# Applied genomics homework 3: RNA-seq

(equal marks for each question)
Due Wednesday 3/7/2018.

**Q1.** An RData file is provided which has the raw counts and gene lengths (length of CDS) for the arabidopsis data used in the RNA-seq lab.

Write an R function to compute the TPM expression level of each gene.

**Q2.** Counts for a simulated small RNA-seq dataset are provided in "smallRNA_simulated.RData". Matrix A stores the counts over 1000 microRNAs (rows), in two conditions (columns). Biologically, the microRNA expression is identical in the two conditions except the first microRNA which has a drastically increased expression under the condition 1.
Note that the total read counts for each column is the same (1E6 reads) representing the fixed sequencing capacity of the high-throughput sequencing assay used.

By reference to the paper Robinson, M. and Oshlack, A. "A scaling normalization method for differential expression analysis of RNA-seq data" (in readings in classes), describe the "real-estate effect" and why the expression of most microRNAs appear to be different between the two conditions in the unnormalised raw data given.

Apply a TMM normalization to the data and compare the expression levels after. Did it improve the normalization, given the biological assumption that most microRNAs had similar expression between conditions?

**Q3.** Starting with the Rmarkdown file from the RNA-seq lab "limma_interaction_RNAseq_contrasts_GSEA7.Rmd", expand on my comments.

In particular, list and number at the start of your document the major statistical issues for HTS we covered in lectures, and for each relevant line or section of the pipeline describe which statistical issue is involved, referencing this list. (The issues are: (1) Poisson count noise, (2) Moderated t-test (3) Normalisation (4) Multiple testing correction).

Then describe each statistical issue in a paragraph in your own words, referring to the relevant subsections of Conesa et al "A survey of best practices for RNA-seq data analysis" and Ritchie et al "limma powers differential expression analyses for RNA-sequencing and microarray studies" in your answer.