

# Gene-set analysis and data integration

Leif Väremo

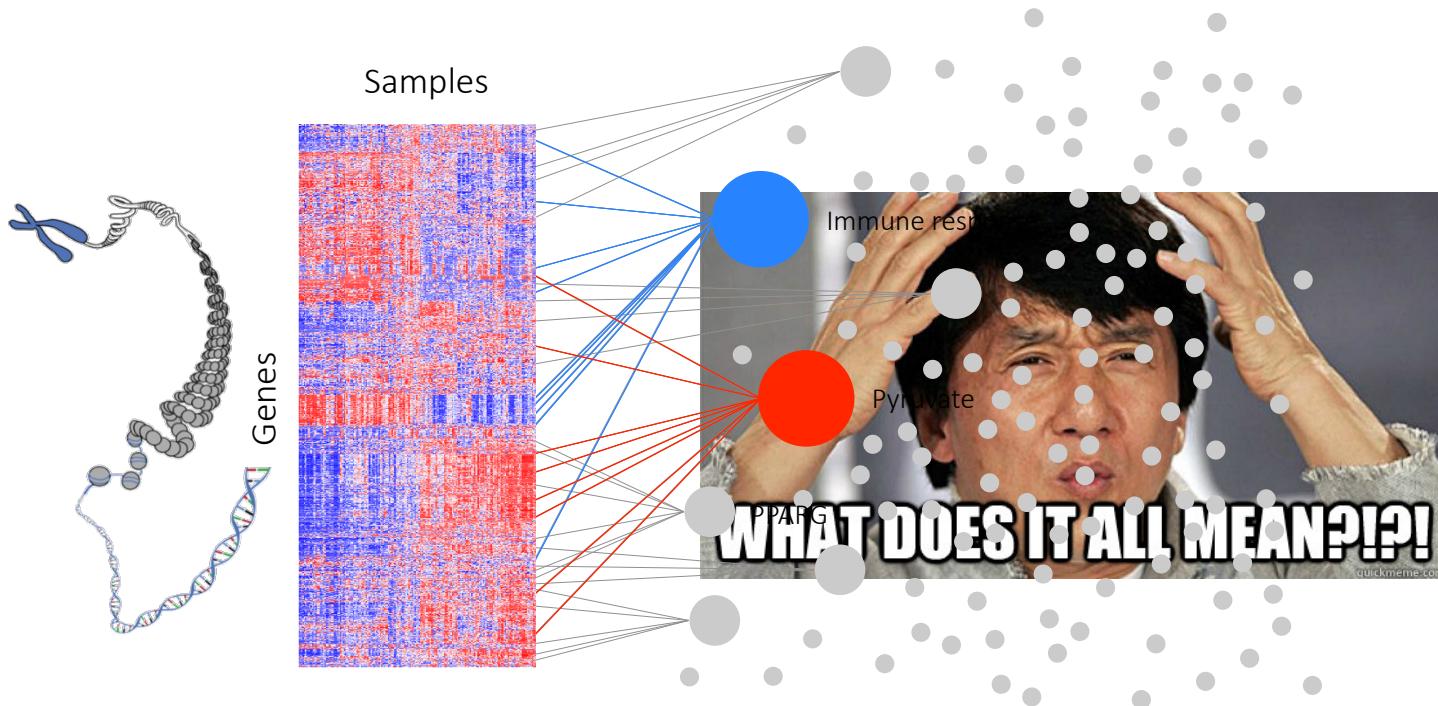
[leif.varemo@scilifelab.se](mailto:leif.varemo@scilifelab.se)

# Outline

- Gene-set analysis - What and why?
- Gene-set collections
- Methods for GSA
- A few words on gene-set directionality and overlap/interactions
- An example
- Things to consider

Will try to be practical, without getting to the detail of code-level

# What is gene-set analysis (GSA)?



GO-terms  
Pathways  
Chromosomal locations  
Transcription factors  
Histone modifications  
Diseases  
etc...



We will focus on transcriptomics and differential expression analysis  
However, GSA can in principle be used on all types of genome-wide data.

# Many names for gene-set analysis

- Pathway analysis
- Gene-set enrichment analysis
- GO-term analysis
- Gene list enrichment analysis
- ...

# Why gene-set analysis (GSA)?

- Interpretation of genome-wide results
- Gene-sets are (typically) fewer than all the genes and have more descriptive names
- Difficult to manage a long list of significant genes
- Integrates external information into the analysis
- Less prone to false-positives on the gene-level
- Top genes might not be the interesting ones, several coordinated smaller changes
- Detect patterns that would be difficult to discern simply by manually going through e.g. the list of differentially expressed genes

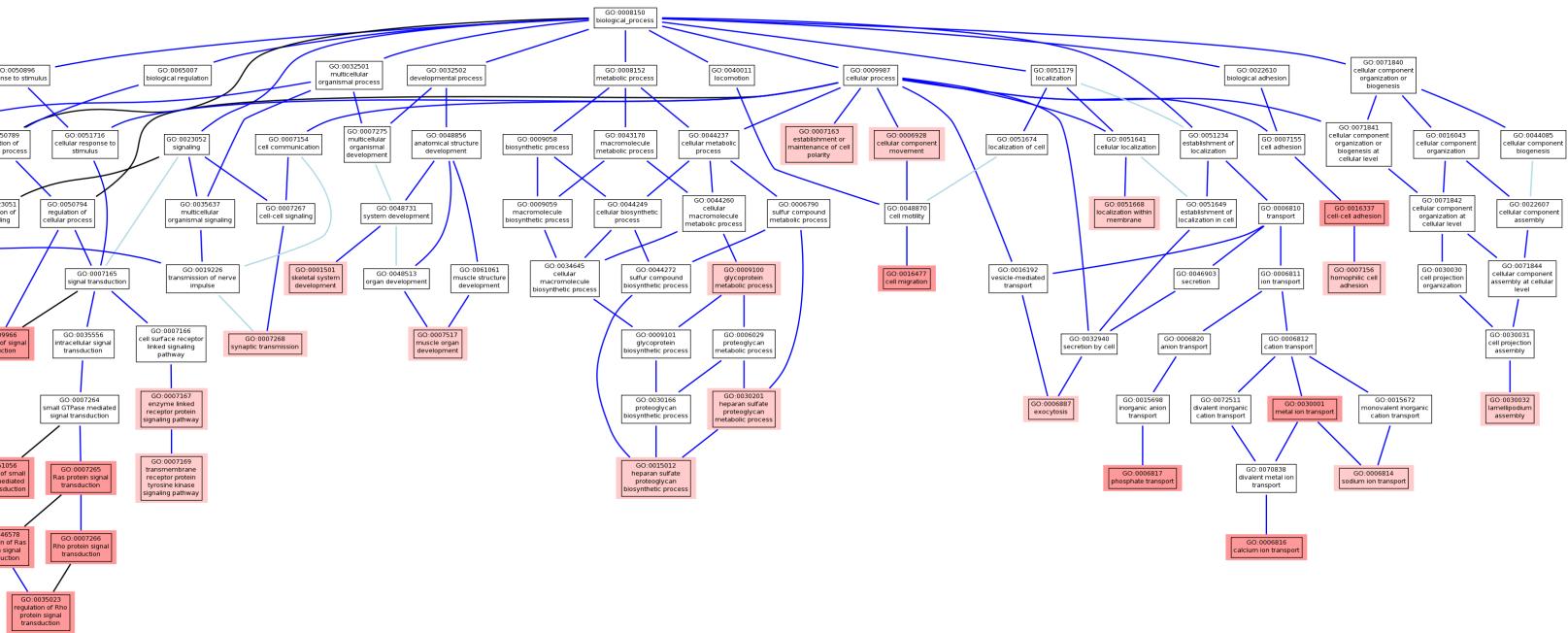
# Gene-sets

# So what about gene-sets?

- Depends on the research question
- Several databases/resources available providing gene-set collections (e.g. MSigDB, Enrichr)
- GO-terms are probably one of the most widely used gene-sets

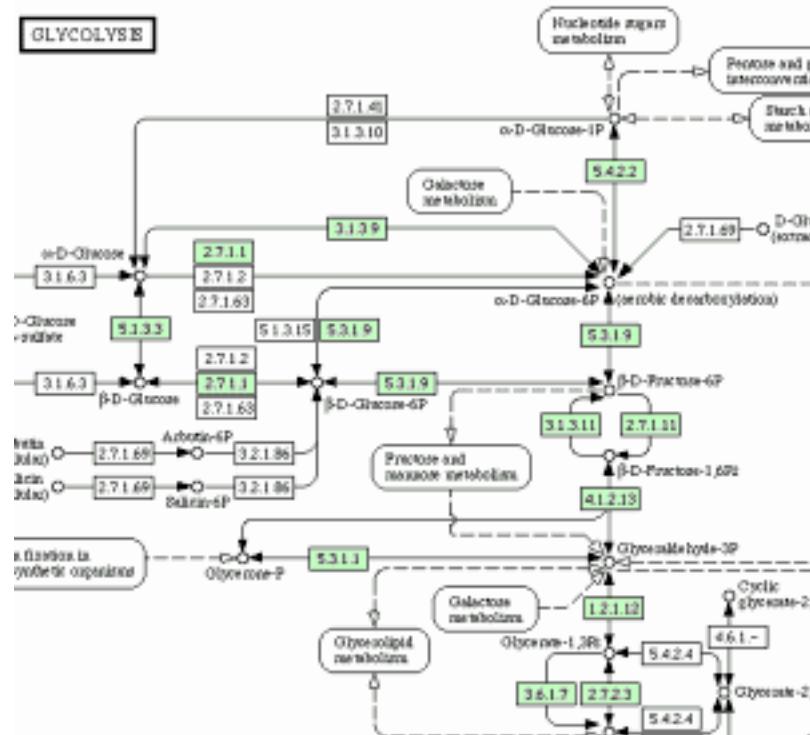
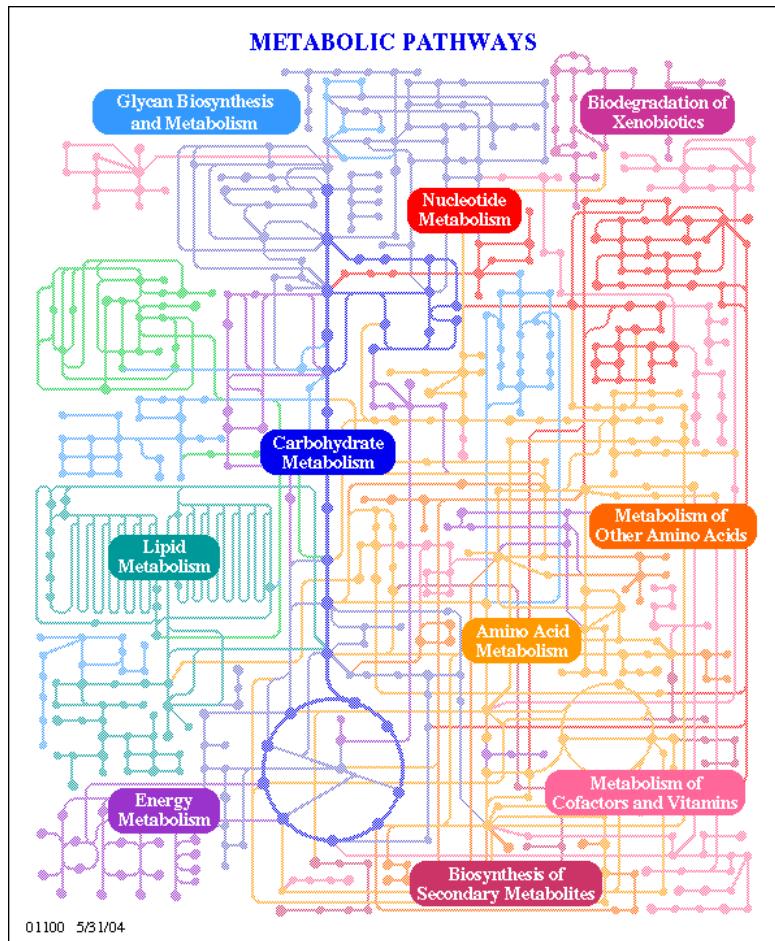
GO-terms  
Pathways  
Chromosomal locations  
Transcription factors  
Histone modifications  
Diseases  
Metabolites  
etc...

# Gene-set example: Gene ontology (GO) terms

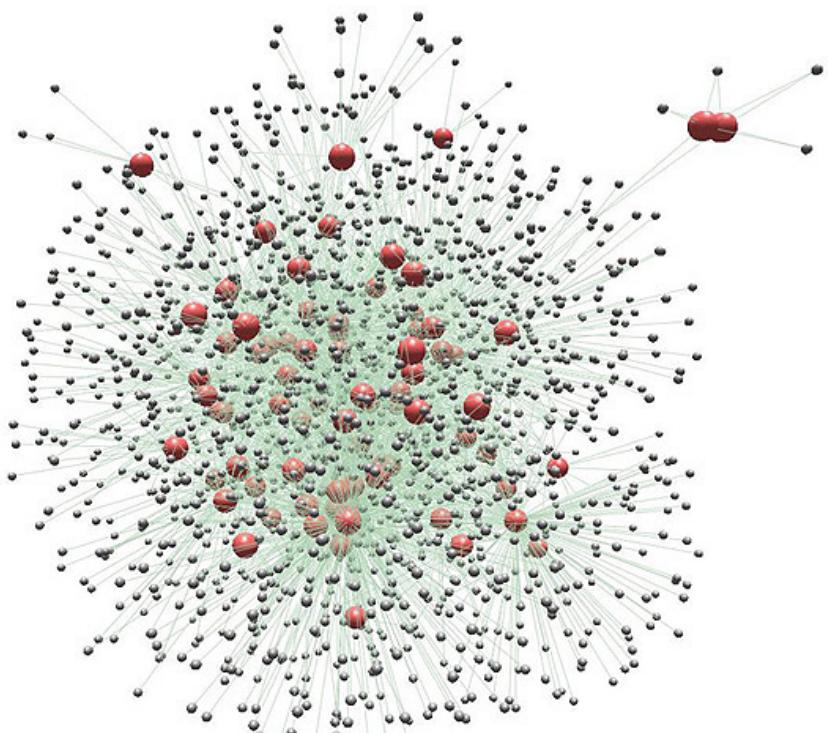
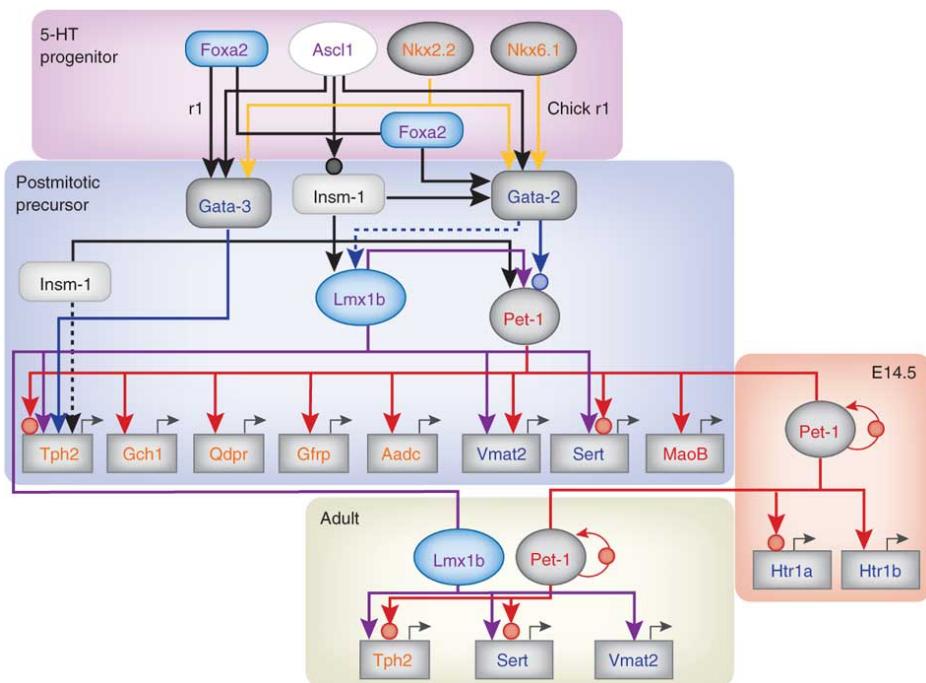


- Hierarchical graph with three categories (or parents): Biological process, Molecular function, Cellular compartment
- Terms get more and more detailed moving down the hierarchy
- Genes can belong to multiple GO terms

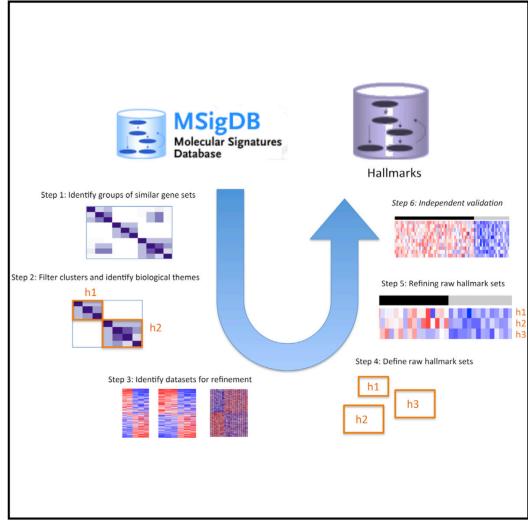
# Gene-set example: Metabolic pathways or metabolites



# Gene-set example: Transcription factor targets



# Gene-set example: Hallmark gene-sets



*"Hallmark gene sets summarize and represent specific well-defined biological states or processes and display coherent expression. These gene sets were generated by a computational methodology based on identifying gene set overlaps and retaining genes that display coordinate expression. The hallmarks reduce noise and redundancy and provide a better delineated biological space for GSEA."*

<http://software.broadinstitute.org/gsea/msigdb/collections.jsp>

Liberzon et al. (2015) Cell Systems 1:417-425

# Where to get gene-set collections?

<http://software.broadinstitute.org/gsea/msigdb/index.jsp>



## 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 appreciate contributions to this shared resource and encourage users to submit their gene sets to [genesets@broadinstitute.org](mailto:genesets@broadinstitute.org). Our thanks to our many contributors.

Funded by: National Cancer Institute, National Institutes of Health, National Institute of General Medical Sciences.

### Collections

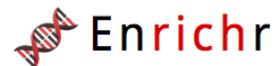
The MSigDB gene sets are divided into 8 major collections:

- 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.

### Citing the MSigDB

To cite your use of the Molecular Signatures Database (MSigDB), please reference Subramanian, Tamayo, et al. (2005, PNAS 102, 15545-15550) and also the source for the gene set as listed on the gene set page.

<http://amp.pharm.mssm.edu/Enrichr/#stats>



Analyze What's New? Libraries Find a Gene About Help

[Login](#) | [Register](#)

1,052,595 lists analyzed

Gene-set Library	Terms	Gene Coverage	Genes per Term
Achilles_fitness_decrease	216	4271	128.0
Achilles_fitness_increase	216	4320	129.0
Aging_Perturbations_from_GEO_down	286	16129	292.0
Aging_Perturbations_from_GEO_up	286	15309	308.0
Allen_Brain_Atlas_down	2192	13877	304.0
Allen_Brain_Atlas_up	2192	13121	305.0
BioCarta_2013	249	1295	18.0
BioCarta_2015	239	1678	21.0
BioCarta_2016	237	1348	19.0
Cancer_Cell_Line_Encyclopedia	967	15797	176.0
ChEA_2013	353	47172	1370.0
ChEA_2015	395	48230	1429.0
Chromosome_Location	386	32740	85.0
CORUM	1658	2741	5.0
dbGaP	345	5613	36.0
Disease_Perturbations_from_GEO_down	839	23939	293.0
Disease_Perturbations_from_GEO_up	839	23561	307.0
Disease_Signatures_from_GEO_down_2014	142	15406	300.0
Disease_Signatures_from_GEO_up_2014	142	15057	300.0
Drug_Perturbations_from_GEO_2014	701	47107	509.0
Drug_Perturbations_from_GEO_down	906	23877	302.0
Drug_Perturbations_from_GEO_up	906	24350	299.0
ENCODE_and_ChEA_Consensus_TFs_from_ChIP-X	104	15562	887.0
ENCODE_Histone_Modifications_2013	109	15852	912.0
ENCODE_Histone_Modifications_2015	412	29065	2123.0
ENCODE_TF_ChIP-seq_2014	498	21493	3713.0
ENCODE_TF_ChIP-seq_2015	816	26382	1811.0
Epigenomics_Roadmap_HM_ChIP-seq	383	22288	4368.0
ESCAPE	315	25651	807.0
Genes_Associated_with_NIH_Grants	32876	15886	9.0
GeneSigDB	2139	23726	127.0
Genome_Browser_PWMs	615	13362	275.0

# Where to get gene-set collections?

- Sooner or later you will run into the problem of matching your data to gene-set collections due to the existence of several gene ID types

```
protein secretion (GO:0009306)      NECAB3  PDIA4  ABCA1  PLEK   NLRC4  LTBP2  PCSK5  ARFGAP3  ARL4D  BACE2  CANX
rRNA transcription (GO:0009303)      GTF3C2  GTF3C3  GTF3C4  GTF3C5  GTF3C6  RNASEK  BRF1   GTF3A   CD3EAP  MKI67IP  GTF3C1
positive regulation of DNA replication (GO:0045740)  INSR    PDGFRα  EPO    TGFB3  SHC1    PLA2G1B  CSF2   TNKS
respiratory burst (GO:0045730)       CD52    NCF2    PGAM1  CYBB   CYBA    NCF1   NOX1   CD24    CD55
positive regulation of protein catabolic process (GO:0045732)  ECLN2  FURIN   HDAC2  F12    TNF    SMAD7  CLN6
positive regulation of DNA repair (GO:0045739)      PRKCG  EYA1    MERIT40 EYA3   CEBPG  H2AFX  BRCC3  BRCA1   RNF8
negative regulation of adenylate cyclase activity (GO:0007194)  CCR2   GABBR2  GABBR1 NPY1R  OPRK1  ADRA2A CORT
DRD2   DRD3   DRD4
inhibition of adenylate cyclase activity by G-protein signaling (GO:0007193)
regulation of transcription factor activity (GO:0051090)      ILL10  NFAM1  SIRT1  PEX14  AGT    SMARCA4 FOXP3
TNF    NLRC3  MTDH   PYCARD  ABRA   STK36  IRAK2  IRAK3  IRAK1  FLNA   NLRP3  RPS3   RIPK1  CARD11 EGLN1  NPM1
BCL10  EDA2R  CREBZF  IKBKB   PRDX3  SUMO1  EP300  ERC1   TNFRSF4 IL6R   MEN1
activation of adenylate cyclase activity (GO:0007190)      CAP2   NTRK2  CAP1   CRHR1  GIPR   P2RY11 NTRK1  AVPR2
positive regulation of transcription factor activity (GO:0051091)  CARD11 NPM1   IL10   NFAM1  ACT    SMARCA4
NOD2   TNF    EDA2R  NLRC3  MTDH   PYCARD  IKBKB  ABRA   PRDX3  IRAK3  EP300  IRAK1  ERC1   RIPK1  IL6R
positive regulation of NF-κappaB transcription factor activity (GO:0051092)  CARD11 NPM1   AGT    IL1B   IL6
PRDX3  IRAK3  IRAK1  ERC1   RIPK1  IL6R
```

```
> head(res)
log2 fold change (MAP): timepoint t24h vs ctrl
Wald test p-value: timepoint t24h vs ctrl
DataFrame with 6 rows and 6 columns
  baseMean log2FoldChange      lfcSE      stat     pvalue      padj
  <numeric>      <numeric>      <numeric>      <numeric>      <numeric>      <numeric>
ENSG000000000003 488.9141058  0.89327988  0.10613362  8.4165589 3.877042e-17 3.077290e-16
ENSG00000000419 816.5442744 -0.19601877  0.09887579 -1.9824748 4.742612e-02 8.740280e-02
ENSG00000000457 81.9349878   0.30293405  0.20363836  1.4876080 1.368543e-01 2.182234e-01
ENSG00000000460 355.7964356 -1.83662295  0.12101968 -15.1762333 5.081360e-52 1.569737e-50
ENSG00000000971  0.5328727 -0.02963864  0.28670478 -0.1033769 9.176639e-01 9.460059e-01
ENSG00000001036 918.3238933 -0.35428837  0.08228014 -4.3058795 1.663236e-05 5.415768e-05
> |
```

# Where to get gene-set collections?

<http://www.ensembl.org/biomart/martview>

**e!Ensembl** BLAST/BLAT | BioMart | Tools | Downloads | Help & Documentation | Blog | Mirrors

New Count Results URL XML Per

**Dataset**  
Homo sapiens genes (GRCh38.p5)

**Filters**  
[None selected]

**Attributes**  
Ensembl Gene ID  
GO Term Name  
Associated Gene Name  
EntrezGene ID

Export all results to File TSV Unique results  
Email notification to

View 200 rows as HTML Unique results only

ENSG00000198763	respiratory electron transport chain	MT-ND2	4536
ENSG00000198763	NADH dehydrogenase (ubiquinone) activity	MT-ND2	4536
ENSG00000198763	mitochondrial electron transport, NADH to ubiquinone	MT-ND2	4536
ENSG00000198763	mitochondrial inner membrane	MT-ND2	4536
ENSG00000198763	cellular metabolic process	MT-ND2	4536
ENSG00000198763	oxidation-reduction process	MT-ND2	4536
ENSG00000198763	integral component of membrane	MT-ND2	4536
ENSG00000198763	mitochondrion	MT-ND2	4536
ENSG00000198763	reactive oxygen species metabolic process	MT-ND2	4536
ENSG00000198763	protein kinase binding	MT-ND2	4536
ENSG00000198763	ionotropic glutamate receptor binding	MT-ND2	4536
ENSG00000198763	postsynaptic density	MT-ND2	4536
ENSG00000198804	respiratory chain complex IV	MT-CO1	4512
ENSG00000198804	aerobic respiration	MT-CO1	4512
ENSG00000198804	oxidative phosphorylation	MT-CO1	4512
ENSG00000198804	gene expression	MT-CO1	4512
ENSG00000198804	small molecule metabolic process	MT-CO1	4512
ENSG00000198804	cytochrome-c oxidase activity	MT-CO1	4512
ENSG00000198804	protein binding	MT-CO1	4512

One way to map different gene IDs to each other, or to assemble a gene-set collection with the gene IDs used by your data

# Gene-set analysis

# Tools and methods for GSA

## OmicsTools (several platforms)

<http://omictools.com/gene-set-analysis-category>

The screenshot shows the OmicsTools website interface. At the top, there's a navigation bar with 'OMIC TOOLS' logo, search bar, and 'Join community' button. Below the navigation, a breadcrumb trail shows 'DIRECTORY > TRANSCRIPTOMICS > GENE SET ANALYSIS'. A note about accepting the end-user license agreement is present. The main content area is titled 'Gene set analysis software tools | Transcriptomics' and includes a sub-section 'Transcriptomic software tools and databases'.

## Bioconductor (R packages)

[https://bioconductor.org/packages/release/BiocViews.html#\\_GeneSetEnrichment](https://bioconductor.org/packages/release/BiocViews.html#_GeneSetEnrichment)

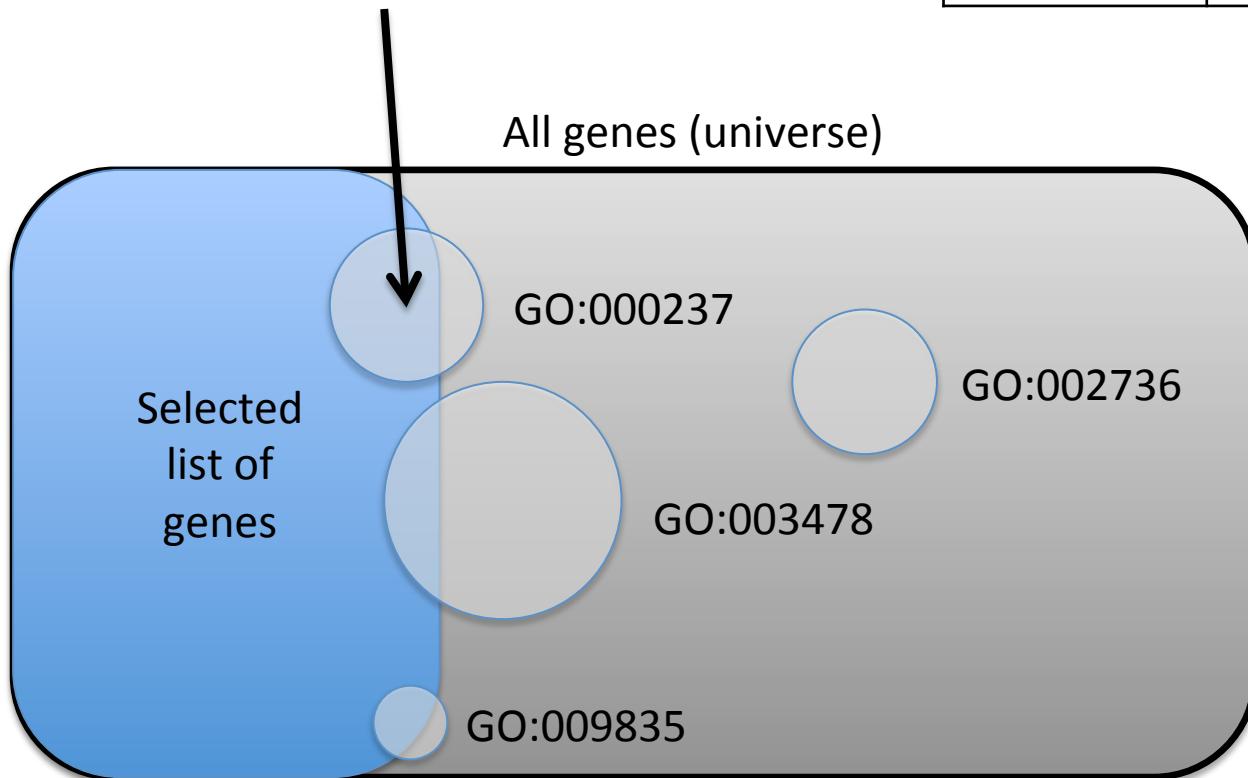
The screenshot shows the Bioconductor website. It features a teal header with the 'Bioconductor OPEN SOURCE SOFTWARE FOR BIOINFORMATICS' logo and navigation links for 'Home', 'Install', 'Help', 'Developers', and 'About'. Below the header, a breadcrumb trail shows 'Home > BiocViews'. The main content area displays the 'BiocViews' page.

- Hypergeometric test / Fisher's exact test (a.k.a overrepresentation analysis)
- DAVID (browser)
- Enrichr (browser)
- GSEA (Java, R)
- Piano (R)

# Overrepresentation analysis

Hypergeometric test  
(Fisher's exact test)  
Is this overlap  
bigger than  
expected by  
random chance?

	Selected	Not selected
In GO-term	8	2
Not in GO-term	92	19768



# Overrepresentation analysis

<http://amp.pharm.mssm.edu/Enrichr/>

 **Enrichr**

Login | Register  
1,052,888 lists analyzed

Analyze What's New? Libraries Find a Gene About Help

## Input data

Choose an input file to upload. Either in BED format or a list of genes. For a quantitative set, add a comma and the level of membership of that gene. The membership level is a number between 0.0 and 1.0 to represent a weight for each gene, where the weight of 0.0 will completely discard the gene from the enrichment analysis and the weight of 1.0 is the maximum.

Try an example BED file.

no file selected

Or paste in a list of gene symbols optionally followed by a comma and levels of membership. Try two examples: [crisp set example](#), [fuzzy set example](#)

```
Nsun3
Polrmt
Nr1x1
Sfxn5
Zc3h12c
Slc25a39
Arsg
Defb29
Ndub6
Zfand1
```

375 gene(s) entered

Enter a brief description for the list in case you want to share it. (Optional)

Contribute

Please acknowledge Enrichr in your publications by citing the following reference:  
Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, Clark NR, Ma'ayan A. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics*. 2013;128(14).

<https://david.ncifcrf.gov/home.jsp>

 **Gene Functional Classification Tool**  
DAVID Bioinformatics Resources 6.7, NIAID/NIH

Home Start Analysis Shortcut to DAVID Tools Technical Center Downloads & APIs Term of Service Why DAVID? About Us

**Upload Gene List**

Demolist 1 Demolist 2  
Upload Help

**Step 1: Enter Gene List**  
A: Paste a list

```
Clear
```

Or  
B: Choose From a File  
Choose File no file selected  
Multi-List File

**Step 2: Select Identifier**  
AFFYMETRIX\_3PRIME\_IVT\_ID

**Step 3: List Type**  
Gene List  
Background

**Step 4: Submit List**  
Submit List

**Gene Functional Classification Tool**

Submit your gene list to start the tool!

Tell us how you like the tool  
Read technical notes of this tool  
Contact us for questions

**What does this tool do?**

- Classify large gene list into functional related gene groups
- Rank the importance of the discovered gene groups
- Summarize the major biology of the discovered gene groups
- Search other functionally related genes from genome, but not in your list
- Visualize genes and their functional annotations in a group by a single 2-D view
- Explore global view of gene groups in a Fuzzy Heat Map visualization
- More

**The advantage of the tool: A novel gene-centric annotation approach**

- Your genes are highly organized so that they are more readable and understandable.
- Your genes are ranked so that you can quickly focus on the most likely important ones.
- Your genes are displayed with their annotation in one single view so that you can cross compare them.
- Your genes can be extended so that you have chance to know other functionally related genes, but not in your list.

**Rational Concepts:**

Grouping genes based on functional similarity can systematically enhance biological interpretation of large lists of genes derived from high throughput studies. The Functional Classification Tool generates a gene-to-term similarity matrix based on shared functional annotation using over 75,000 terms from 14 functional annotation sources. Our novel clustering algorithms classifies highly related genes into functionally related groups.

Tools are provided to further explore each functional gene cluster including listing of the consensus terms? shared by the genes in the cluster, display of enriched terms, and heat map visualization of gene-to-term relationships. A global view of cluster-to-cluster relationships is provided using a fuzzy heat map visualization. Summary information provided by the Functional Classification Tool is extensively linked to DAVID Functional Annotation Tools and to external databases allowing further detailed exploration of gene and term information.

The Functional Classification Tool provides a rapid means to organize large lists of genes into functionally related groups to help unravel the biological content captured by high throughput technologies.[more](#)

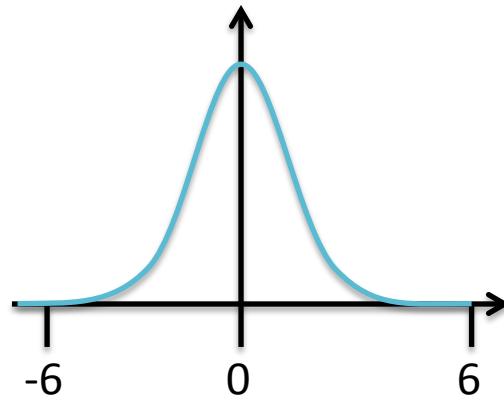
**Fuzzy Heuristic Partition**

We developed a novel heuristic partitioning procedure that allows an object (gene) to participate in more than one cluster. The use of this method in grouping related genes much better reflects the nature of biology in that a given gene may be associated with more than functional group of genes. Two additional advancement included in this algorithm are: 1) the automatic determination of the optimal numbers of clusters (K), and 2) the exclusion of members (genes) that have weak relationships to other members. Users are permitted to change default parameters to set cluster membership similarity stringencies. Fuzzy Heuristic Partitioning of a gene list yields high quality clusters of highly related genes, with some genes participating in more than one function cluster. [more](#)

# Overrepresentation analysis

- Requires a cutoff (arbitrary)
- Omits the actual values of the gene-level statistics
- Good for e.g. overlap of significant genes in two comparisons
- Computationally fast
- In general, it is recommended to use some kind of gene-set analysis. This will use all gene-level data and can detect small but coordinate changes that collectively contribute to some biological process

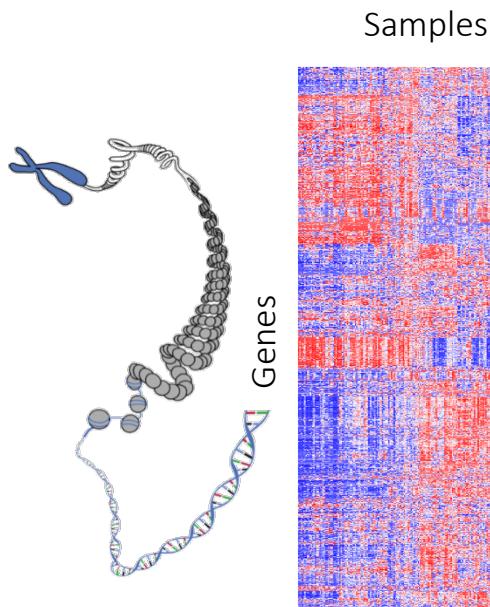
$S \downarrow$ permuted



## GSA: a simple example

- $S$  is the gene-set statistic
- $G$  are gene-level statistics of the genes in the gene-set

$$S \downarrow i = \text{mean}(G \downarrow i)$$



Gene-set 1

$$S \downarrow 1 = -0.1$$

Gene-set 2

$$S \downarrow 2 = 6.2$$

Permute the gene-labels (or sample labels) and redo the calculations over and over again (e.g. 10,000 times)!

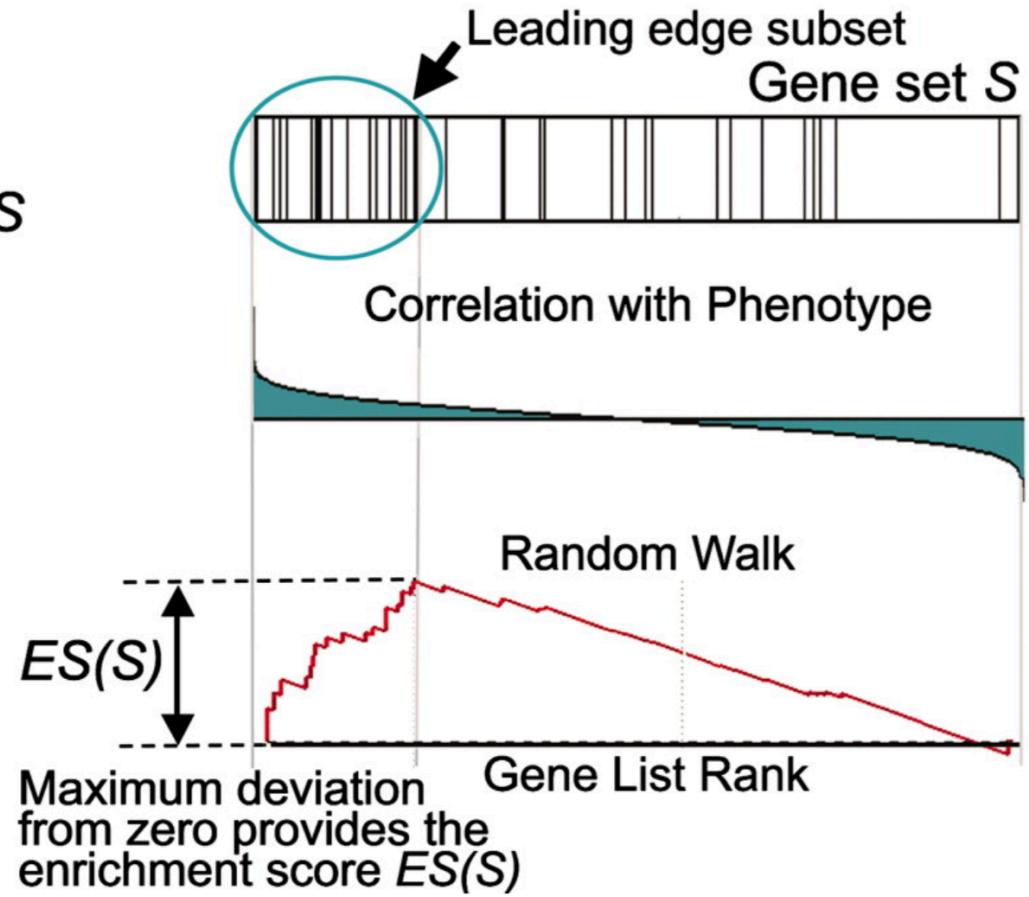
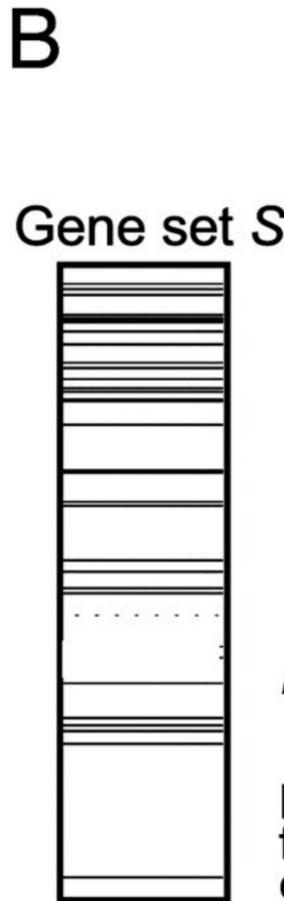
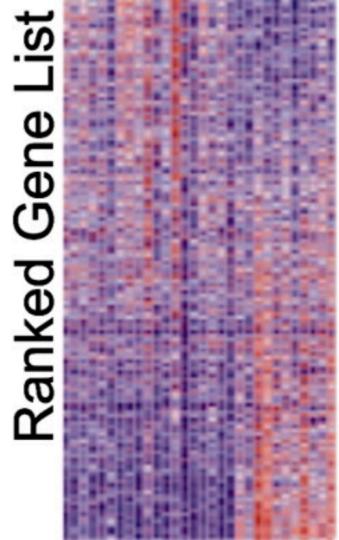
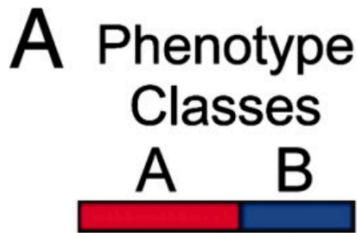
$$p \downarrow i = \text{fraction of } S \downarrow \text{permuted} \text{ that is more extreme than } S \downarrow i$$

# Gene-level statistics

- P-values
- T-values, etc
- Fold-changes
- Correlations
- Signal to noise ratio
- ...

# GSEA

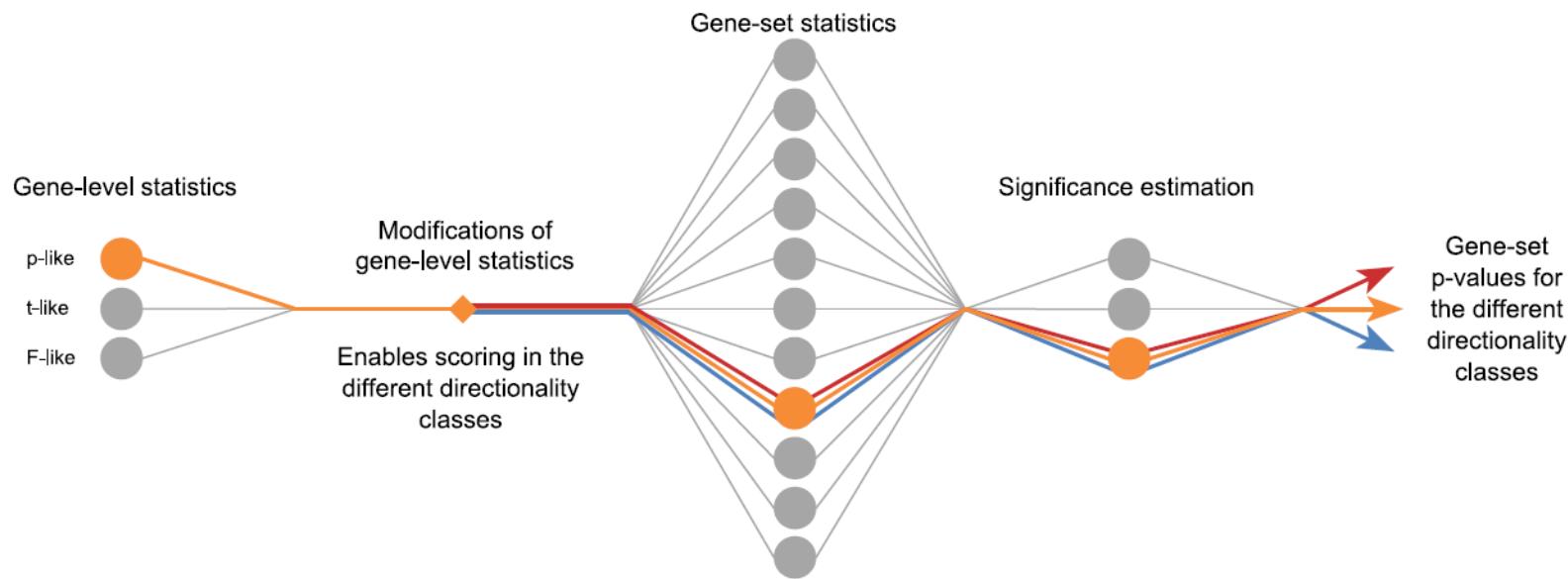
Mootha et al Nature Genetics, 2003; Subramanian PNAS 2005



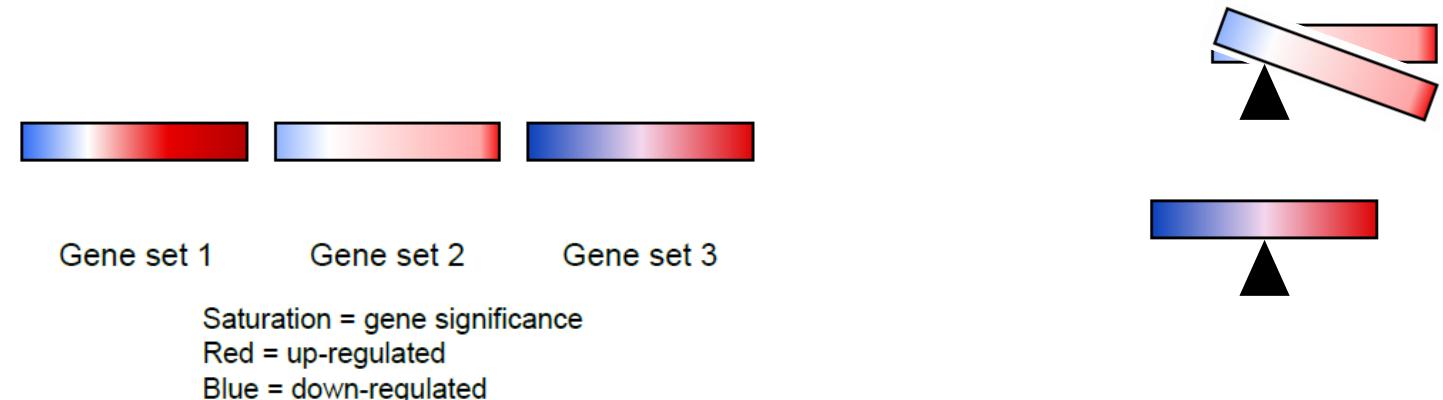
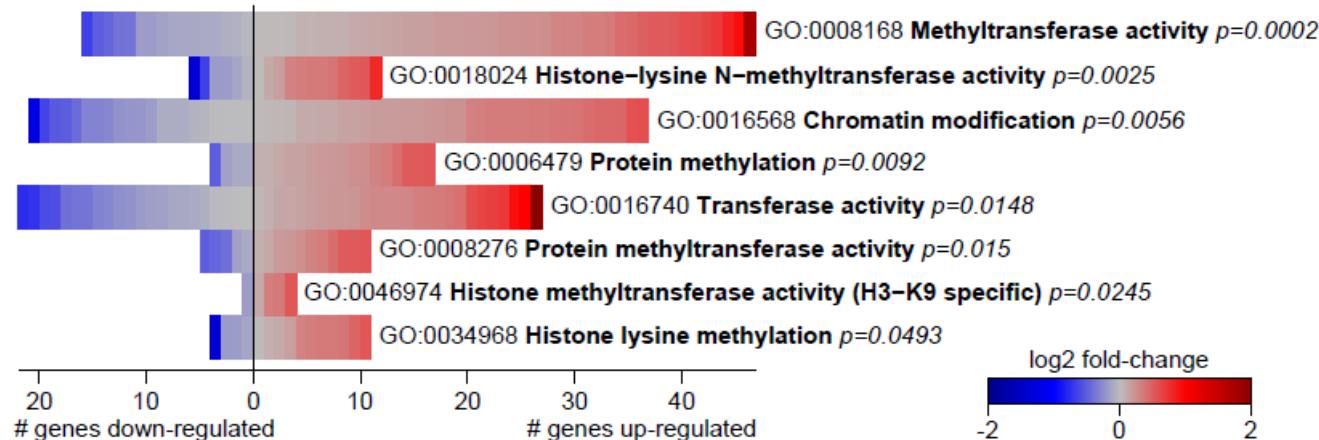
# Piano – a platform for gene-set analysis (in R)

- Reporter features
- Parametric analysis of gene-set enrichment, PAGE
- Tail strength
- Wilcoxon rank-sum test
- Gene-set enrichment analysis, GSEA
- Mean
- Median
- Sum
- Maxmean

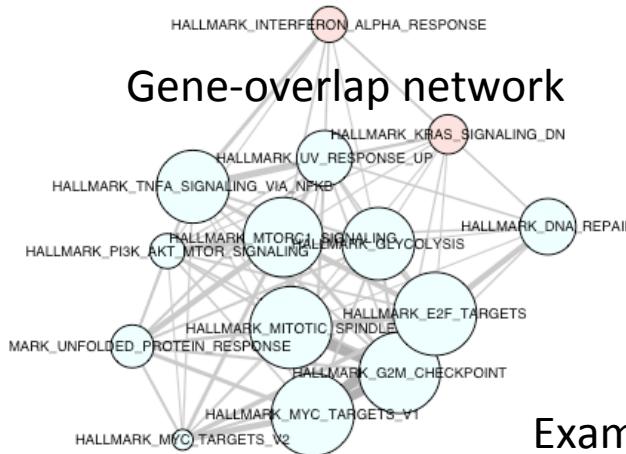
Consensus result



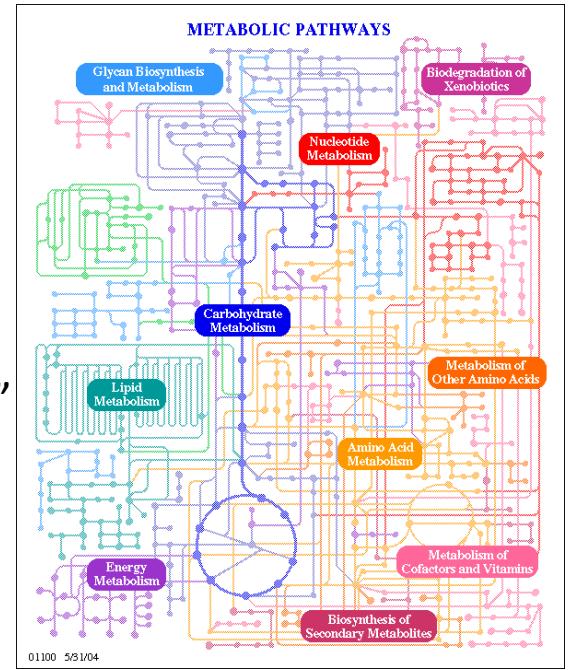
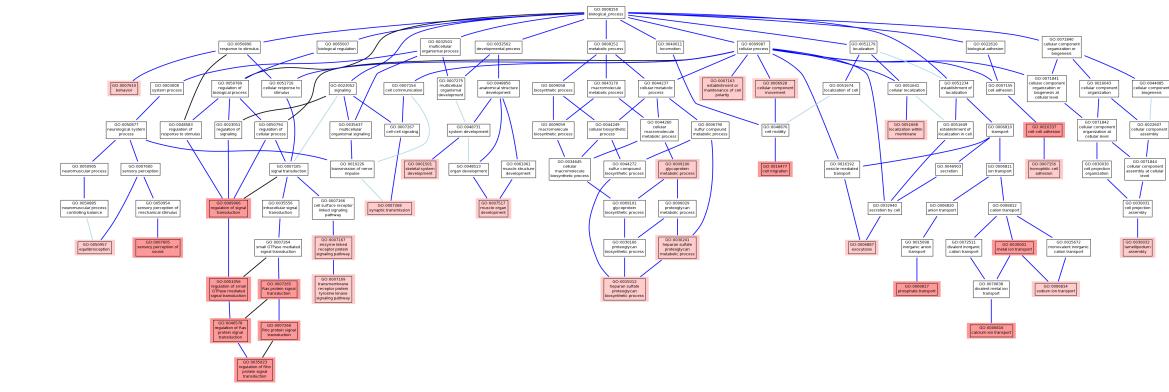
# Directionality of gene-sets



# Gene-set overlap and interaction

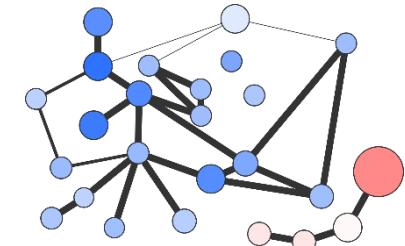
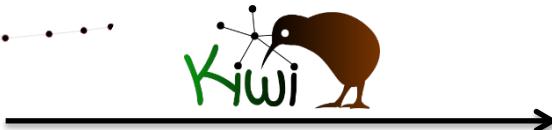
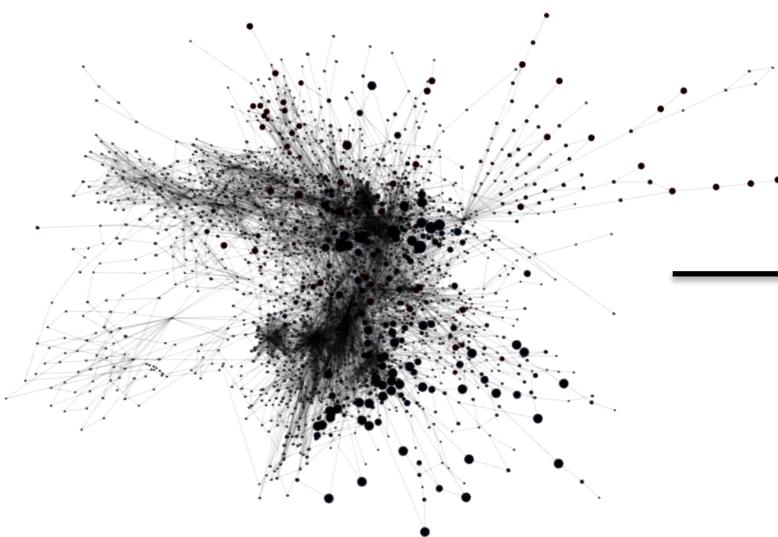
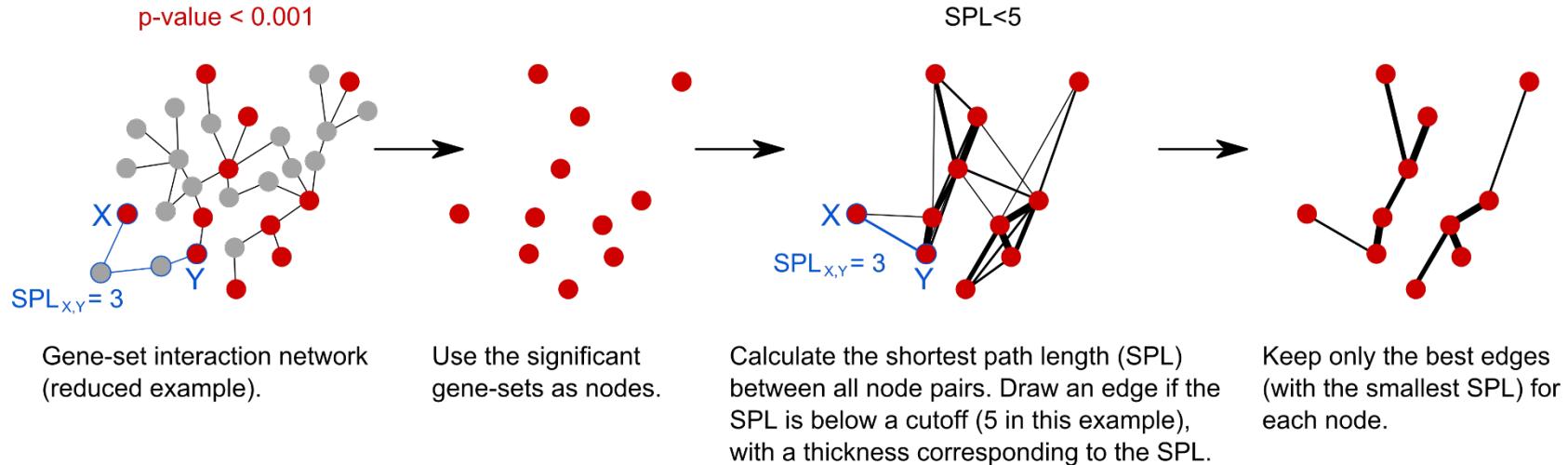


Examples of gene-set “interactions”

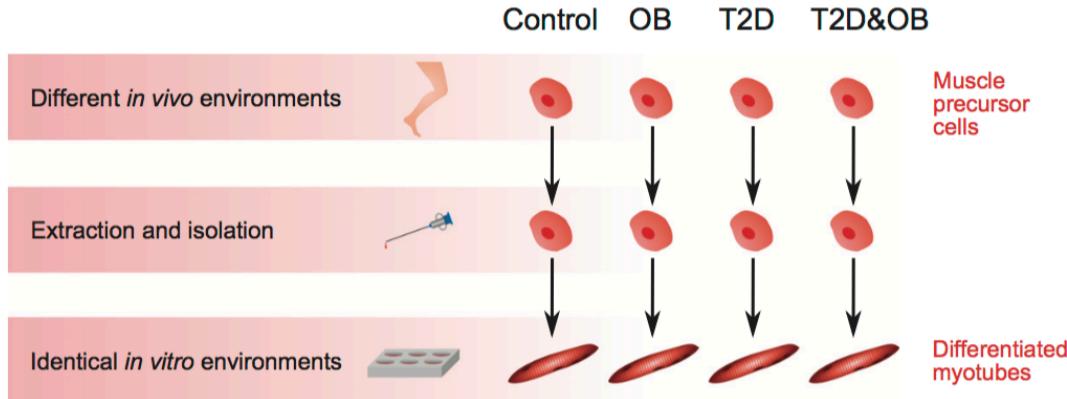


- High number of very overlapping gene-sets (representing a similar biological theme) can bias interpretation and take attention from other biological themes that are represented by fewer gene-sets.
- Can be valuable to take gene-set interaction into account

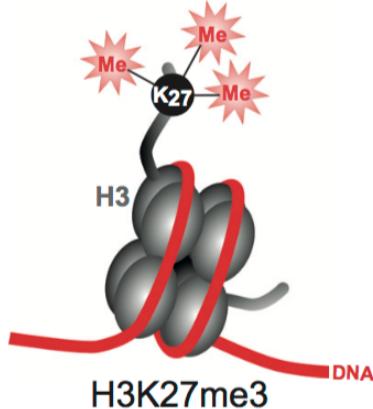
# Exploiting the gene-set interaction network



# Example



Using GSA of histone modification gene-sets to pinpoint a candidate epigenetic mechanism behind observed transcriptional changes.



	T2D	OB	T2D&OB
H3K27ac	Green		
Colon Smooth Muscle		Blue	
H3K27me3	Green		
Duodenum Smooth Muscle			Blue
Rectal Smooth Muscle			Blue
Skeletal Muscle			Blue
Stomach Smooth Muscle	Green	Blue	
H3K36me3			
Colon Smooth Muscle			Blue
Duodenum Smooth Muscle			Blue
Muscle Satellite Cultured Cells			Blue
Rectal Smooth Muscle			Blue
Skeletal Muscle			Blue
Stomach Smooth Muscle			Blue
H3K4me1			
Colon Smooth Muscle			Blue
Duodenum Smooth Muscle			Blue
Muscle Satellite Cultured Cells			Blue
Skeletal Muscle			Blue
Stomach Smooth Muscle			Blue
H3K4me2			
Muscle Satellite Cultured Cells			Blue
H3K4me3			
Colon Smooth Muscle			Blue
Duodenum Smooth Muscle			Blue
Rectal Smooth Muscle			Blue
Skeletal Muscle			Blue
Stomach Smooth Muscle			Blue
H3K9ac			
Colon Smooth Muscle			Blue
Muscle Satellite Cultured Cells			Blue
Rectal Smooth Muscle			Blue
Skeletal Muscle			Blue
Stomach Smooth Muscle			Blue
H3K9me3			
Colon Smooth Muscle			Blue
Duodenum Smooth Muscle			Blue
Rectal Smooth Muscle			Blue
Skeletal Muscle			Blue
Stomach Smooth Muscle			Blue

# Considerations when performing GSA

- Bias in gene-set collections
- Gene-set names can be misleading (revisit the genes!)
- Consider the gene-set size, i.e. number of genes (specific or general)
- Positive and negative association between genes and gene-sets makes gene-level fold-changes tricky to interpret correctly
- (Typically) binary association to gene-sets, does not take into account varying levels of influence from individual genes on the process that is represented by the gene-sets
- Remember to revisit the gene-level data! In particular if a permutation based approach is used for gene-set significance calculation. Are the genes significant? Are they correctly assigned to the specific gene-set?
- Remember to adjust for multiple testing