3000788 Intro to Comp Molec Biol

Lecture 12: Functional enrichment analysis

Fall 2025

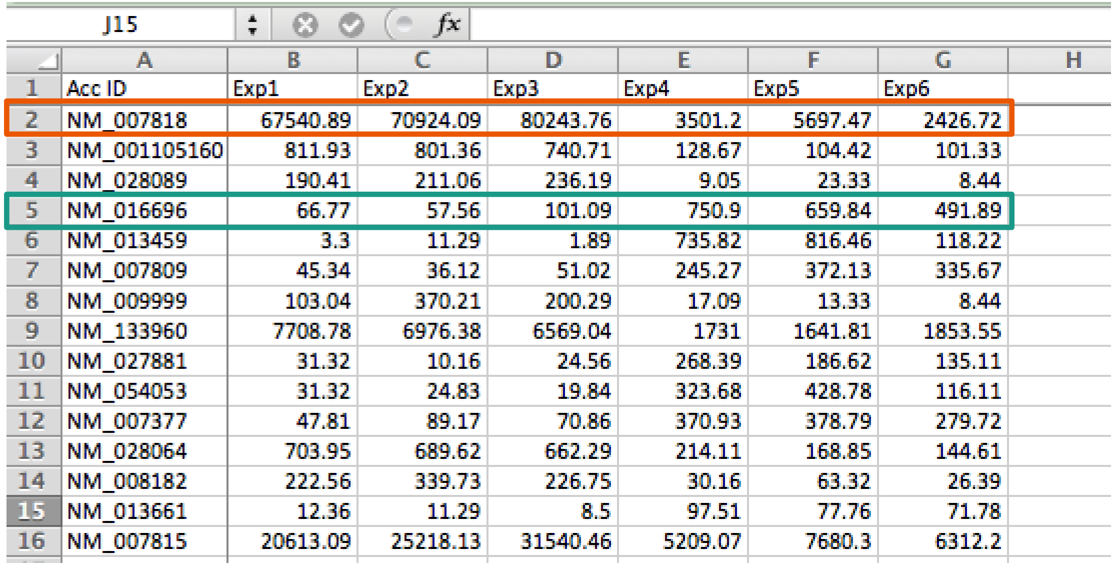
- Research Affairs
- Center of Excellence in Computational Molecular Biology (CMB)
- Center for Artificial Intelligence in Medicine (CU-AIM)

Today’s agenda

- From differential expression to functional enrichment
- Gene annotation systems
- Overrepresentation
- Gene Set Enrichment Analysis (GSEA)
- Network-analysis and permutation test

DEG is univariate

Control Treatment



	A	B	C	D	E	F	G	H
1	Acc ID	Exp1	Exp2	Exp3	Exp4	Exp5	Exp6	
2	NM_007818	67540.89	70924.09	80243.76	3501.2	5697.47	2426.72	
3	NM_001105160	811.93	801.36	740.71	128.67	104.42	101.33	
4	NM_028089	190.41	211.06	236.19	9.05	23.33	8.44	
5	NM_016696	66.77	57.56	101.09	750.9	659.84	491.89	
6	NM_013459	3.3	11.29	1.89	735.82	816.46	118.22	
7	NM_007809	45.34	36.12	51.02	245.27	372.13	335.67	
8	NM_009999	103.04	370.21	200.29	17.09	13.33	8.44	
9	NM_133960	7708.78	6976.38	6569.04	1731	1641.81	1853.55	
10	NM_027881	31.32	10.16	24.56	268.39	186.62	135.11	
11	NM_054053	31.32	24.83	19.84	323.68	428.78	116.11	
12	NM_007377	47.81	89.17	70.86	370.93	378.79	279.72	
13	NM_028064	703.95	689.62	662.29	214.11	168.85	144.61	
14	NM_008182	222.56	339.73	226.75	30.16	63.32	26.39	
15	NM_013661	12.36	11.29	8.5	97.51	77.76	71.78	
16	NM_007815	20613.09	25218.13	31540.46	5209.07	7680.3	6312.2	

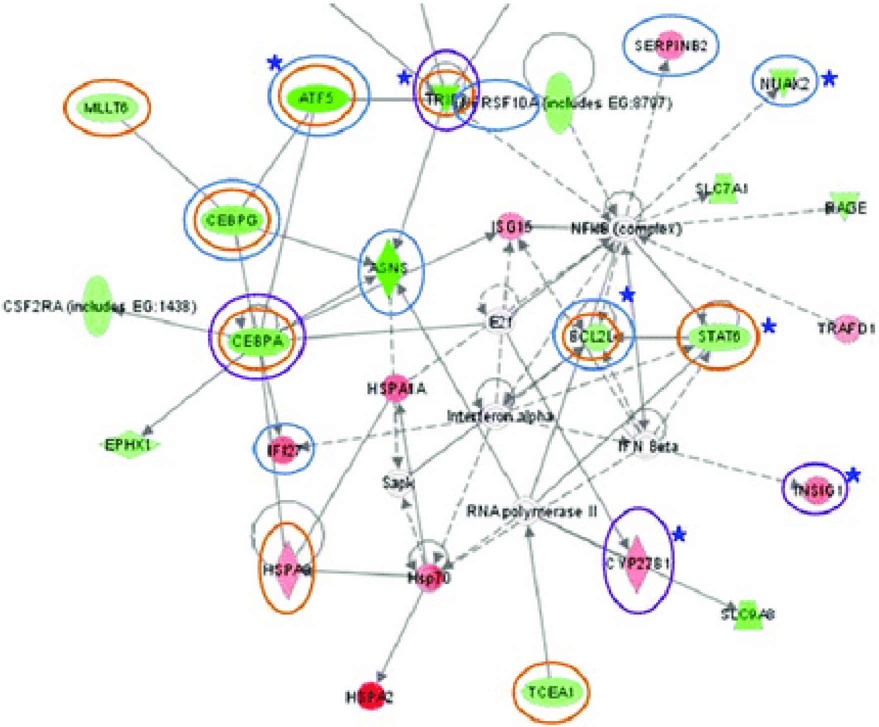
http://homer.ucsd.edu/homer/basicTutorial/clustering.html

- Each gene is tested separately

- But treatment likely **affects a group of genes,** such as those from the same pathway

- No simple multivariate model for gene expression!

Functional enrichment as pseudo-multivariate analysis



- Observing more differentially expressed genes (DEG) from the same pathway → higher confidence

- Singleton DEGs could be noises

Dubey, R. et al. PLoS ONE 6:e28509 (2011)

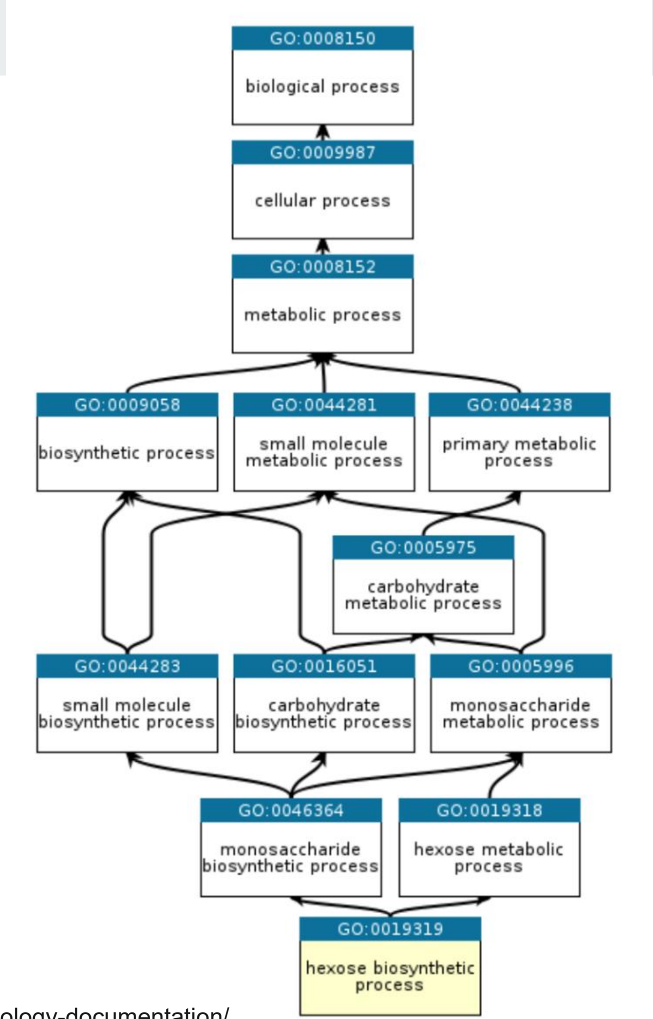
Gene annotation systems

Gene Ontology (GO)

- Three aspects
  - **Biological process**: Broad biological programs
    Ex: DNA repair, Cytosine biosynthesis
  - **Molecular function**: Molecular-level activities
    Ex: Catalytic, Insulin receptor
  - **Cellular component**: Cellular locations
    Ex: Plasma membrane, mitochondrion

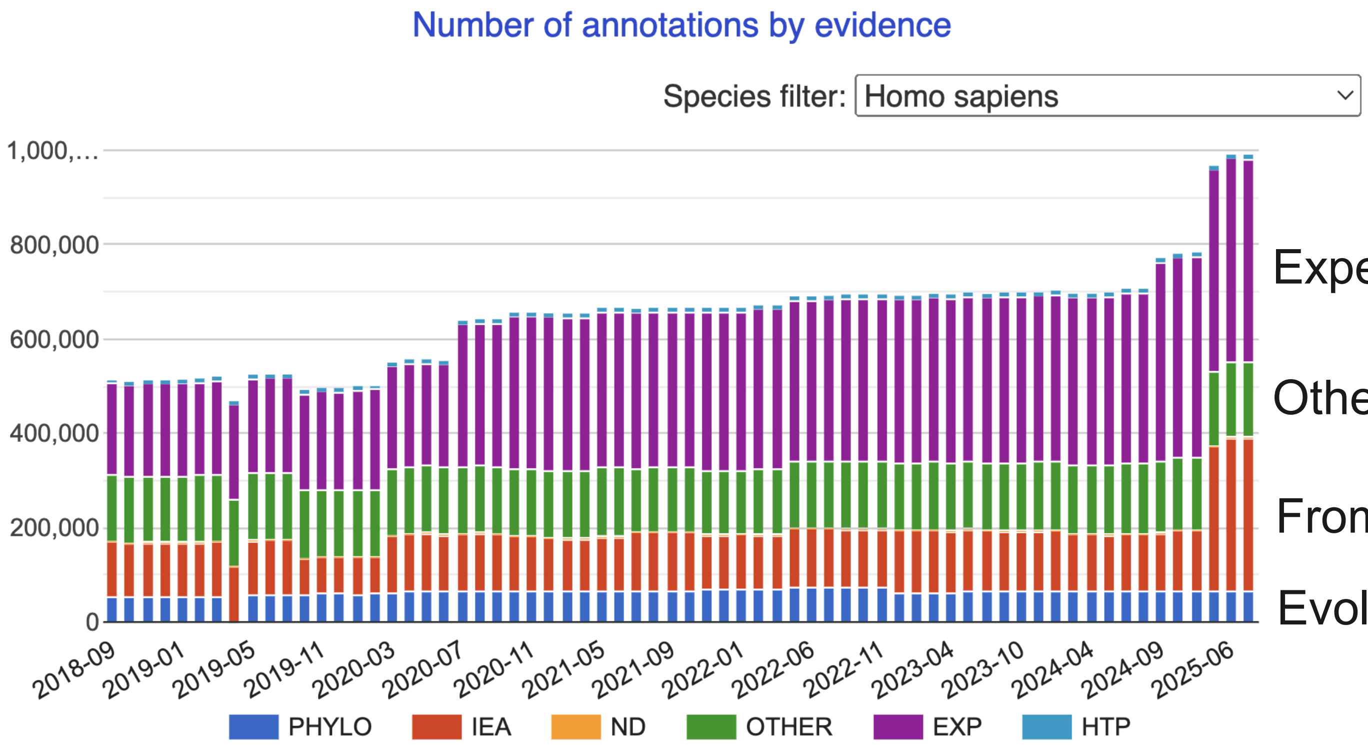
- Tree hierarchy, from broad to detailed

- GO:XXXXXX accession format


GO:0008150
biological process
GO:0009987
cellular process
GO:0008152
metabolic process
GO:0009058
biosynthetic process
GO:0044281
small molecule metabolic process
GO:0044238
primary metabolic process
GO:0005975
carbohydrate metabolic process
GO:0044283
small molecule biosynthetic process
GO:0016051
carbohydrate biosynthetic process
GO:0005996
monosaccharide metabolic process
GO:0046364
monosaccharide biosynthetic process
GO:0019318
hexose metabolic process
GO:0019319
hexose biosynthetic process

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

GO annotations for human genes



Number of annotations by evidence

Species filter: Homo sapiens

Experiment

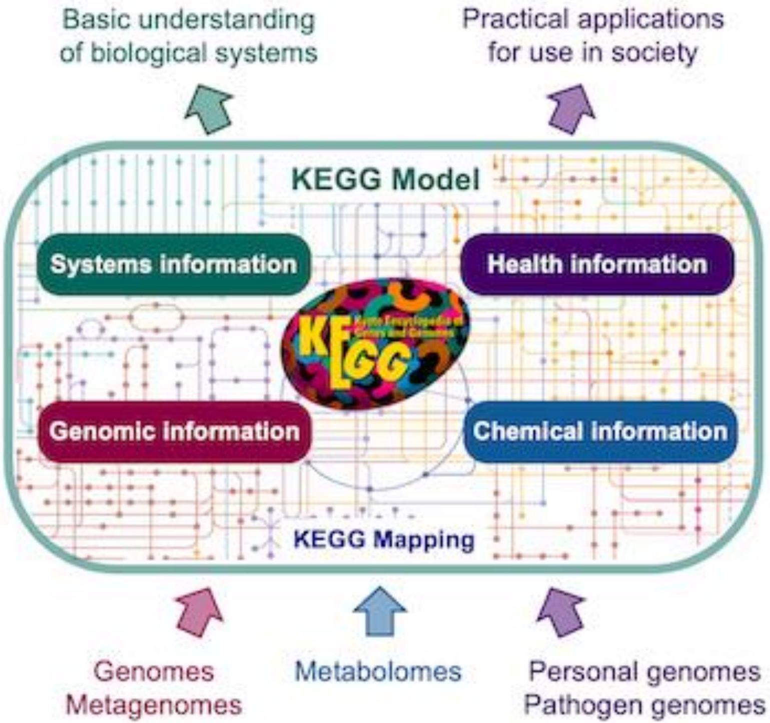
Other

From other DB

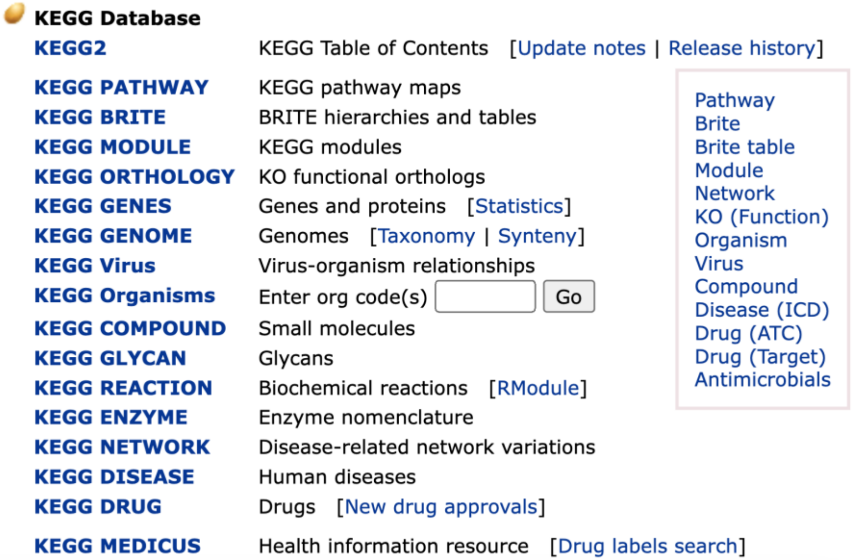
Evolutionary

https://geneontology.org/stats.html

Kyoto Encyclopedia of Genes and Genomes (KEGG)


Basic understanding of biological systems
Practical applications for use in society
KEGG Model
Systems information
Health information
KEGG Kyoto Encyclopedia of Genes and Genomes
Genomic information
Chemical information
KEGG Mapping
Genomes Metagenomes
Metabolomes
Personal genomes Pathogen genomes

- A collection of relationships between genes, proteins, metabolites, etc.

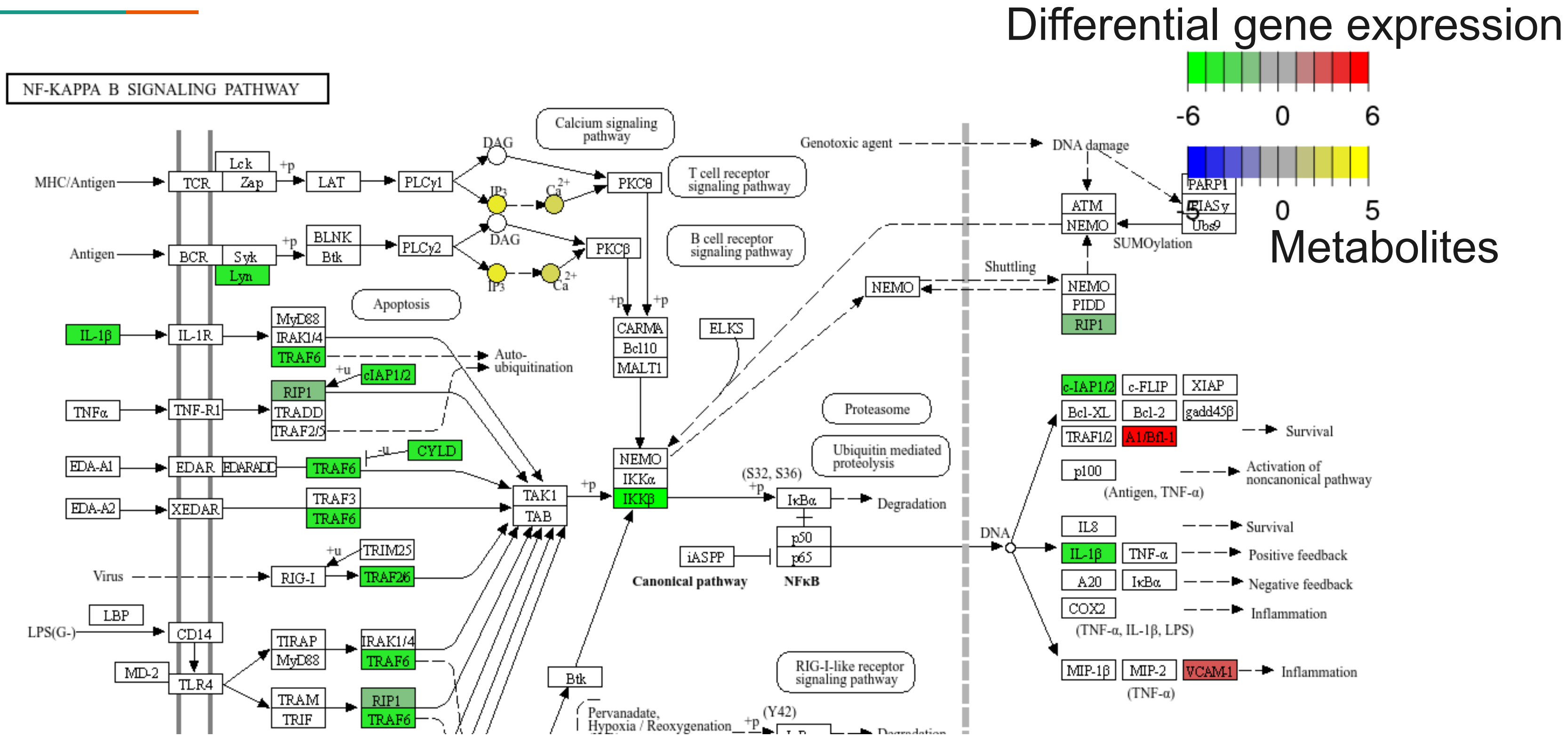


KEGG Database

KEGG2	KEGG Table of Contents [Update notes | Release history]
KEGG PATHWAY	KEGG pathway maps
KEGG BRITE	BRITE hierarchies and tables
KEGG MODULE	KEGG modules
KEGG ORTHOLOGY	KO functional orthologs
KEGG GENES	Genes and proteins [Statistics]
KEGG GENOME	Genomes [Taxonomy | Synteny]
KEGG Virus	Virus-organism relationships
KEGG Organisms	Enter org code(s) [ ] Go
KEGG COMPOUND	Small molecules
KEGG GLYCAN	Glycans
KEGG REACTION	Biochemical reactions [RModule]
KEGG ENZYME	Enzyme nomenclature
KEGG NETWORK	Disease-related network variations
KEGG DISEASE	Human diseases
KEGG DRUG	Drugs [New drug approvals]
KEGG MEDICUS	Health information resource [Drug labels search]

Pathway
Brite
Brite table
Module
Network
KO (Function)
Organism
Virus
Compound
Disease (ICD)
Drug (ATC)
Drug (Target)
Antimicrobials

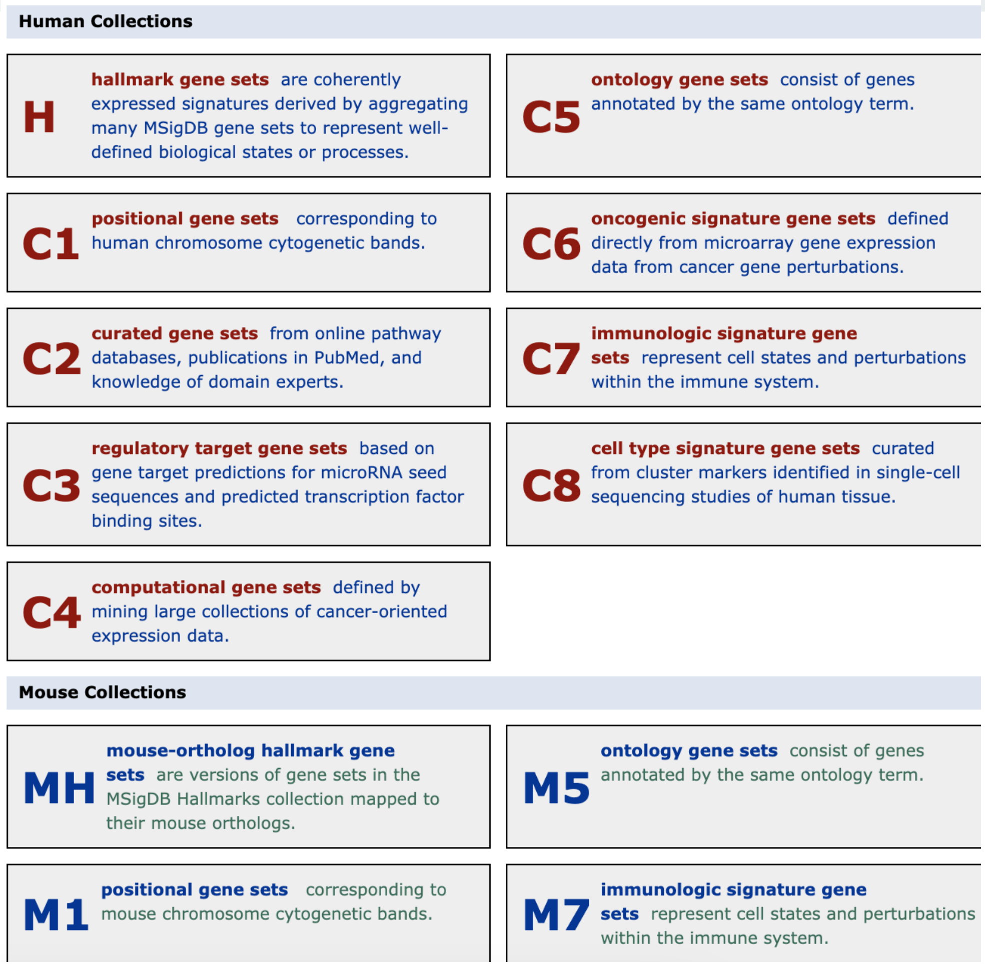
KEGG pathway visualization with omics data



Tang, D. et al. PLoS ONE 18:e0294236 (2023)

Other databases

- WikiPathways
- Panther
- Reactome
- Biocarta
- **MSigDB**: A collection of many other databases (human and mouse only)



Human Collections

H **hallmark gene sets** are coherently expressed signatures derived by aggregating many MSigDB gene sets to represent well-defined biological states or processes.

C5 **ontology gene sets** consist of genes annotated by the same ontology term.

C1 **positional gene sets** corresponding to human chromosome cytogenetic bands.

C6 **oncogenic signature gene sets** defined directly from microarray gene expression data from cancer gene perturbations.

C2 **curated gene sets** from online pathway databases, publications in PubMed, and knowledge of domain experts.

C7 **immunologic signature gene sets** represent cell states and perturbations within the immune system.

C3 **regulatory target gene sets** based on gene target predictions for microRNA seed sequences and predicted transcription factor binding sites.

C8 **cell type signature gene sets** curated from cluster markers identified in single-cell sequencing studies of human tissue.

C4 **computational gene sets** defined by mining large collections of cancer-oriented expression data.

Mouse Collections

MH **mouse-ortholog hallmark gene sets** are versions of gene sets in the MSigDB Hallmarks collection mapped to their mouse orthologs.

M5 **ontology gene sets** consist of genes annotated by the same ontology term.

M1 **positional gene sets** corresponding to mouse chromosome cytogenetic bands.

M7 **immunologic signature gene sets** represent cell states and perturbations within the immune system.

Tips for choosing annotation systems

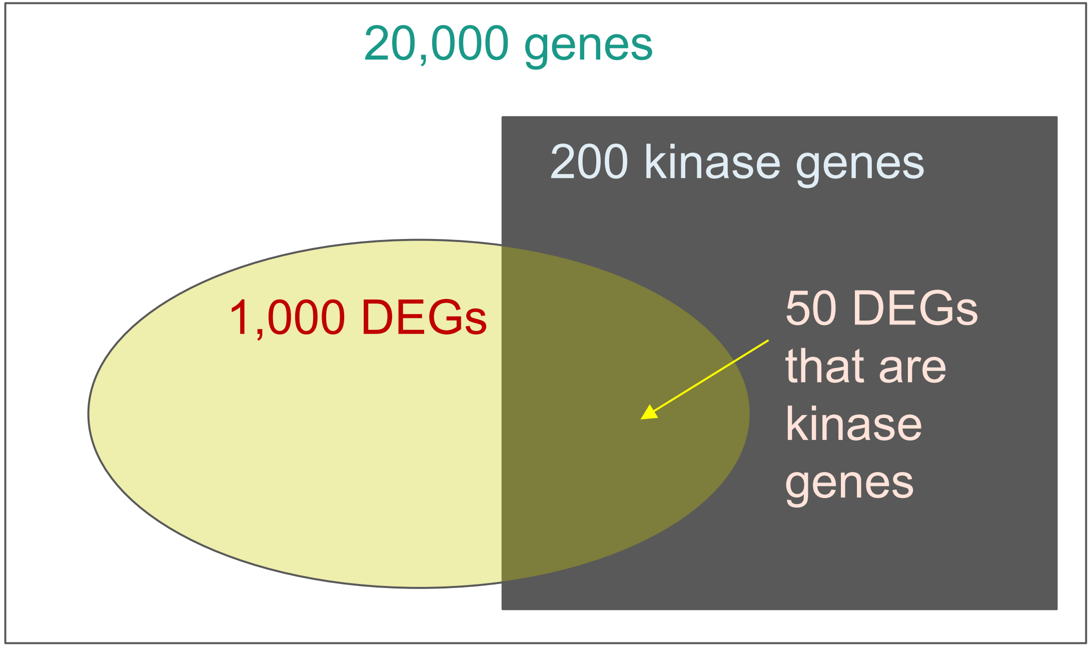
- Try multiple annotations

- If you have some prior knowledge, identify annotation systems containing your pathways, or terms of interest

- GO and KEGG are a good place to start

Overrepresentation analysis (ORA)

Hypergeometric distribution



- Probability of observing this data = $\frac{\binom{200}{50}\binom{20{,}000-200}{1{,}000-50}}{\binom{20{,}000}{1{,}000}}$
- P-value = probability of observing ≥50 DEGs being kinase genes

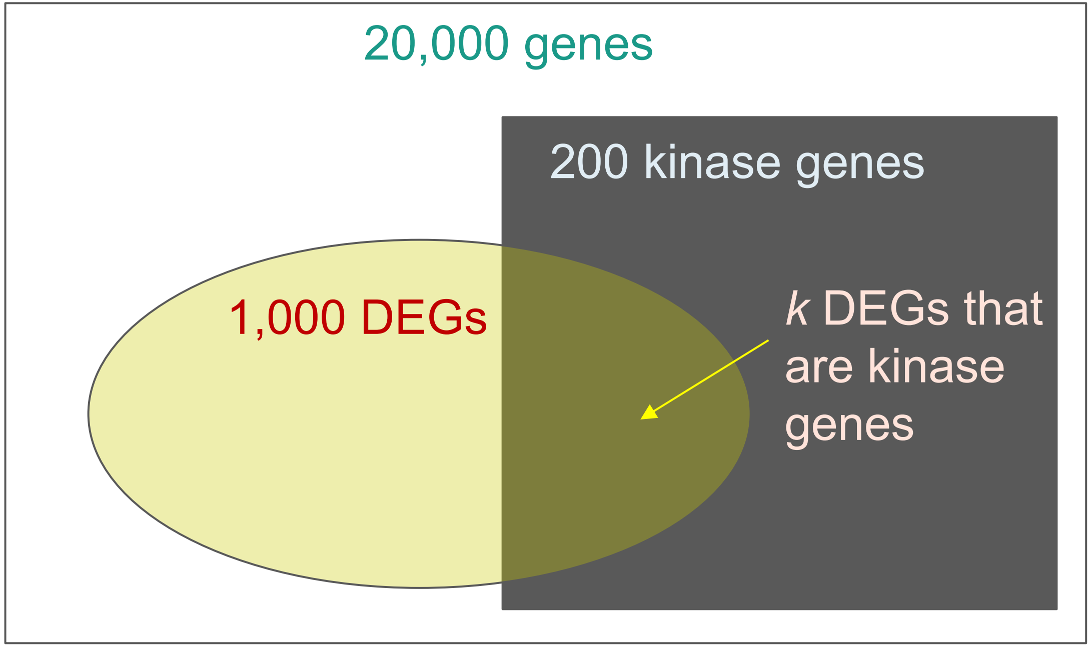
Calculating hypergeometric probability

- How large is $\frac{\binom{200}{50}\binom{20,000-200}{1,000-50}}{\binom{20,000}{1,000}}$? Less than 1.0 ☺
- How large is $\binom{200}{50}$? Larger than 10^{42} ☹

- The trick is to take log-transform of individual terms
  log(200!) = log(200 x 199 x … x 1) = log(200) + log(199) + … + log(1)

- Combine them, and take the exponential back at the very end

Functional enrichment is a one-tailed test



- Expected $k = \frac{1{,}000 \times 200}{20{,}000} = 10$
- If k > 10, p-value reflects enrichment of kinases
- If k < 10, there is no enrichment, and no need to test for negative enrichment

Fisher’s exact test

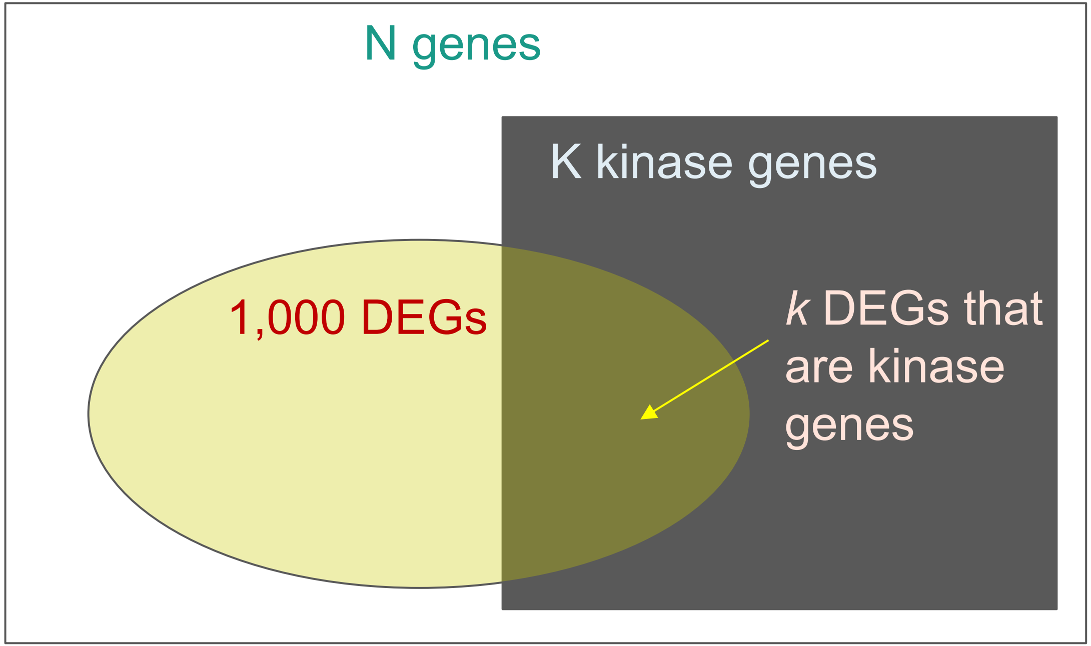
Gene group	Kinase	Not kinase	Total
Differentially expressed	$k \geq 50$	1,000 – k	1,000
Not differentially expressed	200 – k	18,800 + k	19,000
Total	200	19,800	20,000

- P-value follows hypergeometric distribution

ORA procedure

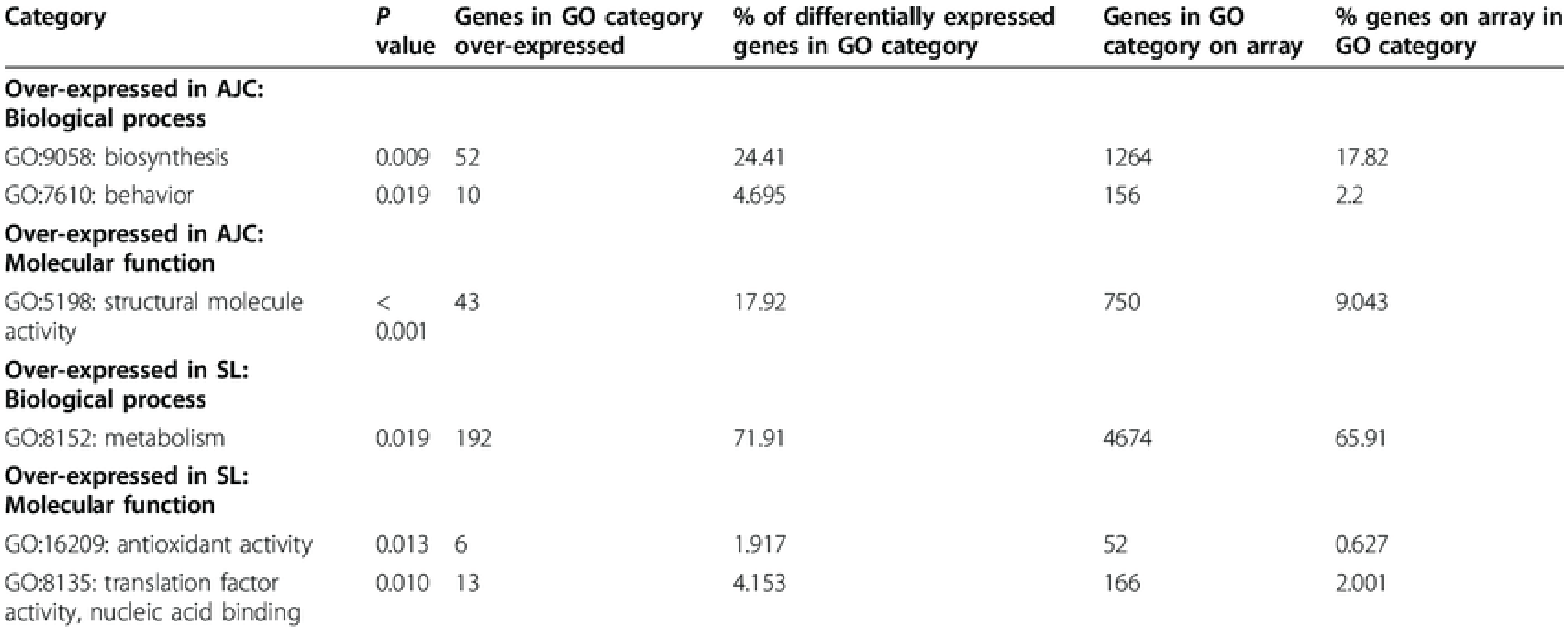
- Perform differential expression analysis
- Apply a p-value cutoff
- Analyze the set of genes that passed
- **Choices**: Analyze up- and down-regulated genes together or separately

Background gene set for functional enrichment



- What if you are using targeted techniques, such as Nanostring?
- Target 800 cancer genes
- All DEGs will be related to cancer!
- Set the background genes!

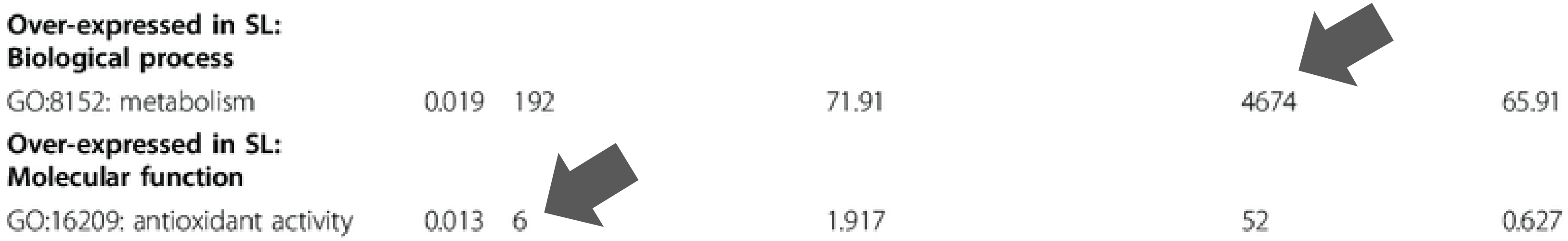
Examples of ORA result



Category	*P* value	Genes in GO category over-expressed	% of differentially expressed genes in GO category	Genes in GO category on array	% genes on array in GO category
Over-expressed in AJC: Biological process					
GO:9058: biosynthesis	0.009	52	24.41	1264	17.82
GO:7610: behavior	0.019	10	4.695	156	2.2
Over-expressed in AJC: Molecular function					
GO:5198: structural molecule activity	< 0.001	43	17.92	750	9.043
Over-expressed in SL: Biological process					
GO:8152: metabolism	0.019	192	71.91	4674	65.91
Over-expressed in SL: Molecular function					
GO:16209: antioxidant activity	0.013	6	1.917	52	0.627
GO:8135: translation factor activity, nucleic acid binding	0.010	13	4.153	166	2.001

Pavey, S.A. et al. BMC Ecology 11:31 (2011)

Filtering ORA enrichment result



Over-expressed in SL: Biological process					
GO:8152: metabolism	0.019	192	71.91	4674	65.91
Over-expressed in SL: Molecular function					
GO:16209: antioxidant activity	0.013	6	1.917	52	0.627

- Annotations that are too broad (GO:8152 metabolism) are not informative

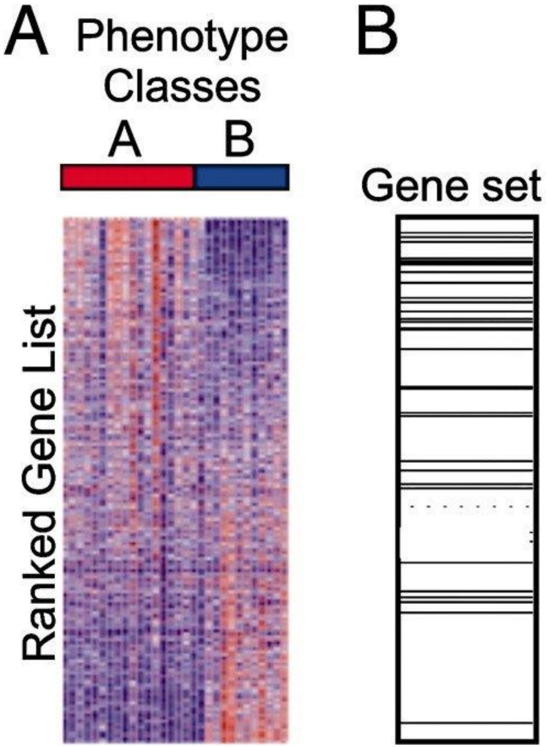
- Enrichment with very few genes (for rare annotations) may be significant just by chance

Tips for ORA

- Don’t change your p-value cutoff to get more or less genes into ORA
- If you have few DEGs, ORA may not be the right choice
- Think carefully about the background gene set of your study

Gene Set Enrichment Analysis (GSEA)

Motivation

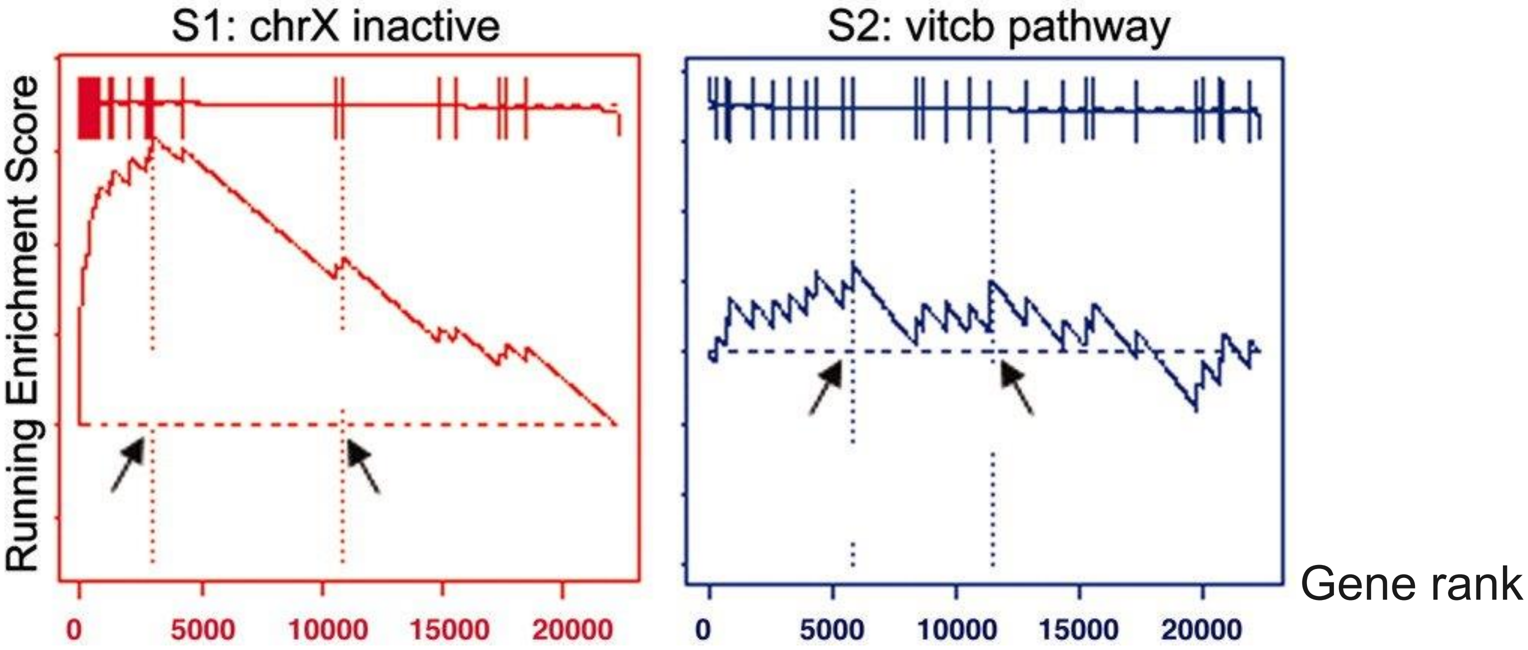


Subramanian *et al*. PNAS. 102:15545-15550 (2005)

- Don’t want to apply subjective p-value cutoff
- Rank genes by DEG scores, such as fold changes or p-values
- Pick a gene set (annotation term)
- Start from top to bottom: +score if found, – score otherwise
- **What patterns do we expect?**

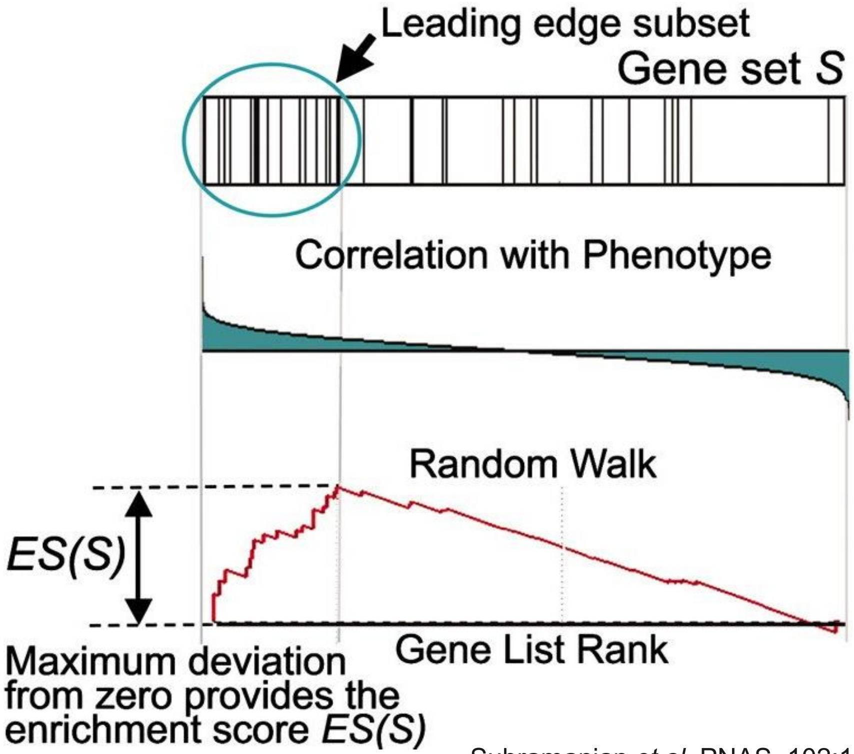
GSEA score examples

Subramanian *et al*. PNAS. 102:15545-15550 (2005)



- High cumulative scores if genes aggregate at the top
- Low cumulative scores if genes are scattered throughout

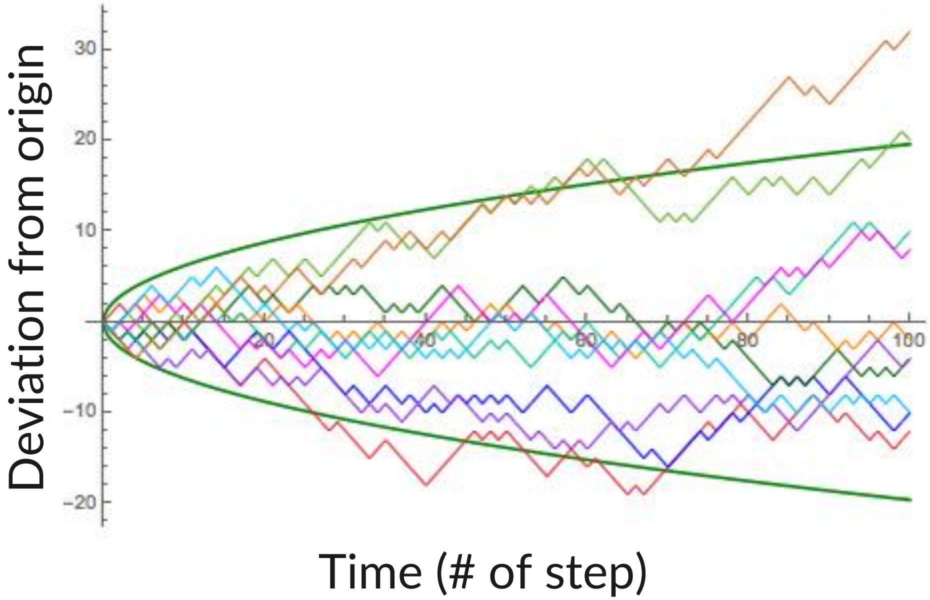
GSEA algorithm


Leading edge subset
Gene set *S*
Correlation with Phenotype
Random Walk
ES(S)
Gene List Rank
Maximum deviation from zero provides the enrichment score *ES(S)*

Subramanian *et al*. PNAS. 102:15545-15550 (2005)

- Calculate cumulative scores going from the top gene to bottom
- Record ES(S) the maximum scores deviation from zero
- Test the significant of ES(S)
- Positive ES(S) = up-regulated
- Negative ES(S) = down-regulated

Random walk



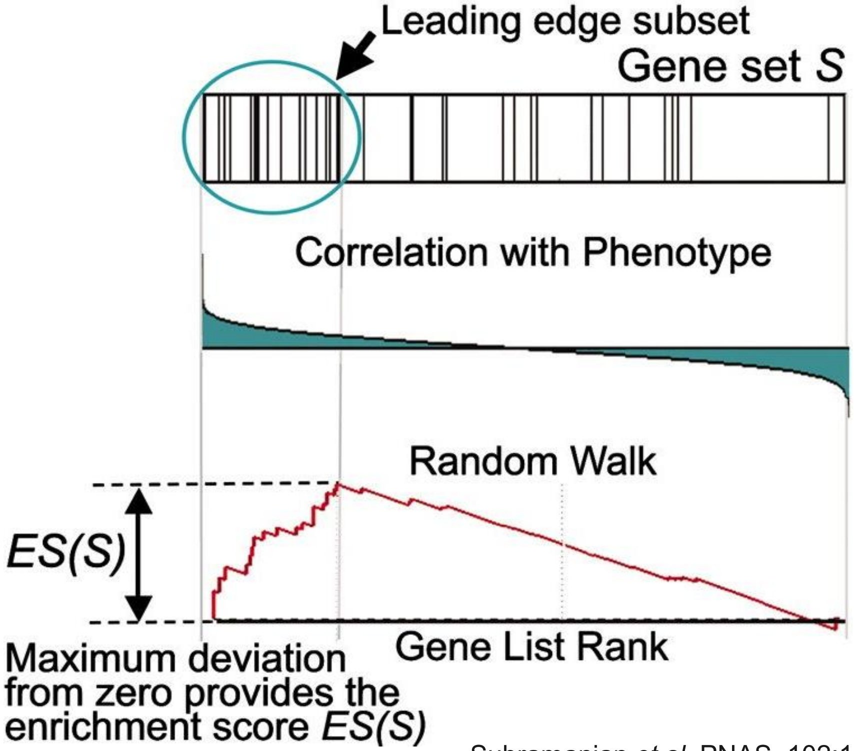
- **1D random walk**: a random process starting from zero, with equal probability of going left or right

- Sometimes stay close to zero, sometimes deviate

- P(maximal deviation > d) ≈ $2\sum_{k=1}^{\infty}(-1)^{k-1}\, e^{-2(kd)^2}$

https://demonstrations.wolfram.com/SimulatingTheSimpleRandomWalk/

Random walk model for GSEA score



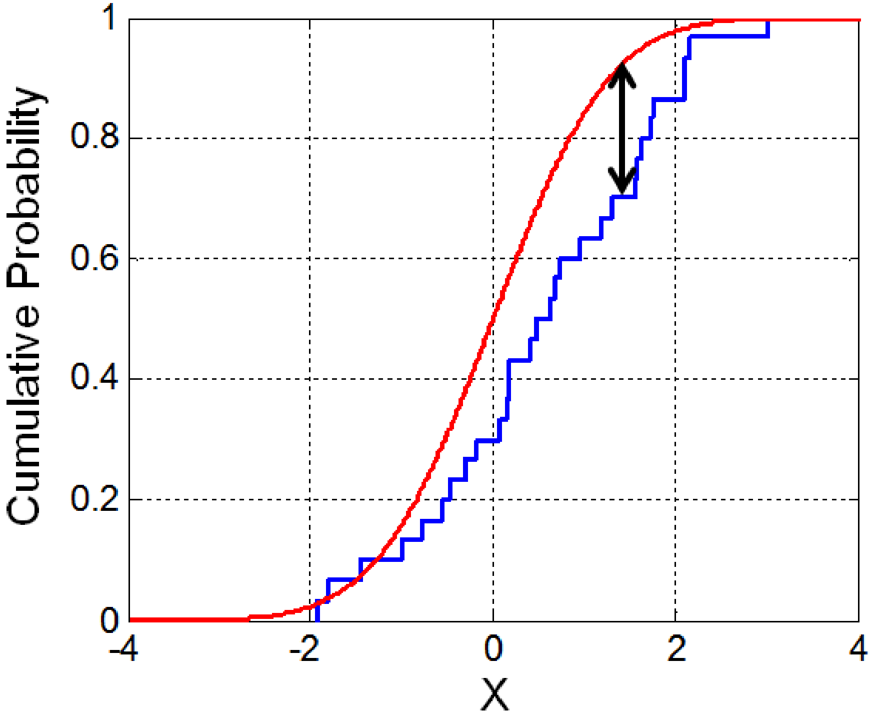
Subramanian *et al*. PNAS. 102:15545-15550 (2005)

- **GSEA Null Hypothesis**: Gene rank is not associated with selected annotation

- Cumulative scores go up and down randomly (with probability of going up = frequency of annotated genes)

- We know the p-value!

Kolmogorov-Smirnov test



https://en.wikipedia.org/wiki/Kolmogorov%E2%80%93Smirnov_test

- Test whether two probability distributions are equal
- Calculate the maximum deviation between cumulative densities
- **Null Hypothesis**: Gap between cumulative densities is random
- P-value according to random walk

Tips for GSEA

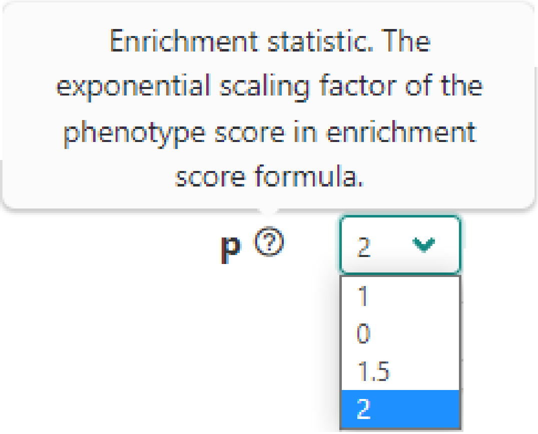
- Which DEG scores to use to rank genes?

- GSEA was optimized for microarray data. For RNA-seq, many people use different scores, such as **log-fold change** or **−10 log p-value**

- GSEA implements exponential factor *p* (score^p)

- Setting *p* = 0, means that score = +/-1

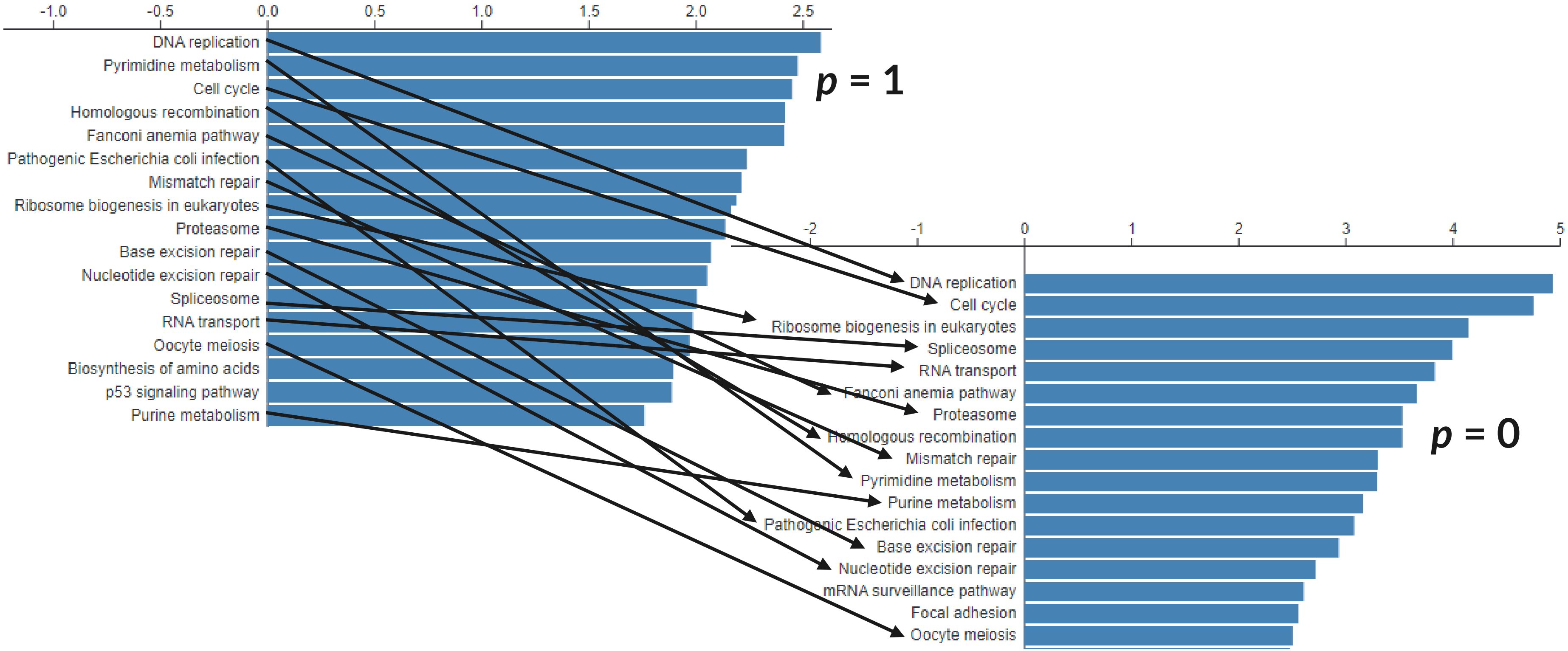
- This can help if you have extreme fold changes



Enrichment statistic. The exponential scaling factor of the phenotype score in enrichment score formula.

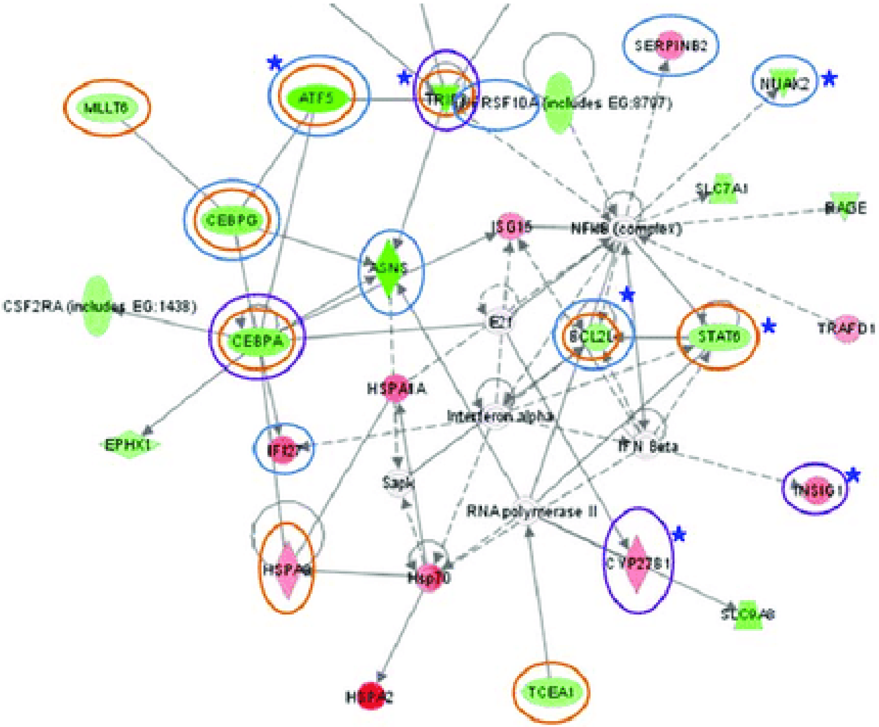
p ⓘ 2

Impact of setting *p*



Gene-gene network

Gene-gene network



Dubey, R. et al. PLoS ONE 6:e28509 (2011)

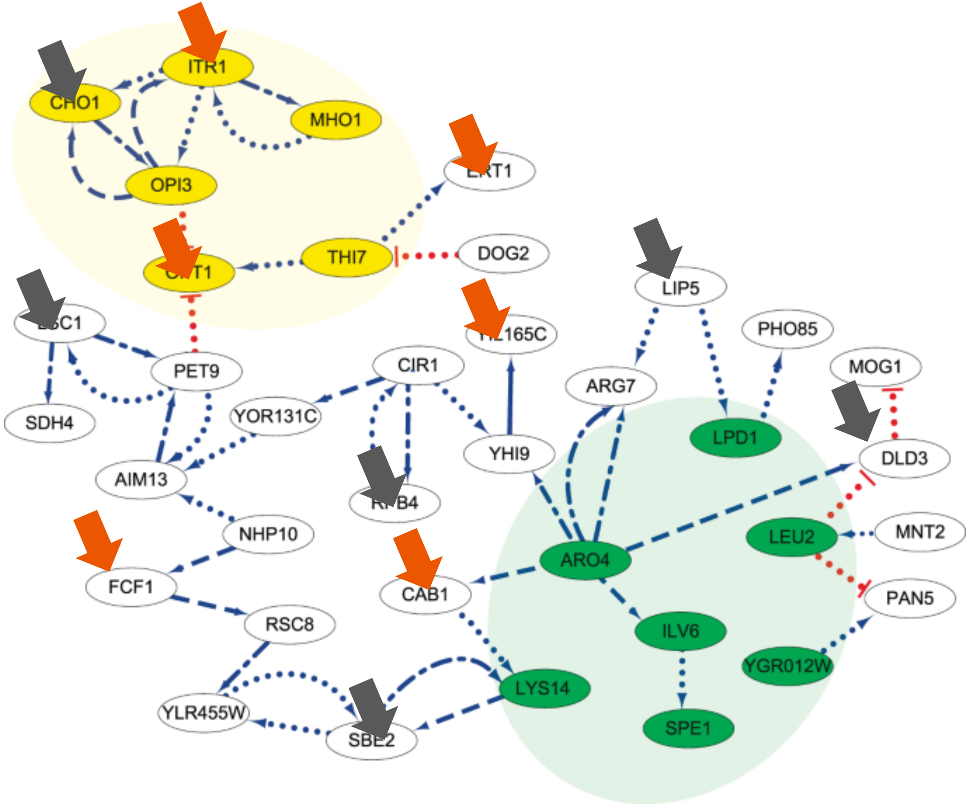
- Observing more DEGs from the same pathway → higher confidence
- Even better if fold-changes match gene regulation
- Even better if DEGs are connected to each other

Functional enrichment on gene-gene network

- Given a group of DEGs, measure their connectivity on the network

- The number of edges needed to connect all of them

- Compared to randomly selected genes



Chen, C. et al. Scientific Reports 9:1197 (2019)

Permutation test

Permutation test as hypothesis testing

- **Alternative Hypothesis**: DEGs are highly connected on the gene-gene network because treatment affects that biological process

- **Null hypothesis**: The connectivity between DEGs is due to chance. Other random sets of genes could yield the same connectivity.

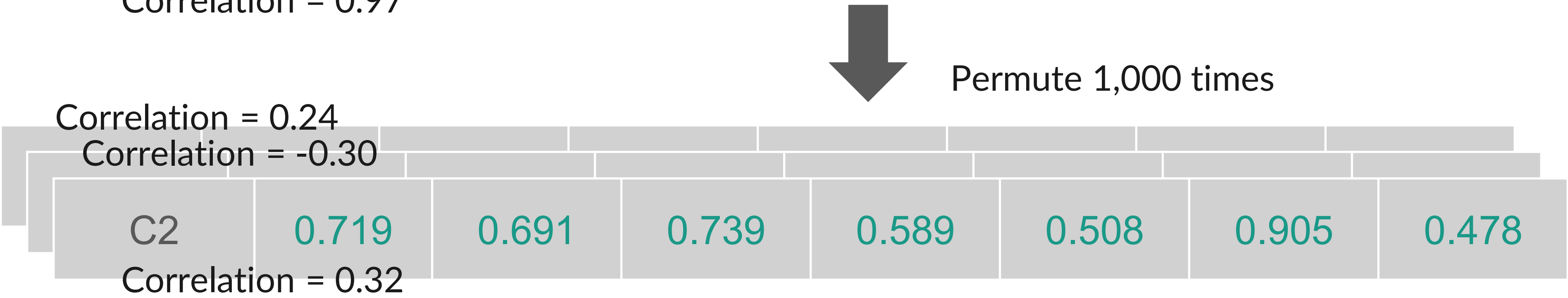
- **P-value** = Probability that a random set of genes has the same or higher connectivity score than the DEGs

- Measure by sampling many random sets of genes (**permutation**)

Examples of permutation test

Cell	Gene 1	Gene 2	Gene 3	Gene 4	Gene 5	Gene 6	Gene 7
C1	0.701	0.503	0.991	0.827	0.623	0.728	0.596
C2	0.691	0.478	0.905	0.739	0.589	0.719	0.508

Correlation = 0.97



Permute 1,000 times

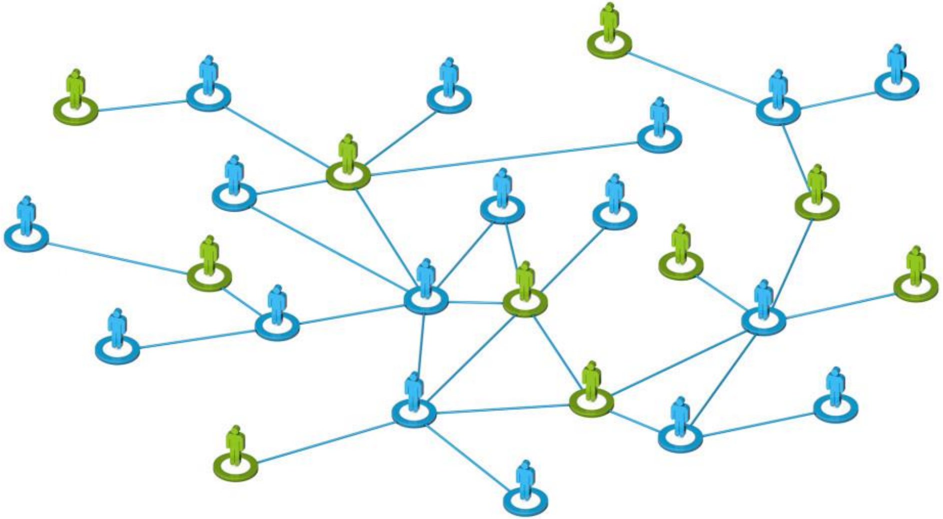
Correlation = 0.24

Correlation = -0.30

C2	0.719	0.691	0.739	0.589	0.508	0.905	0.478

Correlation = 0.32

Examples of permutation test

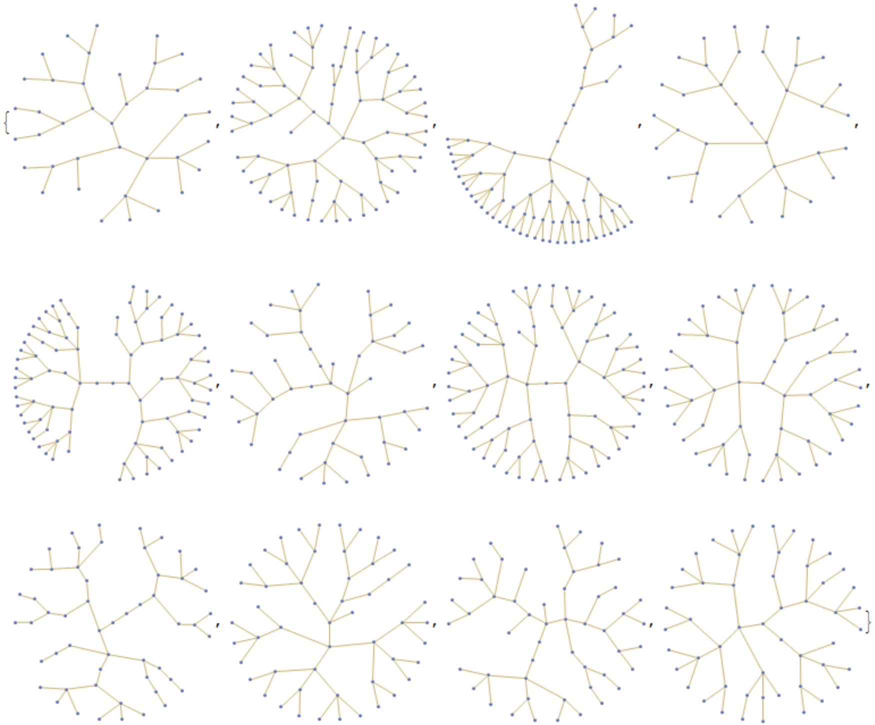


Observation: There are many edges connecting people with different color

Null Hypothesis: The observation is due to chance. Any network with 17 blue nodes, 10 green nodes, and 32 edges would have that property

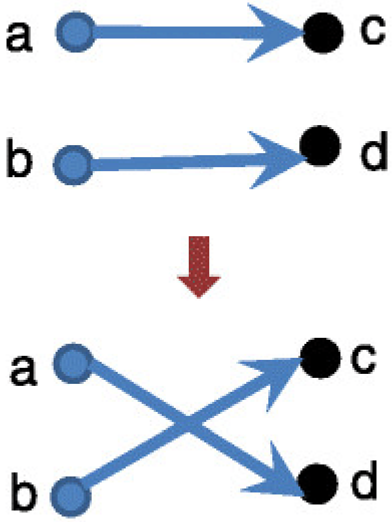
- **Permutation test**: Generate 1,000 random networks with 17 blue nodes, 10 green nodes, and 32 edges and count # edges connect blue-green

Generating random networks



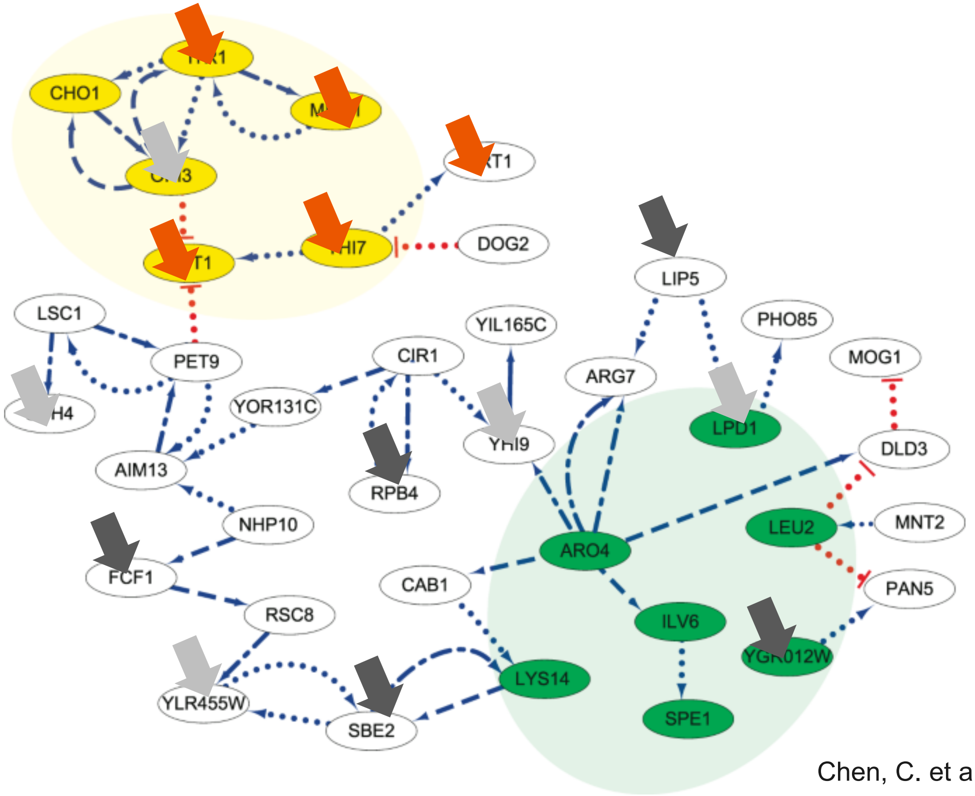
Source: https://mathematica.stackexchange.com/questions/11632/how-to-generate-a-random-tree

Edge switching



Temate-Tiageru *et al*. BMC Genomics, 17:542 (2016)

Permutation test for functional enrichment

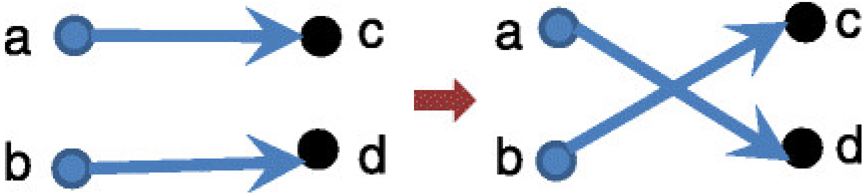


Chen, C. et al. Scientific Reports 9:1197 (2019)

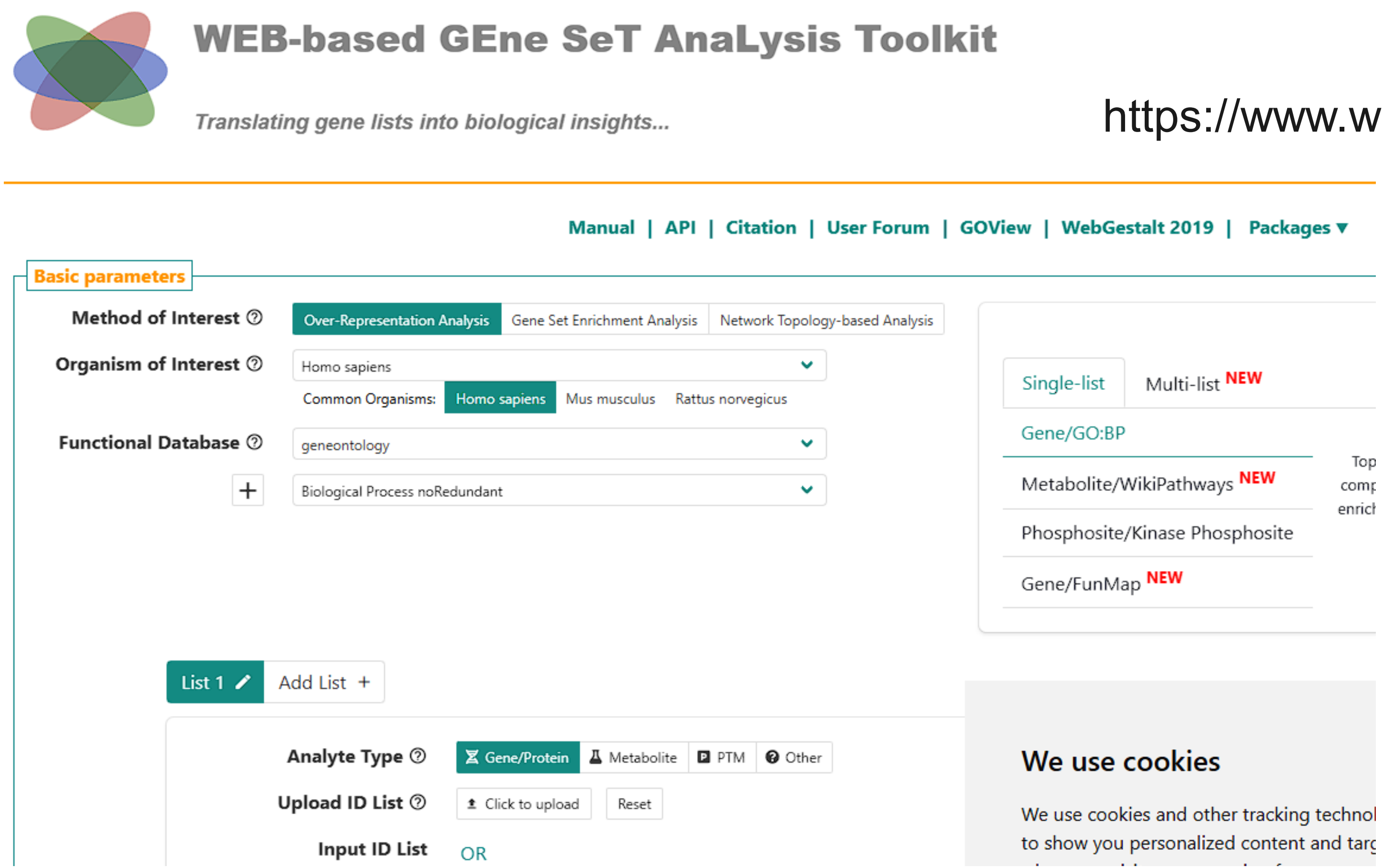
- Fix the network structure
- Randomly select set of genes
- Calculate connectivity score
- **P-value** = # of permutations with better connectivity scores compared to the original score

Tips for permutation test

- Non-parametric, work on any data structure
- Be careful when permuting data: **some structure must be preserved**
- Edge switching preserves the total number of edges coming in and out of each node



Temate-Tiageru *et al*. BMC Genomics, 17:542 (2016) https://www.webgestalt.org/



WEB-based GEne SeT AnaLysis Toolkit

Translating gene lists into biological insights...

Manual | API | Citation | User Forum | GOView | WebGestalt 2019 | Packages ▾

Basic parameters

Method of Interest ⓘ [✓ Over-Representation Analysis] Gene Set Enrichment Analysis Network Topology-based Analysis

Organism of Interest ⓘ Homo sapiens

Common Organisms: [✓ Homo sapiens] Mus musculus Rattus norvegicus

Functional Database ⓘ geneontology

+ Biological Process noRedundant

[✓ Single-list] Multi-list NEW

Gene/GO:BP

Metabolite/WikiPathways NEW

Phosphosite/Kinase Phosphosite

Gene/FunMap NEW

Top
comp
enrich

List 1 ✎ Add List +

Analyte Type ⓘ [✓ Gene/Protein] Metabolite PTM Other

Upload ID List ⓘ Click to upload Reset

Input ID List OR

We use cookies

We use cookies and other tracking technol
to show you personalized content and targ

Any question?

- See you next time