Workplan for RISE program – 2nd Semester 2015-2016

Evaluating quantification and expression methods using *Drosophila Melanogaster* data

Student: Iván L. Jiménez Ruiz

Primary Mentors: Dr. Humberto Ortiz-Zuazaga; Dr. Ricardo González; Alexander Ropelewski; Dr. José Agosto Rivera

Secondary Mentors: Pallavi Ishwad

Abstract:

Even with the abundance of data that has been obtained from bioinformatics, detection of differential expression in two samples is a recurring problem. This detection is affected by the accuracy of the quantification methods used in the process, which in turn depend on the techniques used for assembly. The quantification methods that were compared include Cufflinks, RSEM and eXpress. Serially diluted RNA-seq data from D. melanogaster containing aliquots from both External RNA Control Consortium and Schneider 2 (S2) cells line at different percentages were assembled. We used Sequence Read Archive (SRA) data from the National Center for Biotechnology Information (NCBI) database to retrieve four transcriptomic datasets containing a total of 2,049,901,812 nucleotides (nt). The average number of bases per read in each dataset was 36 nt. Quantification results obtained from the assembled transcripts produced by Trinity and RNA Sequencing by Expectation Maximization (RSEM) [de-novo] were compared with those from TopHat and Cufflinks [reference based] to determine the validity of these protocols. We hypothesize that de-novo assembly is equally as powerful as reference-based assembly in the detection of differential expression. TopHat and Trinity are bioinformatics programs commonly used in the process of RNA assembly; Cufflinks and RSEM are quantification tools that generate Fragments per Kilobase of transcript per Million mapped reads (FPKM) that are used in differential expression detection. Using a heatmap produced by running RSEM, several genes were identified as being differentially expressed at a p-value exceeding 1e-3. Annotation of these genes through the Basic Local Alignment Search Tool (BLAST by NCBI) database identified them as coming from viral sources, specifically Drosophila birnavirus and X virus. These genes were not identified using reference-based approach, as their source was different from the reference used. This approach provides a draft workflow to be used with data produced from de-novo RNA-seq experiments using non-models organisms.

Scientific Objectives:

- Develop a draft scientific paper to be published by the end of the semester.
- Analyze Drosophila and foreign RNA-seq reads by using a variety of bioinformatics tools
- Apply bioinformatics tools for the analysis of spike-in data for the detection of differential expression at the gene level of *D. Melanogaster*

Prior Work: The project is based on the analysis of SRA data files of Drosophila published at the NCBI website, which can be downloaded from: http://www.ncbi.nlm.nih.gov/sra

The specific data files and their respective project websites are:

- SRR039935 - http://www.ncbi.nlm.nih.gov/sra/?term=SRX019236
- SRR039933 - http://www.ncbi.nlm.nih.gov/sra/?term=SRX019234
- SRR039460 - http://www.ncbi.nlm.nih.gov/sra/?term=SRX018872
- SRR039458 - http://www.ncbi.nlm.nih.gov/sra/?term=SRX018870

Deliverables:

- Workflow to be used in the detection of differential expression of reads in a de-novo (as well as reference) based assembly project using *D. Melanogaster* data.
- A draft of a PLOS one paper

Milestones:

January:	Finish de novo assembly using "new" (clean) sequences, establish abundance estimations using RSEM and eXpress
February:	Develop figures for mapping and de novo assemblies
March:	Compare and contrast gene and contig results of both assemblies
April:	Develop a draft PLOS paper for publication
May:	Submit draft PLOS paper for publication

_____

Dr. Humberto Ortiz-Zuazaga
University of Puerto Rico, Rio Piedras Campus
humberto.ortiz@upr.edu