Published online 14 June 2021

Nucleic Acids Research, 2021, Vol. 49, No. 16 ***e91***
https://doi.org/10.1093/nar/gkab481

Cell type hierarchy reconstruction via reconciliation of multi-resolution cluster tree

Minshi Peng[1], Brie Wamsley[2], Andrew G. Elkins[2], Daniel H. Geschwind[2,3], Yuting Wei[1] and Kathryn Roeder[1,4,*]

[1]Department of Statistics and Data Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA, [2]Program in Neurogenetics, Department of Neurology, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA, USA, [3]Program in Neurobehavioral Genetics and Center for Autism Research and Treatment Semel Institute and Department of Human Genetics, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA, USA and [4]Computational Biology Department, Carnegie Mellon University, Pittsburgh, PA 15213, USA

Received April 14, 2021; Revised May 12, 2021; Editorial Decision May 15, 2021; Accepted May 18, 2021

ABSTRACT

A wealth of clustering algorithms are available for single-cell RNA sequencing (scRNA-seq) data to enable the identification of functionally distinct subpopulations that each possess a different pattern of gene expression activity. Implementation of these methods requires a choice of resolution parameter to determine the number of clusters, and critical judgment from the researchers is required to determine the desired resolution. This supervised process takes significant time and effort. Moreover, it can be difficult to compare and characterize the evolution of cell clusters from results obtained at one single resolution. To overcome these challenges, we built Multi-resolution Reconciled Tree (MRtree), a highly flexible tree-construction algorithm that generates a cluster hierarchy from flat clustering results attained for a range of resolutions. Because MRtree can be coupled with most scRNA-seq clustering algorithms, it inherits the robustness and versatility of a flat clustering approach, while maintaining the hierarchical structure of cells. The constructed trees from multiple scRNA-seq datasets effectively reflect the extent of transcriptional distinctions among cell groups and align well with levels of functional specializations among cells. Importantly, application to fetal brain cells identified subtypes of cells determined mainly by maturation states, spatial location and terminal specification.

INTRODUCTION

Single-cell RNA sequencing (scRNA-seq) is a recently developed technology that enables the identification of functionally distinct subpopulations of cells that each possess a different pattern of gene expression activity (1). These subpopulations can indicate different cell types with relatively stable, static behavior or cell states in intermediate stages of a transient process. Unbiased discovery of cell types from scRNA-seq data can be automated using a wealth of unsupervised clustering and integration algorithms (2–7); however, a major challenge regarding clustering algorithms is that they explicitly or implicitly require the resolution parameter, in most cases the number of clusters, to be supplied as an input parameter.

There are some computational methods available to guide the choice of the resolution parameter; however, these methods are often shown to be ineffective. Some testing-based methods are too sensitive to heterogeneity (8), especially for large samples, while other methods tend to favor a coarse resolution, with clearly separated clusters, and fail to identify closely related and overlapping cell types. Therefore, judgment from the researchers is often required to select the desired resolution. A common practice in scRNA-seq data analysis is to run a clustering algorithm repeatedly for a range of resolutions, followed by careful inspections of individual results by examining the cluster compositions and the expression of published marker genes to select the final partition. This supervised process takes significant time and effort and is limited by the current state of the investigator's and field's knowledge about cell type and cell state diversity. It would be a substantial advantage in terms of efficiency and veracity to be able to reach the same level of resolution in an unsupervised manner.

Hierarchical clustering (HC) offers another approach to identify cell-populations (see (9–11) and references therein).

[*] To whom correspondence should be addressed. Tel: +1 412 268 577; Fax: +1 412 268 7828; Email: roeder@andrew.cmu.edu

© The Author(s) 2021. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

HC has the advantage of being able to determine relationships between clusters of different granularities since the result can be visualized as a dendrogram. This hierarchical structure helps identify multiple levels of functional specialization of cells. Classical hierarchical agglomerative clustering (HAC) based on 'average' linkage is adopted in SC3 (4), a popular clustering algorithm. However, an important limitation of HAC is that it is prohibitively expensive in terms of computation for large datasets. A few scRNA-seq tools expand upon the idea of hierarchical clustering; for instance, pcaReduce (12) introduces an agglomerative clustering approach by conducting dimension reduction after each merge, starting from an initial clustering, and CellBIC (13) performs bisecting clustering in a top-down manner leveraging the bi-modal gene expression patterns. But these methods either require a good initial clustering solution, which is implicitly equivalent to the choice of *K*, or are highly dependent on restrictive assumptions.

In this study, we build a useful tool to bridge the gap between two separate lines of inquiry, flat and hierarchical clustering. Empirically, scRNA-seq data analysts observe that the partitions obtained from flat clustering at multiple resolutions, when ordered by increasing resolution, produce a layered structure with a tree-style backbone (14). This produces a useful representation to help visually determine the stability of clusters and relations among them. We build on this idea and propose a method called *Multi-resolution Reconciled Tree (MRtree)* that reconstructs the underlying tree structure by reconciling partitions obtained at different granularities (see Figure 1 for illustration) to produce a coherent hierarchy that is as similar as possible to the original flat clustering at different scales. It can work with many specially designed flat clustering algorithms for single-cell data, such as Louvain clustering from Seurat (2), thus inheriting the scalability and good performance in clustering the single-cell data; meanwhile, it recovers the intrinsic hierarchy structure determined by the cell types and cell states.

Applications of MRtree on a variety of scRNA-seq datasets, including mouse brain (9), human pancreas (15,16) and human fetal brain (17), showed improved performance for clustering of scRNA-seq data over initial flat clustering methods. The hierarchical structure discovered by MRtree easily outperformed a variety of tree-construction methods. Moreover, the results accurately reflect the extent of transcriptional distinctions among cell groups and align well with levels of functional specializations among cells. Particularly, when applied to developing human brain cells, the method successfully identified major cell types and recovered an underlying hierarchical structure that is highly consistent with the results from the original study (17). Subsequent analysis on each major type via MRtree revealed finer sub-structure defining biologically plausible subtypes, determined mainly by maturation states, spatial location and terminal specification.

MATERIALS AND METHODS

MRtree aims to recover a hierarchical tree by denoising and integrating a series of flat clusterings into a coherent tree structure. The algorithm starts by applying a suitable flat clustering algorithm to obtain partitions for a range of resolution parameters. The multi-scale results can be represented using a multi-partite graph, referred to as a *cluster tree*, where the nodes represent clusters, and edges between partitions of adjacent resolutions indicate common cells shared. We propose an efficient optimization procedure to reconcile the incoherent cell assignments across resolutions that produces the optimal underlying tree structure following the hierarchy constraints, while adhering to the initial flat clustering to the maximum extent. Formally, this is achieved by minimizing (among valid hierarchical tree structures), the difference between initial multi-level cluster assignments and the cluster assignments in the resulting tree structure. By representing the partitions as a multi-partite graph, the clustering assignments that violate the hierarchy constraint can be identified as merging directed edges and thus penalized in the objective function. The optimization procedure proceeds by iteratively and greedily identifying those tree nodes, which, when corrected by reassigning the associated conflicting cell lineages, contribute to maximum descent in the defined objective function. The outcome of the proposed optimization procedure is a reconciled tree, named the *hierarchical cluster tree*, representing the optimal tree-based cluster arrangement across scales (see Figure 1).

Our method is motivated by consensus clustering (also known as ensemble clustering); however, instead of gathering information over repeated runs of algorithms at the same resolution, we leverage the cluster structure revealed at multiple scales to build an ensembled hierarchy. The common features across resolutions are identified and averaged to reduce noise, while the distinctions between resolutions are utilized to uncover different scales of geometric structure, which are further reconciled to conform to a robust hierarchical tree. We stress that consensus clustering is essentially a noise-reduction technique that aims to deliver robust, interpretable results.

Another key distinguishing feature of our procedure is that we build a cluster hierarchy directly from the raw partitions in an 'in place' way. This is in comparison to existing methods for which a hierarchical clustering algorithm is applied to the similarity matrix or the co-classification consensus matrix built from an ensemble of partitions. For instance, Dendrosplit (18), a recently developed hierarchical clustering method, deploys HAC on the cell-by-cell distance matrix, followed by additional steps of merging pairs of clusters if their separation score is below a threshold. The performance of the clusters relies heavily on the HAC performance. However, it is widely recognized that no single clustering method will perform best across all datasets. MRtree adapts to this challenge by accepting multi-resolution clustering results from most state-of-the-art clustering methods. It uses an optimization framework to edit the original partitions through a similar voting scheme. At the same time, it aims to preserve the original splitting order of the hierarchy determined by the clustering algorithm. The proposed method is efficient in terms of memory cost and time complexity (Supplementary Information). Moreover, MRtree enables a direct comparison of partitions before and after tree reconciliation to examine the stability of the clustering algorithm at different scales. As a benefit, we are able to trim the tree to the maximum depth within the stable range to obtain reliable final clusters.

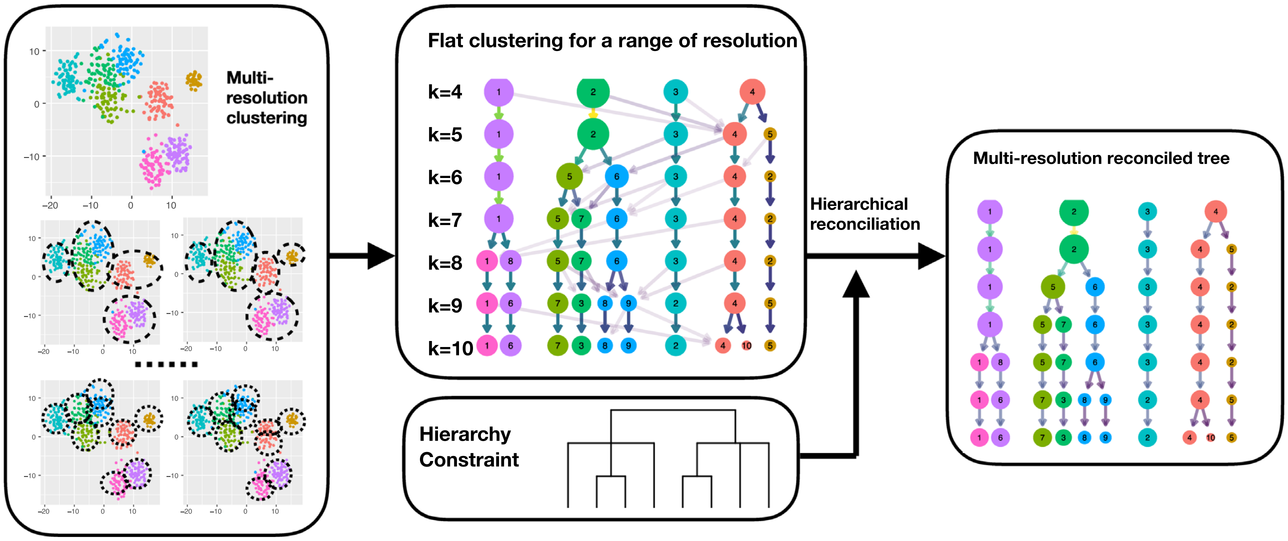
Multi-resolution clustering
Flat clustering for a range of resolution
k=4
k=5
k=6
k=7
k=8
k=9
k=10
Hierarchical reconciliation
Multi-resolution reconciled tree
Hierarchy Constraint

Figure 1. Overview of MRtree framework. The algorithm starts by performing flat clustering on scRNA-seq data for a range of resolutions, where the partitions between adjacent resolutions are matched to form a graph as an entangled cluster tree. Then reconciliation is performed through optimization with the hierarchical structure enforced by constraints. The obtained final optimal solution represents the recovered hierarchical cluster tree.

Optimization scheme

For X, an n-by-p matrix of transcriptomes for n cells on p genes, clustering is performed using algorithm $\mathcal{A}$ at a range of m resolutions. Here the resolution parameters, $\{k_1, \ldots, k_m\}$, are loosely defined where it corresponds explicitly to the number of clusters for some algorithms, while it implicitly determines the number of clusters for other algorithms. To formally state the problem and the hierarchical reconciliation algorithm, we first introduce some notation.

Definition 1. A cluster tree $T_c(k_1, \ldots, k_m)$ at resolution levels $(k_1, \ldots, k_m)$ is a directed m-partite graph with vertex set $V(T_c)$ and edge set $E(T_c)$. Denote the set of all cluster trees as $\mathcal{T}_c(k_1, \ldots, k_m)$.

Here the vertex set $V(T_c)$ is the union of m subsets, where each set $V_j(T_c)$ consists of k_j nodes denoted as $\{v_{j,1}, \ldots, v_{j,k_j}\}$. Each node represents a cluster in the partition of n cells into k_j clusters, namely, $v_{j,k}$ represents the k-th cluster at the j resolution level. Each direct edge $e_{v_{j,k},v_{j+1,k'}}$ is defined between adjacent layers pointing from a lower resolution cluster $v_{j,k}$ to a higher resolution cluster $v_{j+1,k'}$ whenever there are overlapping samples between these two clusters at different resolutions. Further, Let $v_{\text{in}}(e)$ and $v_{\text{out}}(e)$ be the in-vertex and out-vertex of edge e.

Definition 2. We call a cluster tree a hierarchical cluster tree, denoted as $T_h(k_1, \ldots, k_m)$, if it satisfies the following constraint:

Constraint A_1: Each node $v_{j+1,k}$ has one and only one in-vertex edge.

Denote the set of all hierarchical cluster trees at resolution $(k_1, \ldots, k_m)$ as $\mathcal{T}_h(k_1, \ldots, k_m)$.

Condition A_1 ensures that any cluster in a higher resolution belongs to one and only one cluster in the adjacent lower resolution. It further implies that the hierarchical tree can only be a branching tree as the resolution increases (top-down), and the clusters in lower levels should be intact in levels above it. Compared to the cluster trees, hierarchical cluster trees respect the clustering structure at higher resolutions in the sense that they keep samples that are together at higher resolutions in the same cluster for lower resolutions. Similarly, those samples that are far away from each other at a lower resolution do not enter the same clusters at high resolutions. We illustrate an example of a hierarchical cluster tree and its noisy companion in the form of a cluster tree in Supplementary Figure S1.

Arrange the the clustering results at each resolution inside a an n-by-m label matrix $L(T_c) := [L_1, \ldots, L_m]$, where the j-th column denotes the corresponding labels for each data point at resolution k_j.

Definition 3. For each data point x_i, $i = 1, \ldots, n$, define its clustering path $p(x_i) := (v_{1,l_{i1}}, \ldots, v_{m,l_{im}})$ where $v_{j,l_{ij}}$ is the label for x_i at resolution k_j. Let $\mathcal{P}(T_c) := \{p(x_i) \mid i = 1, \ldots, n\}$ be the set of all unique paths.

Let $T_c(k_1, \ldots, k_m)$ be the initial cluster tree by applying clustering algorithm $\mathcal{A}$ on X, and let $T_h^*(k_1, \ldots, k_m)$ be the underlying true hierarchical cluster tree. Further denote the two respective n-by-p label matrices as $L(T_c)$ and $L(T_h^*)$. Our goal is to recover the unknown hierarchical tree from the observed initial cluster tree, working from the multi-resolution flat clustering. Assuming that $T_c(k_j)$ is an estimator of $T_h^*(k_j)$, $j = 1, \ldots, m$, if $T_c(k_1, \ldots, k_m)$ satisfies constraint A_1, it naturally yields an estimator of $T_h^*(k_1, \ldots, k_m)$, though this is rarely the case. Following this idea, we construct the estimator by building a hierarchical cluster tree that mostly preserves the cluster structures from the observed cluster tree $T_c(k_1, \ldots, k_m)$ constructed from the initial flat clustering results. To achieve this, we define a loss function as the distance between the solution tree and the initial flat clusterings $T_c(k_1, \ldots, k_m)$. We seek to minimize the loss under the constraint that the solution tree satisfies constraint A_1. To measure the difference between two trees, which is equivalent to measuring mismatch between two sets of partitions, we adopt hamming distance between the respective label matrices. Hamming distance computes the number of location-mismatches of a pair of matrices, commonly used for measuring the distance between two paired partitions.

The problem formulated above is equivalent to finding the optimum k_m distinct paths from the set of all feasible paths (Def. 3) of a cluster tree, to which all data points are assigned, and the induced multi-scale partitions preserve the most flat clustering structures. It is a combinatorial optimization problem. The complexity grows exponentially with the depth and number of clusters in each layer of the tree, and therefore is computationally intractable. To alleviate the computational burden, we introduce an equivalent objective function and propose a greedy algorithm to solve it. Formally, define $\tilde{V}(T)$ to be the set of 'bad' vertices that have more than one in-vertex edge. Then for any proposed hierarchical cluster tree this set is empty. The hierarchy is therefore estimated by solving the optimization problem respective to the newly formulated constraint,

$$\hat{T}_h = \arg\min_{T} \min_{\pi} \mathcal{D}_{\text{Hamm}}\left(L(T), \pi(L(T_c))\right) \qquad (1)$$

subject to $\tilde{V}(T)$ being an empty set, where $\mathcal{D}_{\text{Hamm}}(\cdot, \cdot)$ represents the hamming distance. The objective is minimized over permutation of labels $\pi = \pi_{k_j}, j = 1, \ldots, m$ within each partition since the error should not be depending on how we label the classes.

We employ a greedy optimization procedure. The formulated problem (1) is first transformed to a soft constraint problem that shares the same solutions which allows for constraint violation during the optimization procedure. This enables initializing the solution with the observed flat cluster tree $T_c(k_1, \ldots, k_m)$. The objective is then minimized by sequentially 'cleaning' one bad vertex at a time. Here 'cleaning' refers to eliminate all but one edge that have this node as its in-vertex, followed by re-routing data points belonging to the eliminated path to remaining nearest viable paths. The increase in the objective as the results of cleaning the node is considered as the cost of eliminating the violation from $\tilde{V}(T)$. In each iteration, the vertex in set $\tilde{V}(T)$ is evaluated for its elimination cost, where the one with the minimum cost is selected. The tree is then updated with the selected node being cleaned and affected data points re-assigned to the nearest remaining paths. The procedure is repeated until no violations remain. The full algorithm is summarized in Algorithm 1. In Supplementary Information we analyze the key properties of the algorithm: Theorem 1 provides the convergence properties, while Theorem 2 describes the memory and time complexity. In addition, we introduce methods for sampling implicit resolution parameters with uniform coverage for modularity-based clustering (Seurat clustering), including linear sampling, exponential sampling, and most preferably, Event Sampling method (Supplementary Figure S2). We also discuss ways of speeding up the algorithm in case of large sample size or a large number of initial flat clusterings through layer-wise reconciliation and performing within-resolution consensus clustering as the first step.

Stability analysis to determine tree cut

We consider clustering stability to determine the tree cut based on a basic philosophy that clustering should be a structure on the dataset that is 'stable'. That is, if applied to datasets from the same underlying model, a clustering algorithm should consistently generate similar results. Higher stability across resolutions is reflected as greater consistency of individual initial flat clustering with the resulting clustering in the reconciled tree. To measure the stability, we calculate the similarity using ARI between clusterings in corresponding layers from the initial cluster tree and the resulted hierarchical cluster tree. This will generate a line plot showing the similarity with increasing resolution. The tree cut can then be determined by finding the 'change point' where the stability is high at the current point and decrease sharply by further increasing the resolution.

Evaluation metrics

To quantify the clustering performance in each layer of the hierarchical tree, we utilize a novel modified version of Adjusted Rand Index (ARI) (19), called Adjusted Multi-resolution Rand Index (AMRI, Supplementary Information, Supplementary Table S1), as the accuracy metric to compare the multi-resolution cluster structures with the true labels. The adjustment allows for comparisons across resolutions, accounting for the reduced ability to uncover details in lower resolutions, thus avoiding a bias towards fine-grained clustering results.

Tree construction accuracy. To evaluate the accuracy of tree construction, given the true tree is known, we reduce the actual and resulted trees to similarity matrices and measure the distance between them. Each entry of the matrix represents the length of branch two data points share. The longer branch a pair share, the more similar they are. In this way, we convert the measurement of the difference between hierarchies (dendrograms) to measure the difference between two similarity matrices. The between-similarity distance is measure with the L_1 norm of the difference, defined by

$$D(T_1, T_2) = \| A_1 - A_2 \|_1 = \sum_{i,j} |A_{1,ij} - A_{2,ij}|, \qquad (2)$$

where A_1, A_2 are the similarity matrices of tree T_1, T_2 respectively. The similarity between pairs of samples is defined by $A_{ij} = 1 - D_{ij}/2$ where D_{ij} is the sum of distances of node i and j from the least common ancestor in the given tree. Given the certain tree structure, the induced similarity metrics can be visualized in Supplementary Figure S3.

Cluster stability. Apart from examining the performance of MRtree for clustering accuracy, we also access the stability of the clusters at multiple resolutions prior to and post to tree reconciliation. Clustering stability has been considered as a crucial indicator of goodness of the clusters, given that well-performed partitions tend to be consistent across different sampling from the same underlying model or of the same data generating process (20). In practice, a large variety of methods has been devised to compute stability scores.

Here we adopt the sub-sampling procedure, where the same clustering method is repeatedly performed on the independently sub-sampled datasets and compute the average similarity among the repetitions. Formally given a dataset of n points S_n, let $\mathcal{C}_k(S_n)$ be the resulted clustering outcome with k clusters. Let $\tilde{S}_n^{(b)}$ be a sub-sampling of S_n by randomly choosing a subset of size τn without replacement. Then the stability score is obtained by averaging the partition similarity on the shared data points,

$$Stab(k, n) = \frac{1}{B}\sum_{b=1}^{B} ARI(C_k(S_n), C_k(\tilde{S}_n^{(b)})). \quad (3)$$

The higher the stability score, the more stable the clustering procedure is regarding the noise in the data. We use $\tau = 0.95$ in our experiments.

RESULTS

Simulation study

To investigate how well MRtree is able to recover the cluster hierarchy and improve the clustering across resolutions, we harness the tools provided by the SymSim package (21) to simulate scRNA-seq data given a known tree structure, using the SymSim parameters estimated from a UMI-based dataset of 3005 mouse cortex cells (9) (Supplementary Information). Motivated by major cell types identified in brain tissues, we constructed a hypothetical tree (Figure 2A,B) as the ground truth representing the hierarchy of the cell types/states.

Repeated simulations were performed by first generating single-cell data with SymSim from a hypothetical tree structure, followed by multi-resolution flat clustering using a variety of clustering methods. Then MRtree was applied to form the hierarchical cluster tree that reconciled the multilevel clusterings. MRtree can be coupled with most flat clustering methods; hence we evaluated the performance using a variety of algorithms, including Seurat (2), SC3 (4), SOUP (22), K-means applied to a UMAP projection, and jSRC (23). The clustering results from MRtree-constructed trees were evaluated and compared with the raw flat clustering results and hierarchical clustering outcomes in three aspects: the accuracy of clustering regarding label assignments at different resolutions, the tree structure estimation accuracy and the clustering stability.

We first sought to quantify how well MRtree performed regarding clustering accuracy, measured using Adjusted Multiresolution Rand Index (AMRI) between the obtained labels and true labels known from the simulation. An AMRI close to 1 indicates perfect clustering given the resolution. MRtree achieved higher accuracy almost uniformly across resolutions compared to raw flat clustering for all five methods (Figure 2C and Supplementary Figure S4). It is worth noticing that the reconciliation procedure even improved upon SC3 results, which already employed an ensemble-based method for each fixed resolution. This demonstrates that applying an ensemble approach across resolutions captures additional structural information within the data. In addition, the gain was more pronounced for coarse clustering and when there was more room for improvement. It worth noting that MRtree is not a competitor to other clustering techniques, such as Seurat, jSRC and SC3, rather it aims to enhance clustering techniques by borrowing strength across resolutions. Since the Seurat and SC3 packages both provide tools for constructing hierarchical trees, we also investigated whether MRtree was able to generate superior clusters compared to existing methods in each layer of the tree structure. For Seurat, an agglomerative hierarchical cluster tree was built starting with the identified Seurat clusters, while for SC3, a full HAC was performed from the consensus similarity matrix constructed by aggregating clustering results with different dimension reduction schemes. Similarly, we built a HAC tree from jSRC clusters with the jSRC low dimensional representations. Compared with these hierarchical results, MRtree-constructed trees were judged significantly more accurate and stable based on the higher AMRI and lower variance observed across repeated simulations.

Next, we evaluated the ability of MRtree to recover the tree structure. For comparison, we again leveraged the tools that build hierarchical trees in Seurat and SC3. MRtree produced a significantly improved tree structure compared to the competing methods, as demonstrated by the reduced error of tree reconstruction (Figure 2D,E; Materials and Methods).

Finally we evaluate the clustering stability before and after tree reconciliation, coupled with multiple clustering methods. The stability score is calculated following the subsample procedure described in Methods. For K less than the true number of clusters, the measured stability is confounded by the instability induced by the incorrect resolution. Therefore, we restrict our comparison to the measured stability at the true resolution. Clustering stability with MRtree is clearly improved compared to the initial clustering across all methods (Supplementary Figure S5), demonstrating the improved robustness of MRtree, which successfully employs the consensus mechanism to denoise the individual clustering with collective information across resolutions.

scRNA-seq data

Mouse brain cells. We illustrate MRtree using a scRNA-seq dataset containing 3005 cells of somatosensory cortex and hippocampal-CA1 region from mice, collected between postnatal 21 and 31 days. We call this the mouse brain data (9). The authors have assigned the cells to seven major types: pyramidal CA1, pyramidal SS, interneurons, astrocytes-ependymal, microglia, endothelial-mural and oligodendrocytes. For comparison, these labels are treated as the gold standard in the following analysis.

We chose SOUP (22) for multi-resolution clustering due to its superior performance on these data. The clustering labels were obtained by varying the resolution parameter for a targeted number of clusters from 2 to 12 (note that SOUP hard-clustering can produce fewer clusters than the supplied resolution parameter if the data clearly fit better with a more parsimonious choice). With MRtree, we were able to construct a hierarchical cluster tree from the flat sequential clusterings. The initial cluster tree is visualized with nodes colored by the major type referencing the gold stan-

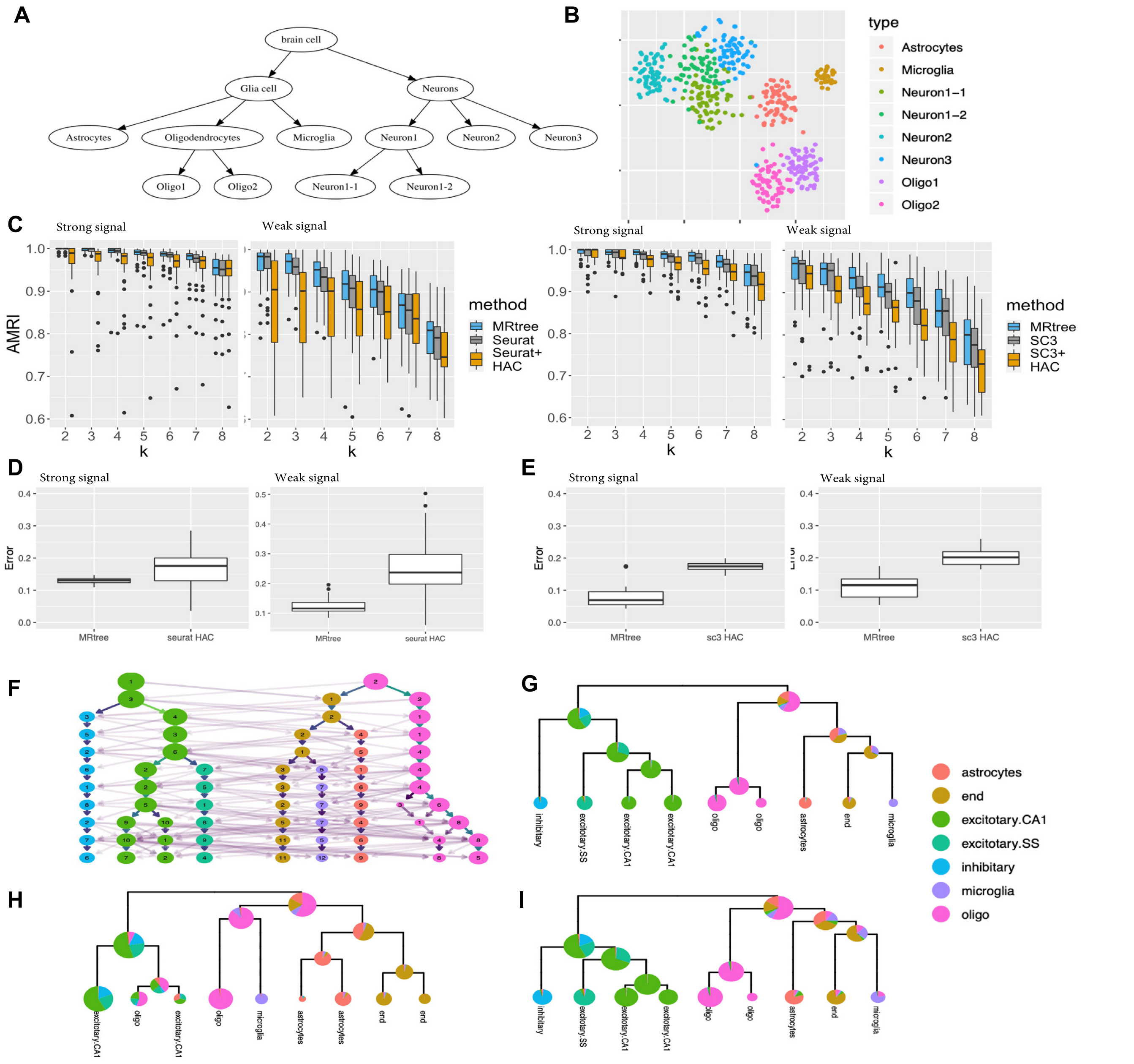

Figure 2. Evaluate the performance of MRtree via simulations and analysis on mouse brain data (9). (**A**) The hypothetical tree structure of the cell states from which cells are generated. (**B**) tSNE plot of the simulated cells in one experiment, colored by the cell types indicated in the leaf of the actual hierarchical cluster tree. The difficulty of the simulation varies from simple (strong signal with higher cluster separation) to challenging (weaker signal with stronger noise) in panels (**C**–**E**) for each method. (C) Comparing the accuracy of MRtree clusters with the clusters from initial flat clustering, and hierarchical clustering at multiple resolutions using Seurat (left panels) and SC3 (right panels). The accuracy is measured by the Adjusted Multi-resolution Rand Index (AMRI). (**D** and **E**) Evaluate tree construction accuracy of MRtree with dendrogram from hierarchical clustering obtained with Seurat (D) and SC3 (E). (**F**–**I**) MRtree applied to scRNA-seq data from the mouse brain. (F) Initial flat clustering by SOUP on 3005 cells (9) by varying the resolution parameter specifying the targeted number of clusters, colored by the gold standard labels. (G) The MRtree-constructed tree from initial SOUP clusterings. The pie charts on tree nodes represent the cell type composition referencing the gold standard. (**H** and **I**) Comparing tree construction and clustering accuracy on mouse brain data using different methods, hierarchical cluster tree generated by HAC starting with SOUP clusters (H) and starting with individual cells (I).

dard, and the recovered tree from MRtree is shown on the right, with the proportion of cell types in each node visualized by a pie chart (Figure 2F,G). The tree successfully split the neurons and glial cells at an early stage, followed by splitting pyramidal cells from two regions (CA1,SS) from the interneurons. Finally, cells from the same type but distinct brain regions were identified. The tree reconciliation step also improved the clustering performance by increasing the accuracy measured by AMRI in multiple layers (Supplementary Figure S6a).

To further compare the performance of MRtree with HAC, we applied HAC using complete linkage on the first 20 principal components, starting either from singletons (individual cells) or the 9 SOUP clusters obtained at the maximum resolution (Figure 2H,I; only the top layers of HAC from singleton are shown for comparison purposes). Compared to MRtree, HAC shared a similar overall tree structure, but it generated clusters at lower accuracy for each layer. The results support the argument that MRtree is able to improve accuracy upon initial clustering by pooling information across resolutions. HAC from singletons performed much worse regarding both accuracy and tree structure, possibly owing to the sensitivity of HAC to outliers and linkage selection. For completeness, we also demonstrate the accuracy from two widely applied clustering methods, Seurat and SC3, where the HAC results were generated from the built-in functions provided as part of the toolkits. In both cases, MRtree outperformed both the initial flat clustering and the HAC (Supplementary Figure S6b,c).

In addition to improving the clustering accuracy, we were able to infer the resolution that achieved the highest stability by inspecting the difference between the initial tree and the reconstructed tree. It indicated that both the SOUP and Seurat algorithms should stop splitting at $K = 7$, which was consistent with the gold standard (Supplementary Figure S7). Stability analysis on SC3 results showed a preferred resolution of 6 clusters. Indeed, we observed steep drop in accuracy for any resolution >6 (Supplementary Figure S6c). By comparison, using available K-selection methods supported in multiple single-cell analysis pipelines, the optimal number of clusters selected varied widely (Supplementary Table S2). For instance, SC3 supported 22 clusters. In addition, the large gap between MRtree and initial Seurat clusterings indicated the inability of Seurat to identify accurate and stable clusters on this dataset. This observation was further supported by the lower accuracy (AMRI < 0.6) of the resulting Seurat clusters (Supplementary Figure S6b).

Human pancreas islet cells. To evaluate performance on cell types that are fairly well separated, we investigated the hierarchical structure identified by MRtree for cells from human pancreatic tissues. We first analyzed single-cell RNA sequencing of 635 cells on islets from Wang *et al.* (15), which come from multiple donors, including children, control adults and individuals with Type 1 or Type 2 diabetes (T1D, T2D). Among them, 430 cells were annotated by the authors into seven cell types, while 205 cells were considered ambiguous and unlabeled. We applied MRtree to construct the hierarchical cluster tree based on SC3 flat clustering with the number of clusters ranging from 2 to 15. The tree was then trimmed to eight leaf nodes based on stability analysis (Figure 3A and Supplementary Figure S8A). The first split created two large interpretable cell groups: gene ontology (GO) shows enrichment of exocrine functions such as terms related to 'Putrescine catabolic process' (adjusted P-value $= 2.3E - 02$) and 'Cobalamin metabolic process' (adjusted P-value $= 5.48E - 05$) for the left branch, and enrichment of endocrine functions such as 'Insulin secretion' (adjusted P-value $= 3.4E - 5$) and 'Enteroendocrine cell differentiation' (adjusted P-value $= 2.1E - 2$) for the right branch. The exocrine group was further divided into acinar (*PRSS1*) and ductal cells (*SPP1*). The right branch further separates a previously undiscovered cluster composed mainly of ambiguous cells and a few previously labeled alpha and mesenchyme cells. This cluster expresses marker genes with significant GO terms such as 'Collagen metabolic process' and 'regulation of endothelial cell migration', pointing to endothelial and stellate cells (Supplementary Table S3) that were not labeled in the original analysis. The remaining endocrine cells were further divided into a group containing α cells (*GCG*) and pancreatic polypeptide cells (*PPY*), and another group containing β (*INS*) and δ cells (*RBP4*) (Supplementary Table S4).

In addition to recapitulating a logical tree for all cell types, the eight clusters improved upon the initial SC3 clusters. In particular, seven of the clusters match well with the identified seven major cell types from Wang *et al.*, achieving AMRI >0.95 (Supplementary Figure S8B–D). By contrast, a competing tree construction method, CellBIC (13), revealed a similar tree structure, but it failed to identify the group of δ cells (13). Finally, because it is well accepted that β cells are heterogeneous, especially in conditions of metabolic stress, such as obesity or type 2 diabetes (15), we further applied MRtree on the subset of 111 β cells. We obtained five β subclusters that corresponded to key biological features, including two clusters composed mainly of cells from T2D individuals, and one control group containing 90% cells from children (Supplementary Figure S8E–G).

Next, we considered a more challenging dataset, again from the human pancreatic islet, produced by merging data from five technologies (16). In total, 14 892 cells were annotated and grouped by respective studies into 13 major cell-types with cluster sizes varying by magnitude. We first integrated the cells using Seurat MNN integration tools using 2000 highly expressed genes (Figure 3B). Despite the observation that SC3 demonstrates superior performance on the smaller datasets, we utilized Seurat graph-based clustering because it demonstrates greater scalability to large-scale analysis. Flat clusterings were obtained for 50 different resolution parameters sampled via Event Sampling in the range of [0.001, 2]. The resulting tree identified all 13 major types with high accuracy and also uncovered many subtypes organized as subtrees (Figure 3C). Very distinct cell types separated early and fall into remote branches, while cell types that share similar functions share internal branches and split later in the process. For instance, endothelial, schwann and stellate cells are very different from other endocrine and exocrine cells and thus split out first. Two types of endocrine cells, acinar and ductal, fall into a common subtree. Likewise, five types of exocrine cells are organized in the same subtree. Finally, subtypes from the

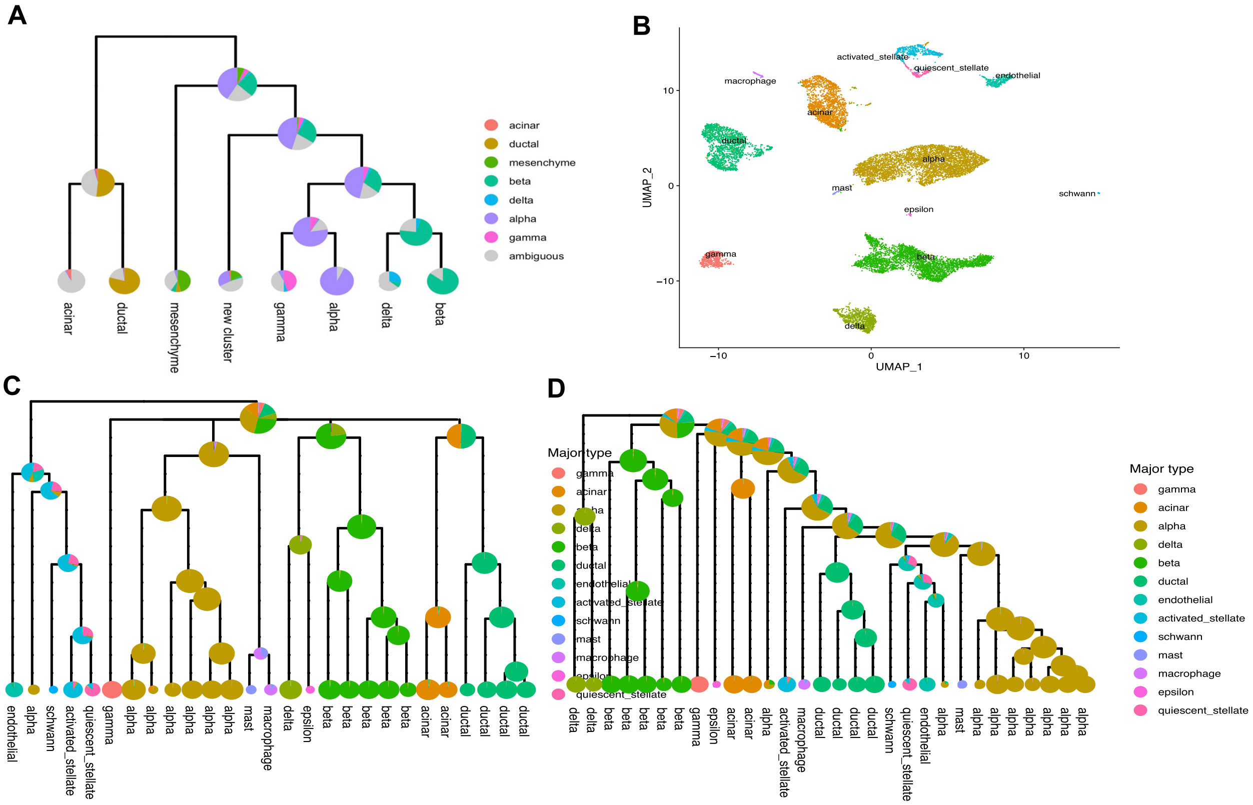

Figure 3. MRtree applied on pancreas islet cells datasets reveals the transcriptional distinctions and similarities between cell types. (**A**) MRtree-constructed tree with SC3 clusterings on 635 cells from Wang *et al.* (15). The tree was trimmed to the layer with eight leaf clusters ($K = 8$). The pie charts overlying on tree nodes represent the cell type composition for corresponding clusters. Colors indicate the cell-type labels by Wang *et al.*, where a fraction of cells (marked in gray) were considered ambiguous cells by the authors and unlabeled. The leaf labels demonstrate the inferred cluster identity. (**B–D**) Jointly constructing the cell type hierarchical tree for pancreas islet cells integrated from five technologies. (B) UMAP project of 14 892 cells integrated from five technologies using Seurat MNN integration tools, colored with the cell type labels from respective studies. (C) MRtree-constructed tree from the integrated data with Seurat initial flat clusterings. Pie charts on tree nodes show the cell-type composition given the referencing labels from the studies. Leaf labels indicate the inferred labels of cells in each leaf node. (D) Hierarchical tree constructed by Seurat agglomerative hierarchical clustering starting from Seurat flat clustering results obtained with the highest resolution, annotated similarly by cell type compositions.

same major type are organized in the same subtree, with one exception. A small subset of α cells was inappropriately placed in the tree. However, evidence suggests these cells represent an anomaly, possibly due to batch correction. These α outliers appear in the UMAP projection separated from other α cells and near the activated stellate cells.

For comparison, we produced a hierarchical tree using Seurat agglomerative clustering (Figure 3D). Given the well-separated cluster structure of cell types in the projected PCA space, it is not surprising that the tree also identifies all the major cell types; however, the hierarchical structure appears less reasonable. For instance, the activated and quiescent stellate cells were placed far from each other in the tree, and two endocrine types were grouped in different subtrees. In summary, MRtree produced a more useful tree than competing methods for both applications, and the interpretable subtree structure observed across applications shows promise for further investigation of the cell subtypes identified here.

Human fetal brain cells. We applied MRtree to cells from the mid-gestational human cortex, which we call the human brain data (17). These data were derived from ∼40 000 cells from germinal zones (ventricular zone and subventricular zone) and developing cortex (subplate (SP) and cortical plate (CP)) separated before single-cell isolation. By performing Seurat clustering (2), the authors assigned the cells into 16 transcriptionally distinct cell groups (Supplementary Table S5). For convenience, here we refer to these expert classifications as the Polioudakis labels.

Our analysis began with the same preprocessing steps as conducted in the study (17) using the pipeline supported by Seurat V3. The multilevel clustering results are visualized by increasing resolution from the top layer (resolution = 0.001) to bottom layer (resolution = 2), where each layer corresponds to one clustering (Supplementary Figure S9A). Notably there were a considerable number of cells assigned to clusters inconsistently over changing resolutions, which made it challenging to determine the optimal resolution and the final cluster memberships. By applying MRtree, we were able to construct the organized hierarchical tree, which was represented by a dendrogram with the cell-type composition of clusters referencing Polioudakis labels shown by pie charts on tree nodes (Figure 4A). MRtree first separated

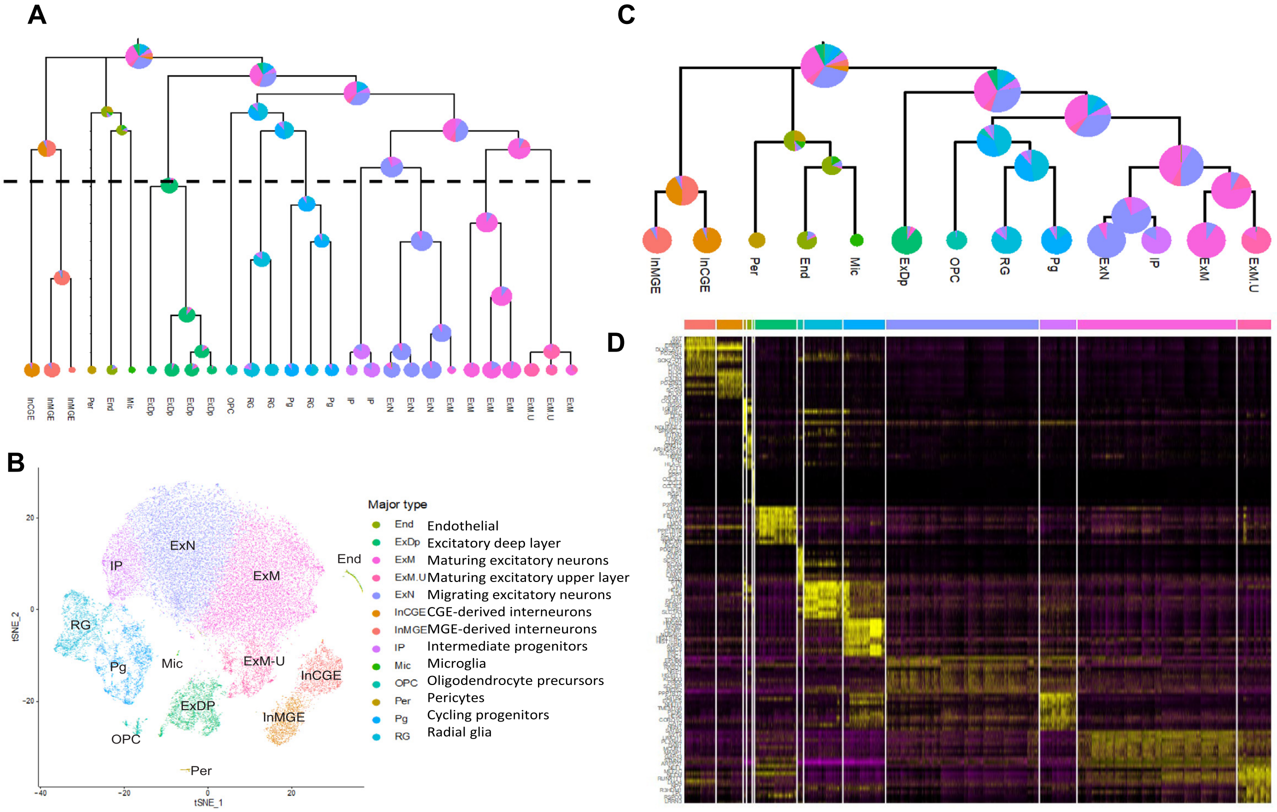

Figure 4. MRtree applied to scRNA-seq data from human brain cells. (**A**) MRtree produces the hierarchical cluster tree from the initial flat clusterings at multiple resolutions obtained from Seurat. The nodes correspond to clusters, with a pie chart displaying the cluster composition referencing the Polioudakis labels. Tree cut is placed at tree layer corresponding to $K = 13$ based on stability analysis, above which the clusters are stable. (**B**) tSNE plot of all 40 000 human brain cells colored by which of the 13 major clusters the cells belong to, with cell-type identities names hovering over clusters in black. (**C**) The 13 major clusters were obtained by cutting the tree at the level indicated by the dashed line in (**A**), indicating the identified major cell types and the associated stable hierarchical structure. (**D**) Heatmap of the top 10 significant marker genes (FDR-adjusted P-value<0.05) for the identified 13 major clusters ranked by average log fold change, arranged according to the display of tree leaf nodes.

interneurons and pericytes, endothelial and microglia, followed by splitting excitatory deep layer neurons from radial glia to maturing excitatory neurons, representing the rest of a closely connected lineage (upper layer enriched). By further increasing the resolution, the radial glia cells and excitatory neurons were isolated, where the intermediate progenitors were more closely connected with maturing excitatory neurons. The finer distinctions of excitatory neurons were subsequently identified as migrating, maturing and maturing upper enriched subtypes supported by differential gene expression and canonical cell markers (Supplementary Table S6). The results were consistent with the group-wise separability visible through a 2-dimensional tSNE projection (Figure 4B). The cluster stability was inspected by comparing the initial Seurat clusters at each resolution with the MRtree results (Supplementary Figure S9B), which suggested that the clusters were stable up to around $K = 15$. We decided to cut the tree at $K = 13$, which corresponded fairly closely to the 16 major gold standard cell types of the midgestational brain by examination of differentially expressed marker genes (Supplementary Table S6, Figure 4C,D). For comparison, an agglomerative hierarchical tree was generated starting from the Seurat clusters obtained at the highest resolution (Supplementary Figure S9C). These results were distance-based and consequently more vulnerable to outliers, which appear to have caused several anomalies: subsets of ExM, ExM-U and ExDp1 were grouped together, and two subsets of IP were separated from each other.

Identify subtypes. Next, we scrutinized the fine-grained structure by re-clustering the 13 major cell types obtained from the hierarchical cluster tree of all cells. The cells were pre-processed from the raw count data as performed in the first iteration, followed by clustering using the Seurat graph-based method. By setting the resolution parameters from 0.05 to 1 and applying MRtree, we obtained one hierarchical tree for each major cell type, determined by trimming the full tree to the stable top layers (ExDP and InMGE are depicted in Figure 5A,B). This resulted in 21 transcriptionally distinct cell types from 7 of the identified major types, expanding IP, ExN, ExM, ExM-U, ExDp, InCGE and InMGE (Figure 5C and Supplementary Table S8). The subtypes' partitions were first evaluated by assessing whether the likely technical and biological co-variation, including brain sample, sequencing run and cortical region, illustrated somewhat even distribution and appropriate overlap within each identified cluster. Results show that the clusterings

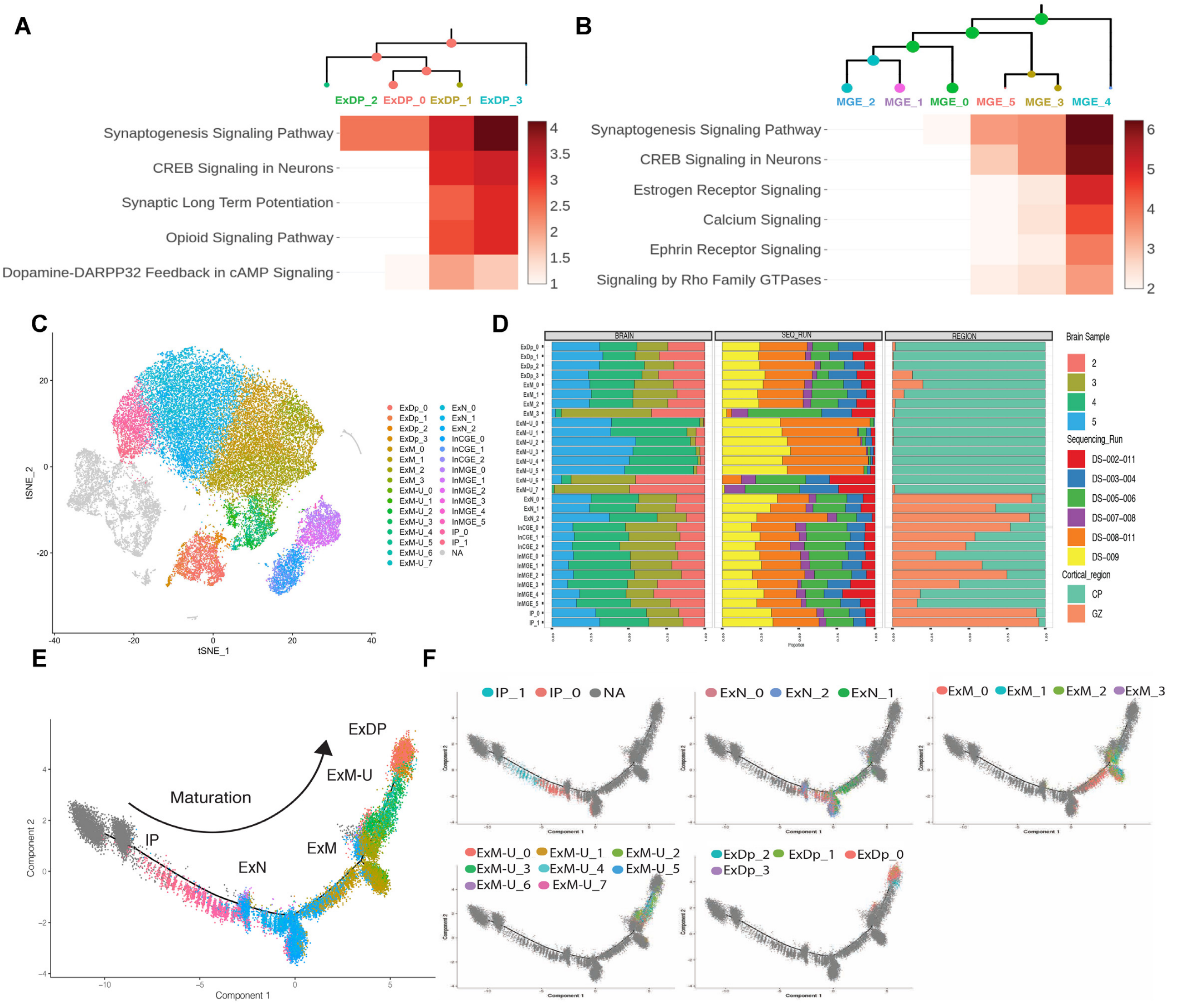

Figure 5. MRtree clusters cells into known cell subtypes and states that underlie known cellular developmental transcriptional trajectories at a higher resolution. (**A**) Hierarchical cluster tree of subplate/ deep layer excitatory neurons (ExDp) with a heatmap of gene expression within canonical gene ontology categories showing a gradually increased enrichment of Synaptogenesis, CREB signaling, synaptic signaling (i.e. synaptic long term potentiation, opioid and dopamine-DARPP32-cAMP signaling) across maturation from ExDP_0 to most mature ExDP_3 cluster. (**B**) Hierarchical cluster tree of MGE-derived interneuron with a heatmap of gene expression within canonical gene ontology categories shows a gradually increased enrichment of Synaptogenesis, CREB signaling and calcium-mediated signaling across maturation from InMGE_2 to most mature InMGE_4 cluster. (**C**) tSNE projection of all cells colored by the MRtree identified subtypes from subsequent analysis of the MRtree major cell types. (**D**) MRtree clusters are driven by biology and not technical co-variation in the data: Histogram of the percentage of cells that each brain sample (left), sequencing run (middle) and cortical region (right) contribute to each cellular cluster identified by MRtree. (**E**) Cells projected onto Monocle pseudotime analysis from Polioudakis *et al*. with cells colored by MRtree cell-types and names hovering above. (**F**) Pseudotime projection of each cluster cell types from MRtree illustrating a continuous developmental trajectory of excitatory neurons, first: top left; intermediate progenitors IP_1, IP_0, top middle; newly born excitatory neurons ExN_0, ExN_2, ExN_1, and, top right; maturing excitatory neurons ExM_0, ExM_1, ExM_2, ExM_3, bottom left; followed by maturing upper layer neurons ExM-U_0 through ExM-U_7 and, bottom left; maturing subplate/ deep layer neurons ExDP_2, ExDP_0, ExDP_1, ExDP_3.

were not driven by these technical features and are likely biologically meaningful (Figure 5D and Supplementary Figure S10).

We focus on the results of excitatory neuronal subtypes, given their critical roles in neurological disorders. Close examination revealed that MRtree clustered cells into well-known cell types and states that underlie known cellular developmental transcriptional trajectories at a higher resolution. Projection of each cluster of cell types from MRtree onto Polioudakis Monocle Psuedotime illustrated a continuous developmental trajectory of excitatory neurons, starting from intermediate progenitors (IP) with IP_1 preceding IP_0. The cells then develop into newly born excitatory neurons in the order of ExN_0, ExN_2, ExN_1, which then grow into maturing excitatory neuron subtypes following the order of ExM_0, ExM_1, ExM_2, ExM_3. The trajectory finally ends at maturing upper layer neurons ExM-U_0 through ExM-U_7 and maturing subplate/deep layer neurons ExDP_2, ExDP_0, ExDP_1, ExDP_3, with ExDP_3 considered as the most mature subtype (Figure 5E,F). The estimated hierarchical tree for subtypes corresponded with gene ontology analysis of differential gene expression between branch cell types. For ExDp, the most distinct subtype was ExDp_3, which was first differentiated from the other subtypes, followed by the split for ExDp_2, and then ExDp_0 and ExDp_1 (Figure 5A). The heatmap of gene expression within canonical gene ontology categories showed a gradual increase in enrichment of Synaptogenesis, *CREB* signaling, synaptic signaling (i.e. synaptic long-term potentiation, opioid and dopamine-DARPP32-cAMP signaling) across maturation from ExDP_0 to the most mature ExDP_3 cluster. We observed similar functional specializations of inhibitory neuron subtypes (Figure 5B). The most mature subtype InMGE_4 was discriminated from the other MGE interneurons first, followed by splitting the second and third most mature subtypes from less mature cells, and finer distinctions were established subsequently in two branches. The heatmap of gene expression within canonical gene ontology categories showed a gradual increase in enrichment of Synaptogenesis, *CREB* signaling, calcium-mediated signaling across maturation from InMGE_2 to most mature InMGE_4 cluster.

MRtree partitioned intermediate progenitor cells into two subtypes (IP_1 and IP_0; Figure 6A) similar to cell types revealed in Polioudakis *et al.*, achieved only after multiple rounds of analysis of flat clustering results. Marker genes for newly born neurons (i.e. *SLA, STMN2* and *NEUROD6*) and intermediate progenitors (i.e. *EOMES, SOX11, SOX4* and *PTN*) showed increased expression markers within IP_0 in contrast to expression of more intermediate progenitors and radial glia genes within IP_1 (i.e. *SLC1A3, VIM, SOX2* and *HES1*) (Figure 6B). Notably, by comparing the significant protein-protein interacting (PPI) networks (Supplementary Table S7) from differential genes (DGE) expressed in IP_1 versus significant PPI network from DGE in IP_0, we observed that IP_1 cells PPI contains a highly connected radial glial genes surrounding *VIM* including *MKi67, SOX2*, for example, whereas, IP_0 cells contain more neuronal-committed genes involving early step in neuronal differentiation including *MAPT, GAP43, CALM2, GRIA2* and *PTPRD* (Figure 6C). Gene ontology analysis further uncovered a switch in the enrichment of *EIF2* signaling, growth factors, and cell cycling pathways (i.e. Sirtuin signaling pathway and *SAPK/JNK* signaling) in IP_1 to more specific neuronal categories like Synaptogenesis, Ephrin Receptor signaling, Reelin signaling underlying migration and neurite pathfinding signaling within IP_0 (Figure 6D).

For ExDP subtypes, a closer examination of the expression of marker genes for layer 5 (i.e., *ETV1, RORB, FOXP1* and *FEZF2*), Layer 6 (i.e., *TBR1, SYT6* and *FOXP2*), shared deep markers (i.e., *RORB, TLE4, LMO3, CRYM* and *THY1*) and subplate makers (i.e., *NR4A2* and *ST18*) showed expression of Layer 5 markers within the least mature cells, ExDP_2 and layer 6 markers within ExDP_0, in contrast to the more mature expression of layer 6-CTIP2 markers within ExDP_3 and mature expression of markers of subplate and layer 6 within ExDP_1 (Figure 6E,F). Surprisingly, ExDP_2 PPI revealed a set of genes and structure similar to an intermediate progenitor with *VIM* at the center of translational control and the expression of neuronally committed genes *SOX4, SOX11, ID2* similar to IP_1, except that neuronal specificity genes within this cluster were linked directly to upper Layer 5 cell fate (i.e., *FEZF2, FOXP1, RORB* and *SYT4*) instead of a general excitatory neuronal lineage seen in IP_1. ExDP_1 subplate cells PPI exhibited a group of connected genes related to more mature cellular properties such as synaptic plasticity and Wnt signaling (i.e., *GRIN2B, CTNNB1, NR2F1* and *NRXN1*) but no energy or translational pathways that were present in ExDP_2. ExDP_3 cells showed the most extensive and unique PPI that illustrated more committed axonal and synaptic pathways underlying specifically Layer 6 *CTIP2*+ cells (i.e.*CALM2, NRCAM, SNCA* and GABAergic postsynaptic machinery) (Figure 6G).

Four other cell types revealed subtypes that were also related to developmental ordering. ExN was partitioned into three subtypes that indicate a gradually increased expression of markers of upper layer excitatory neurons in contrast to no expression of deep layer neuronal programs (Supplementary Figure S11). ExM was partitioned into four subtypes, three of which illustrate gradually increased expression of upper layer markers, in contrast, a fourth that expressed deep layer markers indicating layer 4/5 excitatory neurons (Supplementary Figure S12). InMGE was partitioned into six progressively more mature subtypes (Figure 5B) that demonstrate distinctions in both maturation and terminal specification (Supplementary Figure S13). Finally, the three subtypes of InCGE display a general maturation of CGE interneurons through a gradual decrease in expression of transcription factors along with a gradual increase in expression of axonal-related genes (Supplementary Figure S14). Meanwhile, although the signal was sparse, the PPI network for the allegedly most mature subtype revealed a connection between genes critically involved in post-synaptic glutamate signaling and plasticity, further supporting this conjecture. Additional characteristics of these subtypes can be found in Supplementary Information.

DISCUSSION

In this article, we propose MRtree, a computational approach for characterizing multi-resolution cell clusters ranging from major cell groupings to fine-level subtypes using a hierarchical tree. The approach is based on deriving a multi-resolution reconciled tree to integrate clusterings obtained for a range of different resolutions. The proposed method combines the flat and hierarchical clustering results in a novel manner, inheriting the computational efficiency and scalability from the flat clustering and the interpretability of a hierarchical structure. In comparison, MRtree outperforms bottom-up and top-down hierarchical clustering approaches and provides superior clustering for each level of resolution. MRtree also provides tools for sampling implicit resolution parameters for Louvain clustering. This enables equal coverage of different clustering scales as input for the tree construction process. All clustering methods face the challenge of determining the optimal number of clusters supported by the data. While this problem is in-

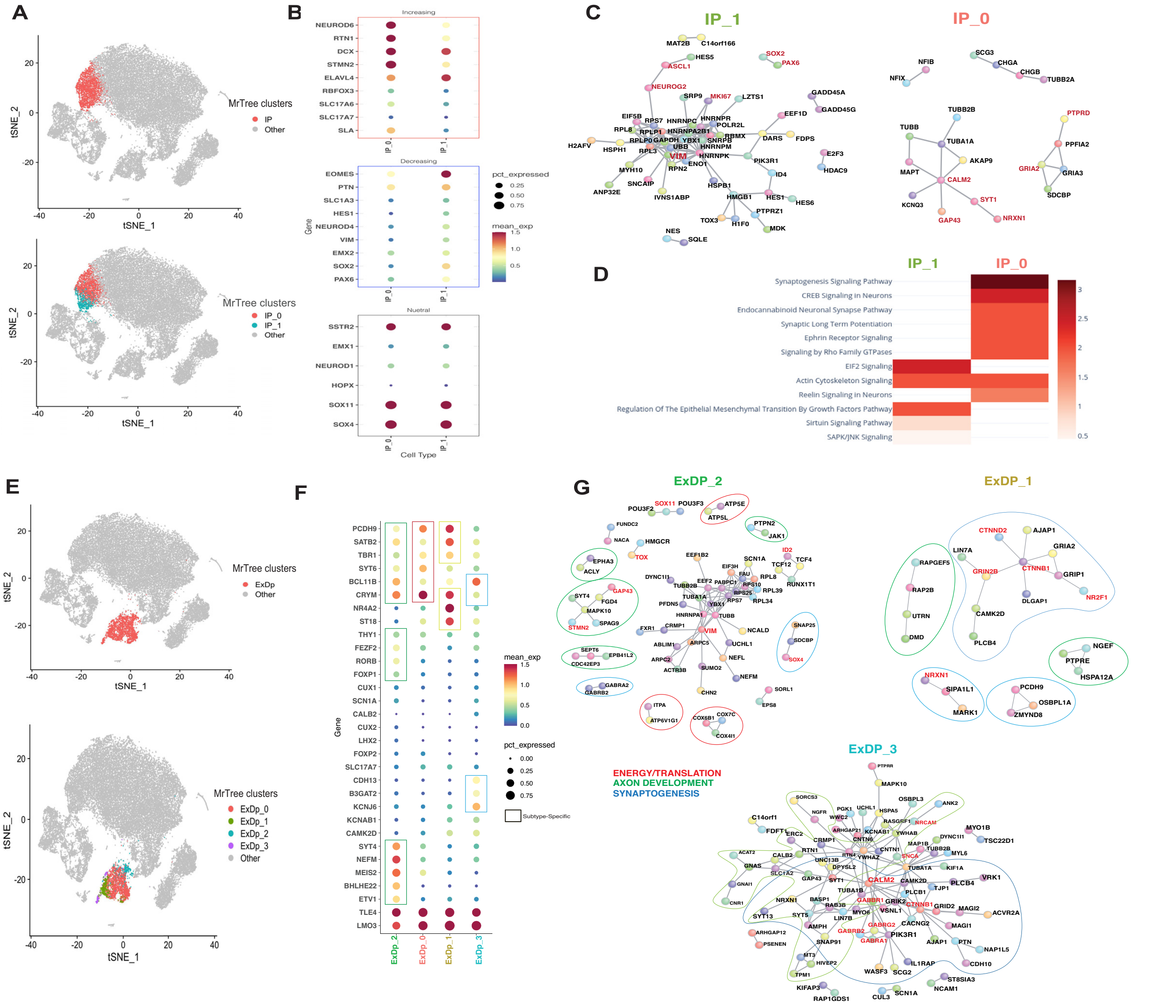

Figure 6. Known and unique biological states identified by MRtree with sub clustering on human fetal brain data: intermediate progenitors and subplate/deep excitatory neurons (**A**) Top: tSNE plot of all cells where intermediate progenitor (IP) cells identified by MRtree are colored by red, bottom: tSNE projection of MRtree clustering where IP is broken into IP_1, colored in blue and IP_0, colored by red. (**B**) Gene expression dot plot showing the normalized mean expression of marker genes for newly born neurons (i.e. *SLA, STMN2* and *NEUROD6*), intermediate progenitors (i.e. *EOMES, SOX11, SOX4* and *PTN*) and radial glia (i.e. *SLC1A3, VIM, SOX2* and *HES1*) within IP_1 (left) and IP_0 (right), grouped by increasing (top), decreasing (middle) and neural (bottom) expressions from IP_1 to IP_0. (**C**) Significant protein-protein interacting (PPI) networks from differential genes expressed in IP_1 on the left versus significant PPI network from DGE in IP_0 on the right. (**D**) Heatmap of IP_1 and IP_0 gene expression within canonical gene ontology categories. (**E**) Top: tSNE projection where subplate and deep excitatory neurons (ExDP) cells identified by MRtree are colored by red; bottom: tSNE projection where ExDP are broken into ExDP_2, colored in blue and ExDP_0, colored by red, ExDP_1 colored by green, and ExDP_3 colored by purple. (**F**) Gene expression dot plot showing the normalized mean expression of marker genes for layer 5 (i.e. *ETV1, RORB, FOXP1* and *FEZF2*), Layer 6 (i.e. *TBR1, SYT6* and *FOXP2*), shared deep markers (i.e. *RORB, TLE4, LMO3, CRYM* and *THY1*) and subplate makers (i.e *NR4A2, ST18*) within ExDP_2, ExDP_0, ExDP_1 and ExDP_3 from left to right. The subtype-specific expressions are marked by brackets. (**G**) Significant protein–protein interacting (PPI) networks from differential genes expressed in ExDP_2 on the top left versus significant PPI network from ExDP_1 top right and PPI from ExDP_3 bottom center.

herently intractable, MRtree uses a stability criterion to determine the maximum resolution level for which stable clustering results can be obtained for a given dataset. Because MRtree is agnostic to the clustering approach, it can readily utilize input from any flat clustering algorithm. Hence MRtree is extremely flexible, immediately incorporating the advantages of available clustering algorithms, while often providing improved clustering at every resolution due to the reconciliation procedure.

To illustrate the performance of our method, we apply MRtree to a variety of scRNA-seq datasets, including cells from the mouse brain, human pancreas and human fetal brain tissues. Coupled with suitable initial flat clustering algorithms, MRtree constructs the hierarchical tree that reveals different levels of transcriptional distinction between cell types and outperforms popular competitors, including bottom-up HAC and divisive methods such as CellBIC (13). For functionally distinct cell types that can be easily identified, the reconciliation process organizes the clusters obtained under different scales into a unified hierarchical structure and suggests a proper tree cut to retain the stable partitions. For instance, the constructed tree from integrated pancreatic islet datasets successfully identified endocrine and exocrine groups and subsequent cell types within each group. The clusters from the tree of mouse cortex datasets accurately recovered the known major cell types organized into subtrees of neurons and glial cells. Application of MRtree on human fetal brain cells uncovered previously recognized main types organized in a tree structure along the maturation trajectories.

Our method has a greater impact in challenging situations where clusters are similar. Apart from validating the method on the widely acknowledged main cell types, we uncovered a list of stable subtypes from the fetal brain dataset that exhibit distinct states and functionality by examining canonical gene ontology categories and significant PPI networks. Specifically, we have shown that two subtypes of intermediate progenitors are well-defined by the expression of radial glia markers versus newly born neurons markers. The subplate/deep layer excitatory neurons are mainly differentiated by the layers the cells will populate. While migrating and maturing excitatory subtypes show a gradual increase of upper layer excitatory neuron markers, upper and deep layer excitatory neuron markers, respectively. Subtypes close in maturation states are reflected in the hierarchical tree as they are split later down the tree. InMGE demonstrates the distinction in both maturation and terminal specification with respect to the engagement of synaptic programs. At the same time, InCGE subtypes differentiate mainly by maturation, which fits nicely with the fact that CGE interneurons are born after MGE interneurons. While both cell types are born in the ventral telencephalon, their terminal specification happens only upon beginning Synaptogenesis when they begin to express subtype-specific markers. Surprisingly, subtypes of ExDP revealed a set of genes and structures similar to an intermediate progenitor that can be further investigated in future work.

It is worth noting that the quality of MRtree's construction relies on the performance of the chosen flat clusterings. If the flat clusterings method inputs unstable or biased clusters, these errors will be largely retained and reflected in the estimated hierarchical cluster tree. Similar to many consensus clustering methods, MRtree can be extended to allow input from multiple sources, each applying different flat clustering methods; however, the quality of the constructed tree depends on the clustering performance of the full spectrum of sources. If the input data provide a disparate signal, then the outcome is likely to be unstable.

Our studies suggest several interesting questions worthy of future investigations. For instance, our method is a general framework that allows for any flat-clustering base procedure. In practice, how to determine which base procedure suits better for different datasets still remains open. In addition, the current framework relies on a rough idea about the range of resolution. Can we automatically decide the range of resolution? How can we select this range when the resolution is not parameterized by the number of clusters? In particular, our current method adopts a stability measure to decide whether to further branch the hierarchical tree. Can we provide theoretical guarantees for the power of this stopping criterion? Furthermore, our work shed light on how major cell types evolve to subtypes, and we would like to further verify these biological findings.

SOFTWARE

MRtree can be constructed using the mrtree R package, which can work directly with Seurat and SingleCellExperiment objects, available on Github (https://github.com/pengminshi/MRtree).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

National Institute of Mental Health (NIMH) [R01MH123184, R37MH057881 to K.R. (in part)]; National Institutes of Health (NIH) [R01-MH116489, R01-MH109912 (to D.G.)]; National Science Foundation (NSF) [CCF-2007911, DMS-2015447 to Y.W. (in part)]. Funding for open access charge: National Institute of Mental Health (NIMH) [R01MH123184].

Conflict of interest statement. None declared.

REFERENCES

1. Buettner,F., Natarajan,K.N., Casale,F.P., Proserpio,V., Scialdone,A., Theis,F.J., Teichmann,S.A., Marioni,J.C. and Stegle,O. (2015) Computational analysis of cell-to-cell heterogeneity in single-cell rna-sequencing data reveals hidden subpopulations of cells. *Nat. Biotechnol.*, **33**, 155.
2. Butler,A., Hoffman,P., Smibert,P., Papalexi,E. and Satija,R. (2018) Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.*, **36**, 411.
3. Stuart,T., Butler,A., Hoffman,P., Hafemeister,C., Papalexi,E., Mauck,W.M. III, Hao,Y., Stoeckius,M., Smibert,P. and Satija,R. (2019) Comprehensive integration of single-cell data. *Cell*, **177**, 1888–1902.
4. Kiselev,V.Y., Kirschner,K., Schaub,M.T., Andrews,T., Yiu,A., Chandra,T., Natarajan,K.N., Reik,W., Barahona,M., Green,A.R. *et al.* (2017) Sc3: consensus clustering of single-cell rna-seq data. *Nat. Methods*, **14**, 483.
5. Wang,B., Zhu,J.J., Pierson,E., Ramazzotti,D. and Batzoglou,S. (2017) Visualization and analysis of single-cell rna-seq data by kernel-based similarity learning. *Nat. Methods*, **14**, 414.
6. Grun,D., Lyubimova,A., Kester,L., Wiebrands,K., Basak,O., Sasaki,N., Clevers,H. and van Oudenaarden,A. (2015) Single-cell messenger rna sequencing reveals rare intestinal cell types. *Nature*, **525**, 251.
7. Peng,M., Li,Y., Wamsley,B., Wei,Y. and Roeder,K. (2021) Integration and transfer learning of single-cell transcriptomes via cFIT. *Proc. Natl. Acad. Sci.*, **118**, e2024383118.
8. Patterson,N., Price,A.L. and Reich,D. (2006) Population structure and eigenanalysis. *PLoS Genet.*, **2**, e190.
9. Zeisel,A., Muñoz-Manchado,A.B., Codeluppi,S., Lönnerberg,P., La Manno,G., Juréus,A., Marques,S., Munguba,H., He,L., Betsholtz,C. *et al.* (2015) Cell types in the mouse cortex and hippocampus revealed by single-cell rna-seq. *Science*, **347**, 1138–1142.
10. Baron,M., Veres,A., Wolock,S.L., Faust,A.L., Gaujoux,R., Vetere,A., Ryu,J.H., Wagner,B.K., Shen-Orr,S. S., Klein,A.M. *et al.*

(2016) A single-cell transcriptomic map of the human and mouse pancreas reveals inter-and intra-cell population structure. *Cell Syst.*, **3**, 346–360.

11. Wilson,N.K., Kent,D.G., Buettner,F., Shehata,M., Macaulay,I.C., Calero-Nieto,F.J., Castillo,M.S., Oedekoven,C.A., Diamanti,E., Schulte,R. *et al.* (2015) Combined single-cell functional and gene expression analysis resolves heterogeneity within stem cell populations. *Cell Stem Cell*, **16**, 712–724.
12. Zurauskiene,J. and Yau,C. (2016) pcareduce: hierarchical clustering of single cell transcriptional profiles. *BMC Bioinformatics*, **17**, 140.
13. Kim,J., Stanescu,D.E. and Won,K.J. (2018) Cellbic: bimodality-based top-down clustering of single-cell rna sequencing data reveals hierarchical structure of the cell type. *Nucleic Acids Res.*, **46**, e124.
14. Zappia,L. and Oshlack,A. (2018) Clustering trees: a visualization for evaluating clusterings at multiple resolutions. *GigaScience*, **7**, giy083.
15. Wang,Y.J., Schug,J., Won,K.J., Liu,C., Naji,A., Avrahami,D., Golson,M.L. and Kaestner,K.H. (2016) Single-cell transcriptomics of the human endocrine pancreas. *Diabetes*, **65**, 3028–3038.
16. Lab,S. (2019) panc8.SeuratData: Eight Pancreas Datasets Across Five Technologies. R package version 3.0.2, https://github.com/satijalab/seurat-data.
17. Polioudakis,D., de la Torre-Ubieta,L., Langerman,J., Elkins,A.G., Shi,X., Stein,J.L., Vuong,C.K., Nichterwitz,S., Gevorgian,M., Opland,C.K. *et al.* (2019) A single-cell transcriptomic atlas of human neocortical development during mid-gestation. *Neuron*, **103**, 785–801.
18. Zhang,J.M., Fan,J., Fan,H.C., Rosenfeld,D. and David,N.T. (2018) An interpretable framework for clustering single-cell rna-seq datasets. *BMC Bioinformatics*, **19**, 93.
19. Hubert,L. and Arabie,P. (1985) Comparing partitions. *J. Classif.*, **2**, 193–218.
20. Von,Luxburg U. (2009) Clustering stability: an overview. *Found. Trends Mach. Learn.*, **2**, 235–274.
21. Zhang,X.W., Xu,C.L. and Yosef,N. (2019) Simulating multiple faceted variability in single cell rna sequencing. *Nat. Commun.*, **10**, 2611.
22. Zhu,L., Lei,J., Klei,L., Devlin,B. and Roeder,K. (2019) Semisoft clustering of single-cell data. *Proc. Natl. Acad. Sci.*, **116**, 466–471.
23. Wu,W., Liu,Z. and Ma,X. (2021) jsrc: a flexible and accurate joint learning algorithm for clustering of single-cell rna-sequencing data. *Brief. Bioinform*, bbaa433.