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 training[**meeta2023hbctraining**].

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**].

3 Analysis

3.1 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**].

3.2 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. Deeptools[**ramirez2014deeptools**] was used to visualize the signal. Figure 1

3.3 Annotation and functional enrichment analysis

3.4 Motif discovery and annotation

4 Conclusion

Future work: inclusion multi-mapped reads[**chung2011discovering**]

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.

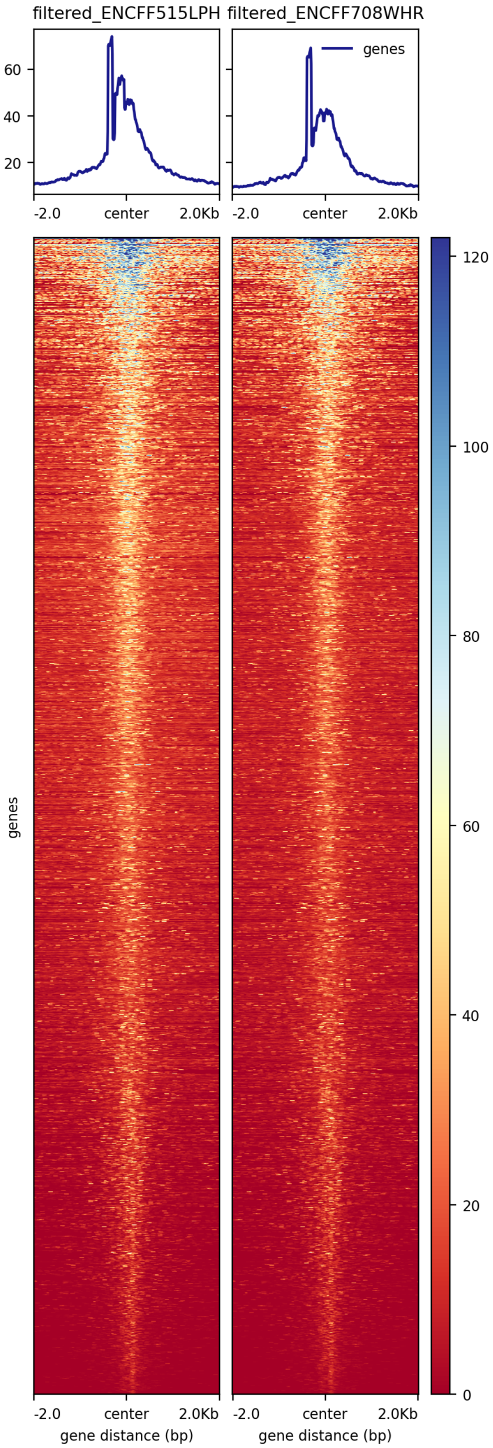
[3] Yang Liao, Gordon K Smyth, and Wei Shi. “The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads”. In: *Nucleic acids research* 47.8 (2019), e47–e47.

[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: Heatmap


filtered_ENCFF515LPH
filtered_ENCFF708WHR
genes
center
2.0Kb
center
2.0Kb
genes
center
2.0Kb
gene distance (bp)
center
2.0Kb
gene distance (bp)