BIOL437 Exercise on Transcriptomics

Shweta Ramdas

November 12, 2019

All files for this are on Canvas, or at https://github.com/shwetaramdas/biol437 You will need two files 1) GTEX_readcounts.txt, and 2) genelengths.txt in the same folder where you're running your R code.

Analysis of RNA-Sequencing data in R

The GTEx project has measured gene expression in different tissues obtained from many different people. We look at read counts for almost 10000 transcripts (in columns) transcripts from 1000 samples (in rows) (these may represent fewer than 1000 individuals, because each individual may have multiple tissues sampled). Individual IDs are in the format 'GTEX-1117F'

Part 1. Pre-processing data

We first read in the data, and see how many individuals are in our dataset. The first column of the data frame contains individual IDs.

```
gtex = read.table("GTEX_readcounts.txt",header=T,sep="\t",stringsAsFactors=F)
dim(gtex) #dim = dimension, giving you the number of rows and columns

length(unique(gtex[,1]))
```

How many individuals are in this dataset? The second column contains the tissue name. How many tissues have been sampled?

1. Genes that have zero read counts across all samples are not informative for our analysis. How many genes are not expressed in any sample?

```
expressions = gtex[,-c(1,2)]
sampleids = gtex[,c(1,2)]
maxes = apply(expressions, 2, max)
#this function calculates the maximum read count per sample
```

We take out those genes from the expressions data frame

```
length(which(maxes == 0))
expressions = expressions[,-which(maxes == 0)]
```

2. In an ideal (and usually, theoretical) experiment, all samples would have the same amount of mRNAs sequenced. In practise, there can be variation in the total amount of RNA sequenced per sample, maybe because the grad student pipetted a little more RNA from one sample. We can measure this by the total number of reads per sample.

```
totalreadcounts = apply(expressions, 1, sum)
plot(sort(totalreadcounts))
```

3. We will convert read counts to transcripts per million (TPM) before we cluster samples

```
#this is a vector of gene lengths for all our initial genes
genelengths = read.table("genelengths.txt",stringsAsFactors=F)
genelengths = genelengths[-which(maxes == 0),] #removing those genes with zero counts

norm1 = sweep(expressions, 2, genelengths[,2], "/")
tpm = apply(norm1, 1, function(x){x*1000000/sum(x)})
tpm = t(tpm) #tranposing to get back to the samples x genes format
```

Part 2. Interpreting biology

1. Clustering samples by PCA.

Colour the samples by 1) individual id 2) tissue type to ask if samples are more likely to cluster by individual or by tissue. I.e., if we sample the lung and the liver from person A and person B to get four samples: LungA, LiverA, LungB and LungB, do we see LungA cluster with Liver A? Or LungA with LungB?

We will use PCA to visualize distances between samples. Distances in the top two principal components can be viewed as a metric for distances between samples.

```
#Running the PCA
pca = prcomp(tpm)

#Naming variables so we can have a different symbol for each tissue in the PCA plot
tissues=sort(unique(sampleids[,2]))
col=rep(1:9,times=3)
sym=rep(1:9,each=3)
names(col)=tissues
names(sym)=tissues

#pca, first coloring by tissue
plot(pca$x[,1:2], col = col[sampleids[,2]], pch=sym[sampleids[,2]], xlab="PC1", ylab="PC2")
legend( "bottomright", tissues, col = col, pch=sym, cex=0.75, ncol=2)

#now coloring by individual ID
plot(pca$x[,1:2], col = as.factor(sampleids[,1]), xlab="PC1", ylab="PC2")
```

2. PARK7 is a gene (ensembl ID ENSG00000116288.12) shown to be differentially expressed in the brains of patients with Parkinson's Disease. The brain is the tissue we're interested in, but blood is a tissue easier to access and measure gene expression in. Is this gene expressed in blood? And specifically, is there a significant difference in expression levels between brain and blood?

```
#First, identify rows in the matrix corresponding to 'Blood' and Brain
bloodrows = which(gtex[,2] == 'Blood')
brainrows = which(gtex[,2] == 'Brain')

#Find the column containing expression values for gene ENSG00000116288.12
colnum = which(colnames(tpm) == "ENSG00000116288.12")
```

```r
#Now use a T-test to test for differences between expression in blood and brain. What is the P-value?
t.test(x = tpm[bloodrows, 236], y = tpm[brainrows, 236])
```

How would you interpret the above result? Can blood be used as a reasonable stand-in for brain tissue to test for biomarkers for Parkinson's?

3. We don't have sample information in this dataset. However, some genes are known to be sex-specific. One of these is Xist–a gene expressed on the X chromosome, that is only expressed in females. Xist (ENSEMBL Id: ENSG00000229807.10) is one of those genes. Can we plot the expression of Xist in each sample to assign their sex?

```r
#Plotting the expression of Xist for each sample, coloring by sample ID
plot((tpm[,'ENSG00000229807.10']), col=as.factor(sampleids[,1]), xlab="Index", ylab="TPM")

#To make the separation between the clusters more distinct, plot the logged values
plot(log10(tpm[,'ENSG00000229807.10']), col=as.factor(sampleids[,1]), xlab="Index", ylab="TPM")

#Now use a cutoff to differentiate likely males from likely females
```