Group 6

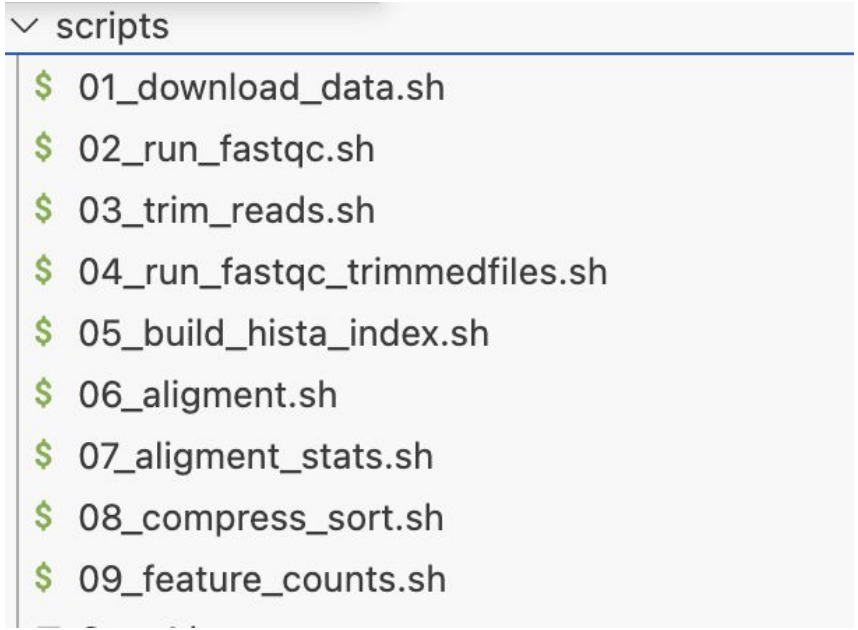
Project 3: Short-read RNA-seq of mice

1. Experiment:

1.1 Understanding how immune challenges elicit in different responses, in mouse tissues when infected with several virus

1.2 samples Case (1,2,3) control (1,2,3) - RNA-seq analysis

1.3 In total we have 9 scripts, we used the parameters from the paper



$ 06_aligment.sh

Alignment by Hisat2

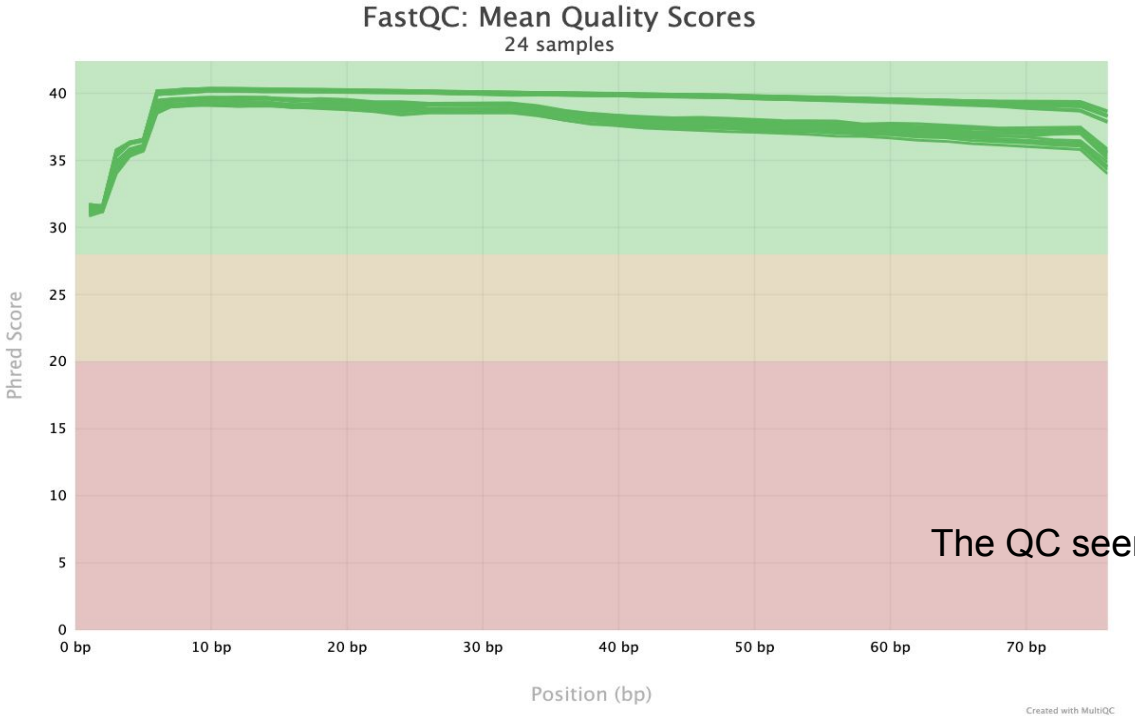
Script was adapted to the samples Case and Control using a loop for each type do data

scripts > $ 06_aligment.sh

```
#!/usr/bin/env bash

TRIMMED_DIR=/group_work/group6/data/fastq/trimmed
REFERENCE_DIR=/group_work/group6/data/reference
ALIGNED_DIR=/group_work/group6/data/alignment

mkdir -p $ALIGNED_DIR


for i in {1..3}; do
    # Construct the case identifier string (e.g., "Case1", "Case2", "Case3")
    case_id="Case${i}" # Using ${i} is slightly safer than $i here
    hisat2 \
    -x $REFERENCE_DIR/Mus_musculus.GRCm38.dna.chromosome.5.fa \
    -1 $TRIMMED_DIR/trimmed_${case_id}_R1.fastq \
    -2 $TRIMMED_DIR/trimmed_${case_id}_R2.fastq \
    > $ALIGNED_DIR/${case_id}.aligned.sam

done

for i in {1..3}; do
    # Construct the case identifier string (e.g., "Case1", "Case2", "Case3")
    case_id="Control${i}" # Using ${i} is slightly safer than $i here
    hisat2 \
    -x $REFERENCE_DIR/Mus_musculus.GRCm38.dna.chromosome.5.fa \
    -1 $TRIMMED_DIR/trimmed_${case_id}_R1.fastq \
    -2 $TRIMMED_DIR/trimmed_${case_id}_R2.fastq \
    > $ALIGNED_DIR/${case_id}.aligned.sam

done

```

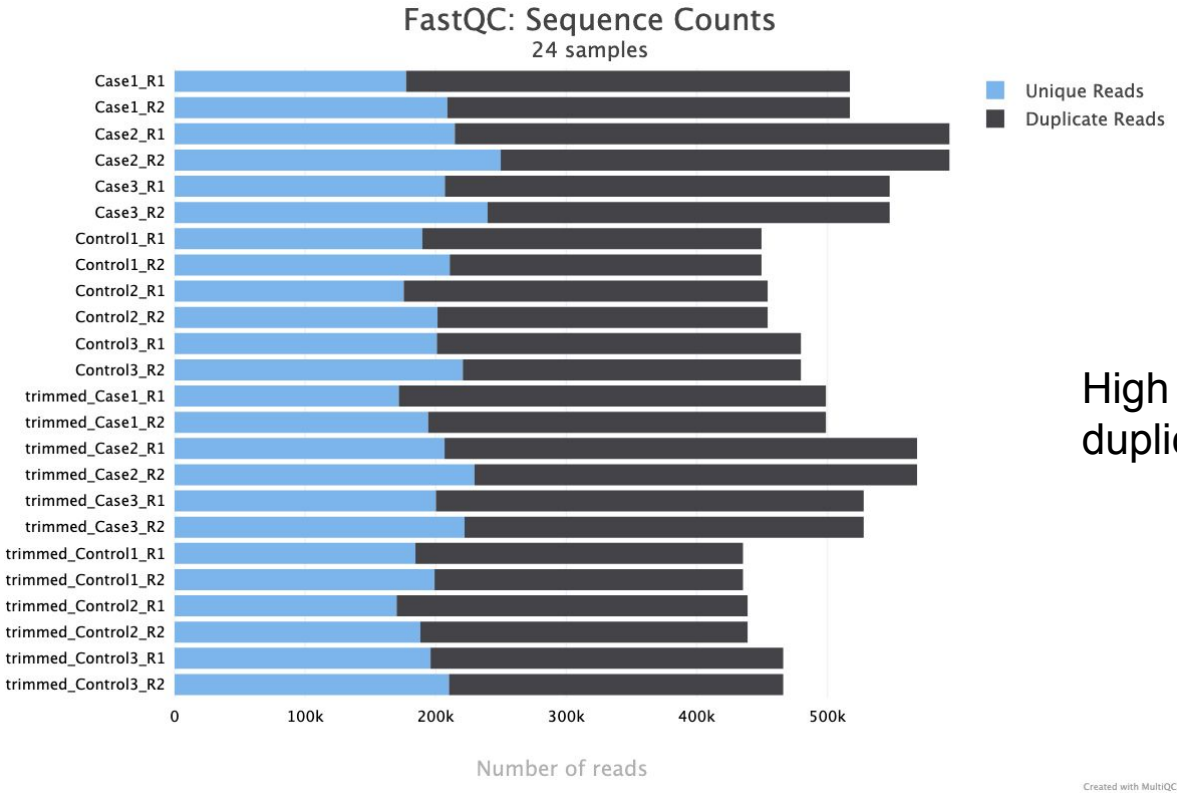
2. MultiQC parameters

MultiQC: Quality scores



The QC seems ok

2. MultiQC parameters - problems



High number of duplicate reads

Overrepresented Sequences

Overrepresented sequences by sample

3 21

Help

21 / 24 samples with warnings

Click bar to fix in place
Highlight these samples
Show only these samples

The total amount of overrepresented sequences found in each lib

24 samples had less than 1% of reads made up of overrepres

Top overrepresented sequences

Top overrepresented sequences across all samples. The table shows 20 most overrepresented sequences across all samples, ranked by the number of samples they occur in.

Copy table | Configure columns | Scatter plot | Violin plot | Export as CSV... | Showing $^{7}/_{7}$ rows and $^{3}/_{3}$ columns. | Summarize table

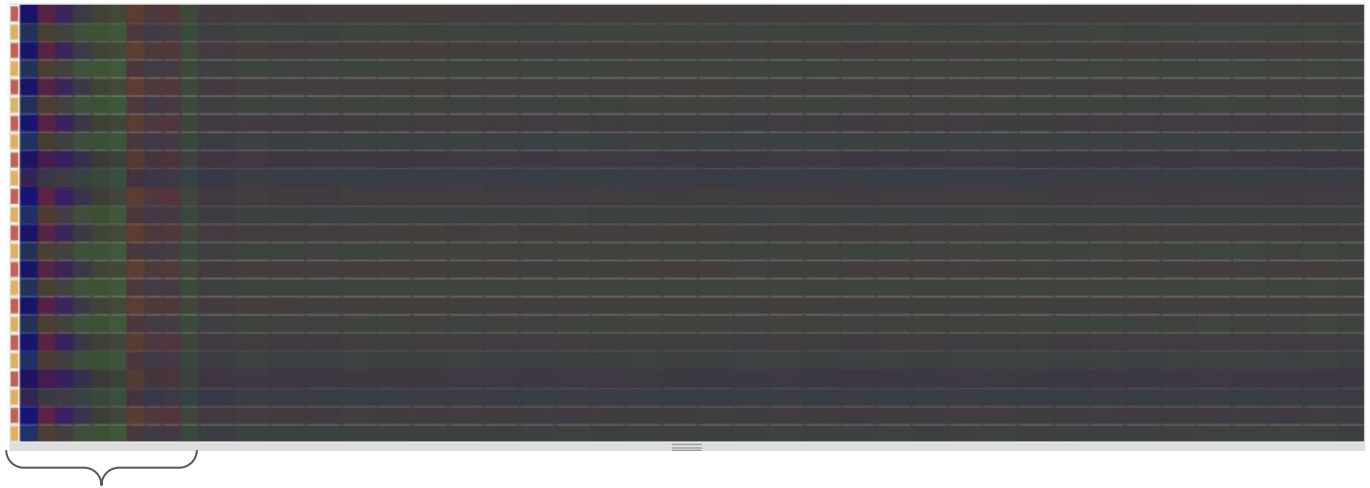
Overrepresented sequence	Reports	Occurrences	% of all reads
CCCGAATCTCAGTGAGGTCCTCCTTGGTGAACACGAAGCCCACGTTCCCC	9	5 638	0.0472 %
CCTCGTTGGAGTGACATCGTCTTTAAACCCCGCGTGGCAATCCCTGACGC	7	3 733	0.0312 %
CTCAACCCCAGACCACAGGACCGGTTCTGCCCAACCCTTTTGAACTACTT	4	4 533	0.0379 %
CCCGGATGTGAGGCAGCAGTTTCTCCAGAGCTGGGTTGTTCTCCAGGTGG	2	1 015	0.0085 %
CTCCGACTCTTCCTTTGCTTCAGCTTTGGCAGGGGCTGCAGCAGCCGCAG	2	1 036	0.0087 %
GGGAGCCTGAGGAGGCAGCAGCAGCTGAGAACTGCACTTGGACCTGTGCT	1	537	0.0045 %
GTCAGCTGCCACTTGACATCCAAGACAAGTGAAACAAAAGGTCCCACAGA	1	610	0.0051 %

Overrepresented Sequences

☑	Mus musculus 8 days embryo whole body cDNA, RIKEN full-length enriched library, clone:5730529P21 product:a…	Mus musculus	93.5	93.5	100%	1e-15	100.00%	1096	AK161683.1
☑	Mus musculus ribosomal protein, large, P0, mRNA (cDNA clone MGC:18649 IMAGE:3982438), complete cds	Mus musculus	93.5	93.5	100%	1e-15	100.00%	1138	BC011106.1
☑	Mus musculus targeted KO-first, conditional ready, lacZ-tagged mutant allele Pxn:tm1a(EUCOMM)Hmgu; transge…	Mus musculus	93.5	93.5	100%	1e-15	100.00%	37953	JN964543.1
☑	Mus musculus bone marrow macrophage cDNA, RIKEN full-length enriched library, clone:I830007N01 product:ac…	Mus musculus	93.5	93.5	100%	1e-15	100.00%	1095	AK150448.1
☑	Mus musculus ribosomal protein, large, P0, mRNA (cDNA clone MGC:107166 IMAGE:30254142), complete cds	Mus musculus	93.5	93.5	100%	1e-15	100.00%	1183	BC089496.1
☑	Mus musculus BAC clone RP23-346B13 from chromosome 3, complete sequence	Mus musculus	93.5	93.5	100%	1e-15	100.00%	221423	AC123859.4
☑	Mus musculus ribosomal protein, large, P0, mRNA (cDNA clone MGC:6264 IMAGE:3600590), complete cds	Mus musculus	93.5	93.5	100%	1e-15	100.00%	1102	BC003833.1
☑	Mus musculus adult male testis cDNA, RIKEN full-length enriched library, clone:4930579L21 product:acidic riboso…	Mus musculus	93.5	93.5	100%	1e-15	100.00%	1097	AK029816.1
☑	Mus musculus ES cells cDNA, RIKEN full-length enriched library, clone:2400003H06 product:acidic ribosomal ph…	Mus musculus	93.5	93.5	100%	1e-15	100.00%	1312	AK010267.1
☑	Mus musculus profilaggrin (Flg) gene, exons 2 and 3 and partial cds	Mus musculus	93.5	93.5	100%	1e-15	100.00%	3716	AF510859.1
☑	Mus musculus blastocyst blastocyst cDNA, RIKEN full-length enriched library, clone:I1C0007K09 product:acidic ri…	Mus musculus	93.5	93.5	100%	1e-15	100.00%	1096	AK166728.1
☑	Mus musculus 11 days embryo whole body cDNA, RIKEN full-length enriched library, clone:2700094E22 product:…	Mus musculus	93.5	93.5	100%	1e-15	100.00%	1109	AK012606.1

- Genes expressed at high level = Blast -> Ribosomal proteins
- Or it could be a problem regarding the library, where we can not distinguish exons from different genes

MultiQC: Per Base Sequence Content

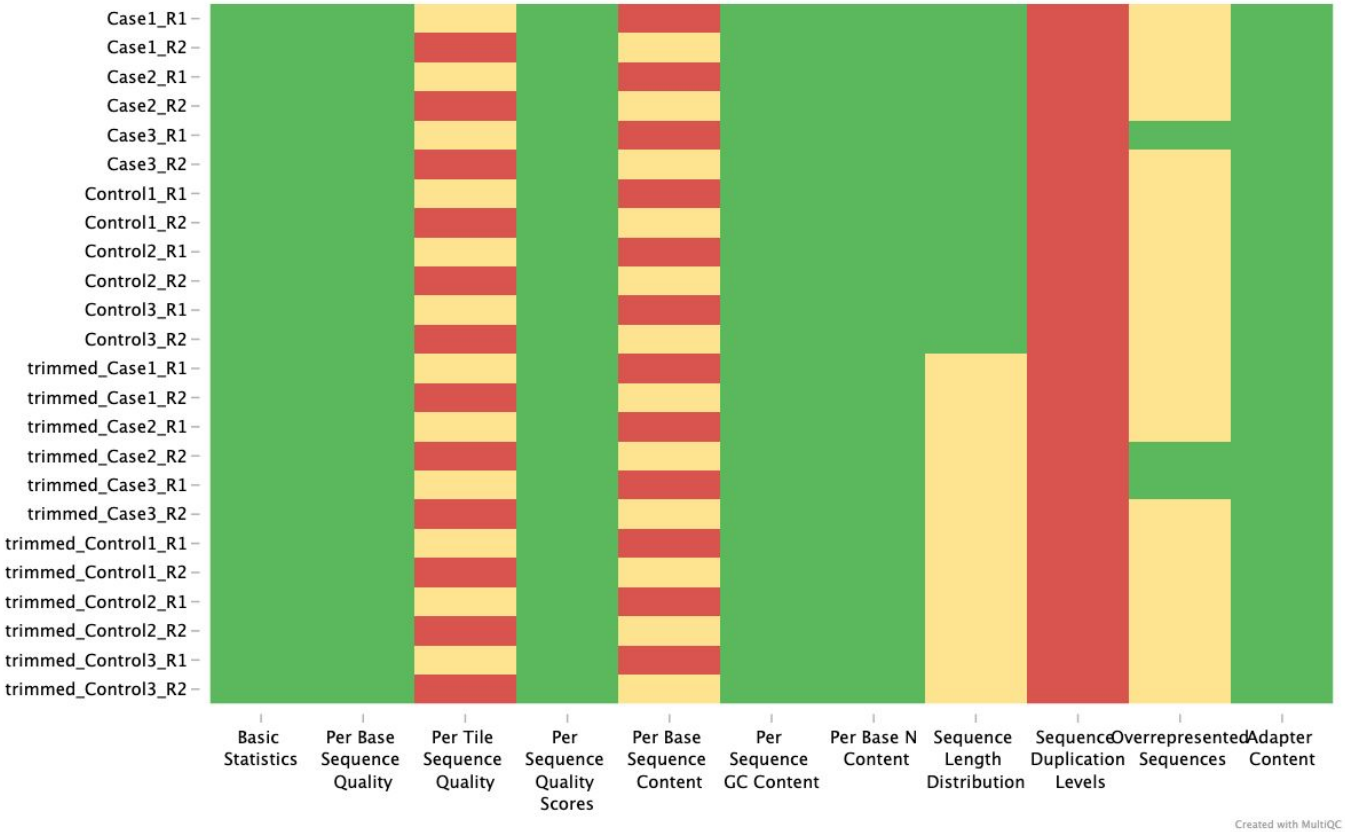


The proportion of each base position were different in the first 10 bp positions.

The use of random hexamers in library preparation can introduce a positional bias in the first few bases of the reads, leading to uneven sequence content.

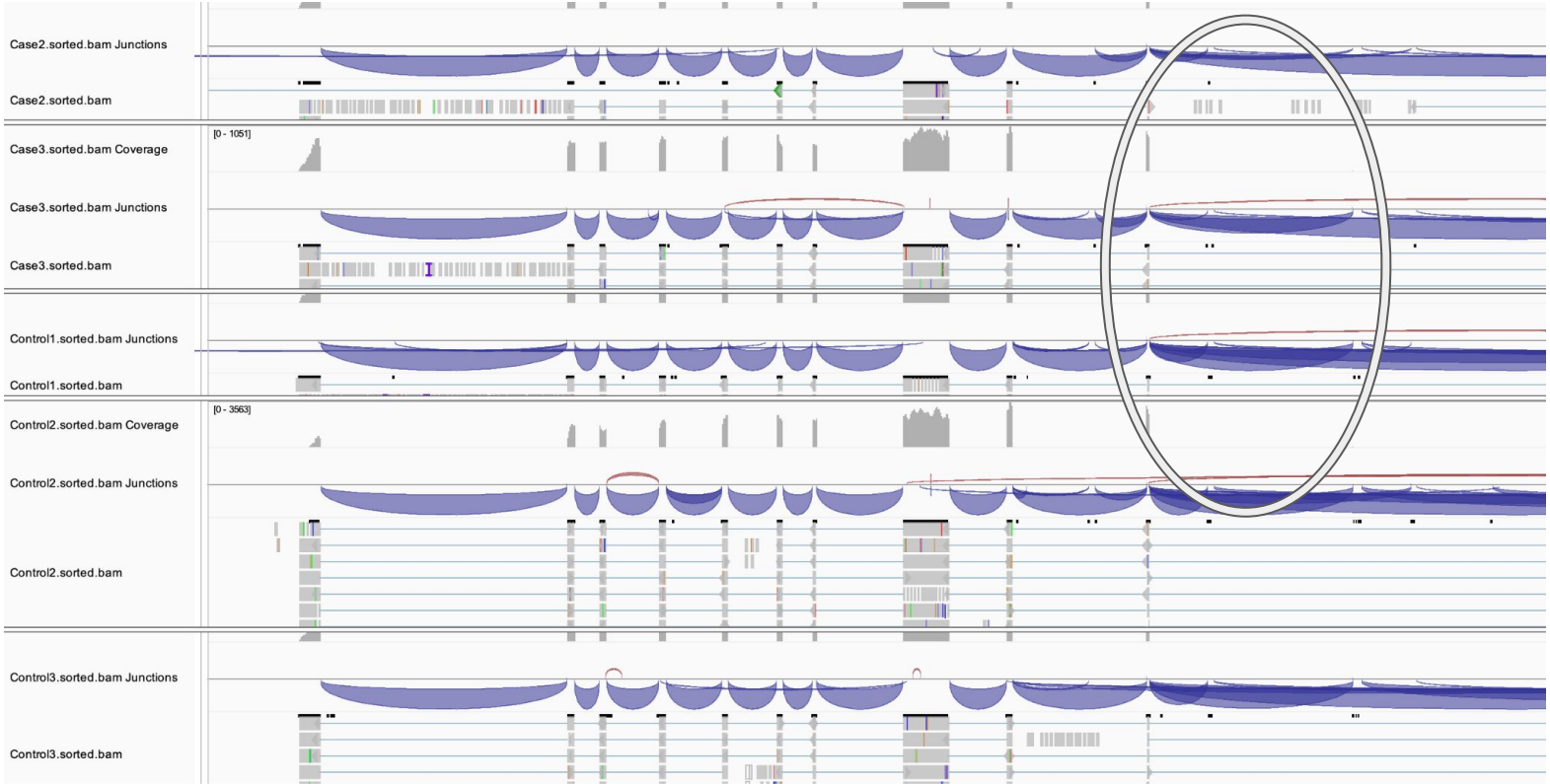
This is due to a biased selection of random primers, but doesn't represent any individually biased sequences

MultiQC: Overview

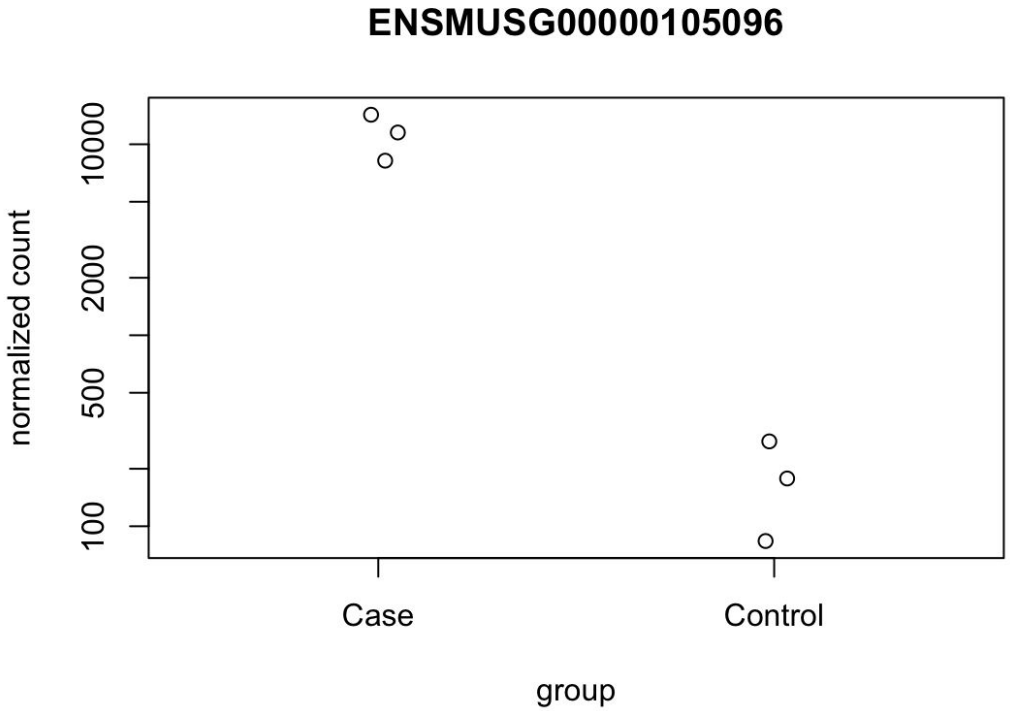


Visualization in IGV

- alternative splicing can be visualized by the arcs (red and blue plus and minus strands)
- we can see alternative splicing between case and control and also, between control samples

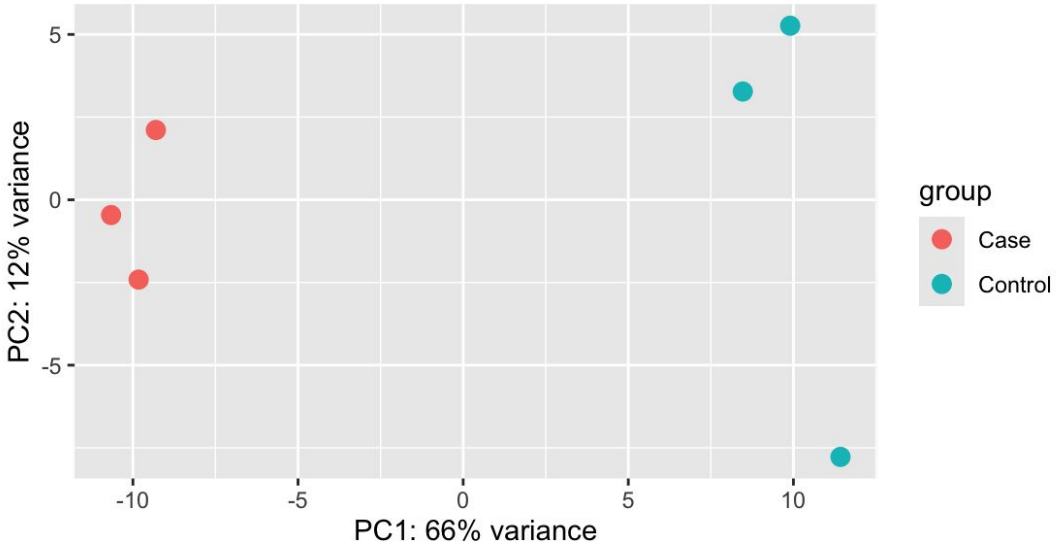


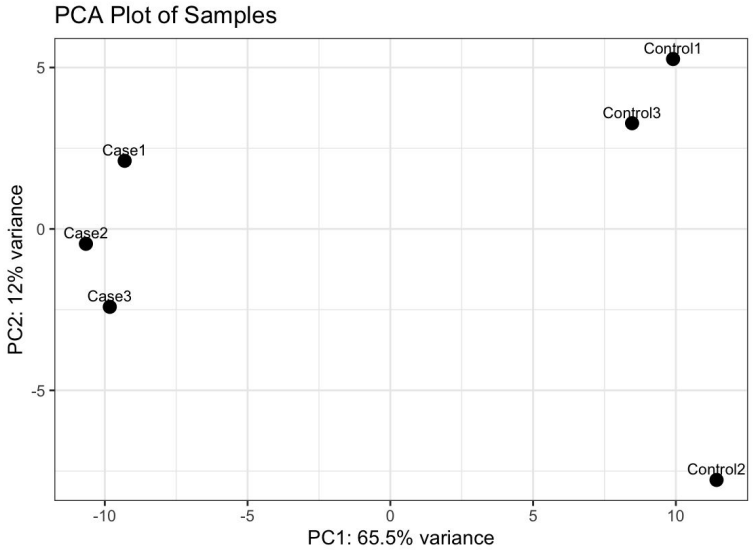
Counts of most differentially expressed gene



- Most differentially expressed gene: Guanylate-binding protein 10

PCA - Differentially expressed genes





Discussion

Sequencing data passed quality checks

Data is ready for downstream analysis

Differentially expressed genes in cases vs controls

Pipeline could be further optimized to address the problems detected