FungiDB: Background on differential expression analysis

Differential Expression Analysis is comparative analysis of transcript abundance

- Assess **quantitative** differences between conditions
- Simplest form – Single-Factor Experiment comparing two conditions
    - E.g wild type vs mutant
- Multi-Factor Experiments comparing multiple conditions / samples
    - E.g Plants infected with fungal pathogen – samples grown at different temperature and with a protective chemical treatment or not.

Experimental design

- Replication – best practice is at least 3 biological replicates which allows for correction of within sample variation
- Biological question – does the design allow you to fully test your hypothesis, do you have the correct controls

Basic statistics that are needed to understand the results

P-value and False discovery rate (FDR) adjusted p values:
A p-value of 0.05 implies that 5% of all tests will result in false positives. An FDR adjusted p-value (or q-value) of 0.05 implies that 5% of significant tests will result in false positives. The latter will result in fewer false positives.

Fold Change:
Fold change is a measure describing how much a quantity changes going from an initial to a final value. For example, an initial value of 30 and a final value of 60 corresponds to a fold change of 1 (or equivalently, a change to 2 times), or in common terms, a one-fold increase.

Types of errors produced by Differential Expression Analysis tools

Type 1 – false POSITIVE:
In a type I error (or false-positive) the null hypothesis is really true (the gene is not differentially expressed) but the statistical test has led you to believe that it is false (there is a difference in expression). This type of error is potentially very dangerous, if a rejected hypothesis allows publication of a scientific finding, a type I error brings a “false discovery”, and the risk of publication of a potentially misleading scientific result.

Type 2 – false NEGATIVE:

In a type II error (or false-negative) the null hypothesis is really false (the gene is differentially expressed) but the test has not picked up this difference. This type of error is less dangerous than the type I but should still be avoided if possible.

Tools involved in the RNA-seq pipeline in the pre-configured workflows available within the EuPathDB Galaxy Instance

1. **Quality Control and Trimming**

FastQC - https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

- Allows for quality control in raw sequence reads
- Can take in BAM, SAM or FastQ files
- Quick overview of problems
- Results in an HTML report

Sickle - https://github.com/ucdavis-bioinformatics/sickle

- Uses sliding windows and quality and length thresholds to trim both 3' and 5' ends of reads
- Can take input from Illumina, Solexa and Sanger sequencing
- Single or paired-end reads

Trimmomatic - http://www.usadellab.org/cms/?page=trimmomatic

- Illumina specific
- Uses sliding window, length and quality thresholds
- Adaptor trimming
- Phred-33 or Phred-64 quality
- FastQ files – compressed or uncompressed

2. **Alignment tools**

GSNAP - http://research-pub.gene.com/gmap/

- Align single or paired-end reads
- can detect short- and long-distance splicing
- permits SNP-tolerant alignment to a reference space of all possible combinations of major and minor alleles.
- can align reads from bisulfite-treated DNA for the study of methylation state.

TopHat2- https://ccb.jhu.edu/software/tophat/index.shtml

- spliced aligner
- can align across fusion breaks
- combines identification of novel spice sites and direct mapping to known transcripts

3. **Differential Expression Tools/Pipelines**

a. Cufflinks suite -- http://cole-trapnell-lab.github.io/cufflinks/

Cufflinks

- Aligned RNA-seq reads as input
- Assembled transcripts

- Estimates their abundances

Cuffmerge

- Combines multiple assemblies to form a master transcriptome

Cuffdiff

- Compares expression levels of genes and transcripts
- can determine which genes are up- or down-regulated between two or more conditions.
- can determine which genes are differentially spliced or are undergoing other types of isoform-level regulation.

b. HTSeq-counts followed by DESeq2

HTSeq-counts - http://htseq.readthedocs.io/en/release_0.9.1/count.html

- *determine counts – how many reads map to a genomic feature – gene, exon etc*

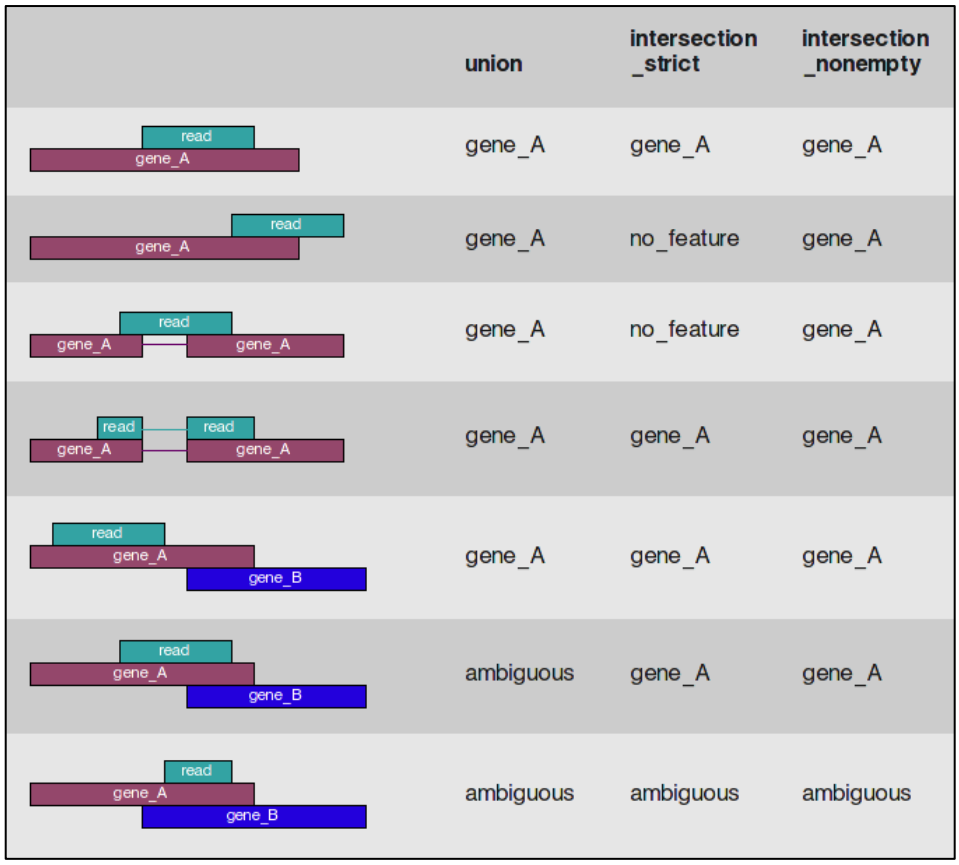

Figure 1: HTSeq-count options for the overlap resolution modes

DESeq2- https://bioconductor.org/packages/release/bioc/html/DESeq2.html

- comparison of expression between features
- takes count data as input

- can deal with pairwise or multiple comparisons
- requires replicates
- based on negative binomial models
- Uses shrinking estimations for dispersion and fold change
- Initial gene-wise dispersion is estimated using maximum likelihood
- A smooth curve is fitted to the data to capture trend of estimates based on average expression strength
- This is then used as a guide for a second round of estimations that shrink the noisy gene-wise estimates towards the mean
- In previous studies this method has been shown to be more conservative than the other two methods

Structure of a FASTQ file

FASTQ file are text files (similar to FASTA) that include sequence quality information and details in addition to the sequence (ie. name, quality scores, sequencing machine ID, lane number etc.). FASTQ files are large and as a result not all sequencing repositories will store this format. However, tools are available to convert, for example, NCBI's SRA format to FASTQ. Sequence data is housed in three repositories that are synchronized on a regular basis.

- The sequence read archive at GenBank
- The European Nucleotide Archive at EMBL
- The DNA data bank of Japan

FPKM

FPKM (fragments per kilobase of exon model per million reads mapped) is a normalised estimation of gene expression based on RNA-seq data. FPKM are calculated from the number of reads that mapped to each particular gene sequence taking into account the gene length (one expects more reads to be produced from longer genes) and the sequencing depth (one expects more reads to be produced from the sample that has been sequenced to a greater depth).

RPKM

RPKM stands for “Reads Per Kilobase of transcript per Million mapped reads”. This formula is commonly used to normalise next-generation sequencing data, accounting for varying sequencing depth (number of total reads per sample) and transcript length (more reads mapping to a longer genomic region).

Reference: https://www.ebi.ac.uk/training/online/glossary