Gene Expression Changes in Toxoplasma-Infected *Mus musculus*

Oriane Kopp

Abstract

Toxoplasma gondii, a parasite with a life cycle involving mice and cats (and humans by mistake), is known for changing mice behaviour to increase predation by cats. Besides behavioural changes, the immune response of the mouse is changed by the infection, as showed by Singhania et al. (2019). Building on this, our study explores differential gene expression in blood and lung tissues from healthy and *T. gondii*-infected mice, using a subset of Singhania et al.'s dataset for RNA-sequencing analysis on 16 samples.

After some quality checks, a differential expression analysis was made. A principal component analysis (PCA) showed that the samples were clustered based on tissue and infection status. In blood, 12384 differentially expressed genes were found (37% upregulated, 63% downregulated), and in lung, 9481 genes were differentially expressed (41% upregulated, 59% downregulated), showing that downregulation is more present than upregulation.

To investigate the immune response, a focus was made on Oas1a and Fcgr1. Both are upregulated in infected tissues, as expected since both genes are linked with immune response. Gene Ontology (GO) analysis highlighted key GO terms like chemotaxis and positive regulation of cell activation, indicating a well establish response to the infection.

Introduction

Toxoplasma gondii, an intracellular protozoan, is a well-known parasite that completes its life cycle by infecting cats (definitive hosts) and mice (intermediate hosts). Importantly, *T. gondii* can also infect humans, with many individuals harbouring the parasite without major consequences, however, it poses a significant threat to those who are immunosuppressed. Upon infecting a mouse, *T. gondii* will change the host's behaviour, such as silencing the fear of novelty, to increase the likelihood of predation by a cat, easing parasite transmission and life cycle completion (Berdoy *et al.,* 2000).

Beyond these behavioural alterations, the host's immune system undergoes a myriad of responses following *T. gondii* infection. One aspect of this immune response involves changes in gene expression as showed by Singhania et al. in their 2019 study. Building upon their data and findings, our study aim is to investigate the complex host's immune response to *T. gondii* infection, focusing on the similarities and differences in gene expression profiles within the lung and blood tissues. Gaining more knowledge it the change in gene expression is essential to understand the immune reaction to a parasite infection.

Utilizing the dataset provided by Singhania et al. (2019), our investigation seeks to identify differentially expressed genes through a differential expression analysis. Specifically, two differential expression analysis will be performed, one between blood samples and one between lung samples from *T. gondii*-infected and healthy mice. Then, both sets of results will be compared to explore differences and similarities between the tissues. A Gene Ontology (GO) analysis will also be performed to identify the most present GO terms in the differentially expressed genes.

In addition to the exploration of immune reaction to *T. gondii*, another goal of this study is to gain knowledge in the tools necessary for RNA-sequencing data analysis and to understand the outcomes obtained.

Material and methods

Dataset:

The dataset consists of 16 bulk-RNA-sequencing samples from mice, a subset of the dataset obtained by Singhania *et al.* (2019). The samples are divided in four groups based on the tissue and infection status: blood control (3 samples), blood case (5 samples), lung control (3 samples) and lung case (5 samples). For each sample, paired-end reads are given.

Quality check:

To address the quality of the reads, a quality check was performed using FastQC (version 0.11.9) and later MultiQC (version 1.8) to summarize all the FastQC reports in one report. The quality was satisfying so no further trimming was made.

Mapping reads to the reference genome:

Before mapping the reads to the reference genome, HISAT2 (version 2.2.1) was used, as well as the *Mus musculus* genome Ensembl GRCm39 (release 110) to create index files.

Using the index files, the reads were mapped to the reference genome, using HISAT2, with the parameter “rna-strandess” set as “RF” since the reads are pair-ends.

The resulting sam files were converted into bam files (binary version of sam files) using SAMtools (version 1.10), and then sorted by genomic coordinates (default coordinates), also using SAMtools. Finally, still using SAMtools, an index file for each bam file was created.

Number of reads per gene:

Then, to count the number of mapped reads per gene, featureCounts (version 2.0.1) was used, along with the Ensembl annotation file corresponding to the assembly, in our case GRCm39 (release 110), in general transfer format (GTF). The settings for featureCounts are “-s” set as “2” because “RF” was used in the mapping, and “-p” because reads are pair-ends. FeatureCounts produces a table of counts containing the number of reads per gene in each sample.

Exploratory data analysis:

The count table from featureCounts was adapted by removing the columns containing “Chr”, “Start”, “End”, “Strand” and “Length”, and renaming the columns. A “column table” was also created with two columns: Sample and Group. The column “Sample” contains the ID of the sample (e.g., SRR7821918) and the column “Group” contains the name of the experimental group in which the sample is from (e.g., Lung_WT_Case).

Then, R (version 4.3.2) in RStudio (version 2023.12.0+369) was used to create the DESeqDataSet (DDS) object using the Bioconductor package DESeq2 (version 1.42.0), inputting the adapted count table, the “column table” and the design set as “Group”.

Still using the package DESeq2, the differential expression analysis was performed. To visualise how the sample clusters looked like based on their gene expression a principal component analysis (PCA) was

made. For that, the dependence of the variance on the mean of the DDS was removed, with the parameter blind set as TRUE to not consider the experimental group.

Differential expression analysis:

The differential expression analysis was run in the previous step. To extract the results, DESeq2 was used to do two comparisons: Blood Case against Blood Control and Lung Case against Lung Control (set in parameter "contrast"). To calculate differentially expressed genes, the Wald test was used in DESeq2. If the log2 fold change (LFC), calculated as follows: $\log_2(\frac{Mean\ Treatment}{Mean\ Control})$, is positive, the gene is upregulated, and if it's negative, the gene is downregulated. To control the false discovery rate (FDR) adjusted p-values were filtered to only keep the ones lower than 0.05 (to guarantee an FDR lower than 5%). Genes symbols were added to the dataset using the package biomaRt (version 2.58.0) and org.Mm.eg.db (version 3.18.0) and then a volcano plot was producing using the package EnhancedVolcano (version 1.20.0) to visualise the differentially expressed genes. Two genes present in the Singhania et al. (2019) article, Oas1a and Fcgr1, were specifically investigated by doing boxplots.

Overrepresentation analysis:

Since many genes were found, a functional analysis was performed to regroup the genes by function. This last step consisted in identifying Gene Ontology (GO) terms that contain more differentially expressed genes than expected by chance. For that, the package clusterProfiler (version 4.10.0) was used. The parameter "ont" (orthogonal ontologies) was set as "BP" for biological processes (all other parameters were default ones). Finally, the top GO terms were extracted by sorting them by GeneRatio (percentage of total DE genes in each GO term) in a dot plot.

Results

Quality check:

To assess the quality of the reads, examination was conducted using MutliQC report (Figure 1). Given the RNA-sequencing data, evaluation included checking only for per base sequence quality, per sequence GC content and the absence of adapter sequences. Except for four reads showing a slightly lower pre sequence GC content, the overall quality of the reads was satisfactory.

The mapping was good quality, based on the output provided by HISAT2. The alignment rates across sample were between 87% and 98%, and between 74% and 90% of the reads were aligned concordantly (meaning that both reads of the pair-ends reads align to the reference genome) exactly one time. There was some multimapping, between 5% and 11% depending on the reads. Since this number is still low, and for most of the reads lower than 10%, it wasn't a problem for downstream analysis.

When counting the number of reads per gene with featureCount the summary output showed that most of the genes overlapped with the annotated genes (between 65% and 85%). On average, 588'304 genes were unassigned due to ambiguity, which means that if a read could correspond to multiple genes, it was left out. Other genes were not assigned due to the multimapping during the alignment.

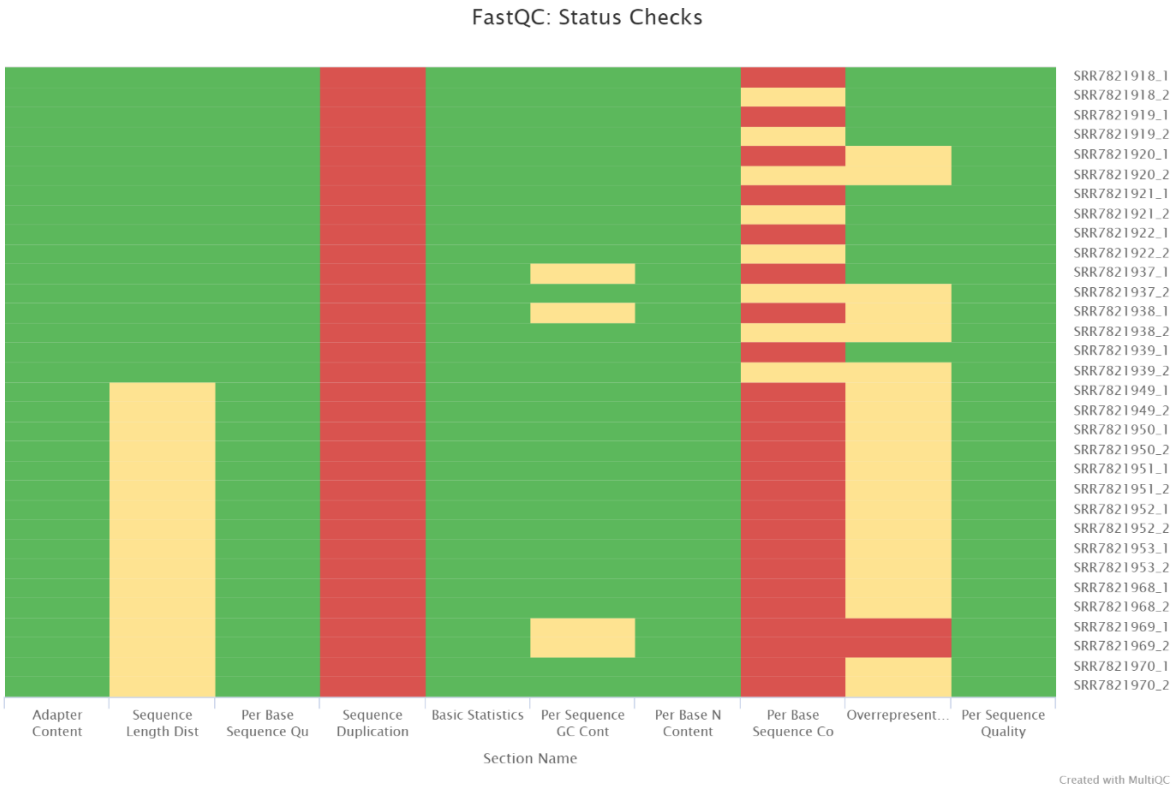

Figure 1: Status Checks from the MultiQC report

Principal component analysis:

In the principal component analysis (PCA) (Figure 2), four clusters based on gene expression profile can be identified. Those clusters correspond to our four experimental groups. 86% of the variance is explained by PC1, which is caused because our samples come from different tissues (blood and lung). However, for both tissues, it seems like they react the same way to the infection; for each tissue, the Case group has a higher value for PC1 but a lower value for PC2 than the Control group.

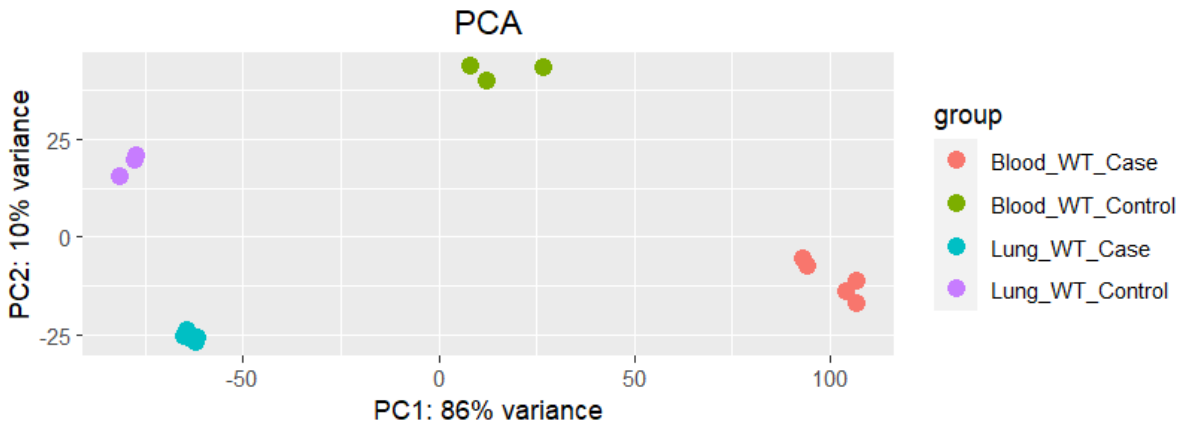

Figure 2: PCA were how the sample are clustered based on their gene expression profile can be seen.

Differentially expressed genes:

In the blood samples, 12384 genes are differentially expressed with an adjusted p-value of 0.05. Among those genes, 4592 (37%) genes are upregulated (LFC > 0) and 7792 (63%) are downregulated (LFC < 0).

In the lung samples, 9481 genes are differentially expressed with an adjusted p-value of 0.05. Among those genes, 3847 (41%) genes are upregulated and 5634 (59%) are downregulated.

As saw in the PCA, both tissues react in an analogous way, more genes are downregulated than upregulated, but in blood, slightly more genes are downregulated. It is more visible in the Volcano plots (Figure 3). For blood tissues, we clearly see a difference between the number of genes downregulated. For lung tissues, it is harder to see because the upregulated genes have a wider range of p-value than the downregulated genes. By looking at the LFC, we can see that in both cases, more downregulated genes have a bigger absolute LFC than the upregulated genes, meaning that the downregulation is stronger than the upregulation.

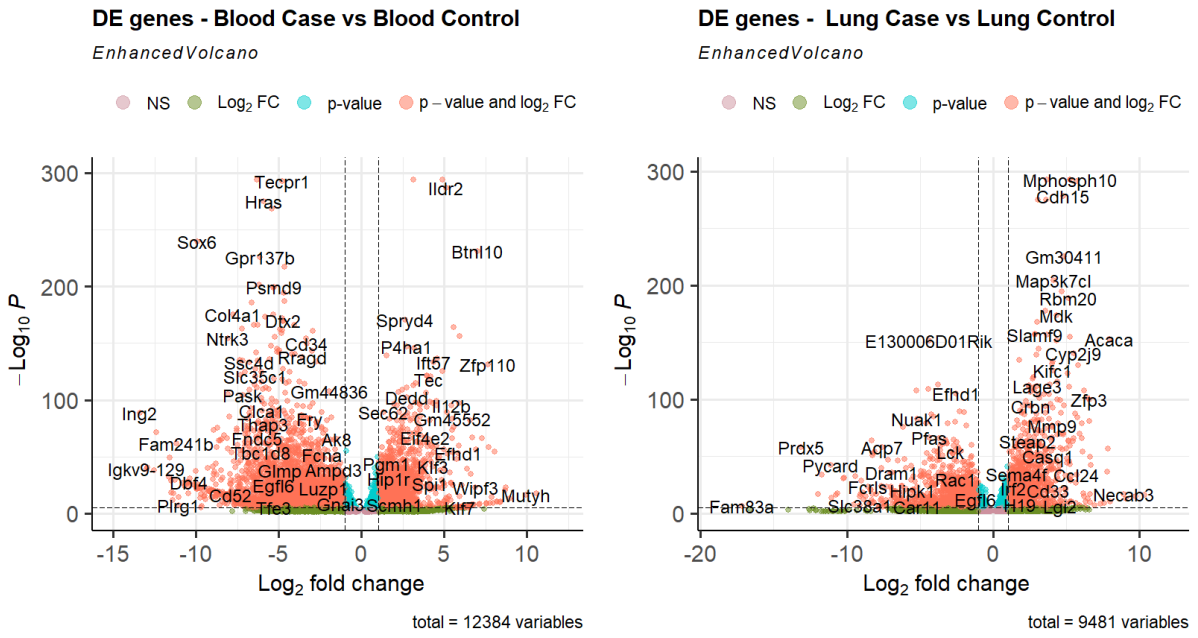

Figure 3: Volcano plots of differentially expressed genes. On the left for blood and on the right for lung.

Investigation of two specific genes

Based on the original publication, two genes were chosen for a deeper investigation. Those genes are Oas1a and Fcgr1. They both take part in immune response. Oas1a is involved in the innate response to viral infections and Fcgr1 in innate and adaptative immune responses. On Figure 4, we can see that both genes are more expressed in Case sample than in Control sample, and more expressed in blood samples. Since they are linked with immune response, it makes sense that they are more present when an infection is present, even if Oas1a is linked with antiviral response because it also plays a role in regulation, so even under a non-viral infection like *T. gondii*, it makes sense that it is over expressed (UniProt).

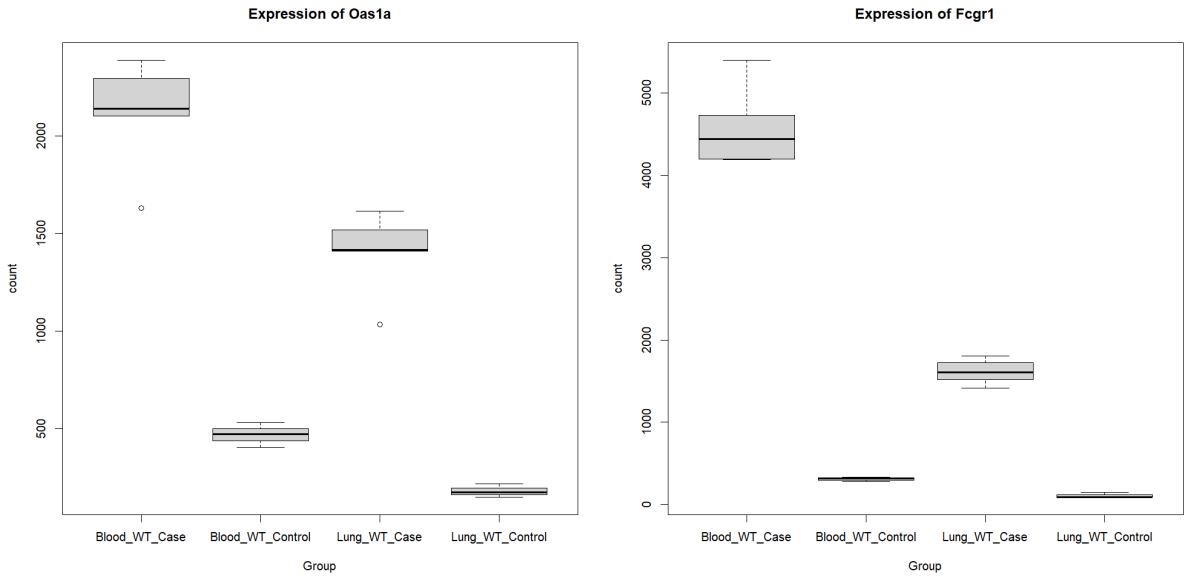

Figure 4: Expression of Oas1a (on the left) and Fcgr1 (on the right) in all groups

Gene Ontology (GO) terms

In the top GO terms of the two groups (lung and blood), four of them are present in both: taxis, chemotaxis, positive regulation of cell activation and positive regulation of leukocyte activation (Figure 5).

Taxis signifies the movement of a cell or organism in response to an external stimulus, while chemotaxis is a subtype where movement is guided by a chemical concentration gradient. Positive regulation of cell activation encompasses processes that activates or amplify the frequency or rate of activation. Positive regulation of leukocyte activation represents any process that triggers an increase in the frequency or rate of leukocyte activation.

In the lung tissues, the main GO term is regulation of innate immune response. This term stands for processes modulating the frequency, rate, or extent of the innate immune response.

The leading GO term in blood is ameboidal-type cell migration, a process involving cell migration through extension and retraction of pseudopodium (GeneOntology).

These GO terms align with the context of an infection. An increase in leukocyte and the regulation of immune response are obvious reaction against the parasite. Taxis and chemotaxis show that the organism react to the infection (stimulus), and the migration of ameboidal-type cells allows immune cells to reach the point of the infection.

We can also see that lung tissues shows more GO terms linked with immune response, while blood tissues show terms associated with positive regulation/activation.

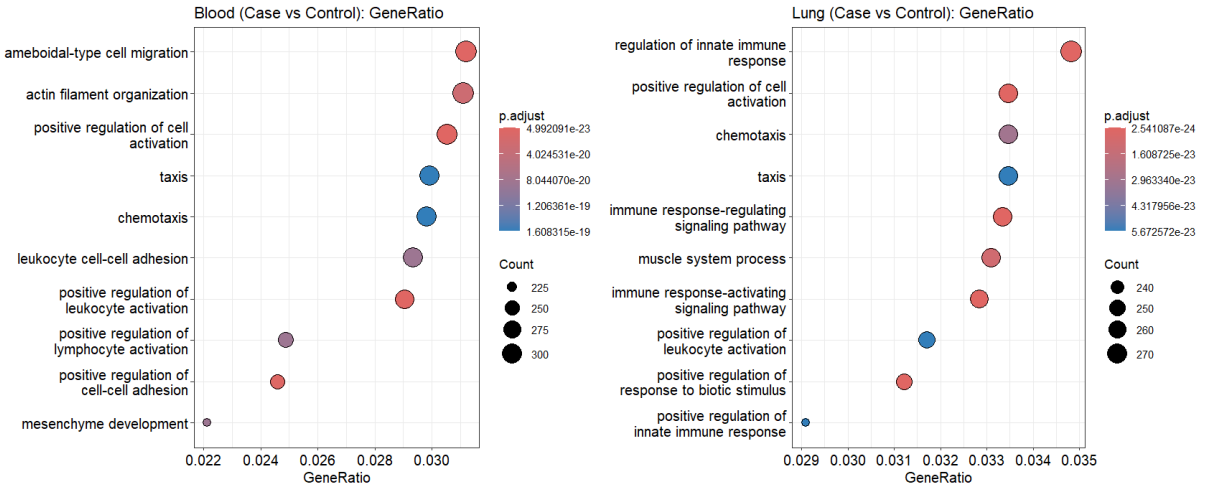

Figure 5: Top GO terms in blood (on the left) and lung (on the right), sorted by GeneRatio

Discussion & Conclusion

Overall, we saw that blood and lung tissues react in a comparable way to the infection in the differential expression patterns: approximatively 2/5 of the genes are upregulated, and 3/5 downregulated. Some of the functions of the differential expressed genes are different between tissues: lung shows more functions linked with immune reaction and blood with positive regulation/activation, but they also share some function, showing a clear reaction against the parasite. The two investigated genes, Oas1a and Fcgr1, aligned with our findings; they are both genes linked with immune response and are more present in infected tissues.

Some findings are different from what expected form the article of Singhania et al. (2019). Mostly, the number of differentially expressed genes. In the article, 3566 and 5052 differentially expressed genes were found, respectively for lung and blood tissues. Here, those numbers are 9'481 and 12'384. This is caused by the fact that different filtering were performed. In the original study, the LFC was filtered to only keep genes with a LFC bigger than 1 or smaller than -1. This filtering was not performed here, that's why way more differentially expressed genes were found.

The number of counts for the Oas1a and Fcgr1 genes are also slightly different in the lung (the information for blood is not given for those genes in the original paper), however, those differences are probably only caused by the different tools used to count the number of reads per gene; they used HtSeq but featureCounts was used here. Other reasons are the fact that different version of the tools and a different version of the *Mus musculus* genome were used (GRCm38 release 86 in the original study and GRCm39 release 110 in this study).

Finally, and most important a lot of knowledge in RNA-sequencing analysis was obtained by doing this study. *I* feel more confident with the tools used and the way to interpret the results obtained.

Supplementary materials

GitHub repository: https://github.com/lieselty/RNA_seq_course

References

Ashburner, M., et al. (2000). Gene ontology: tool for the unification of biology. Nature Genetics, 25(1), 25-29. https://doi.org/10.1038/75556

Berdoy, M., Webster, J. P., & Macdonald, D. W. (2000). Fatal Attraction in Rats Infected with Toxoplasma Gondii. Proceedings of the Royal Society B: Biological Sciences, 267(1452), 1591–1594. https://doi.org/10.1098/rspb.2000.1182

Singhania, A., Graham, C. M., Gabryšová, L., et al. (2019). Transcriptional profiling unveils type I and II interferon networks in blood and tissues across diseases. Nature Communications, 10, 2887. https://doi.org/10.1038/s41467-019-10601-6

The Gene Ontology Consortium. (2023). The Gene Ontology knowledgebase in 2023. Genetics, 224(1), iyad031. https://doi.org/10.1093/genetics/iyad031

The UniProt Consortium. (2023). UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Research, 51, D523–D531. https://doi.org/10.1093/nar/gkac123

Tools

Andrews, S. (2010). FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Kim, D., Paggi, J. M., Park, C., et al. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nature Biotechnology, 37, 907–915. https://doi.org/10.1038/s41587-019-0201-4

Ewels, P., Magnusson, M., Lundin, S., Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics, 32(19), 3047–3048. https://doi.org/10.1093/bioinformatics/btw354

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. GigaScience, 10(2), giab008. https://doi.org/10.1093/gigascience/giab008

Liao, Y., Smyth, G. K., Shi, W. (2014). featureCounts: an efficient general-purpose program for assigning sequence reads to genomic features. Bioinformatics, 30(7), 923–930. https://doi.org/10.1093/bioinformatics/btt656

R and R packages

Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A., & Huber, W. (2005). BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics, 21, 3439-3440. https://doi.org/10.1093/bioinformatics/bti525

Blighe, K., Rana, S., & Lewis, M. (2023). EnhancedVolcano: Publication-ready volcano plots with enhanced coloring and labeling. doi:10.18129/B9.bioc.EnhancedVolcano, R package version 1.20.0, https://bioconductor.org/packages/EnhancedVolcano.

Carlson, M. (2019). org.Mm.eg.db: Genome wide annotation for Mouse. R package version 3.18.

Yu, G., Wang, L., Han, Y., & He, Q. Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS: A Journal of Integrative Biology, 16(5), 284-287.

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology, 15, 550. https://doi.org/10.1186/s13059-014-0550-8

Durinck, S., Spellman, P. T., Birney, E., & Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nature Protocols, 4, 1184-1191. https://doi.org/10.1038/nprot.2009.97

R Core Team. (2023). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

RStudio Team. (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA. URL http://www.rstudio.com/.

Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., Feng, T., Zhou, L., Tang, W., Zhan, L., Bo, X., & Yu, G. (2021). clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. The Innovation, 2(3), 100141.