June, 2016

Analysis of RNA-Seq Alignments using DNA Subway Green Line (computational)

Raymond A Enke

DNA Subway Green Line Analysis of RNA-Seq Alignments

Dr. Ray Enke Bio 480 Advanced Molecular Bio Lab
James Madison University

How to cite this work

Objectives:

- Review basic steps of RNA-Seq bioinformatics analysis in DNA Subway Green Line
- View and run basic analytics of RNA-Seq data set in DNA Subway Green Line

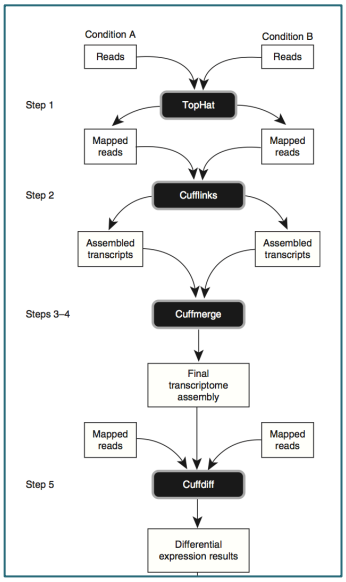
URLs for Lab Activity:

- DNA Subway: *http://dnasubway.iplantcollaborative.org/*
- GoogleSheet for RNA-Seq Preliminary Analysis: *tinyurl.com/hj3zr5s*

I. Overview of RNA-Seq Bioinformatics Analysis in DNA Subway Green Line

DNA Subway is a bioinformatics pipeline designed to make high-level genome analysis broadly available to students & educators and provides easy access to the types of data and informatics tools that drive modern biology. Using the metaphor of a subway map, DNA Subway organizes bioinformatics analysis tools into logical and easy to use workflows. We will view a public RNA-Seq project that I previously created in the DNA Subway Green Line, a tool designed specifically for RNA-Seq data analysis. The RNA seq project that my lab conducted and that we will analyze in this lab is as follows:

- E8 chicken whole retina mRNA X2 replicates
- E18 chicken whole retina mRNA X2 replicates
- E18 chicken whole cornea mRNA X2 replicates



Each of these mRNA samples were prepped and sequenced using an Illumina HiSeq Next Generation Sequencer. This experiment generated ~20-60 million short (150 bp) sequencing reads/sample. I've started a bioinformatics workflow in the Green Line to align each of these short reads to the reference chicken genome (Galgal4). Once reads from each sample are aligned we can then assemble individual transcripts and a whole transcriptome for each sample. We can then compare transcriptomes between samples to determine which individual genes are differentially expressed between samples. Here's an overview of a popular set of software packages that run these jobs specifically for RNA-Seq data called the **Tuxedo Protocol**:

TopHat: aligns millions of reads/sample to a reference genome. You can think of it as performing 20-60 million simultaneous BLAST searches

Cufflinks: merges individual reads mapped by TopHat into full transcripts

Cuffmerge: merges individual transcripts into a sample transcriptome

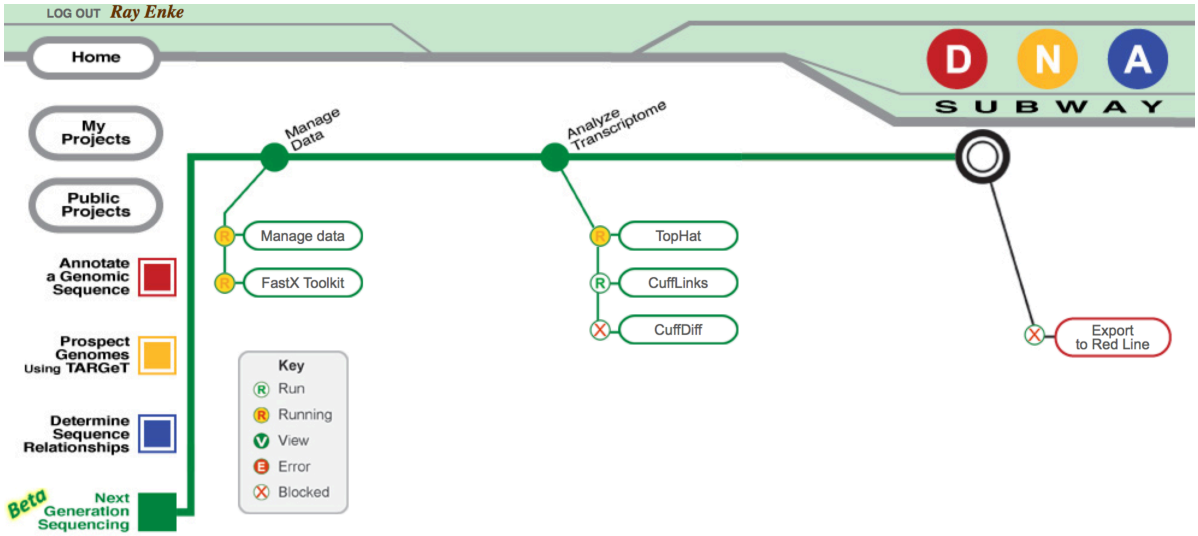
Cuffdiff: compares transcriptomes between 2 sets of sample replicates to ID differentially expressed genes (DEGs)

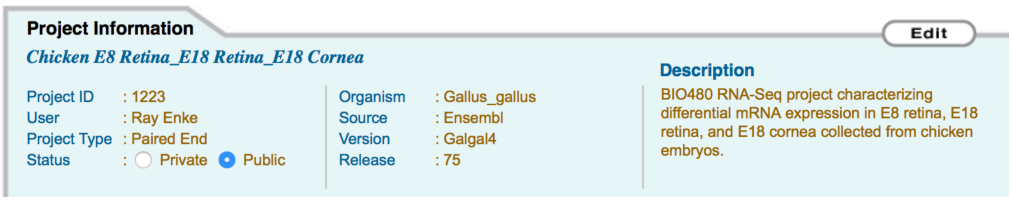
We will discuss each of these software tools more in depth in later labs.

II. TopHat Sequence Alignment Analysis in DNA Subway Green Line

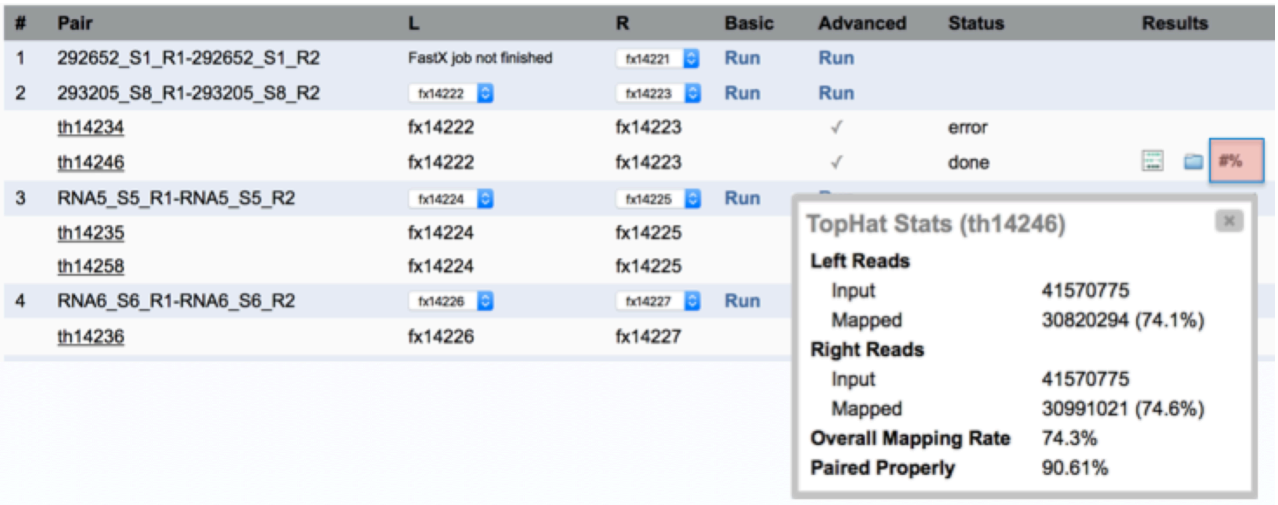
Today you will simply view the results of several **TopHat** jobs that have finished running.

- Navigate to and login to DNA Subway (http://dnasubway.iplantcollaborative.org/)
- Click on the Public Projects and open to Green Line Project #1223 “Chicken E8 Retina E18 Retina E18 Cornea” (Note: this activity can be done for any public Green Line project)

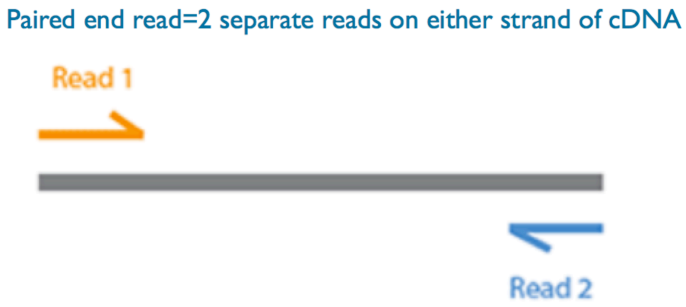




- Click on the **#%** symbol to view some stats for each of the finished TopHat jobs.



Each cDNA fragment in this experiment was actually sequenced 2X, once from the top strand (left) and once from the bottom strand (right). This is referred to as **paired end sequencing** and allows you to generate 2X the amount of sequence data without increasing the number of fragments sequenced.



These TopHat Stats report the number of **paired end reads** for each sample, how many of the reads mapped to the reference genome (chicken in this case), and the amount of left and right reads that were able to be computationally paired up. These stats give us some basic information on how each individual sample transcriptome was built. These transcriptomes will then be compared to each other to ID differentially regulated genes in subsequent bioinformatics steps.

- Navigate to a GoogleSheet that I created to log these TopHat stats for our project: *tinyurl.com/hj3zr5s*
- Download the sheet (you are unable to edit) and log the TopHat stats for the 3 jobs that are currently finished.
- Input this information into your notebook

- Here are the sample identifiers:

FASTQ Filename	Sample	Read End
RNA5_S5_R1.fastq.gz	E8 retina replicate #1	1-left
RNA5_S5_R2.fastq.gz	E8 retina replicate #1	2-right
RNA6_S6_R1.fastq.gz	E8 retina replicate #2	1-left
RNA6_S6_R2.fastq.gz	E8 retina replicate #2	2-right
RNA7_S7_R1.fastq.gz	E18 retina replicate #1	1-left
RNA7_S7_R2.fastq.gz	E18 retina replicate #1	2-right
RNA8_S8_R1.fastq.gz	E18 retina replicate #2	1-left
RNA8_S8_R2.fastq.gz	E18 retina replicate #2	2-right
292652_S1_R1.fastq.gz	E18 cornea replicate #1	1-left
292652_S1_R2.fastq.gz	E18 cornea replicate #1	2-right
293205_S8_R1.fastq.gz	E18 cornea replicate #2	1-left
293205_S8_R2.fastq.gz	E18 cornea replicate #2	2-right

We will further analyze these data in subsequent labs.

Assignment:

Navigate to and complete the TopHat alignment stats GoogleSheet (tinyurl.com/hj3zr5s) and use data to construct a bar chart plotting input reads as well as mapped and paired reads for each sample.