A Guide to Differential RNA-seq Analysis

Nikolas Tsiouplis
Departments of Biology and Genetics, University of North Carolina at Chapel Hill

Introduction

High-throughput, whole-genome approaches to biological questions have been increasingly utilized in biomedical research and other fields as cost has decreased and computing power has increased[1,2]. The uptake and impact has been immense, as researchers and clinicians incorporate genomic data into cancer diagnosis, population health, genetic disease, and personalized medicine[2]. With widespread utility and adoption, many scientists without a background in informatics or computational biology are being tasked with interpreting and analyzing functional genomics assays. This poster aims to provide an accessible and detailed guide into devising, conducting, and analyzing an RNA-sequencing experiment based on experiences from my own research.

RNA-sequencing is a functional genomics experiment in which all the mRNA in a sample are converted to a cDNA library, fragmented to approximately 150-200 base pairs, and ligated with adapter sequences[3]. These cDNA fragments can then be sequenced on a short-read, next-generation sequencing platform like those manufactured by Illumina. Each platform employs a slightly different technique, but the underlying principles are the same. Individual fragments bind to oligos attached to small compartments called flow cells. PCR amplification results in hundreds of millions of copies of each fragment. DNA polymerase adds a modified, complementary base to the chain, and the base is recorded with a fluorescent camera or by other means. The output of these platforms is a large (several GBs) dataset of reads with associated quality scores[3]. This is the point where many biologists find themselves unsure how to proceed. The following pipeline offers a step-by-step protocol for turning RNA-sequencing data into a full differential analysis.

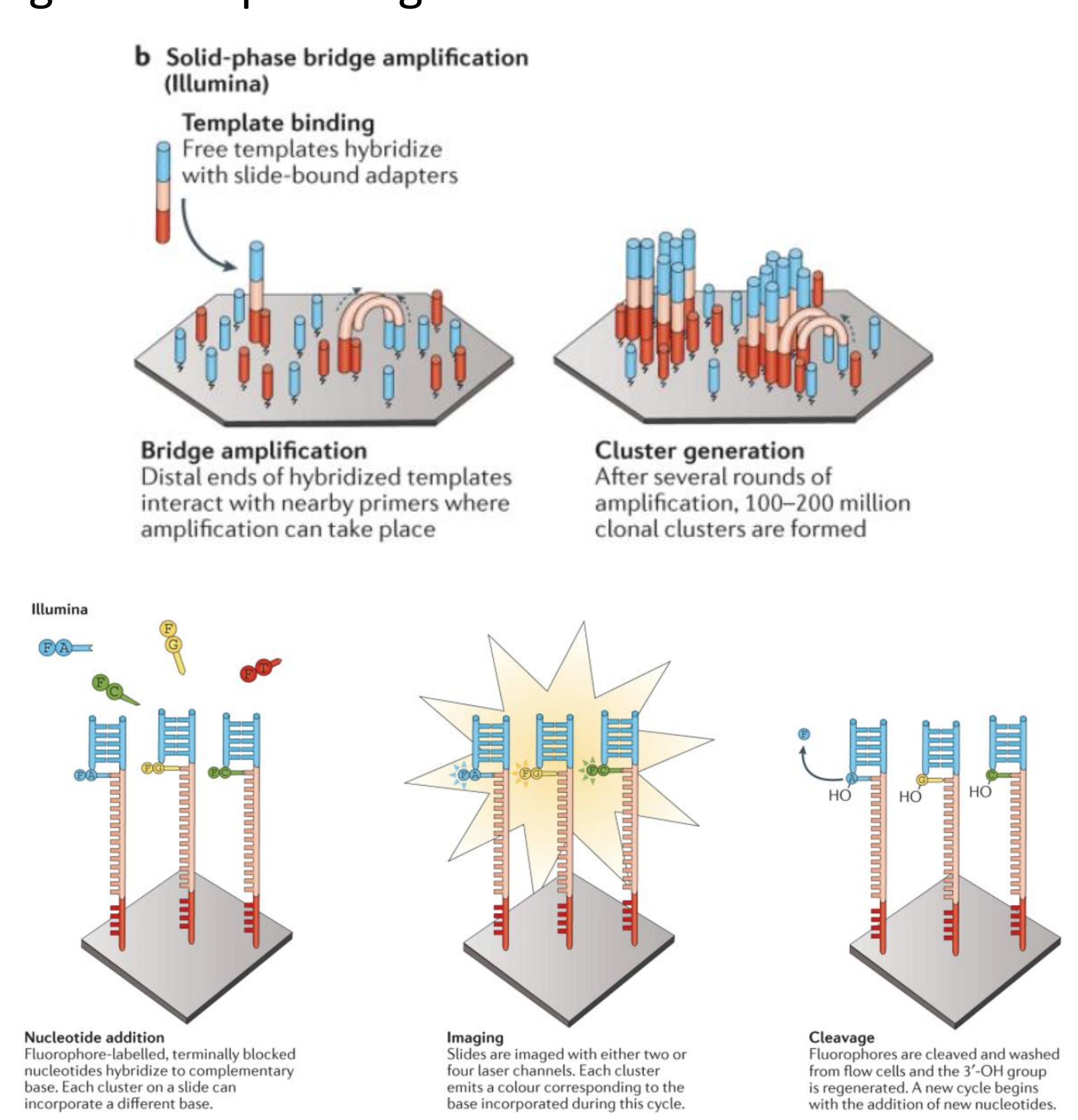

Figure 1. Illumina Method of Next-Generation Sequencing[4]

Pipeline

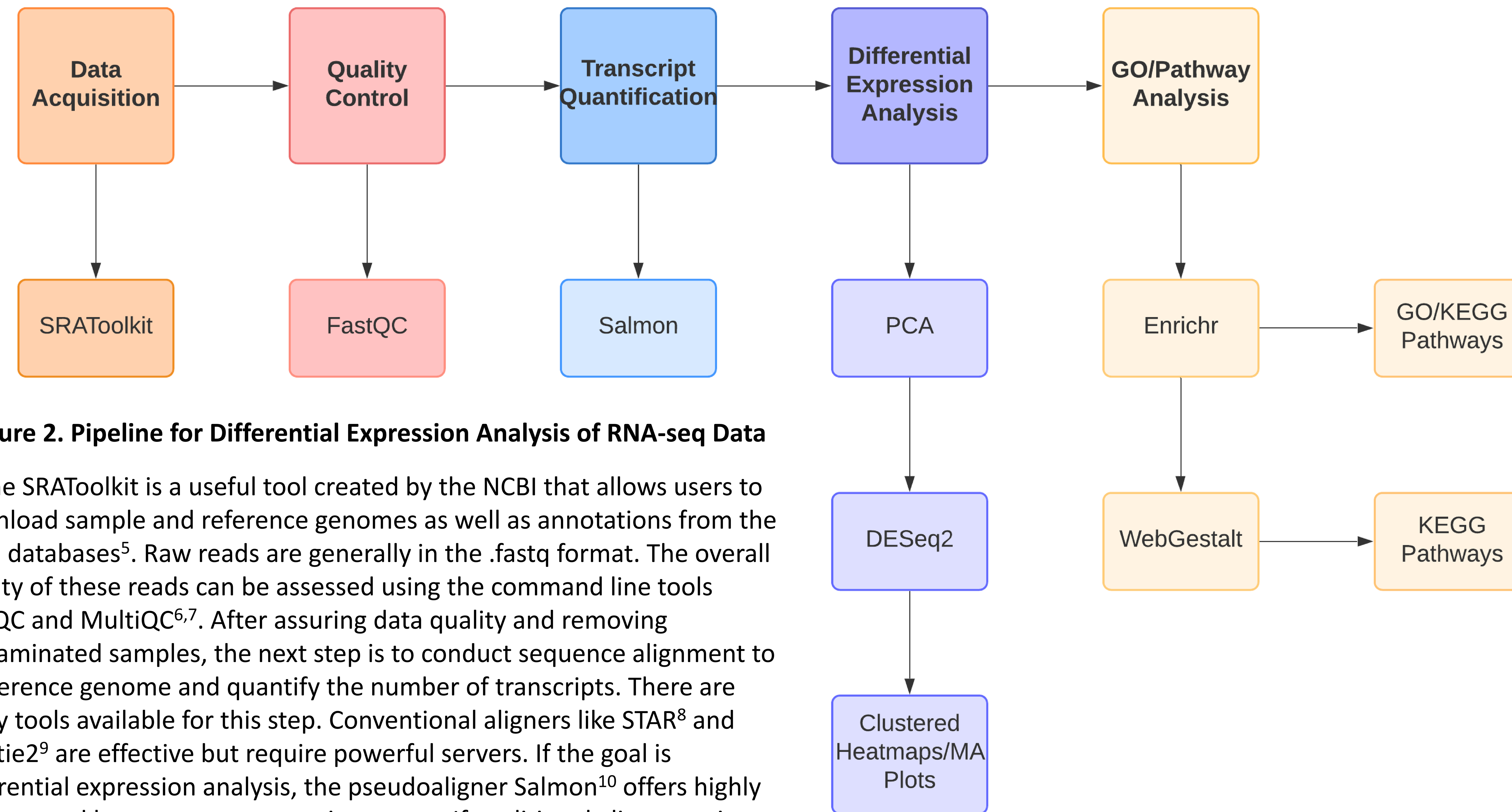

Figure 2. Pipeline for Differential Expression Analysis of RNA-seq Data

The SRAToolkit is a useful tool created by the NCBI that allows users to download sample and reference genomes as well as annotations from the NCBI databases[5]. Raw reads are generally in the .fastq format. The overall quality of these reads can be assessed using the command line tools FastQC and MultiQC[6,7]. After assuring data quality and removing contaminated samples, the next step is to conduct sequence alignment to a reference genome and quantify the number of transcripts. There are many tools available for this step. Conventional aligners like STAR[8] and Bowtie2[9] are effective but require powerful servers. If the goal is differential expression analysis, the pseudoaligner Salmon[10] offers highly efficient and low-compute transcript counts. If traditional alignment is used, another transcript quantifier like HTseq[11] is necessary to proceed further into the workflow.

After transcripts have been identified and counted, the next step is to conduct a differential gene expression analysis. A common and effective package for this is DESeq2[12], available in the programming language R. Importing the transcripts into R will allow you to set up comparisons between samples. Running DESeq2 on each comparison will generate an output of significantly differentially expressed genes. This is the key to the whole process; these significant genes explain the differences between individual conditions or genotypes. For a big-picture understanding of these genes, heatmaps and MA plots can be generated within R. Furthermore, gene ontology and pathway analysis can describe the enrichment of certain characteristics or pathways in the set of significant genes. Two useful web-based tools for this are Enrichr[13] and WebGestalt[14]. The former uses gene names to quickly calculate enrichment, while the latter accounts for gene-level statistics calculated from DESeq2. Following these steps will provide insight into the differences between samples, genotypes, or treatments of interest. The list of significant genes, gene ontologies, and pathways offer new avenues of exploration and can generate new hypotheses for wet lab experiments.

Acknowledgements

I want to thank Dr. Ageliki Tsagaratou, my PI, as well as the rest of the team in the Tsagaratou Lab for their help and support. I also want to thank Dr. Tarmo Aijo for his help in computational methodology. I would also like to thank Dr. Furey for his instruction in functional genomics analysis.

References

1. Epidemiology: From Familial Aggregation to Genomic Sequencing. *Am J Epidemiol*. 2019;188(12):2069-2077. doi:10.1093/aje/kwz193
2. Mattick JS, Dziadek MA, Terrill BN, et al. The impact of genomics on the future of medicine and health. *Med J Aust*. 2014;201(1):17-20. doi:10.5694/mja13.10920
3. Kukurba KR, Montgomery SB. RNA sequencing and analysis. *Cold Spring Harb Protoc*. 2015;2015(11):951-969. doi:10.1101/pdb.top084970
4. Goodwin S, McPherson JD, McCombie WR. Coming of age: Ten years of next-generation sequencing technologies. *Nat Rev Genet*. 2016;17(6):333-351. doi:10.1038/nrg.2016.49
5. Team STD. SRA Toolkit. Published online 2020. http://ncbi.github.io/sra-tools/
6. Andrews S. FastQC. Published online 2010. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
7. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*. 2016;32(19):3047-3048. doi:10.1093/bioinformatics/btw354
8. Dobin A, Davis CA, Schlesinger F, et al. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013;29(1):15-21. doi:10.1093/bioinformatics/bts635
9. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods*. 2012;9(4):357-359. doi:10.1038/nmeth.1923
10. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods*. 2017;14(4):417-419. doi:10.1038/nmeth.4197
11. Anders S, Pyl PT, Huber W. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics*. 2015;31(2):166-169. doi:10.1093/bioinformatics/btu638
12. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 2014;15(12):550. doi:10.1186/s13059-014-0550-8
13. Chen EY, Tan CM, Kou Y, et al. Enrichr: Interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics*. 2013;14(1):128. doi:10.1186/1471-2105-14-128
14. Liao Y, Wang J, Jaehnig EJ, Shi Z, Zhang B. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. *Nucleic Acids Res*. 2019;47(W1):W199-W205. doi:10.1093/nar/gkz401