Report of ATAC-seq Analysis: Effect of Ecc15-induced Immune Response in Drosophila on High-sugar Diet

Abstract

Infection with Erwinia carotovora carotovora 15 (Ecc15) has been shown to elicit an immune response in Drosophila melanogaster under normal dietary condition. However, its effects on high sugar diet states remain unclear. This study employed ATAC-seq analysis to investigate chromatin accessibility changes in high sugar Ecc15-infected and high sugar unchallenged groups. The results showed that while both groups presented metabolic adaptations to high sugar intake, differential analysis revealed immune response genes and pathways specifically enriched in the Ecc15-infected group, like Toll and Imd signaling pathways related to innate immune response. These findings provide insights into the molecular mechanisms underlying the immune response to Ecc15 infection in Drosophila under metabolic stress induced by high sugar diet.

Background

The dataset presented in this study was designed to investigate the impact of Erwinia carotovora carotovora 15 (Ecc15) infection on immune response and gene regulation in Drosophila Melanogaster under high sugar diet. Previous studies have shown that constant high sugar diets can lead to obesity, metabolic disorders, and altered cellular and humoral immune responses in Drosophila (Yu, Zhang and Jin, 2018). While the infection response to Ecc15 has been studied under normal diet conditions, it remains unclear how this response may be affected under a high sugar diet.

In current experiment, two groups were set: unchallenged (UC) group and Ecc15 infection (Ecc15) group. Both groups of fruit flies were fed with a high sugar diet for 5 days. Subsequently, the Ecc15 group received an Ecc15 infection for 8 hours, while the UC group remained untreated. After the infection period, the gut tissue from both groups was collected and subjected to ATAC-seq analysis. Each group had 3 replicates.

By investigating and comparing the ATAC-seq data between the UC and Ecc15 groups, this report aims to verify the characteristics of the high sugar diet and, more importantly, to test the hypothesis that the immune response sign will remain elevated in the Ecc15-infected group under high sugar diet. The findings from this study may contribute to the development of strategies for managing metabolic disorders and infectious diseases in the context of dietary factors.

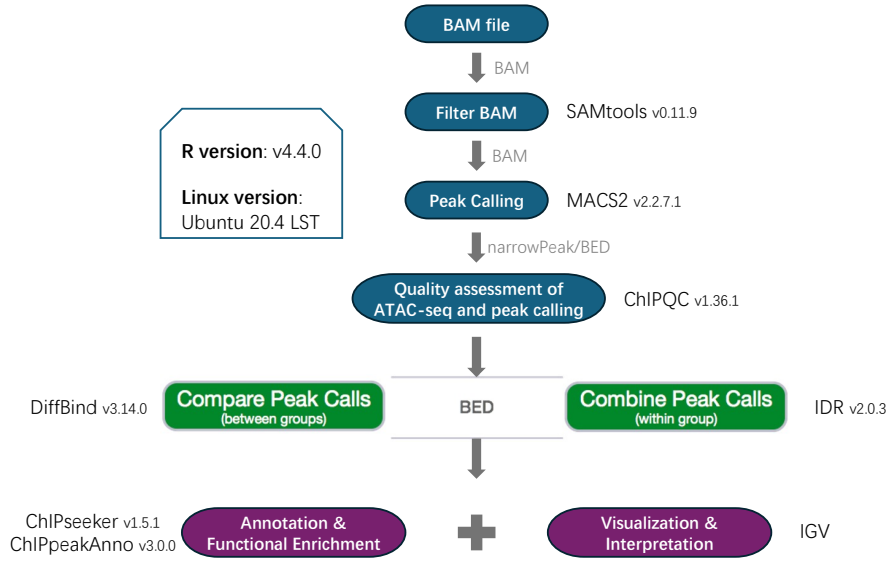


Figure 1 This schematic represents the comprehensive workflow employed for the ATAC-seq data analysis, encompassing key steps from BAM files with initial quality assessment to functional enrichment analysis. All the main file input and output and the software version has been clarified. More details please check Supplement Material -Code.

Sample	Genotype	Diet	Treatment	Replicate#	Renamed file
2	CantonS	HS	UC	2	HS_UC_2.bam
4	CantonS	HS	Ecc15 8hr	2	HS_Ecc15_2.bam
8	CantonS	HS	UC	3	HS_UC_3.bam
10	CantonS	HS	Ecc15 8hr	3	HS_Ecc15_3.bam
15	CantonS	HS	UC	1	HS_UC_1.bam
17	CantonS	HS	Ecc15 8hr	1	HS_Ecc15_1.bam

Table 1 Six files were downloaded for ATAC-seq analysis, including three replicates of high sugar-conditioned Drosophila melanogaster samples challenged with Ecc15 for 8 hours, and three replicates of unchallenged high sugar-conditioned Drosophila melanogaster samples. Sample indexes were 2, 4, 8, 10, 15, 17. For clear demonstration, the sample files were renamed according to their respective conditions (UC, Ecc15) and replicate numbers (1, 2, 3)

Methods

The overall pipeline described in this section contains quality assessment, quality control, peak calling, differential sites analysis, gene annotion, functional enrichment analysis, and ect (Figure 1). Core code sections has been released as supplement material for reproducibility.

Data

The data was accessed from the ZJE server. A file (For_QC_SampleB_1.fastq.gz) was used for quality control, and six BAM files were downloaded for ATAC-seq analysis (Table 1).

Quality Control of raw fastq file

FastQC v0.11.9 (version) was used for quality control with default parameters. The report was checked in the output HTML file.

Filtration of BAM

BAM files were processed with SAMtools v1.15.1 (Li et al., 2009). Statistical report generated by SAMtools revealed that the PCR duplication and mitochondrial reads had been removed. The files were filtered for properly mapped reads (-f 2) with mapping quality score ≥ 30 (-q 30) to improve reliability. The filtered files were sorted, indexed (BAI file), and prepared for peak calling.

Peak calling and related quality assessment

Open chromatin regions were identified using the peak-calling software MACS2 v 2.2.7.1 (Zhang et al., 2008). MACS2 is a widely used tool for identifying significant peaks from ChIP-seq and ATAC-seq.

The genome size was set to be dm (refer to drosophila melanogaster, -g dm). The 'nomodel' parameter was used, which prevents MACS2 from building a shifting model to account for potential biases in replicates. In this case, the '--nomodel' parameter was set to make different datasets comparable. Other parameters were left at their default values. The principal narrowPeak output file for downstream analysis contained peak locations, summits, p-values and q-values.

Before downstream analysis, overall quality of ATAC-seq including the peak calling results were assessed with ChIPQC v1.36.1, a R package for quality assessment, and deeptools v3.5.1, a powerful toolbox for ATAC-seq.

Differential sites/ regions analysis

The differential analysis was conducted with R package DiffBind v3.10.1, which could make use of replicates to reflect the significant different sites. The key method employed in this step was DESeq2, the default option. DiffBind employs a Bayesian approach to estimate the log-fold changes in expression intensity for each site across the different experimental conditions assigning a p-value and FDR. Default statistical thresholds were used, with results in a TXT file.

Merging replicates for overview

IDR v2.0.3 was a commonly used Java software to assess concordance of peak calls between replicates and functioned to statistically merge the replicates. A recommended stepwise merging process was implemented, producing a consensus result that incorporated all three replicates. Default IDR parameters were used.

a.

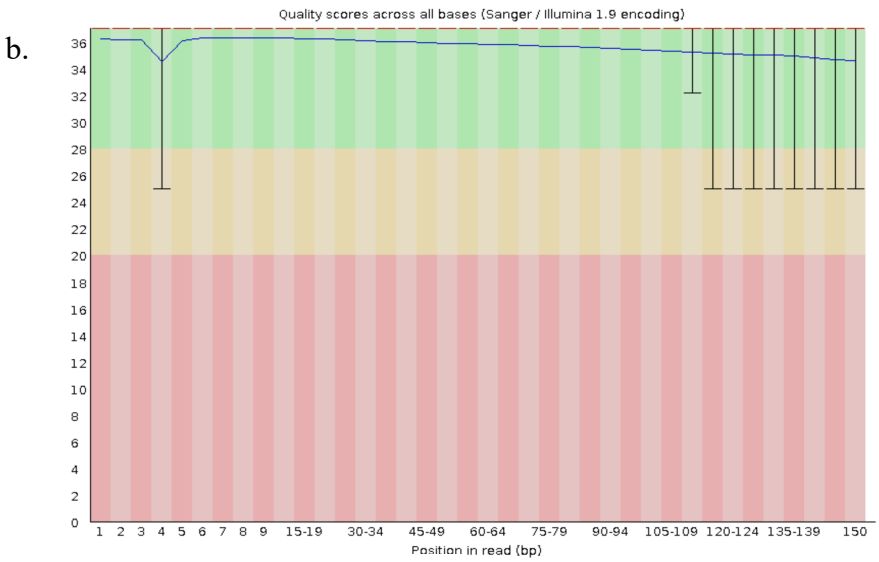
Summary

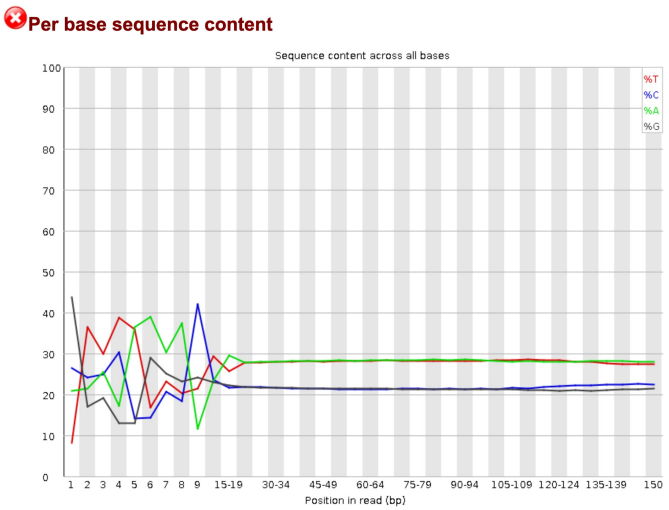
- Basic Statistics
- Per base sequence quality
- Per tile sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content

Basic Statistics

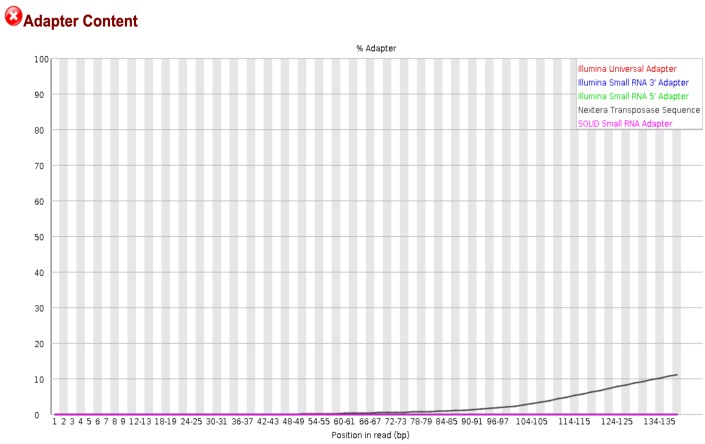
Measure	Value
Filename	For_QC_SampleB_1.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	13783174
Sequences flagged as poor quality	0
Sequence length	150
%GC	43

b.





c.



d.

Figure 2 Quality control assessment of raw For_QC_SampleB_1.fastq.gz file using FastQC. (a) Overview of quality control metrics. Most indicators met the assessment criteria, except 'Per base sequence content' and 'Adapter Content'. (b) 'Per base sequence quality' plot demonstrating a high overall sequencing quality score, within green and yellow sections. (c) Exploring the 'Per base sequence content' plot. The initial 20 base pairs were highlighted for the fluctuation. (d) Exploring the 'Adapter Content' plot. The black line increasing in the last period revealed indicating the presence of adapter sequences within the reads.

Peak annotation and functional enrichment analysis

Peaks can be annotated by the nearest genes or regulatory elements. ChIPseeker v1.40.0 were used for conducting annotation. ChIPseeker found the nearest genes around the peak, and statistically annotate the regions of the peaks. After annotation, functional enrichment analysis can be performed. Gene Ontology (GO) enrichment used clusterProfiler v4.12.0, while KEGG enrichment analysis utilized the STRING website. KEGG pathway visualization utilized Cytoscape app. Statistical visualizations conducted with R packages like ggplot2.

Results

Quality assessment

I. Quality control of the FASTQ file

The raw For_QC_SampleB_1.fastq.gz file was qualified as an example with FastQC. The report indicated that all quality indicators passed the assessment criteria, except for 'Per base sequence content' and 'Adapter Content' (Figure 2a). The basic statistics table in the report showed that the data had a sufficient number of total sequences, appropriate sequence length, and acceptable GC content, suggesting that the overall sequencing quality was satisfactory for downstream analysis (Figure 2a). The ‘Per base sequence quality’ showed relatively high score of the overall sequencing quality (Figure 2b).

To address the two unsatisfactory metrics, further preprocessing steps were considered. The 'Per base sequence content' metric revealed the initial 20 base pairs should be trimmed because of the low quality with fluctuation in base content (Figure 2c). The 'Adapter Content' metric indicated the presence of adapter sequences in the reads; therefore, adapter trimming should be performed (Figure 2d). To verify the understanding of quality control reports and implement the proposed preprocessing steps, the trimming tool Trimmomatic could be employed. Then the trimmed FASTQ file could align to genome with Bowtie2, to get the SAM file, which could be further converted into BAM file with SAMtools.

II. Quality assessment of overall ATAC-seq dataset and replicates

BAM files were processed correctly, statistical verifications were still needed. The overall mapping rate was all 100%, with about 99% of the reads being properly paired. Duplicates and mitochondrial reads were removed. However, a small fraction of reads were singletons, and few reads had mates mapped to a different chromosome. After filtering with higher-quality control, the resulting BAM file contained only properly paired reads with mapping quality ≥30.

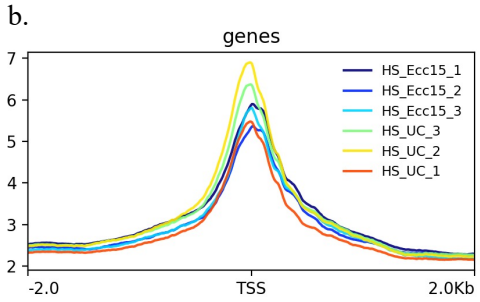
A detailed report (Figure 3a) showed reads per sample were in the normal range, with 24-43% reads in peaks (RiP%), indicating abundant alignment to potentially functional regions. ATAC-seq signal was enriched around transcription start sites (TSS) for all replicates (Figure

a.

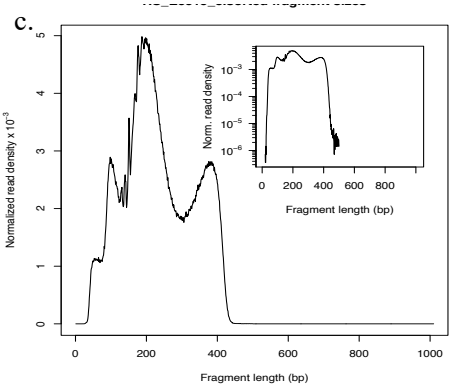
ID	Tissue	Factor	Condition	Replicate	Reads	Dup%	ReadL	FragL	RelCC	SSD	RiP%
HS_Ecc15_1			Ecc15	1	5040975	0	143	290	0.87	25	43
HS_Ecc15_2			Ecc15	2	742113	0	146	300	0.65	8.9	24
HS_Ecc15_3			Ecc15	3	956536	0	140	281	0.94	12	32
HS_UC_1			UC	1	6109971	0	142	285	0.84	25	36
HS_UC_2			UC	2	2555743	0	144	292	0.66	23	43
HS_UC_3			UC	3	4658450	0	144	289	0.79	29	43

- **ID** - Unique sample ID.
- **Tissue/Factor/Condition** - Metadata associated to sample.
- **Replicate** - Number of replicate within sample group
- **Reads** - Number of sample reads within analysed chromosomes.
- **Dup%** - Percentage of MapQ filter passing reads marked as duplicates
- **FragLen** - Estimated fragment length by cross-coverage method
- **SSD** - SSD score (ChIPQC)
- **FragLenCC** - Cross-Coverage score at the fragment length
- **RelativeCC** - Cross-coverage score at the fragment length over Cross-coverage at the read length
- **RiP%** - Percentage of reads within peaks
- **RiBL%** - Percentage of reads within Blacklist regions

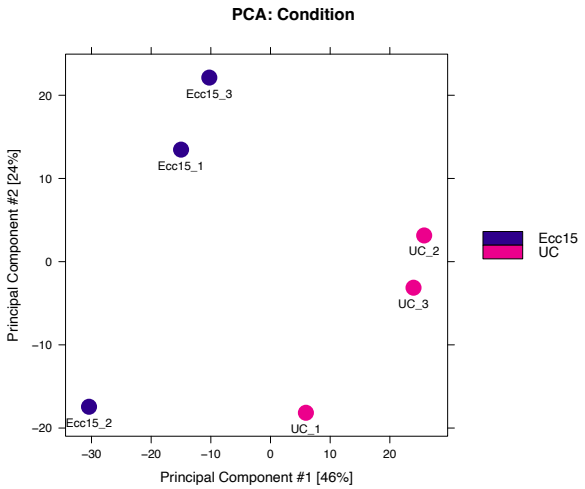
b.



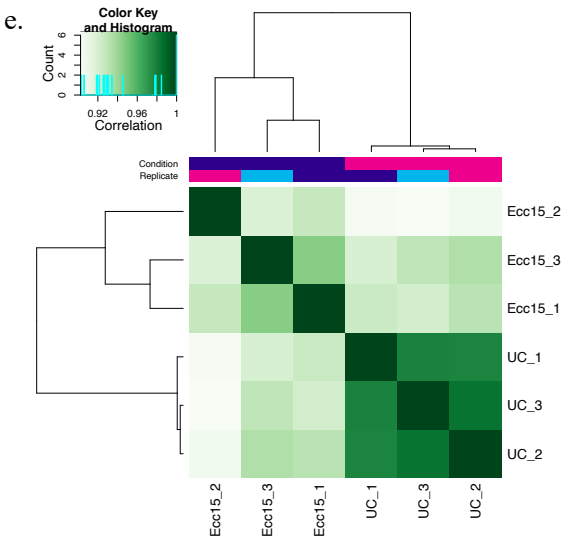
c.



d.



e.



f.

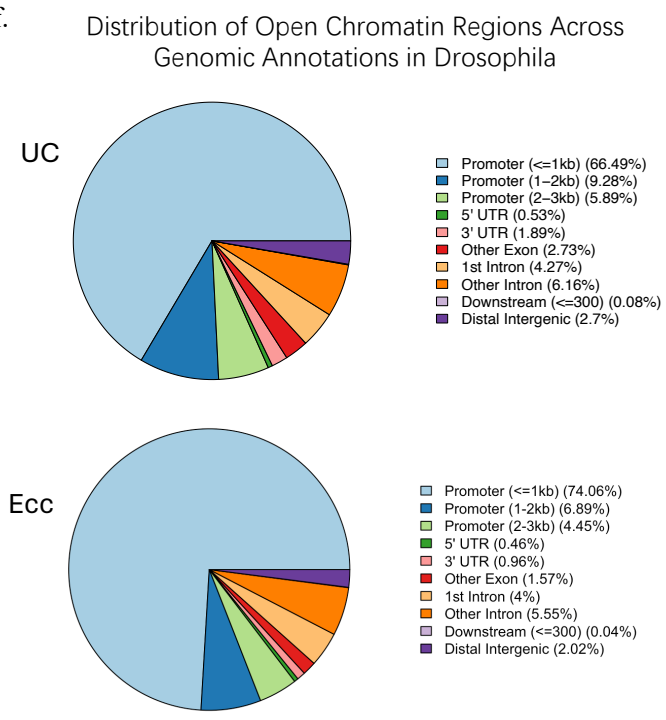


Figure 3 Comprehensive assessment of BAM file quality and replicate analysis. (a) The table summarized by ChIPQC. (b) The distribution of ATAC-seq signal, which showed enrichment around the transcription start sites (TSS) for all replicates. (c) Representative fragment size distribution of HS_Ecc15_3 sample. The distribution showed an expected peak around 200-300 bp for mononucleosome fragments and smaller peaks for polynucleosome fragments. Results from other samples were also check properly, showing in supplement figure. (d) PCA plot from DiffBind illustrating the clustering of replicates within and between groups, with Ecc15_2 acceptably separated, indicating relatively good replicate quality. (e) Pearson correlation heatmap showing the high similarity among replicates within each group and lower correlations between different conditioned group (f) Pie plots of genomic annotation distributions of overall merged groups.

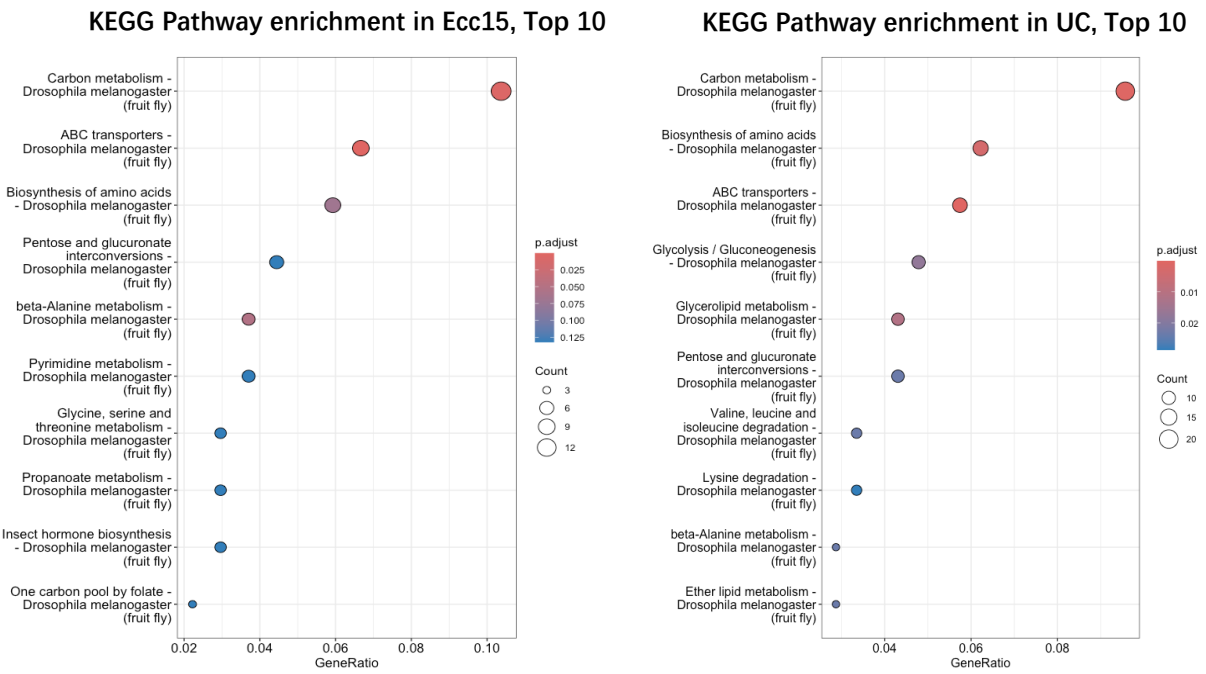


Figure 4 The overview of the top 10 (based on gene ratio) KEGG pathway enriched in each group, seperately. In both group, the metabolism alternations in multiple aspects were highlighted, which revealed the effect of high sugar diet.

3b), associated with active transcriptional regulation. Ditribution of fragment size is also a crucial consideration when assess ATAC-seq data (Figure 3c and Supplement Figure 1). The observed fragments fitted the favoable distribution with a prominent peak around 200-300 bp, representing mononucleosome fragments, with smaller peaks at higher fragment sizes corresponding to polynucleosome fragments (Yan, et al, 2020). Most patterns presented the absense of very short (<50 bp) and excessively long (>1000bp) fragments. These periodic patterns suggests successful capture of the chromatin accessibility landscape.

Replicate quality was verified using PCA plot and Pearson correlation heatmap from DiffBind (Figure 3e, 3f; Supplement Code). The PCA showed replicates clustered within groups, with Ecc15_2 acceptably separated on PC2 axis. The pearson correlation heatmap further confirms the higher similarity among replicates within each group, while lower correlations between different conditioned groups, as expected for comparative analysis. Overall, eplicate quality was appropriate for downstream merged group and differential accessibility analyses.

In summary, the comprehensive assessment evaluated overall data quality and replicate sample quality, covering mapping statistics, signal enrichment patterns, fragment size distributions, and replicate reproducibility.

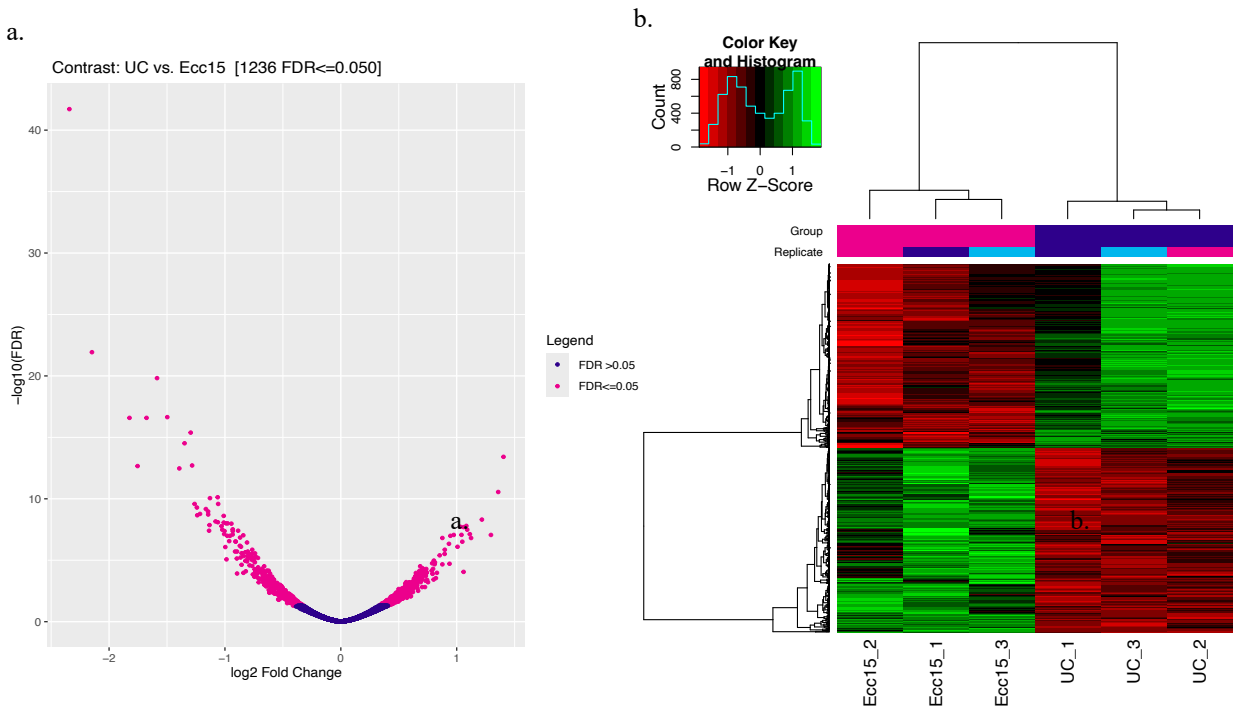


Figure 5 The discovery of differential regions from Ecc15 and UC groups. (a) The volcano plot illustrated the significance (FDR) and fold change of regions related to chromatin assecssibility across experimental conditions. The pink dots represented the differential regions with an FDR below the threshold of 0.05, indicating 1236 differntially accessible regions, while the blue dots represented the regions with the FDR higher than 0.05. (b) The heatmap displays the pattern of differential enrichment for the 1236 differntial regions. Red showed that lack of enrichment while the green section represented enrichment.

Differential Accessibility Analysis

I. Group-specific overview

Before conducting differential accessibility analysis, it was essential to check the overview situation of each group itself (group-specifc overview). Replicates within each group were merged using IDR appraoch (see Method). Merged narrwoPeak files of each group reveived gene annotation and KEGG pathway enrichment analysis.

Pie plots of genomic annotation distributions showed a predominance of promoters and introns in each group, suggesting strong regulatory potential with accessible regions near promoters and within introns (Figure 3g). Notably, the distributions between the two groups were similar and consistent with expected patterns.

KEGG analysis showed signs of high sugar diet effects (Figure 4). Both groups had significant carbon metabolism pathways, potentially compensating for increased sugar influx to maintain energy homeostasis. ABC transporters facilitating molecule transport across membranes were also highlighted, possibly aiding sugar/nutrient uptake and distribution. Other metabolism pathways were altered to fit the diet (Figure 4). High sugar diet indeed

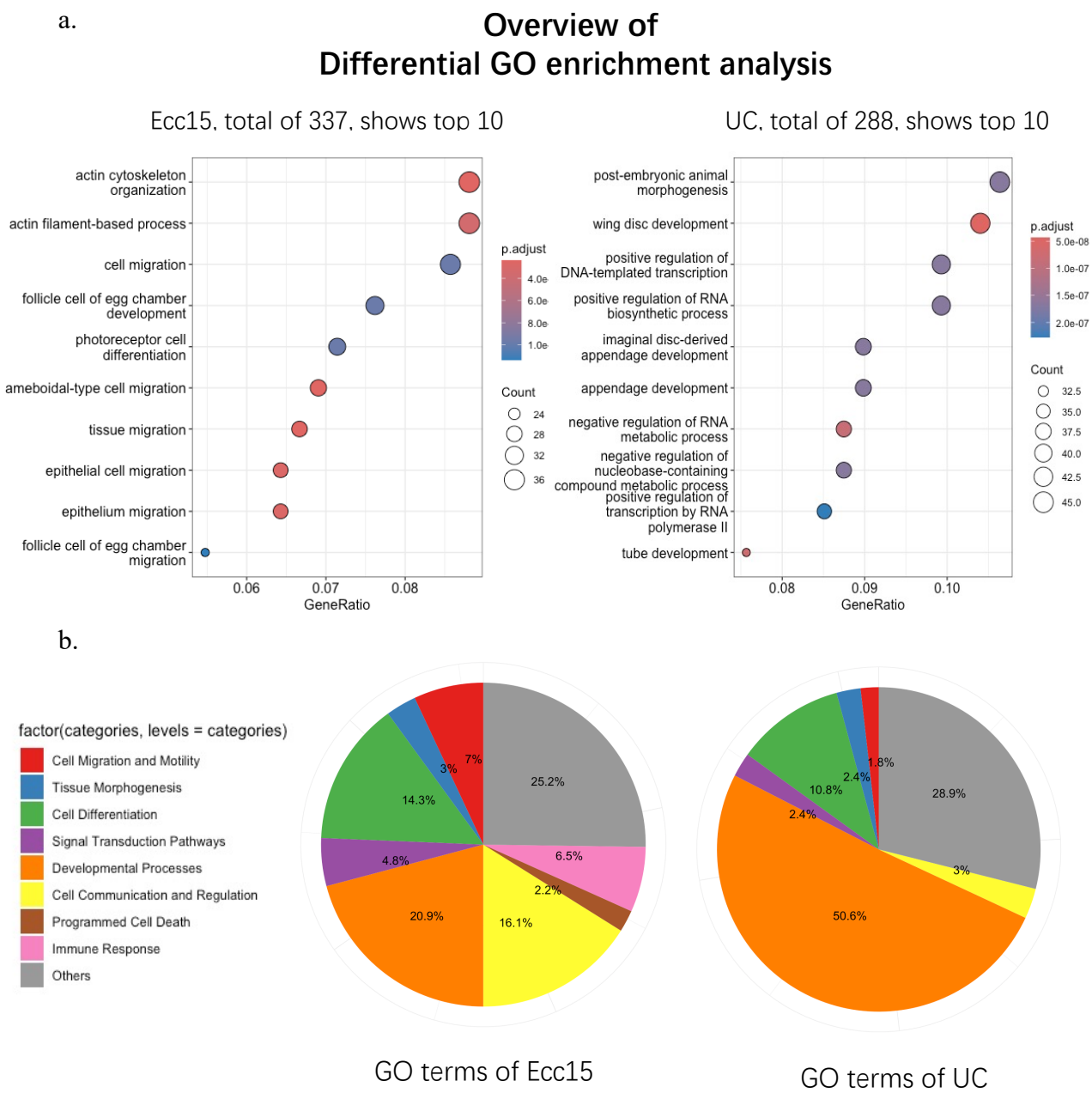


Figure 6 Overview of the result of differential genes GO enrichment analysis. (a) This figure presents the results of Gene Ontology (GO) enrichment analysis performed on the lists of differentially upregulated genes identified in each group. Top 10 GO terms in each group were visualized in bubble plots. In total, Ecc15 group contains 337 enriched GO terms, while UC group contains 288 enriched GO terms. (b) Summarized categories distribution of GO terms in each group, based on the description in GO website, so the categories could be rough (9 categories). Pie plots represented the results. Programmed cell death and immune response associated GO terms enriched in Ecc15 group while completely missed in UC group, which might indicate the alterative immune-related responses to Ecc15 infection.

functioned in fruit fly in both Ecc15-infected and normal state. However, whether it would effect the immune response within infected group still requried differential accessibility analysis

II. Differential gene expression identification

Differential chromatin accessibility regions were identified by comparing the replicates within each group against the other groups using DiffBind (See Method). False discovery rate (FDR) was utilized to define the different regions. Finally, sites were filtered using a pre-defined FDR threshold (0.05 here), and 1236 regions with an FDR below the threshold are defined as significantly differentially regions across the compared experimental conditions (Figure 5a). Upregulated regions have been particularly selected for each group. The heatmap plotted with

the differential regions revealed a clear pattern of differential enrichment, with distinct clusters of regions showing higher accessibility in one set of conditions compared to the others, which indicating the success of identifying differential regions (Figure 5b).

Identified differential regions were annotated with specific genes for further analysis (Supplement Figure). Until now, differential(upregulated) gene list of each group has been uncovered. An overview of GO enrichment analysis was performed on these differential gene lists to explore their functional categories (Figure 5c).

Categorizing the Go term of each group, results revealed that terms related to immune response showed in Ecc15 rather than UC group (Figure 6b). Other differences, like the enrichment of programmed cell death process in Ecc15 group, may also be worth to be deeply researched, yet immune response was the main subject we eagered to figure out.

III. Immune response elevated in Ecc15-infection group

KEGG pathway analysis was conducted to understand the roles of these differentialgenes in biological process (See Method). Three pathways with significant FDR were identified from the upregulated genes in the Ecc15 group, including the 'Toll and Imd signaling pathway' and the 'MAPK signaling pathway' (Figure 6a). MAPK signaling pathway is involoved in several regulation in proliferation, differntiation and survival in furit fly, which has been provided can be elevated when encontering survial pressure (Shilo, 2014). More importantly, upregulated Toll and Imd signaling pathway can activate an innate immune response in Drosophila melanogaster (Tanji et al., 2007). Therefore, this pathway was emerged as a pathway of particular interest, requring further in-depth analysis (Figure 6b).

The genes invloved in this pathway were selected in GO enrichment results, returing 157 GO terms containg at least one of these gene. From these selected GO terms, 15 immune response-related GO terms were highlighted, covering most of the immuno repsonse GO terms in overll Ecc15 upregulated gene (Figure 6d), suggesting their association with the Toll and Imd signaling pathway and specific immune response genes. To further verify the

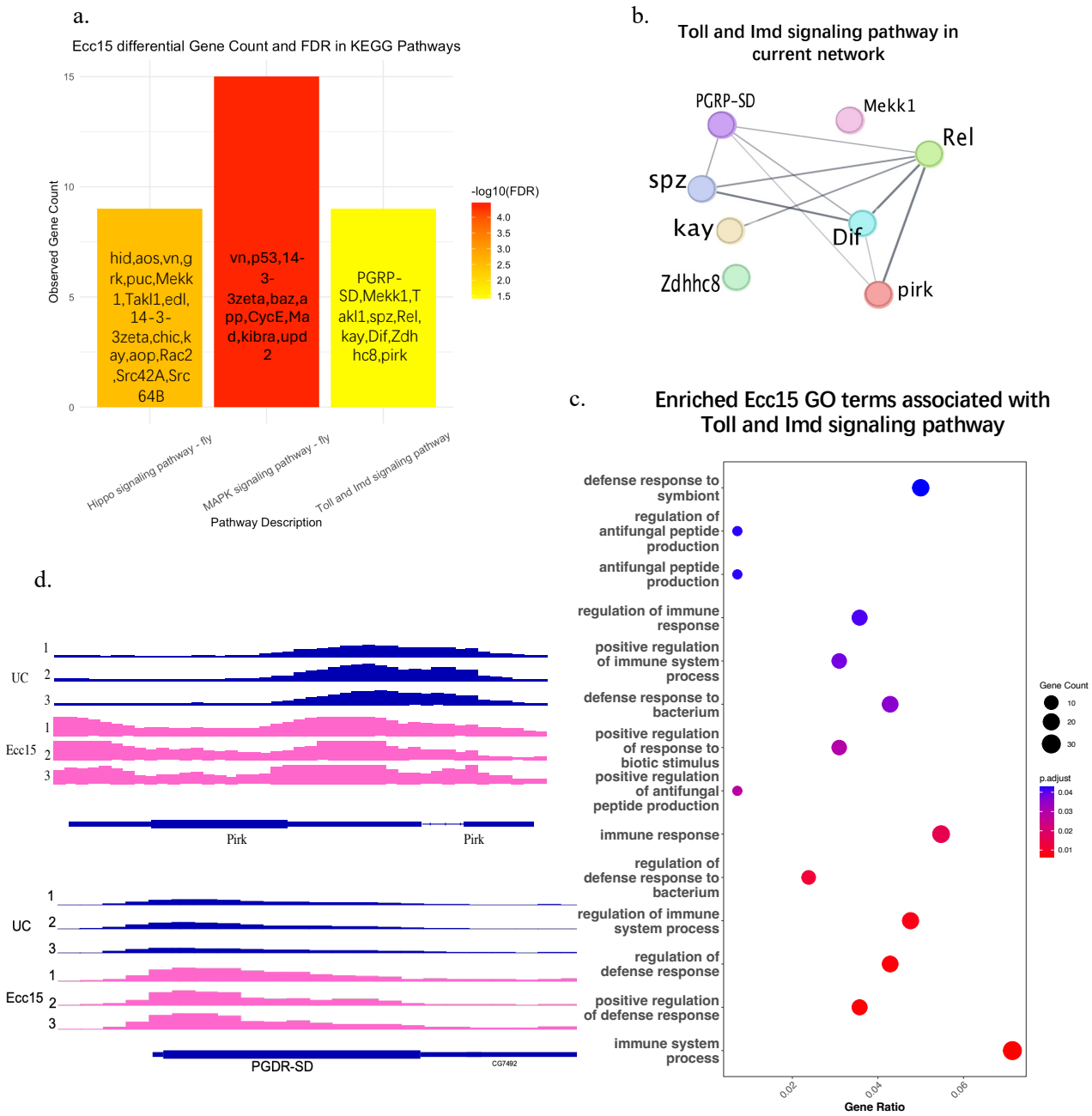


Figure 7 Activation of immune response pathways in Ecc15-infected Drosophila under high sugar diet. (a) KEGG pathway analysis results of upregulated genes in the Ecc15 group, using STRING website default search. The color represented –log (FDR), with the FDR all below the significance threshold 0.05. The pathway-associated genes found in Ecc15 upregulated gene list were texted in the bar (b) Identified schematic representation of the Toll and Imd signaling pathway from current Ecc15 upregulated genes networks. Generated with STRING website and Cytoscape. (c) GO enrichment analysis of immune response-related terms associated with genes in the Toll and Imd signaling pathway, in other word, the enriched Ecc15 GO terms associated with Toll and Imd signaling pathway. The size of the dots represents the number of genes in each GO term, and the color indicates the significance level (FDR). The genes in 7b were used to search in the list of differential GO term of Ecc15 group. (d) IGV browser visualization of two randomly selected genes (Pirk and PGDR-SD) from the Toll and Imd signaling pathway genes in 7b. The Ecc15 (pink) group exhibits increased chromatin accessibility compared to the UC group (blue), as evidenced by the higher signal intensity and the presence of well-defined peaks at the promoter and enhancer regions of these genes.

reliability, IGV brower was utilized to visulize two of the genes in the pathway, randomly, and both patterns represented the higher signal in Ecc15 group as expected (Figure 6e and 6f).

These findings strongly suggest that the immune system is activated in the Ecc15-infected condition under high sugar diet, potentially with the activated Toll and Imd signaling pathway. This observation supports the hypothesis that high sugar intake may induce an immune response in Drosophila during Ecc15 infection.

Discussion and Conclusions

The report demonstrates a systematic bioinformatics pipeline to analyze ATAC-seq data, focusing on the effect of Ecc15 infection on chromatin accessibility in Drosophila under high sugar diet. By employing quality assessment and quality, peak calling, and differential enrichment analysis with widely used tools like FastQC, SAMtools, MACS2, and DiffBind. Annotation followed by GO and KEGG analyses provided a biological-meaningful genomic landscape for further investigation.

The discovery revealed significant activation of immune response, particularly the Toll and Imd signaling pathways, in the Ecc-15-infected group, while the high sugar diet also caused metabolism alternation in all replicates. This suggests a complex regulatory mechanism linking dietary sugar, infection, and immune response, yet the detailed correlation approaches remained unclear. Further research could extend to motif enrichment analysis, footprints discovery and integration with multiomics data.

In conclusion, the ATAC-seq analysis revealed the effect of Ecc15-induced immune response in Drosophila on high-sugar diet, with highlighted innate immune response Toll and Imd signaling pathway.

References

Gáliková, M. and Klepsatel, P. (2018) ‘Obesity and Aging in the Drosophila Model’, International Journal of Molecular Sciences, 19(7), p. 1896. Available at: https://doi.org/10.3390/ijms19071896.

Li, H. et al. (2009) ‘The Sequence Alignment/Map format and SAMtools’, Bioinformatics, 25(16), pp. 2078–2079. Available at: https://doi.org/10.1093/bioinformatics/btp352.

Shilo, B.-Z. (2014) ‘The regulation and functions of MAPK pathways in Drosophila’, Methods, 68(1), pp. 151–159. Available at: https://doi.org/10.1016/j.ymeth.2014.01.020.

Tanji, T. et al. (2007) ‘Toll and IMD Pathways Synergistically Activate an Innate Immune Response in Drosophila melanogaster’, Molecular and Cellular Biology, 27(12), pp. 4578–4588. Available at: https://doi.org/10.1128/MCB.01814-06.

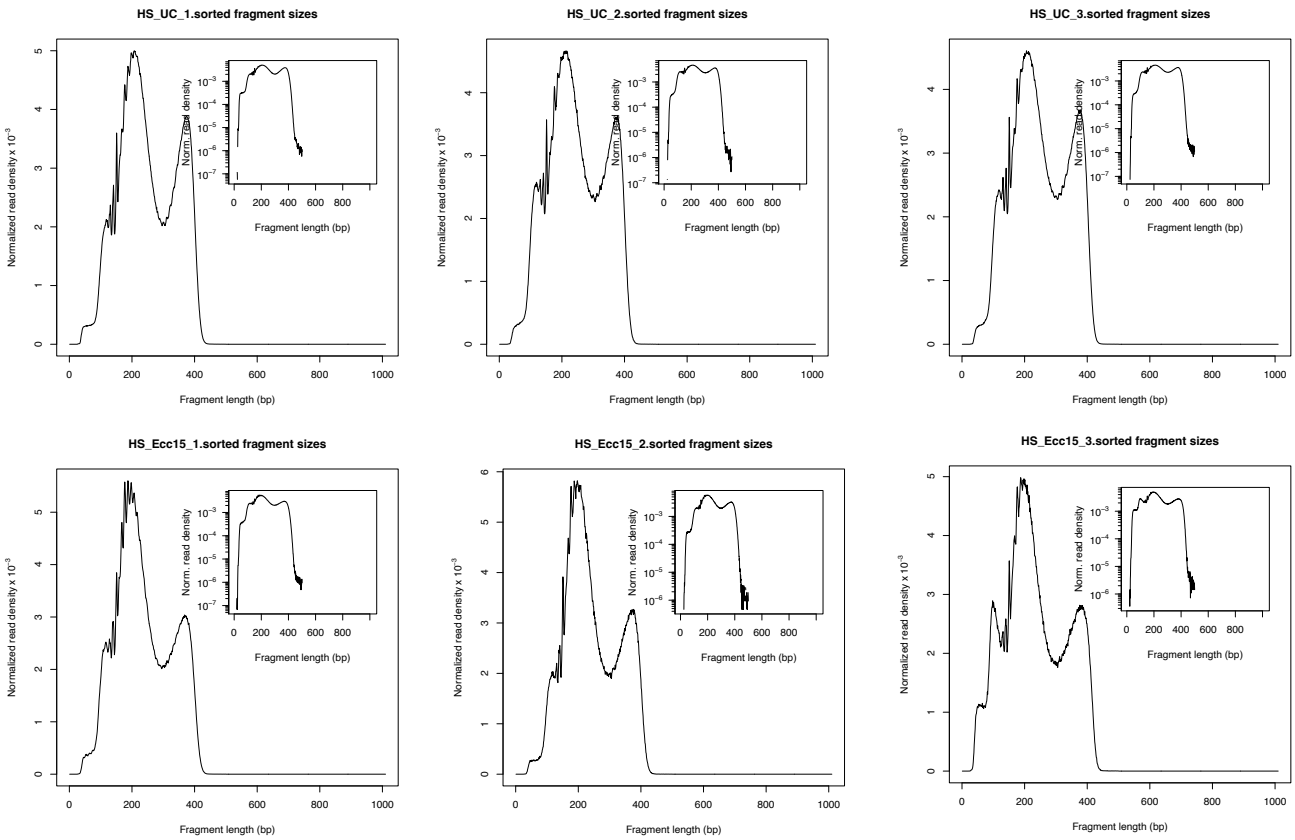
Yan, F. et al. (2020) ‘From reads to insight: a hitchhiker’s guide to ATAC-seq data analysis’, Genome Biology, 21, p. 22. Available at: https://doi.org/10.1186/s13059-020-1929-3.

Yu, S., Zhang, G. and Jin, L.H. (2018) ‘A high-sugar diet affects cellular and humoral immune responses in Drosophila’, Experimental Cell Research, 368(2), pp. 215–224. Available at: https://doi.org/10.1016/j.yexcr.2018.04.032.

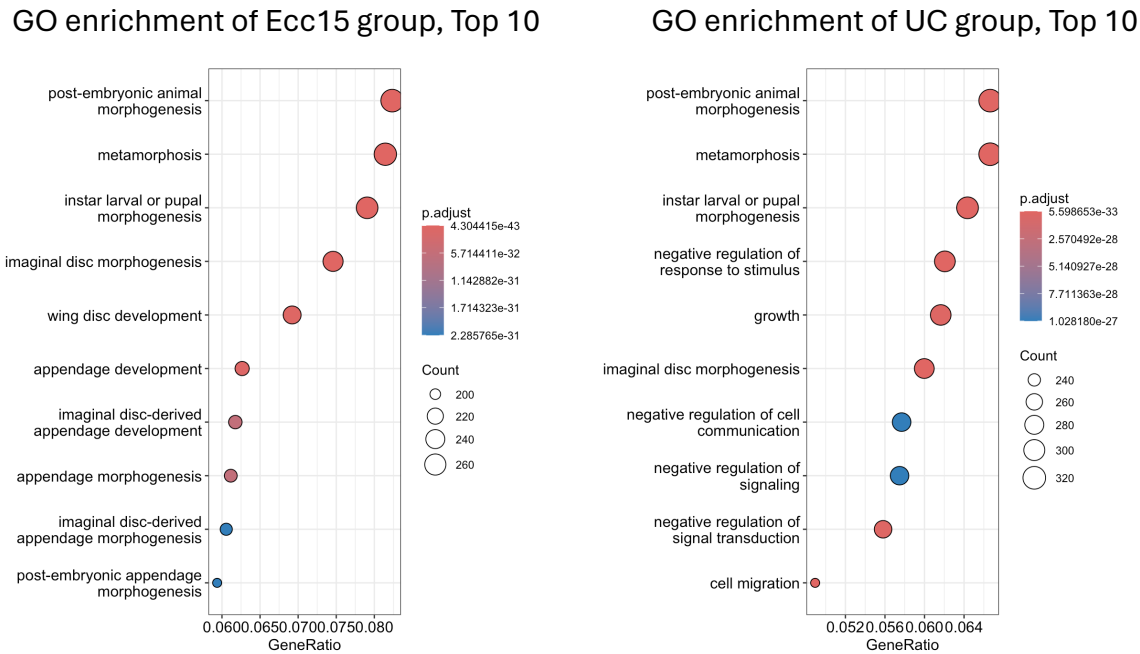
Zhang, Y. et al. (2008) ‘Model-based Analysis of ChIP-Seq (MACS)’, Genome Biology, 9(9), p. R137. Available at: https://doi.org/10.1186/gb-2008-9-9-r137.

Supplement Material -Figures

a.



b.



Supplement figure (a) The distribution of fragment size of all the samples. Patterns were all fits the requirement. (b) The Top 10 GO term enrichment of merged UC and Ecc15 groups, based on gene ratio rank.

Supplement Material -Code

Quality Control, FastQC

```shell
# Run FastQC for quality control on a FASTQ file
fastqc <fastq_file> -o <output_directory>
```

ATAC-seq pre-process, SAMtools

```
raw_dir="<raw_data_directory>"
proc_dir="<processed_data_directory>"
out_dir="<output_directory>"

# Loop through all BAM files in the raw directory
for id in "${raw_dir}"/*.bam; do
echo "Processing file: $id"
sample=$(basename "$id" .bam)
mapped_bam="${proc_dir}${sample}.mapped.bam"
sorted_bam="${out_dir}${sample}.sorted.bam"
samtools flagstat "$id" > "${proc_dir}${sample}.flagstat"
samtools view -h -f 2 -q 30 "$id" > "$mapped_bam"
samtools sort "$mapped_bam" -o "$sorted_bam"
rm "$mapped_bam"
samtools flagstat "$sorted_bam" >
"${out_dir}${sample}.sorted.flagstat"
samtools index "$sorted_bam"
bedtools bamtobed -i "$sorted_bam" > "${out_dir}${sample}.last.bed"
done
```

Peak calling, MACS2

```
# Add Anaconda to the PATH
export PATH=<anaconda_bin_directory>:$PATH
cd <working_directory>
# Loop through all sorted BAM files and call peaks using MACS2
ls *.sorted.bam | while read id; do
echo $id
macs2 callpeak -t $id -g dm --nomodel --shift -100 --extsize 200 -n
$id --outdir .
done
```

ATAC quality assessment, ChIPQC

```
library(ChIPQC)
# This sample sheet stores the metadata of the peak calling and BAM
information
samples <- read.csv('<samplesheet_file>')
## Create ChIPQC object
chipObj <- ChIPQC(samples, annotation="dm3")
## Create ChIPQC report
ChIPQCreport(chipObj, reportName="ChIP QC report", reportFolder="
<output_directory>")
```

ATAC quality assessment, fragment size distribution, ATACseqQC

```
library(ATACseqQC)
bamfile <- "HS_Ecc15_1.sorted.bam"
bamfile.labels <- gsub(".bam", "", basename(bamfile))

fragSize <- fragSizeDist(bamfile, bamfile.labels)
```

ATAC quality assessment, signal around TSS, deeptools

```
computeMatrix reference-point  --referencePoint TSS  --
missingDataAsZero  -b 2000 -a 2000  -R tft_ref.bed  -S HS_Ecc15_1.bw
HS_Ecc15_2.bw HS_Ecc15_3.bw HS_UC_3.bw HS_UC_2.bw HS_UC_1.bw  -o
TSS.gz

plotProfile -m TSS.gz --plotFileFormat pdf --perGroup
```

Differential enrichment analysis, DiffBind

```
library(DiffBind)

# Load the sample sheet containing information about the samples
db_data <- dba(sampleSheet = "<sample_sheet_file>")

# Count reads for each sample and perform summarization
db_data <- dba.count(db_data , bUseSummarizeOverlaps=TRUE)
db_data

# Perform differential binding analysis using DESeq2 method
db_data <- dba.analyze(db_data, method = DBA_DESEQ2)
```

```
# Generate a PCA plot and save it as a PDF file
pdf("<output_file_pca>.pdf")
dba.plotPCA(db_data,  attributes=DBA_CONDITION, label=DBA_ID)
plot(db_data)
dev.off()


# Set up contrasts between conditions, considering only conditions
with at least 2 members
db_data <- dba.contrast(db_data,categories=DBA_CONDITION,minMembers
= 2)

# Perform differential binding analysis on the contrasts using
DESeq2 method
db_data <- dba.analyze(db_data, method=DBA_DESEQ2)

# Show a summary of the differential binding results with a
threshold of 0.1
dba.show(db_data,bContrasts=T,th=0.1)

# Generate a report of the differential binding results for the
first contrast
db_data.DB <- dba.report(db_data,contrast= 1)

# Convert the report to a data frame
out <- as.data.frame(db_data.DB)

# Save the report as a tab-separated text file
write.table(out, file="<output_file_result>.txt", sep="\t", quote=F,
row.names=F)

# Save the report as a text file (redundant with the previous line)
write.table(db_data.DB,"<output_file_result>.txt")

# Generate a heatmap of the read scores and save it as a PDF file
pdf("<output_file_heatmap>.pdf")
hmap <- colorRampPalette(c("red", "black", "green"))(n = 13)
readscores <- dba.plotHeatmap(db_data,
contrast=1,correlations=FALSE,scale="row",colScheme=hmap)
dev.off()

# Generate a volcano plot of the differential binding results and
save it as a PDF file
pdf("<output_file_volcano>.pdf")
dba.plotVolcano(db_data,contrast=1)
```

```
dev.off()

library(tidyverse)

# Filter the results for enriched regions with FDR < 0.05 and
positive fold change
UC_enrich <- out %>%
  filter(FDR < 0.05 & Fold > 0) %>%
  select(seqnames, start, end)

# Save the enriched regions as a BED file
write.table(UC_enrich, file="<output_file_uc_diff>.bed", sep="\t",
quote=F, row.names=F, col.names=F)

# Filter the results for enriched regions with FDR < 0.05 and
negative fold change
Ecc_enrich <- out %>%
  filter(FDR < 0.05 & Fold < 0) %>%
  select(seqnames, start, end)

# Save the enriched regions as a BED file
write.table(Ecc_enrich, file="<output_file_ecc_diff>.bed", sep="\t",
quote=F, row.names=F, col.names=F)
```

Merging replicates, IDR

```
idr --samples <peak_file_1> <peak_file_2> \
    --input-file-type narrowPeak \
    --rank p.value \
    --output-file <output_file> \
    --plot \
    --log-output-file <log_file>
```

Peak annotation and functional enrichment analysis

```
# Define a function to perform ChIP-seq data annotation and GO
enrichment analysis
chipseq_analysis <- function(peak_files, txdb, org_db, output_dir) {
  library(ChIPseeker)
  library(TxDb.Dmelanogaster.UCSC.dm6.ensGene)
  library(clusterProfiler)

  for (peak_file in peak_files) {
    # Read peak file
```

```
    peak <- readPeakFile(peak_file)

    # Add "chr" prefix to chromosome names
    seqlevels(peak) <- paste0("chr", seqlevels(peak))

    # Remove non-standard chromosomes
    keepChr <- !grepl('_', seqlevels(peak))
    seqlevels(peak, pruning.mode = "coarse") <- seqlevels(peak)
[keepChr]

    # Annotate peaks
    peakAnno <- annotatePeak(peak, tssRegion = c(-3000, 3000), TxDb
= txdb, annoDb = org_db)
    peakAnno_df <- as.data.frame(peakAnno)

    # Save annotation results
    anno_output_file <- file.path(output_dir,
paste0(basename(peak_file), "_anno.csv"))
    write.csv(peakAnno_df, anno_output_file, row.names = FALSE)

    # GO enrichment analysis
    entrezids <- peakAnno_df$ENTREZID %>% as.character() %>%
unique()
    ego <- enrichGO(gene = entrezids, keyType = "ENTREZID", OrgDb =
get(org_db), ont = "BP", pAdjustMethod = "BH", qvalueCutoff = 0.05,
readable = TRUE)

    # Save GO enrichment analysis results
    go_output_file <- file.path(output_dir,
paste0(basename(peak_file), "_GO.csv"))
    cluster_summary <- data.frame(ego)
    write.csv(cluster_summary, go_output_file)
  }
}

# Specify peak files
peak_files <- c("path/to/Ecc_diff.bed")

# Run ChIP-seq analysis
result <- chipseq_analysis(peak_files = peak_files,
                           txdb =
TxDb.Dmelanogaster.UCSC.dm6.ensGene,
                           org_db = "org.Dm.eg.db",
                           output_dir = "path/to/output/directory")
```

Example of ggplot2 visualization

```
# Load the required package
library(ggplot2)

# Data
my_colors <- c("#E41A1C", "#377EB8", "#4DAF4A", "#984EA3",
"#FF7F00", "#FFFF33", "#A65628", "#F781BF", "#999999")

categories <- c("Cell Migration and Motility", "Tissue
Morphogenesis", "Cell Differentiation",
                "Signal Transduction Pathways", "Developmental
Processes", "Cell Communication and Regulation",
                "Programmed Cell Death", "Immune Response",
"Others")
counts <- c(26, 19, 20, 21, 21, 9, 8, 15, 18)
percentages <- round(counts/sum(counts)*100, 1)
data <- data.frame(categories, counts, percentages)

# Pie Chart
ggplot(data, aes(x = "", y = counts, fill = categories)) +
  geom_bar(width = 1, stat = "identity") +
  coord_polar("y", start = 0) +
  labs(title = "Distribution of GO Terms",
       x = NULL, y = NULL) +
  geom_text_repel(aes(label = paste0(percentages, "%")),size =8,
position = position_stack(vjust = 0.5)) +
  theme_minimal() +
  theme(axis.text = element_blank(),
        axis.ticks = element_blank(),
        plot.title = element_text(hjust = 1),
        legend.position = "right")
```