Identifying gene function and module connections by the integration of multispecies expression compendia

Hao Li, Daria Rukina, Fabrice P.A. David, Tertyty Yang Li, Chang-Myung Oh, Arwen W. Gao, Elena Katsyuba, Maroun Bou Sleiman, Andrea Komljenovic, Qingyao Huang, Robert W. Williams, Marc Robinson-Rechavi, Kristina Schoonjans, Stephan Morgenthaler, and Johan Auwerx

1Laboratory of Integrative Systems Physiology, Institute of Bioengineering, École Polytechnique Fédérale de Lausanne, Lausanne 1015, Switzerland; 2Institute of Mathematics, École Polytechnique Fédérale de Lausanne, Lausanne 1015, Switzerland; 3Gene Expression Core Facility, École Polytechnique Fédérale de Lausanne, Lausanne 1015, Switzerland; 4SV-IT, École Polytechnique Fédérale de Lausanne, Lausanne 1015, Switzerland; 5Swiss Institute of Bioinformatics, Lausanne 1015, Switzerland; 6Department of Ecology and Evolution, University of Lausanne, Lausanne 1015, Switzerland; 7Laboratory of Metabolic Signaling, Institute of Bioengineering, École Polytechnique Fédérale de Lausanne, Lausanne 1015, Switzerland; 8Department of Genetics, Genomics and Informatics, University of Tennessee, Memphis, Tennessee 38163, USA

The functions of many eukaryotic genes are still poorly understood. Here, we developed and validated a new method, termed GeneBridge, which is based on two linked approaches to impute gene function and bridge genes with biological processes. First, Gene-Module Association Determination (G-MAD) allows the annotation of gene function. Second, Module-Module Association Determination (M-MAD) allows predicting connectivity among modules. We applied the GeneBridge tools to large-scale multispecies expression compendia—700 data sets with over 300,000 samples from human, mouse, rat, fly, worm, and yeast—collected in this study. G-MAD identifies novel functions of genes—for example, DDT in mitochondrial respiration and WDFY4 in T cell activation—and also suggests novel components for modules, such as for cholesterol biosynthesis. By applying G-MAD on data sets from respective tissues, tissue-specific functions of genes were identified—for instance, the roles of EHHADH in liver and kidney, as well as SLC6A1 in brain and liver. Using M-MAD, we identified a list of module-module associations, such as those between mitochondria and proteasome, mitochondria and histone demethylation, as well as ribosomes and lipid biosynthesis. The GeneBridge tools together with the expression compendia are available as an open resource, which will facilitate the identification of connections linking genes, modules, phenotypes, and diseases.

[Supplemental material is available for this article.]
et al. 1998). By using overrepresentation analyses on subnetworks or modules, one can then deduce aspects of gene functions.

However, these approaches generally depend on discrete subsets of genes whose expression correlations exceed either a hard or soft threshold, which would strongly influence the final results. In addition, such analyses typically focus on positive or absolute values of correlations among data sets. The key polarity of interactions is often lost among gene products and linked modules (Warde-Farley et al. 2010; Greene et al. 2015; van Dam et al. 2015; Zhu et al. 2015; Li et al. 2017). Gene set analyses, such as gene set enrichment analysis (GSEA) (Subramanian et al. 2005), have been developed to identify processes or modules that are affected by certain genetic or environmental perturbations (Khatri et al. 2012). While GSEA uses all measured genes in the analysis, its application has mainly been limited to studying G/LOF models or environmental perturbations, where comparisons are inherently among discrete categories. This limits its applicability in most populations, in which variations among individuals are often subtle and continuous (Williams and Auwerx 2015).

Here, we developed the GeneBridge toolkit that uses two interconnected approaches to improve upon the identification of gene function and to bridge genes to phenotypes using large-scale cross-species transcriptome compendia collected for this study. First, we describe a computational approach, named Gene-Module Association Determination (G-MAD), to impute gene function. G-MAD considers expression as a continuous variable and identifies the associations between genes and modules. Second, we developed the Module-Module Association Determination (M-MAD) method to identify connections between modules based on the transcriptome compendia. The data and GeneBridge tools described here are available at systems-genetics.org, an open resource, which will facilitate the identification of novel connections between genes, modules, phenotypes, and diseases.

**Results**

**Gene-Module Association Determination (G-MAD)**

Owing to the fact that a large number of genes are still not well annotated or even uncharacterized (Supplemental Fig. S1), we propose here a new computational strategy, Gene-Module Association Determination (G-MAD), which uses expression data from large-scale cohorts to propose potential functions of genes. We use the term “modules” to refer to the knowledge-based gene sets, ontology terms, and biological pathways from different resources for simplicity in the rest of the paper. The differences between gene sets or directed or undirected pathways are important in many contexts, but for our purpose, they can be treated in the same manner as modules and will not be distinguished. The basic concept is similar to classic pathway/gene set analysis, i.e., genes that possess similar functions tend to have similar expression patterns (Subramanian et al. 2005). However, instead of using binary group settings (e.g., control vs. treatment, or wild type vs. knockout) as commonly used in gene set analysis, we consider the continuous expression levels of the gene-of-interest across a population and determine its possible functions based on its coexpression patterns against all genes.

In this study, we collected transcriptome data sets with over 80 samples obtained from six species (human, mouse, rat, fly, worm, and yeast) from GEO, ArrayExpress, dbGaP, GeneNetwork, and other data repository sources (Supplemental Table S1). For example, 1337 data sets containing over 265,000 human samples with whole-genome transcript levels were analyzed in this study (Supplemental Table S1). The expression data sets were preprocessed using PEER (Stegle et al. 2012) to remove the known and hidden covariates that would influence the analysis (Supplemental Fig. S2). We applied a competitive gene set testing method—Correlation Adjusted Mean RAnk gene set test (CAMERA), which adjusts for inter-gene correlations (Wu and Smyth 2012)—to compute the enrichment between gene-of-interest and biological modules. Gene-module connections with enrichment P-values that survived multiple testing corrections of the gene or module numbers were allocated connection scores of 1 or −1, based on the enrichment direction, and 0 otherwise. The results were then meta-analyzed across data sets, and gene-module association scores (GMASs) were computed as the averages of the connection scores weighted by the sample sizes and inter-gene correlation coefficients within modules (Fig. 1A).

One should be aware of the fact that modules can overlap partially or completely. For example, Gene Ontology (GO) categories have a hierarchical structure (Ashburner et al. 2000), and modules from different sources can be very similar in composition. Therefore, we computed the similarities across all modules and generated a global module similarity network. As expected, redundant modules formed clusters in the network, and we were able to extract 62 distinct module clusters in the human module similarity network (Fig. 1B; Supplemental Table S2). This network can be used as a way to visualize the results of gene-module associations.

We assessed the performance of G-MAD in prioritizing known genes for modules through cross-validation. We then compared the area under the receiver operating characteristic (ROC) curve (AUC) with the ones obtained from WeGET (Szklarczyk et al. 2016) and COXPRESdb (Oabyrin et al. 2019). G-MAD exhibits better predictive performance than WeGET and COXPRESdb, especially for larger modules (e.g., those with more than 50 genes), as well as a much simpler method based on the average of correlation coefficient between gene pairs using the same expression compendium of our method (average r) (Supplemental Fig. S3A, B). To estimate if the performance gained from larger data set numbers in our study, we repeated G-MAD using a subset (800) of the data sets (G-MADsub). We observed that G-MADsub had similar performance as G-MAD and better than COXPRESdb and WeGET, where around 1000 data sets were used. We investigated the influence of the inter-gene correlations within modules (ρ) and module size on the predictive performance of G-MAD and noticed that modules with higher inter-gene correlations and smaller modules tend to have better performances (Supplemental Fig. S3C,D).

Furthermore, in order to determine the threshold of significance of gene-module associations, we computed the GMAS of all the gene-module pairs, including both known and unknown pairs. We then created the ROC curve by varying the threshold of significance and calculating the true positive rate (percentage of known genes above the threshold against all known genes) and false positive rate (percentage of unknown genes above the threshold against all unknown genes) (Fig. 1C). Detecting more true positives by lowering the threshold comes with a cost of a higher false positive rate. Therefore, to be stringent in proposing novel gene-module associations and restrain the likelihood of raising false positives, we considered a true positive rate of 0.1 (only 10% of all the known gene-module pairs as significant), and used a GMAS threshold of 0.268 (Fig. 1C). With this threshold, we saw that only 0.24% of unknown gene-module pairs are significant, which is 40 times less than the known pairs.
The gene-module connections predicted by G-MAD provide a resource that researchers can use as a reference when annotating gene functions. We describe below some examples on how the G-MAD results can be used to facilitate the discovery of novel gene functions or the identification of new members of modules.

WDFY4 was recently annotated as a crucial gene in activating immune T cells in antiviral and antitumor immunity through a functional CRISPR screen (Theisen et al. 2018). Through G-MAD, we found that WDFY4 indeed associated with antigen processing, T cell activation, and immune response in human, mouse, and rat (Fig. 1D; Supplemental Fig. S4A,B), verifying that its function is conserved across species. Cholesterol is critical in cell differentiation and growth. We identified 20 genes (AACS, ACLY, ACSL3, ACSS2, CYB5B, DBI, ELOVL6, ERG28, FADS1, FASN, INSIG1, PANK3, PCSK9, PCYT2, PNPLA3, RHDI1, SLC25A1, STAR4, TMEM41B, TMEM97) associated with cholesterol biosynthesis conserved in human, mouse, and rat (Fig. 1E; Supplemental Fig. S4C–E). Several of these genes, including FASN (Carroll et al. 2018) and TMEM97 (Bartz et al. 2009), have already been described to have relevant functions in cholesterol metabolism.

**Figure 1.** Gene-Module Association Determination (G-MAD). (A) G-MAD methodology. See text and Methods for detailed description. (B) Module similarity network showing the composition similarities across all module pairs. Modules were detected using a community detection algorithm embedded in Gephi and indicated in different colors. The 10 most frequent words of the module terms in each module were used to represent the module and can be found in Supplemental Table S2. (C) Influence of the GMAS threshold (t) on the true positive rate (TPR) and false positive rate (FPR) of G-MAD. Using a threshold of 0.268, G-MAD identified 10% of true positives and 0.24% of false positives (reflected by the red lines intersecting the x- and y-axes). (D) G-MAD revealed the potential role of WDFY4 in T cell activation and immune response. The threshold of significant gene-module association is indicated by the red dashed line. Modules are organized by the module similarities. Known modules connected to WDFY4 from annotations are shown in red dots (there is no known connected module for WDFY4) and other modules with GMAS over the threshold are shown in black dots. Dot sizes reflect the GMAS of WDFY4 against the respective modules. Detailed information of all the modules is available at www.systems-genetics.org/modules_by_gene/WDFY4?organism=human. (E) G-MAD identified the involvement of known as well as 20 novel genes in cholesterol biosynthesis. The threshold of significant gene-module association is indicated by the red dashed line. Genes are organized by the genetic positions across chromosomes. Genes annotated to be involved in cholesterol biosynthesis are shown in red dots and novel genes with GMAS over the threshold are shown in black dots. Novel genes conserved in human, mouse, and rat are highlighted in red bold text.
G-MAD identifies tissue-specific associations

Using the expression compendia, we noticed that genes annotated to some modules have higher coexpression in data sets from certain tissues than others (Fig. 2A), suggesting the tissue-specific activation of these modules. For instance, genes involved in “pancreatic secretion” have much higher coexpressions in data sets obtained from pancreas (Fig. 2B). To predict the tissue specificity of modules, we compared the inter-gene correlations within each module (ρ) in every tissue against those in the other tissues using the nonparametric Kolmogorov–Smirnov (K–S) test. The resulting P-values are used as a measure to indicate tissue specificity of the modules (Fig. 2C). As an example, the “pancreatic secretion” module has higher specificity in the pancreas than in other tissues, such as the blood (Fig. 2D). Similarly, genes belonging to the “collecting duct acid secretion” module are highly coexpressed in kidney (Supplemental Fig. S5A–E), while genes in the “lamellar body” module are highly coexpressed in lung (Supplemental Fig. S5F–J). Therefore, G-MAD can also highlight tissue-specific gene-module associations using data sets from specific tissues. EHHADH is a peroxisomal protein highly expressed in liver and kidney (Fig. 3A; Uhlen et al. 2015). Although best known for its key role in the peroxisomal oxidation pathway, a recent report demonstrated that EHHADH mutations cause renal Fanconi’s syndrome (Klootwijk et al. 2014). G-MAD of EHHADH in liver and kidney identifies its conserved role in peroxisome and fatty acid oxidation and also recovers its specific functions in liver (e.g., bile acid biosynthesis) and kidney (e.g., brush border membrane) (Fig. 3B–E; Supplemental Table S3). SLC6A1 is one of the major gamma-aminobutyric acid (GABA) transporters in the neurotransmitter release cycle in brain (Carvill et al. 2015). However, SLC6A1 is also highly expressed in the liver (Supplemental Fig. S6A), and its function in liver remains poorly understood. G-MAD of SLC6A1 was performed in all data sets and only data sets from brain confirm its function as a neurotransmitter transporter in the GABA release cycle (Supplemental Fig. S6B,C), while G-MAD using data sets from liver identifies its possible role in carboxylic acid transport and metabolism (Supplemental Fig. S6D,E; Supplemental Table S4).

G-MAD determines novel genes linked to mitochondria

Mitochondria are the main powerhouses of cells and harvest energy in the form of ATP through mitochondrial respiration. There are around 1100 genes known to encode mitochondrial-localized proteins (mito-proteins), depending on the source used (e.g., 1158 mito-proteins in MitoCarta [Calvo et al. 2016], 1074 in Human Protein Atlas [Uhlen et al. 2015]); however, many of these genes remain uncharacterized, and the list of mito-proteins is still incomplete (Williams et al. 2018).

By using the genes annotated to be involved in the respiratory electron transport chain (ETC, Reactome: R-HSA-611105), we searched for genes potentially related to respiratory electron transport, by applying G-MAD to expression data sets in human,

![Figure 2](image-url)
mouse, and rat. As expected, genes annotated in the ETC module are strongly enriched; moreover, other known ETC genes that were not included in the module were also positively enriched, providing proof that G-MAD can recover known gene functions (Fig. 4A; Supplemental Fig. S7A,B). Based on G-MAD results from human, mouse, and rat, there were 707 genes showing conserved associations with the ETC (Fig. 4B). Many of these genes—for example, DMAC1 (previously known as C9orf123) (Arroyo et al. 2016; Stroud et al. 2016; Horlbeck et al. 2018), NDUFAF8 (previously known as C17orf89) (Floyd et al. 2016), and FMC1 (previously known as C7orf55) (Lefebvre-Legendre et al. 2001; Li et al. 2017)—were not included in the respiratory electron transport module but have been recently validated to be involved in mitochondrial localization of DDT through immunofluorescence staining (Fig. 4E). The function of DDT was confirmed through RNAi-mediated knockdown in HEK293 cells, which led to reduced transcript levels of genes encoding for the ETC subunits (Supplemental Fig. S7H; Supplemental Table S6) and decreased oxygen consumption rate (OCR) (Fig. 4F; Supplemental Fig. 7I), verifying that DDT impacts mitochondrial respiration. Similarly, we also validated the involvement of BOLA3 in the ETC using G-MAD and further experimental validations (Supplemental Fig. S8; Cameron et al. 2011).

Contrary to most of the existing sources that predict only positive gene-module associations, G-MAD is also able to exploit negative associations. For example, ARID1A exhibits significant negative associations with the respiratory electron transport in human and mouse (Fig. 4A,G,H; Supplemental Fig. S9). ARID1A is a known member of the SWI/SNF family, and the inactivating mutations of SWI/SNF complex genes (mainly SMARCA4 and ARID1A) have recently been linked to increased expression of ETC genes and mitochondrial respiration (Lissanu Deribe et al. 2018). To further validate its regulatory role, we checked an extant public data set from mice with uterus-specific Arid1a knockout (Kim et al. 2015) and confirmed that dysfunction of Arid1a led to the increased expression of mitochondrial genes (Fig. 4I), especially those involved in respiratory electron transport (Fig. 4J).

Module-Module Association Determination (M-MAD)

Biological processes and modules such as metabolism, cellular signaling, biogenesis, and degradation are interconnected and coordinated (Barabási et al. 2011). However, there are few reports exploring the connections between modules in a systematic fashion (Li et al. 2008). Here, we extend G-MAD to develop Module-Module Association Determination (M-MAD) to investigate the connections between modules based on the expression compendia. Results for individual modules against all genes, obtained

---

**Figure 3.** G-MAD identifies tissue-specific associated modules for EHHADH by using data sets from different tissues. (A) Expression patterns of EHHADH across tissues. The figure was adapted from the Human Protein Atlas (www.proteinatlas.org). (B–D) G-MAD of EHHADH in human using data sets from all tissues (B), from liver (C), or from kidney (D). The threshold of significant gene-module association is indicated by the red dashed line. Modules are organized by their similarities. Known modules connected to EHHADH from gene annotations are shown in red dots and other modules with GMAS over the threshold are shown by black dots. (E) Comparison of G-MAD results of EHHADH in liver and kidney. Known modules connected to EHHADH are shown in red dots. The threshold of significant gene-module association is indicated by the red dashed line. Modules significantly associated with EHHADH only in one specific tissue are highlighted. The comparison of the association results of EHHADH in liver and kidney can be found in Supplemental Table S3.
Figure 4. G-MAD predicts novel genes linked to mitochondria. (A) G-MAD Manhattan plot of the respiratory electron transport (Reactome: R-HSA-611105) module in human. Genes are arranged based on their genetic positions and genes annotated to be involved in the module are colored red. Genes with absolute GMAS over 0.268 are considered significantly associated. DDT, BOLA3, and ARID1A are labeled. (B) Venn diagram of novel genes associated with respiratory electron transport module in human, mouse, and rat; 707 genes were predicted to be mito-proteins by G-MAD in all three species, and 351 genes, including DMAC1, NDUFAF8, FMCI, and BOLA3, were recently annotated to be involved in mitochondrial respiration in at least one species, whereas 356 genes, including DDT, C16orf91, C15orf61, FLAD1, and GRHPR, have not been previously linked with mitochondria based on the current annotations. The association results for all genes in human, mouse, and rat can be found in Supplemental Table S5. (C) DDT associates with mitochondrial respiratory chain modules in human. The threshold of significant gene-module association is indicated by the red dashed line. Modules are organized by module similarities. Known modules connected to DDT from annotations are highlighted in red and other modules with GMAS over the threshold are colored in black. Dot sizes reflect the GMAS of DDT against the respective modules. (D) Module similarity network showing the modules associated with DDT. Modules are plotted based on their layout in Figure 1B and colored based on their GMAS against DDT. (E) Mitochondrial localization of DDT in mouse embryonic fibroblasts (MEFs). DDT expression is overlapped with the Mitotracker red label. (F) DDT knockdown leads to the reduction of oxygen consumption rate (OCR) as a reflection of mitochondrial respiration in human HEK293 cells. Addition of specific mitochondrial inhibitors, including the oligomycin (ATPase inhibitor), FCCP (uncoupling agent), and rotenone/antimycin A (electron transport chain inhibitors), are indicated by arrows. (G) ARID1A negatively associates with the mitochondrial respiratory chain in human. The threshold of significant gene-module association is indicated by the red dashed line. Modules are organized by the module similarities. Known modules connected to ARID1A from extant annotations are highlighted in red and other modules with GMAS over the threshold are colored in black. Dot sizes are proportional to GMAS of the respective modules. (H) Module similarity network showing the modules associated with ARID1A. Modules are colored based on their GMAS against ARID1A. (I) Mice with the uterine-specific Arid1a knockout showed positive enrichment in mitochondrial respiration modules. Nominal P-values from the GSEA results are used to plot against normalized enrichment score (NESs), with dot sizes indicating the number of genes in the modules and transparencies indicating the false discovery rate (FDR). (J) Enrichment plot showing the enrichment of genes included in respiratory electron transport in uterus-specific Arid1a knockout mice compared to wild-type controls. Genes are ranked based on the fold change between Arid1a knockout and wild-type mice, and the ranking positions of genes in respiratory electron transport are labeled as vertical black bars. (NES) Normalized enrichment score.
Figure 5. Module-Module Association Determination (M-MAD) reveals module connections. (A) Scheme of the M-MAD methodology in detecting module connections. Intermediate results of G-MAD for all modules are further processed and used as the basis of M-MAD. The $-\log_{10}(P)$ values of G-MAD for the target module against all genes in each data set are used as the gene statistic for the module, and connections between the target module and all modules are calculated using CAMERA. The results are then meta-analyzed by taking the sample sizes and inter-gene correlations of all data sets to compute the module-module association score (MMAS) between modules. (B) Module association network showing the connections across all modules. Colors of nodes represent the modules defined in the global module similarity network in Figure 1B. Module clusters with respective colors are identified and labeled. Modules used as examples in the following figure panels are highlighted with a circle. (C) Comparison of pairwise module connections derived from module similarities in Figure 1B and associations (from M-MAD) in Figure 5B. A red dashed line is plotted when the pairwise module similarity equals association. The distributions of module similarity and association scores are illustrated in the top and at the right of the plot and are colored in red and blue, respectively. Two examples of novel module connections are encircled. (D,E) Subnetworks showing the association between mitochondrial and proteasomal modules (D), and mitochondrial and histone demethylation modules (E). Edge colors indicate the significance of module connections, with red as positive and blue as negative.
from G-MAD, were used to compute their associations against all modules. The enrichment scores of all genes for the target module were used as the gene-level statistics to calculate the enrichment against all modules using CAMERA (Wu and Smyth 2012). The resulting enrichment P-values across modules were transformed to 1, 0, or −1 based on the Bonferroni threshold and then meta-analyzed across all data sets to obtain the module-module association scores (MMASs) (Fig. 5A).

Module–module associations with an absolute MMAS of over 0.268, corresponding to 4% of the total number of module pairs, were considered significant and were used to construct a module association network (Fig. 5B). Modules were represented as nodes with the same colors as the module clusters from Figure 1B. While the module similarity network in Figure 1B is based solely on existing gene annotations, the module association network relies on analyzing the full expression data sets. It can thus reveal new biological connections among modules, which were not included in literature-based annotations. We compared the two networks (Supplemental Fig. S10) obtained from module similarity (Fig. 1B) and module association (Fig. 5B). There are numerous module pairs with no similarity (overlap of annotated genes) but with high association based on expression (M-MAD) (Fig. 5C). Moreover, many module pairs have predicted negative associations (Fig. 5C). Therefore, these results provide a resource for hypothesis generation and validation of the module connections.

By applying M-MAD, we observed a strong positive link between mitochondrial modules and the proteasome (Fig. S5D; Supplemental Fig. S11A–C). Most of the genes encoding for proteasomal subunits exhibit strong association with the ETC in human and mouse (Supplemental Fig. S11G), indicating a conserved regulatory mechanism. Dysfunction of mitochondria and the ubiquitin–proteasome system (UPS) are hallmarks of aging and aging-related neurodegenerative diseases, such as Alzheimer’s, Parkinson’s, and Huntington’s diseases (Ortega and Lucas 2014; Ross et al. 2015; D’Amico et al. 2017). Abnormalities that perturb the crosstalk between these two modules have been demonstrated to contribute to the pathogenesis of these diseases, and several mechanisms have been proposed (D’Amico et al. 2017; Harrigan et al. 2018). It has also been shown that ETC disruption leads to proteasome impairment (D’Amico et al. 2017), while conversely the inhibition of the UPS causes mitochondrial dysfunction (Ross et al. 2015).

Similar to G-MAD, M-MAD can also predict negative connections between modules. For example, we found strong negative connections between histone demethylation processes and mitochondrial modules (Fig. 5E; Supplemental Fig. S11D–F). The link between epigenetics and mitochondria is a research focus for many groups, including ours (Schoeder et al. 2013; Merkwirth et al. 2016; Tian et al. 2016). It has been reported that mitochondrial dysfunction affects histone methylation and, conversely, histone lysine demethylases can impact mitochondrial functions (Merkwirth et al. 2016). Most of the histone lysine demethylases showed negative associations with the ETC in human and mouse (Supplemental Fig. S11G), suggesting a conserved negative connection between histone demethylation and mitochondrial function.

As another example of M-MAD, we investigated modules connected with lipid biosynthetic modules. Of note, ribosome modules exhibited strong negative association with lipid biosynthetic modules (Fig. 6A,B; Supplemental Fig. S12A,B). This is in line with our previous finding that a ribosomal protein, Rpl26, negatively correlates with body weight and fat mass (Li et al. 2018). In support of this connection, liver and adipose transcripts of most ribosomal protein genes negatively correlated with metabolic phenotypes, such as body weight, fat mass, and cholesterol levels, in the BXD mouse cohort (Fig. 6C; Supplemental Fig. S12C; Wu et al. 2014), as well as in a CAST/EJ and C57BL6/J F2 intercross (Fig. 6D; Supplemental Fig. S12D; Schadt et al. 2008). Finally, RNAi targeting nine of the identified ribosomal protein genes out of total 13 tested led to the accumulation of lipid droplets in C. elegans (Fig. 6E; Supplemental Fig. S12E–G), further validating the robustness of the lipid synthesis-ribosome connection across species.

Discussion
Significant efforts in biological research have been devoted to defining the molecular and physiological functions of genes. However, many genes are still not well annotated or even remain uncharacterized (Edwards et al. 2011; Dolgin 2017; Stoeger et al. 2018). Here, we developed an approach, termed G-MAD, to facilitate the identification of novel gene functions and to establish robust connections between genes and modules. Using transcriptome data sets from cohorts ranging from human to mouse, rat, fly, worm, and yeast, we identified millions of gene-module connections, many of which are novel. Unlike most available sources relying on coexpression to predict gene functions, G-MAD can identify not only positive gene-module connections but also negative associations between genes and modules or processes. We illustrated the predictive power of G-MAD by predicting novel mitochondrial protein genes and validated DDT and BOLA3 through experiments. Meanwhile, tissue-specific functions of genes, for example, EHHADH and SLC6A1, can also be identified using data sets derived from respective tissues.

In addition, we extended G-MAD to M-MAD, to uncover connections between modules. Association scores of one module against all genes from G-MAD were used to compute its associations with all modules. Similar to G-MAD, M-MAD can identify both positive and negative module associations. For example, in humans we identified around 2,000,000 associations between all modules, over 170,000 of which were negative. We constructed a module association network based on these connected modules and compared it to the module similarity network. Many of the associated module pairs have low or no similarities in gene compositions. By applying M-MAD on the ETC module, we discovered a conserved connection between mitochondria and the proteasome in various organisms (D’Amico et al. 2017). In addition, we identified negative associations between histone lysine demethylation and mitochondrial modules, underscoring the inverse connection between epigenetic regulation and mitochondrial function (Schroeder et al. 2013; Merkwirth et al. 2016; Tian et al. 2016). Moreover, we discovered and validated a novel negative regulatory role of ribosomal proteins on lipid biosynthesis (Li et al. 2018).

In summary, we described here a set of approaches to identify gene function and module connectivity that we collectively termed GeneBridge to reflect their capacity to bridge genes to biological functions and phenotypes. The GeneBridge toolkit is accessible through our open web resource (systems-genetics.org) to the research community for hypothesis generation or validation. It should be noted that we selected a stringent threshold of 0.268 to limit the probability of detecting false positives. Researchers, however, have the possibility to fully explore the results by altering the thresholds on the open web resource. Although only protein-
coding genes were included in our analysis, the same approach can be applied to noncoding genes to reveal their potential functions. Similarly, GeneBridge can also be utilized to identify novel gene-disease associations based on known disease-associated genes from databases, such as the Human Disease Ontology (DO) (Schröml et al. 2019) or DisGeNET (Piérot et al. 2017). The GeneBridge toolkit could also be applied to large-scale proteomics data sets after correcting for the background of all measured proteins. Integration of GeneBridge with other well-established databases, such as BioGRID (Stark et al. 2006) and STRING (Szklarczyk et al. 2015), will facilitate the investigation of the connections between genes, modules, and diseases.

Methods

Gene-Module Association Determination (G-MAD)

G-MAD makes use of the PEER-resulted expression residuals (Stegle et al. 2012) of transcriptome data sets from large cohorts (data sets with over 80 samples). The expression levels of the gene-of-interest (target gene $T$) are used as a continuous trait to test whether a module $M$ is enriched when $T$ is highly expressed or, alternatively, whether it is depleted. The analysis uses the competitive gene set testing method CAMERA, which adjusts for inter-gene correlations (Wu and Smyth 2012). This adjustment is important because, left unadjusted, too many significant results would emerge. To perform CAMERA, we first regress all genes $G$ on $T$ according to the following relationship:

$$G = \mu + \beta_{T-G}T + \epsilon.$$  

The fitting of this model equation to the observations is done separately for each data set by using the least squares method. The result is one fitted value $\beta_{T-G}$ per gene. These coefficients define a set of statistics numerically characterizing the connection between the target gene $T$ and any gene $G$. CAMERA provides a test of the null hypothesis that the average values of the $\beta$ coefficients for the genes $G$ in the module $M$ are equal to the values for the genes not in the module. In order to correct for the inter-gene correlations, a variance inflation factor is computed based on the average correlation coefficient $\rho_M$ computed from the expression residuals obtained from PEER and only using the genes in the module $M$. When the average association scores between genes in the set and genes outside the set, $\sum_{G \in \text{genes} \setminus M} \beta_{T-G}/|M|$ and $\sum_{G \in \text{genes} \setminus M} \beta_{T-G}/(|\text{genes} \setminus M|)$, are compared on the final step, $\rho_M$ is included in the variance inflation factor. The resulting statistic revealing the association between the target gene $T$ and module $M$ we refer to as the enrichment score $ES_M(T)$. The same procedure was conducted for all the genes in the analyzed data sets to obtain the enrichment $P$-value matrix between genes and modules in all the data sets. Two types of analyses can be applied on the gene-module $P$-value matrix. One can extract the $P$-values for one gene against all modules across the data sets to
obtain the association between this gene and all modules; or extract the P-values for one module against all genes to check the association between this module and all genes. To restrict the final scores into the range of \((-1, 1)\), we converted the P-values to 1/0/\(-1\) based on the significance threshold using Bonferroni corrections for each data set (i.e., the thresholds are either 0.05/\# genes when assessing genes for a given module or 0.05/\# modules when assessing modules for a given gene). Gene-module associations with P-values that survived multiple testing corrections were set to 1 or \(-1\), based on the enrichment direction, and 0 otherwise:

\[
S(p_{GM}) = \begin{cases} 
1, & p_{GM} < -0.05 ~ \text{# modules}, \\
-1, & p_{GM} > 0.05 ~ \text{# modules}, \\
0, & \text{otherwise}.
\end{cases}
\]

where \(p_{GM}\) are one-sided P-values, corresponding to either positive or negative associations. The resulting \(S(p_{GM})\) values were then meta-analyzed across the data sets, and the gene-module association scores (GMAS) were computed as the weighted averages of the scores with the weights functions of the sample sizes combined with the inter-gene correlation coefficients within modules. In this way, data sets with more samples and with higher coexpression of genes in modules are given more weight. Denote \(D_j\), \(j = 1, \ldots, J\) available data sets with corresponding sample sizes \(n_j\), \(j = 1, \ldots, J\), and average inter-gene correlations \(p_{ij}\), \(i, j = 1, \ldots, J\). Let the P-value obtained for the \(j\)th data set be \(p_{GM}(j)\). The final association score is then computed as

\[
\text{GMAS} = \frac{\sum_{j=1}^{J} w_j S(p_{GM}(j))}{\sum_{j=1}^{J} w_j},
\]

where weight for the \(j\)th data set is \(w_j = \sqrt{n_j} \hat{p}_j\). Under the null hypothesis, if we consider the positive and negative associations separately, the random variables \(S(p_{GM}(j))\) follow a Bernoulli distribution with probability of success \((0.05/\#\) modules). Therefore, statistic GMAS is the weighted sum of Bernoulli variables, whose theoretical distribution is hard to establish. The weight is proportional to the square root of the sample size in the \(j\)th data set. Another important component of \(w_j\) is the average correlation coefficients among genes in the module in the \(j\)th data set, \(\hat{p}_j\) which reflects the coexpression or “level of activation” of the module for this data set.

In summary, G-MAD employed PEER to remove the influence of covariates, CAMERA to compute P-values between genes and modules in each data set, Bonferroni correction to transform P-values into the range of \([-1, 1]\), and meta-analysis to summarize the results from all data sets.

For the final decision, we computed the true positive rate (percentage of known genes above the threshold against all known genes) and false positive rate (percentage of unknown genes above the threshold against all unknown genes) by varying the threshold of significance. We noticed that decreasing the threshold would increase the true positive rate (TPR) but also the false positive rate (FPR). Therefore, we selected a very stringent threshold for GMAS of 0.268, where only 10% of the known (TPR) and 0.24% of the unknown (FPR) gene-module connections are recovered.

Module-Module Association Determination (M-MAD)

M-MAD takes the association P-value matrix between a target module and all genes computed by CAMERA in all data sets (Fig. 1A, bottom left) and uses the \(-\log_{10}(P)\) transformed values as a continuous measure to test whether other biological modules are enriched by having genes that are highly associated with the target module. As CAMERA generates P-values that are uniformly distributed, \(-\log_{10}(P)\) transformed values have an exponential distribution skewed toward 0. The following analysis again uses the competitive gene set testing method CAMERA to compute a P-value for testing the equality of the average transformed values for the genes in the other biological modules compared to all other genes. It will result in a small P-value when many of the genes in the other biological modules are relatively highly connected to the target module. The same analysis is performed for all modules to achieve a final association P-value matrix between modules. The Bonferroni correction was used to correct for the multiple testing errors with 0.05/\#modules as the significance threshold. To constrain the final score into the range between \(-1\) and 1, module-module connections with enrichment P-values that survived multiple testing corrections were allocated 1 or \(-1\), based on the enrichment directions, and 0 otherwise. The results were then meta-analyzed across the data sets, and the module-module association scores (MMAS) were computed as the weighted averages of the connection scores by the sample sizes and inter-gene correlation coefficients within modules across data sets.

Cross-validation

In order to test the predictive performance of G-MAD and compare it with the other methods using coexpression (including WeGET, COXPRESdb, and average \(r\)), we performed a cross-validation analysis by removing groups of genes from modules, recomputing the associations between the removed genes and the reduced module, and testing if we can rediscover the removed genes (Szkulczyn et al. 2016). We applied leave-one-out cross-validation for modules with no more than 50 genes, and 10-fold cross-validation for larger modules. The area under the receiver operating characteristic curve is used to estimate the performance of prediction, with an AUC of 1 indicating perfect prediction and 0.5 indicating random guess. Details of the methods are described in the Supplemental Methods.

Validation experiments in vitro and in C. elegans

Some of the novel associations were validated either using cell models or C. elegans. Detailed procedures for the validation experiments are described in Supplemental Methods.

Data access

The data and source codes from this study are available in the Supplemental Code and from https://systems-genetics.org/ and https://github.com/lihaone/GeneBridge, respectively.

Acknowledgments

We thank the research groups who made these data publicly available for systems biology research. We thank the entire J.A. lab for comments and discussions. We thank N. Agarwal for help in data preprocessing. H.L. is the recipient of a doctoral scholarship from the China Scholarship Council. This work was supported by grants from the École Polytechnique Fédér diesel de Lausanne (EPFL), the European Research Council (ERC-AdG-787702), the Swiss National Science Foundation (SNSF 31003A_179435), the GRL grant of the National Research Foundation of Korea (NRF 2017K1A1A2013124), the AgingX program of the Swiss Initiative for Systems Biology (RTD 2013/153), and the National Institutes of Health (R01AG043930).

Author contributions: Conceptualization: H.L. and J.A.; data curation: H.L.; formal analysis: H.L., D.R., S.M., and J.A.; funding acquisition: R.W.W., M.R.-R., K.S., and J.A.; methodology: H.L.,...
Understanding multicellular disease and diseases with human tissue–mammalian organs.

Nature Genet 47: 569–576. doi:10.1038/ng.3259

Harrigan JA, Jacq X, Martin NM, Jackson SP. 2018. Debiquitylating enzymes and drug discovery: emerging opportunities. Nat Rev Drug Discov 17: 57–78. doi:10.1038/nrd.2017.152

Hein MY, Hubner NC, Poser I, Cox J, Nagara N, Toyoda Y, Gak IA, Weisweiss I, Mansfeld J, Buchholz F, et al. 2015. A human interactome in three quantitative dimensions organized by stoichiometries and abundances. Cell 163: 713–723. doi:10.1016/j.cell.2015.09.012

Horber MA, Xu A, Wang M, Bennett NK, Park CY, Bogdanoff D, Adamson B, Chow ED, Kampmann M, Peterson TR, et al. 2018. Mapping the genetic landscape of human cells. Cell 174: 953–967.e22. doi:10.1016/j.cell.2018.06.010

Huttlin EL, Bruckner RJ, Paulo JA, Cannon JR, Ting L, Baltier K, Colby G, Gebreb F, Gygi MP, Parzen H, et al. 2017. Architecture of the human interactome defines protein communities and disease networks. Nature 545: 505–509. doi:10.1038/nature22766

Jiang Y, Oron TR, Clark WT, Bankapur AR, D’Andrea D, Lepore R, Funk CS, Kahanda I, Verspoor KM, Ben-Hur A, et al. 2016. An expanded evaluation of protein function prediction methods shows improvement in accuracy. Genome Biol 17: 184. doi:10.1186/s13059-016-1037-6

Khatri P, Stroia M, Butte AJ. 2012. Ten years of pathway analysis: current approaches and outstanding challenges. PLoS Comput Biol 8: e1002375. doi:10.1371/journal.pcbi.1002375

Kim J, Joon AA, Cho H, Kremer LS, Alston CL, et al. 2016. Mitochondrial protein import deficiency of the mitochondrial stress response. Nat Neurosci 19: 163–171. doi:10.1038/nn.4159

Koh JL, Toufighi K, Mostafavi S, et al. 2010. The genetic landscape of a gene coexpression database for 11 animal species supported by 23 multiple respiratory chain and 2-oxoacid dehydrogenase enzymes. Am J Hum Genet 87: 621–631. doi:10.1016/j.ajhg.2010.05.016

Kremer LS, Minnaert SM, Ahonen J, et al. 2015. MitoCarta2.0: an updated inventory of mammalian mitochondrial proteins. Nucleic Acids Res 43: D1251–D1257. doi:10.1093/nar/gkv1003

Larivièr L, Czeck P, Bairoch A, et al.), pp. 124–131. Springer Berlin Heidelberg, Berlin, Heidelberg.

Li Y, Carvill GL, McMahon JM, Schneider A, Zemel M, Myers CT, Saykally J, Carroll RG, Zaslavsky S, Terenin A, et al. 2016. Bgee: integrating and comparing heterogeneous transcriptional and proteomic data among species. In Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, Eisenberg D. 1999. Detecting protein function and protein-protein interactions from genome sequences. Nucleic Acids Res 27: 423–431. doi:10.1093/database/27.3.423

Lissanu Deribe Y, Sun Y, Terranova C, Khan F, Martinez-Ledesma J, Gay J, Li Y, Agarwal P, Rajagopalan D. 2008. A global pathway crosstalk network. Nucleic Acids Res 36: D620–D627. doi:10.1093/nar/gkm892

Liu Y, Calvo SE, Colve SA, Li S, Mootha VK. 2016. COXPRESdb v7: computational network analysis of human and mouse gene expression and evolutionary conservation for the assembly/stability of yeast mitochondrial F1-ATPase in heat stress conditions. J Biol Chem 281: 6789–6796. doi:10.1074/jbc.M009572000

Lupoglou A, Li Y, Silverman L, Zhou J, Ciriacy-Turcu A, Pressoir G, Mehta A, Kolla R, Parthasarathy S, et al. 2010. A comprehensive systems approach to identify and validate novel therapeutic targets of the ubiquitin–proteasome system. PLoS One 5: e11441. doi:10.1371/journal.pone.0011441

Maitra A, Agarwala P, Bhardwaj SK, et al. 2009. The Gene Ontology Consortium. Genome Biol 10: R1. doi:10.1186/gb-2009-10-1-r1

Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, Eisenberg D. 1999. Detecting protein function and protein-protein interactions from genome sequences. Nucleic Acids Res 27: 423–431. doi:10.1093/database/27.3.423

Marshall KA, Phillippy KH, Sherman PM, Holko M, et al. 2013. NCBI databases. Nucleic Acids Res 41: D1–D32. doi:10.1093/nar/gkt1193

Mathias AC, Ruben JM, Shokat KM, et al. 2010. The unified database data among species. In Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, Eisenberg D. 1999. Detecting protein function and protein-protein interactions from genome sequences. Nucleic Acids Res 27: 423–431. doi:10.1093/database/27.3.423

Meyers RC, Gingeras TR. 2003. cDNA microarray analysis of gene expression. Genome Biol 4: 223. doi:10.1186/gb-2003-4-7-223

Mitalipov SM, Deem LW, et al. 2004. WebQTL: rapid exploration of gene expression and genetic networks for brain and behavior. Nat Neurosci 7: 485–486. doi:10.1038/nn1044

Mukherjee S, Islam MM, Uddin M, et al. 2015. A curated database for mammalian mitochondrial proteins. Nucleic Acids Res 43: D559–D562. doi:10.1093/nar/gku1057

Nguyen J, Robbiano A, Zara F, Specchio N, et al. 2015. Mutations in the α subunit of the cytoplasmic ATP synthase F0-ATPase cause a fatal deficiency of multiple respiratory chain and 2-oxoacid dehydrogenase enzymes. Am J Hum Genet 96: 808–815. doi:10.1016/j.ajhg.2015.02.016

Owen KR, Bibikova M, Chiu H, et al. 2011. Mutations in the ATP synthase F0–F1 complex cause a fatal deficiency of multiple respiratory chain and 2-oxoacid dehydrogenase enzymes. Am J Hum Genet 96: 808–815. doi:10.1016/j.ajhg.2015.02.016

Ortega Z, Lucas JJ. 2014. Ubiquitin-proteasome system involvement in Huntington’s disease. Front Mol Neurosci 7: 505–509. doi:10.3389/fnmol.2014.00505

Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. 1999. Assigning protein functions by comparative genome analysis: protein–protein interaction mapping identifies regulators of respiratory chain function. Mol Cell 3: 621–632. doi:10.1016/S1097-2765(99)80158-4

Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. 1999. Assigning protein functions by comparative genome analysis: protein–protein interaction mapping identifies regulators of respiratory chain function. Mol Cell 3: 621–632. doi:10.1016/S1097-2765(99)80158-4

Qin K, Xie Y, Haider SM, et al. 2016. The mitochondrial proteome in human disease. Trends Biochem Sci 41: 537–549. doi:10.1016/j.tibs.2016.01.005

Rice DJ, McCallum D, et al. 2007. Human protein reference database—2007 update. Nucleic Acids Res 35: D559–D562. doi:10.1093/nar/gkm1155

Ritzau E, Fazio A, et al. 2014. COXPRESdb v7: computational network analysis of human and mouse gene expression and evolutionary conservation for the assembly/stability of yeast mitochondrial F1-ATPase in heat stress conditions. J Biol Chem 281: 6789–6796. doi:10.1074/jbc.M009572000
