Lecture 8: Enrichment Testing

BIOINF3005/7160: Transcriptomics Applications

Dr Stephen Pederson

Bioinformatics Hub,
The University of Adelaide

May 11th, 2020

Databases

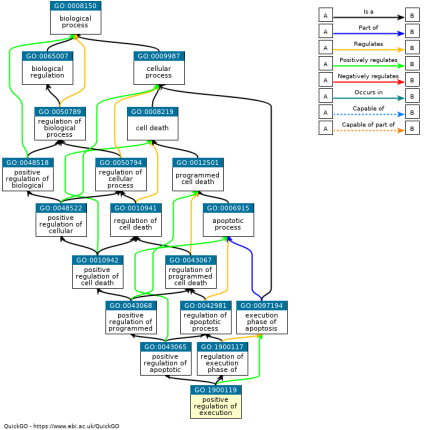
Introduction

- Once we have obtained results from our analysis:
  - How do we summarise the results for hundreds/thousands of genes?
- We look for biological patterns
- How do we even define biological patterns?
- We can use pre-existing databases with defined terms
  - GO, KEGG, Wiki Pathways, MSigDB, JASPAR etc
- We obtain pre-defined **gene sets** and test for enrichment in our dataset

The Gene Ontology Database

- The most commonly used resource for describing biology
  - Also one of the most frustrating
- Has a restricted vocabulary for describing biological features $\Longrightarrow$ GO Terms
- Multiple classification levels for assigning GO terms to genes
- The basic structure is as a Directed Acyclic Graph (DAG)

The Gene Ontology Database



The Gene Ontology Database

The three Ontologies

1. **Molecular Function**: *A molecular function is a process that can be carried out by the action of a single macromolecular machine, via direct physical interactions with other molecular entities*

[1] Paul D. Thomas. "The Gene Ontology and the Meaning of Biological Function". In: *The Gene Ontology Handbook*. Ed. by Christophe Dessimoz and Nives Škunca. New York, NY: Springer New York, 2017, pp. 15–24. ISBN: 978-1-4939-3743-1. DOI: 10.1007/978-1-4939-3743-1_2. URL: https://doi.org/10.1007/978-1-4939-3743-1_2.

The Gene Ontology Database

The three Ontologies

1. **Molecular Function**: *A molecular function is a process that can be carried out by the action of a single macromolecular machine, via direct physical interactions with other molecular entities*
2. **Cellular Component**: *A cellular component is a location, relative to cellular compartments and structures, occupied by a macromolecular machine when it carries out a molecular function*

[1] Paul D. Thomas. "The Gene Ontology and the Meaning of Biological Function". In: *The Gene Ontology Handbook*. Ed. by Christophe Dessimoz and Nives Škunca. New York, NY: Springer New York, 2017, pp. 15–24. ISBN: 978-1-4939-3743-1. DOI: 10.1007/978-1-4939-3743-1_2. URL: https://doi.org/10.1007/978-1-4939-3743-1_2.

The Gene Ontology Database

The three Ontologies

1. **Molecular Function**: *A molecular function is a process that can be carried out by the action of a single macromolecular machine, via direct physical interactions with other molecular entities*
2. **Cellular Component**: *A cellular component is a location, relative to cellular compartments and structures, occupied by a macromolecular machine when it carries out a molecular function*
3. **Biological Process**: *A biological process represents a specific objective that the organism is genetically "programmed" to achieve*

All definitions taken from Thomas (2017)[1]

[1] Paul D. Thomas. "The Gene Ontology and the Meaning of Biological Function". In: *The Gene Ontology Handbook*. Ed. by Christophe Dessimoz and Nives Škunca. New York, NY: Springer New York, 2017, pp. 15–24. ISBN: 978-1-4939-3743-1. DOI: 10.1007/978-1-4939-3743-1_2. URL: https://doi.org/10.1007/978-1-4939-3743-1_2.

The Gene Ontology Database

- Each GO term belongs exclusively to one Ontology
- Contains an ID, Name, Definition
- Browsing our term from the previous image:
  `https://www.ebi.ac.uk/QuickGO/term/GO:1900119`

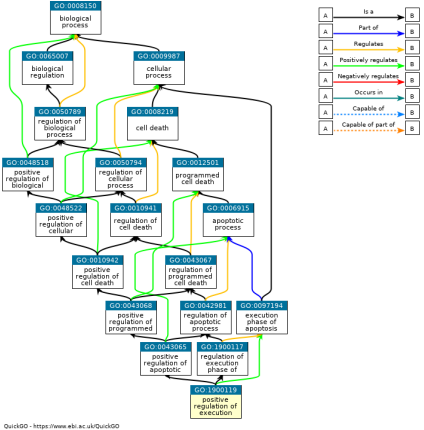
The Gene Ontology Database

- By definition, every term/node in each ontology inherits the properties of the parent node
- Each parent node contains several child terms directly beneath it
  - `http://amigo.geneontology.org/amigo/dd_browse`

The Gene Ontology Database

- By definition, every term/node in each ontology inherits the properties of the parent node
- Each parent node contains several child terms directly beneath it
  - `http://amigo.geneontology.org/amigo/dd_browse`
- Each child node inherits the properties of it's parent node
- Children can have multiple parents
- Edges connect children to parents

The Gene Ontology Database



The Gene Ontology Database

- Once a term is defined, it can be assigned to a gene/protein
- We need evidence ...
  - Multiple evidence codes are defined
  - Each mapping of gene to term includes the level of evidence
  - `http://geneontology.org/docs/guide-go-evidence-codes/`

The Gene Ontology Database

- Once a term is defined, it can be assigned to a gene/protein
- We need evidence ...
  - Multiple evidence codes are defined
  - Each mapping of gene to term includes the level of evidence
  - `http://geneontology.org/docs/guide-go-evidence-codes/`
- *Evidence is species-specific*, but is often mapped across species
- IEA represents the lowest quality
  - In non-model organisms, this might be all we have

A Few Challenges with GO Annotation

1. A set of specific terms are mapped to each gene
   - Parent terms may or may not be
2. There is a high level of redundancy
   - GO terms may overlap parent terms **significantly**
3. Visualisation for hundreds of GO terms from our analysis
   - Can we cluster by semantic similarity
   - Can we cluster by common membership (e.g. community detection)
4. Terms may also appear quite biologically abstract

GO Visualisations

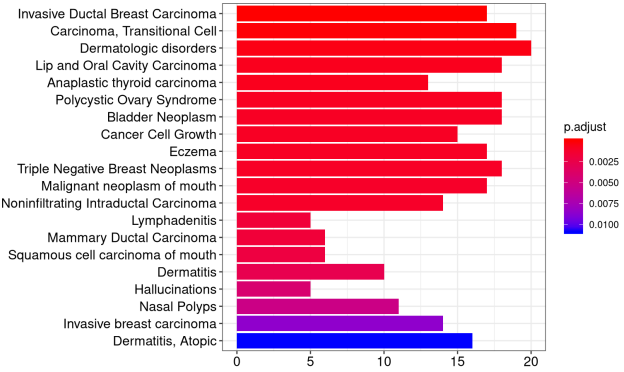


Image taken from clusterProfiler vignette `https://yulab-smu.github.io/clusterProfiler-book/chapter12.html`

GO Visualisations

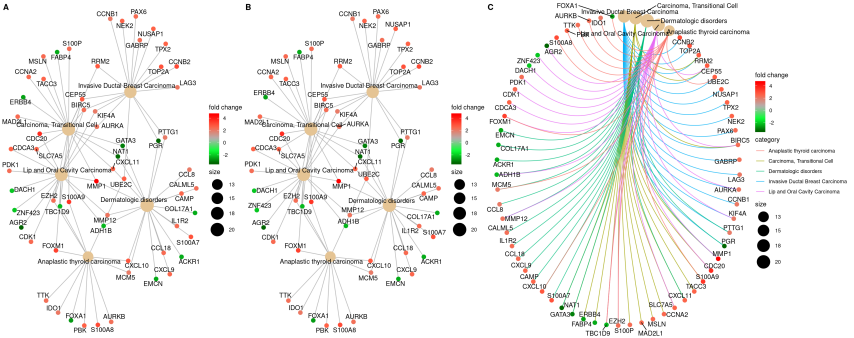


Image taken from clusterProfiler vignette `https://yulab-smu.github.io/clusterProfiler-book/chapter12.html`

GO Visualisations

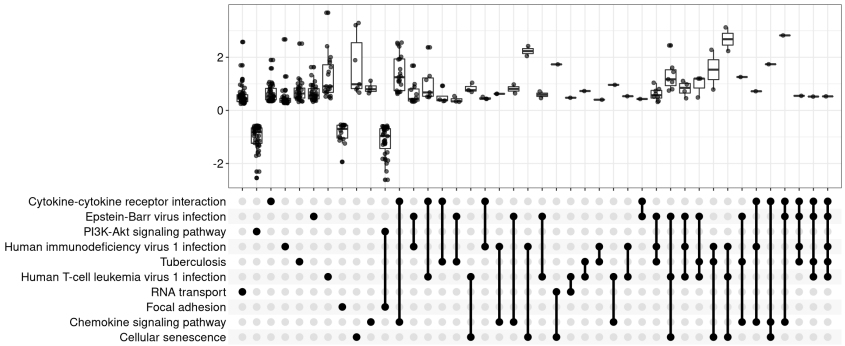


Image taken from clusterProfiler vignette `https://yulab-smu.github.io/clusterProfiler-book/chapter12.html`

KEGG Pathways

- The Kyoto Encyclopedia of Genes and Genomes: *KEGG*
- KEGG Pathways are *manually drawn pathway maps representing our knowledge on the molecular interaction, reaction and relation networks for*[2]:
  1. Metabolism
  2. Genetic Information Processing
  3. Environmental Information Processing
  4. Cellular Processes
  5. Organismal Systems
  6. Human Diseases
  7. Drug Development

[2] Taken from `https://www.genome.jp/kegg/pathway.html`

KEGG Pathways

- Each pathway is considered as a discrete unit $\implies$ no inheritance structure
- Pathways may strongly overlap still:
  `https://www.genome.jp/kegg-bin/show_pathway?map01100`
- Can search by compounds, genes, pathways

KEGG Pathways

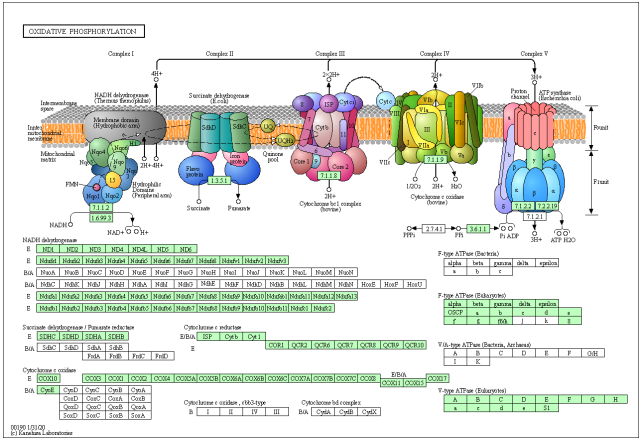
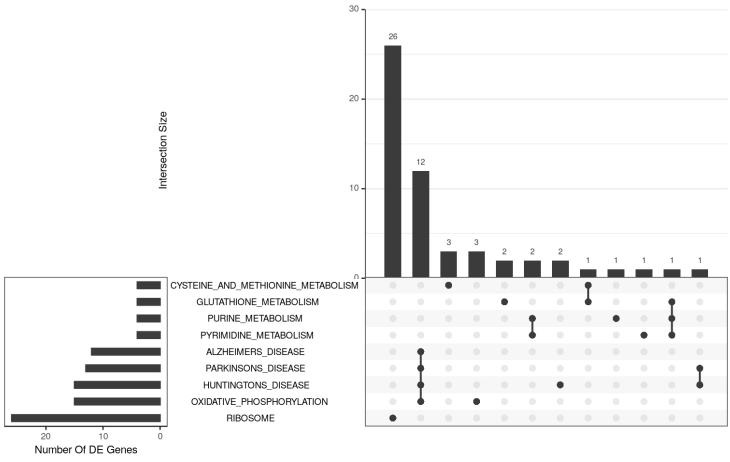


Image downloaded from KEGG

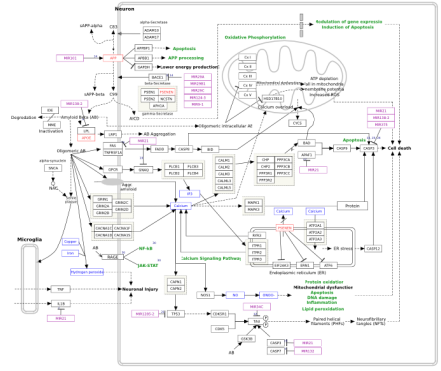
KEGG Pathways



Wiki Pathways

- Wiki Pathways is *maintained by and for the scientific community*
- Not dissimilar to a a publicly maintained KEGG
- Currently holds 2862 pathways

Wiki Pathways



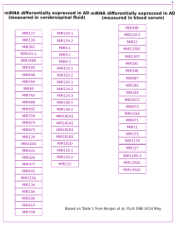


Image taken from https://www.wikipathways.org/index.php/Pathway:WP2059

The Molecular Signatures Database

- The Molecular Signatures Database (MSigDB) collects other databases
  - H: Hallmark Gene Sets
  - C1: Positional Gene Sets
  - C2: Curated Gene Sets (*BioCarta*, *KEGG*, *Reactome*)
  - C3: Regulatory Target Gene Sets (*miRNA targets*, *Transcription Factor targets*)
  - C4: Computational Gene Sets
  - C5: GO Gene Sets
  - C6: Oncogenic Gene Sets
  - C7: Immunologic Gene Sets

The Molecular Signatures Database

- Doesn't use or retain identifiers from original source
- Datasets are supplied as *species-specific* gene sets
- Huge redundancy
- Plays very nicely with R (`msigdbr`)

Transcription Factors

- Transcription factors present their own unique problems
- Genomic binding sites allow for significant flexibility

Binding Sites

(a) FOXP3

(b) FOXA1

(c) FOXO1

Transcription Factors

- Transcription factors present their own unique problems
- Genomic binding sites allow for significant flexibility
- DNA Shape can also play a role in specificity
- There is no 100% match giving a binary Yes/No

Transcription Factors

- Transcription factors present their own unique problems
- Genomic binding sites allow for significant flexibility
- DNA Shape can also play a role in specificity
- There is no 100% match giving a binary Yes/No
  - How do we define the presence of a motif?
  - How do we know which TF binds the motif?
  - Does only one TF bind a genomic locus?
  - How do we define a promoter & which gene(s) does an enhancer influence?

Testing Within DE Genes

Testing Our Data

- The most common test is for enrichment of a *pre-defined gene-set* within an *analytically defined gene-set*
- Our analytically defined geneset could be:
  - DE genes from a two-way comparison
  - Some other group defining a pattern of expression
- Groups can be defined directionally or not
- We usually test for enrichment *in comparison to a reference set of genes*

Testing Our Data

- The most common test is *Fisher's Exact Test*
- Tests H_0: *No association between groups*
- A common reference set of genes is *expressed but not DE* genes
- Far better than a random genomic reference
  - e.g. In brain cells we compare DE in brain against expressed in brain but not DE. This avoids finding enrichment for "brain-expressed genes"
- Is often referred to as a *hypergeometric* test

Testing Our Data

An Example

	DE	notDE
In gene-set	50	50
Not in gene-set	950	15000
Total	1000	15050

Under H_0 we expect $\pi = \frac{50}{15050} = 0.003$ of our DE genes to be in the gene set.

$(50 + 950) \times \frac{50}{15050} = 1000 \times \pi = 3.32$ genes. Clearly $50 \ggg 3.32 \implies p < 2X10^{-16}$

Testing Our Data

- Fisher's Exact Test is two-sided: *test is for association*
  - $p_{FET} = \frac{\text{the number of more extreme tables}}{\text{the total number of possible tables}}$
- Can return results which are **not** enriched
- Still need to use two-sided test, but can also check the observed $>$ expected
- Implemented in `limma` as `goana()` and `kegga()`

Testing Our Data

What about bias?

- Gene-length should be roughly constant between samples
- Long genes have higher counts $\implies$ biases DE
- Would this impact our results using Fisher's Exact Test?

[3] M. D. Young et al. "Gene ontology analysis for RNA-seq: accounting for selection bias". In: *Genome Biol.* 11.2 (2010), R14.

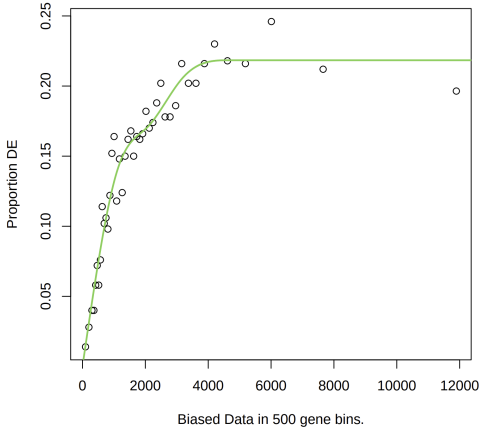
Testing Our Data

What about bias?

- Gene-length should be roughly constant between samples
- Long genes have higher counts $\implies$ biases DE
- Would this impact our results using Fisher's Exact Test?
- *Wallenius' Non-Central Hypergeometric Distribution* allows for sampling **with bias**
  - Also very applicable if GC content varies across samples/groups
- This incorporation of bias is implemented in goseq[3]

[3] M. D. Young et al. "Gene ontology analysis for RNA-seq: accounting for selection bias". In: *Genome Biol.* 11.2 (2010), R14.

Testing Our Data



Taken from https://bioconductor.org/packages/release/bioc/vignettes/goseq/inst/doc/goseq.pdf

Testing Our Data

In all cases:

1. We obtain a set of analytically defined genes (e.g. DE genes)
2. We test multiple predefined gene sets (usually 1000s)
3. We obtain a list of results with p-values
4. We adjust the p-values

Adjusting P-Values

- If there are no DE genes in a GO term (i.e. a gene-set), would we test for enrichment?
  - We could remove these gene-sets from our gene sets to be tested
  - Do we require a minimum number of DE genes in the gene-set to be interested?
- If using GO terms, those near the Ontology root tend to be uninformative
  - Remove terms based on shortest/longest path to root node?
- FDR-adjustment or Bonferroni?
  - Do we care more about Type I or Type II errors
  - Under Bonferroni $p < 0.05$ is a difficult threshold to cross

Using Ranked Lists

Using Ranked Lists

- All of the above looks for enrichment **within** an analytically-derived gene set
- This focusses on genes with the most significantly altered expression
- Are other biological behaviours worth exploring

Using Ranked Lists

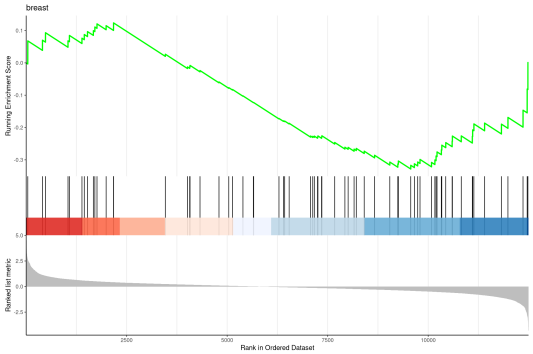
- All of the above looks for enrichment **within** an analytically-derived gene set
- This focusses on genes with the most significantly altered expression
- Are other biological behaviours worth exploring
- What if an entire pathway is up-regulated a very small amount?
- We can use ranked lists to test for "enrichment"
  - We can rank on t-statistic, p-value or any appropriate statistic

Using Ranked Lists

- The first approach proposed for this was *Gene Set Enrichment Analysis*, (GSEA)[4]
- "Takes a walk" down a ranked list and increases the *enrichment score* every time a gene is found from the gene-set
- Find the maximum deviation from zero and considers that the Enrichment Score
- All Enrichment Scores for a gene set are then normalised $\implies$ Normalised Enrichment Score
- The position *up to the maximal ES* is often called the *leading edge*

[4] Aravind Subramanian et al. "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles". In: *Proceedings of the National Academy of Sciences* 102.43 (2005), pp. 15545–15550. ISSN: 0027-8424. DOI: 10.1073/pnas.0506580102. eprint: https://www.pnas.org/content/102/43/15545.full.pdf. URL: https://www.pnas.org/content/102/43/15545.

Using Ranked Lists



- Here, the walk began at the most *downregulated* gene
- The leading edge would be genes to the right of the maximal ES (below the axis)

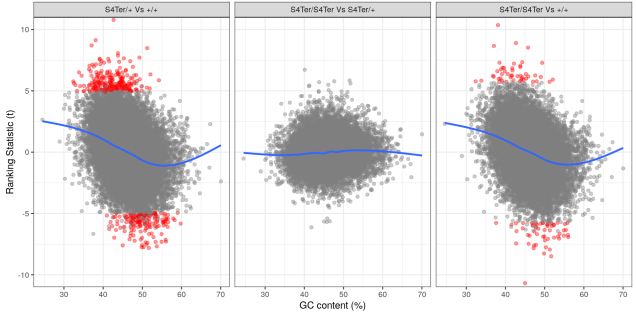
Image taken from clusterProfiler vignette `https://yulab-smu.github.io/clusterProfiler-book/chapter12.html`

Using Ranked Lists

- This approach is independent of any significantly DE genes
- Significance for a gene-set is obtained by comparing the ES to a Null distribution
  - Null distribution is obtained by permutation of samples/genes
- The end result is not dissimilar to the non-parametric Kolgorov-Smirnov test
- **However** this approach is very sensitive to bias and inter-gene correlations

Using Ranked Lists

- If GC or length bias shows strong correlation with treatment groups $\Longrightarrow$ lots of spurious results



Using Ranked Lists

- An alternative is ROAST, which uses rotation testing not permutation
- Inter-gene correlations are *explicitly accommodated*
- A gene-set level T-statistic is obtained, with a p-value by Monte-Carlo (rotation)
- A fast version is implemented in `limma` as `fry()`.
  - No direct equivalent to the leading edge is obtained
  - Crude approximation may be genes with $|T| > 2$

Using Ranked Lists

- Many alternatives exist
  - Wilcoxon Rank Sum Test, Kolgorov-Smirnov
  - Hypergeomteric testing whilst walking down a list
- The package `EGSEA` integrates multiple methods
- We want to capture real biology **not** artefacts from bias

Summary

- Testing within a set of DE genes against non-DE genes, for enrichment *within* the DE genes
- Testing along a ranked list for enrichment at either end
- Multiple testing applies under both approaches $\implies$ strong biological signals only
- Gene-sets can be literally anything (TFBS, miR targets, KEGG pathway)
  - We can also define our own gene-sets, i.e. 3'UTR with IRE
- Visualisations can be very challenging