Gene set testing in *limma*

COMBINE RNA-seq Workshop
Why?

- Sometimes after differential expression testing, we have a long list of 1000’s of genes
- Too difficult to go through one by one
- Or there may be very few / no genes that make statistical significance (small effect sizes + experimental noise)
- Want to understand pathways involved in the biological system being studied
Gene set tests available in *limma*

- Want to test **LOTS of gene sets**?
  - `goana()` function
    - Test Gene Ontology (GO) categories
  - `kegga()` function
    - Test KEGG pathways
  - `camera()` function
    - User specified gene sets

- Want to test just a **few gene sets**?
  - `mroast()` / `fry()` functions
Basic principles behind
gene set testing
“Overlap” analysis: goana, DAVID, ToppFun, GOstats (& most web-based tools)

190 genes in geneset

180

10

60

Is an overlap of 10 significant?

70 significant genes
Problem: this test is biased due to the fact that longer genes tend to have more reads assigned to them.

GO categories have different avg gene lengths

GOseq, Young et al, 2010
Solution: take into account gene length in your GO analysis

- `goana()` has the ability to take into account gene length using the “covariate” argument
- The GOseq bioconductor package contains the original method
CAMERA

• An “overlap” analysis assumes the genes are independent

• CAMERA tests the ranking of the gene set relative to the other genes in the experiment, while taking into account inter-gene correlations

• It also takes into account strength of evidence of DE by using the moderated $t$-statistics
Rank genes and mark signature

Rank genes by differential expression

Negative signature genes

Positive signature genes

Gene 1
Gene 2
Gene 3
Gene 4
Gene 5
Gene 6
Gene 7
Gene 8
Gene 9
Gene 10
Gene 11
Gene 12
Gene 13
Gene 14
Gene 15
Gene 16

Slide courtesy of Gordon Smyth
Rank genes and mark signature

Rank genes by differential expression

Gene 1
Gene 2
Gene 3
Gene 4
Gene 5
Gene 6
Gene 7
Gene 8
Gene 9
Gene 10
Gene 11
Gene 12
Gene 13
Gene 14
Gene 15
Gene 16

Genome-wide barcode plot

Slide courtesy of Gordon Smyth
Visualisation:
Barcodeplot + enrichment worm

Red bars: GMP to neutrophil upregulated genes

Blue bars: GMP to neutrophil downregulated genes

increased upon PU.1 restoration

decreased upon PU.1 restoration

Enrichment

6.3

5.7

Data courtesy of Mark McKenzie
Gene signature collections

The Molecular Signatures Database (MSigDB) is a collection of gene sets for use with GSEA software. From this website, you can:

- Search for gene sets
- Browse gene sets
- View annotations by clicking a gene set name to display its gene set page; for example, AKTPATHWAY
- Download gene sets
- Compute overlaps between your gene set and other gene sets in MSigDB
- Categorize members of a gene set by gene families
- Build an expression signature of the gene set using a compendium of expression profiles

Registration

Please register to download the GSEA software and view the MSigDB gene sets. After registering, you can log in at any time using your email address. Registration is free. Its only purpose is to help us track usage for reports to our funding agencies.

Current Version

GSEA/MSigDB website v2.0 released December 14 2007
MSigDB database v2.5 updated April 7 2008, Release notes.
ROAST gene set test

• The question asked is “Do the genes in this gene set tend to be differentially expressed?”
• It is NOT compared relative to other genes
• It is designed such that if > 25-50% of genes in the gene set are differentially expressed it will be significant
• It uses sophisticated techniques (rotation) to preserve gene-gene dependence in the data.
• fry is a fast implementation of roast that assumes constant gene-wise variance
Summary

• Gene set testing techniques range from simple (overlap analysis) to quite complex (CAMERA and ROAST)

• Which test you choose depends on what your hypothesis is

• Sometimes we just do them all...
Acknowledgements

• Gordon Smyth
• Belinda Phipson