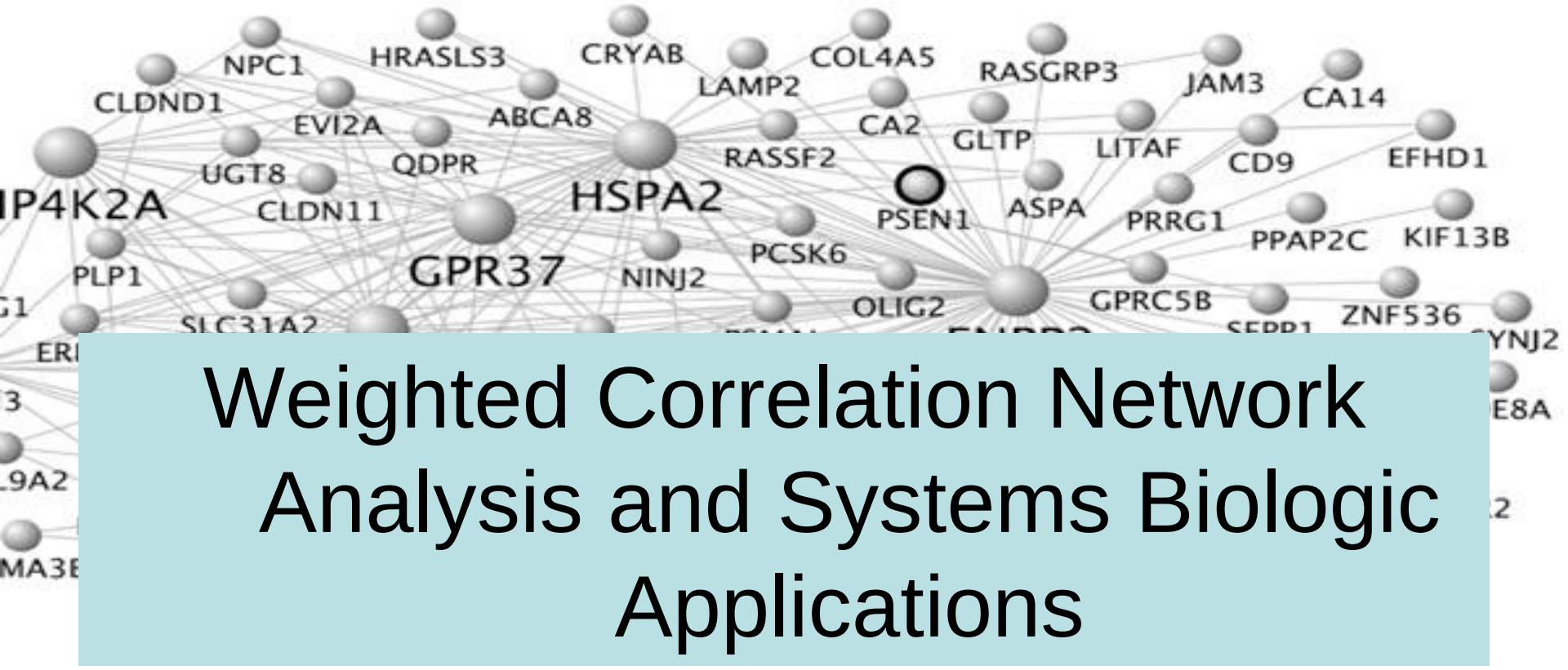


Weighted Correlation Network Analysis and Systems Biologic Applications

Steve Horvath

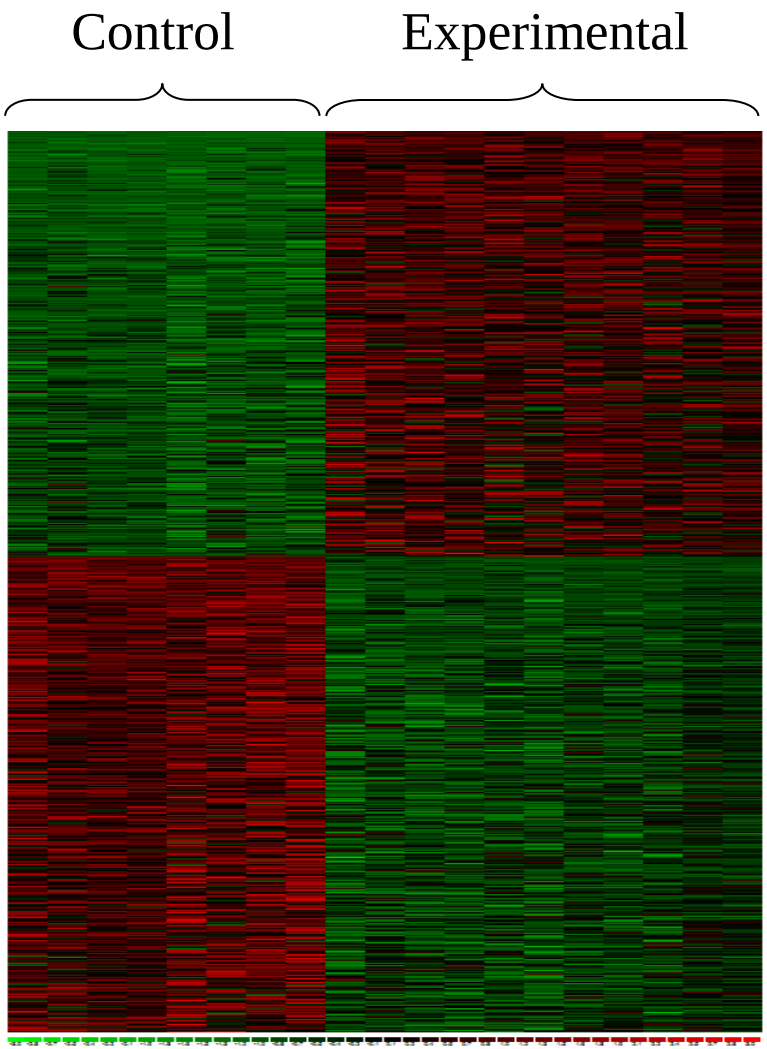
University of California, Los Angeles

Contents

- Weighted correlation network analysis (WGCNA)
- Module preservation statistics
- Applications
  - Atlas of the adult human brain transcriptome
  - Age related co-methylation modules

Bonus material: Epigenetic clock

Standard differential expression analyses seek to identify individual genes



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

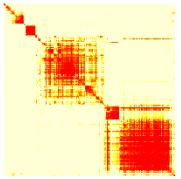
Philosophy of Weighted Gene Co-Expression Network Analysis

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

What is weighted gene co-expression network analysis?

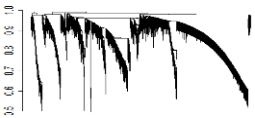
Construct a network

Rationale: make use of interaction patterns between genes



Identify modules

Rationale: module (pathway) based analysis

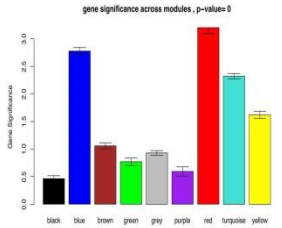


Relate modules to external information

Array Information: Clinical data, SNPs, proteomics

Gene Information: gene ontology, EASE, IPA

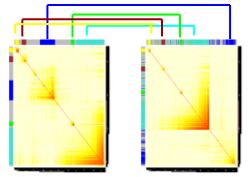
Rationale: find biologically interesting modules



Study Module Preservation across different data

Rationale:

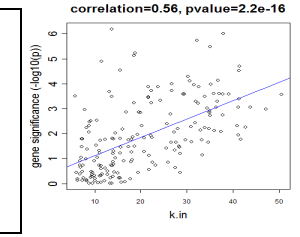
- Same data: to check robustness of module definition
- Different data: to find interesting modules.



Find the key drivers in *interesting* modules

Tools: intramodular connectivity, causality testing

Rationale: experimental validation, therapeutics, biomarkers



Weighted correlation networks are valuable for a biologically meaningful…

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

Review of weighted correlation network analysis

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 node pairs
  - Our convention: diagonal elements of *A* are all 1.

Two types of weighted correlation networks

Unsigned network, absolute value

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

Signed network preserves sign info

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

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

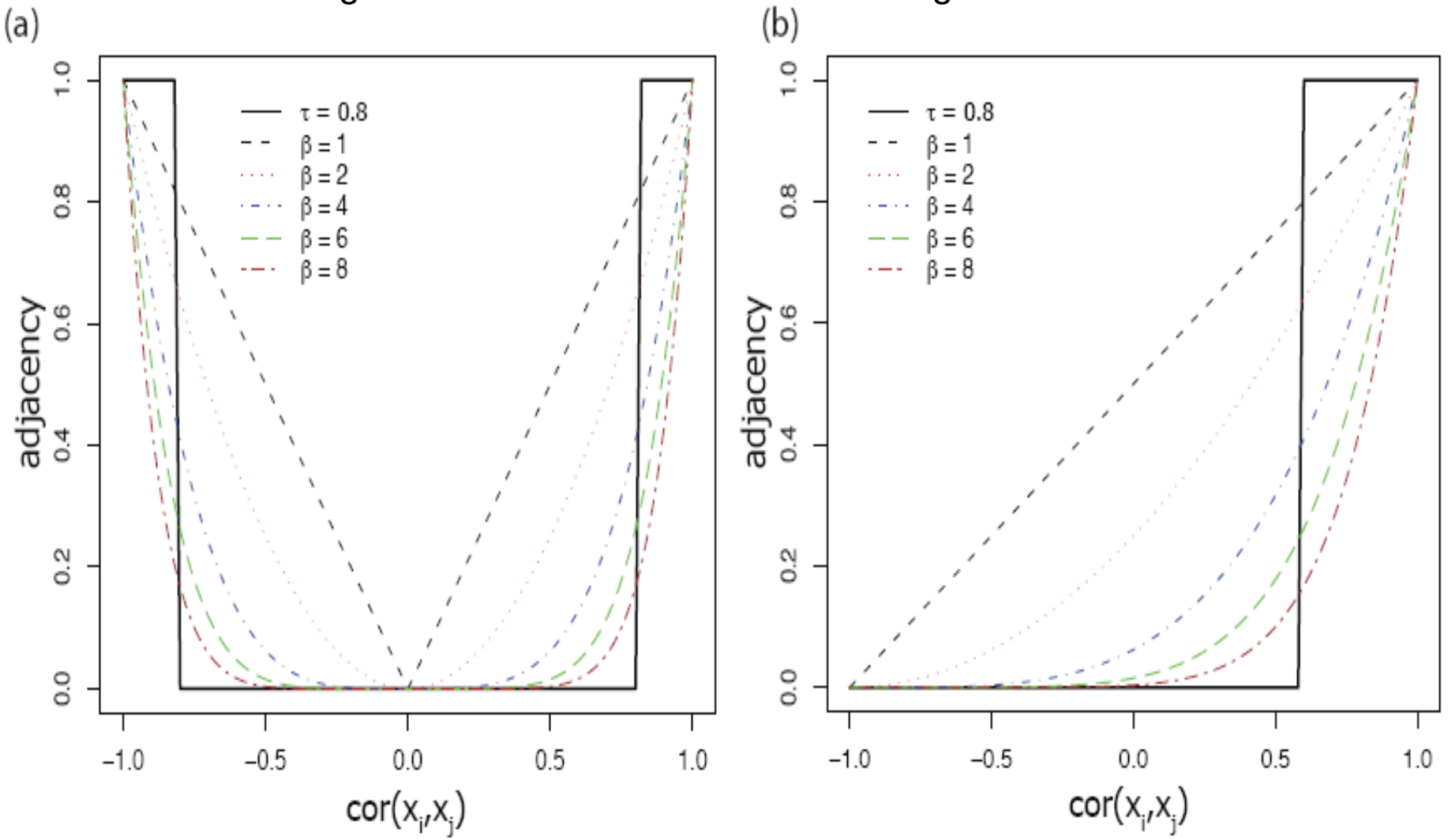
We prefer signed networks…

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

Adjacency versus correlation in unsigned and signed networks

Unsigned Network

Signed Network



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

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

*Biweight midcorrelation (bicor)

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

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

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

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

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

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

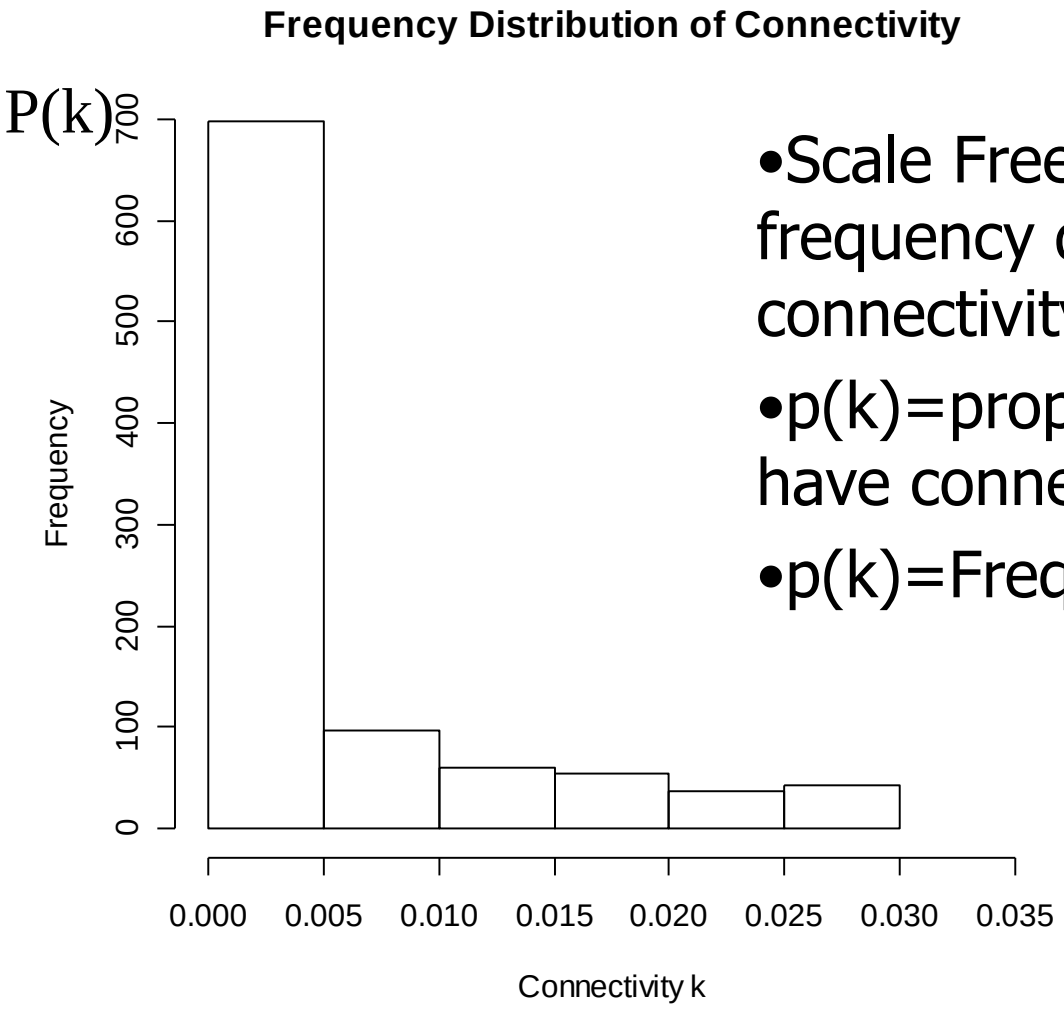
Langfelder et al 2012: Fast R Functions For Robust Correlations And Hierarchical Clustering. *J Stat Softw* 2012, **46**(i11):1–17.

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

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

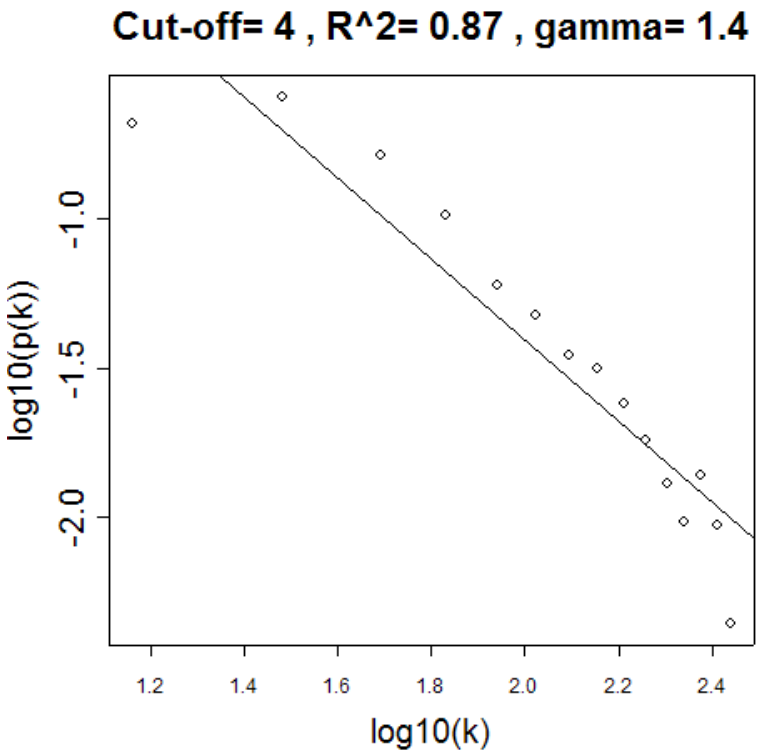
P(k) vs k in scale free networks



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

How to check Scale Free Topology?

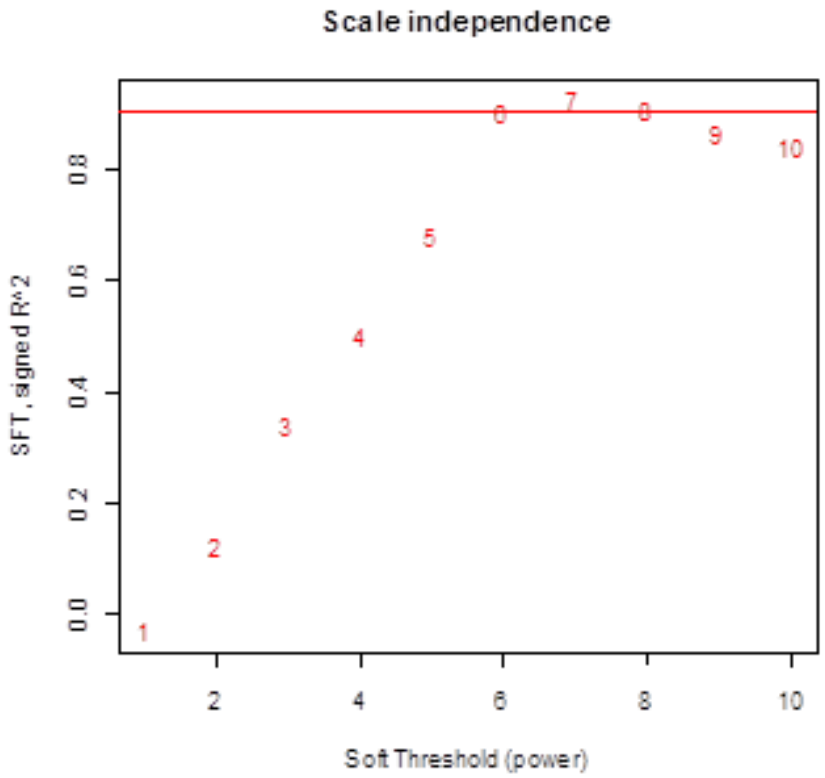
Idea: Log transformation p(k) and k and look at scatter plots



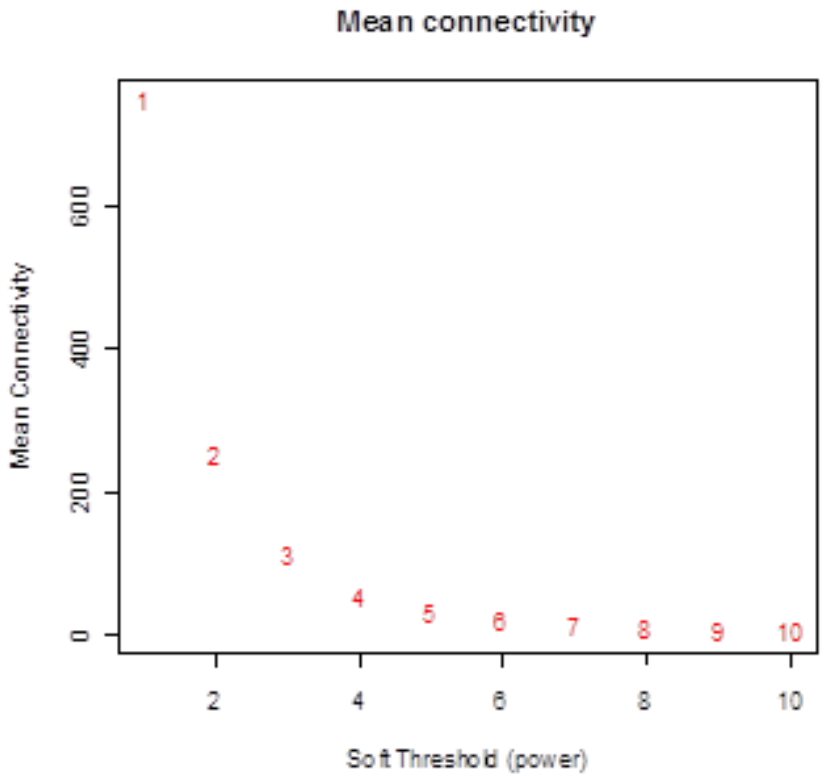
Linear model fitting R^2 index can be used to quantify goodness of fit

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

SFT model fitting index R^2



mean connectivity



From your software tutorial

How to measure interconnectedness in a network?

Answers:

1) adjacency matrix

2) topological overlap matrix

Topological overlap matrix and corresponding dissimilarity

(Ravasz et al 2002)

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

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

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

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

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

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

Advantages of soft thresholding with the power function

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

How to detect network modules (clusters) ?

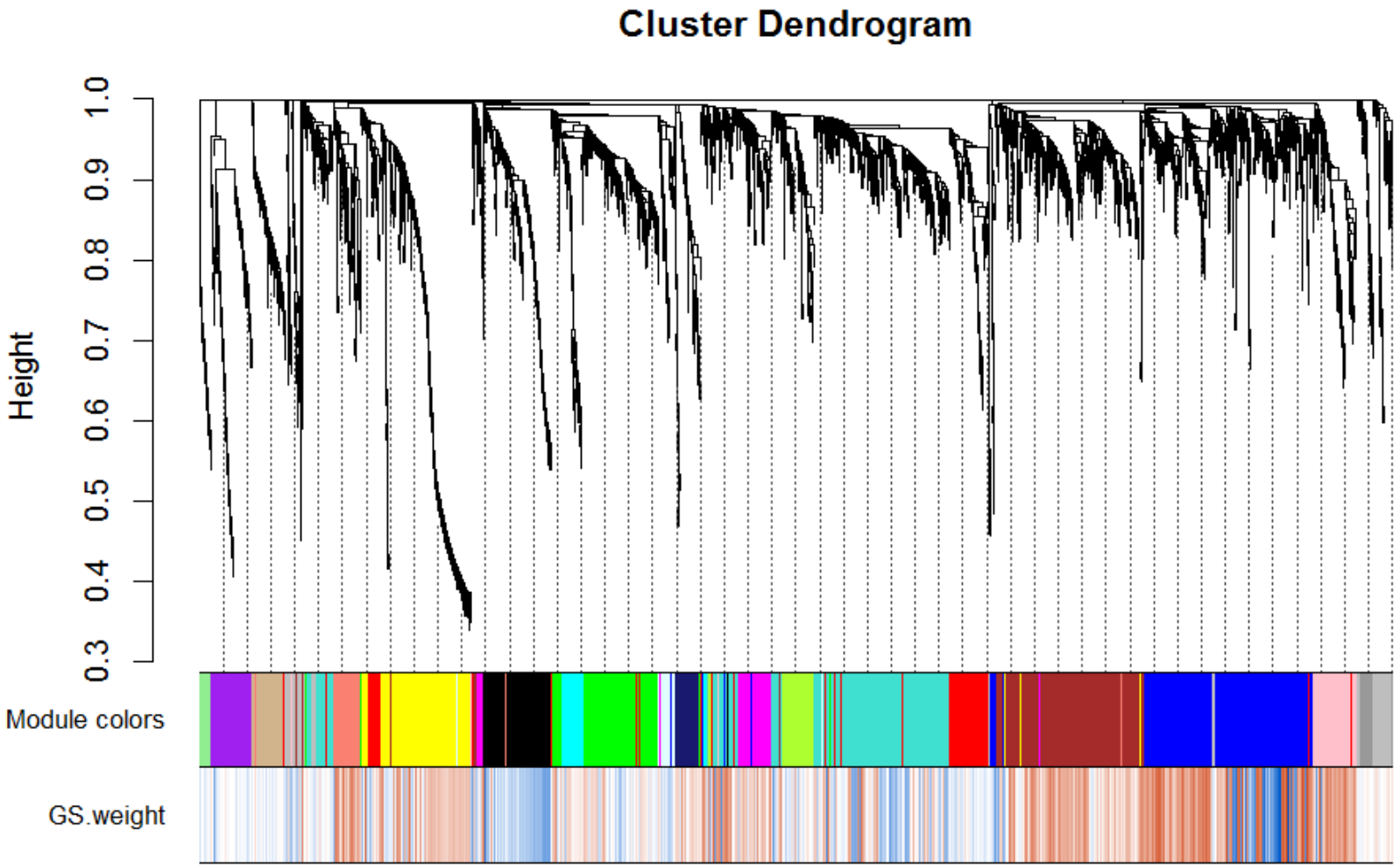
Module Definition

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

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

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

Example:

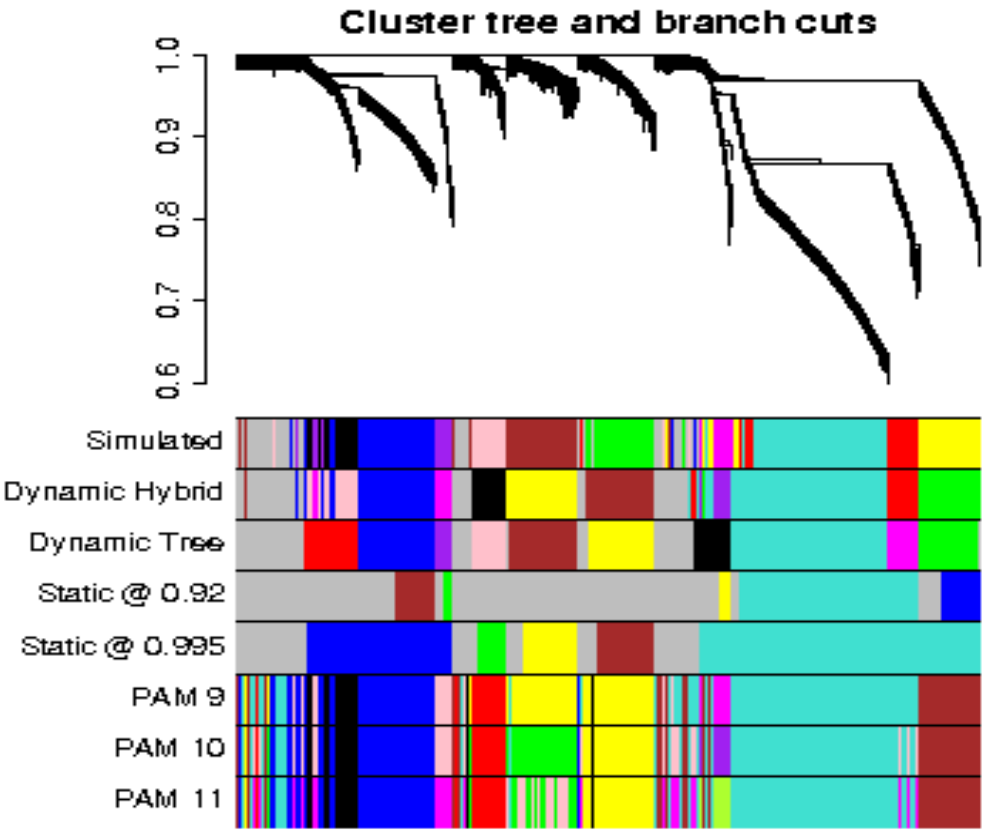


From your software tutorial

Two types of branch cutting methods

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

How to cut branches off a tree?



Module=branch of a cluster tree

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

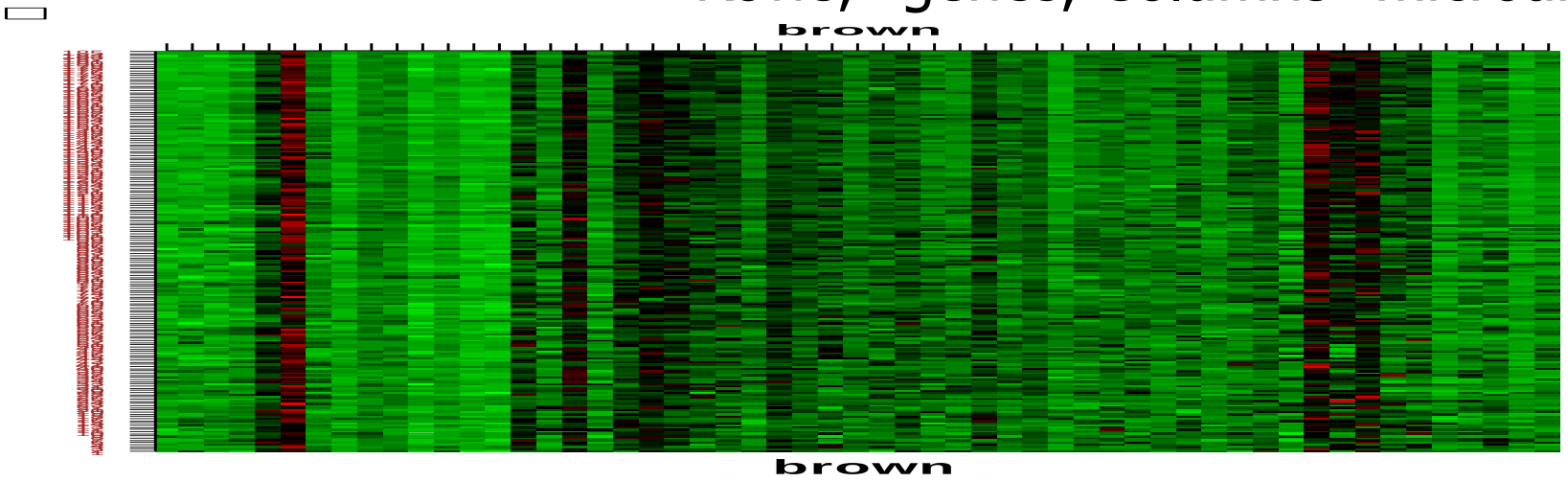
Question: How does one summarize the expression profiles in a module?

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

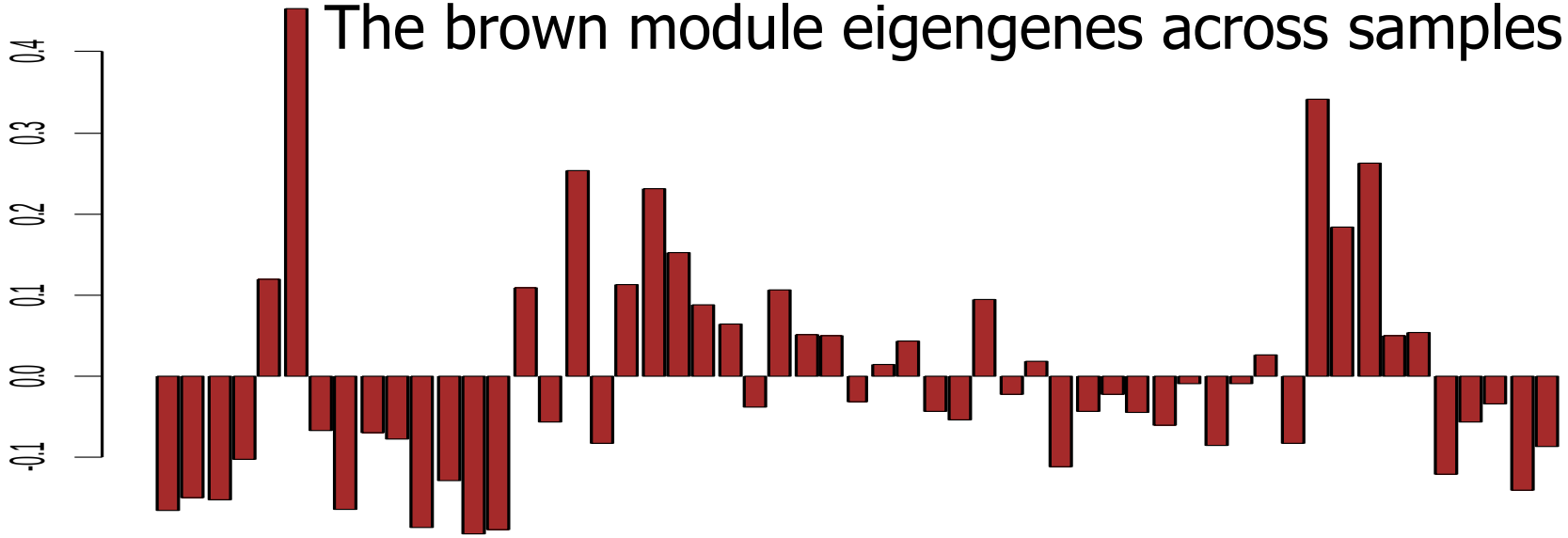
Both turn out to be equivalent

Module Eigengene= measure of over-expression=average redness

Rows,=genes, Columns=microarray



The brown module eigengenes across samples



Module eigengene is defined by the singular value decomposition of X

- X=gene expression data <u>of a module</u> gene expressions (rows) have been standardized across samples (columns)

$$X = \tilde{U}\tilde{D}\tilde{V}^{\tau}$$

$$\tilde{U} = (\tilde{u}_1 \quad \tilde{u}_2 \quad \cdots \quad \tilde{u}_m)$$

$$\tilde{V} = (\tilde{v}_1 \quad \tilde{v}_2 \quad \cdots \quad \tilde{v}_m)$$

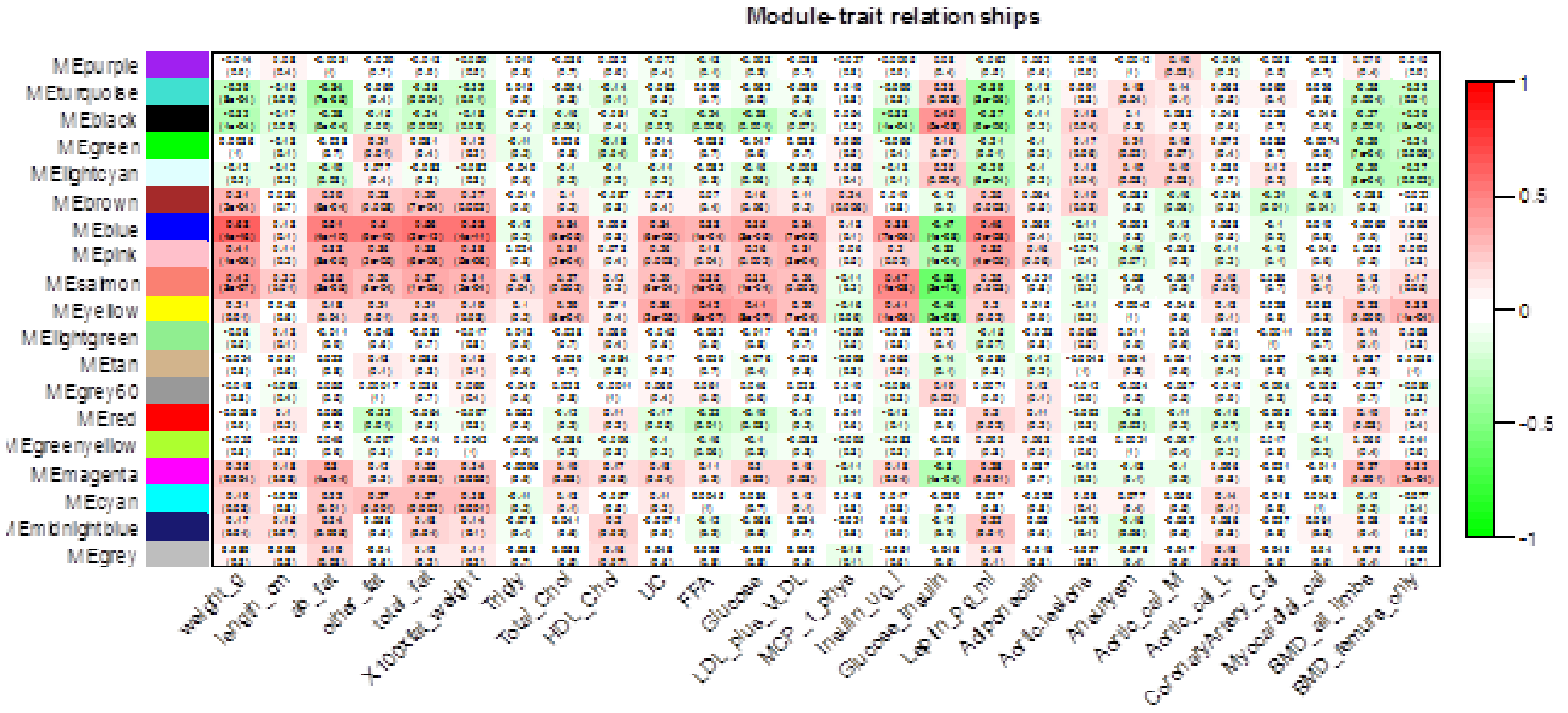
$$\tilde{D} = diag(|\tilde{d}_1|, |\tilde{d}_2|, \ldots, |\tilde{d}_m|)$$

Message: $\tilde{v}_1$ is the module eigengene E

Module eigengenes are very useful

- 1) They allow one to relate modules to each other
  - Allows one to determine whether modules should be merged
- 2) They allow one to relate modules to clinical traits and SNPs
  - -> avoids multiple comparison problem
- 3) They allow one to define a measure of module membership: kME=cor(x,ME)
  - Can be used for finding centrally located hub genes
  - Can be used to define gene lists for GO enrichment

Table of module-trait correlations and p-values.
Each cell reports the correlation (and p-value) resulting from correlating module eigengenes (rows) to traits (columns). The table is color-coded by correlation according to the color legend.



Module detection in very large data sets

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

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

Works for hundreds of thousands of variables

Eigengene based connectivity, also known as kME or module membership measure

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

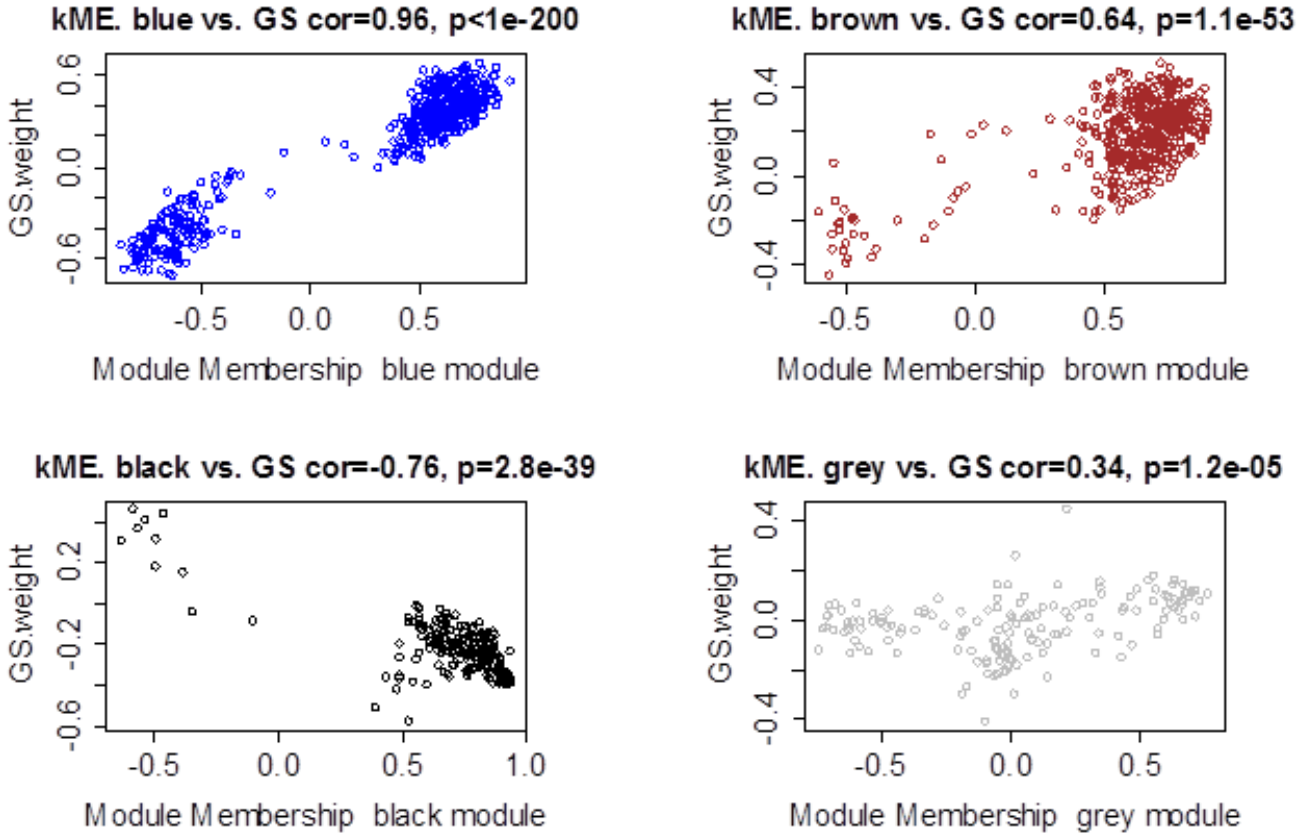
kME(i) is simply the correlation between the i-th gene expression profile and the module eigengene.

kME close to 1 means that the gene is a hub gene

Very useful measure for annotating genes with regard to modules.

Module eigengene turns out to be the most highly connected gene

Gene significance vs kME



Gene significance (GS.weight) versus module membership (kME) for the body weight related modules. GS.weight and MM.weight are highly correlated reflecting the high correlations between weight and the respective module eigengenes.

We find that the brown, blue modules contain genes that have high positive and high negative correlations with body weight. In contrast, the grey "background" genes show only weak correlations with weight.

Intramodular hub genes

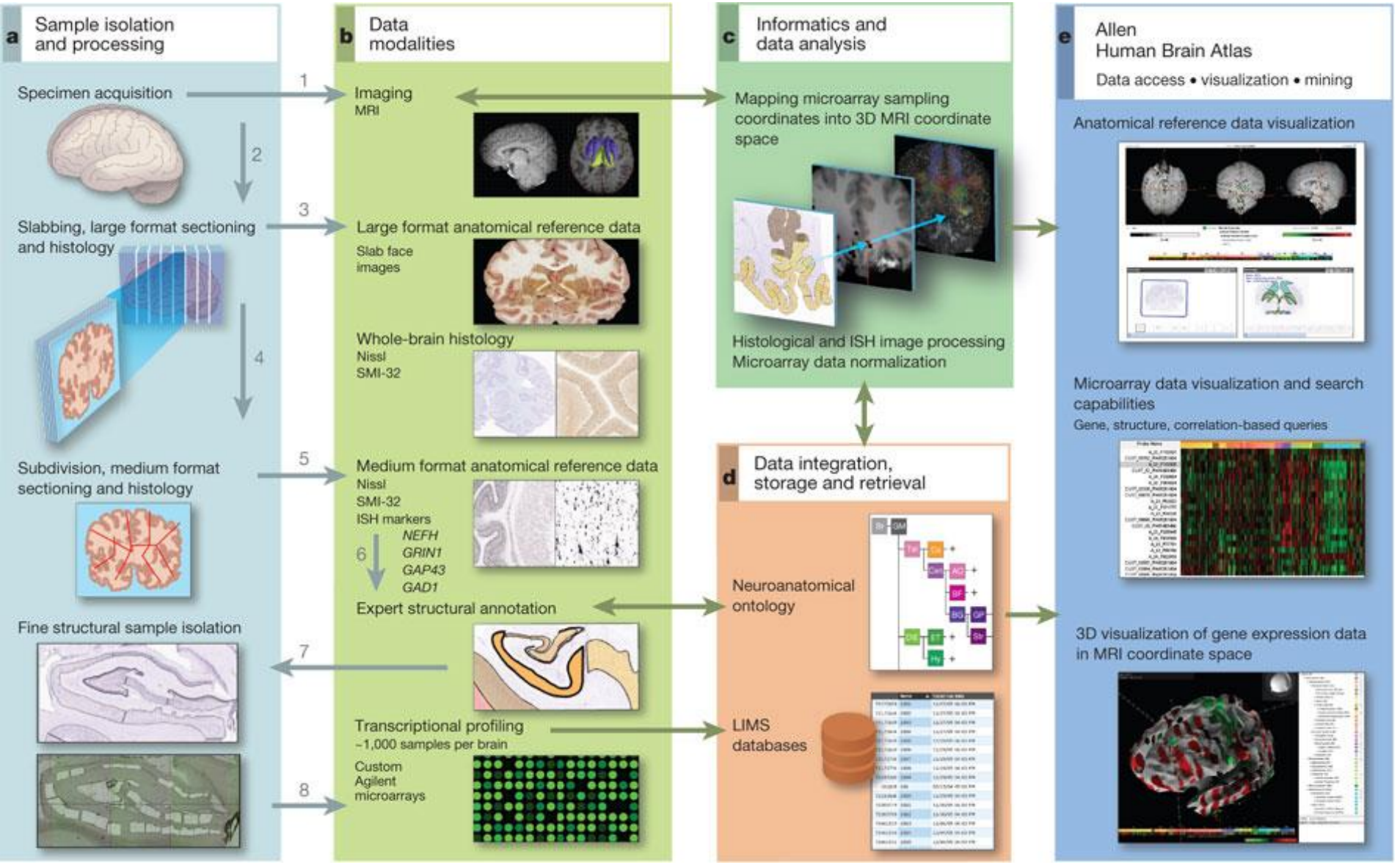
- Defined as genes with high kME (or high kIM)
- Single network analysis: Intramodular hubs in biologically interesting modules are often very interesting
- Differential network analysis: Genes that are intramodular hubs in one condition but not in another are often very interesting

An anatomically comprehensive atlas of the adult human brain transcriptome

MJ Hawrylycz, E Lein,..,AR Jones (2012) Nature 489, 391-399

Allen Brain Institute

Data generation and analysis pipeline



MJ Hawrylycz *et al. Nature* **489**, 391-399 (2012) doi:10.1038/nature11405

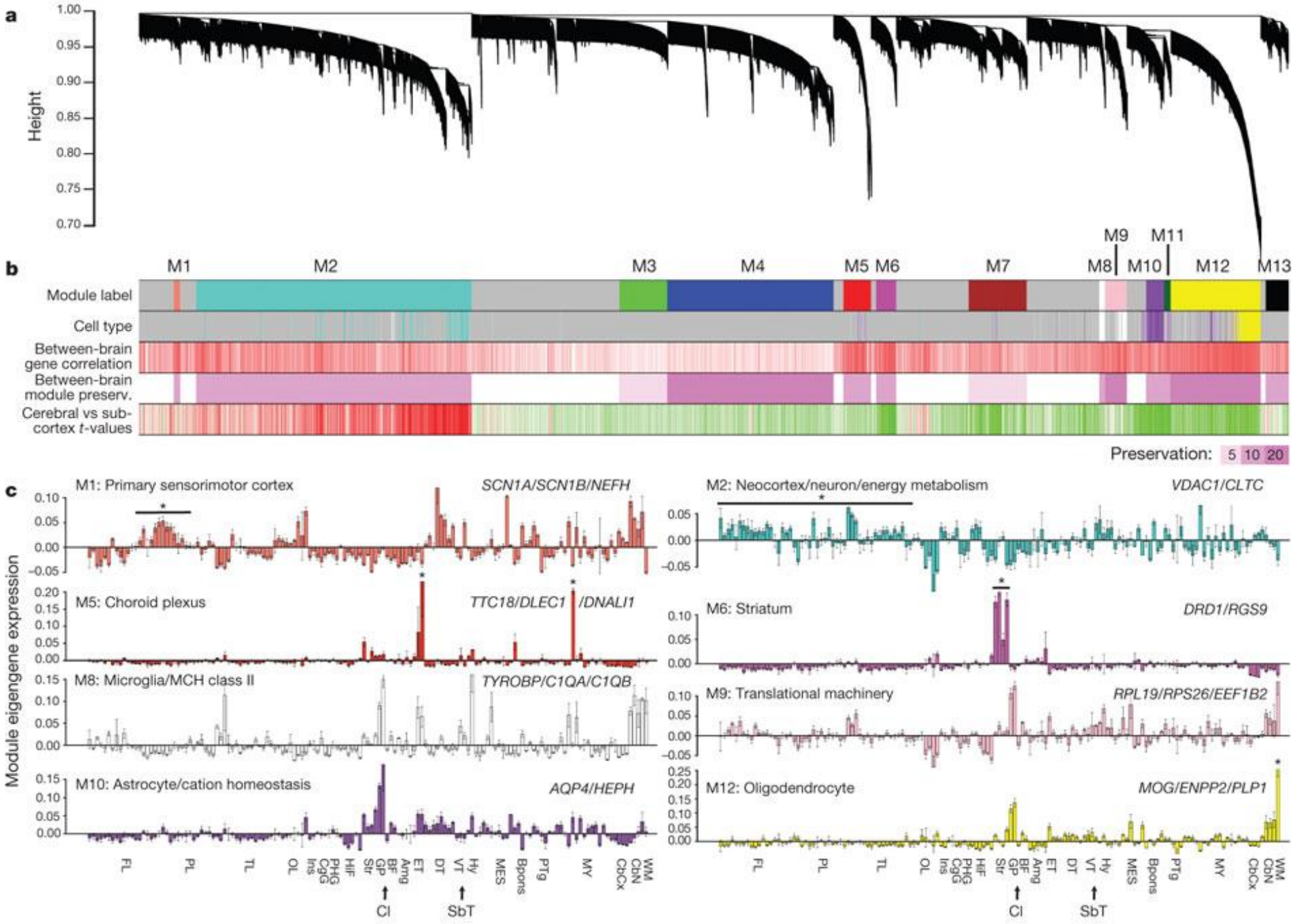
Data

- Brains from two healthy males (ages 24 and 39)
- 170 brain structures
- over 900 microarray samples per individual
- 64K Agilent microarray
- This data set provides a neuroanatomically precise, genome-wide map of transcript distributions

Why use WGCNA?

1. Biologically meaningful data reduction
   - WGCNA can find the dominant features of transcriptional variation across the brain, beginning with global, brain-wide analyses
   - It can identify gene expression patterns related to specific cell types such as neurons and glia from heterogeneous samples such as whole human cortex
     - Reason: highly distinct transcriptional profiles of these cell types and variation in their relative proportions across samples (Oldham et al Nature Neurosci. 2008).
2. Module eigengene
   - To test whether modules change across brain structures.
3. Measure of module membership (kME)
   - To create lists of module genes for enrichment analysis
4. Module preservation statistics
   - To study whether modules found in brain 1 are also preserved in brain 2 (and brain 3).

Modules in brain 1



Caption

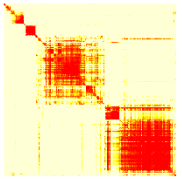
- **a**, Cluster dendrogram using all samples in Brain 1
- **b**, Top colour band: colour-coded gene modules.
- Second band: genes enriched in different cell types (400 genes per cell type) selectively overlap specific modules.
  - Turquoise, neurons; yellow, oligodendrocytes; purple, astrocytes; white, microglia.
  - Fourth band: strong preservation of modules between Brain 1 and Brain 2, measured using a Z-score summary ($Z \geq 10$ indicates significant preservation).
- Fifth band: cortical (red) versus subcortical (green) enrichment (one-side t-test).
- **c**, Module eigengene expression (y axis) is shown for eight modules across 170 subregions with standard error. Dotted lines delineate major regions
- An asterisk marks regions of interest. Module eigengene classifiers are based on structural expression pattern, putative cell type and significant GO terms. Selected hub genes are shown.

Module Preservation

Module preservation is often an essential step in a network analysis

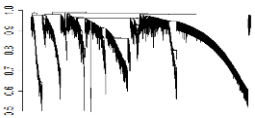
Construct a network

Rationale: make use of interaction patterns between genes



Identify modules

Rationale: module (pathway) based analysis

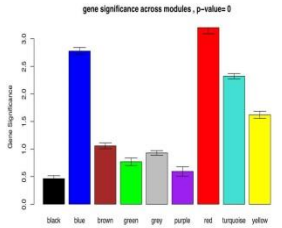


Relate modules to external information

Array Information: Clinical data, SNPs, proteomics

Gene Information: gene ontology, EASE, IPA

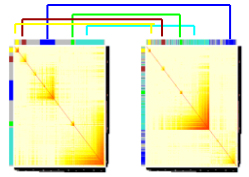
Rationale: find biologically interesting modules



Study Module Preservation across different data

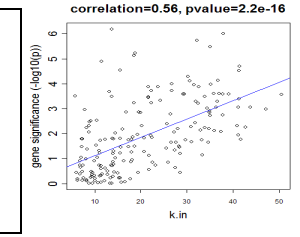
Rationale:

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



Find the key drivers of *interesting* modules

Rationale: experimental validation, therapeutics, biomarkers



Is my network module preserved and reproducible?

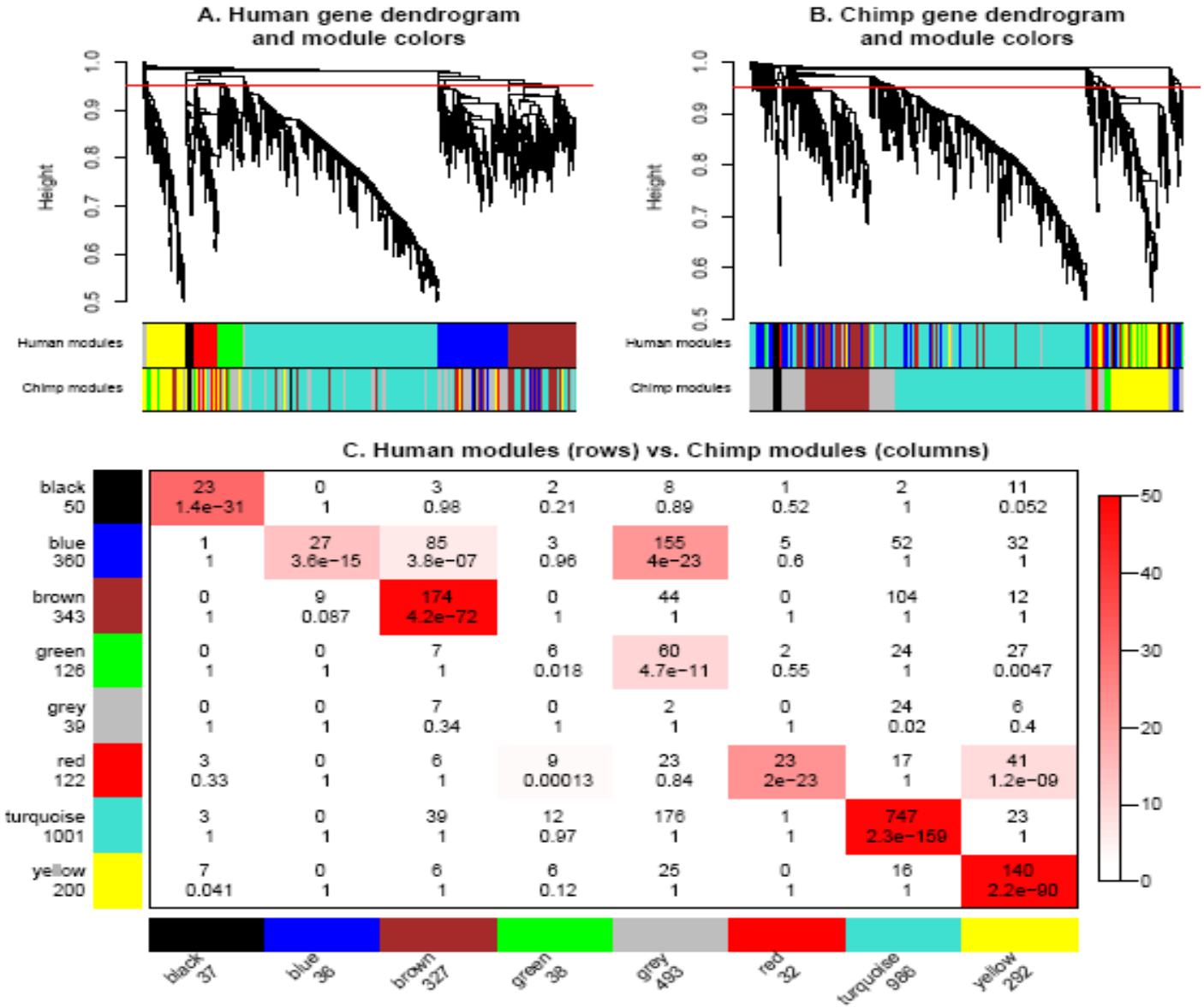
Langfelder et al PloS Comp Biol. 7(1): e1001057.

Motivational example: Studying the preservation of human brain co-expression modules in chimpanzee brain expression data.

Modules defined as clusters
(branches of a cluster tree)

Data from Oldham et al 2006 PNAS

Preservation of modules between human and chimpanzee brain networks



Standard cross-tabulation based statistics have severe disadvantages

Disadvantages

1. only applicable for modules defined via a clustering procedure
2. ill suited for making the strong statement that a module is not preserved

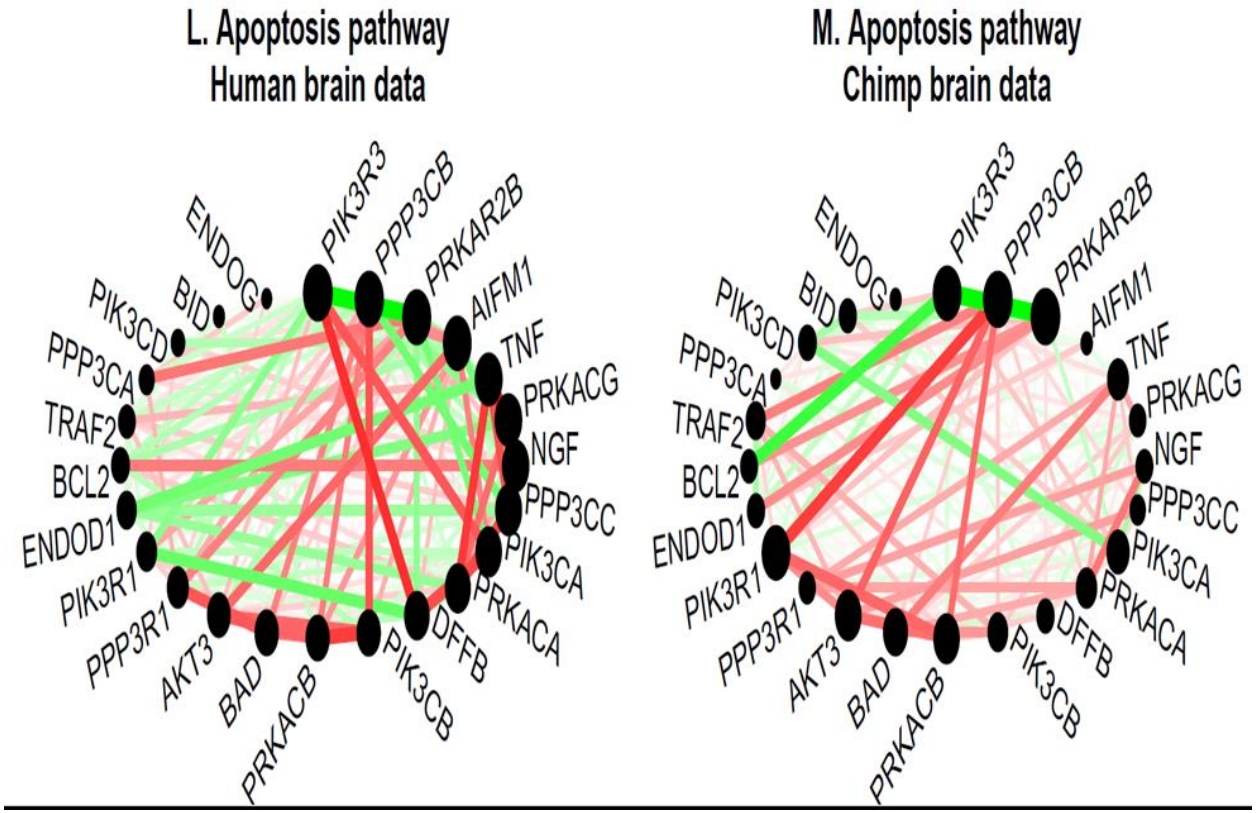
We argue that network based approaches are superior when it comes to studying module preservation

Broad definition of a module

- Abstract definition of module=subset of nodes in a network.
- Thus, a module forms a sub-network in a larger network
- Example: module (set of genes or proteins) defined using external knowledge: KEGG pathway, GO ontology category
- Example: modules defined as clusters resulting from clustering the nodes in a network
- Module preservation statistics can be used to evaluate whether a given module defined in one data set (reference network) can also be found in another data set (test network)

How to measure relationships between different networks?

- Answer: network statistics



Weighted gene co-expression module.
Red lines=positive correlations,
Green lines=negative cor

Connectivity (aka degree)

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

$$Connectivity_i = k_i = \sum_{j \neq i} a_{ij}$$

$$\text{Scaled connectivity}=\mathrm{K_i} = \frac{k_i}{\max(k)}$$

Density

- Density= mean adjacency
- Highly related to mean connectivity

$$Density = \frac{\sum_i \sum_{j \neq i} a_{ij}}{n(n-1)} = \frac{mean(k)}{n-1}$$

where n is the number of network nodes.

Network-based module preservation statistics

- Input: module assignment in reference data.
- Adjacency matrices in reference A^{ref} and test data A^{test}
- Network preservation statistics assess preservation of
  - 1. network density: Does the module remain densely connected in the test network?
  - 2. connectivity: Is hub gene status preserved between reference and test networks?
  - 3. separability of modules: Does the module remain distinct in the test data?

Module preservation in different types of networks

- One can study module preservation in general networks specified by an adjacency matrix, e.g. protein-protein interaction networks.
- However, particularly powerful statistics are available for correlation networks
  - weighted correlation networks are particularly useful for detecting subtle changes in connectivity patterns. But the methods are also applicable to unweighted networks (i.e. graphs)

Several connectivity preservation statistics

For general networks, i.e. input adjacency matrices

- *cor.kIM*=cor(kIM^{ref},kIM^{test})
  - *correlation of intramodular connectivity across module nodes*
- *cor.ADJ*=cor(A^{ref},A^{test})
  - *correlation of adjacency across module nodes*

For correlation networks, i.e. input sets are variable measurements

- cor.Cor=cor(cor^{ref},cor^{test})
- cor.kME=cor(kME^{ref},kME^{test})

One can derive relationships among these statistics in case of weighted correlation network

Choosing thresholds for preservation statistics based on permutation test

- For correlation networks, we study 4 density and 4 connectivity preservation statistics that take on values <= 1
- Challenge: Thresholds could depend on many factors (number of genes, number of samples, biology, expression platform, etc.)
- Solution: Permutation test. Repeatedly permute the gene labels in the test network to estimate the mean and standard deviation under the null hypothesis of no preservation.
- Next we calculate a Z statistic

$$Z = \frac{observed - mean_{permuted}}{sd_{permuted}}$$

Permutation test for estimating Z scores

- For each preservation measure we report the observed value and the permutation Z score to measure significance.

$$Z = \frac{observed - mean_{permuted}}{sd_{permuted}}$$

- Each Z score provides answer to “Is the module significantly better than a random sample of genes?”

- Summarize the individual Z scores into a composite measure called Z.summary

- Zsummary < 2 indicates no preservation, 2<Zsummary<10 weak to moderate evidence of preservation, Zsummary>10 strong evidence

Composite statistic in correlation networks based on Z statistics

Permutation test allows one to estimate Z version of each statistic

$$Z^{(q)}_{cor.Cor} = \frac{cor.Cor^{(q)} - E(cor.Cor^{(q)} \mid null)}{\sqrt{Var(cor.Cor^{(q)} \mid null)}}$$

Composite connectivity based statistics for correlation networks

$$Z^{(q)}_{connectivity} = median(Z^{(q)}_{cor.Cor}, Z^{(q)}_{cor.kME}, Z^{(q)}_{cor.A}, Z^{(q)}_{cor.kIM})$$

Composite density based statistics for correlation networks

$$Z^{(q)}_{density} = median(Z^{(q)}_{meanCor}, Z^{(q)}_{meanAdj}, Z^{(q)}_{propVarExpl}, Z^{(q)}_{meanKME})$$

Composite statistic of density and connectivity preservation

$$Z^{(q)}_{summary} = \frac{Z^{(q)}_{connectivity} + Z^{(q)}_{density}}{2}$$

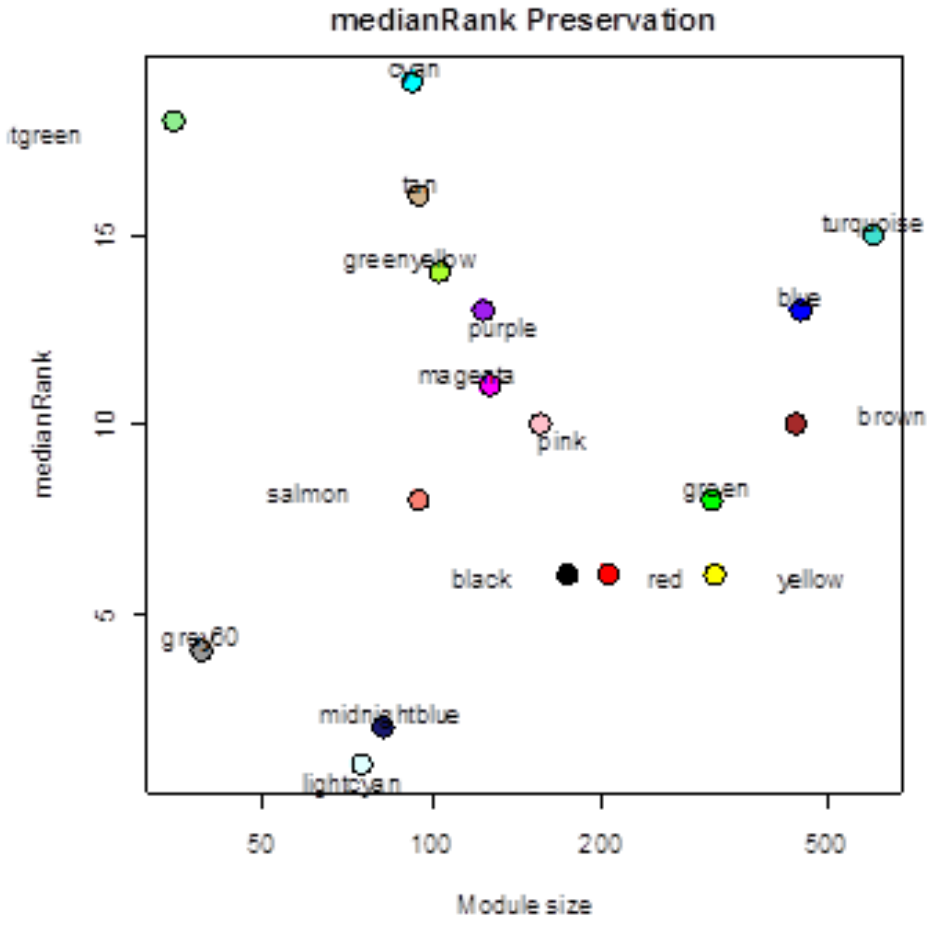
Analogously define composite statistic: medianRank

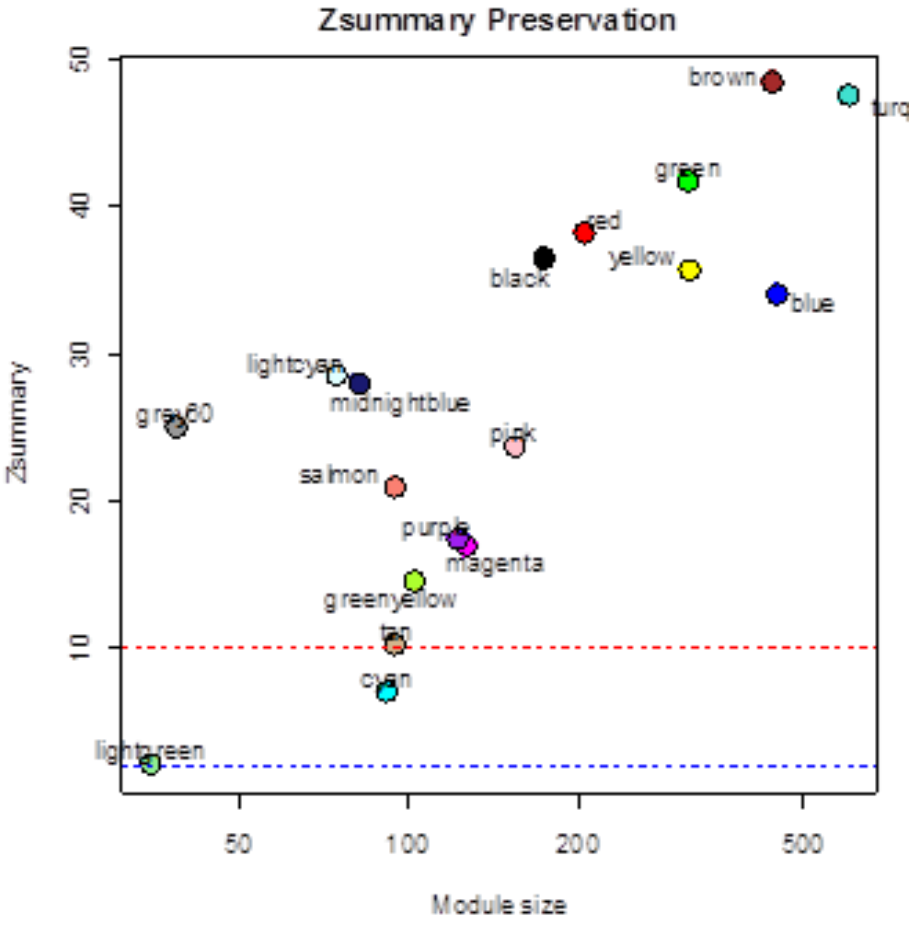
- Based on the ranks of the observed preservation statistics
- Does not require a permutation test
- Very fast calculation
- Typically, it shows no dependence on the module size

Overview
module preservation statistics

- Network based preservation statistics measure different aspects of module preservation
  - Density-, connectivity-, separability preservation
- Two types of composite statistics: Zsummary and medianRank.
- Composite statistic Zsummary based on a permutation test
  - Advantages: thresholds can be defined, R function also calculates corresponding permutation test p-values
  - Example: Zsummary<2 indicates that the module is *not* preserved
  - Disadvantages: i) Zsummary is computationally intensive since it is based on a permutation test, ii) often depends on module size
- Composite statistic medianRank
  - Advantages: i) fast computation (no need for permutations), ii) no dependence on module size.
  - Disadvantage: only applicable for ranking modules (i.e. relative preservation)

Preservation of female mouse liver modules in male livers.

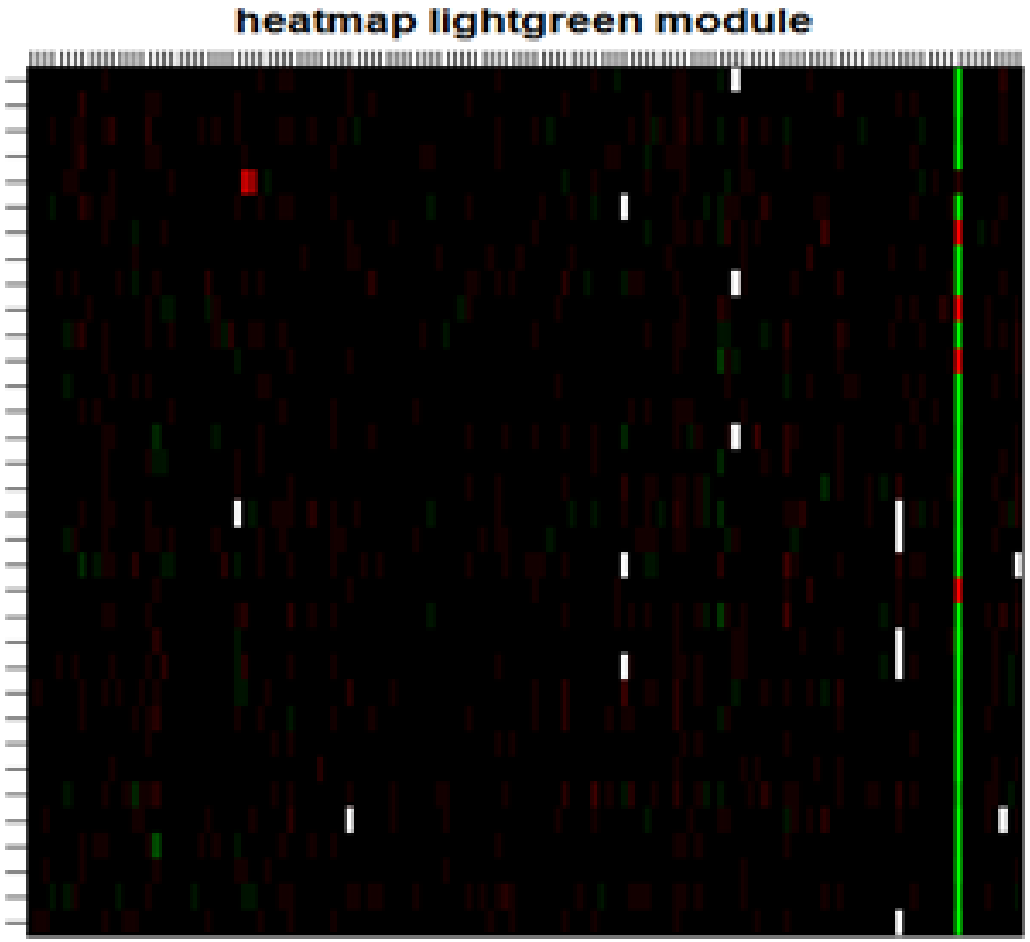




Lightgreen module is not preserved

Heatmap of the lightgreen module gene expressions (rows correspond to genes, columns correspond to female mouse tissue samples).

Note that most genes are under-expressed in a single female mouse, which suggests that this module is due to an array outliers.



Aging effects on DNA methylation modules in human brain and blood tissue

Collaborators:
Yafeng Zhang,
Peter Langfelder,
René S Kahn,
Marco PM Boks,
Kristel van Eijk,
Leonard H van den Berg,
Roel A Ophoff

Genome Biology 13:R97

DNA methylation: epigenetic modification of DNA

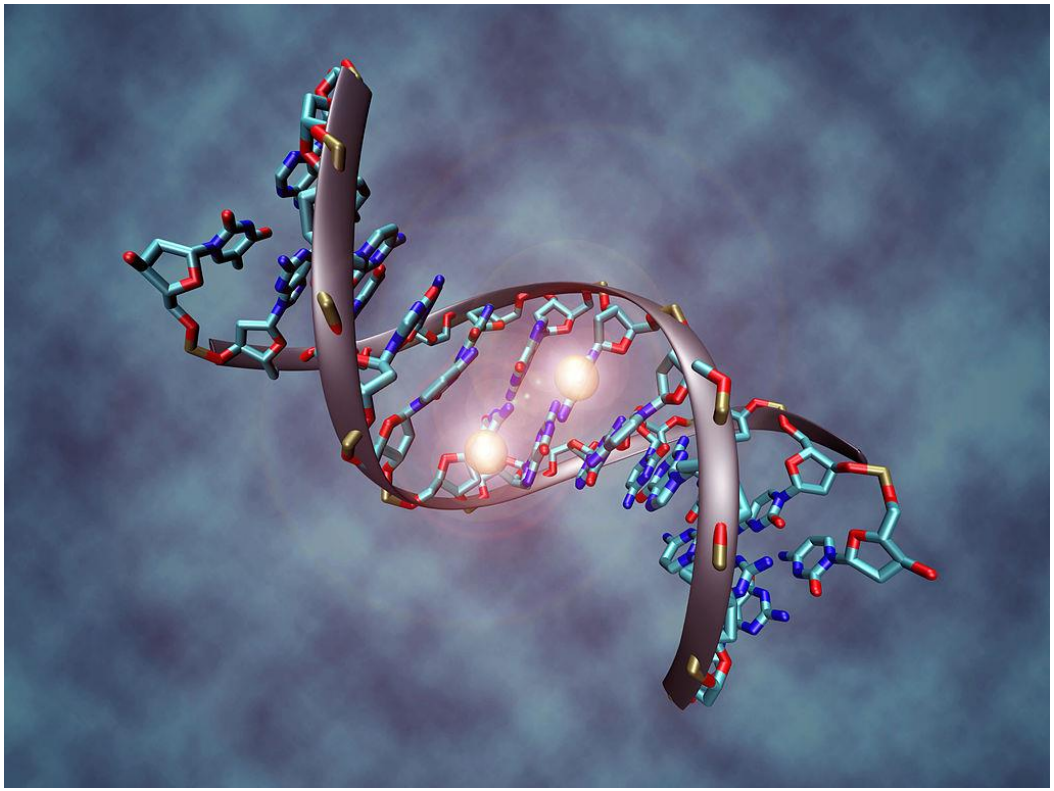
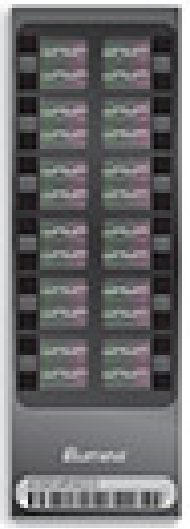


Illustration of a DNA molecule that is methylated at the two center cytosines. DNA methylation plays an important role for epigenetic gene regulation in development and disease.

Ilumina DNA methylation array (Infinium 450K beadchip)

- Measures over 480k locations on the DNA.
- It leads to 480k variables that take on values in the unit interval [0,1]
- Each variable specifies the amount of methylation that is present at this location.



Background

- Many articles have shown that age has a significant effect on DNA methylation levels
- Goals:
  a) Find age related co-methylation modules that are preserved in multiple human tissues
  b) Characterize them biologically
- Incidentally, it seems that this cannot be achieved for gene expression data.

Table 1. Description of DNA methylation data sets

Set No	Analysis	n	Tissue	Description	Mean Age	Age Range	Platform	Citation	Public Availability
1	Cons. Module	92	WB	Dutch controls from ALS study	64	34-88	Infin 27k	novel data	GSE41037
2	Cons. Module	273	WB	Dutch controls from SZ study	33	16-65	Infin 27k	novel data	GSE41037
3	Cons. Module	293	WB	Dutch Cases, SZ	34	17-86	Infin 27k	novel data	GSE41037
4	Cons. Module	190	WB	Type 1 diabetics	44	24-74	Infin 27k	Teschendorff 2010	GSE20067
5	Cons. Module	87	WB	Healthy older women	63	49-74	Infin 27k	Rakyan 2010	GSE20236
6	Cons. Module	261	WB	healthy postmenopausal women from UKOPS	65	52-78	Infin 27k	Teschendorff 2010, Song 2009	GSE19711
7	Cons. Module	132	FCTX	FCTX brain	48	16-101	Infin 27k	Gibbs 2010	GSE15745
8	Cons. Module	126	TCTX	TCTX brain	48	15-101	Infin 27k	Gibbs 2010	GSE15745
9	Cons. Module	123	PONS	PONS brain	46	15-101	Infin 27k	Gibbs 2010	GSE15745
10	Cons. Module	111	CRBLM	CRBLM brain	47	16-96	Infin 27k	Gibbs 2010	GSE15745
11	Validation	94	WB 450k	controls and SZ	32	18-65	Illumina 450	novel data	GSE41169
12	Validation	24	MSC	MSC cells	50	21-85	Infin 27k	Schellenberg 2(	GSE26519+GSE17448
13	Validation	50	CD14+CD4+	CD4+ T-cells and CD14+ monocytes	36	16-69	Infin 27k	Rakyan 2010	GSE20242
14	Validation	398	leukocyte	pediatric population	10	3-17	Infin 27k	Alisch et al 201	GSE27097
15	Validation	72	leukocyte	healthy children	5	1-16	Illumina 450	Alisch et al 201	GSE36064
16	Validation	108	Prefr. Cortex	healthy controls	26	-0.5-84	Infin 27k	Numata 2012	BrainCloudMethyl

(WB) Whole blood, FCTX (Frontal Cortex), TCTX (Temporal Cortex), CRBLM (Cerebellum), MSC (mesenchymal stromal cells)

How does one find “consensus” module based on multiple networks?

1. Consensus adjacency is a quantile of the input
   e.g. minimum, lower quartile, median

$$Consensus_q(A) = pquantile_q(A[[1]], \ldots, A[[no.networks]])$$
$$= pmin(A[[1]], \ldots, A[[no.networks]])$$

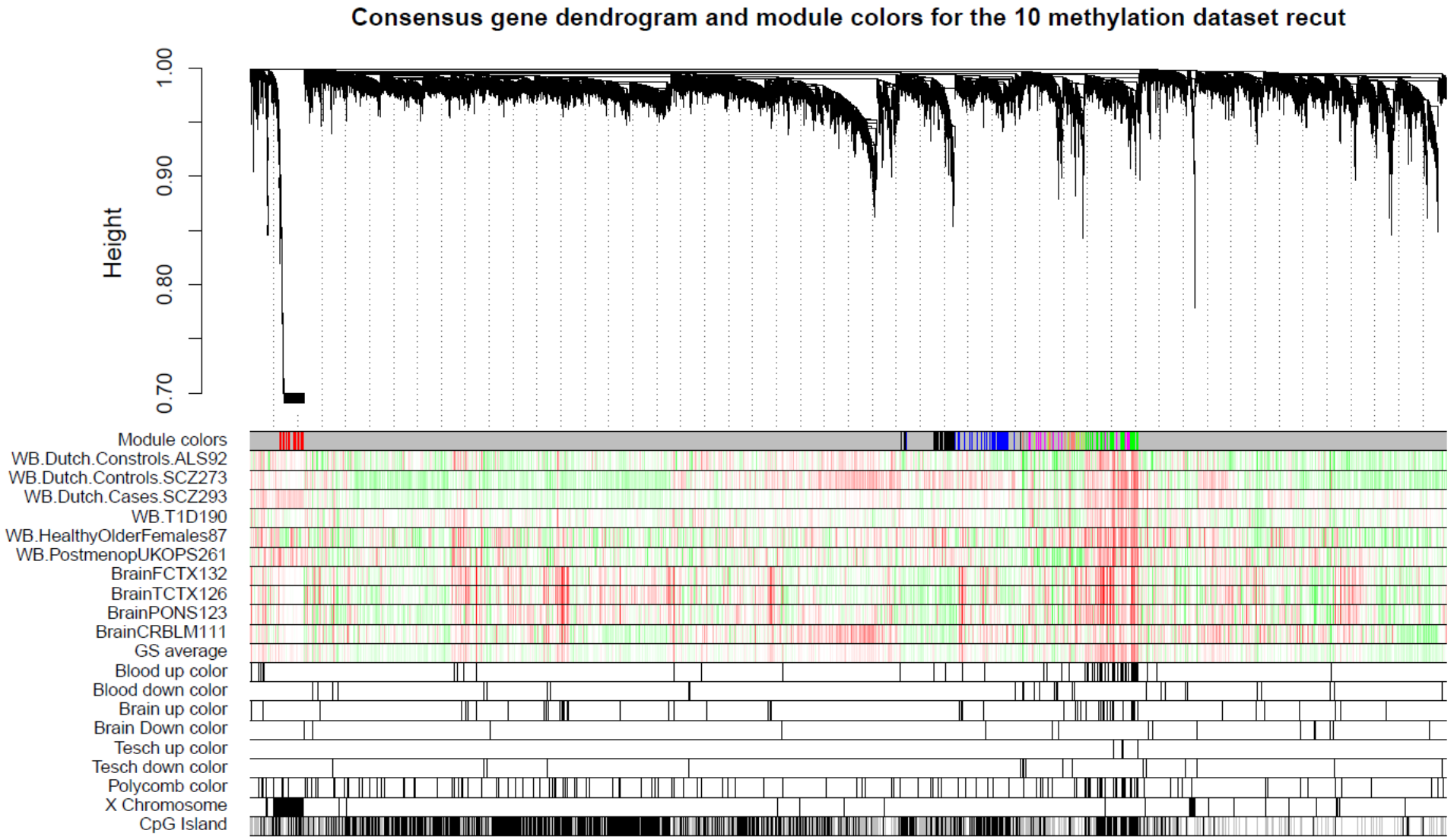
2. Apply usual module detection algorithm

Analysis steps of WGCNA

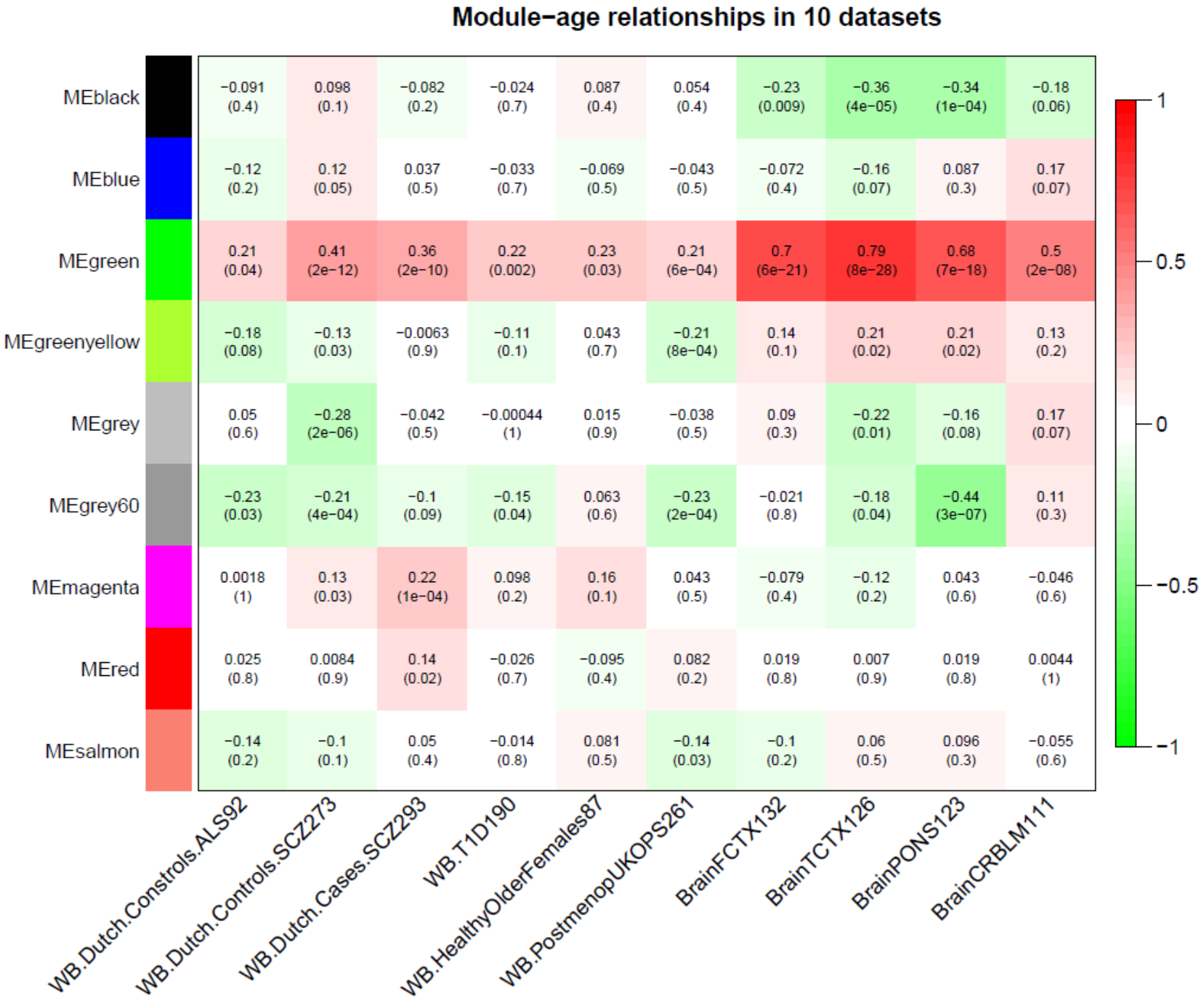
1. Construct a signed weighted correlation network based on 10 DNA methylation data sets (Illumina 27k)
Purpose: keep track of co-methylation relationships

2. Identify consensus modules
Purpose: find robustly defined and reproducible modules

3. Relate modules to external information
Age
Gene Information: gene ontology, cell marker genes
Purpose: find biologically interesting age related modules



Message: green module contains probes positively correlated with age

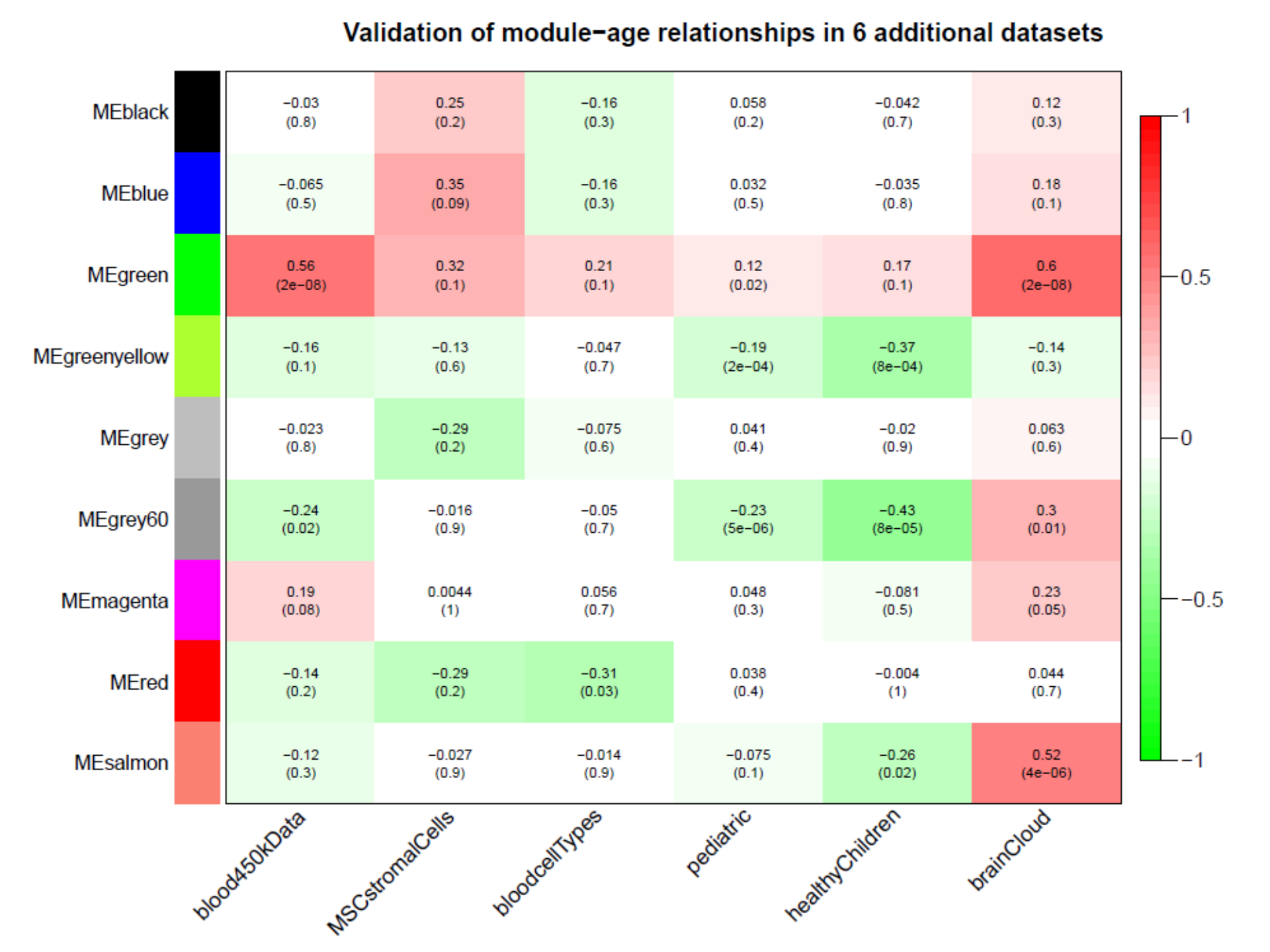

Module−age relationships in 10 datasets
MEblack
MEblue
MEgreen
MEgreenyellow
MEgrey
MEgrey60
MEmagenta
MEred
MEsalmon
−0.091 (0.4)
0.098 (0.1)
−0.082 (0.2)
−0.024 (0.7)
0.087 (0.4)
0.054 (0.4)
−0.23 (0.009)
−0.36 (4e−05)
−0.34 (1e−04)
−0.18 (0.06)
−0.12 (0.2)
0.12 (0.05)
0.037 (0.5)
−0.033 (0.7)
−0.069 (0.5)
−0.043 (0.5)
−0.072 (0.4)
−0.16 (0.07)
0.087 (0.3)
0.17 (0.07)
0.21 (0.04)
0.41 (2e−12)
0.36 (2e−10)
0.22 (0.002)
0.23 (0.03)
0.21 (6e−04)
0.7 (6e−21)
0.79 (8e−28)
0.68 (7e−18)
0.5 (2e−08)
−0.18 (0.08)
−0.13 (0.03)
−0.0063 (0.9)
−0.11 (0.1)
0.043 (0.7)
−0.21 (8e−04)
0.14 (0.1)
0.21 (0.02)
0.21 (0.02)
0.13 (0.2)
0.05 (0.6)
−0.28 (2e−06)
−0.042 (0.5)
−0.00044 (1)
0.015 (0.9)
−0.038 (0.5)
0.09 (0.3)
−0.22 (0.01)
−0.16 (0.08)
0.17 (0.07)
−0.23 (0.03)
−0.21 (4e−04)
−0.1 (0.09)
−0.15 (0.04)
0.063 (0.6)
−0.23 (2e−04)
−0.021 (0.8)
−0.18 (0.04)
−0.44 (3e−07)
0.11 (0.3)
0.0018 (1)
0.13 (0.03)
0.22 (1e−04)
0.098 (0.2)
0.16 (0.1)
0.043 (0.5)
−0.079 (0.4)
−0.12 (0.2)
0.043 (0.6)
−0.046 (0.6)
0.025 (0.8)
0.0084 (0.9)
0.14 (0.02)
−0.026 (0.7)
−0.095 (0.4)
0.082 (0.2)
0.019 (0.8)
0.007 (0.9)
0.019 (0.8)
0.0044 (1)
−0.14 (0.2)
−0.1 (0.1)
0.05 (0.4)
−0.014 (0.8)
0.081 (0.5)
−0.14 (0.03)
−0.1 (0.2)
0.06 (0.5)
0.096 (0.3)
−0.055 (0.6)
WB.Dutch.Controls.ALS92
WB.Dutch.Controls.SCZ273
WB.Dutch.Cases.SCZ293
WB.T1D190
WB.HealthyOlderFemales87
WB.PostmenopUKOPS261
BrainFCTX132
BrainTCTX126
BrainPONS123
BrainCRBLM111
1
0.5
0
−0.5
−1

Age relations in brain regions

The green module eigengene is highly correlated with age in

i) Frontal cortex (cor=.70)
ii) Temporal cortex (cor=.79)
iii) Pons (cor=.68)

But less so in cerebellum (cor=.50).


Validation of module−age relationships in 6 additional datasets
MEblack
MEblue
MEgreen
MEgreenyellow
MEgrey
MEgrey60
MEmagenta
MEred
MEsalmon
−0.03 (0.8)
0.25 (0.2)
−0.16 (0.3)
0.058 (0.2)
−0.042 (0.7)
0.12 (0.3)
−0.065 (0.5)
0.35 (0.09)
−0.16 (0.3)
0.032 (0.5)
−0.035 (0.8)
0.18 (0.1)
0.56 (2e−08)
0.32 (0.1)
0.21 (0.1)
0.12 (0.02)
0.17 (0.1)
0.6 (2e−08)
−0.16 (0.1)
−0.13 (0.6)
−0.047 (0.7)
−0.19 (2e−04)
−0.37 (8e−04)
−0.14 (0.3)
−0.023 (0.8)
−0.29 (0.2)
−0.075 (0.6)
0.041 (0.4)
−0.02 (0.9)
0.063 (0.6)
−0.24 (0.02)
−0.016 (0.9)
−0.05 (0.7)
−0.23 (5e−06)
−0.43 (8e−05)
0.3 (0.01)
0.19 (0.08)
0.0044 (1)
0.056 (0.7)
0.048 (0.3)
−0.081 (0.5)
0.23 (0.05)
−0.14 (0.2)
−0.29 (0.2)
−0.31 (0.03)
0.038 (0.4)
−0.004 (1)
0.044 (0.7)
−0.12 (0.3)
−0.027 (0.9)
−0.014 (0.9)
−0.075 (0.1)
−0.26 (0.02)
0.52 (4e−06)
blood450kData
MSCstromalCells
bloodcellTypes
pediatric
healthyChildren
brainCloud
1
0.5
0
−0.5
−1

Gene ontology enrichment analysis of the green aging module

- Highly significant enrichment in multiple terms related to cell differentiation, development and brain function
  - neuron differentiation (p=8.5E-26)
  - neuron development (p=9.6E-17)
  - DNA-binding (p=2.3E-21).
  - SP PIR keyword "developmental protein" (p-value 8.9E-37)

Polycomb-group proteins

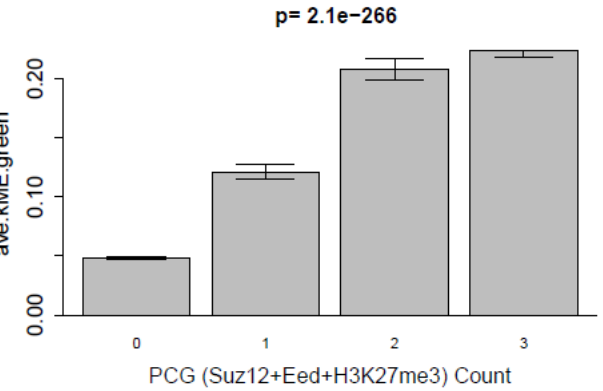
Polycomb group gene expression is important in many aspects of development.

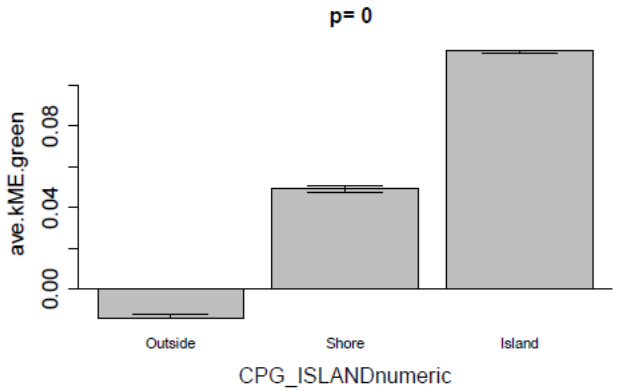
Genes that are hypermethylated with age are known to be significantly enriched with Polycomb group target genes (Teschendorff et al 2010)

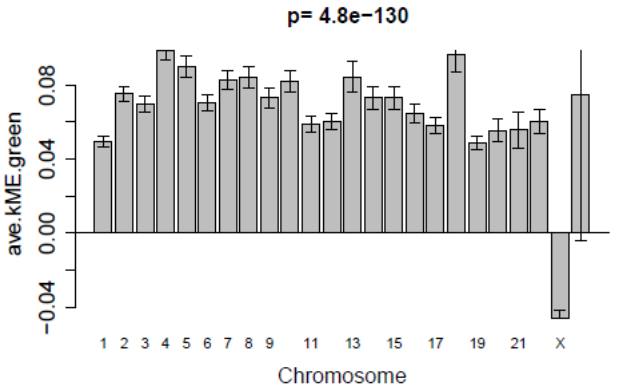
This insight allows us to compare different gene selection strategies.

The higher the enrichment with respect to PCGT genes the more signal is in the data.

Relating module membership (ave. kME.green) to sequence properties







Analysis of variance shows relative contribution:

Source of Variation		ave.kME.green, Total Prop Var Explained=15.8%			
Source	Degrees of Freedom	Sums of Sq	Prop. of Total Var	F statistic	p-value (F-test)
PCG (Suz12+Eed+H3K27me3) OccupancyCount	1	49.35	0.071	2013.0	< 2.2E-16
CPG_Island	2	50.78	0.073	1035.7	< 2.2E-16
X chromosome	1	9.74	0.014	397.4	< 2.2E-16

CpGs that get hypermethylated with age tend to be
- inside CpG islands
- targets of PCGs
- located on autosomes

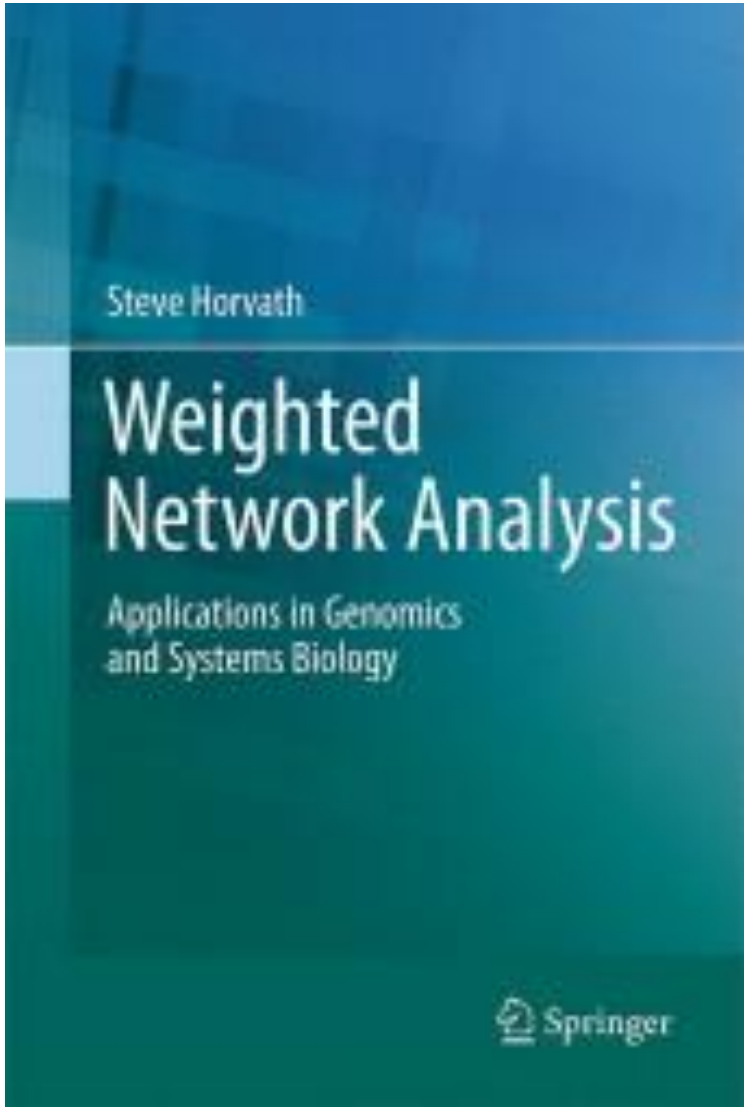
Discussion of aging study

- We confirm the findings of many others
  - age has a profound effects on thousands of methylation probes
- Consensus module based analysis leads to biologically more meaningful results than those of a standard marginal meta analysis
- We used a signed correlation network since it is important to keep track of the sign of the co-methylation relationship
- We used a weighted network b/c
  - it allows one to calibrate the networks for consensus module analysis
  - module preservation statistics are needed to validate the existence of the modules in other data

Implementation and R software tutorials, WGCNA R library

- General information on weighted correlation networks
- Google search
  - “WGCNA”
  - “weighted gene co-expression network”
  - R package: WGCNA
  - R package: dynamicTreeCut
- R function `modulePreservation` is part of WGCNA package

Book on weighted networks



E-book is often freely accessible if your library has a subscription to Springer books

Webpages where the tutorials and ppt slides can be found

- http://www.genetics.ucla.edu/labs/horvath/CoexpressionNetwork/WORKSHOP/
- R software tutorials from S. H, see corrected tutorial for chapter 12 at the following link:

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

Acknowledgement

Students and Postdocs:

- **Peter Langfelder** is first author on many related articles
- Jason Aten, Chaochao (Ricky) Cai, Jun Dong, Tova Fuller, Ai Li, Wen Lin, Michael Mason, Jeremy Miller, Mike Oldham, Anja Presson, Lin Song, Kellen Winden, Yafeng Zhang, Andy Yip, Bin Zhang

- Colleagues/Collaborators
- Neuroscience: Dan Geschwind, Giovanni Coppola
- Methylation: Roel Ophoff
- Mouse: Jake Lusis, Tom Drake