An Introduction to Pathway Enrichment Analysis

Alex Sánchez

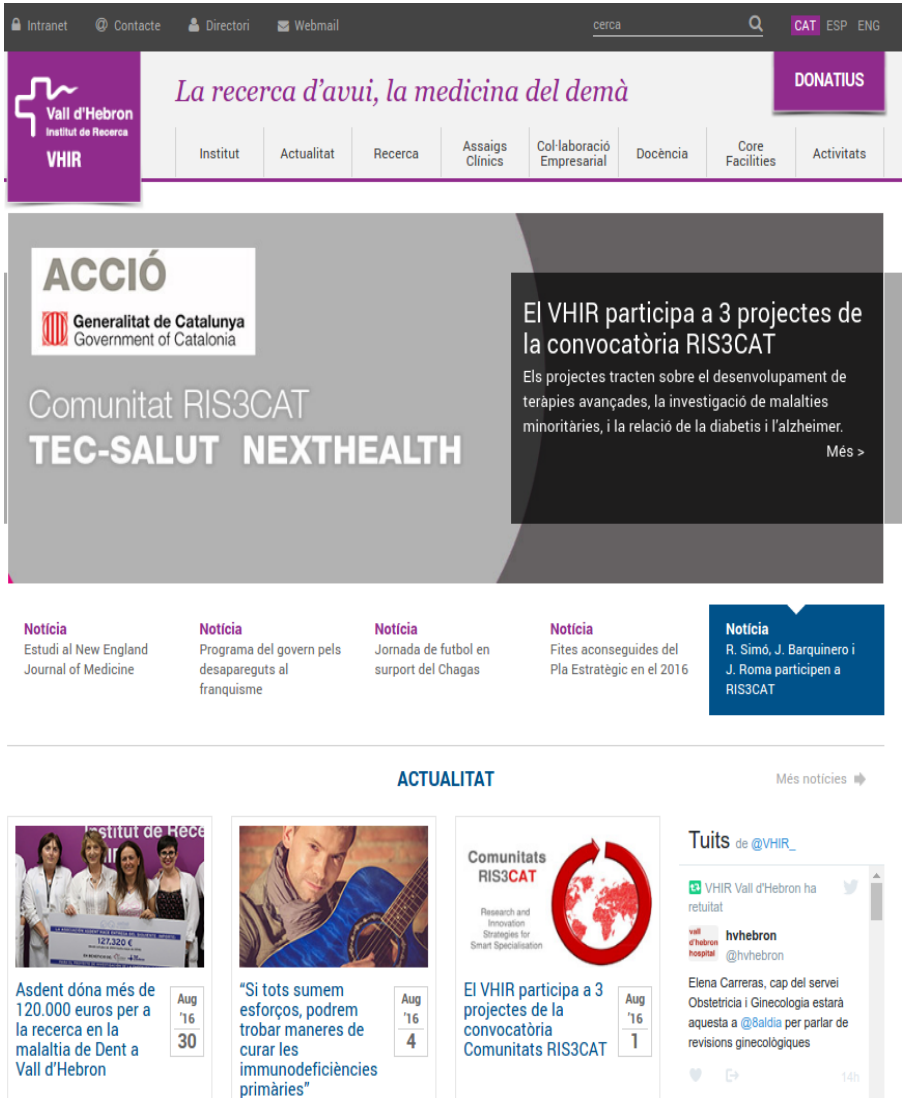
Statistics and Bioinformatics Unit
Vall d'Hebron Institut de Recerca

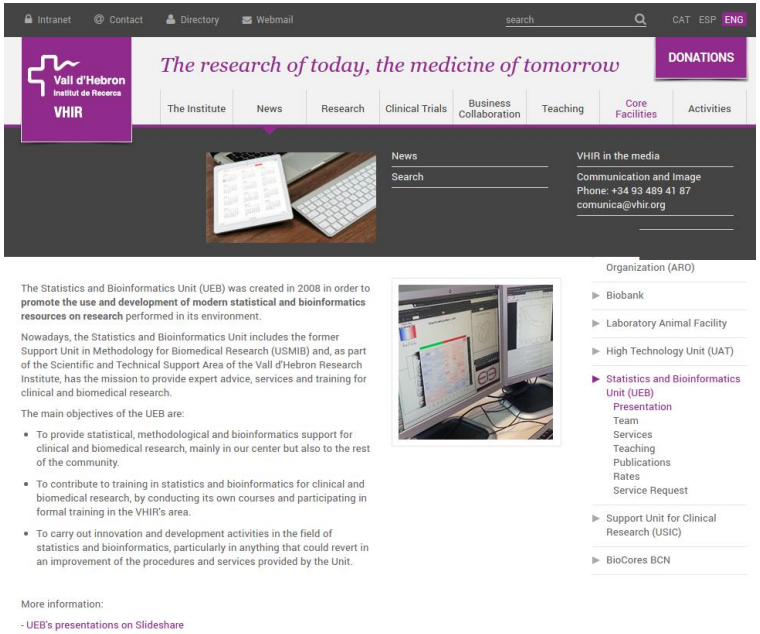
Statistics and Bioinformatics Research Group
Statistics department, Universitat de Barelona

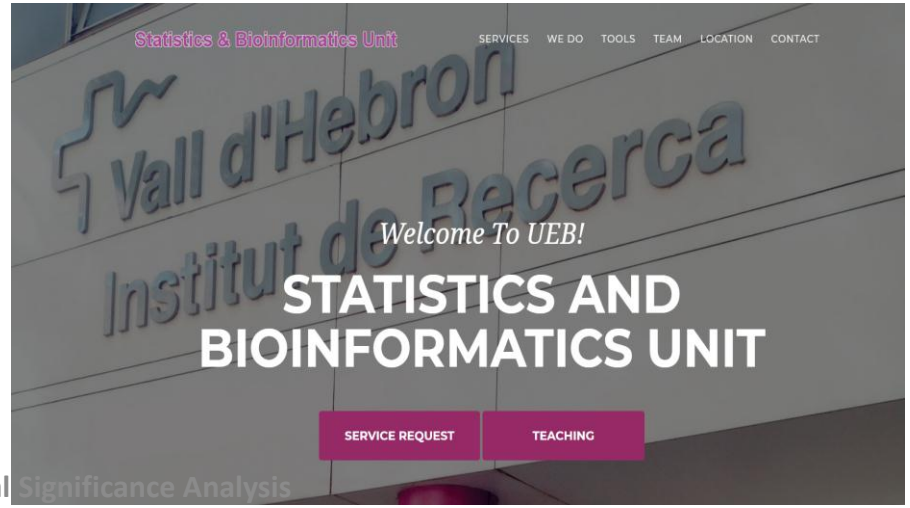
Outline

- Presentation
- Introduction and Background
  - Gene lists, Identifiers and Pathway databases
- Pathway Analysis: Methods and Tools
  - Overrepresentation analysis and GSEA
  - Multiple Testing Adjustments
  - Network Visualization and Enrichment Map
- A protocol for Pathway Enrichment Analysis
- A user experience

The Statistics and Bioinformatics Unit







An Overview of Biological Significance Analysis ("Alternatives to IPA")

UEB people is here to help you!

2009 UEB creation
- Personal: 3

2010 UEB becomes a platform of the UAT

2011
- Personal increases: unit head and 3 technicians
- Begin to work with a Tiki (not normalized protocols)

2014
- UEB becomes a core facility
- UEB + USMIB
- Personal: Unit head and 4 technicians

2017 Certificació ISO 9001:2008

How we can help you

We provide support in …

Clinical Data Analysis

- Biostatistical Analysis
- Clinical Trials
- CRF development (Redcap)
- Epidemiological studies
- *Data Management for Clinical Trials*

Omics Data Analysis & Bioinformatics

- Gene Expression Analysis
- Methylation
- Metagenomics
- Exome sequencing
- Integrative Omics
- Database / applicationsn development
- High performance Computing (HPC) services

Training

- Short Workshops
- Courses
- Official training (MSc)
- Students in practice

Consulting

- Sample size
- Experimental design
- GRANT review
- Statistical parts of papers

Short consultations (< 3hrs) are free

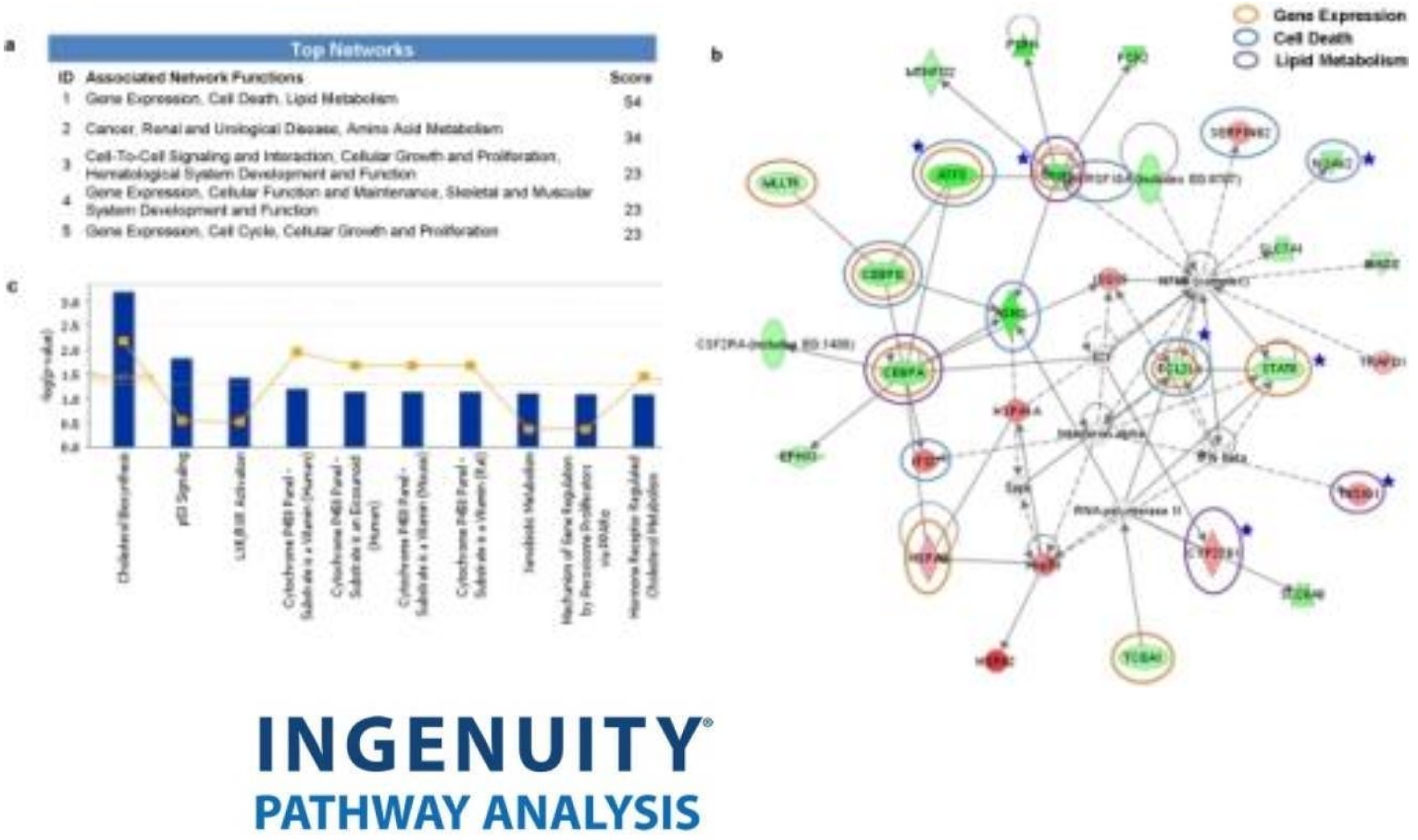
Other services budgeted according our rates:

http://ueb.vhir.org/Services

Acknowledgements and disclaimer

- This presentation has been developed along several years of teaching some lectures on Biological Significance Analysis.
- For this I have borrowed materials from many –now unknown people- and I wish to thank them for providing the materials.
- Two sources that I must acknowledge specifically
  - Bioinformatics.ca for sharing their course materials which I have some times re-used (that is copied) directly.
  - Stéphane Nemours, from the Renal Phisiopathology lab @ VHIR for the help in preparing this sesion and sharing his slides.
- Blame me for any errors, not them.

Once upon a time there was *IPA*



INGENUITY®
PATHWAY ANALYSIS

An Overview of Biological Significance Analysis
("Alternatives to IPA")

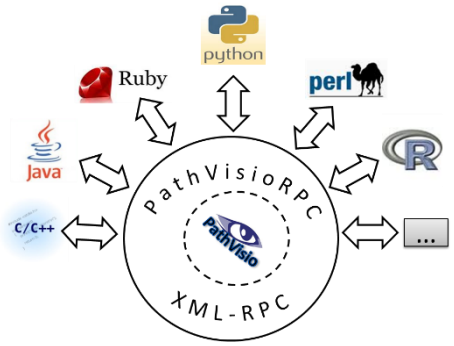
But we don't have it anymore

IPA has been discontinued January 2019

Main reasons

- Problems with license complying
- High Cost

There are many alternatives ...

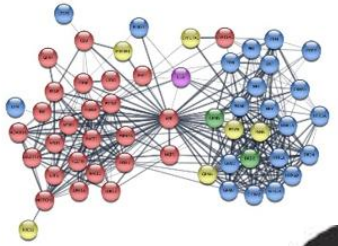


Protocol | Published: 21 January 2019

Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap

Network visualization

A crash course on using Cytoscape



lome » BiocViews

All Packages

ioconductor version 3.9 (Release)

utocomplete biocViews search:

- Software (1741)
  - AssayDomain (698)
  - BiologicalQuestion (708)
  - Infrastructure (382)
  - ResearchField (775)
  - StatisticalMethod (613)
  - Technology (1103)
  - WorkflowStep (936)
- AnnotationData (948)
- ExperimentData (371)
- Workflow (27)

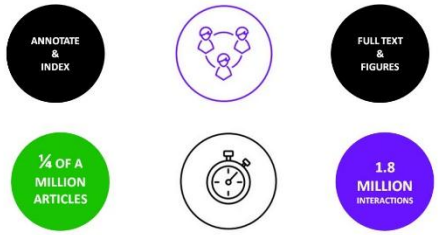
Packages found under GeneSetEnrichment:

Rank based on number of downloads: lower numbers are more frequently downloaded.

Show All entries Search table:

Package	Maintainer	Title	Rank
limma	Gordon Smyth	Linear Models for Microarray Data	10
edgeR	Yunshun Chen, Aaron Lun, Mark Robinson, Davis McCarthy, Gordon Smyth	Empirical Analysis of Digital Gene Expression Data in R	24
DOSE	Guangchuang Yu	Disease Ontology Semantic and Enrichment analysis	44
fgsea	Alexey Sergushichev	Fast Gene Set Enrichment Analysis	45
clusterProfiler	Guangchuang Yu	statistical analysis and visualization of functional profiles for genes and gene clusters	46
GSEABase	Bioconductor Package Maintainer	Gene set enrichment data structures and methods	51

MetaCore™



An Overview of Biological Significance Analysis ("Alternatives to IPA")

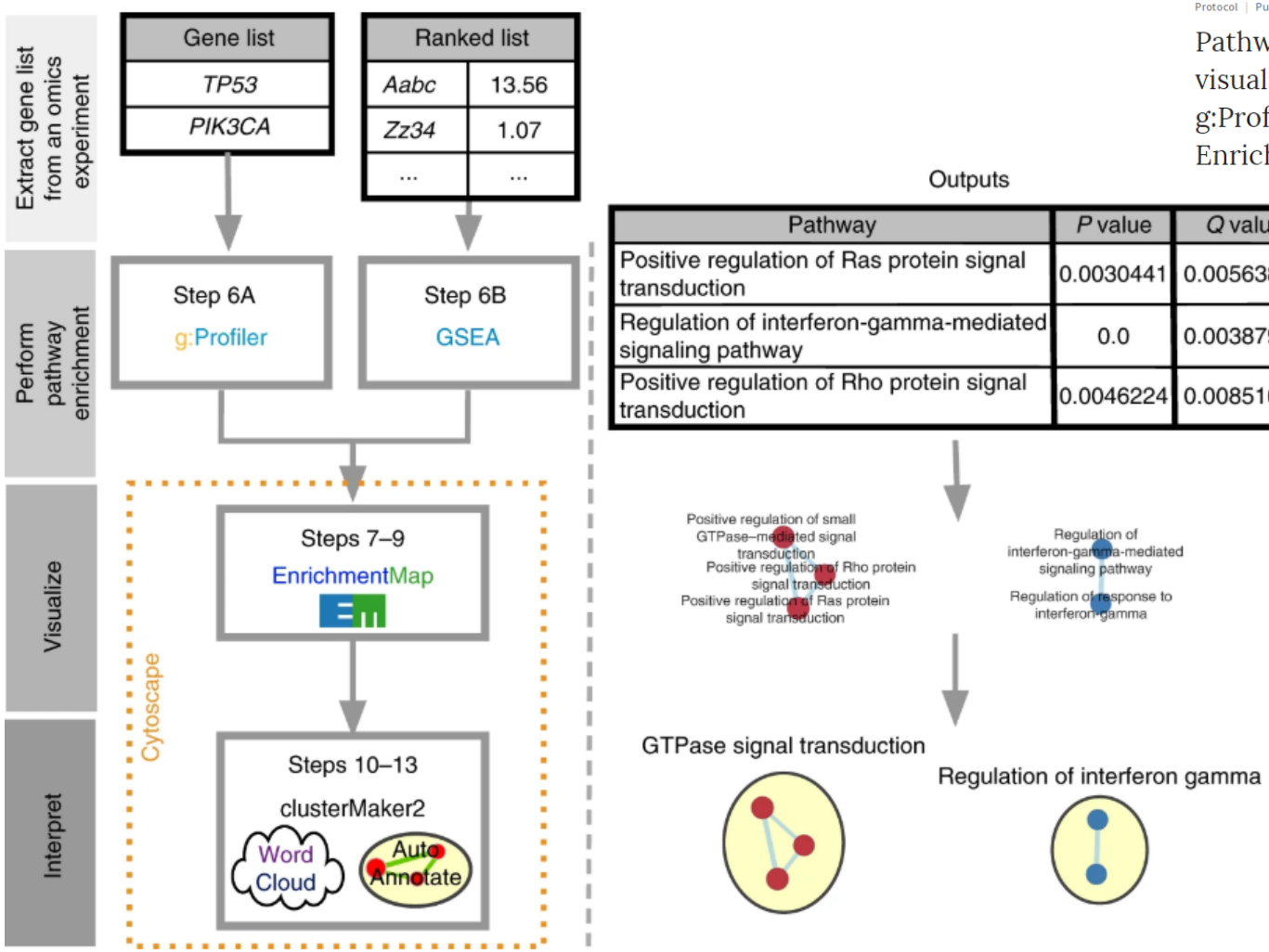
So many, that selecting one is hard

Name	Availability	Reference
ORA tools		
Onto-Express	Web (http://vortex.cs.wayne.edu)	[4,5]
GenMAPP	Standalone (http://www.genmapp.org)	[11,71]
GoMiner	Standalone, Web (http://discover.nci.nih.gov/gominer)	[72,73]
FatiGO	Web (http://babelomics.bioinfo.cipf.es)	[74]
GOstat	Web (http://gostat.wehi.edu.au)	[7]
FuncAssociate	Web (http://llama.mshri.on.ca/funcassociate/)	[6]
GOToolBox	Web (http://genome.crg.es/GOToolBox/)	[10]
GeneMerge	Standalone, Web (http://genemerge.cbcb.umd.edu/)	[9]
GOEAST	Web (http://omicslab.genetics.ac.cn/GOEAST/)	[75]
ClueGO	Standalone (http://www.ici.upmc.fr/cluego/)	[76]
FunSpec	Web (http://funspec.med.utoronto.ca/)	[77]
GARBAN	Web	[78]
GO:TermFinder	Standalone (http://search.cpan.org/dist/GO-TermFinder/)	[8]
WebGestalt	Web (http://bioinfo.vanderbilt.edu/webgestalt/)	[79]
agriGO	Web (http://bioinfo.cau.edu.cn/agriGO/)	[80]
GOFFA	Standalone, Web (http://edkb.fda.gov/webstart/arraytrack/)	[81]
WEGO	Web (http://wego.genomics.org.cn/cgi-bin/wego/index.pl)	[82]
FCS tools		
GSEA	Standalone (http://www.broadinstitute.org/gsea/)	[21,29]
sigPathway	Standalone (BioConductor)	[22]
Category	Standalone (BioConductor)	[24]
SAFE	Standalone (BioConductor)	[30]
GlobalTest	Standalone (BioConductor)	[15]
PCOT2	Standalone (BioConductor)	[17]
SAM-GS	Standalone (http://www.ualberta.ca/~yyasui/software.html)	[83]
Catmap	Standalone (http://bioinfo.thep.lu.se/catmap.html)	[84]
T-profiler	Web (http://www.t-profiler.org)	[85]
FunCluster	Standalone (http://corneliu.henegar.info/FunCluster.htm)	[86]
GeneTrail	Web (http://genetrail.bioinf.uni-sb.de)	[87]
GAzer	Web	[88]
PT-based tools		
ScorePAGE	No implementation available	[37]
Pathway-Express	Web (http://vortex.cs.wayne.edu)	[38,39]
SPIA	Standalone (BioConductor)	[40]
NetGSA	No implementation available	[43]

doi:10.1371/journal.pcbi.1002375.t001

Khatri, Purvesh & Sirota, Marina & Butte, Atul. (2012). Ten Years of Pathway Analysis. PLoS computational biology.

We introduce (a good) one ...



nature protocols

Protocol | Published: 21 January 2019

Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap

An Overview of Biological Significance Analysis ("Alternatives to IPA")

Our plans for today

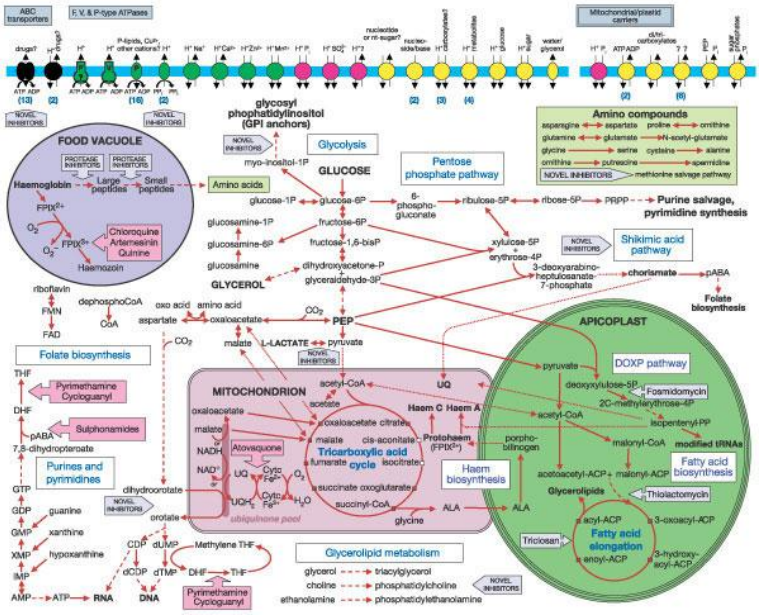
- General goal
  - Have an overview of how to do pathway analysis, from a list of genes to a network visualization that helps/guides the biological interpretation.
- Work plan
  - Background: Gene lists, Gene Sets and Databases
  - Methods and tools for enrichment analysis
  - Visualization and grouping of enrichment results
  - All together: Applying the protocol
- More information, including these slides at *http://uebvhir.github.io/Pathway_Analysis-Guidelines*

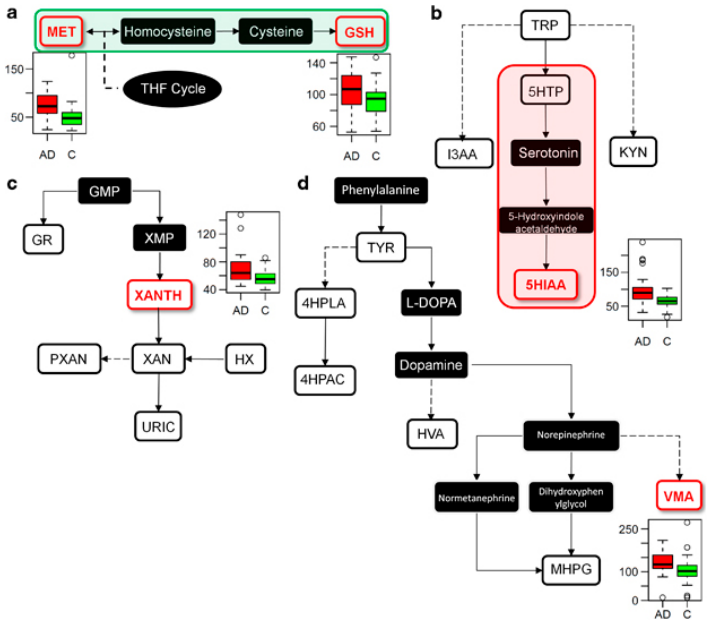
Introduction & Background

Health, disease and pathways

Metabolism is a complex network of chemical reactions within the confines of a cell that can be analyzed in self-contained parts called ***pathways***

One can generally assume that “normal” metabolism is what happens in healthy state or, reciprocally, that disease can *be associated with some type of alteration in meta*bolism.





Pathways altered in ALZHEIMER disease

Characterization of disease can be attempted by studying how this affects or disrupts pathways

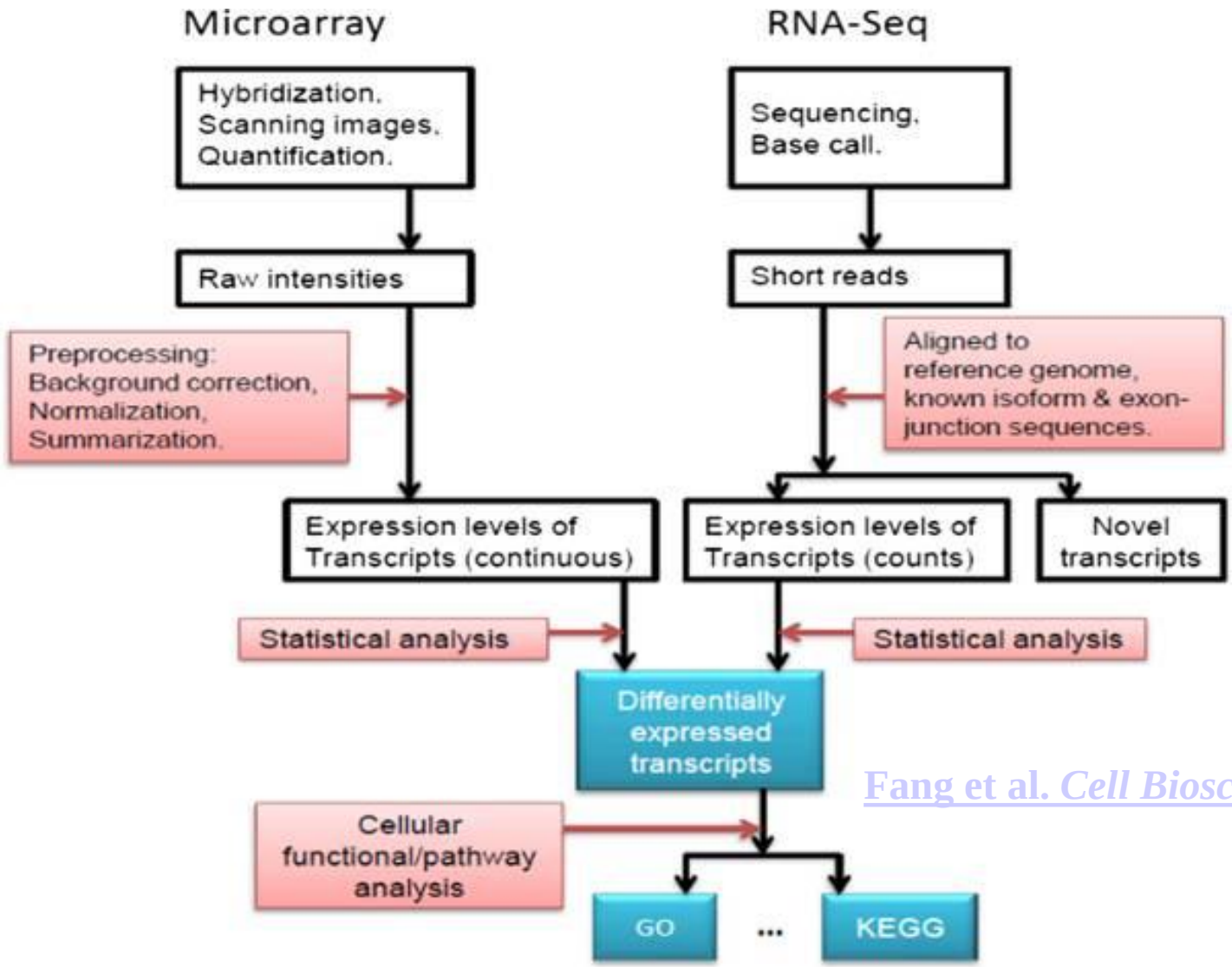
That’s what Pathway Analysis is about (more or less)

Pathway Analysis

- The term Pathway Analysis denotes *any analytic technique that benefits from biological pathway or molecular network information to gain insight into a biological system*. (Creixell et alt., Nature Methods 2015 (12 (7))
- To be more specific, Pathway Analysis methods rely on high throughput information provided by omics technologies to:
  - Contextualize findings to help understand the mechanism of disease
  - Identify genes/proteins associated with the aetiology of a disease
  - Predict drug targets
  - Understand how to therapeutically intervene in disease processes
  - Conduct target literature searches
  - Integrate diverse biological information

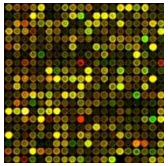
The beginning: *Gene Lists*

The life-cycle of an omics-based study



Fang et al. *Cell Biosci. 2012; 2: 26.*

The (in)famous "*where to now?*" question



- You obtained a list of features. What's next?
  - Select some genes for validation?
  - Follow up experiments on some genes/proteins/…?
  - Publish a huge table with all results?
  - Try to learn about **all** features in the list?

GNAQ
GNAS
DGKZ
GUCY1A3
PDE4B
PDE4D
ATP2A2
ATP2A3
NOS1
CNN1
GSTO1
NOS3
CNN2
MYLK2
CALD1
ACTA1
MYL2

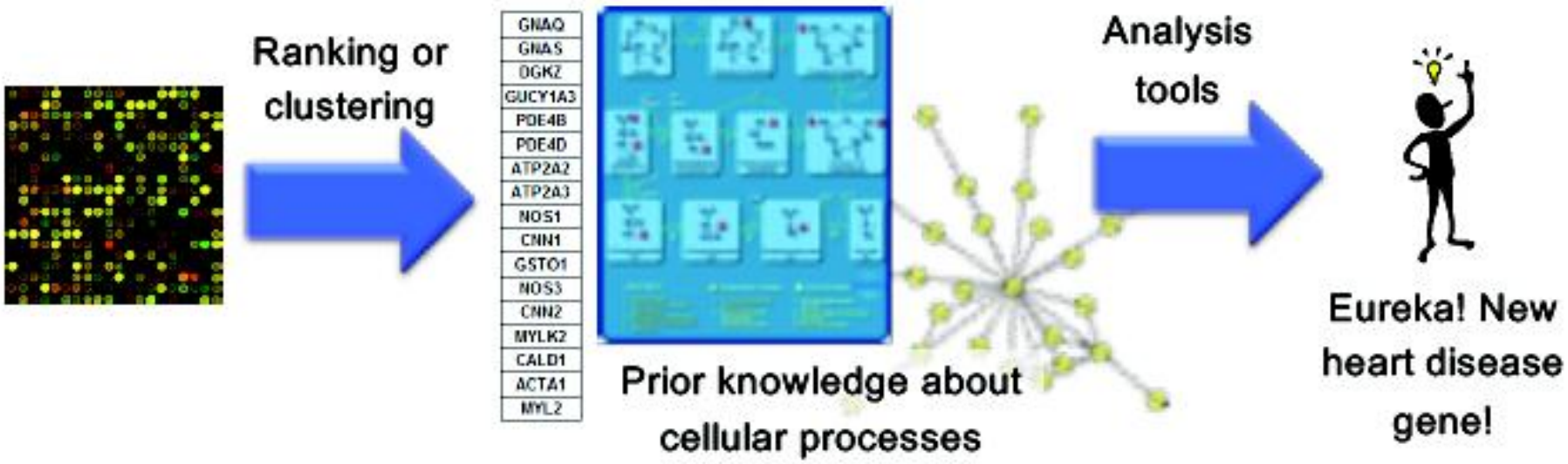
?

From gene lists to *Pathway Analysis*

- Gene lists are made of individual genes
  - Information about each gene can be extracted from databases.
  - Generically described as ***Gene Annotation***
- Besides, we may obtain information from the analysis of *gene sets*
  - Genes don't act individually, rather in groups
    More ***realistic*** approach
  - There are less gene sets than individual genes
    Relatively ***simpler*** to manage.
  - Generically described as ***Pathway Analysis***

Pathway Analysis Wishlist

- Tell me what's interesting about these genes
  - Are they enriched in known pathways, complexes, functions



Example 1

- Genes with frequent somatic SNVs identified in TCGA exome sequencing data of 3,200 tumors of 12 types
- 127 cancer driver genes displaying higher than expected mutation frequencies were detected using the MuSiC software.
- Genes are ranked in decreasing order of significance and mutation frequency

TP53
PIK3CA
PTEN
APC
VHL
KRAS
MLL3
MLL2
ARID1A
PBRM1
NAV3
EGFR
NF1
PIK3R1
CDKN2A
GATA3
RB1
NOTCH1
FBXW7
CTNNB1
DNMT3A
MAP3K1
FLT3
MALAT1
TSHZ3

Example 2

- Second example is a ranked list of genes obtained from TCGA ovarian cancer dataset.
- Two subgroups - immunoreactive and mesenchymal- were compared.
- The list contains **all genes, not only differentially expressed**, ranked by the value of statistic.

rank	GeneName	test statistic
1	IGDCC3	35.5553322839225
2	ANTXR1	35.3770766531836
3	AEBP1	33.0690543534961
4	FBN1	32.1199562790897
5	ANGPTL2	31.8605806216522
6	COL16A1	31.7641267462069
7	BGN	31.533826423921
…	…	…
15201	IRF1	-14.7629673442493
15202	CXCL10	-14.9827363665643
15203	TAP2	-15.1488606179238
15204	UBE2L6	-15.7162058907796
15205	KIAA0319	-15.7796986548781
15206	PSMB8	-15.7846188665582
15207	PSME1	-16.4510045533584
15208	CSAG3	-16.8014265945244
15209	OVGP1	-17.6903158148446
15210	GBP4	-17.9447602030134
15211	TAP1	-18.0549262210415
15212	PSME2	-18.3639448844986
15213	PSMB9	-18.6614452029879

Gene Lists and Annotations

Gene and Protein Identifiers

- Identifiers (IDs) are ideally unique, stable names or numbers that help track database records
  - E.g. Social Insurance Number, Entrez Gene ID 41232
- But, information on features is stored in many databases.
  - Genes have many IDs
- Records for: Gene, DNA, RNA, Protein
  - Important to recognize the correct record type
  - E.g. Entrez Gene records don't store sequence. They link to DNA regions, RNA transcripts and proteins e.g. in RefSeq, which stores sequence.

TP53
PIK3CA
PTEN
APC
VHL
KRAS
MLL3
MLL2
ARID1A
PBRM1
NAV3
EGFR
NF1
PIK3R1
CDKN2A
GATA3
RB1
NOTCH1
FBXW7
CTNNB1
DNMT3A
MAP3K1
FLT3
MALAT1
TSHZ3

Common Identifiers

Gene
Ensembl ENSG00000139618
Entrez Gene 675
Unigene Hs.34012

RNA transcript
GenBank BC026160.1
RefSeq NM_000059
Ensembl ENST00000380152

Protein
Ensembl ENSP00000369497
RefSeq NP_000050.2
UniProt BRCA2_HUMAN or A1YBP1_HUMAN
IPI IPI00412408.1
EMBL AF309413
PDB 1MIU

Species-specific
HUGO HGNC BRCA2
MGI MGI:109337
RGD 2219
ZFIN ZDB-GENE-060510-3
FlyBase CG9097
WormBase WBGene00002299 or ZK1067.1
SGD S000002187 or YDL029W
Annotations
InterPro IPR015252
OMIM 600185
Pfam PF09104
Gene Ontology GO:0000724
SNPs rs28897757
Experimental Platform
Affymetrix 208368_3p_s_at
Agilent A_23_P99452
CodeLink GE60169
Illumina GI_4502450-S

Red = Recommended

Identifier Mapping

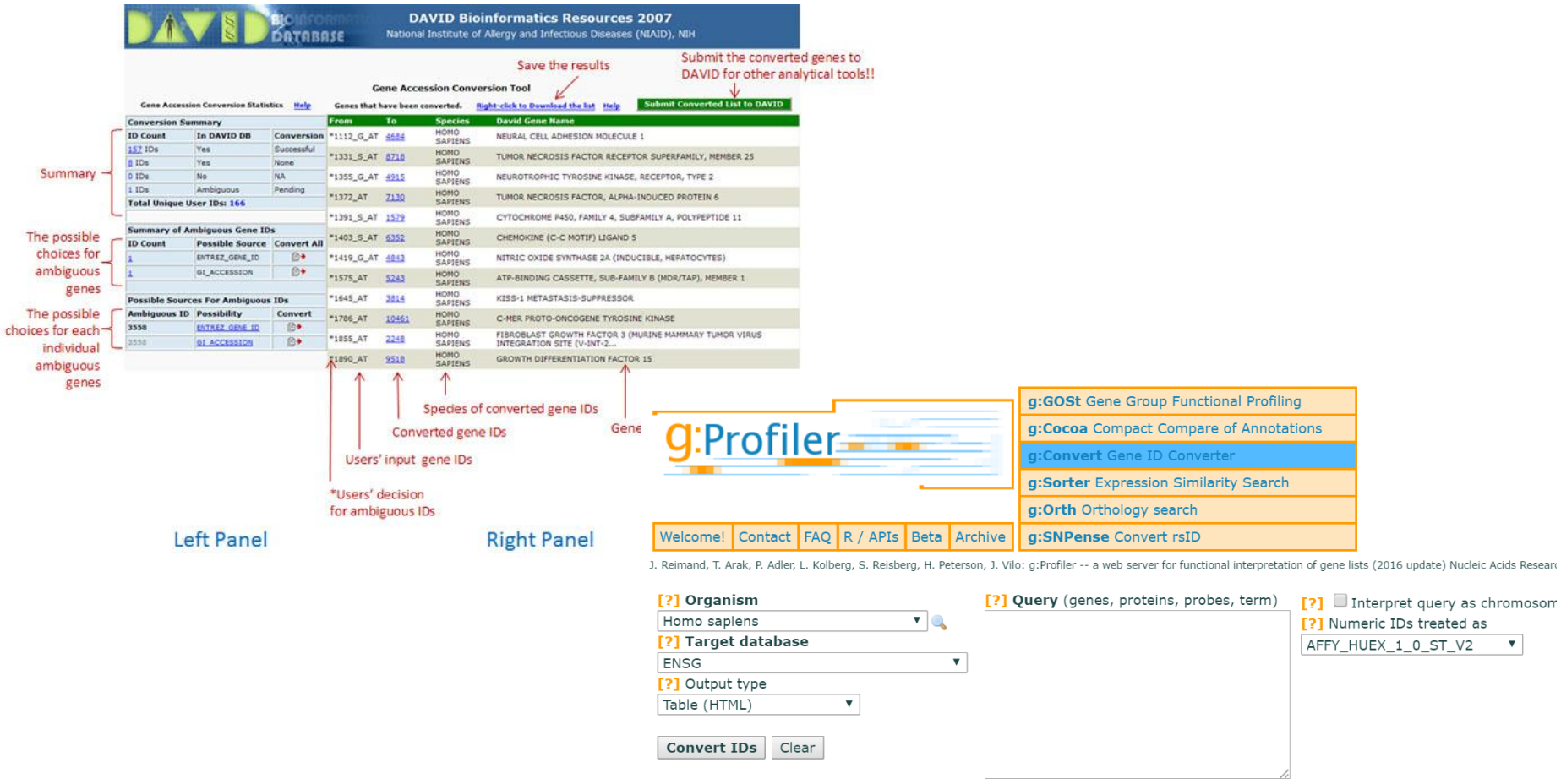
- There are many IDs!
  - Software tools recognize only a handful
  - May need to map from your gene list IDs to standard IDs
- Four main uses
  - Searching for a favorite gene name
  - Link to related resources
  - Identifier translation
    - E.g. Proteins to genes, Affy ID to Entrez Gene
  - Merging data from different sources
    - Find equivalent records

ID Challenges

- Avoid errors: map IDs correctly
  - Beware of 1-to-many mappings
- Gene name ambiguity – not a good ID
  - e.g. FLJ92943, LFS1, TRP53, p53
  - Better to use the standard gene symbol: TP53
- Excel error-introduction
  - OCT4 is changed to October-4 (paste as text)
- Problems reaching 100% coverage
  - E.g. due to version issues
  - Use multiple sources to increase coverage

Zeeberg BR et al. *Mistaken identifiers: gene name errors can be introduced inadvertently when using Excel in bioinformatics*
BMC Bioinformatics. 2004 Jun 23;5:80

Use ID converters to prepare list



Example 1: Gene ID conversion with g:Profiler

Recommendations

- For proteins and genes
  - (doesn't consider splice forms)
  - Map everything to Entrez Gene IDs or Official Gene Symbols using an appropriate tool, such as gProfiler, DAVID or Biomart.
- If 100% coverage desired, manually curate missing mappings using multiple resources
- Be careful of Excel auto conversions – especially when pasting large gene lists!
  - Remember to format cells as 'text' before pasting

Pathway and Gene Sets databases

Where is pathway information? (1)

- Most common sources*
  - Gene Ontology: Biological process,
  - Pathway databases:
    - Reactome : http://reactome.org
    - http://www.pathguide.org
    - MSigDB: http://www.broadinstitute.org/gsea/msigdb/
    - http://www.pathwaycommons.org/

*Comparison of human cell signaling pathway databases—evolution, drawbacks and challenges

Where is pathway information? (2)

- Other annotations
  - Gene Ontology molecular function, cell location
  - Chromosome position
  - Disease association
  - DNA properties (TF binding sites, gene structure (intron/exon), SNPs, …)
  - Transcript properties (Splicing, 3' UTR, microRNA binding sites, …)
  - Protein properties (Domains, 2ry and 3ry structure, PTM sites)
  - Interactions with other genes

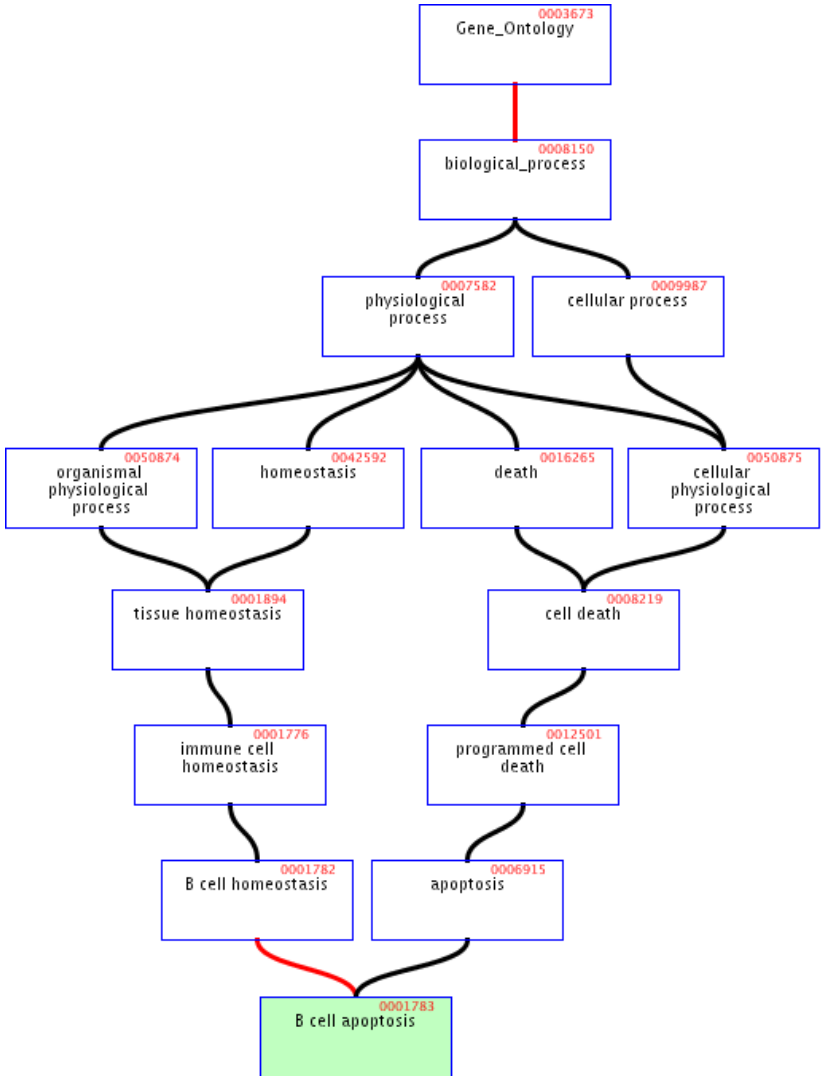
*Comparison of human cell signaling pathway databases—evolution, drawbacks and challenges

What is the Gene Ontology (GO)?

- Set of biological phrases (terms) which are applied to genes:
  - protein kinase, apoptosis, membrane
- An ontology is not a dictionary
  - Dictionary: A collection of term definitions,
    - Alphabetic organization
  - Ontology: A formal system for describing knowledge
    - Hierarchical organization
- http://geneontology.org/

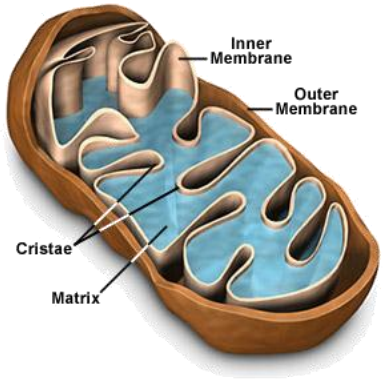
GO Structure

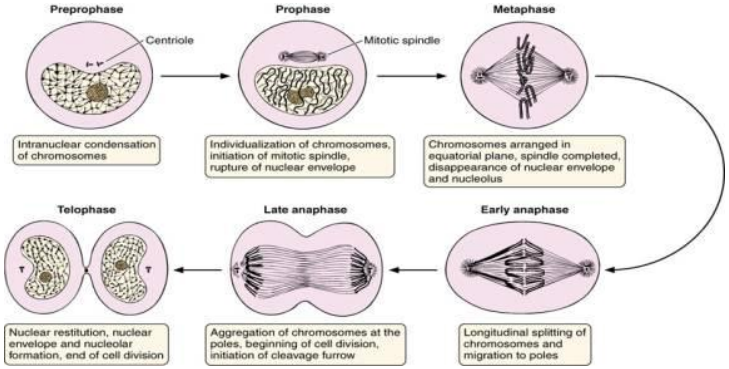
- Terms are related within a hierarchy
  - is-a
  - part-of
- Describes multiple levels of detail of gene function
- Terms can have more than one parent or child



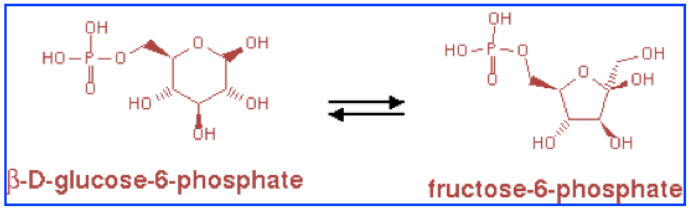
What is covered by the GO?

- GO terms divided into three aspects:
  - cellular component
  - molecular function
  - biological process





Cell division



glucose-6-phosphate isomerase activity

Part 1/2: Terms

- Where do GO terms come from?
  - GO terms are added by editors at EBI and gene annotation database groups
  - Terms added by request
  - Experts help with major development

	Jun 2012	Apr 2015	increase
Biological process	23,074	28,158	22%
Molecular function	9,392	10,835	15%
Cellular component	2,994	3,903	30%
total	**37,104**	**42,896**	**16%**

Part 2/2: Annotations

- Genes are linked, or associated, with GO terms by trained curators at genome databases
  - Known as ‘gene associations’ or GO annotations
  - Multiple annotations per gene
- Some GO annotations created automatically (without human review)

Annotation Sources

- Manual annotation
  - Curated by scientists
    - High quality
    - Small number (time-consuming to create)
  - Reviewed computational analysis
- Electronic annotation
  - Annotation derived without human validation
    - Computational predictions (accuracy varies)
    - Lower ‘quality’ than manual codes
- Key point: be aware of annotation origin

Evidence Types

- Experimental Evidence Codes
  - EXP: Inferred from Experiment
  - IDA: Inferred from Direct Assay
  - IPI: Inferred from Physical Interaction
  - IMP: Inferred from Mutant Phenotype
  - IGI: Inferred from Genetic Interaction
  - IEP: Inferred from Expression Pattern

- Computational Analysis Evidence Codes
  - ISS: Inferred from Sequence or Structural Similarity
  - ISO: Inferred from Sequence Orthology
  - ISA: Inferred from Sequence Alignment
  - ISM: Inferred from Sequence Model
  - IGC: Inferred from Genomic Context
  - RCA: inferred from Reviewed Computational Analysis

- Author Statement Evidence Codes
  - TAS: Traceable Author Statement
  - NAS: Non-traceable Author Statement
- Curator Statement Evidence Codes
  - IC: Inferred by Curator
  - ND: No biological Data available

- IEA: Inferred from electronic annotation

http://www.geneontology.org/GO.evidence.shtml

Contributing Databases

- Berkeley *Drosophila* Genome Project (BDGP)
- dictyBase (*Dictyostelium discoideum)*
- FlyBase (*Drosophila melanogaster)*
- GeneDB (*Schizosaccharomyces pombe*, *Plasmodium falciparum*, *Leishmania major* and *Trypanosoma brucei)*
- UniProt Knowledgebase (Swiss-Prot/TrEMBL/PIR-PSD) and InterPro databases
- Gramene (grains, including rice, *Oryza*)
- Mouse Genome Database (MGD) and Gene Expression Database (GXD) (*Mus musculus)*
- Rat Genome Database (RGD) (*Rattus norvegicus)*
- Reactome
- *Saccharomyces* Genome Database (SGD) (*Saccharomyces cerevisiae)*
- The *Arabidopsis* Information Resource (TAIR) (*Arabidopsis thaliana)*
- The Institute for Genomic Research (TIGR): databases on several bacterial species
- WormBase (*Caenorhabditis elegans)*
- Zebrafish Information Network (ZFIN): (*Danio rerio)*

Pathway Analysis

Overrepresentation Analysis

Gene Set Enrichment Analysis

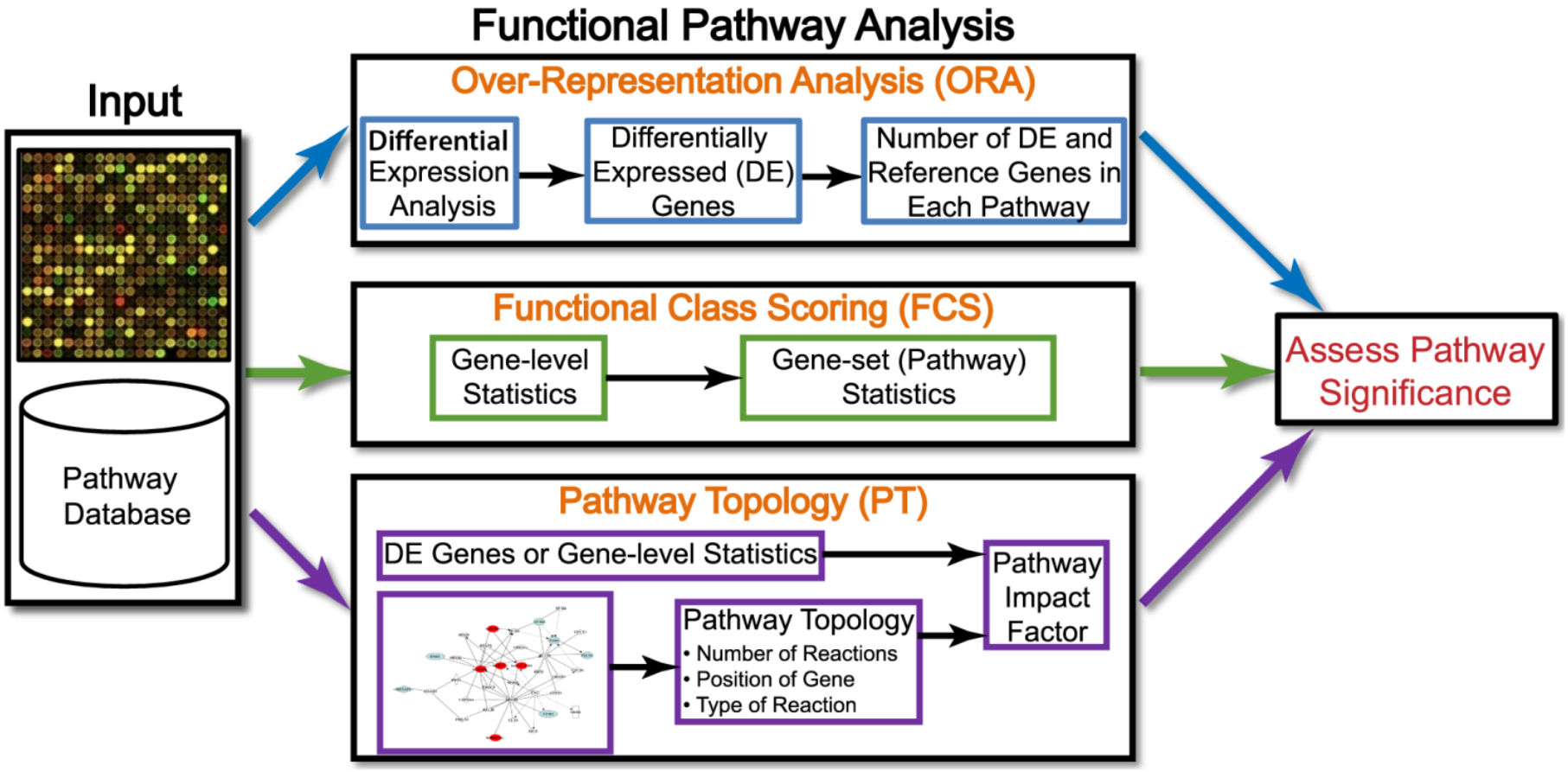
Pathway Analysis

- "*Any type of analysis that involves pathway or information*"
  - Most popular type is ***enrichment analysis***, but many others exist.
- Intended to gain insight into 'omics' data. E.g:
  - Identifying a master regulator gene,
  - Finding drug targets,
  - Characterizing pathways active in a sample.

Benefits of Pathway Analysis

- Relatively easy to interpret
  - Familiar concepts e.g. cell cycle
- Identifies possible causal mechanisms
- Predicts new roles for genes
- Improves statistical power
  - Fewer tests, aggregates data from multiple genes into one pathway
- More reproducible
  - E.g. gene expression signatures
- Facilitates integration of multiple data types

Types of Pathway Analysis



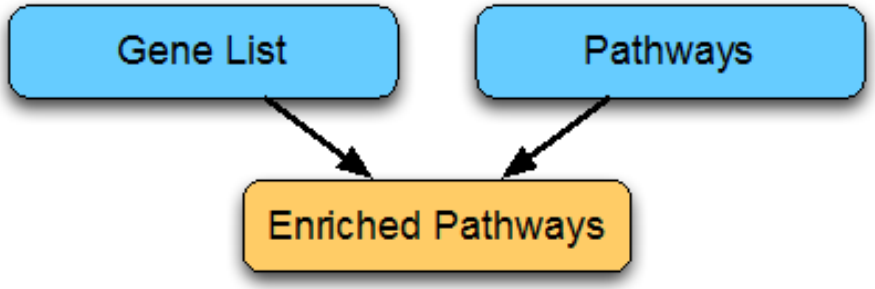
Khatri et alt. *10 years of Pathway Analysis*

Analysis of thresholded lists with *Enrichment Analysis*

(also called Overrepresentation A.)

Over-representation analysis

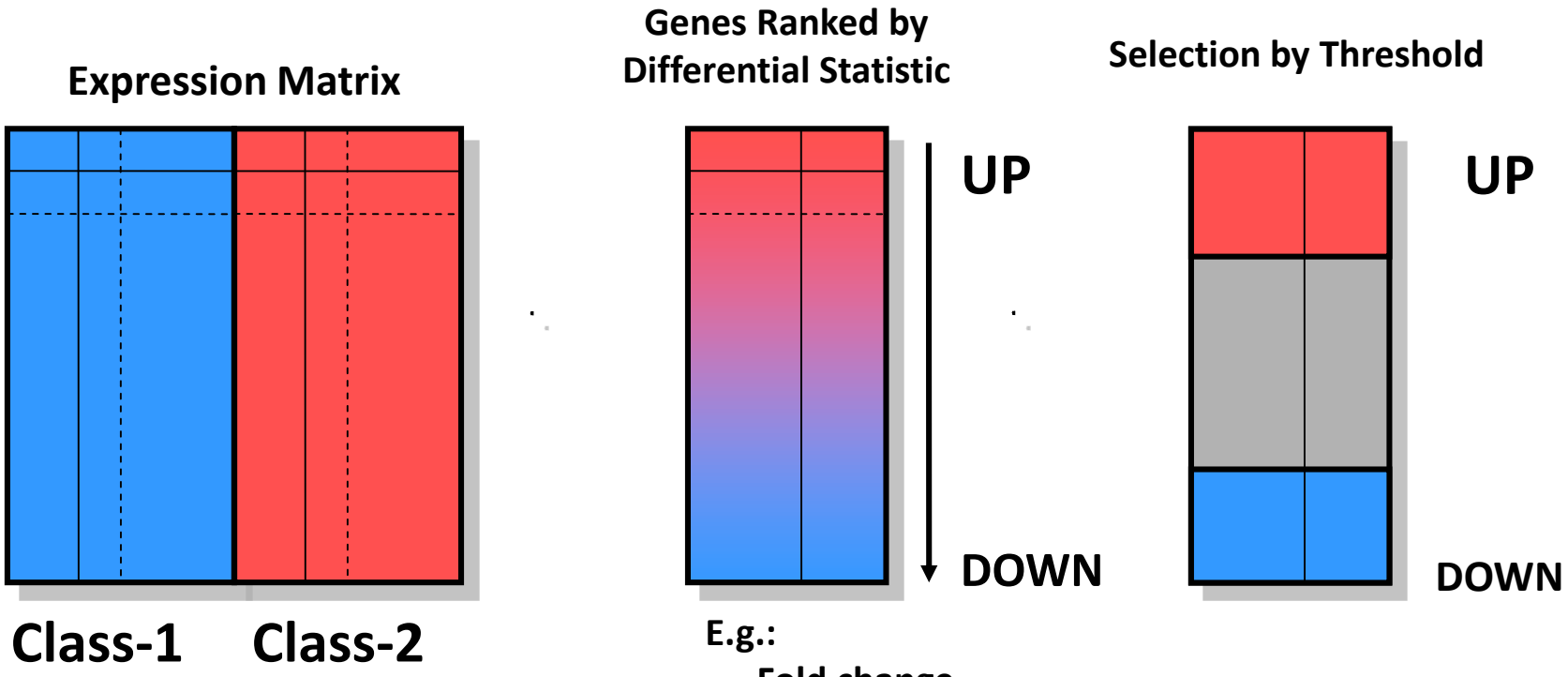
- Combines
  - Gene (feature) lists ← (Gen)omic experiment
  - Pathways and other gene annotations
    - Gene Ontology
    - Reactome
    - Pathway commons



Over-representation analysis

- Given:
  1. Gene list: e.g. RRP6, MRD1, RRP7, RRP43, RRP42 (yeast)
  2. Gene sets or annotations: e.g. Gene ontology, transcription factor binding sites in promoter
- Question: *Are any of the gene annotations surprisingly enriched in the gene list?*
- Details:
  1. Where do the gene lists come from?
  2. How to assess “surprisingly” (statistics)
  3. How to adjust for test multiplicity?

Obtaining the gene lists

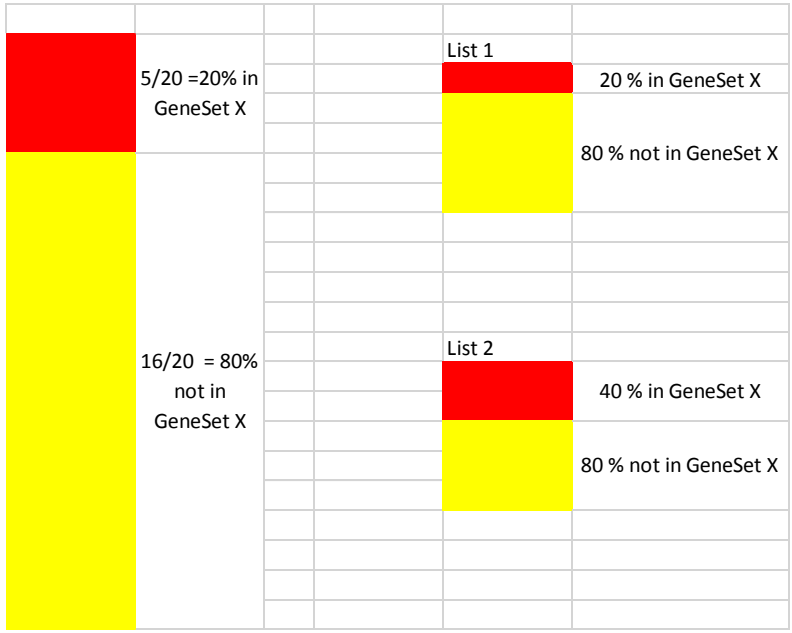


E.g.:

- Fold change
- Log (ratio)
- t-test
- Significance analysis of microarrays

Assessing “surprisingly”

- Given a gene list, “gl”, and a gene set, “GS”, check:
- Is the % of genes in “gl” annotated in “GS” the same as the % of genes globally annotated in “GS”?
  - If both percentages are similar → *No Enrichment*
  - If the % of genes annotated in “GS” is greater in “gl” than in the rest of genes → *“gl” is enriched in “GS”*



Examples

	Differentially expressed (gl_1)	Not differentially expressed	TOTAL
In Gene Set (GS1)	10	30	40
Not In Gene Set	390	3570	3960
TOTAL	400	3600	4000
% of gl1 in GS1	**10/400=0.025**	**30/3600=0.00833**	

0.025 >> 0.00833 → "gl_1" is enriched in "GS_1"

	Differentially expressed (gl_2)	Not differentially expressed	TOTAL
In Gene Set (GS2)	10	30	40
Not In Gene Set	390	1220	1610
TOTAL	400	1500	1650
% of gl_2 in GS_2	**10/400=0.025**	**30/1500=0.2**	

0.025 ≈ 0.02 → Can't say that "gl_2" is enriched in "GS_2"

Assessing significance: Fisher test

- The examples shows two cases
  - One where percentages are quite different
  - Another where percentages are similar
- How can we set a threshold to decide that the difference is “big enough” to call it “Enriched”
  - Use Fisher Test or, equivalently,
  - a test to compare proportions or
  - a hypergeometric test.

Assessing significance: Fisher test (1)

```
> GOnnnnCounts<- matrix(c(10, 30, 390, 3570),
+         nrow = 2, byrow=TRUE,
+         dimnames = list(GeneSet = c("In Gene Set", "Not in Gene Set"),
+                         Test =c("Differentially expressed", "Not Dif. Expr.")))
> GOnnnnCounts
                Test
GeneSet          Differentially expressed Not Dif. Expr.
  In Gene Set                          10             30
  Not in Gene Set                     390           3570
> fisher.test(GOnnnnCounts, alternative = "greater")

        Fisher's Exact Test for Count Data

data:  GOnnnnCounts
p-value = 0.004836
alternative hypothesis: true odds ratio is greater than 1
95 percent confidence interval:
 1.508343      Inf
sample estimates:
odds ratio
  3.049831
```

P-value small, odds-ratio high → List is *surprisingly* enriched in Gene Set

Assessing significance: Fisher test (2)

```
> GOnnnnCounts<-matrix(c(10,30,390,1220), nrow=2, byrow=TRUE,
+                      dimnames=list(
+                        GeneSet=c("In Gene Set", "Not in Gene Set"),
+                        Test=c("Diff.expressed", "Not diff.expr.")))
> GOnnnnCounts
                Test
GeneSet          Diff.expressed Not diff.expr.
  In Gene Set                10             30
  Not in Gene Set           390           1220
> fisher.test(GOnnnnCounts, alternative="greater")

        Fisher's Exact Test for Count Data

data:  GOnnnnCounts
p-value = 0.517
alternative hypothesis: true odds ratio is greater than 1
95 percent confidence interval:
 0.5149828       Inf
sample estimates:
odds ratio
  1.042711
```

P-value not small, odds-ratio approx. 1 → List is not *surprisingly* enriched in Gene Set

Recipe for gene list enrichment test

- **Step 1:** Define gene list (e.g. thresholding analyzed list) and background list,
- **Step 2:** Select gene sets to test for enrichment,
- **Step 3:** Run enrichment tests and correct for multiple testing, if necessary,
- **Step 4:** Interpret your enrichments
- **Step 5:** Publish! ;)

Possible problems with gene list test

- No “natural” value for the threshold
- Possible loss of statistical power due to thresholding
  - No resolution between significant signals with different strengths
  - Weak signals neglected
- Different results at different threshold settings
- Based on the wrong assumption of independent gene (or gene group) sampling, which increases false positive predictions

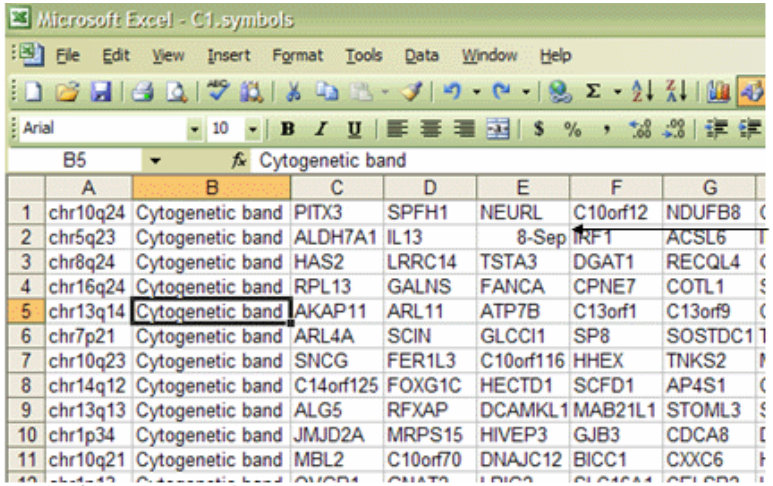
Analysis of ranked gene lists with *Gene Set Enrichment Analysis* (also called Functional Class Scoring)

Gene Sets

- A gene set
  - a group of genes with related functions.
  - sets of genes or pathways, for their association with a phenotype.
  - Examples: metabolic pathway, protein complex, or GO (gene ontology) category.
- Identified from a prior biological knowledge.
- May better reflect the true underlying biology.
- May be more appropriate units for analysis.

Gene Sets

Each row represents one gene set



If editing in excel, watch out for its tendency to auto-format gene sets (SEP8 becomes 8-Sep)

First column are gene set names. Duplicates are not allowed

Second column contains a brief description. Its optional – you can fill in a dummy field (e.g. "na")

Unequal lengths (i.e # of genes) *is* allowed

MSigDB Collection	Subcollection	No. Gene Sets
C1: positional gene sets		326
C2: curated gene sets	CGP: chemical and genetic perturbations	3402
	CP: Canonical pathways KEGG/Biocarta/REACTOME	1320
C3: motif gene sets	MIR: microRNA targets	221
	TFT: transcription factor targets	615
C4: computational gene sets	CGN: cancer gene neighborhoods	427
	CM: cancer modules	431
C5: GO gene sets	BP: GO biological process	825
	CC: GO cellular component	233
	MF: GO molecular function	396
C6: oncogenic signatures		189
C7: immunologic signatures		1910
	Total	10295

Gene Set (Enrichment) Analysis

- Mootha (2003) as an alternative to ORA.
- It aims to identify gene sets with *subtle but coordinated expression changes* that cannot be detected by ORA methods.
  - Weak changes in individual genes gathered to large gene sets can show a significant pattern.
- Results not affected by arbitrarily chosen cutoffs.
- It does not provide information as detailed as ORA

The GSEA method

- Original GSEA method is based on comparing, for each gene group, the distribution of the test statistic within the group with the overall distribution of those statistics, i.e. the calculated for all genes.
- To do this, test statistics are ranked (from biggest to smallest) and **for each gene set** a running sum is computed such that
  - If a gene is in the gene set add a certain quantity (moderate)
  - If a gene is not in the gene set, substract a (small) quantity
- The distribution of the running sum is compared with that of the random walk using a Kolmogorov-Smirnov test (K-S test) statistic
- P-values are computed based on a randomization.

Calculating enrichment score (ES)

Create a running sum statistic based on the following
If gene p is not in set S, then add

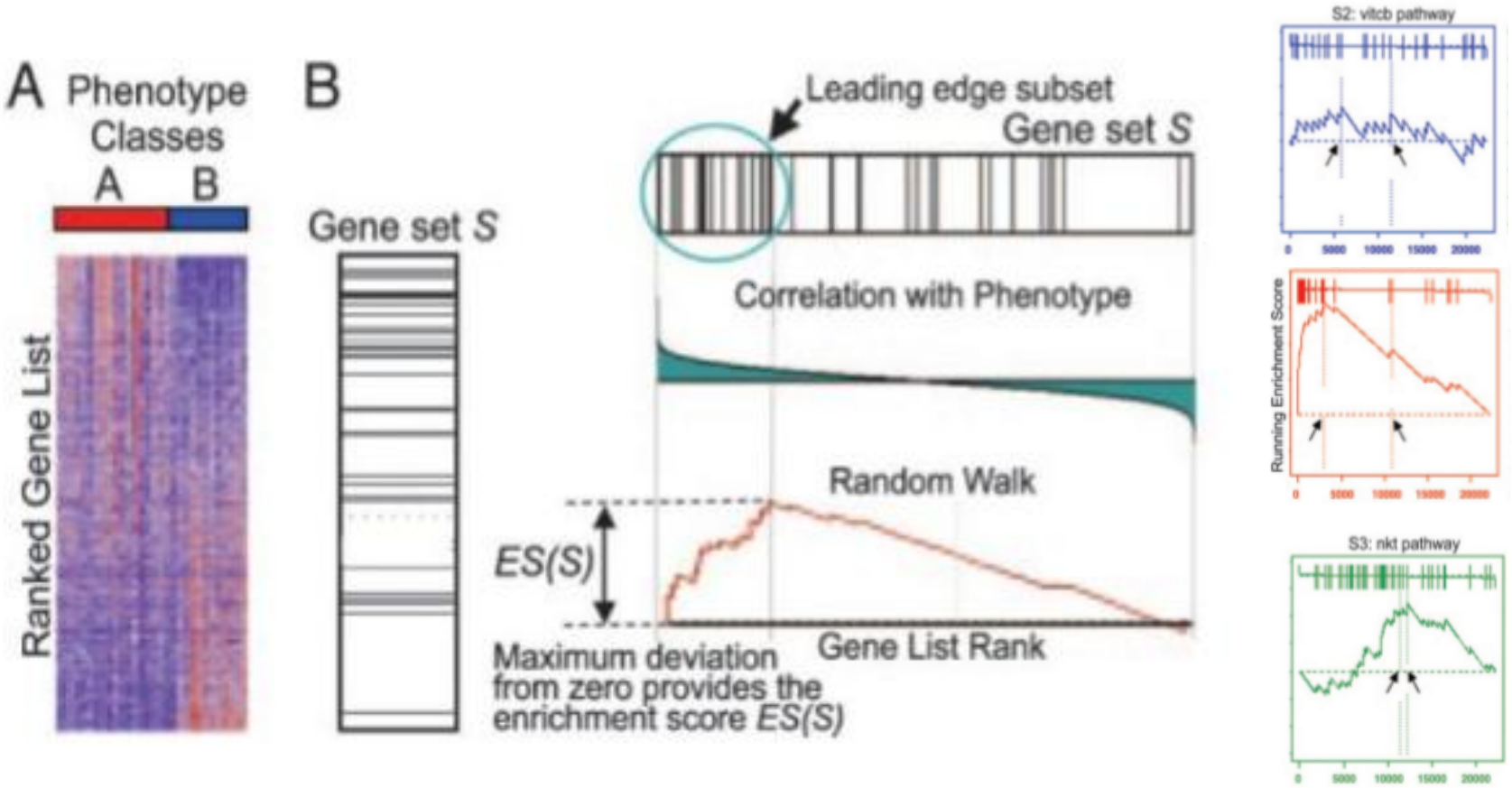
$$X_i = -\sqrt{\frac{N_S}{N - N_S}}$$

If gene p is in set S, then add

$$X_i = \sqrt{\frac{N - N_s}{N_S}}$$

This creates a running sum
The maximum sum over the whole list L is the Enrichment Score
MES

The GSEA method



Recipe for ranked list enrichment test

- **Step 1:** Rank ALL your genes,
- **Step 2:** Select gene sets to test for enrichment,
- **Step 3:** Run enrichment tests and correct for multiple testing, if necessary,
- **Step 4:** Interpret your enrichments
- **Step 5:** Publish! ;)

GSEA variants

- GSEA is not free from criticisms
  - Use of KS test
  - Null hypothesis is not clear
- Many alternative available
  - Efron's GSA
  - Limma's ROAST
  - Irizarry's simple GSA based on Wilcoxon…

Multiple test adjustments

Why we need to “adjust”

- We use a statistical test to decide if a gene list is “surprisingly” enriched in a Gene Set.
  - We use “surprisingly” instead of “significantly”
- Remember that when doing statistical tests one can be right or wrong differently.
  - Right
    - Rejecting the null hypothesis (H_0) when it is false
    - Not rejecting H_0 when it is true
  - Wrong
    - Rejecting the null hypothesis (H_0) when it is true
    - Not rejecting H_0 when it is false

Errors and Successes in tests: Type I and type II errors

Decision	Actual Situation "Truth" H_0 True	Actual Situation "Truth" H_0 False
Do Not Reject H_0	Correct Decision $1 - \alpha$	Incorrect Decision Type II Error β
Rejct H_0	Incorrect Decision Type I Error α	Correct Decision $1 - \beta$

$$\alpha = P(Type\ I\ Error) \quad \beta = P(Type\ II\ Error)$$

Testing repeatedly

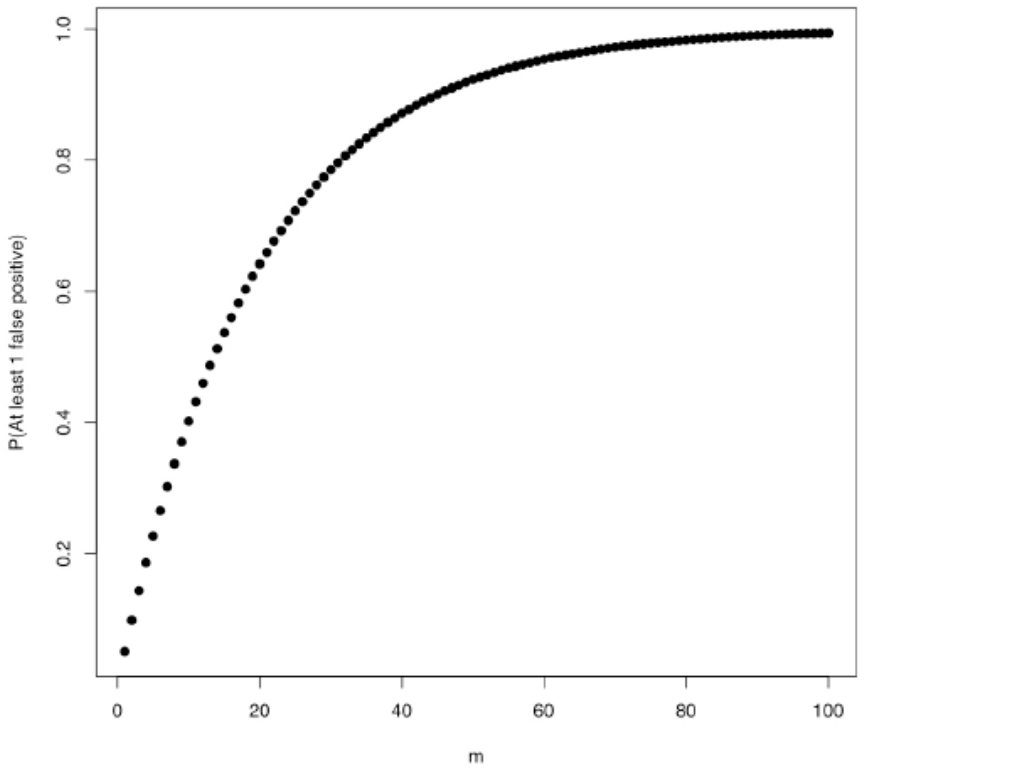
- Omics studies are “high throughput”
  - Selecting genes: One test per each gene
  - Finding enriched gene sets: One test per each gene set
- Doing many tests means facing repeatedly the probability of making one false positive.
  - As the number of tests increases →
  - The chance of observing at least one false positive is going to increase too.

Why multiple testing matters

- The probability of observing one false positive if testing once is:
  - P(Making a type I error) = α
  - P(not making a type I error) = 1- α
- Now imagine we perform *m* tests independently
  - P(not making a type I error in *m* tests) = $(1- \alpha)^m$
  - P(making at least a type I error in *m* tests) = $1-(1- \alpha)^m$
- As m increases the probability of having at least one type error tends to increase

Type I error not useful in multiple testing

Number of tests: m	P(making at least a type I error) = $1-(1- a)^m$
1	0.050000
2	0.097500
3	0.142625
4	0.185494
5	0.226219
6	0.264908
7	0.301663
8	0.336580



How can we deal with this issue?

- Controlling for type I error is not feasible if many tests.
- Idea: Modify α (or alternatively the p-value) so the error probability is ***controlled overall***
- This may mean different things:
  1. The probability of at least one error in *m* tests is $< \alpha$
  2. The expected number of false positives is below a fixed threshold.

…

Controlling the FWER: *Bonferroni*

If *M* = # of annotations tested:

Corrected P-value = *M* x original P-value

Corrected P-value is greater than or equal to the probability that ***one or more of the observed enrichments*** could be due to random draws.
The jargon for this correction is "controlling for the *Family-Wise Error Rate (FWER)"*

Bonferroni correction caveats

- Bonferroni correction is very stringent and can "wash away" real enrichments leading to false negatives,
- Often one is willing to accept a less stringent condition, the "false discovery rate" (FDR), which leads to a gentler correction when there are real enrichments.

False discovery rate (FDR)

- FDR is *the expected **proportion** of "False Positives" that is of the observed enrichments due to random chance.*
- Compare to Bonferroni correction which is a bound on *the probability that **any one** of the observed enrichments could be due to random chance.*
- Typically FDR corrections are calculated using the Benjamini-Hochberg procedure.
- FDR threshold is often called the "q-value"

Benjamini-Hochberg example I

Rank	Category	(Nominal) P-value
1	*Transcriptional*	0.001
2	*regulation*	0.002
3	*Transcription factor*	0.003
4	*Initiation of*	0.0031
5	*transcription*	0.005
…	*Nuclear localization*	…
	Chromatin modification	
52	…	0.97
53	*Cytoplasmic localization*	0.99
	Translation	

Sort P-values of all tests in decreasing order

Benjamini-Hochberg example II

Rank	Category	(Nominal) P-value	Adjusted P-value
1	*Transcriptional regulation*	0.001	0.001 x 53/1 = 0.053
2	*Transcription factor*	0.002	0.002 x 53/2 = 0.053
3	*Initiation of transcription*	0.003	0.003 x 53/3 = 0.053
4	*Nuclear localization*	0.0031	0.0031 x 53/4 = 0.040
5	*Chromatin modification*	0.005	0.005 x 53/5 = 0.053
…	…	…	…
52	*Cytoplasmic localization*	0.97	0.985 x 53/52 = 1.004
53	*Translation*	0.99	0.99 x 53/53 = 0.99

Adjusted P-value is "nominal" P-value times # of tests divided by the rank of the P-value in sorted list

Adjusted P-value = P-value X [# of tests] / Rank

Benjamini-Hochberg example III

Rank	Category	(Nominal) P-value	Adjusted P-value	FDR / Q-value
1	*Transcriptional regulation*	0.001	0.001 x 53/1 = 0.053	0.040
2	*Transcription factor*	0.002	0.002 x 53/2 = 0.053	0.040
3	*Initiation of transcription*	0.003	0.003 x 53/3 = 0.053	0.040
4	*Nuclear localization*	0.0031	0.0031 x 53/4 = 0.040	0.040
5	*Chromatin modification*	0.005	0.005 x 53/5 = 0.053	0.053
…	…	…	…	…
52	*Cytoplasmic localization*	0.97	0.985 x 53/52 = 1.004	0.99
53	*Translation*	0.99	0.99 x 53/53 = 0.99	0.99

Q-value (or FDR) corresponding to a nominal P-value is the smallest adjusted P-value assigned to P-values with the same or larger ranks.

Benjamini-Hochberg example III

P-value threshold for FDR < 0.05

Rank	Category	(Nominal) P-value	Adjusted P-value	FDR / Q-value
1	*Transcriptional regulation*	0.001	0.001 x 53/1 = 0.053	0.040
2	*Transcription factor*	0.002	0.002 x 53/2 = 0.053	0.040
3	*Initiation of transcription*	0.003	0.003 x 53/3 = 0.053	0.040
4	*Nuclear localization*	0.0031	0.0031 x 53/4 = 0.040	0.040
5	*Chromatin modification*	0.005	0.005 x 53/5 = 0.053	0.053
…	…	…	…	…
52	*Cytoplasmic localization*	0.97	0.985 x 53/52 = 1.004	0.99
53	*Translation*	0.99	0.99 x 53/53 = 0.99	0.99

Red: non-significant

Green: significant at FDR < 0.05

P-value threshold is highest ranking P-value for which corresponding Q-value is below desired significance threshold

Reducing adjustment stringency

- The adjustment to the P-value threshold depends on the # of tests that you do,
- So, no matter what, *the more tests you do, the more sensitive the test needs to be*
- Can control the stringency by ***reducing the number of tests***:
  - Don't use all collections of Gene Sets available
  - Restrict testing to the appropriate GO annotations;
  - Filter gene sets by size

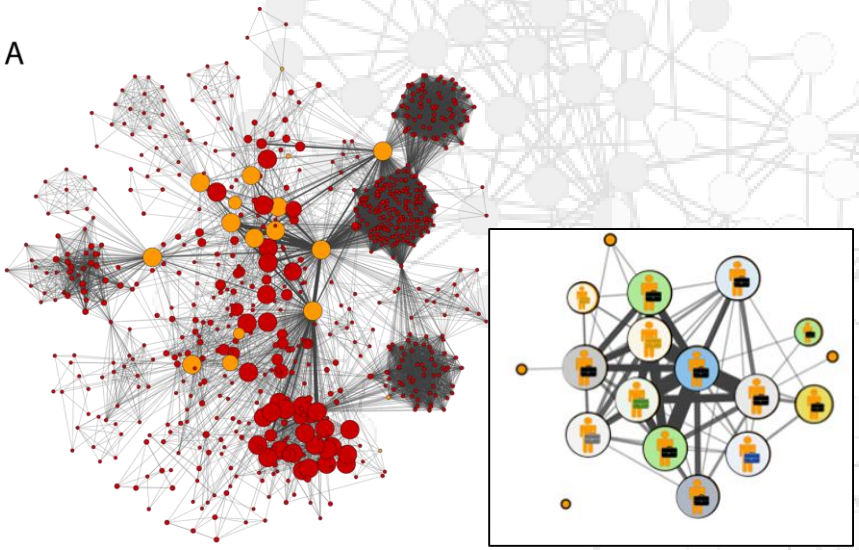
Network Analysis and Network Visualization

There are networks and networks

- Many Pathway Analysis methods make some use of Networks, or even Network Analysis.
- Although network analysis is not a tòpic for this talk it is important to distinguish
  - Network Analysis
    - Where the network structure of the data is exploited to improve understanding of the underlying biological processes.
  - Network Visualization
    - Where relations between the elements of a data set are use to create a network that helps summarizing and visualizing the data and the relations.

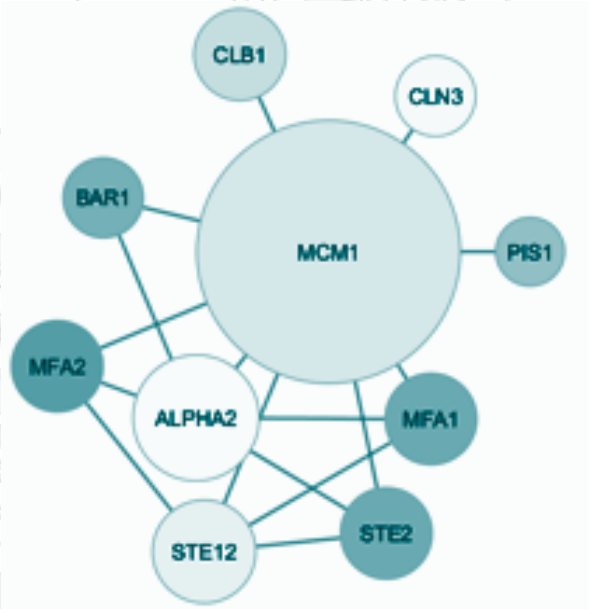
Social vs Biological Data Networks

example of a social network



Cytoscape: SocialNetwork app

example of a biological data network

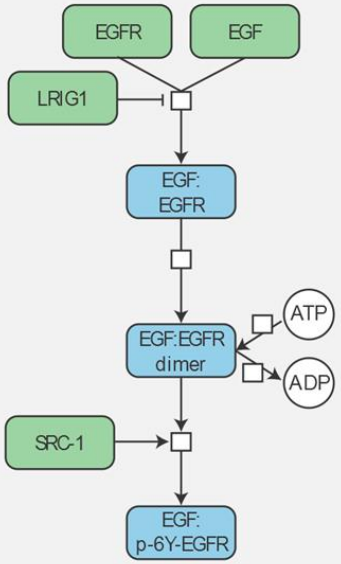


Cytoscape: custom network

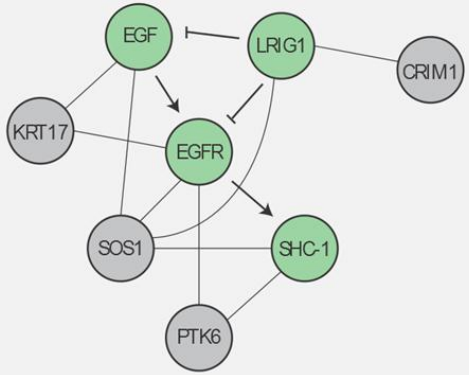
purpose of a network = studying interactions

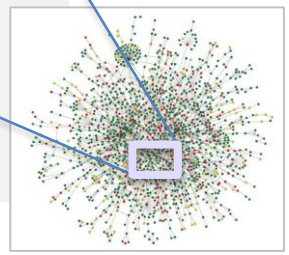
Pathway versus Network

EGFR-centered Pathway



EGFR-centered Network





- Detailed, high-confidence consensus
- Biochemical reactions
- Small-scale, fewer genes
- Concentrated from decades of literature

- Simplified cellular logic, noisy
- Abstractions: directed, undirected
- Large-scale, genome-wide
- Constructed from *omics* data integration

Why Would We Use Network Visualization for Biological Data?

- Represent relationships of biological molecules
  - Physical, regulatory, genetic, functional interactions
- Useful for discovering relationships in large data sets
  - Better than tables in Excel
- Visualize multiple data types together
  - Discover interesting patterns
- Network analysis
  - Finding sub-networks with certain properties (densely connected, co-expressed, frequently mutated, clinical characteristics)
  - Finding paths between nodes (or other network “motifs”)
  - Finding central nodes in network topology (“hub” genes)

Steps in Network Visualization & Analysis

1. Need a software to create the network
2. Upload our data (usually table format) to create the network
3. Navigate through the created network
4. Analyse the network (e.g do you see clusters?)
5. Export network (table) and network image

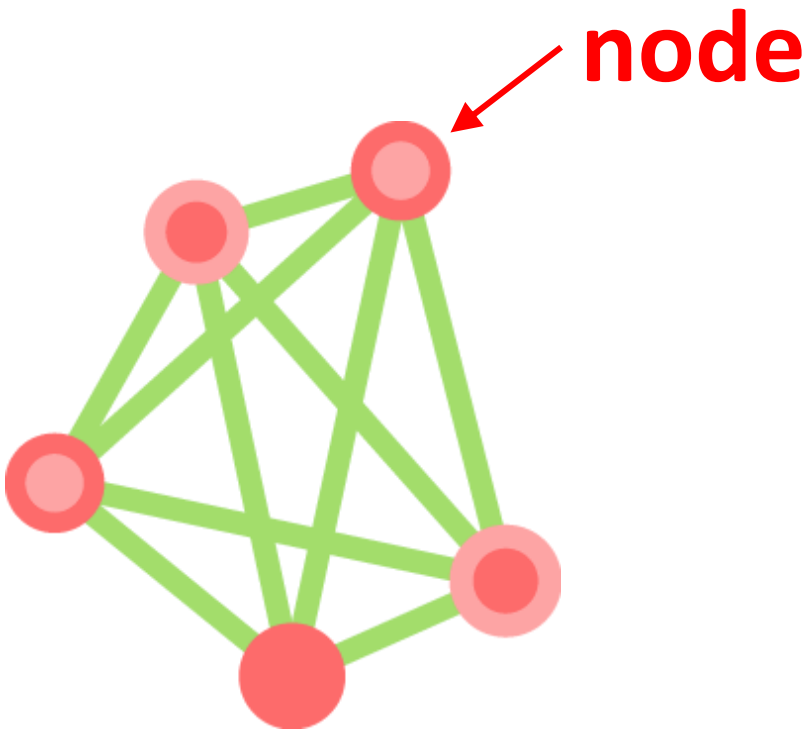
Network Basics

Node (molecule)

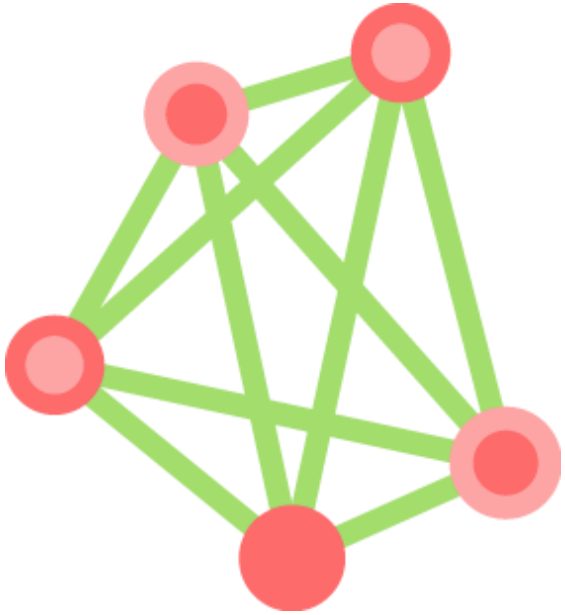
- Gene
- Protein
- Transcript
- Drug
- MicroRNA
- …

Edge (interaction)

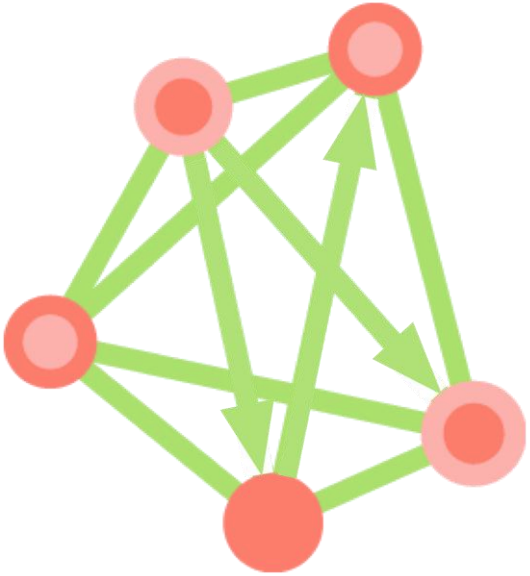
- Genetic interaction
- Physical protein interaction
- Co-expression
- Signaling interaction
- Metabolic reaction
- DNA-binding



Directed or Undirected Graph



undirected

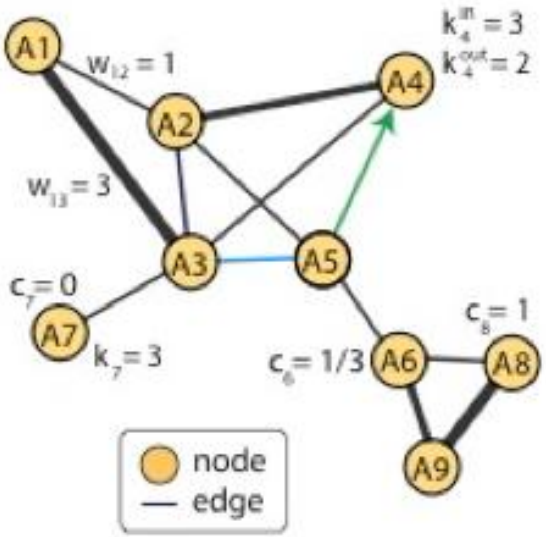


directed

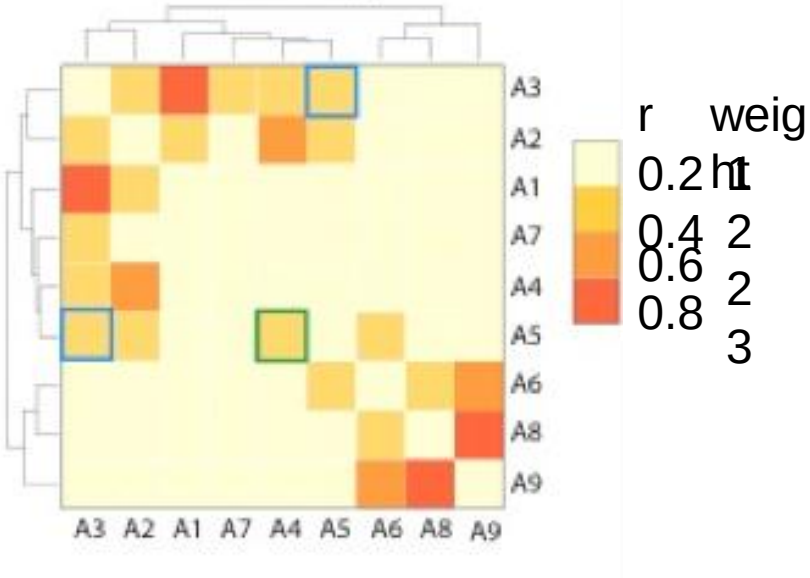
Weighted Edges

Relationships	Optional weight
A1 ↔ A2	1
A1 ↔ A3	3
A2 ↔ A3	1
A2 ↔ A4	2
A2 ↔ A5	1
A3 ↔ A4	1
A3 ↔ A5	1
A3 ↔ A7	1
A5 → A4	1
A5 ↔ A6	1
A6 ↔ A8	1
A6 ↔ A9	2
A8 ↔ A9	3

Network

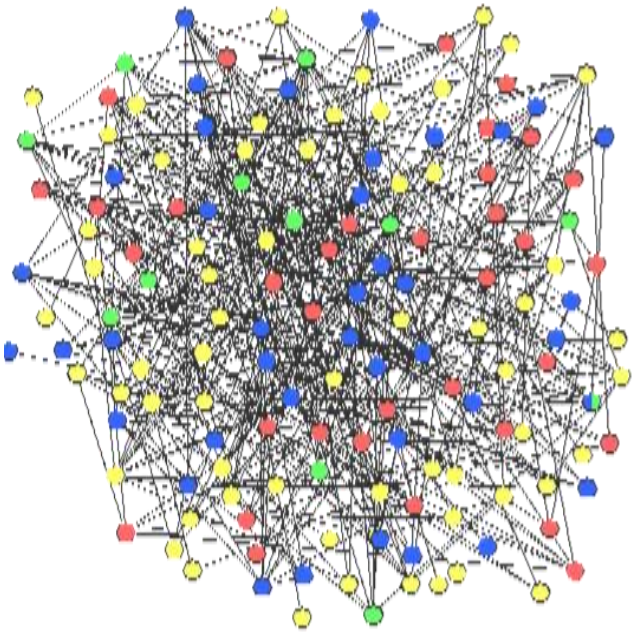


Heatmap

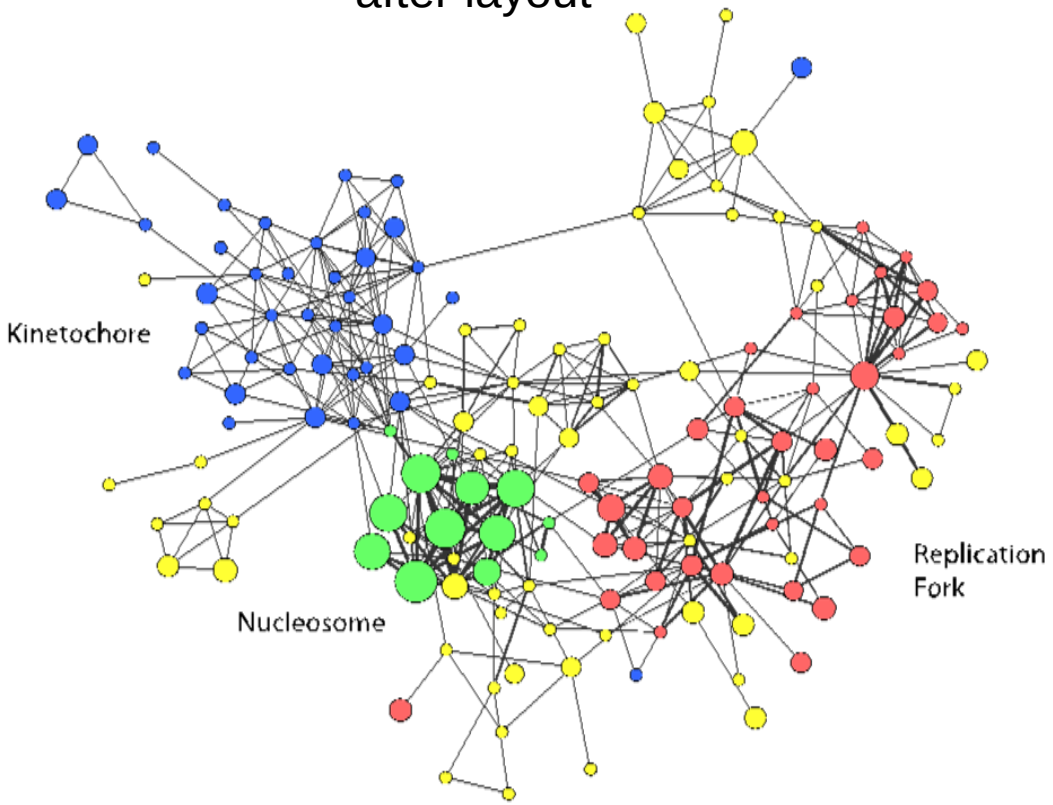


Network Layout

before layout

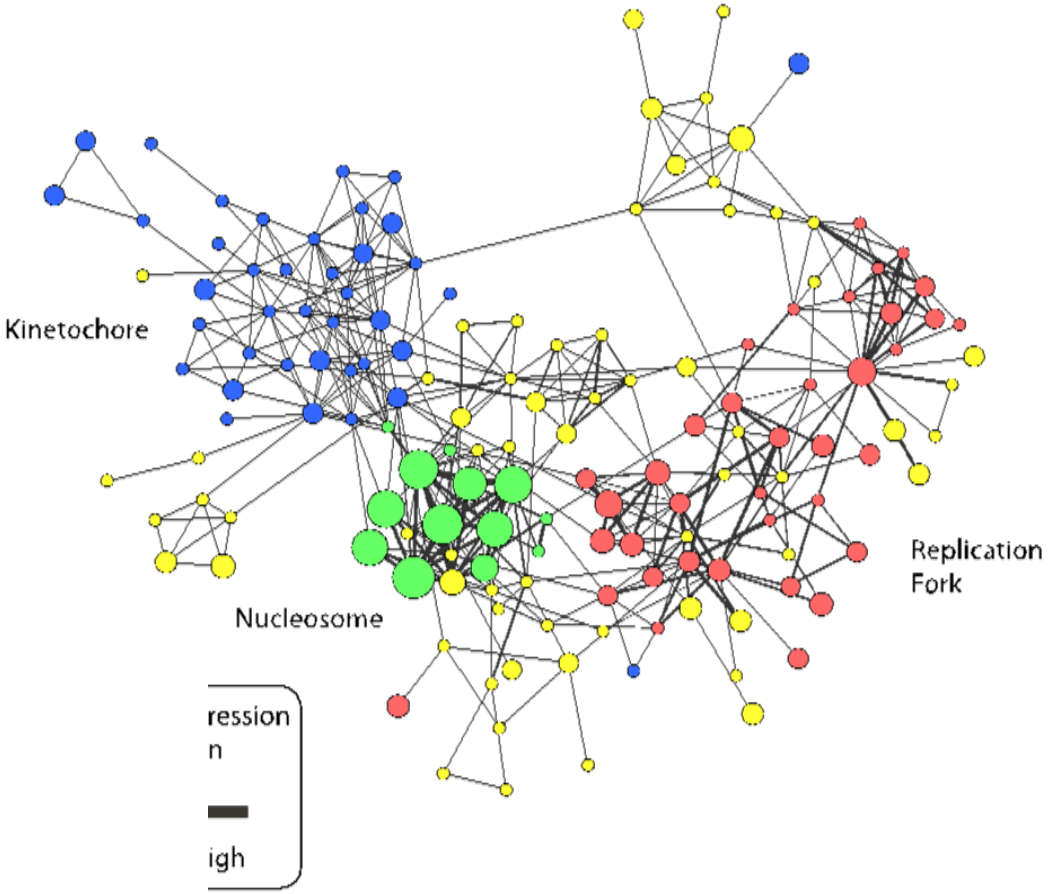


after layout



Network analysis

Use layout to interpret network = network analysis

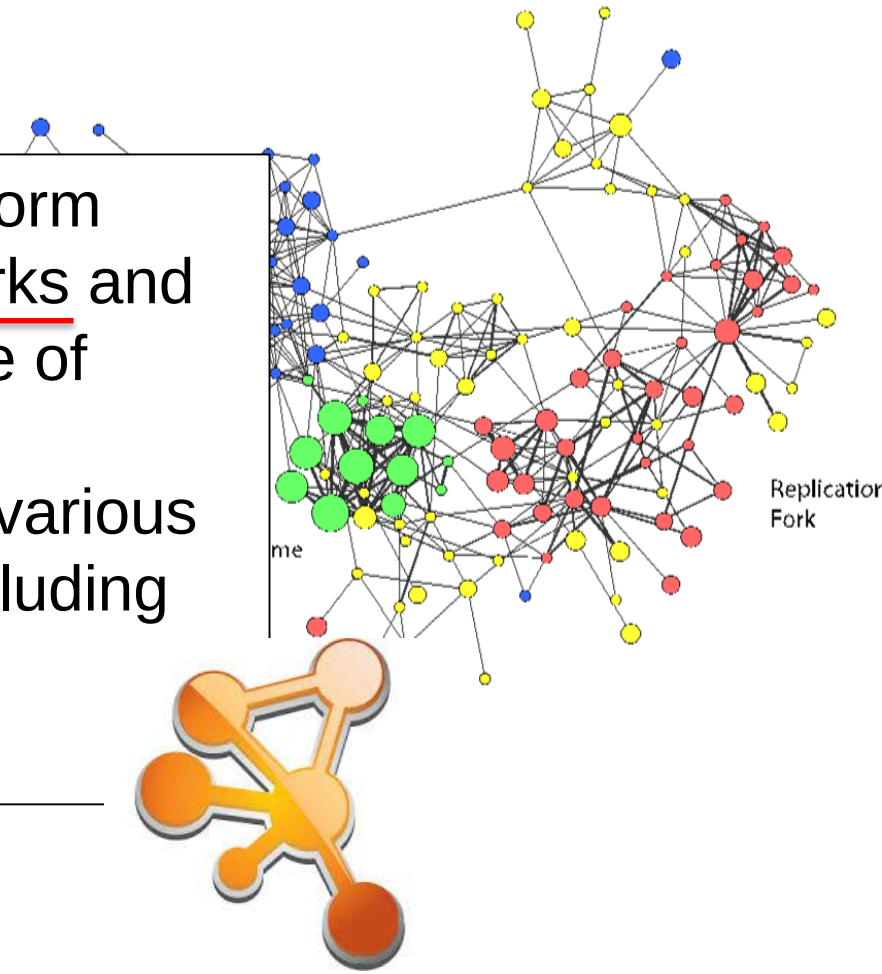


- Finding sub-networks with certain properties (densely connected, co-expressed, frequently mutated, clinical characteristics)
- Finding paths between nodes (or other network "motifs")
- Finding central nodes in network topology ("hub" genes)

Cytoscape

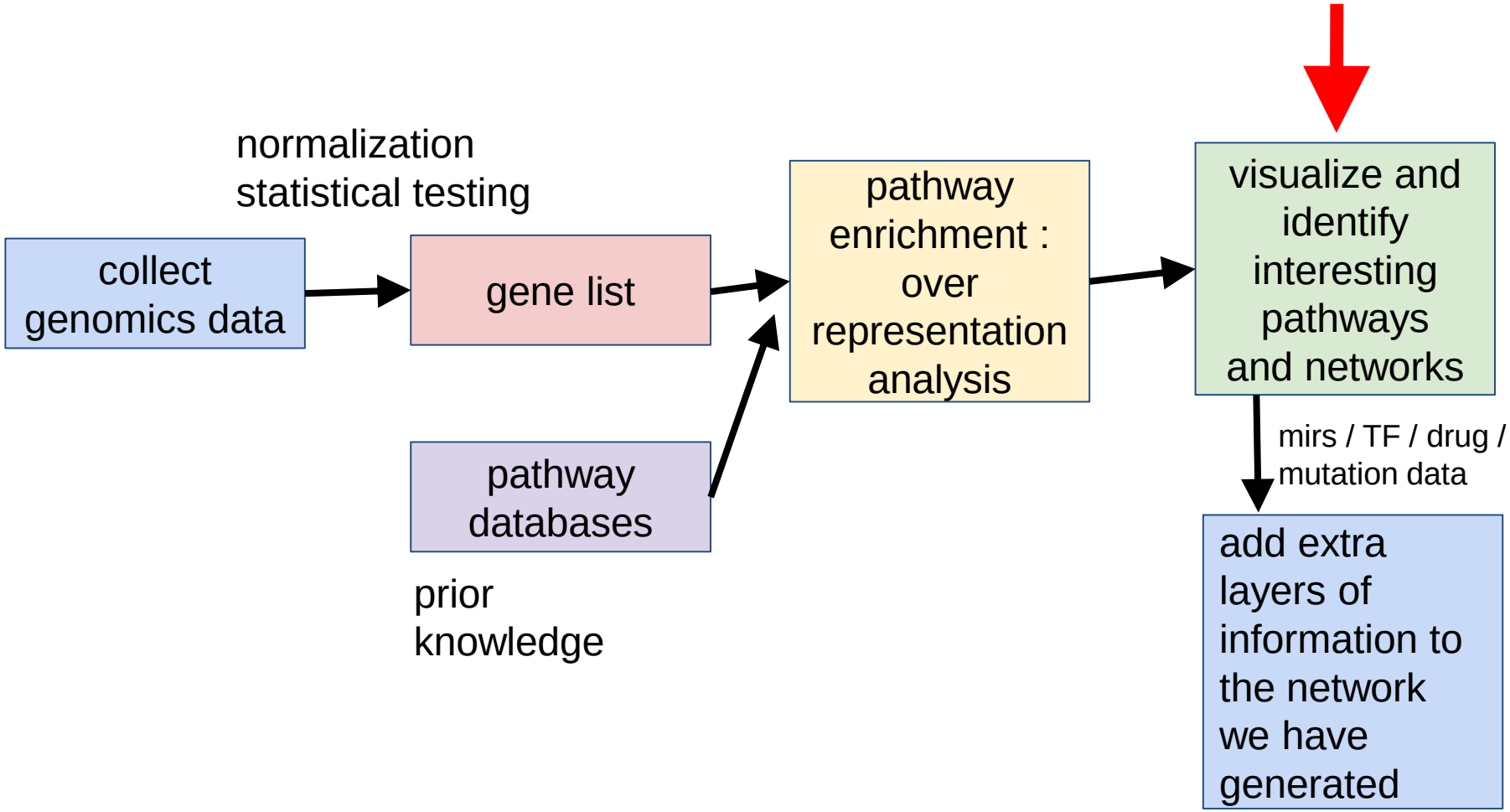
Cytoscape is

- an open source software platform
- for visualizing complex networks and integrating these with any type of attribute data.
- a lot of apps are available for various kinds of problem domains, including bioinformatics, social network analysis, and semantic web.



Network Visualization & Analysis with Cytoscape and EnrichmentMap

Where are we in the workflow?



Creating Networks

gene list

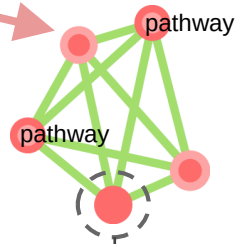
network

large (100- 2000 genes)

medium (100 genes)

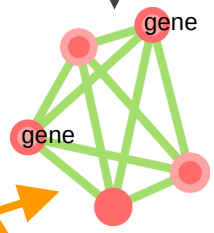
summarize by pathways

- represent as a network of pathways



EnrichmentMap

- represent as a network of genes (gene products)

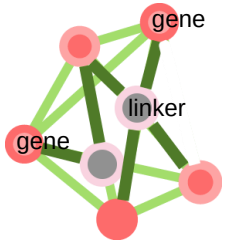


Reactome FI

small (1-50 genes)

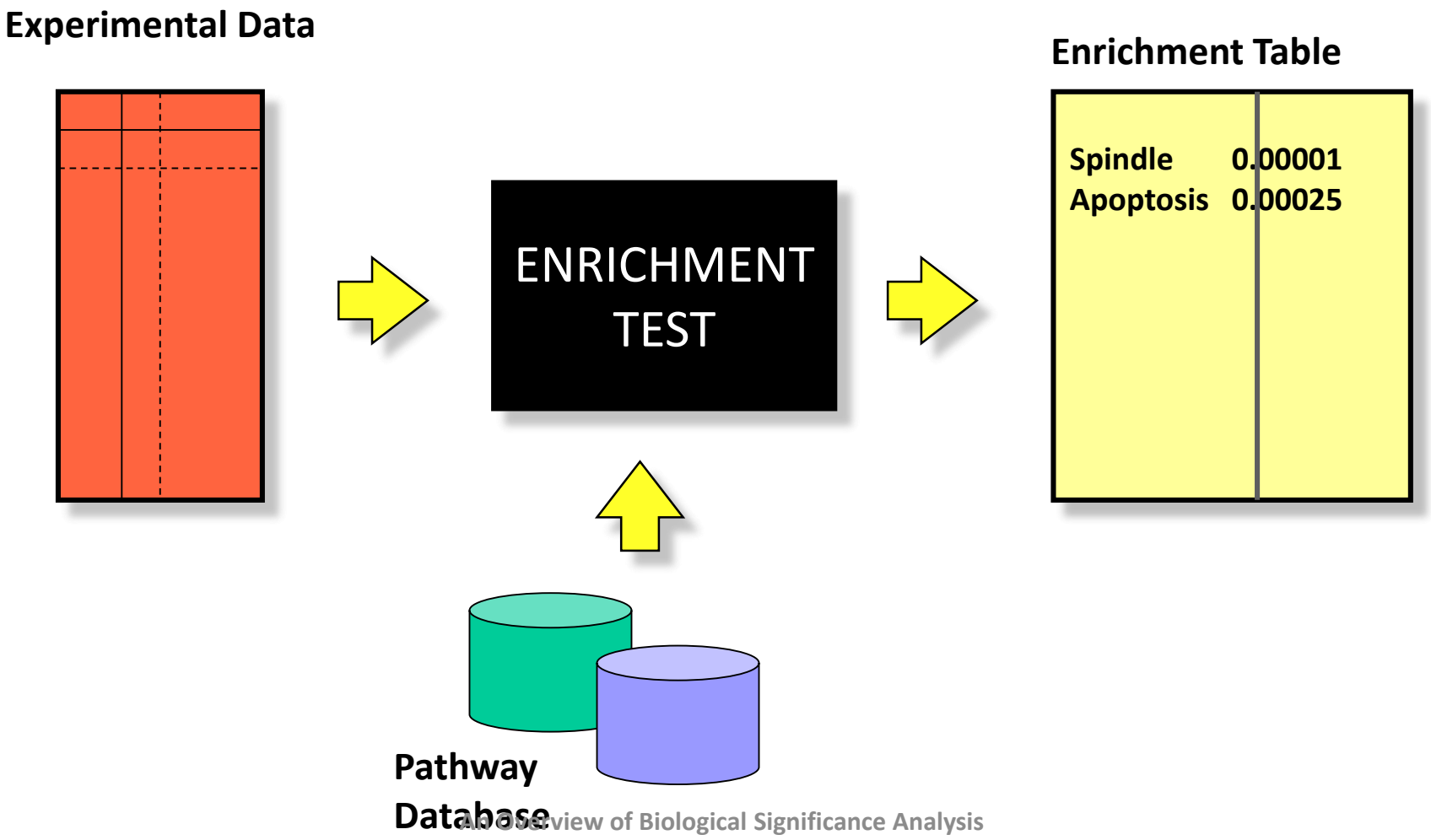
expand the list; use function prediction

- represent as a network of gene (gene products) and add gene linkers

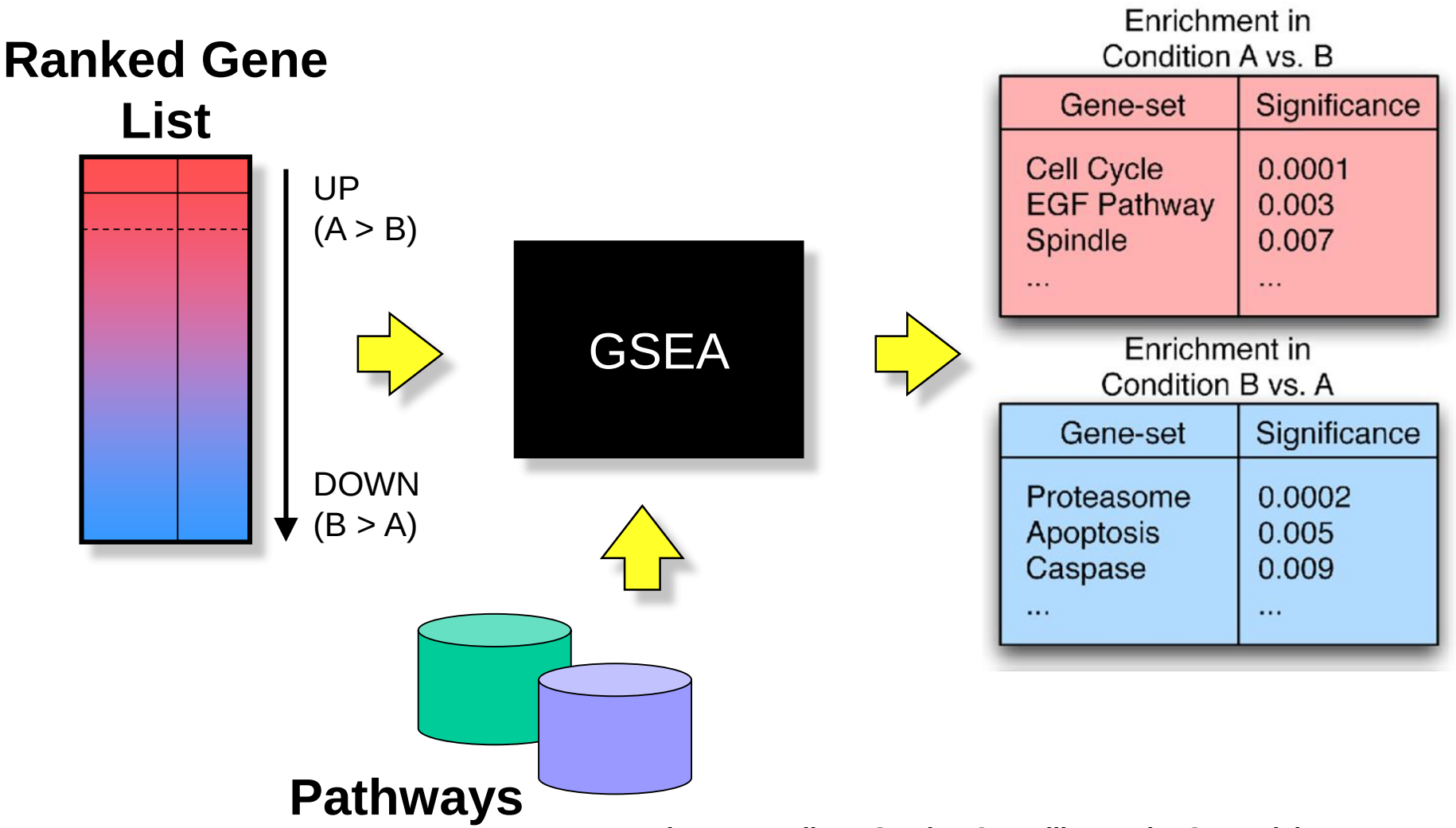


Reactome FI with linkers or geneMANIA

Pathway Enrichment Test: General Framework



Gene Set Enrichment Analysis (GSEA)



Merico D, Isserlin R, Stueker O, Emili A, Bader GD Enrichment map: a network-based method for gene-set enrichment visualization and interpretation PLoS One. 2010 Nov 15;5(11):e13984

Overcoming Gene Set Redundancy

- Pathway Analysis is expected to find a small number of coherent gene sets.
- Increasing number and redundancy of gene sets → noisy results difficult to interpret.
- Different approaches exist to reduce redundancy
  - DAVID groups gene sets by biological themes provides a Tabular representation
  - ClueGO is a cytoscape plugin that does similar and provides network visualization

Enrichment Map

- Network-based visualization method for gene-set enrichment results.
  - Gene-sets are organized in a network, where
    - each set is a node and
    - edges represent gene overlap between sets.
  - Automated network layout groups related gene-sets into network clusters,
  - This enables users to
    - quickly identify the major enriched functional themes
    - and more easily interpret the enrichment results.

Enrichment Map

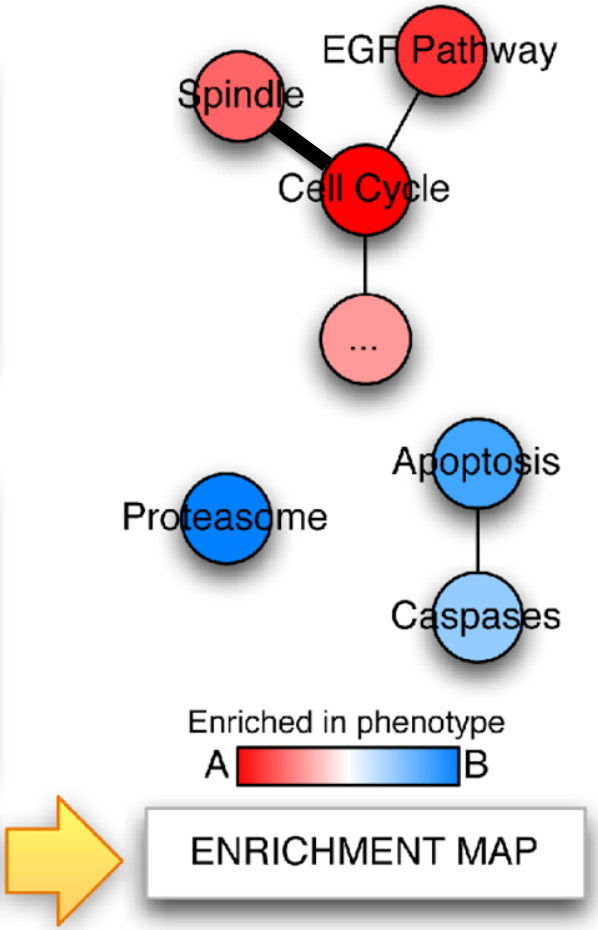
Enrichment in Condition A vs. B

Gene-set	Significance
Cell Cycle	0.0001
EGF Pathway	0.003
Spindle	0.007
...	...

Enrichment in Condition B vs. A

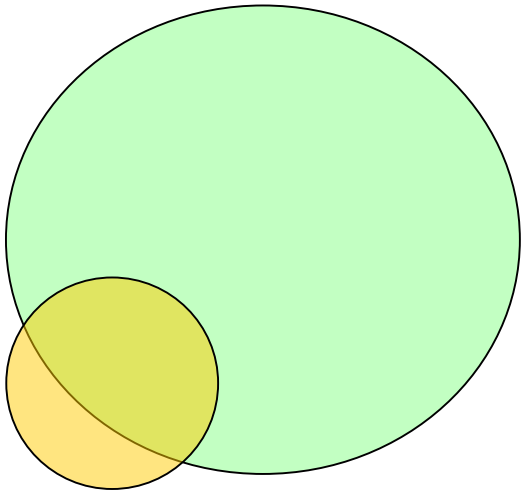
Gene-set	Significance
Proteasome	0.0002
Apoptosis	0.005
Caspase	0.009
...	...

GENE-SET LIST



ENRICHMENT MAP

Overlap



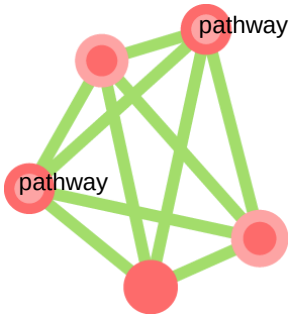
$$\frac{|A \cap B|}{\min(|A|,|B|)}$$

Typical output

	overlap	score	pvalue	FDR
RNA HELICASE ACTIVITY%GO%GO:0003724	18	1.77	0.0041	0.0464386
MRNA SURVEILLANCE PATHWAY%KEGG%HSA03015	82	1.77	0	0.0466167
UBIQUITIN-DEPENDENT DEGRADATION OF CYCLIN D1%REACTOME%REACT_4.1	50	1.77	0.0021	0.0486015
BIOCARTA_CD40_PATHWAY%MSIGDB_C2%BIOCARTA_CD40_PATHWAY	15	1.77	0.0048	0.0483781
IGF1 PATHWAY%PATHWAY INTERACTION DATABASE NCI-NATURE CURATED DATA%IGF1 PATHWAY	29	1.76	0.003	0.0489742
UBIQUITIN-DEPENDENT PROTEIN CATABOLIC PROCESS%GO%GO:0006511	204	1.76	0	0.0488442
PHAGOSOME%KEGG%HSA04145	147	1.76	0	0.0486164
PROTEASOME COMPLEX%GO%GO:0000502	29	1.76	0.007	0.0490215
ANTIGEN PRESENTATION: FOLDING, ASSEMBLY AND PEPTIDE LOADING OF CLASS I MHC%REACTOME%REACT_7	24	1.76	0.0041	0.0505599
ABORTIVE ELONGATION OF HIV-1 TRANSCRIPT IN THE ABSENCE OF TAT%REACTOME%REACT_6261.3	23	1.75	0	0.0529242
DNA DAMAGE RESPONSE, SIGNAL TRANSDUCTION BY P53 CLASS MEDIATOR RESULTING IN CELL CYCLE ARREST%	67	1.75	0	0.052886
REGULATION OF MACROPHAGE ACTIVATION%GO%GO:0043030	11	1.75	0.003	0.0534709
PROTEIN FOLDING%REACTOME%REACT_16952.1	52	1.75	0.002	0.0537717
ENDOPLASMIC RETICULUM UNFOLDED PROTEIN RESPONSE%GO%GO:0030968	73	1.75	0	0.0546052
PROTEIN EXPORT%KEGG%HSA03060	24	1.75	9.75E-04	0.0548699
TRANSCRIPTION INITIATION FROM RNA POLYMERASE II PROMOTER%GO%GO:0006367	64	1.75	0.001	0.0545783
S PHASE%REACTOME%REACT_899.4	110	1.75	0	0.0546003
PROTEASOMAL PROTEIN CATABOLIC PROCESS%GO%GO:0010498	163	1.75	0	0.0550066
ATP-DEPENDENT RNA HELICASE ACTIVITY%GO%GO:0008026	20	1.74	0.0059	0.0556722
ACID-AMINO ACID LIGASE ACTIVITY%GO%GO:0016881	217	1.74	0	0.0560217
GO%GO:0072474	67	1.74	0.002	0.0565978
GO%GO:0035966	107	1.74	0	0.0562957
GO%GO:0072413	67	1.74	9.81E-04	0.05761
BIOCARTA_IL4_PATHWAY%MSIGDB_C2%BIOCARTA_IL4_PATHWAY	11	1.74	0.0082	0.0581508
ASSOCIATION OF TRIC CCT WITH TARGET PROTEINS DURING BIOSYNTHESIS%REACTOME%REACT_16907.2	28	1.74	0.0039	0.0581298
UBIQUITIN-DEPENDENT DEGRADATION OF CYCLIN D%REACTOME%REACT_938.4	50	1.74	0.0029	0.057876
MODIFICATION-DEPENDENT PROTEIN CATABOLIC PROCESS%GO%GO:0019941	207	1.74	0	0.0576579
TRANSLATION INITIATION COMPLEX FORMATION%REACTOME%REACT_1979.1	55	1.74	0.0021	0.0575181
GO%GO:0001906	13	1.74	0.0117	0.0572877
G1 S TRANSITION%REACTOME%REACT_1783.2	107	1.74	0	0.0572618
GO%GO:0034620	73	1.73	0.0021	0.0576606
SIGNALING BY NOTCH%REACTOME%REACT_299.2	19	1.73	0.0069	0.0578565
RESPONSE TO UNFOLDED PROTEIN%GO%GO:0006986	102	1.73	0	0.0583864
SIGNAL TRANSDUCTION INVOLVED IN G1 S TRANSITION CHECKPOINT%GO%GO:0072404	68	1.73	0.002	0.0582213
GO%GO:0072431	67	1.73	0	0.058551
BIOCARTA_PROTEASOME_PATHWAY%MSIGDB_C2%BIOCARTA_PROTEASOME_PATHWAY	19	1.73	0.0099	0.0586655
HOST INTERACTIONS OF HIV FACTORS%REACTOME%REACT_6288.4	117	1.73	0	0.0586888
AUTOPHAGIC VACUOLE ASSEMBLY%GO%GO:0000045	13	1.73	0.0122	0.0588271
CYCLIN A:CDK2-ASSOCIATED EVENTS AT S PHASE ENTRY%REACTOME%REACT_9029.2	66	1.73	0	0.0610099

TABULAR FORMAT

NETWORK VISUALIZATION

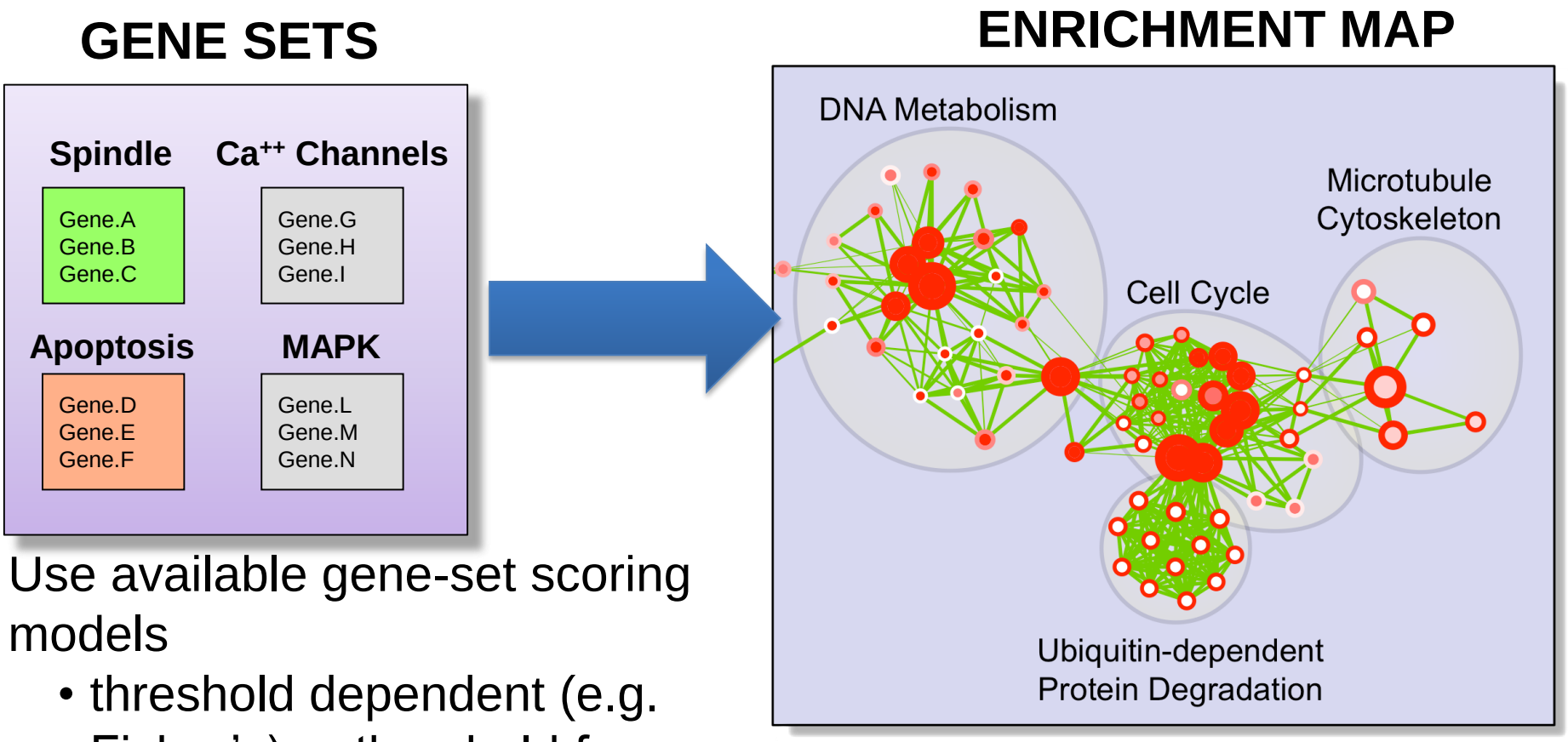


Enrichment Map

Each row is a gene-set (pathway).
It displays:

- a score associated with the magnitude of overlap between gene-set and gene list.
- a pvalue that estimates the significance of the enrichment (by chance or not).
- An adjusted pvalue (FDR) that corrects for multiple hypothesis testing.

Enrichment Map



- Use available gene-set scoring models
  - threshold dependent (e.g. Fisher's) or threshold free (e.g. GSEA)
- Use the network framework to organize gene-sets exploiting their inter-dependencies

http://baderlab.org/Software/EnrichmentMap/

Enrichment Map: use case I
Single enrichment

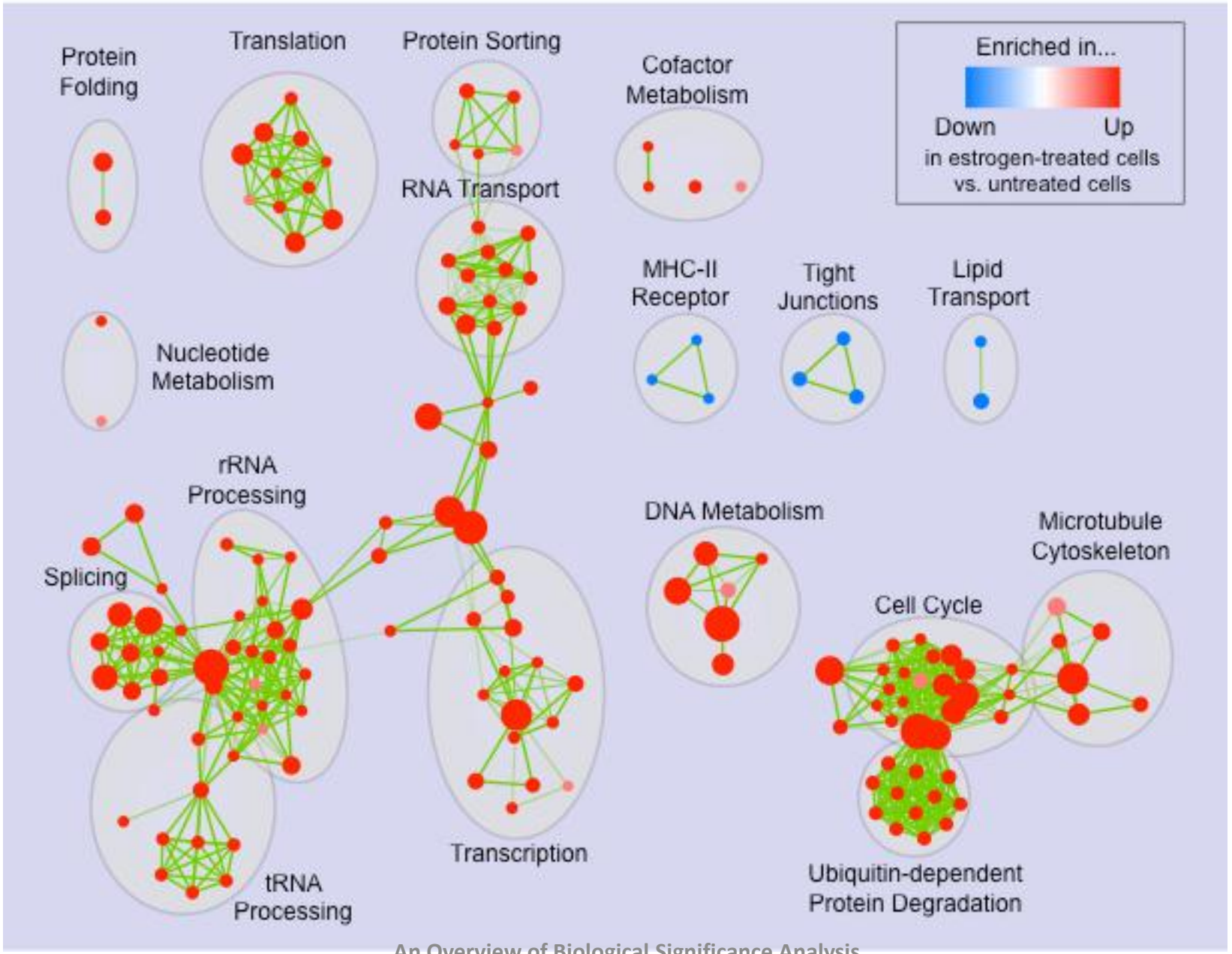
Estrogen treatment of breast cancer cells

- Design:
  2-time points, two-class

	12 hrs	24 hrs
Estrogen-treated	3	3
Untreated	3	3

- Gene set Database:
  Gene Ontology

Lin C-Y, Vega VB, Thomsen JS, Zhang T, Kong SL, et al. (2007) Whole-genome cartography of estrogen receptor alpha binding sites. PLoS Genetics 3:e87


Protein Folding
Translation
Protein Sorting
Cofactor Metabolism
Enriched in...
Down
Up
in estrogen-treated cells vs. untreated cells
RNA Transport
MHC-II Receptor
Tight Junctions
Lipid Transport
Nucleotide Metabolism
rRNA Processing
DNA Metabolism
Microtubule Cytoskeleton
Splicing
Cell Cycle
Transcription
tRNA Processing
Ubiquitin-dependent Protein Degradation

Examples

Step by step examples of
some analyses described
in previous slides

List of Examples

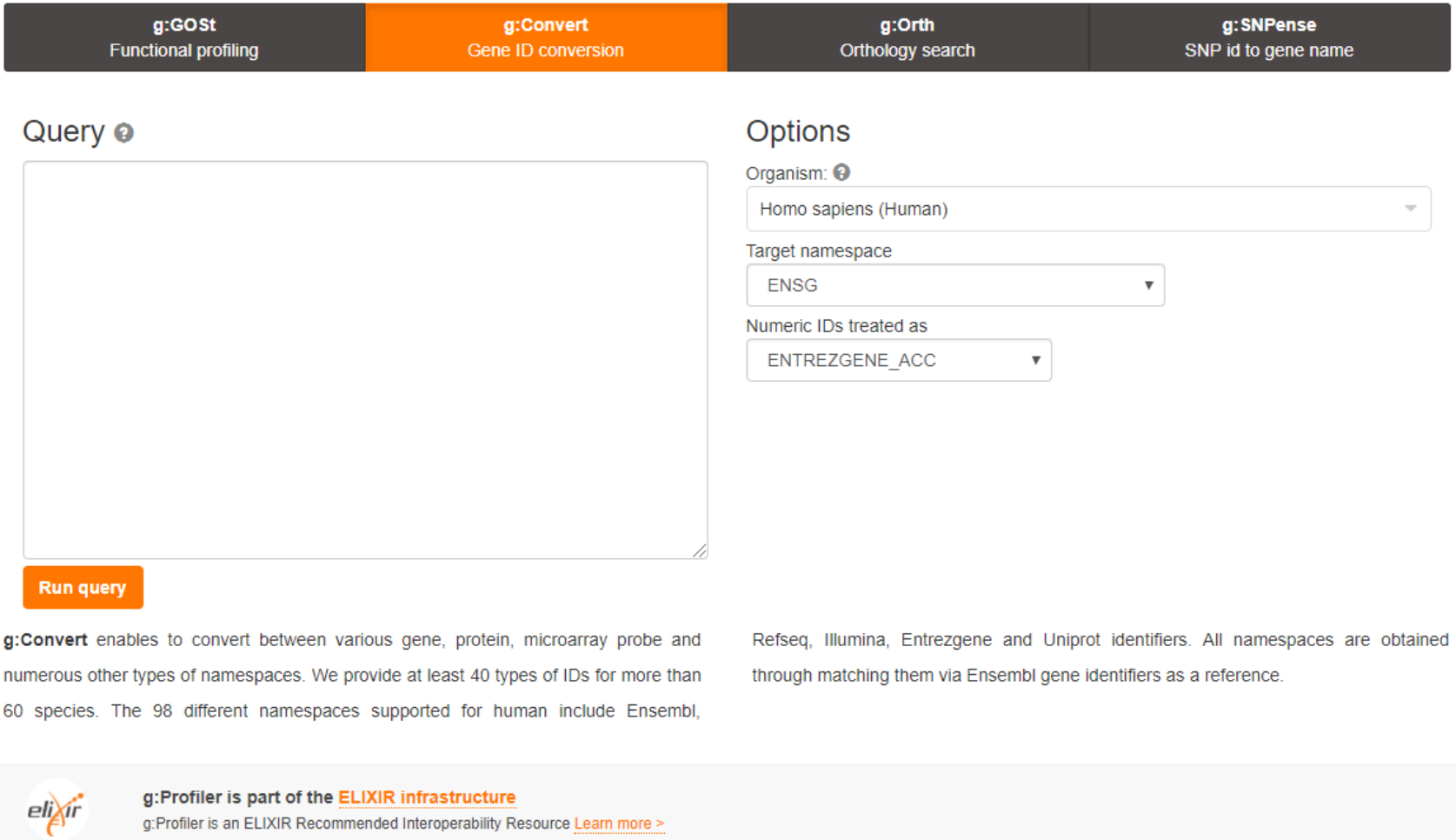
1. Gene ID conversion with g:Profiler
2. Going through the protocol:

   from enrichment analysis to visualization

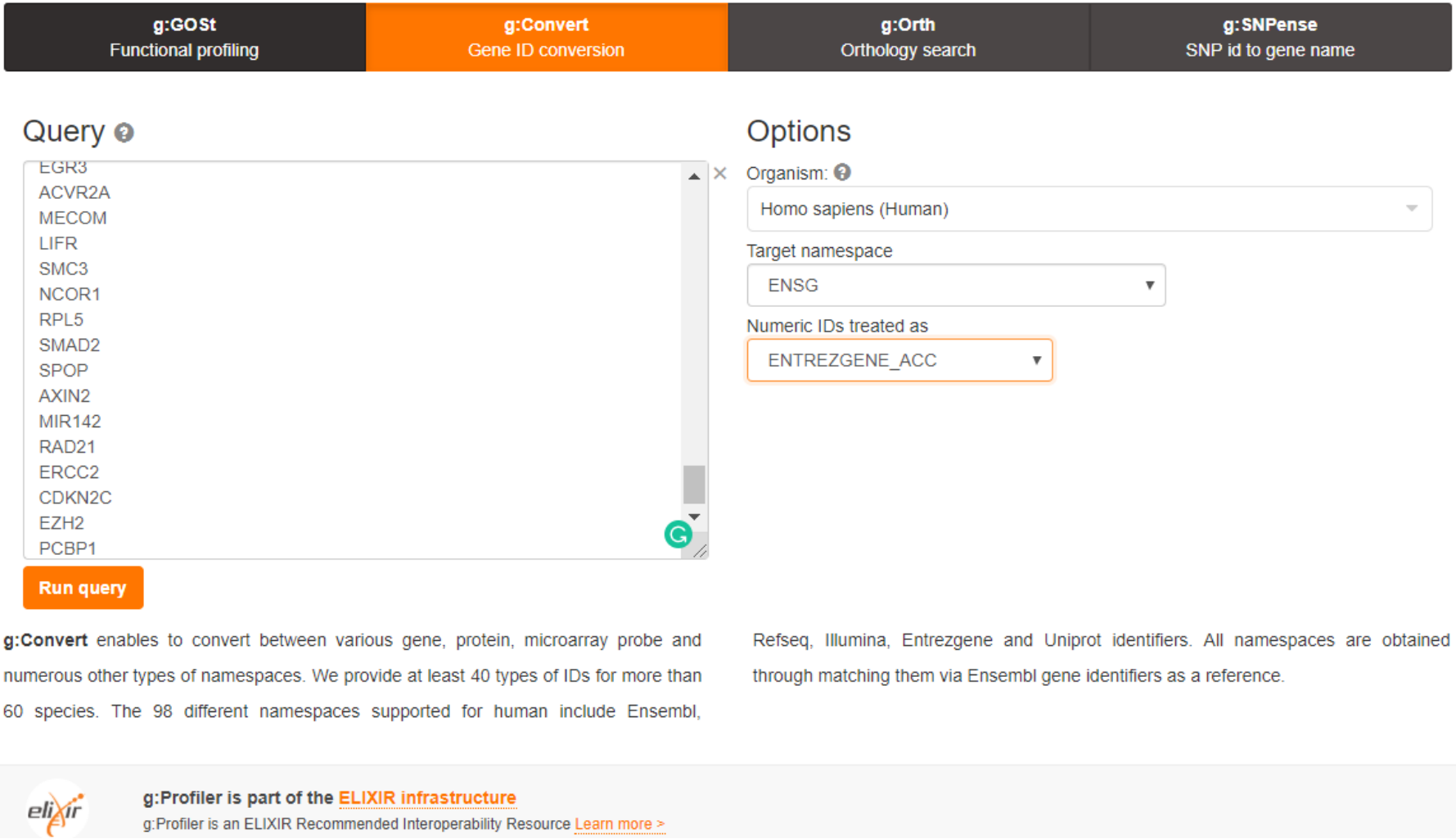
Gene ID conversion

- Gene list in supplementary table 1 is made of gene símbols (HUGO)
- Many programs require identifiers to be provided as ENTREZ Gene or ENSEMBL
- gProfiler can be used to transform identifiers

Gene ID conversion with gProfiler (1)



Gene ID conversion with gProfiler (2)



Gene ID conversion with gProfiler (3)

Input: 127 gene símbols → Output 124 ENSEMBL identifiers

Export to CSV | Show query URL | Show short link

initial alias	converted alias	name	description	namespace
TP53	ENSG00000141510	TP53	tumor protein p53 [Source:HGNC Symbol;Acc:HGNC:11998]	DBASS3, DBASS5, ENTREZGENE, H
PIK3CA	ENSG00000121879	PIK3CA	phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha [Source:HGNC Sy...	ENTREZGENE, HGNC, UNIPROT_G
PTEN	ENSG00000171862	PTEN	phosphatase and tensin homolog [Source:HGNC Symbol;Acc:HGNC:9588]	DBASS5, ENTREZGENE, HGNC, UN
APC	ENSG00000134982	APC	APC, WNT signaling pathway regulator [Source:HGNC Symbol;Acc:HGNC:583]	DBASS3, ENTREZGENE, HGNC, UN
VHL	ENSG00000134086	VHL	von Hippel-Lindau tumor suppressor [Source:HGNC Symbol;Acc:HGNC:12687]	DBASS5, ENTREZGENE, HGNC, UN
KRAS	ENSG00000133703	KRAS	KRAS proto-oncogene, GTPase [Source:HGNC Symbol;Acc:HGNC:6407]	ENTREZGENE, HGNC, UNIPROT_G
MLL3	None	None	None	
MLL2	None	None	None	
ARID1A	ENSG00000117713	ARID1A	AT-rich interaction domain 1A [Source:HGNC Symbol;Acc:HGNC:11110]	ENTREZGENE, HGNC, UNIPROT_G
PBRM1	ENSG00000163939	PBRM1	polybromo 1 [Source:HGNC Symbol;Acc:HGNC:30064]	ENTREZGENE, HGNC, UNIPROT_G
NAV3	ENSG00000067798	NAV3	neuron navigator 3 [Source:HGNC Symbol;Acc:HGNC:15998]	ENTREZGENE, HGNC, UNIPROT_G
EGFR	ENSG00000146648	EGFR	epidermal growth factor receptor [Source:HGNC Symbol;Acc:HGNC:3236]	ENTREZGENE, HGNC, UNIPROT_G
NF1	ENSG00000196712	NF1	neurofibromin 1 [Source:HGNC Symbol;Acc:HGNC:7765]	DBASS3, DBASS5, ENTREZGENE, H

Laboratory seminar

nature protocols

PROTOCOL

https://doi.org/10.1038/s41596-018-0103-9

Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap

Jüri Reimand[1,2,8], Ruth Isserlin[3,8], Veronique Voisin[3], Mike Kucera[3], Christian Tannus-Lopes[3], Asha Rostamianfar[3], Lina Wadi[1], Mona Meyer[1], Jeff Wong[3], Changjiang Xu[3], Daniele Merico[4,5] and Gary D. Bader[3,6,7*]

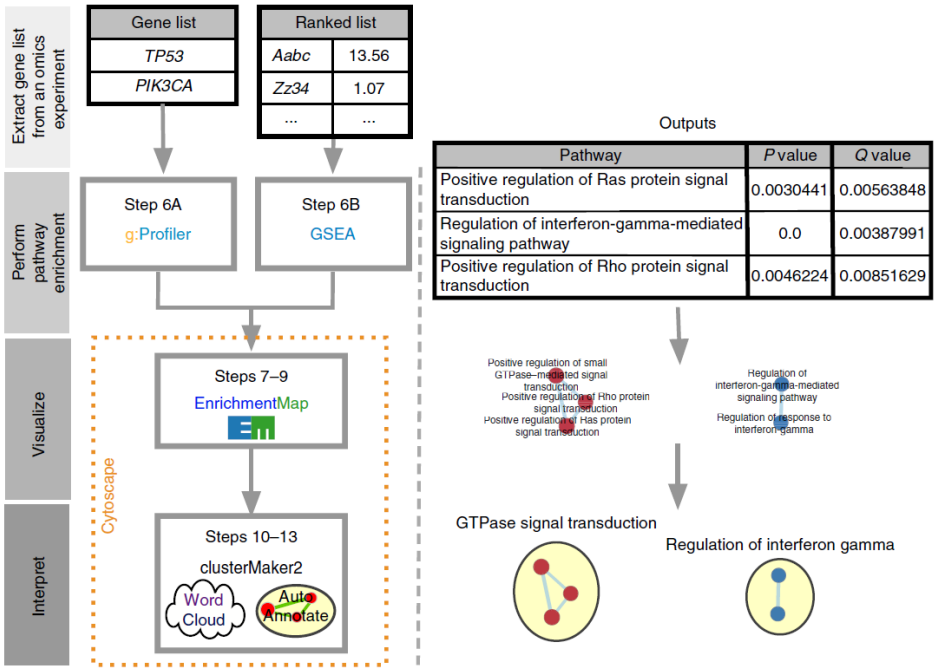
&

Heatmap visualization of IPA

06.05.19

Stéphane Nemours-Fisiopatología Renal-CIBBIM-Nanomedicina

Overview of the procedure



The protocol comprises three major steps:

1. Definition of a gene list from omics data
2. Determination of statistically enriched Pathways
3. Visualization and interpretation of the results.

Stéphane Nemours-Fisiopatología Renal-CIBBIM-Nanomedicina

Stage 1: Definition of a gene list of interest using omics data

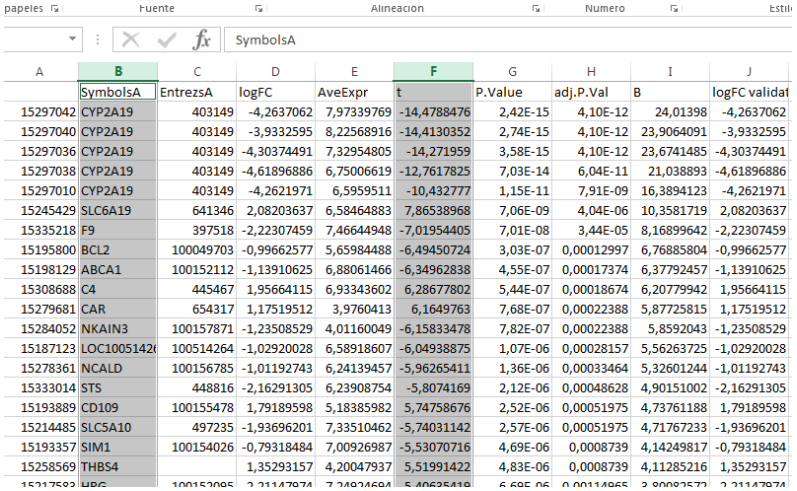
There are two major ways to define a gene list from omics data: list or ranked list.

Certain omics data naturally produce a gene list, such as all somatically mutated genes in a tumor identified by exome sequencing, or all proteins that interact with a bait in a proteomics experiment. Such a list is suitable for direct input into pathway enrichment analysis using **g:Profiler** (Step 6A).

Other omics data naturally produce ranked lists. For example, a list of genes can be ranked by differential gene expression score or sensitivity in a genome-wide CRISPR screen. Some pathway enrichment analysis approaches analyze a ranked gene list filtered by a particular threshold (e.g., FDR-adjusted P value 2). Alternative approaches, such as **GSEA**, are designed to analyze ranked lists of all available genes and do not require a threshold (Step 6B).

Stéphane Nemours-Fisiopatología Renal-CIBBIM-Nanomedicina

List/Ranked gene list



papeles | Fuente | Alineacion | Numero | Estil

SymbolsA

	A	B	C	D	E	F	G	H	I	J
		SymbolsA	EntrezsA	logFC	AveExpr	t	P.Value	adj.P.Val	B	logFC validat
15297042		CYP2A19	403149	-4,2637062	7,97339769	-14,4788476	2,42E-15	4,10E-12	24,01398	-4,2637062
15297040		CYP2A19	403149	-3,9332595	8,22568916	-14,4130352	2,74E-15	4,10E-12	23,9064091	-3,9332595
15297036		CYP2A19	403149	-4,30374491	7,32954805	-14,271959	3,58E-15	4,10E-12	23,6741485	-4,30374491
15297038		CYP2A19	403149	-4,61896886	6,75006619	-12,7617825	7,03E-14	6,04E-11	21,038893	-4,61896886
15297010		CYP2A19	403149	-4,2621971	6,5959511	-10,432777	1,15E-11	7,91E-09	16,3894123	-4,2621971
15245429		SLC6A19	641346	2,08203637	6,58464883	7,86538968	7,06E-09	4,04E-06	10,3581719	2,08203637
15335218		F9	397518	-2,22307459	7,46644948	-7,01954405	7,01E-08	3,44E-05	8,16899642	-2,22307459
15195800		BCL2	100049703	-0,99662577	5,65984488	-6,49450724	3,03E-07	0,00012997	6,76885804	-0,99662577
15198129		ABCA1	100152112	-1,13910625	6,88061466	-6,34962838	4,55E-07	0,00017374	6,37792457	-1,13910625
15308688		C4	445467	1,95664115	6,93343602	6,28677802	5,44E-07	0,00018674	6,20779942	1,95664115
15279681		CAR	654317	1,17519512	3,9760413	6,1649763	7,68E-07	0,00022388	5,87725815	1,17519512
15284052		NKAIN3	100157871	-1,23508529	4,01160049	-6,15833478	7,82E-07	0,00022388	5,8592043	-1,23508529
15187123		LOC1005142(	100514264	-1,02920028	6,58918607	-6,04938875	1,07E-06	0,00028157	5,56263725	-1,02920028
15278361		NCALD	100156785	-1,01192743	6,24139457	-5,96265411	1,36E-06	0,00033464	5,32601244	-1,01192743
15333014		STS	448816	-2,16291305	6,23908754	-5,8074169	2,12E-06	0,00048628	4,90151002	-2,16291305
15193889		CD109	100155478	1,79189598	5,18385982	5,74758676	2,52E-06	0,00051975	4,73761188	1,79189598
15214485		SLC5A10	497235	-1,93696201	7,33510462	-5,74031142	2,57E-06	0,00051975	4,71767233	-1,93696201
15193357		SIM1	100154026	-0,79318484	7,00926987	-5,53070716	4,69E-06	0,0008739	4,14249817	-0,79318484
15258569		THBS4		1,35293157	4,20047937	5,51991422	4,83E-06	0,0008739	4,11285216	1,35293157
15217583		HBG	100152085	2,21147974	7,24924694	5,40635419	6,69E-06	0,00114965	3,89082572	2,21147974

```
##   geneID   t_stat
## 1   TFEC 20.75773
## 2   CD86 17.36037
## 3   CD48 17.30732
## 4  IL2RG 16.87160
## 5   LCP2 16.73821
## 6  GPR65 16.65257
```

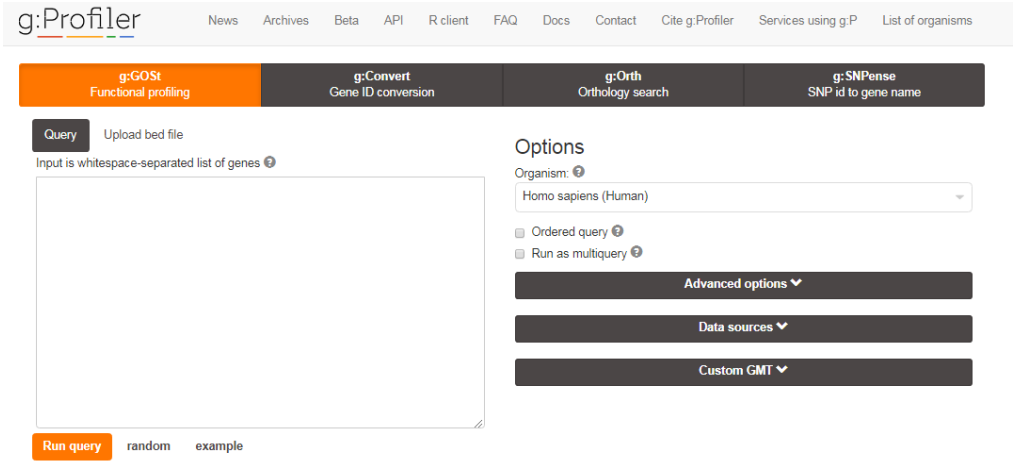
FPSvsFPR.rnk

Stéphane Nemours-Fisiopatología Renal-CIBBIM-Nanomedicina

Stage 2A: pathway enrichment analysis of a gene list using g:Profiler (Step 6A)

The default analysis implemented in g:Profiler and similar web-based tools searches for pathways whose genes are significantly enriched (i.e., over-represented) in the fixed list of genes of interest, as compared to all genes in the genome (Step 6A). The P value of the enrichment of a pathway is computed using a **Fisher's exact test** and **multiple-test correction** is applied.

g:Profiler searches a collection of gene sets representing Gene Ontology (GO) terms, pathways, networks, regulatory motifs, and disease phenotypes. Major categories of gene sets can be selected to customize the search.

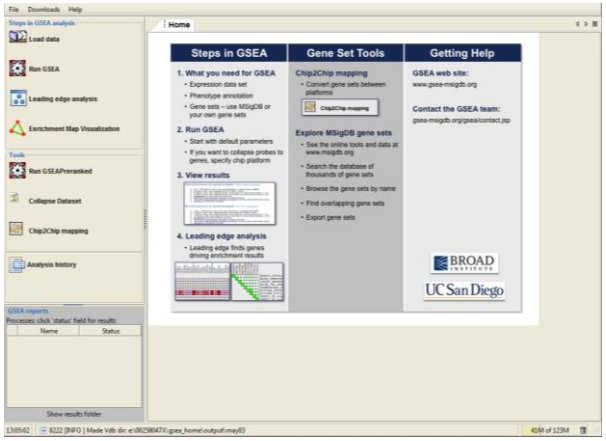


Stéphane Nemours-Fisiopatología Renal-CIBBIM-Nanomedicina

Stage 2B: pathway enrichment analysis of a ranked gene list using GSEA (Step 6B)

Pathway enrichment analysis of a ranked gene list is implemented in the GSEA software14 (Step 6B) (Box 4). GSEA is a threshold-free method that analyzes all genes on the basis of their differential expression rank, or other score, without prior gene filtering. GSEA is particularly suitable **and is recommended when ranks are available** for all or most of the genes in the genome (e.g., for RNA-seq data).

GSEA searches for pathways whose genes are enriched at the top or bottom of the ranked gene list, more so than expected by chance alone. For instance, if the topmost differentially expressed genes are involved in the cell cycle, this suggests that the cell cycle pathway is regulated in the experiment. By contrast, the cell cycle pathway is probably not significantly regulated if the cell cycle genes appear randomly scattered through the whole ranked list.



Stéphane Nemours-Fisiopatología Renal-CIBBIM-Nanomedicina

Enrichment Score (ES) and Normalized Enrichment Score (NES)

To calculate an **enrichment score (ES)** for a pathway, GSEA progressively examines genes from the top to the bottom of the ranked list, increasing the ES if a gene is part of the pathway and decreasing the score otherwise. These running sum values are weighted, so that enrichment in the very top- (and bottom-) ranking genes is amplified, whereas enrichment in genes with more moderate ranks are not amplified. The ES score is calculated as the maximum value of the running sum and normalized relative to pathway size, resulting in a **normalized enrichment score (NES)** that reflects the enrichment of the pathway in the list. Positive and negative NES values represent enrichment at the top and bottom of the list, respectively.

Finally, a permutation-based **P value** is computed and corrected for **multiple testing** to produce a permutation based false-discovery rate (FDR) Q value that ranges from 0 (highly significant) to 1 (not significant). The same analysis is performed starting from the bottom of the ranked gene list to identify pathways enriched in the bottom of the list. Resulting pathways are selected using the FDR Q value threshold (e.g., $Q < 0.05$) and ranked using NES. In addition, the 'leading edge' aspect of the GSEA analysis identifies specific genes that most strongly contribute to the detected enrichment signal of a pathway.

Stéphane Nemours-Fisiopatología Renal-CIBBIM-Nanomedicina

Stage 3: visualization and interpretation of pathway enrichment analysis results

Pathway information is inherently **redundant**, as genes often participate in multiple pathways, and databases may organize pathways hierarchically by including general and specific pathways with many shared genes (e.g., 'cell cycle' and 'M-phase of cell cycle'). Consequently, pathway enrichment analysis often highlights several versions of the same pathway. Collapsing redundant pathways into a single biological theme simplifies interpretation. We recommend addressing such redundancy with visualization methods such as EnrichmentMap16, ClueGO40 and others.

An 'enrichment map' is a network visualization that represents overlaps among enriched pathways (Fig. 1), whereas 'EnrichmentMap' refers to the Cytoscape application that creates the visualization. An enrichment map helps identify interesting pathways and themes.

Finally, pathway enrichment analysis results can be published to support a scientific conclusion (e.g., functional differences of two cancer subtypes), or used for hypothesis generation or planning of experiments to support the identification of novel pathways.

Stéphane Nemours-Fisiopatología Renal-CIBBIM-Nanomedicina

Tools needed before starting…

Hardware

- A personal computer with Internet access and ≥8 GB of RAM. 1 GB of RAM is sufficient to run GSEA analysis; however, Cytoscape (required to run EnrichmentMap software) requires ≥8 GB of RAM.

Software

- A contemporary web browser (e.g., Chrome) for pathway enrichment analysis with g:Profiler (Step 6A).
- g:Profiler (https://biit.cs.ut.ee/gprofiler/)
- Java Standard Edition (http://www.oracle.com/technetwork/java/javase/downloads/index.html) is required to run GSEA and Cytoscape.
- GSEA desktop application (http://software.broadinstitute.org/gsea/downloads.jsp) is used for pathway enrichment analysis (Step 6B).
- The Cytoscape desktop application (http://www.cytoscape.org/download.php), as well as the following Cytoscape applications, is required for enrichment map visualization:
    - EnrichmentMap, v.3.1 or higher;
    - clusterMaker2, v.0.9.5 or higher;
    - WordCloud, v.3.1.0 or higher;
    - AutoAnnotate, v,1.2.0 or higher.

  •These can be conveniently downloaded and installed together by installing the 'EnrichmentMap Pipeline Collection' (http://apps.cytoscape.org/apps/enrichmentmappipelinecollection) from the Cytoscape App Store.

Stéphane Nemours-Fisiopatología Renal-CIBBIM-Nanomedicina

Software installation ● Timing 5 min

Download the required input and output.

● Create two directories, project data folder and results data folder.

● Place all downloaded input and example output files into the project data folder.

● As you progress through the protocol, place any newly generated files into the results data folder.

Stéphane Nemours-Fisiopatología Renal-CIBBIM-Nanomedicina

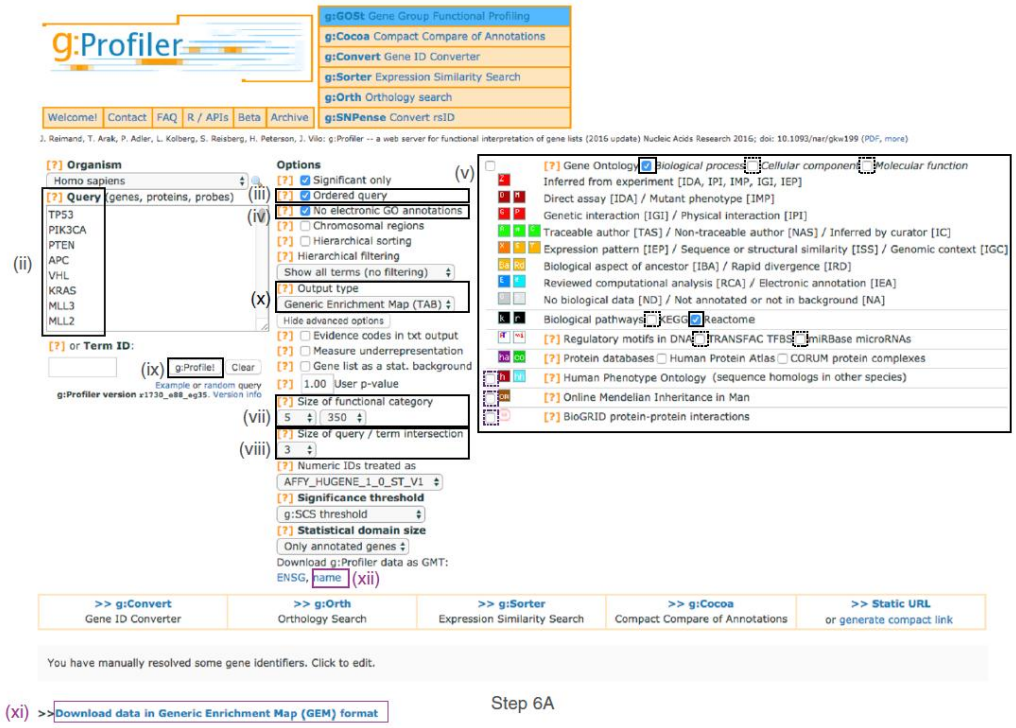
Pathway enrichment analysis ● Timing 3–20 min

Two major types of gene lists are used in pathway enrichment analysis of omics data. Flat (unranked) gene lists of dozens to thousands of genes can be analyzed using g:Profiler (option A). A statistical threshold is required to compile a gene list from omics data.

By contrast, ranked, whole-genome gene lists are suitable for pathway enrichment analysis using GSEA (option B). Gene lists analyzed with GSEA do not require prior filtering using statistical thresholds. Partial, filtered ranked gene lists can also be analyzed with g:Profiler. Select Step 6A or 6B, depending on the type of gene list you have. (A) Pathway enrich

Stéphane Nemours-Fisiopatología Renal-CIBBIM-Nanomedicina

(A) Pathway enrichment analysis of a gene list using g:Profiler ● Timing 3 min



(iii) Check the box next to *Ordered query*. This option treats the input as an ordered gene list and prioritizes genes with higher mutation ESs at the beginning of the list.

(iv) (Optional) Check the box next to *No electronic GO annotations*. This option will discard less reliable GO annotations (IEAs) that are not manually reviewed.

(v) Set filters on gene annotation data using the menu on the right. We recommend that initial pathway enrichment analyses includes only biological processes (BPs) of GO and molecular pathways of Reactome. Keep the two checkboxes checked and uncheck all other boxes in the menu.

(vi) Click on *Show Advanced Options* to set additional parameters.

(vii) Set the values of *Size of functional category* in the dropdown menu to 5 ('min') and 350 ('max'). Large pathways are of limited interpretative value, whereas numerous small pathways decrease the statistical power because of excessive multiple testing.

(viii) Set the *Size of query/term intersection* in the dropdown menu to 3. The analysis will consider only more reliable pathways that have three or more genes in the input gene list.

(ix) Click *g:Profile!* to run the analysis. A graphical heat map image will be shown, with detected pathways shown along the *y* axis (left) and associated genes of the input list shown along the *x* axis (top). Resulting pathways are organized hierarchically into related groups. g:Profiler uses graphical output by default and switches to textual output when a large number of pathways is found. g:Profiler returns only statistically significant pathways with *P* values adjusted for multiple testing correction (called *Q* values). By default, results with *Q* values <0.05 are reported. g:Profiler reports unrecognized and ambiguous gene IDs that can be resolved manually.

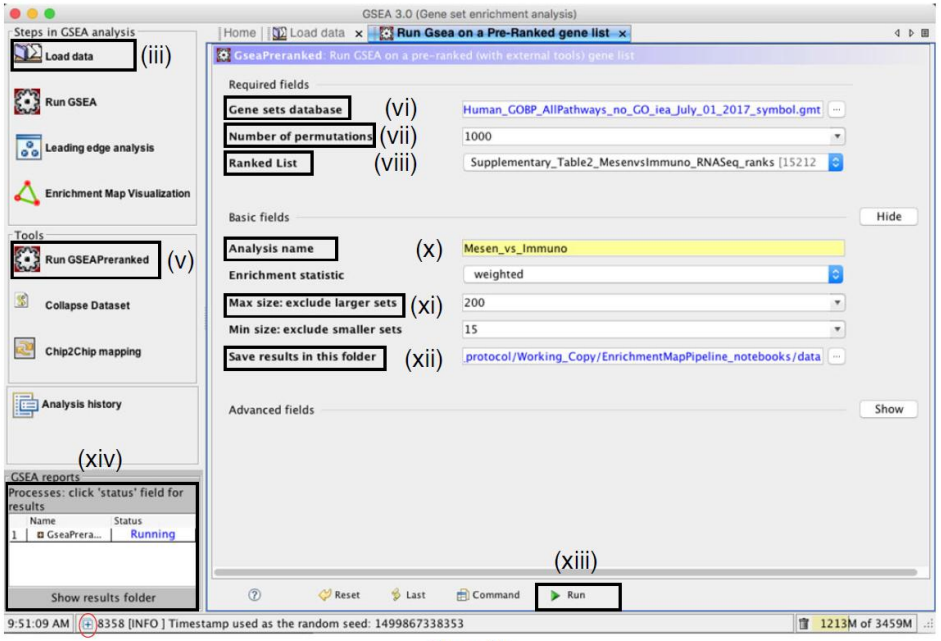
(x) Use the dropdown menu *Output type* and select the option *Generic Enrichment Map (TAB)*. This file is required for visualizing pathway results with Cytoscape and EnrichmentMap.

(xi) Click *g:Profile!* again to run the analysis with the updated parameters. The required link *Download data in Generic Enrichment Map (GEM) format* will appear under the g:Profiler interface. Download the file from the link and save it on your computer in your *result data folder* created in Step 1. Example results are provided in Supplementary Table 4.

(xii) Download the required GMT file by clicking on the link *name* at the bottom of the *Advanced Options* form. The GMT file is a compressed ZIP archive that contains all gene sets used by g:Profiler (e.g., gprofiler_hsapiens.NAME.gmt.zip). The gene set files are divided by data source. Download and uncompress the ZIP archive to your project folder. All required gene sets for this analysis will be in the file hsapiens.pathways.Name.gmt (Supplementary_Table5_hsapiens.pathways.NAME.gmt). Place the saved file in your *result data folder* created in Step 1.

Stéphane Nemours-Fisiopatología Renal-CIBBIM-Nanomedicina

(B) Pathway enrichment analysis of a ranked gene list using GSEA ● Timing ~20 min



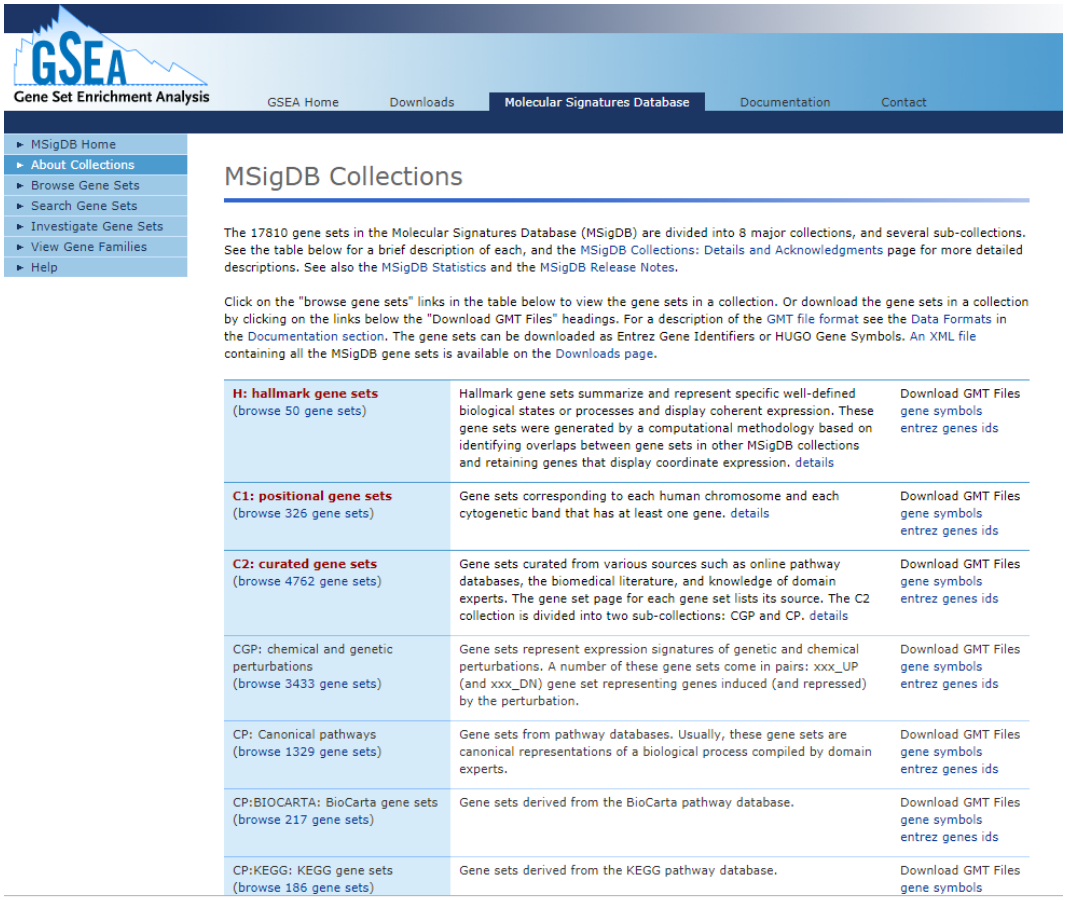
Step 6B

(i) Launch GSEA by opening the downloaded GSEA file (gsea.jnlp) (Fig. 3).
? TROUBLESHOOTING
Loading the required data files into GSEA
(ii) Click on *Load Data* in the top left corner of the *Steps in GSEA Analysis* section.
? TROUBLESHOOTING
(iii) In the *Load Data* tab, click on *Browse for files* ...
(iv) Find your *project data folder* and select the Supplementary_Table2_MesenvsImmuno_RNASeq_ranks.rnk file. Also select the pathway gene set definition (GMT) file using a multiple-select method such as shift-click (Supplementary Table 3). Click the *Choose* button to continue. A message box indicates that the files were loaded successfully. Click the *OK* button to continue.
▲ CRITICAL STEP GSEA also supplies its own gene set files, which are accessible directly through the GSEA interface from the MSigDB resource[80,81]. These files do not need to be imported into GSEA. To define the GMT file, find the MSigDB gene set files in the first tab, *Gene Matrix (from website)*, of the *Select one or more genesets* dialog. The latest versions of the MSigDB gene set files are shown in bold, but the earlier versions can also be accessed. To select multiple gene set files, click on the desired files while holding the control key in Windows or the command key in macOS.
(v) Click on *Run GSEAPreranked* in the side bar under *Tools*. The *Run GSEA on a Pre-Ranked gene list* tab will appear.
? TROUBLESHOOTING

g:profiler GMT files

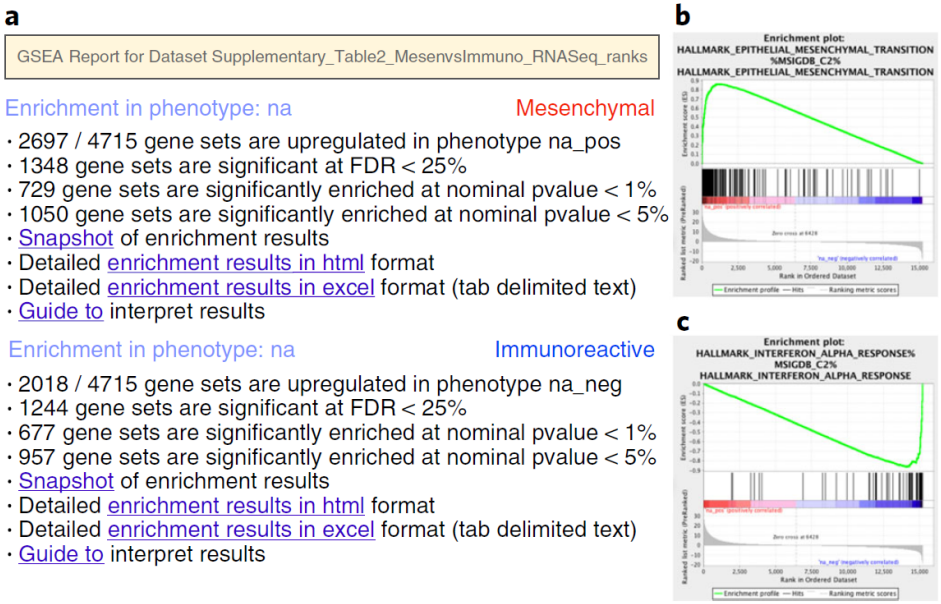
Stéphane Nemours-Fisiopatología Renal-CIBBIM-Nanomedicina

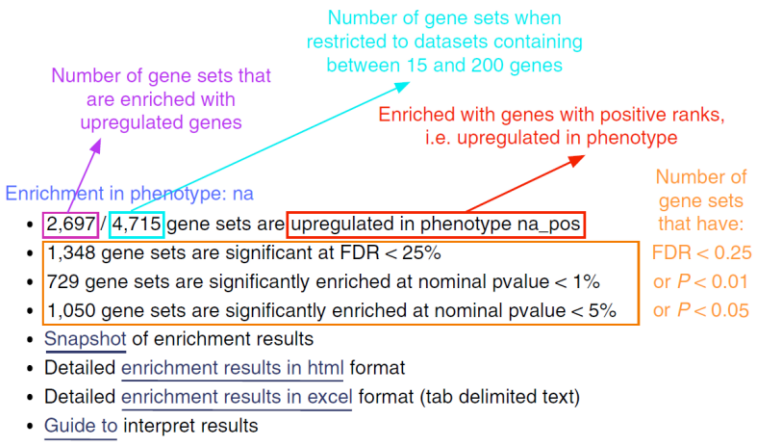
GSEA/ GMT files



Stéphane Nemours-Fisiopatología Renal-CIBBIM-Nanomedicina

Examination of GSEA results



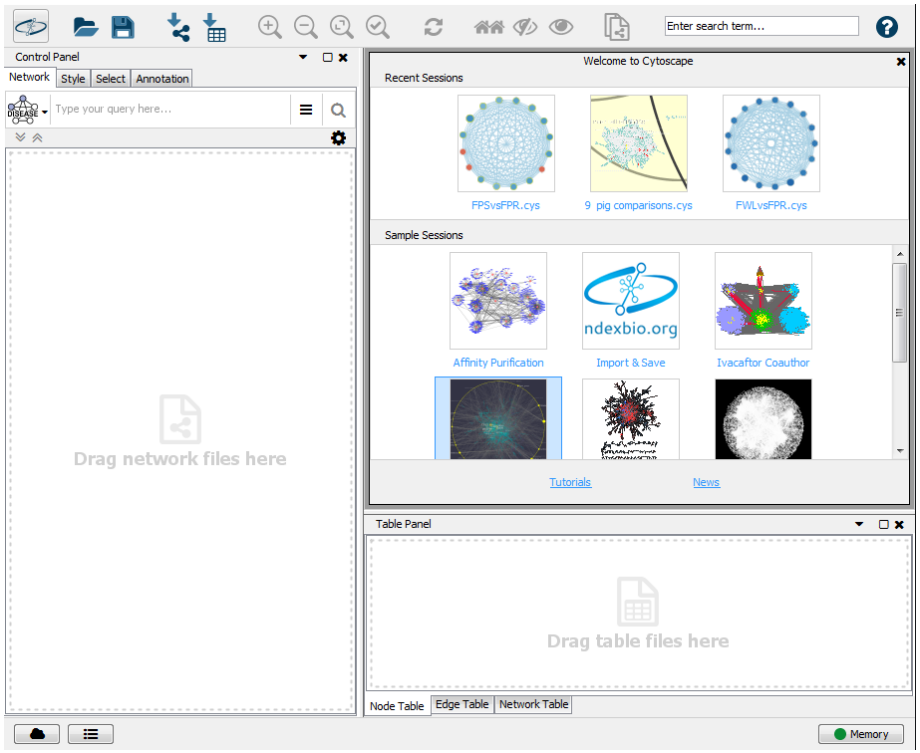


In the web browser results summary, click on Detailed enrichment results in HTML format and use the row numbering to check the number of pathways that have FDR Q values <0.05 to determine appropriate thresholds for EnrichmentMap in the next step of the protocol. If no pathways are reported at Q < 0.05, more lenient thresholds such as Q < 0.1 or Q < 0.25 could be used (Fig. 5). The threshold Q < 0.25 provides very lenient filtering, and it is not uncommon to find thousands of enriched pathways at this level. Robust analyses should use a cutoff of Q < 0.05 or lower. Filtering only by uncorrected P values is inappropriate and not recommended.

Stéphane Nemours-Fisiopatología Renal-CIBBIM-Nanomedicina

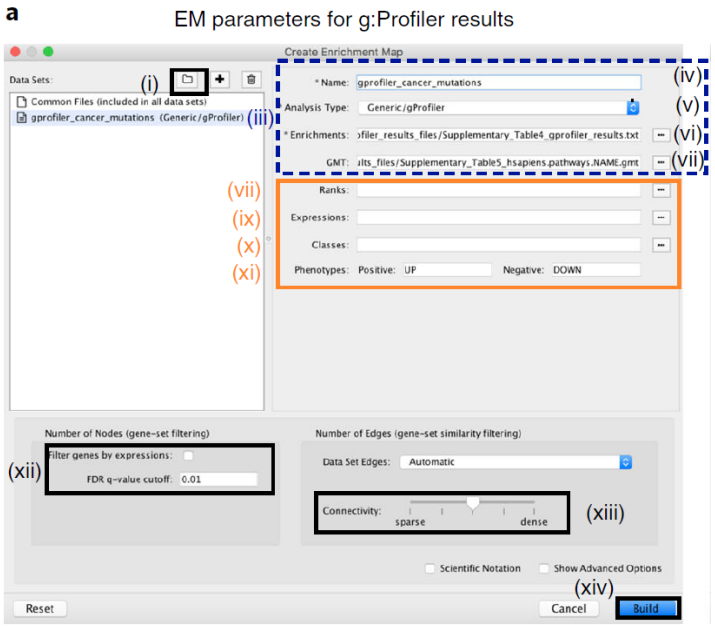
Visualization of enrichment results with EnrichmentMap ● Timing ~5 min

Launch the Cytoscape software



Stéphane Nemours-Fisiopatología Renal-CIBBIM-Nanomedicina

Creation of enrichment maps for g:Profiler results

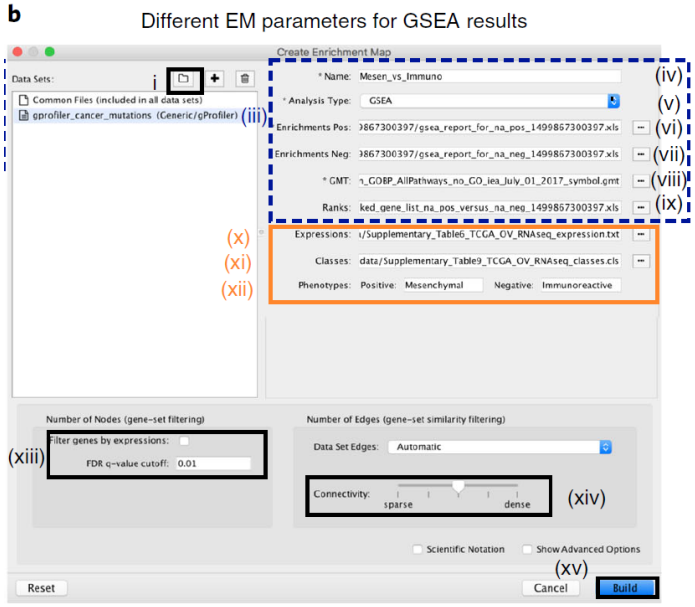


Step 9A

Number of Nodes. By default, g:Profiler returns only statistically significant results (Q < 0.05), so the FDR q-value cutoff parameter can be set to 1 in the EnrichmentMap Input panel, unless a more stringent filtering is desired. For this protocol, set the FDR Q value to 0.01. (Optional)

Stéphane Nemours-Fisiopatología Renal-CIBBIM-Nanomedicina

Creation of enrichment maps from GSEA results generated in Step 6B

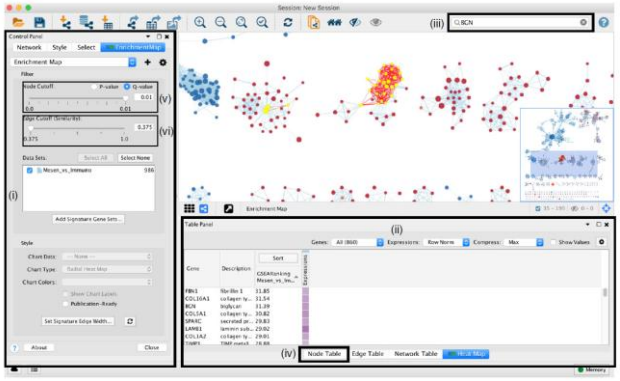


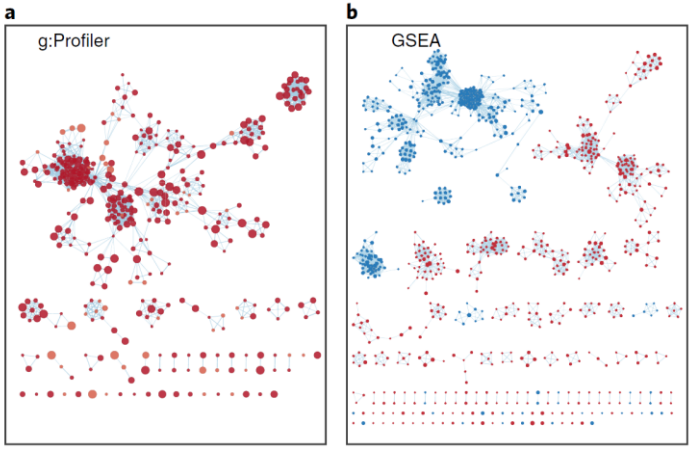
Step 9B

Number of Nodes. Set the FDR Q value cutoff to 0.01. (Optional) Select Filter genes by expressions to exclude any genes in the gene set definition file (i.e., the GMT file) that are not found in the supplied expression file.

Stéphane Nemours-Fisiopatología Renal-CIBBIM-Nanomedicina

Creation of enrichment maps

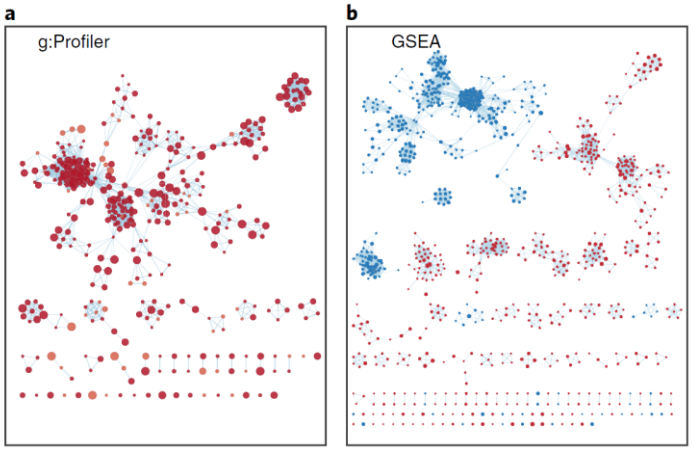




Pathways are shown as circles (nodes) that are connected with lines (edges) if the pathways share many genes. Nodes are colored by ES, and edges are sized on the basis of the number of genes shared by the connected pathways. Network layout and clustering algorithms automatically group similar pathways into major biological themes.

Stéphane Nemours-Fisiopatología Renal-CIBBIM-Nanomedicina

Navigation and interpretation of the enrichment map ● Timing ~4 h



(A) Exploring the Table Panel heat map ● Timing 45 min

(B) Organization and clarification of the network ● Timing 30 min

(C) Defining major biological themes ● Timing 2.5 h

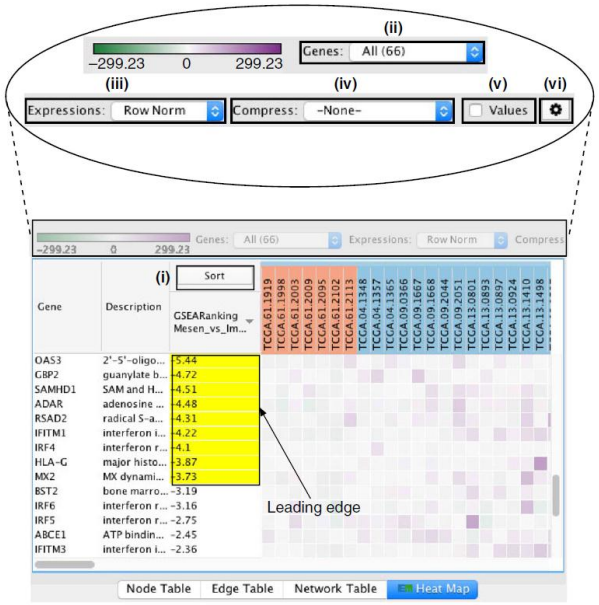
(D) Creation of a simplified network view ● Timing 15 min

(E) Manual arrangement of network nodes and updating of theme labels ● Timing 45 min

(F) Creation of a subnetwork that highlights a specific theme subset ● Timing 10 min

Stéphane Nemours-Fisiopatología Renal-CIBBIM-Nanomedicina

(A) Exploring the Table Panel heat map ● Timing 45 min



When a gene expression matrix is provided as input to EnrichmentMap, we can study the expression pattern of the genes included in enriched pathways.

If the analysis is based on GSEA results and a rank file is supplied, the ‘leading edge’ genes will be highlighted in yellow for individual node selections. Several options for heat map visualization are available.

Stéphane Nemours-Fisiopatología Renal-CIBBIM-Nanomedicina

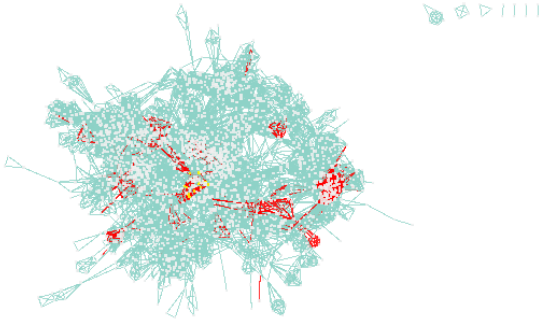
(B) Organization and clarification of the network ● Timing 30 min

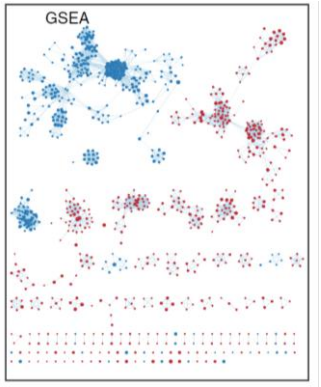
(i) If the network has too many nodes, go to the EnrichmentMap tab in the Control Panel and use the Node Cutoff Q-value threshold slider. Adjusting to a numerical value closer to 0 will remove less significant nodes (Fig. 8 (v)).

(ii) If the network is too interconnected, go to the EnrichmentMap tab in the Control Panel and increase the Edge Cutoff (Similarity) threshold; this will remove connections between less related nodes (Fig. 8 (vi)).

(iii) Apply the network layout again after adjusting the cutoffs (see the Layout menu in Cytoscape). The default layout algorithm is the unweighted Prefuse Force Directed layout.

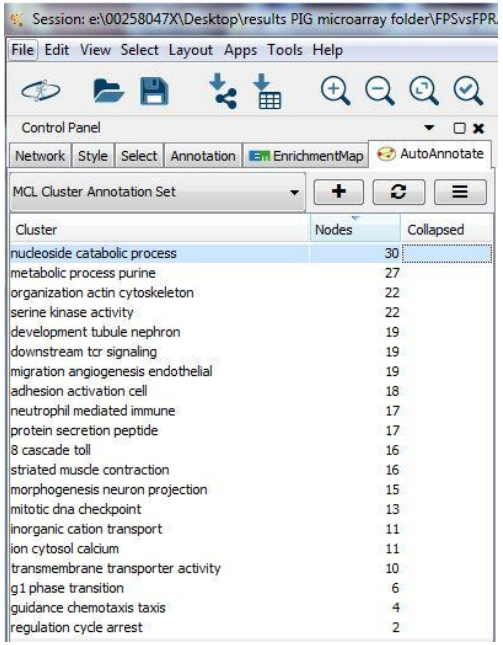
We also recommend that the prefuse force-directed layouts be weighted using the gene set similarity coefficient. Alternative layout algorithms are available and we encourage experimentation with them.

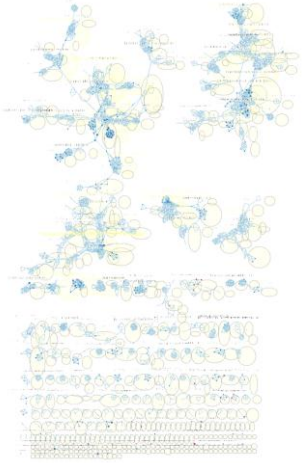




Stéphane Nemours-Fisiopatología Renal-CIBBIM-Nanomedicina

(C) Defining major biological themes ● Timing 2.5 h

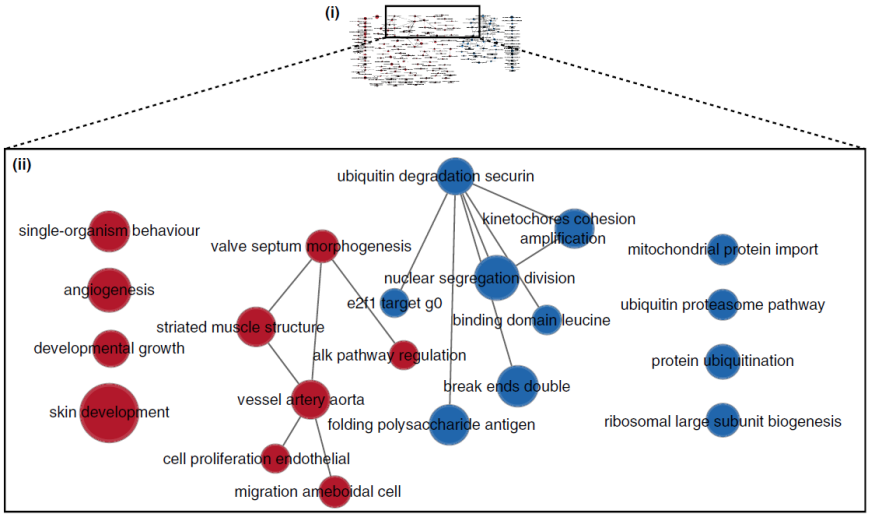




Enrichment maps typically include clusters of similar pathways representing major biological themes. Clusters can be automatically defined and summarized using the AutoAnnotate Cytoscape application. AutoAnnotate first clusters the network using the clusterMaker2 application and then summarizes each cluster on the basis of word frequency within the pathway names via the WordCloud app.

Stéphane Nemours-Fisiopatología Renal-CIBBIM-Nanomedicina

(D) Creation of a simplified network view ● Timing 15 min

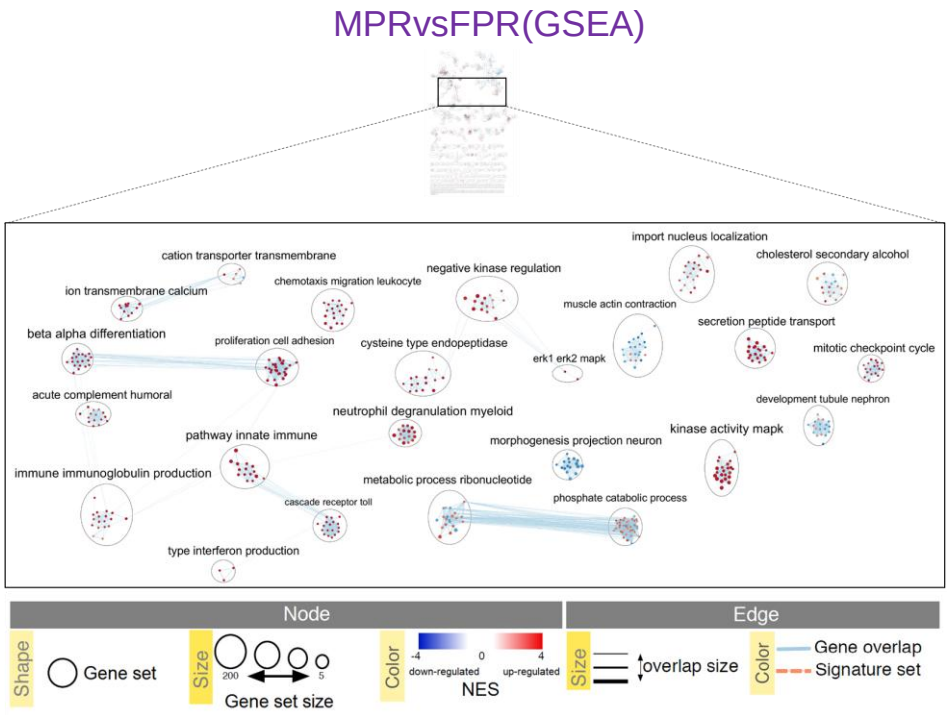


This creates a single group node for each cluster with a summarized name and provides an overview of the enrichment result themes that is useful for enrichment maps containing many nodes

Stéphane Nemours-Fisiopatología Renal-CIBBIM-Nanomedicina

(E) Manual arrangement of network nodes and updating of theme labels

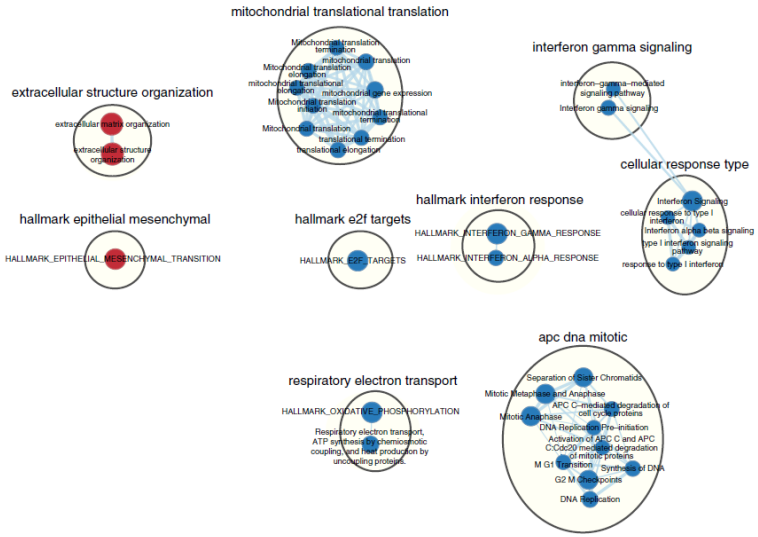
● Timing 45 min



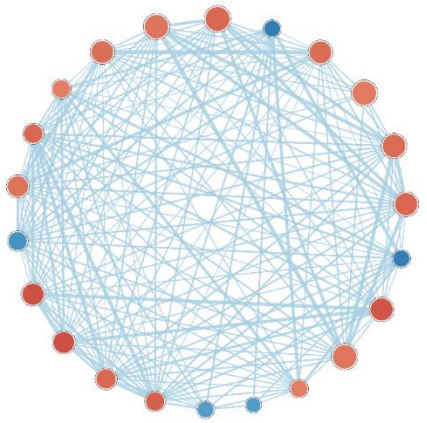
This section is required for the clearest network view and for a publication quality figure. For instance, it is useful to bring together similar themes, such as signaling or metabolic pathways, even if they are not connected in the map. Use of space should be optimized so that large amounts of white space are not present. This is a time-consuming step, but the more effort spent, the higher the quality of the resulting figure will be.

Stéphane Nemours-Fisiopatología Renal-CIBBIM-Nanomedicina

(F) Creation of a subnetwork that highlights a specific theme subset ● Timing 10 min

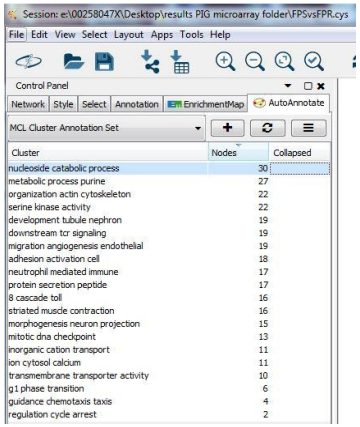


Enrichment maps of rich omics datasets are often large and complicated, and it is often useful to emphasize specific themes or relevant pathways in a final figure.



Stéphane Nemours-Fisiopatología Renal-CIBBIM-Nanomedicina

Alternative visualization of subnetworks (clusters)



SUID	__mclCluster	EnrichmentMap::Colouring (MPSvsMPR.GseaPreranked)	EnrichmentMap::ES (MPSvsMPR.GseaPreranked)	EnrichmentMap::fdr_qvalue (MPSvsMPR.GseaPreranked)	EnrichmentMap::Formatted_name	EnrichmentMap::fwer_qvalue (MPSvsMPR.GseaPrera	EnrichmentMap::Genes	EnrichmentMap::GS_DESCR	EnrichmentMap::gs_size	EnrichmentMap::GS_Type	EnrichmentMap::Name	EnrichmentMap::NES (MPSvsMPR.GseaPreranked)	EnrichmentMap::pvalue (MPSvsMPR.GseaPreranked)	name	selected	shared name	
1426	14	0,7273	0,4352	0,6963	REAC:R-HSA-168176	1	NFKB1|TRAF6|MYD88|IRAK4|FOS	Toll Like Receptor 5 (TLR5) Cascade	15	ENR	REAC:R-HSA-168	1,1613	0,2727	REAC:R-HSA	true	REAC:R-HSA-168176	
1266	14	0,4294	0,3043	0,914	REAC:R-HSA-168138	1	NFKB1|TRAF6|MYD88|CD14|IRAK4	Toll Like Receptor 9 (TLR9) Cascade	23	ENR	REAC:R-HSA-168	0,915	0,5706	REAC:R-HSA	true	REAC:R-HSA-168138	
1331	14	0,5822	0,3625	0,824	REAC:R-HSA-168188	1	NFKB1|TRAF6|MYD88|CD14|IRAK4	Toll Like Receptor TLR6:TLR2 Cascade	19	ENR	REAC:R-HSA-168	1,0329	0,4178	REAC:R-HSA	true	REAC:R-HSA-168188	
1427	14	0,6097	0,3625	0,8247	REAC:R-HSA-168179	1	TRAF6|NFKB1|MYD88|CD14|IRAK4	Toll Like Receptor TLR1:TLR2 Cascade	19	ENR	REAC:R-HSA-168	1,036	0,3903	REAC:R-HSA	true	REAC:R-HSA-168179	
215	14	0,6311	0,3465	0,8127	REAC:R-HSA-166166	1	TRAF6|NFKB1|CD14|UBE2D2|CASP	MyD88-independent TLR4 cascade	26	ENR	REAC:R-HSA-166	1,0486	0,3689	REAC:R-HSA	true	REAC:R-HSA-166166	
858	14	0,5776	0,3625	0,8234	REAC:R-HSA-181438	1	NFKB1|TRAF6|MYD88|CD14|IRAK4	Toll Like Receptor 2 (TLR2) Cascade	19	ENR	REAC:R-HSA-181	1,0305	0,4224	REAC:R-HSA	true	REAC:R-HSA-181438	
1406	14	0,6349	0,3465	0,8049	REAC:R-HSA-168164	1	NFKB1|TRAF6|CD14|UBE2D2|CASP	Toll Like Receptor 3 (TLR3) Cascade	26	ENR	REAC:R-HSA-168	1,0712	0,3651	REAC:R-HSA	true	REAC:R-HSA-168164	
1475	14	0,4009	0,3043	0,9143	REAC:R-HSA-975155	1	TRAF6|NFKB1|MYD88|CD14|IRAK4	MyD88 dependent cascade initiated on endosome	23	ENR	REAC:R-HSA-975	0,9106	0,5991	REAC:R-HSA	true	REAC:R-HSA-975155	
547	14	0,6376	0,3465	0,8016	REAC:R-HSA-937061	1	NFKB1|TRAF6|CD14|UBE2D2|CASP	TRIF(TICAM1)-mediated TLR4 signaling	26	ENR	REAC:R-HSA-937	1,0695	0,3624	REAC:R-HSA	true	REAC:R-HSA-937061	
1383	14	0,7452	0,4352	0,6826	REAC:R-HSA-168142	1	TRAF6|NFKB1|MYD88|IRAK4|FOS	Toll Like Receptor 10 (TLR10) Cascade	15	ENR	REAC:R-HSA-168	1,1796	0,2548	REAC:R-HSA	true	REAC:R-HSA-168142	
1323	14	0,4079	0,3043	0,9184	REAC:R-HSA-168181	1	TRAF6|NFKB1|MYD88|CD14|IRAK4	Toll Like Receptor 7/8 (TLR7/8) Cascade	23	ENR	REAC:R-HSA-168	0,9075	0,5921	REAC:R-HSA	true	REAC:R-HSA-168181	
651	14	0,3067	0,2653	0,9305	REAC:R-HSA-168898	1	TRAF6|NFKB1|CD14|IRAK4|CTSK|L	Toll-Like Receptors Cascades	36	ENR	REAC:R-HSA-168	0,8689	0,6933	REAC:R-HSA	true	REAC:R-HSA-168898	
1516	14	0,419	0,3115	0,896	REAC:R-HSA-975138	1	NFKB1|TRAF6|MYD88|CD14|IRAK4	TRAF6 mediated induction of NFkB and MAP kinases upon	22	ENR	REAC:R-HSA-975	0,93	0,581	REAC:R-HSA	true	REAC:R-HSA-975138	
1708	14	0,6134	0,3625	0,8248	REAC:R-HSA-166058	1	TRAF6|NFKB1|MYD88|CD14|IRAK4	MyD88:Mal cascade initiated on plasma membrane	19	ENR	REAC:R-HSA-166	1,0409	0,3866	REAC:R-HSA	true	REAC:R-HSA-166058	
1772	14	0,5348	0,3092	0,855	REAC:R-HSA-166016	1	TRAF6|NFKB1|CD14|IRAK4|UBE2D	Toll Like Receptor 4 (TLR4) Cascade	31	ENR	REAC:R-HSA-166	0,9927	0,4652	REAC:R-HSA	true	REAC:R-HSA-166016	
973	14	0,7171	0,4352	0,6964	REAC:R-HSA-975871	1	TRAF6|NFKB1|MYD88|IRAK4|MEF2	MyD88 cascade initiated on plasma membrane	15	ENR	REAC:R-HSA-975	1,1591	0,2829	REAC:R-HSA	true	REAC:R-HSA-975871	
1038	14	-0,8649	-0,3742	0,7912	REAC:R-HSA-168928	1	TRAF6|NFKB1|MYD88|IKBKE|UBE2	DDX58/IFIH1-mediated induction of interferon-alpha/beta	23	ENR	REAC:R-HSA-168	-1,2731	0,1351	REAC:R-HSA	true	REAC:R-HSA-168928	

Stéphane Nemours-Fisiopatología Renal-CIBBIM-Nanomedicina