Network Analysis and Systems Biology

Nicholas Wisniewski
Assistant Adjunct Professor
Department of Medicine; Integrative Biology and Physiology
University of California, Los Angeles
nw@ucla.edu

Motivation

Nature **409** (2001), *Science* **291** (2001)

- Human Genome Project promises:
  - Genetic basis of diseases
  - Personalized medicine
  - Genetic basis of mathematical ability, etc.

Manolio, *N Engl J Med* **363** (2010)

- Genome-Wide Association Studies (GWAS) limitations:
  - SNPs confer only small increase in risk (~1.3 fold)
  - SNPs explain only small fraction of genetic component (<20%)
  - SNPs account for only 0.2-5% of explanation of common diseases

- Common diseases are polygenic + environmental = hard

Introduction

- Network methods are used to analyze high-throughput datasets
  - 47,000 transcripts
  - 20,000 genes

- Systems genetics
  - Find gene modules that are interesting (<100 genes)
  - Find eigengene networks (<20 modules)
  - Find hubs/drivers

- Machine Learning
  - Classify patients based on gene networks
  - Geometric insights about disease dynamics

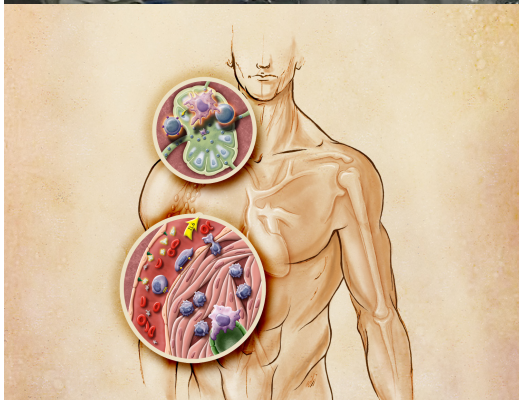
Example Clinical Problem:
Multi Organ Dysfunction (MOD)

- MOD is the **leading cause of mortality** in ICUs
  - MOD develops in 15% of all ICU admissions
  - 80% of all ICU deaths
  - ICU costs of >$100,000 per patient, or $500,000 per survivor.

- **Cause of death** is not understood
  - MOD is an **immune-related** condition
  - Organ tissues looks fine
  - aberrant leukocyte activation
  - systemic inflammatory response

- Established **clinical scoring systems** to predict risk
  - **Limitations**: do not reflect any immune function parameters

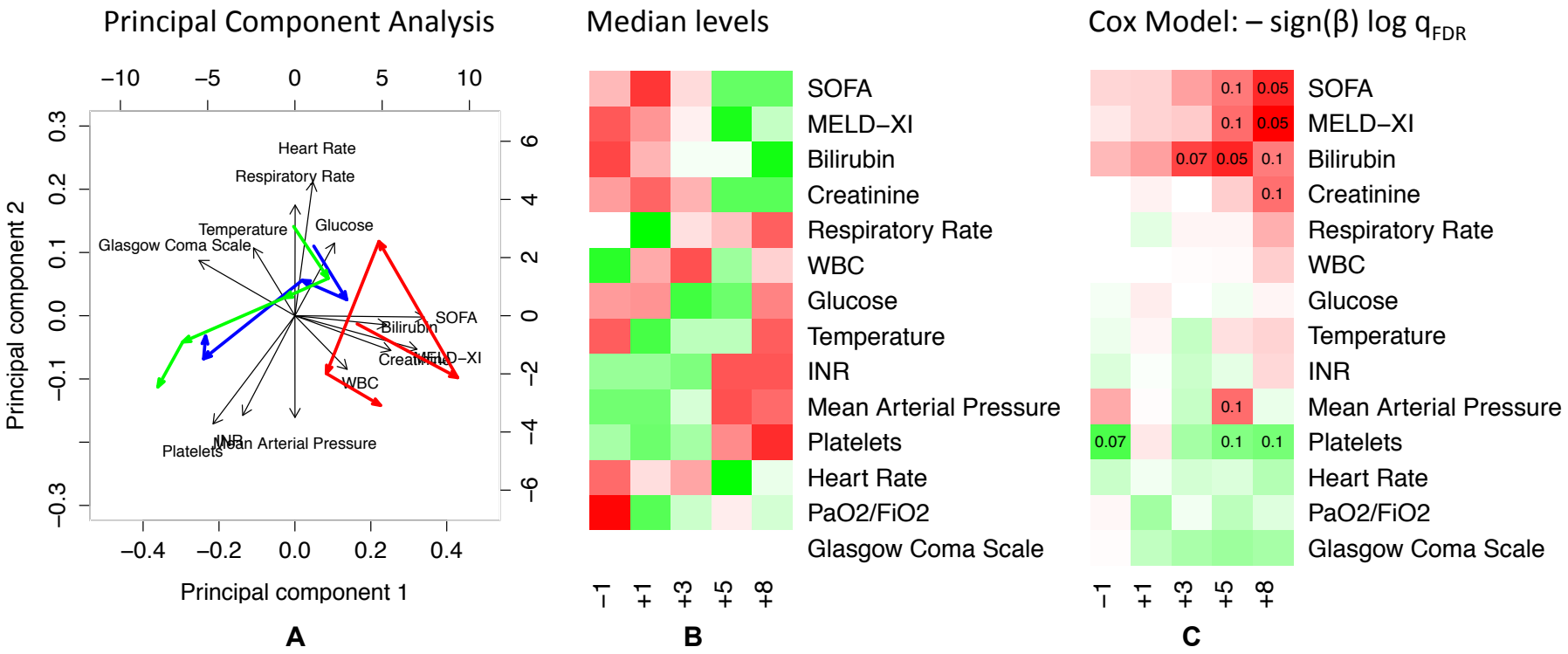
- We **hypothesize**:
  - **Leukocyte gene expression** is indicative of organ dysfunction
  - **Systems-level biomarkers** can detect causes of mortality



Dataset

- 22 AdHF patients
  ICU recovery following mechanical circulatory support device implantation
- Longitudinal blood samples: days -1,1,3,5,8
  PBMC, Next Generation RNA Sequencing, Illumina HiSeq2000 TruSeq
- 14,753 gene expression values per sample
  after filtering on variance and entropy
- 14 clinical phenotypes
  SOFA score, bilirubin, creatinine, glucose, etc.
- 1 survival outcome
  Survival times, 17 live/ 5 dead

Analysis of Clinical Parameters



Bilirubin, creatinine (and corresponding organ dysfunction scores), and platelet count, become more predictive of survival outcomes over time.

Next, we want to be able to make an analogous figure for the genome.

Gene Module Association Study

J. Weiss, A. Karma, W.R. MacLellan, M. Deng, C. Rau, C. Rees, J. Wang, N. Wisniewski, E. Eskin, S. Horvath, Z. Qu, Y. Wang, A.J. Lusis. "Good Enough Solutions" and the Genetics of Complex Diseases. *Circulation Research*, 111: 493-504, 2012.

1. Gene coexpression networks

Infer coexpression network from data

2. Identify gene modules

Cluster and find eigengenes

3. Relate modules to external information

Clinical data, Gene Ontology, Pathways, Literature

4. Find the key genes in interesting modules

Hubs, candidate biomarkers, transcription factor binding sites

5. Predict clinical outcomes

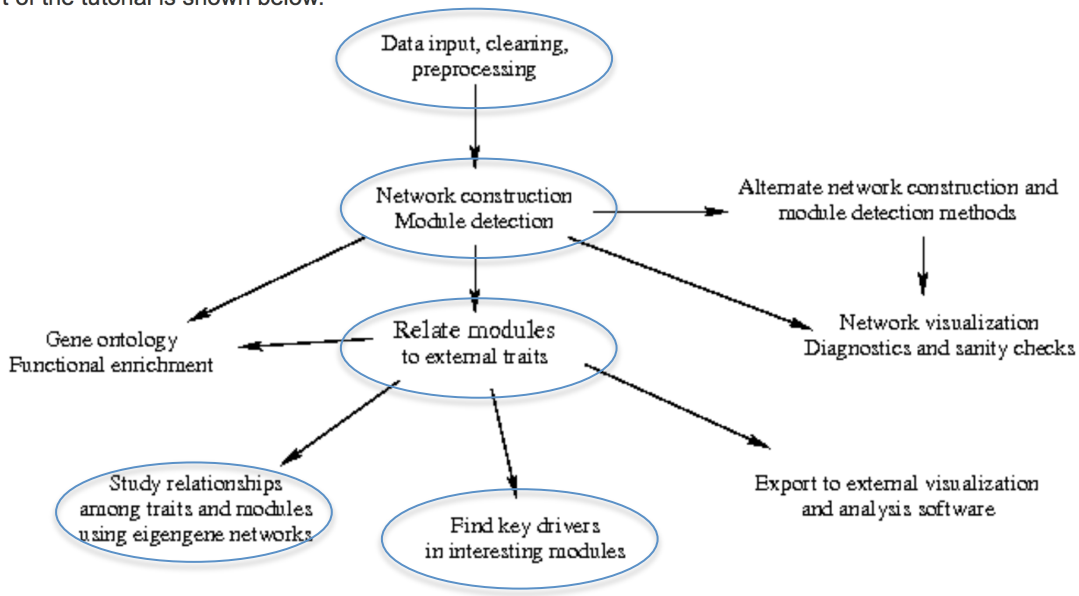
Build predictive models using candidate biomarkers

WGCNA

- Langfelder's tutorials are excellent:
  - https://labs.genetics.ucla.edu/horvath/CoexpressionNetwork/Rpackages/WGCNA/Tutorials/

R Tutorial

The flowchart of the tutorial is shown below.



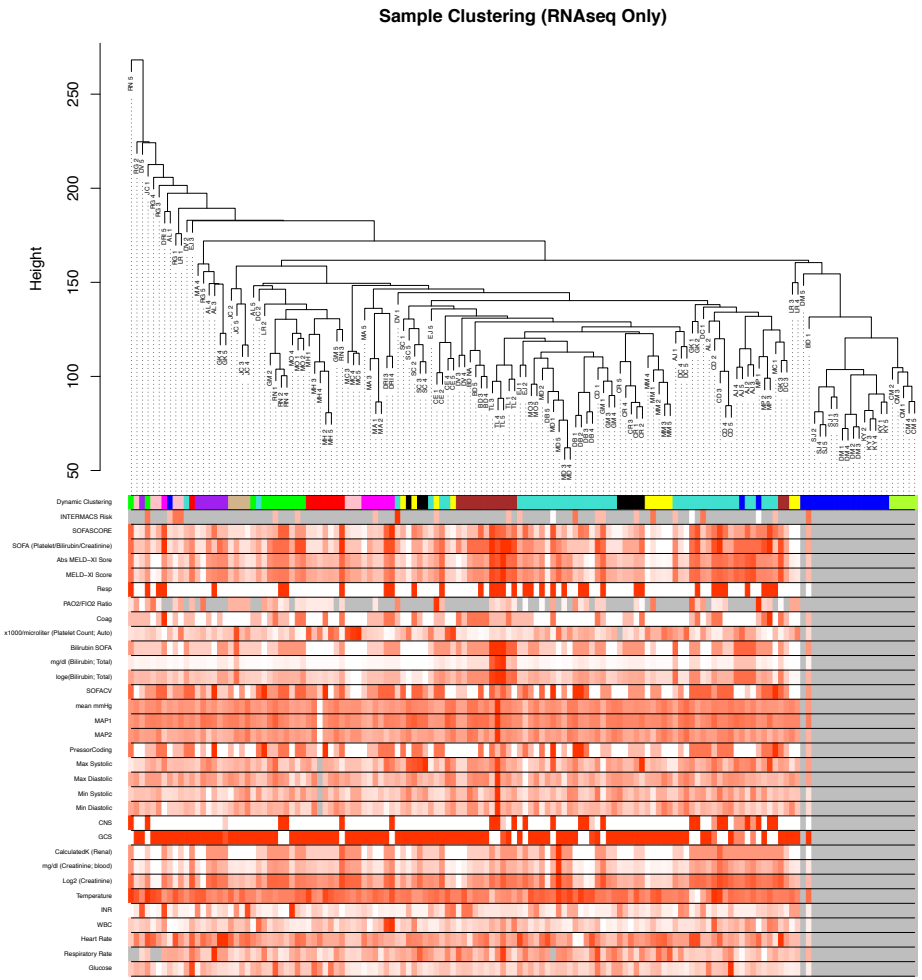
Data Input and Cleaning

Data input and cleaning

- Remove genes and samples with too many missing values
- goodSamplesGenes() checks data for missing entries and zero-variance genes, and returns a list of samples and genes that pass criteria maximum number of missing values. If necessary, the filtering is iterated.

Look for outliers

- Cluster the samples to look for outliers: hclust(dist())
- dist() first converts your array to a distance matrix
- Common distance measures are Euclidean(L2), Manhattan(L1), and Minkowski(LP).



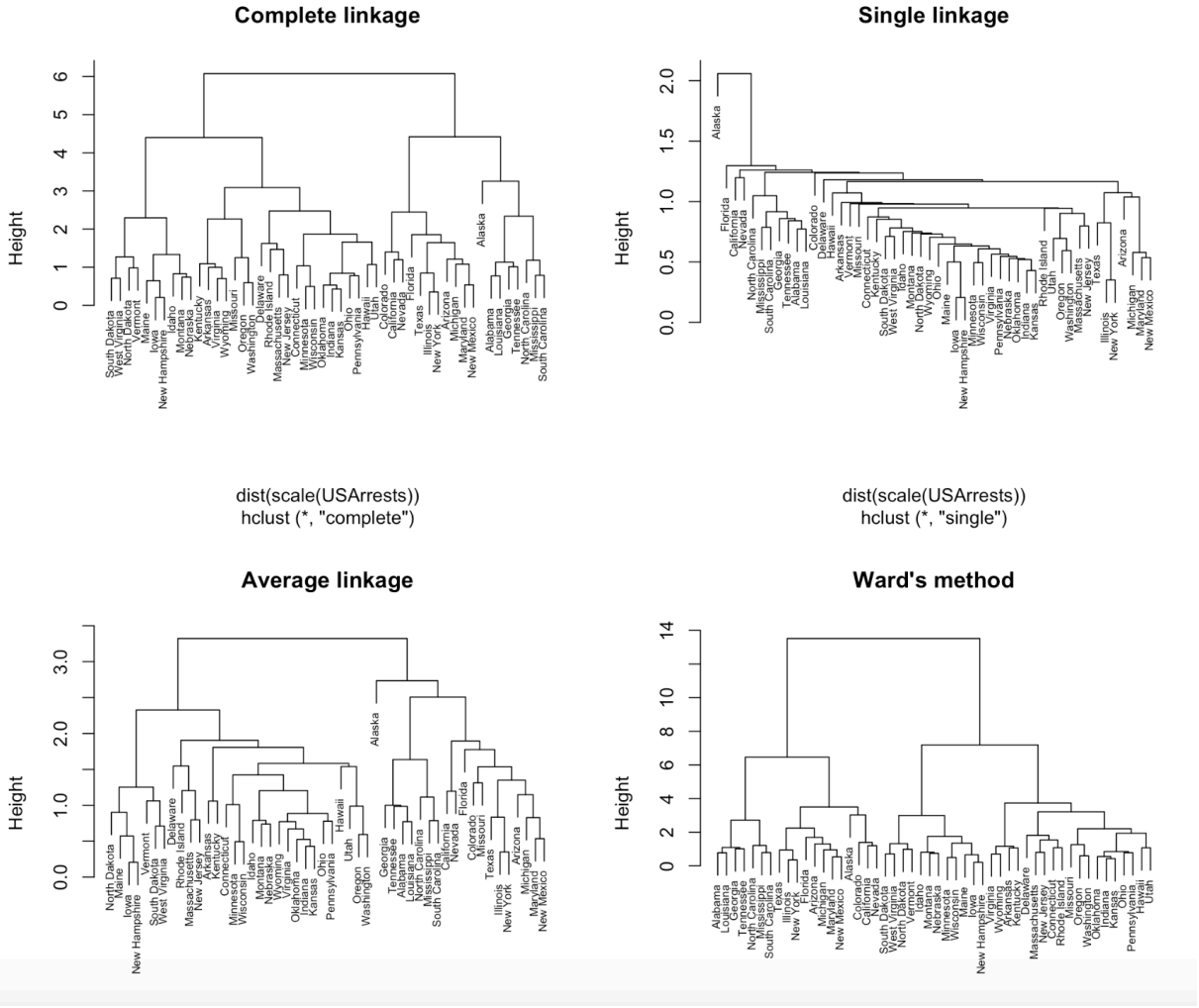
hclust agglomerative methods

- Maximum or **complete linkage clustering**: It computes all pairwise dissimilarities between the elements in cluster 1 and the elements in cluster 2, and considers the largest value (i.e., maximum value) of these dissimilarities as the distance between the two clusters. It tends to produce more compact clusters.
- Minimum or **single linkage clustering**: It computes all pairwise dissimilarities between the elements in cluster 1 and the elements in cluster 2, and considers the smallest of these dissimilarities as a linkage criterion. It tends to produce long, "loose" clusters.
- Mean or **average linkage clustering**: It computes all pairwise dissimilarities between the elements in cluster 1 and the elements in cluster 2, and considers the average of these dissimilarities as the distance between the two clusters.
- **Centroid linkage clustering**: It computes the dissimilarity between the centroid for cluster 1 (a mean vector of length p variables) and the centroid for cluster 2.
- **Ward's minimum variance method**: It minimizes the total within-cluster variance. At each step the pair of clusters with minimum between-cluster distance are merged.

- Complete linkage and Ward's method are generally preferred.

- For a tutorial on clustering, see: http://www.sthda.com/english/wiki/print.php?id=237

Experiment for yourself

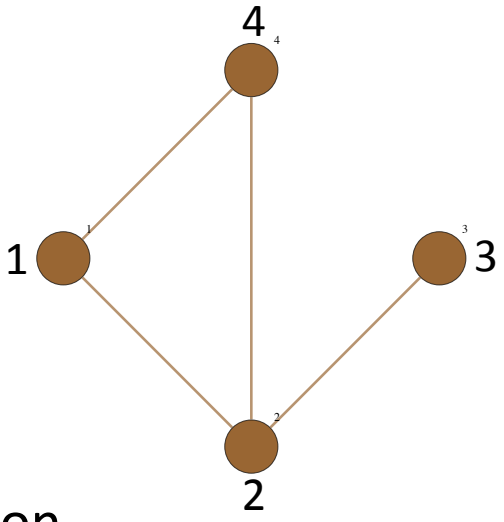


Network Construction and Module Detection

Networks and Graphs

- A network graph is represented by an adjacency matrix
  - Is there a link between nodes i and j?

$$A = (a_{ij}) = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{pmatrix}$$

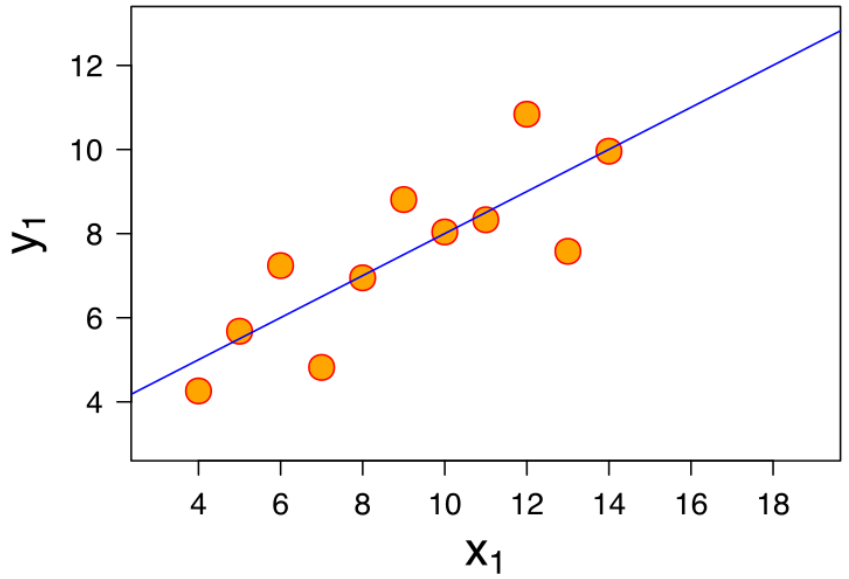


- Unweighted networks
  - {0,1} measures *presence* of a connection

- Weighted networks
  - [0,1] measures *strength* of connection

Inferring Edges

- How to measure strength of connection?
- A common approach is Pearson correlation, but many measures exist.

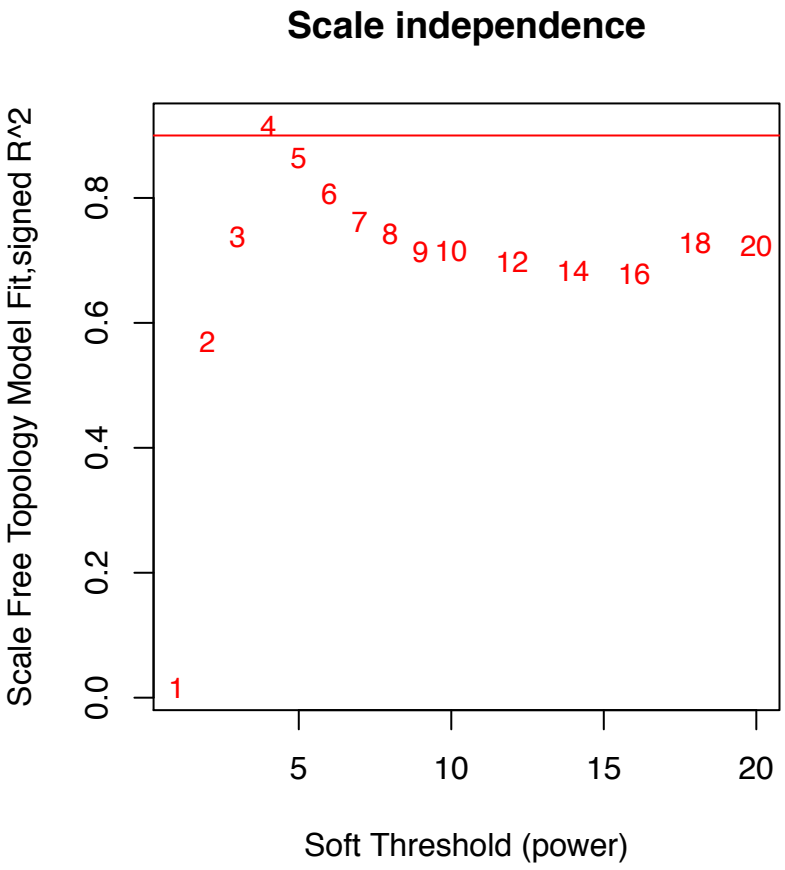


- ✓ Convenient
- ✓ Very fast to compute

- ✓ **adjacency()**

Scale-Free Network Topology

- Biological networks are approximately scale-free (Barabasi-Albert model)
- Correlation networks aren't scale free.
- WGCNA corrects this by exponentiating the correlation coefficients until approximately scale free. (Soft thresholding).
- **pickSoftThreshold()**

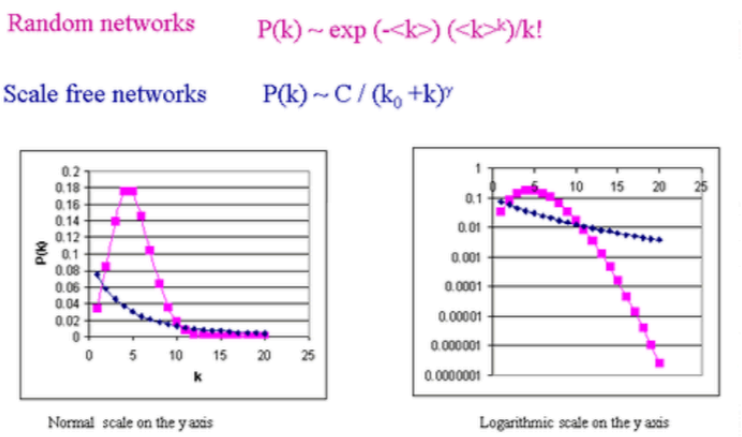


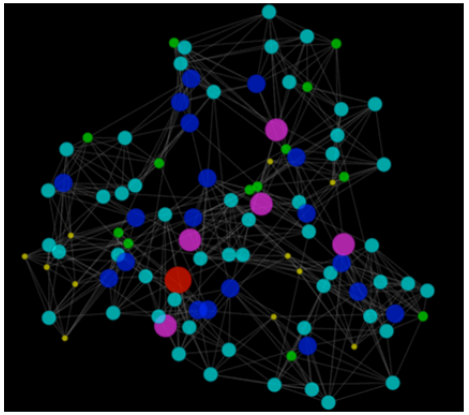
Scale Free Networks

Barabasi-Albert

Preferential attachment: Start with a small network and add new nodes. The more connected a node i is, the more likely it will receive new links.

$$p_i = \frac{k_i}{\sum_j k_j} \tag{3}$$





Degree distribution: power law (scale-free).

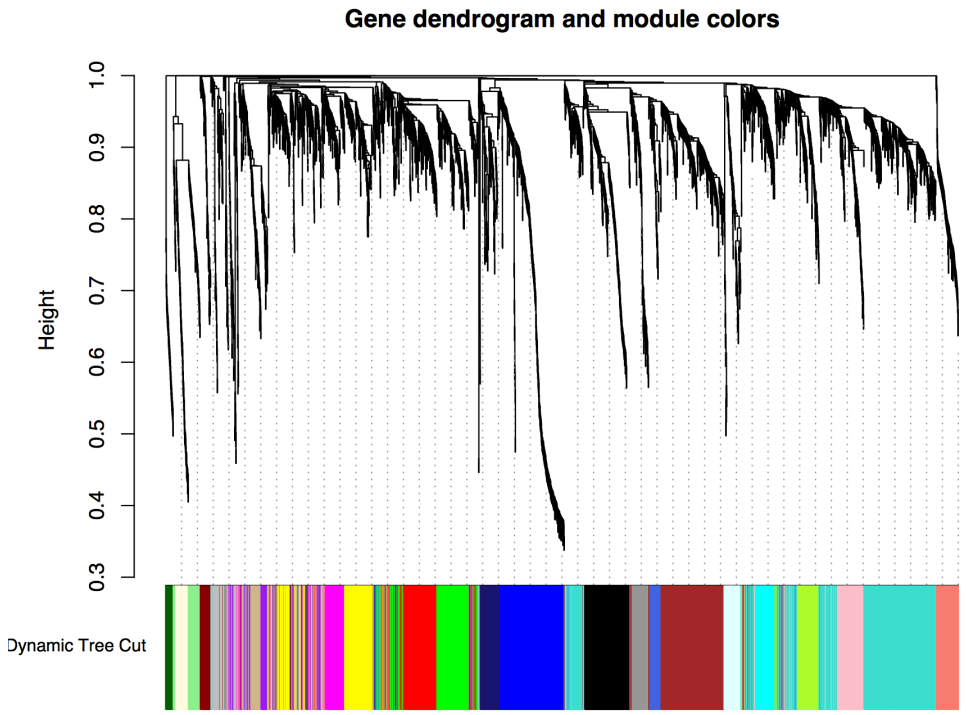
$$p(k) \sim k^{-\alpha} \tag{4}$$

Topological overap

- Naively we can use hierarchical clustering directly on the adjacency matrix
- WGCNA uses “topological overlap” (like mutual friends) to map the adjacency matrix to a stronger similarity matrix for clustering.
- **TOMsimilarity()**

Hierarchical Clustering

- Tree-based structure
- Branches when dissimilar
- Assign colors to modules
- **hclust()**
- **cutreeDynamic()**



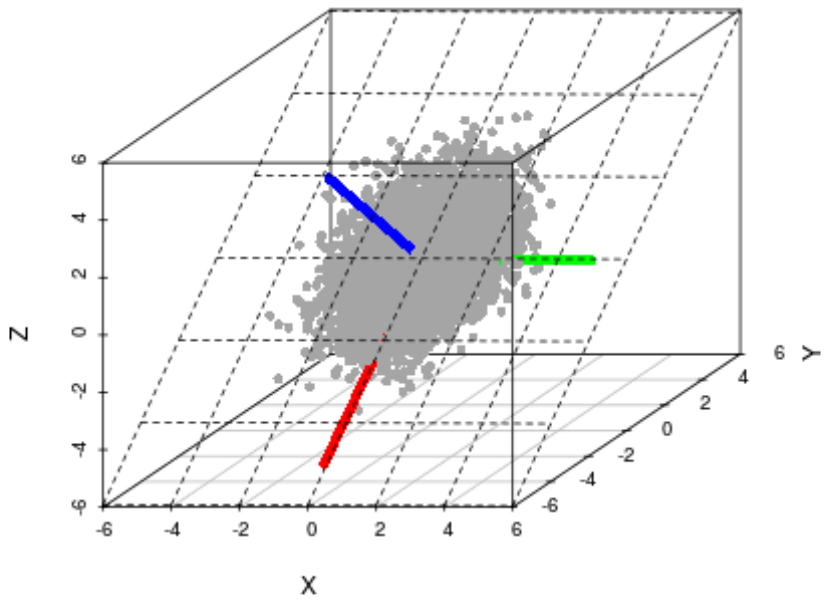
cutreeDynamic

- Module identification amounts to the identification of individual branches (”cutting the branches off the dendrogram”).
- Lots of options to play with:
- set minimum cluster size
- set cut height manually, or
- “hybrid” method allows adjustment through the deepSplit variable

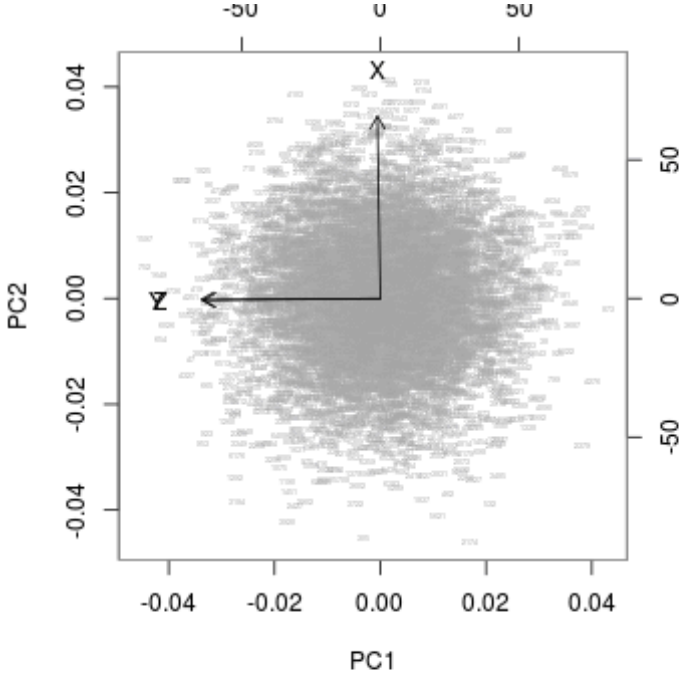
Module Eigengenes

- Data compression technique
- Find 1st principal component of each module, throw out the rest.

Gene space (gene X, Y, Z)

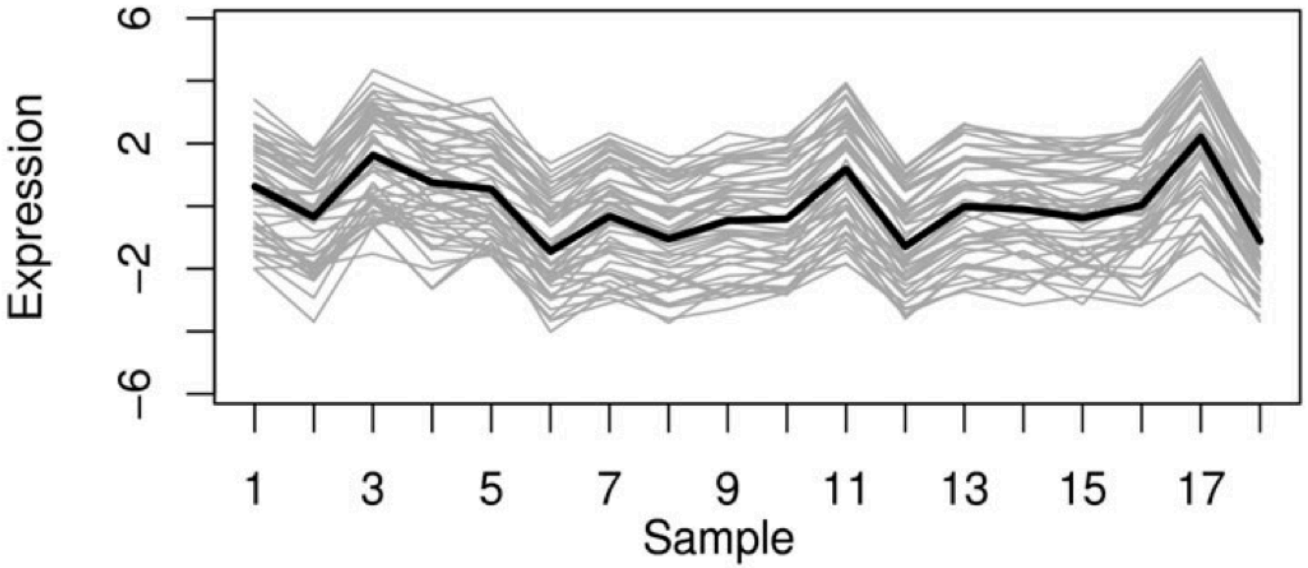


Subspace (PC1, PC2)



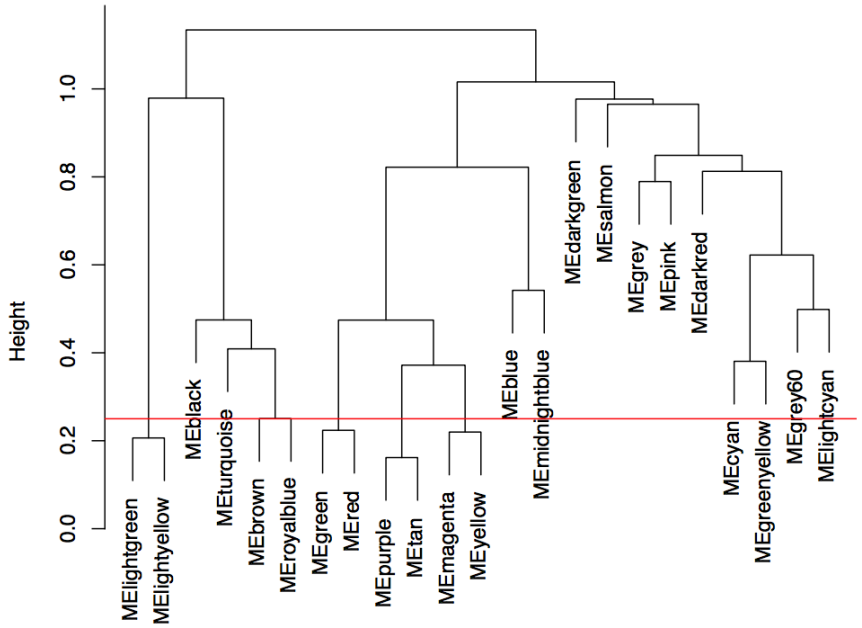
Module Eigengenes

- Looks like the average expression inside the module.
- By reducing the whole module to a single vector, we can compute correlation between modules just like we computed correlations between genes.
- **moduleEigengenes()**



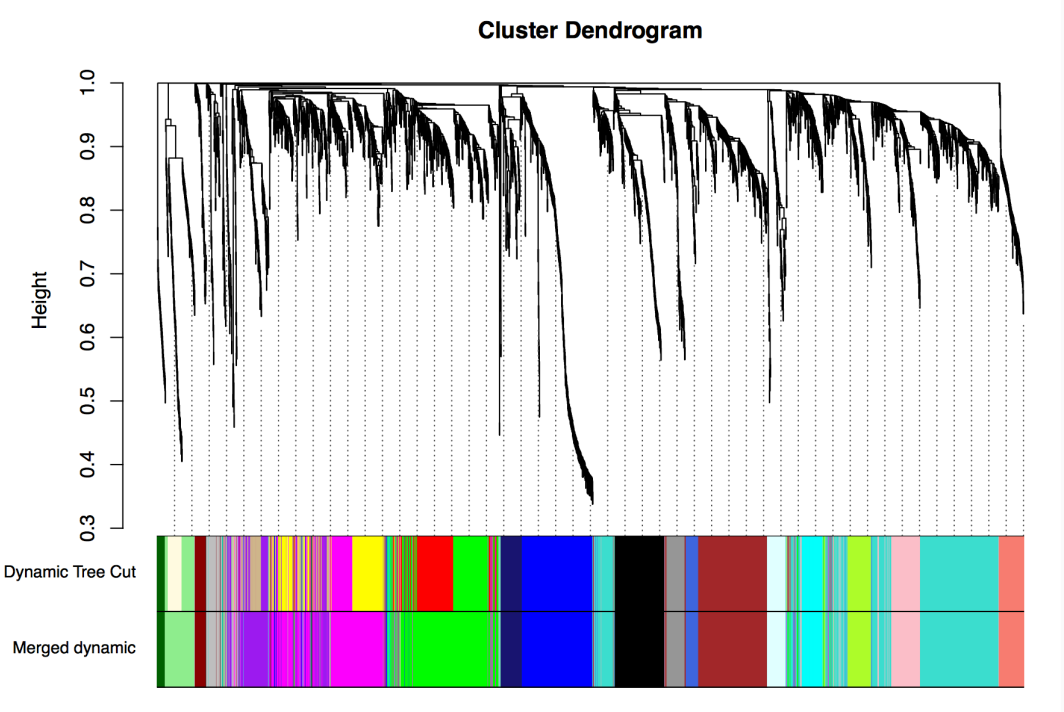
Cluster Eigengenes

- To quantify co-expression similarity of entire modules, we calculate their eigengenes and cluster them on their correlation
- Choose what you think is the best cut height to form merged superclusters



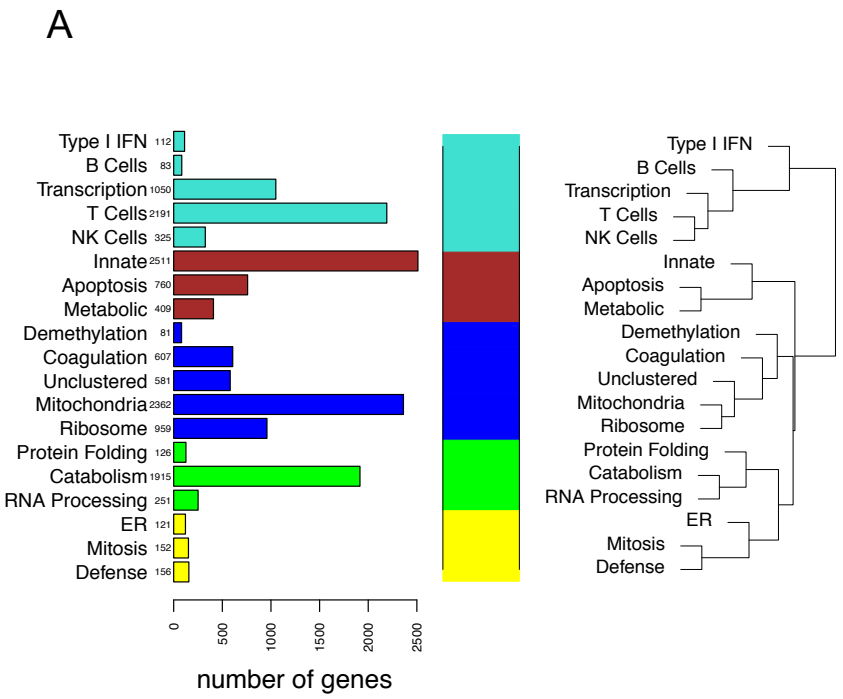
Merge Modules

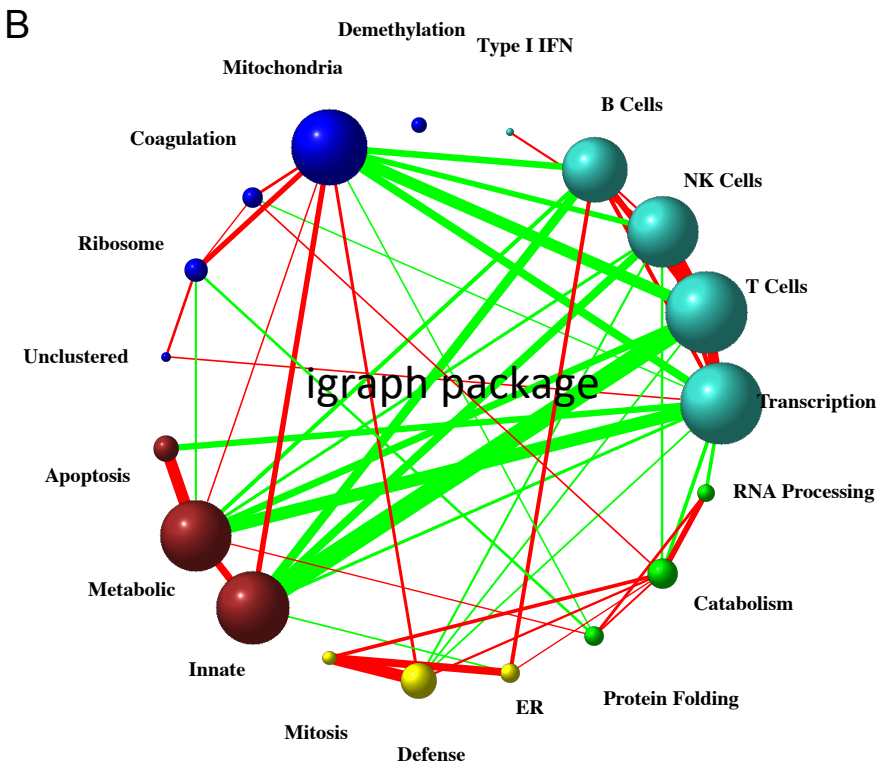
- Compute the resulting gene-module assignments and compare them to the previous module assignments
- **mergeCloseModules()**



Study relationships between traits and modules using eigengene networks

Eigengene Network

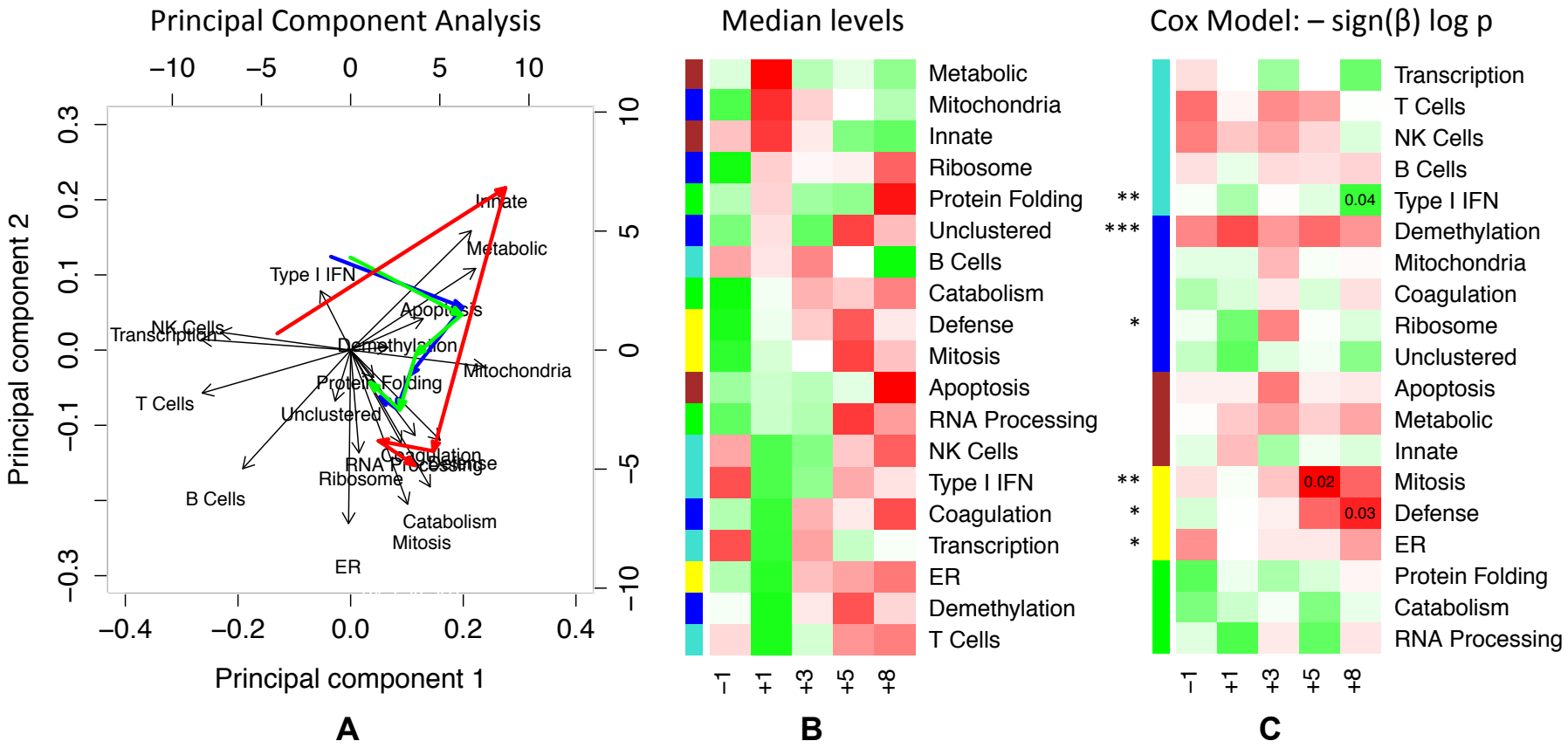




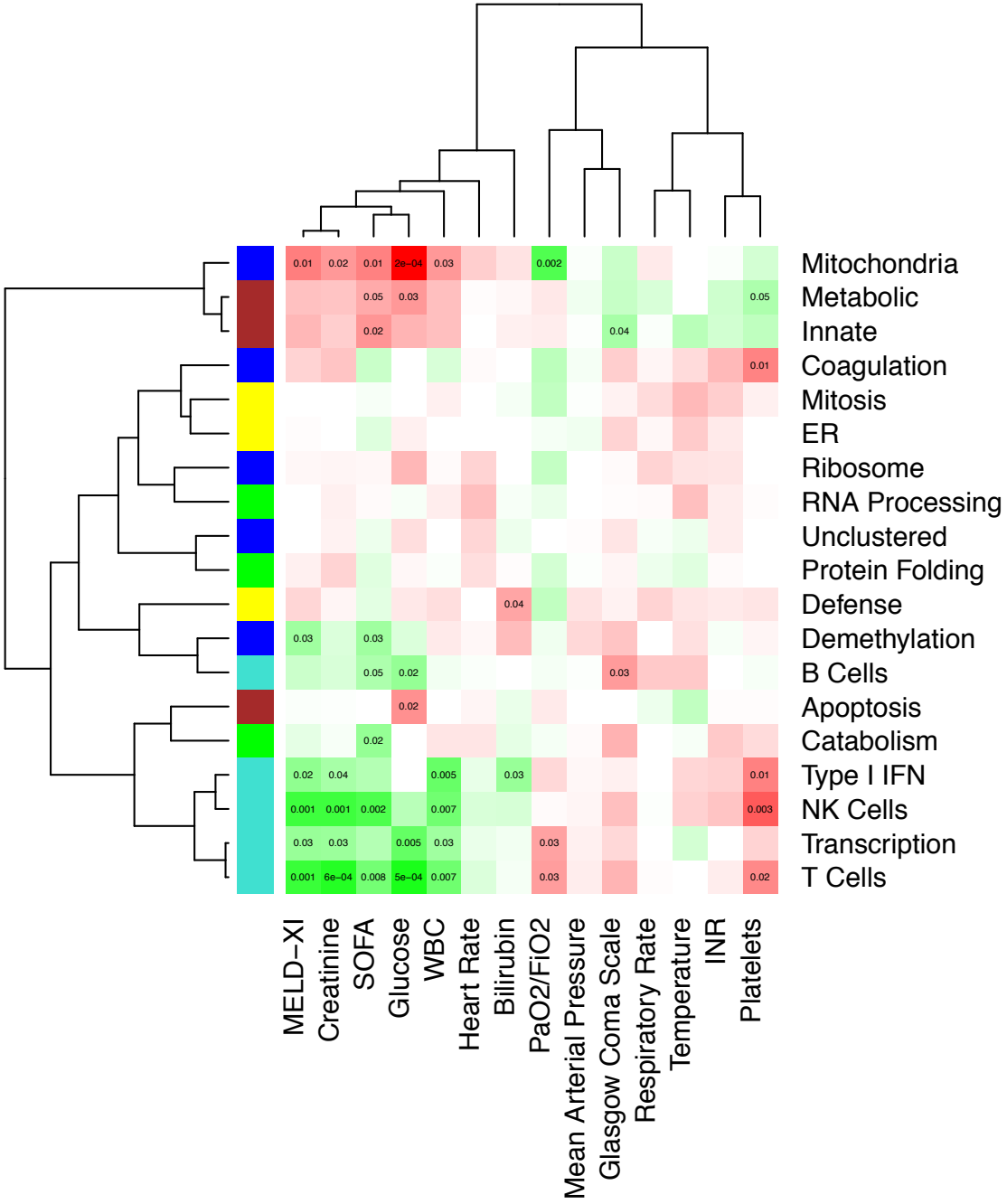
19 modules, 5 superclusters. Strong anticorrelations between innate and adaptive immune system superclusters, and mitochondria

Gene ontology analysis: GOSim package. Webtools like DAVID also exist.

Analysis of Eigengenes



Mitosis and Defense modules are predictive at late timepoints. Looking at GO enrichment terms, these features are related to red blood cell production and extracellular matrix (ECM) degradation, and indicate disseminated coagulopathy.



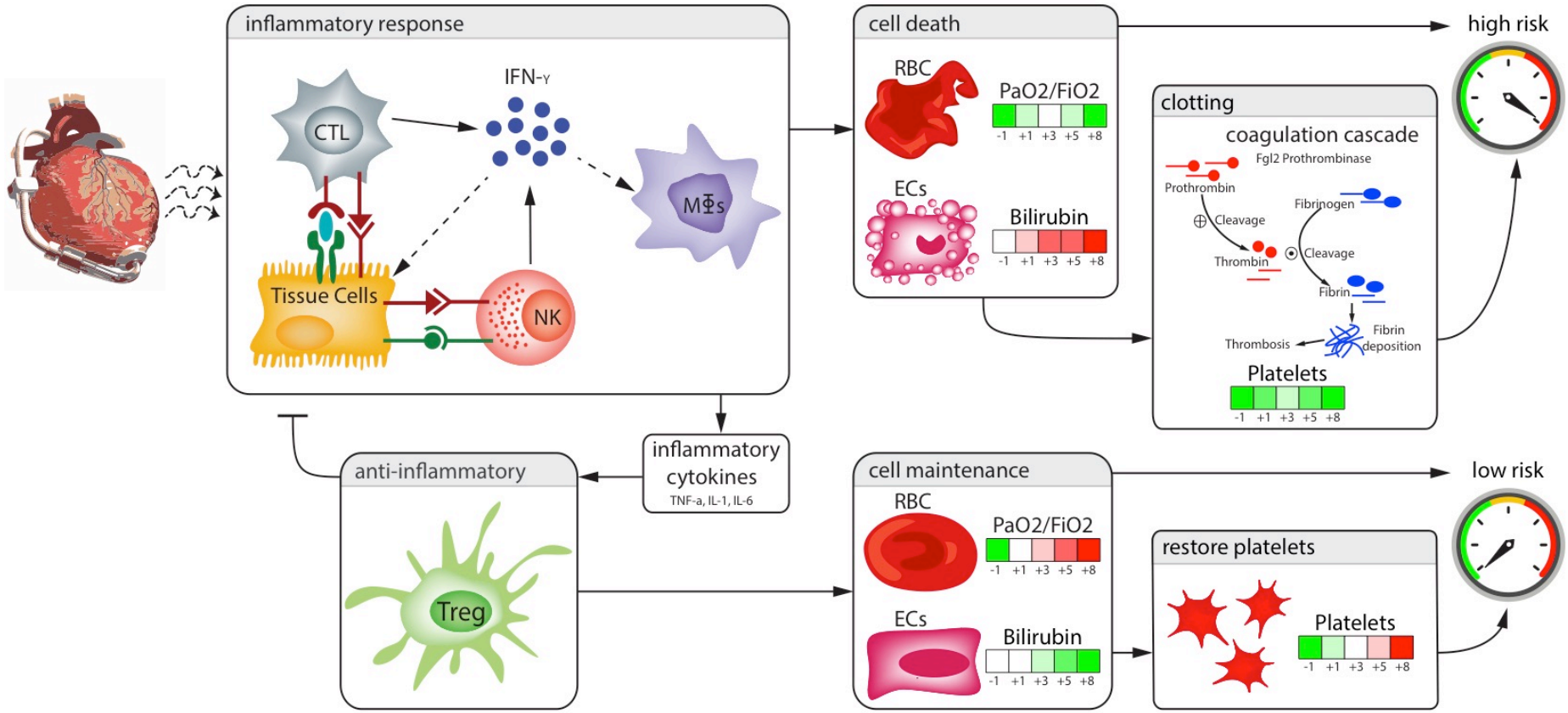
Mapping Eigengenes to Phenotypes

We used a linear mixed effect model lmer() to find associations between eigengenes and clinical parameters.

Innate and adaptive immune system modules have strong associations with organ dysfunction

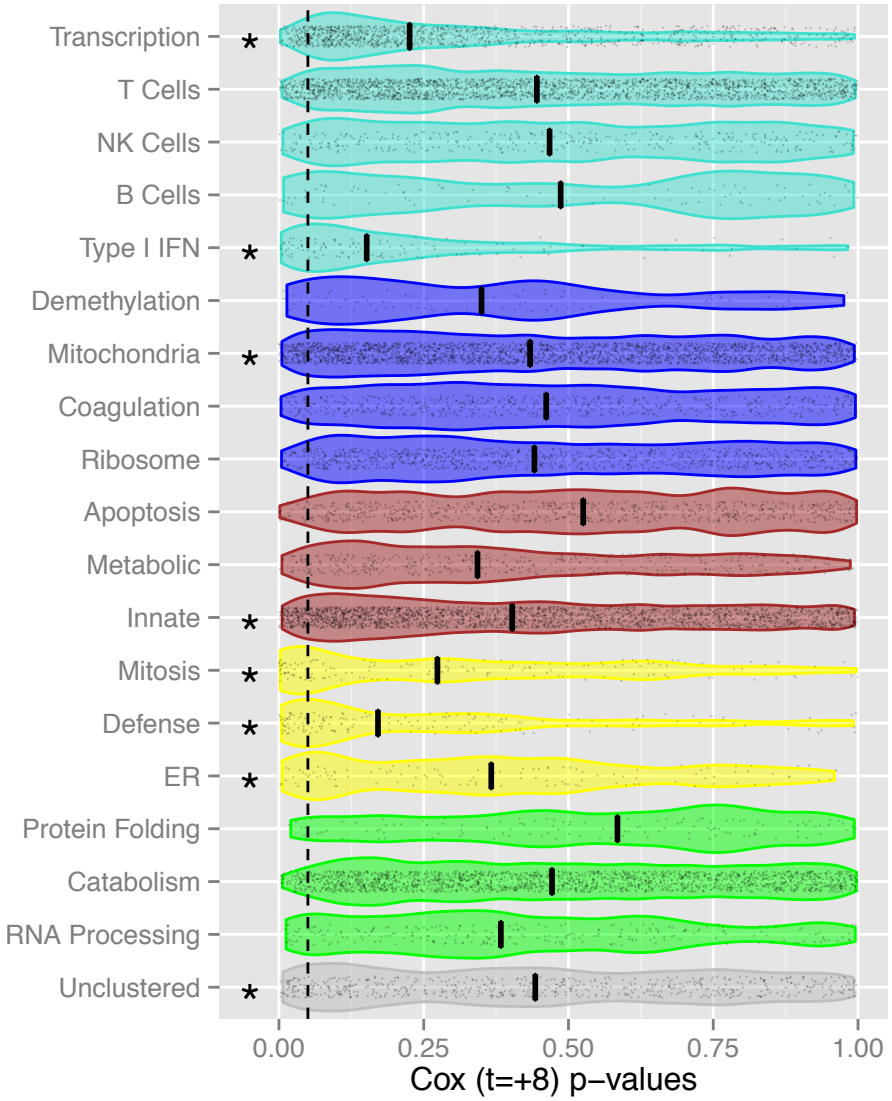
Color = – sign(β) log qval

End Result: Hypothesized Organ Dysfunction Biology



Gene Significance

- Gene significance measures generally look like GS= –log pvalue for an appropriate model (e.g. linear regression, linear mixed effect model, Cox proportional hazards)



Module Membership

- To escape the idea that each gene can only be in a single module, we can compute the correlation between a gene and each module eigengene.
- The result is a “Module Membership” between -1 and 1.

Gene connectivity (centrality)

- There are many measures of centrality
- Degree centrality: Simplest is just the sum of the rows or columns of the adjacency matrix
- There are many others with interesting meaning implemented in the igraph R package.

Degree Centrality

The number of edges connected to a particular node.

$$C_i = k_i = \sum_j a_{ij} \tag{5}$$

- probability of catching whatever is flowing through the network (virus, information)

Closeness Centrality

The distance d between two nodes is their shortest path (number of hops).

$$\text{farness}_i = \sum_j d(i, j) \tag{6}$$

$$C_i = \frac{1}{\sum_j d(i, j)} \tag{7}$$

The more central a node, the lower its total distance to all other nodes.

- how long it takes to spread information to all other nodes sequentially.

Betweenness Centrality

The number of times a node acts as a bridge between two other nodes.

- control of a human on the communication between other humans in a social network.

Eigenvector Centrality

Connections to high scoring nodes count more than lesser nodes.

$$C_i = \frac{1}{\lambda} \sum_j a_{ij} C_j \tag{8}$$

- influence of a node in the network

PageRank

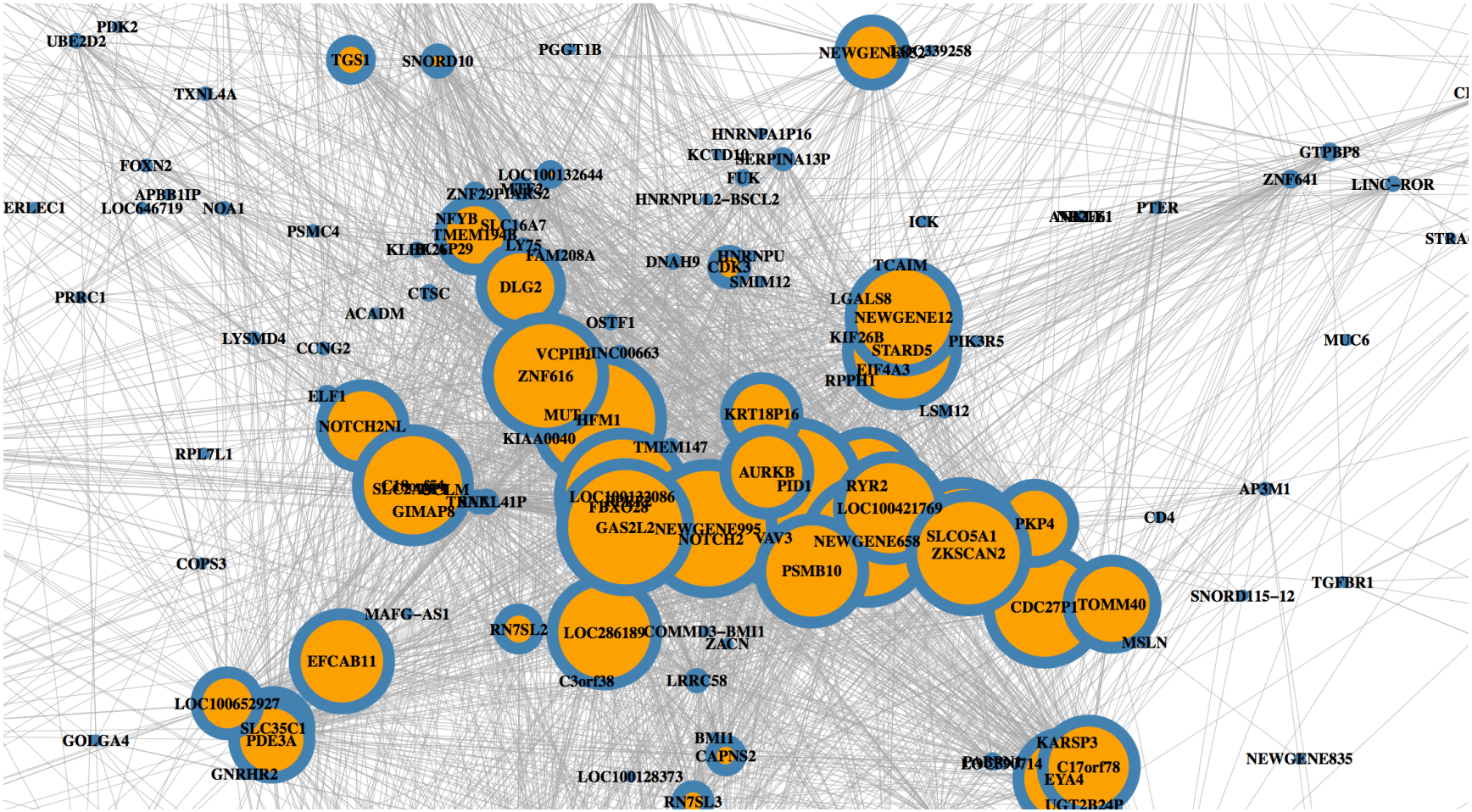
Eigenvector centrality with a different weighting scheme based on the degree centrality of the connections.

$$C_i = \alpha \sum_j a_{ji} \frac{C_j}{\sum_j a_{ji}} + \frac{1-\alpha}{N} \tag{9}$$

Notice it has the same form as the eigenvector centrality, but there is now a "damping factor" α,

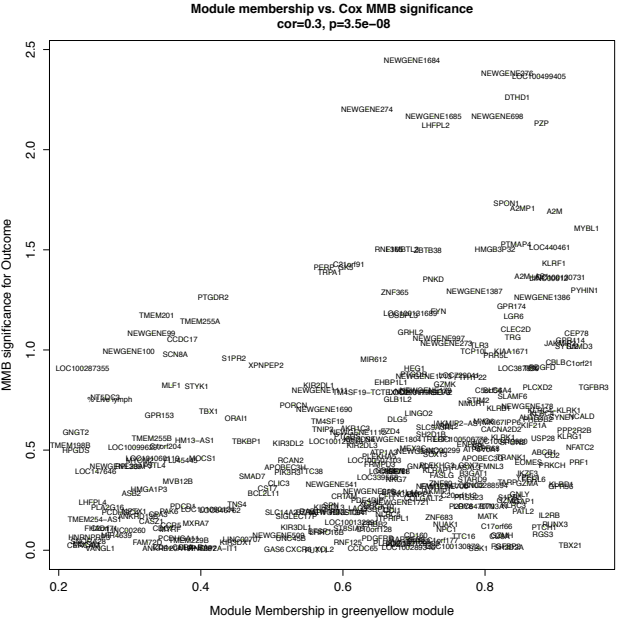
The PageRank theory holds that an imaginary surfer who is randomly clicking on links will eventually stop clicking. The probability, at any step, that the person will continue is a damping factor α. Various studies have tested different damping factors, but it is generally assumed that the damping factor will be set around 0.85.

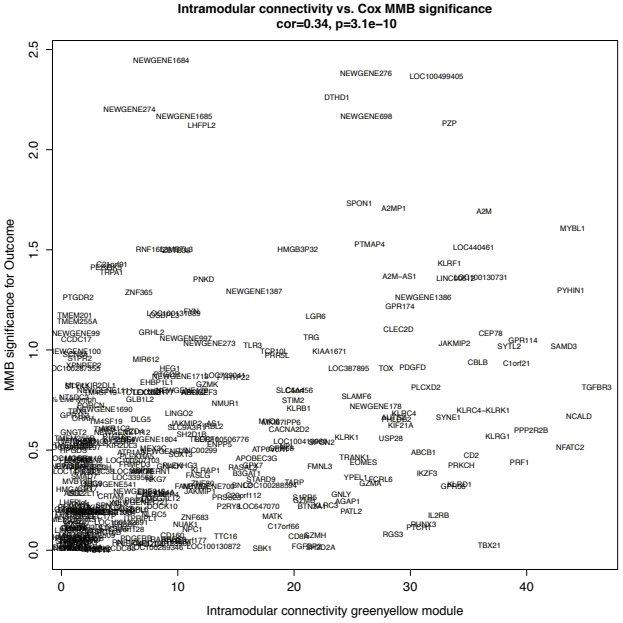
igraph

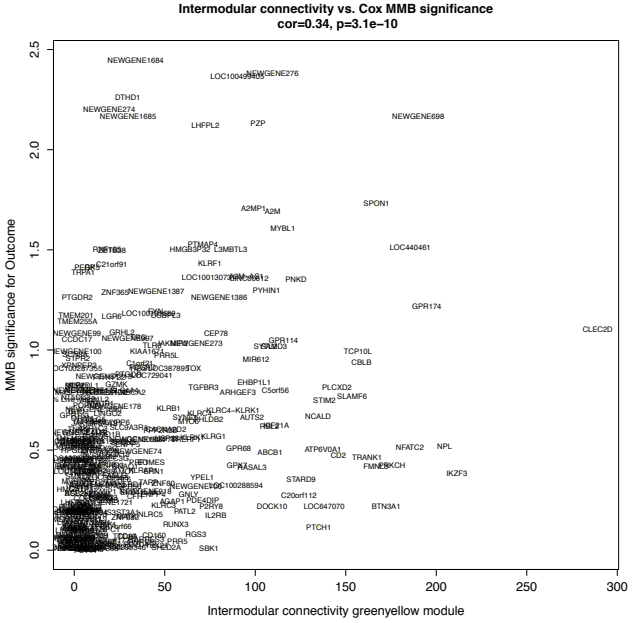

PDK2
UBE2D2
TGS1
SNORD10
PGGT1B
TXNL4A
HNRNPA1P16
KCTD10
SERPINA13P
FOXN2
LOC100132644
FUK
APBB1IP
ERLEC1
LOC646719
NOA1
HNRNPUL2-BSCL2
PSMC4
NFYB
SLC16A7
TMEM194B
LY75
KLHDC26P29
FAM208A
DNAH9
HNRNPU
CDK3
SMIM12
DLG2
PRRC1
CTSC
ACADM
OSTF1
LYSMD4
CCNG2
VCPIP1
LINC00663
ZNF616
ELF1
NOTCH2NL
MUT
HFM1
KIAA0040
TMEM147
KRT18P16
RPL7L1
AURKB
PID1
RYR2
GIMAP8
LOC100133086
FBXO28
GAS2L2
NEWGENE995
NOTCH2
VAV3
LOC100421769
NEWGENE658
SLCO5A1
ZKSCAN2
PKP4
PSMB10
COPS3
MAFG-AS1
RN7SL2
LOC286189
COMMD3-BMI1
ZACN
EFCAB11
C3orf38
LRRC58
LOC100652927
SLC35C1
PDE3A
GOLGA4
GNRHR2
BMI1
CAPNS2
LOC100128373
RN7SL3
TCAIM
LGALS8
NEWGENE12
KIF26B
STARD5
PIK3R5
EIF4A3
RPPH1
LSM12
ICK
PTER
GTPBP8
ZNF641
LINC-ROR
MUC6
AP3M1
CD4
TGFBR1
SNORD115-12
CDC27P1
TOMM40
MSLN
KARSP3
C17orf78
EYA4
NEWGENE835

Identify key genes

- Using the GS, MM, and centrality measures, we can identify genes that have a high statistical significance, as well as high module membership or centrality. These hub genes are likely to be the most interesting.





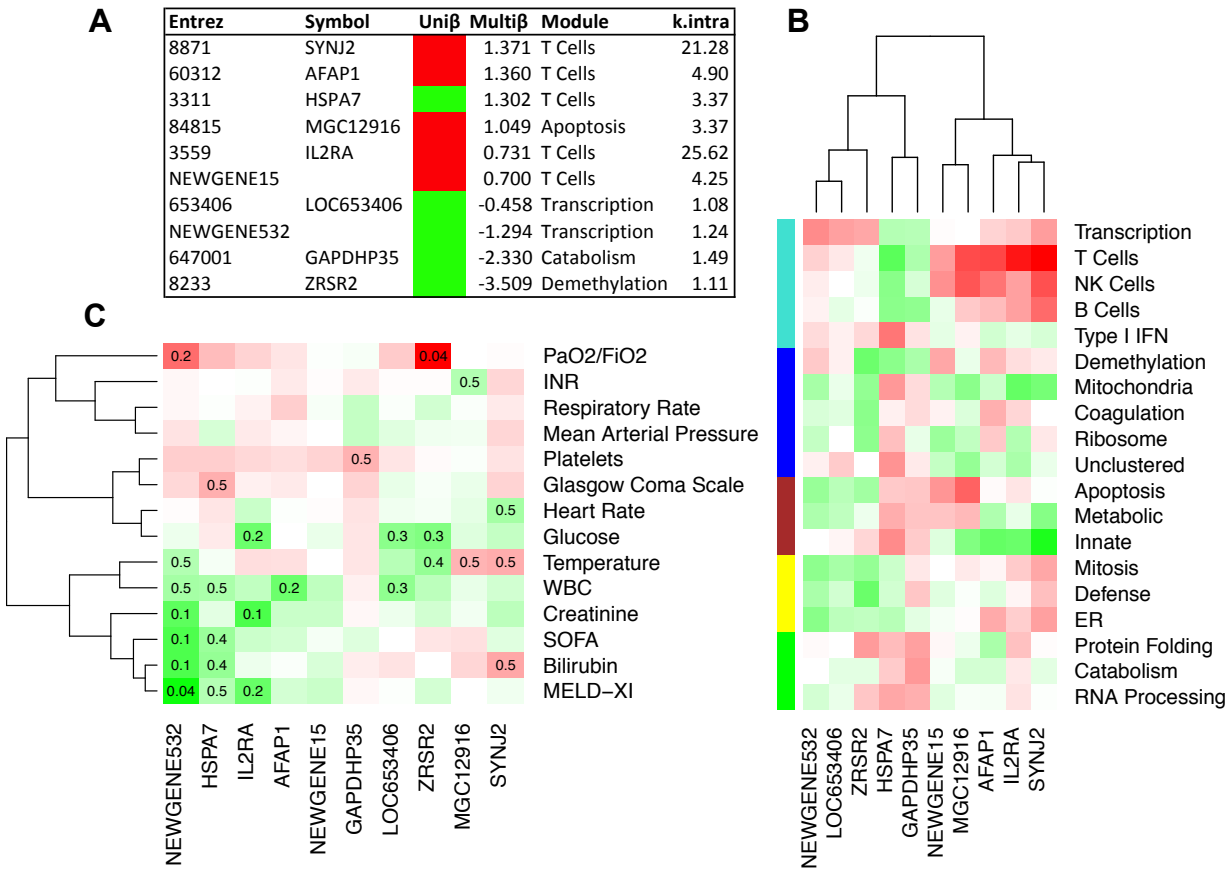


Map any gene set onto the network

Survival Model

A

Entrez	Symbol	Uniβ	Multiβ	Module	k.intra
8871	SYNJ2		1.371	T Cells	21.28
60312	AFAP1		1.360	T Cells	4.90
3311	HSPA7		1.302	T Cells	3.37
84815	MGC12916		1.049	Apoptosis	3.37
3559	IL2RA		0.731	T Cells	25.62
NEWGENE15			0.700	T Cells	4.25
653406	LOC653406		-0.458	Transcription	1.08
NEWGENE532			-1.294	Transcription	1.24
647001	GAPDHP35		-2.330	Catabolism	1.49
8233	ZRSR2		-3.509	Demethylation	1.11



Gene Significance

Module Membership

Other algorithms

- MICA
  - Uses the maximal information coefficient, Bayesian clustering.
  - Linear and Nonlinear gene associations
  - Maximal information component analysis: a novel non-linear network analysis method. *Frontiers in Genetics* 4:28, (2013).
  - https://github.com/ChristophRau/wMICA

- ARACNE
  - uses mutual information, no clustering.
  - related to Bayesian networks
  - Linear and Nonlinear gene associations
  - Reverse engineering cellular networks. *Nature Protocols* **1**, 662 - 671 (2006)
  - minet(), bnlearn(),

Hi-D Multilevel Modeler

- Interactive cloud-based dashboard
- Dozens of machine learning models (supervised and unsuperised)
- Generalized linear mixed effect models
- Correlation networks, Bayesian networks, ARACNE
- Designed for the eigengene + phenotype stage of analysis, or for a reasonable number of genes; not suited yet for full analysis.
- Not publicly available, Contact me if you're interested