Functional enrichment analysis (FEA)

Monash Genomics and Bioinformatics Platform
13/11/2025

Functional enrichment analysis

Functional enrichment analysis is a broad term that refers to various methods used to extract biological or functional insights from lists of biomolecules.

Identify biological functions, pathways, or molecular mechanisms that are significantly associated with a subset of biological molecules, such as those that are differentially expressed or mutated in a particular condition.

Synonyms

- Enrichment analysis
- Pathway analysis
- Pathway enrichment analysis
- Functional annotation analysis
- Functional enrichment analysis

Functional enrichment analysis: when?

Post differential expression (omics datasets)

- Transcriptomics (high-fat diet vs. low-fat diet)
- Proteomics
- Metabolomics/Lipidomics
- Epigenomics (cancer vs. normal cells)
- GWAS
- ...

Functional enrichment analysis: why?

Once a large-scale omics study undertaken:

- Summarise long list of many significant genes/proteins into meaningful biological insights
- Hypothesis generation

FEA doesn't *prove* mechanisms; it *points to them*.
It helps you move from
"What genes changed?" → to "What might be happening biologically, and how can I test it?"

Functional enrichment analysis: Input data?

Data captured from an -omics study

- List of features
- Background set
- Ranked list
- Gene sets

HALLMARK_ADIPOGENESIS	https://	ABCA1	ABCB8	ACAA2		
HALLMARK_ALLOGRAFT_REJECTION	https://	AARS1	ABCE1	ABI1	ACHE	
HALLMARK_ANDROGEN_RESPONSE	https://	ABCC4	ABHD2	ACSL3		
HALLMARK_ANGIOGENESIS	https://	APOH	APP	CCND2	COL3A1	COL5A2
HALLMARK_APICAL_JUNCTION	https://	ACTA1	ACTB	ACTC1	ACTG1	
HALLMARK_APICAL_SURFACE	https://	ADAM10	ADIPOR2	AFAP1L2		

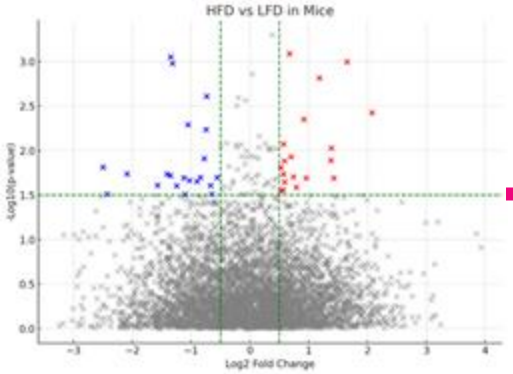
Gene	FC	p-value
ADAR	2.57	2.34E-06
ACADSB	2.53	0.000634
ADH1B	2.48	0.00574
ABCC4	2.17	0.00249
ACLY	2.02	1.98E-05
ACP2	1.94	3.75E-05
ACAD9	1.88	8.50E-06
ACTG2	1.85	0.00507
ACE	1.82	0.025
ADPRHL2	1.79	0.00156
ACP1	1.55	0.00273
ADSL	1.43	0.000453
A2M	1.35	0.00283
AEBP1	1.28	0.002
AAK1	-1.09	0.0238
ACAA2	-1.11	0.0156
ABCF1	-1.28	0.00147
A1BG	-1.34	5.15E-05
ACOX3	-1.41	0.0197
ACIN1	-1.64	2.57E-05
ACAD8	-1.69	0.0116
ABHD10	-1.77	0.00182
ACAA1	-1.84	0.00414
ACSL3	-1.91	0.00166
ABHD14B	-1.97	0.024
ACBD3	-2.12	0.000444
ABAT	-2.15	0.00403
ACSM3	-2.28	0.000703
ACSL1	-2.68	0.000584
ACY1	-2.72	2.61E-05
ACACA	-2.92	0.000124
ACSS1	-3.04	2.16E-05
ABHD11	-3.66	3.81E-06

Gene	Rank
ACLY	12.0898294
ACP1	9.85374061
ABCC4	9.48401106
ACE	7.11197165
ADH1B	6.62515224
ADAR	6.59826379
AEBP1	6.53661481
ACP2	6.37936782
ACAD9	6.28101832
ADPRHL2	6.20646376
ACAA1	6.14771005
ABAT	5.92969843
ACTG2	5.89740654
ABHD11	5.86732359
ADSL	5.74621125
A2M	5.63339695
ACADSB	5.52810629
ABHD14B	-6.2311846
ACAD8	-6.32085 79
ACSM3	-6.386081
ABHD10	-6.40474 45
A1BG	-6.44169 2
ACY1	-6.53491 81
AAK1	-6.64262 54
ACIN1	-6.96494 49
ABCF1	-7.10394 08
ACOX3	-7.17519 47
ACSL3	-7.22030 2
ACAA2	-7.80063 9
ACSL1	-8.01511 74
ACBD3	-8.56038 88
ACSS1	-1.00E+01
ACACA	-10.2042 92

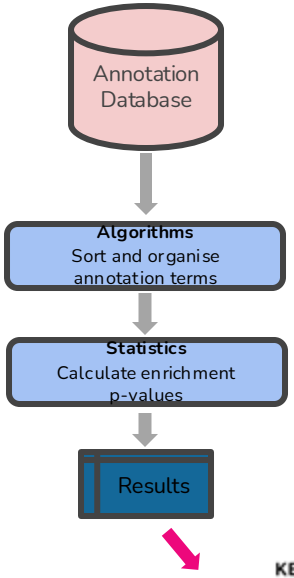
Concepts

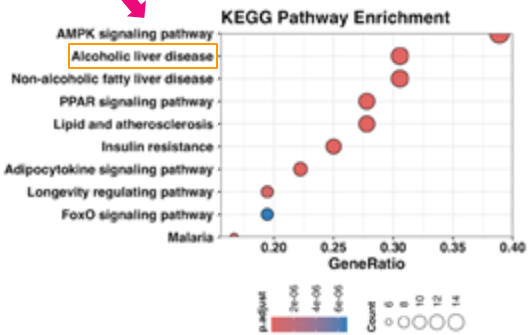
- Gene list (input set)
- Background set
- Gene set
- Annotation database
- P-value and false discovery rate (FDR)
- Regulation
- ID mapping

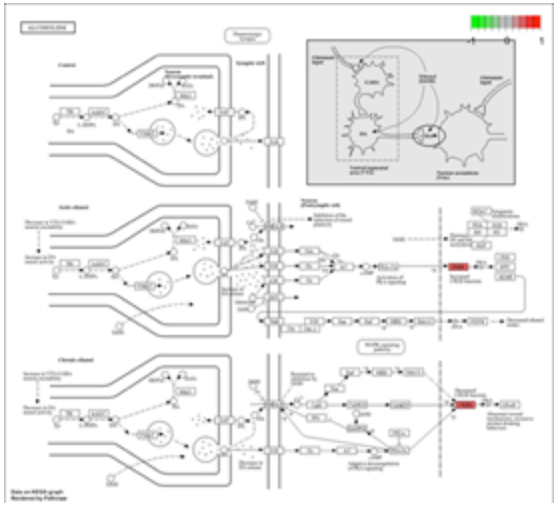
FEA at a glance: Mouse diet experiment



DEGs
Lep
Cd36
Fasn
Srebf1
Ccl2
Tnf
Il6
Il1b
Atf4
Dgat1
Dgat2
Cidea
Fabp4
Scd1
Irf7
Cxcl10
Resistin
Gdf15
Pparg
Adipoq
Cpt1a
Lpl
Ucp1
Slc2a4
Sod2
Irs1
Ppara
Pgc1a
Nrf1
Mc4r
Acadm
Gpr109a
Sod1
Nrf2
Hmgcr
Igf1
Foxo1
Il10
Il4
Eif2ak3





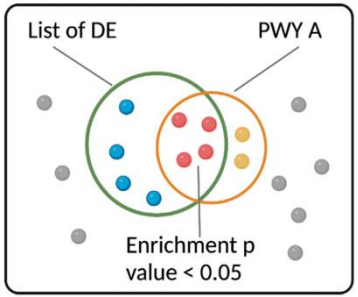


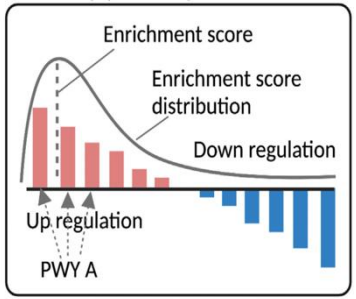
Types of Enrichment Analysis

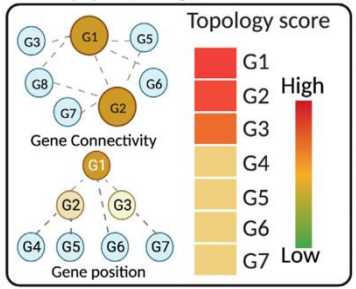
- Over Representation Analysis (ORA)
  - Up-regulated features
  - Down-regulated features
  - Modular Enrichment (clusters)
  - Cell-Specific ORA

- Functional Class Scoring (Rank-based)
  - Pre-ranked GSEA
  - ssGSEA
  - Camera
  - ROAST

- Topology-based Pathway Analysis (TPA)
  - SPIA
  - TopologyGSEA

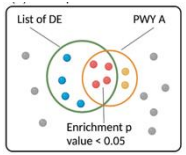


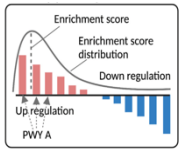


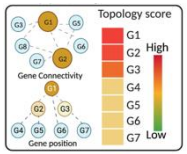


Zhao K, Rhee SY. Interpreting omics data with pathway enrichment analysis. Trends Genet. 2023

Comparison of Enrichment Methods







Method Type	Advantages	Disadvantages
Over-Representation Analysis (**1st Generation**)	• Straightforward • Widely implemented	• Ignores expression magnitude (treats all genes equally). • Requires arbitrary cutoff thresholds (information loss). • Assumes independence of each gene and pathway. • Ignores gene–gene correlation and pathway overlap. • May produce biased significance when pathways share genes.
Functional Class Scoring (**2nd Generation**)	• Uses all genes (no arbitrary cutoff). • Detects subtle, coordinated expression changes. • Incorporates gene-level statistics (fold-change, t-score, etc.). • Considers dependence among genes in a pathway.	• Still analyses each pathway independently. • Overlaps among gene sets can inflate significance. • Many methods use rank-based scores, ignoring magnitude differences. • Does not include interaction directionality or pathway structure.
Pathway Topology–based Analysis (**3rd Generation**)	• Incorporates pathway topology (interactions, direction, activation/inhibition). • Models propagation of perturbations through the network. • Identifies upstream/downstream effects and key nodes. • More mechanistic and biologically realistic.	•Requires experimental evidence for pathway structures and gene–gene interactions, which is largely unavailable for many organisms.

Why/when ora/gsea? Or both?

	ORA	GSEA
Purpose	Is a set of genes overrepresented in a list of significantly expressed genes?	Does a set of genes appear more often at the top of bottom of a ranked list
Inputs	List of significant genes (e.g. based on Pvalue or some cutoff)	Ranked gene list ordered by continuous metric (e.g. p value)
Gene set definition	Predefined genes grouped by function, pathway, phenotype	Same as ORA but applied across whole ranked list
Outputs	Enriched gene sets, p values for enriched sets	Enrichment scores per gene set, normalised enrichment scores, p values/FDR adjusted
Stats used	Fishers exact for over-representation?	Permutation-based for significance testing?
Example research q	What biological pathways are overrepresented among significantly downregulated genes in condition X?	Are immune-related pathways enriched across patients with condition x compared with healthy controls?

Statistics - overview (ora, gsea)

Annotation Databases

Gene Ontology (GO)

- Definition
- Categories
  - Biological Processes
  - Molecular Functions
  - Cellular Components

© 2000 Nature America Inc. • http://genetics.nature.com

commentary

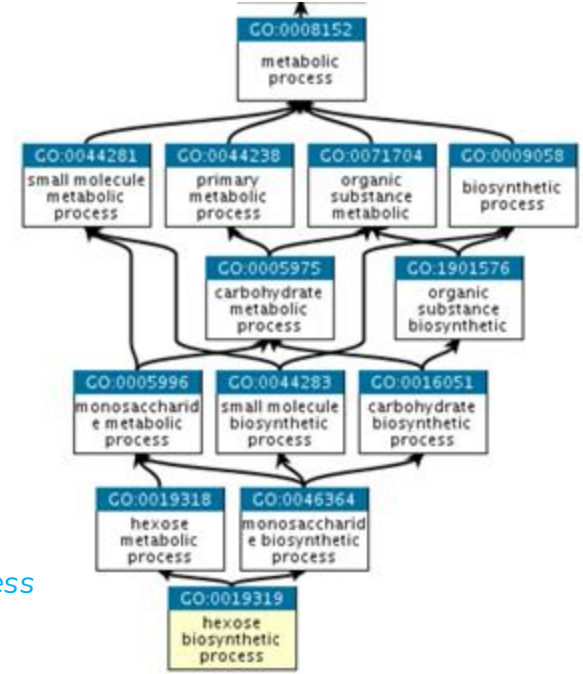
Gene Ontology: tool for the unification of biology

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.

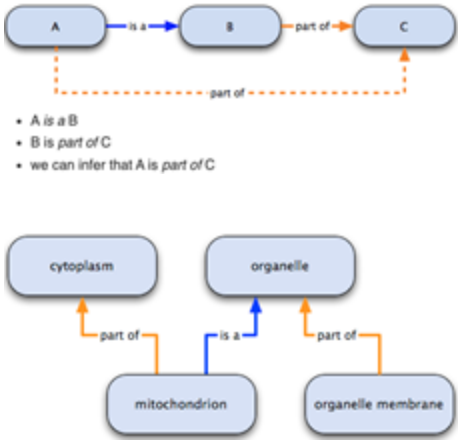
The GO Graph

- Definition
- Domains
  - Biological Processes
  - Molecular Functions
  - Cellular Components



This reflect the fact that:

- *biosynthetic process* is a subtype of *metabolic process*
- a *hexose* is a subtype of *monosaccharide*



mitochondrion has two parents:

- it *is an* organelle
- it is *part of* the cytoplasm

GO Domains

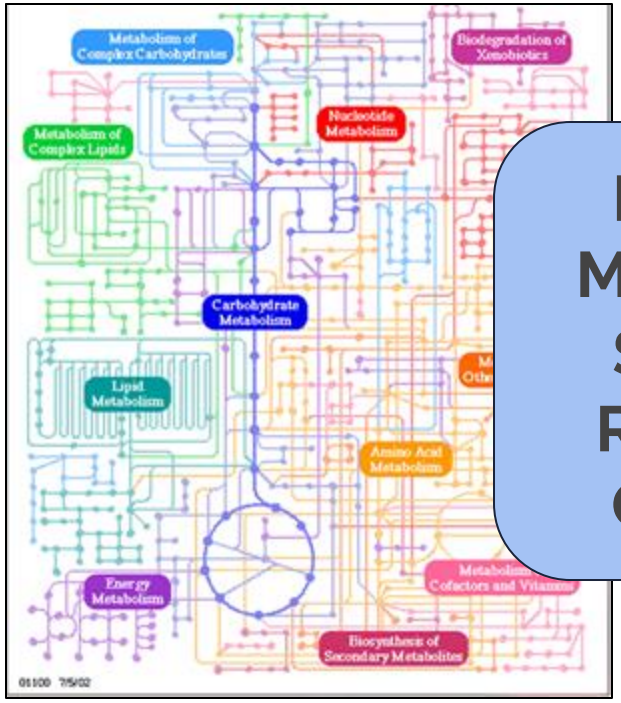
Molecular Function (MF)	Molecular-level activities performed by gene products. *catalytic activity* and *transporter activity*; *adenylate cyclase activity* or *Toll-like receptor binding*. GO molecular functions are often appended with the word "activity" (a *protein kinase* would have the GO molecular function *protein kinase activity*).
Cellular Component (CC)	A location, relative to cellular compartments and structures. *cellular anatomical entities*, includes cellular structures such as the *plasma membrane* and the *cytoskeleton*, as well as membrane-enclosed cellular compartments such as the *mitochondrion*
Biological Process (BP)	The larger processes, or 'biological programs' accomplished by multiple molecular activities. *DNA repair* or *signal transduction*. *pyrimidine nucleobase biosynthetic process* or *glucose transmembrane transport*.

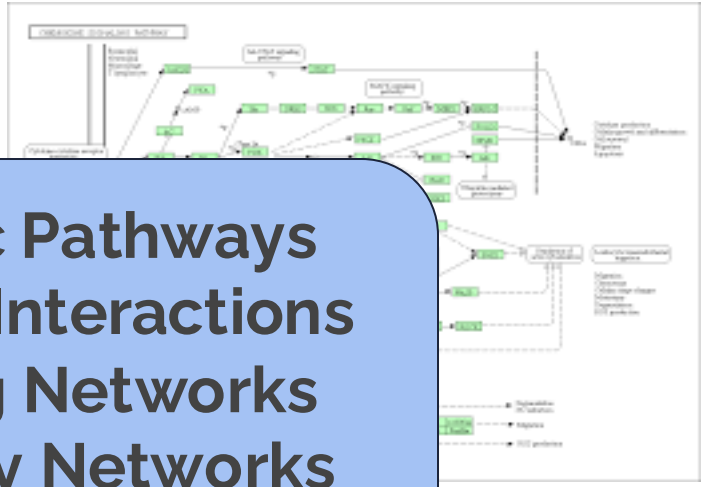
An example of GO annotation: human "cytochrome c":
molecular function *oxidoreductase activity*,
the **biological process** *oxidative phosphorylation*, and
the **cellular component** *mitochondrial intermembrane space*.

Note: a biological process is not equivalent to a pathway.

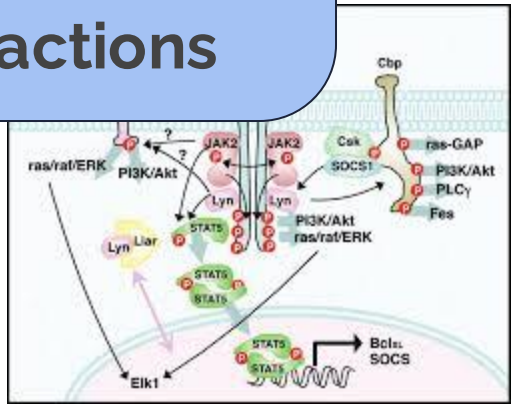
https://geneontology.org/docs/ontology-documentation

Pathways in Biology





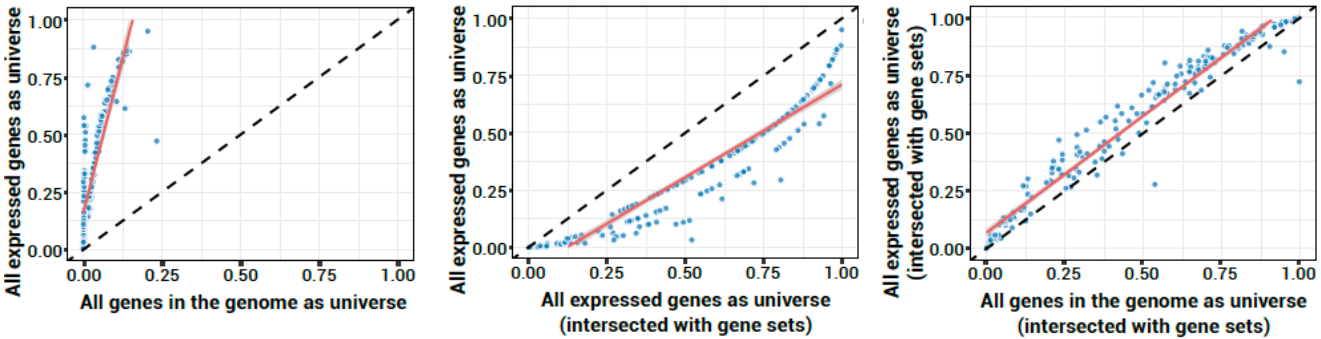
Metabolic Pathways
Molecular Interactions
Signalling Networks
Regulatory Networks
Genetic Interactions



Annotation Databases

Considerations – Background sets

1. Total features of the species
2. Subset of features captured in the experiment
3. Subset of features captured in the experiment ∩ annotation



- Yu G. (2026). Background bias in functional enrichment analysis: Insights from *clusterProfiler*. *The Innovation Life* **4**:100181.

Considerations – Background sets

1. Total features of the species
2. Subset of features captured in the experiment
3. Subset of features captured in the experiment ∩ annotation
4. Subset of research-specific features (metabolism > metabolic genes only)

Considerations – Annotation Databases

Annotation databases use different ontologies to define pathways

- KEGG vs EcoCyc
- KEGG Modules are comparable

Multiple pathway databases or integrative ones

Pathway	Differences	MPath
Notch signalling	WikiPathways = significant ↑ KEGG/Reactome = ↑ (not sig.)	Significant ↓ (direction reversed)
DNA replication	Reactome = ↑ KEGG/WikiPathways = ↓ (not sig.)	Significant ↓
Hedgehog signalling	WikiPathways = ↑ KEGG/Reactome = ↑ (not sig.)	Significant ↓ (direction reversed)
TGF-β signaling	KEGG = activated WikiPathways = inhibited	Inhibited
Estrogen signalling	KEGG = inhibited WikiPathways = activated	Inhibited

- Karp, P.D., Midford, P.E., Caspi, R. et al. Pathway size matters: the influence of pathway granularity on over-representation (enrichment analysis) statistics. BMC Genomics 2021
- Mubeen S, Hoyt CT, et al.. The Impact of Pathway Database Choice on Statistical Enrichment Analysis and Predictive Modeling. Front Genet. 2019 Nov

Special considerations - Proteomics

PSEA: relative abundance (label-free quantification methods)

PSEA-Quant: average abundance and variation between replicates (label-based n label-free)

- Rank proteins > weight > sum of weighted enrichment score of a term constituent proteins

- Cha S, Imielinski MB, Rejtar T, Richardson EA, Thakur D, Sgroi DC, Karger BL. In situ proteomic analysis of human breast cancer epithelial cells using laser capture microdissection: annotation by protein set enrichment analysis and gene ontology. Mol Cell Proteomics. 2010
- Lavallée-Adam M, Rauniyar N, McClatchy DB, Yates JR 3rd. PSEA-Quant: a protein set enrichment analysis on label-free and label-based protein quantification data. J Proteome Res. 2014

Special considerations - Metabolomics

MetPA: relies on compound annotation + network topology

- Annotating metabolites based of chemical standards, searching metabolic features like m/z or fragmentation patterns generated by MS against metabolite libraries.

Mummichog: to bypass annotation limitation

- Using spectral features like m/z and retention time

- Requires downstream analytical chemistry to identify specific metabolites associated with phenotypes of interest

Functional Analysis of Global Metabolomics

- Jianguo Xia, David S. Wishart, MetPA: a web-based metabolomics tool for pathway analysis and visualization, *Bioinformatics*, Volume 26, Issue 18, September 2010, Pages 2342–2344
- Li S, Park Y, Duraisingham S, Strobel FH, Khan N, Soltow QA, Jones DP, Pulendran B. Predicting network activity from high throughput metabolomics. PLoS Comput Biol. 2013;9(7):e1003123. doi: 10.1371/journal.pcbi.1003123. Epub 2013 Jul 4. PMID: 23861661; PMCID: PMC3701697.

5 things to remember when doing FEA

1. **The input gene lists for ORA and GSEA are different!**

   GSEA requires a ranked yet unfiltered gene list, while ORA requires a filtered unranked gene list and experimental background gene list

2. **Always correct for multiple testing!**

   Use adjusted P values/FDR/qvalues, never unadjusted P values

3. **Ensure reproducibility!**

   Record all relevant details including tool and database names and versions, all default and custom parameters and options applied, and background gene list

4. **Different analysis methods *will* return different results!**

   This is expected and OK, as long as your methods are sensible and detailed.

5. **Interpret your results in their biological context!**

   Functional categories are often broad and redundant. Use visualisations to make sense of it all.

Further reading

- Zhao and Rhee 2023: Interpreting omics data with pathway enrichment analysis
- Gable et al 2022: Systematic assessment of pathway databases, based on a diverse collection of user-submitted experiments
- Mubeen et al 2019: The Impact of Pathway Database Choice on Statistical Enrichment Analysis and Predictive Modeling
- Timmons et al 2015: Multiple sources of bias confound functional enrichment analysis of global -omics data
- Wijesooriya et al 2022: Urgent need for consistent standards in functional enrichment analysis
- Reimand et al 2019 (Nature Protocol): Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap
- Geistlinger et al 2020: Toward a gold standard for benchmarking gene set enrichment analysis