RNA Sequencing 2025

Andri L. Widmer 20-105-581

Abstract

1. Introduction

Toxoplasma gondii is an obligate intracellular protozoan parasite capable of infecting a wide range of warm-blooded hosts, including humans. Infection with T. gondii elicits strong host immune responses and can lead to pronounced inflammation, particularly during acute stages of infection (Dubey, 2016). Effective control of this pathogen relies heavily on coordinated innate and adaptive immune mechanisms, with interferon signalling playing a central role in host defence against intracellular parasites.

Although T. gondii can disseminate systemically, the lung represents a particularly relevant tissue for studying infection-induced immune responses. Pulmonary infection is associated with strong immune activation and inflammatory signalling, making lung tissue a suitable model for investigating transcriptional changes during host–pathogen interactions (Dubey, 2016).

Type I and type II interferons are key mediators of anti-parasitic immunity and inflammation. While type I interferons primarily modulate early innate immune responses, type II interferon (interferon-) is essential for activating macrophages and restricting intracellular pathogen replication. Disruption of these pathways is therefore expected to substantially alter transcriptional programmes during infection (Janeway and Medzhitov, 2002).

RNA sequencing (RNA-Seq) enables unbiased, genome-wide quantification of gene expression and has become a standard approach for analysing transcriptional responses across biological conditions (Wang et al., 2009). Owing to its sensitivity and broad dynamic range, RNA-Seq is especially well suited for studying complex immune responses, where coordinated regulation of large gene sets is expected (Conesa et al., 2016).

The RNA-Seq data analysed in this project originate from a study by Singhania et al., which comprehensively investigated immune responses to T. gondii infection in mice with different genetic backgrounds. In that study, wild-type (WT) animals were compared to a double knockout (DKO) model lacking both the type I interferon receptor (Ifnar /) and the type II interferon receptor (Ifngr /), enabling a broad analysis of interferon-dependent transcriptional responses to infection across multiple tissues.

In this practical course, RNA-Seq subsamples derived from the experimental setup of Singhania et al. were provided by the university and re-analysed. The present project focuses specifically on lung tissue and aims to compare bulk transcriptional responses between infected ("case") and uninfected ("control") samples in WT and DKO mice. By analysing the two genotypes separately, this study seeks to assess how the absence of both type I and type II interferon signalling influences the magnitude and nature of infection-induced gene expression changes in the lung.

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. 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. The general workflow could be found in the GitHub repository at ´resources/DEAnalysis-Wordflow.html´.

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 performed 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 and BAM processing with Samtools

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. 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. 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 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.

2.7 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 regularized log transformation (rlog) was applied.

2.8 Exploratory data analysis and Differential expression 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.

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 differentially expressed genes (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.

2.9 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*

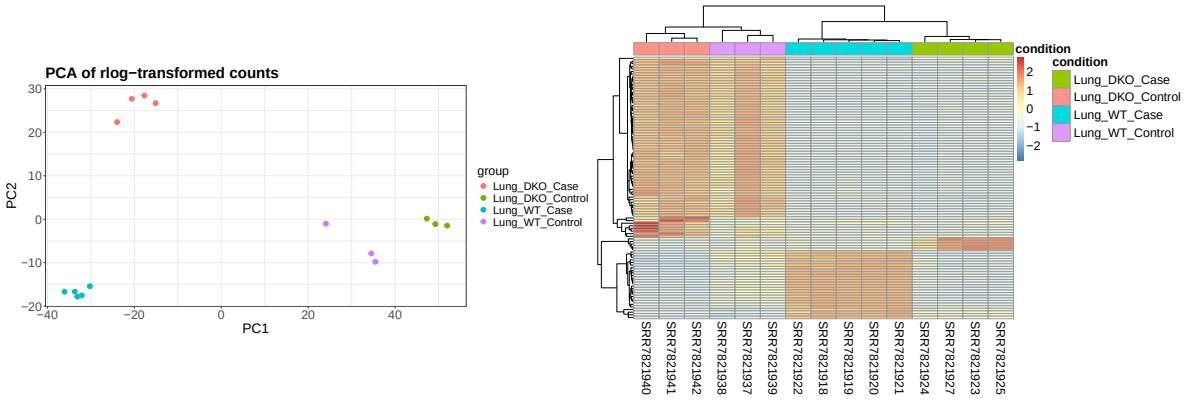
annotation database `org.Mm.eg.db` provided the necessary GO mappings. The Biological Process (BP) ontology was examined, and resulting GO terms were summarised and visualised using standard clusterProfiler plotting functions. Interpretation focused on immune-related processes known to be involved in *Toxoplasma* infection and interferon signalling based on Singhania et al. (2019).

3. Results

3.1 Quality control and read alignment

Quality control of the raw sequencing data was performed using *MultiQC*. Assignment rates ranged from 64 to 76 % for 14 samples, while one sample exhibited a lower assignment rate of 59.9 %. Overall sequence quality metrics were high and no major outlier samples were detected. However, *Per Base Sequence Content* analysis revealed pronounced nucleotide composition bias at the beginning of reads across all samples. Specifically, the first ~10–15 bases showed strong deviations from equal base representation, after which nucleotide frequencies stabilized along the remainder of the read length. This pattern was consistent across all libraries and is characteristic of RNA-seq data, likely reflecting library preparation–associated biases rather than technical artifacts. Consequently, all samples were retained for downstream analyses. The full MultiQC report is available in the GitHub repository at `results/multiqc_report_1.html`.

3.2 Exploratory data analysis



(a) PCA (b) Heatmap of top 100 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.

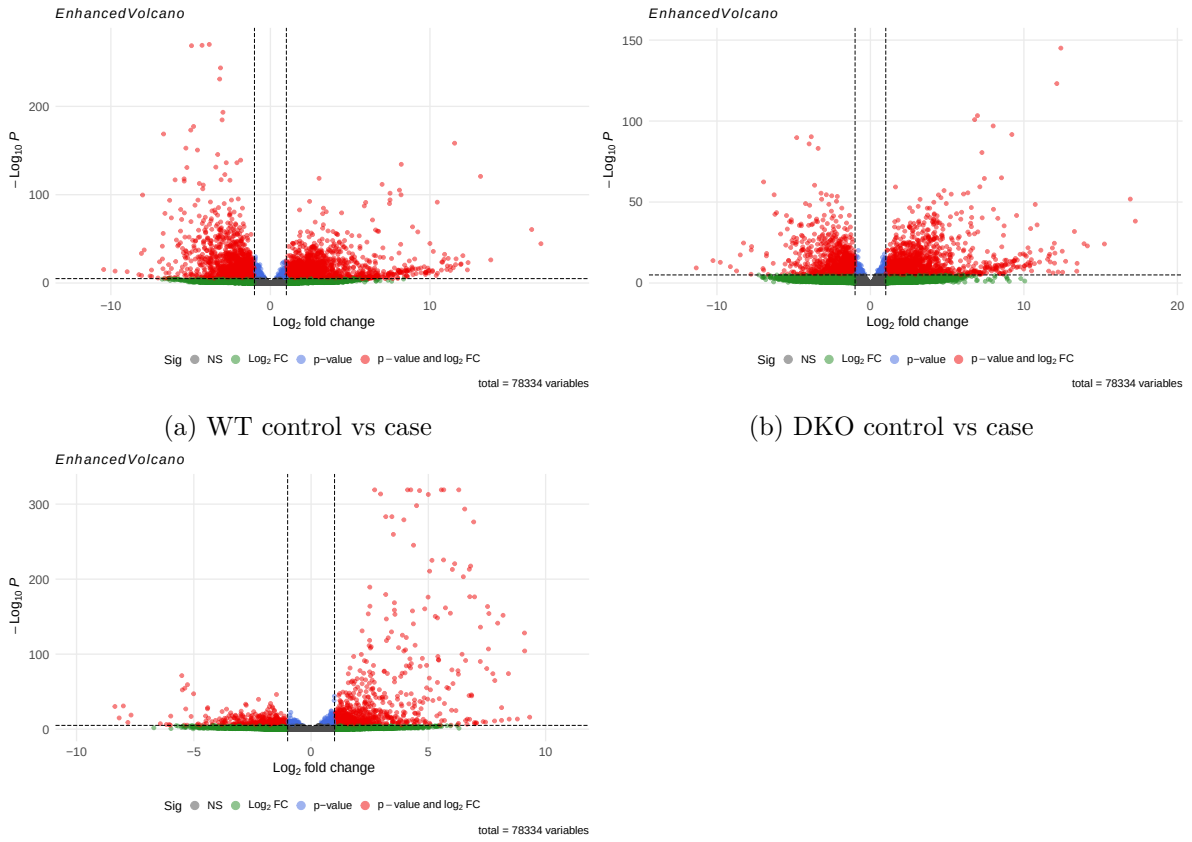
Figure Figure 1 (b) shows the expression patterns of the 20 most variable genes across all samples. Hierarchical clustering reveals a clear separation of samples according to genotype and infection status. Replicates within the same condition cluster closely together, while distinct expression profiles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

3.3 Differential gene expression analysis

Table 1: Number of significantly differentially expressed genes (DEGs; padj < 0.05) identified for infection-induced contrasts within genotypes (WT control vs case, DKO control vs case) and for the direct comparison between infected WT and infected DKO samples. DEGs are further classified as up- or downregulated 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
WT Case vs DKO Case	7689	3536	3536

As summarised in Table 1, infection induced widespread transcriptional changes in lung tissue in both WT and DKO mice, with more than 10,000 genes showing significant differential expression in each genotype. In contrast, the direct comparison between infected WT and infected DKO samples yielded a smaller, yet substantial, set of differentially expressed genes, indicating genotype-dependent differences in transcriptional responses specifically under infectious conditions.



(a) WT control vs case

(b) DKO control vs case

(c) WT case vs DKO case

Figure 2: Volcano plots illustrating differential gene expression in lung tissue following Toxoplasma gondii infection. Shown are (a) WT control vs case, (b) DKO control vs case, and (c) WT case vs DKO case. Each point represents one gene; the x-axis shows log fold change and the y-axis shows −log adjusted p-values (DESeq2 Wald test).

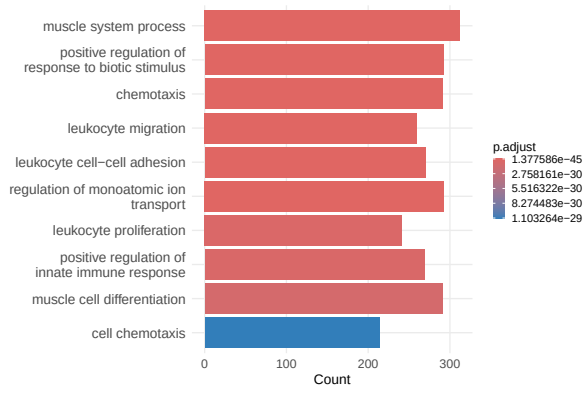
As shown in Figure 2 (a), infection of WT mice leads to widespread transcriptional changes in lung tissue, with numerous genes displaying both large effect sizes and high statistical significance. This reflects a robust transcriptional response to T. gondii infection in WT animals.

In DKO mice Figure 2 (b), infection also results in extensive differential gene expression.

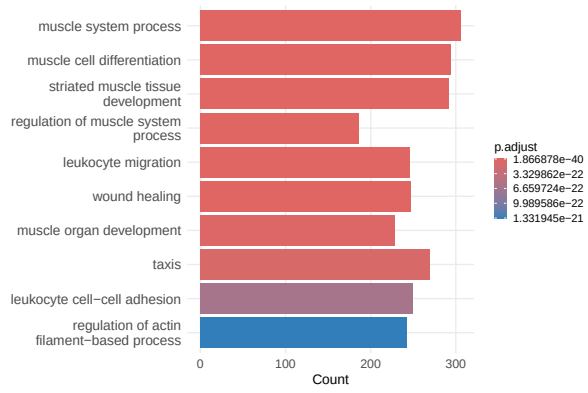
Compared to WT mice, the overall distribution and magnitude of log fold changes differ, suggesting that the absence of type I and type II interferon signalling alters the transcriptional response to infection.

Importantly, the direct comparison of infected WT and infected DKO samples in Figure 2 (c) reveals pronounced genotype-dependent differences under infectious conditions. A substantial number of genes show strong differential expression between the two genotypes, supporting the conclusion that interferon receptor deficiency markedly reshapes infection-induced transcriptional programmes in lung tissue.

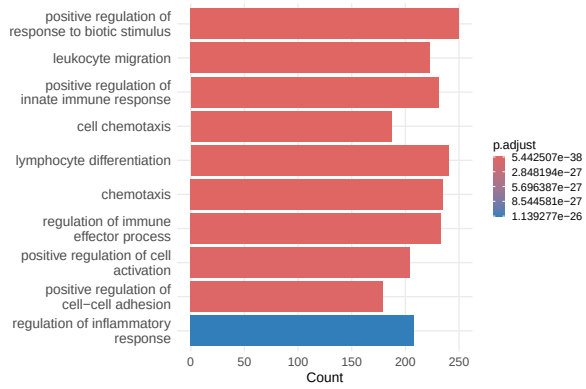
3.4 Functional enrichment analysis



(a) WT control vs case



(b) DKO control vs case



(c) WT case vs DKO case

Figure 3: Gene Ontology (Biological Process) enrichment of significantly differentially expressed genes (padj < 0.05) in WT (a), DKO (b) and WT case vs DKO case (c) 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.

3.5 Comparison to Singhania et al.

Table 2: Selected genes showing significant differential expression in WT control vs case, DKO control vs case, and WT case vs DKO case comparisons. Genes were selected based on adjusted p-values and log fold changes and include interferon- and immune-related candidates previously reported in Singhania et al. (2019). For each gene, the corresponding contrast, Ensembl identifier, gene name, base mean expression, log fold change, and adjusted p-value are shown.

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
WT infected vs DKO infected	ENSMUSG00000037321	Tap1	L7	10336.1841	4.251950	0.0000000
WT infected vs DKO infected	ENSMUSG00000040033	Stat2	L7	5208.4471	2.964662	0.0000000
WT infected vs DKO infected	ENSMUSG00000025498	Irf7	L5	4253.8489	4.997160	0.0000000
WT infected vs DKO infected	ENSMUSG00000034459	Ifit1	L5	2729.7151	6.123893	0.0000000
WT infected vs DKO infected	ENSMUSG00000105504	Gbp5	L7	13707.9481	6.804724	0.0000000
WT infected vs DKO infected	ENSMUSG00000028270	Gbp2	L7	46479.3146	6.746182	0.0000000
WT infected vs DKO infected	ENSMUSG00000028268	Gbp3	L7	8398.9121	6.023623	0.0000000
WT infected vs DKO infected	ENSMUSG00000032690	Oas2	L5	2195.9083	6.492079	0.0000000
WT infected vs DKO infected	ENSMUSG00000045932	Ifit2	L7	4573.5386	4.989932	0.0000000
WT infected vs DKO infected	ENSMUSG00000052776	Oas1a	L5	795.2636	7.524691	0.0000000

Table 2 highlights a subset of representative genes that are significantly differentially expressed across the analysed contrasts. The selected genes include multiple interferon-stimulated and immune-related transcripts, such as Tap1, Stat2, Irf7, and several Gbp family members, which show strong regulation upon infection. In addition, the direct comparison between infected WT and infected DKO samples reveals pronounced genotype-dependent differences for key immune effector genes. These examples illustrate how disruption of type I and type II interferon signalling alters infection-induced transcriptional responses in lung tissue.

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