Date: June 10, 2017

Name: Ben Larson

Re: Bioinformatics Project Results Summary

Motivation:

High throughput RNA sequencing is used to look for actively transcribed genes in samples of human blood. These transcriptomic profiles can be compared to gene sequences associated with various types of cancer, thereby enabling earlier detection compared with conventional methods. However, common genes such as those encoding for ribosomal RNA (rRNA) constitute the bulk of the profiles, and so obscure the signal from genes of interest. Experimental methods have been developed to deplete rRNA, but it is important that the depletion does not otherwise distort the profile. Here, I use pile-ups to characterize transcriptomic profiles. The term pile-ups is a bioinformatics colloquialism for a histogram of sequence reads that map to a given position in a reference genome (the single rRNA gene in the present case). Here I test whether these rRNA pile-ups change during rRNA depletion as a proxy for whether other components of interest are affected by the depletion procedure.

Methods:

Data: RNA Sequence data come from Illumina in a fastq format (1-4 GB when compressed), and data from 5 experiments with 2-4 replicates comprise the dataset for the current exercise.

Sequence Mapping: STAR v. 2.5.3a was used to map the sequence data to the rRNA gene, samtools v. 1.4 was used to sort the aligned data, and finally, the *R* package, Rsamtools was used to index the sorted data and generate the pile-ups. In an attempt to improve run times, this step was carried out variously on a dual core MacBook Pro with 8 GB of memory, a AWS Linux m4.4xLarge with 16 cores, and 64 GB RAM, and finally an AWS Linux m4.16xLarge with 64 cores and 256 GB RAM. Run times ran, respectively from ~150, to 90, to 20 mins. The mapping can be monitored by viewing the mappers log file, for the example, this would be: L11_L1_Log.progress.out. I have left the m4.4xLarge running and given instructions in the ReadMe.txt file to log on and run the example.

K-S Test: The pile-ups were compared with a Kolmogorov–Smirnov test (Fig. 2). Replicates for any given experiment were compared with all other replicates for that experiment to generate a set of 15 K-S statistics under the null hypothesis that the pile-ups are indistinguishable from one another for a given experiment.

Bootstrapping: These 15 K-S stats were bootstrapped (Fig. 2) to generate a distribution of K-S stats against which to compare the K-S stat for a depleted sample relative to an undepleted one.

Results:

Log Files

1. basespace.remote.example.log – logs output from larson_baseCode_remote_example.R
2. basespace.local.log – logs output from larson_baseCode_local.R

Table 1. Shows Experimental ID and Percentage of RNA reads that map to the reference genome (rRNA in this case) for the 4 comparison cases. All cases except for the two lanes of the same experiment (L11, L1 and L2) have a very low probability of coming from the same distribution as the null according to the p values. I therefore reject the hypothesis that pile-ups are the same.

Experiment ID	Percent Mapped to rRNA	K-S Statistic relative to L11_L1	Empirical P value
L11_L1	67.3	-	
L11_L2	67.84	0.0077	0.9996
L12_L1	20.6	0.1951	$< 2\times^{-4}$
Red7_L1	24.3	0.1951	$< 2\times^{-4}$
Red8_L1	46.4	0.0481	$< 2\times^{-4}$

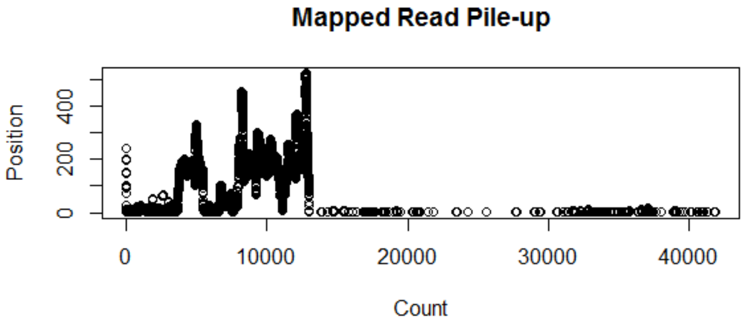

Figure. 1. Example of typical Pile-up for sequence data line to the rRNA gene

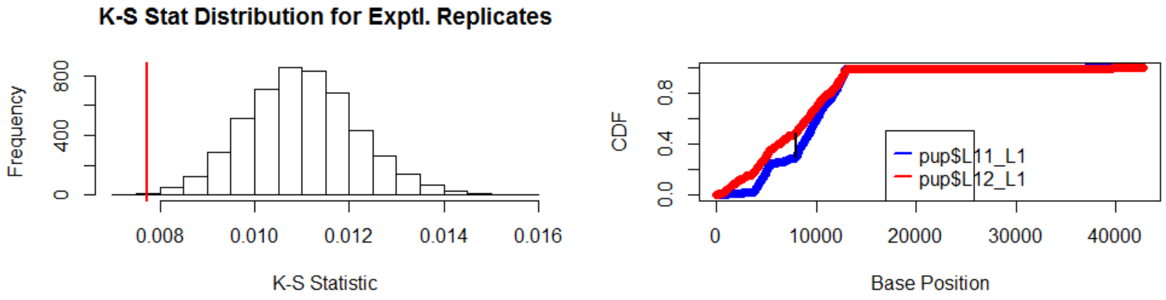

Figure. 2 Showing the distribution of K-S statistics (left) calculated as described in the methods and one example of the comparison of cumulative distribution functions (right). Only the K-S statistic gotten from two replicates of the same experiment (red vertical line, but CDF's not shown here) falls within the null distribution. All other K-S stats (including for the comparison depicted at right) are not on scale.

Conclusion:

Only the experimental replicates show a statistically similar mapped-read pile-up, which could mean the RNA depletion does, in fact, change the transcriptomic profile. However it is also possible that other experimental uncertainties not considered here explain the differences in pile-ups. One way to assess this would be to compare two profiles with a similar amount of depletion to see if the differences in pile-up persist. If so, that would suggest experimental uncertainty is the culprit, if two pile-ups were similar after depletion, such a result would point to changes introduced by the depletion itself. Further such work, however, is beyond the scope of the current project.