Introduction to weighted gene co-expression network analysis (WGCNA)

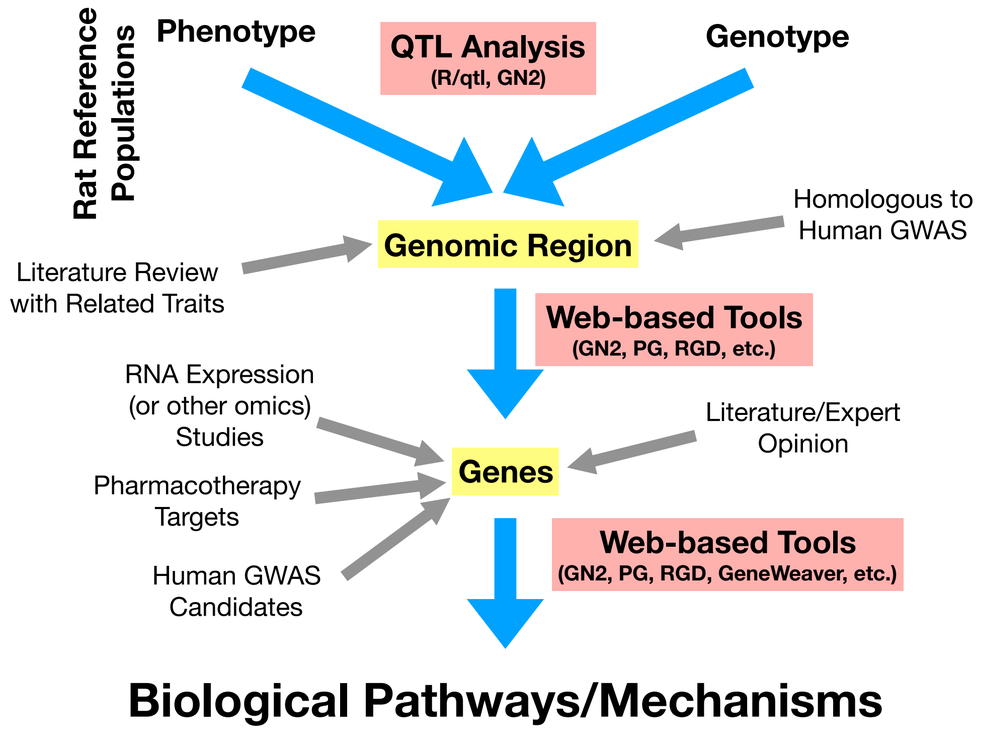
7th Webinar for Quantitative Genetics Tools for Mapping Trait Variation to Mechanisms, Therapeutics, and Interventions

Laura Saba, PhD
University of Colorado Anschutz Medical Campus
NIDA Center of Excellence in Omics, Systems Genetics and the Addictome

Quantitative Genetics Tools for Mapping Trait Variation to Mechanisms, Therapeutics, and Interventions Webinar Series

Goal of the Series:
Transverse the path from trait variance to QTL to gene variant to molecular networks to mechanisms to therapeutic and interventions

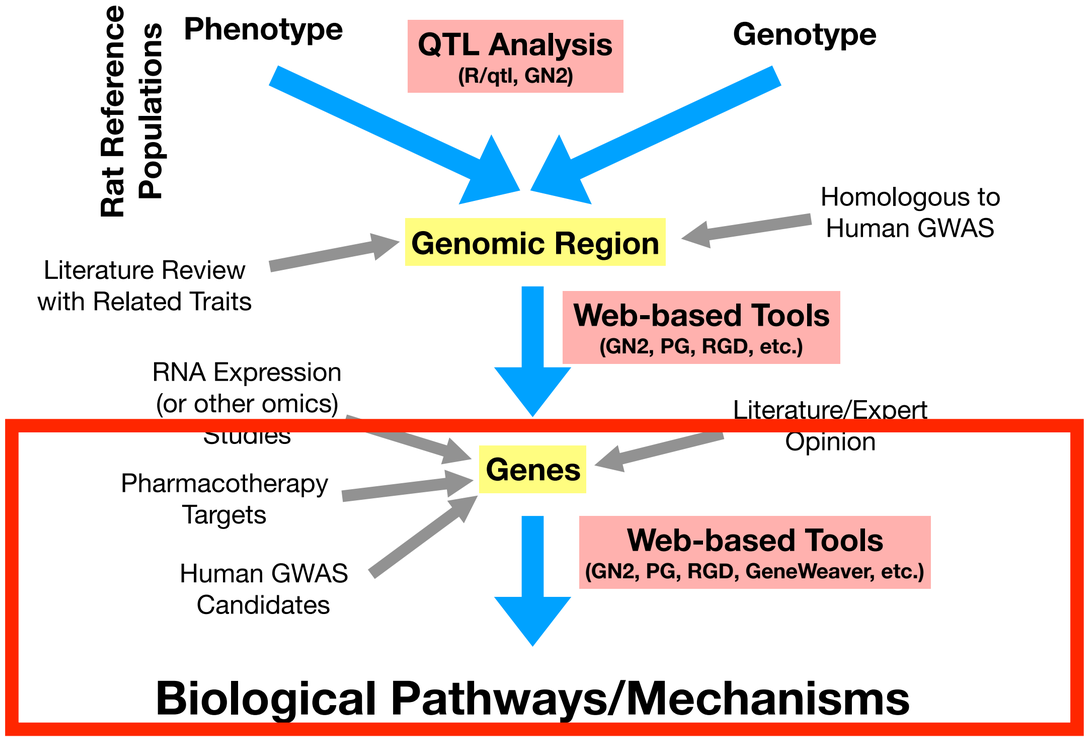
Forward Genetics in Model Organisms



Outline

- Why networks?
- Basics of WGCNA
- WGCNA algorithmic details

Forward Genetics in Model Organisms



Why Networks?

Why networks?

1. No gene product acts independently in the cell.
2. Most tissues, especially brain, are **complex hierarchical networks** that are spatio-temporally linked through structure and functions.
3. Many diseases, including substance use disorders, are often conceptualized as **failure in network regulation**.
4. The generation of a network **provides insight for understanding** predisposition to disease, etiology of organ or behavioral pathology, and response to medications or toxins.

Practical issue with independent gene analyses

- Out of your 100s of ‘candidate’ genes identified from independent tests of each gene, which gene(s) should be pursued further?
  - Do you choose the most significant result?

    ➡Statistical significance ≠ Biological relevance
  - Do you choose the gene with the most literature and annotation? Do you simply pick your favorite?

    ➡Biases towards well-annotated genes
- Once you have a candidate gene what kind of functional validation do you pursue?
  - Some transcripts (protein-coding or non-coding) have little to no functional annotation
  - Some transcripts play many roles depending on the cellular environment

What do we gain by building networks and identifying co-expression modules instead of considering each candidate gene individually?

- Inferred biological function of a gene from other co-expressed genes
  - Helpful for under-annotated or unannotated genes
- Context in which the gene exerts its effect
  - Helpful for genes will multiple functions within the cell
- Identification of multiple therapeutic targets within the same pathway
  - Helpful for finding druggable targets
- Implicate cell types or regions of interest in heterogeneous tissues

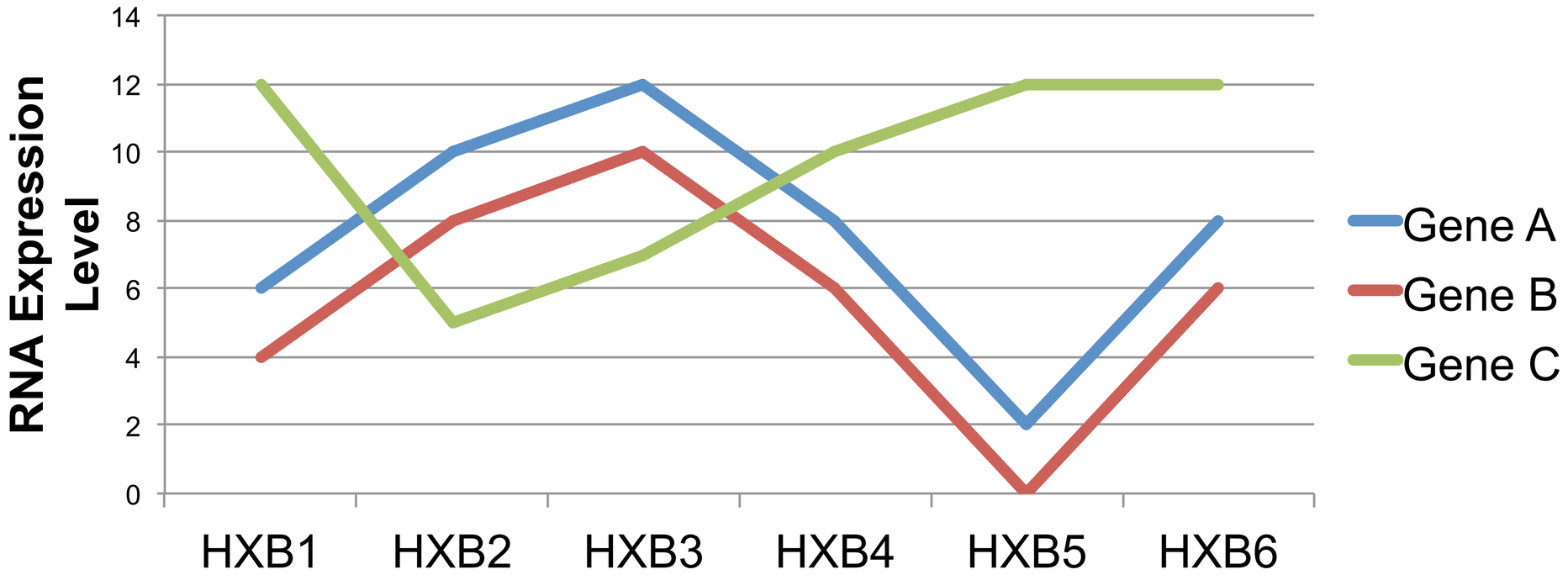
Methods for defining networks of genes

- Protein-protein interactions
  - e.g., STRING database (https://string-db.org/), bioGRID (https://thebiogrid.org/)
- Annotated pathways/gene ontology terms
  - e.g., KEGG Pathways (https://www.genome.jp/kegg/), PANTHER Pathways (http://www.pantherdb.org/), Gene Ontology (http://geneontology.org/)
- RNA co-expression
  - e.g., Weighted Gene Co-Expression Network Analysis, k-means clustering, Bayesian Networks, Gaussian graphical models

Basics of WGCNA

Co-expression as a measure of “connection”

Theory - if the magnitude of RNA expression of two transcripts correlates over multiple “environments” (genomes), then the two transcripts are involved in similar biological processes.



Weighted gene co-expression network analysis (WGCNA)

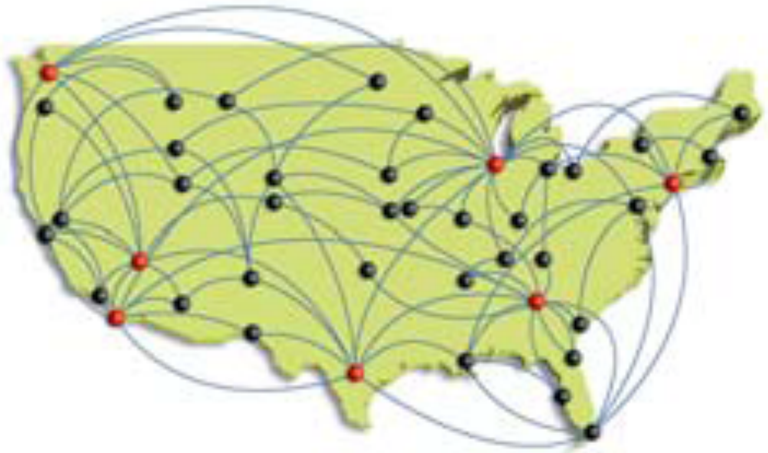
Zhang B, Horvath S (2005) A general framework for weighted gene co-expression network analysis. Stat Appl Genet Mol Biol 4

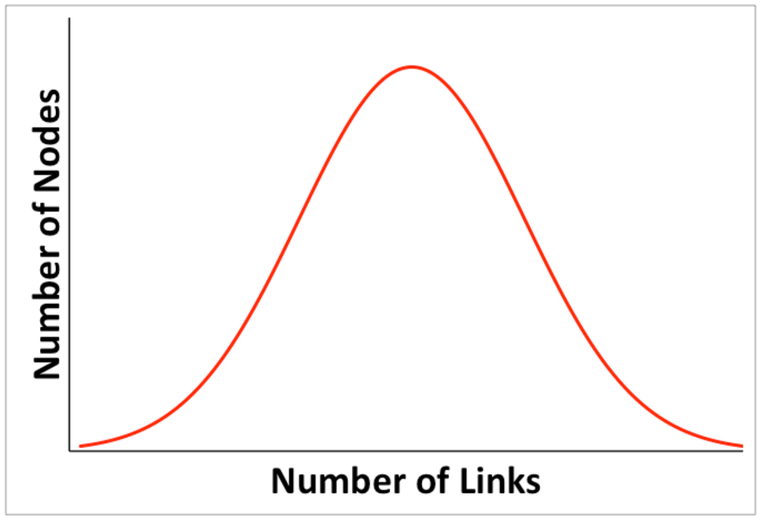
Why not just correlation?

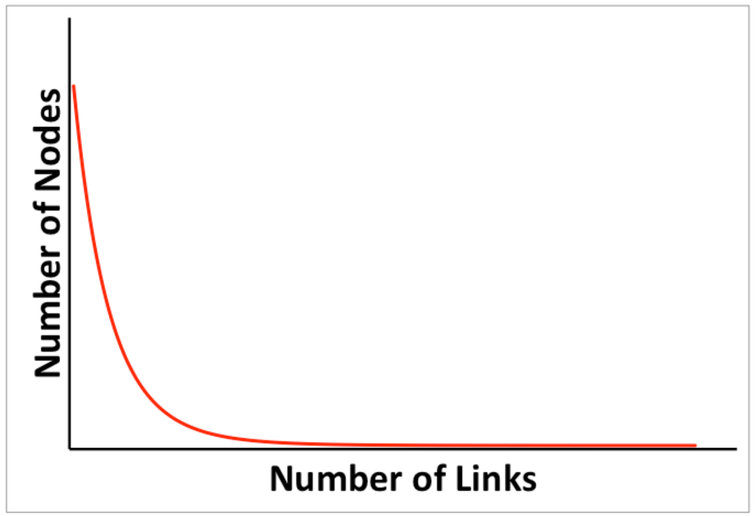
1. Relationships between genes better described using a **scale-free network**

2. We can get a more **robust** measure of co-expression by including a measure of how many “friends” two genes have in common (indirect relationships)

What is a scale-free network?





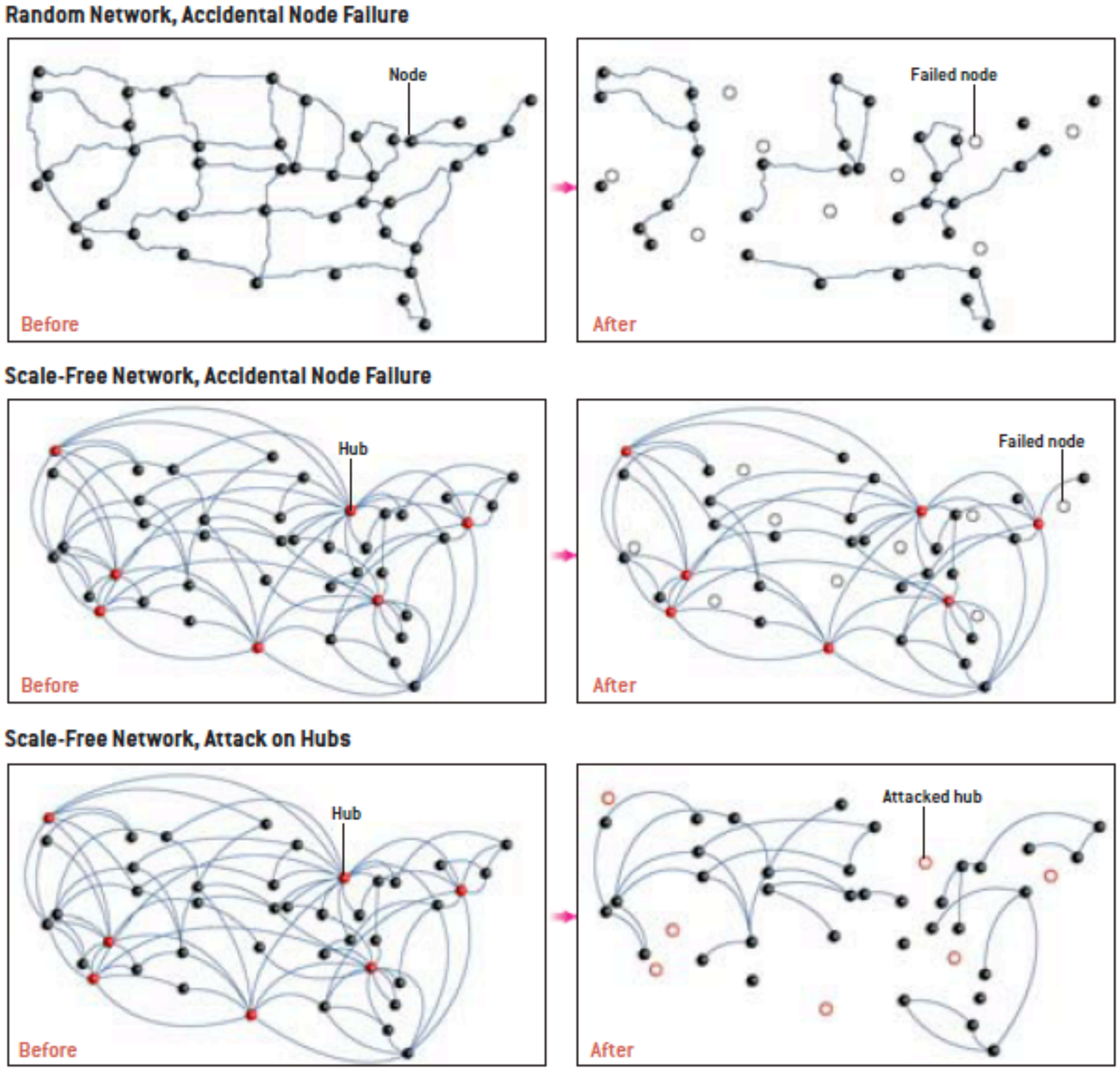


Images taken from *Scale-Free Networks*, Scientific American, May 2003

Advantages of a scale free network

- **Robustness**
  - Random failure of a node does not bring the whole system down
- **Efficiency and Adaptability**
  - Small-world property - The ability to access any individual node from a multitude of alternative pathways makes scale-free networks inherently adaptable to changing environmental conditions.

Advantages of a scale free network - ROBUSTNESS



Advantages of scale free networks - EFFICIENCY

Small world properties - 6 degrees to Kevin Bacon

Advantages of scale free networks - EFFICIENCY

Small world properties - 6 degrees to Kevin Bacon

Get Smart (2008)

Wild Things
(1998)

Advantages of scale free networks - EFFICIENCY

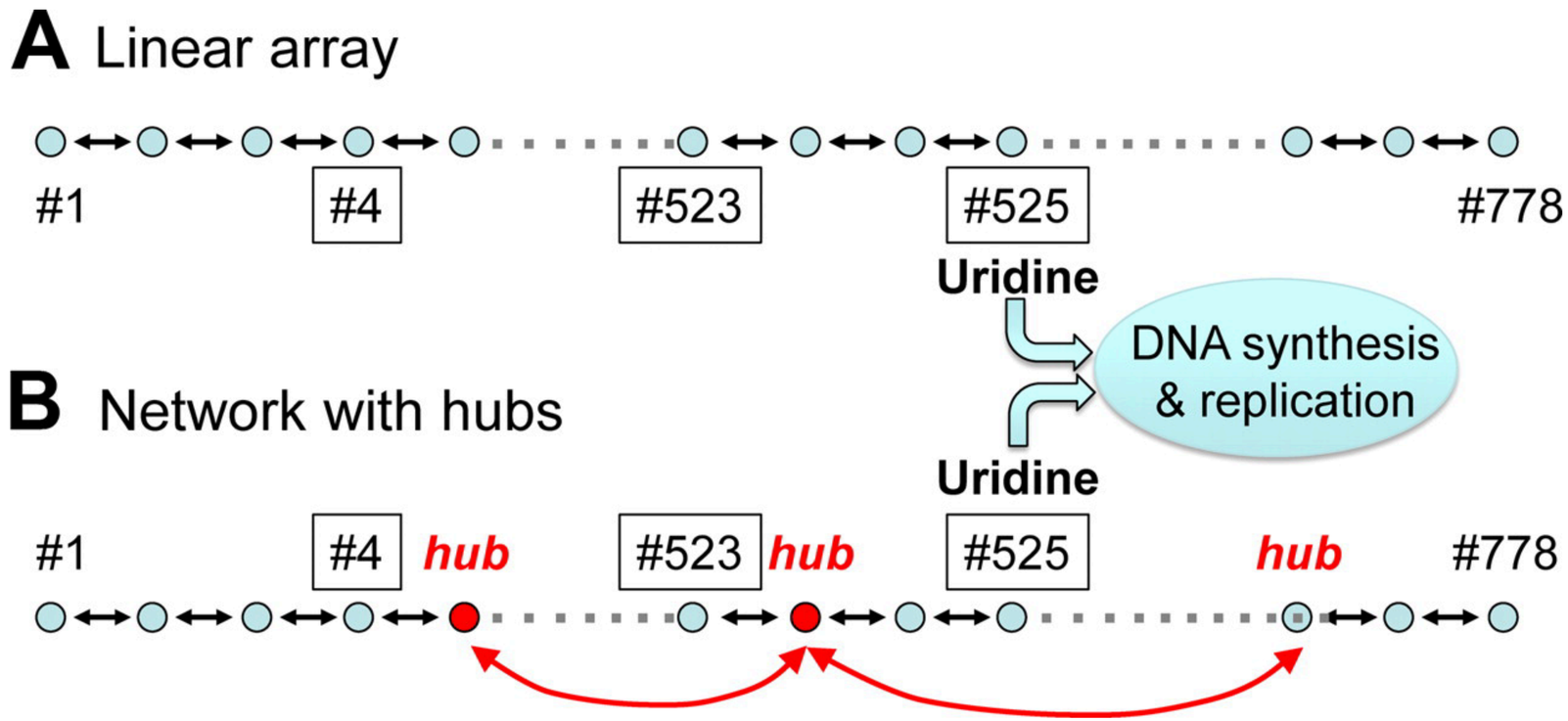
Small world properties - 6 degrees to Kevin Bacon

Advantages of scale free networks - EFFICIENCY

Small world properties - 6 degrees to Kevin Bacon

Advantages of scale free networks - EFFICIENCY

Small world property in metabolic networks



Small-world properties in metabolic networks. A, Hypothetical arrangement of E coli's 778 metabolites in a linear chain. If No. 525 is uridine (required for DNA synthesis and replication), its synthesis from nearby precursor metabolites such as No. 523 is highly efficient, requiring only a few steps. However, if only distant precursors such as No. 4 are available, hundreds of steps are now required, making uridine synthesis slow, inefficient, and energetically costly. B, In contrast, highly connected hub nodes (red) in a scale-free network allow efficient conversion of either No. 523 or No. 4 to uridine in only a few steps.

Weiss et al. Circulation Research. 2012;111:493-504

Selective pressure and scale-free networks

- **Random Network**
  - Majority of nodes have the average number of connections
  - Few nodes have many more or many less than the average
- **Growth of the Network**
  - Random addition of new nodes and new links
  - Old nodes have more links than new nodes (due to time in network)
  - “The strong get stronger”, i.e., preferential attachment

Scale-Free Network

Number of links per node follows a power law distribution

Weighted gene co-expression network analysis (WGCNA)

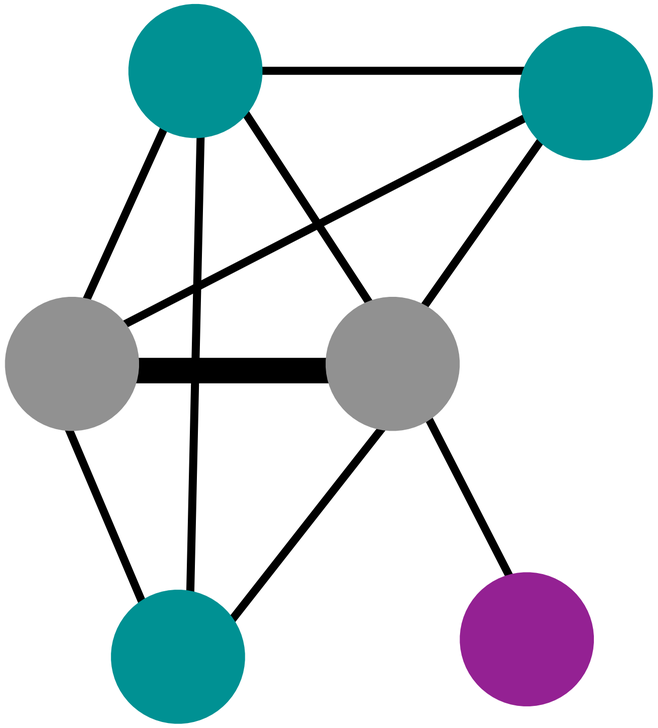
Zhang B, Horvath S (2005) A general framework for weighted gene co-expression network analysis. Stat Appl Genet Mol Biol 4

Why not just correlation?

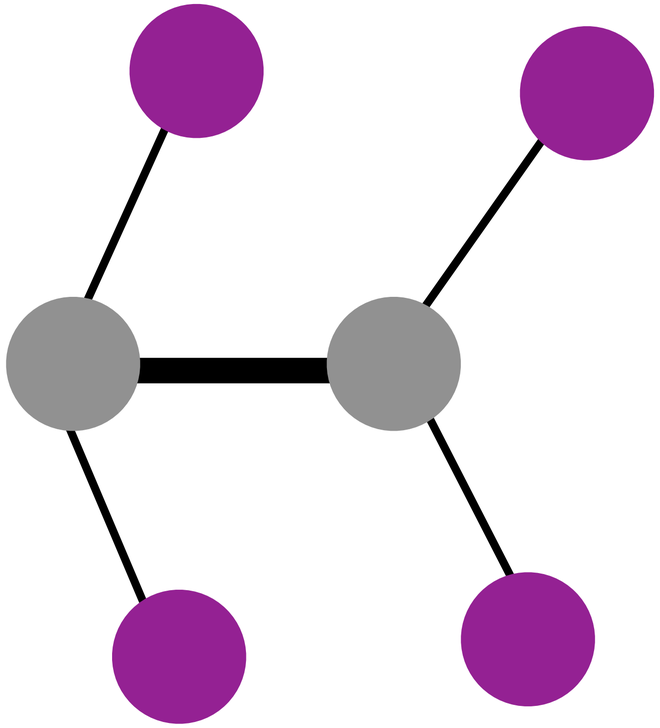
1. Relationships between genes better described using a **scale-free network**

2. We can get a more **robust** measure of co-expression by including a measure of how many “friends” two genes have in common (indirect relationships)

Integration of indirect and direct correlations



High Indirect Correlations

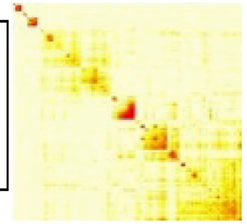


Low Indirect Correlations

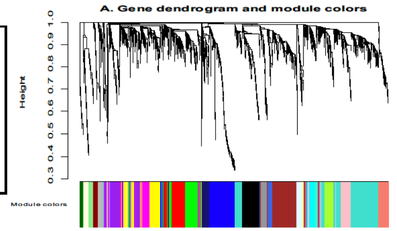
WGCNA Algorithmic Details

Overview of WGCNA workflow

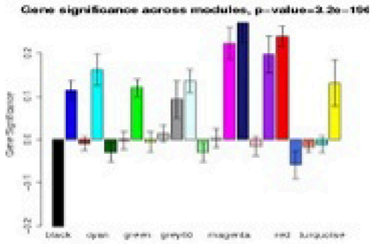
Construct a gene co-expression network
Rationale: make use of interaction patterns among genes
Tools: correlation as a measure of co-expression



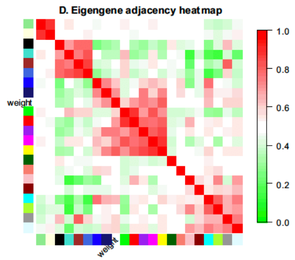
Identify modules
Rationale: module (pathway) based analysis
Tools: hierarchical clustering, Dynamic Tree Cut



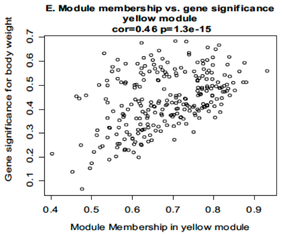
Relate modules to external information
Array Information: clinical data, SNPs, proteomics
Gene Information: ontology, functional enrichement
Rationale: find biologically interesting modules



Study module relationships
Rationale: biological data reduction, systems-level view
Tools: Eigengene Networks



Find the key drivers in *interesting* modules
Rationale: experimental validation, biomarkers
Tools: intramodular connectivity, causality testing



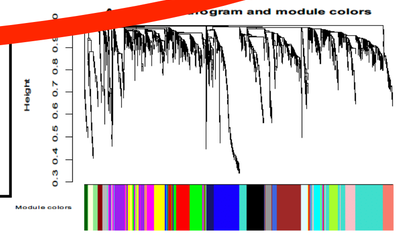
Langfelder and Horvath BMC Bioinformatics 9(1):559 (2008)

Construct a co-expression network

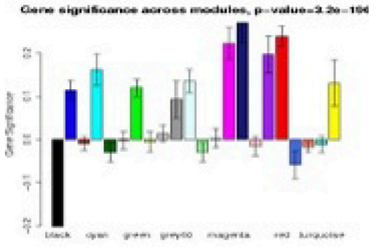
1. Calculate a **similarity matrix** among genes
2. Transform the similarity into an **adjacency matrix**
3. Create **topological overlap matrix** (TOM) from adjacency matrix
4. Transform the TOM to represent **dissimilarity** among genes

Construct a gene co-expression network
Rationale: make use of interaction patterns among genes
Tools: correlation as a measure of co-expression

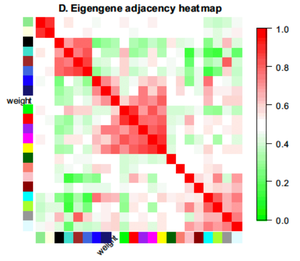
Identify modules
Rationale: module (pathway) based analysis
Tools: hierarchical clustering, Dynamic Tree Cut



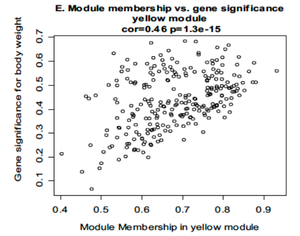
Relate modules to external information
Array Information: clinical data, SNPs, proteomics
Gene Information: ontology, functional enrichement
Rationale: find biologically interesting modules



Study module relationships
Rationale: biological data reduction, systems-level view
Tools: Eigengene Networks

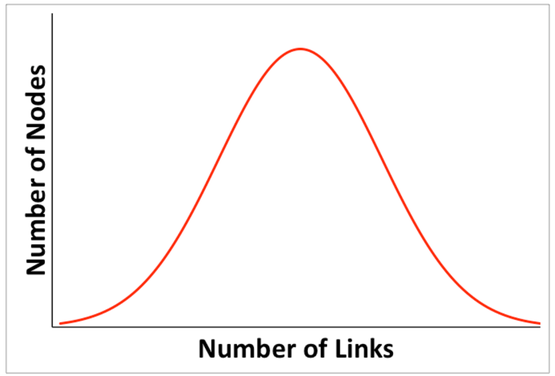
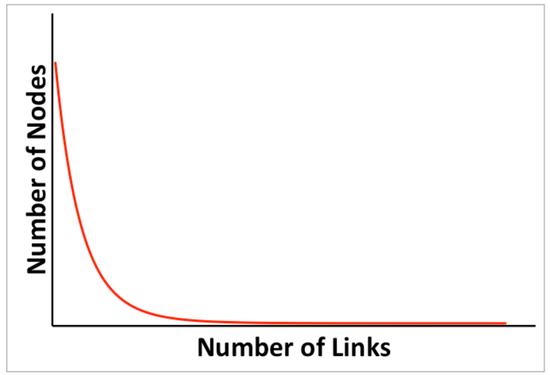
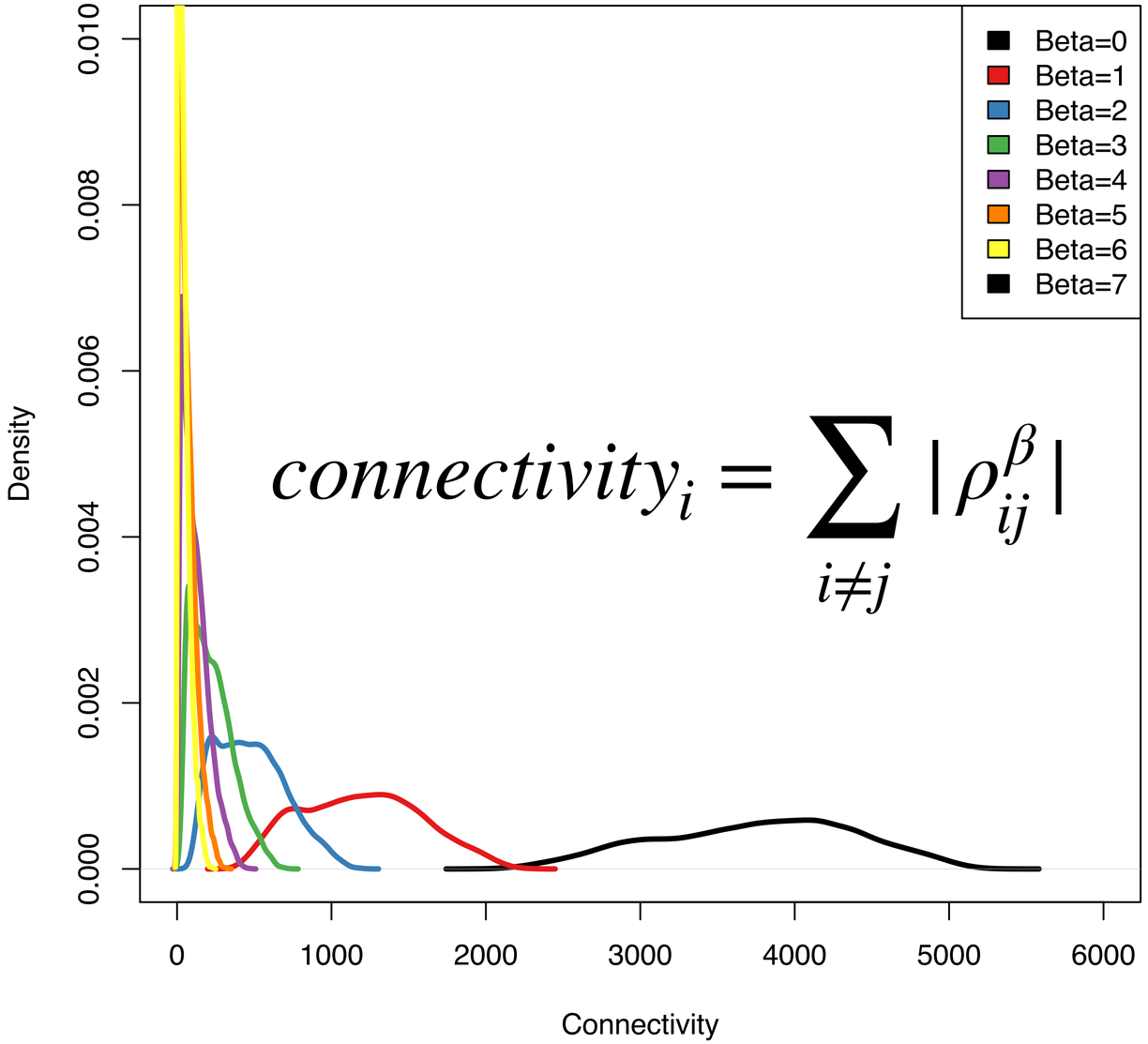
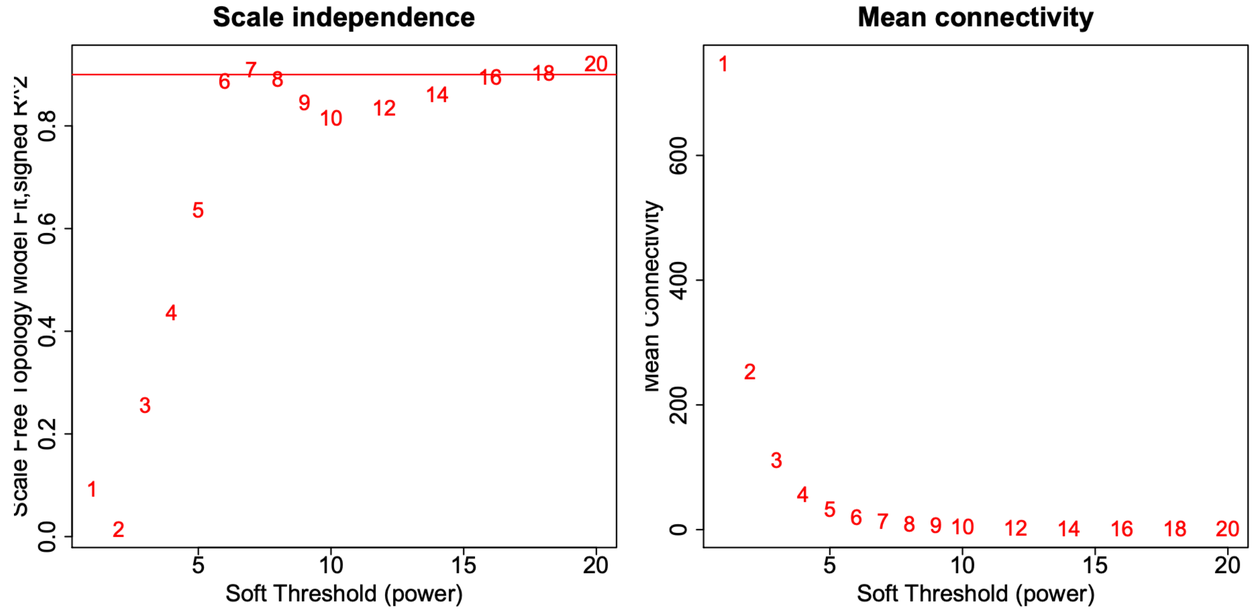


Find the key drivers in *interesting* modules
Rationale: experimental validation, biomarkers
Tools: intramodular connectivity, causality testing



1. Calculate similarity matrix among genes

- Similarity is typically measured using correlation coefficients
- Needs to range from -1 to 1
- Resulting matrix is symmetric with the number of rows and the number of columns equal to the number of genes included in the network.
- R example:

```
```{r similarity}
similarity = cor(datExpr, method="spearman")
```
```

2. Transform the similarity into an adjacency matrix

Goal of this step: To mathematically transform the similarity (e.g., correlation) matrix to represent a scale free network.

Step 1: Choose whether you would like your network to be unsigned, signed, or signed hybrid.

- Unsigned: genes are similar if the magnitude of their association is high, regardless of the direction of that association
- Signed: genes are only similar if they are positively correlates; genes that have a high negative correlation are dissimilar
- Signed hybrid: negative associations are treated as zeros in the similarity matrix

Step 2: Evaluate several soft-threshold betas to determine the most appropriate beta to transform the similarity matrix into a scale free network.

Mimicking a scale free network

Random Network



Scale-free Network





Evaluating soft threshold values

- **Soft threshold** - the choice for the most appropriate value to achieve an approximately scale-free network is data dependent
- Typical values for a signed network are 11 to 16
- Typical values for an unsigned network are 6 to 8



Calculating adjacency

- Once a soft threshold is determined, the similarity matrix can be transformed into an adjacency matrix by simply taking the entire matrix to the power of the soft threshold.
- R example:

```
```{r}
## calculate adjacency from similarity matrix
adjac <- adjacency.fromSimilarity(similarity, type = "unsigned", power = 8)
```
```

3. Create topological overlap matrix (TOM) from adjacency matrix

Goal of this step: To incorporate both direct and indirect relationships into the TOM.

$$\omega_{ij} = \frac{l_{ij} + a_{ij}}{\min\{k_i, k_j\} + 1 - a_{ij}}$$

a_{ij}

adjacency value for gene i and gene j

$l_{ij} = \sum_u a_{iu} a_{uj}$

measure of the indirect relationship between gene i and gene j

$k_i = \sum_u a_{iu}$

connectivity of gene i

Rcode for calculating TOM

```
```{r}
# Turn adjacency into topological overlap
TOM = TOMsimilarity(adjac)
```
```

4. Transform the TOM to represent dissimilarity among genes

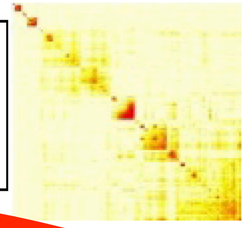
Goal of this step: The TOM is a measure of similarity (i.e., larger values indicate more similarity between two genes), but for hierarchical clustering in the next step we need a measure of dissimilarity (i.e., smaller values indicate more similarity between two genes).

$$dissTOM = 1 - TOM$$

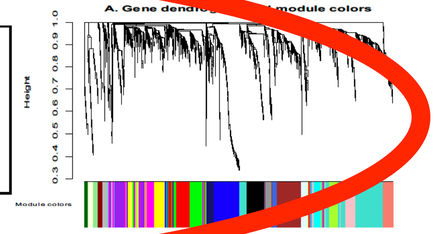
```{r}
## convert similarity TOM into dissimilarity TOM
dissTOM = 1-TOM
```

Identify modules

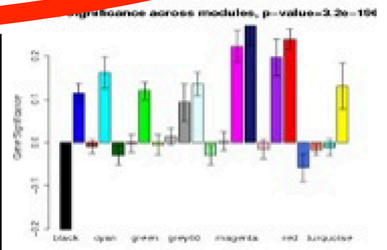
Construct a gene co-expression network
Rationale: make use of interaction patterns among genes
Tools: correlation as a measure of co-expression



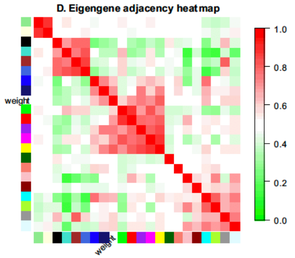
Identify modules
Rationale: module (pathway) based analysis
Tools: hierarchical clustering, Dynamic Tree Cut



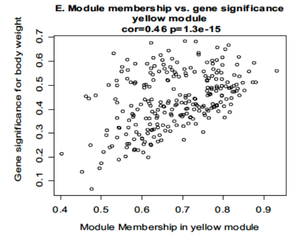
Relate modules to external information
Array Information: clinical data, SNPs, proteomics
Gene Information: ontology, functional enrichement
Rationale: find biologically interesting modules



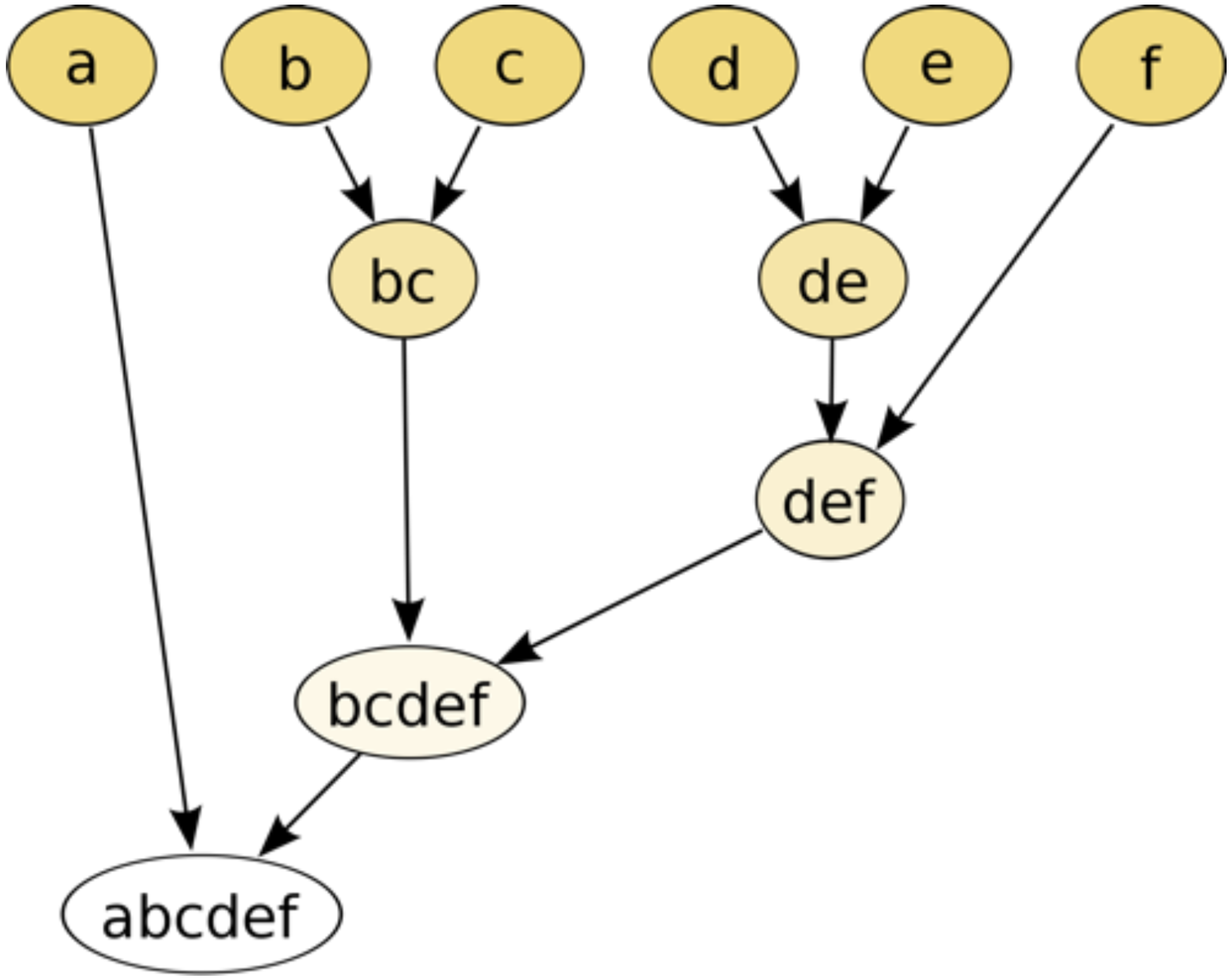
Study module relationships
Rationale: biological data reduction, systems-level view
Tools: Eigengene Networks



Find the key drivers in *interesting* modules
Rationale: experimental validation, biomarkers
Tools: intramodular connectivity, causality testing



Hierarchical Clustering



1. Calculate the all pairwise distances among the nodes (i.e., genes in the case of WGCNA)

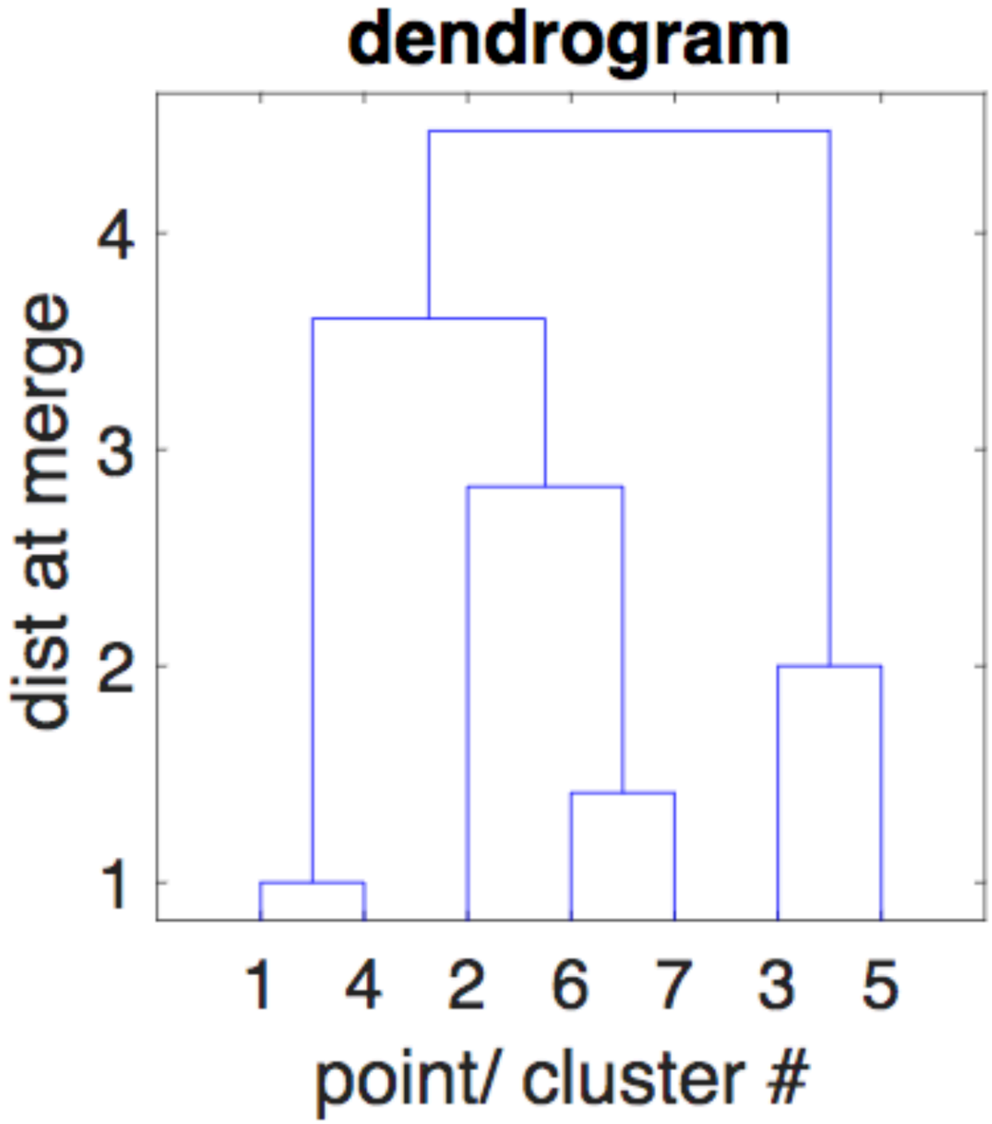
2. Combine the two nodes that have the smallest distance between them.

3. Evaluate the distance between all nodes including the newly combined node.

4. Combine the two nodes or combined nodes that have the smallest distance between them

5. Repeat these steps until all notes have been combined into a single node.

Dendrograms and identifying clusters



- Often times to identify a cluster, we ‘cut’ the dendrogram at a particular height and any nodes connected below that cut form a module/ cluster.

Dynamic Tree Cut

Benefits of using iterative dynamic tree cut:

1. they are capable of identifying nested clusters
2. they are flexible—cluster shape parameters can be tuned to suit the application at hand
3. they are suitable for automation
4. they can optionally combine the advantages of hierarchical clustering and partitioning around medoids, giving better detection of outliers.

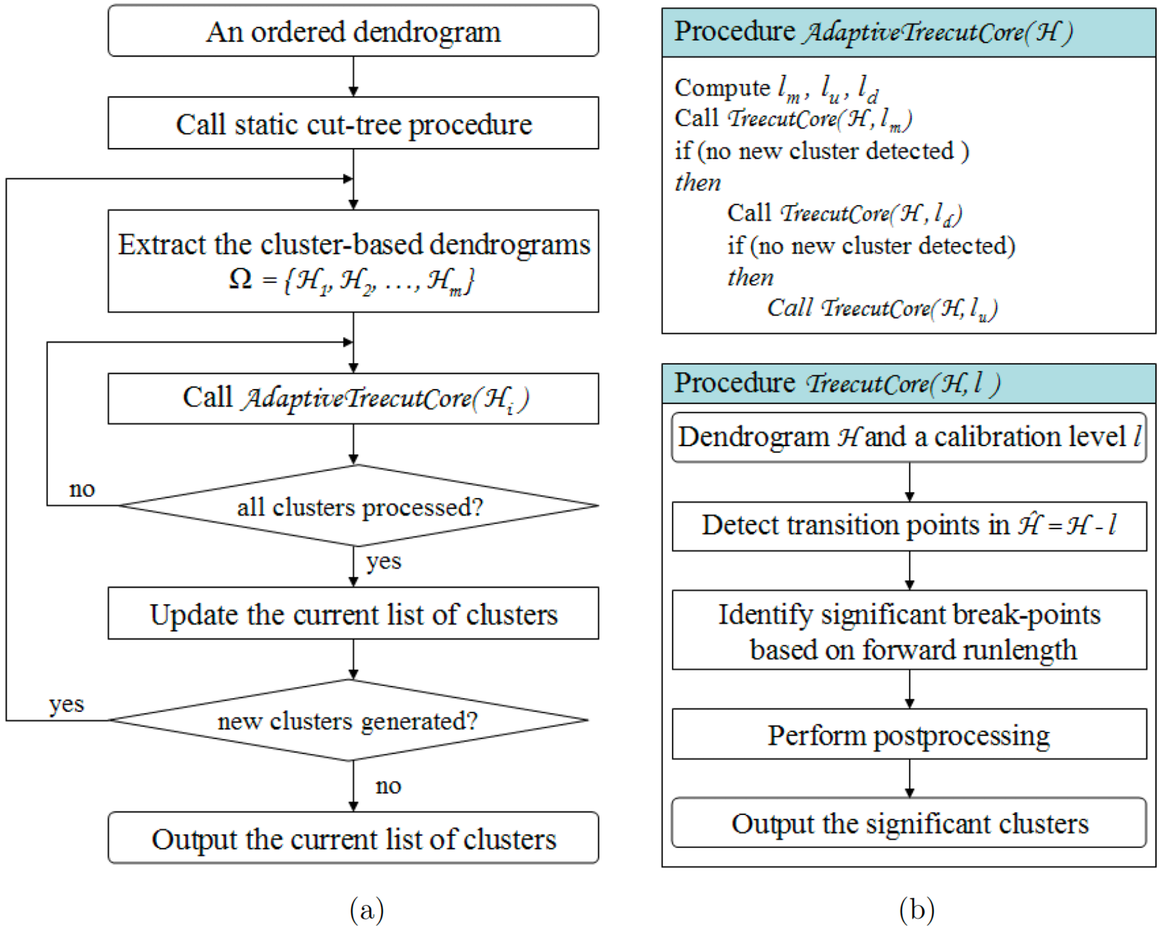
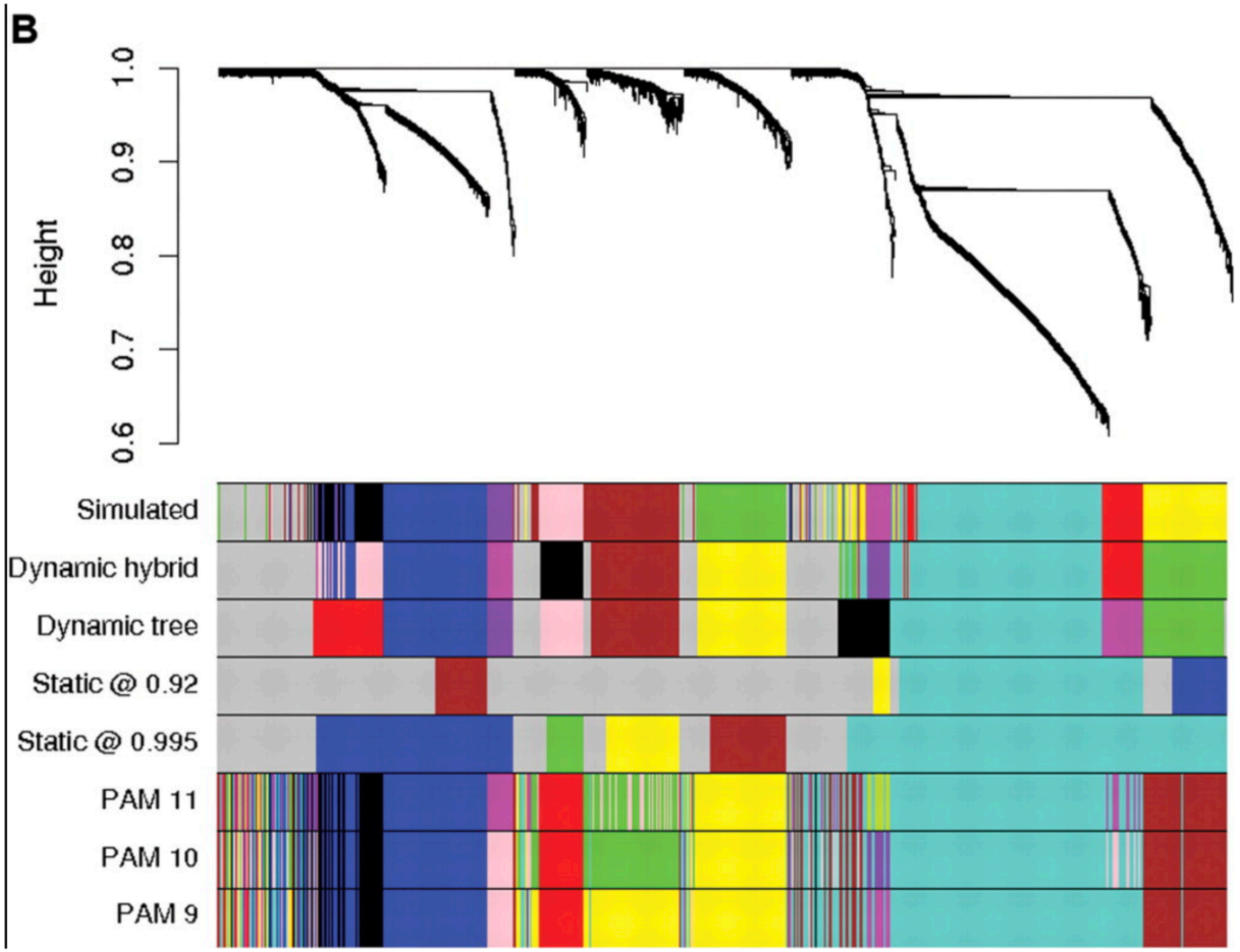


Figure 2: The dynamic cut-tree algorithm: (a) the overall flowchart; (b) the flowcharts of the procedures *AdaptiveTreecutCore* (the top block) and *TreecutCore* (the bottom block).

Langfelder P, Zhang B, Horvath S (2007) Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. Bioinformatics 2008 **24**(5):719-720

Performance of Dynamic Tree Cut



Langfelder P, Zhang B, Horvath S (2007) Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. Bioinformatics 2008 **24**(5):719-720

R code for dynamic tree cut

```r
# Module identification using dynamic tree cut
geneTree = hclust(as.dist(dissTOM), method = "average")
dynamicMods = cutreeDynamic(dendro = geneTree,
                            distM = dissTOM,
                            deepSplit = 2,
                            pamRespectsDendro = FALSE,
                            minClusterSize = 5)
```

R wrapper for constructing a gene co-expression network and identifying modules

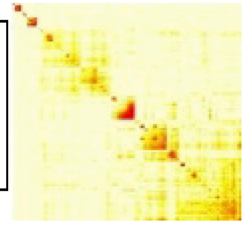
Construct network and identify modules in one function

```{r}
net = blockwiseModules(datExpr,
                       # similarity matrix options
                       corType = "pearson",

                       # adjacency matrix options
                       power = 8,
                       networkType = "unsigned",

                       # TOM options
                       TOMType = "unsigned",

                       # Module identification options
                       minModuleSize = 5,
                       deepSplit = 2,
                       pamRespectsDendro = FALSE)
```

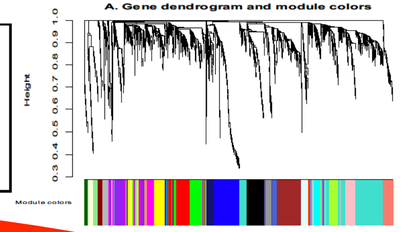
Relate modules to external information

1. Association analyses through module eigengenes
2. Aggregation of gene-level associations across genes
3. Enrichment of functional categories, pathways, specific tissues/region, and specific cell types.

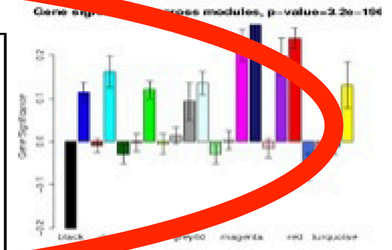
Construct a gene co-expression network
Rationale: make use of interaction patterns among genes
Tools: correlation as a measure of co-expression



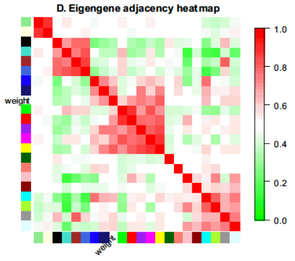
Identify modules
Rationale: module (pathway) based analysis
Tools: hierarchical clustering, Dynamic Tree Cut



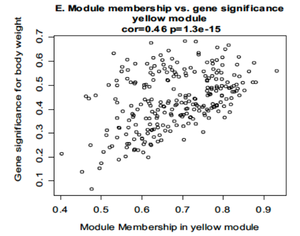
Relate modules to external information
Array Information: clinical data, SNPs, proteomics
Gene Information: ontology, functional enrichement
Rationale: find biologically interesting modules



Study module relationships
Rationale: biological data reduction, systems-level view
Tools: Eigengene Networks

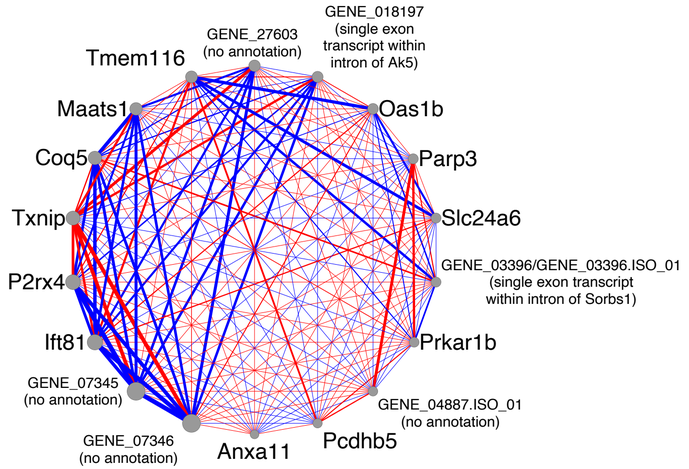


Find the key drivers in *interesting* modules
Rationale: experimental validation, biomarkers
Tools: intramodular connectivity, causality testing



Relate models to external information

DNA Variants



Tissue and cell type information

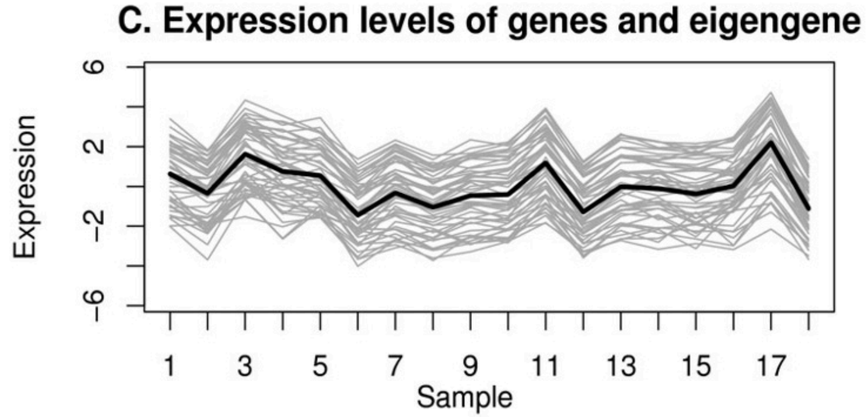
Other Omics Data

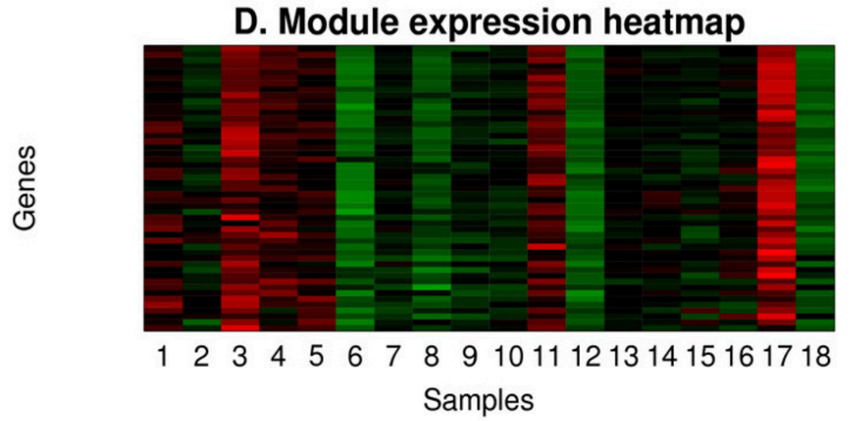
Behavioral and Physiological Phenotypes

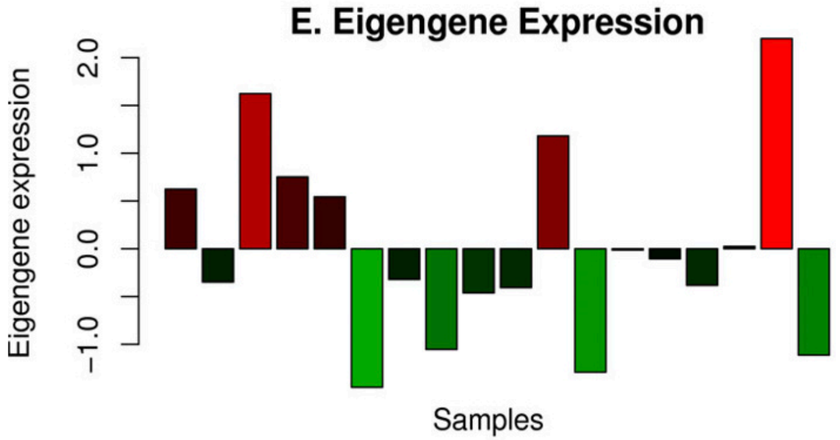
Functional Information

1. Association analyses through module eigengenes

- Module eigengene is used to summarize the expression patterns of the module genes across samples
- First principal component from PCA
- Often the proportion of total variation explained by the module eigengene is used as a measure of robustness
- End results is one quantitative value per sample





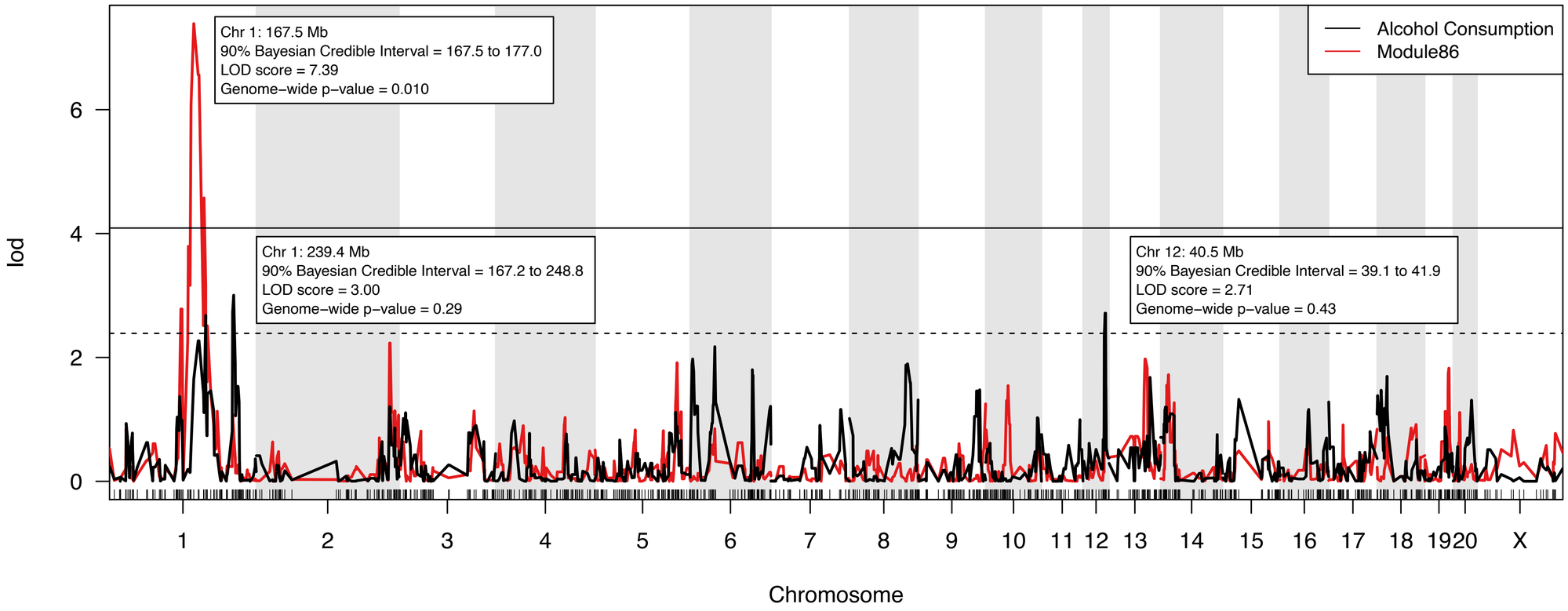


Langfelder, P., Horvath, S. Eigengene networks for studying the relationships between co-expression modules. *BMC Syst Biol* **1,** 54 (2007).

DNA Variants

Module eigengene QTL

Just like a behavioral or physiological phenotype, a module eigengene can be mapped to the genome using a QTL analysis.



Module-trait relationships

Behavioral and Physiological Phenotypes

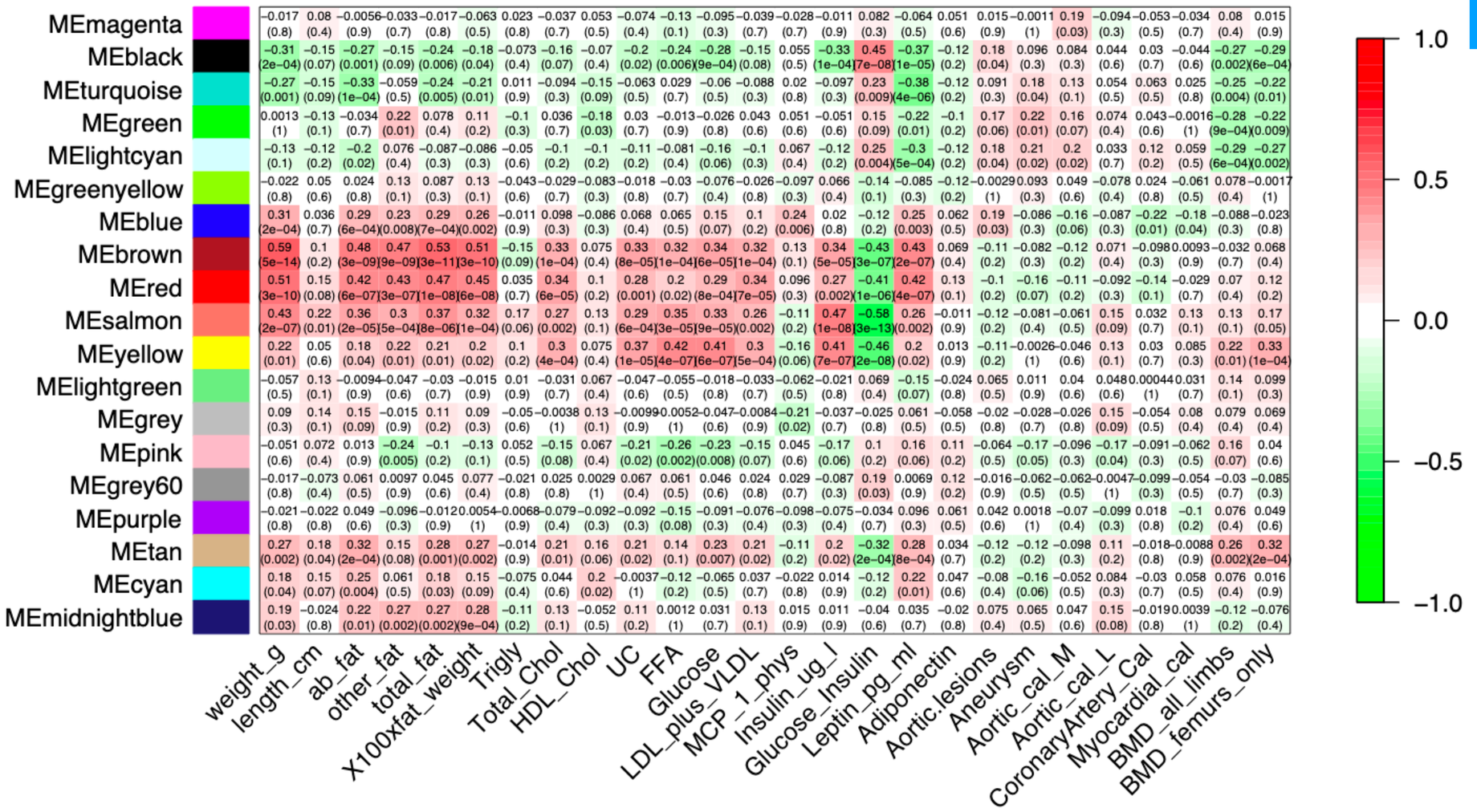


Figure 1: Module-trait associations. Each row corresponds to a module eigengene, column to a trait. Each cell contains the corresponding correlation and p-value. The table is color-coded by correlation according to the color legend.

2. Aggregation of gene-level associations across genes

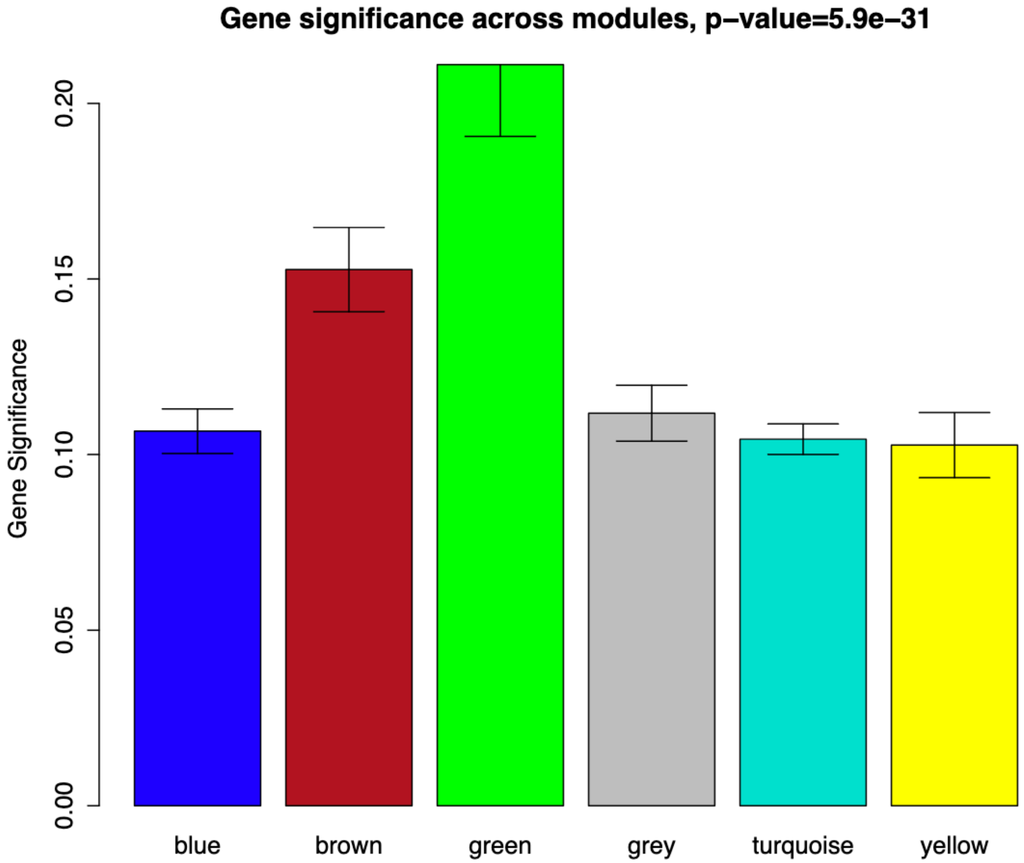


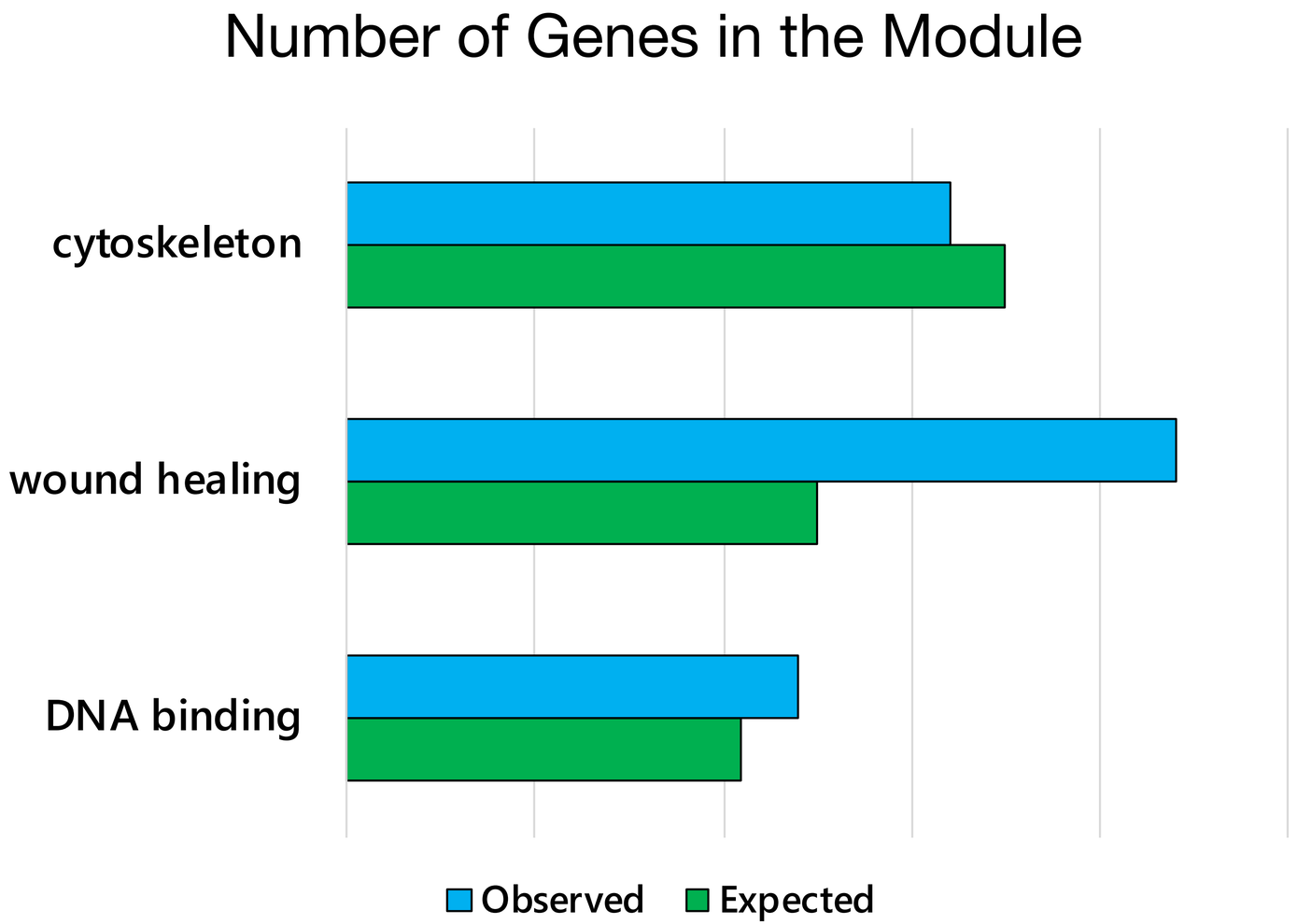
Figure 5: Barplot of module significance defined as the mean gene significance across all genes in the module. The green and brown modules are the most promising.

Example approach from WGCNA tutorial:

- Average association of the phenotype with each gene in the module.

3. Enrichment of functional categories, pathways, specific tissues/region, and specific cell types

Goal of Enrichment Analysis: To determine if the number of genes within the module that are associated with a specific pathway, ontology, cell type etc. is more than you would expect by chance.



Slide courtesy of Cheyret Wood

Functional Information

Tools for functional enrichment

- **Enrichr (https://maayanlab.cloud/Enrichr/)** - web-based enrichment tool that examines over 160 libraries for enrichment; easy to use but does not allow for the definition of a background data set
- **PANTHER (http://www.pantherdb.org/)** - web-based enrichment tool that examines Gene Ontology and the PANTHER Pathways; allows the user to set the background data set; only uses two libraries
- **API tools from Ensembl and KEGG** - these APIs make it easy to retrieve the latest data from the source and simple enrichment statistics can easily be executed in R

Enrichment of tissues, brain regions, and cell types

Tissue and cell type information

ARTICLE

doi:10.1038/nature11405

An anatomically comprehensive atlas of the adult human brain transcriptome

Michael J. Hawrylycz[1]*, Ed S. Lein[1]*, Angela L. Guillozet-Bongaarts[1], Elaine H. Shen[1], Lydia Ng[1], Jeremy A. Miller[1], Louie N. van de Lagemaat[2], Kimberly A. Smith[1], Amanda Ebbert[1], Zackery L. Riley[1], Chris Abajian[1], Christian F. Beckmann[3], Amy Bernard[1], Darren Bertagnolli[1], Andrew F. Boe[1], Preston M. Cartagena[4], M. Mallar Chakravarty[1,5], Mike Chapin[1], Jimmy Chong[1], Rachel A. Dalley[1], Barry David Daly[6], Chinh Dang[1], Suvro Datta[1], Nick Dee[1], Tim A. Dolbeare[1], Vance Faber[1], David Feng[1], David R. Fowler[7], Jeff Goldy[1], Benjamin W. Gregor[1], Zeb Haradon[1], David R. Haynor[8], John G. Hohmann[1], Steve Horvath[9], Robert E. Howard[1], Andreas Jeromin[10], Jayson M. Jochim[1], Marty Kinnunen[1], Christopher Lau[1], Evan T. Lazarz[1], Changkyu Lee[1], Tracy A. Lemon[1], Ling Li[11], Yang Li[1], John A. Morris[1], Caroline C. Overly[1], Patrick D. Parker[1], Sheana E. Parry[1], Melissa Reding[1], Joshua J. Royall[1], Jay Schulkin[12], Pedro Adolfo Sequeira[13], Clifford R. Slaughterbeck[1], Simon C. Smith[14], Andy J. Sodt[1], Susan M. Sunkin[1], Beryl E. Swanson[1], Marquis P. Vawter[13], Derric Williams[1], Paul Wohnoutka[1], H. Ronald Zielke[15], Daniel H. Geschwind[16], Patrick R. Hof[17], Stephen M. Smith[18], Christof Koch[1,19], Seth G. N. Grant[2] & Allan R. Jones[1]

Neuroanatomically precise, genome-wide maps of transcript distributions are critical resources to complement genomic sequence data and to correlate functional and genetic brain architecture. Here we describe the generation and analysis of a transcriptional atlas of the adult human brain, comprising extensive histological analysis and comprehensive microarray profiling of ~900 neuroanatomically precise subdivisions in two individuals. Transcriptional regulation varies enormously by anatomical location, with different regions and their constituent cell types displaying robust molecular signatures that are highly conserved between individuals. Analysis of differential gene expression and gene co-expression relationships demonstrates that brain-wide variation strongly reflects the distributions of major cell classes such as neurons, oligodendrocytes, astrocytes and microglia. Local neighbourhood relationships between fine anatomical subdivisions are associated with discrete neuronal subtypes and genes involved with synaptic transmission. The neocortex displays a relatively homogeneous transcriptional pattern, but with distinct features associated selectively with primary sensorimotor cortices and with enriched frontal lobe expression. Notably, the spatial topography of the neocortex is strongly reflected in its molecular topography—the closer two cortical regions, the more similar their transcriptomes. This freely accessible online data resource forms a high-resolution transcriptional baseline for neurogenetic studies of normal and abnormal human brain function.

The enormous complexity of the human brain is a function of its precise circuitry, its structural and cellular diversity, and, ultimately, the regulation of its underlying transcriptome. In rodents, brain- and transcriptome-wide, cellular-resolution maps of transcript distributions are widely useful resources to complement genomic sequence data[1–3]. However, owing to the challenges of a 1,000-fold increase in size from mouse to human, limitations in post-mortem tissue availability and quality, and the destructive nature of molecular assays, there has been no human counterpart so far. Several important recent studies have begun to analyse transcriptional dynamics during human brain development[4,5], although only in a small number of relatively coarse brain regions. Characterizing the complete transcriptional architecture of the human brain will provide important information for understanding the impact of genetic disorders on different brain regions and functional circuits. Furthermore, conservation and divergence in brain function between humans and other species provide essential information for the understanding of drug action, which is often poorly conserved across species[6].

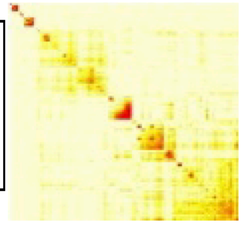
The goal of the Allen Human Brain Atlas is to create a comprehensive map of transcript usage across the entire adult brain, with the emphasis on anatomically complete coverage at a fine nuclear resolution in a small number of high-quality, clinically unremarkable brains profiled with DNA microarrays for quantitative gene-level transcriptome coverage. Furthermore, structural brain imaging data were obtained from each individual to visualize gene expression data in its native three-dimensional anatomical coordinate space, and to allow correlations between imaging and transcriptome modalities. These data are freely accessible via the Allen Brain Atlas data portal (http://www.brain-map.org).

[1]Allen Institute for Brain Science, Seattle, Washington 98103, USA. [2]Genes to Cognition Programme, Edinburgh University, Edinburgh EH16 4SB, UK. [3]MIRA Institute, University of Twente & Donders Institute, Radboud University Nijmegen, Nijmegen, Netherlands. [4]Department of Psychiatry & Human Behavior, University of California, Irvine, California 92697, USA. [5]Kimel Family Translational Imaging-Genetics Laboratory, Centre for Addiction and Mental Health Toronto, Ontario M5S 2S1, Canada. [6]University of Maryland School of Medicine, Department of Diagnostic Radiology, University of Maryland Medical Center, Baltimore, Maryland 21201, USA. [7]Department of Pathology, University of Maryland School of Medicine, Baltimore, Maryland 21201, USA. [8]Department of Radiology, University of Washington, Seattle, Washington 98195, USA. [9]Department of Human Genetics, Gonda Research Center, David Geffen School of Medicine, Los Angeles, California 90095, USA. [10]Banyan Biomarkers, Inc., Alachua, Florida 32615, USA. [11]Office of the Chief Medical Examiner, Baltimore, MD, Department of Pediatrics, University of Maryland, Baltimore, Maryland 21201, USA. [12]Department of Neuroscience, Georgetown University, School of Medicine, Washington DC 20007, USA. [13]Functional Genomics Laboratory, Department of Psychiatry & Human Behavior, School of Medicine, University of California, Irvine, California 92697, USA. [14]Histion LLC, Everett, Washington 98204, USA. [15]The Eunice Kennedy Shriver NICHD Brain and Tissue Bank for Developmental Disorders, University of Maryland, Baltimore, Maryland 21201, USA. [16]Program in Neurogenetics, Department of Neurology and Department of Human Genetics, and Semel Institute, David Geffen School of Medicine-UCLA, Los Angeles, California 90095, USA. [17]Fishberg Department of Neuroscience and Friedman Brain Institute, Mount Sinai School of Medicine, New York, New York 10029, USA. [18]FMRIB, Oxford University, Oxford OX3 9DU, UK. [19]Computation & Neural Systems, California Institute of Technology, Pasadena, California 91125, USA.
*These authors contributed equally to this work.

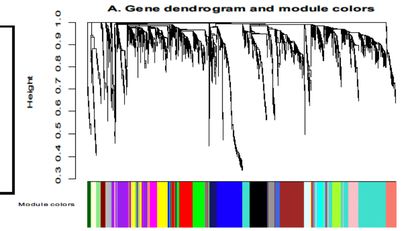
- BrainRegionMarkers, BrainLists, and PWLists are included in the WGCNA package
- With the emergence of single cell RNA-Seq, cell type specific marker are available for many cell types.

Study module relationships

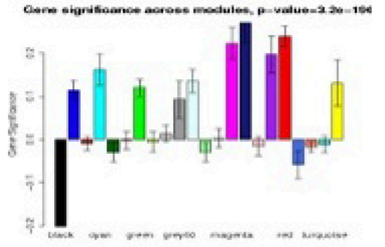
Construct a gene co-expression network
Rationale: make use of interaction patterns among genes
Tools: correlation as a measure of co-expression



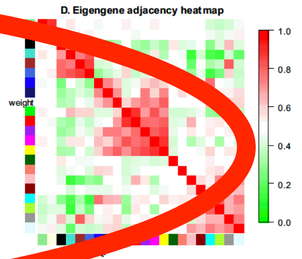
Identify modules
Rationale: module (pathway) based analysis
Tools: hierarchical clustering, Dynamic Tree Cut



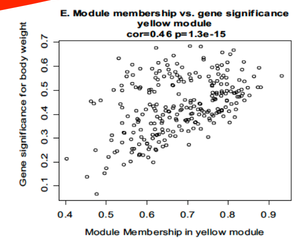
Relate modules to external information
Array Information: clinical data, SNPs, proteomics
Gene Information: ontology, functional enrichement
Rationale: find biologically interesting modules



Study module relationships
Rationale: biological data reduction, systems-level view
Tools: Eigengene Networks



Find the key drivers in *interesting* modules
Rationale: experimental validation, biomarkers
Tools: intramodular connectivity, causality testing



Study module relationships

BMC Systems Biology

BioMed Central

Methodology article

Open Access

Eigengene networks for studying the relationships between co-expression modules

Peter Langfelder[1] and Steve Horvath*[2]

Address: [1]Department of Human Genetics, University of California, Los Angeles, CA 90095, USA and [2]Department of Human Genetics and Department of Biostatistics, University of California, Los Angeles, CA 90095, USA

Email: Peter Langfelder - peter.langfelder@gmail.com; Steve Horvath* - shorvath@mednet.ucla.edu

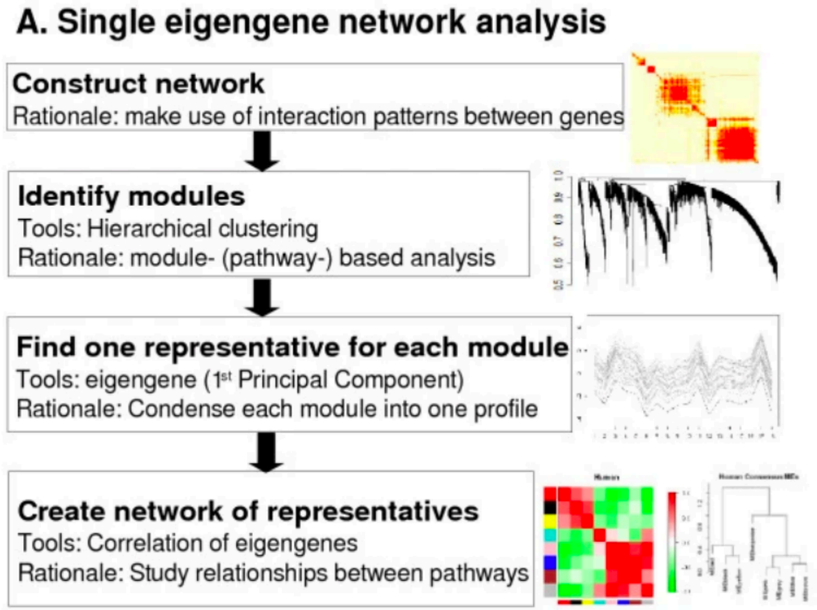
* Corresponding author

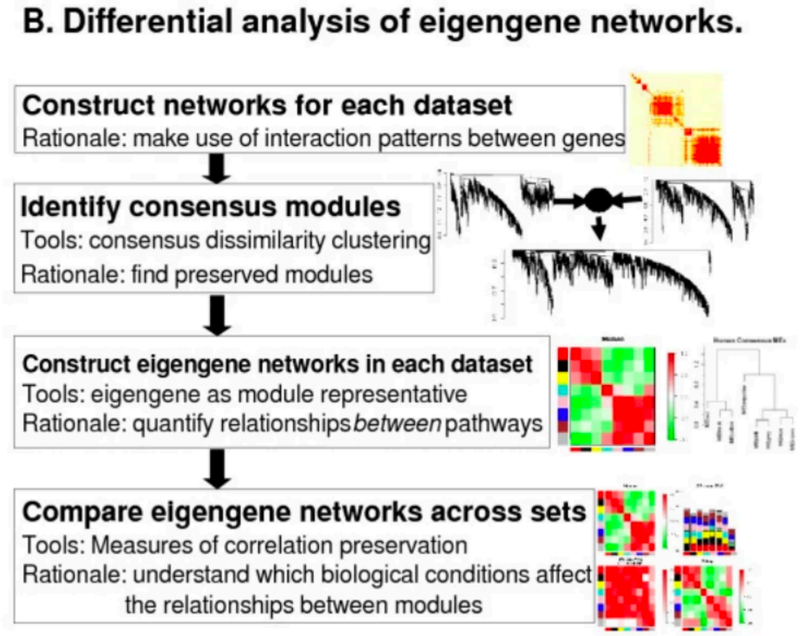
Published: 21 November 2007

BMC Systems Biology 2007, 1:54 doi:10.1186/1752-0509-1-54

Received: 8 May 2007
Accepted: 21 November 2007

This article is available from: http://www.biomedcentral.com/1752-0509/1/54





Example eigengene network

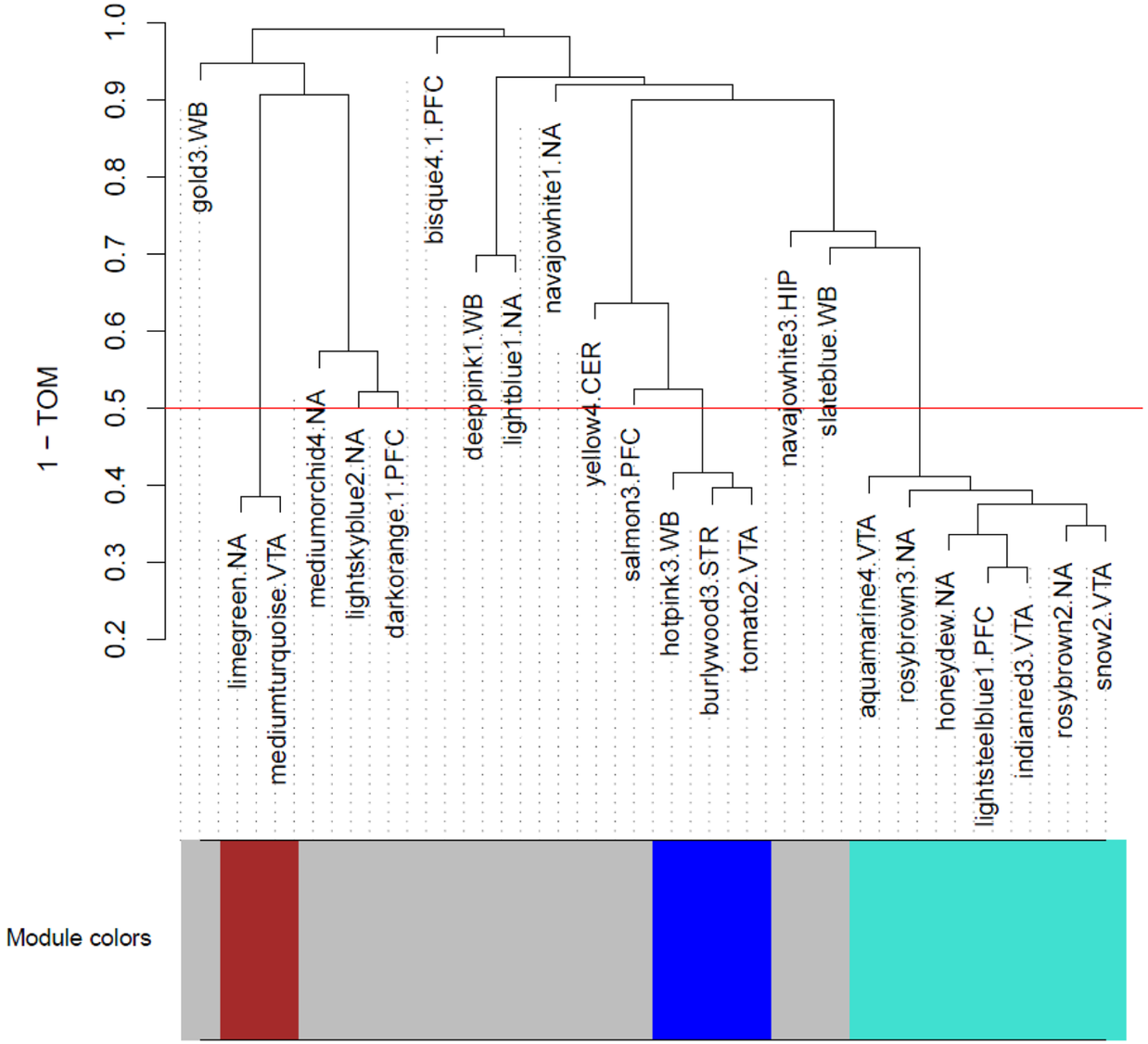
OPEN ACCESS Freely available online

PLOS ONE

Whole Brain and Brain Regional Coexpression Network Interactions Associated with Predisposition to Alcohol Consumption

Lauren A. Vanderlinden[1], Laura M. Saba[1], Katerina Kechris[2], Michael F. Miles[3], Paula L. Hoffman[1], Boris Tabakoff[1]*

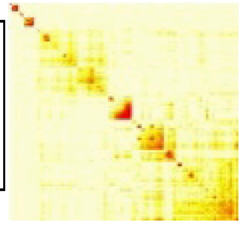
1 Department of Pharmacology, University of Colorado School of Medicine, Aurora, Colorado, United States of America, **2** Department of Biostatistics and Informatics, University of Colorado School of Public Health, Aurora, Colorado, United States of America, **3** Departments of Pharmacology and Neurology and the Center of Study of Biological Complexity, Virginia Commonwealth University, Richmond, Virginia, United States of America



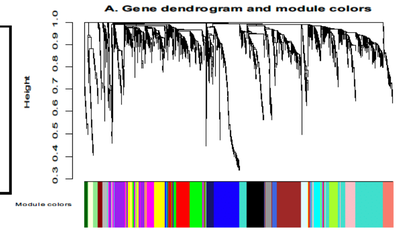
Find the key drivers in *interesting* modules

1. Intramodular connectivity
2. Assessment of module cohesion related to an individual gene
3. Causality testing/Bayesian Networks

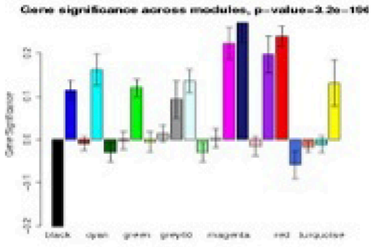
Construct a gene co-expression network
Rationale: make use of interaction patterns among genes
Tools: correlation as a measure of co-expression



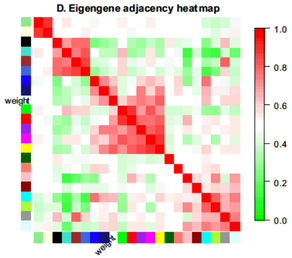
Identify modules
Rationale: module (pathway) based analysis
Tools: hierarchical clustering, Dynamic Tree Cut



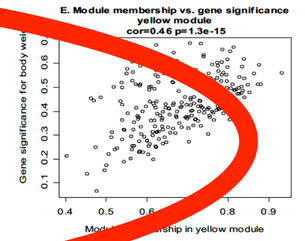
Relate modules to external information
Array Information: clinical data, SNPs, proteomics
Gene Information: ontology, functional enrichement
Rationale: find biologically interesting modules



Study module relationships
Rationale: biological data reduction, systems-level view
Tools: Eigengene Networks



Find the key drivers in *interesting* modules
Rationale: experimental validation, biomarkers
Tools: intramodular connectivity, causality testing



1. Intramodular connectivity

- **Intramodular connectivity** - how connected is a gene to all the other genes within the module
  - Could be calculated as simply the sum of all pairwise correlations of a gene.
- **Hub gene** - the gene with the highest intramodular connectivity
  - some controversy in the field about whether the hub gene is an appropriate gene to target for therapeutics

2. Assessment of module cohesion related to an individual gene

- One important question when examining individual genes in interesting modules is to evaluate the module cohesiveness when the expression of that gene is held constant.

- Theoretically, a connection between the products of two genes that was present in the original module, but was eliminated after adjusting for the effect of a third gene, represents an indirect association between the expression levels of those two gene products that is due to the influence of the third gene on both transcripts.

Genes, Brain and Behavior
Official publication of the International Behavioural and Neural Genetics Society

REVIEW | Open Access

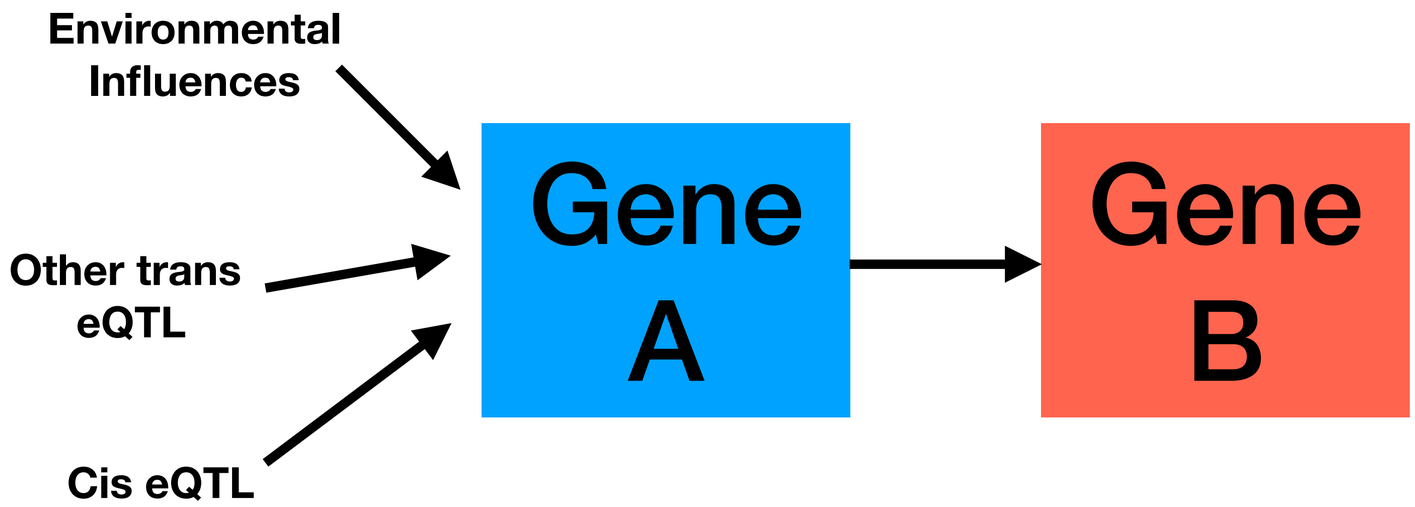
A Long Non-Coding RNA (Lrap) Modulates Brain Gene Expression and Levels of Alcohol Consumption in Rats

Laura M. Saba, Paula L. Hoffman, Gregg E. Homanics, Spencer Mahaffey, Swapna V. Daulatabad, Sarath Chandra Janga, Boris Tabakoff

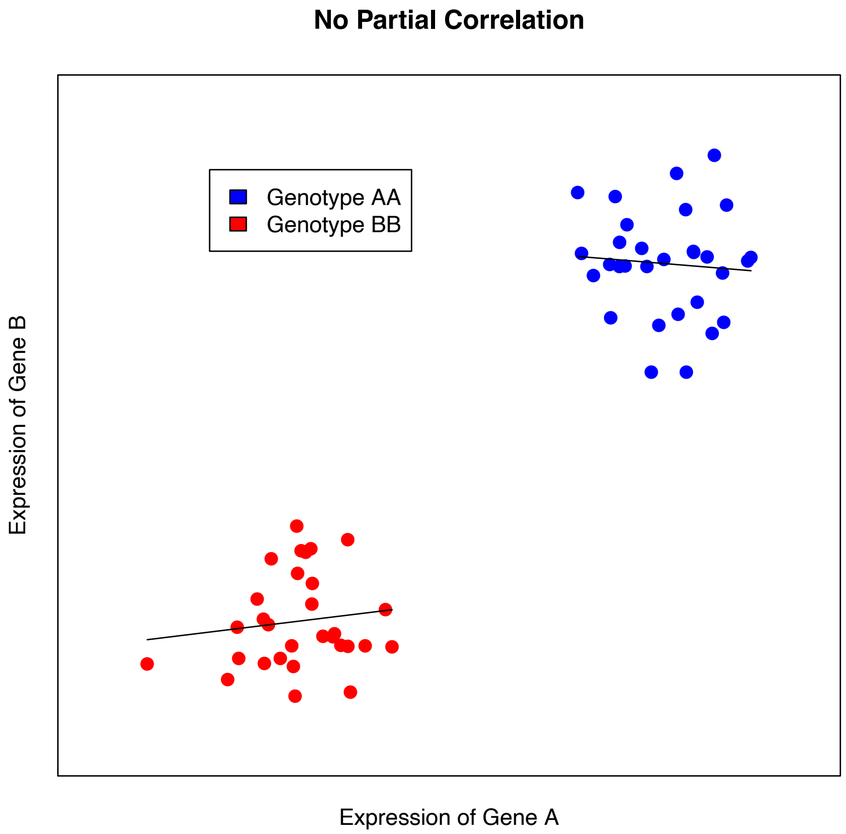
First published: 06 September 2020 | https://doi.org/10.1111/gbb.12698

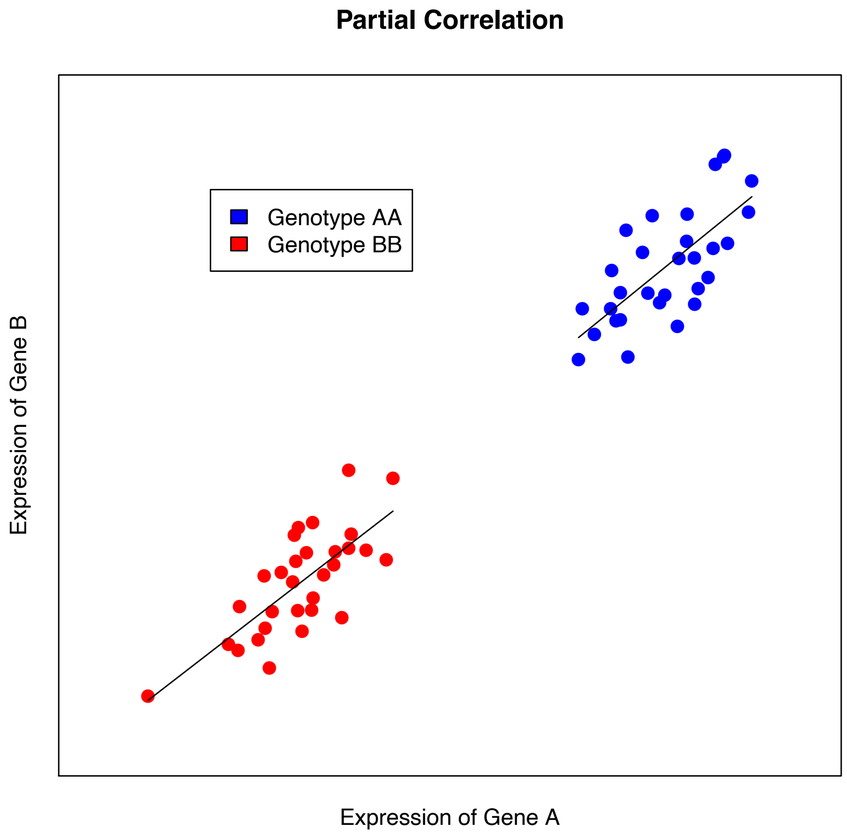
Partial Correlation As a Tool for Causal Inference

Theory - If changes in Gene A “cause” changes in Gene B, then regardless of the source of biological variation in Gene A it should correspond to variation in Gene B



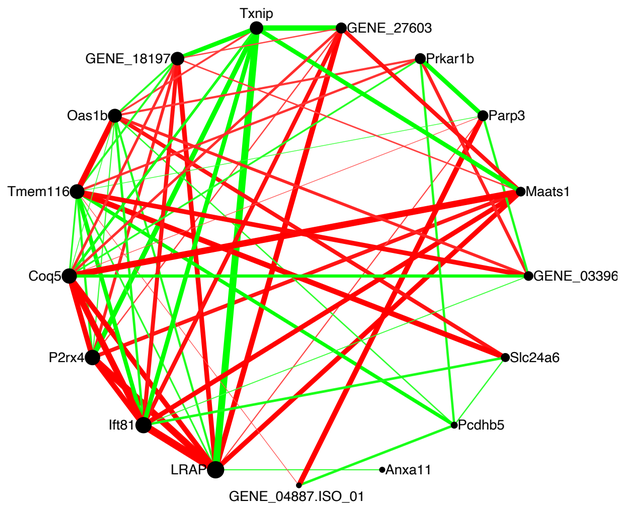
Partial Correlation As a Tool for Causal Inference



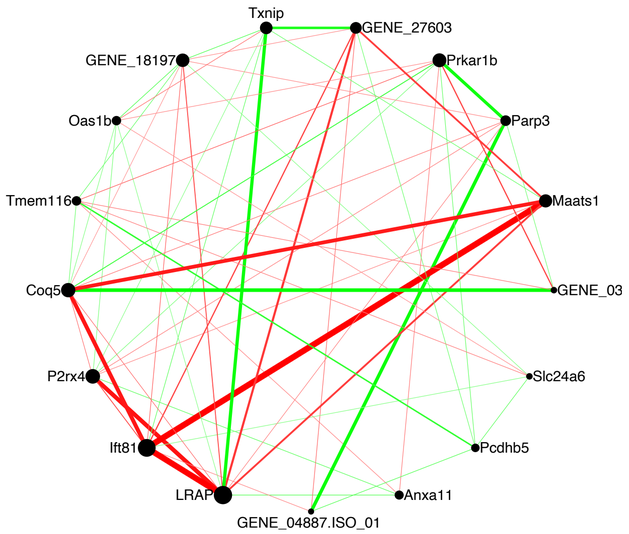


Comparisons with partial correlation

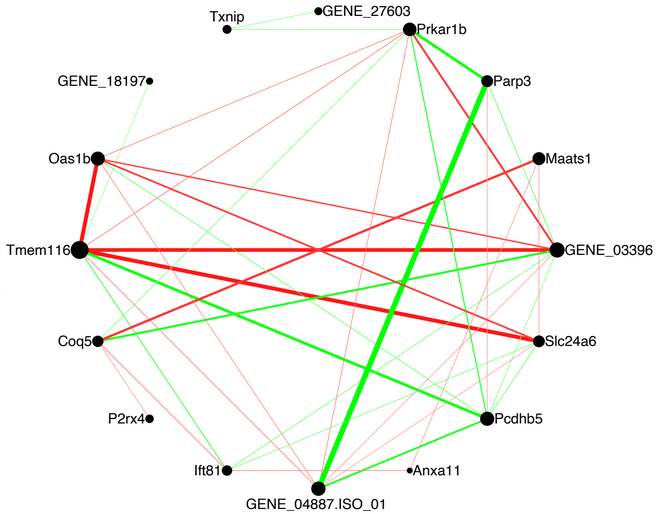
Original Correlations



Partial Correlation After Accounting for the Influence of the Module Eigengene QTL



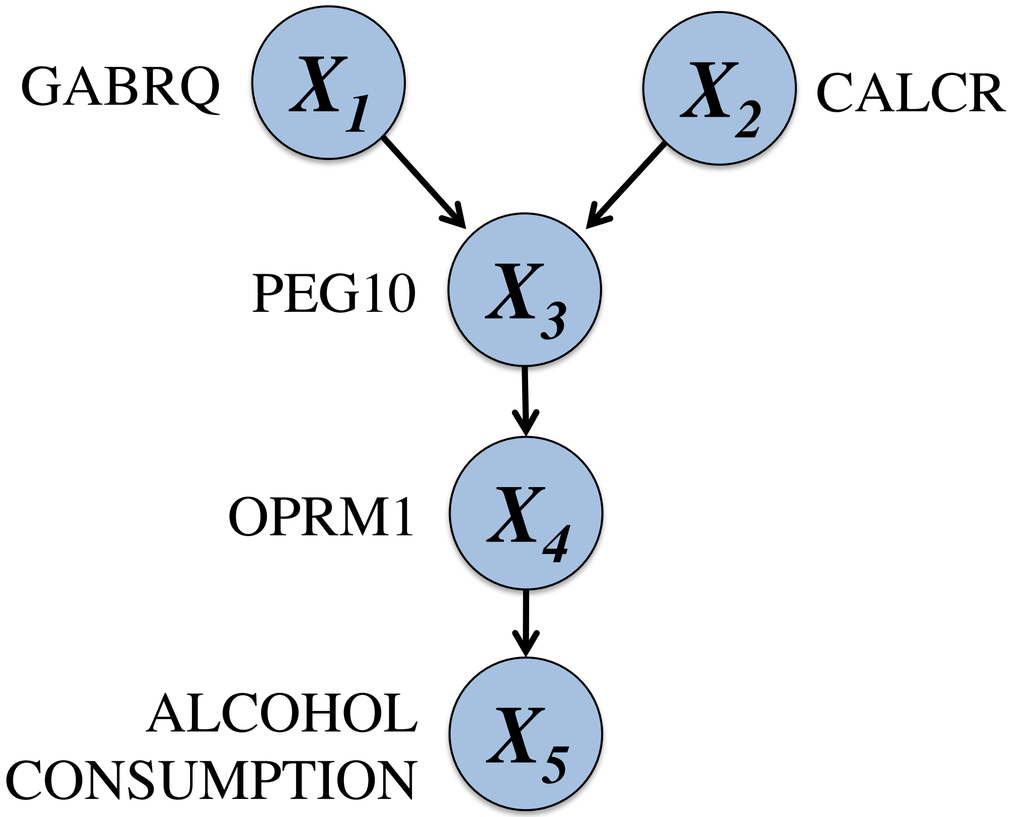
Partial Correlation After Accounting for Lrap Expression



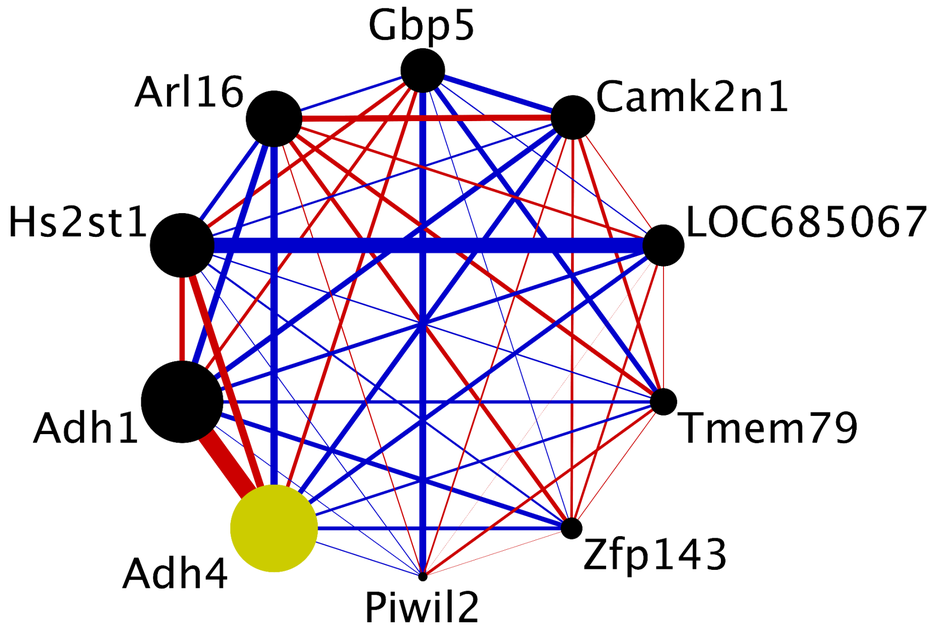
Lrap expression is a major driven of co-expression in the candidate module.

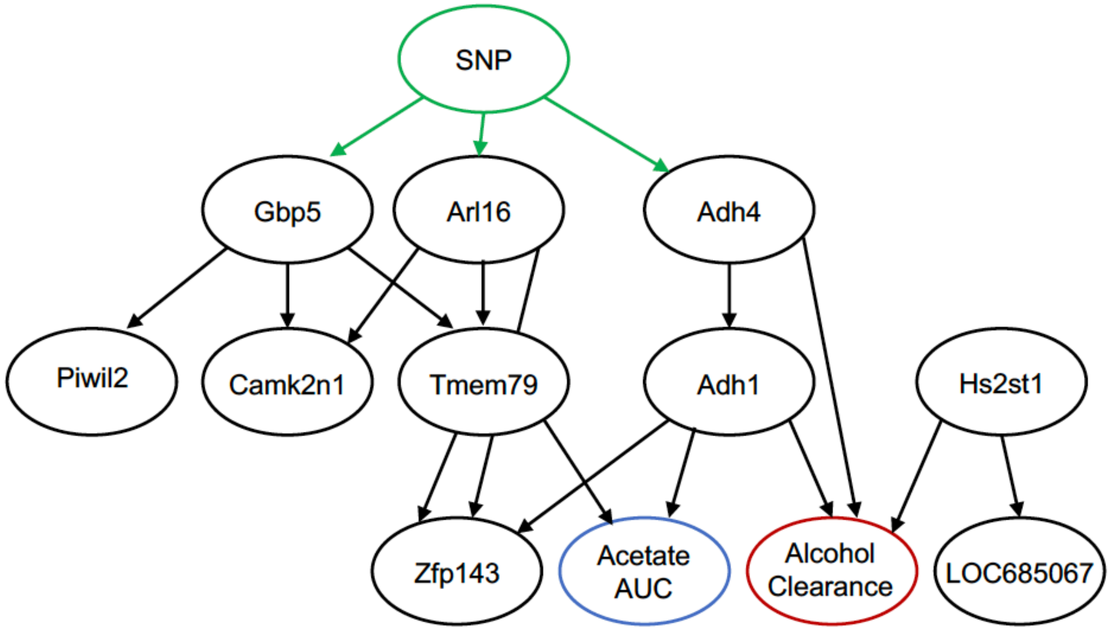
3. Causality testing/Bayesian Networks

- Graphical model of the causal relationships among variables
- Can incorporate multiple data types and prior knowledge
- Can be used for both diagnostic reasoning and predictive reasoning



BN with Alcohol Metabolism





Graphics courtesy of Ryan Lusk

Limitations of WGCNA

- Assumption of **linear or monotonic relationships** among genes
- **Scale-free networks** may be appropriate for relationships among protein-coding genes, but how can/should non-coding transcripts or other omics data be included?
- **Robustness/repeatability** - see Langfelder P, Luo R, Oldham MC, Horvath S (2011) Is My Network Module Preserved and Reproducible?. PLOS Computational Biology 7(1): e1001057 for network-level reproducibility, but what about evaluation of individual genes within module?
- Assumption that co-expression is related to similar biological function rather than **linkage disequilibrium**.
- **Prior information** about interacting genes or established networks is not included in network construction.

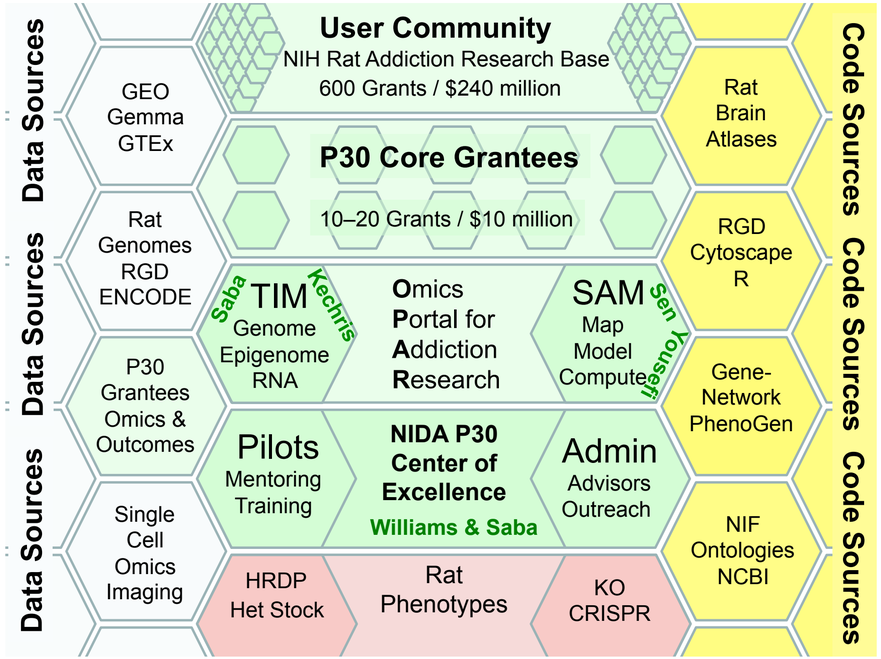
Conclusions

- **Why networks?**
  - Provides context to the role of individual genes in a particular experiment/environment
  - Allows for the development of functional hypotheses
  - Does a better job of mimicking what is going on in the cell
- **Basics of WGCNA**
  - WGCNA goes beyond simple correlation to model the data as an approximately scale-free network and incorporates indirect and direct relationships among genes to derive a more robust measure of co-expression
- **WGCNA algorithmic details**
  - WGCNA can be easily broken down into several concise steps to construct the network, identify co-expression modules, and evaluate the modules for their association with external data, for their association with other modules, and about the roles of individual genes within a module

NIDA Core Center of Excellence in Omics, System Genetics, and the Addictome

Co-Directors: Rob Williams (UTHSC) and Laura Saba (CU-AMC)

The **purpose** of the NIDA P30 Core Center of Excellence in Omics, Systems Genetics, and the Addictome is to empower and train researchers supported by NIH, NIDA, NIAAA, and other federal and state institutions to use more quantitative and testable ways to analyze genetic, epigenetic, and the environmental factors that influence drug abuse risk and treatment.



Our Approach:

- Omics Portal for Addiction Research (OPAR)
- Study design and RNA-Seq analysis services
- Training in Systems Genetics, RNA-Seq, and OPAR usage
- Funding for pilot grants

Acknowledgements

- Saba Lab:
  - Current: Ryan Lusk, Cheyret Wood, and Samuel Rosean
  - Former: Lauren Vanderlinden, Harry Smith, and Sean Hickey
- Boris Tabakoff, Paula Hoffman, and their lab
  - Spencer Mahaffey and Jenny Mahaffey

- Financial Support:
  - NIDA Core “Center of Excellence" in Omics, Systems Genetics and the Addictome (NIDA - P30DA044223; MPIs - Williams, Saba)
  - The heritable transcriptome and alcoholism (NIAAA - R24AA013162; MPIs - Tabakoff, Hoffman, Saba