1. Introduction
   - Gene regulation
   - Genomics and genome analyses
   - Hidden markov model (HMM)

2. Gene regulation tools and methods
   - Regulatory sequences and motif discovery
   - TF binding sites, microRNA target prediction

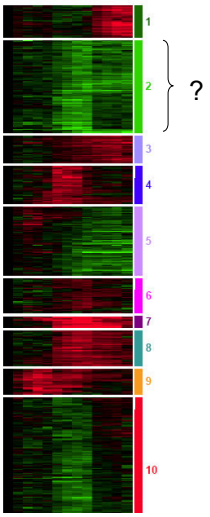
3. Technologies
   - Microarrays
   - Deep sequencing (RNAseq)
   - Single cell RNAseq spatial transcriptomics

4. Clustering
   - Unsupervized clustering (HCA,K-means, PCA, SOM)
   - Supervized clustering (classification)

5. Gene ontology, Pathways
   - Databases, tools
   - Over representation and enrichment analysis

6. Biomolecular networks
   - Network analysis and characteristics

Biological meaning of the gene sets



Co-expressed genes have something in common (guilt by association)

- Co-regulated (by the same TF or other regulators) (Lecture 2)
- Gene ontology/pathway
- Over representation analysis
- Gene set enrichment analysis
- Footprint analysis

Gene Ontology

Gene Ontology (GO)

The Gene Ontology project (http://geneontology.org) provides a **controlled vocabulary** to describe gene and gene product attributes in any organism.

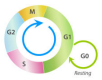
The three organizing principles (categories) of GO are

- cellular component

mitochondrium

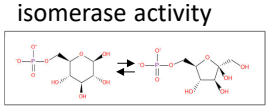
- biological process

cell cycle



- molecular function

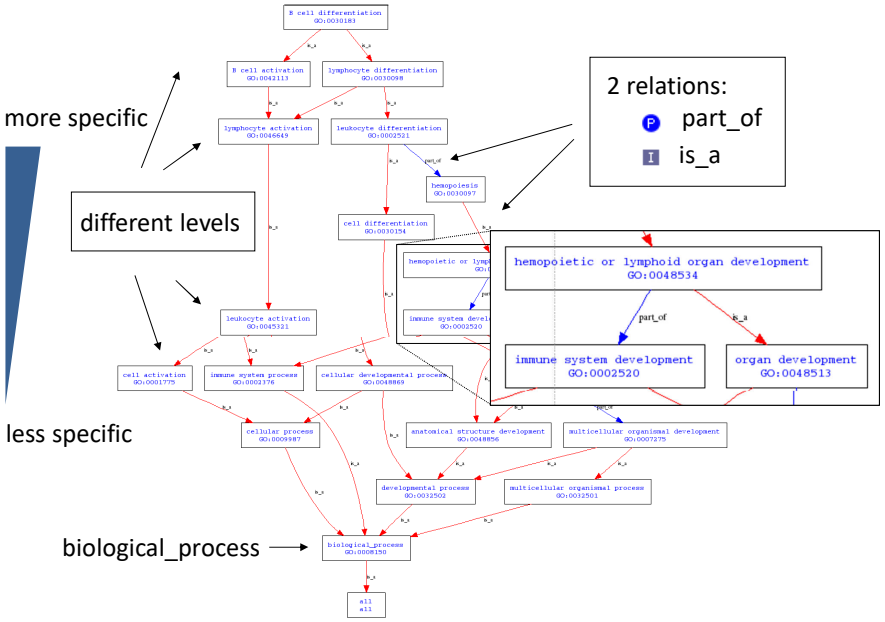
isomerase activity



What's in a GO term?

- **Term**
  transcription initiation
- **ID**
  GO:0006352
- **Definition**
  Processes involved in starting transcription, where transcription is the synthesis of RNA by RNA polymerases using a DNA template.

Parent /child relation in directed acyclic graph (DAG)



Gene Ontology Browser (Amigo2)

http://amigo2.geneontology.org (http://geneontology.org/)

Term information

Accession GO:0006629
Name lipid metabolic process
Ontology biological_process
Synonyms lipid metabolism

Inferred tree view

- GO:0008150 biological_process
  - GO:0008152 metabolic process
  - GO:0044699 single-organism process
    - GO:0071704 organic substance metabolic process
    - GO:0044238 primary metabolic process
    - GO:0044710 single-organism metabolic process
      - **GO:0006629 lipid metabolic process**
        - GO:0044255 cellular lipid metabolic process
        - GO:1900555 emericellamide metabolic process
        - GO:1902898 fatty acid methyl ester metabolic process
        - GO:1903173 fatty alcohol metabolic process
        - GO:0008610 lipid biosynthetic process
        - GO:0016042 lipid catabolic process
        - GO:1903509 liposaccharide metabolic process
        - GO:0045833 negative regulation of lipid metabolic process
        - GO:0045834 positive regulation of lipid metabolic process
        - GO:0019216 regulation of lipid metabolic process
        - GO:0008202 steroid metabolic process

Annotation

Total: 413; showing 11-20 **Results count**

Gene/prod	Gene/product name	Direct annotation	Assigned by	Taxon	Evid
THEM4	Acyl-coenzyme A thioesterase THEM4	fatty acid metabolic process	UniProt	Homo sapiens	IDA
ABHD12	Monoacylglycerol lipase ABHD12	acylglycerol catabolic process	UniProt	Homo sapiens	IDA
APOA5	Apolipoprotein A-V	triglyceride metabolic process	BHF-UCL	Homo sapiens	IDA

...

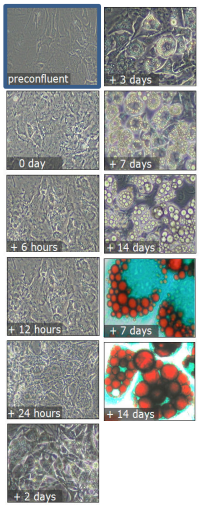
Evidence code for GO annotations

ISS	Inferred from Sequence Similarity
IEP	Inferred from Expression Pattern
IMP	Inferred from Mutant Phenotype
IGI	Inferred from Genetic Interaction
IPI	Inferred from Physical Interaction
IDA	Inferred from Direct Assay
RCA	Inferred from Reviewed Computational Analysis
TAS	Traceable Author Statement
NAS	Non-traceable Author Statement
IC	Inferred by Curator
ND	No biological Data available

Over representation analysis

Case study: fat cell differentiation

Reference (proliferating cells)

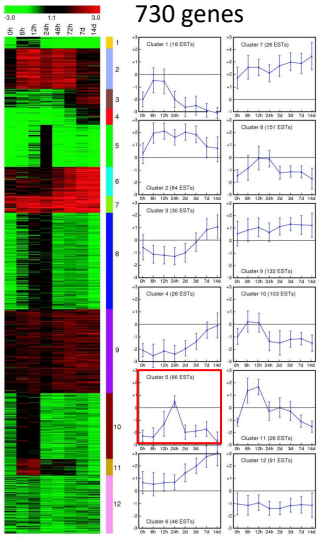


Two color microarray analysis

⇨

Time series
0h-ref
+6h-ref
+12h-ref
…
+14d-ref

730 genes



Biological function of genes in clusters

⇨

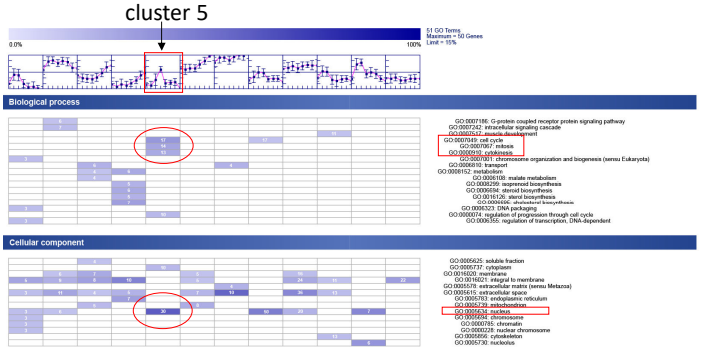
GO Analysis

Heatmap (k-means clustering)

Hackl H, Burkard TR et al. Genome Biol. 2005

GO terms for gene sets

cluster 5



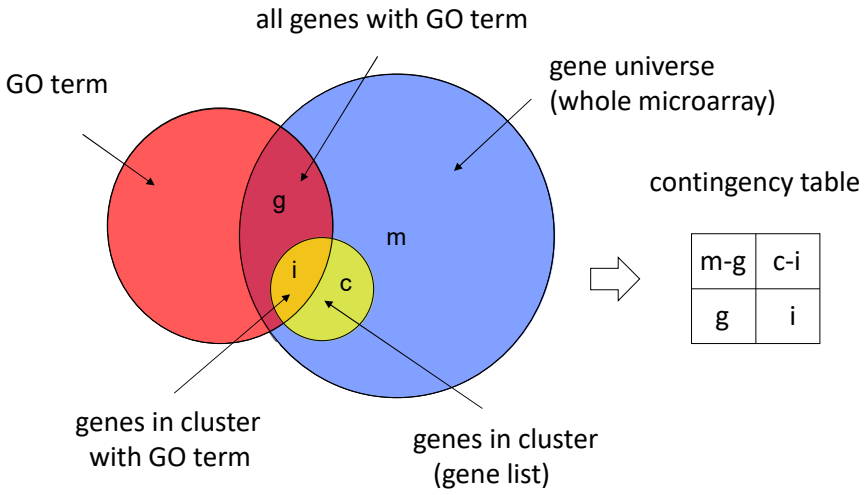
cell cycle (17)
mitosis (14) ⟶
cytokineses (13)

nucleus (30) ⟶

– 3T3-L1 cell line undergoes ≥ 1 cell cycle before terminal adipocyte differentiation around 1 day after induction (clonal expansion)

Are results just by chance?
⇨ Over representation (enrichment) analysis

Over representation analysis

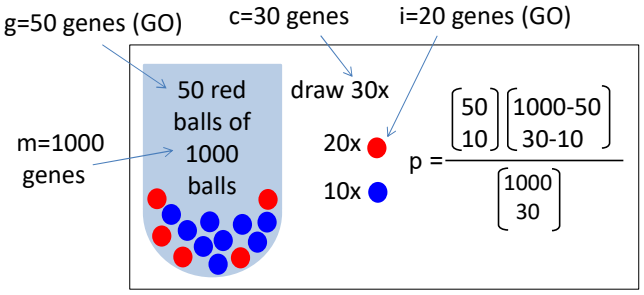


Over representation analysis

- Fisher exact test for contingency table

m-g	c-i
g	i

- Hypergeometric distribution



- Multiple hypothesis testing => adjust p-value
- Not only for GO Terms also for TFBS, pathways,..

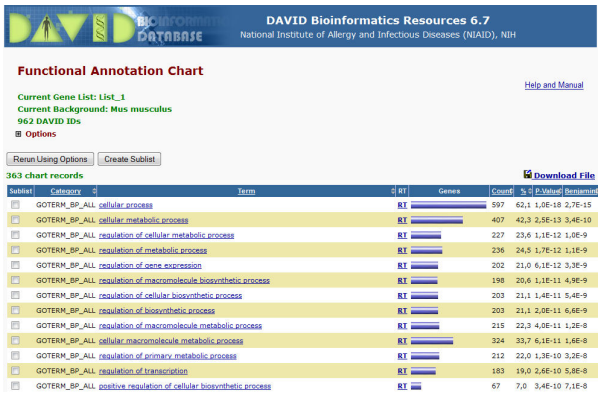
DAVID

- Database for Annotation, Visualization and Integrated Discovery
- https://david.ncifcrf.gov
- Functional annotation tool (over representation analysis)

1019 mouse gene symbols

Dnajb1
Wnt11
Sorbs3
D230025D16Rik
Sfxn3
Hspa5
Golga3
Hgs
Npc1
Mta2
Cnn2
Spg20
Zpr1

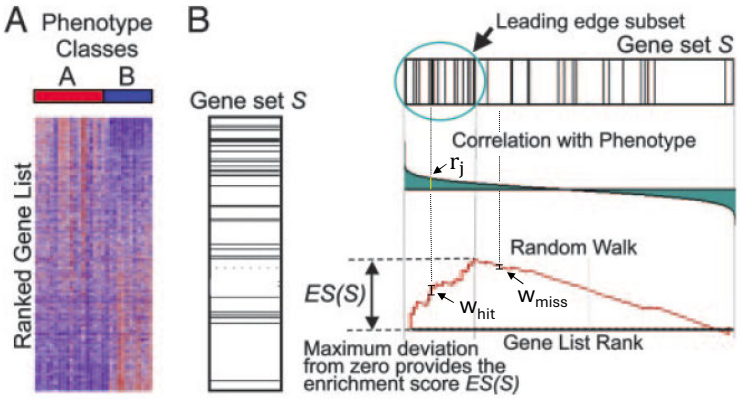
...



...

Gene set enrichment analysis

Gene set enrichment analysis (GSEA)



$$w_{hit} = \frac{|r_j|}{\sum_S |r_j|} \qquad w_{miss} = -\frac{1}{N - N_S}$$

N ... number of all genes
N_S ... number of genes in S
r_j ... ranked statistics (e.g. log2FC) for gene j in gene set S

Subramanian A et al. Proc Natl Acad Sci (2005)

Gene set enrichment analysis (GSEA)

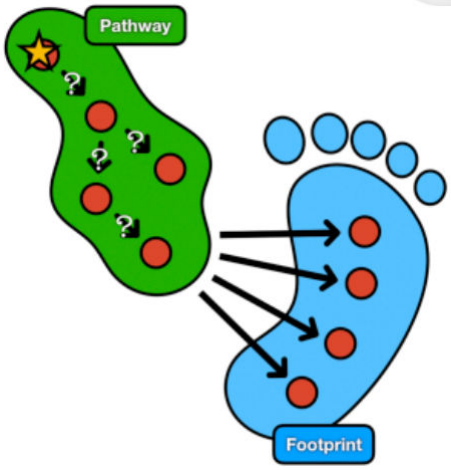
1. Given an *a priori* defined set of genes *S (MSigDB)*
2. Rank genes (e.g. by t-value between 2 groups of microarray samples) → ranked gene list L.
3. Calculation of an enrichment score (*ES*) that reflects the degree to which a set *S* is overrepresented at the extremes (top or bottom) of the entire ranked list *L*.
4. Estimation the statistical significance (nominal *P* value) of the *ES* by using an empirical phenotype-based permutation test procedure.
5. Adjustment for multiple hypothesis testing

http://www.broadinstitute.org/gsea

Footprint analyses

Footprint methods to infer functional activity

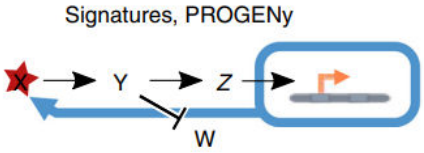
Signalling
Metabolism
Gene expression

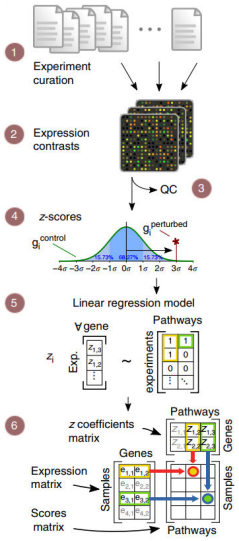


Dugourd and Saez-Rodriguez, Curr Opin Syst Biol, 2019

Footprint methods to infer functional activity

Compendium of perturbation experiments of signaling pathways to infer pathway activity from gene expression readout (14 pathways)

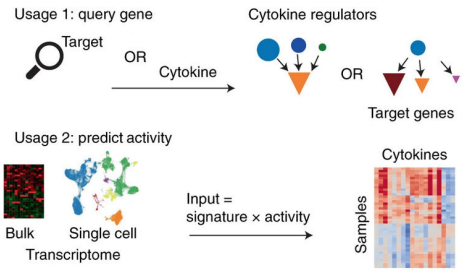




Schubert et al. Nat Commun 2018

CytoSig

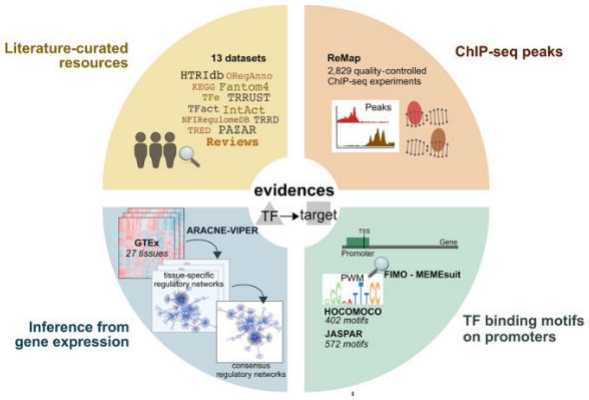
- Collection of >20k transcriptome profiles for human cytokine, chemokine and growth factor responses.
- Prediction of signaling activities in distinct cell populations in infectious diseases, chronic inflammation and cancer using bulk and single-cell transcriptomic data.



Jiang et al. Nat Methods 2021

DoRothEA

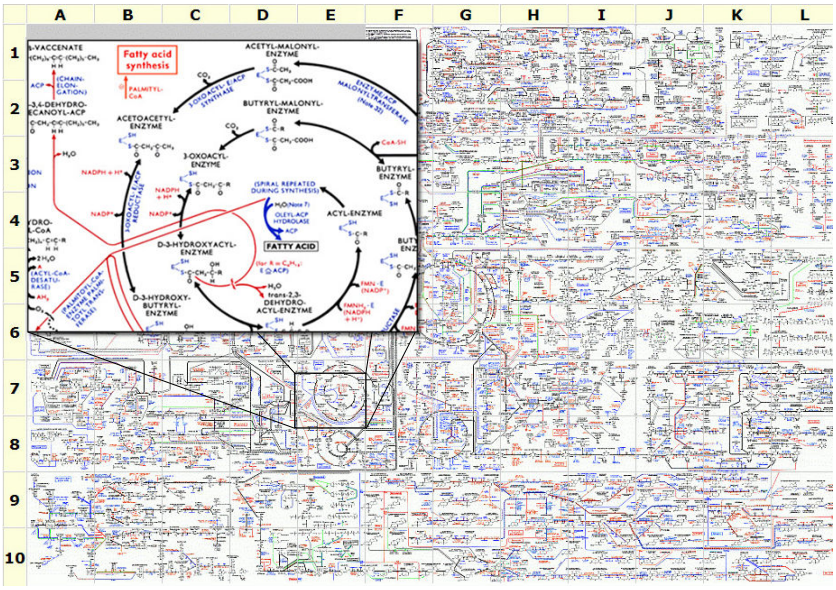
- The prediction of transcription factor (TF) activities from the gene expression of their targets (i.e., TF regulon)



Garcia-Alonso *et al.* Genome Res. 2019

Pathway databases and tools

Biochemical and Metabolic Pathways



Böhringer Mannheim

Pathways

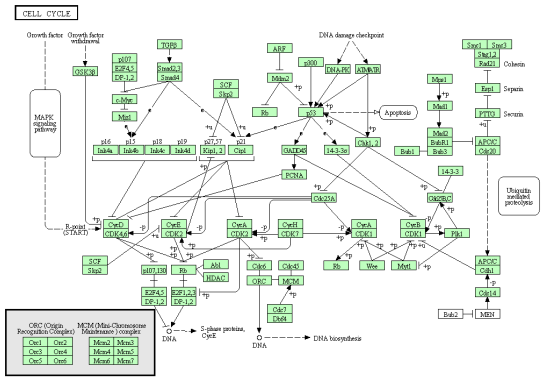
A **biological pathway** is a series of actions among molecules in a cell that leads to a certain product or a change in a cell. Such a pathway can trigger the assembly of new molecules, such as a fat or protein. Pathways can also turn genes on and off, or spur a cell to move (genome.gov/27530687).

- metabolic pathways
- signaling pathways
- gene regulation pathways

Canonical Pathways are idealized or generalized pathways that represent common properties of a particular signaling module or pathway

Pathways

- Kyoto Encylopedia of Genes and Genomes (KEGG)
- Reactome
- Wiki Pathways
- BioCyc
- Biocarta
- PANTHER



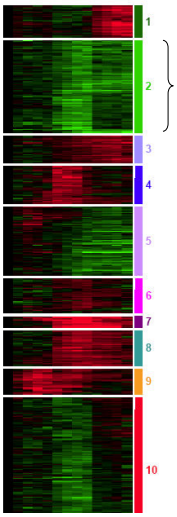
Pathway (and network) analysis

Pathway analysis
- over representation
- mapping gene expression to pathway

Interaction network (direct or indirect interactions)
- from protein-protein interaction databases (HPRD, IntAct)
- known and predicted functional association (STRING)

De novo network construction
- co-expression network
- reversed engineering

Biological meaning of the gene sets



Co-expressed genes have something in common (guilt by association)

- Co-regulated (by the same TF or other regulators) (Lecture 2)
- Gene ontology/pathways
- Over representation analysis
- Gene set enrichment analysis
- Footprint analysis