NUTRIOME

Functional analysis

Part 1: Molecular processes and pathways

Martina Summer-Kutmon

martina.kutmon@maastrichtuniversity.nl

NUTRIOME Workshop 1

Maastricht Centre for Systems Biology (MaCSBio)

30 May 2024

ORCID: 0000-0002-7699-8191

Maastricht University

Funded by the European Union

Marie Skłodowska-Curie Actions

macsbio
MAASTRICHT CENTRE FOR SYSTEMS BIOLOGY

Current knowledge level

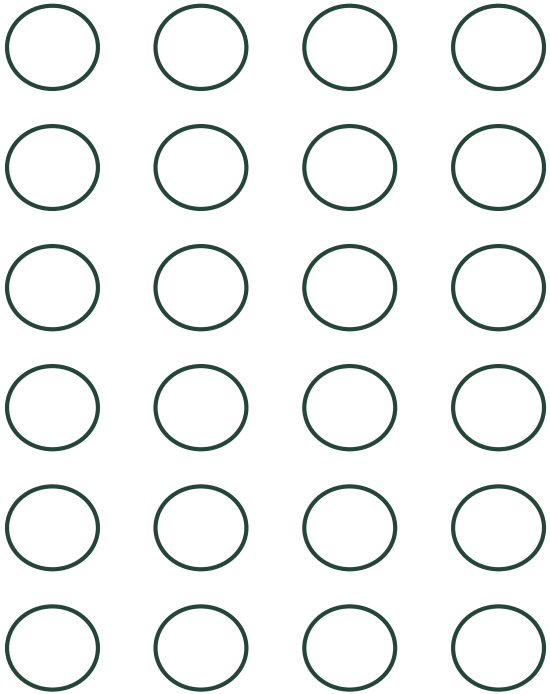
NUTRIOME

Introduction

Funded by
the European Union

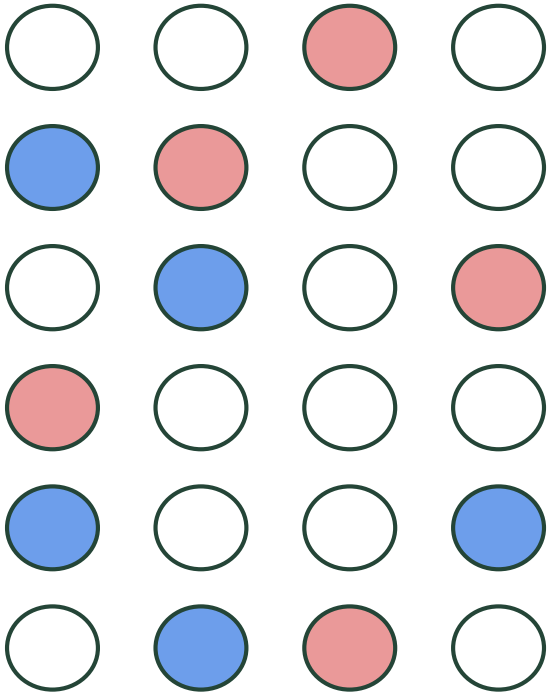
Marie
Skłodowska-
Curie Actions

Introduction enrichment analysis



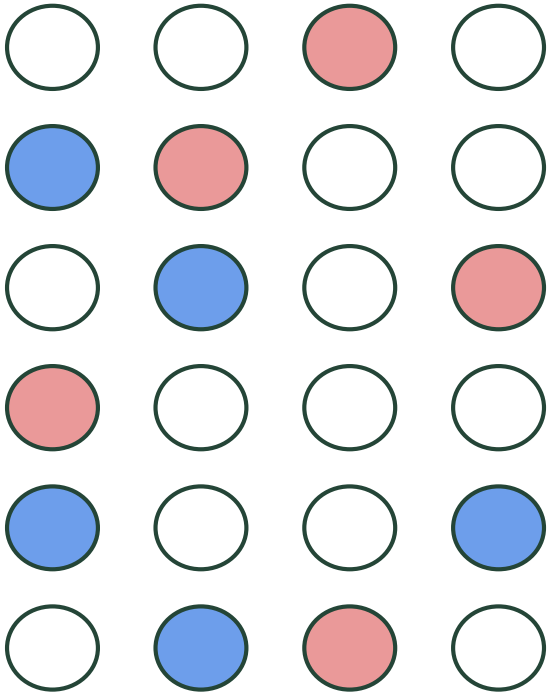
Quantify
Isolated data points

Introduction enrichment analysis



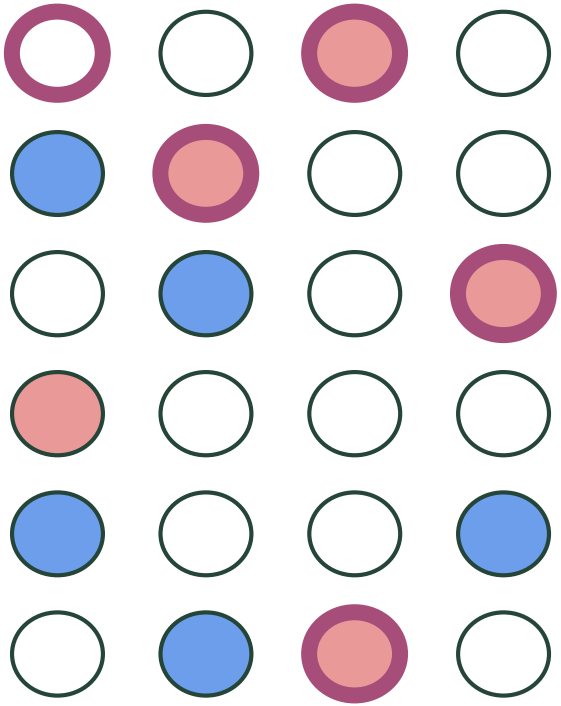
Comparative statistics
Genes of interest (DEseq2)

Introduction enrichment analysis



Enrichment analysis
Pre-defined gene sets → functional groups

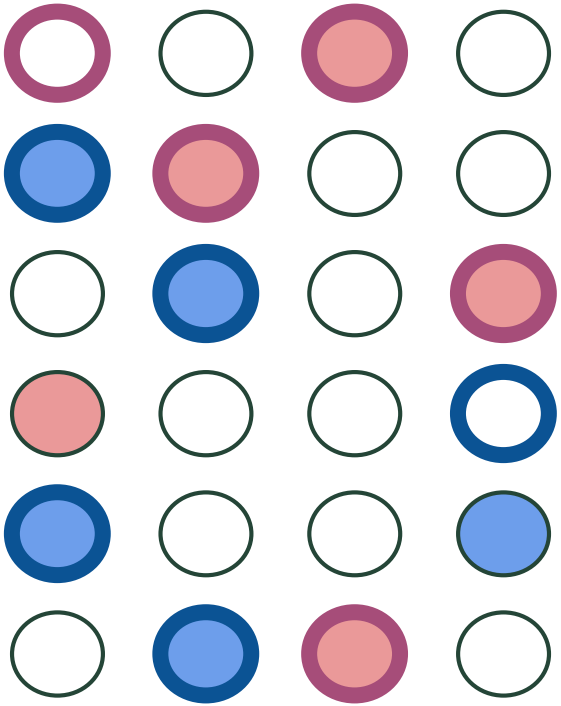
Introduction enrichment analysis



Enrichment analysis
Pre-defined gene sets → functional groups

Apoptosis

Introduction enrichment analysis

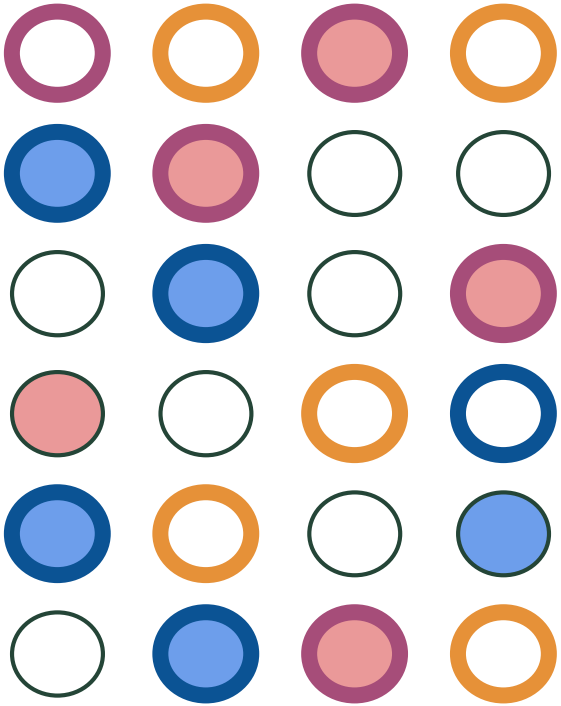


Enrichment analysis
Pre-defined gene sets → functional groups

Apoptosis

Catalytic activity

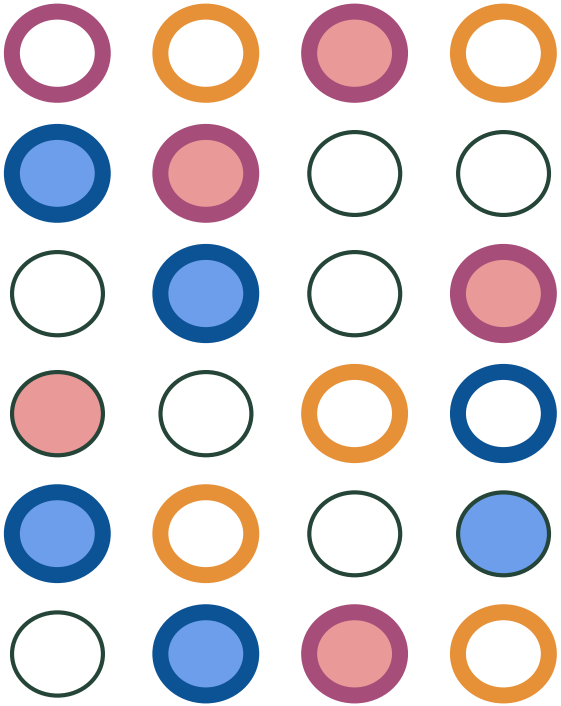
Introduction enrichment analysis



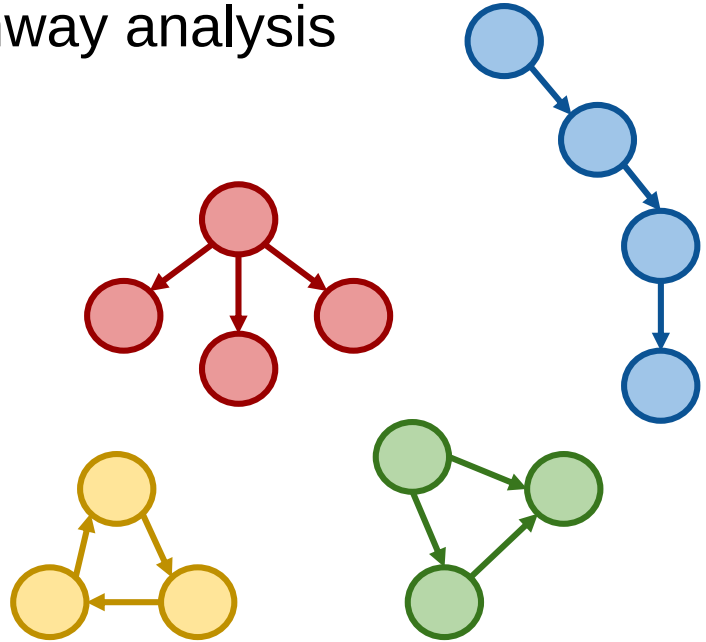
Enrichment analysis
Pre-defined gene sets → functional groups

- Apoptosis
- Catalytic activity
- GATA3 targets

Introduction enrichment analysis

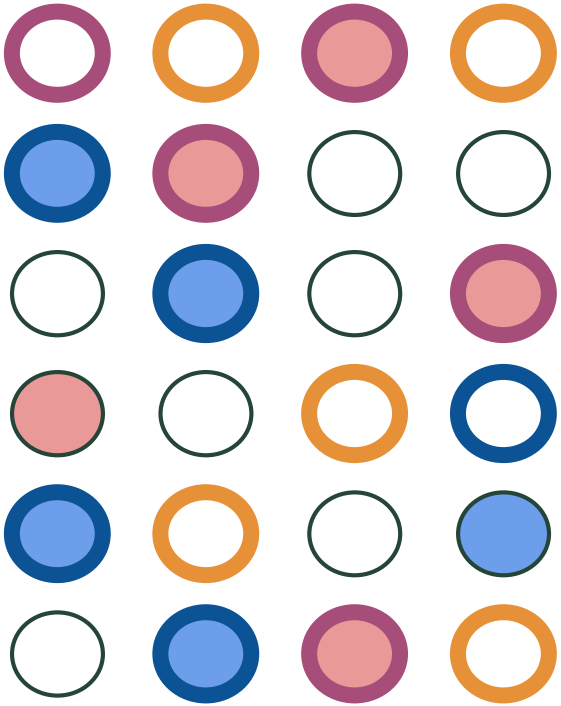


Enrichment analysis
Pathway analysis

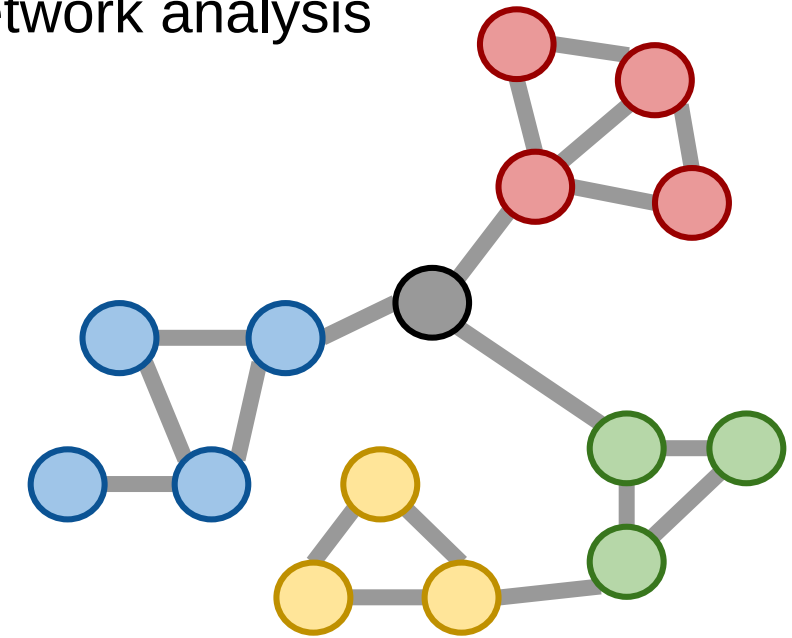


Pathway = gene set with information about relationships

Introduction enrichment analysis



Systems organization
Network analysis



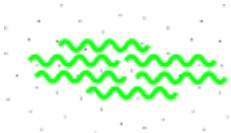
Why enrichment analysis?

“Enrichment” of gene sets

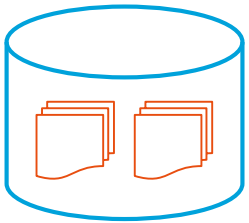
- Statistics
  - Analysis of groups instead of individual genes
  - Increases power and reduces dimensionality
- Biological
  - Analysis on a functional level
  - Higher explanatory power

NUTRIOME

How does it work?



Gene expression
(microarray / RNASeq)



Gene sets
Pathways, GO, gene sets

Enrichment analysis method
Over-representation analysis
Functional class scoring

Gene set significance

NUTRIOME

Gene set collections

Funded by
the European Union

Marie
Skłodowska-
Curie Actions

Gene set collections

Group genes based on some shared characteristic, e.g.

- Molecular processes/pathways
- Molecular function
- Cellular component
- Positional (on chromosomes)
- Hallmark gene sets
- Motif gene sets
- Signature gene sets
- Disease gene sets

MSigDB
Molecular Signatures Database

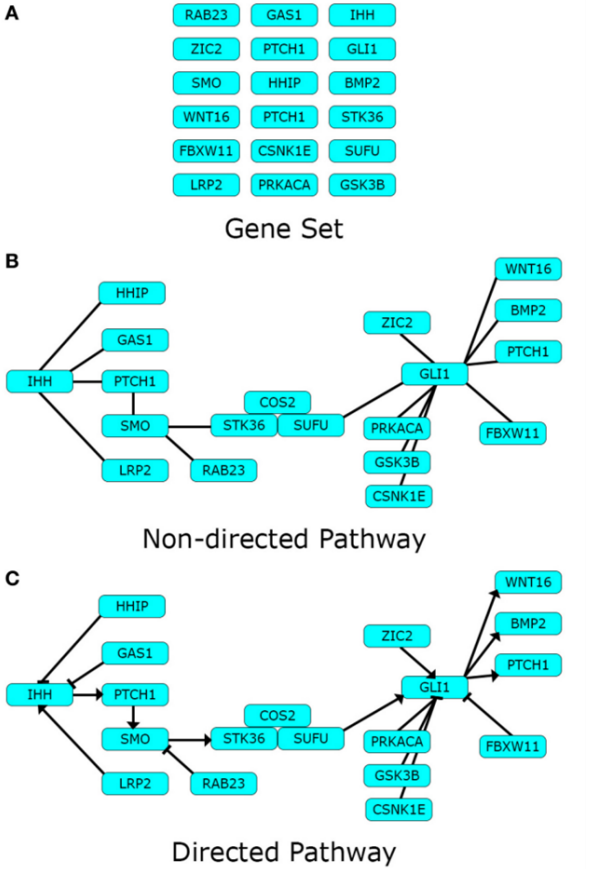
Molecular signature database
https://www.gsea-msigdb.org/gsea/msigdb

Subramanian (2005) PNAS
doi: 10.1073/pnas.0506580102

Gene sets - level of detail

Example: Hedgehog signaling pathway

A. Gene sets
   - Biological components pertaining a definite biological theme
B. Non-directed pathways
   - Describe the existence of definite interactions between the same components in the form of a network
C. Directed pathways
   - Disclose the character of the interactions in the network. Arrows depict an activating impact from the pointer component over the pointed one, and blunt edges an inhibiting one.



García-Campos (2015) Front. Physiol.
doi: 10.3389/fphys.2015.00383

Gene Ontology

- Ontologies provide **controlled, consistent vocabularies** to describe concepts and relationships, thereby enabling **knowledge sharing**” (Gruber 1993)

- Ontologies for **molecular biology domains** developed and supported by the Gene Ontology Consortium for **gene and gene product annotations** for all organisms

The Gene Ontology Consortium (2023) Genetics
doi: 10.1093/genetics/iyad031

Gene ontology vocabularies

- **Molecular Function**
  - What a product ‘does’, precise activity
- **Biological Process**
  - Biological objective, accomplished via one or more ordered assemblies of functions
- **Cellular Component**
  - ‘is located in’ (‘is a subcomponent of’)

Gene Ontology - coverage

Table 1. Changes to GO terms in the past two-year period. The ontology has undergone substantial revision and improvement, with nearly 2,000 terms added or removed.

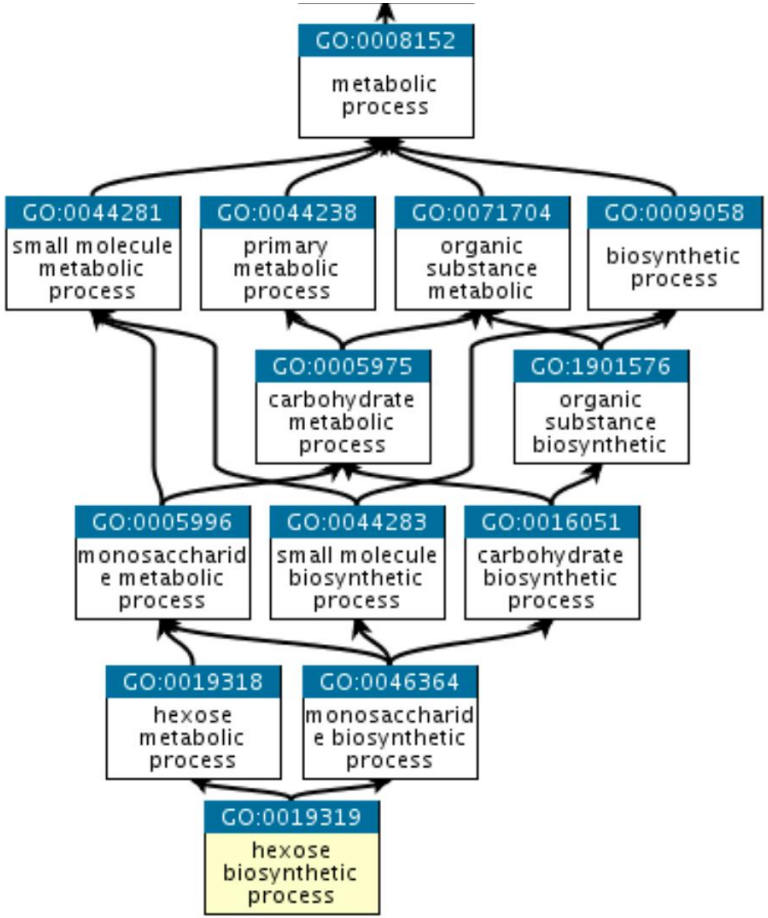
GO aspect	Total number of terms	Added terms	Obsoleted terms	Merged terms[a]
Molecular function	11,271	315	65	143
Cellular component	4,039	34	19	162
Biological process	27,993	217	782	254

Also includes obsoleted terms that have been replaced by another term.

The Gene Ontology Consortium (2023) Genetics
doi: 10.1093/genetics/iyad031

Gene Ontology - structure

- **Directed acyclic graph** (DAG): each child may have one or more parents
- **Relationships** between terms defined
- All terms are **defined**, accession ID associated with definition
- **True Path**: all attributes of children must hold for all parents



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

NUTRIOME

Gene Ontology - annotations

- GO annotations are created by associating **a gene or gene product** with a **GO term**
- Minimal information added by curator
  - Gene product (may be a protein, RNA, etc.)
  - GO term
  - Reference
  - Evidence (ECO ontology)

NUTRIOME

Gene Ontology - annotations

Term Information

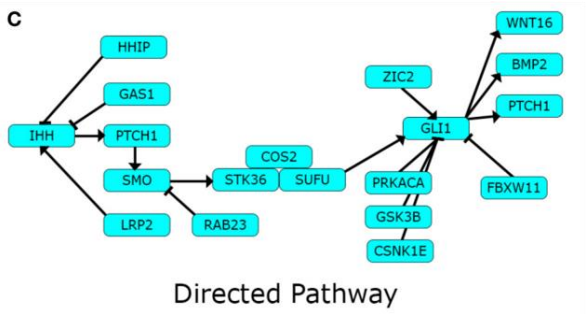
Accession GO:0006207
Name 'de novo' pyrimidine nucleobase biosynthetic
Ontology biological_process

Gene/product	Gene/product name	Annotation qualifier	GO class (direct)	Annotation extension	Contributor	Organism	Evidence
CPS1	Carbamoyl-phosphate synthase [ammonia], mitochondrial		'de novo' pyrimidine nucleobase biosynthetic process		InterPro	Homo sapiens	IEA
CMPK1	UMP-CMP kinase		'de novo' pyrimidine nucleobase biosynthetic process		InterPro	Homo sapiens	IEA
CAD	Multifunctional protein CAD		'de novo' pyrimidine nucleobase biosynthetic process		BHF-UCL	Homo sapiens	ISS
CAD	Multifunctional protein CAD		'de novo' pyrimidine nucleobase biosynthetic process		UniProt	Homo sapiens	IDA
CAD	Multifunctional protein CAD		'de novo' pyrimidine nucleobase biosynthetic process		GO_Central	Homo sapiens	IBA
MTOR	Serine/threonine-protein kinase mTOR		'de novo' pyrimidine nucleobase biosynthetic process		Ensembl	Homo sapiens	IEA
DHODH	Dihydroorotate dehydrogenase (quinone), mitochondrial		'de novo' pyrimidine nucleobase biosynthetic process		GO_Central	Homo sapiens	IBA
UMPS	Uridine 5'-monophosphate synthase		'de novo' pyrimidine nucleobase biosynthetic process		InterPro	Homo sapiens	IEA

Pathway databases

- Pathways are gene sets with **graphical representation** and information about the **relationships between the molecules**

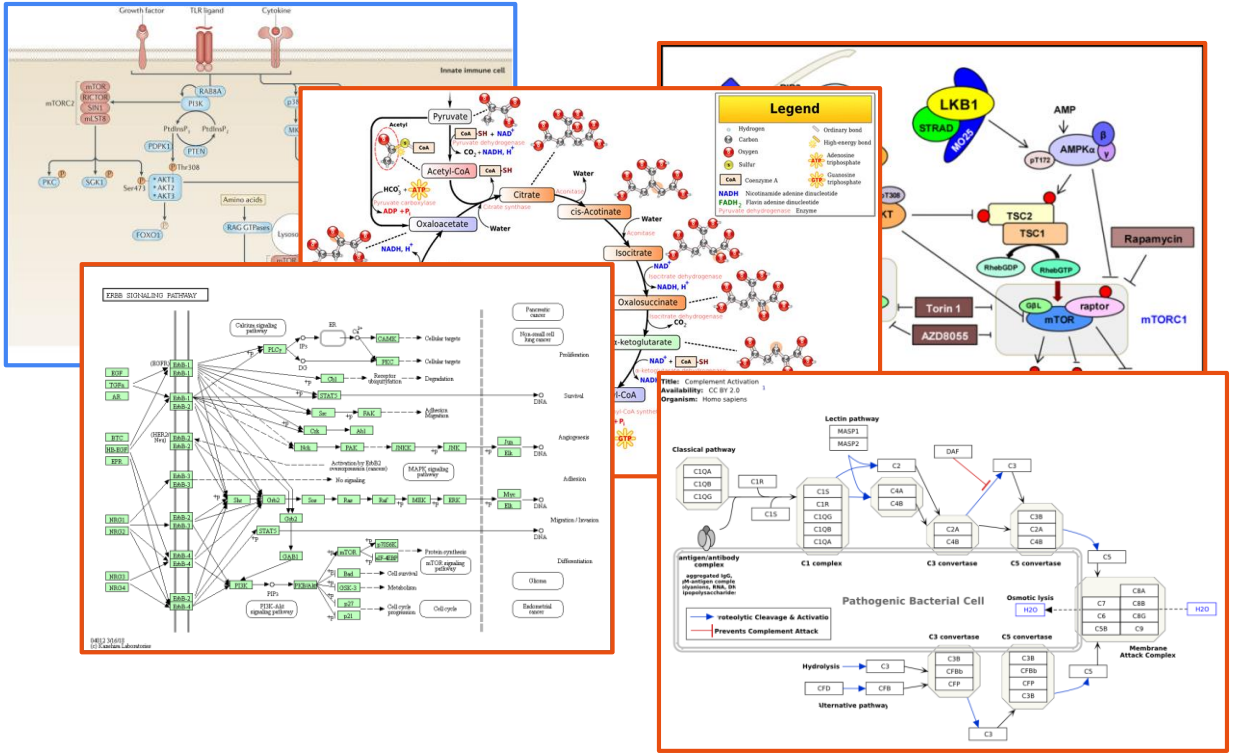
- Many different online databases (different species, biological focus, curation style)



García-Campos (2015) Front. Physiol.
doi: 10.3389/fphys.2015.00383

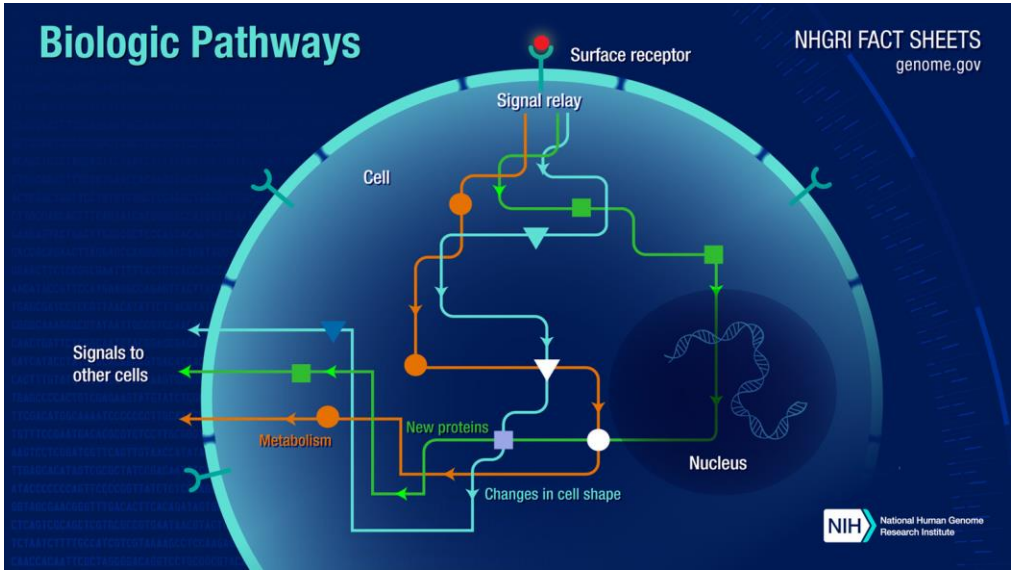
Biological pathways

Pathway diagrams are found everywhere!



Biological pathways

- Signaling pathways
- Metabolic pathways
- Gene regulation pathways



https://www.genome.gov/about-genomics/fact-sheets/Biological-Pathways-Fact-Sheet

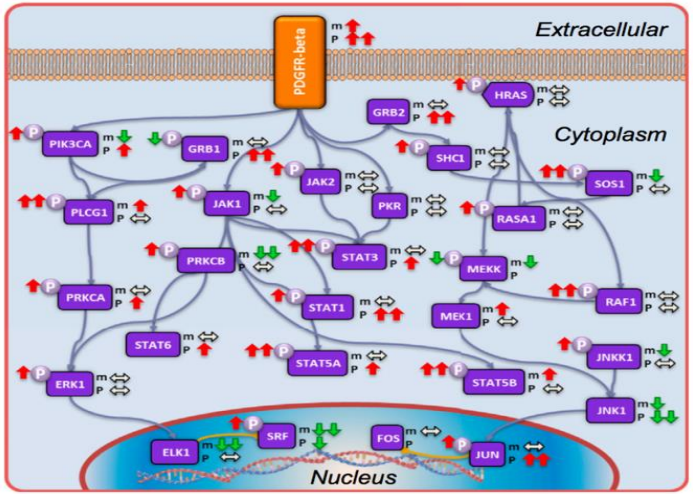
Biological pathways

Pathway diagrams are found everywhere!

Utility to biologists as conceptual models is obvious

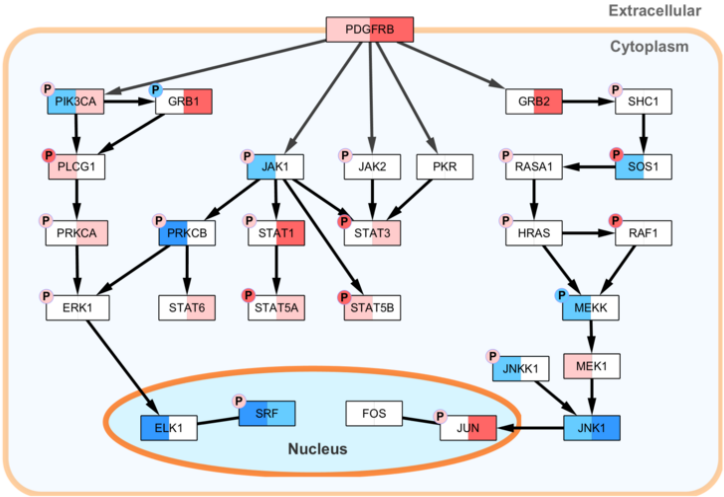
If modeled properly - immensely useful for computational analysis and interpretation of large-scale experimental data

Biological pathways



Static image

Zhang et al, Cell 2016



PDGFR-beta pathway with transcriptomic/phosphoproteomic data

www.wikipathways.org/instance/WP3972

Pathway Databases

PDB Name	Pathway focus	URL	Y.O.R.	Standard formats
EcoCyc	M,S	biocyc.org	1995	SBML, BioPAX
KEGG	M,S,D	kegg.jp	1996	BioPAX
RegulonDB	GR	regulondb.ccg.unam.mx	1997	BioPAX
MetaCyc	M	metacyc.org	1999	SBML, BioPAX
STRINGDB	PPI	string-db.org	2000	PSI-MI
PANTHER	S,D,PS	pantherdb.org	2004	SBML, SBGN, BioPAX
Gene Ontology	PPI,M,S	geneontology.org	2000	
REACTOME	M,S,D	reactome.org	2005	SBML, SBGN, BioPAX, PSI-MI
MSigDb	M,S,GR	broadinstitute.org/gsea/msigdb	2005	
Ingenuity Knowledge Base*	PPI,PCI,M,S,GR,D	ingenuity.com	2005	
NCI PID	S,D	pid.nci.nih.gov	2006	BioPAX
WikiPathways	M,S,D	wikipathways.org	2008	BioPAX
Small Molecule Pathway DB	M,S	smpdb.ca	2009	SBML, BioPAX
ConsensusPathDB	PPI,PCI,M,S,GR	consensuspathdb.org	2009	BioPAX, PSI-MI
Pathway Commons	PPI,PCI,M,S	pathwaycommons.org	2010	BioPAX

*A brief example of the diversity of available PDBs found online. The second column shows the kind of biological focus pursued by each database: (PPI, protein-protein intereactions; PCI, protein-compound interactions; M, metabolic; S, signaling; GR, gene regulation; D, diagrams; PS, protein sequence). The last column addresses the standard pathway languages adopted to provide data. Additionally, in the third column the links to web sites are supplied. YOR, Year of release. *Commercial database.*

García-Campos (2015) Front. Physiol.
doi: 10.3389/fphys.2015.00383

WikiPathways

- Launched in 2008 as an experiment in community-based curation of biological pathways

Too much data!
Difficult to keep knowledge up-to-date, accessible and integrated

Taking advantage of direct participation by a greater portion of the community (**crowdsourcing**)

Image:
https://www.vizioninteractive.com/blog/data-overload-when-it-all-becomes-too-much/

WikiPathways

- Community-curated
- Collaborative
- Open

Content:

- 1,958 pathways
- 27 species
- 600+ editors

www.wikipathways.org

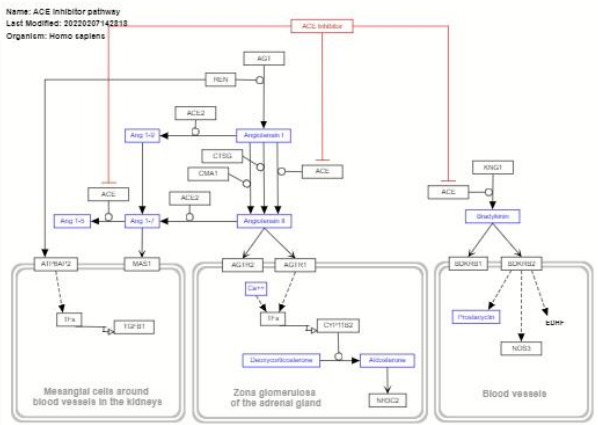
Agrawal (2024) NAR
doi: 10.1093/nar/gkad960

ACE inhibitor pathway (WP554)

Homo sapiens

The core of this pathway was elucidated over a century ago and involves the conversion of angiotensinogen to angiotensin I (Ang I) by renin, its subsequent conversion to angiotensin II (Ang II) by angiotensin converting enzyme. Ang II activates the angiotensin II receptor type 1 to induce aldosterone synthesis, increasing water and salt resorption and potassium excretion in the kidney and increasing blood pressure. Source: PharmGKB

more text



last edited 04 Aug 2022 curation success

view full size download options

Authors

Caroline F. Thorn , Jinwook.seo , Kristina Hanspers , Alex Pico , Thomas Kelder , Martijn Van Iersel , Egon Willighagen , Christine Chichester , Nuno , Denise Slenter , Martina Summer-Kutmon , and Eric Weitz

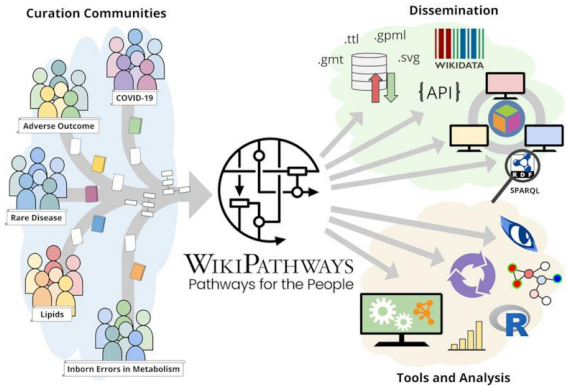
Cited In

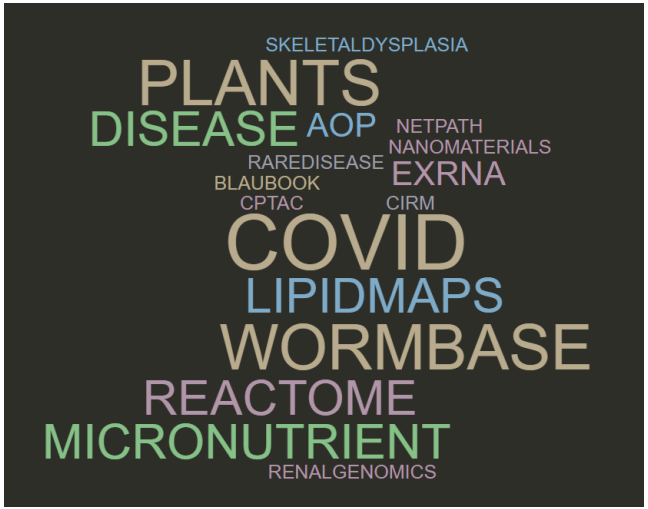
PMC PMC7982796 PMC PMC7360763 PMC PMC4338111

Are you planning to include this pathway in your next publication? See How to Cite and add a link here to your paper once it's online.

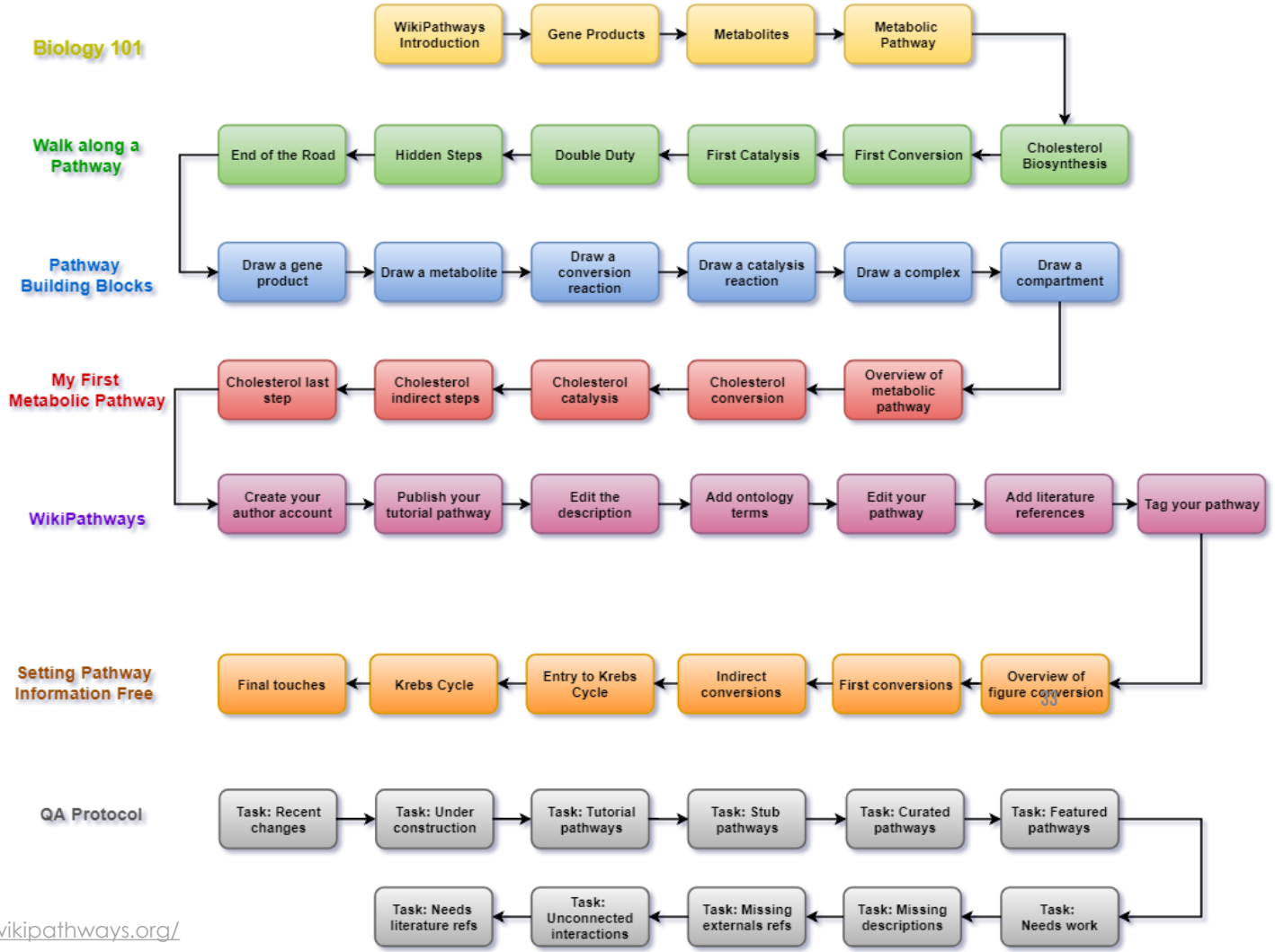
Community portals

- Special interest groups
- Portal pages to highlight communities
- 15 community portals supported





Martens (2021) NAR
doi: 10.1093/nar/gkaa1024



NUTRIOME

https://academy.wikipathways.org/

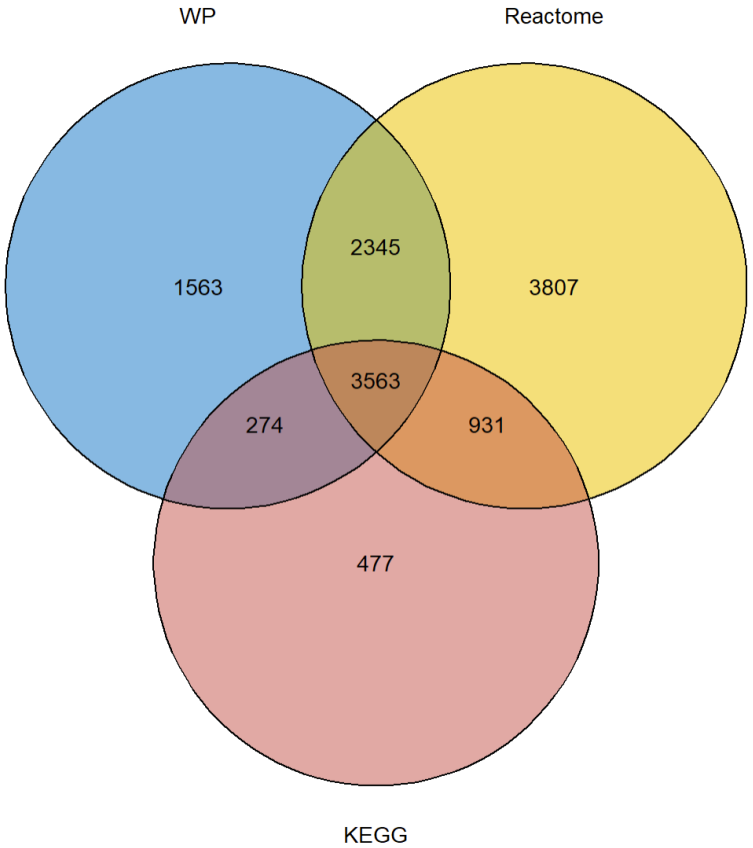
Pathway databases - coverage

MSigDb

Human MSigDB v2023.2.Hs

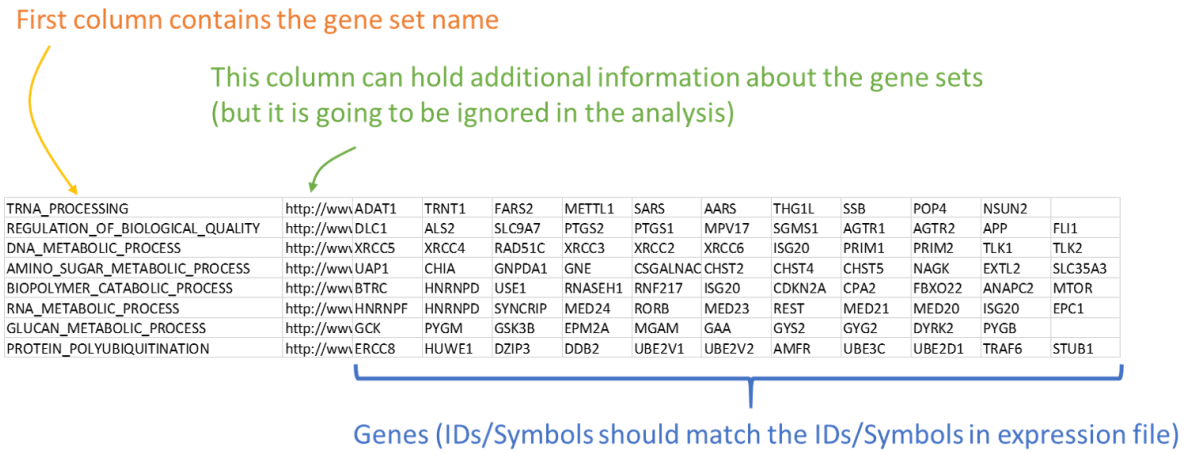
19,846 protein coding genes (Ensembl GRCh38.p14)

Genes in at least one pathway of the three databases
→ 12,960 genes (65%)



File format

- GMT (Gene Matrix Transposed) file format
  - tab delimited file format
  - each row represents a gene set
  - gene set name | description | gene list (one gene per column)



TRNA_PROCESSING	http://ww	ADAT1	TRNT1	FARS2	METTL1	SARS	AARS	THG1L	SSB	POP4	NSUN2	
REGULATION_OF_BIOLOGICAL_QUALITY	http://ww	DLC1	ALS2	SLC9A7	PTGS2	PTGS1	MPV17	SGMS1	AGTR1	AGTR2	APP	FLI1
DNA_METABOLIC_PROCESS	http://ww	XRCC5	XRCC4	RAD51C	XRCC3	XRCC2	XRCC6	ISG20	PRIM1	PRIM2	TLK1	TLK2
AMINO_SUGAR_METABOLIC_PROCESS	http://ww	UAP1	CHIA	GNPDA1	GNE	CSGALNAC	CHST2	CHST4	CHST5	NAGK	EXTL2	SLC35A3
BIOPOLYMER_CATABOLIC_PROCESS	http://ww	BTRC	HNRNPD	USE1	RNASEH1	RNF217	ISG20	CDKN2A	CPA2	FBXO22	ANAPC2	MTOR
RNA_METABOLIC_PROCESS	http://ww	HNRNPF	HNRNPD	SYNCRIP	MED24	RORB	MED23	REST	MED21	MED20	ISG20	EPC1
GLUCAN_METABOLIC_PROCESS	http://ww	GCK	PYGM	GSK3B	EPM2A	MGAM	GAA	GYS2	GYG2	DYRK2	PYGB	
PROTEIN_POLYUBIQUITINATION	http://ww	ERCC8	HUWE1	DZIP3	DDB2	UBE2V1	UBE2V2	AMFR	UBE3C	UBE2D1	TRAF6	STUB1

NUTRIOME

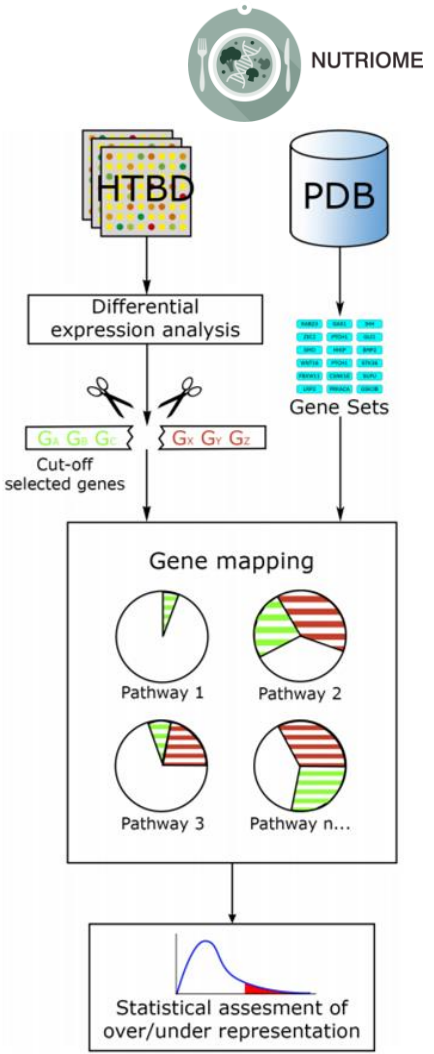
Pathway enrichment

Funded by
the European Union

Marie
Skłodowska-
Curie Actions

Over Representation Analysis (ORA)

- **Methodology**
  - Use parametric statistics to identify differentially regulated molecules, e.g. limma
  - Choose significance level e.g. FDR < 0.05, FC > 1.5
  - Use parametric statistics to identify annotations over represented within your list compared to what was assayed e.g. Fisher's exact test

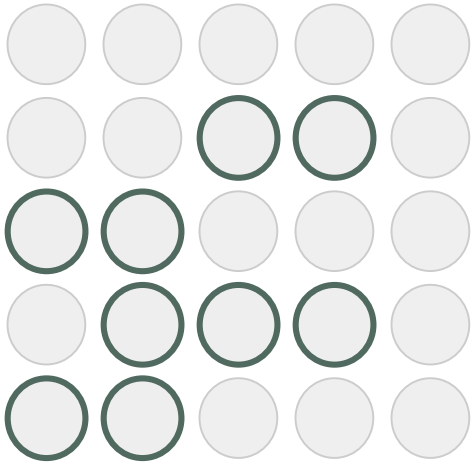


García-Campos (2015) Front. Physiol.
doi: 10.3389/fphys.2015.00383

Over Representation Analysis (ORA)

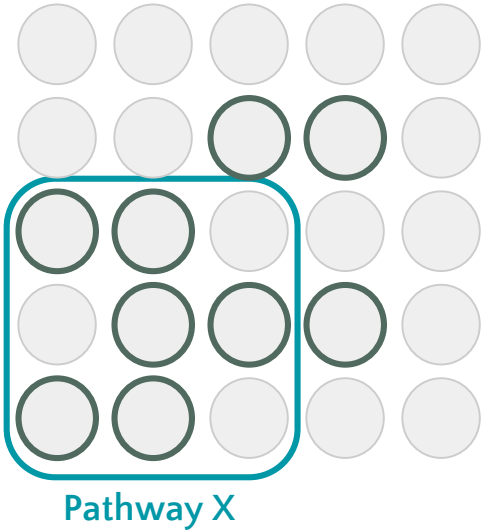
- **N = 25**
  background list (total number of measured genes in experiment)

Over Representation Analysis (ORA)



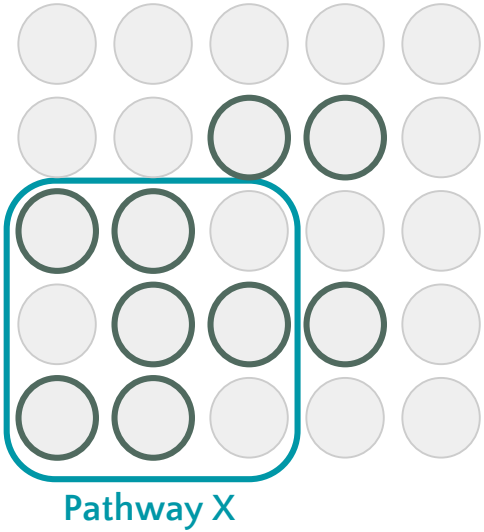
- N = 25
  background list (total number of measured genes in experiment)
- **R = 9**
  input list (number of changed genes in experiment)

Over Representation Analysis (ORA)



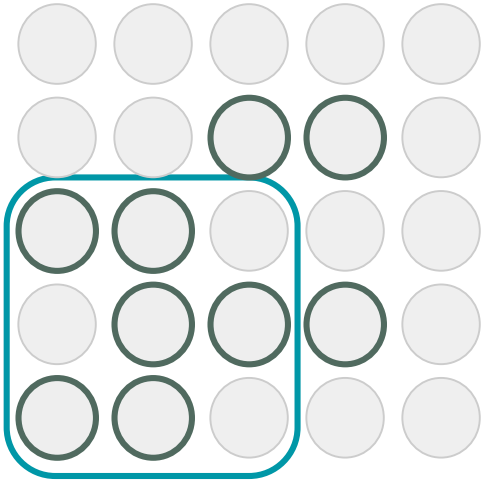
- N = 25
  background list (total number of measured genes in experiment)
- R = 9
  input list (number of changed genes in experiment)
- **n = 9**
  total number of genes in pathway

Over Representation Analysis (ORA)



- N = 25
  background list (total number of measured genes in experiment)
- R = 9
  input list (number of changed genes in experiment)
- n = 9
  total number of genes in pathway
- **r = 6**
  number of changed genes in pathway

Over Representation Analysis (ORA)



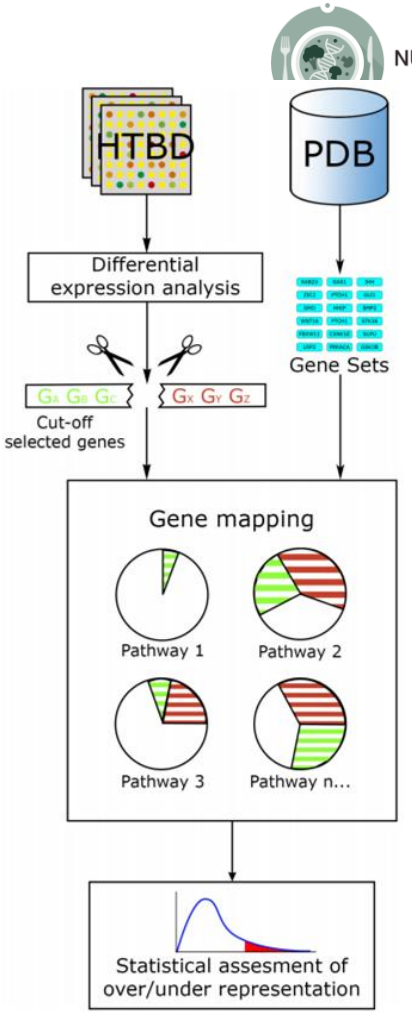
- N = 25
  background list (total number of measured genes in experiment)
- R = 9
  input list (number of changed genes in experiment)
- n = 9
  total number of genes in pathway
- r = 6
  number of changed genes in pathway

$$\text{Z-score} = \frac{(r - n\frac{R}{N})}{\sqrt{n\frac{R}{N}(1-\frac{R}{N})(1-\frac{n-1}{N-1})}}$$

Enrichment score (e.g. Z-score)

Over Representation Analysis (ORA)

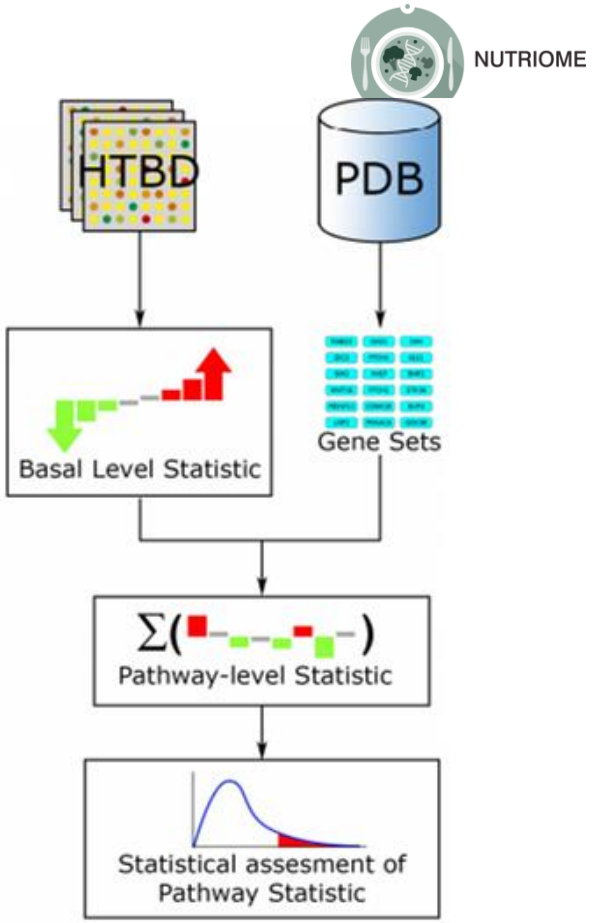
- **Caveats**
  - Threshold
    - what about the transcript with p = 0.050001, FC = 1.4999
  - Equality
    - transcript-X with p = 0.0000001, FC = 100 considered equal to transcript-Y with p = 0.049, FC = 1.51
  - Assumption of independence between both genes and pathways inflates significance
  - Ignores relationships between genes/gene products
  - Significance increases with population size



García-Campos (2015) Front. Physiol.
doi: 10.3389/fphys.2015.00383

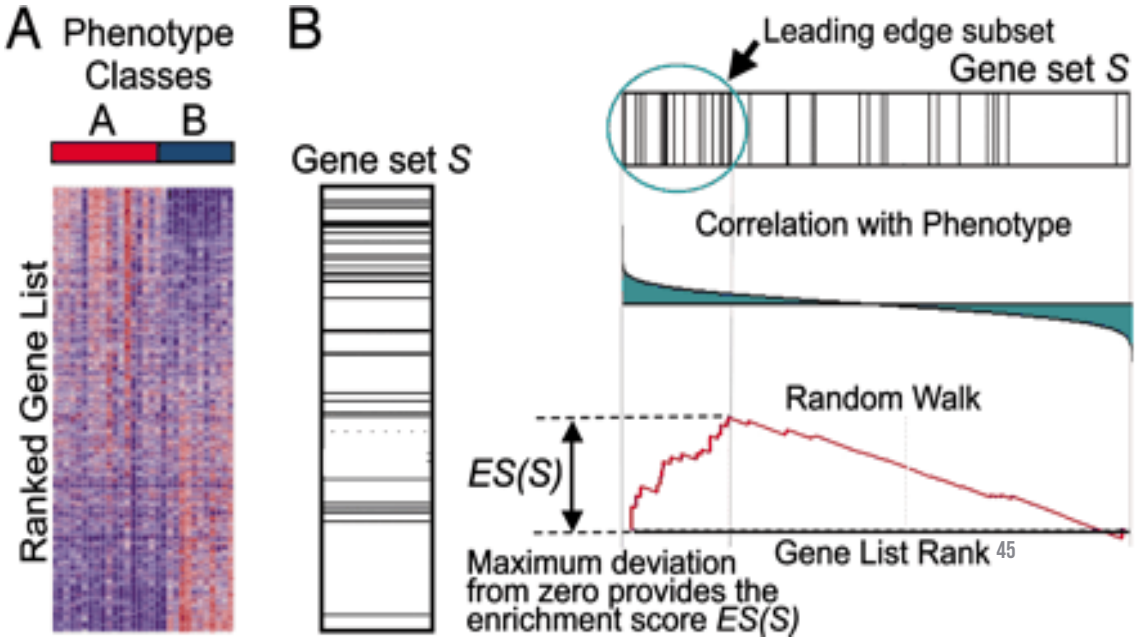
Functional Class Scoring (FCS)

- **Methodology**
  - Use parametric statistics to determine differential regulation for all molecules e.g. t-distribution statistics
  - Use various statistics to combine gene statistics and determine pathway statistics e.g. Wilcoxon rank sum, Kolmogorov-Smirnov
  - Permutes phenotypes and pathways to determine pathway significance
- **Applications**
  - Gene Set Enrichment analysis (GSEA)



García-Campos (2015) Front. Physiol.
doi: 10.3389/fphys.2015.00383

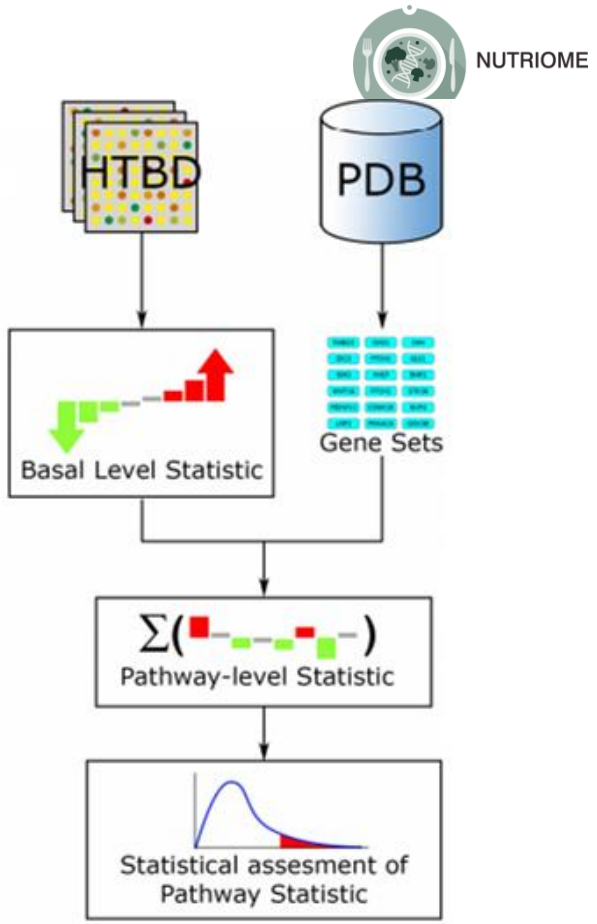
Functional Class Scoring (FCS)



Subramanian (2015) PNAS
doi: 10.1073/pnas.0506580102

Functional Class Scoring (FCS)

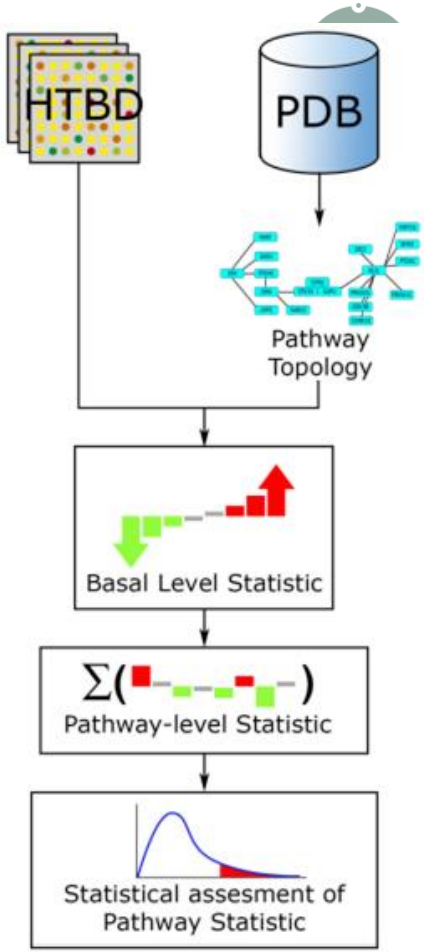
- **Caveats**
  - Assumes independence between pathways
  - Dependence on ranking approaches miss magnitude of changes between phenotypes, i.e., sham FC = 10; treated similar FC = 100
  - Ignores relationships between genes/gene products



García-Campos (2015) Front. Physiol.
doi: 10.3389/fphys.2015.00383

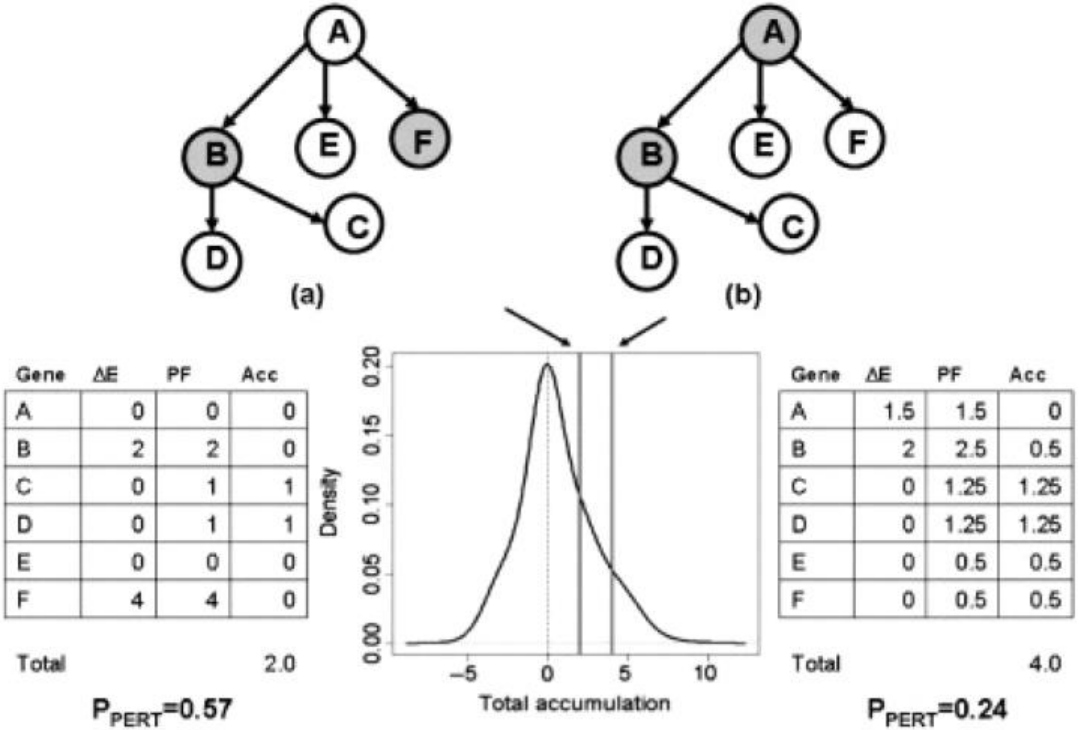
Pathway Topology Based (PTB)

- **Methodology**
  - Use various statistics to determine differences in gene-gene interactions (node-edge-node) for all genes (e.g. Pearson's correlation)
  - Use various statistics to combine gene interaction statistics and determine pathway significance e.g. permutation, hypergeometric distribution
- **Applications**
  - pathfindR
  - SPIA
  - PathNet



García-Campos (2015) Front. Physiol.
doi: 10.3389/fphys.2015.00383

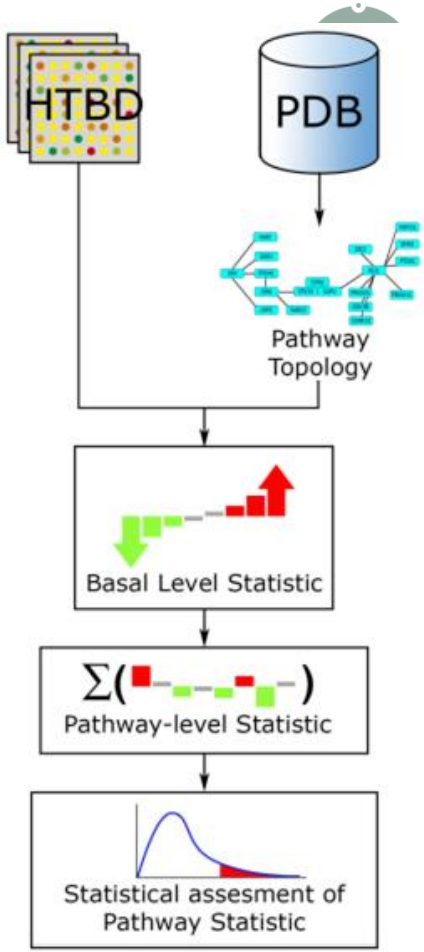
Pathway Topology Based (PTB) - SPIA



Adi Laurentiu Tarca (2009) Bioinformatics
doi: 10.1093/bioinformatics/btn577

Pathway Topology Based (PTB)

- **Caveats**
  - Limited interaction knowledge, i.e., thus hampered by immature interaction databases (KEGG, BioCarta, Reactome, PantherDB etc.)
  - Not to mention a lack of cellular and temporal resolution of interactions.



García-Campos (2015) Front. Physiol.
doi: 10.3389/fphys.2015.00383

Interpretation

NUTRIOME

- **Be aware!**
  - ORA and FCS do not take pathway topology into account!
  - You don’t know yet where the changes occur in the pathway.
  - If you have pathway models > always look at the pathway diagrams and study the changes to make the right conclusions!

Tools

- **Many, many different tools** to perform pathway analysis!
  - Integrated in resources
  - Standalone applications
  - Packages (R / Python / Perl / etc.)

- **Practical**:
  - R-package clusterProfiler
  - implements GSEA and ORA

NUTRIOME

Interpretation and visualization of results

Funded by the European Union

Marie Skłodowska-Curie Actions

Analysis results

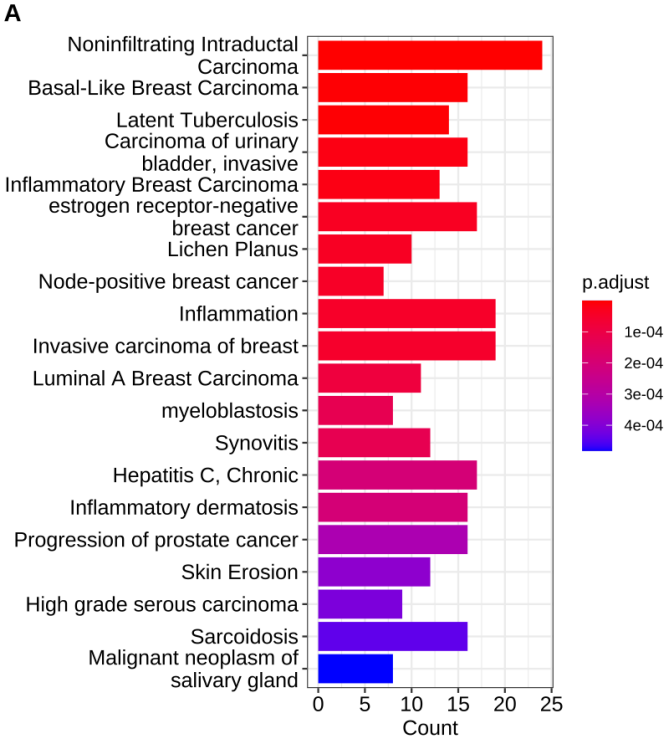
- Table view

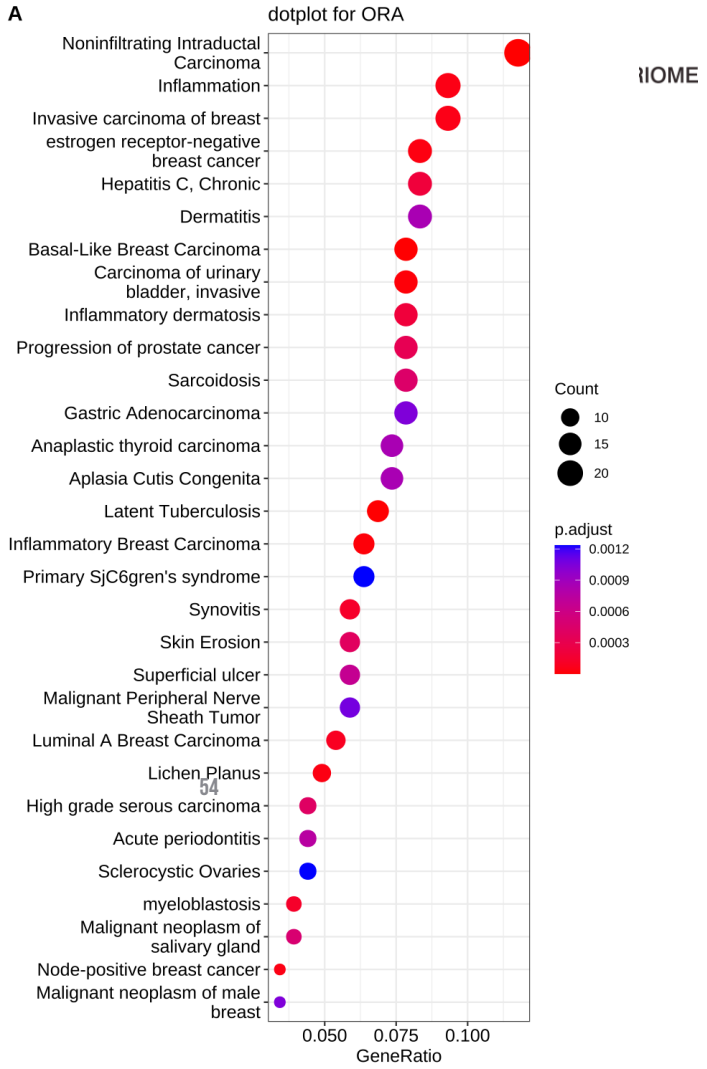
KEGG		stats					
Term name	Term ID	p_{adj}	$-\log_{10}(p_{adj})$ 0 ≤16	T	Q	T∩Q	U
Glycerolipid metabolism	KEGG:00561	5.247×10^{-6}		61	16	5	8014
RNA degradation	KEGG:03018	8.620×10^{-4}		79	16	4	8014
mRNA surveillance pathway	KEGG:03015	1.945×10^{-3}		97	16	4	8014
Glycerophospholipid metabolism	KEGG:00564	2.026×10^{-3}		98	16	4	8014

VEGF-activated neuropilin signaling pathway	2	2	.00	> 100	+	2.08E-05	2.72E-03
↳neuropilin signaling pathway	4	2	.01	> 100	+	5.20E-05	5.43E-03
↳cell surface receptor signaling pathway	2106	17	3.99	4.26	+	8.49E-08	4.75E-05
↳signal transduction	4820	22	9.13	2.41	+	9.64E-06	1.61E-03
↳signaling	5163	22	9.78	2.25	+	4.31E-05	4.73E-03
↳cell communication	5266	23	9.97	2.31	+	1.06E-05	1.69E-03
↳cellular response to stimulus	6479	25	12.27	2.04	+	3.57E-05	4.18E-03
↳response to stimulus	8096	32	15.34	2.09	+	6.23E-08	3.91E-05
↳regulation of cellular process	11275	35	21.36	1.64	+	5.27E-06	1.01E-03
↳regulation of biological process	11735	35	22.23	1.57	+	1.29E-05	1.91E-03
↳biological regulation	12469	35	23.62	1.48	+	9.83E-05	8.81E-03

Analysis results

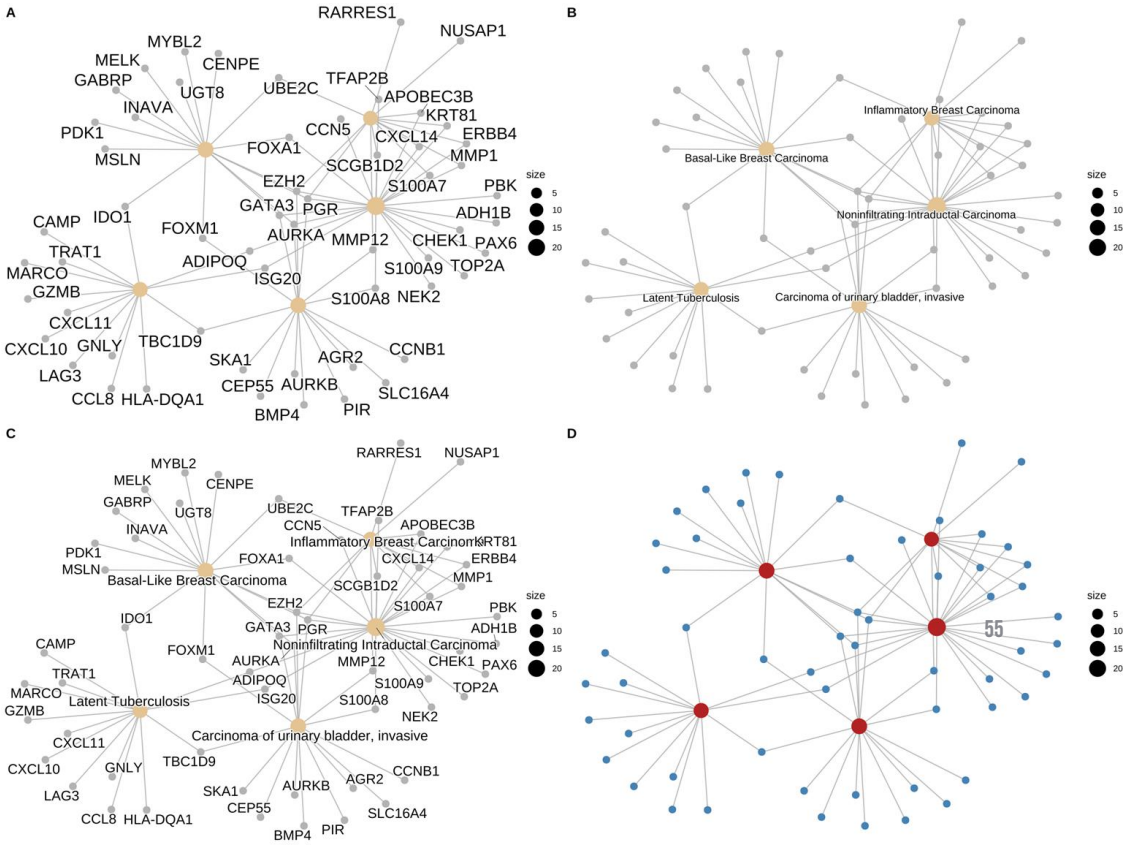
- Bar or Dot plots





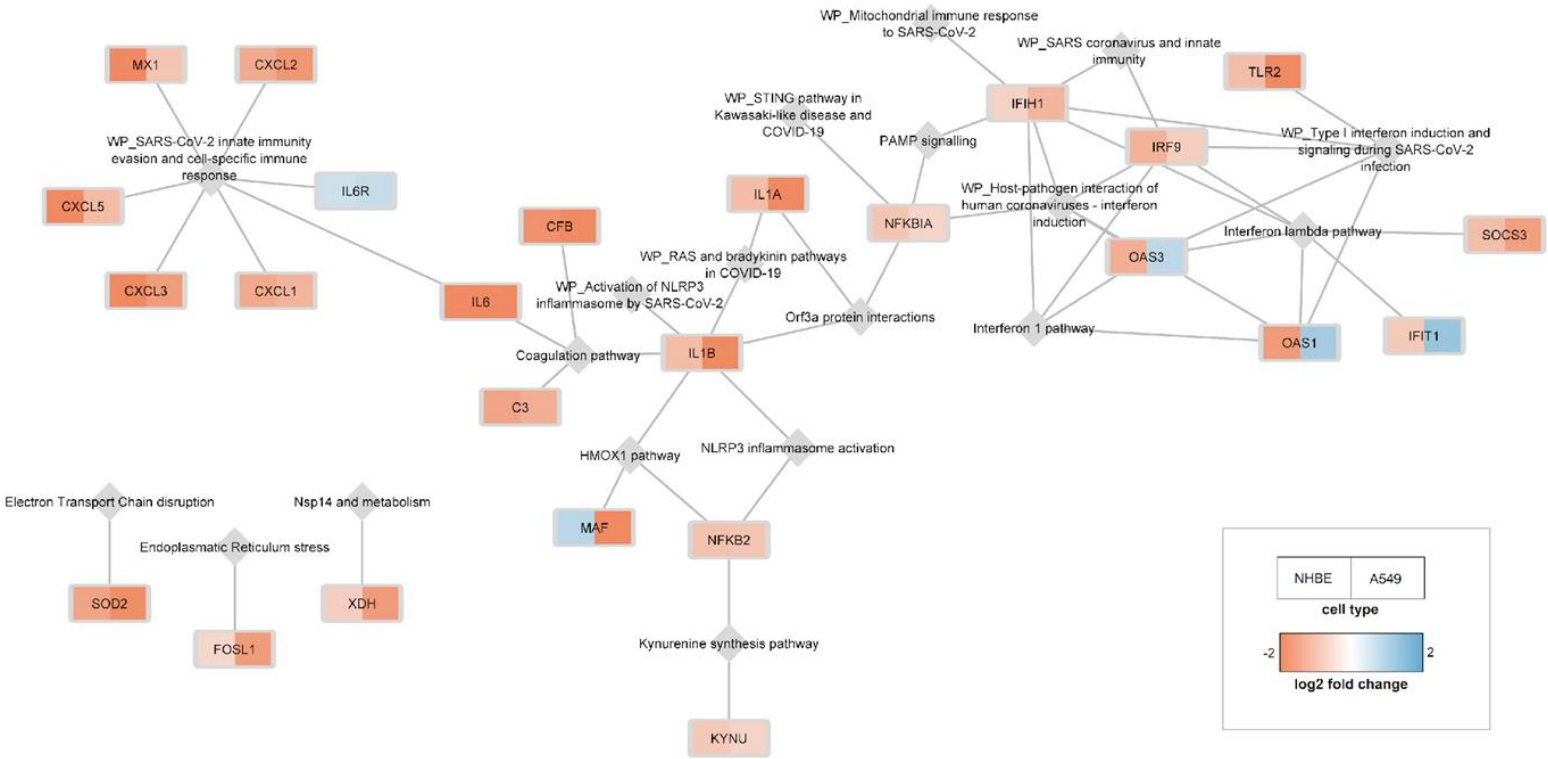
https://yulab-smu.top/biomedical-knowledge-mining-book/enrichplot.html

Gene-concept networks



https://yulab-smu.top/biomedical-knowledge-mining-book/enrichplot.html

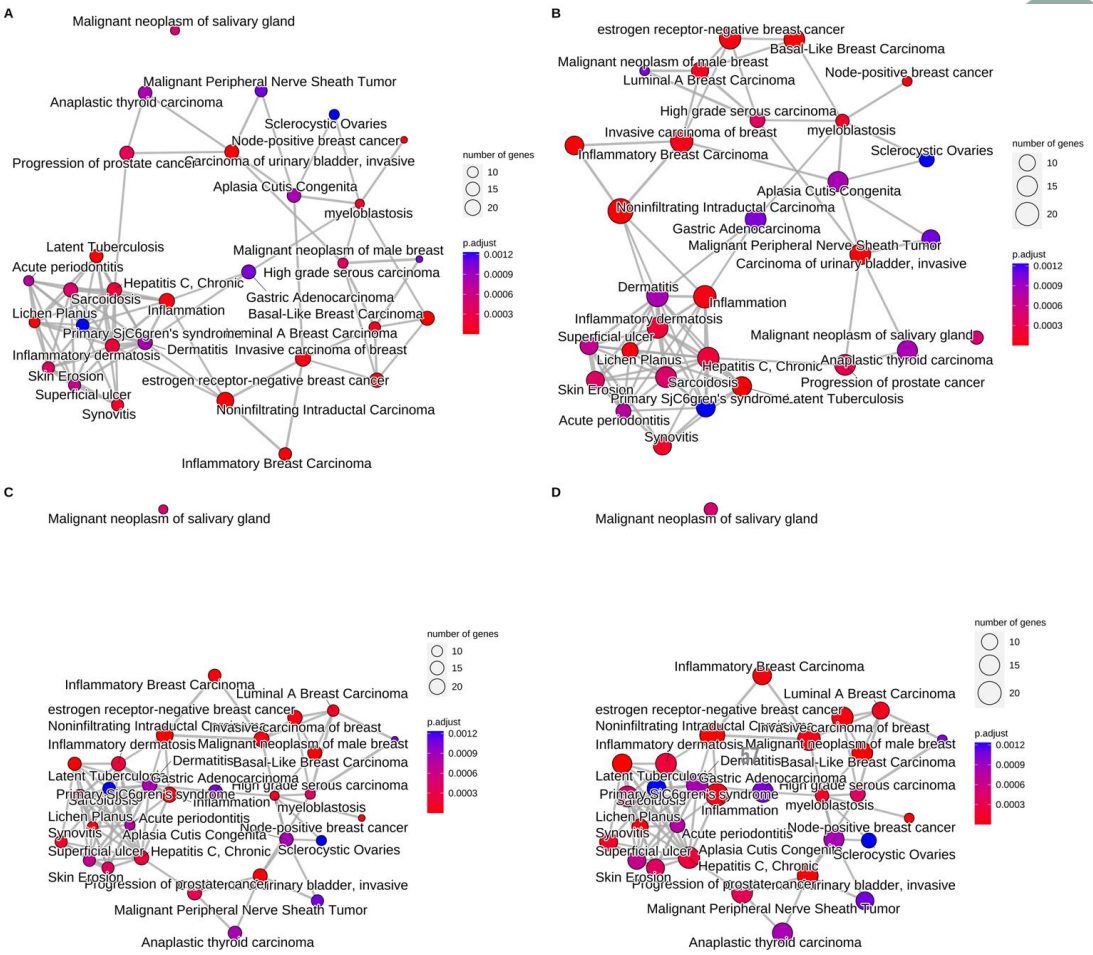
Gene-concept networks



Cytoscape

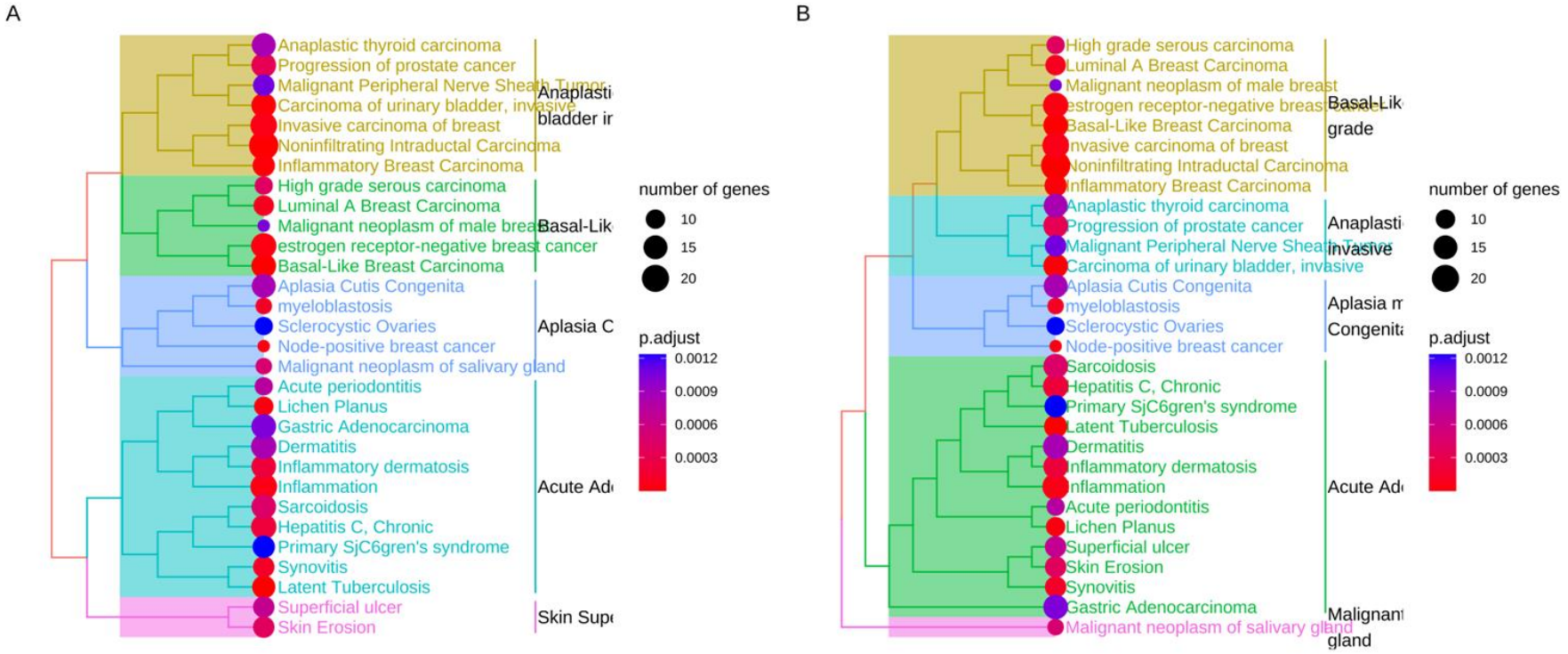
Niarakis (2023) Frontiers Immunology
doi: 10.3389/fimmu.2023.1282859

Enrichment maps



https://yulab-smu.top/biomedical-knowledge-mining-book/enrichplot.html

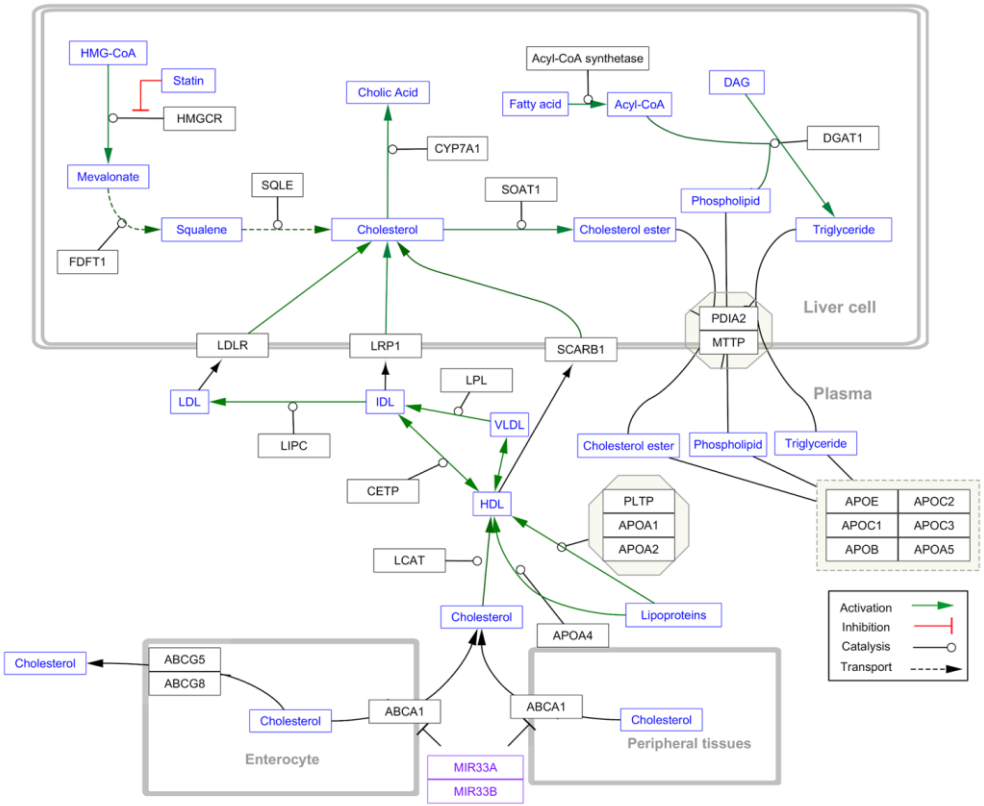
Tree plots



https://yulab-smu.top/biomedical-knowledge-mining-book/enrichplot.html

Data visualization in Cytoscape

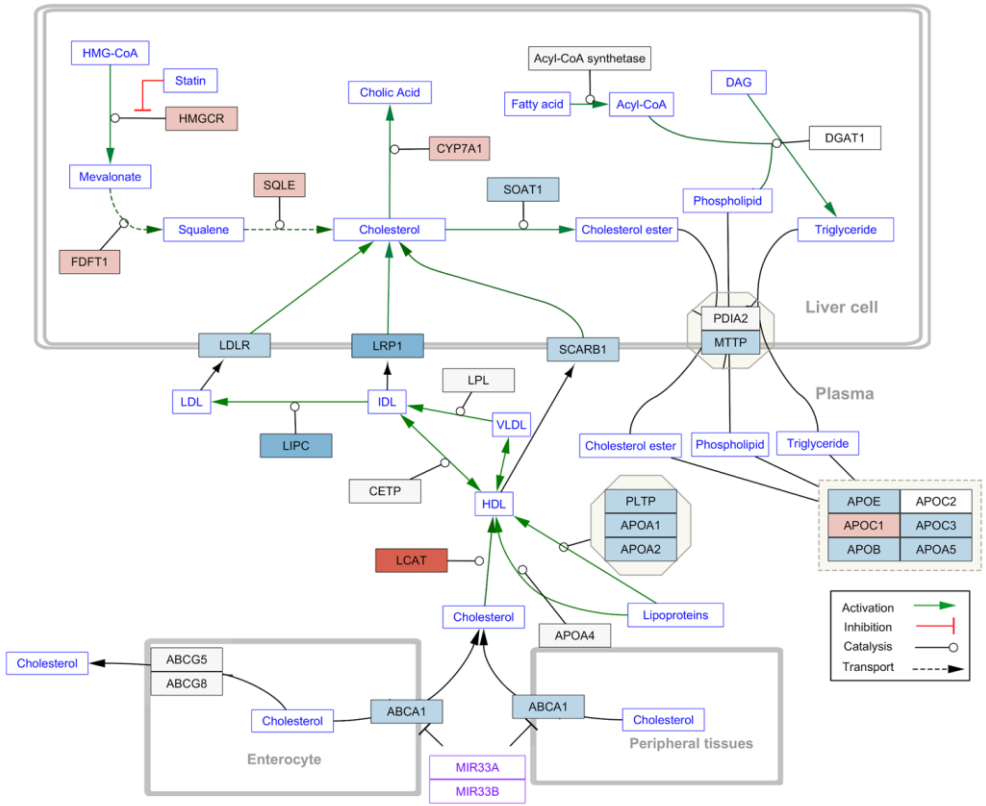
NUTRIOME



Cytoscape

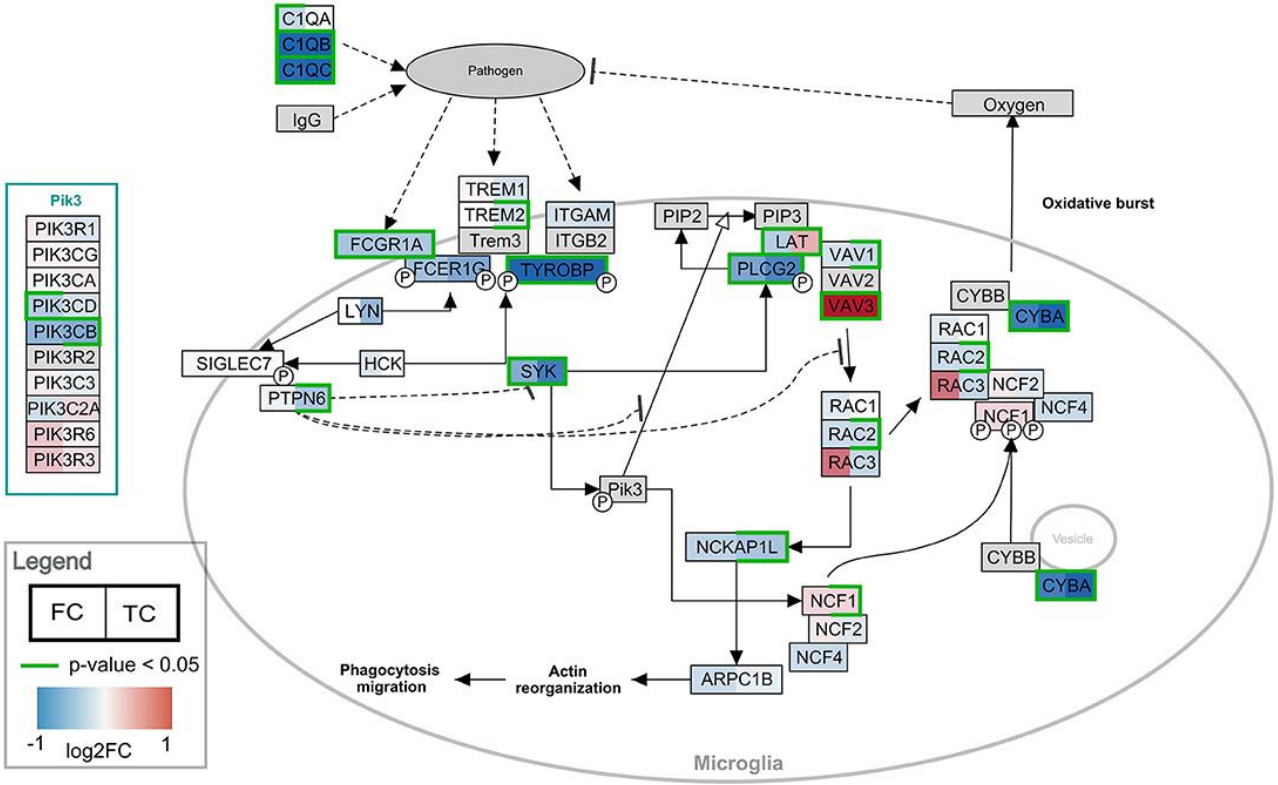
Data visualization in Cytoscape

NUTRIOME



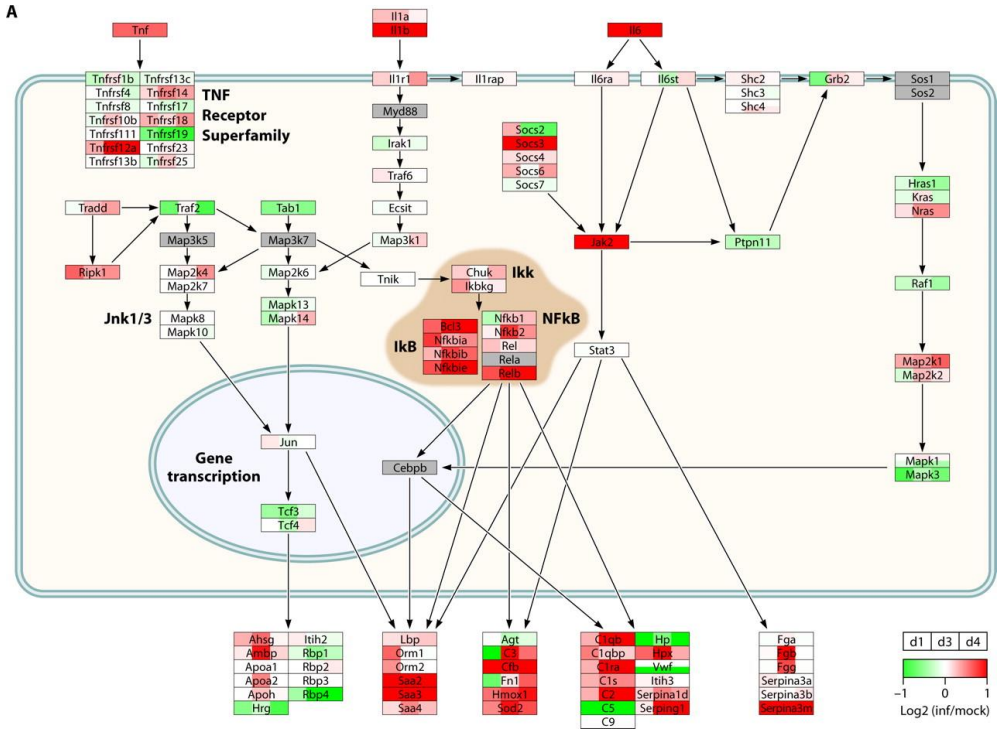
Cytoscape

Multiple comparisons



Miller (2019) Frontiers Genetics
doi: 10.3389/fgene.2019.00059

Time-series data



Tisoncik (2012) Microbiol Mol Biol Rev.
doi: 10.1128/MMBR.05015-11

NUTRIOME

Questions?

Martina Summer-Kutmon
martina.kutmon@maastrichtuniversity.nl
Maastricht Centre for Systems Biology (MaCSBio)

Maastricht University

Funded by the European Union

Marie Skłodowska-Curie Actions

macsbio
MAASTRICHT CENTRE FOR SYSTEMS BIOLOGY