BIOL 6150 Project - Assignment 2

Group No. 3:

Submission on behalf of all team members:
Jiachen(Nicole) Duan
Jacob Feldman
Devishi Kesar
Deepali Kundnani
Stephen Thomas
Gulay Bengu Ulukaya

Short description of datasets

Dataset 1 (GSE120534) contains RNAseq data from an Illumina NextSeq 500 [GPL18573]. The series investigates the effect of silencing p53 in p53 hotspot mutant enriched in lung cancer cell lines: NCI-H2087 (p53 mutant V157F) and NCI-H441 (p53 mutant R158L). We have analyzed differentially expressed genes in each variant cell line for their p53 silenced conditions vs native control cell lines. (Tab 1 and 2 of attached excel sheet)

Dataset 2 (GSE 79051) contains RNASeq data from an Ion Torrent Proton [GPL17303]. The series investigates NGFR knockdowns in lung cancer lines NCI-H1299 and NCI-H460. We have analyzed differentially expressed genes in WTp53 expressing lung cancer cell line (H460) vs p53 lung cancer cell line (H1299). (Tab 3 of attached excel sheet)

**We also performed differentially expressed genes in the Dataset1(both variants combined) versus the Dataset2(Tab 4 of attached excel sheet)

Pipeline Elements

1. Downloading datasets: SRR files for each sample was downloaded using prefetch and fastq-dump from SRA-tools and uploaded to Galaxy
2. Quality control: FastQC(UseGalaxy) was used before and after trimming of ambiguous read ends.

3. Trimming: Trim Galore (UseGalaxy) was used on both datasets to trim 10 bp off of both the 5' and 3' ends. Since dataset 2 had longer reads, the 3' ends needed additional trimming beyond 200 bps, this was done using Trim (UseGalaxy).
4. Alignment: HISAT2 was used for alignment on both datasets. Dataset1 is pair ended, the input used is interleaved fastqsanger files and Dataset 2 has single read fastqsanger reads as input. Reference genome was hg19.
5. Expression counts and DEG analysis:We used an assembled R pipeline consisting of featureCounts and DeSeq2 for getting counts and gaining differentially expressed genes respectively.

Statistical analysis and differential gene expression analysis

Data was normalized using log base 2 transformation. P-values are calculated to determine the statistical significance of the hypothesis test being that the treated cell lines are different from control cell lines. Genes with p-value less than 0.05 were considered having significant differential expression among samples. Probability of false positives are calculated using the False-Discovery-Rate-adjusted p-value from Benjamini & Hochberg adjustment method. While a p-value of 0.05 implies that 5% of all tests result in false positive, an FDR-adjusted p-value of 0.05 implies that 5% of significant tests will result in false positive. Another criteria taken into consideration for selection of significantly upregulated and downregulated genes was the |log2FoldChange| value to be greater than 1.0 for all three datasets(Two variants from Dataset1 and one WT from Dataset2).

Finally, we confirmed the integrity of our data by doing a correlation analysis of absolute log2(fold change) value of variant V157F DEGs found in the paper versus the ones we got from our analysis.(Tab 5 of attached excel sheet). **We were able to find 36 out of 47 genes.**