RNA-seq data analysis

Juan R Gonzalez

BRGE - Bioinformatics Research Group in Epidemiology
Center for Research in Environmental Epidemiology (CREAL)
http://brge.isglobal.org

EXERCISE 1. Library `tweeDEseqCountData` contains data corresponding to an RNA-seq experiment described in Pickrell et al. (2010). Data correspond to lymphoblastoid cell lines about 69 non-related Nigerian individuals. This information as well as phenotypic data is available as an object of class `eSet` that can be loaded after typing:

```
data(pickrell)
```

Annotation (from ENSEMBL and including gene length and GC-content) is available as a *data.frame* that can be loaded by executing:

```
data(annotEnsembl63)
```

a) Select those genes with available gene information (GC-content and gene length from `annotEnsembl63`). For those selected genes:

b) Obtain normalized counts using RPKM method

c) Obtain normalized counts using TMM method

d) Obtain normalized counts using 'GC-content' and 'gene length' information and CQN method

e) Create an MA-plot using each type of normalization and describe the main differences among them

EXERCISE 2. In this exercise we are going to compare the gene expression levels between males and females for a given list of genes (in order to reduce computing time). This list of genes (`geneSubset`) can be obtained by executing:

```
library(tweeDEseqCountData)
data(pickrell)
data(hkGenes)
data(genderGenes)
countsNigerian <- exprs(pickrell.eset)
geneSubset <- unique(c("ENSG00000070031",
                       intersect(rownames(countsNigerian),
                       c(hkGenes, msYgenes, XiEgenes))))
```

This list of genes contains some Housekeeping Genes (genes that should not be differentially expressed between males and females) some genes that are sex-specific selected from chromosome X, genes that scape to X-inactivation (Carrel and Willard, 2005) and some genes from the male-specific region from chromosome Y (Skaletsky et al., 2003).

a) Get normalized counts using TMM method

b) Select count data table of those genes that are in the list `geneSubset`

c) Determine those genes that are differentially expressed between males and females by using `edgeR` `DESeq` `tweeDEseq`.

d) Perform DE analysis using `voom` method. NOTE: Select only p-values for those genes that are in the list `geneSubset`

e) Are there differences among the four methods? Which is the best method? (NOTE: only those gene declare as sex-specific are to be expected as a differentially expressed)