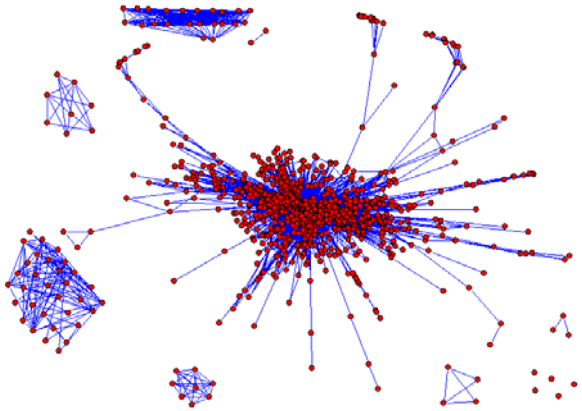


An Overview of Weighted Gene Co-Expression Network Analysis

Steve Horvath

University of California, Los Angeles

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)?
- What is neighborhood analysis?

Philosophy of Weighted Gene Co-Expression Network Analysis

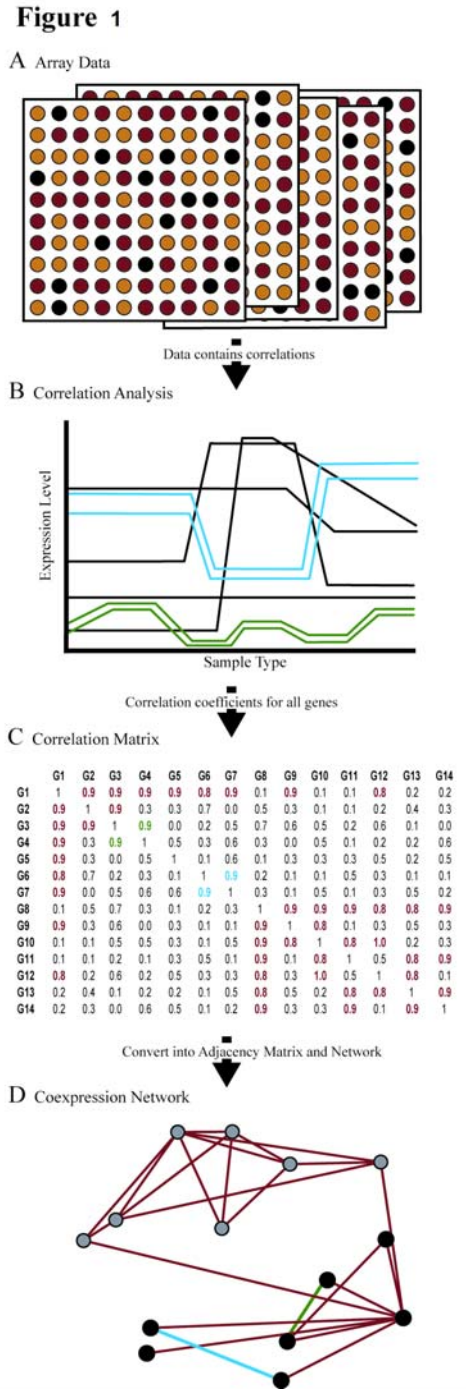
- Understand the “system” instead of reporting a list of individual parts
  - Describe the functioning of the engine instead of enumerating individual nuts and bolts
- Focus on modules as opposed to individual genes
  - this greatly alleviates multiple testing problem
- Network terminology is intuitive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

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) Microarray gene expression data
B) Measure concordance of gene expression with a Pearson correlation
C) The Pearson correlation matrix is either dichotomized to arrive at an adjacency matrix → unweighted network

Or transformed continuously with the power adjacency function → weighted network

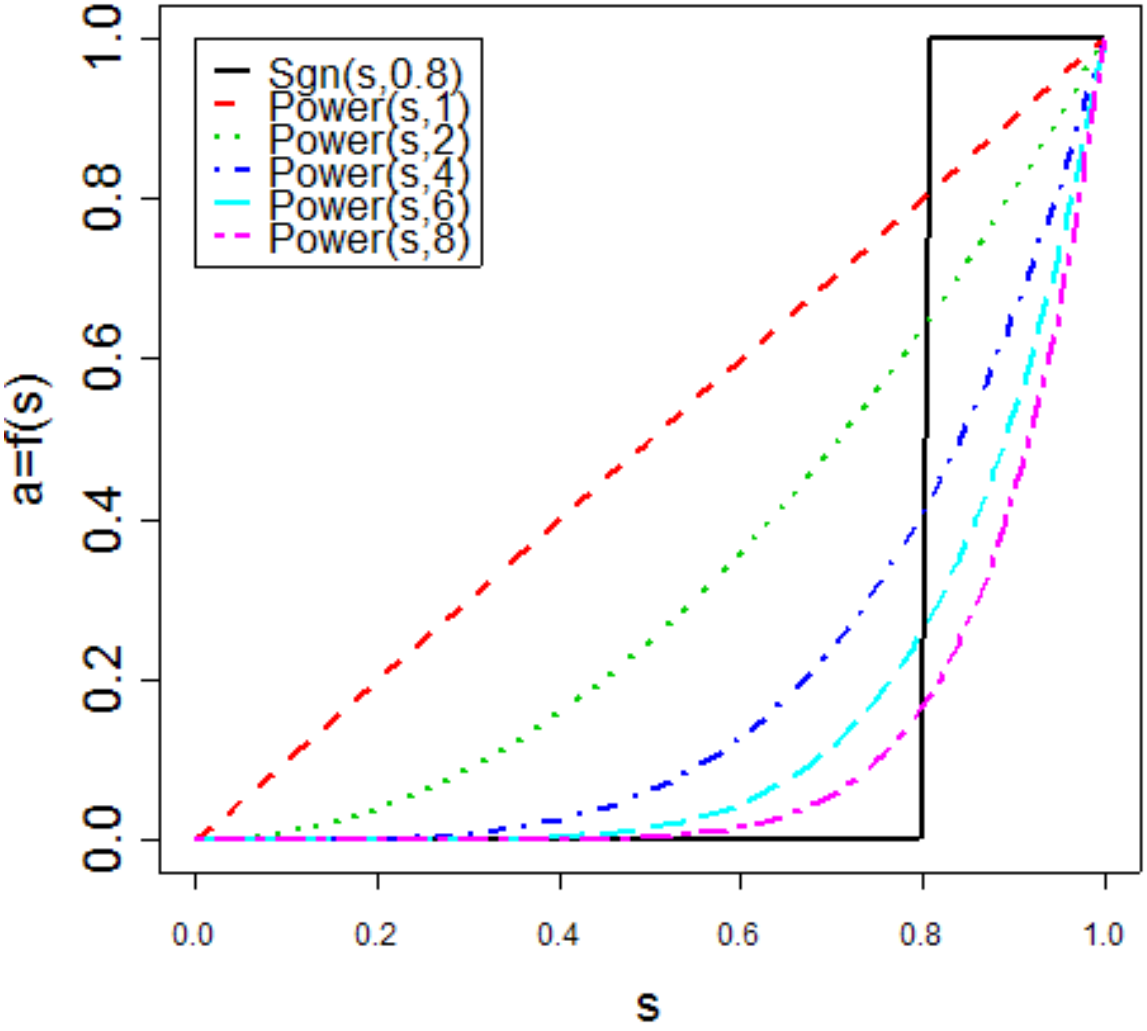
Power adjacency function results in a weighted gene network

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

Often choosing beta=6 works well but in general we use the "scale free topology criterion" described in Zhang and Horvath 2005.

Comparing adjacency functions

Power Adjancy vs Step Function



Comparing the power adjacency function to the step function

- While the network analysis results are usually highly robust with respect to the network construction method there are several reasons for preferring the power adjacency function.
  - Empirical finding: Network results are highly robust with respect to the choice of the power beta
    - Zhang B and Horvath S (2005)
  - Theoretical finding: Network Concepts make more sense in terms of the module eigengene.
    - Horvath S, Dong J (2008) Geometric Interpretation of Gene Co-Expression Network Analysis. PloS Computational Biology

How to detect network modules?

Module Definition

- Numerous methods have been developed
- Here, we use average linkage hierarchical clustering coupled with the topological overlap dissimilarity measure.
- Once a dendrogram is obtained from a hierarchical clustering method, we choose a height cutoff to arrive at a clustering.
- Modules correspond to branches of the dendrogram

The topological overlap dissimilarity is used as input of hierarchical clustering

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

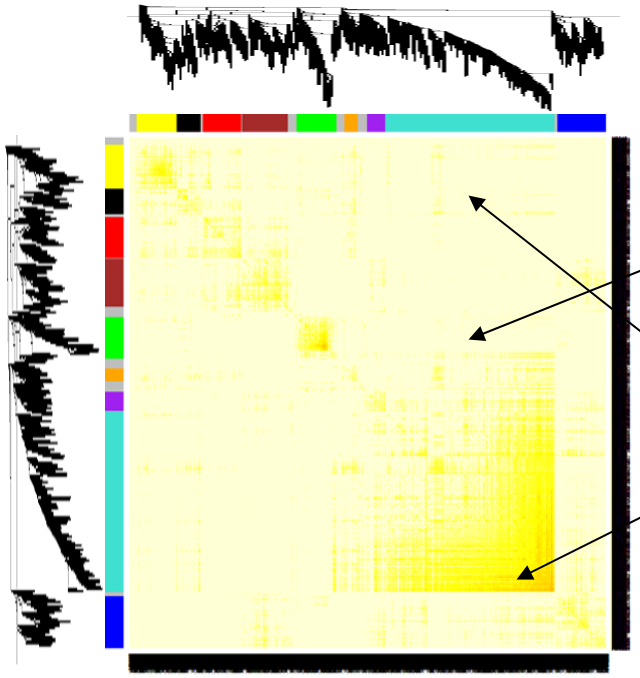
$$DistTOM_{ij} = 1 - TOM_{ij}$$

- Generalized in Zhang and Horvath (2005) to the case of weighted networks
- Generalized in Yip and Horvath (2006) to higher order interactions

Using the topological overlap matrix (TOM) to cluster genes

- Here modules correspond to branches of the dendrogram

TOM plot



Genes correspond to rows and columns

Hierarchical clustering dendrogram

TOM matrix

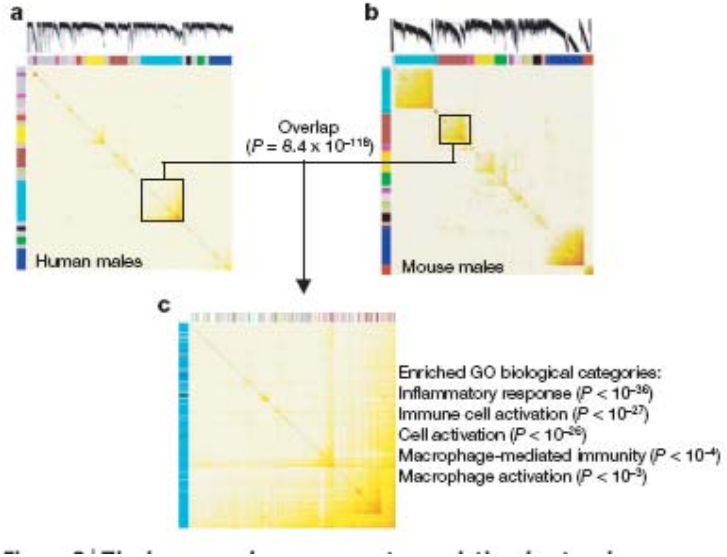
Module: Correspond to branches

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

ARTICLES

Genetics of gene expression and its effect on disease

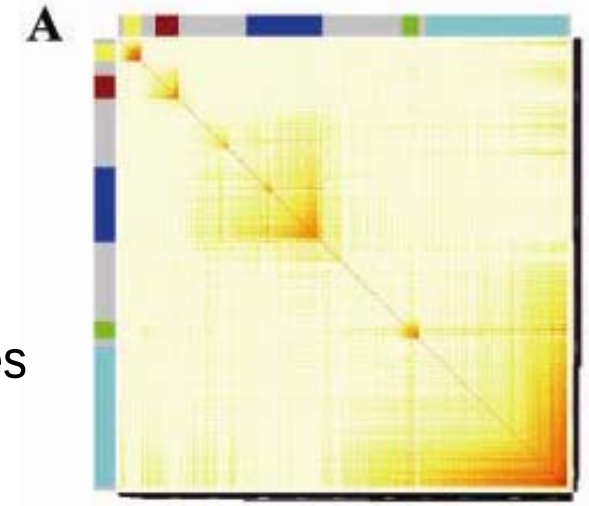
Valur Emilsson[1,2], Gudmar Thorleifsson[1], Bin Zhang[2], Amy S. Leonardson[2], Florian Zink[1], Jun Zhu[2], Sonia Carlson[2], Agnar Helgason[1], G. Bragi Walters[1], Steinunn Gunnarsdottir[1], Magali Mouy[1], Valgerdur Steinthorsdottir[1], Gudrun H. Eiriksdottir[1], Gyda Bjornsdottir[1], Inga Reynisdottir[1], Daniel Gudbjartsson[1], Anna Helgadottir[1], Aslaug Jonasdottir[1], Adalbjorg Jonasdottir[1], Unnur Styrkarsdottir[1], Solveig Gretarsdottir[1], Kristinn P. Magnusson[1], Hreinn Stefansson[1], Ragnheidur Fossdal[1], Kristleifur Kristjansson[1], Hjortur G. Gislason[3], Tryggvi Stefansson[3], Bjorn G. Leifsson[3], Unnur Thorsteinsdottir[1], John R. Lamb[2], Jeffrey R. Gulcher[1], Marc L. Reitman[4], Augustine Kong[1], Eric E. Schadt[2]* & Kari Stefansson[1]*



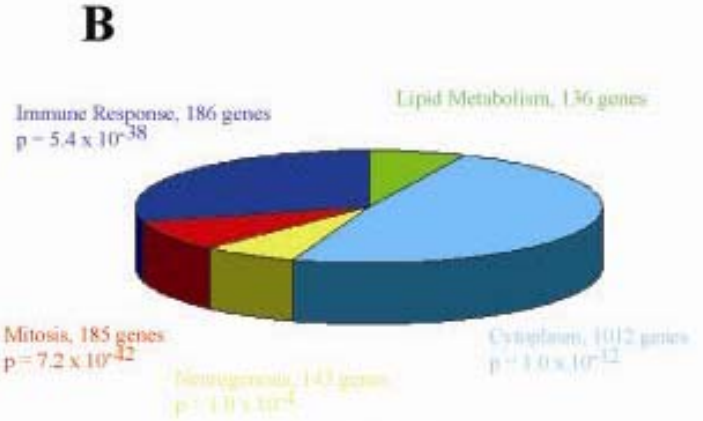
Different Ways of Depicting Gene Modules

Topological Overlap Plot

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

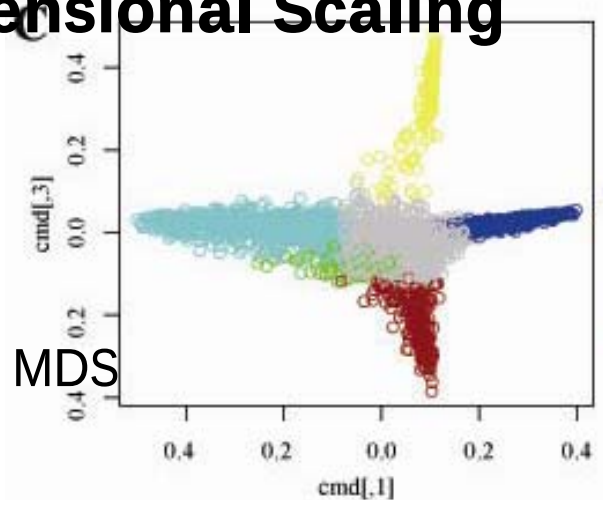


Gene Functions

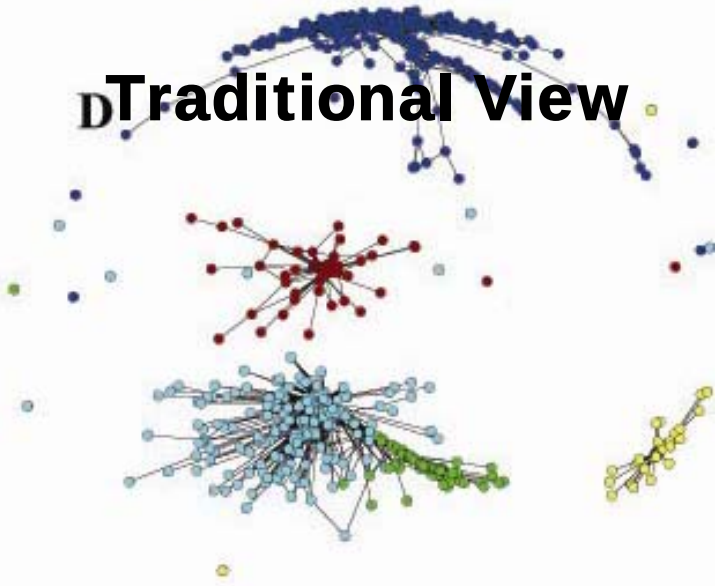


Multi Dimensional Scaling

Idea:
Use network distance in MDS



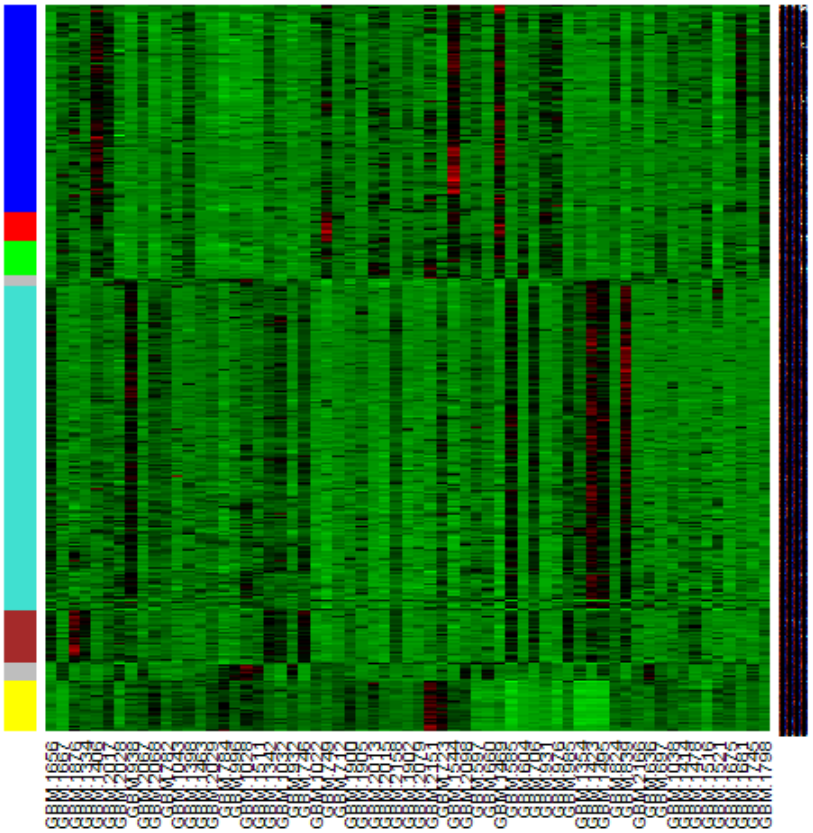
Traditional View



Heatmap view of module

Columns= tissue samples

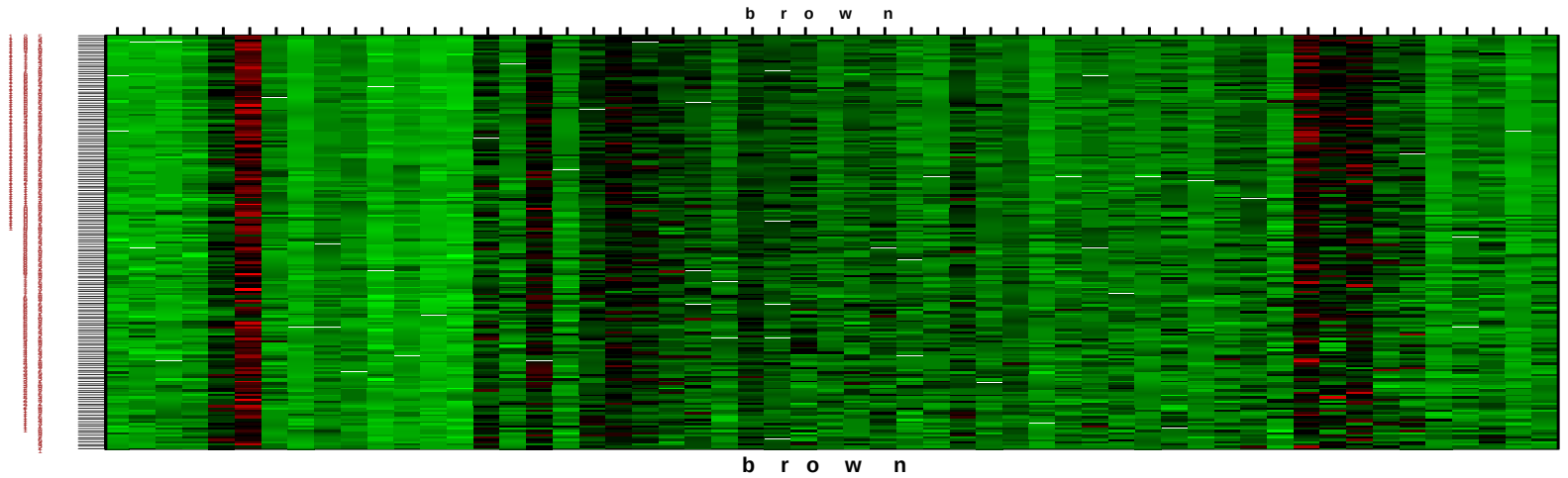
Rows=Genes
Color band indicates module membership



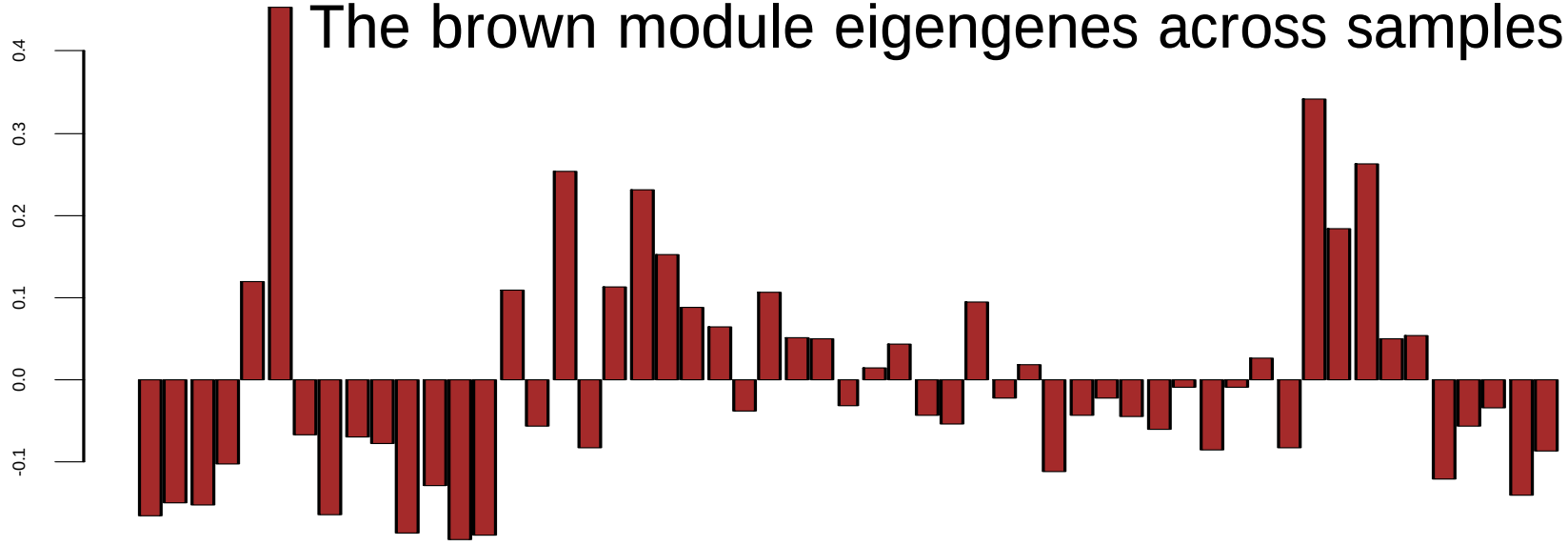
Message: characteristic vertical bands indicate tight co-expression of module genes

Module Eigengene= measure of over-expression=average redness

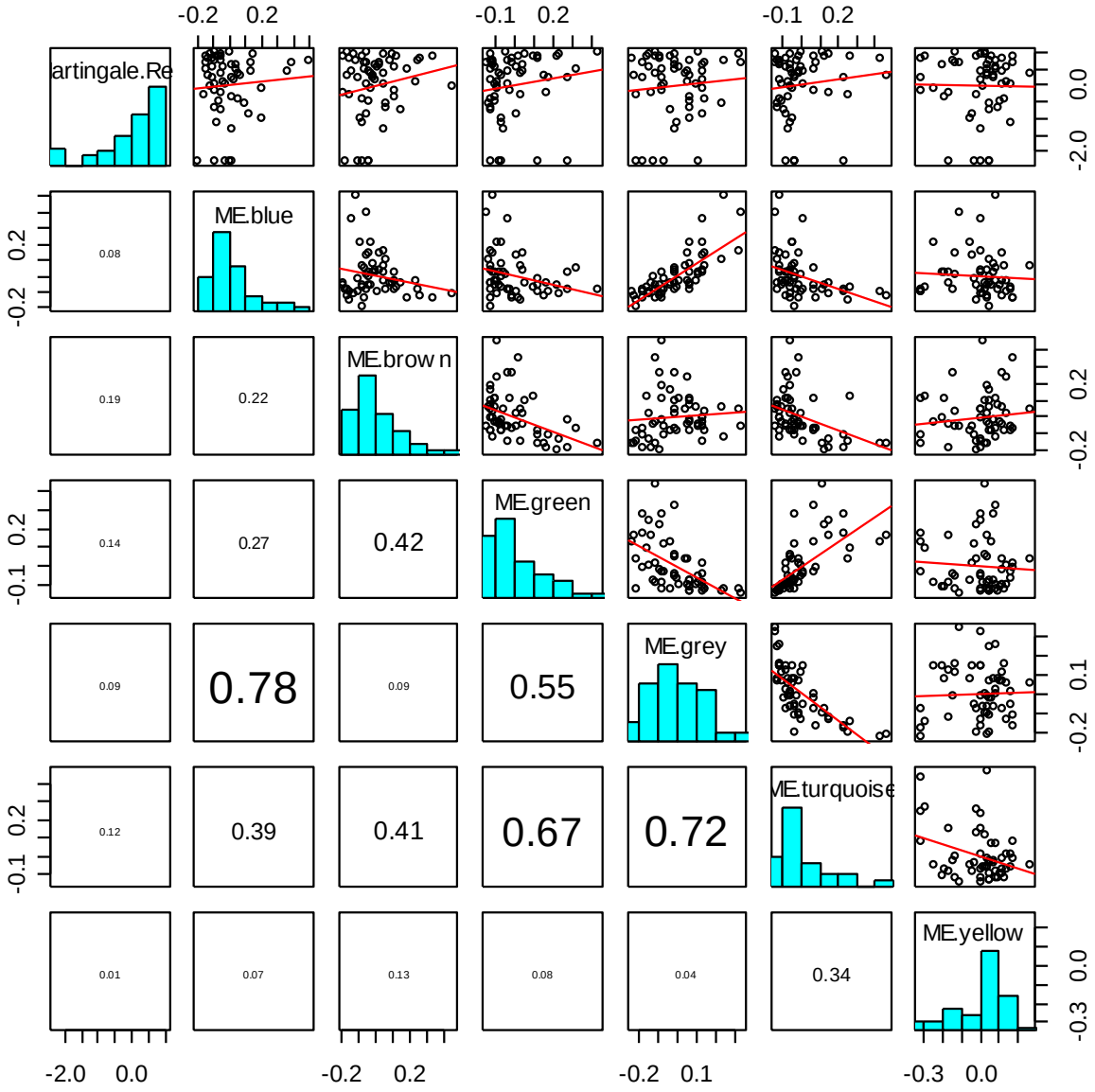
Rows,=genes, Columns=microarray



The brown module eigengenes across samples

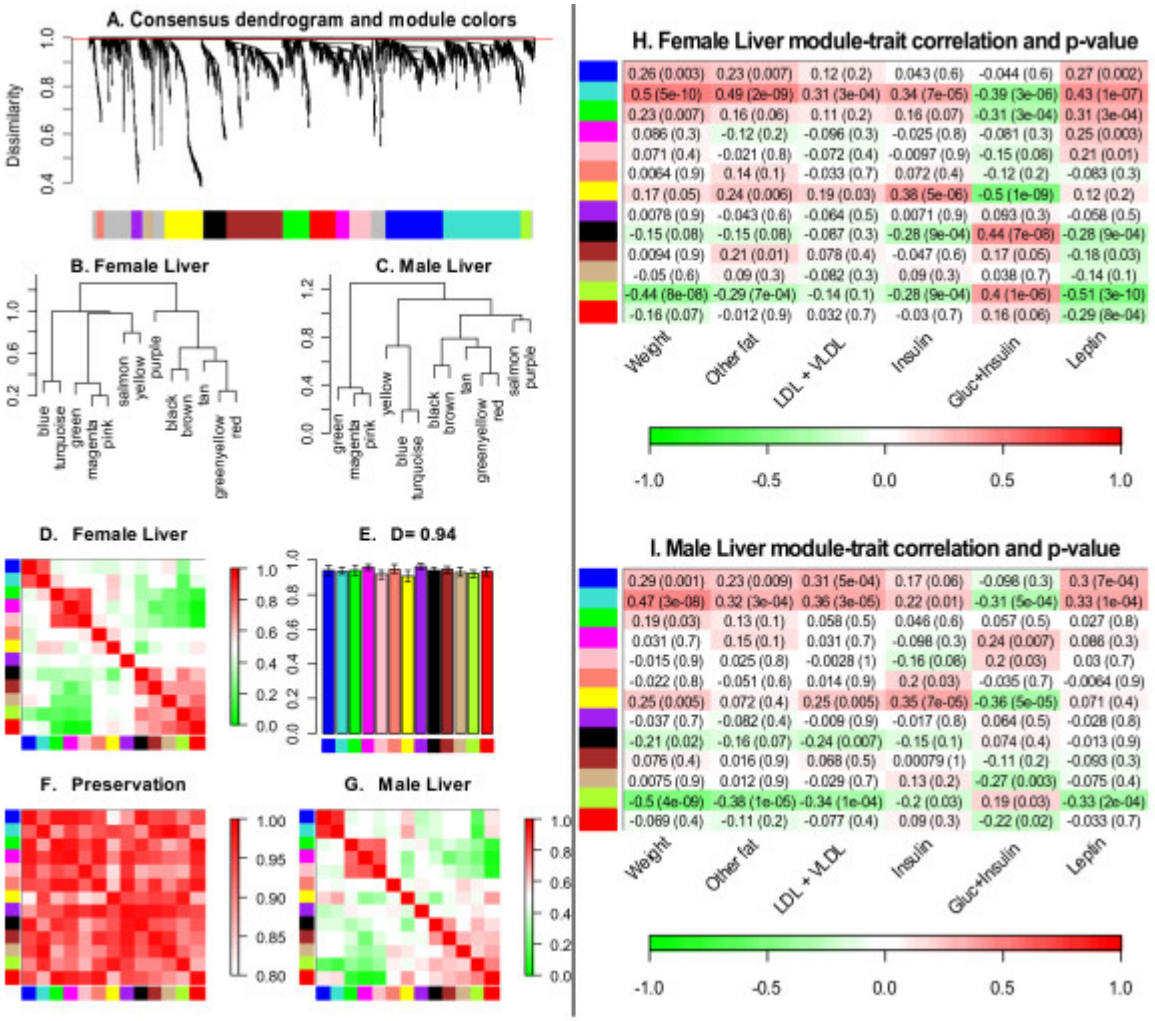


Module eigengenes can be used to determine whether 2 modules are correlated. If correlation of MEs is high-> consider merging.



Eigengenes can be used to build separate networks…

Consensus eigengene networks in male and female mouse liver data and their relationship to physiological traits



Langfelder P, Horvath S (2007) Eigengene networks for studying the relationships between co-expression modules. BMC Systems Biology 2007

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

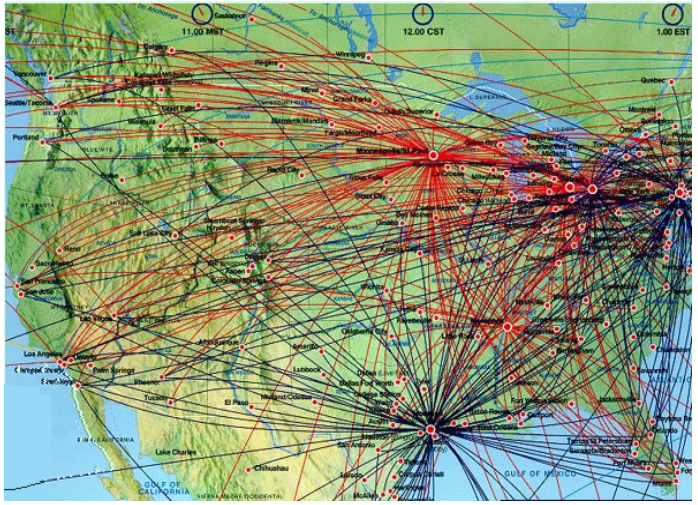
A gene significance naturally gives rise to a module significance measure

- 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?

Flight connections and hub airports



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

****Slide courtesy of A Barabasi**

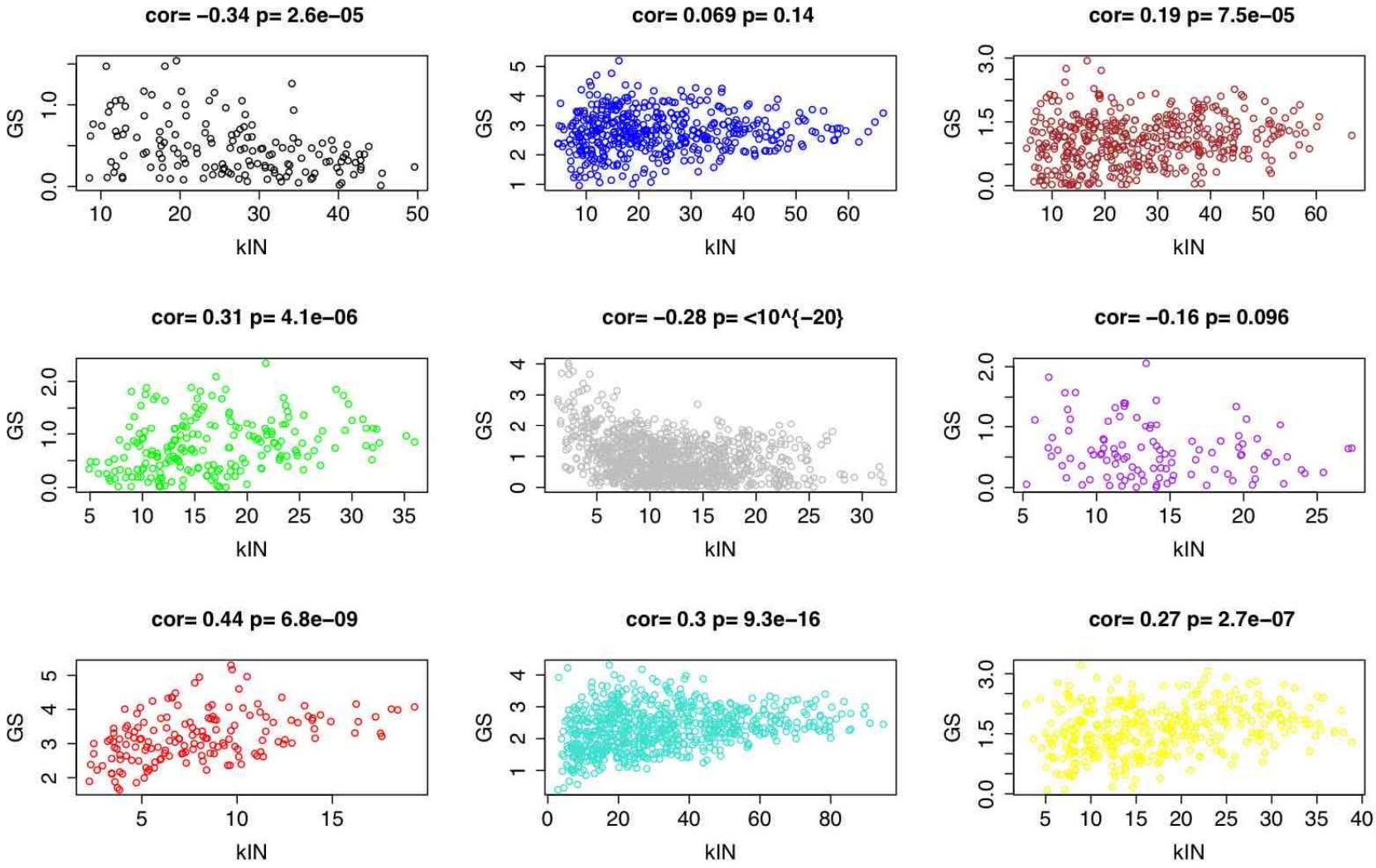
What is intramodular connectivity?

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

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

Gene significance versus intramodular connectivity kIN



How to use networks for gene screening?

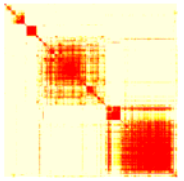
Intramodular connectivity kIN versus gene significance GS

- Note the relatively high correlation between gene significance and intramodular connectivity in some modules
- In general, kIN is a more reliable measure than GS
- In practice, a combination of GS and k should be used
- Module eigengene turns out to be the most highly connected gene (under mild assumptions)

What is weighted gene co-expression network analysis?

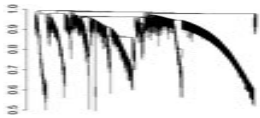
Construct a network

Rationale: make use of interaction patterns between genes



Identify modules

Rationale: module (pathway) based analysis

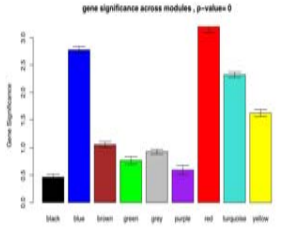


Relate modules to external information

Array Information: Clinical data, SNPs, proteomics

Gene Information: gene ontology, EASE, IPA

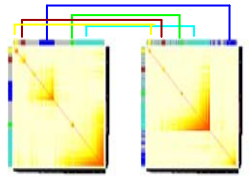
Rationale: find biologically interesting modules



Study Module Preservation across different data

Rationale:

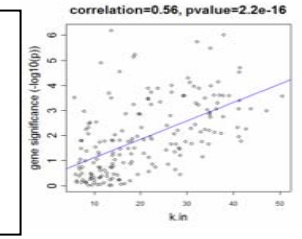
- Same data: to check robustness of module definition
- Different data: to find interesting modules.



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 (pathways) instead of individual genes**
  - Greatly alleviates the problem of multiple comparisons
    - Less than 20 comparisons versus 20000 comparisons
- Use of intramodular connectivity to find key drivers
  - Quantifies module membership (centrality)
  - Highly connected genes have an increased chance of validation
- Module definition is based on gene expression data
  - 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: power of a correlation
  - Rationale:
    - puts different data sets on the same mathematical footing
    - Considers effect size estimates (cor) and significance level
    - p-values are highly affected by sample sizes (cor=0.01 is highly significant when dealing with 100000 observations)
- Technical Details: soft thresholding with the power adjacency function, topological overlap matrix to measure interconnectedness

Case Study 1:
Finding brain cancer genes

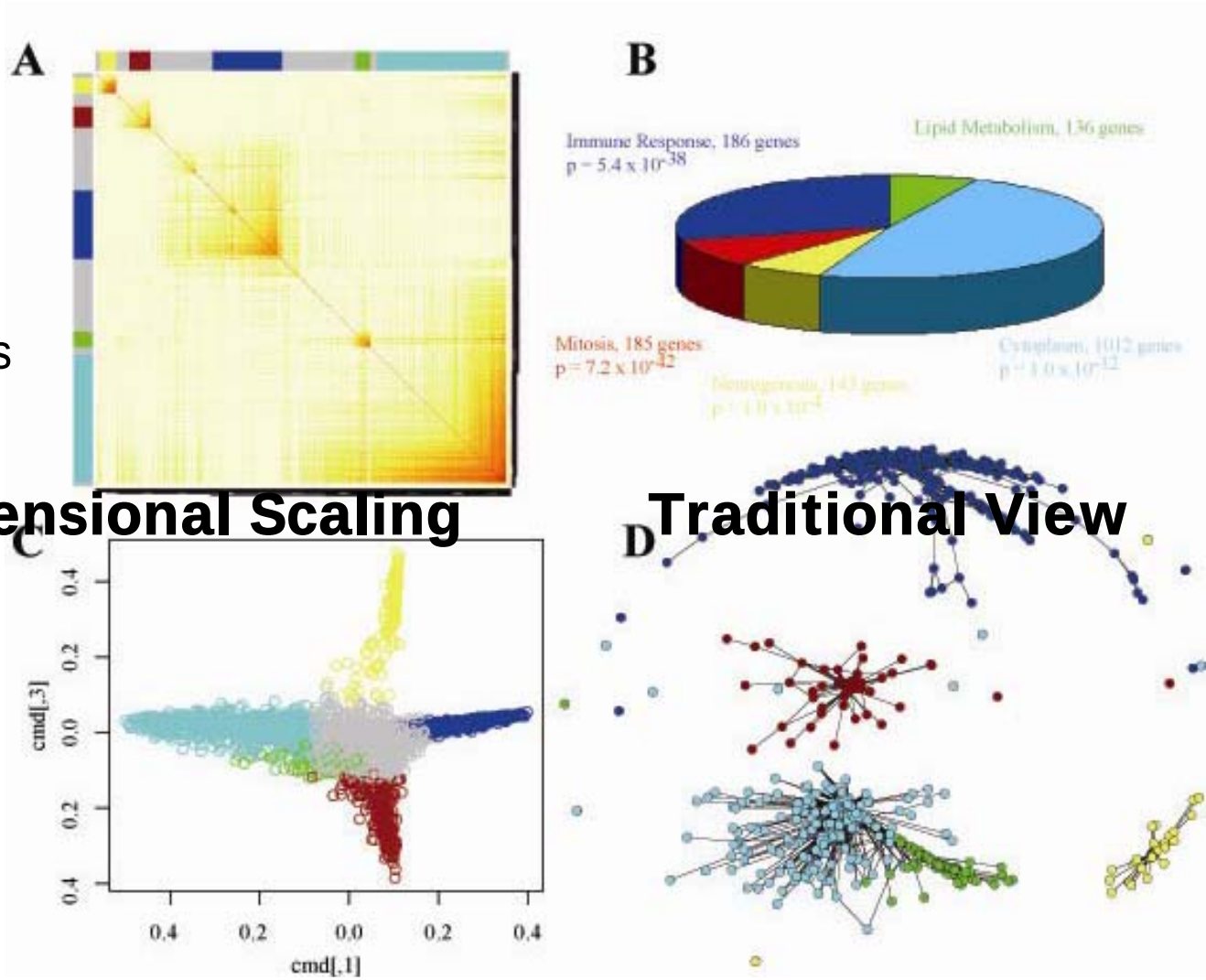
Horvath S, Zhang B, Carlson M, Lu KV, Zhu S, Felciano RM, Laurance MF, Zhao W, Shu, Q, Lee Y, Scheck AC, Liau LM, Wu H, Geschwind DH, Febbo PG, Kornblum HI, Cloughesy TF, Nelson SF, Mischel PS (2006) "Analysis of Oncogenic Signaling Networks in Glioblastoma Identifies ASPM as a Novel Molecular Target",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

Different Ways of Depicting Gene Modules

Topological Overlap Plot

Gene Functions

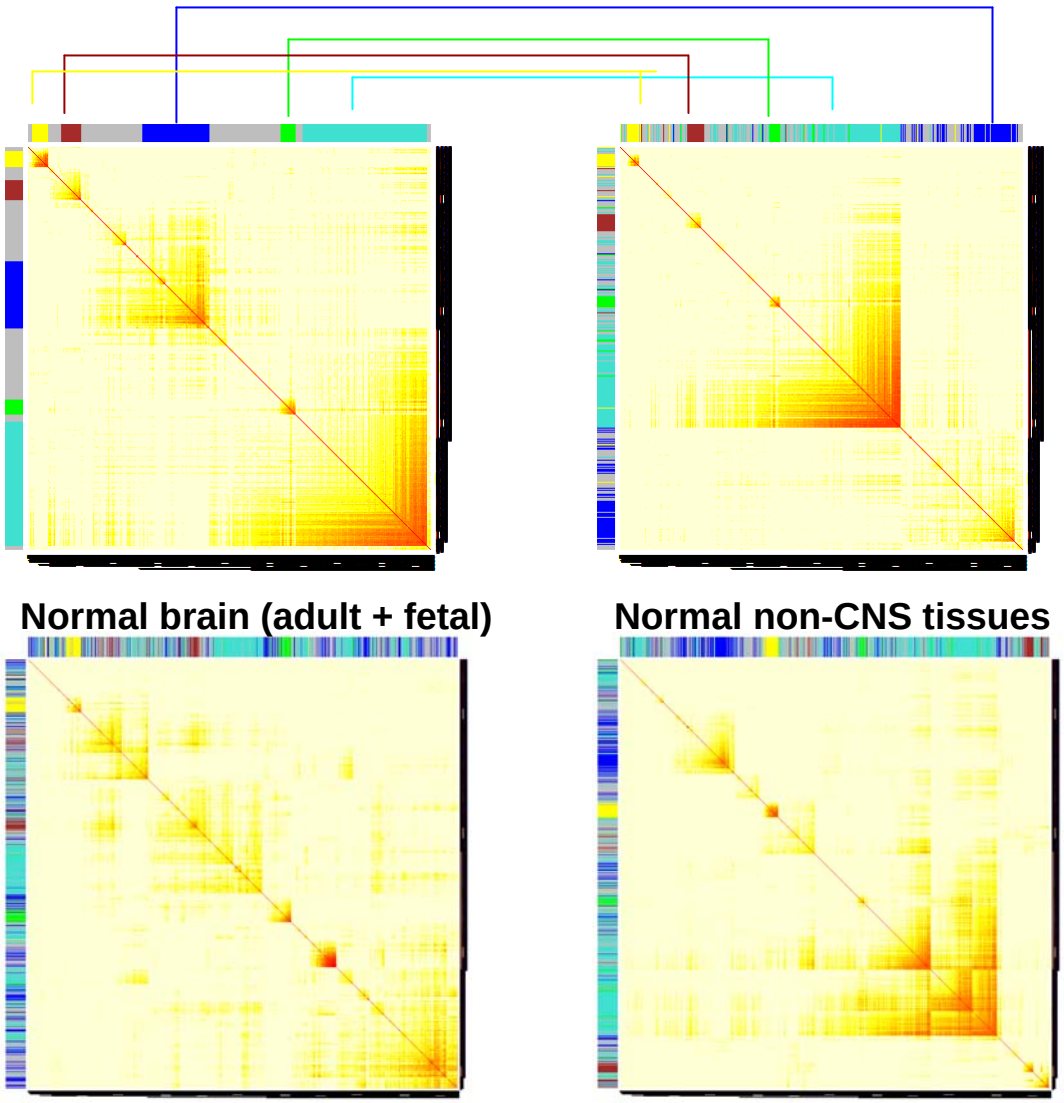
1) Rows and columns correspond to genes
2) Red boxes along diagonal are modules
3) Color bands=modules



Multi Dimensional Scaling

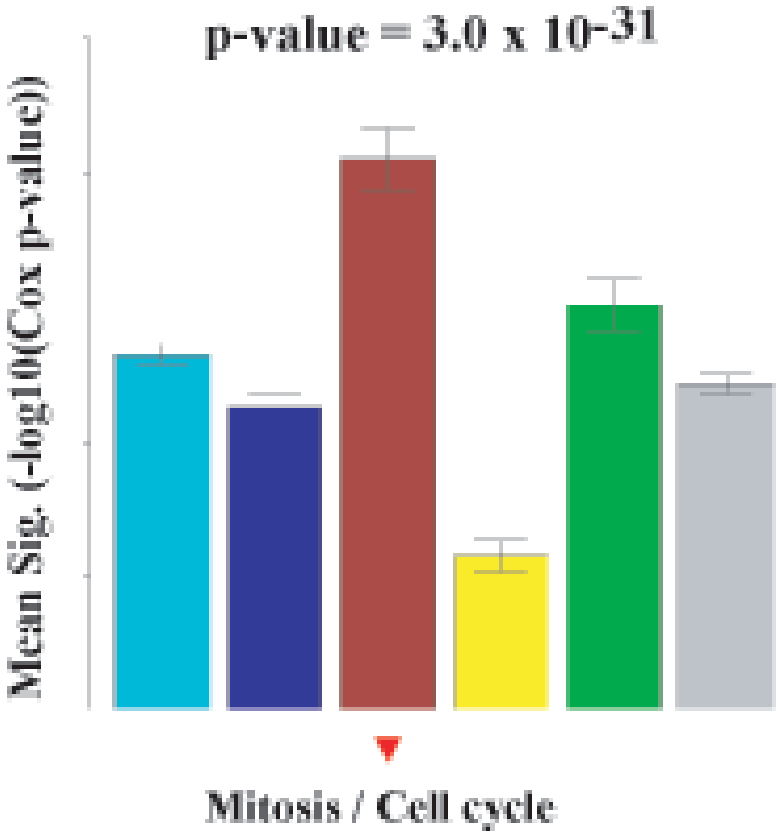
Traditional View

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

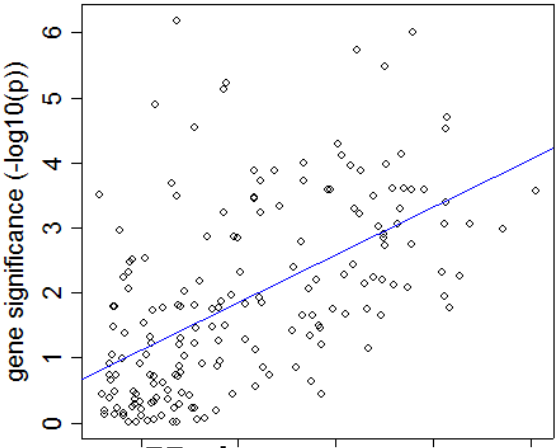
Mean Prognostic Significance of Module Genes



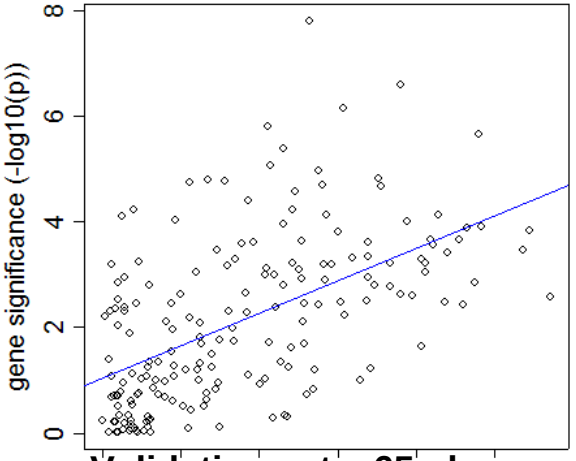
Message: Focus the attention on the brown module genes

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}

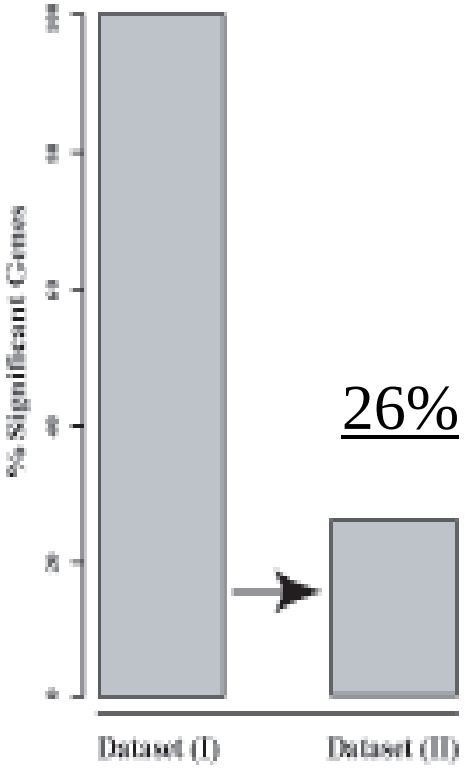
The fact that genes with high intramodular connectivity are more likely to be prognostically significant facilitates a novel screening strategy for finding prognostic genes

- 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

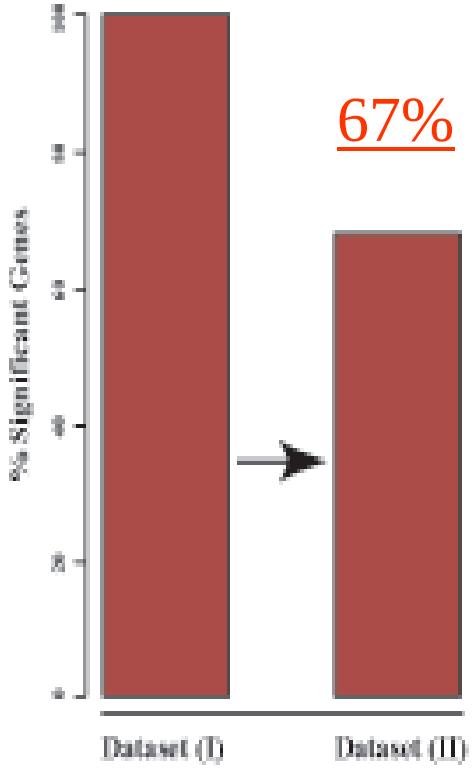
- Validation success rate= proportion of genes with independent test set Cox regression p-value<0.05.
- Validation success rate of network based screening approach (68%)
- Standard approach involving top 300 most significant genes: 26%

Validation success rate of gene expressions in independent data

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



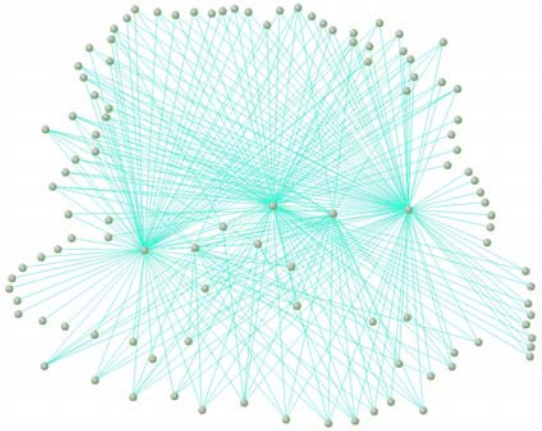
Network based screening
p<0.05 and
high intramodular connectivity

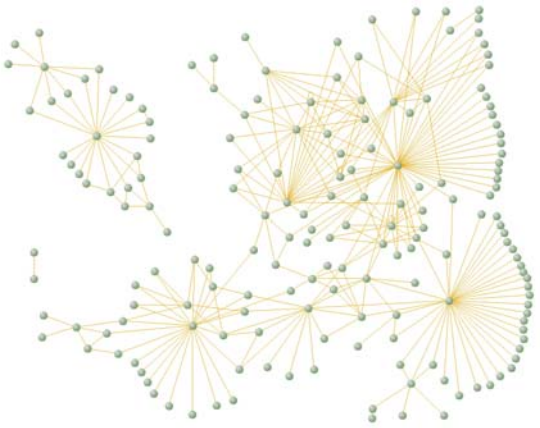


The network-based approach uncovers novel therapeutic targets

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.

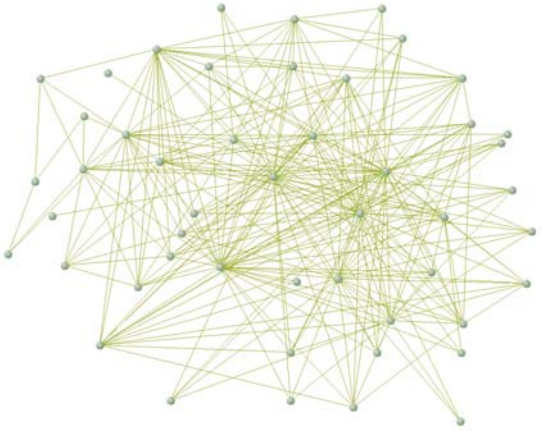
Biological validation: siRNA mediated inhibition of ASPM

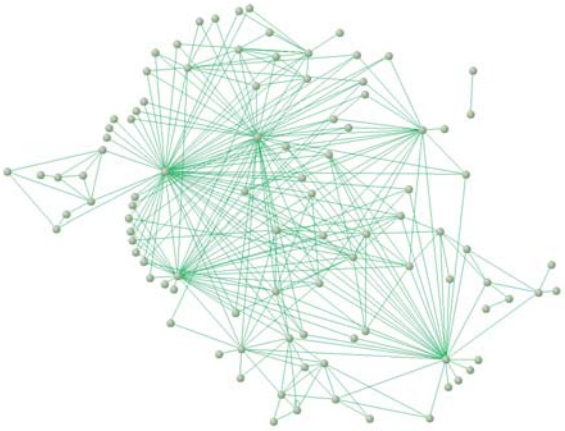




Case Study 2

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)

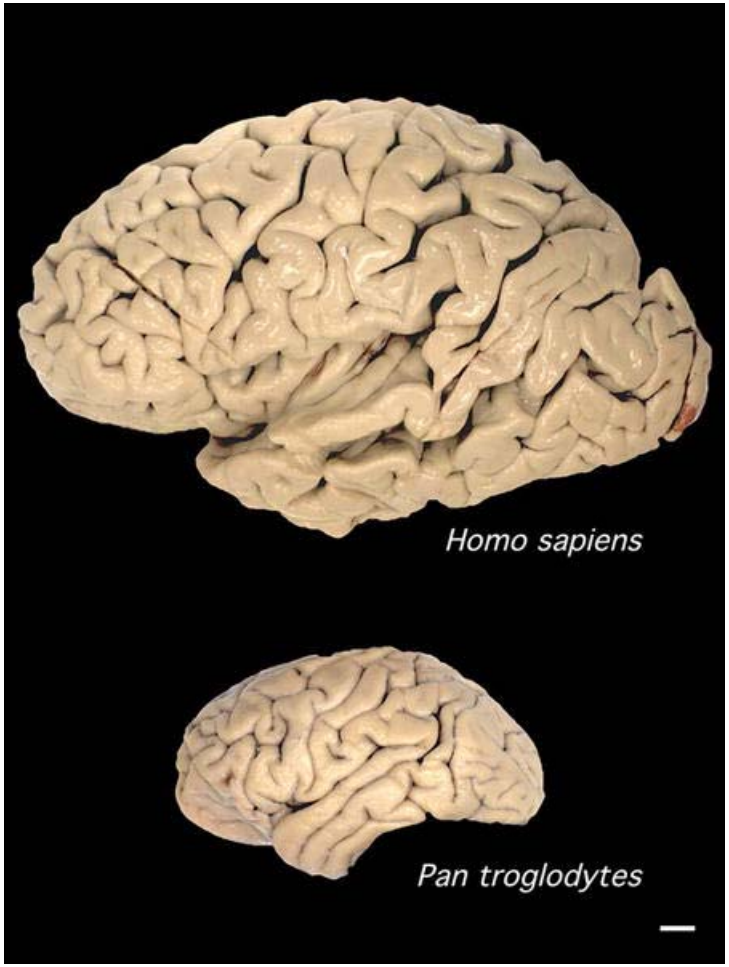


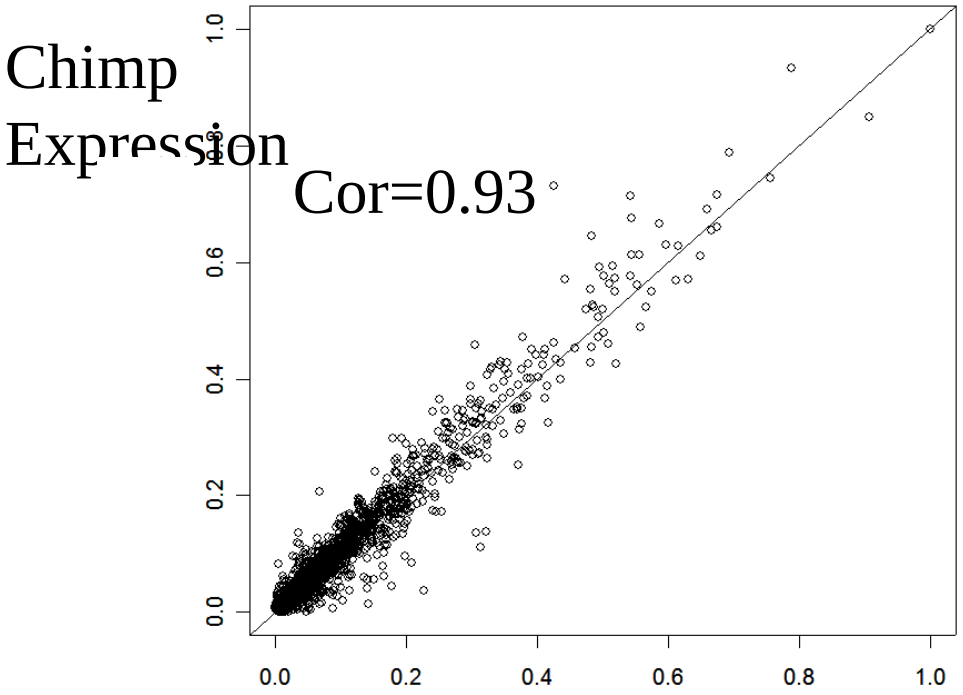
Image courtesy of Todd Preuss (Yerkes National Primate Research Center)

Assessing the contribution of regulatory changes to human evolution

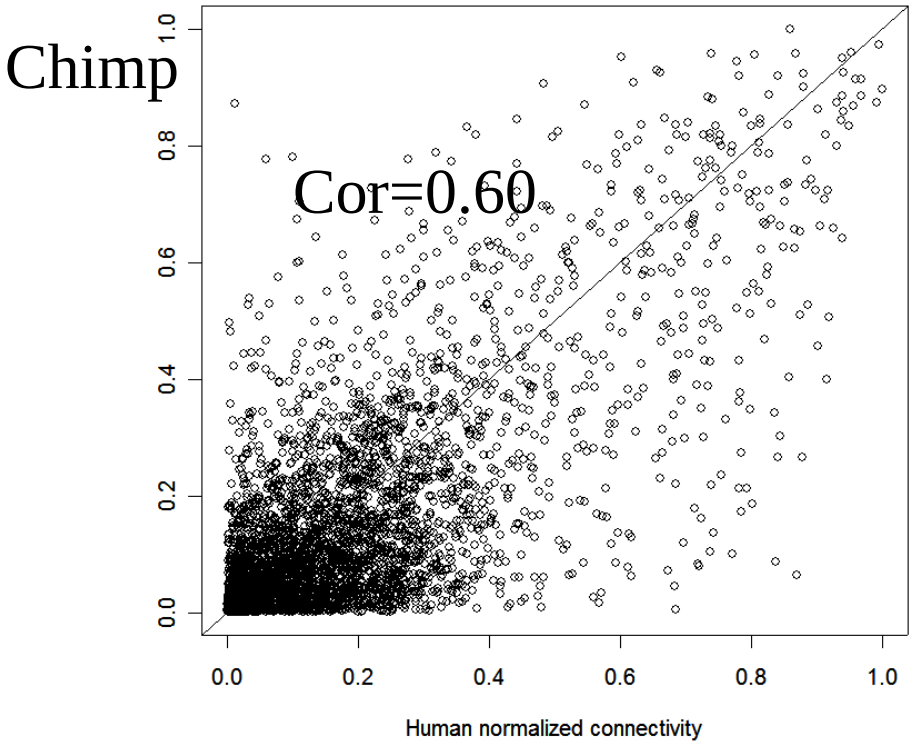
- Hypothesis: Changes in the regulation of gene expression were critical during recent human evolution (King & Wilson, 1975)

- 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

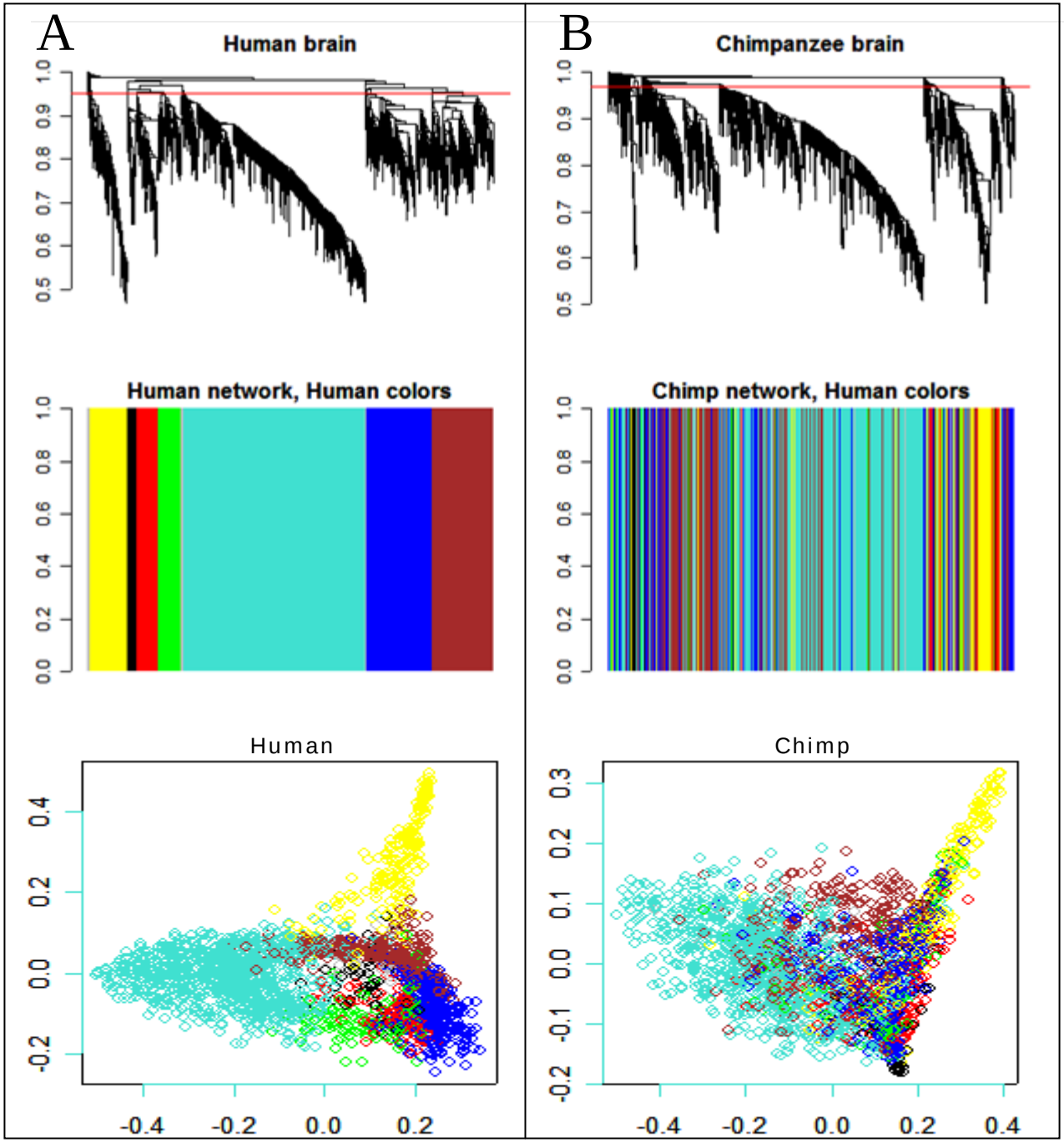


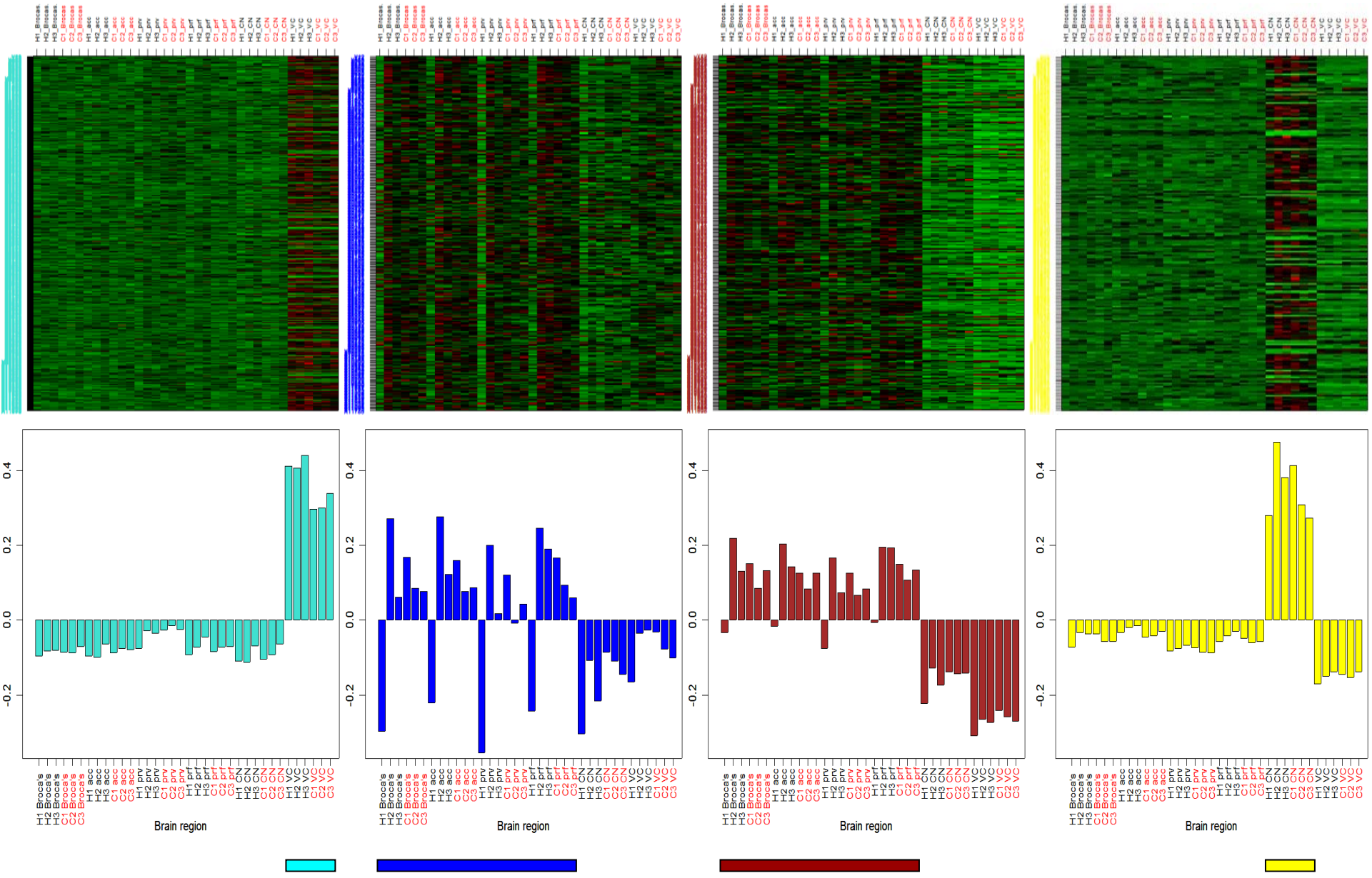
Human Connectivity

Hypothesis: molecular wiring makes us human

Raw data from Khaitovich *et al.*, 2004

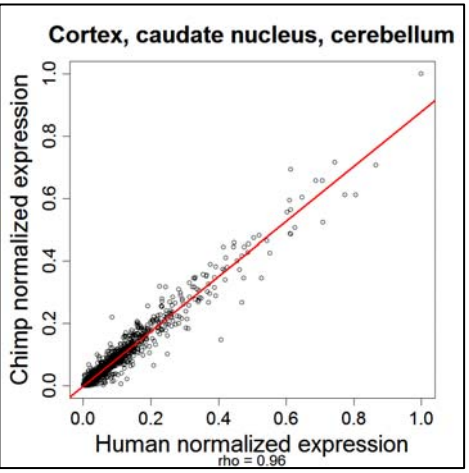
Mike Oldham

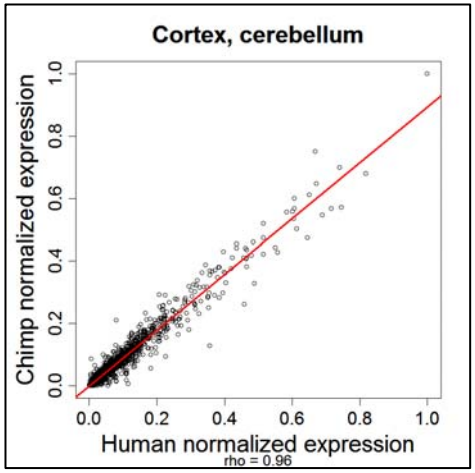

A
Human brain
1.0
0.9
0.8
0.7
0.6
0.5
Human network, Human colors
1.0
0.8
0.6
0.4
0.2
0.0
Human
0.4
0.2
0.0
-0.2
-0.4
-0.2
0.0
0.2
B
Chimpanzee brain
1.0
0.9
0.8
0.7
0.6
0.5
Chimp network, Human colors
1.0
0.8
0.6
0.4
0.2
0.0
Chimp
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.4
-0.2
0.0
0.2
0.4

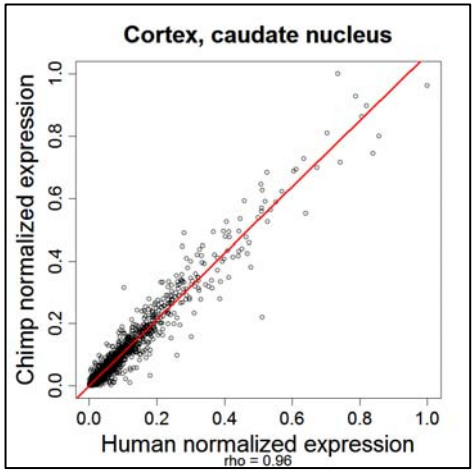


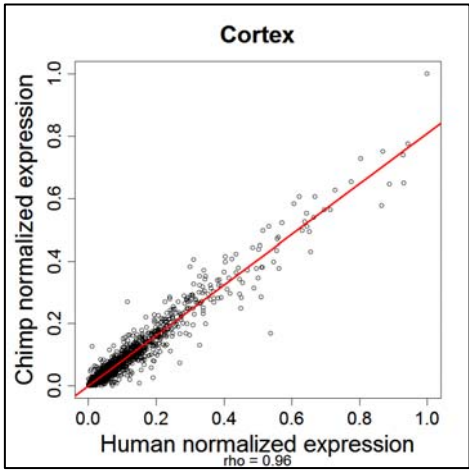
$p = 1.33 \times 10^{-4}$ $p = 8.93 \times 10^{-4}$ $p = 1.35 \times 10^{-6}$ $p = 1.33 \times 10^{-4}$

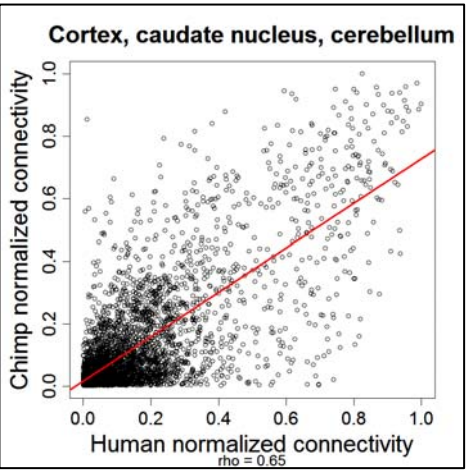
Connectivity diverges across brain regions whereas expression does not

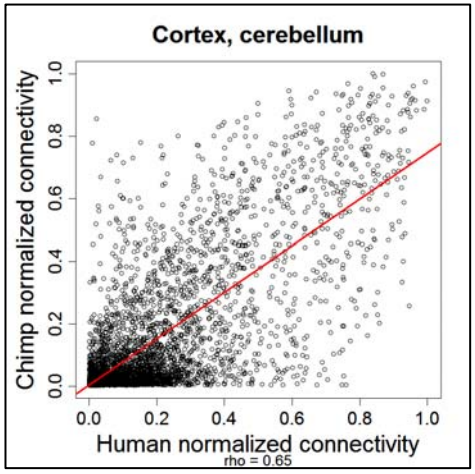


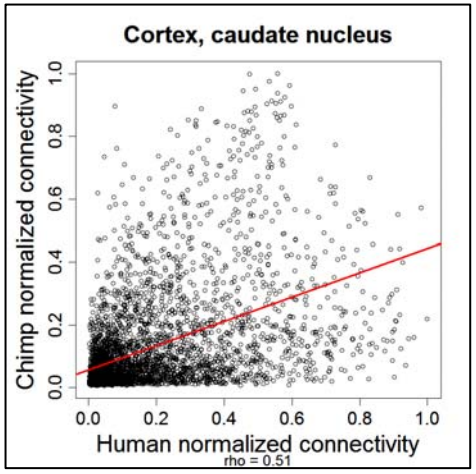


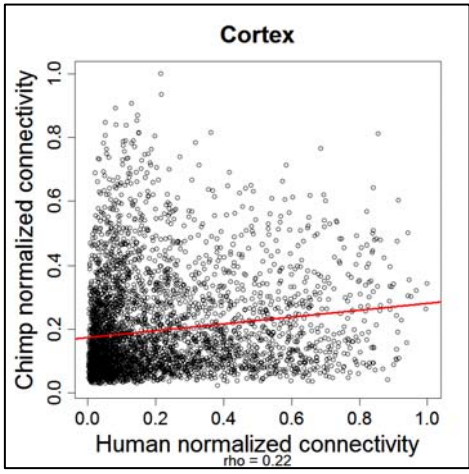












Conclusions: chimp/human

- 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

Software and Data Availability

- Sample data and R software tutorials can be found at the following webpage
- http://www.genetics.ucla.edu/labs/horvath/CoexpressionNetwork
- An R package and accompanying tutorial can be found here:
- http://www.genetics.ucla.edu/labs/horvath/CoexpressionNetwork/Rpackages/WGCNA/
- Tutorial for this R package
- http://www.genetics.ucla.edu/labs/horvath/CoexpressionNetwork/Rpackages/WGCNA/TutorialWGCNApackage.doc

THE END