A quick guide to our RNA-seq Workflow

Earlham Bioinformatics

based on Anders et al. (2013, Nature Protocols, 8: 1765-1786)

Wet lab:

1) Obtain Sample; 2) Isolate RNA, 3) Reverse Transcribe, 4) Barcode, 5) Sequence

Each sample to be sequenced is referred to as a "library" (of, originally, RNA)

Raw Data:

1) Download the raw fastq files (either from database such as SRA or sequencing facility.
    - For this we are using the NCBI SRA database (SRA = short read archive)
    - Short reads, such as from Illumina are used for gene expression and resequencing.
    - Using SRAtools, download using fastq-dump
    - Library IDs can be downloaded from the SRA website.
    - Published papers should have SRA or other accession information to help you find the data.
2) Download the relevant genome and GFF
    - The GFF is necessary as you are assigning the reads to features, such as exons.
    - The genome needs to be indexed for faster computing using the bowtie2 program.
3) Map the reads to the genome.
    - This is done using the tophat2 program.
4) Organize the data
    - The mappings are organized using commands in SAMtools.
    - The organization is necessary for viewing the information using genome viewers, such as IGV.
    - It is a best practice to look at how your data map to features in the genome, such as with IGV, by loading your genome, GFF, and a set of mappings (bam file).
        - You don't have to check everything, just spot check.
    - The indexing of the reads will allow for a more rapid counting of reads per feature.
5) Count the reads per feature.
    - This is done using the htseq-count algorithm.
    - This will ultimately result in a table with each feature/gene as a row and the samples (libraries) as columns.
    - The count data is a matrix that can then be normalized and analyzed.
    * This format of data should now be old hat....

Differential Expression:

1) We will use the R library "edgeR" for this step.
2) The data structure of edgeR objects is a set of lists, with the matrix of counts being among them, but also included are groups and genes, among other things.
3) The first thing to do is a preliminary normalization - convert raw counts to relative counts (counts per million, or CPM).
4) Next discard genes with too few reads. The rule of thumb is:
    - Get rid of the genes without at least one CPM in at least the quantity of samples that has the lowest number of samples.
    - For example: you have two treatments, +nuts and nothing added. There are 4 replicates in each of those groups. So, you would say there needs to be at least 1 CPM in at least 4 samples.
5) Normalize the data again in a much more complicated way, called TMM (trimmed-mean of M-values).
    - In brief, this normalizes the samples based on library size and among library variance.
6) Analyze the data using an linear model (like an ANOVA).
7) Make graphs like ordination plots and heat maps, as well as smear and volcano plots.