RNA-seq Analysis Guide

|  |  |
| --- | --- |
| **Date:** | July 1, 2023 |

table of contents

[table of contents 2](#_Toc139128875)

[1 Retrieving and Organizing files 3](#_Toc139128876)

[1.1 Organizing .fastq files 3](#_Toc139128877)

[2 Quality Control – Raw Data 3](#_Toc139128878)

[2.1 Running FASTQC 3](#_Toc139128879)

[2.2 Running MultiQC 3](#_Toc139128880)

[3 Preprocessing 4](#_Toc139128881)

[3.1 Quality score and adapter trimming 4](#_Toc139128882)

[3.2 rRNA and HBB removal (optional) 4](#_Toc139128883)

[4 Alignment 5](#_Toc139128884)

[4.1 Retrieve and concatenate reference files 5](#_Toc139128885)

[4.2 Generating STAR index 5](#_Toc139128886)

[4.3 Alignment with STAR 5](#_Toc139128887)

[5 Tabulation of counts 6](#_Toc139128888)

[5.1 Tabulation 6](#_Toc139128889)

**List of Tables**

No table of figures entries found.

**List of Appendices**

No table of figures entries found.

# Retrieving and Organizing files

## Organizing .fastq files

### Required files

* 01\_MovingFiles.sh
* 211206\_SampleSheet.csv
* Raw .fastq.gz files from the sequencing core

### Notes

* The sequencing facility uploaded raw .fastq.gz files onto OAK. In this section, files are moved from OAK into my SHERLOCK $SCRATCH directory.
* Each sequencing library has an associated directory in which the library’s .fastq.gz files are placed. The files were generated from paired end sequencing split across two lanes, so each directory has four files (\*\_L001\_R1.fastq.gz; \*\_L001\_R2.fastq.gz; \*\_L002\_R1.fastq.gz; \*\_L002\_R2.fastq.gz).
* The sample sheet associates each fastq.gz file with the proper directory name.

# Quality Control – Raw Data

## Running FASTQC

### Required files

* 02\_FASTQC.sbatch
* 211209\_DirectoryList.csv

### Notes

* SBATCH tasks route SLURM output and error files are to a premade “slurmout” directory
* SHERLOCK has FASTQC installed in the biology module
* Workflow for this task (and certain other batch tasks performed on SHERLOCK) uses the directory list file to associate the SLURM array task ID with a sample directory and carry out the task on the files in that sample directory. Results are routed to a directory for that sample within a new parent directory.

## Running MultiQC

### Required files

* 03\_MultiQC.sbatch
* Output from 2.1 Running FastQC

### Notes

* MultiQC combs a directory and its subdirectories for file extensions that it can use. You can tell it to ignore certain extensions if you don’t want it to generate an aggregate analysis for those files.
* **A word of caution on interpreting the results** – remember that 1) the transcriptome of erythroblasts is unusually biased toward a restricted subset of genes, 2) libraries generated from infected erythroblasts contain transcripts from the parasite which has an AT-rich genome and 3) rRNA depletion was not carried out for the parasite transcriptome.

# Preprocessing

## Quality score and adapter trimming

### Required files

* 05\_Trimming.sbatch
* 211209\_DirectoryList.csv
* Raw fastq.gz files from the sequencing core

### Notes

* Trim Galore is in the biology module on SHERLOCK
* Trim Galore outputs trimmed fastq files and runs FASTQC on the output. You can run MultiQC by adapting the script used in 2.2 if you want to check the aggregated result after trimming.

## rRNA and HBB removal (optional)

### Required files

* 07\_bbMaprRNARemoval.sbatch
* 211209\_DirectoryList.csv
* Trimmed .fastq.gz files
* Concatenation of the following fasta files: Pf\_rRNA\_1; human\_HBA1; human\_HBA\_2; human\_HBB; human\_HBD; human\_mito\_genome; human\_rRNA; phiX

### Notes

* Removing these transcripts speeds up the alignment step and was necessary to demonstrate how the unusual biology of these samples (heavily biased transcriptome, parasite rRNA) contributed to sequencing results.
* This step can be omitted. If so, the STAR alignment time will be longer and rRNA genes will need to be removed post-tabulation of counts.
* FASTQC and MultiQC can also be used to run quality control after removal of sequences
* Reference files for contaminating sequences were sourced from Ensembl and Illumina

# Alignment

## Retrieve and concatenate reference files

### Required files

* 09\_MakeReference.sh

### Notes

* URLs to retrieve genome and genome annotation files are hardcoded and may need to be updated if new versions have been released since this work was completed.

## Generating STAR index

### Required files

* 10\_STARIndex.sbatch
* concatenated\_Hs\_Pf.fa.gz (generated by 4.1)
* concatenated\_Hs\_Pf.gtf.gz (generated by 4.1)

### Notes

* STAR is in the biology module of SHERLOCK.

## Alignment with STAR

### Required files

* 11\_STARAlignA.sbatch
* Trimmed and (optional) decontaminated .fastq.gz files
* 211209\_DirectoryList.csv
* Genome indices generated in 4.2

### Notes

* This is the most time and memory intensive step in the workflow. The largest .fastq.gz files can take 3-4 hours for alignment.
* Tabulation of counts for each sample is calculated using STAR’s quantMode flag.

# Tabulation of counts

## Tabulation

### Required files

* 12\_TabulateCounts.sh
* .tab files generated for each sample in 4.3
* 211209\_DirectoryList.csv

### Notes

* This script combines the count tables for each sample into a single file. It is the final step conducted on the computing cluster.

# Preparing data for DESeq2 (R)

## Preparing data

### Required files

* 2023APR30\_REFERENCE\_SETUP.ipynb
* Pf\_ASM276v2\_Genes.txt
* PfGeneNames\_plasmodb.csv
* 220103\_HsPf\_2020.txt

### Notes

* This notebook prepares data for analysis in DESeq2. The result is separate count tables for human and *Plasmodium* gene counts. These include only protein-coding and lncRNA genes. The gene information for the genes in the final count tables is also saved.
* This notebook additionally calculates the total and percentage of all reads aligned to the human and *P. falciparum* genome using the unedited counts table.
* File paths are hard-coded and may need to be updated if rerunning the notebook.

# DESeq2

## Running DESeq2 and Exporting CPMs

### Required files

* HumanColData\_DESeq2\_Input.csv
* HumanGeneCountsTable\_v2.csv
* PfGeneCountsTable\_v2.csv
* 2023MAY01\_DESEQ2\_PROCESSING.Rmd

### Notes

* This R notebook runs DESeq2 separately for human and parasite counts. Each cell type is processed as a separate DESeq object because large differences in the erythroblast transcriptome at each stage (Note that 1. most genes change between stages or are entirely stage-specific and most changes are in the same direction as the erythroblasts mature. See also <https://support.bioconductor.org/p/p132527/>).
* Results are output as csv files.

# Data Visualization

## PCA

### Required files

* HumanColData\_DESeq2\_Input.csv
* HumanGeneCountsTable\_v2.csv
* 2023MAY07\_PCA.Rmd

### Notes

* Includes PCA with all samples and option to filter by condition or cell type.

## Heatmaps

### Required files

* HumanColData\_DESeq2\_Input.csv
* HumanGeneCountsTable\_v2.csv
* 2023MAY07\_HEATMAPS.Rmd