RNA- SEQ DATA ANALYSIS OF EMBRYONIC STAGES OF BUFFALO GENOME IN CLONED AND IVF INDIVIDUALS

Major Project Dissertation

Submitted by

LIMMALA PRANATHI

For the partial fulfillment of the

Degree of Master of Science in
PLANT BIOTECHNOLOGY

Submitted to

Department of Biotechnology
TERI School of Advanced Studies

July, 2021

DECLARATION

This is to certify that the work embodied in this thesis “RNA- SEQ DATA ANALYSIS OF EMBRYONIC STAGES OF BUFFALO GENOME IN CLONED AND IVF INDIVIDUALS” is an original work carried out by me and has not been submitted anywhere else for the award of any degree.

I certify that all sources of information and data are fully acknowledged in the project thesis.

Limmala Pranathi
Date: 28/ 07/2021

CERTIFICATE

This is to certify that MS. LIMMALA PRANATHI has carried out her major project in partial fulfilment of the requirement for the Degree of Master of Science in Plant Biotechnology on the topic “RNA- SEQ DATA ANALYSIS OF EMBRYONIC STAGES OF BUFFALO GENOME IN CLONED AND IVF INDIVIDUALS” during January 2021 to May 2021. The project was carried out at National Institute of Plant Genome Research (NIPGR), New Delhi.

The <u>Dissertation</u> embodies the original work of the candidate to the best of our knowledge.

Date: 28/07/2021

Kumar

डॉ. शैलेश कुमार, पीएच.डी
Dr. Shailesh Kumar, PhD
स्टाफ वैज्ञानिक II
Staff Scientist II
राष्ट्रीय पादप जीनोम अनुसंधान संस्थान
National Institute of Plant Genome Research
अरुणा आसफ अली मार्ग, नई दिल्ली-110067
Aruna Asaf Ali Marg, New Delhi-110067

Dr. Shailesh Kumar
(External Supervisor)
Staff Scientist- III
NIPGR

Shashi Bhushan Tripl.

Dr. Shashi Bhushan Tripathi
(Internal Supervisor)
Head & Associate Professor
Department of Biotechnology
TERI School of Advanced Studies

Shashi Bhushan Tripl.

Dr. Shashi Bhushan Tripathi
Head and Associate Professor
Department of Biotechnology
TERI School of Advanced Studies

ACKNOWLEDGMENT

I would like to take this opportunity to convey my gratitude to each and every one who has been a part of this project for their constant encouragement, guidance and support, in completing this project, successfully.

This paper and the research wouldn’t have been possible without my external supervisor Dr. Shailesh Kumar, Staff Scientist- III, Bioinformatics laboratory, National Institute of Plant Genome Research (NIPGR), who gave me this opportunity of carrying out my dissertation in his laboratory and for believing in me. His continuous support and encouragement for me during my time at the institute was tremendous. It is my greatest pleasure to have had worked under such an amazing, young scientist with so much expertise in his field. His immense knowledge in research has been extremely helpful in effective completion of my work, despite of the situations arised due to the ongoing Coronavirus Pandemic.

To my internal supervisor Dr. Shashi Bhushan Tripathi, HOD and Associate Professor, Department of Biotechnology, TERI SAS, my sincere gratitude for his exceptional support. His knowledge on the subject of matter here, paved a way for some early insights on the project, helped me to expertly handle the dissertation. I, now, would like to thank all the faculty members of the Department of Biotechnology, TERI SAS, Dr. Pallavi Somvanshi, Dr. Ramakrishnan Sitaraman, Dr. Anandita Singh, Dr. Chaithanya Madhurantakam and Dr. Udit Soni for imparting their knowledge and sharing their experiences, which have been indeed very helpful during this project. They taught me patience and endurance in research which were so much needed to complete the project on time considering the current situation.

I present my sincere thanks to my mentors Ajeet Singh, PhD student, NIPGR and Nikita Poddar, Junior Research Fellow, NIPGR for their assistance and direction in completion of my work. I would like to thank my lab colleagues, A.T. Vivek, Shafaque Zahara, Malini Nemalikanti, Rohan Bharadwaj and Pragya Chitkara for providing a workable environment for my research work.

I would like to extend my gratitude to all the PhD Scholars, Department of Biotechnology, TERI SAS for their valuable assistance and support all this time. Lastly, I would like to thank all my friends and family for believing in me and motivating me to complete my project successfully.

LIST OF ABBREVIATIONS

Abbreviations	**Full forms**
RNA- Seq	RNA- Sequencing
DGE	Differential gene expression
mRNA	messenger RNA
cDNA	Complementary DNA
NDRI	National Dairy Research Institute
NIPGR	National Plant Genome Research
lncRNA	Long non- coding RNA
SBS	Sequencing by synthesis
RTA	Real time analysis
BCL	Base call file
R1	Read 1
R2	Read 2
fastq.gz	Compressed FASTQ file
FM	Ferragina- Manzini
SAM	Sequence Alignment Map
VCF	Variant calling format
BAM	Binary Alignment Map
HISAT	Hierarchical indexing for spliced alignment of transcripts
CRAM	Compressed columnar file format
GTF	Gene transfer files
GFF	Gene finding format
GLM	Generalized linear models
FastP	FAST (Protein)

TABLE OF CONTENTS

ABSTRACT

High throughput sequencing has increased rapidly over the years and of course omics have become that verb, indeed. The intrusion of computational methods in everyday biology has created greater avenues in understanding deeply the subject not only the genetics but also evolutionary relations. Water buffalo is commercially very important and enhancing our understanding on its genomics helps us to increase its growth and productivity. RNA- Seq data analysis performed here, enables us to identify the differentially expressed genes and thereby can be a reference to come up with better strategies for a more efficient yet faster data analysis techniques.

Keywords: RNA- Sequencing, Data analysis, Differential gene expression

CHAPTER 1: INTRODUCTION

RNA sequencing is a technique which was developed more than a decade ago. This technique was emerged just after the next generation sequencing technique, which happens to be in use extensively in biological sciences[1]. With the advent of RNA- Sequencing technique the limitations of expression microarrays were overcome very well. RNA- Sequencing has been playing a major role in enhancing the rate of transcriptomic research very effectively. This is most often used for analyzing the differential gene expression (DGE)[26]. The standard work flow of the expression analysis involves, RNA extraction followed by mRNA enrichment or ribosomal RNA depletion, cDNA synthesis and preparation of adapter-ligated sequencing libraries. A read depth of 10- 30 million reads per sample will be sequenced on a high throughput platform (Ex. Illumina). [26]The final steps in this workflow are Computational involving, aligning and assembling the sequencing reads to a transcriptome, quantifying reads that overlap transcripts, filtering and normalizing between samples and statistical modelling of significant changes in the expression levels of individual genes and/ or transcripts between sample groups[1].

Till date there are almost 100 techniques developed using a standard RNA- Seq protocol. The RNA-Sequencing methods are categorized as short- read sequencing, long- read cDNA sequencing and long-read direct RNA- sequencing. The short- read and the long– read c- DNA methods share similar steps in their protocols but, all methods require an adapter ligation step and all are affected by same quality and computational issues in library preparation both upstream and downstream.

Bubalus bubalis (Water Buffalo) is one of the most economically important animals which are being domesticated, over 3000- 6000 years now. [27]They provide a 5% of world's milk supply. Their milk contains higher amount of fat, lactose, protein and higher mineral content when compared to the cow's milk and so as their meat and other raw materials which have a greater economic importance[27]. Therefore, in order to increase the productivity and its outputs, research is being done tremendously and hence, various techniques are being used in the fields of Biotechnology and other applied sciences. Such, one, important technique is RNA- sequencing technique, applied recently on this buffalo, for more advanced research with respect to its genome[2-3].

However, very little is known about its genes and their expressions and as of now, the Y chromosome of the organism is yet to be sequenced completely as referred from NCBI, database. Considering the above-mentioned reasons, a research group at National dairy research institute (NDRI), Karnal, Haryana, has started working on the buffalo, by conducting transcriptomic studies on different embryonic samples such as, 2- celled, 8- celled and Blastula staged, in both cloned and invitro fertilized genomes[4- 15]. As, a huge chunk of data was collected on the same, the organization and the team collaborated with the National institute of plant genome research (NIPGR), Bioinformatics Department, New Delhi, to help them in further analyzing that data for good.

Therefore, the following are the key objectives for the team (includes me) at NIPGR:

1. Generating, different statistical plots such as MA Plot, Enhanced Volcano Plot and Heatmap.

2. To identify novel/ unique transcripts
3. To identify up and down regulated genes, respectively.
4. List of genes with different fold change ratios, such as, >10 fold, >5 fold and >2 fold.

CHAPTER 2: LITERATURE REVIEW

The type of RNA- sequencing performed on the biological samples of the organism under study here is paired- end Illumina sequencing[24]. Paired- end RNA- sequencing is used for quantification of long transcripts such as mRNA and lncRNA. As the size of the water buffalo genome is slightly smaller than the Human Genome, a paired- end RNA- sequencing had to be done. [28]The output or the reads obtained after performing a paired- end Illumina sequencing was in the form of FASTQ files. This Illumina sequencing technology uses cluster generation and sequencing by synthesis (SBS) chemistry, on millions/ billions of clusters on a flow cell depending on the sequencing platform[28]. The process involves sequencing by Real- time analysis (RTA) software on the instrument, where the base calls are saved and stored for every cell cycle. The RTA stores the data as individual base call file (BCL). Once this sequencing is done, the BCL files are converted into FASTQ files. In short, a FASTQ file contains the sequence data. In a single- end run only one read 1 (R1) is created whereas in paired- end run two reads, read 1 (R1) and read 2 (R2) will be created for each sample in each lane. FASTQ files are then compressed in an extension file as fastq.gz[16].

2.1 Hisat2

RNA- Seq analysis begins with the alignment of these FASTQ reads with its reference genome[19-20] to locate their origin. In the recent times to overcome the long hours of sequencing, HISAT, an alignment tool was developed. HISAT (hierarchical indexing for spliced alignment of transcripts) is a highly efficient system for aligning reads from the RNA- seq experiments. HISAT2 is a successor for HISAT and uses the same indexing scheme based on the Burrows- wheeler transform and the Ferragina- Manzini (FM) index. HISAT2 can align both DNA and RNA sequences using a graph Ferragina- Manzini indexing. In other words, this method is fast, memory efficient, search algorithm which provides an accurate variant analysis than many other methods[18- 19].

2.2 SAM tools

Due to the advancements in genomic sequencing and large-scale sequencing studies, new data formats became necessary for compact data storage and its efficient analysis. SAM is one of the most commonly used data formats these days. This was developed by 1000 Genome Projects in 2008. These specialized formats for storing read alignments (SAM) along with another type of data format VCF for genetic variants are row oriented, tab- delimited text files are easy to process custom scripts. The main part of the SAM tools package is a single executable that offers commands which are useful on the alignment data. The initial releases of the SAM tools could read and write the alignment data in SAM and BAM formats. The 1.0 version of the tool updated recently can now automatically detect the input files and can directly read and write the SAM, BAM and CRAM files[21].

2.3 StringTie

StringTie is a computational method which applies a network flow of algorithm on complex data sets useful in optimization of theory and optional assemblage of de novo transcripts.

StringTie is an efficient assembler than many other computational methods holding the same motive, such as Cufflinks. It involves a novel network flow algorithm in addition to an optional de novo assemblage step for assembling and quantitating the full-length transcripts representing multiple splice variants for each gene locus. Its input includes not only short reads that can be used by other transcript assemblers but also alignments of long sequences from the reads. These StringTie outputs can be processed further using specialized softwares such as Ballgown, Cufflinks and other programs (DESeq2, edgeR etc.) to identify the differentially expressed genes[22].

2.4 DESeq2

The sudden rise in high throughput sequencing technologies recently has resulted in production of huge chunk of diverse data bringing excellent advancements in the fields of Genetics and Molecular Biology. Therefore, in order to do research on the sequenced data, the world is in a dire requirement of simple, free yet swiftly functioning, computational tools/ methods. One such tool is DESeq2, which is being used in R studio to perform statistical analysis on the RNA- Seq count data, extracted from assemblers (StringTie).

The DESeq2 analysis is performed initially on the count data matrix. If we consider K as an example of a count data set matrix with a row for each gene i and another column each for sample j. The sequenced reads which are mapped haphazardly to a particular gene in the sample are considered as the matrix entries K_{ij}. The read counts K_{ij} follows a negative binomial distribution (gamma- Poisson distribution) with mean μ_{ij} and dispersion α_i. The mean considered as quantity q_{ij} proportional to the cDNA fragment concentration from the gene in the sample results in a normalized factor, s_{ij} which is $\mu_{ij} = s_{ij} q_{ij}$

In order to estimate the factors DESeq2 uses the same median- of- ratios method already being used by DESeq previously. To compare between two groups such as control and treated samples, the design matrix elements indicates whether the sample j is treated or not. The GLM fit model used here returns the coefficients indicating the overall expression of the gene and the log2 fold change ratio between the control and treatment. The use of such linear models hence provides us to analyze more complex designs and are being widely used in genomics.

An empirical Bayes shrinkage method is involved in the analysis which lets the shrinkage depends on the following:

- Estimation of the closeness between the true dispersion values to the fit.
- Degrees of freedom

DESeq2 estimates the prior distribution from the data and based on the observations from that data it controls the amount of shrinkage automatically.

CHAPTER 3: RESOURCES AND METHODS

3.1 Filtering the raw reads:

The RNA- Seq data set contained sequenced genomes of 2- celled, 8- celled and blastula stages i.e., the three different biological replicates of the buffalo embryo under study. Each biological replicate had a total of six technical replicates/ samples. There were 3 of each type, i.e., three Cloned and three IVF, samples. The reads were pair- end sequenced and were in FASTQ file format. FastQC is a popular tool to perform quality assessment and as a rule, read quality decreases towards the 3' end of reads, and if it becomes too low, bases should be removed to improve mappability.

For quality control, trimming of adapter sequences, filtering by quality and read pruning of the paired sequences a FASTP software tool was used initially and hence obtained trimmed files. FASTP is a tool devised for a fast all- in- one pre- processing of FASTQ files and developed in C++ with multithreading supported to high performance.

3.2 Aligning the reads:

Transcriptome mapping is very important and useful step in performing differential gene expression. The location of a gene on a particular chromosome is identified by mapping.

Hisat-2 is the software tool which was used on the trimmed FASTP read sequences and thereby the whole process of alignment and mapping with the reference genome i.e., with *Bubalus bubalis* resulted in sequence alignment map (SAM) files.

3.3 Reading the SAM and BAM files:

The SAM tools software was used to read the BAM, SAM files obtained as a result of running the HISAT2. This hence, created BAM files.

3.4 Obtaining the GTF files:

StringTie is a software tool used on the BAM files and hence individual Gene transfer files (GTF) were obtained for all the replicates and eventually merged them using StringTie merge function resulting in a StringTie merged GTF file which holds all the information about mRNA transcripts and exons present and helps us to find out novel transcripts present if any, in the obtained data set. A total gene count matrix and transcript count matrix was obtained finally, upon using StringTie yet again on the StringTie merged re-estimated GTF files.

3.5 GFF compare:

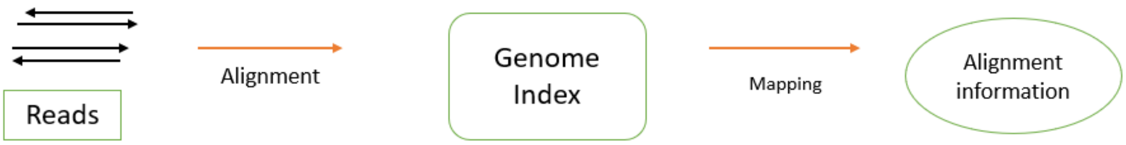
GFF compare software was used on the StringTie merged gtf file to obtain the total information about the amount of RNA transcripts present and hence novel transcripts were identified.

3.6 DESEQ2 analysis:

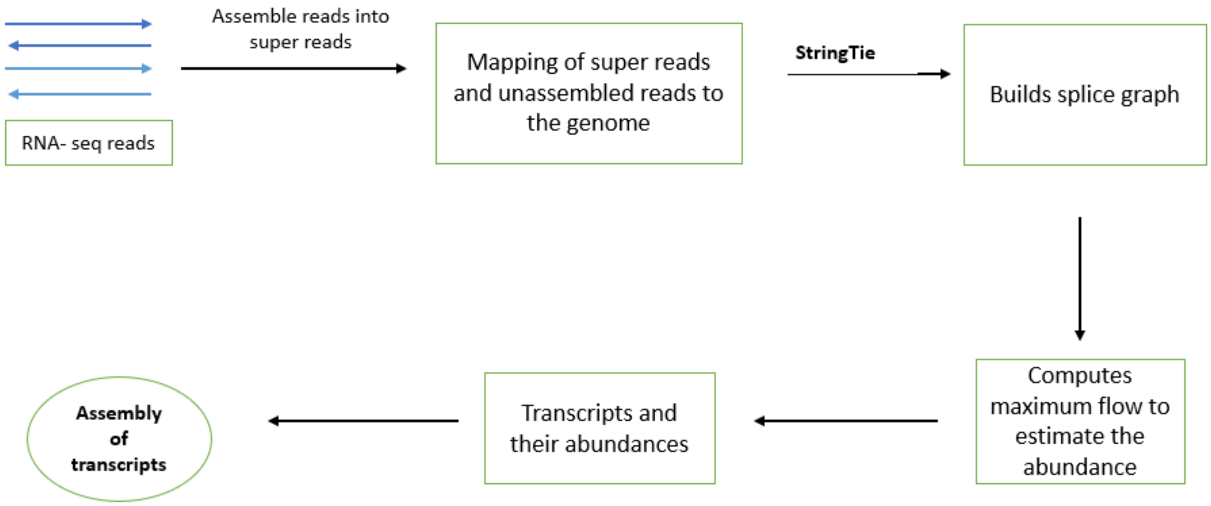
Upon using DESEQ2 package in R studio, on the total gene count matrix obtained, we could able to normalize the gene counts by applying an adjusted p-value of < 0.05 with a Log fold change ratio 2. This is how all the significant genes were obtained, depending on the fold change ratios and adjusted p-values in a xlsx file.

R-plots: Using the xlsx data, plotted MA plot, PCA analysis, Enhanced Volcano plot and Heat map for top 20 significant genes.

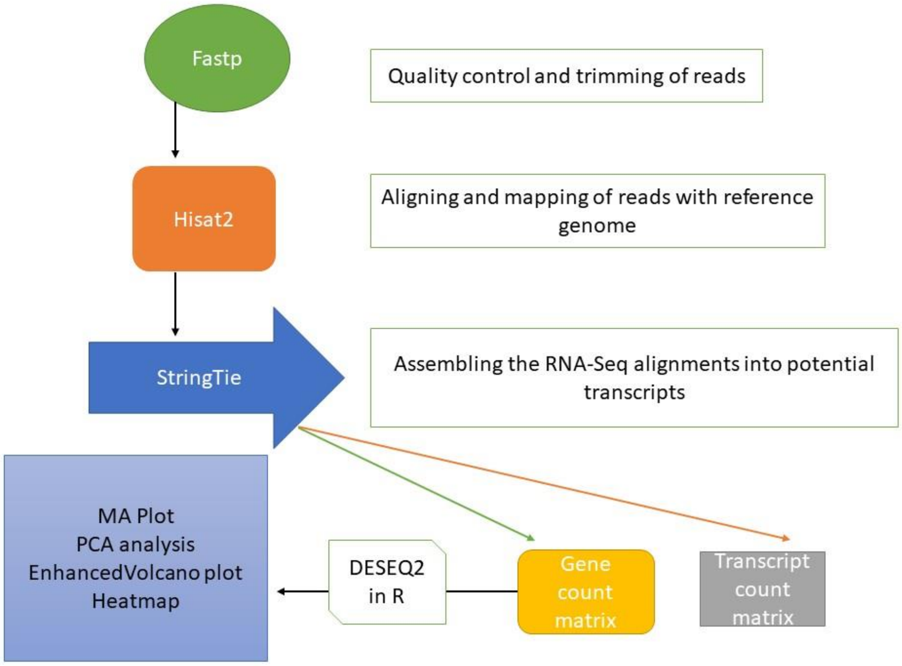
CHAPTER 4: LIST OF FIGURES



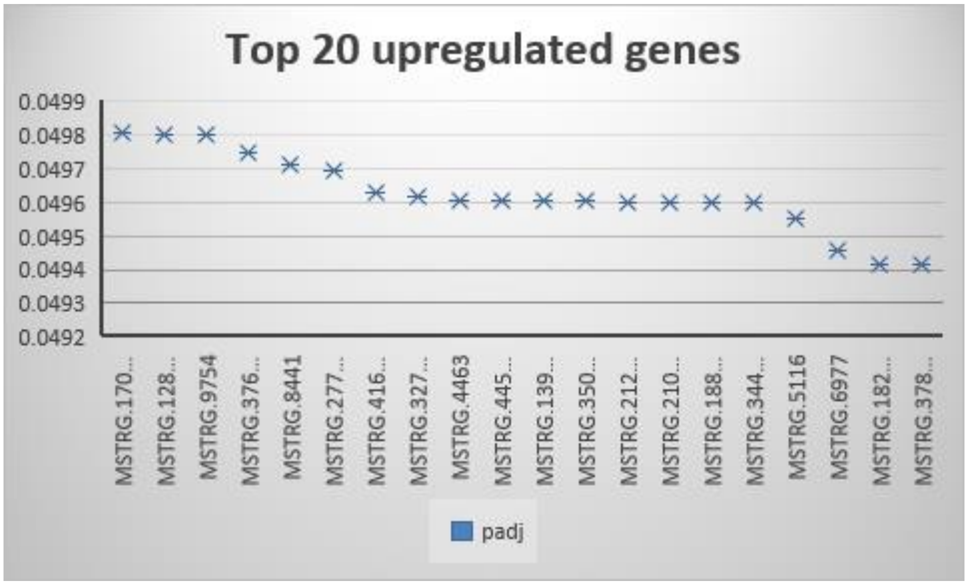
1 HISAT2 workflow


RNA- seq reads
Assemble reads into super reads
Mapping of super reads and unassembled reads to the genome
StringTie
Builds splice graph
Computes maximum flow to estimate the abundance
Transcripts and their abundances
Assembly of transcripts

2 StringTie workflow

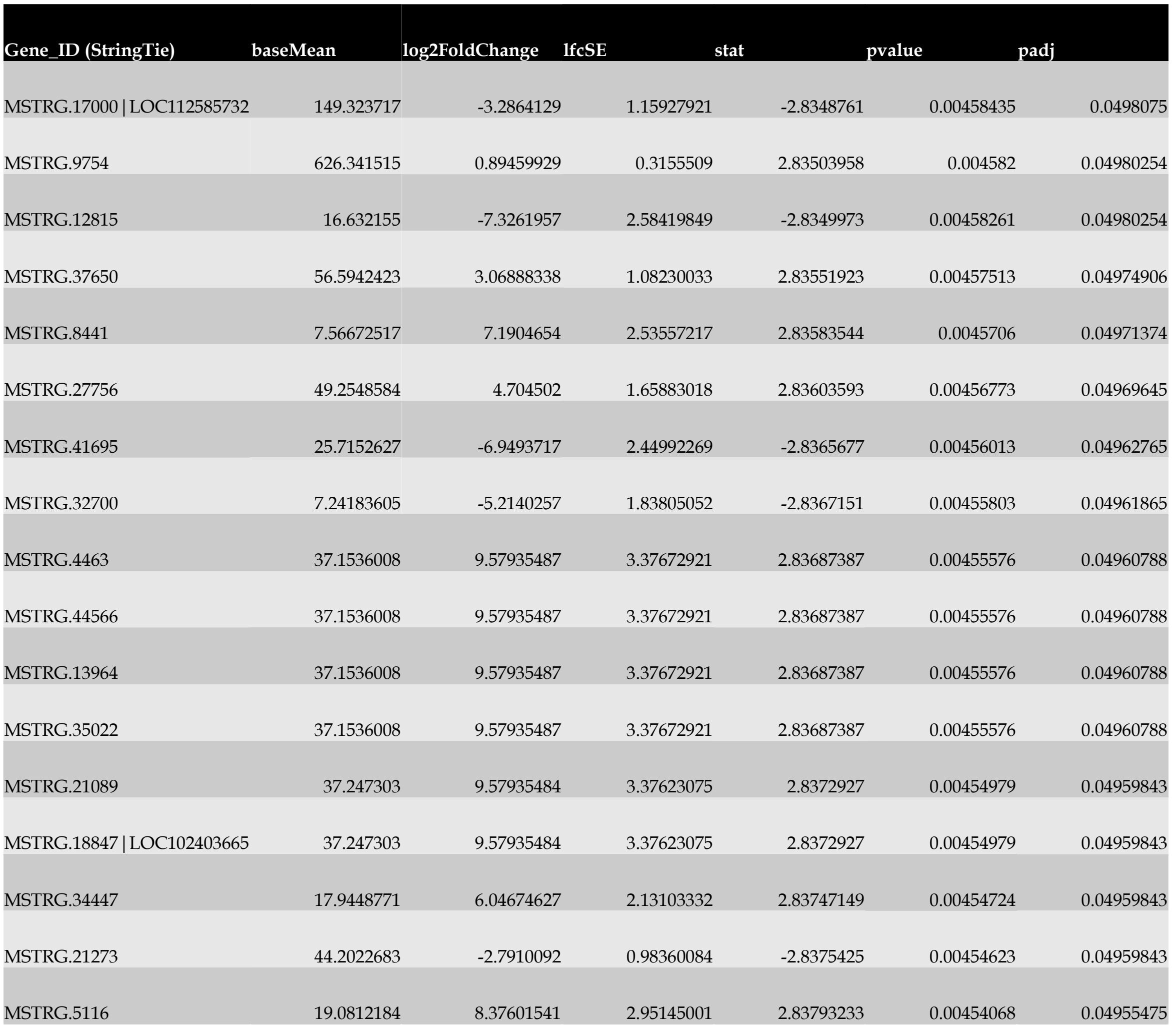

Fastp
Quality control and trimming of reads
Hisat2
Aligning and mapping of reads with reference genome
StringTie
Assembling the RNA-Seq alignments into potential transcripts
MA Plot
PCA analysis
EnhancedVolcano plot
Heatmap
DESEQ2 in R
Gene count matrix
Transcript count matrix

3 Differential gene expression workflow



4 Top 20 genes/ upregulated genes

CHAPTER 5: LIST OF TABLES

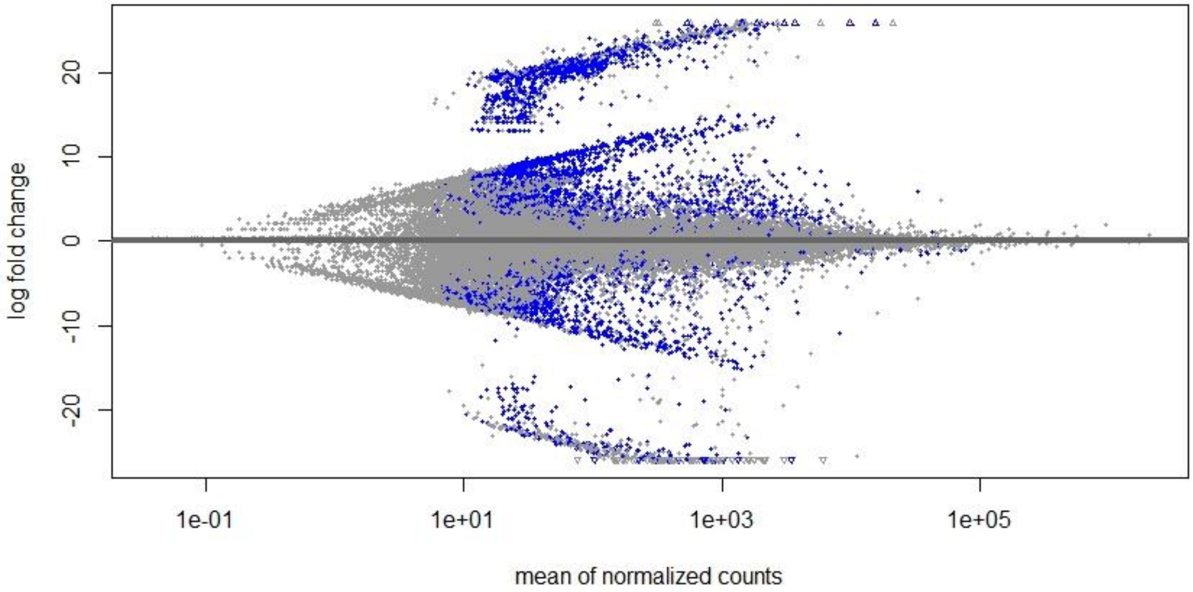


Gene_ID (StringTie)	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
MSTRG.17000 | LOC112585732	149.323717	-3.2864129	1.15927921	-2.8348761	0.00458435	0.0498075
MSTRG.9754	626.341515	0.89459929	0.3155509	2.83503958	0.004582	0.04980254
MSTRG.12815	16.632155	-7.3261957	2.58419849	-2.8349973	0.00458261	0.04980254
MSTRG.37650	56.5942423	3.06888338	1.08230033	2.83551923	0.00457513	0.04974906
MSTRG.8441	7.56672517	7.1904654	2.53557217	2.83583544	0.0045706	0.04971374
MSTRG.27756	49.2548584	4.704502	1.65883018	2.83603593	0.00456773	0.04969645
MSTRG.41695	25.7152627	-6.9493717	2.44992269	-2.8365677	0.00456013	0.04962765
MSTRG.32700	7.24183605	-5.2140257	1.83805052	-2.8367151	0.00455803	0.04961865
MSTRG.4463	37.1536008	9.57935487	3.37672921	2.83687387	0.00455576	0.04960788
MSTRG.44566	37.1536008	9.57935487	3.37672921	2.83687387	0.00455576	0.04960788
MSTRG.13964	37.1536008	9.57935487	3.37672921	2.83687387	0.00455576	0.04960788
MSTRG.35022	37.1536008	9.57935487	3.37672921	2.83687387	0.00455576	0.04960788
MSTRG.21089	37.247303	9.57935484	3.37623075	2.8372927	0.00454979	0.04959843
MSTRG.18847 | LOC102403665	37.247303	9.57935484	3.37623075	2.8372927	0.00454979	0.04959843
MSTRG.34447	17.9448771	6.04674627	2.13103332	2.83747149	0.00454724	0.04959843
MSTRG.21273	44.2022683	-2.7910092	0.98360084	-2.8375425	0.00454623	0.04959843
MSTRG.5116	19.0812184	8.37601541	2.95145001	2.83793233	0.00454068	0.04955475

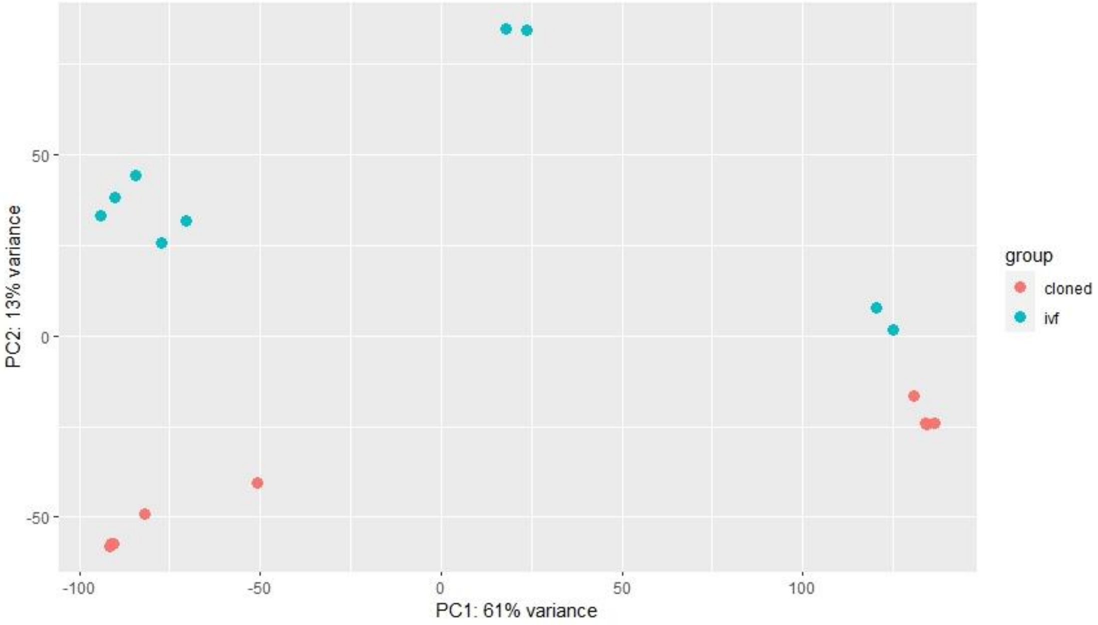
Table 1 Top 20 upregulated genes considering the p- value

CHAPTER 6: RESULTS AND DISCUSSION

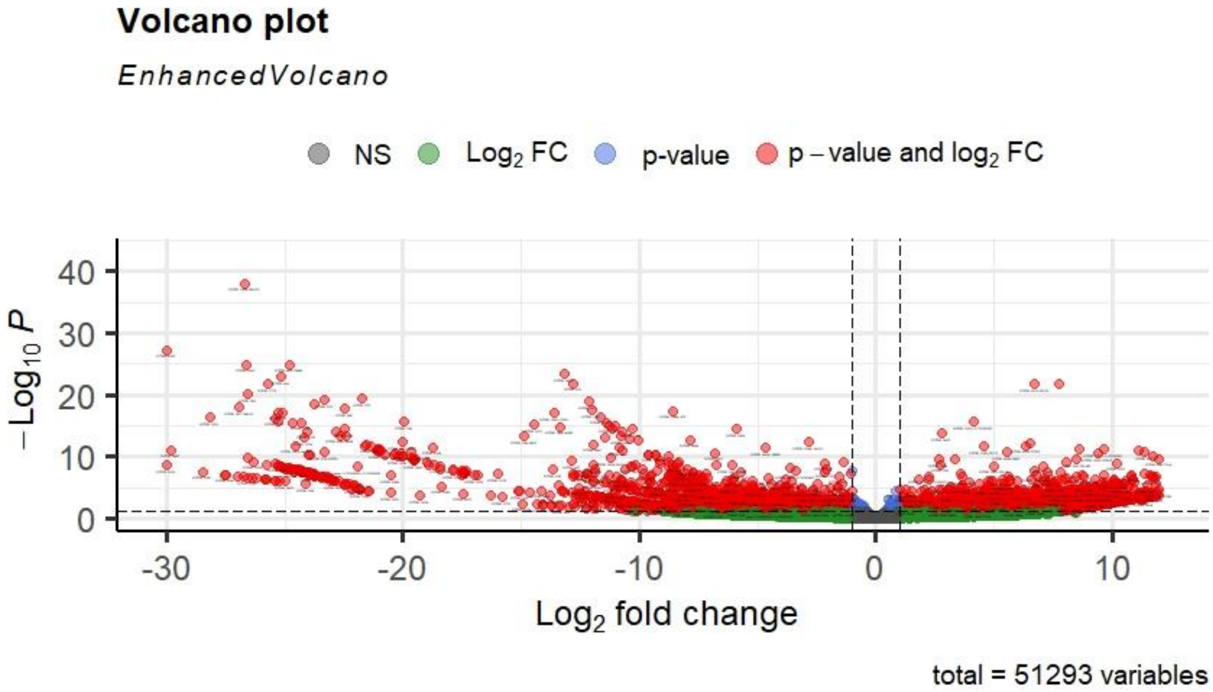
After performing the DESeq2 in R- studio on the total gene count matrix, the following results have been obtained.



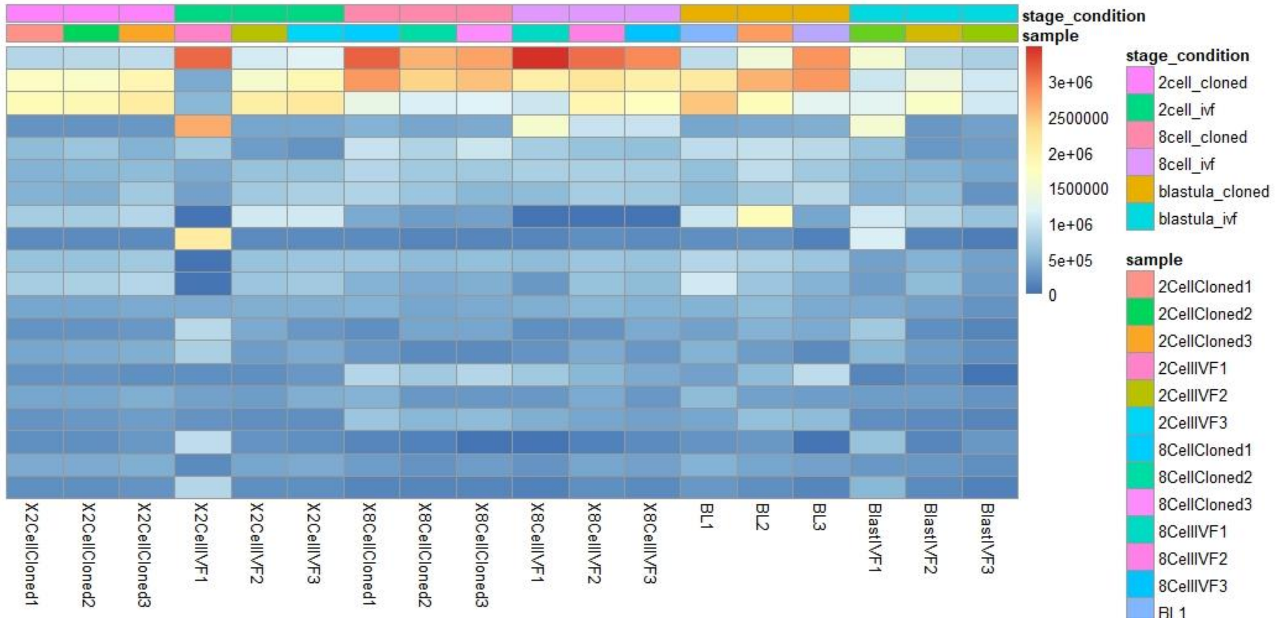
5 MA Plot



6 Principal component analysis (PCA Plot)



7 Volcano Plot



8 Heat map (Top 20 genes)

#Summary for the dataset: stringtie_merged.gtf

#Query mRNAs: 118920 in 57976 loci (86849 multi-exon transcripts)

(14808 multi-transcript loci, ~2.1 transcripts per locus)

#Reference mRNAs: 76339 in 33704 loci (67890 multi-exon)

#Super-loci wrt reference transcripts: 33415

#Levels	**Sensitivity**	**Precision**
Base level	100.0	85.5
Exon level	98.6	85.0
Intron level	100.0	94.2
Intron chain level	100.0	78.2
Transcript level	99.7	64.0
Locus level	99.4	57.4

Matching intron chains: 67890

Matching transcripts: 76114

Matching loci: 33518

Missed exons: 0/281881 (0.0%)

Novel exons: 31437/325858 (9.6%)

Missed introns: 37/242442 (0.0%)

Novel introns: 3852/257259 (1.5%)

Missed loci: 0/33704 (0.0%)

Novel loci: 24561/57976 (42.4%)

***Total union super-loci across all input datasets: 57976**

118920 out of 118920 consensus transcripts written in strtcmp.annotated.gtf (0 discarded as redundant)

Upon performing the DESeq2 software on the total gene count matrix, generated R-Plots such as MA plot, PCA, Volcano plot and Heatmap.

From the MA- plot and Volcano plots shown above i.e., in figures 7 and 9 respectively we can identify the genes which are upregulated and downregulated. The genes which are expressed above the value of zero are upregulated and below are downregulated genes.

MA Plot is an application of Bland- Altman visualization of genomic data. It makes us to understand differences between M (log ratio) and A (mean average) scales and also by plotting them, the values. Whereas a volcano plot helps us to fit more amount of data into the plot window. The visualization is very much similar to that of an MA plot. These plots are hence useful in a clear understanding of the genes which are expressed differentially from the RNA- Seq data set.

The principal component analysis (PCA) helps us to catch the variance in the data. There are axes which are called principal components in a PCA plot. As shown in the above results, in figure 8 we can observe that the normalized gene counts hold two principal components, PCA 1 and PCA 2. We observe that PCA 1 has a 61% and PCA 2 with 13% variance in gene expression. PCA 1 correlates more strongly with cloned individuals than that of PCA 2 which holds IVF individuals more.

From the generated heatmap, we understand the level of gene expression in different stages of the individuals and the kind of genes expressing. We can always carry out divergent studies further on the same. Its multiple colors varies both in the type of stage as well the individual. Heatmap follows a clustering algorithm which reorders both variables and observations made all along. This also computes the distance between each pair of rows and columns and orders them with respect to their similarities. The matrix was constructed considering the samples and stage condition i.e., 2- celled, 8- celled and Blastula stages in both Cloned and IVF individuals respectively.

The novel genes can be identified by visualizing the StringTie merged gtf file where we can observe different mRNA transcripts and can also generate their Ensembl or NCBI IDs accordingly.

CHAPTER 7: CONCLUSION

This way, RNA- Seq data analysis plays a major role in cutting down the work load by involving computational and advanced algorithms as different softwares to visualize the data efficiently in less time. This is one of the pioneer techniques for the emergence and advancements in various other Omics techniques. The curiosity never dies and new science has always evolved and sorting out datasets which were left since many years can be helpful for new discoveries and inventions.

It helps us to answer different scientific questions such as:

- Sample profiling in gene expression studies
- To study alternative splicing events associated with diseases.
- To identify allele- specific expressions, disease associated single nucleotide polymorphisms (SNP' s) and gene fusions to understand issues such as disease casual variants in like cancer.
- In recent days single cell RNA- Seq emerged as a way to study complex biological processes, cellular heterogeneity, and diversity especially in stem- cell biology and also in the field of Neuroscience.

Advantages of RNA- Seq data analysis:

- The analysis is not limited for the organisms which have a prior whole genome information, it can always be performed on new organisms whose genomes aren't sequenced yet or no information on their sequenced genomes is available.
- It has higher sensitivity for genes expressed in both high and low levels and a higher dynamic range of expression levels over which transcripts can be detected.
- A lower technical variation and higher reproducibility is observed.
- This is unbiased unlike microarrays with respect to cross hybridization.

CHAPTER 8: REFERENCES

1. Stark, R., Grzelak, M. and Hadfield, J. (2019) ‘RNA sequencing: the teenage years’, Nature Reviews Genetics, 20(11), pp. 631–656. doi: 10.1038/s41576-019-0150-2.
2. Michelizzi, V. N. et al. (2010) ‘Water Buffalo Genome Science Comes of Age’, International Journal of Biological Sciences, pp. 333–349. doi: 10.7150/ijbs.6.333.
3. Sood, T. J. et al. (2019) ‘RNA sequencing and transcriptome analysis of buffalo (*Bubalus bubalis*) blastocysts produced by somatic cell nuclear transfer and in vitro fertilization’, Molecular Reproduction and Development, 86(9), pp. 1149–1167. doi: 10.1002/mrd.23233.
4. Chen, F. et al. (2020) ‘Maternal transcription profiles at different stages for the development of early embryo in buffalo’, Reproduction in Domestic Animals, 55(4), pp. 503–514. doi: 10.1111/rda.13644.
5. Chitwood, J. L. et al. (2013) ‘RNA-seq analysis of single bovine blastocysts’, BMC Genomics, 14(1), p. 350. doi: 10.1186/1471-2164-14-350.
6. Driver, A. M. et al. (2012) ‘RNA-Seq analysis uncovers transcriptomic variations between morphologically similar in vivo- and in vitro-derived bovine blastocysts’, BMC Genomics, 13(1), p. 118. doi: 10.1186/1471-2164-13-118.
7. Graf, A. et al. (2014) ‘Fine mapping of genome activation in bovine embryos by RNA sequencing’, Proceedings of the National Academy of Sciences, 111(11), pp. 4139–4144. doi: 10.1073/pnas.1321569111.
8. He, X. et al. (2019) ‘Characterization and comparative analyses of transcriptomes of cloned and in vivo fertilized porcine pre-implantation embryos’, Biology Open, p. bio.039917. doi: 10.1242/bio.039917.

9. Huang, W. and Khatib, H. (2010) ‘Comparison of transcriptomic landscapes of bovine embryos using RNA-Seq’, BMC Genomics, 11(1), p. 711. doi: 10.1186/1471-2164-11-711.

10. Jiang, Z. et al. (2014) ‘Transcriptional profiles of bovine in vivo pre-implantation development’, BMC Genomics, 15(1), p. 756. doi: 10.1186/1471-2164-15-756.

11. Mamo, S. et al. (2011) ‘RNA Sequencing Reveals Novel Gene Clusters in Bovine Conceptuses Associated with Maternal Recognition of Pregnancy and Implantation1’, Biology of Reproduction, 85(6), pp. 1143–1151. doi: 10.1095/biolreprod.111.092643.

12. Min, B. et al. (2015) ‘Transcriptomic Features of Bovine Blastocysts Derived by Somatic Cell Nuclear Transfer’, G3 Genes|Genomes|Genetics, 5(12), pp. 2527–2538. doi: 10.1534/g3.115.020016.

13. Pang, C.-Y. et al. (2019) ‘Global transcriptome analysis of different stages of preimplantation embryo development in river buffalo’, PeerJ, 7, p. e8185. doi: 10.7717/peerj.8185.

14. Xu, W. et al. (2015) ‘RNA-Seq transcriptome analysis of porcine cloned and in vitro fertilized blastocysts’, Journal of Integrative Agriculture, 14(5), pp. 926–938. doi: 10.1016/S2095-3119(14)60866-2.

15. Zhang, L. et al. (2020) ‘RNA sequencing revealed the abnormal transcriptional profile in cloned bovine embryos’, International Journal of Biological Macromolecules, 150, pp. 492–500. doi: 10.1016/j.ijbiomac.2020.02.026.

16. Hughes, G. M. et al. (2013) ‘Using Illumina next generation sequencing technologies to sequence multigene families in de novo species’, Molecular Ecology Resources, 13(3), pp. 510–521. doi: 10.1111/1755-0998.12087.

17. Kim, D. et al. (2019) ‘Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype’, Nature Biotechnology, 37(8), pp. 907–915. doi: 10.1038/s41587-019-0201-4.

18. Kim, D., Langmead, B. and Salzberg, S. L. (2015) ‘HISAT: a fast spliced aligner with low memory requirements’, Nature Methods, 12(4), pp. 357–360. doi: 10.1038/nmeth.3317.

19. The RGASP Consortium et al. (2013) ‘Systematic evaluation of spliced alignment programs for RNA-seq data’, Nature Methods, 10(12), pp. 1185–1191. doi: 10.1038/nmeth.2722.

20. Garber, M. et al. (2011) ‘Computational methods for transcriptome annotation and quantification using RNA-seq’, Nature Methods, 8(6), pp. 469–477. doi: 10.1038/nmeth.1613.

21. Danecek, P. et al. (2021) ‘Twelve years of SAMtools and BCFtools’, GigaScience, 10(2), p. giab008. doi: 10.1093/gigascience/giab008.

22. Pertea, M. et al. (2015) ‘StringTie enables improved reconstruction of a transcriptome from RNA-seq reads’, Nature Biotechnology, 33(3), pp. 290–295. doi: 10.1038/nbt.3122.

23. Love, M. I., Huber, W. and Anders, S. (2014) ‘Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2’, Genome Biology, 15(12), p. 550. doi: 10.1186/s13059-014-0550-8.

24. Galipon, J. et al. (2018) ‘High-Quality Overlapping Paired-End Reads for the Detection of A-to-I Editing on Small RNA’, in Ørom, U. A. V. (ed.) miRNA Biogenesis. New York, NY: Springer New York (Methods in Molecular Biology), pp. 167–183. doi: 10.1007/978-1-4939-8624-8_13.

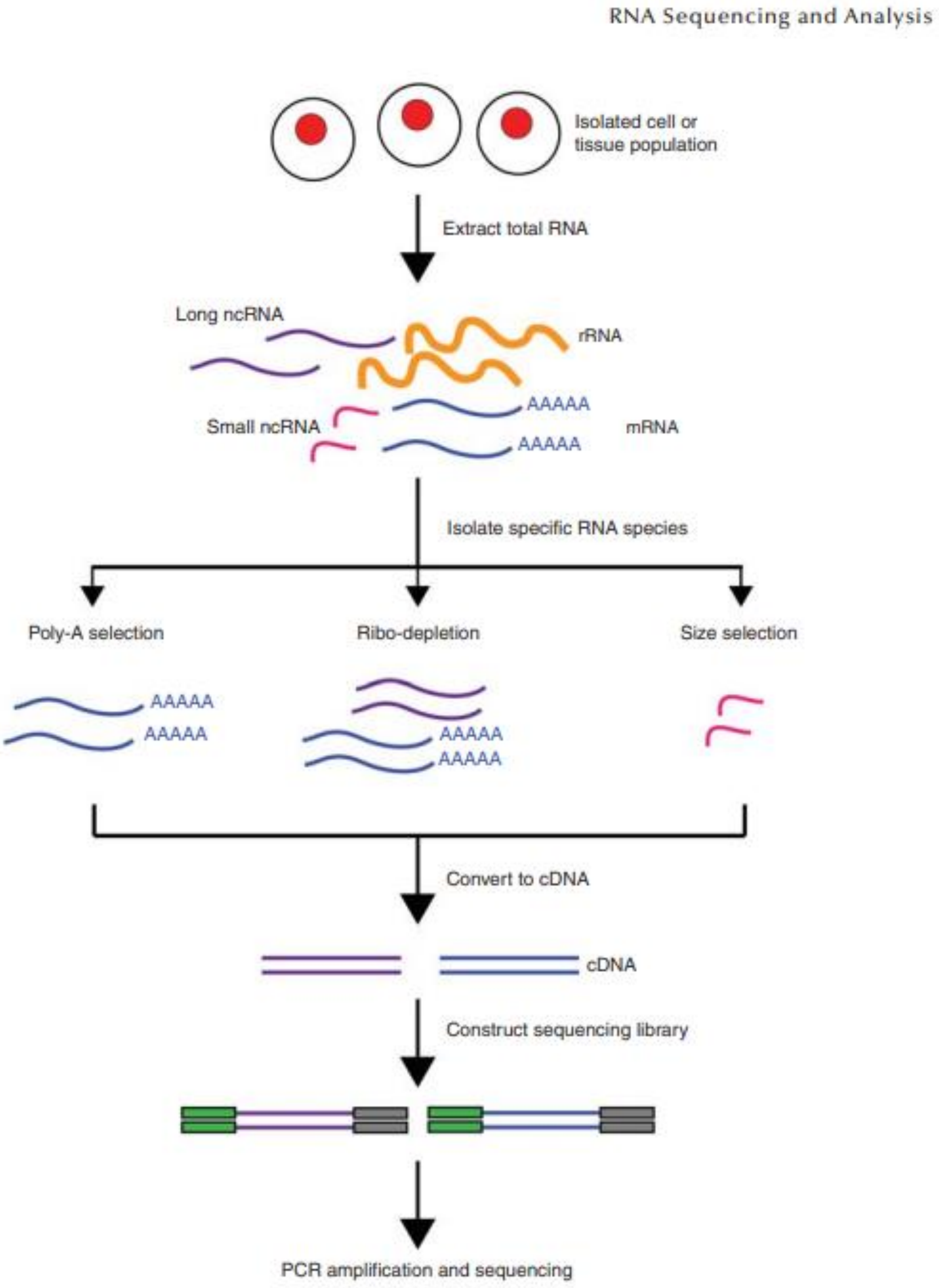
25. Kukurba, K. R. and Montgomery, S. B. (2015) ‘RNA Sequencing and Analysis’, *Cold Spring Harbor Protocols*, 2015(11), p. pdb.top084970. doi: 10.1101/pdb.top084970.

26. Araldi, R. P. et al. (2020) ‘Stem Cell-Derived Exosomes as Therapeutic Approach for Neurodegenerative Disorders: From Biology to Biotechnology’, Cells, 9(12), p. 2663. doi: 10.3390/cells9122663.

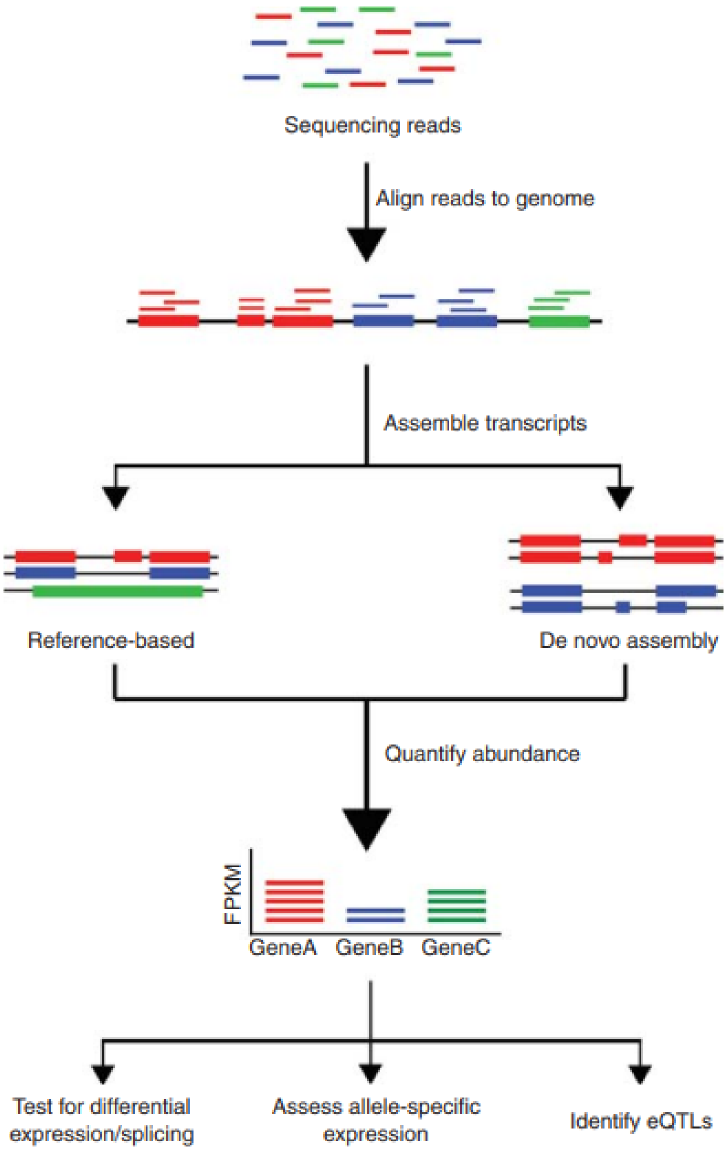
27. Mintoo, A. A. et al. (2019) ‘Draft genome of the river water buffalo’, Ecology and Evolution, 9(6), pp. 3378–3388. doi: 10.1002/ece3.4965.

28. Heather, J. M. and Chain, B. (2016) ‘The sequence of sequencers: The history of sequencing DNA’, Genomics, 107(1), pp. 1–8. doi: 10.1016/j.ygeno.2015.11.003.

ANNEXURES


RNA Sequencing and Analysis
Isolated cell or tissue population
Extract total RNA
Long ncRNA
rRNA
AAAAA
Small ncRNA
mRNA
AAAAA
Isolate specific RNA species
Poly-A selection
Ribo-depletion
Size selection
AAAAA
AAAAA
AAAAA
AAAAA
Convert to cDNA
cDNA
Construct sequencing library
PCR amplification and sequencing

9 Overview of the RNA- Seq workflow[25]

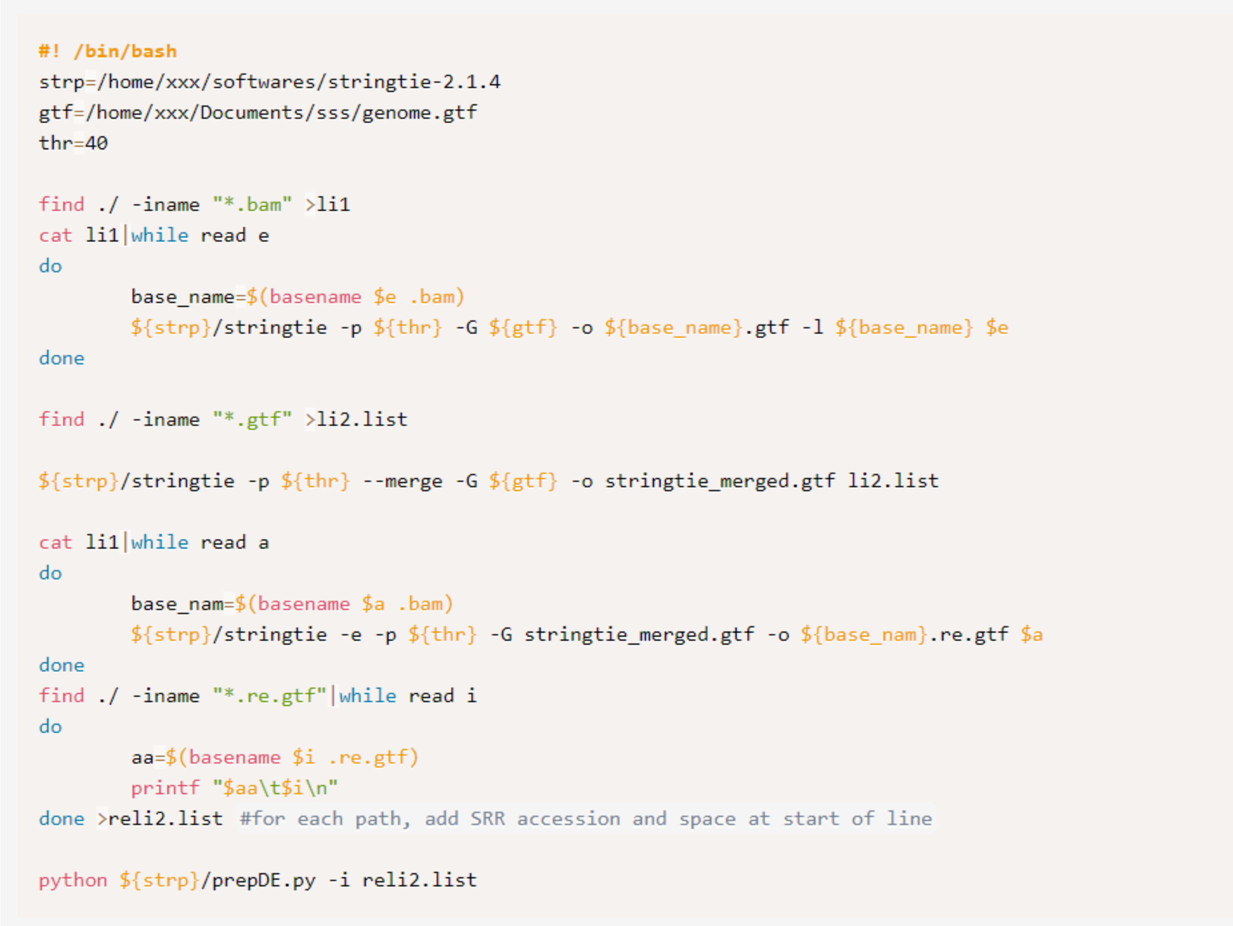


10 Generalized overview of the RNA- Seq data analysis[25]

Primary category	Tool name	Notes
Splice-aware read alignment	GEM	Filtration-based approach to approximate string matching for alignment
	GSNAP	Based on seed and extend alignment algorithm aware of complex variants
	MapSplice	Based on Burrows-Wheeler Transform (BWT) algorithm
	RUM	Integrates alignment tools Blat and Bowtie to increase accuracy
	STAR	Based on seed searching in an uncompressed suffix arrays followed by seed clustering and stitching procedure; fast but memory-intensive
	TopHat	Uses Bowtie, based on BWT, to align reads; resolves spliced reads using exons by split read mapping
Transcript assembly and quantification	Cufflinks	Assembles transcripts to reference annotations or de novo and quantifies abundance
	FluxCapacitor	Quantifies transcripts using reference annotations
	iReckon	Models novel isoforms and estimates their abundance
Differential expression (DE)	BaySeq	Count-based approach using empirical Bayesian method to estimate posterior likelihoods
	Cuffdiff2	Isoform-based approach based on beta negative binomial distribution
	DESeq	Exon-based approach using the negative binomial model
	EdgeR	Count-based approach using empirical Bayes method based on the negative binomial model

	MISO	Isoform-based model using Bayes factors to estimate posterior probabilities
Other tools	HCP	Normalizes expression data by inferring known and hidden factors with prior knowledge
	PEER	Normalizes expression data by inferring known and hidden factors using a probabilistic estimation based on the Bayesian framework
	Matrix eQTL	Fast eQTL detection tool that uses linear models (linear regression or ANOVA)

11 Widely used RNA- Seq softwares[25]



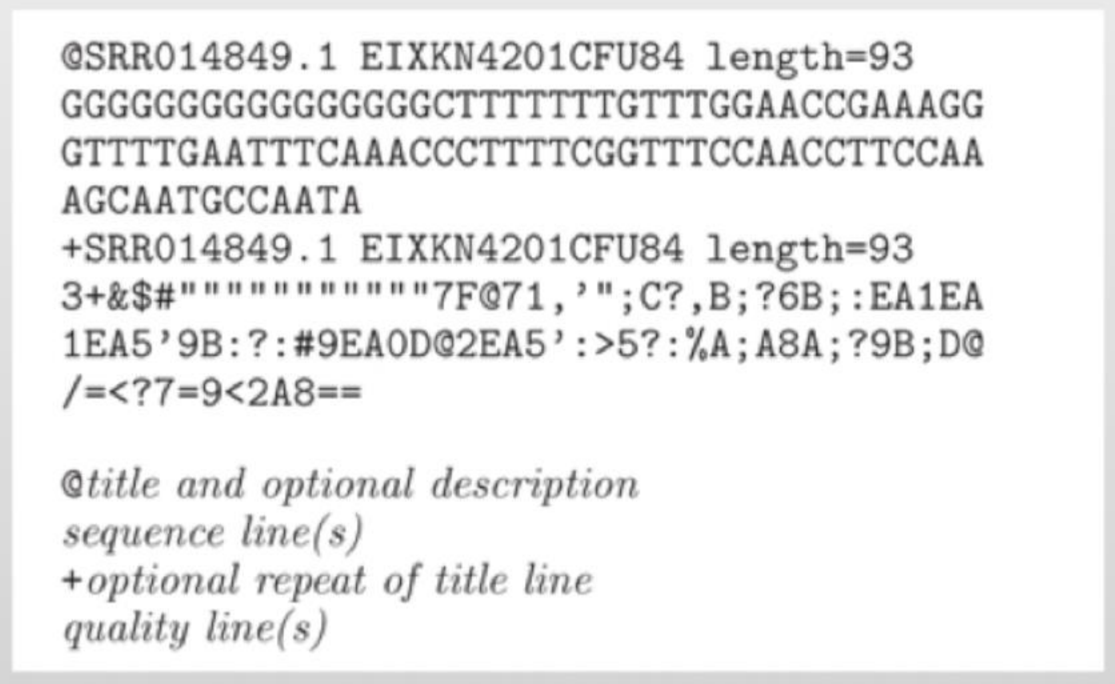
```
#! /bin/bash
strp=/home/xxx/softwares/stringtie-2.1.4
gtf=/home/xxx/Documents/sss/genome.gtf
thr=40

find ./ -iname "*.bam" >li1
cat li1|while read e
do
        base_name=$(basename $e .bam)
        ${strp}/stringtie -p ${thr} -G ${gtf} -o ${base_name}.gtf -l ${base_name} $e
done

find ./ -iname "*.gtf" >li2.list

${strp}/stringtie -p ${thr} --merge -G ${gtf} -o stringtie_merged.gtf li2.list

cat li1|while read a
do
        base_nam=$(basename $a .bam)
        ${strp}/stringtie -e -p ${thr} -G stringtie_merged.gtf -o ${base_nam}.re.gtf $a
done
find ./ -iname "*.re.gtf"|while read i
do
        aa=$(basename $i .re.gtf)
        printf "$aa\t$i\n"
done >reli2.list  #for each path, add SRR accession and space at start of line

python ${strp}/prepDE.py -i reli2.list
```

12 This code was run on the data set provided. The code was for bam files generated after performing FastP, Hisat2 and Sam tools.

R script used to identify the DEG:

```
setwd("C:/Users/Lenovo/OneDrive/Documents")
countData <- read. table ("C:/Users/Lenovo/OneDrive/Documents/gene_count_matrix.csv", header =
TRUE, sep = ",", row.names = 1)
head(countData)
colData <- read.table("phenodata2.csv", header = TRUE, sep = ",")
head(colData)

library (DESeq2)
dds <- makeExampleDESeqDataSet(m=4)
design(dds) <- formula (~ 1)
dds <- DESeqDataSetFromMatrix(countData=countData,
                              colData= colData,
                              design= ~stage_condition)
dds <- dds[rowSums(counts(dds)) > 0]
DESeq(dds)
dds <- DESeq(dds)
res <- results(dds)
head(res)
res_sig <- subset (res, padj<0.05)
res_lfc <- subset (res_sig, abs(log2FoldChange) > 2)
head(res_lfc)

openxlsx::write.xlsx(as.data.frame(res_sig), file="ressig.xlsx",rowNames=T,asTable = TRUE)


# Use the log transform on the data set
rld <- rlog(dds,blind=F)

topVarianceGenes <- head(order(rowVars(assay(rld)), decreasing=T),100)
matrix <- assay(rld)[ topVarianceGenes, ]
matrix <- matrix - rowMeans(matrix)

#Plot MA
plotMA(res)

#Plot for PCA
dds <- estimateSizeFactors(dds)
# Shifted log of normalized counts
se <- SummarizedExperiment(log2(counts(dds, normalized=TRUE) + 1),
                           colData=colData(dds))
# the call to DESeqTransform() is needed to
# Trigger our plotPCA method.
plotPCA( DESeqTransform( se ) )



#Volcano plot
library("EnhancedVolcano")
EnhancedVolcano(res,lab = rownames(res),x = 'log2FoldChange',y = 'pvalue',labSize = 0, pCutoff =
0.05,FCcutoff = 1,xlim=c(-30,12))
```

```
plot (EnhancedVolcano)

#Heatmap
library("RColorBrewer")
library('pheatmap')

variance_dds<- vst(dds, blind=FALSE)
sampleDists<-dist(t(assay(variance_dds)))
sampleDistMatrix<- as.matrix(sampleDists)
rownames(sampleDistMatrix)<- paste (variance_dds$Condition, variance_dds$sample, sep="-")
colnames(sampleDistMatrix) <- NULL
colors <- colorRampPalette(rev(brewer.pal(9, "Blues")) )(255)
pheatmap(sampleDistMatrix,
clustering_distance_rows=sampleDists,clustering_distance_cols=sampleDists, col=colors)

select <- order(rowMeans(counts(dds,normalized=TRUE)),
                decreasing=TRUE) [1:20]
df <- as.data.frame(colData(dds)[,c("sample","stage_condition")])
pheatmap(assay(dds)[select,], cluster_rows=FALSE, show_rownames=FALSE,
         cluster_cols=FALSE, annotation_col=df)
pheatmap(assay(dds)[select,], cluster_rows=FALSE, show_rownames=FALSE,
         cluster_cols=FALSE, annotation_col=df)
```



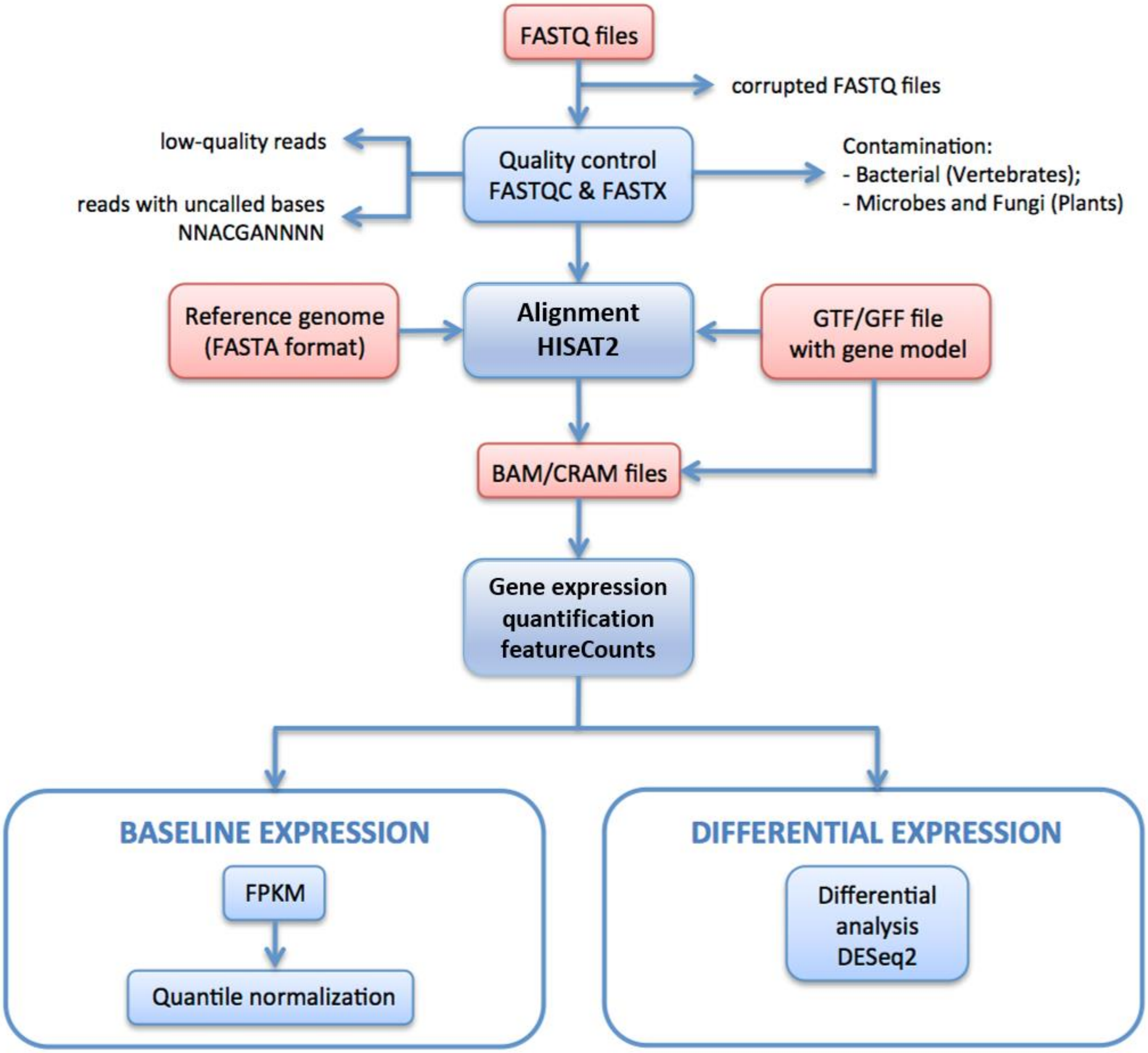
```
@SRR014849.1 EIXKN4201CFU84 length=93
GGGGGGGGGGGGGGGGCTTTTTTTGTTTGGAACCGAAAGG
GTTTTGAATTTCAAACCCTTTTCGGTTTCCAACCTTCCAA
AGCAATGCCAATA
+SRR014849.1 EIXKN4201CFU84 length=93
3+&$#"""""""""""7F@71,’";C?,B;?6B;:EA1EA
1EA5’9B:?:#9EA0D@2EA5’:>5?:%A;A8A;?9B;D@
/=<?7=9<2A8==

@title and optional description
sequence line(s)
+optional repeat of title line
quality line(s)
```

13 A FastQ format file

There are four-line types in the FastQ format:

- A ‘@’ in the title line which often holds just a record identifier.
- Nucleotide sequence of the read.
- A ‘+’ line in the signal and the end of the sequence lines and the start of the quality string.
- Quality score for each nucleotide of the read.

Read counts are affected by transcript length and total number of reads. So, we normalize the gene count to compare the expression levels. The measure **RPKM (reads per kilo base of exon model per million reads)** and its derivative **FPKM (fragments of kilo base of exon model per million reads mapped)** accounts for both the gene length and library size effects.



14 General RNA- Seq pipeline, Source: EMBL- EBI Training website

https://www.ebi.ac.uk/training/online/courses/functional-genomics-ii-common-technologies-and-data-analysis-methods/rna-sequencing/

15 iSeq from Illumina ***to perform RNA- Sequencing***

Some useful links:

https://www.thermofisher.com/in/en/home/life-science/sequencing/next-generation-sequencing.html?cid=csd_ngs_sbu_r01_co_cp1445_pjt7410_csd00000_0se_bng_nt_con_ngsphrillm&s_kwcid=AL!3652!3!!p!!o!!Illumina%20gene%20sequencing&ef_id=0d7da571b436125c9ccac595b0303ee6:G:s&s_kwcid=AL!3652!10!76759803869654!76759857929789

https://www.illumina.com/techniques/sequencing/rna-sequencing.html

https://www.ebi.ac.uk/training/online/courses/functional-genomics-ii-common-technologies-and-data-analysis-methods/rna-sequencing/

https://www.scopus.com/authid/detail.uri?authorId=56499245700

https://pkgs.org/download/fastp

http://daehwankimlab.github.io/hisat2/manual/

https://ccb.jhu.edu/software/stringtie/

http://www.bioconductor.org/packages/release/bioc/html/DESeq2.html