Is my network module preserved and reproducible?

PloS Comp Biol. 7(1): e1001057.

Steve Horvath
Peter Langfelder
University of California, Los Angeles

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.

Review of *some* fundamental network concepts

BMC Systems Biology 2007, 1:24
PLoS Comput Biol 4(8): e1000117

Network concepts are also known as network statistics or network indices

- Network concepts underlie network language and systems biological modeling.
- Abstract definition: function of the adjacency matrix

Connectivity

- 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

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

Hub-nodes: nodes with the largest connectivities

Density

- Density= mean adjacency
- Highly related to mean connectivity

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

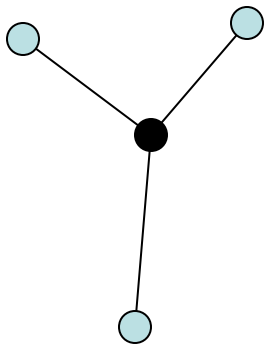
where n is the number of network nodes.

Clustering Coefficient

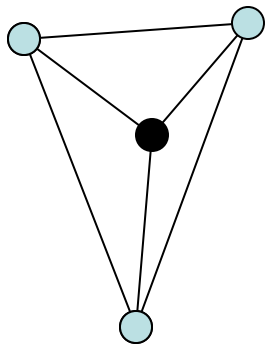
Measures the cliquishness of a particular node
« A node is cliquish if its neighbors know each other »

$$ClusterCoef_i = \frac{\sum_{l \neq i} \sum_{m \neq i,l} a_{il} a_{lm} a_{mi}}{\left(\sum_{l \neq i} a_{il}\right)^2 - \sum_{l \neq i} a_{il}^2}$$

This generalizes directly to weighted networks (Zhang and Horvath 2005)



Clustering Coef of the black node = 0



Clustering Coef = 1

Network module

- Abstract definition of module=a 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)

Modules versus clusters

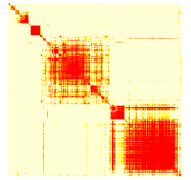
- In general, modules are different from clusters (e.g. KEGG pathways may not correspond to clusters in the network).
- But a cluster is a special case of a module
- In general, studying module preservation is different from studying cluster preservation.
- However, many module preservation statistics lend themselves as powerful cluster preservation statistics
- A limited comparison of module and cluster preservation statistics is provided in the article (for the special case when modules co-incide with clusters).

Module preservation is often an essential step in a network analysis

The following slide provides an overview of many network analyses. Adapted from weighted gene co-expression network analysisWGCNA.

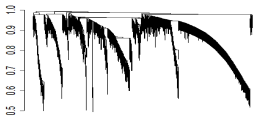
Construct a network

Rationale: make use of interaction patterns between genes



Identify modules

Rationale: module (pathway) based analysis

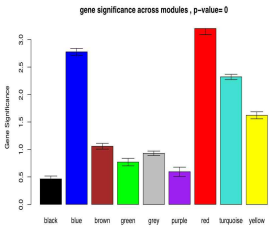


Relate modules to external information

Array Information: Clinical data, SNPs, proteomics

Gene Information: gene ontology, EASE, IPA

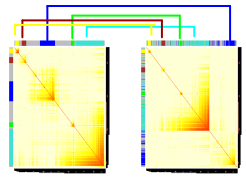
Rationale: find biologically interesting modules



Study Module Preservation across different data

Rationale:

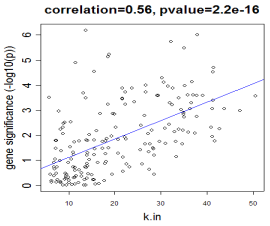
- Same data: to check robustness of module definition
- Different data: to find interesting modules



Find the key drivers in *interesting* modules

Tools: intramodular connectivity, causality testing

Rationale: experimental validation, therapeutics, biomarkers



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)
  - For example: could study differences in large-scale organization of co-expression networks between disease states, genders, related species, ...

Network based module preservation statistics use network concepts for measuring network connectivity preservation

- Quantify whether modules defined in a reference network remain “good” modules in the test network
- Module definition in the test network is not necessary
- A multitude of network concepts can be used to describe the preservation of connectivity patterns
  - Examples:connectivity, clustering coefficient, density

Multiple connectivity preservation statistics

For general networks, i.e. input adjacency matrices

- *cor.kIM=correlation of intramodular connectivity across module nodes*
- *cor.ADJ=correlation of adjacency across module nodes*
- cor.kIM=correlations of intramodular connectivity

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

- cor.cor=Correlations of correlations.
- cor.kME= correlations of eigengene-based connectivity *kME*

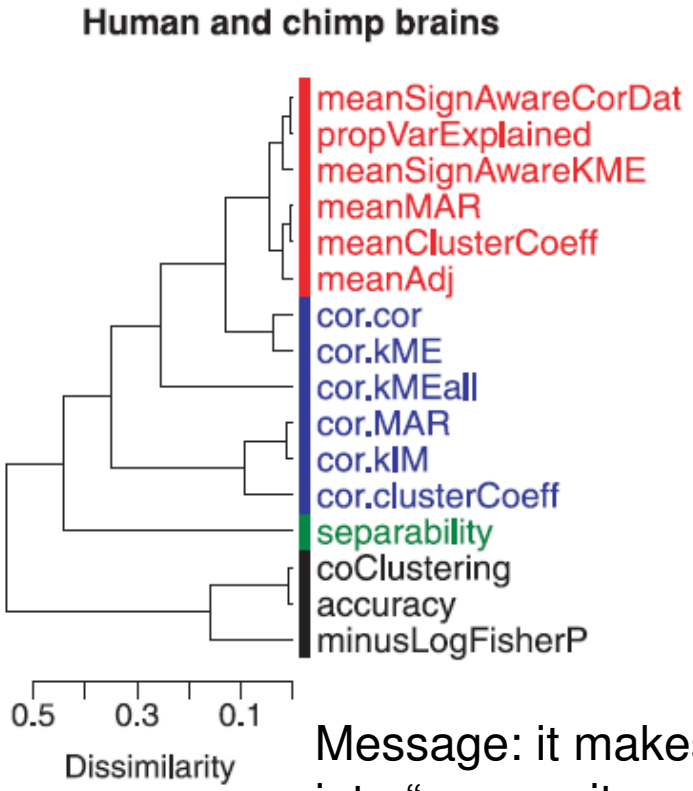
Table 1. Overview of module preservation statistics. Details are provided below and in the paper…

No.	Preservation Statistic			Network	Ref. netw. input			Test netw. input			Used in composite		
	Name	Eq.	Type		Lbl	Adj	*datX*	Lbl	Adj	*datX*	*Zsum.*	*medR.*	*Zsum.A*
1	coClustering	Supp.	Cross-tab	not used	yes	no	no	yes	no	no	no	no	no
2	*accuracy*	Supp.	Cross-tab	not used	yes	no	no	yes	no	no	no	no	no
3	−log(p-value)	Supp.	Cross-tab	not used	yes	no	no	yes	no	no	no	no	no
4	$meanAdj$	8	Density	general	yes	no	no	no	yes	no	no	no	yes
5	$meanClCoef$	9	Density	general	yes	no	no	no	yes	no	no	no	no
6	$meanMAR$	10	Density	general	yes	no	no	no	yes	no	no	no	no
7	$cor.Adj$	11	Connect.	general	yes	yes	no	no	yes	no	yes	yes	yes
8	$corkIM$	12	Connect.	general	yes	yes	no	no	yes	no	yes	yes	yes
9	$cor.ClCoef$	13	Connect.	general	yes	yes	no	no	yes	no	no	no	no
10	$cor.MAR$	14	Connect.	general	yes	yes	no	no	yes	no	no	no	no
11	$separability^{ave}$	27	Separab.	general	yes	yes	no	no	yes	no	no	no	no
12	$meanCor$	19	Den.+Con.	cor	yes	no	yes	no	no	yes	yes	yes	no
13	$cor.cor$	20	Connect.	cor	yes	no	yes	no	no	yes	yes	yes	no
14	$propVarExpl$	21	Density	cor	yes	no	yes	no	no	yes	yes	yes	no
15	$meanKME$	22	Den.+Con.	cor	yes	no	yes	no	no	yes	yes	yes	no
16	$corkME$	23	Connect.	cor	yes	no	yes	no	no	yes	yes	yes	no
17	$corkMEall$	24	Connect.	cor	yes	no	yes	no	no	yes	no	no	no
18	$separability$	28	Separab.	cor	yes	no	yes	no	no	yes	no	no	no
19	$Z_{summary}$	1	Compos.	cor	yes	yes	yes	no	yes	yes			
20	$p_{summary}$		Compos.	cor	yes	yes	yes	no	yes	yes			
21	$medianRank$	34	Compos.	cor	yes	yes	yes	no	yes	yes			
22	$Z_{summaryADJ}$	35	Compos.	general	yes	yes	no	no	yes	no			

The columns report the names, types, and input of individual preservation statistics (Lbl, module label; Adj, general network adjacency; $datX$, numeric data from which a correlation network is constructed). The last 3 columns indicate which of the individual statistics are used in the composite summary statistics $Z_{summary}$, $medianRank$, and $Z_{summaryADJ}$, respectively. The definition of cross-tabulation based statistics can be found in Supplementary Text S1.

Module preservation statistics are often closely related

Clustering module preservation statistics based on correlations across modules



Red=density statistics

Blue: connectivity statistics
Green: separability statistics

Cross-tabulation based statistics

Message: it makes sense to aggregate the statistics into “composite preservation statistics”.

How to define threshold values of network concepts to consider a module "good"?

- We have 4 density and 4 connectivity preservation measures defined such that their values lie between 0 and 1
- However, thresholds will vary depending on many factors (number of genes/probesets, number of samples, biology, expression platform, etc.)
- We determine baseline values by permutation and calculate Z scores

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

Judging modules by their Z scores

- For each 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

Some math equations

Definition of vectorized matrix

$$v.cor^{[test](q)} = vectorizeMatrix(cor^{[test](q)})$$

Connectivity based statistics for measuring correlation preservation

$$cor.cor^{(q)} = cor(v.cor^{[ref](q)}, v.cor^{[test](q)})$$

Density+connectivity based statistics

$$mean.cor^{(q)} = mean(sign(v.cor^{[ref](q)}) * v.cor^{[test](q)})$$

Permutation test allows one to estimate Z version of each statistic

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

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

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

Summary of the methodology

- We take module definitions from a reference network and apply them to a test network
- We ask two basic question:
  - 1. Density: are the modules (as groups of genes) denser than background?
  - 2. Preservation of connectivity: Is hub gene status preserved between reference and test networks?
- We judge modules mostly by how different they are from background (random samples of genes) as measured by the permutation Z score

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

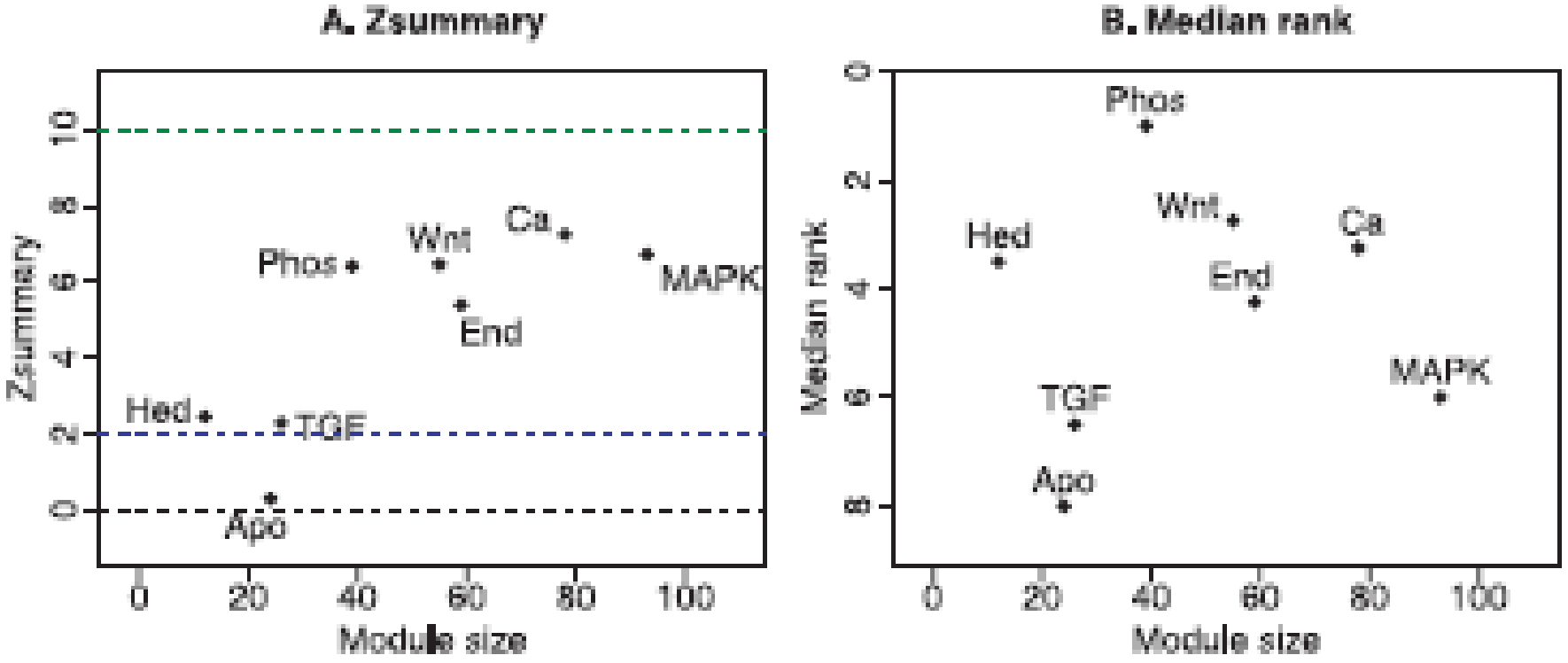
Application: Modules defined as KEGG pathways.

Connectivity patterns (adjacency matrix) is defined as signed weighted co-expression network.

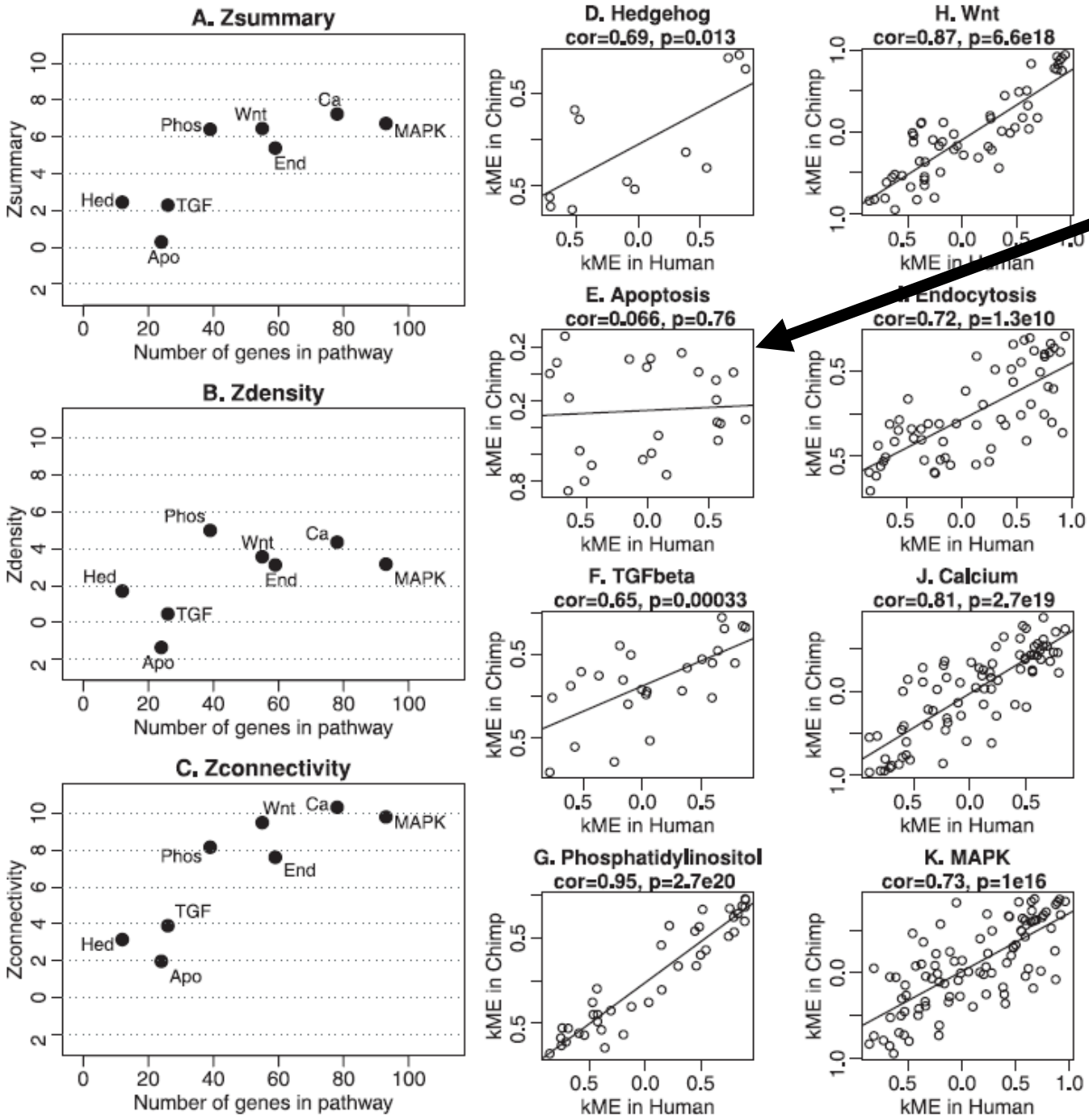
Comparison of human brain (reference) versus chimp brain (test) gene expression data.

Preservation of KEGG pathways measured using the composite preservation statistics Zsummary and medianRank

- Humans versus chimp brain co-expression modules

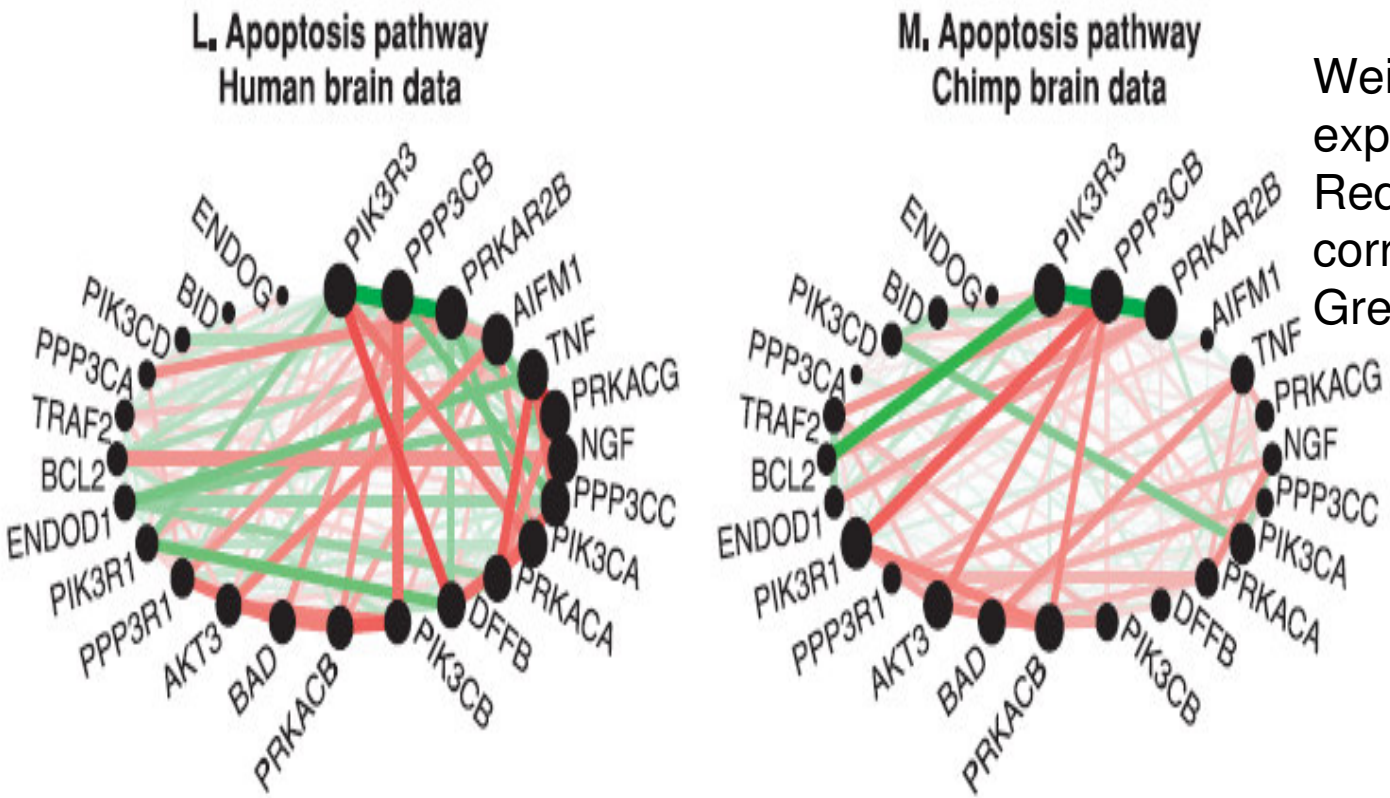


Apoptosis module is least preserved
according to both composite preservation statistics



Apoptosis
module
has low value
of cor.kME=0.066

Visually inspect connectivity patterns of the apoptosis module in humans and chimpanzees



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

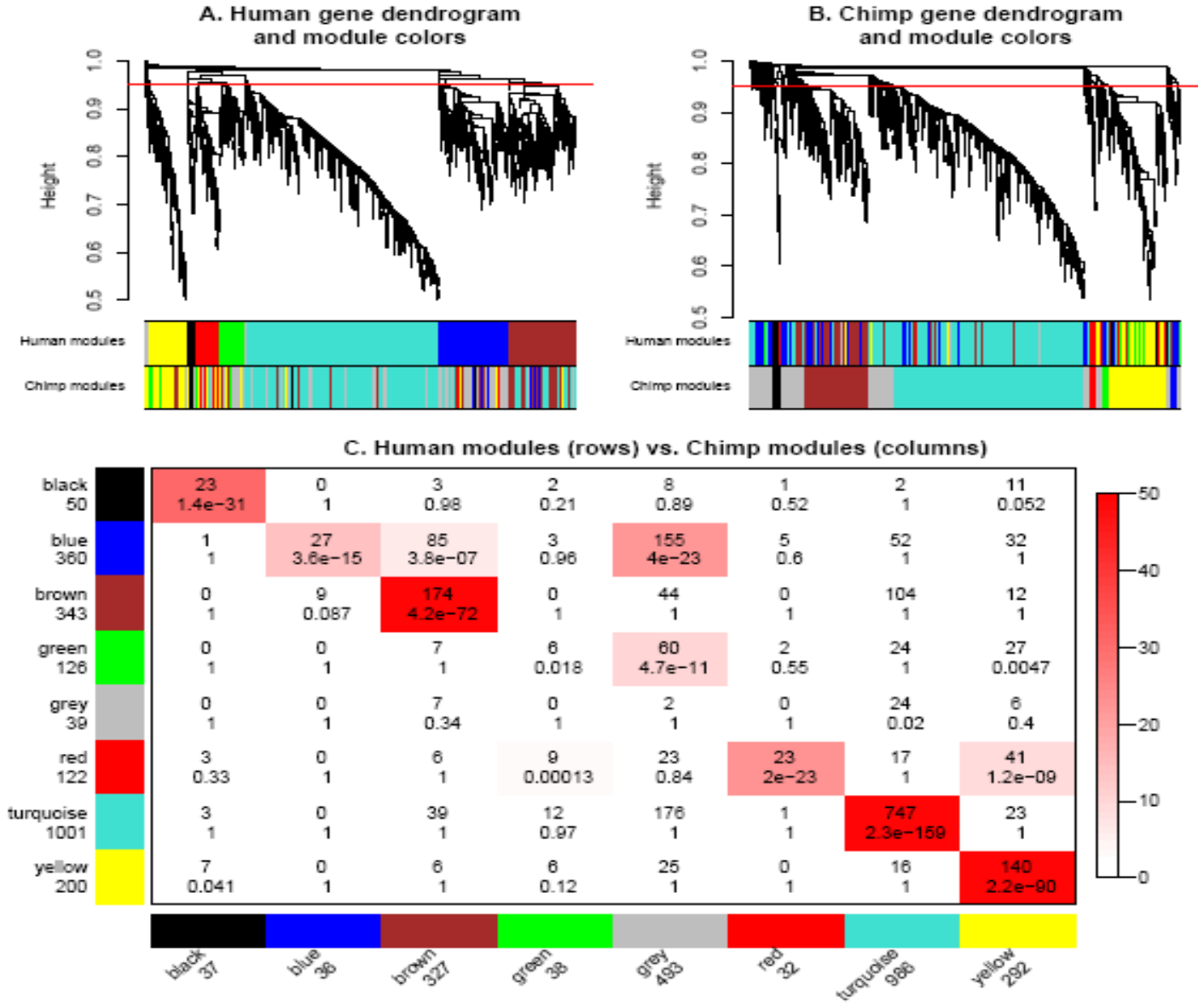
Note that the connectivity patterns look very different.
Preservation statistics are ideally suited to measure differences in connectivity preservation

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

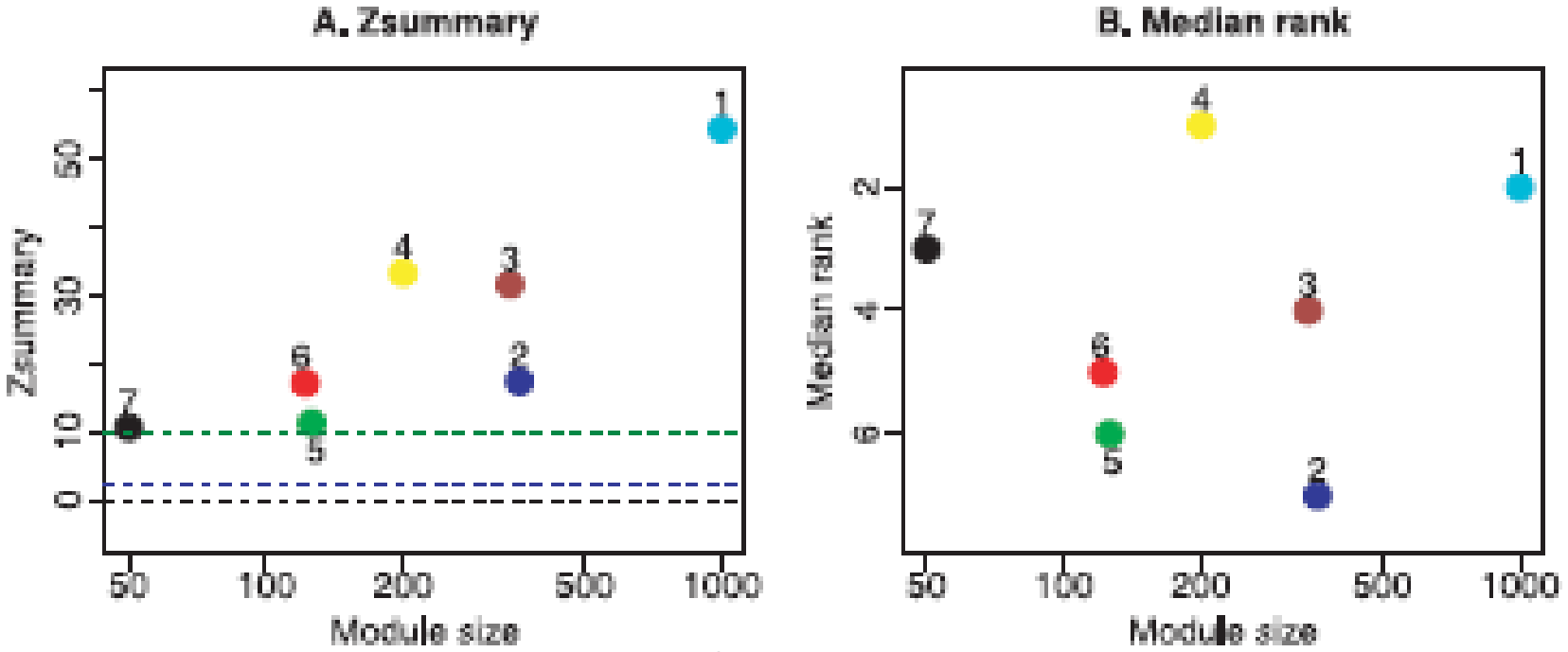
Modules defined as clusters
(branches of a cluster tree)

Data from Oldam et al 2006

Preservation of modules between human and chimpanzee brain networks



2 composite preservation statistics



Zsummary is above the threshold of 10 (green dashed line), i.e. all modules are preserved.

Zsummary often shows a dependence on module size which may or may not be attractive (discussion in paper)

In contrast, the median rank statistic is not dependent on module size.

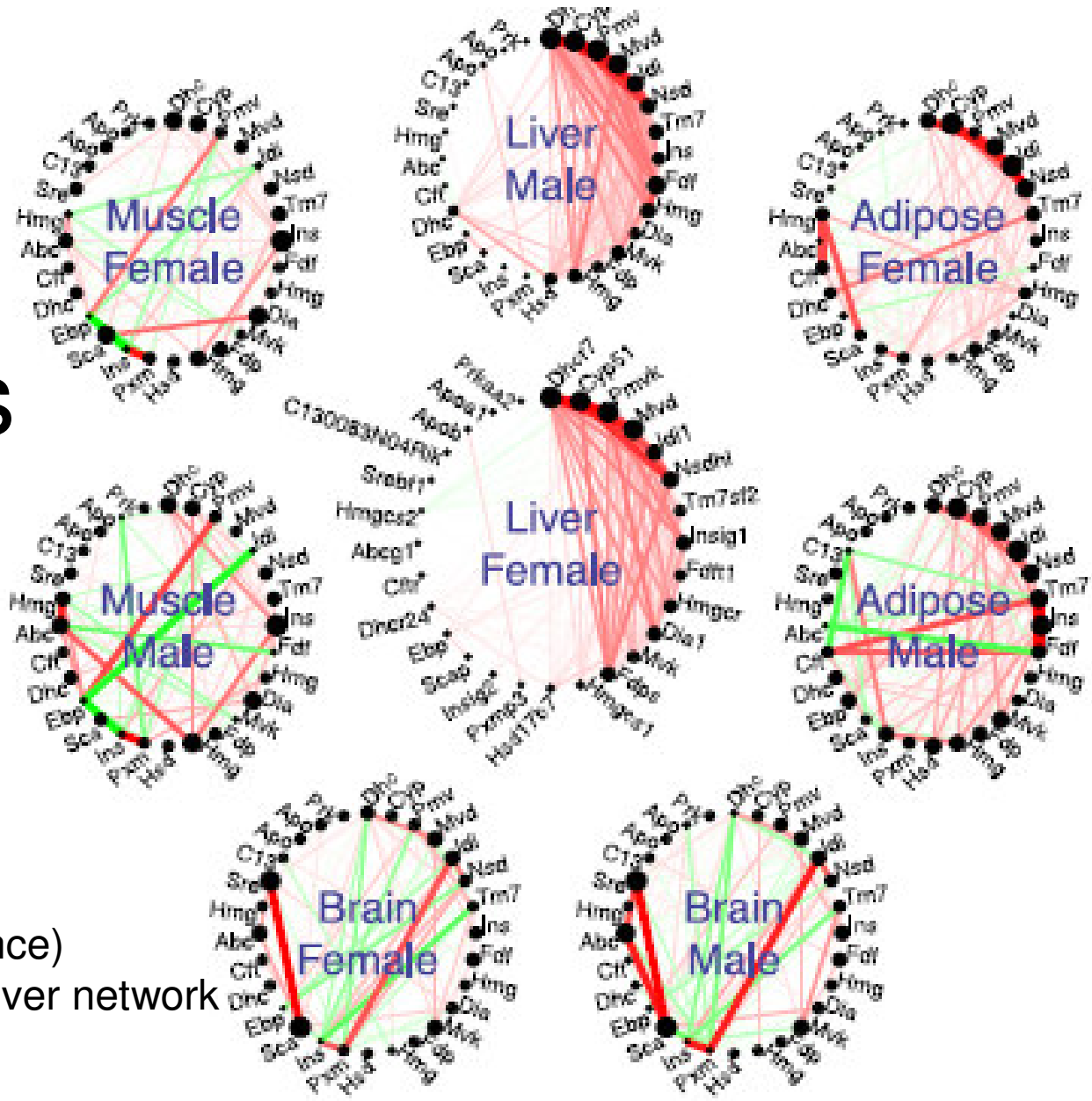
It indicates that the yellow module is most preserved

Application: Studying the preservation of a female mouse liver module in different tissue/gender combinations.

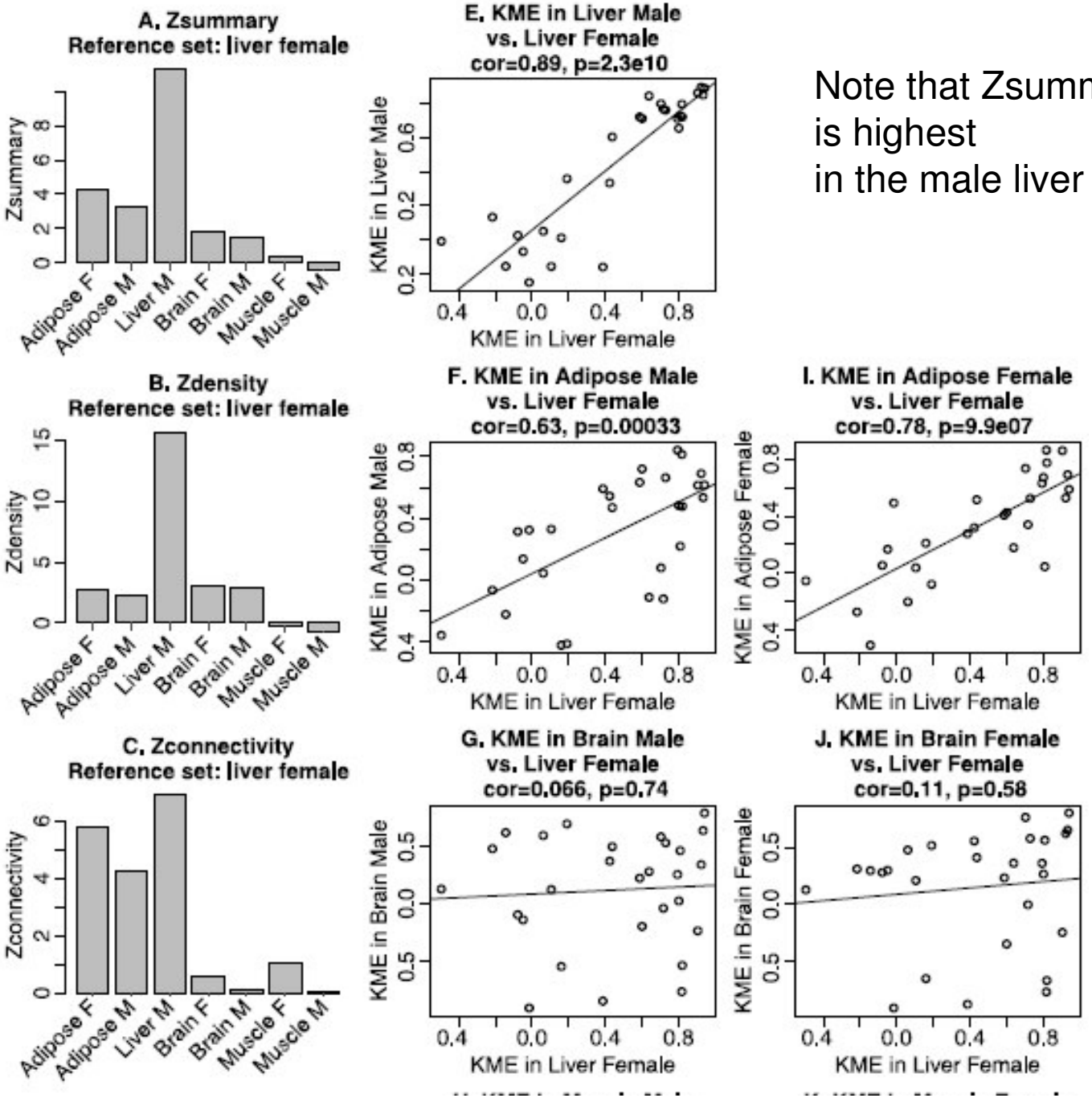
Module: genes of cholesterol biosynthesis pathway
Network: signed weighted co-expression network
Reference set: female mouse liver
Test sets: other tissue/gender combinations

Data provided by Jake Lusis

Network of cholesterol biosynthesis genes



Message:
female liver network (reference)
Looks most similar to male liver network



Note that Zsummary
is highest
in the male liver network

Implementation

- Function modulePreservation is part of WGCNA R package

  http://www.genetics.ucla.edu/labs/horvath/CoexpressionNetwork/Rpackages/WGCNA

- Tutorials: example study of module preservation between female and male liver samples, and preservation between human and chimp brains, at

www.genetics.ucla.edu/labs/horvath/CoexpressionNetwork/ModulePreservation

General information on weighted correlation networks
Google search
"WGCNA"
"weighted gene co-expression network"

Input for the R: function `modulePreservation`

- reference data set in which modules have been defined
  - either raw data datExpr.ref or adjacency matrix A.ref
- module assignments in reference data
- test data set:
  - either datExpr.test or adjacency matrix A.test
  - No need for test set module assignment

Standard cross-tabulation based approach for comparing preservation of modules

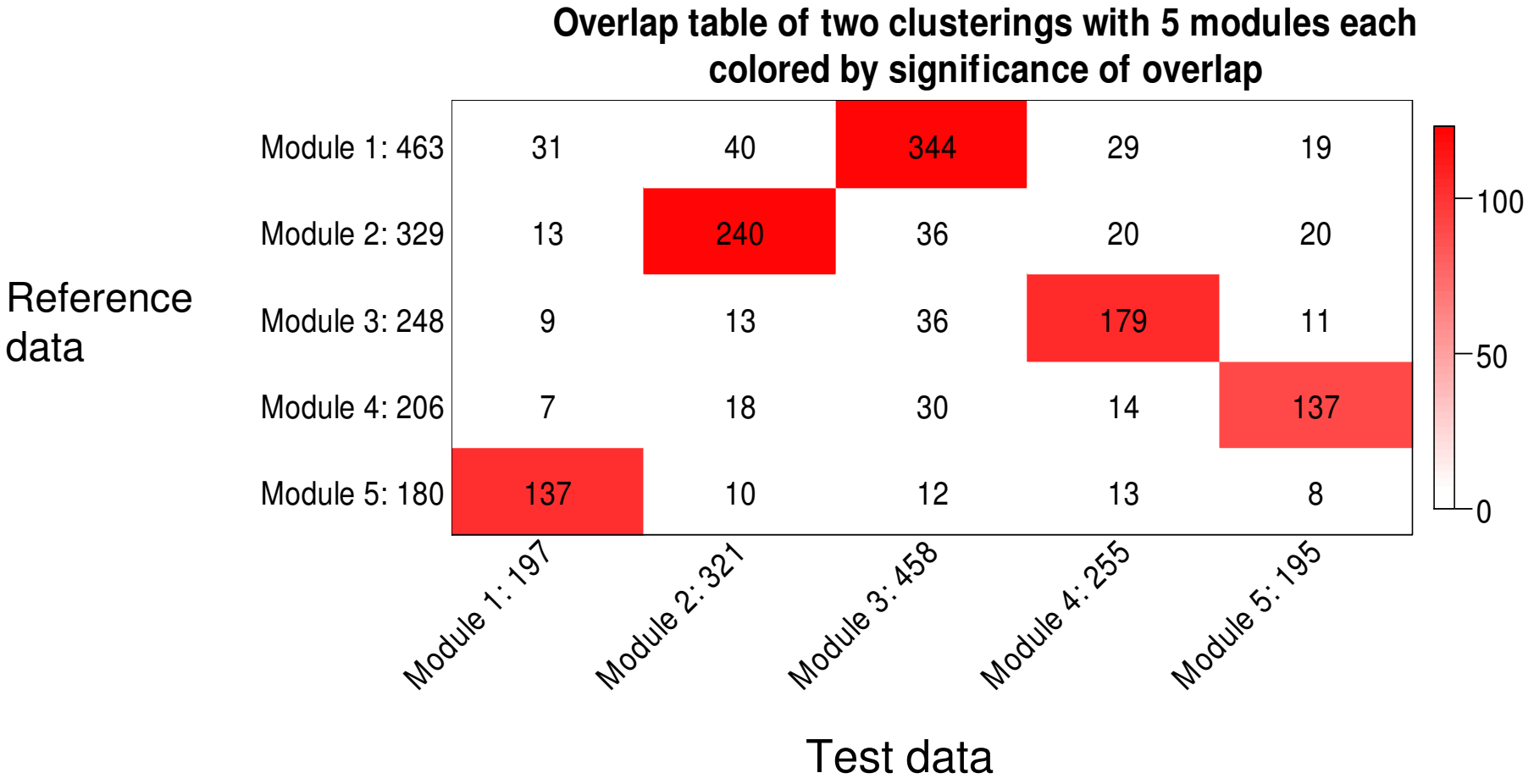
- Applicable when a module detection algorithm was applied to the reference data
- STEPS

  1) Apply the same module detection algorithm to the test data as well

  2) Compare the module labels in the reference and the test data using cross-tabulation

  3) Measure whether the overlap of module labels is significant (e.g. Pearson's chi-square test for contingency tables)

Cross-tabulation table for comparing reference modules to test modules



- Note that the module labels from the reference data don't have to correspond to the labels in the test data

Problems with the standard cross-tabulation based approach

- Requires that module labels are defined in the test data set
- Only useful if a module detection procedure is used to define modules.
- Cross-tabulation statistics are ill-suited for arguing that a reference module is *not* preserved
  - since slightly different parameter choices of the module detection procedure may result in a new module in the test network that overlaps with the original reference module.
- Cross-tabulation based approaches ignore the connectivity pattern among the nodes that form the module. They fail to measure connectivity preservation.

Discussion

- Standard cross-tabulation based statistics are intuitive
  - Disadvantages: i) only applicable for modules defined via a module detection procedure, ii) ill suited for ruling out module preservation
- 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)

Acknowledgement

- Co-authors

  Peter Langfelder, Rui Luo, Mike C Oldham
- Mouse data by A. J. Lusis
- Module preservation applications: Chaochao Cai, Lin Song, Tova Fuller, Jeremy Miller, Dan Geschwind, Roel Ophoff