**Comparing Salmon, Kallisto and CLC Genomic Workbench for RNA-Seq Data Analysis**

STAT 535 (Applied Bioinformatics)

Spring 2018 Final assignment

Prepared by:

Lok Raj Joshi

Samitinjaya Dhakal

Tirth Uprety

**1.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]. 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.

RNA-Seq have been widely used to study viral disease pathogenesis. Zika virus, which is a single stranded RNA virus infects human beings and has been associated with the congenital defects. There are several cases of zika virus associated microcephaly, a condition where babies are born with smaller head (brain), in babies born from Zika virus infected mother. There are reports which say that babies born from the Zika infected mother may show nervous symptoms at later stage of life too. However, due to lack of a good animal a lot of questions about Zika pathogenesis and disease dynamics remains unanswered. Recently Darbelly et al (2017) have tried to develop a pig model for Zika virus infection. They infected pig fetus with Zika virus and studied the disease dynamics. They have deposited reads from their RNA-Seq experiments in NCBI database. Here, we are analyzing that RNA-Seq data using various methods.

**Objective:**

1. Understating gene expression pattern in Zika virus infected piglets
2. Comparing different RNA-Seq data analysis pipeline (Kallisto, Salmon, CLC)

**2. Methods**

***2.1 Experiment method:***

Asian-lineage ZIKV strain PRVABC59 (GenBank: KU501215) was inoculated (supernatant collected after culturing in Vero cells) intra uterine into porcine fetuses. Virus was inoculated Intracerebrally (Zika\_IC), Intraperitoneally and Intra-amniotically  (Zika\_IP) at 50th gestation day. 3 mock treated conceptuses and 3 surgery control (supernatant from normal cells) were used as control. 5 log10 TCID 50 Virus titer used for experimental inoculation.

The objective of the experiment was to identify if ZIKV infection is possible through IP and IC route. Second objective was to test if the offspring born with asymptomatic ZIKV develops mental illness.

For RNA-seq, total RNA was extracted using TRIzol lysis and extraction and cleaned up using Norogen RNA purification kit. Bioanalyzer was used to assess RNA quality and all extracted RNA had RNA integrity number (RIN) of 8.5 or above.  RIN is used to assess the integrity of RNA extracted and is a ratio of 28s to 18s. RIN score ranges from 1-10 with 10 being least degraded RNA. Sequencing library was prepared using Illumina RIBO Gold Kit and TruSeq Stranded Total RNA Library Prep Kit. NextSeq 500 was used to sequence libraries which generated around 30 million paired end reads per sample. Adaptor sequences were trimmed from FASTQ files and Cutadapt was used to filter low quality reads.

***2.2 Analysis Method:***

***Data:*** All the reads were downloaded from the NCBI short read archive (SRA) database for the project [PRJNA407675](https://www.ncbi.nlm.nih.gov/bioproject/PRJNA407675). There were 12 .srr files (SRR6047326-SRR6047337). They were downloaded using SRA Tookit v2.3.5. All the .srr files were splitted into paired .fastq files using SRA Tooklit. The commands used are provided in the index. Read quality was assessed using FastQC tool. Low quality reads and adapter were filtered using btrim. The fastq files obtained after quality trimming were analysed by three methods as given below. Notably, for CLC all the above-mentioned steps from downloading raw files from SRA to quality trimming were done within CLC.

***CLC***: CLC genomic workbench v 11.0.1 was used. Reference genome was downloaded through CLC for Sus scorfa. The fastq files were mapped with the reference genome and the RNA seq analysis was run against the reference genome and reference mRNA. CLC uses progressive alignment technique. All genes are extracted from the reference genome (using a gene track). Then, all annotated transcripts are extracted (using an mRNA track). If there are several annotated splice variants, they are also extracted.

Firstly, for alignments, CLC constructs a multiple alignment of a number of homologs DNA or protein sequences. These algorithms employ a scoring function which incorporates the underlying phylogeny. Then, uses an explicit stochastic model of molecular evolution which makes it possible to compare different types of solutions in a statistically rigorous way.

***Salmon:*** Salmon v 0.8.1 was used for the transcript counts. First of all, an index was created for Sus scorfa (pig) pig transcriptome (Emsembl : Sus scorfa annotation 10.2). Then the paired files for each samples were aligned to the indexed pig genome. No of reads data from the output files were combined into a single matrix for visualization.

***Kallisto :*** Kallisto v 0.44.0 was used to obtain abundance of transcripts. It is based on the idea of pseudo alignment of reads with targets. The reads obtained from RNA sequencing was pseudoaligned with pig reference genome obtained from Ensembl: Sus scorfa 10.2. The read counts of all the samples were combined in one matrix for visualization.

***Data visualization:*** For CLC, all the data visualizations were done natively in CLC whereas iDEP v 0.73 was used for the visualization of data obtained from Kallisto and Salmon.

**3. Results:**

**3.1 Data quality:** Number of reads present in each sample is shown in Figure 1A. All the samples have almost similar number of reads except one sample (Zika\_IC\_2). The quality of the reads after reads filtering and trimming is shown in 1B. All the reads having quality score of 25 were trimmed. Only the representative figure of one sample is shown here in 1B. Principal component analysis shows that all the groups are separated from each other Fig 1C. Samples from Zika\_IC group are closer to the control whereas samples from Zika\_IP cluster far from control group. Similar kind of PCA plots were obtained from both iDEP and CLC.

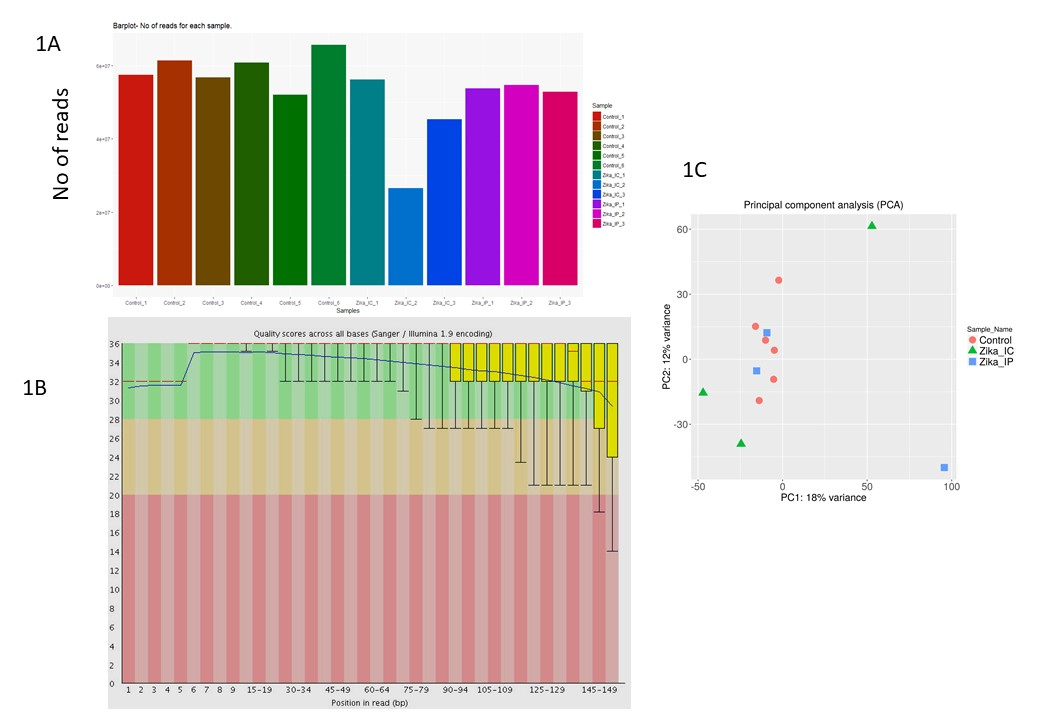


Figure 1. Data quality and filtering and principal component analysis

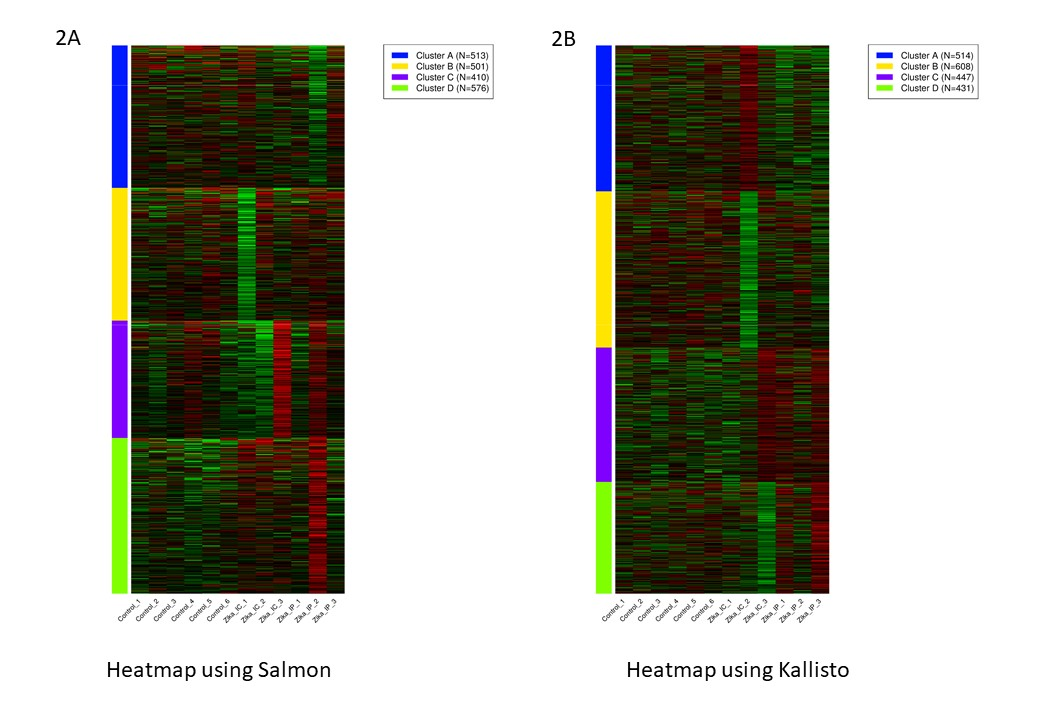
***3.2 Differential gene expression in Zika infected piglets***

Authors reported dysregulation of more than 669 genes at 0.05 FDR based on their analysis. They also stated that ZIKA\_IC group had relatively weaker gene expression for genes associated with Virus response, immune system, interferon and stress response compared to IP+IA. Our analysis both from kallisto and salmon shows upregulation of genes associated with biological pathways as mentioned in the paper. When comparing between Kallisto and Salmon, there is difference in terms of number of genes being upregulated or downregulated which could be due to difference in cut-offs and different FDR applied. One of the interesting observation as seen in the heat map is that controls are also showing up-regulation of certain genes which is similar to experimental ones. One of the possible explanation as mentioned in the paper is that it could be result of fetus to fetus transmission. This means even non- manipulated fetus acquired infection via experimentally infected ones.

The heat-map generated from all 3 methods is shown figure 2. When compared between three methods, Salmon showed that genes in cluster D and cluster C were up-regulated for ZIKA\_IP\_2 whereas ZIKA\_IP\_2 had cluster C and D genes upregulated for Kallisto. CLC heatmap showed that ZIKA\_IP\_3 had genes upregulated in clusters. Heat map for Salmon and Kallisto data showed cluster C genes upregulated in ZIKA\_IC\_3. Number of genes upregulated in each method are shown as Venn-diagram in figure 3. Number of genes that were upregulated is listed below:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Salmon | Kallisto | CLC |
| ZIKA\_IP-Control only | 9 | 139 | 79 |
| ZIKA\_IC-Control only | 38 | 53 | 5 |
| ZIKA\_IP-ZIKA\_IC | 15 | 46 | 30 |
| ZIKA\_IP-ZIKA\_IC-Control | 0 | 1 | 3 |

Some of the up regulated genes are associated with biological processes like defense response to virus, innate immune response, immune effector process, response to other biotic stimuli, negative regulation of of viral process, Response to stress, multicellular organism process etc. Salmon showed both the pathway related to viral infection and organism development. Whereas, Kallisto and CLC only showed the enrichment of the pathway related to nervous system and organism development. It is noteworthy to mention that Zika virus infects brain of the off-springs and severely impairs nervous system development. That is the reason behind birth of babies with microcephaly from the Zika virus infected pregnant mother. It can be seen that not only the genes that are responsible for nervous system are being affected but the genes related to other systems like cardiovascular system are also being affected by the Zika virus. It is surprising that both Kallistio and CLC failed to identify the genes related to virus infection. Some of the major biological pathways associated with enriched genes are shown in Table 2.



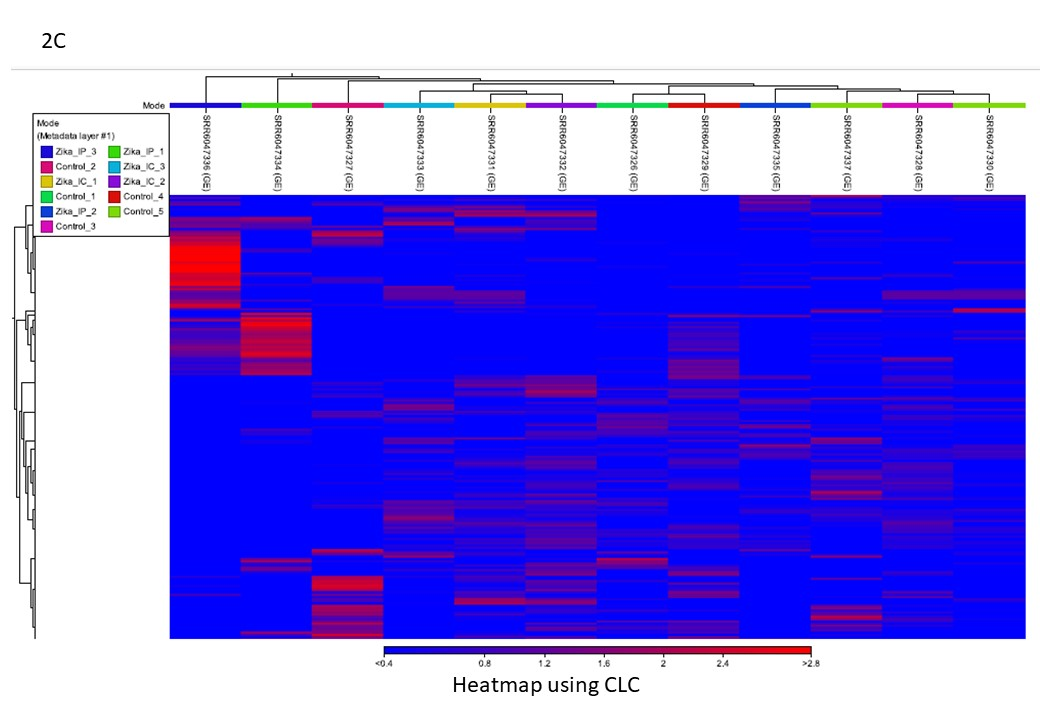
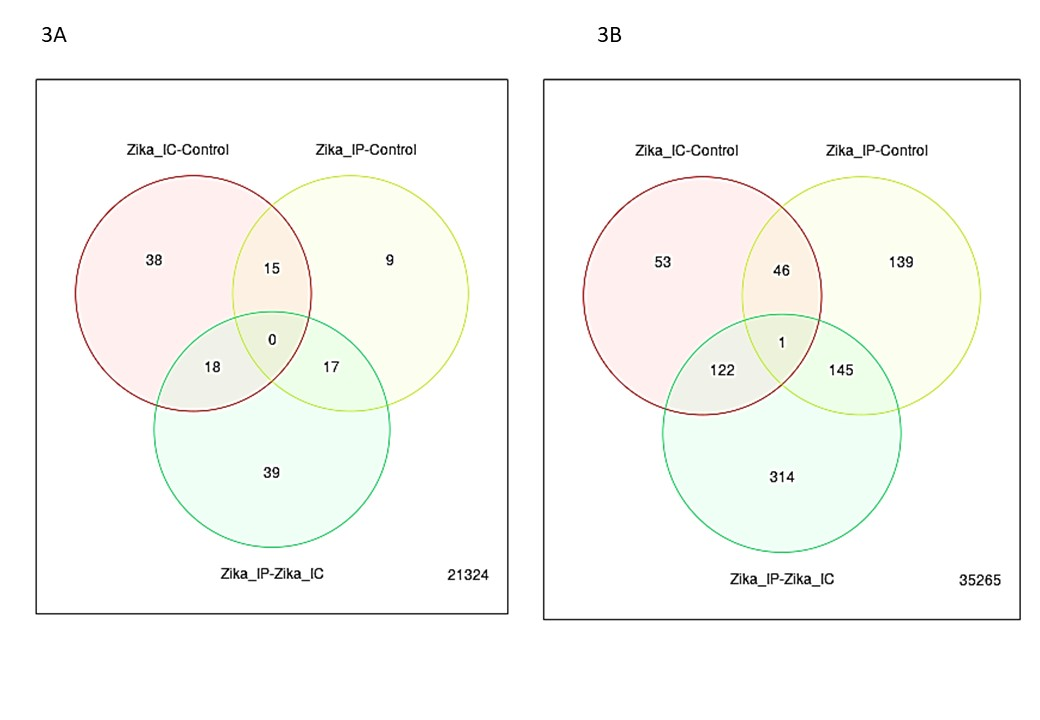


Figure 2. Heatmap generated by using three different tools



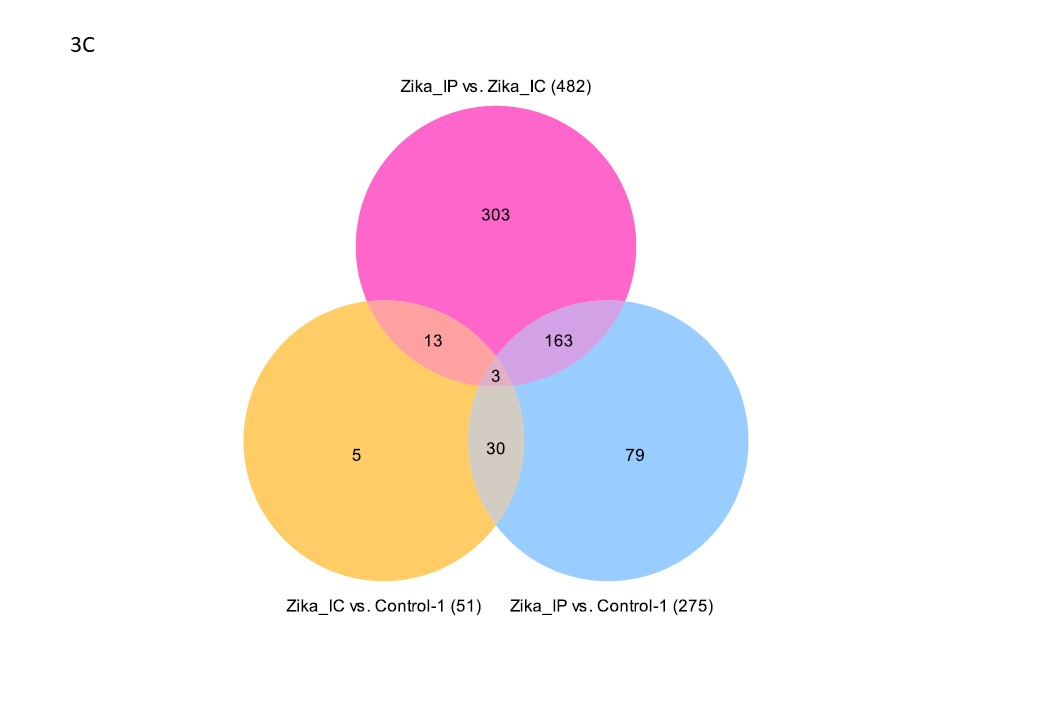


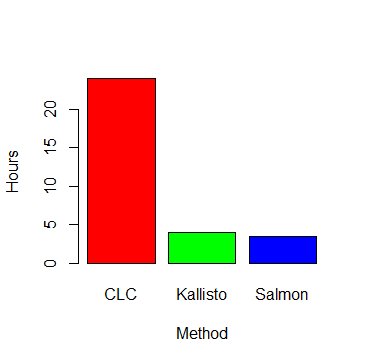
Figure 3. Venn-diagrams of the upregulated genes from different software. 3A- Salmon , 3B- Kallisto , 3C – CLC

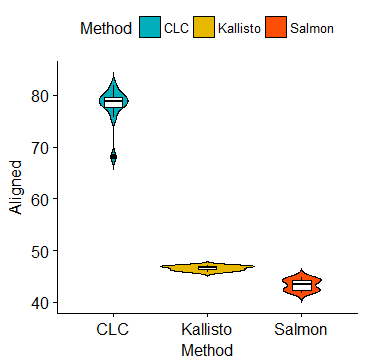
|  |  |  |
| --- | --- | --- |
| Salmon | Kallisto | CLC |
| Defense response to virus | Cellular component organization | Extracellular matrix structural component |
| Innate immune response | Organelle organization | Nervous system development |
| Response to biotic stimulus | Cell projection organization | Regulation of multicellular organism process |
| Negative regulation of viral life cycle | Macromolecule localization | Protein complex biogenesis |
| Immune response | MiRNA loading onto RISC involved in gene silencing by miRNA | Oxygen transporter activity |
| Multicellular organismal process | Regulation of transport | Blood vessel development |
| Multicellular organismal process | Response to light stimulus | Single organism process |
| Nervous system development | Optic nerve development | Neurogenesis development |
| Wound healing | Regulation of localization | Localization |
| System development | Wound healing | Wound healing |

Table 2. Pathway analysis as given by different methods

**3.3 Comparison of three methods:**

The number reads aligned to the reference genome was compared among three tools. The percentage alignment was the highest for the CLC, which aligned 78% reads on an average ranging from 68% to 81% (Fig 4A). Salmon and Kallisto had lower alignment percentage than CLC. Kallisto aligned 46.5% reads while Salmon aligned 43.3 % reads (Fig 4A). Using ANOVA and Tukey HSD test we found significant difference in alignment among all three methods (P<0.05).

Time consumed for the analysis was compared among three groups. CLC took about 24 hours to complete all the analysis while Kallisto and Salmon took 4 hours and 3.5 hours respectively (Fig 4B).

Figure 4- Comparison of alignment percentage between three methods and time taken by the software for analysis by three methods.

1. Discussion:

Although there are differences in number of genes that are shown to be up-regulated in 3 different methods used, the overall gene enriched pathways are same. Kallisto and Salmon showed similar data. It may be due to use of iDEP.70 for analysis and visualization of data. The differences in three methods could be result of differences in analysis settings and cut-off score that resulted in variation in number of genes upregulated among 3 methods.

1. Conclusion: Kallisto and Salmon have similar speed and almost similar output probably owing to similar principle of sequence alignment (pseudo alignment for Kallisto and Quasi mapping for Salmon). It seems Salmon has slightly better speed than Kallisto and bioinformatician claim Salmon to be more accurate [6]. Although Salmon and Kallisto have somewhat similar results, yet there are no methods that will give same results.  Equally important to note is that we used ETS counts to analyse data and this could have resulted in some more variation between Kallisto and Salmon data. When using TPM it has been observed that both Salmon and Kallisto data overlap considerably probably because TPM is used to filter transcripts by some thresholds. CLC took longer time than Kallisto and Salmon.

One inference that can be drawn from our analysis is that none of the methods used in RNA-seq analysis yields exactly same results. From our analysis it seems Kallisto and SAlmon are better in terms of speed and accuracy.

Table 1. Comparison of pathway analysis

**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.

6. Patro, R., et al. (2016). "Salmon provides accurate, fast, and bias-aware transcript expression estimates using dual-phase inference." bioRxiv.

7. Darbellay, J., et al. (2017). "Zika Virus Causes Persistent Infection in Porcine Conceptuses and may Impair Health in Offspring." EBioMedicine 25: 73-86.

8. Nicolas L Bray, Harold Pimentel, Páll Melsted and Lior Pachter, [Near-optimal probabilistic RNA-seq quantification](http://www.nature.com/nbt/journal/v34/n5/full/nbt.3519.html), Nature Biotechnology **34**, 525–527 (2016), doi:10.1038/nbt.3519

**Index :**

**Commands Used:**

Commands for first samples shown. Same commands were repeated using ***pbs script in blackjack*** cluster.

**1) Fastqc**

fastqc -o /home/joshil/ngs/rnaseq/qualitycheck    /home/joshil/ngs/rnaseq/SRR6047334

**2) Trimming :**

btrim64-static -q -t SRR6047326\_1.fastq -o btrimSRR6047326\_1.fastq

**3) Salmon**

1. ***make index files***

./salmon index -t /home/joshil/ngs/rnaseq/Sus\_scorfa.Sscorfa10.2.cdna.all.fa -i pig\_index

(here pig\_index is the index file we gave. this is used for alignment later)

**2) *Run psuedoalignemnt***

./salmon quant -i pig\_index -l A -1 /home/joshil/ngs/rnaseq/SRR6047326\_1.fastq -2 /home/joshil/ngs/rnaseq/SRR6047326\_2.fastq -p 10 -o qunat326

**4) Kallisto**

***1) make index files***

kallisto index -i pig\_index /home/joshil/ngs/rnaseq/Sus\_scorfa.Sscorfa10.2.cdna.all.fa

***2) Run psuedoalignemnt***

kallisto quant -i pig\_index -o qunat326 -1 /home/joshil/ngs/rnaseq/SRR6047326\_1.fastq -2 /home/joshil/ngs/rnaseq/SRR6047326\_2.fastq -p 10