STA 525 - Final Project

Suruchi Ahuja, Hoi Lam Tai, Abbas Rizvi, Jingchen Zhang

May 10, 2016

Introduction

The Cancer Genome Atlas (TCGA) is a publicly available online consortium that offers download access to data from next-generation sequencing (NGS) experiments conducted on human cancer patient tissue samples. TCGA tissue collection is comprised of matched tumor and normal tissues from 11,000 patients in 33 cancer types and subtypes. Researchers throughout the world are accessing these data in order to make and validate discoveries. The essence of TCGA is invaluable to contemporary cancer research and hopefully more projects similar to it arise in the forthcoming years.

The characterization of gene expression in cells via measuring mRNA levels has been an interest to researchers for quite some time now, both in terms of which genes are expressed in tissues, and the level at which they are expressed in said tissues. Before the post-genomics era, gene expression was thought to be a reliable measure of protein abundance as per the Central Dogma of Biology. This is assumption is no longer as definitive, as there is much evidence for post-transcriptional and epigenetic events altering the abundance of protein irrespective of mRNA levels, suggesting a weakened correlation between gene expression and protein level. mRNA can be sequenced effectively and relatively easily, compared to to the field of proteomics which has had difficulty defining the abundance of certain proteins with such specificity due to complexity of the post translational modifications. However, measuring mRNA concentration levels is still a useful tool in determining global changes in the genome-wide transcriptional network (transcriptome) of the cell and how it is affected in the presence of external stimuli (e.g. drug treatments) or how the cell/tissue transcriptome differs between a healthy state and a diseased state. The aforementioned comparison of gene expression between samples is called *differential expression*.

A newer technology to measure relative RNA transcript abundance is called RNA-sequencing (RNA-seq). RNA-seq uses next-generation sequencing (NGS) to identify and measure the presence and quantity of RNA in a biological sample at a given moment in time. TCGA offers RNA-seq data and is available to download from the Cancer Genomics Hub. These RNA-seq data can be used to detect differential expression between normal and tumor tissue samples of many different cancer subtypes. Differential expression is detected by statistical analyses that are typically computed using the open source statistical computing software `R`. And in the field of bioinformatics, three popular `R` packages for detecting differential expression from RNA-seq experiments are `DESeq2`, `edgeR`, and `limma`. `DESeq2` and `edgeR` are newer, more robust packages, while `limma` is the classic, less sophisticated package. `edgeR` and `DESeq2` require unnormalized integer-based read counts and `limma` can be used on normalized non-integer-based counts [2-3].

'Level 3' TCGA RNA-seq data has been aligned to the human reference genome using MapSplice (Wang *et al.*, 2010), quantified at the gene and transcript levels using RSEM (Li and Dewey *et al.*, 2011) and normalized by reads per kilobase of transcript per million mapped reads (RPKM) yielding expression values in the form of non-integer-based counts [1]. Henceforth, we will refer to TCGA RPKM normalized data simply as Level 3 data. This can immediately be seen as problematic as `edgeR` and `DESeq2` require un-normalized read counts that are integer-based. In order to get TCGA RNA-seq expression values that are not RPKM normalized, the FASTQ files (the raw output from the RNA-seq platform) are needed and an alternative method of aligning and processing the data would be needed. TCGA offers the raw FASTQ files, but permission is needed to access them from the Cancer Genomics Hub, rendering the FASTQ files not as accessible as the Level 3 data.

A recent publication in the journal *Bioinformatics*, Rahman *et al.*, 2015 was permitted to download TCGA RNA-seq FASTQ files (Level 1) from Cancer Genomics Hub, and alternatively processed the data using `Rsubread`. `Rsubread` is an open source `R` package that has shown high concordance with other existing methods of alignment and summarization, but is simple to use and takes much less time to process the data

[1]. Rahman *et al.*, 2015, alternatively aligned TCGA RNA-seq data can be found on NCBI GEO Datasets (GEO Accession Number: GSE62944), and henceforth will be referred to as GSE62944.

The purpose of our project was to conduct a comparative analysis between two RNA-seq data sets that comprise of the same samples but were processed and compiled by different pipelines and how the pre-processing of count data affects detection of differential expression outcome using `DESeq2`, `edgeR`, and `limma`. This is important because cancer researchers are using TCGA for making discoveries, and if they are incorrectly identifying differentially expressed genes, it may effect the outcome of their future studies.

Materials and Methods

The Cancer Genome Atlas (TCGA)

1. What is TCGA?
2. Write brief paragraph about TCGA.

RNA Sequencing

1. What is RNA seq?
2. Write brief paragraph of RNA seq and how it is processed (generally speaking) to expression values.
3. Explain expression values and read counts.
4. Talk about overdispersion and the negative binomial assumption of RNA seq.

Data sets

1. The Cancer Genome Atlas Level 3
2. GSE62944 – Alternatively processed and compiled RNA-sequencing data for thousands of samples from TCGA [1]

Level 3 Data

1. What is level 3 data?
2. How is it aligned? How is it normalized?
3. What is RPKM (write formula)?

GSE62944

1. What is this data?
2. How was it aligned?

Pre-processing Data

Level 3 prostate adenocarcinoma (PRAD) data was downloaded from the Broad Institute Firehose Portal. GSE62944 was downloaded from NCBI Gene Expression Omnibus. GSE62944 was parsed down from expression values of 22 different cancer subtypes to just expression values of PRAD. The normal (n=54) and tumor (n=376) expression values come in separate files for both of these studies (a total of 4 files). The samples were matched such that only the patients that had both normal and tumor expression values were selected (n=52 normal, n=52 tumor). These data were then ready for differential expression analysis.

limma

1. What does limma do?
2. How does it model RNA seq?
3. What normalization methods does it use (voom)?
4. What type of statistical test does it do?
5. Whats unique about it from the other packages?
6. Look at my code and see what the limma functions that I used do ... explain some of them... the important ones

edgeR

1. What does edgeR do?
2. What normalization methods does it use (library size –> TMM)?
3. How does it model RNA seq (limiting variance?)
4. What type of statistical test does it do?
5. Whats unique about it from the other packages?
6. Look at my code and see what the edgeR functions that I used do ... explain some of them... the important ones

DESeq2

1. What does DESeq2 do?
2. What normalization methods does it use (does it use a library size normalization)?
3. How does it model RNA seq (limiting variance, variance stabilizing transform?, regularized log transform? what does the package even do? I'm not sure, find out....look at my code too, see what the functions do)?
4. What type of statistical test does it do?
5. Whats unique about it from the other packages?
6. Look at my code and see what the DESeq2 functions that I used do ... explain some of them... the important ones

Volcano Plot

1. What are volcano plots?
2. How can we use them to intepret differential expression?

Venn Diagrams

1. What are venn diagrams?
2. How do we use them as a comparative tool in our analysis?

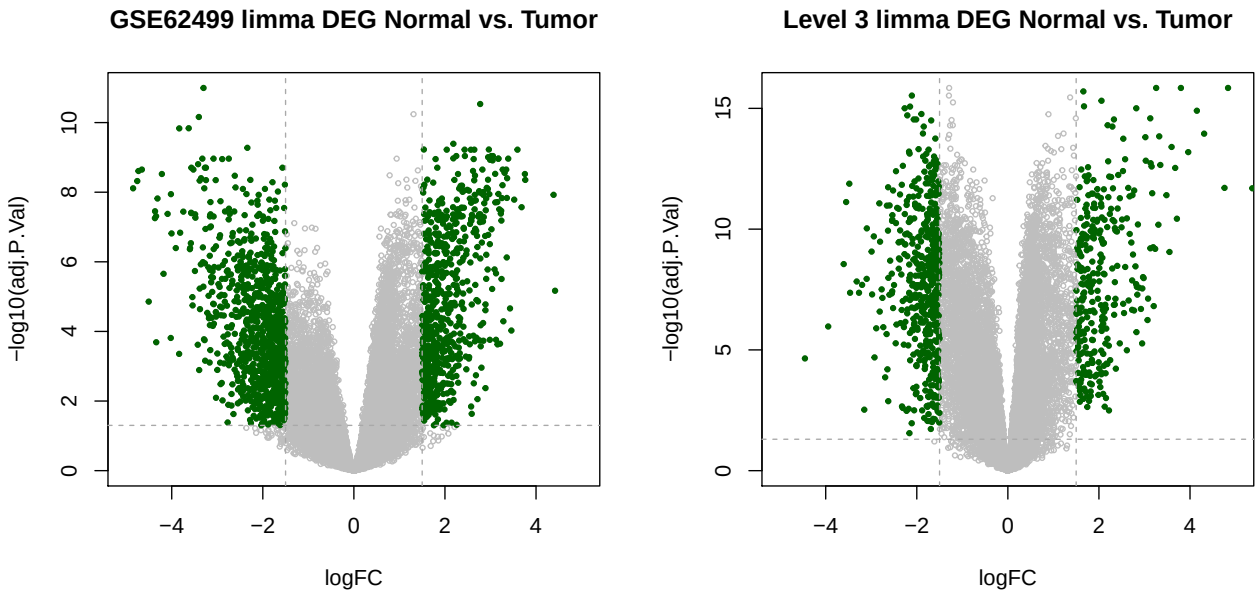

Figure 1: Volcano Plots of Level 3 and GSE62499 Differential Expression from Matched Normal to Tumor Tissue Using Limma. Each dot on the plot represents a single gene. Statistically significant ($|\text{logFC}| > 1.5$; FDR < 0.05) genes are highlighted in green.

Results

Differential expression analyses were conducted using between matched normal to tumor prostate adenocarcinoma (PRAD) tissues using the `R` packages `limma`, `edgeR`, and `DESeq2`. Each of these packages implements similar, but moderated, statistical techniques when calculating differential expression (see Materials and Methods). The threshold that was used to consider differentially expressed genes (DEGs) was a $\left|log(\text{fold change})\right| > 1.5$ and a false discovery rate (FDR) < 0.05. Level 3 data was rounded to the nearest whole integer so it could be analyzed with `DESeq2` and `edgeR` (this will be discussed further in the discussion).

We analyzed differential expression in two ways:

Discussion

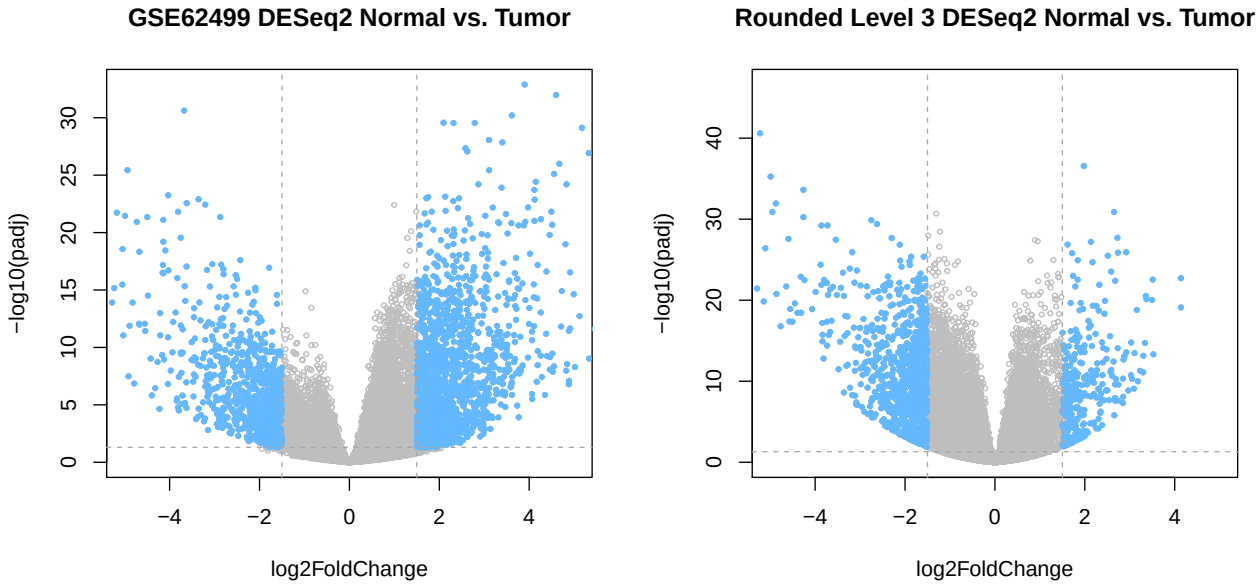

Figure 2: Volcano Plots of Level 3 and GSE62499 Differential Expression from Matched Normal to Tumor Tissue Using DESeq2. Each dot on the plot represents a single gene. Statistically significant ($|\text{logFC}| > 1.5$; FDR < 0.05) genes are highlighted in light blue.

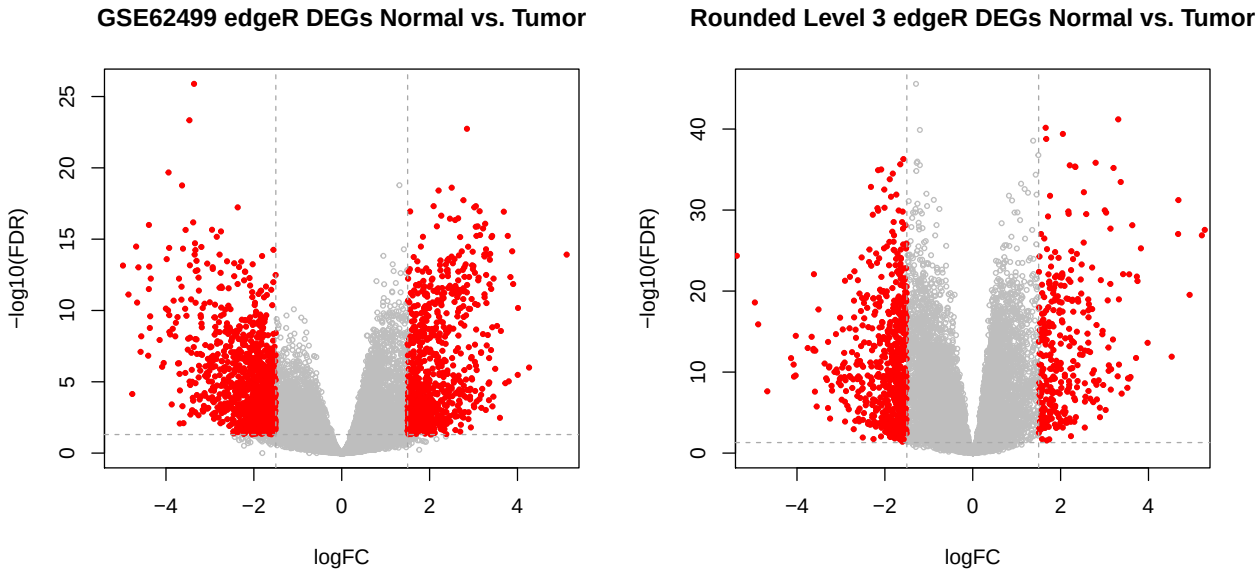

Figure 3: Volcano Plots of Level 3 and GSE62499 Differential Expression from Matched Normal to Tumor Tissue Using edgeR. Each dot on the plot represents a single gene. Statistically significant ($|\text{logFC}| > 1.5$; FDR < 0.05) genes are highlighted in red.

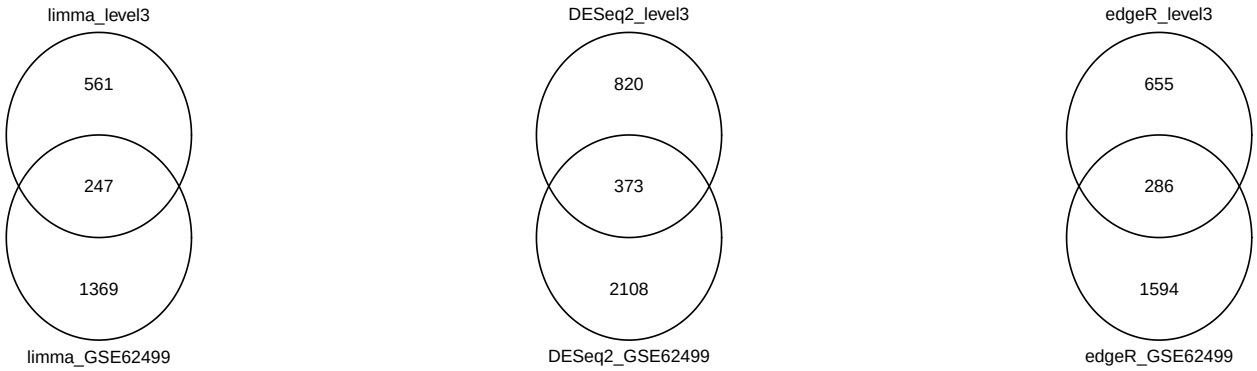

Figure 4: Venn Diagrams of Statistically Significant Genes Detected From GSE62499 vs. Level3 using limma (left panel), DESeq2 (middle panel), edgeR (right panel)

Figure 5: Venn Diagrams of comparing statistically significant DEGs using limma, edgeR, and DESeq2 within the same dataset. GSE62499 data (left panel), Level 3 data (right panel)

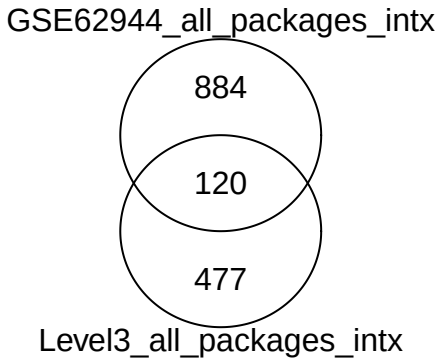

Figure 6: Venn diagram overlapping statistically significant DEGs from level 3 and GSE62499 data intersection between all packages

References

1. Rahman M, Jackson LK, Johnson WE, Li DY et al. Alternative preprocessing of RNA-Sequencing data in The Cancer Genome Atlas leads to improved analysis results. *Bioinformatics* 2015 Nov 15;31(22):3666-72. PMID: 26209429
2. M. I. Love, W. Huber, S. Anders: Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* 2014, 15:550.
3. Robinson, MD, McCarthy, DJ, Smyth, GK (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140.
4. Ritchie, ME, Phipson, B, Wu, D, Hu, Y, Law, CW, Shi, W, and Smyth, GK (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research* 43(7), e47.