Further Analysis of RNA-seq data

Mark Dunning

Last modified: March 11, 2016

Contents

1 Clustering

1.1 Example from the DESeq2 Vignette

The general application of clustering to RNA-seq data is outlined by the DESeq2 vignette

```
library("pasilla")
library("Biobase")
data("pasillaGenes")
countData <- counts(pasillaGenes)
colData <- pData(pasillaGenes)[,c("condition","type")]
```

```
library(DESeq2)
dds <- DESeqDataSetFromMatrix(countData = countData,
colData = colData,
design = ~ condition)
dds
```

```
featureData <- data.frame(gene=rownames(pasillaGenes))
(mcols(dds) <- DataFrame(mcols(dds), featureData))
```

```
dds <- estimateSizeFactors(dds)
```

```
library("pheatmap")
select <- order(rowMeans(counts(dds,normalized=TRUE)),decreasing=TRUE)[1:20]
```

```
nt <- normTransform(dds) # defaults to log2(x+1)
log2.norm.counts <- assay(nt)[select,]
df <- as.data.frame(colData(dds)[,c("condition","type")])
```

```
pheatmap(log2.norm.counts, cluster_rows=FALSE, show_rownames=FALSE,
cluster_cols=FALSE, annotation_col=df)
```

```
rld <- rlog(dds,blind = TRUE)
vsd <- varianceStabilizingTransformation(dds)
head(assay(rld), 3)
```

```
sampleDists <- dist(t(assay(rld)))
sampleDists
```

```
library("RColorBrewer")
sampleDistMatrix <- as.matrix(sampleDists)
rownames(sampleDistMatrix) <- paste(rld$condition, rld$type, sep="-")
colnames(sampleDistMatrix) <- NULL
colors <- colorRampPalette( rev(brewer.pal(9, "Blues")) )(255)
pheatmap(sampleDistMatrix,
clustering_distance_rows=sampleDists,
clustering_distance_cols=sampleDists,
col=colors)
```

```
plotPCA(rld, intgroup=c("condition", "type"))
```

1.2 Re-visiting our ESCC dataset

```
library(DESeq2)
load("Day2/Counts.RData")
#Load data
Counts <- tmp$counts
colnames(Counts) <- c("16N", "16T", "18N", "18T", "19N", "19T") #Rename the columns
Coldata <- data.frame(sampleReplicate=c("16", "16", "18", "18", "19", "19"),
sampleType=c("N", "T", "N", "T", "N", "T"))
rownames(Coldata) <- c("16N", "16T", "18N", "18T", "19N", "19T")

deSeqData <- DESeqDataSetFromMatrix(countData=Counts, colData=Coldata,
              design= ~sampleReplicate + sampleType)
deSeqData
```

```
nrow(counts(deSeqData))
summary(rowSums(counts(deSeqData)))
deSeqData <- deSeqData[rowSums(counts(deSeqData))>1,]
```

```
deSeqData <- estimateSizeFactors(deSeqData)
colData(deSeqData)
head(counts(deSeqData))
```

```
library(genefilter)
nt <- normTransform(deSeqData) # defaults to log2(x+1)
select <- order(rowVars(counts(deSeqData,normalized=TRUE)),decreasing=TRUE)[1:20]

log2.norm.counts <- assay(nt)[select,]
df <- as.data.frame(colData(deSeqData)[,c("sampleReplicate","sampleType")])
pheatmap(log2.norm.counts, cluster_rows=FALSE, show_rownames=FALSE,
cluster_cols=FALSE, annotation_col=df)
```

```
pheatmap(log2.norm.counts, cluster_rows=FALSE, show_rownames=FALSE,
cluster_cols=TRUE, annotation_col=df)
```

```
plotPCA(nt, intgroup=c("sampleType","sampleReplicate"))
```

```
pcData <- plotPCA(nt, intgroup=c("sampleType","sampleReplicate"),returnData=TRUE)
pcData
```

```
library(ggplot2)
percentVar <- round(100*attr(pcData, "percentVar"))
ggplot(pcData, aes(x=PC1,y=PC2,color=sampleType,shape=sampleReplicate))+
  geom_point(size=5)+
  xlab(paste0("PC1: ", percentVar[1], "% variance")) +
  ylab(paste0("PC2: ", percentVar[2], "% variance"))
```

```
sampleDists <- dist(t(assay(rld)))
sampleDists

sampleDistMatrix <- as.matrix(sampleDists)
rownames(sampleDistMatrix) <- paste(rld$sampleReplicate, rld$sampleType, sep="-")
colnames(sampleDistMatrix) <- NULL
colors <- colorRampPalette( rev(brewer.pal(9, "Blues")) )(255)
pheatmap(sampleDistMatrix,
clustering_distance_rows=sampleDists,
clustering_distance_cols=sampleDists,
col=colors)
```

```
deSeqData <- estimateDispersions(deSeqData)
mcols(deSeqData)
deSeqData <- nbinomWaldTest(deSeqData)
res <- results(deSeqData)
```

```
res.sig <- res[which(res$padj < 0.05),]
N <- 100
res.sig.ord <- res.sig[order(res.sig$padj,decreasing = FALSE),]
topNGenes <- rownames(res.sig.ord)[1:N]
```

```
pheatmap(assay(nt)[match(topNGenes, rownames(assay(nt))),],annotation_col=df)
```