

The power of RNA-seq

Interpretation: enrichment, networks etc.

Dick de Ridder



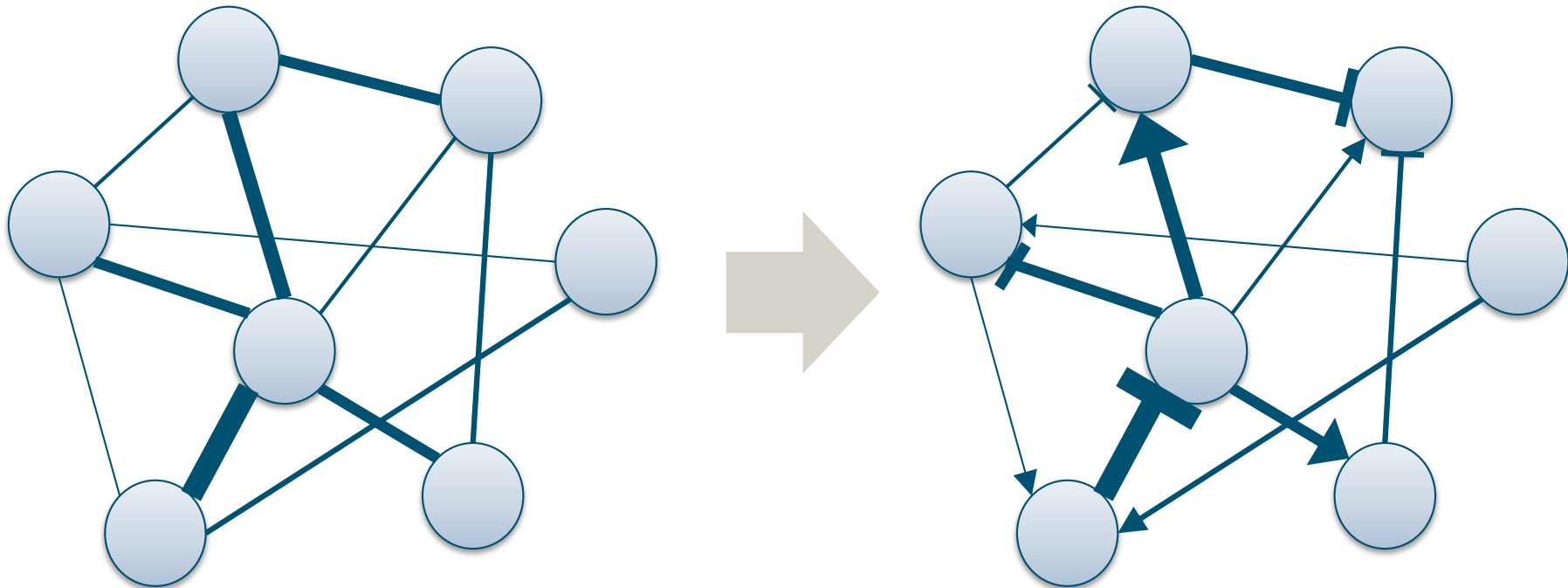
Contents

- Inferring regulation networks

- Interpretation
 - Annotation types
 - Fisher's exact test
 - Gene Set Enrichment Analysis (GSEA)
 - Network-based analysis

Inferring regulation networks

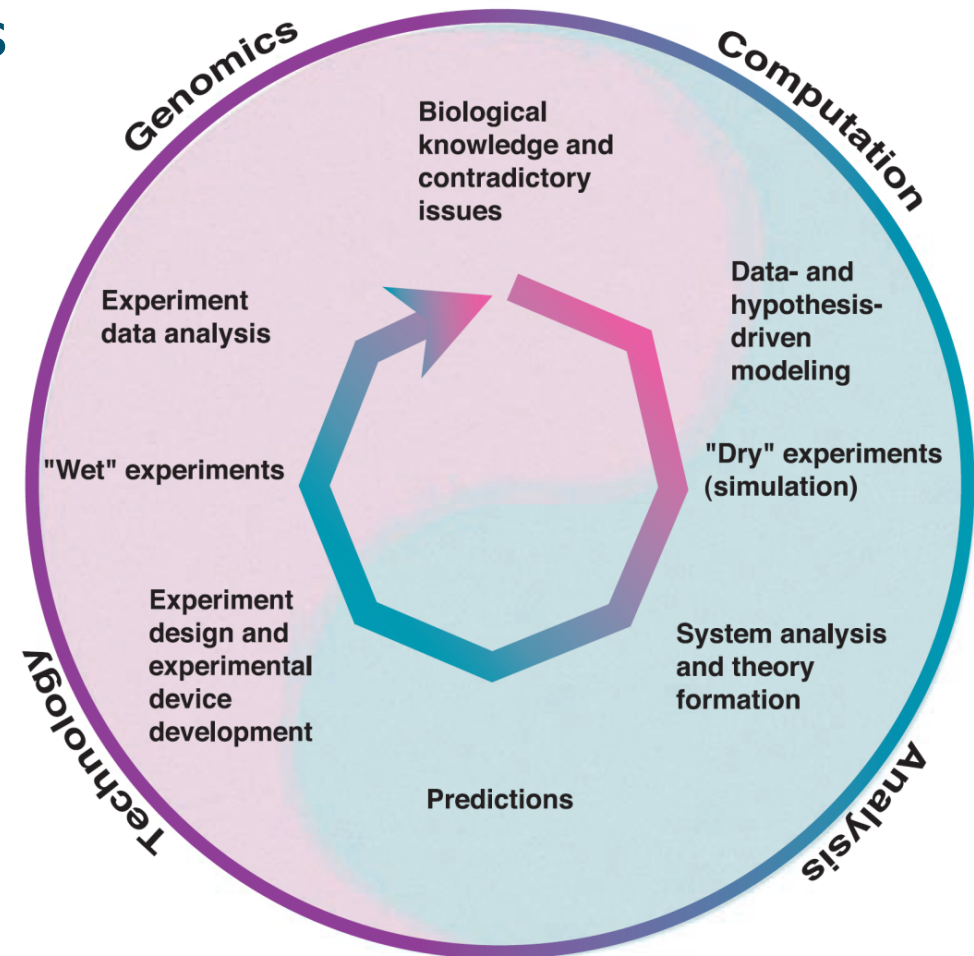
- Yesterday: WGCNA, simple correlation:



- Beyond correlation: reverse engineering of regulation, find connections and parameters from measurement data

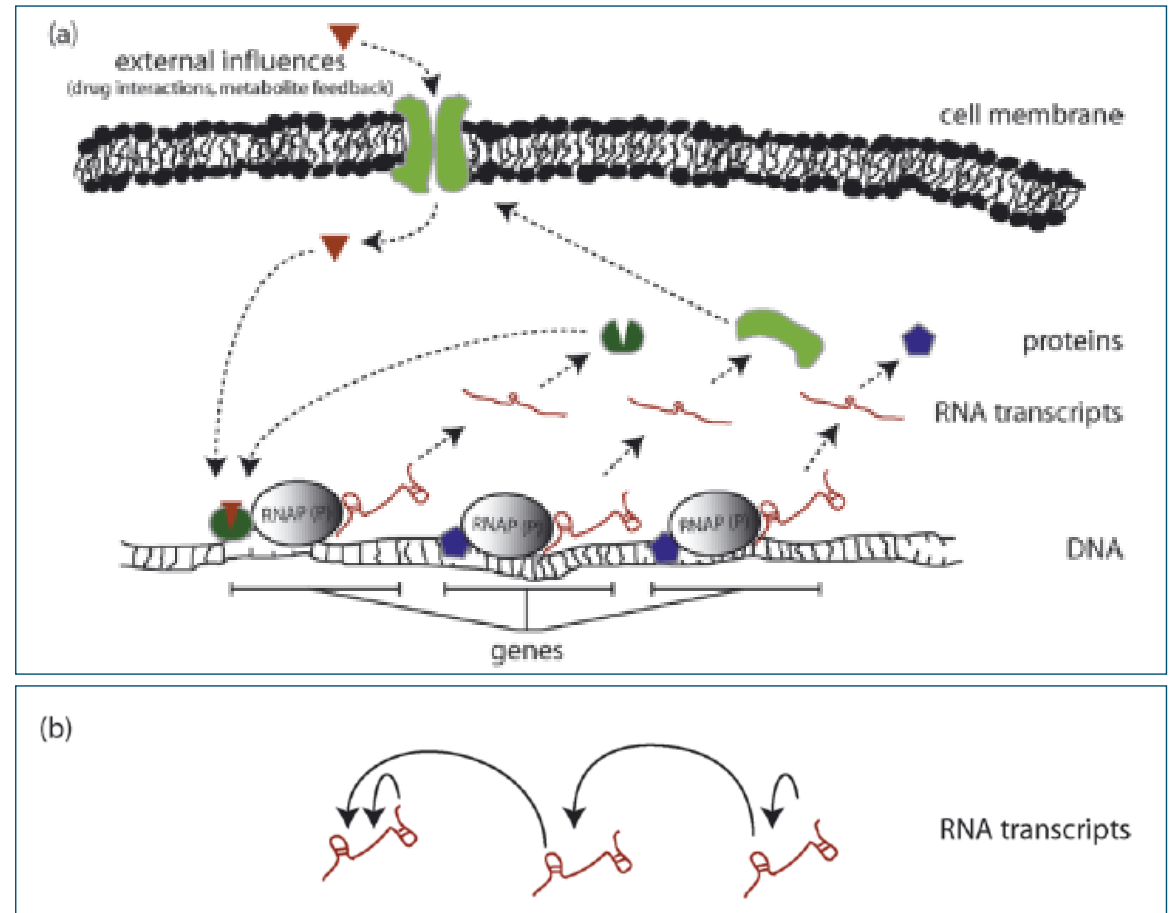
Systems biology

- Model biological processes to better understand life
- Interplay between biologists and computer scientists
- Use networks to
 - create hypotheses
 - guide experiments
 - store knowledge



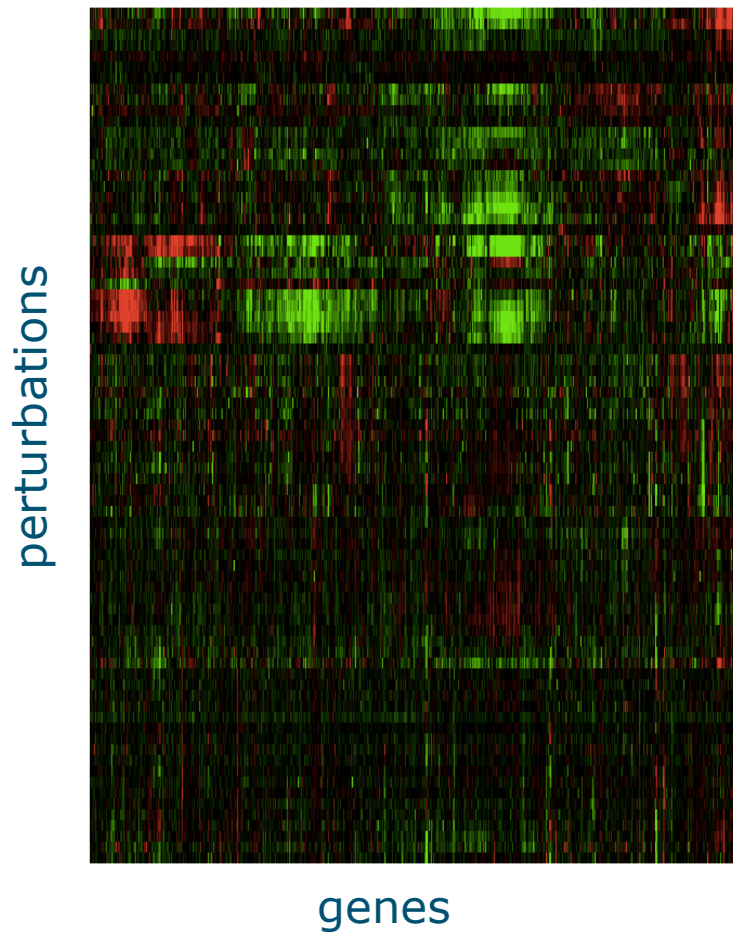
Inferring regulation networks (2)

- Physical approach: identify protein factors regulating transcription (model-based)
- Influence approach: summarize regulatory influences between transcripts

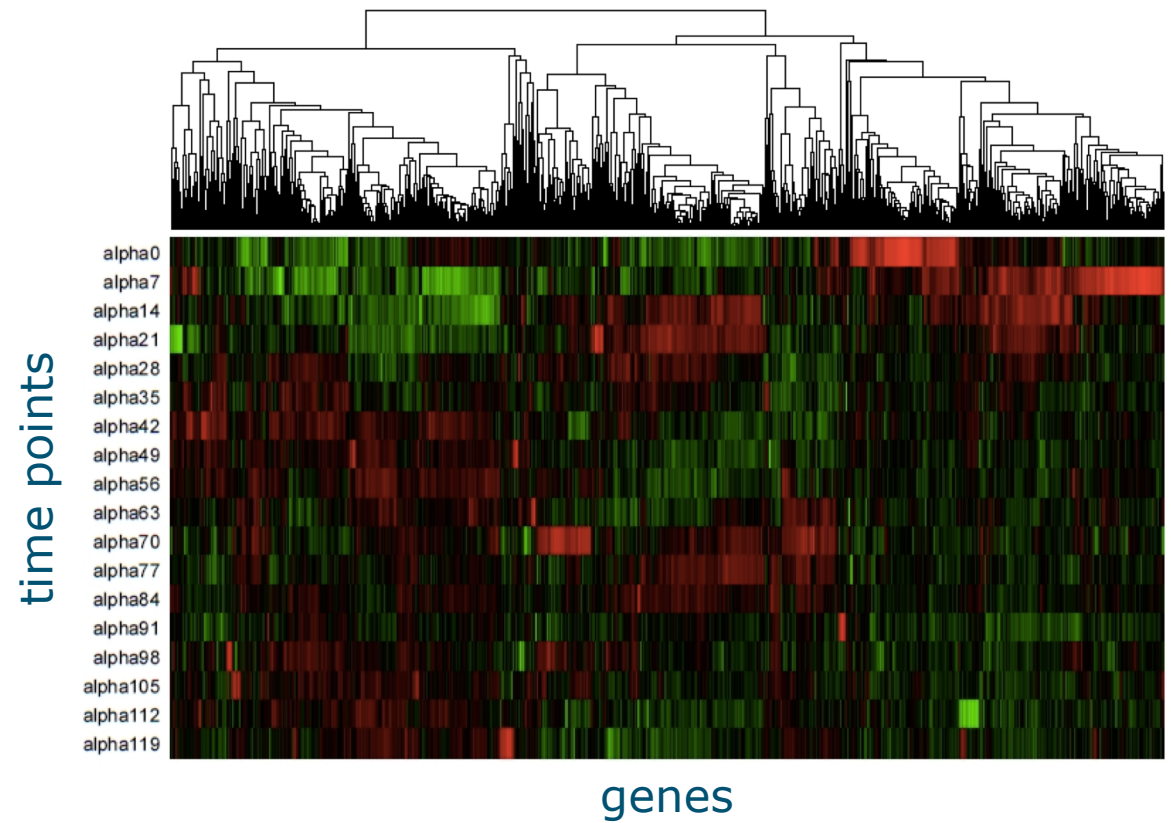


Measurement data

- Steady state

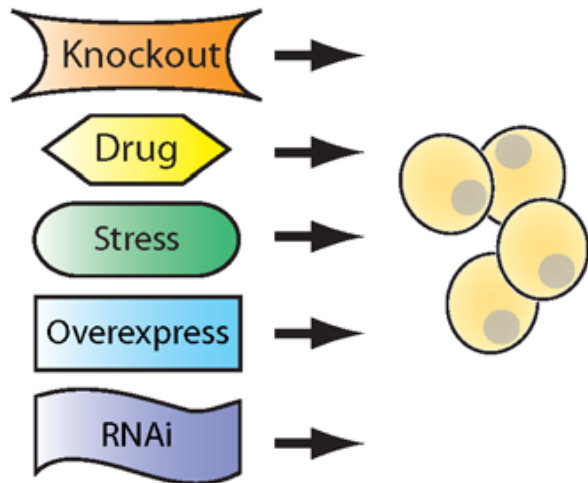


- Time series

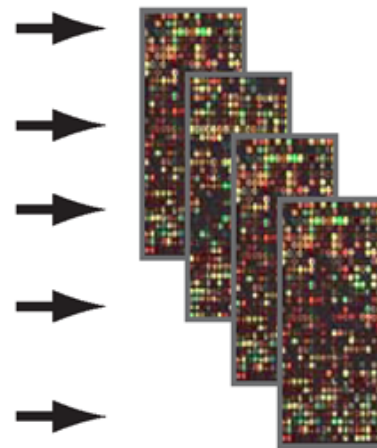


Steady-state data

1) Apply diverse treatments to cells



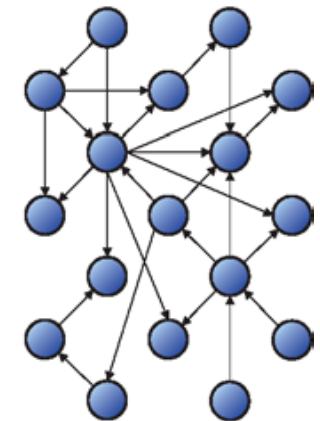
2) Measure RNA expression for each treatment



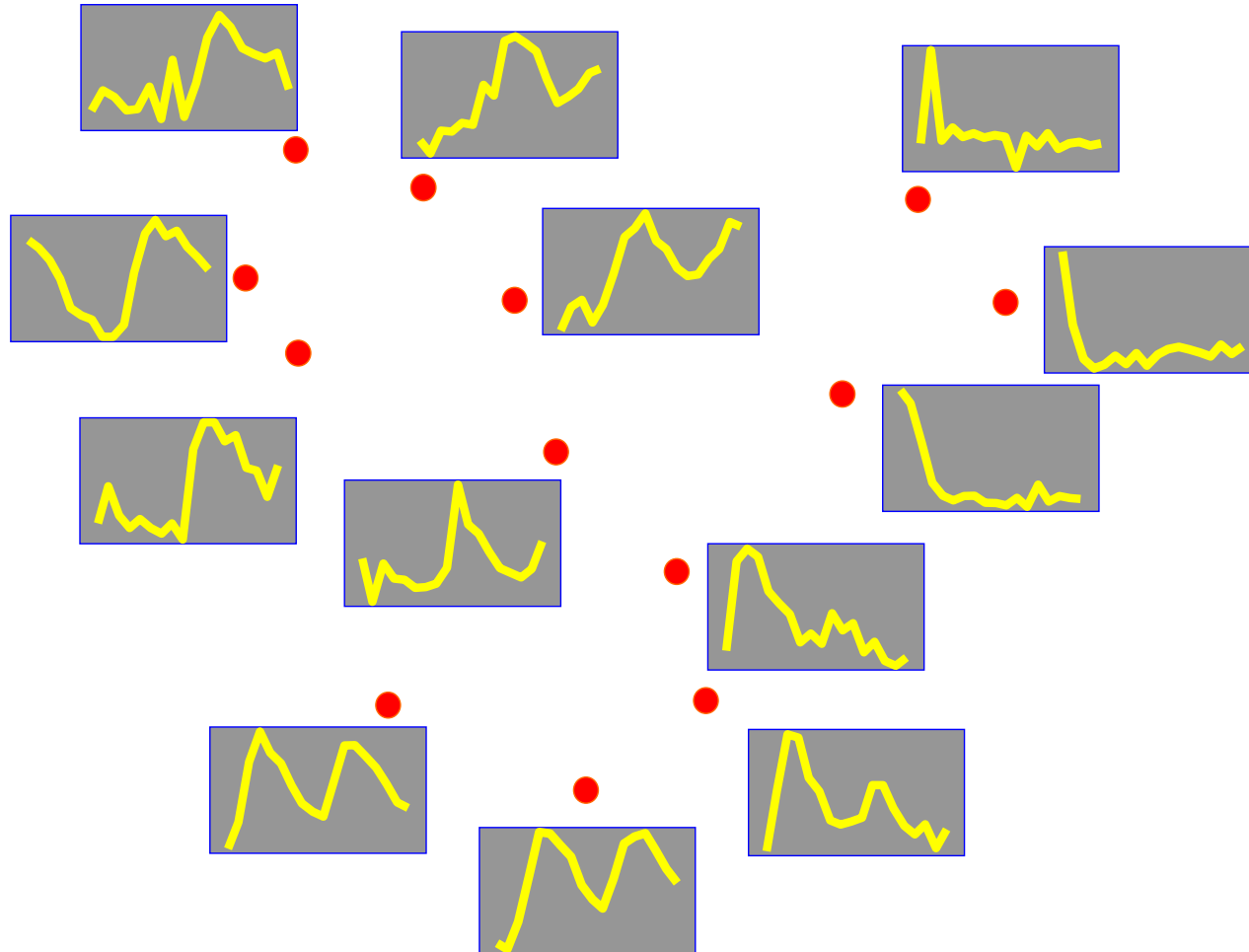
3) Learn model parameters



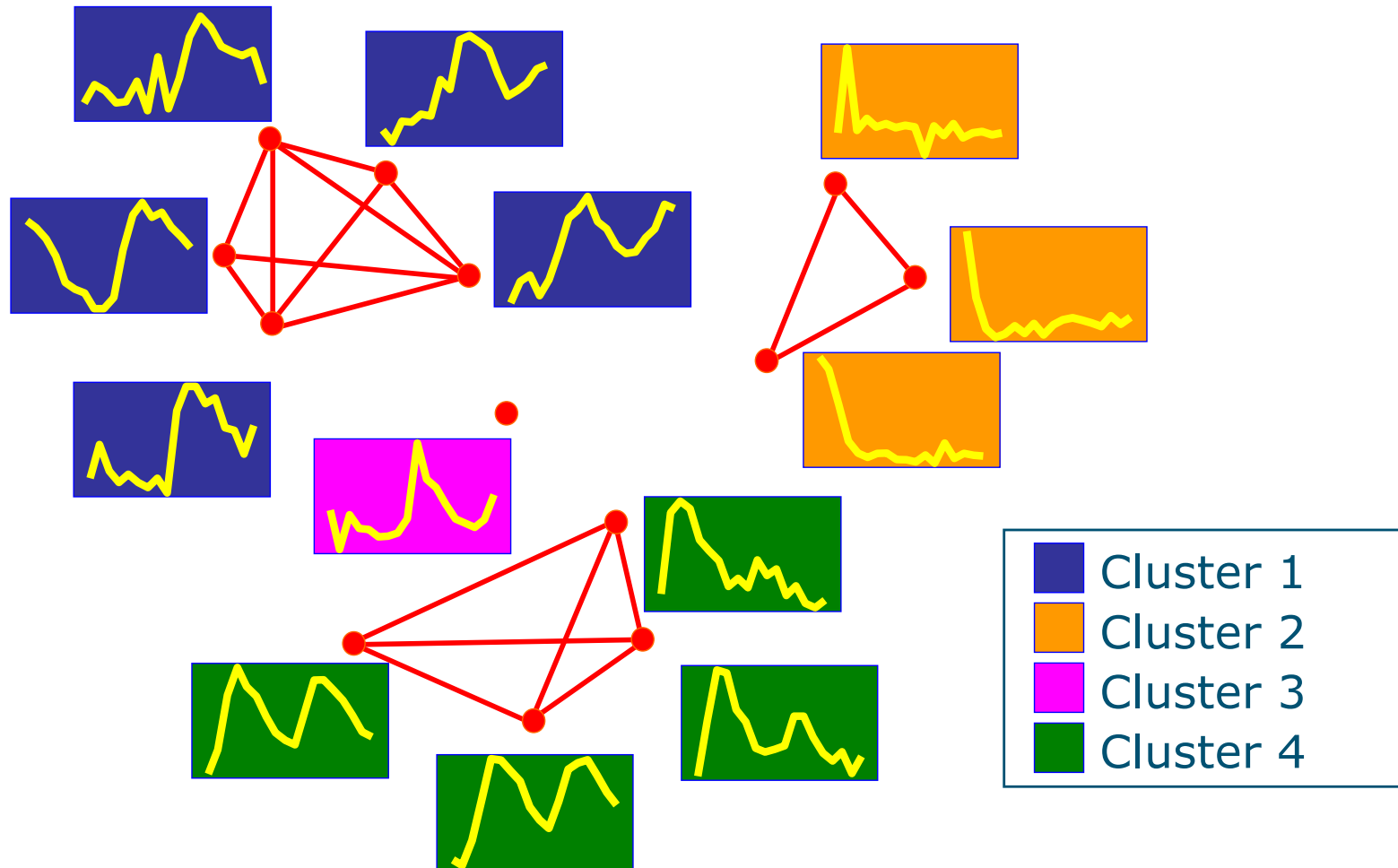
Model of transcription regulation



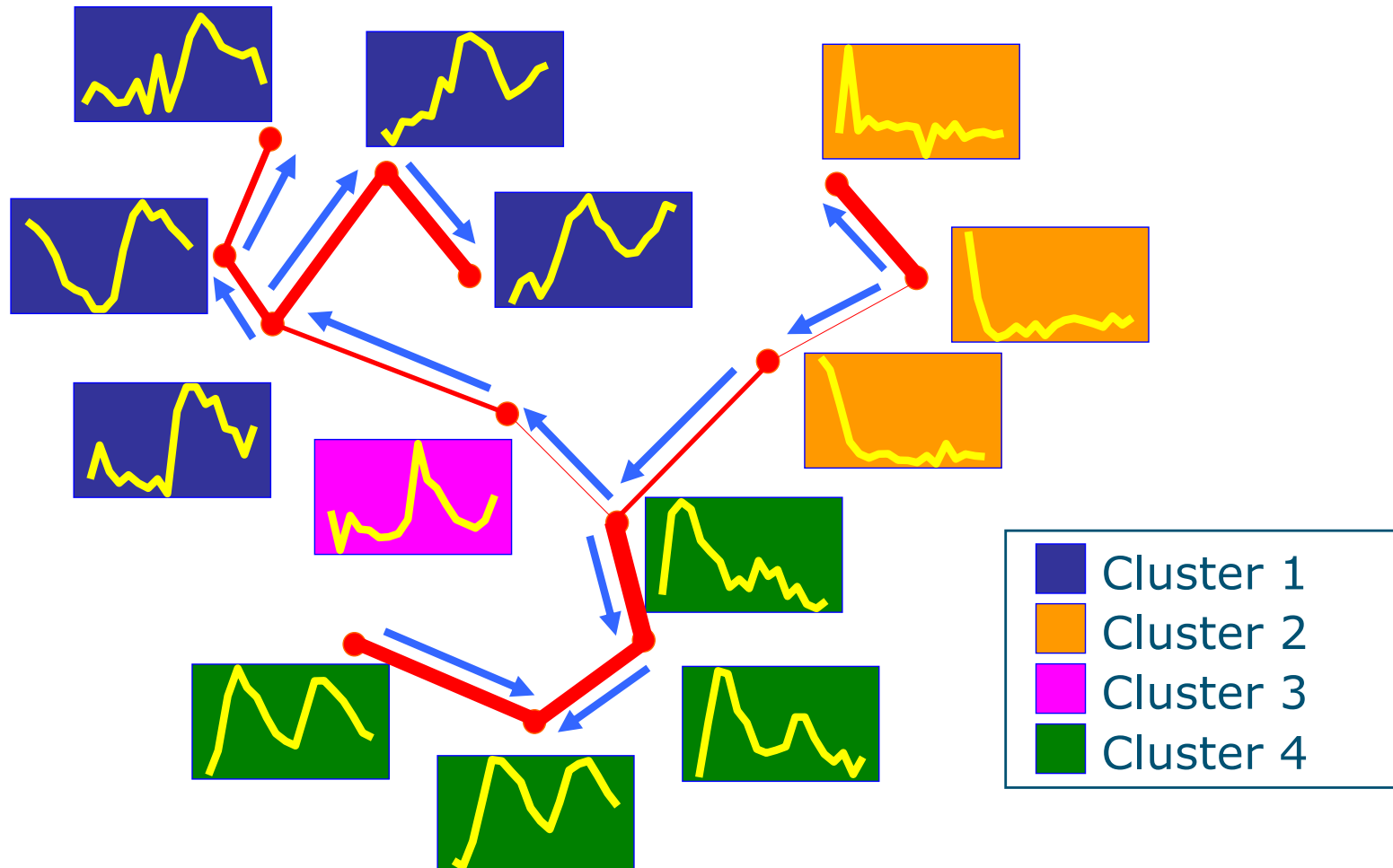
Time series data



Clustering: groups of genes

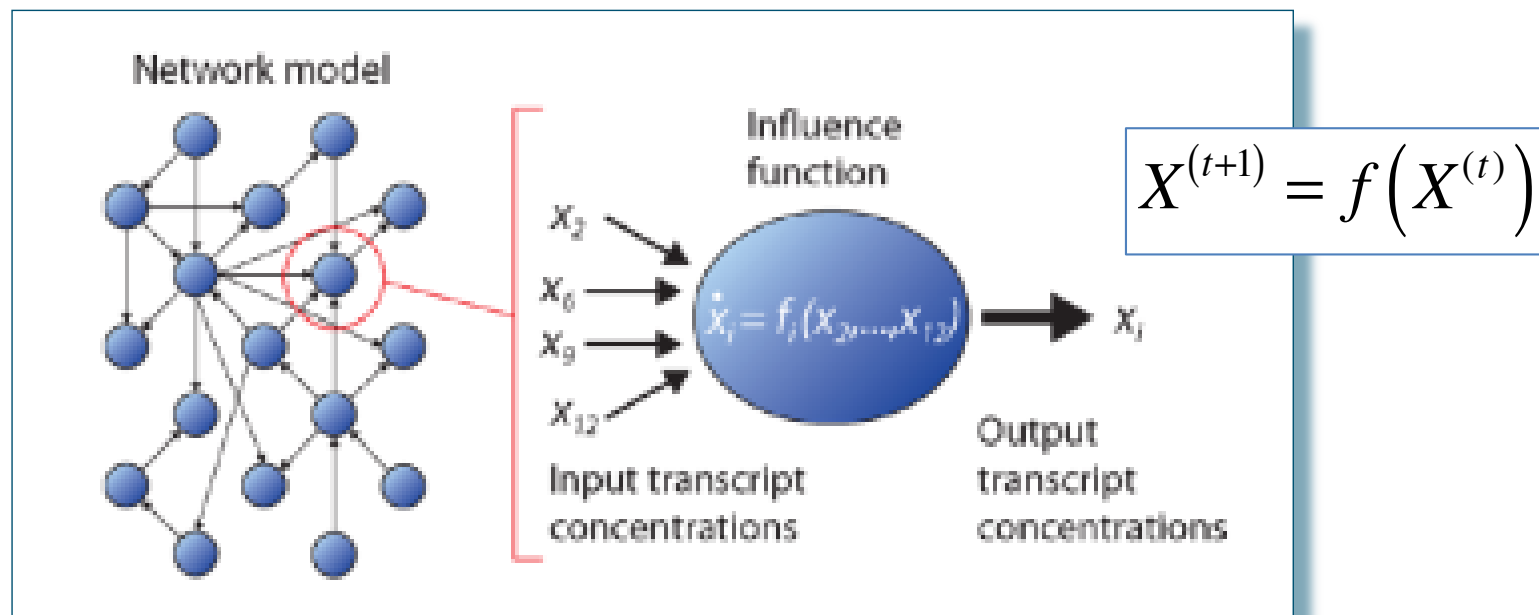


Network inference: derive relationships

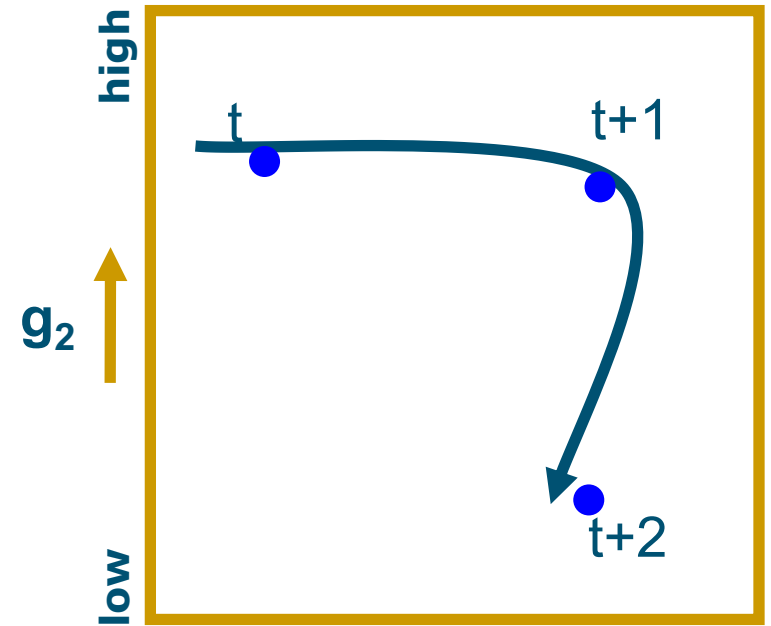
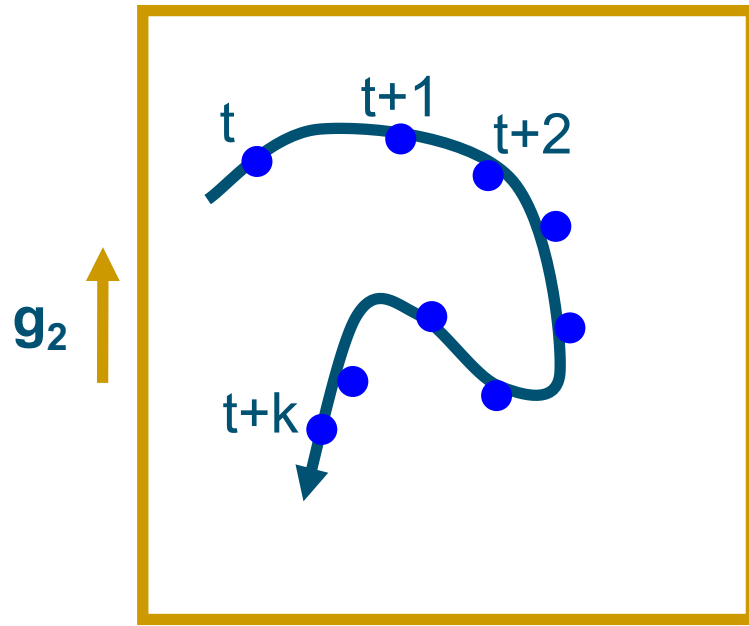


General approach

- Assume (change of) gene expression at time t depends on activity of regulators at time $t-1$
- For regulator activity, take gene expression as proxy
- Number of samples and sampling interval critical for fitting



From time series data to networks



Sample at time t_k :
activity of g_1 & g_2
(each point is one sample)



Trajectory of expression
during the experiment



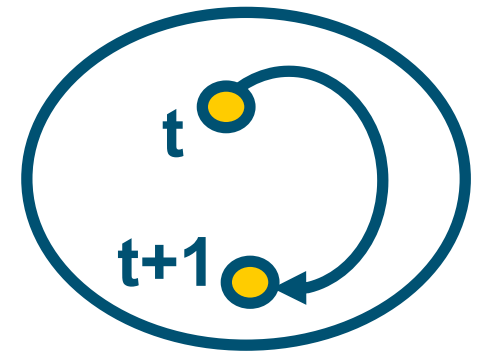
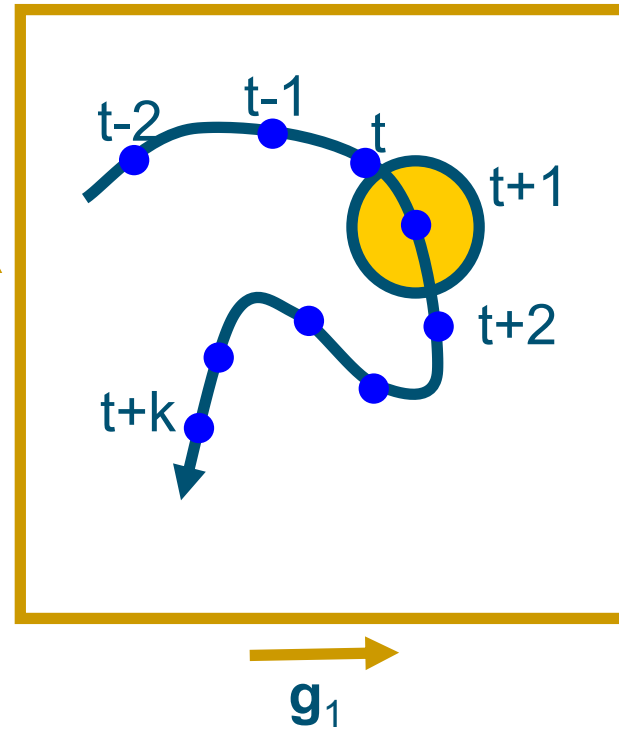
If ($g_2 = \text{high}$)
then ($g_1 \rightarrow \text{high}$)

If ($g_1 = \text{high}$)
then ($g_2 \rightarrow \text{low}$)



State-space model

- State of cell given by expression levels of all genes g_2
- Closed, one-step memory system
- Simplest model: linear, activity of a gene = weighted sum of all genes,



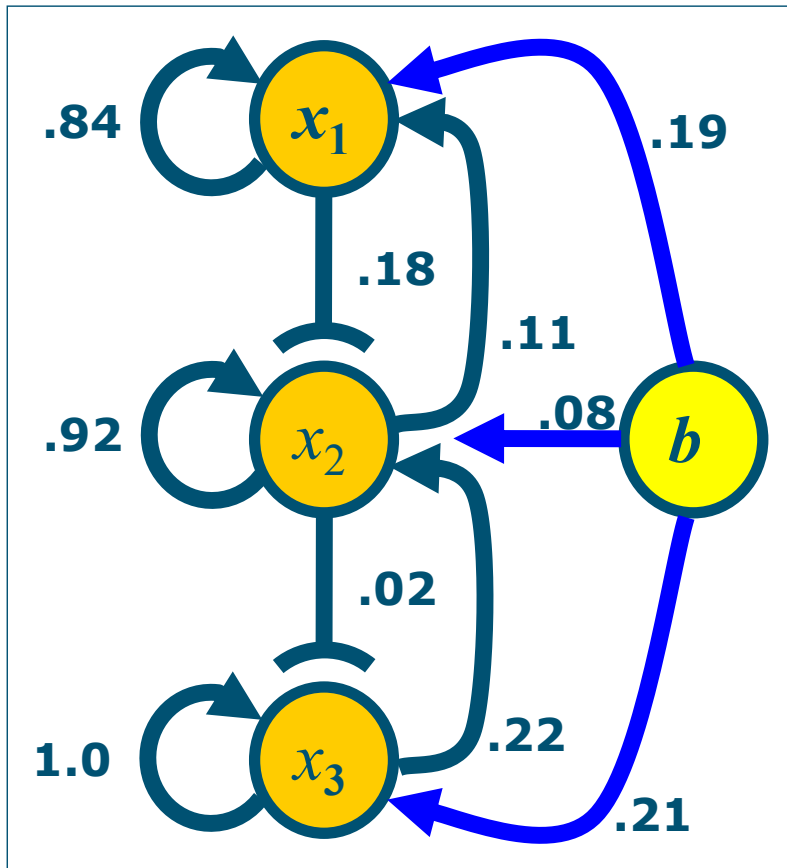
$$g^{(t+1)} = Wg^{(t)} + b$$

find W and b by linear regression

State-space model (2)

- Example: 3 gene network

$$g^{(t+1)} = Wg^{(t)} + b$$



W

.84	-.18	0
.11	.92	-.02
0	.22	1.0

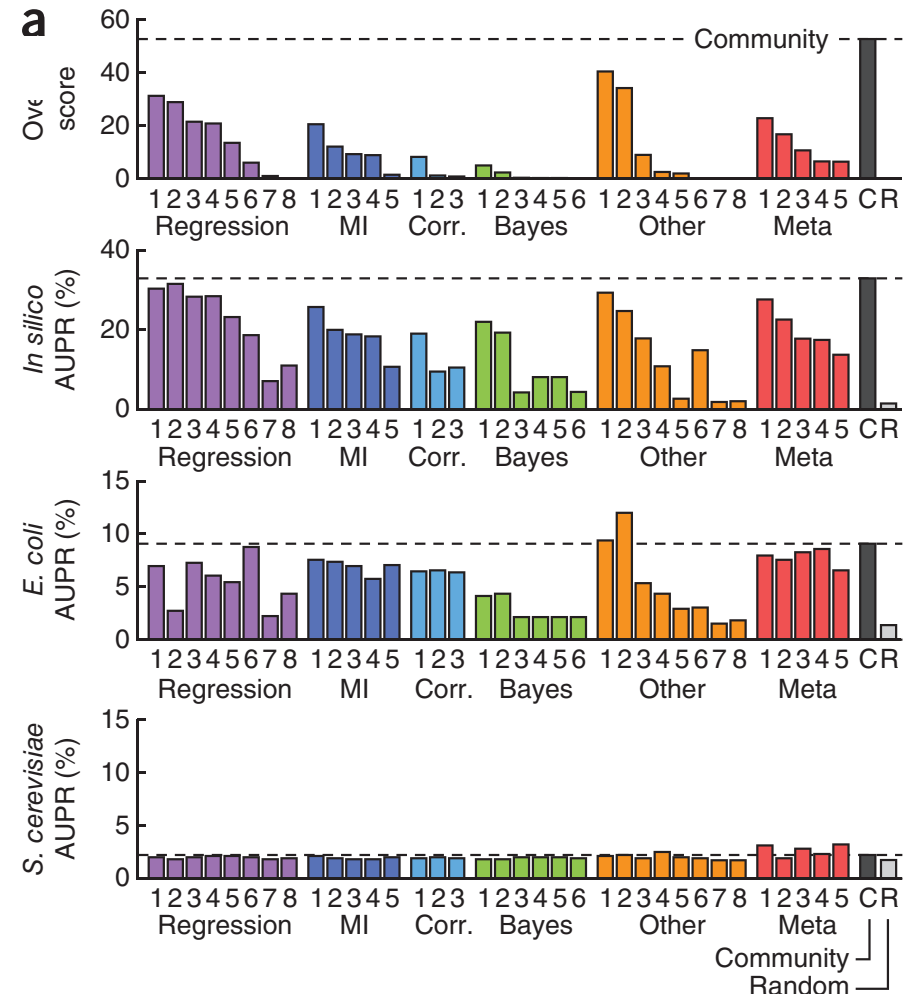
b

.19
.08
-.21



Many other methods

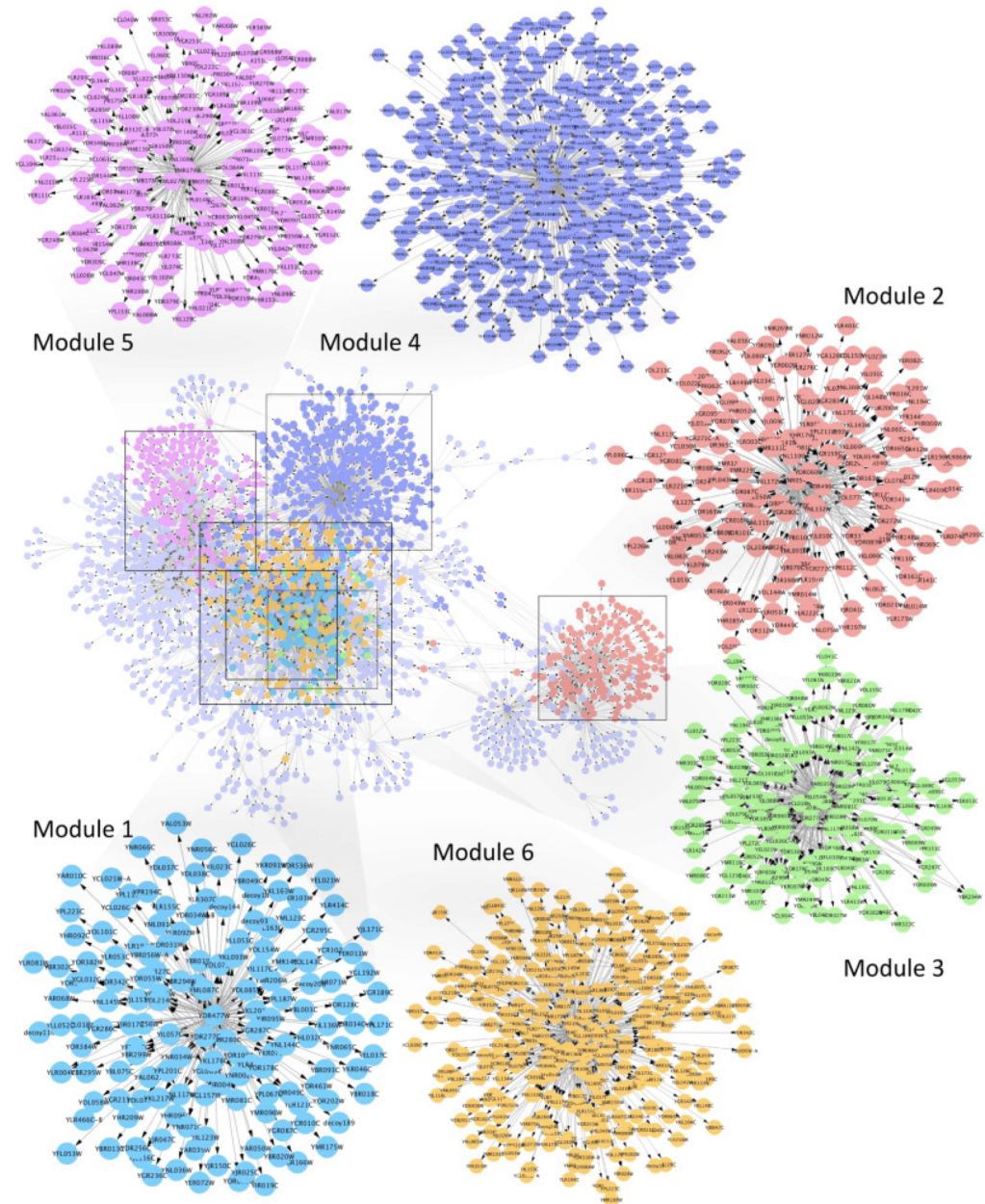
- Association networks:
 - correlation
 - mutual information
- Boolean networks
 - REVEAL
- Bayesian networks
- Dynamical systems (ODEs)
 - Inferelator
- Etc. etc. - see e.g. Hurley, *NAR* 2012 or Huynh-Thu and Sanguinetti, *arXiv* 2018



Marbach *et al.*,
Nature Methods 2012

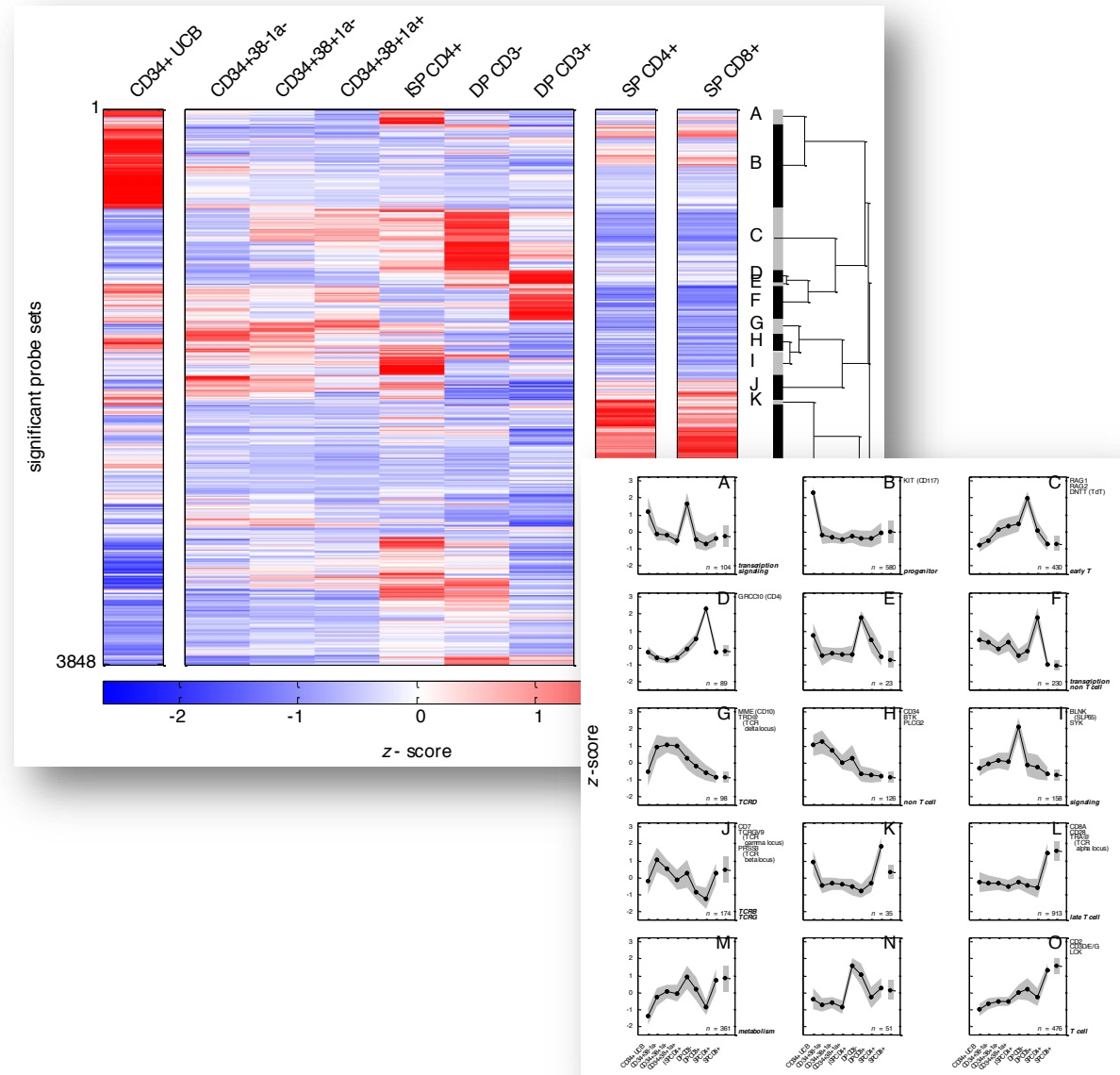
Network mining

- Like in WGCNA: find subnetworks (clusters, modules) that may correspond to specific functions, processes, complexes...
- General idea: clusters/modules have many (high-weight) connections within, and few (low-weight) connections without



Interpretation

- Given a list of significant genes, a cluster or module, what information is available to learn more about it?
- A lot of functional information is known about genes and gene products



Measurement databases

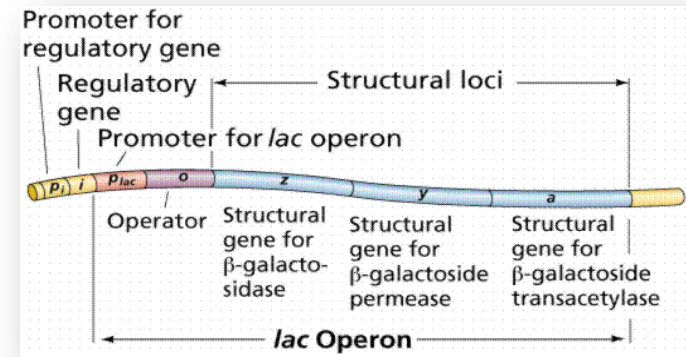
- Genomics: Ensembl, UCSC, RefSeq, Uniprot, GenBank
- Transcriptomics: Gene Expression Omnibus (GEO), EBI ArrayExpress, Stanford Microarray Database (SMD)
- Proteomics: Open Proteomics Database (OPD), Integr8
- Protein-DNA: Biomolecular Network Database (BIND), Encyclopedia of DNA elements (ENCODE)
- Protein-protein: Munich Inf. Center for Prot. Seq. (MIPS), Database of Interacting Proteins (DIP)
- Interactome: General Repository for Interaction Datasets (GRID)
- ...

Databases

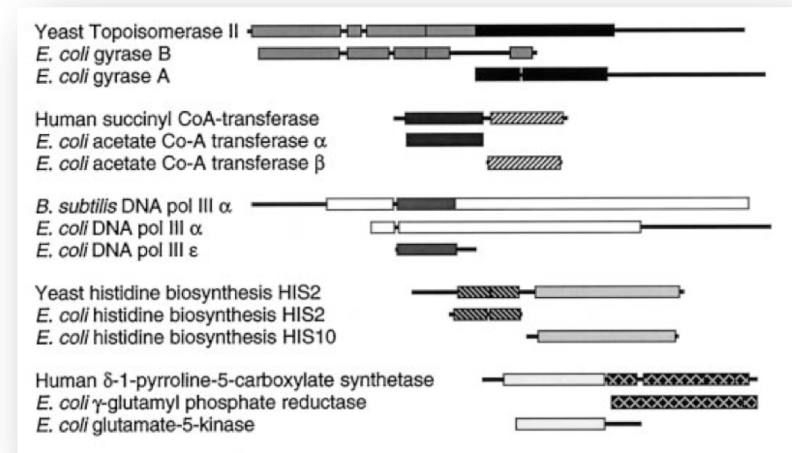
- HuGO: gene names
- Gene Ontology (GO): gene annotations
- TRANSFAC: transcription factors
- TRANSPATH: signalling pathways
- KEGG LIGAND, Brenda: chemical reactions, enzymes
- REACTOME, BIOCARTEA, KEGG: biological pathways
- Saccharomyces Genome Database (SGD)
- PUBMED/MEDLINE: biological references, abstracts
- ...

Sequence features

- Chromosome: genes may be functionally related if...
 - they lie close on the genome
 - example: operons

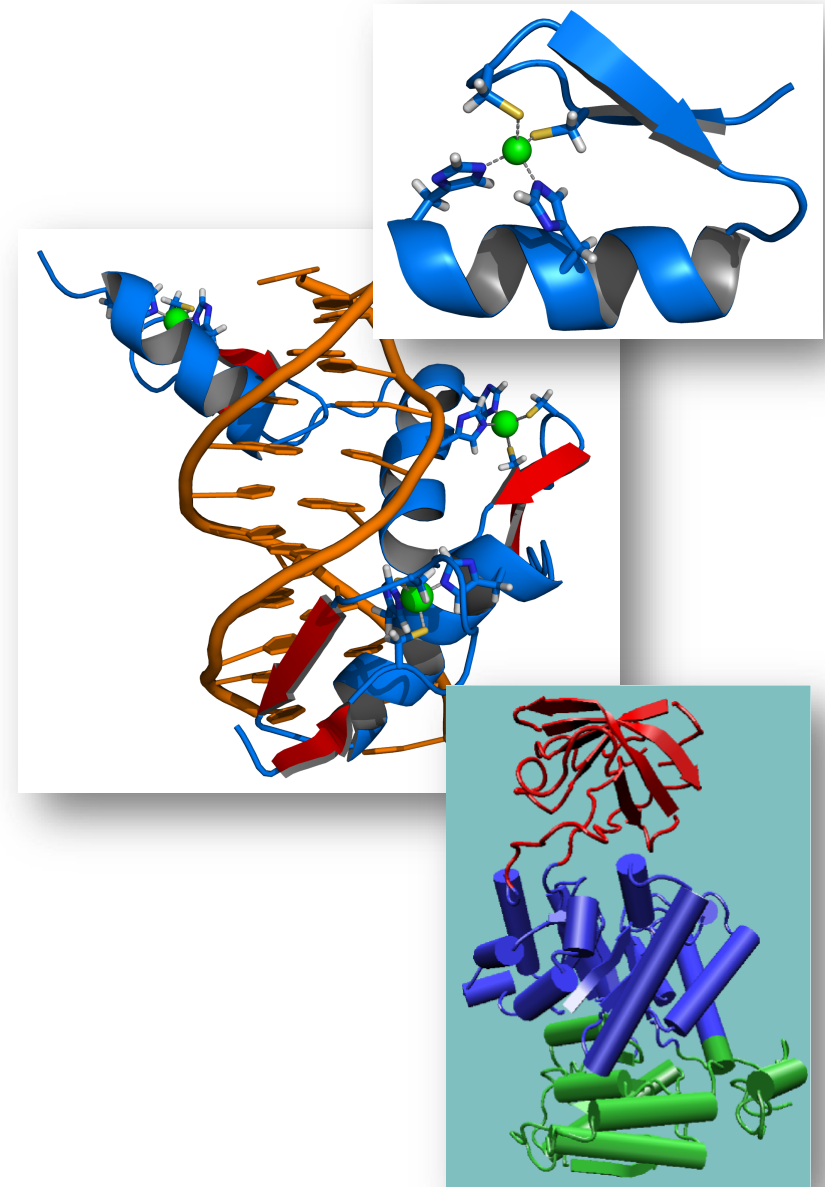


- Sequence: genes may be functionally related if...
 - they share a transcription factor binding site
 - they are homologous to a single other gene
 - example: Rosetta



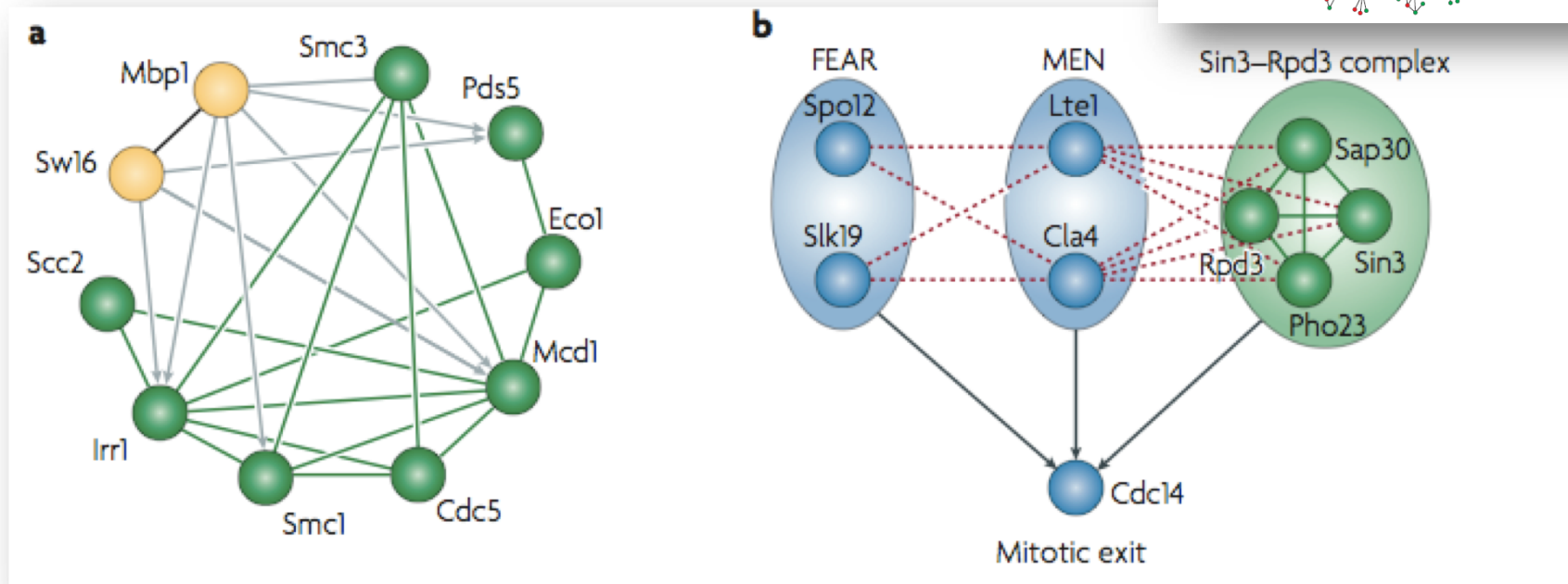
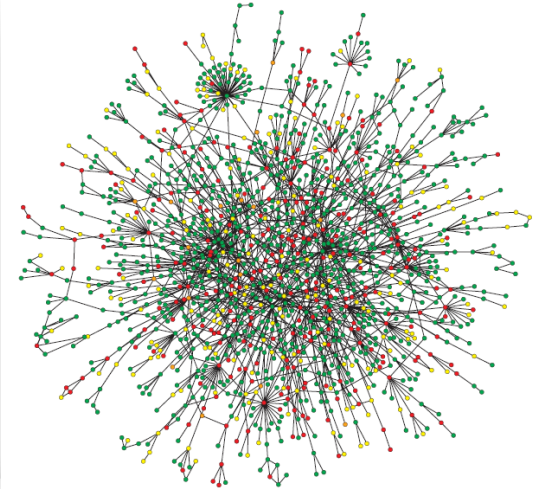
Protein features

- Protein domains: genes may be functionally related if...
 - they share certain structural domains
- Protein families: genes may be functionally related if...
 - their products belong to the same protein (super)family (evolutionary related, but no longer homologous)



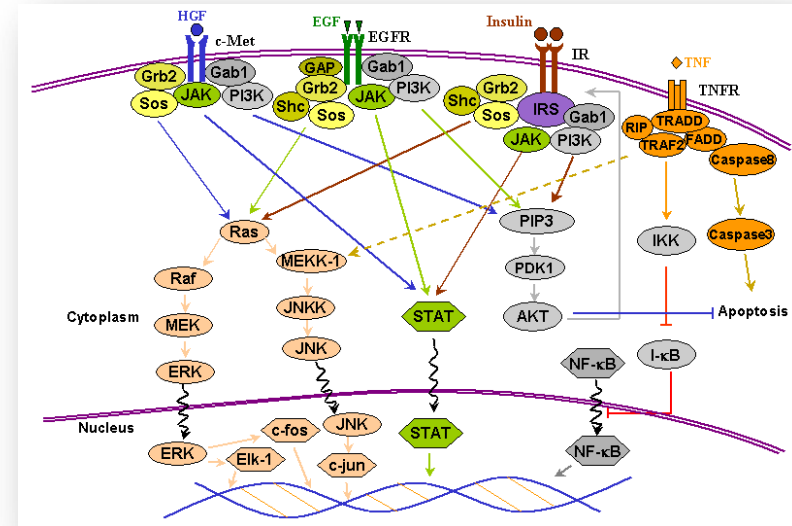
Protein interactions

- Protein interactions: genes may be functionally related if...
 - their products interact in some way, e.g. form a complex

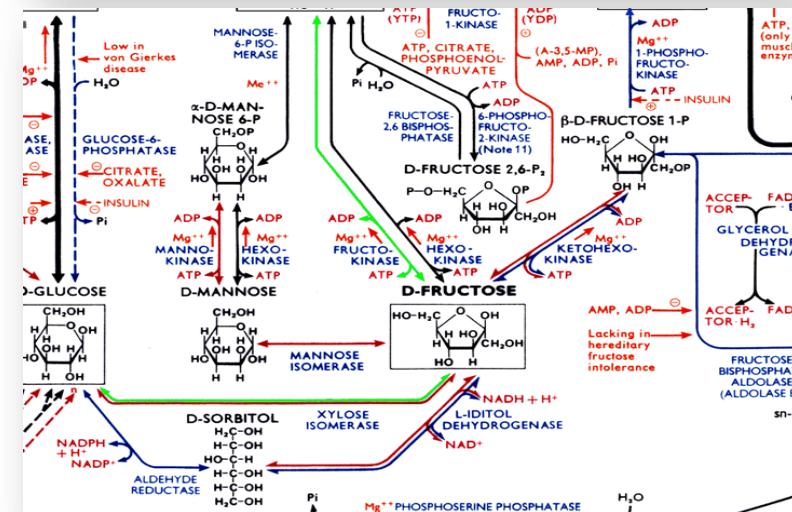


Pathways

- Specific interactions:
 - Signalling pathways



- Metabolic pathways



Phenotypes

- Disease annotation

- Tissue expression

The image shows two overlapping web browser windows. The top window is OMIM (Online Mendelian Inheritance in Man) displaying the entry for MIM #114480, BREAST CANCER, FAMILIAL. The entry includes a search bar, navigation tabs, and a detailed description of the disease, including its genetic basis and clinical features. The bottom window is T1DBase (IGF1 Gene Overview) showing tissue expression data for the IGF1 gene. It features a table with columns for Human, Mouse, and Rat tissues, and a legend for expression levels.

OMIM - BREAST CANCER - Windows Internet Explorer

OMIM #114480
Text
Description
Clinical Features
Other Features
Inheritance
Diagnosis
Clinical Management
Mapping
Cytogenetics
Molecular Genetics
Pathogenesis
Animal Model
See Also
References
Contributors
Creation Date
Edit History

Entrez Gene
Nomenclature
RefSeq
GenBank
Protein

**#114480
BREAST CANCER**

Alternative titles; symbols
**BREAST CANCER, FAMILIAL
BREAST CANCER, FAMILIAL MALE, INCLUDED**

Gene map locus [17q22-q23](#), [17q22](#), [17p13.1](#), [16p12](#), [15q15.1](#), [14q32.3](#), [13q12.3](#), [12p12.1](#), [11q22.3](#), [11p15.5](#), [8q11](#), [5q33.2](#), [3q26.3](#), [2q34-q35](#), [2q33](#), [22q12.1](#)

TEXT

A number sign (#) is used with this entry because of evidence that mutation at more than one locus can be involved in different families or even in the same case. These loci include BRCA1 ([113705](#)) on 17q, BRCA2 ([600185](#)) on 13q12, BRCATA ([600048](#)) on 11q, BRCA3 ([605365](#)) on 13q21, BWSERIA ([602631](#)) on 11p15.5, the TP53 gene ([191170](#)) on 17p, the BRIP1 gene ([605882](#)) on 17q22, and the RB1CC1 gene ([606837](#)) on 8q11. Mutations in the androgen receptor gene (AR; [313700](#)) on the X chromosome have been found in cases of male breast cancer ([313700.0016](#)). Mutation in the RAD51 gene ([179617](#)) was found in patients with familial breast cancer ([179617.0001](#)). Breast cancer susceptibility alleles have been reported in the CHEK2 gene (see [604373.0001](#) and [604373.0012](#)) and in the BARD1 gene (see [601593.0001](#)).

T1DBase - IGF1 Gene Overview - Windows Internet Explorer

Users should be aware that the scale represents a **rank** within an experiment rather than a normalized expression signal.

Human				Mouse			Rat					
ductal cells	exocrine pancreas	pancreatic islets	primary beta cells	Pancreatic Islets MPSS	beta cell line	pancreatic islets	whole pancreas	alpha cell	beta cell line	pancreatic islets	primary beta cells	whole pancreas
				no data								

Expression Legend

Tissue Expression Top Help

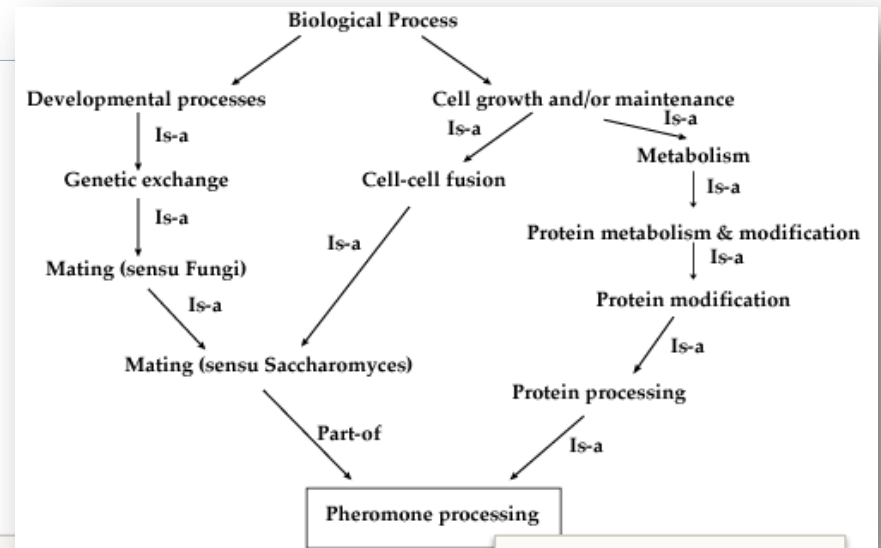
Users should be aware that the scale represents a **rank** within an experiment rather than a normalized expression signal.

PB CD14+ Monocytes	PB CD56+ NK Cells	PB CD34+ B Cells	PB CD4+ T Cells	PB CD8+ T Cells	PB BCR4+ Benthritic Cells	T21 B Lymphoblasts	Lymph Node	Thymus	Salivary Gland	Bone Marrow	BM CD34+	Heart	Skeletal Muscle	Kidney	Liver	Lung	Pancreas	Pancreatic Islets	Trachea	Placenta	Thyroid	Testis	Caudate Nucleus	Cerebellum	Hypothalamus	Spiral cord	Thalamus	Skin	Ovary	Placenta	Prostate	Testis	Uterus

Expression Legend

Gene ontology

- Three ontologies:
 - Biological processes (BP)
 - Molecular functions (MF)
 - Cellular components (CC)



id: GO:0007323

name: peptide pheromone maturation

namespace: biological_process

alt_id: GO:0007324

alt_id: GO:0007326

alt_id: GO:0046613

def: "The generation of a mature, active peptide pheromone via p
unique to its processing and modification. An example of t
is found in *Saccharomyces cerevisiae*." [GOC:elh]

synonym: "a-factor processing (proteolytic)" NARROW []

synonym: "alpha-factor maturation" NARROW []

synonym: "pheromone processing" EXACT []

is_a: GO:0016485 ! protein processing

AXL1

KEX2

KEX2

krp1

mug138

RAM2

RCE1

SPAC1687.02

SPAC3H1.05

STE13

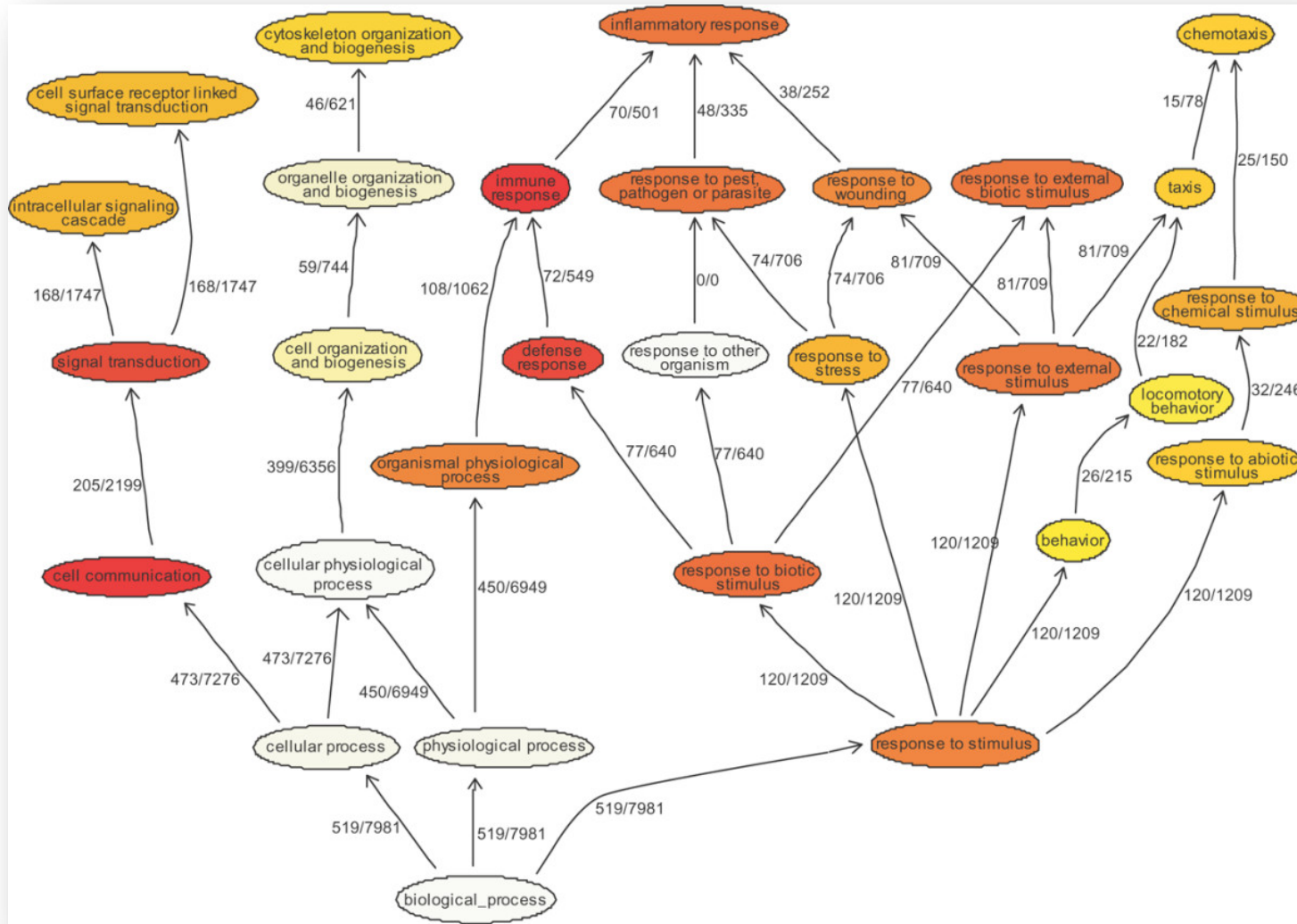
STE14

STE23

STE24

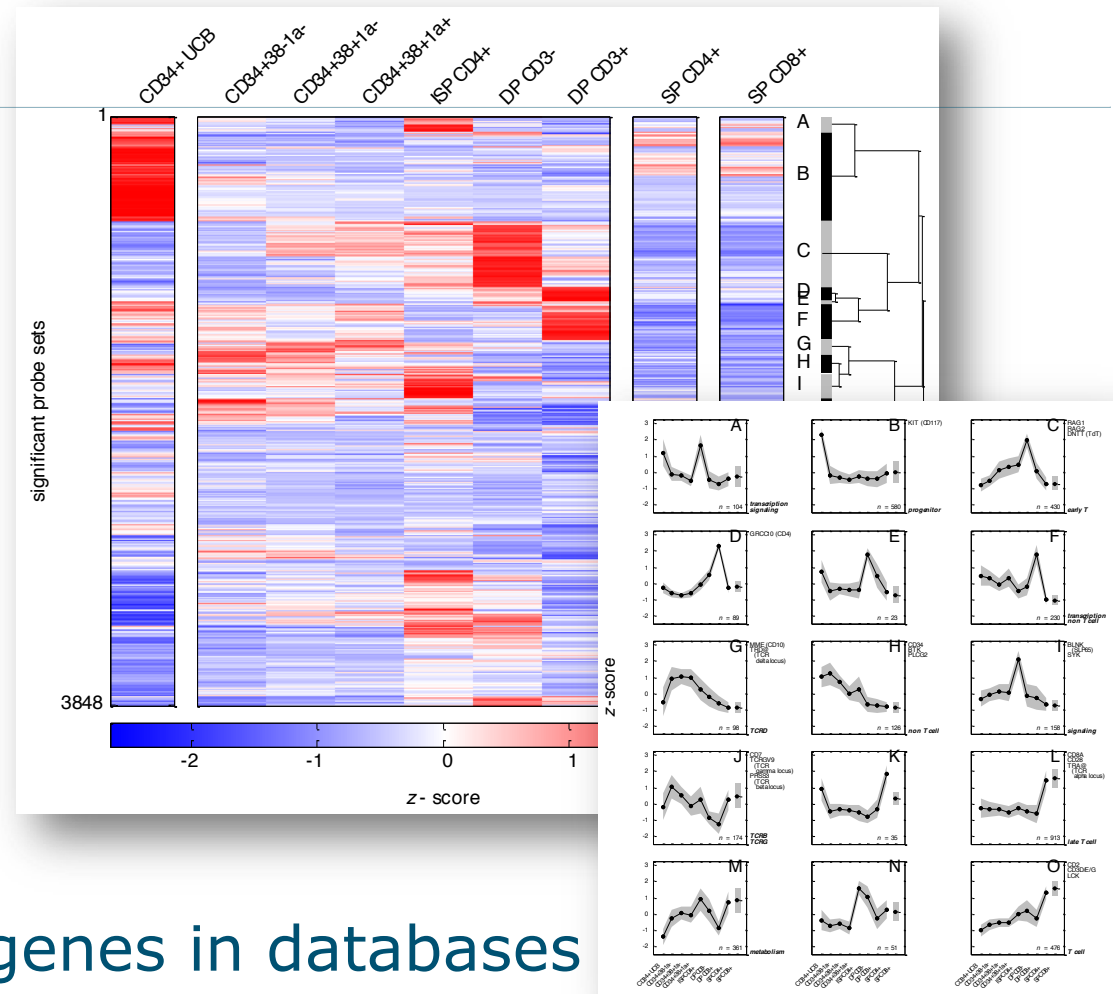


Gene ontology (2)



Annotation (2)

- How do we use this information to help interpret a list of significant genes or a cluster?



- **Annotation:** look up genes in databases
- **Enrichment:** look for significant annotations in gene list
- **Prioritization:** order genes on relation to phenotype

Enrichment: Fisher's exact test

- Statistical test for association due to R.A. Fisher
- Also known as the hypergeometric test



Test setup

- $n = 8$ cups of tea
- $a = 3$ with milk first, $b = 5$ with tea first



Fisher's exact test

- Remember from statistics: hypothesis tests
 - *null hypothesis: assume there is no real association between pouring order and the ladies choice*
 - what is the probability of finding an association in an experiment *by chance*?
 - if this is probability is low, the *assumption is likely incorrect* : she can really tell the difference

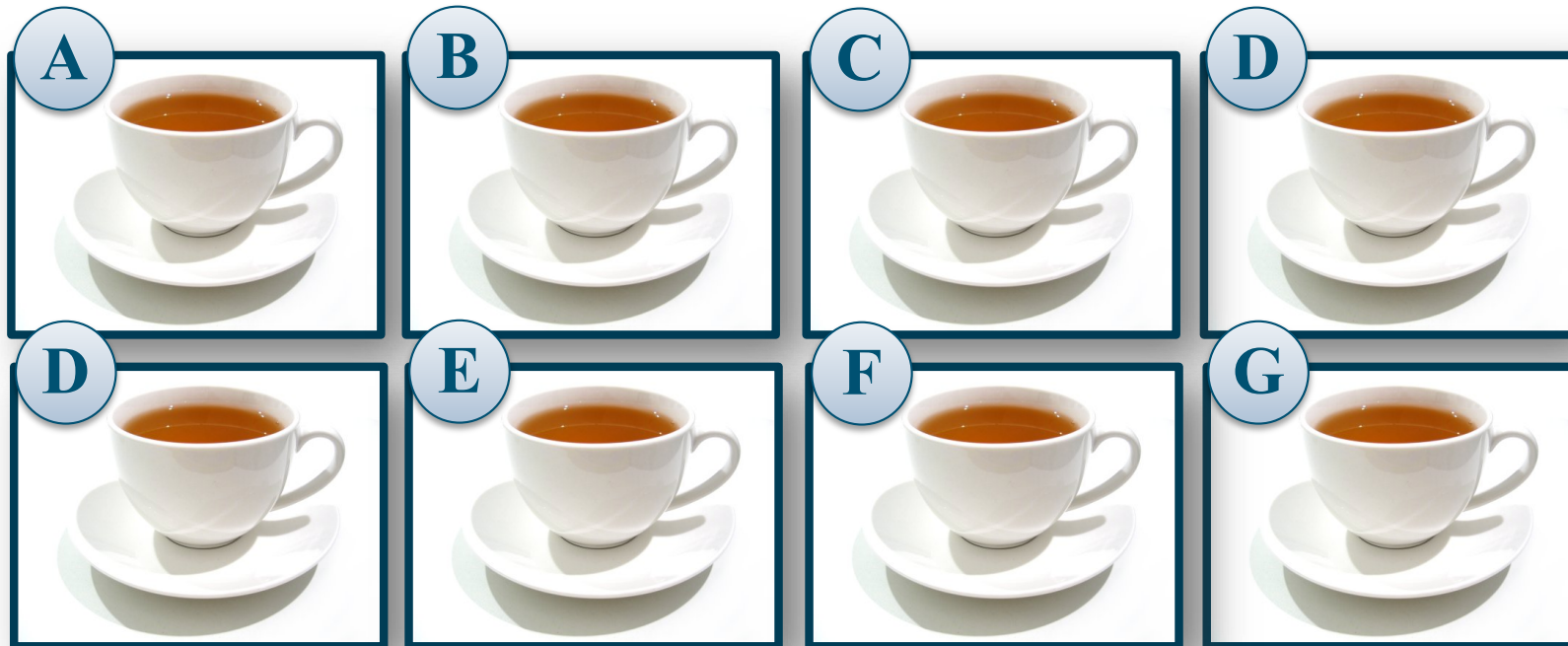
Fisher's exact test (2)

- Test setup:
 - Present cups in random order
 - Ask the lady to pick the three "milk first" cups
 - Null hypothesis: H_0 : choice is random
- The lady picks 2 cups correctly
- What is the probability of this happening under H_0 ?

$$\frac{\text{\#ways of picking 2 "milk first" and 1 "tea first" cups}}{\text{total \#ways of picking 3 cups}}$$

Fisher's exact test (3)

- What is the total number ways in which she could choose 3 cups *in a specific order*?

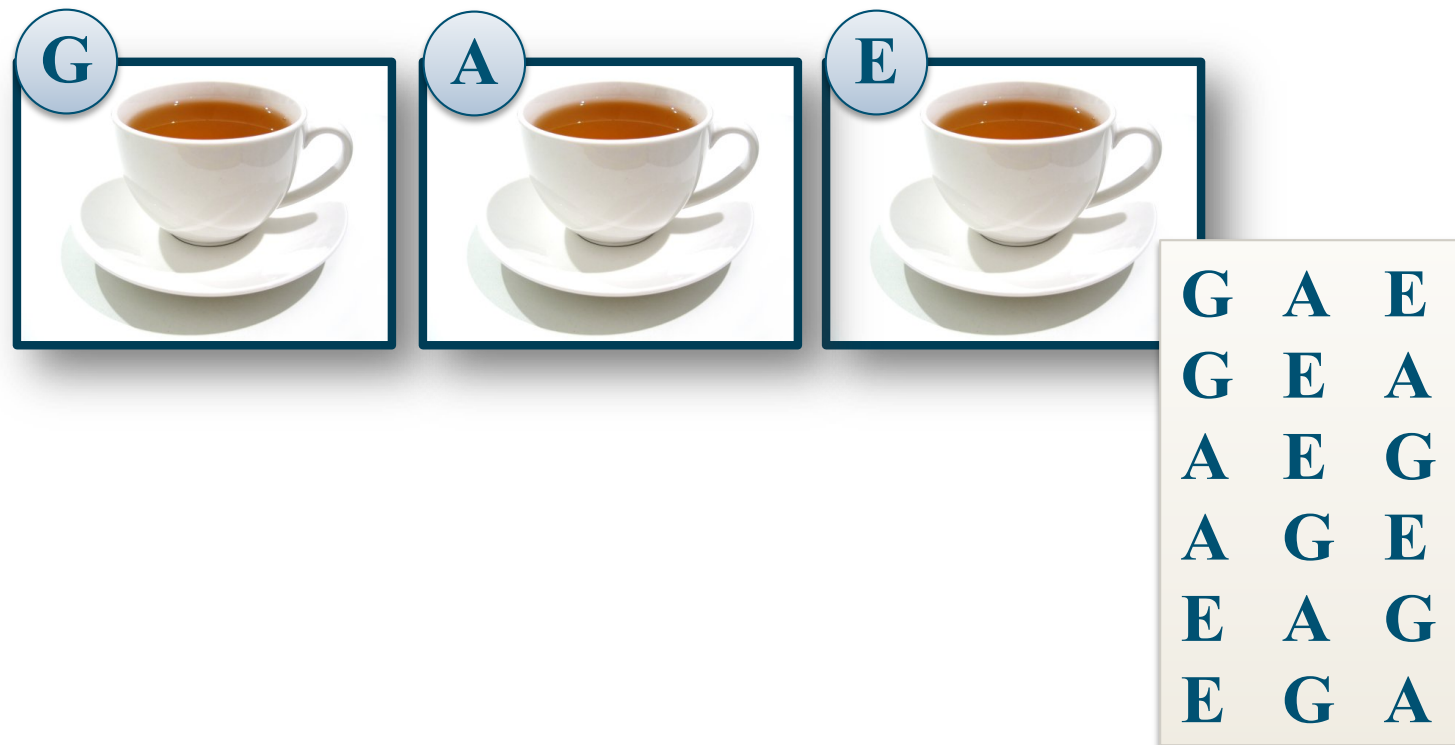


8·7·6



Fisher's exact test (4)

- How many ways are there of ordering 3 cups?



3·2·1

Fisher's exact test (5)

- What is the total number ways in which she could choose 3 cups *in a specific order*?

$$8 \cdot 7 \cdot 6$$

- How many ways are there of ordering 3 cups?

$$3 \cdot 2 \cdot 1$$

- What is the total number ways in which she could choose 3 cups *in any order*?

$$\frac{8 \cdot 7 \cdot 6}{3 \cdot 2 \cdot 1} = \frac{(8 \cdot 7 \cdot 6) \cdot (5 \cdot 4 \cdot 3 \cdot 2 \cdot 1)}{(3 \cdot 2 \cdot 1) \cdot (5 \cdot 4 \cdot 3 \cdot 2 \cdot 1)} = \frac{8!}{3! \cdot 5!} = \binom{8}{5} = \binom{8}{3}$$

Fisher's exact test (6)

- The lady picks 2 cups correctly
- What is the probability of this happening under H_0 ?

$$\frac{\text{\#ways of picking 2 "milk first" and 1 "tea first" cups}}{\text{total \#ways of picking 3 cups}}$$

Fisher's exact test (7)

- What is the total number ways in which she could choose 2 "milk first" cups out of 3 *in any order*?

$$\binom{3}{2}$$

- What is the total number ways in which she could choose 1 "tea first" cup out of 5?

$$\binom{5}{1}$$

Fisher's exact test (8)

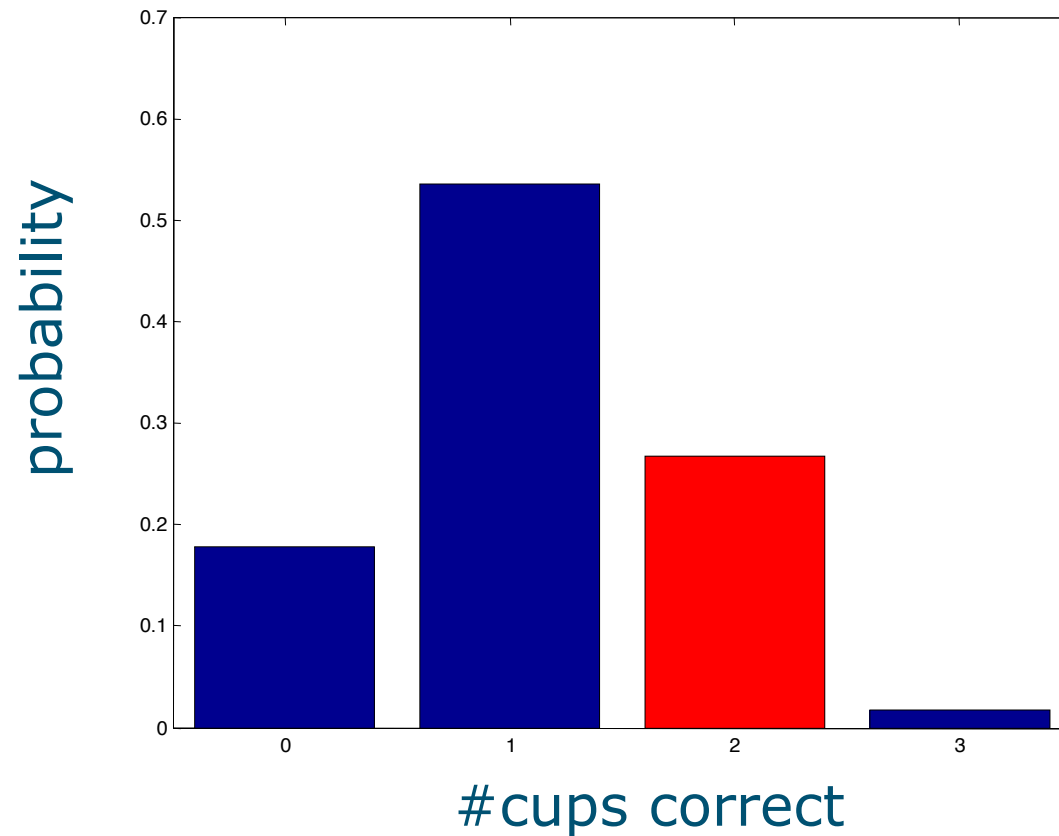
- The lady picks 2 cups correctly
- What is the probability of this happening under H_0 ?

$$\frac{\text{\#ways of picking 2 "milk first" and 1 "tea first" cups}}{\text{total \#ways of picking 3 cups}}$$

$$= \frac{\binom{3}{2} \binom{5}{1}}{\binom{8}{3}} = \frac{3 \cdot 5}{56} \approx 0.27$$

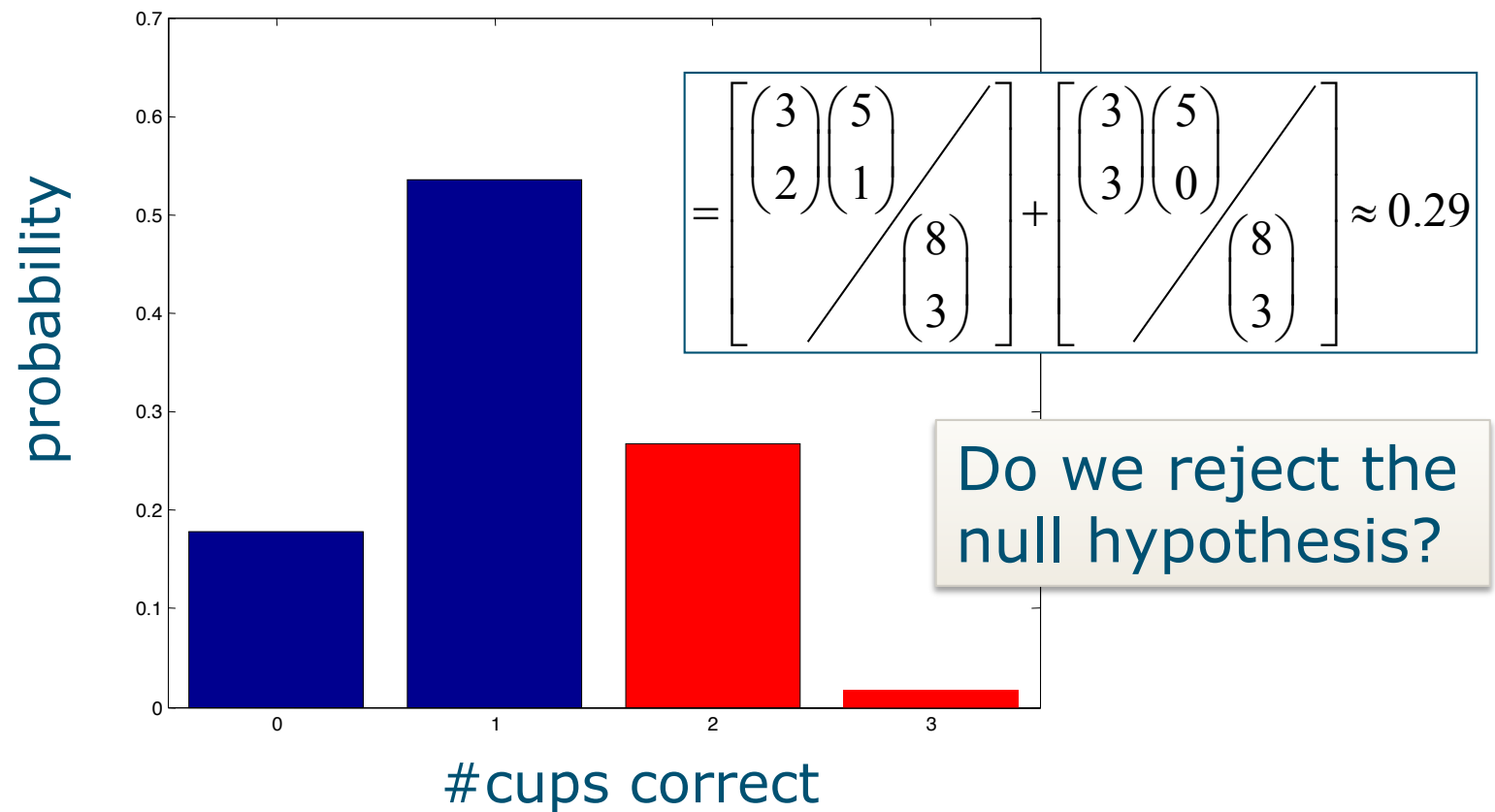
Fisher's exact test (9)

- The lady picks 2 cups correctly
- What is the probability of this happening under H_0 ?



Fisher's exact test (10)

- p -value: what is the probability of picking *at least* 2 cups correctly under H_0 ?



Fisher's exact test (11)

- More generally:
 - n balls
 - a green ones
 - b red ones
 - draw k balls
 - what is the probability of finding at least m green balls by chance?

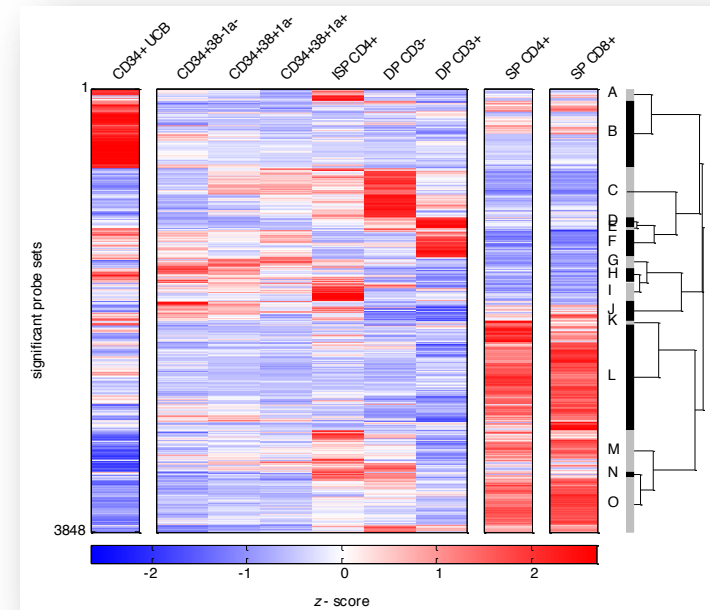


	Drawn	Not drawn	Total
Green	m	$a-m$	a
Red	$k-m$	$b-(k-m)$	b
Total	k	$a+b-k$	n



Fisher's exact test (12)

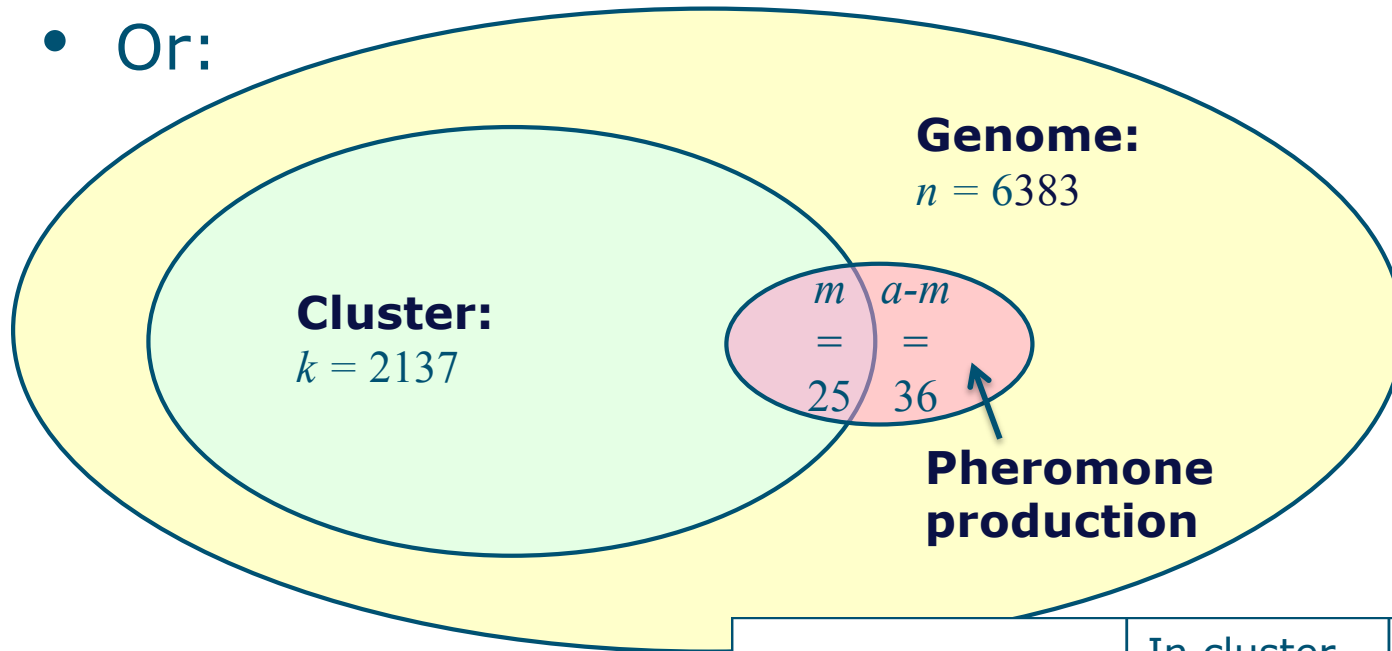
- More generally:
 - n genes
 - a "pheromone production"
 - b not
 - find a cluster of k genes
 - what is the probability of finding at least m "pheromone genes" by chance?



	In cluster	Not in cluster	Total
Pheromone pr.	m	$a-m$	a
Non-pherom. pr.	$k-m$	$b-(k-m)$	b
Total	k	$a+b-k$	n

Fisher's exact test (13)

- Or:



	In cluster	Not in cluster	Total
Pheromone pr.	m	$a - m$	a
Non-pherom. pr.	$k - m$	$b - (k - m)$	b
Total	k	$a + b - k$	n

Fisher's exact test (14)

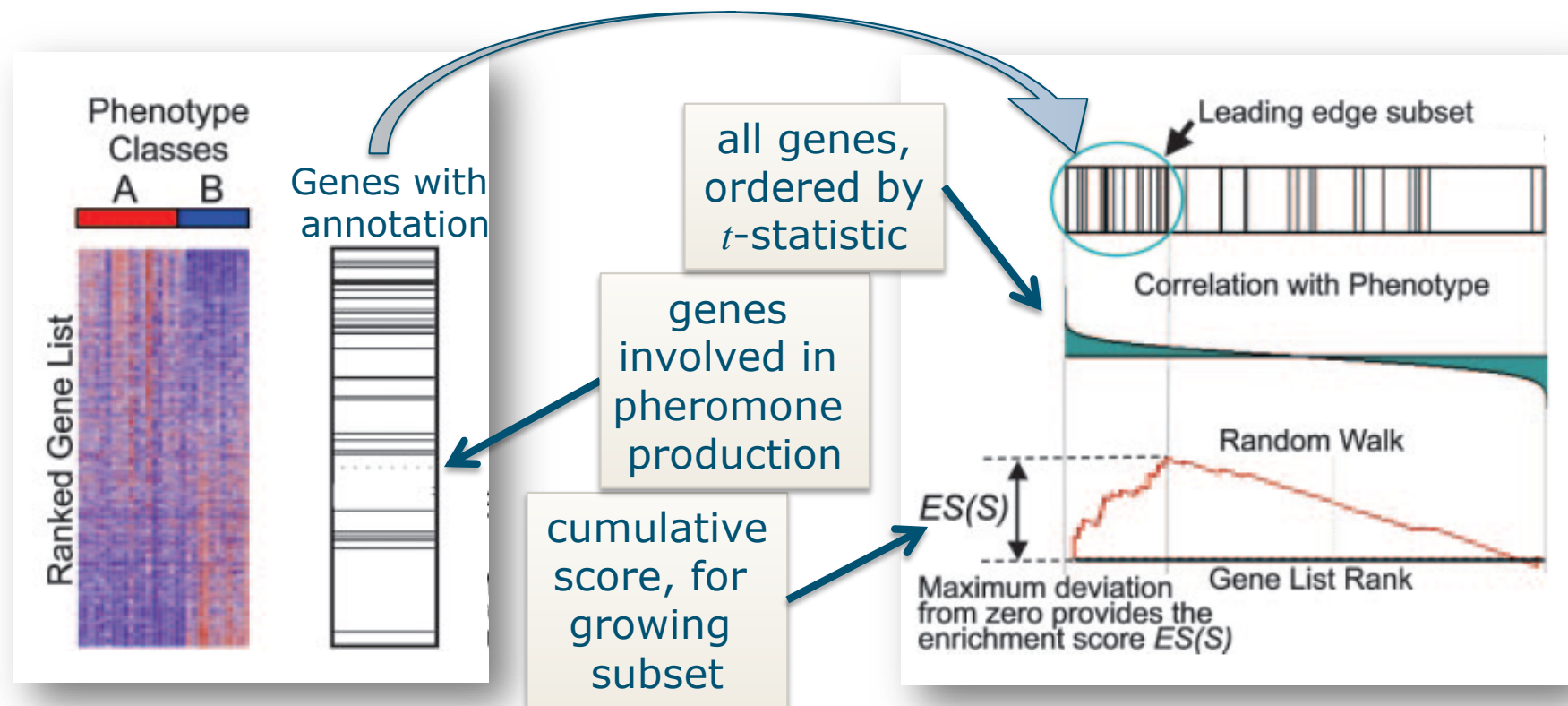
- If you test multiple annotations, adjust for multiple testing, e.g. using FDR or Bonferoni: multiply each p -value by the number of statistical tests
- For example, testing 30,000 GO annotations: significant at $p < 0.05/30,000 = 1.67 \times 10^{-6}$

Gene set enrichment analysis

- Standard high-throughput experiment:
 - Perform an experiment with two conditions, check for significant differential expression, e.g.: perform a t -test for each gene, calculate p -value
 - Adjust for multiple testing (Bonferoni)
 - Select only genes with $p_{\text{adj}} < 0.05$ or $p_{\text{adj}} < 0.01$
- Alternatively:
 - Cluster genes using time series or set of conditions
- Problem:
 - Result is often a very small set of genes
 - Consequently, Fisher's exact test will never give significant enrichments

Gene set enrichment analysis (2)

- Alternative: check whether *ranking* of genes based on *t*-test is associated with a certain annotation (no *p*-value threshold!)



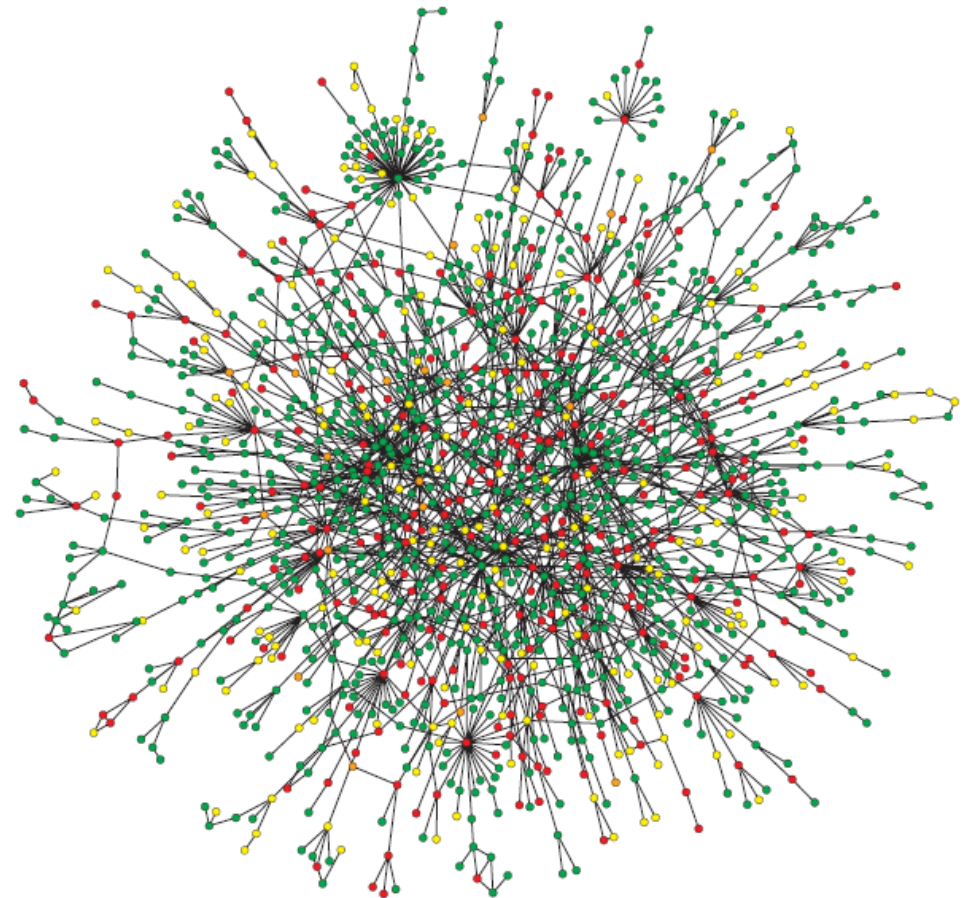
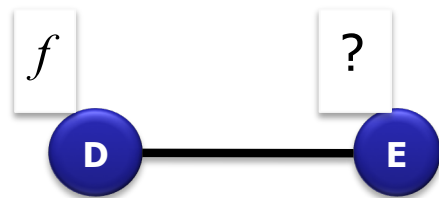
Online tools

- For human and other model organisms:
 - DAVID, <https://david.ncifcrf.gov/>
 - GOrilla, <http://cbl-gorilla.cs.technion.ac.il/>
 - GSEA, <http://software.broadinstitute.org/gsea/>

- For plants:
 - AgriGO, <http://bioinfo.cau.edu.cn/agriGO/>
 - PlantGSEA, <http://structuralbiology.cau.edu.cn/PlantGSEA/>
 - gProfiler, <http://biit.cs.ut.ee/gprofiler/>

Network-based analysis

- Interpret genes and gene lists by looking at their neighbourhood in a network of interacting genes
- “Guilt-by-association”:
if a gene A is linked to another gene B with a known function, it may also have that function

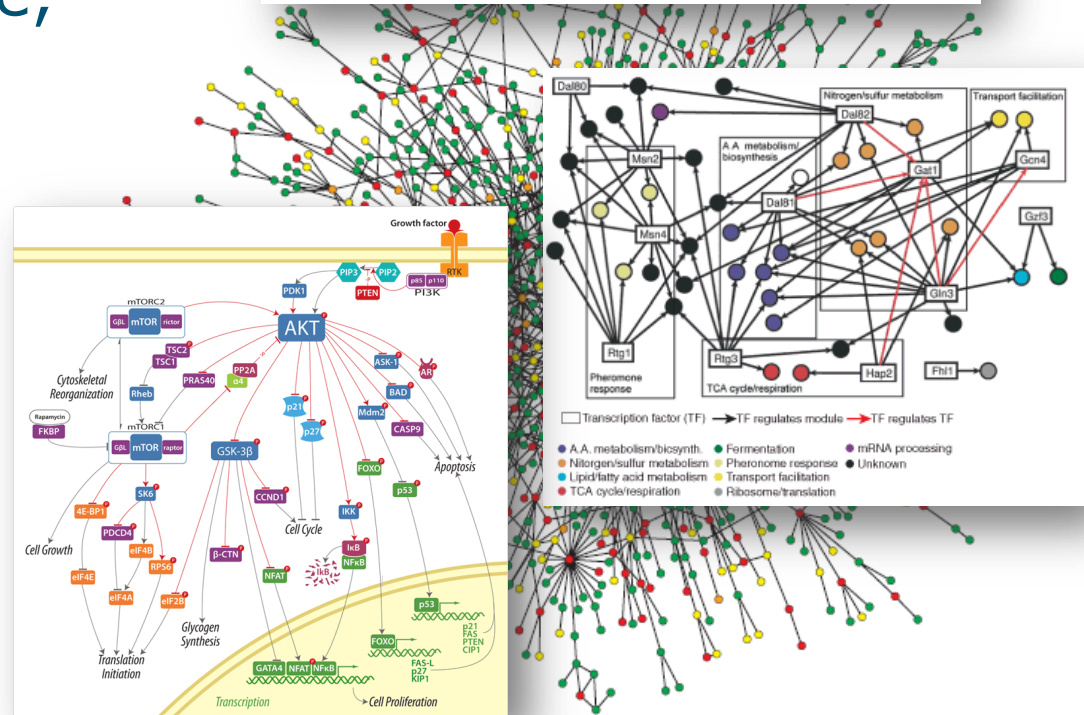
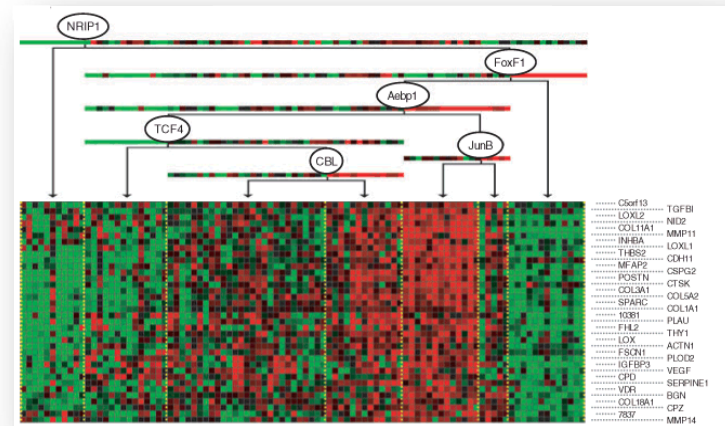


Network-based analysis (2)

- Interaction type is important!

- physical (protein-protein)
- regulatory (protein-DNA)
 - TF2Network
- functional (gene-gene, often predicted)
 - GeneMania (model organisms)
 - STRING
 - AraNet

- Functional interactions most informative

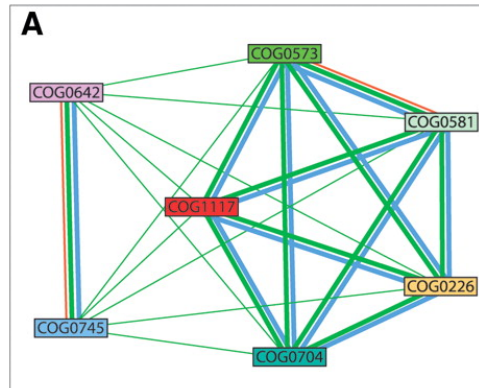


STRING

- Search Tool for the Retrieval of Interacting Genes

- Predicts functional interactions based on co-expression, co-evolution, homology, literature etc.

- <http://string-db.org/>



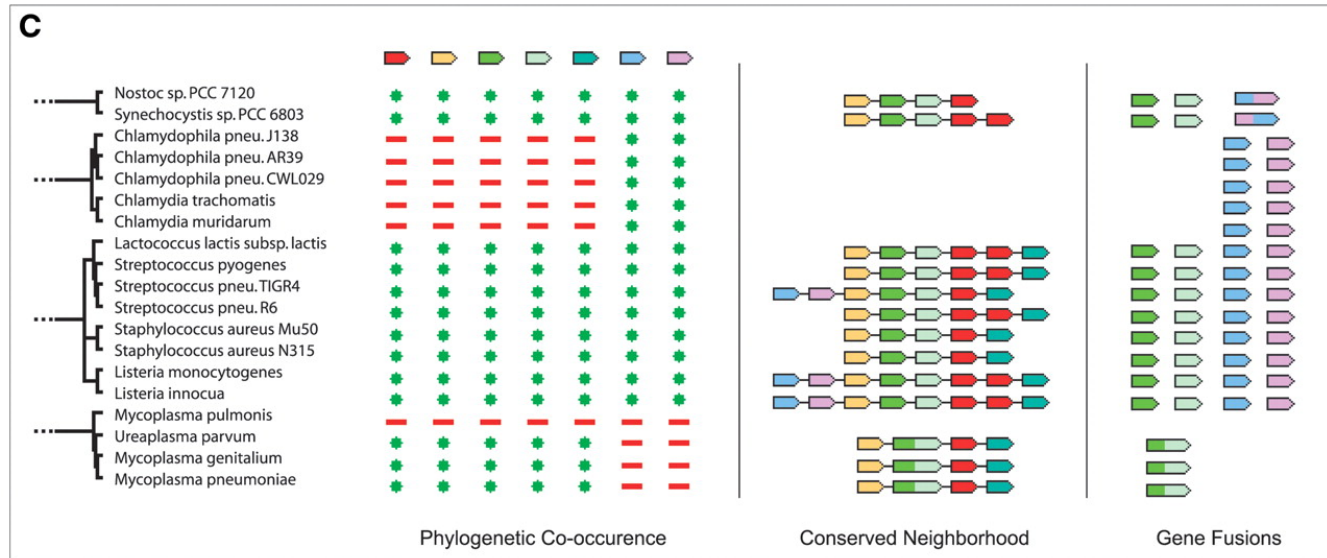
B

Input:

COG117 ABC-type phosphate transport system, ATPase component

Predicted functional associations:

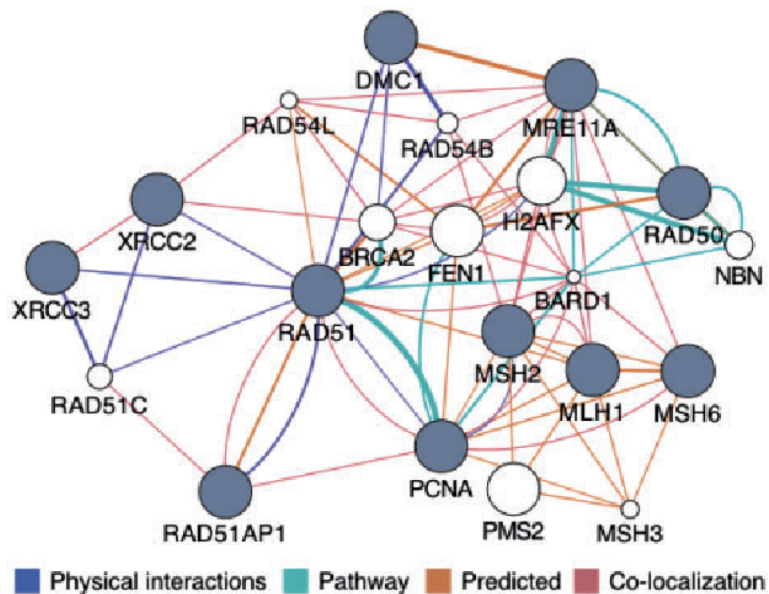
	Score:
COG0226 ABC-type phosphate transport system, periplasmic component	0.997
COG0573 ABC-type phosphate transport system, permease component	0.997
COG0581 ABC-type phosphate transport system, permease component	0.997
COG0704 Phosphate uptake regulator	0.979
COG0745 Response regulator: Che-Y-like receiver domain & DNA-binding domain	0.861
COG0642 Signal transduction histidine kinase	0.799



von Mering *et al.*, *NAR* 2003

GeneMania

- Same principle, other data sources



<http://genemania.org/>

The screenshot shows the GeneMania web interface. At the top, it says "GENEMANIA" and "Find genes in *A. thaliana (arabidopsis)* related to PHYB,ELF3,COP1,SPA1,FUS6,DET1,HYS,OP1,CIP8,P". Below this, there are tabs for "Save", "Actions", "Networks legend", and "Link legend". The main area displays a network of genes with a tooltip for "PRR7" (PSEUDO-RESPONSE REGULATOR 7) showing its description and synonyms. On the right, there is a list of related genes with their scores. At the bottom, there is a "Networks legend" with various interaction types and their corresponding colors.

Gene	Score
PRR7 (PSEUDO-RESPONSE REGULATOR 7; transcription regulator/ two-component response regulator...)	.55
More at SAS or TAB or Index of	
AT5G01780	.55
COI1	.55
AT2G08410	.54
GBF1	.54
AT3G06620	.54
AT4G08900	.54
CK2	.54
AT1G02280	.54
NDPK2	.54
AT1G1880	.54
AT3G06630	.54
PHOT1	.53
FAS5	.53
FCA	.53
NHL8	.53
AT3G06640	.53
AT2G19540	.52
COP13	.52
CPFT5Y	.52
UBC13	.52
UBC7	.52
AT3G06340	.52
VP1	.52

Take-home

- Regulatory networks can be inferred from gene expression data and mined for modules
- Annotation enrichment tests:
 - Fisher's exact test: the basic tool
 - GSEA: needs no subset selection
- Network-based tools can be used to explore interactions

