ChIP-seq data analysis of Sin3A in Drosophila melanogaster

Fien Strijthaegen
Université Libre de Bruxelles
MA1 Bioinformatics and modelling

August 24th, 2023

1 Introduction

Chromatin immunoprecipitation followed by sequencing (ChIP-seq) is a method in epigenomic research for the analysis of binding sites of DNA-associated proteins. DNA and associated proteins on chromatin are crosslinked, then the DNA-protein complexes are sheared into small fragments. The fragments associated with the protein of interest are selectively immunoprecipitated using a protein-specific antibody and then purified and sequenced. Computational analysis of these sequenced reads can reveal epigenomic information[5]. The ENCODE consortium[6] hosts several databases of biological assays, including ChIP-seq of several species.

In this project, an analysis was performed on a data set of the Kc167 cell line of Drosophila melanogaster from the ENCODE ChIP-seq database, targeting the Sin3A gene. This gene encodes the paired amphipathic helix protein of the same Sin3A, a transcriptional regulatory protein. The analysis was guided by practice sessions from this course as well as a HCB [**meeta2023hbctraining**] and CRUK course.

TODO: summary

2 Data sets

This analysis concerns data from Encode experiment ENCSR264MBG, the reads of which were sequences on the Illumina HiSeq 2000 platform. The analysis was performed on isogenic replicate 1 and 2. The reads are single-ended, with a length of 44 nucleotides. The genome assembly of Drosophila melanogaster was acquired from the UCSC genome browser[1]. The soft-masked assembly from the file dm6.fa.gz was used in the analysis. The BED file dm6-blacklist.v2.bed.gz with blacklisted regions was acquired from the Boyle Lab[**amemiya2019encode**]. The database used for MEME-Chip analysis was OnTheFly$_{2}$014$_{D}$*rosophila.meme*[**shazman2014onthefly**], *wh*

3 Code availability

TODO

4 Analysis

4.1 Read mapping

Code for this section is in the R script read_mapping.R. The raw FASTQ files contain reads of length 44nt, 23 million for isogenic replicate 1, and 18 million for isogenic replicate 2. The ends of these reads were trimmed based on quality, ends with a quality score lower than 20 were trimmed. Reads with a length lower than 40nt were filtered out. This filtering was performed using the ShortRead package[4].

Reads were mapped to the assembled genome using the Rsubread package[3] with default parameters. Then filtering was performed using Sambamba[**tarasov2015sambamba**]. Unmapped reads, duplicates and multimapped reads were filtered out, leaving only uniquely mapping reads. While including multi-mapped reads would increase the number of usable reads and might increase the sensitivity of peak detection, they were omitted here, following convention, since the number of false positives might also increase[**chung2011discovering**].

4.2 Peak calling

Peak calling was performed with MACS2[**gaspar2018improved**] using default parameters. Bedtools[**quinlan2010bedtools**] was used to filter out blacklisted regions from the results. Removing these anomalous, unstructured or high in signal independent of cell line or experiment regions improves accuracy[**amemiya2019encode**]. Concordance between replicates was then assessed by finding overlapping regions. Regions that overlap at least 30% between the results of isogenic replicates 1 and 2 were kept for further analysis, this is meant to only keep peaks that are reproducible across the two replicates. The 30% overlap was chosen for simplicity, but other measures of reproducibility exist as well such as the IDR framework[**li2011measuring**]. Deeptools[**ramirez2014deeptools**] was used to visualize the signal. Figure 1 shows that there is a signal for both isogenic replicates.

4.3 Annotation and functional enrichment analysis

ChIPseeker[**yu2015chipseeker**] was used to annotate genomic features. Figures 2 and **??** visualize the distances from the peak to the transcription start site (TSS) of the nearest genes. These figures reveal that a subset of peaks are near promotor regions, thus potentially being of interest in revealing transcriptional regulation.

ReactomePA[**yu2016reactomepa**] was used for functional enrichment analysis. Figure 4 displays the gene sets found, a table with the full names and p-values can be found in the The three pathways with the lowest p-values are related to the Nonsense Mediated Decay (NMD) pathway, which

4.4 Motif discovery and annotation

The CLI version of MEME-Chip[**machanick2011meme**] was used to perform motif discovery.

5 Conclusion

TODO: SAMENVATTING

There are several more possibilities for analysis of this data. Multi-mapped reads were not included, but could yield further insight[**chung2011discovering**]. Peaks from the two isogenic replicates were seen as overlapping if they overlapped 30%, but there are other methods of measuring reproducibility of peaks, such as the Irreproducible Discovery Rate (IDR) [**li2011measuring**].

References

[1] Donna Karolchik et al. “The UCSC genome browser database”. In: *Nucleic acids research* 31.1 (2003), pp. 51–54.

[2] Heng Li et al. “The sequence alignment/map format and SAMtools”. In: *bioinformatics* 25.16 (2009), pp. 2078–2079.

[3] Yang Liao, Gordon K Smyth, and Wei Shi. “The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads”. In: *Nucleic acids research* 47.8 (2019), e47–e47.

[4] Martin Morgan et al. “ShortRead: a bioconductor package for input, quality assessment and exploration of high-throughput sequence data”. In: *Bioinformatics* 25.19 (2009), pp. 2607–2608.

[5] Ryuichiro Nakato and Toyonori Sakata. “Methods for ChIP-seq analysis: A practical workflow and advanced applications”. In: *Methods* 187 (2021), pp. 44–53.

[6] Natalie de Souza. “The ENCODE project”. In: *Nature methods* 9.11 (2012), pp. 1046–1046.

Figure 1: Profile plot and heatmap of both isogenic replicates. The profile plots of both replicates have a similar shape, a peak at the center of the consensus peaks, with that of isogenic replicate 1 being a bit higher because of the bigger number of sequenced reads.

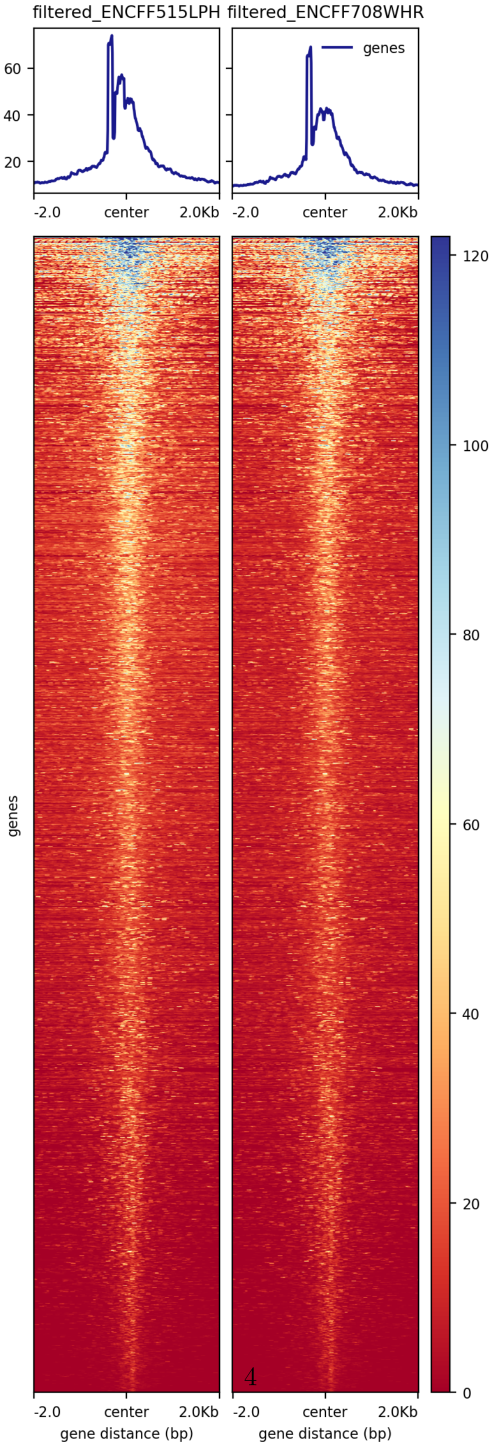


Figure 2: Distribution of locations of peaks relative to nearest TSS. Most peaks are either ¡= 1kb away from the promoter, distal intergenic, or ”‘other intron”’.

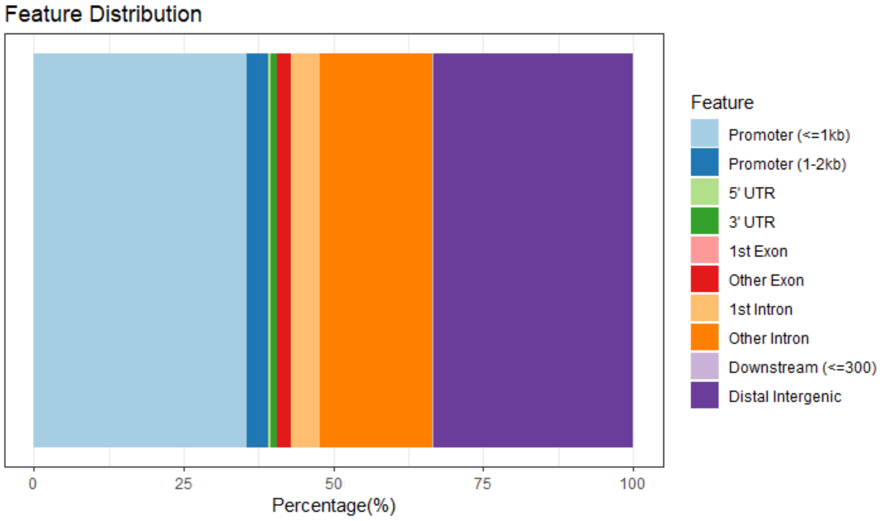


Figure 3: Distribution of transcription factor-binding loci relative to TSS.

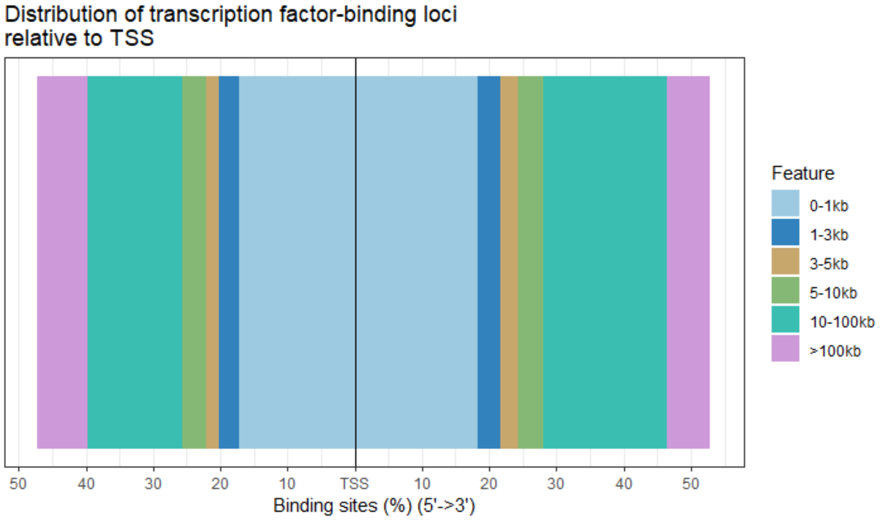
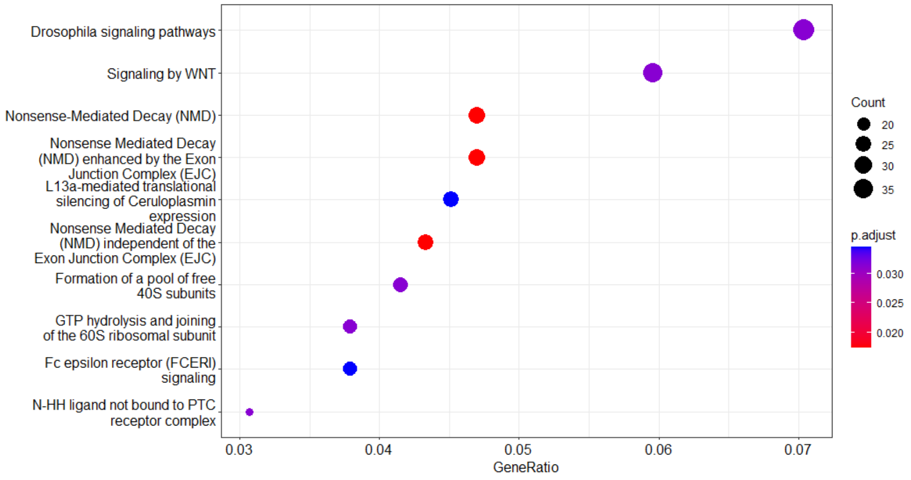


Figure 4: Enriched pathways



A Table of enriched pathways

ID	Description	GeneRatio	BgRatio	pvalue	p.a
R-DME-975956	Nonsense Mediated Decay (NMD) independent of the Exon Junction Complex (EJC)	24/554	85/4593	3.87e-05	0.0
R-DME-927802	Nonsense-Mediated Decay (NMD)	26/554	99/4593	7.28e-05	0.0
R-DME-975957	Nonsense Mediated Decay (NMD) enhanced by the Exon Junction Complex (EJC)	26/554	99/4593	7.28e-05	0.0
R-DME-5252538	Drosophila signaling pathways	39/554	184/4593	0.000233	0.0
R-DME-72706	GTP hydrolysis and joining of the 60S ribosomal subunit	21/554	78/4593	0.000246	0.0
R-DME-209446	N-HH ligand not bound to PTC receptor complex	17/554	58/4593	0.000320	0.0
R-DME-195721	Signaling by WNT	33/554	150/4593	0.000346	0.0
R-DME-72689	Formation of a pool of free 40S subunits	23/554	91/4593	0.000348	0.0
R-DME-2454202	Fc epsilon receptor (FCERI) signaling	21/554	82/4593	0.000516	0.0
R-DME-156827	L13a-mediated translational silencing of Ceruloplasmin expression	25/554	105/4593	0.000523	0.0
R-DME-8951664	Neddylation	31/554	141/4593	0.000527	0.0
R-DME-1799339	SRP-dependent cotranslational protein targeting to membrane	21/554	83/4593	0.000615	0.0
R-DME-72613	Eukaryotic Translation Initiation	26/554	113/4593	0.000712	0.0
R-DME-72737	Cap-dependent Translation Initiation	26/554	113/4593	0.000712	0.0
R-DME-209392	Hedgehog pathway	18/554	69/4593	0.00101	0.0