CSE 180 Final Project

Minh Nguyen

A11302948

**Biological Motivation**

We want to be able to determine the transcription level gene expression. This is useful for determining the pool of candidate genes that may lead to certain functions in the subject from which the sequences being examined were obtained. These expression levels can also be compared to the expression levels in other samples in order to gain an understanding of how the two samples may be related or perhaps not related at all.

By determining the expression levels of genes, we are able to see the pieces of what constitutes a tissue. This is essentially determining the building blocks of life and sorting them into categories so that we might be able to understand what the role of each block is.

A practical application of this tool would be running it once on a mouse heart cell sequence and once on a human heart cell sequence and determining how related they might be by using cuffdiff on the outputs of both in order to determine significant changes or similarities.

**Project Details**

The general workflow of the project is as follows:

1. Inputs: Reference genome in fasta format. If users have a reference genome in fastq the users can use seqtk to convert them to fasta. Rna-seq reads in fastq.gz format (If they are simply fastq the user should gzip them to save space. Eventually this would be an online tool using cloud computing and we would like to reduce the file size)
2. Genome is generated from the fasta sequence information using STAR. This step often fails due to the amount of memory required. When the tasks are clustered onto a cloud architecture we would need servers with at least 32GB RAM and it would be nice to have the ability to handle 8-24 threads.
3. STAR maps the fastq reads to the genome and leaves us with mapping summary statistics, alignments in sequence alignment map (SAM) format, unmapped reads, and logs.
4. Samtools is used to sort the alignments as cufflinks takes in a sorted SAM file. Here we can produce a BAM (binary SAM file) or a SAM. The pipeline chose to produce a SAM file so that it is still human readable.

The sorted SAM file is piped into cufflinks and the cufflinks output is created in the ./cufflinks\_results folder. This contains transcripts.gtf, isoforms.fpkm\_tracking, and genes.fpkm\_tracking. Gtf is gene transfer format (gene structure information) and fpkm is fragments per kilobase of transcript per Million mapped reads.

**Project Results**

Put the inputs (reference fasta and rna-seq read fastq) inside of the inputs folder

Run the following command in order to give read/execution permissions to others

* chmod 755 pipeline.sh

Run ./pipeline.sh –c in order to clear the outputs of the previous run

Run ./pipeline.sh or sh pipeline.sh in order to run the pipeline.

The fastqc outputs can be seen in the fastqc\_results folder. This gives an idea on whether the data used should be reconsidered as good data or bad data. On realistic data, this pipeline can take a long time as STAR can take a long time to run especially on large sets of data. Even on relatively small files the STAR genome generation can fail due to insufficient amounts of memory available.