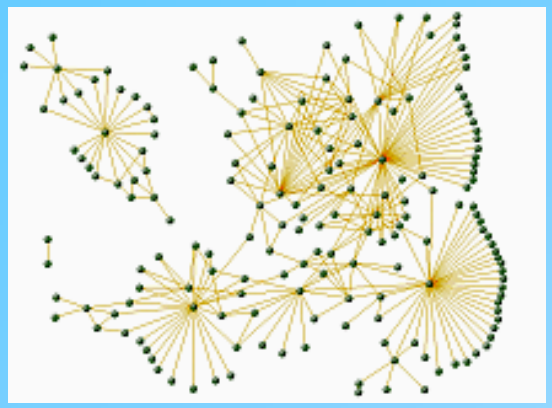
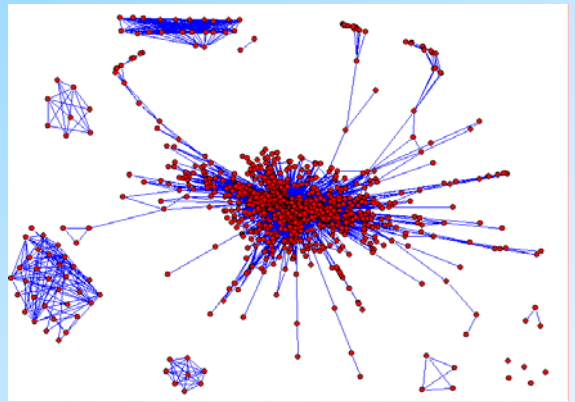
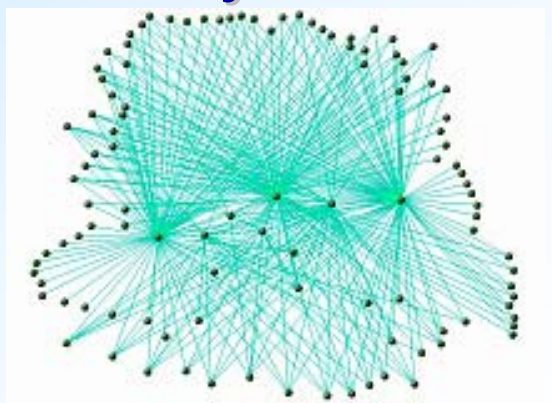




Weighted Gene Co-Expression Network Analysis

Combined Analysis of SNP and Microarray Data

Steve Horvath
Jeanette Papp
UCLA



U'C L A HUMAN Genetics

The Challenge

- Using traditional genetic methods, the major simple mendelian traits have been identified
- Finding genes for complex traits and understanding the underlying biology has proven much more difficult

The Solution

Integration of different data types

- Genetic Data - SNPs
- Expression Data
- Protein Data

The Solution

Integration of different data types

- Genetic Data - SNPs
- Expression Data
- Protein Data

UCLA HUMAN Genetics

The Tool

Network Analysis

Combined Analysis of SNP and Microarray Data

UCLA HUMAN Genetics

Network Analysis

- Brief review of Network Analysis concepts and methods
- Some results from projects at UCLA

UCLA HUMAN Genetics

Connectivity can be an important variable for identifying key nodes

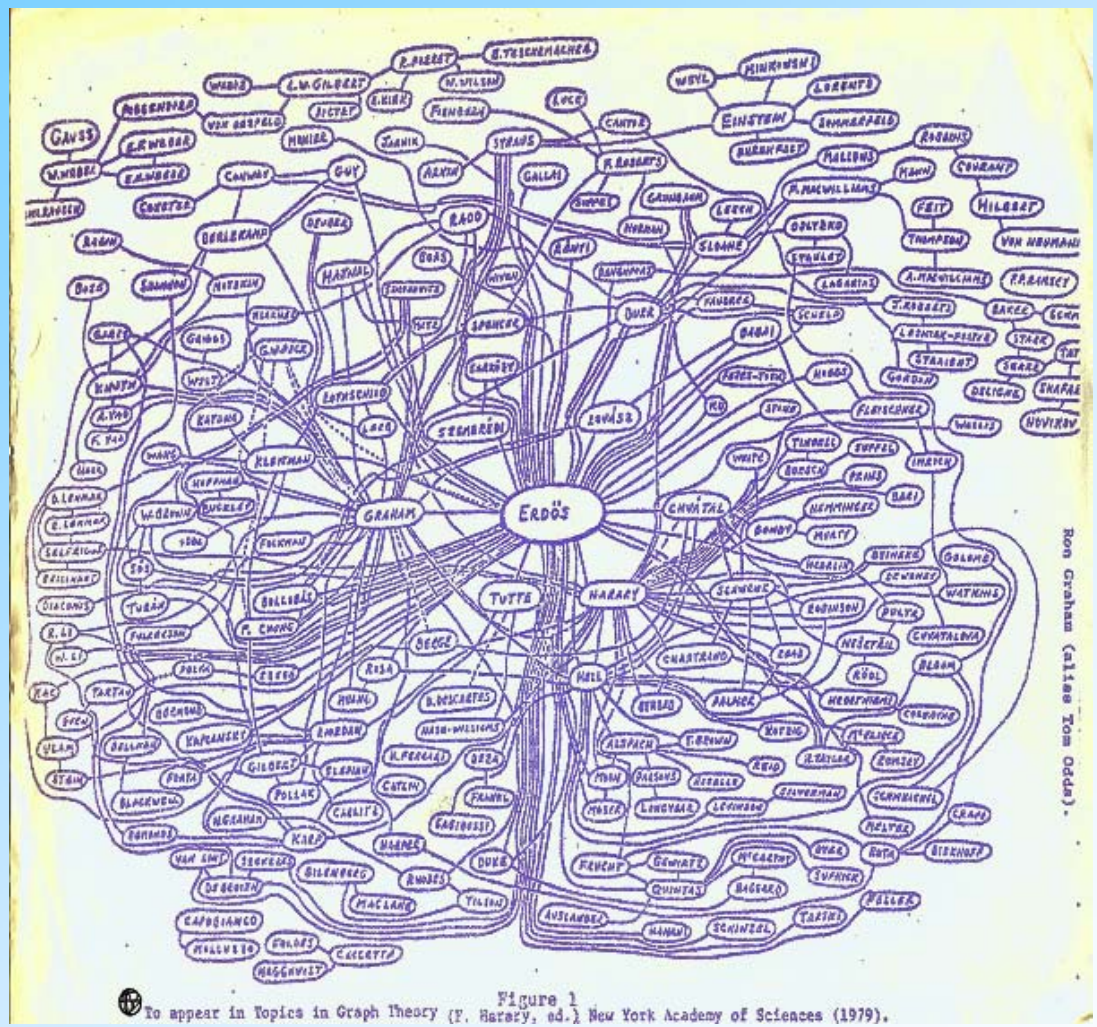


Figure 1
To appear in Topics in Graph Theory (F. Harary, ed.) New York Academy of Sciences (1979).

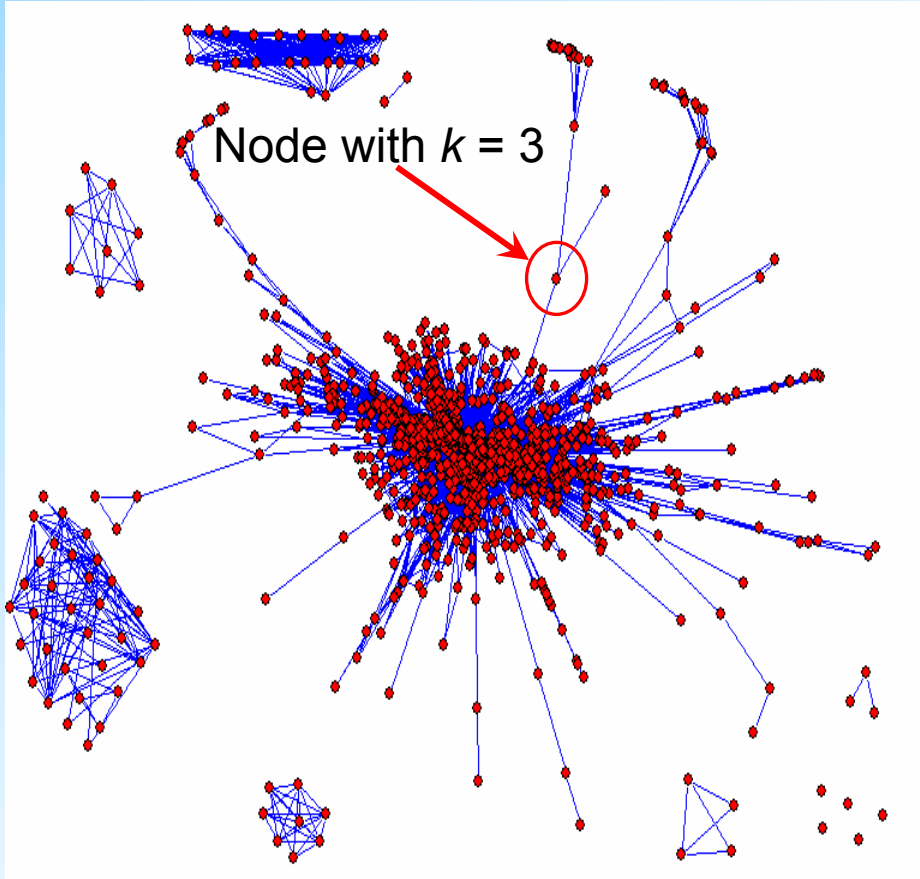
Which of the following mathematicians had the biggest influence on others?

Networks

Network methods are used to model systems such as the internet, social interactions, and biological pathways.

Definitions:

- Node = object (eg. gene)
- Link = line connecting 2 objects
- k = Degree(Node i) = # of links to Node i
- $\Pr(k)$ = probability Node i has k links
- Gene Module = group of co-expressed genes



Two Network Types:

1) Random Network:

$$\Pr(k) \sim Poisson(\lambda) = \frac{e^{-\lambda}\lambda^{k}}{k!}$$

i.e. Each node has approximately the same number of links

2) Scale-Free Network:

$$\Pr(k) \sim \text{Power Law}(\gamma) = k^{-\gamma}$$

i.e. Some nodes are highly connected with thousands or even millions of links = "Hub Nodes"

UCLA HUMAN Genetics

Weighted vs. Unweighted Networks

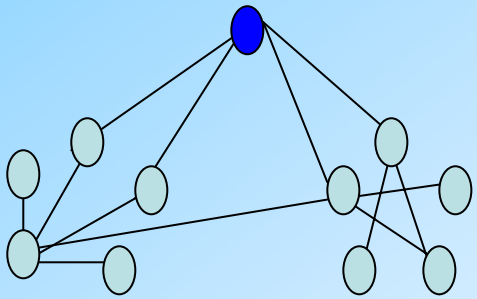
Network = Adjacency matrix

$A=[a_{ij}]$ encodes whether/how a pair of nodes is connected.

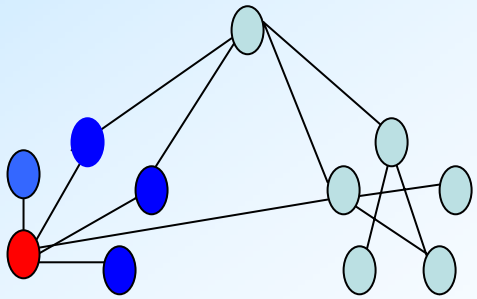
- A is a symmetric matrix with entries in [0,1]
- For **unweighted** networks, $a_{ij} = 1$ if two genes are adjacent (connected) and 0 otherwise.
- For **weighted** networks, the adjacency matrix reports the connection strength between gene pairs

Network Modules

Whole network connectivity



Intramodular connectivity

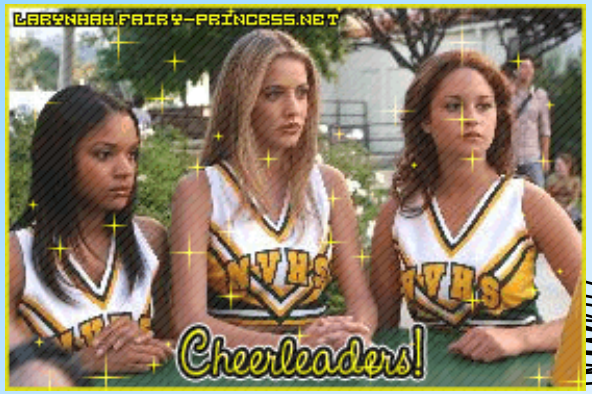


UCLA HUMAN Genetics

Identifying Key Players of Interest

Imagine you wanted to recruit students to your science program. Popularity alone might suggest the head cheerleader or quarterback.

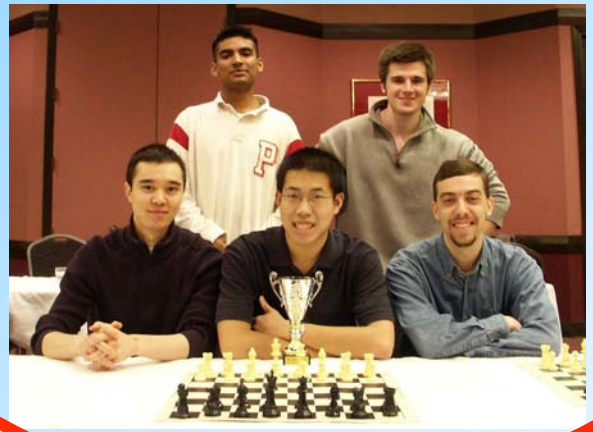
Head Cheerleader



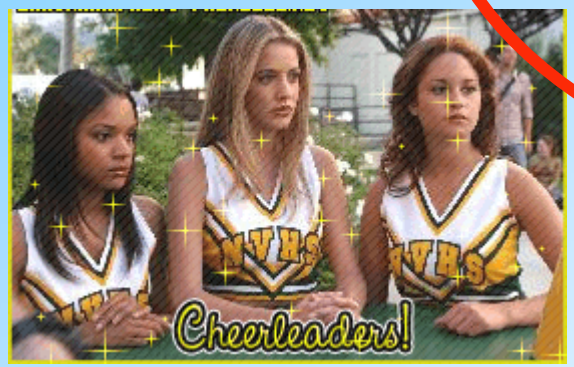
Star Quarterback

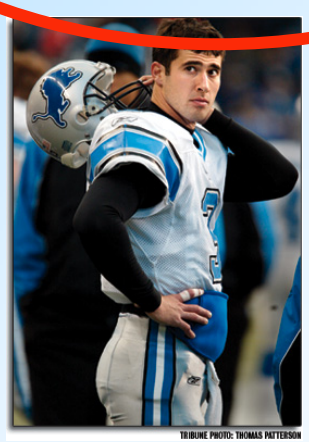
But, the head of the chess club would probably be a better bet!

Chess Club President



Cheerleader





Quarterback

Two Network Definitions

- Number of friends = "**Connectivity**"

  Gene Connectivity = row sum of the adjacency matrix, sum of gene$_i$'s connection strengths

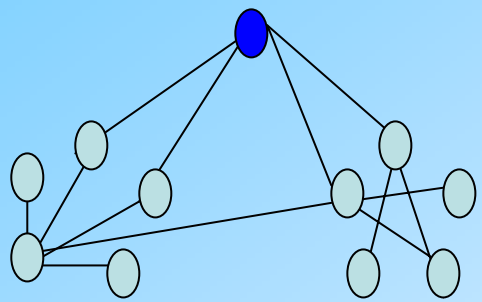
$$k_i = \sum_j a_{ij}$$

- Chess Club, Sport Teams = "**Modules**"

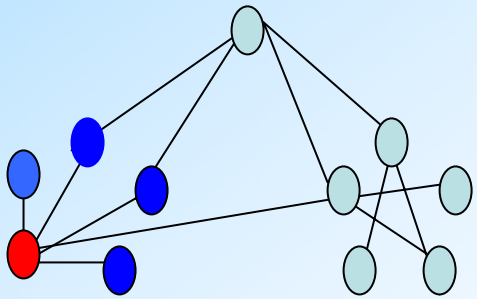
  Gene Module = cluster of highly connected (similarly expressed) genes in a network

Intra-modular connectivity is biologically and mathematically more meaningful than whole network connectivity

Whole network connectivity



Intramodular connectivity



- Hub genes are module genes in co-expression networks
- Genes that are not in modules tend to have low connectivity
- Module genes have relatively high connectivity
- Module genes have high connectivity within their module

Modules, not individual genes, are key drivers of the network

UCLA HUMAN Genetics

Gene Module

- A group of co-expressed genes
- A set of tightly co-regulated genes
- A biological pathway?

Gene Network Analysis

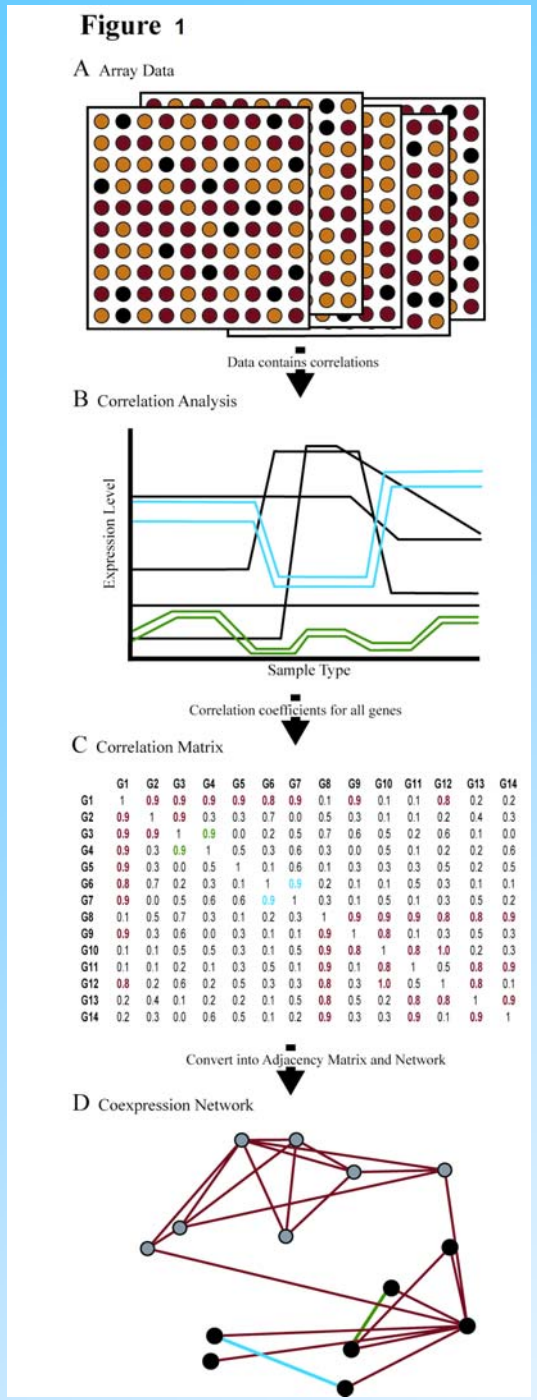
- In gene co-expression networks, each gene corresponds to a node.
- Two genes are connected by an edge if their expression values are highly correlated.
- Describes the presence of **Hub Nodes** that are connected to a large number of other nodes
- Defines **Gene Modules** as sets of tightly co-regulated genes

An Adjacency Function is used to turn co-expression information into a network

- Measure co-expression by the absolute value of the Pearson correlation
- Define an adjacency matrix by using an adjacency function $A(i,j)=AF(|cor(x[i],x[j])|)$
- The adjacency function AF is a monotonic function that maps $[0,1]$ onto $[0,1]$
- We consider 2 classes of AF
  - **Hard Thresholding:**

    Step Function $AF(s) = I(s > tau)$ with parameter tau
  - **Soft thresholding:**

    Continuous Power Adjacency Function $AF(s) = s^b$ with parameter b
- The choice of the AF parameters determines the properties of the network.



Steps for constructing a co-expression network

A. Take microarray gene expression data

B. Measure concordance of gene expression with a Pearson correlation

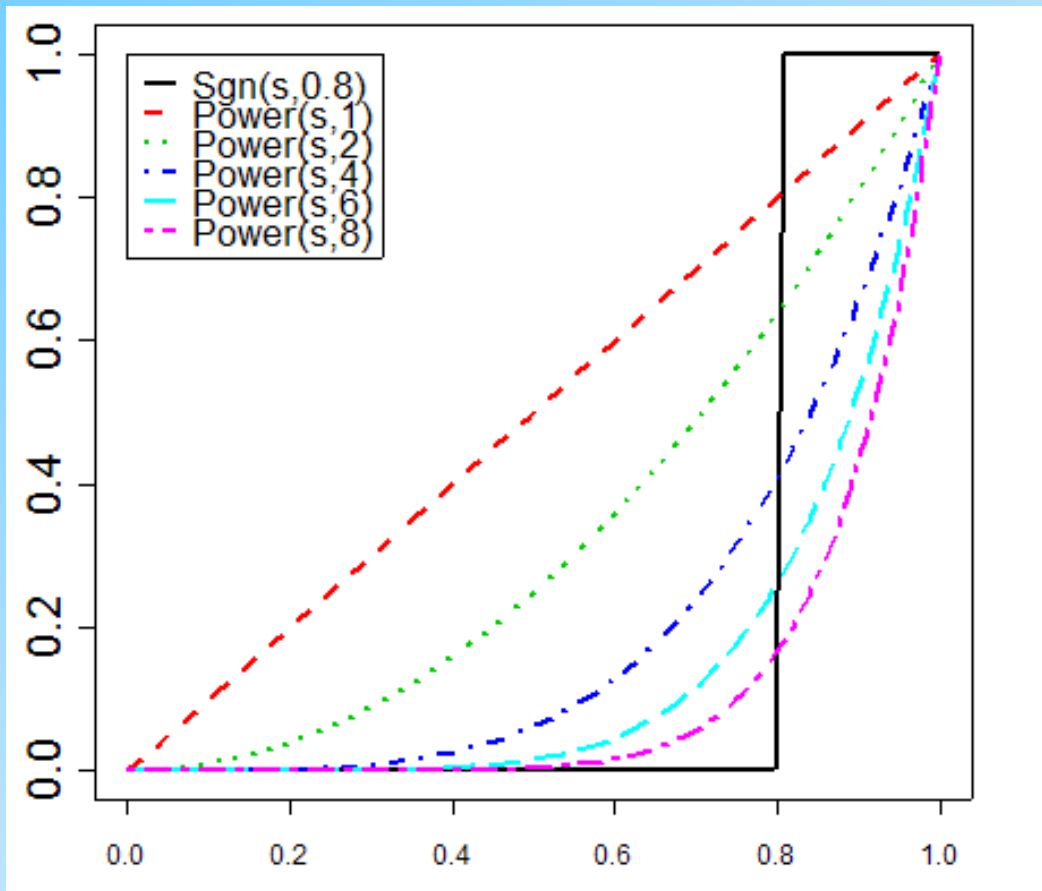
C. The Pearson correlation matrix is either dichotomized to arrive at an adjacency matrix → unweighted network

   Or

D. Transformed continuously with the power adjacency function → weighted network

Adjacency Functions

Connection Strength (Adjacency) vs. Correlation



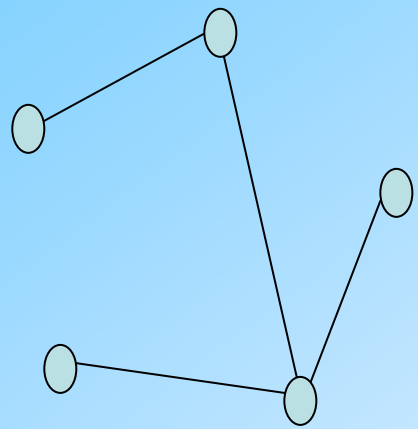
Adjacency a_{ij} =
$|cor(gene_i, gene_j)|^{\beta}$

- Step function (hard thresholding) is indicated by the black, solid line
  → unweighted
- Power adjacency functions (soft thresholding) are indicated by colored, dashed lines
  → weighted

UCLA HUMAN Genetics

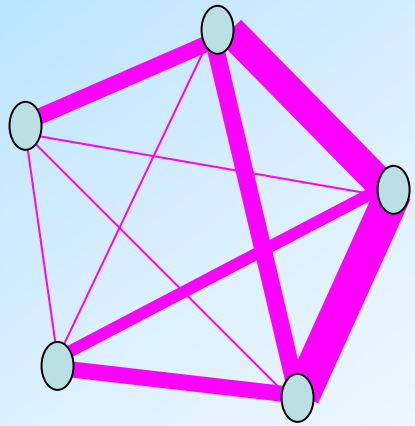
Weighted vs. Unweighted Networks

Unweighted Network View



Some genes are connected
All connections are equal

Weighted Network View

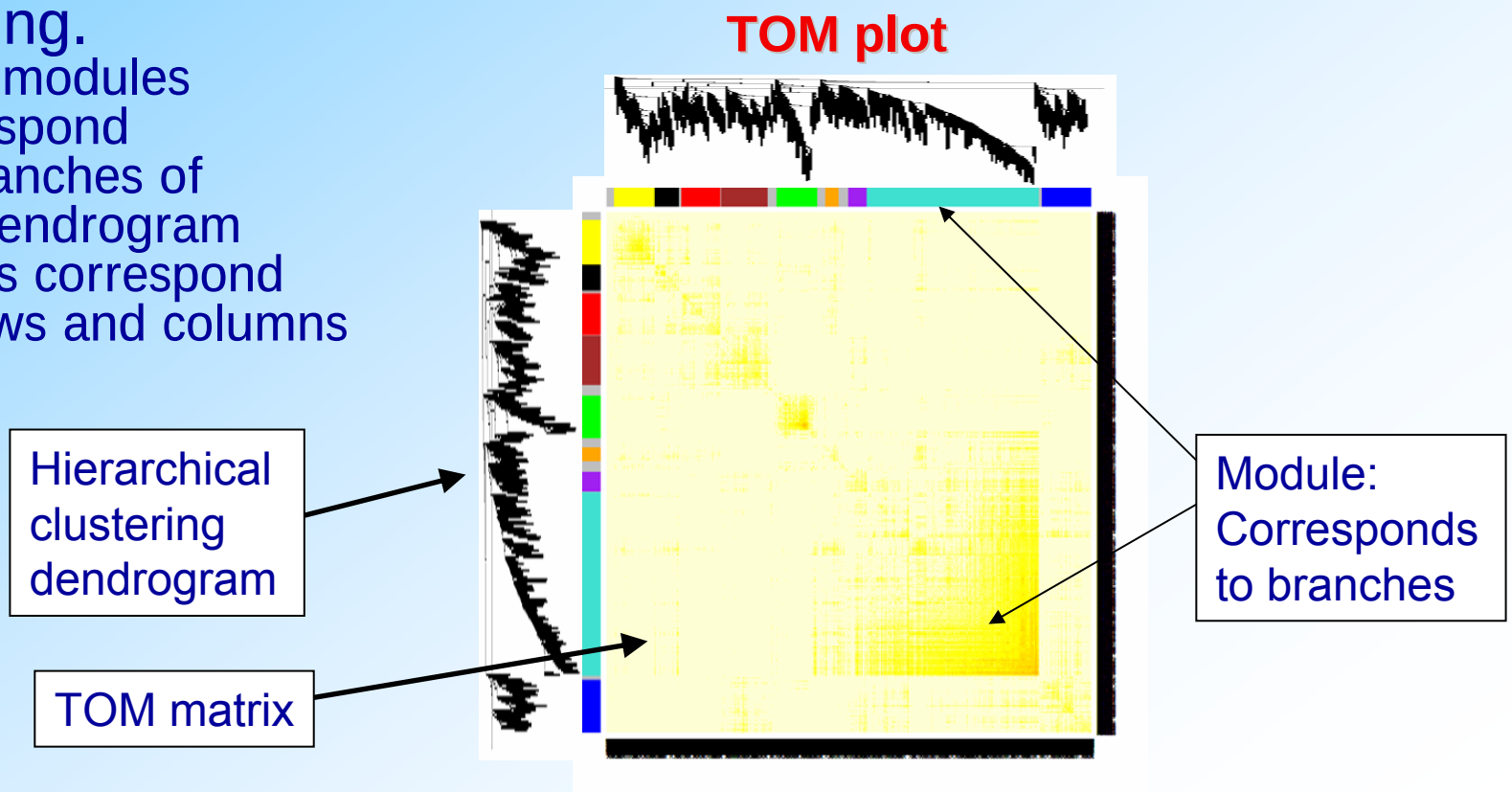


All genes are connected
Connection Width = Connection strength

Hard thresholding may lead to loss of information

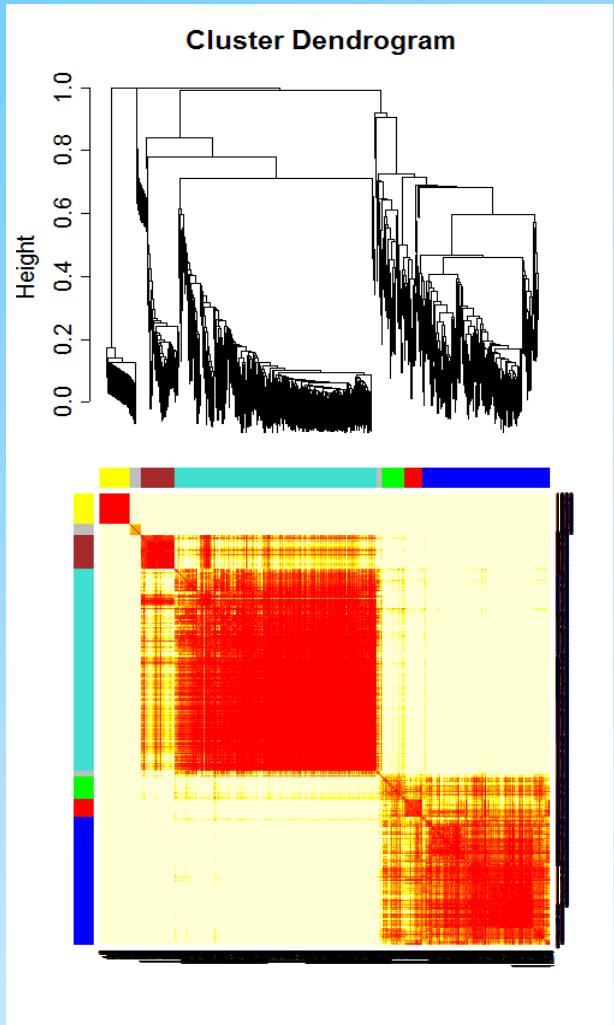
Topological Overlap Matrix Using the TOM to Cluster Genes

- 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
  - Genes correspond to rows and columns

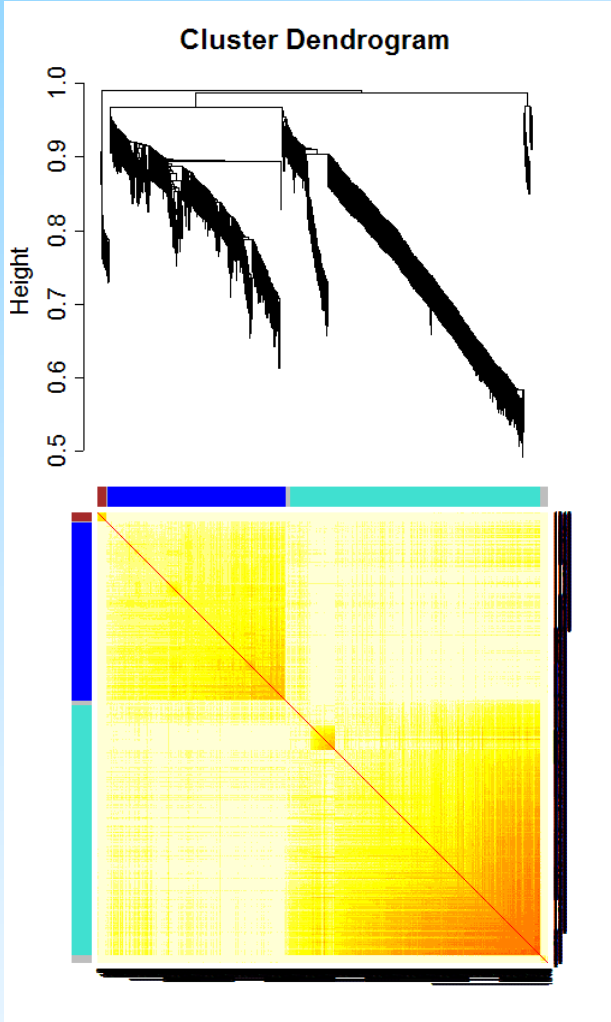


Dendrogram "trimmed" to create modules

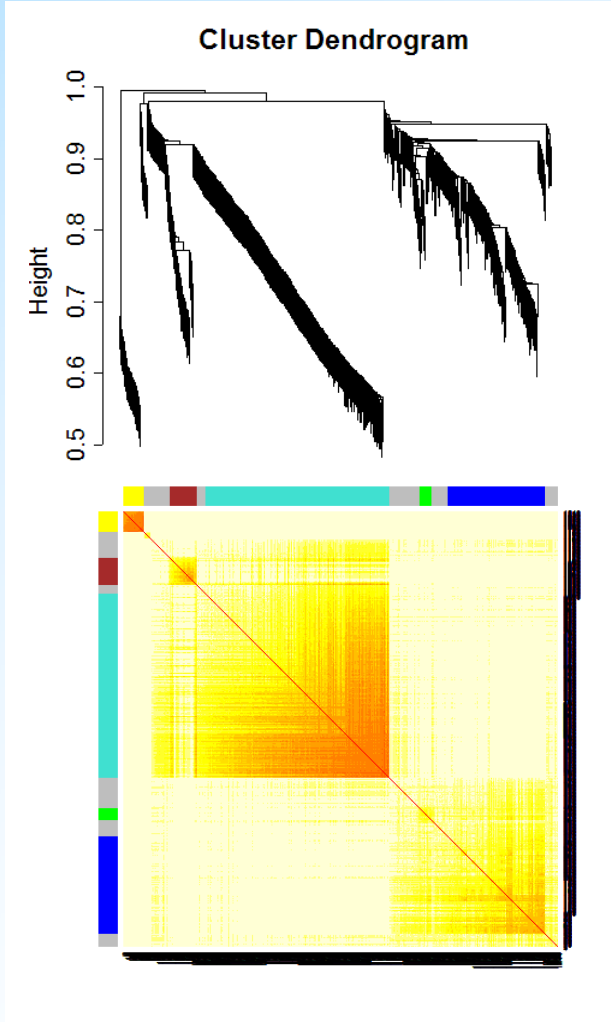
Step AF (tau)



Power AF (b)

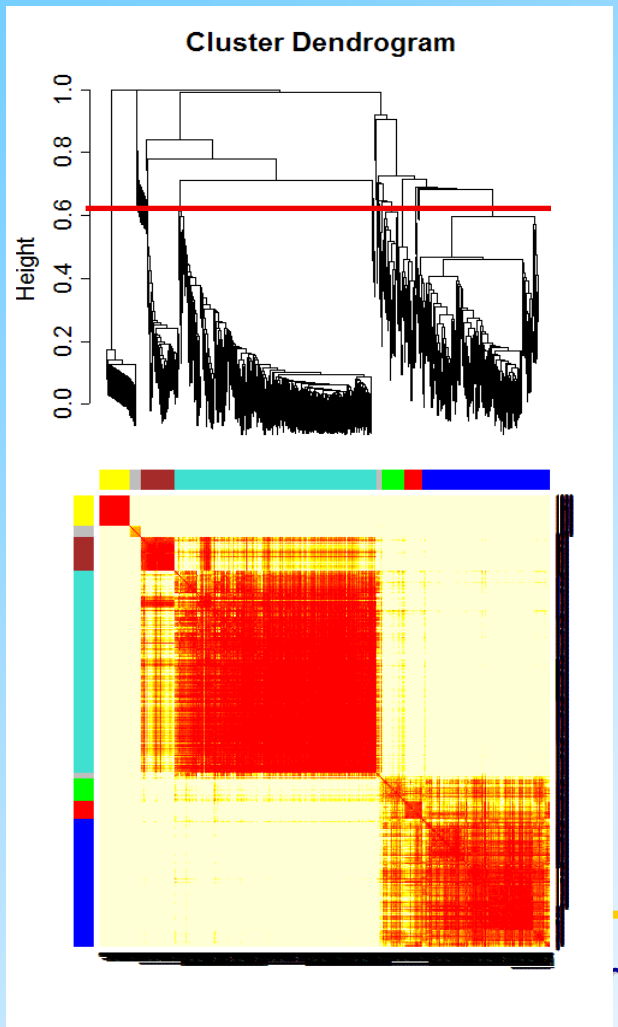


Sigmoid AF

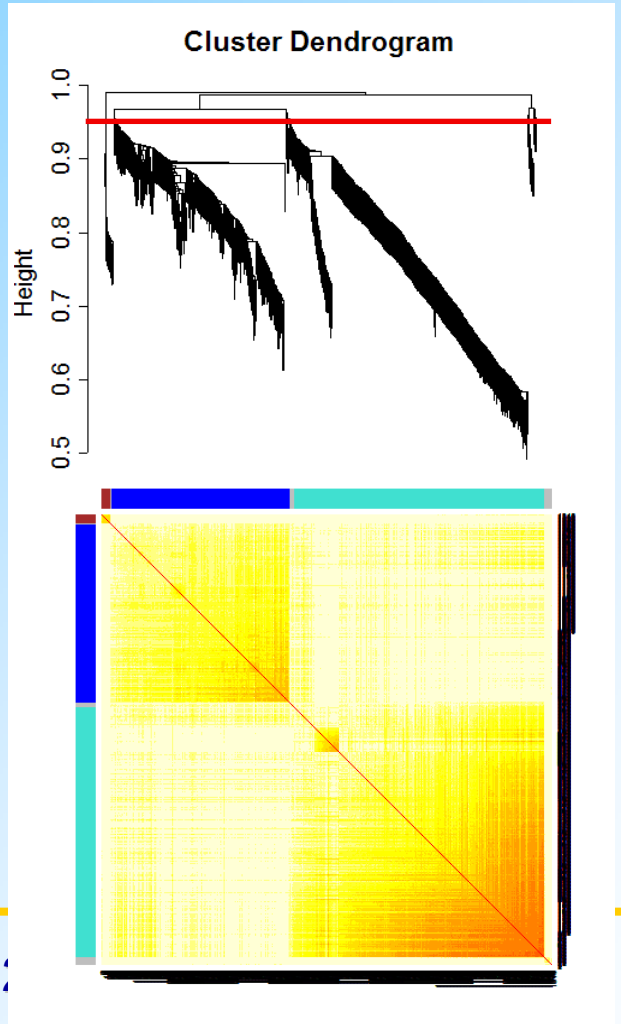


In practice, module detection is relatively robust to choice of Adjacency Function (AF)

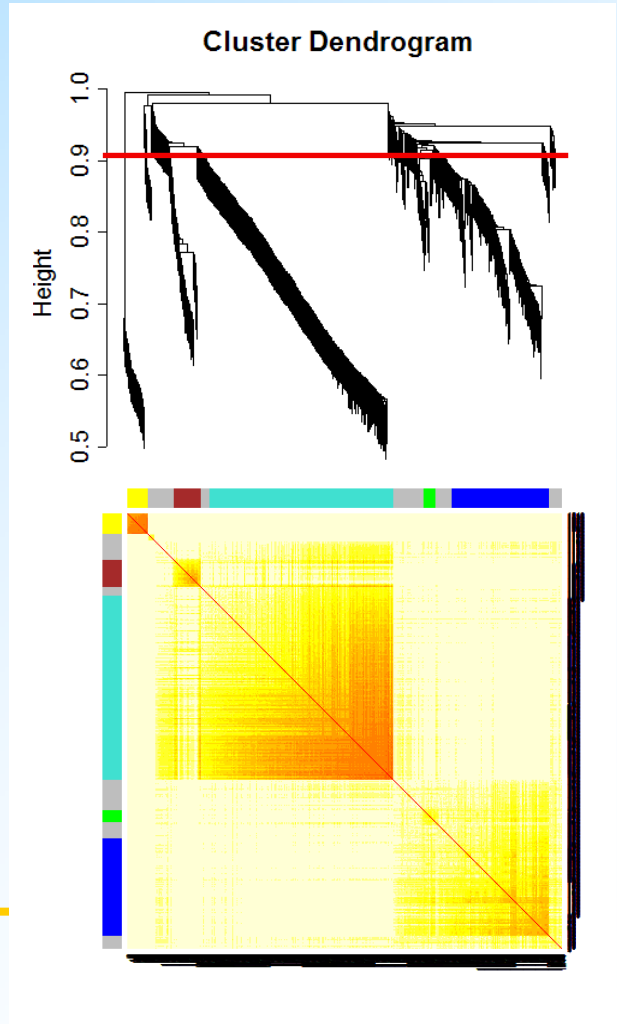
Step AF (tau)



Power AF (b)



Sigmoid AF

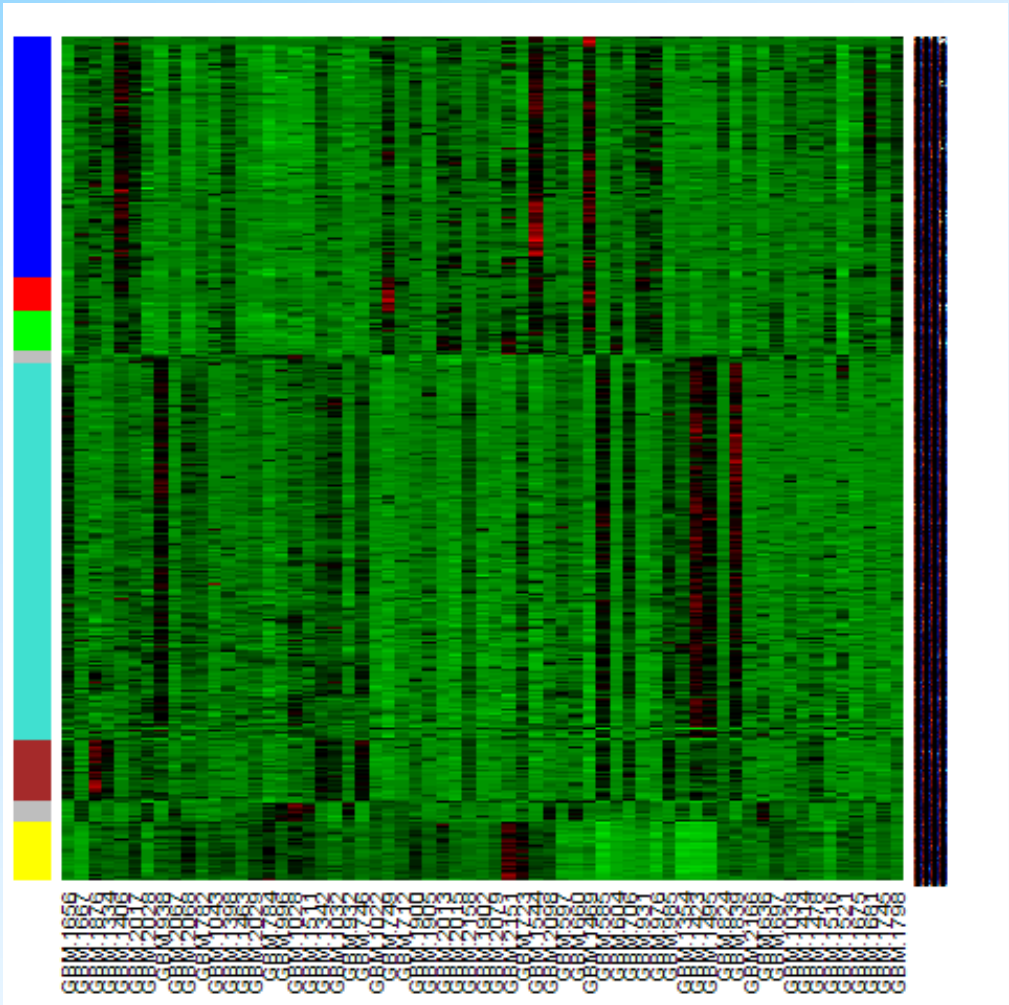


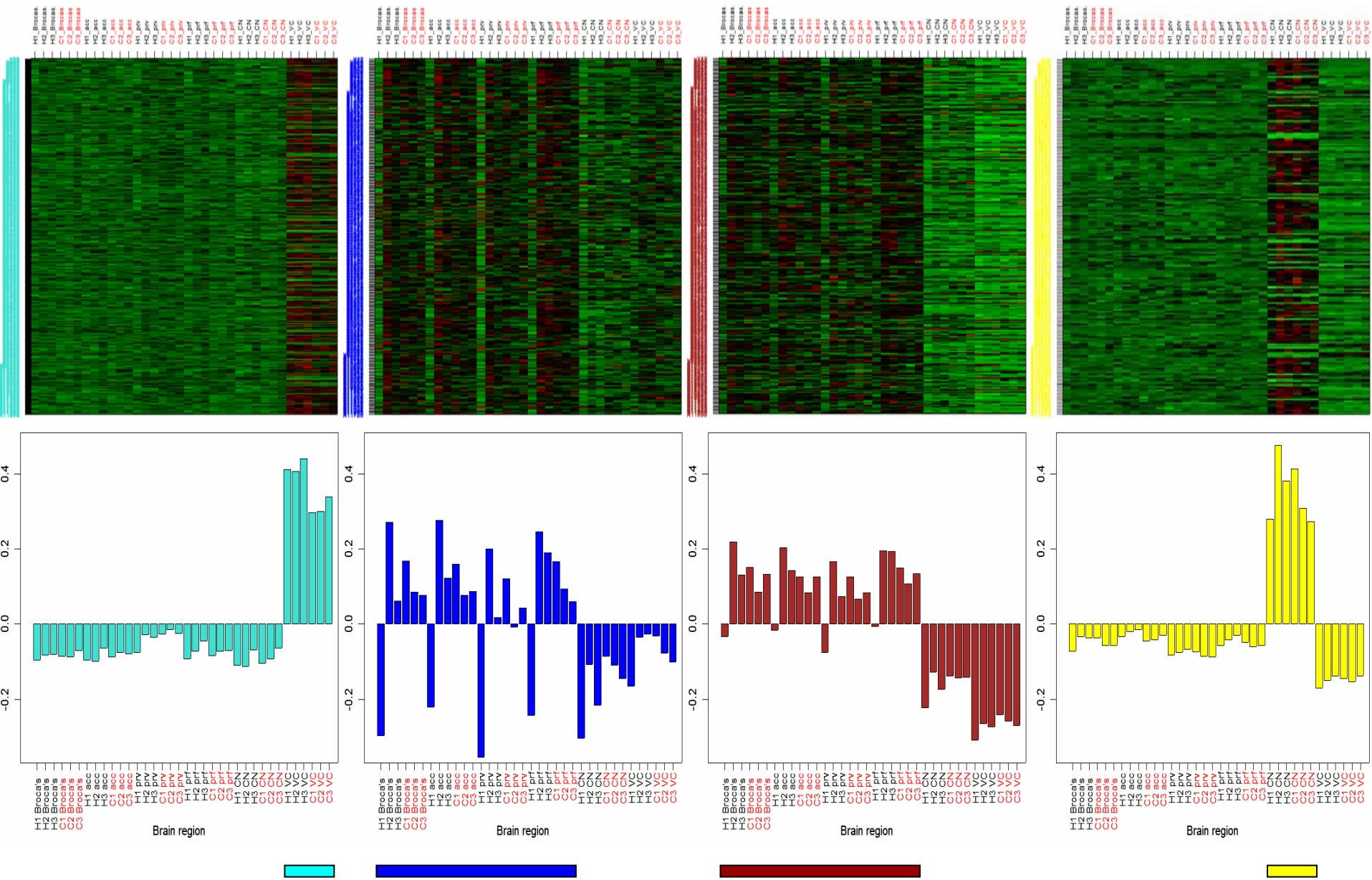
Identifying Gene Co-expression Modules

Columns=Brain tissue samples

Rows = Genes
Color bands indicate modules

Characteristic vertical bands indicate tight co-expression of module genes

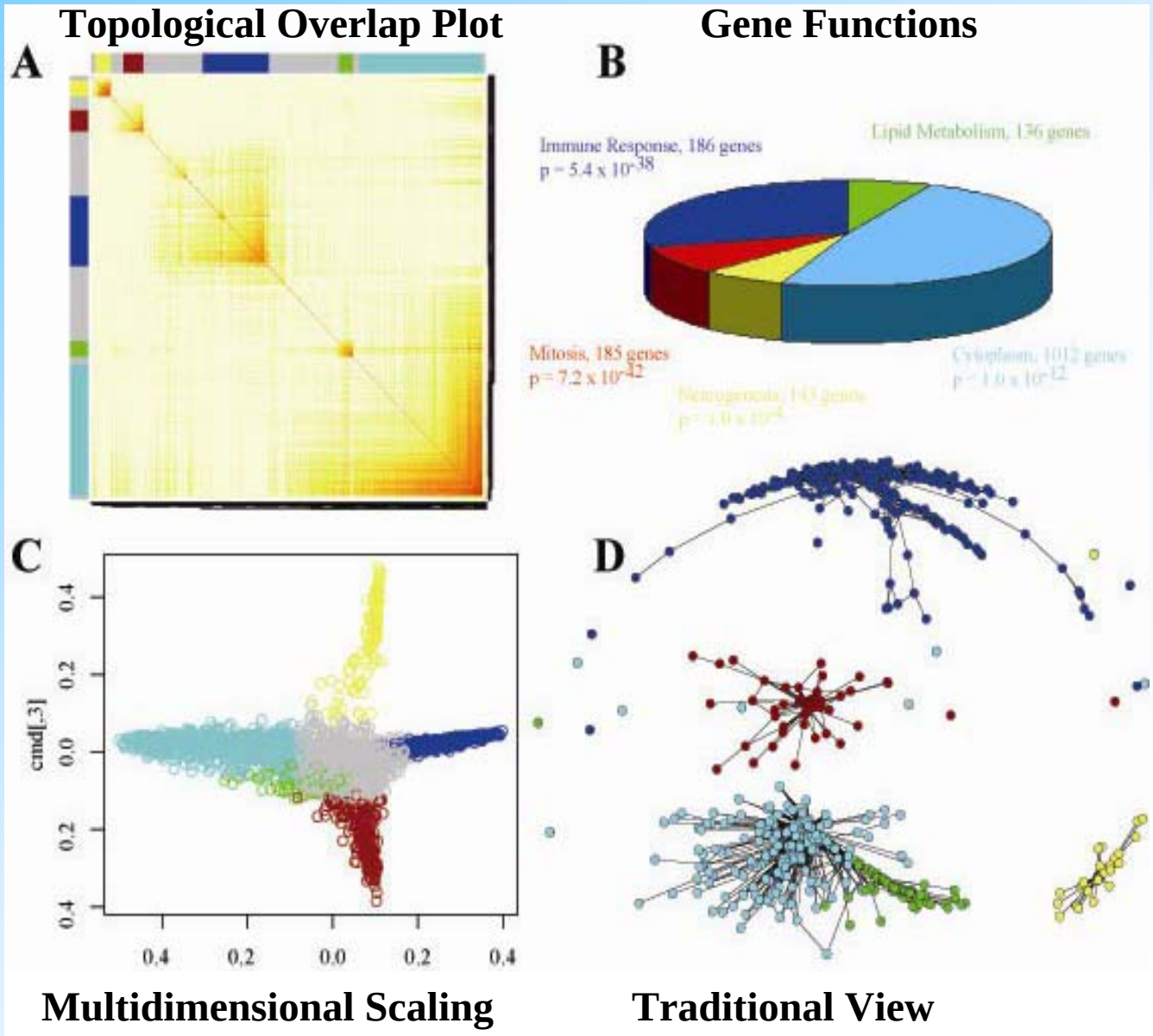




$p = 1.33 \times 10^{-4}$ $p = 8.93 \times 10^{-4}$ $p = 1.35 \times 10^{-6}$ $p = 1.33 \times 10^{-4}$

Different Ways of Depicting Gene Modules

- Rows and columns correspond to genes
- Red boxes along diagonal are modules
- Color bands=modules



Comparing Human and Chimp Brains

Mike Oldham, Steve Horvath, Dan Geschwind

UCLA HUMAN Genetics

Comparing Human and Chimp Brains

- Only six million years separate chimp...

…and man

UCLA HUMAN Genetics

What changed?

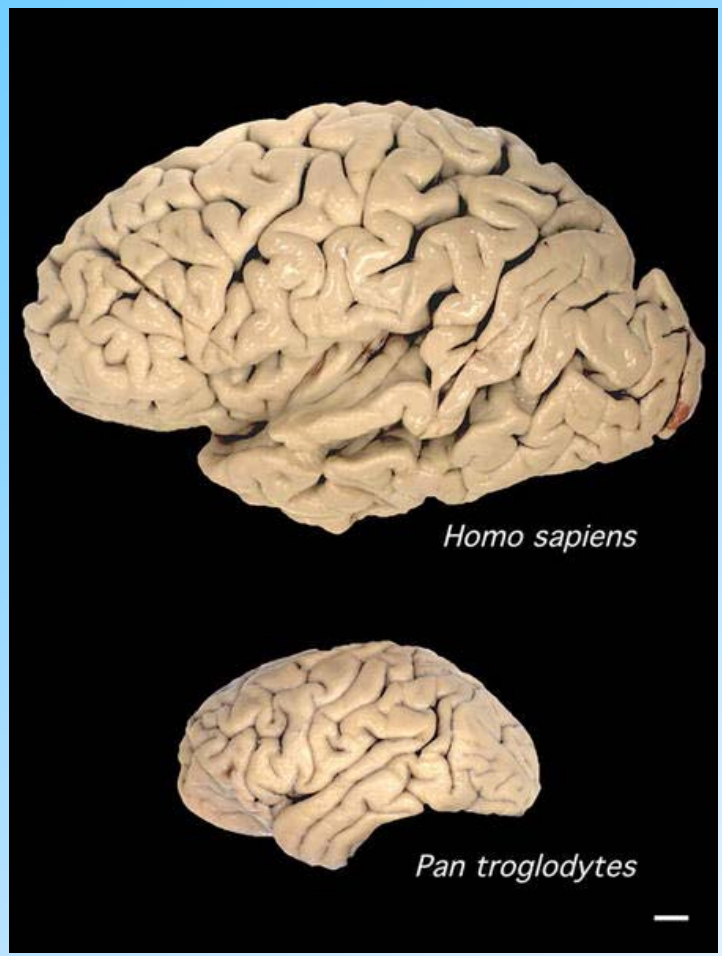


Image courtesy of Todd Preuss
(Yerkes National Primate Research Center)

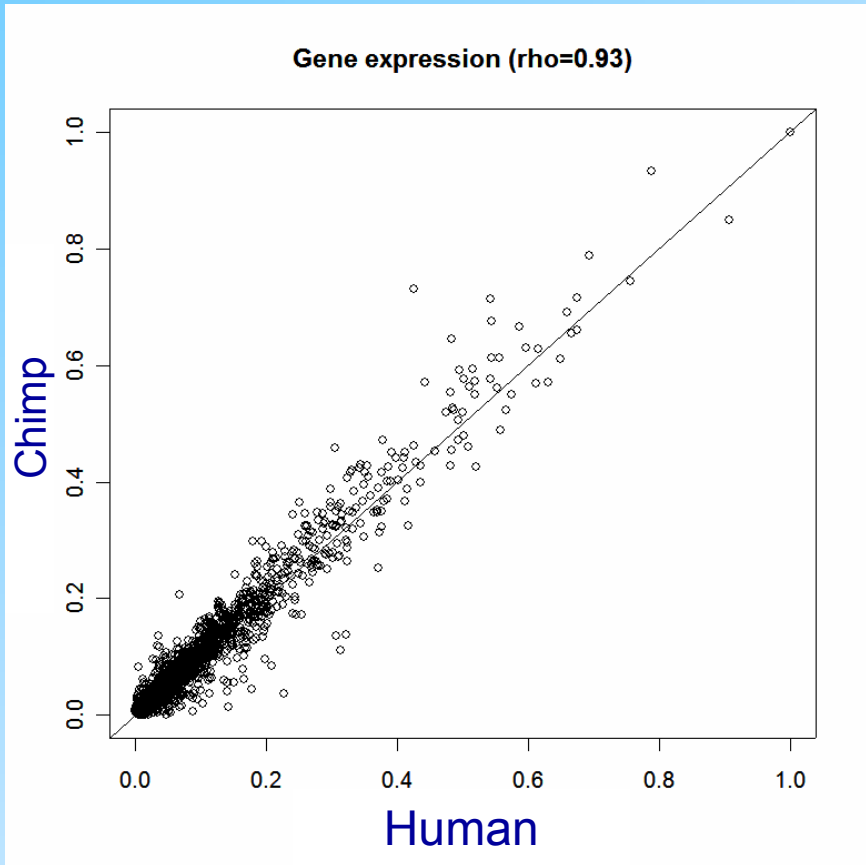
- Despite pronounced phenotypic differences, genomic similarity is ~96% (including single-base substitutions and indels)
- Similarity is even higher in protein-coding regions

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

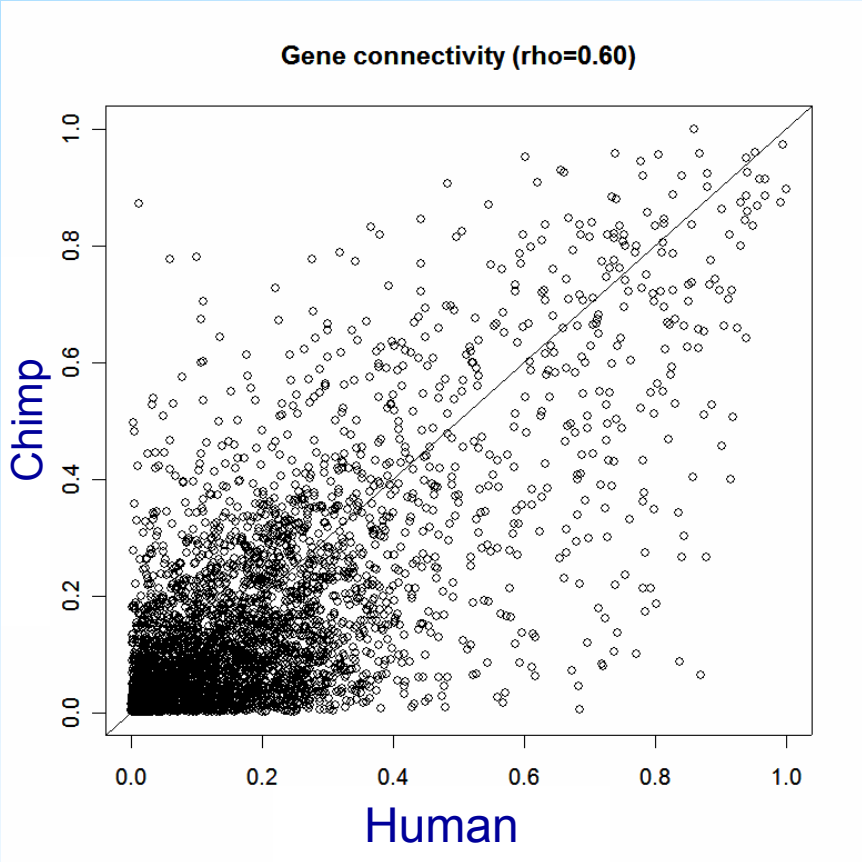
UCLA HUMAN Genetics

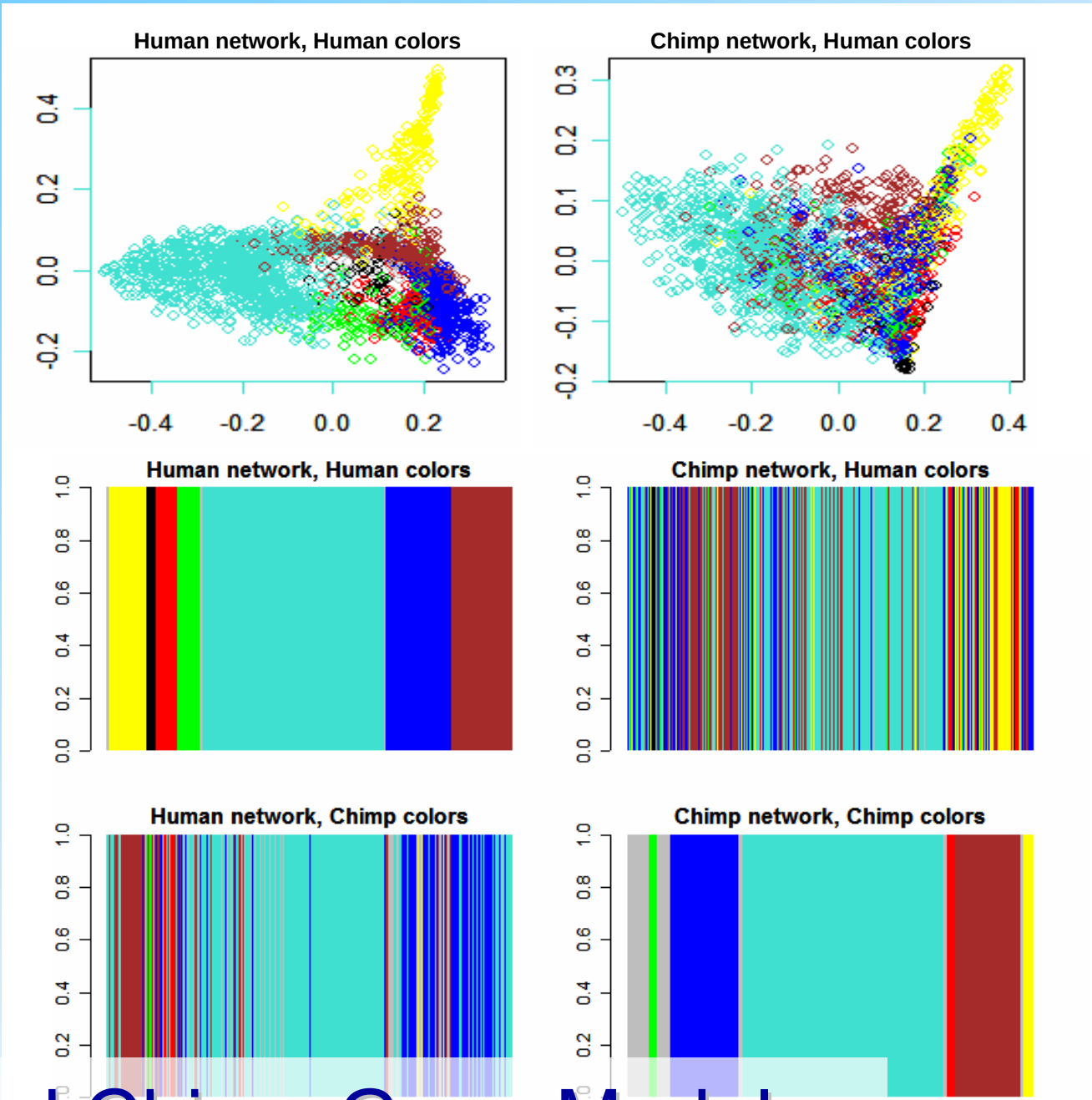
In the Brain Gene Expression is More Strongly Preserved than Gene Connectivity

Expression



Connectivity





Human and Chimp Gene Modules

Comparing Human and Chimp Brains

- Gene Expression is highly preserved across species brains
- Gene Co-expression is less preserved
- Gene modules correspond roughly to brain architecture

Conclusion: Molecular wiring makes us human

Integrating Gene Co-expression Networks With Genetic Marker Data in Study of Chronic Fatigue Syndrome

UCLA HUMAN Genetics

CAMDA

CRITICAL ASSESSMENT OF MICROARRAY DATA ANALYSIS

Publicly available data set -

- Clinical Data
- Expression Data
- SNP Data

UCLA HUMAN Genetics

Chronic Fatigue Syndrome

- Complex Disease
- Diagnosis – a minimum of six months of medically unexplained, debilitating fatigue
- Other symptoms:
  - elevated levels of cortisol due to an overactive hypothalamic-pituitary-adrenal (HPA) axis
  - altered immune response substantiated by high T-cell counts
  - skeletal muscle dysfunction
- May be triggered by viral infection
- Expression studies report over-expression of immune response genes

Chronic Fatigue Syndrome

DNA Level: ~ 50 Pre-selected SNP's

2. Relate SNP data to Expression data

mRNA Level: ~ 20K genes/array

1. Relate Expression data to Clinical Trait data

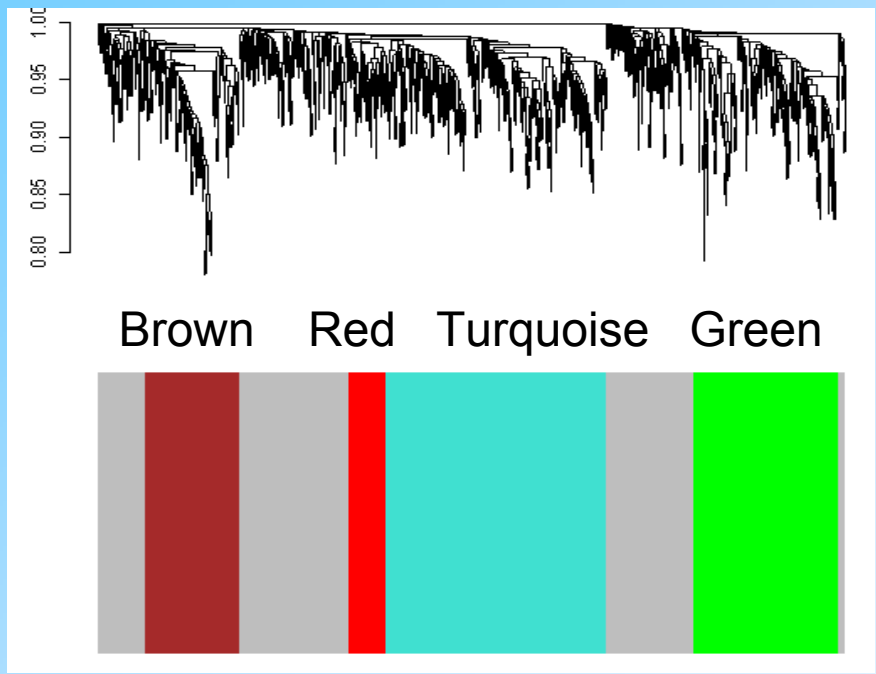
Organism Level: ~ 70 Clinical Traits

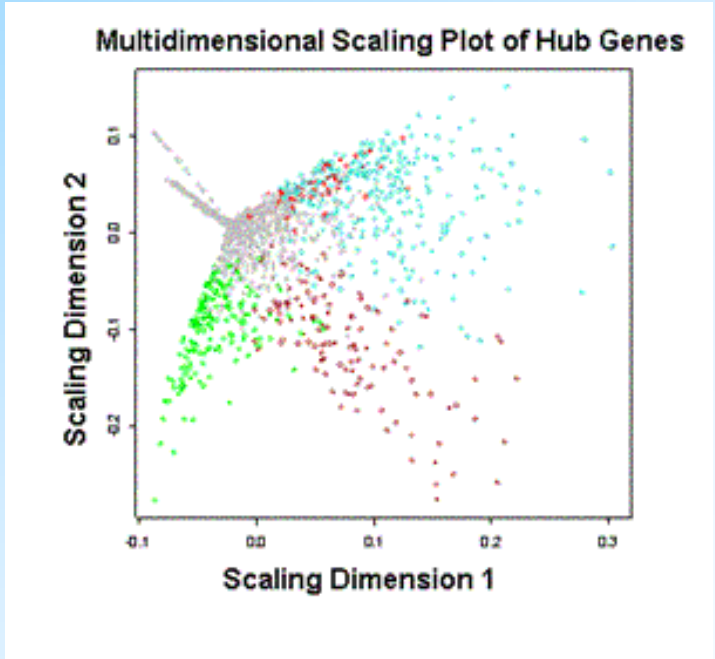
3. Integrate results to find CFS relevant genes.

Analysis Overview

1. Construct gene co-expression network from microarray data
2. Identify module of interest using trait data.
3. Determine informative SNP's and relate them to gene co-expression network.
4. Identify genes with statistical and biological significance.
5. Choose subset of CFS and control samples for validating the candidate biomarker.

Four Modules Identified Using Hierarchical Clustering





- Grey colors indicate genes outside of any module.
- MDS plot indicates clear separation of brown, green, turquoise modules.

UCLA HUMAN Genetics

Analysis Overview

1. Construct gene co-expression network from microarray data. (Zhang and Horvath 2005)
2. Identify module of interest using trait data.
3. Determine informative SNP's and relate them to gene co-expression network.
4. Identify genes with statistical and biological significance.
5. Choose subset of CFS and control samples for validating the candidate biomarker.

A clinical trait gives rise to a "Trait Significance" measure

TraitSignificance(i) = |cor(x(i), TRAIT)|

where $x(i)$ is the gene expression profile of the i^{th} gene.

Module Trait Significance = Average(Trait Significance values for genes in a module)

UCLA HUMAN Genetics

Trait Significance Results

- Table shows trait significance for each module.
- Every module was characterized in terms of a group of clinical traits.
- Interested in CFS severity trait “CLUSTER” a composite score of 14 clinical traits (evaluation responses).
- Focused on the green module (184 genes) since it was related to the CLUSTER trait.

	Module Trait Significance (Standard Error)				
Clinical Traits	Turquoise	Grey	Red	Brown	Green
Shortness of Breath	0.176 (0.003)	0.096 (0.001)	0.162 (0.009)	0.107 (0.005)	0.078 (0.004)
Mental Health	0.189 (0.003)	0.105 (0.002)	0.215 (0.005)	0.188 (0.004)	0.144 (0.004)
Role Emotional	0.292 (0.003)	0.139 (0.002)	0.336 (0.005)	0.217 (0.005)	0.172 (0.004)
Sinus Nasal	0.06 (0.003)	0.076 (0.001)	0.048 (0.005)	0.135 (0.004)	0.133 (0.004)
Muscle Pain	0.108 (0.002)	0.076 (0.001)	0.085 (0.006)	0.116 (0.003)	0.059 (0.002)
Unrefreshing Sleep	0.092 (0.002)	0.071 (0.001)	0.064 (0.004)	0.12 (0.003)	0.054 (0.002)
CLUSTER	0.102 (0.003)	0.105 (0.002)	0.131 (0.008)	0.115 (0.006)	0.216 (0.004)
Abdominal Pain	0.069 (0.003)	0.082 (0.001)	0.091 (0.009)	0.094 (0.005)	0.122 (0.005)

Trait Significance Results

- Table shows trait significance for each module.
- Every module was characterized in terms of a group of clinical traits.
- Interested in CFS severity trait “CLUSTER” a composite score of 14 clinical traits (evaluation responses).

➢ Focused on the green module (184 genes) since it was related to the CLUSTER trait.

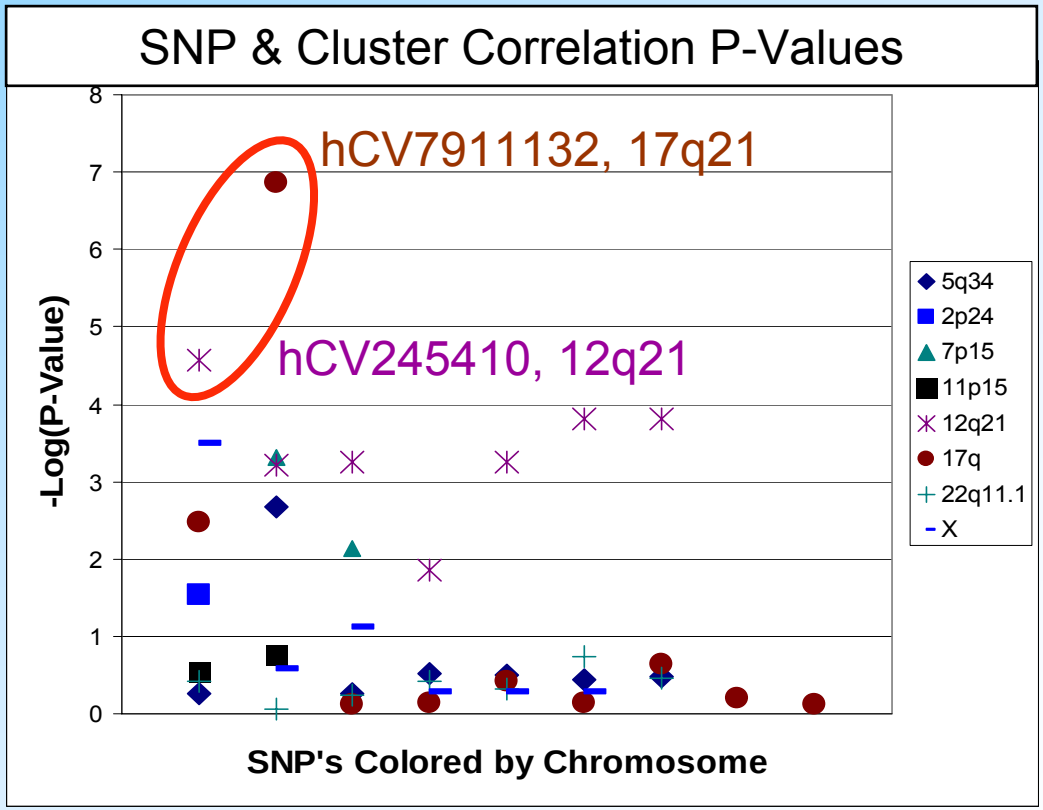
	Module Trait Significance (Standard Error)				
Clinical Traits	Turquoise	Grey	Red	Brown	Green
Shortness of Breath	0.176 (0.003)	0.096 (0.001)	0.162 (0.009)	0.107 (0.005)	0.078 (0.004)
Mental Health	0.189 (0.003)	0.105 (0.002)	0.215 (0.005)	0.188 (0.004)	0.144 (0.004)
Role Emotional	0.292 (0.003)	0.139 (0.002)	0.336 (0.005)	0.217 (0.005)	0.172 (0.004)
Sinus Nasal	0.06 (0.003)	0.076 (0.001)	0.048 (0.005)	0.135 (0.004)	0.133 (0.004)
Muscle Pain	0.108 (0.002)	0.076 (0.001)	0.085 (0.006)	0.116 (0.003)	0.059 (0.002)
Unrefreshing Sleep	0.092 (0.002)	0.071 (0.001)	0.064 (0.004)	0.12 (0.003)	0.054 (0.002)
CLUSTER	0.102 (0.003)	0.105 (0.002)	0.131 (0.008)	0.115 (0.006)	0.216 (0.004)
Abdominal Pain	0.069 (0.003)	0.082 (0.001)	0.091 (0.009)	0.094 (0.005)	0.122 (0.005)

Analysis Overview

1. Construct gene co-expression network from microarray data. (Zhang and Horvath 2005)
2. Identify module of interest using trait data.
3. Determine informative SNP's and relate them to gene co-expression network.
4. Identify genes with statistical and biological significance.
5. Choose subset of CFS and control samples for validating the candidate biomarker.

Finding SNPs Correlated with the CLUSTER Trait

- We chose two SNPs with highest CLUSTER correlation
- SNP12 = hCV245410 on 12q21 (p-value = 0.01)
- SNP17 = hCV7911132 on 17q21 (p-value = 0.001)



Correlation with relevant SNPs defines "SNP Significance" of the ith gene

$$SNPSignificance = |cor(x(i), SNP)|$$

(Where SNP data is additively coded)

- Conceptually related to a LOD score at the SNP marker for the i^{th} gene expression.
- Why correlate SNP and gene expression data?
  - Puts SNP effect on the same footing as trait effect and gene-gene connection strengths. Effect sizes are important in our analysis.

SNP Filtering & Significance Results

- Table shows the average SNP significance for each module.
- Green module genes most correlated with SNP12.
- "SNP12 Sub-sample" = average module correlations with SNP12 among samples that have a particular SNP12 and SNP17 genotype.
- Higher correlation(green module,SNP12) in the sample subset.

		Module SNP Significance (Standard Error)			
SNPs	Turquoise	Grey	Red	Brown	Green
SNP12	0.052 (0.002)	0.077 (0.001)	0.036 (0.004)	0.091 (0.004)	0.128 (0.004)
SNP17	0.056 (0.002)	0.064 (0.001)	0.045 (0.005)	0.039 (0.003)	0.04 (0.002)
SNP12 Sub-sample	0.128 (0.005)	0.144 (0.002)	0.067 (0.009)	0.203 (0.007)	0.186 (0.007)

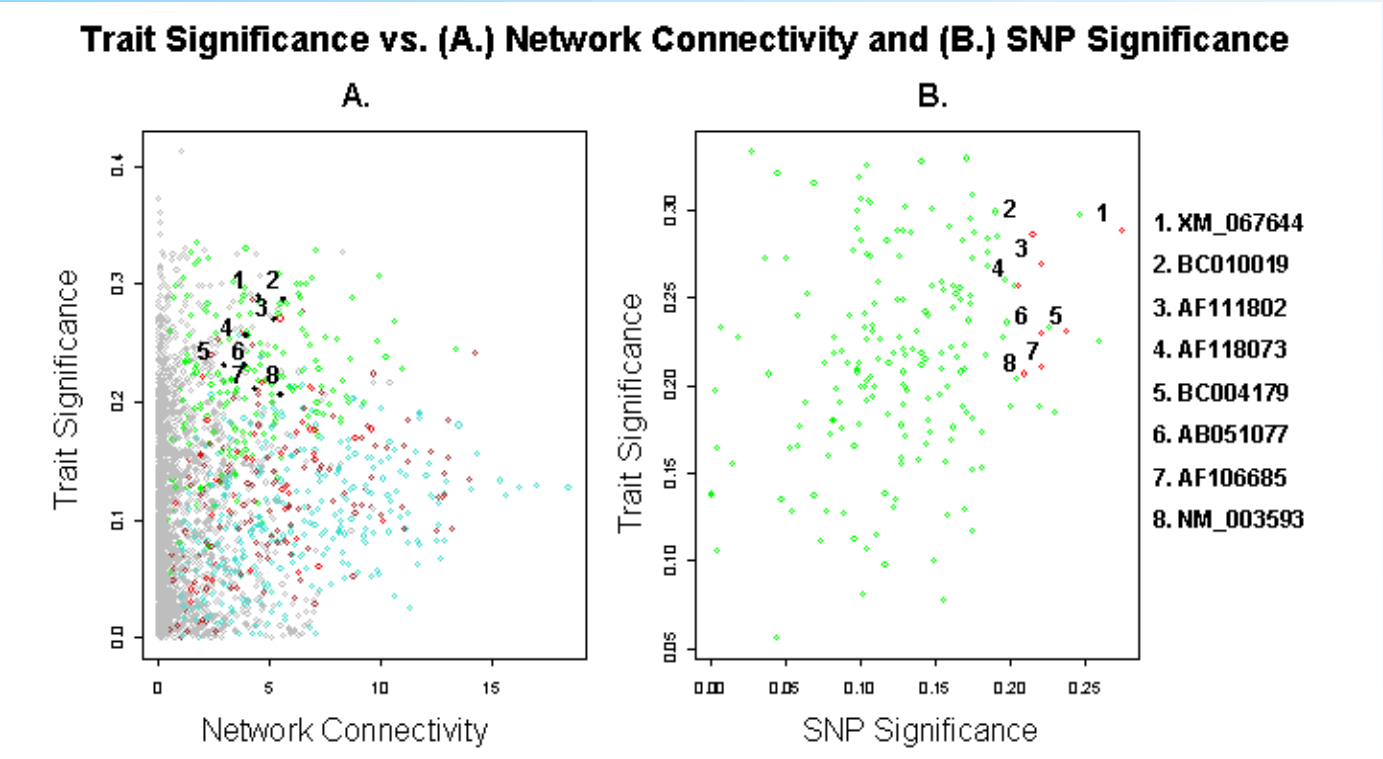
Analysis Overview

1. Construct gene co-expression network from microarray data. (Zhang and Horvath 2005)
2. Identify module of interest using trait data.
3. Determine informative SNP's and relate them to gene co-expression network.
4. Identify genes with statistical and biological significance.
5. Choose subset of CFS and control samples for validating the candidate biomarker.

Integration of Genetic and Network Analysis

Combined Gene Selection Criteria:

1. CLUSTER trait significance > 0.2
2. SNP12 significance > 0.2
3. Genes with high intramodular connectivity (top 50%)
4. Member of the Green Module



Eight Most Significant Genes:

Accession	Gene Symbol (Name) and Information	Locus	P-Value (Correlation)		Biomarker
			CLUSTER	SNP	
NM_003593	FOXN1 (forkhead box N1): Functions in defense response, T-cell immunodeficiency, and known to cause nudity in mice and humans. Expressed in thymus.	17q11-q12	0.055 (-0.21)	0.018 (0.21)	YES
AF118073	PRDX3 (peroxiredoxin 3): Regulates cell proliferation, differentiation, and antioxidant functions.	10q25-q26	0.017 (-0.26)	0.02 (0.21)	YES
AB051077	PEX6 (peroxisomal biogenesis factor 6): absence results in zellweger syndrome (zws), neurological and metabolic defects.	6p21.1	0.032 (-0.23)	0.013 (0.22)	YES
AF106685	MYEF2 (myelin expression factor 2): myoblast cell differentiation and transcription.	15q21.1	0.05 (-0.21)	0.012 (0.22)	YES
AF111802	CRNKL1 (Crn, crooked neck-like 1 (Drosophila)): expressed in testes, involved in mRNA splicing	20p11.2	0.012 (-0.27)	0.013 (0.22)	YES
BC010019	MED8 (mediator of RNA polymerase II transcription, subunit 8 homolog (yeast)): regulates transcription.	1p34.2	0.007 (-0.29)	0.015 (0.22)	YES
XM_067644	Similar to polynucleotide phosphorylase-like protein and 3-5 RNA exonuclease.		0.007 (-0.29)	0.002 (0.27)	NO
BC004179	Unknown (protein for mgc:2780)		0.032 (-0.23)	0.007 (0.24)	NO

Source: NCBI (http://www.ncbi.nlm.nih.gov)

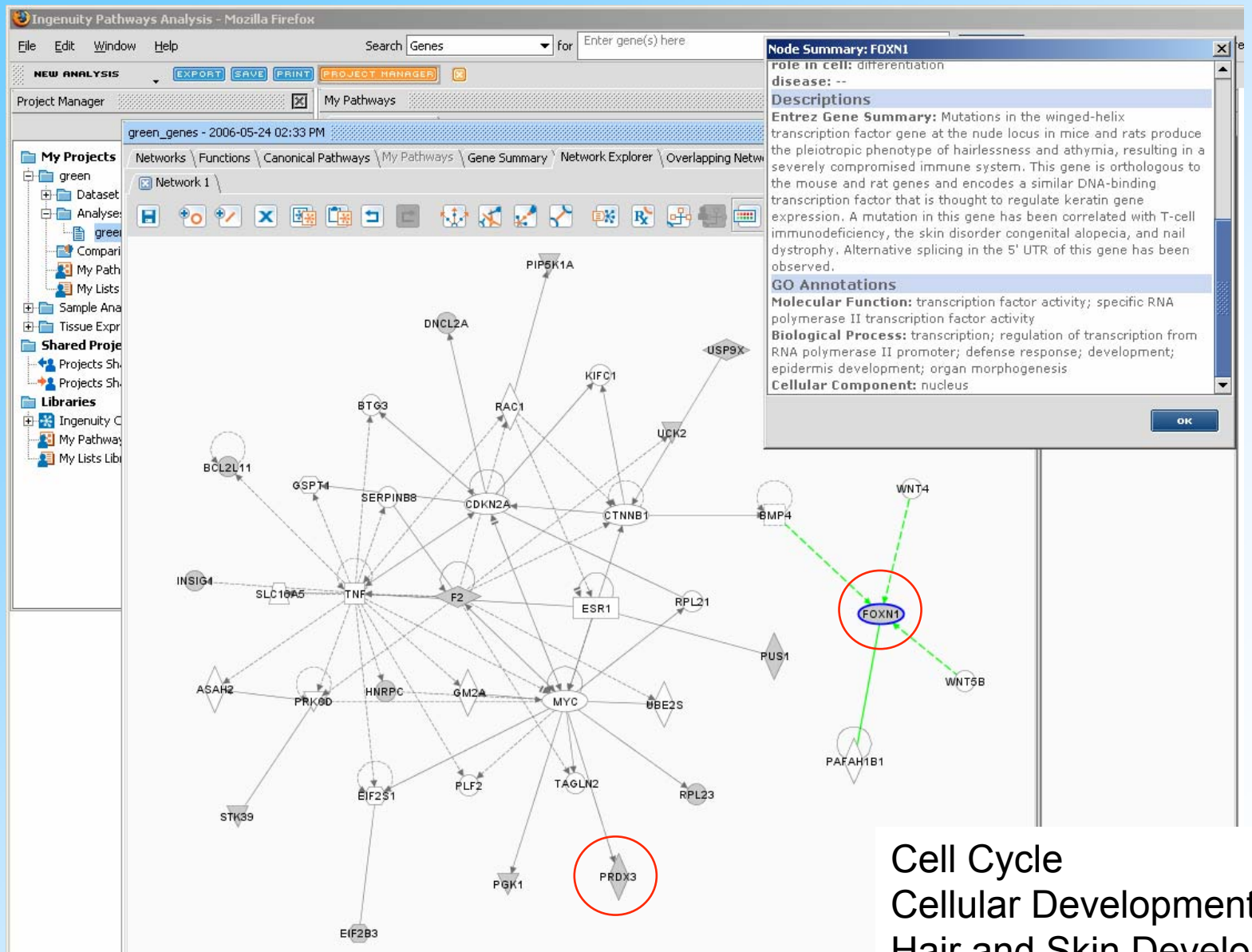
FOXN1 Statistical Significance:

- Member of the green module that is related to the CFS severity trait (CLUSTER)
- Significantly associated with SNP 12 (p-value = 0.0179), which is significantly associated with CLUSTER (p-value = 0.010)
- High intramodular network connectivity
- Moderate direct correlation with the CLUSTER trait

FOXN1 Biological Significance

- Mutations in mice & humans cause:
  - Nudity.
  - Depleted immune system due to dysfunctional T-cells.
- Highly expressed in thymus epithelia cells.
- Thymus involved in immune system:
  - Converts lymphocytes to T-cells.
  - Releases functional T-cells to combat infection.

Ingenuity Pathway Analysis



Cell Cycle
Cellular Development
Hair and Skin Development

FOXN1: Validation for Chronic Fatigue Syndrome

CFS patients have an overactive immune system with high T cell production and T cell abnormalities

⇒ FOXN1 may be highly expressed in CFS.

To further investigate this finding?

⇒ A FOXN1 knockout mouse available

⇒ Explore the relationship between FOXN1 and fatigue in a mouse model

UCLA HUMAN Genetics

Analysis Overview

1. Construct gene co-expression network from microarray data. (Zhang and Horvath 2005)
2. Identify module of interest using trait data.
3. Determine informative SNP's and relate them to gene co-expression network.
4. Identify genes with statistical and biological significance.
5. Choose subset of CFS and control samples for validating the candidate biomarker.

Relationship between FOXN1 and SNP12 & 17 genotypes

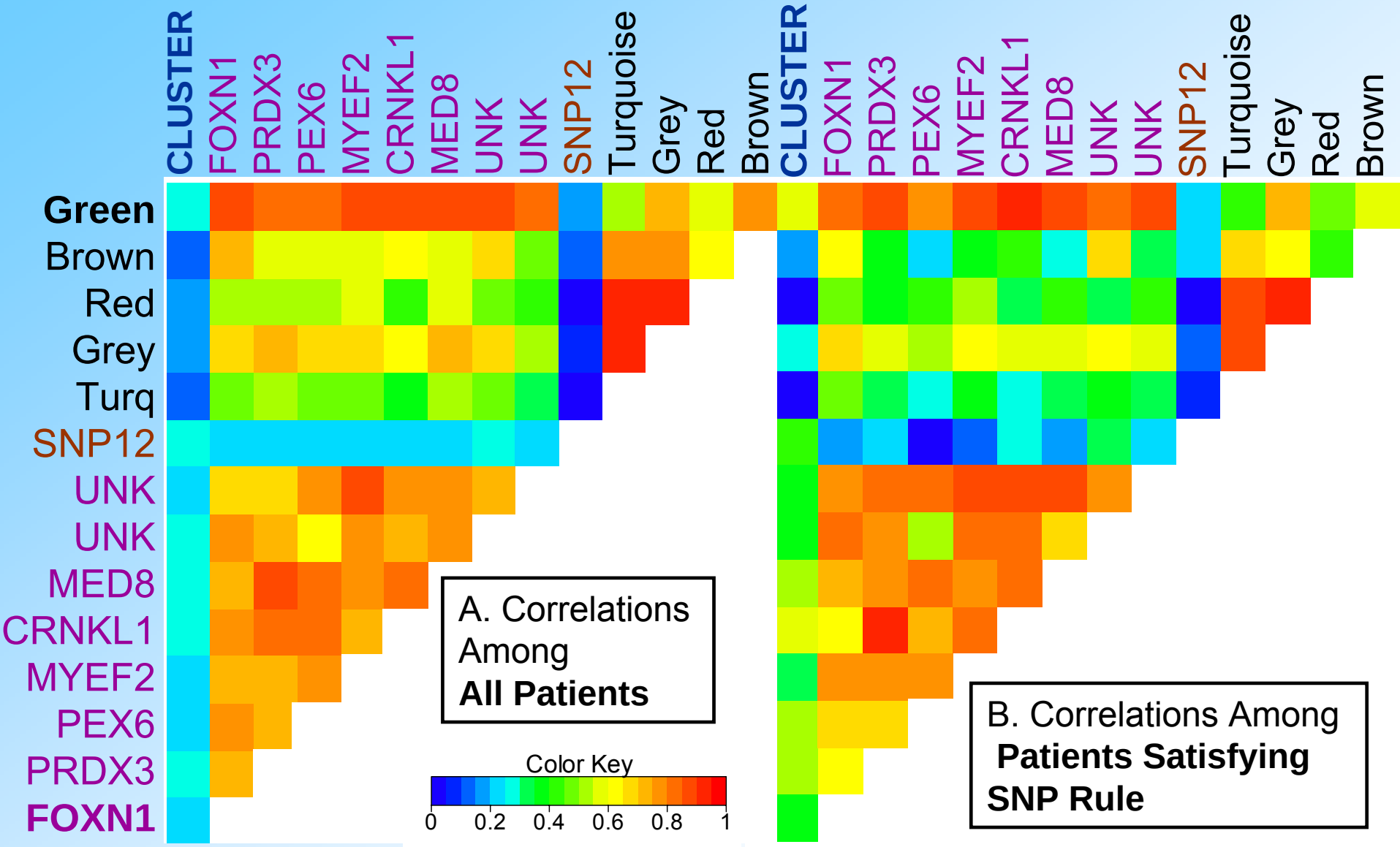
- The two SNP's most correlated with the CLUSTER phenotype – most differentially expressed between cases and controls – identify a sub-phenotype of CFS.
- **SNP rule:**

	SNP 12		SNP 17
	0	+	2
or			
	1	+	2

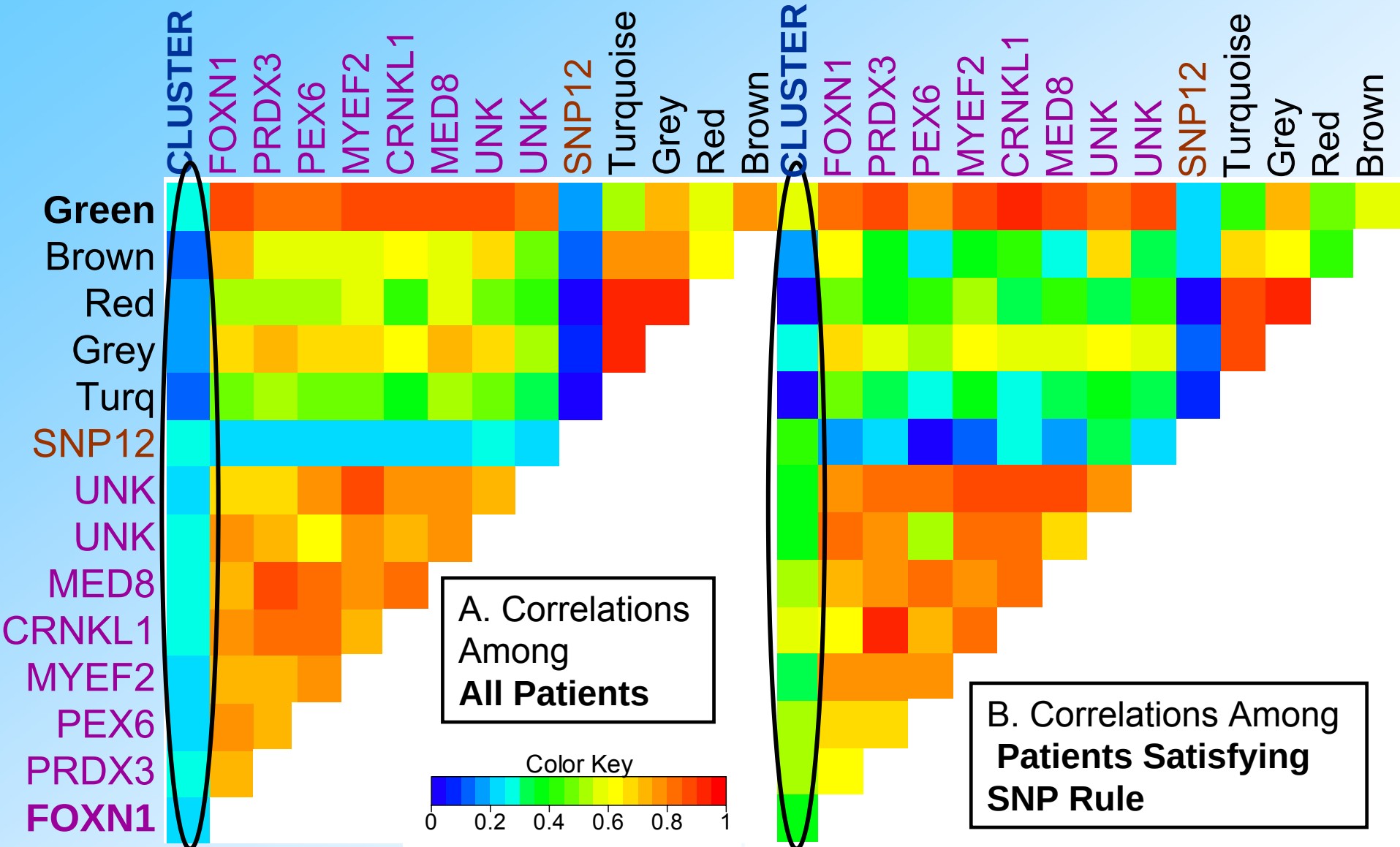
  We define a sample subgroup where all individuals have 0+2 or 1+2 genotypes.
- About 1/3 of the samples satisfy the SNP rule.
  - For these samples FOXN1 is useful for predicting CFS severity.

UCLA HUMAN Genetics

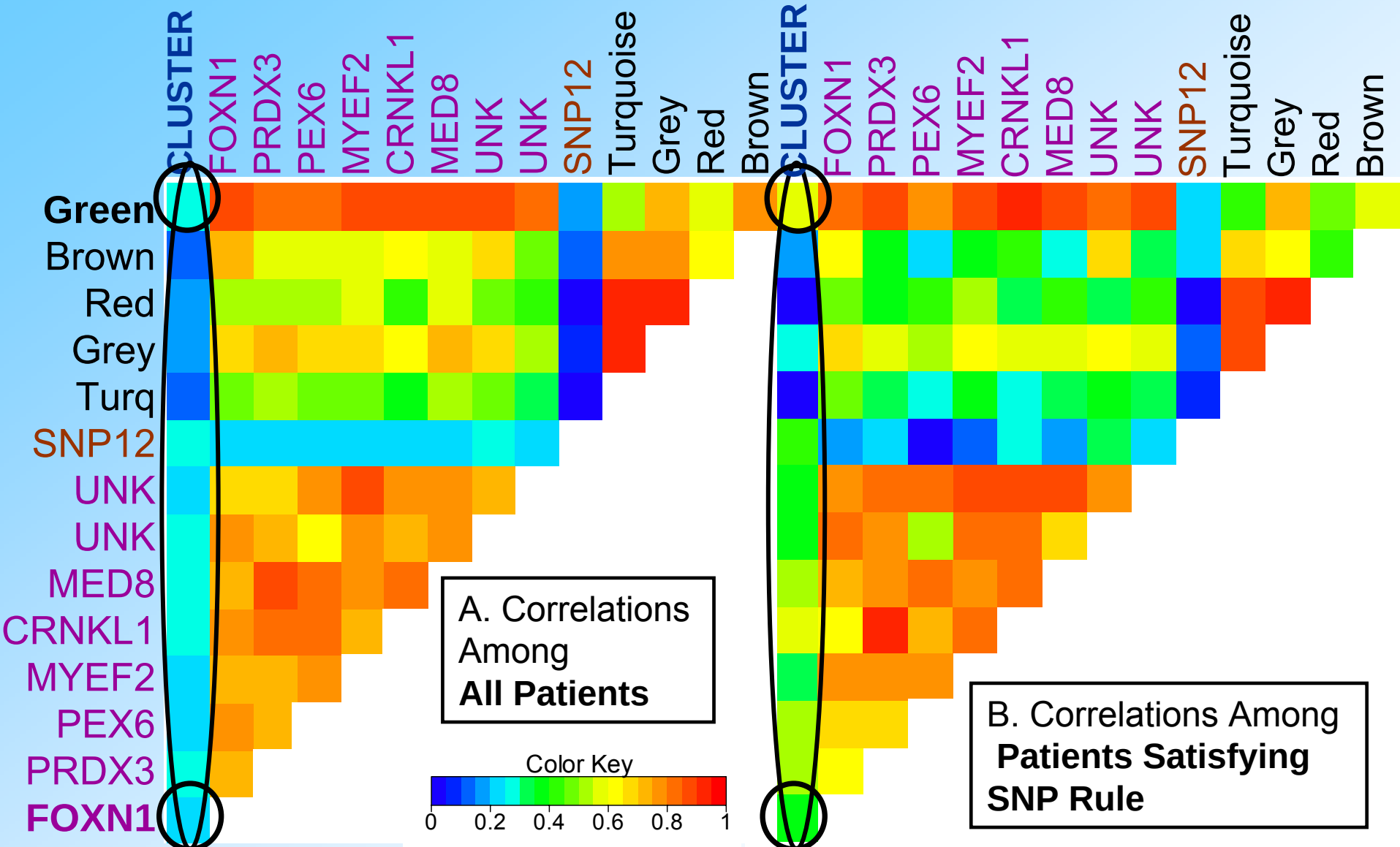
“SNP Rule” Aids in Patient Selection



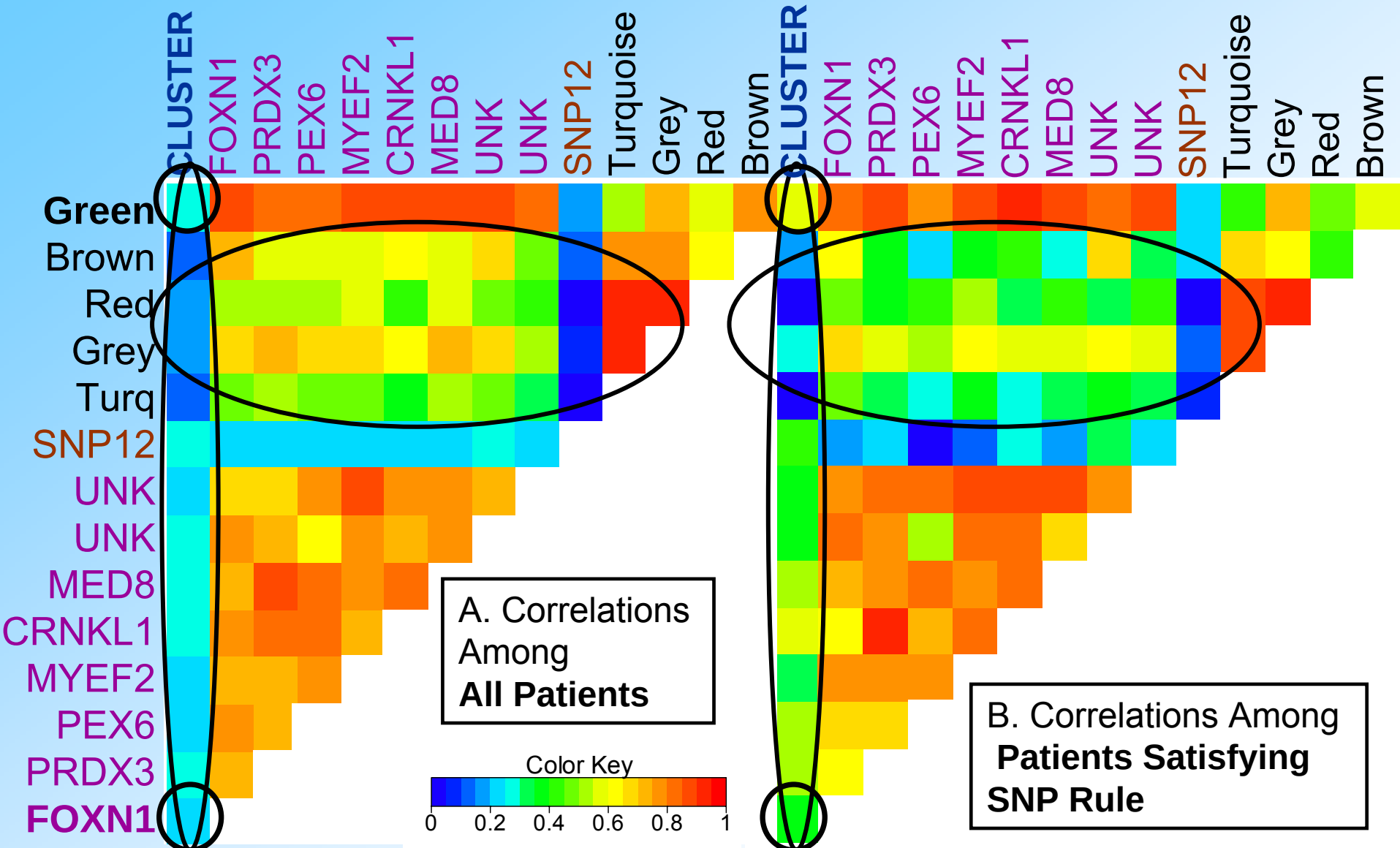
“SNP Rule” Aids in Patient Selection



“SNP Rule” Aids in Patient Selection



“SNP Rule” Aids in Patient Selection



SNP Filtering & Significance Results

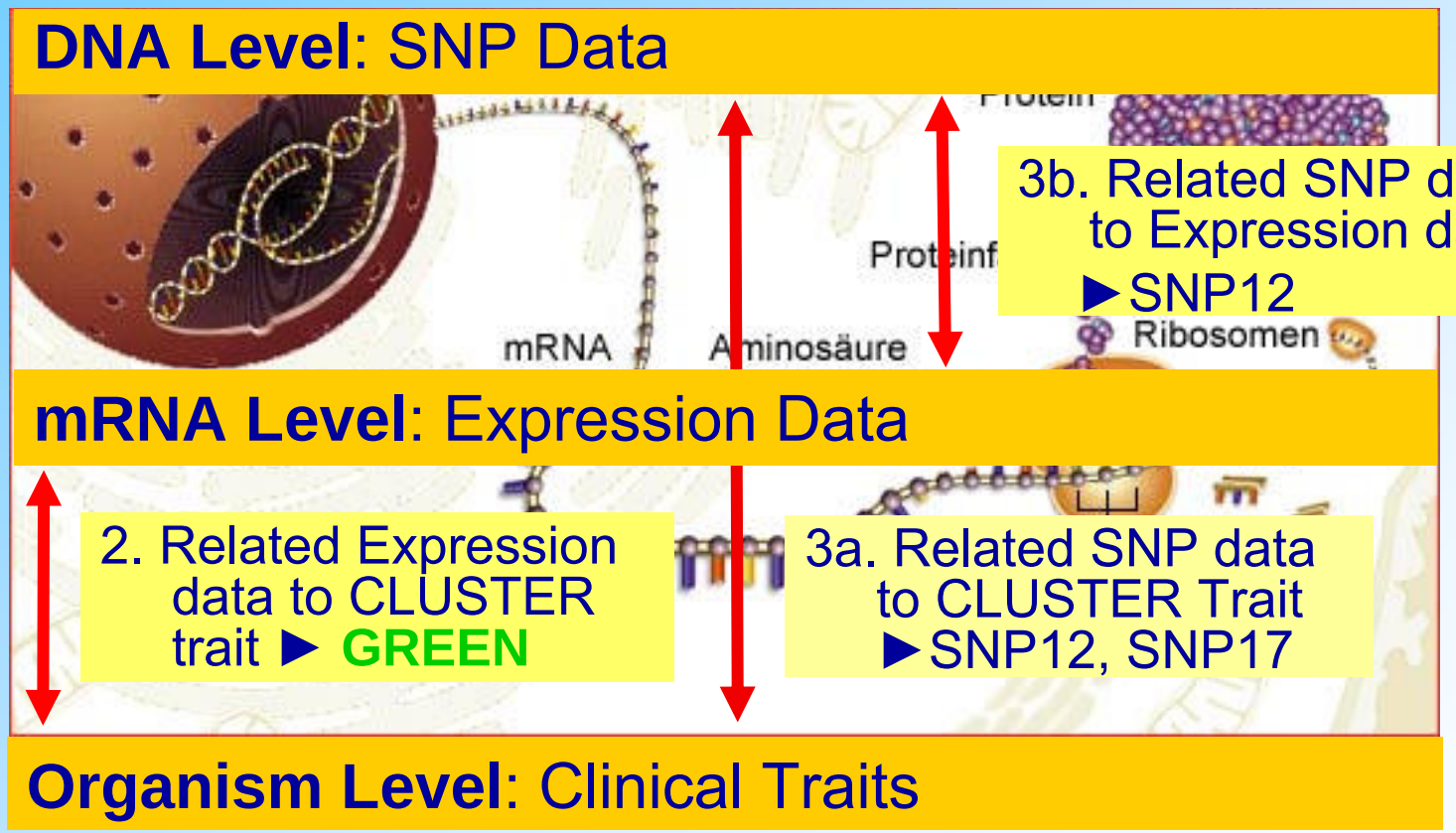
- Table shows the average SNP significance for each module.
- Green module genes most correlated with SNP12.
- “SNP12 Sub-sample” = average module correlations with SNP12 among samples that have a particular SNP12 and SNP17 genotype.
- Higher correlation(green module,SNP12) in the sample subset.

		Module SNP Significance (Standard Error)			
SNPs	Turquoise	Grey	Red	Brown	Green
SNP12	0.052 (0.002)	0.077 (0.001)	0.036 (0.004)	0.091 (0.004)	0.128 (0.004)
SNP17	0.056 (0.002)	0.064 (0.001)	0.045 (0.005)	0.039 (0.003)	0.04 (0.002)
SNP12 Sub-sample	0.128 (0.005)	0.144 (0.002)	0.067 (0.009)	0.203 (0.007)	0.186 (0.007)

UCLA HUMAN Genetics

Summary

1. Constructed gene co-expression network from the microarray data.



4. FOXN1 has statistical and biological significance
5. Highest differential FOXN1 expression in subgroup that has a particular SNP12 & SNP17 genotype

Conclusion

- **Network** approaches provide a means to bridge the gap from individual genes to systems biology.
- **Integrating** gene co-expression networks with genetic marker and trait information helps us understand what factors influence the relationship between gene expression and biological pathways

UCLA HUMAN Genetics

Acknowledgements

- Steve Horvath
  - Lin Wang
  - Jun Dong
  - Chi-ying Lee
  - Ai Li
  - Bin Zhang
  - Wei Zhao

Dan Geschwind
Mike Oldham

Eric Sobel
Jenny Papp
Anja Presson

UCLA HUMAN Genetics

Network Construction

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.

UCLA HUMAN Genetics