Pathway and Functional Enrichment Analysis Methods

Fall 2017

Overview

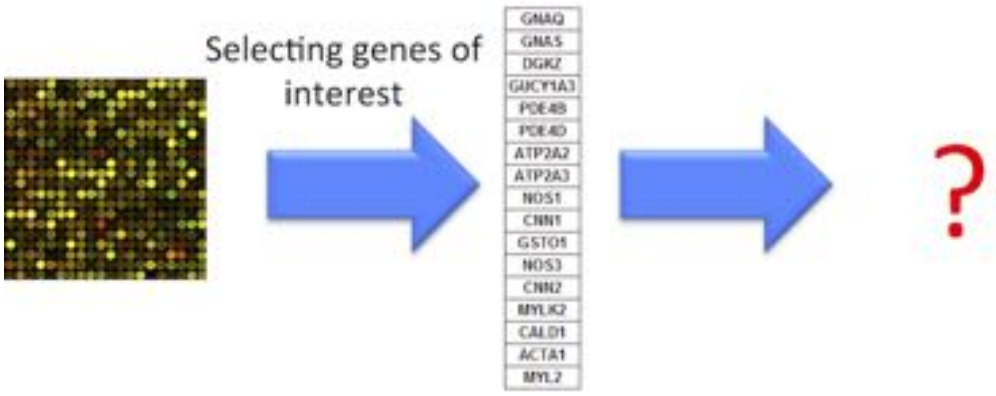
- Why enrichment analysis?
- What is enrichment analysis?
- Gene ontology and pathways
- GENE ontology and pathways enrichment
- Tools and references

Overview

- **Why enrichment analysis?**
- What is enrichment analysis?
- Gene ontology and pathways
- GENE ontology and pathways enrichment
- GENOMIC REGIONS enrichment
- Tools and references

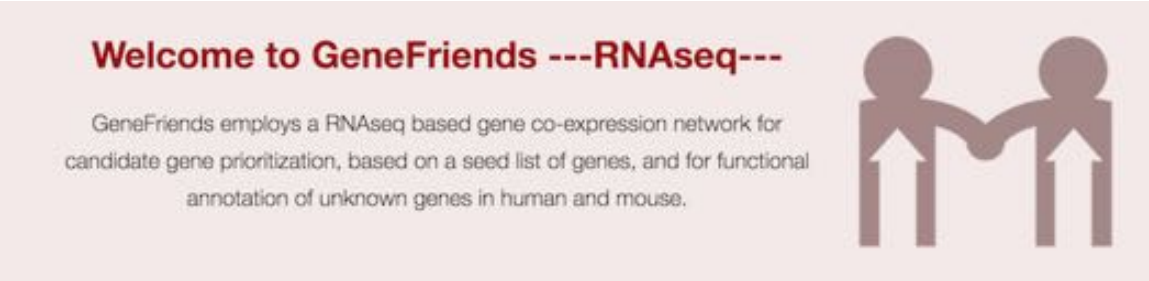
Why enrichment analysis?

- 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



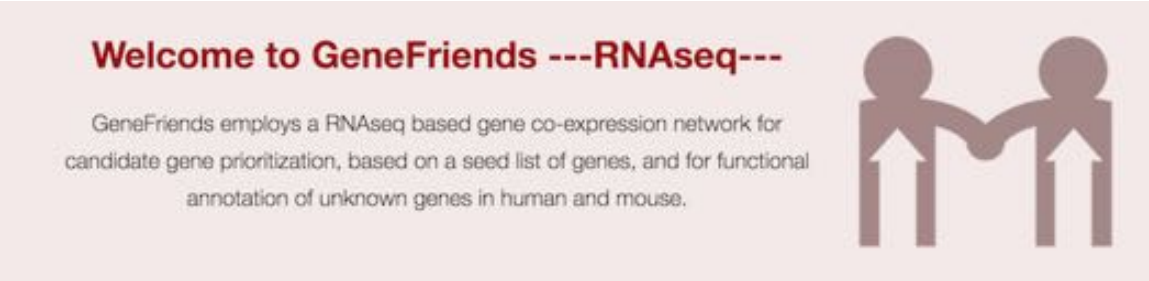
Birds of a feather flock together

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



Birds of a feather flock together

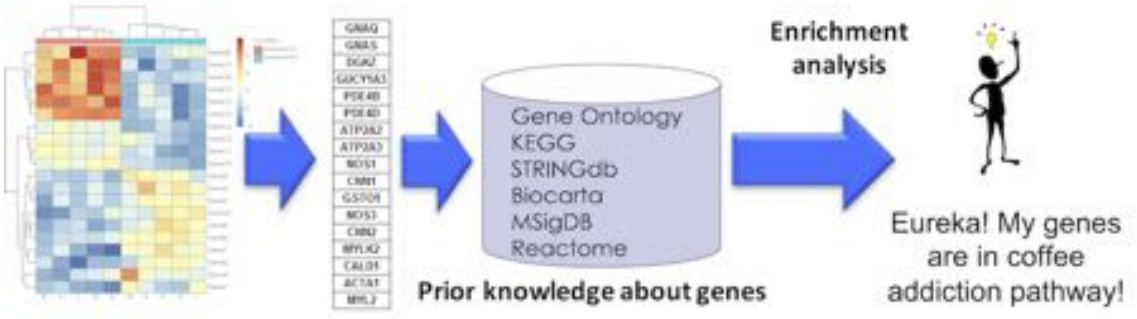
- Genes with similar expression patterns share similar functions
- Similar (common) functions characterize a group of genes



- 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?

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

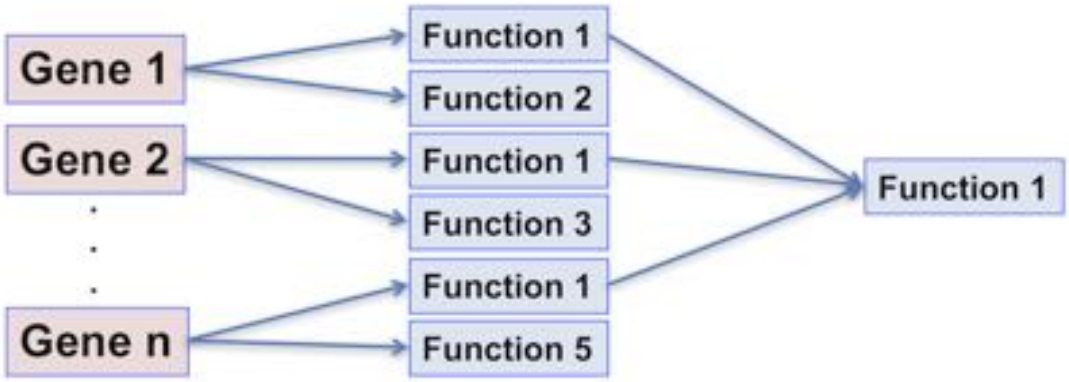


Overview

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

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
- Transcripts expressed in the same cellular compartments
- Co-regulated/co-expressed genes
- Genes located on the same cytogenetic band
- …

Overview

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

Annotation databases and ontologies

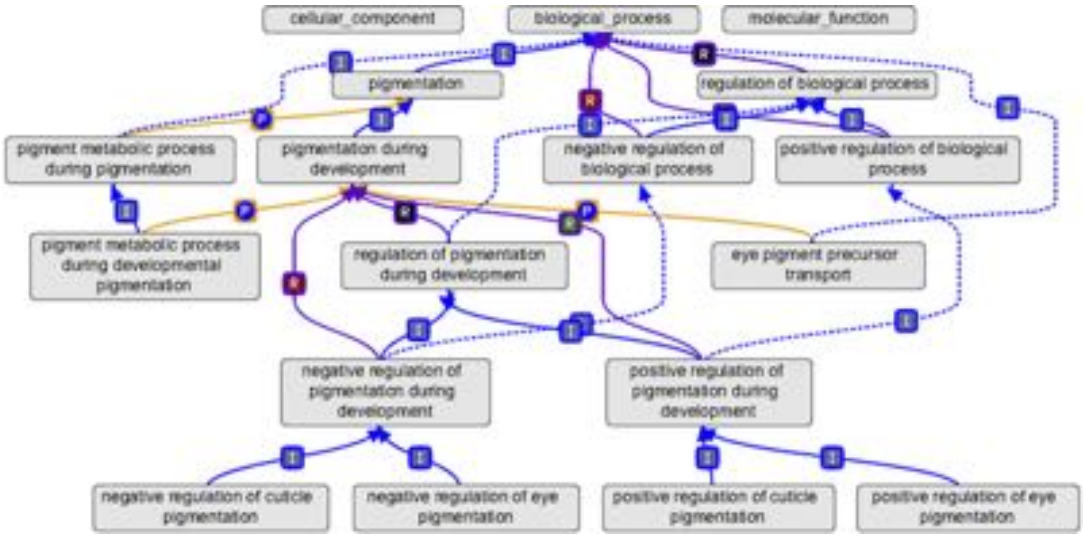
- An annotation database annotates genes with functions or properties - sets of genes with shared functions
- Structured prior knowledge about genes

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 within a hierarchy using "is-a", "part-of" and other connectors



Gene ontology structure

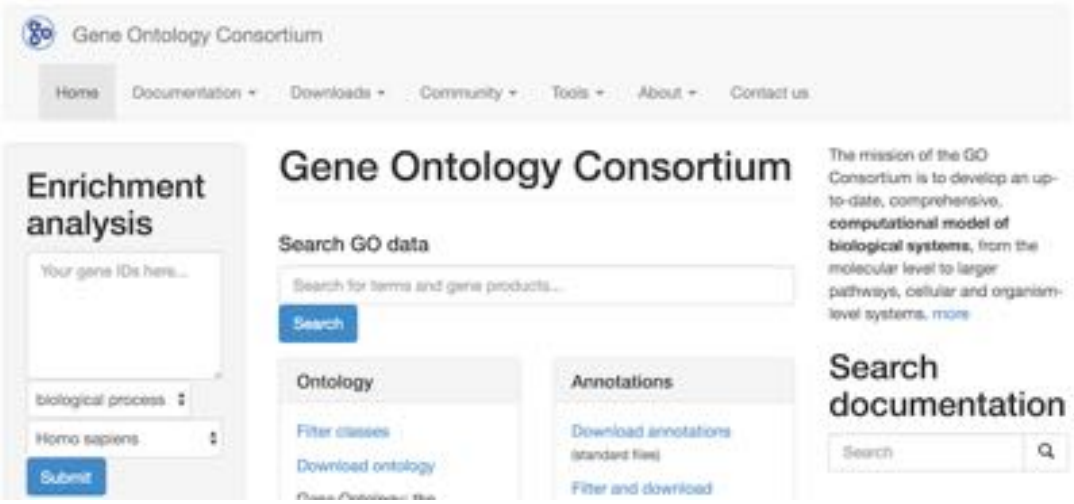
Gene ontology describes multiple levels of detail of gene function.

- **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

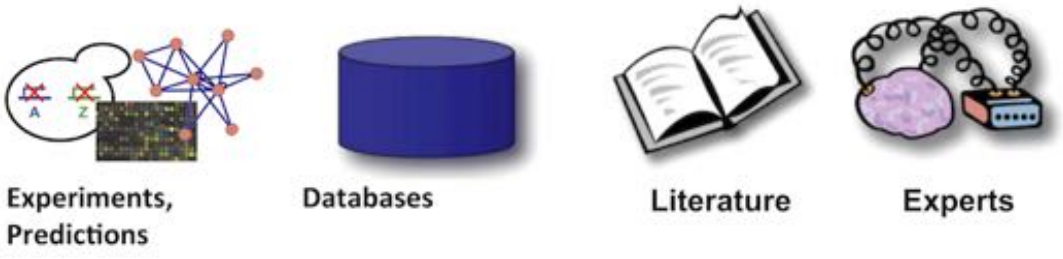
http://geneontology.org/

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



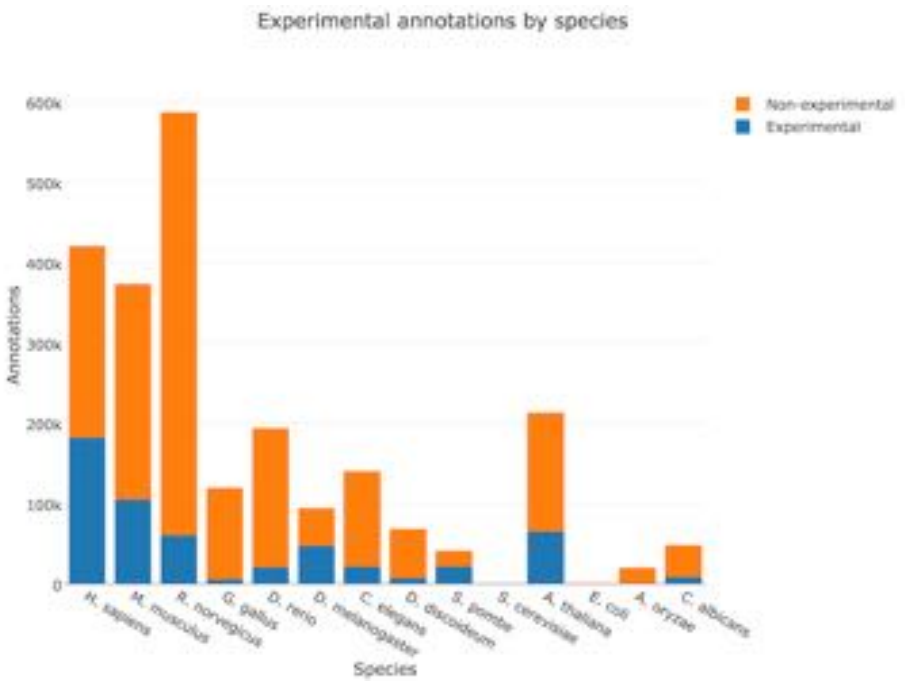
Gene ontologies are not created equal

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



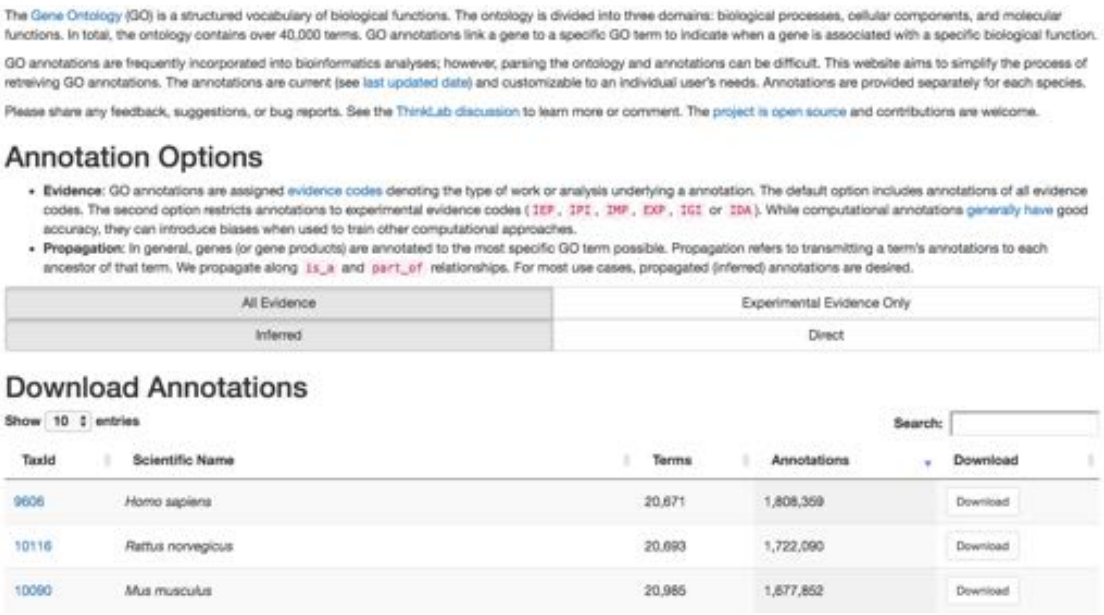
http://geneontology.org/page/evidence-code-decision-tree

Gene ontologies are not created equal



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

User-friendly Gene Ontology annotations



The Gene Ontology (GO) is a structured vocabulary of biological functions. The ontology is divided into three domains: biological processes, cellular components, and molecular functions. In total, the ontology contains over 40,000 terms. GO annotations link a gene to a specific GO term to indicate when a gene is associated with a specific biological function.

GO annotations are frequently incorporated into bioinformatics analyses; however, parsing the ontology and annotations can be difficult. This website aims to simplify the process of retreiving GO annotations. The annotations are current (see last updated date) and customizable to an individual user's needs. Annotations are provided separately for each species.

Please share any feedback, suggestions, or bug reports. See the ThinkLab discussion to learn more or comment. The project is open source and contributions are welcome.

Annotation Options

- **Evidence**: GO annotations are assigned evidence codes denoting the type of work or analysis underlying a annotation. The default option includes annotations of all evidence codes. The second option restricts annotations to experimental evidence codes (`IEP`, `IPI`, `IMP`, `EXP`, `IGI` or `IDA`). While computational annotations generally have good accuracy, they can introduce biases when used to train other computational approaches.
- **Propagation**: In general, genes (or gene products) are annotated to the most specific GO term possible. Propagation refers to transmitting a term's annotations to each ancestor of that term. We propagate along `is_a` and `part_of` relationships. For most use cases, propagated (inferred) annotations are desired.

All Evidence	Experimental Evidence Only
Inferred	Direct

Download Annotations

Show 10 entries Search:

TaxId	Scientific Name	Terms	Annotations	Download
9606	*Homo sapiens*	20,671	1,808,359	Download
10116	*Rattus norvegicus*	20,693	1,722,090	Download
10090	*Mus musculus*	20,985	1,677,852	Download

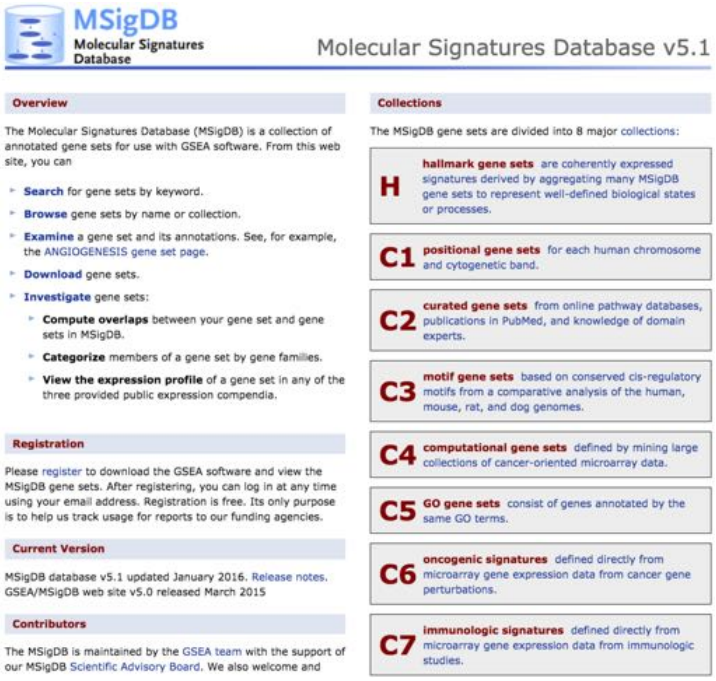
http://git.dhimmel.com/gene-ontology/

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

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



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.

MSigDb - Molecular Signatures Database

https://github.com/stephenturner/msigdf

- **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.

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

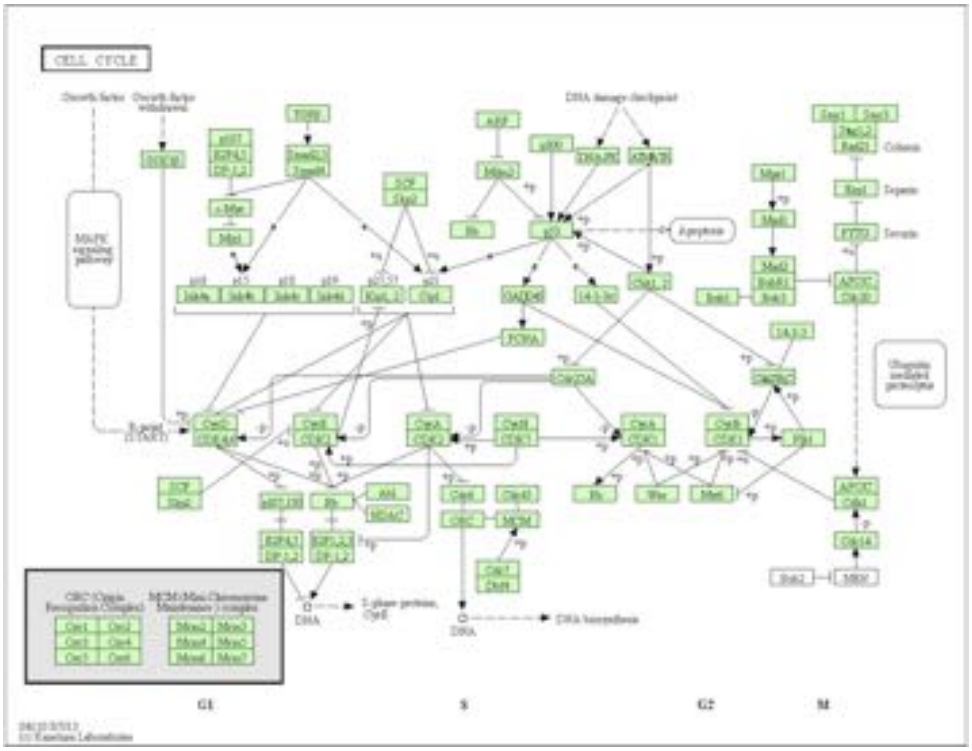
http://biochemical-pathways.com/#/map/1

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

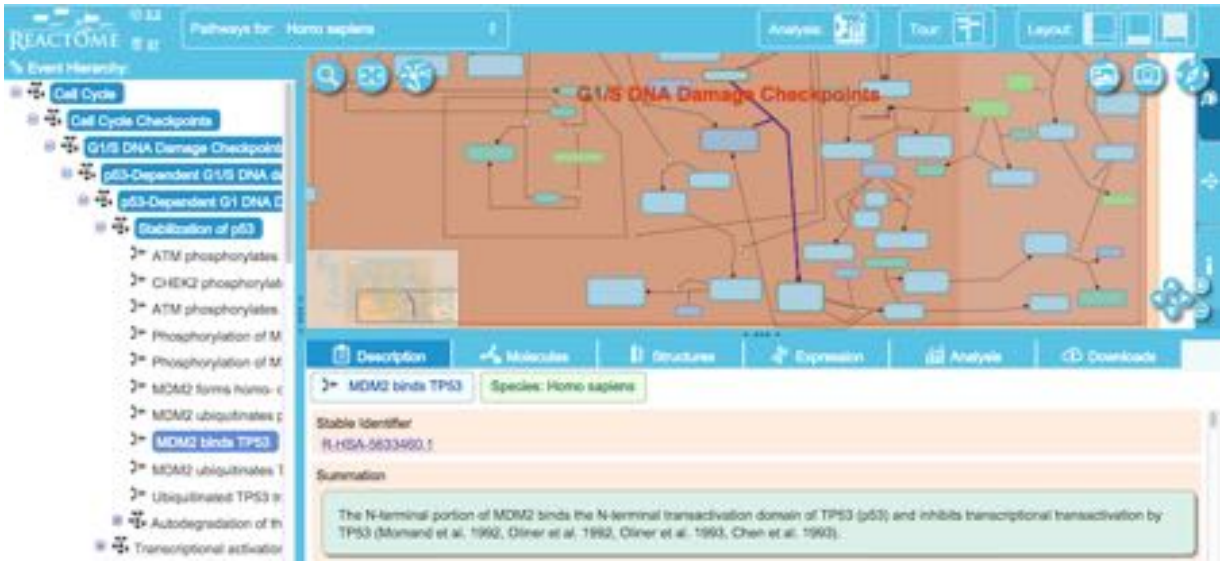


Reactome

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

http://www.reactome.org/

Reactome pathway diagram



Other pathway databases

- **PathwayCommons**, version 8 has over 42,000 pathways from 22 data sources, http://www.pathwaycommons.org/
- **PathGuide**, lists ~550 pathway related databases, http://www.pathguide.org/
- **WikiPathways**, community-curated pathways, http://wikipathways.org/
- **BioCarta**, pathway genes and diagrams, https://cgap.nci.nih.gov/Pathways/BioCarta_Pathways
- **Consensus-PathDB**, pathway interactions, enrichment, data, http://www.consensuspathdb.org/

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/

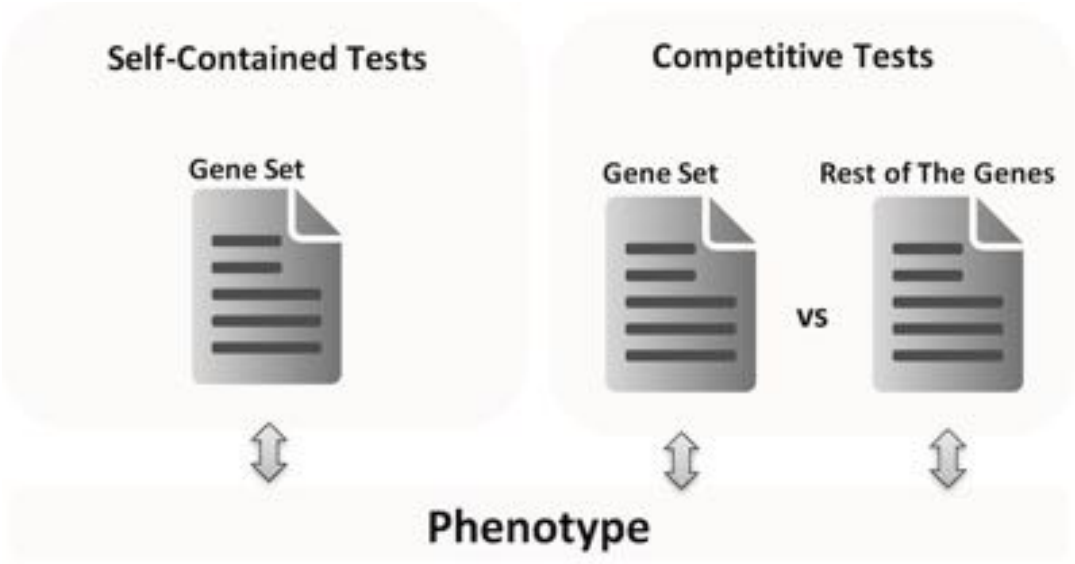
Overview

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

Enrichment analysis

Null hypothesis

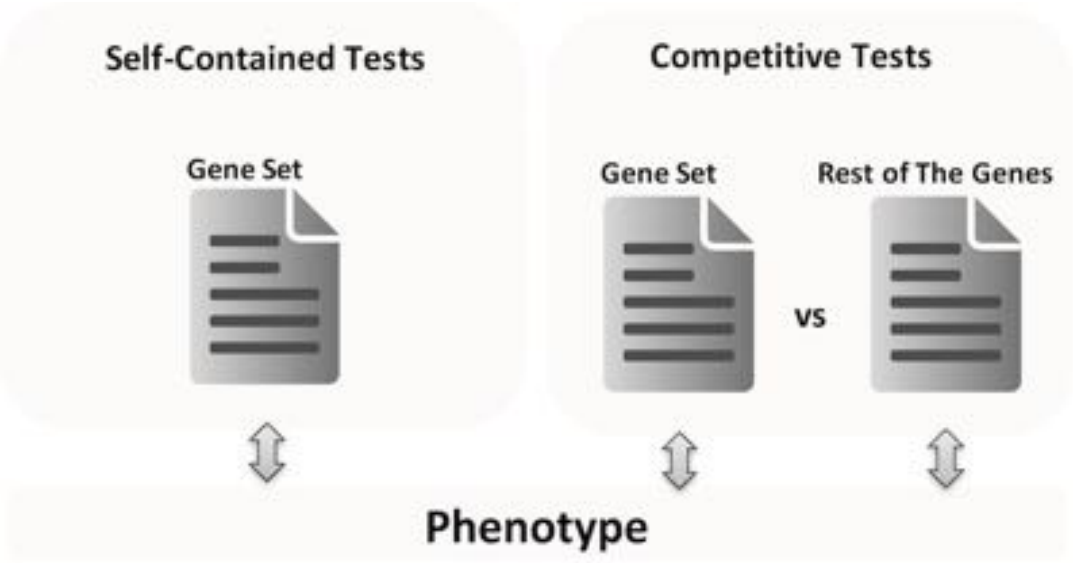
- **Self-contained** H_0: genes in the gene set do not have any association with the pheontype
- Problem: restrictive, use information only from a gene set



Enrichment analysis

Null hypothesis

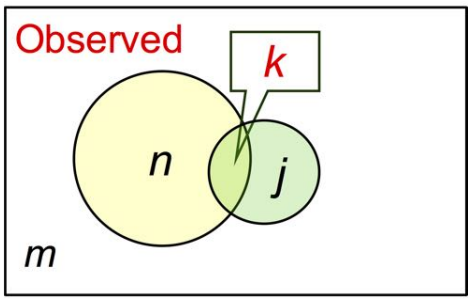
- **Competitive** H_0: genes in the gene set have the same level of association with a given phenotype as genes in the complement gene set
- Problem: wrong assumption of independent gene sampling



Approach 1

Overrepresentation analysis, 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



Approach 1

Overrepresentation analysis, 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 $r = k/k_e$

Approach 1

Overrepresentation analysis, 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

Approach 1

Overrepresentation analysis, 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?

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

Approach 1

Overrepresentation analysis, 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?

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

Approach 1

Overrepresentation analysis (ORA)

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

Approach 1

Problems with Fisher's exact 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

Correcting for pathway overlap

A

	DE	NDE	Total
P_i	n_i	m_i	$n_i + m_i$
P_i^c	$n - n_i$	$m - m_i$	$(n+m) - (n_i + m_i)$
Total	n	m	$n+m$

B

	DE	NDE	Total
$P_{i\setminus j}$	$n_{i\setminus j}$	$m_{i\setminus j}$	$n_{i\setminus j} + m_{i\setminus j}$
$P_{i\setminus j}^c$	$n - n_{i\setminus j}$	$m - m_{i\setminus j}$	$(n+m) - (n_{i\setminus j} + m_{i\setminus j})$
Total	n	m	$n+m$

Figure 9. A comparison of the classical overrepresentation analysis (*A*) with the crosstalk matrix analysis proposed here (*B*). (*A*) The standard overrepresentation approach contingency table: $n_i + m_i$ and $n + m$ represent, respectively, the number of genes belonging to pathway P_i and the total number of genes. n_i and n represent, respectively, the number of differentially expressed genes belonging to pathway P_i and the total number of DE genes. (*B*) Contingency table for the overrepresentation approach, taking into account the overlap between pairs of pathways; $P_{i\setminus j}$ represents the set of elements in P_i excluding the intersection with P_j; with the notations $n_{i\setminus j} + m_{i\setminus j}$ we represent the total number of genes that are in pathway P_i but not in pathway P_j, and with $n_{i\setminus j}$ the number of DE genes that are in pathway P_i but not in pathway P_j.

https://www.ncbi.nlm.nih.gov/pubmed/23934932

Many GO enrichment tools

- GOStat, http://gostat.wehi.edu.au/
- GOrilla, Gene Ontology enRIchment anaLysis and visuaLizAtion tool http://cbl-gorilla.cs.technion.ac.il/
- g:Profiler, http://biit.cs.ut.ee/gprofiler/
- Metascape, http://metascape.org/
- ToppGene, https://toppgene.cchmc.org/
- WebGestals - WEB-based GEne SeT AnaLysis Toolkit, http://www.webgestalt.org/
- GeneTrails2 - gene-, protein, miRNA, genomic enrichment analysis, https://genetrail2.bioinf.uni-sb.de/
- R packages, clusterProfiler, https://www.bioconductor.org/packages/devel/bioc/html/clusterProfiler.html

Approach 2

Functional Class Scoring (FCS)

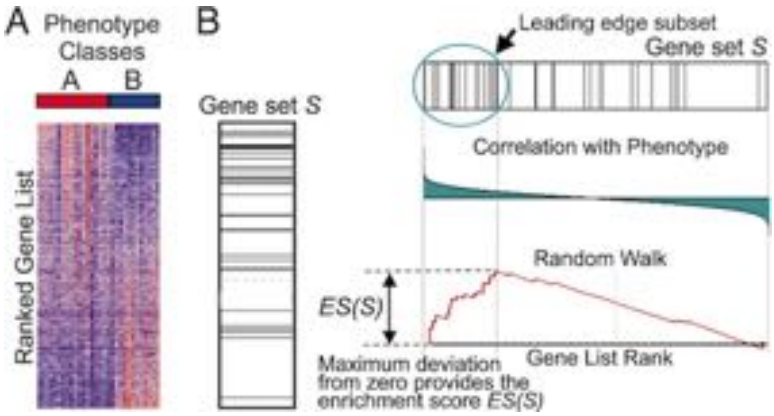
- **Gene set analysis (GSA)**. Mootha et al., 2003; modified by Subramanian, et al. "**Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles.**" PNAS 2005 http://www.pnas.org/content/102/43/15545.abstract
- Main rationale – functionally related genes 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 $R_1, \ldots, R_N$ 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

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

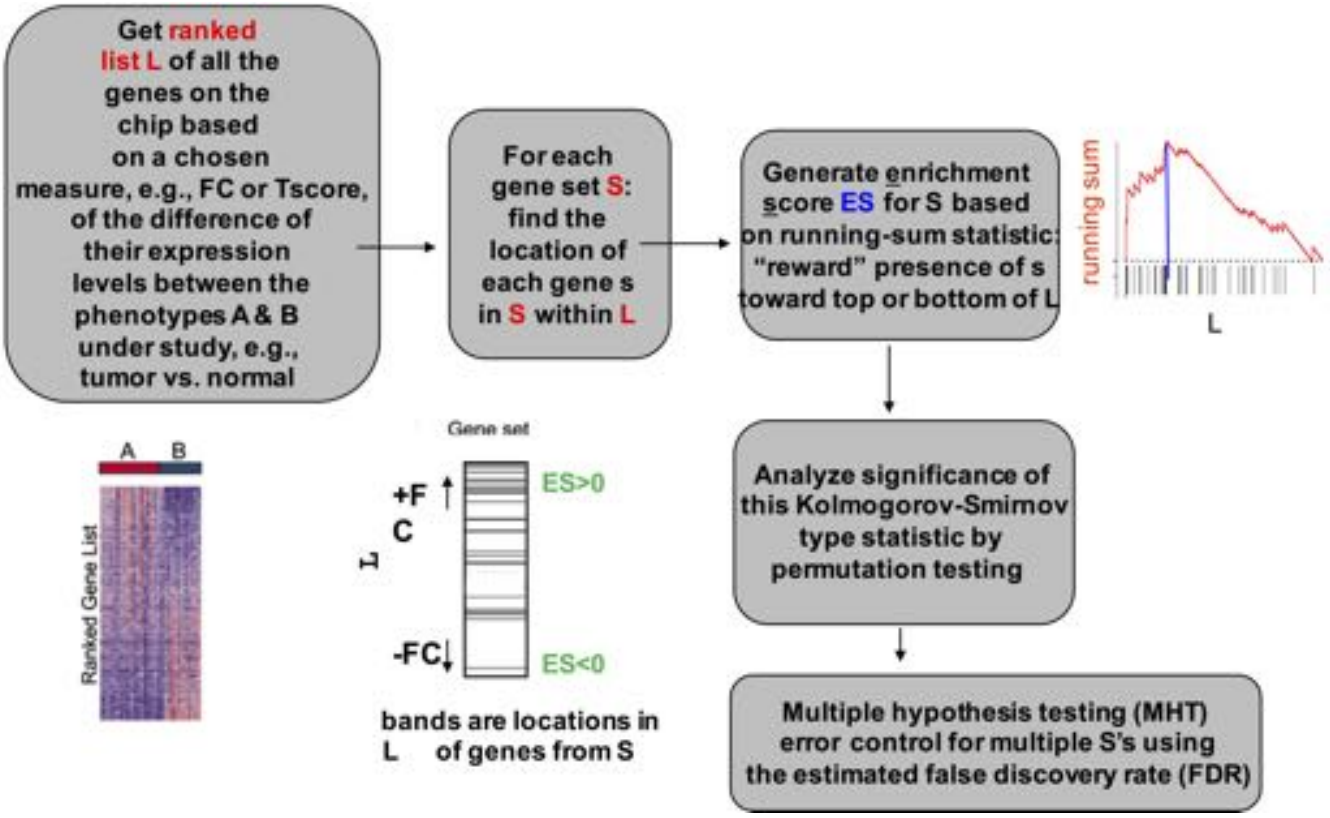
GSEA: Gene set enrichment analysis

Enrichment Score

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

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

- 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
running sum
L
A
B
Ranked Gene List
Gene set
+FC
-FC
ES>0
ES<0
L
bands are locations in L of genes from S
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)

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

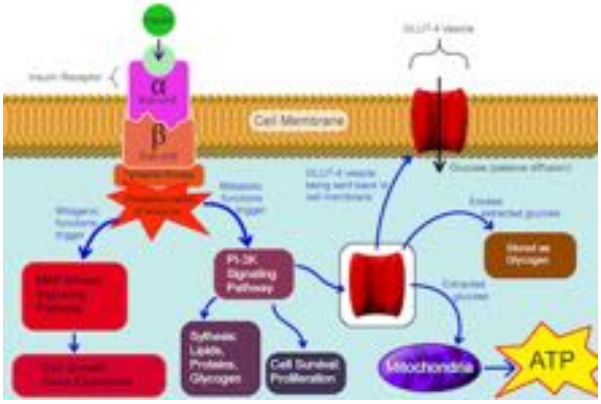
Linear model-based

- **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.

Other approaches

Impact analysis - incorporates topology of the pathway.

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



Other approaches

- **Pathway-Express**, http://vortex.cs.wayne.edu/projects.htm#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/release/bioc/html/SPIA.html

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

Overview

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

Gene set enrichment analysis

Web

- **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
- **FRY** (http://shiny.bioinf.wehi.edu.au/giner.g/FRY_GeneSetExplorerApp/) - Fast Interactive Biological Pathway Miner, from WEHI group

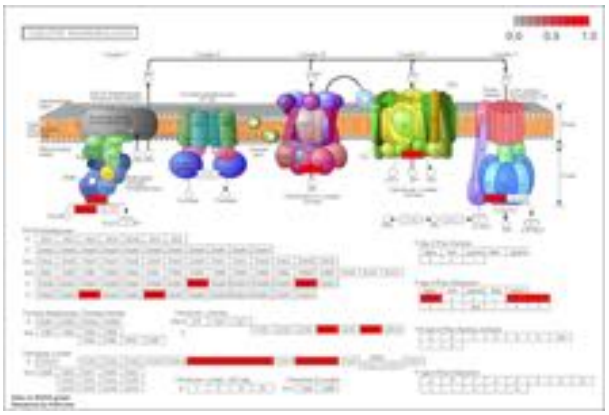
Gene set enrichment analysis

DIY

- **clusterProfiler** (https://bioconductor.org/packages/release/bioc/html/clusterProfiler.html) - statistical analysis and visualization of functional profiles for genes and gene clusters
- **limma** (https://bioconductor.org/packages/release/bioc/html/limma.html) - Linear Models for Microarray Data, includes functional enrichment functions `goana`, `camera`, `roast`, `romer`
- **GOstats** (https://www.bioconductor.org/packages/2.8/bioc/html/GOstats.html) - tools for manimpuating GO and pathway enrichment analyses. https://github.com/mdozmorov/MDmisc/blob/master/R/gene_enrichment.

Gene annotation databases

- **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://www.bioconductor.org/packages/devel/bioc/html/pathview.html) - a tool set for pathway based data integration and visualization



Genomic regions enrichment analysis

- **GREAT** predicts functions of cis-regulatory regions, http://bejerano.stanford.edu/great/public/html/
- **Enrichr**, gene- and genomic regions enrichment analysis tool, http://amp.pharm.mssm.edu/Enrichr/#
- **GenomeRunner**, Functional interpretation of SNPs (any genomic regions) within regulatory/epigenomic context, http://integrativegenomics.org/

Learn more

- Dave's blog (http://davetang.org/muse/) search for "Gene ontology enrichment analysis"
- Nam D., and Seon-Young K.. "**Gene-Set Approach for Expression Pattern Analysis.**" *Briefings in Bioinformatics* 2008 https://www.ncbi.nlm.nih.gov/pubmed/18202032
- Mutation Consequences and Pathway Analysis working group. "**Pathway and Network Analysis of Cancer Genomes.**" *Nature Methods* 2015 https://www.ncbi.nlm.nih.gov/pubmed/26125594
- Khatri, P. et.al. "**Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges.**" *PLoS Computational Biology* 2012 https://www.ncbi.nlm.nih.gov/pubmed/22383865
- de Leeuw, C. et.al. "**The Statistical Properties of Gene-Set Analysis.**" *Nature Reviews* 2016 https://www.ncbi.nlm.nih.gov/pubmed/27070863

FINE