INF-BIOx121 2017

RNA-seq

differential expression analysis

Arvind Sundaram
Sep 18-20, 2017

RNA-seq analysis

Differential expression (DE)

Arvind Sundaram

Sep 19, 2017

Differential expression

- Genes
- Transcripts (Isoforms)
- Allele specific expression
- Exon level expression

Counting

- Feature - genes, transcript or exon
- How many reads aligned to each feature of interest?
- What is the length of the feature?

- Raw count calculated from BAM files using featureCounts, HTSeq, etc
- Most (all) DE tools would require raw count file and not (pre) scaled data.

Normalisation

- Normalistion within and across samples
- Count gets converted to RPKM, FPKM or TPM
- RPKM (Reads Per Kilobase Million)

  Scaling factor = Total number reads / 1,000,000

  RPM = Read count per feature / scaling factor

  RPKM = RPM / Feature length in kilo bases
- FPKM (Fragments Per Kilobase Million)

  FPM = Fragment count per feature / scaling factor

  FPKM = FPM / Feature length in kilo bases
- TPM (Transcripts Per Kilobase Million)

  RPK = Read count per feature / Feature length in kilo bases

  Scaling factor = sum of RPK / 1,000,000

  TPM = RPK / Scaling factor

DESeq2 (or edgeR) is different!!

https://www.youtube.com/watch?v=UFB993xufUU

https://www.youtube.com/watch?v=TTUrtCY2k-w

DESeq2

- Generalised linear model fit
  - Using negative binomial distortion (aka gamma-Poisson distribution)
- Empirical Bayes shrinkage
  - for within-group variability, i.e., variability between replicates
- Fold change estimation
- Not just pair-wise comparison. Allows for complicated nested designs to be compared

DESeq2

- plotDispEsts(): To look at the dispersion plots
- plotPCA(): To find outliers
- plotMA(): Exploring DE results

Multiple hypothesis testing and FDR

Multiple hypothesis testing

- Thousands of genes = thousands of hypothesis tests (simultaneously)
- Increased chance of false positives! (Type I error)
  - e.g. you test for differential expression in 1000 genes that are not differentially expressed
  - You would expect 1000 x 0.05 = 50 of them to have a *P*-value < 0.05
- Individual *P*-values not useful : Need multiple testing statistic instead

False Discovery date (Benjamini & Hochberg 1995)

- The expected proportion of Type I errors among the rejected hypotheses
  - i.e. the proportion of false positives
  - Tends to be conservative if many genes are DE
    - FDR = 0.05 common for exploratory/broad scope studies
    - FDR < 0.05 common for medical applications and hunts for candidate genes