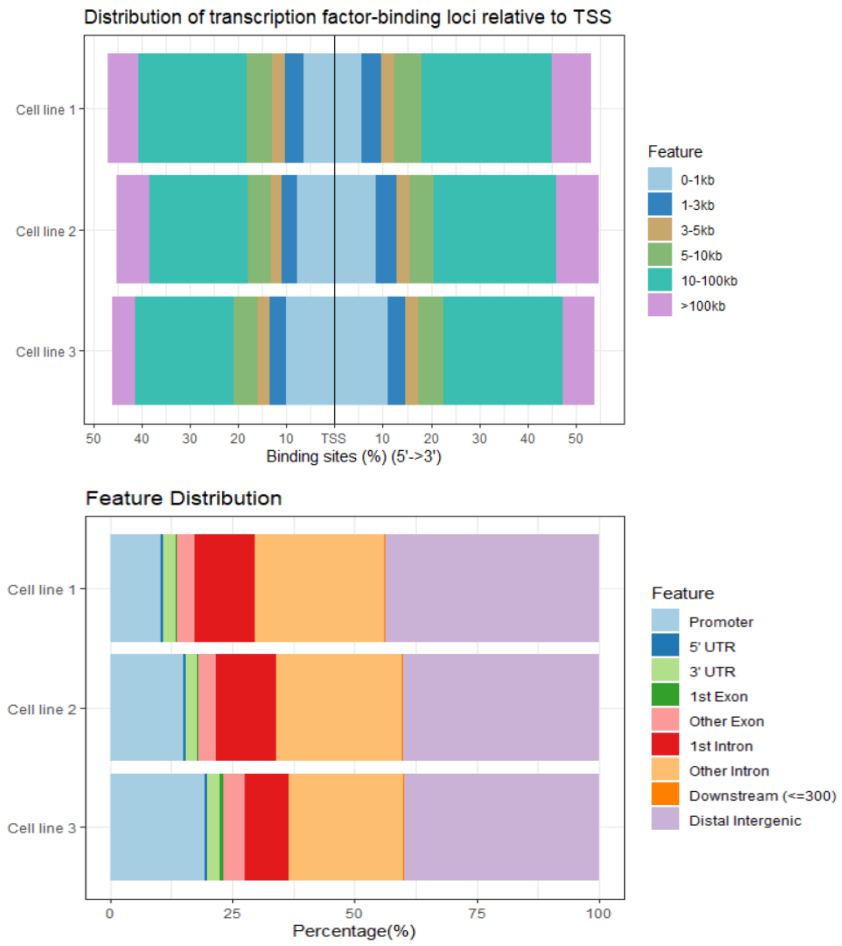


Figura 22. Anotación funcional de los picos de cada línea celular. Incluye el solapamiento con elementos genómicos (abajo) y distancias TSS (arriba) (ChIPseeker)

Para el análisis de enriquecimiento funcional, ChIPseeker no incluye funciones integradas con esta función, por lo que tenemos que recurrir al paquete clusterProfiler. Mediante este paquete podemos realizar un análisis tanto de GO Terms como de KEGG Pathway, para las rutas biológicas. Podemos crear los diagramas de puntos de la Figura 23, donde observamos cuales son las funciones y rutas biológicas más significativas.

Finalmente, el análisis de motivos sí podemos realizarlo sobre los picos comunes a las tres líneas celulares, obteniendo la Tabla 6. En la Figura 24 observamos cuántos genes solapan por cada uno de estos motivos y a qué distancia se sitúan los motivos de los TSS (Figura 18).

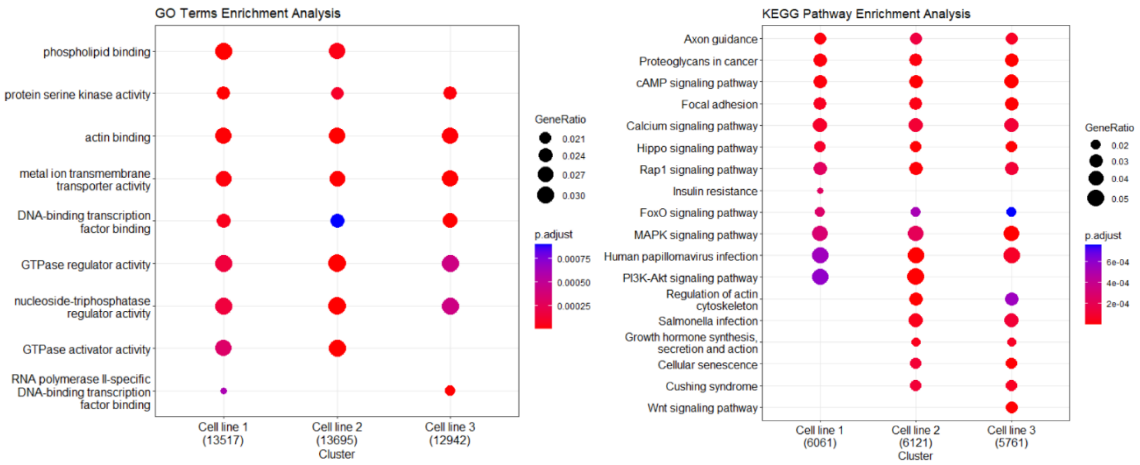


Figura 23. Enriquecimiento funcional de los picos de cada línea celular con los GO Terms más significativos (izquierda) y los KEGG Pathways (derecha)

Nombre	Binom Raw P-Value	Binomial Adj P-Value
Motif GGGCGGR matches SP1: Sp1 transcription factor	3.607082e-49	2.218355e-46
Motif GGGGCGGGGC matches SP1: Sp1 transcription factor	1.426937e-24	4.387831e-22
Motif ACWTCCK matches ETV4: ets variant gene 4 (E1A enhancer binding protein, E1AF)	2.385696e-24	4.890677e-22
Motif GGGGAGGG matches MAZ: MYC-associated zinc finger protein (purine-binding transcription factor)	2.833600e-23	4.356660e-21

Tabla 6. Motivos de ADN más significativos en los picos comunes

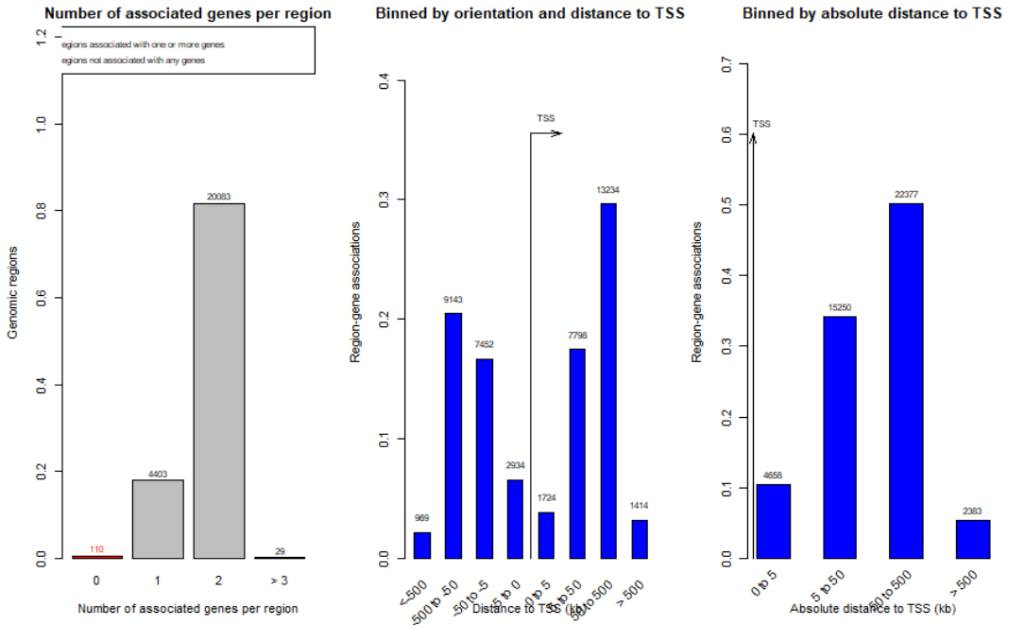


Figura 24. Genes asociados a motivos y distancias de los motivos a los TSS

5. Discusión

Los pasos preliminares a la anotación funcional son muy similares, ya sean con ChIPpeakAnno o ChIPseeker. La principal fortaleza de ChIPpeakAnno en esta parte del análisis es el diagrama de barras mediante la función *genomicElementDistribution* (ej.: Figuras 7 y 9), ya que aporta bastante información de manera específica para cada conjunto de datos. Sus ventajas se resumen en: 1) generar estos gráficos directamente desde los ficheros BED/NARROWPEAK/BROADPEAK sin pasar por la anotación funcional; 2) poder personalizar la región en la que consideramos solapamiento y 3) poder visualizar los porcentajes a nivel exónico, intergénico, génico y de promotor para los picos comunes (Figura 9) [20]. La figura 22 (abajo) es el gráfico más parecido que podemos generar con ChIPseeker, sin la posibilidad de observar los solapamientos en promotores con tanto detalle ni una división tan clara en la escala de los solapamientos.

Sin embargo, con ChIPpeakAnno no es posible generar un gráfico de distancias picos-TSS para más de un fichero, por ello la Figura 14 (izquierda) solo se aplica a los picos comunes. En cambio, esto si es posible en ChIPseeker mediante la función *plotAvgProf*, pudiendo además remuestrear los datos con cálculo de intervalo de confianza (Figura 14, derecha) [21].

En la fase de anotación funcional, si se busca una mayor personalización y especificidad de la anotación de los picos, la función *annotatePeakInBatch* de ChIPpeakAnno es la mejor opción. Mientras que, si se opta más por una visualización de la anotación más llamativa, aunque no necesariamente por ello menos informativa, la función *annotatePeak* de ChIPseeker es más útil.

annotatePeakInBatch permite diferenciar entre anotación de picos solo con elementos genómicos solapantes (*output="overlapping*"), solo con los más cercanos dentro de una región establecida por el usuario (*output="nearestLocation*") o con ambos casos (*output="both"*) [20].

Adicionalmente, este argumento permite anotar picos solapantes con promotores bidireccionales (*output="nearestBiDirectionalPromoters"*) [20]. También es poisble configurar calcular las distancias desde el inicio, mitad o final de cada pico o cada elemento genómico con los argumentos *PeakLocForDistance = c("start", "middle", "end", "endMinusStart")* y *FeatureLocForDistance = c("TSS", "middle", "start", "end", "geneEnd")* [20]. Ninguna de estas tres características se encuentran en la función *annotatePeak* de ChIPseeker*,* aunque si permite generar gráficos de distancias a TSS y diagramas de barras específicos de cada fichero (Figura 22), más informativos que la generada por ChIPpeakAnno (ej.: Figura 16).

En la última fase del análisis, las funciones integradas en ChIPpeakAnno para enriquecimiento de los datos y su visualización (*getEnrichedGO*, *getEnrichedPATH* y *enrichmentPlot)* son una opción de fácil ejecución y rápida visualización, mostrando las tres categorías de GO Terms (bp, cc y fm) y en cada una las 20 funciones biológicas más significativas estadísticamente (Figuras 11 y 17) [20].

Por otra parte, al no ofrecer estas funciones ChIPseeker, clusterProfilern ofrece la función *compareCluster*, con la que permite analizar tanto GO Terms como KEGG Pathways. *compareCluster* no muestra la clasificación bp, cc y fm pero el formato de diagrama de puntos es bastante compacto y fácil de entender (Figura 23) [22, 23]. También ofrece la posibilidad de establecer un umbral del P-Value a partir del cual considerar la anotación significativa y qué método utilizar para el ajuste del P-Value (ej.: Bonferroni o BH) [22, 23].

Finalmente, en cuanto al análisis de motivos de ADN, el paquete rGREAT cumple con las funciones básicas esperadas para identificar motivos ya conocidos como posibles dianas de factores de transcripción y obtener los más significativos (ej.: Tablas 3 y 4). Supone una alternativa de menor coste computacional en comparación con otros paquetes del proyecto Bioconductor, como el paquete rGADEM [36], el cual se especializa en la identificación de motivos *de novo*. La principal desventaja del paquete rGREAT puede ser que no incluya funciones dedicadas a la visualización de logos de los motivos, por

lo que para ello puede ser necesario recurrir a otras herramientas, por ejemplo, servidores online como MEME-Chip o FIMO [37]. Este puede ser el aspecto más débil de este pipeline, siendo necesario utilizar herramientas alternativas para un análisis de motivos de ADN más exhaustivo.

6. Conclusiones

6.1 Conclusiones del trabajo

- ChIPpeakAnno ofrece funcionalidades más flexibles, personalizables y específicas para la anotación funcional de picos.

- ChIPseeker ofrece mayor variedad de funciones dedicadas a la visualización de la anotación funcional de picos.

- ChIPpeakAnno es menos flexible y ofrece menos posibilidades de personalización del análisis de enriquecimiento funcional que clusterProfiler.

- ChIPseeker no es un paquete apropiado para análisis de enriquecimiento funcional.

- Ambos paquetes son eficientes para evaluar datos extraídos de diferentes contextos experimentales.

- Ambos paquetes compensan sus fortalezas y debilidades, por lo que puede ser necesario usarlos conjuntamente según los datos de interés.

- rGREAT cumple los requerimientos básicos para realizar el análisis de motivos de ADN, aunque ofrece un catálogo muy limitado de funciones.

7. Bibliografía

1. Johnson, D. S., Mortazavi, A., Myers, R. M., & Wold, B. (2007) Genome-wide mapping of in vivo protein-DNA interactions. Science (New York, N.Y.), 316(5830), 1497–1502. https://doi.org/10.1126/science.1141319

2. Park, P. (2009) ChIP–seq: advantages and challenges of a maturing technology. Nat Rev Genet 10, 669–680 https://doi.org/10.1038/nrg2641

3. Furey T. S. (2012). ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. Nature reviews. Genetics, 13(12), 840–852. https://doi.org/10.1038/nrg3306

4. Farh, K. K., Marson, A., Zhu, J., Kleinewietfeld, M., Housley, W. J., Beik, S., Shoresh, N., Whitton, H., Ryan, R. J., Shishkin, A. A., Hatan, M., Carrasco-Alfonso, M. J., Mayer, D., Luckey, C. J., Patsopoulos, N. A., De Jager, P. L., Kuchroo, V. K., Epstein, C. B., Daly, M. J., Hafler, D. A., … Bernstein, B. E. (2015). Genetic and epigenetic fine mapping of causal autoimmune disease variants. Nature, 518(7539), 337–343. https://doi.org/10.1038/nature13835

5. Heinz, S., Romanoski, C. E., Benner, C., & Glass, C. K. (2015). The selection and function of cell type-specific enhancers. Nature reviews. Molecular cell biology, 16(3), 144–154. https://doi.org/10.1038/nrm3949

6. Mundade, R., Ozer, H. G., Wei, H., Prabhu, L., & Lu, T. (2014). Role of ChIP-seq in the discovery of transcription factor binding sites, differential gene regulation mechanism, epigenetic marks and beyond. Cell cycle (Georgetown, Tex.), 13(18), 2847–2852. https://doi.org/10.4161/15384101.2014.949201

7. Dirks, R. A., Stunnenberg, H. G., & Marks, H. (2016). Genome-wide epigenomic profiling for biomarker discovery. Clinical epigenetics, 8, 122. https://doi.org/10.1186/s13148-016-0284-4

8. Nakato, R., & Sakata, T. (2021). Methods for ChIP-seq analysis: A practical workflow and advanced applications. Methods (San Diego, Calif.), 187, 44–53. https://doi.org/10.1016/j.ymeth.2020.03.005

9. Yan, H., Tian, S., Slager, S. L., & Sun, Z. (2016). ChIP-seq in studying epigenetic mechanisms of disease and promoting precision medicine: progresses and future directions. Epigenomics, 8(9), 1239–1258. https://doi.org/10.2217/epi-2016-0053

10. Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A. J., Sawitzki, G., … Zhang, J. (2004). Bioconductor: open software development for computational biology and bioinformatics. Genome biology, 5(10), R80. https://doi.org/10.1186/gb-2004-5-10-r80

11. Huber, W., Carey, V. J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B. S., Bravo, H. C., Davis, S., Gatto, L., Girke, T., Gottardo, R., Hahne, F., Hansen, K. D., Irizarry, R. A., Lawrence, M., Love, M. I., MacDonald, J., Obenchain, V., Oleś, A. K., Pagès, H., … Morgan, M. (2015). Orchestrating high-throughput genomic analysis with Bioconductor. Nature methods, 12(2), 115–121. https://doi.org/10.1038/nmeth.3252

12. Thomas Samuel Carroll, Ziwei Liang, Rafik Salama, Rory Stark, Ines de Santiago: Impact of artefact removal on ChIP quality metrics in ChIP-seq and ChIP-exo data. Frontiers in Genetics, in press. https://doi.org/doi:10.18129/B9.bioc.ChIPQC

13. Stark R and Brown G (2011). DiffBind: differential binding analysis of ChIP-Seq peak data. Bioconductor. https://doi.org/doi:10.18129/B9.bioc.DiffBind

14. Edgar, R., Domrachev, M., & Lash, A. E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Research, 30 (1), 207–210. https://doi.org/10.1093/nar/30.1.207

15. Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., Holko, M., Yefanov, A., Lee, H., Zhang, N., Robertson, C. L., Serova, N., Davis, S., & Soboleva, A. (2012). NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Research, 41 (D1), D991–D995. https://doi.org/10.1093/nar/gks1193

16. Davis, C. A., Hitz, B. C., Sloan, C. A., Chan, E. T., Davidson, J. M., Gabdank, I., Hilton, J. A., Jain, K., Baymuradov, U. K., Narayanan, A. K., Onate, K. C., Graham, K., Miyasato, S. R., Dreszer, T. R., Strattan, J. S., Jolanki, O., Tanaka, F. Y., & Cherry, J. M. (2017). The Encyclopedia of DNA elements (ENCODE): data portal update. Nucleic Acids Research, 46 (D1), D794–D801. https://doi.org/10.1093/nar/gkx1081

17. Davis, H., C. A. (2018). The encyclopedia of dna elements (encode): Data portal update. Nature, 489, 57–74. https://doi.org/10.1093/nar/gkx108137

18. Liu, T., Ortiz, J. A., & Taing, L. (2011). Cistrome: An integrative platform for transcriptional regulation studies. Genome Biol, 12, R83. https://doi.org/10.1186/gb-2011-12-8-r83

19. Zheng, R., Wan, C., Mei, S., Qin, Q., Wu, Q., Sun, H., Chen, C.-H., Brown, M., Zhang, X., Meyer, C. A., & Liu, X. S. (2018). Cistrome Data Browser: expanded datasets and new tools for gene regulatory analysis. Nucleic Acids Research, 47 (D1), D729–D735. https://doi.org/10.1093/nar/gky1094

20. Zhu, L. J., Gazin, C., Lawson, N. D., Pagès, H., Lin, S. M., Lapointe, D. S., & Green, M. R. (2009). ChIPpeakAnno: A bioconductor package to annotate chip-seq and chip-chip data. BMC Bioinformatics, 11, 237–237. https://doi.org/10.1186/1471-2105-11-237

21. Yu, G., Wang, L.-G., & He, Q.-Y. (2015). ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization.

Bioinformatics, 31 (14), 2382–2383. https://doi.org/10.1093/bioinformatics/btv145

22. Yu, G., Wang, L.-G., Han, Y., & He, Q.-Y. (2012). ClusterProfiler: An r package for comparing biological themes among gene clusters. Omics : A Journal of Integrative Biology, 16 (5), 284—287. https://doi.org/10.1089/omi.2011.0118

23. Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., Feng, T., Zhou, L., Tang, W., Zhan, L., Fu, X., Liu, S., Bo, X., & Yu, G. (2021). ClusterProfiler 4.0: A universal enrichment tool for interpreting omics data. The Innovation, 2 (3), 100141. https://doi.org/10.1016/j.xinn.2021.100141

24. R Special Interest Group on Databases (R-SIG-DB), Wickham, H., & Müller, K. (2021). DBI: R database interface. https://CRAN.R-project.org/package=DBI

25. Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, et al. (2013) Software for Computing and Annotating Genomic Ranges. PLOS Computational Biology 9(8): e1003118. https://doi.org/10.1371/journal.pcbi.1003118

26. Carlson M (2019). org.Hs.eg.db: Genome wide annotation for Human. R package version 3.8.2. https://doi.org/doi:10.18129/B9.bioc.org.Hs.eg.db

27. Crook, O. M., Breckels, L. M., Lilley, K. S., Kirk, P., & Gatto, L. (2019). A Bioconductor workflow for the Bayesian analysis of spatial proteomics. F1000Research, 8, 446. https://doi.org/10.12688/f1000research.18636.1

28. Ligtenberg W (2019). reactome.db: A set of annotation maps for reactome. R package version 1.68.0. https://doi.org/doi:10.18129/B9.bioc.reactome.db

29. Yu, G., & He, Q.-Y. (2016). ReactomePA: An r/bioconductor package for reactome pathway analysis and visualization. Molecular BioSystems, 12 (12), 477–479. https://doi.org/10.1039/C5MB00663E

30. Gu Z (2021). rGREAT: Client for GREAT Analysis. https://github.com/jokergoo/rGREAT, http://great.stanford.edu/public/html/.

31. Carlson M, Maintainer BP (2015). TxDb.Hsapiens.UCSC.hg18.knownGene: Annotation package for TxDb object(s). R package version 3.2.2. https://doi.org/doi:10.18129/B9.bioc.TxDb.Hsapiens.UCSC.hg18.knownGene

32. Carlson M, Maintainer BP (2015). TxDb.Hsapiens.UCSC.hg19.knownGene: Annotation package for TxDb object(s). R package version 3.2.2. https://doi.org/doi:10.18129/B9.bioc.TxDb.Hsapiens.UCSC.hg19.knownGene

33. Kasowski, M., Grubert, F., Heffelfinger, C., Hariharan, M., Asabere, A., Waszak, S. M., Habegger, L., Rozowsky, J., Shi, M., Urban, A. E., Hong, M. Y., Karczewski, K. J., Huber, W., Weissman, S. M., Gerstein, M. B., Korbel, J. O., & Snyder, M. (2010). Variation in transcription factor binding among humans. Science (New York, N.Y.), 328(5975), 232–235. https://doi.org/10.1126/science.1183621

34. Nevedomskaya, E., Wessels, L., & Zwart, W. (2014). Genome-wide epigenetic profiling of breast cancer tumors treated with aromatase inhibitors. Genomics data, 2, 195–198. https://doi.org/10.1016/j.gdata.2014.06.023

35. ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature, 489(7414), 57–74. https://doi.org/10.1038/nature11247

36. Arnaud Droit, Raphael Gottardo, Gordon Robertson and Leiping Li (2021). rGADEM: de novo motif discovery. R package version 2.42.0. https://doi.org/doi:10.18129/B9.bioc.rGADEM

37. Timothy L. Bailey, James Johnson, Charles E. Grant, William S. Noble, (2015) The MEME Suite, Nucleic Acids Research, Volume 43, Issue W1, 1 July, Pages W39–W49, https://doi.org/10.1093/nar/gkv416

8. Anexos

- Pipeline íntegro con los tres análisis expuestos en la memoria escrito en un archivo RMarkdown en formato pdf.