UGWU PASCHAL

+2348166207095 | ugwupaschal@gmail.com

https://www.linkedin.com/in/paschal-ugwu-52abb6229/

Melbourne Bioinformatics

Topic: RNA-Seq differential gene expression (DGE) analysis using datasets from common fruit fly, *Drosophila melanogaster*.

August, 2020

Abstract

RNA sequencing (RNA-Seq) makes use of high-throughput sequencing techniques to shed light on a cell's transcriptome. It is a commonly used technique for examining the activity of genes under various biological circumstances and a potential new technology for precisely monitoring gene expression levels. The goal of this project was to use RNA-Seq to measure the amount of mRNA in several samples of the *Drosophila melanogaster* and statistically evaluate the variations in expression per-gene between the samples using differential gene expression analyses. I carried out the following tasks: QC (Quality Control) of the raw sequence data, trimming of the RNA reads for amount and adapter sequence, QC for the RNA-Seq alignment data, alignment of quality RNA reads, visualization of the RNA-Seq alignment data with IGV and JBrowse, and discovery of differentially expressed genes in the samples. Trim Galore, HISAT, HTSEQ, Deseq2, and FeatureCounts were the tools I utilized. By testing for differential expression analysis for samples of *Drosophila melanogaster*, this study has given me the opportunity to comprehend the fundamental workflow of alignment quantification.

Background

RNA sequencing (RNA-Seq) uses the capabilities of high-throughput sequencing methods to provide insight into the transcriptome of a cell. Beyond quantifying gene expression, the data generated by RNA-Seq facilitate the discovery of novel transcripts, identification of alternatively spliced genes, and detection of allele-specific expression (Kukurba & Montgomery, 2015).

Also, RNA-Seq is a promising new technology for accurately measuring gene expression levels. Expression estimation with RNA-Seq requires the mapping of relatively short sequencing reads to a reference genome or transcript set. Because reads are generally shorter than transcripts from which they are derived, a single read may map to multiple genes and isoforms, complicating expression analyses (Li *et al., 2010).*

Furthermore, RNA-Seq is a widely used method for studying the behavior of genes under different biological conditions. An essential step in an RNA-Seq study is normalization, in which raw data are adjusted to account for factors that prevent direct comparison of expression measures. Errors in normalization can have a significant impact on downstream analysis, such as inflated false positives in differential expression analysis. An underemphasized feature of normalization is the assumptions on which the methods rely and how the validity of these assumptions can have a substantial impact on the performance of the methods (Evans *et al.,* 2018).

Noise in gene expression is a main determinant of phenotypic variability. Increasing experimental evidence suggests that genome-wide cellular constraints largely contribute to the heterogeneity observed in gene products. It is still unclear, however, which global factors affect gene expression noise and to what extent. Since eukaryotic gene expression is an energy demanding process, differences in the energy budget of each cell could determine gene expression differences (Guantes *et al.,* 2015).

High-throughput data production has revolutionized molecular biology. However, massive increases in data generation capacity require analysis approaches that are more sophisticated, and often very computationally intensive. Thus, making sense of high-throughput data requires informatics support. Galaxy (http://galaxyproject.org) is a software system that provides this support through a framework that gives experimentalists simple interfaces to powerful tools, while automatically managing the computational details. Galaxy is distributed both as a publicly available Web service, which provides tools for the analysis of genomic, comparative genomic, and functional genomic data, or a downloadable package that can be deployed in individual laboratories. Either way, it allows experimentalists without informatics or programming expertise to perform complex large-scale analysis with just a Web browser (Blankenberg *et al.,* 2010).

Aim of Project

To use differential gene expression studies to exploit RNA-Seq to quantify the amount of mRNA in different samples of *Drosophila melanogaster* and statistically test the differences in expression per-gene between the samples.

Objectives of Project

1. QC (Quality Control) of the raw sequence data.
2. Trim RNA-reads for quantity and for adapter sequence.
3. QC for RNA-Seq alignment data.
4. Alignment of quality RNA-reads.
5. Visualize RNA-Seq alignment data with IGV and JBrowse.
6. Find differentially expressed genes.

Method

In this project I used a subset of the data from published experiment by Hateley *et al*. in 2016. In practice, full-sized dataset would be much larger and take longer to run.

The sequence data I worked on was from *Drosophila melanogaster* pupae. The experiment had two (2) conditions, g3 where pupae underwent development in three times Earth's gravity (i.e. 3g), and g1, the control, where pupae developed in the standard gravitational acceleration felt on the surface of the surface of the Earth (i.e. 1g).

There were three (3) samples in each conditions and the sequencing data is paired-end so I had two files for each of the six (6) samples. My aim was to find differentially expressed genes in g1 vs g3.

I created a pipeline on Galaxy Europe software with the following bioinformatics tools:

1. Trim Galore: was used to automate quality and adapter trimming as well as quality control.
2. HISAT: provided several alignment strategies specifically designed for mapping different types of RNA-seq reads. All these together, HISAT enabled extremely fast and sensitive alignment of reads, in particular those spanning two exons or more. HISAT was selected because it uses the Bowtie2 implementation to handle most of the operations. In addition to

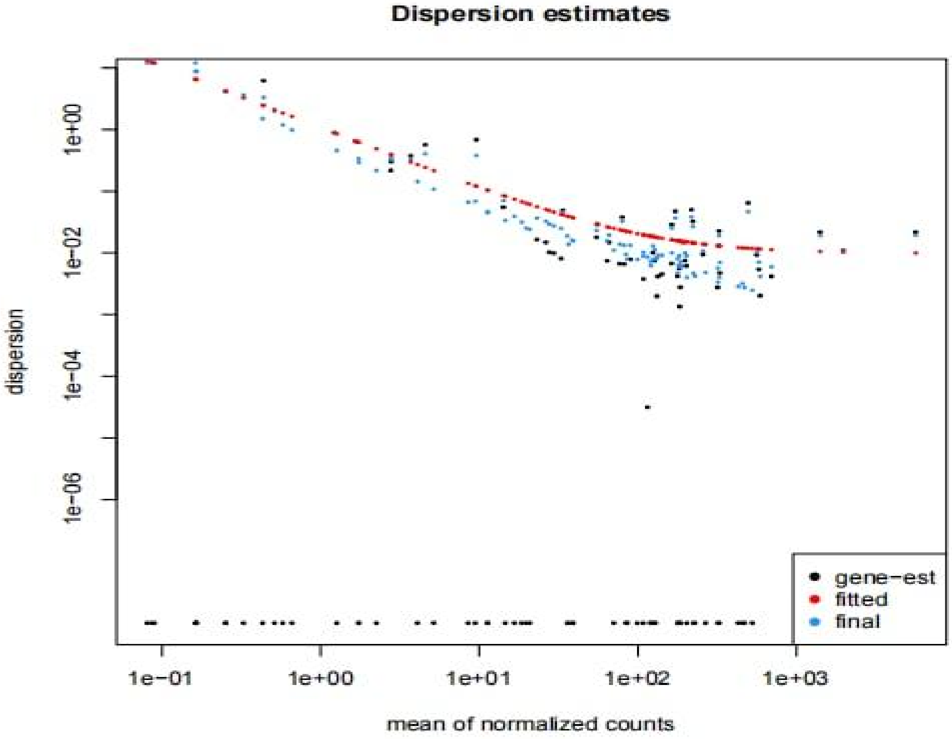
spliced alignment, HISAT handles reads involving indels and supports a paired-end alignment mode. Multiple processors can be used simultaneously to achieve greater alignment speed. HISAT outputs alignments in SAM format, enabling interoperation with a large number of other tools that use SAM.

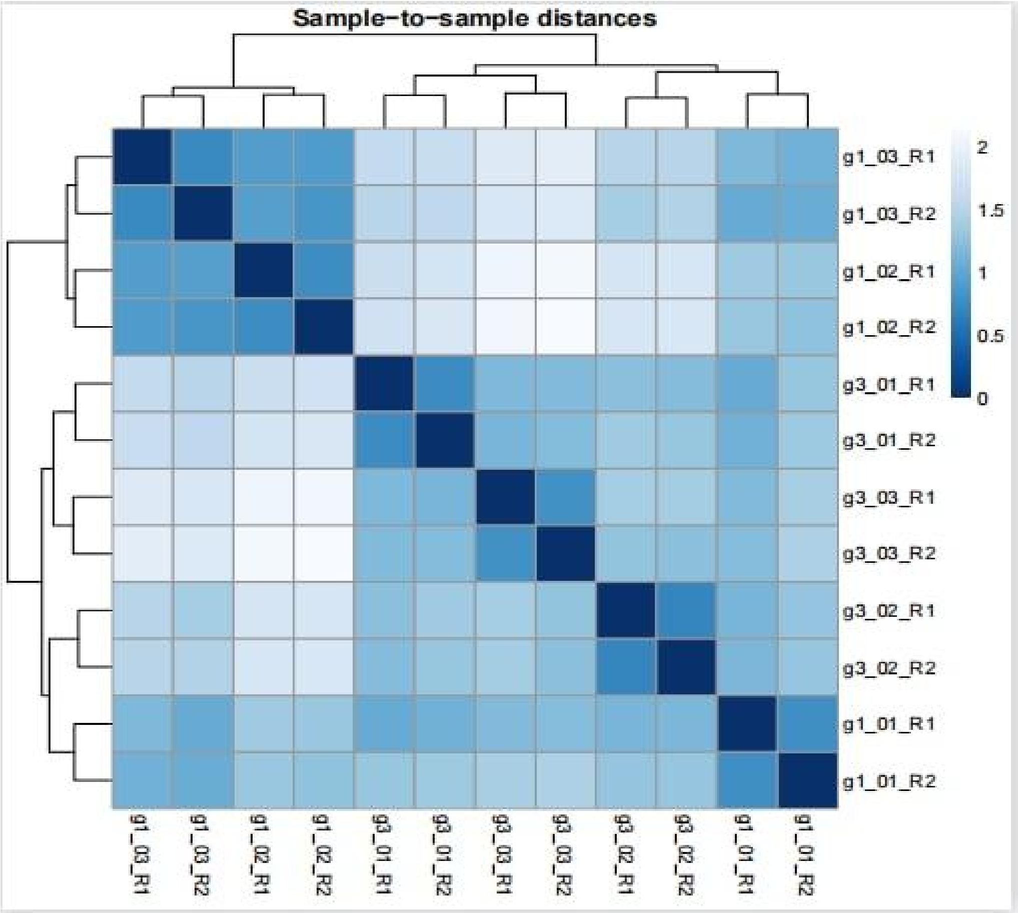
3. HTSEQ-count: This tool took the alignment file in BAM format and feature file in GFF format and calculated the number of reads mapping to each feature.

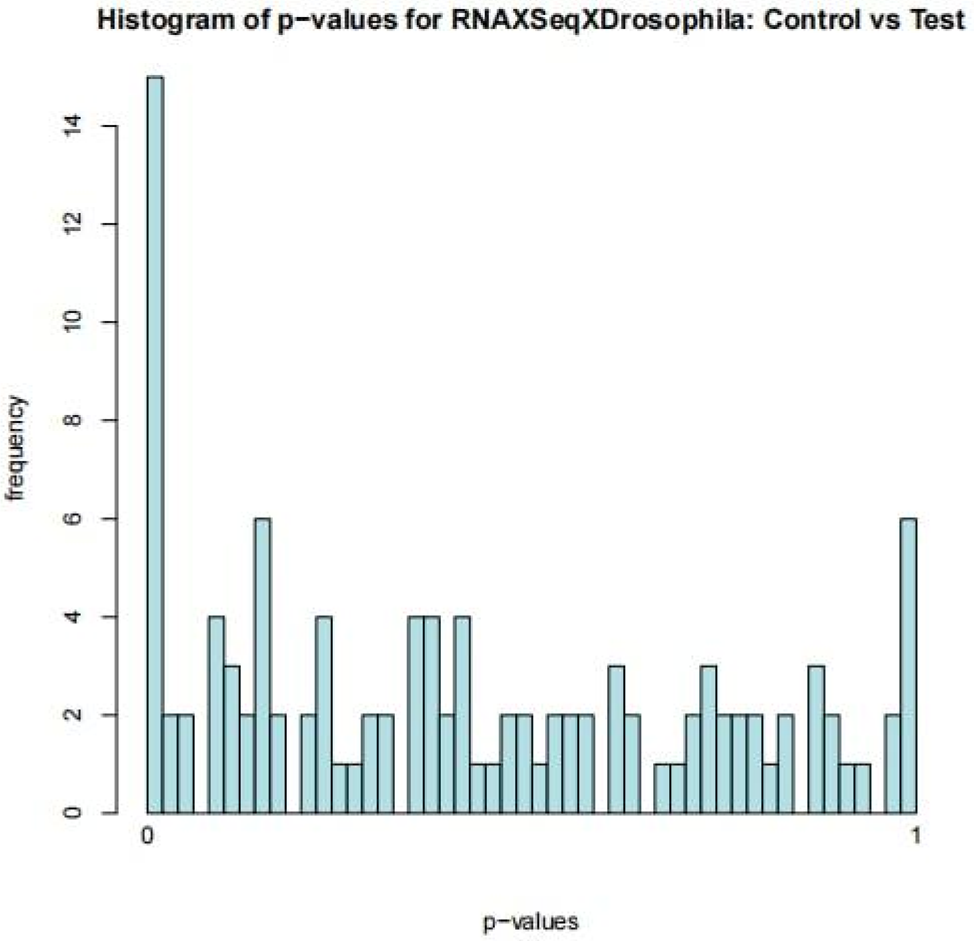
4. Deseq2 (Determines differentially expressed features from count tables): Estimated variance-mean dependence in count data from high-throughput sequencing assays and tested for differential expression based on a model using the negative binomial distribution.

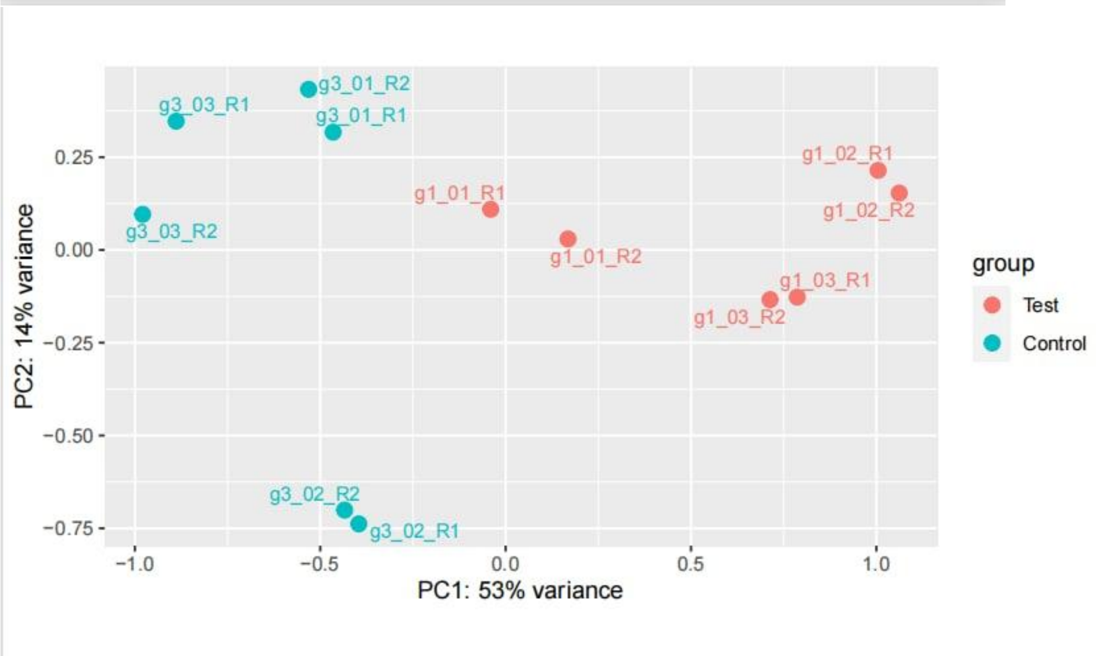
5. FeatureCounts: FeatureCounts is a light-weight read counting program written entirely in the C programming language that was used to count both gDNA-seq and RNA-seq reads for genomic features in in BAM files. Inputs were provided in the BAM format. After running it produced a table containing counted reads, per genes, per row.

Results


Dispersion estimates
dispersion
1e+00
1e-02
1e-04
1e-06
1e-01
1e+00
1e+01
1e+02
1e+03
mean of normalized counts
● gene-est
● fitted
● final


Sample-to-sample distances
g1_03_R1
g1_03_R2
g1_02_R1
g1_02_R2
g3_01_R1
g3_01_R2
g3_03_R1
g3_03_R2
g3_02_R1
g3_02_R2
g1_01_R1
g1_01_R2
2
1.5
1
0.5
0


Histogram of p−values for RNAXSeqXDrosophila: Control vs Test
frequency
0
2
4
6
8
10
12
14
0
1
p−values


g3_01_R2
g3_03_R1
g3_01_R1
g3_03_R2
g1_01_R1
g1_01_R2
g1_02_R1
g1_02_R2
g1_03_R1
g1_03_R2
g3_02_R2
g3_02_R1
PC2: 14% variance
0.25
0.00
−0.25
−0.50
−0.75
−1.0
−0.5
0.0
0.5
1.0
PC1: 53% variance
group
Test
Control

Conclusion

This project has enabled me to understand the basic work flow of alignment quantification, by testing for differential expression analysis for samples of *Drosophila melanogaster*.

Also, I was able to process raw RNA-Sequence data into a list of differentially expressed genes.

Finally, I have understood the relationship between the number of biological replicates in experiment and the statistical power available to detect the differentially expressed genes.

References

Blankenberg, D., Von Kuster, G., Coraor, N., Ananda, G., Lazarus, R., Mangan, M., Nekrutenko, A., & Taylor, J. (2010). Galaxy: a web-based genome analysis tool for experimentalists. *Current protocols in molecular biology*, *Chapter 19*, Unit–19.10.21. https://doi.org/10.1002/0471142727.mb1910s89

Evans, C., Hardin, J., & Stoebel, D. M. (2018). Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. *Briefings in bioinformatics*, *19*(5), 776-792.

Guantes, R., Rastrojo, A., Neves, R., Lima, A., Aguado, B., & Iborra, F. J. (2015). Global variability in gene expression and alternative splicing is modulated by mitochondrial content. *Genome research*, *25*(5), 633-644.

Hateley, S., Hosamani, R., Bhardwaj, S. R., Pachter, L., & Bhattacharya, S. (2016). Transcriptomic response of Drosophila melanogaster pupae developed in hypergravity. *Genomics*, *108*(3-4), 158-167.

Kukurba, K. R., & Montgomery, S. B. (2015). RNA Sequencing and Analysis. *Cold Spring Harbor protocols*, *2015*(11), 951–969. https://doi.org/10.1101/pdb.top084970

Li, B., Ruotti, V., Stewart, R. M., Thomson, J. A., & Dewey, C. N. (2010). RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics*, *26*(4), 493-500.