Pathway and Functional Enrichment Analysis Methods

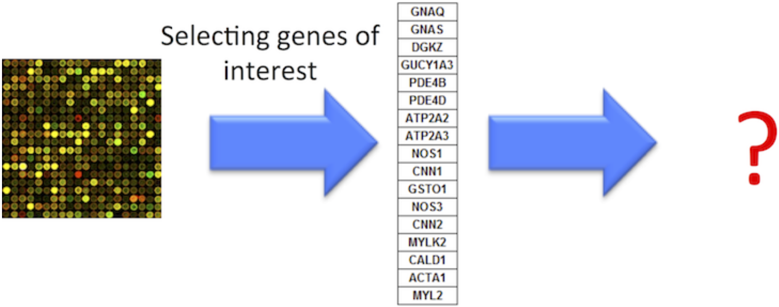
Mikhail Dozmorov

Overview

- Why enrichment analysis?
- What is enrichment analysis?
- Gene ontology and pathways enrichment
- Tools and references

Why enrichment analysis?

- Human genome contains ~20,000-25,000 genes
- Each gene has multiple functions
- If 1,000 genes have changed in an experimental condition, it may be difficult to understand what they do

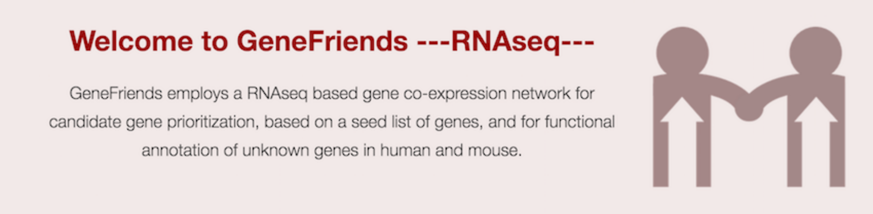


Selecting genes of interest

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

Birds of a feather flock together

- Genes with similar expression patterns share similar functions
- Similar (common) functions characterize a group of genes



Welcome to GeneFriends ---RNAseq---

GeneFriends employs a RNAseq based gene co-expression network for candidate gene prioritization, based on a seed list of genes, and for functional annotation of unknown genes in human and mouse.

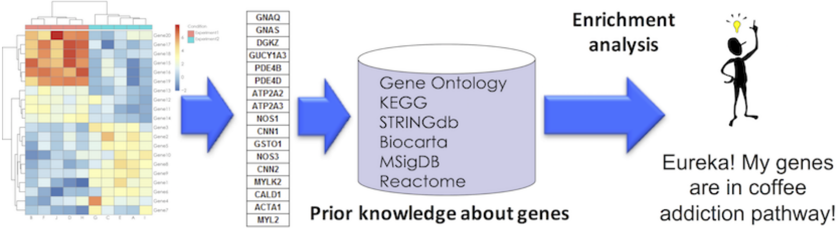
https://genefriends.org/

- People with similar genetic patterns are likely friends

Christakis NA, Fowler JH. "Friendship and natural selection." PNAS 2014 https://www.ncbi.nlm.nih.gov/pubmed/25024208

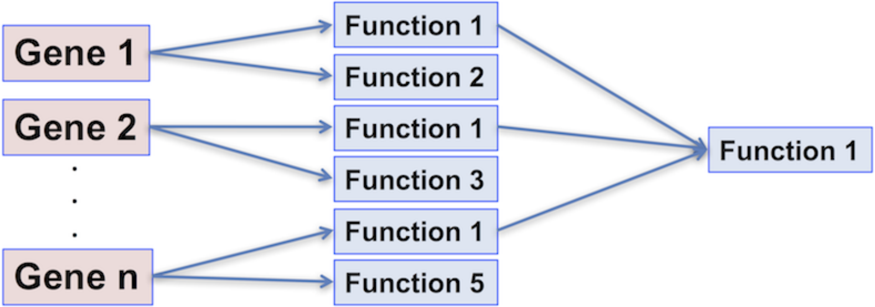
Why enrichment analysis?

- Translating changes of **hundreds/thousands of differentially expressed genes** into a few biological processes (reducing dimensionality)
- High level understanding of the biology behind gene expression – **Interpretation!**



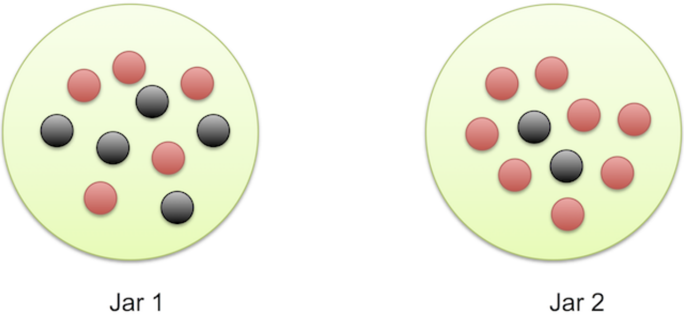
What is enrichment analysis

- **Enrichment analysis** - summarizing common functions associated with a group of objects



What is enrichment analysis? – statistical definition

Enrichment analysis – detection whether a group of objects has certain properties more (or less) frequent than can be expected by chance



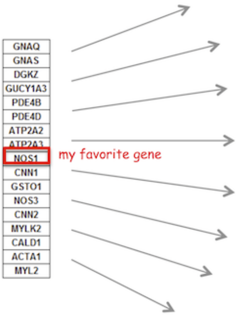
Classification of genes

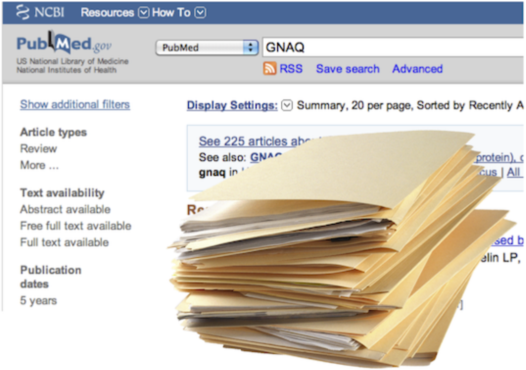
Gene set - *a priori* classification of genes into biologically relevant groups (sets)

- Members of the same biochemical pathways
- Genes annotated with the same molecular function (gene signatures)
- Transcripts expressed in the same cellular compartments
- Co-regulated/co-expressed genes
- Genes located on the same cytogenetic band
- ...

Annotation databases and ontologies

- An annotation database annotates genes with functions or properties - sets of genes with shared functions
- Structured prior knowledge about genes




NCBI
Resources
How To
PubMed.gov
US National Library of Medicine
National Institutes of Health
PubMed
RSS
Save search
Advanced

GNAQ

Show additional filters

Display Settings: Summary, 20 per page, Sorted by Recently A

Article types
Review
More ...

Text availability
Abstract available
Free full text available
Full text available

Publication dates
5 years

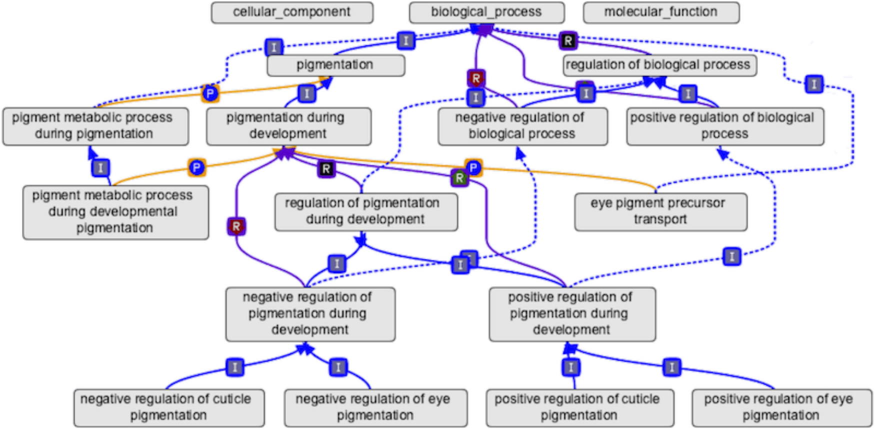
See 225 articles ab
See also: GNA
gnaq in
protein), c
cus | All
Re
sed b
elin LP,

Gene ontology

- An ontology is a formal (hierarchical) representation of concepts and the relationships between them.
- The objective of GO is to provide controlled vocabularies of terms for the description of gene products.
- These terms are to be used as attributes of gene products, facilitating uniform queries across them.

Gene ontology hierarchy

- Terms are related using “is-a”, “part-of” and other connectors


cellular_component
biological_process
molecular_function
pigmentation
regulation of biological process
pigment metabolic process during pigmentation
pigmentation during development
negative regulation of biological process
positive regulation of biological process
pigment metabolic process during developmental pigmentation
regulation of pigmentation during development
eye pigment precursor transport
negative regulation of pigmentation during development
positive regulation of pigmentation during development
negative regulation of cuticle pigmentation
negative regulation of eye pigmentation
positive regulation of cuticle pigmentation
positive regulation of eye pigmentation
I
P
R

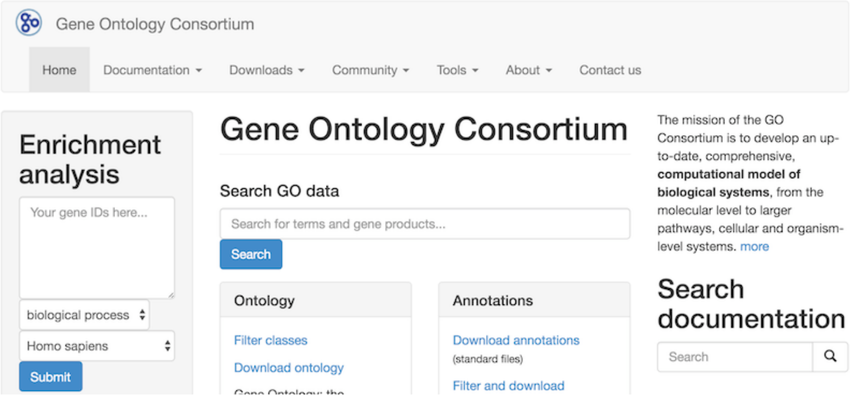
http://geneontology.org/docs/ontology-relations/

Gene ontology structure

Gene ontology describes multiple levels of detail of gene function.

- **Molecular Function** - the tasks performed by individual gene products; examples are *transcription factor* and *DNA helicase*
- **Biological Process** - broad biological goals, such as *mitosis* or *purine metabolism*, that are accomplished by ordered assemblies of molecular functions
- **Cellular Component** - subcellular structures, locations, and macromolecular complexes; examples include *nucleus*, *telomere*, and *origin recognition complex*

Gene ontology database



Enrichment analysis

Your gene IDs here...

biological process

Homo sapiens

Gene Ontology Consortium

Search GO data

Search for terms and gene products...

Ontology

Filter classes

Download ontology

Gene Ontology: the

Annotations

Download annotations (standard files)

Filter and download

The mission of the GO Consortium is to develop an up-to-date, comprehensive, **computational model of biological systems**, from the molecular level to larger pathways, cellular and organism-level systems. more

Search documentation

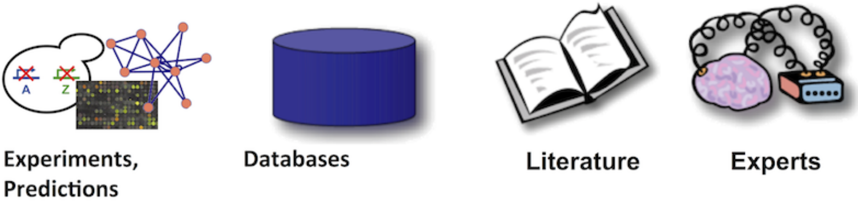
Search

http://geneontology.org/

https://www.ebi.ac.uk/QuickGO/

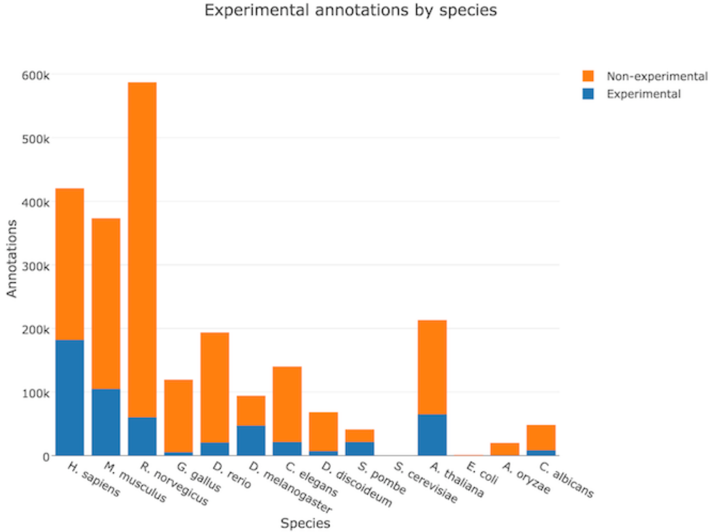
Gene ontologies are not created equal

- Different levels of evidence:
  - Experimental
  - Computational analysis
  - Author Statement
  - Curator Statement
  - Inferred from electronic annotation



http://geneontology.org/docs/guide-go-evidence-codes/

Gene ontologies are not created equal

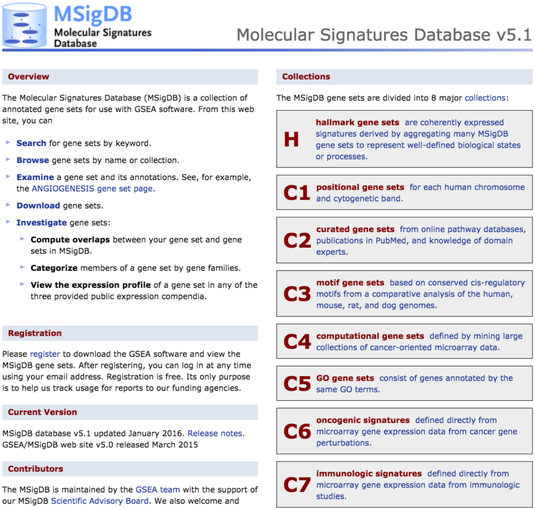


http://amigo.geneontology.org/amigo/base_statistics

Gene ontologies for model organisms

- **Mouse Genome Database** (MGD) and Gene Expression Database (GXD) (Mus musculus) http://www.informatics.jax.org/
- **Rat Genome Database** (RGD) (Rattus norvegicus) http://rgd.mcw.edu/
- **FlyBase** (Drosophila melanogaster) http://flybase.org/
- **Berkeley Drosophila Genome Project** (BDGP) http://www.fruitfly.org/
- **WormBase** (Caenorhabditis elegans) http://www.wormbase.org/
- **Zebrafish Information Network** (ZFIN) (Danio rerio) http://zfin.org/
- **Saccharomyces Genome Database** (SGD) (Saccharomyces cerevisiae) http://www.yeastgenome.org/
- **The Arabidopsis Information Resource** (TAIR) (Arabidopsis thaliana) https://www.arabidopsis.org/
- **Gramene** (grains, including rice, Oryza) http://www.gramene.org/
- **dictyBase** (Dictyostelium discoideum) http://dictybase.org/
- **GeneDB** (Schizosaccharomyces pombe, Plasmodium falciparum, Leishmania major and Trypanosoma brucei) http://www.genedb.org/

MSigDb - Molecular Signatures Database



MSigDB Molecular Signatures Database

Molecular Signatures Database v5.1

Overview

The Molecular Signatures Database (MSigDB) is a collection of annotated gene sets for use with GSEA software. From this web site, you can

- **Search** for gene sets by keyword.
- **Browse** gene sets by name or collection.
- **Examine** a gene set and its annotations. See, for example, the ANGIOGENESIS gene set page.
- **Download** gene sets.
- **Investigate** gene sets:
  - **Compute overlaps** between your gene set and gene sets in MSigDB.
  - **Categorize** members of a gene set by gene families.
  - **View the expression profile** of a gene set in any of the three provided public expression compendia.

Registration

Please register to download the GSEA software and view the MSigDB gene sets. After registering, you can log in at any time using your email address. Registration is free. Its only purpose is to help us track usage for reports to our funding agencies.

Current Version

MSigDB database v5.1 updated January 2016. Release notes.
GSEA/MSigDB web site v5.0 released March 2015

Contributors

The MSigDB is maintained by the GSEA team with the support of our MSigDB Scientific Advisory Board. We also welcome and

Collections

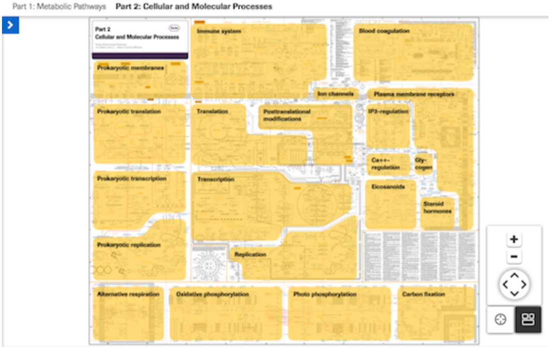
The MSigDB gene sets are divided into 8 major collections:

H	**hallmark gene sets** are coherently expressed signatures derived by aggregating many MSigDB gene sets to represent well-defined biological states or processes.
C1	**positional gene sets** for each human chromosome and cytogenetic band.
C2	**curated gene sets** from online pathway databases, publications in PubMed, and knowledge of domain experts.
C3	**motif gene sets** based on conserved cis-regulatory motifs from a comparative analysis of the human, mouse, rat, and dog genomes.
C4	**computational gene sets** defined by mining large collections of cancer-oriented microarray data.
C5	**GO gene sets** consist of genes annotated by the same GO terms.
C6	**oncogenic signatures** defined directly from microarray gene expression data from cancer gene perturbations.
C7	**immunologic signatures** defined directly from microarray gene expression data from immunologic studies.

http://software.broadinstitute.org/gsea/msigdb/

Pathways

- An ordered series of molecular events that leads to the creation new molecular product, or a change in a cellular state or process.
- Genes often participate in multiple pathways – think about genes having multiple functions



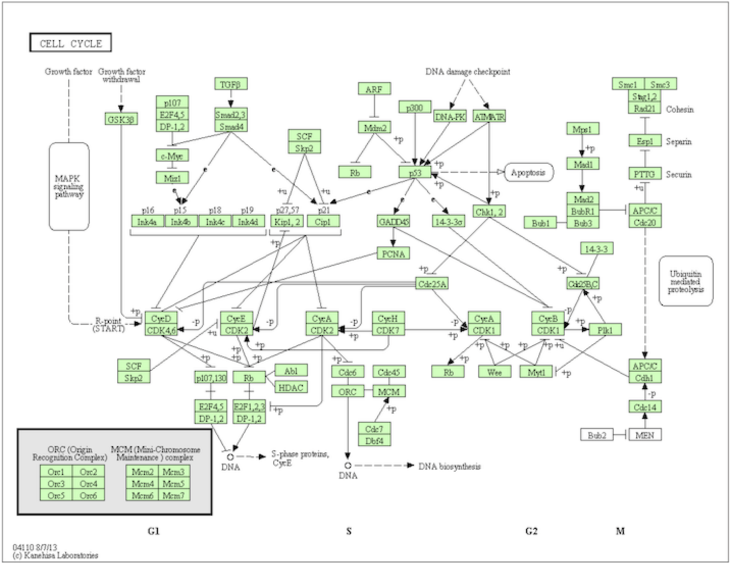
http://biochemical-pathways.com/#/map/1

KEGG pathway database

- **KEGG: Kyoto Encyclopedia of Genes and Genomes** is a collection of biological information compiled from published material = curated database.
- Includes information on genes, proteins, metabolic pathways, molecular interactions, and biochemical reactions associated with specific organisms
- Provides a relationship (map) for how these components are organized in a cellular structure or reaction pathway.

http://www.genome.jp/kegg/

KEGG pathway diagram

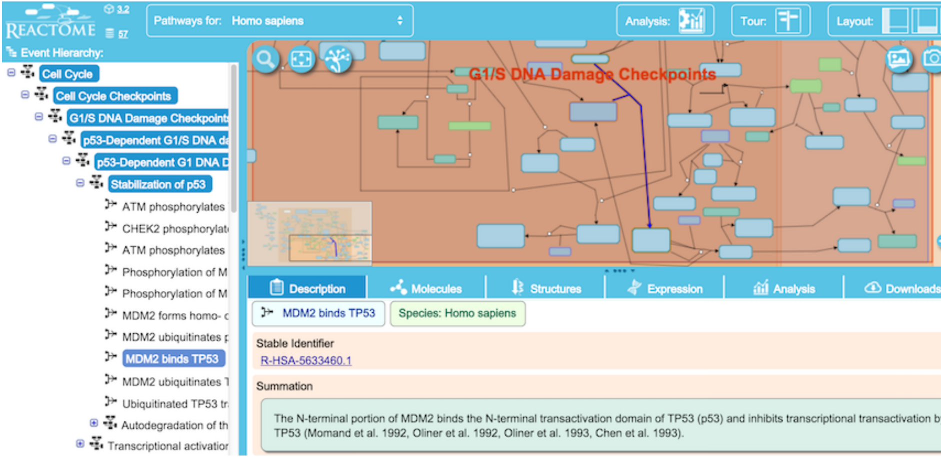


Reactome

- Curated human pathways encompassing metabolism, signaling, and other biological processes.
- Every pathway is traceable to primary literature.

http://www.reactome.org/

Reactome pathway diagram



MDM2 binds TP53

Species: Homo sapiens

Stable Identifier

R-HSA-5633460.1

Summation

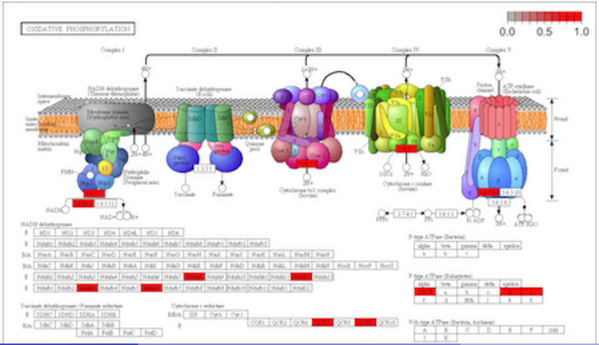
The N-terminal portion of MDM2 binds the N-terminal transactivation domain of TP53 (p53) and inhibits transcriptional transactivation b TP53 (Momand et al. 1992, Oliner et al. 1992, Oliner et al. 1993, Chen et al. 1993).

Other pathway databases

- **PathwayCommons**, version 12 has over 5,770 pathways from 22 data sources, http://www.pathwaycommons.org/
- **PathGuide**, lists over 700 pathway related databases, http://www.pathguide.org/
- **WikiPathways**, community-curated pathways, http://wikipathways.org/
- **Consensus-PathDB**, pathway interactions, enrichment, data, http://www.consensuspathdb.org/

Gene annotation databases in R

- **annotables** (https://github.com/stephenturner/annotables) - R data package for annotating/converting Gene IDs
- **msigdf** (https://github.com/stephenturner/msigdf) - Molecular Signatures Database (MSigDB) in a data frame
- **pathview** (https://bioconductor.org/packages/pathview/) - a tool set for pathway based data integration and visualization



Genes to networks

- **GeneMania**, networks based on different properties, http://genemania.org
- **STRING**, protein-protein interaction networks, http://string-db.org
- **Genes2Networks**, protein-protein interaction networks, http://amp.pharm.mssm.edu/X2K/#g2n
- **IntAct**, protein-protein interaction data and networks, https://www.ebi.ac.uk/intact/
- **HPRD**, protein-protein interaction database, http://www.hprd.org/

Section 1

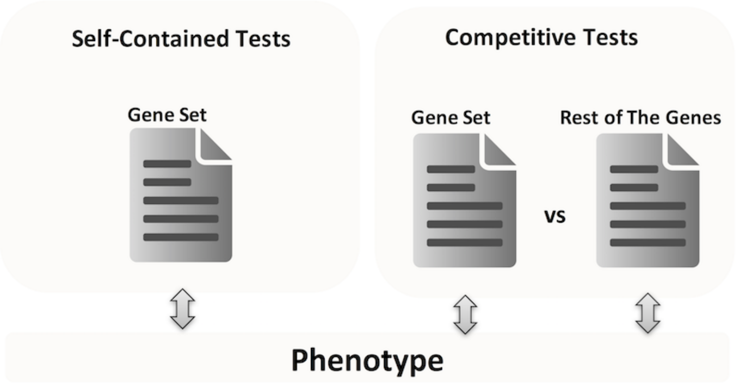
Enrichment analysis

Types of enrichment analyses

- **First generation** - traditional overrepresentation analyses, hypergeometric distribution-based test whether genes of interest (i.e., differentially expressed) are overrepresented in functional gene sets.
- **Second generation** - tests the tendency of gene set members to appear rather at the top or bottom of the ranked list of all measured genes.
- **Third generation** - network- or topology-based tests, consider relationships among genes.

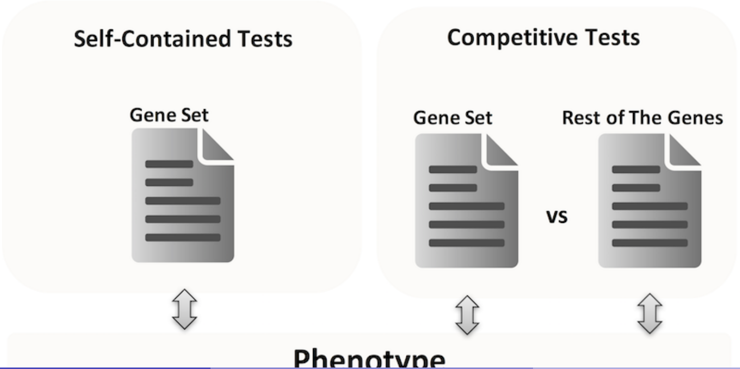
First generation enrichment analysis: Null hypothesis

- **Self-contained** H_0: genes in the gene set do not have any association with the pheontype
- Problem: restrictive, use information only from a gene set



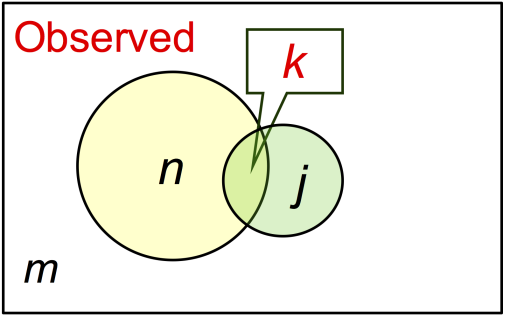
First generation enrichment analysis: Null hypothesis

- **Competitive** H_0: genes in the gene set have the same level of association with a given phenotype as genes in the complement gene set
- Problem: wrong assumption of independent gene sampling



Hypergeometric test

- m is the total number of genes
- j is the number of genes are in the functional category
- n is the number of differentially expressed genes
- k is the number of differentially expressed genes in the category



Hypergeometric test

- m is the total number of genes
- j is the number of genes are in the functional category
- n is the number of differentially expressed genes
- k is the number of differentially expressed genes in the category

The expected value of k would be $k_e = (n/m) * j$.

If $k > k_e$, functional category is said to be enriched, with a ratio of enrichment $r = k/k_e$

Hypergeometric test

- *m* is the total number of genes
- *j* is the number of genes are in the functional category
- *n* is the number of differentially expressed genes
- *k* is the number of differentially expressed genes in the category

	Diff. exp. genes	Not Diff. exp. genes	Total
In gene set	k	j-k	j
Not in gene set	n-k	m-n-j+k	m-j
Total	n	m-n	m

Hypergeometric test

- m is the total number of genes
- j is the number of genes are in the functional category
- n is the number of differentially expressed genes
- k is the number of differentially expressed genes in the category

What is the probability of having k or more genes from the category in the selected n genes?

$$P = \sum_{i=k}^{n} \frac{\binom{m-j}{n-i}\binom{j}{i}}{\binom{m}{n}}$$

Hypergeometric test

- m is the total number of genes
- j is the number of genes are in the functional category
- n is the number of differentially expressed genes
- k is the number of differentially expressed genes in the category

$k < (n/m) * j$ - underrepresentation. Probability of k or less genes from the category in the selected n genes?

$$P = \sum_{i=0}^{k} \frac{\binom{m-j}{n-i}\binom{j}{i}}{\binom{m}{n}}$$

Hypergeometric test

1. Find a set of differentially expressed genes (DEGs)
2. Are *DEGs in a set* more common than *DEGs not in a set*?

- Fisher test `stats::fisher.test()`
- Conditional hypergeometric test, to account for directed hierachy of GO `GOstats::hyperGTest()`

Example: https://github.com/mdozmorov/MDmisc/blob/master/R/gene_enrichment.R

Problems with Hypergeometric test

- The outcome of the overrepresentation test depends on the significance threshold used to declare genes differentially expressed.
- Functional categories in which many genes exhibit small changes may go undetected.
- Genes are not independent, so a key assumption of the Fisher's exact tests is violated.
- Pathways overlap

Secong generation: Gene set enrichment analysis (GSEA)

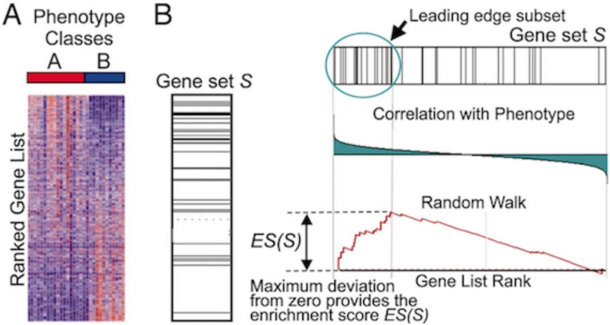
- **Gene set analysis (GSA)**. Mootha et al., 2003; modified by Subramanian, et al. "**Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles.**" PNAS 2005 http://www.pnas.org/content/102/43/15545.abstract
- Main rationale – functionally related genes often display a coordinated expression to accomplish their roles in the cells
- Aims to identify gene sets with "subtle but coordinated" expression changes that would be missed by DEGs threshold selection

GSEA: Gene set enrichment analysis

- The null hypothesis is that the **rank ordering** of the genes in a given comparison is **random** with regard to the case-control assignment.
- The alternative hypothesis is that the **rank ordering** of genes sharing functional/pathway membership is **associated** with the case-control assignment.

GSEA: Gene set enrichment analysis

1. Sort genes by log fold change
2. Calculate running sum - increment when gene in a set, decrement when not
3. Maximum of the runnig sum is the enrichment score - larger means genes in a set are toward top of the sorted list
4. Permute subject labels to calculate significance p-value



GSEA: Gene set enrichment analysis

- Compute a statistic (difference between 2 clinical groups) for each gene that measures the degree of differential expression between treatments.
- Create a list L of all genes ordered according to these statistics.
- Given a set of genes S we can see if these genes are non-randomly distributed in our list L
- If the experiment produced random results, we don't expect gene order to have biological coherence

GSEA: Gene set enrichment analysis

- Calculate an enrichment score (ES) that reflects the degree to which a set S is overrepresented at the extremes (top or bottom) of the entire ranked list L.
- The score is calculated by walking down the list L and ...
  - Increase a running-sum statistic when we encounter a gene in S
  - Decrease it when we encounter genes not in S.
- The magnitude of the increment depends on the correlation of the gene with the phenotype.
- The final enrichment score is the maximum deviation from zero encountered in the random walk
  - Corresponds to a weighted Kolmogorov–Smirnov-like statistics

GSEA: Gene set enrichment analysis

Enrichment Score

- Consider genes $R_1, ..., R_N$ ordered by the difference metric
- Consider a gene set S of size G, containing functionally similar genes or pathway members.
- If R_i is not a member of S, define

$$X_{Ri} = -\sqrt{\frac{G}{N-G}}$$

- If R_i is a member of S, define

$$X_{Ri} = \sqrt{\frac{N-G}{G}}$$

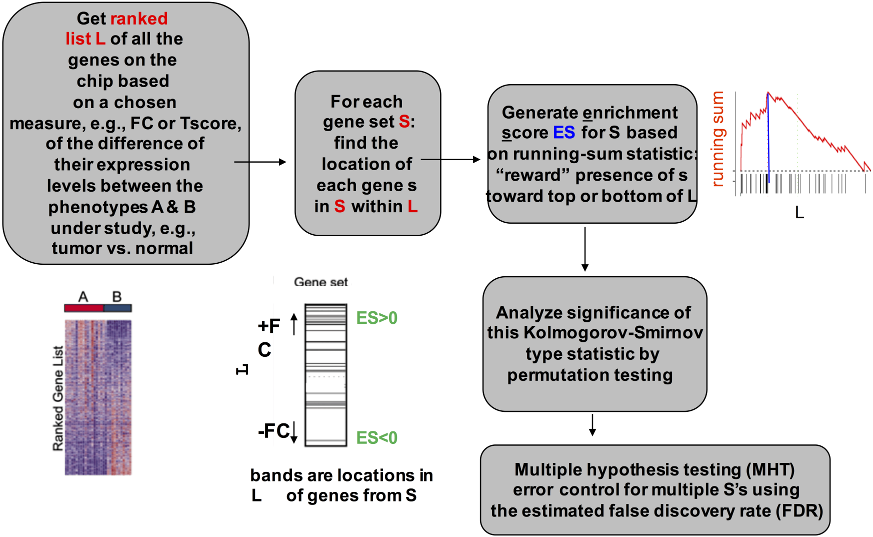
GSEA: Gene set enrichment analysis

Enrichment Score

- Compute running sum across all N genes. The ES is defined as

$$\max_{1 \leq j \leq N} \sum_{i=1}^{j} X_{Ri}$$

- or the maximum observed positive deviation of the running sum.
- ES is measured for every gene set considered. To determine whether any of the given gene sets shows association with the class phenotype distinction, permute the class labels 1,000 times, each time recording the maximum ES over all gene sets.



Get **ranked list L** of all the genes on the chip based on a chosen measure, e.g., FC or Tscore, of the difference of their expression levels between the phenotypes A & B under study, e.g., tumor vs. normal

For each gene set **S**: find the location of each gene s in **S** within **L**

Generate enrichment score **ES** for S based on running-sum statistic: “reward” presence of s toward top or bottom of L

Analyze significance of this Kolmogorov-Smirnov type statistic by permutation testing

Multiple hypothesis testing (MHT) error control for multiple S’s using the estimated false discovery rate (FDR)

bands are locations in L of genes from S

“Using the fast preranked gene set enrichment analysis (fgsea) package”, https://davetang.org/muse/2018/01/10/using-fast-preranked-gene-set-enrichment-analysis-fgsea-package/

Other approaches

Linear model-based

- **CAMERA** (Wu and Smyth 2012)
- **C**orrelation-**A**djusted **ME**an **RA**nk gene set test
- Estimating the variance inflation factor associated with inter-gene correlation, and incorporating this into parametric or rank-based test procedures

Other approaches

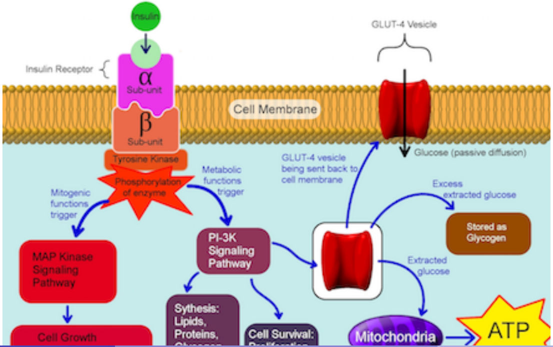
Linear model-based

- **ROAST** (Wu et.al. 2010)
- Under the null hypothesis (and assuming a linear model) the residuals are independent and identically distributed $N(0, \sigma_g^2)$.
- We can *rotate* the residual vector for each gene in a gene set, such that gene-gene expression correlations are preserved.

Third generation: network- or topology-based analyses

Impact analysis - incorporates topology of the pathway.

- Gene's fold change
- Classical enrichment statistics
- The topology of the signaling pathway


Insulin
Insulin Receptor
α
Sub-unit
β
Sub-unit
Tyrosine Kinase
GLUT-4 Vesicle
Cell Membrane
Glucose (passive diffusion)
Phosphorylation of enzyme
Mitogenic functions trigger
Metabolic functions trigger
GLUT-4 vesicle being sent back to cell membrane
Excess extracted glucose
Stored as Glycogen
MAP Kinase Signaling Pathway
PI-3K Signaling Pathway
Extracted glucose
Cell Growth
Sythesis: Lipids, Proteins,
Cell Survival:
Mitochondria
ATP

Third generation: network- or topology-based analyses

- **Pathway-Express**

Sorin Draghici et al., "A Systems Biology Approach for Pathway Level Analysis," *Genome Research*. 2007. https://www.ncbi.nlm.nih.gov/pubmed/17785539

- **SPIA**: Signaling Pathway Impact Analysis, https://bioconductor.org/packages/SPIA/

Adi Laurentiu Tarca et al., "A Novel Signaling Pathway Impact Analysis," *Bioinformatics*. 2009

Tools for Gene set enrichment analysis

- **GSEA** (https://www.broadinstitute.org/gsea/index.jsp) - Better way of doing enrichment analysis
- **g:Profiler** (http://biit.cs.ut.ee/gprofiler/) - gene ID converter, GO and pathway enrichment, and more
- **ToppGene** (https://toppgene.cchmc.org) - Quick gene enrichment analysis in multiple categories
- **Metascape** (http://metascape.org/) - Enrichment analysis of multiple gene sets
- **DAVID** (https://david.ncifcrf.gov/) - Newly updated gene enrichment analysis

Tools for Gene set enrichment analysis

- **clusterProfiler** (https://bioconductor.org/packages/clusterProfiler/) - statistical analysis and visualization of functional profiles for genes and gene clusters
- **limma** (https://bioconductor.org/packages/limma/) - Linear Models for Microarray Data, includes functional enrichment functions `goana`, `camera`, `roast`, `romer`
- **GOstats** (https://bioconductor.org/packages/GOstats/) - tools for manimpuating GO and pathway enrichment analyses.