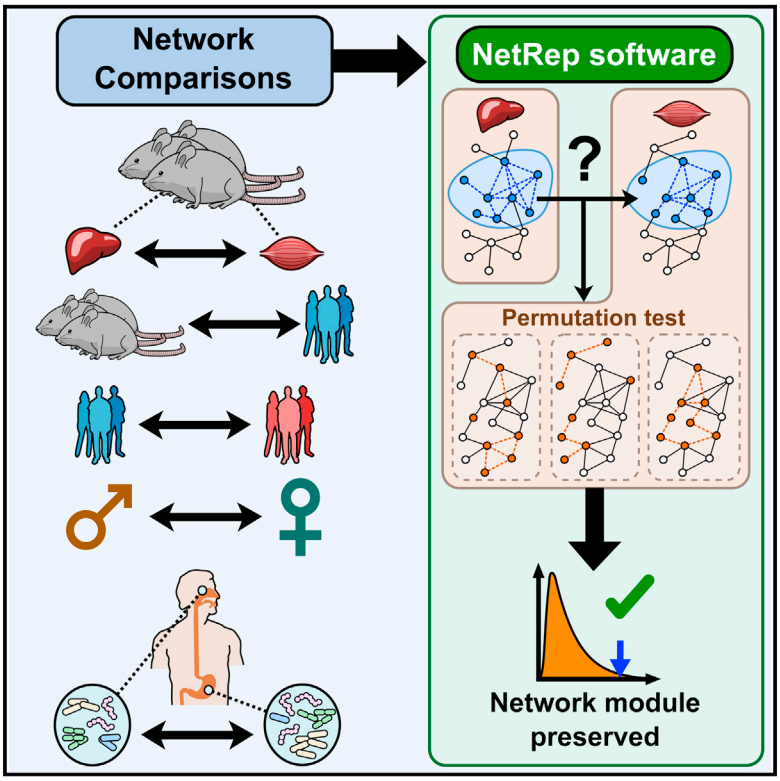
Article

Cell Systems

A Scalable Permutation Approach Reveals Replication and Preservation Patterns of Network Modules in Large Datasets

Graphical Abstract



Authors

Scott C. Ritchie, Stephen Watts, Liam G. Fearnley, Kathryn E. Holt, Gad Abraham, Michael Inouye

Correspondence

scottr@student.unimelb.edu.au (S.C.R.), minouye@unimelb.edu.au (M.I.)

In Brief

Ritchie et al. present NetRep, an open-source tool for statistically testing replication and preservation of network modules. NetRep is shown to identify multi-tissue gene networks in mice associated with body weight as well as gut microbial community networks preserved between men and women.

Highlights

- Common network preservation statistics are non-normal
- A fast permutation-based framework and software, NetRep, is presented and tested
- Using NetRep, we identify preserved gene networks across diverse tissues
- We further identify preserved networks of human gut microbiota across genders

Ritchie et al., 2016, Cell Systems *3*, 71–82
July 27, 2016
http://dx.doi.org/10.1016/j.cels.2016.06.012

CellPress

Cell Systems
Article

A Scalable Permutation Approach Reveals Replication and Preservation Patterns of Network Modules in Large Datasets

Scott C. Ritchie,[1,2,*] Stephen Watts,[1,3] Liam G. Fearnley,[1,2,4] Kathryn E. Holt,[1,3] Gad Abraham,[1,2,4] and Michael Inouye[1,2,4,*]

[1]Centre for Systems Genomics, The University of Melbourne, Parkville, VIC 3010, Australia
[2]Department of Pathology, The University of Melbourne, Parkville, VIC 3010, Australia
[3]Department of Biochemistry and Molecular Biology, Bio21 Molecular Science and Biotechnology Institute, The University of Melbourne, Parkville, VIC 3010, Australia
[4]School of BioSciences, The University of Melbourne, Parkville, VIC 3010, Australia
*Correspondence: scottr@student.unimelb.edu.au (S.C.R.), minouye@unimelb.edu.au (M.I.)
http://dx.doi.org/10.1016/j.cels.2016.06.012

SUMMARY

Network modules—topologically distinct groups of edges and nodes—that are preserved across datasets can reveal common features of organisms, tissues, cell types, and molecules. Many statistics to identify such modules have been developed, but testing their significance requires heuristics. Here, we demonstrate that current methods for assessing module preservation are systematically biased and produce skewed p values. We introduce NetRep, a rapid and computationally efficient method that uses a permutation approach to score module preservation without assuming data are normally distributed. NetRep produces unbiased p values and can distinguish between true and false positives during multiple hypothesis testing. We use NetRep to quantify preservation of gene coexpression modules across murine brain, liver, adipose, and muscle tissues. Complex patterns of multi-tissue preservation were revealed, including a liver-derived housekeeping module that displayed adipose- and muscle-specific association with body weight. Finally, we demonstrate the broader applicability of NetRep by quantifying preservation of bacterial networks in gut microbiota between men and women.

INTRODUCTION

Modern high-throughput technologies generate a large amount of genomic, transcriptomic, metabolomic, and proteomic data. Rather than consider each measurement in isolation, network inference techniques integrate these -omic data to identify meaningful biological relationships between components. In general, these approaches represent each measured variable as a node and the relationships between variables as edges that connect nodes; in aggregate, the connected edges and nodes comprise the network. Statistical analysis of these networks can identify and characterize gene modules, gene regulatory networks, protein-protein interactions, microbial networks and predict diverse molecular interactions (Abraham et al., 2014; Barabási et al., 2011; Dagan, 2011; Faust and Raes, 2012; Lusis and Weiss, 2010; Schadt, 2009).

Typically, a research project investigates one or more subgraphs of these inferred networks, for example, a group of genes associated with disease pathogenesis. These are commonly referred to as network "modules" (Gustafsson et al., 2014; Rotival and Petretto, 2014). The next step for many studies is to assess whether a network module(s) is wholly or partially preserved in an independent dataset(s); preservation is taken as an indication that the module is biologically relevant. Module preservation analysis can be used to quantify the replication of modules (Emilsson et al., 2008; Fuller et al., 2007; Hawrylycz et al., 2012; Miller et al., 2010; Ritchie et al., 2015; Xia et al., 2006), to determine their changes across conditions (Fuller et al., 2007; Keller et al., 2008; van Nas et al., 2009), to examine their tissue specificity (Cai et al., 2010; Keller et al., 2008), and to identify modules conserved across different species (Boyle et al., 2014; Gerstein et al., 2014; Stuart et al., 2003).

Module preservation analyses are both timely and increasingly common, given recent concerns about the reproducibility and generality of research findings (Collins and Tabak, 2014). However, rigorous statistical methodology for assessing module preservation has received little attention. Module preservation is typically assessed via visual inspection and/or tabulation of module composition after application of the same network inference and module detection algorithms in the second (i.e., test) dataset (Boyle et al., 2014; Gerstein et al., 2014; Keller et al., 2008; Miller et al., 2010; van Nas et al., 2009; Xia et al., 2006). A major limitation to these approaches is that they cannot systematically capture information about the network topology, i.e., the relationships between nodes in the module of interest. These relationships encode important biological information. For example, node degree (how many other nodes any given node is connected to) is a common metric analyzed in network studies, as it can indicate relative importance to the network. Genes that are highly connected are often essential to an organism's survival (Carlson et al., 2006; Jeong et al., 2001), and within a module, node degree can be used as a measure of relative

biological importance (Horvath and Dong, 2008; Langfelder et al., 2013).

To address this problem, Langfelder et al. developed a suite of statistics for quantifying the preservation of a module's topology in an independent dataset where the same module or a subset of nodes has been measured (Langfelder et al., 2011). Their module preservation statistics were primarily designed for networks inferred through weighted gene coexpression network analysis (WGCNA) (Zhang and Horvath, 2005). These are weighted, complete networks, which are defined through a power transform on the correlation structure calculated between all pairs of genes. Each gene is connected to every other gene with an edge weight between 0 and 1 denoting connection strength. Modules can either be defined as genes belonging to a pathway of interest or discovered de novo through clustering of the network (Zhang and Horvath, 2005). Seven module preservation statistics are used to quantify module preservation. For convenience, their definitions are given in the Experimental Procedures and their biological interpretation in the Supplemental Experimental Procedures. Broadly speaking, they measure whether the density and connectivity of a module are preserved in a second test dataset. The density statistics assess whether nodes composing a module remain strongly connected in the test dataset and whether measurements in each sample are similar across the module's nodes. The connectivity statistics assess whether the pattern of node-node relationships are similar between the discovery and test datasets (Langfelder et al., 2011). This approach uses Z scores to determine whether any test statistic is significant. The null hypothesis is that the module of interest is not preserved, and thus the value of each statistic should not be higher than expected by chance, assuming each statistic follows a normal distribution. However, Langfelder et al. found that Z scores were typically abnormally large, leading to many false positives in simulation. Consequently, heuristic tests for significance were formulated (Langfelder et al., 2011).

The number of modules and datasets undergoing module preservation analyses is increasing as large multi-omic datasets with dozens of tissues, cell lines, conditions, and corresponding metadata become common and openly available. In particular, studies are now performing unbiased discovery of preserved gene coexpression modules (Cai et al., 2010; Melé et al., 2015) and identifying tissue-specific and multi-tissue modules (Pierson et al., 2015; GTEx Consortium, 2015). Multiple testing becomes a problem for studies assessing preservation of many modules across many datasets; false positives may be detected due to the large number of statistical tests. Typically this is addressed through multiple testing adjustment of p values or thresholds for significance. It is therefore crucial that preservation p values are unbiased and accurately calibrated in order to control type I (false positive) and type II (false negative) errors (Bender and Lange, 2001). Heuristic tests cannot be adjusted for multiple testing, and thus such studies are susceptible to increased type I and II error rates.

Robust and unbiased p values should be determined in the absence of distributional assumptions by permutation testing. When this principle is applied to module preservation analyses, the module preservation statistics are calculated when shuffling node identifiers in the test dataset to determine their distributions under the null hypothesis. The true value of each statistic is then compared to the empirical null distribution to obtain a permutation test p value. However, at least w permutations are required to estimate significance at a threshold of $1/w$ (Phipson and Smyth, 2010). The analysis of large datasets, together with the concomitant multiple testing correction burden, necessitates increasingly stringent significance thresholds, making permutation-based significance testing computationally challenging.

Here, we address this challenge by developing a rapid and efficient approach for assessing module preservation, available as an R package, NetRep. We use NetRep to create and assess the empirical null distributions of Langfelder et al.'s suite of module preservation statistics when inferring weighted gene coexpression networks in a publicly available resource of mouse adipose, brain, liver, and muscle tissue expression (Yang et al., 2006). We show the majority of these statistics have non-normal distributions and are thus in need of a permutation approach. Next, we demonstrate NetRep's scalability to large-scale module preservation analysis by performing permutation tests to quantify cross-tissue gene coexpression module preservation. We identify and characterize multi-tissue modules associated with mouse body weight. Consequently, we uncover a body weight-associated module with differential adipose and muscle tissue expression. Finally, we explore the broader applicability of Langfelder et al.'s suite of module preservation statistics by using NetRep to quantify the preservation of gut microbial community networks between men and women from publicly available 16S rRNA gene sequence data (Human Microbiome Project Consortium, 2012).

RESULTS

Rapid Module Preservation Analysis

We have developed a time- and memory-efficient method for massively parallel calculation of module preservation statistics. The software is available as an R package, NetRep, which can be downloaded from https://github.com/InouyeLab/NetRep. Implementation details are provided in the Supplemental Experimental Procedures.

To examine the null distributions of the module preservation statistics in an empirical setting, we applied NetRep to publicly available gene expression data for brain, adipose, liver, and muscle tissues from a BxH mouse cross (Yang et al., 2006). From 334 total mice, there were 249 brain, 295 adipose, 306 liver, and 319 muscle tissue samples available for analysis (Experimental Procedures). Figure 1 illustrates the workflow of network construction, module detection, and module preservation analysis. We inferred weighted gene coexpression networks (Experimental Procedures; Zhang and Horvath, 2005) for each tissue, identifying 38, 66, 29, and 32 distinct coexpression modules in the brain, liver, adipose, and muscle tissues, respectively (Figure S1). For a module, we refer to the tissue it was initially identified in as its "discovery" tissue, and other tissues where its preservation is being tested as "non-discovery" tissues.

A runtime comparison of NetRep versus WGCNA's modulePreservation function for calculating permutations for these modules is provided in the Supplemental Experimental Procedures (see also Figure S2). Briefly, NetRep was, on

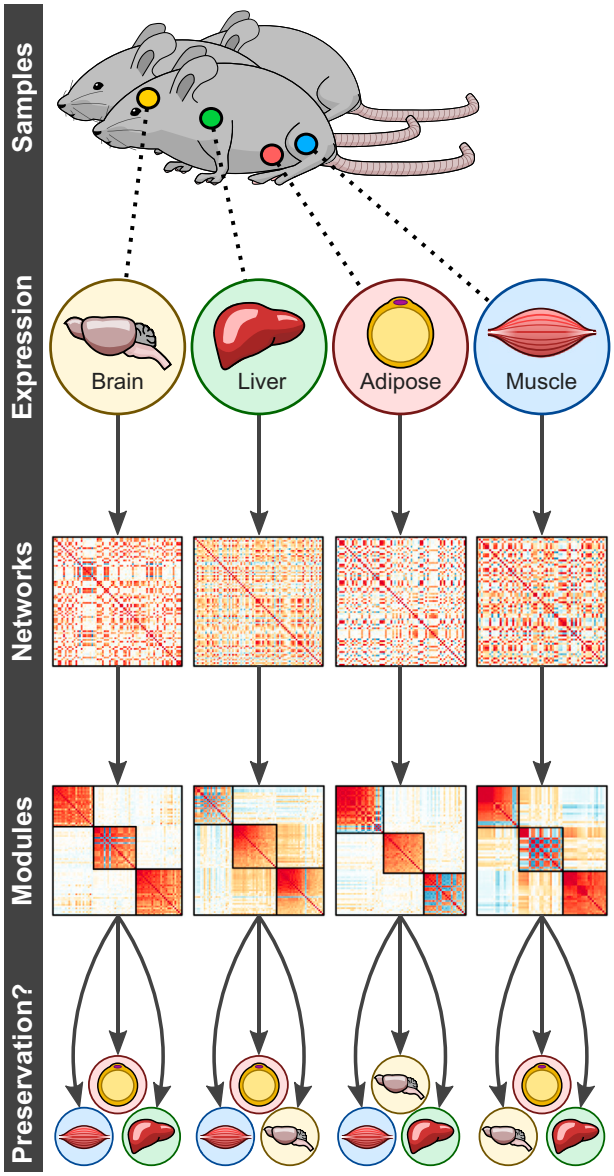


Figure 1. Workflow of Network Construction, Module Detection, and Module Preservation Analysis Workflow on the BxH Mice Tissue Expression

First, the correlation structure (coexpression) between probes was calculated from the gene expression in each tissue. Next the interaction network was inferred and modules were detected using weighted gene coexpression network analysis (WGCNA) (Experimental Procedures). Finally, the topological preservation of each module was assessed in each non-discovery tissue using their respective gene expression, correlation structure, and interaction network matrices.

average, 11 times faster than WGCNA and used less memory when run in parallel. The runtime of NetRep depended on multiple factors, as follows: the number of dataset comparisons, number of permutations for each comparison, sample size, and the sum of the squared sizes of modules included in the analysis. Pairwise comparison of the 165 modules across 4 mouse tissues with 100,000 permutations took approximately 8 days when NetRep was parallelized over 40 cores.

Null Distributions of Module Preservation Statistics

We investigated the normality of the seven module preservation statistics by comparing permutation-based null distribution quantiles to theoretical normal distribution quantiles (Figure S3). For each discovered module, module preservation statistics were calculated on 100,000 random gene sets of identical size in each non-discovery tissue (Experimental Procedures). Across all 38 brain modules, 66 liver modules, 29 adipose modules, and 32 muscle modules, 495 null distributions were generated for each module preservation statistic. We observed strong non-normality of null distributions generated for the average edge weight, density of correlation structure, and concordance of correlation structure statistics (Figure S3). Moderate non-normality was also observed in null distributions for the other four statistics. We also observed increasing non-normality with decreasing module size, particularly for modules of <100 probes (Figure S3). This shows that the assumption of normality required for Z score statistics is often violated.

To determine the consequences of non-normality, we matched Z score p values and permutation p values calculated for each module and preservation test statistic. Substantial inflation of the Z scores and corresponding deflation of p values was observed for average edge weight, density of correlation structure, and concordance of correlation structure (Figure 2). Moderate inflation was also observed for module coherence and average node contribution (Figure 2). These results are consistent with the extremely low p values that motivated heuristic significance thresholds (Langfelder et al., 2011), indicating that a non-parametric approach is necessary to produce unbiased p values.

We further investigated the test statistic with the most extreme deviation from normality, the average edge weight statistic. Edge weights in interaction networks inferred with WGCNA are calculated through a power transform on the absolute correlation coefficient. This power transform acts as a soft-threshold: it penalizes weak correlation coefficients toward zero. The soft-threshold power is chosen under the assumption that the resulting network should be scale free (Experimental Procedures; Zhang and Horvath, 2005). Under this assumption, the distribution of the weighted degree of the network follows an inverse power law where a few hub genes are strongly connected in the network, while most genes are only weakly connected (Barabási and Albert, 1999; Stumpf and Porter, 2012). To test the impact of the scale-free assumption on the null distribution normality, we calculated the average edge weight statistic on 10,000 permutations in the muscle tissue for the 66 liver modules, varying the soft-threshold exponent used to define the edge weights. We observed a trend toward non-normality as the exponent increased, indicating that the scale-free assumption contributes to non-normality for this statistic (Figure S4). We similarly generated null distributions for the concordance of weighted degree as it is also calculated from the edge weights in the interaction network; however, its deviation from normality was only mild and did not change as the exponent increased (Figure S5). These analyses indicate that the non-normality of preservation test statistics can be influenced by the distribution of node degree.

To assess the performance of NetRep and the permutation approach for quantifying module preservation, we simulated a

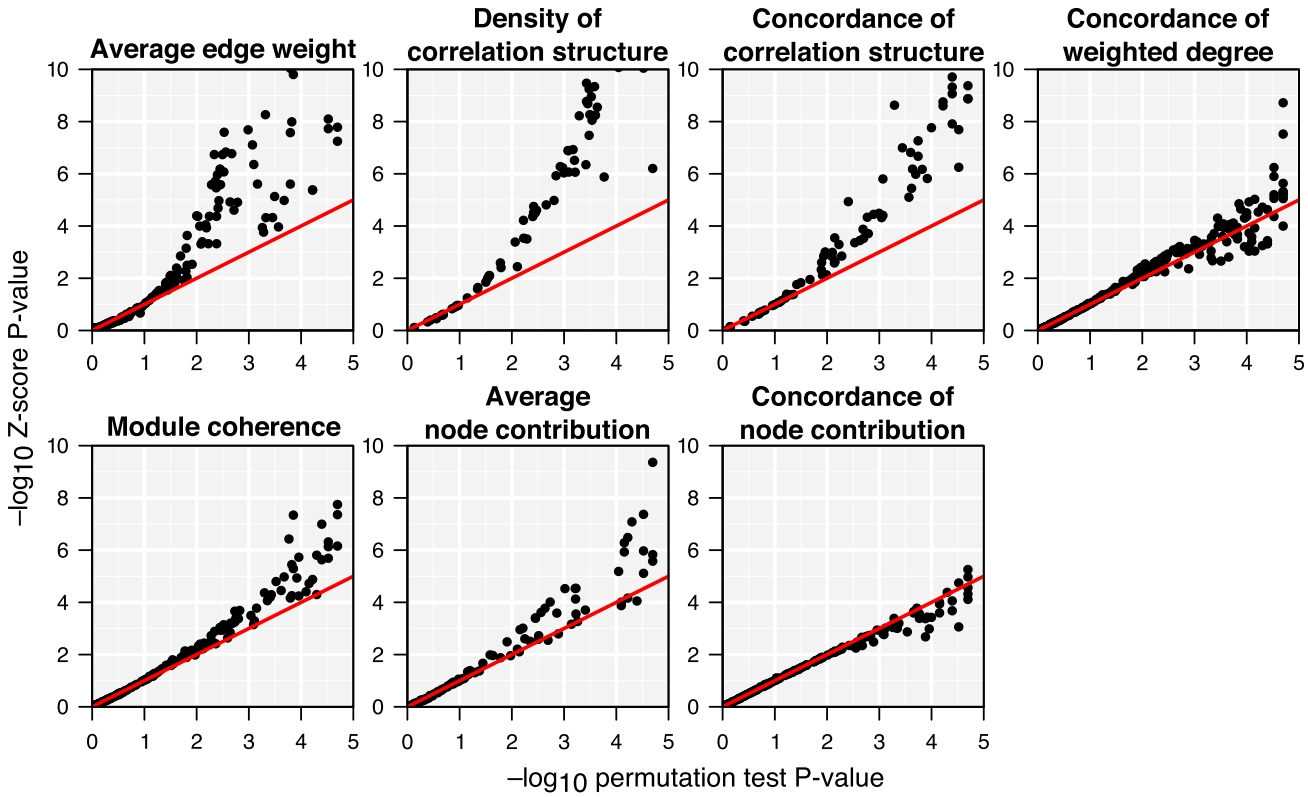


Figure 2. Comparison of Permutation p Values to Z Test p Values for Each of the 165 BxH Mice Modules when Tested in Their Three Non-discovery Tissues

The mean and standard deviation of the 495 null distributions were used to calculate the Z test p values. Null distributions were generated from 100,000 permutations. p values were plotted on a −log10 scale. Tests where the permutation test p value was incomparable to the Z test p values (2,229 of 3,465 tests where permutation test $p \leq 1 \times 10^{-5}$, the smallest permutation p value possible with 100,000 permutations) are not shown. Z tests with $p < 1 \times 10^{-10}$ are not plotted (25 data points for the average edge weight statistic with a minimum $p = 1 \times 10^{-63}$, 11 data points for the concordance of correlation structure statistic with a minimum $p = 1 \times 10^{-19}$, and 21 data points for the density of correlation structure statistic with a minimum $p = 1 \times 10^{-28}$).

discovery gene expression dataset containing negative and positive control modules of varying sizes and three test datasets with varying noise levels (Supplemental Experimental Procedures). Positive control modules were simulated to have identical topology in each test dataset, while negative control modules were simulated as random. We estimated permutation p values for each simulated module in each test dataset using NetRep. In total, we performed 3,000 tests for each module preservation statistic (10 modules × 3 test datasets × 100 simulations), estimating permutation p values with 10,000 permutations per test. At a significance threshold of $p \leq 1 \times 10^{-4}$ (the smallest possible p value that can be obtained from 10,000 permutations; Supplemental Experimental Procedures), NetRep was able to successfully detect preservation of positive control modules while being robust to false positives (Figure S6). Sensitivity varied by statistic but decreased as noise increased or module size decreased (Figure S6). The module preservation statistics were nearly always robust to false positives, with the exception of the module coherence statistic, which falsely detected preservation for large negative control modules (≥500 genes) in the presence of low and medium levels of simulated noise (Figure S6). These results indicate that NetRep is sensitive and can distinguish between true and false positives under most conditions.

Cross-Tissue Module Preservation in Mouse Transcriptomic Data

We next examined the preservation of each discovered module in other tissues by evaluating the permutation p values for each of the 495 null distributions (Experimental Procedures). We defined strong evidence for a module's preservation in another tissue as all test statistics achieving $p < 0.0001$, weak evidence if one or more, but not all, test statistics were $p < 0.0001$, and no evidence if no test statistics are $p < 0.0001$. The significance threshold of 0.0001 was chosen to Bonferroni adjust for the 495 tests performed for each preservation statistic. Figure 3 provides a summary view of the cross-tissue module preservation in the BxH mice, and Figure 4 shows the preservation evidence for each module in each non-discovery tissue.

We observed widespread preservation for modules in all four tissues (Figures 3A and 3B). 85 of 165 modules (52%) had strong evidence of preservation in at least one other non-discovery tissue, and 41 modules (25%) had strong evidence of preservation in all non-discovery tissues (Figure 3A). In contrast, only 21 modules (13%) had no evidence for preservation in any other tissue, suggesting tissue specificity of these modules (Figures 3A and 3C). In comparing NetRep results with those obtained through summary Z score and heuristic thresholds (Supplemental

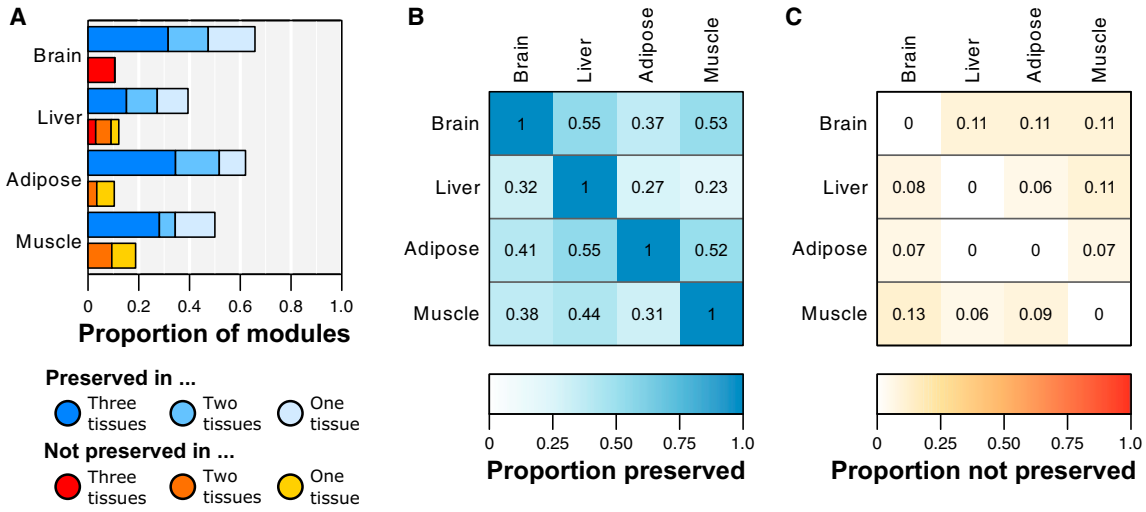


Figure 3. Summary of BxH Mice Cross-Tissue Module Preservation

(A) Summary of preservation evidence for BxH mice modules discovered in each tissue. Here, a module was considered preserved if it had strong evidence of preservation in another tissue, and not preserved if it had no evidence of preservation in another tissue.

(B) Tissue similarity based on the proportion of modules discovered in the row tissue with strong evidence of preservation in the column tissue.

(C) Tissue uniqueness based on the proportion of modules discovered in the row tissue with no evidence of preservation in the column tissue. Note that the heatmap entries in (B) and (C) cannot be read vertically.

Experimental Procedures; Langfelder et al., 2011), the two approaches mostly obtained similar levels of evidence for preservation (Table S1). However, differences in preservation were observed for 120 of the 495 (24%) module preservation tests. In terms of module preservation, 55 (24%) of the modules found to be strongly preserved by heuristics were classified as weakly preserved by NetRep. Similarly, 44 (54%) of the modules found to be not preserved by heuristic were classified as weakly preserved by NetRep (Table S1). Therefore, NetRep was more stringent in the evidence required to call a module either strongly preserved or not preserved.

The liver had the lowest proportion of modules preserved in at least one other tissue; however, it had twice as many modules in total than any other tissue. Many of these were small (<100 genes) and had only weak evidence for preservation (Figure 4). Only the brain and liver had any modules with no evidence for preservation in all three non-discovery tissues (Figures 3 and 4). These results were broadly consistent with recent results from the GTEx consortium, who observed high similarity between coexpression network modules across nine human tissues (including adipose and muscle tissues) (GTEx Consortium, 2015).

In total, NetRep found that 41 modules (10 discovered in adipose tissue, 12 in brain, 10 in liver, and 9 in muscle) were preserved in all non-discovery tissues. Analysis of Gene Ontology (GO) terms and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways for each module (Supplemental Experimental Procedures) showed that these were putative housekeeping modules, which were most frequently enriched for genes involved in translation, and to a lesser extent transcription and basic cellular functions, e.g., cell cycle, apoptosis, and DNA repair (Table S2). The putative housekeeping modules were most frequently enriched for genes coding for ribosomal proteins with 10 of 41 putative housekeeping modules enriched for the Ribosome pathway in KEGG (Table S2).

Multi-tissue Weight-Associated Modules

The BxH mice were bred to enhance differentiation of cardiovascular disease risk traits such as obesity and circulating lipids (Yang et al., 2006). Previous coexpression network analysis of the BxH mice focused on the identification of modules associated with mouse weight (Chen et al., 2008; Ghazalpour et al., 2006). We asked whether modules with strong evidence of preservation in more than one tissue ("multi-tissue" modules) were associated with obesity. We therefore tested each multi-tissue module's summary expression (first principal component; Experimental Procedures) for an association with mouse weight in any tissue for which there was strong evidence for its preservation (Supplemental Experimental Procedures). Significant association with weight was defined as $p < 0.0001$ (Bonferroni correction). Each regression model was adjusted for sex due to its strong effect on both mouse weight and gene expression (Fuller et al., 2007; Ghazalpour et al., 2006; Yang et al., 2006).

Of the 85 multi-tissue modules, 43 modules were significantly associated with mouse body weight in either their discovery tissue or in a tissue where it was strongly preserved, comprising 57 body weight associations in total (Table S3). Twenty-seven (32%) of these multi-tissue modules were also putative housekeeping modules. Weight was most frequently associated with modules in adipose tissue (28 of 57 associations) and liver tissue (24 of 57 associations). Notably, there were many cases where multi-tissue modules were not associated with weight in the tissue they were identified in, but displayed a significant association in one or more of the tissues in which they were preserved (Table S3). In total, 13 multi-tissue modules were associated with mouse weight in multiple tissues (Table S3). Notably, we observed different directions of weight association across tissues for five modules: i.e., in tissue A, an increase in weight was associated with a decrease in module summary expression, but in tissue B an increase in weight was associated with an

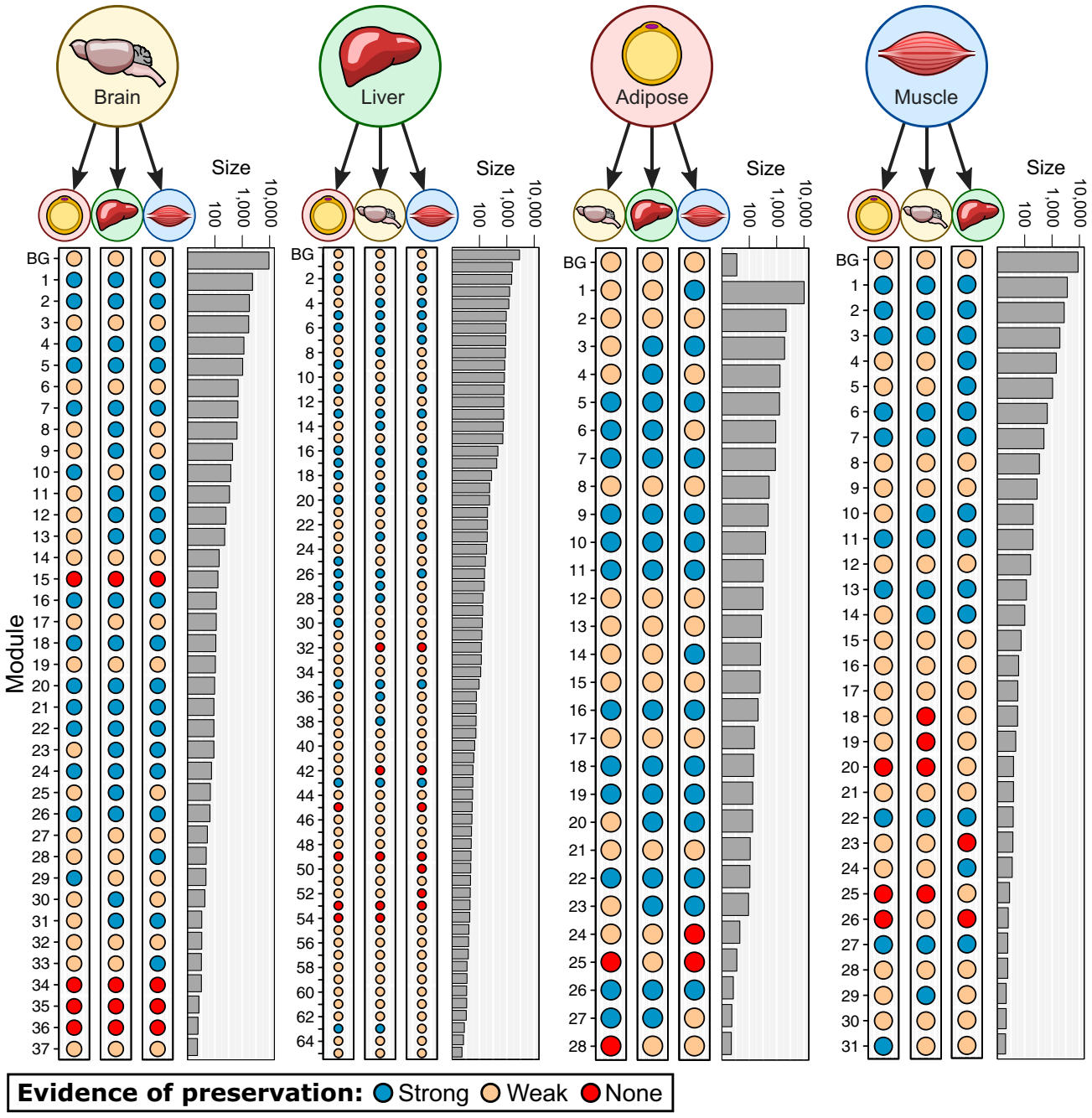


Figure 4. Evidence for Preservation for Each BxH Mice Module in Each Non-discovery Tissue

Module sizes (horizontal bar plots) are shown on a $\log_{10}$ scale. Modules (dots) are sorted according to module size and colored according to preservation evidence ("strong" evidence is blue, "weak" is yellow, "none" is red. "BG": the background module, which contains all nodes that did not cluster into a coexpression module in that tissue.).

increase in module summary expression or vice versa (Table S3). For these five modules, visualization of the network topology indicated the weight-associated differential summary expression reflected differential whole-module expression for two modules: liver module 35 and brain module 20. We focus on liver module 35 (LM35) for subsequent investigation (Figure 5).

LM35 is a putative housekeeping module consisting of 99 genes (permutation test $p \leq 1 \times 10^{-5}$ for all statistics in the brain, adipose, and muscle tissues). Consistent with previous analysis of the putative housekeeping modules, GO term and KEGG pathway enrichment indicated LM35 was primarily enriched for ribosomal genes involved in translation (Table S4). While a majority of probes in LM35 were specific to the custom microarray design and thus lacking gene annotation, 17 of its 24 annotated genes coded for ribosomal proteins (Table S5). Increased body weight was associated with increased LM35

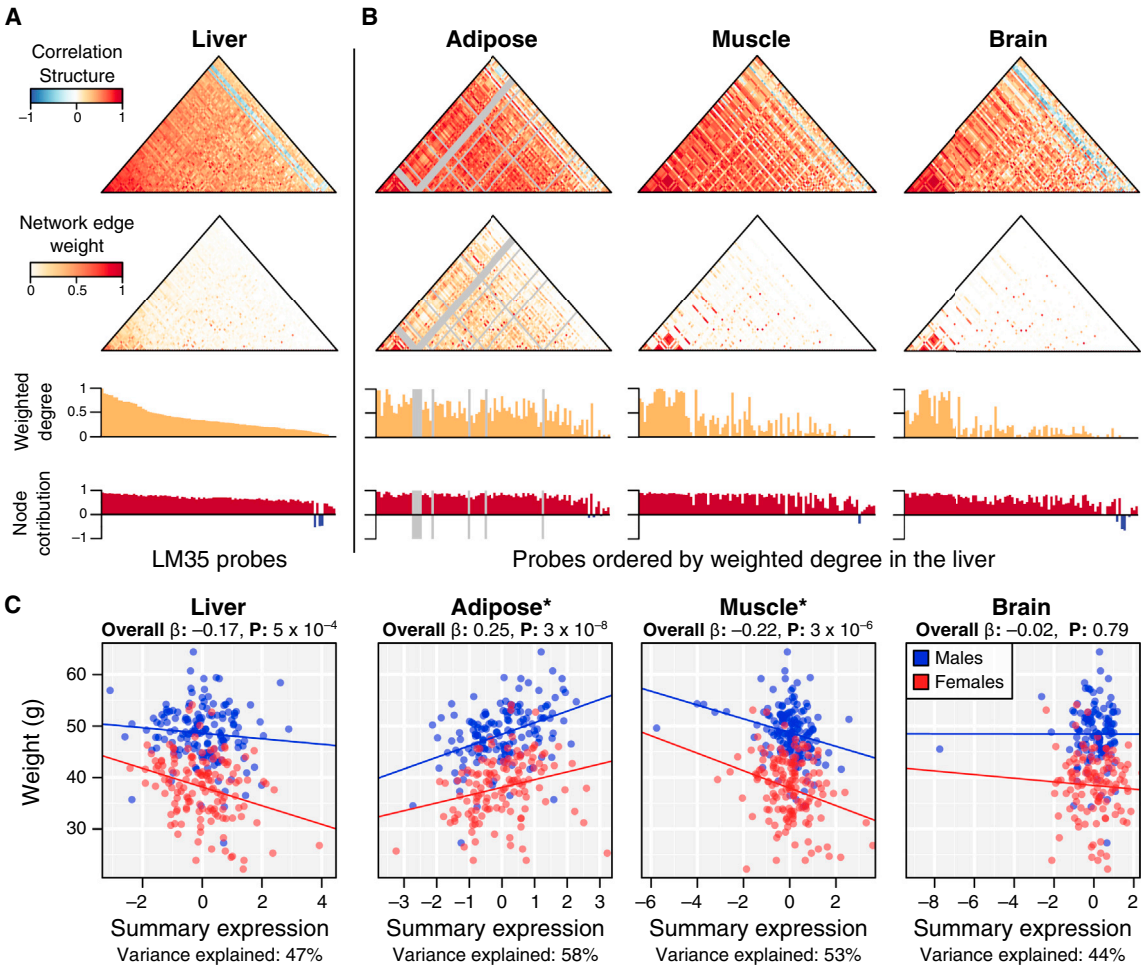


Figure 5. BxH Mice Liver Module 35

(A) Network topology in the liver (discovery tissue). From top to bottom: heatmap of the correlation structure (Pearson correlation), heatmap of the interaction network edge weights, normalized weighted degree (calculated within the module and normalized by the maximum value), and node contribution (Pearson correlation between each probe and the summary expression profile). Probes are ordered by descending order of weighted degree.

(B) Network topology in adipose, muscle, and brain tissues. Probes are ordered as in the liver tissue. Grey bars denote probes either missing, or not passing quality control, in the respective tissue.

(C) Scatter plots of standardized LM35 summary expression versus body weight. Points on the scatter plot are colored by sex (males in blue, females in red), and linear regression models were adjusted for gender (lines shown are fitted within genders). An "*" next to the tissue name indicates significant weight association in Table S3. Models were robust to outliers (mice with summary expression < −3 SD). Variance explained indicates the proportion of variance in liver module 35 (LM35) expression explained by the summary expression vector in each tissue (i.e., the module coherence).

expression in adipose tissue ($p = 3 \times 10^{-8}$) and decreased LM35 expression in muscle tissue ($p = 3 \times 10^{-6}$) (Figure 5). Consistent with this, its summary expression profile was negatively correlated between the adipose and muscle tissues (Pearson's $\rho = -0.13$) and the expression of 64 of its 99 probes were negatively correlated across the two tissues, suggesting that the relationship between body weight and LM35 was tissue specific—genes associated with weight were simultaneously upregulated in the adipose tissue and downregulated in the muscle tissue.

We subsequently tested 20 other cardiometabolic traits for association with LM35 expression in adipose and muscle tissues (Table S6). Consistent with the direction of the weight associations, increased insulin, total cholesterol, and total fat were associated with increased adipose expression and decreased muscle expression (false discovery rate [FDR] $q < 0.025$; Table S6). These changes in LM35 expression were also associated with a decrease in the ratio of glucose over insulin (Table S6). Increased LM35 adipose expression was associated with increased glucose, other fat, body length, and monocyte chemotactic protein-1 (MCP-1) (Table S6). On the other hand, decreased LM35 muscle expression was associated with increased abdominal fat, free fatty acids, total cholesterol, and LDL+VLDL, but a decreased ratio of HDL to LDL+VLDL (Table S6). Our findings indicate the tissue specificity of LM35 function and its relationships with phenotypes.

Overall, these analyses highlight that NetRep can be used to determine whether the relationships between genes are preserved, but separate investigation of preserved modules is required to determine whether module function is preserved. In

the case of multi-tissue analyses, this may elucidate differential inter-tissue module regulation.

Preservation of Gut Microbial Community Networks

To demonstrate the broader applicability of NetRep, we inferred microbial community networks in gut samples of 62 healthy adult men and 65 healthy adult women from the Human Microbiome Project (HMP) Consortium (Human Microbiome Project Consortium, 2012). The nodes in these networks corresponded to operational taxonomic units (OTUs), and we generated OTU networks with the commonly utilized SparCC approach (Friedman and Alm, 2012) (Experimental Procedures). From 152 distinct OTUs, we identified 17 and 21 communities of co-occurring OTUs in the male and female gut samples, respectively (Figures 6A and 6B). Using NetRep, we subsequently tested the preservation of the male gut communities in the female network and vice versa. Permutation p values were estimated from null distributions drawn from 10,000 permutations of OTU labeling in the respective test datasets (Figure 6C). We considered each module preservation statistic significant where $p < 0.001$ to Bonferroni adjust for the 38 tests performed for each statistic.

Unlike weighted gene coexpression networks, where all individuals in a population have more or less the same genes, OTU networks are relatively sparse due to the variable presence/absence of microbial taxa in the gut. Module sizes in OTU networks were also substantially smaller (range: 2–12 nodes). Thus, in applying the module preservation statistics to OTU networks, it was clear that some statistics would be more appropriate than others. We found that concordance of node contribution, concordance of correlation structure, and concordance of weighted degree statistics were not suitable for assessing preservation of these OTU modules. In addition to their small size, OTU modules tended to have uniform structure across nodes in terms of their SparCC correlation coefficient, node contribution, and weighted degree. This led to low values for these statistics in cases where the node contribution, SparCC correlation coefficient, and weighted degree were high across all nodes, due to dramatic changes in node rank caused by tiny variations in these values. Further, these module preservation statistics could not be evaluated where the node contribution, SparCC correlation coefficient, and weighted degree were identical for all nodes in a module. This always occurred for modules composed of two OTUs (Figure 6C), for which the weighted degree was always identical for both nodes and there was only one SparCC correlation coefficient. The sparsity of the network also meant the concordance of weighted degree could often not be calculated. This occurred where a module had no edges between any nodes in the test network (e.g., male module 7 in the female gut samples; Figure 6), which occurred frequently when generating null distributions for all modules, reducing the power of the permutation tests.

Therefore, in applying the module preservation statistics and NetRep to sparse networks and small modules, we recommend assessing module coherence, average node contribution, density of correlation structure, and average edge weight. Using these four statistics, we defined strong evidence for module preservation where all four were significant ($p < 0.001$), weak evidence if one or more were significant, and no evidence if none of these four were significant. Ignoring modules composed of only two OTUs, for which obvious false negatives were prevalent, we observed widespread preservation of microbial communities found in the men's gut samples, with 50% (4 of 8) of their OTU modules having strong evidence of preservation in the women's gut samples (Figure 6D). The women's gut microbial communities were less preserved, with 20% (2 of 10) of their OTU modules having strong evidence of preservation in the men's gut samples (Figure 6D). However, four of six women's gut OTU modules that had weak evidence of preservation (modules 2, 3, 9, and 10) were almost identical to OTU modules identified in the men's gut samples, suggesting comparative levels of preservation between women's and men's gut microbial communities.

DISCUSSION

Accurate and unbiased assessment of the replication and preservation of network modules requires permutation testing of network feature similarity. However, the current approach employs heuristics to assess significance due to the computational burden of these calculations (Langfelder et al., 2011). While heuristics may be appropriately employed for a small number of modules, the scale of module preservation and replication analyses now requires a rapid and statistically rigorous method to enable adjustment for multiple hypothesis testing, consequently allows confident investigation of the underlying biology. In this study, we have empirically shown that module preservation statistics are typically non-normal under the null hypothesis of non-preservation and thus have developed a rapid and efficient approach for assessing module preservation through permutation testing: NetRep.

In addition to assessment of reproducibility, module preservation analysis can be used to ask questions about conserved biological interactions and functions across spatial locations or species (Langfelder et al., 2011). Application of NetRep to a multi-tissue gene expression dataset showed widespread preservation of gene coexpression network modules across brain, adipose, liver, and muscle tissues in a BxH mouse cross. Housekeeping modules, those preserved in all four tissues, were enriched for genes involved in basic cellular processes, most notably ribosomal genes involved in translation. Subsequent investigation of multi-tissue modules associated with body weight revealed that preserved modules can exhibit differential intramodule expression across tissues, and we have identified a housekeeping module linked to obesity and insulin resistance with increased adipocyte expression and decreased muscle expression in overweight mice.

Previous studies have identified multi-tissue modules driving obesity in mice and humans, with concordant expression across tissues (Chen et al., 2008; Emilsson et al., 2008). Here, we found that multi-tissue modules may be differentially expressed across tissues with corresponding phenotypic differences. The liver module LM35 exhibited negative, positive, and negative associations with body weight in liver, adipose, and muscle tissues, respectively. Perhaps consistent with its tissue-specific directions of body weight association, LM35 was enriched for genes encoding ribosomal proteins, which maintain putative housekeeping functions. However, the gene set comprising LM35

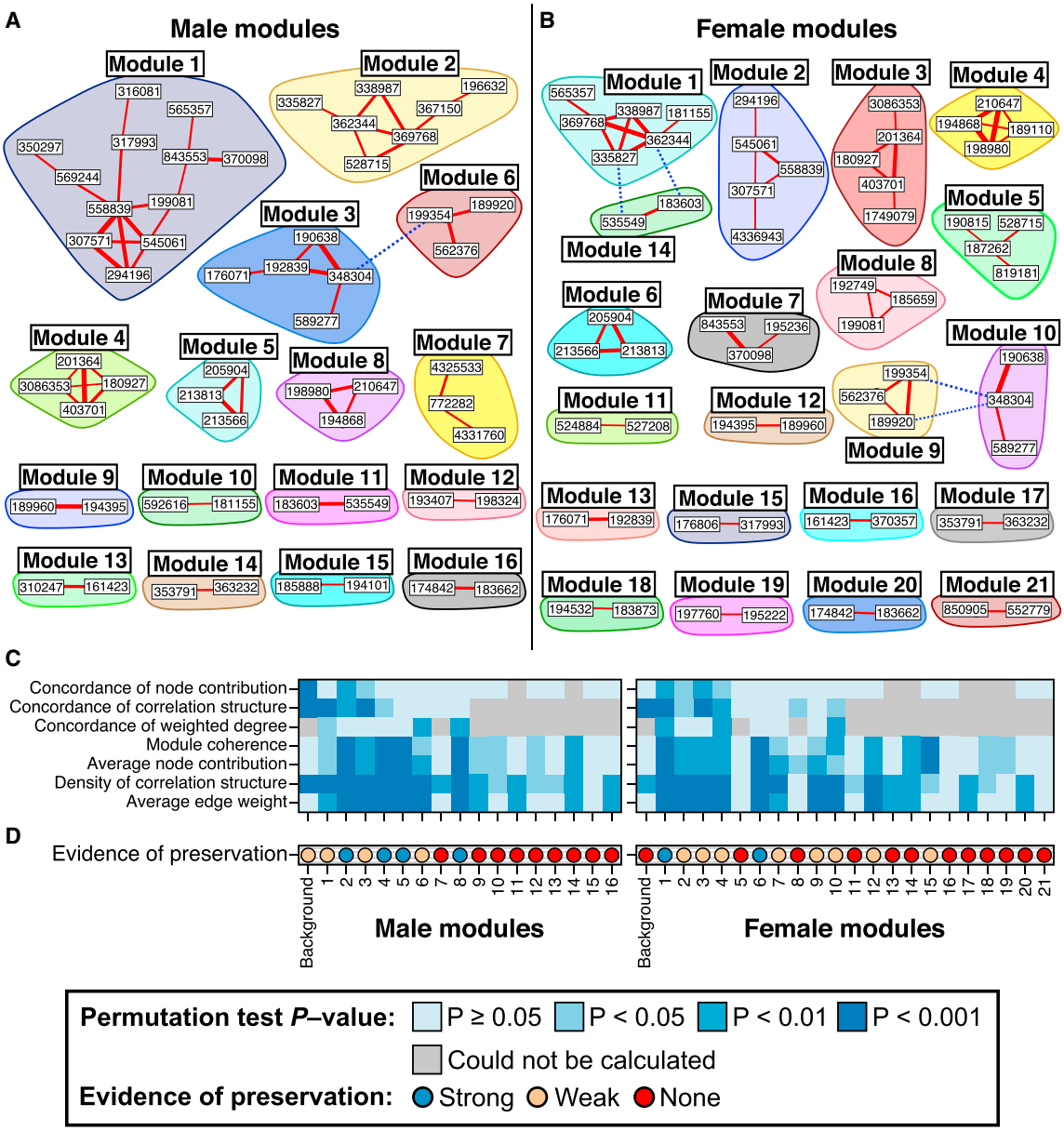


Figure 6. Preservation of Gut Microbial Communities across Males and Females Participating in the Human Microbiome Project

(A and B) Microbial communities inferred from the male (A) and female (B) gut samples. Nodes correspond to operational taxonomic units (OTUs) and numeric labels indicate their unique identifier. The colored shapes drawn around OTUs indicate community (module) assignment. Edge widths indicate strength of the correlation coefficient and color and linetype indicate positive (red, solid line) or negative (blue, dashed line) coefficients. Edge weights were defined as the absolute value of the correlation coefficient for the purpose of module detection (Experimental Procedures). Note that many OTUs present in male modules are also present in female modules and vice versa, and some male and female modules overlap.

(C) Heatmaps of permutation p values when assessing module preservation in the other sex. Permutation p values were estimated from null distributions generated from 10,000 permutations. Grey cells indicate tests where the permutation p value could not be calculated.

(D) Evidence of preservation for each OTU module in the other sex ("strong" evidence is blue, "weak" is yellow, "none" is red). See Table S7 for the taxonomic assignments of OTUs participating in modules with strong evidence of preservation.

was the only multi-tissue housekeeping module that exhibited significant patterns of differential body weight association. Furthermore, LM35 was associated with several obesity related traits, including a decreased ratio of glucose over insulin, suggesting an association with decreased insulin sensitivity. The link between insulin sensitivity, obesity, and adipocytes is well established (Hotamisligil et al., 1993; Kahn and Flier, 2000; Kahn et al., 2006), and consistent with this link, the adipose expression of LM35 was associated with circulating MCP-1 levels. MCP-1 has been shown to be secreted by adipocyte cells as well as overexpressed in obese mice, and it has been shown to decrease insulin-stimulated glucose uptake in vitro (Kanda

et al., 2006; Sartipy and Loskutoff, 2003). The phenotypic associations of LM35 across tissues may be explained by possible coexpression with obesity-linked genes in the adipose and muscle tissue, which did not coexpress with the module in the liver tissue where the module was identified.

We also showed that NetRep can be successfully applied to OTU networks derived for 16S microbiome data and have offered recommendations for dealing with the relative sparsity of these networks. In doing so, we identified several gut OTU modules that were preserved between men and women in the HMP data. Consistent with expectation, preserved modules largely involved multiple OTUs from the same genus (e.g., *Dialister*, *Bacteroides*, and *Ruminococcus* modules). A more diverse module comprising OTUs from the *Clostridiales* order, particularly *Faecalibacterium prausnitzii*, *Coprococcus*, *Butyrivibrio*, and *Clostridium*, was also preserved (male module 2, female module 5). *F. prausnitzii* has been linked to various human diseases, including inflammatory bowel disease, celiac disease, and obesity, and has been the subject of intense research to understand its specific functions, both individually and as part of communities, in the human gut (Miquel et al., 2013). Our analyses suggest that *F. prausnitzii* is part of a broader preserved *Clostridiales* community that may have functional consequences. Further studies in larger sample sizes may offer more power to detect additional members of this community, its variation across sexes, and its relevance to disease. Identification of gut microbial communities that change in composition between sexes may offer insight into diseases, such as irritable bowel syndrome, which have different prevalences in males and females (Canavan et al., 2014; Kassinen et al., 2007).

In recent years, studies have begun generating and analyzing datasets containing gene expression measured in dozens of tissues and cell types, for example, the Genotype-Tissue Expression (GTEx) Consortium (GTEx Consortium, 2015), the Immunological Genome (ImmGen) (Shay and Kang, 2013), and the Immune Variation (ImmVar) projects (De Jager et al., 2015). Similar scale studies are investigating microbiota spatiotemporally and in conjunction with other -omics data (Alivisatos et al., 2015; Human Microbiome Project Consortium, 2012; Integrative HMP (iHMP) Research Network Consortium, 2014). Already, multiple module preservation analyses have been performed on the GTEx pilot data (Melé et al., 2015; Pierson et al., 2015; GTEx Consortium, 2015), and here we have performed an initial preservation analysis of microbiome network modules between men and women. With large-scale expression studies increasing in scale and complexity, and the emergence of other types of datasets of similar scale, there is an urgent need for powerful and accurate statistical methodologies that quantify module replication and preservation. Here, we have presented an approach for rapid assessment of network module preservation and reproducibility that makes possible unbiased large-scale comparative analysis.

EXPERIMENTAL PROCEDURES

Full experimental procedures and data details can be found in the Supplemental Experimental Procedures. For the Humap Microbiome Project, details of institutional review boards are given in Human Microbiome Project Consortium (2012). For the mouse data, these are given in Yang et al. (2006).

Network Inference and Module Detection

Network inference and module detection were performed on a per-tissue basis for the BxH mouse cross using WGCNA v.1.43.1 with the default parameters (Langfelder and Horvath, 2008). First, the correlation structure (coexpression) for each tissue was calculated as the Pearson correlation coefficient between all probes passing quality control (Supplemental Experimental Procedures). Next, the network of interactions between probes was constructed by taking the element-wise absolute value of the correlation coefficient and exponentiating it to the smallest power such that the distribution of the weighted node degree (i.e., the sum of all edge weights for each node) of the resulting network was approximately scale free (scale-free topology criterion $R^2 > 0.85$) (Zhang and Horvath, 2005). This results in a dense, complete network where edge weights can take values between 0 and 1, most pairs of probes are connected with extremely small edge weights, and the comparatively few strongly correlated probes are connected with strong edge weights. The automated selection procedure selected the exponents of 12, 5, 4, and 12 for the brain, liver, adipose, and muscle tissues, respectively (Figure S7). Subsequently, the topological overlap dissimilarity (Zhang and Horvath, 2005) between probes was calculated and hierarchically clustered using the average linkage method. Hierarchically nested modules were identified from the results dendrogram using the dynamic tree cut algorithm with default parameters (Langfelder et al., 2008). Similar modules were merged together using an iterative process in which modules whose summary expression profile (first principal component, see below) clustered together (hierarchical average linkage) below a height of 0.2 were joined.

Network inference and module detection were performed separately for the HMP male and female gastrointestinal samples. First, 16S rDNA reads were clustered by sequence similarity ($\geq$97%) to representative sequences with known taxonomic assignments (Supplemental Experimental Procedures). Subsequent OTU tables were filtered to gastrointestinal samples collected on the first visit for 127 individuals. Next, the correlation structure between the 152 non-rare OTUs (Supplemental Experimental Procedures) was calculated using SparCC, a method for calculating unbiased correlation coefficients in sparse, compositional data (Friedman and Alm, 2012). The interaction network between OTUs was defined as the magnitude of the correlation where the SparCC correlation coefficient was significant at $p < 0.005$ in a bootstrap test. Modules were subsequently defined as groups of OTUs connected with significant positive SparCC correlation coefficients. The bootstrap p values were calculated using the estimator described by Phipson and Smyth (2010) (Supplemental Experimental Procedures), and the threshold $p < 0.005$ was chosen as it provided the best separation of OTUs into distinct modules for testing with NetRep.

Module Preservation

Seven statistics were used to quantify whether the relationships and correlation structure between nodes composing each module were replicated or preserved when measured in a different dataset (Langfelder et al., 2011). Here, we have renamed the statistics so that they are accessible to a wider audience and meaningful when applied to networks inferred from data sources (Table 1). Each module preservation statistic—their biological interpretation, application to different data sources, and network inference methods—are discussed in the Supplemental Experimental Procedures.

A permutation procedure was employed to characterize the distribution of each statistical test under the null hypothesis of non-replication and non-preservation. Specifically, each module preservation statistic was re-calculated when shuffling the node labels in the test dataset. The node labels in the discovery dataset were left unchanged. Nodes that were not present in both the discovery and test dataset were ignored both when calculating the module preservation statistics and when shuffling the node labels in the test dataset. Under the alternate hypothesis of replication/preservation, the test statistics calculated on the non-permuted dataset were expected to be higher than when calculated on random sub-graphs in the test dataset. Permutation p values were then calculated from these null distributions using the estimator described by Phipson and Smyth (2010), which provides a conservative estimate of the p value appropriate for multiple testing adjustment (Supplemental Experimental Procedures).

Module Summary Profiles

The summary profile for each module was calculated as the first principal component of module *w*. Specifically, each summary profile was calculated

Table 1. Definitions of the Module Preservation Statistics

NetRep Name	WGCNA Name	Definition
Module coherence	proportion of variance explained	$mean((cor(g_i^{[t](w)}, Eig_1^{[t](w)}))^2)$
Average node contribution	mean sign-aware module membership	$mean(sign(cor(g_i^{[d](w)}, Eig_1^{[d](w)})) \cdot cor(g_i^{[t](w)}, Eig_1^{[t](w)}))$
Concordance of node contributions	correlation of module membership	$cor(cor(g_i^{[d](w)}, Eig_1^{[d](w)}),\ cor(g_i^{[t](w)}, Eig_1^{[t](w)}))$
Density of correlation structure	mean sign-aware coexpression	$mean(sign(C^{[d](w)}) \cdot C^{[t](w)})$
Concordance of correlation structure	correlation of coexpression	$cor_{i \neq j}(C^{[d](w)}, C^{[t](w)})$
Average edge weight	mean adjacency	$mean_{i \neq j}(a_{ij}^{[t](w)})$
Concordance of weighted degree	correlation of intramodular connectivities	$cor((\sum_{i \neq j}^{j} a_i)^{[d](w)}, (\sum_{i \neq j}^{j} a_i)^{[t](w)})$

The NetRep name indicates the name of the statistic in the main text, while the WGCNA name indicates the name given to the statistics by Langfelder et al. (2011). Mathematical symbols are as follows: for n variables measured across m samples, G refers to the $m \times n$ matrix of observations, C refers to the $n \times n$ square matrix containing the pairwise correlation coefficients between variables, and A refers to the $n \times n$ square adjacency matrix denoting the connection strength (edge weight) between each pair of variables (nodes). Lowercase g, c, and a refer to individual elements of the matrices denoted by their respective uppercase letter. The superscripts [d] and [t] indicate whether the respective entity, formula, or network is obtained/calculated from the discovery or test dataset, respectively. The subscript letters i and j denote individual variables/nodes in module w. The superscript (w) indicates that the entity/formula that it is attached to is obtained/calculated on all nodes j (or all pairs of nodes i,j) in module w. For example, $g_i^{[t](w)}$ denotes a vector of observations for node i (which belongs to module w) in the test dataset, and $a_{ij}^{[d](w)}$ indicates the edge weight between nodes i and j (both of which belong to module w) in the discovery dataset. $Eig_1^{(w)}$ refers to the summary profile of the module w (first principal component; Experimental Procedures). The *sign* function evaluates to 1 if its argument is a positive value or −1 if its argument is a negative value. The *cor* function calculates Pearson's correlation coefficient between two vectors.

as the first eigenvector of a singular value decomposition of $G^{(w)}$. Two solutions exist for every eigenvector; both contain the same values, but with opposite signs. Thus, the summary profile is chosen as the eigenvector that is positively correlated with the average of $G^{(w)}$ across samples. This ensures that the eigenvector is oriented in the same direction as the data.

For interpretability, we refer to the summary profile as the "summary expression" profile when calculated on the BxH mice gene coexpression network modules, although this vector is commonly referred to as the "module eigengene" in the weighted gene coexpression network literature (Langfelder and Horvath, 2008). For the HMP gut community modules, we refer to the summary profiles as the community "summary abundance."

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, seven figures, and seven tables and can be found with this article online at http://dx.doi.org/10.1016/j.cels.2016.06.012.

AUTHOR CONTRIBUTIONS

M.I. and S.C.R. conceived and designed the study. S.C.R., S.W., G.A., L.G.F., K.E.H., and M.I. performed analyses. S.W. and K.E.H. contributed data and methods. S.C.R. and M.I. wrote the paper with input from all authors.

ACKNOWLEDGMENTS

This study was supported by funding from National Health and Medical Research Council (NHMRC) grant APP1062227. M.I. was supported by an NHMRC and Australian Heart Foundation Career Development Fellowship (no. 1061435). K.E.H. was supported by an NHMRC Career Development Fellowship (no. 1061409). S.R. was supported by an Australian Postgraduate Award and a PhD student top-up award from Victorian Life Sciences Computation Initiative (VLSCI). S.W. was supported by an Australian Postgraduate Award. G.A. was supported by an NHMRC Early Career Fellowship (no. 1090462).

The Mouse Model of Sexually Dimorphic Atherosclerotic Traits data were generated and contributed by Jake Lusis, Eric Schadt, and Merck Pharmaceutical through the Sage Bionetworks Repository.

Received: October 20, 2015
Revised: February 9, 2016
Accepted: June 29, 2016
Published: July 27, 2016

REFERENCES

Abraham, G., Bhalala, O.G., de Bakker, P.I.W., Ripatti, S., and Inouye, M. (2014). Towards a molecular systems model of coronary artery disease. Curr. Cardiol. Rep. *16*, 488.

Alivisatos, A.P., Blaser, M.J., Brodie, E.L., Chun, M., Dangl, J.L., Donohue, T.J., Dorrestein, P.C., Gilbert, J.A., Green, J.L., Jansson, J.K., et al.; Unified Microbiome Initiative Consortium (2015). MICROBIOME. A unified initiative to harness Earth's microbiomes. Science *350*, 507–508.

Barabási, A.-L., and Albert, R. (1999). Emergence of scaling in random networks. Science *286*, 509–512.

Barabási, A.-L., Gulbahce, N., and Loscalzo, J. (2011). Network medicine: a network-based approach to human disease. Nat. Rev. Genet. *12*, 56–68.

Bender, R., and Lange, S. (2001). Adjusting for multiple testing–when and how? J. Clin. Epidemiol. *54*, 343–349.

Boyle, A.P., Araya, C.L., Brdlik, C., Cayting, P., Cheng, C., Cheng, Y., Gardner, K., Hillier, L.W., Janette, J., Jiang, L., et al. (2014). Comparative analysis of regulatory information and circuits across distant species. Nature *512*, 453–456.

Cai, C., Langfelder, P., Fuller, T.F., Oldham, M.C., Luo, R., van den Berg, L.H., Ophoff, R.A., and Horvath, S. (2010). Is human blood a good surrogate for brain tissue in transcriptional studies? BMC Genomics *11*, 589.

Canavan, C., West, J., and Card, T. (2014). The epidemiology of irritable bowel syndrome. Clin. Epidemiol. *6*, 71–80.

Carlson, M.R.J., Zhang, B., Fang, Z., Mischel, P.S., Horvath, S., and Nelson, S.F. (2006). Gene connectivity, function, and sequence conservation: predictions from modular yeast co-expression networks. BMC Genomics *7*, 40.

Chen, Y., Zhu, J., Lum, P.Y., Yang, X., Pinto, S., MacNeil, D.J., Zhang, C., Lamb, J., Edwards, S., Sieberts, S.K., et al. (2008). Variations in DNA elucidate molecular networks that cause disease. Nature *452*, 429–435.

Collins, F.S., and Tabak, L.A. (2014). Policy: NIH plans to enhance reproducibility. Nature *505*, 612–613.

Dagan, T. (2011). Phylogenomic networks. Trends Microbiol. *19*, 483–491.

De Jager, P.L., Hacohen, N., Mathis, D., Regev, A., Stranger, B.E., and Benoist, C. (2015). ImmVar project: Insights and design considerations for future studies of "healthy" immune variation. Semin. Immunol. *27*, 51–57.

Emilsson, V., Thorleifsson, G., Zhang, B., Leonardson, A.S., Zink, F., Zhu, J., Carlson, S., Helgason, A., Walters, G.B., Gunnarsdottir, S., et al. (2008). Genetics of gene expression and its effect on disease. Nature *452*, 423–428.

Faust, K., and Raes, J. (2012). Microbial interactions: from networks to models. Nat. Rev. Microbiol. *10*, 538–550.

Friedman, J., and Alm, E.J. (2012). Inferring correlation networks from genomic survey data. PLoS Comput. Biol. *8*, e1002687.

Fuller, T.F., Ghazalpour, A., Aten, J.E., Drake, T.A., Lusis, A.J., and Horvath, S. (2007). Weighted gene coexpression network analysis strategies applied to mouse weight. Mamm. Genome *18*, 463–472.

Gerstein, M.B., Rozowsky, J., Yan, K.-K., Wang, D., Cheng, C., Brown, J.B., Davis, C.A., Hillier, L., Sisu, C., Li, J.J., et al. (2014). Comparative analysis of the transcriptome across distant species. Nature *512*, 445–448.

Ghazalpour, A., Doss, S., Zhang, B., Wang, S., Plaisier, C., Castellanos, R., Brozell, A., Schadt, E.E., Drake, T.A., Lusis, A.J., and Horvath, S. (2006). Integrating genetic and network analysis to characterize genes related to mouse weight. PLoS Genet. *2*, e130.

GTEx Consortium (2015). Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science *348*, 648–660.

Gustafsson, M., Nestor, C.E., Zhang, H., Barabási, A.-L., Baranzini, S., Brunak, S., Chung, K.F., Federoff, H.J., Gavin, A.-C., Meehan, R.R., et al. (2014). Modules, networks and systems medicine for understanding disease and aiding diagnosis. Genome Med. *6*, 82.

Hawrylycz, M.J., Lein, E.S., Guillozet-Bongaarts, A.L., Shen, E.H., Ng, L., Miller, J.A., van de Lagemaat, L.N., Smith, K.A., Ebbert, A., Riley, Z.L., et al. (2012). An anatomically comprehensive atlas of the adult human brain transcriptome. Nature *489*, 391–399.

Horvath, S., and Dong, J. (2008). Geometric interpretation of gene coexpression network analysis. PLoS Comput. Biol. *4*, e1000117.

Hotamisligil, G.S., Shargill, N.S., and Spiegelman, B.M. (1993). Adipose expression of tumor necrosis factor-alpha: direct role in obesity-linked insulin resistance. Science *259*, 87–91.

Human Microbiome Project Consortium (2012). A framework for human microbiome research. Nature *486*, 215–221.

Integrative HMP (iHMP) Research Network Consortium (2014). The Integrative Human Microbiome Project: dynamic analysis of microbiome-host omics profiles during periods of human health and disease. Cell Host Microbe *16*, 276–289.

Jeong, H., Mason, S.P., Barabási, A.L., and Oltvai, Z.N. (2001). Lethality and centrality in protein networks. Nature *411*, 41–42.

Kahn, B.B., and Flier, J.S. (2000). Obesity and insulin resistance. J. Clin. Invest. *106*, 473–481.

Kahn, S.E., Hull, R.L., and Utzschneider, K.M. (2006). Mechanisms linking obesity to insulin resistance and type 2 diabetes. Nature *444*, 840–846.

Kanda, H., Tateya, S., Tamori, Y., Kotani, K., Hiasa, K., Kitazawa, R., Kitazawa, S., Miyachi, H., Maeda, S., Egashira, K., and Kasuga, M. (2006). MCP-1 contributes to macrophage infiltration into adipose tissue, insulin resistance, and hepatic steatosis in obesity. J. Clin. Invest. *116*, 1494–1505.

Kassinen, A., Krogius-Kurikka, L., Mäkivuokko, H., Rinttilä, T., Paulin, L., Corander, J., Malinen, E., Apajalahti, J., and Palva, A. (2007). The fecal microbiota of irritable bowel syndrome patients differs significantly from that of healthy subjects. Gastroenterology *133*, 24–33.

Keller, M.P., Choi, Y., Wang, P., Davis, D.B., Rabaglia, M.E., Oler, A.T., Stapleton, D.S., Argmann, C., Schueler, K.L., Edwards, S., et al. (2008). A gene expression network model of type 2 diabetes links cell cycle regulation in islets with diabetes susceptibility. Genome Res. *18*, 706–716.

Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics *9*, 559.

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

Langfelder, P., Luo, R., Oldham, M.C., and Horvath, S. (2011). Is my network module preserved and reproducible? PLoS Comput. Biol. *7*, e1001057.

Langfelder, P., Mischel, P.S., and Horvath, S. (2013). When is hub gene selection better than standard meta-analysis? PLoS ONE *8*, e61505.

Lusis, A.J., and Weiss, J.N. (2010). Cardiovascular networks: systems-based approaches to cardiovascular disease. Circulation *121*, 157–170.

Melé, M., Ferreira, P.G., Reverter, F., DeLuca, D.S., Monlong, J., Sammeth, M., Young, T.R., Goldmann, J.M., Pervouchine, D.D., Sullivan, T.J., et al.; GTEx Consortium (2015). Human genomics. The human transcriptome across tissues and individuals. Science *348*, 660–665.

Miller, J.A., Horvath, S., and Geschwind, D.H. (2010). Divergence of human and mouse brain transcriptome highlights Alzheimer disease pathways. Proc. Natl. Acad. Sci. USA *107*, 12698–12703.

Miquel, S., Martín, R., Rossi, O., Bermúdez-Humarán, L.G., Chatel, J.M., Sokol, H., Thomas, M., Wells, J.M., and Langella, P. (2013). Faecalibacterium prausnitzii and human intestinal health. Curr. Opin. Microbiol. *16*, 255–261.

Phipson, B., and Smyth, G.K. (2010). Permutation p-values should never be zero: calculating exact p-values when permutations are randomly drawn. Stat. Appl. Genet. Mol. Biol. *9*, Article39.

Pierson, E., Koller, D., Battle, A., Mostafavi, S., Ardlie, K.G., Getz, G., Wright, F.A., Kellis, M., Volpi, S., and Dermitzakis, E.T.; GTEx Consortium (2015). Sharing and specificity of co-expression networks across 35 human tissues. PLoS Comput. Biol. *11*, e1004220.

Ritchie, S.C., Würtz, P., Nath, A.P., Abraham, G., Havulinna, A.S., Fearnley, L.G., Sarin, A.-P., Kangas, A.J., Soininen, P., Aalto, K., et al. (2015). The biomarker GlycA is associated with chronic inflammation and predicts long-term risk of severe infection. Cell Syst. *1*, 293–301.

Rotival, M., and Petretto, E. (2014). Leveraging gene co-expression networks to pinpoint the regulation of complex traits and disease, with a focus on cardiovascular traits. Brief. Funct. Genomics *13*, 66–78.

Sartipy, P., and Loskutoff, D.J. (2003). Monocyte chemoattractant protein 1 in obesity and insulin resistance. Proc. Natl. Acad. Sci. USA *100*, 7265–7270.

Schadt, E.E. (2009). Molecular networks as sensors and drivers of common human diseases. Nature *461*, 218–223.

Shay, T., and Kang, J. (2013). Immunological Genome Project and systems immunology. Trends Immunol. *34*, 602–609.

Stuart, J.M., Segal, E., Koller, D., and Kim, S.K. (2003). A gene-coexpression network for global discovery of conserved genetic modules. Science *302*, 249–255.

Stumpf, M.P.H., and Porter, M.A. (2012). Mathematics. Critical truths about power laws. Science *335*, 665–666.

van Nas, A., Guhathakurta, D., Wang, S.S., Yehya, N., Horvath, S., Zhang, B., Ingram-Drake, L., Chaudhuri, G., Schadt, E.E., Drake, T.A., et al. (2009). Elucidating the role of gonadal hormones in sexually dimorphic gene coexpression networks. Endocrinology *150*, 1235–1249.

Xia, K., Xue, H., Dong, D., Zhu, S., Wang, J., Zhang, Q., Hou, L., Chen, H., Tao, R., Huang, Z., et al. (2006). Identification of the proliferation/differentiation switch in the cellular network of multicellular organisms. PLoS Comput. Biol. *2*, e145.

Yang, X., Schadt, E.E., Wang, S., Wang, H., Arnold, A.P., Ingram-Drake, L., Drake, T.A., and Lusis, A.J. (2006). Tissue-specific expression and regulation of sexually dimorphic genes in mice. Genome Res. *16*, 995–1004.

Zhang, B., and Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. Stat. Appl. Genet. Mol. Biol. *4*, e17.

Cell Systems, Volume *3*

Supplemental Information

A Scalable Permutation Approach Reveals Replication and Preservation Patterns of Network Modules in Large Datasets

Scott C. Ritchie, Stephen Watts, Liam G. Fearnley, Kathryn E. Holt, Gad Abraham, and Michael Inouye

Supplemental Materials

Supplemental Experimental Procedures

Data collection and quantification

The BxH mouse cross is a publicly available dataset comprising samples from 334 mice bred on an Apolipoprotein E null background in order to enhance the differentiation of cardiovascular disease traits. Data collection protocols are extensively described in (Yang et al., 2006). Briefly, mice were fed on a high-fat, high-cholesterol diet from 8 weeks of age for 16 weeks and sacrificed at 24 weeks of age after fasting for 4 hours. Gonadal adipose (epididymal fat pad in males, perimetrial fat pad in females), whole brain, liver, and skeletal hamstring muscle were collected and immediately frozen in liquid nitrogen. Tissue samples were homogenised and RNA was extracted, prepared, and hybridised on a custom Agilent array as previously described (Schadt et al., 2003, 2008; Yang et al., 2006). Extensive physiological measurements were taken for each mouse, and metabolic measurements were quantified from plasma. Descriptions of trait measurement protocols can be found in (Ghazalpour et al., 2006), (Estrada-Smith et al., 2004), and (Meng et al., 2007). Gene expression data are available through GEO through the following identifiers. Brain: GSE3087, liver: GSE2814, adipose: GSE3086, and muscle: GSE3088. The full dataset was obtained from Sage BioNetworks at https://www.synapse.org/#!Synapse:syn4497.

The Human Microbiome Project (HMP) is a publicly available dataset comprising 16S rRNA gene sequence data quantified from 18 distinct body sites on 300 healthy adult volunteers aged 18–40 (Human Microbiome Project Consortium, 2012). For this study, we downloaded the raw sequence data from variable regions 1-3 of the 16S rRNA gene sequenced using the Roche-454 FLX Titanium platform across multiple body sites and visits for 173 individuals for a total of 3,363 samples. Data was downloaded from http://hmpdacc.org/HMQCP/healthy but is also available from the Sequence Read Archive at NCBI at http://www.ncbi.nlm.nih.gov/sra/?term=SRP002395.

Data processing and quality control

Individual transcript intensities were corrected for experimental variation and normalised within each tissue of the BxH mouse cross, then reported as the mean $\log_{10}$ ratio of each individual experiment relative to a pool of RNA comprised of equal aliquots of RNA from the respective tissues of 150 randomly selected mice (He et al., 2003; Yang et al., 2006). Gene expression microarray probes were subsequently quality controlled on a per-tissue basis. First, probes with more than 5% missingness were excluded, followed by samples with >5% probe-level missingness. Remaining missing information was imputed use a *K*-nearest neighbours algorithm using the R package *impute* (version 1.38.1). 22,808 probes and 295 samples from the adipose tissue, 22,950 probes and 249 samples from the brain tissue, 22,863 probes and 306 samples from the liver tissue, and 22,999 probes and 319 samples in the muscle tissue passed quality control. In total, 20,367 probes for 18,787 genes

corresponded to genes from the NCBI build 36/mm8 annotation release. 539 of the genes have been subsequently withdrawn, as they were not predicted in a later annotation. 3,205 probes on the array corresponded to UniGene clusters lacking annotation (*i.e.*, no corresponding Entrez Identifier) (Schadt et al., 2008).

Physiological and metabolic measurements with > 10% missingness were excluded from the analyses of the BxH mouse cross. 21 traits passed quality control: length, weight, abdominal fat, other fat, total fat, total cholesterol, unesterified cholesterol, free fatty acids, glucose, insulin, triglycerides, high density lipoprotein cholesterol (HDL), low density lipoprotein + very low density lipoprotein (LDL + VLDL), monocyte chemotactic protein-1 (MCP-1 / CCL2), aortic lesion size, aneurysm severity, medial aortic calcification, lateral aortic calcification, the ratio of glucose over insulin, the ratio of 100 x total fat over weight, the ratio of HDL over LDL + VLDL. The downloaded data were provided normalised using a natural log transform (MCP-1, insulin, triglycerides, HDL, the ratio of glucose over insulin, and the ratio of HDL over LDL + VLDL) or a square root transform for physiological traits with measurements of zero for many samples (aneurysm severity, medial aortic calcification, and lateral aortic calcification).

16S rDNA reads were clustered into operational taxonomic units (OTUs) at 97% similarity on all 3,363 samples using the *QIIME* pipeline (v1.9.1) (Caporaso et al., 2010a). Briefly; reads were clustered to reference sequences with known taxonomy from the Greengenes reference database (v3.8) (McDonald et al., 2012; Werner et al., 2012) where their sequences were ≥ 97% using *uclust* (Edgar, 2010). Next, representative sequences were picked for each OTU and aligned to the Greengenes Core reference alignment with *PyNAST* (Caporaso et al., 2010b), then chimeric sequences were detected with *ChimeraSlayer* and subsequently removed (Haas et al., 2011). The data was subsequently filtered to include only the first gastrointestinal (stool) sample collected for each individual and excluded if they had less than 500 reads. 62 males and 65 females were included in the final analysis. Rare OTUs were subsequently excluded following the recommendations of (Friedman and Alm, 2012) due to limitations of the network inference algorithm. OTUs were excluded if they were not abundant in at least 3 male and 3 female samples or had less than 2 reads on average in samples in which they were present. 152 OTUs were included in the final analysis.

Permutation and bootstrap test p-value estimation

Permutation test *p*-values for the module preservation statistics and bootstrap test *p*-values for the *SparCC* correlation coefficients were calculated using the estimator described by (Phipson and Smyth, 2010), which is implemented in the *statmod* R package:

$$p = \frac{b+1}{\mathrm{v}+1} - \int_0^{0.5/v_t+1} F(b;\mathrm{v},v_t)dv_t$$

Where v is the total number of permutations or bootstraps performed, b is the total number of permuted/bootstrapped test statistics as or more extreme than the observed value for the test statistic,

and v_t is the total number of possible unique permutations. The number of permutations required to perform a test at significance level α with this estimator is $\frac{1}{\alpha}$. This estimator provides a biased, yet, conservative upper bound estimate of significance. The commonly used estimator, $p = \frac{b}{v}$, systematically underestimates *p*-values, most notably by evaluating to 0 where $b = 0$ (Phipson and Smyth, 2010).

For the permutation tests on the module preservation statistics, *v* corresponded to the total number of permutations performed. Specifically, each module preservation statistic was calculated *v* times when permuting the node labels in the test dataset *v* times. The node labels in the discovery dataset were left unchanged, and nodes that were not present in both the discovery and test datasets were ignored both when calculating the module preservation statistics and when shuffling the node labels in the test dataset. Here, *b* was the number of tests in which the module preservation statistic was higher than when calculated on the non-permuted data. Calculation of v_t depended on whether the order of nodes within a permuted module affected the value of the module preservation statistic. For the *average edge weight* and *module coherence* statistics, the order of nodes did not matter, thus v_t was calculated as the number of possible $n^{(w)}$-combinations out of *n* nodes. For the other five module preservation statistics v_t was calculated as the number of possible $n^{(w)}$-permutations out of *n* nodes.

For the bootstrap tests on the *SparCC* correlation coefficients, *v* was the total number of bootstrapped OTU tables generated for each gender (1,000). Specifically, bootstrapped OTU tables were created by sampling each OTU with replacement across the gut samples separately for males and females. Next, bootstrapped SparCC correlation coefficients were calculated for each bootstrapped OTU table. Here, *b* was calculated separately for each *SparCC* correlation coefficient (i.e. each pair of OTUs) as the number of bootstrapped *SparCC* correlation coefficients for the respective pair of OTUs that were more extreme than the *SparCC* correlation coefficient calculated from the original OTU table. This was performed as a two-sided test: the absolute value of all correlation coefficients was used to calculate *b*. Due to sparsity in the OTU tables v_t was calculated for each SparCC correlation coefficient separately, as the product of total number of possible permutations for the respective pair of OTUs. The total number of possible permutations for each OTU was calculated accounting for repetition and indistinguishable elements (e.g. abundances of 0 across multiple samples). For our application, v_t varied between 1×10^{10} and 5×10^{125}.

For all permutation and bootstrap tests performed v_t was sufficiently large in comparison to *v* such that the estimator reduced to $p = \frac{b+1}{v+1}$ as the integral term in the estimator, at maximum, evaluated to numerical values many orders of magnitudes smaller than the smallest possible *p*-value that could be obtained with *v* permutations/bootstraps.

Module preservation statistics

Each of the seven module preservation statistics captures the preservation of a different aspect of a module's network topology (Langfelder et al., 2011). Although Langfelder *et al.* provided an aggregate score across all seven statistics to call module preservation (see below) (Langfelder et al.,

2011), similar aggregation of permutation test statistics is not meaningful due to the heterogeneity of each statistics null distribution. Further, it is important to consider the utility and appropriateness of each statistic for each analysis, which depends on the type of network inferred, the type of data the network was inferred from, and whether the topological property it captures is meaningful to the study. For example, in our analysis of the BxH mice modules we examined the preservation of weighted gene coexpression network modules, for which the statistics were designed (Langfelder et al., 2011), and we had no strong prior as to the particular importance of any statistic. Thus, we took a conservative approach and required all seven to be significant when calling preservation. To aid statistic selection we provide a discussion on their biological interpretation, as well as appropriateness for different types of networks and data types. Definitions for each statistic are provided in **Table 1**.

The *module coherence*, *average node contribution,* and *concordance of node contribution* are all calculated from a module's *summary profile*, which is the eigenvector of the 1st principal component of all observations for a module's nodes across samples (**Experimental Procedures**). For gene coexpression modules this can be interpreted as the summary expression profile, while for modules defined as OTU communities this can be interpreted as a summary of community abundance.

The *module coherence* measures the proportion of variance in the $m \times n^{(w)}$ matrix of observations made for the module's nodes across *m* samples in the test dataset explained by the module's summary profile. *I.e.* it measures whether the module's data are more coherent than expected by chance in the test dataset. It has previously been referred to as the *proportion of variance explained* (Langfelder et al., 2011).

The *node contribution* is calculated as the Pearson correlation coefficient between each node's vector of observations and the module's *summary profile*, *i.e.* it is a measure of how strongly each node contributes to the module's summary profile. It has previously been referred to as the *module membership* (Langfelder et al., 2011). For modules identified through weighted gene coexpression network analysis (WGCNA) (Zhang and Horvath, 2005) the *node contribution* is typically positive. Negative *node contributions* occur where a gene is differentially expressed in comparison to the rest of the module.

The *average node contribution* measures the average contribution of each node to the module's *summary profile* in the test dataset. Importantly, each node's *node contribution* in the test dataset is multiplied by the sign of its *node contribution* in the discovery dataset (**Table 1**). Thus, nodes that have strong negative *node contribution* scores contribute to a high *average node contribution*, while nodes that are negatively correlated with the *summary profile* in one dataset but not the other penalise the *average node contribution* towards zero. Previously it has been referred to as the *mean sign-aware module membership* (Langfelder et al., 2011). A high *average node contribution* has a similar interpretation to a high *module coherence*: that the data remains coherent in the test dataset.

The *concordance of node contribution* measures the Pearson correlation coefficient between the vectors of *node contribution* calculated the discovery and test datasets (**Table 1**). Previously it has

been referred to as the *correlation of module membership* (Langfelder et al., 2011). A high *concordance of node contribution* indicates that nodes contribute similarly to the *summary profile* in both datasets, *i.e.* that the *summary profile* summarises the nodes similarly in both datasets, and that associations with the *summary profile* have similar biological meaning (in terms of relative *node contribution*) in both datasets.

Caution should be used when applying and interpreting these statistics to sparse data, *e.g.* 16S rDNA read counts and OTU abundances. First, the singular value decomposition used by NetRep to calculate the *summary profile* can sometimes fail, throwing an error. This can reduce the number of observations in the null distributions of the *module coherence*, *average node contribution*, and *concordance of node contribution* statistics. Second, the calculation of the *node contribution* relies on the Pearson correlation coefficient, which is inappropriate for sparse data. The *node contribution* will therefore be systematically underestimated for nodes with sparse observations (*i.e.* a value of 0 for most samples). This can lead to underestimated values for the *average node contribution* and *module coherence* (which is calculated as the sum of *node contribution* squared; **Table 1**, (Langfelder et al., 2011)).

The *concordance of node contribution* should also be interpreted in the context of the *module coherence* and *average node contribution* statistics. Where the *module coherence* and *average node contribution* are low, a high *concordance of node contribution* is unlikely to be biologically meaningful. Conversely, if the *module coherence* and *average node contribution* are very high, the *concordance of node contribution* will be low and have a high permutation test *p*-value where the *node contribution* have similar values across all nodes composing a module. In this case tiny variations in the *node contribution* can lead to dramatic changes in the relative rank of the *node contribution*, leading to a very low *concordance of node contribution*. This tiny variations is unlikely to be biologically meaningful, thus the *concordance of node contribution* may be incorrectly classified as not preserved.

The *concordance of correlation structure* and *density of correlation structure* are both calculated from the user-provided $n \times n$ square matrix containing the correlation coefficients between each pair of nodes in the dataset. When applied to WGCNA defined modules, this matrix is typically referred to as the *coexpression* matrix, and the statistics are referred to as the *correlation of coexpression* and *mean sign-aware coexpression* respectively (Langfelder et al., 2011). The *concordance of correlation structure* measures whether the correlation coefficients are similar between the discovery and test datasets. This is calculated by flattening the $n^{(w)} \times n^{(w)}$ matrix into a single vector of correlation coefficients (ignoring the diagonal), then calculating the Pearson correlation coefficient between both vectors (**Table 1**). The *density of correlation structure* measures whether the nodes are strongly correlated in the test dataset. This is calculated as the mean correlation coefficient in the test dataset, multiplied by the sign of the correlation coefficient in the discovery dataset (**Table 1**). Thus, strong negative correlation contribute to a high *density of correlation structure*, while pairs of nodes that are

positively correlated in one dataset and negatively correlated in the other penalise the *density of correlation structure* towards zero.

The *concordance of correlation structure* should be interpreted in the context of the *density of correlation structure*. A high *concordance of correlation structure* is not likely to be biologically meaningful where the *density of correlation structure* is low and has a high permutation test *p*-value. Conversely, a low *concordance of correlation structure* may arise where the *density of correlation structure* is high where all correlation coefficients are large. In this case tiny variations between correlation coefficients can lead to dramatic changes in the relative rank of node pairs, leading to a low *concordance of correlation structure*. This tiny variations is unlikely to be biologically meaningful, thus the *concordance of correlation structure* may be incorrectly classified as not preserved.

The *average edge weight* and *concordance of weighted degree* are both calculated from the network inferred from the data. In the case of weighed gene coexpression networks, this is defined through a power transform on the *correlation structure* (Zhang and Horvath, 2005). The power transform acts as a soft-threshold on the correlation coefficients, thus the resulting networks are weighted and complete: every node has an edge to every other node, with a weight between 0 and 1 indicating correlation strength.

The *average edge weight* measures the average connection strength between nodes within a module in the test dataset. For modules defined through weighted gene coexpression network analysis (WGCNA), this measures the average gene–gene interaction strength, and is typically called the *mean adjacency* or *module density* (Langfelder et al., 2011). For modules detected through WGCNA's clustering algorithm (Langfelder et al., 2008), one would expect the *average edge weight* to be higher than expected by chance in the test dataset where the module is preserved (Langfelder et al., 2011). This indicates that the module is more tightly connected on average than the rest of the interaction network. Small modules in weighted gene coexpression networks tend to be very tightly connected and embedded within larger, more loosely connected modules.

The *concordance of weighted degree* measures whether the *weighted degree* for each node in a module is correlated across the discovery and test datasets. The *weighted degree* of each node is calculated as the sum of the node's edge weights to all other nodes in the network. This is typically referred to as the *connectivity* when calculate on modules identified through WGCNA, and the module preservation statistic is referred to as the *correlation of intramodular connectivities* (Langfelder et al., 2011). For those familiar with WGCNA, we note that *connectivity* typically has a completely different meaning in network theory (Newman, 2010). For modules defined through WGCNA, the *weighted degree* is typically used as a metric of relative biological importance to a module. Genes with a high *weighted degree* typically play a central role in the module's biological function (Horvath and Dong, 2008; Langfelder et al., 2011; Zhang and Horvath, 2005). The *concordance of weighted degree* therefore measures whether genes that are important to a module remain important in the test dataset

(Langfelder et al., 2011). For modules that are preserved, taking the average *weighted degree* across datasets can provide a more robust measure of gene importance (Langfelder et al., 2013).

Both the *average edge weight* and *concordance of weighted degree* are calculated assuming edge weights have positive values. Importantly, networks with a mixture of positive and negative edge weights the negative edges will penalise both statistics towards zero, leading to artificially low permutation test *p*-values where a module is preserved. Note that the *average edge weight* counts edges with zero weights when calculating the *mean*, thus in sparse weighted networks it provides a combined metric of the proportion of node pairs connected by an edge as well as the average strength of those connections. For unweighted networks (*i.e.* edges have weight 1 where they exist and 0 otherwise) the *average edge weight* instead quantifies the proportion of node pairs that are connected with an edge, and the *weighted degree* instead becomes the node degree (*i.e.* the number of nodes each node is connected to). For directed networks the interpretation of the *average edge weight* remains unchanged, while the *concordance of weighted degree* (as implemented in *NetRep*) will only measure the *in-degree* of each node. The permutation test *p*-values quantify whether the *average edge weight* and *concordance of weighted degree* are higher than expected by chance, *i.e.* the test assumes the module is more densely connected in terms of both edge numbers and edge strength than the rest of the network.

Caution should be used when applying and interpreting either statistic to sparse networks, particularly those inferred via a hard threshold. The choice of threshold strongly influences both the *average edge weight* and *concordance of weighted degree*. For example, a threshold too strict can lead to an artificially low *average edge weight* and *weighted degree* (and artificially high permutation test *p*-values) due to edges that only just fail to pass the threshold in the test dataset. It may be beneficial to relax the thresholds used in network inference when assessing module preservation.

The *concordance of weighted degree* has several additional limitations to be aware of when applied to sparse networks. First, it relies on the Pearson's correlation coefficient, which is inappropriate when comparing sparse vectors, which may arise when calculating the *weighted degree* on sparse networks. For example, if many nodes in a module have a *weighted degree* of 0, then the correlation coefficient will be biased (towards zero). In practice, this bias will be present for most null distribution observations as well, so should have little effect on the permutation test *p*-value. Note that the same is not true for the *node contribution*-based statistics, as there may be a mixture of nodes with sparse and dense data. Sparse networks also lead to problems when calculating permutation test *p*-values due to missing values in the null distribution where all nodes have a *weighted degree* of 0 in the permuted subset. These are ignored in calculations of the permutation test *p*-value, leading to high *p*-values in cases of module preservation, simply due to lack of non-missing observations in the null distribution. This is also a problem for modules where every node has the same *weighted degree*, for which the *concordance of weighted degree* will always be a missing value.

The *concordance of weighted degree* should be interpreted in the context of the *average edge weight*. Where the *average edge weight* is low and has a high permutation test *p*-value, a high *concordance of*

weighted degree is not likely to be biologically meaningful. Conversely, a low *concordance of weighted* may arise where the *average edge weight* is high where the *weighted degree* is similar for all nodes. In this case tiny variations in each node's *weighted degree* can lead to dramatic changes in the relative rank of nodes leading to a low *concordance of weighted degree.* This tiny variations is unlikely to be biologically meaningful, thus the *concordance of weighted degree* may be incorrectly classified as not preserved.

When assessing module preservation we recommend visualisation of the module topology in both datasets as well as calculation of the permutation test *p*-values. Visualisation helps to interpret module preservation where the permutation tests *p*-values are significant, and can help to identify which statistics are appropriate. For example, visualisation can be used to determine whether the *concordance of node contribution*, *concordance of correlation structure,* and *concordance of weighted degree* are likely to be informative or spuriously non-significant due to lack of variability in the *node contribution*, *correlation coefficients*, and *weighted degree* respectively.

Heuristic thresholds for module preservation

For comparison with previous methodology (**Table S1**), Z-scores, summary statistic, and heuristic thresholds were calculated as previously described (Langfelder et al., 2011) for each test of the BxH mouse cross modules. Standardised Z-scores for each module preservation statistic were calculated using the mean and standard deviation estimated from the null distributions, and the combined summary score was calculated as:

$$Z_{Summary} = \frac{median\begin{pmatrix} average\ edge\ weight, \\ module\ coherence, \\ density\ of\ correlation\ structure, \\ average\ node\ contribution \end{pmatrix} + median\begin{pmatrix} concordance\ of\ correlation\ structure, \\ concordance\ of\ weighted\ degree, \\ concordance\ of\ node\ contributions \end{pmatrix}}{2}$$

The heuristic thresholds for significance were defined by (Langfelder et al., 2011) as: strong evidence for module preservation if $Z_{summary} > 10$, weak evidence if $2 < Z_{summary} < 10$, and no evidence of preservation if $Z_{summary} \leq 2$.

Simulations

We simulated gene expression data containing positive and negative control modules then tested their preservation in simulated datasets with varying amounts of noise. For 100 simulations, we simulated four datasets, each containing 10,000 genes and 100 samples. A discovery dataset, in which positive and negative control modules of sizes 10, 50, 100, 500, and 1,000 genes were simulated as:

$$G_{simulated}^{(w)} = E^{(w)}\, r_i + \sqrt{1 - r_i^2}\, \varepsilon$$

Where $E^{(w)}$ is the simulated module's summary vector, r is the vector of simulated node contributions for each gene in the simulated module, and ε is the error term, drawn from a normal distribution with mean of 0 and standard deviation of 1. $E^{(w)}$ and r were simulated by bootstrapping (sampling with

replacement) the corresponding vectors in a randomly selected BxH mice liver module of similar size to the simulated module. Genes not in any module were simulated by bootstrapping 100 samples from the liver tissue expression data for 6,680 randomly selected genes then adding the same level of statistical noise, ε, to each observation. Three test datasets were simulated with "low", "medium", and "high" levels noise, where ε was drawn from a normal distribution with mean of 0, and standard deviation of 1, 2, and 5 respectively. In these datasets, genes composing the positive control modules were simulated using identical node contributions r, and a summary vector $E^{(w)}$ drawn by bootstrapping 100 samples from the summary vector of the same liver module selected when simulating the discovery dataset. Genes composing the negative control modules were simulated by bootstrapping liver tissue expression levels of randomly selected genes then adding the statistical noise. Permutation test *p*-values were then estimated using null distributions drawn from 10,000 permutations. Results of the simulations are given in **Figure S6.**

Module GO term and KEGG pathway enrichment

Enrichment of Gene Ontology (GO) terms (Ashburner et al., 2000) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways (Kanehisa and Goto, 2000) for each BxH mice module were determined through over-representation analysis (**Tables S2 and S4**). Briefly, a hypergeometric test was performed for each GO term and KEGG Pathway annotating at least two module genes. Annotations were considered significant at the false discovery rate (FDR) corrected significance threshold of $P < 0.05$ within each annotation type (KEGG pathway, GO biological process, GO molecular function, GO cellular component).

GO term annotations for *Mus musculus* genes were obtained from the Genome-wide Annotation for Mouse Database provided through the R package *org.Mm.eg.db* (version 2.14.0). Definitions for each GO term were obtained from the GO database through the *GO.db* R package (version 2.14.0). KEGG Pathway annotations for *Mus musculus* genes were retrieved from the KEGG database using the *KEGGREST* package (version 1.4.1).

Statistical tests

Associations between body weight and each multi-tissue module in the BxH mice were assessed through linear regression of weight on the module's *summary profile* (1st principal component, **Experimental Procedures**) adjusting for sex (**Table S3**). Multi-tissue modules were those with strong evidence for preservation in any other tissue. Associations with body weight were assessed in the module's discovery tissue as well as any tissues for which the module had strong evidence of preservation. Effect sizes correspond to change in standard deviation (SD) of body weight per SD increase in the module's summary profile in the corresponding test tissue. An association was considered significant at a Bonferroni-corrected threshold of $P < 0.0001$, adjusting for the 273 tests.

Associations between BxH mice liver module 35 (LM35) and the cardiometabolic traits, excluding weight, were assessed through a linear regression model of each trait on the module's summary profile calculated in the adipose and muscle tissue, adjusting for sex (**Table S6**). In total 20 associations were assessed through linear regression of each trait on LM35 adipose expression, and 20

associations were assessed through linear regression of each trait on LM35 muscle expression. Effect sizes correspond to change in LM35 tissue expression (SD-units) per SD-increase of each trait. False-discovery rate (FDR) correction was applied within each tissue to the resulting linear model *p*-values. We considered an association significant at FDR $q < 0.025$ (adjusting for the two tissues).

Implementation and performance of NetRep

The permutation procedure in NetRep is implemented as a multi-threaded C++ program that makes use of highly optimised linear algebra subroutines provided by the C++ library armadillo (Sanderson, 2010). This is wrapped by an R package which interfaces with the C++ code through the Rcpp and RcppArmadillo packages (Eddelbuettel and François, 2011; Eddelbuettel and Sanderson, 2014). The permutation procedure is run on datasets in which the correlation structure and interaction networks have already been computed by the user: these are required as input. NetRep dynamically loads and unloads each dataset so that only the data, correlation structure, and interaction network matrices for one dataset are ever in memory at any point in time. These are stored in heap memory during the permutation procedure so that they can be simultaneously accessed by all parallel threads. Each thread thus requires only additional memory for calculation of the network properties in the test dataset at each permutation. Taken together, NetRep is scalable: its memory usage is independent of the number of datasets analysed, runtime can be divided linearly across multiple cores, and memory usage is fixed for each dataset: each thread requires only a small amount of memory in most cases.

Computational performance of NetRep was evaluated through comparison of runtime and memory usage to the modulePreservation function implemented in the R package WGCNA (Langfelder et al., 2011). Both methods were used to generate null distributions for the 38 BxH mice brain modules from permutations of node labels in the BxH liver tissue (**Figure S2**). Module sizes ranged from 23 nodes to 9,487 nodes (median: 106 nodes, **Figure S1**) and the total network size was 22,528 nodes (probes present in both datasets).

NetRep was 11-times faster than WGCNA both when run on a single core (**Figure S2a**) and when run in parallel (**Figure S2b**). NetRep took an average of 7.4 seconds per permutation per core compared to 93 seconds for WGNCA. For all points in both panels, 4 minutes, 20 seconds were taken by NetRep prior to the permutation procedure to load and unload each dataset during input checks and calculation of network properties in the brain tissue. Runtime of permutation procedure decreased linearly with the number of cores used. Memory usage of NetRep was 8.13 GB for a single core, and 12.28 GB for 64 cores: Each thread used 68 MB for storage of the network properties of permuted modules in the liver tissue at each permutation. In contrast, WGCNA used 800 MB of RAM per core to store the gene expression data and dynamic calculations of the correlation structure and interaction networks for each permutation. No memory was shared between cores by WGCNA. Memory usage of WGCNA exceeded that of NetRep when >16 cores were in use. Although the correlation structure and interaction network matrices are not required for WGCNA's permutation procedure, these will typically be pre-calculated and saved by the user for downstream analyses prior to assessing module preservation. NetRep is more flexible since it requires these matrices as input: the user may use any algorithm for network construction.

The overall runtime, RAM usage, and disk requirements of NetRep are a function of the number of permutations required, the total size of each network, sample size within each dataset, and distribution of module sizes. Disk requirements are a function of total network size of each dataset, as are the RAM requirements to store the matrices of the test network in the memory shared between threads. The runtime of each permutation and the additional RAM usage per thread are a function of the module sizes within each dataset. Both are proportional to the sum of module sizes squared: although runtime is also influenced by the sample size of the test dataset. This means that large modules can dramatically increase runtime and memory usage. Our worst-case scenario was that of the adipose tissue, for which module 1 contained 10,119 genes (the largest module in any other tissue was 3,612 genes). For adipose tissue modules, NetRep took an average of 48 seconds per permutation and required 500 MB of RAM per thread. Across all tissue comparisons the median time taken for each permutation was 15 seconds.

In practice, total runtime depends on the total number of tests performed: *i.e.* the total number of modules multiplied by the number of datasets their preservation is tested in. This number dictates the number of permutations required to appropriately control for multiple testing, although we recommend running at least 10,000 permutations for each test. In a pairwise comparison setting runtime exponentially increases due to the multiplication of the number of tissue comparisons by the number of permutations required for each comparison (**Figure S2c**).

Several approaches can be taken to reduce runtime of NetRep. Runtime is most dramatically reduced by excluding large modules or filtering to the top most connected genes. Pairwise analysis of the four tissues restricted to modules with fewer than 250 nodes (109 of the 165 modules) reduced total runtime from 19 hours to 2 hours on 40 cores. This was due to a 14-fold reduction in the average time taken per permutation from 22 seconds to 1.5 seconds. Dimensionality reduction prior to network inference will also achieve this goal, with the added benefit of a reduction in disk space and memory usage. If memory usage is not a concern, then there may be a benefit to running multiple instances of NetRep, each using 1 core, instead of parallelising the procedure across multiple cores. In this case each instance will have its own copy of the test dataset in memory (*e.g.* each instance requires 8.13 GB of RAM). On some systems we tested, this provided a two-fold speed increase. Finally, some calculations may be faster when installing non-default BLAS and LAPACK libraries: however, this requires some experimentation on each machine, and reinstallation of R from source. All tests described above were performed with the default libraries installed with Ubuntu 12.0.5 LTS.

Software and hardware

Analyses were performed using the statistical computing software, R version 3.1.3 (http://www.r-project.org/). Network module preservation was assessed using the NetRep package version 0.23.0 (https://github.com/InouyeLab/NetRep/releases/tag/v0.23.0). Runtime analyses were performed using the latest version of NetRep: version 0.60.0. The latest stable version of the software can be found at http://github.com/InouyeLab/NetRep.

Analyses were performed on a Dell R910 with 40 cores (Intel Xeon 2.00 GHz), 512 Gb of RAM, 16x 1 Tb hard drives with 7200 RPM, running the 64 bit version of Ubuntu 12.0.5 LTS.

Supplemental Tables

		Permutation testing			
Evidence for preservation		Strong	Weak	None	Totals
Heuristic threshold on $Z_{summary}$	Strong	171	55	0	226
	Weak	17	167	4	188
	None	0	44	37	81
	Totals	188	266	41	495

Table S1, related to the experimental procedures and the main text: Comparison of module preservation evidence between *NetRep* and heuristic approach when testing the preservation of the 165 BxH mice modules in other tissues. For the permutation test, we defined strong evidence for a module's preservation in another tissue as all test statistics achieving $P < 0.0001$, weak evidence if one or more, but not all, test statistics were $P < 0.0001$, and no evidence if no test statistics are $P < 0.0001$. The significance threshold of 0.0001 was chosen to Bonferroni adjust for the 495 tests performed for each preservation statistic. The heuristic threshold is defined on $Z_{summary}$, a weighted combination of the Z-scores for all seven statistics (**Supplemental Experimental Procedures**). There is strong evidence for module preservation where $Z_{summary} > 10$, weak evidence if $2 < Z_{summary} < 10$, and no evidence of preservation if $Z_{summary} \leq 2$ (Langfelder et al., 2011).

Type	Term ID	Term	N	Total
KEGG	path:mmu03010	Ribosome	10	13
	path:mmu03013	RNA transport	4	5
	path:mmu03015	mRNA surveillance pathway	3	3
	path:mmu04120	Ubiquitin mediated proteolysis	3	4
	path:mmu04141	Protein processing in endoplasmic reticulum	3	4
GO:(BP)	GO:0006412	Translation	8	11
	GO:0006397	mRNA processing	6	7
	GO:0008380	RNA splicing	6	7
	GO:0015031	Protein transport	6	7
	GO:0006364	rRNA processing	5	5
	GO:0006974	Cellular response to DNA damage stimulus	4	9
	GO:0007049	Cell cycle	4	9
	GO:0051301	Cell division	4	10
	GO:0006281	DNA repair	3	7
	GO:0006355	Regulation of transcription, DNA-templated	3	5
	GO:0006511	Ubiquitin-dependent protein catabolic process	3	3
	GO:0006810	Transport	3	4
	GO:0006915	Apoptotic process	3	4
	GO:0007067	Mitosis	3	8
	GO:0016568	Chromatin modification	3	5
	GO:0051028	mRNA transport	3	3
GO:(MF)	GO:0044822	Poly(A) RNA binding	10	28
	GO:0003735	Structural constituent of ribosome	9	12
	GO:0003723	RNA binding	6	12
	GO:0000166	Nucleotide binding	4	9
	GO:0003676	Nucleic acid binding	4	5
	GO:0016874	Ligase activity	3	6
	GO:0046872	Metal ion binding	3	6
GO:(CC)	GO:0030529	Ribonucleoprotein complex	11	16
	GO:0005840	Ribosome	10	14
	GO:0022625	Cytosolic large ribosomal subunit	9	12
	GO:0005730	Nucleolus	5	9
	GO:0005622	Intracellular	4	6
	GO:0005634	Nucleus	4	8
	GO:0000775	Chromosome, centromeric region	3	6
	GO:0000785	Chromatin	3	5
	GO:0000808	Origin recognition complex	3	3
	GO:0005681	Spliceosomal complex	3	5
	GO:0005694	Chromosome	3	6
	GO:0005739	Mitochondrion	3	7
	GO:0016605	PML body	3	3
	GO:0022627	Cytosolic small ribosomal subunit	3	3
	GO:0071013	Catalytic step 2 spliceosome	3	5

Table S2, related to the experimental procedures and the main text: GO term and KEGG pathway enrichment of putative housekeeping BxH mice modules. 'Type': annotation type.

‘KEGG’: KEGG pathway, ‘GO:(BP)’: GO biological process, ‘GO:(MF)’: GO molecular function, ‘GO:(CC)’: GO cellular compartment. ‘Housekeeping’: the number of putative housekeeping modules (of a total of 41) significantly enriched for the associated term (**Supplemental Experimental Procedures**). ‘Total’: total number of modules (out of 165) significantly enriched for the associated term. Terms associated with 3 or more modules are listed in this table. 9 of the 41 putative housekeeping modules were not significantly enriched for any term.

Module	Var expl	Test tissue	Var expl	Effect size	95% confidence interval	*P*-value
Brain 2	51%	liver	13%	-0.26	-0.34– -0.18	1×10^{-9}
Brain 7	56%	liver	25%	0.20	0.11– 0.29	3×10^{-5}
Brain 16	59%	adipose	46%	0.25	0.16– 0.33	1×10^{-8}
Brain 22	57%	adipose	36%	0.18	0.093– 0.27	6×10^{-5}
Brain 24	72%	adipose	43%	-0.25	-0.33– -0.16	2×10^{-8}
Brain 31	63%	muscle	67%	-0.18	-0.26– -0.095	3 v 10^{-5}
Liver 2	35%	liver	35%	-0.34	-0.44– -0.23	2 v 10^{-9}
Liver 6	33%	liver	33%	0.22	0.13– 0.31	5×10^{-6}
Liver 7	33%	liver	33%	0.22	0.12– 0.31	1×10^{-5}
Liver 16	37%	adipose	29%	0.26	0.18– 0.34	2×10^{-9}
Liver 20	39%	liver	39%	-0.17	-0.26– -0.089	6×10^{-5}
Liver 28	48%	adipose	31%	0.18	0.097– 0.27	4×10^{-5}
Adipose 6	41%	adipose	41%	-0.20	-0.28– -0.11	6×10^{-6}
Adipose 7	45%	adipose	45%	-0.28	-0.37– -0.18	3×10^{-8}
Adipose 9	50%	adipose	50%	-0.23	-0.32– -0.15	1×10^{-7}
Adipose 10	55%	adipose	55%	0.26	0.18– 0.34	1×10^{-9}
Adipose 11	55%	liver	38%	-0.24	-0.32– -0.15	1×10^{-7}
Adipose 16	58%	liver	57%	0.19	0.10– 0.28	5×10^{-5}
Adipose 22	65%	adipose	65%	0.21	0.13– 0.30	1×10^{-6}
Adipose 23	45%	adipose	45%	0.20	0.11– 0.28	1×10^{-5}
Adipose 26	70%	adipose	70%	0.21	0.12– 0.29	4×10^{-6}
Muscle 2	57%	liver	12%	-0.27	-0.35– -0.19	7×10^{-10}
Muscle 3	57%	liver	13%	-0.25	-0.33– -0.17	5×10^{-9}
Muscle 5	49%	liver	12%	-0.32	-0.43– -0.21	9×10^{-9}
Muscle 7	51%	adipose	39%	-0.27	-0.35– -0.19	3×10^{-10}
Muscle 10	64%	liver	17%	-0.18	-0.27– -0.098	3×10^{-5}
Muscle 11	64%	adipose	40%	0.22	0.13– 0.30	6×10^{-7}
Muscle 14	73%	liver	21%	0.30	0.22– 0.39	4×10^{-11}
Muscle 24	76%	muscle	76%	-0.18	-0.27– -0.10	2×10^{-5}
Muscle 31	72%	adipose	61%	0.18	0.090– 0.27	9×10^{-5}
Brain 4	56%	liver	22%	-0.20	-0.29– -0.12	6×10^{-6}
		adipose	34%	-0.20	-0.28– -0.11	6×10^{-6}
Liver 5	35%	liver	35%	0.25	0.16– 0.35	2×10^{-7}
		adipose	32%	0.45	0.36– 0.53	2×10^{-21}
Liver 11	34%	liver	34%	-0.20	-0.29– -0.12	5×10^{-6}
		adipose	39%	-0.21	-0.29– -0.12	3×10^{-6}
Liver 18	37%	liver	37%	0.31	0.23– 0.39	1×10^{-13}
		adipose	34%	0.33	0.25– 0.41	7×10^{-14}
Adipose 3	44%	liver	15%	0.31	0.21– 0.40	4×10^{-10}
		adipose	44%	0.47	0.38– 0.56	2×10^{-21}
Adipose 19	58%	liver	28%	-0.20	-0.29– -0.12	4×10^{-6}
		adipose	58%	-0.24	-0.32– -0.15	7×10^{-8}
Adipose 20	61%	liver	38%	0.26	0.18– 0.34	1×10^{-9}
		adipose	61%	0.32	0.24– 0.40	2×10^{-13}
Muscle 6	66%	liver	32%	-0.20	-0.28– -0.11	9×10^{-6}
		adipose	47%	-0.20	-0.29– -0.12	3×10^{-6}
Brain 20	**58%**	**adipose**	**51%**	**0.30**	**0.22– 0.38**	$\mathbf{8 \times 10^{-12}}$
		muscle	**48%**	**-0.18**	**-0.27– -0.093**	$\mathbf{6 \times 10^{-5}}$

Liver 25	**43%**	**adipose**	**32%**	**-0.35**	**-0.45– -0.26**	**4×10^{-12}**
		liver	**43%**	**0.33**	**0.25– 0.40**	**5×10^{-15}**
Liver 35	**47%**	**adipose**	**58%**	**0.25**	**0.16– 0.33**	**3×10^{-8}**
		muscle	**53%**	**-0.21**	**-0.30– -0.13**	**3×10^{-6}**
Adipose 4	**45%**	**liver**	**11%**	**0.54**	**0.42– 0.66**	**1×10^{-17}**
		adipose	**45%**	**-0.44**	**-0.53– -0.36**	**4×10^{-21}**
Muscle 27	**79%**	**liver**	**56%**	**-0.18**	**-0.27– -0.093**	**9×10^{-5}**
		adipose	**71%**	**0.30**	**0.22– 0.38**	**3×10^{-12}**
		muscle	**79%**	**-0.25**	**-0.35– -0.15**	**1×10^{-6}**

Table S3, related to the experimental procedures and the main text: Preserved BxH mice modules associated with body weight. Linear regression of weight on module summary expression profiles which have strong evidence for preservation. Models were fit in each module's discovery tissue and in each tissue for which the module had strong evidence of preservation. Models were adjusted for sex. "Var expl" indicates the variance in module expression explained by the summary expression profile in the discovery tissue and test tissue respectively. Associations shown are significant at $P < 0.0001$; Bonferroni correcting for the total number of tests (273). Effect sizes denote the difference in SD-units of weight per SD increase of the module's summary expression profile in the test tissue. 95% CI: 95% confidence interval of the effect size. Emphasis indicates modules with a change in effect size sign across different test tissues (*i.e.* an association with differential module expression).

Type	Term ID	Term	#A	Module genes	*Q*-value
KEGG	path:mmu03010	Ribosome	145	*Rpl7a*, *Rpl18a*, *Rpl21*, *Rpl36*, *Rpl18*, *Rplp2*, *Rps24*, *Rpl8*	7×10^{-12}
GO:(BP)	GO:0006412	Translation	279	*Rpl18a*, *Rpl21*, *Rpl36*, *Rpl18*, *Rps24*, *Rpl8*	1×10^{-5}
	GO:0006414	Translational elongation	58	*Rplp2*, *Rps24*	0.01
	GO:2001141	Regulation of RNA biosynthetic process	82	*Naca*, *Eid1*	0.02
	GO:0006364	rRNA processing	108	*Rps24*, *Exosc7*	0.03
GO:(MF)	GO:0003735	Structural constituent of ribosome	128	*Rpl18a*, *Rpl21*, *Rpl36*, *Rpl18*, *Rplp2*, *Rps24*, *Rpl8*	1×10^{-9}
	GO:0044822	Poly(A) RNA binding	1,065	*Rpl7a*, *Rpl18a*, *Rpl21*, *Rps24*, *Rpl8*	0.04
GO:(CC)	GO:0022625	Cytosolic large ribosomal subunit	51	*Rpl7a*, *Rpl18a*, *Rpl21*, *Rpl18*, *Rplp2*, *Rpl8*	5×10^{-10}
	GO:0005840	Ribosome	154	*Rpl18a*, *Rpl21*, *Rpl36*, *Rpl18*, *Rplp2*, *Rps24*, *Rpl8*	5×10^{-9}
	GO:0030529	Ribonucleoprotein complex	308	*Snrpe, Rpl18a*, *Rpl21*, *Rpl36*, *Rpl18*, *Rplp2*, *Rps24*, *Rpl8*	1×10^{-8}
	GO:0016235	Aggresome	23	*Mvb12a, Eid1*	0.001
	GO:0005622	Intracellular	1,329	*Rpl21*, *Rpl36*, *Rpl18*, *Rplp2*, *Rps24*, *Rpl8*	0.02

Table S4, related to the experimental procedures and the main text: Significant enrichment of GO terms and KEGG pathways for BxH mice liver module 35 (LM35). GO terms and KEGG pathways significantly over-represented in the 24 annotated genes comprising LM35 (**Supplemental Experimental Procedures**). 'Type': annotation type. 'KEGG': KEGG pathway, 'GO:(BP)': GO biological process, 'GO:(MF)': GO molecular function, 'GO:(CC)': GO cellular compartment. '#A': total number of genes annotated for the term. In total there were 18,737 annotated genes on the microarray. *Q*-value: false discovery rate corrected hypergeometric test *p*-value.

Table S5, related to the experimental procedures and the main text: Probe annotations for LM35. Probes are listed in descending order by weighted degree in the liver. The relative rank of the weighted degree indicates importance to the module within each tissue. Note that the values cannot be directly compared across tissues. Blank spaces in the adipose column indicate probes that did not pass quality control in the adipose tissue. Gene symbols with a preceding * indicate entry has been withdrawn by Entrez and is not in the current annotation release. Gene symbols starting with MMT indicate custom probes designed from mouse Unigene clusters lacking annotation (**Supplemental Experimental Procedures**). Start and end indicate the start and end base position of the gene on the corresponding chromosome.

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

Table S6, related to the experimental procedures and the main text: Associations between cardiometabolic traits and LM35 summary expression in the adipose and muscle tissues. Associations were calculated via linear regression of each trait on LM35 summary expression. In total, associations for 20 cardiovascular risk traits, not including weight, were assessed in both the adipose and muscle tissue (**Supplemental Experimental Procedures**). The summary expression profiles were calculated in the respective test tissue for probes comprising LM35. The summary expression profile explained 58% and 53% of the LM35's variance in the adipose and muscle tissues respectively. Models were adjusted for sex, and *p*-values were false-discovery rate corrected within each test tissue. Associations were considered significant at FDR < 0.025. A log transform was applied to insulin, monocyte chemotactic protein 1 (MCP-1/CCL2), and the ratio of glucose/insulin. Effect sizes denote the difference in SD-units of each trait per SD increase of the summary expression profile in the respective tissue. *Q*-value: false discovery rate adjusted *p*-value. Traits significantly associated with LM35 in both tissues in bold.

Module	OTU	Order	Family	Genus	Species
Female 1	565357	Clostridiales			
	362344	Clostridiales	Lachnospiraceae		
	369768	Clostridiales	Lachnospiraceae		
	335827	Clostridiales	Lachnospiraceae	*Butyrivibrio*	
	338987	Clostridiales	Lachnospiraceae	*Coprococcus*	
	181155	Clostridiales	Ruminococcaceae		
Female 6	205904	Bacteroidales	Bacteroidaceae	*Bacteroides*	
	213566	Bacteroidales	Bacteroidaceae	*Bacteroides*	
	213813	Bacteroidales	Bacteroidaceae	*Bacteroides*	
Male 2	196632	Clostridiales	Clostridiaceae	*Clostridium*	
	367150	Clostridiales	Lachnospiraceae		
	362344	Clostridiales	Lachnospiraceae		
	369768	Clostridiales	Lachnospiraceae		
	335827	Clostridiales	Lachnospiraceae	*Butyrivibrio*	
	338987	Clostridiales	Lachnospiraceae	*Coprococcus*	
	528715	Clostridiales	Ruminococcaceae	*Faecalibacterium*	*prausnitzii*
Male 4	180927	Clostridiales	Veillonellaceae	*Dialister*	
	201364	Clostridiales	Veillonellaceae	*Dialister*	
	3086353	Clostridiales	Veillonellaceae	*Dialister*	
	403701	Clostridiales	Veillonellaceae	*Dialister*	
Male 5	205904	Bacteroidales	Bacteroidaceae	*Bacteroides*	
	213566	Bacteroidales	Bacteroidaceae	*Bacteroides*	
	213813	Bacteroidales	Bacteroidaceae	*Bacteroides*	
Male 8	194868	Clostridiales	Ruminococcaceae	*Ruminococcus*	
	198980	Clostridiales	Ruminococcaceae	*Ruminococcus*	
	210647	Clostridiales	Ruminococcaceae	*Ruminococcus*	

Table S7, related to Figure 6: Taxonomical assignments for preserved HMP gut microbial communities. Blank cells indicate taxonomical assignment at the respective level could not be assigned. "Module": the preserved HMP gut microbial community each OTU participated in. Only modules with strong evidence for preservation in the other sex's gut microbial network are shown.

Supplemental figures

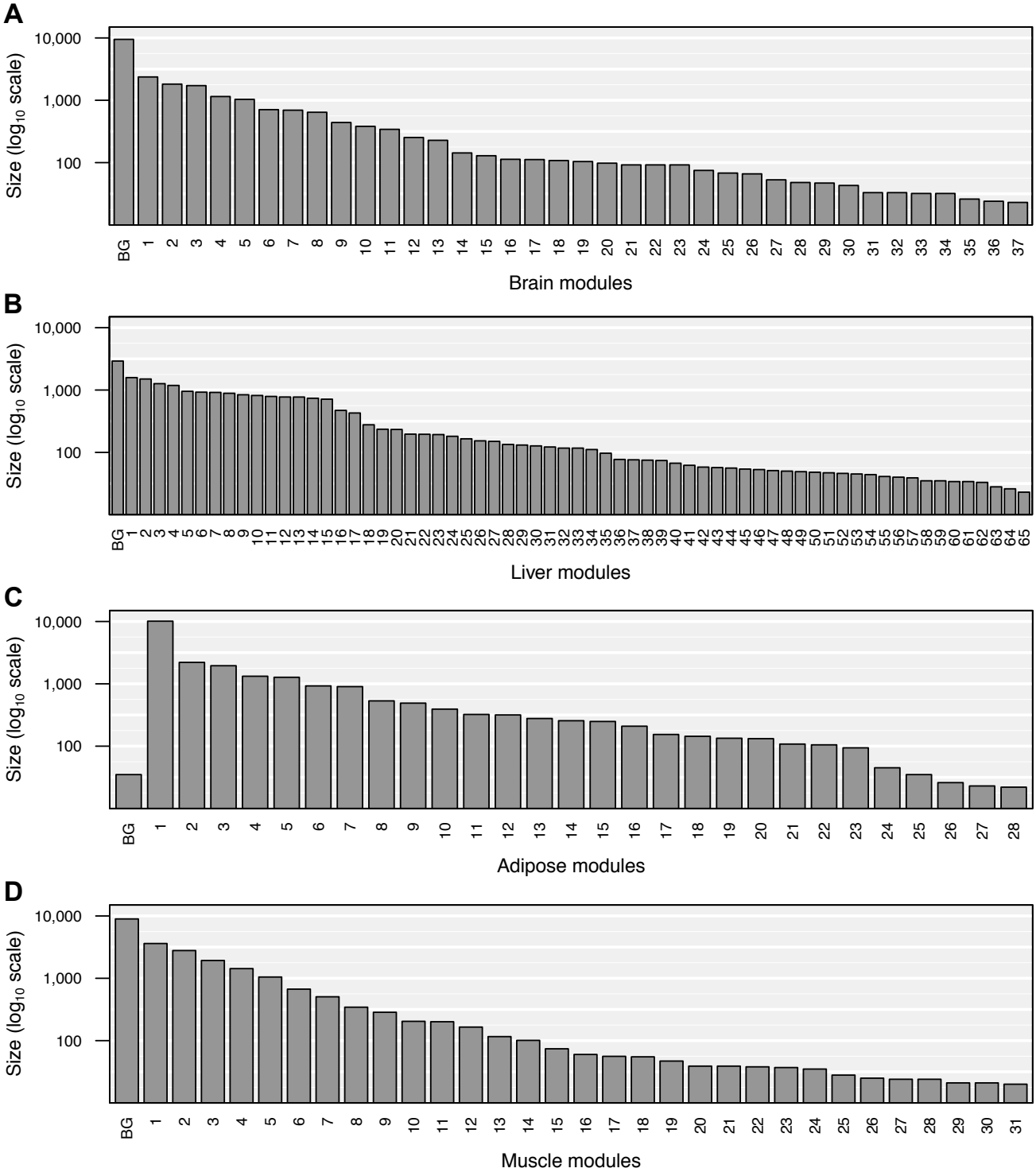


Figure S1, related to the experimental procedures and the main text: Network modules identified in the four BxH mice tissues. Number of probes clustering into each inferred network module (**Experimental Procedures**) in the brain (A), liver (B), adipose (C), and muscle (D) tissues of the BxH mouse cross. 'BG' denotes the background module, which contains probes otherwise not assigned to any module.

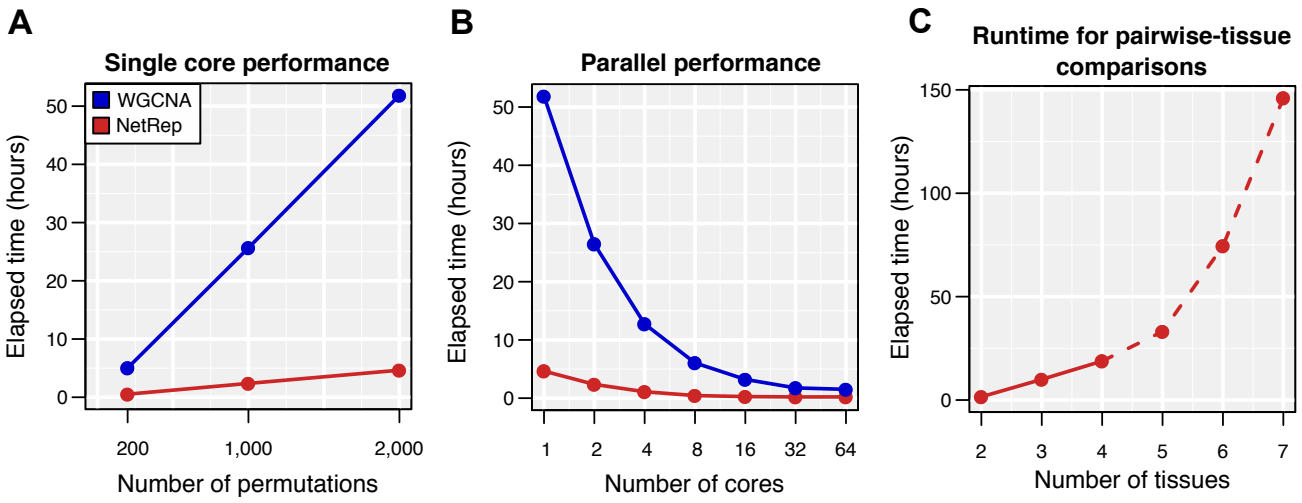


Figure S2, related to the experimental procedures and the main text: Computational performance of *NetRep*. (A) Total elapsed runtime when calculating module preservation statistics for the BxH mice brain modules on 200, 1,000, and 2,000 permutations of the liver tissue using *NetRep* (red) and when calculating the same number of permutations with the *WGCNA* R package (blue). (B) Total elapsed runtime to calculate 2,000 permutations when parallelised across multiple cores. (C) Total elapsed runtime when parallelised over 40 cores to perform a pairwise comparison of two tissues (brain and liver), three tissues (brain, liver, and adipose), and all four tissues with 10,000 permutations per tissue comparison. Runtime for pairwise comparison of 5–7 tissues was extrapolated from the median time taken per permutation in the pairwise analysis of the four tissues (15 seconds) with Bonferroni corrected significance thresholds assuming an average of 40 modules per tissue ($P < 6 \times 10^{-5}$, $P < 4 \times 10^{-5}$, and $P < 3 \times 10^{-5}$, requiring 16,000, 24,000, and 33,600 permutations *per tissue*, respectively).

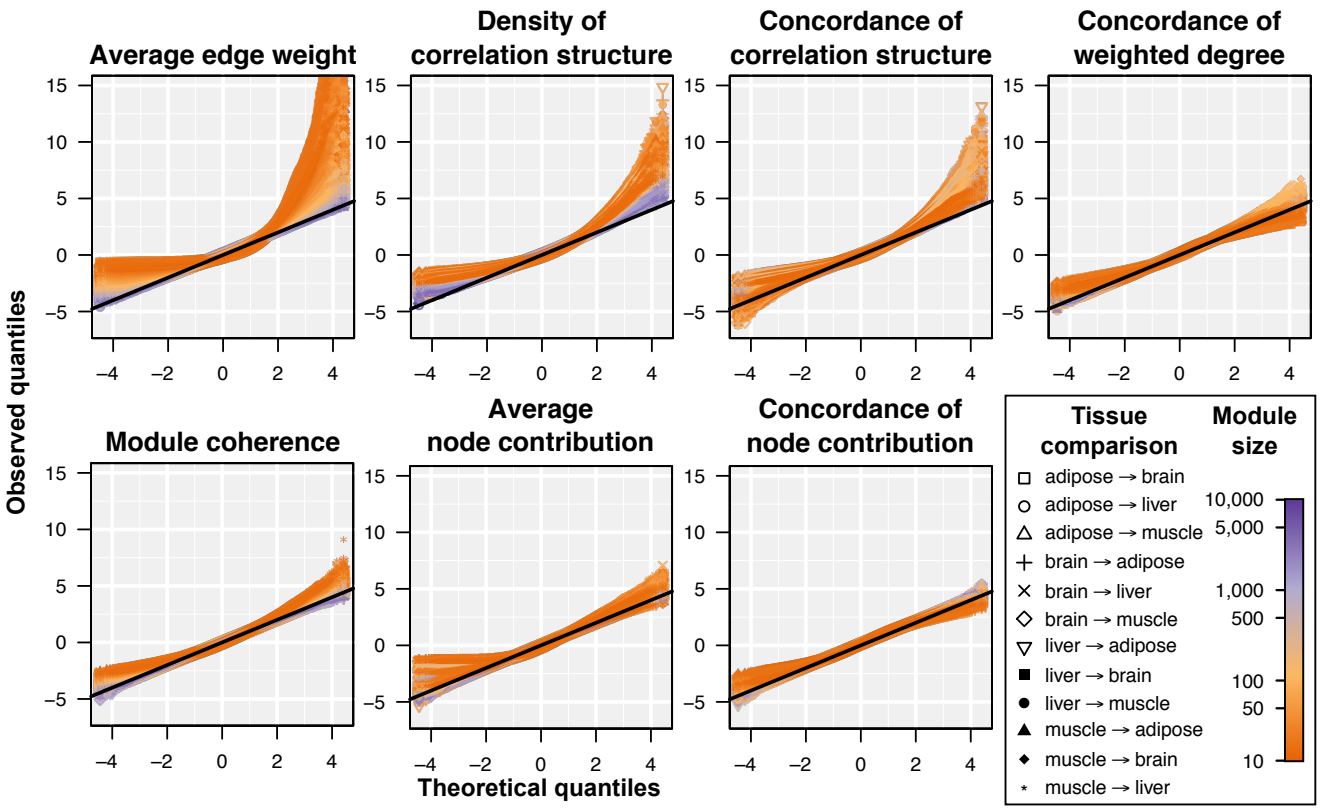


Figure S3, related to the experimental procedures and the main text: Non-normality of module preservation statistics when drawing null distributions for 165 BxH mice modules in their three non-discovery tissues. Quantile-Quantile plots (QQ-plots) show the theoretical quantiles of a normal distribution in comparison to the observed quantiles from each of the 495 empirical null distributions for each module preservation statistic. Observed quantiles were generated from 100,000 permutations (**Experimental Procedures**). Points in each panel were given different symbols to denote each of the twelve tissue comparisons, and coloured to denote module size (on a $\log_{10}$ scale). Points were overlaid in descending order of tissue comparison and module size shown in the legend. The QQ-plot for the *average edge weight* is truncated: the maximum observed quantile is 38.

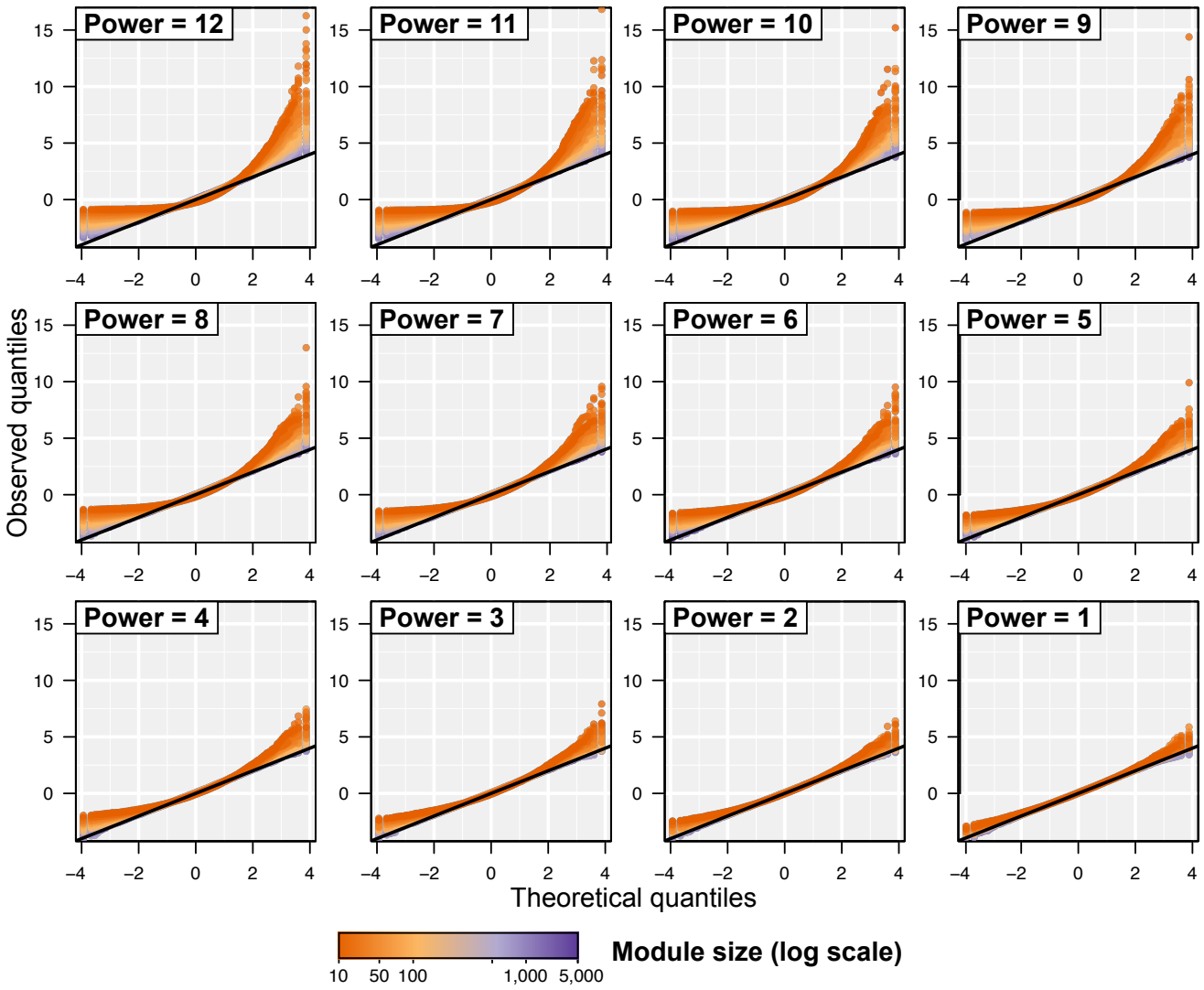


Figure S4, related to the experimental procedures and the main text: The scale-free assumption affects non-normality of the *average edge weight* statistic. Quantile-Quantile plots comparing theoretical quantiles of a normal distribution to observed quantiles for the null distributions of the *average edge weight* statistic, when varying the soft-threshold exponent used to define the interaction network. Null distributions were generated for the 66 BxH mice liver modules from 10,000 permutations in the muscle tissue. The exponent was varied from 12, the power used to define the interaction network in the muscle tissue (**Experimental Procedures**, **Figure S7**), to 1. Quantiles were coloured to denote module size (on a $\log_{10}$ scale). Points were overlaid in descending order of module size.

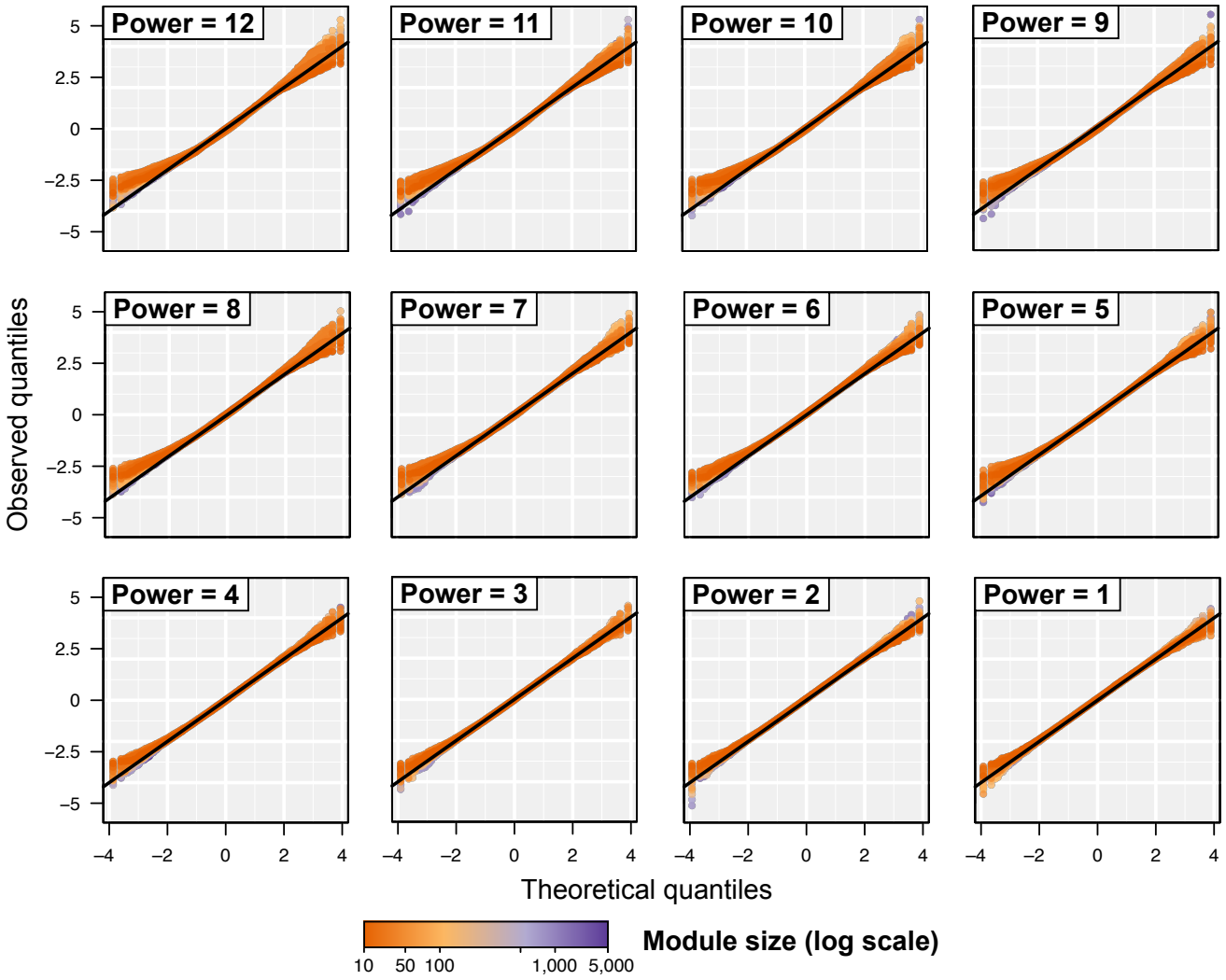


Figure S5, related to the experimental procedures and the main text: The scale-free assumption has no effect on the *concordance of weighted degree* statistic. Quantile-Quantile plots comparing theoretical quantiles of a normal distribution to observed quantiles for the null distributions of the *concordance of weighted degree* statistic, when varying the soft-threshold exponent used to define the interaction network. Null distributions were generated for the 66 BxH mice liver modules from 10,000 permutations in the muscle tissue. The exponent was varied from 12, the power used to define the interaction network in the muscle tissue (**Experimental Procedures**, **Figure S7**), to 1. Quantiles were coloured to denote module size (on a $\log_{10}$ scale). Points were overlaid in descending order of module size.

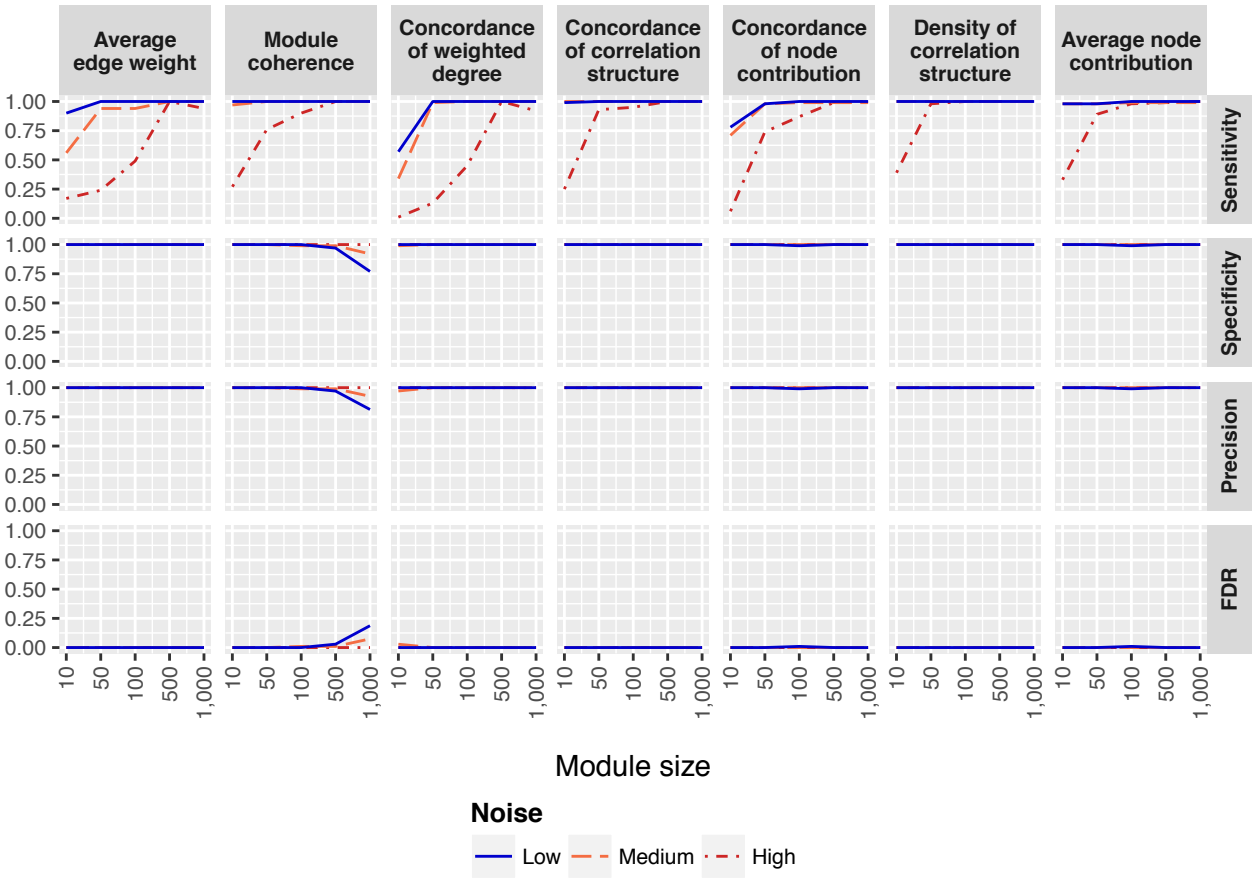


Figure S6, related to the experimental procedures and the main text: Performance of *NetRep* in a simulation study. Sensitivity (true positive rate), specificity (true negative rate / 1 – false positive rate), precision, and false discovery rate (FDR) of the seven module preservation statistics when assessing preservation of positive control and negative control modules simulated with varying degrees of noise (**Supplemental Experimental Procedures**). Points within each line (one point per module size) were calculated from 100 simulated discovery and test datasets. Module preservation statistics were considered significant where $P < 1 \times 10^{-4}$. *p*-values were estimated from null distributions generated from 10,000 permutations. "Low", "Medium" and "High" noise test datasets were simulated using error terms drawn from a normal distribution with mean of 0 and standard deviations of 1, 2, and 5 respectively.

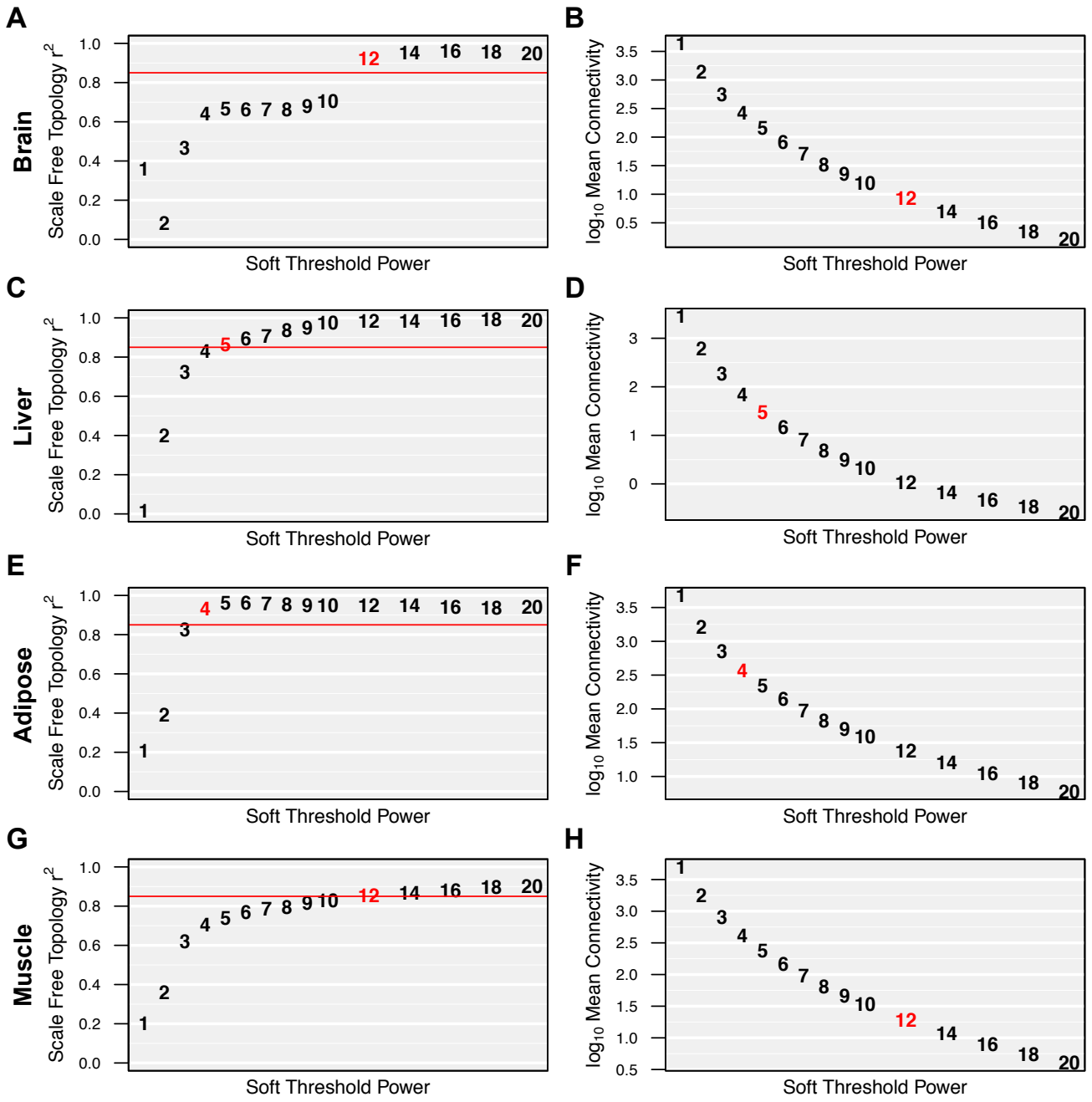


Figure S7, related to the experimental procedures and the main text: Soft-threshold powers used to define the BxH mice interaction networks. Scale free topology criterion R^2 for the brain (A), liver (C), adipose (E), and muscle (G) interaction networks when defined using various soft threshold powers. The horizontal red line (A, C, E, G) shows the R^2 cut-off of 0.85 recommended by the scale-free topology criterion (Zhang and Horvath, 2005). Average weighted degree (on a $\log_{10}$ scale) for the whole interaction networks inferred for the brain (B), liver (D), adipose (F), and muscle (H) tissues when defined using various soft threshold powers. The power labelled in red shows the power selected for the power transform function (**Experimental Procedures**).

Supplemental References

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. *25*, 25–29.

Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., Fierer, N., Peña, A.G., Goodrich, J.K., Gordon, J.I., et al. (2010a). QIIME allows analysis of high-throughput community sequencing data. Nat. Methods *7*, 335–336.

Caporaso, J.G., Bittinger, K., Bushman, F.D., DeSantis, T.Z., Andersen, G.L., and Knight, R. (2010b). PyNAST: a flexible tool for aligning sequences to a template alignment. Bioinformatics *26*, 266–267.

Eddelbuettel, D., and François, R. (2011). Rcpp : Seamless R and C++ integration. J. Stat. Softw. *40*.

Eddelbuettel, D., and Sanderson, C. (2014). RcppArmadillo: Accelerating R with high-performance C++ linear algebra. Comput. Stat. Data Anal. *71*, 1054–1063.

Edgar, R.C. (2010). Search and clustering orders of magnitude faster than BLAST. Bioinformatics *26*, 2460–2461.

Estrada-Smith, D., Castellani, L.W., Wong, H., Wen, P.Z., Chui, A., Lusis, A.J., and Davis, R.C. (2004). Dissection of multigenic obesity traits in congenic mouse strains. Mamm. Genome *15*, 14–22.

Friedman, J., and Alm, E.J. (2012). Inferring correlation networks from genomic survey data. PLoS Comput. Biol. *8*, e1002687.

Haas, B.J., Gevers, D., Earl, A.M., Feldgarden, M., Ward, D. V, Giannoukos, G., Ciulla, D., Tabbaa, D., Highlander, S.K., Sodergren, E., et al. (2011). Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. Genome Res. *21*, 494–504.

He, Y.D., Dai, H., Schadt, E.E., Cavet, G., Edwards, S.W., Stepaniants, S.B., Duenwald, S., Kleinhanz, R., Jones, A.R., Shoemaker, D.D., et al. (2003). Microarray standard data set and figures of merit for comparing data processing methods and experiment designs. Bioinformatics *19*, 956–965.

Human Microbiome Project Consortium (2012). A framework for human microbiome research. Nature *486*, 215–221.

Kanehisa, M., and Goto, S. (2000). Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. *28*, 27–30.

McDonald, D., Price, M.N., Goodrich, J., Nawrocki, E.P., DeSantis, T.Z., Probst, A., Andersen, G.L., Knight, R., and Hugenholtz, P. (2012). An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. ISME J. *6*, 610–618.

Meng, H., Vera, I., Che, N., Wang, X., Wang, S.S., Ingram-Drake, L., Schadt, E.E., Drake, T.A., and Lusis, A.J. (2007). Identification of Abcc6 as the major causal gene for dystrophic cardiac calcification in mice through integrative genomics. Proc. Natl. Acad. Sci. *104*, 4530–4535.

Newman, M. (2010). Networks: an introduction (Oxford University Press).

Sanderson, C. (2010). Armadillo: an open source C++ linear algebra library for fast prototyping and computationally intensive experiments.

Schadt, E.E., Monks, S.A., Drake, T.A., Lusis, A.J., Che, N., Colinayo, V., Ruff, T.G., Milligan, S.B., Lamb, J.R., Cavet, G., et al. (2003). Genetics of gene expression surveyed in maize, mouse and man.

Nature *422*, 297–302.

Schadt, E.E., Molony, C., Chudin, E., Hao, K., Yang, X., Lum, P.Y., Kasarskis, A., Zhang, B., Wang, S., Suver, C., et al. (2008). Mapping the genetic architecture of gene expression in human liver. PLoS Biol. *6*, 1020–1032.

Werner, J.J., Koren, O., Hugenholtz, P., DeSantis, T.Z., Walters, W.A., Caporaso, J.G., Angenent, L.T., Knight, R., and Ley, R.E. (2012). Impact of training sets on classification of high-throughput bacterial 16s rRNA gene surveys. ISME J. *6*, 94–103.

Yang, X., Schadt, E.E., Wang, S., Wang, H., Arnold, A.P., Ingram-Drake, L., Drake, T.A., and Lusis, A.J. (2006). Tissue-specific expression and regulation of sexually dimorphic genes in mice. Genome Res. *16*, 995–1004.