Group 2 Final Project Rough Draft

Introduction:

We will introduce RNA-sequencing and The Cancer Genome Atlas (TCGA). We will introduce how the TCGA project has pre-processed the available Level 3 (RPKM normalization). We will introduce the GSE62944 study and how they re-aligned the TCGA RNA-seq data by accessing the FASTQ files from the actual TCGA RNA-seq experiments from the permission-needed Cancer Genomic Hub. And we will talk about experimental design of RNA-seq, and how the end points could be different, but for purposes of this study we are interested in differential expression analysis. The most popular differential expression analysis packages on R are DESeq2 and edgeR (and the older package, limma), which both require raw (non-normalized) integer-based counts. The problem with TCGA data is that they do not provide integer based counts, and they are already normalized. In light of this, we will introduce different normalizations that are typically used for RNA-seq data (e.g. library size normalization, gene length normalization) and the importance (or unimportance) of certain normalizations.

We are conducting a comparative analysis between TCGA Level 3 data and GSE62944 Rsubread aligned data. For this analysis, we are grabbing only the primary prostate adenocarcinoma (PRAD) data. We are only using the matched tissue samples, so only patients (whom have a unique bar code) that have both expression data from their normal tissue and their tumor tissue. We are interested in seeing if the statistically significant results obtained from differential expression analysis are different between the two differently aligned dataset.

Objective: Comparative analysis of TCGA RPKM Level 3 Data with realigned TCGA data using Rsubread feature counts

Specific Aims:

1. Research different alignment methods used in RNA-seq and understand how count data are presented after alignment
2. Research statistical assumptions of RNA-seq data and how they are modeled to detect differential expression
3. Conduct comparative analysis of statistically significant differentially expressed transcripts between TCGA RPKM Level 3 data and TCGA feature counts aligned by *Rahman et al., 2015*

Methods:

* Download TCGA level 3 data from Broad Institute’s firehose portal
  + Match normal-tissue samples
* Download GSE62944 normal and tumor datasets
  + Subset PRAD data
    - Match normal-tumor tissue samples
* Introduce different normalization methods and the math behind them
* Conduct differential expression analysis using limma, edgeR, DESeq2 packages in R
* Create venn diagrams comparing the statistically significant results
* Create heatmaps showing the expression patterns from both studies across patients
* Also create plots showing the relationship between adjusted p-values and gene length (if possible)

Results:

Our results will involve an exploratory data analysis stage, validating some of the model assumptions that the field has a consensus on. Our preliminary results reveal that different approaches to analyze differentially expressed genes result in distinct outcome. The inputs at the beginning of differential analysis are often overlooked, yet they require an appropriate input for interpretation of the results to be valid.

Discussion:

We will write a detailed discussion on our results, interpretation of results, conclusions, and future directions.