# Effect of toxoplasma infection on gene expression in mouse blood and lung

Saskia Perret-Gentil

January 2021

## Abstract

## 1 Introduction

## 2 Material and methods

### 2.1 Data

Data used is a subset from Singhania et al. 2019 [5], composed of fastq files which were downloaded through the Gene Expression Omnibus (GEO), accession GSE119855. The library preparation protocol was strand-specific. The libraries were sequenced on an Illumina HiSeq 4000 in paired-end mode.

The subset includes samples from blood and lung of five mice infected with toxoplasma and three uninfected controls.

The quality of the subset were assess with FastQC [4] (v. 0.11.7).

### 2.2 Mapping reads to the reference genome.

The reference genome (assembly GRCm38) and associated annotation were download from the Ensembl ftp site. Checksums were computed and compared to the values in the CHECKSUMS file on the ftp server.

Files were indexed with SAMtools [2] (v. 1.10). The reads were mapped to the reference genome using Hisat2 [1] (v. 2.2.1). SAMtools was then used to convert the resulting sam files to bam format. Bam files were sorted and indexed, also using SAMtools.

A table of counts containing the number of reads per genes was produced using featureCounts [3] , the bam files and the annotation file.

# 3 Results

## 3.1 Quality checks

## 3.2 Mapping reads to the reference genome.

## 3.3 Count the number of reads per gene

## 3.4 Exploratory data analysis

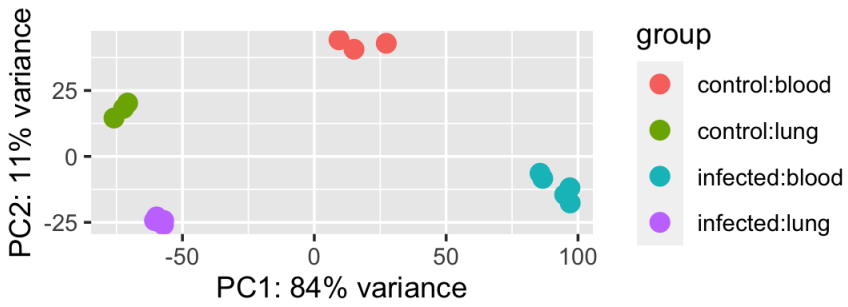


Figure 1: Principal Component Analysis

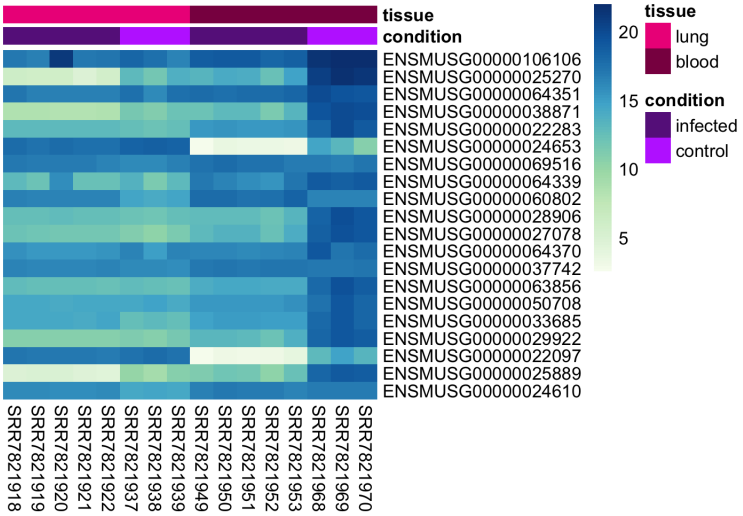


Figure 2: Heatmap

### 3.5 Differential expression analysis

### 3.6 Overrepresentation analysis

## 4 Discussion

## Supplementary materials

All scripts used for this project can be found here: `https://github.com/saskia-droid/toxoplasma_de`.

## References

[1] Daehwan Kim, Ben Langmead, and Steven L Salzberg. "HISAT: a fast spliced aligner with low memory requirements". In: *Nature methods* 12.4 (2015), pp. 357–360.

[2] H Li et al. "The Sequence Alignment/Map format and SAMtools". In: *Bioinformatics* 25.16 (Aug. 2009), pp. 2078–2079. DOI: 10.1093/bioinformatics/btp352. URL: https://www.ncbi.nlm.nih.gov/pubmed/19505943.

[3] Y. Liao, G. K. Smyth, and W. Shi. "featureCounts: an efficient general purpose program for assigning sequence reads to genomic features". In: *Bioinformatics* 30.7 (Nov. 2013), pp. 923–930. DOI: 10.1093/bioinformatics/btt656. URL: https://doi.org/10.1093/bioinformatics/btt656.

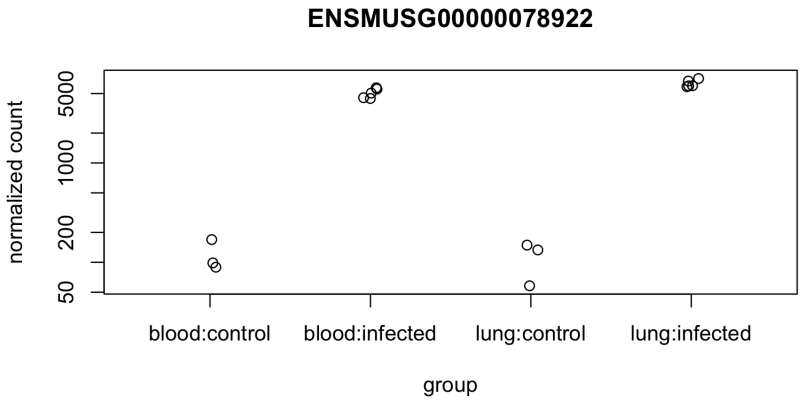


Figure 3: Expression level of Tgtb1 gene.

[4] Andrews S. *FASTQC. A quality control tool for high throughput sequence data.* URL: `http://www.bioinformatics.babraham.ac.uk/projects/fastqc/`.

[5] Akul Singhania et al. "Transcriptional profiling unveils type I and II interferon networks in blood and tissues across diseases". In: *Nature Communications* 10.1 (June 2019), p. 2887. ISSN: 2041-1723. DOI: `10.1038/s41467-019-10601-6`. URL: `https://doi.org/10.1038/s41467-019-10601-6`.

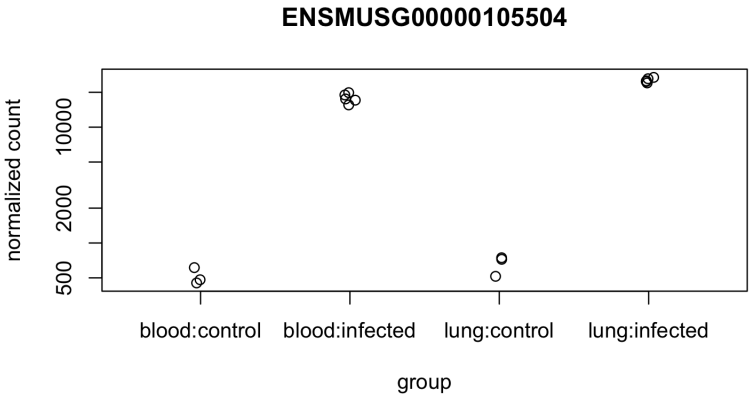


Figure 4: Expression level of Gbp5 gene.

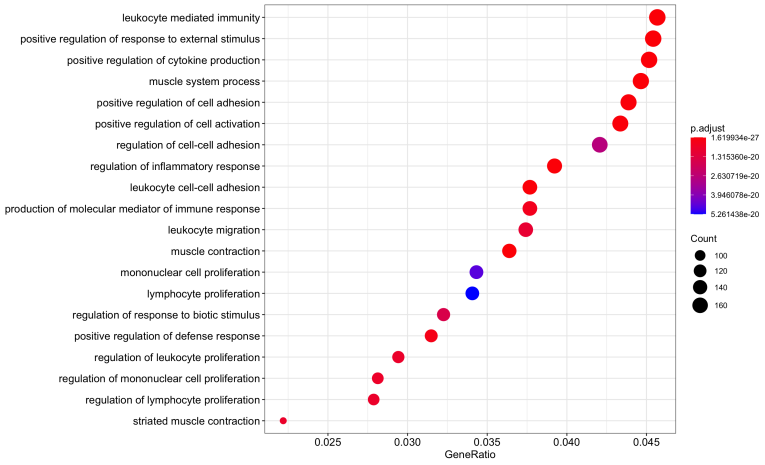


Figure 5: Dot plot