Introduction

In the study of Brooks et al. 2011, the authors wanted to identify exons regulated by the Pasilla gene (the Drosophila homologue of the mammalian splicing regulators Nova-1 and Nova-2 proteins) using RNA sequencing data. They depleted the Pasilla (PS) gene in Drosophila melanogaster by RNA interference (RNAi).

Total RNA was then isolated and used to prepare both single-end and paired-end RNA-seq libraries for treated (PS depleted) and untreated samples. These libraries were sequenced to obtain RNA-seq reads for each sample. The RNA-seq data for the treated and the untreated samples can be then compared to identify the effects of Pasilla gene depletion on splicing events.

In this problem, you will analyze this dataset using the online toolkit Galaxy. Different from our in-class tutorial, we will use the public Galaxy.eu because it has more up-to-date tools. You will need to establish an account on their website in order to perform the analyses.

The original data is available at NCBI Gene Expression Omnibus (GEO) under accession number GSE18508.

In this problem we will use the first 7 samples:

4 untreated samples: GSM461176, GSM461177, GSM461178, GSM461182 3 treated samples (Pasilla gene depleted by RNAi): GSM461179, GSM461180, GSM461181 Each sample constitutes a separate biological replicate of the corresponding condition (treated or untreated). Moreover, two of the treated and two of the untreated samples are from a paired-end sequencing assay, while the remaining samples are from a single-end sequencing experiment.

The raw RNA-seq reads have been extracted from the Sequence Read Archive (SRA) files and converted into FASTQ files. In the first part of this problem, we will use the files for 2 of the 7 samples. Later you will use all available data to perform differential gene expression analysis.

Useful references

1. The RNA-seq tutorial on ICON produced by the Minnesota Supercomputing Institute, named “MSI_DGE_Galaxy_Human_2017_Spring.pdf”
2. Galaxy Tutorial. This problem is developed based on this site.

Important! Please read the following instructions regarding what need to be submitted for this question

> In addition to answering the questions in this problem, make sure that you share your history with me. In practice, when you have

satisfactorily finished all the questions in this problem, click the cog icon on the top of the history pane, and choose “share and publish”. In the resulting page, choose “Share with a user” and type in “bin-he@uiowa.edu”. Note, sharing your history with your classmates is considered academic misconduct and will result in serious consequences. If you don’t share your history, the relevant points will be deducted.

Data upload

- Create a new history for this RNA-seq exercise called final-RNAseq
- Go to “Shared Data”->“Data Libraries/Galaxy Courses/RNA-Seq (on second page)”
- Select all files in the “input sequences” folder and use the “To History” button on the top to import them into the new history buffer you just created.
- Check (using the pencil icon next to each dataset) to make sure that the fastq files are of the correct type, i.e. “fastqsanger”
- Click on the name of each dataset and choose the “tag” function in the expanded screen. Add a hash tag to each dataset, following the rule “#LastName”. For example, I would label all my datasets as “#He”. Please be sure to do this. Otherwise your results will not belong to you at the time of grading.
- Now go back to the “RNA-seq” folder and enter the “annotation” folder. Import the “Drosophila_melanogaster.BDGP5.78.gtf” into history. Check and make sure that its datatype is “gtf”, not “gff”. If necessary, change the datatype and make sure to click the “change datatype” button on the top to implement the change.

Data QC

Use FastQC to assess the quality of the data and discuss what you found. Specifically, use statistics and key figures from the FastQC result to explain which datasets concern you? What parameter in particular? What would you do to remedy this problem, if possible?

To answer the above questions, it may be helpful to “aggregate” all the FastQC results. Learn about the “MultiQC” tool available in usegalaxy.eu. Briefly, select “FastQC” in the dropdown menu for “Which tool was used to generate logs”, and choose “Raw data” for the type of FastQC output. Then in the data selection menu, hold down the Ctrl (Windows) or Command (Mac) key while doing multiple selection. You can also use “Shift” to select consecutive datasets. View the output and cite useful figures to discuss your QC results.

Genome mapping

1. What is the purpose of genome mapping?
2. How does mapping software deal with introns? Write an example procedure for how to align reads that map across splice junctions.
3. There are quite a few genome mapping softwares, some general while others specialized for certain applications. We have learned to use TopHat in the class. This time we will again use it, and compare its results to another mapper. You can use the