Functional Enrichment

Dr Stevie Pederson (They/Them) stevie.pederson@thekids.org.au

Black Ochre Data Labs
The Kids Research Institute Australia

Acknowledgement Of Country

I'd like to acknowledge the Kaurna people as the traditional owners and custodians of the land we know today as the Adelaide Plains, where I live & work.

I also acknowledge the deep feelings of attachment and relationship of the Kaurna people to their place.

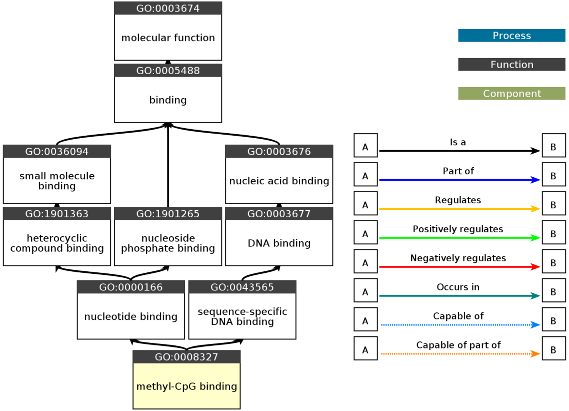
I pay my respects to the cultural authority of Aboriginal and Torres Strait Islander peoples from other areas of Australia, and pay my respects to Elders past, present and emerging, and acknowledge any Aboriginal Australians who may be with us today

Functional Enrichment

- Once we have a list of DE genes ⟹ what's the biological story?
  - Approach is dependent on the biology under investigation
- Compare to databases which define
  - *Pathways*: KEGG; WikiPathways; Reactome
  - *Structured Ontologies*: Gene Ontology (BP, MF & CC)
  - *Existing Experimental Results*: ImmuneSigDB; VAX
  - Multiple other options

We use enrichment testing on suitable *gene-sets*

The Gene Ontology Database



QuickGO - https://www.ebi.ac.uk/QuickGO

- Ontologies: Strictly controlled descriptive syntax
  - Members of a term (i.e. genes) inherit all parent terms
  - Inheritance → repetitive gene-sets
- 3 Primary Ontologies
  - Biological Process
  - Molecular Function
  - Cellular Components

Figure 1: https://www.ebi.ac.uk/QuickGO/term/GO:0008327

The Kyoto Encyclopedia of Genes and Genomes

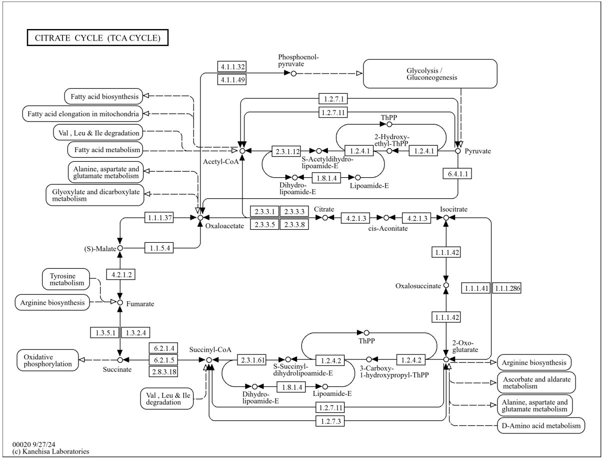


Figure 2: https://www.kegg.jp/pathway/map00020

- KEGG database:
  - Pathway topologies
  - Hierarchical relationship between genes
- Pathways (generally) distinct units

Fisher's Exact Test

- Are genes from a *gene-set* in my DE genes at random or unexpectedly often
  - H_0 : No enrichment of genes from a gene-set
  - H_A: Some enrichment of genes from a gene-set
- Fisher's Exact Test $\implies$ is there an association between groups in a table
  - Formally tests for independence of rows & columns in a table
  - Usually applied to a 2×2 table
- Similar to a χ^2-test but no model assumptions
  - Computationally more demanding

Fisher’s Exact Test

- A one-tailed Fisher’s Exact Test also defines the *hypergeometric distribution*
- If we have white & red marbles in an urn
  - Randomly draw one at a time (without replacing it)
  - Probability of drawing (at least) the exact number of red marbles
  - If we replaced the marble $\implies$ *binomial distribution*

Fisher's Exact Test

	In Gene-set	Not In Gene-set	Row Totals
DE Genes	a	b	$a+b$
Not DE	c	d	$c+d$
Column Totals	$a+c$	$b+d$	$a+b+c+d$

- p-value is the probability of observing a 'more-extreme' table than observed *holding marginal totals fixed*

Fisher's Exact Test

	In Gene-set	Not In Gene-set	Row Totals
DE Genes	a	b	$a+b$
Not DE	c	d	$c+d$
Column Totals	$a+c$	$b+d$	$a+b+c+d$

- p-value is the probability of observing a 'more-extreme' table than observed *holding marginal totals fixed*
  - This is an exact number $\implies$ no model assumptions

Fisher's Exact Test

	In Gene-set	Not In Gene-set	Row Totals
DE Genes	a	b	$a+b$
Not DE	c	d	$c+d$
Column Totals	$a+c$	$b+d$	$a+b+c+d$

- p-value is the probability of observing a 'more-extreme' table than observed *holding marginal totals fixed*
  - This is an exact number $\implies$ no model assumptions
- Perform this test on all selected gene-sets or pathways

Fisher's Exact Test

	In Gene-set	Not In Gene-set	Row Totals
DE Genes	a	b	$a+b$
Not DE	c	d	$c+d$
Column Totals	$a+c$	$b+d$	$a+b+c+d$

- p-value is the probability of observing a 'more-extreme' table than observed *holding marginal totals fixed*
  - This is an exact number $\implies$ no model assumptions
- Perform this test on all selected gene-sets or pathways
- Multiple-testing: FWER control may be better than FDR control

Fisher's Exact Test

	In Gene-set	Not In Gene-set	Row Totals
DE Genes	a	b	$a + b$
Not DE	c	d	$c + d$
Column Totals	$a + c$	$b + d$	$a + b + c + d$

- p-value is the probability of observing a 'more-extreme' table than observed *holding marginal totals fixed*
  - This is an exact number $\implies$ no model assumptions
- Perform this test on all selected gene-sets or pathways
- Multiple-testing: FWER control may be better than FDR control
  - FWER (e.g. Bonferroni): $\alpha \implies$ probability of any errors at all

Fisher's Exact Test

	In Gene-set	Not In Gene-set	Row Totals
DE Genes	a	b	$a+b$
Not DE	c	d	$c+d$
Column Totals	$a+c$	$b+d$	$a+b+c+d$

- p-value is the probability of observing a 'more-extreme' table than observed *holding marginal totals fixed*
  - This is an exact number $\implies$ no model assumptions
- Perform this test on all selected gene-sets or pathways
- Multiple-testing: FWER control may be better than FDR control
  - FWER (e.g. Bonferroni): $\alpha \implies$ probability of any errors at all
  - FDR (e.g. Benjamini Hochberg): $\alpha \implies$ rate of errors within 'significant' results

Fisher's Exact Test

	In Gene-set	Not In Gene-set	Row Totals
DE Genes	100	900	100 + 900 = 1000
Not DE	100	13900	100 + 13900 = 14000
Column Totals	100 + 100 = 20050% DE	900 + 13900 = 148006% DE	**1000 + 14000 = 150006.7% DE**

- Looks intuitively enriched:
  - 50% of gene set is DE
  - 6% outside of gene set is DE
- The p-value here is 3.80e-64
- Same result if matrix can be transposed (rows ↔ columns)
  - Same marginal totals held fixed

Fisher’s Exact Test

- Fisher’s Exact Test requires a list of DE genes
    - Also known as an *over-representation analysis*
- Sensitive to hard cutoffs (e.g. $p < 0.05$)
    - Is a gene with p $=$ 0.0499 DE whilst one with p $=$ 0.0501 isn’t?
- Might also choose top-ranked 100/200/500 if preferable

Gene Set Enrichment Analysis

- Gene Set Enrichment Analysis (GSEA) (Subramanian et al. 2005)
    - Uses complete ranked list (e.g. t-statistic)
- Walks along ranked list
    - Enrichment score ↑ if gene in gene-set
    - Enrichment score ↓ if not in gene-set
- Walk usually performed from either end of list
    - Results implicitly directional

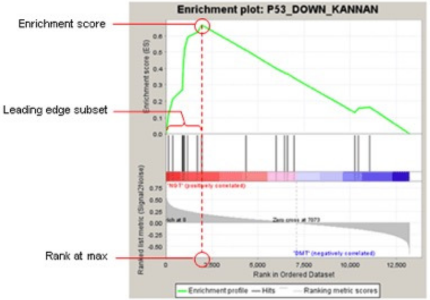


Figure 3: https://docs.gsea-msigdb.org/

Gene Set Enrichment Analysis

- The Enrichment Score depends on the gene-set size
    - Calculate Normalised Enrichment Score (NES)
- p-value obtained uses permutation-based methods
    - Random shuffling of genes and/or samples
    - Can subtly change results between analyses
- Alternatives, e.g. `camera` (Wu and Smyth 2012), incorporate inter-gene correlations

Additional Approaches

- ROAST (& FRY) perform *Rotation Gene Set Testing* (Wu et al. 2010)
  - *Is there differential expression within the gene-set*?
- Topology-based methods where logFC is propagated through a pathway topology
  - Only suitable for KEGG, WikiPathways etc which contain topologies

RNA-Seq Specific Challenges

- All above methods developed for microarray technology
- RNA-Seq contains additional biases
  - Longer genes $\implies$ higher counts $\implies$ bias in p(DE)
  - Modifications to hypergeometric models incorporate bias (Young et al. 2010)
- Most databases classify function at the *gene-level*
  - Transcript-specific function/pathways still poorly understood $\implies$ rarely used

References

Subramanian, Aravind, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, et al. 2005. "Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles." *Proc. Natl. Acad. Sci. U. S. A.* 102 (43): 15545–50.

Wu, Di, Elgene Lim, François Vaillant, Marie-Liesse Asselin-Labat, Jane E Visvader, and Gordon K Smyth. 2010. "ROAST: Rotation Gene Set Tests for Complex Microarray Experiments." *Bioinformatics* 26 (17): 2176–82.

Wu, Di, and Gordon K Smyth. 2012. "Camera: A Competitive Gene Set Test Accounting for Inter-Gene Correlation." *Nucleic Acids Res.* 40 (17): e133.

Young, Matthew D, Matthew J Wakefield, Gordon K Smyth, and Alicia Oshlack. 2010. "Gene Ontology Analysis for RNA-Seq: Accounting for Selection Bias." *Genome Biol.* 11 (2): R14.