Template for RNA-seq bioinformatics support

February 15, 2016

Issue number:	#9999
Request by:	Jan User
Principal Investigator:	Maria Investigator
Organisation:	Stockholm University
BILS staff:	Olga Dethlefsen

Contents

1 Support request

To answer our question we have performed RNA-seq, total RNA ribosomal RNA depleted. We had 4 groups and would like help to finding differentially expressed genees

2 Materials and Methods

2.1 Available data

Data were delivered to Inbox on Uppnex b29999 in fastq format using Illumina 1.8 quality scores. Data were from a paired-end run, with one file for the forward reads and one file for the reverse reads following a naming convention: [LANE_[DATE]_[FLOWCELL]_[SCILIFE NAME]_[READ].fastq.gz.

2.2 Data processing

Raw sequencing reads were processed to obtain counts per genes for each samples. This included:

1. `FastQC/0.11.2` quality check on raw sequencing reads
2. `trimmomatic/0.32` reads filtering for quality score and read length. Reads with average quality below 20 (within 4-base wide sliding window) and/or shorter than 36 bases were removed
3. `star/2.4.1c` was used to align the reads to the reference genome `Mus_musculus.GRCm38.dna.primary_assembly.fa` using the annotation `Mus_musculus.GRCm38.81.gtf`, with reference genome and annotation downloaded from `http://www.ensembl.org/index.html`
4. `featureCounts/1.5.0` from `subread/1.4.5` was used to count the fragments in the exon regions as defined in the `Mus_musculus.GRCm38.81.gtf` file, using default parameters. Specifically, for paired-end reads, a fragment is said to overlap a feature if at least one read base is found to overlap the feature. Fragments overlapping with more than one feature and multi-mapping reads are not counted
5. Counts from multiple lanes were added, if applicable
6. `samtools/0.1.19` were used to sort and index the BAM files containing the aligned reads, e.g. for visualisation in IGV genome browser
7. `MultiQC/0.3.1` was used to aggregate results from `FastQC/0.11.2`, `star/2.4.1c` and `featureCounts/1.5.0` across many samples into a single report

2.3 Differential expression

All analyses were performed under `R`, a programming language and software environment for statistical computing and graphics. Details on the `R` version and packages used can found at the end of this document in R session info

1. `biomaRt` package was used to annotate Ensembl gene identifiers with chromosome name, official gene symbol and description.
2. low count reads were filtered by keeping reads with at least 1 read per million in at least 2 samples
3. `edgeR` package was used to normalise for the RNA composition by finding a set of scaling factors for the library sizes that minimize the log-fold changes between the samples for most genes, using a trimmed mean of M values (TMM) between each pair of samples.
4. the normalized counts were used to examine the samples for outliers and relationships, using Multidimensional Scaling and heatmap based on the Pearson correlation coefficient between every sample pair
5. the normalized counts were used to examine the samples for outliers and relationships, using Multidimensional Scaling and heatmap based on the Pearson correlation coefficient between every sample pair
6. `edgeR` package was to define design matrix based on the experimental design, fitting gene-wise glms model and conducting likelihood ratio tests for the selected group comparisons

2.4 Exon usage

1. `DEXseq` package in `R` was used to infer differential exon usage
2. the provided with the `DEXseq` package `Python` scripts were used to prepare a flattened GTF file based on the `Mus_musculus.GRCm38.81.gtf` and to obtain counts per each exon given the aligned BAM files
3. size factors measuring the relative sequencing depth were estimated to adjust for coverage biases
4. variability of the data was then estimated to be able to distinguish technical and biological variation from real effects on exon usage due to the different conditions. Briefly, per-exon dispersions are calculated using a Cox-Reid adjusted profile likelihood estimation, then a dispersion-mean relation is fitted to this individual dispersion values and finally, the fitted values are taken as a prior in order to shrink the per-exon estimates towards the fitted values
5. having the dispersion estimates and the size factors, differential exon usage was tested. For each gene, `DEXSeq` fits a generalized linear model with the formula $\sim samples + exon + condition : exon$ and compares it to the null model $\sim samples + exon$. The deviances of both fits are compared using a χ^2-distribution, giving rise to a p value, indicative wether the null model is sufficient to explain the data or whether it may be rejected

in favour of the alternative containing an interaction coefficient for condition:exon. The latter means that the fraction of the gene's reads that fall onto the exon under the test differs significantly between the experimental conditions.

6. the obtained p-values were BH adjusted for multiple comparison

2.5 HOMER Motif Analysis

1. `homer/4.7.2` was used to analyze the promoters of genes and look for motifs that are enriched in the target gene promoters relative to other pomoters. The analyses followed the 'Analyzing lists of genes with promoter motif analysis' tutorial
2. briefly, for each comparison, the analyses were run for the differentially expressed genes, separately for down- and up-regulated genes, were differentially expressed genes were defined at 5% FDR and absolute minimum log 2 fold change of 1.
3. the analyses included Gene Ontology enrichment calculations, *de novo* motif analysis and known motif enrichment analysis

3 Work log

A brief project history containing key points

2015-09-15 first meeting with Jan to discuss experimental design, available data and desired results. As first results, Jan would like to receive lists of differentially expressed (DE) genes between the two time points for the two groups

2015-10-09 meeting with Jan to go over the DE results. Agreed that Jan will go over the DE results and try running gene set enrichment analyses using DAVID website.

2015-11-06 I have run and emailed Jan the exon usage results for the 4 comparisons

2015-12-01 I have run and emailed Jan the motif discovery Homer results

2015-12-10 meeting with Jan to discuss the additional results

4 Important practical information

4.1 Data responsibility

Unfortunately, we do not have resources to keep any files associated with the support request. We kindly suggest that you store safely the results delivered by us. In addition, we kindly ask that you remove the files from UPPMAX/UPPNEX b29999. The main storage at UPPNEX is optimized for high-speed and parallell access, which makes it expensive and not the right place for longer time archiving. Please consider others by not taking up the expensive space.

4.2 Acknowledgments

If you are presenting the results in a paper, at a workshop or conference, we kindly ask you to acknowledge us.

BILS staff are encouraged to be co-authors when this is merited in accordance to the ethical recommendations for authorship, e.g. ICMJE recommendations. If applicable, please include Olga Dethlefsen, Bioinformatics Infrastructure for Life Sciences, Science for Life Laboratory, Stockholm University as co-author. In other cases, BILS would be grateful if support by us is acknowledged in publications according to this example: "Support by BILS (Bioinformatics Infrastructure for Life Sciences) is gratefully acknowledged."

Uppmax kindly asks you to acknowledge UPPMAX and SNIC. If applicable, please add: The computations were performed on resources provided by SNIC through Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX) under Project b29999.

5 Results

5.1 Data exploration

Table count, containing counts measured across many genes and samples, is a typical example of a multidimensional dataset, where N objects (samples) were measured on p numeric variables. Hence, to examine the samples for outliers and other relationship one can look at the several multivariate techniques that aim to reduce the dataset dimensions and to reveal the data structure by plotting samples in one or two dimensions. Multidimensional Scaling provides a visual representation of the pattern of proximities among a set of objects, here samples. Another method of visualisation is a heatmap based on the Pearson correlation coefficient, calculated between every sample pair.

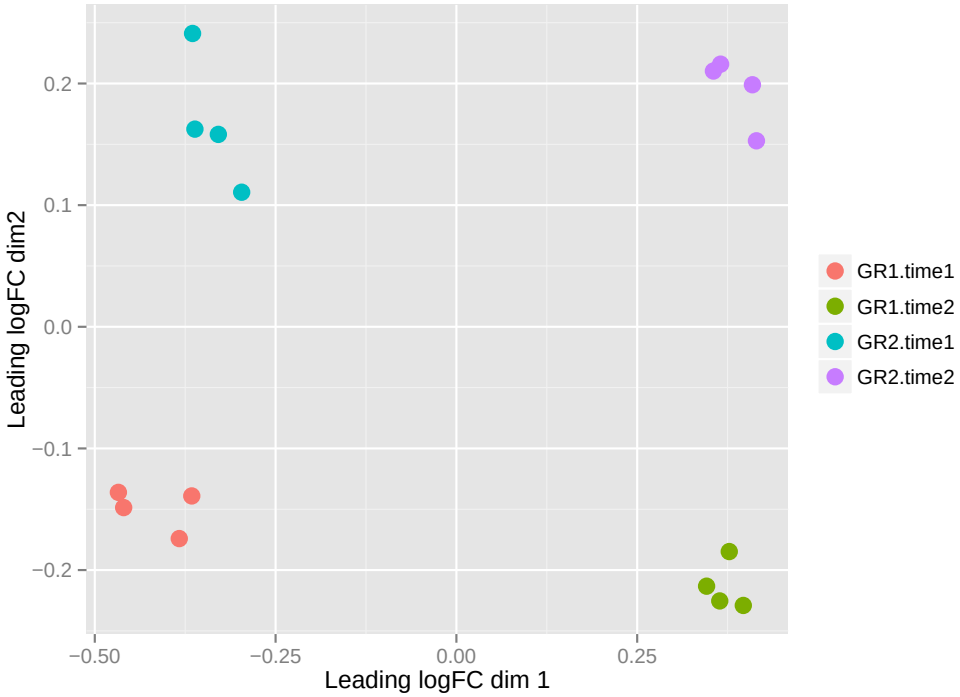

Figure 1: MDS plot on the normalized and filtered counts colour-coded by samples groups

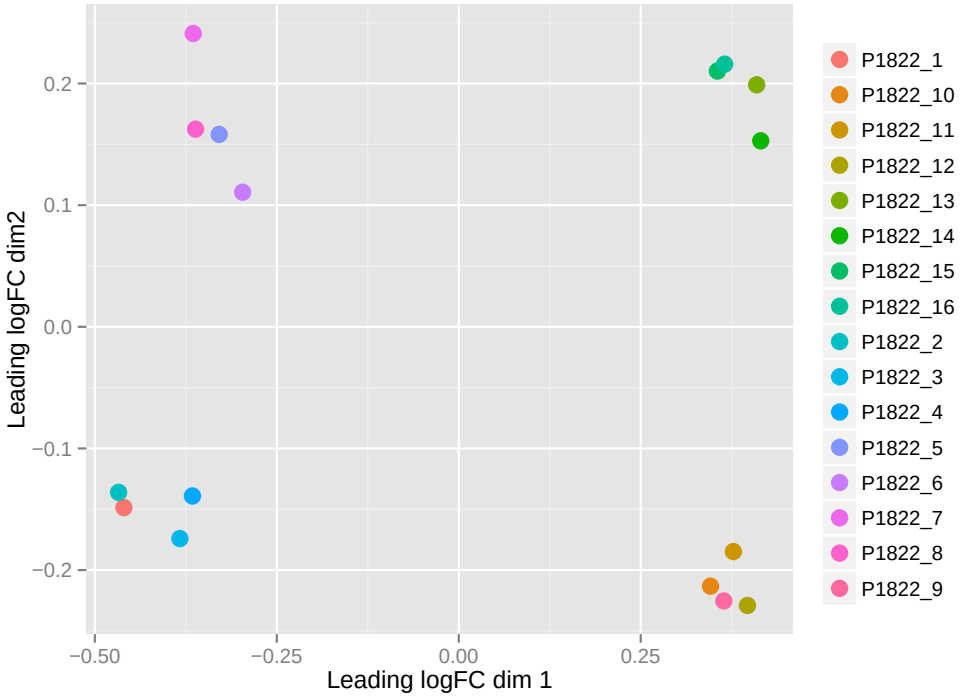

Figure 2: MDS plot on the normalized and filtered counts colour-coded by individual samples

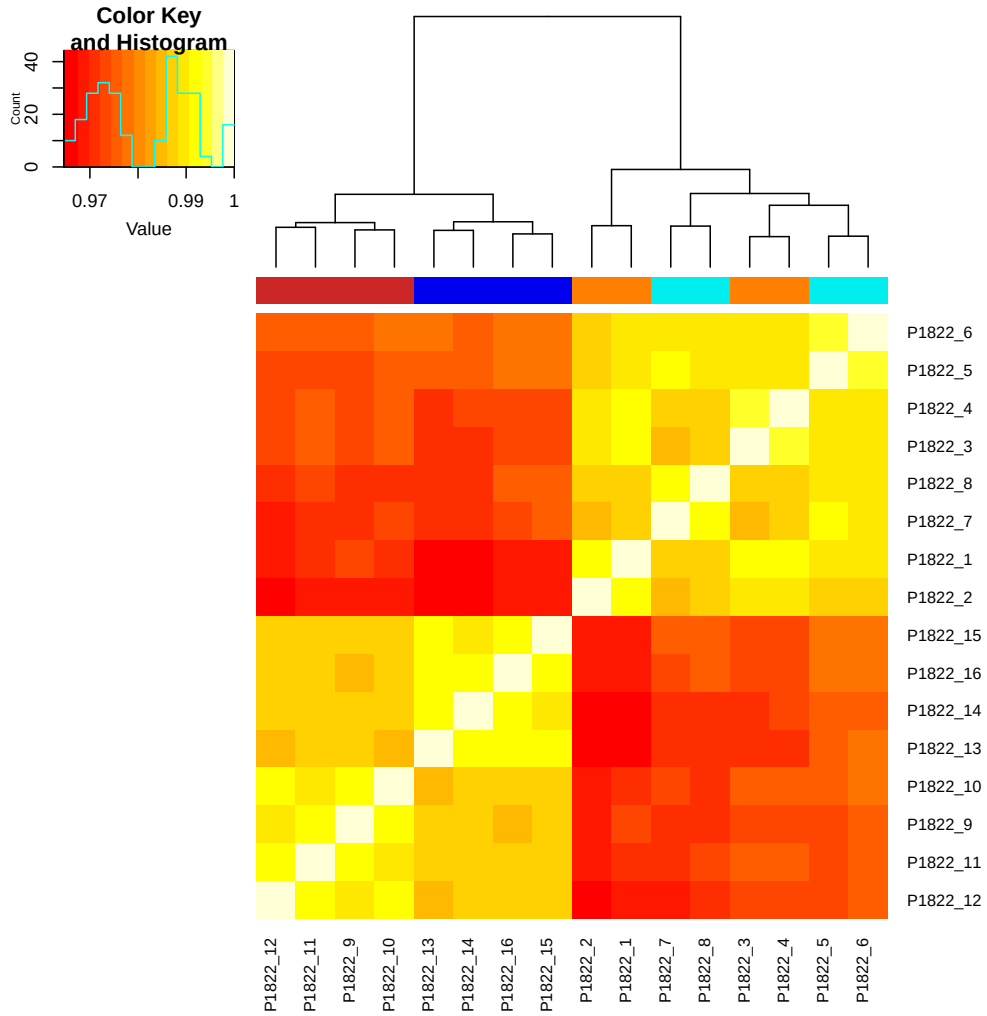

Figure 3: Heatmap based on the pair-wise Pearson correlation coefficient between samples

5.2 Differential expression

5.2.1 Estimating BCVs

Two levels of variation, technical and biological, can be distinguished in any RNA-Seq experiment. Biological coefficient of variation (BCV) is the coefficient of variation with which the (unknown) true abundance of the gene varies between replicate RNA samples. It represents the CV that would remain between biological replicates if sequencing depth could be increased indefinitely. The technical CV decreases as the size of the counts increases. BCV on the other hand does not. BCV is therefore likely to be the dominant source of uncertainty for high-count genes, so reliable estimation of BCV is crucial for realistic assessment of differential expression in RNA-Seq experiments. `edgeR` uses empirical Bayes methods that permit the estimation of gene-specific biological variation, even for experiments with minimal levels of biological replication.

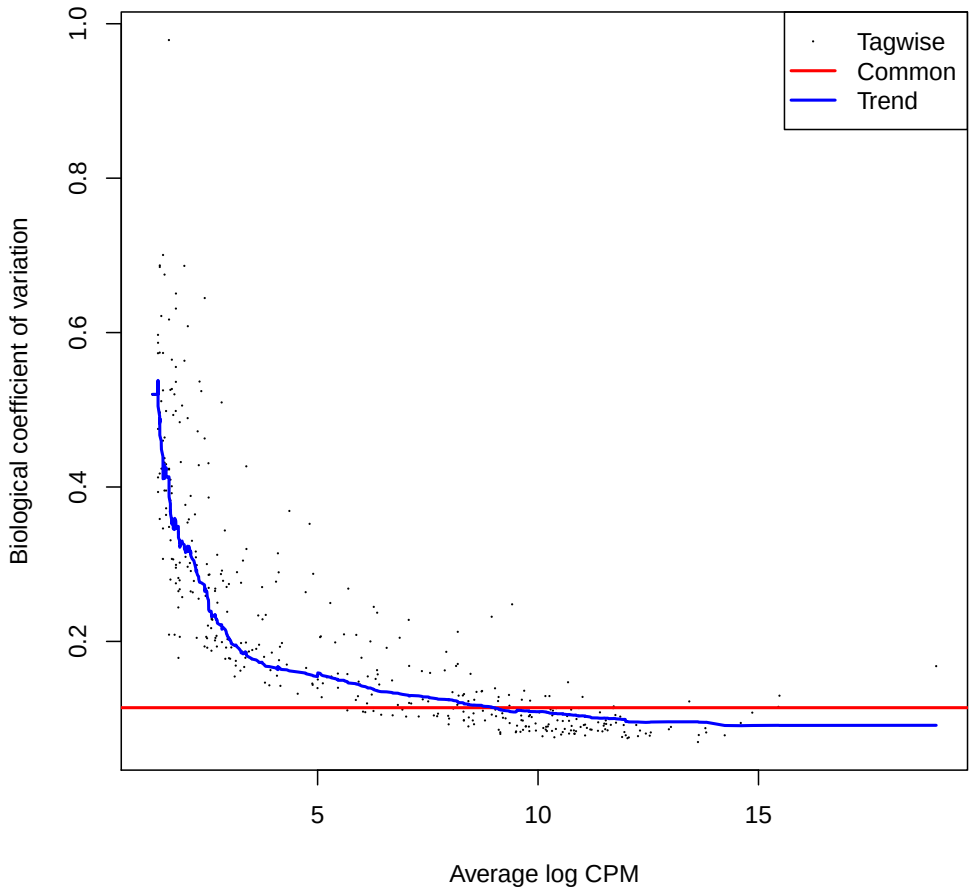

Figure 4: Biological coefficient of variation plot showing the dispersion estimates

5.2.2 GR1: time2 vs. time1

	Down-regulated	Non-significant	Up-regulated
Genes	40	363	36

Table 1: Number of down- and up-regulated differentially expressed genes given 5% FDR and absolute minimum log-fold-change of 1

ensembl_gene_id	mgi_symbol	GR1.logFC	GR1.FDR
ENSMUSG00000056870	Gulp1	2.33	3.33E-109
ENSMUSG00000025911	Adhfe1	2.87	8.75E-82
ENSMUSG00000026064	Ptp4a1	-1.77	9.34E-75
ENSMUSG00000041859	Mcm3	-1.88	1.12E-63
ENSMUSG00000064294	Aox3	4.29	4.36E-53
ENSMUSG00000026023	Cdk15	3.67	6.06E-46
ENSMUSG00000038305	Spats2l	-1.61	6.02E-44
ENSMUSG00000026069	Il1rl1	-1.63	4.74E-40
ENSMUSG00000026070	Il18r1	-2.63	3.72E-37
ENSMUSG00000051951	Xkr4	3.28	1.31E-31

Table 2: 10 genes with the smallest FDR values

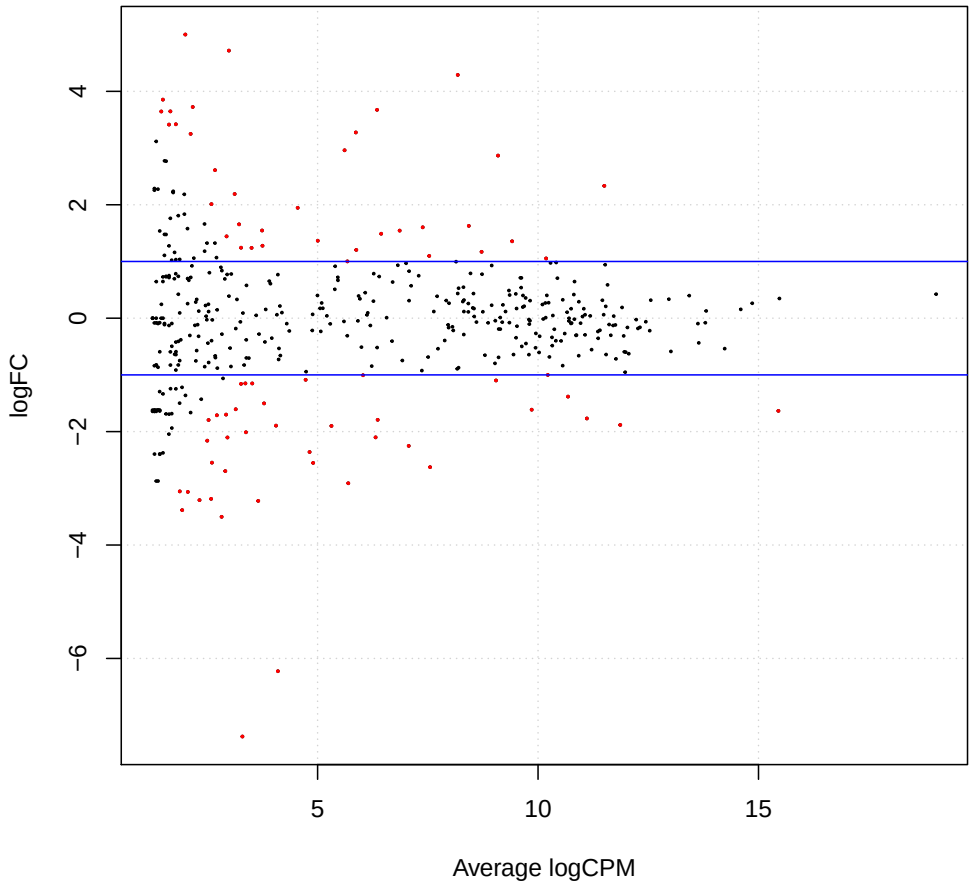

Figure 5: Plot log-fold change against log-counts per million, with DE genes highlighted

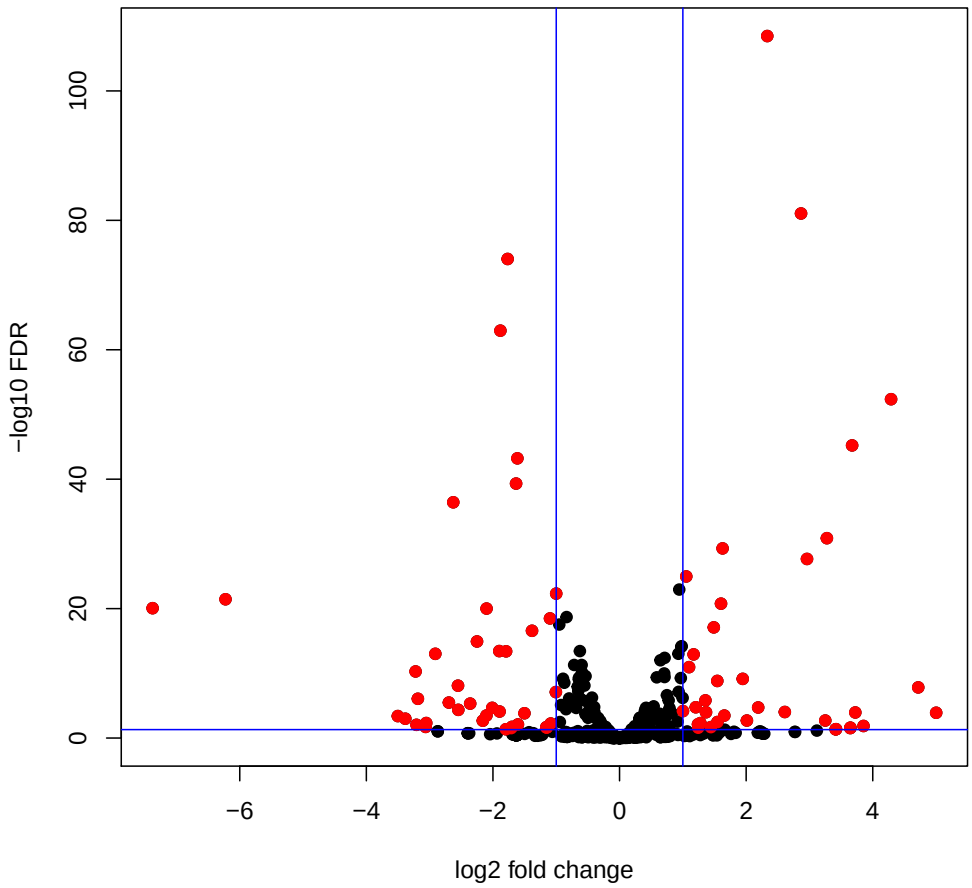

Figure 6: Volcano plot with DE genes highlighted in red. Horizontal blue line corresponds to FDR=0.05 and vertical blue lines correspond to absolute log2 fold change of 1

5.2.3 GR2: time2 vs. time1

	Down-regulated	Non-significant	Up-regulated
Genes	36	368	35

Table 3: Number of down- and up-regulated differentially expressed genes given 5% FDR and absolute minimum log-fold-change of 1

ensembl_gene_id	mgi_symbol	GR2.logFC	GR2.FDR
ENSMUSG00000025993	Slc40a1	3.20	3.50E-128
ENSMUSG00000056870	Gulp1	2.34	3.02E-112
ENSMUSG00000026069	Il1rl1	-2.14	1.05E-65
ENSMUSG00000051951	Xkr4	3.88	2.38E-58
ENSMUSG00000025911	Adhfe1	2.31	4.00E-56
ENSMUSG00000026023	Cdk15	4.07	5.65E-47
ENSMUSG00000064294	Aox3	3.65	3.65E-43
ENSMUSG00000026070	Il18r1	-2.62	9.94E-39
ENSMUSG00000041859	Mcm3	-1.35	5.03E-34
ENSMUSG00000026064	Ptp4a1	-0.90	5.12E-20

Table 4: 10 genes with the smallest FDR values

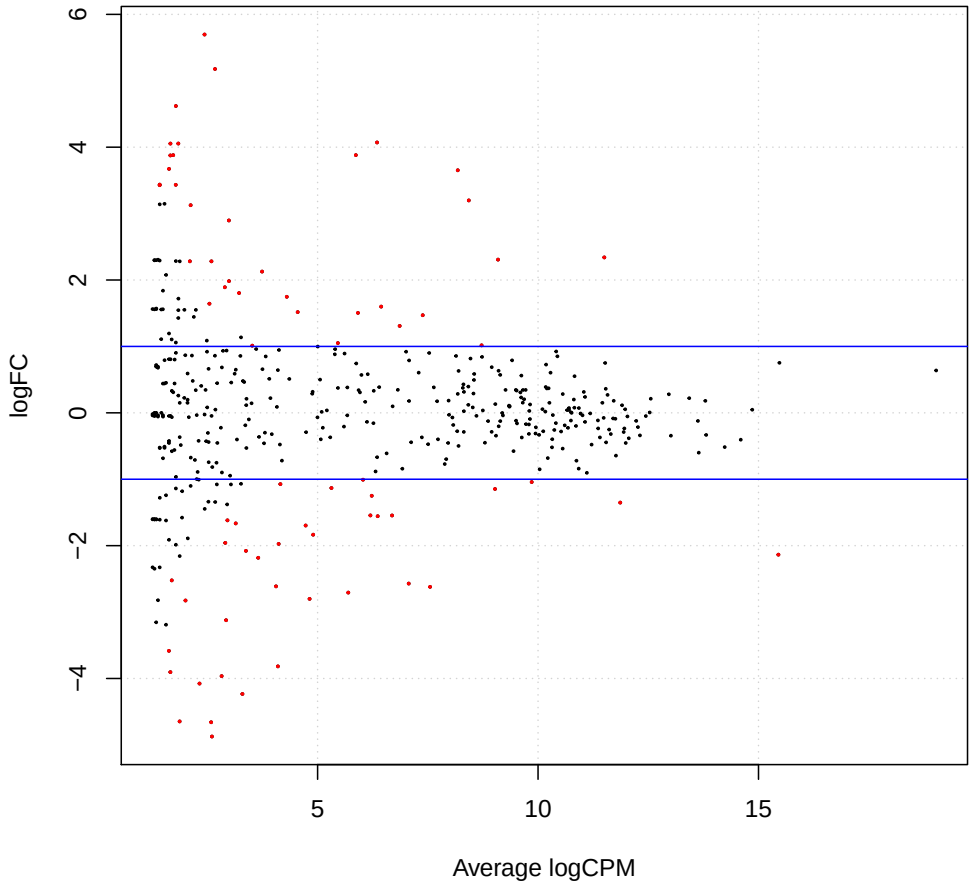

Figure 7: Plot log-fold change against log-counts per million, with DE genes highlighted

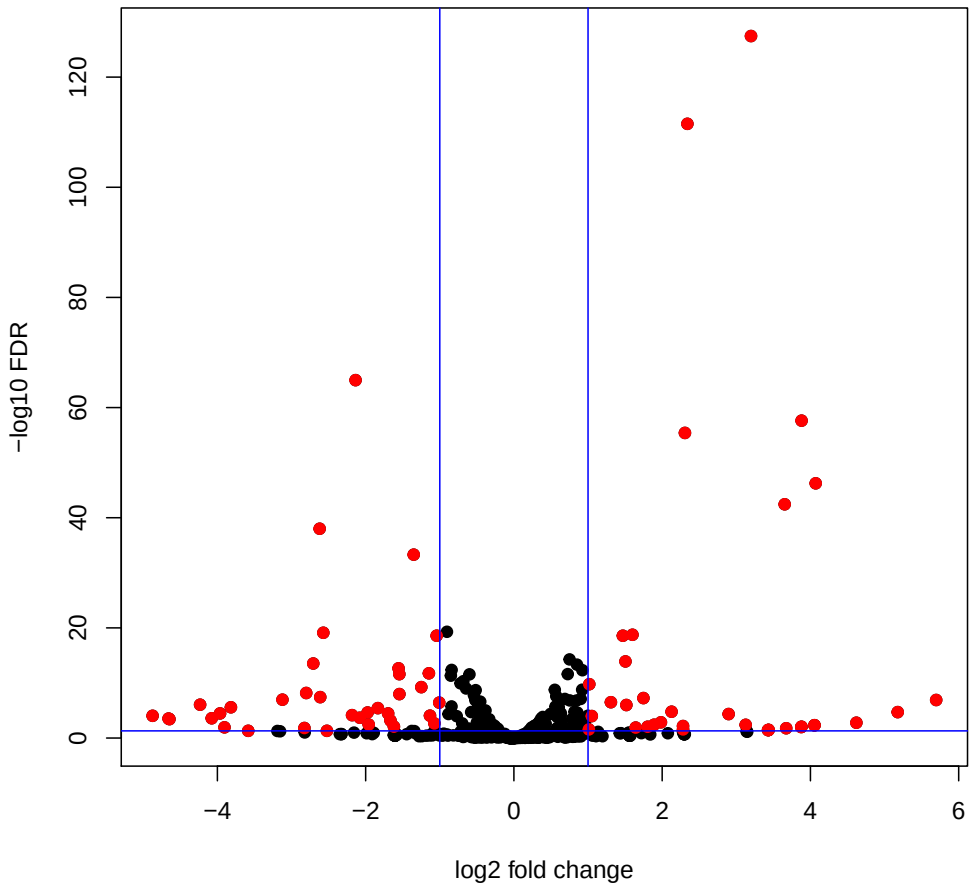

Figure 8: Volcano plot with DE genes highlighted in red. Horizontal blue line corresponds to FDR=0.05 and vertical blue lines correspond to absolute log2 fold change of 1

6 Deliverables

Below is the list of key tab-delimited text files containing the key results from the described data analyses

```
[1] "DE.txt"                          "count_table.txt"
[3] "count_table_annotations.txt" "norm_table.txt"
[5] "norm_table_annotations.txt"
```

where,

DE.txt contains the differential expression results for all the analyses

count_table.txt contains the raw genes counts

count_table_annotations.txt contains the annotations for the genes in the count_table.txt

norm_table.txt contains filtered and normalized genes expression values (TMM)

norm_table_annotations.txt contains the annotations for the norm_table.txt

7 R session info

```
R version 3.2.2 (2015-08-14)
Platform: x86_64-apple-darwin13.4.0 (64-bit)
Running under: OS X 10.10.5 (Yosemite)

locale:
[1] C

attached base packages:
[1] stats4    parallel  stats     graphics  grDevices utils     datasets
[8] methods   base

other attached packages:
 [1] DEXSeq_1.14.2             DESeq2_1.8.2
 [3] RcppArmadillo_0.6.200.2.0 Rcpp_0.12.2
 [5] GenomicRanges_1.20.8      GenomeInfoDb_1.4.3
 [7] IRanges_2.2.9             S4Vectors_0.6.6
 [9] Biobase_2.28.0            BiocGenerics_0.14.0
[11] BiocParallel_1.2.22       xtable_1.8-0
[13] biomaRt_2.24.1            gplots_2.17.0
[15] ggplot2_1.0.1             edgeR_3.10.5
[17] limma_3.24.15

loaded via a namespace (and not attached):
 [1] genefilter_1.50.0    statmod_1.4.22       gtools_3.5.0
 [4] locfit_1.5-9.1       reshape2_1.4.1       splines_3.2.2
 [7] lattice_0.20-33      colorspace_1.2-6     XML_3.98-1.3
[10] survival_2.38-3      foreign_0.8-66       DBI_0.3.1
[13] RColorBrewer_1.1-2   lambda.r_1.1.7       plyr_1.8.3
[16] zlibbioc_1.14.0      stringr_1.0.0        Biostrings_2.36.4
[19] munsell_0.4.2        gtable_0.1.2         futile.logger_1.4.1
[22] hwriter_1.3.2        caTools_1.17.1       labeling_0.3
[25] latticeExtra_0.6-26  geneplotter_1.46.0   AnnotationDbi_1.30.1
[28] proto_0.3-10         acepack_1.3-3.3      KernSmooth_2.23-15
[31] scales_0.3.0         gdata_2.17.0         Hmisc_3.17-0
[34] annotate_1.46.1      XVector_0.8.0        Rsamtools_1.20.5
[37] gridExtra_2.0.0      digest_0.6.8         stringi_1.0-1
[40] grid_3.2.2           tools_3.2.2          bitops_1.0-6
[43] magrittr_1.5         RCurl_1.95-4.7       RSQLite_1.0.0
[46] Formula_1.2-1        cluster_2.0.3        futile.options_1.0.0
```

```
[49] MASS_7.3-45          rpart_4.1-10         nnet_7.3-11
```

8 Where to go next

There is a wide selection of online user-friendly tools available for investigating the interesting genes or list of DE genes. Few recommended below

ClustVist for creating Principal Component Analysis plots and heatmaps

Venny helps to prepare Venn diagram showing relations between a finite collection of different sets, e.g. between list of DE genes from different comparisons

DAVID for a comprehensive set of functional annotation tools for investigators to understand biological meaning behind large list of genes, including identification of enriched Gene Ontology terms, discovering enriched functional-related gene groups, visualizing genes on BioCarta & KEGG pathway and many more

REVIGO for summarizing long list of Gene Ontology terms and visualisation in semantic similarity-based scatterplots, interactive graphs or tag clouds

FunCoup for inferring genome-wide functional couplings or associations, that is an unspecific form of association that encompasses direct physical interaction but also more general types of direct or indirect interaction like regulatory interaction or participation the same process or pathway.