DECODE-ing sparsity patterns in single-cell RNA-seq

Shahin Mohammadi[1,*], Jose Davila-Velderrain[1], Manolis Kellis[1] & Ananth Grama[2]

[1]*Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA.*

[2]*Department of Computer Science, Purdue University, West Lafayette, IN 47907, USA.*

An inherent challenge of interpreting single-cell transcriptomic profiles is the high frequency of zero values. This phenomenon has been attributed to both biological and technical sources, although the extent of the contribution of each remains unclear. Here, we show that the underlying gene presence/absence sparsity patterns are themselves highly information-rich. We develop an algorithm, called DECODE, to assess the extent of joint presence/absence of genes across different cells. We show that this network captures biologically-meaningful pathways, cell-type specific modules, and connectivity patterns characteristic of complex networks. We develop a model that uses this network to discriminate biological vs. technical zeros, by exploiting each gene's local neighborhood. For non-biological zeros, we build a predictive model to impute the missing value using their most informative neighbors. In three diverse datasets, we show that our framework accurately infers gene-gene functional dependencies, pinpoints technical zeros, and biologically-meaningful missing values.

Single-cell technologies are revolutionizing the depth that we can investigate cell type/tissue heterogeneity in health and disease. Among these technologies, single-cell RNA-seq (scRNA-seq) aims to quantify this heterogeneity within individual cells at the transcriptome level. This task, however, is challenging due to both biological reasons, including the intrinsically stochastic nature of transcription and its spatiotemporal dependency, as well as technical limitations, such as insufficiently deep sampling of transcripts within single

cells [1, 2, 3]. These factors jointly influence the structure of single-cell transcriptomic profiles and result in sparse gene expression matrices, with a high number of genes showing no observed scRNA-seq transcripts (zero values), even for transcripts that show high expression in bulk samples [4]. To overcome this challenge, we need new methods to distinguish biological vs. technical sources of such sparsity, and to predict the likely value of expression of undetected genes [2].

Here, we focus on characterizing the sparsity pattern of scRNA-seq profiles, and exploiting them for missing value imputation. Rather than considering the sparsity as a limitation, we show that it contains information that reflects biologically meaningful associations. We develop a new framework to capture the inherent sructure embedded in the scRNA-seq sparsity patterns, and use it to infer a gene-gene network relating functionally-linked genes, which we then use to infer the cell-specific activity for each gene. As the basis of our framework, we introduce a statistical measure to quantify statistical dependencies in the presence/absence patterns for all pairs of genes across individual cells. We demonstrate that the co-dependency network provides meaningful biological information, and uncovers functional modules that are highly enriched for cell-type specific functions. We exploit the network structure to quantify the probability of functional activity for each gene in each cell, based on aggregate evidence from its first-degree neighbors, with the information flow adaptively adjusted for each gene-cell pair. Our framework provides several predictions regarding the context-specific activity of genes and their functional modularity. We use three independent and diverse datasets to test our predictions across different technologies and heterogeneous tissues/cell types, and we find in each case strong support of their biological relevance. Lastly, we apply of our infered single-cell gene-gene networks to the study of Type 2 diabetes (T2D) in the context of pancreas single-cell transcriptomics, and identify cell-specific and T2D-specific pathways supported by functional networks, protein-protein interactions, and distict gene expression patterns. Our results demonstrate that our

framework successfully exploits sparsity patterns in single-cell data to impute missing values and to enable data integration for the study of human biology and disease.

Deciphering the sparsity structure of single-cell transcriptomic profiles

To detect potential biological signals in the sparsity structure stemming from zeros in single-cell transcriptional profiles, we studied dependencies in the presence-absence patterns of pairs of genes. We introduce a new measure, a gene co-dependency index (CDI), to quantify the degree of deviation of the joint sparsity pattern of the gene pairs from the independence model. For each pair of genes, CDI formally quantifies in an information-theoretic framework how informative a non-zero observation for one gene is on the probability of a non-zero observation for the other gene, based on the deviation of their joint co-occurrence frequencies from that expected given their marginal probabilities, quantified as the negative log 10 of the binomial statistic P-value (see Methods). This CDI metric results in a weighted undirected network connecting all pairs of genes based on the degree of dependency in their activity profiles.

We applied our CDI framework and constructed gene co-dependency networks for three separate datasets, spanning different cell types, technologies, and species (Supplementary file 1). We thresholded these networks at a CDI value of 5 (P-value ¡ 10^{-5}), and only considered gene-gene edges above that threshold. We studied several topological properties of the resulting thresholded network, including the network clustering coefficient, average path length, modularity, density, and maximum connected component size (Supplementary Table S1). We also studied how these properties changed with varying CDI threshold values (Supplementary Figure S1), and found that each of these properties was stable and changed smoothly until the CDI=5 threshold, though some became unstable for more stringent CDI thresholds given the in-

creasing sparsity of highly-stringent threshold.

The resulting networks for human pancreas, human PBMC, and mouse brain single-cell data displayed several properties characteristic of real-world networks in their clustering coefficient, degree distribution, connectivity preference patterns, and modularity. Their degree distribution showed an approximately linear decay, in the log-frequency as a function of log-degree, which indicates a heterogeneous distribution characterized by a majority of genes having low connectivity and a small group of highly connected genes or hubs (Figure 2a - top). In addition, genes with higher degrees connected with genes with similarly high degree, while genes of lower degree did not show a clear preference. Although the networks showed qualitatively similar properties, the mouse brain network deviated the most, showing higher density and more highly connected genes.

We then sought to recognize whether the resulting gene-gene networks show modules of genes involved in shared functional processes, and distinctive of cellular states of cell subpopulations. We used a modularity maximization algorithm [5] to identify gene modules in each dataset (Supplementary file 2), and we then characterized their functional relevance. The gene network in the mouse brain dataset partitioned into 5 gene clusters, of which 4 correspond to cell type-specific genes and pathways related to the underlying cell types in the population, including neuronal, microglial, astrocyte, and oligodendrocyte genes. Genes with glial functions showed higher interconnectivity among each other and higher separation from genes with neural functions. The pancreatic gene network partitioned into five modules of genes, of which the top four mapped to endocrine genes (active in alpha and beta islets) and exocrine genes (active in acinar and ductal celltypes). The PBMC gene network partitioned into 10 modules of genes, corresponding to genes active in monocytes and in dendritic cells. The remaining genes in this dataset formed large module that encompassed diverse cell-type specific functions. Notably, we identified a gene module corresponding

to genes active in hematopoietic stem cells (HSCs), suggesting that HSCs are represented in the dataset, although these were missed during the annotation phase of the original data.

We next examined the co-dependency network inferred from PBMC single-cell data in great detail. We first visualized the network and color-coded the modules according to their neighborhoods (Figure 2b). Experimentally, the PBMC dataset was annotated to contain four cell types: NK-cells, T-cells, B-cells, and myeloid-derived cells (monocytes and dendritic cells). Indeed, the gene modules recovered in the co-dependency network correspond to genes with known functions in each of the four cell types, and these functions are significantly enriched in each case. In addition, we studied the hub genes of each module, corresponding to those genes with the most connections, and thus the largest number of co-dependencies with other genes. We found that these hubs correspond to known canonical marker genes commonly used to characterize and isolate the corresponding cell types, even though they were inferred entirely *de novo* by our methods. In addition to the four annotated cell types, we found one module of genes associated with HSCs, with canonical marker CD34 as a hub, suggesting the presence of a subpopulation of cells with HSC state.

We next compared our co-dependency index (CDI) measure to two alternative measures commonly-used of gene functional interaction inference: correlation, and mutual information (Figure S3). We compared the ability of each method to recover pairs of genes that are expressed significantly in a similar group of cell types (see Methods), as the gold standard of related pairs. We found that correlation showed the lowest recovery (XX%), followed by mutual information (XX%), and our CDI measure showed the strongest recovery (XX%) in the PBMC dataset.

In summary, this establishes DECODE as a novel method that can identify functionally relevant pairs

of genes that are complementary to other measures. These interactions form a modular network that resembles characteristics of complex networks.

Inferring gene activities using gene-gene dependency network

Having established that the sparsity patterns of gene activity are able to infer biologically-meaningful gene-gene networks and modules, we next sought to use them to infer the true functional activity of a gene, and to distinguish between biological and technical zeros in a single-cell transcriptome.

Guilt-by-association has been shown to enable transfer of information between proximal genes in different networks [6]. Using the same principle, we developed a statistical model to infer whether or not a gene is active within a given network context by looking at its closest neighbors in the co-dependency network. If among these genes there is a significant fraction that is active, we infer that the target gene is active. While the co-dependency network is universally used across all samples, the optimal neighborhood size for each gene in a cell is automatically adjusted. For each gene in a given cell, we first rank its network neighbors based on their decreasing CDI values. Then, we assign a binary vector to this list, based on the activity of the genes in the given cell. We assess the over-representation of "active" genes among neighborhoods of increasing sizes while tracking the changes in over-representation of active genes among top-ranked neighbors. Finally, we use a dynamic-programming approach to assign an exact p-value to this optimal neighborhood size (see Methods section). This process generates a matrix of the same size as the original expression, and with inferred p-values for functional gene activity as entries. Following this process, we operationalize the principle of 'guilt-by-association' in a hypothesis testing scenario to infer from the network context whether a gene is likely to be active or not.

Our framework predicts that a primarily inactive network neighborhood is likely to reveal true biological zeros of a cell. We do not have a gold standard for such true biological zeros; however, if we know cell-types within the population of cells that are functionally distinct, we can identify cell type-specific gene markers that are unique to each cell type (see Methods section). By definition, these highly-specific genes are unlikely to play a distinctive function in the other cell-type contexts. We thus expect that the inferred activity values should recover the pattern of specificity of marker genes. Figure 3a shows the subspace of marker genes in which samples are permuted according to corresponding cell types. In this framework, diagonal blocks correspond to positive gene/cell-type pairs; that is, genes whose activity is discriminatory of the cell-type, and thus, in part, underlie its functional specificity. On the other hand, off-diagonal elements correspond to genes not expected to be functionally active – i.e., more likely to be a biological zero. Given the sparse nature of scRNA-seq, the diagonal blocks are not perfectly filled following this specificity pattern. As such, it provides an appropriate scenario to test our predictions. To that end, if the proposed framework is consistent with function, we aim to be able to predict missing values (fill in the diagonal blocks) without perturbing biological zeros (off-diagonal elements). Figure 3a clearly shows that our approach succeeds in complementing the activity configurations, while respecting the pattern of specificity, thus supporting our hypothesis.

We next assessed the performance of the inference task and compare it with an alternate approach, *scImpute* [7], which is a recent method designed to impute missing values in scRNA-seq. Its main assumption is that missing values are mainly due to low levels of expression. We applied this method to the same data in order to contrast with our results, and to test whether our principle of gene collectivity in co-dependency networks enables a clearer distinction in terms of putative functional activity (see Supplementary Figure S2 for the comparison of methods). Additionally, we quantitatively assessed the performance of the *DECODE*

versus *scImpute* method, by considering diagonal elements as true-positives (TP) and off-diagonal elements as true-negatives (TN) and computing *sensitivity* and *specificity* (Figure 3b). In all cases, our network-based approach outperforms *scImpute*. However, depending on the sparsity of the dataset, the difference in performance is biased in one or the other direction. For dense datasets, such as **MouseBrain**, the difference is mostly observed in improved specificity of the predictions; whereas in the case of sparse datasets, the difference is mainly observed in sensitivity. In other words, if the majority of non-zero markers were known, our approach does not attempt to falsely declare additional non-zeros. At the same time, if non-zeros are sparse, our method accurately identifies them.

Supervised learning of technical zeros using gene dependency network

Thus far, we used our proposed framework to infer the biological state of genes by aggregating the activity pattern of their co-dependent network neighbors. We showed that the inferred state enables discriminating biological and technological zeros, by defining which of the observed zeros have a significantly low probability of activity. Next, we aim to infer the actual expression value of the identified technical zeros or missing values. To this end, we again exploit the collective information embedded in the co-dependency network structure. For each gene, we partition the expression profile into two classes, one with non-zero values and one with missing values. Next, we fit a predictive model (using support-vector regression – ϵ-SVR) on the observed values, using the expression values of the top-ranked network neighbors of the target gene as features, and use it to predict the missing values. In order to evaluate the predictions, we devised the following scenario.

We split rows and columns of a shuffled version of the original expression matrix, creating 25 total

gene-cell blocks. Then, we perform a 5x5-fold cross-validation experiment, where each time we first mask one of the blocks and then predict the expression values for the entire matrix. Since the main purposes of single-cell expression analysis is to identify relationships among cells and/or genes, we use the ability to recover such associations as criteria to evaluate the predicted values. More specifically, we compute gene-gene and sample-sample correlation values for the corresponding block before masking (i.e., original expression values) and after the prediction. Subsequently, we quantify the similarity of the co-expression patterns. Two other existing methods, *scImpute* and *MAGIC*, also aim to predict missing expression values for scRNAseq. We perform this evaluation experiment using these two methods to test whether our framework and logic better recovers gene and cell associations. Note that since the different methods, by design, produce different levels of sparsity after prediction, we only focused on predicted values for non-zeros for the comparison (see Methods).

Figure 4 shows the results of the evaluation. *scImpute* showed the lowest recovery rate for both gene- and cell-level structure. When the dataset is dense, such as the **MouseBrain**, all the methods are able to recover the relationship between cells. However, only *MAGIC* and *DECODE* are capable of high-quality recovery of gene-level relationships. On the other hand, for sparse datasets, the values predicted by our framework still faithfully recover the structure, while those of other methods cannot. This demonstrates that the idea of network-based collective transfer of information underlying our framework enables an accurate and robust prediction of missing expression values.

Application to Type 2 Diabetes (T2D)

We next sought to study the impact of our imputation approach on elucidating disease-related gene modules in specific cell-types, focusing on Type 2 diabetes, a chronic metabolic disorder characterized by altered glucose processing resulting from pancreatic islet damage. To this end, we first constructed a gene-correlation network using our imputed transcriptional profile of human pancreatic cells. We then incorporated a set of cell type-specific, differentially expressed genes, between T2D patients and healthy controls, from a recent study [8] , and extracted the vertex-induced subset of the global pancreas gene correlation network. The distribution of correlation scores exhibits a clear separation near 0.2, which we used as a threshold value to binarize our network (Supplementary Figure S6). We also constructed for the same genes a physical interactome using protein-protein interactions obtained from InWeb_IM [9]. Finally, we used binarized gene correlation network as a mask to select context-specific protein-protein interactions involved in cell type-specific mediation of T2D pathology. We identified three different gene modules that are differentially expressed in T2D and also supported by physical interactions (Figure 5).

Module *M1* includes key elements previously identified as key regulators of a trans-differentiation process from acinar to insulin secreting β-cells [10]. Cadherin-1 (*CDH1*), the central hub in this module, is a calcium-dependent cell-cell adhesion glycoprotein that is necessary for in-vivo pancreatic exocrine-to-endocrine cell transdifferentiation [10]. Consistently, two other genes in the module, *CLDN2* and *NDRG1*, are also associated with cell-to-cell adhesion. This transdifferentiation process is also dependent on the phosphatidylinositol 3-kinase (PI3K) signaling pathway. Interestingly, we found an additional gene (*FER*) within the same module, which encodes a tyrosine-protein kinase that is involved in the insulin receptor signaling pathway via PI3K. From a different point of view, this transdifferentiation module also includes the

early growth response gene (EGR-1), a transcription factor that is rapidly induced in response to glucose in β-cells. This rapid expression change is directly linked to insulin secretion and is also elicited by cyclic AMP (cAMP) agonists [11]. Finally, three additional genes in the module (*FOS, FOSB*, and *CREM*) mediate cAMP response. Collectively, these our results suggest a mechanistic role for this module in the transdifferentiation of exocrine to endocrine cells. On the other hand, its discovery through T2D differential expression, suggest a potential role in disease etiology. The genes in the other two discovered modules are also associated with key processes underlying T2D biology. Module *M2* is involved in glucose-stimulated insulin secretion (GSIS) dependent on MAPK (e.g., TIAM1 and Rac1) [12], whereas module *M3* contains genes that are annotated with insulin resistance (*PPP1CB* and *PPP1R3B*). Overall, these results demonstrate the utility of the information gained through network-based transfer of cell-specific gene activity.

In summary, our framework infers functional expression profiles that can be effectively integrated with common differential and network-based computational analyses to uncover specific gene modules relevant for disease pathobiology and useful for subsequent experimental interrogation.

Methods

Datasets

FANTOM 5 gene panel

Focusing on a curated set of condition-specific genes significantly enhances the accuracy of cell type predictions [13]. Following this logic, a subset of reference genes for mouse and human was defined and used

throughout the study. The RLE normalized CAGE peak-based expression tables for human and mouse were downloaded from FANTOM5 [14] release (11-Nov-2014), from `http://fantom.gsc.riken.jp/5/datafiles/latest/extra/CAGE_peaks`. All gene IDs were converted to standard gene symbols using HUGO [15] and MGI [16] databases. The final dataset contains $18,869$ genes ($1,073$ samples) and $18,852$ genes ($1,829$ samples), in mouse and human, respectively. These samples include primary known cell types, tissues, and cell lines. Each gene was preprocessed individually and retained in the feature set if its expression in any of the samples is at least an order of magnitude greater than the median of the rest of cell types. This resulted in $6,596$ genes for human and $4,875$ genes for mouse.

Single-cell RNA-seq datasets

1. **Mouse Brain[17]:** was obtained from NCBI GEO under accession number GSE82187. Among the $1,208$ single striatal cells assayed, 532 were selected that were measured using *microfluidic single-cell RNAseq (Mic-scRNAseq)* technology and span 4 major brain cell types: *neurons, astrocytes, oligodendrocytes*, and *microglial cells*. The original dataset contained $18,840$ genes (24% non-zeros). After preprocessing and limitting to the FANTOM panel, the final data included $2,334$ genes (32.47% non-zeros).

2. **PBMC:** filtered 4k PBMC dataset was downloaded from the 10x Genomics website (`https://support.10xgenomics.com/single-cell-gene-expression/datasets/pbmc4k`). Cell annotations were transfered from pure cells in a previous study[18] by matching to the closest cell. Different subtypes of *T cells* were collapsed. The dataset spans cells involved in both innate and adaptive immunity, including *monocytes* and *dedritic cells*, as well as T, B, and NK cells. The initial dataset contained $13,704$ genes and $5,825$ samples (9.3% non-zeros). Filtering genes and retaining

genes in FANTOM5 gene panel resulted in $2,398$ genes (5.6% non-zeros).

3. **Pancreas[19]:** processed dataset was downloaded from (`https://shenorrlab.github.io/bseqsc/vignettes/pages/data.html`) [19]. It measures transcriptomic profile of cells from 3 healthy individuals using inDrop. The four most well-characterized cell types, namely *alpha, beta, ductal* and *acinar cells*, were selected. Samples (a total of $4,352$) span both endocrine and exocrine cells . The dataset contained $17,434$ genes (10.7% non-zeros). The final dataset after preprocessing contains $3,437$ genes (5.8% non-zeros).

Gene co-dependency index (CDI)

Co-dependency of gene activity was measured by quantifying how informative observing a non-zero for a gene g_i is with respect to observing a non-zero for a gene g_j. Assume that N represents the total number of samples, $\mathbb{1}_i$ the total number of observed non-zeros for gene i across all samples, and $\mathbb{1}_{ij}$ the number of samples in which both gene i and gene j are jointly active (have non-zero values). Then, $P(g_i = 1)$ will denote the probability of observing a sample with non-zero value for gene i, and it is computed as $\frac{\mathbb{1}_i}{N}$. Similarly, the joint probability distribution for a pair of given genes, g_i and g_j is $P(g_i = 1, g_j = 1) = \frac{\mathbb{1}_{ij}}{N}$, indicating the probability of simultaneously observing both g_i and g_j to be non-zeros. Assume a null model were genes g_i and g_j are independent, in which case $P(g_i = 1, g_j = 1)$ can be factored as $P(g_i = 1)P(g_j = 1)$. The co-dependency of genes is measured based on the deviation of these two terms – i.e., the joint probability from the product of factored marginal distribution. Finally, consider the factorized probability of observing both gene i and gene j active as $\pi_{ij} = P(g_i = 1)P(g_j = 1) = \frac{\mathbb{1}_i \mathbb{1}_j}{N}$, and assume a binomial distribution as a null model.

Putting the previous pieces together, the co-dependency index (CDI) is defined as follows:

$$
\begin{aligned}
CDI &= -log_{10}\left[p_e(g_i, g_j)\right] && (1)\\
p_e(g_i, g_j) &= \sum_{\mathbb{1}_{ij} \leq x \leq N} \mathbf{Bino}(N, x, \pi_{ij}) \\
&= \sum_{\mathbb{1}_{ij} \leq x \leq N} \binom{N}{x} \pi_{ij}^{x} (1 - \pi_{ij})^{(N-x)}, && (2)
\end{aligned}
$$

where $\mathbf{Bino}(N, x, p)$ is the probability of observing x successes in N trials, if the probability of success is p.

Inferring functional activity of genes

Network-based guilt-by-association approach was implemented as follows. A set of significant neighbor genes was selected to act as proxy for gene i and to "vote" for its activity status. This set consists of the genes with the most significant co-dependency pattern with gene i, not exceeding a total of k_{max}. Define this set by $\mathcal{N}(i)$, and the total number of genes with non-zero values within this set with $\mathbb{1}_{\mathcal{N}} = \sum_{s \in \mathcal{N}(i)} \mathbb{1}\big(s \neq 0)\big)$. Denote with $\mathbb{1}_k$ the total number of non-zeros in sample k and let m be the the total number of genes. The evidence of activity of proxy genes is aggregated using the hypergeometric tail, testing the null hypothesis:"$H_0 : g_i =$ biological zero", as follows

$$
\begin{aligned}
p\text{-val}(g_i = \text{biological zero}) &= \sum_{x=\mathbb{1}_{\mathcal{N}(i)}}^{\mathbf{min}(\mathbb{1}_j, m)} \mathbf{Hyge}(x, m, \mathbb{1}_k, |\mathcal{N}(i)|) \\
&= \sum_{x=\mathbb{1}_{\mathcal{N}(i)}}^{\mathbf{min}(\mathbb{1}_j, m)} \frac{\binom{\mathbb{1}_k}{x}\binom{m-\mathbb{1}_k}{|\mathcal{N}(i)|-x}}{\binom{m}{|\mathcal{N}(i)|}}, && (3)
\end{aligned}
$$

where $\mathbf{Hyge}(x, m, k, n)$ represents the hypergeometric probability distribution, which calculates the probability of observing x white balls when sampling n balls from a population of size m, and where there are k white and $m - k$ black balls.

Note that all neighbors of gene i are not equally informative and using all $|\mathcal{N}(i)|$ neighbors might not always be optimal. To remedy this problem, multiple neighborhood sizes were tested, computing for each the minimum hypergeometric (mHG) score. The set size minimizing the mHG score was selected. A dynamic programming approach was subsequently used for multiple testing correction, and to assign a *p*-value to the mHG score. This process was originally proposed for geneset enrichment analysis without prespecifying the number of top-ranked genes [20]. In our experiment, a max degree of 100 was chosen to build the network. For computed activity *p*-values, multiple hypothesis testing was corrected using Family-wise Discovery Rate (FDR) adjustment [21].

Computing expression domain of genes and identifying cell type-specific marker genes

For a gene i in a cell type c, we use Fisher's exact test to assess the deviation of non-zeros in this context from the background null distribution of- all non-zeros. Let N_c show the total number of cells in cell type c, and p_1 be the probability of observing non-zeros in the background distribution. Thus, the expected number of non-zeros in a sample of size N_c is $p_1 \times N_c$. Positive deviation from this expectation, measured using the tail of the hypergeometric distribution, measures the over-enrichment of non-zeros in that cell type. We use a *p*-value of 0.05 to binarize these memberships. Given a total of k cell types, we then assign a binary vector of size k to each gene that represents its *expression domain*. Pairs of genes with similar expression

domain are used here as standard gold pair of genes that are related in a cell type-specific fashion.

On the other hand, to identify cell type-specific marker genes for cell type c, we select all genes whose expression domain has a 1 at position c, and zero everywhere else. This provides a list of markers for each of the constituting cell types in the population, which are used to assess sensitivity/specificity of different methods in predicting true negatives in the single-cell expression profile.

Imputing missing values using ϵ-Support Vector Regression (SVR)

Gene missing values were imputed across samples by regressing over the values of the network neighbors of the gene. Support Vector Regression (SVR)(SVR) [22] is a popular machine learning regression technique, which maximizes the margins around the estimated separator hyperplane from the closest data points on each side. This margin defines an "insensitive" region in which errors are not penalized. In particular, ϵ-SVR defines the insensitive region by employing an ϵ-insensitive loss function, in which errors smaller than ϵ are ignored, and errors passing ϵ are penalized linearly. Formally, ϵ-insensitive loss is defined as follows:

$$\begin{aligned} \boldsymbol{\mathcal{L}}_{\epsilon}^{(\epsilon)}(r_i) &= \mathbf{max}(0, |r_i| - \epsilon) \\ &= \begin{cases} 0, & \text{if } |r_i| \leq \epsilon \\ |r_i| - \epsilon, & \text{otherwise} \end{cases} \end{aligned}$$

Using this loss function, the primal form of ϵ-SVR can be defined as the following convex optimization problem:

$$\mathbf{min}_{\boldsymbol{w}}\{\sum_{i=1}^{m}\mathcal{L}_{\epsilon}(y_i - \boldsymbol{w}^T\boldsymbol{x}_i) + \lambda\boldsymbol{\mathcal{R}}_2(\boldsymbol{w})\}, \tag{4}$$

where $\boldsymbol{\mathcal{R}}_2$ is a $\|\|_2$ regularizer that smooths the weights, while the parameter λ balances the importance of fitness versus regularization. This balance is particularly useful in scRNASeq, where the high number of zeros is prone to produce ill-conditioned predictions. To apply our framework, the expression values were first normalized between [0-1] using the 95% percentile. We chose an $\epsilon = 0.01$, giving a margin of 1% error. The optimal regularizer (λ) was chosen based on cross-validation with exponentially-separated values in the range $\{2^0, 2^1, \cdots 2^{20}\}$ (Figure S5). $\lambda = 2^{10}$ was found as an optimal trade-off between regularization and minimization of the loss function, and it was used throughout the study.

Evaluation of Imputed expression values

A 5x5-fold cross-validation experiment was performed in a shuffled expression matrix, splitted across rows and columns in 25 blocks. In each step, one block was masked, and the resulting matrix was used as input to the different methods for prediction. For each block, two matrices, $\mathbf{C}_{genes}$ and $\mathbf{C}_{cells}$, were created before and after masking. Similarly, two matrices were created after imputation, denoted by $\hat{\mathbf{C}}_{genes}$ and $\hat{\mathbf{C}}_{cells}$, respectively. The conservation of the correlation structure between genes/cells was quantified using a well-established measure for correlation matrix comparison [23]:

$$sim(\mathbf{C}, \hat{\mathbf{C}}) = \frac{\mathbf{trace}(\mathbf{C}\hat{\mathbf{C}}|}{\parallel \mathbf{C} \parallel_F \parallel \hat{\mathbf{C}} \parallel_F} \tag{5}$$

Availability

Codes and datasets are available to download from `http://compbio.mit.edu/projects/DECODE`.

1. Bacher, R. & Kendziorski, C. Design and computational analysis of single-cell rna-sequencing experiments. *Genome biology* **17**, 63 (2016).

2. Vallejos, C. A., Risso, D., Scialdone, A., Dudoit, S. & Marioni, J. C. Normalizing single-cell rna sequencing data: challenges and opportunities. *Nature methods* (2017).

3. Tanay, A. & Regev, A. Scaling single-cell genomics from phenomenology to mechanism. *Nature* **541**, 331–338 (2017).

4. Hicks, S. C., Townes, F. W., Teng, M. & Irizarry, R. A. Missing Data and Technical Variability in Single-Cell RNA- Sequencing Experiments. *bioRxiv* (2017).

5. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* **2008**, P10008 (2008).

6. Cowen, L., Ideker, T., Raphael, B. J. & Sharan, R. Network propagation: a universal amplifier of genetic associations. *Nature Reviews Genetics* (2017).

7. Li, W. V. & Li, J. J. scimpute: accurate and robust imputation for single cell rna-seq data. *bioRxiv* 141598 (2017).

8. Lawlor, N. *et al.* Single-cell transcriptomes identify human islet cell signatures and reveal cell-typespecific expression changes in type 2 diabetes. *Genome Research* **27**, 208–222 (2017).

9. Li, T. *et al.* A scored human proteinprotein interaction network to catalyze genomic interpretation. *Nature Methods* **14**, 61–64 (2016).

10. Minami, K., Okano, H., Okumachi, A. & Seino, S. Role of Cadherin-mediated Cell-Cell Adhesion in Pancreatic Exocrine-to-Endocrine Transdifferentiation. *Journal of Biological Chemistry* **283**, 13753–13761 (2008).

11. Josefsen, K., Sørensen, L. R., Buschard, K. & Birkenbach, M. Glucose induces early growth response gene (Egr-1) expression in pancreatic beta cells. *Diabetologia* **42**, 195–203 (1999).

12. Veluthakal, R., Madathilparambil, S. V., McDonald, P., Olson, L. K. & Kowluru, A. Regulatory roles for Tiam1, a guanine nucleotide exchange factor for Rac1, in glucose-stimulated insulin secretion in pancreatic β-cells. *Biochemical Pharmacology* **77**, 101–113 (2009).

13. Li, H. *et al.* Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. *Nature Genetics* **49**, 708–718 (2017).

14. Lizio, M. *et al.* Gateways to the FANTOM5 promoter level mammalian expression atlas. *Genome Biology* **16**, 22 (2015).

15. Gray, K. A., Yates, B., Seal, R. L., Wright, M. W. & Bruford, E. A. Genenames.org: the HGNC resources in 2015. *Nucleic acids research* **43**, D1079–85 (2015).

16. Blake, J. A. *et al.* Mouse Genome Database (MGD)-2017: community knowledge resource for the laboratory mouse. *Nucleic Acids Research* **45**, D723–D729 (2017).

17. Gokce, O. *et al.* Cellular Taxonomy of the Mouse Striatum as Revealed by Single-Cell RNA-Seq. *Cell Reports* **16**, 1126–1137 (2016).

18. Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nature Communications* **8**, 14049 (2017).

19. Baron, M. *et al.* A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. *Cell Systems* **3**, 346–360.e4 (2016).

20. Eden, E., Navon, R., Steinfeld, I., Lipson, D. & Yakhini, Z. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC bioinformatics* **10**, 48 (2009).

21. Yekutieli, D. & Benjamini, Y. Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *Journal of Statistical Planning and Inference* **82**, 171–196 (1999).

22. Cortes, C. & Vapnik, V. Support-vector networks. *Machine Learning* **20**, 273–297 (1995).

23. Herdin, M., Czink, N., Ozcelik, H. & Bonek, E. Correlation Matrix Distance, a Meaningful Measure for Evaluation of Non-Stationary MIMO Channels. In *IEEE 61st Vehicular Technology Conference*, vol. 1, 136–140 (IEEE, 2005).

Acknowledgements This work has been supported through NIH Grants 5R01AI114814-02 and U01-CA198941 and NSF grant CCR-0939370.

Competing Interests The authors declare that they have no competing financial interests.

Correspondence Correspondence and requests for materials should be addressed to S.M. (email: atcg@mit.edu) or M.K. (email: manoli@mit.edu).

Authorship M.K. and A.G. and co-last author on the paper.

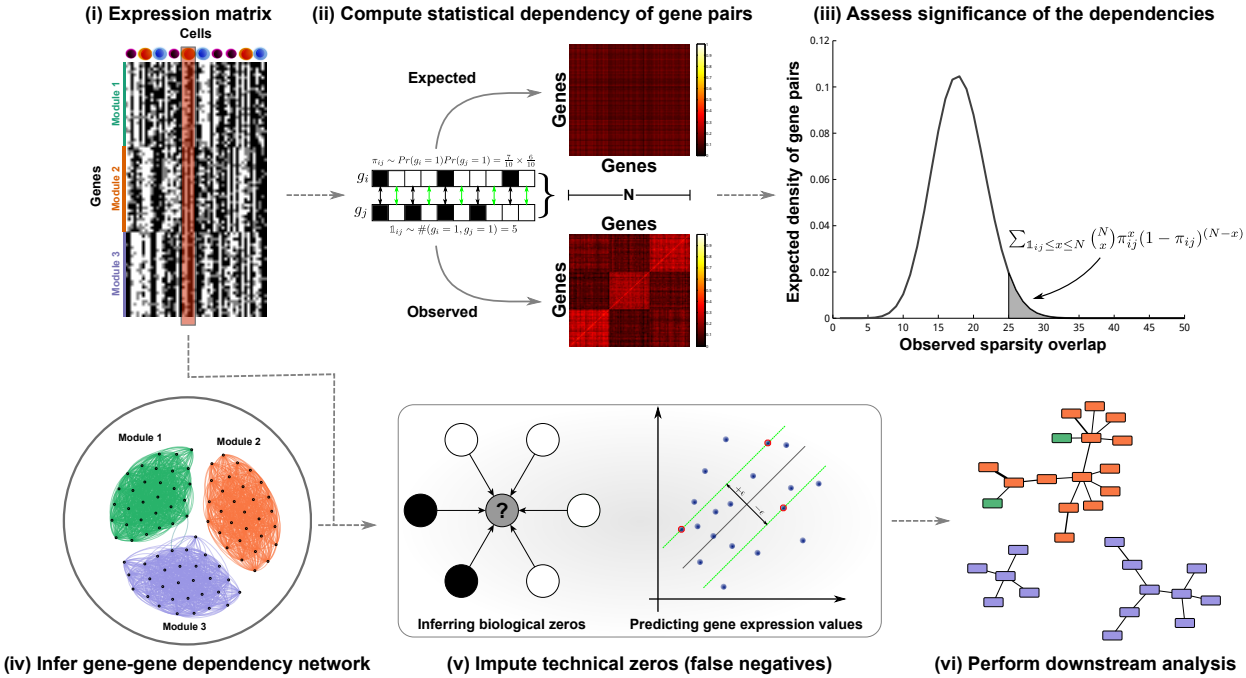
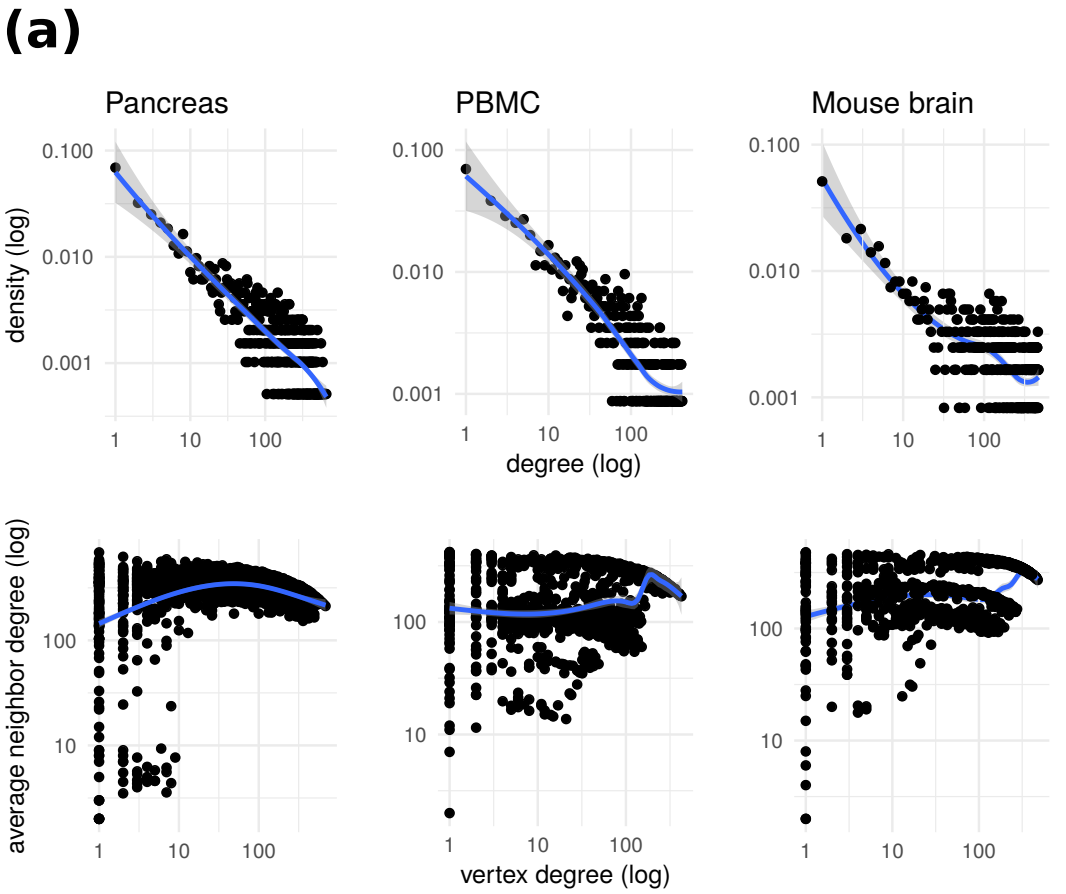


Figure 1: **Summary of our approach**



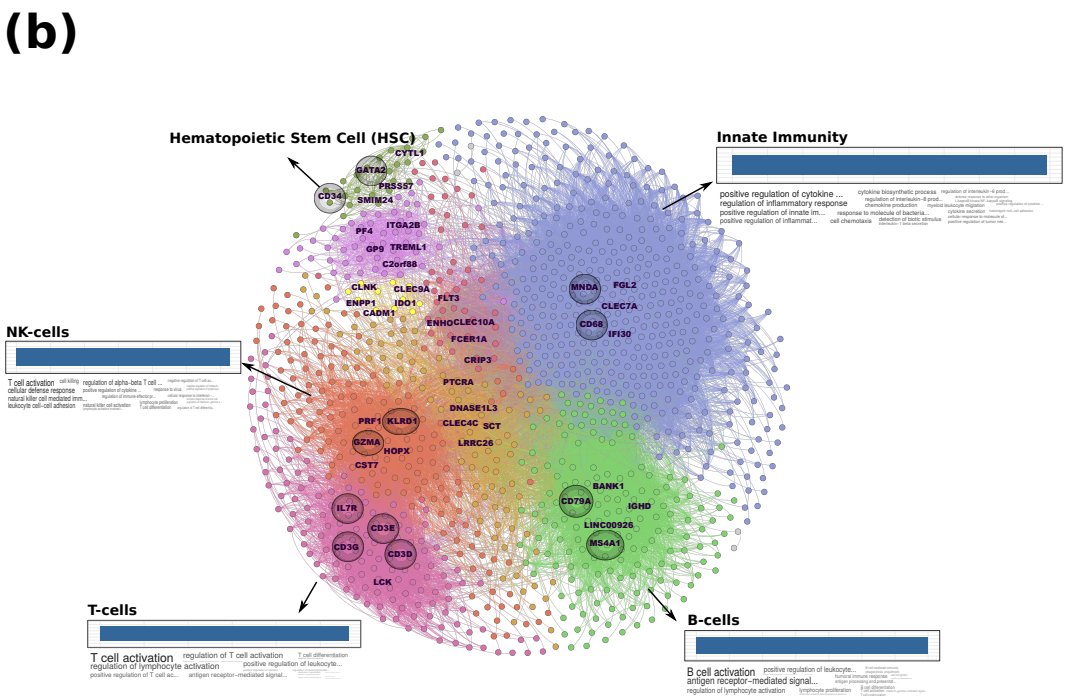


Figure 2: **Characterization of gene-gene co-dependency network**

Figure 2 *(previous page)*: (a) Degree distribution and the neighbor connectivity preferences of genes with different degrees, (b) Co-dependency network of the PBMC dataset. Each module is color coded accordingly and top four largest modules are annotated with their enriched functions. These modules coincide with the underlying cell types in the PBMC dataset. Top-5 high-degree nodes within each module are marker. Many of these genes (marked with shaded circle) are well-established cell type markers.

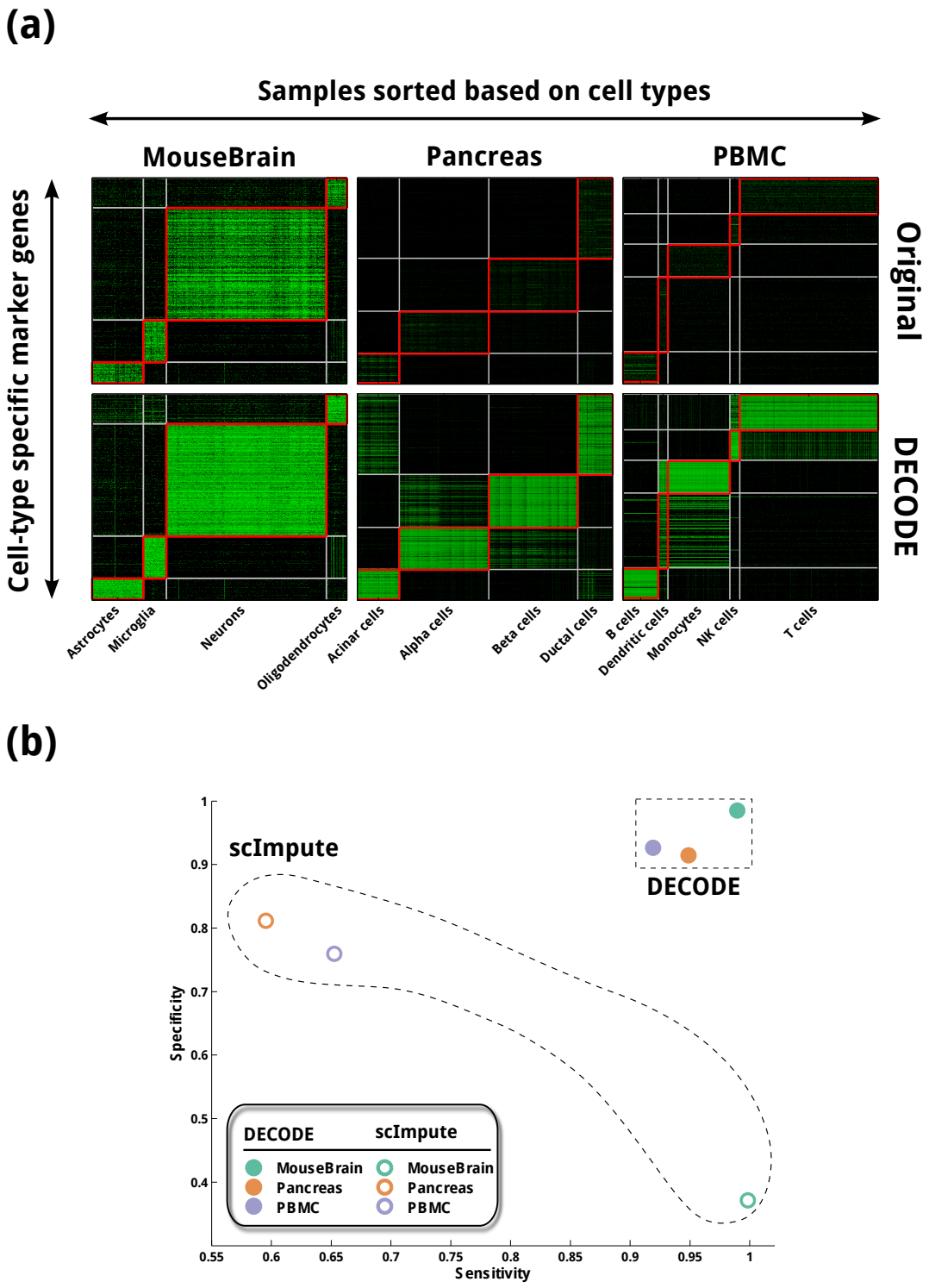


Figure 3: **Performance of network-based approach in predicting the functional activity of genes within the subspace of markers.**

Figure 3 *(previous page)*: (a) Distribution of non-zeros in the original expression matrix, restricted to the subspace of marker genes, and after applying DECODE. Block diagonal elements, marked with red rectangles, show matching markers/cell types. Off-diagonal entries are treated as true biological zeros, or true negatives. (b) Sensitivity and specificity of DECODE versus scImpute. Our method outperforms scImpute in terms of both measure in all datasets.

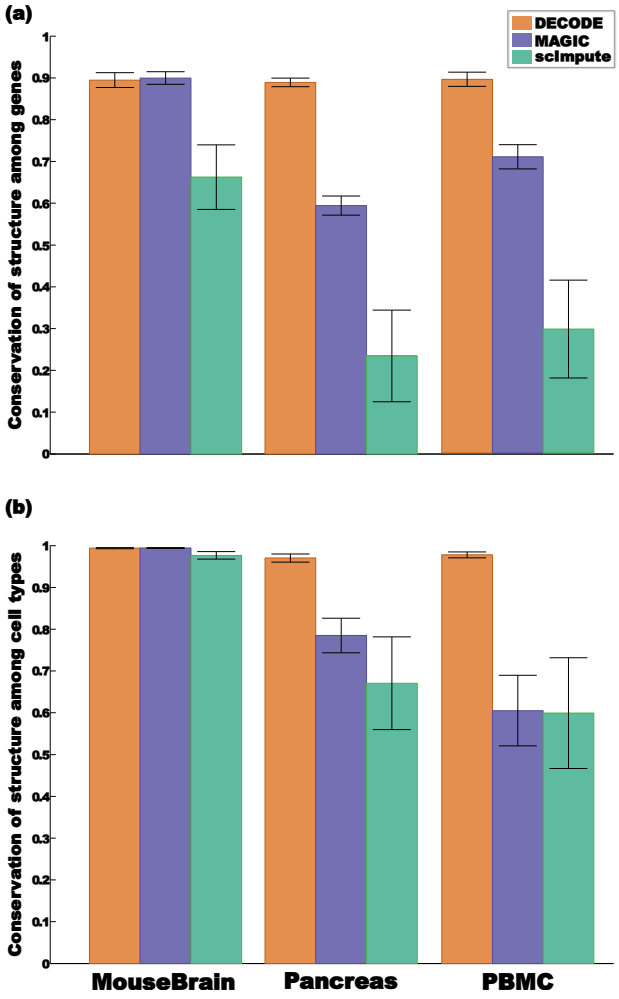


Figure 4: **Cross-validation missing value imputation using 5x5 fold cross-validation.** We randomly permuted rows and columns of the expression matrix and split each dimension into 5 equally-size sections. Then, we masked each block at a time and predicted its values using the rest of matrix with different methods. Finally, we assessed the ability of predictions from each method to reconstruct the underlying correlation structure between genes and cells.

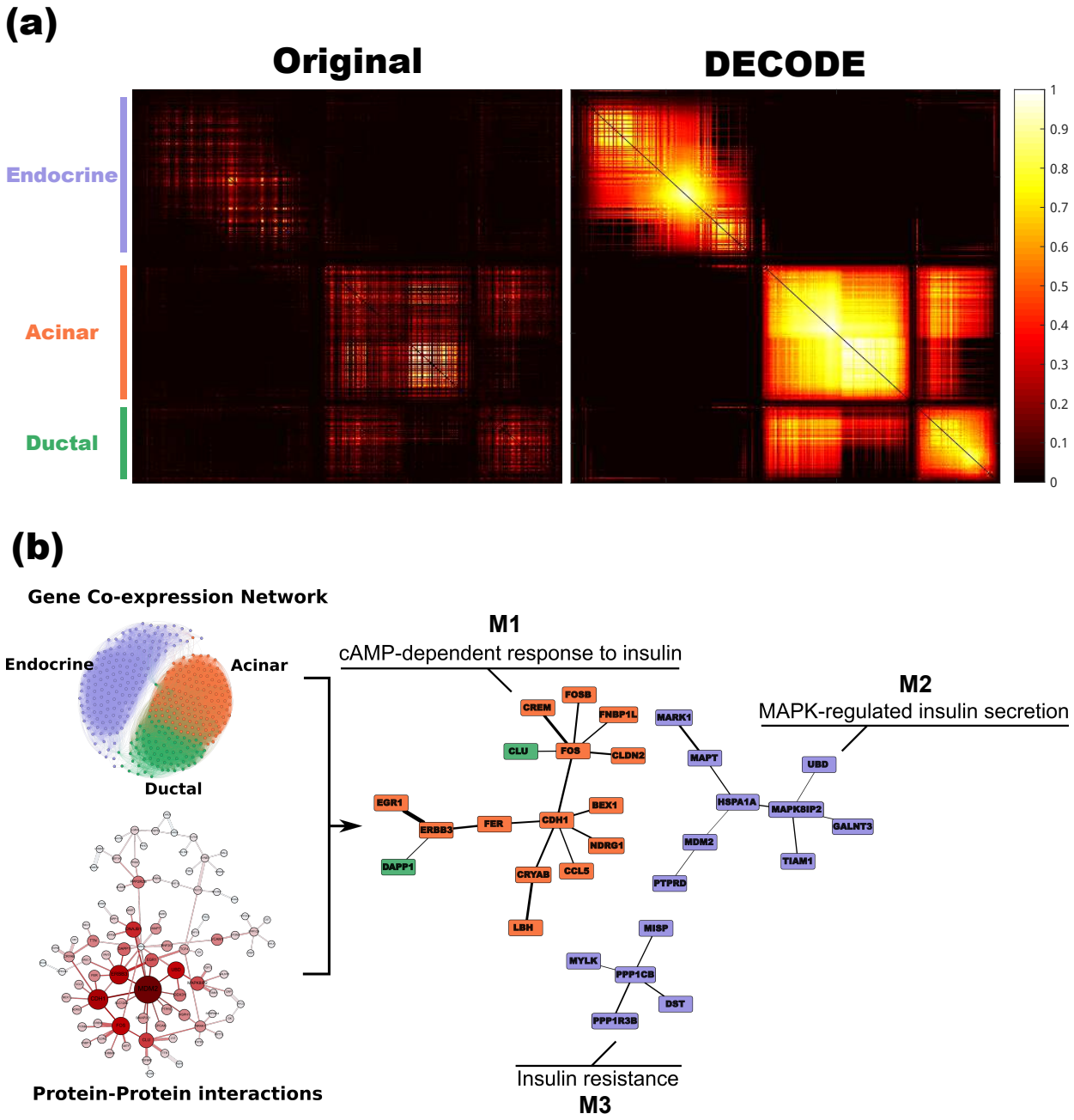


Figure 5: **Combining Gene Co-expression Network and Protein-Protein Interactions to identify T2D-specific pathways.**

Figure 5 *(previous page)*: (a) Gene co-expression network among T2D-associated/cell type-specific gene, constructed using the original expression profile and the one imputed using DECODE framework. (b) Combining the co-expression network computed from imputed profile with the protein-protein interactions (PPIs) downloaded from the InWeb_IM database. We used co-expressed genes as a filter to select for active/context-specific PPIs. We identify three different pathways that are directly related to insulin signaling, secretion, and resistance.

Supplementary Figures

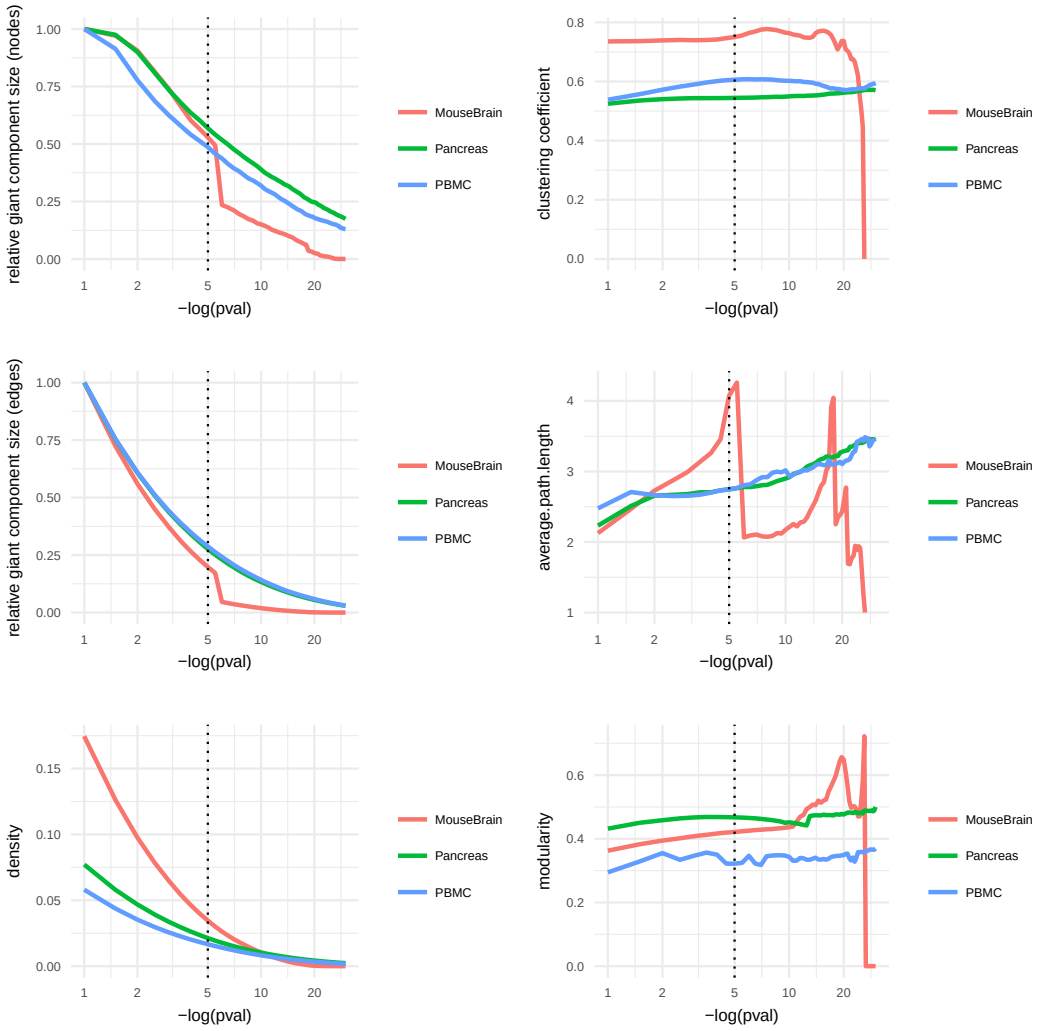


Figure S1: **Different network properties as a function of dependency *p*-value.** We tried different *p*-value thresholds, ranging from 10^{-1} to 10^{-30} and traced six different network measures characteristics of complex networks as a function of this threshold. We found threshold value of 10^{-5} to be a good cutoff, as the three networks largely converge to a similar size and degree of clustering without displaying abrupt shifts.

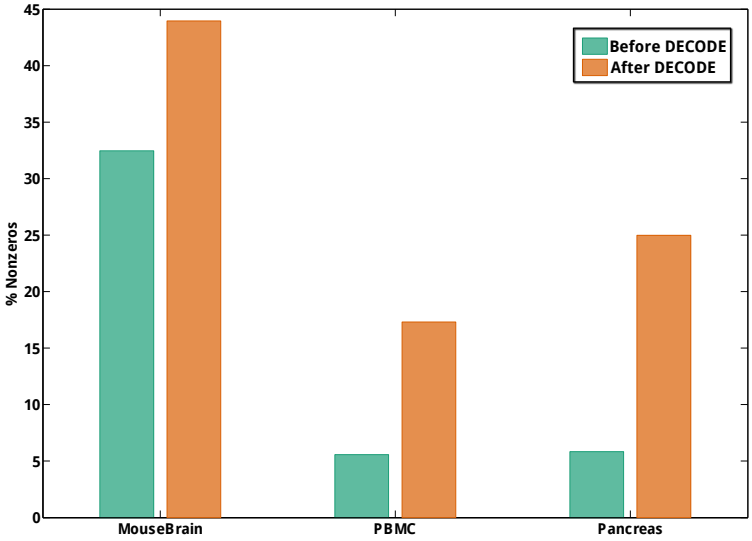


Figure S2: **Percent of non-zero values before and after DECODE.** For each dataset, percent of imputed non-zeros are different. For the **MouseBrain** dataset, which is relatively dense to begin with, we observe close to 50% increase in number of non-zeros. For the other two datasets, which are much sparser, we observe 3-5 fold increase in the number of non-zeros.

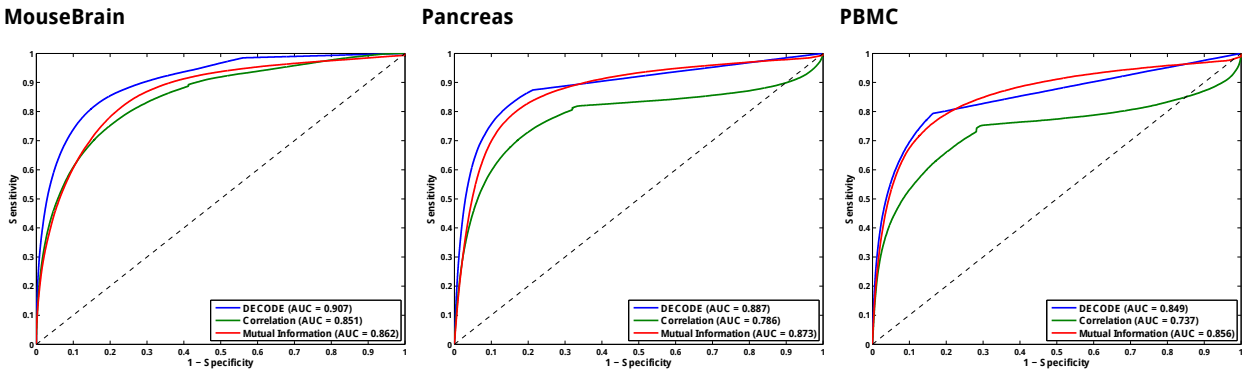


Figure S3: **Comparison of CDI versus other measures.** Given the known cell types, gene pairs with similar expression domains (expressed in similar cell types) are considered as true-positive set. For each measure, we computed the area under the curve (AUC) of the ROC curve as an indicator of performance. Pairwise gene correlation performs worse in all datasets. DECODE outperforms the other two measures in two out of three datasets, while performs marginally worse in the **PBMC** dataset. In all datasets, especially **Pancreas** and **PBMC**, DECODE plot flattens out, which is due to pairs of gene that are deemed insignificant (with p-value 1).

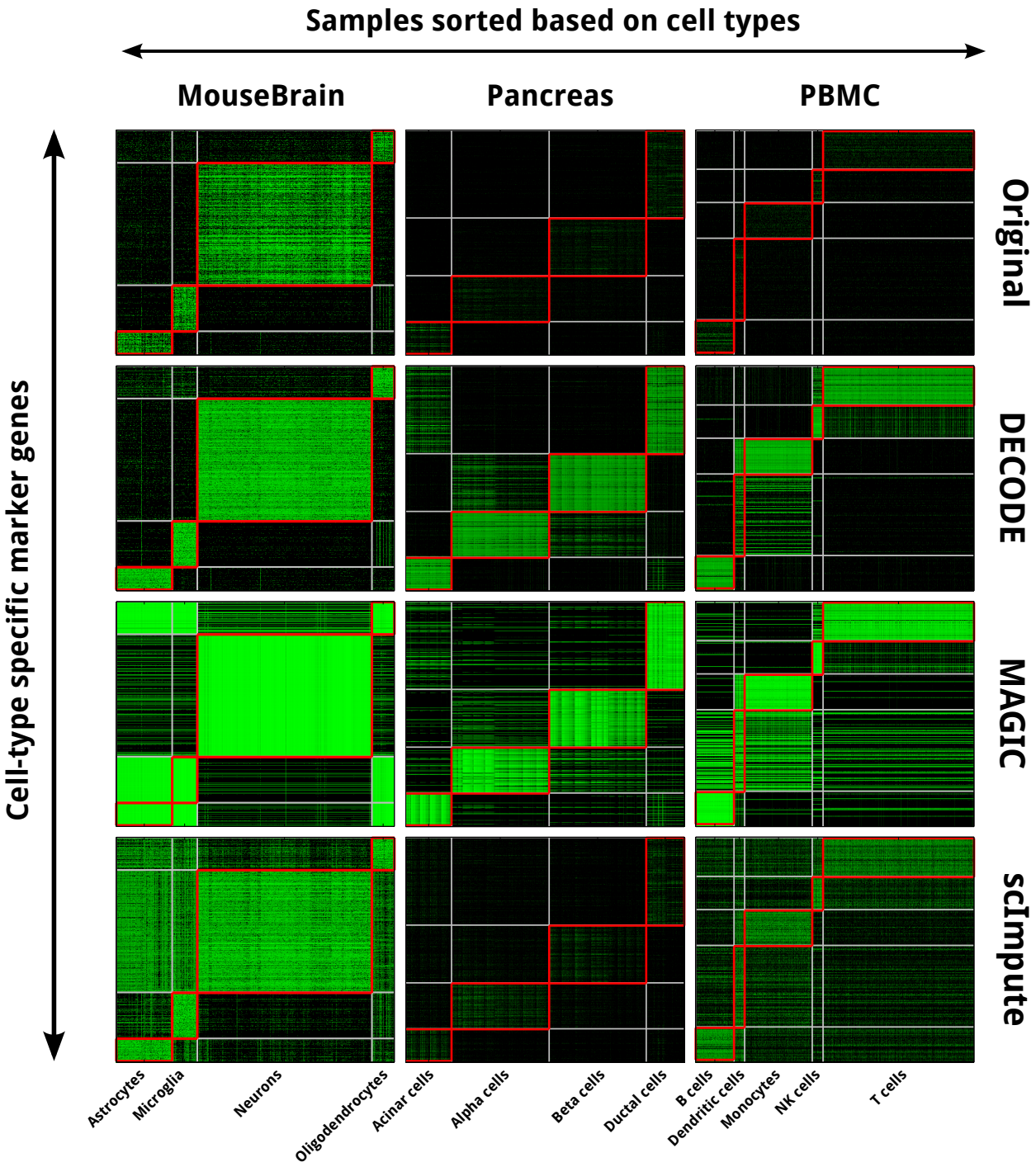


Figure S4: **Expression of marker genes in different cell types after imputation.**

Figure S4 *(previous page)*: Rows correspond to cell type-specific marker genes and columns are cells, sorted according to their cell types. Ideally, all markers should lie on the diagonal blocks, while off-diagonal elements are treated as true biological zeros. *MAGIC* does not have a mechanism to distinguish between biological versus technical zeros, thus it assigns a value to every zero entities in the expression matrix. *scImpute* only assigns values that it deems to be technical zeros. However, these values are over-estimated in the **MouseBrain** dataset, while are highly under-estimated in the other two datasets.

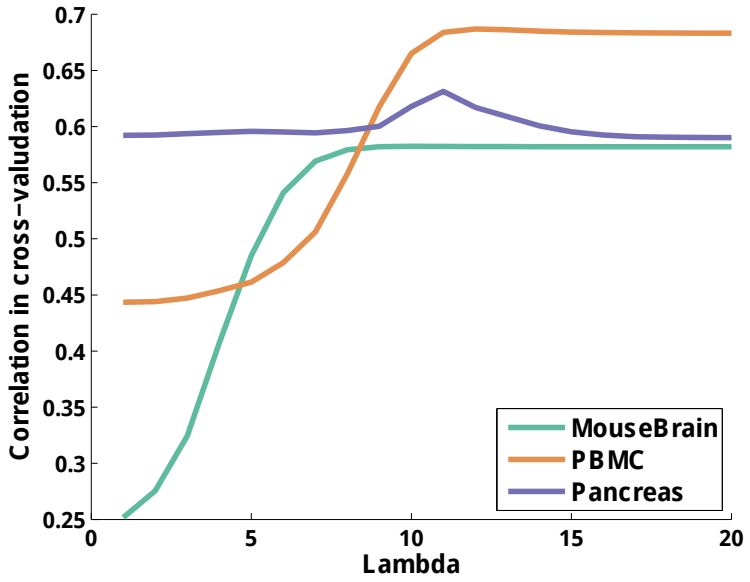


Figure S5: **Cross-validation to find optimal regularizer parameter.** We tried different values of λ, the weight of $\mathcal{L}_2$-regularization in SVR, and for each case we measured the correlation of predicted values with actual values. X-axes is the log_2 of lambda. Values of λ close to 2^{10} show a good balance of regularization-versus-loss minimization.

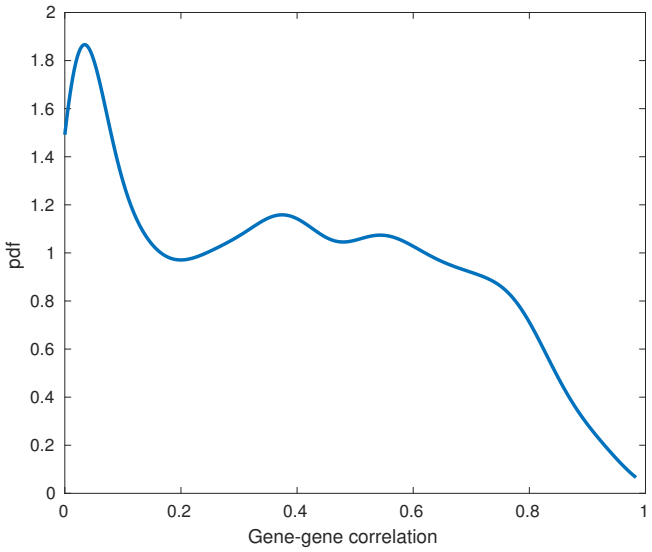


Figure S6: **Distribution of gene-correlation scores in the subnet of T2D genes.** Distribution of positive correlations has been smoothed using Gaussian kernel and visualized. There is a population shift around 0.2, which we used for thresholding T2D gene-correlation subnetwork.

Supplementary Tables

Measures	Networks		
	Pancreas	**PBMC**	**Mouse Brain**
nodes	1954	1146	1211
edges	125605	47907	94079
mean degree	128.56	83.61	155.37
clustering coef.	0.54	0.61	0.75
mean shortest path	2.75	2.74	4.07
diameter	10	7	10
density	0.07	0.07	0.13

Table S1: **Networks topological statistics.** Main topological measures characteristics of complex networks after thresholding dependency networks constructed by DECODE.

Supplementary Materials

Supplementary file 1

This zip file contains co-dependency network between selected genes in each of three datasets in this study

Supplementary file 2

This zip file contains network modules for co-dependency networks.

Supplementary file 3

This zip file contains list of cell type specific markers identified in this study.

Supplementary file 4

This zip file contains predicated gene expression matrices and their corresponding *p*-value matrix for being non-zeros.