Published online 25 May 2022 *Nucleic Acids Research, 2022, Vol. 50, No. 15* ***e86***
https://doi.org/10.1093/nar/gkac412

Inference of cell state transitions and cell fate plasticity from single-cell with MARGARET

Kushagra Pandey[1] **and Hamim Zafar** [1,2,3,*]

[1]Department of Computer Science and Engineering, Indian Institute of Technology Kanpur, Kanpur 208016, India, [2]Department of Biological Sciences and Bioengineering, Indian Institute of Technology Kanpur, Kanpur 208016, India and [3]Mehta Family Centre for Engineering in Medicine, Indian Institute of Technology Kanpur, Kanpur 208016, India

Received January 21, 2022; Revised April 27, 2022; Editorial Decision May 06, 2022; Accepted May 17, 2022

ABSTRACT

Despite recent advances in inferring cellular dynamics using single-cell RNA-seq data, existing trajectory inference (TI) methods face difficulty in accurately reconstructing the cell-state manifold and cell-fate plasticity for complex topologies. Here, we present MARGARET (https://github.com/Zafar-Lab/Margaret) for inferring single-cell trajectory and fate mapping for diverse dynamic cellular processes. MARGARET reconstructs complex trajectory topologies using a deep unsupervised metric learning and a graph-partitioning approach based on a novel connectivity measure, automatically detects terminal cell states, and generalizes the quantification of fate plasticity for complex topologies. On a diverse benchmark consisting of synthetic and real datasets, MARGARET outperformed state-of-the-art methods in recovering global topology and cell pseudotime ordering. For human hematopoiesis, MARGARET accurately identified all major lineages and associated gene expression trends and helped identify transitional progenitors associated with key branching events. For embryoid body differentiation, MARGARET identified novel transitional populations that were validated by bulk sequencing and functionally characterized different precursor populations in the mesoderm lineage. For colon differentiation, MARGARET characterized the lineage for BEST4/OTOP2 cells and the heterogeneity in goblet cell lineage in the colon under normal and inflamed ulcerative colitis conditions. Finally, we demonstrated that MARGARET can scale to large scRNA-seq datasets consisting of ∼ millions of cells.

INTRODUCTION

Dynamic cellular processes such as differentiation involve cell-state transitions that are characterized by cascades of epigenetic and transcriptional changes (1,2). High-throughput single-cell RNA sequencing (scRNA-seq) datasets allow us to identify cellular identities at a single-cell resolution (3,4) and thus can be utilized for elucidating the cellular heterogeneity of a dynamic cellular process and tracking cell fate decisions in normal as well as pathological development (5). Despite recent advances (6,7) in inferring cellular dynamics from the underlying developmental process, existing computational trajectory inference (TI) methods (7–11) face several critical challenges. Most TI methods have largely overlooked the importance of dimensionality reduction by focusing more on trajectory modelling and relying on generalized dimension reduction techniques such as UMAP (12), local linear embedding (11), or diffusion maps (9) which may obscure the identification of some intermediate cell states. Moreover, most TI methods impose strong assumptions on the topology of the trajectory and cannot generalize to disconnected or hybrid topologies without imposing further restrictions (13). Lastly, accurate detection of terminal cell states remain difficult as only a few methods (e.g. Slingshot (10), Palantir (9), Monocle3 (7), VIA (14)) can automatically identify cell fates (Supplementary Table S1). Existing gene-expression similarity-based TI methods also focus mostly on reconstructing the order of cell states, how cell fate choices evolve along a dynamic process remain less explored. Recently, a Markov chain model has been introduced by Palantir (9) to quantify the plasticity of cell fates along a trajectory. However this approach predominantly assumes a connected trajectory and cannot generalize to disconnected trajectories.

To overcome these challenges, we developed MARGARET (Metric leARned Graph pARtitionEd Trajectory) which provides an end-to-end framework that utilizes scRNA-seq data for inferring the cell state trajectory and dynamics of cell fate plasticity and thereby characterizes the differentiation landscape. MARGARET employs an unsupervised metric learning-based approach for inferring the cell-state manifold where the distinct cell states are represented by compact cell clusters. To capture complex trajectory topologies, MARGARET employs the inferred cellular embeddings and the cell clusters to construct a cluster

[*]To whom correspondence should be addressed. Tel: +91 8737992101; Email: hamim@iitk.ac.in

connectivity graph by using a novel measure of connectivity between cell clusters. The cluster connectivity graph is used in conjunction with the cell-nearest-neighbor graph to compute a pseudotime ordering of cells. To identify terminal states in the trajectory, MARGARET introduces a shortest-path betweenness-based measure. Finally, MARGARET refines the absorbing Markov chain model of Palantir and introduces a local random walk-based novel algorithm for computing cell fate probabilities which in turn generalizes the quantification of the cell fate plasticity for complex trajectory topologies.

We demonstrate the performance of MARGARET in trajectory inference and cell-fate prediction across a variety of synthetic and experimental scRNA-seq datasets. For simulated datasets with known ground-truth (13) consisting of diverse topologies and a real benchmark consisting of datasets from placenta trophoblast differentiation (15), mouse cell atlas (15), oligodendrocyte differentiation (16) and planaria parenchyme differentiation (17), MARGARET outperformed state-of-the-art TI methods both in terms of capturing the global topology and recovering the underlying pseudotime ordering. Using simulated datasets with ground-truth pseudotime ordering, we further showed that MARGARET's quantification of cell fate plasticity is superior to that of Palantir and generalizes for complex disconnected trajectories. When applied to real biological datasets (9,18,19) representing human hematopoiesis, embryogenesis and colon differentiation, MARGARET accurately identified all major lineages along a pseudotemporal order that epitomized the expression trends of canonical cell-type markers in these processes. For hematopoiesis, in the myeloid and erythroid-megakaryocytic lineages, MARGARET helped identify transitional progenitors associated with key branching events, which were also characterized by a drastic shift in MARGARET inferred cell-fate plasticity. For embryoid body differentiation, MARGARET accurately characterized all ectodermal, endodermal and mesodermal lineages; identified novel transitional populations that were validated by bulk sequencing; and functionally characterized different precursor populations in the mesoderm lineage. For colon differentiation, MARGARET delineated the secretive and absorptive cell lineages for human colon differentiation; characterized the lineage for BEST4/OTOP2 cells and the heterogeneity in goblet cell lineage in the colon under normal and inflamed ulcerative colitis conditions. Finally, using a 1.3 million neuronal cells dataset (20), we demonstrate that MARGARET can scale to large scRNA-seq datasets making it suitable for analyzing atlas-level scRNA-seq datasets.

MATERIALS AND METHODS

Preprocessing scRNA-seq data

We downloaded the filtered, normalized and log-transformed count matrices for early Human Hematopoiesis datasets (replicates 1 and 2) provided by (9) (see Data Availability). Pre-processed replicate 1 consisted of 5780 cells and 14 651 genes while replicate 2 consisted of 6501 cells and 14 913 genes. For each replicate, we performed imputation using MAGIC (21), and then computed 300-dimensional PCA embeddings on the imputed data as suggested in (9). For the embryoid body (EB) dataset, we downloaded the filtered, normalized and square-root transformed count matrix as provided in (18) (see Data Availability). The EB dataset consisted of 16 821 cells and 17 845 genes. We then performed initial denoising on the dataset to extract 50-dimensional PCA embeddings as performed in (18). For studying colon differentiation under normal and UC conditions, we downloaded the preprocessed count matrix provided by (19). Top 2000 highly variable genes were identified in this dataset and then PCA was performed to compute 300-dimensional embeddings, which were used for further analysis. For real benchmarking datasets (Placenta Trophoblast differentiation, Mouse cell atlas, Oligodendrocyte differentiation and Planaria parenchyme differentiation), we downloaded filtered, normalized and log-transformed count matrices and then computed 10-dimensional PCA embeddings for subsequent analysis using other methods. We downloaded the real datasets utilized in embedding quality evaluations (PBMC-8k, PBMC-4k, Heart Cell Atlas and CORTEX) and runtime benchmarking (1.3M neuronal cells) using scvi-tools (22) and performed filtering, normalization and log-transformation followed by z-score normalization. We then performed PCA with 100 components on the 1.3M dataset and 10 components on the other real datasets. All preprocessing was performed using Scanpy (8).

Overview of MARGARET

The overall framework of MARGARET consists of two main steps. We first infer the lower-dimensional cell-state manifold using preprocessed scRNA-seq data. This is followed by trajectory modelling which encompasses constructing an undirected graph on the cell clusters inferred from the cellular embeddings followed by the inference of pseudotime ordering of the cells (Figure 1). Each of these computational stages are described in more detail in the following sections.

Inference of lower-dimensional cell-state manifold

For inferring the cell-state manifold, given preprocessed scRNA-seq data, we propose an unsupervised metric-learning-based approach to learn a meaningful lower-dimensional representation of each cell in the scRNA-seq dataset (see Figure 1A). The central idea is to learn a nonlinear manifold over cell representations such that the cells which belong to the same cell type (cluster) are packed compactly while the cells belonging to different clusters are far apart in the manifold. The key steps of our proposed approach are as follows:

(1) **Computing initial clusters:** Given the preprocessed scRNA-seq data with N cells and an initial embedding of dimension D, we perform initial clustering on the data. The cluster label assigned to each cell is then treated as a pseudo-label, which is used in subsequent steps. The pseudo-labelled dataset is denoted as $\mathcal{D}_{PL} = \{(x_i, y_i)\}_{i=1}^{N}$, where the pair (x_i, y_i) represents the initial embedding and the pseudo-label of the ith cell in the dataset respectively. It is worth noting that the

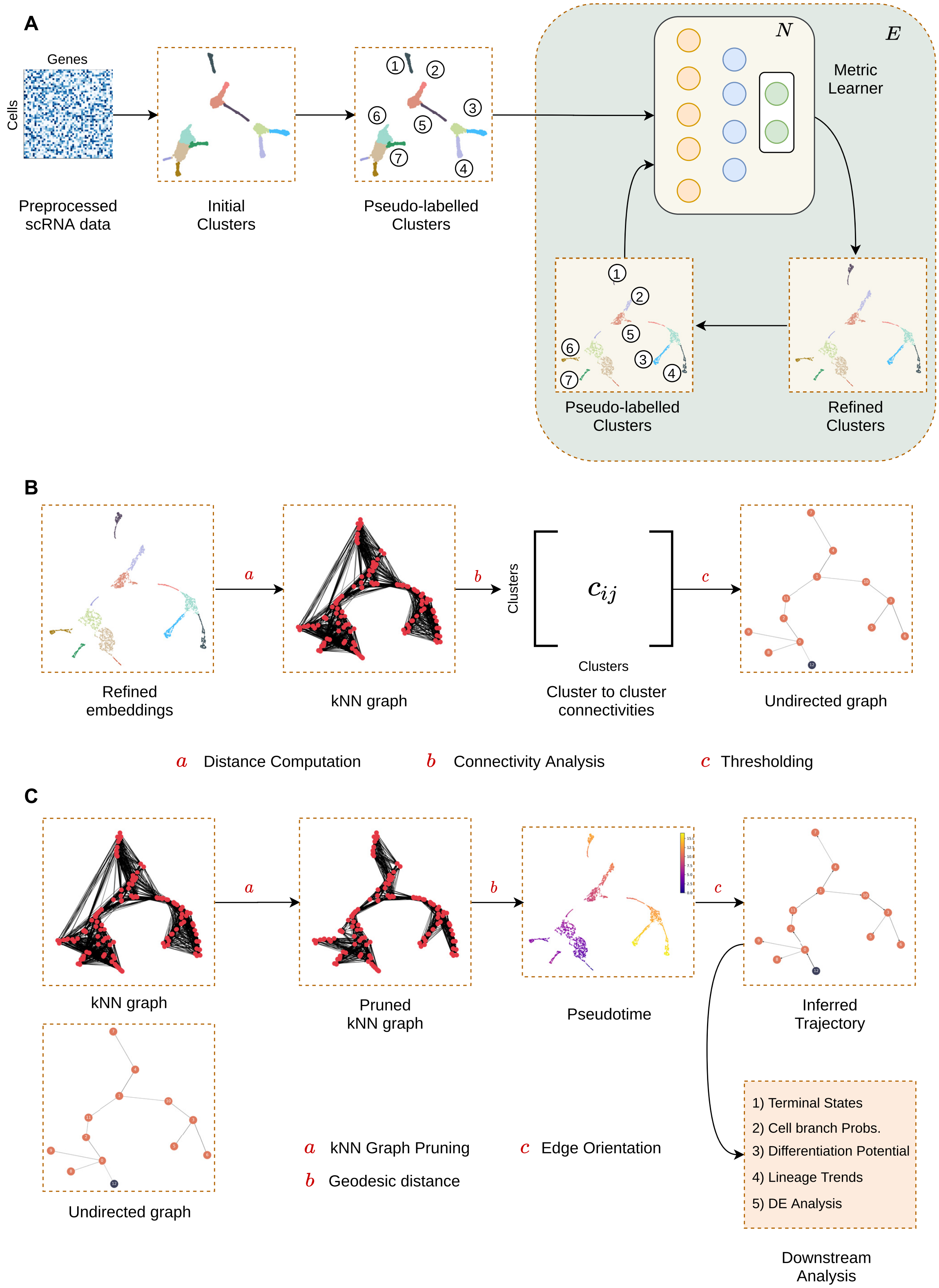

Figure 1. Overview of MARGARET. (**A**) Given a preprocessed scRNA-seq dataset, MARGARET uses an unsupervised metric-learning-based approach to learn compact cell-state representations from pseudolabels generated from an initial cell embedding. The cell cluster assignments and the MARGARET embedding are refined through an episodic training which results in the final refined embeddings and refined cell type clusters. (**B**) MARGARET infers the connectivity graph between the refined cell partitions by computing connectivities between clusters using a cell-level k-nearest-neighbor (kNN) graph. (**C**) MARGARET prunes the kNN graph by removing *short-circuit* edges and infers cell pseudotime from the pruned kNN graph by computing geodesic distances from the start cell(s). The pseudotime values are then utilized to infer a directed trajectory. MARGARET utilizes the inferred cell-state embeddings and the trajectory for several downstream tasks.

choice of the clustering method in this step can be arbitrary and the users can choose any single-cell clustering method of their choice. However, in this work, we primarily use community detection-based clustering methods such as Louvain and Leiden clustering (23) due to their widespread use in the single-cell literature (24) and availability of efficient implementations in Scanpy (8).

(2) **Metric learning:** We then learn a low-dimensional representation of each pseudo-labelled cell using a non-linear mapping $f_\theta : \mathbb{R}^D \rightarrow \mathbb{R}^d$ parameterized by θ where $d \leq D$. In this work, we represent f_θ using a feed-forward deep neural network with parameters θ. Given this parameterization and a pseudo-labelled dataset $\mathcal{D}_{PL}$, we learn a low-dimensional representation of each cell as follows:
Given an initial representation (x_i, y_i) for the ith cell, two additional cells (x_j, y_j) and (x_k, y_k) are sampled from $\mathcal{D}_{PL} - \{(x_i, y_i)\}$ such that $y_i = y_j$ and $y_i \neq y_k$. Let us assume that f_a, f_p and f_n denote the low-dimensional embeddings of x_i, x_j and x_k such that:

$$f_a = f_\theta(x_i)$$
$$f_p = f_\theta(x_j)$$
$$f_n = f_\theta(x_k)$$

θ is updated such that the Triplet-Margin loss function (25) $\mathcal{L}(f_a, f_p, f_n)$ is minimized, where the loss $\mathcal{L}(f_a, f_p, f_n)$ is given by:

$$\mathcal{L}(f_a, f_p, f_n) = \max\{d(f_a, f_p) - d(f_a, f_n) + margin, 0\} \quad (1)$$

$$d(f_i, f_j) = \| f_i - f_j \|_p \quad (2)$$

The value of *margin* and *p* are chosen as 1 and 2, respectively. Intuitively, for each cell x_i (denoting the anchor), we sample positive (x_j) and negative samples (x_k) such that the distance between the anchor and the positive sample is minimized while the distance between the anchor and the negative sample is maximized. This training step is repeated for *e* epochs, where *e* is a hyperparameter.

(3) **Cluster refinement:** We generate the low-dimensional representations of all cells in the dataset using pretrained f_θ from step 2. Using the updated low-dimensional embedding $f_\theta(x_i)$ for all N cells, cells are clustered again to generate refined cluster assignments, which act as new pseudo-labels for the cells $\{x_i\}_{i=1}^N$.

(4) **Episodic training:** The sequence of steps 2 and 3 form a single *episode*. We repeat steps 2 and 3 alternatively for a total of *E* episodes, where *E* is a hyperparameter. We can also alternate between the two steps until convergence which can be assessed by monitoring the quality of the clusters generated from step 3. For example, when using Leiden or Louvain clustering, convergence can be assessed by monitoring the *modularity* score of the refined clusters.

(5) **Inferring final low-dimensional manifold:** After training, the final low-dimensional representation, f_i^* for cell i can be obtained by $f_{\theta^*}(x_i)$ where θ^* denotes the trained parameters of the deep neural network f. As a by-product of training, we also obtain the refined cluster assignments, y_i^* for each cell in the scRNA-seq dataset. The refined embeddings $\mathcal{M}_f = \{f_i^*\}_{i=1}^N$ and cluster assignments $\{y_i^*\}_{i=1}^N$ are used in subsequent steps of MARGARET.

Network architecture and training hyperparameters

We use a simple feed-forward deep neural network architecture in MARGARET consisting of 2-fully connected layers of sizes 128 and 64, respectively. The size of the input layer depends on the size of the embedding of the preprocessed scRNA-seq dataset. The number of neurons in the final output layer is the same as the size of the desired low-dimensional embedding which is a hyperparameter. In addition, each fully connected layer in MARGARET is followed by a Batch Normalization (BN) (26) layer followed by the ReLU activation. To regularize and enrich the intermediate representations, we use Dropout (27) after each Linear-BN-ReLU module. The dropout rate is set to 0.3 for all the experiments. Moreover, the layer sizes were also kept fixed for all the experiments.

During training, we used Stochastic Gradient Descent (SGD) with an initial learning rate of 0.01 to update the metric learner parameters. To adjust the learning rate during training, we used a poly-learning rate scheduler with the following update schedule:

$$lr(t) = lr(0) * (1 - \frac{t}{num\_epochs})^\alpha \quad (3)$$

$\alpha = 0.9$ was used during training.

Trajectory modelling

Given a low dimensional representation of cells, $\mathcal{M}_f$ and refined cluster assignments $\{y_i^*\}_{i=1}^N$, MARGARET learns a connectivity graph over the set of refined clusters similar to PAGA (28) to model the trajectory over the underlying dynamic process. The key steps involved in learning the trajectory are described below.

(1) **Learning an undirected graph:** MARGARET first learns an undirected graph $\mathcal{U}$ over the refined partition of cells y_i^* (Figure 1B). This learned graph $\mathcal{U}$ models the connectivity of the cell clusters and identifies the connected and disconnected neighborhoods in the cell-state manifold. To learn $\mathcal{U}$, we first compute the k-nearest-neighbor (kNN) graph, $\mathcal{G}$ at the single-cell level using the learned low-dimensional manifold $(\mathcal{M}_f)$. The resulting graph, $\mathcal{G}$ is represented as a $N \times N$ sparse adjacency matrix. We then assess the connectivity between two clusters c_i and c_j by introducing a novel measure of connectedness. Formally, we define the connectivity between two clusters c_i and c_j as:

$$\psi_{ij} = \frac{e_{ij} + e_{ji} - e_{rand}}{e_i + e_j - e_{rand}} \quad (4)$$

where e_{ij} denotes the number of kNN graph edges from cluster c_i to c_j, e_{ji} denotes the number of kNN graph edges from cluster c_j to c_i, e_i denotes the number of outgoing edges from cluster c_i, e_j denotes the number of outgoing edges from cluster c_j, and e_{rand} denotes the number of edges from cluster c_i to c_j and vice-versa under the random assignment of edges. Intuitively, when computing the connectivity between two clusters, we adjust the connectivity score to account for the random assignment of edges from cluster c_i to c_j to prevent spurious connections in $\mathcal{U}$.

Following PAGA (28), we model the random assignment of edges between two clusters using a binomial distribution. In this scenario, e_{rand} is given by:

$$e_{rand} = \frac{e_i n_j + e_j n_i}{N - 1}, \tag{5}$$

where n_i and n_j represent the number of cells (size) in the clusters c_i and c_j respectively and N represents the total number of nodes in the kNN graph (i.e. the number of cells). Given a threshold (lower-bound for ψ_{ij}) t_c, two clusters c_i and c_j are said to be connected when $\psi_{ij} > t_c$. In this work, we define a simple statistical test and compute the z-score between two clusters c_i and c_j given by:

$$z_{ij} = \frac{e_{ij} + e_{ji} - e_{rand}}{\sigma_{rand}}, \tag{6}$$

where σ_{rand} denotes the standard deviation of the binomial model and can be specified as:

$$\sigma_{rand} = \frac{e_i n_j (N - n_j - 1) + e_j n_i (N - n_i - 1)}{(N - 1)^2} \tag{7}$$

The statistical test proposed in 6 is a direct consequence of the fact that under sufficiently large partitions, binomial random variables can be well approximated by a normal distribution. Therefore, the connectivity between two clusters c_i and c_j is given by ψ_{ij} when $z_{ij} > t_c$ where t_c is a user-defined threshold.

(2) **Pseudotime computation:** As in prior works (7,9,28), MARGARET learns a temporal ordering over cells to uncover the dynamics of the underlying dynamic process. To determine the temporal order of the cells, for each cell, we infer *pseudotime*, which denotes the position of the cell in the underlying cell-state manifold representing the dynamic process (Figure 1C).

Given the kNN graph $\mathcal{G}$ and a prior starting cell index s, one possible way to estimate the pseudotime can be to compute the shortest-path distance of each cell from the starting cell s because the shortest-path distances better approximate the geodesic distances in a non-linear manifold (29). However, the kNN graph $\mathcal{G}$ can be inherently noisy due to spurious connections between cells. Hence, directly computing the shortest-path distances using $\mathcal{G}$ would give inaccurate estimates of the distance of each cell in the manifold from s. To mitigate this problem, we prune the kNN graph $\mathcal{G}$ by using our undirected graph $\mathcal{U}$ as a reference model.

Formally, given an undirected graph $\mathcal{U}$ and the kNN graph $\mathcal{G}$, we prune an edge between two cells x_{c_i} and x_{c_j} belonging to two clusters c_i and c_j respectively, iff $\exists\ e_{ij} \in \mathcal{G}$ and $\psi_{ij} = 0$ where e_{ij} represents a (short-circuit) edge between cells x_{c_i} and x_{c_j} in the kNN graph $\mathcal{G}$. Given a pruned kNN graph $\mathcal{G}^*$, we compute the shortest-path distance of each cell from a user-defined starting cell s to infer the pseudotime for each cell in the scRNA-seq dataset.

(3) **Orientation of edges in trajectory**: Given an undirected graph $\mathcal{U}$ and the pseudotime $\{\tau_n^p\}_{n=1}^N$, we compute the mean-pseudotime for each cluster c_i as:

$$m_{c_i} = \frac{\sum_{n=1}^N \mathbb{1}[x_n \in c_i]\tau_n^p}{\sum_{n=1}^N \mathbb{1}[x_n \in c_i]} \tag{8}$$

We then orient an edge from cluster c_i to c_j iff: $m_{c_i} lt; m_{c_j}$ and $\psi_{ij} \neq 0$ to obtain the final trajectory $\mathcal{T}$ with node-set $\mathcal{V}$ and directed edge-set $\mathcal{E}$ (Figure 1C).

Prediction of terminal states

Given a trajectory $\mathcal{T} = (\mathcal{V}, \mathcal{E})$, we compute the shortest-path betweenness (30) of every node in $\mathcal{T}$ (Supplementary Figure S1A). Formally, the shortest-path betweenness can be defined as:

$$b(v) = \sum_{(u,w) \in \mathcal{V}} \frac{d_{sp}(u, w|v)}{d_{sp}(u, w)} \tag{9}$$

where $b(v)$ is the betweenness for node v, $d_{sp}(u, w|v)$ represents the shortest-path distance between nodes u and w that passes through node v and $d_{sp}(u, w)$ represents the shortest-path distance between nodes u and w. Intuitively, the betweenness for any node $v \in \mathcal{V}$ is the sum of the fraction of all-pairs shortest paths that pass through node v thus indicating its importance in the network. Given the shortest-path betweenness values $b(v)_{v \in \mathcal{V}}$, we compute the median (b_{med}) and the median absolute-deviation (MAD) (b_{mad}) of the betweenness values respectively. A node v is added to the set of terminal states if

$$b(v) < (b_{med} - t_{TS} \times b_{mad}), \tag{10}$$

where t_{TS} is a user-defined scalar multiplier. Higher values of t_{TS} typically lead to nodes with no outgoing edges in $\mathcal{V}$ being selected as the terminal states. At the single-cell level, we select the cells having the maximum pseudotime value in each terminal state as the terminal cell for the underlying developmental process.

Inferring cell branch probabilities and differentiation potential

Similar to (9), we model differentiation as a stochastic process on our learned cell-state manifold where the cells can follow the paths in the pruned kNN graph $\mathcal{G}^*$ to reach any of the terminal states. Following (9), we model this stochastic process using an absorbing Markov Chain with the terminal cells acting as the absorbing states. Essentially, this formulation enables us to calculate the differentiation potential (DP) for each cell, a quantity that represents the potency of a cell to differentiate into specialized cell types. Given a set of terminal cells $\mathbb{T}_c$, for each cell i, we compute the branch probabilities $p_{b_{ij}} (j \in \mathbb{T}_c)$, which represents the probability of cell i reaching a terminal cell j (Figure S1B). Since DP of a cell (d_{p_i}) quantitatively characterizes the potency of a cell to maturate to different terminal states it can be obtained by computing the entropy of the branch probabilities of each cell as follows (Figure S1C):

$$d_{p_i} = -\sum_{j \in \mathbb{T}_c} p_{b_{ij}} \log p_{b_{ij}} \tag{11}$$

This is a reasonable way of modelling DP because cells with heterogeneous branch probabilities can be expected to have lesser potency of differentiating into diverse cell types. In the remainder of this section, we discuss the key steps involved in computing the branch probabilities for each cell *i*.

(1) **Waypoint sampling for scalable modelling of DP**: Given the scRNA-seq dataset with N cells, fitting an absorbing Markov Chain can be computationally intractable for large N. To scale our absorbing Markov Chain model to large datasets, we sample a subset of cells M, from the scRNA-seq dataset such that $M < <N$. We call these subset of cells as *landmarks* or *waypoints*. Specifically, given a set of K disjoint partitions of the embedding space y^*, we sample k waypoints per cluster by applying k-means++ initialization for each cluster. We use the refined clusters obtained from the Metric-learning step for sampling the waypoints. Using such a scheme for waypoint sampling has two main advantages. Firstly, using a kmeans++ like scheme in each cluster ensures high intra-cluster waypoint coverage. Secondly, sampling waypoints from each cluster guarantees coverage of the entire embedding landscape. In contrast, (31,32) use random sampling to compute waypoints which provides no coverage guarantees. Palantir (9) uses Max-min sampling (33) to compute waypoints which is more efficient than random sampling but might require a large number of waypoints to cover the embedding space.

(2) **Computing waypoint to terminal cell(s) probabilities**: Given a set of waypoints $\mathcal{W}$ consisting of M waypoints, we compute a nearest-neighbor graph $\mathcal{G}_w$ using the low-dimensional representations of $\mathcal{W}$. We then prune $\mathcal{G}_w$ using the short-circuit edge pruning (see Pseudotime computation) with the undirected connectivity graph $\mathcal{U}$ as a reference model. Furthermore, following (9), we also remove edges in $\mathcal{G}_w$ which violate the pseudotime ordering between waypoint cells. Formally, an edge $w_{ij} \in \mathcal{G}_w$ from waypoint w_i to w_j is pruned iff:

$$\tau_i^p > \tau_j^p + \alpha\sigma_{ij} \tag{12}$$

where τ_i^p and τ_j^p represent the pseudotime for waypoints w_i and w_j respectively and σ_{ij} represents the scaling factor for cell w_i given by the distance of w_i to its lth neighbor in $\mathcal{G}_w$. The statistical test in Equation (12) differs from the formulation in (9) in terms of the parameter α. We found that parameterizing Equation (12) with the user-defined parameter α provides an additional flexibility in controlling the number of edges that are pruned, which is an important aspect of fitting an absorbing Markov Chain.

Given a pruned waypoint nearest-neighbor graph $\mathcal{G}_w^*$ and a set of terminal cells $\mathbb{T}_c$, we row-normalize $\mathcal{G}_w^*$ to obtain the transition matrix T. For two waypoints w_i and w_j, the entry t_{ij} in T represents the probability of transitioning from waypoint w_i to w_j. An absorbing Markov Chain is specified by a transition matrix of the form $\begin{bmatrix} Q & R \\ 0 & I \end{bmatrix}$, where Q represents the transition probabilities of moving between intermediate states and R represents the transition probabilities of moving from intermediate states to the terminal states. We represent our transition matrix T using this formulation and compute the waypoint to terminal cell branch probabilities as follows:

$$P_{\mathcal{W}\mathbb{T}_c} = FR \tag{13}$$

where F is the fundamental matrix given by $F = (I - Q)^{-1}$. Since our Transition matrix T can be sparse, we recommend computing the fundamental matrix using the Moore–Penrose pseudoinverse to avoid numerical issues.

(3) **Computing cell to waypoint connectivity**: Given the pruned nearest neighbor graph $\mathcal{G}^*$, we compute cell to cell connectivity using a local random walk (LRW) (34) formulation. LRW is a quasi-local method to estimate connectivity between nodes in a graph based on limiting a random-walk to a fixed number of steps. Hence, the approach is computationally much more efficient than using a global random walk until convergence. Formally, the LRW connectivity $\zeta^{(i,j)}$ between two cells indexed by i and j is given by:

$$\zeta^{(i,j)}(t) = k_i\, p_{ij}(t) + k_j\, p_{ji}(t) \tag{14}$$

where $p_{ij}(t)$ and $p_{ji}(t)$ represent the probabilities obtained when moving from cell i to j and vice-versa at time t respectively. The constants k_i and k_j are set to $\frac{k}{|\mathcal{E}_k^*|}$ where k is the number of nearest neighbors and $|\mathcal{E}_k^*|$ is the total number of edges in $\mathcal{G}^*$. Given the similarity matrix $\mathcal{Z}$ representing LRW-based connectivities between different cells we can index $\mathcal{Z}$ to compute cell to waypoint similarities, $\mathcal{Z}_w$.

Given the cell to waypoint similarities $\mathcal{Z}_w$, and waypoint to terminal state probabilities $P_{\mathcal{W}\mathbb{T}_c}$, we compute the cell to terminal states branch probabilities $p_{b_{ij}}$ by a simple projection:

$$P_b = \mathcal{Z}_w P_{\mathcal{W}\mathbb{T}_c} \tag{15}$$

where P_b is a $N \times |\mathbb{T}_c|$ matrix representing branch probabilities for each cell. We then compute the DP for each cell using Equation (11).

Inference of gene expression trends along lineages

To visualize the variation in the expression of a gene *g* across different lineages with pseudotime, we fit generalized additive models (GAMs) to the gene expressions and pseudotime values (g_i, τ_i^p) of cells along a particular lineage j weighted by their branch probabilities $p_{b_{ij}}$ for that lineage (see Supplementary Note 2 in (9) for more details). We imputed the preprocessed expression value of gene g using MAGIC (21), when computing the lineage trends for the hematopoiesis and EB datasets. No imputation was performed for the colon IBD dataset. We used the LinearGAM implementation available in the *pyGAM* Python package (https://doi.org/10.5281/zenodo.1208724) to fit the lineage trends with the regularization penalty set to 10 and the order of the splines set to 4 for all the experiments.

Differential expression and gene ontology analysis

We used the Wilcoxon-rank sum test available in *Scanpy* (8) to estimate differentially expressed (DE) genes for each cluster and the Benjamini-Hochberg correction for adjusting the p-values. All genes were ranked in the DE analysis, which is the default behavior in Scanpy 1.7.2. To assess the functional significance of MARGARET inferred clusters, we performed Gene Ontology (GO) analysis using the *gprofiler-official* (35) package. To determine the GO terms for a cluster, we selected the top DE genes with a log fold change value greater than 1.0 and the adjusted p-value less than 0.05. In case more than 500 genes were included, we selected the top 500 genes to obtain a list of GO terms associated with that cluster.

Performance metrics for trajectory inference

Here we describe the quantitative performance metrics used for evaluating the trajectory inferred by a TI method:

(1) *Ipsen–Mikhailov (IM) similarity*: We used the IM distance (36) metric for global topology comparison between PAGA and MARGARET. Before computing the IM distance, the trajectories inferred from both the methods (including the ground-truth trajectory) were coerced to an undirected graph. Formally, the Laplacian spectra for a graph $\mathcal{G}$ can be specified as a mixture of Lorentz distributions (Eqn. 16) with the same half-width at half-maximum γ and centered at the frequencies ω_k given by $\omega_k = \sqrt{|\lambda_k|}$, where λ_k is the k^{th} eigenvalue of the graph Laplacian of $\mathcal{G}$. The constant C is a normalization constant for the resulting probability distribution.

$$\rho(\omega) = C \sum_{k=1}^{N-1} \frac{\gamma}{(\omega - \omega_k)^2 + \gamma^2} \tag{16}$$

The IM distance measures the difference between the Laplacian spectra of two graphs as follows:

$$IM(\rho_1, \rho_2) = \sqrt{\int_0^\infty (\rho_1(\omega) - \rho_2(\omega))^2 d\omega} \tag{17}$$

The IM similarity between two graphs can then be computed as $IS(G_1, G_2) = 1 - IM(G_1, G_2)$.
Since, among competing methods, only PAGA and VIA output a connected graph, we adapt the cell landscape in Monocle3 to the milestone framework by implementing the coarse trajectory inference procedure outlined in (7). Similarly, for Palantir (which outputs a continuous cell landscape), we apply PAGA using the cell embeddings inferred using Palantir to convert a continuous trajectory to a network of milestones.

(2) *Rank correlation metrics*: We used two rank correlation metrics to compare the ground truth ordering and pseudotime orderings inferred by the TI methods. More specifically, we used the *Kendall's tau (KT)* and the *Spearman's rank (SR)* correlation coefficients to estimate the similarity in rank orderings of the data. The KT correlation coefficient can be specified as follows:

$$\tau_b = \frac{P - Q}{\sqrt{(P + Q + T) * (P + Q + U)}} \tag{18}$$

where P and Q represent the number of concordant and discordant pairs, respectively. T and U represent the number of ties in the two orderings, respectively. The SR correlation coefficient is simply defined as the Pearson's correlation coefficient applied to the ranks of the variables measured in the two orderings. We used the *scipy.stats* package to compute both KT and SR correlation coefficients.

(3) *Combined score*: We benchmarked all candidate methods based on their performance in capturing the global topology and the pseudotime ordering. We first computed the IM similarity between the inferred trajectory and the ground truth trajectory. We set the IM similarity score for a method as the best score that can be obtained when using either Louvain or Leiden clustering at resolution 1.0. Similarly, we also computed the KT correlation between the predicted and the ground-truth pseudotime orderings. We then perform min-max normalization of the IM and KT scores obtained by all candidate methods separately for the multifurcating and the disconnected datasets. The combined score for a method is then obtained by averaging the normalized IM and KT scores of that method applied across the simulated datasets.

(4) *Clustering metrics*: We used *adjusted Rand index* (ARI) and *normalized mutual information* (NMI) metrics to assess the clustering performance of MARGARET embeddings. The ARI metric corrects the Rand index for chance and is specified by the following formulation:

$$ARI = \frac{\sum_{ij} \binom{n_{ij}}{2} - [\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}]/\binom{n}{2}}{\frac{1}{2}[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}] - [\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}]/\binom{n}{2}} \tag{19}$$

where a_i, b_j and n_{ij} are values from the contingency table which is used to estimate the overlap between two partitionings. The NMI metric between two clusterings C_1 and C_2 can be formulated as follows:

$$NMI = \frac{\mathbb{I}(C_1, C_2)}{\sqrt{\mathbb{H}(C_1)\mathbb{H}(C_2)}} \tag{20}$$

where $\mathbb{I}$ is the mutual information between the two clusterings C_1 and C_2 and $\mathbb{H}$ is the Shannon-Entropy of the clustering.

Visualization of embedding

We used UMAP (12) (with default parameters) for visualization of all the datasets except the early Hematopoiesis scRNA-seq data provided by (9), for which we used tSNE (37) visualization with a perplexity value of 180 for replicate 1 and 150 for replicate 2.

Computing Crypt-axis score

As suggested in (19), we used the expression of the following genes to define the crypt-axis (CA) score: *SEPP1*, *CEACAM7*, *PLAC8*, *CEACAM1*, *TSPAN1*, *CEACAM5*, *CEACAM6*, *IFI27*, *DHRS9*, *KRT20*, *RHOC*, *CD177*, *PKIB*, *HPGD* and *LYPD8*. The final crypt-axis score for each cell was then computed by summing over the normalized expression (between 0 and 1) values of each gene included in our set.

Simulation of benchmark datasets

To benchmark MARGARET's performance on several aspects of trajectory inference, we generated a suite of synthetic single-cell gene expression datasets representing different complex trajectory models. We used *dyntoy* (13) (https://github.com/dynverse/dyntoy) to generate the synthetic benchmark (see Supplementary Table S3 for the details of different synthetic datasets) spanning multifurcating and disconnected topologies. For benchmarking MARGARET, PAGA (28) and Monocle3 (7), all the simulated datasets were subjected to the same data preprocessing steps following Seurat (24): removal of genes expressed in less than 3 cells, normalization, log transformation with a pseudo count of 1.0, and finally z-score normalization. For benchmarking *Palantir*, the steps proposed by the authors in (9) were used to preprocess the simulated datasets.

RESULTS

MARGARET outperforms other TI methods on a diverse simulated benchmark

We benchmarked MARGARET's performance on a variety of synthetic datasets (Materials and Methods) consisting of multifurcating and disconnected trajectories (with complex multifurcating components)(Supplementary Table S3) against state-of-the-art TI methods—Partition-Based Graph Abstraction (PAGA) (28), Palantir (9), Monocle3 (7) and VIA (14).

Figure 2A and B qualitatively compares the trajectories inferred by the algorithms on a multifurcating and a disconnected dataset sampled from our simulated benchmark (Figure 2A). Monocle3 correctly captured the overall topology of the disconnected dataset but underestimated the number of clusters in the two components severely. PAGA and Palantir failed to capture the disconnected topology altogether as PAGA represented two components as cyclic topologies while Palantir represented the entire embedding landscape as a single component (Figure 2B). VIA failed to capture the correct number of components in the disconnected trajectory. Moreover, the directionality in the VIA inferred trajectory suggested a *converging* topology in the larger of the two components leading to highly inaccurate pseudotime estimates. MARGARET outperformed the other algorithms on this dataset by most closely recovering the underlying global topology accompanied by accurate detection of terminal state clusters. For the multifurcating dataset, PAGA and Palantir failed to capture the correct branching topology, whereas Monocle3 failed to recover the global topology as it inferred the overall trajectory as two separate components (Figure 2B). In contrast, VIA and MARGARET were able to capture the underlying multifurcating topology accurately. However, VIA failed to detect multiple branchings and terminal states in the trajectory and the MARGARET captured topology was qualitatively the most similar to the ground truth topology.

We quantitatively benchmarked MARGARET's performance against other methods on two main aspects of TI: *accuracy in inferring the global topology* and *accuracy of pseudotime ordering*. We assessed the effectiveness of a TI method in recovering the underlying global topology of cells by evaluating the Ipsen-Mikhailov (IM) similarity (Materials and Methods) between the ground truth and the inferred trajectory graph. To compare two undirected trajectories (represented as graphs), we adopted the network of 'milestones' representation framework as proposed by (13). Next, we assessed the accuracy in recovering the pseudotime ordering of cells by computing the Kendall's tau (KT) and the Spearman rank (SR) correlation (Materials and Methods) between the ground truth and the inferred ordering. We then assign a combined score (see Materials and Methods) to each TI method which represents the efficacy of the method in capturing both global topology and ordering information. Figure 2C presents a comparison between MARGARET and other methods on our simulated benchmark.

For the disconnected benchmark, MARGARET outperformed other methods on all the five datasets with up to 19.2 % improvement over the next best method. Similarly, for the multifurcating benchmark, MARGARET achieved up to 18.57% improvement over the next best method. The superior performance of MARGARET on both benchmarks suggests its ability to perfectly capture diverse trajectory types. In contrast, other methods performed well for only one type of trajectory (either disconnected or multifurcating) but poorly for the other trajectory type. For example, PAGA and Palantir performed well for the multifurcating datasets but their combined score was lower compared to other methods for the disconnected datasets. On the other hand, Monocle3 and VIA, while performing better than PAGA and Palantir for the disconnected datasets, performed poorly for the complex multifurcating trajectories. Moreover, qualitative analysis of the results on the mutlifurcating benchmark suggests Monocle3 to be biased towards capturing disconnected components as it partitioned several multifurcating datasets in our benchmark into independent disconnected components (data not shown). The detailed comparison of the TI methods for the global topology and pseudotime ordering tasks (Supplementary Figure S2) demonstrated MARGARET's superior performance over other methods in reconstructing the underlying global topology and recovering the pseudotime order for the majority of the datasets in our benchmarks. Across a range of clustering resolutions and clustering methods, MARGARET consistently performed well (Supplementary Figures S3 and S4) illustrating its robustness towards these parameters. It is worth noting that for evaluating VIA on the global topology and the pseudotime ordering tasks, we used PARC clustering (38) which can also be used with MARGARET. Moreover, for VIA, we performed coarse analysis as we found the VIA inferred IM and correlation scores to worsen after the fine analysis.

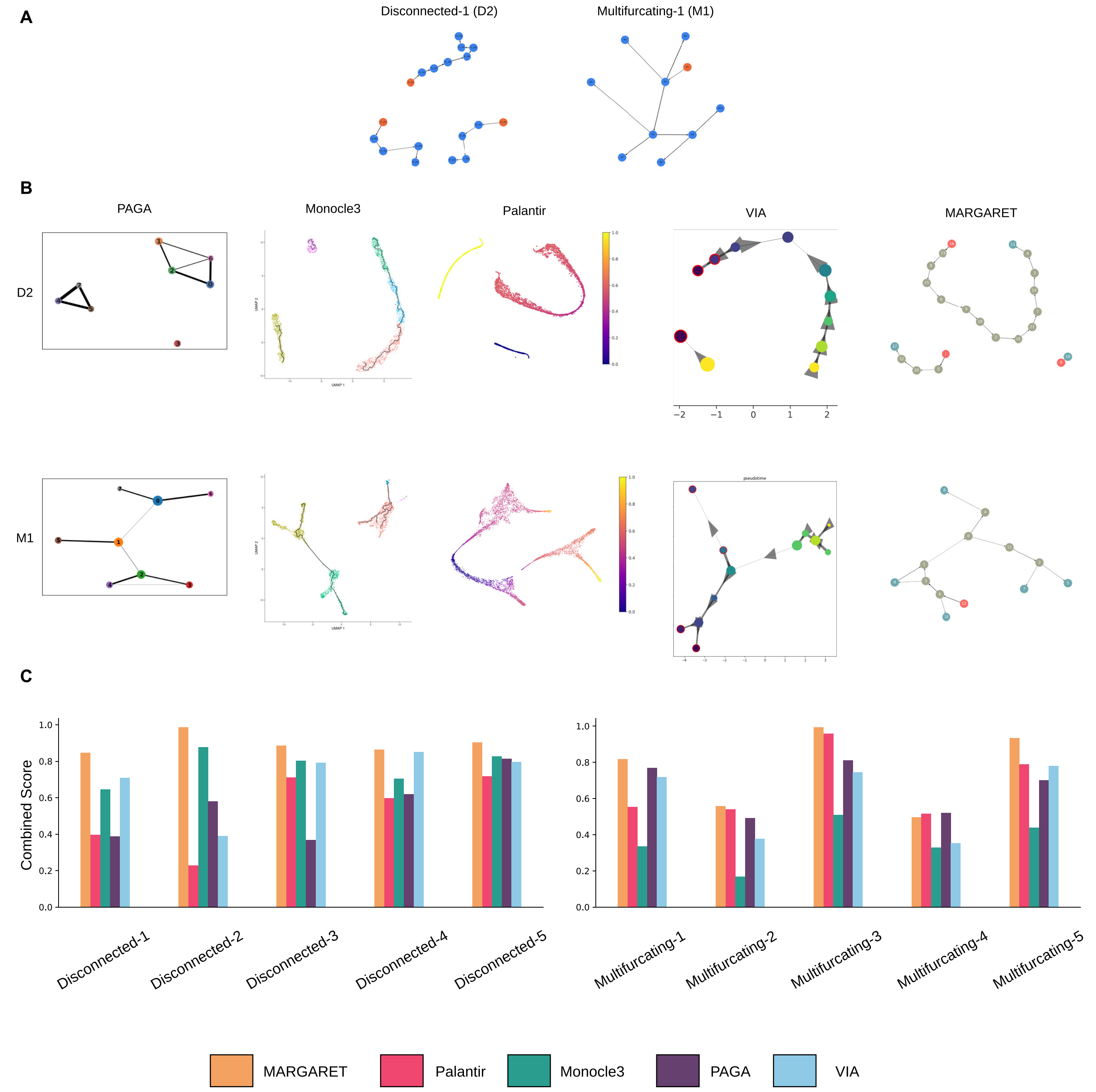

Figure 2. MARGARET outperforms state-of-the-art TI methods on a simulated benchmark. MARGARET outperformed state-of-the-art TI methods on qualitative and quantitative metrics when applied to a simulated benchmark consisting of ten datasets with diverse trajectory types. (**A**) (left) A sample disconnected dataset (4929 cells), and (right) a sample multifurcating dataset (5000 cells) (**B**) Visualization of the cell embedding landscape inferred by different TI methods on the disconnected (Top) and multifurcating (bottom) datasets. Palantir embedding landscapes shown with projected pseudotime values. Monocle3 inferred trajectory graph projected on the embedding with colors denoting cluster information. PAGA, VIA and MARGARET outputs shown as connectivity graphs. PAGA, Monocle3 and MARGARET inferred graphs were computed using Leiden clustering while VIA inferred graphs were computed using PARC clustering. The clustering resolution was set to 0.4 for all methods (except Monocle3 which selects the best resolution). For the MARGARET connectivity graph, the red nodes denote starting cell clusters and the cyan nodes denote detected terminal states. (**C**) Combined score (higher score is better, see Materials and Methods) comparison between MARGARET and other TI methods for the disconnected (left) and the multifurcating (right) datasets (see Supplementary Figure S3 and Supplementary Table S2 for the detail comparison of IM similarity, Kendall's tau (KT), and SpearmanRank (SR) correlation metrics.

We also applied MARGARET to a simulated dataset with a cyclic ground-truth trajectory (Cyclic_1) (Supplementary Figure S5A, Supplementary Table S3A). For this dataset, MARGARET, VIA and PAGA were able to capture the global topology accurately while Palantir and Monocle3 failed to capture the cyclic structure (Supplementary Figure S5B).

Thus, our analysis shows the broad applicability of MARGARET to a diverse suite of trajectory types compared to other TI methods that usually perform well on one type of trajectory but fail to model other trajectory types.

MARGARET outperforms competing TI methods on a real dataset benchmark

We next compared the performance of different TI methods on real biological datasets. For this evaluation, we selected two multifurcating (oligodendrocyte differentiation (16) and planaria parenchyme differentiation (17)) and two disconnected (placenta trophoblast differentiation (15) and mouse cell atlas (15)) datasets (see Supplementary Table S3 for dataset statistics) with ground-truth trajectories available as a network of milestones representation (13).

Figure 3A shows MARGARET inferred trajectories and Supplementary Figure S6 shows the qualitative performance of other state-of-the-art TI methods on the real dataset benchmark. For the datasets with disconnected reference trajectories (*placenta trophoblast differentiation* and *mouse cell atlas*), only MARGARET inferred trajectories correctly captured the number of disconnected components while accurately preserving different cell types specific to each component (see cell-type annotations in Figure 3A (i-ii)). VIA and Palantir were unable to capture the underlying disconnected topology for both the datasets. PAGA was able to capture the number of components in the mouse cell atlas dataset but failed to capture the disconnected topology for trophoblast differentiation. While Monocle 3 was able to infer disconnected trajectory for both these datasets, it also overestimated the number of disconnected components for both datasets. Quantitative comparison on these datasets (Figure 3B) revealed MARGARET to be the best performer which achieved 30.56–97.13% improvement in combined score over the next best performing method indicating its superiority for both global topology and cell ordering tasks. For topology inference task, PAGA performed the worst and exhibited the lowest IM similarity (or highest IM distance) scores while VIA, Monocle 3 and Palantir performed similarly (Supplementary Figure S7A). MARGARET achieved up to 21.04% improvement (over the next best method) in lowering IM distance. For the cell ordering task, MARGARET outperformed the other methods by a large margin (KT correlation of MARGARET for *placenta trophoblast differentiation* was 0.53 as compared to 0.02 by the next best method Palantir, ∼ 23% improvement for mouse cell atlas). Monocle 3 exhibited the lowest KT correlation scores among all methods (Supplementary Figure S7B). Particularly for placenta trophoblast differentiation, only MARGARET was able to recover the temporal order of cells which other methods failed to capture leading to near zero or negative KT and SR correlation scores (Supplementary Figure S7B).

For the datasets with multifurcating reference trajectories (*oligodendrocyte differentiation* and *planaria parenchyme differentiation*), Monocle 3 incorrectly captured the underlying trajectory as a disconnected graph further justifying our observation that Monocle 3 might be biased towards disconnected trajectories. While all other methods were able to capture the multifurcations in the underlying trajectory (Supplementary Figure S6), MARGARET outperformed all the methods based on the combined score (Figure 3B, up to 13.37% improvement over the next best method). PAGA performed the worst on the global-topology task with other methods exhibiting comparable performance while Monocle3 performed the worst on the cell ordering task with consistent negative KT scores on both multifurcating datasets. MARGARET consistently exhibited high IM similarity, KT, and SR scores on both datasets (Supplementary Figure S7). We further visualized the spectrum, $\rho(\omega)$, for the ground-truth trajectory and the trajectory inferred by different TI methods for one multifurcating (oligodendrocyte differentiation, Supplementary Figure S7C) and one disconnected (mouse cell atlas, Supplementary Figure S7D) dataset which further demonstrated superior overlap of the spectrum for MARGARET inferred trajectory with that of the ground-truth trajectory spectrum as compared to that of other methods. These results demonstrate that while other TI methods are more suited towards specific trajectory types, MARGARET can generalize much better for different types of trajectories.

MARGARET generalizes the quantification of differentiation potential for complex trajectories

Differentiation potential (DP) as introduced by Palantir measures cell fate plasticity along a trajectory and can also characterize key events in the underlying dynamic process (9). However, Palantir is able to quantify DP only for connected trajectories as it models the trajectory as a continuum of states. To evaluate whether MARGARET's DP formulation can generalize to more complex disconnected trajectories, we benchmarked MARGARET's DP inference against that of Palantir for two complex simulated disconnected trajectories (Disconnected_4 and Disconnected_6) (Figure 4, Supplementary Table S3A). For both the datasets, MARGARET accurately captured the DP trends in different disconnected components and MARGARET inferred DP showed high negative Spearman-Rank correlation with the ground-truth pseudotime. DP is expected to have negative correlation with pseudotime (9) since as cells differentiate and proceed towards terminal states, pseudotime increases while the DP of the cell decreases. In contrast, Palantir's DP inference performed poorly for disconnected trajectories as observed from the correlation analysis between DP and ground truth pseudotime (Figure 4). The poor performance of Palantir could be the result of its single component assumption when inferring the DP which does not extend to multiple independent disconnected components within the same dataset. Hence, for both the disconnected datasets, the DP inference by MARGARET was superior to that of Palantir indicating

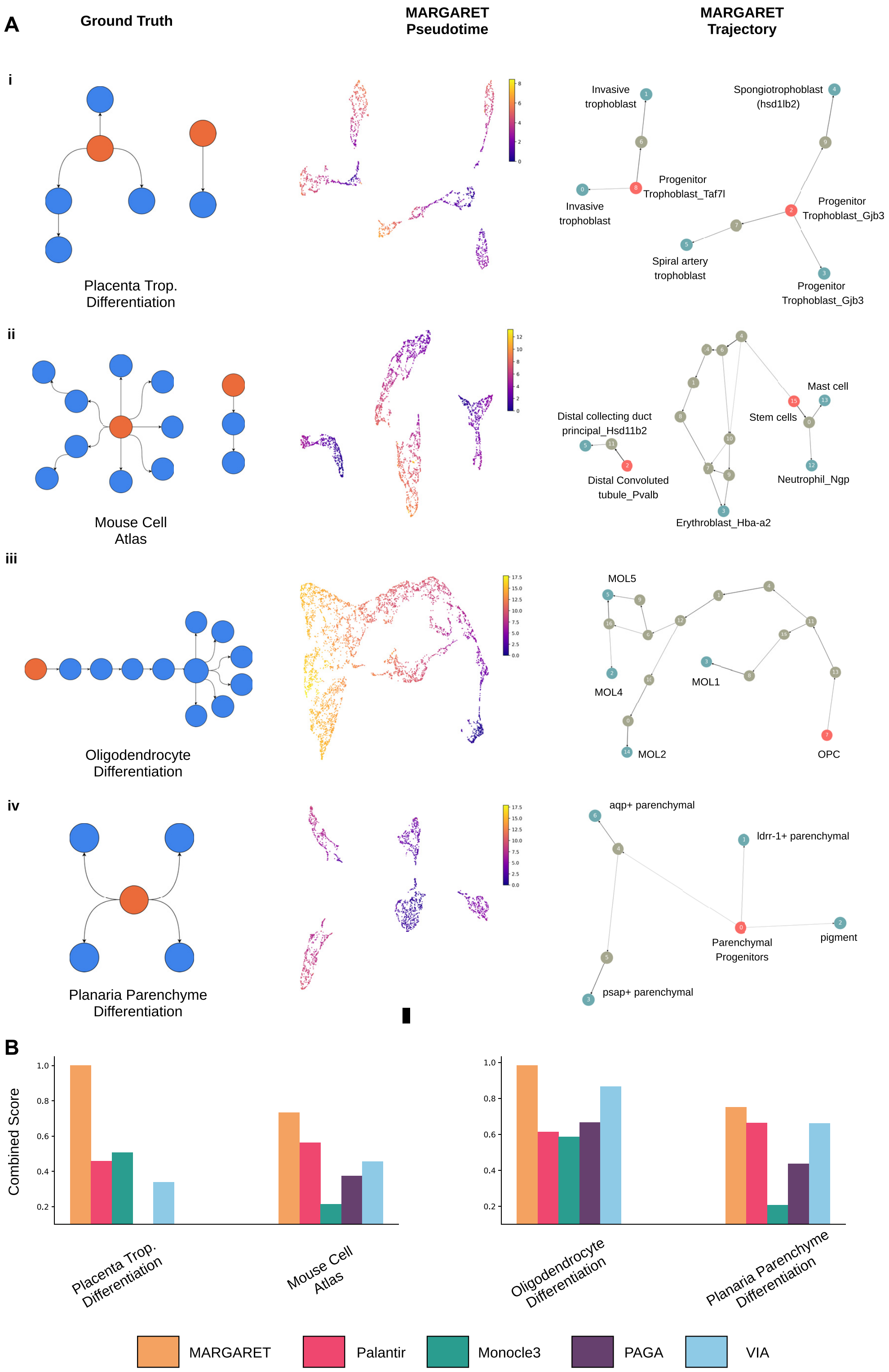

Figure 3. MARGARET outperforms state-of-the-art TI methods on a real benchmark. MARGARET outperformed state-of-the-art TI methods on qualitative and quantitative metrics when applied to a real benchmark consisting of four datasets with disconnected (placenta trophoblast differentiation and mouse cell atlas) and multifurcating (oligodendrocyte differentiation and planaria parenchyme differentiation) trajectories. (**A**) Visualization of MARGARET inferred pseudotime and trajectories on the real dataset benchmark. The first column shows the ground-truth trajectory represented as network-of-milestones. The second column shows MARGARET inferred pseudotime projected on the 2D embeddings. The third column shows MARGARET inferred trajectories where initial and terminal states were annotated with ground-truth cell-type information. (**B**) Combined score (higher score is better, see Materials and Methods) comparison between MARGARET and other TI methods for the disconnected (left) and the multifurcating (right) real datasets (see Supplementary Figure S7 and Supplementary Table S4 for a detailed comparison of IM distance, Kendall's tau (KT), and Spearman-rank (SR) correlation metrics).

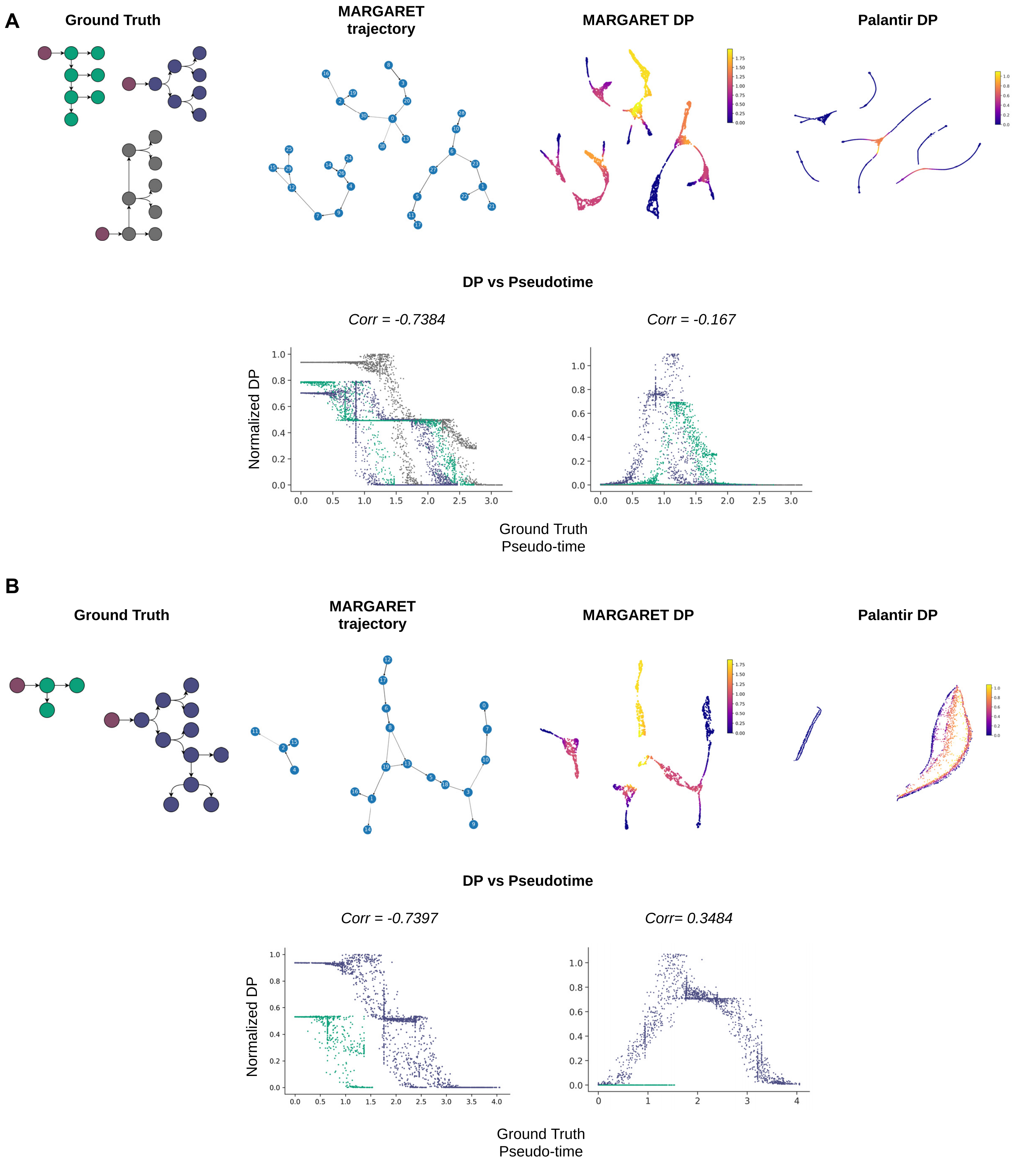

Figure 4. MARGARET inferred DP generalizes to disconnected trajectories. (**A**) (top-left) Schematic of a ground truth simulated dataset with three disconnected components (Disconnected-4, 7500 cells). The starting cells are shown in red color while the trajectories can be uniquely identified by three different colors. (Top-Middle-1) MARGARET inferred trajectory. (Top-middle-2) MARGARET inferred DP for this dataset. Inferred DP trends are qualitatively consistent across independent components. (Top-right) Palantir inferred DP. The results are qualitatively inconsistent as Palantir is unable to capture the disconnected nature of the underlying trajectory. (Bottom) Scatter plot of Normalized DP vs ground truth pseudotime for both MARGARET (left) and Palantir (right). MARGARET shows a higher negative Spearman-rank correlation with the ground truth pseudotime as compared with Palantir. The cells in the scatter-plot are colored by their disconnected component id in the ground-truth trajectory. (**B**) Same as (A) but for a disconnected dataset with two components (Disconnected-6, 2500 cells).

the applicability of MARGARET's DP inference to a wider variety of trajectories.

MARGARET learns modular cluster representations during episodic training

Since MARGARET's metric-learning approach is aimed at inferring a lower-dimensional cell-state manifold where the distinct cell states are represented by compact cell clusters, we evaluated the compactness of the inferred clusters by tracking the clustering modularity scores and the number of clusters inferred at the end of each episode during training of two single-cell RNA sequencing human hematopoiesis datasets (9) using Phenograph (39) across Louvain and Leiden clustering backends. For both datasets, at the end of each training episode, MARGARET improved upon the modularity score of the previous episode before finally converging (Supplementary Figure S8). The clustering modularity score can also be used for estimating MARGARET's convergence. Interestingly, the number of inferred clusters did not always increase monotonically with the number of episodes (Supplementary Figure S8A–C), suggesting that the increase in modularity is due to the improved clustering quality at each episode.

MARGARET can refine the single-cell embeddings learned using other methods

To investigate the ability of MARGARET's unsupervised metric learning-based approach, we initialized MARGARET with *d*-dimensional cell embeddings (as obtained from a linear or nonlinear dimension reduction method such as PCA or scVI (40)) to infer MARGARET-refined *d*-dimensional embeddings. The quality of MARGARET-inferred cell embeddings was evaluated by comparing its cell-type clustering performance (as measured by adjusted rand index (ARI) and normalized mutual information (NMI) metrics) against that of the initial embeddings on a suite of biological datasets for which the ground-truth clustering annotations were available (Methods). As compared to scVI-inferred cell embeddings, MARGARET's refined embeddings achieved higher ARI and NMI scores across all the datasets (Supplementary Figure S9A). Similar results were observed using a PCA-based initialization (Supplementary Figure S9B), which suggests that MARGARET's metric learning-based approach can refine the latent representations captured during earlier dimension reduction stages. We also assessed the impact of using different initial clustering methods (Supplementary Note 1, Supplementary Figure S10) and training hyperparameters with MARGARET (Supplementary Note 1, Supplementary Figure S11) and our analysis showed that MARGARET's metric learning approach is robust to the choice of initial clustering method and training hyperparameters.

MARGARET correctly predicts human hematopoietic differentiation trajectory and associated gene expression changes

Due to the availability of established lineage-specific markers, hematopoiesis (41) has been used as a model biological system by several trajectory inference methods (9,31). Using MARGARET, we first explored early human hematopoiesis where hematopoietic stem cells (HSCs), through a hierarchy of progenitors and bifurcation events, give rise to different mature cell types (41). We applied MARGARET to two human bone marrow scRNA-seq datasets (10X Chromium) (9) (replicates 1 and 2 consisting of 5780 and 6501 cells, respectively). For both datasets, MARGARET correctly identified all major hematopoietic cell types, including hematopoietic stem cells (HSCs), common lymphoid and myeloid progenitors (CLPs and CMPs respectively), as well as cells committed towards erythroid (erythrocytes and megakaryocytes), monocytic and dendritic cell (classical and plasmacytoid dendritic cells (cDCs and pDCs)) lineages (Figure 5A, Supplementary Figure S12A). The cell type clusters inferred by MARGARET were characterized by the expression of key marker genes, obtained by manually curating a set of marker genes for major hematopoietic cell types through prior literature review (9,42–44) (Figure 5D, Supplementary Figures S12F and 13). The expression of the marker genes corresponding to major hematopoietic cell types correlated well with the topology of the MARGARET inferred trajectory (Supplementary Figure S14). We utilized the starting cell information provided by (9) for pseudotime inference (Figure 5B, Supplementary Figure S12C). For both replicates, MARGARET inferred pseudotime follows expected progression, where the pseudotime increases as cells progress towards more specialized cell types from *CD34* enriched stem cells. Moreover, the probability of cells branching to different lineages diminishes as cells commit towards specific lineages (Supplementary Figures S15–S17). Consequently, as expected, MARGARET inferred DP (Figure 5C, Supplementary Figure S12B) decreases as we move towards terminal states in the trajectory since commitment towards a specific lineage is accompanied by a gradual reduction in cell plasticity. Figure 5E and Supplementary Figure S12D represent the annotated hematopoietic trajectory inferred by MARGARET for the two replicates, where the arrows represent transition between cell types.

To validate MARGARET inferred trajectories, we computed expression trends for essential marker genes for all major hematopoietic lineages (Figure 5f). As expected, expression of *CD34* decreases with increasing pseudotime as cells commit to particular lineages (41). In contrast, *CD79B* is selectively upregulated in the lymphoid lineage (45) while *MPO* and *IRF8* are upregulated in the monocyte (46) and dendritic cell (DC) (47) lineages, respectively. *ITGA2B* and *GATA1* are selectively upregulated in the megakaryocytic (48) and erythroid (49) lineages, respectively. Similar expression trends were observed for replicate 2 demonstrating MARGARET's robustness (Supplementary Figure S12E).

MARGARET characterized progenitor populations for monocytic and dendritic cell lineages

Interestingly, we observed an initial upregulation in *MPO* expression in both the monocyte and DC lineages (Figure 5F). However, with pseudotime progression, *MPO* expression is upregulated in the monocyte lineage but gets downregulated in the DC lineage. To explore the branching of monocyte and DC lineages from CMPs in replicate

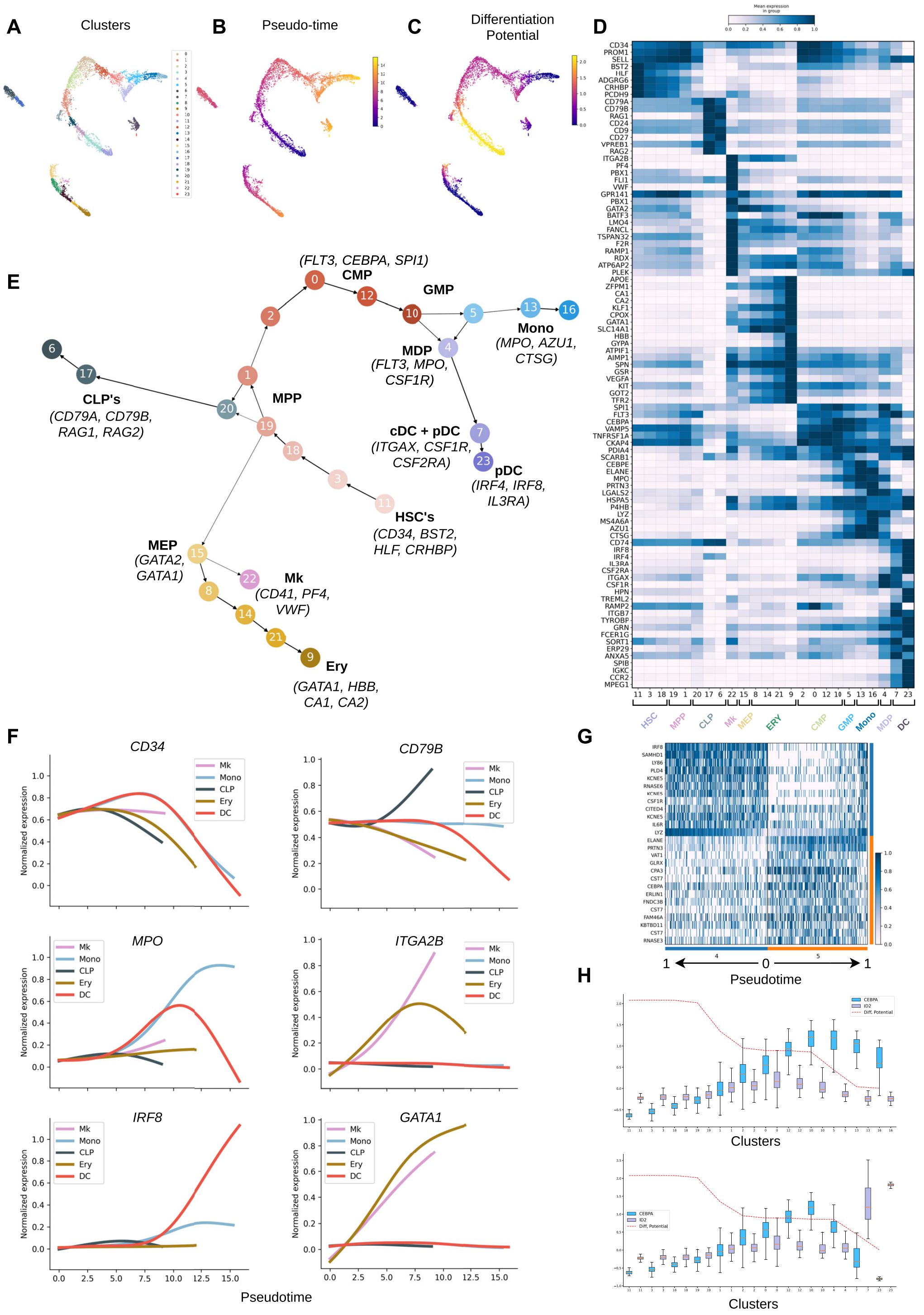

Figure 5. MARGARET delineates major lineages and identifies important transcriptional switches in early human hematopoiesis. Analysis of scRNA-seq data for human hematopoiesis replicate 1 by MARGARET. (**A**) tSNE plot of cell-state embedding inferred by MARGARET for the human hematopoiesis dataset, cells are colored by MARGARET inferred clusters. (**B**) MARGARET pseudotime and (**C**) differentiation potential calculated using one HSC as a start cell. (**D**) Heat map for marker genes for all MARGARET inferred clusters. (**E**) MARGARET inferred trajectory annotated with cell-type and lineage information (important marker genes are mentioned within parantheses with the cell type annotation). Ery: Erythrocyte; Mk: Megakaryocyte; MEP: Megakaryocyte-Erythroid Progenitors; HSC: Hematopoeitic Stem Cells; CLP: Common Lymphoid Progenitors; CMP: Common Myeloid Progenitors; GMP: Granulocyte-Monocyte Progenitors; MDP: Monocyte-Dendritic Cell Progenitors; cDC: Classical Dendritic Cells; pDC: Plasmacytoid Dendritic Cells; Mono: Monocytes. (**F**) Gene expression trends for essential genes for major inferred lineages. (**G**) Differential expression (DE) analysis between cluster 4 (MDP) and cluster 5 (GMP) (**H**) Variation of *CEBPA* and *ID2* gene expressions in the monocyte (Mono) lineage (top) and the dendritic cell (DC) lineage (bottom) for replicate 1. The boxplots summarize the expression of the gene in each cluster in the lineage, where the box depicts the interquartile range (IQR, the range between the 25th and 75th percentile) with the median value, whiskers indicate the maximum and minimum value within 1.5 times the IQR. The red dotted line represents the mean differentiation potential for each cluster in the lineage.

1, we investigated the transitions from cluster 10 to cluster 4 and cluster 10 to cluster 5 that are also associated with substantial changes in DP indicating that these transitions accompany important molecular events corresponding to lineage commitment (9). At the transition from CMPs to monocytes, we observed elevated expressions of monocyte markers ((42)) including *CEBPA*, and *CEBPE* (Figure 5D, H (top), Figure 6A), which correlated with a decrease in DP. The transcription factor (TF) *CEBPA* plays a crucial role in cell fate decisions in granulocyte-monocyte progenitors (GMPs) (50,51) to differentiate into granulocyte and monocyte (42). Apart from elevated expression of *CEBPE* and monocyte markers *MPO*, *LYZ* and *MS4A6A* (43), the monocyte clusters also strongly expressed granule genes such as *CTSG*, *PRTN3* and *ELANE* (Figure 5D), each of which were also highly correlated (>0.92) with the MARGARET inferred monocytic branch probabilities (Figure 6C). Monocytes derived from granulocyte-monocyte progenitors (GMPs) are known to express these granule proteases (52). Therefore, expression of *CEBPA*, *CEBPE* and granule proteases and reduced *FLT3* expression in cluster 5 (Figure 6A (left)) indicates the presence of GMPs in cluster 5 (52). We also observed similar trends for replicate 2 (Supplementary Figure S18).

In contrast, we observed elevated expressions of *FLT3* in cluster 4 (Figure 6A (right)) which also correlated with a decrease in DP on the transition from cluster 10 ($FLT3^+$ CMP) to cluster 4. Moreover, this cluster showed elevated expression of *CSF1R* (*CD115*), dendritic cell marker *ITGAX* (*CD11c*) (Figure 6B), and *MPO* (Figure 5D). Altogether, the $FLT3^+CD115^{hi}$ signature of this cluster suggests the presence of monocyte-dendritic cell progenitors (MDPs) in cluster 4 which gives rise to both monocytes and dendritic cells (52). Therefore, MARGARET inferred branching structure in the myeloid lineage characterized the differentiation of monocytes and dendritic cells from GMPs and MDPs respectively and these important events were also marked by a decrease in MARGARET inferred DP. These progenitor populations and their lineage branchings were not characterized in the original study (9) that used Palantir.

MARGARET characterized the heterogeneity in DC lineage

In the DC lineage for replicate 1, MARGARET inferred cluster 7 expressed markers for both cDCs (*ITGAX, CLEC10A*) and pDCs (*IRF7, IRF8, IL3RA*) (Figure 5D) while cluster 23 only expressed pDC markers. We made a similar observation in replicate 2 with clusters 1 and 20 equivalent to clusters 7 and 23 in replicate 1 , respectively. To characterize the heterogeneity in the DC lineage at a finer resolution, we combined the cells in the DC lineage from clusters 4, 7 and 23 in replicate 1 and clusters 5, 1 and 20 in replicate 2 and analyzed the resulting 1406 cells using MARGARET. Figure 6D shows the MARGARET inferred trajectory, consisting of 10 clusters (DC0-DC9). Since MDPs give rise to DC populations, based on the expression of MDP-specific markers *CSF1R*, and *ITGAX* (52) (Figure 6B) we inferred cluster DC7 as the starting cluster for our analysis. Based on manually curated set of markers specific to pDCs and cDCs through prior literature review (53–55), we then identified pDCs marked by high expression of *E2-2* (*TCF4*) TF, its target TFs *IRF7*, *SPIB*, and pDC-specific marker genes *LILRA4* (*ILT7*) (56) and *PACSIN1* (57), in cluster DC0 (Figure 6E (top), F) which was a terminal state. *E2-2* expression serves as a key event in pDC cell fate choice (58), *IRF7* is a key regulator of IFN expression and has been shown to be highly expressed in pDCs as compared to other cell types (59). Similarly, we localized the cDC lineage by observing high expression of *ITGAX* (*CD11b*), *ID2* and *CD1c* (60) in clusters DC3 and DC8 (Figure 6F). Furthermore, we observed high expression of cDC2-specfic TFs *NR4A3, SREBF2* and marker genes *CLEC10A, CD1E, CLEC12A, CX3CR1* (54) (Figure 6E (bottom), F) and negligible expression of cDC1 marker gene *CLEC9A* (61) in clusters DC3 and DC8 (data not shown) indicating the presence of cDC2 and absence of cDC1 cells in these clusters.

MARGARET characterized erythroid-megakaryocytic lineage

We also characterized the erythroid-megakaryocytic lineage branching in MARGARET inferred trajectory. Both erythroid and megakaryocytic commitment were associated with a sharp decrease in DP (Supplementary Figure S19). In the erythroid lineage, this decrease in DP was concordant with the elevated expressions of TFs *GATA1*, *KLF1* and *MYB*, which are known to play crucial roles in erythropoiesis: *GATA1* is indispensable for erythropoiesis (49), *KLF1* modulates erythroid cell differentiation by regulating erythroid precursor genes and also antagonizes megakaryocyte differentiation (42,62), *MYB* enhances erythropoiesis by suppressing megakaryopoiesis (63). Expression of these TFs also highly correlated with the erythroid branch probabilities (>0.9) indicating their crucial regulatory role in erythroid commitment (Supplementary Figure S20A). In the megakaryocyte lineage, the drop in DP was concordant with increasing expression of transcription factors *PBX1*, *FLI1*, and *MEIS1* (Supplementary Figure S19), which were also closely correlated (>0.9) with megakaryocytic branch probabilities (Supplementary Figure S20B). These TFs are known to play central role in megakaryopoiesis: *FLI1* and *PBX1* are essential TFs for megakaryocyte differentiation (42,64), and *MEIS1* is essential for fetal megakaryopoiesis (65). The cluster 15 (in replicate 1) at which the erythroid and megakaryocytic lineages diverged, expressed both *GATA2* (driver of erythroid commitment (66)) and *CD41* (responsible for megakaryocytic lineage commitment (42)), as well as genes like *SLC14A1* and *VWF*, which are responsible for a continuous transition from megakaryocyte-erythroid progenitors (MEP) to erythroid and megakaryocyte progenitors respectively (44) suggesting the presence of MEPs in cluster 15.

MARGARET applied to early human embryogenesis data

To investigate MARGARET's ability in extracting novel insights from a complex biological system, we applied MARGARET to an scRNA-seq dataset generated from embryoid bodies (EB) (18), which recapitulates the differentiation process in early embryogenesis where pluripotent em-

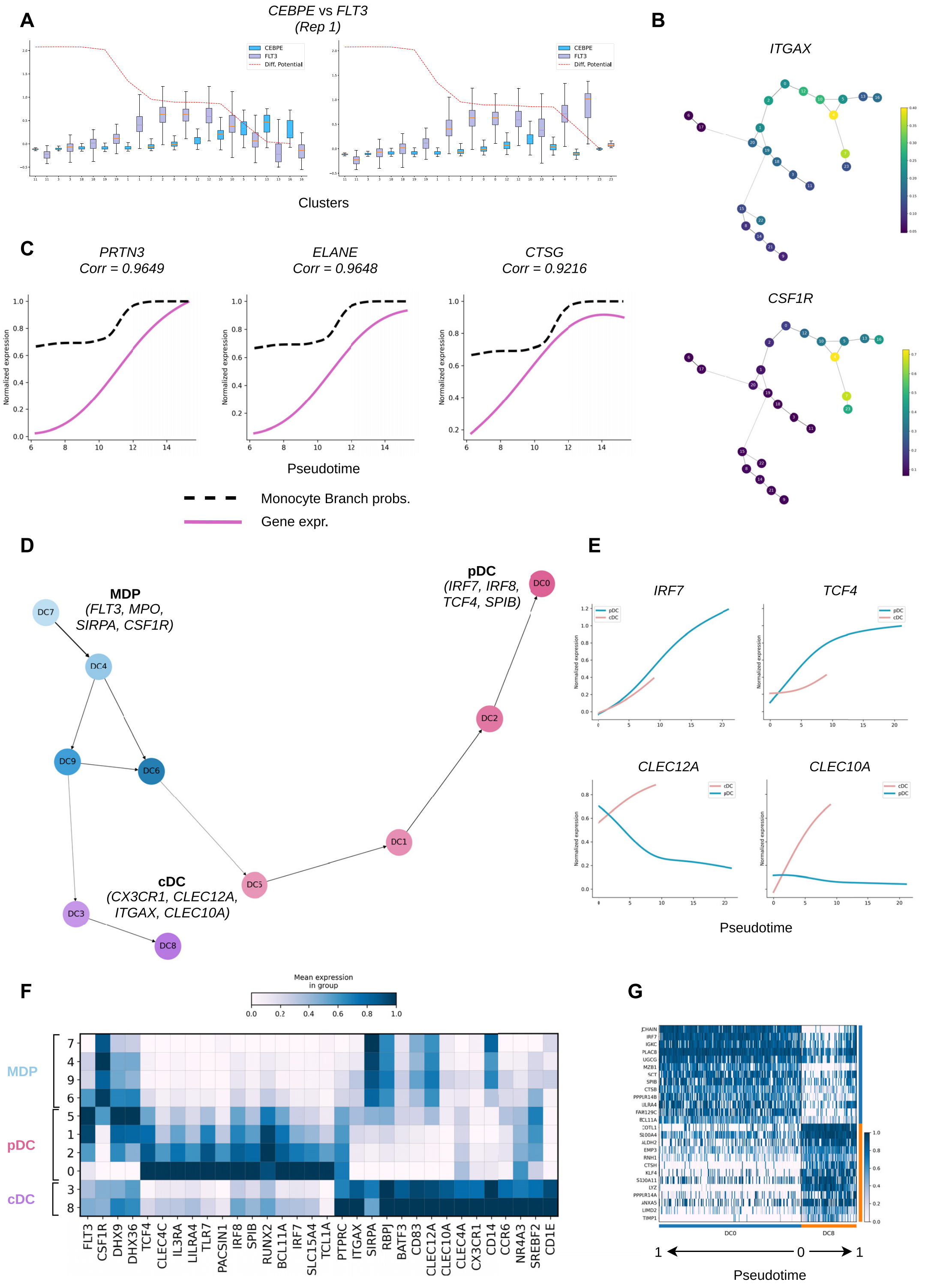

Figure 6. MARGARET characterizes cellular heterogeneity in the monocytic and dendritic cell lineages. (**A**) Variation of the expression of *CEBPE* and *FLT3* across the monocyte (Mono) lineage (left) and the DC lineage (right) for replicate 1. The boxplots summarize the expression of the gene in each cluster in the lineage, where the box depicts the interquartile range (IQR, the range between the 25th and 75th percentile) with the median value, whiskers indicate the maximum and minimum value within 1.5 times the IQR. The red dotted line represents the mean differentiation potential for each cluster in the lineage. (**B**) Mean expression of *ITGAX* (top) and *CSF1R* (bottom), projected on the MARGARET inferred connectivity graph. (**C**) Gene expression trends for neutrophil-like monocyte marker genes: *PRTN3*, *ELANE*, *CTSG* are highly correlated with monocyte branch probability (dotted black line). (**D**) MARGARET inferred trajectory for the DC sub-lineage obtained from the combined analysis of replicates 1 and 2. MARGARET accurately recovers the cDC and pDC lineages. (**E**) Gene expression trends for pDC markers (top): *IRF7*, *TCF4* and cDC markers (bottom): *CLEC12A*, *CLEC10A*. (**F**) Heat map of DC-lineage specific marker genes for MARGARET inferred DC sub-lineage clusters. (**G**) Comparison of differentially Expressed genes between cDCs (cluster DC8) and pDCs (cluster DC0).

bryonic stem cells (ESCs) give rise to early lineage precursors. Even though EB differentiation has been successfully utilized to drive a diverse set of differentiation protocols (67,68), the cellular and molecular states associated with the early lineage precursors as well as their differentiation trajectories from human ESCs remain fairly less explored. The dataset consisted of 16821 cells, sampled at 3-day intervals over a 27-day differentiation time course (18). Even though the sampling time information was not utilized to learn the cell-state manifold, MARGARET inferred embedding that consisted of 26 clusters (Figure 7A) retained the time trend accurately (Figure 7B), thus preserving the global topology associated with the data. Next, we identified the major lineages recovered by MARGARET by examining the expression of essential marker genes for major lineages previously reported in the literature for this dataset (18) (Figure 7D, Supplementary Figure S21A,B). This preliminary analysis revealed the presence of endoderm (EN), mesoderm (ME), neural crest (NC), neuroectoderm (NE) and neuronal subtypes (NS) (including neural progenitors (NP)) lineages along with ESCs in the data. To further validate the inferred lineages, we grouped MARGARET clusters based on their lineage information to obtain five major clusters (Supplementary Figure S22A) for which we performed differential expression (DE) analysis (Supplementary Figure S22B). Gene ontology (GO) analysis (Materials and Methods, Data Availability) of the DE genes for these combined clusters (Supplementary Figure S22C) revealed major functional differences between these clusters, with the enrichment of GO terms corresponding to these major lineages suggesting the validity of the inferred lineages (Supplementary File 1).

Due to the presence of ESCs in cluster 6 as marked by the high expression of *POU5F1*, *NANOG* (essential for maintaining pluripotency in ESCs (69)), and *DDPA2/4* (Figure 7D, F), we selected a starting cell from cluster 6 for further trajectory analysis, including the inference of pseudotime and DP. MARGARET inferred pseudotime (Figure 7C) followed the progression of cell types, where the pseudotime increased as cells progressed towards more specialized cell types from *POU5F1* enriched ESCs.

Furthermore, MARGARET recovered a detailed lineage specification map of embryoid bodies in a fully unsupervised manner (Figure 7E). We further characterized the MARGARET inferred clusters for specific cell types. In the ectoderm lineage: neuroectoderm, neural crest, and neuronal subtype clusters were detected as terminal states. In the mesoderm lineage, MARGARET identified hemangioblasts (H), cardiac precursors (CPs), and smooth muscle precursors (SMPs) as the terminal states. Lastly, a single cluster in the endoderm lineage was also identified as a terminal state. To characterize the recovered lineage map, we explored the expression trends of key marker genes for the terminal cell types (Figure 7F). The expression of ESC marker gene *POU5F1* decreased with pseudotime in all lineages. *CD34* was selectively upregulated in hemangioblasts, while *TNNT2*, and *TBX18* were upregulated in the CPs and SMPs, respectively. *KLF5* and *SOX10* were upregulated in the EN, and NC lineages respectively. *LHX5* was initially upregulated in the NS, NC and NE lineages but was subsequently downregulated in the NS and NC lineages suggesting its importance in NE lineage commitment. While we identified five NS clusters, *ONECUT1* was upregulated in NS-5, the terminal neuronal subtype cluster.

The probability of cells branching towards a specific lineage as inferred by MARGARET increased towards later stages in cell differentiation (Supplementary Figure S23). For this dataset also, DP decreased with an increase in pseudotime (Supplementary Figure S24A) with the ESC cluster having the highest DP, followed by the transitional cell types and the terminal states having the lowest DP (Supplementary Figures S24 and S25). Thus decrease in DP was concordant with the major lineage commitments in EB differentiation.

In the ectodermal lineage, ESCs differentiate into preneuroectoderm cells (showing downregulation of *POU5F1* (Figure 7F)), which give rise to neuroectoderm cells (expressing *LHX2/5, SIX3* (Figure 5D,F)). Similar to (18), MARGARET was able to identify the bipotent precursors (cluster 9 expressing *HOXA2, HOXB1* and *OLIG3* (Figure 7D)) that originated from the neuroectoderm cells expressing *GBX2* and bifurcated from cluster 9 into the neural crest and neuronal sub-lineages (Figure 7E). Further characterization of this branching revealed correlation between a decrease in MARGARET inferred DP and the up/down regulation of important TFs in the neural crest and neuronal lineages (Supplementary Figure S26). The DP drop in the neural crest lineage was concordant with the upregulation of canonical TFs *SOX9/10* (70,71) (Figure 7D,F, Supplementary Figure S26A,B (right)) while neuronal-subtype cluster 3 exhibited upregulation of TFs *SOX1* and *LHX2* (Figure 7D, Supplementary Figure S26A, B (left)), which have been shown to be important for subtype specification in certain types of neurons (72,73). GO analysis at finer resolution (Figure 7G, Data Availabililty) also revealed the enrichment of both neural crest and neuronal differentiation-related functions in cluster 9 further validating its bi-potency. We next validated the neural crest subbranch detected by MARGARET using the bulk RNA-seq data provided by (18) for FACS purified $CD49d^{+}CD63^{-}$ cells. The Spearman correlation analysis between scRNA-seq profiles of cells in the EB dataset with the bulk RNA-seq expressions corresponding to $CD49^{+}$ cells revealed the highest correlation in the neural crest lineage which also showed high *ITGA4* expression (Supplementary Figure S27A), suggesting the accurate localization of the neural crest cells in MARGARET trajectory.

In the mesoderm lineage, differentiation proceeds via the primitive streak progenitor cells (cluster 22 expressing *EOMES* and *T*), towards a number of sub-lineages within the mesoderm. MARGARET trajectory (Figure 7E) identified a series of intermediate precursors expressing (*CER1, GATA1*), (*GATA6, HOXB4*) and (*GATA5/6, HAND1*), respectively which finally gave rise to cardiac precursors (expressing *TNNT2*). Bulk-RNA analysis for FACS purified $CD82^{+}CD142^{+}$ cells revealed the highest correlation of the single-cell expression profiles in the mesoderm sublineage harboring cardiac precursors (cells with high *CD82* and *CD142* expressions) (Supplementary Figure S27B), suggesting that MARGARET accurately detects this sublineage. In the mesoderm lineage, we found three types of

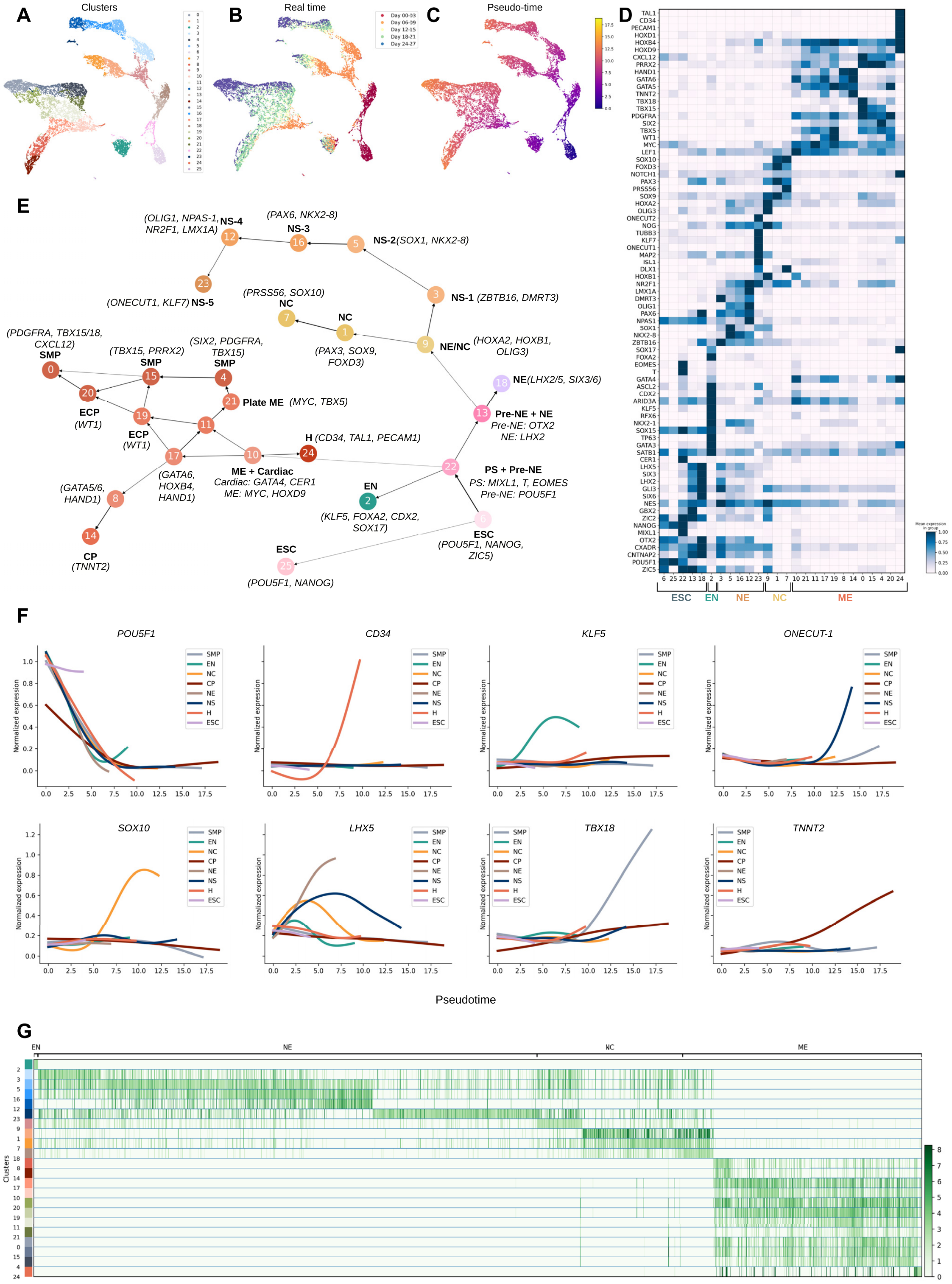

Figure 7. MARGARET characterizes the differentiation trajectory in human embryoid bodies. MARGARET reconstructs a detailed lineage map and identifies novel cell types and differentiation intermediates in the embryoid body (EB) dataset. (**A**) UMAP plot of cell-state embedding inferred by MARGARET for EB dataset. (**B**) MARGARET inferred cell embedding preserves real time information. (**C**) MARGARET pseudotime projected on the cell embeddings. (**D**) Heat map of marker genes for all MARGARET inferred clusters. (**E**) MARGARET inferred trajectory annotated with cell-type and lineage information (characteristic marker genes are mentioned within parantheses with the cell-type annotation). ESC: Embryoid stem cells; PS: primitive streak; NE: neuroectoderm; NC: neural crest; NS: neuronal subtypes; EN: endoderm; H: hemangioblasts; ME: mesoderm; SMP: smooth muscle precursors; ECP: epicardial precursors; CP: cardiac precursors. (**F**) Gene expression trends for the inferred lineages. (**G**) Gene ontology (GO) analysis of MARGARET inferred clusters (grouped by major lineages). The heatmap value for a GO term was set to $\sqrt{-\log(p_{val})}$, where p_{val} is the P-value for the corresponding GO term.

smooth muscle precursors (SMPs) which expressed different markers (*SIX2*, *PRRX2* and *TBX15/18*) suggesting that SMP differentiation in the mesoderm lineage proceeds via one or more differentiation intermediates from plate ME cells. Thus, MARGARET accurately resolved the underlying trajectory structure in the EB dataset by inferring branchings and intermediate cell populations essential for cell lineage commitment in early embryogenesis.

Analysis of colon differentiation using MARGARET

To investigate the epithelial differentiation in colon, we applied MARGARET to a scRNA-seq dataset (11175 cells) comprising three conditions - healthy (4249 cells), clinically inflamed ulcerative colitis (UC) (2848 cells), and non-inflamed UC (4078 cells) across three replicates (19). For the healthy colon, MARGARET identified 15 clusters (Figure 8A) and the inferred trajectory accurately delineated the absorptive and the secretory lineages (Figure 8F). Projection of crypt-axis (CA) scores (19) (see Materials and Methods) for each cell (obtained from the combined expression of 15 gene markers expressed in absorptive and secretory cells) on the 2D representation of the MARGARET-inferred cell embedding (Figure 8D) revealed the cells in the absorptive lineage to have higher CA scores as compared to stem cells or cells in the secretory lineage indicating the presence of these cells towards the crypt-top indicating that MARGARET correctly localized the cell-types within the trajectory. We then investigated the expression of key marker genes (curated from prior literature (19,74)) for different cell types in the absorptive and secretory lineages (Figure 8E). We identified absorptive progenitors, colonocytes, crypt-top (CT) colonocytes, and BEST4/OTOP2 cells in the absorptive lineage; and secretory progenitors, goblet cells, and enteroendocrine cells (EECs) in the secretory lineage. Based on high expression of stem cell markers genes *MLEC* and *LGR5* in cluster 0 (Figure 8E), we selected a cell from this cluster as the starting cell for the subsequent inference of pseudotime (Figure 8B) and DP (Figure 8C). Terminal state prediction using MARGARET further revealed four terminal states namely: EECs (cluster 14) and goblet cells (cluster 5) in the secretory cell lineage, and BEST4/OTOP2 cells (cluster 13) and CT colonocytes (cluster 3) in the absorptive cell lineage. MARGARET inferred branch probabilities (Supplementary Figure S28) and gene expression trends (Figure 8H) in the detected lineages further validated the cell populations detected in the absorptive and the secretory cell lineages as the marker genes *MUC2*, *SCGN*, *AQP8* and *BEST4*, were selectively upregulated in the goblet (75), EECs (76), CT colonocytes (77) and BEST4/OTOP2 cell (19) lineages respectively.

In the absorptive cell lineage, BEST4/OTOP2 cells were detected as a terminal state that branched from the absorptive progenitors (APs) before they gave rise to colonocytes. To characterize this novel branching point, we performed GO analysis of BEST4/OTOP2 cells, CT colonocytes, colonocytes, and APs, which revealed the role of BEST4/OTOP2 cells in maintaining metal-ion transport and homeostasis in the colonic epithelium (Figure 8G). The GO analysis of the BEST4/OTOP2 cell cluster did not include any overlapping GO terms with CT colonocytes and colonocytes, suggesting the functional variability between these cell types. However, BEST4/OTOP2 cells exhibited high CA scores (Figure 8D) and expressed several colonocyte-specific markers *GUCA2A*, *CEACAM1* and *CEACAM7*, which suggests that these cells are similar to mature colonocytes and lie towards the crypt top. Moreover, APs showed enrichment of GO terms corresponding to both BEST4 cells and colonocytes (Figure 8G). The branching of BEST4/OTOP2 cells was also marked by the downregulation of the TF *ESRRA* and the high expression of TFs *CDX1* and *PPARG* (Figure 8I). We observed a similar branching in the absorptive lineage under non-inflamed UC (Supplementary Figure S29) with similar GO enrichment for the BEST4/OTOP2 cells, CT colonocytes, and colonocytes. Therefore, our findings suggest that BEST4/OTOP2 cells are mature colonocytes with a different functional profile as compared to CT colonocytes and originate from APs as a different sub-lineage within the absorptive lineage.

Given the crucial role of goblet cells in colonic barrier maintenance (78), we further characterized the transcriptional landscape of the goblet cell lineages under healthy and inflamed UC conditions using MARGARET (Supplementary Figure S30). Under both conditions, MARGARET reconstructed a linear trajectory in the goblet cell lineage from immature goblet cells to mature goblet cells (Supplementary Figure S30G, H) (maturity of cells inferred by pseudotime order (Supplementary Figure S30C, F)). We observed relatively higher expression levels of *WFDC2* in immature goblet cells than in mature goblet cells under both healthy and inflamed conditions. In contrast, *MUC2* was more expressed in mature goblet cells (Supplementary Figure S30I). Moreover, we found mature goblet cells to have higher CA scores than immature goblet cells, suggesting that the mature cells reside at the top of the colonic crypt (Supplementary Figure S30B, E). The functional characterization (Supplementary Figure S30J) of the goblet cell clusters revealed the enrichment of GO terms related to wound healing, immune, and stress response in the mature goblet cells under inflamed UC condition indicating their potential role in inflammatory responses to inflammatory bowel disease (IBD). The goblet cells in inflamed UC further showed spatial and crypt-wide transcriptional heterogeneity. The mature goblets in inflamed UC that reside in the crypt-top showed elevated expressions of *SPINK1* and *SPINK4* (Supplementary Figure S31A), genes that are normally expressed by immature goblets residing at the crypt bottom in healthy colon. Moreover, we also observed transcriptional dysregulation of interferon-regulated cytokines including *CD164*, *CD55* and *IRF7* (79) throughout the goblet cell lineage in inflamed UC (Supplementary Figure S31B).

MARGARET can scale to large scRNA-seq datasets

To assess MARGARET's scalability to large scRNA-seq datasets, we measured the runtimes of different computational stages in MARGARET on scRNA-seq datasets of different sizes subsampled from a 1.3 million neuronal cells dataset of 10× Genomics (20) (Supplementary Figure S32A). For a large scRNA-seq dataset consisting of 500 000 cells, training MARGARET's neural network-

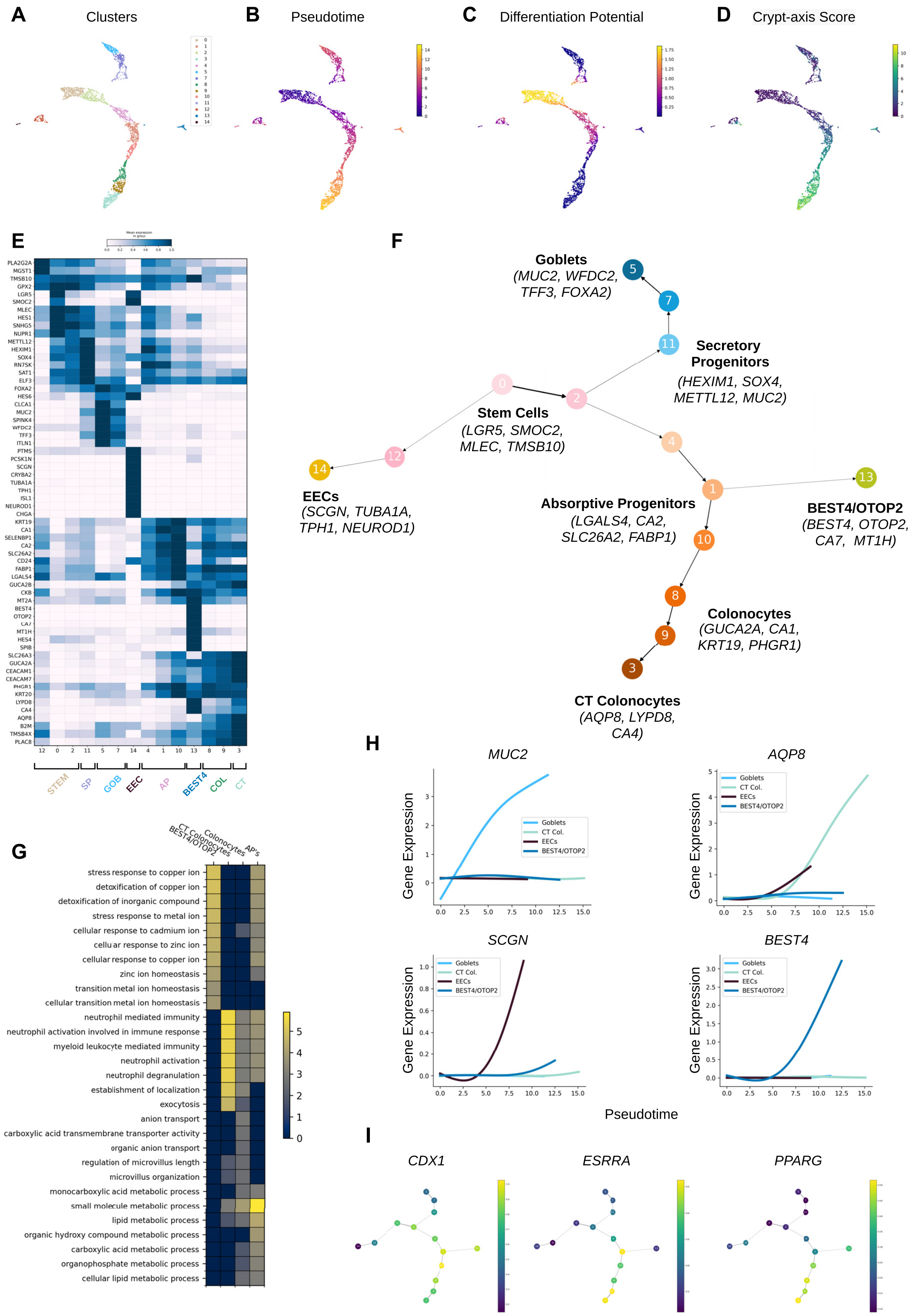

Figure 8. MARGARET delineates the differentiation trajectory for colonic epithelial cells for healthy humans. (**A**) UMAP plot of cell-state embedding inferred by MARGARET for healthy human colonic epithelium, cells are colored by clusters inferred by MARGARET. (**B**) MARGARET pseudotime and (**C**) differentiation potential for each cell projected on the cell embeddings. (**D**) Crypt-axis score (See Methods) projected on MARGARET cell embeddings. (**E**) Heat map showing marker genes for each MARGARET inferred cluster. Marker genes were curated from (19). (**F**) MARGARET inferred trajectory annotated with cell type and lineage information (important marker genes for each cell type are mentioned within parantheses with the cell type information). CT: Crypt-top; EEC: enteroendocrine cells (**G**) Gene Ontology (GO) analysis for major inferred cell types in the absorptive cell lineage. Top GO:BP terms were included for each cell type. The value for a GO term in the heat map was set to $\sqrt{-\log(p_{val})}$, where p_{val} is the *P*-value for the corresponding GO term. (**H**) Gene expression trends for essential genes for major inferred lineages. **(I)** Mean expression of TFs *CDX1*, *ESRRA* and *PPARG*, projected on the MARGARET inferred connectivity graph.

based encoder in the metric learning stage took around 4 min per epoch (on a server with one Nvidia Quadro RTX 5000 GPU). Moreover, due to extensive hardware support in modern computation libraries, MARGARET can also utilize multiple GPUs during the training stage, thus making it scalable for inferring cell-state manifold even for very large scRNA-seq datasets. For the same dataset, undirected graph construction in MARGARET took only ∼20 s making exploratory analysis and visualization of large datasets extremely fast. For pseudotime inference, MARGARET took around 2 h for the 500k cells dataset thus exhibiting scalability across all computational stages when run on a large scRNA-seq dataset. We also compared MARGARET's runtime for trajectory inference (undirected graph construction and pseudotime inference) with that of other TI methods on two relatively large subsampled datasets consisting of 50k and 100k cells and found that MARGARET's runtimes were comparable to that of other methods (Supplementary Figure S32B (left)). Moreover, on two smaller datasets sampled from our real dataset benchmark, MARGARET outperformed all other methods in terms of runtime for trajectory inference (Supplementary Figure S32B (right)).

DISCUSSION

As single-cell datasets grow in size and complexity, MARGARET addresses the need for scalable and accurate detection of cell-state lineages, prediction of cell fates, and the inference of cell fate plasticity in complex topologies underlying dynamic cellular processes. The end-to-end computational framework of MARGARET alleviates the challenges faced by existing TI methods: inability of the classical dimensionality reduction methods to accurately recover the underlying topology, insufficient generalizability to connected and disconnected graph trajectories beyond tree-structured topologies, accurate detection of less-sampled cell fates, and inference of cell fate plasticity for complex trajectory types. Our analysis of a diverse simulated benchmark as well as real benchmark consisting of challenging topologies showed that MARGARET generalizes to complex trajectories and can accurately infer the underlying topology and pseudotime order of cells in the trajectory while outperforming state-of-the-art methods on the same. Specifically, the benchmarking with the real biological datasets showed MARGARET's ability to recapitulate the cell pseudotime order for complex trajectories where other methods completely failed (e.g. placenta trophoblast differentiation). Moreover, using synthetic disconnected trajectories, we showed that MARGARET can accurately infer DP for each component in the disconnected trajectories whereas Palantir's DP inference is mostly incorrect for such datasets. (Figure 4).

Using multiple biological datasets, we also showed that MARGARET's metric learning-based approach can *refine* the cellular latent space inferred by other dimensionality reduction methods. On a variety of important biological systems, we showed that MARGARET identified all the major lineages and the established cell fates and recovered the marker gene expression trends. In early human hematopoiesis, MARGARET identified progenitor populations associated with the branching points (MDP and GMP in the myeloid and MEP in the erythroid lineages respectively), which were not characterized in the original study that used Palantir. MARGARET also identified cDC2 as a terminal state in the dendritic cell lineage along with pDCs. For embryoid body differentiation, MARGARET reconstructed a detailed lineage map of all major and sub-lineages and further validated the presence of novel differentiation intermediates in the neuroectoderm and mesoderm lineages and identified novel smooth muscle precursor populations. For colonic epithelial differentiation, MARGARET helped identify a branch point for BEST4/OTOP2 cells in the absorptive lineage, while also uncovering the dysregulation of essential marker genes and cytokines in goblet cells under the inflamed UC condition. Our runtime experiment with a 1.3 million cells dataset (20) further demonstrated that MARGARET's trajectory inference scales even to datasets with millions of cells.

Similar to other pseudotime methods, MARGARET assumes unidirectional cell differentiation where immature stem cells differentiate into more mature cell types. This assumption is violated for trans-differentiation and de-differentiation, events that can lead to ancestral cell-states and scRNA-seq data alone might be insufficient in characterizing such events (9). Recent methods (80–82) utilized naturally occurring somatic mutations or synthetic mutations for lineage tracing. It would be an interesting direction to extend MARGARET by incorporating auxiliary signals like lineage information and real-time information for elucidating reprogramming. Lastly, while we focus on scRNA-seq datasets in this study, it is worth noting that our framework can easily be extended for other single-cell omics as well as multi-omics datasets. The modular structure of our method also allows for effortless integration with other dimension reduction, clustering and omics-integration methods. The implementation of MARGARET also allows the users to provide their own clustering as an input to MARGARET and thus supports the inclusion of domain knowledge in subsequent stages of trajectory inference. Given the explosion of single-cell datasets fuelled by collaborative efforts such as Human Cell Atlas (HCA) project (83) and Human Biomolecular Atlas Program (HubMAP) (84), we anticipate MARGARET to be a valuable tool for a scalable and multifaceted exploration of dynamic cellular processes from varied biological systems.

DATA AVAILABILITY

All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. The human hematpoiesis dataset is available through the Human Cell Atlas portal at https://prod.data.humancellatlas.org/explore/projects/29f53b7e-071b-44b5-998a-0ae70d0229a4. The scRNA-seq and bulk RNA-seq datasets for the embryoid body dataset can be accessed via the Mendeley Data repository at https://doi.org/10.17632/v6n743h5ng.1. The scRNA-seq data for colon differentiation can be accessed using the GEO accession number GSE116222. All the real datasets used for demonstrating the clustering efficiency of MARGARET (PBMC-8k, PBMC-4k, Heart Cell Atlas and CORTEX) and runtime benchmarking (1.3M neuron dataset) can be accessed via the *scvi-tools* package (22). The Gene-Ontology terms used for the EB dataset can be accessed via Zenodo at https://doi.org/10.5281/zenodo.5751235. All simulated and real datasets (Placenta Trophoblast differentiation, Mouse cell atlas, Oligodendrocyte differentiation and Planaria parenchyme differentiation) used for comparing MARGARET to other TI methods can be accessed via Zenodo at https://doi.org/10.5281/zenodo.5850114.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank the funding agencies for the research grants supporting this study.

FUNDING

Science and Engineering Research Board (SERB), Government of India [SRG/2020/001333] to H.Z. Funding for open access charge: IIT Kanpur Research I Foundation grant 20030163.
Conflict of interest statement. None declared.

REFERENCES

1. Bendall,S.C., Davis,K.L., Amir,E.-A.D., Tadmor,M.D., Simonds,E.F., Chen,T.J., Shenfeld,D.K., Nolan,G.P. and Pe'er,D. (2014) Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. *Cell*, **157**, 714–725.
2. Tanay,A. and Regev,A. (2017) Scaling single-cell genomics from phenomenology to mechanism. *Nature*, **541**, 331–338.
3. Poulin,J.-F., Tasic,B., Hjerling-Leffler,J., Trimarchi,J.M. and Awatramani,R. (2016) Disentangling neural cell diversity using single-cell transcriptomics. *Nat. Neurosci.*, **19**, 1131.
4. Halpern,K.B., Shenhav,R., Matcovitch-Natan,O., Tóth,B., Lemze,D., Golan,M., Massasa,E.E., Baydatch,S., Landen,S., Moor,A.E. *et al.* (2017) Single-cell spatial reconstruction reveals global division of labour in the mammalian liver. *Nature*, **542**, 352.
5. Etzrodt,M., Endele,M. and Schroeder,T. (2014) Quantitative single-cell approaches to stem cell research. *Cell Stem Cell*, **15**, 546–558.
6. Wagner,D.E., Weinreb,C., Collins,Z.M., Briggs,J.A., Megason,S.G. and Klein,A.M. (2018) Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science*, **360**, 981–987.
7. Cao,J., Spielmann,M., Qiu,X., Huang,X., Ibrahim,D.M., Hill,A.J., Zhang,F., Mundlos,S., Christiansen,L., Steemers,F.J. *et al.* (2019) The single-cell transcriptional landscape of mammalian organogenesis. *Nature*, **566**, 496–502.
8. Wolf,F.A., Angerer,P. and Theis,F.J. (2018) SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.*, **19**, 15.
9. Setty,M., Kiseliovas,V., Levine,J., Gayoso,A., Mazutis,L. and Pe'er,D. (2019) Characterization of cell fate probabilities in single-cell data with Palantir. *Nat. Biotechnol.*, **37**, 451–460.
10. Street,K., Risso,D., Fletcher,R.B., Das,D., Ngai,J., Yosef,N., Purdom,E. and Dudoit,S. (2018) Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics*, **19**, 477.
11. Chen,H., Albergante,L., Hsu,J.Y., Lareau,C.A., Bosco,G.L., Guan,J., Zhou,S., Gorban,A.N., Bauer,D.E., Aryee,M.J. *et al.* (2019) Single-cell trajectories reconstruction, exploration and mapping of omics data with STREAM. *Nat. Commun.*, **10**, 1903.
12. McInnes,L., Healy,J., Saul,N. and Großberger,L. (2018) UMAP: Uniform Manifold Approximation and Projection, *Journal of Open Source Software*, **3**, 861.
13. Saelens,W., Cannoodt,R., Todorov,H. and Saeys,Y. (2019) A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.*, **37**, 547–554.
14. Stassen,S.V., Yip,G. G.K., Wong,K. K.Y., Ho,J. W.K. and Tsia,K.K. (2021) Generalized and scalable trajectory inference in single-cell omics data with VIA. *Nat. Commun.*, **12**, 5528.
15. Han,X., Wang,R., Zhou,Y., Fei,L., Sun,H., Lai,S., Saadatpour,A., Zhou,Z., Chen,H., Ye,F. *et al.* (2018) Mapping the Mouse Cell Atlas by Microwell-Seq. *Cell*, **172**, 1091–1107.
16. Marques,S., Zeisel,A., Codeluppi,S., van Bruggen,D., Falcão,A.M., Xiao,L., Li,H., Häring,M., Hochgerner,H., Romanov,R.A. and et,al. (2016) Oligodendrocyte heterogeneity in the mouse juvenile and adult central nervous system. *Science*, **352**, 1326–1329.
17. Plass,M., Solana,J., Wolf,F.A., Ayoub,S., Misios,A., Glažar,P., Obermayer,B., Theis,F.J., Kocks,C. and Rajewsky,N. (2018) Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics. *Science*, **360**, 6391.
18. Moon,K.R., van Dijk,D., Wang,Z., Gigante,S., Burkhardt,D.B., Chen,W.S., Yim,K., Elzen,A.v.d., Hirn,M.J., Coifman,R.R. *et al.* (2019) Visualizing structure and transitions in high-dimensional biological data. *Nat. Biotechnol.*, **37**, 1482–1492.
19. Parikh,K., Antanaviciute,A., Fawkner-Corbett,D., Jagielowicz,M., Aulicino,A., Lagerholm,C., Davis,S., Kinchen,J., Chen,H.H., Alham,N.K. *et al.* (2019) Colonic epithelial cell diversity in health and inflammatory bowel disease. *Nature*, **567**, 49–55.
20. Zheng,G. X.Y., Terry,J.M., Belgrader,P., Ryvkin,P., Bent,Z.W., Wilson,R., Ziraldo,S.B., Wheeler,T.D., McDermott,G.P., Zhu,J. *et al.* (2017) Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.*, **8**, 14049
21. van Dijk,D., Sharma,R., Nainys,J., Yim,K., Kathail,P., Carr,A.J., Burdziak,C., Moon,K.R., Chaffer,C.L., Pattabiraman,D. *et al.* (2018) Recovering gene interactions from single-cell data using data diffusion. *Cell*, **174**, 716–729.
22. Gayoso,A., Lopez,R., Xing,G., Boyeau,P., Amiri,V.V.P., Hong,J., Wu,K., Jayasuriya,M., Mehlman,E., Langevin,M. *et al.* (2022) A Python library for probabilistic analysis of single-cell omics data. *Nat. Biotechnol.*, **40**, 163–166.
23. Traag,V.A., Waltman,L. and van Eck,N.J. (2019) From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Rep.*, **9**, 5233.
24. Satija,R., Farrell,J.A., Gennert,D., Schier,A.F. and Regev,A. (2015) Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.*, **33**, 495–502.
25. Balntas,V., Riba,E., Ponsa,D. and Mikolajczyk,K. (2016) Learning local feature descriptors with triplets and shallow convolutional neural networks. In: *Procedings of the British Machine Vision Conference 2016*. British Machine Vision Association.
26. Ioffe,S. and Szegedy,C. (2015) Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In: Bach,F. and Blei,D. (eds). *Proceedings of the 32nd International Conference on Machine Learning*. PMLR Vol.37 of Proceedings of Machine Learning Research, Lille, France, pp. 448–456.
27. Srivastava,N., Hinton,G., Krizhevsky,A., Sutskever,I. and Salakhutdinov,R. (2014) Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, **15**, 1929–1958.
28. Wolf,F.A., Hamey,F.K., Plass,M., Solana,J., Dahlin,J.S., Göttgens,B., Rajewsky,N., Simon,L. and Theis,F.J. (2019) PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biol.*, **20**, 59.
29. Tenenbaum,J.B., Silva,V.d. and Langford,J.C. (2000) A global geometric framework for nonlinear dimensionality reduction. *Science*, **290**, 2319–2323.
30. Brandes,U. (2001) A faster algorithm for betweenness centrality. *J. Math. Soc.*, **25**, 163–177.
31. Setty,M., Tadmor,M.D., Reich-Zeliger,S., Angel,O., Salame,T.M., Kathail,P., Choi,K., Bendall,S., Friedman,N. and Pe'er,D. (2016) Wishbone identifies bifurcating developmental trajectories from single-cell data. *Nat. Biotechnol.*, **34**, 637–645.
32. Bendall,S.C., Davis,K.L., ad David Amir,E., Tadmor,M.D., Simonds,E.F., Chen,T.J., Shenfeld,D.K., Nolan,G.P. and Pe'er,D. (2014) Single-cell trajectory detection uncovers progression and

regulatory coordination in human B cell development. *Cell*, **157**, 714–725.
33. Silva,V. and Tenenbaum,J. (2004) Sparse multidimensional scaling using landmark points. Technical report, Stanford University.
34. Liu,W. and Lü,L. (2010) Link prediction based on local random walk. *EPL (Europhys. Lett.)*, **89**, 58007.
35. Raudvere,U., Kolberg,L., Kuzmin,I., Arak,T., Adler,P., Peterson,H. and Vilo,J. (2019) g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res.*, **47**, W191–W198.
36. Ipsen,M. and Mikhailov,A.S. (2002) Evolutionary reconstruction of networks. *Phys. Rev. E*, **66**, 046109.
37. van der Maaten,L. and Hinton,G. (2008) Visualizing Data using t-SNE. *J. Mach. Learn. Res.*, **9**, 2579–2605.
38. Stassen,S.V., Siu,D.M.D., Lee,K.C.M., Ho,J.W.K., So,H.K.H. and Tsia,K.K. (2020) PARC: ultrafast and accurate clustering of phenotypic data of millions of single cells. *Bioinformatics*, **36**, 2778–2786.
39. Levine,J.H., Simonds,E.F., Bendall,S.C., Davis,K.L., Amir,E.D., Tadmor,M.D., Litvin,O., Fienberg,H.G., Jager,A., Zunder,E.R. *et al.* (2015) Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. *Cell*, **162**, 184–197.
40. Lopez,R., Regier,J., Cole,M.B., Jordan,M.I. and Yosef,N. (2018) Deep generative modeling for single-cell transcriptomics. *Nat. Methods*, **15**, 1053–1058.
41. Orkin,S.H. and Zon,L.I. (2008) Hematopoiesis: an evolving paradigm for stem cell biology. *Cell*, **132**, 631–644.
42. Paul,F., Arkin,Y., Giladi,A., Jaitin,D.A., Kenigsberg,E., Keren-Shaul,H., Winter,D., Lara-Astiaso,D., Gury,M., Weiner,A. *et al.* (2015) Transcriptional heterogeneity and lineage commitment in myeloid progenitors. *Cell*, **163**, 1663–1677.
43. Pellin,D., Loperfido,M., Baricordi,C., Wolock,S.L., Montepeloso,A., Weinberg,O.K., Biffi,A., Klein,A.M. and Biasco,L. (2019) A comprehensive single cell transcriptional landscape of human hematopoietic progenitors. *Nat. Commun.*, **10**, 2395.
44. Lu,Y.-C., Sanada,C., Xavier-Ferrucio,J., Wang,L., Zhang,P.-X., Grimes,H.L., Venkatasubramanian,M., Chetal,K., Aronow,B., Salomonis,N. *et al.* (2018) The molecular signature of megakaryocyte-erythroid progenitors reveals a role for the cell cycle in fate specification. *Cell Rep.*, **25**, 2083–2093.
45. Benschop,R.J. and Cambier,J.C. (1999) B cell development: signal transduction by antigen receptors and their surrogates. *Curr. Opin. Immunol.*, **11**, 143–151.
46. Yang,J., Zhang,L., Yu,C., Yang,X.-F. and Wang,H. (2014) Monocyte and macrophage differentiation: circulation inflammatory monocyte as biomarker for inflammatory diseases. *Biomarker Res.*, **2**, 1.
47. Lee,J., Zhou,Y.J., Ma,W., Zhang,W., Aljoufi,A., Luh,T., Lucero,K., Liang,D., Thomsen,M., Bhagat,G. *et al.* (2017) Lineage specification of human dendritic cells is marked by IRF8 expression in hematopoietic stem cells and multipotent progenitors. *Nat. Immunol.*, **18**, 877–888.
48. Psaila,B., Barkas,N., Iskander,D., Roy,A., Anderson,S., Ashley,N., Caputo,V.S., Lichtenberg,J., Loaiza,S., Bodine,D.M. *et al.* (2016) Single-cell profiling of human megakaryocyte-erythroid progenitors identifies distinct megakaryocyte and erythroid differentiation pathways. *Genome Biol.*, **17**, 83.
49. Ferreira,R., Ohneda,K., Yamamoto,M. and Philipsen,S. (2005) GATA1 Function, a Paradigm for Transcription Factors in Hematopoiesis. *Mol. Cell. Biol.*, **25**, 1215–1227.
50. Porse,B.T., Pedersen,T.Å., Xu,X., Lindberg,B., Wewer,U.M., Friis-Hansen,L. and Nerlov,C. (2001) E2F repression by C/EBPα is required for adipogenesis and granulopoiesis in vivo. *Cell*, **107**, 247–258.
51. Akashi,K., Traver,D., Miyamoto,T. and Weissman,I.L. (2000) A clonogenic common myeloid progenitor that gives rise to all myeloid lineages. *Nature*, **404**, 193–197.
52. Yáñez,A., Coetzee,S.G., Olsson,A., Muench,D.E., Berman,B.P., Hazelett,D.J., Salomonis,N., Grimes,H.L. and Goodridge,H.S. (2017) Granulocyte-monocyte progenitors and monocyte-dendritic cell progenitors independently produce functionally distinct monocytes. *Immunity*, **47**, 890–902.
53. Reizis,B., Bunin,A., Ghosh,H.S., Lewis,K.L. and Sisirak,V. (2011) Plasmacytoid dendritic cells: recent progress and open questions. *Ann. Rev. Immunol.*, **29**, 163–183.
54. Brown,C.C., Gudjonson,H., Pritykin,Y., Deep,D., Lavallée,V.-P., Mendoza,A., Fromme,R., Mazutis,L., Ariyan,C., Leslie,C. *et al.* (2019) Transcriptional basis of mouse and human dendritic cell heterogeneity. *Cell*, **179**, 846–863.
55. Merad,M., Sathe,P., Helft,J., Miller,J. and Mortha,A. (2013) The dendritic cell lineage: ontogeny and function of dendritic cells and their subsets in the steady state and the inflamed setting. *Ann. Rev. Immunol.*, **31**, 563–604.
56. Cao,W., Bover,L., Cho,M., Wen,X., Hanabuchi,S., Bao,M., Rosen,D.B., Wang,Y.-H., Shaw,J.L., Du,Q. *et al.* (2009) Regulation of TLR7/9 responses in plasmacytoid dendritic cells by BST2 and ILT7 receptor interaction. *J. Exp. Med.*, **206**, 1603–1614.
57. Robbins,S.H., Walzer,T., Dembélé,D., Thibault,C., Defays,A., Bessou,G., Xu,H., Vivier,E., Sellars,M., Pierre,P. *et al.* (2008) Novel insights into the relationships between dendritic cell subsets in human and mouse revealed by genome-wide expression profiling. *Genome Biol.*, **9**, R17.
58. Cisse,B., Caton,M.L., Lehner,M., Maeda,T., Scheu,S., Locksley,R., Holmberg,D., Zweier,C., den Hollander,N.S., Kant,S.G. *et al.* (2008) Transcription factor E2-2 is an essential and specific regulator of plasmacytoid dendritic cell development. *Cell*, **135**, 37–48.
59. Crozat,K., Guiton,R., Guilliams,M., Henri,S., Baranek,T., Schwartz-Cornil,I., Malissen,B. and Dalod,M. (2010) Comparative genomics as a tool to reveal functional equivalences between human and mouse dendritic cell subsets. *Immunol. Rev.*, **234**, 177–198.
60. Collin,M. and Bigley,V. (2018) Human dendritic cell subsets: an update. *Immunology*, **154**, 3–20.
61. Huysamen,C., Willment,J.A., Dennehy,K.M. and Brown,G.D. (2008) CLEC9A is a novel activation C-type lectin-like receptor expressed on BDCA3+ dendritic cells and a subset of monocytes. *J. Biol. Chem.*, **283**, 16693–16701.
62. Siatecka,M. and Bieker,J.J. (2011) The multifunctional role of EKLF/KLF1 during erythropoiesis. *Blood*, **118**, 2044–2054.
63. Bianchi,E., Bulgarelli,J., Ruberti,S., Rontauroli,S., Sacchi,G., Norfo,R., Pennucci,V., Zini,R., Salati,S., Prudente,Z. *et al.* (2015) MYB controls erythroid versus megakaryocyte lineage fate decision through the miR-486-3p-mediated downregulation of MAF. *Cell Death Differ.*, **22**, 1906–1921.
64. Zhu,F., Feng,M., Sinha,R., Seita,J., Mori,Y. and Weissman,I.L. (2018) Screening for genes that regulate the differentiation of human megakaryocytic lineage cells. *Proc. Nat. Acad. Sci.*, **115**, E9308–E9316.
65. Azcoitia,V., Aracil,M., Martínez-A,C. and Torres,M. (2005) The homeodomain protein Meis1 is essential for definitive hematopoiesis and vascular patterning in the mouse embryo. *Dev. Biol.*, **280**, 307–320.
66. Tusi,B.K., Wolock,S.L., Weinreb,C., Hwang,Y., Hidalgo,D., Zilionis,R., Waisman,A., Huh,J.R., Klein,A.M. and Socolovsky,M. (2018) Population snapshots predict early haematopoietic and erythroid hierarchies. *Nature*, **555**, 54–60.
67. Nakano,T., Kodama,H. and Honjo,T. (1996) In vitro development of primitive and definitive erythrocytes from different precursors. *Science*, **272**, 722–724.
68. Rohwedel,J., Maltsev,V., Bober,E., Arnold,H.-H., Hescheler,J. and Wobus,A. (1994) Muscle cell differentiation of embryonic stem cells reflects myogenesis in vivo: developmentally regulated expression of myogenic determination genes and functional expression of ionic currents. *Dev. Biol.*, **164**, 87–101.
69. Pan,G. and Thomson,J.A. (2007) Nanog and transcriptional networks in embryonic stem cell pluripotency. *Cell Res.*, **17**, 42–49.
70. Carney,T.J., Dutton,K.A., Greenhill,E., Delfino-Machin,M., Dufourcq,P., Blader,P. and Kelsh,R.N. (2006) A direct role for Sox10 in specification of neural crest-derived sensory neurons. *Development*, **133**, 4619–4630.
71. Cheung,M. and Briscoe,J. (2003) Neural crest development is regulated by the transcription factor Sox9. *Development*, **130**, 5681–5693.
72. Muralidharan,B., Khatri,Z., Maheshwari,U., Gupta,R., Roy,B., Pradhan,S.J., Karmodiya,K., Padmanabhan,H., Shetty,A.S., Balaji,C. *et al.* (2017) LHX2 interacts with the NuRD complex and regulates cortical neuron subtype determinants Fezf2 and Sox11. *J. Neurosci.*, **37**, 194–203.
73. Panayi,H., Panayiotou,E., Orford,M., Genethliou,N., Mean,R., Lapathitis,G., Li,S., Xiang,M., Kessaris,N., Richardson,W.D. *et al.*

(2010) Sox1 is required for the specification of a novel p2-derived interneuron subtype in the mouse ventral spinal cord. *J. Neurosci.*, **30**, 12274–12280.

74. Fawkner-Corbett,D., Antanaviciute,A., Parikh,K., Jagielowicz,M., Gerós,A.S., Gupta,T., Ashley,N., Khamis,D., Fowler,D., Morrissey,E. *et al.* (2021) Spatiotemporal analysis of human intestinal development at single-cell resolution. *Cell*, **184**, 810–826.
75. Johansson,M.E.V., Larsson,J.M.H. and Hansson,G.C. (2011) The two mucus layers of colon are organized by the MUC2 mucin, whereas the outer layer is a legislator of host–microbial interactions. *Proc. Nat. Acad. Sci. U.S.A.*, **108**, 4659–4665.
76. Wagner,L., Oliyarnyk,O., Gartner,W., Nowotny,P., Groeger,M., Kaserer,K., Waldhäusl,W. and Pasternack,M.S. (2000) Cloning and expression of secretagogin, a novel neuroendocrine- and pancreatic islet of Langerhans-specific Ca2+-binding protein. *J. Biol. Chem.*, **275**, 24740–24751.
77. Laforenza,U., Cova,E., Gastaldi,G., Tritto,S., Grazioli,M., LaRusso,N.F., Splinter,P.L., D'Adamo,P., Tosco,M. and Ventura,U. (2005) Aquaporin-8 is involved in water transport in isolated superficial colonocytes from rat proximal colon. *J. Nutr.*, **135**, 2329–2336.
78. Birchenough,G. M.H., Johansson,M.E., Gustafsson,J.K., Bergström,J.H. and Hansson,G.C. (2015) New developments in goblet cell mucus secretion and function. *Mucosal Immunol.*, **8**, 712–719.
79. Andreou,N.P., Legaki,E. and Gazouli,M. (2020) Inflammatory bowel disease pathobiology: the role of the interferon signature. *Ann. Gastroenterol.*, **33**, 125–133.
80. Zafar,H., Tzen,A., Navin,N., Chen,K. and Nakhleh,L. (2017) SiFit: inferring tumor trees from single-cell sequencing data under finite-sites models. *Genome Biol.*, **18**, 178.
81. Zafar,H., Navin,N., Chen,K. and Nakhleh,L. (2019) SiCloneFit: Bayesian inference of population structure, genotype, and phylogeny of tumor clones from single-cell genome sequencing data. *Genome Res.*, **29**, 1847–1859.
82. Zafar,H., Lin,C. and Bar-Joseph,Z. (2020) Single-cell lineage tracing by integrating CRISPR-Cas9 mutations with transcriptomic data. *Nat. Commun.*, **11**, 3055.
83. Regev,A., Teichmann,S.A., Lander,E.S., Amit,I., Benoist,C., Birney,E., Bodenmiller,B., Campbell,P., Carninci,P., Clatworthy,M. *et al.* (2017) The Human Cell Atlas. *eLife*, **6**, e27041.
84. Snyder,M.P., Lin,S., Posgai,A., Atkinson,M., Regev,A., Rood,J., Rozenblatt-Rosen,O., Gaffney,L., Hupalowska,A., Satija,R. *et al.* (2019) The human body at cellular resolution: the NIH Human Biomolecular Atlas Program. *Nature*, **574**, 187–192.