Functional enrichment analysis (FEA)

Today's webinar

Background and statistical concepts

- Introduction to functional enrichment analysis
- Key statistical concepts
- FEA workflow

PART 1

Functional enrichment analysis in practice

- Platforms and tools for FEA
- Tool choice considerations

PART 2

Part 1

Background and statistical concepts

Functional enrichment analysis

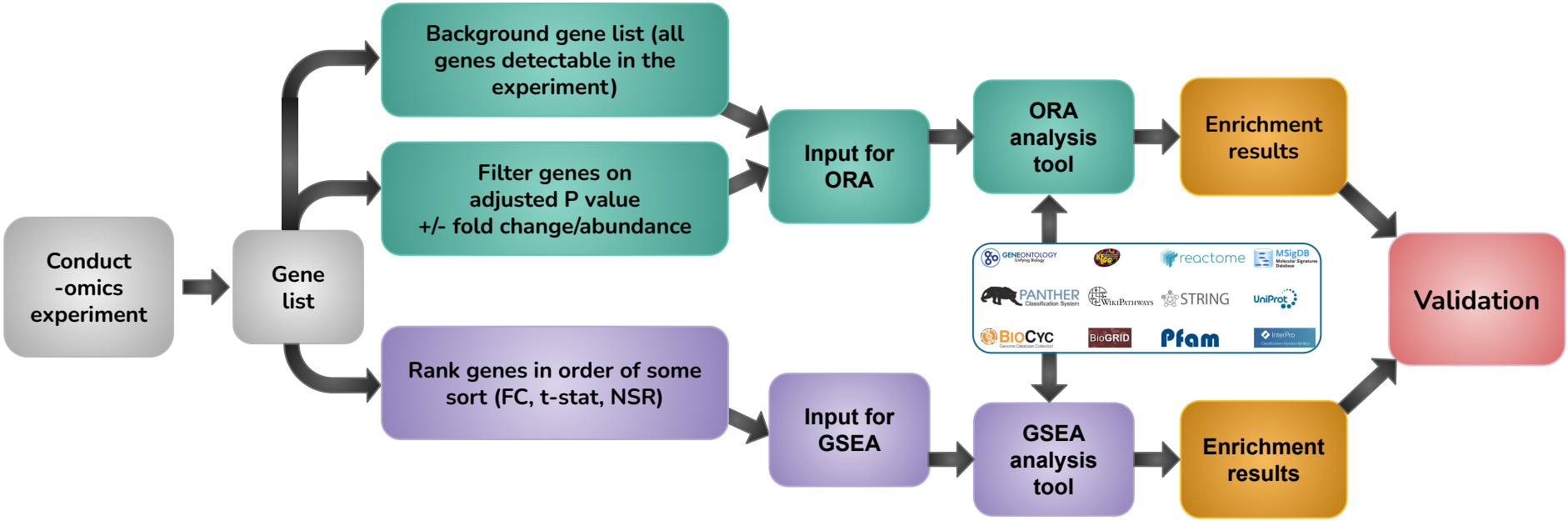
Functional enrichment analysis is a broad term that refers to various methods used to extract biological or functional insights from lists of biomolecules.

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

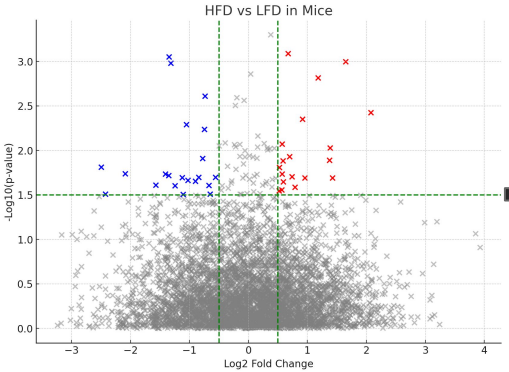
Synonyms

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

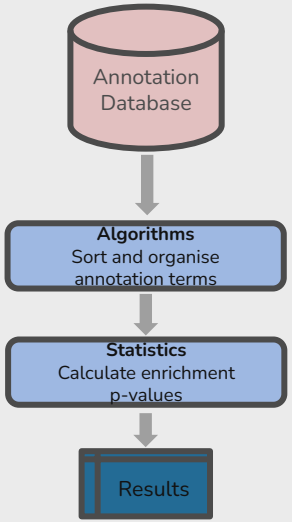
FEA workflow

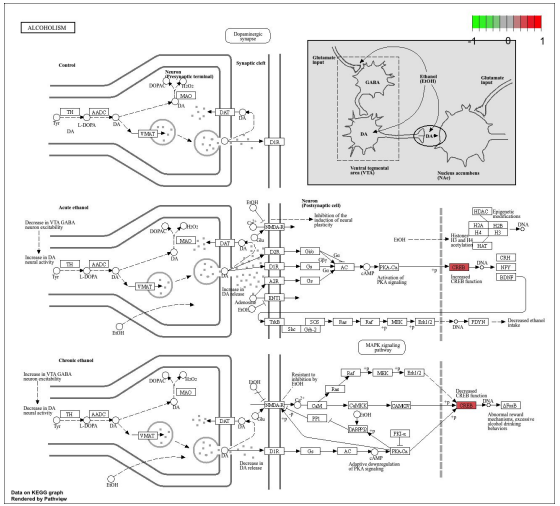


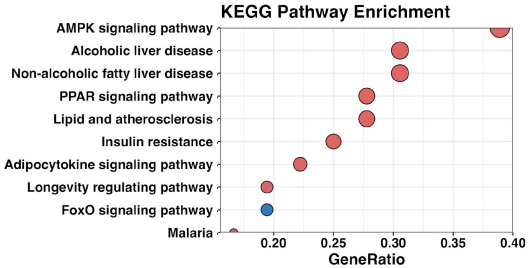
FEA at a glance: Mouse diet experiment



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







Australian BioCommons
MONASH University
THE UNIVERSITY OF SYDNEY

Functional enrichment analysis: When?

Post-differential expression analysis

- Transcriptomics (eg high-fat diet vs. low-fat diet)
- Proteomics (eg tumor tissue vs. healthy tissue)
- Lipidomics (eg disease vs. healthy state)
- Metabolomics (eg diabetic vs. non-diabetic patients)
- Epigenomics (eg smokers vs. non-smokers)

Functional enrichment analysis: Why?

Once a large-scale omics study undertaken

- Summarise long list of many significant genes/proteins
- Extract meaningful bi
- Hypothesis generatio

Monitor systems by observing the behaviour in the order of 100s and 10^6 molecules per experiment
Results in the order of 10^2- 10^4 features, as a list

Dysregulation of DNA repair mechanisms is a key driver in tumor progression in this specific cancer subtype.

Functional enrichment analysis: How?

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		

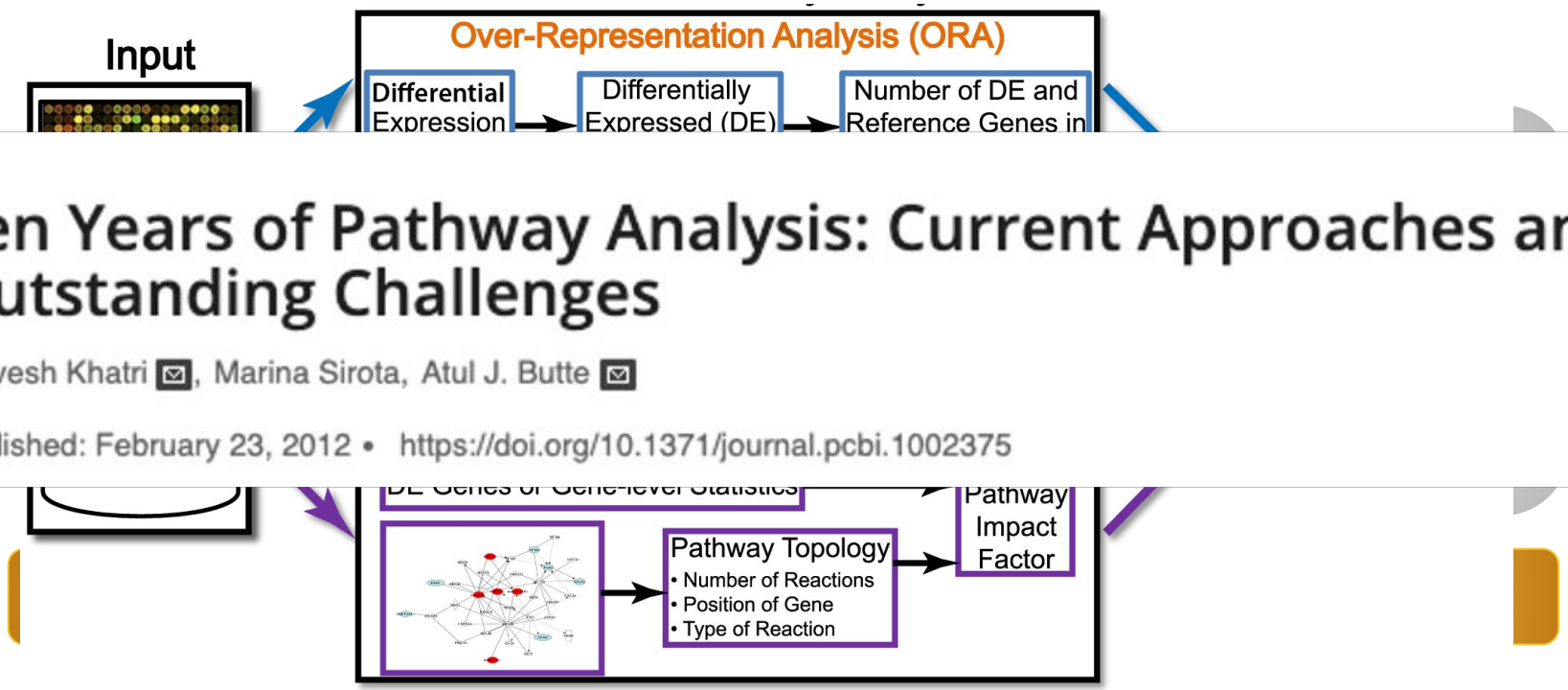
Gene	FC	p-value
ADAR	2.57	2.34E-06
ACADSB	2.53	0.000634
ADH1B	2.48	0.00574
ABCC4	2.17	0.00249
ACLY	2.02	1.98E-05
ACP2	1.94	3.75E-05
ACAD9	1.88	8.50E-06
ACTG2	1.85	0.00507
ACE	1.82	0.025
ADPRHL2	1.79	0.00156
ACP1	1.55	0.00273
ADSL	1.43	0.000453
A2M	1.35	0.00283
AEBP1	1.28	0.002
AAK1	-1.09	0.0238
ACAA2	-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	-6.3208579
ACSM3	-6.386081
ABHD10	-6.4047445
A1BG	-6.441692
ACY1	-6.5349181
AAK1	-6.6426254
ACIN1	-6.9649449
ABCF1	-7.1039408
ACOX3	-7.1751947
ACSL3	-7.220302
ACAA2	-7.800639
ACSL1	-8.0151174
ACBD3	-8.5603888
ACSS1	-1.00E+01
ACACA	-10.204292

Concepts

- Gene list
- Gene set
  - GO Term (Apoptotic process - genes involved in programmed cell death)
  - KEGG Pathway (MAPK signaling pathway - genes involved in cell proliferation, differentiation, and survival)
  - Reactome Pathway (Mitochondrial protein import - genes that help in importing proteins into mitochondria)
  - Hallmark Gene Sets (HALLMARK_HYPOXIA - genes involved in the cellular response to low oxygen levels)
  - Transcription Factor Targets (TP53_TARGETS - genes regulated by the tumor suppressor TP53)
  - Cell-Type Specific Gene Sets (Neurogenesis gene set – genes that control the formation of neurons during brain development)
- P-value and false discovery rate (FDR)
- Regulation
- Background set
- ID mapping
- Annotation database

Types of enrichment analysis



Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges

Purvesh Khatri, Marina Sirota, Atul J. Butte

Published: February 23, 2012 • https://doi.org/10.1371/journal.pcbi.1002375

Types of enrichment analysis

- Over Representation Analysis (ORA)
  - Modular Enrichment (WGCNA)
  - Cell-Specific ORA
- Gene Set Enrichment Analysis (GSEA)
  - Pre-ranked GSEA
  - ssGSEA
- Topology-based Pathway Analysis (TPA)
  - SPIA
  - TopologyGSEA

*Note: TPA is not covered in our workshop.

- Requires good understanding of network biology and pathway topology
- Needs high-quality pathway topology information (detailed pathway data), which might not be always available for all organisms or conditions
- ORA and GSEA have been mostly used by researchers

Why/when ORA/GSEA? Or both?

	ORA	GSEA
Inputs	Predefined gene list (e.g., DEGs, gene modules)	Full ranked gene list ordered by some sort of statistical method
Cutoffs	Requires cutoffs (logFC, p-value)	No cutoffs needed
Statistical methods	Fishers' Exact test, Hypergeometric test	Permutation tests like Kolmogorov–Smirnov (K-S) test to calculate p values
Outputs	Enriched terms with p values or FDR	Enrichment scores (ES) per gene set, normalised enrichment scores (NES), p values or FDR
Pros	Simple, computationally efficient, easier interpretation	Captures subtle effects and coordinated trends across all features
Cons	Missing subtle yet biologically important patterns, independence assumption of genes	Rank bias, gene set size bias, complex statistical framework, computationally intensive (depends on permutations)
Research question	Which biological pathways are over-represented in genes upregulated in response to a specific drug treatment?	Is there enrichment of genes involved in immune response pathways across the entire ranked gene list in patients with a specific disease?

Statistics overview - ORA

Fisher's Exact Test

Ronald A. Fisher
Lady Tasting Tea

	..in term	..not in term	Total
..in gene list	50	100	150
..not in gene list (but in background)	200	15900	16100
Total	250	16000	16250

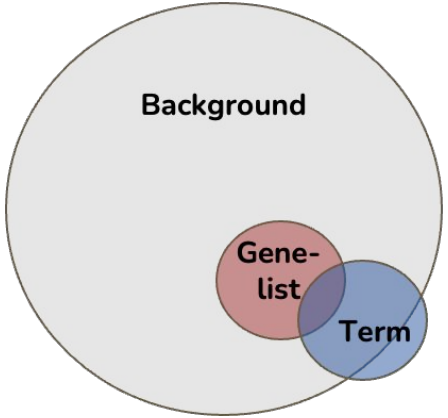
Contingency Table

Hypergeometric Test

Probability theory

more precise for small samples

more computationally efficient



	Category 1	Category 2	Total
Group 1	a	b	$a+b$
Group 2	c	d	$c+d$
Total	$a+c$	$b+d$	$a+b+c+d=N$

$$P = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!N!}$$ = 7.34e-54

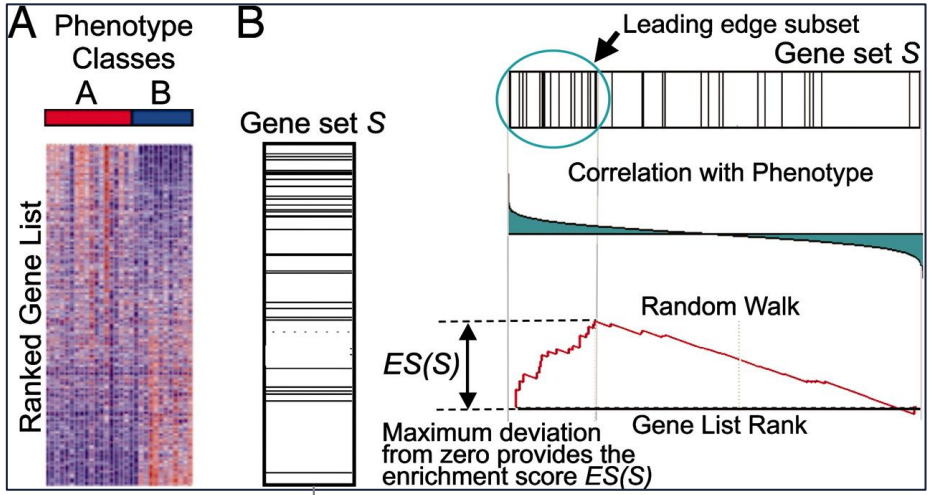
	In Gene Set	Not in Gene Set	Total
In Gene List	k	$n-k$	n
Not in Gene List	$K-k$	$N-K-(n-k)$	$N-n$
Total	K	$N-K$	N

$$P(X=k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$ = 7.52e-54

Australian BioCommons
MONASH University
THE UNIVERSITY OF SYDNEY

Statistics overview - GSEA

1. Calculating enrichment score (ES)
   - Walk down the ranked list of genes *L*, increment the running sum by $\sqrt{((N-N_h)/N_h)}$ and decrement it by $\sqrt{(N_h/(N-N_h))}$, similar to the Kolmogorov-Smirnov
2. Permutations
   - Randomly assign phenotype labels to samples, re-order genes, re-compute the ES of a gene set to generate a null distribution of ES.
   - Using this null, compute an empirical, nominal *p* value for any observed ES
3. Normalising enrichment score (NES)
   - Adjust for variation in gene set size
4. Multiple hypothesis testing
   - False Discovery Rate(FDR)
   - Family-Wise Error Rate (FWER)

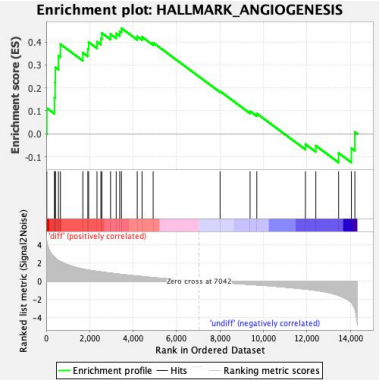


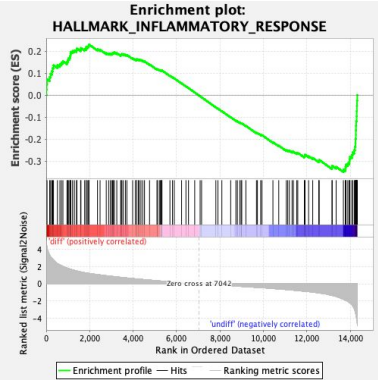
- Mitochondria
- MAP kinase signalling pathway
- Cell cycle control
- .

https://www.pnas.org/doi/epdf/10.1073/pnas.0506580102

https://github.com/ctlab/fgsea/issues/128

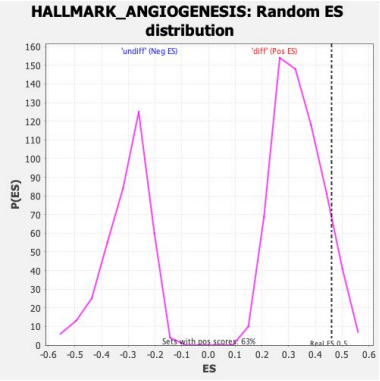
GSEA (null distribution)

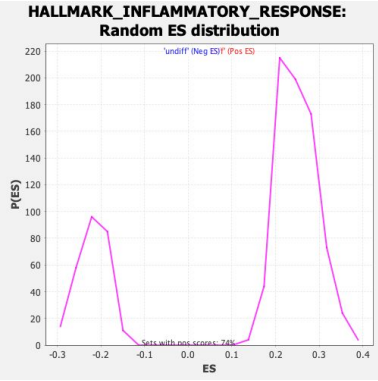




Enrichment Score (ES)	0.46041018
Normalized Enrichment Score (NES)	1.3703138
Nominal p-value	0.0955414
FDR q-value	0.094618164
FWER p-Value	0.729

Enrichment Score (ES)	-0.34908563
Normalized Enrichment Score (NES)	-1.5962827
Nominal p-value	0.0
FDR q-value	0.012562769
FWER p-Value	0.066





Australian BioCommons
MONASH University
THE UNIVERSITY OF SYDNEY

Annotation Databases

Annotation Databases

	Data	Application
GENEONTOLOGY Unifying Biology	Gene annotations and ontologies	Gene ontology mappings
KEGG	Biological pathways	Pathway mapping, system biology, drug development
reactome	Curated biological pathways, mainly human-focused	Cancer biology, immunology, cell signaling
MSigDB Molecular Signatures Database	Gene sets, pathways, and transcriptional signatures	Gene set enrichment analysis
PANTHER Classification System	Gene ontology, protein classification, pathways, and protein families	Gene ontology mappings, evolutionary analysis
WikiPathways	Community-curated biological pathways	Collaborative pathway curation, cross-species analysis
STRING	Protein-protein interaction networks, functional associations	Protein interaction network analysis
UniProt	Protein sequences, functional annotations, curated and predicted data	Protein sequence analysis, functional annotation, gene ontology integration
BioCyc Genome Database Collection	Metabolic pathways, genomes, gene regulatory networks	Metabolic network analysis
BioGRID	Protein, genetic, and chemical interactions	Protein interaction networks, systems biology
Pfam	Protein families, domains, and functional sites	Protein structure and function prediction
InterPro Classification of protein families	Protein families, domains, functional sites, protein sequence features	Functional domain identification, protein classification

Pathways in Biology

Biological pathways are series of interconnected biochemical reactions or molecular events that occur within cells, tissues, organs, or entire organisms.

These pathways describe the flow of biological information, matter, or energy that leads to specific biological outcomes.

- Metabolic Pathways
- Genetic Pathways
- Signal Transduction Pathways
- Immune Response Pathways
- Cell Cycle Pathways

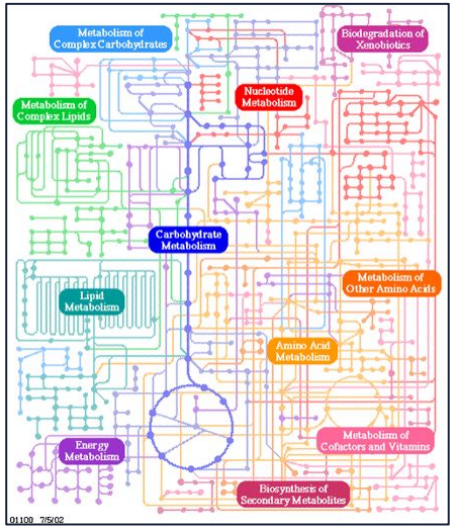
.

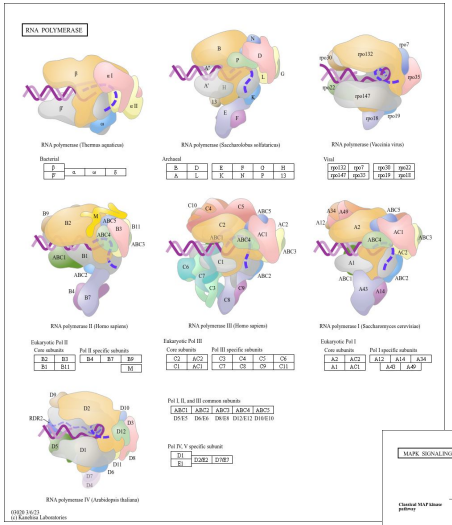
.

.

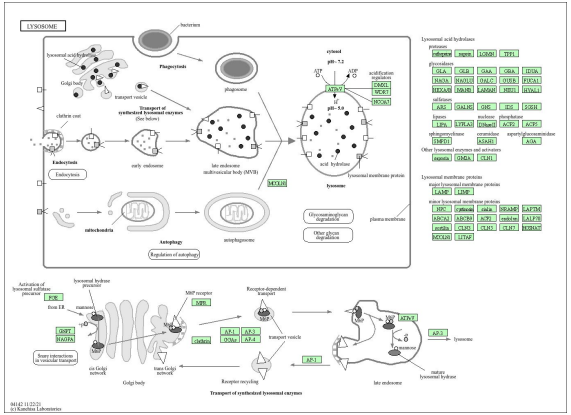
Pathways in Biology

Metabolic pathway

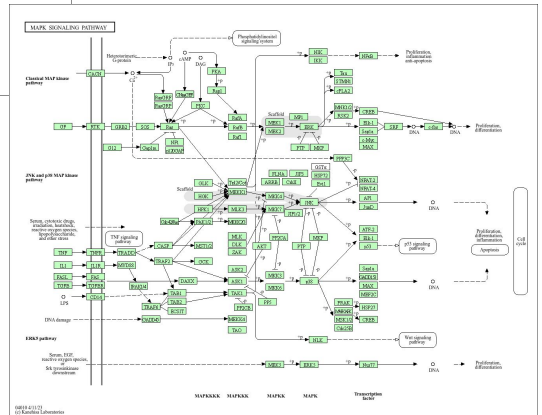




Genetic pathway



Genetic pathway



Signal transduction

Gene Ontology (GO)

- Controlled and structured hierarchical vocabulary for describing the properties and functions of gene products

- Hierarchical- parents and child terms establish more-general and more-specific descriptors of function

- Domains
  - Biological Processes
  - Molecular Functions
  - Cellular Components

 commentary

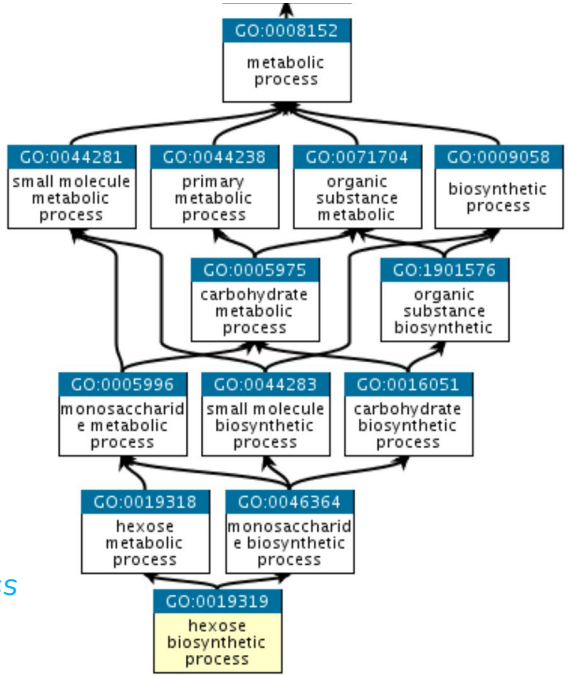
Gene Ontology: tool for the unification of biology

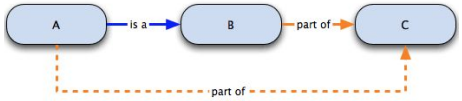
The Gene Ontology Consortium*

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.

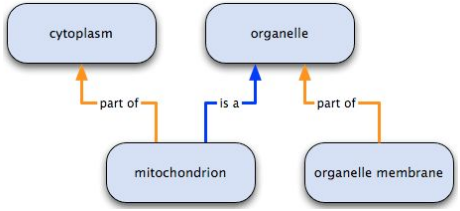
Directed Acyclic Graph (DAG)

GO classes (terms) are composed of a definition, a label, a unique identifier, and several other elements.





- A *is a* B
- B is *part of* C
- we can infer that A is *part of* C



mitochondrion has two parents:

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

This reflect the fact that:

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

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.

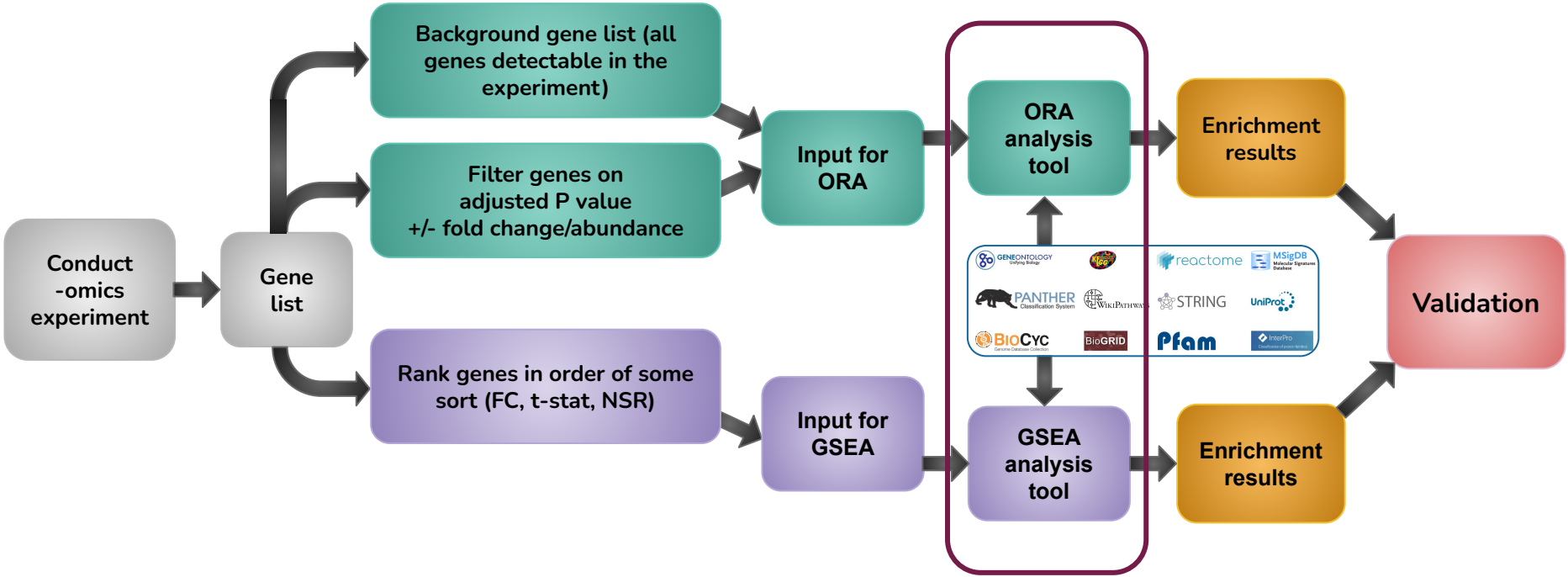
https://geneontology.org/docs/ontology-documentation

GO Evidence Codes	Evidence Code	Example
Experimental	Inferred from Experiment (EXP) Inferred from Direct Assay (IDA) Inferred from Physical Interaction (IPI) Inferred from Mutant Phenotype (IMP) Inferred from Genetic Interaction (IGI) Inferred from Expression Pattern (IEP)	Experimental results support annotation Enzyme assays, Immunofluorescence Co-purification Mutation assays Phenotype suppression or enhancement Expression experiments
Phylogenetically-inferred	Inferred from Biological aspect of Ancestor (IBA) Inferred from Biological aspect of Descendant (IBD) Inferred from Key Residues (IKR) Inferred from Rapid Divergence (IRD)	Ancestral gene Descendant gene Lack of key sequence residues Divergence from ancestral sequence
Computational analysis	Inferred from Sequence or structural Similarity (ISS) Inferred from Sequence Orthology (ISO) Inferred from Sequence Alignment (ISA) Inferred from Sequence Model (ISM) Inferred from Genomic Context (IGC) Inferred from Reviewed Computational Analysis (RCA)	BLAST Phylogenetic analysis Alignment between a query to a reference Predicted statistical model of a sequence Proximity to other genes (like operons) Predictions based on computational analyses of large-scale experimental data sets
Author statement	Traceable Author Statement (TAS) Non-traceable Author Statement (NAS)	Review articles UniProt Knowledgebase records
Curator statement	Inferred by Curator (IC) No biological Data available (ND)	Not supported by any direct evidence New gene sequenced, no biological evidence
Electronic annotation	Inferred from Electronic Annotation (IEA)	Computational methods, no human review

Part 2

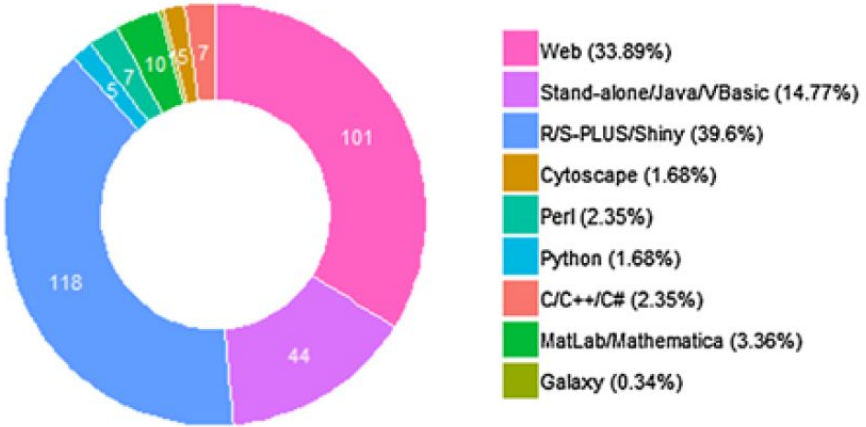
Functional enrichment analysis in practice

FEA workflow



Graphical or command line interface

- FEA can be performed via:
  - Graphical user interface (GUI) - web or desktop application
  - R statistical programming language
- Key considerations:
  - Type of analysis (ORA, GSEA)
  - Database integration
  - Ease of use
  - Which species you are studying



FEA tools published 2001-2021 by platform

Xie et al 2021

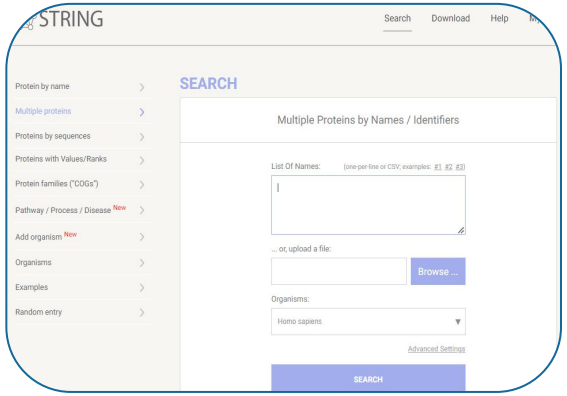
Web platforms for FEA

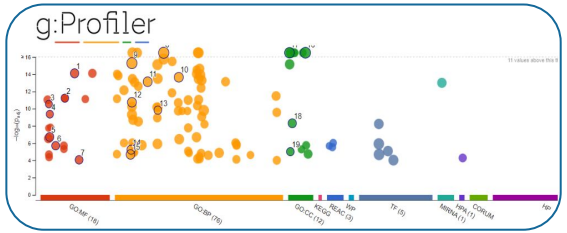
Key **advantages** for using web platforms include:

- Simple user interface
- Database integration

Key **disadvantages**:

- Limited visualisation flexibility
- ‘Black box’ can affect reporting and reproducibility
- Not available to all species
- Data security





GUI tool	Databases	Interface	ORA	ORA BG	GSEA	Network analysis	Species	Notable for
g:Profiler	Several	Web	✓	✓	–	–	984	Intuitive interface
STRING	STRING + several	Web	✓	✓	✓	✓	>12.5 K + any proteome	Protein interaction networks; non-model species
Reactome	Reactome	Web	✓	╳	FIViz app	–	16	Pathways curated on experimental data
GSEA	MSigDB	App	–	–	✓	–	3	Curated gene sets
GenePattern	MSigDB	Web	–	–	✓	–	3	Curated gene sets; many functions
WebGestalt	Several	Web	✓	✓	✓	✓	12 + custom	Visualisations; TPA
Enrichr	Many	Web	✓	✓	–	–	7	Extensive gene sets
Metascape	Several	Web	✓	✓	–	✓	10	Visualisations
DAVID	Several	Web	✓	✓	–	–	Some non-model	Outdated interface
PANTHER	PANTHER, Reactome, GO	Web	✓	✓	✓	–	144	Curated pathways inferred from phylogeny
IPA	IKB + others	Licensed app	✓	✓	–	✓	3	Curated database

Common pitfalls

- Most web platforms for ORA do not emphasise the option for provision of background gene list

- Can be difficult to obtain details sufficient for reporting and reproducibility

Urgent need for consistent standards in functional enrichment analysis

Kaumadi Wijesooriya[1], Sameer A. Jadaan[2], Kaushalya L. Perera[1], Tanuveer Kaur[1], Mark Ziemann[1]*

1 Deakin University, School of Life and Environmental Sciences, Geelong, Australia, 2 College of Health and Medical Technology, Middle Technical University, Baghdad, Iraq

* m.ziemann@deakin.edu.au

Abstract

Gene set enrichment tests (a.k.a. functional enrichment analysis) are among the most frequently used methods in computational biology. Despite this popularity, there are concerns that these methods are being applied incorrectly and the results of some peer-reviewed publications are unreliable. These problems include the use of inappropriate background gene lists, lack of false discovery rate correction and lack of methodological detail. To ascertain the frequency of these issues in the literature, we performed a screen of 186 open-access research articles describing functional enrichment results. We find that 95% of analyses using over-representation tests did not implement an appropriate background gene list or did not describe this in the methods. Failure to perform p-value correction for multiple tests was identified in 43% of analyses. Many studies lacked detail in the methods section about the tools and gene sets used. An extension of this survey showed that these problems are not associated with journal or article level bibliometrics. Using seven independent RNA-seq datasets, we show misuse of enrichment tools alters results substantially. In conclusion, most published functional enrichment studies suffered from one or more major flaws, highlighting the need for stronger standards for enrichment analysis.

https://doi.org/10.1371/journal.pcbi.1009935

R programming language for FEA

Key **advantages** for using R:

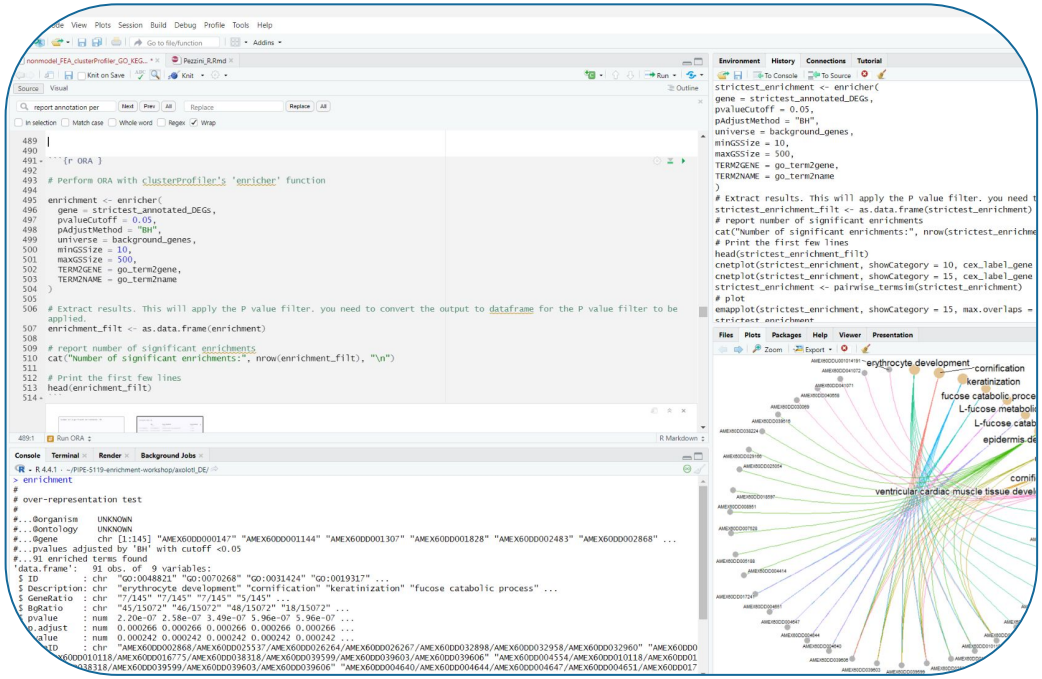
- Saved code provides thorough reproducibility
- Dedicated FEA packages available to simplify analysis
- High flexibility and parameter control
- Comprehensive plot options
- Can be used for non-model species which lack available web databases
- Can be performed offline, for maximum data security

Key **disadvantage**:

- Steeper learning curve
- Can be slower than native web tools (external database calls depend on your local internet speed)

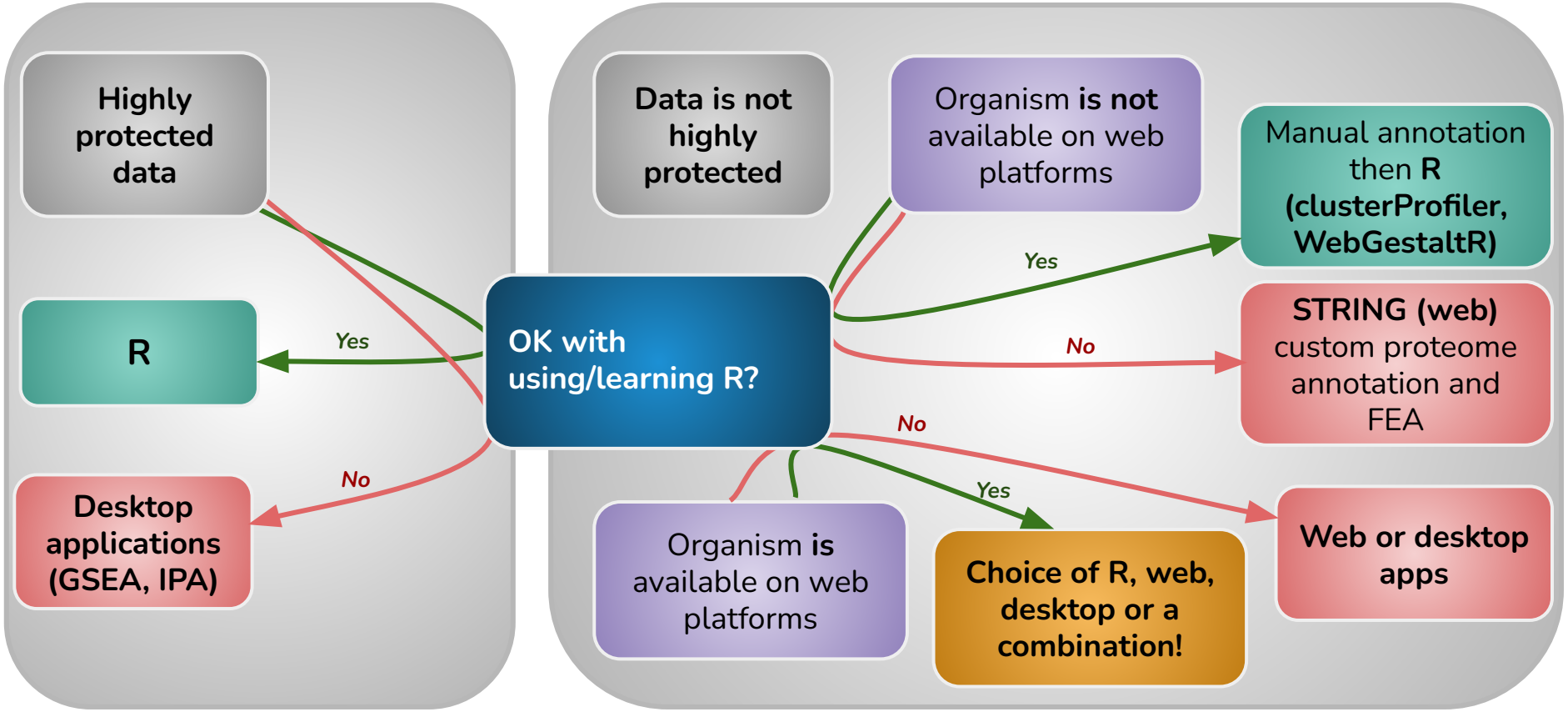
Integrated development environments for R

- IDEs help simplify working in R
- Popular choices include:
  - RStudio
  - VS Code
  - Jupyter Notebook



R Tool	Databases	ORA	ORA BG	GSEA	Network analysis	Species	Notable for
clusterProfiler	Several	✓	✓	✓	–	>10 K (KEGG)	Many functions for integrated DBs; companion plotting tool 'enrichplot'; novel species
gprofiler2	Several	✓	✓	–	–	984	Quick enrichment over many DBs in one command
enrichR	Several	✓	✗	–	–	7	Extensive gene sets
WebGestaltR	Several	✓	✓	✓	✓	12 + custom	Topology-based pathway analysis (TPA); visualisations and reports; novel species
fgsea	MSigDB	–	–	✓	–	3	Curated gene set analysis of human and mouse
STRINGdb	STRING	✓	✓	–	✓	>12.5 K	Protein interaction networks; non-model species
ReactomePA	Reactome	✓	✓	✓	–	16	Reactome DB analysis of model species
topGO	GO	✓	✓	–	–	20	Improve the specificity of GO enrichment results

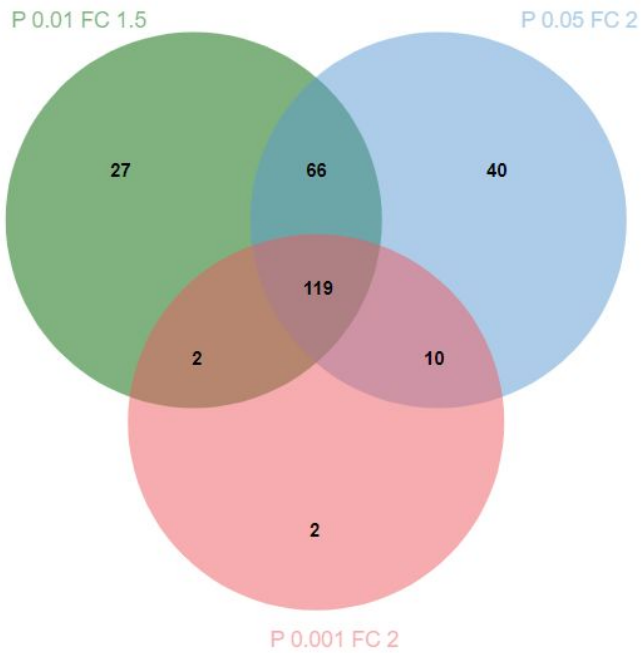
Suggested decision tree: GUI or R

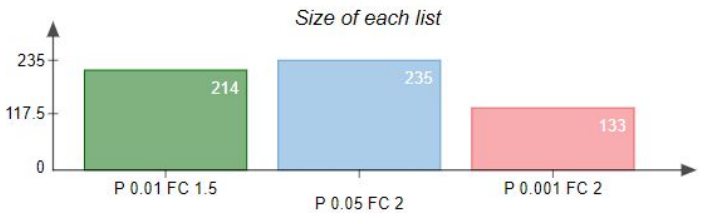


Tool choice will impact results

- Like any statistical analysis, small changes in method can lead to different results
- All analysis tools are doing things slightly differently, eg
    - Different underlying statistical analysis method
    - Different databases
    - Different database versions
    - Different P value adjustment method
    - Different default parameters
- Your gene list processing and arbitrary filtering choices will also impact results

Different input filtering affects results

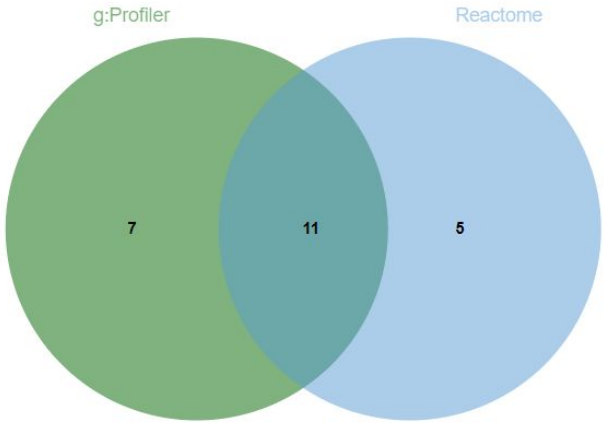




g:Profiler over-representation analysis of GO terms

jvenn: Bardou et al 2014
data: Pezzini et al 2016

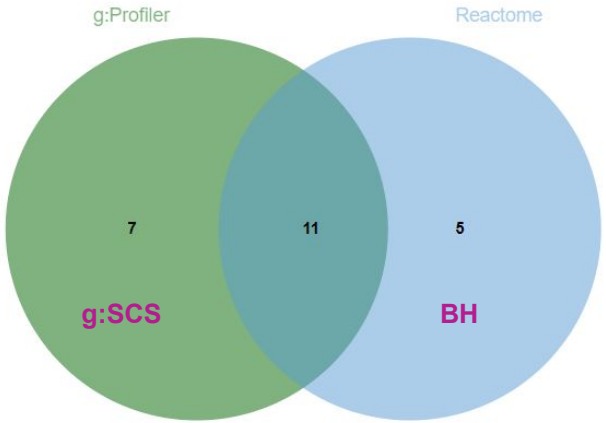
Different tools running on the same database can also give different results



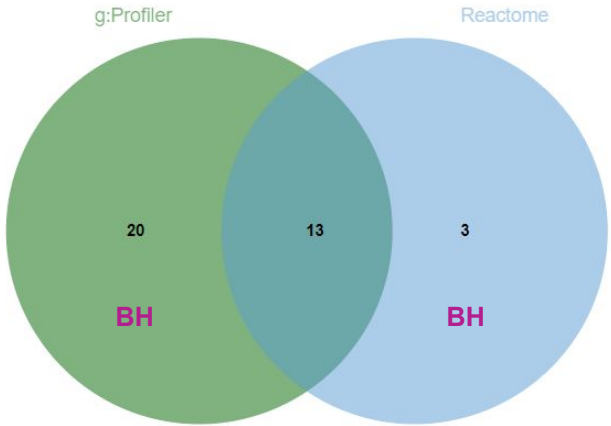
Tool	g:Profiler	Reactome
Statistical test	Hypergeometric	Hypergeometric
Padj cutoff	Default (0.05)	Default (0.05)
FDR method	Default (g:SCS)	Default (BH)
Background	Default (annotated genes)	Default (?)

jvenn: Bardou et al 2014
data: Pezzini et al 2016

Different tools running on the same database can also give different results



Tool	g:Profiler	Reactome
Statistical test	Hypergeometric	Hypergeometric
Padj cutoff	Default (0.05)	Default (0.05)
FDR method	Default (g:SCS)	Default (BH)
Background	Default (annotated genes)	Default (?)



jvenn: Bardou et al 2014
data: Pezzini et al 2016

How to manage conflicting results

This can make it hard to know which results to trust

If you:

- Apply **robust methods**
- Use **sensible parameter choices**
- Interpret your results in their biological context

> *"We believe the right attitude on the functional enrichment analysis is to treat it as a **guidance** to filter and rank pathways and processes, but **not to religiously believe in the absolute numbers**"*

Metascape, 2019

the results will be as valid as any other regardless of which platform you use

Robust and reproducible methods

To ensure robust methods, keep these things in mind:

Use the **right gene list:**
Pre-filter (eg P*adj*) for ORA, don't filter but sort (eg fold change) for GSEA

Choose **actively maintained tools:**
Those that are regularly updated and use the latest databases

Use the **right background**:
Always reduce the genes to the list of those that are detectable in the experiment (eg expressed in your tissue, present on your microarray...) to avoid bias

Use **adjusted P value** to account for multiple testing, never raw P value

Report all methodological details in your methods to ensure reproducibility, eg:

- Tool and tool version
- DB and DB version
- Filter thresholds
- P*adj* method
- Optional parameters applied
- Include background gene list
- Copy of R code/link to repository

Validating FEA results

- Cross-check findings with existing biological knowledge
    - Are the enriched pathways relevant to your tissue type or biological condition?
- Validate results with independent datasets or alternative methods
    - Consistent significant enrichment across multiple methods supports validity
- Reduce redundancy in terms (eg through REVIGO for GO terms) to help highlight the most relevant processes
- Explore FEA workflow benchmarking, a complex but worthy topic, eg GSEABenchmarkeR (Geistlinger at al 2020)
- Where possible, use laboratory validation (e.g., qPCR, Western blotting, knockout studies)

5 things to remember when doing FEA

1. ORA and GSEA are different statistical analyses, and their inputs differ!

GSEA: Kolmogorov-Smirnov test, requires a ranked yet unfiltered gene list

ORA: Hypergeometric or Fisher's Exact test, requires a filtered unranked gene list and experimental background gene list

2. Always correct for multiple testing!

Never unadjusted P values

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

This is expected and OK, as long as your methods are robust, sensible and reproducible. All results should be validated!

4. Ensure reproducibility!

Record all methodological details

5. Interpret your results in their biological context!

Functional categories are often broad and redundant. Use the FEA results as a guide, not the end point. Use visualisations to make sense of it all. Validate!

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