Pathway Analysis

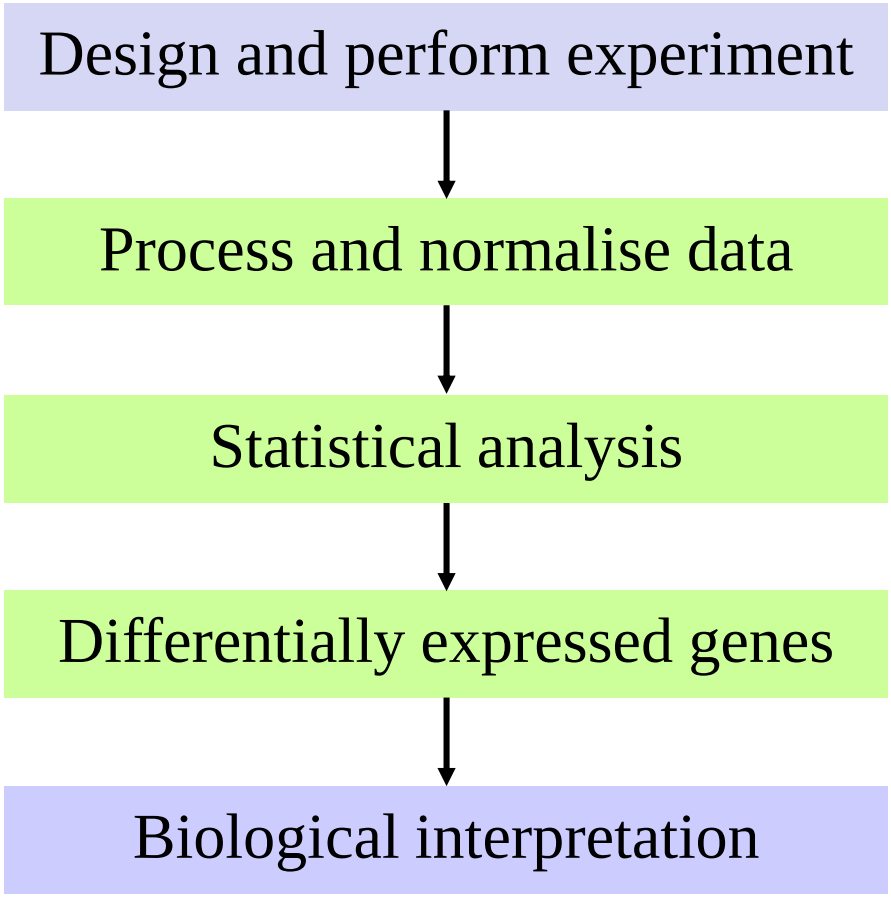
Concepts and Main Aspects of RNA-Seq

28th April 2016

Helen Lockstone

Head of Functional Data Analysis, WTCHG

Interpreting a gene expression study



- Getting from a gene list to biological insight – obvious way to do this is look at differentially expressed genes in terms of their known function(s)
- Known as functional profiling or pathway analysis

Early functional analyses

- Manually annotate list of differentially expressed (DE) genes
- Extremely time-consuming, not systematic, user-dependent
- Group together genes with similar function
- Conclude functional categories with most DE genes important in disease/condition under study
- BUT may not be the right conclusion

Biological Interpretation

- Looking for *enrichment* or *over-representation* of particular categories of genes among the set that are differentially expressed (DE)

- Requires ALL genes to be annotated and a statistical assessment of the number of genes for a given gene set in the list of DE genes compared to the number expected by chance (e.g. Fisher's exact test)

Major bioinformatic developments

- Required standardised and comprehensive annotations for all known genes (www.geneontology.org)

commentary

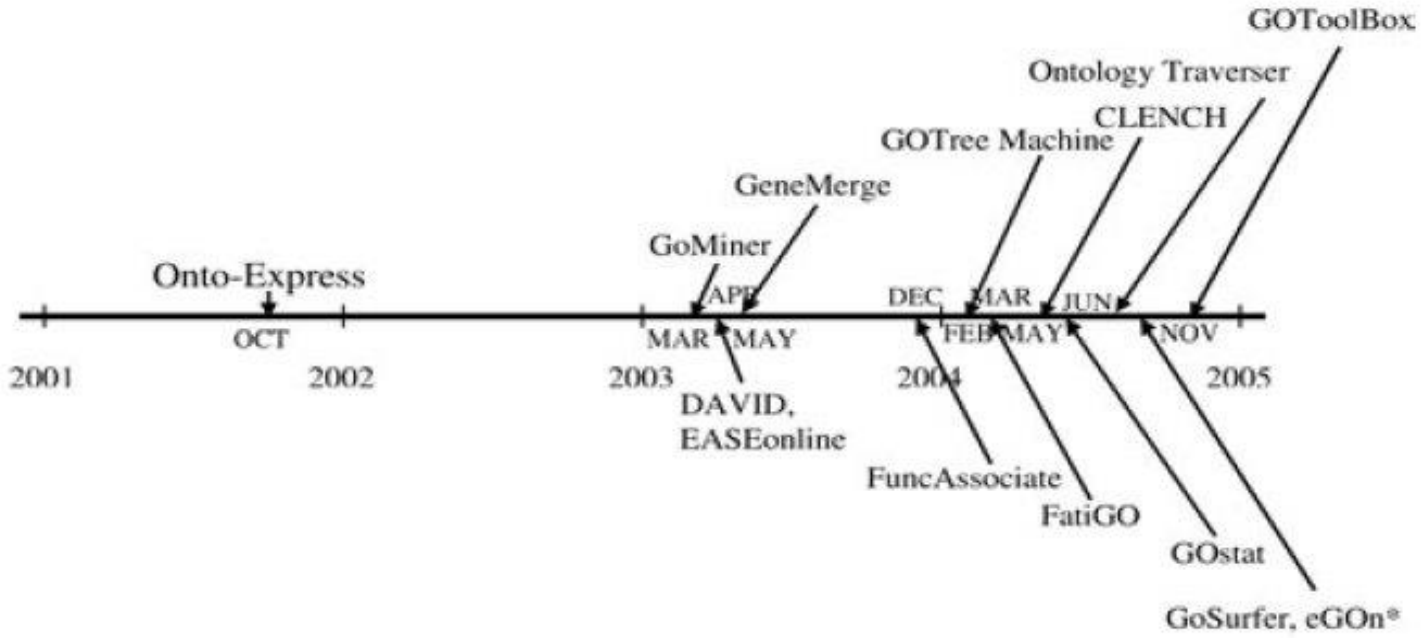
Gene Ontology: tool for the unification of biology

The Gene Ontology Consortium*

Genomic sequencing has made it clear that a large fraction of the genes specifying the core biological functions are shared by all eukaryotes. Knowledge of the biological role of such shared proteins in one organism can often be transferred to other organisms. The goal of the Gene Ontology Consortium is to produce a dynamic, controlled vocabulary that can be applied to all eukaryotes even as knowledge of gene and protein roles in cells is accumulating and changing. To this end, three independent ontologies accessible on the World-Wide Web (http://www.geneontology.org) are being constructed: biological process, molecular function and cellular component.

- Enabled the development of automated, statistical approaches for annotating gene lists and performing functional enrichment analysis

Functional profiling tools



Khatri and Draghici. Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics* (2005) 21(18):3587-95

Functional profiling tools

- Majority of tools based on idea of identifying GO categories significantly enriched in list of differentially expressed genes
  - Requires some threshold to define genes as ‘significant’
- Reduces a large number of DE genes to a smaller number of significantly enriched GO categories
  - more easily interpreted in biological context
- Freely-available stand-alone/web-based tools
  - User-friendly graphical interface and simple to use
  - Extensive documentation, plus tutorials/technical support

Further considerations

- Reference list must be appropriate for accurate statistical analysis
- Arrays usually have multiple probes per gene – don't count as multiple hits for a gene set as inflates enrichment statistic
- Unannotated genes cannot be used in the analysis; gene ontology evolving; well-studied systems over-represented
- Due to GO hierarchy, several related categories may contain a subset of genes that is driving the significant enrichment score so will all be significant

Defining gene sets

- GO categories are just one way to group functionally-related sets of genes
- Many other ways:
  - Other databases (KEGG pathways etc)
  - Target genes of transcription factors or miRNAs
  - Custom e.g. gene list from similar experiment

Gene set enrichment analysis

- GSEA (developed at Broad Institute) takes a different approach by considering all assayed genes
- Uses information from genes even if they do not meet standard significance criteria
- Ranks genes from up-regulated through unchanged through to down-regulated
- Can ‘trick’ the program if wanted to look at up and down-regulated genes simultaneously – rank genes by p-value so all at the top and only inspect the enrichment for the top of the list

GSEA: Key Features

- Ranks all genes assayed based on their differential expression
- Identifies gene sets whose member genes are clustered either towards top or bottom of the ranked list (i.e. up- or down regulated)
- Enrichment score calculated for each category
- Permutation test to identify significantly enriched categories
- Extensive gene sets provided via MolSig DB – GO, chromosome location, KEGG pathways, transcription factor or microRNA target genes

GSEA

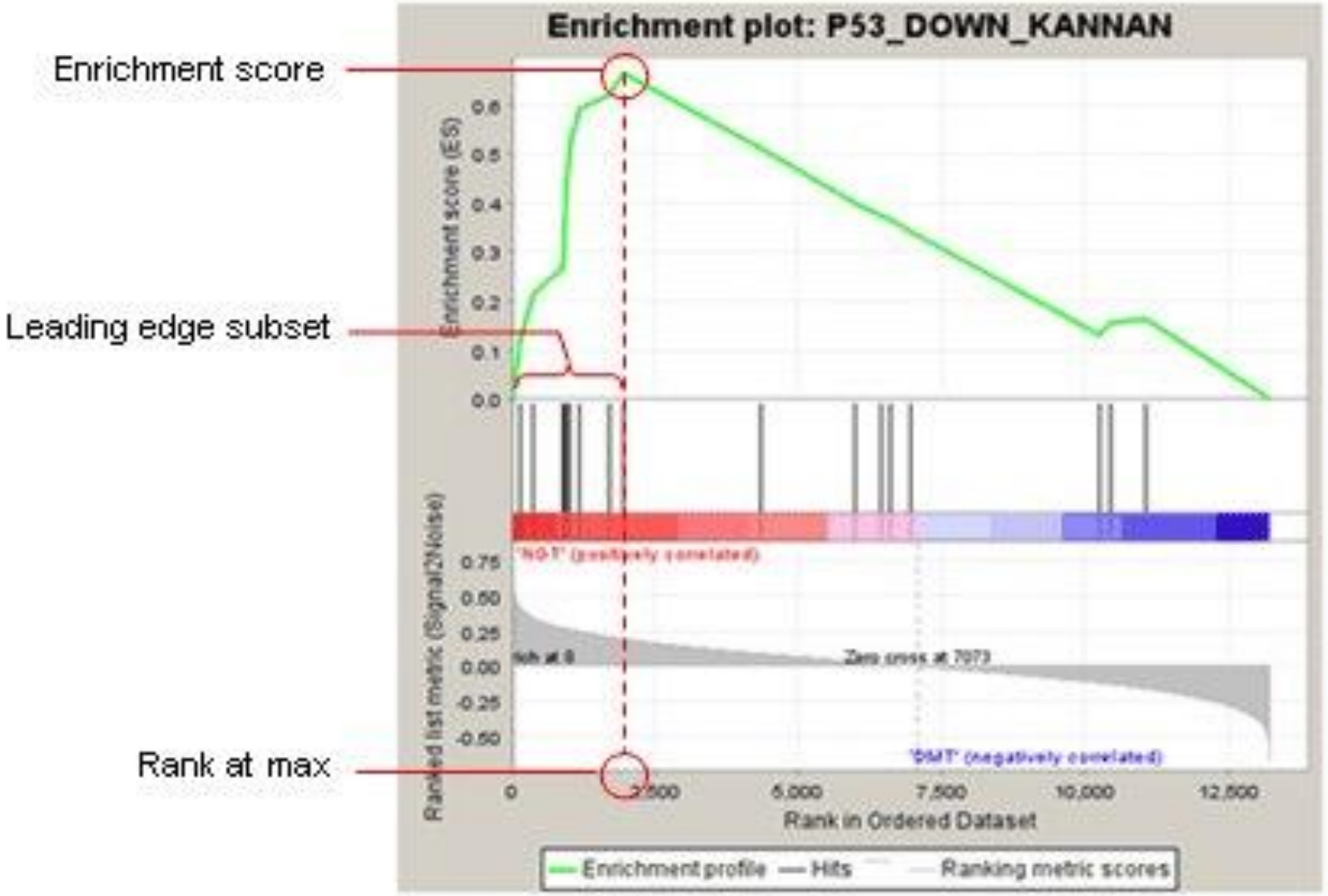
- Each gene category tested by traversing ranked list
- Enrichment score starts at 0, weighted increment when a member gene encountered, weighted decrement otherwise
- Enrichment score – point where most different from zero

Most significantly up-regulated genes

Unchanged genes

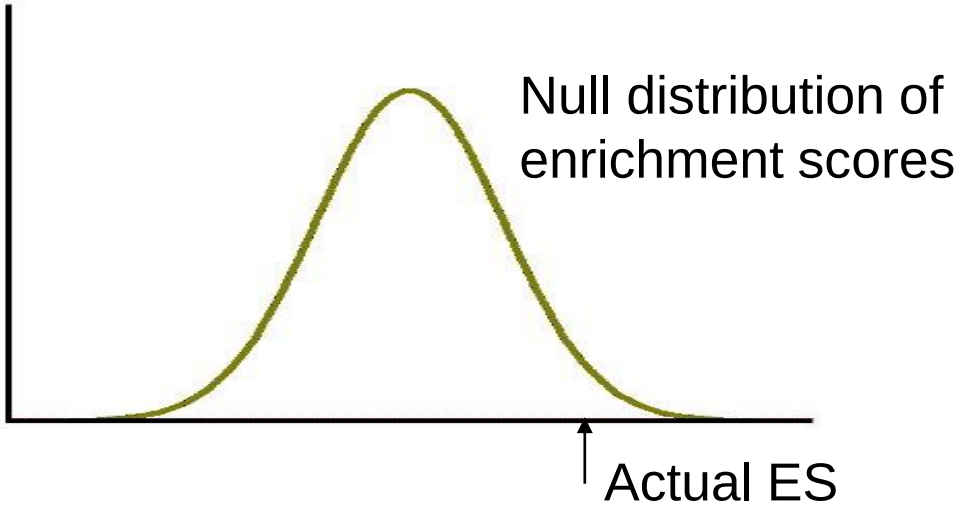
Most significantly down-regulated genes

GSEA algorithm



GSEA: Permutation Test

- Randomise data (groups), rank genes again and repeat test 1000 times
- Null distribution of 1000 ES for geneset



- FDR q-value computed – corrected for gene set size and testing multiple gene sets

Biological Interpretation

- Too many categories found significant
  - Size filter
  - More stringent significance threshold
  - Related categories (redundancy)
- No significant categories
  - Relax significance level slightly
  - e.g. 0.25 recommended by GSEA as exploratory analysis
- No significant genes in differential expression analysis
  - GSEA most suitable for exploratory work

Commercial Tool Suites

- Ingenuity Pathway Analysis (Ingenuity Systems, CA)
  - Developed own extensive ontology over past 10 years
  - Includes gene interactions, disease/drug information
  - PhD-level curators mining the literature
  - Used by many pharmaceutical companies

Enrichment analysis for RNA-Seq

- Another example of application where tools were originally developed for microarray data
- Many can accept RNA-Seq data but there may be caveats to be aware of:
  - Selection bias in RNA-Seq (longer/more highly expressed genes have higher counts and more likely to detect differential expression): GOSeq
  - Ranking metrics in GSEA were designed for continuous data – preranked mode may be better, however caveats to this as well
  - Recent review paper and active discussion on forums

Briefings in Bioinformatics, 2015, 1–15

doi: 10.1093/bib/bbv069
Paper

Gene set analysis approaches for RNA-seq data: performance evaluation and application guideline

Yasir Rahmatallah, Frank Emmert-Streib and Galina Glazko

Corresponding author. Galina Glazko, Department of Biomedical Informatics, University of Arkansas for Medical Sciences, Little Rock, AR 72205, USA. Tel: +1-501-603-1759, Fax: +1-501-526-5964.
E-mail: gvglazko@uams.edu

For more information

- Gene Ontology: http://www.geneontology.org
- GSEA: http://www.broad.mit.edu/gsea/
- Ingenuity: http://www.ingenuity.com/products/pathways_analysis.html
- GOSeq: http://bioconductor.org/packages/release/bioc/html/goseq.html
- Gene set analysis approaches for RNA-seq data: performance evaluation and application guideline: http://www.ncbi.nlm.nih.gov/pubmed/26342128