Gene-set analysis and data integration

Leif Väremo

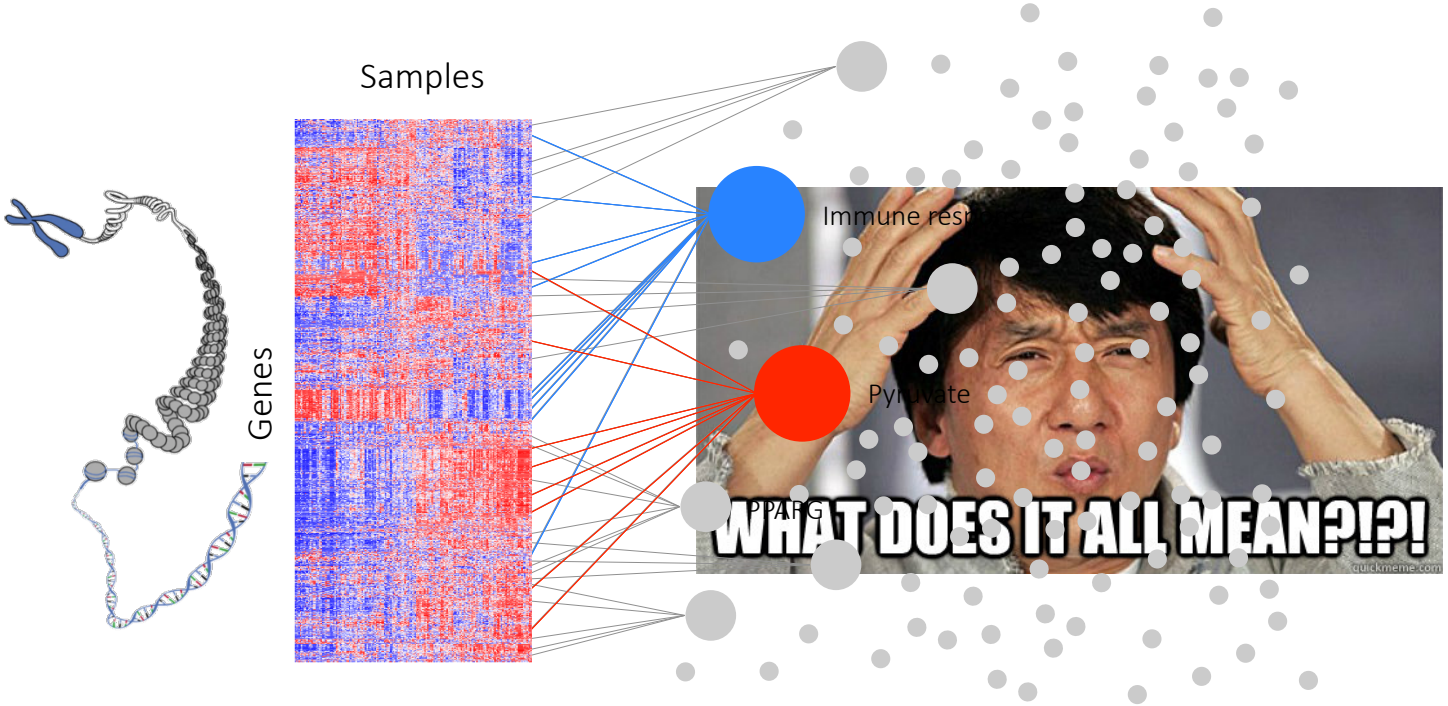
leif.varemo@scilifelab.se

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…

Gene-level data → Gene-set analysis → Gene-set data (results)

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

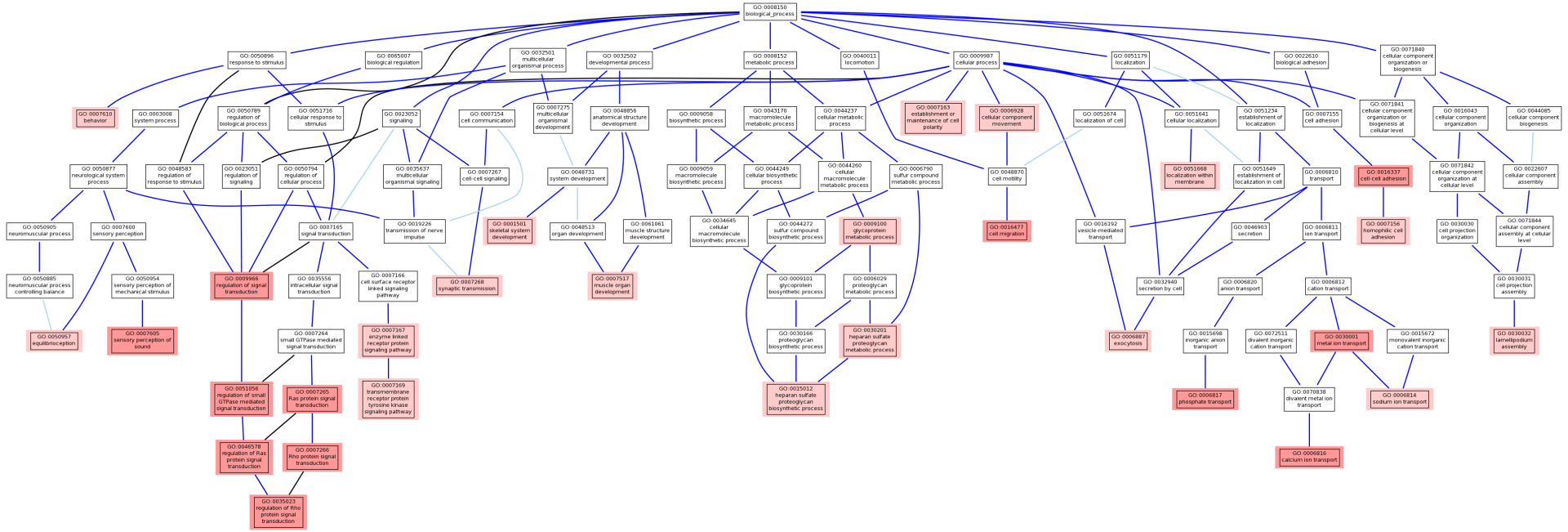
Gene-sets

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

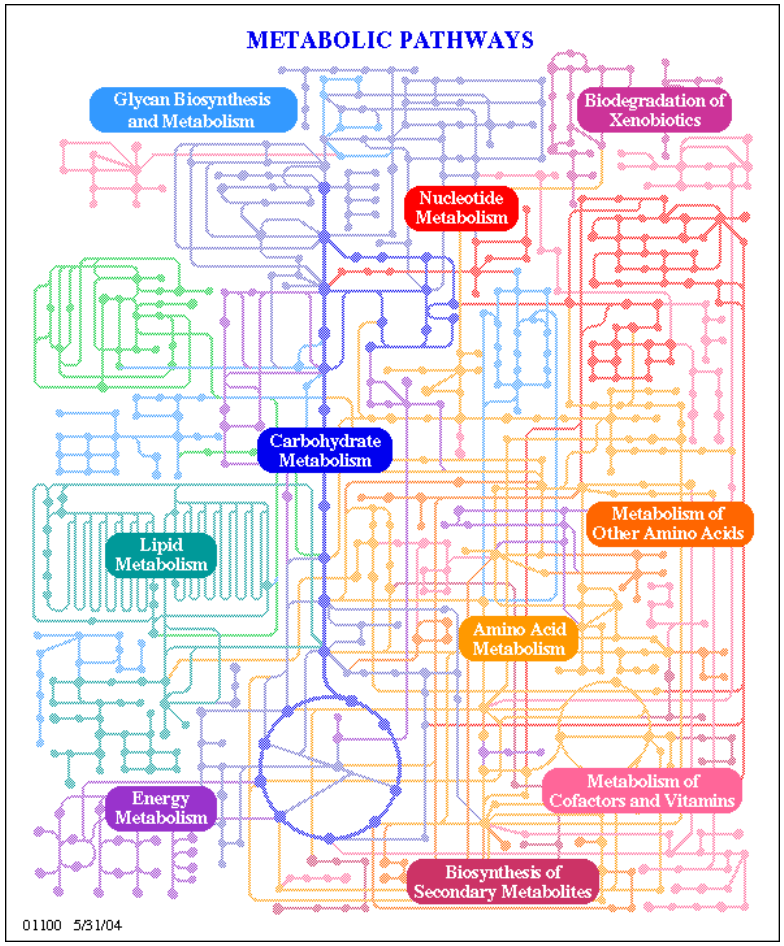
GO-terms
Pathways
Chromosomal locations
Transcription factors
Histone modifications
Diseases
Metabolites
etc…

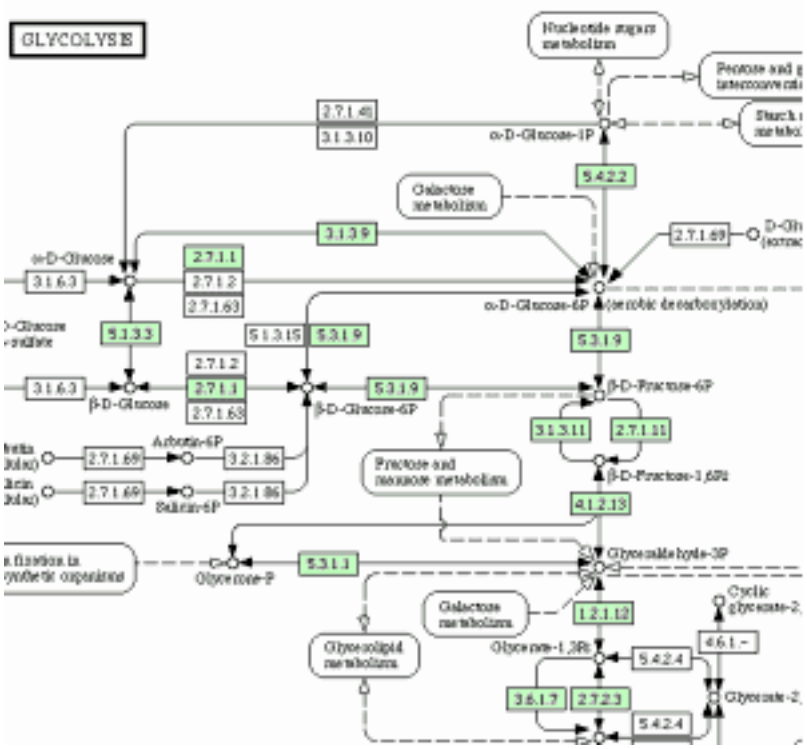
Gene-set example: Gene ontology (GO) terms



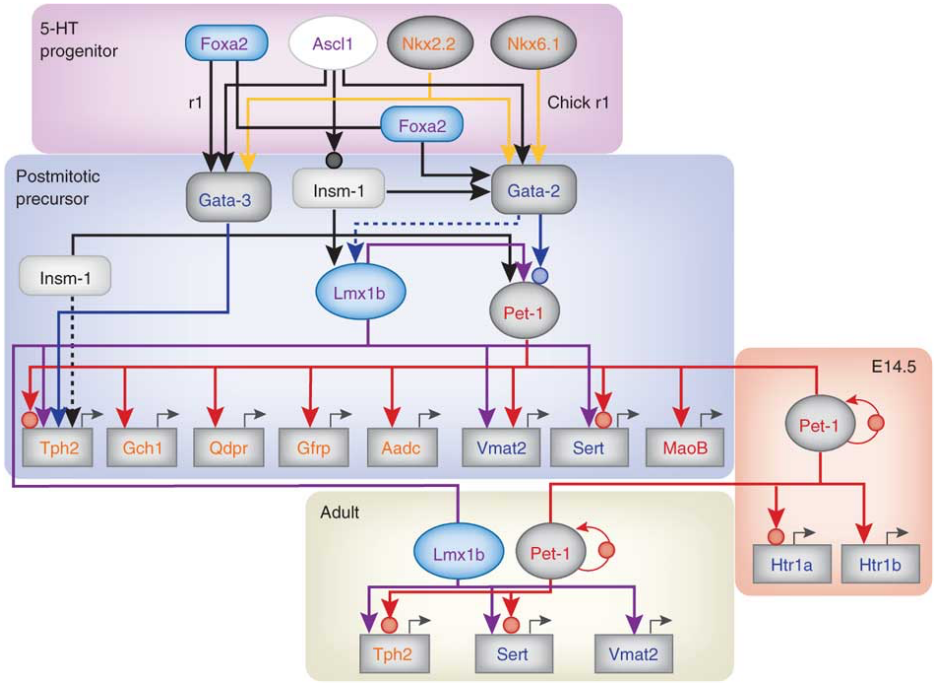
- Hierarchical graph with three categories (or parents): Biological process, Molecular function, Cellular compartment
- Terms get more and more detailed moving down the hierarchy
- Genes can belong to multiple GO terms

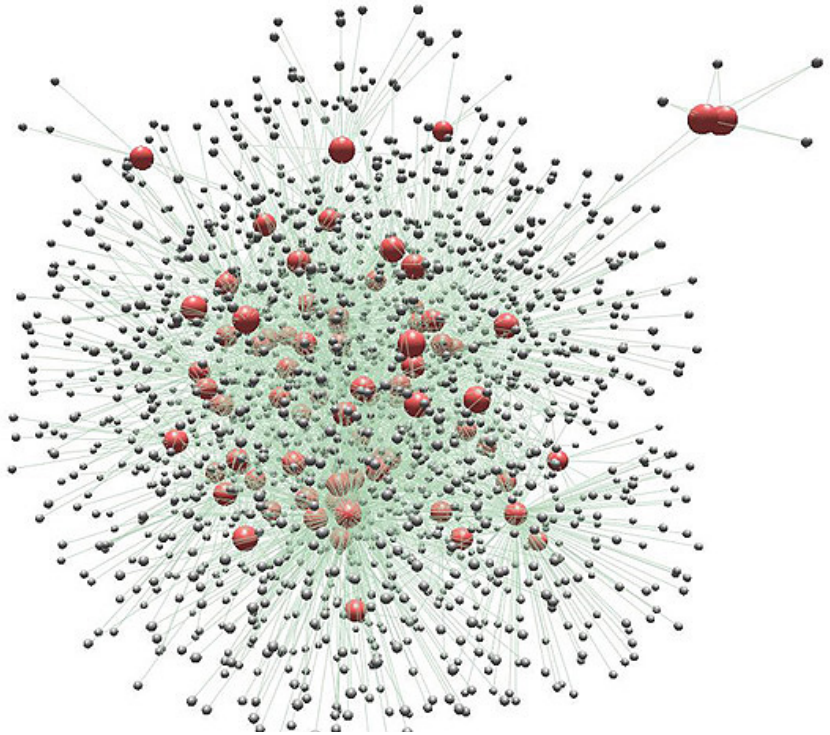
Gene-set example: Metabolic pathways or metabolites



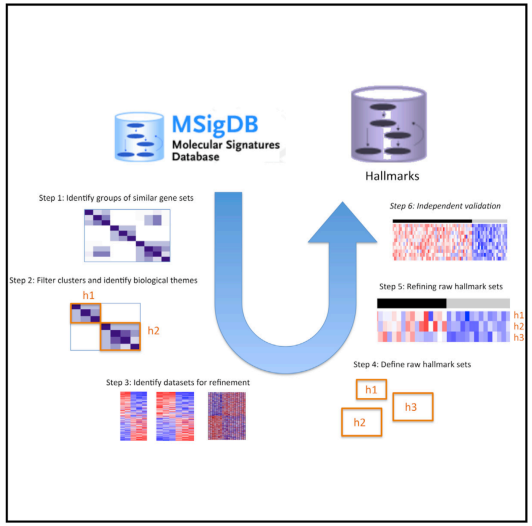


Gene-set example: Transcription factor targets





Gene-set example: Hallmark gene-sets



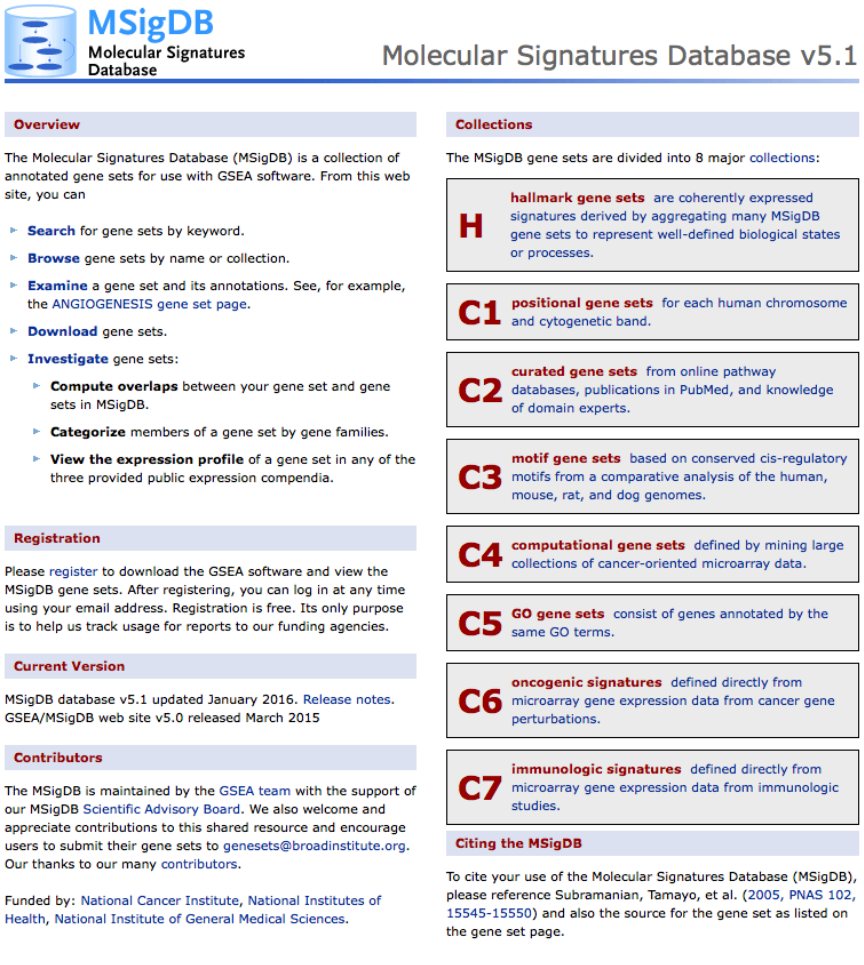
"Hallmark gene sets summarize and represent specific well-defined biological states or processes and display coherent expression. These gene sets were generated by a computational methodology based on identifying gene set overlaps and retaining genes that display coordinate expression. The hallmarks 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



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 appreciate contributions to this shared resource and encourage users to submit their gene sets to 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.

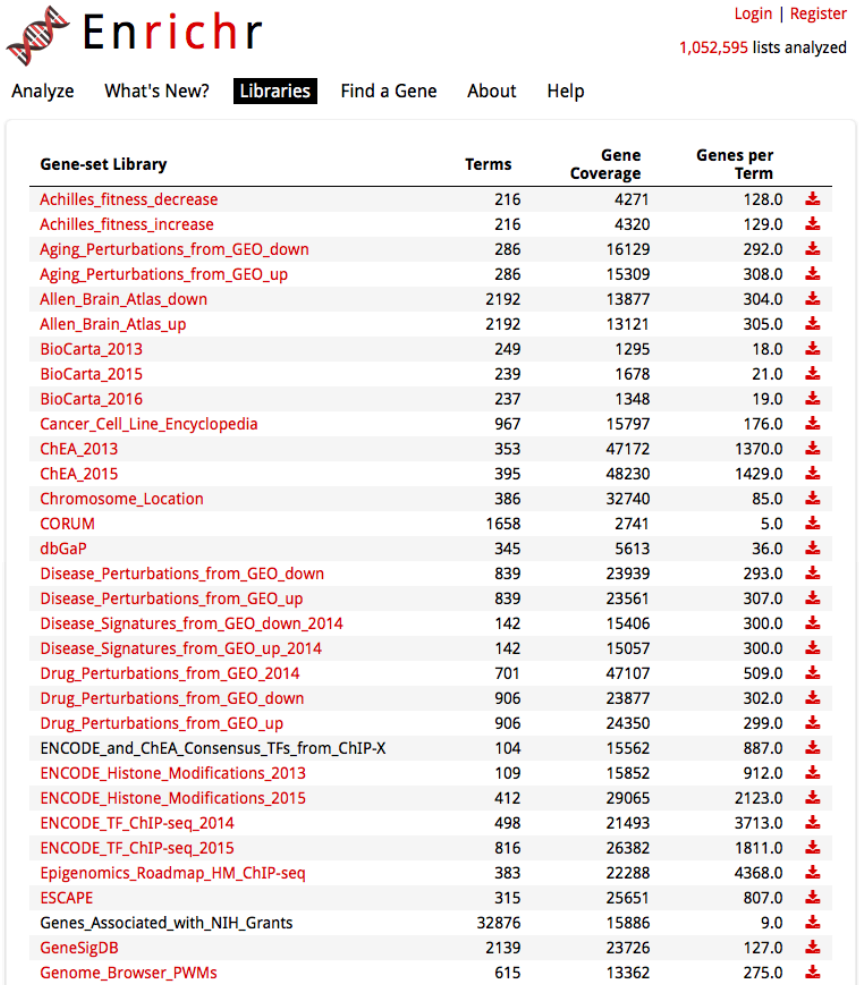
C6 **oncogenic signatures** defined directly from microarray gene expression data from cancer gene perturbations.

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



Enrichr

Login | Register

1,052,595 lists analyzed

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

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

SciLifeLab

Karolinska Institutet

KTH Vetenskap och Konst

Stockholms universitet

Uppsala Universitet

Where to get gene-set collections?

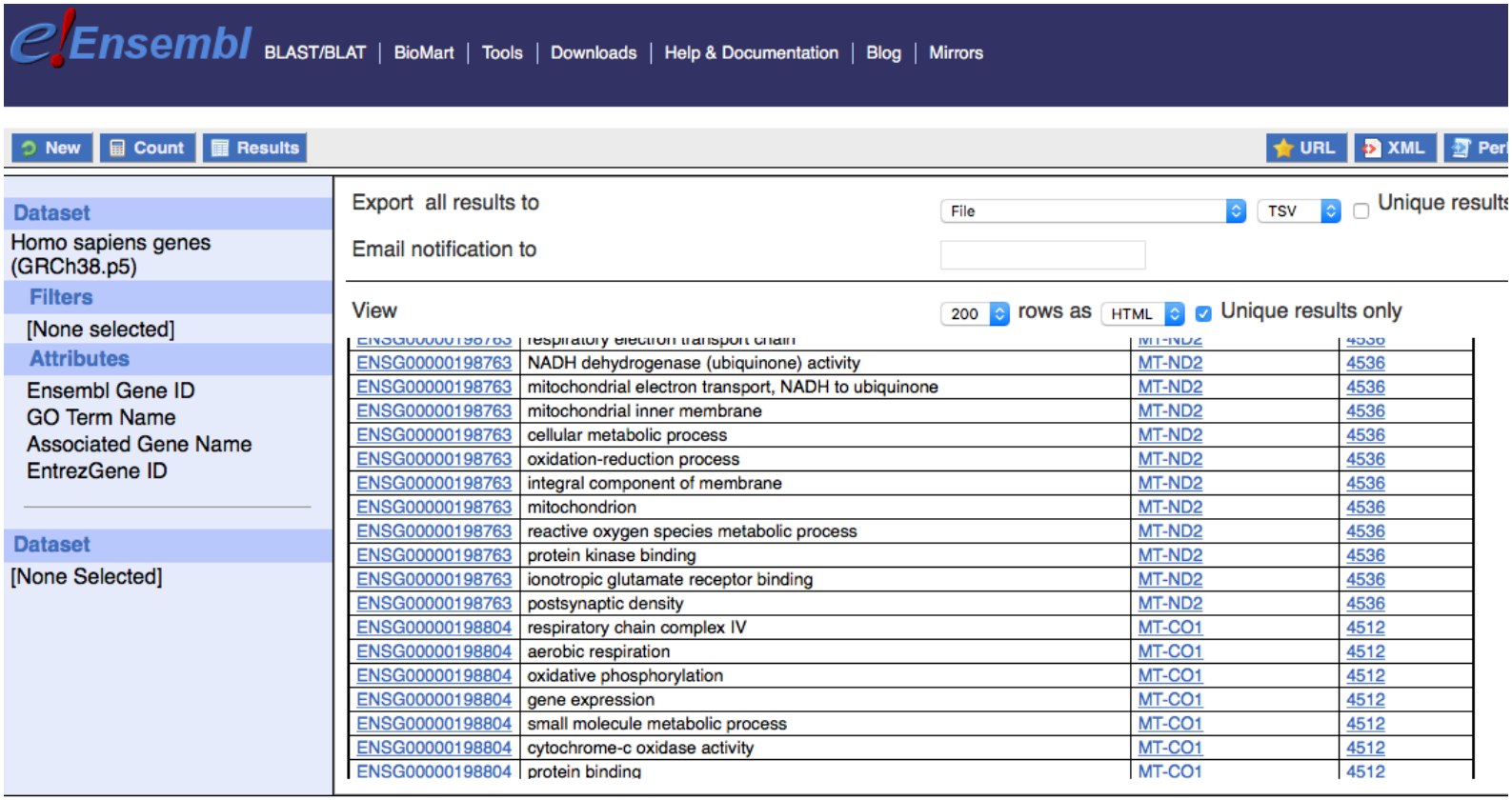
- Sooner or later you will run into the problem of matching your data to gene-set collections due to the existence of several gene ID types

```
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    PDGFRA  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)           EGLN2   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)            CHRM5   NPY2R   NPY1R   OPRK1   OPRL1
regulation of transcription factor activity (GO:0051090)                IL10    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   AGT     SMARCA4
NOD2    TNF     EDA2R   NLRC3   MTDH    PYCARD  IKBKB   ABRA    PRDX3   IRAK3   EP300   IRAK1   ERC1    RIPK1   IL6R
positive regulation of NF-kappaB 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>
ENSG00000000003 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



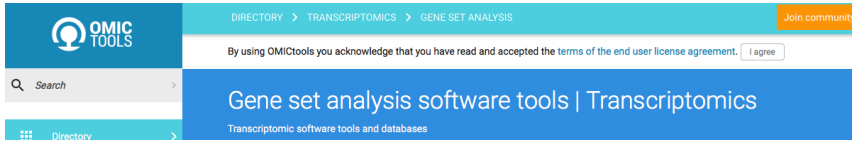
ENSG00000198763	NADH dehydrogenase (ubiquinone) activity	MT-ND2	4536
ENSG00000198763	mitochondrial electron transport, NADH to ubiquinone	MT-ND2	4536
ENSG00000198763	mitochondrial inner membrane	MT-ND2	4536
ENSG00000198763	cellular metabolic process	MT-ND2	4536
ENSG00000198763	oxidation-reduction process	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



Bioconductor (R packages)
https://bioconductor.org/packages/release/BiocViews.html#___GeneSetEnrichment

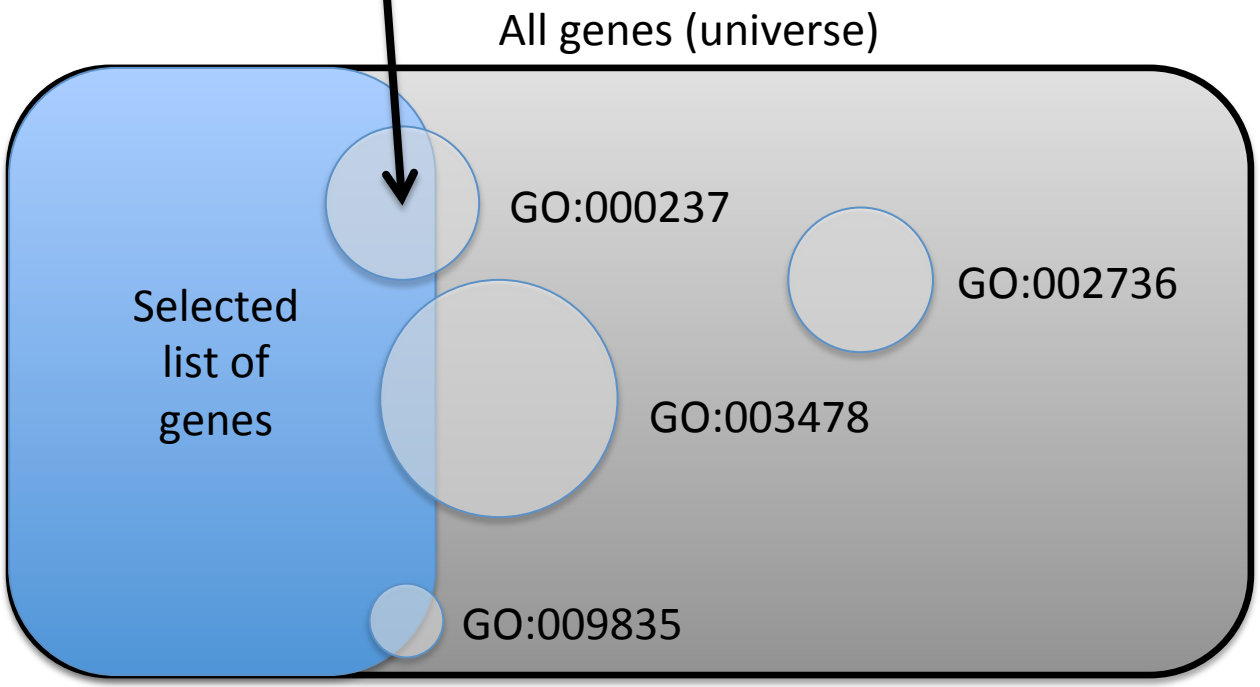
- Hypergeometric test / Fisher's exact test (a.k.a overrepresentation analysis)
- DAVID (browser)
- Enrichr (browser)
- GSEA (Java, R)
- Piano (R)

Overrepresentation analysis

Hypergeometric test
(Fisher's exact test)

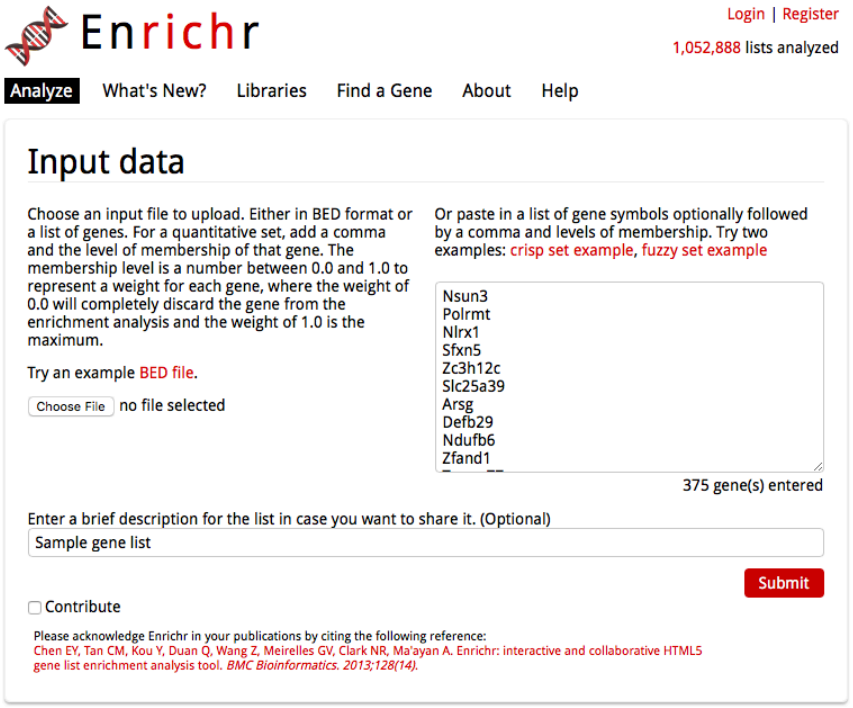
Is this overlap bigger than expected by random chance?

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



Overrepresentation analysis

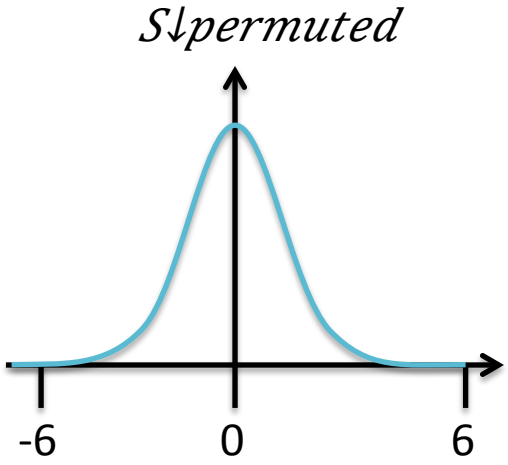
http://amp.pharm.mssm.edu/Enrichr/



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

Overrepresentation analysis

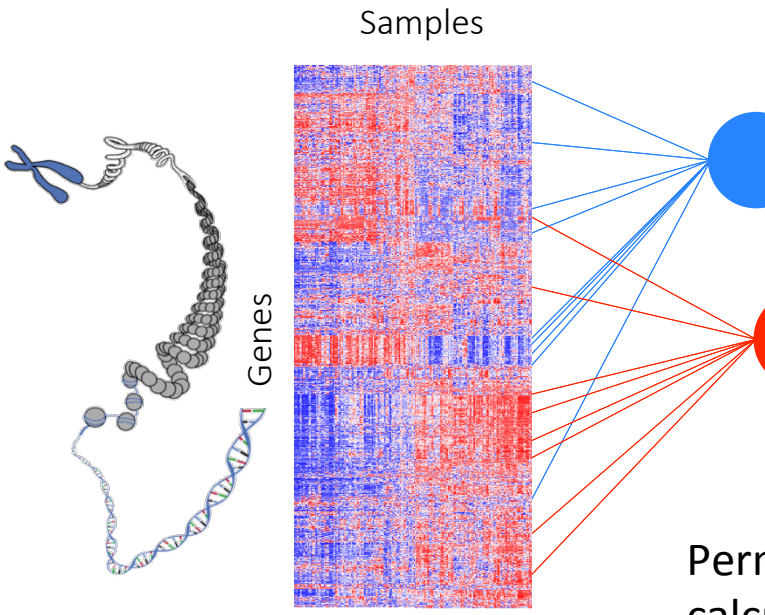
- Requires a cutoff (arbitrary)
- Omits the actual values of the gene-level statistics
- Good for e.g. overlap of significant genes in two comparisons
- Computationally fast
- In general, it is recommended to use some kind of gene-set analysis. This will use all gene-level data and can detect small but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



GSA: a simple example

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

$$S\downarrow i = mean(G\downarrow i)$$



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)!

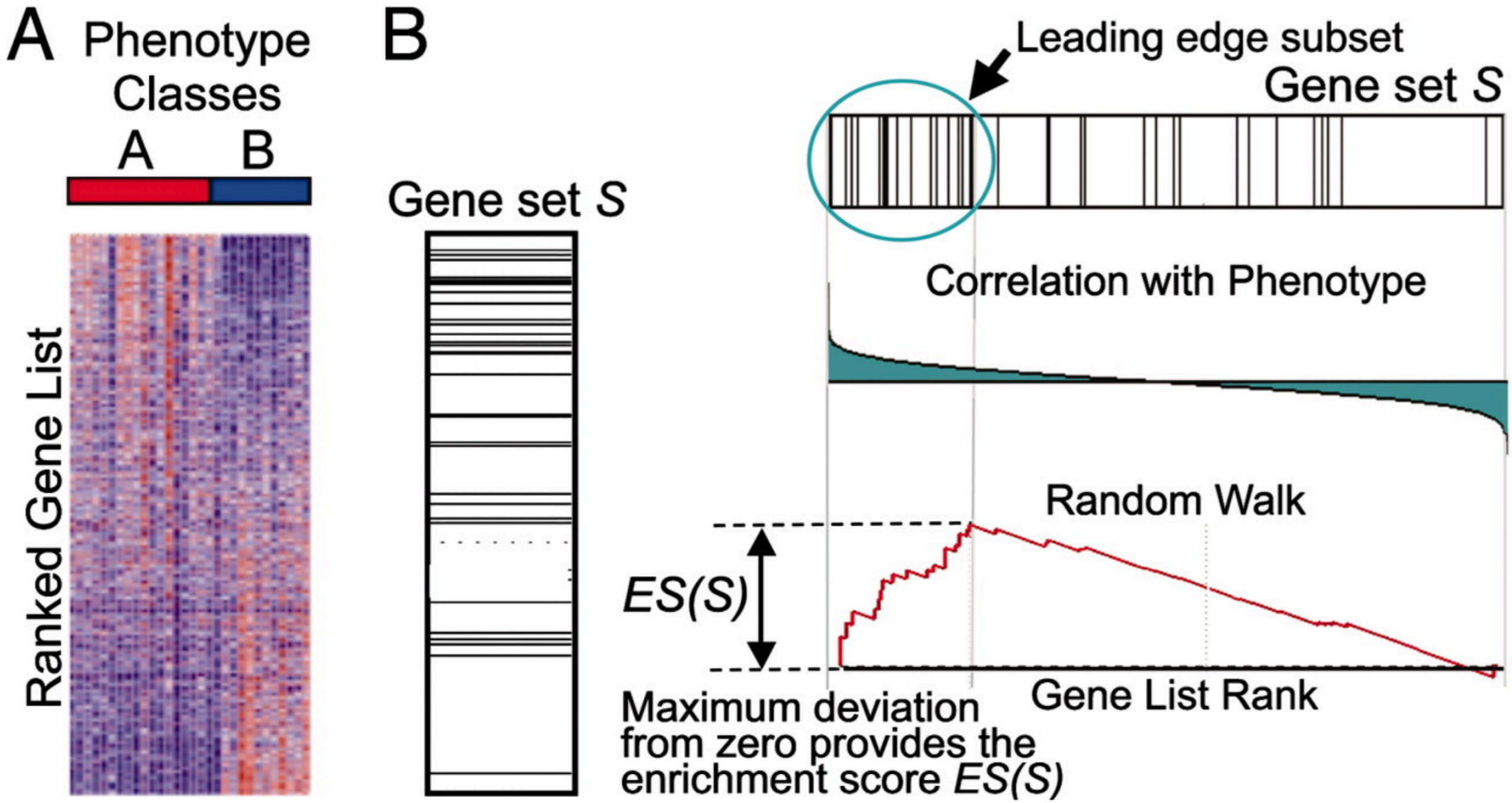
$$p\downarrow i = fraction\ of\ S\downarrow permuted\ \ that\ is\ more\ extreme\ than\ S\downarrow i$$

Gene-level statistics

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

GSEA

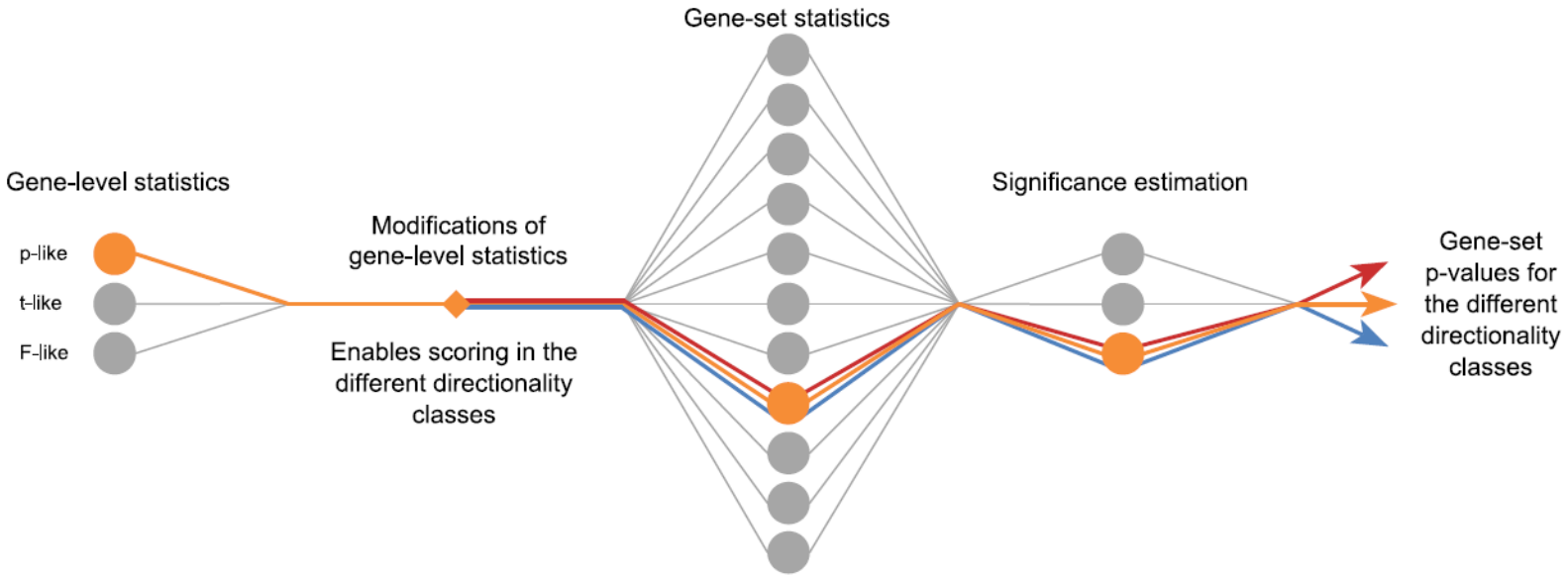
Mootha et al Nature Genetics, 2003; Subramanian PNAS 2005



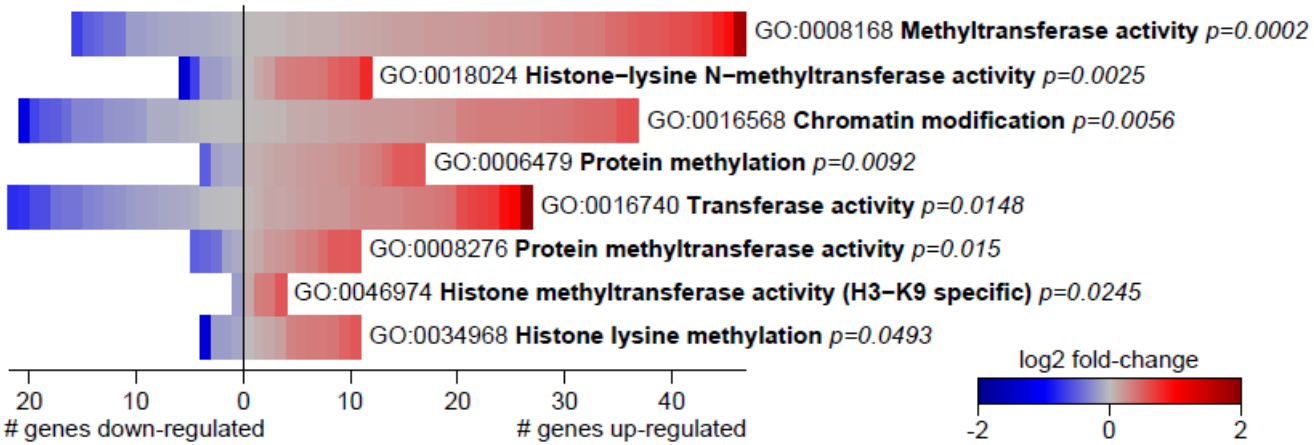
Piano – a platform for gene-set analysis (in R)

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

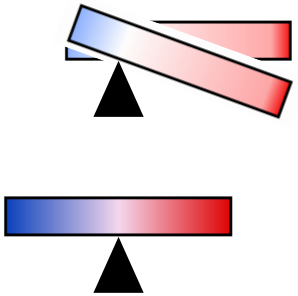
} Consensus result



Directionality of gene-sets

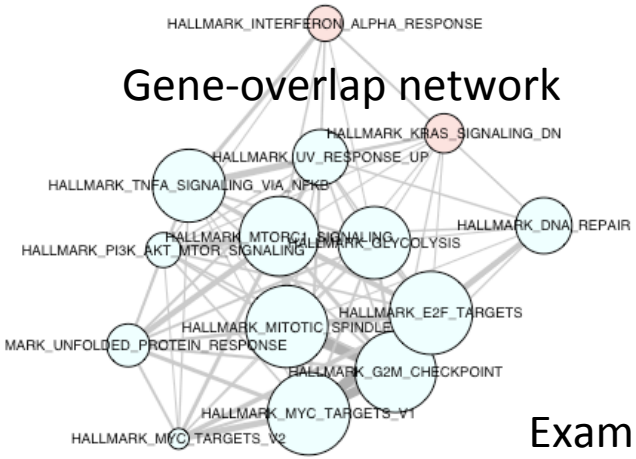


Gene set 1 Gene set 2 Gene set 3



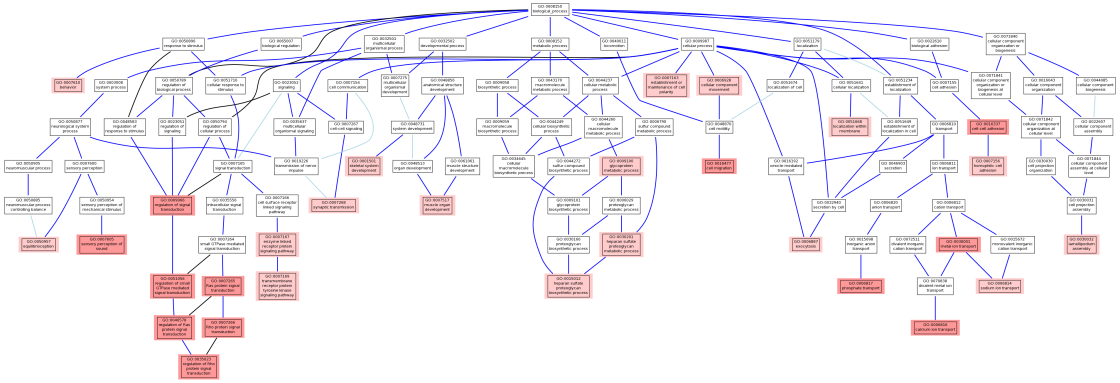
Saturation = gene significance
Red = up-regulated
Blue = down-regulated

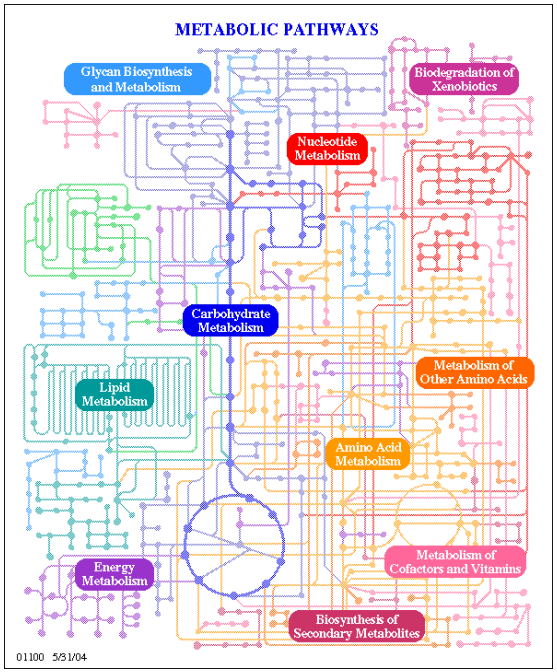
Gene-set overlap and interaction



Gene-overlap network

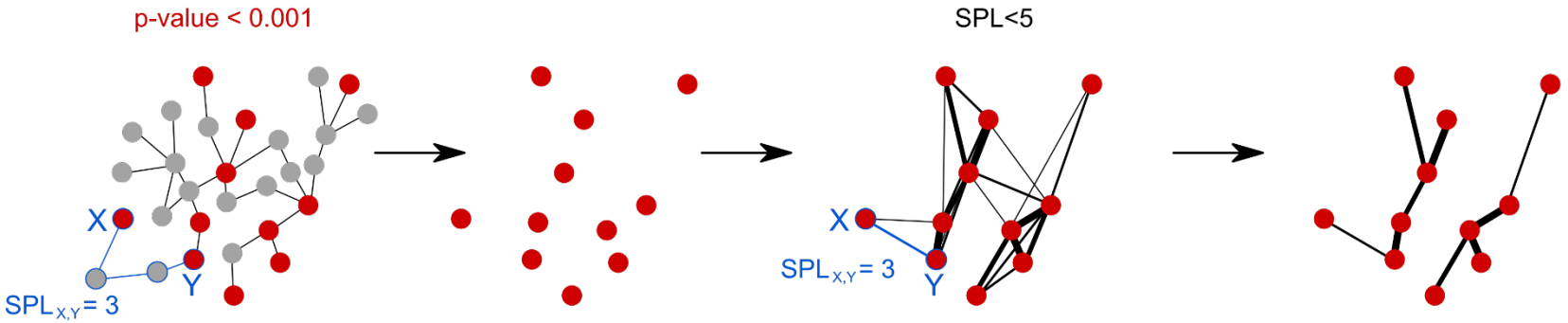
Examples of gene-set “interactions”





- 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

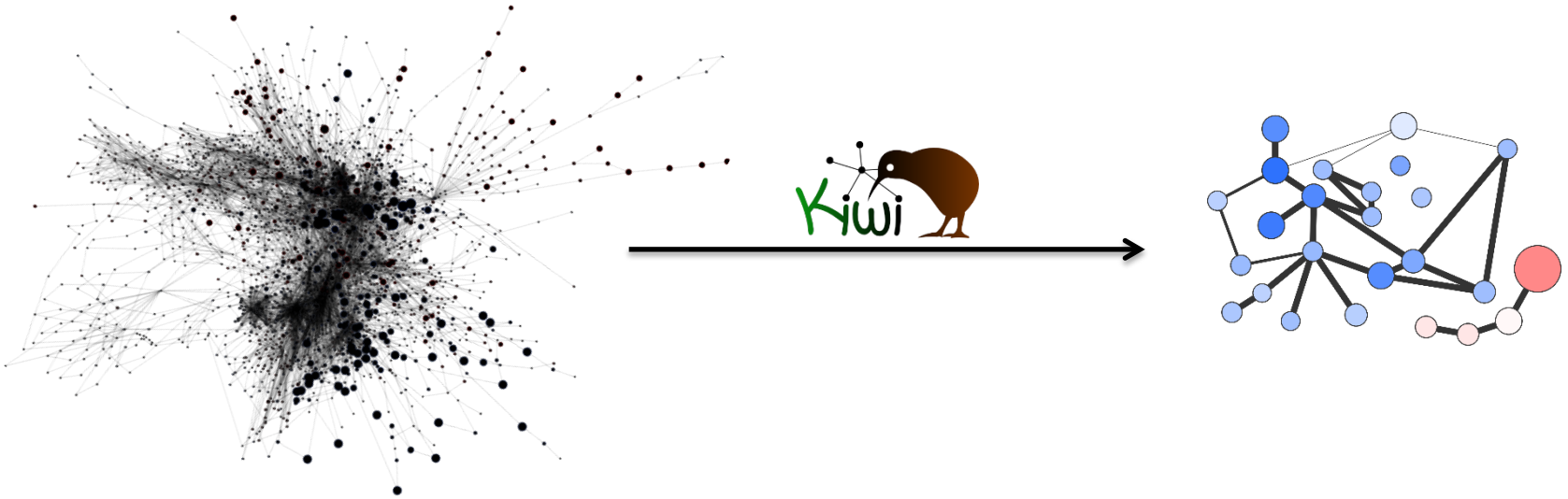


Gene-set interaction network (reduced example).

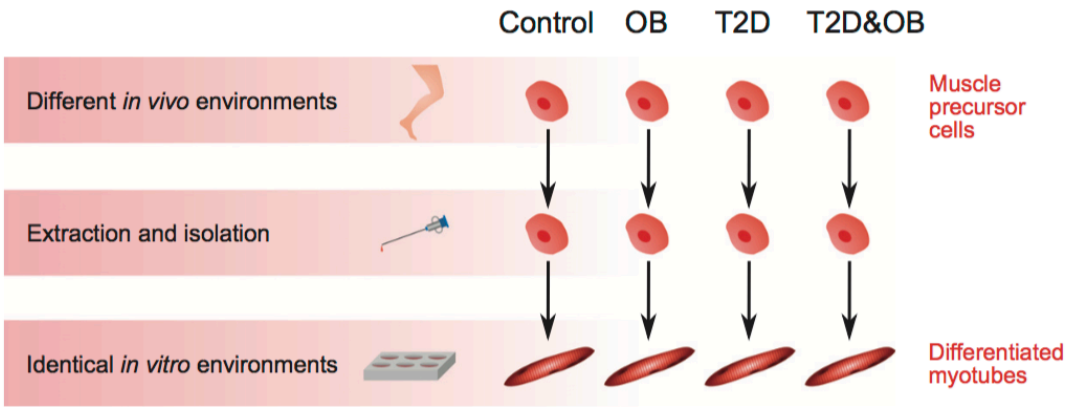
Use the significant gene-sets as nodes.

Calculate the shortest path length (SPL) between all node pairs. Draw an edge if the SPL is below a cutoff (5 in this example), with a thickness corresponding to the SPL.

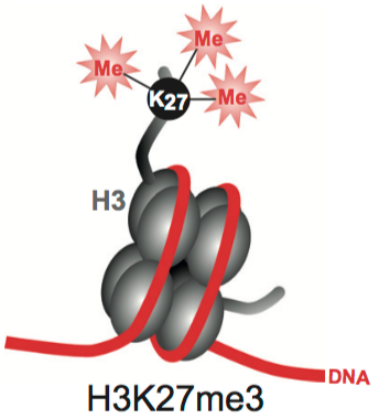
Keep only the best edges (with the smallest SPL) for each node.

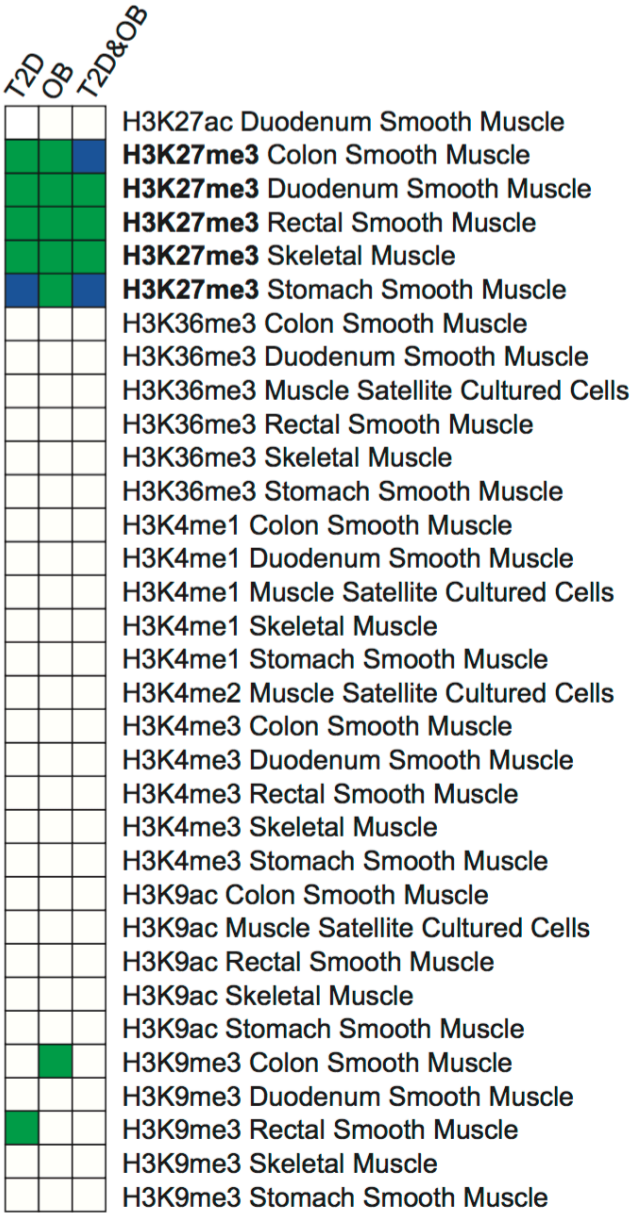


Example



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





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