Grenze International Journal of Engineering and Technology, June Issue

Analysis of Gene Expressed Data of Mouse Strains using Bio-Conductor Packages: A Case Study

Santhosh K[1] and Dr. Ajitha S[2]
[1]Department of Computer Science, JSS College for Women(Autonomous), Mysore, India (Affiliated to University of Mysore)
Email ID: santhukprem@gmail.com
[2]Department of Computer Applications, M S Ramaiah Institute of Technology, Bangalore, India (Affiliated to VTU)
Email ID: ajithasankar@gmail.com

***Abstract*—Analysis of gene expressed data can be performed by various methods. Recently Bio-conductor packages are doing vital role in analysis of gene expressed data. Here we demonstrated how these Bio-conductor packages gain key role in genome analytics. We use C57BL and DBA/2J mouse strains in our case study for gene expression analysis using bio-conductor packages. Mean while, the other methods currently available for mouse genome analysis are B6 strain sequence and oligonucleotide microarray probes. Both these methods are give results in average statistics compared with the bio-conductor packages methods. We show that, by bio-conductor packages with multiple replicates using stringent data, we will get more expressions which are very helpful for identifying of particular matching genome. Finally we conclude that, additional information is gained from Limma compared to edgeR and Deseq2 bio-conductor technique.**

***Index Terms*— replicates, hetero scedasticity, bio-conductor, expression level.**

I. INTRODUCTION

Gene expression can be defined as, the process in which information from a gene is used in the combination of a functional gene product. These products are repeatedly proteins, but in the non-protein coding genes like transfer RNA (tRNA) or small nuclear RNA (snRNA) genes, the merchandise may be a functional RNA, author explains in [1]. The process of gene expression is used in well-known life: prokaryotes (bacteria and archaea), eukaryotes (together with multicellular organism), and utilized by viruses—to generate the macromolecular machinery for life, author mentioned in [2] and [3].

Genome databases will allow us by permitting for storing, comparison and sharing of data across all data types, across organisms, across research studies and across individuals. The Genome Database (GDB) is the authorized central repository for genomic mapping data to ensure from the Human Genome Initiative. In 1990, Johns Hopkins University in Baltimore, Maryland, USA was established the Genome Database. The Human Genome Initiative is a worldwide research effort to analyze the structure of human DNA and determine the location and sequence of the estimated 100,000 human genes, author describes in [4], [5] and [6].

It is very difficult task to find the interested information in a given database and in each database amount of data is highly enormous. In the next process of analysis, we need to be considered the hidden patterns, detection of associations, and other features of interest to get proper interpretation. Computer programs should be create to

Grenze ID: 01.GIJET.8.2.515_2

searching, retrieving and analyzing from various data from different database. It's the primary task of website host to give functions like searching, retrieving and analyze the expressed data, author explains future analysis tools in [7] and [8].

II. Literature Survey

Below list is the common procedure for differential gene expression analysis.

1. Sequencing of genes
    - ✓ RNA Extraction
    - ✓ Library Preparation
2. Bioinformatics
    - ✓ Processing of sequencing reads
    - ✓ Gene expression levels
    - ✓ Normalization
    - ✓ Differentially Expressed (DE) genes

Each cell consists of proteins and DNA will be extracted from the cellular atmosphere then the RNA will be sequenced. Liquid-liquid extractions with acidic phenol-chloroform or silica gel based membranes will be used for RNA isolation. Silica-gel membrane is used to bind RNA and the remaining cellular components will be washed away. Ethanol is used to bind with Silica gel membranes. The quantity of ethanol directly influences to transcripts which are bound to the membrane, so we consider higher volume for better results.

The cellular components are categorized into fallowing types when we use with phenol-chloroform extraction

- ✓ Organic phase
- ✓ Interphase and
- ✓ Aqueous phase

Phenol-chloroform extraction is naturally followed by an alcohol precipitation in the direction of de-salt and concentrate the RNA. Ethanol or Isopropanol is used to perform an alcohol precipitation and salt is necessarily required by both. Different salts can be used to different precipitation efficiencies and these will produce an different RNA populations; e.g., lithium chloride is mainly used salt and will be reported as loss of snRNAs, tRNAs, and rRNAs. Multitude of factors is recommended to overall outcome of RNA extraction in normal procedure. Processing of RNA is very significant with the standardized manner and highly controlled. RNA isolation is a very knowledgeable process in genomics data. Authors explained in [9] and [10] that, RNA extraction methods will highly influence RNA-seq data in different ways.

Libraries are used for high-throughput sequencing terms with respect to DNA fragments to sequence with matching protocol. We observed that, Illumina will provide acceptable high throughput sequencing for all platforms. DNA polymerase is one of the important property carried in Illumina methods. These methods will confirms the base specific records in all time DNA polymerase for the purpose of growing DNA with new nucleotide copy. Fluorescent signals are used for selection of Illumina platforms as they witnessing the process. Incorporation of A, C, T, G in each nucleotide are labeled for specific fluorophore may give different emission signals. Highly sophisticated microscopes programmed cameras are used to store all these signals, explained in [11] and [12].

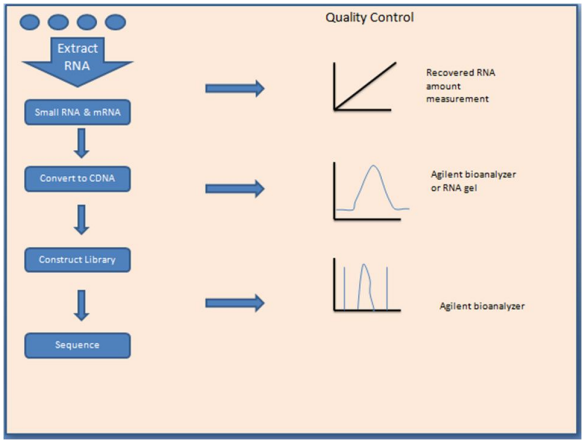

Figure 1: RNA library preparation workflow

One of the huge repository for nucleic acid sequence is the Sequence Read Archive and its capacity is around 25,000 tera bases, author described in [13] and [14]. NCBI, DNA Databank of Japan and European Bioinformatics Institute respectively are the three storage repository to store three copies of SRA, they use multiple mirrors to store, surf and download the data. Majority of databanks are using compressed formats for downloading the genome files, only few databanks will give in normal text format that is text files with FTP.
Htseq-count and feature Counts are the two important accepted tools for gene quantification which are used for largest tool packages. Among these two, htseq-count gives three different modes to adjust its behavior to define overlap instances, author explained in [15].
Individual gene expression levels are always gives absolute proxies when specified gene overlapping of more reads will interpreted directly. Effectiveness of amplification and sequencing of DNA fragments is directly influence the abundant factors. These will be analyzing always with single gene in a single samples on the number of reads. Importance has to give for RNA-sequence that, even given a uniform sampling of adverse transcript pool, the number of sequenced reads mapped and uniform sampling of adverse transcript to a gene are depends on:

- ✓ Gene expression levels
- ✓ Expression length
- ✓ Sequencing depth,
- ✓ Gene expressions within the sample.

Evaluating the gene expressed from two conditions is having too many methods. We calculate fraction of reads assigned to total number of reads in each gene with respect to the overall RNA databank and these may vary drastically from sample to sample. Normalization is applied for getting systematic effects from biological differences of interested for twist exploratory analyses. Again normalization is used to integer read counts with statistical tests conducted for gene expressed data, he explained in [16], [17] & [18].
Based on normalization procedure we can analyze the different gene of expression levels, there will be multiple ways to optimize the statistical tests to consider whether the selected gene expression is differs with two or more replicates. Here the two fundamental responsibilities of all DGE packages are mentioned below:

1. Considering the read counts with replicated samples we can Estimate the magnitude of differential expression, i.e. considering the differences in sequencing depth and variability we can calculate the fold change of read counts.
2. Expressed data will be analyzed with significance difference.

The top performing packages are explained in edgeR[19], DESeq/DESeq2[20] and [21], and limma-voom[22]. DESeq and limma-voom are more complex compare to edgeR but edgeR is good at experimentally 12 replicates and lesser. All these packages are working based on R language with at most features of statistical functions. From last two decades R language improved by advanced functions especially with small numbers of replicates. Regression-based models are used for gene expression difference with these statistical functions defined in R language with bio-conductor packages. We mentioned the important properties of DGE packages in below table.

TABLE I: COMPARISON OF PROGRAMS FOR DIFFERENTIAL GENE EXPRESSION IDENTIFICATION

Feature	Edger	DESeq2	limmaVoom
normalization of sequence	Trimmed median of means	Sample-wise size Factor	Trimmed median of means with gene-wise
Estimate of Dispersion	Cox-Reid approximate conditional inference moderated towards the mean	Cox-Reid approximate conditional inference with focus on maximum individual dispersion estimate	squeezes gene-wise residual variances towards the global variance
Assumed distribution	Neg. binomial	Neg. binomial	log-normal
Differential Expression testing	exact test for 2 factors; LRT for multiple factors	Multiple Factors LRT	t-test
False positives cases	Low	Low	Low
Differential Isoforms Detection	X	X	X
Supporting nature with multi-factor Experiments	√	√	√
Multiple replicates during runtime	50 sec to 4-5 min	40 sec to 3-4 min	30 sec to 2-4 min

III. Case Study & Discussion

Case Study for Digital Gene Expression of mouse strains using different Bio-conductor packages
We have considered the dataset of mouse strains C57BL & DBA/2J [26] for our case study. We used these three packages namely edgeR, Deseq2 and Limma for finding how genes are expressed using Rstudio as IDE and R Programming with statistical functions of bio-conductor for different replicates.

Using edgeR package
In Bio-conductor software packages edgeR is used to replicated count data with differential expression. In this case, we have taken different replicates for our experimental study and observed how genes are expressing when we apply statistical functions on dataset.

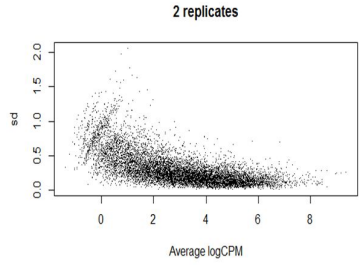

Figure 2: 2 replicates

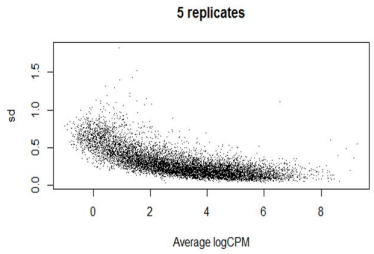

Figure 3: 5 replicates

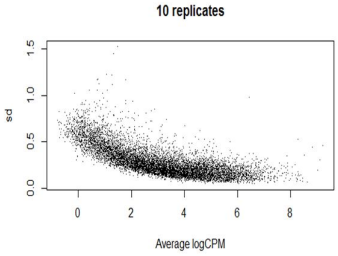

Figure 4:10 replicates

Discussion:
We observed that, larger average expression of genes are having average larger observed variances from corner to corner samples, that is, they vary in expression from sample to sample more than other genes with lesser regular expression. This phenomenon of having different scatter is known as data hetero scedasticity. In figure 2, we considered 2 replicates and observed that the typical deviation getting varies. When we considered 5 replicates then standard deviation becomes lesser values showed in figure 3. In figure 4 we taken 10 replicates and the standard deviation values are getting smaller.
Here we noticed that, heteroscedasticity somewhat large log-counts now encompass smaller scatter than small log-counts. Significantly, we notice the scatter tends to reduce as we enlarge the number of replicates. Therefore, if we add more data (replicates) then we obtain increasingly accurate estimates of group means. With more accurate estimates, we are more easily able to notify apart means which are close together.

Using DESeq2 package
DESeq2 is a one of the biocunductor package to find how the genes are expressed. Here we found that, high-throughput sequencing assays are generated with counting the data using variance-mean and negative binomial distribution is used to make a model for testing the differential expression.

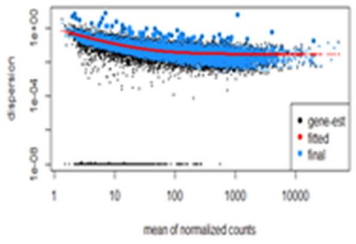

Figure 5: full dataset used

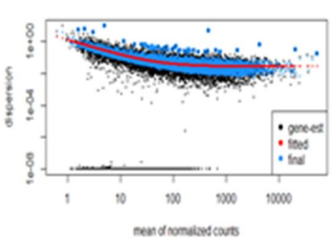

Figure 6: only 5 replicates used

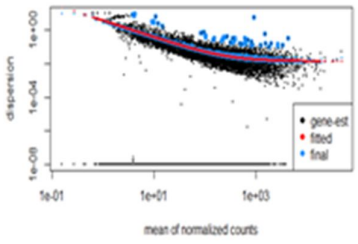

Figure 7: only 2 replicates used

Discussion:
The BCV (Biological Co-efficient Variation) value can be interpreted as the percentage of variation observed in the largedataset of a gene between replicates. In the above figure 5 we have taken full dataset for case study and we observed common BCV is average number which indicates that the large dataset for a gene can vary by small percentage up or down between replicates. In figure 6 we have taken 5 replicates and found that the BCV number increases as the number of biological replicates decreases. In figure 7 we have taken only 2 replicates

and found further increase in BCV number. We observed that, in less number of replicates we found large scatter and scatter will decrease when we consider more number of replicates.

Using Limma-Voom packages

Limma-Voom is a Bioconductor software package which can be used to integrated result for examine gene expression experiments with basic data. Limma package have very usefull features to conduct higher end experimental modules. From last few years, we can conclude that gene discovery through differential expression analyses of micro array and high-throughput PCR data can be achieved by the package Limma-Voom. This package will consist with particularly tough facilities for normalizing, reading and exploring the data.

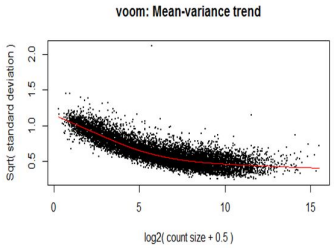

Figure 8: with full dataset

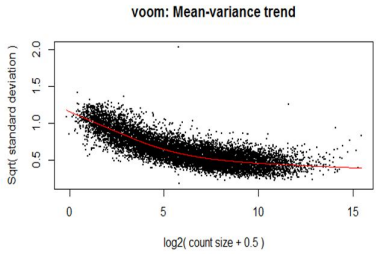

Figure 9: with 5 replicates

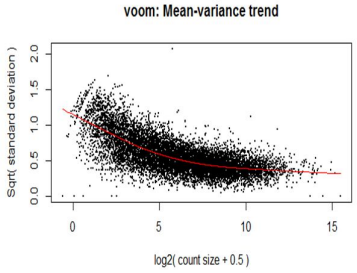

Figure 10: with 2 replicates

Discussion:

The Limma Voom transformation is applied to the read counts in our case study. We take the variance of genes to create weights for use in linear models then the log-cpm values with normalization are fitted by gene wise linear models. Each gene will make a residual standard deviation. In above figure 8, we considered full dataset for our experimental study and found smaller size of scattered in the graph. Scatter size is expanded when we consider five replicates that are represented in figure 9. In figure 10, we took 2 replicated and found large scatter in the graph. By observing these output, we conclude that, for less replicates we get large scatter and for more number of replicates we get small scatter in the graph.

IV. Future Work

- Preprocessing of the gene expression data for analysis
- To develop efficient algorithms for the analysis of gene expression data.
- To develop a frame work for gene expression data analytics using regression algorithm.

V. Conclusion

Differential gene expression analysis is an important stage where we can get many predictions on expressed genes. In the above case study, we considered three Bio-conductor packages (edgeR, DESeq2 and Limma) for gene expression analysis and found how genes are expressed with respect to the different replicates. Among these three packages Limma package will give better prediction and we will use this package for our future work.

Reference

[1] Soumiya Hamena and Souham Meshoul, "Multi-class classification of gene expression data using deep learning for cancer prediction", IJMLC, 2018.

[2] F. Alexander Wolf, Philipp Angerer and Fabian J. Theis, "SCANPY: Large-scale Single-cell gene expression data analysis", Genome Biology, 2018.

[3] Emily clough and Tanya Barrett, "The Gene Expression Omnibus Database", Springer Science, 2016.

[4] Anna Bernasconi, Arif Canakoglu, Marco Masseroli and Stefano Ceri, "META-BASE: a Novel Architecture for Large-Scale Genomic Metadata Integration", IEEE, 2020.

[5] Amalio Telenti and Xiaoqian Jiang, "Treating medical data as a durable asset", Nature Genetics, 2020.

[6] Lincoln D. stein, jean Thierry-mieg, "ACEDB: A Genome Database Management System", IEEE, 1999.

[7] B.D. Singh and A.K. Singh, "Bioinformatics Tools and Databases for Genomics Research", 2015.

[8] Michael Hackenberg and Rune matthiesen, "Annotation Modules: A tool for finding significant combinations for multisource annotations for gene lists", Bioinformatics Group, 2008.

[9] Lorenza Vitale, Allison Pioveson, Francesca Antonaros, “Dataset of differential gene expression between total normal human thyroid and histologically normal thyroid adjacent to papillary thyroid carcinoma” Elsevier Inc, 2019.
[10] Robinson MD and Oshlack A, “A scaling normalization method for differential expression analysis of RNA-seq data”, Genome Biology, 2010.
[11] Marc Sultan, VyacheslavAmstislavskiy, Thomas Risch, Moritz Schuette, Simon Döke , MeryemRalser, “Influence of RNA extraction methods and library selection schemes on RNA-seq data”, BMC Genomics, 2014.
[12] Leinonen R, Sugawara H, and Shumway M. “The sequence read archive”, Nucleic Acids Research, 2011.
[13] Liao Y, Smyth GK, and Shi W, “featureCounts: an efficient general purpose program for assigning sequence reads to genomic features”, Bioinformatics, 2014.
[14] Bullard JH, Purdom E, Hansen KD, and Dudoit S, “Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments”, BMC Bioinformatics, 2010.
[15] Law CW, Chen Y, Shi W, and Smyth GK, “voom: precision weights unlock linear model analysis tools for RNA-seq read counts”, Genome Biology, 2014.
[16] Koyel Mandal, Rosy Sarmah, and Dhruba Kumar Bhattacharyya , “POPBic: Pathway-based Order Preserving Biclustering algorithm towards the analysis of gene expression data”, IEEE Xplore, 2020.
[17] Christopher Heje Grønbech, Maximillian Fornitz Vording, Pascal N Timshel, Casper Kaae Sønderby, Tune H Pers, Ole Winther, “scVAE: variational auto-encoders for single-cell gene expression data”, Bioinformatics, 2020.
[18] Dong Li, Zhisong Pan, Guyu Hu, Graham Anderson and Shan He, “Active module identification from multilayer weighted gene co-expression networks: a continuous optimization approach”, IEEE Xplore, 2020.
[19] Robinson MD, McCarthy DJ, and Smyth GK, “edgeR: a Bioconductor package for differential expression analysis of digital gene expression data”, Bioinformatics, 2010.
[20] Anders S and Huber W, “DESeq: Differential expression analysis for sequence count data”, Genome Biology, 2010.
[21] Love MI, Huber W, and Anders S, “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2”, Genome Biology, 2014
[22] Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, and Smyth GK, “limma powers differential expression analyses for RNA-sequencing and microarray studies”, Nucleic Acids Research, 2015.
[23] https://github.com/mistrm82/msu_ngs2015/raw/master/bottomly_eset.RData
[24] https://www.hsls.pitt.edu
[25] http://www.ncbi.nih.gov
[26] https://github.com
[27] http://www.bioconductor.org
[28] http://genome.ucsc.edu
[29] https://www.rdocumentation.org/