MuDCoD: Multi-Subject Community Detection in Personalized Dynamic Gene Networks from Single Cell RNA Sequencing

Ali O. B. Şapcı[1], Shan Lu[2], Oznur Tastan[1⋆], and Sündüz Keleş[2,3⋆]

[1] Faculty of Engineering and Natural Sciences, Sabancı University, Istanbul, Turkey
[2] Department of Statistics, University of Wisconsin, Madison, WI, USA
[3] Department of Biostatistics and Medical Informatics, University of Wisconsin, Madison, WI, USA

Abstract. Developments in single-cell RNA sequencing (scRNA-seq) advanced our understanding of transcriptional programs of different cell types and cellular stages at the individual cell level. Single-cell RNA-seq datasets across multiple individuals and time points are now routinely generated for different conditions. Analysis of personalized dynamic gene networks constructed from these datasets could unravel subject-specific network-level variation critical for phenotypic differences. While there have been developments in the gene module discovery methods on networks estimated from scRNA-seq data, these have mostly focused on static gene networks. In this work, we develop MuDCoD to cluster genes in personalized dynamic gene networks and identify gene modules that vary not only across time but also among subjects. To this end, MuDCoD extends the global spectral clustering framework of the previously developed method, PisCES, to promote information sharing among the subject as well as the time domain. Our computational experiments across a wide variety of settings indicate that, when present, MuDCoD leverages shared signals among networks of the subjects, and performs robustly when subjects do not share any apparent information. An application to human-induced pluripotent stem cell scRNA-seq data during dopaminergic neuron differentiation indicates that MuDCoD enables robust inference for identifying time-varying personalized gene modules. Our results illustrate how personalized dynamic community detection can aid the exploration of subject-specific biological processes that vary across time. MuDCoD is publicly available at https://github.com/bo1929/MuDCoD.

Keywords: Single cell analysis · Dynamic community detection · Gene modules · Spectral clustering.

1 Introduction

The advent of single-cell RNA sequencing has provided unparalleled insights into the transcriptional programs of different cell types, and cellular stages at the individual cell level [2, 19, 21]. Single-cell RNA-seq datasets across multiple individuals and time points (or conditions) are becoming increasingly available [7, 8, 14]. These datasets enable constructing personalized dynamic gene networks that vary across time. Community structures emerge in these dynamic networks when there is strong local clustering of genes that are synchronized to function together. Unraveling the differentiation dynamics or deciphering pathways affected in various diseases requires discovering which genes cooperate in specific cellular processes. Thus, a key inference from the analysis of gene networks is detecting time/condition varying gene communities that might correspond to genes acting together in various biological processes. While these gene communities change over time due to the dynamic nature of the relations among the genes in the processes, they may also vary among different subjects. Variation among subjects at the network-level, when combined with subject-level phenotype information, can reveal critical information such as the gene modules driving specific phenotypes. Thus, identifying gene modules by taking into account the personalized nature of the gene modules is of paramount importance.

Gene module detection methods for scRNA-seq data have mostly focused on the construction of static gene networks that capture a snapshot of time or a developmental epoch [3, 11]. Recently, Liu et al. [13] developed a biologically and computationally appealing global spectral clustering framework named PisCES for inferring

⋆ Co-corresponding authors.

persistent communities in dynamic networks. PisCES detects gene modules at each time point by imposing module smoothness across time. They show that such information sharing across the time domain improves community detection in dynamic networks [13]. However, PisCES does not accommodate joint analysis of personalized networks from multiple subjects, potentially overlooking information that could boost the signal for community detection. Here, we generalize the PisCES framework to cluster personalized dynamic gene networks and identify modules that vary not only across time but also among subjects. Our method Multi-subject Dynamic Community Detection (MuDCoD) infers such modules by extending the smoothness assumption over subjects.

We evaluated MuDCoD with simulation experiments in a wide variety of scenarios. To this end, we extended the Dynamic Degree Corrected Block Model (Dynamic DCBM) [15], which provides a setting for time-varying networks, to multi-subject scenarios. This multi-subject Dynamic DCBM setting accommodates the heterogeneity of scRNA-seq-based personalized gene networks not only across the time but also the subject dimension. We observed that MuDCoD outperforms PisCES when subjects share information and performs robustly otherwise. We applied MuDCoD on a scRNA-seq study of long-term human induced pluripotent stem cell (iPSC) across multiple donors (i.e., subjects) [8]. Our results highlight that MuDCoD is able to promote information sharing across subjects and infers gene modules that contribute to network variation across subjects and time points. Downstream gene set enrichment analysis of inferred modules highlights persistent biological processes across subjects as well as biological processes that are specific to subsets of subjects. A detailed analysis leveraging the differentiation efficiency of subjects reveals that the gene modules of high differentiation efficiency subjects tend to exhibit higher variation across the differentiation time points compared to subjects with low differentiation efficiency.

2 Methods

2.1 Problem Formulation

We define a multi-subject dynamic gene co-expression network for discrete time steps $t = 0, \ldots, T-1$ and for subject $s = 0, \ldots, S-1$ as a time series of undirected and unweighted graphs $\mathcal{G}_{s1}, \ldots \mathcal{G}_{sT}$ for each subject s. Given a multi-subject dynamic gene co-expression network, the key task is to infer the *community structure* for each time point and subject. The community structure of a gene co-expression network of size G is defined as a partition of genes into K disjoint cohesive modules/subsets, where K is a hyperparameter determining the number of communities. Each community is essentially a group of gene nodes densely connected inside and loosely connected outside the community. The problem structure is shown in Fig. 1. In what follows, we use "community" and "module" interchangeably.

2.2 Promoting Signal Sharing Simultaneously in Subject and Time Dimensions

Let $A_{s1}, \ldots, A_{s(T-1)}$ denote a series of symmetric $G \times G$ adjacency matrices of networks $\mathcal{G}_{s1}, \ldots, \mathcal{G}_{sT}$ varying across time, $t = 1, \ldots, T-1$, for subjects $s = 0, \ldots, S-1$. Let L_{st} denote the degree-normalized Laplacian of A_{st} as defined by:

$$L = D^{-1/2} A D^{-1/2} \text{ where } D_{i,j} = \begin{cases} \deg(v_i) & \text{if } i = j, \\ 0 & \text{otherwise,} \end{cases} \tag{1}$$

where $\deg(v_i)$ denotes the degree of node i. Let K be fixed, and $V_{st} \in \mathbb{R}^{G \times K}$ denote a matrix with columns corresponding to the K leading eigenvectors of L_{st}. A baseline strategy for inferring communities from these adjacency matrices is finding communities separately at each snapshot of time step and for each individual separately. To detect communities that dynamically change longitudinally, PisCES [13] applies smoothing to the eigenvectors of L_{st} across time. We develop a novel extension of this framework that we name Multi-subject Dynamic Community Detection (MuDCoD). MuDCoD applies eigenvector smoothing across both the subject and the time dimensions to promote signal sharing across the subjects in addition to the time dimension. Let $U_{st} = V_{st} V_{st}^T$ be the projection matrix onto the column space of V_{st} and define

$$\left[\bar{U}_t\right]_{ij} = \frac{1}{S} \sum_{s=0}^{S-1} \left[\overline{U}_{st}\right]_{ij}, \tag{2}$$

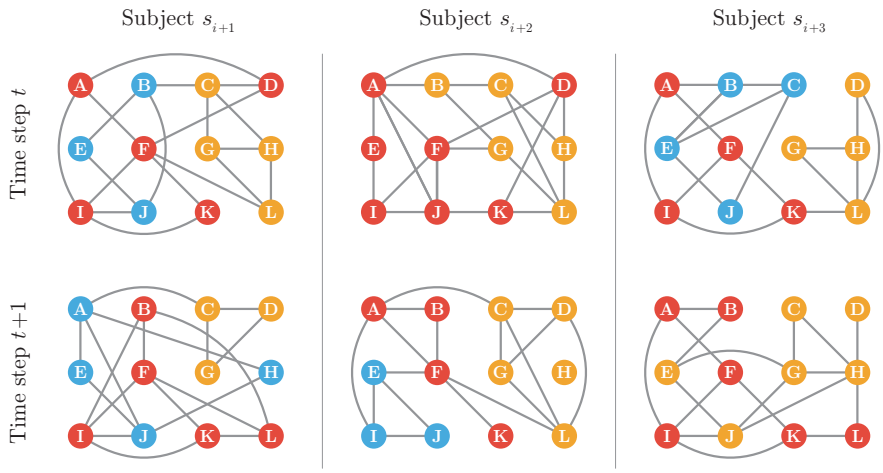


Fig. 1. A schematic for multi-subject dynamic gene networks. A gene network is observed for each subject at each time step among a common set of nodes. The sets of edges vary among both the subjects and the time steps. These networks are estimated from scRNA-seq data and are expected to harbor communities that are conserved at varying levels among the subject and time dimensions. Different colors mark distinct communities where the nodes (genes) within the same communities are depicted with the same color. MuDCoD assumes that communities change smoothly across both the subject and the time dimensions.

where $\overline{U}_{st}$ is the smoothed version of U_{st}. In order to estimate $\overline{U}_{st}$, we propose the following optimization problem:

$$\begin{aligned} \min_{\substack{\overline{U}_{st} \\ s=0,\ldots,S-1 \\ t=0,\ldots,T-1}} & \sum_{t=0}^{T-1} \left(\|U_{st} - \overline{U}_{st}\|_F^2 + \beta \|\overline{U}_{st} - \bar{U}_t\|_F^2 \right) + \alpha \sum_{t=0}^{T-2} \|\overline{U}_{st} - \overline{U}_{s(t+1)}\|_F^2 \\ \text{subject to } & \overline{U}_{st}, \in \left\{ VV^T \ : \ V \in \mathbb{R}^{G \times K}, \ V^T V = I \right\} \ \forall s, \forall t. \end{aligned} \tag{3}$$

Here, the penalty term with parameter α enforces smoothness over the time dimension, whereas the term with parameter β constrains subject-specific variation from the mean time-dependent projection matrix $\bar{U}_t$. The solution of Eq. 3 yields a series of smoothed mean projection matrices for each subject. Similar to PisCES, we propose to solve this non-convex optimization problem with the following iterative method:

$$\begin{aligned} \overline{U}_{s0}^{\ell+1} &= \Pi_K \left(U_{s0} + \alpha \overline{U}_{s1}^{\ell} + \beta \bar{U}_0^{\ell} \right) \\ \overline{U}_{s(T-1)}^{\ell+1} &= \Pi_K \left(\alpha \overline{U}_{s(T-2)}^{\ell} + U_{s(T-1)} + \beta \bar{U}_{T-1}^{\ell} \right) \\ \overline{U}_{st}^{\ell+1} &= \Pi_K \left(\alpha \overline{U}_{s(t-1)}^{\ell} + U_{st} + \alpha \overline{U}_{s(t+1)}^{\ell} + \beta \bar{U}_t^{\ell} \right), \quad t = 1, \ldots, T-2 \\ \overline{U}_{st}^{0} &= U_{st}, \qquad\qquad\qquad\qquad\qquad\qquad t = 0, \ldots, T-1, \end{aligned} \tag{4}$$

where $s = 0, \ldots, S-1$ for all iterations ℓ. The mapping $\Pi_K(M)$ extracts the K leading eigenvectors, and is given as follows for a matrix M:

$$\Pi_K(M) = \sum_{k=1}^{K} v_k v_k^T, \tag{5}$$

where $v_1, \ldots, v_k$ are the K leading eigenvectors of M. This allows model order K to be unknown and possibly varying over time. We utilize the eigengap statistics to select the number of modules, and α and β are chosen with a re-sampling-based cross-validation strategy by [12] as we discuss in the next section. Appendix Fig. S2 illustrates the convergence status for different α and β values.

2.3 Hyperparameter Selection

Selecting the number of communities $\mathcal{K}$: We allow the number of communities, K, to be unknown and possibly varying over time by exploiting the eigengap statistics. Results of Shen and Cheng [17] demonstrated that the degree-normalized Laplacian matrix and the correlation matrix significantly outperform the adjacency matrix, the standard Laplacian matrix, and the modularity matrix at identifying the community structure of networks. Therefore, we use eigenvalues of degree-normalized Laplacian matrix L to detect K as we describe in Section S1.1 of Appendix.

Cross-validation to tune α and β: We adapt the re-sampling based network cross-validation strategy of Li et al. [12] to tune the penalty parameters α and β. This approach splits node pairs rather than nodes to avoid deleting edges and changing the network structure. We apply a grid search strategy to find the best combination of the tuning parameters α and β, where we define the best with the DCBM likelihood function. Further details on this strategy are available in Section S1.2 of Appendix.

3 Experimental Set up

3.1 Simulations

To evaluate MuDCoD, we extended the dynamic degree corrected block model (Dyn-DCBM) simulation set up [16] to multi-subject setting (multi-subject dynamic degree corrected block model (MuS-Dyn-DCBM)). This extension captures the heterogeneity of scRNA-seq-based personalized gene networks both in time and subject dimensions. In our computational experiments, we simulated modules from this model and assessed the performances of different methods in recovering the true modules.

The entries of the adjacency matrix A_{st} under MuS-Dyn-DCBM is generated by the following Bernoulli distribution:

$$\begin{aligned} [A_{st}]_{ij} \sim &\text{Bernoulli}\left(\Psi_i^{(s,t)}\Psi_j^{(s,t)} B^{(s,t)}_{z_j^{(s,t)} z_j^{(s,t)}}\right), \\ &i,j \in [G], \quad j > i, \quad t = 0,\ldots,T-1, \quad s = 0,\ldots,S-1, \end{aligned} \tag{6}$$

with $[A_{st}]_{ij} = [A_{st}]_{ji}$. Here $z^{(s,t)} \in [K]^G$ are vectors of class labels. More specifically, i-th entry $z_i^{(s,t)}$ of $z^{(s,t)}$ denotes the module assignment of gene i for subject s at time point t. $\Psi^{(s,t)} \in \mathbb{R}^G$ are degree parameters and $B^{(s,t)} \in [0,1]^{K\times K}$ is a connectivity matrix at time t for subject s. As we describe below with specific cases, variations on the generating process of $z_i^{(s,t)}$ as a function of s and t provide a systematic way of modulating shared similarities across the subject and time dimensions.

The degree parameters and the connectivity matrix ($\Psi^{(s,t)}$ and $B^{(s,t)}$) for subject s at time t are randomized as follows:

$$B_{lk} = \begin{cases} \text{Uniform}(p_{\text{in}}^{(0)},\, p_{\text{in}}^{(1)}) & \text{if } l = k, \\ p_{\text{out}} & \text{if } l \neq k, \end{cases} \tag{7}$$

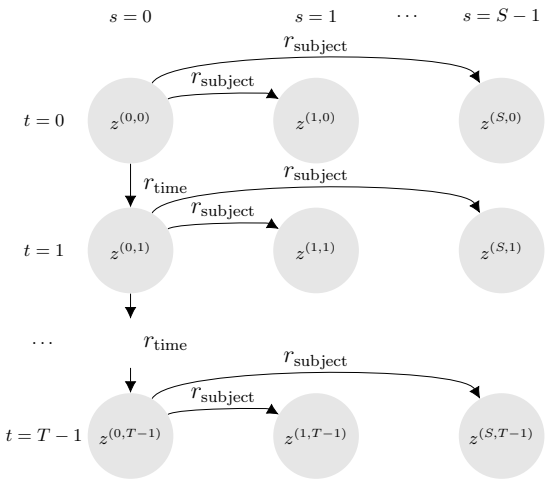
$$\Psi^{(s,t)} = \gamma_0 + \gamma_1 \pi^{(s,t)}/G, \tag{8}$$

where $p_{in} = (p_{in}^{(1)}, p_{in}^{(2)})$ and p_{out} are in-cluster and between-cluster density parameters, respectively, π is a random permutation of 1:G, and γ_0 and γ_1 are determined based on the value of G and the desired degree distribution. For the baseline network ($s = 0$, $t = 0$), class labels are initialized with the following Multinomial distribution:

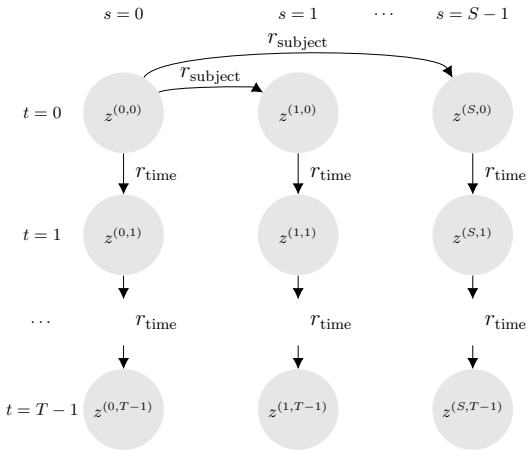
$$z_i^{(0,0)} \sim \text{Multinomial}\left(\frac{1}{K},\ldots,\frac{1}{K}\right). \tag{9}$$

Our extension MuS-Dyn-DCBM spans two distinct multi-subject settings: Signal-Sharing-over-Subjects (SSoS) and Signal-Sharing-over-Time (SSoT) (Fig. 2). Both of these vary the degree of similarity between subject-specific networks while maintaining their time-dependent components and developing corresponding

generating processes for class labels across the time and subject dimensions. Signal-Sharing-over-Subjects (SSoS) enables modulation of signal sharing across subjects. For both settings, each subject's network at time t is generated based on the corresponding module labels.



(a) *SSoS setting:* subjects evolve from a common ancestor at each time step t; and only the ancestor's evolution over time is parameterized.



(b) *SSoT setting:* subjects evolve from a common ancestor at $t = 0$; and then they evolve independently over time.

Fig. 2. Multi-subject dynamic degree corrected block models (MuS-Dyn-DCBM) for the two proposed settings.

Signal-Sharing-over-Subjects(SSoS): In SSoS, gene modules at baseline progress over time with controllable similarity among time points, and each subject's set of gene modules at time t is a realization from this progressing baseline (Fig. 2a). Let $s = 0$ index the subject for the baseline network, and z_{0t} denote its class labels for $t = 1, \ldots, T-1$. In SSoS, z_{0t} progresses over time according to

$$z_i^{(0,t)} = \begin{cases} z_i^{(0,t-1)} & \text{with probability } 1 - r_{\text{time}}, \\ \text{Multinomial}(\frac{1}{K}, \ldots, \frac{1}{K}) & \text{otherwise,} \end{cases} \quad \text{for } i \in [G], \quad t = 1, \ldots, T-1, \tag{10}$$

where r_{time} denotes the probability that node i changes its previous class label, $z_i^{(0,t-1)}$. Then, class labels of other subjects for each time step t are generated as follows:

$$z_i^{(s,t)} = \begin{cases} z_i^{(0,t)} & \text{with probability } 1 - r_{\text{subject}}, \\ \text{Multinomial}(\frac{1}{K}, \ldots, \frac{1}{K}) & \text{otherwise,} \end{cases} \quad \text{for } i \in [G], \quad t = 0, \ldots, T-1, \quad s = 1, \ldots, S-1, \tag{11}$$

where r_{subject} denotes the probability that node i changes its class label while progressing from $z_i^{(0,t)}$.

Signal-Sharing-over-Time (SSoT): In contrast to SSoS, in SSoT, each subject's gene modules start out from a common network at time $t = 0$ and progress over time independent of other subjects. In this setting, gene modules of a subject share similarities between time points (Fig. 2b). Let $s = 0$ index the subject for

the baseline network, and z_{00} denote its class labels at $t = 0$. The class labels of other subjects progress from the baseline network only at time $t = 0$ according to:

$$z_i^{(s,0)} = \begin{cases} z_i^{(0,0)} & \text{with probability } 1 - r_{\text{subject}}, \\ \text{Multinomial}(\frac{1}{K}, \ldots, \frac{1}{K}) & \text{otherwise,} \end{cases} \quad \text{for } i \in [G], \quad s = 1, \ldots, S-1. \tag{12}$$

Then, each subject progresses independently over time by:

$$z_i^{(s,t)} = \begin{cases} z_i^{(s,t-1)} & \text{with probability } 1 - r_{\text{time}}, \\ \text{Multinomial}(\frac{1}{K}, \ldots, \frac{1}{K}) & \text{otherwise,} \end{cases} \quad \text{for } i \in [G], \quad t = 1, \ldots, T-1, \quad s = 0, \ldots, S-1. \tag{13}$$

3.2 Experiments on Real Biological Dataset

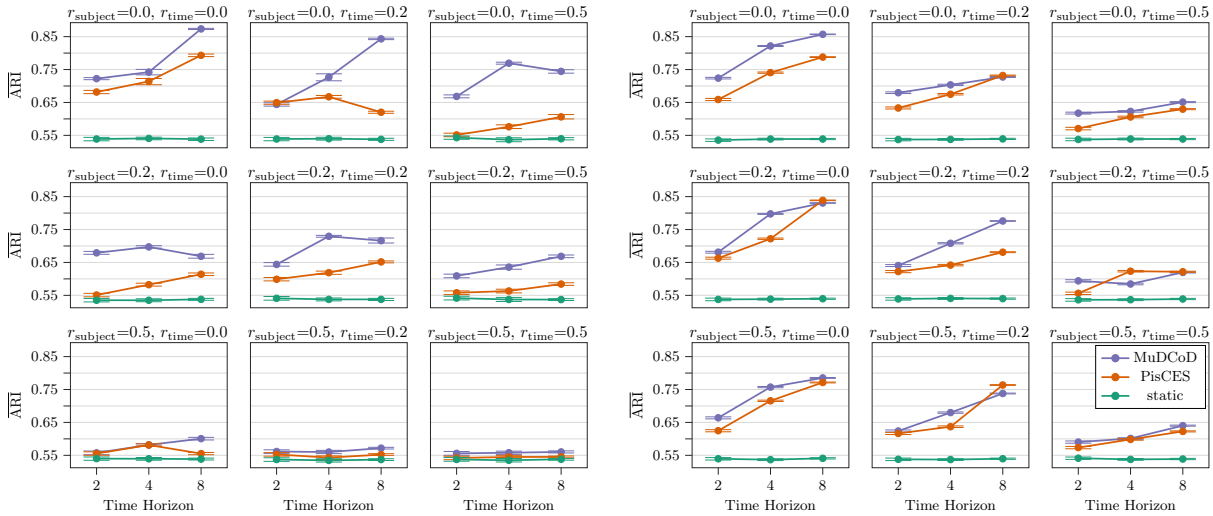
We applied MuDCoD to single cell experimental data. In this section, we explain the data-processing steps for this experiments. The Jerber-2021 dataset [8] contains single cell RNA-Seq data of cells from 215 iPSC lines derived from the Human Induced Pluripotent Stem Cell Initiative (HipSci) and differentiating toward a mid-brain neural fate. Cells were collected for scRNA-seq profiling on days 11, 30, and 52 (day-11, day-30, and day-52). We considered the scRNA-seq count matrices from the three time points together for pre-processing. We retained genes expressed in at least 0.5% of the cells, and selected top $3,000$ highly variable genes using the function `pp.filter_genes_dispersion` from Python package `Scanpy`. We normalized the count matrix for each donor using the R package `SCTransform` [5] and adjusted for "`pool_id`", "`time_point`", and "`treatment`" for batch correction. Leveraging the cell type annotations in [8], we excluded day-11 from the analysis as it predominantly harbored progenitor cells which matured into young and mature neurons in the following time points. We focused on day-30 and day-52 both of which had dopaminergic neurons (DA), serotonin transporter (Sert), and ependymal-like 1 (Epen1) cell-types. Furthermore, donors (i) that do not express, e.g., have zero expression counts across the donor cells, the $3,000$ highly variable genes for one or more "time point-cell type" combinations; (ii) have fewer than 500 cells for any specific "time point-cell type" combination were filtered out for a robust estimation of gene-gene correlations. This resulted in 16, 22, and 8 donors, respectively, for DA, Sert, and Epen1 cell-types. Finally, we constructed cell-type specific gene-gene Pearson correlation matrices for each donor at each time point. These matrices were then converted into unweighted adjacency matrices by keeping the top 5% of the edges (as commonly practiced in gene co-expression network construction from scRNA-seq [10]) based on absolute values of their correlations.

4 Results

4.1 Performance comparison with simulated experiments

We compared MuDCoD, PisCES, and baseline static spectral clustering ("static") in terms of their ability to identify the modules of the networks over subject and time dimensions. The number of communities, K, was selected using the same selection procedure (Section S1 of Appendix) for each method and we used the network cross-validation approach to determine the tuning parameters α and β for MuDCoD, and α for PisCES. Our choice of PisCES and "static" to compare MuDCoD with is motivated by PisCES's overall established advantages over the two additional methods that [13] had considered. Furthermore, to measure the similarity between the inferred and the ground truth gene modules of each method, we used the adjusted Rand index (ARI) [9]) as in [13].

We conducted simulations under the SoSS and SSoT settings of MuS-Dyn-DCBM with 16 subjects and for $T \in \{2, 4, 8\}$. The rest of the parameters were set as follows: network sizes $G = 100$, number of communities $K = 2$, in-cluster and out-cluster density parameters $p_{\text{in}} = (0.2, 0.25)$ and $p_{\text{out}} = 0.1$. We further considered varying r_{time} and r_{subject} levels. Here, r_{time} is the probability of a node changing its module label over two



(a) *The SSoS setting.* (b) *The SSoT setting.*

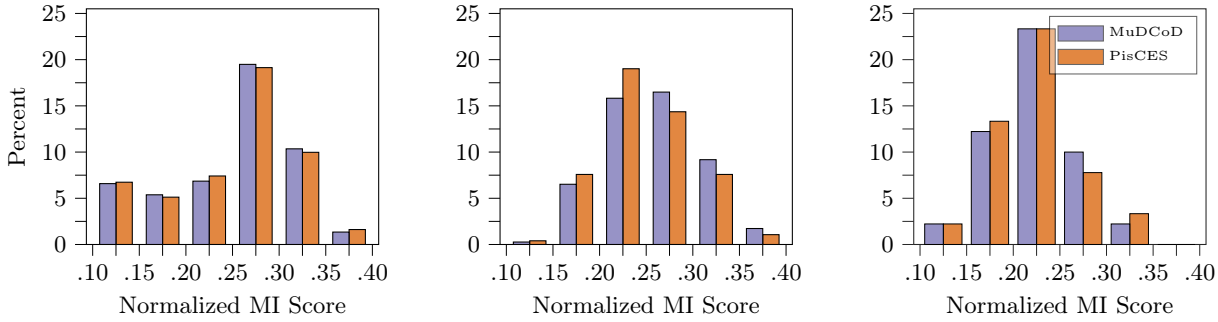
Fig. 3. Evaluation of the identified communities under two different MuS-Dyn-DCBM settings, (a) SSoS and (b) SSoT. Description of these settings are in Section 3.1. The simulation parameters were set as follows: network size $G = 100$, number of class labels $K = 2$, in-cluster and out-cluster density parameters $p_{\text{in}} = (0.2, 0.25)$ and $p_{\text{out}} = 0.1$, number of subjects $S = 16$, and number of time points $T \in \{2, 4, 8\}$. x-axis denotes T, and y axis is the mean ARI of the inferred modules for all subjects and time steps across 500 simulation replicates.

adjacent time points, and, similarly, r_{subject} is the probability of a node changing its module label across two subjects. For example, setting $r_{\text{subject}} = 0$ corresponds to the case where modules are similar across subjects, while $r_{\text{subject}} = 0.5$ corresponds to the setting where the modules of the subjects vary considerably and do not share any signal. Similarly, $r_{\text{time}} = 0$ and $r_{\text{time}} = 0.5$ yield constant and rapidly changing module labels across time, respectively. We conducted 500 simulation replicates and compared methods based on the average ARI values over these replicates.

Fig. 3 summarizes the results of these two general settings. Specifically, for the SSoS setting, Fig. 3a shows that as r_{subject} and/or r_{time} increases, ARIs of MuDCoD and PisCES tend to decrease and, eventually, they perform similar to the baseline "static". This is expected, because, as the heterogeneity in the communities increases along the subject and/or time dimensions, promoting information sharing across either dimension is no longer advantageous. MuDCoD significantly outperforms PisCES for relatively small r_{subject} values. The average increase in ARI is 10% and 7% for $r_{\text{subject}} = 0$ and $r_{\text{subject}} = 0.2$, respectively. The mean ARI increase is markedly high for some settings; for example, when $r_{\text{subject}} = 0$ and $r_{\text{time}} = 0.2$, increase in the mean ARI reaches to 22% with $T = 8$. Even with an extreme value of $r_{\text{subject}} = 0.5$, which corresponds to high levels of heterogeneity between the subject modules, MuDCoD outperforms the other methods, yielding up to 5% increase in ARI. Both PisCES and MuDCoD exhibit increasing performances as T increases, supporting the benefit of information sharing across the time dimension.

For the SSoT setting, Fig. 3b reveals that even for the case with the most module heterogeneity along the subject and time dimensions ($r_{\text{subject}} = 0.5$ and $r_{\text{time}} = 0.5$), both PisCES and MuDCoD outperform "static". Moreover, MuDCoD consistently performs better even when there are varying levels of information sharing among the subjects ($r_{\text{subject}} \in \{0, 0.2, 0.5\}$) and the modules are completely conserved across the time dimension $r_{\text{time}} = 0.0$), yielding up to 7% increase in the ARI with $T = 8$.

In addition to the experiments reported here, we conducted further simulations with a larger network size ($G = 500$) and a higher number of communities ($K = 10$) (Appendix Fig. S2). Overall, the results from this higher dimensional setting agreed with those presented here.



(a) NMI between subjects on day-30. (b) NMI between subjects on day-52. (c) NMI between day-30 and day-52 for each subject.

Fig. 4. Normalized MI scores between subject and time point pairs aggregated across all three cell types. (a) and (b) quantify normalized MI scores between every pair of subjects on day-30 and day-52, respectively, and the y-axis denotes the percentage of subject pairs. (c) Normalized MI scores between day-30 and day-52 across all the subjects, where the y-axis correspond to the percentage of subjects.

4.2 Application of MuDCoD to discover personalized gene communities

We next applied MuDCoD along with PisCES to discover personalized gene communities from the scRNA-seq data of human-induced pluripotent stem cells [8]. We specifically focused on cell types DA (16 donors, $1,955$ retained genes), Sert (22 donors and $2,234$ retained genes) and Epen1 (8 donors and $2,475$ retained genes) on days 30 and 52. Overall, the typical runs times of MuDCoD and PisCES were comparable, ranging between 30 minutes to 1.5 hours for about $2,000$ genes on a cluster with Intel Xeon Gold 6140 CPU (2.30GHz) and 64 GB RAM.

We first evaluated whether the smoothness promoted by MuDCoD over subjects resulted in more shared modules between the donors within each time point compared to PisCES. Fig. 4(a), (b) display the histograms of normalized mutual information (normalized MI) scores [18] between every pair of donors on day-30 and day-52, respectively. We observe that MuDCoD tends to yield higher, but not statistically significant, normalized MI scores between subjects for day-30 (Fig. 4a) and it leads to significantly higher (Wilcoxon rank-sum test p-value 0.0003) scores at day-52 (Fig. 4b), highlighting the impact of information sharing across the subjects. Furthermore, this does not incur at the cost of additional smoothing over the time domain as MuDCoD displays a comparable heterogeneity to PisCES across the time points (Fig. 4c).

Table 1 reports characteristics of MuDCoD inferred gene modules across each time point for each cell-type. While the numbers of donors available for each cell-type were different, the numbers of inferred communities, with the exception of DA at day-30, were similar in size. The mean number of communities across subjects ranges between 9.25 and 17.8. The low standard deviations in the number of communities indicate that the number of communities across subjects does not drastically change. We computed the density of the communities by dividing the number of edges by the total number of possible edges. The median community densities range between 0.30 and 0.39 for all cell-types. DA at day-30 leads to the least dense communities with larger community sizes. The median density of the communities at day-52 is either the same or larger than the communities discovered at day-30.

Following these general observations, we investigated the biological implications of MuDCoD results. To elucidate the biological processes that the modules are part of, we conducted a gene set enrichment analysis using Gene Ontology (GO) biological processes on each of the discovered modules at each time point and for each subject with `clusterProfiler` [20]. We only focused on communities with sizes larger than > 20. Our analysis limited the background gene sets in `clusterProfiler` to only those genes that were used in the construction of the adjacency matrices and used the Benjamini-Hochberg procedure [1] for multiplicity correction by controlling the false discovery rate. Appendix Fig. S3 highlights that some general biological processes, e.g., general cell-cycle and cell division related processes, are enriched across the communities of all

	DA		Sert		Epen1	
	Day-30	Day-52	Day-30	Day-52	Day-30	Day-52
Number of communities mean ± SD	9.25 ± 2.2	14.6 ± 3.7	15.7 ± 2.7	16.4 ± 3.8	16.5 ± 3.9	17.8 ± 4.2
Size of comm. $P_{10\%}$/median/$P_{90\%}$	51/143/458	25/95/293	41/98/280	3/96/291	3/125/259	3/91/306
Density of comm. $P_{10\%}$/median/$P_{90\%}$	.03/.30/.46	.04/.35/.51	.03/.38/.55	.02/.39/.55	.02/.39/.59	.01/.39/.64

Table 1. Summary of MuDCoD inferred communities for each cell-type and time point. The mean and standard deviation of the total number of communities across subjects are reported in the first row. In the second and third rows, median, 10-th, and and 90-th percentiles of the community sizes and densities are reported, respectively.

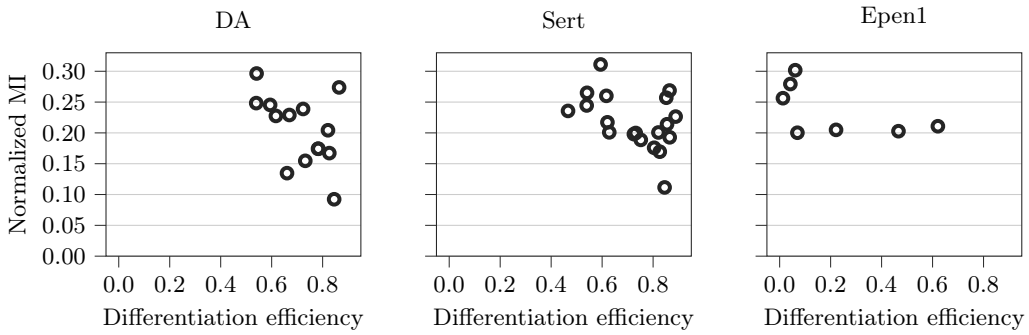


Fig. 5. Normalized MI scores between the inferred modules on day-30 and on day-52 for each subject with respect to the differentiation efficiency of the corresponding subject. The similarities between the modules inferred on day-30 and day-52, quantified with the normalized MI score, tends to decrease with increasing differentiation efficiency.

subjects. Additionally, we also observe processes specific to subsets of subjects, highlighting the personalized nature of the communities.

Next, we assessed whether donors with low and high differentiation efficiency exhibited varying levels of similarity among their modules inferred at different time points. For this, we leveraged the grouping of the donors based on the differentiation efficiency of their cells during dopaminergic neuron differentiation as defined in [8]. Fig. 5 displays the normalized MI scores between the inferred modules of day-30 and day-52 for each subject and suggests that donors with lower differentiation efficiency tend to have higher levels of similarity between their gene communities inferred at the two time points. We then focused on cell-type Epen1, which is not used in the definition of differentiation efficiency by [8], and asked whether differences in the donor-specific communities associated with the differentiation efficiency. This resulted in markedly higher module similarity within each group and lower across groups (Fig. 6a). Furthermore, consistent with the differentiation dynamics, we also observed relatively higher heterogeneity within the high group compared to the low group. Collectively, these support that personalized modules inferred by MuDCoD recapitulate the underlying differentiation efficiencies of donor iPS cells.

Focusing further on the cell type Epen1, we queried the groups of high and low differentiation efficiency donors for biological process enrichments of their communities. First, separately for each day, we identified *co-occurrent gene sets* (i) that are members of the same community in all of the low differentiation efficiency donors and (ii) that do not co-occur in any of the high differentiation efficiency donor communities. Specifically, we utilized the FP-Max [4] frequent item set generation algorithm, which is a variant of FP-Growth [6] (`mlxtend` Python implementation) to find such co-occurrent gene sets. Fig. 6b displays the GO enrichment analysis results for one of the largest co-occurrent gene sets (40 genes) (adjusted p-value ≤ 0.05, q-value ≤ 0.2). Interestingly, this analysis revealed many skin and neurological processes, including epidermal cell differentiation, skin development, keratinocyte differentiation and neutrophil degranatulation, all of which are consistent with the cell-type, as enriched for this gene set. The other co-occurent gene sets were also associated with neuronal processes (data not shown). While these results require further scrutiny to reveal specific biological insights, our framework highlights how personalized module detection promoting module

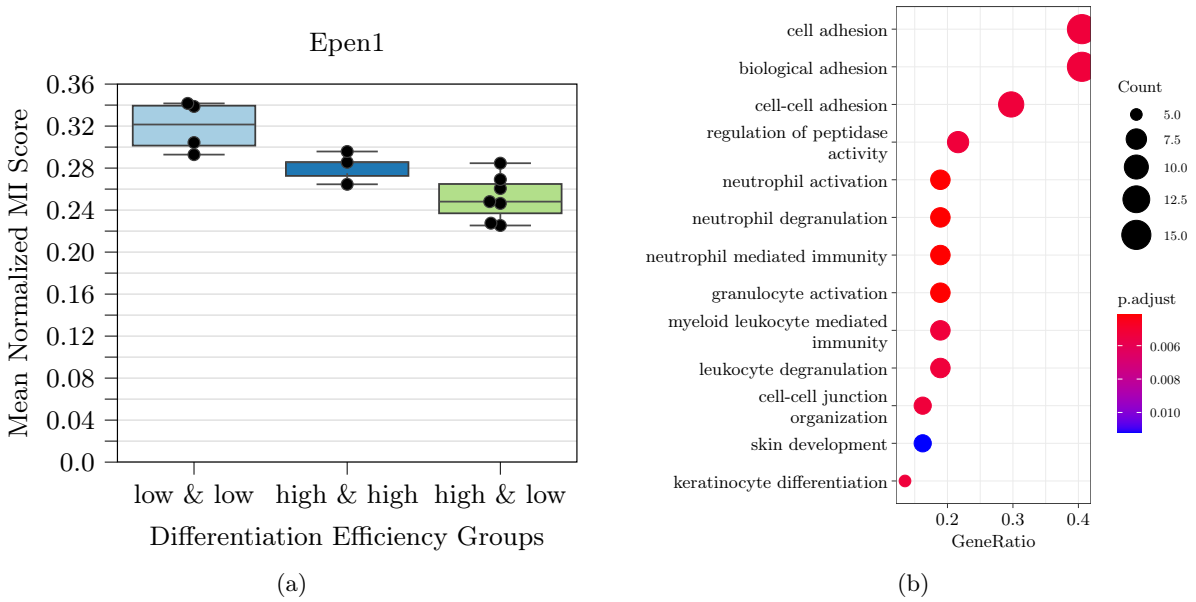


Fig. 6. (a) Mean normalized MI score comparison within and between the low and high differentiation efficiency groups of Epen1 cells. Each point stands for a donor and the y-axis denotes the mean of normalized MI scores between that donor and other donors in the corresponding group. (b) Set of enriched biological processes for a co-occurrent gene set that is specific to low differentiation efficiency donors in Epen1 cells at day-52. Displayed are significant biological processes (adjusted p-value ≤ 0.05) of one of the largest co-occurrent gene sets (40 genes).

smoothness over the subject and time dimensions with simple smoothness constraints can yield gene modules that associate with specific phenotypic characteristics of the subjects.

5 Discussion

As advances in single-cell profiling technologies allow charting transcriptomes of individuals at unprecedented resolutions across developmental and cellular stages, analysis of personalized dynamic gene networks is poised to emerge as a critical tool for elucidating network-level variations among subjects that might explain phenotypic variation. In this work, we developed a global spectral clustering framework that promotes information sharing among the subject and time dimensions. While our current framework encourages information sharing among all subjects, it can be extended to accommodate groupings among subjects where the communities are persistent only within the group members. Such an extension could be useful for diseases with sub-types. Another promising extension could incorporate prior information about the smoothness across the time dimension, for example, to allow for abrupt changes due to possible treatments or other experimental factors. Finally, we expect this multi-subject setting to be useful in other domains to infer network community structures with the availability of multiple samples.

Bibliography

[1] Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B (Methodological) **57**(1), 289–300 (1995), ISSN 00359246, URL http://www.jstor.org/stable/2346101

[2] Chen, Y.J.J., Friedman, B.A., Ha, C., Durinck, S., Liu, J., Rubenstein, J.L., Seshagiri, S., Modrusan, Z.: Single-cell rna sequencing identifies distinct mouse medial ganglionic eminence cell types. Scientific reports **7**(1), 1–11 (2017)

[3] Dai, H., Li, L., Zeng, T., Chen, L.: Cell-specific network constructed by single-cell RNA sequencing data. Nucleic Acids Research **47**(11), e62–e62 (Jun 2019), ISSN 0305-1048, 1362-4962, https://doi.org/10.1093/nar/gkz172, URL https://academic.oup.com/nar/article/47/11/e62/5377474

[4] Grahne, G., Zhu, J.: Efficiently using prefix-trees in mining frequent itemsets. In: Goethals, B., Zaki, M.J. (eds.) Proceedings of the IEEE ICDM Workshop on Frequent Itemset Mining Implementations (FIMI'03), CEUR Workshop Proceedings, vol. 90, Melbourne, Florida, USA (19 November 2003)

[5] Hafemeister, C., Satija, R.: Normalization and variance stabilization of single-cell rna-seq data using regularized negative binomial regression. Genome biology **20**(1), 1–15 (2019)

[6] Han, J., Pei, J., Yin, Y.: Mining frequent patterns without candidate generation. In: Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, p. 1–12, SIGMOD '00, Association for Computing Machinery, New York, NY, USA (2000), ISBN 1581132174, https://doi.org/10.1145/342009.335372, URL https://doi.org/10.1145/342009.335372

[7] HipSci Consortium, Cuomo, A.S.E., Seaton, D.D., McCarthy, D.J., Martinez, I., Bonder, M.J., Garcia-Bernardo, J., Amatya, S., Madrigal, P., Isaacson, A., Buettner, F., Knights, A., Natarajan, K.N., Vallier, L., Marioni, J.C., Chhatriwala, M., Stegle, O.: Single-cell RNA-sequencing of differentiating iPS cells reveals dynamic genetic effects on gene expression. Nature Communications **11**(1), 810 (Dec 2020), ISSN 2041-1723, https://doi.org/10.1038/s41467-020-14457-z, URL http://www.nature.com/articles/s41467-020-14457-z

[8] HipSci Consortium, Jerber, J., Seaton, D.D., Cuomo, A.S.E., Kumasaka, N., Haldane, J., Steer, J., Patel, M., Pearce, D., Andersson, M., Bonder, M.J., Mountjoy, E., Ghoussaini, M., Lancaster, M.A., Marioni, J.C., Merkle, F.T., Gaffney, D.J., Stegle, O.: Population-scale single-cell RNA-seq profiling across dopaminergic neuron differentiation. Nature Genetics **53**(3), 304–312 (Mar 2021), ISSN 1061-4036, 1546-1718, https://doi.org/10.1038/s41588-021-00801-6, URL http://www.nature.com/articles/s41588-021-00801-6

[9] Hubert, L., Arabie, P.: Comparing partitions. Journal of classification **2**(1), 193–218 (1985)

[10] Iacono, G., Massoni-Badosa, R., Heyn, H.: Single-cell transcriptomics unveils gene regulatory network plasticity. Genome biology **20**(1), 1–20 (2019)

[11] Jackson, C.A., Castro, D.M., Saldi, G.A., Bonneau, R., Gresham, D.: Gene regulatory network reconstruction using single-cell RNA sequencing of barcoded genotypes in diverse environments. eLife **9**, e51254 (Jan 2020), ISSN 2050-084X, https://doi.org/10.7554/eLife.51254, URL https://elifesciences.org/articles/51254

[12] Li, T., Levina, E., Zhu, J.: Network cross-validation by edge sampling. Biometrika **107**(2), 257–276 (04 2020), ISSN 0006-3444, https://doi.org/10.1093/biomet/asaa006, URL https://doi.org/10.1093/biomet/asaa006

[13] Liu, F., Choi, D., Xie, L., Roeder, K.: Global spectral clustering in dynamic networks. Proceedings of the National Academy of Sciences **115**(5), 927–932 (2018), ISSN 0027-8424, https://doi.org/10.1073/pnas.1718449115, URL https://www.pnas.org/content/115/5/927

[14] Mathys, H., Davila-Velderrain, J., Peng, Z., Gao, F., Mohammadi, S., Young, J.Z., Menon, M., He, L., Abdurrob, F., Jiang, X., Martorell, A.J., Ransohoff, R.M., Hafler, B.P., Bennett, D.A., Kellis, M., Tsai, L.H.: Single-cell transcriptomic analysis of Alzheimer's disease. Nature **570**(7761), 332–337 (Jun 2019), ISSN 0028-0836, 1476-4687, https://doi.org/10.1038/s41586-019-1195-2, URL http://www.nature.com/articles/s41586-019-1195-2

[15] Matias, C., Miele, V.: Statistical clustering of temporal networks through a dynamic stochastic block model. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **79**(4), 1119–1141 (2017), https://doi.org/https://doi.org/10.1111/rssb.12200, URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssb.12200

[16] Matias, C., Miele, V.: Statistical clustering of temporal networks through a dynamic stochastic block model. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **79**(4), 1119–1141 (2017), https://doi.org/https://doi.org/10.1111/rssb.12200, URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssb.12200

[17] Shen, H.W., Cheng, X.Q.: Spectral methods for the detection of network community structure: a comparative analysis. Journal of Statistical Mechanics: Theory and Experiment **2010**(10), P10020 (Oct 2010), ISSN 1742-5468, https://doi.org/10.1088/1742-5468/2010/10/p10020, URL http://dx.doi.org/10.1088/1742-5468/2010/10/P10020

[18] Strehl, A., Ghosh, J.: Cluster ensembles — a knowledge reuse framework for combining multiple partitions. J. Mach. Learn. Res. **3**(null), 583–617 (Mar 2003), ISSN 1532-4435, https://doi.org/10.1162/153244303321897735, URL https://doi.org/10.1162/153244303321897735

[19] Travaglini, K.J., Nabhan, A.N., Penland, L., Sinha, R., Gillich, A., Sit, R.V., Chang, S., Conley, S.D., Mori, Y., Seita, J., Berry, G.J., Shrager, J.B., Metzger, R.J., Kuo, C.S., Neff, N., Weissman, I.L., Quake, S.R., Krasnow, M.A.: A molecular cell atlas of the human lung from single-cell RNA sequencing. Nature **587**(7835), 619–625 (Nov 2020), ISSN 0028-0836, 1476-4687, https://doi.org/10.1038/s41586-020-2922-4, URL http://www.nature.com/articles/s41586-020-2922-4

[20] Yu, G., Wang, L.G., Han, Y., He, Q.Y.: clusterprofiler: an r package for comparing biological themes among gene clusters. OMICS: A Journal of Integrative Biology **16**(5), 284–287 (2012), https://doi.org/10.1089/omi.2011.0118, URL https://doi.org/10.1089/omi.2011.0118, pMID: 22455463

[21] Zeisel, A., Muñoz-Manchado, A.B., Codeluppi, S., Lönnerberg, P., La Manno, G., Juréus, A., Marques, S., Munguba, H., He, L., Betsholtz, C., et al.: Cell types in the mouse cortex and hippocampus revealed by single-cell rna-seq. Science **347**(6226), 1138–1142 (2015)

S1 Hyperparameter Selection

S1.1 Selecting the Number of Communities $\mathcal{K}$:

Given $L \in \mathbb{R}^{G\times G}$ with eigenvalues $|\lambda_1|, \geq \cdots \geq |\lambda_G|$, define $\mathcal{K} : \mathbb{R}^{G\times G} \mapsto \mathbb{N}$ as a function that computes the number of modules, i.e., the number of eigenvectors, as:

$$\mathcal{K}(M) = \min\left\{K \ : \ |\lambda_i| - |\lambda_{(i+1)}| < \delta \text{ for all } i > K\right\}. \tag{14}$$

Here, δ denotes the threshold for the "noise" eigenvalues. Let $v_1, \ldots, v_{\mathcal{K}(M)}$ denote the eigenvectors of M corresponding to the $\mathcal{K}(M)$ largest eigenvalues in absolute value. Then, $\mathit{\Pi}_K(M)$ can be replaced with $\mathit{\Pi}(M)$, which is defined as:

$$\mathit{\Pi}(M) = \sum_{k=1}^{\mathcal{K}(M)} v_k v_k^T. \tag{15}$$

Based on the assumption of independent observation noise across dyads (i.e., node/gene pairs), [13] simulates Erdős-Renyi normalized Laplacian matrices $L^{(ER)}$ with sizes and densities matching to observed L to determine the threshold δ for the corresponding networks generated with DCBM. We adapt the same approach and compute δ by

$$\delta = \text{quantile}_{0.95}\left[max\left\{|\lambda_i^{(ER)}| - |\lambda_{i+1}^{(ER)}|, \ i \geq 2\right\}\right]. \tag{16}$$

We note that the assumption of dyadically independent observation noise may not be appropriate for the real scRNA-seq co-expression networks. To accommodate this possibility, we compute the threshold δ as the largest eigengap excluding λ_1:

$$\delta = \max\left\{|\lambda_i| - |\lambda_{(i+1)}|, \ i \geq 2\right\}. \tag{17}$$

S1.2 Cross-validation to Tune Hyperparameters α and β:

We implemented a cross-validation scheme based on Li et al. [12] to select the tuning parameters α and β as follows.

1. Randomly split the $\binom{G}{2} \cdot T \cdot S$ dyads of $A_{11}, \ldots, A_{ST}$ into F folds. Let Q denote the set of candidate pairs for (α, β).
2. For each fold $f = 1, \ldots, F$ and time-subject pair (s, t), define $A_{st}^{(f)} \in \mathbb{R}^{G\times G}$ as:

$$\left[A_{st}^{(f)}\right]_{ij} = [A_{st}]_{ij}\, \mathbf{I}\left\{(i, j, s, t) \in f\right\}, \tag{18}$$

   where $\mathbf{I}$ denotes the indicator function.
3. Apply a low-rank matrix completion algorithm [12] to $A_{st}^{(f)}$ to impute the zeroed entries, and let $\hat{A}_{st}^{(f)}$ denote the resulting matrix. Let $\overline{U}_{11}^{(\alpha,\beta,f)}, \ldots, \overline{U}_{ST}^{(\alpha,\beta,f)}$ denote MuDCoD's output with smoothing parameters α, β and imputed $\hat{A}_{11}^{(f)}, \ldots, \hat{A}_{ST}^{(f)}$.
4. Evaluate $\mathcal{L}^{(\alpha,\beta,f)}$, the log likelihood of the fitted DCBM on the fth fold:

$$\mathcal{L}^{(\alpha,\beta,f)} = \sum_{(i,j,s,t)\in f} [A_{st}]_{ij} \log(\left[P_{(s,t)}\right]_{ij}) + (1 - [A_{st}]_{ij}) \log(1 - \left[P_{(s,t)}\right]_{ij}), \tag{19}$$

   where $\left[P_{(s,t)}\right]_{ij} \equiv \left[P_{(s,t)}^{(f)}\right]_{ij}$ is the estimated probability of connection between node i and node j at time t for subject s under the fitted DCBM, and is given by:

$$\left[P_{(s,t)}\right]_{ij} = d_i^{(s,t)} d_j^{(s,t)} B^{(s,t)}(\hat{z}_i^{(s,t)}, \hat{z}_i^{(s,t)}). \tag{20}$$

Here, $\hat{z}_i^{(s,t)} \equiv \hat{z}_i^{(s,t,f)}$ is the estimated community node i at time t as given by clustering on the eigenvectors of $\overline{U}_{st}^{(\alpha,\beta,f)}$ with $\mathcal{K}(\overline{U}_{st}^{(\alpha,\beta,f)})$ classes where $d_i^{(s,t)} \equiv d_i^{(s,t,f)}$ is the estimated DCBM degree parameter for node i, and is given by:

$$d_i^{(s,t)} = \sum_j \left[\hat{A}_{st}^{(f)}\right]_{ij}, \tag{21}$$

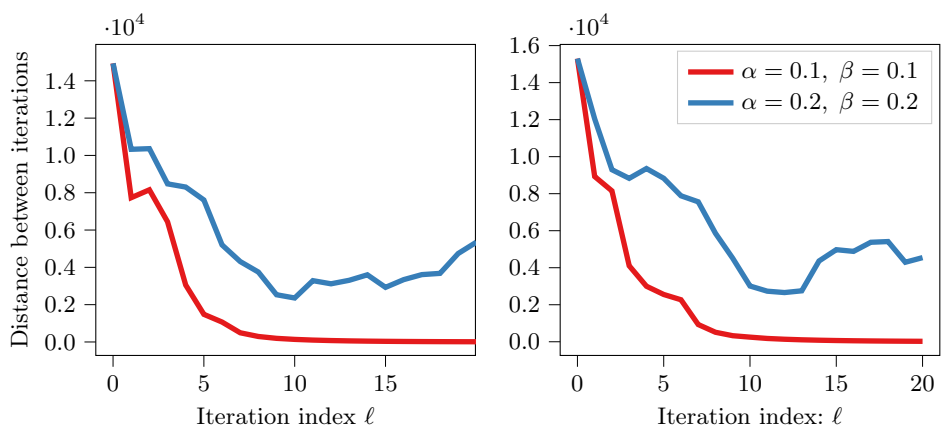
and $B^{(s,t)} \equiv B^{(s,t,f)} \in \mathbb{R}^{K\times K}$ is the estimated DCBM density parameter matrix, and is given by:

$$\left[B^{(s,t)}\right]_{kl} = \left(\sum_{(i,j)} \left[\hat{A}_{st}^{(f)}\right]_{ij} \mathbf{I}\left\{\hat{z}_i^{(s,t)} = k,\ \hat{z}_j^{(s,t)} = l\right\}\right) / \left(\sum_{(i,j)} d_i^{(s,t)} d_j^{(s,t)} \mathbf{I}\left\{\hat{z}_i^{(s,t)} = k,\ \hat{z}_j^{(s,t)} = l\right\}\right). \tag{22}$$

5. Let $\mathcal{L}^{(\alpha,\beta)} = \frac{1}{F}\sum_f \mathcal{L}^{(\alpha,\beta,f)}$ and return the tuning parameter pair $(\alpha,\beta) \in Q$ giving the best value for $\mathcal{L}^{(\alpha,\beta)}$.

S2 Convergence in Simulation Experiments

Fig. S1 illustrates the convergence of MuDCoD with different α and β values by quantifying the ℓ_1 norm of the difference of the estimated eigenvectors between consecutive iterations.

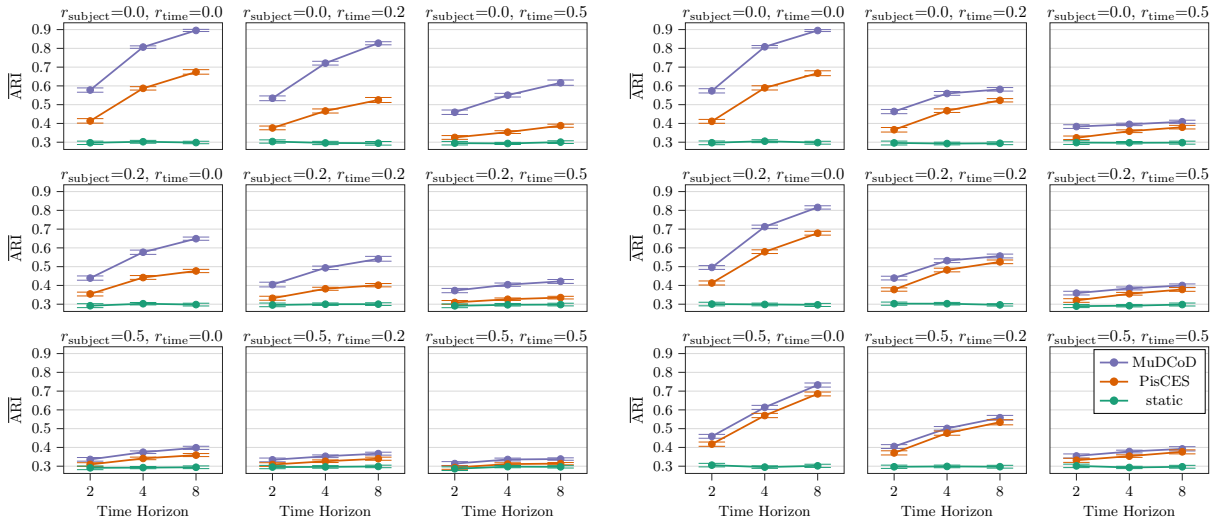


(a) SSoS setting. (b) SSoT setting.

Fig. S1. *Convergence status of MuDCoD with different α and β values.* y-axis quantifies the ℓ_1 norm of the difference of estimated eigenvectors between consecutive iterations: $\sum_{s=0}^{S-1}\sum_{t=0}^{T-1} \|\overline{U}_{st}^{\ell} - \overline{U}_{st}^{\ell-1}\|_1$. Other parameters of the MuSDyn-DCBM were set as follows: $G = 500$, $K = 10$, $p_{\text{in}} = (0.3, 0.3)$, $p_{\text{out}} = 0.1$, $r_{\text{time}} = 0.2$, $r_{\text{subject}} = 0.2$, $S = 4$, $T = 4$.

S3 Performance Comparison on MuS-Dyn-DCBM with $K = 10$ and $G = 500$

Simulations with a larger network size G and a higher number of communities K reveal that MuDCoD performs comparably or better than PisCES and outperforms static spectral clustering.



(a) *The SSoS setting.* (b) *The SSoT setting.*

Fig. S2. Evaluation of the identified communities under two different MuS-Dyn-DCBM settings, (a) SSoS and (b) SSoT. The simulation parameters were set as follows: the network size is $G = 500$, number of class labels $K = 10$, the in-cluster and out-cluster density parameters $p_{\text{in}} = (0.2, 0.4)$ and $p_{\text{out}} = 0.1$, number of subjects $S = 8$, and number of time points $T \in \{2, 4, 8\}$. x-axis is the number of time points T, and y axis is the mean ARI of the inferred modules for all subjects and time steps across 100 simulation replicates.

S4 Gene Set Enrichment Analysis of Communities Detected for Individual Subjects Across the Time Points

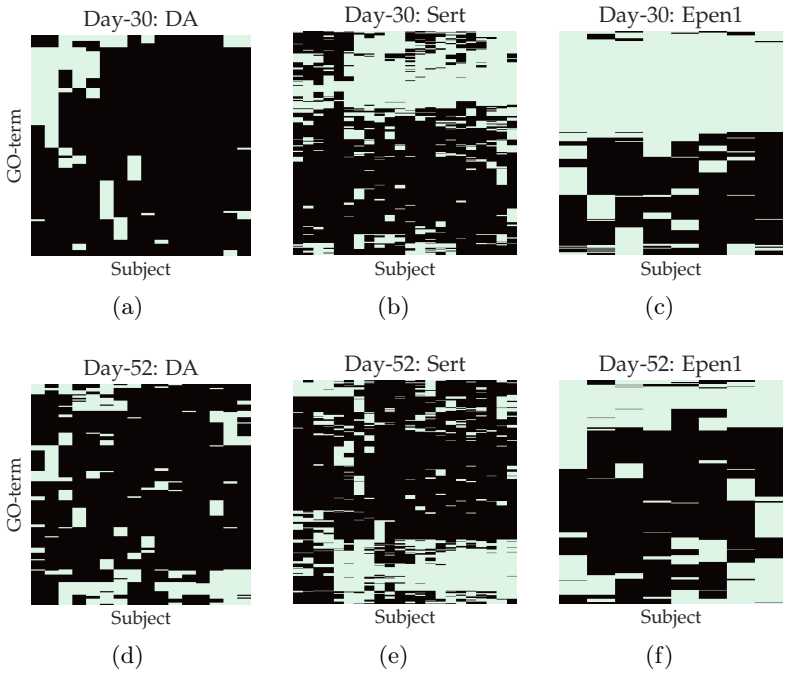


Fig. S3. Gene set enrichment analysis using GO biological processes was carried out for each of the discovered modules (with a size of at least 20 genes) at each time point and for each subject with `clusterProfiler`. Background gene sets were limited to those used in the construction of the adjacency matrices and FDR was controlled at level 0.05 with the Benjamini-Hochberg procedure. Enriched GO terms (rows) are depicted with the green color and the columns denote the subjects. Heatmaps highlight the over patterns of enriched GO terms across the subjects for each "time point-cell type" combination and the actual names of the GO terms on the rows are suppressed to improve visualization.