

1 Cell Type Hierarchy Reconstruction via Reconciliation of 2 Multi-resolution Cluster Tree

3 Minshi Peng¹, Brie Wamsley², Andrew Elkins², Daniel M Geschwind^{2,3}, Yuting Wei¹, and Kathryn
4 Roeder^{1,4,*}

5 ¹Department of Statistics and Data Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA.

6 ²Program in Neurogenetics, Department of Neurology, David Geffen School of Medicine, University of California
7 Los Angeles, Los Angeles, CA, USA.

8 ³Program in Neurobehavioral Genetics and Center for Autism Research and Treatment Semel Institute and
9 Department of Human Genetics, David Geffen School of Medicine, University of California Los Angeles, Los
10 Angeles, CA, USA.

11 ⁴Computational Biology Department, Carnegie Mellon University, Pittsburgh, PA 15213, USA

12 *Corresponding author

13 **Abstract**

14 A wealth of clustering algorithms are available for Single-cell RNA sequencing (scRNA-
15 seq), but it remains challenging to compare and characterize the features across different scales
16 of resolution. To resolve this challenge Multi-resolution Reconciled Tree (MRtree), builds a
17 hierarchical tree structure based on multi-resolution partitions that is highly flexible and can
18 be coupled with most scRNA-seq clustering algorithms. MRtree out-performs bottom-up or
19 divisive hierarchical clustering approaches because it inherits the robustness and versatility of
20 a flat clustering approach, while maintaining the hierarchical structure of cells. Application to
21 fetal brain cells yields insight into subtypes of cells that can be reliably estimated.

22 **Keywords**

23 Single Cell RNA-seq, Clustering, Cluster hierarchy, Cell type identification, Brain cells

24 Background

25 Single-cell RNA sequencing (scRNA-seq) is a recently developed technology that is being widely
26 deployed to collect unprecedented catalogues detailing the transcriptomes of individual cells. The
27 ability to capture the molecular heterogeneity of tissues at high resolution underlies its increasing
28 popularity in discovering cellular and molecular underpinnings of complex and rare cell popula-
29 tions², developing a “parts list” for complex tissues^{13,20}, and studying various diseases and cell
30 development or lineage. One essential component involves the utility of scRNA-seq data to enable
31 the identification of functionally distinct subpopulations that each possess a different pattern of
32 gene expression activity. These sub-populations can indicate different cell types with relatively sta-
33 ble, static behavior, or cell states in intermediate stages of a transient process. Unbiased discovery
34 of cell types from scRNA-seq data can be automated using a wealth of unsupervised clustering
35 algorithms, among them the most widely-applied include the Louvain graph-based algorithm incor-
36 porated as part of Seurat^{3,10} pipeline, k-means clustering and its derivatives by consensus clustering
37 as performed in SC3⁷ and SIMLR¹², and other methods that address issues caused by rare types
38 such as RaceID⁴.

39 A major challenge regarding clustering algorithms mentioned above is that they explicitly or
40 implicitly require the number of clusters to be supplied as an input parameter. Determining the
41 number of cell types in the population presents a significant challenge given the large number of
42 distinct cell types, which is further complicated by substantial biological and technical variation.
43 There are some computational methods available to guide the choice of K ; however, these methods
44 are shown to be flawed in one aspect or another. For example, methods developed on the basis
45 of statistical testing are shown to be too sensitive to heterogeneity, especially for large samples,
46 while other methods tend to favor a fairly coarse resolution, with clearly separated clusters and
47 fail to identify closely related and overlapping cell types. Therefore, judgment from the researchers
48 is required to choose the desired resolution. A common practice in scRNA-seq data analysis is to
49 run a clustering algorithm repeatedly for a range of resolutions, followed by careful inspections of
50 individual results by examining the cluster compositions and the expression of published marker
51 genes to select the final partition. This supervised process takes significant time and effort and is

52 limited by the current state of the investigator's/field's knowledge about cell type and cell state
53 diversity. It would be a substantial advantage in terms of efficiency and veracity to be able to reach
54 the same level of resolution in an unsupervised manner.

55 Hierarchical clustering (HC) is another popular general-purpose clustering method commonly
56 used to identify cell-populations^{18,1,16}. HC has the advantage of being able to determine relation-
57 ships between clusters of different granularities since the result can be visualized as a dendrogram.
58 This dendrogram is then “cut” at different heights to generate different numbers of clusters. This
59 hierarchical structure helps identify multiple levels of functional specialization of cells. For instance,
60 neurons share specific functional characteristics distinct from those of various glial cell types and
61 contain distinct subtypes with more specialized functions, such as excitatory or inhibitory proper-
62 ties. Different variants of hierarchical clustering make different assumptions, the most common ones
63 used in classical hierarchical agglomerative clustering (HAC) is Ward's¹⁵ and “average” linkage as
64 adopted in SC3⁷. An important limitation of HAC is that both time and memory requirements
65 scale at least quadratically with the number of data points, which is slower than many flat cluster-
66 ing methods like K-means and prohibitively expensive for large data sets. A few scRNA-seq tools
67 expand upon the idea of hierarchical clustering, for instance, pcaReduce²² introduces an agglom-
68 erative clustering approach by conducting dimension reduction after each merge, starting from an
69 initial clustering, and CellBIC⁶ performs bisecting clustering in a top-down manner leveraging the
70 bi-modal gene expression patterns. However, these methods either require a good initial start that
71 is implicitly equivalent to the choice of K , or are highly dependent on the assumption of bimodal
72 expression pattern at each iteration, which is not appropriate for multi-modal data.

73 In this study, we build a tool to bridge the gap between two separate lines of inquiry, flat and
74 hierarchical clustering. Empirically, scRNA-seq data analysts observe that the partitions obtained
75 from flat clustering at multiple resolutions, when ordered by increasing resolution, produce a lay-
76 ered structure with a tree-style backbone¹⁷. This produces a useful representation to help visually
77 determine the stability of clusters and relations among them. We build on this idea and propose
78 a method called Multi-resolution Reconciled Tree (MRtree) that reconstructs the underlying tree
79 structure by reconciling partitions obtained at different granularities (Figure 1) to produce a co-

80 herent hierarchy that is as similar as possible to the original flat clustering at different scales. It
81 can work with many specially-designed flat clustering algorithms for single-cell data, such as Lou-
82 vain clustering from Seurat³, thus inheriting the scalability and good performance in clustering
83 the single-cell data; meanwhile, it recovers the intrinsic hierarchy structure determined by the cell
84 types and cell states.

85 Applications of MRtree on a variety of scRNA-seq data sets, including mouse brain¹⁸, human
86 pancreas^{14,8} and human fetal brain⁹, showed improved performance for clustering of scRNA-seq
87 data over initial flat clustering methods. The hierarchical structure discovered by MRtree easily
88 outperformed a variety of tree-construction methods. Moreover, the results accurately reflect the
89 extent of transcriptional distinctions among cell groups and align well with levels of functional
90 specializations among cells. Particularly, when applied to developing human brain cells, the method
91 successfully identified major cell types and recovered an underlying hierarchical structure that is
92 highly consistent with the results from the original study⁹. Subsequent analysis on each major
93 type via MRtree revealed finer sub-structure defining biologically plausible subtypes, determined
94 mainly by maturation states, spatial location, and terminal specification.

95 Results

96 Methods overview

97 MRtree aims to recover a hierarchical tree by denoising and integrating a series of flat clusterings
98 into a coherent tree structure. The algorithm starts by applying a suitable flat clustering algorithm
99 to obtain partitions for a range of resolution parameters. The multi-scale results can be represented
100 using a multi-partite graph, referred to as a *cluster tree*, where the nodes represent clusters, and
101 edges between partitions of adjacent resolutions indicate common cells shared. We propose an
102 efficient optimization procedure to reconcile the incoherent cell assignments across resolutions, that
103 produces the optimal underlying tree structure following the hierarchy constraints, while adhering
104 to the initial flat clustering to the maximum extent. Formally, this is achieved by minimizing
105 (among valid hierarchical tree structures), the difference between initial multi-level cluster assign-

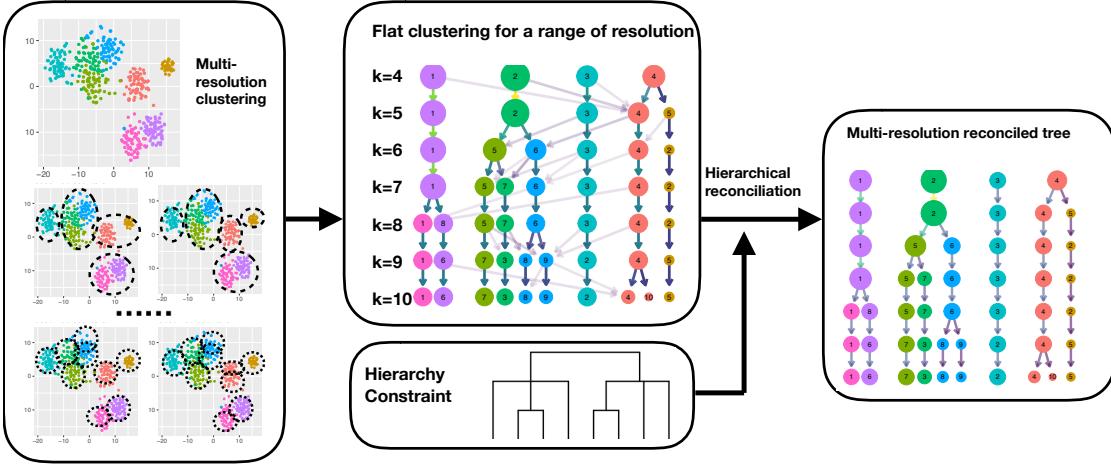


Figure 1: Overview of MRtree framework. The algorithm starts by performing flat clustering on scRNA-seq data for a range of resolutions, where the partitions between adjacent resolutions are matched to form a graph as an entangled cluster tree. Then reconciliation is performed through optimization with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

106 ments and the cluster assignments in the resulting tree structure. By representing the partitions
 107 as a multi-partite graph, the clustering assignments that violate the hierarchy constraint can be
 108 identified as merging directed edges and thus penalized in the objective function. The optimization
 109 procedure proceeds by iteratively and greedily identifying those tree nodes, which, when corrected
 110 by reassigning the associated conflicting cell lineages, contribute to maximum descent in the de-
 111 fined objective function. The outcome of the proposed optimization procedure is a reconciled tree,
 112 named the *hierarchical cluster tree*, representing the optimal tree-based cluster arrangement across
 113 scales (Figure 1, Figure S1, Supplemental Information).

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

121 Another key distinguishing feature of our procedure, compared to existing consensus methods,

122 is that we build a cluster hierarchy directly from the flat partitions in an “in place” way. This
123 is in comparison to existing methods for which an alternative hierarchical clustering algorithm is
124 applied to the co-classification consensus matrix built from an ensemble of partitions. MRtree uses
125 an optimization framework to edit the original partitions through a similar voting scheme. At
126 the same time, it aims to preserve the original splitting order of the hierarchy determined by the
127 clustering algorithm. The proposed method is efficient in terms of memory cost and time complexity
128 (Supplemental Information). Moreover, MRtree enables a direct comparison of partitions before
129 and after tree reconciliation, to examine the stability of the clustering algorithm at different scales.
130 As a benefit, we are able to trim the tree to the maximum depth within the stable range to obtain
131 reliable final clusters.

132 To summarize, we present a computationally efficient method to generate a hierarchical tree
133 from flat clustering results for a range of resolutions, by an iterative greedy optimization scheme.
134 It enables us to take advantage of various flat clustering approaches, while improving over these
135 flat clustering results by averaging over membership assignments across resolutions. The resulting
136 hierarchical structure captures the relations between cell types and at the same time helps circum-
137 vent the problems of choosing the optimal resolution parameter. It turns out that the proposed
138 method improves the clustering accuracy over the initial partition across scales, and outperforms
139 a variety of alternative tree construction methods for recovering the underlying tree structure. To
140 facilitate identifying stable tree layers, we propose a stability measure that compares the initial flat
141 clustering with the reconciled tree. We also implement tools to sample implicit resolution param-
142 eters for Seurat clustering that enable equal coverage of different clustering granularities. Finally,
143 our optimization procedure is made efficient for potential use in big data analyses.

144 Simulation study

145 **Simulated data** To evaluate how well MRtree is able to recover the cluster hierarchy and improve
146 the clustering across resolutions, we harness the tools provided by the SymSim package¹⁹ to simulate
147 scRNA-seq data given a known tree structure, using the SymSim parameters estimated from a UMI-
148 based dataset of 3,005 mouse cortex cells¹⁸ (Supplemental Information). Motivated by major cell

149 types identified in brain tissues, we constructed a hypothetical tree (Figure 2A,B) as the ground
150 truth representing the hierarchy of the cell types/states.

151 Repeated simulations were performed by first generating single-cell data with SymSim from the
152 hypothetical tree structure, followed by multi-resolution flat clustering using a variety of clustering
153 methods. Then MRtree was applied to form the hierarchical cluster tree that reconciled the multi-
154 level clusterings. MRtree can be coupled with most flat clustering methods; hence we evaluated the
155 performance using a variety of algorithms, including Seurat³, SC3⁷, SOUP²¹, and K-means applied
156 to a UMAP projection. The clustering results were evaluated and compared with the raw clusters
157 obtained from flat clustering in three aspects: the accuracy of clustering regarding label assignments
158 at different resolutions, the tree structure estimation accuracy, and the clustering stability.

159 We first sought to quantify how well MRtree performs regarding clustering accuracy, measured
160 using Adjusted Multiresolution Rand Index (AMRI, Methods) between the obtained labels and
161 true labels known from the simulation. An AMRI close to 1 indicates perfect clustering given the
162 resolution. MRtree achieved higher accuracy almost uniformly across resolutions and for various
163 clustering methods (Figures 2C and S2). It is worth noticing that the reconciliation procedure
164 even improved upon SC3 results, which already employed an ensemble-based method for each fixed
165 resolution. This demonstrates that applying an ensemble approach across resolutions captures
166 additional structural information within the data. In addition, the gain was more pronounced for
167 coarse clustering and when there was more room for improvement.

168 Next, we evaluated the ability of MRtree to recover the tree structure. For comparison, we
169 leveraged the tools that build hierarchical trees in Seurat and SC3. For Seurat, an agglomerative
170 hierarchical cluster tree was built starting with the identified Seurat clusters, while for SC3, a full
171 HAC was performed from the consensus similarity matrix constructed by aggregating clustering
172 results with different dimension reduction schemes. MRtree produced a significantly improved tree
173 structure estimation compared to the competing methods, as demonstrated by the reduced error
174 of tree reconstruction (Figure 2D, Methods).

175 Finally we evaluate the clustering stability before and after tree reconciliation, coupled with
176 multiple clustering methods. The stability score is calculated following the subsample procedure

177 described in Methods. For K less than the true number of clusters, the measured stability is
178 confounded by the instability induced by the incorrect resolution. Therefore we restrict our com-
179 parison to the measured stability at the true resolution. Clustering stability with MRtree is clearly
180 improved compared to the initial clustering across all methods (Figure S3), demonstrating the im-
181 proved robustness of MRtree, which successfully employs the consensus mechanism to denoise the
182 individual clustering with collective information across resolutions.

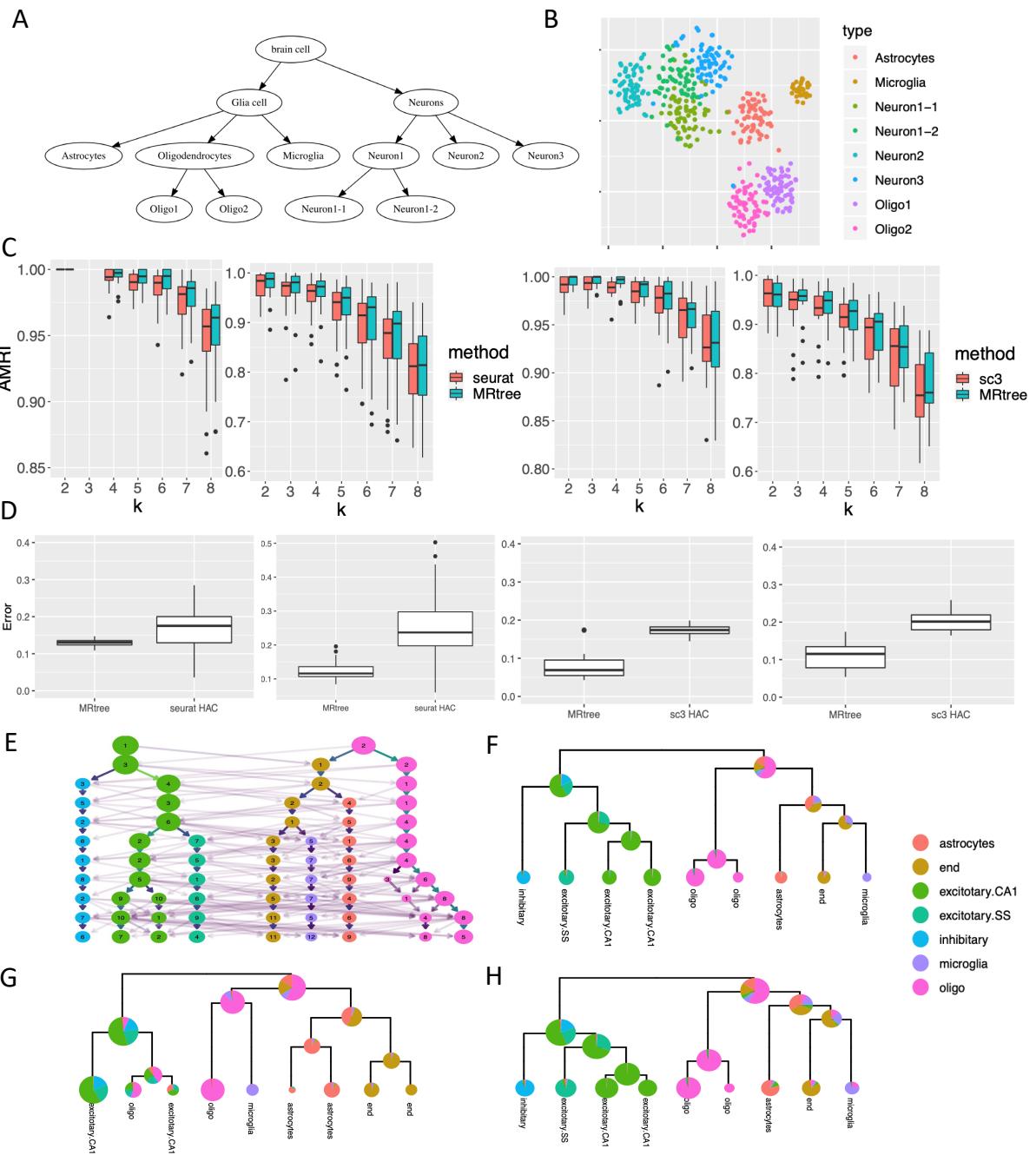


Figure 2: Evaluate the performance of MRtree via simulations and analysis on mouse brain data ¹⁸. (A) The hypothetical tree structure of the cell states from which cells are generated. (B) tSNE plot of the simulated cells in one experiment, colored by the cell types indicated in the leaf of the actual hierarchical cluster tree. The difficulty of the simulation varies from simple (left, smaller within-cluster noise) to challenging (right, stronger noise) in panels (C) and (D) for each method. (C) Comparing the accuracy of MRtree clusters with the clusters from initial flat clustering at multiple resolutions using Seurat and SC3. The accuracy is measured by the Adjusted Multi-resolution Rand Index (AMRI). (D) Evaluate tree construction accuracy of MRtree with dendrogram from hierarchical clustering obtained with Seurat and SC3. (E-H) MRtree applied to scRNA-seq data from the mouse brain. (E) Initial flat clustering by SOUP on 3,005 cells¹⁸ by varying K , the resolution parameter specifying the number of clusters, colored by the gold standard labels. (F) The MRtree-constructed tree from initial SOUP clusterings. The pie charts on tree nodes represent the cell type composition referencing the gold standard. (G, H) Comparing tree construction and clustering accuracy on mouse brain data using different methods, hierarchical cluster tree generated by HAC starting with SOUP clusters (G) and starting with individual cells (H).

184 scRNA-seq data

185 **Mouse brain cells** We illustrate MRtree using a scRNA-seq data set containing 3,005 cells of
186 somatosensory cortex and hippocampal-CA1 region from mice, collected between postnatal 21-31
187 days. We call this the mouse brain data¹⁸. The authors have assigned the cells to seven major
188 types: pyramidal CA1, pyramidal SS, interneurons, astrocytes-ependymal, microglia, endothelial-
189 mural, and oligodendrocytes. For comparison, these labels are treated as the gold standard in the
190 following analysis.

191 We chose SOUP²¹ for multi-resolution clustering due to its superior performance on these data.
192 The hard clustering labels were obtained by varying K , the resolution parameter specifying the
193 desired number of clusters, from 2 to 12. With MRtree, we were able to construct a hierarchical
194 cluster tree from the flat sequential clusterings. The initial cluster tree is visualized with nodes
195 colored by the major type referencing the gold standard, and the recovered tree from MRtree is
196 shown on the right, with the proportion of cell types in each node visualized by a pie chart (Figure
197 2E,F). The tree successfully split the neurons and glial cells at an early stage, followed by splitting
198 pyramidal cells from two regions (CA1,SS) from the interneurons. Finally, cells from the same type
199 but distinct brain regions were identified. The tree reconciliation step also improved the clustering
200 performance by increasing the accuracy measured by AMRI in multiple layers (Figure S4a).

201 To further compare the performance of MRtree with HAC, we applied HAC using complete link-
202 age on the first 20 principal components, starting either from singletons (individual cell) or the 9
203 SOUP clusters obtained at the maximum resolution (Figure 2G,H; only the top layers of HAC from
204 singleton are shown for comparison purposes). Compared to MRtree, HAC shared a similar overall
205 tree structure, but it generated clusters at lower accuracy for each layer. The results support the
206 argument that MRtree is able to improve accuracy upon initial clustering by pooling information
207 across resolutions. HAC from singletons performed much worse regarding both accuracy and tree
208 structure, possibly owing to the sensitivity of HAC to outliers and linkage selection. For complete-
209 ness, we also demonstrate the accuracy from two widely applied clustering methods, Seurat and
210 SC3, where the HAC results were generated from the built-in functions provided as part of the
211 toolkits. In both cases, MRtree outperformed both the initial flat clustering and the HAC (Figure
212 S4b,c).

213 In addition to improving the clustering accuracy, we were able to infer the resolution that achieved
214 the highest stability by inspecting the difference between the initial tree and the reconstructed tree.
215 It indicated that both the SOUP and Seurat algorithms should stop splitting at $K = 7$, which was
216 consistent with the gold standard (Figure S6). Stability analysis on SC3 results showed a preferred
217 resolution of 6 clusters. Indeed, we observed steep drop in accuracy for any resolution greater
218 than 6 (Figure S4c). By comparison, using available K -selection methods supported in multiple
219 single-cell analysis pipelines, the optimal number of clusters selected varied widely (Table S1). For
220 instance, SC3 supported 22 clusters. In addition, the large gap between MRtree and initial Seurat
221 clusterings indicated the inability of Seurat to identify accurate and stable clusters on this dataset.
222 This observation was further supported by the lower accuracy (AMRI < 0.6) of the resulting Seurat
223 clusters (Figure S4b).

224 **Human pancreas islet cells** To evaluate performance on cell types that are fairly well separated,
225 we investigated the hierarchical structure identified by MRtree for cells from human pancreatic
226 tissues. We first analyzed single-cell RNA sequencing of 635 cells on islets from Wang et al.¹⁴,
227 which come from multiple donors, including children, control adults, and individuals with type 1 or
228 type2 diabetes (T1D, T2D). Among them, 430 cells were annotated by the authors into seven cell

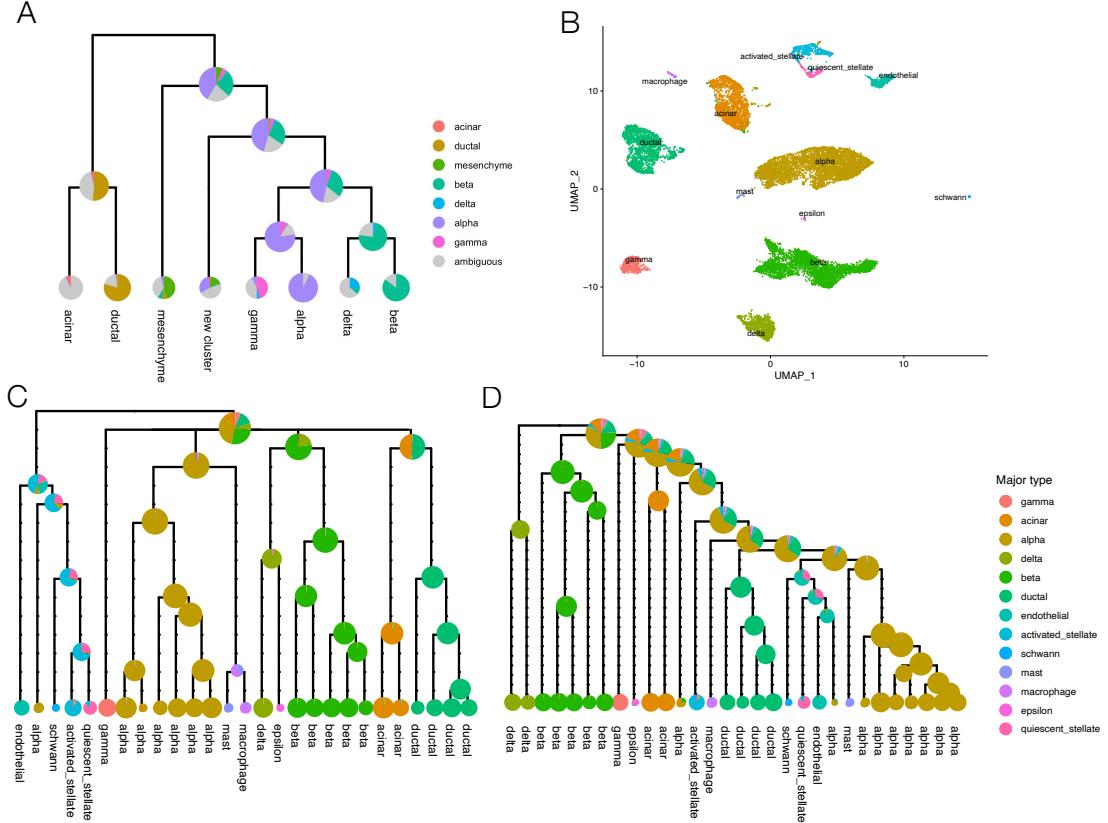


Figure 3: MRtree applied on pancreas islet cells data sets reveals the transcriptional distinctions and similarities between cell types. (A) MRtree-constructed tree with SC3 clusterings on 635 cells from Wang et al.¹⁴. The tree was trimmed to the layer with eight leaf clusters ($K = 8$). The pie charts 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.

229 types, while 205 cells were considered ambiguous and unlabeled. We applied MRtree to construct
230 the hierarchical cluster tree based on SC3 flat clustering with the number of clusters ranging
231 from 2 to 15. The tree was then trimmed to eight leaf nodes based on stability analysis (Figure
232 3A, Figure S7A). The first split created two large interpretable cell groups: gene ontology (GO)
233 shows enrichment of exocrine functions such as terms related to “Putrescine catabolic process”
234 (adjusted p-value=2.3E – 02) and “Cobalamin metabolic process” (adjusted p-value=5.48E – 05)
235 for the left branch, and enrichment of endocrine functions such as “Insulin secretion” (adjusted
236 p-value=3.4E – 5) and “Enterendoctrine cell differentiation” (adjusted p-value=2.1E – 2) for the
237 right branch. The exocrine group was further divided into acinar (*PRSS1*) and ductal cells (*SPP1*).
238 The right branch further separates a previously undiscovered cluster composed mainly of ambiguous
239 cells and a few previously labeled alpha and mesenchyme cells. This cluster expresses marker genes
240 with significant GO terms such as “Collagen metabolic process” and “regulation of endothelial cell
241 migration”, pointing to endothelial and stellate cells (Table S2) that were not labeled in the original
242 analysis. The remaining endocrine cells were further divided into a group containing α cells (*GCG*)
243 and pancreatic polypeptide cells (*PPY*), and another group containing β (*INS*) and δ cells (*RBP4*)
244 (Table S3).

245 In addition to recapitulating a logical tree for all cell types, the eight clusters improved upon the
246 initial SC3 clusters. In particular, seven of the clusters match well with the identified seven major
247 cell types from Wang et al., achieving AMRI greater than 0.95 (Figure S7B-D). By contrast, a
248 competing tree construction method, CellBIC⁶, revealed a similar tree structure, but it failed to
249 identify the group of δ cells⁶. Finally, because it is well accepted that β cells are heterogeneous,
250 especially in conditions of metabolic stress, such as obesity or type 2 diabetes¹⁴, we further applied
251 MRtree on the subset of 111 β cells. We obtained five β subclusters that corresponded to key
252 biological features, including two clusters composed mainly of cells from T2D individuals, and one
253 control group containing 90% cells from children (Figure S7E-G).

254 Next, we considered a more challenging data set, again from the human pancreatic islet, produced
255 by merging data from five technologies⁸. In total, 14,892 cells were annotated and grouped by
256 respective studies into 13 major cell-types with cluster sizes varying by magnitude. We first in-

257 integrated the cells using Seurat MNN integration tools using 2,000 highly expressed genes (Figure
258 3B). Despite the observation that SC3 demonstrates superior performance on the smaller data
259 sets, we utilized Seurat graph-based clustering because it demonstrates greater scalability to large-
260 scale analysis. Flat clusterings were obtained for 50 different resolution parameters sampled via
261 Event-Sampling in the range of [0.001, 2]. The resulting tree identified all 13 major types with high
262 accuracy and also uncovered many subtypes organized as subtrees (Figure 3C). Very distinct cell
263 types separated early and fall into remote branches, while cell types that share similar functions
264 share internal branches and split later in the process. For instance, endothelial, schwann, and
265 stellate cells are very different from other endocrine and exocrine cells and thus split out first. Two
266 types of endocrine cells, acinar and ductal, fall into a common subtree. Likewise, five types of
267 exocrine cells are organized in the same subtree. Finally, subtypes from the same major type are
268 organized in the same subtree, with one exception. A small subset of α cells was inappropriately
269 placed in the tree. However, evidence suggests these cells represent an anomaly, possibly due to
270 batch correction. These α outliers appear in the UMAP projection separated from other α cells
271 and near the activated stellate cells.

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

280 **Human fetal brain cells** We applied MRtree to cells from the mid-gestational human cortex,
281 which we call the human brain data⁹. These data were derived from \sim 40,000 cells from germinal
282 zones (ventricular zone, subventricular zone) and developing cortex (subplate (SP) and cortical
283 plate (CP)) separated before single-cell isolation. By performing Seurat clustering³, the authors
284 assigned the cells into 16 transcriptionally distinct cell groups (Table S4). For convenience, here

285 we refer to these expert classifications as the Polioudakis labels.

286 Our analysis began with the same preprocessing steps as conducted in the study⁹ using the pipeline
287 supported by Seurat V3. The multilevel clustering results are visualized by increasing resolution
288 from the top layer (resolution=0.001) to bottom layer (resolution=2), where each layer corresponds
289 to one clustering (Figure S8A). Notably there were a considerable number of cells assigned to clus-
290 ters inconsistently over changing resolutions, which made it challenging to determine the optimal
291 resolution and the final cluster memberships. By applying MRtree, we were able to construct the
292 organized hierarchical tree, which was represented by a dendrogram with the cell-type composition
293 of clusters referencing Polioudakis labels shown by pie charts on tree nodes (Figure 4A). MRtree
294 first separated interneurons and pericytes, endothelial and microglia, followed by splitting excita-
295 tory deep layer neurons from radial glia to maturing excitatory neurons, representing the rest of a
296 closely connected lineage (upper layer enriched). By further increasing the resolution, the radial glia
297 cells and excitatory neurons were isolated, where the intermediate progenitors were more closely
298 connected with maturing excitatory neurons. The finer distinctions of excitatory neurons were
299 subsequently identified as migrating, maturing, and maturing upper enriched subtypes supported
300 by differential gene expression and canonical cell markers (Table S5). The results were consistent
301 with the group-wise separability visible through a 2-dimensional tSNE projection (Figure 4B). The
302 cluster stability was inspected by comparing the initial Seurat clusters at each resolution with the
303 MRtree results (Figure S8B), which suggested that the clusters were stable up to around $K = 15$.
304 We decided to cut the tree at $K = 13$, which corresponded fairly closely to the 16 major gold
305 standard cell types of the midgestational brain by examination of differentially-expressed marker
306 genes (Table S5, Figure 4C,D). For comparison, an agglomerative hierarchical tree was generated
307 starting from the Seurat clusters obtained at the highest resolution (Figure S8C). These results
308 were distance-based and consequently more vulnerable to outliers, which appear to have caused
309 several anomalies: subsets of ExM, ExM-U, and ExDp1 were grouped together, and two subsets of
310 IP were separated from each other.

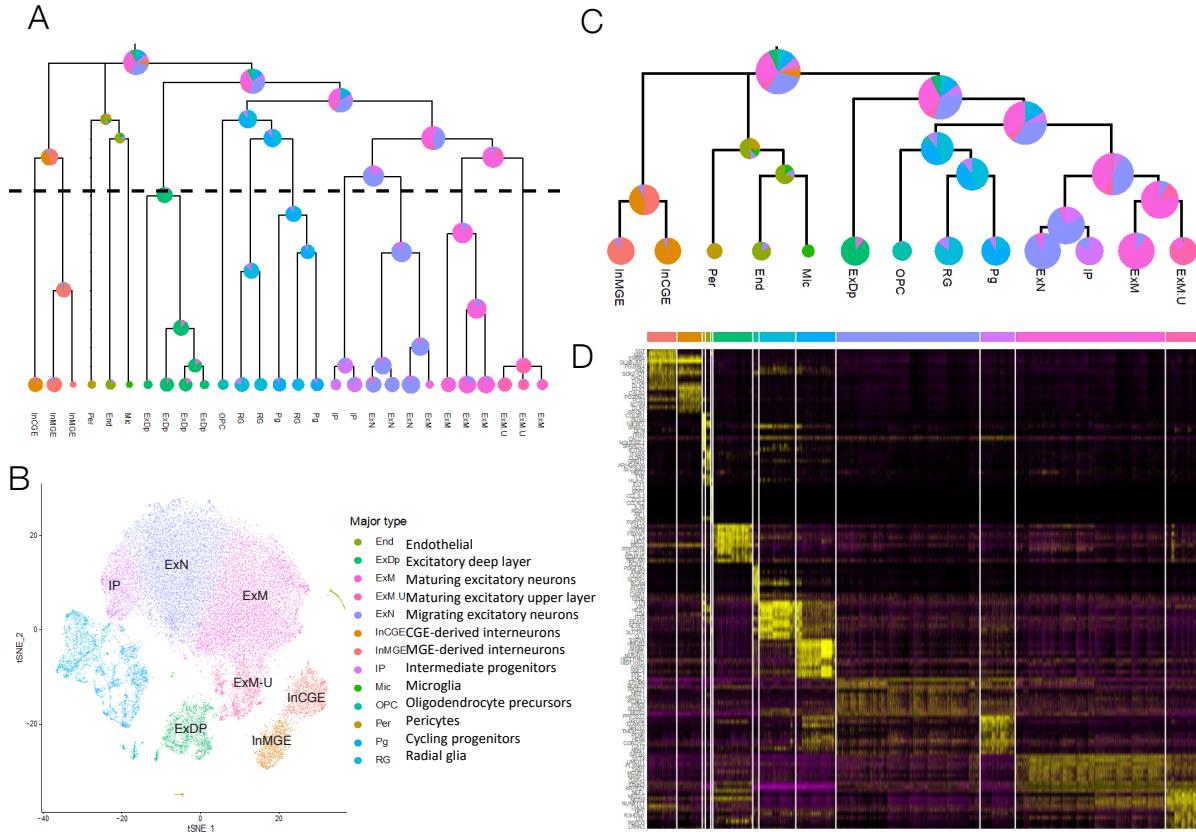


Figure 4: MRtree applied to scRNA-seq data from human brain cells. (A) MRtree produces the hierarchical cluster tree from the initial flat clusterings at multiple resolutions obtained from Seurat V3. The nodes correspond to clusters, with a pie chart displaying the cluster composition referencing the 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 (B), 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.

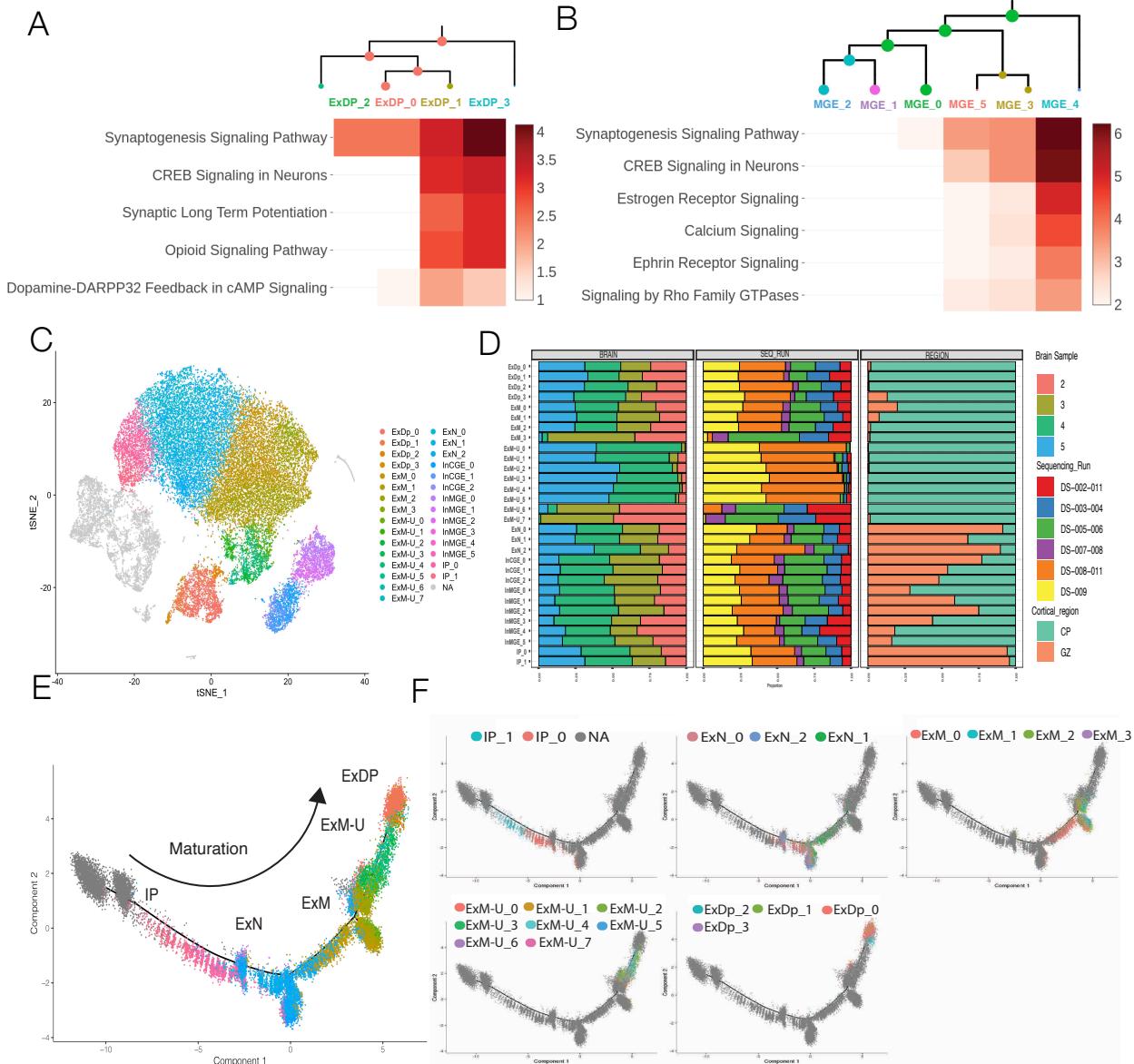


Figure 5: MRtree clusters cells into known cell subtypes and states that underlie known cellular developmental transcriptional trajectories at a higher resolution. (A) Hierarchical cluster tree of subplate/ deep layer excitatory neurons (ExDp) with a heatmap of gene expression within canonical gene ontology 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.

312 **Identify subtypes** Next, we scrutinized the fine-grained structure by re-clustering the 13 major
313 cell types obtained from the hierarchical cluster tree of all cells. The cells were pre-processed from
314 the raw count data as performed in the first iteration, followed by clustering using the Seurat
315 graph-based method. By setting the resolution parameters from 0.05 to 1 and applying MRtree,
316 we obtained one hierarchical tree for each major cell type, determined by trimming the full tree
317 to the stable top layers (ExDP and InMGE are depicted in Figure 5A,B). This resulted in 21
318 transcriptionally distinct cell types from 7 of the identified major types, expanding IP, ExN, ExM,
319 ExM-U, ExDp, InCGE, and InMGE (Figure 5C, Table S8). The subtypes' partitions were first
320 evaluated by assessing whether the likely technical and biological co-variation, including brain
321 sample, sequencing run, and cortical region, illustrated somewhat even distribution and appropriate
322 overlap within each identified cluster. Results show that the clusterings were not driven by these
323 technical features and are likely biologically meaningful (Figure 5D, Figure S9).

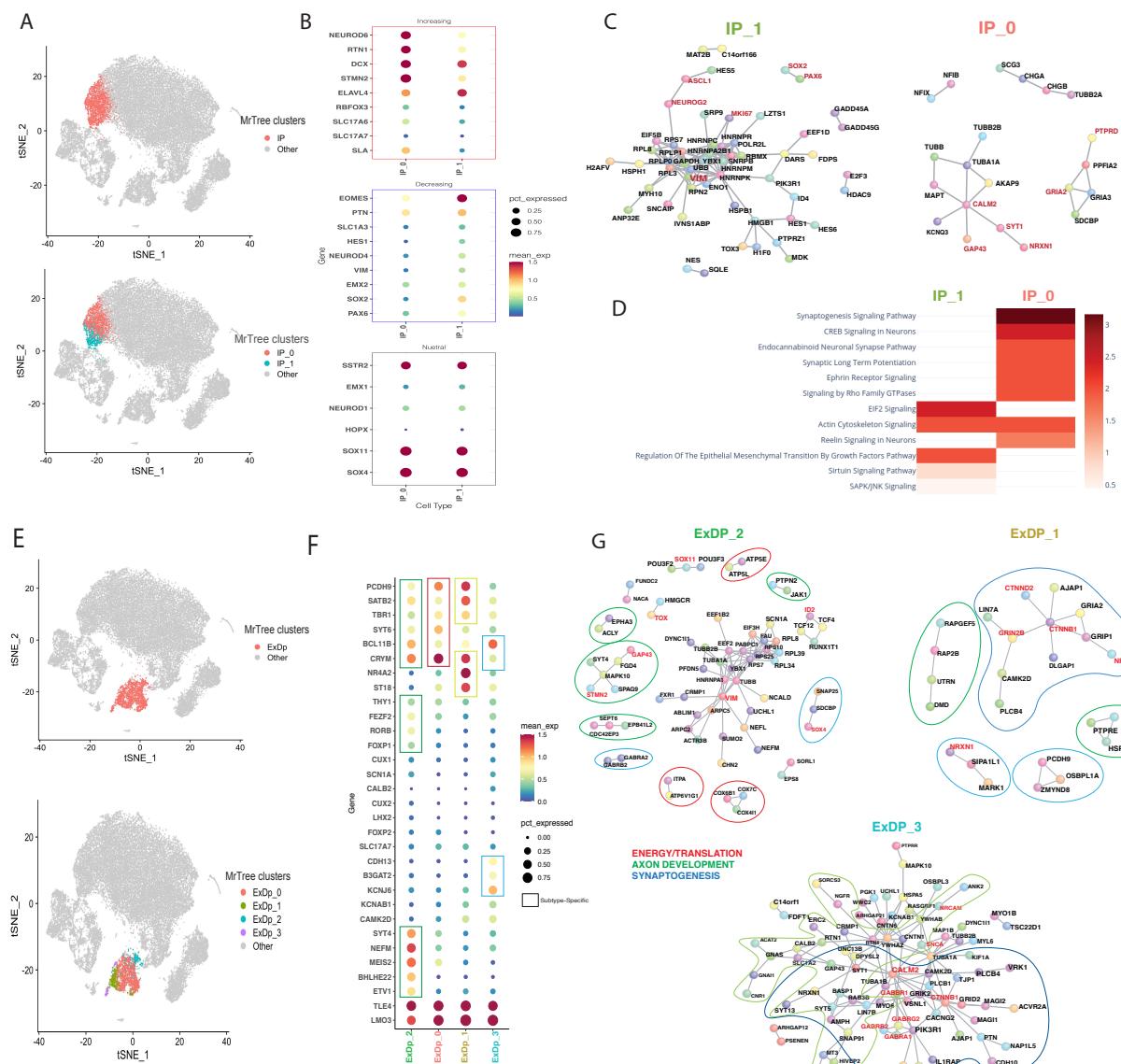


Figure 6: Known and unique biological states identified by MRtree with sub clustering on human fetal brain data: intermediate progenitors and subplate/ deep excitatory neurons (A) top: tSNE plot of all cells where intermediate progenitor (IP) cells identified by MRtree are colored by red, bottom: tSNE 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*, *NEUROD6*), intermediate progenitors (i.e. *EOMES*, *SOX11*, *SOX4*, *PTN*), and radial glia (i.e. *SLC1A3*, *VIM*, *SOX2*, *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*, *FEZF2*), Layer 6 (i.e. *TBR1*, *SYT6*, *FOXP2*), shared deep markers (i.e. *RORB*, *TLE4*, *LMO3*, *CRYM*, *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.

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,

340 synaptic signaling (i.e., Synaptic long-term potentiation, Opioid, and Dopamine-DARPP32-cAMP
341 signaling) across maturation from ExDP_0 to the most mature ExDP_3 cluster. We observed simi-
342 lar functional specializations of inhibitory neuron subtypes (Figure 5B). The most mature subtype
343 InMGE_4 was discriminated from the other MGE interneurons first, followed by splitting the sec-
344 ond and third most mature subtypes from less mature cells, and finer distinctions were established
345 subsequently in two branches. The heatmap of gene expression within canonical gene ontology
346 categories showed a gradual increase in enrichment of Synaptogenesis, *CREB* signaling, Calcium
347 mediated signaling across maturation from InMGE_2 to most mature InMGE_4 cluster.

348 MRtree partitioned intermediate progenitor cells into two subtypes (IP_1 and IP_0; Figure 6A)
349 similar to cell types revealed in Polioudakis et al., achieved only after multiple rounds of analysis of
350 flat clustering results. Marker genes for newly born neurons (i.e. *SLA*, *STMN2*, *NEUROD6*) and
351 intermediate progenitors (i.e. *EOMES*, *SOX11*, *SOX4*, *PTN*) showed increased expression mark-
352 ers within IP_0 in contrast to expression of more intermediate progenitors and radial glia genes
353 within IP_1 (i.e. *SLC1A3*, *VIM*, *SOX2*, *HES1*) (Figure 6B). Notably, by comparing the significant
354 protein-protein interacting (PPI) networks from differential genes (DGE) expressed in IP_1 versus
355 significant PPI network from DGE in IP_0, we observed that IP_1 cells PPI contains a highly con-
356 nected radial glial genes surrounding *VIM* including *MKI67*, *SOX2*, for example, whereas, IP_0 cells
357 contain more neuronal-committed genes involving early step in neuronal differentiation including
358 *MAPT*, *GAP43*, *CALM2*, *GRIA2*, *PTPRD* (Figure 6C). Gene ontology analysis further uncov-
359 ered a switch in the enrichment of *EIF2* signaling, growth factors, and cell cycling pathways (i.e.
360 Sirtuin signaling pathway and *SAPK/JNK* signaling) in IP_1 to more specific neuronal categories
361 like Synaptogenesis, Ephrin Receptor signaling, Reelin signaling underlying migration and neurite
362 pathfinding signaling within IP_0 (Figure 6D).

363 For ExDP subtypes, a closer examination of the expression of marker genes for layer 5 (i.e., *ETV1*,
364 *RORB*, *FOXP1*, *FEZF2*), Layer 6 (i.e., *TBR1*, *SYT6*, *FOXP2*), shared deep markers (i.e., *RORB*,
365 *TLE4*, *LMO3*, *CRYM*, *THY1*) and subplate makers (i.e., *NR4A2*, *ST18*) showed expression of
366 Layer 5 markers within the least mature cells, ExDP_2, and layer 6 markers within ExDP_0, in
367 contrast to the more mature expression of layer 6-CTIP2 markers within ExDP_3 and mature

368 expression of markers of subplate and layer 6 within ExDP_1 (Figure 6F). Surprisingly, ExDP_2
369 PPI revealed a set of genes and structure similar to an intermediate progenitor with *VIM* at the
370 center of translational control and the expression of neuronally committed genes *SOX4*, *SOX11*,
371 *ID2* similar to IP_1, except that neuronal specificity genes within this cluster were linked directly
372 to upper Layer 5 cell fate (i.e., *FEZF2*, *FOXP1*, *RORB*, *SYT4*) instead of a general excitatory
373 neuronal lineage seen in IP_1. ExDP_1 subplate cells PPI exhibited a group of connected genes
374 related to more mature cellular properties such as synaptic plasticity and Wnt signaling (i.e.,
375 *GRIN2B*, *CTNNB1*, *NR2F1*, *NRXN1*) but no energy or translational pathways that were present in
376 ExDP_2. ExDP_3 cells showed the most extensive and unique PPI that illustrated more committed
377 axonal and synaptic pathways underlying specifically Layer 6 *CTIP2*+ cells (i.e., *CALM2*, *NRCAM*,
378 *SNCA*, GABAergic postsynaptic machinery) (Figure 6G).

379 Four other cell types revealed subtypes that were also related to developmental ordering. ExN
380 was partitioned into 3 subtypes that indicate a gradually increased expression of markers of upper
381 layer excitatory neurons in contrast to no expression of deep layer neuronal programs (Figure
382 S10). ExM was partitioned into 4 subtypes, 3 of which illustrate gradually increased expression
383 of upper layer markers, in contrast, a fourth that expressed deep layer markers indicating layer
384 4/5 excitatory neurons (Figure S11). InMGE was partitioned into 6 progressively more mature
385 subtypes (Fig 5B) that demonstrate distinctions in both maturation and terminal specification
386 (Fig S12). Finally, the 3 subtypes of InCGE display a general maturation of CGE interneurons
387 through a gradual decrease in expression of transcription factors along with a gradual increase in
388 expression of axonal-related genes (Figure S13). Meanwhile, although the signal was sparse, the
389 PPI network for the allegedly most mature subtype revealed a connection between genes critically
390 involved in post-synaptic glutamate signaling and plasticity, further supporting this conjecture.
391 Additional characteristics of these subtypes can be found in Supplemental Information.

392 Discussion and conclusion

393 In this article, we propose MRtree, a computational approach for characterizing multi-resolution
394 cell clusters ranging from major cell groupings to fine-level subtypes using a hierarchical tree. The

395 approach is based on deriving a multi-resolution reconciled tree to integrate clusterings obtained for
396 a range of different resolutions. The proposed method combines the flat and hierarchical clustering
397 results in a novel manner, inheriting the computational efficiency and scalability from the flat
398 clustering and the interpretability of a hierarchical structure. In comparison, MRtree outperforms
399 bottom-up and top-down hierarchical clustering approaches and provides superior clustering for
400 each level of resolution. MRtree also provides tools for sampling implicit resolution parameters for
401 Louvain clustering. This enables equal coverage of different clustering scales as input for the tree
402 construction process. All clustering methods face the challenge of determining the optimal number
403 of clusters supported by the data. While this problem is inherently intractable, MRtree uses a
404 stability criterion to determine the maximum resolution level for which stable clustering results
405 can be obtained for a given dataset. Because MRtree is agnostic to the clustering approach, it
406 can readily utilize input from any flat clustering algorithm. Hence MRtree is extremely flexible,
407 immediately incorporating the advantages of available clustering algorithms, while often providing
408 improved clustering at every resolution due to the reconciliation procedure.

409 To illustrate the performance of our method, we apply MRtree to a variety of scRNA-seq data sets,
410 including cells from the mouse brain, human pancreas, and human fetal brain tissues. Coupled
411 with suitable initial flat clustering algorithms, MRtree constructs the hierarchical tree that reveals
412 different levels of transcriptional distinction between cell types and outperforms popular competi-
413 tors, including bottom-up HAC and divisive methods such as CellBIC⁶. For functionally distinct
414 cell types that can be easily identified, the reconciliation process organizes the clusters obtained
415 under different scales into a unified hierarchical structure, and suggests a proper tree cut to retain
416 the stable partitions. For instance, the constructed tree from integrated pancreatic islet data sets
417 successfully identified endocrine and exocrine groups and subsequent cell types within each group.
418 The clusters from the tree of mouse cortex data sets accurately recovered the known major cell
419 types organized into subtrees of neurons and glial cells. Application of MRtree on human fetal brain
420 cells uncovered previously recognized main types organized in a tree structure along the maturation
421 trajectories.

422 Our method has a greater impact in challenging situations where clusters are similar. Apart from

423 validating the method on the widely-acknowledged main cell types, we uncovered a list of stable
424 subtypes from the fetal brain dataset that exhibit distinct states and functionality by examining
425 canonical gene ontology categories and significant PPI networks. Specifically, we have shown that
426 two subtypes of intermediate progenitors are well-defined by the expression of radial glia markers
427 versus newly born neurons markers. The subplate /deep layer excitatory neurons are mainly differ-
428 entiated by the layers the cells will populate. While migrating and maturing excitatory subtypes
429 show a gradual increase of upper layer excitatory neuron markers, upper and deep layer excitatory
430 neuron markers, respectively. Subtypes close in maturation states are reflected in the hierarchical
431 tree as they are split later down the tree. InMGE demonstrates the distinction in both matu-
432 ration and terminal specification with respect to the engagement of synaptic programs. At the
433 same time, InCGE subtypes differentiate mainly by maturation, which fits nicely with the fact that
434 CGE interneurons are born after MGE interneurons. While both cell types are born in the ventral
435 telencephalon, their terminal specification happens only upon beginning synaptogenesis when they
436 begin to express subtype-specific markers. Surprisingly, subtypes of ExDP revealed a set of genes
437 and structures similar to an intermediate progenitor that can be further investigated in future work.
438 It is worth noting that the quality of MRtree's construction relies on the performance of the chosen
439 flat clusterings. If the flat clusterings method inputs unstable or biased clusters, these errors
440 will be largely retained and reflected in the estimated hierarchical cluster tree. Similar to many
441 consensus clustering methods, MRtree can be extended to allow input from multiple sources, each
442 applying different flat clustering methods; however, the quality of the constructed tree depends on
443 the clustering performance of the full spectrum of sources. If the input data provides a disparate
444 signal, then the outcome is likely to be unstable.
445 Our studies suggest several interesting questions worthy of future investigations. For instance,
446 our method is a general framework that allows for any flat-clustering base procedure. In practice,
447 how to determine which base procedure suits better for different data sets still remains open. In
448 addition, the current framework relies on a rough idea about the range of resolution. Can we
449 automatically decide the range of resolution? How can we select this range when the resolution is
450 not parameterized by the number of clusters? In particular, our current method adopts a stability

451 measure to decide whether to further branch the hierarchical tree. Can we provide theoretical
452 guarantees for the power of this stopping criterion? Furthermore, our work shed light on how
453 major cell types evolve to subtypes, and we would like to further verify these biological findings.

454 Methods

455 We briefly formulate the optimization problem and introduce the algorithm we employed. A more
456 detailed description can be found in Supplemental Information. Given the transcriptomes of n cells
457 on p genes, denoted as $X \in \mathbb{R}^{n \times p}$, suppose clustering is performed using algorithm \mathcal{A} at a range
458 of m resolutions with parameters $\{k_1, \dots, k_m\}$. Here the resolution parameters are loosely defined
459 where it corresponds explicitly to the number of clusters for some algorithms, while it implicitly
460 determines the number of clusters for other algorithms. To formally state the problem and the
461 hierarchical reconciliation algorithm, we first introduce some notation.

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

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

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

473 *Constraint A₁: Each node $v_{j+1,k}$ has one and only one in-vertex edge.*

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

475 Condition A_1 ensures that any cluster in a higher resolution belongs to one and only one cluster
 476 in the adjacent lower resolution, that is, $\forall v_{j,k}, \exists k'$ such that $v_{j,k} \subseteq v_{j-1,k'}$. It further implies
 477 that the hierarchical tree can only be a branching tree as the resolution increases (top-down), and
 478 the clusters in lower levels should be intact in levels above it. Compared to the cluster trees,
 479 hierarchical cluster trees respect the clustering structure at higher resolutions in the sense that
 480 they keep samples that are together at higher resolutions in the same cluster for lower resolutions.
 481 Similarly, those samples that are far away from each other at a lower resolution do not enter the
 482 same clusters at high resolutions. We illustrate an example of A hierarchical cluster tree and its
 483 noisy companion in the form of a cluster tree in Figure S1.

484 Arrange the the clustering results at each resolution inside a label matrix

$$L(T_c) := [L_1, \dots, L_m] \in \mathbb{R}^{n \times m}, \quad (1)$$

485 where the j -th column denotes the corresponding labels for each data point at resolution k_j .

486 **Definition 3** For each data point $x_i, i = 1, \dots, n$, define its clustering path $p(x_i) :=$
 487 $(v_{1,l_{i1}}, \dots, v_{m,l_{im}})$ where $v_{j,l_{ij}}$ is the label for x_i at resolution k_j . Let $\mathcal{P}(T_c) := \{p(x_i) \mid i = 1, \dots, n\}$
 488 be the set of all unique paths.

489 **Optimization scheme** Let $T_c(k_1, \dots, k_m; \mathcal{A}, X)$ be the initial cluster tree by applying clustering
 490 algorithm \mathcal{A} on X , and let $T_h^*(k_1, \dots, k_m; \mathcal{A}, X)$ be the underlying true hierarchical cluster tree.
 491 Further denote the two respective n -by- p label matrices as $L(T_c; \mathcal{A}, X)$ and $L(T_h^*; \mathcal{A}, X)$. Our goal
 492 is to recover the unknown hierarchical tree from the observed initial cluster tree from the multi-
 493 resolution flat clustering. For ease of notation, we drop \mathcal{A}, X and replace (k_1, \dots, k_m) with \mathbf{k}^m .
 494 Assuming that $T_c(k_j)$ is an estimator of $T_h^*(k_j), j = 1, \dots, m$, if $T_c(\mathbf{k}^m)$ satisfy constraint A_1 , it
 495 naturally yields an estimator of $T_h^*(\mathbf{k}^m)$, though this is rarely the case. Following this idea, we
 496 construct the estimator by building a hierarchical cluster tree that mostly preserves the cluster
 497 structures from the observed cluster tree $T_c(\mathbf{k}^m)$ constructed from the initial flat clustering results.
 498 To achieve this, we define a loss function as the distance between the solution tree and initial flat
 499 clusterings $T_c(\mathbf{k}^m)$. We seek to minimize the loss under the constraint that the solution tree satisfies

500 constraint A_1 . To measure the difference between two trees, which is equivalent to measuring
 501 mismatch between two sets of partitions, we adopt hamming distance between the respective label
 502 matrices (defined in Eq. (1)). Hamming distance computes the number of location-mismatches of
 503 a pair of matrices, commonly used for measuring the distance between two paired partitions.
 504 The problem formulated above is equivalent to finding the optimum k_m distinct paths from the set
 505 of all feasible paths (Def. 3) of a cluster tree, to which all data points are assigned, and the induced
 506 multi-scale partitions preserve the most flat clustering structures. It is a combinatorial optimization
 507 problem. The complexity grows exponentially with the depth and number of clusters in each layer
 508 of the tree, and therefore is computationally intractable. To alleviate the computational burden,
 509 we introduce an equivalent objective function and propose a greedy algorithm to solve it. Formally,
 510 define $\tilde{V}(T)$ to be the set of “bad” vertices that have more than one in-vertex edge. Then for any
 511 proposed hierarchical cluster tree, i.e. $T \in \mathcal{T}_h$, we have $|\tilde{V}(T)| = 0$. The hierarchy is therefore
 512 estimated by solving the optimization problem respective to the newly-formulated constraint,

$$\hat{T}_h = \arg \min_T \min_{\pi} \mathcal{D}_{\text{Hamm}}(L(T), \pi(L(T_c))) \quad \text{subject to } |\tilde{V}(T)| = 0. \quad (2)$$

513 where $\mathcal{D}_{\text{Hamm}}(\cdot, \cdot)$ represents the hamming distance. The objective is minimized over permutation
 514 of labels $\pi = \pi_{k_j}, j = 1, \dots, m$ within each partition since the error should not be depending on
 515 how we label the classes.

516 We employ a greedy optimization procedure. The formulated problem (2) is first transformed to a
 517 soft constraint problem that shares the same solutions to allow for constraint violation during the
 518 optimization procedure. This enables initializing the solution with the observed flat cluster tree
 519 $T_c(\mathbf{k}_m)$. The objective is then minimized by sequentially “cleaning” one bad vertex in set $|\tilde{V}(T)|$ at
 520 a time. “Cleaning” the node refers to eliminate all but one edge that have this node as its in-vertex,
 521 followed by re-routing data points belonging to the eliminated path to remaining nearest viable
 522 paths. The increase in the objective as the results of cleaning the node is considered as the cost
 523 of eliminating the node from $\tilde{V}(T)$. In each iteration, the vertex in set $\tilde{V}(T)$ is evaluated for its
 524 elimination cost, where the one with the minimum cost is selected. The tree is then updated with
 525 the selected node being cleaned and affected data points re-assigned to the nearest remaining paths.

526 The procedure is repeated until $|\tilde{V}(T)| = 0$, which generates the desired hierarchical cluster tree.
527 The full algorithm is summarized in Algorithm 1. In Supplemental Information we analyze the
528 key properties of the algorithm: Theorem 1 provides the convergence properties, while Theorem 2
529 describes the memory and time complexity. In addition, we introduce methods for sampling implicit
530 resolution parameters with uniform coverage for modularity-based clustering (Seurat clustering),
531 including linear sampling, exponential sampling, and most preferably, *Event Sampling* method. We
532 also discuss ways of speeding up the algorithm in case of large sample size or a large number of
533 initial flat clusterings through layer-wise reconciliation and performing within-resolution consensus
534 clustering as the first step.

535 **Stability analysis to determine tree cut** We consider clustering stability to determine the
536 tree cut based on a basic philosophy that clustering should be a structure on the data set that is
537 “stable”. That is, if applied to data sets from the same underlying model, a clustering algorithm
538 should consistently generate similar results. Higher stability across resolutions is reflected as greater
539 consistency of individual initial flat clustering with the resulting clustering in the reconciled tree.
540 To measure the stability, we calculate the similarity using ARI between clusterings in corresponding
541 layers from the initial cluster tree and the resulted hierarchical cluster tree. This will generate a
542 line plot showing the similarity with increasing resolution. The tree cut can then be determined by
543 finding the “change point” where the stability is high at the current point and start to decrease by
544 further increasing the resolution.

545 **Clustering accuracy** To quantify the clustering performance in each layer of the hierarchical
546 tree, we utilize a novel modified version of Adjusted Rand Index (ARI)⁵, called Modified Multi-
547 resolution Rand Index (AMRI, Supplemental Information), as the accuracy metric to compare the
548 multi-resolution cluster structures with the true labels. The adjustment allows for comparisons
549 across resolutions, accounting for the reduced ability to uncover details in lower resolutions, thus
550 avoiding a bias towards fine-grained clustering results.

551 **Tree construction accuracy** To quantify the performance of hierarchical tree construction,
 552 given the true tree is known, we reduce the tree to a similarity matrix. Each entry of the matrix
 553 represents the length of branch two data points share. The longer branch a pair share, the more
 554 similar they are. In this way, we convert the measurement of the difference between hierarchies
 555 (dendograms) to measure the difference between two similarity matrices. The between-similarity
 556 distance is measure with the L_1 norm of the difference, defined by

$$D(T_1, T_2) = \|A_1 - A_2\|_1 = \sum_{i,j} |A_{1,ij} - A_{2,ij}|, \quad (3)$$

557 where A_1, A_2 are the similarity matrices of tree T_1, T_2 respectively. Given the certain tree structure,
 558 the induced similarity metrics can be visualized in Figure S14.

559 **Cluster Stability** Apart from examining the performance of MRtree for clustering accuracy, we
 560 also access the stability of the clusters at multiple resolutions prior to and post to tree reconciliation.
 561 Clustering stability has been considered as a crucial indicator of goodness of the clusters, given that
 562 well-performed partitions tend to be consistent across different sampling from the same underlying
 563 model or of the same data generating process¹¹. In practice, a large variety of methods has been
 564 devised to compute stability scores. Here we adopt the sub-sampling procedure, where the same
 565 clustering method is repeatedly performed on the independently sub-sampled data sets and compute
 566 the average similarity among the repetitions. Formally given a data set of n points S_n , let $\mathcal{C}_k(S_n)$
 567 be the resulted clustering outcome with k clusters. Let $\tilde{S}_n^{(b)}$ be a sub-sampling of S_n by randomly
 568 choosing a subset of size τn without replacement. Then the stability score is obtained by averaging
 569 the partition similarity on the shared data points,

$$Stab(k, n) = \frac{1}{B} \sum_{b=1}^B ARI(C_k(S_n), C_k(\tilde{S}_n^{(b)})). \quad (4)$$

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

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

575 **Declarations**

576 • Ethics approval and consent to participate

577 Not applicable.

578 • Consent for publication

579 Not applicable.

580 • Availability of data and materials

581 Not applicable.

582 • Competing interests

583 The authors declare no conflict of interest.

584 • Funding

585 This work was supported in part by National Institute of Mental Health (NIMH) grants
586 R01MH123184 and R37MH057881 to K.R., National Institutes of Health (NIH) grants R01-
587 MH116489 and R01-MH109912 to D.G., and Y.W. was partially supported by the NSF grants
588 CCF-2007911 and DMS-2015447.

589 • Authors' contributions

590 M.P. designed the work, performed the analysis, created the new software used in the work
591 and drafted the work, B.W. performed the analysis and drafted the work, A.E. performed
592 the analysis, D.G. designed the work, Y.W. designed the work and drafted the work, K.R.
593 designed the work and drafted the work.

594 • Acknowledgements

595 None at this time.

596 References

- 597 1. Maayan Baron, Adrian Veres, Samuel L Wolock, Aubrey L Faust, Renaud Gaujoux, Amedeo
598 Vetere, Jennifer Hyoje Ryu, Bridget K Wagner, Shai S Shen-Orr, Allon M Klein, et al. A
599 single-cell transcriptomic map of the human and mouse pancreas reveals inter-and intra-cell
600 population structure. *Cell systems*, 3(4):346–360, 2016.
- 601 2. Florian Buettner, Kedar N Natarajan, F Paolo Casale, Valentina Proserpio, Antonio Scialdone,
602 Fabian J Theis, Sarah A Teichmann, John C Marioni, and Oliver Stegle. Computational analy-
603 sis of cell-to-cell heterogeneity in single-cell rna-sequencing data reveals hidden subpopulations
604 of cells. *Nature biotechnology*, 33(2):155, 2015.
- 605 3. A. Butler, P. Hoffman, P. Smibert, E. Papalexi, and R. Satija. Integrating single-cell transcrip-
606 tomic data across different conditions, technologies, and species. *Nature Biotechnology*, 36(5):
607 411–+, 2018. ISSN 1087-0156.
- 608 4. D. Grun, A. Lyubimova, L. Kester, K. Wiebrands, O. Basak, N. Sasaki, H. Clevers, and A. van
609 Oudenaarden. Single-cell messenger rna sequencing reveals rare intestinal cell types. *Nature*,
610 525(7568):251–+, 2015. ISSN 0028-0836.
- 611 5. Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of classification*, 2(1):
612 193–218, 1985. ISSN 0176-4268.
- 613 6. Junil Kim, Diana E Stanescu, and Kyoung Jae Won. Cellbic: bimodality-based top-down
614 clustering of single-cell rna sequencing data reveals hierarchical structure of the cell type.
615 *Nucleic acids research*, 46(21):e124–e124, 2018.
- 616 7. V. Y. Kiselev, K. Kirschner, M. T. Schaub, T. Andrews, A. Yiu, T. Chandra, K. N. Natarajan,
617 W. Reik, M. Barahona, A. R. Green, and M. Hemberg. Sc3: consensus clustering of single-cell
618 rna-seq data. *Nature Methods*, 14(5):483–+, 2017. ISSN 1548-7091.
- 619 8. Satija Lab. *panc8.SeuratData: Eight Pancreas Datasets Across Five Technologies*, 2019. R
620 package version 3.0.2.

621 9. Damon Polioudakis, Luis de la Torre-Ubieta, Justin Langerman, Andrew G Elkins, Xu Shi,
622 Jason L Stein, Celine K Vuong, Susanne Nicterwitz, Melinda Gevorgian, Carli K Opland,
623 et al. A single-cell transcriptomic atlas of human neocortical development during mid-gestation.
624 *Neuron*, 103(5):785–801, 2019.

625 10. Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexis,
626 William M Mauck III, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. Com-
627 prehensive integration of single-cell data. *Cell*, 177(7):1888–1902, 2019.

628 11. Ulrike Von Luxburg. Clustering stability: an overview. 2010.

629 12. B. Wang, J. J. Zhu, E. Pierson, D. Ramazzotti, and S. Batzoglou. Visualization and analysis
630 of single-cell rna-seq data by kernel-based similarity learning. *Nature Methods*, 14(4):414–+,
631 2017. ISSN 1548-7091.

632 13. X. R. Wang, J. Park, K. Susztak, N. R. Zhang, and M. Y. Li. Bulk tissue cell type deconvolution
633 with multi-subject single-cell expression reference. *Nature Communications*, 10, 2019. ISSN
634 2041-1723.

635 14. Yue J Wang, Jonathan Schug, Kyoung-Jae Won, Chengyang Liu, Ali Naji, Dana Avrahami,
636 Maria L Golson, and Klaus H Kaestner. Single-cell transcriptomics of the human endocrine
637 pancreas. *Diabetes*, 65(10):3028–3038, 2016.

638 15. Joe H Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the
639 American statistical association*, 58(301):236–244, 1963.

640 16. Nicola K Wilson, David G Kent, Florian Buettner, Mona Shehata, Iain C Macaulay, Fer-
641 nando J Calero-Nieto, Manuel Sánchez Castillo, Caroline A Oedekoven, Evangelia Diamanti,
642 Reiner Schulte, et al. Combined single-cell functional and gene expression analysis resolves
643 heterogeneity within stem cell populations. *Cell stem cell*, 16(6):712–724, 2015.

644 17. Luke Zappia and Alicia Oshlack. Clustering trees: a visualization for evaluating clusterings at
645 multiple resolutions. *GigaScience*, 7(7):giy083, 2018.

646 18. Amit Zeisel, Ana B Muñoz-Manchado, Simone Codeluppi, Peter Lönnerberg, Gioele La Manno,
647 Anna Juréus, Sueli Marques, Hermany Munguba, Liqun He, Christer Betsholtz, et al. Cell
648 types in the mouse cortex and hippocampus revealed by single-cell rna-seq. *Science*, 347(6226):
649 1138–1142, 2015.

650 19. X. W. Zhang, C. L. Xu, and N. Yosef. Simulating multiple faceted variability in single cell rna
651 sequencing. *Nature Communications*, 10, 2019. ISSN 2041-1723.

652 20. L. X. Zhu, J. Lei, B. Devlin, and K. Roeder. A unified statistical framework for single cell and
653 bulk rna sequencing data. *Annals of Applied Statistics*, 12(1):609–632, 2018. ISSN 1932-6157.

654 21. Lingxue Zhu, Jing Lei, Lambertus Klei, Bernie Devlin, and Kathryn Roeder. Semisoft clustering
655 of single-cell data. *Proceedings of the National Academy of Sciences*, 116(2):466–471, 2019.

656 22. J. Zurauskiene and C. Yau. pcareduce: hierarchical clustering of single cell transcriptional
657 profiles. *Bmc Bioinformatics*, 17, 2016. ISSN 1471-2105.