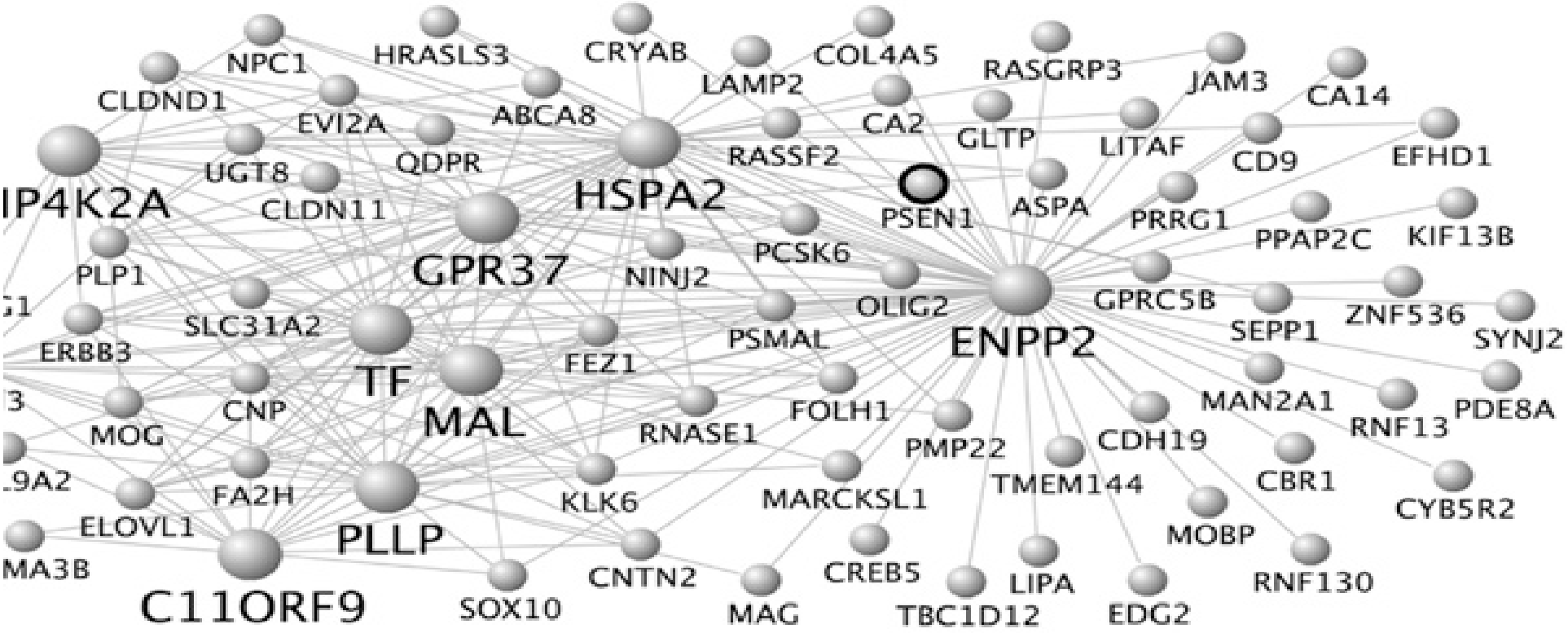
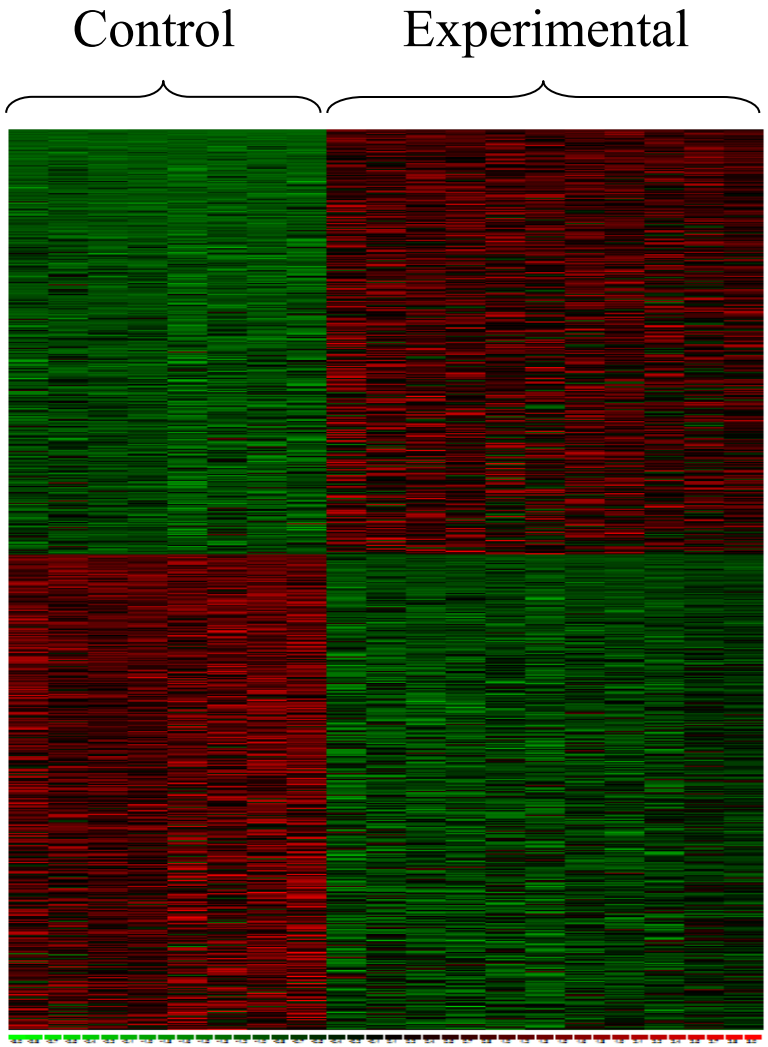


Weighted Correlation Network Analysis and Systems Biologic Applications

Standard differential expression analyses seek to identify individual genes

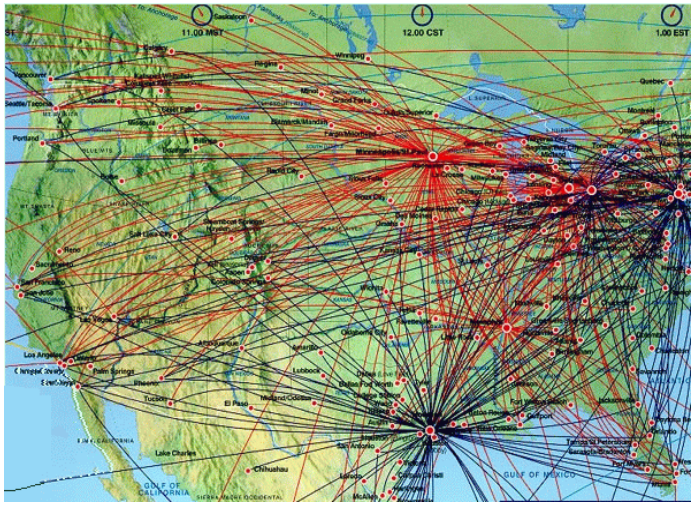


- Each gene is treated as an individual entity
- Often misses the forest for the trees: Fails to recognize that thousands of genes can be organized into relatively few modules

Philosophy of Weighted Gene Co-Expression Network Analysis

- Understand the “system” instead of reporting a list of individual parts
  - Describe the functioning of the engine instead of enumerating individual nuts and bolts
- Focus on modules as opposed to individual genes
  - this greatly alleviates multiple testing problem
- Network terminology is intuitive to biologists

Flight connections and hub airports



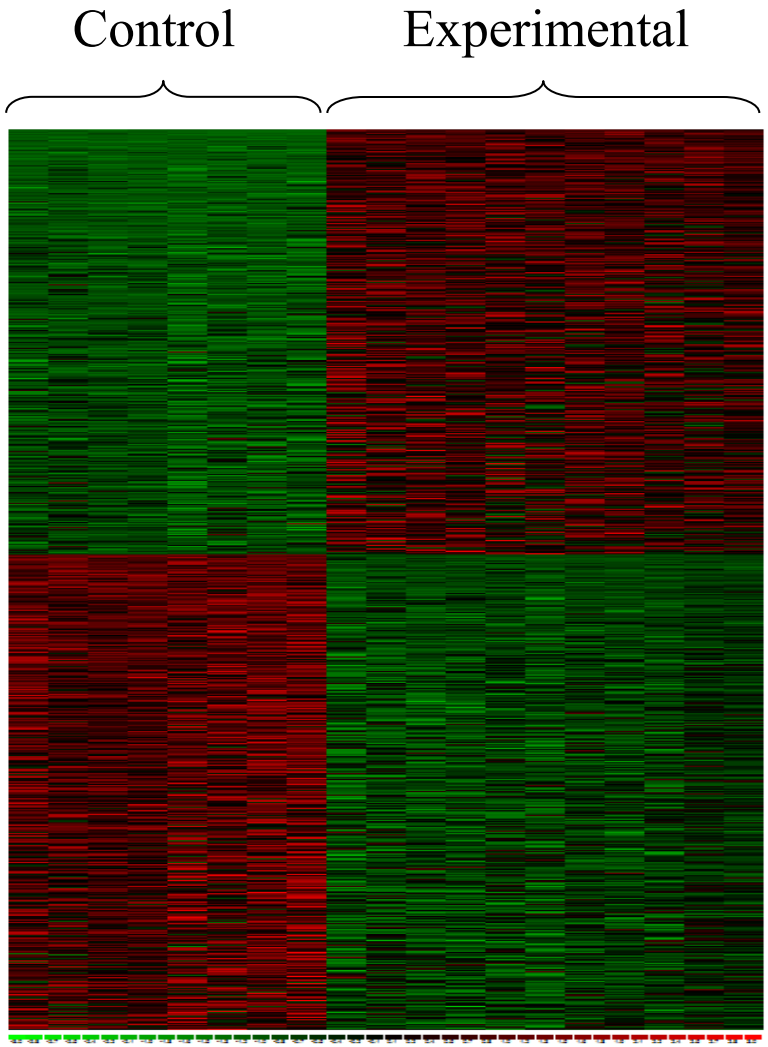
The nodes with the largest number of links (connections) are most important!

****Slide courtesy of AL Barabasi**

Contents

- How to construct a weighted gene co-expression network?
- Why use soft thresholding?
- How to detect network modules?
- How to relate modules to an external clinical trait?
- What is intramodular connectivity?
- How to use networks for gene screening?
- How to integrate networks with genetic marker data?
- What is weighted gene co-expression network analysis (WGCNA)?

Standard differential expression analyses seek to identify individual genes



- Each gene is treated as an individual entity
- Often misses the forest for the trees: Fails to recognize that thousands of genes can be organized into relatively few modules

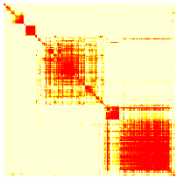
Philosophy of Weighted Gene Co-Expression Network Analysis

- Understand the “system” instead of reporting a list of individual parts
  - Describe the functioning of the engine instead of enumerating individual nuts and bolts
- Focus on modules as opposed to individual genes
  - this greatly alleviates multiple testing problem
- Network terminology is intuitive to biologists

What is weighted gene co-expression network analysis?

Construct a network

Rationale: make use of interaction patterns between genes



Identify modules

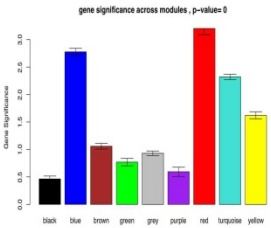
Rationale: module (pathway) based analysis

Relate modules to external information

Array Information: Clinical data, SNPs, proteomics

Gene Information: gene ontology, EASE, IPA

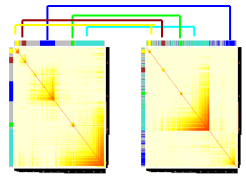
Rationale: find biologically interesting modules



Study Module Preservation across different data

Rationale:

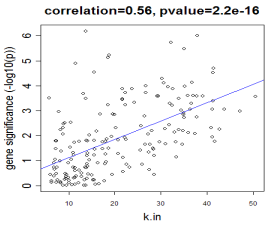
- Same data: to check robustness of module definition
- Different data: to find interesting modules.



Find the key drivers in *interesting* modules

Tools: intramodular connectivity, causality testing

Rationale: experimental validation, therapeutics, biomarkers



Weighted correlation networks are valuable for a biologically meaningful…

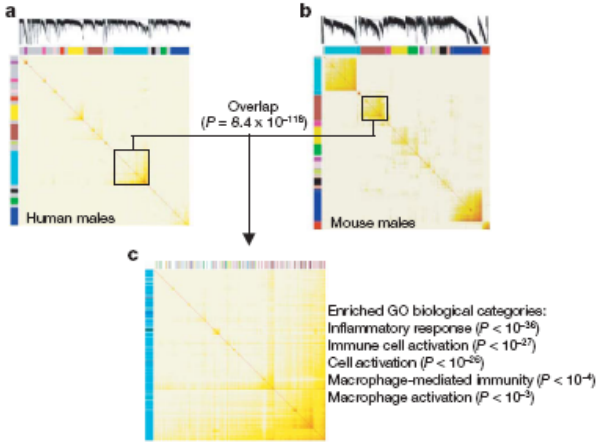
- reduction of high dimensional data
  - expression: microarray, RNA-seq
  - gene methylation data, fMRI data, etc.
- integration of multiscale data:
  - expression data from multiple tissues
  - SNPs (module QTL analysis)
  - Complex phenotypes

Vol 452|27 March 2008|doi:10.1038/nature06758 nature

ARTICLES

Genetics of gene expression and its effect on disease

Valur Emilsson[1,2], Gudmar Thorleifsson[1], Bin Zhang[2], Amy S. Leonardson[2], Florian Zink[1], J
Agnar Helgason[1], G. Bragi Walters[1], Steinunn Gunnarsdottir[1], Magali Mouy[1], Valgerdur
Gudrun H. Eiriksdottir[1], Gyda Bjornsdottir[1], Inga Reynisdottir[1], Daniel Gudbjartsson[1], An
Aslaug Jonasdottir[1], Adalbjorg Jonasdottir[1], Unnur Styrkarsdottir[1], Solveig Gretarsdottir[1],
Hreinn Stefansson[1], Ragnheidur Fossdal[1], Kristleifur Kristjansson[1], Hjortur G. Gislason[3],
Bjorn G. Leifsson[3], Unnur Thorsteinsdottir[1], John R. Lamb[2], Jeffrey R. Gulcher[1], Marc L. Rei
Eric E. Schadt[2]* & Kari Stefansson[1]*



LETTER

doi:10.1038/nature10110

Transcriptomic analysis of autistic brain reveals convergent molecular pathology

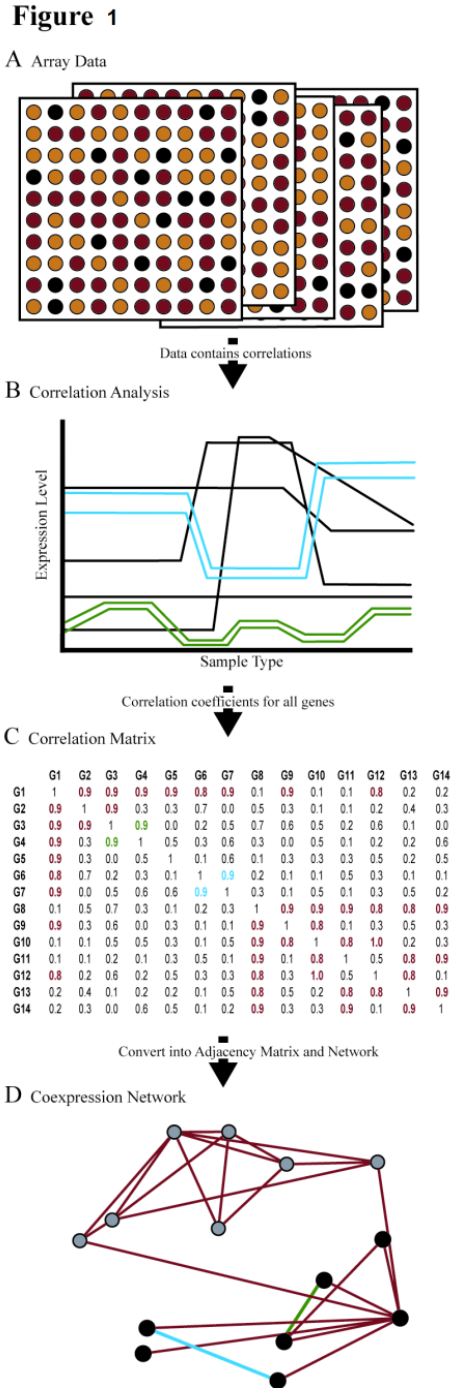
Irina Voineagu[1], Xinchen Wang[2], Patrick Johnston[3], Jennifer K. Lowe[1], Yuan Tian[1], Steve Horvath[4], Jonathan Mill[3], Rita M. Cantor[4], Benjamin J. Blencowe[2] & Daniel H. Geschwind[1,4]

How to construct a weighted gene co-expression network?

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.

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



Steps for constructing a co-expression network

A) Gene expression data (array or RNA-seq)

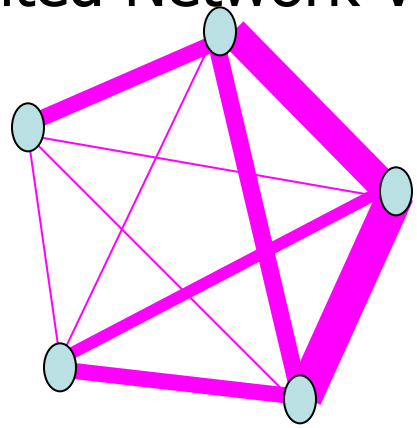
B) Measure co-expression with a correlation coefficient

C) The correlation matrix is either dichotomized to arrive at an adjacency matrix → unweighted network

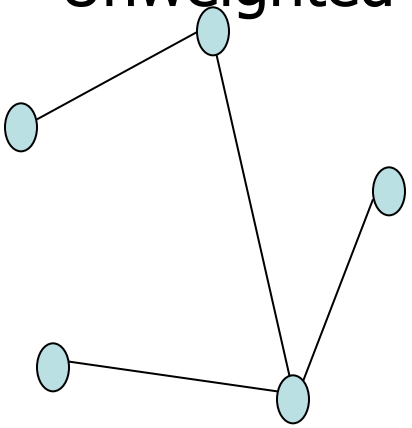
Or transformed continuously with the power adjacency function → weighted network

The `holistic' view of a weighted network

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.

If two genes are correlated with r=0.79, they are deemed unconnected with regard to a hard threshold of tau=0.8

Two types of weighted correlation networks

Unsigned network, absolute value

$$a_{ij} = | cor(x_i, x_j) |^{\beta}$$

Signed network preserves sign info

$$a_{ij} = | 0.5 + 0.5 \times cor(x_i, x_j) |^{\beta}$$

Default values: β=6 for unsigned and β =12 for signed networks.

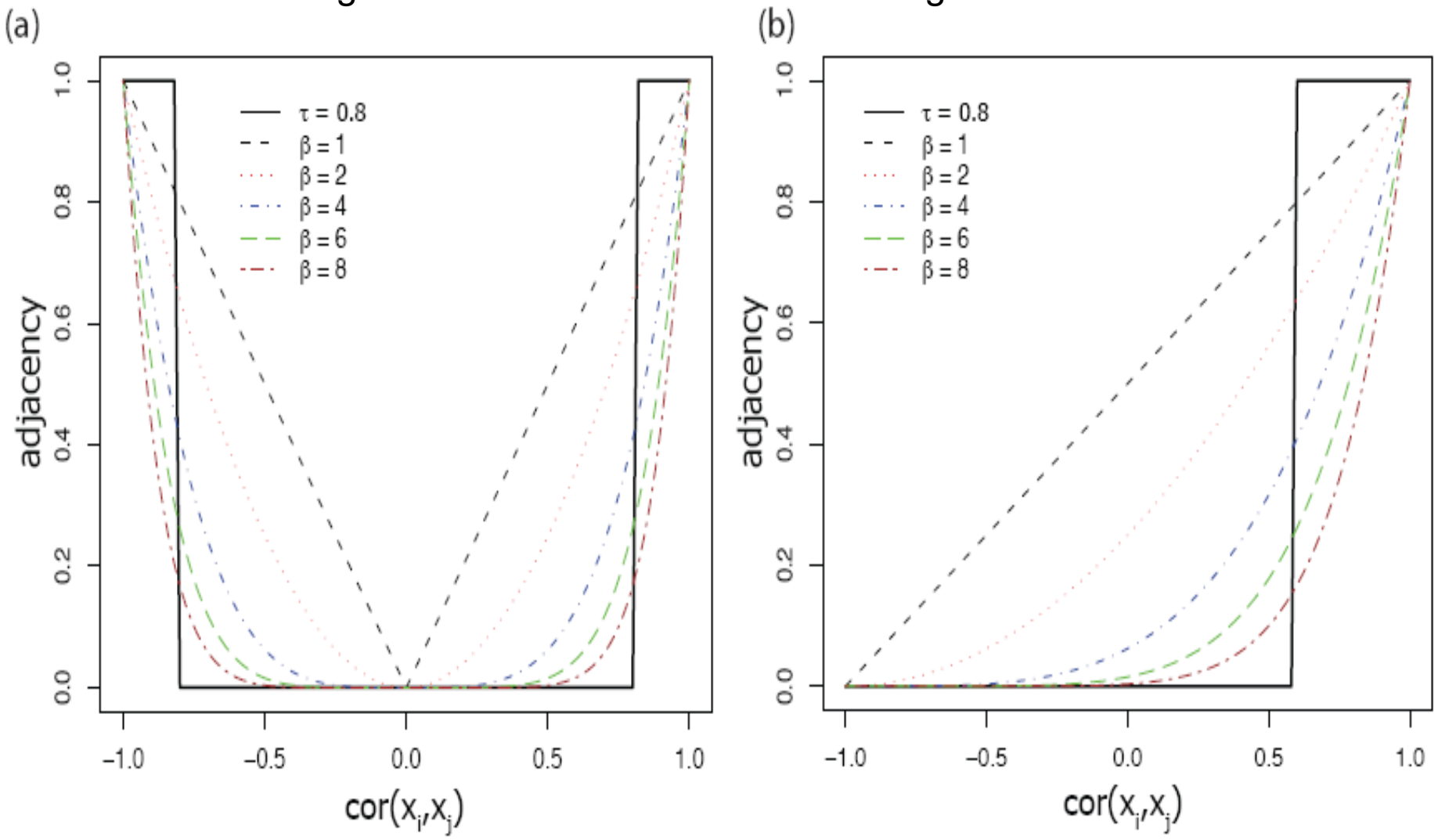
We prefer signed networks…

Zhang et al SAGMB Vol. 4: No. 1, Article 17.

Adjacency versus correlation in unsigned and signed networks

Unsigned Network

Signed Network



Question 1:
Should network construction account for the sign of the co-expression relationship?

Answer: Overall, recent applications have convinced me that signed networks are preferable.

- For example, signed networks were critical in a recent stem cell application
- *Michael J Mason, Kathrin Plath, Qing Zhou, et al (2009) Signed Gene Co-expression Networks for Analyzing Transcriptional Regulation in Murine Embryonic Stem Cells. BMC Genomics 2009, 10:327*

Why construct a co-expression network based on the correlation coefficient ?

1. Intuitive
2. Measuring linear relationships avoids the pitfall of overfitting
3. Because many studies have limited numbers of arrays→ hard to estimate non-linear relationships
4. Works well in practice
5. Computationally fast
6. Leads to reproducible research

*Biweight midcorrelation (bicor)

- A robust alternative to Pearson correlation.
- Definition based on median instead of mean.
- Assign weights to observations, values close to median receive large weights.
- Robust to outliers.

$$covMedianWeighted(x,y) = \frac{sum((x - median(x))w.x(y - median(y))w.y)}{\sqrt{sum(w.x^2)sum(w.y^2)}}, \quad (5.16)$$

where the weights are given by $w.x = weight(robustScale(x))$.

Using these function, the **biweight midcorrelation** between x and y is defined as follows:

$$bicor(x,y) = \frac{covMedianWeighted(x,y)}{\sqrt{covMedianWeighted(x,x)covMedianWeighted(y,y)}}. \quad (5.17)$$

Book: "Data Analysis and Regression: A Second Course in Statistics", Mosteller and Tukey, Addison-Wesley, 1977, pp. 203-209

Langfelder et al 2012: Fast R Functions For Robust Correlations And Hierarchical Clustering. *J Stat Softw* 2012, **46**(i11):1–17.

Comparison of co-expression measures: mutual information, correlation, and model based indices.

- Song et al 2012 BMC Bioinformatics;13(1):328. PMID: 23217028

Result: biweight midcorrelation + topological overlap measure work best when it comes to defining co-expression modules

Why soft thresholding as opposed to hard thresholding?

1. Preserves the continuous information of the co-expression information
2. Results tend to be more robust with regard to different threshold choices

But hard thresholding has its own advantages:
In particular, graph theoretic algorithms from the computer science community can be applied to the resulting networks

Advantages of soft thresholding with the power function

1. Robustness: Network results are highly robust with respect to the choice of the power β (Zhang et al 2005)
2. Calibration of different networks becomes straightforward, which facilitates consensus module analysis
3. Module preservation statistics are particularly sensitive for measuring connectivity preservation in weighted networks
4. Math reason: Geometric Interpretation of Gene Co-Expression Network Analysis. PloS Computational Biology. 4(8): e1000117

Questions:
How should we choose the power beta or a hard threshold?
Or more generally the parameters of an adjacency function?

IDEA: use properties of the connectivity distribution

Generalized Connectivity

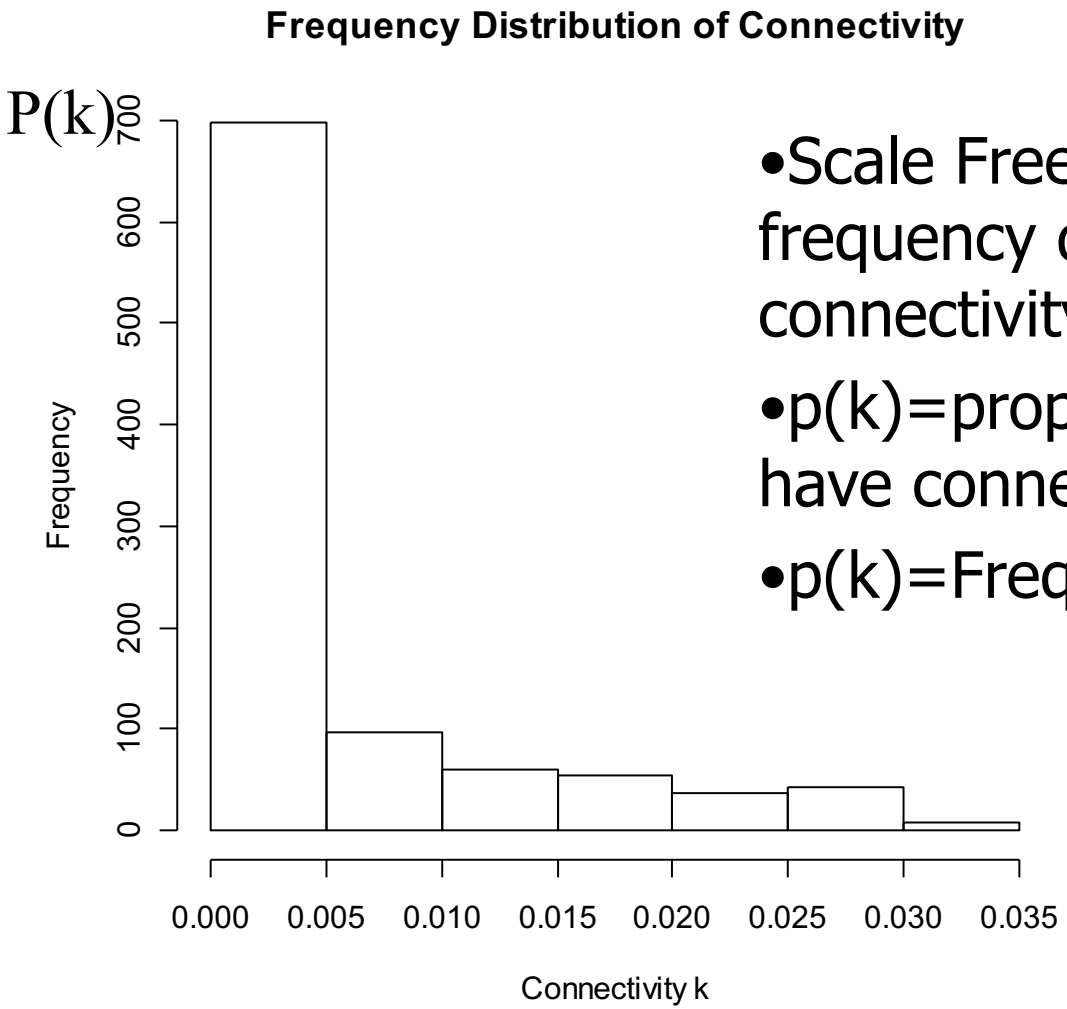
- Gene connectivity = row sum of the adjacency matrix
  - For unweighted networks=number of direct neighbors
  - For weighted networks= sum of connection strengths to other nodes

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

Approximate scale free topology is a fundamental property of such networks (L. 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.

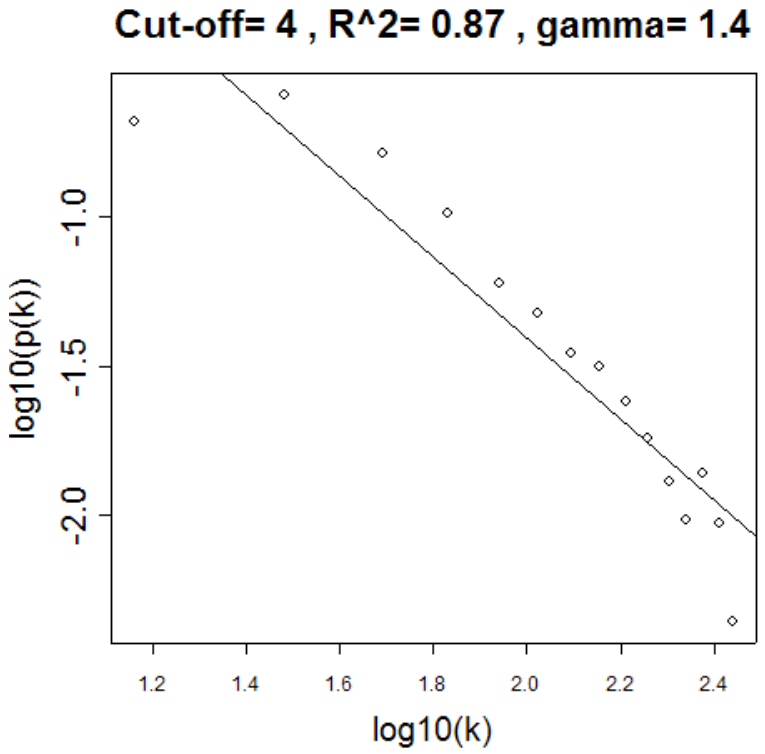
P(k) vs k in scale free networks



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

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

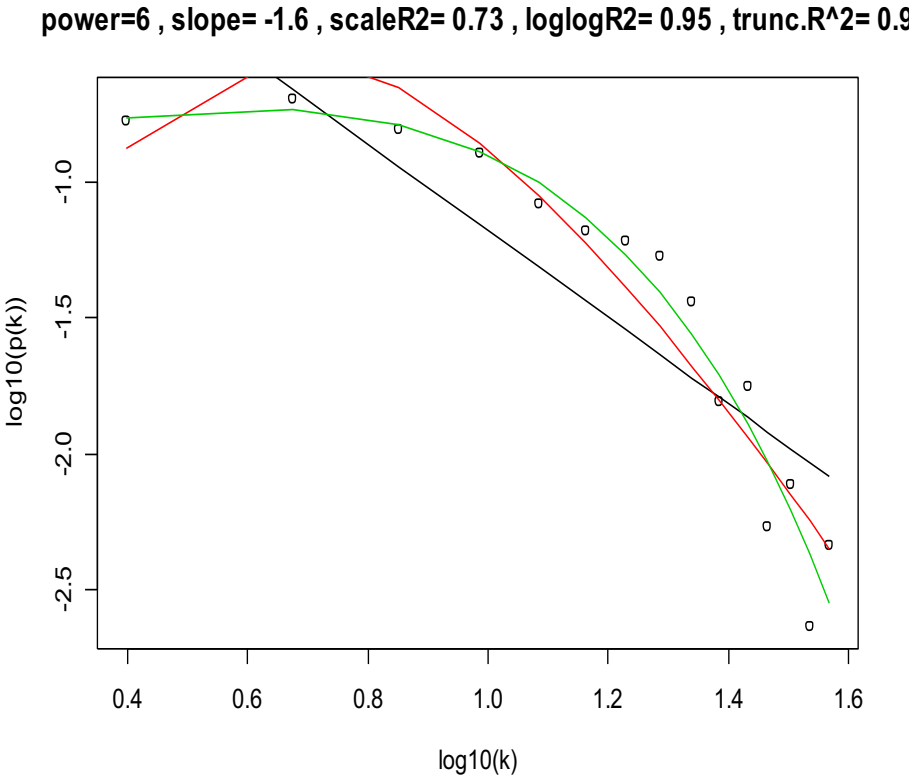
Motivation of generalizations: using weak general assumptions, we have proven that gene co-expression networks satisfy these distributions approximately.

Barabasi (1999) ScaleFree Topology means $\log(p(k)) = c_0 + c_1 \log(k)$

Csanyi-Szendroi (2004) ExponentiallyTrunated SFT means $\log(p(k)) = c_0 + c_1 \log(k) + c_2 k$

Horvath, Dong (2005) LogLogSFT means $\log(p(k)) = c_0 + c_1 \log(k) + c_2 \log(\log(k))$

Checking Scale Free Topology in the Yeast Network



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

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

CONSIDER ONLY THOSE PARAMETER VALUES IN THE ADJACENCY FUNCTION THAT RESULT IN APPROXIMATE SCALE FREE TOPOLOGY, i.e. high scale free topology fitting index R^2

In practice, we use the lowest value where the curve starts to "saturate"

Rationale:

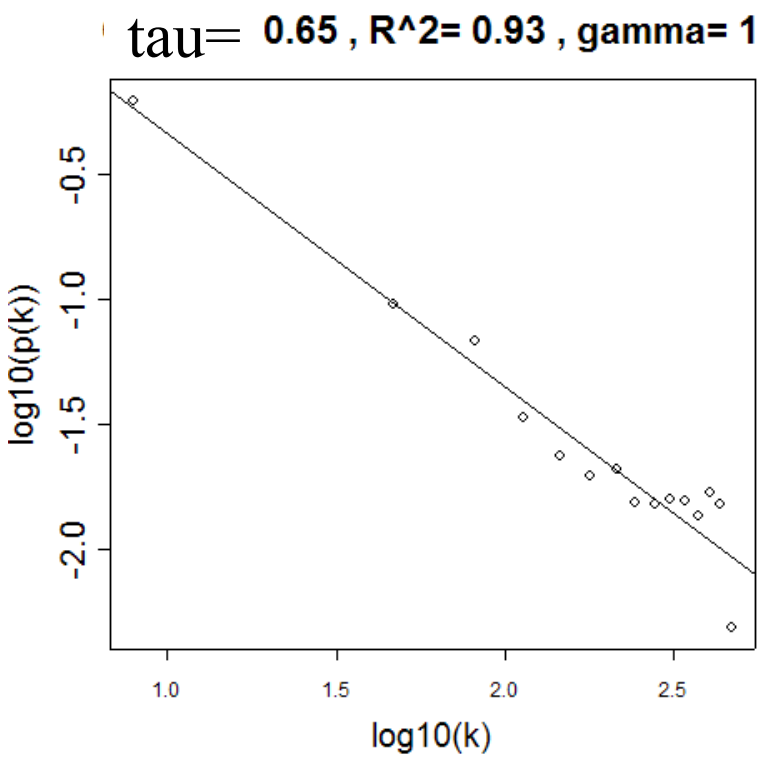
Empirical finding: Many co-expression networks based on expression data from a single tissue exhibit scale free topology

Many other networks e.g. protein-protein interaction networks have been found to exhibit scale free topology

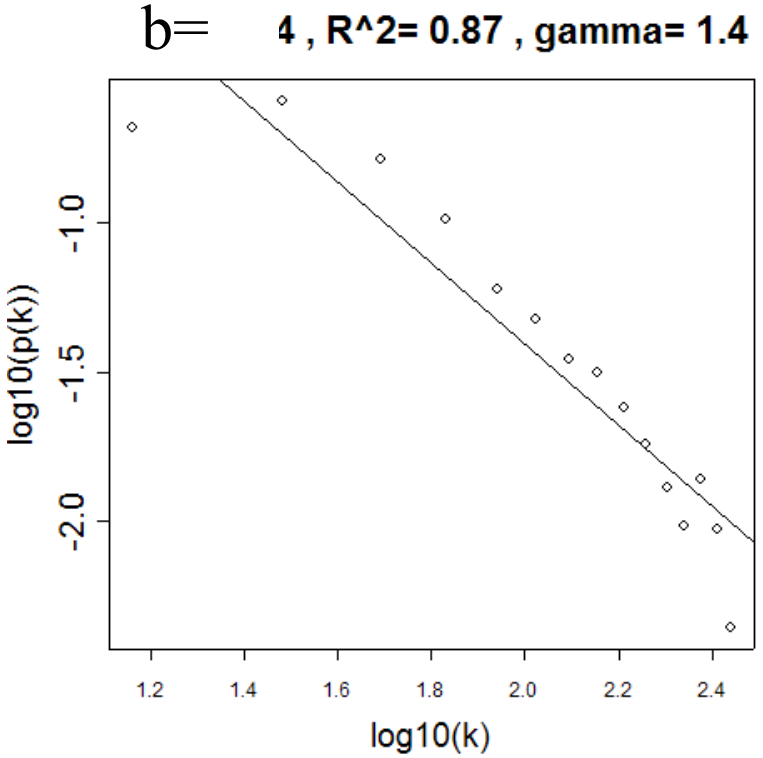
Caveat: When the data contains few very large modules, then the criterion may not apply. In this case, use the default choices.

Scale free topology is measured by the linear model fitting index R^2

Step AF (tau)

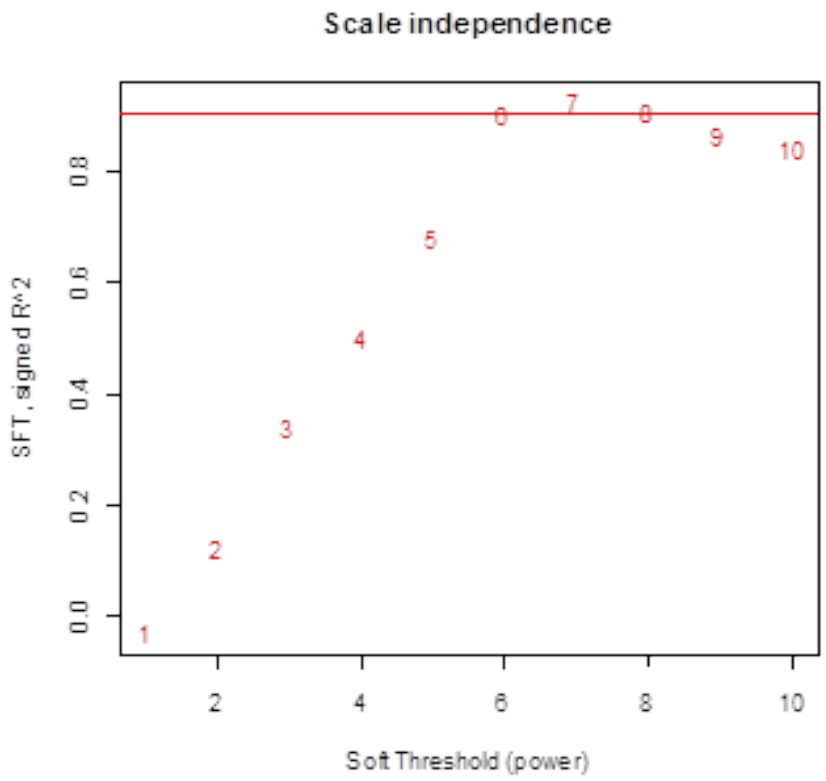


Power AF (b)

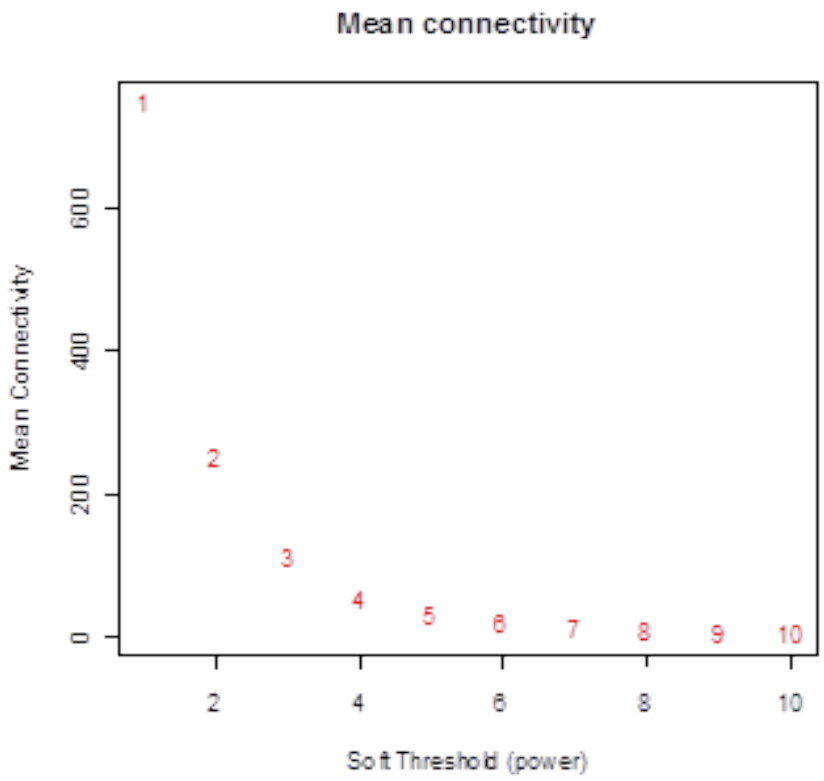


Scale free fitting index (R^2) and mean connectivity versus the soft threshold (power beta)

SFT model fitting index R^2



mean connectivity



How to measure interconnectedness in a network?

Answers:

1) adjacency matrix

2) topological overlap matrix

Topological overlap matrix and corresponding dissimilarity

(Ravasz et al 2002)

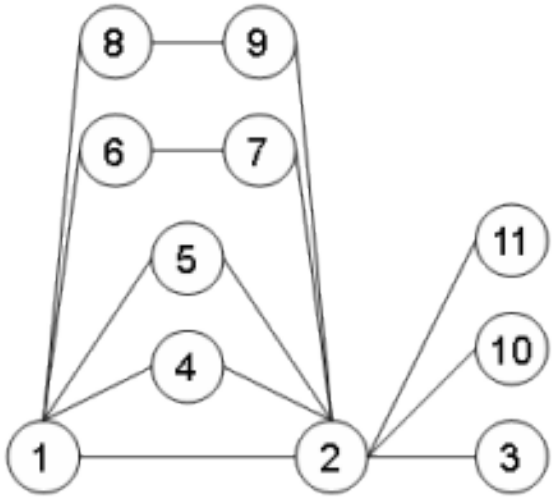
$$TOM_{ij} = \frac{\sum_u a_{iu} a_{uj} + a_{ij}}{\min(k_i, k_j) + 1 - a_{ij}}$$

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

- k=connectivity=row sum of adjacencies
- Generalization to weighted networks is straightforward since the formula is mathematically meaningful even if the adjacencies are real numbers in [0,1] (Zhang et al 2005 SAGMB)
- Generalized topological overlap (Yip et al (2007) BMC Bioinformatics)

Set interpretation of the topological overlap matrix

a.



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

$N_1(i)$ denotes the set of 1-step (i.e. direct) neighbors of node i

| | measures the cardinality

Adding 1-a(i,j) to the denominator prevents it from becoming 0.

Generalizing the topological overlap matrix to 2 step neighborhoods etc

- Simply replace the neighborhoods by 2 step neighborhoods in the following formula

$$GTOM2(i,j) = \frac{|N_2(i) \cap N_2(j)| + a_{ij}}{\min(|N_2(i)|, |N_2(j)|) + 1 - a_{ij}}$$

where $N_2(i)$ denotes the set of nodes within 2 steps of node i

- ***www.genetics.ucla.edu/labs/horvath/GTOM***

Yip A et al (2007) BMC Bioinformatics 2007, 8:22

How to detect network modules (clusters) ?

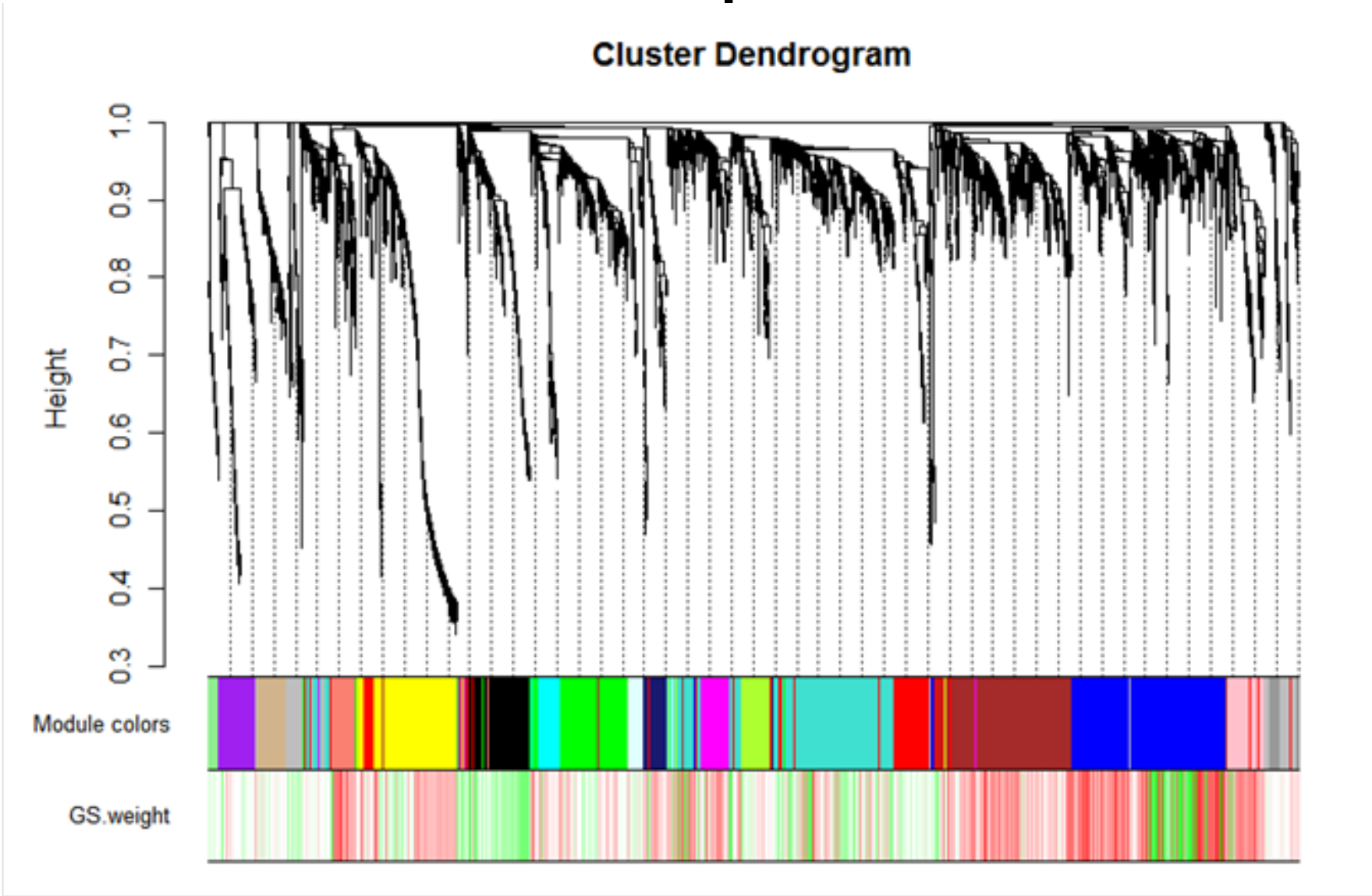
Module Definition

- We often use average linkage hierarchical clustering coupled with the topological overlap dissimilarity measure.
- Based on the resulting cluster tree, we define modules as branches
- Modules are either labeled by integers (1,2,3…) or equivalently by colors (turquoise, blue, brown, etc)

Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut library for R.

Langfelder P, Zhang B et al (2007)
Bioinformatics 2008 24(5):719-720

Example:

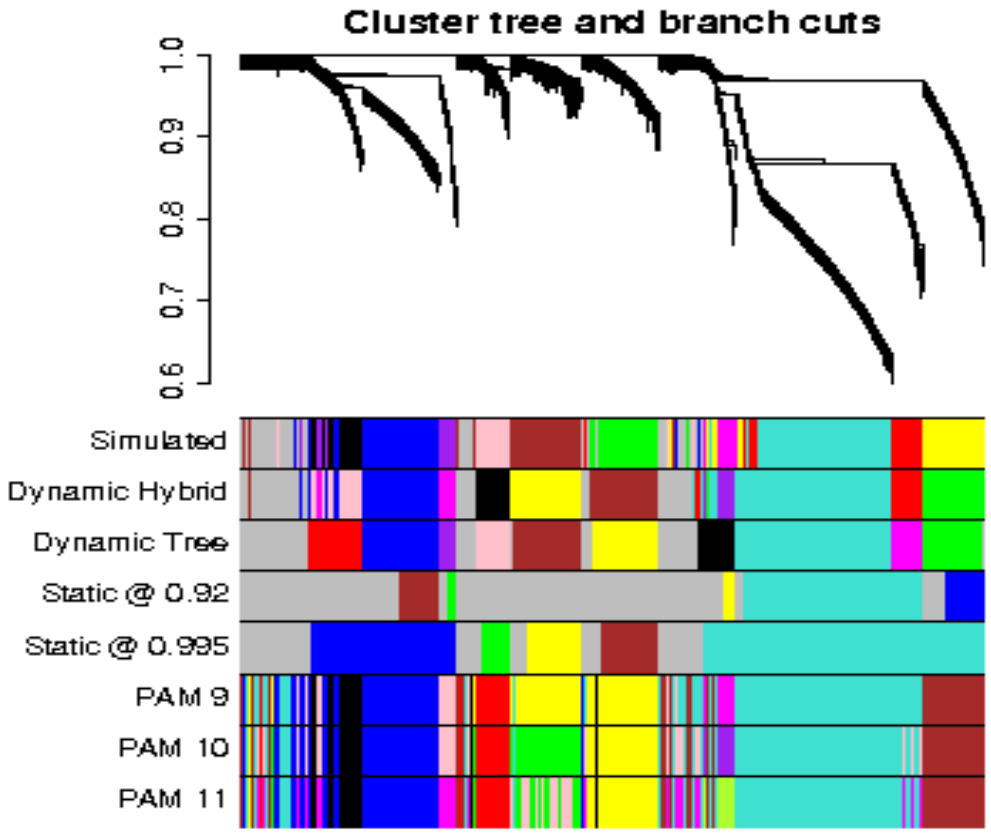


From: Ghazalpour et al (2006), *PLoS Genetics Volume 2 Issue 8*

Two types of branch cutting methods

- Constant height (static) cut
  - `cutreeStatic(dendro,cutHeight,minsize)`
  - based on R function `cutree`
- Adaptive (dynamic) cut
  - `cutreeDynamic(dendro, ...)`
- Getting more information about the dynamic tree cut:
  - `library(dynamicTreeCut)`
  - `help(cutreeDynamic)`
- More details:
  www.genetics.ucla.edu/labs/horvath/CoexpressionNetwork/BranchCutting/

How to cut branches off a tree?



Module=branch of a cluster tree

Dynamic hybrid branch cutting method combines advantages of hierarchical clustering and partitioning aruond medoid clustering

Summary

- Static tree cut: simple, but requires careful choice of height and not suitable for complicated dendrograms with nested clusters.
- Dynamic tree cut: Two versions, Tree and Hybrid
- Both look at the shape of the branches on the dendrogram, height and size information. Small clusters can be merged with neighboring large clusters
- Hybrid combines dendrogram cutting and PAM and retains advantages of both
  - no need to specify number of cluster
  - robustness

Summary (cont'd)

- Advantages of Dynamic Tree Cut methods over the constant height one:
  - More flexible: can deal with complicated dendrograms
  - Better outlier detection (Hybrid best, Tree not as good)
  - Suitable for automation (Tree possibly somewhat better because of fewer parameter settings)
  - Less sensitive to small changes in parameters, but user beware: defaults aren't always appropriate.

How to visualize networks?

Answer:

1) Topological overlap matrix plot aka. connectivity plot

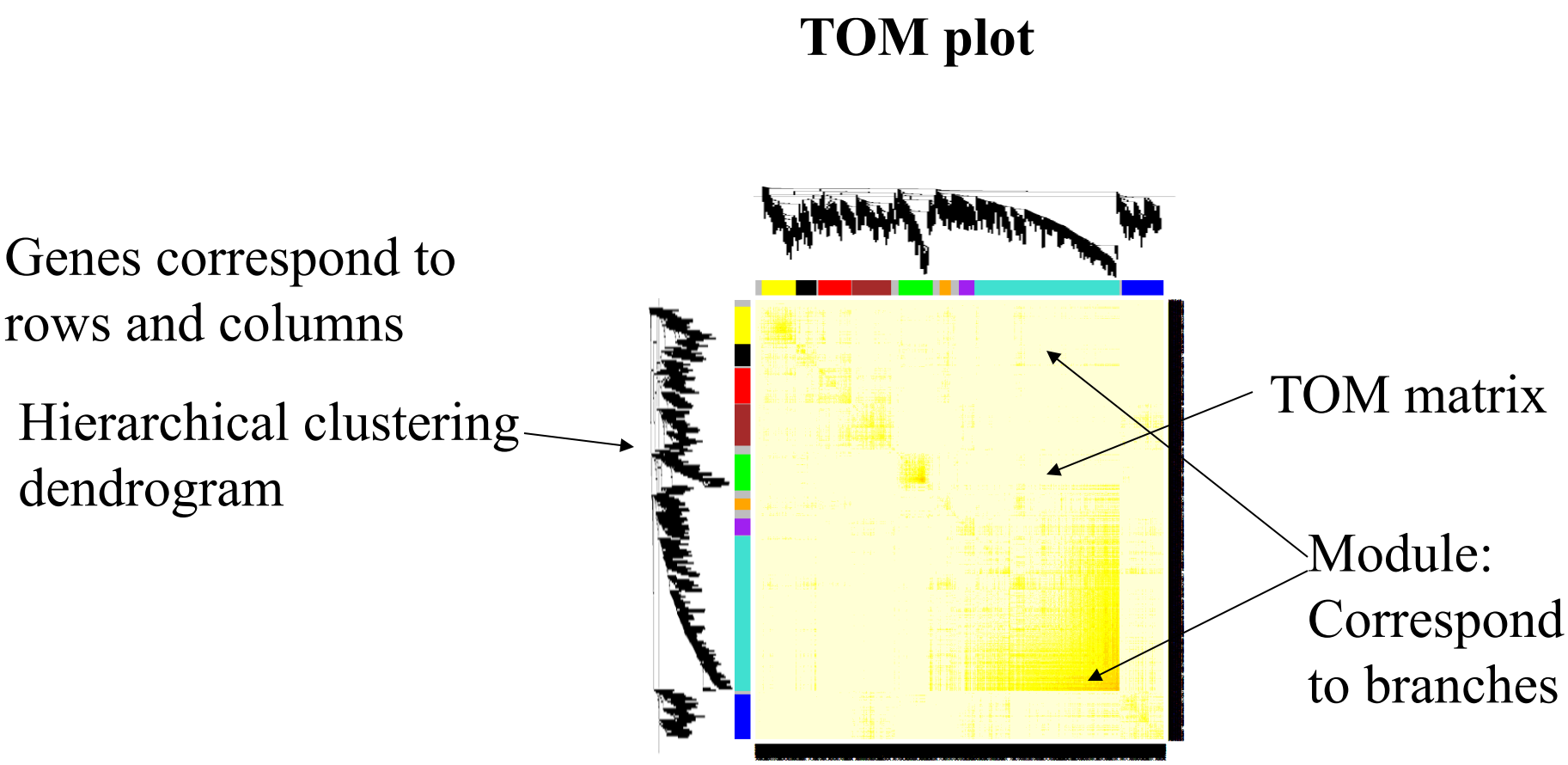
2) Multidimensional scaling

3) heatmaps of modules

4) external software: ViSANT,Cytoscape etc

Using the topological overlap matrix (TOM) to cluster genes

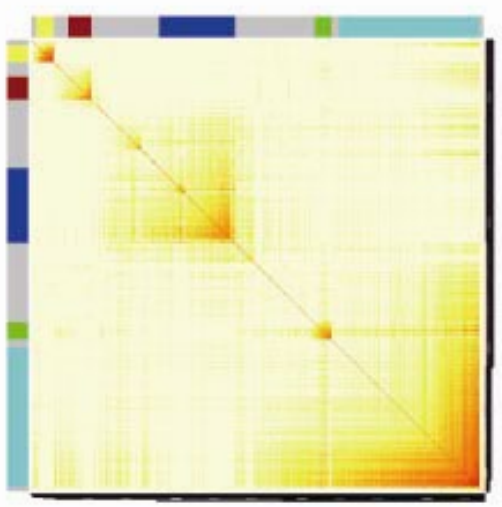
- Here modules correspond to branches of the dendrogram



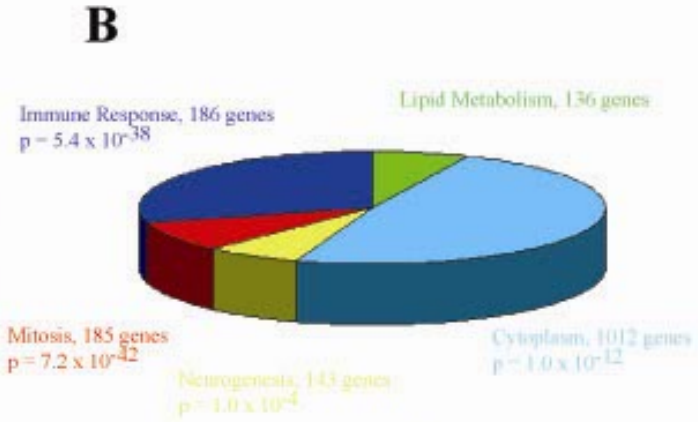
Different Ways of Depicting Gene Modules

Topological Overlap Plot

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

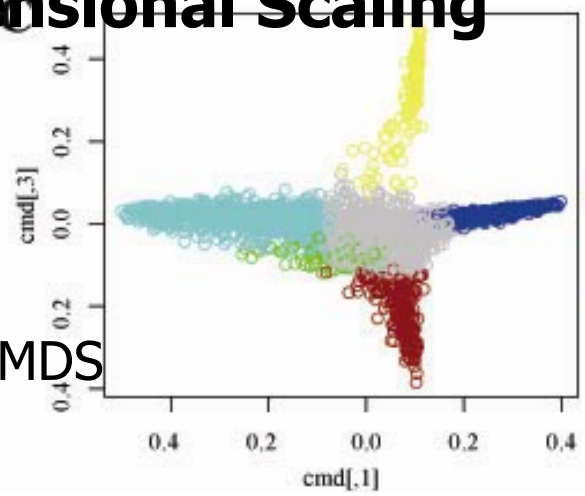


Gene Functions

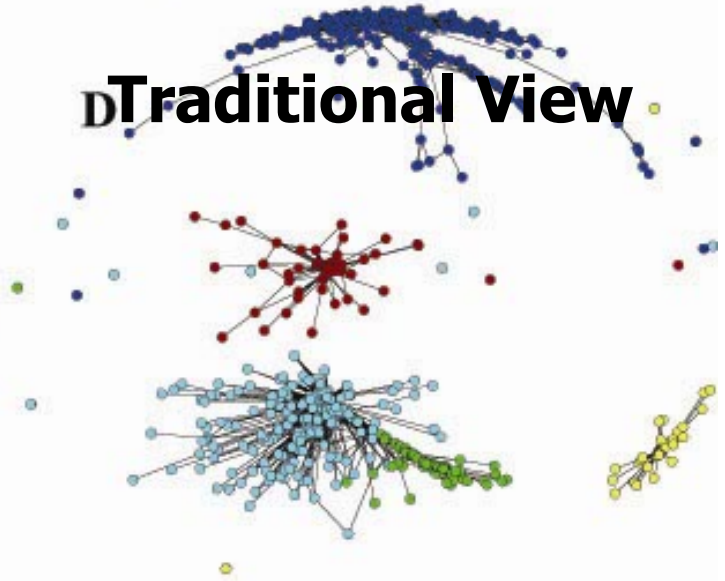


Multi Dimensional Scaling

Idea:
Use network distance in MDS



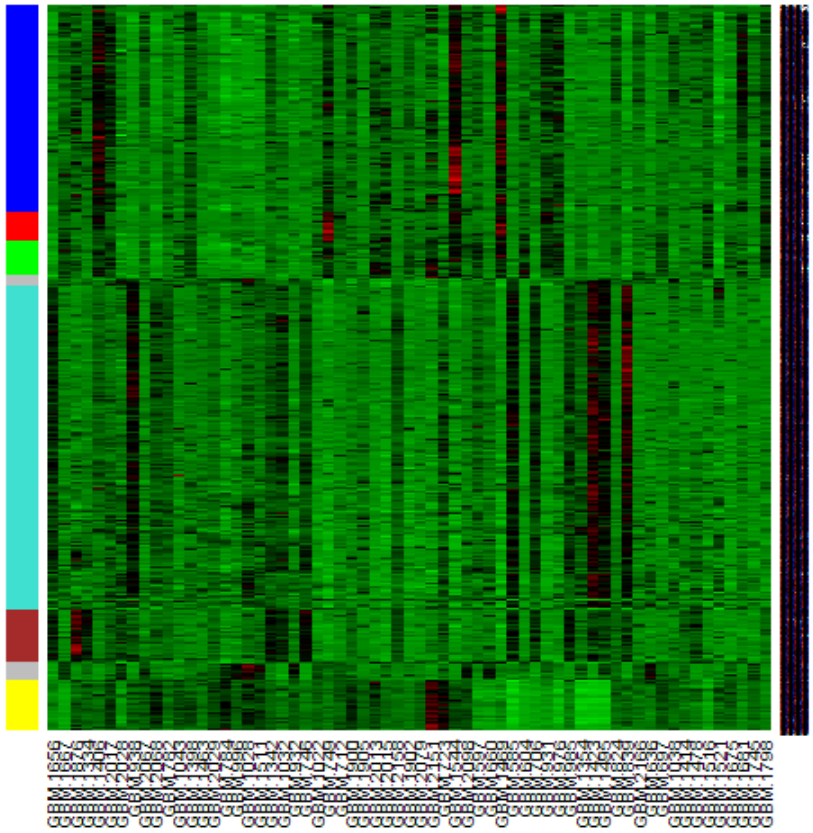
Traditional View



Heatmap view of module

Columns= tissue samples

Rows=Genes
Color band indicates module membership



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

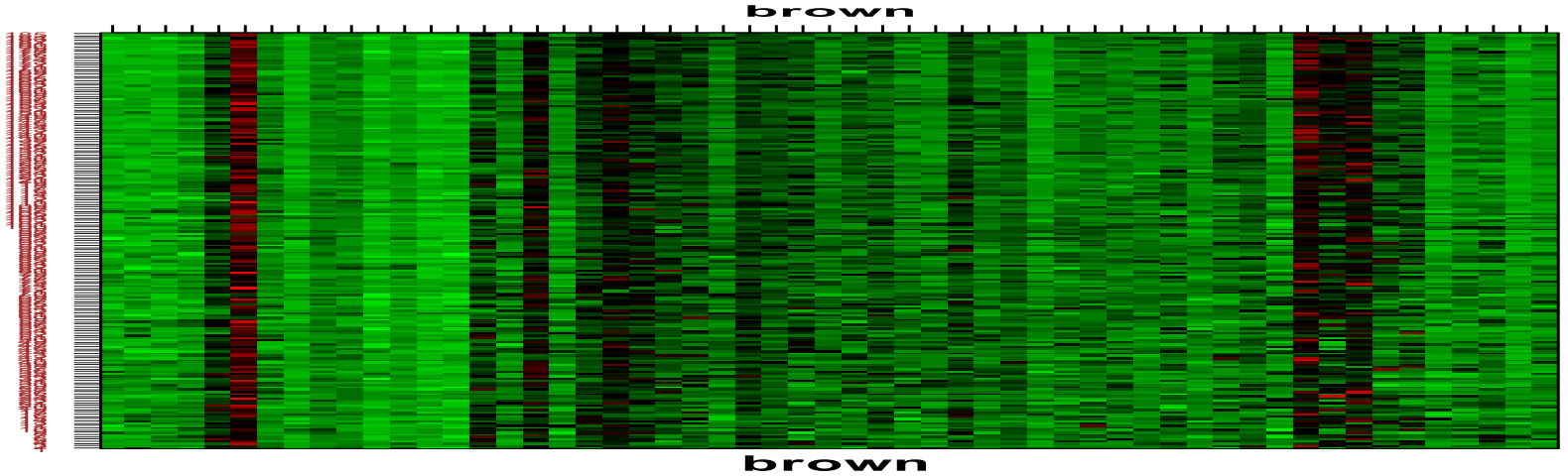
Question: How does one summarize the expression profiles in a module?

Answer: This has been solved.
Math answer: module eigengene = first principal component
Network answer: the most highly connected intramodular hub gene

Both turn out to be equivalent

Module Eigengene= measure of over-expression=average redness

Rows,=genes, Columns=microarray



The brown module eigengenes across samples

Using the singular value decomposition to define (module) eigengenes

Scale the gene expressions profiles (columns)

$$datX = scale(datX)$$

$$datX = UDV^T$$

$$U = (u_1 \quad u_2 \quad \text{L} \quad u_m)$$

$$V = (v_1 \quad v_2 \quad \text{L} \quad v_m)$$

$$D = diag(|d_1|, |d_2|, \text{K} \ , |d_m|)$$

Message: u_1 is the (first) eigengene E

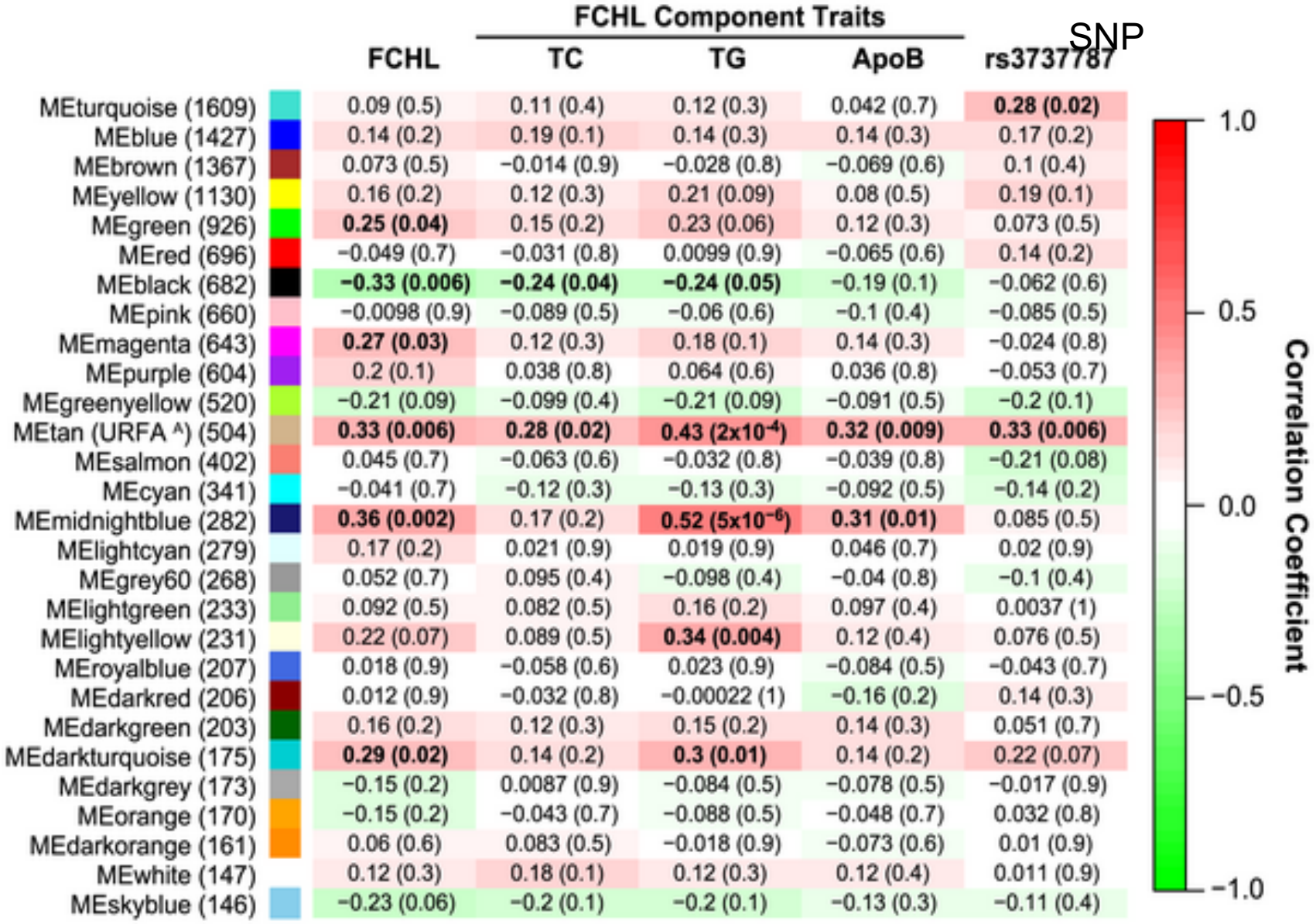
If $\text{datX}^{(q)}$ corresponds to the q-th module then

$\text{E}^{(q)}$ is the q-th module eigengene.

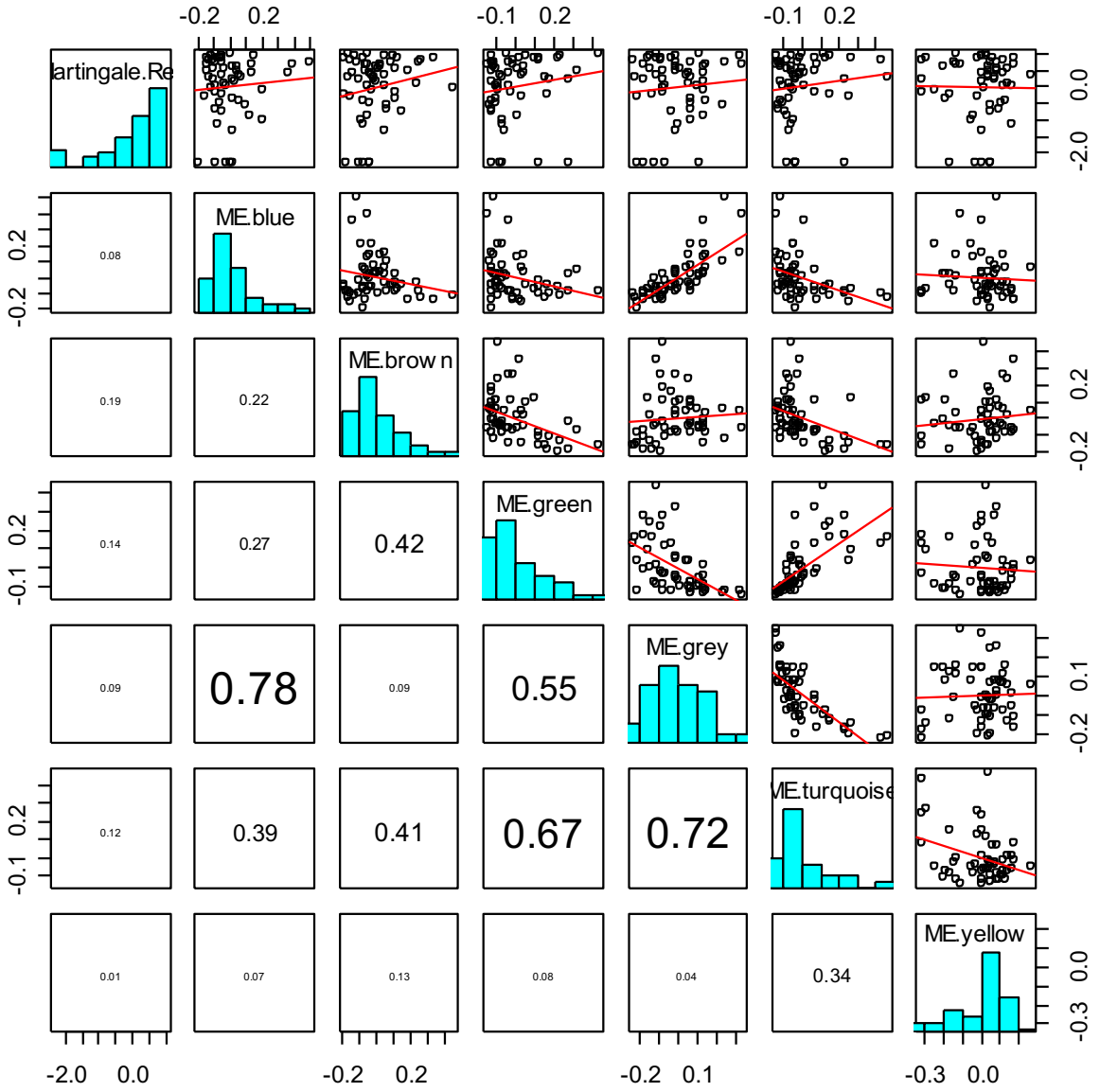
Module eigengenes are very useful

- 1) They allow one to relate modules to each other
  - Allows one to determine whether modules should be merged
  - Or to define eigengene networks
- 2) They allow one to relate modules to clinical traits and SNPs
  - -> avoids multiple comparison problem
- 3) They allow one to define a measure of module membership: kME=cor(x,ME)

Eigengenes correlated with lipid traits and a disease related SNP Plaisier, Pajukanta 2009 Plos Genet



Module eigengenes can be used to determine whether 2 modules are correlated. If correlation of MEs is high-> consider merging.



Eigengene networks Langfelder, Horvath (2007) BMC Systems Biology

Module detection in very large data sets

R function `blockwiseModules` (in WGCNA library) implements 3 steps:

1. Variant of k-means to cluster variables into blocks
2. Hierarchical clustering and branch cutting in each block
3. Merge modules across blocks (based on correlations between module eigengenes)

Works for hundreds of thousands of variables

How to relate modules to external data?

Clinical trait (e.g. case-control status) gives rise to a gene significance measure

- Abstract definition of a gene significance measure
  - GS(i) is non-negative,
  - the bigger, the more *biologically* significant for the i-th gene

Equivalent definitions

- GS.ClinicalTrait(i) = |cor(x(i),ClinicalTrait)| where x(i) is the gene expression profile of the i-th gene
- GS(i)=|T-test(i)| of differential expression between groups defined by the trait
- GS(i)=-log(p-value)

A SNP marker naturally gives rise to a measure of gene significance

$$GS.SNP(i) = |cor(x(i), SNP)|.$$

- Additive SNP marker coding: AA->2, AB->1, BB->0
- Absolute value of the correlation ensures that this is equivalent to AA->0, AB->1, BB->2
  - Dominant or recessive coding may be more appropriate in some situations
  - Conceptually related to a LOD score at the SNP marker for the i-th gene expression trait

A gene significance naturally gives rise to a module significance measure

- Define module significance as mean gene significance
- Often highly related to the correlation between module eigengene and trait

Important Task in Many Genomic Applications:

Given a network (pathway) of interacting genes how to find the central players?

Which of the following mathematicians had the biggest influence on others?

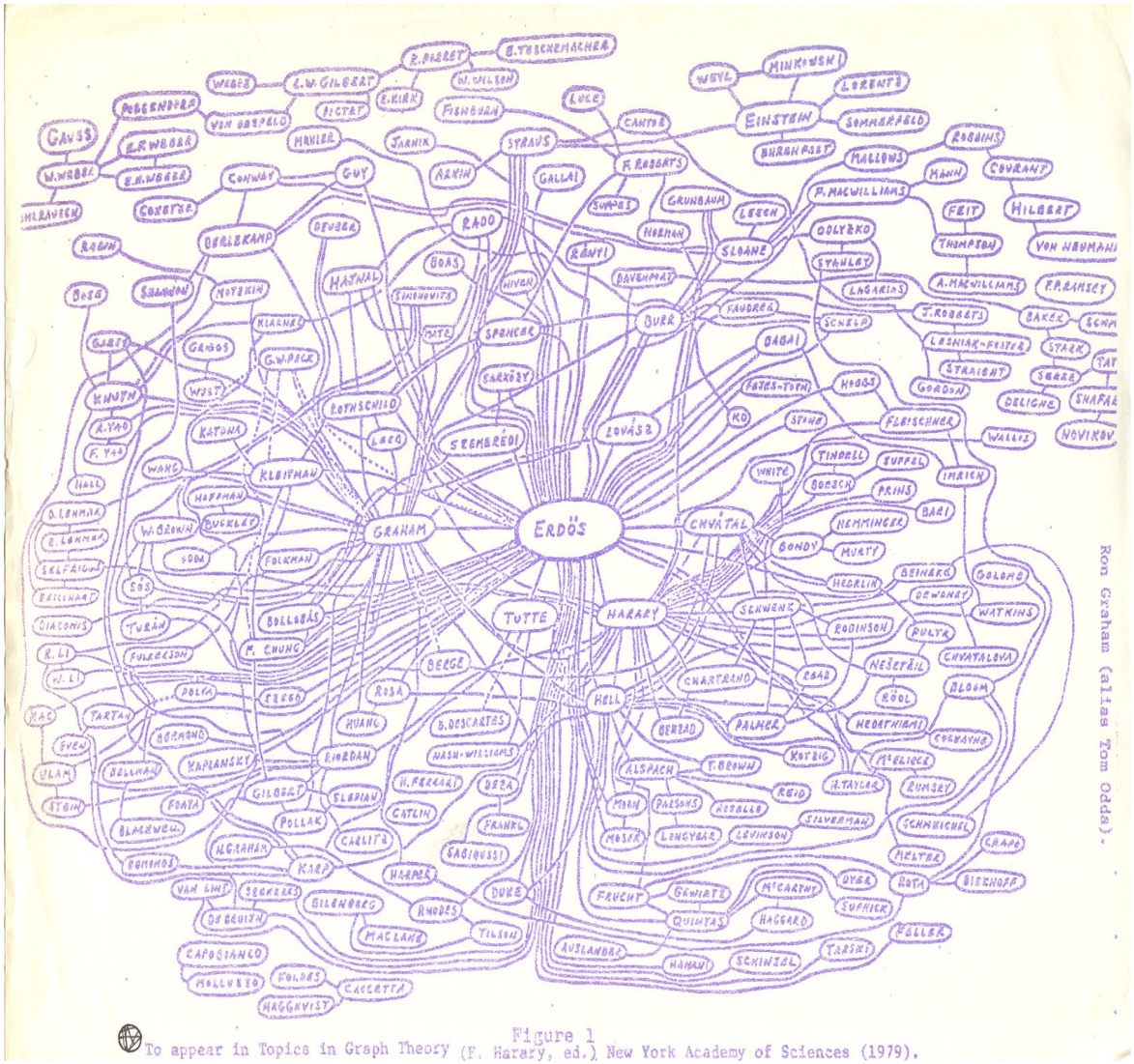
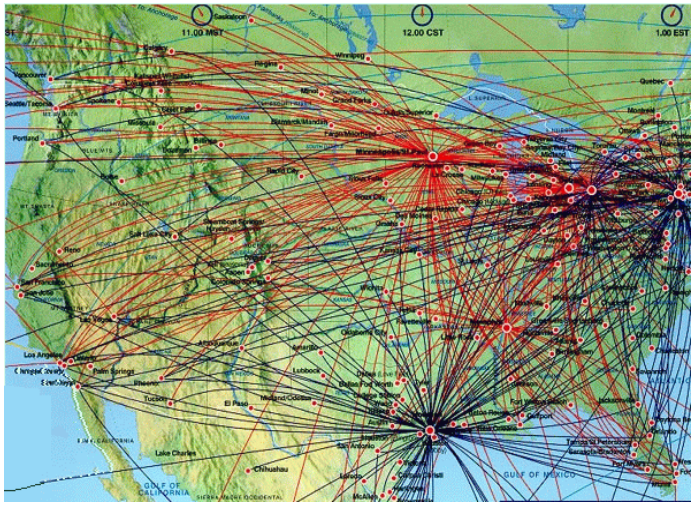


Figure 1
To appear in Topics in Graph Theory (F. Harary, ed.) New York Academy of Sciences (1979).

Connectivity can be an important variable for identifying important nodes

Flight connections and hub airports



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

****Slide courtesy of AL Barabasi**

Hub genes with respect to the whole network are often uninteresting (especially in coexpression networks)…

- but genes with high connectivity in interesting modules can be very interesting.
- Citations:
  1) PNAS 2006 PMC1635024
  2) Langfelder et al (2013) When Is Hub Gene Selection Better than Standard Meta-Analysis? PLoS ONE 8(4): e61505.

Define 2 alternative measures of intramodular connectivity for finding intramodular hubs.

Intramodular connectivity kIN

- Row sum across genes inside a given module

$$kIN(i) = \sum_{j \in ModuleSet} a_{ij}$$

- Advantages: defined for any network based on adjacency matrix.
- Disadvantage: strong depends on module size

Module eigengene based connectivity, kME, also known as module membership measure

$$kME_i = ModuleMembership(i) = cor(x_i, ME)$$

- kME(i) is simply the correlation between the i-th gene expression profile and the module eigengene.
- kME close to 1 means that the gene is a hub gene
- Very useful measure for annotating genes with regard to modules.
- Can be used to find genes that are members of two or more modules (fuzzy clustering).
- Module eigengene can be interpreted as the most highly connected gene.

PloS Computational Biology. 4(8): e1000117. PMID:18704157

"Group conform behavior leads to a lot of friends."

When dealing with a network comprised of module genes,

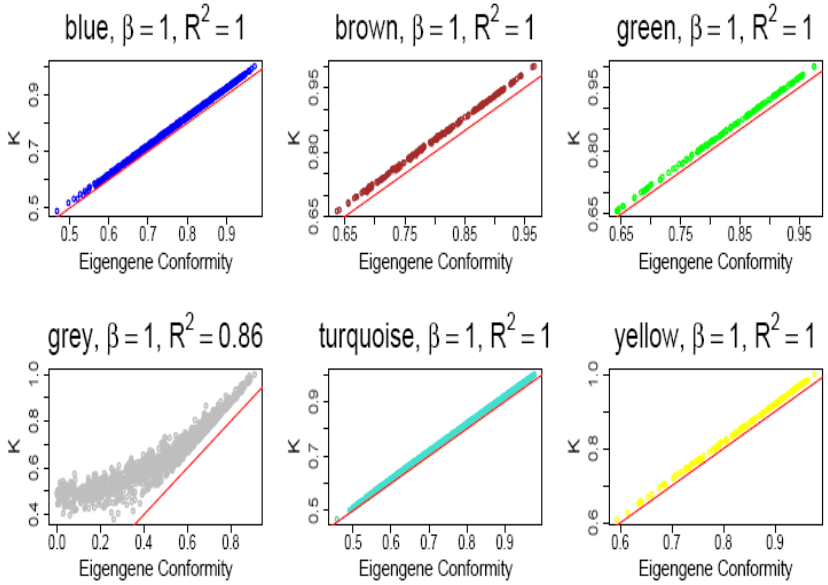
the scaled intramodular connectivity is determined by kME

$$\frac{\mathrm{kIM}_i}{\max(\mathrm{kIM})} \approx |\, cor(x_i, E)\,|^{\beta} = |\, kME(i)\,|^{\beta} \ .$$

where $|\, kME(i)\,|^{\beta}$ measures group conform behavior

Derivation requires an unsigned weighted correlation network

PLoS Comput Biol 4(8): e1000117

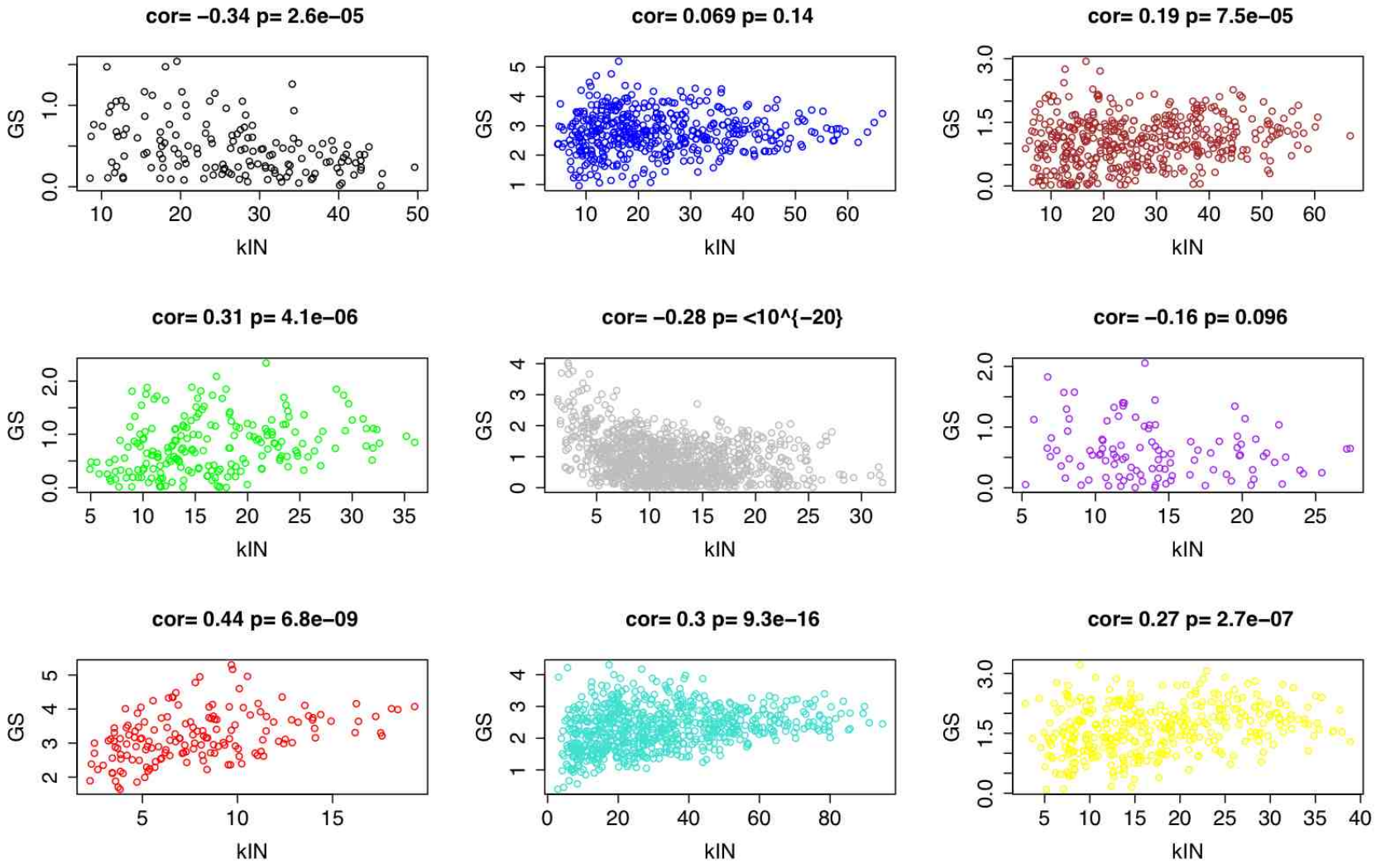


Intramodular hub genes

- Defined as genes with high kME (or high kIM)
- Single network analysis: Intramodular hubs in biologically interesting modules are often very interesting
- Differential network analysis: Genes that are intramodular hubs in one condition but not in another are often very interesting

How to use networks for gene screening?

Gene significance versus intramodular connectivity kIN



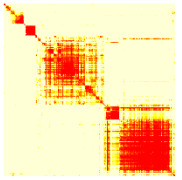
Intramodular connectivity versus gene significance GS

- Note the relatively high correlation between gene significance and intramodular connectivity in some modules
- In practice, a combination of GS and intramodular connectivity is used to select important hub genes.
- Module eigengene turns out to be the most highly connected gene (under mild assumptions)

What is weighted gene co-expression network analysis?

Construct a network

Rationale: make use of interaction patterns between genes



Identify modules

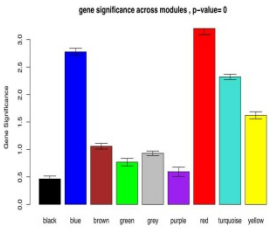
Rationale: module (pathway) based analysis

Relate modules to external information

Array Information: Clinical data, SNPs, proteomics

Gene Information: gene ontology, EASE, IPA

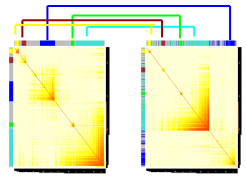
Rationale: find biologically interesting modules



Study Module Preservation across different data

Rationale:

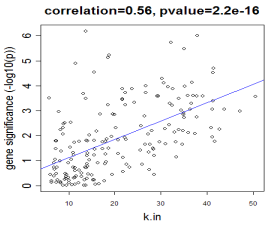
- Same data: to check robustness of module definition
- Different data: to find interesting modules.



Find the key drivers in *interesting* modules

Tools: intramodular connectivity, causality testing

Rationale: experimental validation, therapeutics, biomarkers



What is different from other analyses?

- **Emphasis on modules instead of individual genes**
  - Greatly alleviates the problem of multiple comparisons
- Use of intramodular connectivity to find key drivers
  - Quantifies module membership
  - Module definition is only based on interconnectedness
  - No prior pathway information is used for module definition
  - Two module (eigengenes) can be highly correlated
- Emphasis on a unified approach for relating variables
  - Default: correlation (biweight midcorrelation)
  - Rationale:
    - puts different data sets on the same mathematical footing
- Technical Details: soft thresholding with the power adjacency function, topological overlap matrix to measure interconnectedness

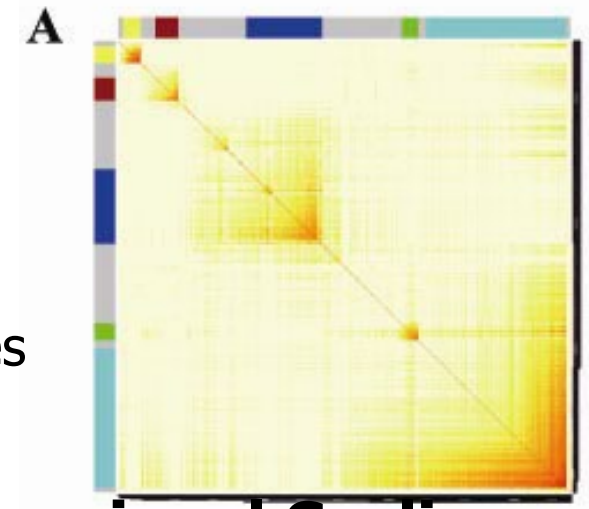
Case Study 1:
Finding brain cancer genes

Horvath S, Zhang B, Carlson M, Lu KV, Zhu S, Felciano RM, Laurance MF, Zhao W, Shu, Q, Lee Y, Scheck AC, Liau LM, Wu H, Geschwind DH, Febbo PG, Kornblum HI, Cloughesy TF, Nelson SF, Mischel PS (2006) "Analysis of Oncogenic Signaling Networks in Glioblastoma Identifies ASPM as a Novel Molecular Target", PNAS | November 14, 2006 | vol. 103 | no. 46

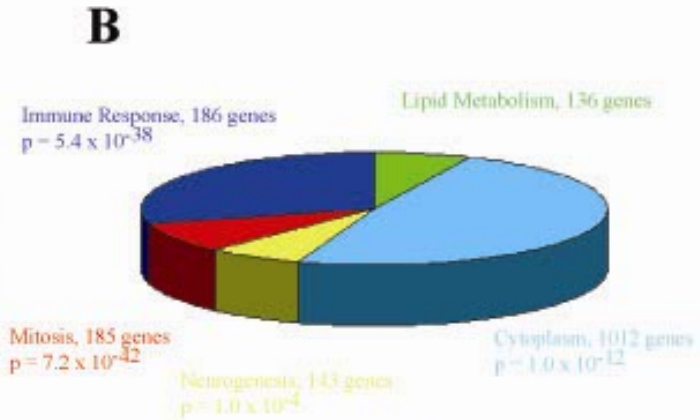
Different Ways of Depicting Gene Modules

Topological Overlap Plot

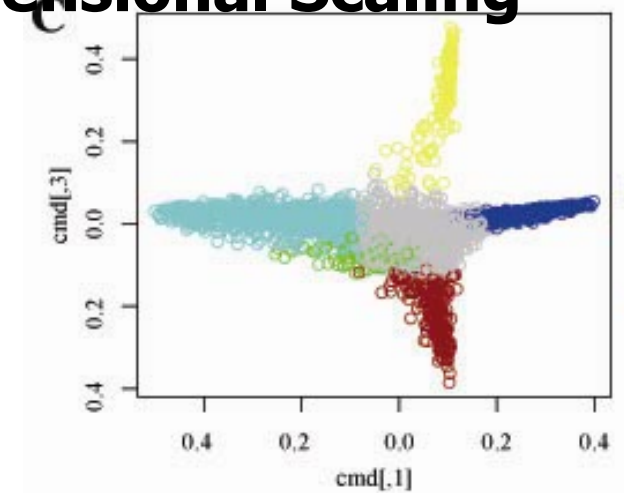
1) Rows and columns correspond to genes
2) Red boxes along diagonal are modules
3) Color bands=modules



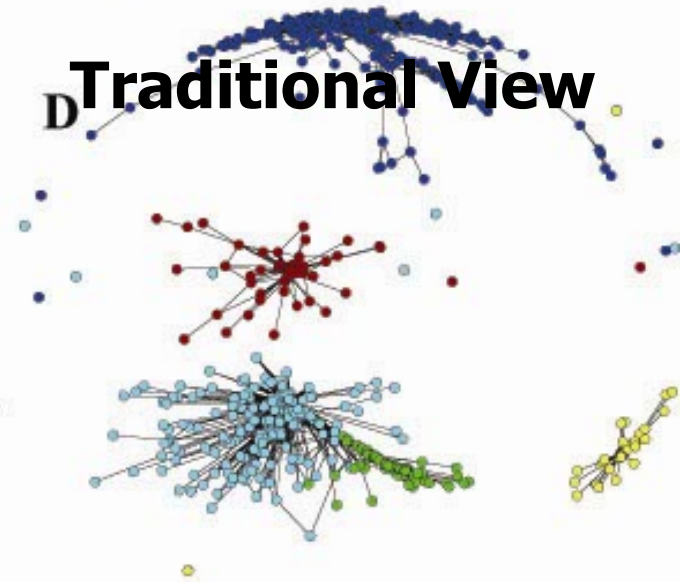
Gene Functions



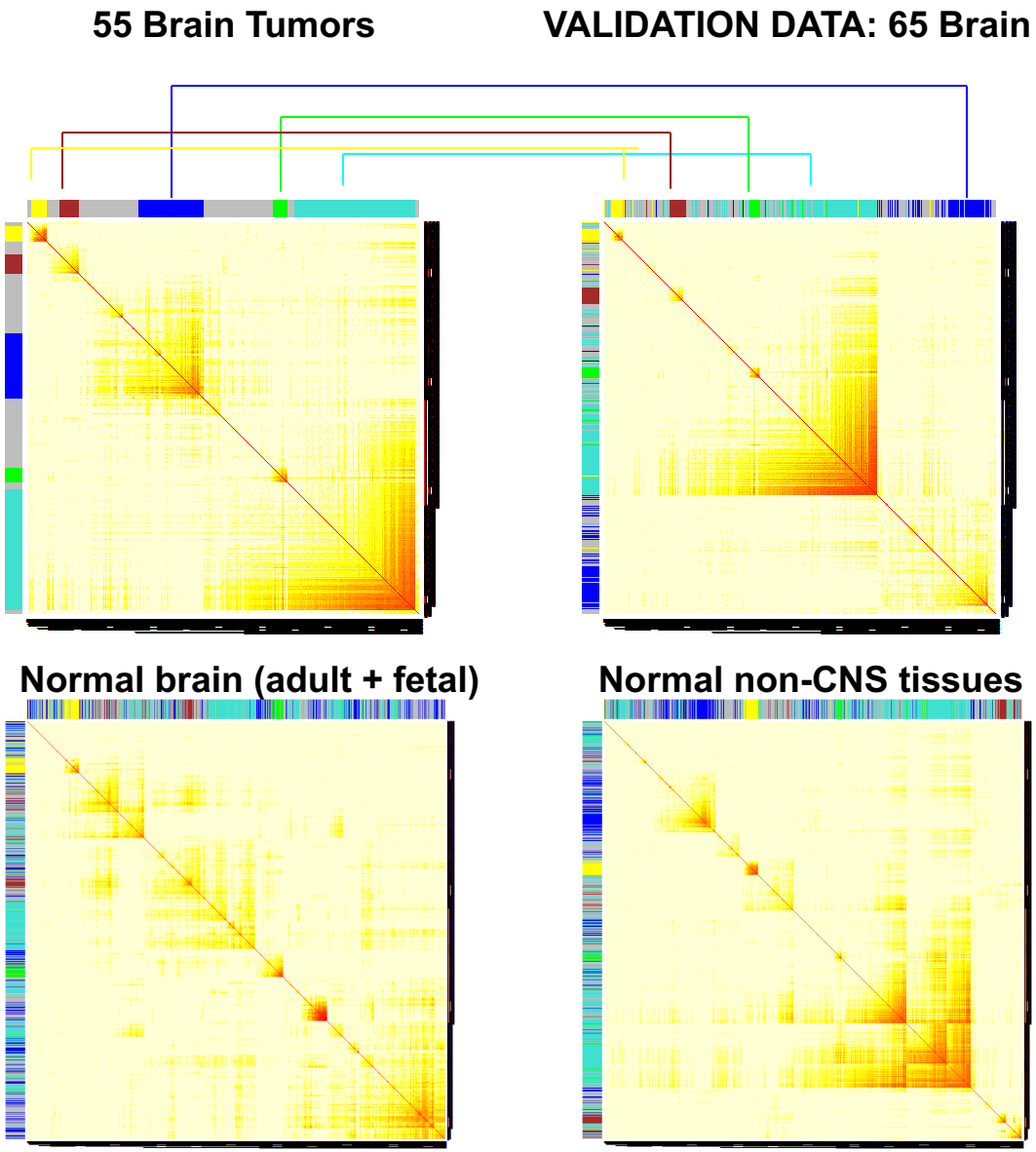
Multi Dimensional Scaling



Traditional View

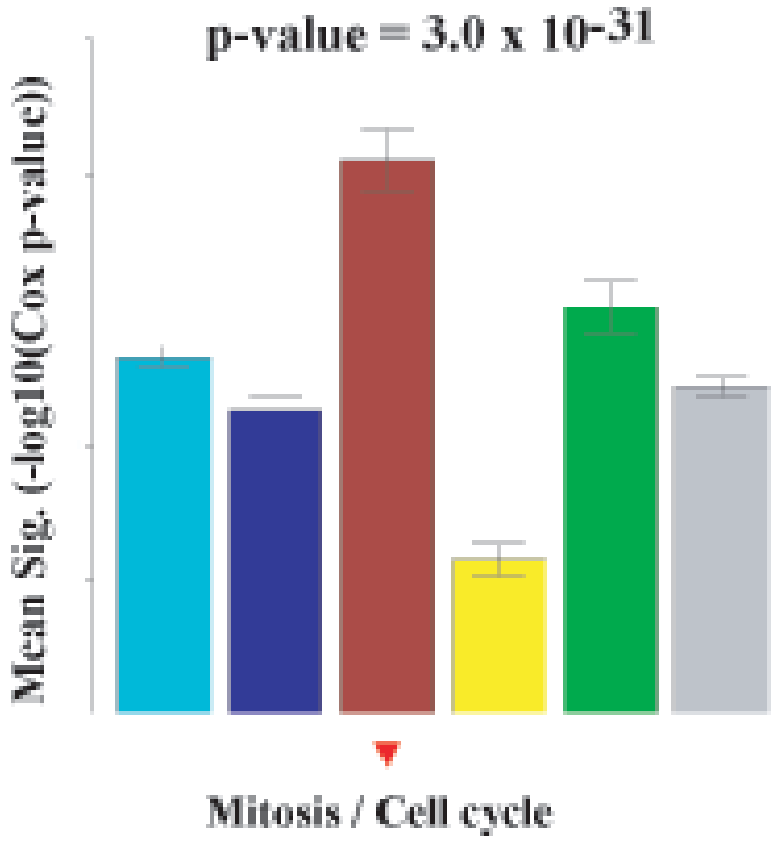


Comparing the Module Structure in Cancer and Normal tissues



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

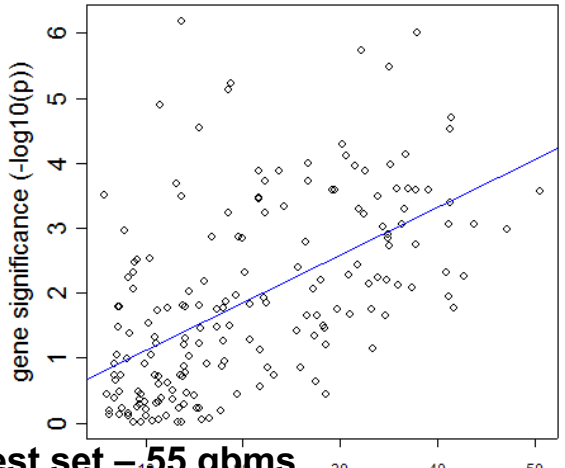
Mean Prognostic Significance of Module Genes



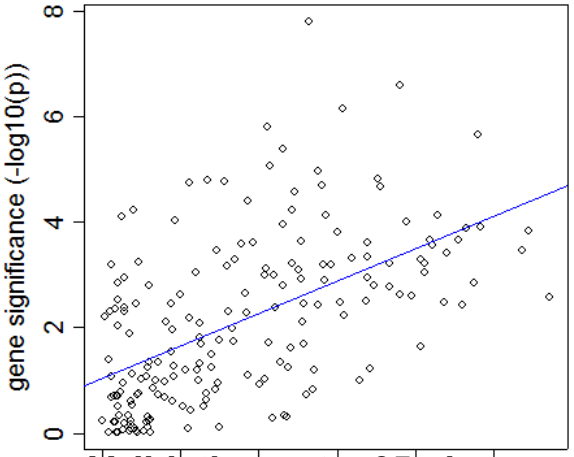
Message: Focus the attention on the brown module genes

Module hub genes predict cancer survival

1. Cox model to regress survival on gene expression levels
2. Defined prognostic significance as –log10(Cox-p-value) the survival association between each gene and glioblastoma patient survival
3. *A module-based measure of gene connectivity significantly and reproducibly identifies the genes that most strongly predict patient survival*



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



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

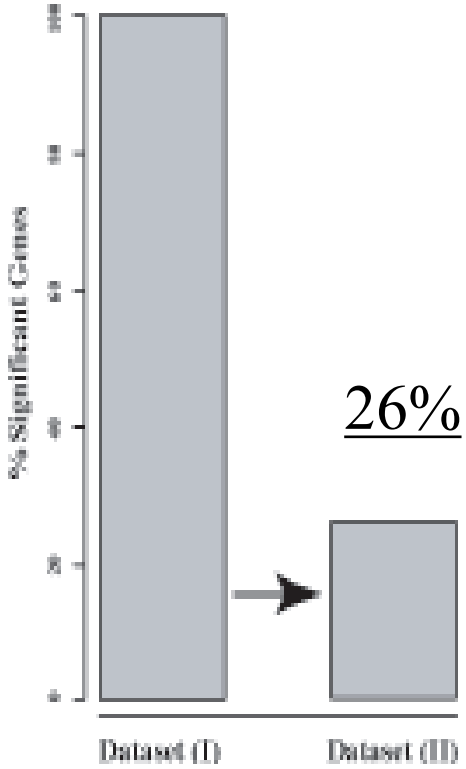
The fact that genes with high intramodular connectivity are more likely to be prognostically significant facilitates a novel screening strategy for finding prognostic genes

- 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 disease related module

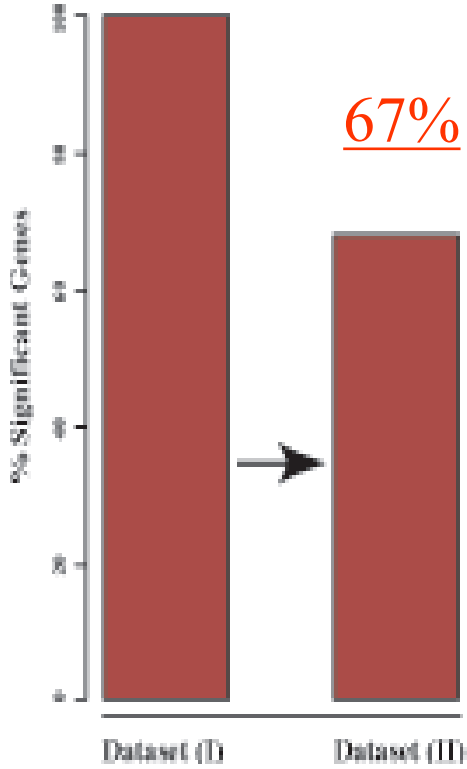
- Validation success rate= proportion of genes with independent test set Cox regression p-value<0.05.
- Validation success rate of network based screening approach (68%)
- Standard approach involving top 300 most significant genes: 26%

Validation success rate of gene expressions in independent data

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



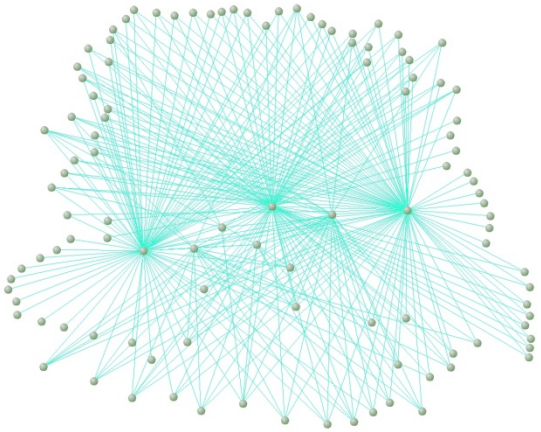
Network based screening
p<0.05 and
high intramodular connectivity



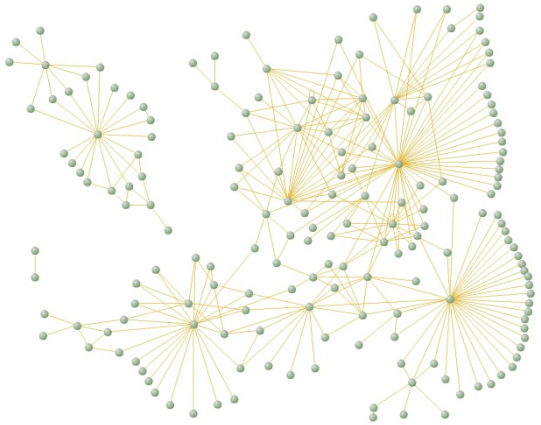
The network-based approach uncovers novel therapeutic targets

Five of the top six hub genes in the mitosis module are already known cancer targets: topoisomerase II,
Rac1, TPX2, EZH2 and KIF14.
We hypothesized that the 6-th gene ASPM gene is novel therapeutic target. ASPM encodes the human ortholog of a drosophila mitotic spindle protein.

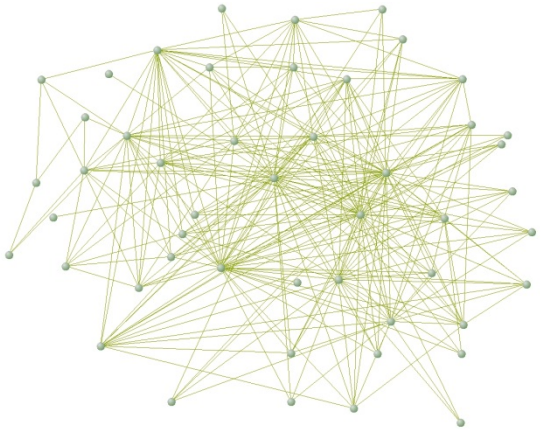
Biological validation: siRNA mediated inhibition of ASPM

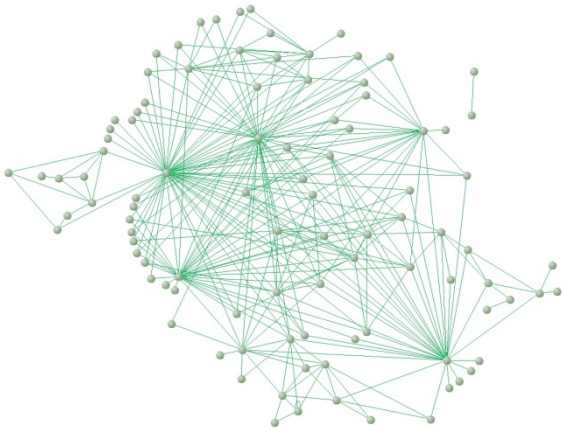


Case Study 2



MC Oldham, S Horvath, DH Geschwind (2006) Conservation and evolution of gene co-expression networks in human and chimpanzee brain. PNAS





What changed?

- Despite pronounced phenotypic differences, genomic similarity is ~96% (including single-base substitutions and indels)[1]
  - Similarity is even higher in protein-coding regions

[1] Cheng, Z. *et al. Nature* **437**, 88-93 (2005)

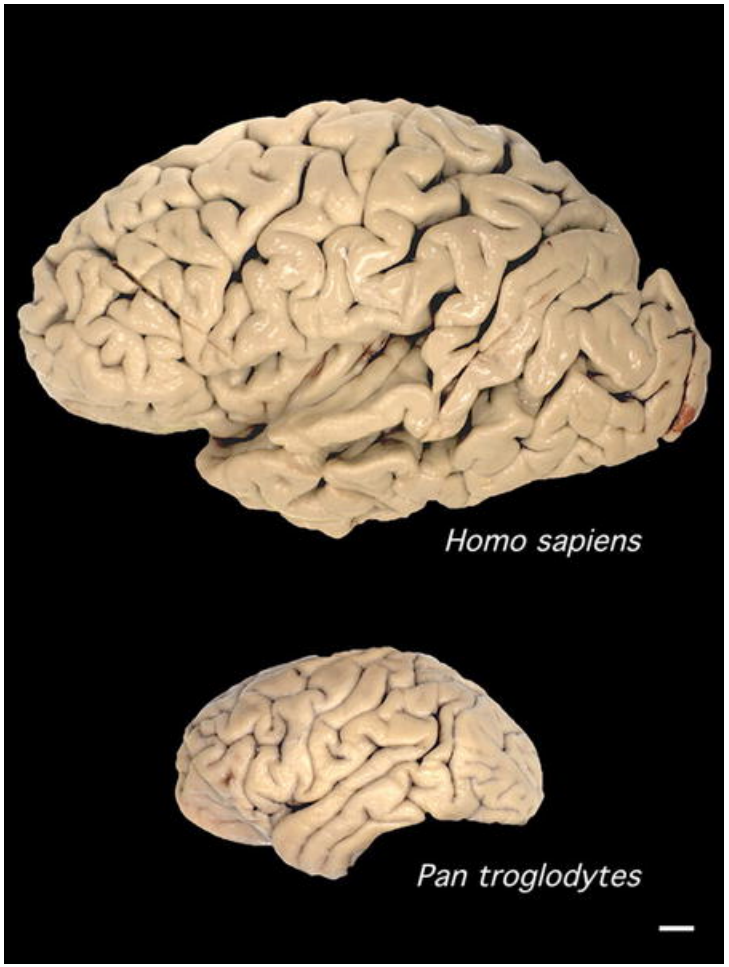


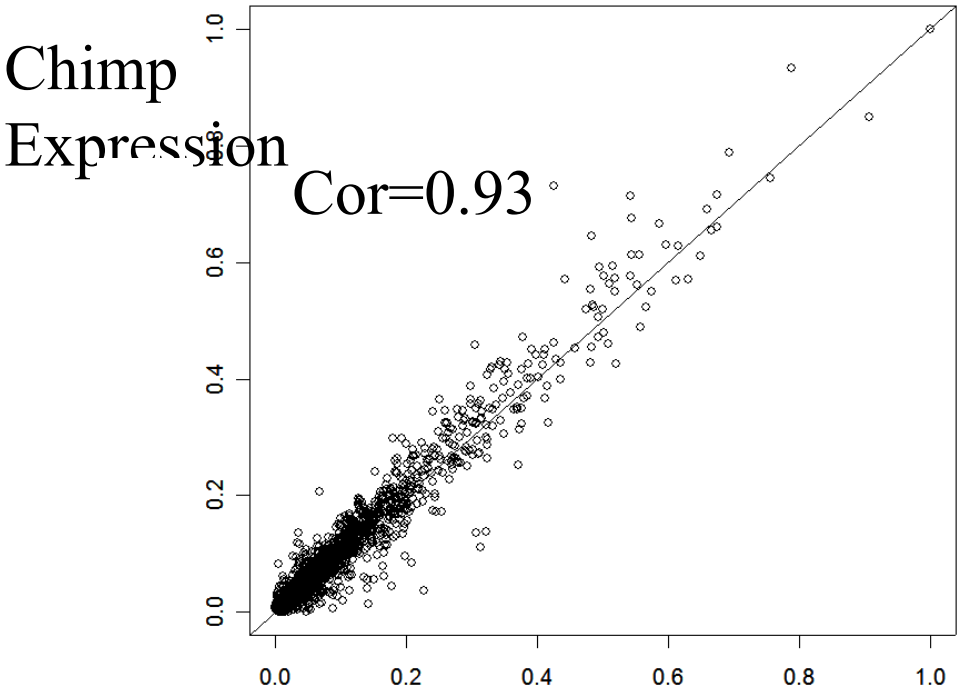
Image courtesy of Todd Preuss (Yerkes National Primate Research Center)

Assessing the contribution of regulatory changes to human evolution

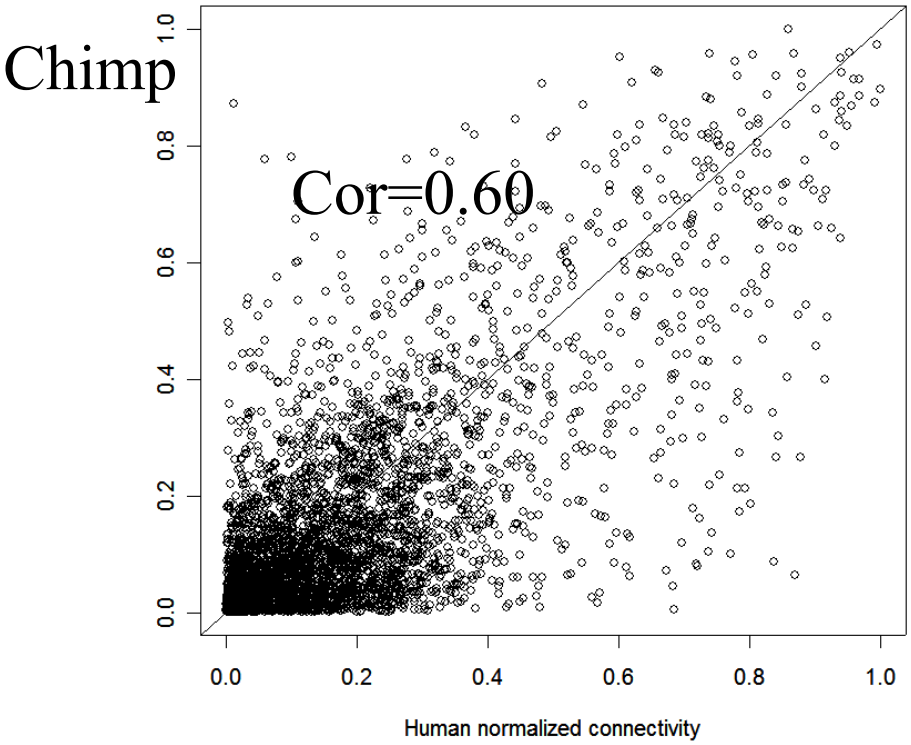
- Hypothesis: Changes in the regulation of gene expression were critical during recent human evolution (King & Wilson, 1975)

- Microarrays are ideally suited to test this hypothesis by comparing expression levels for thousands of genes simultaneously

Gene expression is more strongly preserved than gene connectivity



Human Expression

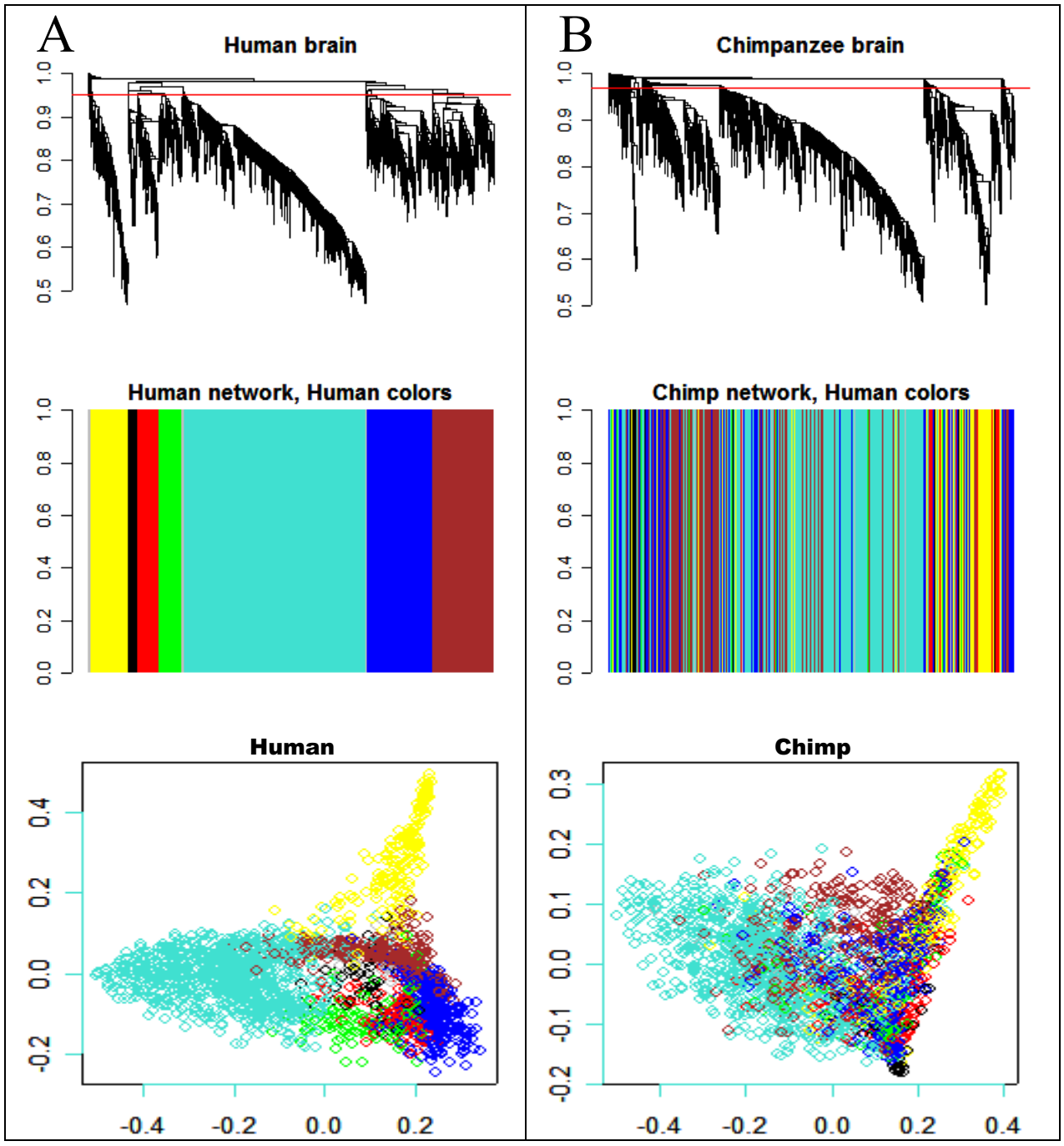


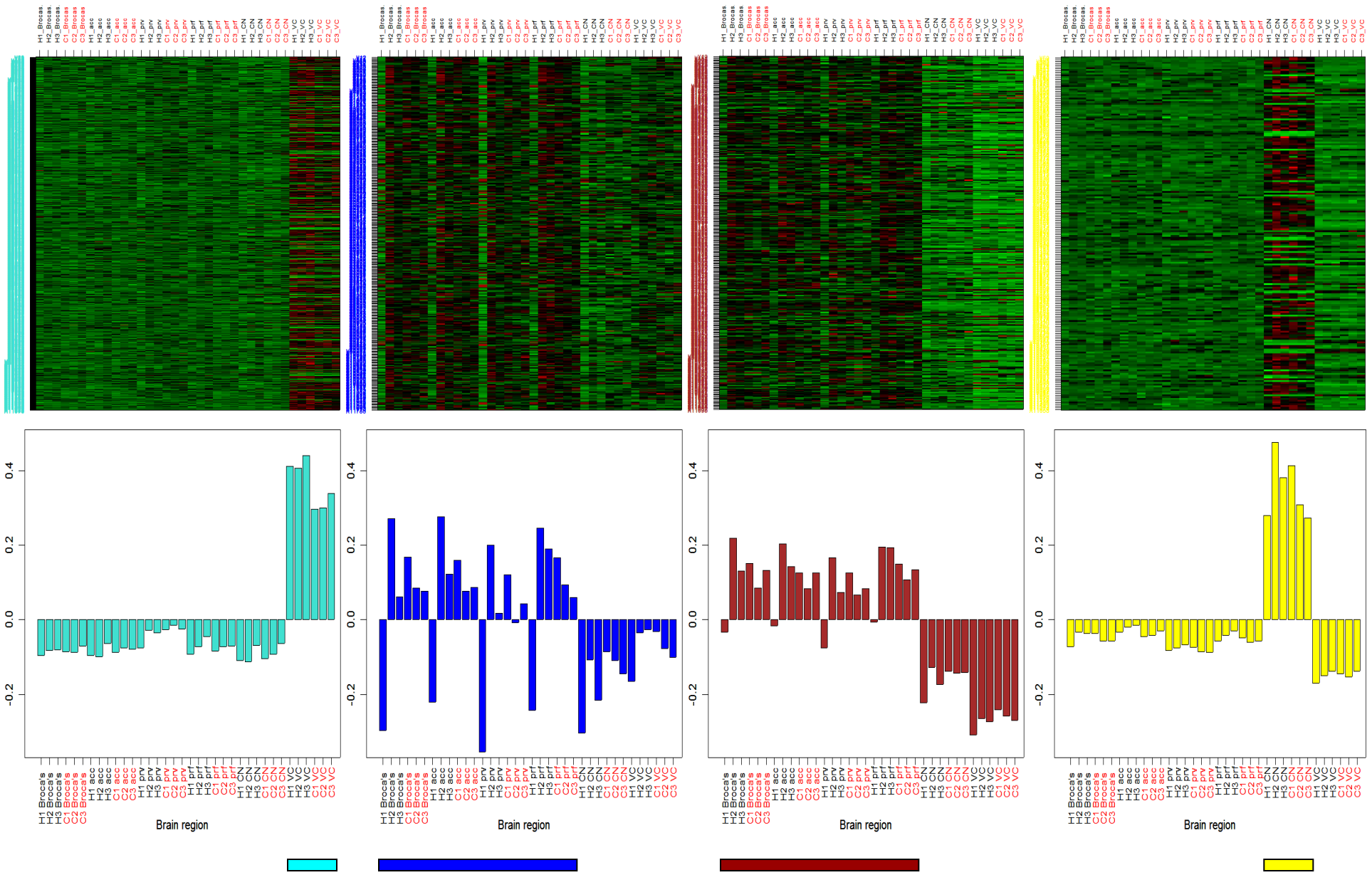
Human Connectivity

Hypothesis: molecular wiring makes us human

Raw data from Khaitovich *et al.*, 2004

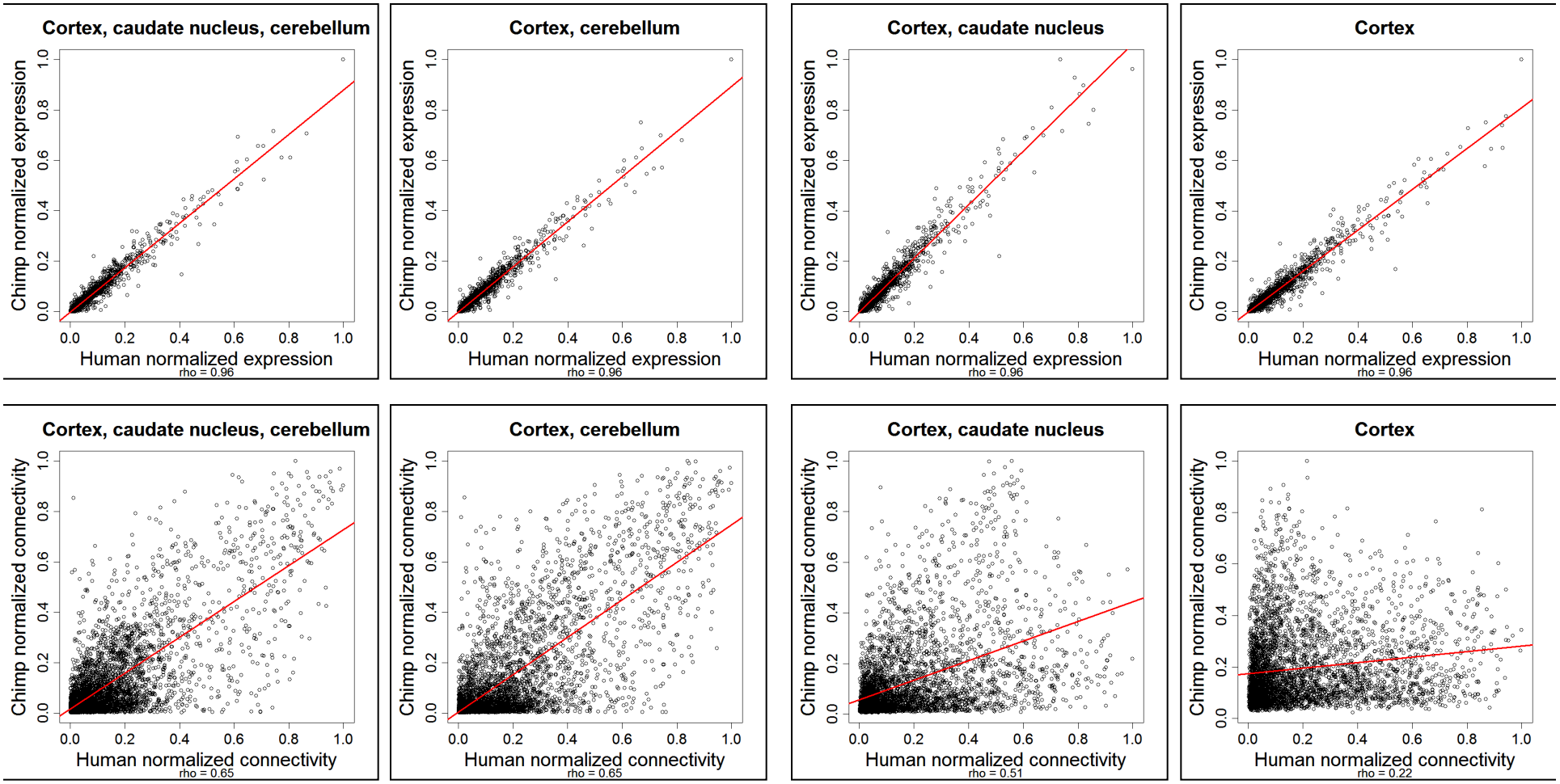
Mike Oldham


A
Human brain
1.0
0.9
0.8
0.7
0.6
0.5
Human network, Human colors
1.0
0.8
0.6
0.4
0.2
0.0
Human
0.4
0.2
0.0
-0.2
-0.4
-0.2
0.0
0.2
B
Chimpanzee brain
1.0
0.9
0.8
0.7
0.6
0.5
Chimp network, Human colors
1.0
0.8
0.6
0.4
0.2
0.0
Chimp
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.4
-0.2
0.0
0.2
0.4



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

Connectivity diverges across brain regions whereas expression does not



Conclusions: chimp/human

- Gene expression is highly preserved across species brains
- Gene co-expression is less preserved
- Some modules are highly preserved
- Gene modules correspond roughly to brain architecture
- Species-specific hubs can be validated in silico using sequence comparisons

A short methodological summary of the publications.

- WGCNA methods
  - Horvath S (2011) Weighted Network Analysis. Applications in Genomics and Systems Biology. Springer Book. ISBN: 978-1-4419-8818-8
  - Zhang B, Horvath S (2005) "A General Framework for Weighted Gene Co-Expression Network Analysis", Statistical Applications in Genetics and Molecular Biology: Vol. 4: No. 1, Article 17
  - Langfelder P, Horvath S (2008) WGCNA: an R package for Weighted Correlation Network Analysis. BMC Bioinformatics. 2008 Dec 29;9(1):559. PMID: 19114008 PMCID: PMC2631488
  - Langfelder P et al (2011) Is my network module preserved and reproducible? PloS Comp Biol. 7(1): e1001057. PMID: 21283776
- Math and WGCNA:
  - Horvath S, Dong J (2008) Geometric Interpretation of Gene Co-Expression Network Analysis. PloS Computational Biology. 4(8): e1000117. PMID: 18704157
- Empirical evaluation of WGCNA
  - Langfelder P, et al (2013) When Is Hub Gene Selection Better than Standard Meta-Analysis? PLoS ONE 8(4): e61505.
  - Song L, Langfelder P, Horvath S. (2012) Comparison of co-expression measures: mutual information, correlation, and model based indices.BMC Bioinformatics;13(1):328. PMID: 23217028
- What is the topological overlap measure? Empirical studies of the robustness of the topological overlap measure:
  - Yip A, Horvath S (2007) Gene network interconnectedness and the generalized topological overlap measure. BMC Bioinformatics 8:22
- Dynamic branch cutting:
  - Langfelder P, Zhang B, Horvath S (2008) Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. Bioinformatics.;24(5):719-20. PMID: 18024473
- Gene screening based on intramodular connectivity identifies brain cancer genes that validate.
  - Horvath S, Zhang B, Carlson M, Lu KV, Zhu S, Felciano RM, Laurance MF, Zhao W, Shu, Q, Lee Y, Scheck AC, Liau LM, Wu H, Geschwind DH, Febbo PG, Kornblum HI, Cloughesy TF, Nelson SF, Mischel PS (2006) "Analysis of Oncogenic Signaling Networks in Glioblastoma Identifies ASPM as a Novel Molecular Target", PNAS | November 14, 2006 | vol. 103 | no. 46 | 17402-17407
- How to integrate SNP markers into weighted gene co-expression network analysis?
  - Plaisier CL et al Pajukanta P (2009) A systems genetics approach implicates USF1, FADS3 and other causal candidate genes for familial combined hyperlipidemia. PloS Genetics;5(9):e1000642 PMID: 19750004
- Differential network analysis:
  - Fuller TF, Ghazalpour A, Aten JE, Drake TA, Lusis AJ, Horvath S (2007) "Weighted Gene Co-expression Network Analysis Strategies Applied to Mouse Weight", Mammalian Genome.
- Neuroscience
  - Oldham M, Horvath S, Geschwind D (2006) Conservation and Evolution of Gene Co-expression Networks in Human and Chimpanzee Brains. 2006 Nov 21;103(47):17973-8
  - Oldham MC, Konopka G, Iwamoto K, Langfelder P, Kato T, Horvath S, Geschwind DH (2008) Functional organization of the transcriptome in human brain. Nature Neuroscience. 11(11):1271-82. PMID: 18849986
  - Voineagu I, Wang X, Johnston P, Lowe JK, Tian Y, Horvath S, Mill J, Cantor R, Blencowe BJ, Geschwind DH (2011) Transcriptomic analysis of autistic brain reveals convergent molecular pathology. Nature. 25;474(7351):380-4 PMID: 21614001

THE END