METHODS

Gene communities in co-expression networks across different tissues

Madison Russell[1], **Alber Aqil**[2], **Marie Saitou**[3], **Omer Gokcumen**[2], **Naoki Masuda**[1,4]*

1 Department of Mathematics, State University of New York at Buffalo, Buffalo, New York, United States of America, **2** Department of Biological Sciences, State University of New York at Buffalo, Buffalo, New York, United States of America, **3** Faculty of Biosciences, Norwegian University of Life Sciences, Ås, Norway, **4** Institute for Artificial Intelligence and Data Science, State University of New York at Buffalo, Buffalo, New York, United States of America

* naokimas@gmail.com

Citation: Russell M, Aqil A, Saitou M, Gokcumen O, Masuda N (2023) Gene communities in co-expression networks across different tissues. PLoS Comput Biol 19(11): e1011616. https://doi.org/10.1371/journal.pcbi.1011616

Editor: Feng Fu, Dartmouth College, UNITED STATES

Received: May 19, 2023

Accepted: October 19, 2023

Published: November 17, 2023

Data Availability Statement: All data that are used in the study can be found publicly. The references and databases are provided in the manuscript. Direct link to GTEx TPM data: https://storage.cloud.google.com/adult-gtex/bulk-gex/v8/rna-seq/GTEx_Analysis_2017-06-05_v8_RNASeQCv1.1.9_gene_tpm.gct.gz. Direct link to GTEx eQTL data: https://storage.cloud.google.com/adult-gtex/bulk-qtl/v8/single-tissue-cis-qtl/GTEx_Analysis_v8_eQTL.tar. Code is available in a GitHub repository: https://github.com/russell-madison/corr_comm_detection.

Abstract

With the recent availability of tissue-specific gene expression data, e.g., provided by the GTEx Consortium, there is interest in comparing gene co-expression patterns across tissues. One promising approach to this problem is to use a multilayer network analysis framework and perform multilayer community detection. Communities in gene co-expression networks reveal groups of genes similarly expressed across individuals, potentially involved in related biological processes responding to specific environmental stimuli or sharing common regulatory variations. We construct a multilayer network in which each of the four layers is an exocrine gland tissue-specific gene co-expression network. We develop methods for multilayer community detection with correlation matrix input and an appropriate null model. Our correlation matrix input method identifies five groups of genes that are similarly co-expressed in multiple tissues (a community that spans multiple layers, which we call a generalist community) and two groups of genes that are co-expressed in just one tissue (a community that lies primarily within just one layer, which we call a specialist community). We further found gene co-expression communities where the genes physically cluster across the genome significantly more than expected by chance (on chromosomes 1 and 11). This clustering hints at underlying regulatory elements determining similar expression patterns across individuals and cell types. We suggest that *KRTAP3-1*, *KRTAP3-3*, and *KRTAP3-5* share regulatory elements in skin and pancreas. Furthermore, we find that *CELA3A* and *CELA3B* share associated expression quantitative trait loci in the pancreas. The results indicate that our multilayer community detection method for correlation matrix input extracts biologically interesting communities of genes.

Author summary

Genes that are similarly expressed across individuals (i.e., co-expressed) are potentially involved in related biological processes. Therefore, the identification and biological analysis of co-expressed genes may be useful for revealing genes associated with specific diseases or other phenotypes. Because gene co-expression depends on the tissue in general,

Funding: O.G. and N.M. acknowledge support from National Institute of General Medical Sciences (under grant no.1R01GM148973-01). N.M. also acknowledges support from the Japan Science and Technology Agency (JST) Moonshot R&D (under grant no.JPMJMS2021), the National Science Foundation (under grant no.2052720), and JSPS KAKENHI (under grant nos.JP 21H04595 and 23H03414). O.G. acknowledges support from the National Science Foundation (under grant nos.2049947 and 2123284). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

we compared co-expression patterns across four different exocrine gland tissues. This problem lends itself to multilayer network analysis in which each layer of the multilayer network is a tissue-specific gene co-expression network. The nodes in the network represent genes, and a pair of genes is directly connected by an edge if the two genes are co-expressed. We developed a method to detect groups of co-expressed genes in the multilayer gene co-expression network using correlational tissue-specific gene expression data. We found some groups of genes that are co-expressed in all four tissues and other groups of genes that are only co-expressed in one tissue. We also found that some of these groups of genes contain genes that are physically clustered across the genome. Our methods reveal groups of genes with potentially different mechanisms of gene co-expression.

1 Introduction

In networks, communities, or modules, are broadly defined as groups of nodes with higher internal than external density of edges compared to a null model [1, 2]. There have been proposed numerous objective functions to be optimized and algorithms for community detection in networks. Because edges in networks represent a relationship between the nodes, it follows that these communities are groups of nodes that likely share common properties or play a similar role within the network. Many real-world networks naturally divide into communities, including biological networks, and studying communities is expected to help us better understand complex biological interactions [3–8].

Communities in gene networks are often called gene modules [4–6]. Methods to find functional gene modules are useful tools for discovering how the genes interact and coordinate to perform specific biological functions [9–12]. Furthermore, studying the relationships between gene modules may reveal a higher-order organization of the transcriptome [13, 14]. Biological analyses of gene modules can suggest genes that play a regulatory role in disease or other phenotypes, or identify novel therapeutic target genes for future intervention studies [15–18]. Additionally, one can study gene modules across evolutionary time to find biologically important groups of co-regulated genes because genes that must be co-expressed together will be under evolutionary pressure to maintain their coordinated expression [19, 20].

While there are various definitions of gene modules, or communities, in gene co-expression networks, gene modules are sets of genes that are similarly expressed across individuals and, therefore, potentially involved in related biological processes [16, 19, 21]. In such networks, the nodes represent genes, and a pair of nodes is directly connected with each other by an undirected edge if the two genes are co-expressed, i.e., if they show a similar expression pattern across samples [9, 15, 21–23]. Biologically, co-expressed genes may occur because transcription factors may have unique DNA binding sites located in promoter regions of distinct sets of genes [24, 25], polymerase binding may cause synchronous transcription of several genes [26], physically closeby genes may cluster within similarly regulated topologically associated domains [27–29], or particular environmental factors may concurrently affect genes in a particular pathway [30–32], among other reasons [33]. Non-biological effects such as batch processing and RNA quality also contribute to gene co-expression [34, 35]. In general, one cannot distinguish between the biological and non-biological sources of co-expression from the expression data alone; thus, interpreting co-expression networks is challenging [33, 36]. However, gene co-expression network analysis may be able to clarify novel molecular mechanisms that are relevant to disease and facilitate identification of potential targets for intervention studies [16, 33]. Crucially, gene co-expression and gene expression carry different information.

For example, differential co-expression analysis identified the alpha synuclein variant (aSynL) in several Parkinson's disease data sets. In contrast, differential expression analysis alone did not identify this variant since aSynL was highly differentially co-expressed but not highly differentially expressed [37]. Gene co-expression analyses can provide novel insights that are likely overlooked or undetected in traditional gene expression analyses [33].

Gene expression and co-expression may depend on regulatory elements in the genome, which are often specific to different cell types [17, 38–41]. The increased availability of tissue-specific gene expression data allows us to compare and contrast gene expression and co-expression and their communities across different tissues. A challenge for deciphering such data is integrating and distinguishing between communities found in various cell types, determining their biological relevance, and identifying regulatory elements maintaining these communities. For example, a simultaneous analysis of both generic multi-tissue co-expression (derived from aggregated gene expression data from multiple tissues) and tissue-specific co-expression resulted in a more efficient prediction of human disease genes than the use of generic multi-tissue co-expression alone [38]. It has also been found that modules conserved across different types of tissues are likely to have functions common to those tissues [39, 42]. In contrast, modules upregulated in a particular tissue are often involved in tissue-specific functions [39].

One can regard a set of co-expression networks of genes constructed for multiple tissues as a multilayer network. As we schematically show in Fig 1, each layer of the multilayer network is a tissue-specific gene co-expression network. The edges within a layer (i.e., intralayer edges) represent tissue-specific co-expression. The edges between the layers (i.e., interlayer edges) connect the same gene across tissues. Multilayer network analysis, particularly multilayer community detection [43, 44], is becoming an increasingly popular tool in biological data analysis given that biological systems are often multi-dimensional and involve complex interactions

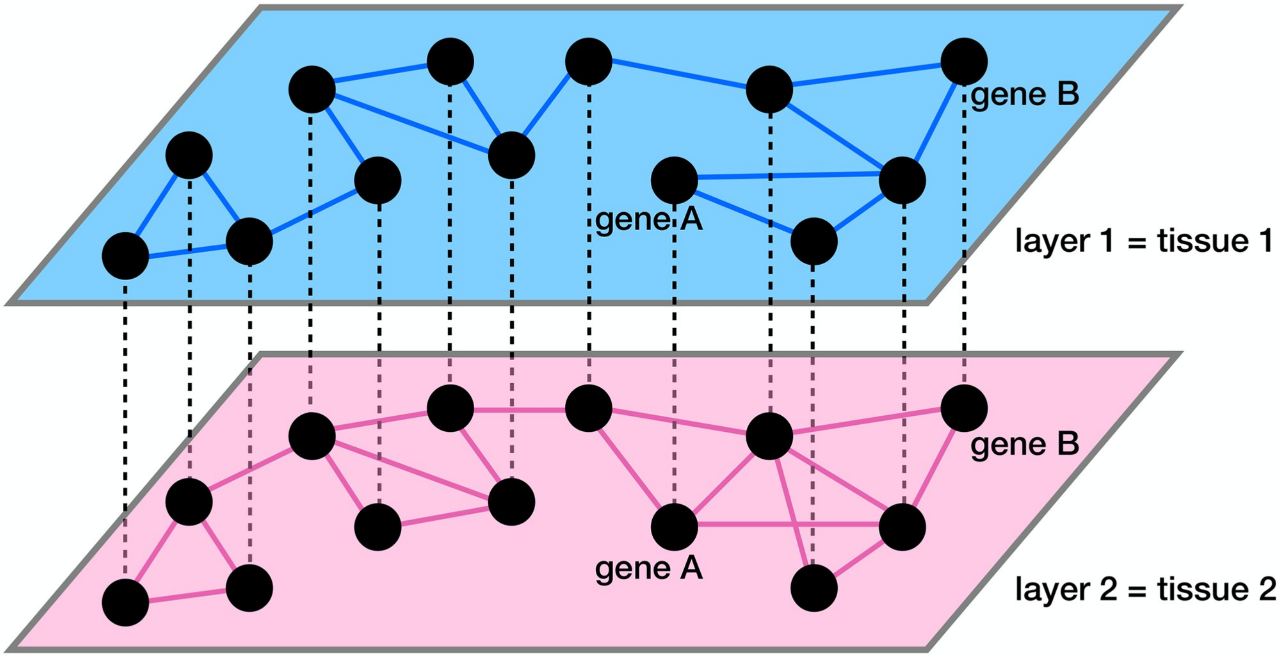

Fig 1. Schematic of a multilayer gene co-expression network. The intralayer edges, shown by the solid lines, represent co-expression. The interlayer edges, shown by the dashed lines, connect the same gene across layers.

https://doi.org/10.1371/journal.pcbi.1011616.g001

[45–48]. Analyzing single-layer networks separately may be insufficient to reveal the patterns of these complex biological interactions [47]. For example, multilayer gene co-expression networks, in which each layer consists of a subset of gene pairs with a similar co-expression level, were constructed for comparing healthy and breast cancer co-expression patterns [49]. In the healthy multilayer co-expression network, the layers gradually attain hub nodes as one goes towards the top layer, whereas in the breast cancer multilayer network, the majority of layers contain no hub nodes and only a few top layers abruptly start to contain hub nodes [49]. In another application to breast cancer data, a multilayer gene co-expression network in which each layer corresponds to a clinical stage of breast cancer was analyzed [50]. A community detection algorithm designed to identify layer-specific modules in multilayer networks finds gene modules in the breast cancer network significantly associated with the survival time of patients [50]. Community detection in multilayer stochastic block models, in which each layer is a gene co-expression network at a specific developmental time, reveals different biological processes active at different stages of a monkey's brain development [51, 52]. A Higher-Order Generalized Singular Value Decomposition method allows for simultaneous identification of both "common" and "differential" modules across several tissue-specific gene co-expression networks [53]. A study of the relationships between communities across different tissue-specific layers of a multilayer gene co-expression network provides promise for our better understanding of inter-tissue regulatory mechanisms through both intra-tissue and inter-tissue transcriptome analysis [41].

Another application for multilayer approaches is to categorize diseases and drug targets. For instance, analyses of densely connected subgraphs that consistently appear in different layers have revealed disease modules (i.e., groups of diseases extracted from a four-layer disease similarity network in which a node is a disease and the four layers are constructed from protein-protein interaction (PPI), a symptom data set, Gene Ontology, and Disease Ontology) [54] and drug-target modules (i.e., groups of genes extracted from a multilayer network in which each layer is a tissue-specific PPI network) [55]. Groups of diseases that have molecular and phenotypic similarities were discovered in an analysis of a bilayer network of human diseases consisting of a genotype-based and phenotype-based layers [56]. A multilayer network analysis in which each layer is a similarity matrix among 26 different populations for a given structural variant revealed evolutionarily adaptive structural variants [57]. Regulatory and signaling mechanisms associated with a given cellular response were discovered using a multilayer community detection method designed for identifying active modules in weighted gene co-expression networks [58]. Community detection on tissue-specific multilayer networks composed of a co-expression network, transcription factor co-targeting network, microRNA co-targeting network, and PPI network revealed candidate driver cancer genes [59].

As discussed above, the study of co-expression networks can lead to various biological insights [22, 33, 60]. However, there are some limitations to this approach. Edges of co-expression networks are correlational in nature. In general, creating unweighted or weighted networks from correlation data can be straightforward (e.g., thresholding on the edge weight and/or assuming no edges between negatively correlated node pairs). However, such straightforward methods are subject to various problems such as false positives [61, 62], arbitrariness in setting the parameter value such as the threshold on the edge weight [63, 64], and loss of information by subthreshold or negative correlation values [63, 65]. Existing methods to estimate sparse networks from correlation matrix data, such as graphical lasso [66–68] or estimation of sparse covariance matrices [69–71], mitigate some of these problems. In contrast to constructing sparse networks, in the present study, we explore the adaptation of network analysis methods to directly work on correlation matrix input. Such methods have been developed for community detection via modularity maximization [72–74] and clustering coefficients [75]. A key observation exploited in these studies is that one needs to use appropriate null models for correlation matrices, which are different from those for general networks. In particular, the standard null model for general networks called the configuration model is not a correlation or covariance matrix in general [72]. In this study, we expand this line of approach to the case of multilayer correlation matrix data. In particular, we develop a method for community detection by combining multilayer modularity maximization and a configuration model of correlation matrices. We also develop statistical methods to calculate the significance of each detected community. We apply our methods to multilayer Pearson correlation matrices representing co-expression of genes in four tissues to compare communities of genes across different tissues. Code for running our multilayer community detection method with covariance matrix input is available at Github [76].

2 Methods

2.1 Data

The Genotype-Tissue Expression (GTEx) portal provides open-access tissue-specific gene expression data [77]. For the analyses in the present work, we use the gene transcripts per million (TPM) data from release V8 for four exocrine glands: pancreas, minor salivary gland, mammary gland, and skin (not sun exposed). In this pilot study, we limit our analysis to four tissues. We chose these tissues because they are all tissues that interact with the outside world and may have adaptively evolved to different environmental conditions. Specifically, the pancreas plays a vital role in the digestive system, secreting digestive enzymes [78]. The salivary gland is the main gatekeeper of our body and contributes to the oral proteome [79]. The mammary gland produces milk containing immunologic agents to nourish and protect young offspring [80]. The skin protects the body against pathogens, regulates body temperature, and has changed most drastically in human lineage [81, 82]. Consequently, we hypothesized that these tissues would retain a high level of variation in gene expression levels.

There are 328 samples from the pancreas, 162 samples from the minor salivary gland, 459 samples from the mammary gland, and 604 samples from the skin (not sun exposed) in this TPM data. Each sample contains gene expression data for 56, 200 different genes.

The number of genes is much larger than the number of samples for all tissues. Therefore, we focused on a subset of genes for our analysis around the same size as the number of samples in our data, as in [39, 83]. To subset the genes, we identified the top 75 genes with the highest variance of TPM across all samples [22], separately for each tissue. We chose the number 75 because the union of the top 75 genes in terms of the variance of TPM across the four tissues contains 203 genes, which is not much larger than the smallest number of samples (162 samples). It is well known that estimation of correlation matrices from data is unreliable if the number of elements (i.e., genes in the present case) is comparable with or larger than the number of samples [84]. Nevertheless, to further validate our choice for the number of genes, we repeated some analysis on an expanded network with 371 genes. We found that the expanded network produces a similar type of partition as the original network, supporting the robustness of our analysis with respect to the number of genes selected for our analysis (see Text A in S1 Text).

We looked at the most variable genes because, again, our goal is to understand the underlying genetic and environmental bases of gene expression variation. In fact, most of the highly variably expressed genes are also highly expressed genes. To show this, for each tissue, we calculate the Jaccard index between the top 75 genes in terms of average TPM and the top 75 genes in terms of variance of TPM. The Jaccard index is defined as the size of the intersection of two finite sets A and B divided by the size of the union of A and B [85]. The range of the

Table 1. Similarity between the highly variable genes and the highly expressed genes in each tissue.

Tissue	Jaccard index	Average rank
pancreas	0.685	52.81
salivary gland	0.531	64.31
mammary gland	0.402	133.5
skin	0.442	167.8

We calculate the Jaccard index between the top 75 genes in terms of average TPM and the top 75 genes in terms of variance of TPM. We calculate the average rank of the top 75 genes in variance, where the rank is in terms of average TPM.

https://doi.org/10.1371/journal.pcbi.1011616.t001

Jaccard index is 0 to 1, and a larger Jaccard index implies a greater overlap between the two sets of genes. We also examine the average rank of the top 75 genes in variance among all 56, 200 genes. We compute the rank in terms of the average TPM. Therefore, if the average rank is high (i.e., a low number), then the highly variable genes are also relatively highly expressed. We show in Table 1 the Jaccard index and the average rank of the top 75 genes for each tissue. The table indicates that the Jaccard index is at least 0.402 and the average rank is at most 167.8. These results suggest that the top 75 genes in terms of variance of TPM are overall highly expressed genes as well because we have calculated these indices for 75 genes in comparison to the 56, 200 genes. This finding is consistent with an established understanding that sequence read count data follows a negative binomial distribution [86–88].

We analyze four-layer networks composed of the 203 genes in the union of the top 75 genes in terms of the variance of TPM across the four tissues. We note that the number of nodes must be the same in each layer for our multilayer community detection method described in section 2.4.

2.2 Multilayer network construction

For each of the four tissues, we generate a 203 × 203 gene co-expression matrix in which the (i, j)-th entry is the Pearson correlation coefficient between the log-normalized TPM of gene i and the log-normalized TPM of gene j across all samples from that tissue. We take the logarithm of TPM before calculating the Pearson correlation coefficient to suppress the effect of outliers; TPM is extremely large for some samples. Let S denote the number of samples from tissue α. We denote by $x_{i,\alpha,s}$ and $x_{j,\alpha,\,s}$ the TPM value for gene i and j, respectively, for sample $s \in \{1, 2, \ldots, S\}$ in tissue α. Then, we calculated the Pearson correlation coefficient between log $(x_{i,\alpha,s} + 1)$ and $\log(x_{j,\alpha,\,s} + 1)$ across the S samples as the co-expression between gene i and gene j in tissue α. In other words, we calculate

$$r_{\alpha}(i,j) = \frac{\sum_{s=1}^{S}[\log(x_{i,\alpha,s}+1) - m_{i,\alpha}][\log(x_{j,\alpha,s}+1) - m_{j,\alpha}]}{\sqrt{\sum_{s=1}^{S}[\log(x_{i,\alpha,s}+1) - m_{i,\alpha}]^2 \sum_{s=1}^{S}[\log(x_{j,\alpha,s}+1) - m_{j,\alpha}]^2}}, \tag{1}$$

where

$$m_{i,\alpha} = \frac{1}{S}\sum_{s=1}^{S}\log(x_{i,\alpha,s}+1) \tag{2}$$

and

$$m_{j,\alpha} = \frac{1}{S}\sum_{s=1}^{S}\log(x_{j,\alpha,s}+1). \tag{3}$$

We took the logarithm of $x_{i,\alpha,s}$ + 1 because, in this manner, $x_{i,\alpha,s} = 0$ is mapped to 0.

To compare the gene co-expression patterns across the different tissues, we view the four correlation matrices as a four-layer correlation matrix, or categorical layers of a multilayer gene co-expression network. Because the set of genes is the same in the four layers, we place an interlayer edge between the same gene in each pair of layers (i.e., tissues) as shown by the dashed lines in Fig 1. Therefore, our network is a multiplex network with diagonal and categorical interlayer couplings, where, by definition, the interlayer edges connect each gene with itself in each other layer [89, 90].

We denote the strength of the interlayer coupling that connects node i in layer α to node i in layer β as $\omega_{i\alpha\beta}$ [43]. One typically assumes that $\omega_{i\alpha\beta}$ takes binary values $\{0, \omega\}$, where ω is a parameter indicating the absence (i.e., 0) or presence (i.e., ω) of interlayer edges [43]. However, how to set and interpret the ω value is not straightforward [91]. In this work, we use the empirical co-expression (i.e., Pearson correlation coefficient) of gene i between tissues α and β as $\omega_{i\alpha\beta}$. Specifically, $\omega_{i\alpha\beta}$ is equal to the right-hand side of Eq (1) with $x_{j,\alpha,s}$ and $m_{j,\alpha}$ being replaced by $x_{i,\beta,s}$ and $m_{i,\beta}$, respectively, and with S being interpreted as the number of samples common to tissues α and β. Since the majority of studies on multilayer modularity maximization assume non-negative interlayer edge weights, if the obtained $\omega_{i\alpha\beta}$ is negative, we force $\omega_{i\alpha\beta} = 0$. However, note that some studies do include negative interlayer edge weights [92].

2.3 Community detection in conventional multilayer networks

We are interested in detecting communities (also called modules and gene sets) in our multilayer networks to find sets of genes that are similarly expressed across individuals and therefore potentially involved in related biological processes. Some algorithms can detect communities that span between multiple layers as well as communities that lie within just one layer. We are interested in these different types of communities and their biological implications. A common method to find such communities in multilayer networks is to maximize an objective function called the multilayer modularity [43]. However, our multilayer gene networks are based on correlation. Therefore, we develop multilayer modularity for multilayer correlation matrices. In this section, we review multilayer modularity for usual multilayer networks as a primer to the multilayer modularity for correlation matrices.

The modularity for single-layer undirected networks, which may be weighted, is given by [93, 94]

$$Q = \frac{1}{2M}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(A_{ij} - \gamma\frac{k_i k_j}{2M}\right)\delta(g_i, g_j), \tag{4}$$

where N is the number of nodes in the given network; A_{ij} is the (i, j)-th entry of the adjacency matrix and we assume $A_{ii} = 0 \; \forall i \in \{1, \ldots, N\}$; $M = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} A_{ij}$ is the number of edges in the case of unweighted networks and the total weight of all edges in the case of weighted networks; γ is the resolution parameter controlling the size of typical communities found by modularity maximization [95]; a large γ tends to lead to relatively many small communities; $k_i k_j/2M$ is equal to the probability that an edge exists, or alternatively the expected edge weight, between nodes i and j under the configuration model; $k_i = \sum_{j=1}^{N} A_{ij}$ is the (weighted) degree of node i; g_i is the community to which node i belongs; $\delta(g_i, g_j) = 1$ if $g_i = g_j$ and $\delta(g_i, g_j) = 0$ otherwise.

To generalize the modularity to the case of multilayer networks, let $\mathcal{L}$ be the number of layers in the multilayer network. We let $A_{ij\alpha}$ be the (i, j)-th entry of the intralayer adjacency matrix, which may be weighted, in network layer α. We assume $A_{ii\alpha} = 0$ $\forall i \in \{1, \ldots, N\}$ and $\forall \alpha \in \{1, \ldots, \mathcal{L}\}$. We remind that $\omega_{i\alpha\beta}$ is the weight of the interlayer coupling between node i in layer α and node i itself in layer β. The multilayer modularity is given by [43]

$$Q = \frac{1}{2\mu}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{\alpha=1}^{\mathcal{L}}\sum_{\beta=1}^{\mathcal{L}}\left[\underbrace{\left(A_{ij\alpha} - \gamma_\alpha \frac{k_{i\alpha}k_{j\alpha}}{2m_\alpha}\right)\delta_{\alpha\beta}}_{\text{intralayer}} + \underbrace{\omega_{i\alpha\beta}\delta_{ij}}_{\text{interlayer}}\right]\delta(g_{i\alpha}, g_{j\beta}), \tag{5}$$

where $k_{i\alpha} = \sum_{j=1}^{N} A_{ij\alpha}$ is the strength (i.e., weighted degree) of node i in layer α, and $m_\alpha = \frac{1}{2}\sum_{i=1}^{N} k_{i\alpha}$ is the total edge weight in layer α. We set $2\mu = \sum_{i=1}^{N}\sum_{\alpha=1}^{\mathcal{L}}(k_{i\alpha} + \sum_{\beta=1}^{\mathcal{L}}\omega_{i\alpha\beta})$, which is equal to twice of the total edge weight. Let γ_α be the resolution parameter in layer α; $\delta_{\alpha\beta} = 1$ if $\alpha = \beta$ and $\delta_{\alpha\beta} = 0$ otherwise; δ_{ij} is defined in the same manner; and $g_{i\alpha}$ is the community to which node i in layer α belongs. Eq (5) implies that communities that contain interlayer edges are rewarded with higher modularity values.

We will discuss the selection of γ_α in section 2.5. We use the Louvain algorithm for multilayer modularity maximization. Specifically, we use the iterated GenLouvain function from GenLouvain version 2.2, which repeatedly implements GenLouvain until convergence to an output partition (i.e., until the output partition does not change between two successive iterations) [96, 97].

The modularity function Q typically has many local maxima [98]. Reflecting this fact, most modularity maximization algorithms are stochastic and do not output a unique answer. A common approach to combine the results from multiple partitions of nodes is consensus clustering to obtain a consensus partition [99]. We use the consensus clustering algorithm described in [100] and implemented in the Python package netneurotools version 0.2.3 [101].

2.4 Community detection in multilayer correlation matrices

In this section, we expand modularity maximization for correlation matrices [72, 73] to the case of multilayer correlation matrices.

Let $\rho = (\rho_{ij})$ be an $N \times N$ correlation matrix and $\langle\rho\rangle$ be a null model of the correlation matrix of the same size. The modularity for a single correlation matrix is given by

$$Q = \frac{1}{C_{\text{norm}}}\sum_{i=1}^{N}\sum_{j=1}^{N}(\rho_{ij} - \langle\rho_{ij}\rangle)\delta(g_i, g_j), \tag{6}$$

where $C_{\text{norm}} = \sum_{i=1}^{N}\sum_{j=1}^{N}\rho_{ij}$ is a normalization constant. One can use a modularity maximization algorithm to maximize Q given $\langle\rho\rangle$.

We generalize Eq (6) to the case of a multilayer correlation matrix by writing down an equation in the same form as Eq (5). We will use the term node to refer to a gene in a specific layer of the four-layer correlation matrix. Let $\rho_{ij\alpha}$ be the empirical Pearson correlation coefficient between nodes i and j in layer α, and let $\langle\rho_{ij\alpha}\rangle$ be the correlation between nodes i and j in layer α in the null model of the correlation matrix. Then, the modularity of a multilayer correlation matrix is

$$Q = \frac{1}{C_{\mathrm{norm}}}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{\alpha=1}^{\mathcal{L}}\sum_{\beta=1}^{\mathcal{L}}\left[(\rho_{ij\alpha} - \gamma_{\alpha}\langle\rho_{ij\alpha}\rangle)\delta_{\alpha\beta} + \omega_{i\alpha\beta}\delta_{ij}\right]\delta(g_{i\alpha}, g_{j\beta}), \tag{7}$$

where $C_{\mathrm{norm}} = \sum_{i=1}^{N}\sum_{\alpha=1}^{\mathcal{L}}(\sum_{j=1}^{N}\rho_{ij\alpha} + \sum_{\beta=1}^{\mathcal{L}}\omega_{i\alpha\beta})$. Parameter γ_α represents the resolution in layer α [95], and we will discuss the selection of γ_α in section 2.5. We remind that $\omega_{i\alpha\beta}$ is the empirical co-expression of gene i between tissues α and β. We double-count (i, j) and (j, i), with $i \neq j$, in Eq (7) following previous literature [72, 73].

We use a configuration model for correlation matrices [74] as the null model, while other null models are also possible, such as the H-Q-S algorithm [102] and those derived from random matrix theory [72]. The configuration model [74], implemented in the configcorr package [103], generates the correlation matrix maximizing the entropy under the constraint that the strength (i.e., weighted degree) of each node of the input correlation matrix is conserved. The model assumes normality of the input data. While the algorithm accepts a covariance matrix or a correlation matrix as input, if the input is a covariance matrix, it is first transformed to the correlation matrix before being fed to the configuration model. To maximize Q given by Eq (7), we feed the supra-modularity matrix B, where $B_{i\alpha j\beta} = (\rho_{ij\alpha} - \gamma_\alpha\langle\rho_{ij\alpha}\rangle)\delta_{\alpha\beta} + \omega_{i\alpha\beta}\delta_{ij}$, to GenLouvain. Again, we use the iterated GenLouvain function [97] and a consensus clustering technique to obtain a final partition [100] but by inputting 200 partitions of the same network.

Prior studies developed methods to assess statistical significance of the detected communities in single-layer networks [104–106]. Here, we extend this approach to the case of multilayer correlation matrices and multilayer networks. We do this by comparing a detected community to the same set of nodes in a random graph (or null model) in terms of some quality measure. For each detected community and given quality measure, we calculated the Z score defined by

$$z = \frac{x - \mu}{\sigma}, \tag{8}$$

where x is the quality measure calculated for the empirical community, and μ and σ are the expected value and the standard deviation, respectively, of the same quality measure for the same community but under a null model. In the following text, we explain this method for multilayer correlation matrices, which we primarily use for our gene data analysis. We show the details of our methods for general multilayer networks in Text B in S1 Text.

We introduce a quality measure of a community that is analogous to the total weight of the intralayer edges within the community. Let W be the total weight of intralayer edges within the set of nodes S in a multilayer correlation matrix. In the remainder of this section, we use the covariance matrices instead of correlation matrices for analytical tractability. This assumption is not detrimental to the application of our methods to multilayer correlation matrix data because a correlation matrix is a covariance matrix in general. Let $C^{\mathrm{org}}_{ij\alpha}$ be the (i, j)-th element of $C^{\mathrm{org}}_{\alpha}$, an empirical covariance matrix for layer α. Then, we have

$$W = \sum_{\alpha=1}^{\mathcal{L}} \sum_{\substack{i=1 \\ (i,\alpha)\in S}}^{N} \sum_{\substack{j=1 \\ (j,\alpha)\in S}}^{i-1} C^{\mathrm{org}}_{ij\alpha}, \tag{9}$$

where (i, α) represents gene i in layer α, and the summation is over all node pairs $((i, \alpha), (j, \alpha))$ in S. We exclude the diagonal elements, i.e., $C^{\mathrm{org}}_{ii\alpha}$ in Eq (9) because they are equal to 1 for correlation matrices.

Let $C^{\mathrm{con}}_{\alpha}$ be a sample covariance matrix for layer α generated by the configuration model for correlation matrices [74]. Let $C^{\mathrm{con}}_{ij\alpha}$ be the (i, j)-th element of $C^{\mathrm{con}}_{\alpha}$. Using $\mathrm{E}[C^{\mathrm{con}}_{\alpha}] = C_{\alpha}$, where C_{α} is the covariance matrix for the estimated multivariate normal distribution for layer α [74], we obtain

$$\mathrm{E}\left[\sum_{\alpha=1}^{\mathcal{L}} \sum_{\substack{i=1\\(i,\alpha)\in S}}^{N} \sum_{\substack{j=1\\(j,\alpha)\in S}}^{i-1} C^{\mathrm{con}}_{ij\alpha}\right] = \sum_{\alpha=1}^{\mathcal{L}} \sum_{\substack{i=1\\(i,\alpha)\in S}}^{N} \sum_{\substack{j=1\\(j,\alpha)\in S}}^{i-1} \mathrm{E}\left[C^{\mathrm{con}}_{ij\alpha}\right] = \sum_{\alpha=1}^{\mathcal{L}} \sum_{\substack{i=1\\(i,\alpha)\in S}}^{N} \sum_{\substack{j=1\\(j,\alpha)\in S}}^{i-1} C_{ij\alpha}. \quad (10)$$

We obtain

$$\mathrm{Var}\left[\sum_{\alpha=1}^{\mathcal{L}} \sum_{\substack{i=1\\(i,\alpha)\in S}}^{N} \sum_{\substack{j=1\\(j,\alpha)\in S}}^{i-1} C^{\mathrm{con}}_{ij\alpha}\right] = \frac{1}{L}\left[\sum_{\alpha=1}^{\mathcal{L}} \sum_{\substack{i=1\\(i,\alpha)\in S}}^{N} \sum_{\substack{j=1\\(j,\alpha)\in S}}^{i-1} \sum_{\substack{k=1\\(k,\alpha)\in S}}^{N} \sum_{\substack{r=1\\(r,\alpha)\in S}}^{k-1} \left(C_{ik\alpha}C_{jr\alpha} + C_{ir\alpha}C_{jk\alpha}\right)\right]. \quad (11)$$

We show the derivation of Eq (11) in Text C in S1 Text. Note that

$$\mathrm{Var}\left[\sum_{\alpha=1}^{\mathcal{L}} \sum_{\substack{i=1\\(i,\alpha)\in S}}^{N} \sum_{\substack{j=1\\(j,\alpha)\in S}}^{i-1} C^{\mathrm{con}}_{ij\alpha}\right] \propto \frac{1}{L}, \quad (12)$$

which is consistent with the central limit theorem.

2.5 Determining a resolution parameter value

For simplicity, we assume γ_α to be common for all layers and denote the common value by γ. We use the Convex Hull of Admissible Modularity Partitions (CHAMP) algorithm version 2.1.0 [107, 108] to determine the γ value. The CHAMP algorithm takes a set of partitions generated by any community detection method as input and identifies the parameter regions in which each partition attains the largest modularity among all the partitions. The algorithm then obtains a pruned subset of admissible partitions and allows one to select parameter values corresponding to more robust community structures, which are large parameter regions in which the same partition maximizes the modularity.

Because we inform the interlayer coupling strength values by the empirical data as we described in section 2.2, we only need to tune the γ value. Therefore, using 15 evenly spaced γ values ranging from $\gamma = 1$ to $\gamma = 4$, we run a multilayer community detection method to obtain 15 partitions, one for each γ value, for a given multilayer network. Then, we employ the one-dimensional CHAMP on the 15 corresponding partitions to identify the ranges of γ in which the same partition maximizes the modularity. The wider ranges of γ correspond to more robust ranges of γ, so we choose a γ value in the two widest ranges according to CHAMP.

2.6 Specialist and generalist communities

The communities in multilayer correlation matrices and multilayer networks determined by the maximization of multilayer modularity may span multiple layers. We refer to a community containing genes belonging to various layers, i.e., tissues, as a generalist community. We refer to a community that contains genes in mostly just one tissue as a specialist community. The genes in a generalist community are general in the sense that they are co-expressed similarly across multiple tissues, whereas the genes in a specialist community are specialist in the sense that they are uniquely co-expressed in a single tissue. We will give the precise definitions of a generalist community and a specialist community in the following text. These different types of communities occur due to the similarity or difference between gene co-expression patterns across different tissues. In particular, some pairs of genes show co-expression across individuals in only specific tissues and others in multiple tissues. We are interested in whether our community detection method can detect these different types of communities. Therefore, we need a measure to classify each detected community as a generalist community or a specialist community.

We define a measure called the specialist fraction to quantify how specialized any multilayer community is as follows. For a given community, we first find the number of genes unique to each tissue α, i.e., the genes i for which node (i, α) belongs to the community and node (i, β) does not for any $\beta \neq \alpha$. Second, we define the specialist tissue of the community as the tissue that has the largest number of unique genes. The specialist fraction is the number of genes unique to the specialist tissue divided by the total number of nodes in the community. If the community lies within one layer, the specialist fraction is equal to 1. A large value of the specialist fraction suggests that the community is a specialist community. Genes unique to a specialist community may have functions specific to the tissue. In contrast, if all genes belong to at least two tissues, the specialist fraction is equal to 0. If many genes belong to different tissues in the community, the specialist fraction is low, suggesting that the community is relatively a generalist community. Genes in a generalist community may have functions expressed across various tissues.

2.7 Gene set enrichment analysis

To explore the biological processes associated with the set of genes constituting a detected community, we carried out a gene set enrichment analysis. It is a standard method for detecting statistically significant enriched biological processes, pathways, regulatory motifs, protein complexes, and disease phenotypes in the given gene set. We use g:Profiler (version e109_eg56_p17_1d3191d) for this purpose [109] and restrict our analysis to the Gene Ontology biological process (GO:BP) release 2023–03-06 [110, 111] and Human phenotype ontology (HP) release 2023–01-27 [112] results. We use a Benjamini-Hochberg FDR significance threshold [113] of 0.05.

2.8 Localization of genes on chromosomes

We developed statistical methods to investigate whether the genes in a community detected by our community detection method are physically clustered across the genome. To this end, we first ask whether a group of genes are more frequently located on the same chromosome than a control. Consider a group of genes, denoted by c. Let n be the number of genes in group c. We define the fraction of pairs of genes on the same chromosome as

$$x_c = \frac{\text{number of pairs of genes in group } c \text{ on the same chromosome}}{n(n-1)/2}. \tag{13}$$

The denominator of x_c is equal to the number of pairs of genes in group c and gives the normalization. For the control, we uniformly randomly shuffle the association between the $N = 203$ genes that we initially selected for our analysis and the chromosome to which each of the N genes belongs. After this random shuffling, the n genes are randomly distributed on various chromosomes as the $N = 203$ genes are distributed on those chromosomes. Then, we calculate x_c^{rand} according to this random distribution of the n genes using Eq (13). We repeat this randomization 100 times and calculate the average and standard deviation of x_c^{rand}, and then the Z score. If the Z score is significantly positive, then we say that the group of genes c has more pairs of genes on the same chromosome than the control.

Second, we tested whether the genes in c are located closer to each other on the chromosome than a control, given the number of genes in c on each chromosome. To this end, we define the physical distance measured in base pairs between gene i and gene j on the same chromosome, $d(i, j)$, as follows. Without loss of generality, assume that the end position of gene i is less than the start position of gene j. Then, we set

$$d(i,j) = (\text{start position of gene } j) - (\text{end position of gene } i). \tag{14}$$

Furthermore, we define the average distance between genes in group c as

$$d_c = \frac{\sum_{i,j \text{ in group } c \text{ on the same chromosome}} d(i,j)}{\text{number of pairs of genes in group } c \text{ on the same chromosome}}. \tag{15}$$

Denote by n_k the number of genes in group c that are on chromosome k. Note that the denominator in Eq (15) is equal to $\sum_k n_k(n_k - 1)/2$. For the control, for each k, we choose n_k genes uniformly at random out of all genes on chromosome k from the N genes. We carry out this procedure for all chromosomes k on which there are at least two genes in group c (i.e., $n_k \geq 2$). Then, we calculate d_c for this random distribution of genes, which we refer to as d_c^{rand}. We repeat this randomization 100 times and calculate the average and standard deviation of d_c^{rand}, and then the Z score. If the Z score is significantly negative, then we say that the genes in group c are localized on the chromosomes.

Third, we test whether the genes in c are located closer to each other than a control on a given chromosome. We define the average distance between genes in group c on chromosome k as

$$\tilde{d}_{c,k} = \frac{\sum_{i,j \text{ in group } c \text{ on chromosome } k} d(i,j)}{n_k(n_k - 1)/2}. \tag{16}$$

For the control, we choose n_k genes uniformly at random out of all genes that are among the N genes and on chromosome k. Then, we calculate $\tilde{d}_{c,k}$ for this random distribution of genes, which we refer to as $\tilde{d}_{c,k}^{\text{rand}}$. We repeat this randomization 100 times and calculate the average and standard deviation of $\tilde{d}_{c,k}^{\text{rand}}$, and then the Z score. We carry out this procedure for each chromosome k on which there are at least two genes in group c (i.e., $n_k \geq 2$). We apply the Bonferroni correction [114] separately to each c to determine which communities have a significantly smaller average distance between pairs of genes on a specific chromosome than the control. We chose to apply the Bonferroni correction because it is a more conservative statistical method than others, such as FDR.

2.9 Pancreas-specific cis-eQTL analysis

Expression quantitative trait loci (eQTL) analysis identifies variants that have significant associations with expression levels of specific genes. We hypothesize that changes in expression levels of a pair of co-expressed genes are associated with the same set of variants. If true, we expect to identify variants that are associated with the expression of both genes in the pair. To investigate gene pairs with shared eQTL single nucleotide polymorphisms (SNPs) in the pancreas, we downloaded the cis-eQTL data set from GTEx release V8. This data set involves SNP-gene pairs with association significance indicated with a nominal p value. The changes in the expression levels of a given gene may be associated with one or multiple SNPs.

Alternatively, it may have no eQTLs, meaning that no SNPs are associated with its gene expression. Using this data set, we searched for SNPs that were associated with both of the genes in a given gene pair of interest. Given that we are interested in whether co-expressed genes share common SNPs, we only investigate gene pairs with co-expression (as defined by Eq (1)) greater than 0.5 in the pancreas.

3 Results

3.1 Communities in the multilayer correlation matrix

We compare the gene communities obtained from the multilayer correlation matrix and those obtained from multilayer gene networks constructed using graphical lasso. For a brief review of graphical lasso, see Text D in S1 Text.

We run iterated GenLouvain [97] on the multilayer correlation matrix to approximately maximize the multilayer modularity at each value of the resolution parameter, γ, which we assume to be common for all layers. We then use the CHAMP algorithm to determine optimal values of γ [107, 108]. We show the results of CHAMP in Fig 2(a). The figure indicates that robust ranges of γ, which are relatively wide ranges of γ in which the optimal partition is the same and correspond to relatively long straight line segments in the figure, are approximately $0 < \gamma < 1.2$ or $2.9 < \gamma < 4.4$. Therefore, we examine the node partitions with one arbitrary γ value from each of these two stable regions of γ, i.e., $\gamma = 1$ and $\gamma = 3$.

We show the composition of the resulting node partitions with $\gamma = 1$ and $\gamma = 3$ in Fig 3(a) and 3(b), respectively. As expected, the number of communities increases when γ increases. We show the Z score for the total intralayer weight within each community detected with $\gamma = 1$ and $\gamma = 3$ in Table 2. With $\gamma = 3$, communities 8 through 12 contain no intralayer edges such that one cannot run the randomization, leading to a null Z score. These communities contain only one gene; communities 8, 9, 10, and 11 detected with $\gamma = 3$ contain two nodes representing the same gene in two different tissues, and community 12 contains only one node. We omitted these trivial communities in Table 2. The table indicates that all the communities detected with $\gamma = 1$ and all the communities containing at least two genes detected with $\gamma = 3$ (i.e., communities 1 through 7) are statistically significant.

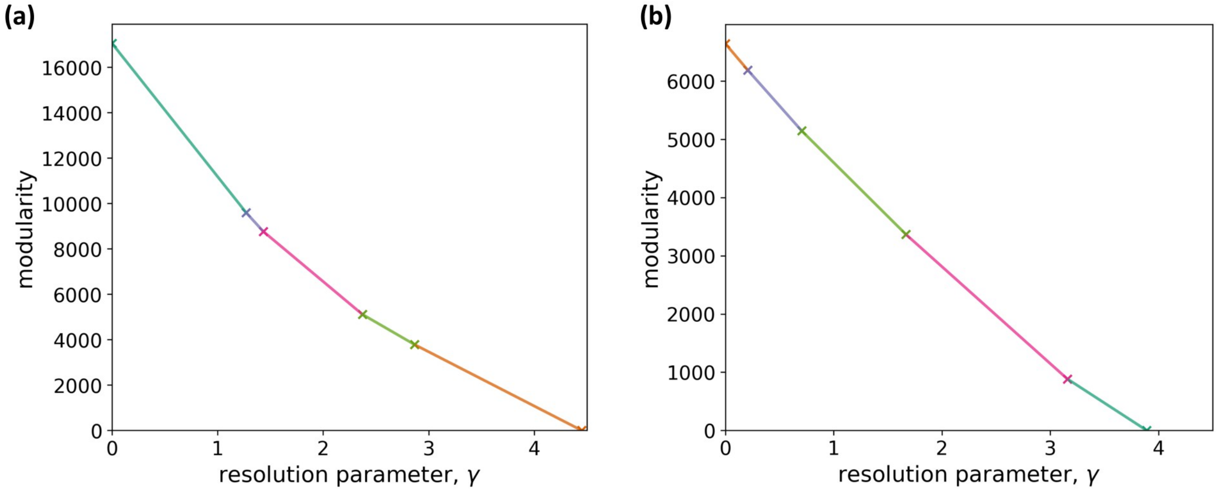

Fig 2. Determination of the resolution parameter value by CHAMP. (a) Multilayer correlation matrix. (b) Multilayer network obtained by graphical lasso. The convex hull of the lines in the (γ, Q) plane, each of which corresponds to a node partitioning, is a piecewise linear curve with the transition values indicated by a cross and change in the line color. Each line segment corresponds to the optimal node partitioning in the corresponding range of γ.

https://doi.org/10.1371/journal.pcbi.1011616.g002

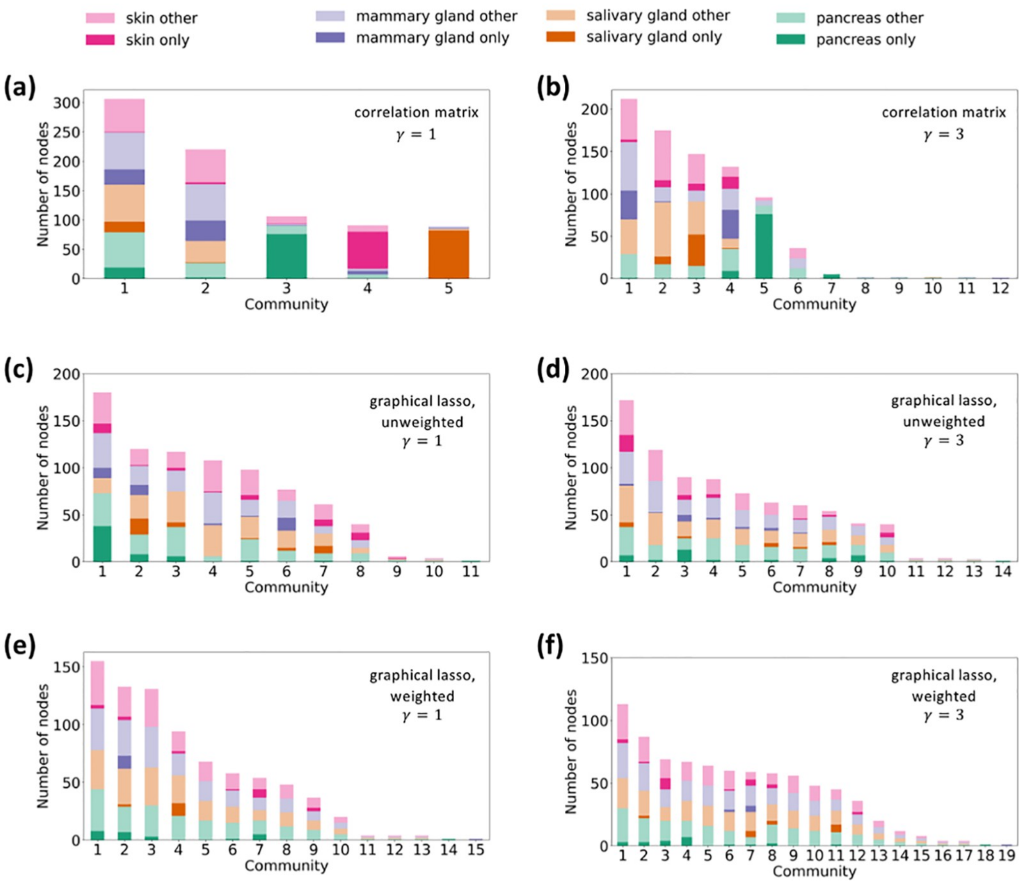

Fig 3. Composition of each community by layer, i.e., tissue. (a) Multilayer correlation matrix, $\gamma = 1$. (b) Multilayer correlation matrix, $\gamma = 3$. (c) Unweighted multilayer network obtained by graphical lasso, $\gamma = 1$. (d) Unweighted multilayer network obtained by graphical lasso, $\gamma = 3$. (e) Weighted multilayer network obtained by graphical lasso, $\gamma = 1$. (f) Weighted multilayer network obtained by graphical lasso, $\gamma = 3$. The darker shades indicate nodes corresponding to genes that only appear in one layer in the given community. The lighter shades indicate nodes corresponding to genes that appear in multiple layers in the community.

https://doi.org/10.1371/journal.pcbi.1011616.g003

Table 2. Z scores for the total intralayer weight within each community detected in the multilayer correlation matrix.

$\gamma = 1$		$\gamma = 3$	
Comm.	**Z score**	**Comm.**	**Z score**
1	24.432	1	68.526
2	73.282	2	55.267
3	62.569	3	72.008
4	14.972	4	19.071
5	65.318	5	124.080
		6	40.288
		7	14.699

Comm. denotes community.

https://doi.org/10.1371/journal.pcbi.1011616.t002

Both node partitions contain some communities that appear to be generalist communities and other communities that appear to be specialist communities. We remind that a generalist community indicates genes that are similarly co-expressed in multiple tissues and that a specialist community indicates genes that are uniquely co-expressed in one tissue. To quantify these findings, we show in Table 3 the specialist fraction for each community in the partition

Table 3. Specialist fraction and the corresponding tissue for each community detected in the multilayer correlation matrix and for each community detected in the multilayer networks obtained by graphical lasso.

Multilayer correlation matrix									
$\gamma = 1$					$\gamma = 3$				
Comm.	No. genes	No. specialist genes	Specialist fraction	Specialist tissue	Comm.	No. genes	No. specialist genes	Specialist fraction	Specialist tissue
1	153	26	0.085	mammary gland	1	96	34	0.160	mammary gland
2	104	35	0.159	mammary gland	2	84	9	0.051	salivary gland
3	92	76	0.717	pancreas	3	87	37	0.252	salivary gland
4	80	63	0.692	skin	4	88	34	0.258	mammary gland
5	86	82	0.921	salivary gland	5	86	76	0.792	pancreas
					6	12	0	0.000	N/A
					7	5	5	1.000	pancreas
Unweighted multilayer network obtained by graphical lasso									
$\gamma = 1$					$\gamma = 3$				
Comm.	No. genes	No. specialist genes	Specialist fraction	Specialist tissue	Comm.	No. genes	No. specialist genes	Specialist fraction	Specialist tissue
1	102	38	0.211	pancreas	1	80	18	0.105	skin
2	62	17	0.142	salivary gland	2	37	2	0.017	pancreas
3	48	6	0.051	pancreas	3	47	13	0.144	pancreas
4	36	2	0.019	mammary gland	4	33	4	0.045	skin
5	35	5	0.051	skin	5	21	2	0.027	mammary gland
6	35	14	0.182	mammary gland	6	24	4	0.063	salivary gland
7	32	8	0.131	salivary gland	7	18	2	0.033	salivary gland
8	17	8	0.200	skin	8	23	4	0.074	pancreas
9	3	1	0.167	skin	9	18	7	0.171	pancreas
					10	15	5	0.125	skin
Weighted multilayer network obtained by graphical lasso									
$\gamma = 1$					$\gamma = 3$				
Comm.	No. genes	No. specialist genes	Specialist fraction	Specialist tissue	Comm.	No. genes	No. specialist genes	Specialist fraction	Specialist tissue
1	51	8	0.052	pancreas	1	35	3	0.027	pancreas
2	54	11	0.083	mammary gland	2	30	3	0.034	pancreas
3	38	3	0.023	pancreas	3	31	9	0.130	skin
4	38	11	0.117	salivary gland	4	23	7	0.104	pancreas
5	17	0	0.000	N/A	5	16	0	0.000	N/A
6	16	1	0.017	pancreas	6	19	2	0.033	mammary gland
7	24	7	0.130	skin	7	32	5	0.085	salivary gland
8	12	0	0.000	N/A	8	23	3	0.052	salivary gland
9	12	3	0.081	skin	9	14	0	0.000	N/A
10	5	0	0.000	N/A	10	12	0	0.000	N/A
					11	19	6	0.133	salivary gland
					12	11	2	0.056	skin
					13	5	0	0.000	N/A
					14	3	0	0.000	N/A
					15	2	0	0.000	N/A

Comm. denotes community and No. denotes "number of".

https://doi.org/10.1371/journal.pcbi.1011616.t003

with $\gamma = 1$. Communities 3, 4, and 5 have specialist fractions greater than 0.5, so we regard them as specialist communities. In contrast, because communities 1 and 2 have specialist fractions substantially less than 0.5, we regard them as generalist communities. The same table also shows the specialist fraction for each significant community found with $\gamma = 3$. Communities 5 and 7 have specialist fractions greater than 0.5. Both of them are pancreas specialist communities. We regard communities 1, 2, 3, 4, and 6, whose specialist fraction is substantially less than 0.5, as generalist communities.

3.2 Communities in the multilayer networks obtained by graphical lasso

For comparison purposes, we run the iterated GenLouvain on the multilayer networks that we constructed using graphical lasso (see Text D in S1 Text for the methods). The results of CHAMP on the detected node partition of the unweighted network, shown in Fig 2(b), indicate that the optimal ranges of γ are approximately $0.7 < \gamma < 1.7$ or $1.7 < \gamma < 3.2$. Therefore, we use the same γ values as those for our multilayer correlation matrix, i.e., $\gamma = 1$ and $\gamma = 3$.

We show the composition of the resulting node partitions of the unweighted network obtained using graphical lasso with $\gamma = 1$ and $\gamma = 3$ in Fig 3(c) and 3(d), respectively. With $\gamma = 1$, we find eleven communities, nine of which are significant. With $\gamma = 3$, we find fourteen communities, ten of which are significant. See Text B in S1 Text for the statistical results. We also show the composition of the node partitions of the weighted multilayer network obtained using graphical lasso with $\gamma = 1$ and $\gamma = 3$ in Fig 3(e) and 3(f), respectively.

Fig 3(c)–3(f) suggests that these partitions apparently contain generalist communities only. Table 3 shows the specialist fraction for each significant community in the unweighted network and each community in the weighted network. Note that we have not evaluated the significance of the communities detected for the weighted multilayer network because the configuration model for weighted networks, which is necessary for constructing a significance test, is not a straightforward concept [115, 116]. For the unweighted network, with both $\gamma = 1$ and $\gamma = 3$, all the significant communities have specialist fractions at most 0.211. For the weighted network, with both $\gamma = 1$ and $\gamma = 3$, all the communities with more than one gene have specialist fractions at most 0.133. Therefore, we conclude that there are no specialist communities for either the unweighted or weighted network and with either $\gamma = 1$ or $\gamma = 3$.

In sum, our community detection method on correlation matrices finds tissue-specific gene co-expression patterns, evident by the detection of specialist communities, whereas the graphical lasso does not. Because we are interested in comparing the biological implications of specialist communities versus generalist communities, in the following sections, we only analyze the communities detected for our multilayer correlation matrix. In particular, we will carry out tissue-specific analysis to investigate the specialist communities detected by our method.

3.3 Localization of genes on chromosomes

To investigate the possible localization of genes in the detected communities on the chromosomes, we first analyze whether the $N = 203$ among the 56, 200 genes that we are analyzing in the GTEx data set are already localized in the genome. The Z score for a fraction of pairs of genes on the same chromosome is 6.735 ($p < 10^{-6}$), which suggests that the $N = 203$ genes are distributed on different chromosomes in a highly biased manner relative to how all the 56, 200 genes are distributed. The Z score for the average distance between pairs of genes on the same chromosome is −6.059 ($p < 10^{-6}$). Therefore, the average distance between pairs of genes among the $N = 203$ genes is significantly smaller than by chance. This result is expected given that highly expressed genes in glandular tissues cluster in specific loci [40]. We show the Z

Table 4. Z score for the average distance between pairs of genes on each chromosome for the $N = 203$ genes.

Chr	Z score	Chr	Z score
1	−1.881	14	0.085
2	−3.540	15	−0.404
3	1.293	16	0.599
4	−4.736	17	−4.842
5	−1.858	18	N/A
6	−0.422	19	−2.545
7	−0.873	20	−1.458
8	0.691	21	−0.317
9	0.276	22	−0.564
10	0.552	X	2.103
11	−1.112	Y	N/A
12	−3.512	M	−0.858
13	N/A		

M stands for the mitochondrial chromosome. Chr denotes chromosome.

https://doi.org/10.1371/journal.pcbi.1011616.t004

scores for the average distance between pairs of genes on each chromosome, analyzed separately, in Table 4. At a significance level of $p = 0.05$, there is significant localization of genes on chromosomes 2 ($p = 0.0088$; Bonferroni corrected; same for the following p values), 4 ($p < 10^{-4}$), 12 ($p = 0.0098$), and 17 ($p < 10^{-4}$).

Next, we run the same localization analysis for each community in the multilayer correlation matrix detected with $\gamma = 1$ and $\gamma = 3$. For a generalist community, we only included the genes in the community that appear in at least three out of the four tissues in this analysis. This is because such genes may play functional roles, which the generalist community represents, across many types of tissues. With this restriction, each gene is present in at most one generalist community. Note that, without this restriction, a gene may appear in multiple generalist communities because the four nodes in the multilayer network representing the same gene may belong to different communities. We exclude this case for simplicity.

For each community, we show in Table 5 the Z score for the fraction of pairs of genes in the community that are on the same chromosome. With $\gamma = 1$, communities 2 ($p < 10^{-4}$) and 3

Table 5. Analysis of localization of genes in each community detected in the multilayer correlation matrix.

$\gamma = 1$			$\gamma = 3$		
Comm.	Z score for x_c	Z score for d_c	Comm.	Z score for x_c	Z score for d_c
1	1.520	−1.095	1	6.190	−3.251
2	7.094	−2.710	2	−0.388	0.652
3	3.940	−1.245	3	0.879	−0.405
4	0.160	0.179	4	−0.281	0.924
5	−0.102	−0.344	5	4.485	−1.175
			6	6.634	−2.255
			7	−0.845	N/A

Note that x_c is the normalized fraction of pairs of genes in the community on the same chromosome and that d_c is the normalized distance between two genes in the community on the same chromosome. Comm. denotes community.

https://doi.org/10.1371/journal.pcbi.1011616.t005

($p = 4.05 \cdot 10^{-4}$) have significantly more genes among the $N = 203$ genes on the same chromosome than by chance. The same table also shows the Z score for the average distance on the chromosome between pairs of genes in the same community for each community. We find that, with $\gamma = 1$, community 2 has a significantly smaller average gene-to-gene distance than by chance ($p = 0.0336$). With $\gamma = 3$, communities 1 ($p < 10^{-4}$), 5 ($p < 10^{-4}$), and 6 ($p < 10^{-4}$) have significantly more pairs of genes on the same chromosome than by chance, and community 1 ($p = 0.0069$) has a significantly smaller average gene-to-gene distance than by chance (see Table 5).

We then compute the Z score for the average distance between pairs of genes separately for each chromosome in addition to each community. We exclude the community-chromosome pairs that have less than three genes from this analysis. With both $\gamma = 1$ and $\gamma = 3$, no group of genes on a specific chromosome in a specific community is significantly clustered when we impose the Bonferroni correction over all the community-chromosome pairs (45 and 23 pairs with $\gamma = 1$ and $\gamma = 3$, respectively; see Tables B and C in S1 Text for the Z scores). With the Bonferroni correction applied to each community separately, there are still no significant clusters in the partition with $\gamma = 1$. However, with $\gamma = 3$, we find that the genes in community 1 on chromosome 1 ($p = 0.0199$) and those in community 5 on chromosome 11 ($p = 0.0371$) are significantly clustered.

3.4 Functional analysis of selected communities

For the communities detected for our multilayer correlation matrix, we found clusters of physically localized genes within two communities with $\gamma = 3$ but none with $\gamma = 1$. Because we are interested in exploring biological implications of localized clusters of genes, we carry out further analysis on the node partition with $\gamma = 3$ in this section. A table showing which nodes (i.e., genes) belong to which communities in this partition is available on GitHub [76].

First, we conducted an enrichment analysis of the communities identified with $\gamma = 3$. We started with an enrichment analysis for the top 50 genes that have the highest expression out of the 203 genes in the network in each tissue. We find that, in all tissues, the top 50 highly expressed genes are enriched significantly in well-established housekeeping categories, such as oxidative phosphorylation and aerobic electron transport chain (FDR <0.05; see Table D in S1 Text). Echoing this finding, one of the modules that we identified (community 1) shows similar enrichment for mitochondrial function, such as aerobic electron transport chain ($p = 1.05 \cdot 10^{-10}$) and oxidative phosphorylation ($p = 5.90 \cdot 10^{-11}$) (see Table E in S1 Text). However, in the other six communities, our network approach identifies novel gene modules with functional enrichments in epidermis development (community 2, $p = 1.90 \cdot 10^{-24}$), keratinization (community 5, $p = 1.95 \cdot 10^{-19}$), positive regulation of respiratory burst (community 6, $p = 5.36 \cdot 10^{-8}$), and adaptive thermogenesis (community 7, $p = 1.73 \cdot 10^{-2}$). Furthermore, these modules are enriched with diseases relevant to the tissues examined, such as hyperkeratosis (community 2, $p = 3.05 \cdot 10^{-7}$) and recurrent pancreatitis (community 1, $p = 1.78 \cdot 10^{-19}$). In addition, we analyzed the top 50 highly connected genes (i.e., top 50 genes in terms of the weighted degree, or in other words, top 50 hub genes) in each of the single-layered networks for each tissue. Not surprisingly, this analysis identified genes that are enriched for functions and diseases that are specific to each tissue (see Table F in S1 Text). However, we found that most of the genes that are identified in our multilayer network approach are different from those identified with single-layer analysis (see Text G in S1 Text). We also found that the functional enrichments of these two network approaches were different (see Table E versus Table F in S1 Text). Overall, our method provides additional biological insights than simple expression-level filtering and single-layer network analysis.

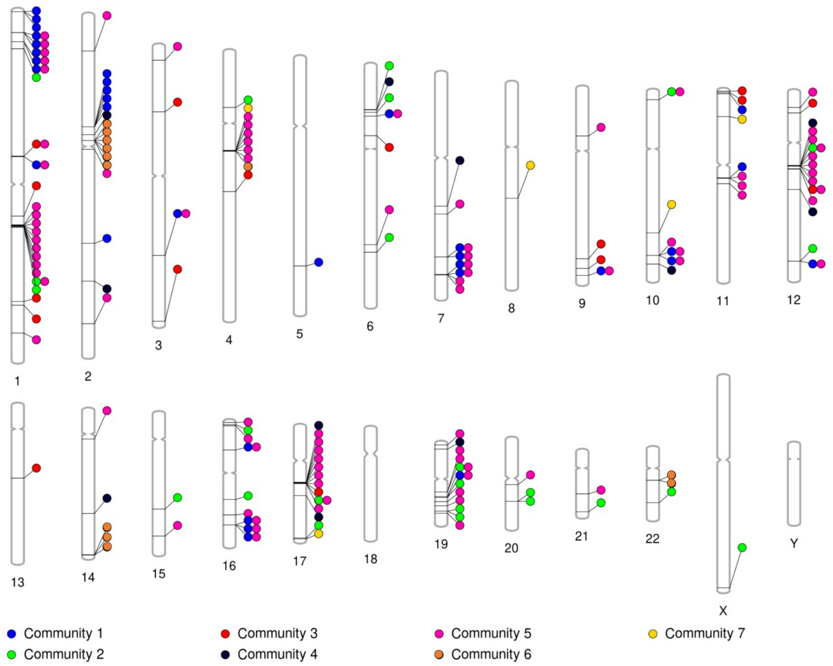

Fig 4. Location of genes on chromosomes, colored by community. There is a colored circle for each associated community pointing to each gene. Note that a gene can belong to more than one community, denoted by multiple colored circles next to each other horizontally pointing to the same gene. This figure allows us to visually see clusters of genes on specific chromosomes and their associated community.

https://doi.org/10.1371/journal.pcbi.1011616.g004

Our multilayer network analysis allowed us to investigate genes that are co-expressed in multiple tissues. We surmised that membership of genes in the same community can be facilitated by shared regulatory sequences affecting multiple genes at the same time. Given that regulatory regions affect gene expression in cis (i.e., nearby regions), we hypothesize that genes in the same multilayer community may be physically close to each other. To investigate this, we visualize in Fig 4 the location of the genes in the different communities on the chromosomes. As in the localization analysis presented in section 3.3, for a generalist community, we only show in Fig 4 the genes in the community that are present in at least three tissues. In Fig 4, a color of the circles represents a community. Note that a gene can belong to more than one community, denoted by multiple colored circles next to each other horizontally pointing to the same gene. It happens to be the case that a gene is associated with a maximum of two different communities, hence a maximum of two colored circles pointing to the same gene. Visually, Fig 4 suggests some tight clusters of genes, especially in community 5.

In section 3.3, we found significantly localized clusters of genes in community 1 on chromosome 1 and in community 5 on chromosome 11 in the partition with $\gamma = 3$. It is somewhat surprising that only these two community-chromosome pair gene sets are significantly localized because there appear to be more localized clusters in Fig 4. A possible reason for this discrepancy is that, besides the genes in the community-chromosome pair of interest, there are so few other genes on the chromosome that the random shuffling of gene associations does not provide sufficient randomization. In this case, the empirical average distance between genes in the community-chromosome pair will not be statistically different from the average distance for the randomized data. Therefore, here we directly compared the average distance between pairs of genes on each community-chromosome pair, as defined by Eq (16), to that for community 5 on chromosome 11. We decided to analyze community 5 because it is a pancreas specialist community while community 1 is a generalist community, as we discussed in regards to functional enrichment earlier in this section.

We denote the average distance between the pairs of genes among the three genes in community 5 on chromosome 11 by $\tilde{d}_{5,11}$, calculated using Eq (16). We looked for any community-chromosome pair, containing all the genes in the selected community on the selected chromosome, with at least three genes whose average distance between genes is less than $\tilde{d}_{5,11}$. There are five such additional gene clusters: community 1 on chromosome M, which contains 15 genes, community 5 on chromosome 4, which contains 6 genes, community 5 on chromosome 17, which contains 9 genes, community 6 on chromosome 2, which contains 6 genes, and community 6 on chromosome 14, which contains 3 genes. Among all these community-chromosome pairs, we focused on the three gene clusters in community 5, including the gene cluster on chromosome 11. We opted to do so because community 5 is a pancreas specialist community, whereas communities 1 and 6 are generalist communities.

After initial investigation of the three gene clusters in community 5, i.e., one each on chromosome 4, 11, and 17, we further analyzed the one on chromosome 17, because keratin loci have been discussed in the context of human evolution [117, 118]. We show in Fig 5A and 5B the expression of each gene in this gene cluster in the skin and pancreas, respectively. We found that gene expression trends vary between the two tissues. Specifically, our method identified community 5 because of co-expression trends in the pancreas. However, in terms of the sheer expression level, the present gene cluster is expressed multiple folds higher in the skin

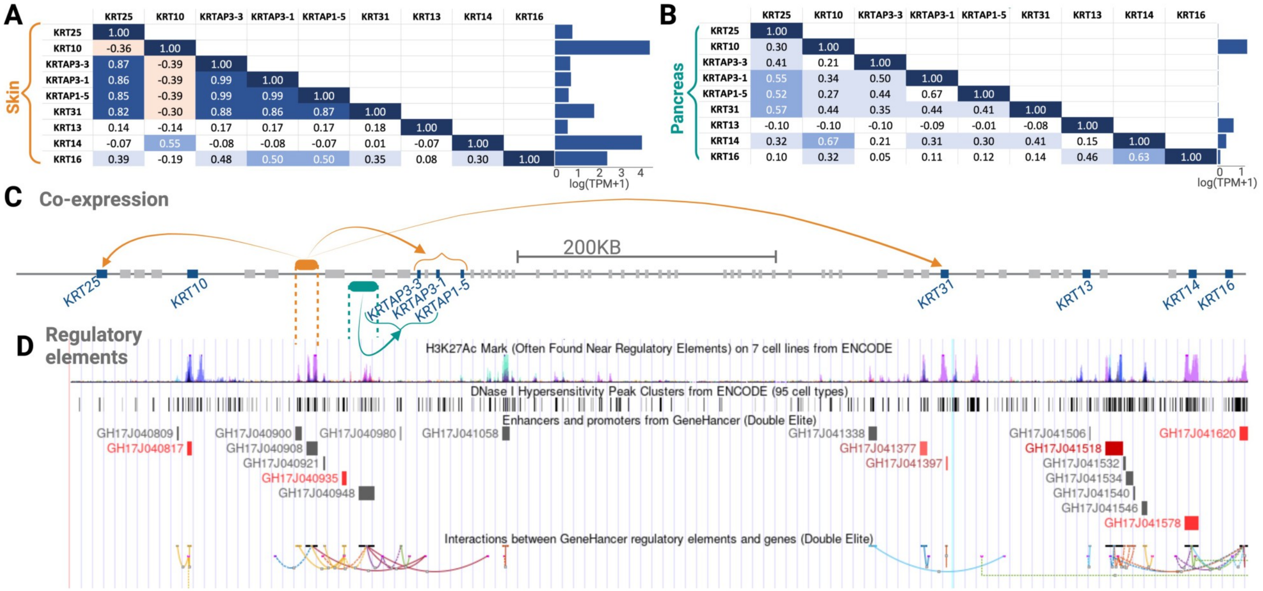

Fig 5. Expression and co-expression analysis of a cluster of genes in community 5 on chromosome 17. The co-expression matrices for these genes in (A) skin and in (B) pancreas are shown. The average expression for each gene in these tissues is shown in the bar graphs. The location of these genes on chromosome 17 is shown in (C), with arrows (colored according to the associated tissue) pointing from putative regulatory elements to highly co-expressed genes. (D) The panel shows different measures of the regulatory potential of this genome section. From top to bottom: 1. H3K27AC modification to histone H3 within the region, which often correlates with activation of transcription and is associated with active enhancers in a given tissue available through ENCODE database [119]. 2. DNAse1 hypersensitivity sites. They are sections of the genome that are cut by DNAse1 enzyme. Given that the chromatin has to be "open" for the DNAse to access the sequence, the sequences that are cut by DNAse indicate open chromatin, which is in turn associated with regulatory activity. Data are available through ENCODE database [119]. 3. Enhancer/promoters. These are sequences that are predicted as enhancers (gray) and promoters (red) from the GeneHancer database [120]. 4. Established interactions between regulatory regions and genes as documented by GeneHancer database [120]. These data sets combined with our co-expression analysis provide a novel outlook into potential topologically associated domains that may be regulated by specific sequences in a tissue-specific manner.

https://doi.org/10.1371/journal.pcbi.1011616.g005

than pancreas. Further, we found that the co-expression patterns for some gene pairs within this gene cluster are common between the skin and pancreas but differ for other gene pairs. The physical clustering of the genes that are co-expressed implicates genetic variation in shared gene regulatory factors as the main basis for co-expression. For example, a search of the GTEx eQTL database showed that the common single nucleotide polymorphism rs12450846 is significantly ($p < 10^{-18}$) associated with lower expression of *KRT31* in the skin but higher expression of this gene in ovaries ($p < 0.005$). Unfortunately, this analysis was not conducted in the pancreas. Regardless, this polymorphism and the haplotype linked to it regulate multiple other keratins and keratin-associated protein genes in this particular locus in a tissue and gene-specific manner according to the GTEx database. Thus, genetic variation that affects the efficacy of regulatory regions (Fig 5C) or the formation of topologically associated domains (Fig 5D) in a tissue-specific manner may underly the co-expression of the genes in community 5 on chromosome 17. Indeed, we found several topologically associated domains, enhancers, transcription factor binding sites, and open chromatins within this region, affecting co-expressed genes in a similar fashion (Fig 5). Overall, our analysis provides several exciting hypotheses for future work to investigate regulatory regions that target multiple nearby genes and explain tissue-specific co-expression trends.

Another interesting community we identified is community 7. The genes in this community are located on different chromosomes and are enriched for response to temperature change (adaptive thermogenesis; see Table E in S1 Text). Because they exist on different chromosomes, it is unlikely that these genes share any common regulatory sequences or topologically associating domains. Instead, their co-expression may be due to environmental stimuli that are shared among the samples at the time of sampling (e.g., warm or cold environments). If true, the co-expression is due to a response to environmental stimuli that is controlled by specific regulatory sequences with broad effects across the genome, such as transcription factors. Thus, our network analysis may be useful for identifying gene clusters that respond to different environments.

3.5 Gene pairs with shared associated SNPs in pancreas

As described in section 2.9, we hypothesized that genetic variation that affects gene expression in a tissue-specific manner can explain some of the co-expression trends we observed. Identifying such variation is challenging because of the huge amount of combinations that are possible between genetic variants and gene expression levels. To overcome this challenge and identify examples of where genetic variation may explain the co-expression trends and chromosomal clustering, we conduct an eQTL analysis considering only cis variants that are physically close to genes of interest. This analysis provides a list of variants (SNPs in this case) that are significantly associated with expression levels of nearby genes. We will refer to these SNPs as eQTLs. Using this approach, we identified three gene pairs (i.e., *CELA3B* and *CELA3A*; *AMY2B* and *AMY2A*; *REG3G* and *REG1B*) that share associated eQTLs in the pancreas out of all the gene pairs in the network of 203 genes with co-expression greater than 0.5.

Notably, out of these three gene pairs, two pairs, i.e., the *CELA3B-CELA3A* and *AMY2B-AMY2A* pairs, are not composed of hub genes within the pancreas single-layered network and are only identified through our multilayer network approach. Both pairs are within community 5. For example, if we searched for the top 86 genes in terms of the weighted degree in the pancreas to match the number of genes in community 5, we were not able to identify the *CELA3B-CELA3A* or *AMY2B-AMY2A* pairs. In contrast, the other gene pair with shared eQTLs (i.e., *REG3G* and *REG1B*) consists of two hub genes in the single-layered pancreas co-expression network. Therefore, we would have missed two out of three gene pairs that may be biologically interesting if we simply investigated hub genes in the pancreas.

Next, to identify the biological relevance of this putatively genetically determined co-expression pattern, we investigated the CELA3 locus. We identified a set of 96 variants from statistically significant eQTLs for both *CELA3A* and *CELA3B* in the pancreas. CELA3A and CELA3B, which are proteases, are produced as zymogens in the pancreas. They then perform their digestive function in the intestine once they have been transported there. It has previously been speculated that the presence of two *CELA3* copies provides a functional substitute for the lack of pancreatic expression of CELA1 in humans relative to pigs [121]. The 96 variants are present in the genomic region spanned by *HSPG2*, *CELA3A*, and *CELA3B*. The minor allele for each of these 96 variants is associated with a decreased expression of *CELA3A* and an increased expression of *CELA3B* in the pancreas. This observation may hint at a possible constraint on the combined expression level of *CELA3A* and *CELA3B* in the pancreas, further supporting the idea that CELA genes may have compensatory roles for the functions of other members in this gene family. To understand the population genetics trends affecting the regulatory variants that we identified, we analyzed 83 SNPs that are associated with gene expression of *CELA3A* and *CELA3B* and genotyped in the 1000 Genomes Project Phase-3 data set. We found that these variants form a single linkage-disequilibrium (LD) group in Europeans at an r^2 threshold of 0.6 [122]. The minor alleles of 10 of these variants are associated with a decreased blood phosphate concentration [123, 124] (see Fig 6). In order to identify putative causal variants in the LD group, we investigated whether any of these variants lie in a regulatory region. We find that four (rs57030248, rs59134693, rs113385886, and rs111651468) of these variants lie in an enhancer (ENSR00000350171), identified by ENSEMBL's variant effect predictor [125] (Fig 6B), which is active in the pancreas. Three of these four variants (rs57030248, rs59134693, and rs113385886) are both present in the enhancer region and associated with decreased blood phosphate levels. It is likely that one or more of these three variants are causal in the context of differences in the expression levels of *CELA3A* ($p = 1.1 \cdot 10^{-8}$, normalized effect size = −0.43) and *CELA3B* ($p = 4.5 \cdot 10^{-10}$, normalized effect size = +0.43). Our results allowed us to construct a hypothetical model (Fig 6). Our multilayer network approach facilitated the narrowing down of putatively causal genetic variants that affect the expression levels of negatively co-expressed gene pairs within the context of protein and phosphate metabolism.

4 Discussion

We developed a multilayer community detection method for Pearson correlation matrix data. We applied the proposed method to gene co-expression data from four tissues in humans to identify gene modules (i.e., communities). Some detected communities spanned multiple layers, which we refer to as generalist communities. Other communities lay mostly within one layer, specifically the pancreas layer, which we refer to as specialist communities. We then found that both generalist and specialist communities were localized on a smaller number of chromosomes than the expectation of random distribution of genes. As a case study, we closely looked into two groups of genes (i.e., the KRTAP cluster in community 5 and community 7 as a whole) and suggested that the detected multilayer communities may imply gene regulatory factors shared across different tissues or environmental stimuli shared among samples. Finally, we found three gene pairs that share associated eQTLs in the pancreas, identifying examples in which genetic variation may explain the co-expression trends and chromosomal clustering.

Various mutually inclusive factors can explain co-expression of genes [16, 33, 34]. We explored two such factors in our case study. First, it is possible that the regulatory regions

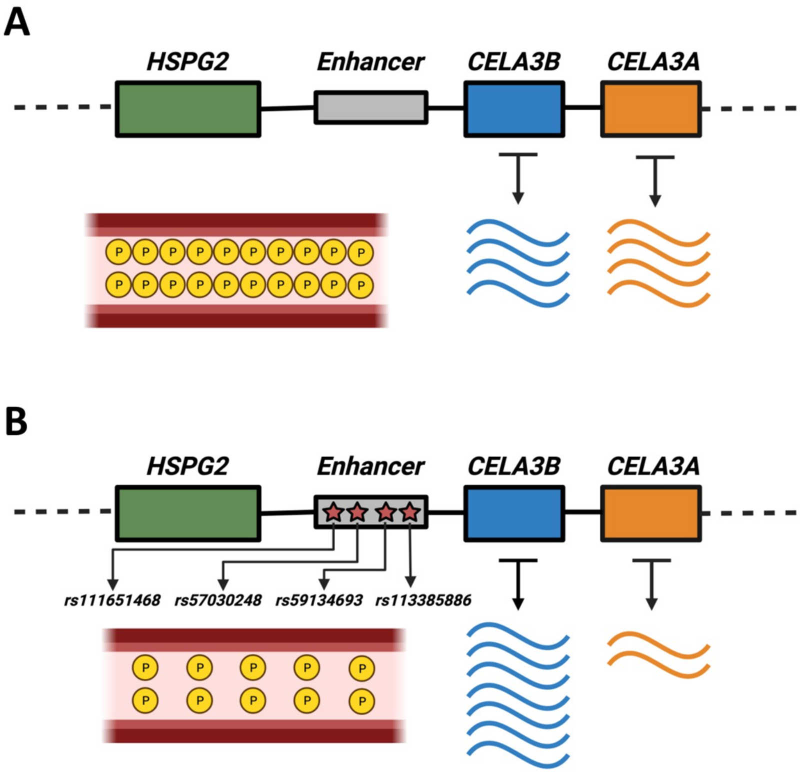

Fig 6. A schematic of SNPs in an enhancer region (gray box) that affect the expression of *CELA3A* (blue box) and *CELA3B* (orange box) in the pancreas and are associated with blood phosphate concentration. (A) Expression levels of *CELA3A* and *CELA3B*, and blood phosphate concentration when the derived alleles for the putatively causal SNPs are absent. (B) The presence of the derived alleles for the putatively causal SNPs decreases the expression level of *CELA3A*, increases the expression level of *CELA3B*, and decreases the blood phosphate concentration.

https://doi.org/10.1371/journal.pcbi.1011616.g006

control the expression of multiple genes in certain tissues [24, 26–29]. In this case, individuals who share genetic variations in these regulatory regions will have similar expression levels in these tissues where these regulators are active. If genetic variation underlies the co-expression of genes and the regulatory elements are cis (i.e., close physical proximity), we expect the co-expressed genes to cluster across the genome. We suggested that *KRTAP3–3*, *KRTAP3–1*, and *KRTAP1–5* share regulatory elements in skin and pancreas. Indeed, several recent studies highlight topologically associating domains as potential sites underlying co-expression of multiple proximate genes [29, 40]. Our approach integrated with chromatin accessibility (e.g., ATAC-seq) data is expected to facilitate identifying such loci where regulatory architecture may underlie the gene expression trends of multiple nearby genes in a tissue-specific manner. Second, it is possible that co-expressed genes have similar or complementing functions that respond to particular environmental conditions [30, 31]. For example, we suggested that the genes in community 7 detected in the multilayer correlation matrix with $\gamma = 3$ may be involved in response to temperature change and co-expressed because samples were subjected to respective environmental conditions at the time of sampling. We argue that the response to environmental stimuli may underlie co-expression in these genes and thus indicate phenotypic plasticity for related traits [126], where an individual can respond to different environmental cues by adjusting the expression levels of multiple genes [127]. Our approach can provide a systematic framework to study phenotypic plasticity using animal models comparing different environmental stimuli (e.g., temperature, pathogenic pressure, diet, xenobiotic substances).

Another particularly relevant study using GTEx data to construct tissue-specific gene co-expression networks compared the community structure across different tissues [41]. While the present study also uses GTEx data to construct tissue-specific gene co-expression networks and compare community structure across layers, the details of the methods differ in the following noteworthy ways. Azevedo et al. [41] apply a thresholding method to the correlation matrices to construct networks and use signed modularity [128] as the quality function for community detection, whereas the present study uses the correlation matrices directly with an appropriate correlation matrix null model in the quality function, as described in section 2.4. Additionally, we perform a multilayer community detection method that incorporates interlayer coupling strength information, whereas Azevedo et al. perform single layer community detection on each layer separately and then compare the community structure across networks using the global multiplexity index [129]. The global multiplexity index quantifies how many times two genes belong to the same communities across all the layers. To connect terminology in their study [41] and the present study, we point out that a group of genes with global multiplexity index equal to $\mathcal{L}$ (i.e., the total number of layers) corresponds to a generalist community that spans all layers of the multilayer network. This type of community is also called a pillar community [91]. A group of genes with global multiplexity index equal to 1 corresponds to a specialist community. Finally, a group of genes with global multiplexity index greater than 1 but less than $\mathcal{L}$ is a generalist community that spans a subset of the layers (of size equal to the global multiplexity index) in the multilayer network. This type of community is also called a semi-pillar community [91]. Both Azevedo et al. [41] and the present study employ enrichment analysis on the communities to identify known biological processes corresponding to the discovered gene communities. Systematic comparison between multilayer community detection methods, such as the present work, and single layer community detection methods with multilayer analysis, such as [41], warrants future work.

We employed multilayer modularity maximization. By design, modularity maximization consists of finding an optimal partition of nodes into non-overlapping communities, and therefore each node belongs to exactly one community. This feature is inherited to multilayer modularity maximization such that each node (i, α), where i represents a gene and α represents a layer, belongs to exactly one community. Multilayer modularity maximization has been used on biological networks to extract groups of proteins or genes that may be functionally related. For example, this technique was used on multilayer networks composed of transcription factor co-targeting, microRNA co-targeting, PPI, and gene co-expression networks as four layers for revealing candidate driver cancer genes [59] and on a multilayer network composed of pathways, co-expression, PPIs, and complexes networks for obtaining groups of disease-related proteins [130]. However, it is not straightforward to interpret the obtained multilayer communities as gene module because, within a single multilayer community, different genes appear in different sets of layers. For example, in a generalist community spanning all the four layers, some genes i may be present in all the layers, whereas other genes j may be present in only one layer. Then, although i and j belong to the same community and connected by group-level co-expression relationships, it may be difficult to argue that i and j share biological functions or environmental factors because how their co-expression depends on layers is different between genes i and j. One option to mitigate this problem is to focus on the resulting gene set in a given multilayer community and ignore the layer identity for simplicity [59, 130]. In contrast, we limited our analysis of generalist communities to the genes that appear in at least three out of the four layers in the community. In this manner, we argued that the genes in the generalist communities used in our localization and biological analyses may have functions common across different tissue types. For the two specialist communities that we analyzed in depth (with $\gamma = 3$), we did not need to select genes because all genes were present in the pancreas and only a small fraction of genes were also present in a different tissue type.

The GTEx Consortium portal provides gene expression data from 30 types of tissues [77]. It is computationally straightforward to extend this analysis to more than four layers (i.e., tissues). Then, however, the results would quickly become much more complicated to interpret. With a number of layers much larger than four, it is likely that our method would no longer discover specialist communities. This is an important limitation of the present analysis. Developing methods more directly tailored to multilayer gene co-expression networks and correlation matrices with a larger number of tissues warrants future work. A suitable method should depend on biological questions. For example, enforcing pillar or semi-pillar communities such that all the genes belonging to the same multilayer community are present in the same set of layers [44, 91] may facilitate biological interpretation of obtained results. Allowing overlapping of communities [131, 132] and genes not belonging to any community may be another choice. For example, overlapping community detection in single-layer networks has been shown to be better at identifying biologically relevant disease modules than non-overlapping community detection [131].

We only analyzed co-expression among $N = 203$ out of the 56, 200 genes because it is difficult to reliably estimate covariance matrices when the number of samples is small [39, 83, 133–135]. Justifiable methods for analyzing co-expression matrices or networks of a larger number of genes are desirable. Such methods will enable us to reduce bias involved in choosing a small subset of genes to analyze. In contrast, a different approach is to formulate the estimation of large correlation networks from big data as a computational challenge and work on efficient algorithms and application to complex biological data [136]. Systematically investigating biological performance of network community detection as a function of the number of samples [135, 137, 138] will help us to better understand potentials and limitations of both single-layer and multilayer community detection in gene and other related networks, which is left as future work.

Supporting information

S1 Text. Fig A. Composition of each community by layer, i.e., tissue, for the multilayer correlation matrix originating from the 150 genes with the highest variance of TPM in each tissue, detected with our community detection method for multilayer correlation matrices with $\gamma = 3$. Although there are 50 communities detected, we only show the communities with more than one gene in this figure. The darker shades indicate nodes corresponding to genes that only appear in one layer in the given community. The lighter shades indicate genes corresponding to genes that appear in multiple layers in the community. **Fig B. Jaccard index between the set of tissue-specific hub genes and the set of genes in a community**. Each row corresponds to the top 50 hub genes in each layer (i.e., tissue), where "panc" denotes pancreas, "sal" denotes salivary gland, "mamm" denotes mammary gland, and "skin" denotes skin (not sun exposed). Each column corresponds to a community identified with $\gamma = 3$. **Table A. Z scores for the number of intralayer edges within each community and for the conductance of each community detected in the unweighted multilayer network obtained by graphical lasso with $\gamma = 1$ and $\gamma = 3$**. Comm. denotes community and no. denotes "number of". **Table B. Z scores for the average distance between pairs of genes on each chromosome and each significant community detected with $\gamma = 1$**. Comm. denotes community and Chr denotes chromosome. **Table C. Z scores for the average distance between pairs of genes on each chromosome and each significant community detected with $\gamma = 3$**. Comm. denotes community and Chr denotes chromosome. **Table D. Results of the gene set enrichment**

analysis for the top 50 highly expressed genes out of the 203 genes in the network in each tissue. Table E. Results of the gene set enrichment analysis for the communities of the multilayer correlation matrix with $\gamma = 3$. Comm. denotes community. **Table F. Results of the gene set enrichment analysis for the top 50 highly connected genes out of the 203 genes in the single-layer network of each tissue. Text A. Analysis of an expanded multilayer correlation matrix. Text B. Significance of communities detected in general multilayer networks. Text C. Derivation of the variance of the total intralayer weight for a community in a multilayer correlation matrix. Text D. Graphical lasso. Text E. Z scores for the average distance between pairs of genes on each chromosome separately in each community. Text F. Results of the gene set enrichment analysis. Text G. Tissue-specific hub genes versus gene communities.**
(PDF)

Author Contributions

Conceptualization: Marie Saitou, Omer Gokcumen, Naoki Masuda.

Data curation: Madison Russell, Alber Aqil, Marie Saitou.

Formal analysis: Madison Russell, Alber Aqil, Omer Gokcumen.

Funding acquisition: Omer Gokcumen, Naoki Masuda.

Investigation: Madison Russell, Alber Aqil, Marie Saitou, Omer Gokcumen, Naoki Masuda.

Methodology: Madison Russell, Alber Aqil, Naoki Masuda.

Project administration: Naoki Masuda.

Resources: Marie Saitou, Omer Gokcumen.

Software: Madison Russell.

Supervision: Naoki Masuda.

Validation: Madison Russell.

Visualization: Madison Russell, Alber Aqil, Omer Gokcumen.

Writing – original draft: Madison Russell, Alber Aqil, Omer Gokcumen, Naoki Masuda.

Writing – review & editing: Madison Russell, Alber Aqil, Marie Saitou, Omer Gokcumen, Naoki Masuda.

References

1. Newman MEJ. Detecting community structure in networks. Eur Phys J B. 2004; 38:321–330. https://doi.org/10.1140/epjb/e2004-00124-y
2. Fortunato S. Community detection in graphs. Phys Rep. 2010; 486(3-5):75–174. https://doi.org/10.1016/j.physrep.2009.11.002
3. Hartwell LH, Hopfield JJ, Leibler S, Murray AW. From molecular to modular cell biology. Nature. 1999; 402(Suppl 6761):C47–C52. https://doi.org/10.1038/35011540 PMID: 10591225
4. Snel B, Bork P, Huynen MA. The identification of functional modules from the genomic association of genes. Proc Natl Acad Sci USA. 2002; 99(9):5890–5895. https://doi.org/10.1073/pnas.092632599 PMID: 11983890
5. Spirin V, Mirny LA. Protein complexes and functional modules in molecular networks. Proc Natl Acad Sci USA. 2003; 100(21):12123–12128. https://doi.org/10.1073/pnas.2032324100 PMID: 14517352
6. Barabási AL, Oltvai ZN. Network biology: understanding the cell's functional organization. Nat Rev Genet. 2004; 5:101–113. https://doi.org/10.1038/nrg1272 PMID: 14735121

7. Barabási AL, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. Nat Rev Genet. 2011; 12:56–68. https://doi.org/10.1038/nrg2918 PMID: 21164525

8. Loscalzo J. Network Medicine. Illustrated ed. Cambridge: Harvard University Press; 2017. https://doi.org/10.4159/9780674545533

9. Butte AJ, Kohane IS. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. Pac Symp Biocomput. 2000; 5:418–429. https://doi.org/10.1142/9789814447331_0040 PMID: 10902190

10. Bar-Joseph Z, Gerber GK, Lee TI, Rinaldi NJ, Yoo JY, Robert F, et al. Computational discovery of gene modules and regulatory networks. Nat Biotechnol. 2003; 21:1337–1342. https://doi.org/10.1038/nbt890 PMID: 14555958

11. Saelens W, Cannoodt R, Saeys Y. A comprehensive evaluation of module detection methods for gene expression data. Nat Commun. 2018; 9:1090. https://doi.org/10.1038/s41467-018-03424-4 PMID: 29545622

12. Kakati T, Bhattacharyya DK, Barah P, Kalita JK. Comparison of methods for differential co-expression analysis for disease biomarker prediction. Comput Biol Med. 2019; 113:103380. https://doi.org/10.1016/j.compbiomed.2019.103380 PMID: 31415946

13. Langfelder P, Horvath S. Eigengene networks for studying the relationships between co-expression modules. BMC Syst Biol. 2007; 1:54. https://doi.org/10.1186/1752-0509-1-54 PMID: 18031580

14. Oldham MC, Konopka G, Iwamoto K, Langfelder P, Kato T, Horvath S, et al. Functional organization of the transcriptome in human brain. Nat Neurosci. 2008; 11:1271–1282. https://doi.org/10.1038/nn.2207 PMID: 18849986

15. Gargalovic PS, Imura M, Zhang B, Gharavi NM, Clark MJ, Pagnon J, et al. Identification of inflammatory gene modules based on variations of human endothelial cell responses to oxidized lipids. Proc Natl Acad Sci USA. 2006; 103(34):12741–12746. https://doi.org/10.1073/pnas.0605457103 PMID: 16912112

16. van Dam S, Võsa U, van der Graaf A, Franke L, de Magalhães JP. Gene co-expression analysis for functional classification and gene–disease predictions. Brief Bioinform. 2018; 19(4):575–592. https://doi.org/10.1093/bib/bbw139 PMID: 28077403

17. Gerring ZF, Gamazon ER, Derks EM, for the Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium. A gene co-expression network-based analysis of multiple brain tissues reveals novel genes and molecular pathways underlying major depression. PLoS Genet. 2019; 15(7): e1008245. https://doi.org/10.1371/journal.pgen.1008245 PMID: 31306407

18. Li N, Zhan X. Identification of clinical trait–related lncRNA and mRNA biomarkers with weighted gene co-expression network analysis as useful tool for personalized medicine in ovarian cancer. EPMA J. 2019; 10:273–290. https://doi.org/10.1007/s13167-019-00175-0 PMID: 31462944

19. Wong DJ, Chang HY. Learning more from microarrays: insights from modules and networks. J Invest Dermatol. 2005; 125(2):175–182. https://doi.org/10.1111/j.0022-202X.2005.23827.x PMID: 16098025

20. Ovens K, Eames BF, McQuillan I. Comparative analyses of gene co-expression networks: Implementations and applications in the study of evolution. Front Genet. 2021; 12:695399. https://doi.org/10.3389/fgene.2021.695399 PMID: 34484293

21. Stuart JM, Segal E, Koller D, Kim SK. A gene-coexpression network for global discovery of conserved genetic modules. Science. 2003; 302(5643):249–255. https://doi.org/10.1126/science.1087447 PMID: 12934013

22. Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. Stat Appl Genet Mol Biol. 2005; 4(17). https://doi.org/10.2202/1544-6115.1128 PMID: 16646834

23. Chowdhury HA, Bhattacharyya DK, Kalita JK. (Differential) co-expression analysis of gene expression: a survey of best practices. IEEE/ACM Trans Comput Biol Bioinform. 2020; 17(4):1154–1173. https://doi.org/10.1109/tcbb.2019.2893170PMID: 30668502

24. Allocco DJ, Kohane IS, Butte AJ. Quantifying the relationship between co-expression, co-regulation and gene function. BMC Bioinformatics. 2004; 5:18. https://doi.org/10.1186/1471-2105-5-18 PMID: 15053845

25. Ribeiro DM, Rubinacci S, Ramisch A, Hofmeister RJ, Dermitzakis ET, Delaneau O. The molecular basis, genetic control and pleiotropic effects of local gene co-expression. Nat Commun. 2021; 12:4842. https://doi.org/10.1038/s41467-021-25129-x PMID: 34376650

26. Ebisuya M, Yamamoto T, Nakajima M, Nishida E. Ripples from neighbouring transcription. Nat Cell Biol. 2008; 10:1106–1113. https://doi.org/10.1038/ncb1771 PMID: 19160492

27. Hurst LD, Pál C, Lercher MJ. The evolutionary dynamics of eukaryotic gene order. Nat Rev Genet. 2004; 5:299–310. https://doi.org/10.1038/nrg1319 PMID: 15131653

28. Sproul D, Gilbert N, Bickmore WA. The role of chromatin structure in regulating the expression of clustered genes. Nat Rev Genet. 2005; 6:775–781. https://doi.org/10.1038/nrg1688 PMID: 16160692

29. Perry BW, Gopalan SS, Pasquesi GIM, Schield DR, Westfall AK, Smith CF, et al. Snake venom gene expression is coordinated by novel regulatory architecture and the integration of multiple co-opted vertebrate pathways. Genome Res. 2022; 32:1058–1073. https://doi.org/10.1101/gr.276251.121 PMID: 35649579

30. Holter NS, Mitra M, Maritan A, Cieplak M, Banavar JR, Fedoroff NV. Fundamental patterns underlying gene expression profiles: Simplicity from complexity. Proc Natl Acad Sci USA. 2000; 97(15):8409–8414. https://doi.org/10.1073/pnas.150242097 PMID: 10890920

31. Carter SL, Brechbühler CM, Griffin M, Bond AT. Gene co-expression network topology provides a framework for molecular characterization of cellular state. Bioinformatics. 2004; 20(14):2242–2250. https://doi.org/10.1093/bioinformatics/bth234 PMID: 15130938

32. Barah P, Naika BN M, Jayavelu ND, Sowdhamini R, Shameer K, Bones AM. Transcriptional regulatory networks in *Arabidopsis thaliana* during single and combined stresses. Nucleic Acids Res. 2016; 44 (7):3147–3164. https://doi.org/10.1093/nar/gkv1463 PMID: 26681689

33. Gaiteri C, Ding Y, French B, Tseng GC, Sibille E. Beyond modules and hubs: the potential of gene coexpression networks for investigating molecular mechanisms of complex brain disorders. Genes Brain Behav. 2014; 13(1):13–24. https://doi.org/10.1111/gbb.12106 PMID: 24320616

34. Leek JT, Storey JD. Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genet. 2007; 3(9):e161. https://doi.org/10.1371/journal.pgen.0030161 PMID: 17907809

35. Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, et al. Tackling the widespread and critical impact of batch effects in high-throughput data. Nat Rev Genet. 2010; 11:733–739. https://doi.org/10.1038/nrg2825 PMID: 20838408

36. Nguyen ND, Blaby IK, Wang D. ManiNetCluster: a novel manifold learning approach to reveal the functional links between gene networks. BMC Genomics. 2019; 20:1003. https://doi.org/10.1186/s12864-019-6329-2 PMID: 31888454

37. Rhinn H, Qiang L, Yamashita T, Rhee D, Zolin A, Vanti W, et al. Alternative α-synuclein transcript usage as a convergent mechanism in Parkinson's disease pathology. Nat Commun. 2012; 3:1084. https://doi.org/10.1038/ncomms2032 PMID: 23011138

38. Piro RM, Ala U, Molineris I, Grassi E, Bracco C, Perego GP, et al. An atlas of tissue-specific conserved coexpression for functional annotation and disease gene prediction. Eur J Hum Genet. 2011; 19:1173–1180. https://doi.org/10.1038/ejhg.2011.96 PMID: 21654723

39. Pierson E, Consortium G, Koller D, Battle A, Mostafavi S. Sharing and specificity of co-expression networks across 35 human tissues. PLoS Comput Biol. 2015; 11(5):e1004220. https://doi.org/10.1371/journal.pcbi.1004220 PMID: 25970446

40. Saitou M, Gaylord EA, Xu E, May AJ, Neznanova L, Nathan S, et al. Functional specialization of human salivary glands and origins of proteins intrinsic to human saliva. Cell Rep. 2020; 33(7):108402. https://doi.org/10.1016/j.celrep.2020.108402 PMID: 33207190

41. Azevedo T, Dimitri GM, Lió P, Gamazon ER. Multilayer modelling of the human transcriptome and biological mechanisms of complex diseases and traits. NPJ Syst Biol Appl. 2021; 7:24. https://doi.org/10.1038/s41540-021-00186-6 PMID: 34045472

42. Ritchie SC, Watts S, Fearnley LG, Holt KE, Abraham G, Inouye M. A scalable permutation approach reveals replication and preservation patterns of network modules in large datasets. Cell Syst. 2016; 3 (1):71–82. https://doi.org/10.1016/j.cels.2016.06.012 PMID: 27467248

43. Mucha PJ, Richardson T, Macon K, Porter MA, Onnela JP. Community structure in time-dependent, multiscale, and multiplex networks. Science. 2010; 328(5980):876–878. https://doi.org/10.1126/science.1184819 PMID: 20466926

44. Magnani M, Hanteer O, Interdonato R, Rossi L, Tagarelli A. Community detection in multiplex networks. ACM Comput Surv. 2021; 54(3):48. https://doi.org/10.1145/3444688

45. Li W, Liu CC, Zhang T, Li H, Waterman MS, Zhou XJ. Integrative analysis of many weighted co-expression networks using tensor computation. PLoS Comput Biol. 2011; 7(6):e1001106. https://doi.org/10.1371/journal.pcbi.1001106 PMID: 21698123

46. Haas R, Zelezniak A, Iacovacci J, Kamrad S, Townsend S, Ralser M. Designing and interpreting 'multi-omic' experiments that may change our understanding of biology. Curr Opin Syst Biol. 2017; 6:37–45. https://doi.org/10.1016/j.coisb.2017.08.009 PMID: 32923746

47. Gosak M, Markovič R, Dolenšek J, Rupnik MS, Marhl M, Stožer A, et al. Network science of biological systems at different scales: A review. Phys Life Rev. 2018; 24:118–135. https://doi.org/10.1016/j.plrev.2017.11.003 PMID: 29150402

48. Hammoud Z, Kramer F. Multilayer networks: aspects, implementations, and application in biomedicine. Big Data Anal. 2020; 5:2. https://doi.org/10.1186/s41044-020-00046-0

49. Dorantes-Gilardi R, García-Cortés D, Hernández-Lemus E, Espinal-Enríquez J. Multilayer approach reveals organizational principles disrupted in breast cancer co-expression networks. Appl Netw Sci. 2020; 5:47. https://doi.org/10.1007/s41109-020-00291-1

50. Ma X, Zhao W, Wu W. Layer-specific modules detection in cancer multi-layer networks. IEEE/ACM Trans Comput Biol Bioinform. 2023; 20(2):1170–1179. https://doi.org/10.1109/TCBB.2022.3176859 PMID: 35609099

51. Lei J, Lin KZ. Bias-adjusted spectral clustering in multi-layer stochastic block models. J Am Stat Assoc. 2022; p. 1–13. https://doi.org/10.1080/01621459.2022.2054817

52. Zhang J, Li C, Wang J. A stochastic block Ising model for multi-layer networks with inter-layer dependence. Biometrics. 2023; p. 1–10. https://doi.org/10.1111/biom.13885 PMID: 37284764

53. Xiao X, Moreno-Moral A, Rotival M, Bottolo L, Petretto E. Multi-tissue analysis of co-expression networks by higher-order generalized singular value decomposition identifies functionally coherent transcriptional modules. PLoS Genet. 2014; 10(1):e1004006. https://doi.org/10.1371/journal.pgen.1004006 PMID: 24391511

54. Yu L, Yao S, Gao L, Zha Y. Conserved disease modules extracted from multilayer heterogeneous disease and gene networks for understanding disease mechanisms and predicting disease treatments. Front Genet. 2019; 9:745. https://doi.org/10.3389/fgene.2018.00745 PMID: 30713550

55. Yu L, Shi Y, Zou Q, Wang S, Zheng L, Gao L. Exploring drug treatment patterns based on the action of drug and multilayer network model. Int J Mol Sci. 2020; 21(14):5014. https://doi.org/10.3390/ijms21145014 PMID: 32708644

56. Halu A, De Domenico M, Arenas A, Sharma A. The multiplex network of human diseases. NPJ Syst Biol Appl. 2019; 5:15. https://doi.org/10.1038/s41540-019-0092-5 PMID: 31044086

57. Saitou M, Masuda N, Gokcumen O. Similarity-based analysis of allele frequency distribution among multiple populations identifies adaptive genomic structural variants. Mol Biol Evol. 2022; 39(3): msab313. https://doi.org/10.1093/molbev/msab313 PMID: 34718708

58. Li D, Pan Z, Hu G, Anderson G, He S. Active module identification from multilayer weighted gene co-expression networks: a continuous optimization approach. IEEE/ACM Trans Comput Biol Bioinform. 2021; 18(6):2239–2248. https://doi.org/10.1109/TCBB.2020.2970400 PMID: 32011261

59. Cantini L, Medico E, Fortunato S, Caselle M. Detection of gene communities in multi-networks reveals cancer drivers. Sci Rep. 2015; 5:17386. https://doi.org/10.1038/srep17386 PMID: 26639632

60. Horvath S, Dong J. Geometric interpretation of gene coexpression network analysis. PLoS Comput Biol. 2008; 4(8):e1000117. https://doi.org/10.1371/journal.pcbi.1000117 PMID: 18704157

61. Barzel B, Barabási AL. Network link prediction by global silencing of indirect correlations. Nat Biotechnol. 2013; 31:720–725. https://doi.org/10.1038/nbt.2601 PMID: 23851447

62. Fiecas M, Ombao H, van Lunen D, Baumgartner R, Coimbra A, Feng D. Quantifying temporal correlations: a test–retest evaluation of functional connectivity in resting-state fMRI. NeuroImage. 2013; 65:231–241. https://doi.org/10.1016/j.neuroimage.2012.09.052 PMID: 23032492

63. Rubinov M, Sporns O. Weight-conserving characterization of complex functional brain networks. NeuroImage. 2011; 56(4):2068–2079. https://doi.org/10.1016/j.neuroimage.2011.03.069 PMID: 21459148

64. Garrison KA, Scheinost D, Finn ES, Shen X, Constable RT. The (in)stability of functional brain network measures across thresholds. NeuroImage. 2015; 118:651–661. https://doi.org/10.1016/j.neuroimage.2015.05.046 PMID: 26021218

65. De Vico Fallani F, Latora V, Chavez M. A topological criterion for filtering information in complex brain networks. PLoS Comput Biol. 2017; 13(1):e1005305. https://doi.org/10.1371/journal.pcbi.1005305 PMID: 28076353

66. Meinshausen N, Bühlmann P. High-dimensional graphs and variable selection with the Lasso. Ann Stat. 2006; 34(3):1436–1462. https://doi.org/10.1214/009053606000000281

67. Yuan M, Lin Y. Model selection and estimation in the Gaussian graphical model. Biometrika. 2007; 94 (1):19–35. https://doi.org/10.1093/biomet/asm018

68. Friedman J, Hastie T, Tibshirani R. Sparse inverse covariance estimation with the graphical lasso. Biostatistics. 2008; 9(3):432–441. https://doi.org/10.1093/biostatistics/kxm045 PMID: 18079126

69. Schäfer J, Strimmer K. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Stat Appl Genet Mol Biol. 2005; 4(1):1–32. https://doi.org/10.2202/1544-6115.1175 PMID: 16646851

70. Bien J, Tibshirani RJ. Sparse estimation of a covariance matrix. Biometrika. 2011; 98(4):807–820. https://doi.org/10.1093/biomet/asr054 PMID: 23049130

71. Kojaku S, Masuda N. Constructing networks by filtering correlation matrices: A null model approach. Proc R Soc A. 2019; 475(2231):20190578. https://doi.org/10.1098/rspa.2019.0578 PMID: 31824228

72. MacMahon M, Garlaschelli D. Community detection for correlation matrices. Phys Rev X. 2015; 5 (2):021006. https://doi.org/10.1103/physrevx.5.021006

73. Bazzi M, Porter MA, Williams S, McDonald M, Fenn DJ, Howison SD. Community detection in temporal multilayer networks, with an application to correlation networks. Multiscale Model Simul. 2016; 14 (1):1–41. https://doi.org/10.1137/15M1009615

74. Masuda N, Kojaku S, Sano Y. Configuration model for correlation matrices preserving the node strength. Phys Rev E. 2018; 98(1):012312. https://doi.org/10.1103/PhysRevE.98.012312 PMID: 30110768

75. Masuda N, Sakaki M, Ezaki T, Watanabe T. Clustering coefficients for correlation networks. Front Neuroinform. 2018; 12:7. https://doi.org/10.3389/fninf.2018.00007 PMID: 29599714

76. GitHub repository for multilayer community detection with covariance matrix input. [cited 2023 May 10]. Available from: https://github.com/russell-madison/corr_comm_detection.

77. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, et al. The Genotype-Tissue Expression (GTEx) project. Nat Genet. 2013; 45:580–585. https://doi.org/10.1038/ng.2653

78. Brannon PM. Adaptation of the exocrine pancreas to diet. Annu Rev Nutr. 1990; 10:85–105. https://doi.org/10.1146/annurev.nu.10.070190.000505 PMID: 2200477

79. Thamadilok S, Choi KS, Ruhl L, Schulte F, Kazim AL, Hardt M, et al. Human and nonhuman primate lineage-specific footprints in the salivary proteome. Mol Biol Evol. 2020; 37(2):395–405. https://doi.org/10.1093/molbev/msz223 PMID: 31614365

80. McClellan HL, Miller SJ, Hartmann PE. Evolution of lactation: nutrition *v.* protection with special reference to five mammalian species. Nutr Res Rev. 2008; 21(2):97–116. https://doi.org/10.1017/S0954422408100749 PMID: 19087365

81. Quillen EE, Norton HL, Parra EJ, Lona-Durazo F, Ang KC, Illiescu FM, et al. Shades of complexity: New perspectives on the evolution and genetic architecture of human skin. Am J Phys Anthropol. 2019; 168(S67):4–26. https://doi.org/10.1002/ajpa.23737 PMID: 30408154

82. Starr I, Seiffert-Sinha K, Sinha AA, Gokcumen O. Evolutionary context of psoriatic immune skin response. Evol Med and Public Health. 2021; 9(1):474–486. https://doi.org/10.1093/emph/eoab042 PMID: 35154781

83. Lyu Y, Xue L, Zhang F, Koch H, Saba L, Kechris K, et al. Condition-adaptive fused graphical lasso (CFGL): An adaptive procedure for inferring condition-specific gene co-expression network. PLoS Comput Biol. 2018; 14(9):e1006436. https://doi.org/10.1371/journal.pcbi.1006436 PMID: 30240439

84. Bun J, Bouchaud JP, Potters M. Cleaning large correlation matrices: Tools from Random Matrix Theory. Phys Rep. 2017; 666:1–109. https://doi.org/10.1016/j.physrep.2016.10.005

85. Jaccard P. The distribution of the flora in the alpine zone. New Phytol. 1912; 11(2):37–50. https://doi.org/10.1111/j.1469-8137.1912.tb05611.x

86. Robinson MD, Smyth GK. Moderated statistical tests for assessing differences in tag abundance. Bioinformatics. 2007; 23(21):2881–2887. https://doi.org/10.1093/bioinformatics/btm453 PMID: 17881408

87. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010; 11: R106. https://doi.org/10.1186/gb-2010-11-10-r106 PMID: 20979621

88. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, et al. A survey of best practices for RNA-seq data analysis. Genome Biol. 2016; 17:13. https://doi.org/10.1186/s13059-016-0881-8 PMID: 26813401

89. Kivelä M, Arenas A, Barthélemy M, Gleeson JP, Moreno Y, Porter MA. Multilayer networks. J Complex Netw. 2014; 2(3):203–271. https://doi.org/10.1093/comnet/cnu016

90. Bianconi G. Multilayer Networks: Structure and Function. First ed. New York: Oxford University Press; 2018. https://doi.org/10.1093/oso/9780198753919.001.0001

91. Hanteer O, Magnani M. Unspoken assumptions in multi-layer modularity maximization. Sci Rep. 2020; 10:11053. https://doi.org/10.1038/s41598-020-66956-0 PMID: 32632217

92. Zhang H, Wang CD, Lai JH, Yu PS. Modularity in complex multilayer networks with multiple aspects: a static perspective. Appl Inform. 2017; 4:7. https://doi.org/10.1186/s40535-017-0035-4

93. Newman MEJ, Girvan M. Finding and evaluating community structure in networks. Phys Rev E. 2004; 69(2):026113. https://doi.org/10.1103/PhysRevE.69.026113 PMID: 14995526

94. Newman MEJ. Analysis of weighted networks. Phys Rev E. 2004; 70(5):056131. https://doi.org/10.1103/PhysRevE.70.056131 PMID: 15600716

95. Fortunato S, Barthélemy M. Resolution limit in community detection. Proc Natl Acad Sci USA. 2007; 104(1):36–41. https://doi.org/10.1073/pnas.0605965104 PMID: 17190818

96. Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. J Stat Mech. 2008; 2008:P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008

97. Jeub LGS, Bazzi M, Jutla IS, Mucha PJ. A generalized Louvain method for community detection implemented in MATLAB. Version 2.2 [software]. 2011-2019 [downloaded 2022 May 16]. Available from: https://github.com/GenLouvain/GenLouvain.

98. Good BH, de Montjoye YA, Clauset A. Performance of modularity maximization in practical contexts. Phys Rev E. 2010; 81(4):046106. https://doi.org/10.1103/PhysRevE.81.046106 PMID: 20481785

99. Strehl A, Ghosh J. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. J Mach Learn Res. 2002; 3:583–617.

100. Bassett DS, Porter MA, Wymbs NF, Grafton ST, Carlson JM, Mucha PJ. Robust detection of dynamic community structure in networks. Chaos. 2013; 23(1):013142. https://doi.org/10.1063/1.4790830

101. netneurotools package. Version 0.2.3 [software]. 2021 Aug 31 [downloaded 2022 Aug 3]. Available from: https://github.com/netneurolab/netneurotools.

102. Hirschberger M, Qi Y, Steuer RE. Randomly generating portfolio-selection covariance matrices with specified distributional characteristics. Eur J Oper Res. 2007; 177(3):1610–1625. https://doi.org/10.1016/j.ejor.2005.10.014

103. configcorr package. [software]. 2018 Sep 27 [downloaded 2022 May 18]. Available from: https://github.com/naokimas/config_corr.

104. Lancichinetti A, Radicchi F, Ramasco JJ. Statistical significance of communities in networks. Phys Rev E. 2010; 81(4):046110. https://doi.org/10.1103/PhysRevE.78.046110

105. Yang J, Leskovec J. Defining and evaluating network communities based on ground-truth. Knowl Inf Syst. 2015; 42:181–213. https://doi.org/10.1007/s10115-013-0693-z

106. Kojaku S, Masuda N. A generalised significance test for individual communities in networks. Sci Rep. 2018; 8:7351. https://doi.org/10.1038/s41598-018-25560-z PMID: 29743534

107. Weir WH, Emmons S, Gibson R, Taylor D, Mucha PJ. Post-processing partitions to identify domains of modularity optimization. Algorithms. 2017; 10(3):93. https://doi.org/10.3390/a10030093 PMID: 29046743

108. Weir WH, Gibson R, Mucha PJ. CHAMP package: Convex Hull of Admissible Modularity Partitions in Python and MATLAB. Version 2.1.0 [software]. 2019 May 22 [downloaded 2022 May 30]. Available from: https://github.com/wweir827/CHAMP.

109. Raudvere U, Kolberg L, Kuzmin I, Arak T, Adler P, Peterson H, et al. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). Nucleic Acids Res. 2019; 47 (W1):W191–W198. https://doi.org/10.1093/nar/gkz369 PMID: 31066453

110. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. Nat Genet. 2000; 25:25–29. https://doi.org/10.1038/75556 PMID: 10802651

111. The Gene Ontology Consortium, Aleksander SA, Balhoff J, Carbon S, Cherry JM, Drabkin HJ, et al. The Gene Ontology knowledgebase in 2023. Genetics. 2023; 224(1):iyad031. https://doi.org/10.1093/genetics/iyad031 PMID: 36866529

112. Köhler S, Gargano M, Matentzoglu N, Carmody LC, Lewis-Smith D, Vasilevsky NA, et al. The Human Phenotype Ontology in 2021. Nucleic Acids Res. 2021; 49(D1):D1207–D1217. https://doi.org/10.1093/nar/gkaa1043 PMID: 33264411

113. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Series B Stat Methodol. 1995; 57(1):289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

114. Dunn OJ. Multiple comparisons among means. J Am Stat Assoc. 1961; 56(293):52–64. https://doi.org/10.1080/01621459.1961.10482090

115. Squartini T, Garlaschelli D. Analytical maximum-likelihood method to detect patterns in real networks. New J Phys. 2011; 13:083001. https://doi.org/10.1088/1367-2630/13/8/083001

116. Mastrandrea R, Squartini T, Fagiolo G, Garlaschelli D. Enhanced reconstruction of weighted networks from strengths and degrees. New J Phys. 2014; 16:043022. https://doi.org/10.1088/1367-2630/16/4/043022

117. Eaaswarkhanth M, Pavlidis P, Gokcumen O. Geographic distribution and adaptive significance of genomic structural variants: an anthropological genetics perspective. Hum Biol. 2014; 86(4):260–275. https://doi.org/10.13110/humanbiology.86.4.0260 PMID: 25959693

118. Ho M, Thompson B, Fisk JN, Nebert DW, Bruford EA, Vasiliou V, et al. Update of the keratin gene family: evolution, tissue-specific expression patterns, and relevance to clinical disorders. Hum Genomics. 2022; 16:1. https://doi.org/10.1186/s40246-021-00374-9 PMID: 34991727

119. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012; 489:57–74. https://doi.org/10.1038/nature11247

120. Fishilevich S, Nudel R, Rappaport N, Hadar R, Plaschkes I, Iny Stein T, et al. GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. Database. 2017; 2017:bax028. https://doi.org/10.1093/database/bax028 PMID: 28605766

121. Boros E, Szabó A, Zboray K, Héja D, Pál G, Sahin-Tóth M. Overlapping specificity of duplicated human pancreatic elastase 3 isoforms and archetypal porcine elastase 1 provides clues to evolution of digestive enzymes. J Biol Chem. 2017; 292(7):P2690–2702. https://doi.org/10.1074/jbc.M116.770560 PMID: 28062577

122. Takeuchi F, Yanai K, Morii T, Ishinaga Y, Taniguchi-Yanai K, Nagano S, et al. Linkage disequilibrium grouping of single nucleotide polymorphisms (SNPs) reflecting haplotype phylogeny for efficient selection of tag SNPs. Genetics. 2005; 170(1):291–304. https://doi.org/10.1534/genetics.104.038232 PMID: 15716494

123. UK BioBank data. [cited 2023 Aug 23]. Available from: http://www.nealelab.is/uk-biobank.

124. Aqil A, Speidel L, Pavlidis P, Gokcumen O. Balancing selection on genomic deletion polymorphisms in humans. eLife. 2023; 12:e79111. https://doi.org/10.7554/eLife.79111 PMID: 36625544

125. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biol. 2016; 17:122. https://doi.org/10.1186/s13059-016-0974-4 PMID: 27268795

126. Fusco G, Minelli A. Phenotypic plasticity in development and evolution: facts and concepts. Phil Trans R Soc B. 2010; 365(1540):547–556. https://doi.org/10.1098/rstb.2009.0267 PMID: 20083631

127. Gibert JM, Mouchel-Vielh E, De Castro S, Peronnet F. Phenotypic plasticity through transcriptional regulation of the evolutionary hotspot gene *tan* in *Drosophila melanogaster*. PLoS Genet. 2016; 12(8): e1006218. https://doi.org/10.1371/journal.pgen.1006218 PMID: 27508387

128. Gómez S, Jensen P, Arenas A. Analysis of community structure in networks of correlated data. Phys Rev E. 2009; 80(1):016114. https://doi.org/10.1103/PhysRevE.80.016114 PMID: 19658781

129. Hristova D, Rutherford A, Anson J, Luengo-Oroz M, Mascolo C. The international postal network and other global flows as proxies for national wellbeing. PloS ONE. 2016; 11(6):e0155976. https://doi.org/10.1371/journal.pone.0155976 PMID: 27248142

130. Didier G, Brun C, Baudot A. Identifying communities from multiplex biological networks. PeerJ. 2015; 3:e1525. https://doi.org/10.7717/peerj.1525 PMID: 26713261

131. Tripathi B, Parthasarathy S, Sinha H, Raman K, Ravindran B. Adapting community detection algorithms for disease module identification in heterogeneous biological networks. Front Genet. 2019; 10:164. https://doi.org/10.3389/fgene.2019.00164 PMID: 30918511

132. Riccio-Rengifo C, Finke J, Rocha C. Identifying stress responsive genes using overlapping communities in co-expression networks. BMC Bioinformatics. 2021; 22:541. https://doi.org/10.1186/s12859-021-04462-4 PMID: 34743699

133. Dempster AP. Covariance selection. Biometrics. 1972; 28(1):157–175. https://doi.org/10.2307/2528966

134. Johnstone IM. On the distribution of the largest eigenvalue in principal components analysis. Ann Stat. 2001; 29(2):295–327. https://doi.org/10.1214/aos/1009210544

135. Ballouz S, Verleyen W, Gillis J. Guidance for RNA-seq co-expression network construction and analysis: safety in numbers. Bioinformatics. 2015; 31(13):2123–2130. https://doi.org/10.1093/bioinformatics/btv118 PMID: 25717192

136. Becker M, Nassar H, Espinosa C, Stelzer IA, Feyaerts D, Berson E, et al. Large-scale correlation network construction for unraveling the coordination of complex biological systems. Nat Comput Sci. 2023; 3:346–359. https://doi.org/10.1038/s43588-023-00429-y

137. Guo W, Calixto CPG, Tzioutziou N, Lin P, Waugh R, Brown JWS, et al. Evaluation and improvement of the regulatory inference for large co-expression networks with limited sample size. BMC Syst Biol. 2017; 11:62. https://doi.org/10.1186/s12918-017-0440-2 PMID: 28629365

138. Ovens KL, Eames BF, McQuillan I. The impact of sample size and tissue type on the reproducibility of gene co-expression networks. In: Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics; 2020;1–10. https://doi.org/10.1145/3388440.3412481