RNA-Seq Pre-processing

Jessica Murphy

March 15, 2019

Overview

The following data was obtained from a 2x2 design using fly models. The two variables of interest are strain and tissue. The strains are white eyed (W) and sevenless (S) and the tissues are optic lope (O) and retina (R). There are four sample types: SO#, SR#, WO#, WR#.

Raw Reads

The raw reads are located on Yampa under `/BIOS6660/Homework6/rawReads/`. They are pair-ended reads with a read length of 150. A text file with the number of reads per sample is under `/home/murphjes/BIOS6660/Homework_6/Data/RawReadsCounts.txt.` and the bash script to count them is under `/home/murphjes/BIOS6660/Homework_6/Code/CountRawReads.sh`.

code:
```
chmod -R u+wrx *
./CountRawReads.sh
```

Trim Reads

The first 15 bases were trimmed off the reads using cudadapt version 1.18, which gives us a read length slightly less than 135. The trimmed reads are located on Yampa under `/home/murphjes/BIOS6660/Homework_6/TrimmedReads/` and the bash script to trim them is located under `/home/murphjes/BIOS6660/Homework_6/Code/TrimReads.sh`. A text file with the number of trimmed reads per sample is under `/home/murphjes/BIOS6660/Homework_6/Data/TrimmedReadsCount.txt.` and the bash script to count them is under `/home/murphjes/BIOS6660/Homework_6/Code/CountTrimmedReads.sh`.

code:
```
chmod -R u+wrx *
./TrimReads.sh ./CountTrimmedReads.sh
```

Check for Quality

FastQC was used to evaluate the quality of the reads and check that the trimming went well. Ideally, we would like green checks on the Per Sequence Quality Scores and the Per Base Sequence Content. According to the reports, Sample A looked good, B was okay, and C and D were not very good. Since this data is a small subset of the original data, the quality is not as good. The reports are located on Yampa under `/home/murphjes/BIOS6660/Homework_6/FastQC/`.

code: `fastqc *.fastq.gz`

Align Reads

The trimmed reads were then aligned to the dm6 genome using hisat2 version 2.1.0. The reference files are located on Yampa under `/BIOS6660/Homework6/indexFiles/` and the aligned reads are located under `/home/murphjes/BIOS6660/Homework_6/AlignedReads/`.

code: `export PATH=/usr/local/bin/samtools-1.3:$PATH hisat2 -x /BIOS6660/Homework6/indexFiles/dm6.hisat -1 /path/sample_R1_trimmed.fastq.gz -2 /path/sample_R2_trimmed.fastq.gz | samtools view -bS - > alignedSample.bam` (repeat for each pair of samples)

Quantitate Reads

The trimmed files were also quantified to the Ensembl dm6 transcriptome using RSEM version 1.3.1. The reference files are located on Yampa under `/BIOS6660/Homework6/indexFiles/` and the quantitation files are located under `/home/murphjes/BIOS6660/Homework_6/Quantitation/`.

code: `export PATH=/usr/local/bin/bowtie2:$PATH rsem-calculate-expression -p 8 --time --seed 2020 --bowtie2 --paired-end --seed-length 20 /path/sample_R1_trimmed.fastq.gz /path/sample_R2_trimmed.fastq.gz /BIOS/Homework6/indexFiles/dm6.ensembl sampleName` (repeat for each pair of samples)

Sample Counts

The following table shows the sum of reads for each sample at each point in the data processing.

Sample	Raw	Trim	RSEM	Trim Percent	RSEM Percent from Trim
A1_SO1_S1	135,093	134,245	75,549	99.37%	56.28%
B1_SR1_S5	137,427	136,554	77,279	99.36%	56.59%
C1_WO1_S9	166,997	165,917	69,142	99.35%	41.67%
D1_WR1_S13	213,363	212,228	85,839	99.47%	40.45%