SOUTH DAKOTA STATE UNIVERSITY

RNA-Seq Analysis for provided data on PS 792

PS 792 Team Report

<u>Instructor</u>

Dr. Qin Ma

Next Generation Sequencing Data Analysis

<u>Submitted by</u>

Lok Raj Joshi (Xsede input, Rstudio data analysis, Write up)

Sudeep Ghimire (Xsede input, Rstudio data analysis, Write up)

Surendra Neupane (Rstudio data analysis, Data formatting and iDEP data input, Write up)

Contribution: Equal

Fall semester, 2017

Introduction

RNA-Sequencing (RNA seq) has been a standard tool for studying qualitative and quantitative gene expression assay providing information on transcript abundance with their variation [1]. It is the study of transcriptome which uses next generation sequencing (NGS) technique to elucidate both sequence and relative quantity measures of RNA in a biological sample in a given time [2] and is better than Sanger sequencing and microarray based methods in terms of resolution [3]. It is highly sensitive tool to identify the changes occurring in different disease stages, therapeutics responses and different environmental conditions. This tool allows the detection of genes and their expression status, single nucleotide variants, isoforms and gene fusions which is impossible with DNA level analysis. Even though with all these advantages, RNA sequence quality and quantity, presence of large introns, capturing all sizes of RNA and its easily degradable nature makes it difficult than sequencing DNA. Understanding of transcriptome allows interpretation of functional characteristics of the genome, relating to molecular elements even to the different aspects of health and disease condition [4].

RNA seq is very sensitive and accurate technique which produces cDNA (complementary DNA) from entire RNA molecule forming library formation and performs deep sequencing. It can act as both assessing RNA content and exploring genes in novel transcripts and non-coding RNA. Moreover, RNA seq data has proven to be boon for the differential gene expression analysis, alternative splicing analysis, pathways analysis and co-expression network analysis [5] and does not require prior information. Here, in this report we were presented the count files of the genes after analysis which were used to view the differential gene expression of the two samples in two different time points.

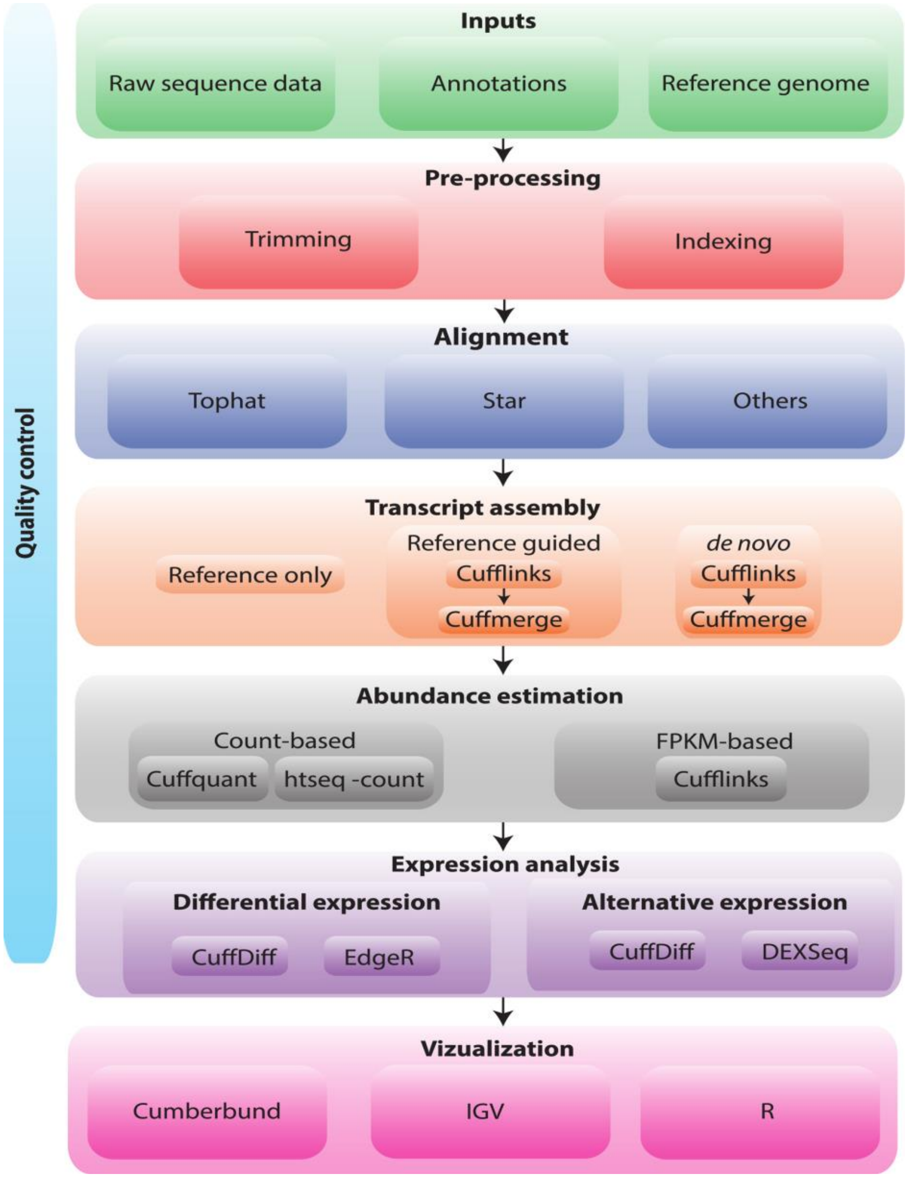


Figure 1: General RNA-seq analysis flow chart [1]

Methodology

RNA Sequencing data were provided in FASTQ format. The quality of the sequences was checked using FastQC tool. After which, poor quality reads were trimmed using BTRIM tool. Then filtered and trimmed sequences were mapped with reference genome using HISAT2 tool. This HISAT2 tools builds a index from a genome and aligns the reads based on the index and gives a .sam file as output. Then, this .sam file was used for final assembly with the help of annotation file. HTseq tool was used for assembly. This tool gives .count file which contains the number of reads for each gene. These count files were analysed using R Studio.

Diferențial gene expression analysis was done using DESeq2 package in RStudio. All the count files were categorized according to strain and time and then converted into a matrix. Finally a .csv files containing all the gene expression data and differential gene expression data were exported.

For visualization of data iDEP (Integrated Differential Expression and Pathway Analysis) web application was used (http://bioinformatics.sdstate.edu/idep/). iDEP can assess number of reads and distribution of RNA Seq data. It can detect differentially expressed genes using DESeq package. Heatmaps, correlation curve, volcano plot and principal component analysis graph were generated using iDEP web application. The parameters used for generating these graphs are provided in Appendix 1.

Graphical representation of the pipeline and tools used for the analysis of RNA Seq data is presented in Figure 1.

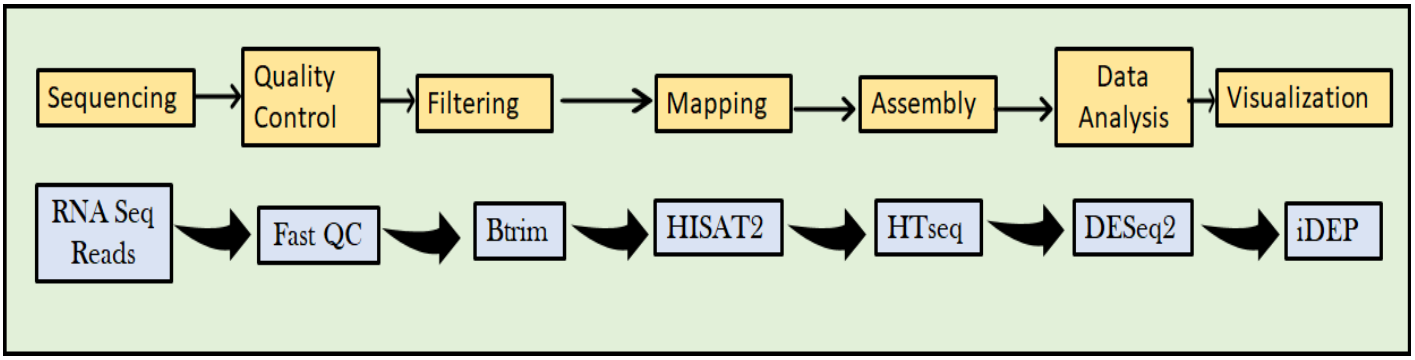


Figure 2: RNA Seq data analysis pipeline.

Results and Discussion

1. Experiment design

There are two strains and the RNA Seq data is available for two time points for each strain. There are two replications for each strain at each time point. All the experimental groups are summarized in Table 1 below.

S.N	**Group**	**Description**
1	S1T1_1	Strain 1, time point 1, first replication
2	S1T1_2	Strain 1, time point 1, second replication
3	S1T2_1	Strain 1, time point 2, first replication
4	S1T2_2	Strain 1, time point 2, second replication
5	S2T1_1	Strain 2, time point 1, first replication
6	S2T1_2	Strain 2, time point 1, second replication

7	S2T2_1	Strain 2, time point 2, first replication
8	S2T2_2	Strain 2, time point 2, second replication

Table 1. Experimental design

2. Data quality

Data quality was visualized using iDEP tool. First of all, number of reads per treatment was calculated. All the treatments has 3-4 million reads as shown in Figure 3. Then data distribution was analyzed to check whether the data is normal or not. As shown in the Figure 3A, we can see the data is not normalized. The correlation between S1T1_1 and S1T1_2 is shown is the Figure 3B, which clearly indicates that the data does not have normal distribution. The data was subjected to log transformation. After log transformation the data became normal as shown in figure 3C. Also, the correlation curve indicated less variability of the data between S1T1_1 and S1T1_2 which is shown in Figure 3D.

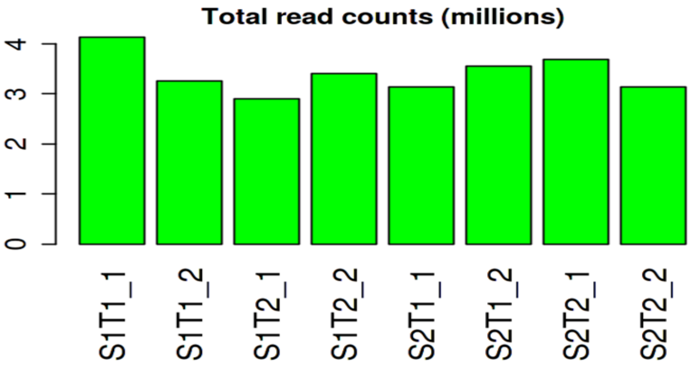


Figure 3: Reads per sample

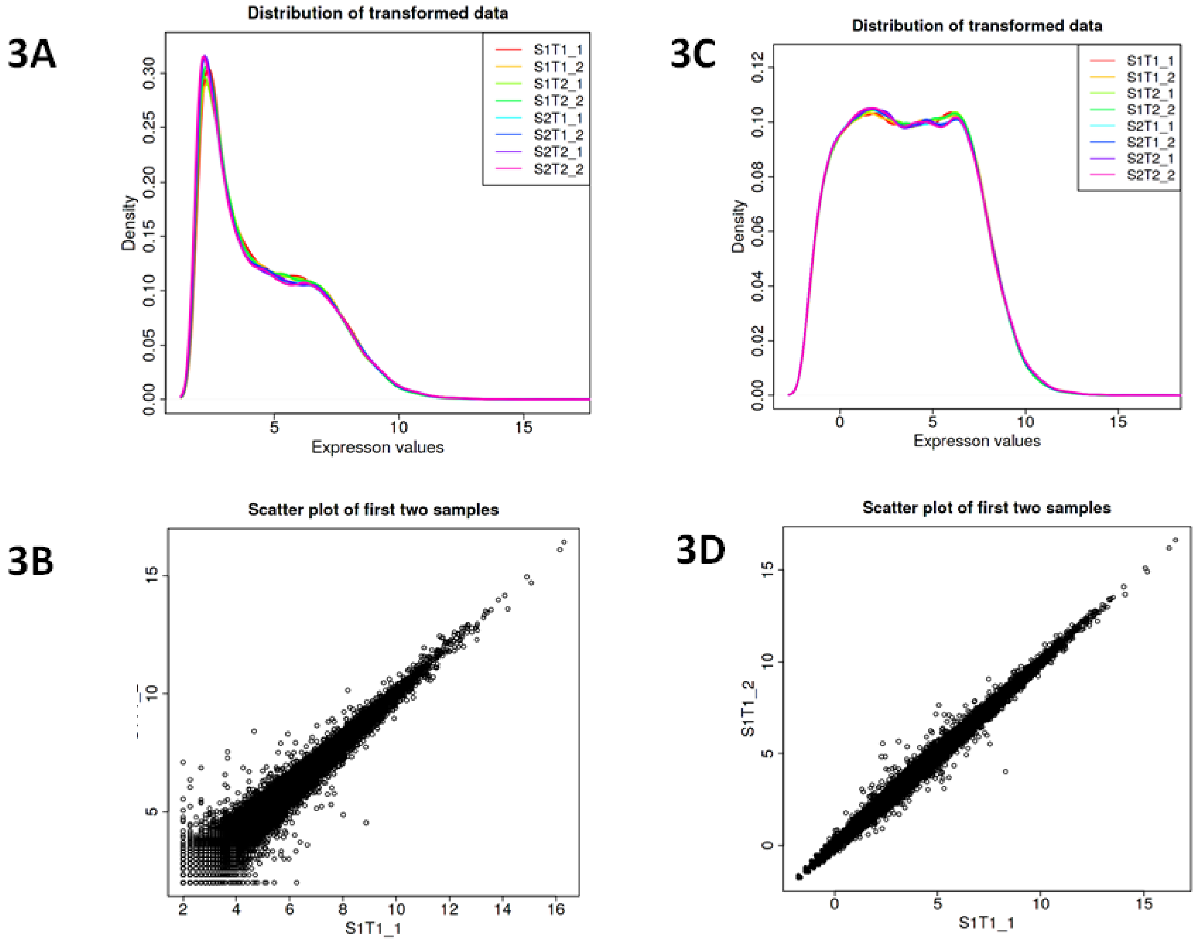


Figure 3. Data normalized by log transformation.

3. Assessment of outliers

We checked if some of the samples were outliers through Principal component analysis (PCA). The PCA analysis did not show any outlier data as the replicated data in different time point were clustered to their respective strains. Also, the correlation matrix shows that the replicates for each time point in respective strains are highly correlated to each other (=1).

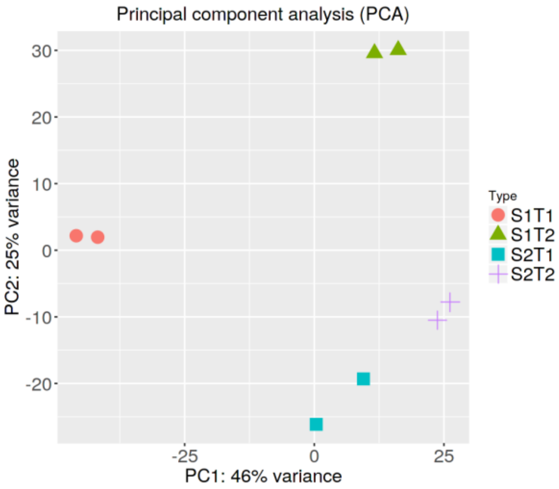
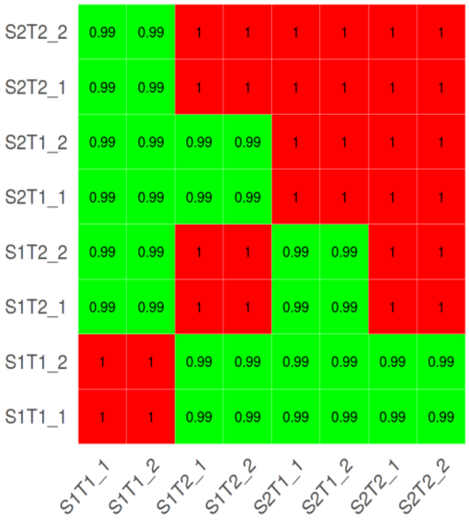


Figure 4: Principal Component Analysis



	S1T1_1	S1T1_2	S1T2_1	S1T2_2	S2T1_1	S2T1_2	S2T2_1	S2T2_2
S2T2_2	0.99	0.99	1	1	1	1	1	1
S2T2_1	0.99	0.99	1	1	1	1	1	1
S2T1_2	0.99	0.99	0.99	0.99	1	1	1	1
S2T1_1	0.99	0.99	0.99	0.99	1	1	1	1
S1T2_2	0.99	0.99	1	1	0.99	0.99	1	1
S1T2_1	0.99	0.99	1	1	0.99	0.99	1	1
S1T1_2	1	1	0.99	0.99	0.99	0.99	0.99	0.99
S1T1_1	1	1	0.99	0.99	0.99	0.99	0.99	0.99

Figure 5: Correlation between strains and time points

4. Heat-maps of transformed reads:

After quality and outlier check, we observed general trend of gene expression in different strains at different time points via plotting the heat-maps. We divided the genes with transformed reads into four different clusters. Cluster A consisted of 2184 genes, B consisted 2766 genes, cluster C consisted 186 genes and D consisted 864 genes. The cluster C consisted of genes with highly variable reads (either very high or very low amount of reads) in all treatment conditions. The moderately expressed genes were clustered each other to form cluster D (864 genes). The cluster B contains both high and low expressed genes (2766 genes). This clustering pattern also shows uniformity in replicates among the treatments. Heat maps of most varied 60 genes in all conditions

Allowed us to see the general expression pattern of these genes in different strains in different time points. For example, the gene, MDP0000007682GDRv10196 which is highly expressed in S1T1 is significantly decreased in S2T1 and S2T2 conditions.

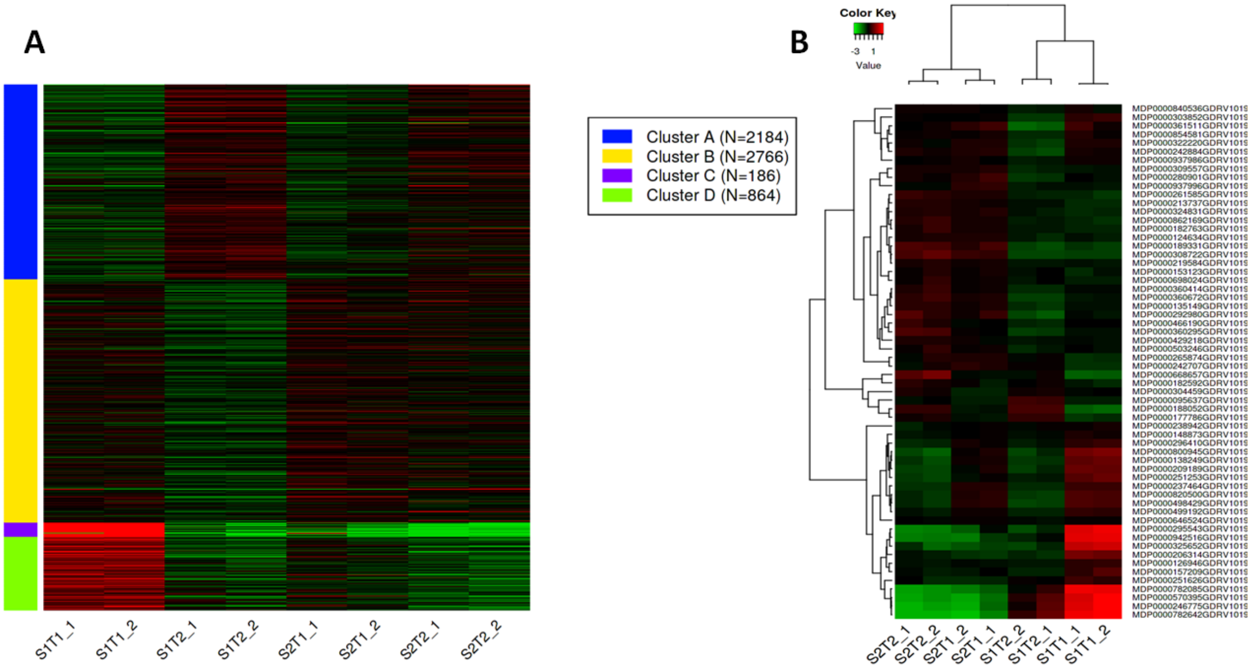


Fig 6. Heat maps A. The k-means clustering of all genes involved in gene expression in different strains at different time points B. Heat maps of most varied 60 genes in all conditions.

5. Differences in gene expression numbers in different groups.

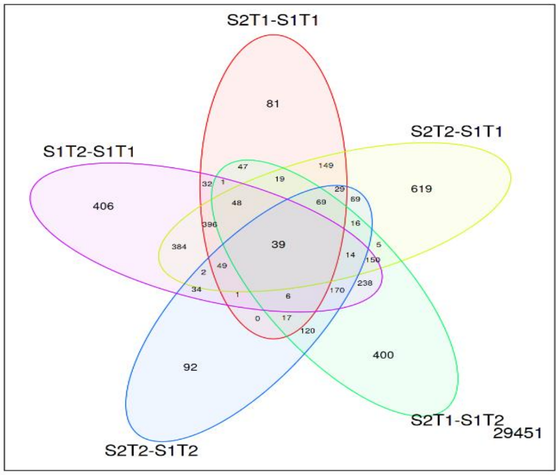


Figure 7: Venn diagram representing total number of differentially expressed genes compared to different strains at different times. Overlapping regions represent co-expressed genes.

Total number of genes expressed differentially in each condition revealed that high number of genes were differentially expressed in S2T2-S1T1 (619), S2T1-S1T2 (400) and S1T2-S1T1 (406) whereas 39 genes were co-expressed in all the conditions. This pattern of specific high gene expression in one condition could attribute to related biological phenomenon.

6. Differential expression patterns of genes in different strains and times.

Strain 1 has high expression of 1310 genes at time 1 (T1) compared to 660 genes at time 2 (T2) (fig a). Similar pattern of differences in the number of expressed genes are seen in fig b,c,d,e, and f. Thus, differential expression of genes varies with time and strain.

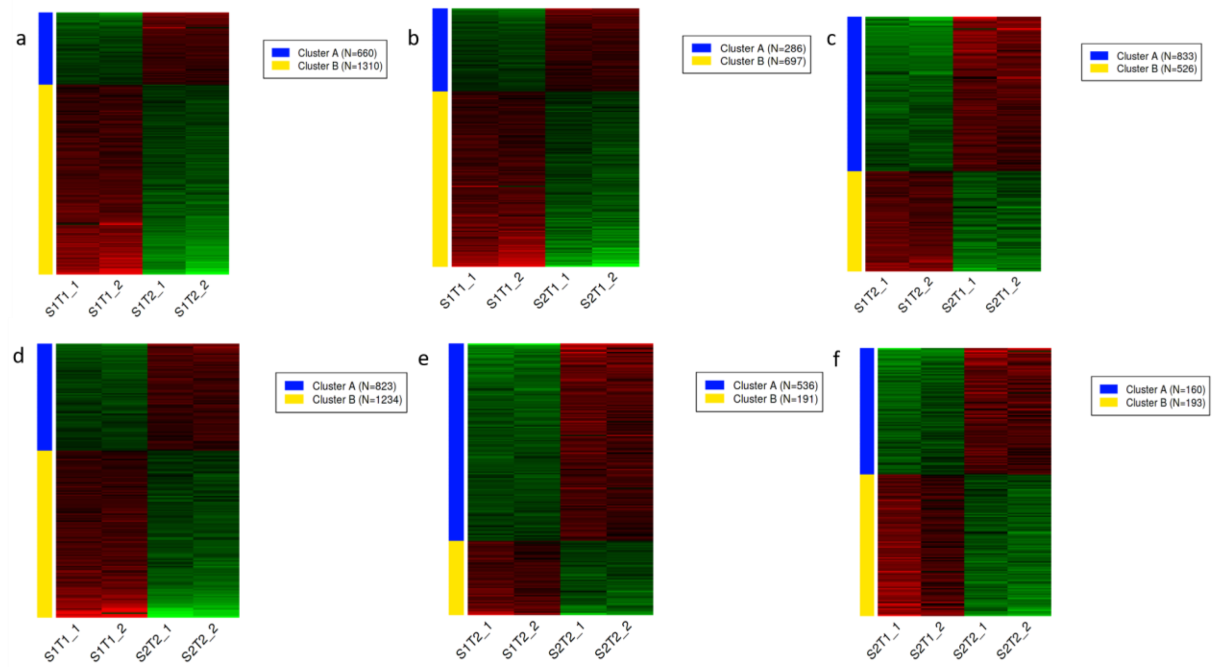


Fig 8: Heatmaps depicting expression of genes on the basis of similarity in gene expression clustered together in two strains (S1 and S2) and two time points (T1 and T2) with one replicate each. FDR cut off and minimum fold change were set at 0.05 and 2 respectively.

Corresponding with Fig 8, volcano plots A,B,C,D,E and F show the difference in gene expression patterns in relation to different strains and time points. Upregulation of genes in one combination compared to another could potentially relate to biological condition produced over time or because of difference of the strains.

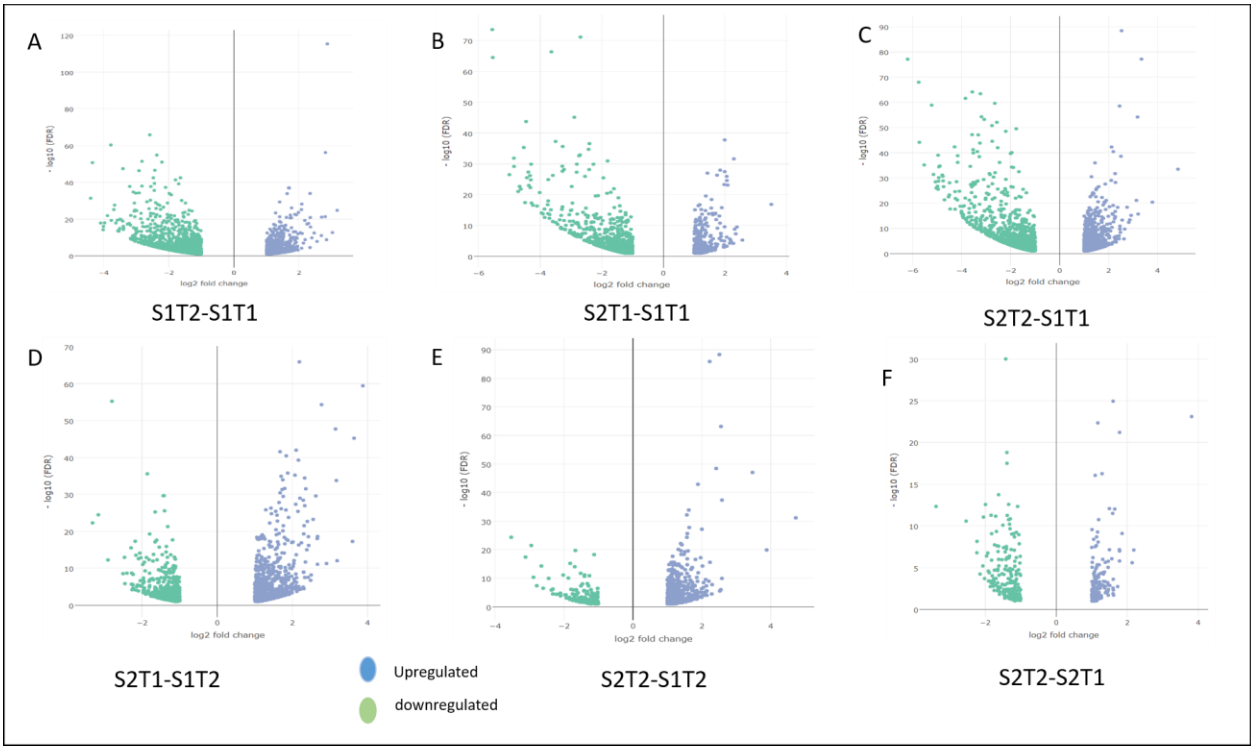


Fig 9: Volcano plots showing expression of differential genes for each combination of strain and time. FDR values were set at 0.05 and log fold change of less than 2 were discarded.

Conclusion

RNA Seq has been widely used tool for quantitative and qualitative measure of gene expression. So, it is valuable to have capability to analyze RNA Seq data. From this assignment we were able to run various programs for RNA Seq data analysis on cluster. Moreover, we became familiar with various R packages like DESeq2, edgeR for RNA for differential gene expression analysis. We were able to generate various graphs and plots that provided very clear understanding of this complex data. Here we don’t have history of the experiment to pinpoint the differences precisely, however we can see a number of genes are being upregulated while some of them are downregulated with respect to strain and time.

References

1. Griffith, M., et al., *Informatics for RNA Sequencing: A Web Resource for Analysis on the Cloud.* Plos Computational Biology, 2015. **11**(8).

2. Conesa, A., et al., *A survey of best practices for RNA-seq data analysis.* Genome Biol, 2016. **17**: p. 13.

3. Nagalakshmi, U., K. Waern, and M. Snyder, *RNA-Seq: a method for comprehensive transcriptome analysis.* Curr Protoc Mol Biol, 2010. **Chapter 4**: p. Unit 4 11 1-13.

4. Byron, S.A., et al., *Translating RNA sequencing into clinical diagnostics: opportunities and challenges.* Nat Rev Genet, 2016. **17**(5): p. 257-71.

5. Han, Y., et al., *Advanced Applications of RNA Sequencing and Challenges.* Bioinform Biol Insights, 2015. **9**(Suppl 1): p. 29-46.

Appendix

Analysis Details

Analysis were conducted using the iDEP 0.39, hosted at http://ge-lab.org on Sun Oct 15 00:49:15 2017.
Data
Species:
Number of samples: 8
Number of genes converted and filtered: 33153
4 sample groups detected.
Input file type: RNA-seq read count file

Pre-processing and exploratory data analysis settings:
Min. counts: minCounts= 2
Counts data transformation method: Started log: log2(x+c)
Pseudo count: c= 4
Method for differential expression: CountsDEGMethod= 3 (DESeq2)
number of genes in heatmap: nGenes= 60
number of genes in k-means clustering: nGenesKNN= 6000
number of clusters in k-means clustering: nClusters= 4
Promoter analysis for k-means clustering: radioPromoterKmeans= 300 bp

Differential expression settings:
FDR cutoff: limmaPval= 0.05
Fold-change cutoff: limmaFC= 2
Promoter analysis for DEGs: radio.promoter= 300 bp

Pathway analysis settings:
Pathway analysis methods: pathwayMethod= GAGE
FDR cutoff: pathwayPvalCutoff= 0.1
Min size for gene set: minSetSize= 15
Max size for gene set: maxSetSize= 2000

PREDA settings:
FDR cutoff: RegionsPvalCutoff= 0.01
FDR cutoff: StatisticCutoff= 0.5