mouse_de

A complete transcriptomics workflow with differential expression analysis for publicly available single-read RNA-seq samples for *Mus musculus.*

Samples

Single read RNA-seq libraries were retrieved from the following public repositories:
https://www.ebi.ac.uk/ena/data/view/PRJEB11951, and
https://www.ebi.ac.uk/ena/data/view/PRJEB11951

Processing analysis

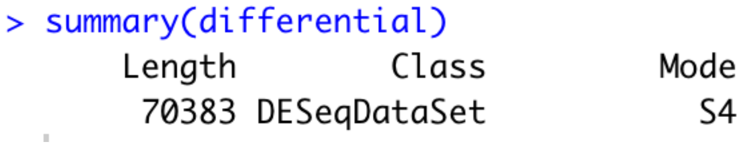
fastq files were assessed using FastQC and multiQC. Four ERR embryonic samples had a mean read length of 125 reads, and between 1.1 to 2.2 million reads. Read duplication rates averaged to around 50%. For the two SRR immune cell samples, approximately 12 million reads were obtained for each sample with a shorter read length of 50 reads, and slightly lower levels of read duplication (averaging 40%). GC content for all samples averaged to 50%. All ERR samples had small (0.5-1.2%) percentages of over-represented sequences, with samples ERR1147326 and ERR1147327 being flagged for sequence duplication levels. All 6 samples passed the MultiQC threshold for basic statistics.

Pseudo-alignment

fastq files were pseudo-aligned to a *Mus musculus* reference using kallisto quant in single read mode. A custom python script was developed to produce different alignment parameters depending on the library metrics generated from FastQC. All ERR samples were aligned using the read length of 125 and an estimated read standard deviation of 20 reads; conversely, SRR samples were aligned with a read length of 50 and a correspondingly smaller standard deviation of 10 reads.

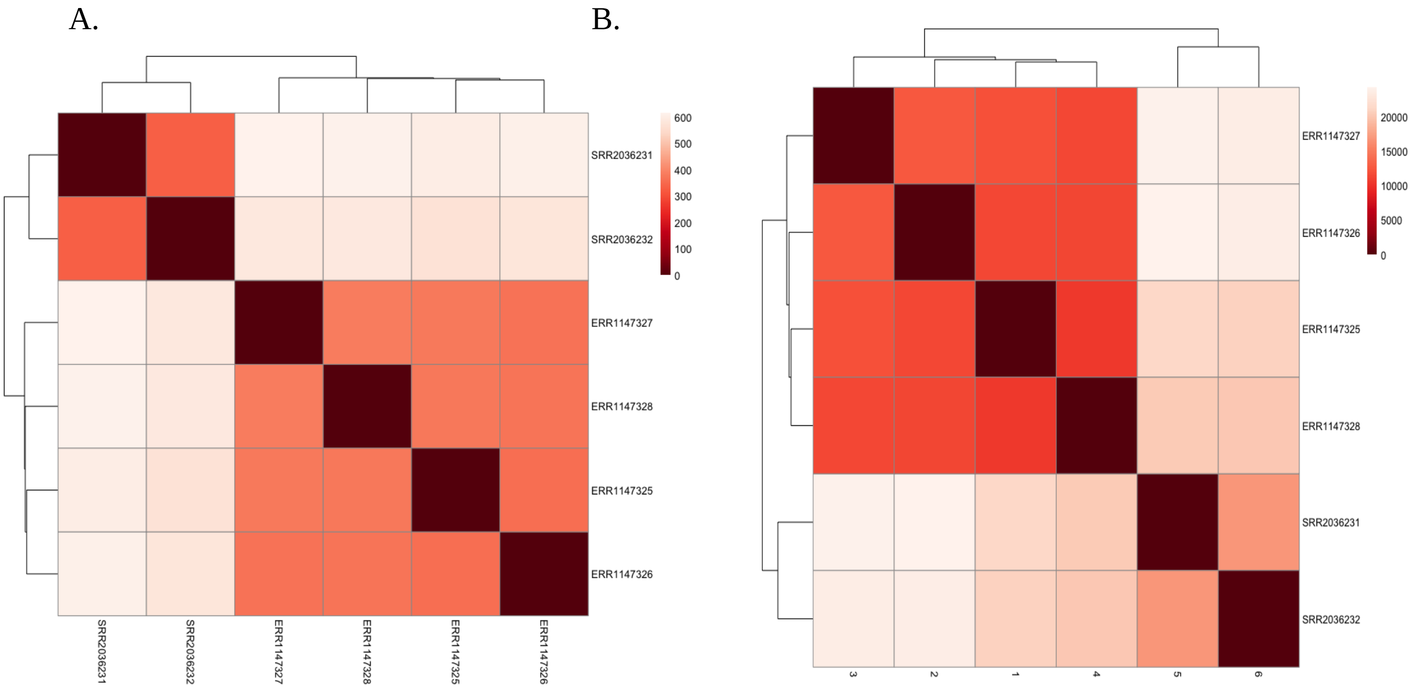
Differential expression analysis using DESeq2

Differential expression analysis was conducted in R using DeSeq2. Abundance files generated from kallisto quant for each library were imported into R using a bulk function from tximport. Differential expression analysis was conducted using embryonic (for ERR) and immune cell (for SRR) conditions as factors. Upon creating a DESEq2 count matrix, rows were filtered to ensure at least 1 read for at least 1 sample would be used for analysis. Of the 118489 gene features used in kallisto count, this filtering step kept 70383 features.



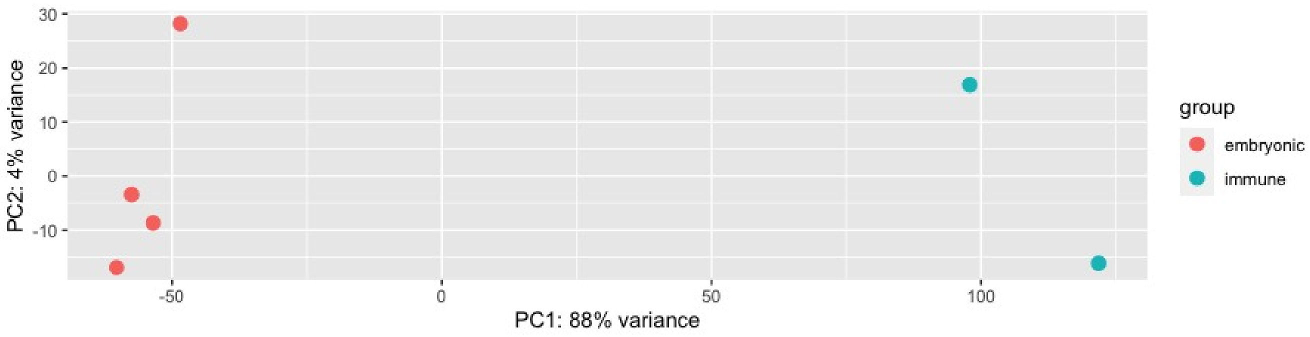
```
> summary(differential)
     Length        Class          Mode
      70383 DESeqDataSet            S4
```

Samples distances (both Euclidean and Poisson) were established for all samples to generate a correlation matrix for similarity across all retained features. The heat maps for these distance measurements can be seen below.



From this pair of heat maps it can be seen that the ERR and SRR samples have greater degrees of similarity to each other than to the other group; furthermore, it is evident that the 4 ERR samples possess a greater degree of similarity among each other than the 2 SRR samples have to each other. This is not surprising since the 2 SRR samples represent within themselves two different treatment options for immune cells, whereas there is no treatment variation across the embryonic ERR samples. Since there is only 1 replicate for each of these treatments within the SRR group, differential expression analysis of this subgroup would not be possible.

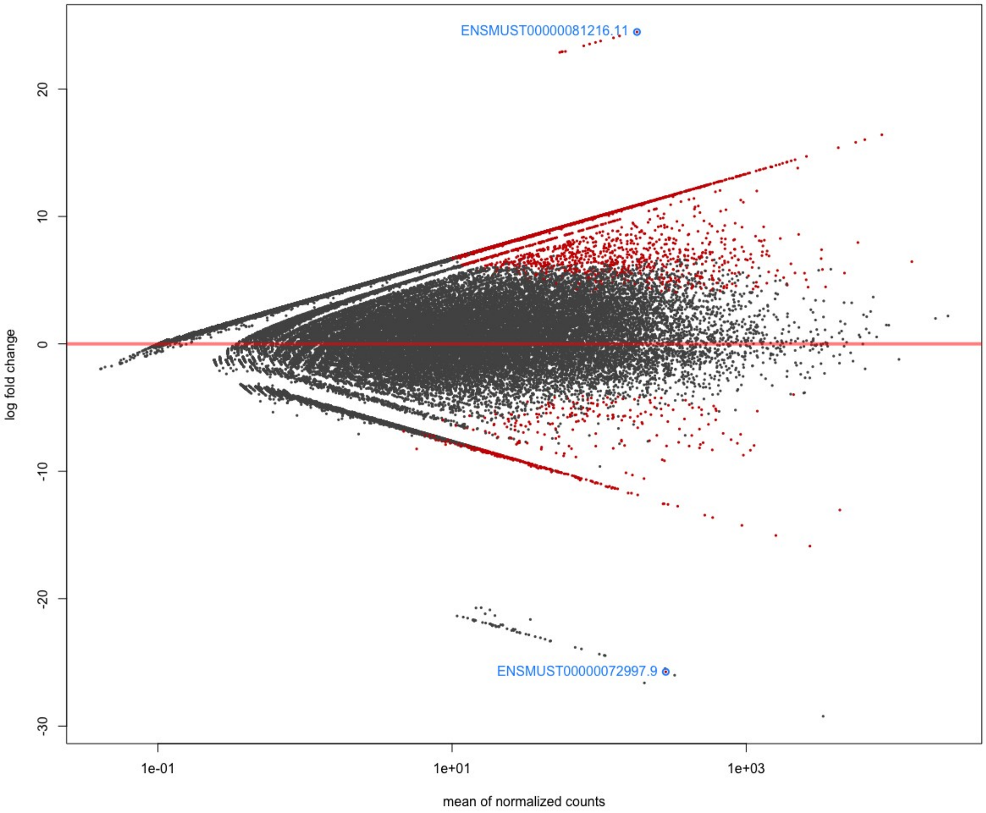
A principal component analysis further confirms the similarity and clustering of the ERR embryonic and SRR immune cells together. The first principal component is observed to represent a large portion (88%) of the variance in gene expression values across the treatments.



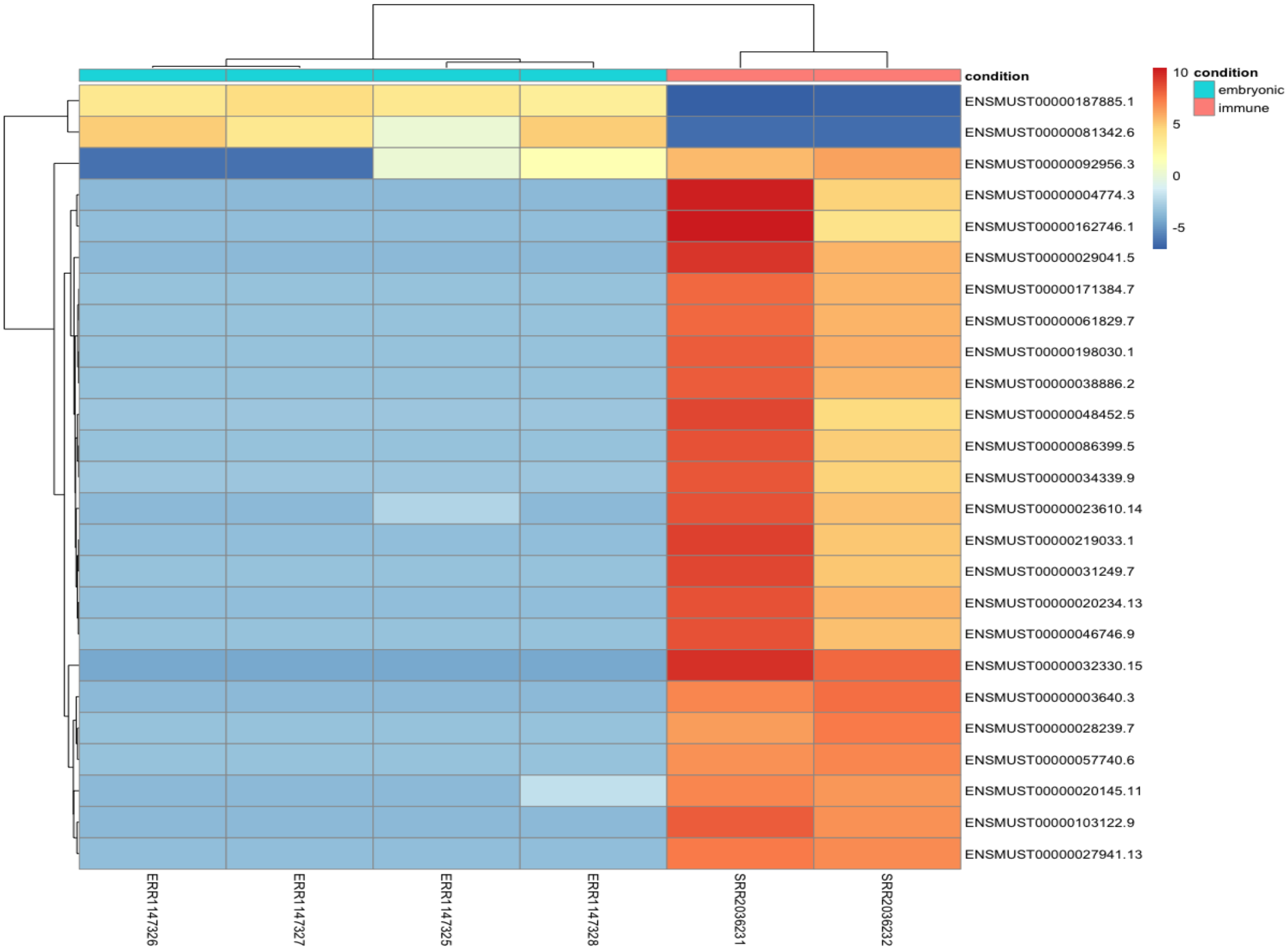
A results object for differential analysis if created, and log fold changes are arranged by the significance of the differential expression as measured by the adjusted p-value (padj). When a typical threshold of 5% is applied as the significance cut-ff, 5128 genes are found to have significant differential expression across treatments.

An MA plot of the average log transformed counts with the average log fold changes between conditions can show the ratio of signal intensity for differential expression of each gene. Points in red are genes with significant differential expression based on padj, and demonstrate that lowly expressed genes or genes without significant fold changes do not represent significant statistical findings. Furthermore, we can see that the measurements for low abundance genes may be created due to Poisson noise.

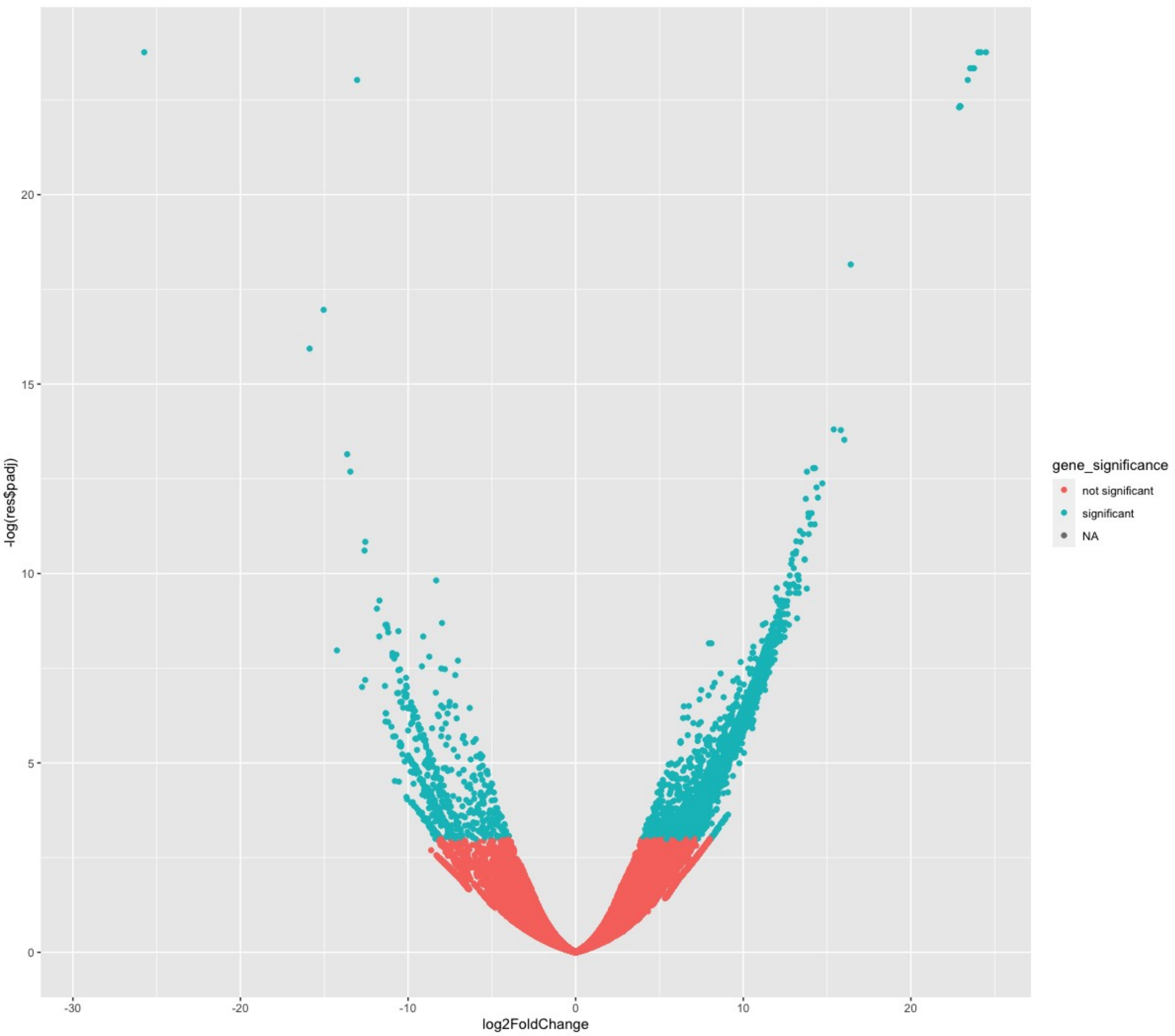
When the 2 genes with the greatest statistical significance (lowest padj) are added to the plot, we can see that these genes are very differentially expressed across the two conditions. They may represent genes of interest to assess the transcriptional differences between the sample groups.



To perform more gene-specific clustering and visualize a subset of the differential analysis, a heat map can be constructed of the top 25 genes that display the most variance across the treatment groups. We can see a distinct subset of genes that are differentially expressed across the embryonic and immune cell libraries, with a consistent level of expression for these genes within the ERR samples, with the exception of ENSMUST00000092956.3. When comparing the SRR samples, we see more variation within this treatment group of these genes, with the exception of both 187885.1 and 81342.6, which show identical expression values. This heat map gives us a subset of genes which to evaluate based on scientific literature and prevailing knowledge of differential expression between these two tissues types.



Finally, a volcano plot of gene log fold changes responses as compared with -log transformed padj shows a typical distribution of values; genes with no visible fold changes responses do not typically have a significant p-value, while the genes that are most differentially expressed often have very significant p-values, corresponding to very large log transformed values. From these analyses, we can export a subset of genes based on large variance values or padj to apply a BioMart or NCBI search to identify gene functions and co-expression patterns. These searches identify genes related to


-log(res$padj)
0
5
10
15
20
-30
-20
-10
0
10
20
log2FoldChange
gene_significance
not significant
significant
NA