Applications of weighted gene coexpression analysis

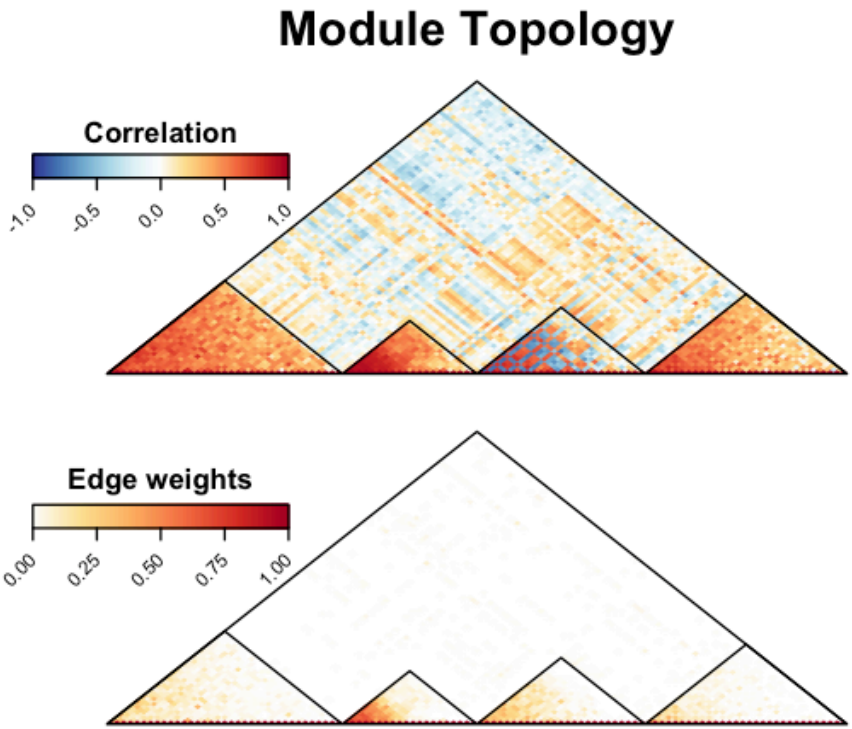
Michael Inouye
Baker Heart and Diabetes Institute
Univ of Melbourne / Monash Univ

Summer Institute in Statistical Genetics 2017
Integrative Genomics Module
Seattle

@minouye271
www.inouyelab.org

Gene co-expression networks

- **Weighted, undirected complete gene network**
  - **Nodes**: genes/probes
  - **Edges**: $|cor(node\_i, node\_j)|^{\gamma}$
    - Scale-free assumption and [0,1]
- **Identify subnets (modules/clusters)**
  - Typically subnets represent known biological pathways
  - Various methods and tools for clustering



Strategies for testing association of a subnet with a phenotype

- **Univariate**
  - For each subnet gene, perform a test
- **Eigenvector**
  - Calculate 1st principal component
  - With vector of PC1 sample loadings, perform a test
- **Multivariate**
  - Simultaneously test for association of phenotype with all genes
  - Example: Canonical correlation analysis (CCA)
- **Considerations**
  - Multiple testing burden
  - Sensitivity and specificity

Interpretation of subnets

- **Pathway analysis and gene set statistics**
- **If subnet is small enough, manual interpretation is possible (with proper literature support)**
- **Correlation vs Causation**
  - Confounding, causality and reactivity
    - It is more useful (and more difficult) to know the underlying structure of relationships b/n genes than *clusters* of co-regulation
  - How can causality be tested?
    - Perturbation techniques
    - Mendelian randomisation (genetic variation has a special role in determining causality)

OPEN ACCESS Freely available online PLOS GENETICS

An Immune Response Network Associated with Blood Lipid Levels

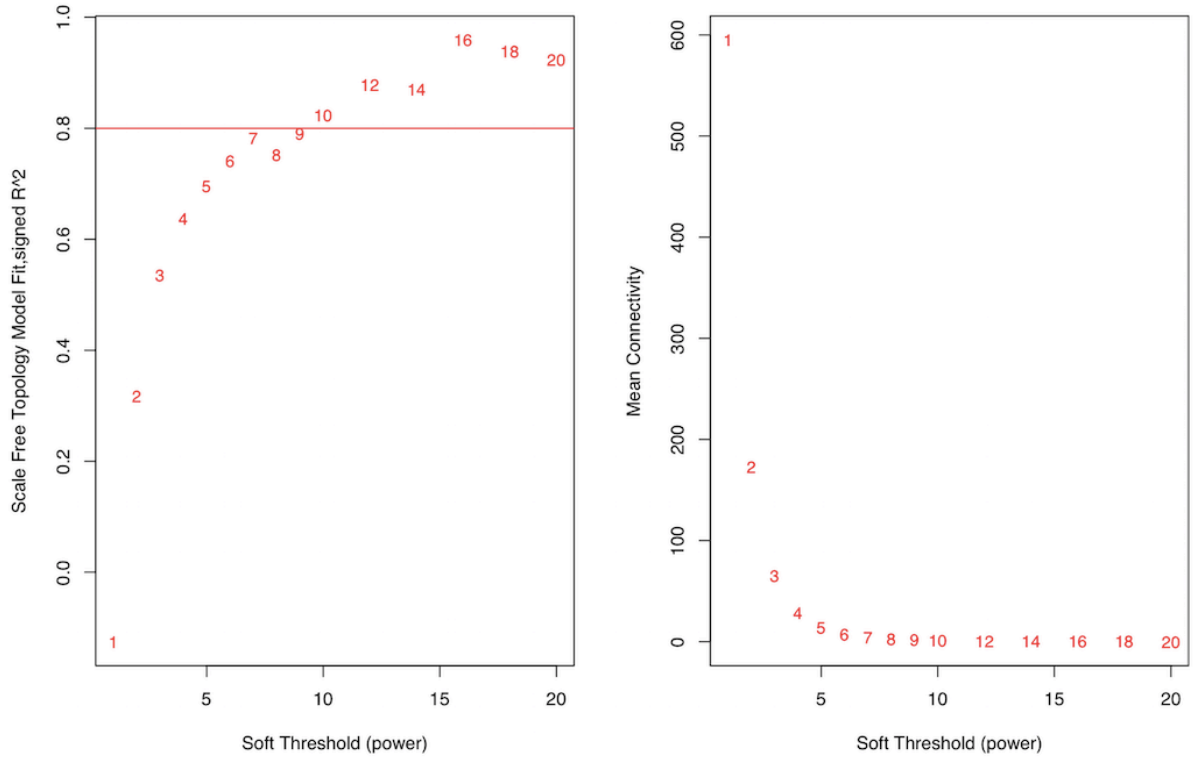
Michael Inouye[1,2¤*], Kaisa Silander[3], Eija Hamalainen[1], Veikko Salomaa[4], Kennet Harald[4], Pekka Jousilahti[4], Satu Männistö[4], Johan G. Eriksson[4,5,6,7,8], Janna Saarela[3,9], Samuli Ripatti[3], Markus Perola[3], Gert-Jan B. van Ommen[2], Marja-Riitta Taskinen[10], Aarno Palotie[1,3,11,12], Emmanouil T. Dermitzakis[1,13], Leena Peltonen[1,3,11†]

518 randomly sampled individuals

Fasting whole blood

Transcriptome

Selection of soft power threshold for adjacency matrix



Better differentiate strong vs weak correlations

Approximate scale-free network topology (signed $R^2 > 0.80$) but maximize connectivity

Detect gene modules

- Goal: Get the most coherent gene subnetworks as possible
- Instead of using the correlation-based edges, WGCNA is calculating a distance measure called topological similarity (TOM):

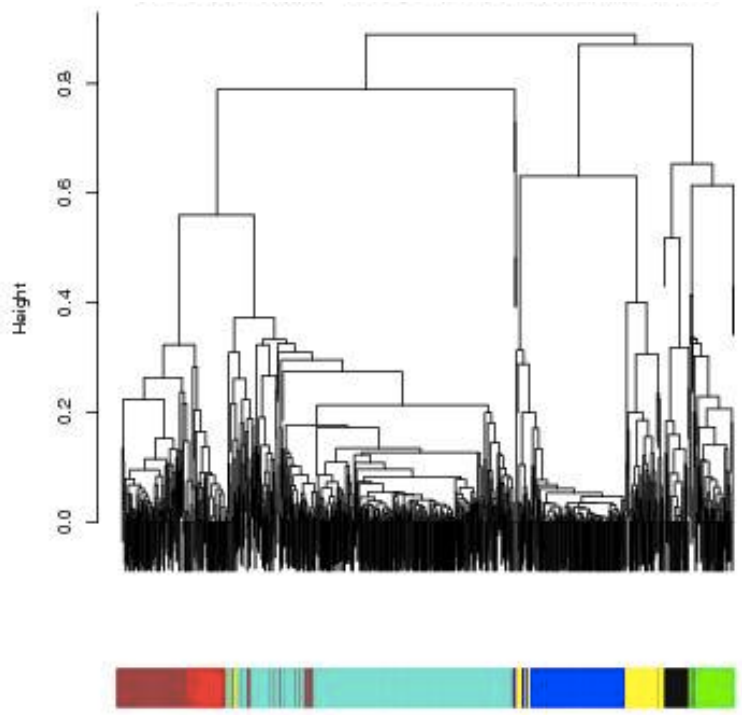
$$t_{ij} = \begin{cases} \dfrac{|N_1(i) \cap N_1(j)| + a_{ij}}{\min\{|N_1(i)|, |N_1(j)|\} + 1 - a_{ij}} & \text{if } i \neq j \\ 1 & \text{if } i = j. \end{cases} \quad (1)$$

where $N_1(i)$ denotes the set of direct neighbors of i excluding i itself and $|\cdot|$ denotes the number of elements (cardinality) in its argument. The quantity $|N_1(i) \cap N_1(j)|$ measures the number of common neighbors that nodes i and j share whereas $|N_1(i)|$ gives the number of neighbors of i. The topological overlap t_{ij} assumes a minimal value of 0 if there is no direct linkage between the two nodes and if they share no common direct neighbors. It assumes a maximum value of 1 if there is a direct link between the two nodes and if one set of direct neighbors is a subset of the other. The fact that t_{ij} is bounded between 0 and 1 is used in the definition of the topological overlap based dissimilarity measure (see Eq. 4).

Yip & Horvath, BMC Bioinf 2007

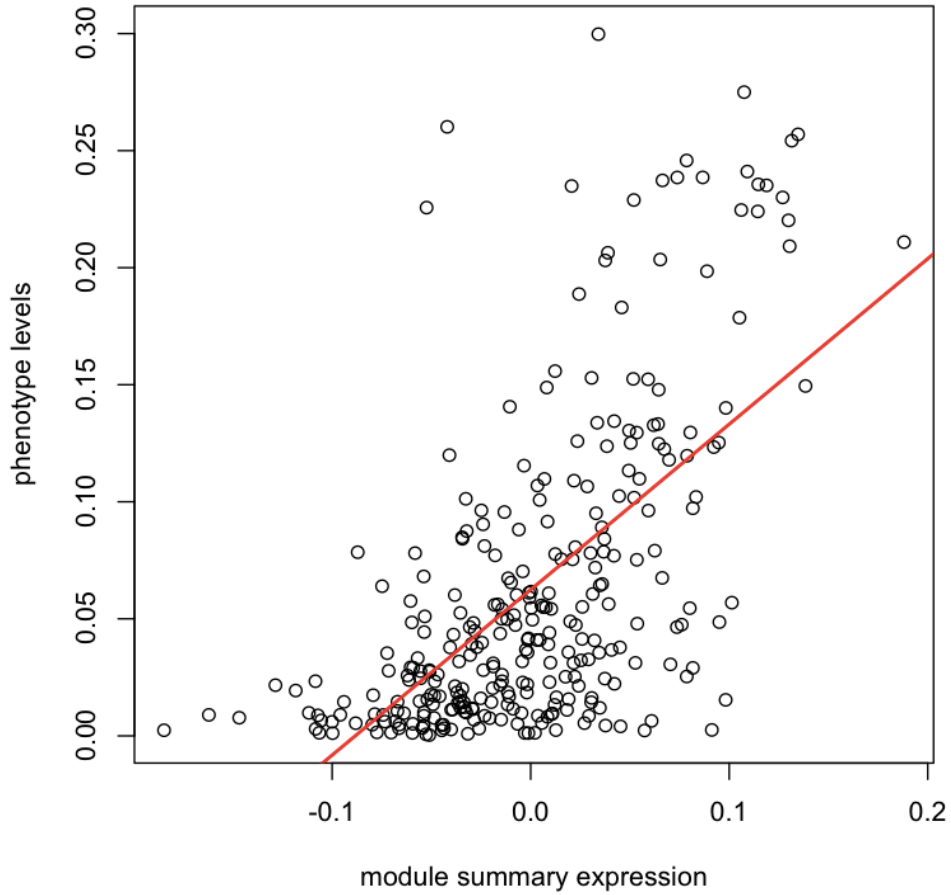
Detect gene modules

- Hierarchical clustering of TOM matrix
- Move through the dendrogram with a dynamic cutting algorithm

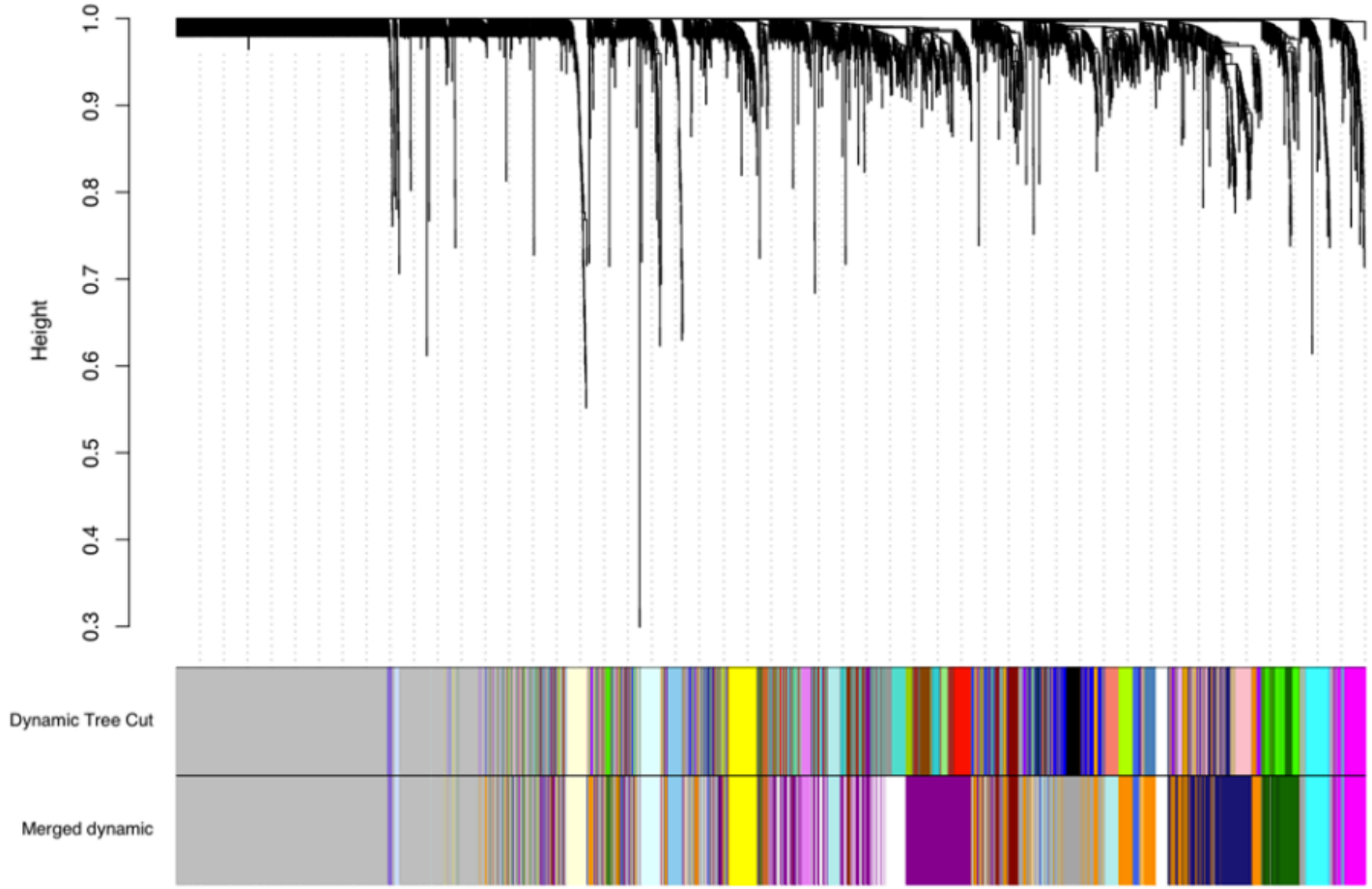


Yip & Horvath, BMC Bioinf 2007

Phenotype association analysis



Detect gene modules – Real data



Lipid association analysis – Real data

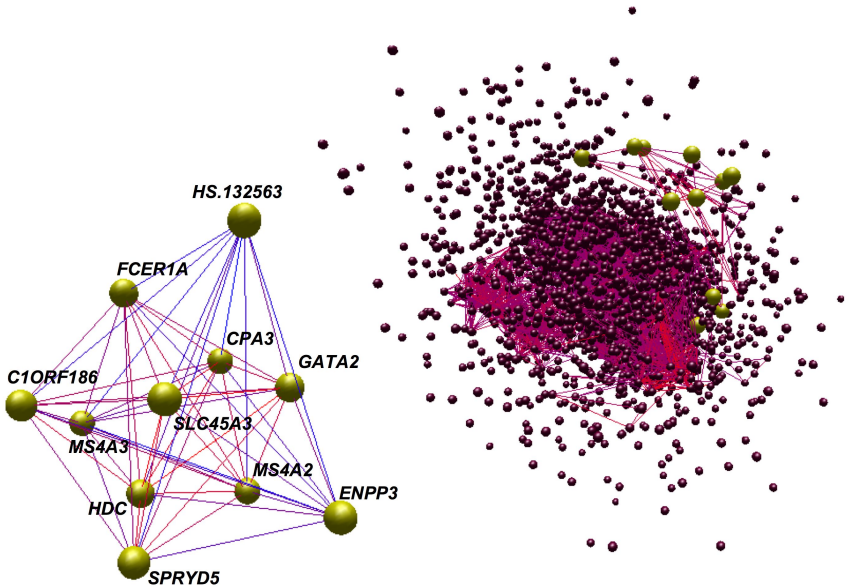
Module appears to be involved in immune response

Genes

- ***FCER1A*** – high affinity IgE receptor
- ***MS4A2*** – high affinity IgE receptor
- ***HDC*** – enzyme for histamine synthesis
- ***CPA3*** – mast cell secreted peptidase
- ***GATA2*** – TF crucial for mast cell dev
- ***SLC45A3*** - *?*
- ***SPRYD5*** - *?*
- ***MS4A3*** - *?*
- ***ENPP3*** - *?*
- ***C1ORF186*** - *?*
- ***HS.132563*** - *?*

Immune markers

- IL-1ra (P=3.1x10^{-6})
- C-reactive protein (P=2.6x10^{-4})
- HMW adiponectin (P=1.6x10^{-5})
- Total IgE (P>0.05)

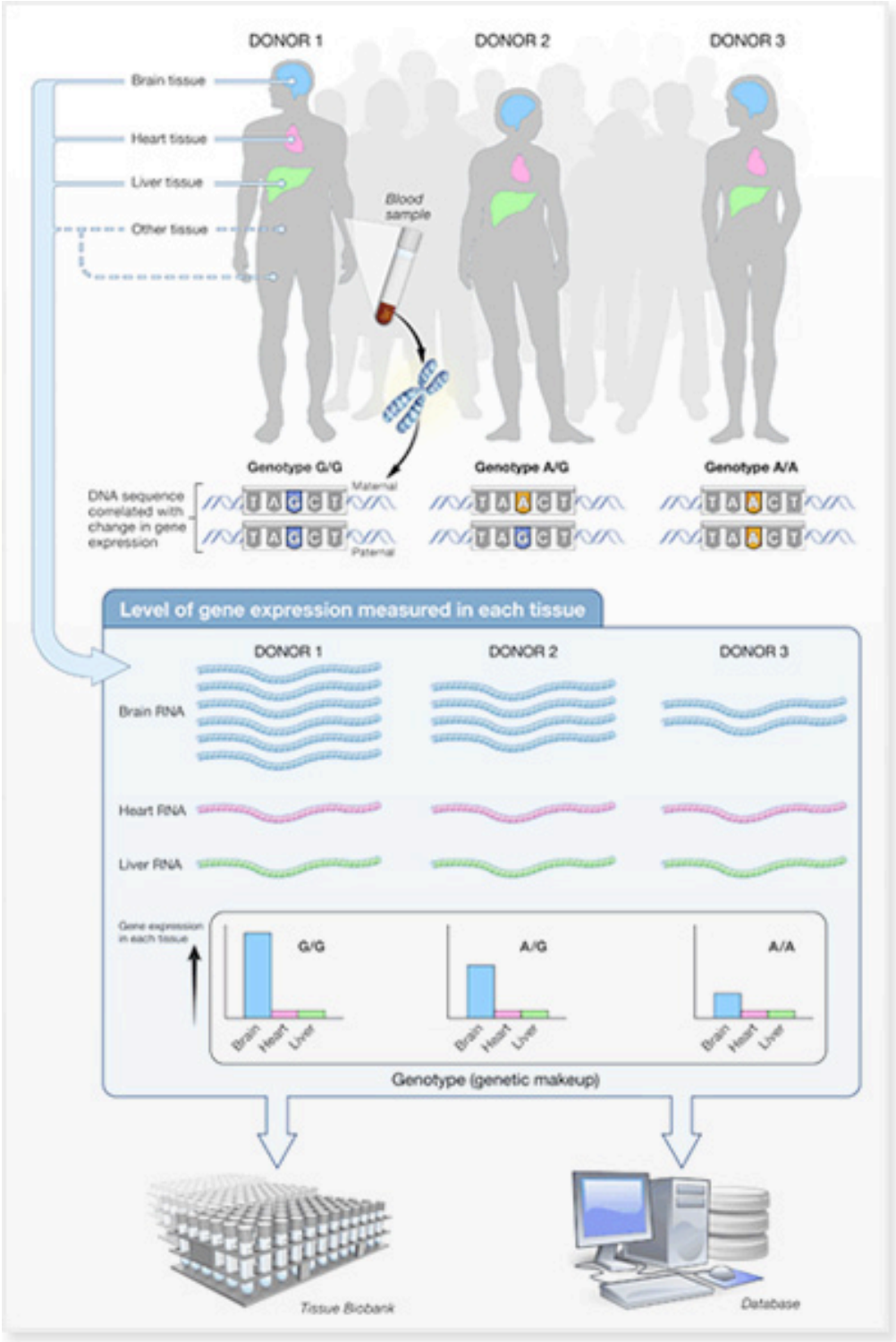


HUMAN GENOMICS

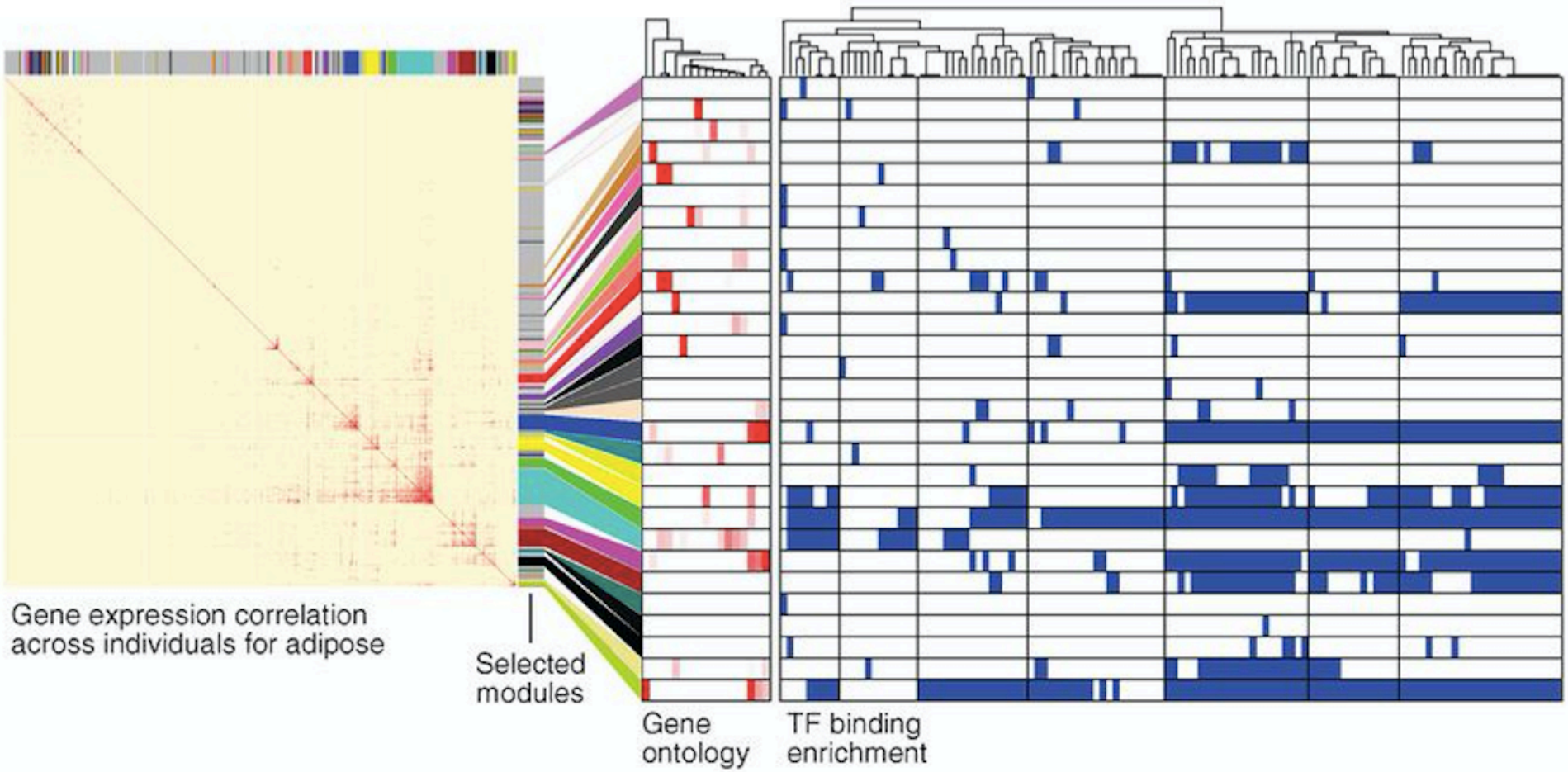
The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans

The GTEx Consortium*†

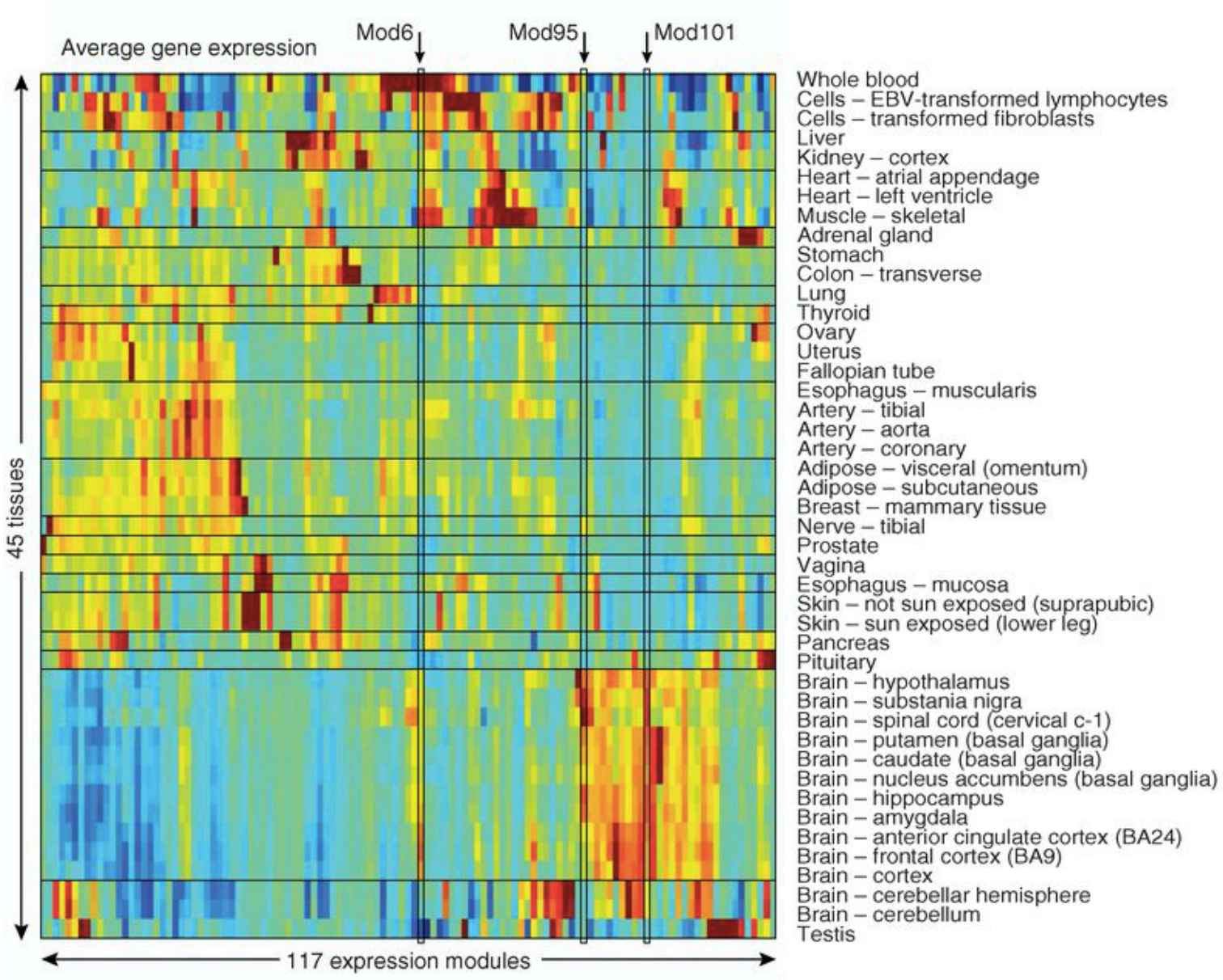
Science 8 MAY 2015 • VOL 348 ISSUE 6235



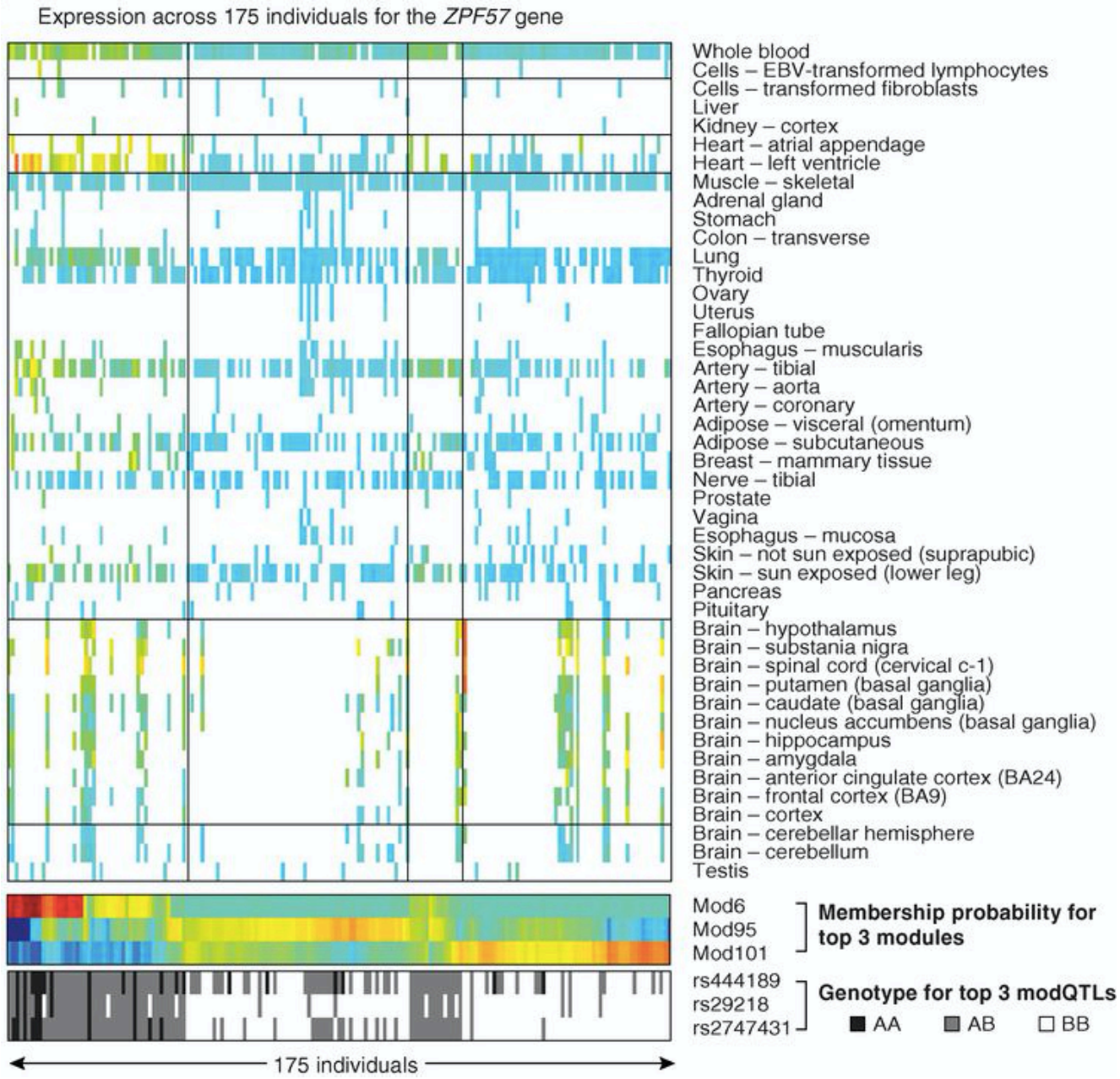
Adipose tissue: Differential pathway enrichment and TF binding profiles



Expression levels of modules across tissues



Expression of a gene (*ZPF57*) between tissues/genotypes



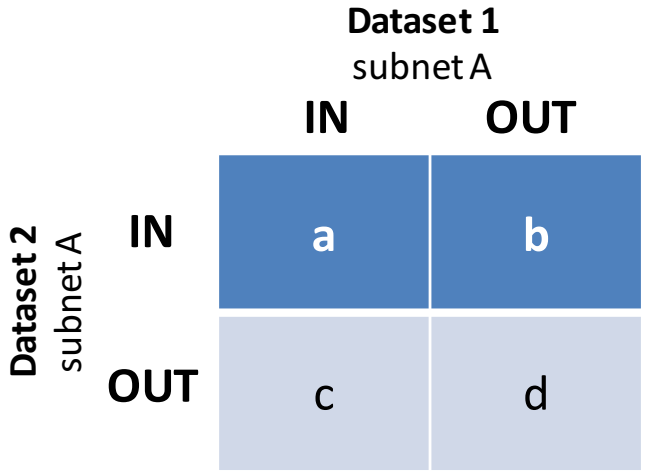
Preservation of subnets

- **Given a subnet (nodes, edges), is to preserved in a separate dataset?**

- **Examples**
  - Replication
    - Given N datasets generated under identical/similar settings, does a subnet ‘replicate’?
  - Cross-tissue gene network preservation
    - Is a subnet derived from liver data preserved in adipose data?
  - Microbial communities between body sites
    - Is an operational taxonomic unit (OTU) subnet preserved between skin and upper airway samples?

Approaches to subnet preservation

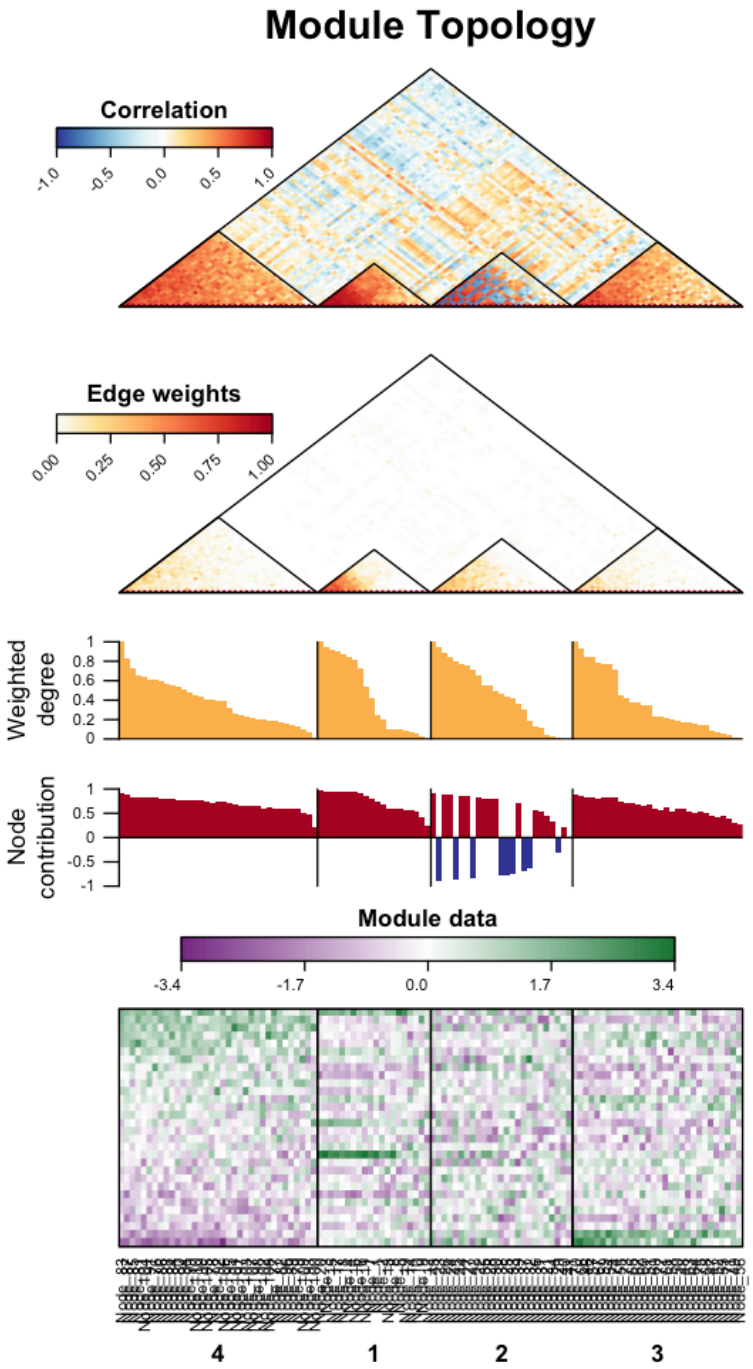
- **Tabulation**
  - Make a table of features in a given subnet and those not. Test for deviation from null (e.g. Fisher Exact Test).



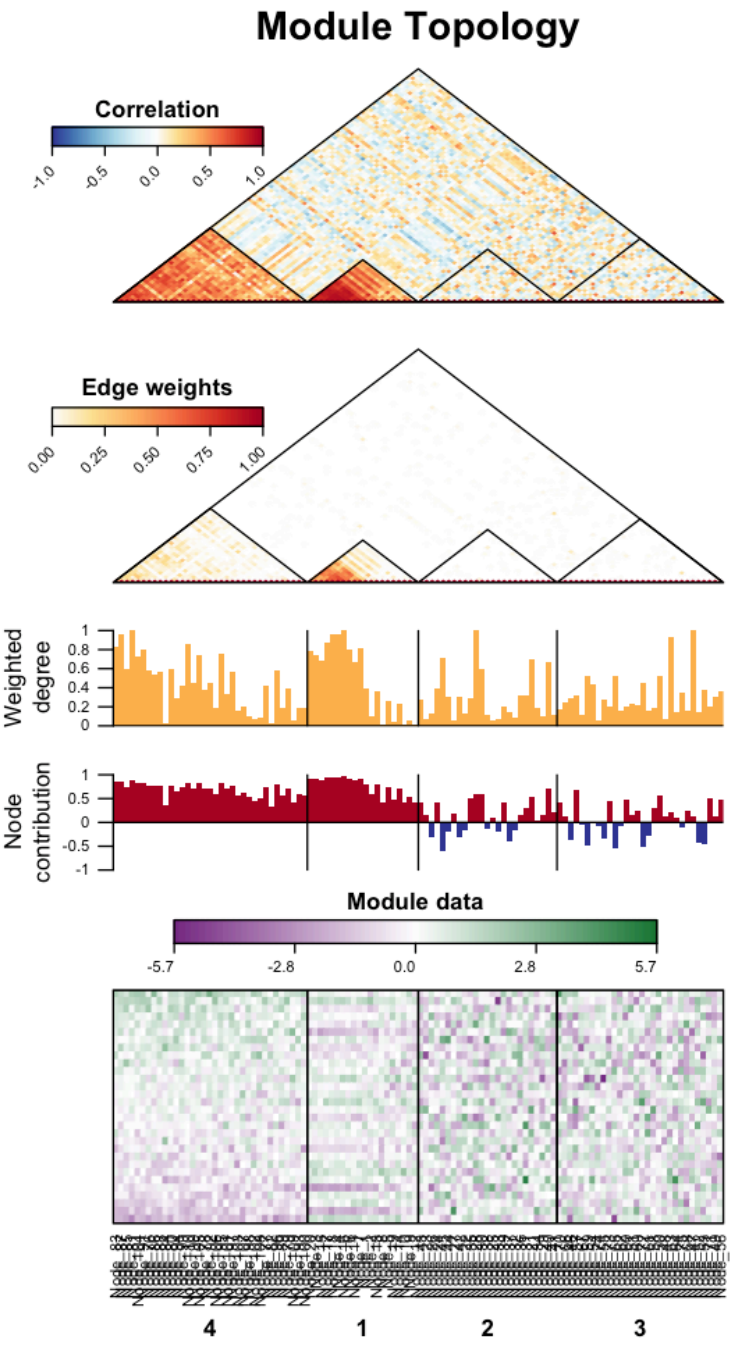
$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}}$$

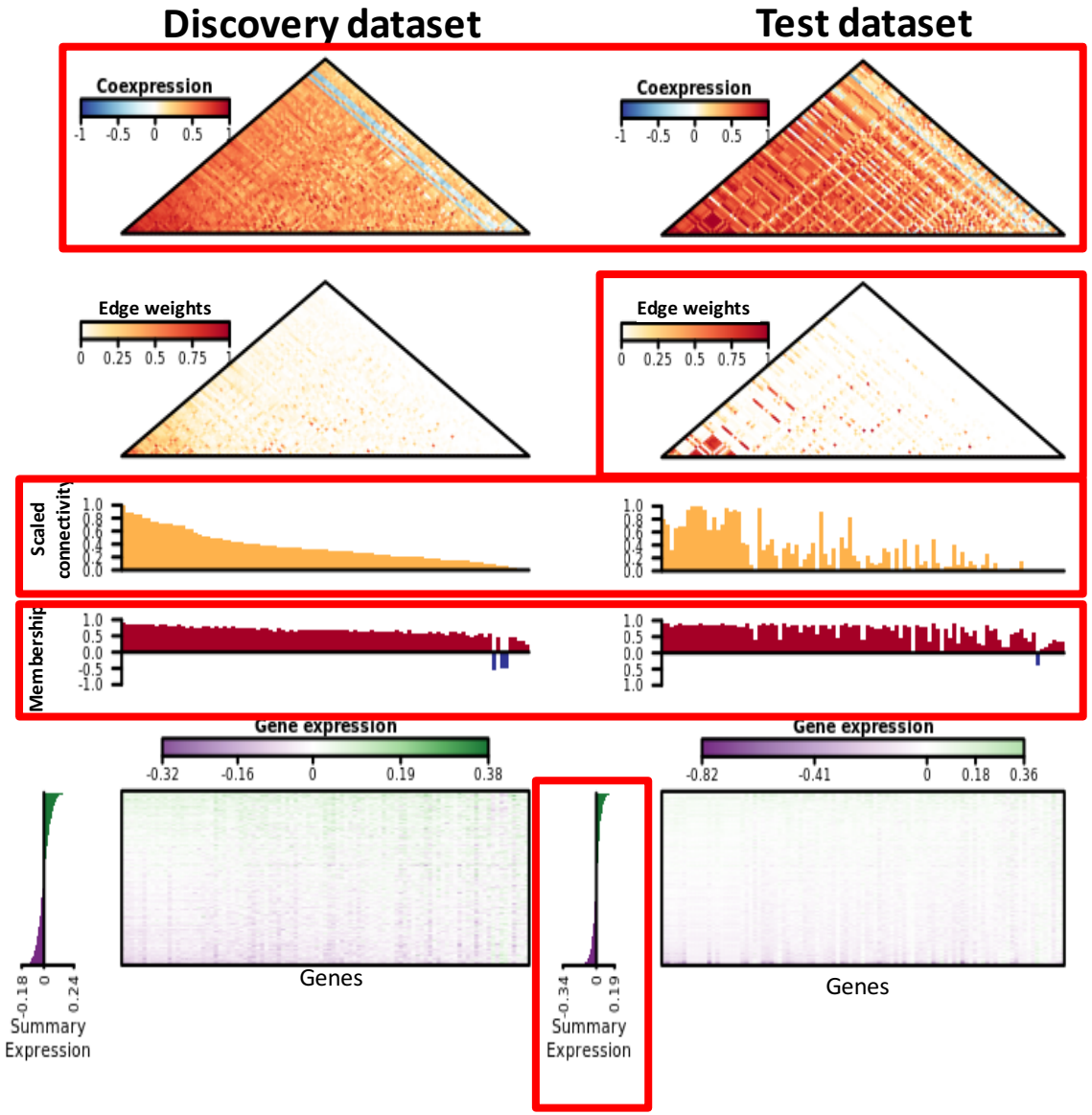
- **Topological properties**
  - Edge patterns (for simplicity, assume no missing nodes)

Dataset 1 (discovery)



Dataset 2 (replication)





Null Hypothesis:

Indistinguishable from comparisons to random gene sets in test dataset.

Module preservation statistics

How distinguishable is the module?

- Density / average edge weight
- Proportion of variance explained

How similar is the module topology?

- Similarity of correlation structure
- Correlation of connectivity / degree
- Correlation of membership / contribution

Combination:

- Mean correlation structure
- Average membership / contribution

Preservation of topology

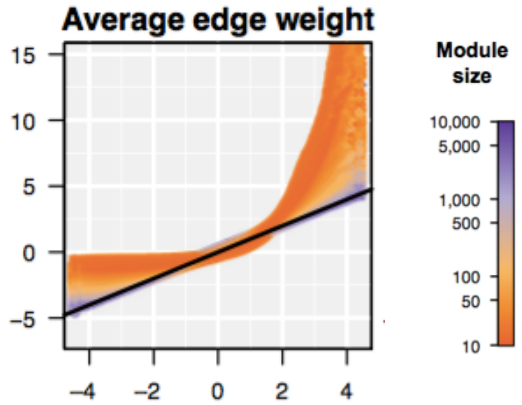
- Langfelder & Horvath, *PLOS Comp Bio* 2011
- Ritchie et al, *Cell Systems* 2016

	General name of test statistic	WGCNA	Calculation
(1)	Module coherence	Proportion of variance explained	$mean\left(\left(cor\left(g_i^{[t](w)}, Eig_1^{[t](w)}\right)\right)^2\right)$
(2)	Average node contribution	Mean sign-aware module membership	$mean\left(sign\left(cor\left(g_i^{[d](w)}, Eig_1^{[d](w)}\right)\right) \cdot cor\left(g_i^{[t](w)}, Eig_1^{[t](w)}\right)\right)$
(3)	Concordance of node contributions	Correlation of module membership	$cor\left(cor\left(g_i^{[d](w)}, Eig_1^{[d](w)}\right), cor\left(g_i^{[t](w)}, Eig_1^{[t](w)}\right)\right)$
(4)	Density of correlation structure	Mean sign-aware coexpression	$mean\left(sign\left(C^{[d](w)}\right) \cdot C^{[t](w)}\right)$
(5)	Concordance of correlation structure	Correlation of coexpression	$cor_{i \neq j}\left(C^{[d](w)}, C^{[t](w)}\right)$
(6)	Average edge weight	Mean adjacency	$mean_{i \neq j}\left(a_{ij}^{[t](w)}\right)$
(7)	Concordance of weighted degree	Correlation of intramodular connectivities	$cor\left(\left(\sum_{i \neq j}^{j} a_i\right)^{[d](w)}, \left(\sum_{i \neq j}^{j} a_i\right)^{[t](w)}\right)$

a	edge weight
g	feature vector
cor	correlation
C	correlation matrix
Sign	+ / -
Eig	1st principal component

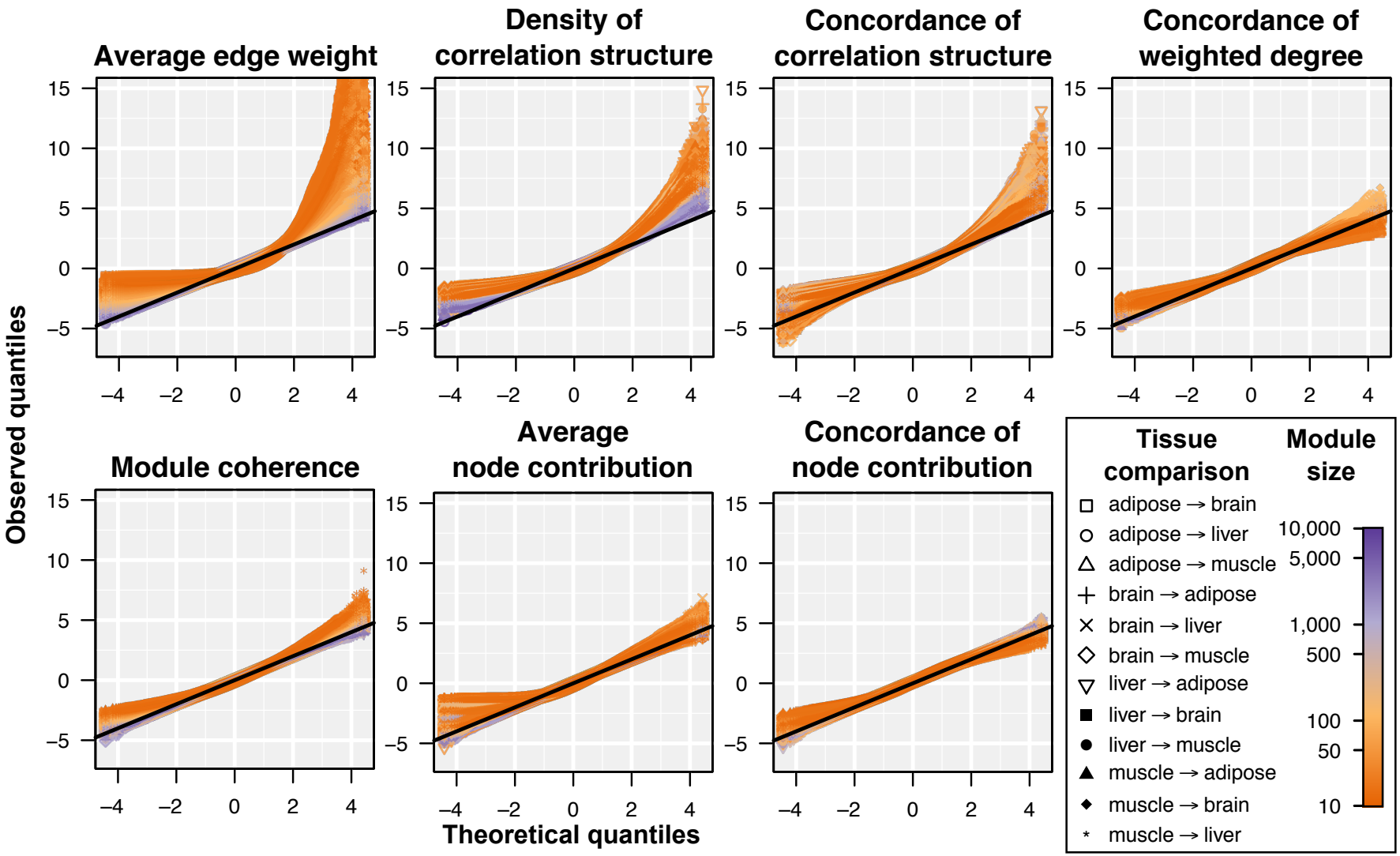
When in doubt, permute the data

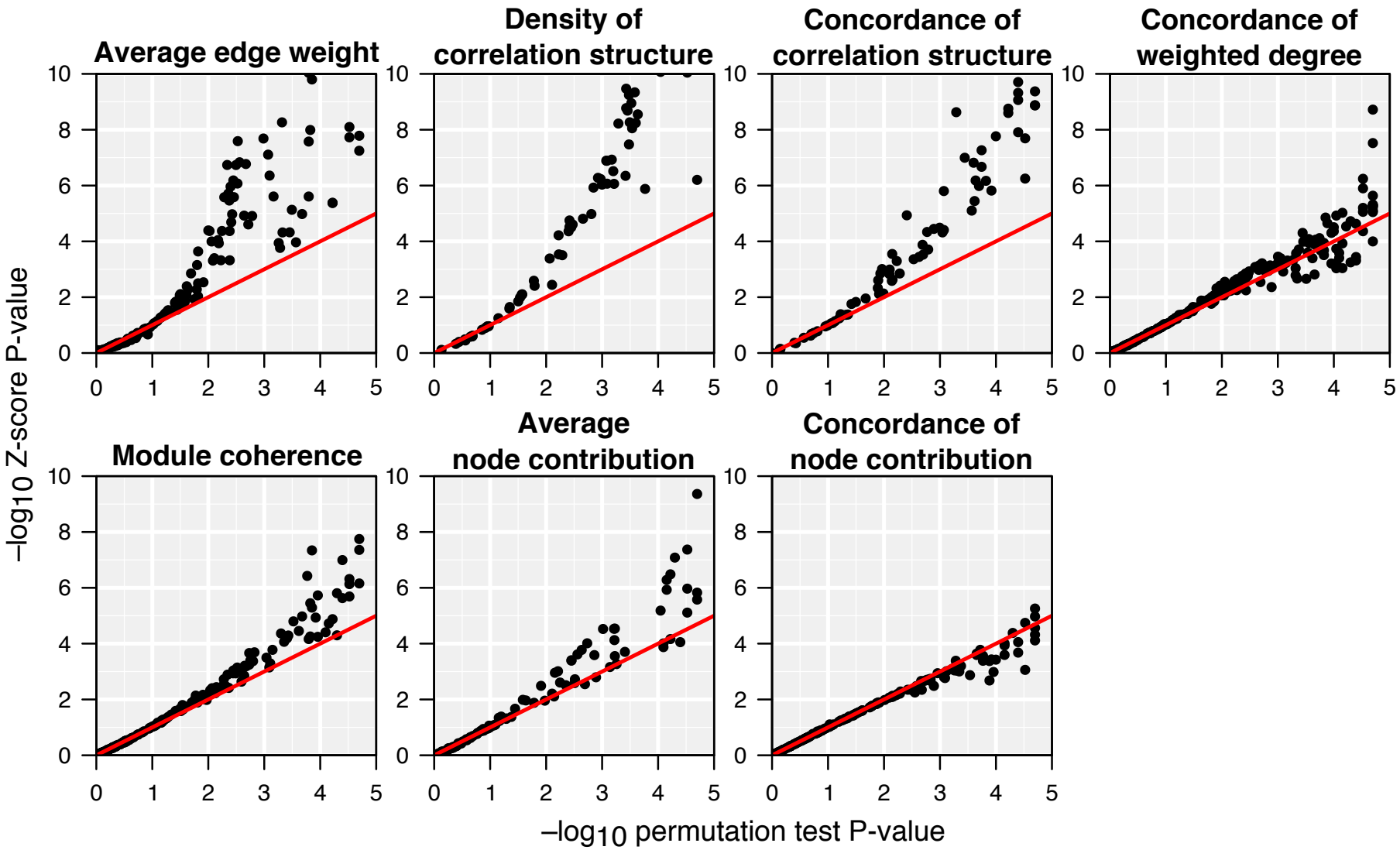
- **In network analysis, the complex relationships amongst nodes can make it difficult to assume a given test statistic follows a particular distribution**



- **It is common (and good practice) to create an empirical (permuted) distribution of the test statistic to assess the original observation's significance**

- **E.g. for a given module of with M nodes, with a given test statistic…**
  - Randomly draw M nodes from the overall network
  - Compute the test statistic of these random M nodes
  - Repeat many times
  - Compare the observed module value to the distribution of permuted values


Average edge weight
Density of correlation structure
Concordance of correlation structure
Concordance of weighted degree
Module coherence
Average node contribution
Concordance of node contribution
Observed quantiles
Theoretical quantiles
Tissue comparison
Module size
adipose → brain
adipose → liver
adipose → muscle
brain → adipose
brain → liver
brain → muscle
liver → adipose
liver → brain
liver → muscle
muscle → adipose
muscle → brain
muscle → liver
10,000
5,000
1,000
500
100
50
10


Average edge weight
Density of correlation structure
Concordance of correlation structure
Concordance of weighted degree
Module coherence
Average node contribution
Concordance of node contribution
−log10 Z-score P-value
−log10 permutation test P-value

Effect of scale-free'edness on preservation

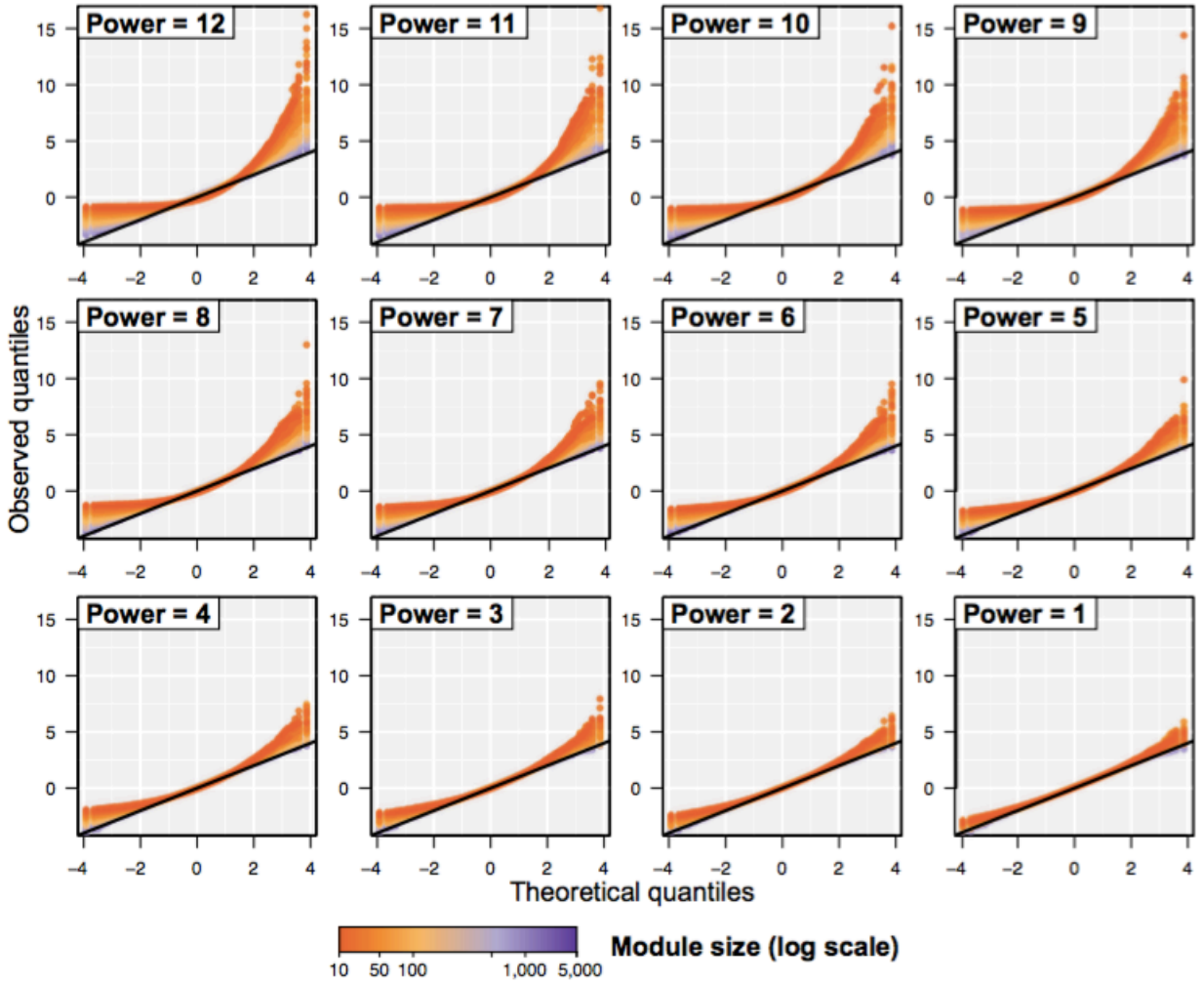
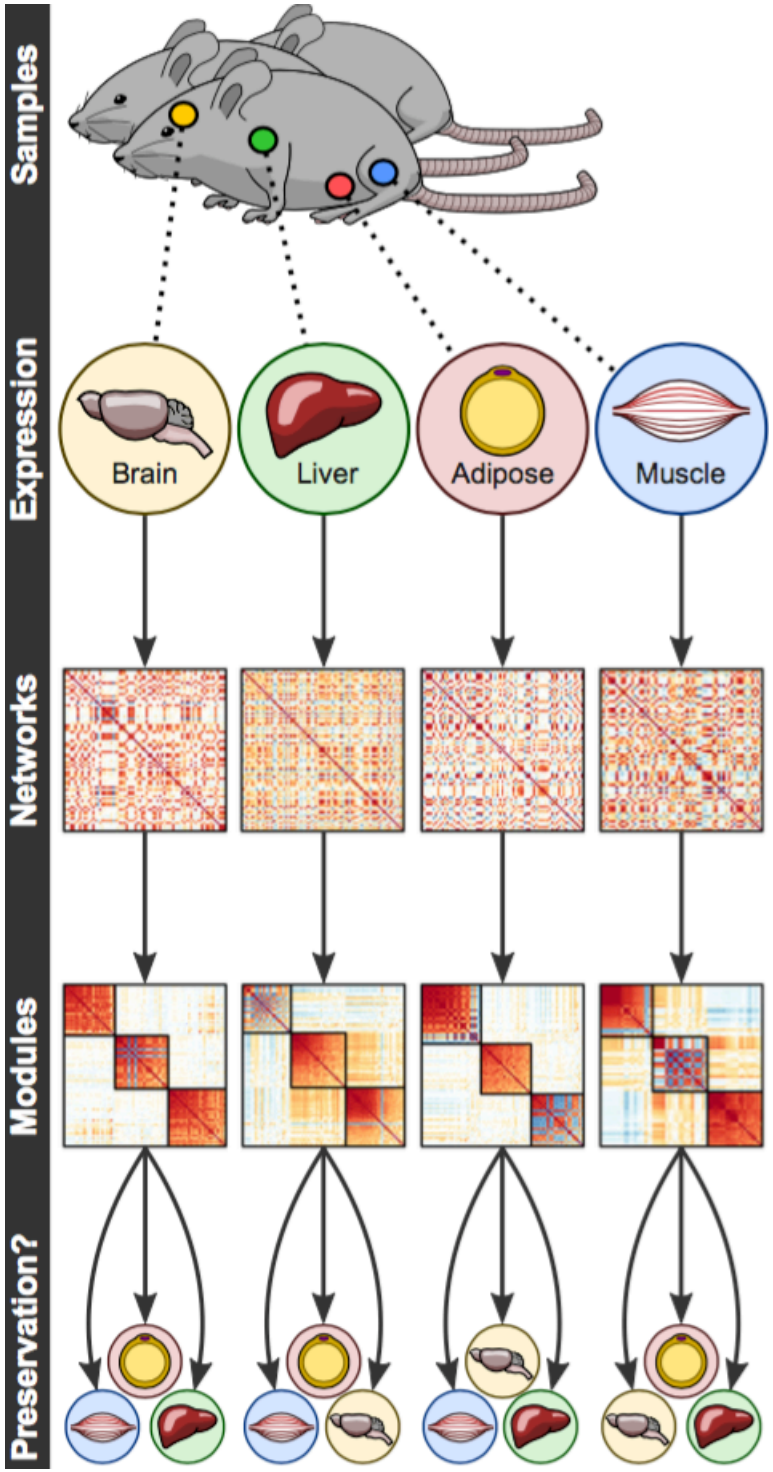
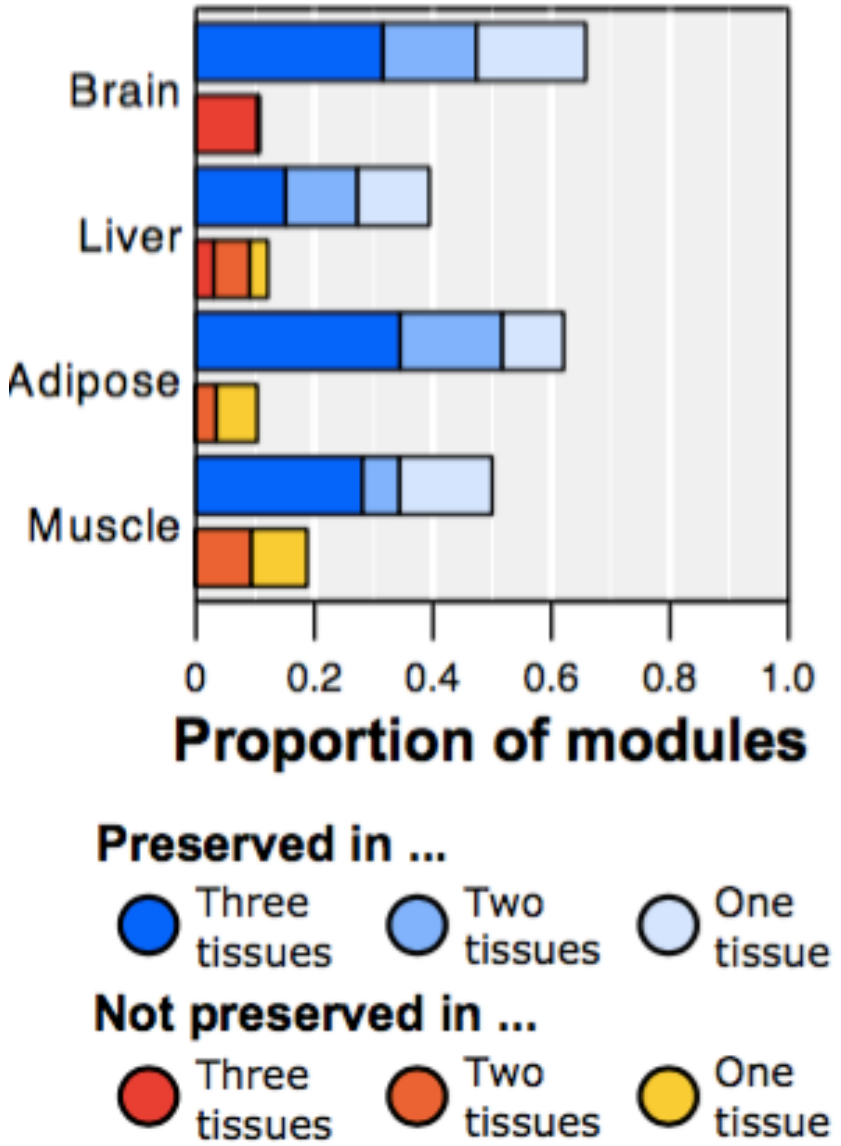
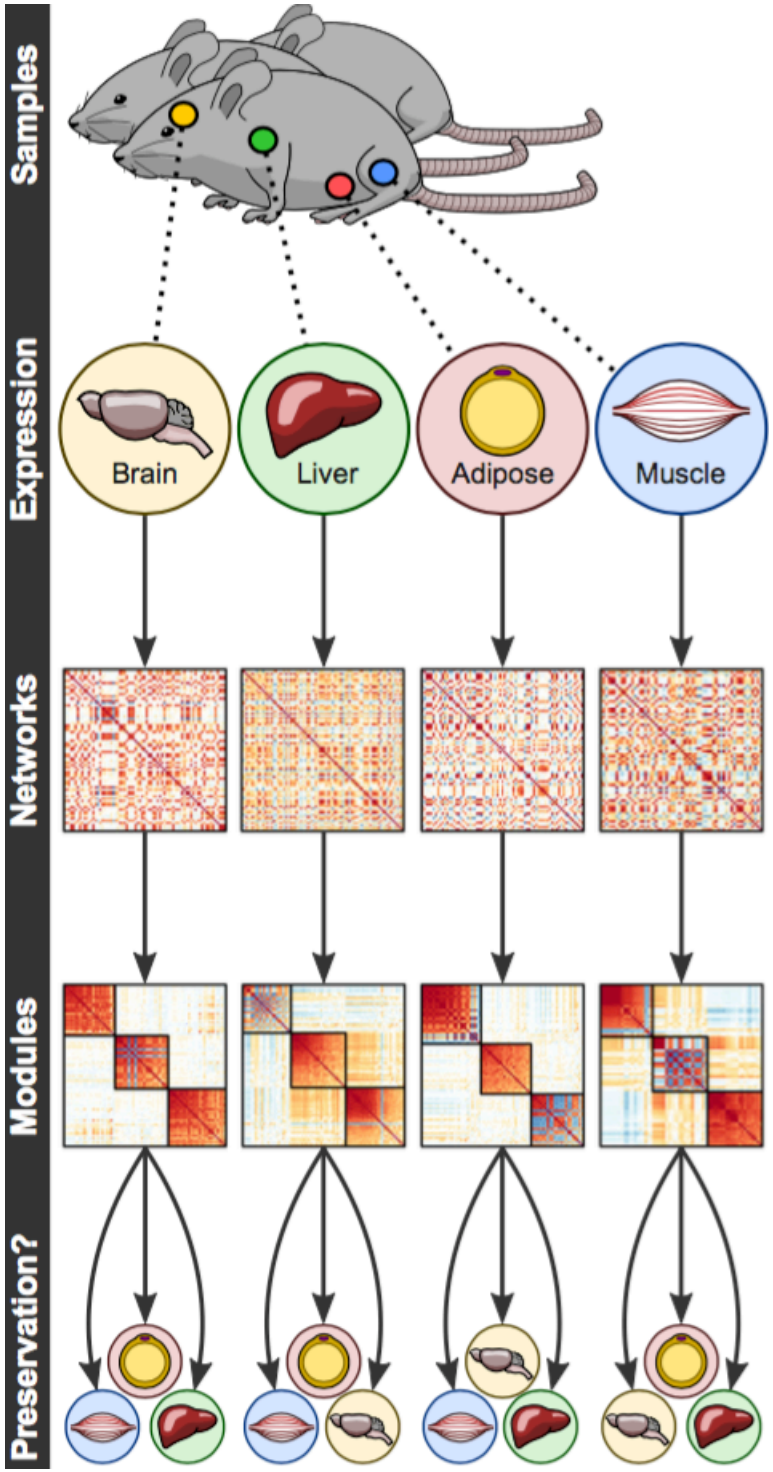


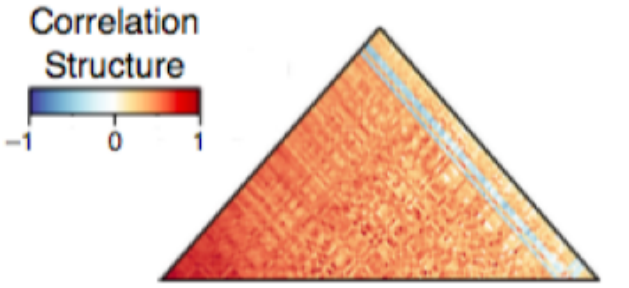
Figure S4, related to the experimental procedures and the main text: The scale-free assumption affects non-normality of the *average edge weight* statistic. Quantile-Quantile plots comparing

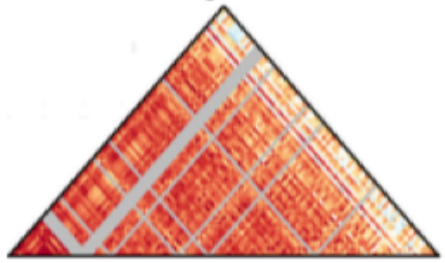


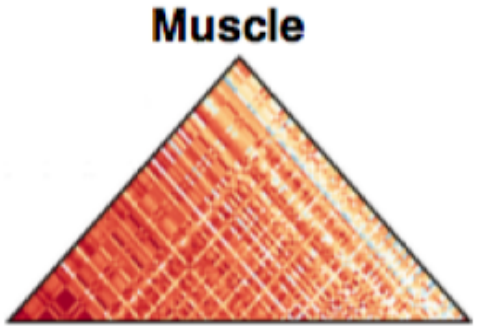


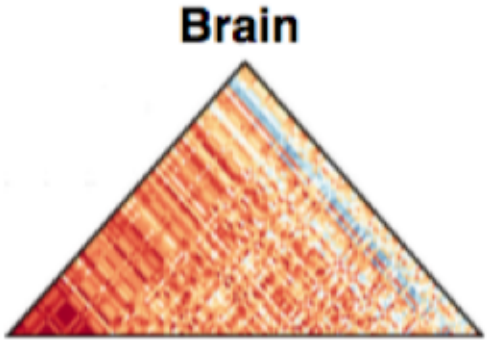
Ritchie et al, *Cell Systems* 2016


Samples
Expression
Networks
Modules
Preservation?
Brain
Liver
Adipose
Muscle


Liver Module A
Correlation
Structure
−1
0
1

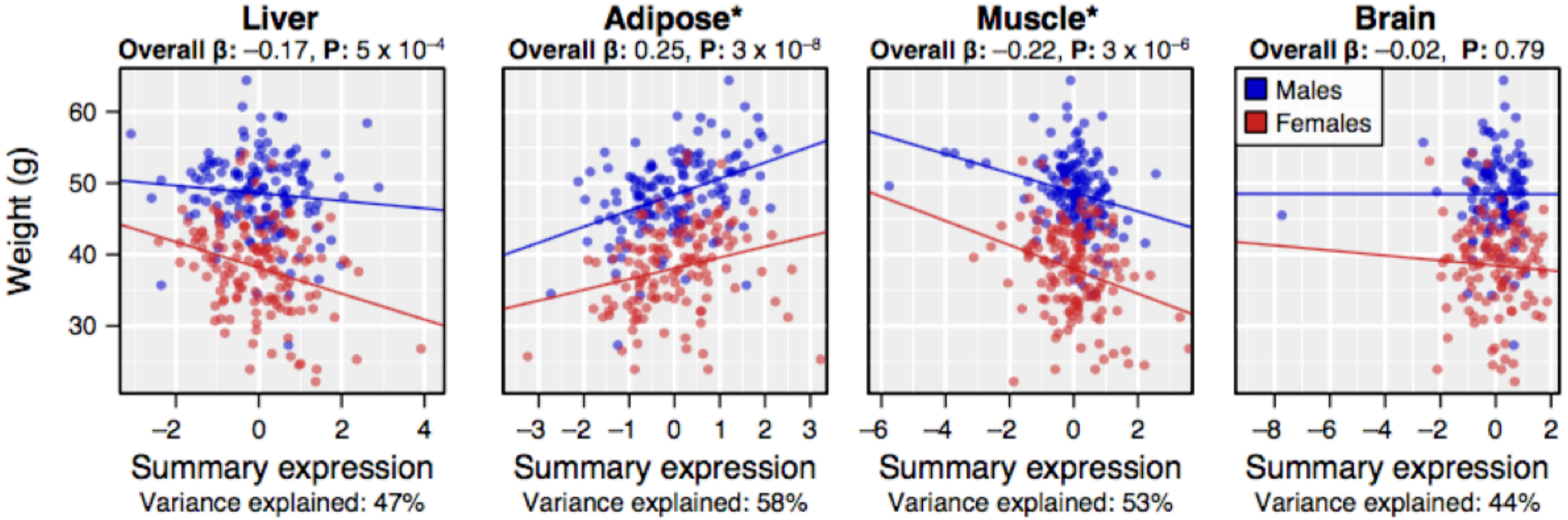

Adipose


Muscle


Brain

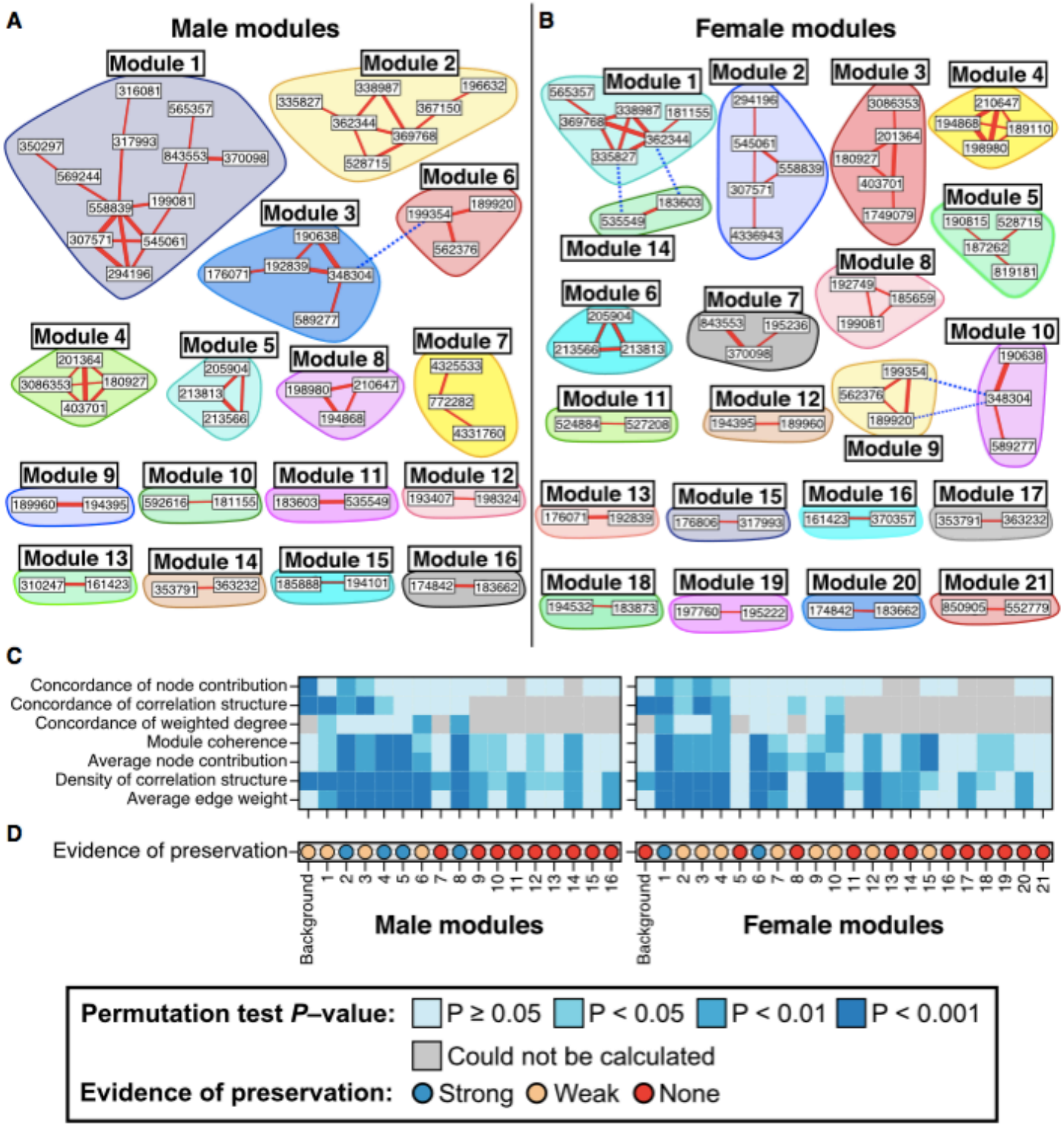
Gene 1 … n

Phenotypic association (body weight)



Test tissue	Trait	Effect size	95% confidence interval	*P*-value	*Q*-value
Adipose	**Weight**	**0.25**	**0.16– 0.33**	**3×10^{-8}**	**-**
	Insulin	**0.23**	**0.14– 0.32**	**1×10^{-6}**	**2×10^{-5}**
	Glucose/Insulin	**-0.21**	**-0.30– -0.12**	**7×10^{-6}**	**7×10^{-5}**
	Other fat	0.23	0.11– 0.35	1×10^{-4}	8×10^{-4}
	Total fat	**0.19**	**0.081– 0.30**	**7×10^{-4}**	**0.004**
	Length	0.17	0.069– 0.27	0.001	0.004
	MCP-1 (CCL2)	0.18	0.064– 0.29	0.002	0.007
	Glucose	0.18	0.064– 0.30	0.003	0.007
	Unesterified cholesterol	**0.18**	**0.061– 0.29**	**0.003**	**0.007**
Muscle	**Weight**	**-0.21**	**-0.30– -0.13**	**3×10^{-6}**	**-**
	Unesterified cholesterol	**-0.21**	**-0.34– -0.092**	**6×10^{-4}**	**0.01**
	Insulin	**-0.16**	**-0.25– -0.061**	**0.001**	**0.01**
	Total fat	**-0.19**	**-0.31– -0.072**	**0.002**	**0.01**
	Abdominal fat	-0.17	-0.27– -0.061	0.002	0.01
	Glucose/Insulin	**0.14**	**0.048– 0.24**	**0.003**	**0.01**
	Free fatty acids	-0.18	-0.31– -0.059	0.004	0.01
	LDL+VLDL	-0.18	-0.30– -0.056	0.005	0.01
	HDL/LDL+VLDL	0.17	0.051– 0.29	0.005	0.01
	Total cholesterol	-0.17	-0.29– -0.049	0.006	0.01

Gut 16S microbial community networks (SparCC)



Microbiome communities present in both men and women