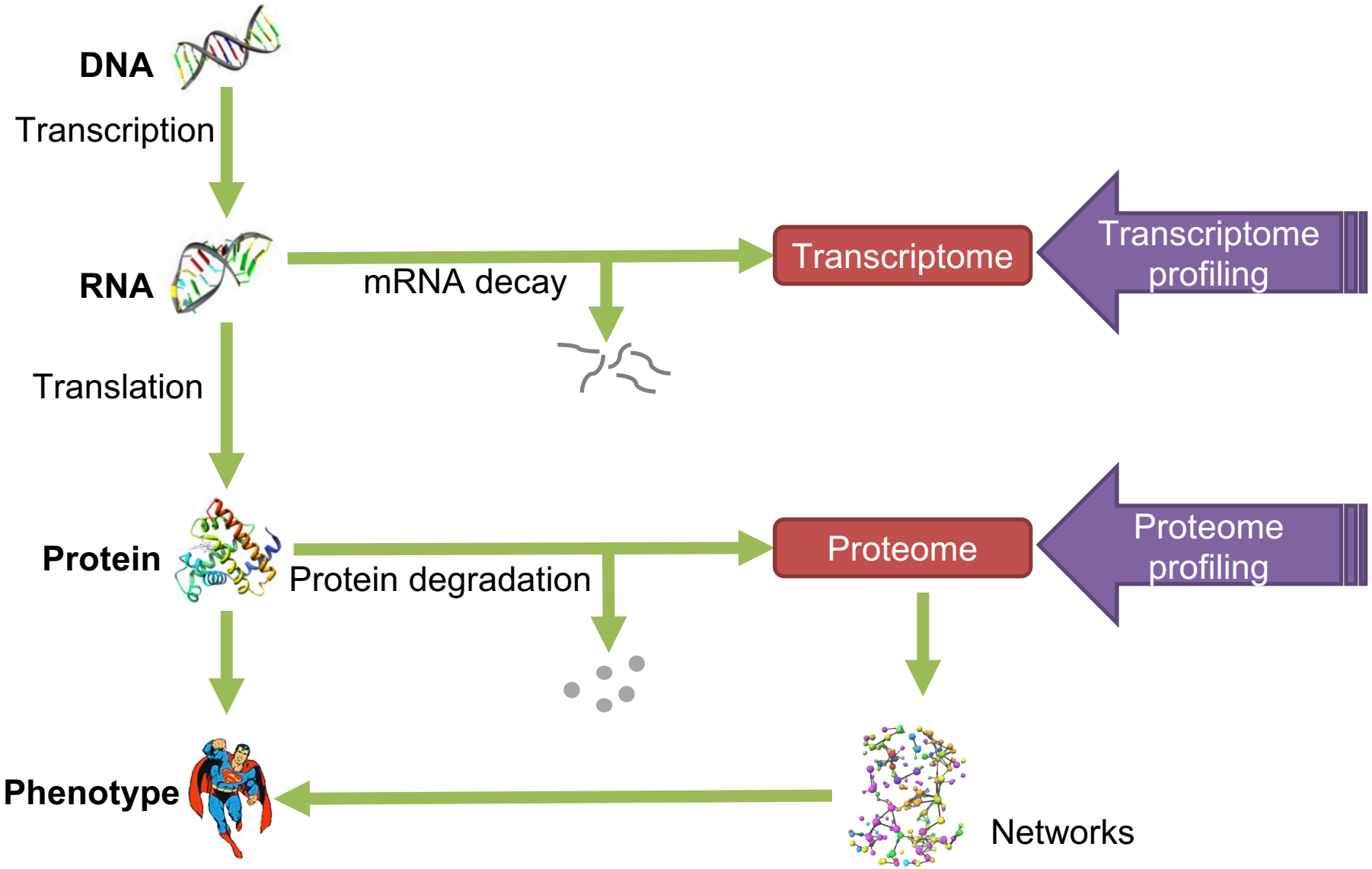
Pathway and network analysis for mRNA and protein profiling data

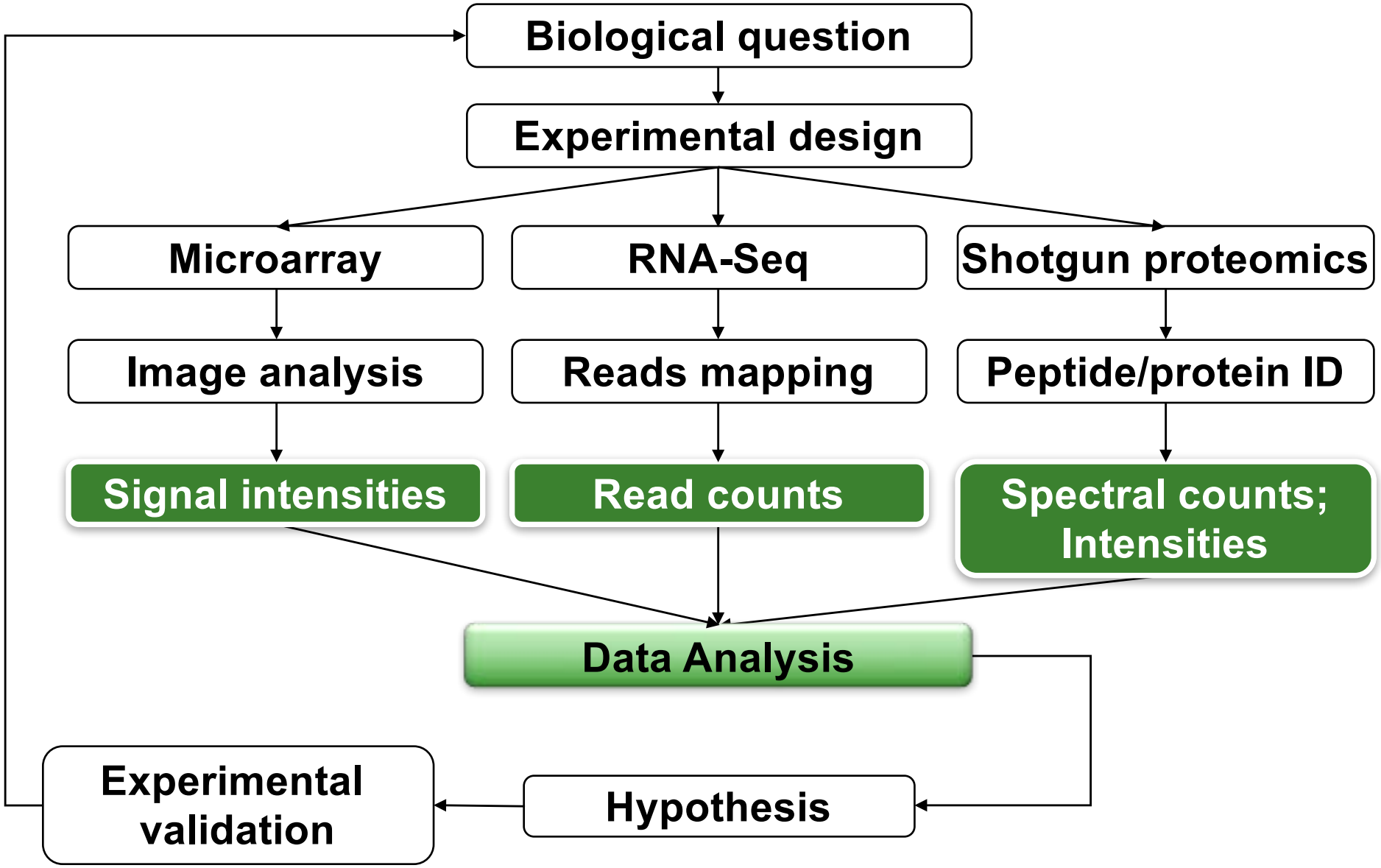
Bing Zhang, Ph.D.
Professor of Molecular and Human Genetics
Lester & Sue Smith Breast Center
Baylor College of Medicine
bing.zhang@bcm.edu

VU workshop, 2016

Gene expression



Overall workflow of gene expression studies



Data matrix

Samples

Genes

probe_set_id	HNE0_1	HNE0_2	HNE0_3	HNE60_1	HNE60_2	HNE60_3
1007_s_at	8.6888	8.5025	8.5471	8.5412	8.5624	8.3073
1053_at	9.1558	9.1835	9.4294	9.2111	9.1204	9.2494
117_at	7.0700	7.0034	6.9047	9.0414	8.6382	9.2663
121_at	9.7174	9.7440	9.6120	9.7581	9.7422	9.7345
1255_g_at	4.2801	4.4669	4.2360	4.3700	4.4573	4.2979
1294_at	6.3556	6.2381	6.2053	6.4290	6.5074	6.2771
1316_at	6.5759	6.5330	6.4709	6.6636	6.6438	6.4688
1320_at	6.5497	6.5388	6.5410	6.6605	6.5987	6.7236
1405_i_at	4.3260	4.4640	4.1438	4.3462	4.3876	4.6849
1431_at	5.2191	5.2070	5.2657	5.2823	5.2522	5.1808
1438_at	7.0155	6.9359	6.9241	7.0248	7.0142	7.0971
1487_at	8.6361	8.4879	8.4498	8.4470	8.5311	8.4225
1494_f_at	7.3296	7.3901	7.0886	7.2648	7.6058	7.2949
1552256_a_at	10.6245	10.5235	10.6522	10.4205	10.2344	10.3144
1552257_a_at	10.3224	10.1749	10.1992	10.2464	10.2191	10.2405

Signal intensities

Read counts

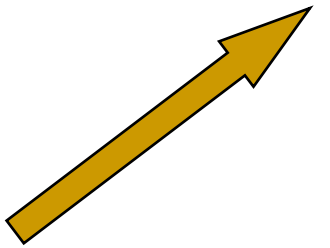
Spectral counts; Intensities

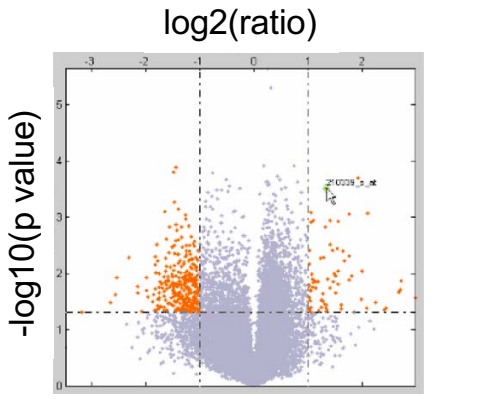
Differential expression and clustering

Microarray

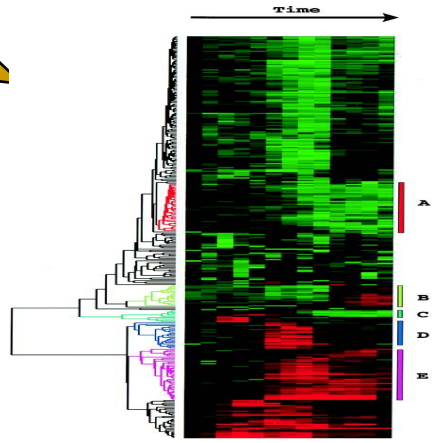
RNA-Seq

Proteomics



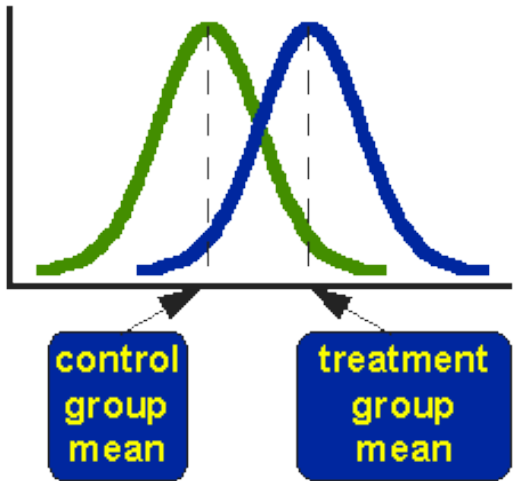


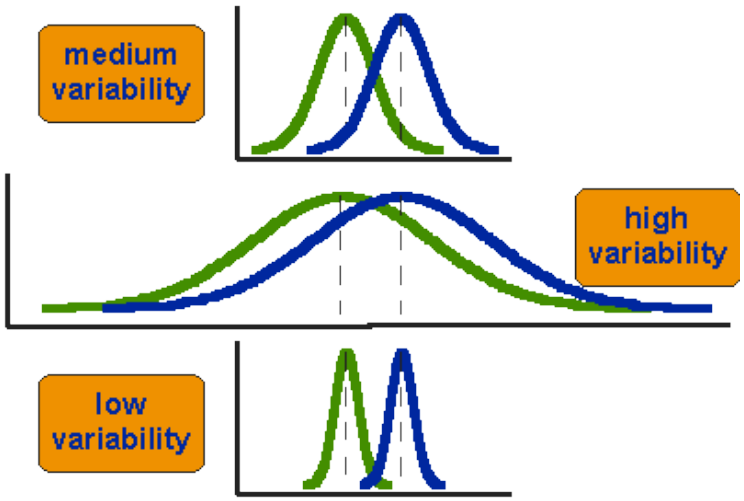
Differential expression



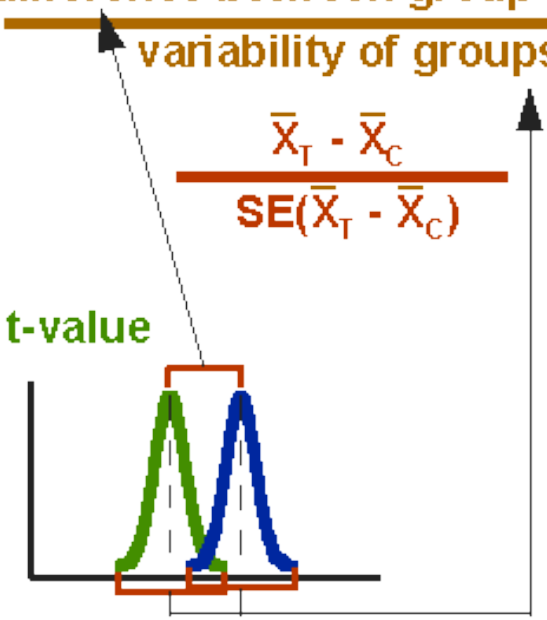
Clustering

Differential expression: *t*-test





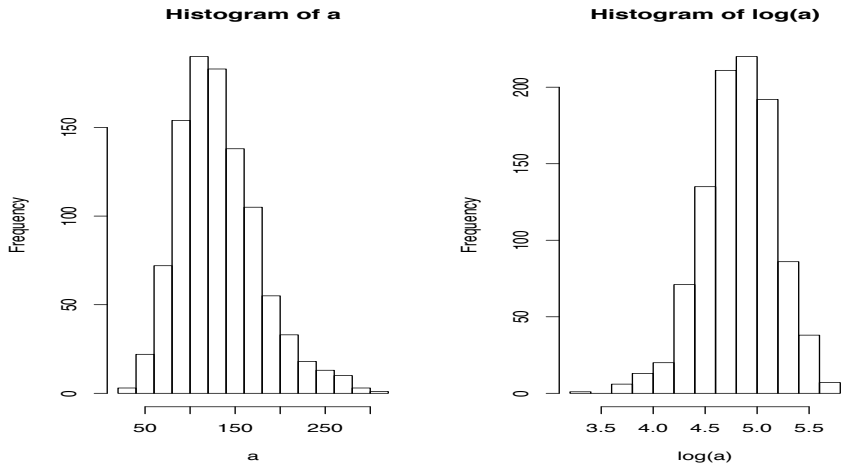
$$\frac{\text{signal}}{\text{noise}} = \frac{\text{difference between group means}}{\text{variability of groups}} = \frac{\bar{X}_T - \bar{X}_C}{SE(\bar{X}_T - \bar{X}_C)} = \text{t-value}$$



graph courtesy of www.socialresearchmethods.net

Data are not normally distributed

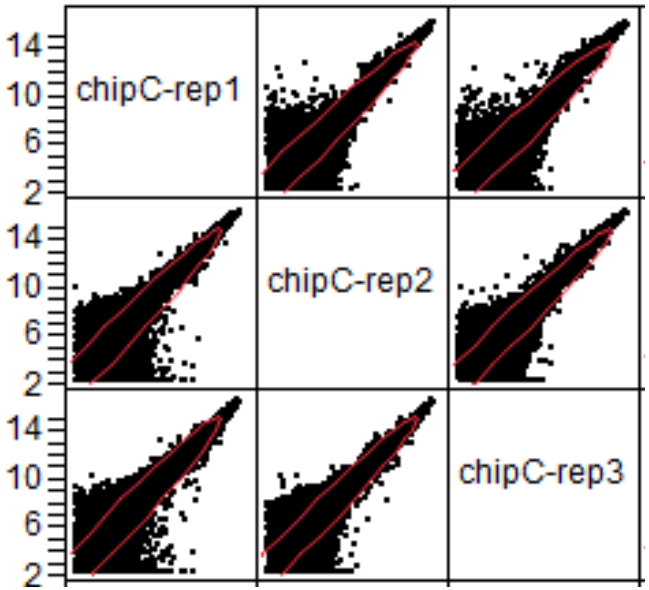
- Log transformation



- Mann-Whitney U test (Wilcoxon rank-sum test)
  - Based on ranks of observations
  - Does not rely on data belonging to any particular distribution

Sample size is small

- Moderated *t*-test (limma)
  - Designed for experiments with small sample size
  - The standard errors are moderated across genes, effectively borrowing information from the ensemble of genes to aid with inference about each individual gene.



Count data

- Negative binomial distribution
  - Cuffdiff
  - DESeq
  - edgeR
- Log-transformed counts
  - Voom/limma

Correction for multiple testing

- Why
  - In an experiment with a 10,000-gene array in which the significance level *p* is set at 0.05, 10,000 x 0.05 = 500 genes would be inferred as significant even though none is differentially expressed
- How
  - **Control the family-wise error rate (FWER)**, the probability that there is a **single** type I error in the entire set (family) of hypotheses tested. *e.g.*, Standard Bonferroni Correction
  - **Control the false discovery rate (FDR)**, the expected proportion of false positives among the number of rejected hypotheses. *e.g.*, Benjamini and Hochberg correction.

Clustering (unsupervised analysis)

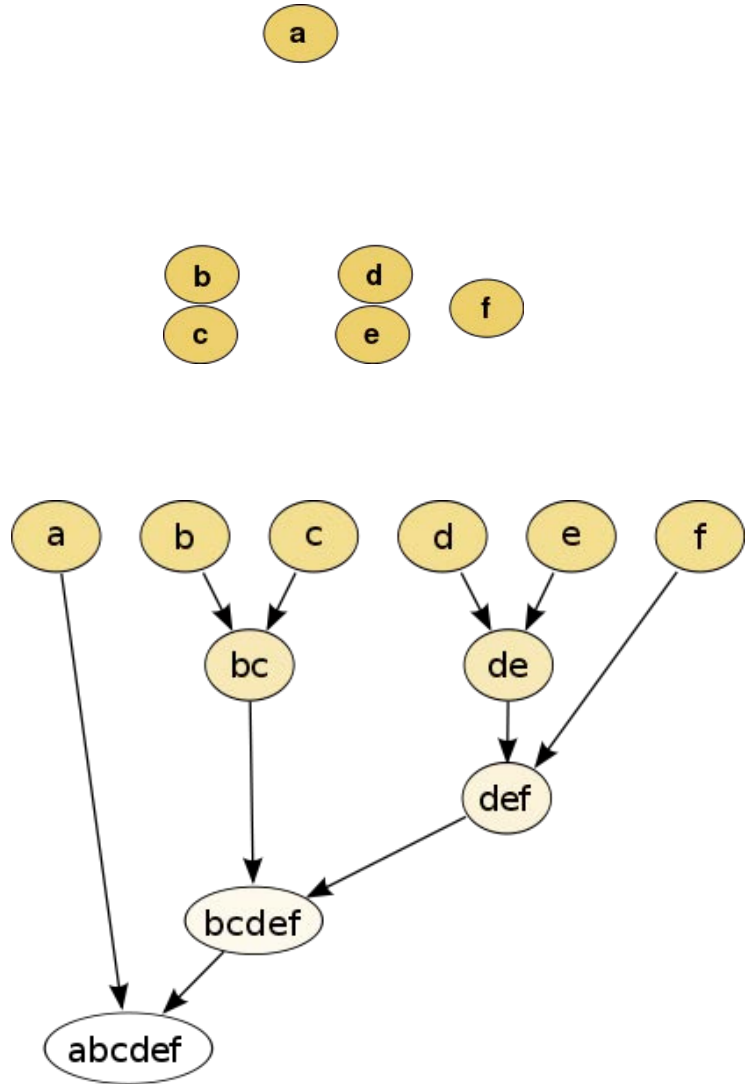
- Clustering algorithms are methods to divide a set of *n* objects (genes or samples) into *g* groups so that within group similarities are larger than between group similarities
- Unsupervised techniques that do not require sample annotation in the process
- Identify candidate subgroups in complex data. e.g. identification of novel sub-types in cancer, identification of co-expressed genes

Samples

Genes	Sample_1	Sample_2	Sample_3	Sample_4	Sample_5	.....
TNNC1	14.82	14.46	14.76	11.22	11.55	.....
DKK4	10.71	10.37	11.23	19.74	19.73	.....
ZNF185	15.20	14.96	15.07	12.57	12.37	.....
CHST3	13.40	13.18	13.15	11.18	10.99	.....
FABP3	15.87	15.80	15.85	13.16	12.99	.....
MGST1	12.76	12.80	12.67	14.92	15.02	.....
DEFA5	10.63	10.47	10.54	15.52	15.52	.....
VIL1	11.47	11.69	11.87	13.94	14.01	.....
AKAP12	18.26	18.10	18.50	15.60	15.69	.....
HS3ST1	10.61	10.67	10.50	12.44	12.23	.....
.....	.....	.....	.....	.....	.....	.....

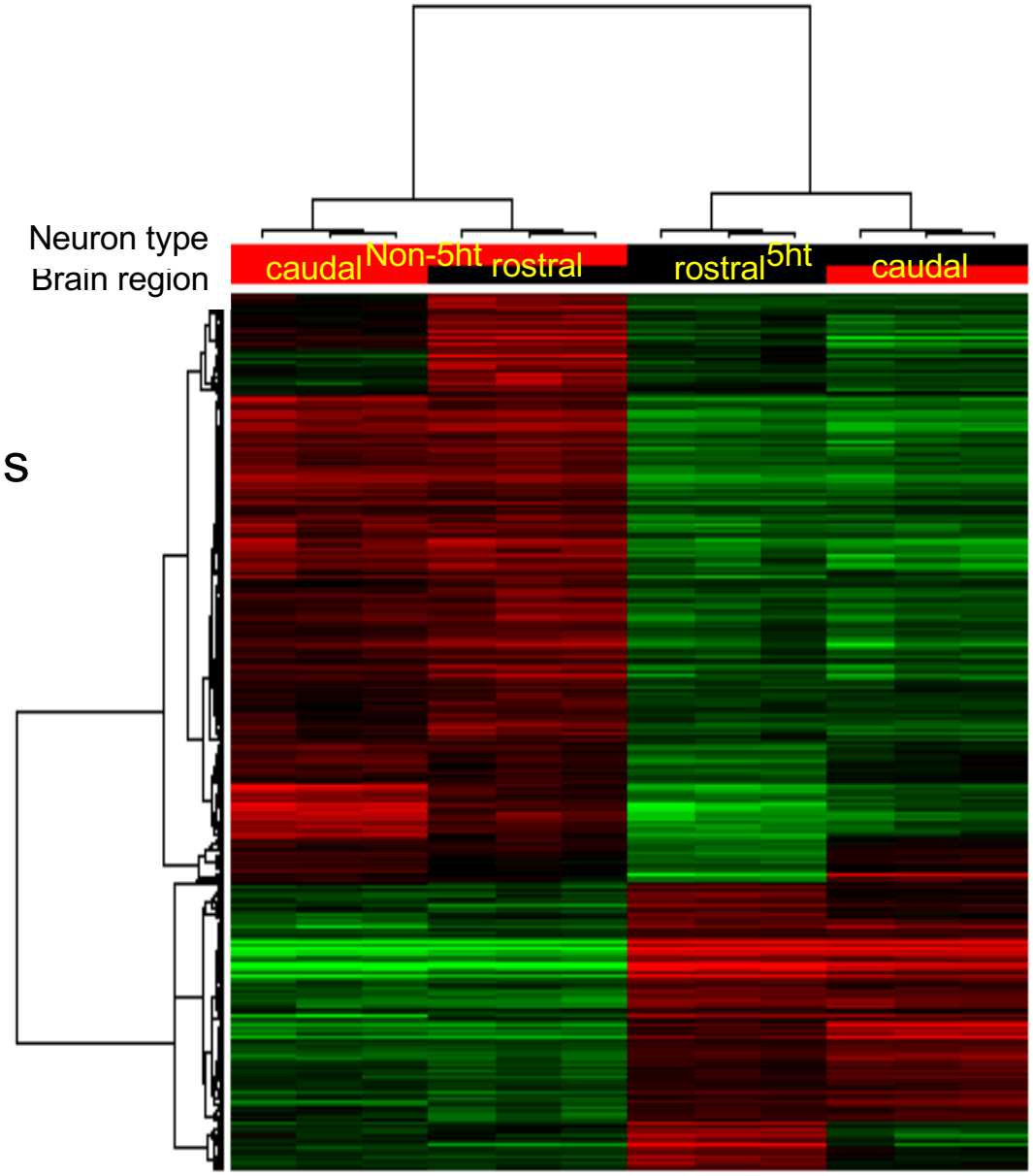
Hierarchical clustering

- Agglomerative hierarchical clustering
  - Start out with all sample units in *n* clusters of size 1.
  - At each step of the algorithm, the pair of clusters with the shortest distance are combined into a single cluster.
  - The algorithm stops when all sample units are combined into a single cluster of size *n*.



Visualization of hierarchical clustering results

- Dendrogram
  - Output of a hierarchical clustering
  - Tree structure with the genes or samples as the leaves
  - The height of the join indicates the distance between the branches
- Heat map
  - Graphical representation of data where the values are represented as colors.

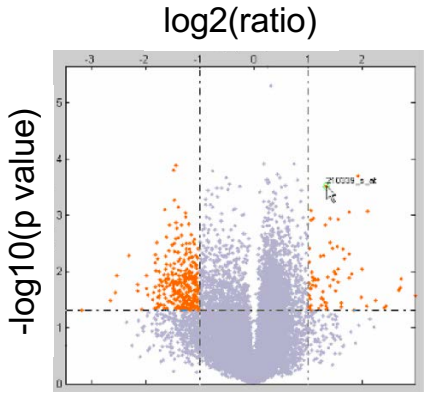


Omics studies generate lists of interesting genes

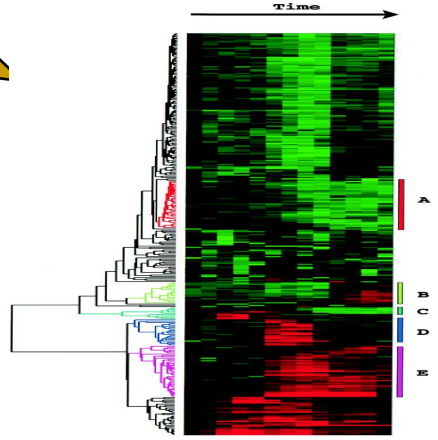
Microarray

RNA-Seq

Proteomics



Differential expression



Clustering

92546_r_at
92545_f_at
96055_at
102105_f_at
102700_at
161361_s_at
92202_g_at
103548_at
100947_at
101869_s_at
102727_at
160708_at
…...

Lists of genes with potential biological interest

Organizing genes based on pathways

- Gene 1
  - Pathway 1
  - Pathway 2
- Gene 2
  - Pathway 1
  - Pathway 3
- Gene 3
  - Pathway 2
  - Pathway 3
- Gene 4
  - Pathway 3

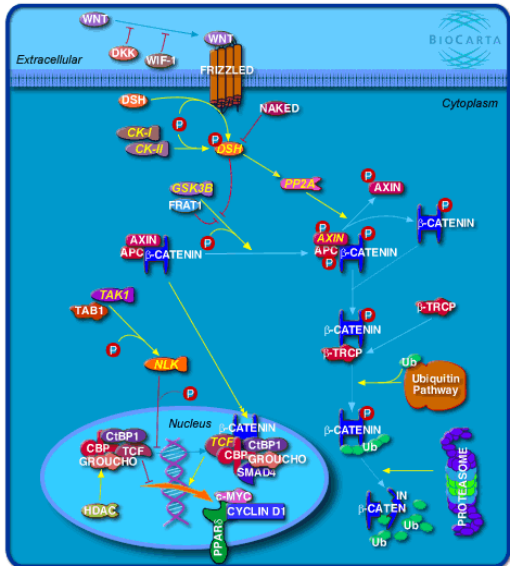
- Pathway 1
  - Gene 1
  - Gene 2
- Pathway 2
  - Gene 1
  - Gene 3
- Pathway 3
  - Gene 2
  - Gene 3
  - Gene 4

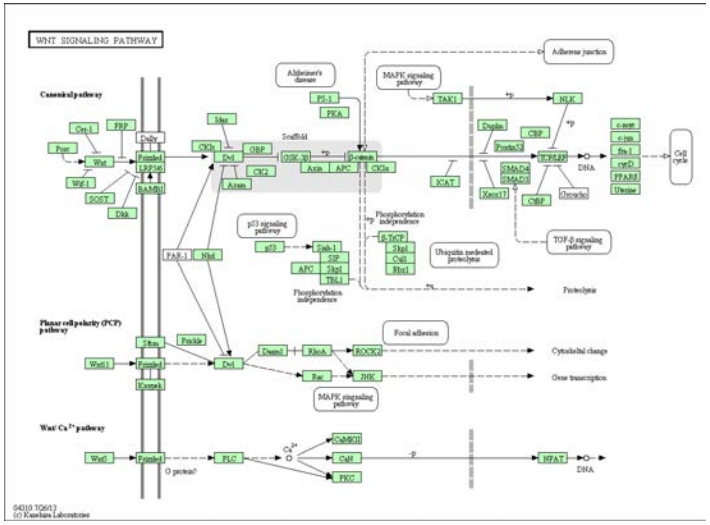
Advantages of pathway analysis

- Better interpretation
  - From interesting genes to interesting biological themes
- Improved robustness
  - Robust against noise in the data
- Improved sensitivity
  - Detecting minor but concordant changes in a pathway

Pathway databases

- Databases
  - BioCarta (http://www.biocarta.com/genes/index.asp)
  - KEGG (http://www.genome.jp/kegg/pathway.html)
  - MetaCyc (http://metacyc.org)
  - Pathway commons (http://www.pathwaycommons.org)
  - Reactome (http://www.reactome.org)
  - STKE (http://stke.sciencemag.org/cm)
  - Signaling Gateway (http://www.signaling-gateway.org)
  - Wikipathways (http://www.wikipathways.org)
- Limitation
  - Limited coverage
  - Inconsistency among different databases
  - Relationship between pathways is not defined

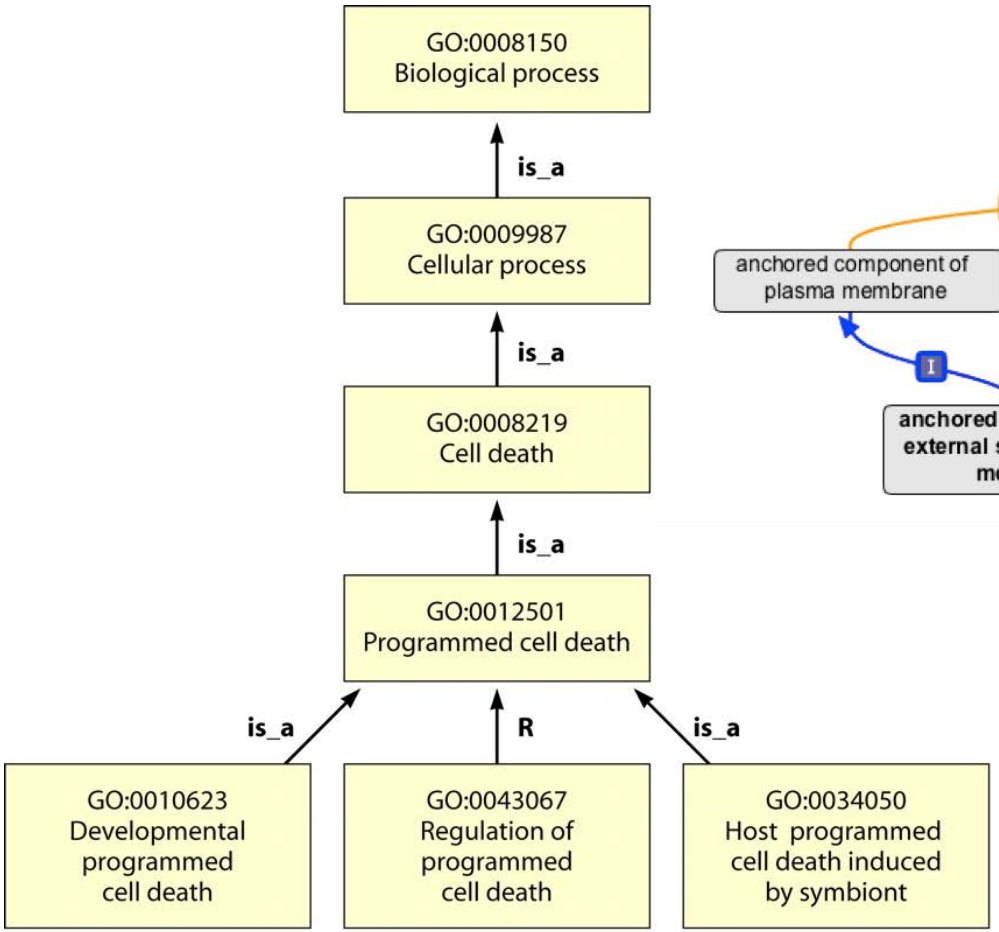


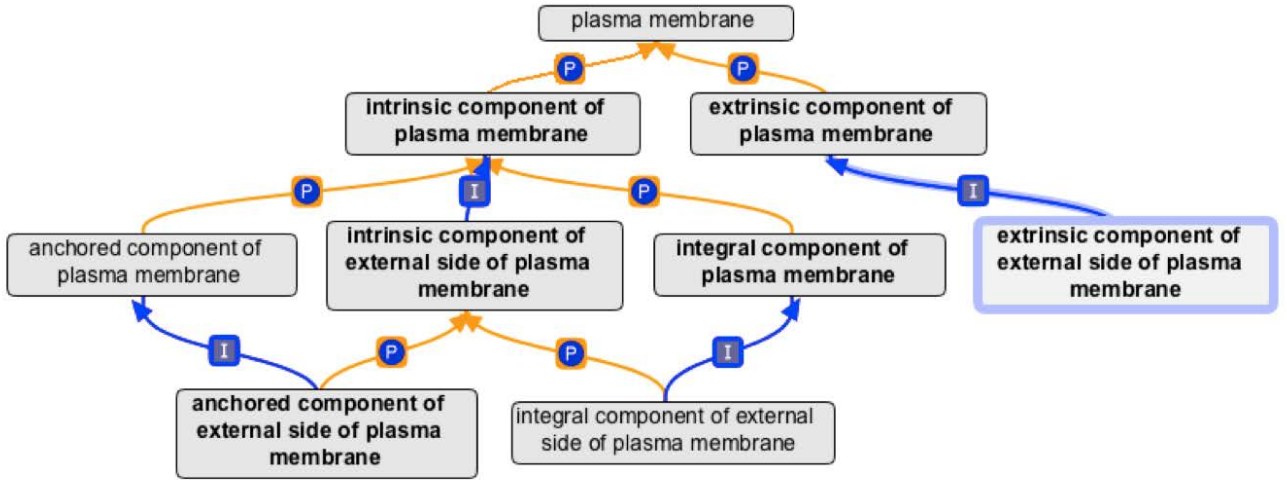


Gene Ontology

- Structured, precisely defined, controlled vocabulary for describing the roles of genes and gene products
- Three organizing principles: molecular function, biological process, and cellular component
  - Dopamine receptor D2, the product of human gene DRD2
  - molecular function: dopamine receptor activity
  - biological process: synaptic transmission
  - cellular component: plasma membrane
- Terms in GO are linked by several types of relationships
  - Is_a (e.g. plasma membrane is_a membrane)
  - Part_of (e.g. membrane is part_of cell)
  - Has part
  - Regulates
  - Occurs in

Gene Ontology



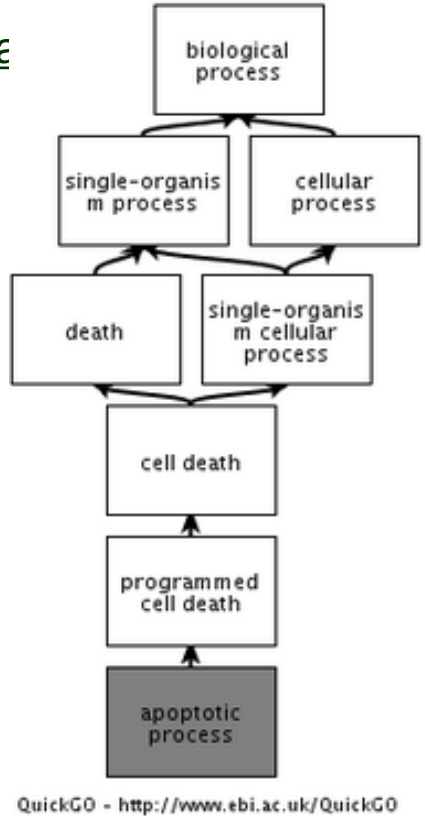


Annotating genes using GO terms

- Two types of GO annotations
  - Electronic annotation
  - Manual annotation
- All annotations must:
  - be attributed to a source
  - indicate what evidence was found to support the GO term-gene/protein association
- Types of evidence codes
  - Experimental codes - IDA, IMP, IGI, IPI, IEP
  - Computational codes - ISS, IEA, RCA, IGC
  - Author statement - TAS, NAS
  - Other codes - IC, ND

Access GO

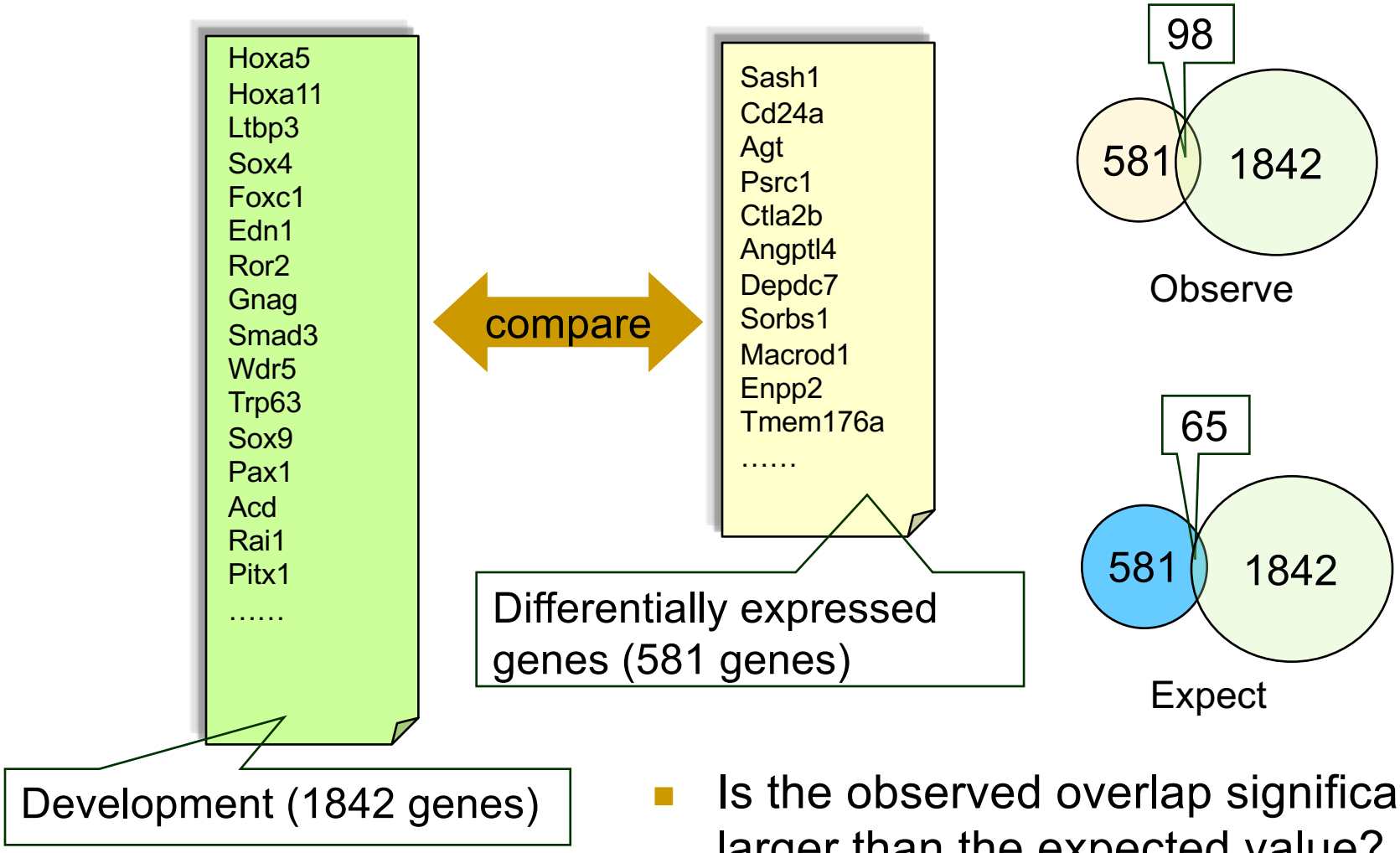
- Downloads (http://www.geneontology.org)
  - Ontologies
    - http://www.geneontology.org/page/download-ontology
  - Annotations
    - http://www.geneontology.org/page/download-a
- Web-based access
  - AmiGO: http://www.godatabase.org
  - QuickGO: http://www.ebi.ac.uk/QuickGO



Coverage of GO annotations

	Homo sapiens		*Mus musculus*	
	#term	#gene	#term	#gene
GO/BP	6502	15228	6227	15709
GO/MF	3144	16389	2961	17287
GO/CC	947	16765	882	16801

Over-representation analysis: concept



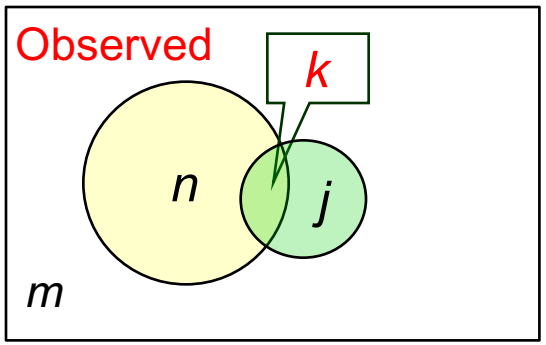
- Is the observed overlap significantly larger than the expected value?

Over-representation analysis: method

	Significant genes	Non-significant genes	Total
genes in the group	*k*	*j-k*	*j*
Other genes	*n-k*	*m-n-j+k*	*m-j*
Total	*n*	*m-n*	*m*

Hypergeometric test: given a total of *m* genes where *j* genes are in the functional group, if we pick *n* genes randomly, what is the probability of having *k* or more genes from the group?

$$p = \sum_{i=k}^{\min(n,j)} \frac{\binom{m-j}{n-i}\binom{j}{i}}{\binom{m}{n}}$$



Zhang et.al. Nucleic Acids Res. 33:W741, 2005

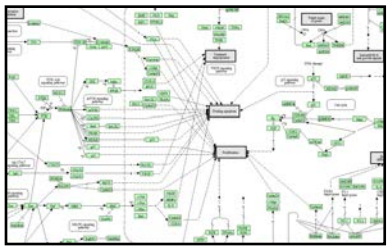
Over-representation analysis: WebGestalt

Gene list

92546_r_at
92545_f_at
96055_at
102105_f_at
102700_at
……

~200 ID types

~60K Functional categories



Pathways/ functional categories

8 organisms
Human, Mouse, Rat, Dog, Fruitfly, Worm, Zebrafish, Yeast

Microarray Probe IDs
- Affymetrix
- Agilent
- Codelink
- Illumina

Genetic Variation IDs
- dbSNP

Gene IDs
- Gene Symbol
- GenBank
- Ensembl Gene
- RefSeq Gene
- UniGene
- Entrez Gene
- SGD
- MGI
- Flybase ID
- Wormbase ID
- ZFIN

Protein IDs
- UniProt
- IPI
- RefSeq Peptide
- Ensembl Peptide

196 ID types with mapping to Entrez Gene ID

WebGestalt

Analysis/ visualization

59,278 functional categories with genes identified by Entrez Gene IDs

Gene Ontology
- Biological Process
- Molecular Function
- Cellular Component

Pathway
- KEGG
- Pathway Commons
- WikiPathways

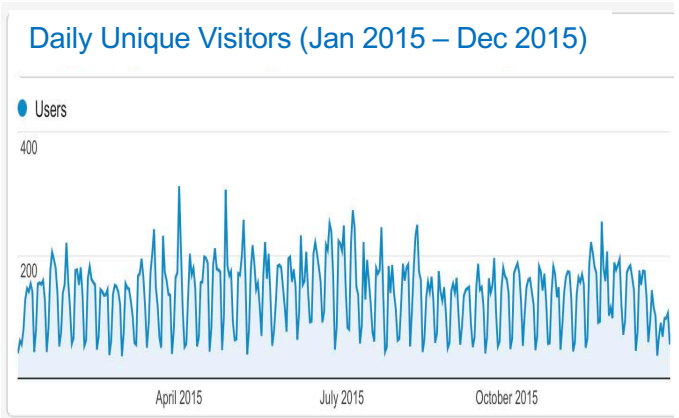
Network module
- Transcription factor targets
- microRNA targets
- Protein interaction modules

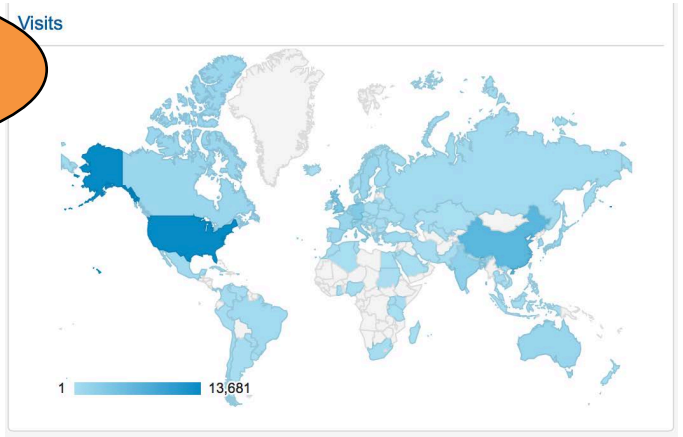
Disease and Drug
- Disease association genes
- Drug association genes

Chromosomal location
- Cytogenetic bands

Zhang et.al. Nucleic Acids Res. 33:W741, 2005
Wang et al. Nucleic Acids Res. 41:W77, 2013

http://www.webgestalt.org



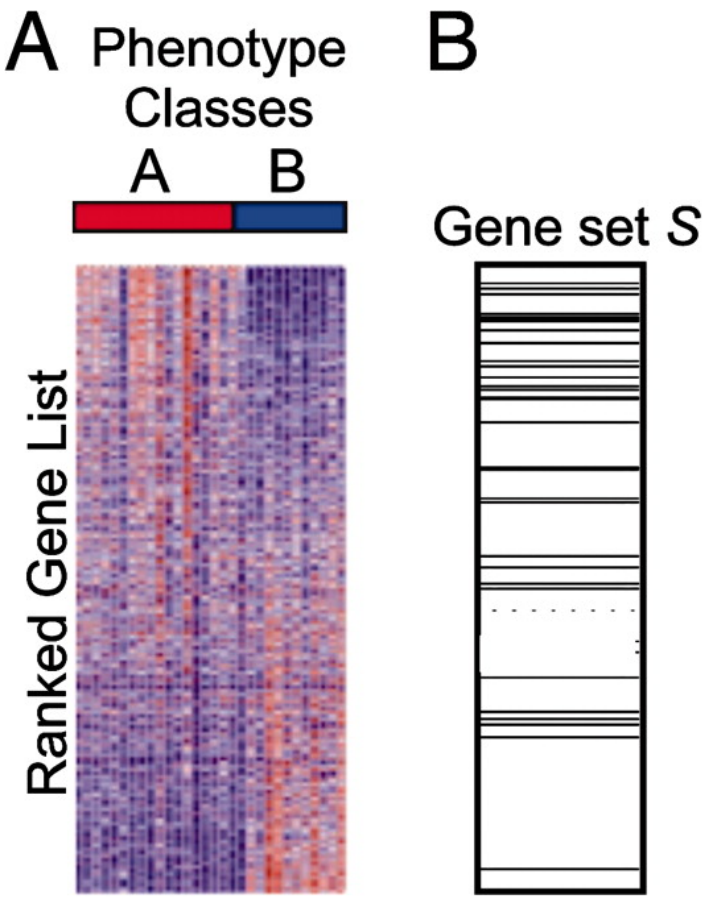


Jan. 1, 2015 – Dec. 31, 2015
63,932 visits from 27,409 visitors
>300 citations

Over-representation analysis: limitations

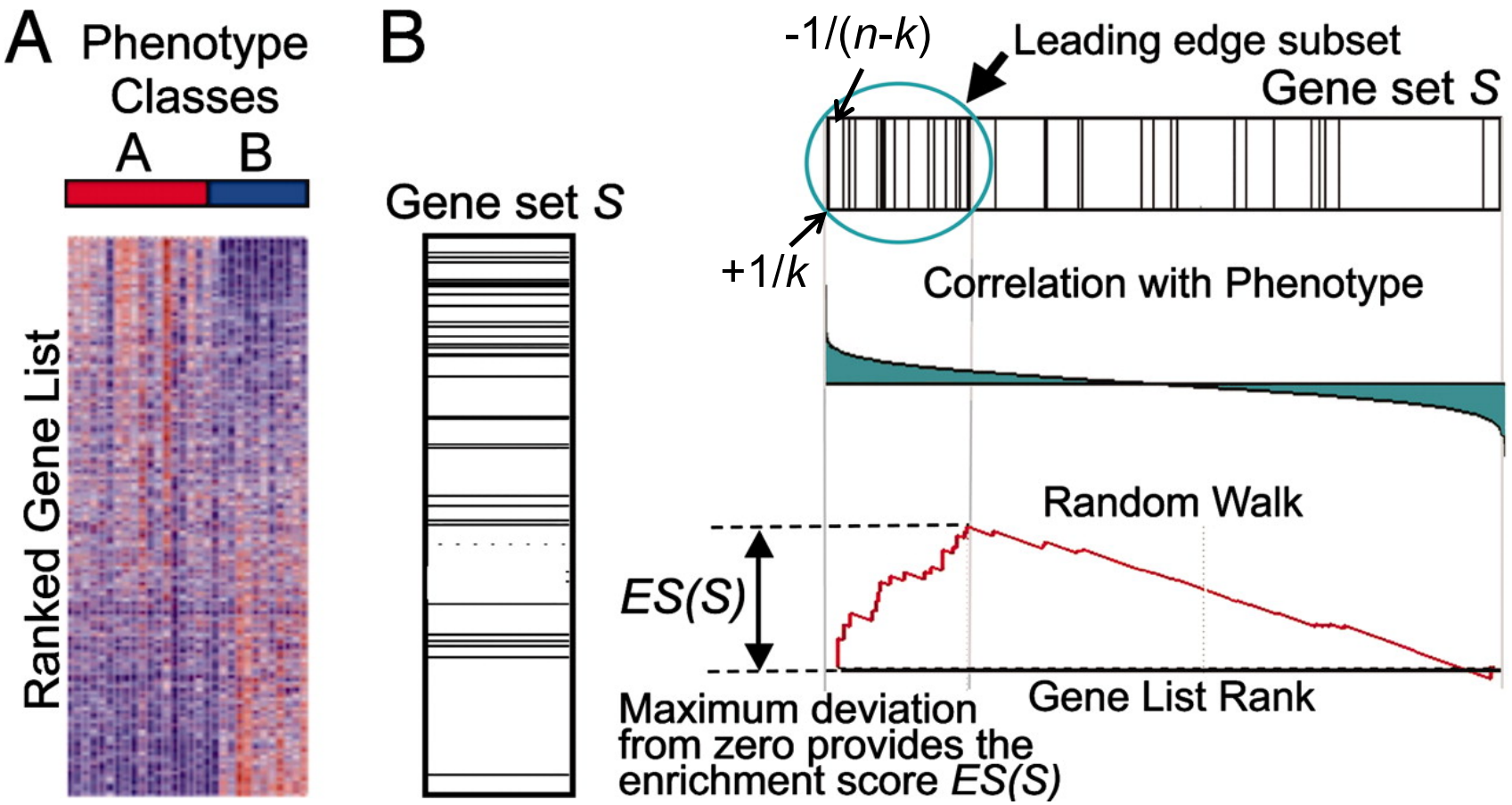
- Arbitrary thresholding
- Ignoring the order of genes in the significant gene list

Gene Set Enrichment Analysis: concept



- Do genes in a gene set tend to locate at the top or bottom of the ranked gene list?

Gene Set Enrichment Analysis: method



Subramanian et.al. PNAS 102:15545, 2005

http://www.broad.mit.edu/gsea/

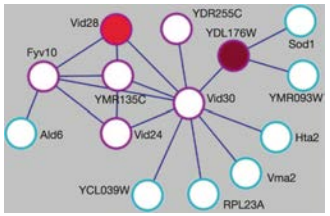
k: Number of genes in the gene set S
n: Number of all genes in the ranked gene list

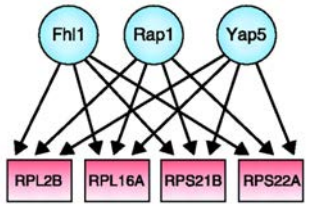
Pathway-based analysis

- Organizing genes by
  - Pathways
  - Gene Ontology
  - Other functional categories
- Enrichment analysis methods
  - Over-representation analysis
  - Gene Set enrichment analysis
- Major limitation
  - Existing knowledge on pathways or gene functions is far from complete
  - Connections between genes are not considered

Biological networks

Networks		Nodes	Edges
Physical interaction networks	Protein-protein interaction network	Proteins	Physical interaction, undirected
	Signaling network	Proteins	Modification, directed
	Gene regulatory network	TFs/miRNAs Target genes	Physical interaction, directed
	Metabolic network	Metabolites	Metabolic reaction, directed
Functional association networks	Co-expression network	Genes/proteins	Co-expression, undirected
	Genetic network	Genes	Genetic interaction, undirected

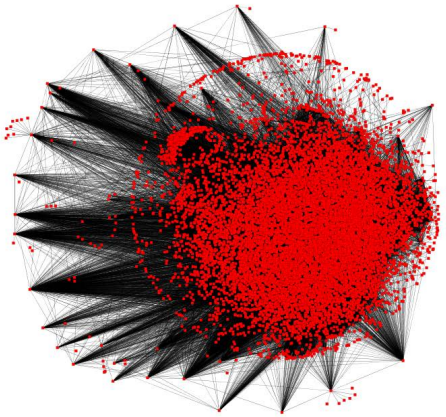




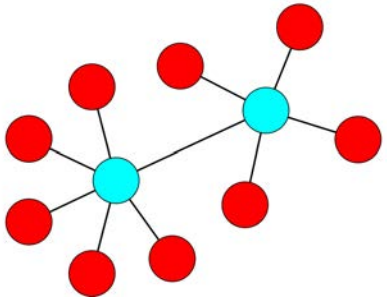
Protein-protein interaction databases

- Database of Interacting Proteins (DIP)
  http://dip.doe-mbi.ucla.edu/
- The Molecular INTeraction database (MINT)
  http://mint.bio.uniroma2.it/mint/
- The Biomolecular Object Network Databank (BOND)
  http://bond.unleashedinformatics.com/
- The General Repository for Interaction Datasets (BioGRID)
  http://www.thebiogrid.org/
- Human Protein Reference Database (HPRD)
  http://www.hprd.org
- Online Predicted Human Interaction Database (OPHID)
  http://ophid.utoronto.ca
- iRef
  http://wodaklab.org/iRefWeb
- The International Molecular Exchange Consortium (IMEX)
  http://www.imexconsortium.org

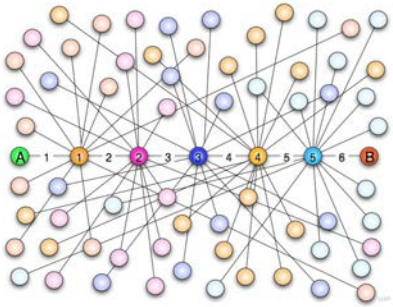
Properties of complex networks



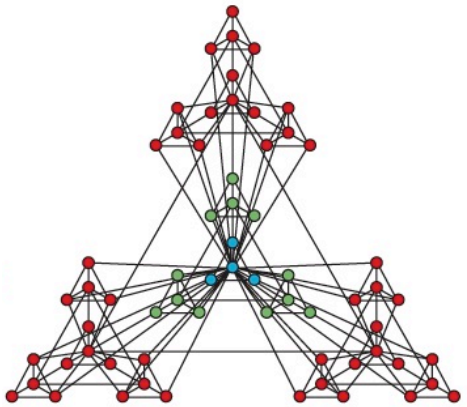
Human protein-protein interaction network
9,198 proteins and 36,707 interactions



Scale-free
(hubs)

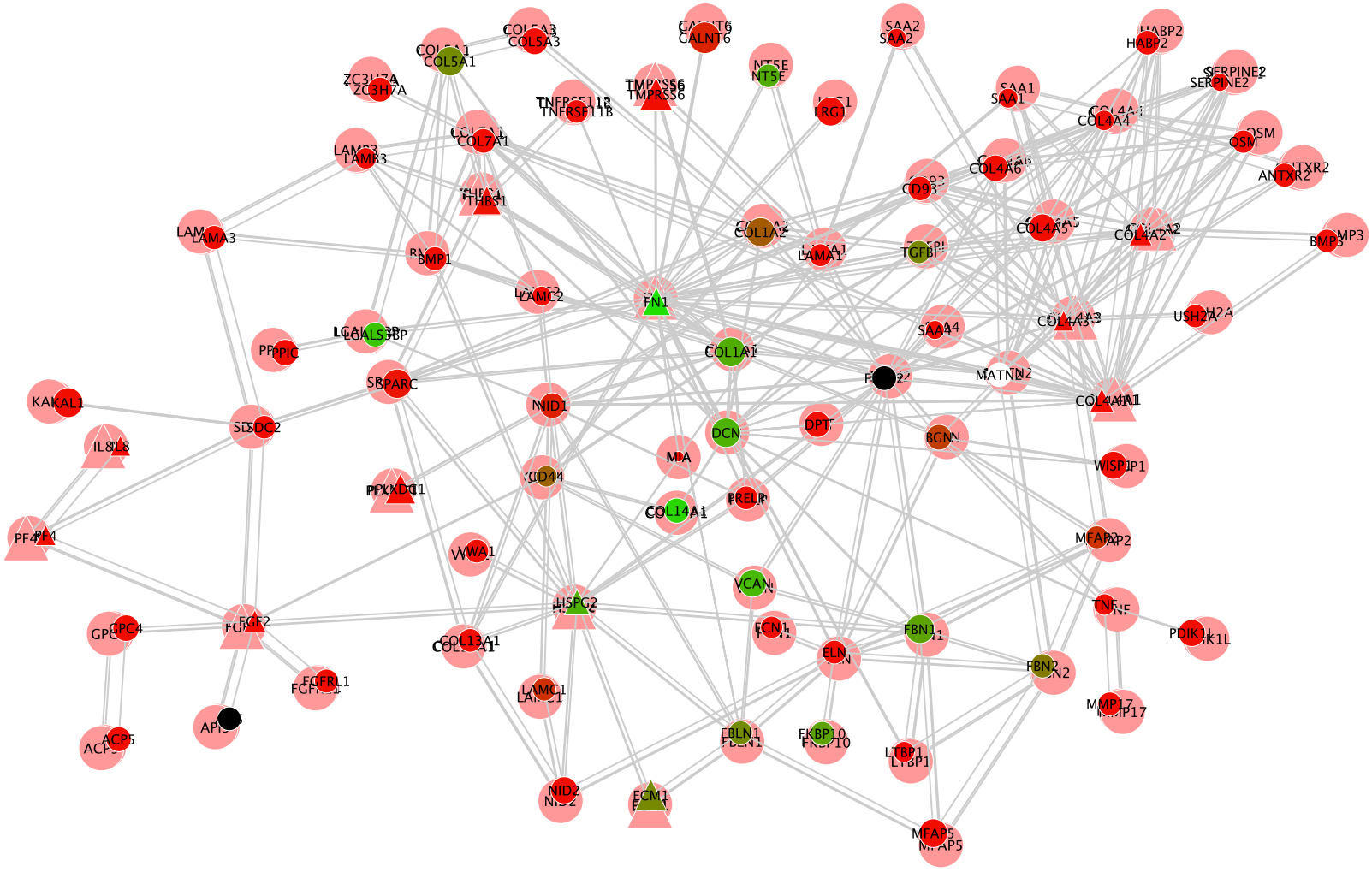


Small world
(six degree separation)



Hierarchical modular

Network visualization



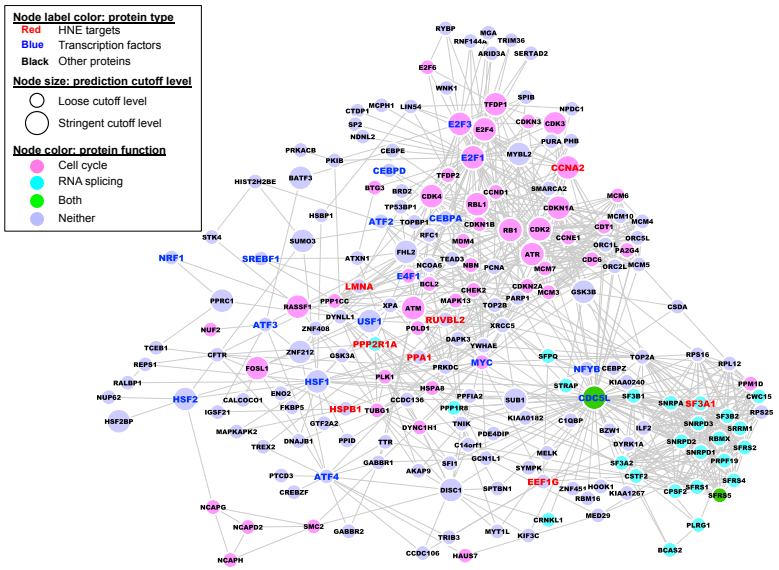
Network visualization

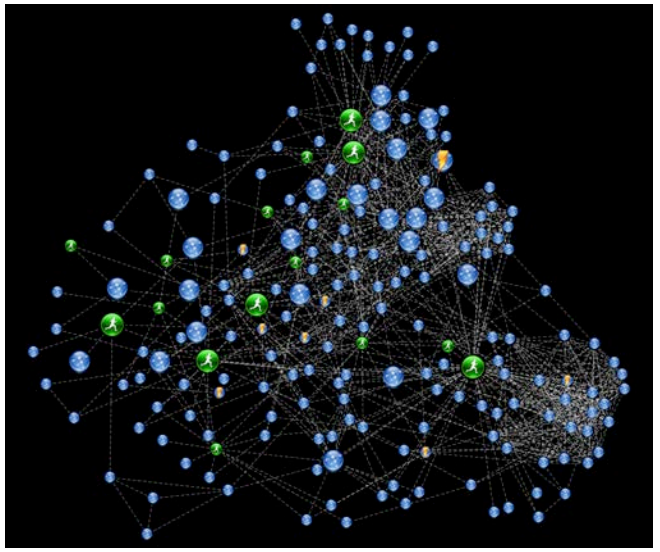
Network visualization tools

Name	Cost	OS	Description	URL
Stand-alone				
Arena 3D[63]	Free	Win, Mac, Linux	Visualization of biological multi-layer networks in 3D	http://www.arena3d.org/
BiNA[81]	Free	Win, Mac, Linux	Exploration and interactive visualization of pathways	http://www.bnplusplus.org/bina/
BioLayout Express 3D[37]	Free	Win, Mac, Linux	Generation and cluster analysis of networks with 2D/3D visualization	http://www.biolayout.org/
BiologicalNetworks[82]	Free	Win, Mac, Linux	Analysis suite; visualizes networks and heat map; abundance data	http://www.biologicalnetworks.org/
Cytoscape[20, 83]	Free	Win, Mac, Linux	Network analysis; extensive list of plug-ins for advanced visualization	http://www.cytoscape.org/
GENeVis[36]	Free	Win, Mac, Linux	Network and pathway visualization; abundance data	http://tinyurl.com/genevis/
Medusa[84]	Free	Win, Mac, Linux	Basic network visualization tool	http://coot.embl.de/medusa/
N-Browse[85]	Free	Win, Mac, Linux	Network visualization software for heterogeneous interaction data	http://www.gnetbrowse.org/
NAViGaTOR[23, 86]	Free	Win, Mac, Linux	Visualization of large protein-interaction data sets; abundance data	http://tinyurl.com/navigator1/
Ondex[87]	Free	Win, Mac, Linux	Integrative workbench: large network visualizations; abundance data	http://www.ondex.org/
Osprey[88]	Free	Win, Mac, Linux	Tool for visualization of interaction networks	http://tinyurl.com/osprey1/
Pajek[89]	Free	Win	Generic network visualization and analysis tool	http://pajek.imfm.si/
ProViz	Free	Win, Mac, Linux	Software for visualization and exploration of interaction networks	http://tinyurl.com/proviz/
SpectralNET[90]	Free	Win	Network visualizations; scatter plots for dimensionality reduction methods	http://tinyurl.com/spectralnet/
Tulip[91]	Free	Win, Mac, Linux	Generic visualization tool; extremely large networks; 3D support	http://tulip.labri.fr/TulipDrupal/
VANTED[21]	Free	Win, Mac, Linux	Combined visualization of abundance data, networks and pathways	http://tinyurl.com/vanted/
yEd	Free	Win, Mac, Linux	Generic network visualization software; offers many layout algorithms	http://tinyurl.com/yEdGraph/
Cytoscape plug-in				
BiNoM[92]	Free	Win, Mac, Linux	Extensive support for common systems biology network formats	http://tinyurl.com/binom1/
BioModules[24]	Free	Win, Mac, Linux	Detects modules in networks; maps abundance data onto nodes and modules	http://tinyurl.com/biomodules/
Cerebral[26, 78]	Free	Win, Mac, Linux	Biologically motivated layout algorithm; maps abundance data; clustering	http://tinyurl.com/cerebral1/
MCODE[18]	Free	Win, Mac, Linux	Network clustering algorithm; support for manual cluster refinement	http://tinyurl.com/MCODE123/
VistaClara[42]	Free	Win, Mac, Linux	Mapping of abundance data to nodes and 'heat strips'; provides heat map	http://tinyurl.com/cytoplugins/
Web-based				
Graphle[93]	Free		Distributed client/server network exploration and visualization tool	http://tinyurl.com/graphle/
Lichen	Free		Library for web-based visualization of network and abundance matrix data	http://tinyurl.com/Lichen1/
MAGGIE Data Viewer	Free		Visualization of networks; abundance data in heat maps and profile plots	http://maggie.systemsbiology.net/
STITCH[31]	Free		Construction and visualization of networks from a wide range of sources	http://stitch.embl.de/
VisANT[22]	Free	Win, Mac, Linux	Analysis, mining and visualization of pathways and integrated omics data	http://visant.bu.edu/

Gehlenborg et al. Nature Methods, 7:S56, 2010

Cytoscape (http://www.cytoscape.org)

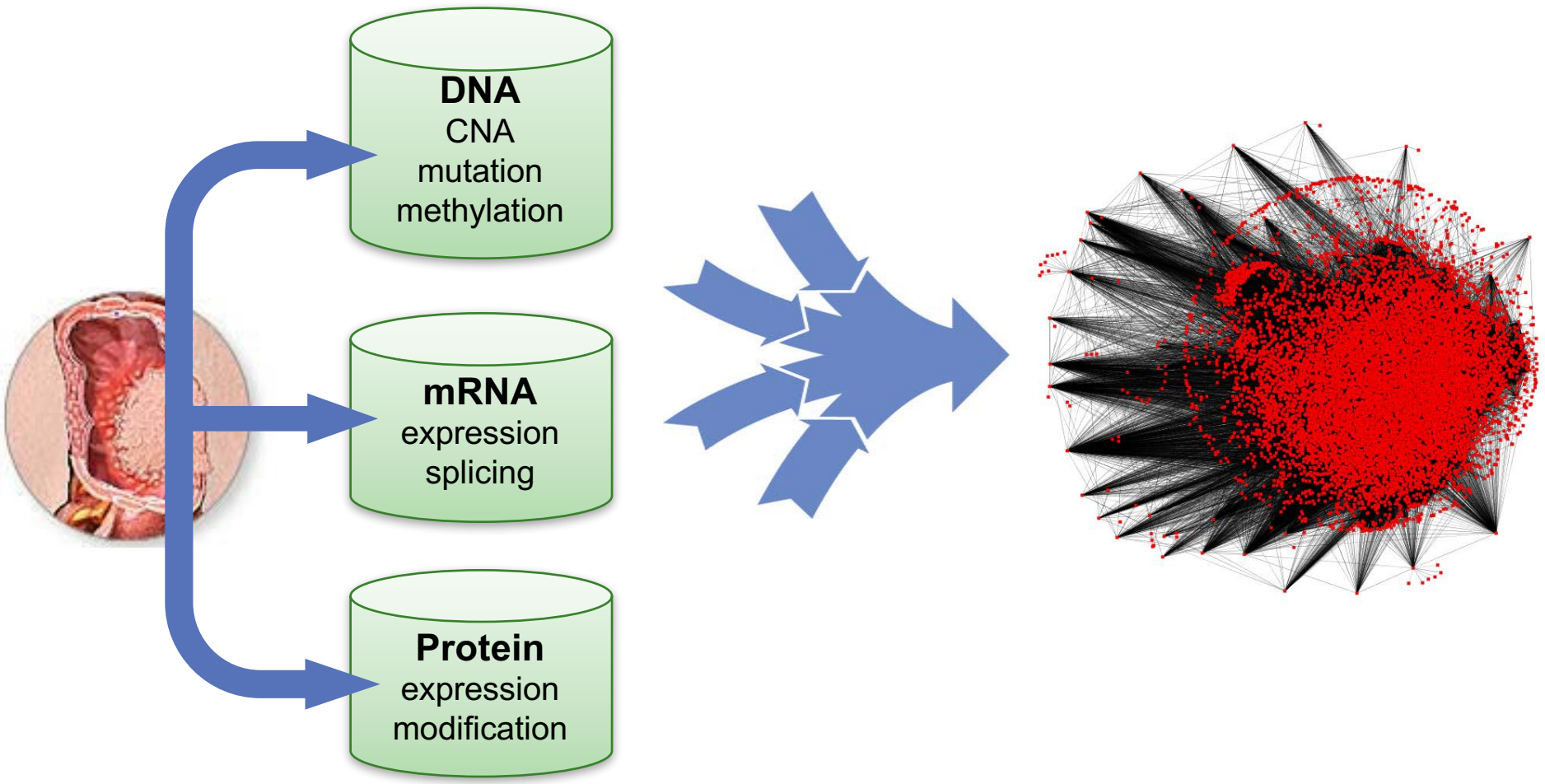




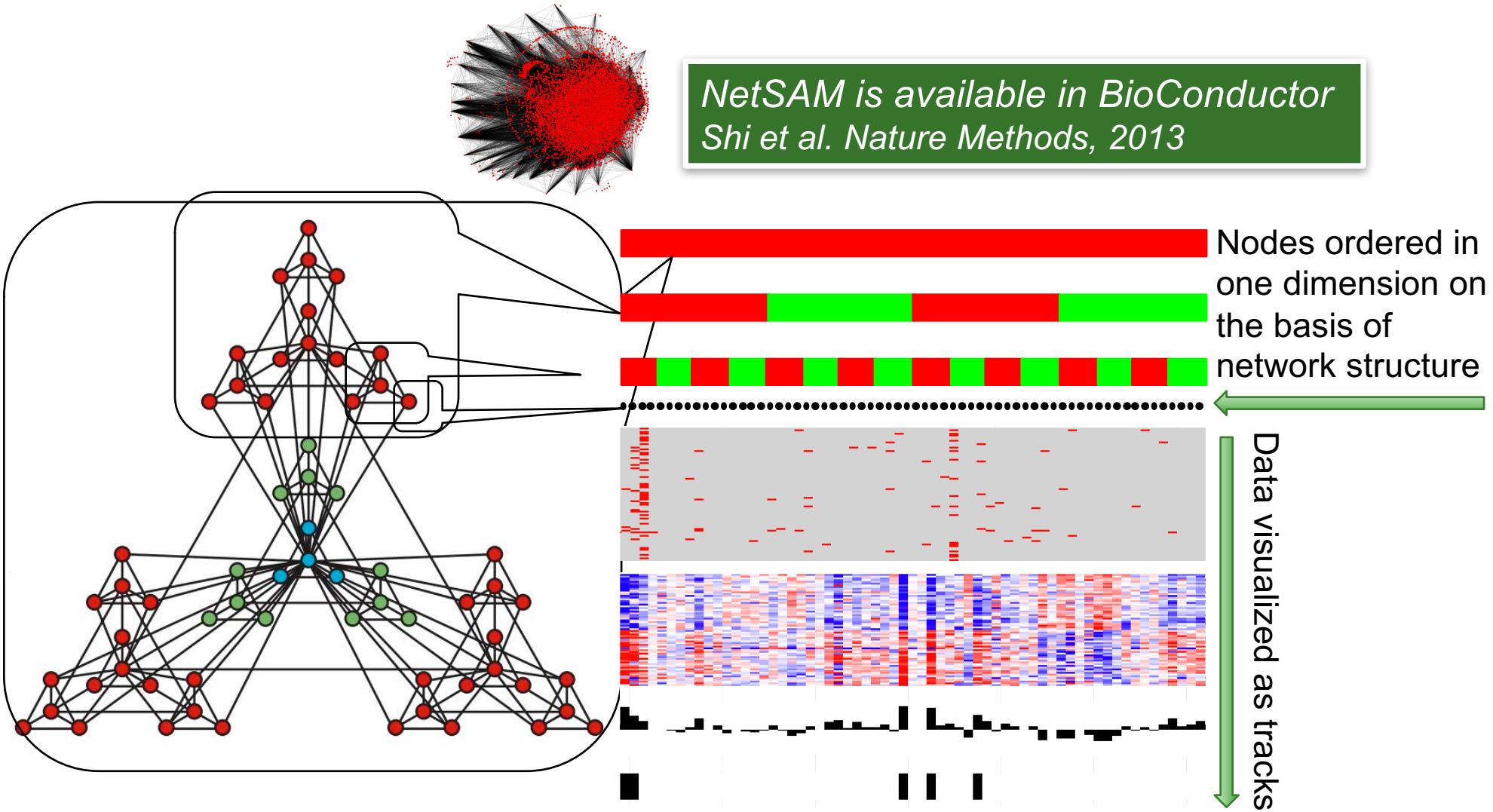
Scalability challenges

Data complexity

Network size



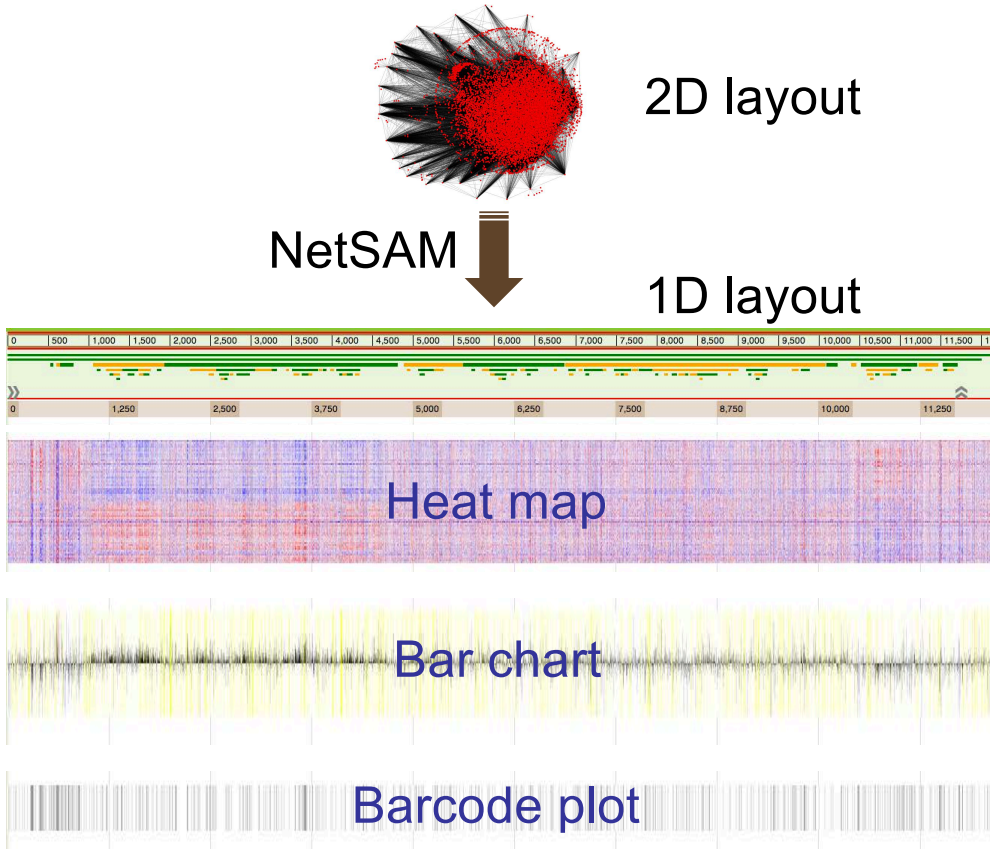
Making sense of the hairballs



Barabasi and Oltvai, Nat Rev Genet, 2004

NetGestalt: exploring multidimensional omics data over biological networks

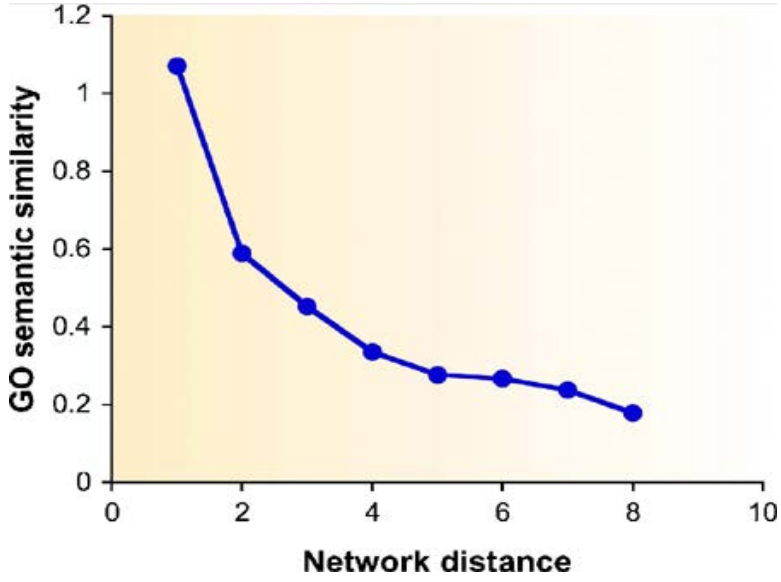
- Genes ordered in the horizontal dimension
- Data visualized as tracks along the vertical dimension
  - Composite continuous track (e.g. gene expression matrices)
  - Single continuous track (e.g. fold changes, *p* values)
  - Single binary track (e.g. significant genes, GO annotations)
- Interactive data visualization and analysis
  - Search, zoom, pan, filter
  - Track comparison
  - Statistical analysis
  - Network analysis and visualization
  - Pathway enrichment analysis
  - Upload user networks and tracks



http://www.netgestalt.org
Shi et al. Nature Methods, 2013

Network distance *vs* functional similarity

- Proteins that lie closer to one another in a protein interaction network are more likely to have similar function and involve in similar biological process.
- Network-based gene function prediction
- Network-based gene prioritization



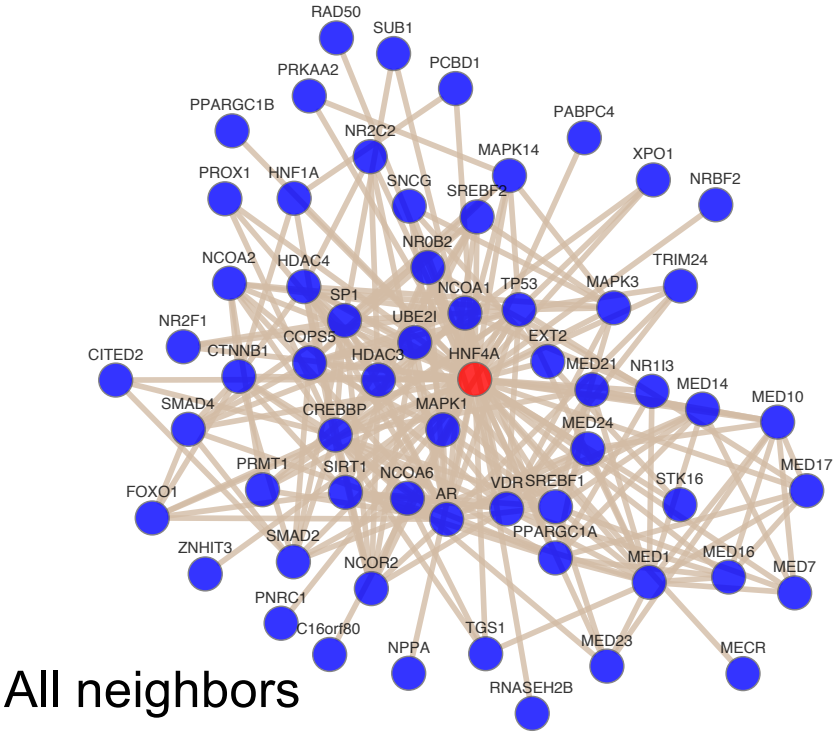
Sharan et al. Mol Syst Biol, 3:88, 2007

Network-based analysis

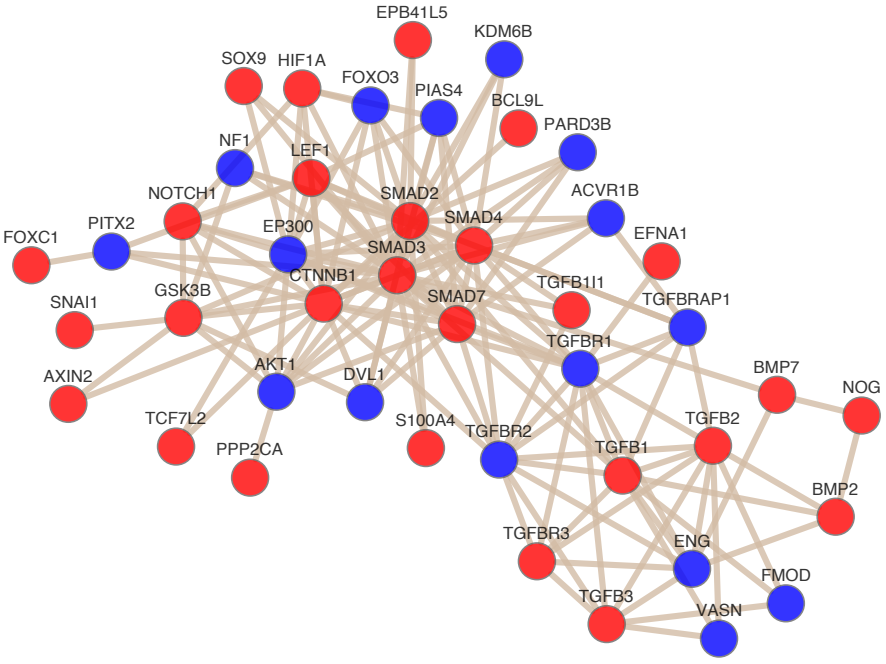
- Direct neighbor-based analysis
- Network module-based analysis
- Network diffusion analysis

Direct neighbor-based analysis

- **Network expansion:** to expand a list of "seed" genes to include other related genes in the network
  - **All neighbors:** to identify all direct neighbors of these seed genes in the network
  - **Enriched neighbors:** to identify all non-seed genes significantly enriched with seed neighbors



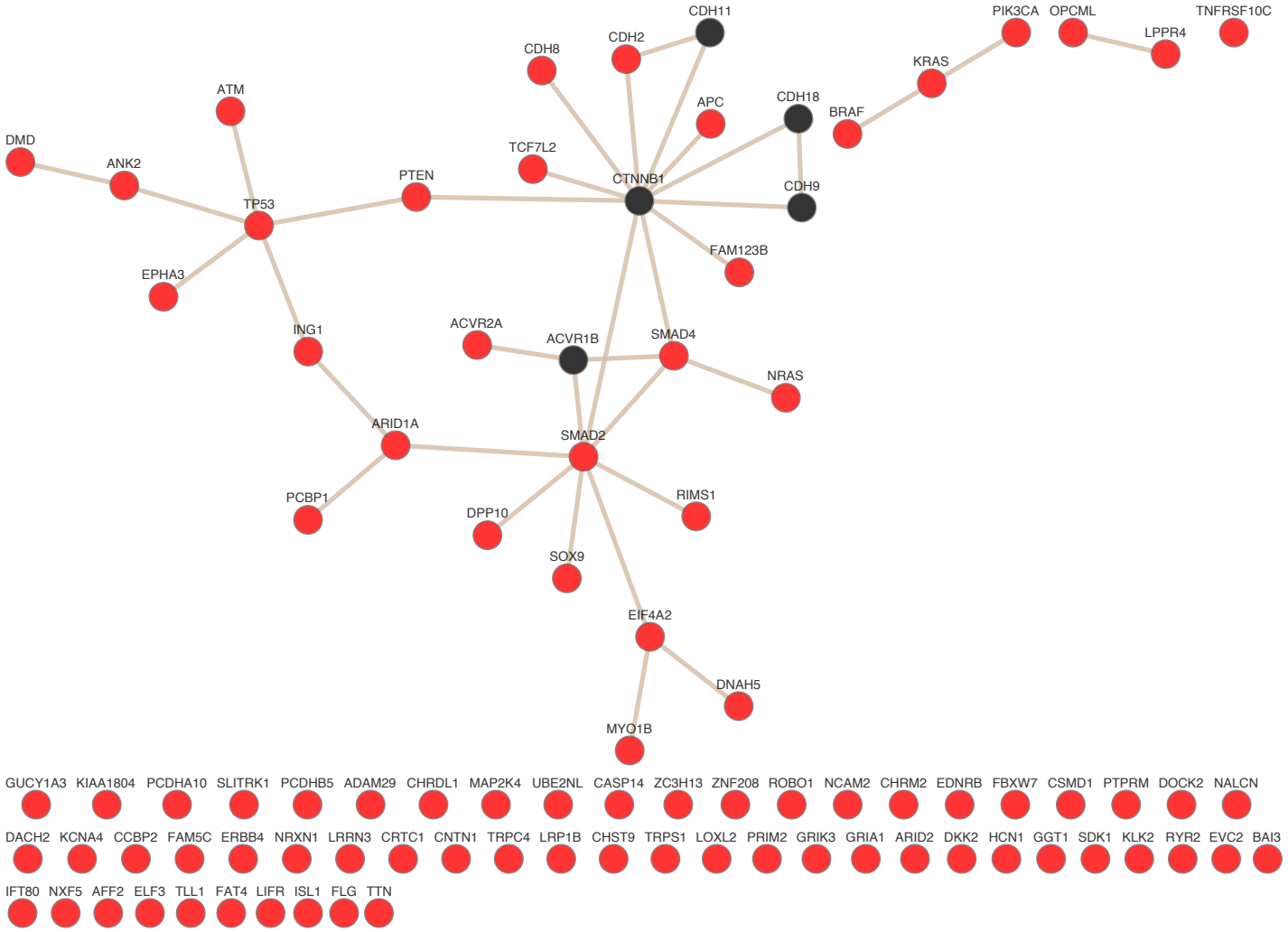
All neighbors



Enriched neighbors

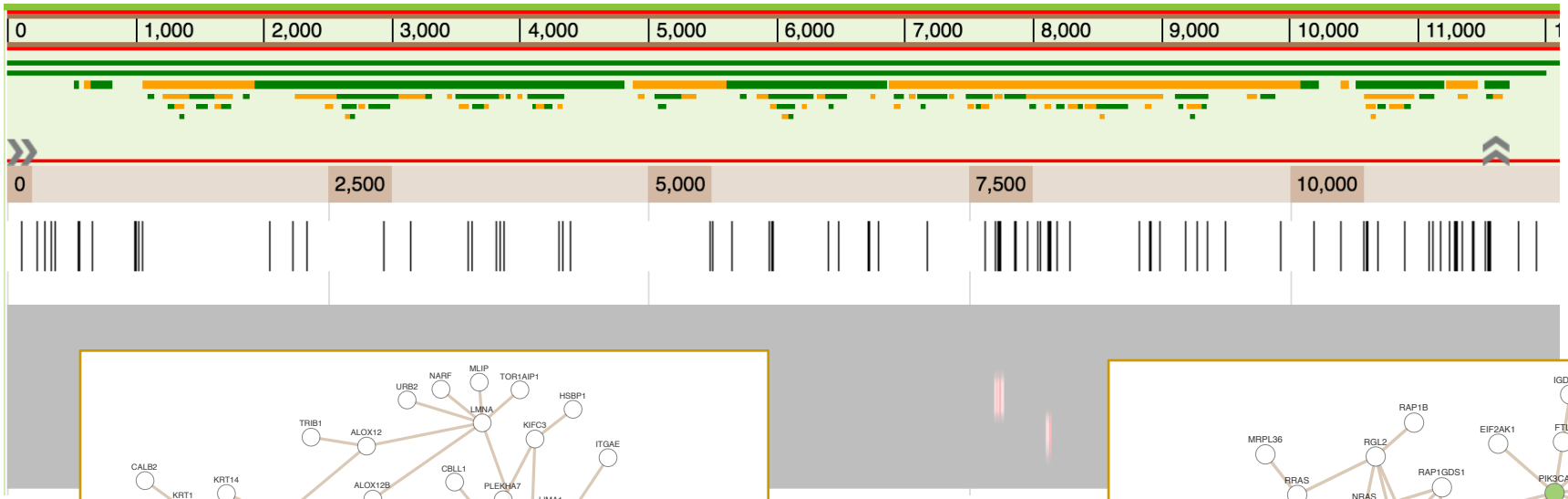
Direct neighbor-based analysis

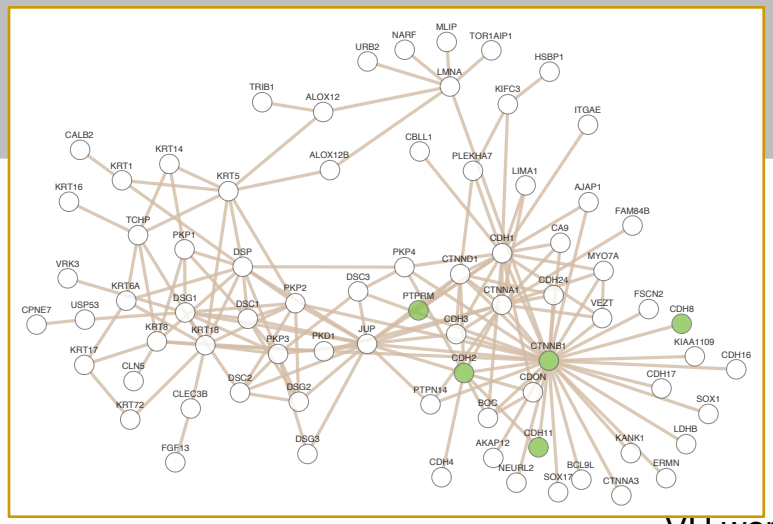
- **Gene prioritization:** to prioritize genes in a list of candidate genes

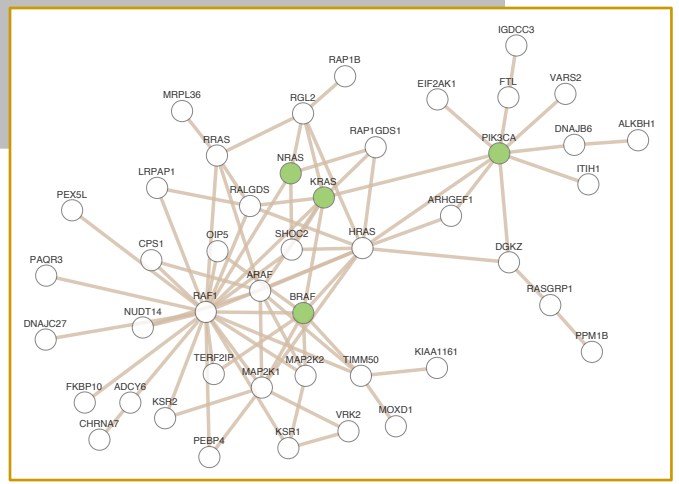


Network module-based analysis

- Identify network modules from a network
  - *e.g.*, NetSAM
- Use network modules as gene sets for enrichment analysis
  - Over-representation analysis; GSEA

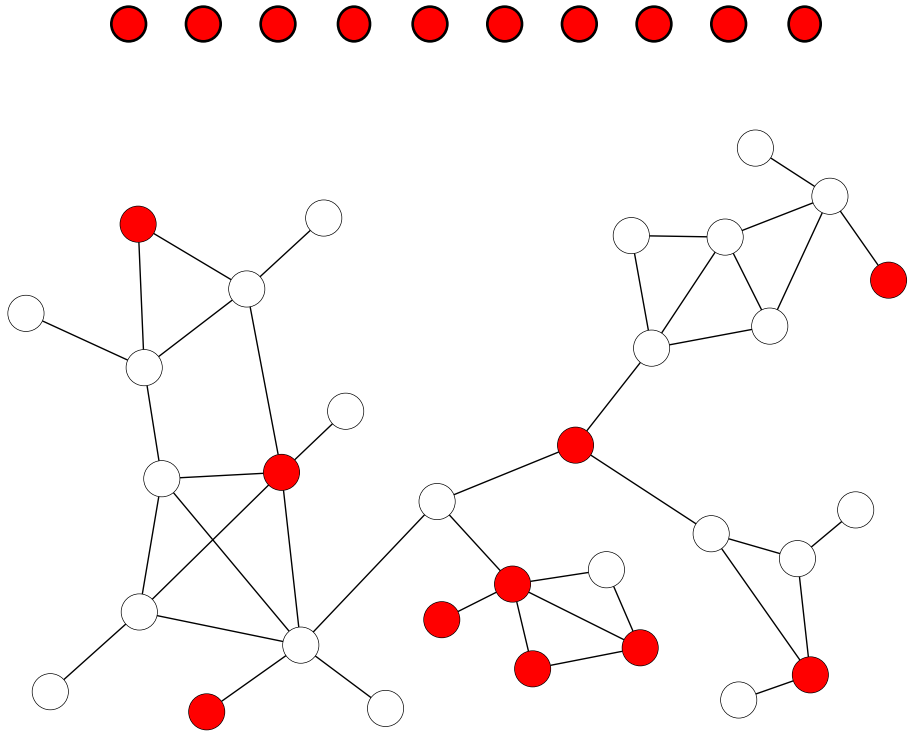






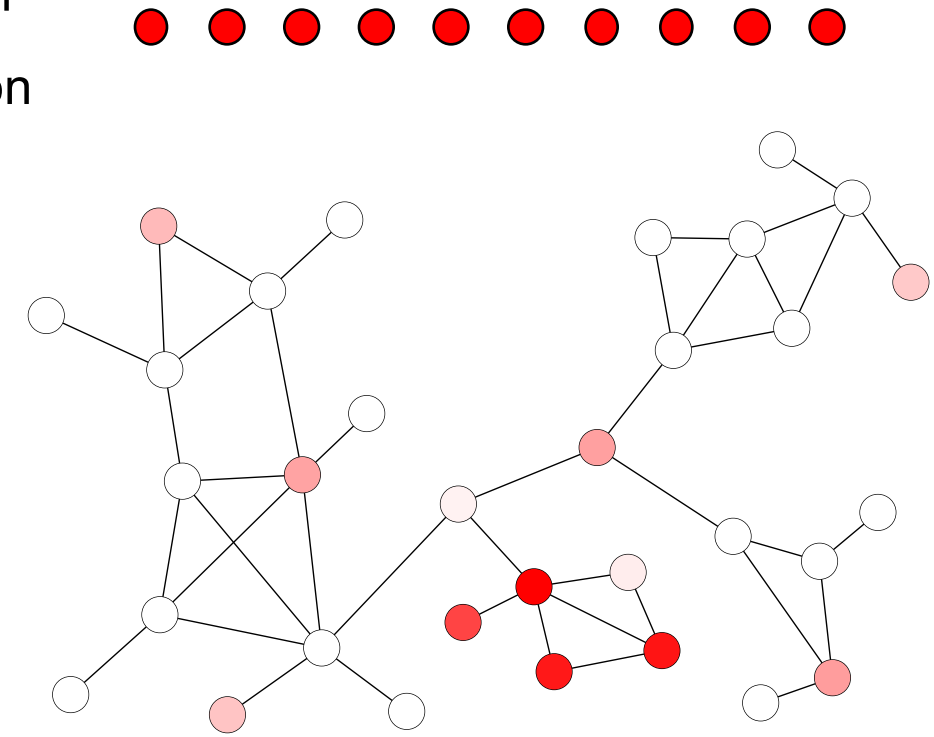
Network diffusion analysis

- Simulating a random walker's behavior on a network (with restart)



Network diffusion analysis

- Simulating a random walker's behavior on a network (with restart)
  - Gene prioritization
  - Network expansion



http://www.gene2net.org

Summary

- Biological networks
  - Physical interaction networks
  - Functional association networks
- Properties of complex networks
  - Scale-free
  - Small world
  - Hierarchical modular
- Network visualization
  - Node-link diagrams
  - NetGestalt
- Network-based analysis
  - Direct neighbor-based analysis
  - Network module-based analysis
  - Network diffusion analysis

Hands-on session

- WebGestalt
  - Pathway analysis for gene lists
  - http://www.webgestalt.org
- NetGestalt
  - Pathway and network analysis for gene lists and data matrices
  - http://www.netgestalt.org
- Gene2Net
  - network diffusion analysis for gene lists
  - http://www.gene2net.org