WORLD JOURNAL OF PHARMACEUTICAL RESEARCH
SJIF Impact Factor 8.084
Volume 10, Issue 1, 993-1003. Research Article ISSN 2277– 7105

DIFFERENTIAL EXPRESSION ANALYSIS IN RNA-SEQUENCING DATA OF *SACCHAROMYCES CEREVISIAE* AFTER TREATMENT WITH METHYL METHANESULFONATE (MMS)

Meenakshi Lal, Sanjeev Ranjan* and Rashi Gupta

Depatment of Bioinformatics, Amcon Biotech, Ranchi, India.

Article Received on
13 November 2020,

Revised on 04 Dec. 2020,
Accepted on 25 Dec. 2020

DOI: 10.20959/wjpr20211-19071

***Corresponding Author**
Sanjeev Ranjan
Depatment of
Bioinformatics, Amcon
Biotech, Ranchi, India.

ABSTRACT

The transcriptome is the complete set of transcripts in a cell, and their quantity, for a specific developmental stage or physiological condition. Understanding the transcriptome is essential for interpreting the functional elements of the genome and revealing the molecular constituents of cells and tissues, and also for understanding development and disease.Translation of aberrant mRNAs induces ribosomal collisions, thereby triggering pathways for mRNA and nascent peptide degradation and ribosomal recycling.[1] Ribosome stalling is demonstrated by the local accumulation of ribosomes at specific codon positions of mRNAs. This study demonstrates a computational workflow for the detection of differentially expressed genes and pathways from RNA-Seq data by providing a complete analysis of an RNA-Seq experiment profiling *Saccharomyces cerevisiae* after treatment with Methyl methanesulfonate (MMS) on colliding ribosomes. The study shows the differentiated gene of paired-end (upregulated and downregulated gene), biological pathway enrichment and network analysis between the control and treated samples.

KEYWORDS: RNA-Seq, Transcriptome, Differential expression.

INTRODUCTION

RNA sequencing (RNA-seq) has become a standard procedure to investigate transcriptional changes between conditionsand is routinely used in research and clinics[2] including identification of regulatory elements such as enhancer RNAs or long non-coding RNAs, biomarkers, responses to a signal, as well as to capture and model whole biological processes by time course (TC) data.[3–7] RNA sequencing (RNA-Seq) protocols have been continuously improved.

Translation elongation plays essential roles in diverse aspects of protein biogenesis, such as differential expression, co-translational folding, covalent modification, and secretion.[8] Cellular, tissue and organismal health face challenges of continuous changing conditions and exposure to extrinsic proteotoxic stressors. So, cellular protein homeostasis must be adapted to these changes. Proteotoxic stress arises during translation where mRNA processing errors can result in the translation of defective or truncated proteins and lead to the accumulation of toxic nascent protein products.[9] These deleterious proteins can lead to aggregation; which contribute to human pathologies including a wide range of neurodegenerative disorders.[10] Acellular quality control pathways have evolved to guard against the accumulation of these aberrant mRNA or nascent polypeptides to maintain homeostasis.[11] Histones protein is essential for packaging the genomic DNA of all eukaryotes into nucleosomes to form chromatin.[12] These proteins are encoded by multiple genes resulting in the excess production. The positively charged histones have a very high affinity for negatively charged molecules such as DNA, and any excess of histone proteins results in deleterious effects on genomic stability and cell viability.[13,14] Actively elongating ribosomal complexes whose progression is halted due to defective mRNA or emerging nascent chain is identified by ribosome-associated quality control (RQC) pathway.[15] After the initial recognition events, the RQC response catalyzes the degradation of both the mRNA and nascent polypeptide, followed by ribosomal subunit recycling. It has been shown that RQC failures result in the production of aberrant protein products and an eventual accumulation of protein aggregates.[16] RQC pathways have been genetically well-characterized in *S. Cerevisia*. The initial recognition event requires the ribosomal protein Asc1 and the ubiquitin ligase Hel2.[17,18]

RNA- Sequencing methods involve sequencing of RNA from a sample, quantification by mapping to reference genome and comparing between conditions comprising control and treated. A population of RNA is converted to a library of cDNA fragments by the process of reverse transcription with adaptors attached to one or both ends. Each molecule, with or without amplification, is then sequenced in a high-throughput manner to obtain short sequences from one end (single-end sequencing) or both ends (pair-end sequencing). The reads are typically 30–400 bp, depending on the DNA-sequencing technology used. This method can be used to identify both known and novel transcripts as well as assembling a trancriptome*de novo* without any prior knowledge of samples.[19,20] RNA- Sequencing can be done through Paired-end sequencing and Single-end sequencing. Paired-end sequencing allows to sequence both ends of a fragment and generate high-quality, alignable sequence data.[21] The alignment with the reference genome in paired-end sequencing is more accurate because the gap size between the ends of the fragment can be estimated.[22] In single-end reading, the sequencer reads a fragment from only one end to the other, generating the sequence of base pairs. RNA-Seq does not have an upper limit for quantification, which correlates with the number of sequences obtained. Consequently, it has a large dynamic range of expression levels over which transcripts can be detected: a greater than 9,000-fold range was estimated in a study that analysed 16 million mapped reads in *Saccharomyces cerevisiae.*[23]

Transcriptome assembly and differential expression analysis process overview:

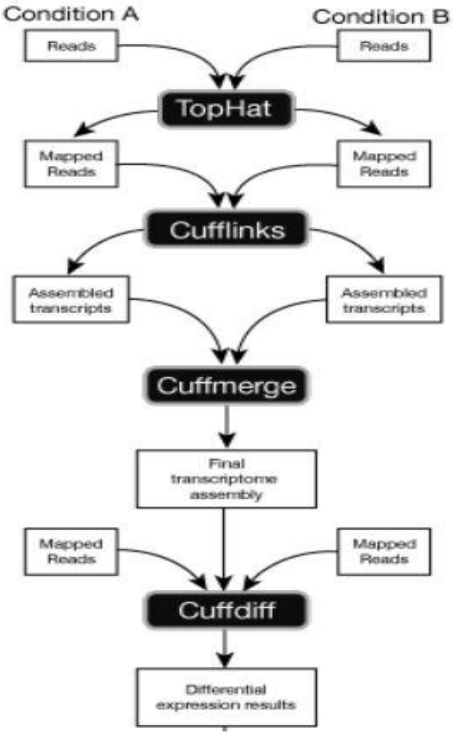

Fig 1: The Cufflinks RNA-Seq workflow (Reprint from: http://cole-trapnell-ab.github.io/cufflinks/manual/)

MATERIALS AND METHODS

Sample Data: In this study, following datasets were used:

2 control samples:**GSM4558776, GSM4558780**

2 treated samples: **GSM4558777, GSM4558781**

Data Upload

Data set were taken from National Center for Biotechnology Information (NCBI) and then fastq files were uploaded from European Bioinformatics Institute (EMBL-EBI) for each sample. The data were checked for quality and then mapping with reference genome was done before the analysis by using various bioinformatics tools. The gtf and genome files of reference genome were downloaded from Ensembl.

Quality Control

During sequencing, errors are introduced which can leads to analysis bias and misinterpretation of data. This is due to technical limitations of sequencing platform or presence of adapters if reads are longer than the fragments sequenced. So, trimming these reads is very essential to the number of reads mapped. In this study, Fastqc and cutadapt tools were used for this purpose.[24,25]

Script

```
#Runningfastqc
fastqc *.fastq
#trimming and cleaning (cutadapt)
cutadapt -b GATCGGAAGAGCACACGTCTGAACTCCAGTCACCCTTCGTTGCATCTCGT -
B GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG -
q 30,30 -m 20 -o trimmed_Treated1_Forward.fastq -p trimmed_Treated1_Reverse.fastq
control1_forward.fastq control1_reverse.fastq

cutadapt -b GATCGGAAGAGCACACGTCTGAACTCCAGTCACGTAGTACAGTATCTCGT -
B GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG -
q 30,30 -m 20 -o trimmed_control2_forward.fastq -p trimmed_control2_reverse.fastq
control2_forward.fastq control2_reverse.fastq

cutadapt -b GATCGGAAGAGCACACGTCTGAACTCCAGTCACCAGGTATTGAATCTCGT -
B GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG -
q 30,30 -m 20 -o trimmed_treated_forward.fastq -p trimmed_treated_reverse.fastq
treated1_forward.fastq treated1_reverse.fastq

cutadapt -b GATCGGAAGAGCACACGTCTGAACTCCAGTCACCAGGTATTGAATCTCGT -
B GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG -
```

```
q 30,30 -m 20 -o trimmed_treated2_forward.fastq -p trimmed_treated2_reverse.fastq treated2_forward.fastq treated2_reverse.fastq
```

```
cutadapt -b GATCGGAAGAGCACACGTCTGAACTCCAGTCACCAGGTATTGAATCTCGT -B GGGGGGGGGGGGGGGGGGGGGGGGAAAAAAAGGGGGGGGGGGGGGGGG -q 30,30 -m 20 -o trimmed2_treated2_forward.fastq -p trimmed2_treated2_reverse.fastq trimmed_treated2_forward.fastq trimmed_treated2_reverse.fastq.
```

Mapping

The reads were mapped against the genome to figure out which gene did they belong to. The process of aligning or mapping of the reads to the reference genome of the organism is called mapping. Reference genome of *Saccharomyces cerevisiae* was used to map the quality-controlled sequences using*TopHat 2* tool.

Script

alignment of reads on ref genome using tophat.

```
tophat2 -o Tophat_Control1 -G genome.gtf genome trimmed_control1_forward.fastq trimmed_control1_reverse.fastq
tophat2 -o Tophat_Control2 -G genome.gtf genome trimmed_control2_forward.fastq trimmed_control2_reverse.fastq
tophat2 -o Tophat_Treated1 -G genome.gtf genome trimmed_treated1_forward.fastq trimmed1_treated_reverse.fastq
tophat2 -o Tophat_Treated2 -G genome.gtf genome trimmed2_treated2_forward.fastq trimmed2_treated2_reverse.fastq.
```

The outputs were generated in BAM files. A BAM (Binary Alignment Map) file is a compressed binary file storing the read sequences, whether they have been aligned to a reference sequence (e.g. a chromosome), and if so, the position on the reference sequence at which they have been aligned.

A BAM file (or a SAM file, the non-compressed version) consists of:

- ➢ A header section (the lines starting with @) containing metadata particularly the chromosome names and lengths (lines starting with the @ symbol)
- ➢ An alignment section consisting of a table with 11 mandatory fields, as well as a variable number of optional fields:

Table: BAM SAM file format (Data from source: https://felixfan.github.io/bam-sam/).

Col	Field	Type	Brief Description
1	QNAME	String	Query template NAME
2	FLAG	Integer	Bitwise FLAG
3	RNAME	String	References sequence NAME
4	POS	Integer	1- based leftmost mapping POSition
5	MAPQ	Integer	MAPping Quality
6	CIGAR	String	CIGAR String
7	RNEXT	String	Ref. name of the mate/next read
8	PNEXT	Integer	Position of the mate/next read
9	TLEN	Integer	Observed Template LENgth
10	SEQ	String	Segment SEQuence
11	QUAL	String	ASCII of Phred-scaled base QUALity+33

Cufflinks

Cufflinks tool suit were used to assemble transcriptomes from RNA-Seq data and quantify their expression.

Script

```
#assembling all transcript using cufflinks
cufflinks -o cufflinks_Control1 -G genome.gtf /home/hp/Documents/Publication/Control_1/Tophat_Control1/accepted_hits.bam
cufflinks -o cufflinks_Control2 -G genome.gtf /home/hp/Documents/Publication/Control_2/Tophat_Control2/accepted_hits.bam
cufflinks -o cufflinks_Treated1 -G genome.gtf /home/hp/Documents/Publication/Treated_1/Tophat_Treated1/accepted_hits.bam
cufflinks -o cufflinks_Treated2 -G genome.gtf /home/hp/Documents/Publication/Treated_2/Tophat_Treated1/accepted_hits.bam
```

Cuffmerge

The assembled transcriptomes from multiple RNA-seq libraries were merged into a master transcriptome by the tool cuffmerge. This step is required for the differential analyses of the new transcript which had been assembled.

Script

```
#assemble all .gtf file generated from cufflinks program we have to use cuffmerge. here assemled.txt file contains path of "transcript.gtf" file for every sample
cuffmerge -o cuffmerge_output -g genome.gtf -s genome.fa assemblies.txt
```

CuffDiff

CuffDiff tool was finally used to analyse the differential expression of the RNA-Seq data. It is a highly accurate tool which could perform the comparison of expression levels of gene and transcripts in RNA-Seq experiment. This would provide information regarding which genes were up-regulated or down-regulated between two conditions, which genes are differentially spliced or are undergoing other types of isoform-level regulation.

RESULTS AND CONCLUSION

Gene Expression by HeatMap2

A heat map is a graphical representation of data where the individual values contained in a matrix are represented as colors. HeatMap2 is commonly used for the visualization of expression of gene across smaples. The heatmap gives an overview of similarities and dissimilarities between samples: the color represents the distance between the samples.

Create Heatmaps using R

R is a language and environment for statistical computing and graphics.

Script for creating HeatMap:

```
library(gdata)
library(gplots)
data=read.xls("/home/hp/Documents/heatmap_input_data.xlsx",sheet=1)
head(data)
ind = which(data[,6]< 0.05)
data1 = as.matrix(data[ind,c(2,4)])
rownames(data1) = data[ind,1]
data1[is.nan(data1)] = 0
class(data1) = "numeric"
data1 <- data1[is.finite(rowSums(data1)),]
head(data1)
heatmap.2(data1[1:30,], trace = "none", density.info = "none" ,col = redgreen(75), cexRow
= 0.9 , cexCol = 0.9 , offsetCol = 0 , offsetRow = 0, srtCol = 330 , adjCol = c(0, 0.3))
```

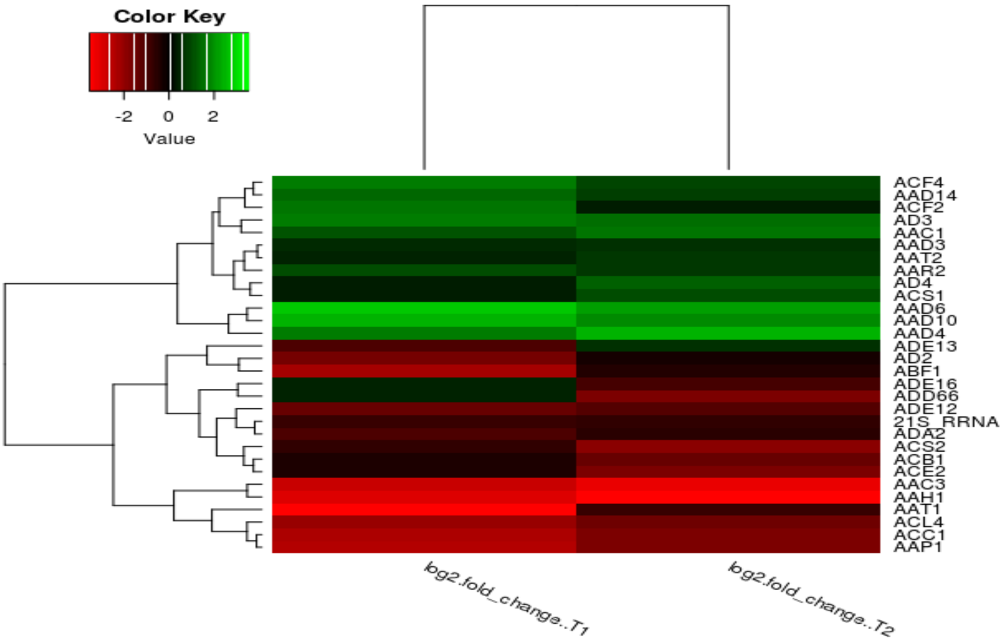

Fig2: The expression patterns of the ~7000 predicted yeast genes in response to MMS treatment, as measured by RNA-Sequencing technology over 4 samples (2 control, 2 treated).

A clustering method was used to organize genes according to their similarity in expression such that genes with similar expression patterns were "clustered" together. The data were graphically displayed in tabular format. The variations in transcript abundance for each gene are represented by a color scale, in which shades of red represent increases and shades of green represent decreases in mRNA levels, relative to the control sample. The saturation of the color is proportionate to the magnitude of the variation in transcript levels. In addition, the clustering algorithm generates a dendrogram that indicates the relationships between the expression patterns of genes. Genes with similar patterns of expression over multiple experiments were thus grouped together on a common branch of the dendrogram and can also be recognized by an obvious pattern of contiguous patches of color in the cluster diagram.

The results of clustering revealed that MMS treated samples triggered rapid and extensive changes in the genomic expression as compared to the controlled samples. Approximately, 931 genes were upregulated and 1435 were downregulated.

Network Analysis by Cystoscope

Cystoscope tool helps in both the visualization and analysis of biological data. Large biological data are easily visualized as graphs, i.e., a set of nodes and edges. Nodes are representations of biological molecules and edges connect the nodes depicting some kind of relationship.

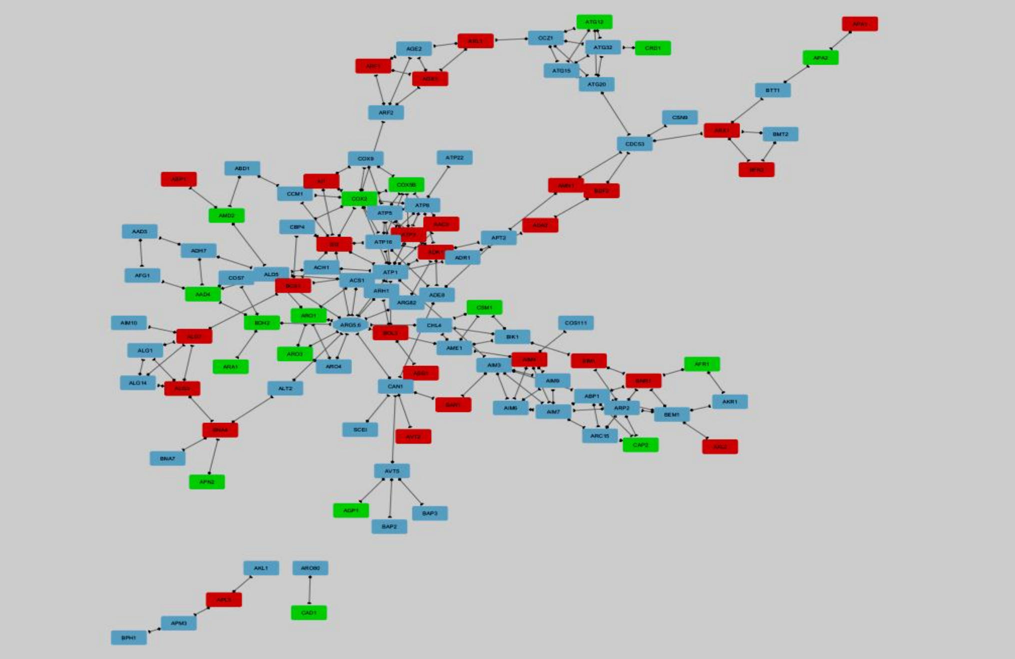

Fig. 3: Showing cystoscope result for the dataset under study.

From the analysis, we came to the conclusion that 200 were significant genes from which 48 genes were showing upregulation while 45 genes were downregulated on the basis of log fold change values. Gene ARG5 was considered as the hub node of the network.

REFERENCES

1. Niladri K Sinha, Alban Ordureau, Katharina Best, et. al. EDF1 coordinates cellular responses to ribosome collisions. Elife, Aug 3; 2020; 9: e58828. doi: 10.7554/eLife.58828.
2. Daniel Spies, Peter F. Renz, Tobias A. Beyer and Constance Ciaudo., Comparative analysis of differential gene expression tools for RNA sequencing time course data. Briefings in Bioinformatics, 2017; 1–11. doi: 10.1093/bib/bbx115
3. Acerbi E, Vigano` E, Poidinger M, et al. Continuous time Bayesian networks identify Prdm1 as a negative regulator of TH17 cell differentiation in humans. Sci Rep., 2016; 6: 23128.
4. doAmaral MN, Arge LW, Benitez LC, et al. Comparative transcriptomics of rice plants under cold, iron, and salt stresses. FunctIntegr Genomics, 2016; 16: 567–79.

5. Giannopoulou EG, Elemento O, Ivashkiv LB. Use of RNA sequencing to evaluate rheumatic disease patients. Arthritis Res Ther., 2015; 17: 167.
6. Sudmant PH, Alexis MS, Burge CB. Meta-analysis of RNA-seq expression data across species, tissues and studies. Genome Biol., 2015; 16: 287.
7. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet, 2009; 10: 57–63.
8. Richmond TJ, Davey CA. The structure of DNA in the nucleosome core. Nature, 2003; 423: 145–150.
9. Brandman, O., Hegde, R.S. Ribosome-associated protein quality control, 2016; 23: 7–15.
10. Gestwicki, J.E., and Garza, D.Protein quality control in neurodegenerative disease, 2012; 107: 327–353.
11. Lykke-Andersen, J., and Bennett, E.J. Protecting the proteome: Eukaryotic cotranslational quality control pathways, 2014; 204: 467–476.
12. vanHolde KE Chromatin; van Holde KE, editor. 1988; New York: Springer-Verlag.
13. Gunjan A, Verreault A., A Rad53 kinase-dependent surveillance mechanism that regulates histone protein levels in S. cerevisiae. Cell, 2003; 115: 537–549.
14. Singh RK, Kabbaj MH, Paik J, Gunjan A., Histone levels are regulated by phosphorylation and ubiquitylation-dependent proteolysis. Nat Cell Biol., 2009; 11: 925–933.
15. Brandman, O., and Hegde, R.S., Ribosome-associated protein quality control, 2016; 23: 7–15.
16. Choe, Y.J., Park, S.H., Hassemer, T., Korner, R., Vincenz-Donnelly, L., Hayer-Hartl, M., and Hartl, F.U., Failure of RQC machinery causes protein aggregation and proteotoxic stress, 2016; 531: 191–195.
17. Brandman, O., Stewart Ornstein, J., Wong, D., Larson, A., Williams, C.C., Li, G.W., Zhou, S., King, D., Shen, P.S., Weibezahn, J., et al., A ribosome-bound quality control complex triggers degradation of nascent peptides and signals translation stress, 2012; 151: 1042–1054.
18. Kuroha, K., Akamatsu,M.,Dimitrova,L.,Ito,T.,Kato,Y.,Shirahige,K.,andInada,T.Receptor for activated C kinase 1 stimulates nascent polypeptide-dependent translation arrest, 2010; 11: 956–961.
19. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq.Nat Methods, 2008; 5(7): 621–628. doi: 10.1038/nmeth.1226.

20. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol, 2010; 28(5): 511–515. doi: 10.1038/nbt.1621.
21. Hu Y, Wang K, He X, Chiang DY, Prins JF, Liu J. A probabilistic framework for aligning paired-end RNA-seq data. Bioinformatics, 2010; 26(16): 1950–1957. doi: 10.1093/bioinformatics/btq336.
22. Li H, Homer N. A survey of sequence alignment algorithms for next-generation sequencing. Brief Bioinform, 2010; 11(5): 473–483. doi: 10.1093/bib/bbq015.
23. Nagalakshmi U, et al. The transcriptional landscape of the yeast genome defined by RNA sequencing. Science, 2008; 320: 1344–1349.
24. Trapnell, C., L. Pachter, and S. L. Salzberg, TopHat: discovering splice junctions with RNA-Seq. Bioinformatics, 2009; 25: 1105–1111.
25. Marcel, M., Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet. journal, 2011; 17910: 10-12.