Data summary and analysis procedure: dual RNA-seq of *Eimeria falciformis* infected *Mus musculus*

Totta Kasemo, Simone Spork,
Christoph Dietrich, Richard Lucius, Emanuel Heitlinger

June 23, 2016

1 Overview

This document contains data summaries and parts of analysis from the dual RNA-seq experiments of *Eimeria falciformis* infected mice carried our by Simone Spork. Basic read-outs such as number of transcripts detected in each sample from mouse and parasite are summarized, as well as compilations of the outcome of statistical tests for differential mRNA abundance from comparisons of different experimental groups. Mouse data was compared (correlated) with previously published (schmid12) microarray data from *E. falciformis* infected mice. Hierarchical clustering illustrated by heatmaps is shown for different comparisons.
The largest mRNA abundance differences on the parasite side are seen in comparisons between different time points post infection. Mouse strain, including Rag1-/- (adaptive immunity compromised mice) and mouse immune status due to previous parasite exposure (second/challenge infection) seems to have less effects on mRNA abundance profiles in the parasite.

2 Methods

2.1 Mice and infection procedure

Three strains of mice were used in our experiments: NMRI (Charles River Laboratories, Sulzfeld, Germany), C57BL/6 (), and Rag1-/- on C57BL/6 background (gift from Susanne Hartmann, FU?). Animal procedures were performed according to the German Animal Protection Laws as directed and approved by the overseeing authority Landesamt fuer Gesundheit und Soziales (Berlin, Germany). Animals where infected as described by Schmid et al., (Schmid12), but

tapwater was used instead of PBS for administration of oocysts. Briefly, NMRI mice were infected two times, which will be referred to as first and second infection. For the first infection, 150 sporulated oocysts were administered in 100 L by oral gavage. During the first infection of 60 mice, all animals were weighed every day. On day zero, before infection, as well as on day three, five and seven days post infection, dpi, caeca from 3-4 sacrificed mice per time point were collected. Epithelial cells were isolated as described in Schmid et al.(schmid12). For challenge infection, mice recovered for four weeks before second infection. Recovery was monitored by weighing and visual inspection of fur. For the second infection, 1500 sporulated oocysts were applied by oral gavage. Three mice were used as non-second infection control, referred to as day 0, second infection.

2.2 Oocyst purification for infection and sequencing

Sporulated oocysts were purified by flotation from feces stored in potassium dichromate and administered orally in 100 uL tapwater. One *E. falciformis* isolate, *E. falciformis* Bayer Haberkorn 1970, was used for all infections and parasite samples. The strain is maintained through passage in NMRI mice in our facilities as described elsewhere (schmid12).

2.3 Sporozoite isolation

Sporozoites were isolated from sporocysts by excystment. For this, sporocysts were incubated at 37C in DMEM containing 0.04% tauroglycocholate (MP Biomedicals) and 0.25% trypsin (Applichem) for 30 min. Sporozoites were purified by the method of Schmatz et al (schmatz–).

2.4 RNA extraction

Total RNA was isolated from infected epithelial cells, sporozoites and sporulated oocysts using Trizol according to the manufacturers protocol (Invitrogen). High quality *what is the meaning of 'high quality' here?* RNA was used to produce an mRNA library using the Illuminas TruSeq RNA Sample Preparation guide. *stolen from genome paper* Sporozoites were stored in 1 mL Trizol until RNA-isolation. Total RNA was isolated using the PureLink RNA Mini Kit (Invitrogen).

2.5 Sequence quality assessment and alignment

Fastq_quality_filter was applied to Illumina Hiseq 2000 sequenced samples. Since this is not easily applicable to pair-end sequencing data, a low threshold was used on the hiseq data. A phred score of 10 was used, i.e., the probability of false base calling is one in ten. We further set q = 60, i.e., nine out of ten bases or more is required to be correct in at least 60% of the bases in each read for the read sequence to be kept for further analysis. This resulted in.....

2.5.1 Alignment and reference genomes

We used the published *Mus musculus* mm10 assembly (Genome Reference Consortium Mouse Build 38, GCA_000001635.2) as reference genome including annotations for mouse data. The *E. falciformis* genome (Heitlinger14) was downloaded from ToxoDB (Gajria07). For the alignment, the mouse and parasite genome files were merged into a dual reference genome, and files including mRNA sequences from both species were aligned against the dual reference genome using TopHat2 (version 2.0.14, Trapnell09)/ Bowtie2 (version 1.1.2, Langmead12). Single-end and pair-end sequence samples were aligned separately with library type 'fr-unstranded' specified for pair-end samples. Import into R was enabled by the R package Ballgown, which requires bam files to be processed by Tablemaker (Frazee15). Tablemaker in turn makes use of Cufflinks (version 2.1.1, Trapnell10).

3 Table of reads per sample

Table 1: Samples are listed from highest to lowest percentace of *E. falciformis* transcripts per total number of transcripts (mouse plus *E. falciformis*). Corresponding number of *E. falciformis* genes are shown. Gray indicates that samples are excluded from analysis (see Methods for details).

Samples	Mouse trancripts	*E. falciformis* transcripts	Percentage *E. falciformis*	*#E.falciformis* genes
NMRI_oocysts_rep1	11676.000	108477484.000	99.989	5734.000
NMRI_oocysts_rep2	19024.000	126543533.000	99.985	5774.000
NMRI_sporozoites_rep1	13800.000	92259539.000	99.985	5808.000
NMRI_sporozoites_rep2	8702.000	21508353.000	99.960	5564.000
NMRI_1stInf_7dpi_rep1	12532238.000	79648900.000	86.405	5894.000
NMRI_1stInf_7dpi_rep2	94310278.000	154343046.000	62.072	5897.000
NMRI_1stInf_5dpi_rep3	334671421.000	54022504.000	13.899	5794.000
NMRI_2ndInf_7dpi_rep1	97221189.000	9927803.000	9.265	5865.000
NMRI_1stInf_5dpi_rep1	204647381.000	18549727.000	8.311	5739.000
C57BL6_1stInf_5dpi_rep2	32887009.000	250954.000	0.757	3946.000
NMRI_2ndInf_5dpi_rep1	589643389.000	3752923.000	0.632	5602.000
NMRI_1stInf_5dpi_rep2	316721928.000	1609009.000	0.505	5439.000
C57BL6_2ndInf_5dpi_rep1	62053975.000	311900.000	0.500	4610.000
NMRI_1stInf_3dpi_rep1	209723287.000	815820.000	0.388	5466.000
C57BL6_1stInf_5dpi_rep1	65523435.000	217606.000	0.331	4259.000
Rag_2ndInf_5dpi_rep1	85288323.000	224273.000	0.262	4251.000
Rag_1stInf_5dpi_rep2	57192380.000	113673.000	0.198	2969.000
NMRI_1stInf_3dpi_rep2	515516785.000	903330.000	0.175	5101.000
Rag_1stInf_5dpi_rep1	71173382.000	73445.000	0.103	2748.000
NMRI_2ndInf_7dpi_rep2	122999109.000	46700.000	0.038	2174.000
NMRI_2ndInf_3dpi_rep2	106446839.000	34699.000	0.033	1901.000
NMRI_1stInf_0dpi_rep1	229165002.000	31459.000	0.014	1380.000
NMRI_2ndInf_5dpi_rep2	113957667.000	12692.000	0.011	539.000
NMRI_2ndInf_3dpi_rep1	91352242.000	5355.000	0.006	121.000
NMRI_2ndInf_0dpi_rep2	90681785.000	3738.000	0.004	62.000
Rag_0dpi_rep1	43414004.000	474.000	0.001	2.000
C57BL6_0dpi_rep1	53877840.000	491.000	0.001	2.000
C57BL6_0dpi_rep2	76753491.000	657.000	0.001	2.000
Rag_0dpi_rep2	80702547.000	730.000	0.001	3.000
NMRI_2ndInf_0dpi_rep1	285032128.000	326.000	0.000	2.000

4 Experimental overview

Table 2: Replicate average of transcripts as order of magnitude for *E. falciformis*/mouse. Columns represent mouse strains used in infection experiments. Rows represent timepoints post infection plus oocyst and sporozoite samples. The upper part of the table shows data for first infection, and oocyst and sporozoite data. The lower part shows data for challenge infection. Averages were calculated after sample exclusions (see Methods). For exact values, see Table 1.

Day, 1st infection	NMRI	C57BL/6	Rag1-/-
0 (control)	10^4 / 10^8	10^2 / 10^8	10^2 / 10^7
3	10^5 / 10^8	NA	NA
5	10^7 / 10^8	10^5 / 10^7	10^5 / 10^7
7	10^8 / 10^7	NA	NA
Oocysts	10^8 / NA	NA	NA
Sporozoites	10^7 / NA	NA	NA
Day, 2nd infection			
0 (control)	10^3 / 10^8	NA	NA
3	10^4 / 10^8	NA	NA
5	10^4 / 10^8	10^5 / 10^7	10^5 / 10^7
7	10^6 / 10^8	NA	NA

5 Transcript coverage distribution and sample exclusions

5.1 Mouse transcript distributions

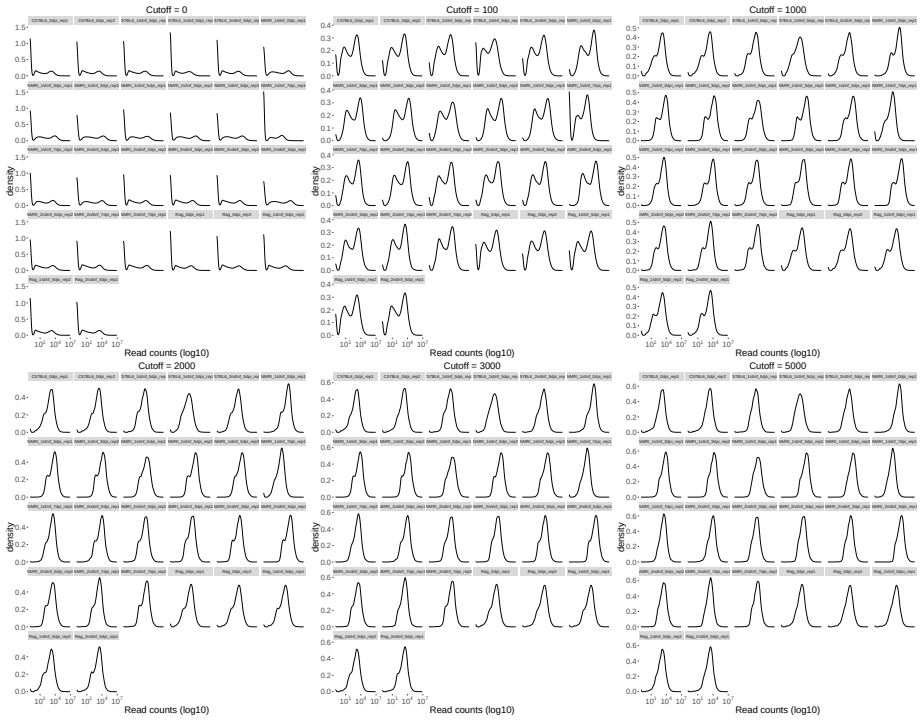

Figure 1: Transcript coverage distribution (density) for mouse samples. Without any requirement (cutoff) for a minimum read coverage per transcript, all samples display a bimodal distribution trend and an additional peak at zero by visual inspection. In our analyses, a cutoff of 3000 was applied. *(2000 enough? No big diff...?).* With this cutoff we can assume a negative binomial coverage distribution as required in the analysis pipline. Non-normalised read counts were used. For this visualization 0.1 was added to all read counts to allow to plot absent transcripts (zero reads) on a log-scale.

5.2 *E. falciformis* transcript distributions

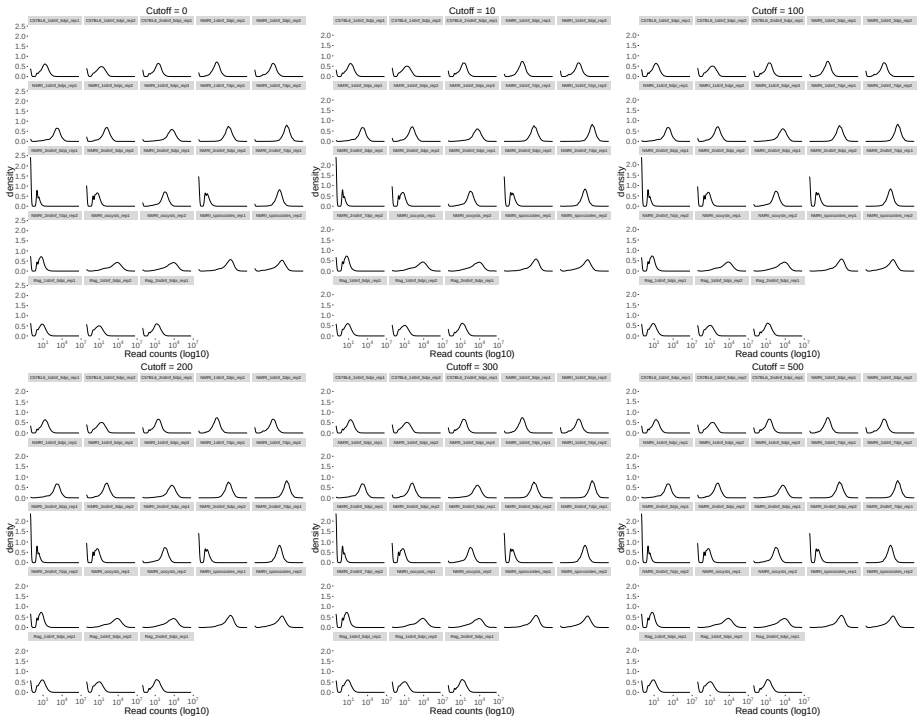

Figure 2: Transcript coverage distribution (density) for *E. falciformis* samples. Without any requirement (cutoff) for a minimum read coverage per transcript, most samples are in agreement with the assumed negative binomial distribution. However, two samples clearly contradict this assumption and a cutoff was applied to evaluate whether peaks in these samples can be removed by excluding transcripts with very low coverage. The cutoff does not have a large effect on distributions of *E. falciformis* data and therefore two samples were removed from further analysis: NMRI_1st_3dpi_rep1 and NMRI_2nd_5dpi_rep1. On all kept samples, a cutoff of 100 was applied. *(but I guess we could also skip it altogether?)*. Non-normalised read counts were used. To all read values, 0.1 was added to allow to plot density of absent transcripts on a log-scale.

6 Mouse RNA-seq data compared with mouse microarray data

Figure 3: Comparison of mouse data from RNA-seq, day 7, (y-axis) and microarray data, day 6 (x-axis, (schmid14)). a. Data normalised with upperquartile method implemented in R package edgeR (version ...). R^2 = yy. b. Normalised data as i a. and adjusted with RUV method as implemented in R package RUVseq (version....). R^2 = xx. Each axis shows log fold changes compared to control.

7 Analysing variance in data by multidimensional scaling

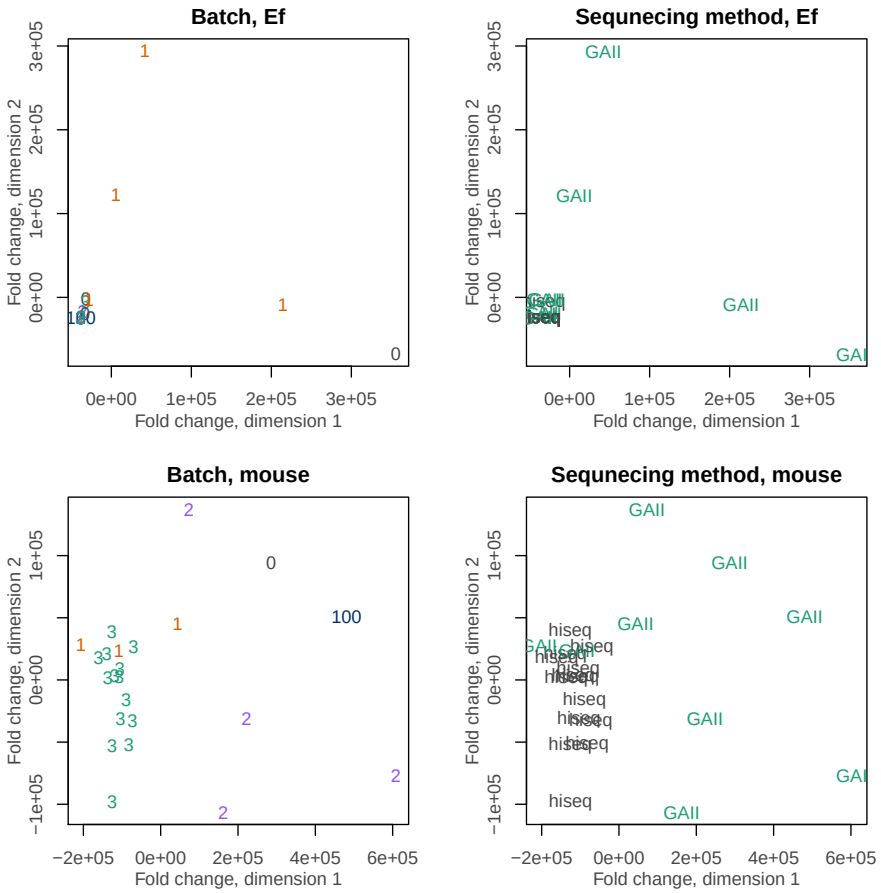

Figure 4: Multidimensional scaling shows no clear sample pattern due to technical variation. Effect on scaling is visualized for batch (left) and sequencing method (right). Upper panel displays *E. falciformis* data and lower panel mouse data. Distances on plot approximate log2 fold changes. Scaling was done with Euclidean distance and "pairwise" gene selection method as implemented in R package limma (version ...).

8 mRNA abundance differences between different experimental groups

Table 3: *E. falciformis* data overview of pairwise comparisons and number of genes with differently abundant mRNAs per comparison. NMRI followed by number indicates day post infection (e.g. NMRI3 = *E. falciformis* genes from NMRI mouse day 3 post infection). Genes with Benjamini-Hochberg corrected p-values ≤0.01 as implemented in edgeR are indluced. NAs are missing samples or not applicable for the species. NA* is due to missing NMRI day 0 sample from first infection.

Day post infection comparisons	*Ef* genes different (FDR≤0.01)	Mouse genes different (FDR≤0.01/0.05)
NMRI 0 vs NMRI 3	NA	274
NMRI 0 vs NMRI 5	NA	1736
NMRI 0 vs NMRI 7	NA	2802
NMRI 3 vs NMRI 5	111	1
NMRI 3 vs NMRI 7	1385	1407
NMRI 5 vs NMRI 7	1895	873
C57BL/6 0 vs C57BL/6 5	NA	914
Rag1-/- 0 vs Rag1-/- 5	NA	45
First and second infection comparisons		
NMRI 3 1st vs NMRI 3 2nd	0	5
NMRI 5 1st vs NMRI 5 2nd	0	1
NMRI 7 1st vs NMRI 7 2nd	0	902
C57BL/6 1st vs C57BL/6 2nd (day 5)	0	mouse
Rag1-/- 1st vs Rag1-/- 2nd (day 5)	0	mouse
Mouse strain comparisons		
NMRI vs C57BL/6	22	NA*
NMRI vs Rag1-/-	0	NA*
C57BL/6 vs Rag1-/-	0	356
Day post infection, parasite relevant comparisons		
Oocysts vs NMRI 3	3310	NA
Oocysts vs NMRI 5	3605	NA
Oocysts vs NMRI 7	3085	NA
Oocysts vs sporozoites	3421	NA
Sporozoites vs NMRI 3	1663	NA
Sporozoites vs NMRI 5	1605	NA
Sporozoites vs NMRI 7	2473	NA

8.1 Differentially abundant mRNAs used for hierarchical clustering

Table 4: A selection of differently abundant mRNAs are used for hierarchical clustering of *E. falciformis* life cycle relevant genes. In each comparison (see Table 3), the 500 differentially abundant mRNAs with lowest FDR are selected. In the next step, the 500 mRNAs from each comparison (or less) are joined (4935 genes) and only unique genes are selected (1618 genes). The 22 genes in the NMRI vs C57BL/6 comparison are not included in the *E. falciformis* life cycle analysis.

Data description	Number of genes
Sum of 1st infection NMRI sample differences (including oocysts and sporozoites)	4935
Used in hierarchical clustering (heatmap)	1618

9 Hierarchical clustering...

All heatmaps shown are now with cutoff 0.05, even though we previously used 0.01 for the life cycle part. I will change it, but the difference seems small and I am really tired... Therefore, I also didn't write the legends for the 1st VS 2nd and mouse strain effect heatmaps. The 1st-2nd looks rediculous for Ef, for instance (it could only compile with cutree=2 or 1). So sorry abot the ugly end.

9.1 Time post infection, *E. falciformis*

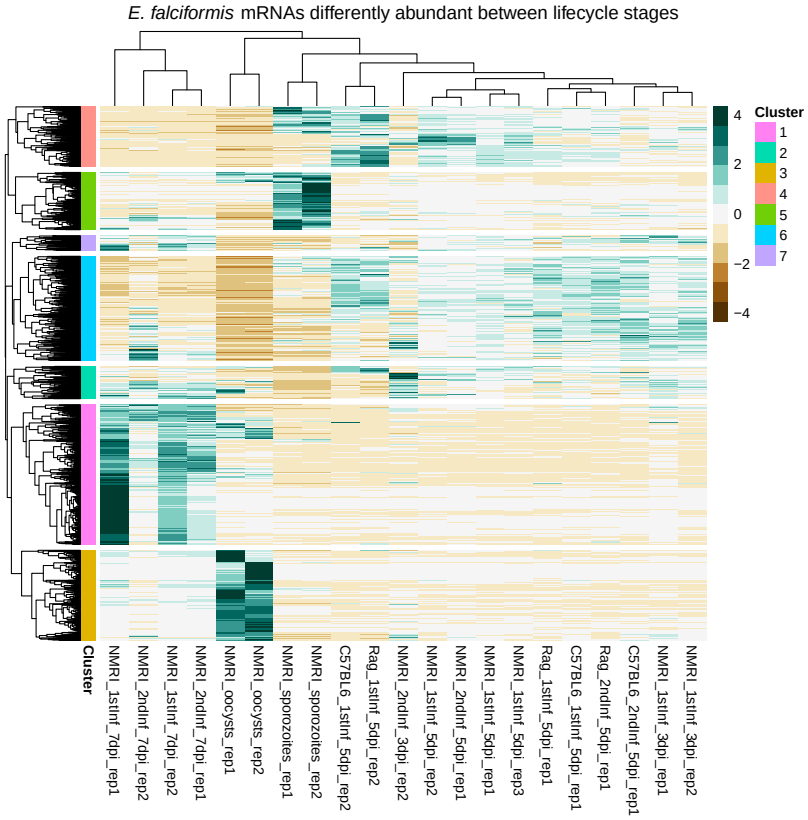

Figure 5: *E. falciformis* mRNAs with significantly different abundance at different times post infection in NMRI mice (see Table 3). All (parasite) samples were included in heatmap. *E. falciformis* samples from seven days p.i. cluster together (NMRI mice only). For day 7, NMRI the pattern is most pronounced in cluster 1 (up). Withing this cluster, the second replicate from the challenge infection has a deviant profile (compare mouse data in Figure 6). Distinct groups of genes also define sporozoites (cluster 6: up) and oocysts (cluster 3: up, cluster 7: down). mRNA profiles on days three and five p.i. from all three mouse strains cluster together. These samples are distinct from oocysts, NMRI day 7 p.i., and sporozoites. The latter cluster closer with days 3 and 5 than with oocysts or day 7. On scale bar, 0 is mean mRNA abundance for each gene (row). Up (green) and down-regulation (brown) numbers denote number of standard deviations from row mean. Hierarchical clustering was performed using with Euclidean distances, method 'complete' (R package limma).

9.2 Time post infection, mouse

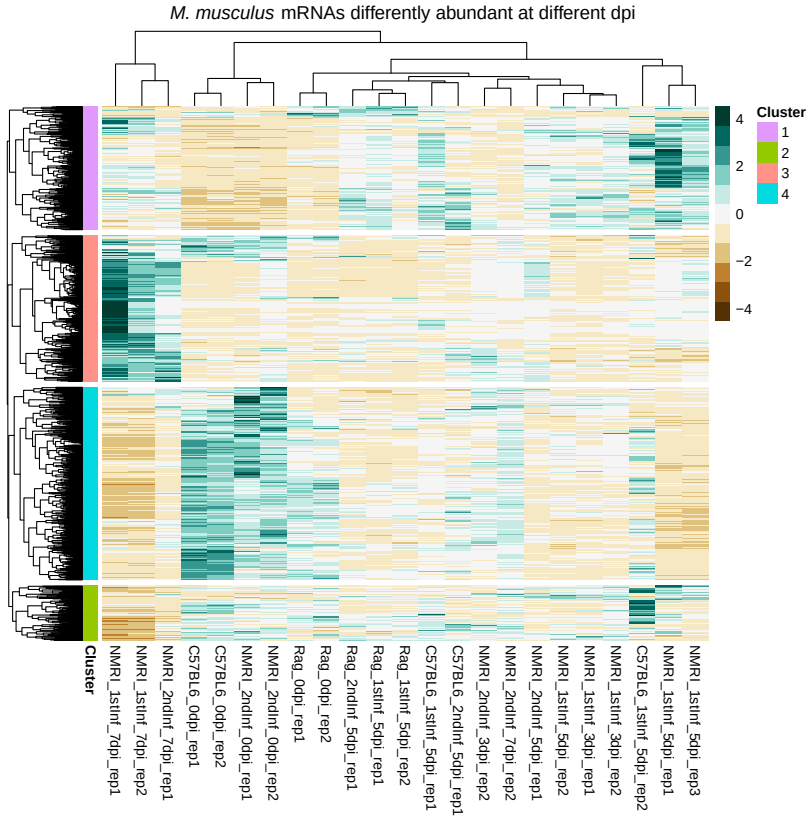

Figure 6: Mouse mRNAs with significantly different abundance at different times post infection in NMRI mice (see Table 3). Three out of four samples from day 7 p.i. (NMRI only) cluster together. These samples are characterized by genes in cluster 2 (up) as well as 3 and 4 (down). The fourth day 7 sample (challenge infection, second replicate) clusters with day 3 and 5 samples but upon visual inspection of cluster 4diplays a profile similar to non-infected mice. The same sample is abnormal in the parasite profile (see Figure 5). NMRI and C57BL/6 uninfected samples cluster together, defined by clusters 3, 4 (up), and 1 (down). Day 3 and 5 samples cluster together, with Rag1-/- forming a separate group and Rag1-/- non-infected most distant in this sample cluster. All non-infected samples share a high mRNA abundance in cluster 3. On scale bar, 0 is mean mRNA abundance for each gene (row). Up (green) and down-regulation (brown) numbers denote number of standard deviations from row mean. Hierarchical clustering was performed using with Euclidean distances, method 'complete' (R package limma).

9.3 First versus second infection, *E. falciformis*

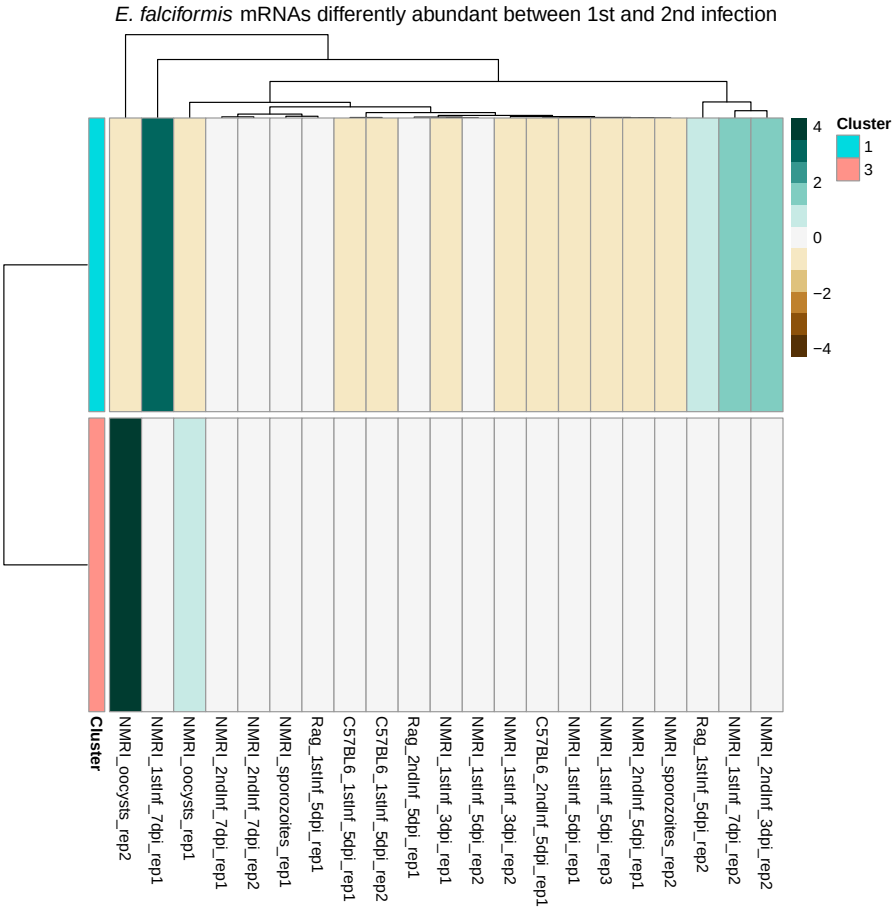

Figure 7: *E. falciformis* with Euclidean distance .....as implemented in R package limma.

9.4 First versus second infection, mouse

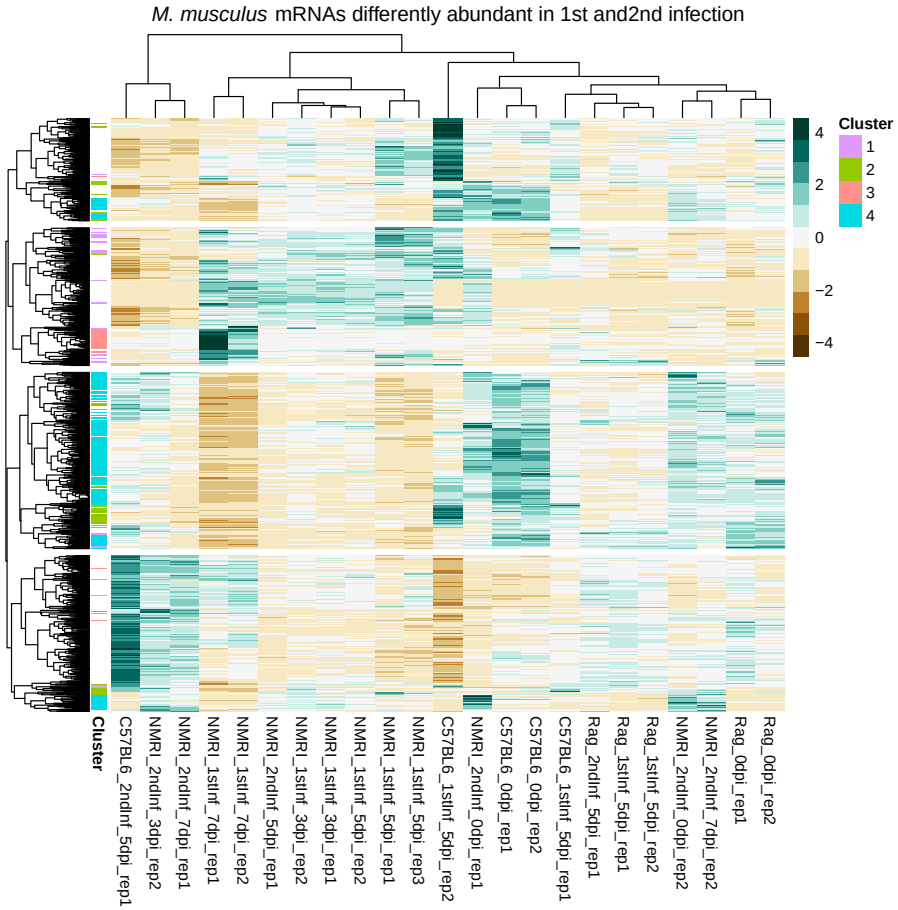

Figure 8: Mouse.... with Euclidean distance .....as implemented in R package limma.

9.5 Differences between mouse strains, *E. falciformis*

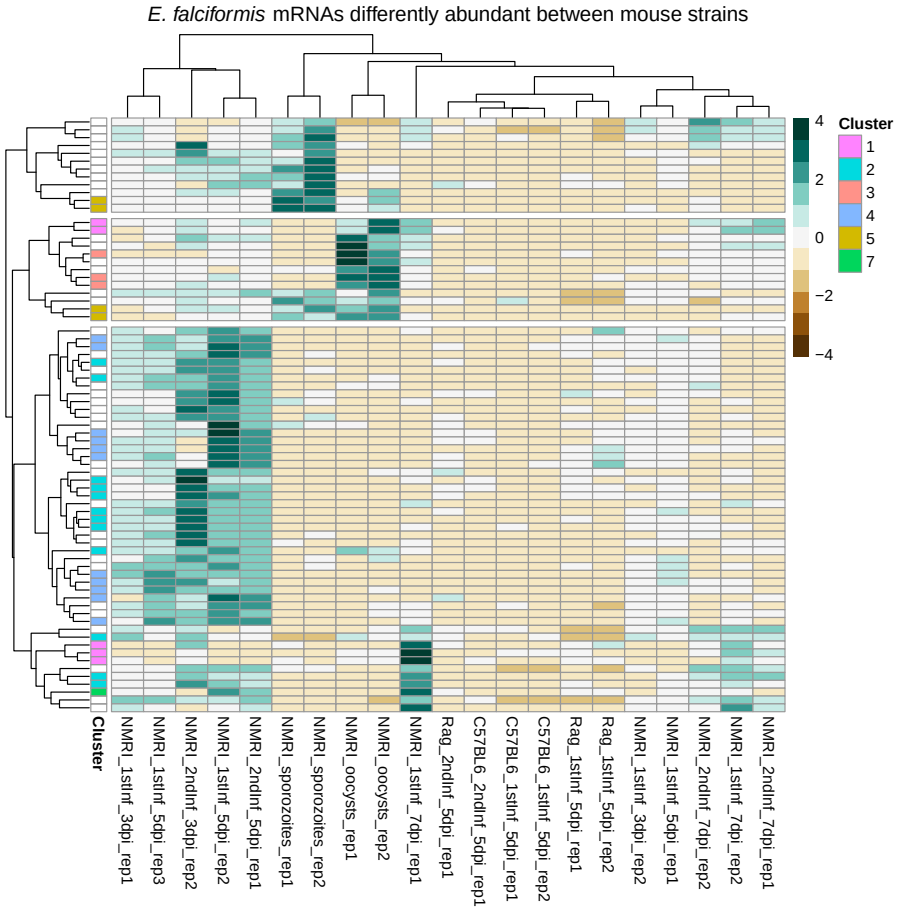

Figure 9: *E. falciformis* with Euclidean distance .....as implemented in R package limma.

9.6 Differences between mouse strains, mouse

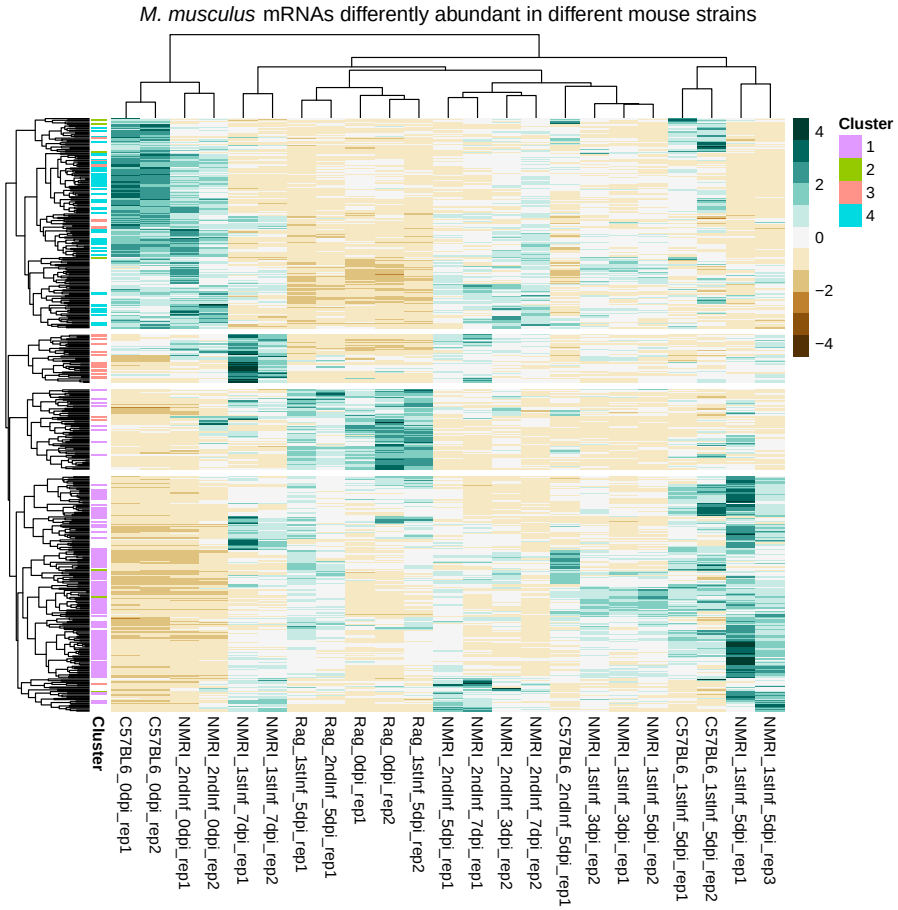

Figure 10: Mouse.... with Euclidean distance .....as implemented in R package limma.