GeneFishing to reconstruct context specific portraits of biological processes

Ke Liu[a], Elizabeth Theusch[b], Yun Zhou[a], Tal Ashuach[c], Andrea C. Dose[b], Peter J. Bickel[a,c,1], Marisa W. Medina[b,1], and Haiyan Huang[a,c,1]

[a]Department of Statistics, University of California, Berkeley, CA 94720; [b]Children's Hospital Oakland Research Institute, Oakland, CA 94609; and [c]Center for Computational Biology, University of California, Berkeley, CA 94720

Contributed by Peter J. Bickel, July 23, 2019 (sent for review December 7, 2018; reviewed by Sunduz Keles and Nancy R. Zhang)

Rapid advances in genomic technologies have led to a wealth of diverse data, from which novel discoveries can be gleaned through the application of robust statistical and computational methods. Here, we describe GeneFishing, a semisupervised computational approach to reconstruct context-specific portraits of biological processes by leveraging gene–gene coexpression information. GeneFishing incorporates multiple high-dimensional statistical ideas, including dimensionality reduction, clustering, subsampling, and results aggregation, to produce robust results. To illustrate the power of our method, we applied it using 21 genes involved in cholesterol metabolism as "bait" to "fish out" (or identify) genes not previously identified as being connected to cholesterol metabolism. Using simulation and real datasets, we found that the results obtained through GeneFishing were more interesting for our study than those provided by related gene prioritization methods. In particular, application of GeneFishing to the GTEx liver RNA sequencing (RNAseq) data not only reidentified many known cholesterol-related genes, but also pointed to glyoxalase I (*GLO1*) as a gene implicated in cholesterol metabolism. In a follow-up experiment, we found that *GLO1* knockdown in human hepatoma cell lines increased levels of cellular cholesterol ester, validating a role for *GLO1* in cholesterol metabolism. In addition, we performed pantissue analysis by applying GeneFishing on various tissues and identified many potential tissue-specific cholesterol metabolism-related genes. GeneFishing appears to be a powerful tool for identifying related components of complex biological systems and may be used across a wide range of applications.

context-specific gene functional groups | cholesterol metabolism | gene prioritization | gene pathways | pantissue analysis

Systems biology was first introduced into the language of modern biology in the early 21st century (1, 2). It is an interdisciplinary research field that focuses on understanding a big picture of how small cell components (such as RNAs and proteins) interact in complex biological systems. Over the past 2 decades, along with the rapid development of high-throughput experimental and computational tools, the field of systems biology has advanced greatly. This advance has been driven to a considerable extent by the collaboration of researchers in biology and quantitative fields. Large collaborative efforts have made significant contributions to systems biology research under many aspects: experimental, computational, and philosophical (3–5). In this article, we propose a tool that should be helpful in one of these aspects—the study of reconstructing comprehensive context-specific portraits of biological processes using gene expression data and the change in such portraits across different contexts (such as tissue types, disease status, and so on). The variability in such portraits gives rise to the diverse functional behaviors of biological systems.

The types of questions that we tackle here have been and continue to be considered extensively in the literature under the heading of "gene prioritization" for a specific biological process or pathway. An excellent review of developments up to 2012 and exposition of developments to come may be found in referenced literatures (6). In particular, they point to the general "guilt by association" principle and extensions of the principle that they call "edge prioritization" or "generating hypotheses about potential interactions between top candidates and seed (bait) genes." Although many tools [such as GIANT (7) and ENDEAVOUR (8)] have used this principle and been successful in many applications (7–12), some issues need to be addressed further. One issue is the low signal-to-noise ratio in data. On the one hand, it is believed that the great majority of genes across the whole genome have no relationship with the process of interest. On the other hand, gene–gene coexpression, one of the most highly used measures in guilt by association procedures, often generates results with high false positive rates. Thus, when whole-genome gene–gene coexpression is considered, the sheer number of gene pairs that are coexpressed randomly may outweigh that of gene pairs with coexpression that reflects underlying biology. Another issue is the selection of "seed (bait)" genes. Although this is not a statistical issue (i.e., it largely depends on the biological question of interest), it calls for an assessment of the sensitivity of the conclusions to choice of the bait set. The current literature lacks such systematic sensitivity analysis. Our view is that a successful statistical method needs to produce scientific conclusions that are in part unexpected on the basis of current (maybe

Significance

Biological systems function through the interaction of numerous molecules influencing a variety of biochemical reactions. However, most biological systems are still only partially understood. This paper introduces GeneFishing, a method for "fishing out" candidate genes in a biological process. The method is "semisupervised" using a set of "bait" genes (i.e., ones previously known to be relevant to the same process). GeneFishing effectively combines modern and traditional statistical ideas for analyzing both big and small data. We applied this method to cholesterol-related genes and identified several interesting phenomena. GeneFishing has the potential for pointing to functional importance in known but poorly studied genes, and its underlying framework is broadly applicable inside and outside biology.

Author contributions: K.L., E.T., P.J.B., M.W.M., and H.H. designed research; K.L., E.T., A.C.D., P.J.B., M.W.M., and H.H. performed research; K.L., E.T., Y.Z., T.A., P.J.B., M.W.M., and H.H. analyzed data; and K.L., E.T., P.J.B., M.W.M., and H.H. wrote the paper.

Reviewers: S.K., University of Wisconsin–Madison; and N.R.Z., University of Pennsylvania.

The authors declare no conflict of interest.

Data deposition: GeneFishing is freely available at https://github.com/tomwhoooo/GeneFishingPy.

[1]To whom correspondence may be addressed. Email: bickelp@berkeley.edu, mwmedina@chori.org, or hyh0110@berkeley.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1820340116/-/DCSupplemental.

Published online September 4, 2019.

imperfect) knowledge and in part conform to well-known biology. These 2 issues call for gene prioritization methods that are effective for the analysis of large noisy datasets with sparse signal and are also robust against possible noise in the seed (bait) genes.

To address these issues, we combine several high-dimensional statistical techniques, including dimensionality reduction, clustering, subsampling, and aggregation of results (motivated by a bagging-like idea), in a way to develop a method that we call GeneFishing. We attempt to identify all genes in the genome with an expression activity pattern related to that of most of the bait genes under the same conditions. We note that, while datasets may be "large" in terms of the number of measured variables (i.e., genome-wide genetic features), the sample size of these data could be limited (e.g., there might be only hundreds of individuals per dataset). Thus, powerful techniques for big data, such as deep learning, may be an overkill, while approaches that effectively combine modern and traditional statistical ideas for analyzing both big and small data may be more effective.

We evaluated our method through an application to cholesterol metabolism, a highly characterized biological process. Using a set of preidentified 21 "bait genes," all of which have known roles in cholesterol metabolism, we applied GeneFishing to 3 independent RNAseq (RNA sequencing) datasets of human lymphoblastoid cell lines and found that our approach not only identified other genes with known roles in cholesterol metabolism but also, did so with high levels of consistency across the 3 datasets. Additional application of this approach to the GTEx (Genotype-Tissue Expression project) human liver RNAseq data identified 56 genes, of which 11 were prioritized for functional validation studies in human hepatoma cell lines. From this analysis, we identified gene glyoxalase I (*GLO1*), with expression levels that are highly correlated with known cholesterol-related genes. More importantly, in a follow-up wet laboratory experiment, we found that *GLO1* knockdown increased levels of cellular cholesterol esters. In addition, we performed pantissue analysis by applying GeneFishing to GTEx expression data from a large collection of tissues and identified many potential tissue-specific cholesterol metabolism-related genes. These findings demonstrate the ability of GeneFishing to identify genes relevant to previously defined biological pathways in a context-specific manner. The strategy is obviously generalizable to the study of other aspects of biological systems and may be used across a wide range of applications outside biology.

Since our approach falls into the guilt by association paradigm, we recognize that, as usual, association does not imply causation, and many of the genes that we point to may well be involved with other functions of cholesterol than ones governing metabolism.

Results

A Motivating Example. A major challenge of genome-wide analyses is how to extract sparse signals from large-scale datasets, which tend to be heterogeneous and noisy. To illustrate how the level of noise in the data increases the complexity of detecting genes involved in a specific biological process, we performed a simple study of the cholesterol metabolic process using transcriptomic measures from 426 LCLs (lymphoblastoid cell lines) derived from participants of the CAP (Cholesterol and Pharmacogenetics) statin clinical trial (13) (CAP-LCLs). This is one of the major datasets that we use in this paper to demonstrate the performance of our GeneFishing method.

From Ensembl BioMart (https://www.ensembl.org/biomart/martview/7f44660a1147fceb60a6845325da0ca5), we extracted 120 genes that are annotated with the GO BP (Gene Ontology biological process) term "GO:0008203 cholesterol metabolic process," of which 82 are expressed in the CAP-LCL dataset. We first measured coexpression of all gene pairs as the absolute value of Spearman rank correlation of gene expression values across subjects. Thus, our data can be thought of as a T × T gene coexpression matrix (here, T = 82). We next performed spectral analysis based on the coexpression matrix to project each gene onto the space of the first 2 non-0 eigenvectors of the normalized Laplacian matrix and identified a tight cluster of 21 genes (Fig. 1*A*), 18 of which encode enzymes in the cholesterol biosynthesis pathway (14), with the remaining 3 genes known to be involved in the transcriptional regulation of these 18 genes (i.e., *INSIG1* and *SREBF2*) or complementary functions (*LDLR*, the key regulator of low density lipoprotein [LDL] uptake) (*SI Appendix*, Fig. S1 and Table S1). To test whether this tight cluster persisted in the context of other genes, we repeated the analysis using gene sets composed of the 21 genes as well as 100, 1,500, and 2,000 random genes (Fig. 1 *B* to *D*). Since the majority of genes should be unrelated to cholesterol metabolism, we expect that the sheer number of pairs of such genes outweighs those that show patterned relations among our subjects. As shown in Fig. 1*B*, the 21 genes created an obvious cluster when mixed with 100 random genes. However, this cluster became obscured in the presence of larger sets of

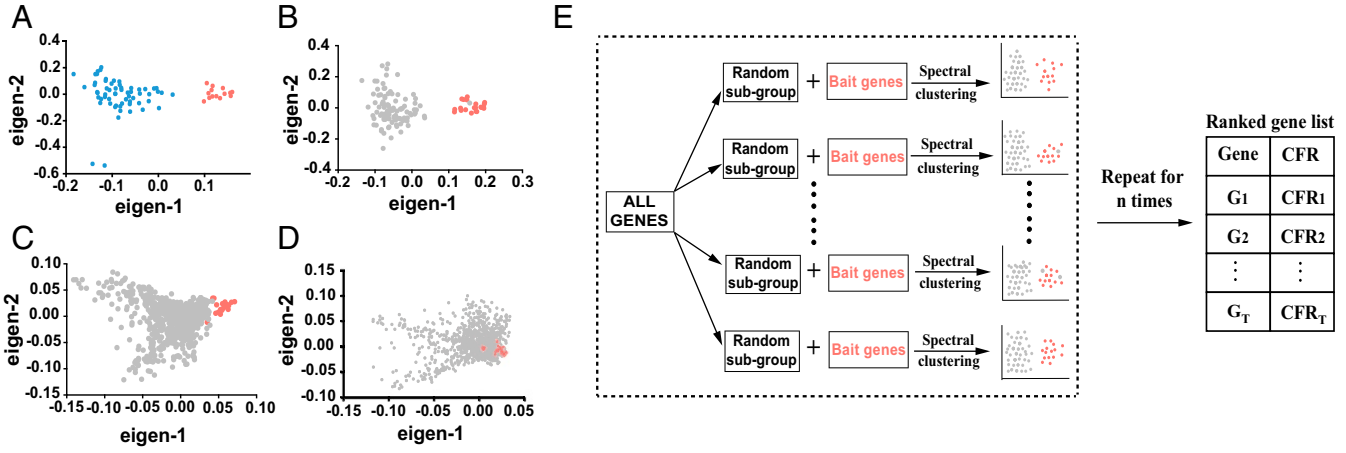

Fig. 1. Motivation and workflow of GeneFishing. (*A* to *D*) Spectral clustering plot of the 21 bait genes (colored in red) with another 61 genes (colored in blue) associated with the GO BP term "cholesterol metabolic process" (*A*), and 100 (*B*), 1,500 (*C*), and 2,000 (*D*) random genes (colored in gray). (*E*) Workflow of GeneFishing.

random genes as shown in Fig. 1 *C* and *D*. These results illustrate how the information provided by the 21 cholesterol genes is progressively obscured by random noise patterns with increasing numbers of random genes.

The GeneFishing Procedure. Our goal is to develop an effective procedure to identify genes relevant to known biological processes using transcriptomic data. Leveraging the clustering of the 21 cholesterol-related genes observed above, we develop GeneFishing, a semisupervised, nonparametric clustering procedure based on a bagging-like idea to reconstruct portraits of biological processes of interest in varying context. The input data of GeneFishing are an M × T matrix representing the normalized expression values of T genes across M subjects together with a small set of preidentified genes known to be relevant to the biological process of interest (such as the 21 genes mentioned in the motivating example). This set of genes can be used as "bait" genes to guide our search for additional genes that are potentially relevant to the biological process.

The GeneFishing flowchart is shown in Fig. 1*E*. Given bait genes, step 1, reduction of search space, is key in our method, facilitating the pulling of "signal" out of the "noise." In particular, the candidate genes are randomly split into many subsearch spaces of m genes each (e.g., m = 100). The bait genes are then added to each of the candidate gene subsets. In step 2, coexpression matrices are constructed for gene pairs contained within each subsearch space, and the spectral clustering algorithm was applied to each matrix separately. The current implementation uses Spearman rank correlation to generate gene coexpression matrices. Other coexpression measures can be more appropriate in other contexts as discussed in refs. 15–17. While in most cases, the bait genes cluster separately from the candidate genes, in some instances candidate gene(s) will cluster with the bait genes (e.g., when a gray dot clusters within the red dots as shown in Fig. 1*B*). When this occurs, we consider the candidate gene to be "fished out." Since a candidate gene may randomly cocluster with the bait genes, we repeat steps 1 and 2 (defining 1 round of GeneFishing) n times (e.g., n = 1000). In step 3, the results from all rounds are aggregated. The final output is a table that records the "capture frequency rate" (CFR; the ratio of the number of times that each candidate gene has been fished out in the n rounds of GeneFishing to n). We consider fished out genes with large CFR values as "discoveries." Note, however, that we can only conclude that these discoveries are likely functionally related to the bait genes, not that they perform a specific or similar function as the bait genes. Complete technical details of the GeneFishing procedure as well as the computation of approximate P values and false discovery rates (FDRs) are provided in *Methods* and *SI Appendix*.

Evaluating GeneFishing with Real and Simulated Datasets. All statistical models (or methods) in genomics are crude approximations to reality. They are used to generate procedures and provide measures using model-based inferences of the potential validity of perceived findings. In the usual case when we lack reliable models for some of the biological systems of interest, we focus on 3 minimum requirements: interpretability, replicability, and stability (18). By interpretability, we mean that some of the results can be related to known biology and ideally, guide further experimental study. Replicability refers to the stability of conclusions when the same methodology is applied to similar independent datasets. Stability means that conclusions should vary little under small statistical perturbations of the data and the model.

Interpretability. We first assessed whether the discoveries derived from GeneFishing were biologically plausible. Since genes involved in sterol metabolism are themselves well known to be transcriptionally coregulated, we used the 21 genes discussed in our motivating example as bait genes and applied GeneFishing to the CAP-LCL dataset. We noted that the CFR distribution of GeneFishing was strongly bimodal, which indicates a very natural cutoff for CFR (Fig. 2*A*). Finally, we identified 27 genes with CFR ≥ 0.99 (*SI Appendix*, Table S2). Interestingly, 10 of these had known roles in lipid or sterol metabolism and included *TMEM55B*, which we had previously identified as a cholesterol-regulatory gene based on its very high degree of coexpression with *HMGCR*, 1 of the 21 bait genes (19).

Replicability. To assess replicability, we tested the performance of GeneFishing in 2 other independent LCL datasets: the GEUVADIS-LCL (20) (462 lymphoblastoid cell lines from Genetic European Variation in Disease project) dataset and the GTEx-LCL (4) dataset (118 lymphoblastoid cell lines from GTEx project). We first checked the expression of the 21 bait genes in both datasets and observed clear clustering of the 21 genes again by spectral analysis (*SI Appendix*, Fig. S2 *A* and *B*). We then applied GeneFishing to each dataset using the 21 genes as bait and tested the overlap within the top t fished out genes (ordered by CFR values with t varying from 20 to 100) between the 3 (CAP, GEUVADIS, and GTEx). For benchmarking purposes, we also compared GeneFishing with other methods, including WGCNA (21) (weighted correlation network analysis, an unsupervised approach for finding gene coexpression clusters) and 3 different versions of guilt by association approaches (i.e., the association between a candidate gene and the set of bait genes is evaluated by the mean, median, and maximum of the Spearman rank correlations between the candidate and each of the bait genes, respectively). Of the methods tested, GeneFishing had the best (or equally good) replicability (Fig. 2*B* and *SI Appendix*, Fig. S2*C*).

Stability. Using the CAP-LCL dataset, we assessed the stability of GeneFishing under the following 3 scenarios: (*i*) when random genes are included in the set of bait genes (i.e., there is noise in the bait set), (*ii*) when only a subset of the 21 genes is used as baits, and (*iii*) when the method is applied to subsamples of all subjects (e.g., 80% of subjects were used to construct a gene–gene coexpression matrix when performing GeneFishing). As shown in Fig. 2*C*, the CFR values of each scenario were reasonably correlated with that derived from original CAP-LCL dataset, especially for high CFRs (e.g., when CFR > 0.9). This suggests that GeneFishing is quite robust to small perturbations of the input dataset. We also performed a simulation study to further investigate the stability of GeneFishing, and the results are presented in *SI Appendix*.

Application of GeneFishing to Liver and a Follow-Up Wet Laboratory Experiment Implicate *GLO1* as a Cholesterol Metabolism Regulator. Since the liver is the major organ that impacts plasma cholesterol, we applied GeneFishing to the GTEx human liver RNAseq dataset (119 samples). After confirming clear clustering of the 21 bait genes (*SI Appendix*, Fig. S3*A*), we identified 56 genes with a CFR ≥ 0.99 (*SI Appendix*, Table S3). GO term enrichment analysis (with R package GOStats) (22) identified substantial enrichment for multiple GO terms related to sterol metabolism, including "lipid metabolic process" (FDR = 7.56E-09) and "lipid biosynthetic process" (FDR = 5.29E-07). Next, since many genes involved in cholesterol metabolism are themselves transcriptionally regulated by cellular sterols, we sought to determine if any of the 56 genes showed evidence of sterol regulation. We performed transcriptome-wide sequencing on HepG2 cells that were first sterol depleted (incubated with 2 μM simvastatin + 10% lipoprotein-deficient serum for 24 h), after which 50 μg/mL low-density lipoprotein cholesterol (LDLC) was added back and incubated for an additional 24 h. Of the 56 genes, transcript levels of 28 genes were changed in response to sterol depletion (adjusted P value < 0.05), with effects reversed by LDLC add back (*SI Appendix*, Table S3). Interestingly, 13 of the 56

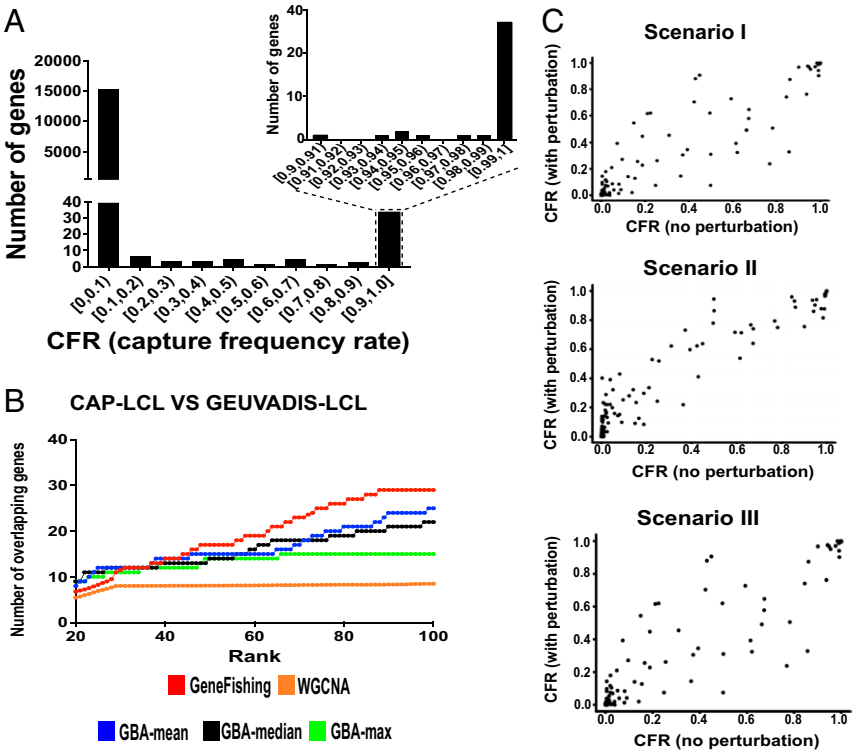

Fig. 2. Evaluation of GeneFishing. (*A*) Distribution of CFR values when GeneFishing was applied to the CAP-LCL dataset. (*B*) For each method, 2 ranked gene lists were generated by applying the method to the CAP-LCL and GEUVADIS-LCL datasets. Each colored curve corresponds to a gene prioritization method, plotting the number of overlapped genes between the 2 lists up to a rank position (*y* axis) against the rank (*x* axis). GBA is short for guilt-by-association. (*C*) Scatterplots of the CFR values when GeneFishing was applied to the raw CAP-LCL dataset and 3 randomly perturbed datasets.

genes did not appear to be changed in response to sterol depletion (P value > 0.5); 6 of the 56 genes were not expressed at a high-enough level in the HepG2 cells to meet the minimum threshold for expression. Several of the genes identified (e.g., *MMAB*, *SNAI3-AS1*) appeared to share promoter elements with 1 of the 21 bait genes (*SI Appendix*, Fig. S3*B*).

Of the genes not previously implicated in cholesterol metabolism, we tested the effect of knockdown of 11 of these genes on measures of intracellular cholesterol. We purposefully selected some genes that showed no evidence of sterol regulation (e.g., *GLO1*, *TDRKH*, *TTC39B*, and *C2orf82*) (Fig. 3*A*), as the reason why and/or how these genes may have been identified by GeneFishing was unclear. Huh7 cells were reverse transfected with siRNAs (silencing RNAs) targeting each gene of interest or a nontargeting control siRNA, and after 48 h, changes in gene expression and cellular cholesterol were quantified by qPCR and via the Amplex Red Cholesterol assay, respectively (Fig. 3*A*). Knockdown of 2 genes, *GLO1* and *RDH11*, significantly impacted transcript levels of *SQLE*, which encodes an enzyme in the cholesterol synthesis pathway (Fig. 3*B*). This change was confirmed in a second human hepatoma cell line, HepG2 (Fig. 3*C*). In addition, consistent with the increase in *SQLE* levels, we found that *GLO1* knockdown significantly increased cellular cholesterol esters in both Huh7 and HepG2 cells (Fig. 3*D*).

Pantissue GeneFishing Analysis. The cholesterol metabolic process functions widely in different human tissues. Motivated by the success of GeneFishing in the application to the GTEx liver data, we next sought to determine if the strong clustering of the 21 bait genes was also observed in other tissue types. In more detail, given a tissue, we performed the same spectral clustering analysis as in Fig. 1*A* and computed 2 statistics: tightness (defined as the ratio between within-cluster sum-of-squares and total sum-of-squares) of the cluster that contains most of the 21 genes and Jaccard index between the cluster and the 21 bait genes. Most tissues exhibited the 21 genes as a tight cluster. However, the 21-gene module was not apparent in some tissues due either to stronger coexpression with genes outside of the 21-gene module (e.g., adrenal gland) or to complete absence of coexpression (e.g., skeletal muscle) (Fig. 4*B*). Although it is well established that genes in the cholesterol synthesis pathway are coregulated, the change in their coexpression pattern that we observed across different tissues indicates an unexpectedly high degree of tissue specificity of such coregulation and meanwhile, may inform their unknown functions (or interesting connections of the cholesterol synthesis pathway to other biological processes).

To construct a somewhat global picture of cholesterol metabolism as well as its potential cross-talk with other biological processes, we next applied GeneFishing to the 17 GTEx tissues in which the coexpression pattern of the 21 genes was well maintained. In the previous sections, when generating candidate gene lists for experimental validation, we used a very strict CFR ≥ 0.99 threshold; here, we loosened the cutoff to 0.9, as the coexpression strength between bait genes and genes that are functionally linked to lipid metabolism are strongest in the liver as compared to other tissues. We discuss in *SI Appendix* that much lower cutoff points than 0.9 are still likely to correspond to very low FDR. In total, 329 genes were identified with a CFR larger than 0.9 in at least 1 tissue (*SI Appendix*, Table S4). Almost 74% (246 genes) of these were identified in only 1 tissue, while only 7.5% (28 genes) were identified in at least 8 tissues, illustrating that there is a high degree of tissue specificity. Tissue-specific GO enrichment analysis of the 329 genes identified 52 GO BP terms, each of which is significant in at least 1 tissue (FDR < 0.001). Interestingly, all of the 52 GO BP terms were child terms of the

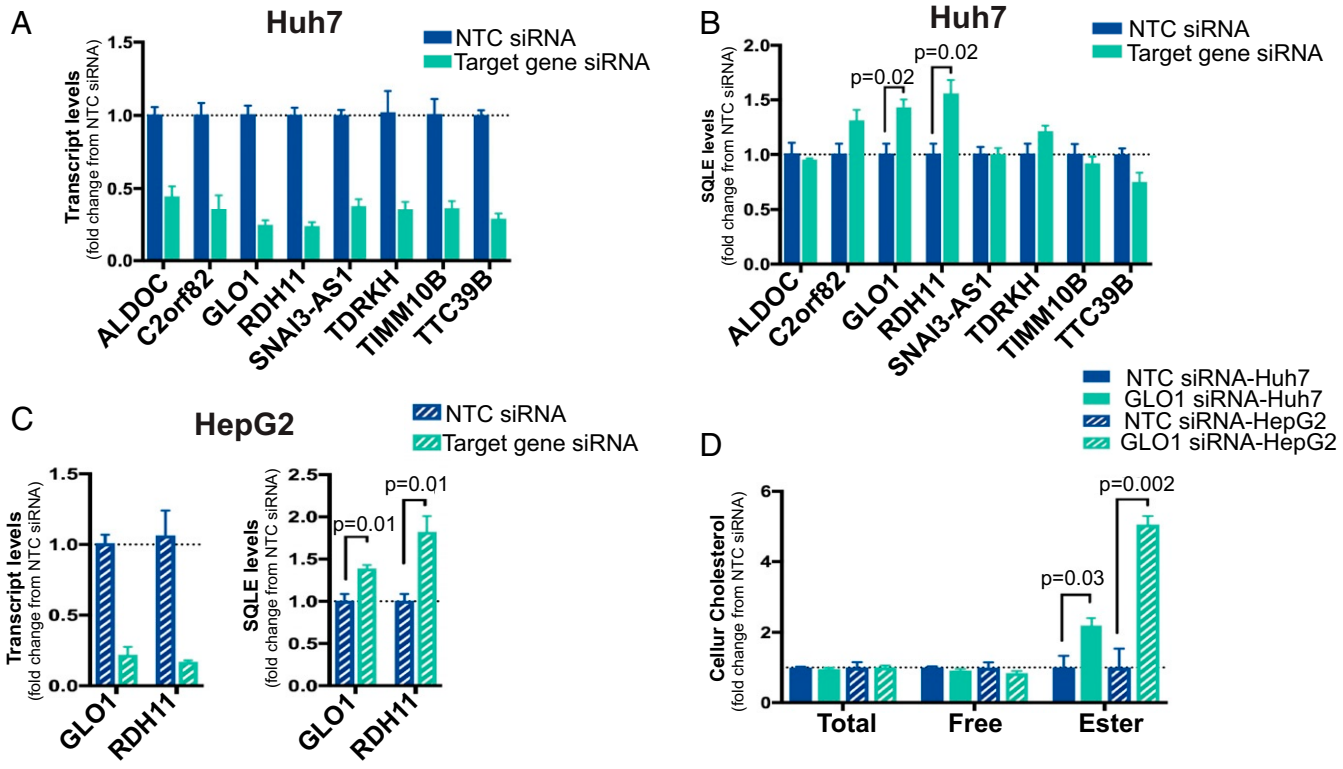

Fig. 3. Effect of candidate gene knockdown on transcript levels of cholesterol related genes. (*A*) Transcript levels (in the Huh7 cell line) of candidate genes were quantified by SYBR Green assay via qPCR to assess the degree of gene knockdown. (*B*) Transcript level of *SQLE* (in the Huh7 cell line) was quantified by SYBR Green assay to test whether candidate genes knockdown modulated its expression level. (*C*) Transcript levels (in the HepG2 cell line) of *GLO1* and *RDH11* were quantified by SYBR Green assay via qPCR to assess the degree of gene knockdown. Transcript level of *SQLE* (in the HepG2 cell line) was quantified by SYBR Green assay to test whether *GLO1* and *RDH11* knockdown modulated its expression level. In *A* to *C*, data were analyzed using the delta Ct (cycle threshold) method and normalized to *CLPTM1* transcript levels as a loading control. All qPCR assays were performed in triplicate. (*D*) Cellular cholesterol levels were quantified using the Amplex Red Cholesterol Assay kit with values normalized to total cellular protein quantified via Bradford assay. There are 3 to 6 replicates per treatment condition. NTC, nontargeting control.

"GO:0008152 metabolic process" (*SI Appendix*, Table S5). As expected, "GO:0006629 lipid metabolic process" was enriched in the genes identified in all of the 17 tissues. We also performed hierarchical clustering based on the GO enrichment profile and found that 6 tissues (artery–aorta, artery–tibial, whole blood, thyroid, pancreas, and stomach) seemed to be distinct from the remaining 11 tissues due to a depletion of the GO terms that were broadly enriched in other tissues (*SI Appendix*, Fig. S4). For example, while "GO:0006641 triglyceride metabolic process" was identified in 10 of the other 11 tissues, it was not enriched in any of the 6 tissues mentioned above.

Comparing GeneFishing with GIANT and ENDEAVOUR. Two popular methods, GIANT and ENDEAVOUR, were proposed before our study, and both of them have been widely used for gene prioritization. Although differing in key aspects from GeneFishing, the 3 methods share identical input–output schema: they all accept a group of seed (or bait) genes that are related to a biological process as input and return a list of genes that have been ranked according to computed functional relevance. We ranked all GTEx liver-expressed genes with GIANT and ENDEAVOUR. Since liver is the tissue that plays an important role in lipid metabolism and the 21 bait genes are all related to cholesterol metabolism, it is reasonable to expect that, in the returned gene list from any of the 3 gene prioritization methods, lipid metabolism-related genes should have high rankings. We found that GeneFishing captured the highest number of genes associated with the GO BP term "lipid metabolic process" among its top-ranked genes, demonstrating its superiority to the other 2 methods, at least in this application (Fig. 5). When compared with ENDEAVOUR, GeneFishing did substantially better in the identification of lipid-related genes. Although a similarly high number of lipid-related genes is found among the first 25 genes as ranked by our method and GIANT separately, our method outperforms GIANT substantially from then on. Interestingly, we found that gene *PCSK9*, a promising drug target to lower the LDLC level (which is also an *SREBF2* target gene) (23), was fished out (with CFR = 1) by GeneFishing, while its priority rank in the ranked list of candidate genes by GIANT was low (rank 6,102). In addition, the distribution of functional relevance measure returned by GIANT did not show as strong of bimodality as GeneFishing, suggesting that the calibration of the GIANT scores seems quite inferior to ours (*SI Appendix*, Fig. S5). We note that GIANT and ENDEAVOUR attempt to incorporate multiple sources of data (such as gene expression, protein–protein interaction, DNA sequence) to perform gene prioritization. They thus have large advantages in terms of broad applicability. However, as we demonstrate here, the generality of the information that they use may lead them to miss patterns specifically related to the biological question of interest. This is consistent with the phenomenon that we observed in Fig. 1 (in which inclusions of too much input data or noisy candidate genes obscure signal) and that we believe accounts for the mediocre performance of "all-purpose systems" in this task.

Discussion

In this paper, we developed a method we call GeneFishing. Our goal is to reveal potential relations between genes and gene

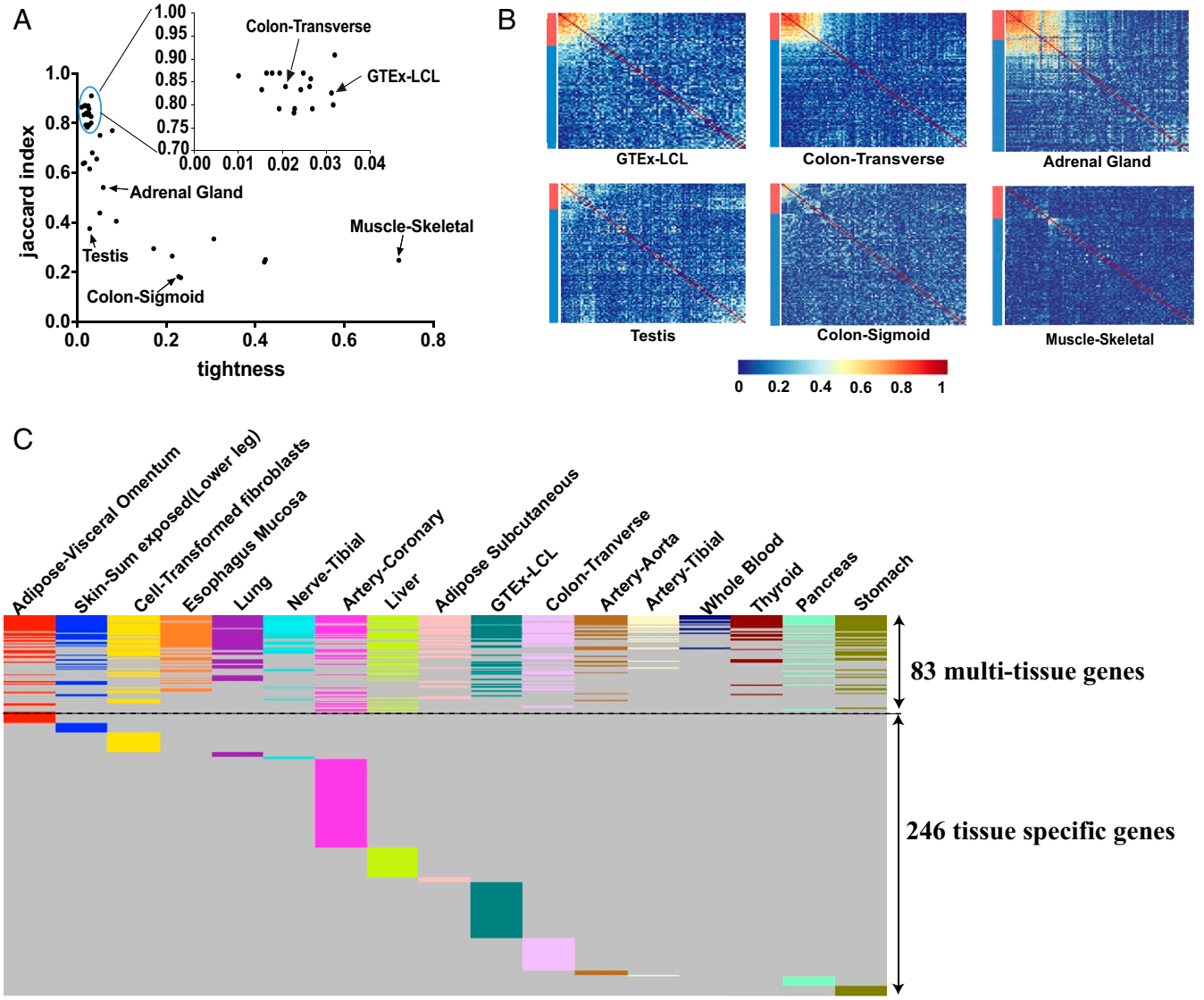

Fig. 4. Pantissue GeneFishing analysis. (*A*) Examination of modularity of the 21 bait genes across GTEx tissues. GeneFishing was applied to the 17 tissues inside the blue circle. The *Inset* shows the detailed coordinates of the 17 tissues. (*B*) The coexpression pattern of the genes associated with the GO BP term cholesterol metabolic process in 6 representative tissues. In each heat map, the row and column have identical gene orders, and the side bar indicates whether the gene belongs to the 21 bait genes (red means yes). (*C*) Visualization of pantissue GeneFishing results. Each row is associated with a gene, and each column is associated with a tissue (labeled with different colors). If the color of an entry is not gray, then it means that the CFR of the corresponding gene is higher than 0.9 in the corresponding tissue.

pathways. We applied this method to cholesterol-related genes and identified several interesting phenomena.

Applying GeneFishing to the GTEx liver dataset, we identified *GLO1* as a gene not previously implicated in cholesterol metabolism. Notably, murine models of *GLO1* knockdown and overexpression have reported conflicting results in regard to whether *GLO1* alters cholesterol metabolism in vivo. A *GLO1* transgenic ApoE$^{-/-}$ model was reported to have increased plasma cholesterol (24), while no change in lipids was observed in a second transgenic model in which *GLO1* was knocked down (25). Thus, our findings of increased cholesterol ester on *GLO1* knockdown in human liver-derived cell lines demonstrated that additional study is warranted to evaluate the role of *GLO1* in cholesterol metabolism.

We made an interesting observation when applying GeneFishing to the GTEx tissue dataset. Unlike most tissues, there is a striking lack of coexpression of the 21 cholesterol-related bait genes in the skeletal muscle dataset. In fact, they were not coexpressed with any of the 120 cholesterol metabolism genes annotated from BioMart (Fig. 4*B*). Since the 21 bait genes are well-known targets of *SREBF2*, a cholesterol-regulated transcription factor, this lack of coexpression suggests that, unlike other tissues, *SREBF2* may not be the major driver of expression of these 21 genes in the skeletal muscle. Statins, a class of cholesterol-lowering drug, function in part through the activation of *SREBF2*. Thus, the potential lesser importance of *SREBF2* in the regulation of the 21 bait genes may be relevant to the molecular mechanism underlying statin-induced myopathy, one of the most common adverse effects of statin treatment. Since statins also inhibit the production of isoprenoids and ubiquinone (i.e., coenzyme Q), our findings support further mechanistic studies that look beyond the role of *SREBF2*-mediated effects during the development of statin-induced myopathy.

The success of GeneFishing in the study of the cholesterol metabolic process illustrates how our method may reveal

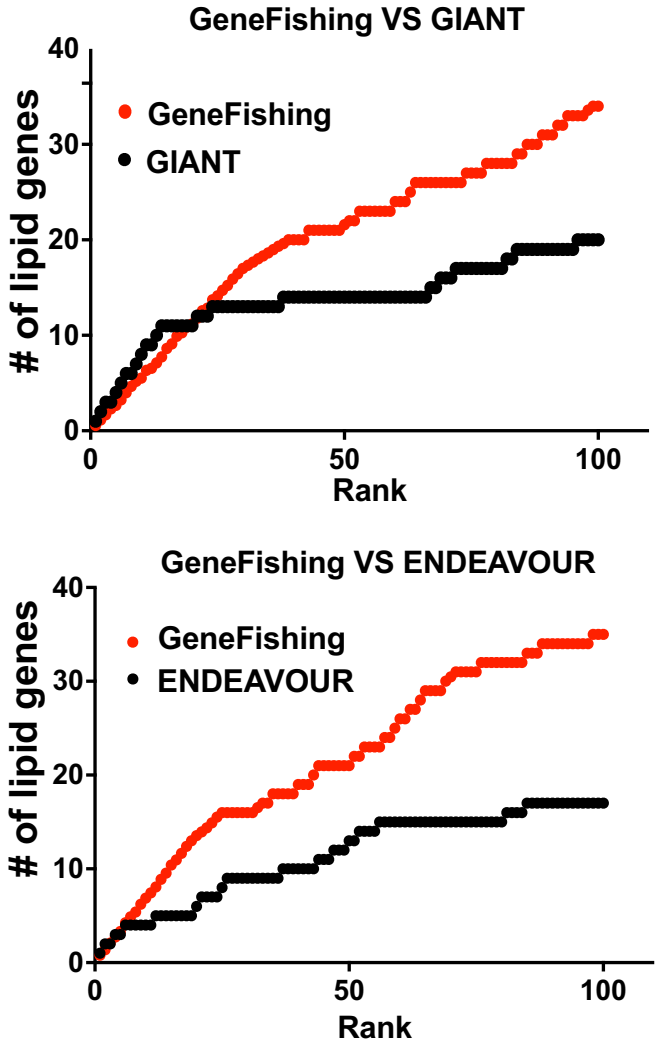

Fig. 5. In both panels, each colored curve corresponds to a method, with *x* axis representing the rank and the *y* axis representing the number of lipid metabolism genes among the top-ranked genes.

biology worthy of additional investigation. Importantly, compared with other methods, GeneFishing has the following advantages. (*i*) It is robust against noise in the bait genes. Through our evaluation on real and simulated data, as long as the majority of the bait genes are functionally active in relation to the biological process of interest, our procedure is reasonably effective in finding other relevant active genes. (*ii*) It provides reliable, interpretable measures of importance. (*iii*) It is computationally cheap and easy to parallelize and therefore, can easily handle genome-wide analyses. (*iv*) It is flexible, requiring only an appropriate set of bait genes and expression (or other measurements) on all genes for a set of subjects (or conditions). (*v*) It is simple and can easily incorporate other information, such as genetic variants or measures from other assays, at the dimensionality reduction or clustering stage to improve the relevance assessment of candidate genes to bait genes.

Our method is easily extendable to an iterative version (that we could call "iterative GeneFishing"), in which the discoveries may be added back to the bait set and the algorithm continued until some stopping criterion is satisfied. This might provide a tool for continued exploration of possible relations between processes in the same or different tissues.

In summary, GeneFishing is a powerful tool for reconstructing comprehensive context-specific portraits of biological processes and should be usable across a wide range of applications inside and outside of biology.

Methods

Data Collection and Processing. Multiple RNAseq datasets were used in this study. Preprocessed RNAseq data (version v6p) were downloaded from the GTEx data portal (https://gtexportal.org/home/) (4). RNAseq data of the LCL cell lines from GEUVADIS were downloaded from ArrayExpress (https://www.ebi.ac.uk/arrayexpress/experiments/E-GEUV-1/) (20). In addition to the above 2 datasets, we obtained the RNAseq data (of LCL cell lines) generated by the CAP project (13) from dbGap (accession no. phs000481.v2.p1). For each dataset, we removed genes with median RPKM (reads per kilobase per million mapped reads) or FPKM (fragments per kilobase per million mapped reads) value that is less than 0.1 and then, applied the PEER (26) software (number of hidden factors is 25) on log-transformed RPKM (or FPKM) value (with a pseudocount 0.01 added) to normalize the data.

Spectral Analysis for Dimension Reduction in GeneFishing. Given an $M \times T$ matrix, which contains the expression values of T genes across M samples, we first compute a $T \times T$ similarity matrix A with entry A[i,j] representing the absolute value of Spearman rank correlation between gene i and gene j across the M samples (note that alternative gene coexpression or association measures, such as those introduced and discussed in refs. 15–17, can be used per users' choice and study goal). Next, we performed an eigen decomposition of the normalized graph Laplacian $L = I - D^{-1/2}AD^{-1/2}$ and formed a $T \times K$ matrix G with column G[, j] representing the eigenvector corresponding to the jth smallest non-0 eigenvalue of L. Here, D is a diagonal matrix in which entry D[i,i] is the sum of the ith row of A. The matrix G thus provides a representation of the T genes in a space with reduced dimension (i.e., K dimensions). For the results presented in this article, we used K = 2. The spectral analysis method used here is based on the method proposed by Ng et al. in 2001 (27). The 2 clusters were determined by a K-means algorithm (*SI Appendix* has more details).

The "*P* value" for Individual Genes. The random partitioning of the space of genes that we perform gives us an initial solid basis for assigning *P* values (e.g., using an approximated binomial distribution) or some other measure of importance to fished out genes that can be used for prioritization. With the partitioning procedure being carried out multiple times, a natural measure of the relevance of a gene is the CFR. In some situations, the number of genes prioritized is not sensitive to the choice of CFR threshold. For example, in the case of CAP-LCL (Fig. 1*A*), choosing arbitrarily high CFR thresholds would result in a similar number of fished out genes. This is the case for most but not all tissues. An example, artery–coronary tissue, is discussed in *SI Appendix* (*SI Appendix*, Fig. S6). We also discuss various possible ways of calculating *P* values and FDR values in such cases in *SI Appendix*. In the analysis of liver tissue data, we found that a cutoff at CFR = 0.99 was a safe choice to select the candidates to follow-up. In the pantissue analysis, a cutoff at CFR = 0.9 seemed adequate.

ACKNOWLEDGMENTS. We thank Prof. Bora E. Baysal for allowing us to include the cholesterol synthesis pathway figure in our manuscript. We also thank Dr. James Ben Brown for providing us the computational clusters. We thank Zhiyue Tom Hu for implementing GeneFishing in Python (code available at https://github.com/tomwhoooo/GeneFishingPy). This work was supported by NIH Grants U01 HG007031, HL139902, and GM115318.

1. T. Ideker, T. Galitski, L. Hood, A new approach to decoding life: Systems biology. *Annu. Rev. Genom. Hum. Genet.* **2**, 343–372 (2001).
2. H. Kitano, Systems biology: A brief overview. *Science* **295**, 1662–1664 (2002).
3. I. Dunham *et al.*, An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
4. J. Lonsdale *et al.*, The genotype-tissue expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).
5. W. Y. S. Wang, B. J. Barratt, D. G. Clayton, J. A. Todd, Genome-wide association studies: Theoretical and practical concerns. *Nat. Rev. Genet.* **6**, 109–118 (2005).
6. Y. Moreau, L. C. Tranchevent, Computational tools for prioritizing candidate genes: Boosting disease gene discovery. *Nat. Rev. Genet.* **13**, 523–536 (2012).
7. C. S. Greene *et al.*, Understanding multicellular function and disease with human tissue-specific networks. *Nat. Genet.* **47**, 569–576 (2015).
8. L. C. Tranchevent *et al.*, Candidate gene prioritization with Endeavour. *Nucleic Acids Res.* **44**, W117–W121 (2016).

9. A. Krishnan *et al.*, Genome-wide prediction and functional characterization of the genetic basis of autism spectrum disorder. *Nat. Neurosci.* **19**, 1454–1462 (2016).
10. D. Guala, E. L. L. Sonnhammer, A large-scale benchmark of gene prioritization methods. *Sci. Rep.* **7**, 46598 (2017).
11. S. Aerts *et al.*, Gene prioritization through genomic data fusion. *Nat. Biotechnol.* **24**, 537–544 (2006).
12. O. Tzfadia, D. Amar, L. M. T. Bradbury, E. T. Wurtzel, R. Shamir, The MORPH algorithm: Ranking candidate genes for membership in arabidopsis and tomato pathways. *Plant Cell* **24**, 4389–4406 (2012).
13. J. A. Simon *et al.*, Phenotypic predictors of response to Simvastatin therapy among African-Americans and Caucasians: The cholesterol and pharmacogenetics (CAP) study. *Am. J. Cardiol.* **97**, 843–850 (2006).
14. C. B. Wilcox *et al.*, Coordinate up-regulation of tmem97 and cholesterol biosynthesis genes in normal ovarian surface epithelial cells treated with progesterone: Implications for pathogenesis of ovarian cancer. *BMC Canc.* **7**, 223 (2007).
15. Y. R. Wang *et al.*, Inferring gene–gene interactions and functional modules using sparse canonical correlation analysis. *Ann. Appl. Stat.* **9**, 300–323 (2015).
16. Y. R. Wang, M. S. Waterman, H. Huang, Gene coexpression measures in large heterogeneous samples using count statistics. *Proc. Natl. Acad. Sci. U.S.A* **111**, 16371–16376 (2014).
17. Y. R. Wang *et al.*, Generalized correlation measure using count statistics for gene expression data with ordered samples. *Bioinformatics* **34**, 617–624 (2017).
18. B. Yu, Stability. *Bernoulli* **19**, 1484–1500 (2013).
19. M. W. Medina *et al.*, Transmembrane protein 55B is a novel regulator of cellular cholesterol metabolism. *Arterioscler. Thromb. Vasc. Biol.* **34**, 1917–1923 (2014).
20. T. Lappalainen *et al.*, Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506–511 (2013).
21. P. Langfelder, S. Horvath, WGCNA: An R package for weighted correlation network analysis. *BMC Bioinf.* **9**, 559 (2008).
22. S. Falcon, R. Gentleman, Using gostats to test gene lists for go term association. *Bioinformatics* **23**, 257–258 (2006).
23. R. T. Dadu, C. M. Ballantyne, Lipid lowering with pcsk9 inhibitors. *Nat. Rev. Cardiol.* **11**, 563–575 (2014).
24. M. Geoffrion *et al.*, Differential effects of glyoxalase 1 overexpression on diabetic atherosclerosis and renal dysfunction in streptozotocin-treated, apolipoprotein E-deficient mice. *Physiol. Rep.* **2**, 1–17 (2014).
25. M. Wortmann *et al.*, A Glyoxalase-1 knockdown does not have major short term effects on energy expenditure and atherosclerosis in mice. *J. Diabetes Res.* **2016**, 1–8 (2016).
26. O. Stegle, L. Parts, M. Piipari, J. Winn, R. Durbin, Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat. Protoc.* **7**, 500–507 (2012).
27. A. Y. Ng, M. I. Jordan, Y. Weiss, "On spectral clustering: Analysis and an algorithm" *in Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic, NIPS'01*, T. G. Dietterich, S. Becker, Z. Ghahramani, Eds. (MIT Press, Cambridge, MA, 2001), pp. 849–856.