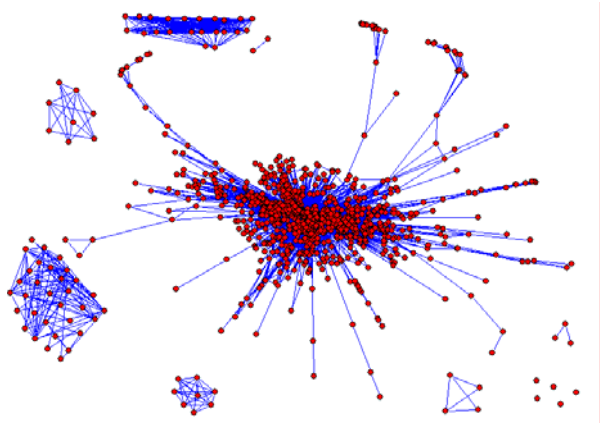


A General Framework for Weighted Gene Co-Expression Network Analysis

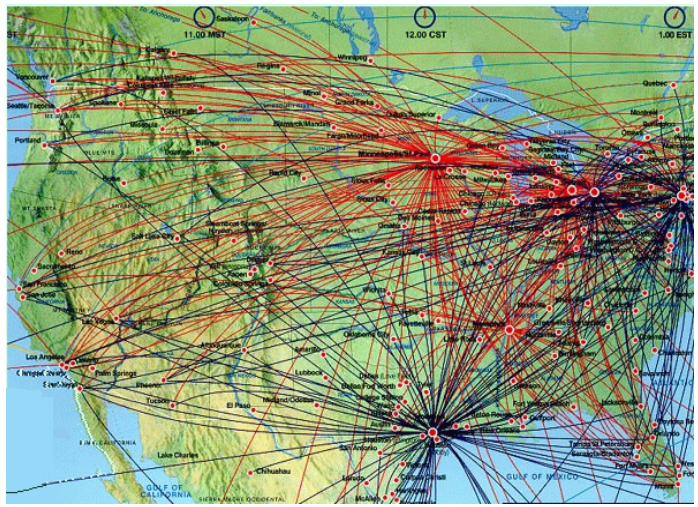
Steve Horvath
Human Genetics and Biostatistics
University of CA, LA

Content

- Novel statistical approach for analyzing microarray data:
  - weighted gene co-expression network analysis
- Empirical evidence that it matters in practice
  - Application 1: identifying cancer genes
  - Application 2: comparing chimp and human brain

Does this map tell you which cities are important?

This one does!



The nodes with the largest number of links (connections) are most important!

****Slide courtesy of Paul Mischel and AL Barabasi**

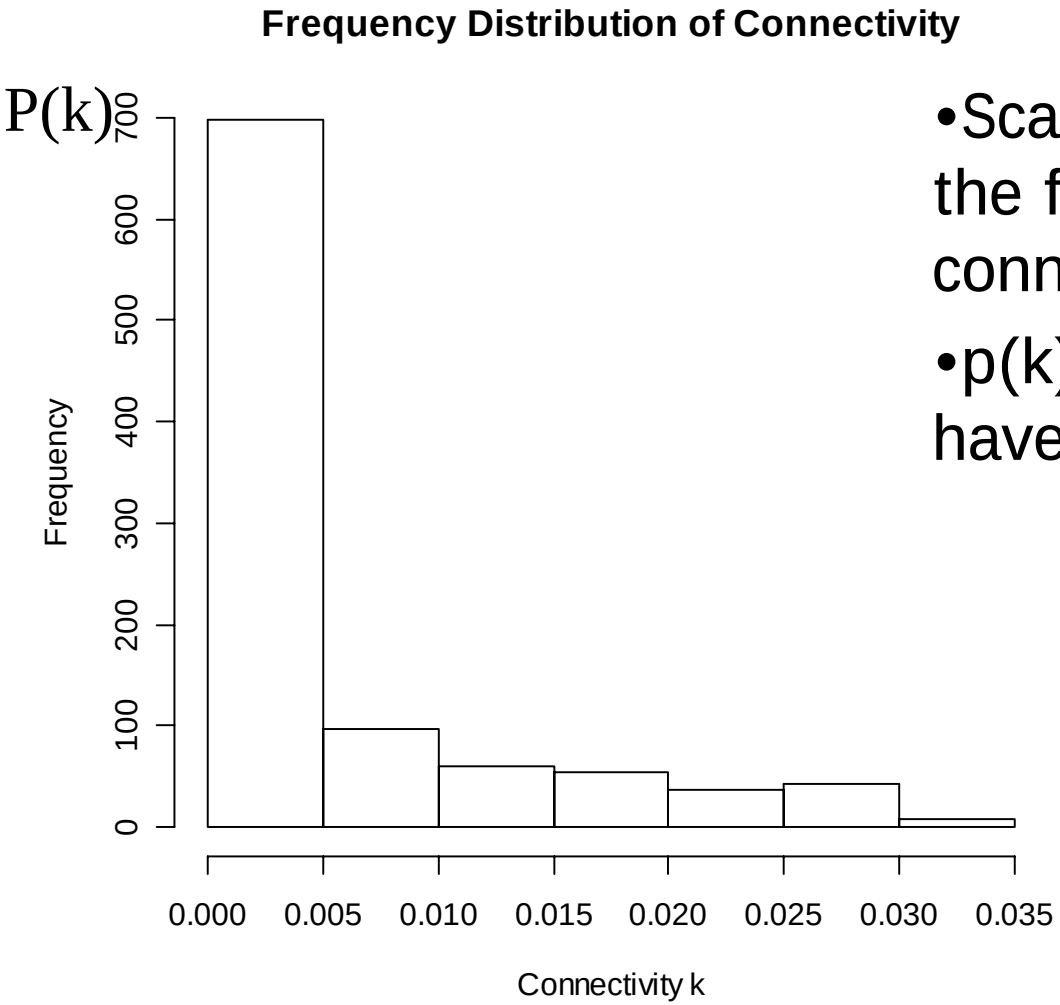
Background

- Network based methods have been found useful in many domains,
  - protein interaction networks
  - the world wide web
  - social interaction networks
  - OUR FOCUS: gene co-expression networks

Scale free topology is a fundamental property of such networks (Barabasi et al)

- It entails the presence of hub nodes that are connected to a large number of other nodes
- Such networks are robust with respect to the random deletion of nodes but are sensitive to the targeted attack on hub nodes
- It has been demonstrated that metabolic networks exhibit scale free topology at least approximately.

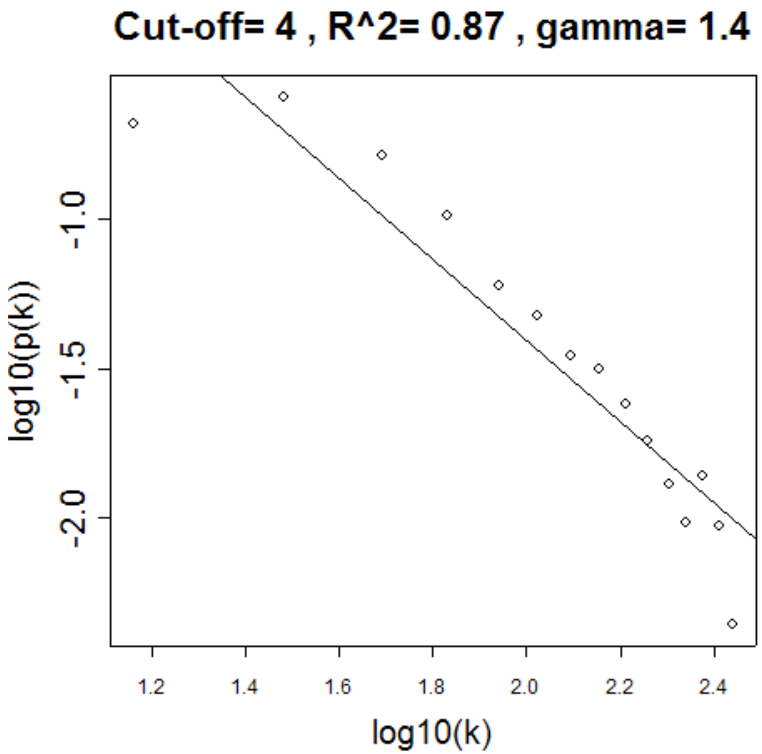
P(k) vs k in scale free networks



- Scale Free Topology refers to the frequency distribution of the connectivity k
- p(k)=proportion of nodes that have connectivity k

How to check Scale Free Topology?

Idea: Log transformation p(k) and k and look at scatter plots



Linear model fitting R^2 index can be used to quantify goodness of fit

Generalizing the notion of scale free topology

Motivation of generalizations: using weak general assumptions, we have proven that gene co-expression networks satisfy these distributions approximately.

Barabasi (1999)

$$ScaleFree \hat{=} \log(p(k)) = c_0 + c_1 \log(k)$$

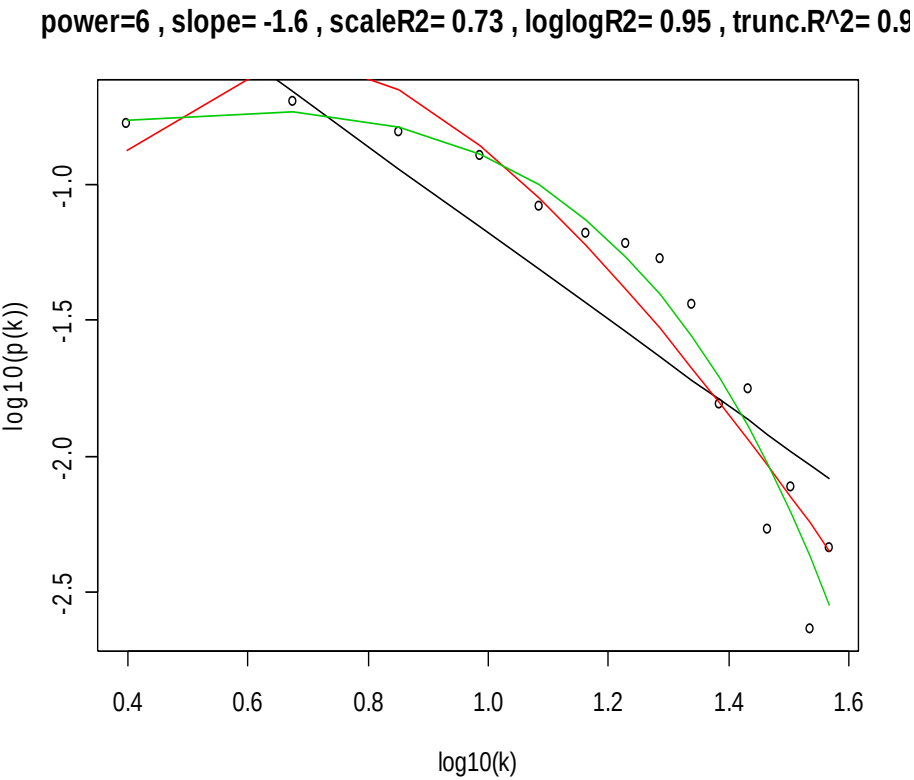
Csanyi-Szendroi (2004)

$$ExponentiallyTrunatedSFT \hat{=} \log(p(k)) = c_0 + c_1 \log(k) + c_2 k$$

Horvath, Dong (2005)

$$LogLogSFT \hat{=} \log(p(k)) = c_0 + c_1 \log(k) + c_2 \log(\log(k))$$

Checking Scale Free Topology in the Yeast Network

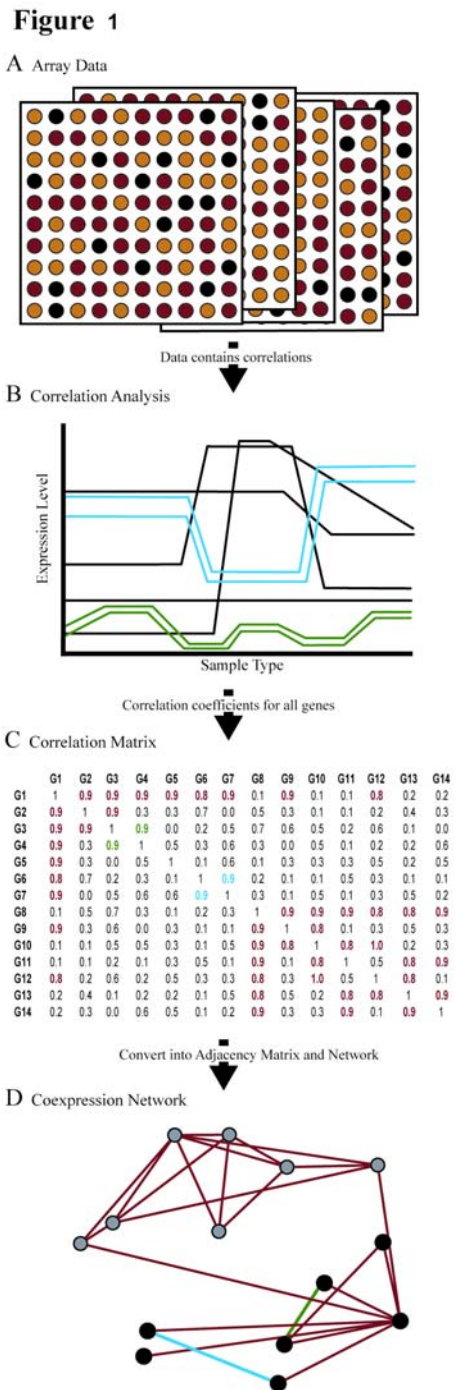


- Black=Scale Free
- Red=Exp. Truncated
- Green=Log Log SFT

How to define a gene co-expression network?

Gene Co-expression Networks

- In gene co-expression networks, each gene corresponds to a node.
- Two genes are connected by an edge if their expression values are highly correlated.
- Definition of “high” correlation is somewhat tricky
  - One can use statistical significance...
  - But we propose a criterion for picking threshold parameter: scale free topology criterion.



Steps for constructing a simple, unweighted co-expression network

A) Microarray gene expression data

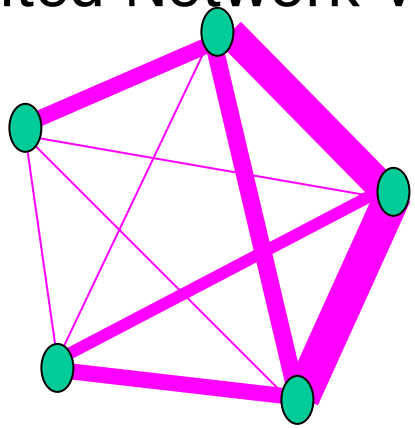
B) Measure concordance of gene expression with a Pearson correlation

C) The Pearson correlation matrix is dichotomized to arrive at an adjacency matrix. Binary values in the adjacency matrix correspond to an unweighted network.

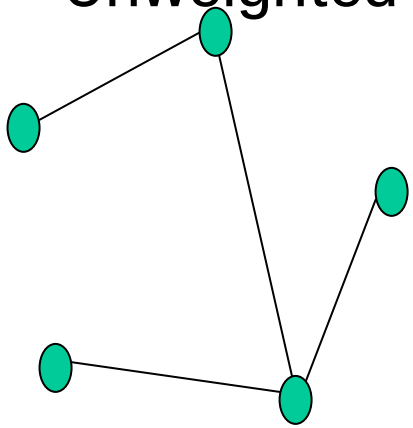
D) The adjacency matrix can be visualized by a graph.

Our `holistic' view....

Weighted Network View



Unweighted View



- All genes are connected
- Connection Widths=Connection strenghts

Some genes are connected
All connections are equal

Hard thresholding may lead to an information loss.

Mathematical Definition of an Undirected Network

Network=Adjacency Matrix

- A network can be represented by an adjacency matrix, $A=[a_{ij}]$, that encodes whether/how a pair of nodes is connected.
  - A is a symmetric matrix with entries in [0,1]
  - For unweighted network, entries are 1 or 0 depending on whether or not 2 nodes are adjacent (connected)
  - For weighted networks, the adjacency matrix reports the connection strength between gene pairs

Generalized Connectivity

- Gene connectivity = row sum of the adjacency matrix
  - For unweighted networks=number of direct neighbors
  - For weighted networks= sum of connection strengths to other nodes

$$k_i = \sum_j a_{ij}$$

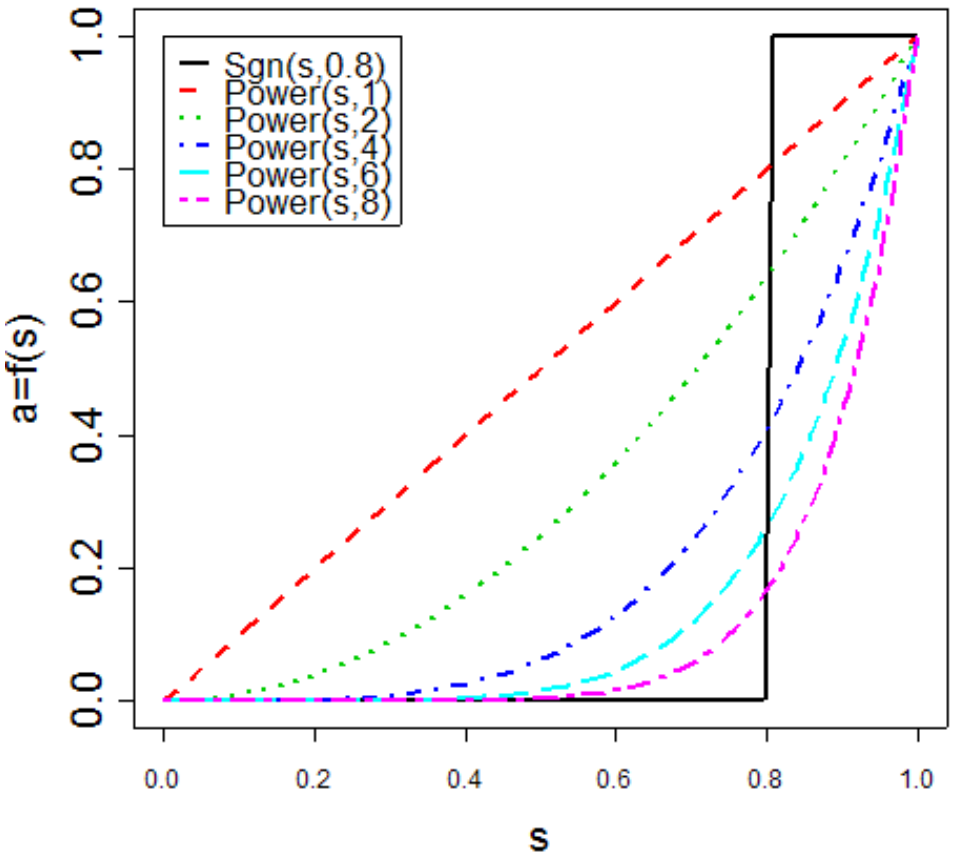
How to construct a weighted gene co-expression network?

Using an adjacency function to define a network

- Measure co-expression by a similarity s(i,j) in [0,1] e.g. absolute value of the Pearson correlation
- Define an adjacency matrix as A(i,j) using an adjacency function AF(s(i,j))
- AF is a monotonic function from [0,1] onto [0,1]
- Here we consider 2 classes of AFs
  - Step function AF(s)=I(s>tau) with parameter tau (unweighted network)
  - Power function $AF(s)=s^b$ with parameter b
- The choice of the AF parameters (tau, b) determines the properties of the network.

Comparing the power adjacency functions with the step function

Adjacency
=connection strength



Gene Co-expression Similarity

The scale free topology criterion for choosing the parameter values of an adjacency function.

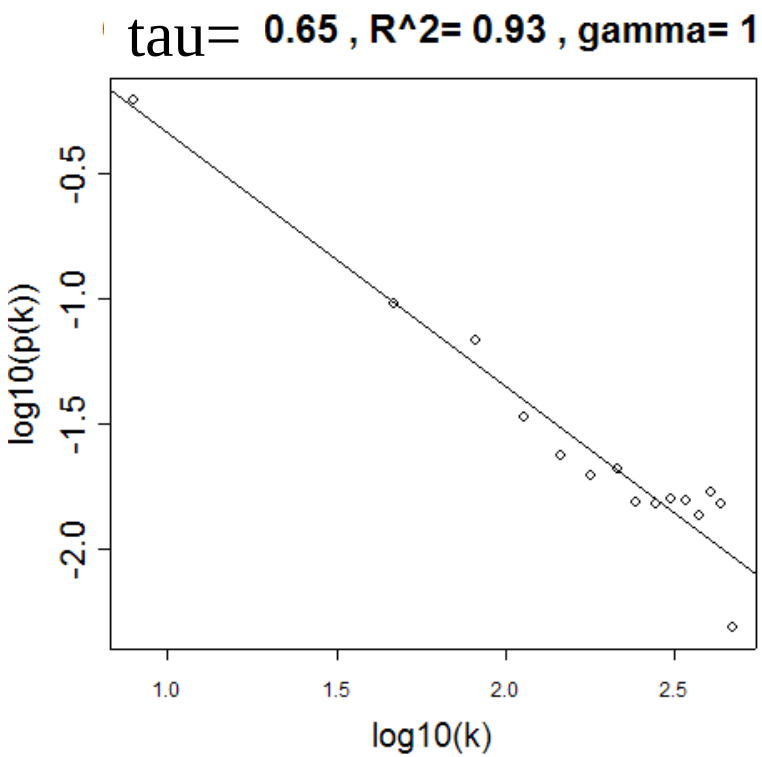
A) CONSIDER ONLY THOSE PARAMETER VALUES THAT RESULT IN APPROXIMATE SCALE FREE TOPOLOGY

B) SELECT THE PARAMETERS THAT RESULT IN THE HIGHEST MEAN NUMBER OF CONNECTIONS

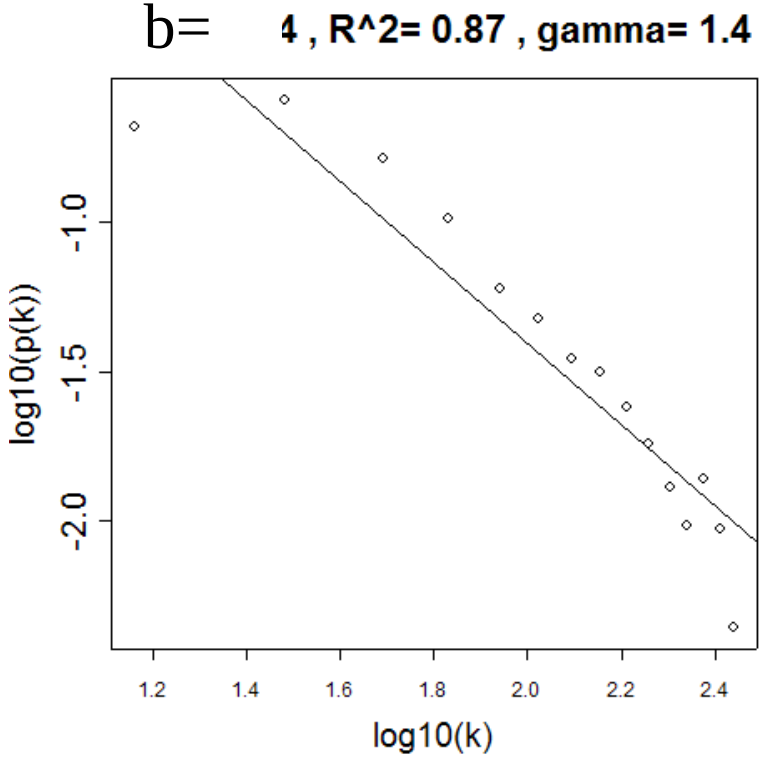
- Criterion A is motivated by the finding that most metabolic networks (including gene co-expression networks, protein-protein interaction networks and cellular networks) have been found to exhibit a scale free topology
- Criterion B leads to high power for detecting modules (clusters of genes) and hub genes.

Criterion A is measured by the linear model fitting index R^2

Step AF (tau)



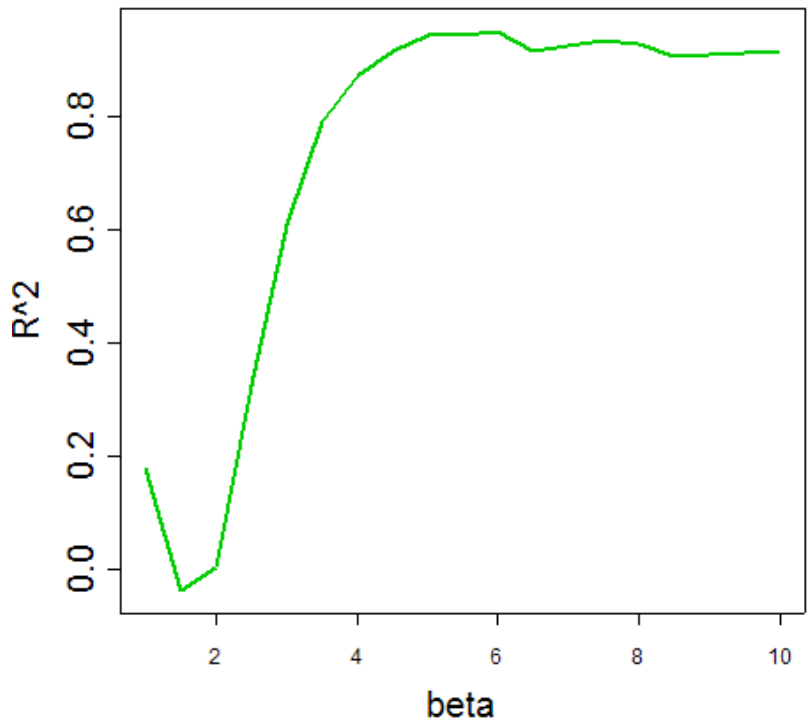
Power AF (b)



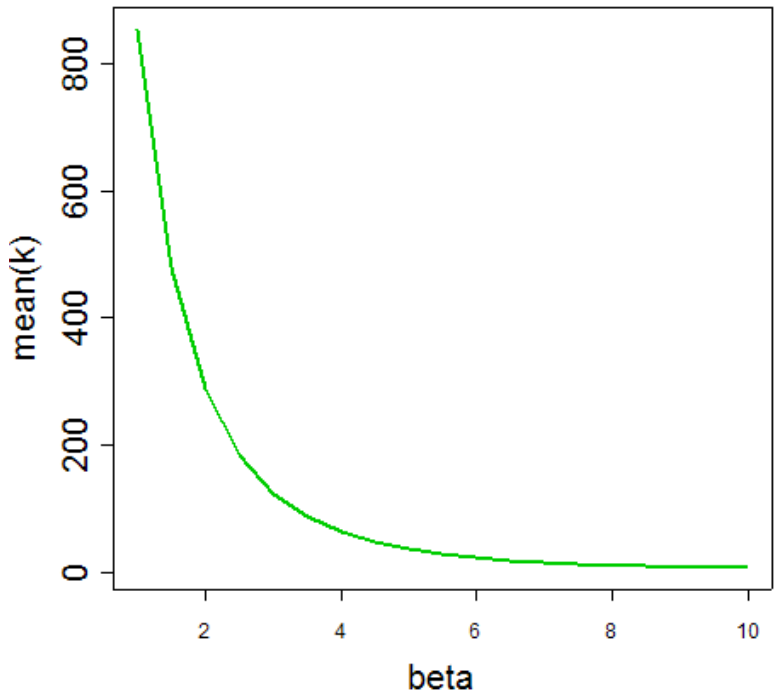
Trade-off between criterion A (R^2) and criterion B (mean no. of connections) when varying the power b

Power $AF(s)=s^b$

criterion A: SFT model fit R^2



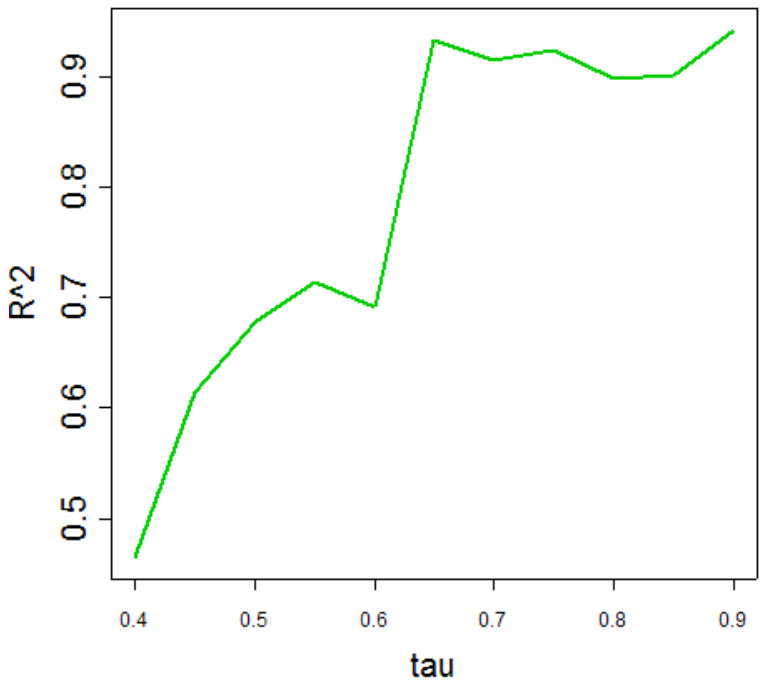
criterion B: mean connectivity



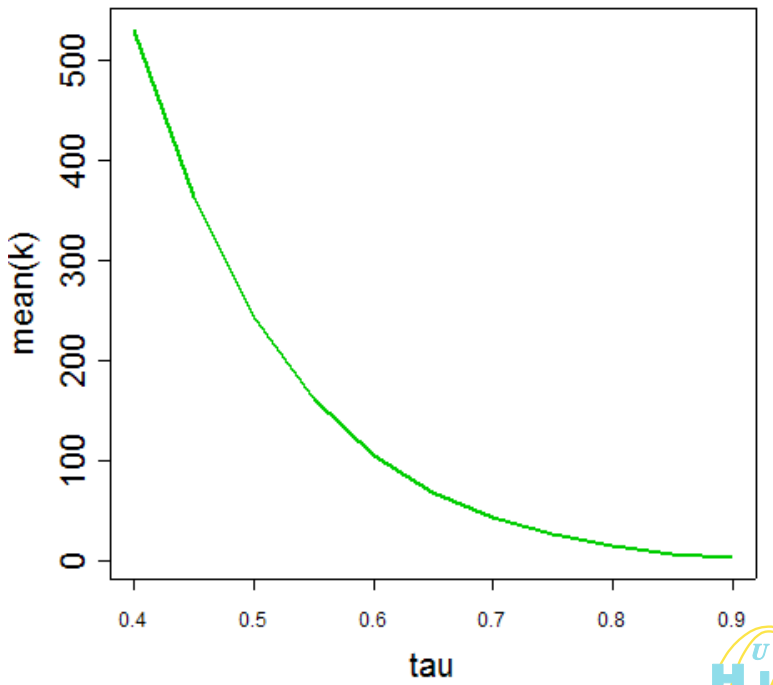
Trade-off between criterion A and B when varying tau

Step Function: I(s>tau)

criterion A

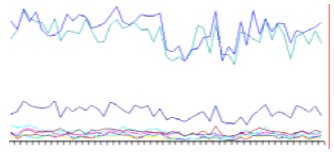


criterion B

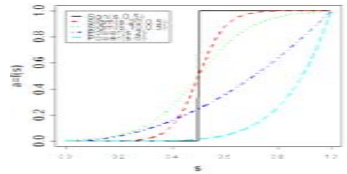


General Framework for Network Analysis

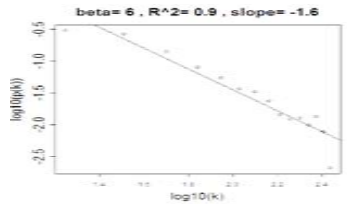
Define a Gene Co-expression Similarity



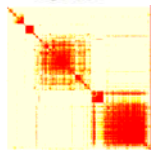
Define a Family of Adjacency Functions



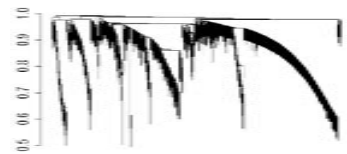
Determine the AF Parameters



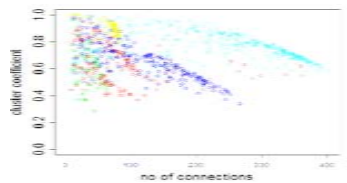
Define a Measure of Node Dissimilarity



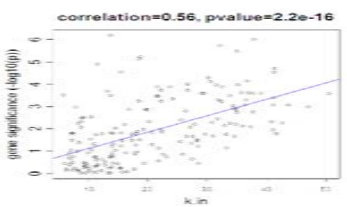
Identify Network Modules (Clustering)



Relate Network Concepts to Each Other



Relate the Network Concepts to External Gene or Sample Information



How to measure distance in a network?

- Mathematical Answer: Geodesics
  - length of shortest path connecting 2 nodes
- Biological Answer: look at shared neighbors
  - Intuition: if 2 people share the same friends they are close in a social network
  - Use the topological overlap measure based distance proposed by Ravasz et al (2002)

Topological Overlap leads to a network distance measure (Ravasz et al 2002)

$$TOM_{ij} = \frac{\sum_u a_{iu} a_{uj} + a_{ij}}{\min(k_i, k_j) + 1 - a_{ij}}$$

$$DistTOM_{ij} = 1 - TOM_{ij}$$

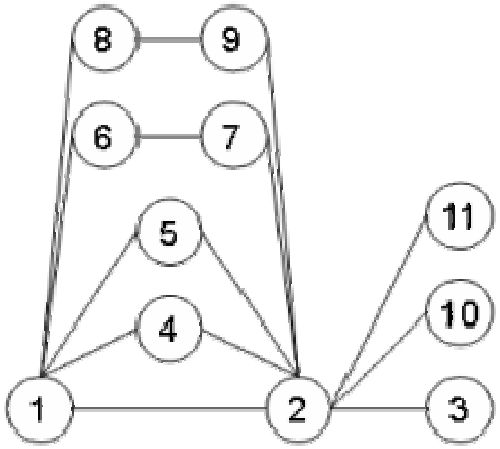
- Generalized in Zhang and Horvath (2005) to the case of weighted networks

The Generalized Topological Overlap

Andy Yip, S.Horvath

The general topological overlap matrix

a.



b.

$N_m(i)$	$i=1$	$i=2$	$i=3$
$m=1$	2,4,5,6,8	1,3,4,5,7,9,10,11	2
$m=2$	2,3,4,5,6,7,8,9,10,11	1,3,4,5,6,7,8,9,10,11	1,2,4,5,7,9,10,11

c.

$t_{ij}^{[m]}$	$(i,j)=(1,2)$	$(i,j)=(1,3)$	$(i,j)=(2,3)$
$m=0$	1	0	1
$m=1$	3/5	1/2	1
$m=2$	1	7/9	1

$$TOM(i,j)=\frac{|N_1(i)\cap N_1(j)|+a_{ij}}{\min(|N_1(i)|,|N_1(j)|)+1-a_{ij}}$$

$N_1(i)$ denotes the set of neighbors of node i

|*| measures the cardinality

We have re-interpreted the TOM measure as the “normalized” proportion of genes that are in both node neighborhoods. This allows for a straightforward generalization to “larger” neighborhoods.

Yip, Horvath (2005)

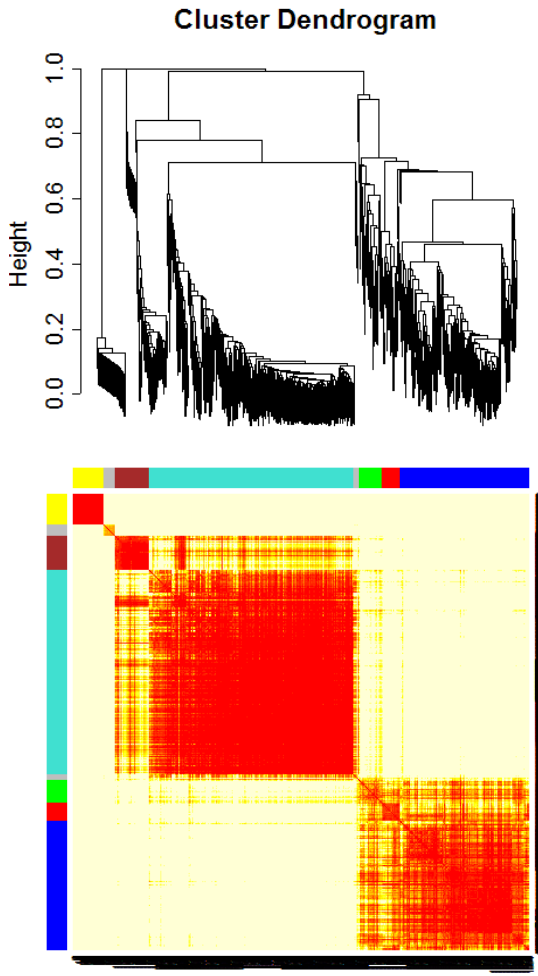
Defining Gene Modules
=sets of tightly co-regulated genes

Module Identification based on the notion of topological overlap

- One important aim of metabolic network analysis is to detect subsets of nodes (modules) that are tightly connected to each other.
- We adopt the definition of Ravasz et al (2002): modules are groups of nodes that have high topological overlap.

Steps for defining gene modules

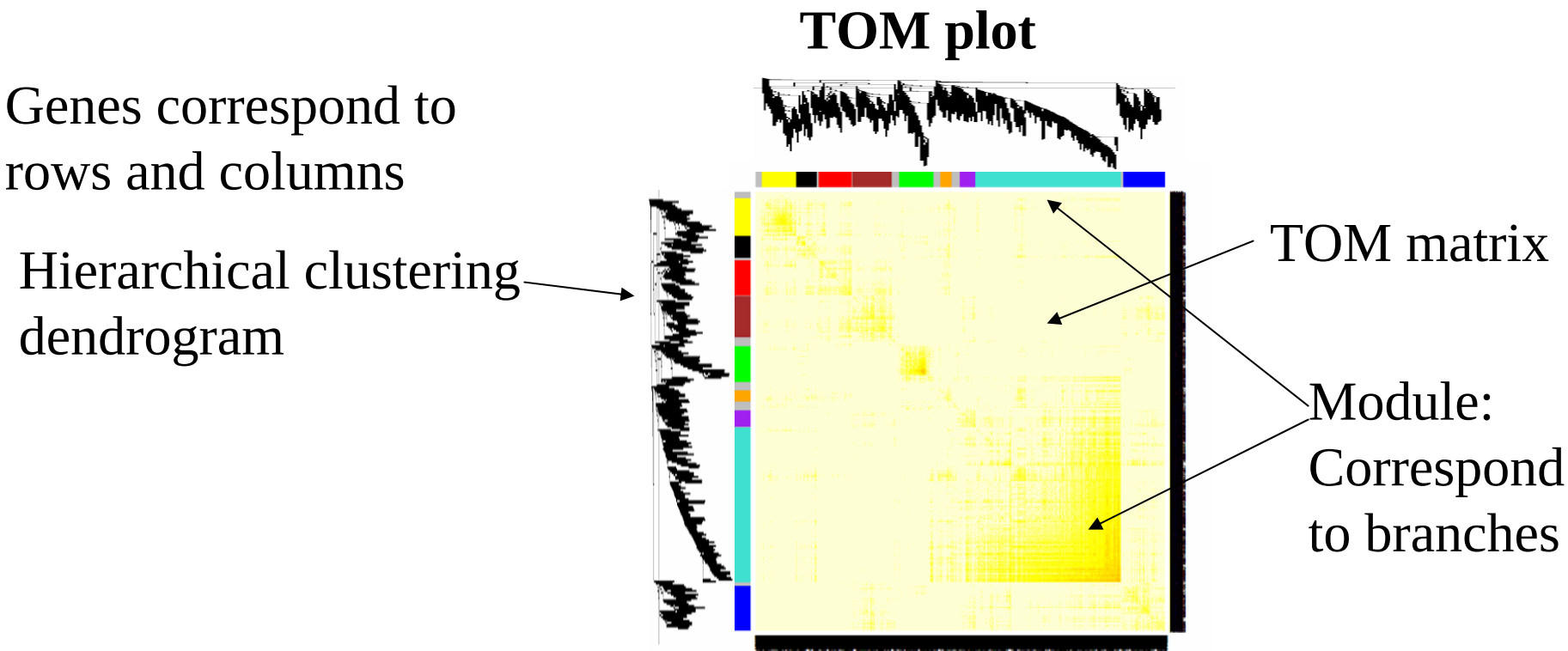
- Define a dissimilarity measure between the genes.
  - Standard Choice: dissim(i,j)=1-abs(correlation)
  - Choice by network community=1-Topological Overlap Matrix (TOM)
    - Used here
- Use the dissimilarity in hierarchical clustering
- Define modules as branches of the hierarchical clustering tree
- Visualize the modules and the clustering results in a heatmap plot



Heatmap

Using the TOM matrix to cluster genes

- To group nodes with high topological overlap into modules (clusters), we typically use average linkage hierarchical clustering coupled with the TOM distance measure.
- Once a dendrogram is obtained from a hierarchical clustering method, we choose a height cutoff to arrive at a clustering.
  - Here modules correspond to branches of the dendrogram

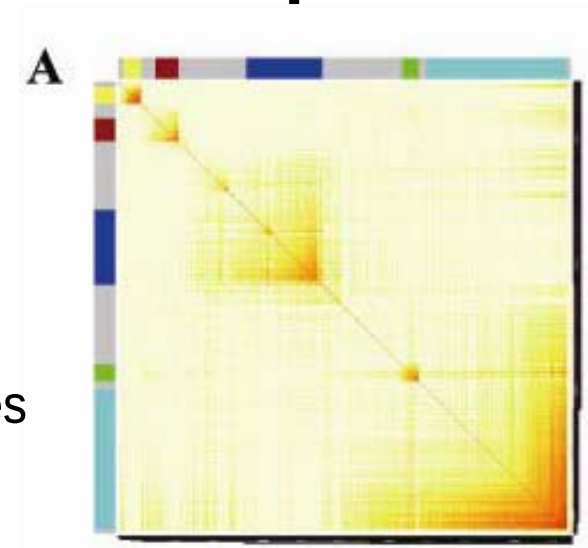


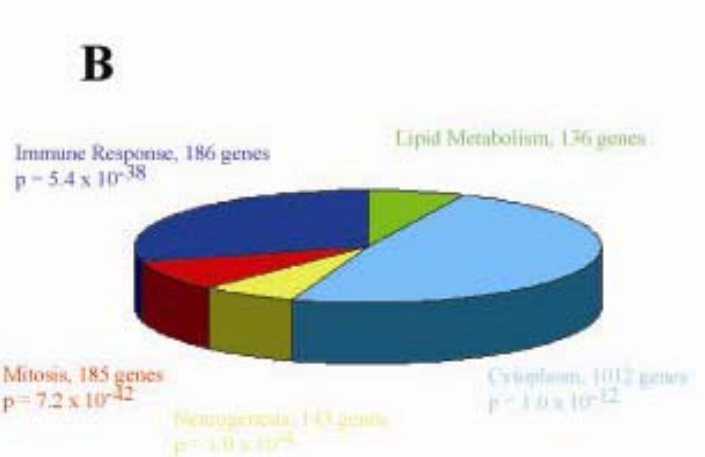
Different Ways of Depicting Gene Modules

Topological Overlap Plot

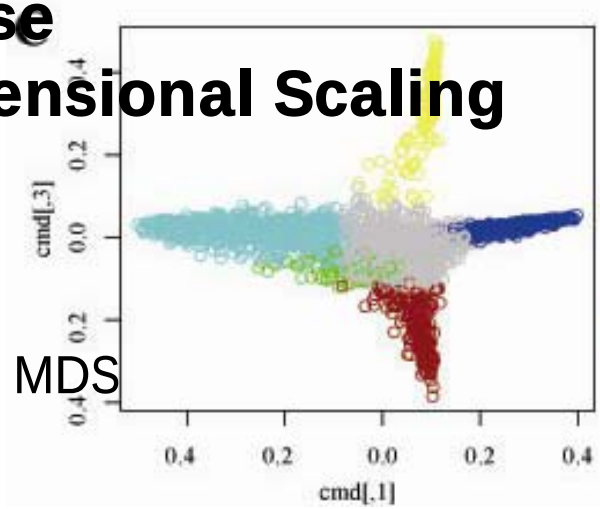
Gene Functions

1) Rows and columns correspond to genes
2) Red boxes along diagonal are modules
3) Color bands=modules

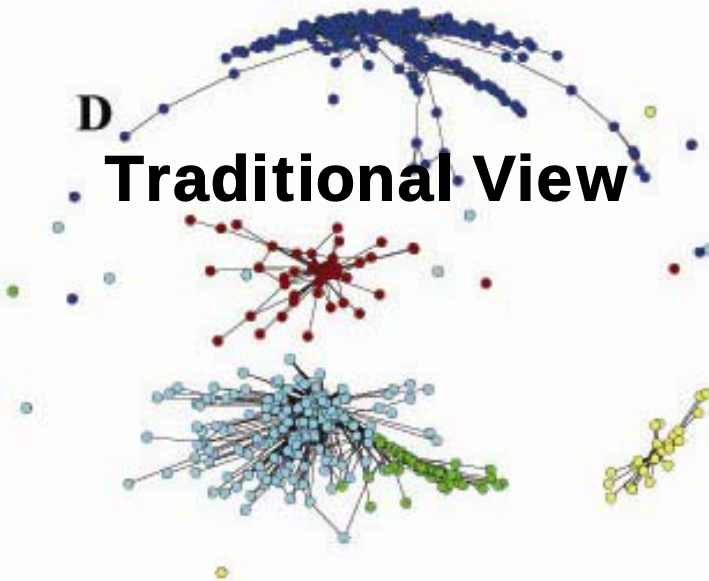




We propose Multi Dimensional Scaling



Traditional View

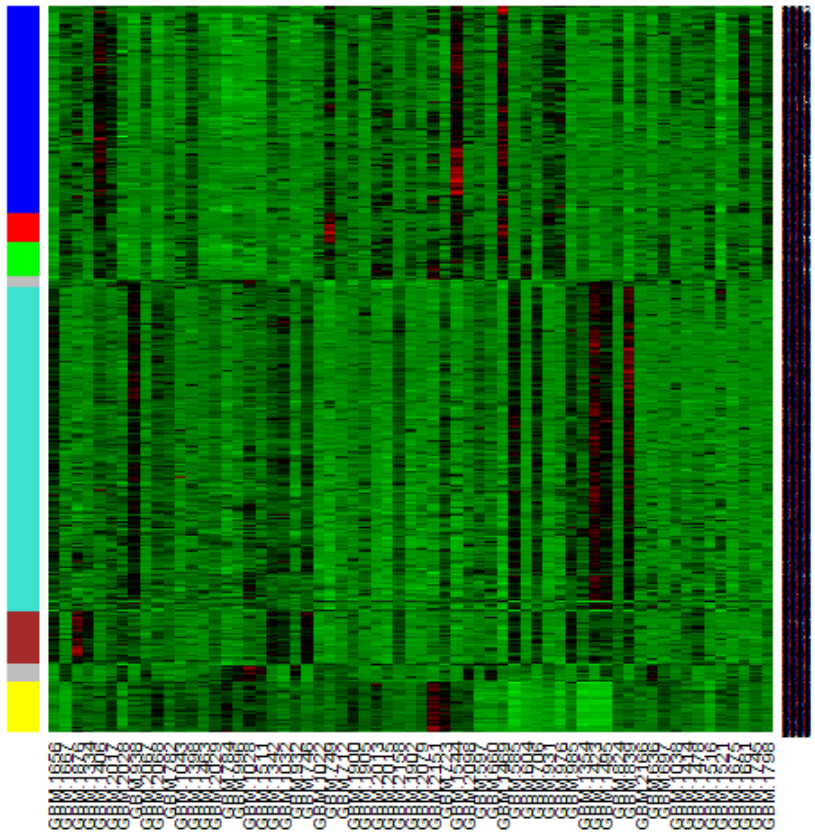


Idea:
Use network distance in MDS

More traditional view of module

Columns=Brain tissue samples

Rows=Genes
Color band indicates module membership

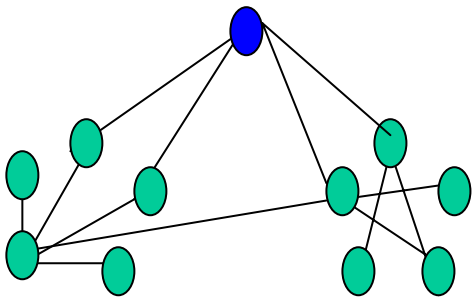


Message: characteristic vertical bands indicate tight co-expression of module genes

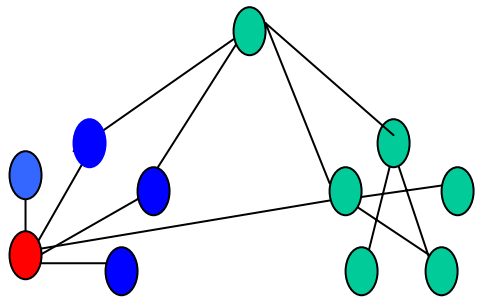
Module-Centric View of Networks

Our team (module)-centric view v.s. traditional prima-donna (hub) centric view

- Traditional view based on whole network connectivity



- Module view based on within module connectivity



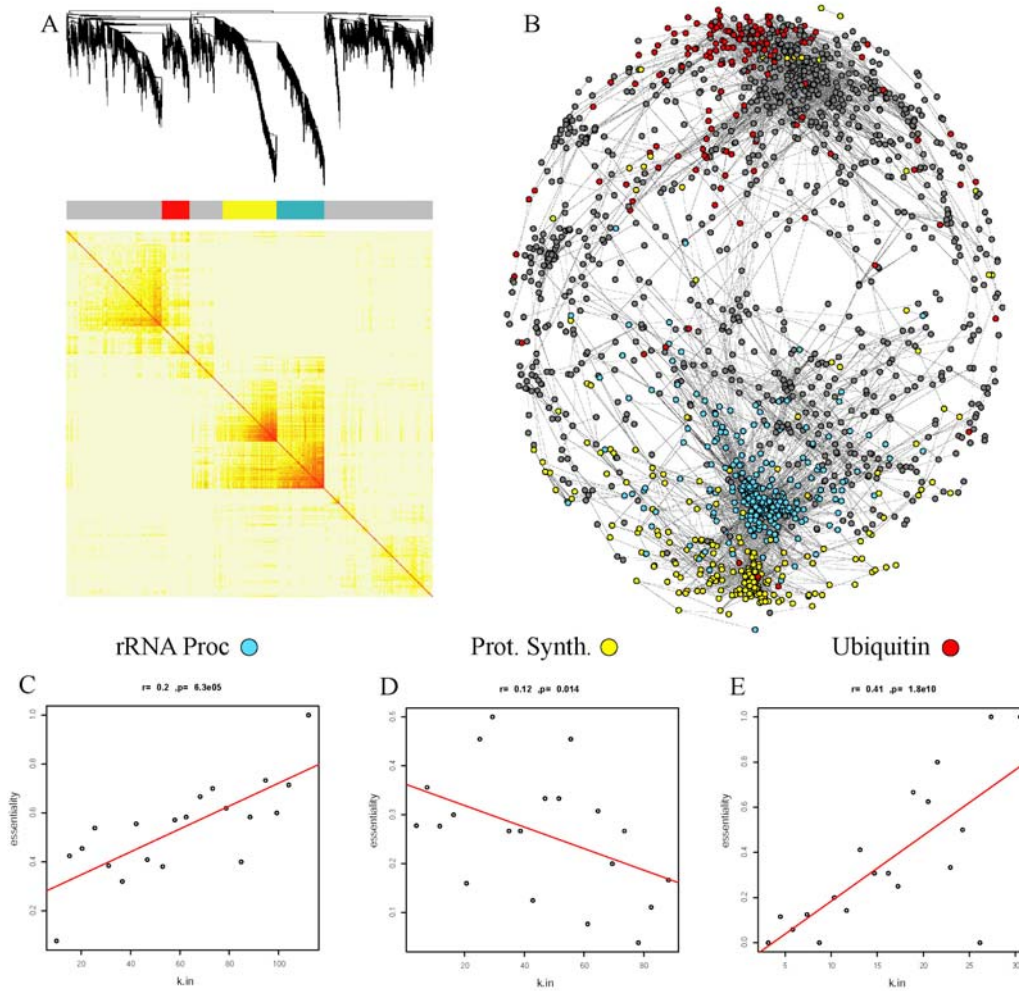
In many applications, we find that intramodular connectivity is biologically and mathematically more meaningful than whole network connectivity

Mathematical Facts (Horvath, Dong, Yip 2005)
Hub genes are always module genes in co-expression networks.
Most module genes have high connectivity.

Yeast Data Analysis
Marc Carlson

Within Module Analysis

Findings
1) The intramodular connectivities are related to how essential a gene is for yeast survival
2) Modules are highly preserved across different data sets
3) Hub genes are highly preserved across species



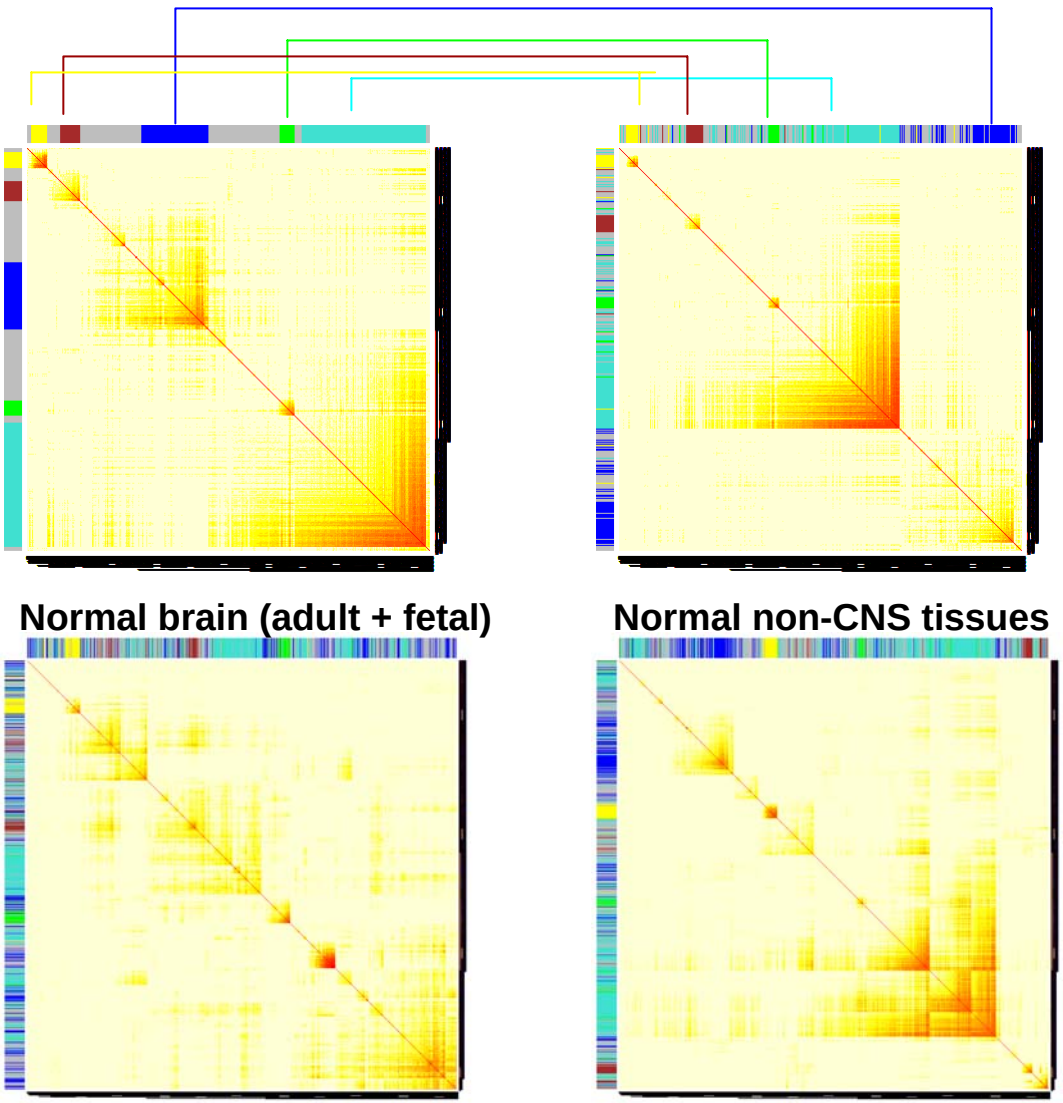
Prob(Essential)

Connectivity k

Hub Genes Predict Survival for Brain Cancer Patients

Mischel PS, Zhang B,et al, Horvath S, Nelson SF.

Module structure is highly preserved across data sets



Messages:
1) Cancer modules can be independently validated
2) Modules in brain cancer tissue can also be found in normal, non-brain tissue.
-->
Insights into the biology of cancer

Gene prognostic significance

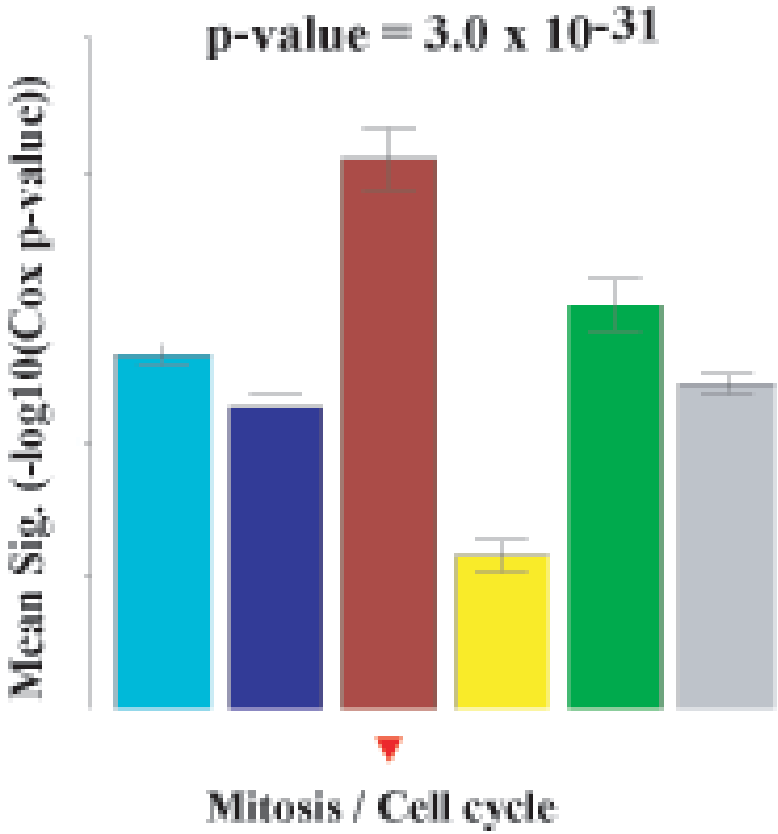
Definition

1) Regress survival time on gene expression information using a univariable Cox regression model
2) Obtain the score test p-value
3) Gene significance=-log10(p-value)
   - Roughly speaking
     Gene significance~no of zeroes in the p-value.

Goal

Relate gene significance to intramodular connectivity

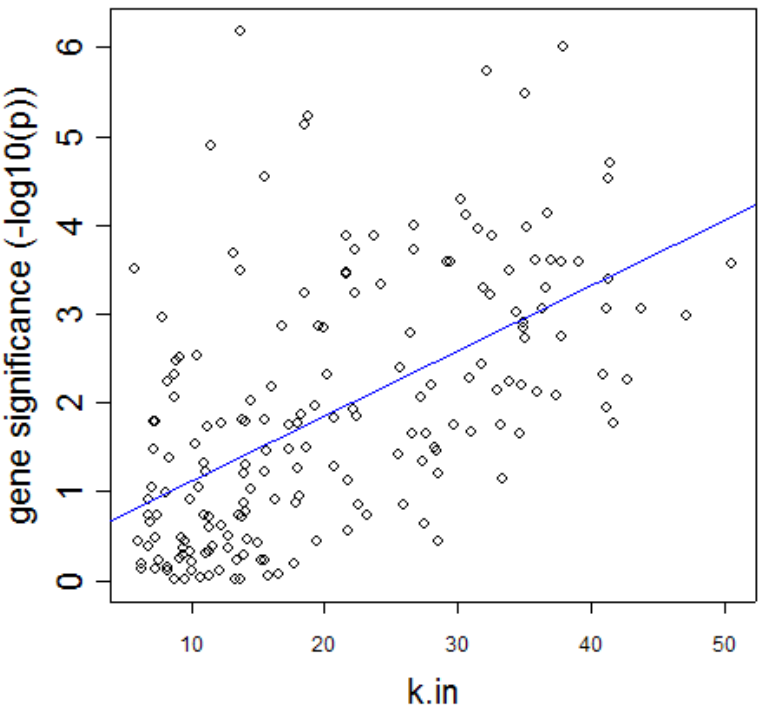
Mean Prognostic Significance of Module Genes



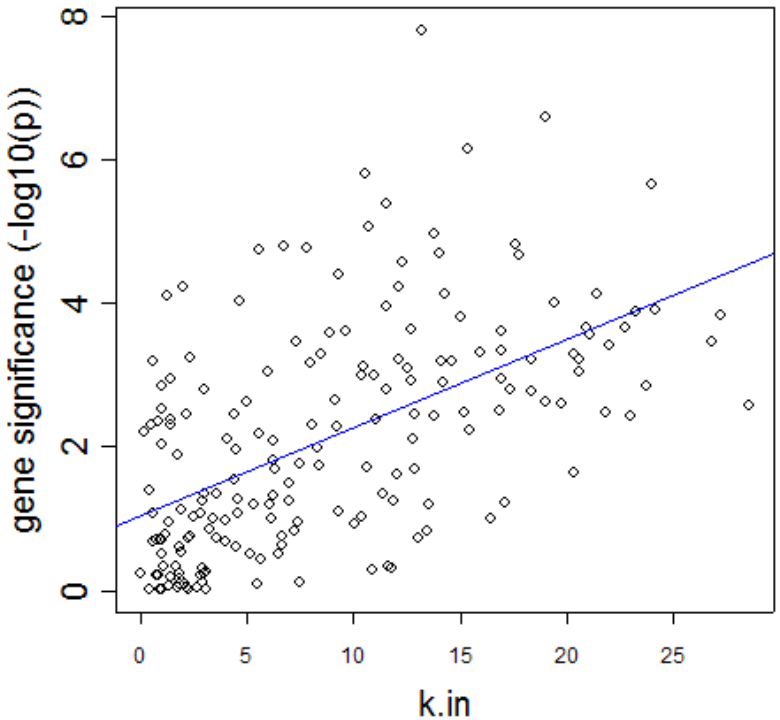
Message: Focus the attention on the brown module genes

Module hub genes predict cancer survival

1. **Intramodular connectivity is highly correlated with gene significance**
2. **Recall prognostic significance as –log10(Cox-p-value)**



Test set: 55 samples
r = 0.56; p-2.2 x 10^{-16}



Validation set: 65 samples
r = 0.55; p-2.2 x 10^{-16}

The fact that genes with high intramodular connectivity are more likely to be prognostically significant facilitates a novel screening strategy for finding prognostic genes

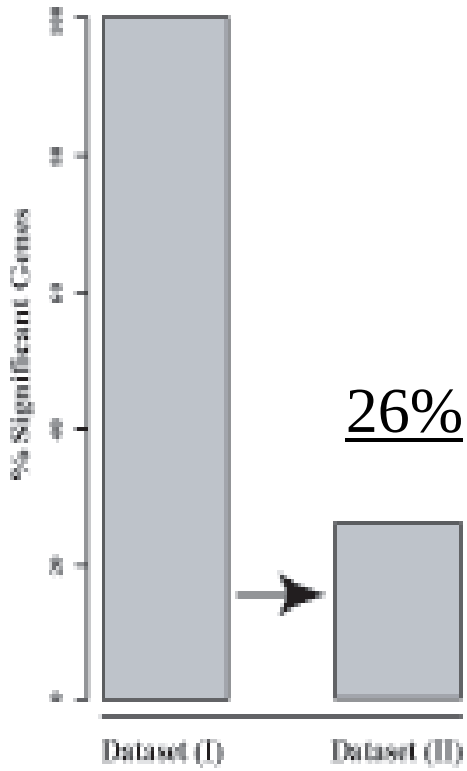
- Focus on those genes with significant Cox regression p-value and high intramodular connectivity.
  - It is essential to to take a module centric view: focus on intramodular connectivity of module that is enriched with significant genes.

Gene screening strategy that makes use of intramodular connectivity is far superior to standard approach

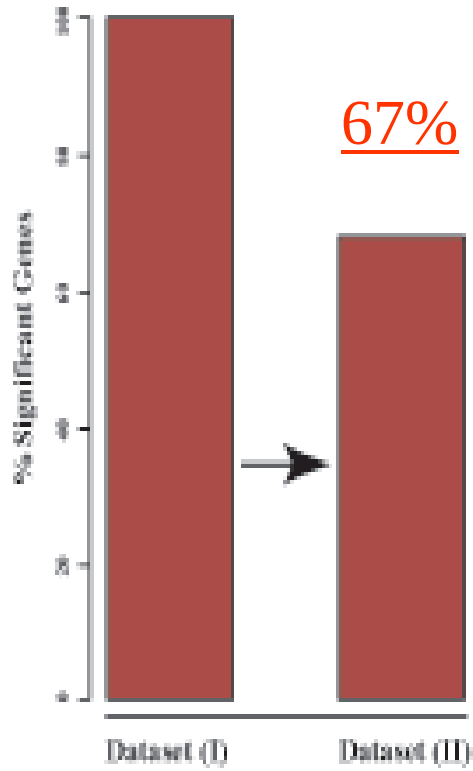
- Validation success rate= proportion of genes with independent test set Cox regression p-value<0.05.
- Validation success rate of network based screening approach (68%)
- Standard approach involving top 300 most significant genes: 26%

Validation success rate of gene expressions in independent data

300 most significant genes
(Cox p-value<$1.3*10^{-3}$)



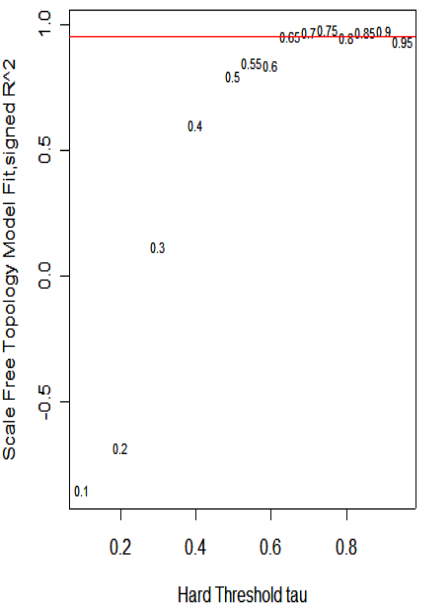
Network based screening
p<0.05 and
high intramodular connectivity

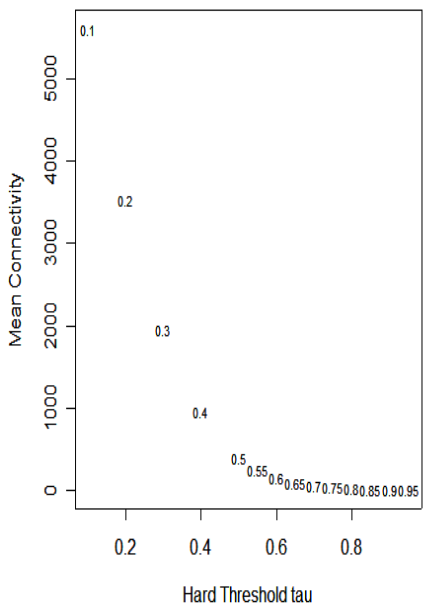


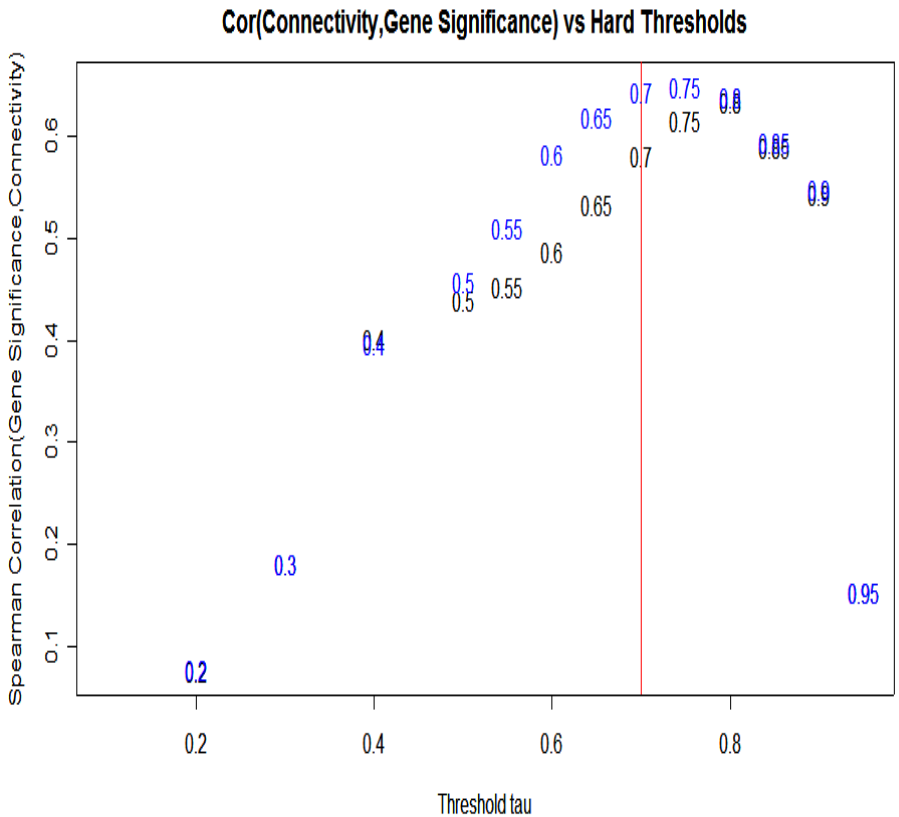
The biological signal is much more robust in weighted than in unweighted networks.

- Biological signal = Spearman correlation between brown intramodular connectivity and prognostic significance,
  - Biological Signal=cor(Gene Signif ,K)
- Robustness analysis
  - Explore how this biological signal changes as a function of the adjacency function parameters tau (hard thresholding) and b (=power=soft thresholding).

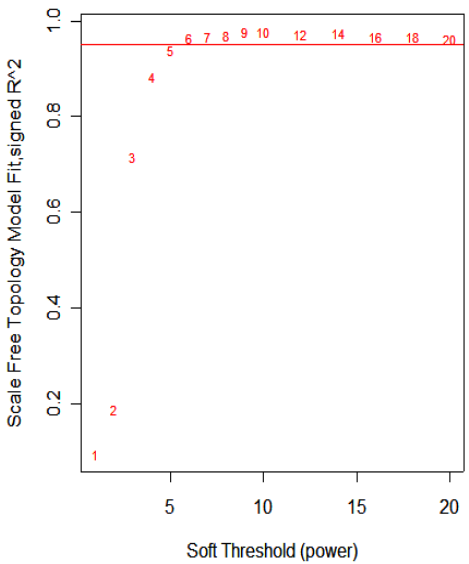
Scale Free Topology fitting index and biological signals for different hard thresholds

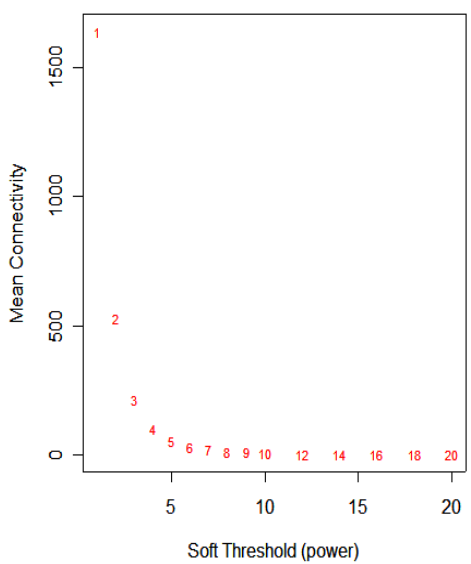


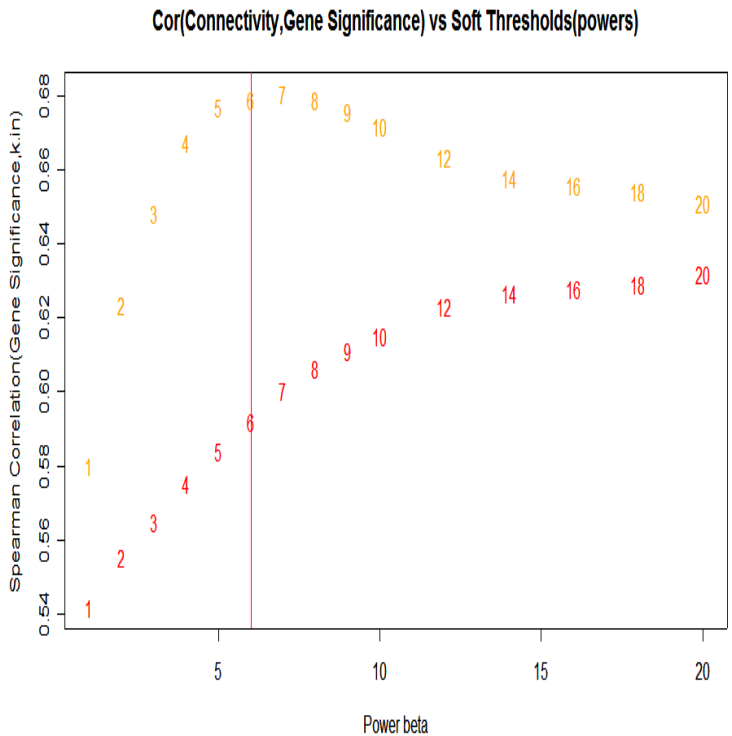




Scale Free Topology fitting index and biological signals for different SOFT thresholds (powers)







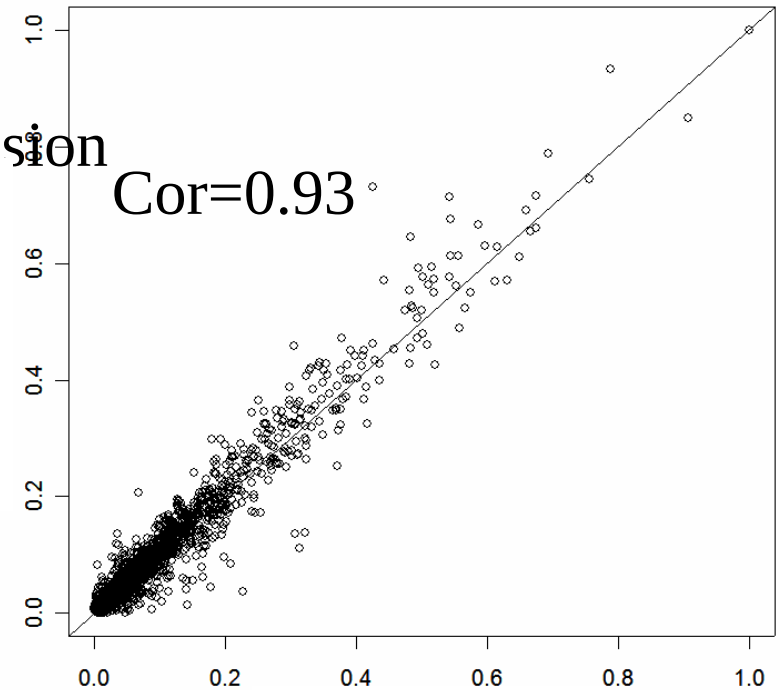
Soft thresholding leads to more robust results

- The results of soft thresholding are highly robust with respect to the choice of the adjacency function parameter, i.e. the power b
- In contrast, the results of hard thresholding are sensitive to the choice of tau
- In this application, the biological signal peaks close to the adjacency function parameter that was chosen by the scale free topology criterion.

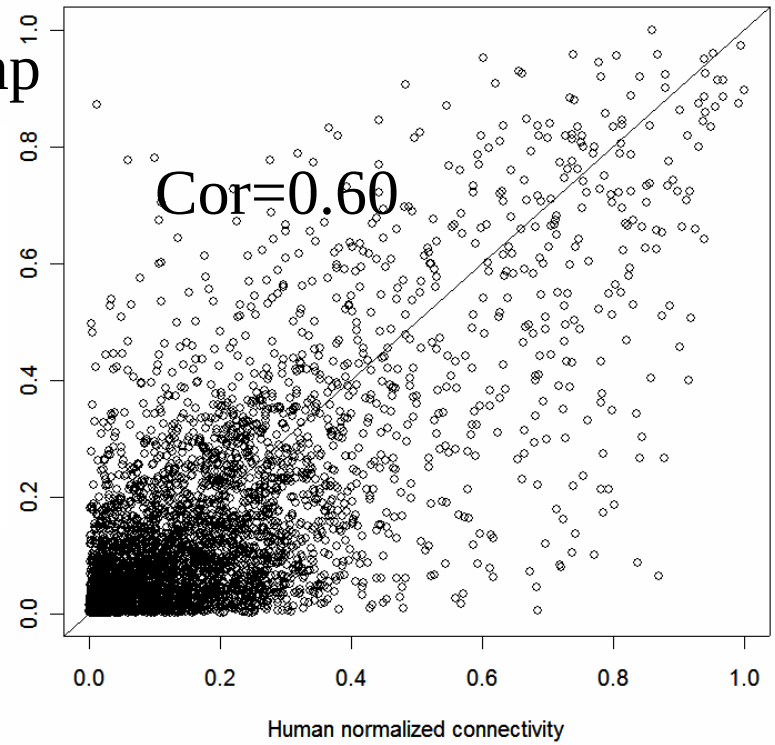
Application II
Comparing Human and Chimp Brains

Mike Oldham, S.Horvath, Dan Geschwind D

Gene expression is more strongly preserved than gene connectivity



Human Expression

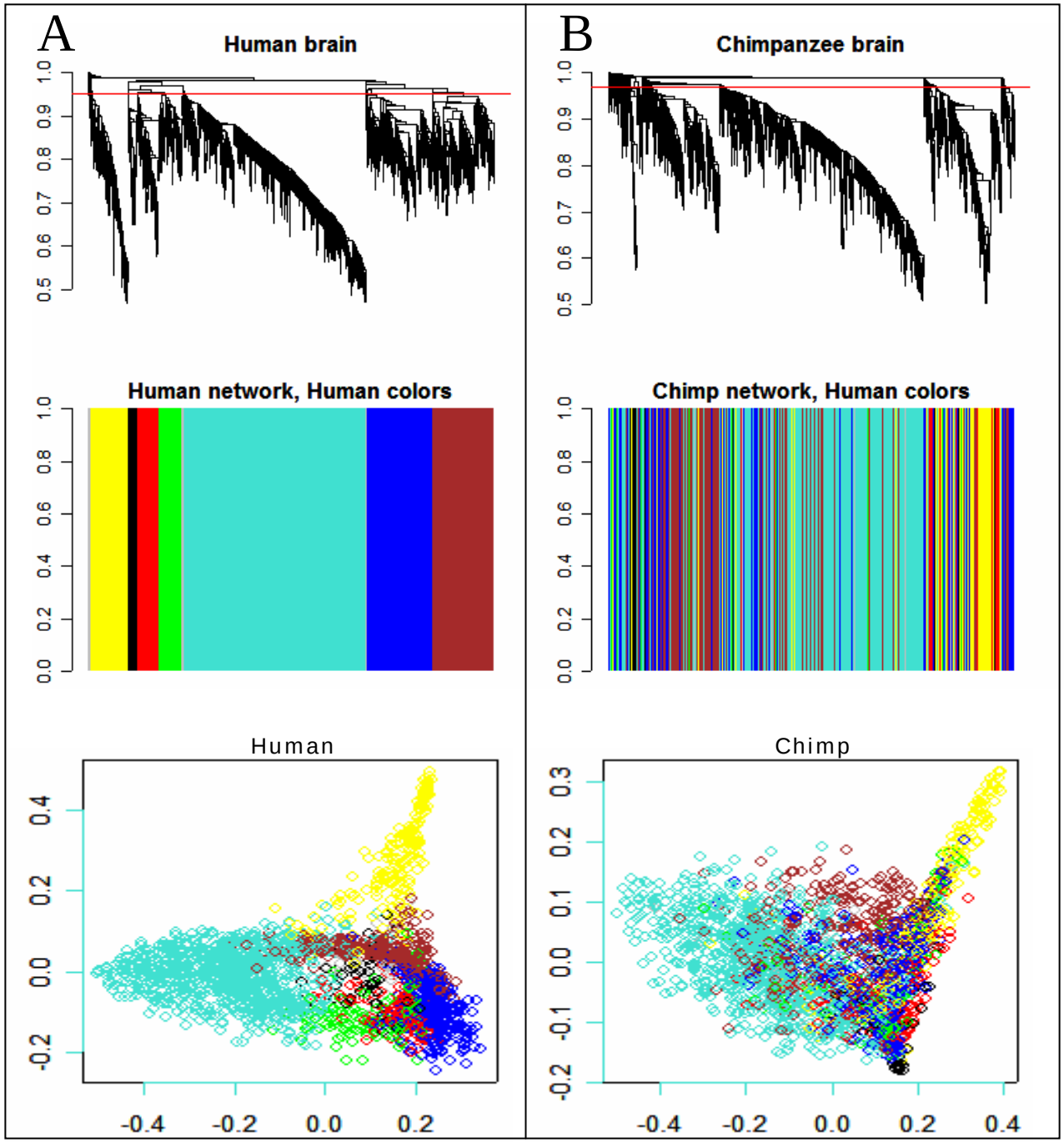


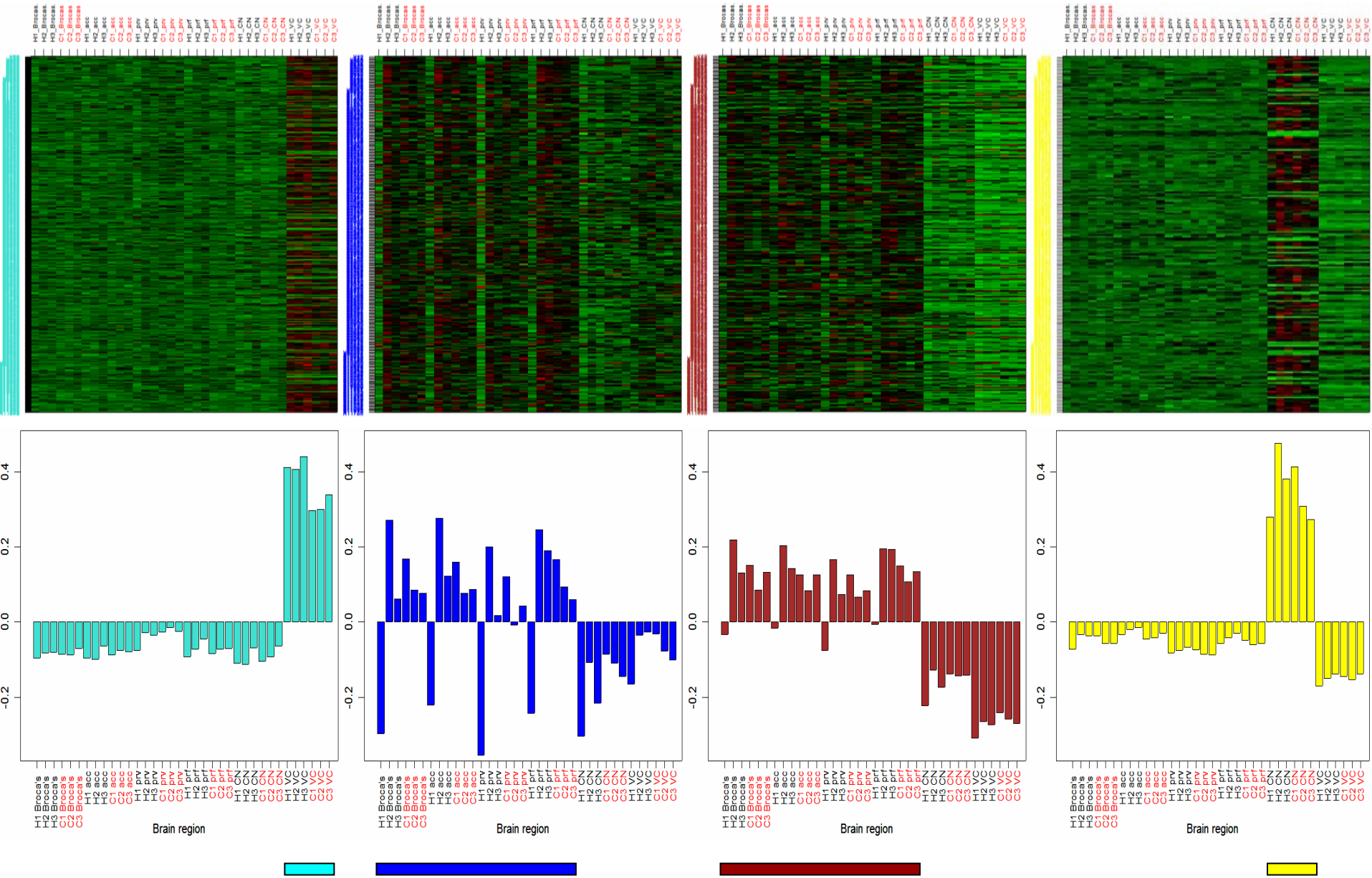
Human Connectivity

Hypothesis: molecular wiring makes us human

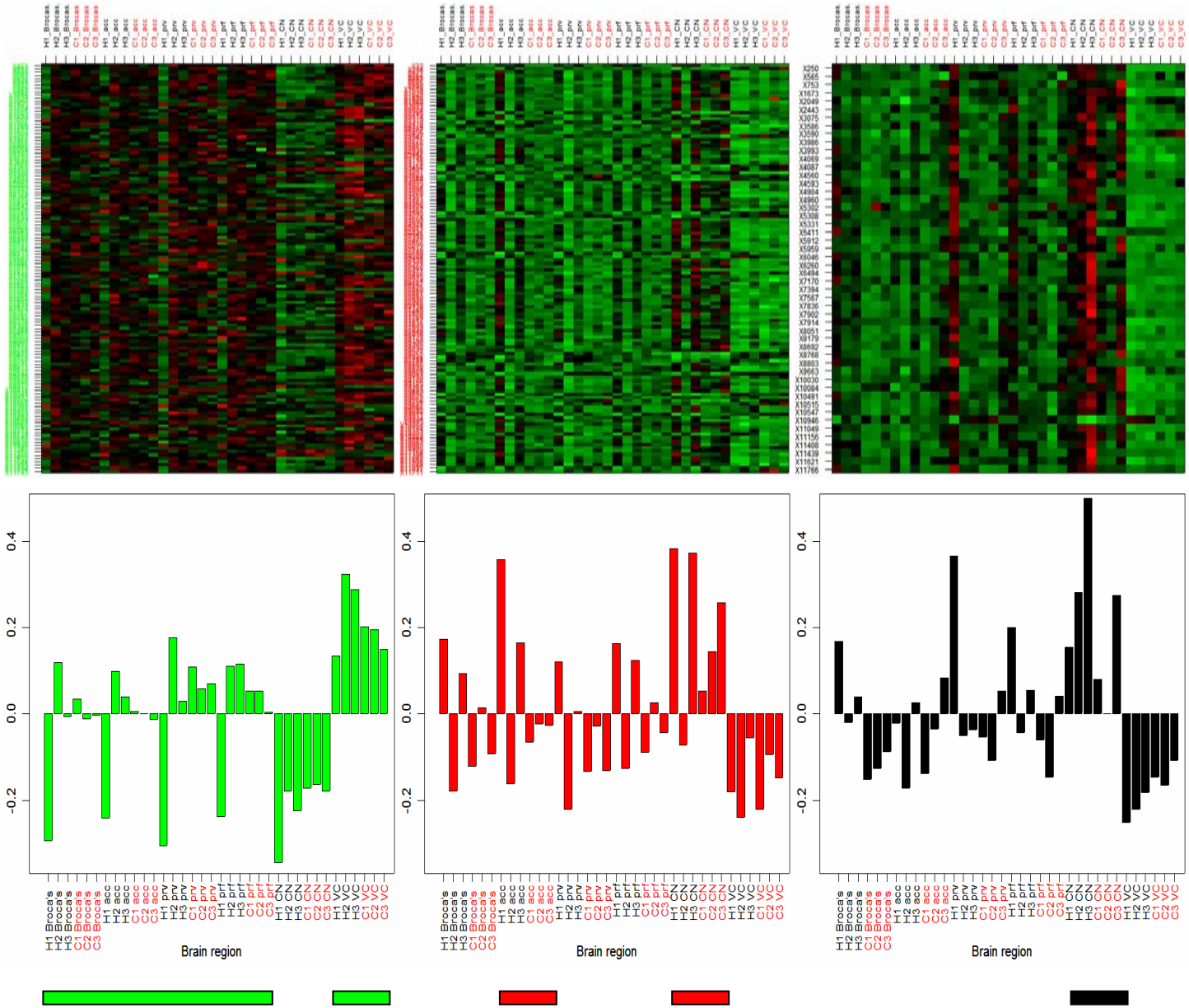
Raw data from Khaitovich *et al.*, 2004

Mike Oldham


A
Human brain
B
Chimpanzee brain
Human network, Human colors
Chimp network, Human colors
Human
Chimp

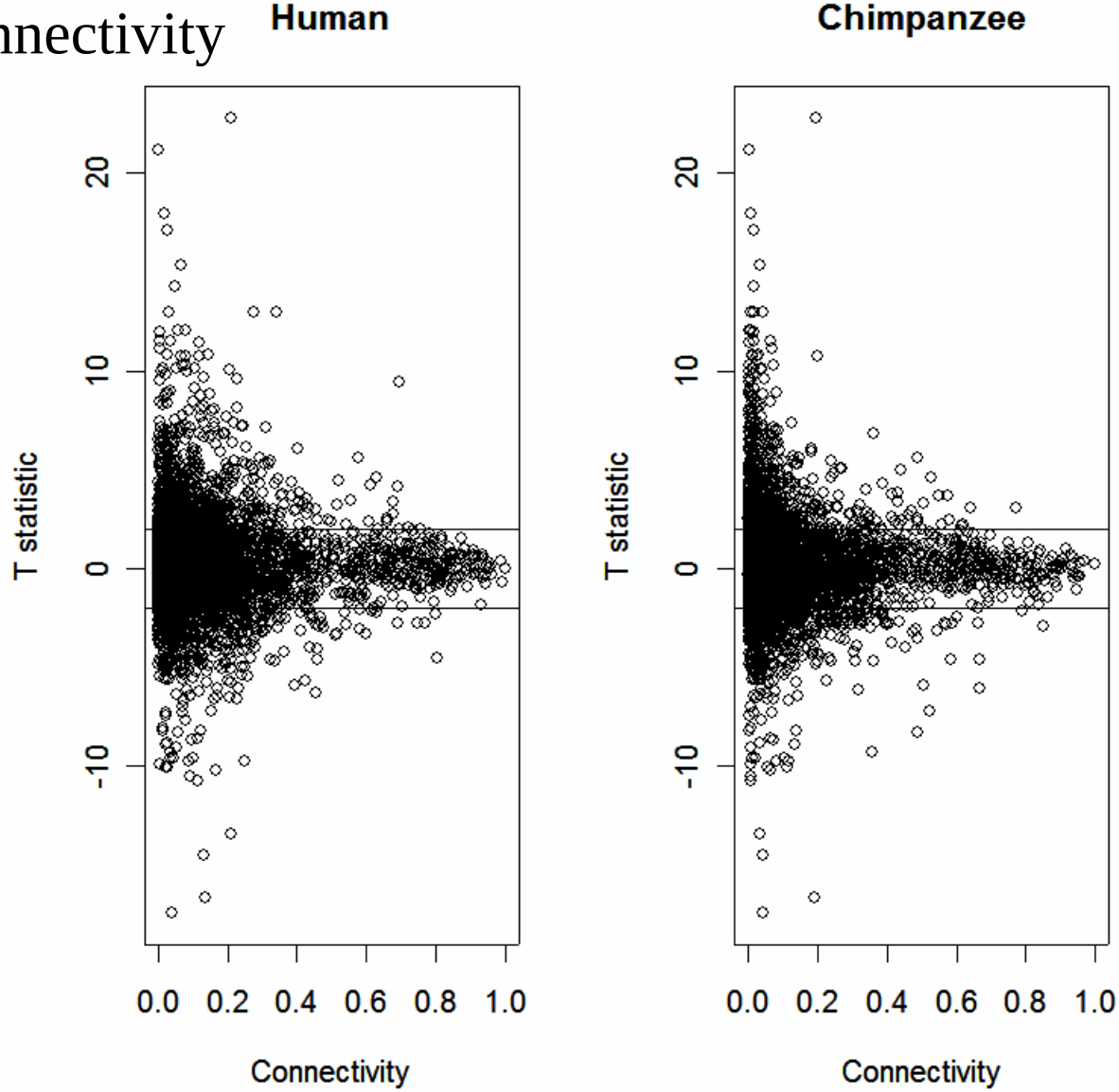


$p = 1.33 \times 10^{-4}$ $p = 8.93 \times 10^{-4}$ $p = 1.35 \times 10^{-6}$ $p = 1.33 \times 10^{-4}$



$p = 2.97 \times 10^{-3}$ $p = 8.02 \times 10^{-3}$ $p = 2.24 \times 10^{-3}$

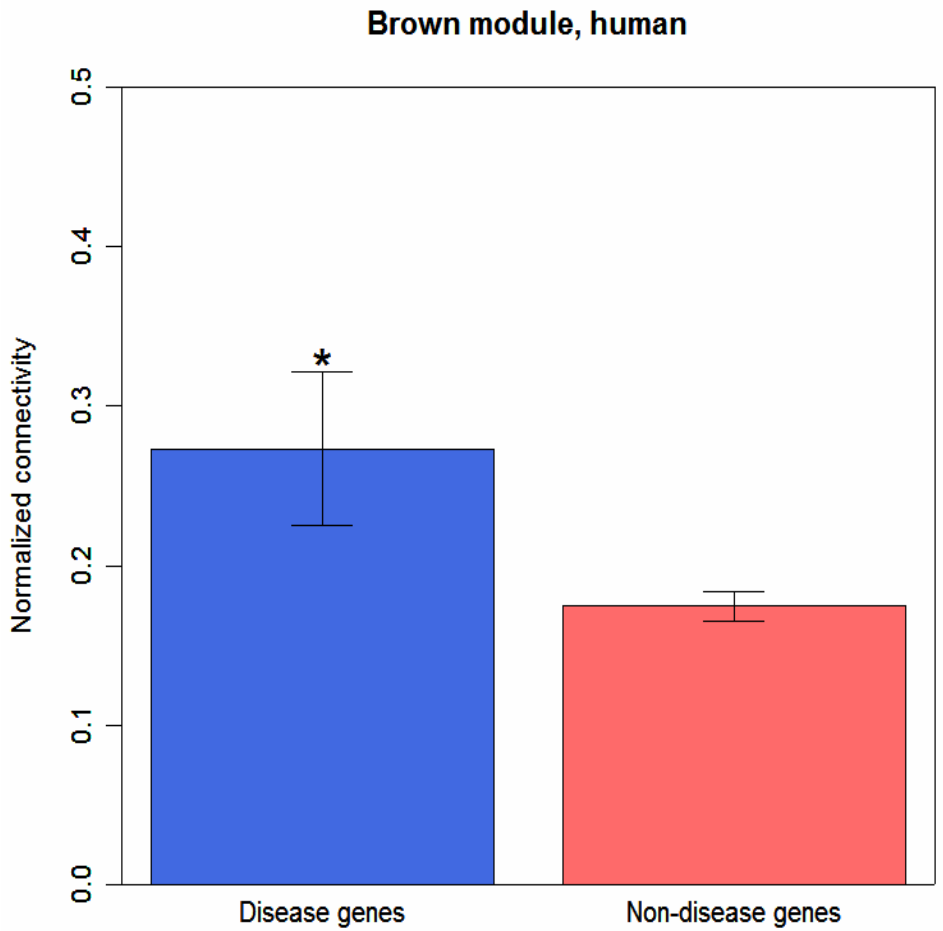
Relationship between T statistic (differential gene expression) and connectivity



Message: highly connected genes have preserved gene expressions

Neurodegenerative disease genes have a higher connectivity than non-disease genes.

- Congruent with association between connectivity and lethality in lower organisms
- OMIM database was queried with the term neurodegeneration, yielding genes causing or related to neurodegenerative diseases.



Conclusions: chimp/human

- Gene expression is highly preserved across species brains
- Gene co-expression is less preserved
- Gene modules correspond roughly to brain architecture
  - Modules that correspond to ‘old’ architectural features are highly preserved across species
  - Cortex module is not preserved.
- We find evidence that Mendelian disease genes tend to have more connections than anonymous genes.

Conclusion

- Gene co-expression network analysis can be interpreted as the study of the Pearson correlation matrix.
- Key insight: connectivity can be used to single out important genes.
- Weak relationship with principal or independent component analysis
  - Network methods focus on “local” properties
- Open questions:
  - What is the mathematical meaning of the scale free topology criterion
    - Starting point: noise suppression in modules.
  - Alternative connectivity measures, network distance measures
  - Which and how many genes to target to disrupt a disease module?

References for this talk

•*Bin Zhang and Steve Horvath (2005) "A General Framework for Weighted Gene Co-Expression Network Analysis", Statistical Applications in Genetics and Molecular Biology: Vol. 4: No. 1, Article 17.*
http://www.bepress.com/sagmb/vol4/iss1/art17

•R software tutorials at

http://www.genetics.ucla.edu/labs/horvath/CoexpressionNetwork/

•Notion of general topological overlap matrix
http://www.genetics.ucla.edu/labs/horvath/GTOM/

•For people who like math, check out
http://www.genetics.ucla.edu/labs/horvath/ModuleConformity/

•Cancer network:
Mischel PS, Zhang B, Carlson M, Fang Z, Freije W, Castro E, Scheck AC, Liau LM, Kornblum HI, Geschwind DH, Cloughesy TF, Horvath S, Nelson SF (2005) A Network Approach to Detecting Individual Prognostic Genes and Therapeutic Targets in Brain Cancer. Submitted.

Acknowledgement

Biostatistics/Bioinformatics

- Jun Dong, Postdoc
- Ai Li, graduate student
- Andy Yip, graduate student
- Bin Zhang, senior statistician
- Chris Plaisier, Access student

Brain Cancer/Yeast

- Paul Mischel, Prof
- Stan Nelson, Prof
- Marc Carlson, Postdoc

Comparison Human-Chimp

Dan Geschwind, Prof

Mike Oldham, grad student

General REFERENCES

- Albert R, Barabási AL (2002) Statistical mechanics of complex networks, Reviews of Modern Physics 74, 47 (2002).
- Almaas E, Kovacs B, Vicsek T, Z.N. Oltvai and A.-L. Barabási (2004) Global organization of metabolic fluxes in the bacterium. Escherichia coli. Nature 427, 839-843
- Balázsi G, Kay KA, Barabási AL, Oltvai Z (2003) Spurious spatial periodicity of co-expression in mocroarray data due to printing design. Nucleic Acids Research 31, 4425-4433 (2003)
- Barabási AL, Bonabeau E (2003) Scale-Free Networks. Scientific American 288, 60-69
- Barabási AL, Oltvai ZN (2004) Network Biology: Understanding the Cells's Functional Organization. Nature Reviews Genetics 5, 101-113
- Bergman S, Ihmels J, Barkai N (2004) Similarities and Difference in Genome-Wide Expression Data of Six Organisms. PLOS Biology. Jan 2004. Vol 2, Issue 1, pp0085-0093
- Davidson, G. S., Wylie, B. N., & Boyack, K. W. (2001). Cluster stability and the use of noise in interpretation of clustering. Proc. IEEE Information Visualization 2001, 23-30.
- Dezso Z, Oltvai ZN, Barabási AL (2003) Bioinformatics analysis of experimentally determined protein complexes in the yeast saccharomyces cerevisiae. Genome Research 13, 2450-2454 (2003)
- Dobrin R, Beg QK, Barabási AL (2004) Aggregation of topological motifs in the Escherichia coli tranascriptional. BMC Bioinformatics 5: 10 (2004)
- Farkas I, Jeong H, Vicsek HT, Barabasi AL, Oltvai ZN (2003) The topology of transcription regulatory network in the yeast, Saccharomyces cerevisiae. Physica A 318, 601-612 (2003)
- Giaever G, Chu AM, Ni L, Connelly C, Riles L, et al. (2002) Functional profiling of the Saccharomyces cerevisiae genome. Nature 418(6896): 387-391.
- Ihaka R, Gentleman R (1996) R: a language for data analysis and graphics. J. Comput. Graphical Statistics, 5, 299-314.
- Jeong H, Tombor B, Albert R, Oltvai ZN, Barabási AL (2000) The large-scale organization of metabolic networks. Nature 407, 651-654 (2000).
- Jeong H, Mason S, Barabási AL and Oltvai ZN (2001) Lethality and centrality in protein networks. Nature 411, 41-42 (2001)
- Kaufman, L. and Rousseeuw, P.J. (1990), Finding Groups in Data: An Introduction to Cluster Analysis (New York: John Wiley & Sons, Inc.)
- Klein, J. P. and Moeschberger, M. L. (1997) Survival Analysis: Techniques for Censored and Truncated Data, Springer-Verlag, New York.
- Li C, Wong WH (2001) Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection, Proc. Natl. Acad. Sci. Vol. 98, 31-36
- Podani J, Oltvai ZN, Jeong H, Tombor B, Barabási AL, E. Szathmáry E (2001) Comparable system-level organization of Archaea and Eukaryotes. Nature Genetics 29, 54-56 (2001)
- Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabasi AL (2002) “Hierarchical organization of modularity in metabologic networks”. Science Vol 297 pp1551-1555
- Stuart JM et al. Science 2003. A gene-coexpression network for global discovery of conserved genetic modules.
- van Noort V, Snel B, Huynen MA (2003) Predicting gene function by conserved co-expression. Trends Genet 19(5): 238-242.
- van Noort V, Snel B, Huynen MA (2004) The yeast coexpression network has a small-world, scale-free architecture and can be explained by a simple model. EMBO Rep 5(3): 280-284
- Wuchty S, Ravasz E, Barabási AL (2003) The Architecture of Biological Networks in T.S. Deisboeck, J. Yasha Kresh and T.B. Kepler (eds.) Complex Systems in Biomedicine. Kluwer Academic Publishing, New York (2003)
- Yook SY, Oltvai ZN and Barabási AL (2004) Functional and topological characterization of protein interaction networks. Proteomics 4, 928-942 (2004)
- Bin Zhang and Steve Horvath (2005) "A General Framework for Weighted Gene Co-Expression Network Analysis", Statistical Applications in Genetics and Molecular Biology: Vol. 4: No. 1, Article 17. http://www.bepress.com/sagmb/vol4/iss1/art17