Multilayer modelling and analysis of the human transcriptome

Tiago Azevedo[1,†], Giovanna Maria Dimitri[1,2,3,†], Pietro Lio[1,2,*], Eric R. Gamazon[2,4,5,*]

[1]Department of Computer Science and Technology, University of Cambridge, Cambridge, UK
[2]Clare Hall, University of Cambridge
[3]Department of Engineering, University of Siena, Italy
[4]Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, TN, USA
[5]Data Science Institute, Vanderbilt University, Nashville, TN, USA
[†] These authors contributed equally to this work.
[*]Correspondences should be addressed to: pl219@cam.ac.uk, ericgamazon@gmail.com.

In the present work, we performed a comprehensive intra-tissue and inter-tissue network analysis of the human transcriptome. We generated an atlas of communities in co-expression networks in each of 49 tissues and evaluated their tissue specificity. UMAP embeddings of gene expression from the identified communities (representing nearly 18% of all genes) recovered biologically meaningful tissue clusters, based on tissue organ membership or known shared function. We developed an approach to quantify the conservation of global structure and estimate the sampling distribution of the distance between tissue clusters via bootstrapped manifolds. We found not only preserved local structure among clearly related tissues (e.g., the 13 brain regions) but also a strong correlation between the clustering of these related tissues relative to the remaining ones. Interestingly, brain tissues showed significantly higher variability in community size than non-brain ($p = 1.55 \times 10^{-4}$). We

identified communities that capture some of our current knowledge about biological processes, but most are likely to encode novel and previously inaccessible functional information. For example, we found a 17-member community present across all of the brain regions, which shows significant enrichment for the nonsense-mediated decay pathway (adjusted $p = 1.01 \times 10^{-37}$). We also constructed multiplex architectures to gain insights into tissue-to-tissue mechanisms for regulation of communities in the transcriptome, including communities that are likely to play a functional role throughout the central nervous system (CNS) and communities that may participate in the interaction between the CNS and the enteric nervous system. Notably, new gene expression data can be embedded into our models to accelerate discoveries in high-dimensional molecular datasets. Our study provides a rich resource of co-expression networks, communities, multiplex architectures, and enriched pathways in a broad collection of tissues, to catalyse research into inter-tissue regulatory mechanisms and enable insights into their downstream phenotypic consequences.

Introduction

The modern science of networks has contributed to notable advances in a range of disciplines, facilitating complex representations of biological, social, and technological systems (*1*). A key aspect of such systems is the existence of *community structures*, wherein groups of nodes are organised into dense internal connections with sparser connections between groups. Community structure detection in genome-wide gene expression data may enable detection of regulatory relationships between regulators (e.g., transcription factors or microRNAs) (*2*) and their targets and capture novel tissue biology otherwise difficult to reach. Furthermore, it offers opportuni-

ties for data-driven discovery and functional annotation of biological pathways (*3*).

We hypothesise that community structure is an important organising principle of the human transcriptome, with critical implications for biological discovery and clinical application. Co-expression networks, in fact, encode functionally relevant relationships between genes, including gene interactions and coordinated transcriptional regulation, and provide an approach to elucidating the molecular basis of disease traits. Therefore, reconstructing communities of genes in the transcriptome may uncover novel relationships between genes, facilitate insights into regulatory processes, and improve the mapping of the human diseasome.

In this work, we develop a model of the human transcriptome as a multilayer network, and we perform a comprehensive analysis of the communities obtained with this modelling in order to further our understanding of its wiring diagram. We conduct a systematic analysis of the tissue or cell-type specificity of the communities in the transcriptome in order to gain insights into gene function in the genome and improve our ability to identify disease-associated genes. Our study represents an effort to fill an important gap in our understanding of the role of gene expression in complex traits, i.e., how a gene's phenotypic consequence on disease or trait (*4*) is mediated by its membership in tissue-specific biological modules as molecular substrates. Finally, the inter-tissue analysis of the transcriptome that we present holds promise for identifying novel regulatory mechanisms and enhancing our understanding of trait variation and pleiotropy (*5*).

Methods

GTEx Dataset. The GTEx V8 dataset (*6–8*) is a genomic resource consisting of 948 donors and 17,382 RNA-Seq samples from 52 tissues and two cell lines. The resource provides a catalog of genetic effects on the transcriptome and a broad survey of individual- and tissue-specific gene expression. Of the 54 tissues and cell lines, 49 include samples with at least 70

subjects, forming the basis of the analysis of genetic regulatory effects (*8*). In this study, we leveraged the 49 tissues because of their sample size (see Table S1) and our interest in shared transcriptional regulatory programs for co-expressed genes (*9*).

Data Preprocessing. We restricted our analyses to protein-encoding genes based on the GENCODE (*10*) annotation. Although the GTEx dataset had annotated genes with ENSEMBLE IDs, we removed duplicates (using GENE IDs) and unmapped genes from downstream analyses. After this preprocessing step, the resulting dataset is characterised by the following count statistics:

- Unique genes across all tissues: 18,364
- Genes present in only 1 tissue: 412
- Genes present in all 49 tissues: 12,557

Accounting for Unmodelled Factors. In order to correct for batch effects and other unwanted variation in the gene expression data, we used the *sva* R package, which is specifically targeted for identifying surrogate variables in high-dimensional data sets (*11*). For each tissue gene expression matrix, the number of components (latent factors) was estimated using a permutation procedure, as described in (*12*).

Subsequently, using the function *sva_network*, residuals were generated after regressing out the surrogate variables. The residual values, rather than the original gene expression values, were used in the downstream analyses. For convenience, we refer to the residual values as the 'gene expression data', since they represent the expression levels that have been corrected for (unwanted) confounders.

Tissue-dependent Correlation and Adjacency Matrices. For each tissue, a correlation matrix $C = [z_{ij}]$ was created by calculating the Pearson correlation coefficient r_{ij} for every pair (i, j) of genes. Fisher z-transformation was then applied:

$$z_{ij} = 0.5 \times \ln\left(\frac{1 + r_{ij}}{1 - r_{ij}}\right) \tag{1}$$

where ln is the natural logarithm function.

For each correlation matrix, we retained only the strongest correlations (i.e., transformed z_{ij} less than -0.8 and greater than 0.8) to generate a co-expression network. An adjacency matrix $A = [A_{ij}]$ was defined, for each tissue, such that A_{ij} is equal to z_{ij} if gene i and gene j are co-expressed (retained), and zero otherwise.

Community Detection. We sought to detect groups of genes in each tissue, with the aim of finding communities whose internal connections are denser than the connections with the rest of the co-expression network (*13*). We applied the Louvain community detection method (*14*) in each tissue to generate a comprehensive atlas of communities. An asymmetric treatment for the negative correlations was used, thus inducing negatively correlated genes to belong to different communities (*15*). The algorithm identifies communities by maximising the modularity index (*16*), Q^*, as the algorithm progresses:

$$Q^* = \frac{1}{v^+} \sum\nolimits_{ij} \left({w_{ij}}^+ - {e_{ij}}^+\right) \delta_{M_i M_j} - \frac{1}{v^+ + v^-} \sum\nolimits_{ij} \left({w_{ij}}^- - {e_{ij}}^-\right) \delta_{M_i M_j} \tag{2}$$

Here a positive connection between nodes i and j is denoted as ${w_{ij}}^+$ and has a value between 0 and 1; likewise, a negative connection is represented ${w_{ij}}^-$ and can also have a value between 0 and 1. ${e_{ij}}^\pm$ is the chance-expected within-module connection weight and calculated, for each positive/negative correspondent, as $\frac{s_i^\pm s_j^\pm}{v^\pm}$, where $s_i^\pm$ is the sum of positive or negative connection weights of node i. $v^\pm$ is the sum of all positive or negative edges, and $\delta_{M_i M_j} = 1$ when nodes i and j are in the same module or zero otherwise. In particular, the Louvain method iteratively

evaluates the gain in modularity if one node is moved from one formed community to another of its neighborhood. We leveraged the Brain Connectivity Toolbox Python package (available on github: aestrivex/bctpy). The resolution parameter γ was set to its default value, 1.

UMAP Embeddings of Community-defined Gene Expression. To produce a lower dimensional representation of the original dataset, we applied to it Uniform Manifold Approximation and Projection (UMAP) (*17*), a manifold learning technique. Our goal was to generate a map that reveals embedded structures and test whether biologically relevant clusters can be recovered from the gene expression data. Towards this end, we analysed both the full master matrix $\mathbf{M}$ of scaled gene expression (in the range $[0, 1]$), consisting of all genes (i.e., $18,364$), and a submatrix consisting of only those genes that belong to a community in at least one tissue (i.e., $3,259$). (Similarly to all of the results in the rest of the paper, we considered only Louvain communities with at least 4 genes.)

We chose UMAP because of the substantial improvement in running time on our data (compared to t-SNE, with its known computational and memory complexity that is quadratic in the sample size (*18*)), and UMAP's theoretical grounding in manifold theory (*17*). UMAP can also capture non-linear effects in gene expression, and this was another reason why we chose it, over more traditional dimensionality reduction techniques such as Principal Component Analysis (*19*). Additional implementation details can be found in Figure S1.

Persistence of the UMAP Global Structure. We quantified the conservation of, and variability in, the UMAP structure, including the relation among biologically meaningful clusters, e.g., tissues. We characterised such a structure using the matrix $[\mathrm{d}(i, j)]$ of pairwise distances for clusters i and j in $\{1, 2, \ldots, L\}$. For the actual (original) gene expression data, we define $\mathbf{V}_{(\mathbf{0})} = [v_{ij,0}]$ as the resulting matrix of pairwise distances. Note $\mathbf{V}_{(\mathbf{0})}$ is a symmetric matrix with zeros along the diagonal. We sought to:

- estimate the sampling distribution of $\mathrm{d}(i,j)$, and calculate its standard error and a confidence interval

- correlate the matrix $\mathbf{V}_{(\mathbf{0})}$ and the resulting matrix W from a perturbation of the original structure

We approached the quantification problem through a (non-parametric) bootstrapping procedure. From the master matrix $\mathbf{M}$ of gene expression, we generated a total of B bootstrapped manifolds, each of equal size (here, each such sample was randomly drawn from 80% of the data points, i.e., rows, in $\mathbf{M}$). For the *k*-th sample, we constructed the matrix $\mathbf{V}_{(\mathbf{k})} = [\widehat{\mathrm{d}(i,j)_{(k)}}]$ of pairwise distances derived from the UMAP embeddings for tissues i and j. Here we used the "induced metric" $\mathrm{d} : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}$ from the embedding $\phi : M_g \hookrightarrow \mathbb{R}^m$ of the Riemannian manifold M_g into Euclidean space, but our treatment here generalises to the intrinsic metric $\mathrm{g} : M_g \times M_g \to \mathbb{R}$, with $\mathrm{g}(\phi^{-1}(i), \phi^{-1}(j))$, of the original manifold. The set $\{\widehat{\mathrm{d}(i,j)_{(k)}}\}_{k=1}^{B}$ allows us to calculate the mean and variance of the UMAP-derived estimator for $\mathrm{d}(i,j)$:

$$\overline{\mathrm{d}(i,j)} = \frac{\sum_{k=1}^{B} \widehat{\mathrm{d}(i,j)_{(k)}}}{B} \tag{3}$$

$$\widehat{\sigma^2_{d(i,j)}} = \frac{\sum_{k=1}^{B} \widehat{\mathrm{d}(i,j)_{(k)}}^2}{B-1} - \left(\frac{\sum_{k=1}^{B} \widehat{\mathrm{d}(i,j)_{(k)}}}{B-1} \right)^2 \tag{4}$$

This approach provides a maximum likelihood estimate, i.e., $\widehat{\sigma_{d(i,j)}}$, of the standard error (*20*). We used a heatmap to visualise $\overline{\mathrm{d}(i,j)}$ for each tissue pair (i,j). An alternative could have been to use a normalised "metric" (which is more robust to the scale from the embedding ϕ):

$$\mathrm{d}^*(i,j)_{(k)} = \frac{\mathrm{d}(i,j)_{(k)}}{\mathrm{median}_{s,t \in 1,\ldots,L} \mathrm{d}(s,t)_{(k)}} \tag{5}$$

but we found this normalisation to be unnecessary in the GTEx data.

For two tissues i_0 and i_1, we define a "clustering conservation coefficient" to quantify the preservation of the clustering of tissues i_0 and i_1 relative to all tissues $\{j\}$:

$$C_{(i_0,i_1)} = \text{corr}(\overline{\text{d}(i_0,j)}, \overline{\text{d}(i_1,j)}) \tag{6}$$

where *corr* is the correlation operator. In particular, this statistic allows us to formally test the null hypothesis of no conservation of global structure for a given pair of tissues; under the null hypothesis,

$$\sqrt{\frac{L-3}{1.06}}\text{arctanh}(C_{(i_0,i_1)}) \sim N(0,1) \tag{7}$$

This coefficient can be extended to a larger set of tissues, $i_0, \ldots, i_l$ (e.g., the 13 brain regions), using the first order statistic:

$$C_{i_0,\ldots,i_l} = \min_{s,t \in 1,\ldots,l} C_{(i_s,i_t)} \tag{8}$$

Furthermore, we calculated the relationship between the original $\mathbf{V}_{(\mathbf{0})}$ and "perturbed" $\mathbf{V}_{(\mathbf{k})}$ for each sample k:

$$r_k = \text{corr}(\mathbf{V}_{(\mathbf{0})}, \mathbf{V}_{(\mathbf{k})}) \tag{9}$$

and the resulting empirical distribution of the correlation values r_k. We note that UMAP has a stochastic element since it utilises stochastic approximate nearest neighbor search and stochastic gradient descent for optimisation; however, the r_k derived from a different run $\mathbf{V}_{(\mathbf{k})}$ (rather than from bootstrapping) quantifies the stability of the global structure in the presence of stochasticity. Collectively, our approach provides a way to perform statistical inference on the UMAP embedded structures.

Embedding new transcriptome data into UMAP learned space. To evaluate the relevance of the trained model generated from the GTEx communities, we passed previously unseen data D_{test} to the model for embedding into the learned latent map (from the UMAP embedding of GTEx training data, $\phi : D_{train} \subset M_g \hookrightarrow \mathbb{R}^m$). We used The Cancer Genome Atlas (TCGA)

gene expression data in acute myeloid leukemia (*21*), breast cancer (*22*), and lung adenocarcinoma (*23*) as test sets.

Prediction Power of Communities for Tissues. We investigated the extent to which each community's gene expression profile is predictive of each of the tissues. The master matrix $\mathbf{M}$, representing the entire dataset under analysis, has $15,201$ rows representing each RNA-Seq sample from each tissue collected from all subjects, and $18,364$ columns representing the total number of genes available. If a value was non-existent (which may be due to the gene's expression being tissue-specific), we assumed a zero value, conveying no expression in that tissue.

For each community, the expression values of the member genes were selected from $\mathbf{M}$. With this sliced table, 49 binary classifications were performed using Support Vector Machine (SVM) (*24*), wherein for each classification, we predicted each tissue. Essentially, the sliced table, which comprises the training data, for a k-member community can be viewed as a collection of vectors $\{(\vec{x_1}, y_1), \ldots, (\vec{x_n}, y_n)\}$, where $\vec{x_i} \in \mathbf{R}^k$ is the gene expression profile of the k genes for the i-th sample and $y_i \in \{1, 0\}$ indicates membership in the tissue to be predicted. The goal of the classification is to separate the tissue to be predicted from the other tissues via the largest margin hyperplane, which can be generically written as $\vec{w} \cdot \vec{x} + b = 0$, where $\vec{w}$ is normal to the hyperplane. SVM was used with a linear kernel and weights were adjusted to be inversely proportional to class frequencies in the input data (this corresponds to setting the *class_weight* parameter in *scikit-learn* to "balanced") (*25*). To avoid overfitting, each classification was performed using a stratified 3-fold cross-validation procedure, in which the F_1 score metric

$$F_1 = \frac{2}{(precision)^{-1} + (recall)^{-1}} \tag{10}$$

was used to report the prediction power across the three folds. We decided to use the F_1 score instead of other metrics, given that each binary classification was highly unbalanced, i.e., a given tissue is the positive outcome and all the other 48 tissues are the negative outcome. (Louvain communities with less than 4 genes were filtered out from this analysis.)

Prediction of Tissues Given Reactome Pathways. For comparison, we investigated the extent to which biologically meaningful sets of genes encoding current biological knowledge are predictive of tissues. For each Reactome (*26*) pathway, we selected the expression of member genes from the master matrix $\mathbf{M}$. If a gene from a Reactome pathway was not present, that gene was ignored. Using the same stratified 3-fold cross validation procedure described in the previous section, we performed 49 binary classifications.

Enrichment Analysis. To evaluate the degree to which a community corresponds to well-known biological pathways, we performed enrichment analyses using the Reactome 2016 as reference. We used the *gseapy* python package to make calls on the *Enrichr* web API (*27*), and considered significant those pathways with a Benjamini-Hochberg-adjusted p-value below 0.05. (Louvain communities with less than 4 genes were considered "not enriched.")

Multilayer Analysis. In order to investigate the tissue-shared profiles of gene communities, as well as the relationships between gene expression traits across tissues, we proceeded to model our system as a multilayer network (*28*). Formally, a multilayer network is defined as a pair $\mathbf{\Lambda} = (\mathbf{G}; \mathbf{D})$, where $\mathbf{G} := \{G_1, \ldots, G_L\}$ is a set of graphs and $\mathbf{D}$ consists of a set of interlayer connections existing between the graphs and connecting the different layers. Each graph $l \in \mathbf{G}$ is a "network layer" with its own associated adjacency matrix A_l. Thus, $\mathbf{G}$ can be specified by the vector of adjacency matrices of the L layers: $\mathbf{A} := (A_1, \ldots, A_L)$. Multilayer networks allow us to represent complex relationships which would otherwise be impossible to describe

using single-layer graphs separately considered.

A special case of multilayer networks is a multiplex network, which we used to model the GTEx transcriptome data. In this case, all layers are composed of the same set of nodes but may exhibit very different topologies. The degree of node i is the vector $d^{[i]} = (d_1^{[i]}, \ldots, d_L^{[i]})$, and $d_l^{[i]}$ may vary across the layers. Interlayer connections are established between corresponding nodes across different layers. Layers represent different tissues, nodes represent genes, and edges between two nodes are weighted according to the correlation weights. In the GTEx data, the correlation matrices, previously described, define an adjacency matrix A_l for each layer l of the multiplex network. We applied community detection analysis to each layer separately to identify communities of co-expressed genes in each tissue, using the Louvain community detection method previously described.

Having identified communities of co-expressed genes for each tissue separately, we then computed the so-called *global multiplexity index* (*29*) to investigate the relationships of communities across different layers. This index quantifies how many times two nodes (genes) are clustered in the same communities across different layers. If, for example, gene i and gene j are clustered together in the layer of tissue T_1 and of tissue T_2, then the global multiplexity index is two. In the matrix $[\text{gmi}(i,j)]$ of global multiplexity indices for a multiplex architecture, each element represents the number of times that two given genes, i and j, are clustered in the same community. More formally, if L is the number of layers, g is the generic layer, and N the number of nodes for each layer, then the global multiplexity index $\text{gmi}(i,j)$ for gene i and gene j, with i and $j \in \{0, \ldots, N\}$ is defined as follows:

$$\text{gmi}(i,j) = \sum_{g=1}^{L} \delta(c_i^g, c_j^g) \tag{11}$$

where $\delta(c_i^g, c_j^g)$ represents the Kroenecker delta function. The value of $\text{gmi}(i,j)$ therefore increases by 1 if the two nodes are found to be part of the same community in a layer. If two genes

share a high value of global multiplexity index, this may indicate a greater level of connectivity and suggest greater functional similarity, as they appear multiple times in the same community across different layers. We can define the individual probabilities:

$$\pi_k = p(\text{ gmi} = k \mid \theta \text{ }) \tag{12}$$

where θ represents all of the parameters of the model. π_k is the probability that two genes are clustered in the same communities across k layers, where $k \leq L$ and L is the number of layers in the network. We estimated the probability distribution of gmi(i, j) in the GTEx data.

We tested whether the UMAP embeddings of the transcriptome profiles of the communities in a multiplex architecture – a subset of all communities previously interrogated – could also recover biologically meaningful clusters. This analysis allowed us to estimate the topology of the high-dimensional transcriptome data and test whether additional clusters could be uncovered at increasingly finer scales.

Results

Spurious Co-expression and Confounding due to Unmodelled Factors. Disambiguating true co-expression from artefacts is an important concern in the presence of hidden variables. We therefore applied *sva* analysis to investigate unmodelled and unmeasured sources of expression heterogeneity. The number of factors or components identified by this analysis was significantly correlated ($r \approx 0.95$, $p \approx 5.4 \times 10^{-26}$) with the number of samples across tissues (see Figure 1a). Notably, the greater number of such surrogate variables that we regressed out for tissues with larger sample sizes recapitulates the approach used by the GTEx Consortium (using a related adjustment method (*30*)) of using more inferred factors for tissues with larger sample sizes (i.e., from 15 factors for tissue sample size $N < 150$ to 60 factors for $N \geq 350$) in order to optimise the number of eGenes from the eQTL analysis (*8*).

We then quantified the impact of confound correction (see Figure 1b) in co-expression analysis. The distribution of correlation (Pearson) values is closer to zero with less variance after correction, suggesting that unmodelled factors may induce spurious (or artificially inflate) correlations in gene expression. The effect of unmodelled factors is further illustrated in Figure 1, Panels (b) and (c), where the distribution of correlation values for the covariate *Age* is shown for whole blood. Before correction, those values are spread between around -0.4 and 0.4, whereas after correction the corresponding values move towards the centre (zero) and become less dispersed. Notably, the variable *Cohort* (with possible values being *Postmortem* and *Organ Donor* in available tissues, except for some which also have *Surgical* values) seems to have undergone the largest change in the correction process. This suggests that estimation of cohort effect on gene expression can be substantially improved by accounting for unmodelled factors.

Atlas of Communities across Human Tissues. For each tissue, we identified communities in the co-expression networks, using the Louvain algorithm (see Methods), to develop an atlas across human tissues. On average, a tissue was found to have 108 communities (standard deviation [SD] = 31) (see Figure 2). We observed the highest number of communities ($n = 251$) in "Kidney Cortex" and the lowest number ($n = 73$) in "Muscle Skeletal". The nonsolid tissues, consisting of "Cells EBV" and "Whole Blood", have the highest number of genes (i.e., at least 4,300 for each) that belong to a community. The size of a community varies considerably within each tissue and its distribution differs across tissues. (See Table S2 and Figure S2 for the distribution in all tissues.) Interestingly, the brain tissues (median SD = 9.9) show significantly higher variability (Mann-Whitney U test $p = 1.55 \times 10^{-4}$) than non-brain tissues (median SD = 5.18). Thus, tissues and tissue classes may differ in the overall topology of the communities in co-expression networks, which likely contains a lot of tissue information.

We noticed that after removing the weaker correlations ($-0.80 < z_{ij} < 0.80$), most of the

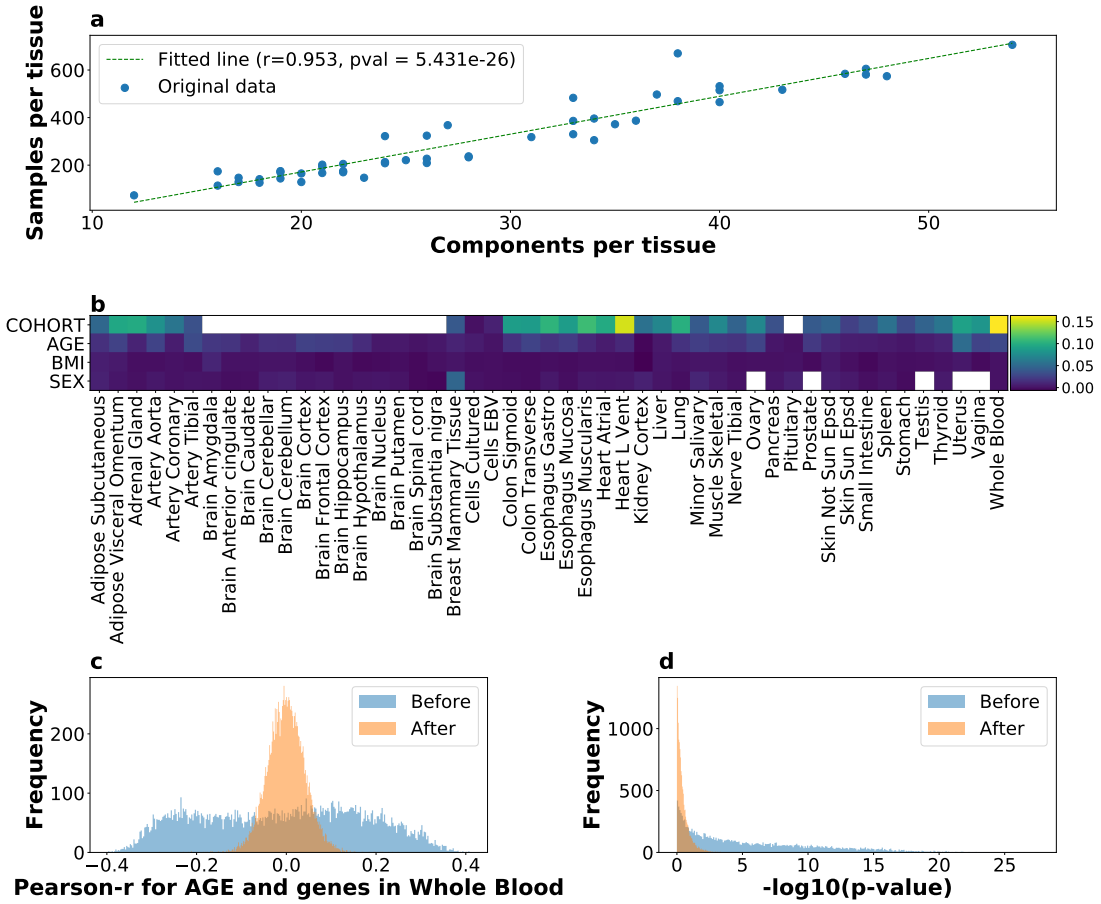

Figure 1: **Confounding due to unmodelled factors. a.** Relationship between the number of inferred factors and tissue sample size. Fitted line corresponds to a linear least-squares regression. The two-sided p-value is based on the null hypothesis that the slope is zero, using the Wald Test with t-distribution for the test statistic (*31*). **b.** The difference in the variance of the distribution of Pearson correlation values for each tissue over all genes, before and after correction. Empty cells correspond to tissues in which only one value of the confound is available. The "Cohort" variable undergoes the most substantial change after the correction across all tissues. **c.** Distribution of Pearson correlation between the expression of a gene in whole blood and age, before and after correction. After the correction, the correlation values move towards zero and show considerably less dispersion. **d.** The p-value distribution from Panel (a)'s values, in logarithmic space. The enrichment for significant (low) p-values is greatly attenuated after the correction, suggesting that unmeasured variables can induce spuriously significant correlations.

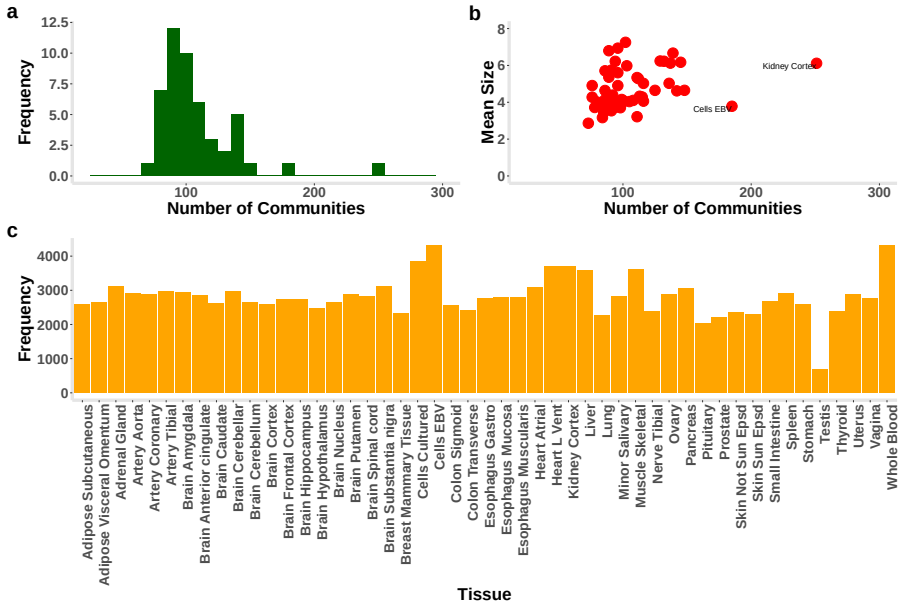

Figure 2: **Summary statistics on identified communities. a.** Histogram shows the distribution of community count in the various tissues (mean = 108, SD = 31). **b.** The scatter plot displays the community count and mean community size for each tissue, showing a significant correlation (Spearman $\rho = 0.39$, $p = 0.006$). The highest number of communities was observed in "Kidney Cortex" ($n = 251$). **c.** Plot provides the number of genes that belong to a community in each tissue. The nonsolid tissues, "Cells EBV" and "Whole Blood", show the highest number of genes with membership in a community.

subnetworks were already highly segregated from the rest of the entire network, indicating that the Louvain communities could almost be completely formed by just this removal process. In order to evaluate the segregation of such communities, we calculated the number of connections coming out of communities of each size. We found that for every tissue the mode was zero, and the maximum number was never over 17. Given the thousands of genes in each tissue's co-expression network, the observed maximum number of connections between different communities (i.e., at most 17) illustrates how strong the segregation is prior to the application of the Louvain community analysis.

More information on these communities is available on github: notebook *09_community_info*.

UMAP of Community-defined Gene Expression Manifold Reveals Tissue Clusters. Nearly 18% of the genes belong to a community in at least one tissue. Notably, gene expression from this subset was able to recover the tissue clusters (see Figure 3) as fully as the complete set of genes analysed here (see Figure S1).

Drawing conclusions about relationships between clusters (tissues) from UMAP (and related approaches) must be done with caution due to some known caveats (*32*) (see next section for more details). However, starting from known relationships between tissues, we found that the subset of community-based genes yielded biologically consistent embeddings from UMAP. Indeed, the clustering of related tissues (based on organ membership), such as the 13 brains regions, or the clustering of other related tissues (based on shared function), such as the hypothalamus-pituitary complex (which controls the endocrine system (*33*)), could be observed for the genes that belong to communities. Taken together, these results show that gene expression from the identified communities encodes sufficient information to distinguish the various tissues in a biologically meaningful low-dimensional representation. We note, however, that not all sets of genes with correlated expression produce the distinct separation of tissues observed

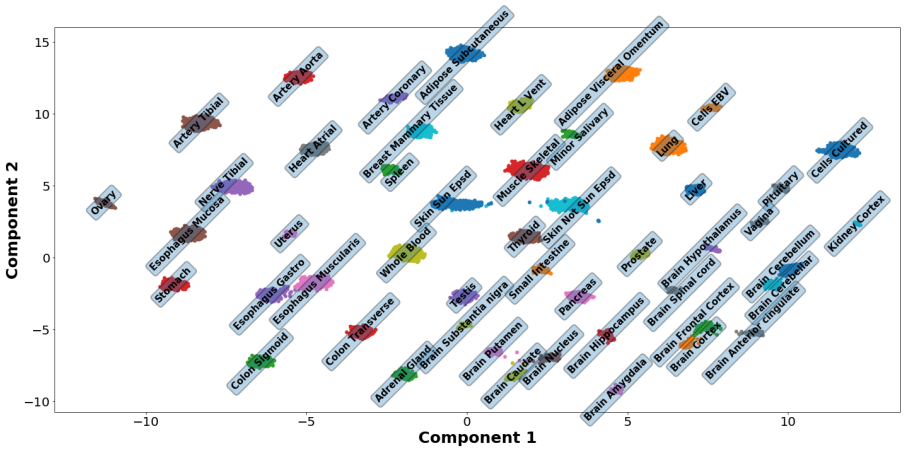

Figure 3: **Lower-dimensional representation of the transcriptome data restricted to the communities.** UMAP generates embedded structures through a low-dimensional projection of the submatrix consisting of only the genes ($n = 3,259$) that belong to a community in at least one tissue. This subset of genes (17.7% of total) contains sufficient information to recover the tissue clusters. In addition, known relationships between tissues, based on organ membership and, separately, on shared function, are reflected in the UMAP projection.

for the set of genes that belong to the communities (see below).

In theory, additional clusters may be present at different scales, e.g., within a tissue. Therefore, we performed UMAP analysis on the single-tissue "Whole Blood" to test for the presence of additional clusters. Notably, no well-defined clustering was observed with respect to cohort (Figure S3), BMI (Figure S4), and the other covariates, indicating that the *sva* analysis was successful in removing potential confounders (see Figure 1b).

Persistence of the UMAP Embeddings. We developed an approach to quantify the conservation and variability of the UMAP global structure and estimate the sampling distribution of the local structure, e.g., the distance $\mathrm{d}(i, j)$ for a given pair of tissues i and j (see Methods). Using 500 bootstrapped manifolds, we found, for example, that, on average, (a) the 13 brain regions, (b) the colonic and esophageal tissues, and (c) various artery tissues tended to cluster closely together (see Figure 4). Figure S5 shows the relationship between the average distance

between tissue clusters and the variance in the distance, showing a significant positive correlation (Spearman $\rho \approx 0.38$, $p < 2.2 \times 10^{-16}$). Reassuringly, the tissue pairs ("Brain Cerebellum", "Brain Cerebellar") and ("Skin Not Sun Epsd", "Skin Sun Epsd") had the lowest average distance between clusters among all tissue pairs; the first pair consists of known duplicates of a brain region in the GTEx data (*7*) and is thus expected to cluster together. Among the tissue pairs with the highest average distance, "Adipose Subcutaneous" with each of the colonic tissues ("Colon Sigmoid" and "Colon Transverse", each with average distance greater than 17) had low variance comparable to tissue pairs with some of the smallest average distance. Additional global patterns can be easily observed. For example, the relationships of related tissues (e.g., "Skin Sun Epsd" and "Skin Not Sun Epsd" with $C_{(i_0,i_1)} \approx 0.62$, $p = 3.4 \times 10^{-5}$) to the remaining tissues were found to be strongly preserved, using our clustering conservation coefficient (see Methods).

Embedding of test set into learned latent space. Interestingly, embedding of each of the TCGA datasets into the learned latent space from the communities showed clustering with the testis tissue. This result recapitulates two known results: (a) the GTEx finding that the testis is an outlier relative to the other GTEx tissues in transcriptome profile (*6*) and (b) the role of the so-called cancer-testis (CT) genes (*34*), which function as driver genes in cancer (*35, 36*), evoke immune responses in cancer patients as immunogenic antigens across a range of cancers (*37*), and contribute to various neoplastic phenotypes. The implementation is available on github: notebook *12_tcga*.

Prediction of Tissues by Communities. We then tested individual communities for their ability to predict a tissue. By definition, we consider that a set of genes can predict a tissue when the average F_1 score is above 0.80 (see Methods). Some broad patterns are noteworthy. Most of the communities from "Whole Blood" do not have prediction power for the other tissues (Figure 5a)

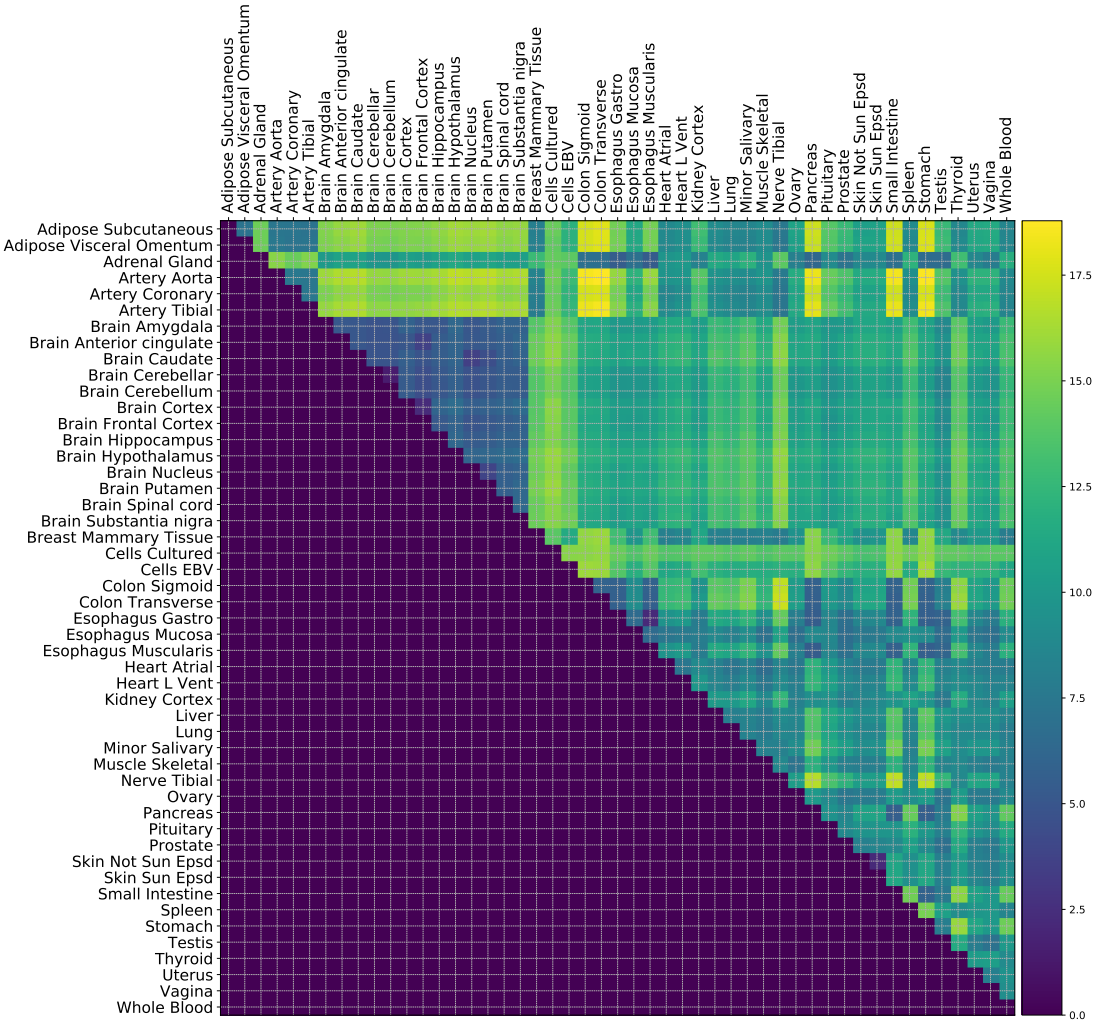

Figure 4: **Conservation of UMAP Global Structure.** Using bootstrapped manifolds (see Methods), we estimated the persistence of the global structure and pairwise relationships across tissue clusters. Here we show the upper-triangular matrix of the average pairwise distances across the bootstrapped manifolds. We found consistent clustering of known related tissues, including the 13 brain regions, the colonic and esophageal tissues, and various artery tissues. Additional patterns were observed. For example, as reflected in the heatmap, we found a highly correlated relationship, i.e., high "clustering conservation coefficient" ($C_{(i_0,i_1)} \approx 0.62$, $p = 3.4 \times 10^{-5}$) (see Methods), of the two skin tissues to all the other tissues.

partly due to the stringency of our F_1 threshold, which is likely to produce false negatives. This observation indicates that the member genes in each such community from the source tissue ("Whole Blood") cannot "separate" the test tissue (say, "Lung") from the remaining tissues possibly due to lack of tissue specificity of the gene expression profile of the community. (Here we tested a linear classifier, and the so-called "kernel trick" (*38*) may work in the non-linearly separable gene expression profiles, though perhaps at the expense of biological interpretability.) However, a community of only five genes can predict the brain region nucleus accumbens (basal ganglia). For these communities, the member genes, collectively, are "differentially expressed" between the test brain region and the remaining tissues. Thus, although the genes are present in "Whole Blood" (as a community), the expression profile in the test brain region is substantially different or tissue-specific. "Cells cultured fibroblasts" is the tissue which can be predicted by the largest number of "Whole Blood" communities (three) and, consistently, the largest number from the other tissues.

Relationship between Reactome Pathways and Tissues. Consistent with our observations for the communities, most Reactome pathways are not sufficient to predict any tissue (available on github: output *output_06_02*), while many are tissue-specific (i.e., can predict only one tissue). However, we identified Reactome pathways that can predict more than half of the tissues: *GPCR_LIGAND_BINDING*, *GPCR_DOWNSTREAM_SIGNALING*, and *SIGNALING_BY_GPCR* predict 34, 33, and 32 tissues, respectively. This observation is perhaps expected: G-protein-coupled receptors (GPCRs) comprise a large family of cell surface receptors that form the essential sites of communication between the internal and external environments of cells, with a central and widespread role in human physiology (*39*). Their gene expression profile in each of the predicted tissues differs from the remaining tissues, potentially reflecting their broad but tissue-specific function.

Prediction of tissues by Reactome pathways varies substantially. The brain tissues "Brain Caudate" (basal ganglia), "Brain Frontal Cortex", "Brain Hippocampus", and "Brain Nucleus" accumbens (basal ganglia) are not predicted by any Reactome pathway, likely reflecting the fact that our current understandings (as encoded in these pathways) have been hampered by the relative inaccessibility of these tissues. In contrast, the tissues cells cultured fibroblasts and whole blood are the tissues most highly predicted (269 and 330 Reactomes, respectively). Some tissues are predicted by less than 5 Reactome pathways, including the tissues "Brain Amygdala" (two), "Brain Anterior Cingulate" cortex (BA24) (one), "Brain Cortex" (two), and "Brain Hypothalamus" (two). More information on the relationship between communities and enriched pathways is available on github: notebook *10_reactomes_per_tissue*.

Enrichment of Communities for Known Biological Processes. We quantified the extent to which the communities in the various tissues reflect current biological knowledge (as encoded in the Reactome pathways). We identified 114 communities (8.28% of all the communities with more than 3 member genes) enriched for some Reactome pathway (i.e., at an adjusted $p < 0.05$ for level of enrichment), thus contributing in complex ways to multiple biomolecular processes. "Whole Blood" was the only tissue without any community enriched for known pathways, and the "Esophagus Mucosa" was the tissue with the most communities enriched for known pathways, with a total of 5 communities. Since the entire set of communities could fully recover all tissues as clusters in the UMAP embeddings, these results suggest that the remainder of the communities are likely to capture previously inaccessible and novel tissue biology.

Notably, our analysis may uncover the role of these communities in human diseases. For example, a community of 15 genes in the "Brain Hippocampus" showed a significant enrichment for diseases associated with glycosaminoglycan metabolism (adjusted p-value = 0.026; see Figure 5). Interestingly, glycosaminoglycans, which are major extracellular matrix components

whose interactions with tissue effectors can alter tissue integrity, have been shown to play a role in brain development (*40, 41*), modulating neurite outgrowth and participating in synaptogenesis. Alterations of glycosaminoglycan structures from Alzheimer's disease hippocampus have been implicated in impaired tissue homeostasis in the Alzheimers disease brain (*42*). They have also been found to be altered in the hippocampus of patients with temporal lobe epilepsy (*43*).

Multiplex Analysis of the Transcriptome. We analysed five multiplex networks to model the various tissue interactions, of clear biological interest, in the GTEx dataset. For each multiplex architecture, only the specific component tissues were used to construct the multiplex network, and consequently we calculated the global community index for each multiplex architecture, using only the component tissues of the multiplex network.

- **All Tissues**: Each layer represents one of the 49 tissues analysed. This architecture allows us to investigate gene communities that are shared across all tissues, with potentially universal function.

- **Brain Tissues**: The 13 layers correspond to the various brain regions. This architecture facilitates identification of communities that may play a functional role throughout the central nervous system (CNS).

- **Brain Tissues and Whole Blood**: This multiplex model consists of the 14 layers corresponding to these tissues. This architecture allows us to study brain-derived communities for which the easily accessible whole blood can serve as a proxy tissue.

- **Brain and Gastrointestinal Tissues**: The 16 layers correspond to the brain tissues and 3 gastrointestinal tissues. This architecture may provide insights into the gut-brain axis (*44*), which has attracted recent attention in the literature, e.g., in the study of neu-

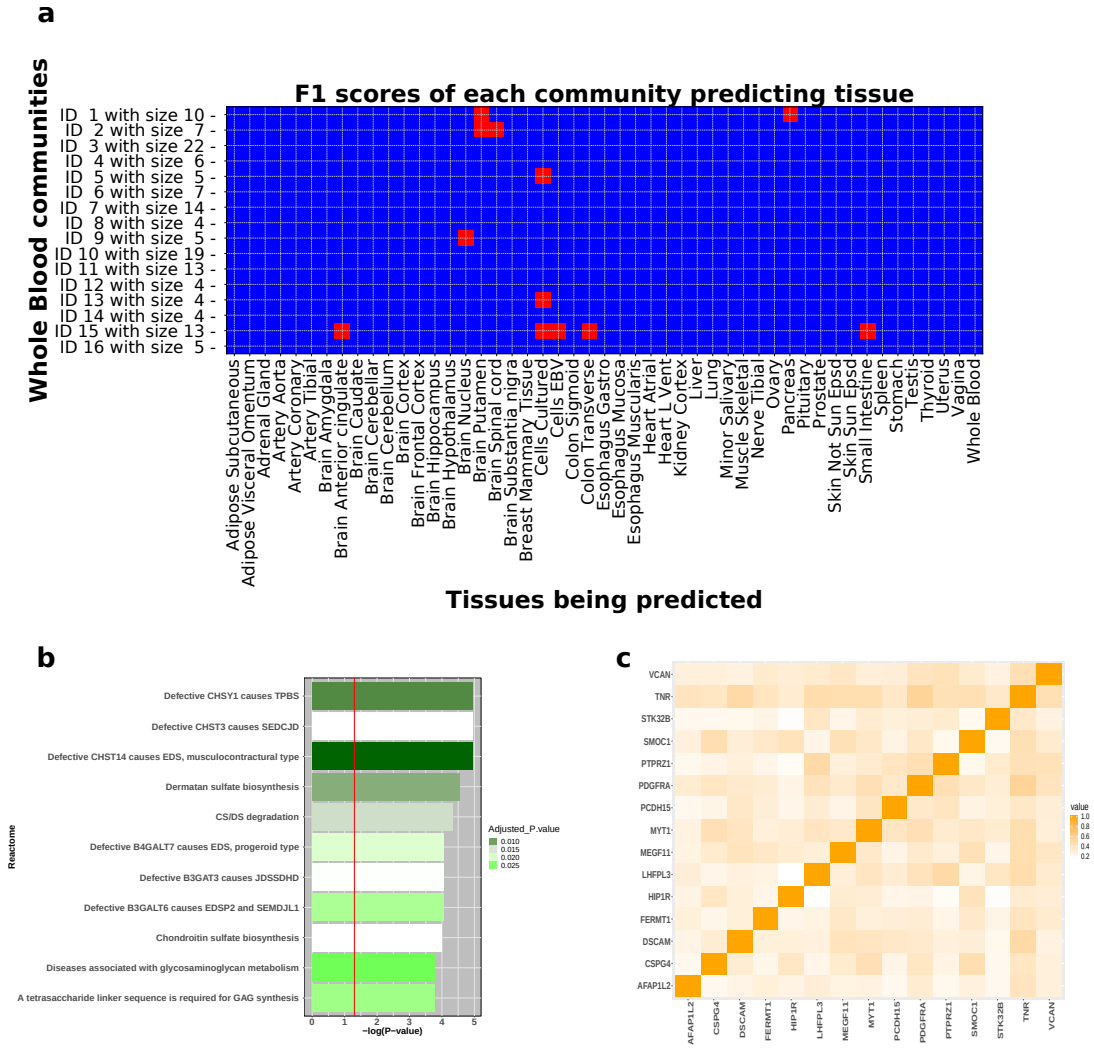

Figure 5: **Communities in co-expression networks, their prediction power for tissues, and enrichment for known biological processes. a.** Prediction power of "Whole Blood" communities, in F1 scores thresholded over 0.8. Most communities in "Whole Blood" do not have prediction power for the remaining tissues. There are notable exceptions, such as a 13-member community that can predict multiple tissues, including "Brain Anterior cingulate", "Small Intestine", and "Colon Transverse". **b.** Enrichment analysis was performed on all communities to determine the extent to which each community would capture known biological processes. For example, a 15-member hippocampal community (characterised here) was found to be significantly enriched for Reactome pathways. P-value refers to raw p-value. Note the x-axis shows the minus log-transformed p-value. Red line corresponds to the raw $p < 0.05$ threshold. Colour gradient reflects the adjusted p-value. All Reactome pathways shown are the ones that meet adjusted $p < 0.05$. **c.** Heatmap displays the gene-gene correlation values for the member genes of the community in panel (b). While the relationships of individual member genes to the pathways are known, their organisation into a community structure within a tissue-dependent co-expression network is a novel finding, suggesting coordinated function.

ropsychiatric processes (*45, 46*) and the interaction between the CNS and the enteric nervous system in neurological disorders (*47*).

- **Non-Brain Tissues**: The 36 layers consist of all tissues outside the brain. This architecture may stimulate investigations into developmental and pathophysiological processes outside the CNS.

Multiplex analysis provides an inter-tissue framework for the analysis of high-dimensional molecular traits such as gene expression. The global multiplexity matrix was obtained for each of the five proposed architectures. We extracted from the global multiplexity matrices the groups of genes with the maximal global multiplexity index in the five architectures, i.e., the groups of genes that share a value of 49, 13, 14, 16, and 36 respectively, equal to the number of layers in the respective architectures. Among these groups of genes with the highest global multiplexity index, we obtained the sub-clusters for each architecture, identifying the groups of genes that always appear in the same community across the various layers. Revealing the shared community structure across the layers improves our understanding of the functional and disease consequences of the clusters of genes. We then investigated the biological pathways (Reactome) in which such subgroups were involved for each architecture. Our goal was to test the communities for enrichment for known biological pathways and therefore quantify the degree to which the communities capture current understanding of biological processes as encoded in the knowledge base.

We illustrate this approach here. (The complete results for all 5 architectures can be found on our github page in the jupyter notebook *11_multiplex_enrichment.ipynb*.) In Figure 6a, we show an example of a multiplex network. In this case, the multiplex network is constructed from data in two brain tissues: "Brain Hippocampus" and "Brain Putamen" (basal ganglia). Each layer represents a tissue, and nodes are labelled as the genes to which they correspond.

In the structure, we can see the presence of two communities (in this case, randomly drawn from the full set) formed by the groups of genes: {*FGA*, *ORM1*, *FGB*} and {*MYH11*, *ACAT2*, *MYL9*}. Corresponding genes are connected through the layers as shown, through interlayer connections. We note that, based on our analysis, these two communities are indeed part of the "Brain Tissues" multiplex architecture, present in all 13 component layers (brain regions). Notably, all three genes that belong to the first community have been previously implicated as biomarker and therapeutic target candidates for intracerebral hemorrhage (*48*). This observation is interesting given our Reactome pathway analysis results; the community is enriched for "Common Pathway of Fibrin Clot Formation" (adjusted $p = 8.8 \times 10^{-4}$) and "Formation of Fibrin Clot (Clotting Cascade)" (adjusted $p = 1.7 \times 10^{-3}$), indicating the genes' involvement in coagulation. All three members of the second community are myelin-associated genes (*49*), with *MYH11* expression in putamen (basal ganglia) showing nominal association with increased risk of myocardial infarction ($p = 6.5 \times 10^{-3}$) from PrediXcan analysis (*50*) of a large-scale meta-analysis (*51*). Myelination-related genes have been associated with decreased white matter integrity in schizophrenia (*52*). We found a 17-member community in the "Brain Tissues" multiplex that is significantly enriched for the nonsense-mediated decay (NMD) pathway (adjusted $p = 1.01 \times 10^{-37}$), which is known to be a critical modulator of neural development and function (*53*). The pathway accelerates the degradation of mRNAs with premature termination codons, limiting the expression of the truncated proteins with potentially deleterious effects. The community's presence in all brain regions underscores its crucial protective function throughout the central nervous system.

The multiplex analysis we performed can also be used to investigate the relationship between two distinct systems. Here we illustrate this using the CNS and the gastrointestinal system, possibly reflecting a coordinated transcriptional regulatory mechanism between the CNS and the enteric nervous system (ENS). The ENS is a large part of the autonomic nervous sys-

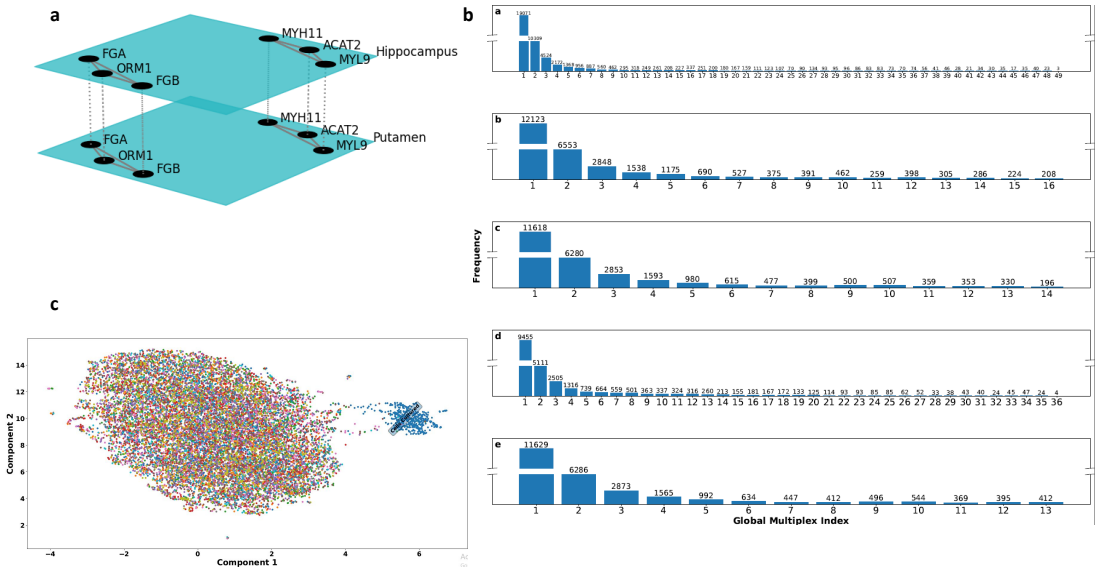

Figure 6: **Multiplex analysis. a.** An example of a multiplex network constructed from data in "Brain Hippocampus" and "Brain Putamen" (basal ganglia). Each layer denotes a tissue, and nodes are the genes, which are connected via the interlayer connections. In the multiplex example, we can see the presence of two communities. These two communities are indeed part of the "Brain Tissues" multiplex architecture, present in all 13 component layers (brain regions). All genes in the community $\{FGA, ORM1, FGB\}$ have been previously implicated as biomarker and therapeutic target candidates for intracerebral hemorrhage. **b.** Histograms show the empirical distribution of the global multiplexity index for each multiplex architecture, for when the index is positive. The index quantifies how many times two genes belong to the same communities across different layers. The proportion or frequency at each value k of the index is an estimate of the π_k (see Methods). All the 5 multiplex architectures follow a similar, skewed pattern. The maximum global multiplexity index corresponds to the number of layers or tissues used to define the multiplex network. Histograms, starting from the top, correspond to the following multiplex architectures: All tissues; Brain tissues; Brain tissues and whole blood; Brain and gastrointestinal tissues; Non-brain tissues. **c.** We performed UMAP on the subset of communities that exist across all layers of the central nervous system ("Brain Tissues") multiplex. This set of communities does not yield complete clustering of all tissues, but it uncovers a different pattern of tissue clustering.

tem that can control gastrointestinal behaviour (*47*). We found a 14-member community in the "Brain and Gastrointestinal Tissues" multiplex, whose presence in all 16 layers suggests a strong interaction between the CNS and ENS. Consistent with this hypothesis, the community was found to be significantly enriched for the "metabolism of vitamins and cofactors" (adjusted $p = 6.5 \times 10^{-7}$), which has been shown to be responsible for altered functioning of the CNS and ENS (*54*). Although the involvement of the individual member genes in this pathway is known, the finding that the genes are organised as a community structure, within co-expression networks, which persists across the entire 16 layers of the various brain regions and the gastrointestinal tissues examined here is a novel one.

The empirical distribution of the global multiplexity index is presented in Figure 6b for each of the five architectures. The maximal global multiplexity index in the five architectures represents the groups of genes that share a value of 49, 13, 14, 16, and 36 respectively, equal to the number of layers (i.e., tissues) in the respective architectures. These genes appear in the same community across *all* layers of the respective architectures.

For comparison with the UMAP embeddings of the set of all communities, we performed similar analyses in the various multiplex networks. For example, we tested whether the complete tissue clustering could be observed using just the subset of communities that exist across all layers of the central nervous system multiplex. Interestingly, we discovered a different clustering pattern, with cultured fibroblasts clustering separately from the rest of the tissues, which no longer show well-defined clustering (Figure 6c). This finding suggests the presence of a hierarchy of clusters in the transcriptome at increasingly finer scales.

Finally, a freely available genomics resource of communities, their predictive power for tissues, and multiplex networks, and the software to construct a similar resource on a transcriptome dataset can be accessed at `https://github.com/tjiagoM/gtex-transcriptome-modelling`.

Discussion

We develop an inter-tissue multiplex framework for the analysis of the human transcriptome. Given the complexity of pathophysiological processes underlying complex diseases, intra-tissue and inter-tissue transcriptome analysis should enable a more complete mechanistic understanding. For these phenotypes, studying the interaction among tissues may provide greater insights into disease biology than an intra-tissue approach. Communities in co-expression networks are here shown to be enriched for some known pathways, encoding current understandings of biological processes; however, we identify other communities that are likely to contain novel or previously inaccessible functional information.

UMAP embeddings of the transcriptome of the entire set of communities (representing only 18% of all genes) fully reveal the tissue clusters. Low-dimensional representation of the subset of communities that are in the multiplex networks does not recover the tissue clusters, but uncovers other clustering patterns, suggesting a hierarchy of clusters at increasingly finer scales. We develop an approach to quantify the conservation of, and uncertainty in, the UMAP global structure and estimate the sampling distribution of the local structure (e.g., distances among tissue clusters), with broad relevance to other applications (such as cell population identification in single-cell transcriptome studies (*55*)). Importantly, new gene expression data can be embedded into our models, facilitating integrative analyses of the large volume of transcriptome data that are increasingly available. We provide a publicly available resource of co-expression networks, communities, multiplex architectures, and enriched pathways, and code to stimulate research into network-based studies of the transcriptome.

Using the global multiplexity index, we investigate the tissue-sharedness of identified communities. In fact, communities that are shared across multiple tissues may suggest the presence of a tissue-to-tissue mechanism that controls the activity of member genes across the layers

in the network. Such regulatory mechanisms have been relatively understudied in comparison with intra-tissue controls.

We identified tissue-dependent communities that are enriched for human diseases. For example, we found a 15-member community in the "Brain Hippocampus" that is enriched for diseases associated with glycosaminoglycan metabolism. These genetic disorders are due to mutations in the biosynt"hetic enzymes, e.g., glycosyltransferases and sulfotransferases, for glycosaminoglycans. Sulfated glycosaminoglycans include the chondroitin sulfate and dermatan sulfate chains that are covalently bound to the core proteins of proteoglycans, which are present in the extracellular matrices and at cell surfaces. Mutations affecting the biosynthesis of these chains may lead to genetic diseases that are characterized by craniofacial dysmorphism and developmental delay (*56*). The community structure we identified proposes a cooperative role for these genes, and the fact that they span multiple chromosomes suggests the presence of coordinated transcriptional regulation.

Some of the communities are shared across multiple tissues; their dysregulation may thus lead to pleiotropic effects and contribute to known and novel comorbidities. Here we modeled these communities as belonging to layers of multiplex networks. For example, we identified a 17-member community in the "Brain Tissues" multiplex network (i.e., spanning across all brain regions sampled here), consisting primarily of ribosomal proteins (therefore, enriched for proteins involved in translation [adjusted $p = 2.53 \times 10^{-35}$]), with significant overrepresentation for the viral mRNA translation pathway (adjusted $p = 1.06 \times 10^{-38}$), NMD (adjusted $p = 1.01 \times 10^{-37}$), and other Reactome pathways. Viral mRNAs in the cytoplasm can be translated by the host cell ribosomal translational apparatus, and indeed viruses have evolved strategies to recruit the host translation initiation factors necessary for the translation initiation by host cell mRNAs (*57*). NMD, a surveillance pathway that targets mRNAs with aberrant features for degradation (*58*), may interfere with the hijacking of the host translational machinery (*59*).

In the brain, NMD, as a post-transcriptional mechanism, affects neural development, neural stem cell differentiation decisions, and synaptic plasticity, and thus defects in the pathway can cause aberrant neuronal activation and neurodevelopmental disorders (*53*). Detecting this co-expression network of ribosomal proteins therefore provides a sanity check to our approach, but the identified community structure and the presence of this in the multiplex may suggest a highly coordinated regulatory mechanism across the tissues (*60*).

In summary, we performed network analysis on the most comprehensive human transcriptome dataset available to gain insights into how structures in co-expression networks may contribute to biological pathways and mediate disease processes. The rich resource we generated and the network approach we developed may prove useful to other omics datasets, facilitating studies of inter-tissue and intra-tissue regulatory mechanisms, with important implications for our mechanistic understanding of human disease.

Data Availability

The protected data for the GTEx project (for example, genotype and RNA-sequence data) are available via access request to dbGaP accession number phs000424.v8.p2. Processed GTEx data (for example, gene expression and eQTLs) are available on the GTEx portal: https://gtexportal.org. Results and code (for reproducibility) are available on github at https://github.com/tjiagoM/gtex-transcriptome-modelling

References

1. S. Fortunato, *Physics Reports* **486**, 75 (2010).

2. R. Shalgi, D. Lieber, M. Oren, Y. Pilpel, *PLoS computational biology* **3** (2007).

3. Q. Zhu, *et al.*, *Nature Methods* **12**, 211 (2015).

4. E. R. Gamazon, *et al.*, *Nature genetics* **50**, 956 (2018).

5. K. Watanabe, *et al.*, *Nature genetics* **51**, 1339 (2019).

6. G. Consortium, *et al.*, *Science* **348**, 648 (2015).

7. G. Consortium, *et al.*, *Nature* **550**, 204 (2017).

8. F. Aguet, *et al.*, *BioRxiv* p. 787903 (2019).

9. M. M. Babu, N. M. Luscombe, L. Aravind, M. Gerstein, S. A. Teichmann, *Current opinion in structural biology* **14**, 283 (2004).

10. J. Harrow, *et al.*, *Genome biology* **7**, S4 (2006).

11. P. Parsana, *et al.*, *Genome Biology* **20** (2019).

12. A. Buja, N. Eyuboglu, *Multivariate Behavioral Research* **27**, 509 (1992).

13. F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, D. Parisi, *Proceedings of the national academy of sciences* **101**, 2658 (2004).

14. V. D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre, *Journal of statistical mechanics: theory and experiment* **2008**, P10008 (2008).

15. M. Rubinov, O. Sporns, *NeuroImage* **56**, 2068 (2011).

16. M. E. Newman, M. Girvan, *Physical review E* **69**, 026113 (2004).

17. L. McInnes, J. Healy, J. Melville, *arXiv preprint arXiv:1802.03426* (2018).

18. L. v. d. Maaten, G. Hinton, *Journal of machine learning research* **9**, 2579 (2008).

19. M. Ringnér, *Nature biotechnology* **26**, 303 (2008).

20. J. Friedman, T. Hastie, R. Tibshirani, *The elements of statistical learning*, vol. 1 (Springer series in statistics New York, 2001).

21. C. G. A. R. Network, *New England Journal of Medicine* **368**, 2059 (2013).

22. C. G. A. Network, *et al.*, *Nature* **490**, 61 (2012).

23. C. G. A. R. Network, *et al.*, *Nature* **511**, 543 (2014).

24. C. Cortes, V. Vapnik, *Machine learning* **20**, 273 (1995).

25. E. E. Osuna, Support vector machines: Training and applications, Ph.D. thesis, Massachusetts Institute of Technology (1998).

26. A. Fabregat, *et al.*, *Nucleic Acids Research* **46**, D649 (2017).

27. M. V. Kuleshov, *et al.*, *Nucleic Acids Research* **44**, W90 (2016).

28. M. Kivelä, *et al.*, *Journal of complex networks* **2**, 203 (2014).

29. D. Hristova, A. Rutherford, J. Anson, M. Luengo-Oroz, C. Mascolo, *PloS one* **11** (2016).

30. O. Stegle, L. Parts, M. Piipari, J. Winn, R. Durbin, *Nature protocols* **7**, 500 (2012).

31. P. Virtanen, *et al.*, *Nature Methods* **17**, 261 (2020).

32. A. Diaz-Papkovich, L. Anderson-Trocmé, S. Gravel, *PLoS genetics* **15** (2019).

33. A. Rijnberk, *Clinical endocrinology of dogs and cats* (Springer, 1996), pp. 11–34.

34. Y. Chang, *et al.*, *Cancer medicine* **8**, 3511 (2019).

35. A. J. Simpson, O. L. Caballero, A. Jungbluth, Y.-T. Chen, L. J. Old, *Nature Reviews Cancer* **5**, 615 (2005).

36. C. Wang, *et al.*, *Nature communications* **7**, 1 (2016).

37. M. J. Scanlan, A. Simpson, L. J. Old, *et al.*, *Cancer Immun* **4**, 1 (2004).

38. B. Schölkopf, *Advances in neural information processing systems* (2001), pp. 301–307.

39. D. M. Rosenbaum, S. G. Rasmussen, B. K. Kobilka, *Nature* **459**, 356 (2009).

40. R. Margolis, R. Margolis, L. Chang, C. Preti, *Biochemistry* **14**, 85 (1975).

41. K. R. Long, W. B. Huttner, *Royal Society Open Biology* **9**, 180216 (2019).

42. M. B. Huynh, *et al.*, *PloS one* **14** (2019).

43. S. R. Perosa, *et al.*, *Epilepsia* **43**, 159 (2002).

44. M. Carabotti, A. Scirocco, M. A. Maselli, C. Severi, *Annals of gastroenterology: quarterly publication of the Hellenic Society of Gastroenterology* **28**, 203 (2015).

45. E. R. Gamazon, A. H. Zwinderman, N. J. Cox, D. Denys, E. M. Derks, *Nature genetics* **51**, 933 (2019).

46. G. Rogers, *et al.*, *Molecular psychiatry* **21**, 738 (2016).

47. M. Rao, M. D. Gershon, *Nature Reviews Gastroenterology & Hepatology* **13**, 517 (2016).

48. G.-c. Li, *et al.*, *Clinical proteomics* **14**, 14 (2017).

49. S. B. Siems, *et al.*, *Elife* **9**, e51406 (2020).

50. E. R. Gamazon, *et al.*, *Nature genetics* **47**, 1091 (2015).

51. M. Nikpay, *et al.*, *Nature genetics* **47**, 1121 (2015).

52. I. Chavarria-Siles, *et al.*, *European Journal of Human Genetics* **24**, 381 (2016).

53. S. R. Jaffrey, M. F. Wilkinson, *Nature Reviews Neuroscience* **19**, 715 (2018).

54. M. Majewski, A. Kozlowska, M. Thoene, E. Lepiarczyk, W. Grzegorzewski, *J Physiol Pharmacol* **67**, 3 (2016).

55. C. Hafemeister, R. Satija, *Genome Biology* **20**, 1 (2019).

56. S. Mizumoto, S. Ikegawa, K. Sugahara, *Journal of Biological Chemistry* **288**, 10953 (2013).

57. M. Bushell, P. Sarnow, *The Journal of cell biology* **158**, 395 (2002).

58. Y.-F. Chang, J. S. Imam, M. F. Wilkinson, *Annu. Rev. Biochem.* **76**, 51 (2007).

59. R. E. Rigby, J. Rehwinkel, *Trends in immunology* **36**, 179 (2015).

60. S. Xue, M. Barna, *Nature reviews Molecular cell biology* **13**, 355 (2012).

Acknowledgements

E.R.G. is grateful to Clare Hall, University of Cambridge for the Fellowship support. This research is supported by the National Institutes of Health Genomic Innovator Award (NHGRI R35HG010718). T.A. is funded by the W. D. Armstrong Trust Fund, University of Cambridge, UK. We thank The Genotype-Tissue Expression (GTEx) project for the use of v8 data and The Cancer Genome Atlas (TCGA) project for the use of gene expression data from the cancer samples.

Author contributions

T.A. and G.M.D. conducted the analysis. E.R.G, T.A., G.M.D. wrote the manuscript. E.R.G. and P.L. designed and supervised the study. All authors contributed to the editing of the manuscript.

Competing interests

E.R.G. receives an honorarium from the journal Circulation Research of the American Heart Association, as a member of the Editorial Board. He performed consulting on pharmacogenetic analysis with the City of Hope / Beckman Research Institute. The other authors declare no competing interests.

Supplementary materials

Figs. S1 to S5

Tables S1 to S2