RNA-Seq data analysis project

Course: Functional Genomics and Proteomics

Instructor: Panagiotis Moulos

Student: Alexandros Mitsis

Part 1 - Alignment of a FASTQ file

1. Align the file to the reference genome and produce a BAM file.

2. Provide the HISAT2 command used to generate the BAM file.

Run the alignment program:

hisat2 -x ./hg19/genome -U ./human.fastq.gz -S HISAT_OUT.sam

Convert the sam file:

samtools view -bS HISAT_OUT.sam > HISAT_OUT.bam

3. Provide the HISAT2 stats output from the screen. What is the total alignment rate?

Total alignment rate: 96.66%

```
7152258 reads; of these:
  7152258 (100.00%) were unpaired; of these:
    238896 (3.34%) aligned 0 times
    6265698 (87.60%) aligned exactly 1 time
    647664 (9.06%) aligned >1 times
96.66% overall alignment rate
```

Part 2 - QC and differential expression analysis

1. Create and deliver the targets file. The BAM files contain paired-end reads. Take this into account in your target file.

samplename	filename	condition	paired	stranded
Pre_1	C:/Users/mitsis/Desktop/Moulos/bam/Pre_1.bam	Pre	paired	forward
Pre_2	C:/Users/mitsis/Desktop/Moulos/bam/Pre_2.bam	Pre	paired	forward
Pre_3	C:/Users/mitsis/Desktop/Moulos/bam/Pre_3.bam	Pre	paired	forward
Post_1	C:/Users/mitsis/Desktop/Moulos/bam/Post_1.bam	Post	paired	forward
Post_2	C:/Users/mitsis/Desktop/Moulos/bam/Post_2.bam	Post	paired	forward
Post_3	C:/Users/mitsis/Desktop/Moulos/bam/Post_3.bam	Post	paired	forward

2. Perform a differential expression analysis with metaseqR2 for the contrast Post vs Pre. The normalization method as well as the statistical testing method should be "deseq2". The following QC plots should be reported: "mds", "biodetection","countsbio","saturation","correl","boxplot","meandiff","meanvar", "volcano","mastat". A report should be created. Please deliver the command you used.

```
# Set file path
targetsFile <- file.path("C:/Users/mitsis/Desktop/Moulos/targets.txt")

# Define statistical comparisons
theContrasts <- c("Post_vs_Pre")

# Perform analysis using the metaseqr2 function
metaseqr2(
  sampleList=targetsFile,
  contrast=theContrasts,
  org="hg19",
  countType="exon",
  normalization="deseq2",
  statistics="deseq2",
  figFormat="png",
  qcPlots=c(
    "mds","biodetection","countsbio","saturation","correl",
    "boxplot","meandiff","meanvar", "volcano","mastat"
  ),
  exportWhere=file.path("C:/Users/mitsis/Desktop/Moulos/task_2"),
  pcut=0.05,
  restrictCores=0.25,
  exportWhat=c("annotation","p_value","adj_p_value","fold_change",
```

```
                "counts","flags"),
  exportScale=c("natural","log2","rpgm"),
  exportValues="normalized",
  saveGeneModel=TRUE,
  reportTop=0.1,
  localDb=file.path("C:/Users/mitsis/Desktop/Moulos/annotation.sqlite")
)
```

3. Inspect the report. Are there any samples that could be excluded in a second round of analysis? Why?

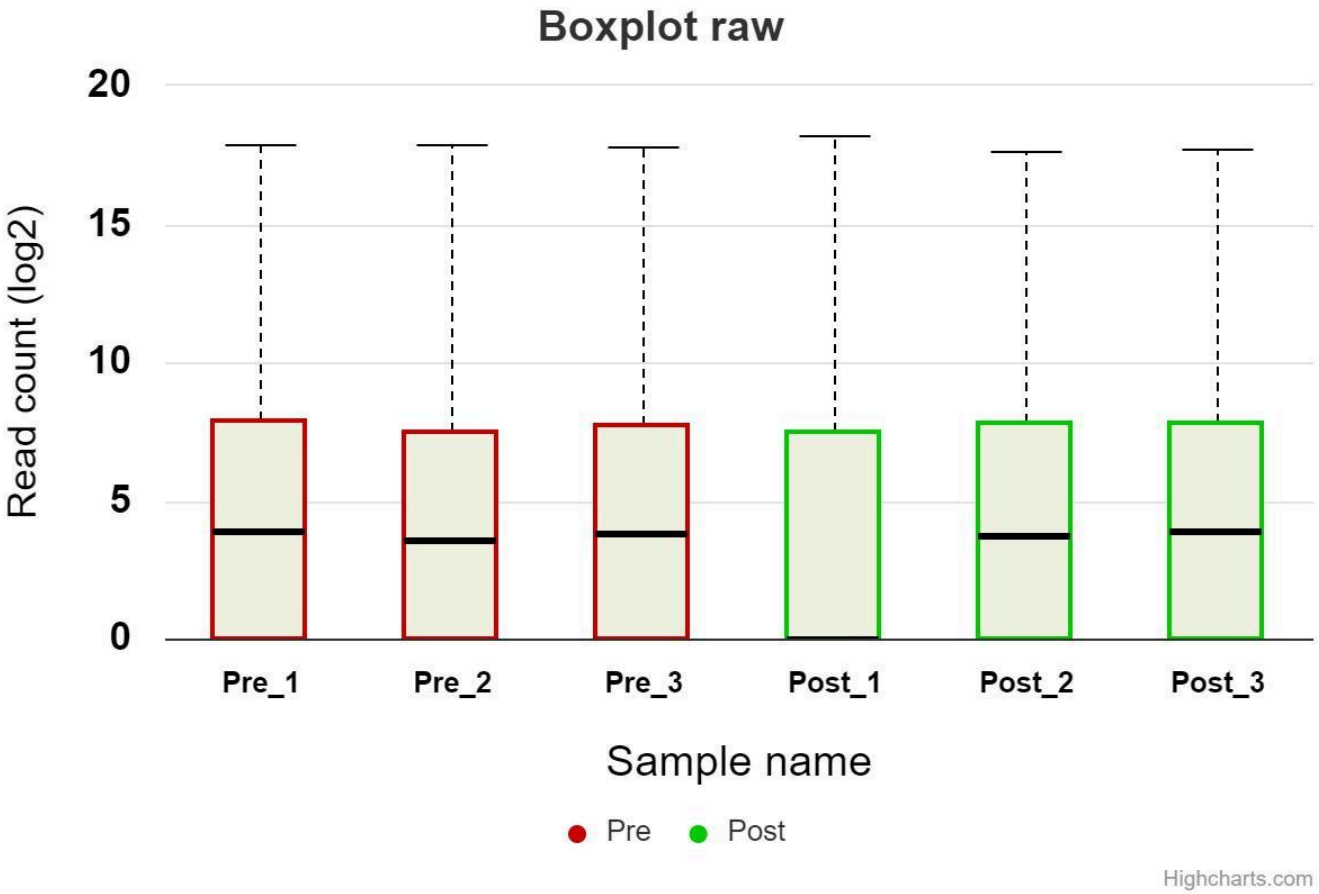

Figure 1: Normalization assessment boxplot showing whether the data from the different samples follow the same underlying distribution.

Upon inspecting the report, one sample stands out as a candidate for exclusion in a subsequent round of analysis: Post_1. This is primarily due to discrepancies observed in the normalization plot (Figure 1), where the boxplots diverge significantly. Such dissimilarity

indicates suboptimal normalization quality, with the median of Post_1 notably deviating from the other samples.

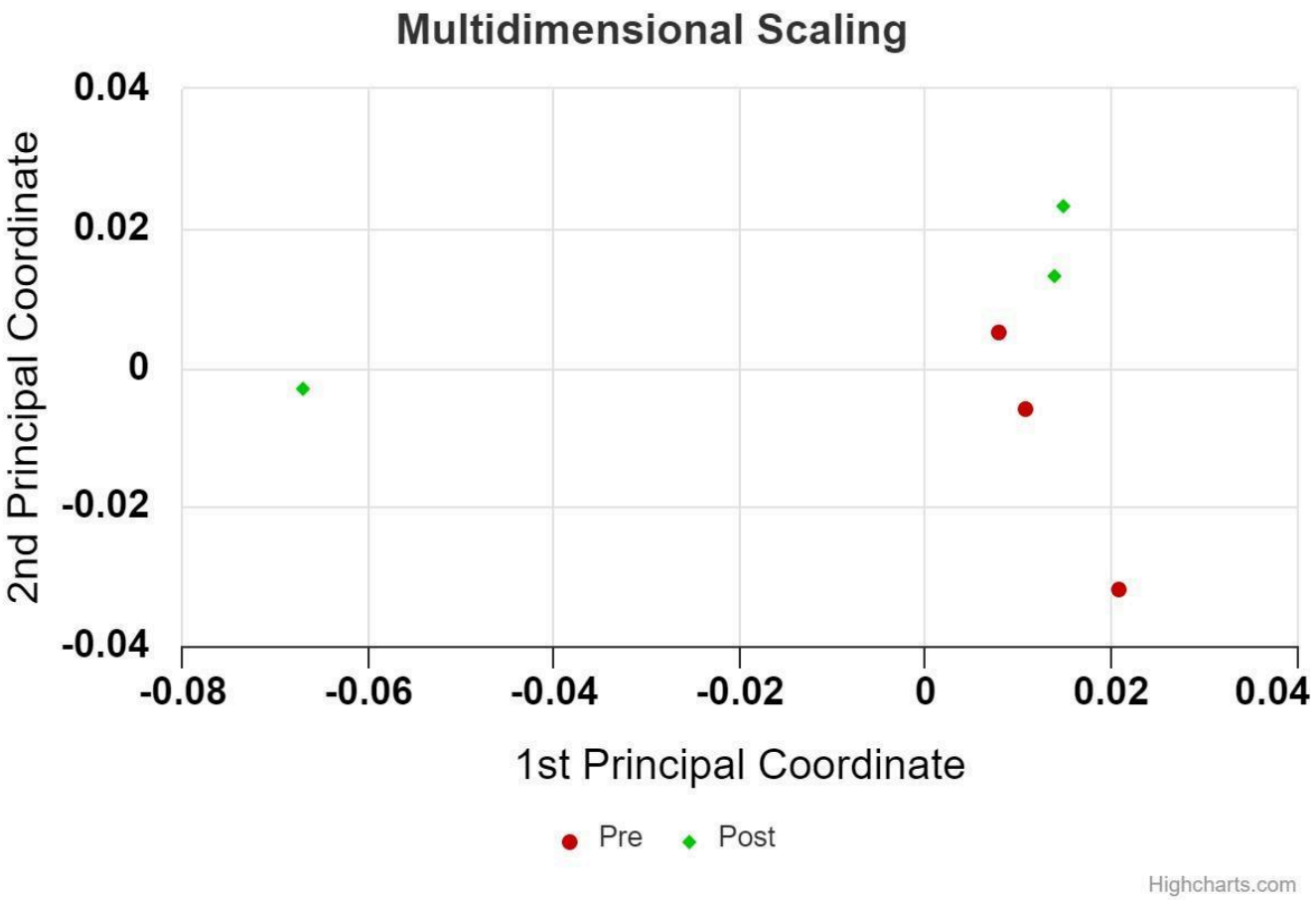

Figure 2: Multidimensional Scaling (MDS) plot depicting the degree of similarity among individual cases within our dataset.

The irregularity observed in the boxplot of Post_1 extends its impact to the next Quality Control Figures (2-4). In Figure 2, a considerable distance between Post_1 and other samples within the same biological condition suggests poor correlation and reproducibility.

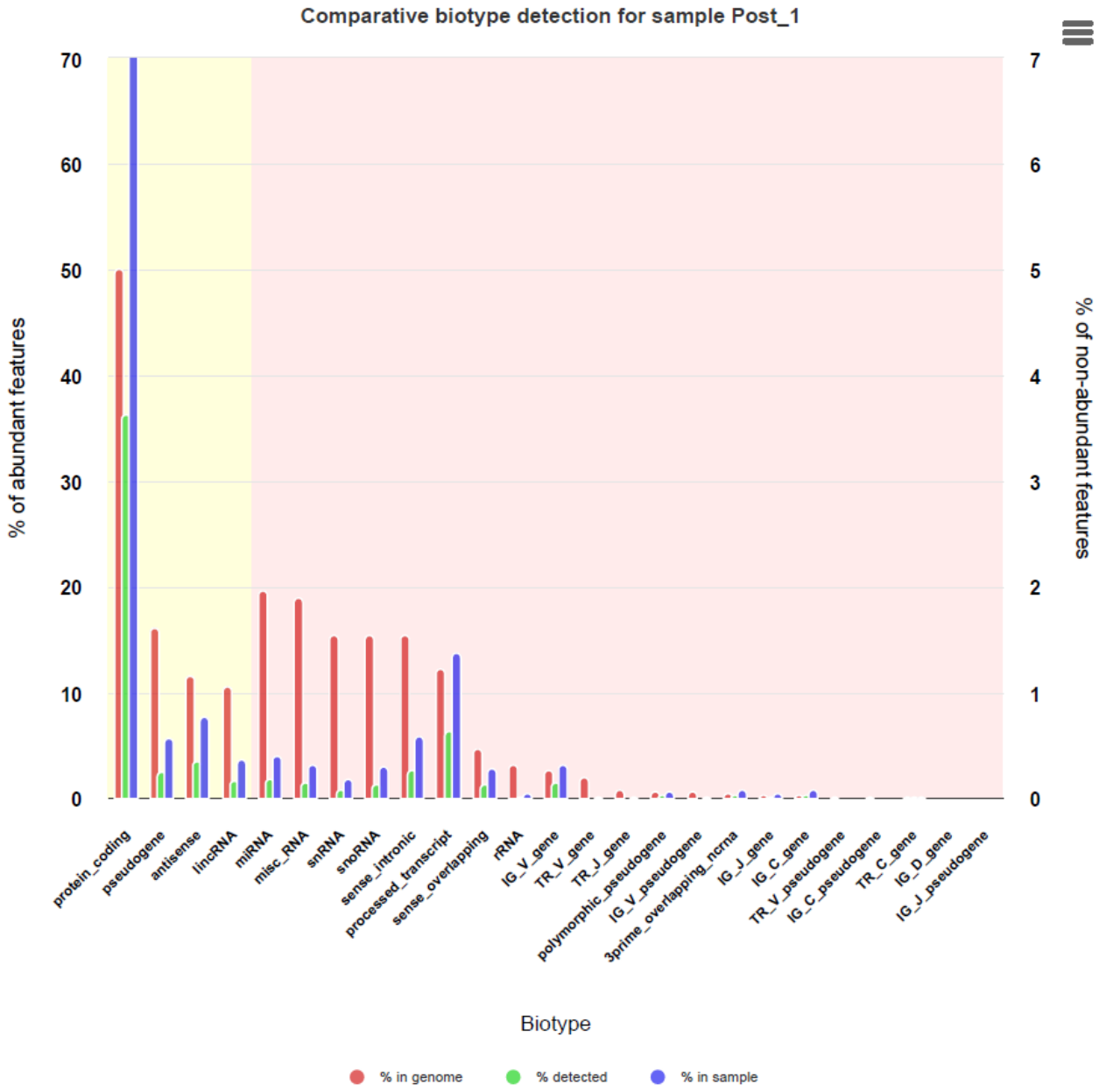

Figure 3: Bar diagram illustrating biotype detection, showcasing the percentage of each biotype in the genome represented by red bars. The green bars Indicate the proportion of biotypes detected in a sample pre-normalization and after basic filtering. The diagram also illustrates the percentage of each biotype within the sample through blue bars.

Figure 3 shows a lower percentage of detected protein coding genes for Post_1, suggesting potential quality problems. The pairwise sample correlations display inconsistencies, with samples from different biological conditions occasionally mixing, indicating broader dataset quality issues. It's advisable to either exclude Post_1 from the analysis or explore alternative normalization algorithms beyond deseq2 to address potential issues.

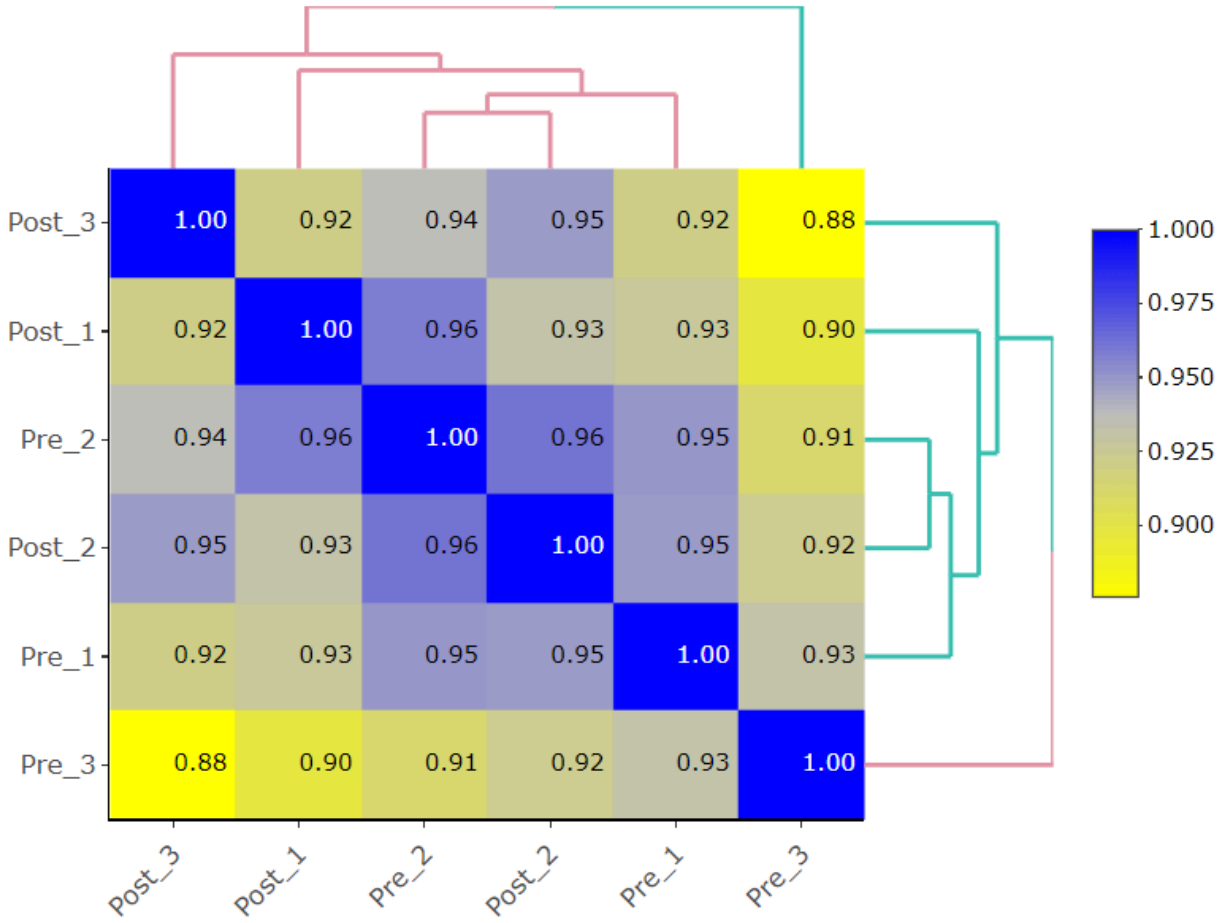

Figure 4: Sample correlation plot, portraying the pairwise correlations among each pair of samples through a clustered heatmap displayed as color-scaled images.

The correlation plot above further showcases a potential quality issue, as samples from the same group aren't clustered together.

4. Report the number of differentially expressed genes based on the report.

The counts of differentially expressed genes per contrast are as follows:

For the contrast between Post and Pre, 927 (46) statistically significant genes were identified using a p-value (FDR or adjusted p-value) threshold of 0.05. Among these, 345 (13) genes were up-regulated, 407 (33) were down-regulated, and 175 (0) were not deemed differentially expressed, considering an absolute fold change cutoff value of 1 in log2 scale. (The numbers within parentheses indicate genes with p value ≤ 0.0001 ($-\log10(\text{p value}) \geq 4$)).

5. Create and deliver the alignment metrics using the R scripts and the metaseqR2 database as described in the last part of the tutorial, using the targets file you created for the differential expression analysis.

sample_name	total_reads	total_paired_reads	total_read_pairs	aligned_reads	proper_paired_aligned_reads	proper_paired_aligned_read_pairs
Pre_1	52003302	52003302	26001651	50609590	48546564	24273282
Pre_2	53935182	53935182	26967591	52795825	51303100	25651550
Pre_3	51956088	51956088	25978044	51104119	49861334	24930667
Post_1	54160748	54160748	27080374	52980251	51228734	25614367
Post_2	53625048	53625048	26812524	52432681	50763272	25381636
Post_3	55282922	55282922	27641461	54102433	52649694	26324847

uniquely_aligned_reads	proper_paired_uniquely_aligned_reads	proper_paired_uniquely_aligned_read_pairs	chimeric_reads	uniquely_aligned_chimeric_reads
47701649	45914834	22957417	221194	150305
49638901	48340180	24170090	225948	171467
47826508	46761468	23380734	167580	113648
50514163	48994132	24497066	227668	162937
49590138	48123120	24061560	193918	140394
51450056	50171382	25085691	206368	158285

on_target_reads	on_target_uniquely_aligned_reads	total_bases	total_paired_bases	aligned_bases	proper_paired_aligned_bases
42019462	40475517	4680297180	4680297180	4554863100	4369190760
35143229	33794477	4854166380	4854166380	4751624250	4617279000
41423838	39799153	4676047920	4676047920	4599370710	4487520060
40909793	39525215	4874467320	4874467320	4768222590	4610586060
41977991	40500540	4826254320	4826254320	4718941290	4568694480
40722302	39328066	4975462980	4975462980	4869218970	4738472460

uniquely_aligned_bases	proper_paired_uniquely_aligned_bases	on_target_bases	on_target_uniquely_aligned_bases
4293148410	4132335060	3781751580	3642796530
4467501090	4350616200	3162890610	3041502930
4304385720	4208532120	3728145420	3581923770
4546274670	4409471880	3681881370	3557269350
4463112420	4331080800	3778019190	3645048600
4630505040	4515424380	3665007180	3539525940