RNA Sequencing 2025

Andri L. Widmer 20-105-581

Abstract

1. Introduction

Host–pathogen interactions induce complex transcriptional responses that are essential for effective immune defence and strongly influence disease outcome (Janeway and Medzhitov, 2002). The intracellular parasite Toxoplasma gondii infects a wide range of warm-blooded hosts and can colonise multiple tissues, including the lung. Pulmonary infection is associated with pronounced immune activation and inflammation, making lung tissue a suitable model for studying host transcriptional responses to T. gondii infection (Dubey, 2016).

The RNA-Seq data analysed in this project originate from a study by Singhania et al., which investigated immune responses to T. gondii infection in mice with different genetic backgrounds. In particular, the study compared wild-type (WT) animals to a double knockout (DKO) model lacking both the type I interferon receptor (Ifnar /) and the type II interferon receptor (Ifngr /). Interferon signalling plays a central role in host defence against intracellular pathogens, and disruption of both interferon pathways is expected to substantially alter transcriptional responses to infection. In the context of this practical course, RNA-Seq subsamples derived from this experimental setup were provided by the university and re-analysed to address defined analytical questions.

RNA sequencing (RNA-Seq) enables quantitative, genome-wide profiling of gene expression and has become a standard method for analysing transcriptional changes across biological conditions (Wang et al., 2009; Conesa et al., 2016). Its unbiased nature makes RNA-Seq particularly suitable for investigating complex processes such as infection-induced immune responses, where coordinated regulation of many genes is expected.

In this project, RNA-Seq data derived from lung tissue were analysed to characterise host transcriptional changes associated with T. gondii infection. For both WT and DKO mice, infected (“case”) samples were compared to uninfected (“control”) samples to identify genes that are differentially expressed upon infection. By analysing the two genotypes separately, this study aims to assess how the absence of interferon signalling influences the magnitude and nature of the transcriptional response to T. gondii.

The overall aim of this project was to compare lung-specific transcriptional responses to Toxoplasma gondii infection between wild-type and interferon receptor–deficient mice. Rather than testing predefined hypotheses, the analysis follows an exploratory approach to characterise how disruption of type I and type II interferon signalling affects infection-induced gene expression in lung tissue. To this end, a standard RNA-Seq analysis workflow was applied, including quality control of sequencing data, read alignment to a reference genome, gene-level quantification, and statistical testing for differential expression. Emphasis was placed on reproducibility, appropriate statistical analysis, and careful assessment of data quality to enable a robust comparison of transcriptional responses between genotypes.

2. Materials and Methods

2.1 Dataset and experimental design

This project analysed strand-specific, paired-end Illumina HiSeq 4000 RNA-seq data derived from lung tissue of mice representing two genotypes: wildtype (WT) and interferon alpha/gamma receptor double knockout (DKO). For each genotype, both uninfected control animals and *Toxoplasma gondii*–infected mice were included, with 3–5 biological replicates per condition. The FASTQ files used in this analysis correspond to a subset of samples generated in the study by Singhania et al. (2019) and were accessed on the IBU cluster under `/data/courses/rnaseq_course/toxoplasma_de`. All scripts and code used throughout the workflow are available in the public GitHub repository: https://github.com/awidmer123/RNA_Seq.

2.2 Computational environment and workflow organisation

All computationally intensive steps were executed on the IBU high-performance computing cluster using the Slurm workload manager. A structured project directory was created under `/data/users/awidmer/RNA_Seq`, following a consistent organisation by analysis step (quality control, alignment, BAM processing, and read counting).

To ensure full reproducibility, all tools were run inside predefined Apptainer containers, which provide stable versions of all required software independent of system configuration. Containerised versions of FastQC, HISAT2, SAMtools, and featureCounts were used throughout the workflow. MultiQC was used to aggregate quality control metrics across samples.

Detailed software versions for all cluster-based tools as well as R, Bioconductor, and the R packages used for downstream analysis are documented in the project repository under `results/versions/`. This includes the exact versions of FastQC, HISAT2, SAMtools, featureCounts, MultiQC, R, Bioconductor, and all relevant R packages to ensure full reproducibility of the analysis.

2.3 Quality control of raw sequencing reads

Initial quality assessment of the raw FASTQ files was performed using FastQC, which provided information on sequencing depth, per-base quality profiles for both mates, GC content, and the presence of adapter-derived sequences. The individual reports were merged using MultiQC to obtain an overview across all samples. No severe quality issues were detected, and all samples exhibited consistently high base qualities along the read length. As a result, no trimming or preprocessing was required prior to alignment.

2.4 Reference genome acquisition and index preparation

The *Mus musculus* reference genome (GRCm39) and corresponding GTF annotation were downloaded from the Ensembl FTP server, following workflow recommendations. Integrity of both files was verified via checksum comparison using the `sum` utility. The FASTA file was subsequently used to build a HISAT2 genome index, which is required for performing splice-aware alignment of the RNA-seq reads.

2.5 Read alignment with HISAT2

Reads from each sample were aligned individually to the reference genome using HISAT2 executed via its Apptainer container. Resource usage followed the recommended cluster settings. Alignment produced SAM output files, which were converted into BAM format using Samtools. Mapping quality metrics—such as alignment rate, proportion of properly paired reads, and frequency of multimappers—were inspected for each sample to confirm the overall quality of the alignment process.

2.6 BAM processing with Samtools

Post-alignment processing involved three standard Samtools operations: (i) conversion of SAM to BAM using `samtools view`; (ii) coordinate sorting of each BAM file using `samtools sort`; and (iii) index generation using `samtools index`. Sorted and indexed BAM files are required for efficient downstream processing, in particular for accurate read counting. All operations completed successfully using the recommended memory allocations for the IBU cluster.

2.7 Gene-level quantification with featureCounts

Gene-level read quantification was performed using featureCounts from the Subread package. The Ensembl GTF annotation was provided to define gene boundaries, and counting was conducted in paired-end and strand-specific mode to match the library preparation protocol. The resulting count table included one column per sample and reported the number of reads

unambiguously assigned to each gene. FeatureCounts summary statistics—such as the number of assigned and unassigned reads and the fraction of multimapping or ambiguous alignments—were reviewed to ensure successful quantification.

2.8 Data processing and normalization in R

Downstream analysis was conducted locally in RStudio using the DESeq2 package. The featureCounts output was imported, and non-quantitative annotation columns (chromosome, start, end, strand, length) were removed. A `DESeqDataSet` object was created using a design formula modelling the experimental condition, which encoded each sample as WT_Control, WT_Case, DKO_Control, or DKO_Case.

DESeq2's internal filtering removed genes with insufficient read support. Library size differences were normalised via DESeq2's size-factor estimation. For exploratory analysis, a variance stabilising transformation (VST) was applied with `blind = TRUE`, ensuring that no group information biased variance estimation.

2.9 Exploratory data analysis

Principal component analysis (PCA) and sample-to-sample distance heatmaps were generated using VST-transformed counts. These visualisations revealed how samples cluster based on their global gene expression profiles and allowed the assessment of replicate consistency and separation of experimental groups. Lung samples clustered according to both infection status and genotype, in line with expectations for immune-related transcriptomic responses.

2.10 Differential expression analysis

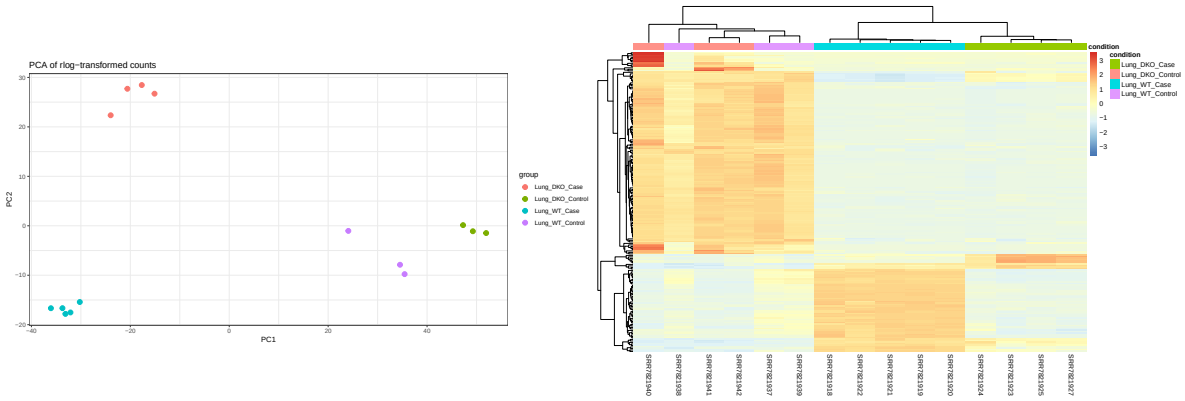
Differential expression analysis was performed using DESeq2's Wald test. Two contrasts were evaluated independently: (i) WT Case vs WT Control, and (ii) DKO Case vs DKO Control. Genes were considered differentially expressed at a false discovery rate of `padj < 0.05`. No additional log2 fold-change threshold was applied. For each contrast, the total number of significant DEGs as well as the counts of up- and down-regulated genes were recorded. Normalised counts of selected biologically relevant genes were inspected to support interpretation of genotype-specific responses.

2.11 Gene Ontology enrichment analysis

Functional enrichment analysis was conducted using the clusterProfiler package, specifically the `enrichGO` function. The list of differentially expressed genes (Ensembl IDs) served as input, while the set of all detected genes formed the universe background. The *Mus musculus*

annotation database `org.Mm.eg.db` provided the necessary GO mappings. The Biological Process (BP) ontology was examined, and resulting GO terms were summarised and visualised using standard clusterProfiler plotting functions. Interpretation focused on immune-related processes known to be involved in *Toxoplasma* infection and interferon signalling.

3. Results



(a) PCA (b) Heatmap of top 200 variable genes

Figure 1: PCA and Heatmap

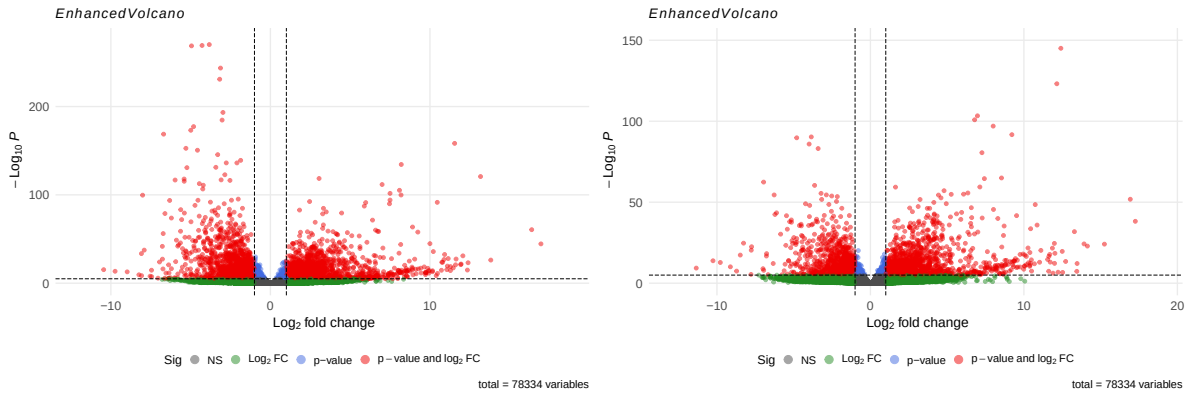
As shown in Figure 1 (a), samples cluster according to both genotype and infection status. WT and DKO samples separate along the first principal component, while infected and control samples show additional separation along the second component. Biological replicates group closely together, indicating good reproducibility and low within-group variability.

Figure Figure 1 (b) shows the expression patterns of the 20 most variable genes across all samples. Hierarchical clustering reveals a clear separation of samples according to genotype and infection status. Replicates within the same condition cluster closely together, while distinct expression profiles are observed between WT and DKO samples.

Table 1: The number of significantly differentially expressed genes (DEGs; padj < 0.05) was quantified separately for WT case vs control and DKO case vs control, and further split into upregulated and downregulated transcripts based on the sign of the log fold change.

Comparison	DE_genes	Upregulated	Downregulated
WT Control vs Case	10618	5663	4955
DKO Control vs Case	11059	5713	5346

As shown in Table 1, both contrasts (WT and DKO) yield a substantial number of significantly differentially expressed genes. In each comparison, DEGs include both upregulated and downregulated transcripts, indicating that infection is associated with broad transcriptional shifts rather than changes restricted to a single direction. This table provides a compact overview of DEG counts used to contextualize the downstream functional enrichment analyses.



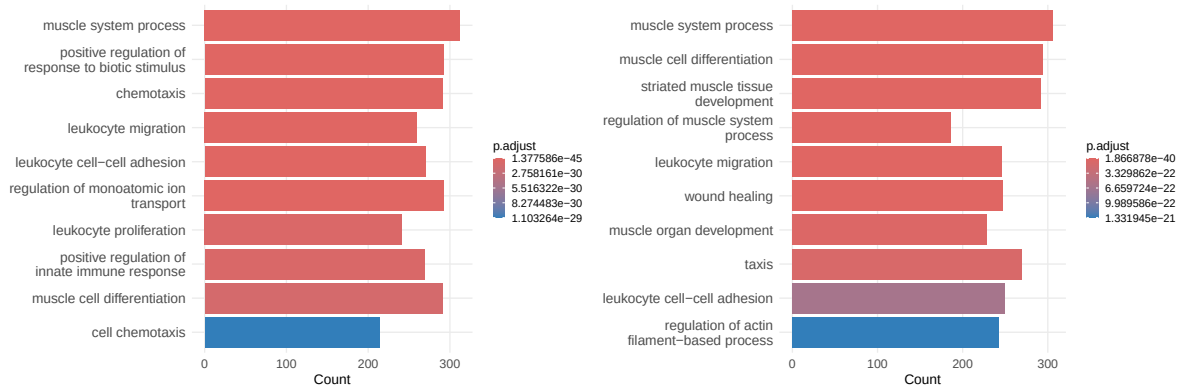
(a) WT control vs case

(b) DKO control vs case

Figure 2: Volcano plots for WT and DKO lung samples following infection.

As shown in Figure 2 (a), infection of WT mice leads to widespread changes in gene expression. Both upregulated and downregulated genes are observed, with a substantial number of transcripts reaching statistical significance. The distribution of log fold changes indicates a strong transcriptional response to infection in WT lung tissue.

As shown in Figure 2 (b), infection of DKO mice results in differential expression of a large number of genes. Compared to the WT contrast, the overall distribution and magnitude of log fold changes differs, indicating an altered transcriptional response in the absence of interferon signalling. Numerous genes reach statistical significance, reflecting a strong but distinct response to infection.



(a) WT control vs case

(b) DKO control vs case

Figure 3: Gene Ontology (Biological Process) enrichment of significantly differentially expressed genes (padj < 0.05) in WT (left) and DKO (right) lung samples. Terms were simplified to reduce redundancy (Wang similarity, cutoff $= 0.7$).

As shown in Figure 3 (a), differentially expressed genes in WT samples are significantly enriched for multiple Gene Ontology Biological Process terms. The enriched categories reflect biological processes associated with the transcriptional response to infection. Only the most significant non-redundant GO terms are displayed.

As shown in Figure 3 (b), differentially expressed genes in DKO samples are enriched for multiple Gene Ontology Biological Process terms. The enrichment profile differs from that observed in WT samples, indicating altered functional responses to infection. Only the most significant non-redundant GO categories are displayed.

Table 2: Selected genes significantly up- or downregulated in DKO case versus control samples. Genes were selected based on adjusted p-value and log fold change criteria and include interferon- and immune-related candidates reported in Singhania et al. (2019).

contrast	ensembl_id	gene_name	module	baseMean	log2FoldChange	padj
WT: infected vs control	ENSMUSG00000037321	Tap1	L7	10336.1841	-3.178026	0.0000000
WT: infected vs control	ENSMUSG00000040033	Stat2	L7	5208.4471	-2.096848	0.0000000
WT: infected vs control	ENSMUSG00000025498	Irf7	L5	4253.8489	-3.418282	0.0000000
WT: infected vs control	ENSMUSG00000105504	Gbp5	L7	13707.9481	-5.408531	0.0000000
WT: infected vs control	ENSMUSG00000028270	Gbp2	L7	46479.3146	-5.402905	0.0000000
WT: infected vs control	ENSMUSG00000045932	Ifit2	L7	4573.5386	-3.656165	0.0000000
WT: infected vs control	ENSMUSG00000024338	Psmb8	L7	6274.8895	-2.528669	0.0000000
WT: infected vs control	ENSMUSG00000027514	Zbp1	L5	4615.9741	-4.462999	0.0000000
WT: infected vs control	ENSMUSG00000105096	Gbp10	L7	2025.9729	-6.207112	0.0000000
WT: infected vs control	ENSMUSG00000028268	Gbp3	L7	8398.9121	-3.724148	0.0000000
DKO: infected vs control	ENSMUSG00000041827	Oasl1	L5	547.9058	-1.829155	0.0000000
DKO: infected vs control	ENSMUSG00000032690	Oas2	L5	2195.9083	-1.699226	0.0000001
DKO: infected vs control	ENSMUSG00000032661	Oas3	L5	1955.2858	-2.254911	0.0000002
DKO: infected vs control	ENSMUSG00000034855	Cxcl10	L7	4153.8464	-2.041226	0.0001443
DKO: infected vs control	ENSMUSG00000000386	Mx1	L5	745.8786	1.359264	0.0008799
DKO: infected vs control	ENSMUSG00000105096	Gbp10	L7	2025.9729	-2.171321	0.0037565
DKO: infected vs control	ENSMUSG00000029417	Cxcl9	L7	6678.1878	-1.738857	0.0060771
DKO: infected vs control	ENSMUSG00000020641	Rsad2	L5	1852.1554	1.117234	0.0213775
DKO: infected vs control	ENSMUSG00000031972	Acta1	L24	9001.3732	12.409696	0.0000000
DKO: infected vs control	ENSMUSG00000044041	Krt13	L24	13661.5094	12.147003	0.0000000

Table 2 lists a subset of genes that are significantly differentially expressed in WT and DKO case versus control samples. The table highlights both upregulated and downregulated transcripts, including several genes previously implicated in immune-related responses. These genes were selected to illustrate representative expression changes observed in the DKO condition.

4. Discussion

We see that the major differences between the doublo knock out mutants and their control reference are on genes linked to this and that pathway. placeholder placeholder placeholder placeholder placeholder

5. Conclusion

In conclusion this study shows that bibedibabedibubedi. placeholder placeholder placeholder placeholder placeholder placeholder