# RNA-Seq Analysis of Tumor Resistance in Knockout (KT) vs Wild-Type (WT) Mice

- Lohit Marla

# Agenda

1. Steps Accomplished
2. Data Quality Reports
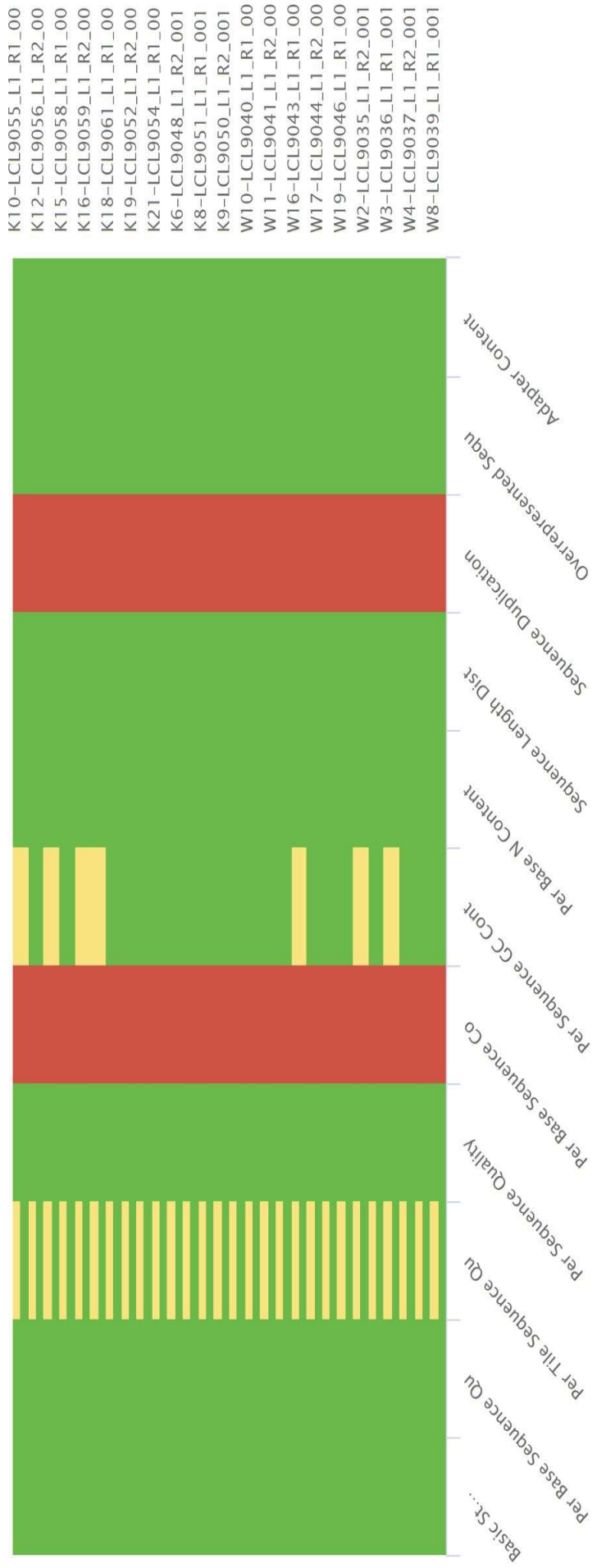3. Results
4. Current Progress
5. Achievements

# Steps Accomplished - High Level

1. Transferred 757 GB Data to HPC Clusters
2. Unzipping and FastQC, MultiQc reports
3. Trimming using multiple techniques
4. Verifying with data quality reports
5. Read Alignment using STAR
6. Collected statistics of Bam using Samtools
7. Generated featurecounts by GRCm39
8. Loaded the data into Rstudio
9. Processed the raw counts for analysis
10. Applied Deseq2
11. Filtered significant genes
12. Visualized the genes
13. Performed the Gene Ontology
14. Applied Kegg Pathway Analysis

# Data Quality Checks - Raw Data
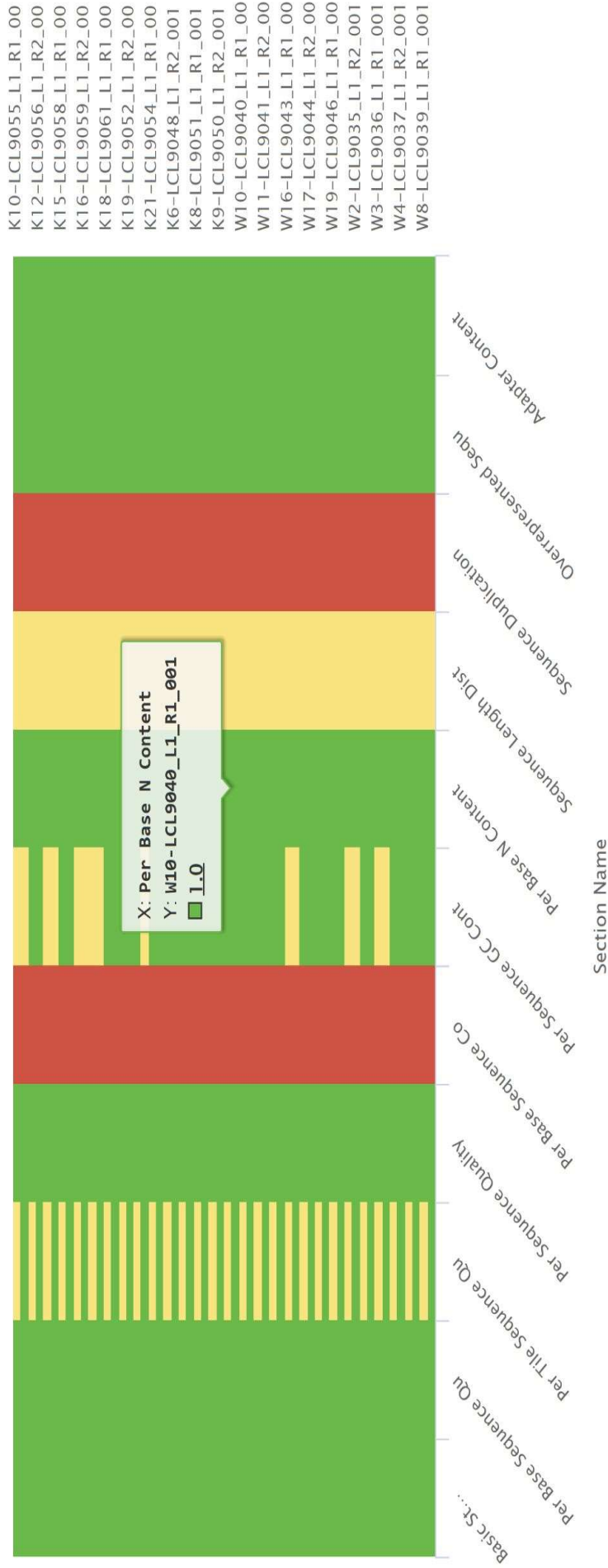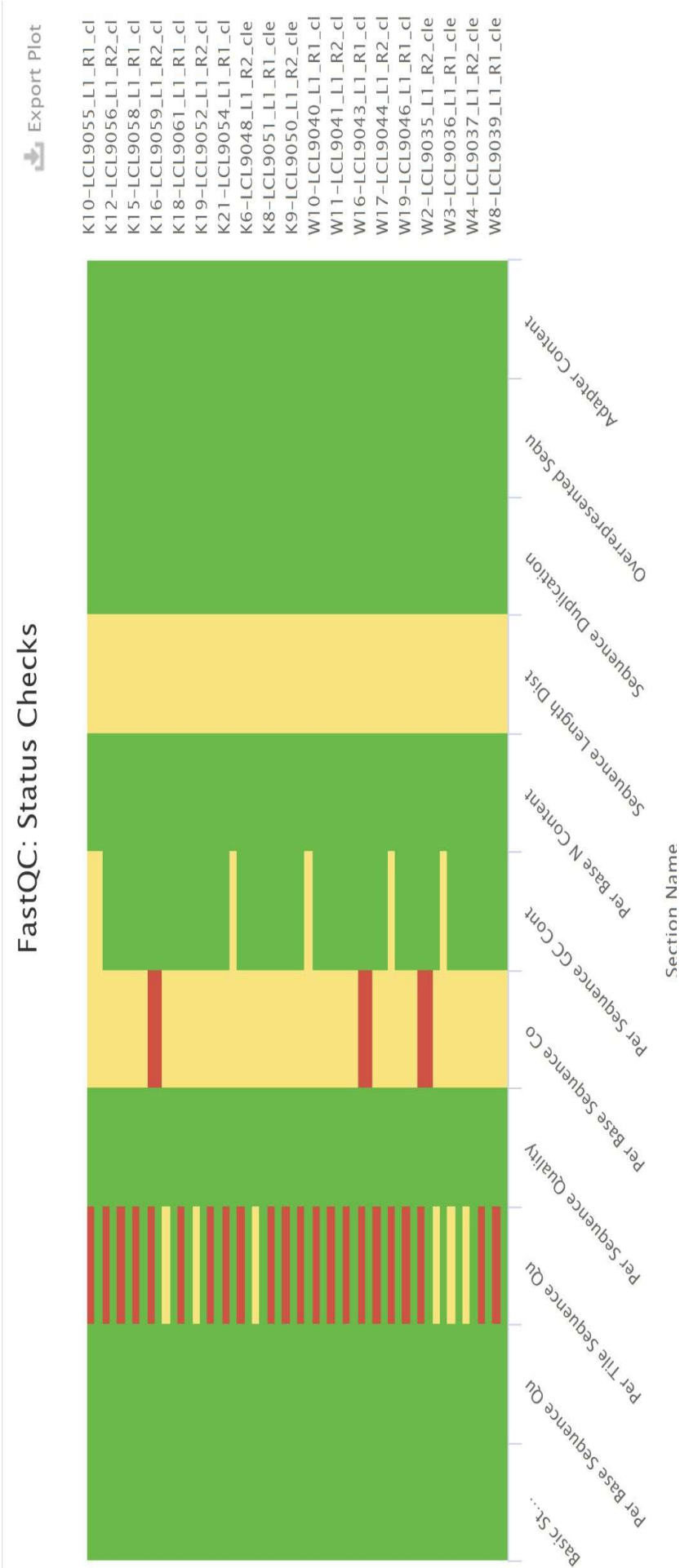
## FastQC: Status Checks

# Data Quality Checks - Dedup Reads

FastQC: Status Checks

Sample labels:
K10-LCL9055_L1_R1_00
K12-LCL9056_L1_R2_00
K15-LCL9058_L1_R1_00
K16-LCL9059_L1_R2_00
K18-LCL9061_L1_R1_00
K19-LCL9052_L1_R2_00
K21-LCL9054_L1_R1_00
K6-LCL9048_L1_R2_001
K8-LCL9051_L1_R1_001
K9-LCL9050_L1_R2_001
W10-LCL9040_L1_R1_00
W11-LCL9041_L1_R2_00
W16-LCL9043_L1_R1_00
W17-LCL9044_L1_R2_00
W19-LCL9046_L1_R1_00
W2-LCL9035_L1_R2_001
W3-LCL9036_L1_R1_001
W4-LCL9037_L1_R2_001
W8-LCL9039_L1_R1_001

Section Name:
Adapter Content
Overrepresented Sequ
Sequence Duplication
Sequence Length Dist
Per Base N Content
Per Sequence GC Cont
Per Base Sequence Co
Per Sequence Quality
Per Tile Sequence Qu
Per Base Sequence Qu
Basic St...

Tooltip:
X: Per Base N Content
Y: W10-LCL9040_L1_R1_001
■ 1.0

# Data Quality Checks - Trimmomatic Min Len - 80%

FastQC: Status Checks



Export Plot

Section Name

Columns (top labels):
K10-LCL9055_L1_R1_cl
K12-LCL9056_L1_R2_cl
K15-LCL9058_L1_R1_cl
K16-LCL9059_L1_R2_cl
K18-LCL9061_L1_R1_cl
K19-LCL9052_L1_R2_cl
K21-LCL9054_L1_R1_cl
K6-LCL9048_L1_R2_cle
K8-LCL9051_L1_R1_cle
K9-LCL9050_L1_R2_cle
W10-LCL9040_L1_R1_cl
W11-LCL9041_L1_R2_cl
W16-LCL9043_L1_R1_cl
W17-LCL9044_L1_R2_cl
W19-LCL9046_L1_R1_cl
W2-LCL9035_L1_R2_cle
W3-LCL9036_L1_R1_cle
W4-LCL9037_L1_R2_cle
W8-LCL9039_L1_R1_cle

Row labels (section names):
Adapter Content
Overrepresented Sequ
Sequence Duplication
Sequence Length Dist
Per Base N Content
Per Sequence GC Cont
Per Base Sequence Co
Per Sequence Quality
Per Tile Sequence Qu
Per Base Sequence Qu
Basic St...

# Data Quality Check - Filtered Reads



FastQC: Status Checks

Export Plot

K21-LCL9054_L1_R1_fi

K21-LCL9054_L1_R2_fi

Section Name

Adapter Content
Overrepresented Sequ
Sequence Duplication
Sequence Length Dist
Per Base N Content
Per Sequence GC Cont
Per Base Sequence Co
Per Sequence Quality
Per Tile Sequence Qu
Per Base Sequence Qu
Basic St...
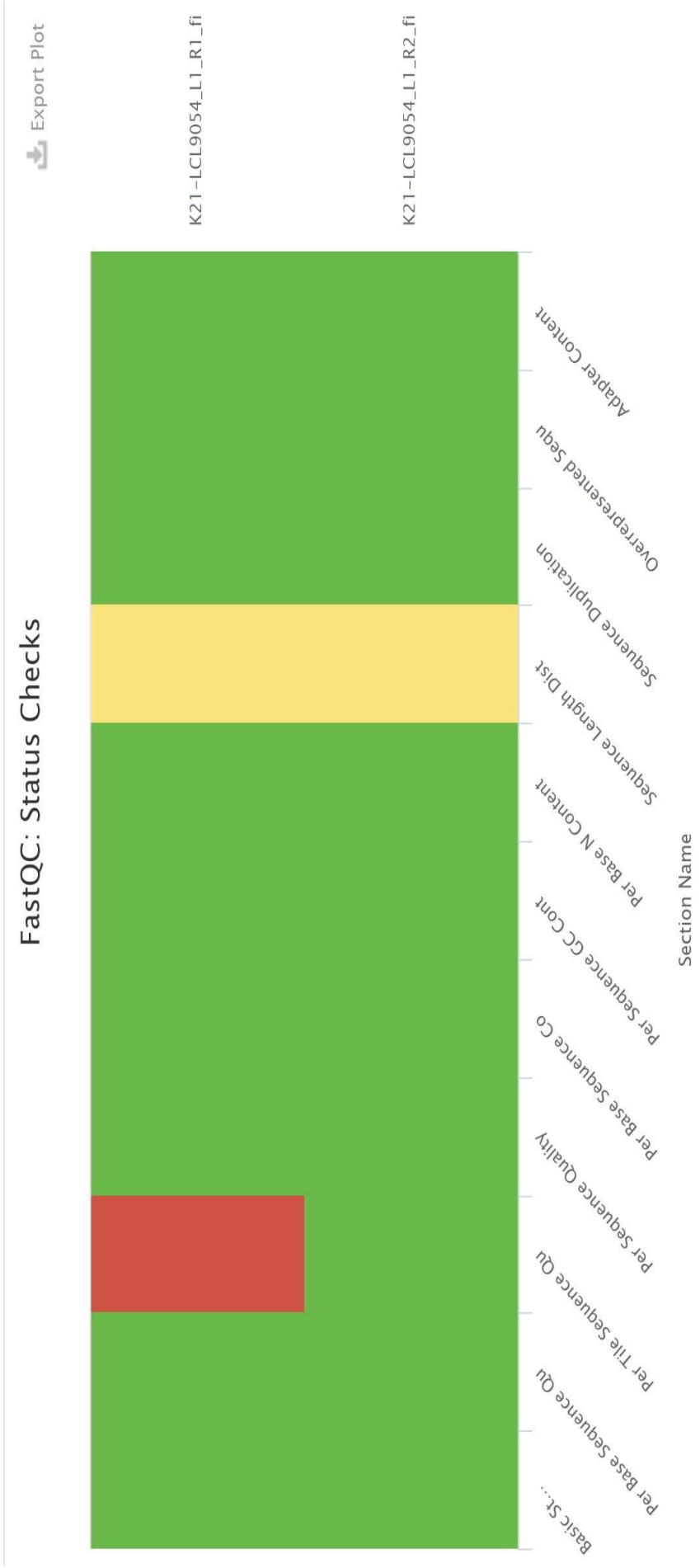
# Data Quality Check - Recalibrated Reads

## FastQC: Status Checks

K21-LCL9054_L1_R1_re

Adapter Content

Overrepresented Sequ

Sequence Duplication

Sequence Length Dist

Per Base N Content

Per Sequence GC Cont

Per Base Sequence Co

Per Sequence Quality

Per Tile Sequence Qu

Per Base Sequence Qu

Basic St...

Section Name
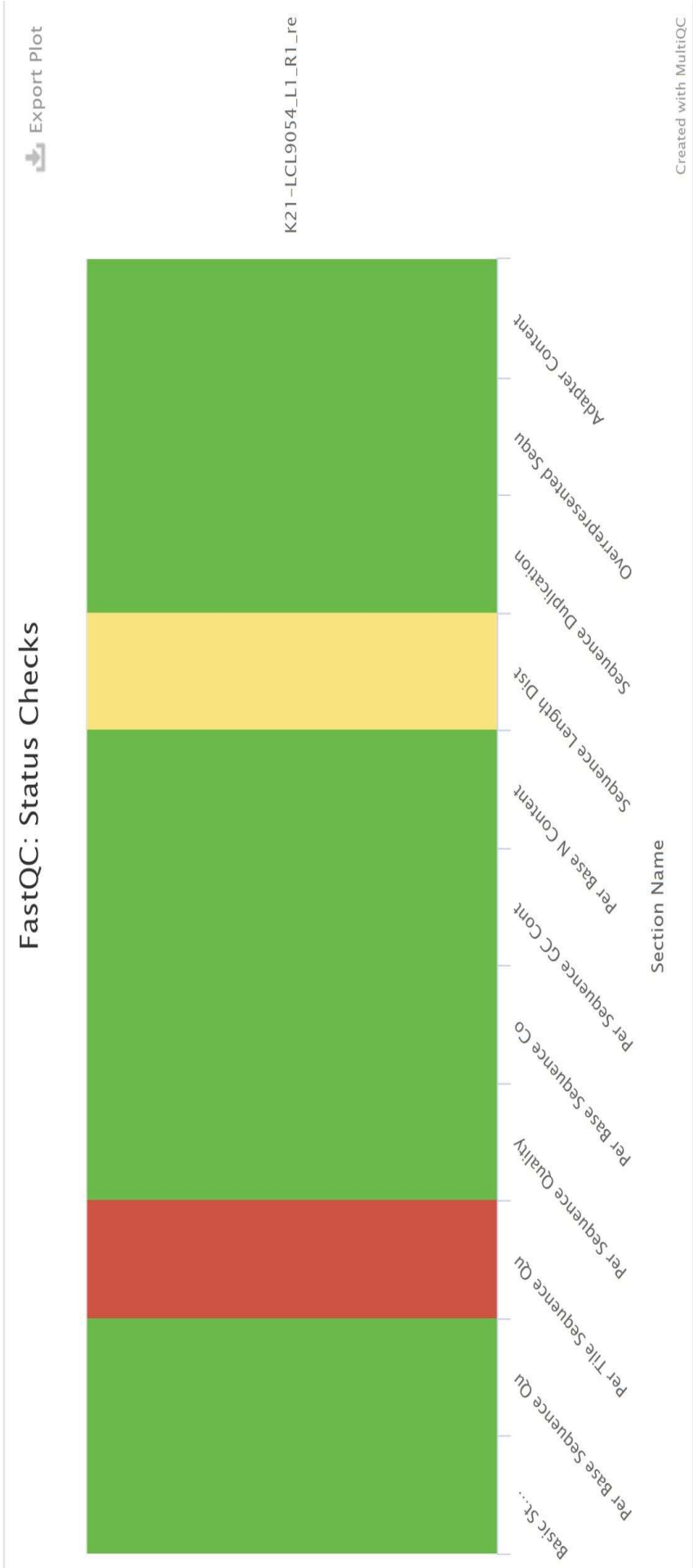
# Data Quality Check - Tile Filtered

## FastQC: Status Checks

K21-LCL9054_L1_R1_ti

Section Name

Adapter Content

Overrepresented Sequ

Sequence Duplication

Sequence Length Dist

Per Base N Content

Per Sequence GC Cont

Per Base Sequence Co

Per Sequence Quality

Per Tile Sequence Qu

Per Base Sequence Qu

Basic St...

# Data Quality Checks - Final

## FastQC: Status Checks

Section Name

Sections (left to right): Basic St..., Per Base Sequence Qu..., Per Tile Sequence Qu..., Per Sequence Quality, Per Base Sequence Co..., Per Sequence GC Cont, Per Base N Content, Sequence Length Dist, Sequence Duplication, Overrepresented Sequ, Adapter Content

Samples:
- K10-LCL9055_L1_R1_fi
- K12-LCL9056_L1_R2_fi
- K15-LCL9058_L1_R1_fi
- K16-LCL9059_L1_R2_fi
- K18-LCL9061_L1_R1_fi
- K19-LCL9052_L1_R2_fi
- K21-LCL9054_L1_R1_fi
- K6-LCL9048_L1_R2_fin
- K8-LCL9051_L1_R1_fin
- K9-LCL9050_L1_R2_fin
- W10-LCL9040_L1_R1_fi
- W11-LCL9041_L1_R2_fi
- W16-LCL9043_L1_R1_fi
- W17-LCL9044_L1_R2_fi
- W19-LCL9046_L1_R1_fi
- W2-LCL9035_L1_R2_fin
- W3-LCL9036_L1_R1_fin
- W4-LCL9037_L1_R2_fin
- W8-LCL9039_L1_R1_fin

# Data Quality Check - Samtools

Samtools is a suite of programs for interacting with high-throughput sequencing data.

## Samtools Flagstat

This module parses the output from `samtools flagstat`. All numbers in millions.



*Hover over a data point for more information*
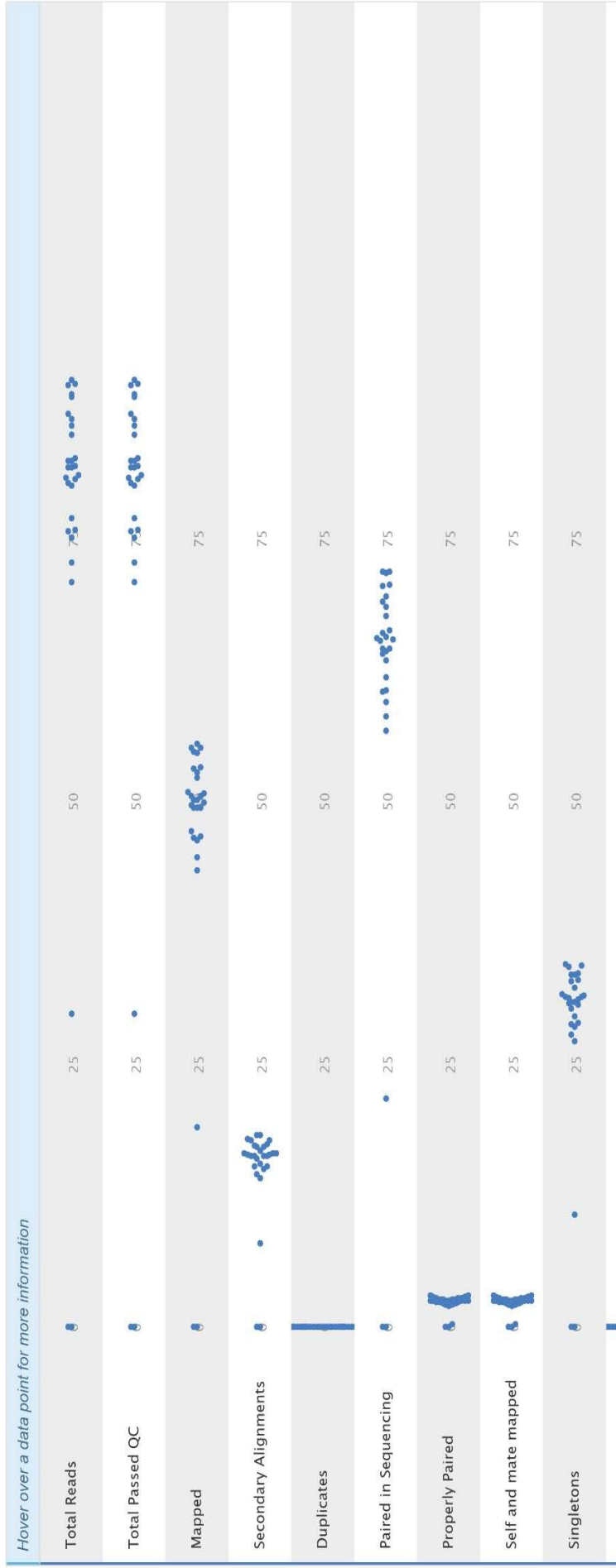
# Results

## Volcano Plot of Differentially Expressed Genes

*EnhancedVolcano*
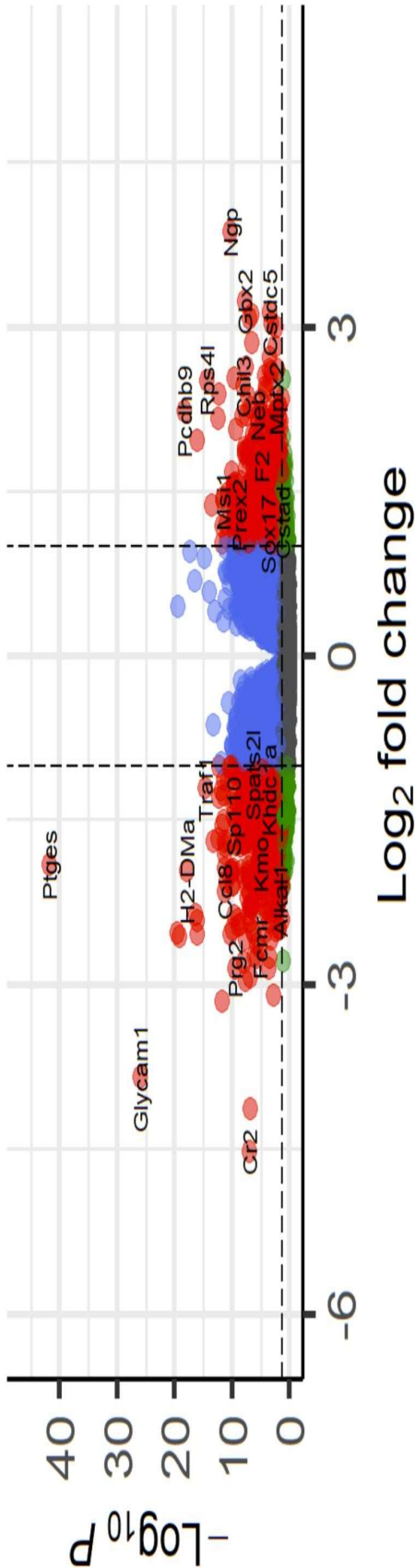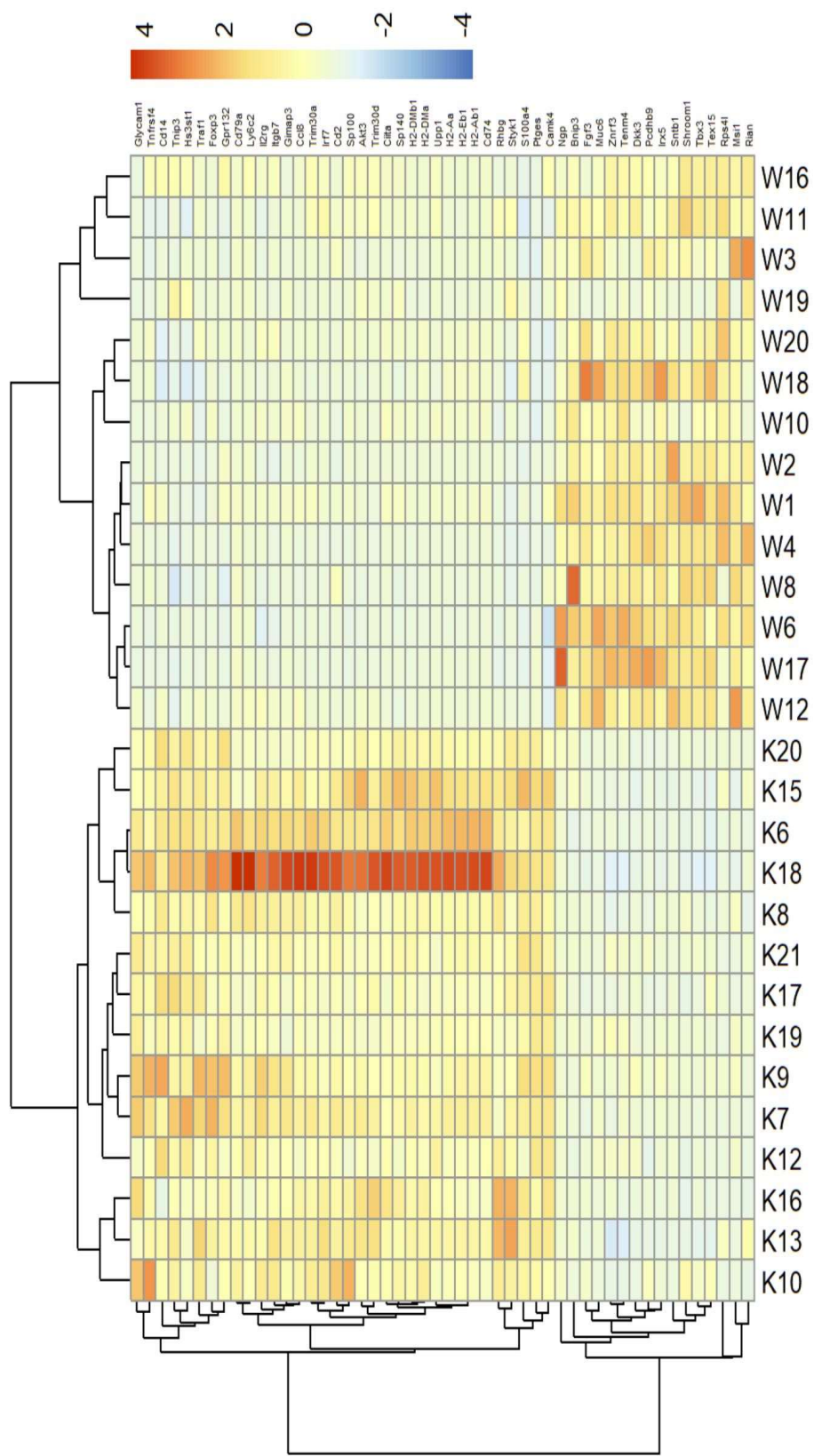
Top 10 Enriched GO Terms

GO Term

-Log10 Adjusted P-Value

response to virus

defense response to virus

antigen processing and presentation

leukocyte proliferation

lymphocyte proliferation

regulation of innate immune response

mononuclear cell proliferation

antigen processing and presentation of peptide antigen

positive regulation of innate immune response

positive regulation of response to biotic stimulus

GO Biological Process Enrichment
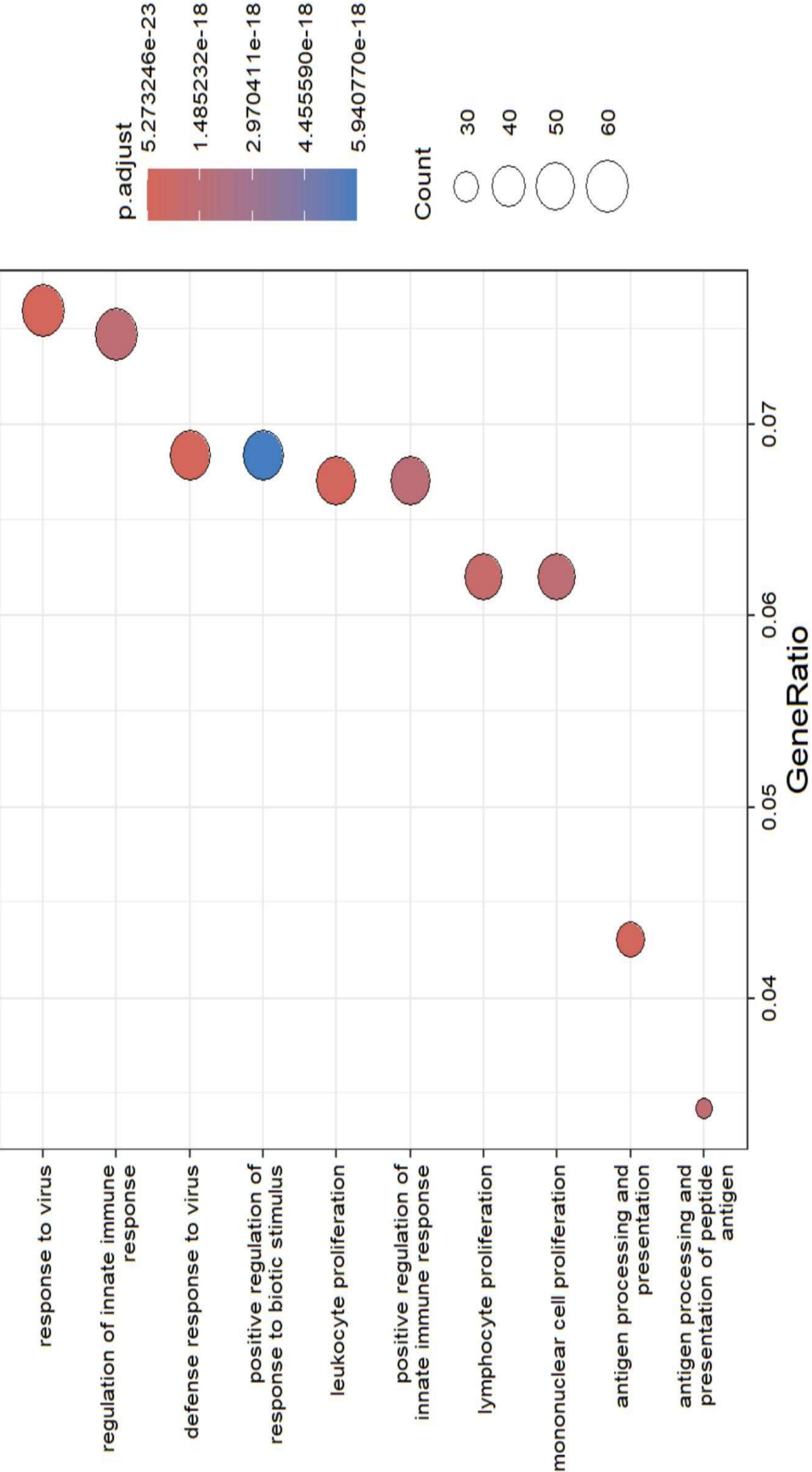
# Gene Ontology (GO) Analysis

**Key findings:**

- "GO analysis revealed a **strong immune system activation signature.**
- The top enriched biological processes included:
  - **Defense response to virus**
  - **Antigen processing and presentation**
  - **Leukocyte and lymphocyte proliferation**
  - **Regulation of innate immune response**
  - **Leukocyte cell-cell adhesion"**

**Interpretation:**

- "These results suggest that the **KO gene plays a role in immune regulation.**
- The upregulation of antigen presentation and leukocyte activity indicates a **heightened immune response**, possibly due to **autoimmune activation or infection susceptibility."**

KEGG Pathway Enrichment - Mouse
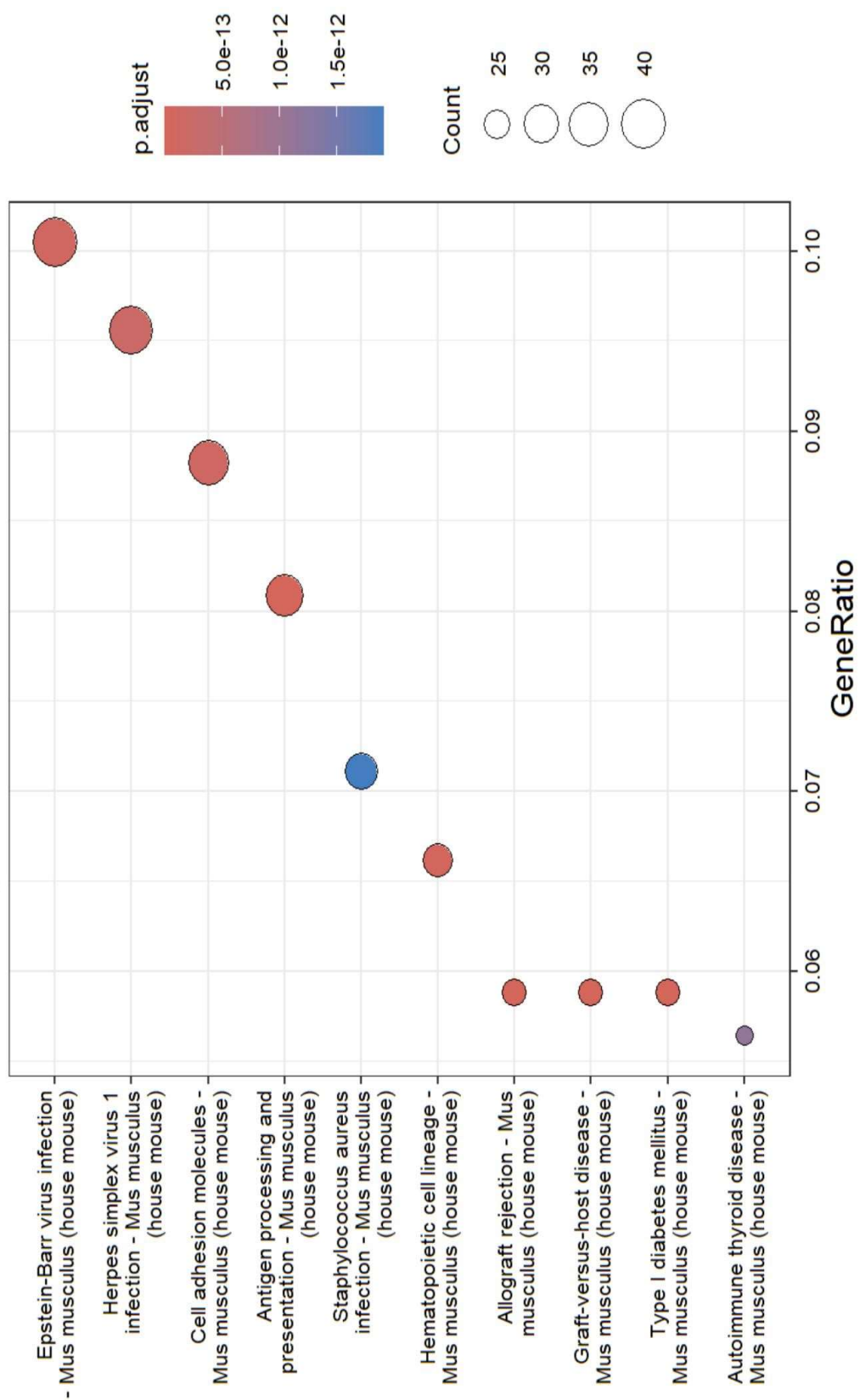
# KEGG Pathway Enrichment Analysis

## 1. Epstein-Barr Virus Infection (Top Pathway)

- "This pathway is highly enriched, suggesting that the **KO gene may impact the immune response to viral infections**."
- "This could indicate an altered susceptibility to **latent viral infections**."

## 2. Herpes Simplex Virus Infection

- "Similar to Epstein-Barr virus, this pathway suggests that the **KO gene may play a role in antiviral defense mechanisms**."
- "Changes in this pathway may **influence susceptibility or immune response to herpesviruses**."

## 3. Cell Adhesion Molecules (CAMs)

- "This pathway is critical for **immune cell communication and migration**."
- "The enrichment of this pathway suggests that the KO might be affecting **immune cell adhesion, trafficking, or signaling**."

## 4. Antigen Processing and Presentation

- "This is a key pathway in **adaptive immunity**, affecting how the immune system recognizes and responds to foreign antigens."
- "If the KO gene disrupts antigen presentation, it could impact **T-cell activation and immune surveillance**."

## 5. Staphylococcus aureus Infection

- "This indicates a potential change in **bacterial immune defense mechanisms**."
- "The KO may alter how the host immune system recognizes and clears bacterial infections."
  - ○

# Why KO Mice Are Protected from Cancer?

Based on pathway enrichment findings, we can hypothesize that KO mice have:

- **Stronger immune surveillance** via **enhanced antigen presentation and T-cell activation**
- **Increased immune system activation**, reducing tumor immune escape
- **Reduced viral oncogenesis**, preventing infection-driven cancers
- **A pro-inflammatory immune environment**, making it harder for tumors to establish
- **A potential autoimmunity-tumor trade-off**, where increased immune activation protects against cancer

# Current Progress

1. Revisit the data quality checks and improve the score on fastqc reports.

2. Rerun the pipeline once the data quality checks are well established.

3. Resume the analysis on the data for finding insights.

4. Analyse the data on different scenarios such as:

   a. With genes starting with "Ig".

   b. WT small tissue samples VS WT large tissue samples.

   c. KO small tissue samples VS KO large tissue samples.

   d. WT small tissue samples VS KO small tissue samples.

   e. WT large tissue samples VS KO large tissue samples.

# Achievements

- Globus Transfer to transfer 757 GB data in minutes from hours(scp, rsync).

- End to End pipeline development from reading data to visualising insights.

- On average 60% mapping unique read with Genome (yet to improve more).

- Reduced fastqc generation on all files from 1.5 hours to under 26 mins.

  ○ Utilised parallel module to achieve this feet on the HPC clusters

Thank You