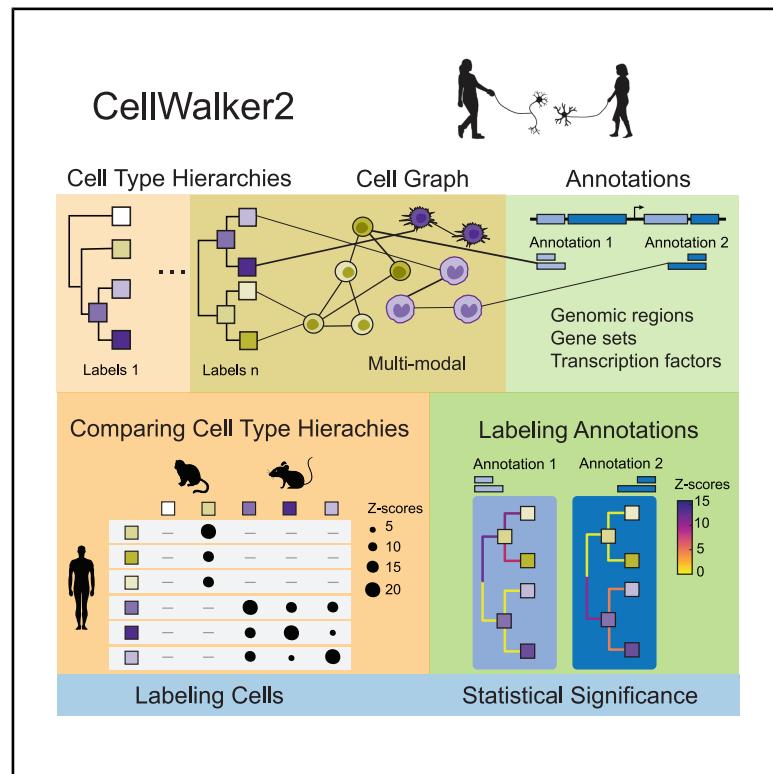


CellWalker2: Multi-omic discovery using hierarchical cell type relationships

Graphical abstract



Authors

Zhirui Hu, Paweł F. Przytycki,
Katherine S. Pollard

Correspondence

katherine.pollard@gladstone.ucsf.edu

In brief

Hu et al. present CellWalker2, a graph-based model that improves comparisons of cell types across contexts and species. By integrating multimodal single-cell data and modeling hierarchical cell-type relationships, CellWalker2 annotates cells, genomic regions, and gene sets while assessing the statistical significance of these mappings.

Highlights

- Hierarchical cell-type relationships improve cell-type mapping
- Multi-modal data link genomic regions to cell types via chromatin accessibility
- CellWalker2 assigns cell-type labels to regulatory elements and TFs
- CellWalker2 quantifies conserved and divergent cell-type relationships across species

Technology

CellWalker2: Multi-omic discovery using hierarchical cell type relationships

Zhirui Hu,¹ Paweł F. Przytycki,^{1,2} and Katherine S. Pollard^{1,3,4,5,*}

¹Gladstone Institute of Data Science & Biotechnology, 1650 Owens Street, San Francisco, CA 94158, USA

²Faculty of Computing & Data Sciences, Boston University, 665 Commonwealth Avenue, Boston, MA 02215, USA

³Department of Epidemiology & Biostatistics, University of California, San Francisco, 1650 Owens Street, San Francisco, CA 94158, USA

⁴Chan Zuckerberg Biohub SF, 499 Illinois Street, San Francisco, CA 94158, USA

⁵Lead contact

*Correspondence: katherine.pollard@gladstone.ucsf.edu

<https://doi.org/10.1016/j.xgen.2025.100886>

SUMMARY

Tissues are composed of cells with a wide range of similarities to each other, yet existing methods for single-cell genomics treat cell types as discrete labels. To address this gap, we developed CellWalker2, a graph diffusion-based model for the annotation and mapping of multi-modal data. With our open-source software package, hierarchically related cell types can be probabilistically matched across contexts and used to annotate cells, genomic regions, or gene sets. Additional features include estimating statistical significance and enabling gene expression and chromatin accessibility to be jointly modeled. Through simulation studies, we show that CellWalker2 performs better than existing methods in cell-type annotation and mapping. We then use multi-omics data from the brain and immune system to demonstrate CellWalker2's ability to assign high-resolution cell-type labels to regulatory elements and TFs and to quantify both conserved and divergent cell-type relationships between species.

INTRODUCTION

Single-cell technologies are revealing the diversity of cells within multi-cellular organisms and differences in cellular heterogeneity between tissues. Genomic assays, such as single-cell RNA sequencing (scRNA-seq) and single-cell ATAC sequencing (scATAC-seq), can be used to group cells within a tissue into cell types that may represent developmental lineages, functional specializations, or dynamic responses to the microenvironment. This knowledge is propelling discoveries about cellular diversity across evolution, development, and disease. One major limitation of existing single-cell analysis methods is their treatment of cell types as discrete and unrelated labels, despite cell types having varying relationships to each other. We hypothesized that modeling highly distinct cell types (e.g., from two different germ layers) in the same way as closely related subtypes leads loss of power and interpretability in single-cell analysis.

Downstream analyses, such as identifying differentially expressed genes or differentially accessible regions (DARs), rely on accurately annotating cells to cell types using scRNA-seq and/or scATAC-seq. Many methods have been developed for this task, including some based on classical machine learning methods (e.g., Seurat,¹ Signac,² ArchR,³ CellTypist,⁴ SIMBA,⁵ cisTopic,⁶ snapATAC,⁷ and LIGER⁸), and others that leverage deep learning (e.g., GLUE,⁹ MARS,¹⁰ scArches,¹¹ scTGCN,¹² and scANVI¹³). Beyond discrete cell types, methods such as velocity¹⁴ and CellRank¹⁵ infer cell trajectories or fates using RNA velocity, while MIRA¹⁶ does so using expression and accessibility.

As more and more single-cell data are generated, it has become imperative to be able to compare cell-type annotations across studies. Single-cell datasets from the same tissue often have distinct cell-type labels due to biological variation across samples, different modalities of the data (e.g., scRNA-seq, scATAC-seq, or multi-ome data), variable sequencing depths, different computational methods or tuning parameters (e.g., clustering resolution), and divergent choices when naming cell clusters. Although cell-type labels can be manually compared, few computational methods can automatically match cell-type labels and provide a probabilistic measure of the mapping. Methods that directly map cell types include MARS, which trains a neural network model for cell-type classification; treeArches,¹⁷ which uses kNN classifiers after embedding cells via a deep learning model; and CellHint,¹⁸ which uses a predictive clustering tree algorithm. Existing methods for comparing cell types across contexts do not take into account the hierarchical relationships of cell types. While treeArches and CellHint do build cell-type hierarchies upon integrating multiple datasets, they cannot compare existing cell-type hierarchies directly.

Integrating data from different omics modalities also facilitates interpretation of annotations from bulk data at the single-cell level. For example, scATAC-seq data can be used to assess cell type activities of transcription factor (TF) motifs, regulatory regions identified from bulk experiments (chromatin immunoprecipitation sequencing [ChIP-seq] or ATAC-Seq),^{19–22} or SNPs from expression quantitative trait locus (eQTL) experiments²³ or genome-wide association studies.²⁴ Most methods (e.g.,

Signac, snapATAC, and ArchR) identify cell-type-specific annotations after clustering, annotating cells and identifying DARs by testing for enrichment in cell-type-specific peaks. Because every step loses some information from the original sequencing data of each cell, cell-type labeling and DAR identification can have large uncertainties, especially in complex tissues, thereby complicating the calculation of statistical significance for associations between annotations and cell types. cisTopic uses a topic model to simultaneously cluster cells and regions, enabling users to identify TF motifs for each topic, but these are not directly linked to cell types.

In contrast, CellWalker²² provides a framework that directly assigns bulk-derived labels to cell types by constructing a graph of cells and cell types using scATAC-seq data. While this enables regions to be mapped to cell types, the statistical significance of these mappings is not established, making it difficult to compare cell type-specific annotations across different conditions or species. SIMBA is another graph-based method that incorporates cells and features (genes, peaks, motifs, k-mers) into the same graph and outputs embeddings for all these elements. SIMBA measures how close cells and features are to each other (e.g., nearest TFs to a cell), but, as it does not incorporate cell type labels in the graph or perform clustering, it does not directly map cell types to cells or features and does not directly output the relationships between cell-type labels. In addition, although using a cell-type hierarchy can increase power to detect cell-type-specific regions, existing methods can only map bulk-derived annotations to a single level of a cell-type hierarchy.

Motivated by these gaps in the single-cell toolkit, we sought to combine the individual strengths of existing methods into a single integrative modeling framework while ensuring that the resulting method provides robust performance and estimates of statistical significance. We significantly extended the CellWalker graph-diffusion model to add (1) hierarchical relationships between cell types, (2) a permutation null distribution for estimating statistical significance, (3) flexibility to use scRNA-seq, scATAC-seq, or multi-ome data, and (4) functionality for comparing cell types across contexts. Using a graph enables us to avoid assuming that cells are independent, which is important when computing statistical significance. Our open-source software, CellWalker2, can be used to assign cell-type labels to either cells or annotations (e.g., gene sets, TF binding sites, or genetic variants) using any cell-type ontology (hierarchical or not). The model also enables statistical comparisons between cell types from two or more ontologies, allowing users to assess the similarity of cell types across species, disease states, and research groups.

DESIGN

CellWalker2 serves as a modeling and statistical inference tool to be used after processing raw sequencing reads and calling candidate regulatory elements, allowing it to naturally plug in downstream of existing single-cell quantification software tools (e.g., Seurat, Signac and ArchR). The inputs are (1) count matrices from scRNA-seq (gene by cell) and/or scATAC-seq (peak by cell), (2) one or more cell type ontologies (e.g., tree of

cell type relationships with marker genes for each leaf node), and optionally (3) regions of interest (e.g., genetic variants, regulatory elements, gene sets) (Figure 1A). CellWalker2 builds a single heterogeneous graph that integrates these inputs (Figure 1B). Then, the algorithm conducts a random walk with restarts on the graph and computes an influence matrix. From sub-blocks of the influence matrix, CellWalker2 learns relationships between different nodes. For instance, label-to-label similarities enable users to compare different cell-type ontologies by learning how cell types in one context (e.g., lab, disease state, species) map to cell types in another. With hierarchical ontologies, CellWalker2 provides relationships not only for cell types with marker genes but also for internal nodes that represent broader cell types. As additional applications, cells can be mapped to cell types using cell-to-label similarities, and bulk-derived genomic elements and genetic variants can be mapped to cell types using annotation-to-label similarities. Finally, CellWalker2 performs permutations to estimate the statistical significance (Z scores) of these learned associations (Figure 1C).

CellWalker2 is notably different from CellWalker, which does not model hierarchical relationships or assess statistical significance, uses only open chromatin data and not gene expression to quantify similarity, and uses an *ad hoc* method to map genome coordinates to labels rather than including coordinates as nodes in the graph. CellWalker2 is also distinct from clustering methods that define cell types *de novo*; it requires marker genes from one of these methods or an expert curator as input, and hence it is not designed to compete with these methods. Instead, it focuses on using reference cell types to annotate a query dataset and comparing different sets of reference cell types with each other in the context of single-cell data. In the following, we highlight these new functionalities by first describing the CellWalker2 model and then demonstrating (1) cell annotation using scRNA-seq data, (2) comparing cell-type hierarchies using scRNA-seq data, and (3) mapping bulk-derived regulatory regions to cell types using multi-omics data.

Constructing cell graphs

The nodes in a CellWalker2 graph represent three types of entities: cells, cell types (labels), and regions of interest (if provided; annotations). Cell nodes are associated with scATAC-seq, scRNA-seq, or multi-omics data, and the cells can come from different studies, conditions, or species. Label nodes possess marker genes defining the cell type, which are predefined using one of the many available approaches. The data at annotation nodes are genomic coordinates or gene names. Nodes are connected by edges, derived from the single-cell count matrices, and edge weights quantify relationships among cells, among cell types, between cells and labels, and between cells and annotations (Figure 1B). Cell-to-cell edges are computed based on each cell's nearest neighbors in terms of genome-wide similarity between cells (STAR methods). Cell-to-label edge weights are based on the expression (or accessibility) of each label's marker genes in each cell. Annotation-to-cell edges are based on accessibility of the genome regions or expression of genes in the gene set. Label-to-label edges are an input to CellWalker2 that is either part of the cell type ontology or pre-computed using gene expression similarity.

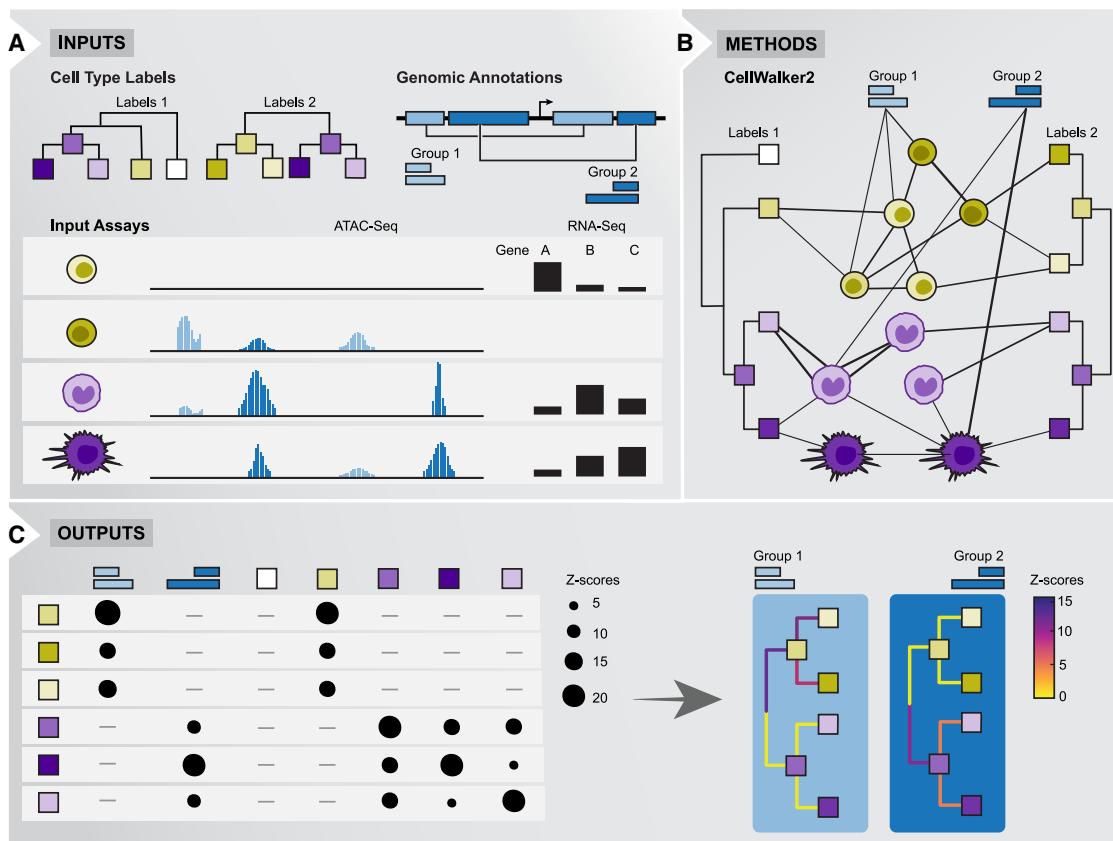


Figure 1. Overview of CellWalker2

(A) The inputs to CellWalker2 are (1) one or more sets of cell type labels with marker genes and an optional hierarchical structure; (2) cells with RNA-Seq and/or ATAC-Seq data; and (3) optionally, gene sets or annotations with genome coordinates that may be derived from bulk assays (e.g., genes for which the proteins form a complex, TF motifs).

(B) CellWalker2 constructs a graph with labels, cells, and annotations as nodes. Cells are connected to each other, to labels, and to annotations with edge weights that are computed based on the available assays for the cell. For example, cell-to-cell edge weights are based on genome-wide expression and/or chromatin accessibility ([STAR methods](#)). The edge weight between a label and a cell is based on expression of the label's marker genes in the cell (no edge if the cell does not have RNA-seq). The edge weight between an annotation and a cell is based on chromatin accessibility of the genome coordinates in the cell (no edge if the cell does not have ATAC-seq), while gene set to cell edge weights are based on gene expression. A random walk on the graph is performed to calculate the influence scores between all pairs of nodes.

(C) CellWalker2 outputs Z scores that measure the statistical associations between each cell type label and (1) every annotation and (2) all other cell types. This general framework is flexible and can be modified for different applications by generating graphs with different combinations of assays, labels, and annotations ([Figure S3](#)).

Computing influence scores and Z scores

CellWalker2 does a random walk with restarts on the graph, initiating walks from all nodes and computes the influence score matrix that represents the steady-state probability of reaching each node from all nodes²² ([STAR methods](#)). This matrix summarizes how strongly each node in the graph is associated with all other nodes given the topology and edge weights. To test if CellWalker2 can integrate scRNA-seq and scATAC-seq cells using influence scores, we benchmarked it against GLUE and SIMBA using multi-omics data from peripheral blood mononuclear cells (PBMCs; [STAR methods](#)). Using this as ground truth, we simulated single-modality data by assigning the RNA-seq and ATAC-seq modalities to separate groups of cells. CellWalker2 was competitive or better at connecting cells of the same cell type ([Figure S1; STAR methods](#)). This shows that

our heterogeneous graph model encourages information flow among similar cells with different modalities.

To quantify the statistical significance of relationships in the influence matrix, CellWalker2 computes a Z score for each entry by comparing the observed value with its expectation and variance under a permutation null distribution. This distribution is estimated by permuting graph edges while maintaining node degree, and the set of permuted edges depends on the type of relationship being evaluated ([STAR methods](#)). Controlling for node degree is critical, because labels that are prevalent and connected to many cells tend to have larger influence scores just by chance. In contrast, our Z scores are robust to variability in node degrees and quantify the statistical significance of node associations ([Figure S2](#)). The larger the Z score, the more significant the association. When the underlying data are normally

distributed, as will be the case for nodes with high degrees, the Z score reflects the number of SDs an observation is from the mean and can be easily converted to a p value using the standard normal distribution. Here, we focus on Z scores rather than p values to avoid distributional assumptions.

CellWalker2 use

Labeling cells

If a user has cells from a single-cell experiment (query dataset) and would like to annotate them using a cell-type ontology (reference dataset), they would build a CellWalker2 graph containing cell nodes with gene expression data plus label nodes with marker genes defined in the reference dataset (Figure S3A). Optionally, cells from the experiment that generated the reference labels may also be included (Figure S3B). To map labels to query cells, CellWalker2 computes the cell-to-label normalized influence scores (STAR Methods). It assigns each cell to the cell-type label with the largest normalized influence score, and the vector of normalized scores across all labels can be used as a probabilistic assignment.

Comparing cell types

To compare cell-type labels between two or more ontologies, the user builds a CellWalker2 graph that includes nodes for all labels and nodes for cells from the datasets from which the labels were defined (Figure S3C). The graph includes edges between cells from different datasets, creating paths that connect the ontologies. The resulting influence matrix includes measurements of the information that flows from labels in one ontology to labels in another ontology. These mappings are converted to Z scores using a null distribution obtained by permuting cell-to-label edges (STAR methods). A label's highest Z score indicates the most significant corresponding cell type in the other ontology, and any pair of labels with a high Z score is more connected than expected by chance. Users can compare a cell type's Z scores for nodes in the other ontology to evaluate if there is a single best mapping versus a more general mapping to a group of related nodes (e.g., clade in a cell-type tree).

Labeling bulk-derived annotations

Users can assign cell-type labels to annotations that lack cellular resolution, such as disease-associated variants, TF ChIP-seq peaks, or eQTLs from bulk experiments, by building a CellWalker2 graph that includes nodes for labels, annotations, and cells. If the annotations are genomic regions, some cells must have multi-ome data, because annotation-to-cell edges are based on accessibility and label-to-cell edges are based on marker gene expression (Figure S3D). However, the graph may also include cells with only scRNA-seq or only scATAC-seq, which provide additional information about cell types (Figure S3E). In this case, cells with multi-omics data serve as bridges connecting cells with only one modality, creating paths between annotations and labels; a similar idea is exploited in Hao et al.²⁵ A high Z score between an annotation and a cell-type label means that the genome coordinates are more highly accessible in that cell type compared with other cell types. For example, TFs specific to one or a small subset of cell types can be prioritized using Z scores for their motifs or ChIP-seq peaks.

Software

We implemented CellWalker2 in R by extending the CellWalkR package.²⁶ The open-source software, available at Github <https://github.com/PFPPrzytycki/CellWalkR/tree/cellwalker2>, includes documentation and vignettes. CellWalker2's functions and pipelines are shown in Figure S4.

RESULTS

To demonstrate how the robustness and flexible functionality of CellWalker2 enables biological discovery, we first benchmarked CellWalker2's performance on different tasks and then analyzed single-cell data from three contexts encompassing different complex tissues, developmental stages, and species.

Benchmarking CellWalker2 on different tasks

Simulations to evaluate cell labeling

To compare the performance of CellWalker2 versus Seurat, a commonly used tool for assigning reference cell-type labels to cells in a query dataset, we designed a series of simulation scenarios where the correct cell labels are known (STAR methods) (Figure S5A). CellWalker2 and Seurat perform equally well when cell types are distinct and no batch effects or dropout are simulated (easy scenario) (Figure 2A). With batch effects and dropouts (medium scenario) (Figure S5B), CellWalker2 performs better than Seurat (Figure 2A). The gap in performance is greater when the CellWalker2 graph includes cells from the reference dataset, but outperforming Seurat is possible using a subset of cells from only the query dataset. This provides a computational advantage, especially when the reference dataset has many cells (e.g., a cell atlas), because Seurat needs to integrate cells from both datasets.

When we probed performance as a function of cell composition by altering the number of cells in the smallest cell population in the reference dataset, we observed that Seurat incorrectly annotates cells to a related and more abundant cell type, whereas CellWalker2 is robust to the change (Figure 2B). When the cell types are less well separated (hard scenario) (Figure S5C), CellWalker2's performance decreases, but remains better than Seurat's. CellWalker2 also performs better at identifying differentially expressed genes because of its superior performance in cell annotation (Figure S6). Finally, using a hierarchical cell-type ontology provides a small performance advantage over discrete cell-type labels.

Simulations to evaluate cell-type comparisons

To compare CellWalker2 with treeArches and MARS, we simulated four scenarios (STAR Methods). In each scenario, three cell types are shared between dataset 1 (DS1) and dataset 2 (DS2), but the fourth cell type in DS2 has a different relationship to the cell-type hierarchy in DS1 representing a divergent cell type, an ancestral cell type, an altered cell type, and a convergent cell type (Figure S7). In all cases, CellWalker2 assigns the largest Z score to the closest cell type, and Z scores decrease as the true similarity diminishes (Figures 2C and S8). Moreover, the Z scores effectively differentiate between the four scenarios. In comparison, MARS accurately detected the ancestral cell type but could not distinguish the other three scenarios, while treeArches struggled to differentiate the divergent and

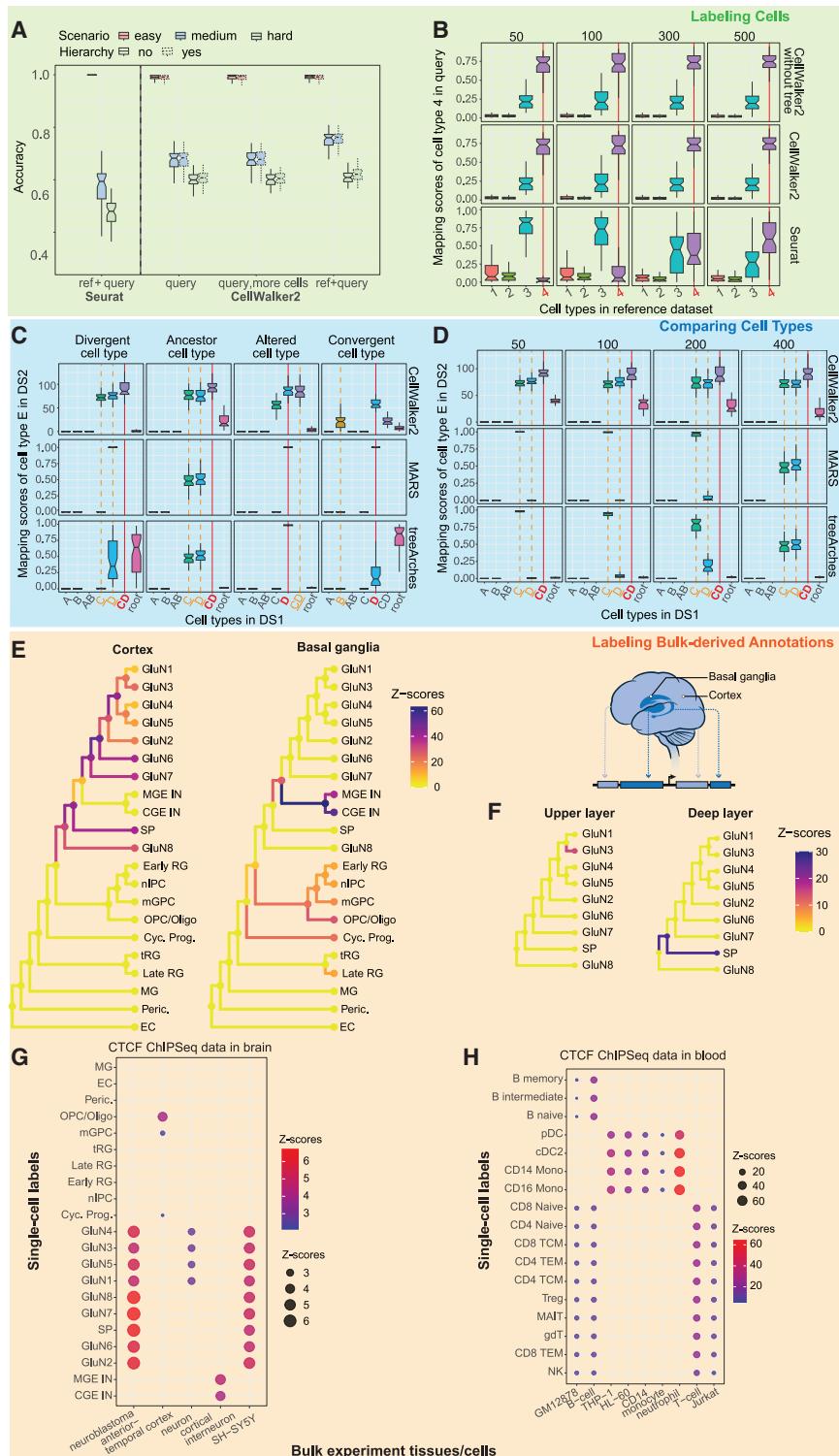


Figure 2. Benchmarking CellWalker2's functionalities: labeling cells, comparing cell types, and labeling bulk-derived annotations

(A and B) For labeling cells, CellWalker2 outperforms Seurat in simulations. For (B–D), the expected output given the cell type relationships is in red and bold, while the second closest cell types are in orange and underlined (see Figure S7 for details on cell type relationships). The vertical lines above the labels are shown in the same colors as well. (A) Performance was evaluated across different simulation scenarios: easy (red, no batch effect or dropout), medium (blue, batch effect and dropout), hard (green, batch effect and more dropout). Hierarchy indicates whether edges between cell types are included. For CellWalker2, different alternatives are implemented: query (only includes cells from the query dataset to construct the graph), query, more cells (only includes cells from the query dataset but increases the number of cells to be equal to the total number of cells in the query and reference datasets), and ref+query (includes cells from both the reference and query datasets in the cell graph).

(B) CellWalker2 is more accurate on rare cell populations. The percentage of cells from the query dataset mapped to each reference cell type (vertical axis) is shown as the size of the correct cell type in the reference dataset (label 4) increases from 50 (3%) to 500 (32%) cells. CellWalker2 without tree, running CellWalker2 without the tree structure of reference cell types.

(C and D) For comparing cell types between simulated DS1 and DS2, CellWalker2 outperforms treeArches and MARS. (C) Boxplots show mapping cell type E of DS2 to different cell types in DS1 in four simulation scenarios (columns). The scores on the vertical axes are Z scores for CellWalker2 and probabilities for MARS and treeArches. In each section, the horizontal axis represents different cell types in DS1. AB, parent node of A and B; CD, parent of C and D. CellWalker2 can map to all nodes in the cell type hierarchy, treeArches to tip and root nodes but not internal nodes, and MARS only to tips. (D) CellWalker2 is robust to rare cell populations. The boxplots show mapping Z scores or probability (treeArches and MARS) for cell type E in DS2 to cell types in DS1 as the size of the correct cell type in DS1 (cell type D) increases from 50 (4%) to 400 (32%) cells. For (A–D), each simulation scenario is repeated 50 times.

(E and F) pREs from different brain regions of human developing telencephalon²⁷ show differential chromatin accessibility across cell types. The branches and nodes are colored by CellWalker2's Z scores.

(G) Mapping basal ganglia versus cortex specific pREs to the cell type hierarchy in Trevino et al.²⁸

(H) Mapping upper versus deep layer specific pREs from PFC onto the subtree of excitatory neurons.

(G and H) CellWalker2 maps CTCF ChIP-seq peaks from cell lines and primary cells to the corresponding cell types in single cell data. (G) Z scores using developing human cortex data to map brain ChIP-seq peaks. (H) Z scores using 10× Genomics multi-omics PBMC data to map blood ChIP-seq peaks.

(legend continued on next page)

convergent cell types. We also observed that MARS frequently provides a non-zero probability to only one cell type so that equal and secondary relationships are not detected, and treeArches often assigns a high probability to the root node, meaning that it does not detect any specific cell-type relationships. To see how cell composition affects performance, we ran slightly altered divergent cell-type simulations varying the size of one cell type in DS1. CellWalker2's Z scores were robust to changes in cell-type proportions, while treeArches and MARS tended to assign the new cell type to a the label of the most prevalent cell type (Figure 2D). Additionally, we find that CellWalker2 is robust to erroneous cell labels (Figure S9).

Using biological knowledge to benchmark labeling of bulk-derived annotations

To compare CellWalker2's annotation labeling to commonly used statistical tests and the original CellWalker method,²² we utilized predicted regulatory elements (pREs) identified from bulk ATAC-seq of different micro-dissected cortical regions²⁷ as annotations and assessed labeling results based on known differences in cell-type composition between regions of the human developing cortex. We constructed a CellWalker2 graph using 19,151 pREs as annotations, cell-type labels from a study of human developing cortex,²⁸ and cells from the same study that were assayed with scRNA-seq or scATAC-seq (multiple developmental stages) or multi-omics (21 post-conception weeks).

Focusing initially on one of the most striking differences in cell-type composition, we compared CellWalker2's cell-type labeling for 6,941 pREs from the cortical plate versus 3,463 from the basal ganglia. Consistent with expectations, excitatory neuron Z scores were higher for cortical plate, while those for inhibitory neurons, progenitors, and radial glia (RG) were highest for basal ganglia (Figure 2E). Wilcoxon tests comparing the distribution of edge weights to basal ganglia versus cortical plate pREs for each cell type defined in Trevino et al.²⁸ shows a similar cell-type specificity pattern as CellWalker2 (Figure S10A), but this method can only be applied after labeling every cell with a cell-type label. As a third method, we tested for enrichment of cortical plate and basal ganglia pREs overlapping with cell-type-specific DARs. It detected enrichment of DARs for cycling progenitors and RG in cortex-specific pREs, but these cell types do not enter the cortical plate in early development (Figure S10B). Finally, the original *ad hoc* method implemented in CellWalkR does not assign distinct cell types to pREs from cortex versus basal ganglia (Figure S10C). These results demonstrate that using a cell-type hierarchy and modeling genome regions as graph nodes notably boost the performance of CellWalker2.

To evaluate CellWalker2 on smaller sets of pREs in a context where cell type differences are more subtle than cortex versus basal ganglia, we zoomed in on excitatory neurons of the dorso-lateral prefrontal cortex (PFC) and repeated the above analysis using 2,333 pREs from the upper cortical layers versus 445 from deep layers. Consistent with expectations, CellWalker2 Z scores map upper layer pREs to glutameteric neurons and deep layer pREs to subplate (Figure 2F). Adding scATAC-seq cells to the graph revealed significant mappings between upper layer pREs and additional subtypes of glutameteric neurons (e.g., GluN5) (Figure S11A), and these mappings varied as expected when using scATAC-seq from an earlier developmental

stage (post-conception week 16) or a more fine-grained cell-type ontology (Figures S11B and S11C). These findings show that CellWalker2 Z scores accurately annotate bulk-derived regulatory elements with the cell types in which they are most likely to be active.

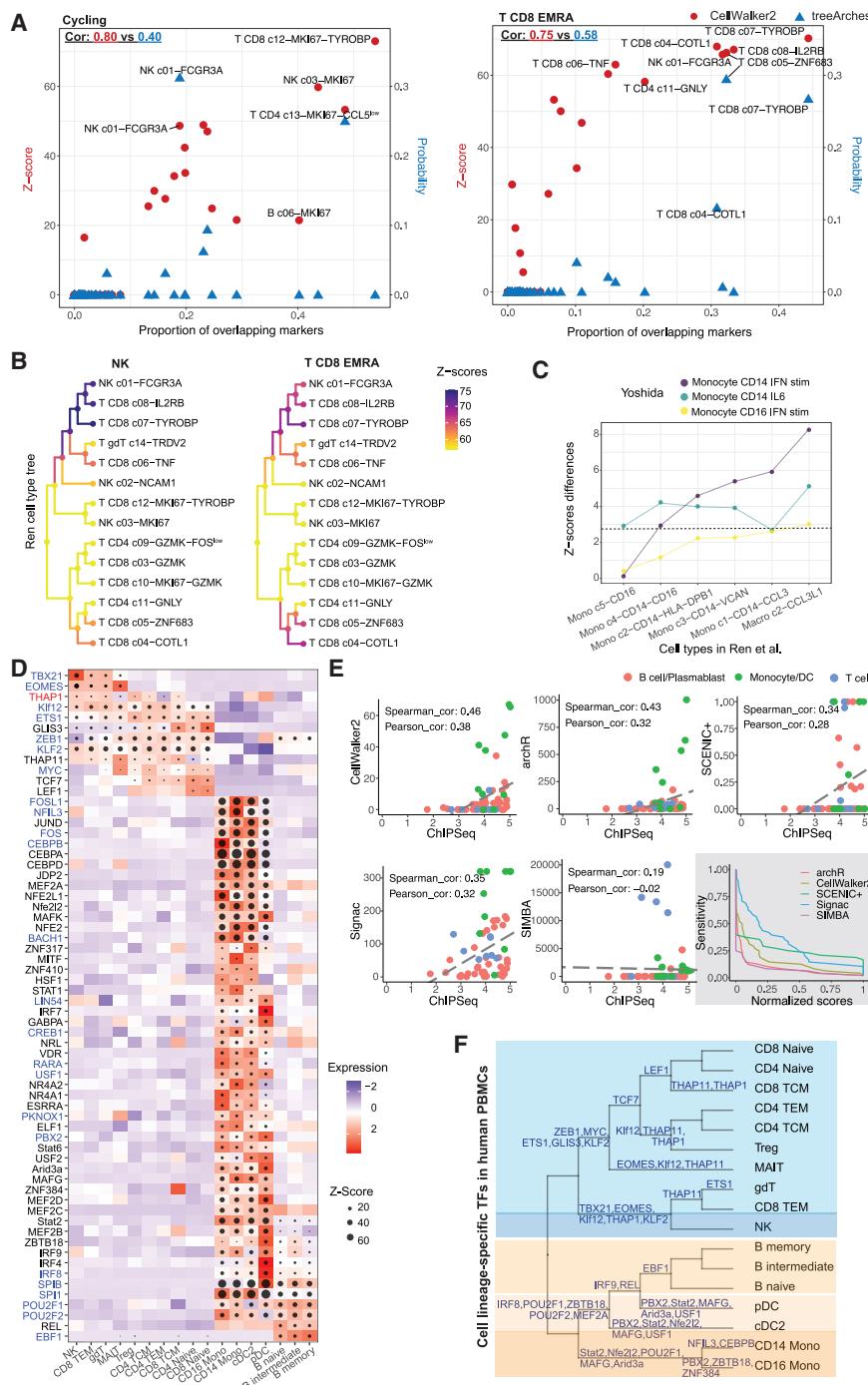
As a second evaluation of CellWalker2's annotation labeling, we quantified how well the method labels CTCF peaks from different ChIP-seq experiments (Figures 2G and 2H). We chose CTCF because it is ubiquitous, but its binding sites vary across cell types and often correspond with regions of open chromatin. CellWalker2 correctly labels CTCF peaks from neurons and interneurons as glutamatergic and GABAergic neurons, respectively (Figure 2G). Repeating this analysis for CTCF ChIP-seq peaks from B cell, T cell, and monocyte cell lines and primary cells²⁹ using multi-omics data from PBMCs ([STAR methods](#)), we again observe the highest Z scores in the corresponding cell types (Figure 2H). In summary, given a set of genome coordinates, CellWalker2 can identify the particular cell types in which the chromatin in these regions is most accessible using single cell multi-omics data and/or scATAC-seq data.

Human PBMCs

PBMCs are heterogeneous and contain many closely related cell types, exemplified by various kinds of immune cells that transition into alternative states upon stimulation. In this dynamic landscape, TFs govern gene expression and cellular functions. Consensus on cell-type definitions across studies is lacking, as is a comprehensive list of activating TFs for cell types and lineages. Here, we use CellWalker2 to address these gaps and compare its results with other single-cell analysis tools.

Comparing cell-type hierarchies

We analyzed two human PBMC datasets^{30,31} in which different strategies were used to define cell types at a high resolution: marker gene expression³⁰ versus cellular functions,³¹ complicating direct comparisons between the two cell type ontologies. We compared the ontologies using CellWalker2, MARS, and treeArches. MARS fails to map 38% of cell types, including known correspondences between platelets and megakaryocytes and between plasmacytoid dendritic cells (DCs) (plasmacytoid DC and DC c4-LILRA4), which are correctly mapped by CellWalker2 and treeArches (Figure S12). Furthermore, only 55% of MARS's top mappings overlap with the top three CellWalker2 mappings (Figure S12), whereas treeArches and CellWalker2 are much more correlated and concordant (82% of top mappings). Some differences between CellWalker2 and treeArches stem from treeArches being biased toward more prevalent cell types, as we saw in simulations. For example, treeArches maps CD4 cytotoxic T lymphocytes to more prevalent CD8 cytotoxic T lymphocytes (Figure S12). TreeArches also misses the correspondence between plasma cells and B c05-MZB1-XBP1, where XBP1 is a marker of plasma cells.³² CellWalker2 identifies both of these biologically supported mappings, and it also performs well on cell types with multiple markers, such as associating Mono c4-CD14⁻CD16 with both CD14 and CD16 monocytes (Figure S12). Finally, CellWalker2's Z scores are correlated with the proportion of overlapping markers (Figure S13), with greater marker concordance than treeArches (Figure 3A). In sum, CellWalker2 provides a statistical



different thresholds. Greater sensitivity means that the method can recover more ChIP-seq validated TF-to-cell type mappings at a given threshold. The horizontal axis shows the scores from various methods normalized by their maximum values.

(F) CellWalker2's top TFs for tip and internal nodes of the cell type tree from 10 \times Genomics multi-omics PBMC dataset. Shaded regions on the cell type tree reflect different classes of cell types.

IFN, interferon.

mapping between PBMC ontologies that includes expected relationships among prevalent cell types, in agreement with treeArches, plus several unique yet biologically plausible associations between rare and complex cell types.

Because CellWalker2 associates labels in one ontology with multiple hierarchically related labels in another dataset, the resulting vectors of Z scores can be used to cluster cell-type labels into groups based on their similarity to the second

Figure 3. Comparing cell types and identifying cell type-specific TFs in human PBMCs using CellWalker2

(A) Cycling cells (left) and T CD8 EMRA cells (right) from Yoshida et al.³¹ were mapped to the PBMC types in Ren et al.³⁰ using CellWalker2 and treeArches. For each plot, the left vertical axis and red symbols are CellWalker2's Z score, while the right axis and blue symbols are treeArches' probability score. The horizontal axis shows the proportion of overlapping positive markers. The Spearman's rank correlation for each method (red, CellWalker2; blue, treeArches) is shown in the top left corner of each graph. CellWalker2's scores are more highly correlated with marker gene overlap.

(B) Instead of trying to pinpoint a single tip within each clade, which could be unrealistic due to technical or biological variability, CellWalker2 maps NK to a single clade (multiple related labels and their ancestral nodes) and T CD8 EMRA to multiple clades of the cell-type tree from Ren et al.³⁰. The branches and nodes are colored by Z scores.

(C) Z scores capture differences between stimulated and unstimulated cell states. Cell types from Yoshida et al.³¹ were mapped to monocyte-related cell types in Ren et al.³⁰ (horizontal axis). Vertical axis shows differences in Z scores for three stimulated monocyte cell types (colors) versus their unstimulated counterparts. Dashed line, Bonferroni-adjusted p value < 0.05.

(D) CellWalker2 identifies cell-type-specific TFs. Each row is a TF expressed in PBMCs and each column is a cell type from 10 \times Genomics multi-omics PBMC dataset. The size of the dot represents Z score. The color of each square is the standardized gene expression of a TF in a particular cell type. TF names colored red are universal stripe factors and blue are other stripe factors.³³

(E) CellWalker2's Z scores are more correlated with the number of ChIP-seq peaks in various cell types and show competitive sensitivity in detecting cell type-specific TFs. The upper left corner shows the Spearman's rank and Pearson correlation coefficients between log10 number of ChIP-seq peaks (horizontal axis) and TF mapping scores from each software tools: Z scores (CellWalker2), log10 adjusted p values (ArchR and Signac), area under the curve (SCENIC+), or rank (SIMBA). For all the methods, a larger score means greater specificity of that TF to a given cell type. Each dot is the maximum score of a TF in a cell class and the color reflects the cell class. The last graph shows the sensitivity of each method at

ontology. Applying this approach to PBMC Z scores reveals four clusters corresponding to different lineages and states: plasma and B cells, monocytes, natural killer (NK) cells, cytotoxic T cells, and other T cells (Figure S12). Z score vectors also help users to interpret uncertainty about how each cell-type label relates to the second ontology (Figure 3B). When there is a strong 1:1 mapping, CellWalker2's Z score is highest for that leaf node (e.g., platelets and pDCs), whereas an ancestral node representing a broad cell type scores higher when there is more ambiguity (e.g., cycling cells). Furthermore, Z scores capture differences between stimulated cell types and their unstimulated counterparts (e.g., baseline versus interferon-stimulated monocytes and NK cells in Figure 3C), whereas treeArches assigns similar probabilities. We see this flexibility as an advantage, but users should be aware that it decreases the chance of seeing one highly specific cell-type mapping.

Mapping TFs to cell types

We next applied CellWalker2 to the human PBMC multi-omics data from 10 \times Genomics with the goal of identifying cell-type-specific TFs. We analyzed these results alongside TFs discovered using ArchR, Signac, SCENIC+, and SIMBA, validating cell-type associations with known TF roles, TF expression patterns, and blood ChIP-seq data.²⁹ ArchR and Signac are examples of traditional TF motif enrichment analyses that call DARs and then treat them equally. In contrast, CellWalker2 assigns higher Z scores to TFs whose motifs are in regions connected to many cells in a given cell type. SCENIC+ and SIMBA identify cell-type-specific regulatory modules by combining information about TF expression, motif accessibility, and target gene expression.

We first assessed the differences among methods using known relationships of TFs to cell types (Figures 3D and S14). CellWalker2 and SCENIC+ identify many of the same cell-type-specific TFs, including well-known regulators, such as *TBX21* and *EOMES* in CD8 TEM and NK cells, *LEF1* for T cells, *TCF7* in T cells, and *EBF1* in B cells (Figure S15), but SCENIC+ better distinguishes related motifs, such as *SPI1* and *SPIB*, having higher scores in monocytes and B cells, respectively. Signac and ArchR, in contrast, show much broader cell-type mappings for lineage-specific TFs like *TBX21*, *EOMES*, and *EBF1*. In contrast, SIMBA is more conservative in mapping TFs, failing to identify any TFs specific to B cells and DCs and missing many known regulators.

Next, we compared TF expression levels with cell-type mappings (Figures 3D and S14), finding that CellWalker2 and ArchR scores tend to correlate with expression, while those from Signac and SIMBA are less correlated. SCENIC+ shows reasonable correlation, but some TFs have very low expression in the associated cell type (e.g., *ETS1*, *KLF2*, and *LEF1* in B cells), suggesting these could be false positives or TFs that function at very low expression levels.

Third, we evaluated the methods' TF mappings using ChIP-seq data.²⁹ CellWalker2 and ArchR have the highest correlation with the number of ChIP-seq peaks in various cell types, while SIMBA has the lowest (Figure 3E). CellWalker2 has competitive sensitivity, though Signac's is higher since it makes the most mappings (Figure S14), some of which are likely false-positive

mappings given their low correlation with TF expression and known roles.

A distinctive feature of CellWalker2 is its ability to place TFs on the cell type hierarchy, going beyond mappings to individual cell types (Figure 3F). Ancestral nodes receive high Z scores when TFs play regulatory roles in multiple related subtypes or the resolution from scATAC-seq data is insufficient to distinguish between related cell types despite their showing distinct marker gene expression in the data used to build the cell-type ontology. For example, CellWalker2 identifies *EBF1* as an active regulator for the ancestral node of all B cells, *POU2F1* in B cells, DCs, and monocytes, *LEF1* and *TCF7* in different clades of T cells, and *TBX21* and *EOMES* in the ancestor of NK and effector memory T cell types. In other cases, the activity of TFs does not fully correspond with the hierarchical structure of the cell-type ontology. For instance, *THAP11* and *THAP1* are significant in multiple types of T cells and NK cells. Collectively, these hierarchical mappings between TFs and PBMC types reveal similarities between TFs that are active in the same cell types, and between cell types that share many TFs.

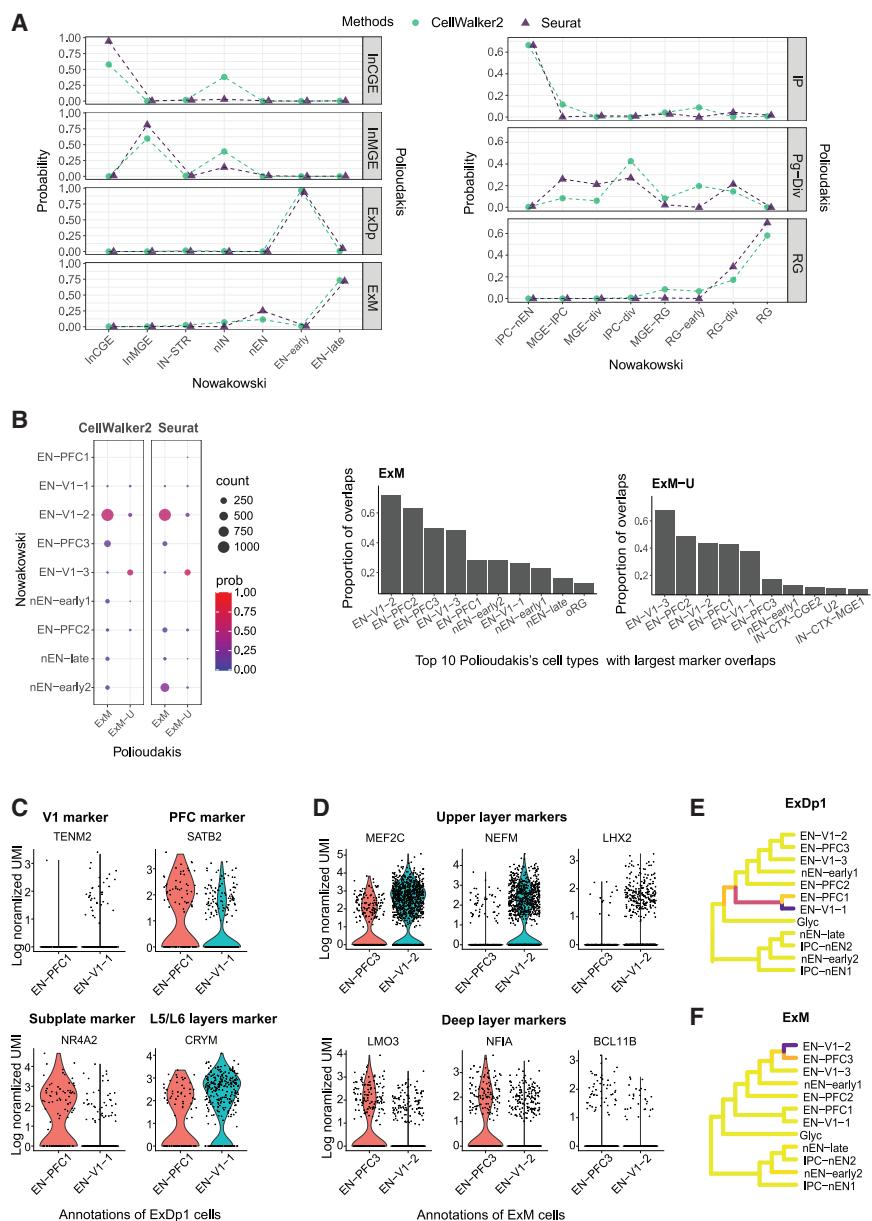
Human developing cortex

The developing cortex is complex organ with numerous interrelated cell types. Cell-type labels often vary across different studies, impeding comparisons and integrative analyses. Here, we use two independently collected human developing cortex scRNA-seq datasets^{34,35} with different cell type classification criteria (**STAR methods**) to illustrate how CellWalker2 addresses this challenge.

Labeling cells

We used CellWalker2 and Seurat to annotate cells in the Polioudakis et al.³⁵ scRNA-seq dataset using the cell types defined in Nowakowski et al.³⁴ The resulting cell annotations were compared with the original cell type labels in Polioudakis et al.³⁵ As Seurat cannot use a hierarchical cell type ontology, we first ran CellWalker2 without considering cell-type relationships. Although the top-mapped cell types are mostly consistent between methods (Figure 4A), CellWalker2 additionally provides probabilistic mappings to related cell types (Figure S15A).

We used excitatory neurons to further explore the challenges of cross-study cell annotation because both datasets have multiple subtypes of excitatory neurons from different layers, areas, and developmental stages. The Polioudakis et al.³⁵ dataset includes five types of excitatory neurons: two deep layer subtypes (ExDp1 and ExDp2), maturing (ExM), upper layer enriched maturing (ExM-U), and migrating (ExN). The Nowakowski et al.³⁴ dataset additionally separates excitatory neurons from different brain areas, such as early-born deep layer/subplate excitatory neurons in visual versus PFC (EN-V1-1 and EN-PFC1, respectively). CellWalker2 successfully mapped ExM to early and late-born excitatory neurons (EN-PFC3 and EN-V1-2), while mapping most ExM-U cells to late-born excitatory neurons (EN-V1-3) (Figure 4B), consistent with upper layer neurons developing later³⁴ and high marker overlap between these pairs of cell types. In contrast with CellWalker2, Seurat maps a large portion of ExM cells to newborn excitatory neurons (nEN-early2) (Figure 4B), which is questionable given that these cells are in different maturation stages and the fact that ExM shares more



marker genes with EN-V1 and EN-PFC cell types compared with nEN-early2 (Figure 4B). This is consistent with our simulation results showing that Seurat is biased toward prevalent cell types, because nEN-early2 has the highest number of cells among all the excitatory neurons (~27%).

When the reference labels contain information absent from the query cell ontology, CellWalker2's cell-to-label mappings can refine our understanding of cell types in the query dataset. For example, ExDp1 cells are divided into deep layer excitatory neurons from different brain areas (EN-V1-1 versus EN-PFC1) based on their mappings to the Nowakowski et al.³⁴ dataset (Figure S16A). Although all of these labels indicate early-born deep-layer excitatory neurons, the more than 1.5-fold expression differences of area markers, including *KCNJ6* and *SATB2*

Figure 4. Labeling cells with a reference cell type hierarchy using human developing cortex scRNA-seq data

(A) CellWalker2 (emerald circles) and Seurat (violet triangles) were used to label cells from Polioudakis et al.³⁵'s dataset with the cell types from Nowakowski et al.³⁴ Vertical axes show the proportion of cells in each Polioudakis et al.³⁵ cell type (rows) that are mapped to each label from Nowakowski et al.³⁴ (horizontal axis). Similar cell types are grouped together (Tables S1 and S2). (Left) Cell types are arranged from inhibitory to excitatory neurons. (Right) Cell types are arranged from intermediate progenitors (IPs) to RG.

(B) CellWalker2's cell annotations show expected differences across subtypes of maturing excitatory neurons and high proportions of marker gene overlap between query cells and the annotated cell types. (Left) Heatmap showing annotations for ExM and ExM-U cells. The color of each dot represents the proportion of mapped cells and the size represents the number of mapped cells (labels with less than five cells not shown). (Right) Barplots showing the proportion of positive markers of ExM and ExM-U that overlap with positive markers of cell types in Nowakowski et al.³⁴ The top 10 cell types with largest marker overlaps are shown.

(C) ExDp1 cells annotated to either EN-PFC1 or EN-V1-1 by CellWalker2 show differential expression of layer or region specific marker genes.

(D) ExM cells annotated to either EN-PFC3 or EN-V1-2 by CellWalker2 show differential expression of upper or deep layer specific genes.

(E and F) CellWalker2 annotates ExDp1 and ExM cells in Polioudakis et al.³⁵'s dataset by the cell type hierarchy from Nowakowski et al.³⁴ Colors of the branches represent the percentage of cells annotated to each cell type on the tree. Explanation of abbreviations in Tables S1 and S2.

for PFC and *TENM2* for V1, validate the annotation of these cells as originating from distinct areas (Figure 4C). Moreover, these area-divided subgroups also express different laminar layer markers, including a 4-fold change for the subplate

marker *NR4A2* and a 2-fold change for the layer 5/6 marker *CRYM* (Figure 4C). They also show different timings of neuronal cell birth (Figure 4D in³⁴). Another example is ExM, for which subgroups of cells are annotated as two different cell types that show differential expression of markers for upper versus deep layer clusters identified in Polioudakis et al.³⁵ (Figure 4D). By taking account of the cell-type hierarchy, CellWalker2 can identify cells at an intermediate state between cell types. For instance, some ExDp1 and ExM cells were mapped to the ancestor cell types (Figures 4E and 4F), and the UMAP and expression level of markers support that this group of cells represents an intermediate state between two tip cell types (Figures S16B and S16C). Together, these results highlight the power of our cross-study cell annotation for refining

cell classifications and underscore the relationship between neuronal cell types and their migration patterns during development.

Comparing cell type hierarchies

We next applied CellWalker2, treeArches, and MARS to map cell-type hierarchies between the two developing human cortex datasets. We mapped the cell types in the Polioudakis et al.³⁵ dataset onto those from Nowakowski et al.³⁴ and also flipped them to evaluate whether our findings were consistent in both directions. CellWalker2 and treeArches showed similar results overall, but MARS was quite different (Figure S17), making errors for several rare, non-neuronal cell types.

Places where results from CellWalker2 and treeArches differed revealed the effects of several of our modeling choices. First, these tools handle cell types in the query ontology that do not strongly match a single cell type in the reference ontology differently; treeArches treats this as a new cell type and assigns it to the root node, whereas CellWalker2 generates a Z score for all ancestral and leaf nodes. For example, CellWalker2 maps (highest Z score) ExN to an ancestor node (Figure 5A), suggesting that excitatory neurons are more broadly defined in Polioudakis et al.³⁵ than in Nowakowski et al.³⁴ This is supported by cell-to-cell distances from ExN to other cell types (Figures 5B and S18). We observed that such mappings to internal nodes depend on CellWalker2's use of a null distribution, because influence scores do not show the same adaptability and are maximal at the leaf nodes (Figure S19). In contrast, treeArches maps ExN to the root node (highest probability), and terminal cell types that have higher prevalence. For both methods, similar behaviors are observed for two types of inhibitory neurons (InCGE and InMGE) (Figure 5A). If we manipulate the cell type resolution of the Nowakowski et al.³⁴ ontology, for instance by amalgamating excitatory neuron cell types, treeArches assigns higher scores to the combined cell type and reduces its score at the root node (Figures S20A and S20B). Alternatively, if we remove cell types with the highest scores for ExN, treeArches' scores increase for other cell types, including a prevalent cell type from a different developmental stage (Figures S20C and S21A). In contrast, CellWalker2's top cell types stay the same with only minor decreases in Z scores (Figures S20C and S21B). This suggests a second difference between CellWalker2 and treeArches: inclusion of closely related cell types in a dataset may lower treeArches' mapping scores to each of them, particularly rare cell types due to treeArches' sensitivity to compositional bias, but encourages CellWalker2's mapping to all these cell types and their ancestor. These findings indicate that treeArches may be better able to map cell types when the two ontologies have comparable resolution, while CellWalker2's use of internal nodes enables mappings between fine-resolution and broad cell types.

As another example, CellWalker2 maps two types of RG in Nowakowski et al.³⁴ (oRG, vRG) to all nodes in the RG subtree with roughly equal Z scores, but maps another type of RG (tRG) weakly in the Polioudakis et al.³⁵ ontology, which does not have a tRG cell type (Figure 5C). In contrast, the ancestor nodes of all RG in these two datasets show strong correspon-

dence with each other (Figures 5C and S22), indicating agreement at this broader level of classification. These findings indicate that CellWalker2's ability to map to internal nodes of cell-type hierarchies resolves problems that arise when cell-type ontologies contain different cell types and non-unique cell-type relationships. For the two groups of RG-like intermediate progenitor cells in Nowakowski et al.³⁴ (IPC-div1 and IPC-div2), although the Nowakowski et al.³⁴ labels do not reference the cell cycle, the mapping result from CellWalker2 suggests that G2M versus S phase is one differentiating factor between them (Figure 5C). The reverse mapping of PgG2M and PgS to the cell types in Nowakowski et al.³⁴ supports this conclusion, while also identifying a few other types of dividing and progenitor cell types with potential enrichment for cells in G2M or S phase (Figure S22). Thus, CellWalker2's cell-type comparisons facilitate the interpretation of cell types in one ontology when the other ontology carries additional information about cell state.

Finally, since the threshold for calling marker genes is usually a subjective user choice, we showed that CellWalker2 is not very sensitive to the number of marker genes for each cell type, and that including more markers does not substantially alter the results (Figure S23A). As single-cell data are often noisy, we also showed the robustness of CellWalker2 upon varying cell-type composition and adding different intensities of noise to the cell-cell similarity matrix (Figure S23B). Altogether, these analyses of scRNA-seq from the developing human brain demonstrate the robustness of CellWalker2, the importance of using a null distribution to compute Z scores, and the flexibility and potential for new understanding of cell states that arises from using all nodes of cell-type hierarchies.

Mapping TFs to cell types

We mapped TF motifs to cell types, finding that CellWalker2 shows greater cell-type specificity than Signac (Figure S24A and S24B). Many of the top TFs on different lineages of the cell-type tree (Figure 5D) have known roles in those cell types, including DLX2, a modulator of neuron versus oligodendrocyte development⁴³ that maps to inhibitory neurons, EBF1, which contributes to pericyte cell commitment⁴⁴ and maps to pericytes, and EMX2, a promoter of neurogenesis³⁹ that maps to progenitor cells. However, only some TFs have expression levels correlated with CellWalker2's Z scores (Figures S24A and S24C), as TFs with similar motifs and/or co-occur within similar pREs often have similar Z score profiles (Figure S25A and S25B). Therefore, we repeated the mapping of TFs to cell types using ChIP-seq peaks rather than motifs as the annotations (Figures 5E and S26A). Compared with the motif results, the ChIP-seq based mappings were more correlated with gene expression (Figure S26B), especially for TFs like OLIG2 and SOX21 that had a high number of motifs outside of ChIP-seq peaks (Figures S26C–S26E). ChIP-seq peaks also helped CellWalker2 to differentiate TFs with similar motifs. For example, GATA2 and GATA3 motifs are highly overlapping in the genome, but their ChIP-seq peaks are more distinct and hence map to different cell types (Figure S26F). Thus, when ChIP-seq data are available, we recommend it over motif instances for mapping TFs to cell types. However, since ChIP-seq experiments in brain are limited, motifs provide a means to explore a larger set of TFs (Figure S24C).

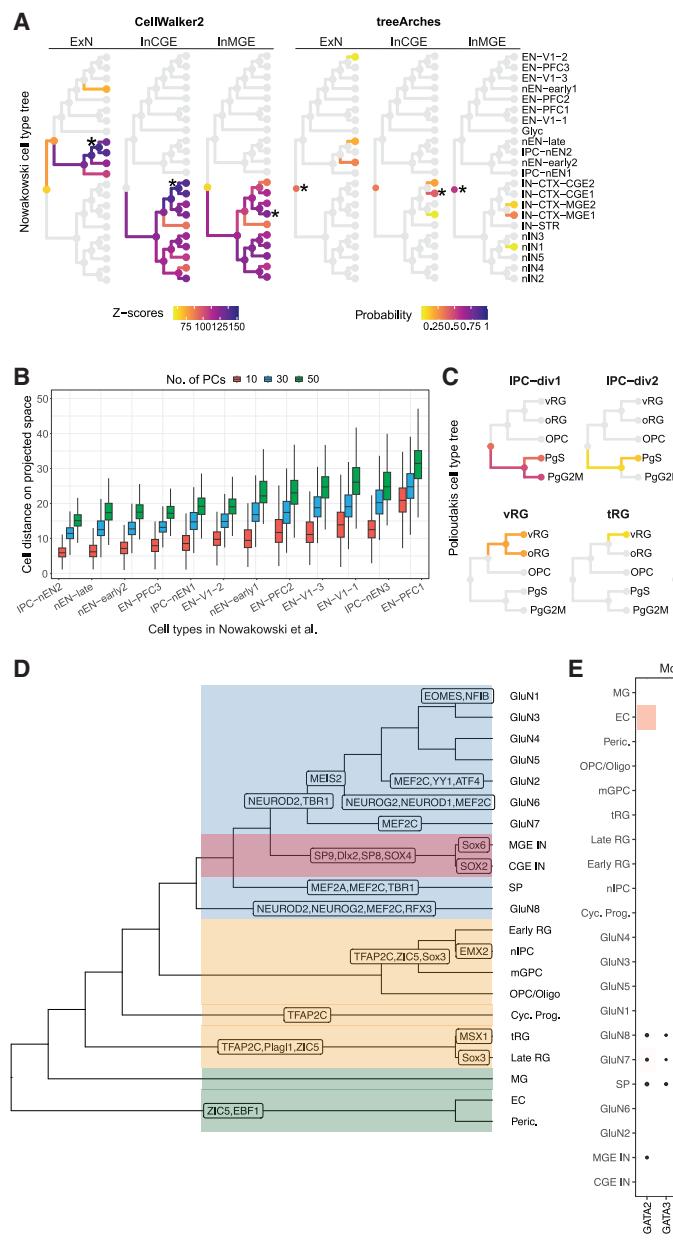


Figure 5. Comparing cell-type hierarchies using scRNA-seq data in human developing cortex

(A) CellWalker2 and treeArches were used to map the ExN, InCGE, and InMGE cell types in Polioudakis et al.³⁵'s dataset onto the cell type hierarchy from Nowakowski et al.³⁴ The nodes with the highest scores are indicated by *. CellWalker2 maps ExN to multiple similar cell types, while treeArches maps ExN to the root node, but assigns no probability to IPC-nEN2, most likely because it is composed of fewer cells (4% versus >20% for other cell types in the same clade). CellWalker2 maps InMGE to one type of MGE inhibitory neurons, whereas InCGE is mapped to the ancestor of two types of CGEs. For treeArches, the root node has a high probability for both InMGE (56%) and InCGE (33%) probably because of the uncertainty about 1:1 mapping. The treeArches rejection probabilities (not shown) are 0.1, 0.03, and 0.02 for ExN, InCGE, and InMGE, respectively.

(B) ExN is similar to the cell types mapped by CellWalker2, i.e., IPC-nEN2, nEN-late and nEN-early2. Euclidean distance between ExN cells from Polioudakis et al.³⁵ and excitatory neurons of various types from Nowakowski et al.³⁴ Boxplot shows the distribution of cell-to-cell distances computed on the top principal components (PCs) using a subsample of 2,000 ExN cells. The cell types on the horizontal axis are ordered by median cell-to-cell distance using 30 PCs.

(C) Mapping RG and intermediate progenitor (IPC-div) cell types in Nowakowski et al.³⁴ onto the cell-type hierarchy of Polioudakis et al.³⁵ using CellWalker2. IPC-div1 and IPC-div2 are two types of RG-like intermediate progenitor cells, with IPC-div as their parent; vRG, tRG and oRG are three types of RG, with RG as their ancestor. For (A) and (C), the branches and nodes are colored by Z scores for CellWalker2 and by mapping probabilities for treeArches (Z score > 75 and probability > 0.005 are shown). Explanation of abbreviations in Tables S1 and S2.

(D) Cell-type-specific TFs in human developing cortex based on sequence motifs. Top expressed TFs with the largest CellWalker2 Z scores for tips and internal nodes of the cell type tree are shown. Many of the TFs predicted to be cell

type specific have known roles regulating neurodevelopment in specific brain regions.^{36–42} Shaded regions on the cell-type tree reflect different classes of cell types.

(E) Cell-type-specific TFs in human developing cortex based on ChIP-seq peaks. Each column is a TF and each row is a cell type. The size of the dot is Z score, and only Z scores of >5 are shown. The color of the box is the log normalized gene expression of a TF in a particular cell type. As expected, CellWalker2's labeling is more specific for ChIP-seq peaks compared with sequence motifs, and it better matches TF gene expression.

Cross-species comparison of neurons in the motor cortex

We applied CellWalker2 to BICCN scRNA-seq and SNARE-seq data from adult human, marmoset, and mouse motor cortex samples.⁴⁵ Although Bakken et al.⁴⁵ mapped the cell types from each species to a cross-species consensus taxonomy, they did not assess statistical significance, nor did they investigate relationships across levels of the cell-type hierarchy. To explore these gaps, we applied CellWalker2 to investigate spe-

cies individually and jointly, using labels based on two levels of granularity (cell types and subclasses) (STAR methods). These analyses showed that CellWalker2 can identify pathways and TFs with shared versus divergent evolutionary patterns of cell type specificity.

Mapping subtypes of inhibitory neurons across species

First, we ran CellWalker2 to compare human and marmoset inhibitory neurons at the subclass level (Figure 6A), observing that cells from one species map to the corresponding subclass

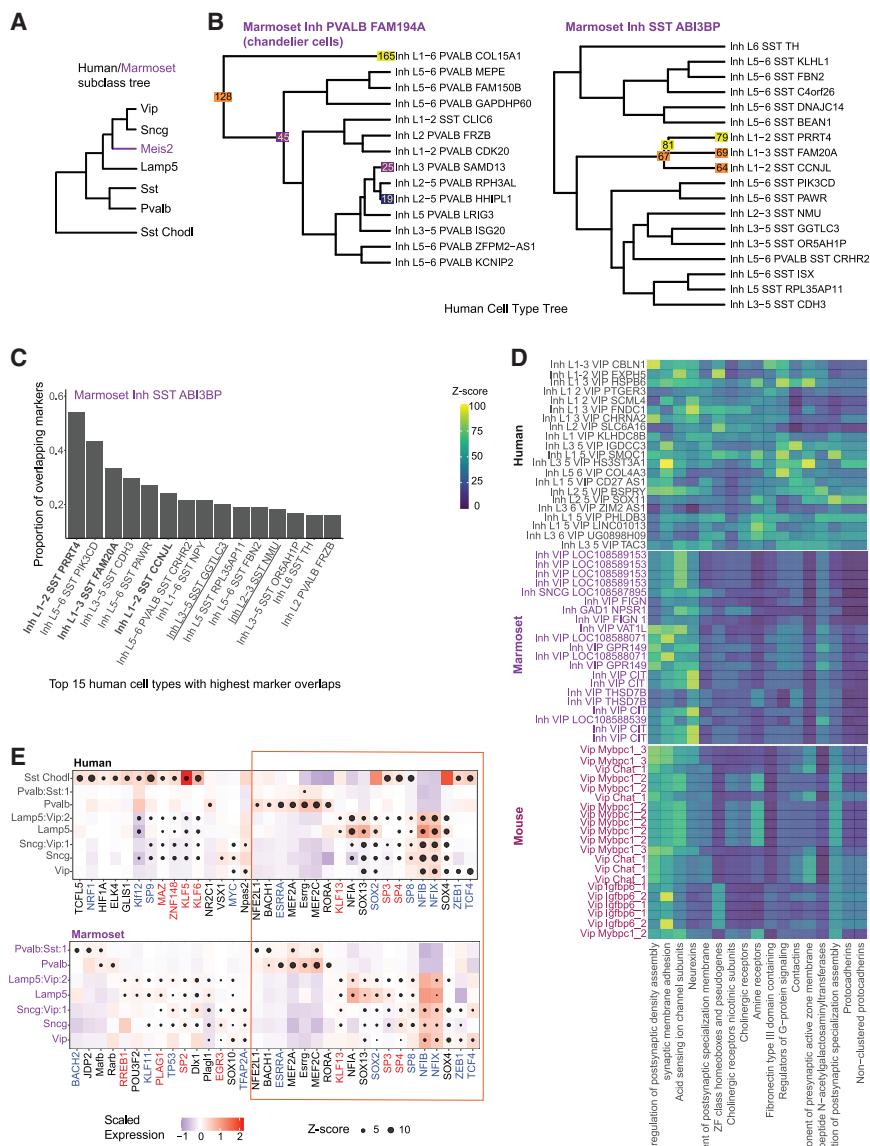


Figure 6. Cross-species analyses with CellWalker2

(A) The marmoset and human consensus taxonomy of cell subclasses.⁴⁵ Marmoset has Meis2 cells (purple), which are not present in the human ontology. Other subclasses are shared.

(B) Marmoset cell types map onto the human cell type tree with either clear one-to-one relationships or with similar Z scores for all nodes in a subtree of related cell types (e.g., Inh PVALB FAM194A and Inh SST ABI3BP). Human Inh PVALB (left) and SST (right) subtrees are shown. The top 5 cell types (largest Z scores) are shown. The number and color of each node both reflect the magnitude of the node's Z score. Cell-type names are based on two marker genes.⁴⁵ Human cell-type names also contain laminae layer information.

(C) High proportions of cell-type markers overlap between marmoset Inh PVALB FAM194A cells and mapped human cell types. The top 15 human cell types with the greatest overlap of positive marker genes are shown. The three cell types with the highest Z scores are shown as bold and the other two cell types in the human and marmoset consensus cluster (Sst_3) defined in Bakken et al.⁴⁵ are underlined.

(D) Gene sets that are activated in human VIP cells show different cell-type specificity across species. Heatmaps show the Z scores (color scale) for mappings between gene sets and cell types in human (top), marmoset (middle), and mouse (bottom). Some of the pathways are active in all three species (e.g., neurexins, acid-sensing ion channel subunits, synaptic and postsynaptic related pathways), while others are only active in human (e.g., protocadherins, cholinergic receptors, amine receptors, contactins, regulation of G-protein signaling and presynaptic active zone membrane). Columns, gene sets; rows, cell types. For each human cell type, the marmoset and mouse cell types with the highest CellWalker2 Z scores are shown.

(E) CellWalker2 identifies both consensus and unique cell subclass specific TFs (**STAR Methods**) in human and marmoset inhibitory neurons. Columns, TFs; orange box, consensus TFs between human and marmoset (TFs to the left are divergent); rows, node in cell subclass tree with internal nodes named by their two descendant nodes and depth in the tree. Rows are ordered identically in

both species, except the additional Sst chodl cell type in human. Sst is not shown as no significantly associated TFs. The color of each square is the standardized gene expression of a TF in a particular cell subclass. Red font, universal stripe factors; blue font, other stripe factors.

in the other species whenever one exists, and subclasses cluster into CGE-derived versus MGE-derived subclasses based on their Z scores (Figure S27). Then, we applied CellWalker2 to compare finer-grained cell types across species within each subclass. Some cell types in marmoset can be mapped to a single cell type in humans with a high Z score. For example, Inh PVALB FAM194A in marmoset mapped to Inh PVALB COL15A1 in humans, both of which are chandelier cells (Figure 6B). Although these two cell types have different cell-type markers in each species, they share similar expression pro-

files that enable CellWalker2 to connect them and distinguish them from other Pvalb cells. In contrast, some cell types receive similar Z scores for multiple cell types in the other species (Figure 6B). For instance, Inh SST ABI3BP is mapped to a subtree with three tips, all of which belong to a Sst cell cluster (Sst_3) in the consensus taxonomy of Bakken et al.⁴⁵ These high-scoring cell types are all present in upper layers and share many markers with Inh SST ABI3BP, while the other cell types in Sst_3 are associated with deep layers and share fewer markers (Figure 6C), suggesting that CellWalker2 refined the consensus

taxonomy. Overall, cell-type mapping results within subclasses show nested block-wise structure, indicating that consensus subgroups within subclasses of inhibitory neurons exist between human and marmoset (Figure S28). These results demonstrate that CellWalker2 can provide a more nuanced and hierarchical mapping between cell types than is possible with a consensus taxonomy.

Next, we mapped the entire cell-type hierarchies in human and marmoset (Figure S29). Cell types within the same subclass have higher mapping scores in general, but some cell types from different subclasses have detectable similarity (Figure S30). For example, human Sst chodl cells, grouped together with other Sst cell types in the human ontology, not only map to Sst chodl in marmoset, but also to another subtree of Sst cell types in a separate lineage, in which Inh SST MPP5 is also labeled as Sst chodl in the consensus taxonomy.⁴⁵ Another example is a marmoset Sst cell type (Inh PVALB SST LRRC6) that expresses both *PVALB* and *SST* and shows similarity not only to the human Sst subtree but even greater similarity to the Pvalb subtree, consistent with this cell type having features of both Pvalb and Sst cells. Thus, CellWalker2 successfully identifies cell type similarities across different lineages of a cell-type tree, including cases where a cell type in one species evolved to have features of multiple separate lineages in the other species.

Finally, we used CellWalker2 to integrate scRNA-seq data from human, marmoset, and mouse. By including cells from all three species in the same graph, we could compare cell type similarity between different pairs of species. We observed that the human-marmoset Z scores are in general stronger than the human-mouse ones (Figure S31A and S31B), as expected given evolutionary relationships. However, there are a few cases where a human cell type is more similar to some cell types in mouse than to any in marmoset (Figure S31C). This suggests that CellWalker2 could be used to nominate the best animal model to study a particular cell type (e.g., for disease research).

Comparison of cell-type-specific gene sets

We next used CellWalker2 to investigate gene sets with greater average expression in a particular cell type⁴⁵ and compare these between species (STAR methods). Compared with gene set enrichment analysis, our approach does not need to label the cells first, which can be difficult for closely related cell types. The Z score matrix from CellWalker2 shows a similar pattern across species (Figures S32A–S32C), meaning that the majority of gene sets have conserved cell-type specificity. However, we also observed differences in gene set activity across species. Some of the differences come from cell types that do not exist in all species (e.g., the Meis2 subclass in Figure S32C). Other differences come from gene sets with divergent expression within cell types that are shared between species (Figure S33). For example, VIP shows divergent cell-type specificity in human compared with both marmoset and mouse (Figure 6D). In contrast, the Pvalb subclass has strong conservation in gene set specificity, except that the small integrin binding ligand N-linked glycoprotein (SIBLING) family has higher Z score in human (Figure S33C). The SIBLING family has been shown to affect cellular proliferation, differentiation, and apoptosis, including survival of dopaminergic neurons.⁴⁶ These results highlight how CellWalker2 identifies gene sets with higher average

expression in a particular cell type without needing to select a differential expression threshold and in small sample sizes where power to detect differential expression is low.

Comparison of cell-subclass-specific regulatory regions and TFs

To compare regulatory elements across species, we applied CellWalker2 to label DARs from cell types in the motor cortex using human and marmoset single-nucleus chromatin accessibility and mRNA expression sequencing (SNARE-seq2) data.⁴⁵ We observed that human DARs have high Z scores for multiple cell types of the corresponding marmoset cell subclass (Figure S34). The limited number of marmoset cells makes it hard to do the reverse analysis, but our labeling of DARs suggests that human and marmoset have similar open chromatin signatures on the subclass level.

Next, we sought to investigate the upstream TFs that bind open chromatin regions and to identify conserved and divergent regulators in human versus marmoset. We used CellWalker2 to score the cell-type specificity of TF motifs within open chromatin regions for each species, filtering out TFs not expressed in the corresponding cell subclass, and compared the results between species (Figure 6E). Since the marmoset dataset has a much smaller sample size, Z scores are smaller in marmoset than that in human. Still, a lot of top TFs are shared between human and marmoset within similar cell subclasses (Figure 6E). Many of these TFs are stripe factors, which occupy regulatory regions broadly and have key roles in tissue-specific transcription.³³ CellWalker2 also discovered unique cell-type-specific TFs in each species. For human, these included *NPAS2*, *MYC*, and *VSX1* in Vip and Sncg cells (Figure 6E). *NPAS2* regulates GABAergic neurotransmission and associated with psychiatric disorders.⁴⁷ *MYC* promotes neuronal differentiation in developing neural tube,⁴⁸ and *VSX1* contributes to interneuron development.^{49,50} Performing a similar analysis with ArchR showed that CellWalker2 identifies more cell subclass-specific TFs (Figure S35).

DISCUSSION

CellWalker2 has several features that in combination enhance its performance. First, we estimate statistical significance via permutations that preserve edge distributions, so the algorithm overcomes the bias other methods have toward prevalent cell types (or annotations). The resulting Z scores are robust to cell-type composition, sequencing depth, number of marker genes, cell-to-cell graph uncertainty, and cell-type definitions. Second, we represent cell types as hierarchies and compute Z scores for all nodes in the hierarchy. This enables CellWalker2 to leverage cell type similarities across different lineages and to identify a broad cell-type mapping when a fine-resolution one does not exist. Hence, although CellWalker2 requires a label hierarchy as input, cell-type labels can be preliminary or loosely defined. CellWalker2's outputs can be used to refine the initial cell-type tree. Third, CellWalker2 incorporates a strategy to automatically tune model parameters to optimize performance. By tuning the weight of cell-to-label versus cell-to-cell edges, CellWalker2 can account for the unknown reliability of cell-to-label edges. While CellWalker2 is fairly robust to erroneous cell

labeling, comparing Z scores for different sets of labels can help a user to identify unusual labels and evaluate the labeling accuracy. Altogether, these design choices help CellWalker2 to mitigate various sources of noise in single-cell experiments.

CellWalker2 can integrate single-cell data from different experiments into a single graph. We showed that CellWalker2 is more robust to batch effects and dropout than competing methods and is unbiased in detecting associations (Figure S36). However, if strong batch effects exist, as might be observed when comparing tumors from different patients, the cells would be less connected between batches, creating a bottleneck of information flow between labels, which could make the method under-powered. Extreme batch effects could affect the identification of accurate marker genes, upon which CellWalker2 depends. But for most cross-species and cross-tissue comparisons, batch effects and dropout will not heavily influence CellWalker2's outputs.

CellWalker2's graph model was designed to be highly flexible. For example, it can utilize cells from different studies or with a mixture of different modalities measured. Furthermore, CellWalker2 does not assign cells to specific cell types; instead, it treats cells as nodes in an interconnected graph. This approach proves advantageous for complex tissues, such as overlapping cell states within the developing brain or fine-grained cell states in blood. Probabilistic labeling is also useful if the query dataset contains cell types not present in the reference, as might be observed when comparing tumor and healthy samples.

Looking ahead, it will be exciting to apply CellWalker2 to arbitrary cell type graphs, including discretized and continuous cell type trajectories. In the future, we plan to scale CellWalker2 to handle millions of cells, versus tens of thousands of cells in this study (Table S3). To mitigate false positive TF cell-type mappings due to shared motifs or motifs co-occurring in DARs, we used ChIP-seq peaks and/or required that TFs be expressed in given cell type with a positive correlation between expression and Z score. A more systematic approach is another focus for future work.

Limitations of the study

CellWalker2 does not output regulatory module components, such as regulatory regions and downstream target genes of a TF. Other methods, like SCENIC+, SIMBA, GLUE, FigR,⁵¹ and Pando,⁵² can identify regulatory modules. Another caveat is that the permutation schema in CellWalker2 does not directly utilize relationships between cell types, i.e., the edges to a cell are randomly assigned rather than preferentially assigned to related cell types, which may underestimate the SD of the influence score as permutations do not preserve correlations between cell types. We also found that CellWalker2 could potentially assign higher Z scores to rare cell types. When too few edges connect to a cell type label, the influence score may be close to zero and its SD could be underestimated. We suggest that users increase the rounds of permutation in this case. Finally, when labeling bulk-derived annotations, CellWalker2 relies on multi-omics data as a bridge to connect scRNA-seq and scATAC-seq. This may limit its use due to data availability, but with the growing popularity of multi-omics data, this limitation

is becoming less significant. In the absence of multi-omics data, existing methods combining scRNA-seq with scATAC-seq data by using chromatin accessibility in the gene regions (as in the original CellWalker²²) can be adopted.

RESOURCE AVAILABILITY

Lead contact

Requests for further information and resources should be directed to and will be fulfilled by the lead contact, Katherine S. Pollard (katherine.pollard@gladstone.ucsf.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- This paper analyzes existing, publicly available data. Accession numbers are listed in the [key resources table](#).
- All original code has been deposited at Zenodo: <https://doi.org/10.5281/zenodo.15106832> and is publicly available as of the date of publication.
- Any additional information required to reanalyze the data reported in this paper is available from the [lead contact](#) upon request.

ACKNOWLEDGMENTS

This research was supported by National Institute of Mental Health (NIMH) grant numbers U01MH116438 (to K.S.P.), R01MH109907 (to K.S.P.), R01MH123179 (to K.S.P.), and Additional Ventures. We also thank Alex Pollen, Tomasz Nowakowski, and Nadia Roan for discussions on the results; Ryan Corces for help on using ArchR; Sean Whalen for providing information on pREs; Amanda Everitt for help on ChIP-seq data; and all Pollard lab members for suggestions on this project.

AUTHOR CONTRIBUTIONS

K.S.P. conceptualized the study, contributed to data analysis and interpretation, and contributed to manuscript writing. Z.H. developed the method, conducted the experiments, analyzed the data, and contributed to manuscript writing. P.F.P. initialized the study, contributed to the study design, provided critical feedback on data analysis, and revised the manuscript. All authors read and approved the final version of the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

STAR METHODS

Detailed methods are provided in the online version of this paper and include the following:

- [KEY RESOURCES TABLE](#)
- [METHOD DETAILS](#)
 - Specifying edge weights for the graph
 - Computing influence score matrix and labeling cells
 - Computing Z score from the influence score matrix
 - Benchmarking single-cell multiple modality integration
 - Details for simulating scRNA-Seq data
 - Processing scRNA-Seq data in human developing cortex
 - Processing multimodal data in human developing cortex
 - Processing multispecies scRNA-Seq and SNARE-Seq2 data
 - Human PBMC scRNA-Seq datasets
 - 10x genomics multiomics human PBMC dataset
 - Comparing different methods with ChIP-Seq data in human PBMC
- [QUANTIFICATION AND STATISTICAL ANALYSIS](#)

- Using Signac to identify DARs and motifs enrichment
- Using archR to identify DARs and motifs enrichment
- Using Seurat to identify DEGs
- Statistical tests for labeling bulk-derived annotations

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at <https://doi.org/10.1016/j.xgen.2025.100886>.

Received: June 10, 2024

Revised: November 4, 2024

Accepted: April 28, 2025

Published: May 22, 2025

REFERENCES

1. Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W.M., 3rd, Zheng, S., Butler, A., Lee, M.J., Wilk, A.J., Darby, C., Zager, M., et al. (2021). Integrated analysis of multimodal single-cell data. *Cell* 184, 3573–3587.e29. <https://doi.org/10.1016/j.cell.2021.04.048>.
2. Stuart, T., Srivastava, A., Madad, S., Lareau, C.A., and Satija, R. (2021). Single-cell chromatin state analysis with signac. *Nat. Methods* 18, 1333–1341. <https://doi.org/10.1038/s41592-021-01282-5>.
3. Granja, J.M., Corces, M.R., Pierce, S.E., Bagdatli, S.T., Choudhry, H., Chang, H.Y., and Greenleaf, W.J. (2021). ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nat. Genet.* 53, 403–411. <https://doi.org/10.1038/s41588-021-00790-6>.
4. Domínguez Conde, C., Xu, C., Jarvis, L.B., Rainbow, D.B., Wells, S.B., Gomes, T., Howlett, S.K., Suchanek, O., Polanski, K., King, H.W., et al. (2022). Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science* 376, eabl5197. <https://doi.org/10.1126/science.abl5197>.
5. Chen, H., Ryu, J., Vinyard, M.E., Lerer, A., and Pinello, L. (2024). SIMBA: single-cell embedding along with features. *Nat. Methods* 21, 1003–1013. <https://doi.org/10.1038/s41592-023-01899-8>.
6. Bravo González-Blas, C., Minnoye, L., Papasokrati, D., Aibar, S., Hulselmans, G., Christiaens, V., Davie, K., Wouters, J., and Aerts, S. (2019). cis-topic: cis-regulatory topic modeling on single-cell ATAC-seq data. *Nat. Methods* 16, 397–400. <https://doi.org/10.1038/s41592-019-0367-1>.
7. Fang, R., Preissl, S., Li, Y., Hou, X., Lucero, J., Wang, X., Motamedi, A., Shiao, A.K., Zhou, X., Xie, F., et al. (2021). Comprehensive analysis of single cell ATAC-seq data with SnapATAC. *Nat. Commun.* 12, 1337. <https://doi.org/10.1038/s41467-021-21583-9>.
8. Liu, J., Gao, C., Sodicoff, J., Kozareva, V., Macosko, E.Z., and Welch, J.D. (2020). Jointly defining cell types from multiple single-cell datasets using LIGER. *Nat. Protoc.* 15, 3632–3662. <https://doi.org/10.1038/s41596-020-0391-8>.
9. Cao, Z.J., and Gao, G. (2022). Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nat. Biotechnol.* 40, 1458–1466. <https://doi.org/10.1038/s41587-022-01284-4>.
10. Brbić, M., Zitnik, M., Wang, S., Pisco, A.O., Altman, R.B., Darmanis, S., and Leskovec, J. (2020). MARS: discovering novel cell types across heterogeneous single-cell experiments. *Nat. Methods* 17, 1200–1206. <https://doi.org/10.1038/s41592-020-00979-3>.
11. Lotfollahi, M., Naghipourfar, M., Luecken, M.D., Khajavi, M., Büttner, M., Wagenstetter, M., Avsec, Ž., Gayoso, A., Yosef, N., Interlandi, M., et al. (2022). Mapping single-cell data to reference atlases by transfer learning. *Nat. Biotechnol.* 40, 121–130. <https://doi.org/10.1038/s41587-021-01001-7>.
12. Kan, Y., Qi, Y., Zhang, Z., Liang, X., Wang, W., and Jin, S. (2025). Integration of unpaired single cell omics data by deep transfer graph convolutional network. *PLoS Comput. Biol.* 21, e1012625. <https://doi.org/10.1371/journal.pcbi.1012625>.
13. Xu, C., Lopez, R., Mehlman, E., Regier, J., Jordan, M.I., and Yosef, N. (2021). Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. *Mol. Syst. Biol.* 17, e9620, arXiv. <https://doi.org/10.1525/msb.20209620>. <https://www.embopress.org/doi/abs/10.1525/msb.20209620>. <https://www.embopress.org/doi/pdf/10.1525/msb.20209620>.
14. La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K., Kastriti, M.E., Lönnberg, P., Furlan, A., et al. (2018). RNA velocity of single cells. *Nature* 560, 494–498. <https://doi.org/10.1038/s41586-018-0414-6>.
15. Lange, M., Bergen, V., Klein, M., Setty, M., Reuter, B., Bakhti, M., Lickert, H., Ansari, M., Schniering, J., Schiller, H.B., et al. (2022). CellRank for directed single-cell fate mapping. *Nat. Methods* 19, 159–170. <https://doi.org/10.1038/s41592-021-01346-6>.
16. Lynch, A.W., Theodoris, C.V., Long, H.W., Brown, M., Liu, X.S., and Meyer, C.A. (2022). MIRA: joint regulatory modeling of multimodal expression and chromatin accessibility in single cells. *Nat. Methods* 19, 1097–1108. <https://doi.org/10.1038/s41592-022-01595-z>.
17. Michielsen, L., Lotfollahi, M., Strobl, D., Sikkema, L., Reinders, M.J.T., Theis, F.J., and Mahfouz, A. (2023). Single-cell reference mapping to construct and extend cell-type hierarchies. *NAR Genom. Bioinform.* 5, lqad070, arXiv. <https://doi.org/10.1093/nargab/lqad070>. <https://academic.oup.com/nargab/article-pdf/5/3/lqad070/51052048/lqad070.pdf>.
18. Xu, C., Prete, M., Webb, S., Jardine, L., Stewart, B.J., Hoo, R., He, P., Meyer, K.B., and Teichmann, S.A. (2023). Automatic cell-type harmonization and integration across human cell atlas datasets. *Cell* 186, 5876–5891.e20. <https://doi.org/10.1016/j.cell.2023.11.026>. <https://www.sciencedirect.com/science/article/pii/S0092867423013120>.
19. Heumos, L., Schaaf, A.C., Lance, C., Litinetskaya, A., Drost, F., Zappia, L., Lücke, M.D., Strobl, D.C., Henao, J., Curion, F., et al. (2023). Best practices for single-cell analysis across modalities. *Nat. Rev. Genet.* 24, 550–572. <https://doi.org/10.1038/s41576-023-00586-w>.
20. Schep, A.N., Wu, B., Buenrostro, J.D., and Greenleaf, W.J. (2017). chrom-var: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat. Methods* 14, 975–978. <https://doi.org/10.1038/nmeth.4401>.
21. Bravo González-Blas, C., De Winter, S., Hulselmans, G., Hecker, N., Matetovic, I., Christiaens, V., Poovathingal, S., Wouters, J., Aibar, S., and Aerts, S. (2023). SCENIC+: single-cell multiomic inference of enhancers and gene regulatory networks. *Nat. Methods* 20, 1355–1367. <https://doi.org/10.1038/s41592-023-01938-4>.
22. Przytycki, P.F., and Pollard, K.S. (2021). Cellwalker integrates single-cell and bulk data to resolve regulatory elements across cell types in complex tissues. *Genome Biol.* 22, 61. <https://doi.org/10.1186/s13059-021-02279-1>.
23. Przytycki, P.F., and Pollard, K.S. (2024). Hierarchical annotation of eQTLs by H-eQTL enables identification of genes with cell type-divergent regulation. *Genome Biol.* 25, 299. <https://doi.org/10.1186/s13059-024-03440-2>.
24. Wen, C., Margolis, M., Dai, R., Zhang, P., Przytycki, P.F., Vo, D.D., Bhattacharya, A., Matoba, N., Jiao, C., Kim, M., et al. (2023). Cross-ancestry, cell-type-informed atlas of gene, isoform, and splicing regulation in the developing human brain. Preprint at medRxiv. 2023.03.03.23286706, arXiv. <https://doi.org/10.1101/2023.03.03.23286706>. <https://www.medrxiv.org/content/early/2023/03/06/2023.03.03.23286706>. <https://www.medrxiv.org/content/early/2023/03/06/2023.03.03.23286706.full.pdf>.
25. Hao, Y., Stuart, T., Kowalski, M., Choudhary, S., Hoffman, P., Hartman, A., Srivastava, A., Molla, G., Madad, S., Fernandez-Granda, C., and Satija, R. (2022). Dictionary learning for integrative, multimodal, and scalable single-cell analysis. Preprint at bioRxiv. <https://doi.org/10.1101/2022.02.24.481684>. <https://www.biorxiv.org/content/early/2022/02/26/2022.02.24.481684>. <https://www.biorxiv.org/content/early/2022/02/26/2022.02.24.481684.full.pdf>.
26. Przytycki, P.F., and Pollard, K.S. (2022). CellWalkR: an R package for integrating and visualizing single-cell and bulk data to resolve regulatory

- elements. *Bioinformatics* 38, 2621–2623, arXiv: <https://academic.oup.com/bioinformatics/article-pdf/38/9/2621/4987446/btac150.pdf>. <https://doi.org/10.1093/bioinformatics/btac150>.
27. Markenscoff-Papadimitriou, E., Whalen, S., Przytycki, P., Thomas, R., Biyyameen, F., Nowakowski, T.J., Kriegstein, A.R., Sanders, S.J., State, M.W., Pollard, K.S., and Rubenstein, J.L. (2020). A chromatin accessibility atlas of the developing human telencephalon. *Cell* 182, 754–769.e18. <https://doi.org/10.1016/j.cell.2020.06.002>.
28. Trevino, A.E., Müller, F., Andersen, J., Sundaram, L., Kathiria, A., Shcherbina, A., Farh, K., Chang, H.Y., Paşa, A.M., Kundaje, A., et al. (2021). Chromatin and gene-regulatory dynamics of the developing human cerebral cortex at single-cell resolution. *Cell* 184, 5053–5069.e23. <https://doi.org/10.1016/j.cell.2021.07.039>.
29. Hammal, F., de Langen, P., Bergon, A., Lopez, F., and Ballester, B. (2022). ReMap 2022: a database of human, mouse, drosophila and arabidopsis regulatory regions from an integrative analysis of DNA-binding sequencing experiments. *Nucleic Acids Res.* 50, D316–D325. <https://doi.org/10.1093/nar/gkab996>.
30. Ren, X., Wen, W., Fan, X., Hou, W., Su, B., Cai, P., Li, J., Liu, Y., Tang, F., Zhang, F., et al. (2021). COVID-19 immune features revealed by a large-scale single-cell transcriptome atlas. *Cell* 184, 1895–1913.e19. <https://doi.org/10.1016/j.cell.2021.01.053>.
31. Yoshida, M., Worlock, K.B., Huang, N., Lindeboom, R.G.H., Butler, C.R., Kumasaka, N., Dominguez Conde, C., Mamanova, L., Bolt, L., Richardson, L., et al. (2022). Local and systemic responses to SARS-CoV-2 infection in children and adults. *Nature* 602, 321–327. <https://doi.org/10.1038/s41586-021-04345-x>.
32. Shaffer, A.L., Shapiro-Shelef, M., Iwakoshi, N.N., Lee, A.H., Qian, S.B., Zhao, H., Yu, X., Yang, L., Tan, B.K., Rosenwald, A., et al. (2004). XBP1, downstream of blimp-1, expands the secretory apparatus and other organelles, and increases protein synthesis in plasma cell differentiation. *Immunity* 21, 81–93. <https://doi.org/10.1016/j.immuni.2004.06.010>.
33. Zhao, Y., Vartak, S.V., Conte, A., Wang, X., Garcia, D.A., Stevens, E., Kyoung Jung, S., Kieffer-Kwon, K.R., Vian, L., Stodola, T., et al. (2022). “stripe” transcription factors provide accessibility to co-binding partners in mammalian genomes. *Mol. Cell* 82, 3398–3411.e11. <https://doi.org/10.1016/j.molcel.2022.06.029>.
34. Nowakowski, T.J., Bhaduri, A., Pollen, A.A., Alvarado, B., Mostajo-Radji, M.A., Di Lullo, E., Haeussler, M., Sandoval-Espinosa, C., Liu, S.J., Velmeshev, D., et al. (2017). Spatiotemporal gene expression trajectories reveal developmental hierarchies of the human cortex. *Science* 358, 1318–1323. <https://doi.org/10.1126/science.aap8809>.
35. Polioudakis, D., de la Torre-Ubieta, L., Langerman, J., Elkins, A.G., Shi, X., Stein, J.L., Vuong, C.K., Nichterwitz, S., Georgian, M., Opland, C.K., et al. (2019). A single-cell transcriptomic atlas of human neocortical development during mid-gestation. *Neuron* 103, 785–801.e8. <https://doi.org/10.1016/j.neuron.2019.06.011>. <https://www.sciencedirect.com/science/article/pii/S0896627319305616>.
36. Harrington, A.J., Raissi, A., Rajkovich, K., Berto, S., Kumar, J., Molinaro, G., Raduazzo, J., Guo, Y., Loerwald, K., Konopka, G., et al. (2016). Mef2c regulates cortical inhibitory and excitatory synapses and behaviors relevant to neurodevelopmental disorders. *eLife* 5, e20059. <https://doi.org/10.7554/eLife.20059>.
37. Hevner, R.F., Shi, L., Justice, N., Hsueh, Y., Sheng, M., Smiga, S., Bulfone, A., Goffinet, A.M., Campagnoni, A.T., and Rubenstein, J.L. (2001). Tbr1 regulates differentiation of the preplate and layer 6. *Neuron* 29, 353–366. [https://doi.org/10.1016/S0896-6273\(01\)00211-2](https://doi.org/10.1016/S0896-6273(01)00211-2). <https://www.sciencedirect.com/science/article/pii/S0896627301002112>.
38. Bedogni, F., Hodge, R.D., Elsen, G.E., Nelson, B.R., Daza, R.A.M., Beijer, R.P., Bammler, T.K., Rubenstein, J.L.R., and Hevner, R.F. (2010). Tbr1 regulates regional and laminar identity of postmitotic neurons in developing neocortex. *Proc. Natl. Acad. Sci. USA* 107, 13129–13134, arXiv: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2910823/>. <https://doi.org/10.1073/pnas.1002285107>.
39. Brancaccio, M., Pivetta, C., Granzotto, M., Filippis, C., and Mallamaci, A. (2010). Emx2 and Foxg1 Inhibit Gliogenesis and Promote Neuronogenesis. *Stem Cell* 28, 1206–1218. <https://doi.org/10.1002/stem.443>. https://academic.oup.com/stemcl/article-pdf/28/7/1206/41951632/stemcl_28_7_1206.pdf.
40. Merzdorf, C.S. (2007). Emerging roles for zic genes in early development. *Dev. Dyn.* 236, 922–940, arXiv: <https://anatomypubs.onlinelibrary.wiley.com/doi/pdf/10.1002/dvdy.21098>. <https://doi.org/10.1002/dvdy.21098>. <https://anatomypubs.onlinelibrary.wiley.com/doi/abs/10.1002/dvdy.21098>.
41. Pastor, W.A., Liu, W., Chen, D., Ho, J., Kim, R., Hunt, T.J., Lukianchikov, A., Liu, X., Polo, J.M., Jacobsen, S.E., and Clark, A.T. (2018). Tfap2c regulates transcription in human naive pluripotency by opening enhancers. *Nat. Cell Biol.* 20, 553–564. <https://doi.org/10.1038/s41556-018-0089-0>.
42. Keck, M.K., Sill, M., Wittmann, A., Kumar, P.J., Stichel, D., Sievers, P., Wefers, A.K., Roncaroli, F., Hayden, J., McCabe, M.G., et al. (2022). OTHR-41. Amplification of the PLAG family genes—PLAGL1 and PLAGL2—is a key feature of a novel embryonal CNS tumor type. *Neuro Oncol.* 24, i156, arXiv: https://academic.oup.com/neuro-oncology/article-pdf/24/Supplement_1/i156/43945176/noac079.579.pdf. <https://doi.org/10.1093/neuonc/noac079.579>.
43. Petryniak, M.A., Potter, G.B., Rowitch, D.H., and Rubenstein, J.L.R. (2007). Dlx1 and dlx2 control neuronal versus oligodendroglial cell fate acquisition in the developing forebrain. *Neuron* 55, 417–433. <https://doi.org/10.1016/j.neuron.2007.06.036>.
44. Paganí, F., Tratta, E., Dell'Era, P., Cominelli, M., and Poliani, P.L. (2021). Ebf1 is expressed in pericytes and contributes to pericyte cell commitment. *Histochem. Cell Biol.* 156, 333–347. <https://doi.org/10.1007/s00418-021-02015-7>. <https://europepmc.org/articles/PMC8550016>.
45. Bakken, T.E., Jorstad, N.L., Hu, Q., Lake, B.B., Tian, W., Kalmbach, B.E., Crow, M., Hodge, R.D., Krienen, F.M., Sorensen, S.A., et al. (2021). Comparative cellular analysis of motor cortex in human, marmoset and mouse. *Nature* 598, 111–119. <https://doi.org/10.1038/s41586-021-03465-8>.
46. Bellahcène, A., Castronovo, V., Ogbureke, K.U.E., Fisher, L.W., and Fedarko, N.S. (2008). Small integrin-binding ligand n-linked glycoproteins (SIBLINGs): multifunctional proteins in cancer. *Nat. Rev. Cancer* 8, 212–226. <https://doi.org/10.1038/nrc2345>.
47. Ozburn, A.R., Kern, J., Parekh, P.K., Logan, R.W., Liu, Z., Falcon, E., Becker-Krai, D., Purohit, K., Edgar, N.M., Huang, Y., and McClung, C.A. (2017). NPAS2 regulation of anxiety-like behavior and GABA_A receptors. *Front. Mol. Neurosci.* 10, 360. <https://doi.org/10.3389/fnmol.2017.00360>.
48. Zinin, N., Adameyko, I., Wilhelm, M., Fritz, N., Uhlén, P., Ernfors, P., and Henriksson, M.A. (2014). MYC proteins promote neuronal differentiation by controlling the mode of progenitor cell division. *EMBO Rep.* 15, 383–391. <https://doi.org/10.1002/embr.201337424>.
49. Francius, C., Hidalgo-Figueroa, M., Debrulle, S., Pelosi, B., Rucchin, V., Ronellenfitsch, K., Panayiotou, E., Makrides, N., Misra, K., Harris, A., et al. (2016). Vsx1 transiently defines an early intermediate V2 interneuron precursor compartment in the mouse developing spinal cord. *Front. Mol. Neurosci.* 9, 145. <https://doi.org/10.3389/fnmol.2016.00145>.
50. Debrulle, S., Baudouin, C., Hidalgo-Figueroa, M., Pelosi, B., Francius, C., Rucchin, V., Ronellenfitsch, K., Chow, R.L., Tissir, F., Lee, S.K., and Clotman, F. (2020). Vsx1 and chx10 paralogs sequentially secure V2 interneuron identity during spinal cord development. *Cell. Mol. Life Sci.* 77, 4117–4131. <https://doi.org/10.1007/s00018-019-03408-7>.
51. Kartha, V.K., Duarte, F.M., Hu, Y., Ma, S., Chew, J.G., Lareau, C.A., Earl, A., Burkett, Z.D., Kohlway, A.S., Lebofsky, R., and Buenrostro, J.D. (2022). Functional inference of gene regulation using single-cell multi-omics. *Cell Genom.* 2, 100166. <https://doi.org/10.1016/j.xgen.2022.100166>. <https://www.sciencedirect.com/science/article/pii/S2666979X22001082>.

52. Fleck, J.S., Jansen, S.M.J., Wollny, D., Zenk, F., Seimiya, M., Jain, A., Okamoto, R., Santel, M., He, Z., Camp, J.G., and Treutlein, B. (2023). Inferring and perturbing cell fate regulomes in human brain organoids. *Nature* 621, 365–372. <https://doi.org/10.1038/s41586-022-05279-8>.
53. Fornes, O., Castro-Mondragon, J.A., Khan, A., van der Lee, R., Zhang, X., Richmond, P.A., Modi, B.P., Correard, S., Gheorghe, M., Baranasić, D., et al. (2020). JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* 48, D87–D92, arXiv: <https://academic.oup.com/nar/article-pdf/48/D1/D87/31697271/gkz1001.pdf>. <https://doi.org/10.1093/nar/gkz1001>.
54. Zappia, L., Phipson, B., and Oshlack, A. (2017). Splatter: simulation of single-cell RNA sequencing data. *Genome Biol.* 18, 174. <https://doi.org/10.1186/s13059-017-1305-0>.
55. Kuhn, H.W. (1955). The Hungarian method for the assignment problem. *Nav. Res. Logist. Q.* 2, 83–97. <https://doi.org/10.1002/nav.3800020109>.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Human developing cortex scRNA-Seq	Nowakowski et al. ³⁴	https://cells.ucsc.edu/cortex-dev/
Human developing cortex scRNA-Seq	Polioudakis et al. ³⁵	http://solo.bmap.ucla.edu/shiny/webapp/
Human developing cortex 10x Genomics multi-omics	Trevino et al. ²⁸	GEO: GSE162170
Mammalian motor cortex scRNA-Seq	BICCN, Bakken et al. ⁴⁵	https://data.nemoarchive.org/publication_release/Lein_2020_M1_study_analysis
Mammalian motor cortex SNARE-Seq2	BICCN, Bakken et al. ⁴⁵	https://data.nemoarchive.org/publication_release/Lein_2020_M1_study_analysis
Human PBMC scRNA-Seq	Ren et al. ³⁰ ; Yoshida et al. ³¹	https://celltypist.cog.sanger.ac.uk/Resources/Organ_atlas/Blood/Blood.h5ad
Human PBMC 10x Genomics multiomics data	10x Genomics	https://www.10xgenomics.com/datasets/pbmc-from-a-healthy-donor-granulocytes-removed-through-cell-sorting-10-k-1-standard-1-0-0
ChIP-Seq peaks	ReMap2022 ²⁹	https://remap.univ-amu.fr/storage/remap2022/hg38/MACS2/remap2022_nr_macs2_hg38_v1_0.bed.gz
JASPAR2020	Fornes et al. ⁵³	https://www.bioconductor.org/packages/release/data/annotation/html/JASPAR2020.html
CellWalker2 and codes for simulations and data analysis	This paper	Zenodo: http://doi.org/10.5281/zenodo.15106832
Software and algorithms		
CellWalker2	This paper	https://github.com/PFPrzytycki/CellWalkR/tree/cellwalker2 ; Zenodo: https://doi.org/10.5281/zenodo.15106832
Codes for simulations and data analysis	This paper	https://github.com/xyz111131/CellWalker2_supple_codes/tree/main ; Zenodo: https://doi.org/10.5281/zenodo.15106832
Seurat 4.3.0	Stuart et al. ⁵⁴ ; Hao et al. ¹	https://satijalab.org/seurat/
MARS	Brbić et al. ¹⁰	https://snap.stanford.edu/mars/
treeArches	Michielsen et al. ¹⁷	https://docs.scarches.org/en/latest/treeArches_pbmc.html
Signac 1.10.0	Stuart et al. ²	https://stuartlab.org/signac/
ArchR	Granja et al. ³	https://www.archrproject.com/
SCENIC+	Bravo Gonzalez-Bias et al. ²¹	https://github.com/aertslab/scenicplus
SIMBA	Chen et al. ⁵	https://simba-bio.readthedocs.io/en/latest/
GLUE	Cao et al. ⁹	https://github.com/gao-lab/GLUE
Splatter 1.12.1	Zappia et al. ⁵⁵	https://github.com/Oshlack/splatter

METHOD DETAILS

Specifying edge weights for the graph

The graph in CellWalker2 includes four types of edges: cell-to-cell, cell-to-label, cell-to-annotation and label-to-label (Figure S37). The label-to-label edges are optional (i.e., users can work with discrete cell types without defining their relationships to each other), and in this study we focused on the commonly used label-to-label graph structure of a binary, hierarchical tree, although CellWalker2 can utilize ontologies with other topologies.

For cell-to-cell edges, CellWalker2 first computes the cell-to-cell similarity based on gene expression and/or chromatin accessibility profile of the cells and constructs a K nearest neighbor (KNN) graph of all cells ($K = 200$ by default). The cell-to-cell edge weight is based on shared neighbors on the KNN graph. For cells with RNA-Seq data, the cell-to-cell similarity is based on gene expression profiles. CellWalker2 projects the gene expression data onto a low dimensional space ($\text{dim} = 30$ by default) using PCA and the cell-to-cell distance is the Euclidean distance in the latent space. The low dimensional space is defined using all the cells with RNA-Seq data.

We normalize the distances by their largest value to make them between 0 and 1 (denoted as d), and the cell-to-cell similarity is defined as $-\log(d)$. We also standardize the similarities to make them comparable with other data modalities. For cells with ATAC-Seq data, the cell-to-cell similarity is computed as the Jaccard or Cosine similarity of the vectors of peak presence/absence in each cell. We include peaks that appear in 0.2%–20% cells. Then we take the logarithm and standardize the similarities to make it comparable with other modalities. Although CellWalker2 provides several distance metrics for the similarity of scATAC-Seq profiles, including Jaccard, Cosine and latent semantic indexing (LSI), in our experiments, we observe that these metrics perform similarly. For cells with multiomics data (i.e., both RNA-Seq and ATAC-Seq), the cell-to-cell similarity is a weighted average of RNA-Seq and ATAC-Seq similarity (Default: 0.7 for RNA-Seq). If we have both unpaired scATAC-Seq and/or scRNA-Seq data and multiomics data, we use multiomics data as a bridge to connect cells with unpaired scATAC-Seq and/or scRNA-Seq data. We compute cell-to-cell similarity between cells from multiomics and unpaired scRNA-Seq data using gene expression profiles and from multiomics and unpaired scATAC-Seq data using chromatin accessibility profiles. Then we combine the cell-to-cell similarity matrices from various modalities of the data, construct a single KNN graph, and obtain a joint cell-to-cell graph as described above (Figure S38).

The cell-to-label edge weights are based on the gene expression of the marker genes of each cell type. The edge weight between a cell and a cell type label is defined as: $\frac{\sum_g w_g Z_g}{\sum_g |w_g|}$. The summation is over all marker genes for a cell type, and w_g is the weight of each marker gene which users can specify (by default, log fold change between mean expression level in one cell type versus the rest), and Z_g is the standardized gene expression in the cell. If cell type labels have graph structure, such as a hierarchical tree, we include internal nodes of the cell type ontology into the graph and connect internal nodes with tips. The edges reflect the (hierarchical) relationships between cell types (by default equally weighted).

Annotations can be gene sets or any genome coordinates of interest, such as a group of regulatory elements, genetic variants, TF sequence motif instances, or TF ChIP-Seq peaks. The edge weight between a cell and a gene set is the average standardized expression level of all genes in the gene set in the cell. On the other hand, the edge weight between a cell and genome coordinates is the overall accessibility of the genomic regions in the cell (i.e., we sum up all the reads in the ATAC-Seq peaks within those regions and normalize by the total number of reads). In our analyses of TFs, we used sequence motifs from JASPAR2020,⁵³ optionally intersected with bulk or cell type specific open chromatin regions, as well as ChIP-Seq peaks from ReMap2022.²⁹ Users may link a TF to cells based on accessibility of its motifs or its ChIP-Seq peaks. We connected a TF to a cell by summing up that cell's ATAC-seq reads within genomic regions that contain the TF motif or peak, normalized by the total number of reads.

Lastly, we tune the weights between cell-to-cell and other types of edges. If the task is to compare cell type labels, we optimize the cell homogeneity score²² such that labels can classify all the cells well. If the task is to label annotations, we minimize the entropy of influence scores between labels to cell types ('label entropy') such that the scores are more specific to certain cell types. The weight between label-to-label and cell-to-label edges is also tunable, depending on how deeply users wish to map cells on a hierarchical cell type ontology. If the weight between labels is large, the random walk will be more likely to reach the internal nodes of the tree or even the root node. For a binary tree, we recommend making the edge weights for going up versus down the tree unequal (e.g., up weight = 1, down weight = 0.1) to limit the random walks from going through the root node and back down toward distant tip nodes, passing information from one side of the tree to the other.

Computing influence score matrix and labeling cells

CellWalker2 does a random walk with restarts on the graph. It initiates walks from all nodes. During the random walk, the transition probability between steps is proportional to the edge weights. CellWalker2 computes the influence score matrix that represents the steady-state probabilities of reaching each node from every other node. Influence scores can be used to derive relationships between labels from two or more ontologies in the context of the cells, from labels to cells or annotations, and among cells. In particular, CellWalker2 assigns labels to query cells using a cell-to-label normalized influence score, which is the influence score for a given cell to a given label normalized by the sum of scores over all cells for that label.

Computing Z score from the influence score matrix

To estimate the Z score between any pair of cell type labels, we compare the observed value for the corresponding entry of the influence matrix with its null distribution assuming independence between cells and labels. To generate this null distribution, we permute the edges between cells and cell type labels. In detail, we re-sample the edge weights between cells and labels while keeping the marginal distributions of edge weights of each cell and each label stay close to its empirical distribution. First, to estimate the marginal distribution, we discretize the edge weights into 0 plus 10 equally-spaced intervals between 0 and 1 for each cell type label and 5 intervals for each cell as the number of labels connecting to each cell is smaller than the number of cells connecting to each label. We compute the proportion of edge weights in each interval for each cell or label. We tested discretizing into quantiles of edge weights, but the results were worse because the edge weight distributions usually have long tails. Second, for each cell type label, we resample the cells connecting to that label while preserving the number of cells in each edge weight interval. The probability to sample a cell is proportional to the probability that the edge weights from that cell lie in such interval. Finally, we uniformly sample values within each interval as the new edge weights. We also implemented different permutation strategies, either permuting the edges from cells to one of the cell type hierarchies or both. Permuting both generally generates larger Z-scores but the rankings do not

change much. If cell type labels of the reference dataset are closely related, we recommend users permute edges to the cell type hierarchy of the query dataset so that the null distribution under permutation maintains the correlation between cell types. As an alternative, we tested sub-sampling cells to estimate the standard deviation of influence scores for computing Z-scores, but the results were similar or worse than the current approach, as the cells are interconnected in the graph and hence not properly connected in sub-samples.

To generate a null distribution for labeling annotations, we permute the edges between cells and cell type labels using the strategy above. To estimate the Z-scores for cell type-specific TFs, we permute the genomic coordinates of TF motif instances. For each TF motif, we resample the genomic regions that contain it. The probability of choosing each genomic region is proportional to the number of motifs in that region. Then we recompute the edge weight between each TF and a cell. Compared to permuting cells to cell type labels, we found that this strategy performs better at identifying cell type-specific TFs. The reason might be that as connections from regions to cell types are kept when generating null distribution, any cell type bias of the chromatin accessibility of input regions will also be maintained in the null distribution so they will not be reflected in Z-scores. Moreover, by permuting TF-to-region edges when identifying cell type-specific TFs, CellWalker2 ensures the preservation of the correlation between cell types in the cell type tree.

We recommend sampling 50–100 permuted graphs. For each one, we compute an influence matrix using each randomized graph and estimate the Z score by $\frac{x - \text{mean}(x_{\text{perm}})}{\text{sd}(x_{\text{perm}})}$ for entry x in the influence matrix. The Z score reflects the statistical significance of the entries of influence matrix. To make a single cell type assignment, users can take the cell type with the maximum Z score. To generate a short list of probable cell types, users can choose a cutoff. The cutoff can be based on Z-scores or p-values that are generated from Z-scores assuming a standard normal distribution (as was done for Figure 3D). When comparing between conditions, such as in Figure 3C where cell type mapping is compared between stimulated and unstimulated cell states, the difference in Z-scores indicates the statistical significance of the differential mapping results.

Benchmarking single-cell multiple modality integration

We used 10x Genomics multiomics PBMC data with 10K cells for benchmarking. We split ATAC and RNA data of 6K cells to be single modality as input to SIMBA, GLUE and CellWalker2, and varied the number of dual-modality cells from 1K to 5K for CellWalker2. For SIMBA and GLUE, we followed their vignettes https://simba-bio.readthedocs.io/en/latest/multiome_10xpbmc10k_integration.html and <https://scglue.readthedocs.io/en/latest/tutorials.html>. For CellWalker2, we computed the normalized influence matrix from RNA cells to ATAC cells as the similarity between each RNA and each ATAC cell, which is the influence score normalized by the sum of scores from all RNA cells. We computed four metrics to show how well ATAC and RNA cells are integrated: 1) MAP: mean average precision, measuring the percentage of the neighboring cells sharing the same cell type labels, reflecting biological conservation of the integration results (adopted from GLUE); 2) Anchoring rank: for each cell, rank the distance between its paired cell (i.e., the same cell but split into ATAC-Seq and RNA-Seq) versus unpaired cells (adopted from SIMBA); 3) Anchoring distance: the distance ratio between paired ATAC and RNA cells versus unpaired cells (adopted from SIMBA); 4) Silhouette index: for each cell, compute the distance ratio between its paired cells to its closest unpaired cells (adopted from SIMBA).

Details for simulating scRNA-Seq data

In the first simulation without batch effect (easy), we use Splatter 1.21.1⁵⁴ to simulate 4000 cells with 1000 cells in each of four cell types, 1400 non-marker genes, 200 marker genes between cell type (1,2) and (3,4), 150 marker genes which are differentially expressed between 1 and 2 but not in 3 and 4, and another 150 marker genes between 3 and 4. Then, we split the cells into two equal-sized reference and query datasets. In the simulation scenario 2 (medium) with batch effect and dropouts, we simulated 1000 non-marker genes, 200 marker genes between cell type (1,2) and (3,4), 200 marker genes between 1 and 2 and between 3 and 4 respectively. We added batch effects between these two datasets using the default parameters in Splatter. We set ‘dropout.mid = 0, 2’ and ‘dropout.shape = -1, -0.5’ in Splatter for the reference and the query dataset respectively, such that the query dataset has more dropouts than the reference one. We also varied the number of cells of cell type 4 in the reference dataset (50, 100, 300 and 500 cells), but kept the number of cells as 500 for others in order to assess performance on rare cell populations. In the simulation scenario 3 (hard) with batch effect and more dropouts, we simulated 1200 non-marker genes, 100 marker genes between cell type (1,2) and (3,4), 150 marker genes between 1 and 2 and between 3 and 4 respectively with ‘de.facScale = 0.2’. We added batch effects between these two datasets using the default parameters in Splatter. We set ‘dropout.mid = 1, 2’ and ‘dropout.shape = -1, -0.5’ in Splatter.

To apply CellWalker2 for labeling cells, we either used cells from the query dataset only or combined cells from both datasets to generate cell graph. For the simulation varying the number of cells, we used cells from both datasets. The marker genes for each cell type were computed using Seurat given the true cell type labels of cells in the reference dataset. We filtered for genes with $|\log_2 FC| > 0.5$ and adjusted p -value < 0.05 using the default two-sided Wilcoxon Rank-Sum test with Bonferroni correction. We input the true cell type hierarchy when used. For Seurat, we integrated cells from both datasets and used the ‘TransferData’ function to transfer labels from reference to query dataset. As Seurat computes a joint embedding of cells from both reference and query dataset, we denoted its results as “ref+query” in Figure 2. We repeated each scenario 50 times to obtain the boxplot of annotation accuracy.

We also compared the differentially expressed genes (DEGs) per cell type using the true cell type label versus using the cell labels output by either Seurat or CellWalker2 (with or without utilizing the cell type tree). Single-cell RNA-seq data was simulated following

the “medium scenario”. We used Seurat’s default two-sided Wilcoxon Rank-Sum test for DEG statistical tests. For each cell type, in each of 50 simulation repetitions, we identified genes with $|\log_2 FC| > 0.5$ and adjusted p -value < 0.05 with Bonferroni correction. Then we computed the Jaccard index to assess the agreement between DEGs using the true labels versus those estimated by each method, obtaining the average Jacard index across cell types. We repeated the simulation 50 times to get the mean and standard error.

We designed four other simulation scenarios where cell types in dataset 2 (DS2) differ from those in dataset 1 (DS1) (Figure S7) in order to test performance for mapping cell types. We simulated batch effects between these two datasets with ‘batch.facLoc = 0.01’ and ‘batch.facScale = 0.1’, used default dropout rates for both. We simulated cell types A,B,C,D in DS1 and A,B,C,E in DS2 with 400 cells per cell type in each dataset. Cell type E was added to the cell type hierarchy in DS1 in several different ways. For ‘Divergent cell type’, cell type E is a new cell type in the lineage of C and D. We simulated 1000 non-marker genes, 200 marker genes between (A,B) and (C,D,E), 200 marker genes between A and B and 300 marker genes among C, D, and E. For ‘Ancestor cell type’, cell type E is the ancestor cell types of C and D a.k.a. cell type CD. We simulated 1000 non-marker genes, 200 marker genes between (A,B) and (C,D,E), 200 marker genes between A and B and 300 marker genes between C and D. For ‘Altered cell type’, cell type E being a slightly altered cell state of cell type D, i.e., E is more similar to D than C. We simulated 900 non-marker genes, 200 marker genes between (A,B) and (C,D,E), 300 marker genes between A and B, 200 marker genes between C and (D,E), and 100 marker genes between D and E with smaller fold changes. For ‘Convergent cell type’, cell type E sharing cell type markers from both cell type D and B from different lineages, i.e., E is in the lineage of C and D but share some features with B. We simulated 1000 non-marker genes, 100 marker genes between (A,B) and (C,D,E), 300 marker genes between A and (B,E) and 100 marker genes between C and (D,E). In the simulation where we varied cell numbers, we adjusted the number of cells of cell type D in DS1 to 50, 100, 200, and 400 in the ‘Divergent cell type’ case, while maintaining 400 cells for the others. To isolate the effect of cell numbers, we only included cell type A,B and E in DS2 for this simulation, so cell type E should have equal distance to cell type C and D in DS1. Each simulation scenario is repeated 50 times.

To apply CellWalker2 for cell type mapping, we combined single-cell data from both datasets to generate cell graphs and obtained cell type markers for both datasets using the procedure described above. We connected all the cells to both sets of labels. For treeArches, we input the combined data from both datasets and used scVI to remove batch effect and project onto a 20 dimensional latent space, as in the default pipeline of treeArches. For MARS, we input integrated and scaled data, after removing batch effects using Seurat as MARS does not have a batch effect removal step integrated. We set pretrain epochs to 50 and used 500 and 200 for hidden dimensions 1 and 2, respectively, in the neuron network model. For both CellWalker2 and treeArches, we input the true cell type labels in each dataset, but for MARS, we can only input cell type labels in DS1 and it did *de novo* clustering for DS2. Therefore, for MARS, we mapped its clustering results with the cell type labels of DS2 by the Hungarian algorithm⁵⁵ then obtained the relationships among cell types across datasets. We input cell type trees from both datasets for CellWalker2, which allows cell types to map to ancestral nodes of the cell type tree.

We also ran a set of simulations to evaluate how CellWalker2’s cell type comparison results are affected by the accuracy of the labels (i.e., marker gene sets). Starting with the medium scenario of the first simulation with four matched cell types in two datasets, we permuted cell type labels in both datasets, varying the percentage of cells being permuted from 0 (using true cell annotations) to 40%. Then we identified marker genes for each cell type based on permuted cell annotations. We ran the simulation 50 times and computed the proportion of times that all cell type labels (including ancestor nodes) are correctly mapped based on the largest Z-scores (using the Hungarian algorithm).

Processing scRNA-Seq data in human developing cortex

Nowakowski et al.³⁴’s dataset sequenced around 4000 cells using Fluidigm C1 in developing human brain across multiple stages (from 6 to 37 post-conception weeks) and areas (PFC, V1 and MGE etc.). Polioudakis et al.³⁵’s dataset contains around 30,000 cells from developing human brain during mid-gestation. The Nowakowski et al.³⁴ dataset has better coverage over low expressed genes, contains more developmental stages, and collected samples from different brain areas and cortical layers. On the other hand, the Polioudakis et al.³⁵ dataset has more cells, which might capture cells in transient states between cell types. In addition to these factors, the cell type classification criteria of the two studies were based on different clustering algorithms.

We downloaded the scRNA-Seq raw counts from Polioudakis et al.³⁵ and Nowakowski et al.³⁴ We obtained cell type hierarchies from the original publication. We selected marker genes with $|\text{avg_diff}| > 1$ from the supplementary material of Nowakowski et al.³⁴ For Polioudakis et al.³⁵ dataset, we computed differentially expressed genes per cell type using Seurat and selected marker genes with $|\log_2 FC| > 0.5$ and adjusted p -value < 0.01 using two-sided Wilcoxon Rank-Sum test with Bonferroni correction.

For cell type annotation, we treated Nowakowski et al.³⁴ dataset as reference and Polioudakis et al.³⁵ as query. We subsampled around 7000 random cells from Polioudakis et al.³⁵ for demonstration. For CellWalker2, we used cells from Polioudakis et al.³⁵ to build the cell-cell graph and marker genes of each cell type in Nowakowski et al.³⁴ as labels. For Seurat, we integrated cells from both datasets and applied ‘TransferData’ function with latent dimension set to 30 to transfer labels from Nowakowski et al.³⁴ to Polioudakis et al.³⁵

For mapping cell types between these two datasets, we integrated all the cells from both datasets, removed batch effects using Seurat, and incorporated cell type labels from both datasets into the graph. To obtain Z-scores, we permuted the edges from cells to cell type labels of Polioudakis et al.³⁵ We tried different permutation schema (e.g., permute cells to cell type labels of Polioudakis

et al.,³⁵ cell type labels of Nowakowski et al.;³⁴ or both), but the result did not change much (Figure S39). Since we showed the results of mapping cell type of Polioudakis et al.³⁵ to the cell type hierarchy of Nowakowski et al.³⁴ in the main text, we used the first permutation schema as it maintains the correlation of cell types on the tree we mapped to. For treeArches, we input combined raw counts from the two datasets then used scVI as default to integrate the data. For MARS, we input integrated and scaled data after removing batch effects using Seurat, and set hidden dimensions 1 and 2 to be 1000 and 100 respectively in the neuron network model. We treated Nowakowski et al.³⁴ as annotated data and Polioudakis et al.³⁵ as unannotated data and mapped MARS identified clusters with the original labels in Polioudakis et al.³⁵ to get the final cell type mapping results.

Processing multimodal data in human developing cortex

We downloaded scATAC-Seq, scRNA-Seq, and 10x Genomics multiomics data from GEO: GSE162170. We integrated scRNA-Seq and multiomics data from week 21, subsampled 9,000 cells from scRNA-Seq data, and used all the 8981 multiomics cells to construct the cell graph in CellWalker2. When further integrating with scATAC-Seq data from week 21, we subsampled 9000 cells with ATAC-Seq from week 21 and added them to the cell graph. When further integrating with scATAC-Seq data from week 16, we used all 6423 cells with ATAC-Seq from week 16. We used Seurat to integrate scRNA-Seq and the RNA-Seq part of multiomics data. We used the cell type labels from scRNA-Seq combining different ages during mid-gestation from the supplementary information of²⁸ for CellWalker2, and we built the cell type tree using ‘BuildClusterTree’ function in Seurat based on first 50 PCs of the gene expression data. We generated a cell type tree using the scRNA-Seq data, using only the leaf nodes in CellWalker but the full hierarchy in CellWalker2. We used the cell type labels from the multiome data for Wilcoxon and Fisher’s exact tests. A comparison for cell type labels between multiome and scRNASEq data is described in Figure S40. One-sided Wilcoxon tests compare the distribution of edge weights between cells from each cell type to basal ganglia versus cortical plate pREs. As a third method, we first identified differentially accessible regions (DARs) using cells with scATAC-seq for each cell type and then tested for enrichment of cortical plate and basal ganglia pREs overlapping these DARs with one-sided Fisher’s exact test. We identified DARs with Bonferroni adjusted *p*-values < 0.05 using Signac’s likelihood ratio test (LRT). Lastly, we created a CellWalker graph using only cells with scATAC-seq data and mapped pREs to discrete cell type labels using the original *ad hoc* method implemented in CellWalkR.

We obtained predicted regulatory elements (pREs) from different brain regions (i.e., basal ganglia versus cortical plate, upper versus deeper layer pre-frontal cortex) in²⁷, and connected region-specific pREs to cells by the overall accessibility in each set of regions. To compute Z-scores of the association between region-specific pREs and cell types, we permuted the edge weights between cells and cell type labels.

To identify cell type-specific TFs, we used Signac to find the presence of 697 TF motifs from JASPAR2020 within all the pREs, and applied CellWalker2 to identify cell type-specific TFs. We used all the scRNA-Seq (12557 cells), scATAC-Seq (12675 cells) and multiomic cells (8981 cells) from week 21. We treated each TF as a bulk label, and we connected a TF to a cell by summing up the reads within pREs that contain the TF motif, normalized by the total number of reads. To get Z-scores, we permuted pREs containing each TF motif 100 times. To further filter for TFs that are expressed in the cell types, we computed the average normalized gene expression in each cell type using the scRNA-Seq data from week 21. For internal node, we averaged the gene expression across all the cells in its descendants. For each cell type, either tips or ancestors, we selected TFs with *Z* score > 5 and standardized expression > 0.5. To show high-scoring TFs on the cell type tree, we selected top five TFs ranked by *Z* score for each node. If a TF is selected for multiple nodes in the same lineage, it is only shown on the most ancestral node. To compare with motif enrichment results by Signac, we identified DARs for each cell type that overlap with pREs and tested for enriched TF motifs in these regions. In detail, for DARs (LRT *p*-value < 0.001) that overlap with pREs, we identified 10,000 background regions from all ATAC-seq peaks that overlap with pREs. Then, we compared the number of foreground regions containing the motif to the total number of background regions containing the motif using a one-sided hypergeometric test.² Benjamini-Hochberg procedure was applied for multiple testing correction.

To compare with using ChIP-Seq peaks as input to CellWalker2, we downloaded non-redundant peaks from all available ChIP-Seq experiments in cortex or neuron from ReMap2022.²⁹ The number of peaks for each TF in each experiment is shown in Figure S41A. For each TF, we identified pREs overlapping with ChIP-Seq peaks as input to CellWalker2. For benchmarking cell type-specific CTCF binding from different ChIP-Seq experiments, we identified pREs overlapping with ChIP-Seq peaks in each experiment as input to CellWalker2. We assigned CTCF ChIP-Seq experiments as labels using multiomics data from the developing cortex.²⁸ We did 100 permutations, as above, to obtain Z-scores.

Processing multispecies scRNA-Seq and SNARE-Seq2 data

Comparing cell types across species by CellWalker2 using scRNA-Seq data We downloaded scRNA-Seq data from human, mouse and marmoset from the BICCN data portal (https://data.nemoarchive.org/publication_release/Lein_2020_M1_study_analysis). By analyzing scRNA-Seq data in each species separately, Bakken et al.⁴⁵ identified 72, 52 and 59 inhibitory neuron cell types in human, marmoset and mouse, respectively. These are grouped into 6 subclasses in human and 7 subclasses in marmoset and mouse, which have an additional Meis2 subclass. We obtained cell type hierarchy from the three species from the original publication,⁴⁵ and identified cell type or subclass specific marker genes using Seurat after balancing the number of cells to 100 in each group. We used ‘roc’ test and selected markers with power > 0.65 for each species. We computed cell-to-label edge weights within each species based on the log fold-change of marker genes and gene expression profiles of each cell. Then, we constructed the

cell-to-cell graph based on the integrated gene expression data using consensus genes across species provided by.⁴⁵ The graph contains around 10,000 cells per species.

Identifying genes expressed specifically in different cell types We used 767 expressed gene sets from HGNC and SynGO⁴⁵ as labels, and CellWalker2 to label these gene annotations using scRNA-seq data from.⁴⁵ The weights of edges connecting each gene set and cells are the average standardized expression level of all genes in the gene set. The cell-to-cell graph and cell-to-label edges are the same as above. The edge weight between a cell and a gene set is the average standardized expression level in the cell, and we filtered for gene set that express in at least 20 cells. To obtain Z-scores, we permuted the edges between cells and cell type labels.

Comparison of cell type-specific regulatory regions in human and marmoset We downloaded marmoset SNARE-Seq2 data from BICCN portal, which includes 1451 marmoset inhibitory neurons with both gene expression and chromatin accessibility. We used the same cell type labels and marker genes derived from scRNA-Seq as above. We used around 50K DARs from 18 human cell clusters identified from scATAC-Seq as labels.⁴⁵ To map human DARs to marmoset cell types, we used liftOver to transfer human DARs to the marmoset genome and removed DARs that are mapped to more than 3 regions. To run CellWalker2, we used marmoset cells to generate a cell-to-cell graph, treated human DARs from different cell types as annotation nodes and linked these nodes to each marmoset cell by computing the proportion of reads in the peaks overlapping with that group of DARs. To obtain Z-scores, we permuted the edges between cells and marmoset cell type labels.

Identifying cell type-specific transcription factors in human and marmoset We downloaded human and marmoset SNARE-Seq2 data with 22217 and 1451 cells, respectively, from the BICCN portal. We used the cell subclass label and identified markers by Seurat using the scRNA-Seq data, as described above. We treated TFs as annotation nodes in CellWalker2. We obtained subclass DARs from the supplementary information in⁴⁵ and used Signac to find the presence of 697 TF motifs from JASPAR2020 within these DARs. We further selected expressed TFs to run CellWalker2, which resulted in 379 TFs for human and 405 TFs for marmoset. Each TF is connected to cells through the chromatin accessibility of the DARs containing the TF motif. To obtain Z-scores, we permuted the genomic regions that contain the motif of each TF. We ran ArchR with the same peaks set and followed ArchR's pipeline to identify TF motifs for each cell subclass. In detail, we identified DARs per cell subclass using a binomial test after binarizing the count matrix, selected DARs with FDR (≤ 0.1 and $\log_2 FC \geq 1$, and tested for TF motif enrichment using hypergeometric tests.

We further computed the average standardized gene expression in each cell subclass using the RNA-Seq part of the SNARE-Seq2 data. For internal nodes, we averaged the gene expression across all the cells in its descendants. For each tip or internal node on the cell subclass tree, we selected TFs with Z score > 2.5 and standardized expression > 0.2 or 0.15 for human and marmoset respectively to compare between human and marmoset. The threshold for calling expressed TFs of a cell type for marmoset is lower because we observed less variation of marmoset data. We varied these cutoffs to maximize the percentage of shared TFs between human and marmoset.

Human PBMC scRNA-Seq datasets

For mapping cell types in PBMCs, we downloaded the scRNA-Seq dataset of healthy adult blood tissue from https://cellypist.cog.sanger.ac.uk/Resources/Organ_atlas/Blood/Blood.h5ad. We subsampled 20% of cells to run CellWalker2, resulting in 33123 cells from 46 cell types in Ren et al.³⁰ and 9361 cells from 33 cell types in Yoshida et al.³¹ We identified cell type markers and built a cell type tree using Seurat. We kept differentially expressed genes per cell type with $|\log_2 FC| > 0.5$ for Ren et al.³⁰ and 0.25 for Yoshida et al.³¹ and adjusted p-value < 0.05 based on two-sided Wilcoxon Rank Sum tests with Bonferroni correction, so that the number of cell type markers are in a similar range for both datasets. We used Seurat to integrate the two datasets based on the top 30 canonical directions and generated a cell-to-cell graph based on K nearest neighbors with $k = 20$. As we illustrated mapping cell types from Yoshida et al.³¹ to the cell type tree in Ren et al.,³⁰ we permuted cells to cell type labels from Yoshida et al.³¹ 100 times to get Z-scores. For treeArches, we input combined raw counts from two datasets then used scVI as default to integrate the data. For MARS, we input integrated and scaled data after removing batch effects using Seurat, set hidden dimensions 1 and 2 to be 1000 and 100 respectively in the neuron network model, and trained for 100 epochs. We treated Ren et al.³⁰ as annotated data and Yoshida et al.³¹ as unannotated data and mapped MARS identified clusters with the original labels in Yoshida et al.³¹ to get the final cell type mapping results. For Z score differences between mapping stimulated and un-stimulated monocytes from Yoshida et al.³¹ to monocyte related cell types in Ren et al.,³⁰ we used Monocyte CD14 IFN stim - Monocyte CD14, Monocyte CD16 IFN stim - Monocyte CD16, and Monocyte CD14 IL6 - Monocyte CD14.

10x genomics multiomics human PBMC dataset

For identifying cell type-specific TFs, we downloaded 10x Genomics multiomics data from <https://www.10xgenomics.com/datasets/pbmc-from-a-healthy-donor-granulocytes-removed-through-cell-sorting-10-k-1-standard-1-0-0>. We followed the Signac pipeline https://stuartlab.org/signac/articles/pbmc_multiomic.html for calling peaks and cell-type annotation. Signac identified 131,364 peaks in total. We removed cell types with less than 30 cells to identify cell type-specific genes and peaks. We followed https://stuartlab.org/signac/articles/motif_vignette.html for identifying cell type-specific peaks and motifs from JASPAR2020 within these peaks. 58,385 cell type-specific peaks are identified using Signac by likelihood ratio test (adjusted p-value < 0.05 with Bonferroni correction). We selected TFs that were expressed in at least 100 cells, which resulted in 303 TF motifs. Then we ran CellWalker2 using the cell type markers from Seurat, cell type-specific peaks, and motifs within these peaks (defined using Signac). We built a cell graph in which TF motif instances in the genome are annotation nodes and then computed Z-scores for their annotation-to-label mappings.

We further computed the standardized expression level for each TF across cell types and selected TFs with standardized expression level > 0.5 and adjusted P-value < 0.01 (after converting Z-scores to p -values and applying Bonferroni correction) for each cell type, as well as Spearman's rank correlation coefficient between expression level and Z score > 0.4 . For an internal node, the expression level is averaged across all the cells of its descendants. To show high-scoring TFs on the cell type tree, we selected the top five TFs ranked by the correlation between Z score and expression level for each node. If a TF was selected for multiple nodes in the same lineage, it was only shown on the most ancestral node. For benchmarking cell type-specific CTCF binding from different ChIP-Seq primary cells or cell lines, we identified cell type-specific peaks overlapping with ChIP-Seq peaks in each experiment as input to CellWalker2. We did 100 permutations as above to obtain Z-scores.

We also ran motif enrichment analysis by Signac directly as a comparison. We compared the number of cell type-specific peaks (adjusted p -value < 0.05 by LRT with Bonferroni correction) containing the motif with the total number of 50,000 accessible peaks containing the motif using a one-sided hypergeometric test.² The Benjamini-Hochberg procedure was applied for multiple testing correction. We used the same criteria as in CellWalker2 for filtering TFs after obtaining the enrichment p -values. We ran ArchR according to <https://www.archrproject.com/bookdown/index.html> except using the cell type label by Signac for direct comparison. ArchR identified 169,218 peaks and 97,920 cell type specific peaks (using a binomial test after binarizing the count matrix, FDR < 0.05 and log2FC > 1.5). Then, we used hypergeometric tests to identify enriched motifs for each set of cell type-specific peaks. We used the same criteria as in CellWalker2 for filtering TFs after obtaining the enrichment p -values. We ran SCENIC+ according to https://scenicplus.readthedocs.io/en/latest/pbmc_multomeTutorial.html#Tutorial:-10x-multome-pbmc except using the cell type labels by Signac for direct comparison. We selected cell type-specific TFs with rho > 0.4 . We ran SIMBA according to https://simba-bio.readthedocs.io/en/latest/multome_shareseq_GRN.html. We relaxed the cutoffs for selecting master regulators to gene_max = 1, gene_gini = 0.2, motif_max = 2, motif_gini = 0.5, and ranked TFs by the distance between their gene and motif on the embedding space from largest to smallest. As SIMBA only outputs master regulators but not the specific cell types in which they function, we queried the top 30 nearest cells for each TF in the embedding space generated by SIMBA, and obtained the most prevalent cell type. Querying more or fewer nearest cells resulted in poorer performance.

Comparing different methods with ChIP-Seq data in human PBMC

We obtained non-redundant ChIP-Seq peaks in blood primary cells or cell lines from ReMap, and compared outputs from different methods using 10x Genomics multiomics human PBMC data as described in the last section. We filtered for TFs with the proportion of peaks containing their motifs > 0.6 (Figure S41B). We used CTCF ChIP-Seq data for benchmarking “Using biological knowledge to benchmark labeling of bulk-derived annotations”, and used the rest TFs for computing correlations with scores and power of different methods. We computed both Spearman's rank and Pearson correlation coefficient between Z-scores (CellWalker2), log10 adjusted P-values (ArchR and Signac), AUC (SCENIC+), rank (SIMBA) and log10 number of ChIP-Seq peaks. For all these methods, larger scores mean more cell type specificity for a TF. As ChIP-Seq data do not have the same cell type granularity as the results obtained from single cell data, we grouped cell types and ChIP-Seq experiments into three major cell classes: B cell/plasmablast, T cells and monocytes/dendritic cells, and obtained the maximum scores/number of peaks in each class. We excluded TFs in each cell class that do not have ChIP-Seq data. This procedure ends up with 58 data points for the correlation computation. To compare the power of detecting TF binding in each cell class, we treated TFs with more than 2000 ChIP-Seq peaks as ground truth, and computed the sensitivity of each method at different thresholds. As the ranges of scores are different for each method, we normalized the scores by their maximum values.

QUANTIFICATION AND STATISTICAL ANALYSIS

Using Signac to identify DARs and motifs enrichment

We used the likelihood ratio test implemented in Signac to identify DARs. It constructs a logistic regression model predicting cell types based on each peak individually and compares this to a null model without the peak. P-value adjustment is performed using Bonferroni correction based on the total number of peaks in the dataset. We provided the number of cells and p -value cutoffs for selecting DARs in each subsection of [STAR Methods](#) that uses the statistical test and in [Table S3](#). To identify enriched motifs, we used the one-sided hypergeometric test implemented in Signac, comparing the number of foreground regions containing the motif with the total number of background regions containing the motif. The Benjamini-Hochberg FDR procedure was applied for multiple testing correction. The number of regions and motifs are provided in each subsection of [STAR Methods](#) that uses the statistical test.

Using archR to identify DARs and motifs enrichment

To identify DARs with archR, we used binomial tests after binarizing the count matrix, as implemented in archR. We adjusted p -values to control the FDR, and FDR cutoffs for selecting DARs are provided in each subsection of [STAR Methods](#) that uses this test. We identified enriched motifs in DARs using hypergeometric tests. The number of cells, regions and motifs are provided in each subsection of [STAR Methods](#) that uses the statistical test and in [Table S3](#).

Using Seurat to identify DEGs

To identify DEGs or cell type markers, we used the default two-sided Wilcoxon Rank-Sum test in Seurat with Bonferroni correction. Cutoffs for p -values and log fold changes for selecting DEGs, as well as the number cells, are provided in each subsection of [STAR Methods](#) that uses the statistical test or in [Table S3](#).

Statistical tests for labeling bulk-derived annotations

We applied one-sided Wilcoxon tests to compare the distribution of edge weights between cells from each cell type to basal ganglia versus cortical plate pREs ([Figure S10A](#)). We also tested for enrichment of cortical plate and basal ganglia pREs overlapping these DARs with a one-sided Fisher's exact test ([Figure S10B](#)). The number of cells and the number of regions are provided in each subsection of [STAR Methods](#) that uses the statistical test.

Supplemental information

**CellWalker2: Multi-omic discovery using
hierarchical cell type relationships**

Zhirui Hu, Paweł F. Przytycki, and Katherine S. Pollard

Supplemental Tables

Cell type	Group	Cell type full name
ExDp1	ExDp	Excitatory neuron deep layer 1
ExDp2	ExDp	Excitatory neuron deep layer 2
ExM	ExM	Maturing excitatory neuron
ExM-U	ExM	Maturing excitatory neuron (upper layer enriched)
ExN	ExN	Migrating excitatory neuron
InCGE	InCGE	Interneuron CGE
InMGE	InMGE	Interneuron MGE
IP	IP	Intermediate Progenitor
PgG2M	Pg-Div	Cycling progenitor G2/M phase
PgS	Pg-Div	Cycling progenitor S phase
oRG	RG	Outer Radial Glia
vRG	RG	Ventricular Radial Glia

Table S1: Polioudakis's cell type groupings with abbreviations, related to Figures 4 and 5.

Cell type	Group	Cell type full name
EN-V1-1	EN-early	Early Born Deep Layer/subplate Excitatory Neuron V1
EN-PFC1	EN-early	Early Born Deep Layer/subplate Excitatory Neuron PFC
EN-PFC2	EN-late	Early and Late Born Excitatory Neuron PFC
EN-V1-3	EN-late	Excitatory Neuron V1 - late born
EN-PFC3	EN-late	Early and Late Born Excitatory Neuron PFC
EN-V1-2	EN-late	Early and Late Born Excitatory Neuron V1
IN-STR	IN-STR	Striatal neurons
IN-CTX-CGE1/2	InCGE	CGE/LGE-derived inhibitory neurons
IN-CTX-MGE1/2	InMGE	MGE-derived Ctx inhibitory neuron, Cortical Plate-enriched
IPC-div1	IPC-div	Dividing Intermediate Progenitor Cells RG-like
IPC-div2	IPC-div	Intermediate Progenitor Cells RG-like
IPC-nEN1/2/3	IPC-nEN	Intermediate Progenitor Cells EN-like
MGE-div	MGE-div	dividing MGE Progenitors
MGE-IPC1/2/3	MGE-IPC	MGE Progenitors
MGE-RG1/2	MGE-RG	MGE Radial Glia 1/2
nEN-early1/2	nEN	Newborn Excitatory Neuron - early born
nEN-late	nEN	Newborn Excitatory Neuron - late born
nIN1-5	nIN	MGE newborn neurons
vRG	RG	Ventricular Radial Glia
oRG	RG	Outer Radial Glia
tRG	RG	Truncated Radial Glia
RG-div2	RG-div	Dividing Radial Glia (G2/M-phase)
RG-div1	RG-div	Dividing Radial Glia (S-phase)
RG-early	RG-early	Early Radial Glia

Table S2: Nowakowski's cell type groupings with abbreviations, related to Figures 4 and 5.

Datasets	Number of cells	Time	Memory	Tasks
Human developing cortex scRNA-Seq	6,768	~ 3mins	< 2.5G	Label cells
Three species cortex scRNA-Seq	29,486	~ 1hr	< 40G	Compare cell types
Human developing cortex scRNA-Seq	38,105	~ 1.7hrs	< 66G	Compare cell types
PBMC scRNA-Seq datasets	42,484	~ 2.1hrs	< 75G	Compare cell types
10X PBMC Multiome	11,331	~ 1hr	< 27G	Map TFs to cell types
Human cortex SNARE-Seq	22,217	~ 18hrs*	< 75G	Map TFs to cell types
Human developing cortex (RNA-Seq + ATAC-seq + Multimodal)	34,213	~ 4.5hrs	< 85G	Map TFs to cell types

Table S3: Runtime and memory benchmarking of datasets analyzed in the manuscript, related to the STAR Methods. Running time is recorded using 32 cores CPU (100 permutations to compute Z-scores) except “Label cells” task which uses 1 core and no permutation. *: this dataset has many more ATAC-seq peaks resulting in a longer run time.

Supplemental Figures

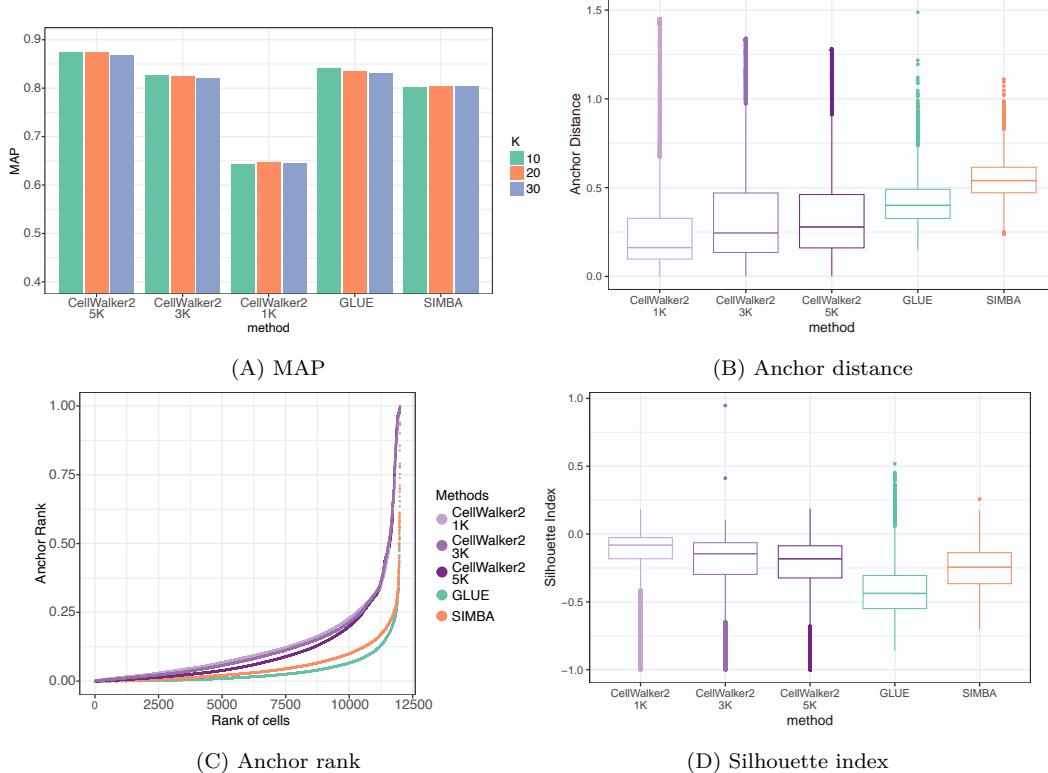


Figure S1: Benchmarking result for integrating scRNA-Seq with scATAC-Seq data on 10x Genomics PBMCs, related to Figure 1. CellWalker2: we varied the number of dual-modality cells from 1K to 5K as labeled. (A) shows the mean percentage of the neighboring ATAC cells with the same cell type label across all RNA cells, higher MAP is better; (B) show the distribution of anchor distance across all cells, smaller scores are better; (C) ranks the cell by ‘anchor rank’ from smallest to largest, smaller scores are better; (D) show the distribution of silhouette index across all cells, higher scores are better. With enough dual-modal cells as a bridge (i.e. more than 3K in this case), The cell graph generated by CellWalker2 is similar or better in conserving cell types (in terms of MAP). We also observed that MAP increases with the number of dual-modal cells. For matching cells with the same barcode between ATAC-Seq and RNA-Seq, CellWalker2 is better in terms of silhouette index and anchor distance but not as good as SIMBA and GLUE in terms of anchor rank. It suggests that although CellWalker2 may not be able to distinguish the matched pairs of cells at the single cell granularity, similar cells have much larger influence scores than others, which encourages information flow within similar cell types. The results indicate that the generated cell graph by CellWalker2 captures the similarity between cells.

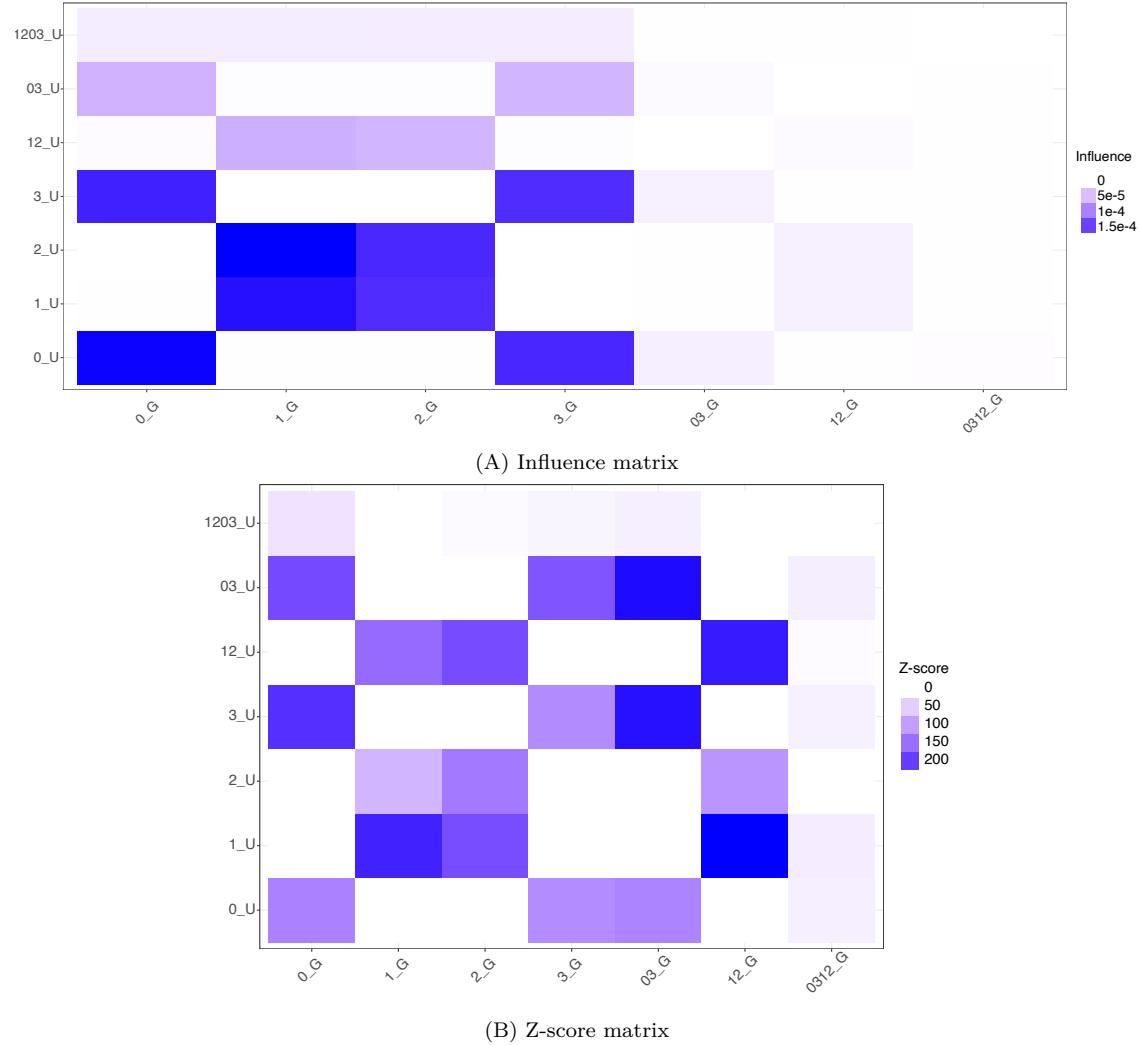


Figure S2: Influence and Z-score matrices between two sets of cell type labels in one dataset from Simulation-easy scenario, related to Figure 1. (A) Influence score; (B) Z-score. This simulation has identical cell types in two scRNA-Seq datasets. We used suffix ‘U’ to denote the cell types in the first data set and ‘G’ for the second. Leaf nodes in general have larger influence scores even for mapping the ancestor cell type. Thus it is hard to tell if a cell type should be mapped to leaves or internal nodes in the other hierarchy by influence scores. However, Z-score can map the cell types to the correct level as the ancestor node has a larger Z-score for mapping the corresponding ancestor node in the other dataset. Some of the leaf cell types also have relatively large Z-scores for mapping to their ancestor cell type.

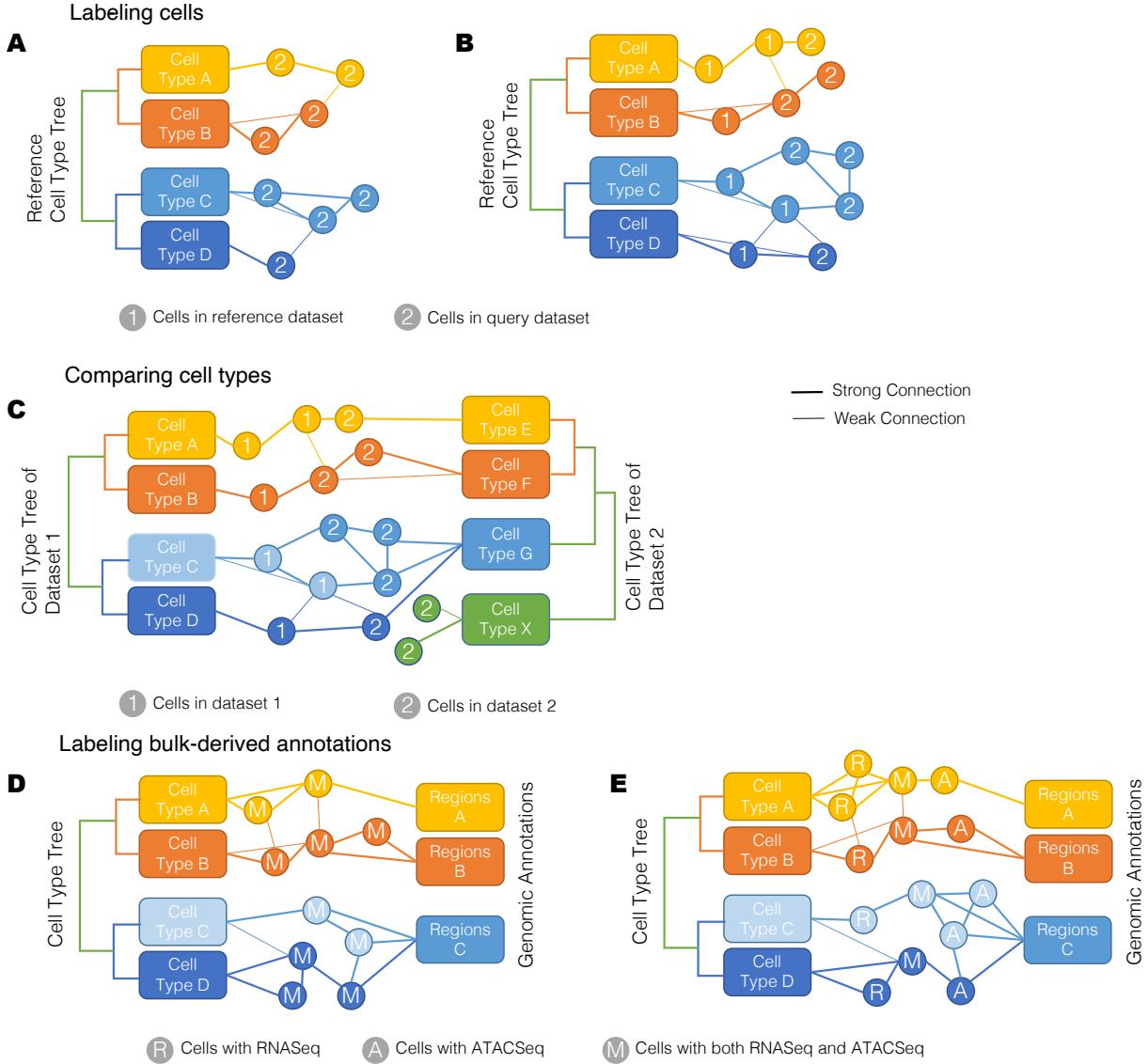


Figure S3: Inputs, outputs, and graphs for different applications of CellWalker2, related to Figure 1. (A-B) Examples of models for labeling cells. Input: reference cell type labels with optional hierarchical structure and cells with RNA-Seq or ATAC-Seq data. The edge weight is specified based on the assay. Cell type label is unknown; cells are colored for illustration only. Output: influence matrix and Z-scores for mapping cells to labels. (A) Using cells from only the query dataset. (B) Using cells from both reference and query dataset. (C) Model for comparing cell type labels between two or more ontologies. Input: cell type labels from each dataset with optional hierarchical structure and cells with RNA-Seq or ATAC-Seq data. In this case, cell type E \approx A, F \approx B, G \approx CD (the parent of C and D) and X is not seen in dataset 1 so it is not mapped to any cell type. The correspondence between labels in the two datasets is unknown; nodes are colored for illustration only. Output: influence matrix and Z-scores for mapping cell type labels between datasets. (D-E) Example models for labeling bulk-derived annotations to cell types. Input: cell type labels with optional hierarchical structure, annotations with genome coordinates, and cells with multiomic, scRNA-seq, or scATAC-seq data. The cell types in which the annotations are active are unknown; nodes are colored for illustration only. In these examples, annotation sets A and B are assigned to cell types A and B, respectively. Annotation set C is activated in cells from both cell type C and D, so it would be mapped to their parent cell type (CD). (D) Using cells with multiomic data. Annotations are connected to cells with multiomic data using the ATAC-Seq channel, while cell type labels are connected via the RNA-Seq channel. (E) Integrating cells with multiomic data, only RNA, or only ATAC. Multiomic cells serve as bridge to connect the cells with only one type of data.

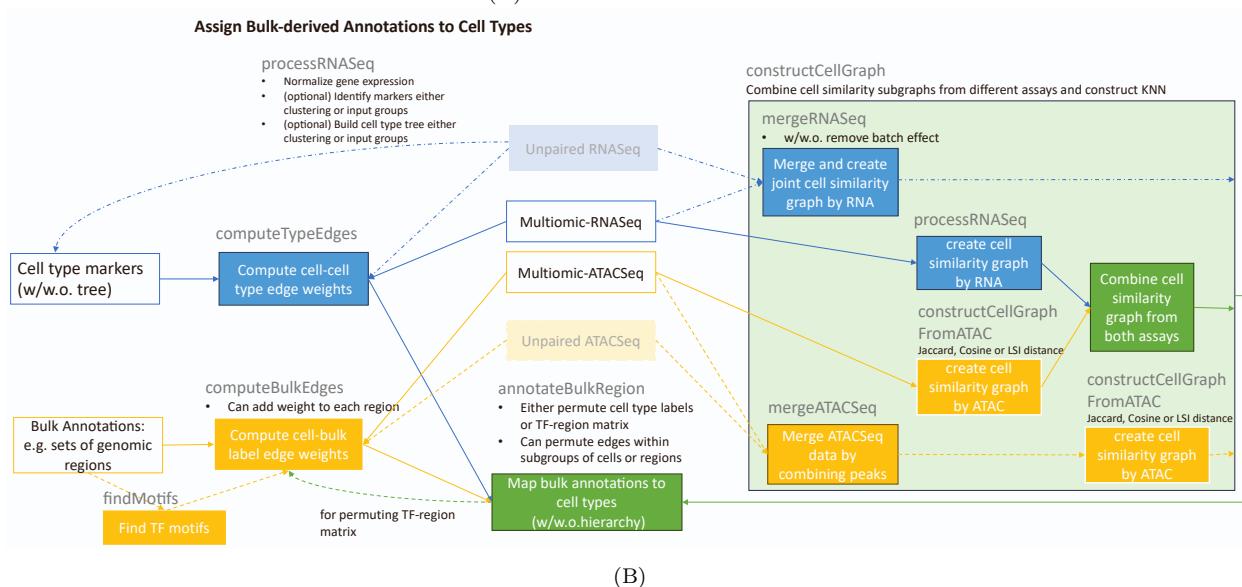
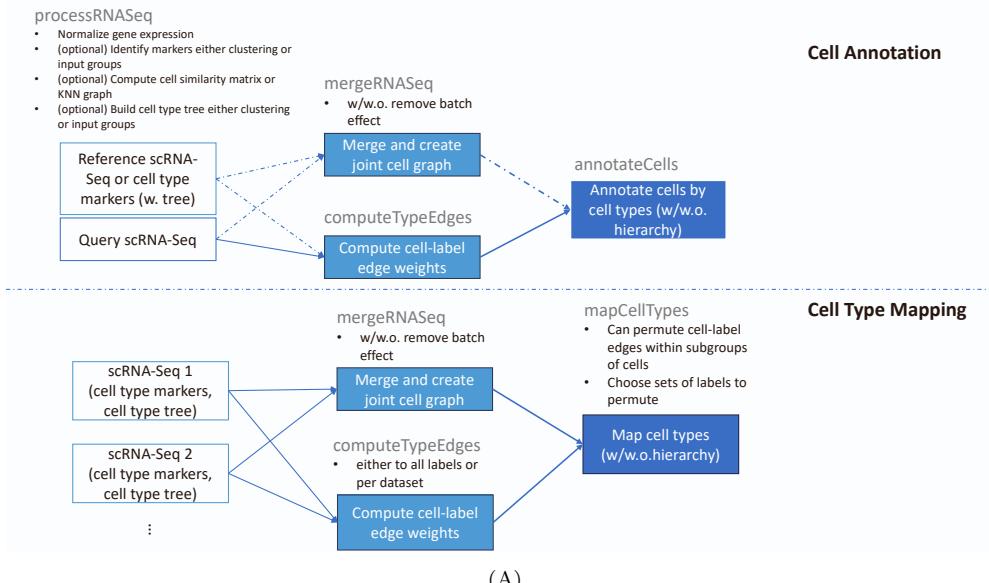


Figure S4: CellWalker2 pipelines in CellWalkR package, related to Figure 1. (A) Cell annotation and cell type mapping pipeline. (B) Assigning bulk-derived annotations to cell types pipeline. Solid line, required steps; dashed line, optional steps. Box with no filling, required input data; Box with shaded filling, optional input data. CellWalker2 functions for RNA-Seq data, blue box; CellWalker2 functions for ATAC-Seq data, orange box; CellWalker2 functions for both assays, green box. The function name is labeled above each box.

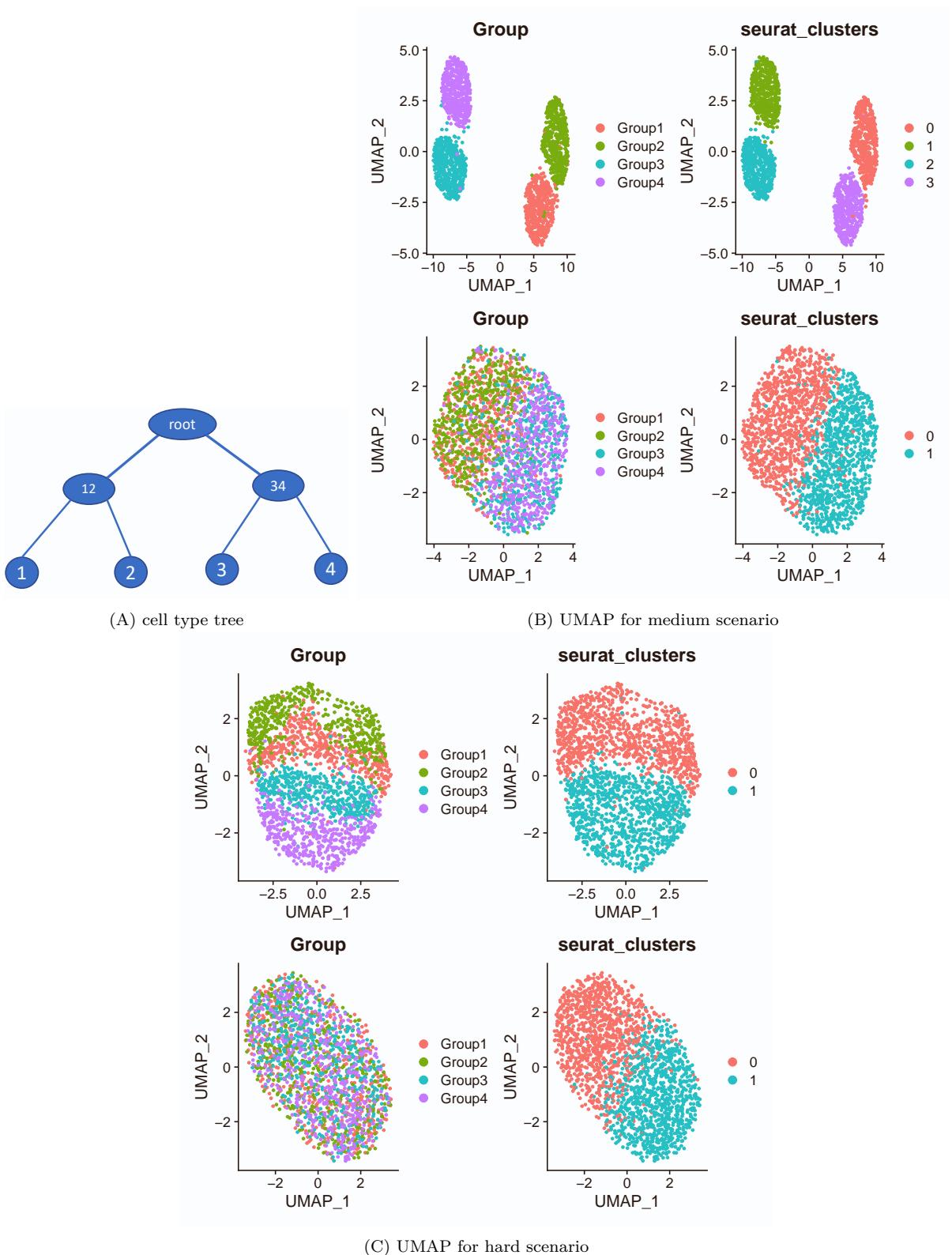


Figure S5: Cell type tree and UMAPs for labeling cells simulations, related to Figure 2A and B. (A) Cell type tree for all simulation scenarios. (B) and (C) UMAP of the simulated data from medium and hard scenarios, respectively. Top and bottom: simulated reference and query datasets. Left panel is colored by true cell type and right is colored by Seurat clusters. In the medium scenario, the query dataset has more dropouts than does the reference dataset, and hence only two cell subclasses are visible on UMAPs generated by Seurat in the query dataset. In the hard scenario, we added more dropouts in reference datasets and cell types are less distinguishable even for the reference data.

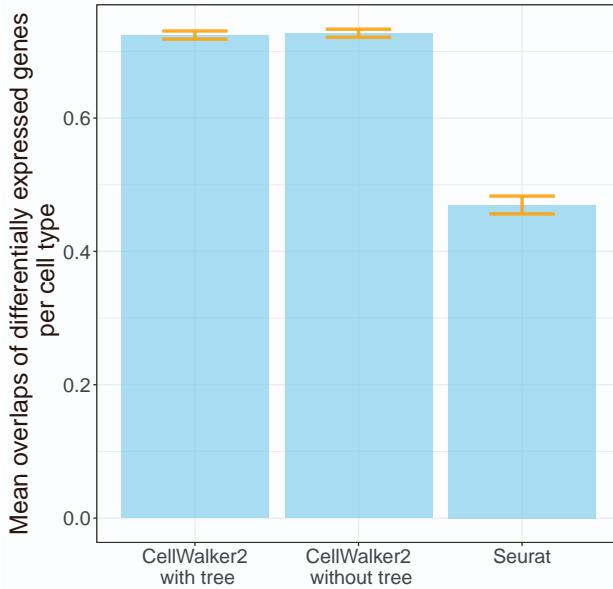


Figure S6: CellWalker2 performs better than Seurat in identifying differentially expressed genes (DEGs) as a downstream task after labeling cells, related to Figure 2A. We simulated single-cell RNA-seq data following the “medium scenario” (details in STAR Methods) and then assigned cell type labels to cells using Seurat or CellWalker2 (with or without utilizing the cell type tree). Following Seurat’s default methodology for differential gene expression tests, we identified DEGs using the true cell type labels, Seurat labels, and both versions of CellWalker2 labels. Seurat and CellWalker2 DEGs were assessed for agreement with the DEGs based on the true cell type labels using the Jaccard index. The barplot shows the mean DEG Jaccard index across cell types, with error bars representing standard errors across 50 repetitions of the simulation. Compared to Seurat, CellWalker2 generates DEGs that agree better with DEGs based on the true cell type labels. This is a consequence CellWalker2’s better performance in labeling cells when batch effects and dropouts are present (Figure 2A).

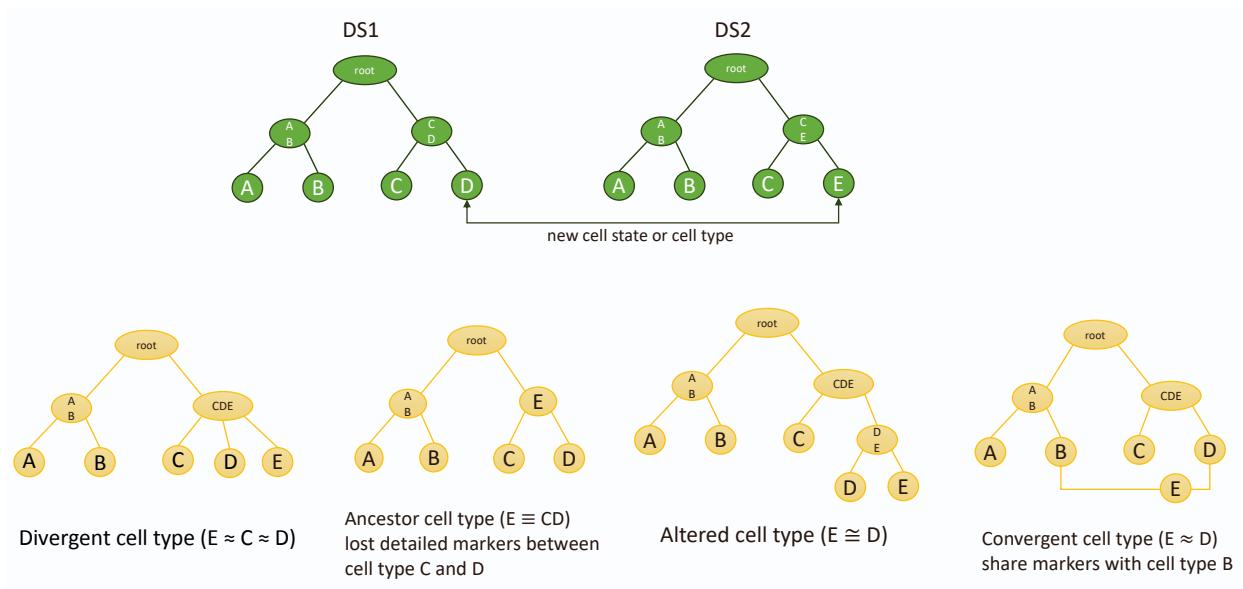
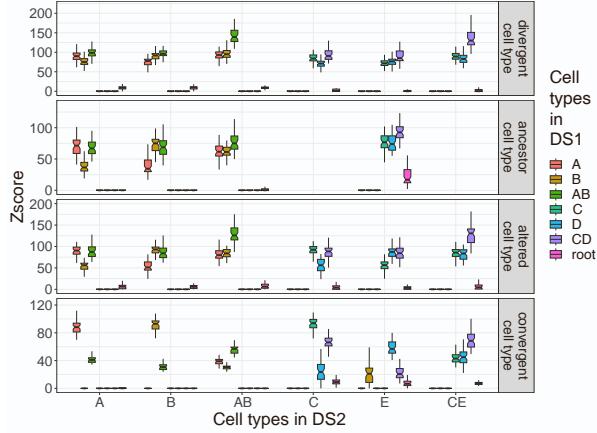
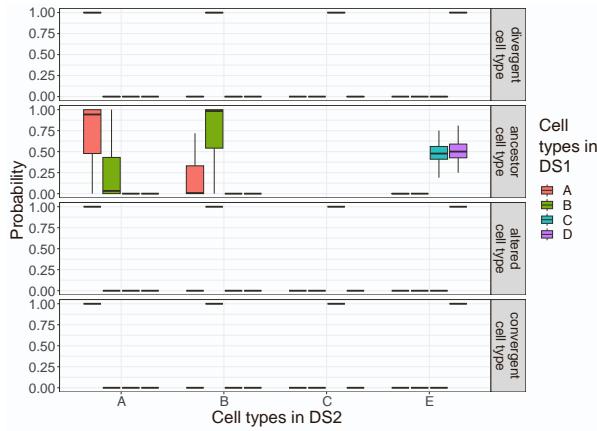


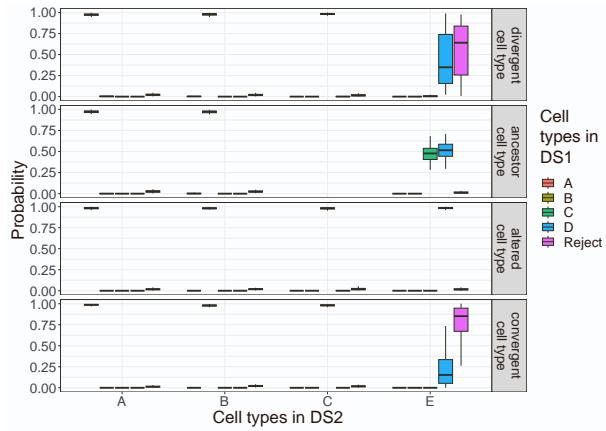
Figure S7: Simulation scenarios for comparing cell types between two hierarchical ontologies, related to Figure 2C and D. Top: both datasets have four cell types that form a balanced tree. But cell type E in dataset2 (DS2) is a new cell state or cell type compared to DS1. Bottom: four simulation scenarios where E has a different hierarchical relationship with the other cell types. We simulated some marker genes for every split of the cell type tree (see STAR Methods).



(A) CellWalker2



(B) MARS



(C) treeArches

Figure S8: Comparison of performance of CellWalker2, MARS, and treeArches for comparing the cell type ontology in dataset 2 (DS2) to the one in dataset 1 (DS1), related to Figure 2C. Four simulation scenarios with different placements of cell type E in DS2 are shown. For both the ‘Ancestor’ and ‘Divergent’ cell type cases, CellWalker2 assigned a high Z-score for mapping cell type E to cell type CD, but in the former case, the Z-score to the root node was also large as CD connects to the root on the cell type tree; for the ‘Convergent cell type’ case, CellWalker2 assigned a lower Z-score to the cell type CD but a larger Z-score to cell type B. MARS assigned a probability near 1 to cell type D in all three cases except ‘Ancestor cell type’. treeArches could not distinguish the ‘Divergent’ and ‘Convergent’ cell type cases, as it fails to detect the similarity to cell type B that shares features with E.

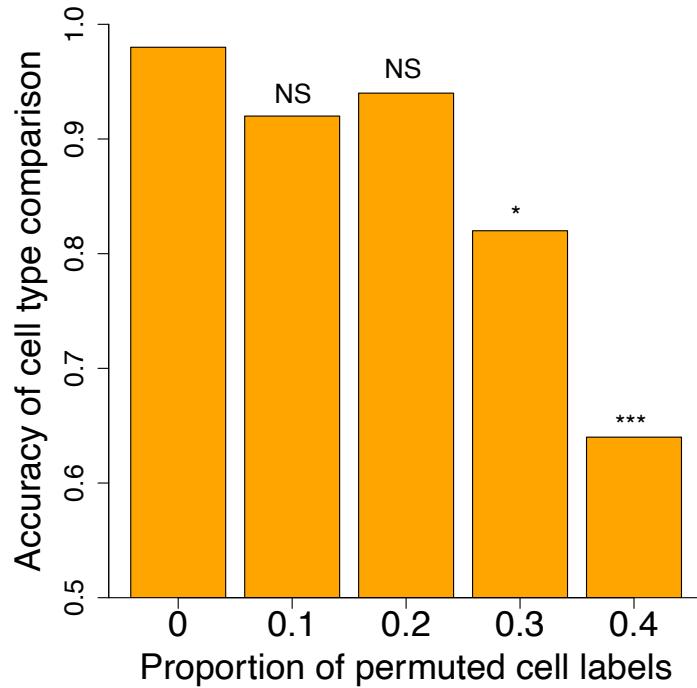


Figure S9: Evaluation of how CellWalker2’s cell type comparison results are affected by the accuracy of the cell labels (i.e., marker gene sets) by simulations, related to Figure 2. Starting with the medium scenario of the simulation with four matched cell types in two data sets (STAR Methods), we permuted cell type labels in both data sets, varying the percentage of cells being permuted from 0 (using true cell annotations) to 40% (X-axis). Then we identified marker genes for each cell type based on permuted cell annotations. We ran each simulation 50 times and computed the proportion of times that all cell type labels (including ancestor nodes) are correctly mapped (Y-axis) based on the largest Z-scores. We compared the accuracy between each permuted scenario and the non-permuted scenario using Chi-squared tests. The significant level is shown on top of each bar. NS: not significant; *: p-value < 0.05; ***: p-value < 1e-4. The accuracy of mapping stays above 0.9 when up to 20% of cell type annotations are permuted and the differences are not significant, but decreases to 0.64 when 40% of cell type annotations are permuted. We observed that when more cells are labeled incorrectly, the cell types are less distinguishable and fewer cell type markers could be found. Thus, CellWalker2 is fairly robust to erroneous labels, but the cell-to-label edges become less reliable as the quality of the labeling degrades notably.

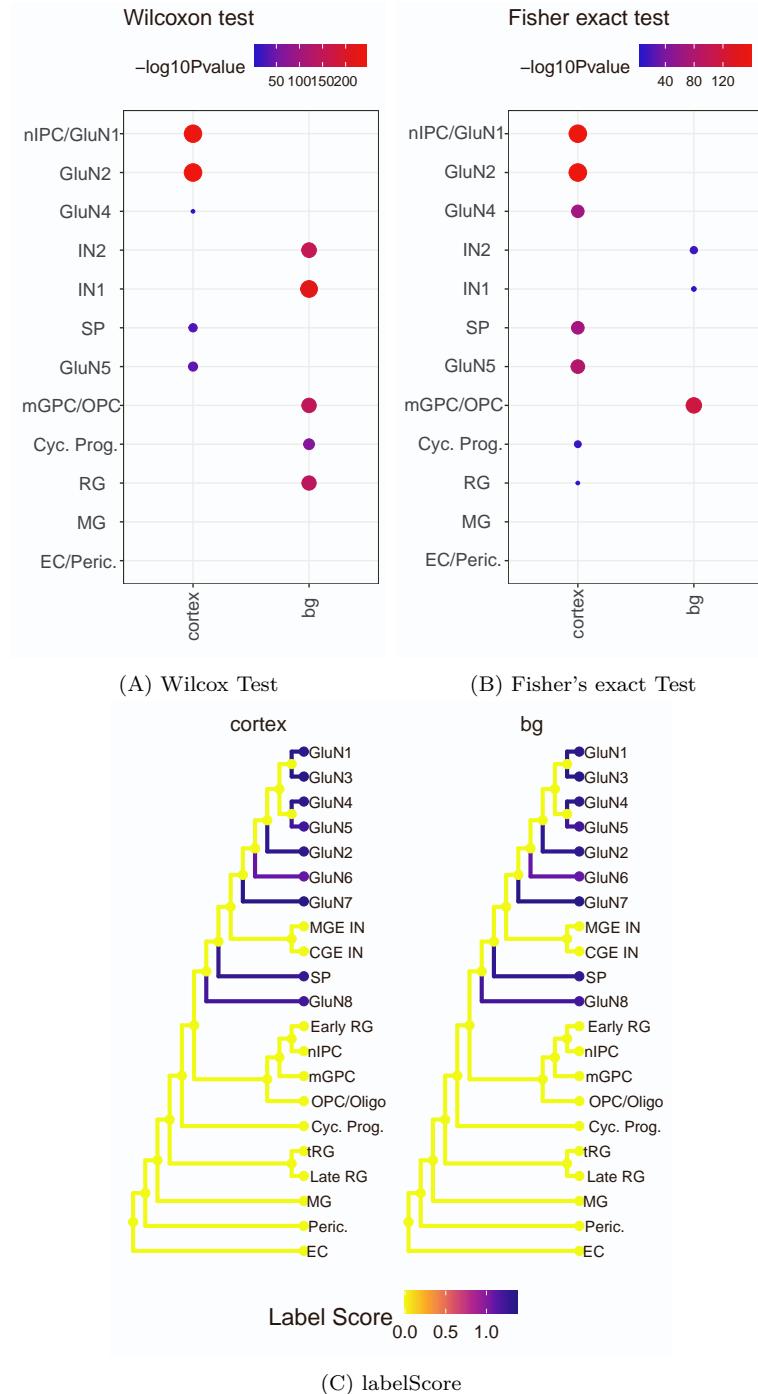


Figure S10: Comparison of different methods mapping basal ganglia versus cortex specific pREs [S1] to different cell types in Trevino et al. [S2], related to Figure 2E. (A) Mapping region specific pREs to cell types identified from transcriptomic profile in the multiomic data using the Wilcoxon test. We compared the distribution of edge weights to basal ganglia versus cortex specific pREs for each cell type. (B) Mapping region specific pREs to cell types identified from transcriptomic profile in the multiomic data using Fisher's exact test. We identified differentially accessible regions (DARs) for each cell type first, and then tested if region specific pREs are over-represented in these DARs. Errors may be due to difficulties calling DARs using scATAC-seq data. For (A) and (B), mappings with one-sided $p\text{-value} < 0.01$ are shown. (C) labelScore from CellWalker mapping region specific pREs to cell type hierarchy in [S2] using both multiomic and RNA-Seq data. Label scores greater than 1 are shown.

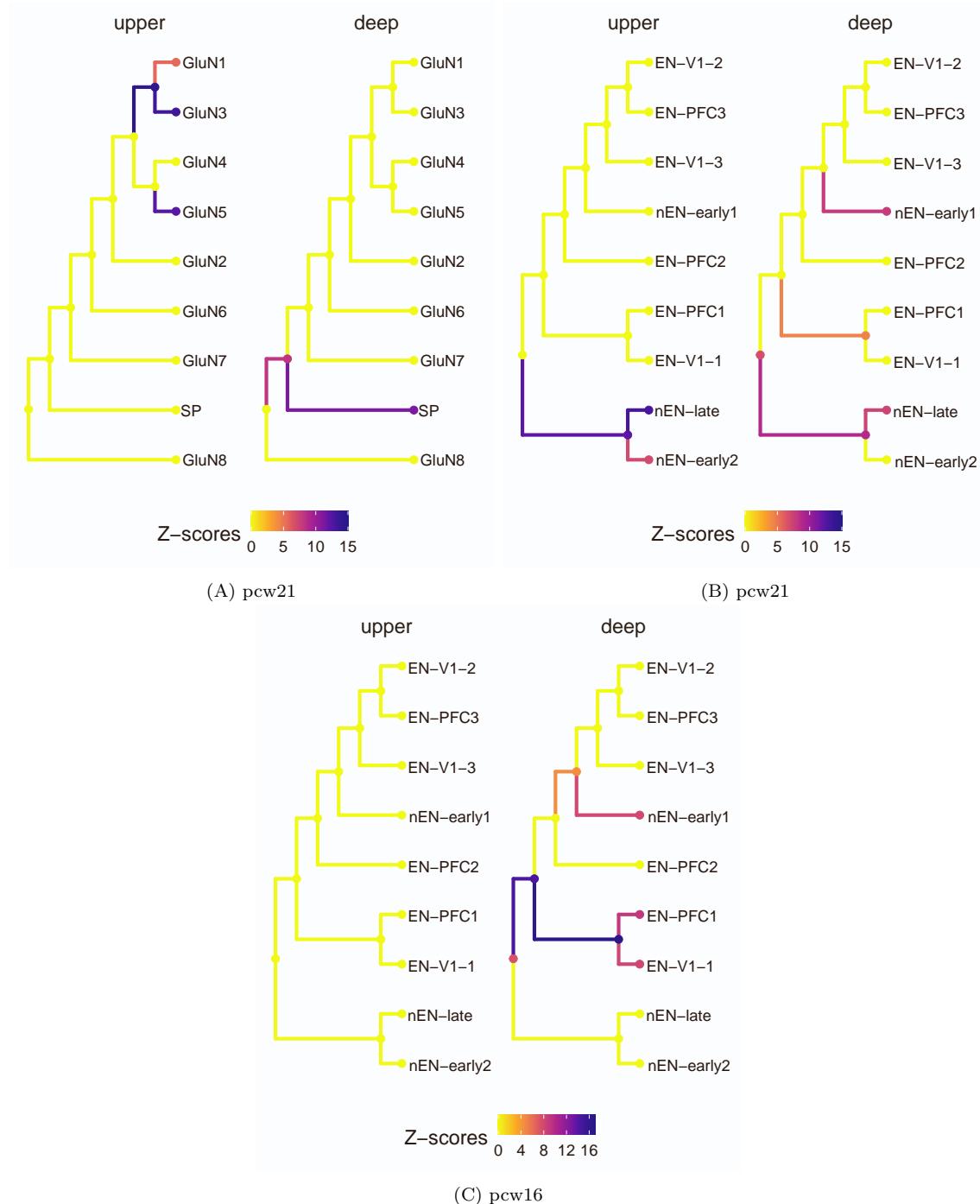


Figure S11: Mapping layer-specific pREs to different cell type lineages by integrating ATAC-Seq from different developmental ages and using different cell type ontologies, related to Figure 2F. As neurons from deep laminae layers develop at an earlier age than those from upper layers, we incorporated either scATAC-Seq cells from a later stage (week 21) or an earlier stage (week 16) into the cell graph. (A) Z-scores mapping upper vs deep layer specific pREs to cell type hierarchy in [S2]. (B) and (C) Z-scores mapping upper vs deep layer specific pREs onto cell type hierarchy of excitatory neurons in [S3], which annotates more fine-grained cell types. scATAC-Seq from post-conception week (pcw) 21 and pcw 16 was used, respectively. As expected, pREs from the upper layer have higher chromatin accessibility in late-born newborn excitatory neurons (nEN-late) and this increased accessibility appears in week 21. On the other hand, pREs from deep layer have higher chromatin accessibility in early-born deep layer/subplate excitatory neurons (EN-PFC-1 and EN-V1-1) as well as early-born newborn excitatory neurons (nEN-early1), and the cell type specificity is weakened at week 21 with smaller Z-scores.

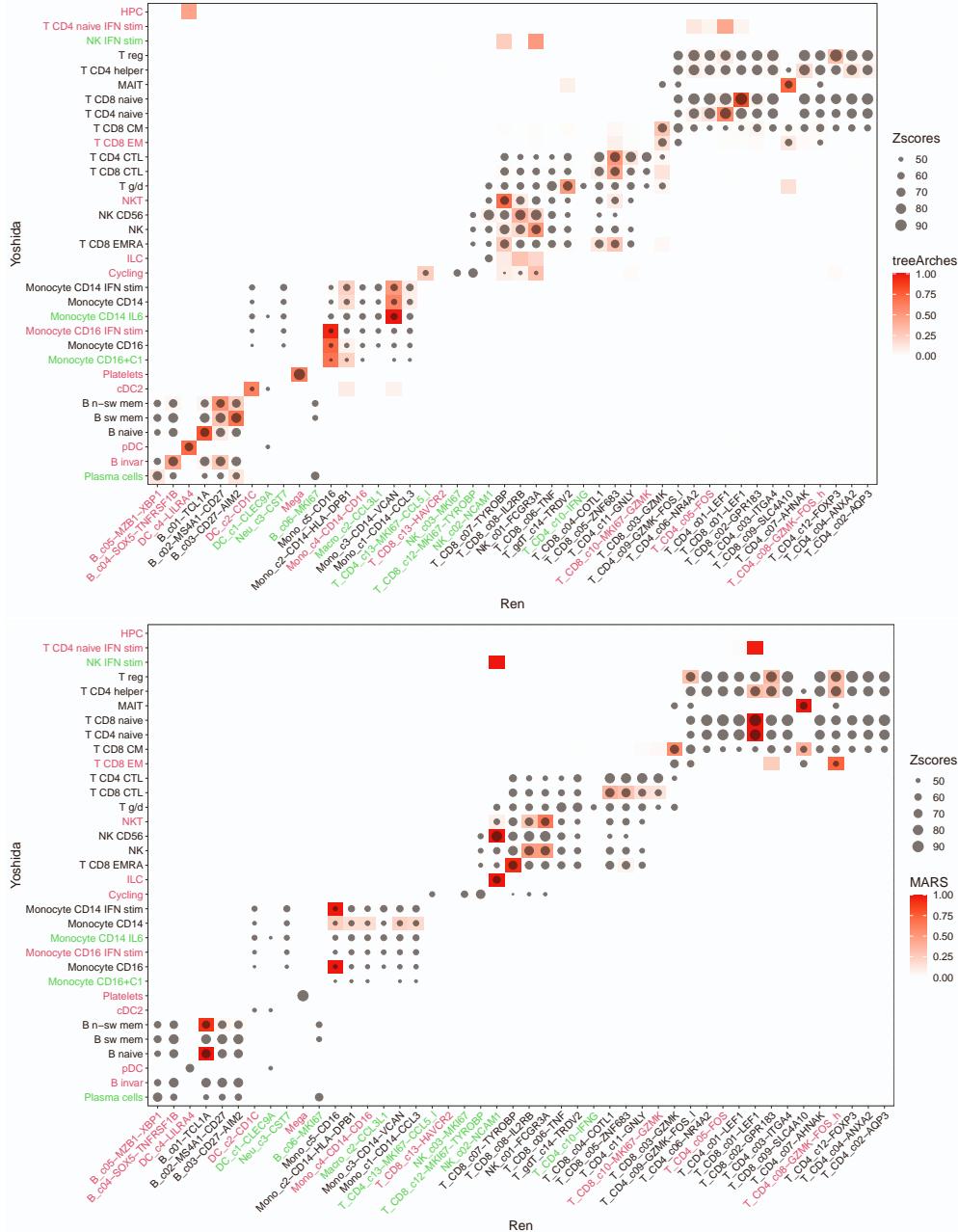


Figure S12: Comparing cell types between Yoshida et al.[S4] and Ren et al.[S5] datasets using CellWalker2, treeArches, and MARS, related to Figure 3A-C. The top panel shows CellWalker2 versus treeArches, and the bottom panel shows CellWalker2 versus MARS. In each panel, columns are cell types from Ren et al. and rows are cell types from Yoshida et al.. Ren's cell types are ordered by the cell type tree. Yoshida's cell types are ordered by the matching cell types of Ren's. Only tips are shown. The size of the dots represent CellWalker2's Z-scores, and these are the same in both panels. The colored boxes represent mapping probabilities from treeArches (top) and MARS (bottom), so the boxes are different in the two panels. MARS made many errors, such as mapping all naive T cells (CD4/CD8/IFN stimulated CD4) to naive CD8 T cells (T CD8 c01-LEF1). treeArches was biased towards more prevalent cell types, e.g., treeArches maps CD4 cytotoxic T lymphocytes (T CD4 CTL) to T CD8 c05-ZNF683 (CD8 cytotoxic T lymphocytes), which is one of the four most prevalent cell types and a common map for treeArches, rather than to the less prevalent T CD4 c11-GNLY, which matches expectations given that both are CD4 cells and GNLY is a marker of cytotoxic T lymphocytes. Z-score > 47 are shown (set the threshold to include all treeArches hits with probability > 0.5). Green: cell type $< 0.1\%$, Red: cell type $< 1\%$. CM, central memory; CTL, cytotoxic T lymphocytes; EM, effector memory; EMRA, effector memory re-expressing CD45RA; invar, invariant; n-sw, non-switched; sw, switched; IFN-stim: interferon-stimulated subpopulations. For Ren et al., cell types are named based on marker genes. “high” or “low” labels indicate the relative expression levels and genes without high or low labels are specifically expressed in the cell type.

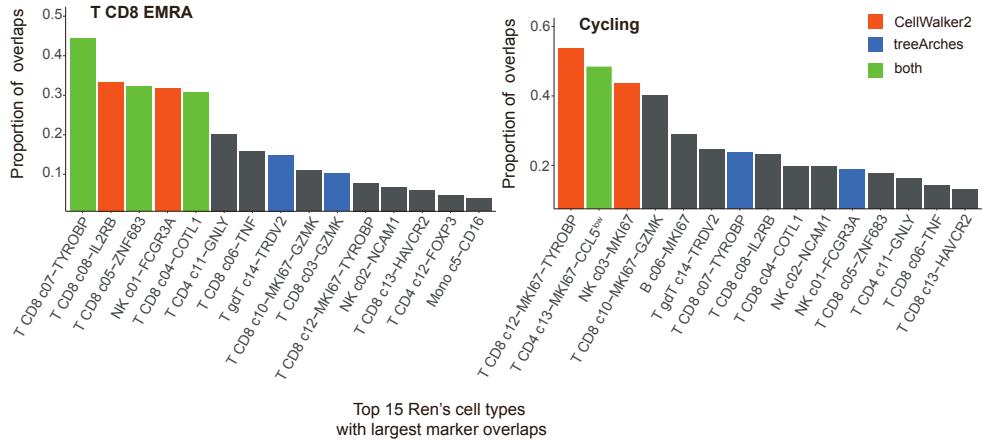


Figure S13: Proportion of positive markers overlapping between T CD8 EMRA (left) and Cycling (right) in Yoshida et al. [S4] versus different cell types in Ren et al. [S5], related to Figure 3A. Only the top 15 cell types with the largest overlaps are shown. Orange: Top 5 (T CD8 EMRA) or 3 (Cycling) mapped cell types by CellWalker2; Dark blue: Top mappings by treeArches; Green: Top mappings by both methods.

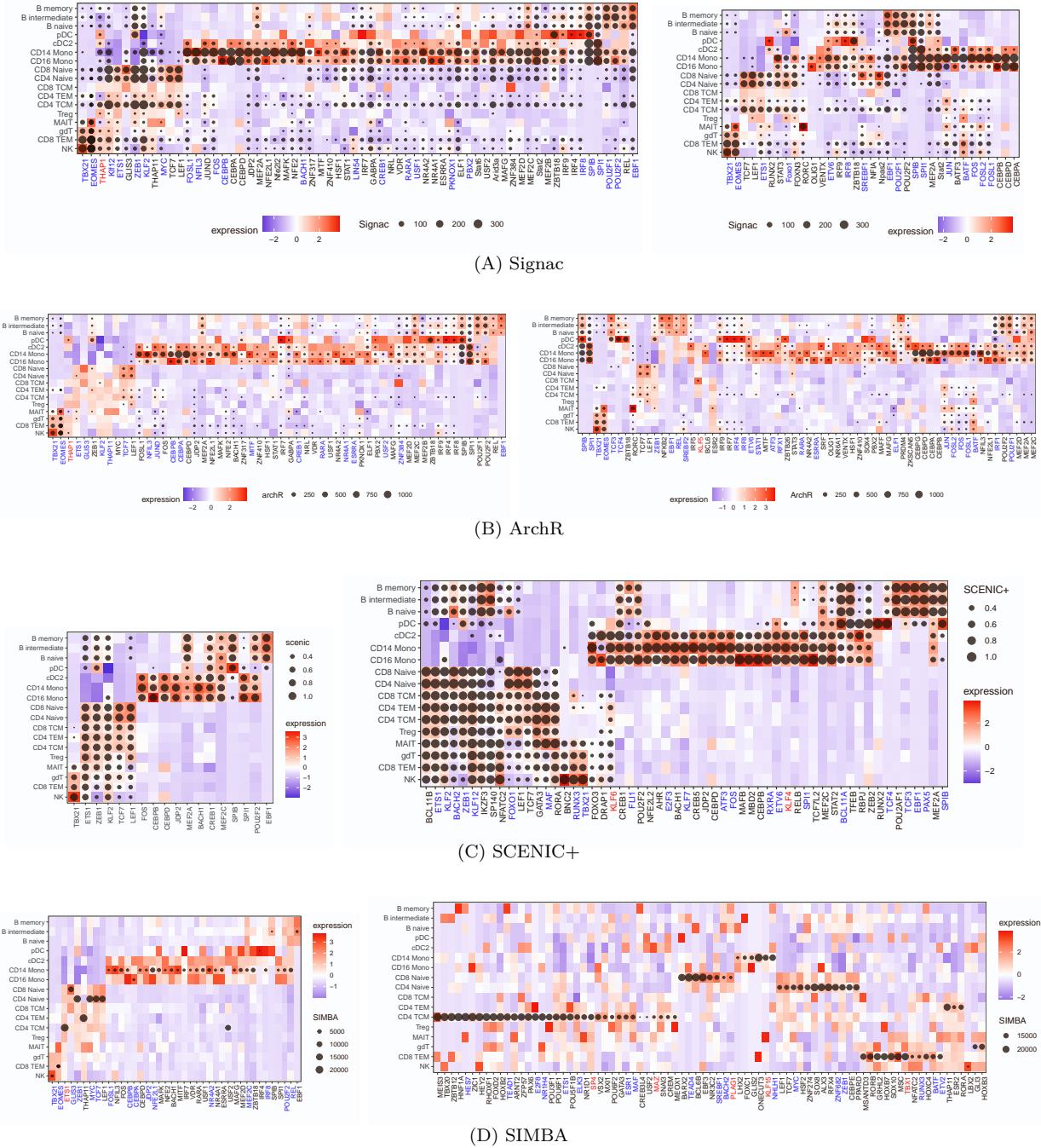


Figure S14: Cell type-specific transcription factors (TFs) in PBMC by different methods, related to Figure 3D-E. Each row is a cell type and each column is a TF. The color represents normalized expression level for each TF across cell types. The size of the dot represents the score by different methods: -log10 P-values for Signac (A) and ArchR (B), AUC for SCENIC+ (C), and gene rank for SIMBA (D). In each panel, the left figure shows TFs selected by CellWalker2, and the right figure shows TFs selected by one of the other methods. TFs in red font are human universal stripe factors, in blue font are other stripe factors. As SCENIC+ and CellWalker2 use different TF motif databases, not all TFs selected by CellWalker2 are included in the SCENIC+ analysis, and hence the left heatmap in (C) shows fewer TFs. For SIMBA, we only included the top cell type for each TF. Although we relaxed the thresholds for calling master regulators (top 50% genes by the rank), SIMBA did not identify any TFs specific to B cells and dendritic cells and does not have many overlaps with CellWalker2 (D). TFs show more cell type specificity with CellWalker2 compared to Signac and ArchR, which show broad enrichment patterns for several TFs that are known to be regulators of specific lineages. For example, *TBX21* and *EOMES* are significant in B cells in addition to T and NK cells for both methods, and the B cell-specific TF *EBF1* is significant in monocytes and T cells for Signac.

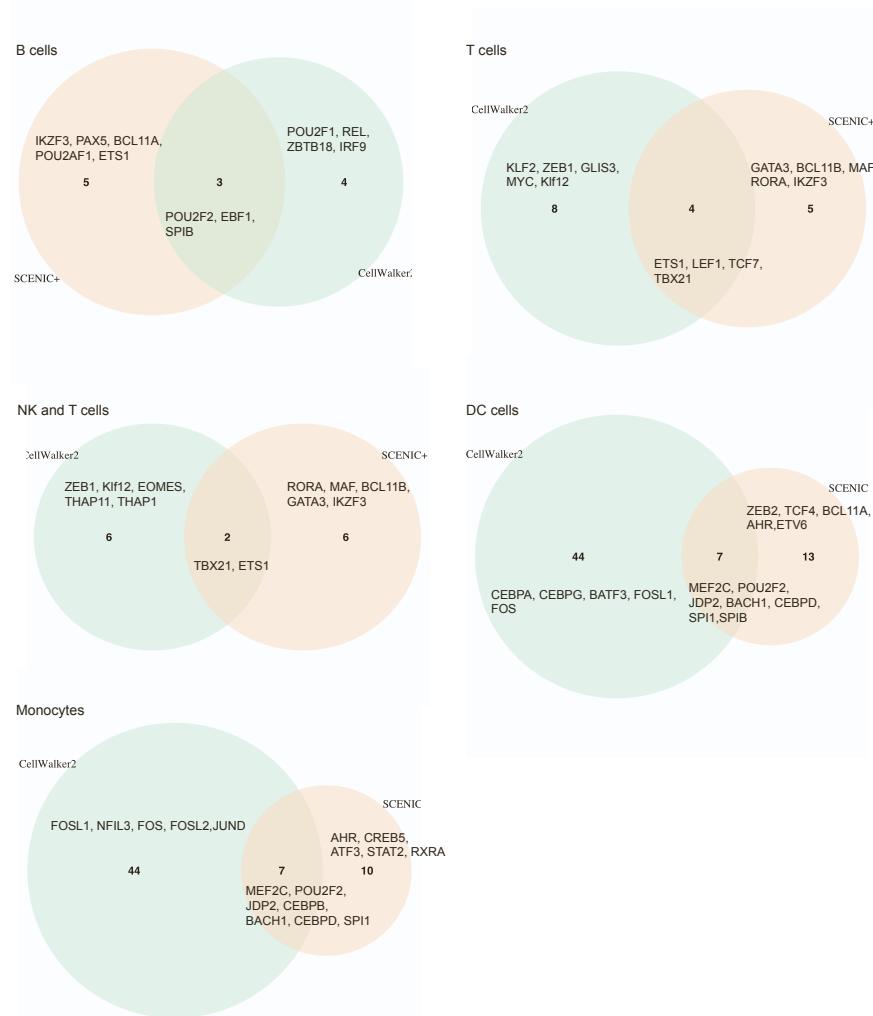
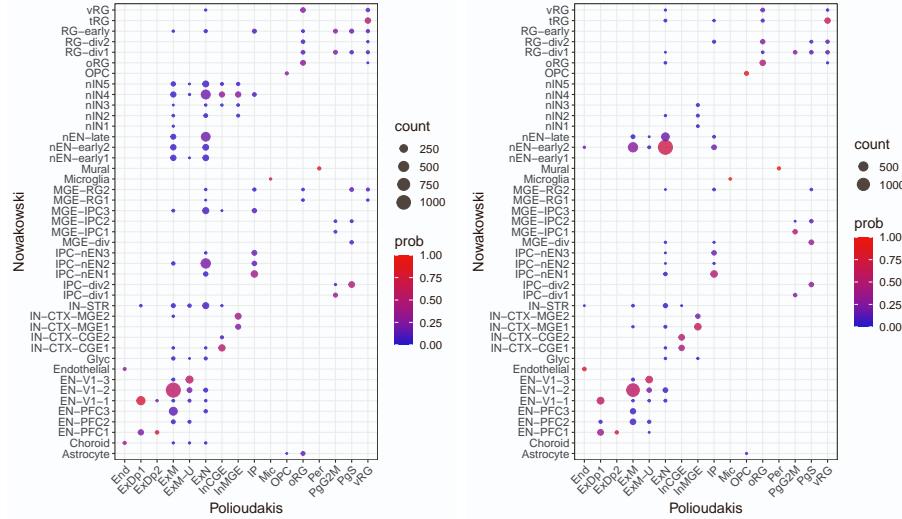
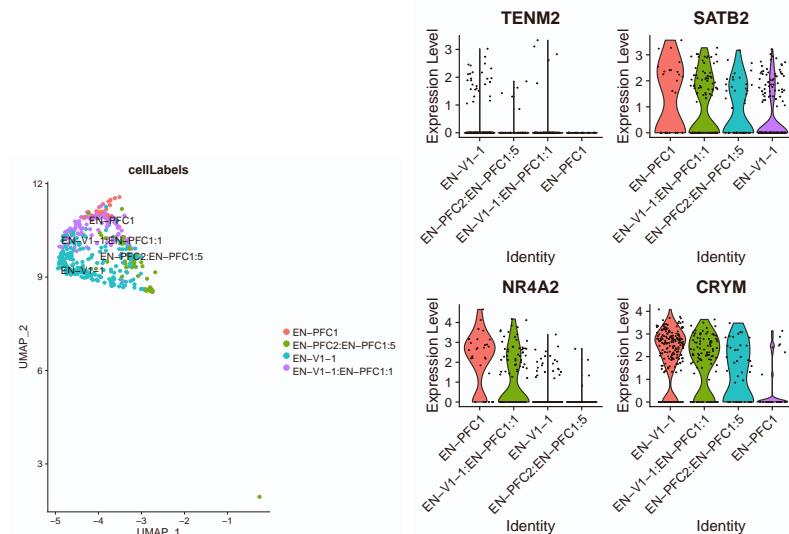


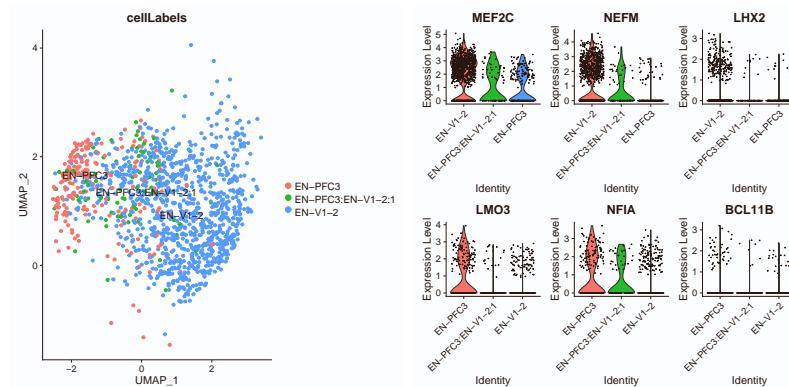
Figure S15: Comparison of transcription factors identified by SCENIC+ and CellWalker2 in major PBMC cell types, related to Figure 3D. The number of TFs uniquely identified and found in common by each method are shown. For cell types with too many uniquely identified, only the top 5 TFs with largest Z-scores or AUC score are named.



(A) CellWalker2 vs Seurat



(B) ExDp1



(C) ExM

Figure S16: Labeling cells by CellWalker2 and Seurat using human developing cortex scRNA-Seq data, related to Figure 4. (A) Results of annotating cells in Polioudakis et al.'s dataset [S6] (query) with the cell types from Nowakowski et al. [S3] (reference). The color of each dot represents the probability of mapping each cell type in query dataset to reference cell type labels, and the size of each dot represents the number of cells mapped (mappings with at least 5 cells are shown). (b-c) UMAP and expression level of markers of ExDp1 (B) and ExM (C). UMAP is computed using all cells from Polioudakis et al. [S6] but only a random subset of cells are plotted.

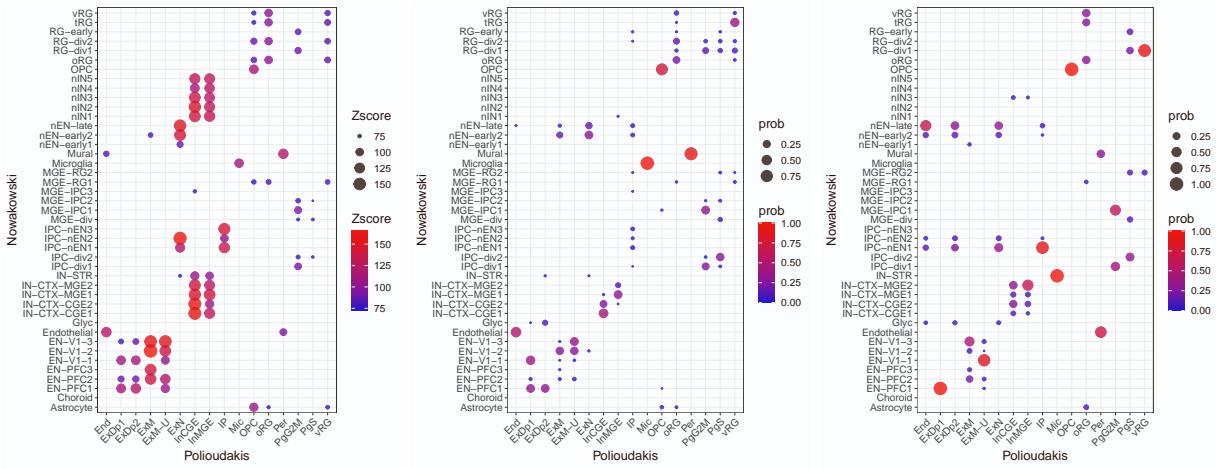


Figure S17: Comparison of mapping cell types between Polioudakis et al. [S6] and Nowakowski et al. [S3] dataset, related to Figure 5A. Z-scores from CellWalker2 (left), probabilities from treeArches (middle), and probabilities from MARS (right). Each column is a cell type from Polioudakis et al. [S6], and each row is a cell type from Nowakowski et al. [S3]. As treeArches and MARS do not consider hierarchical structure of the cell types, we only include terminal cell types as well for CellWalker2 for this figure. MARS made errors for several rare, non-neuronal cell types including pericytes (Per; mapped to endothelial cells) and microglia (Mic; mapped to Striatal neurons (IN-STR)). Only Z-scores > 70 and probabilities > 0.003 are shown.

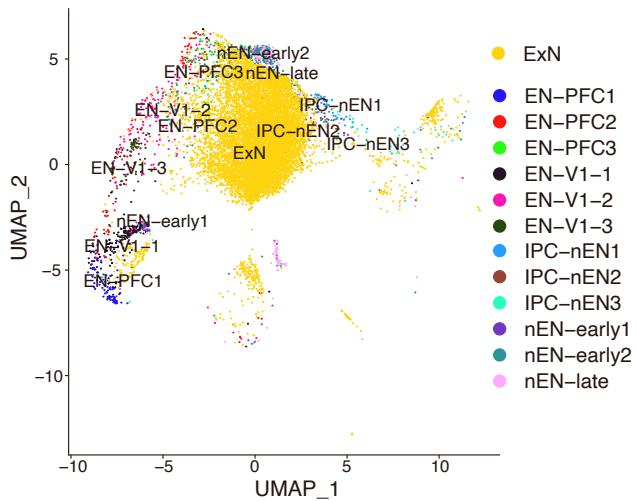


Figure S18: ExN shows proximity to the cell types mapped by CellWalker2, related to Figure 5B. In a UMAP plot, ExN is close to IPC-nEN2, nEN-late, and nEN-early2. The UMAP is drawn using integrated data including ExN from Polioudakis et al. [S6] and excitatory neuron related cell types from Nowakowski et al. [S3].

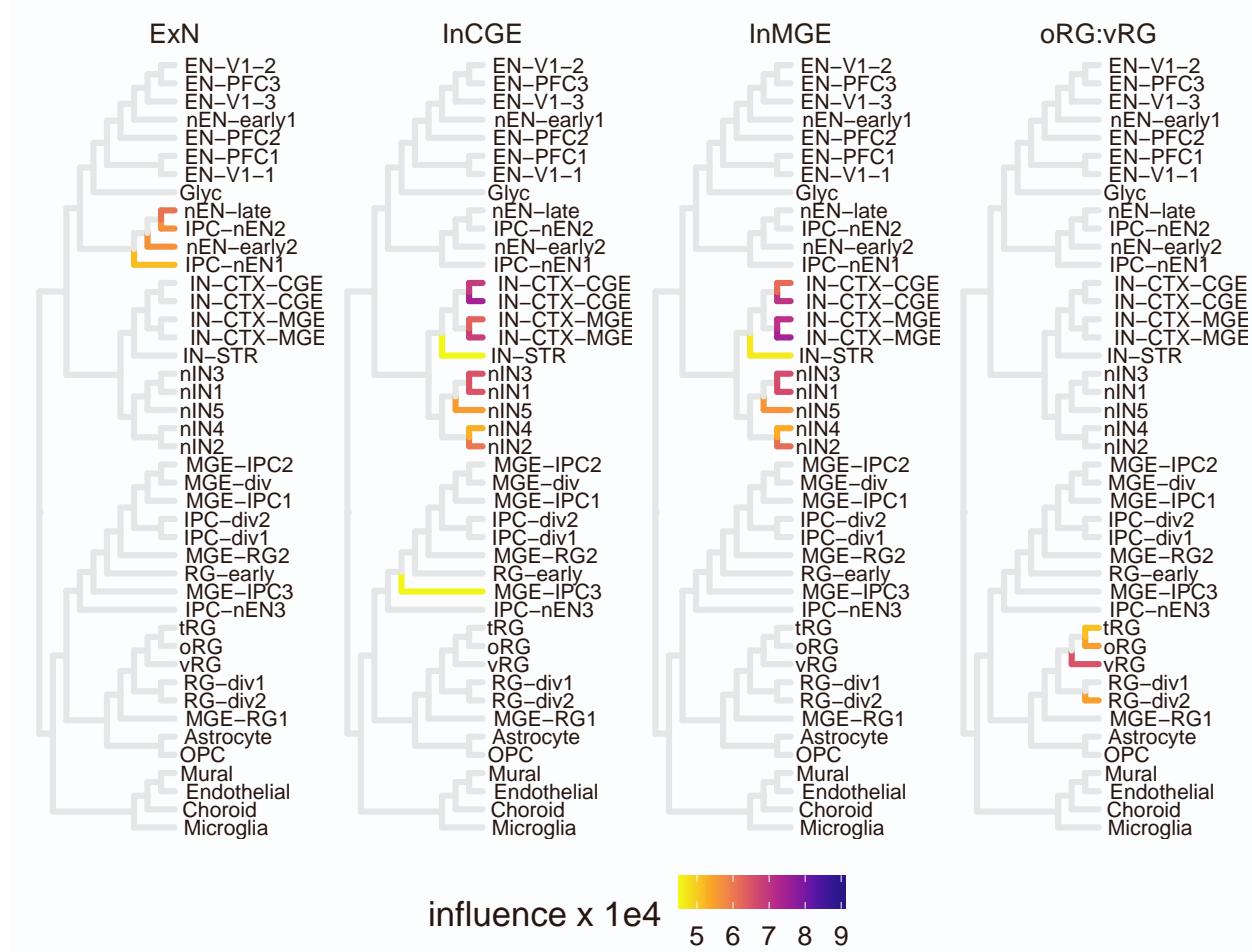


Figure S19: Mapping ExN, InCGE, InMGE and the ancestral cell type oRG:vRG in Polioudakis et al. [S6] onto the cell type hierarchy in Nowakowski et al. [S3] by influence scores ($> 1e - 4$ are shown), related to Figure 5A.

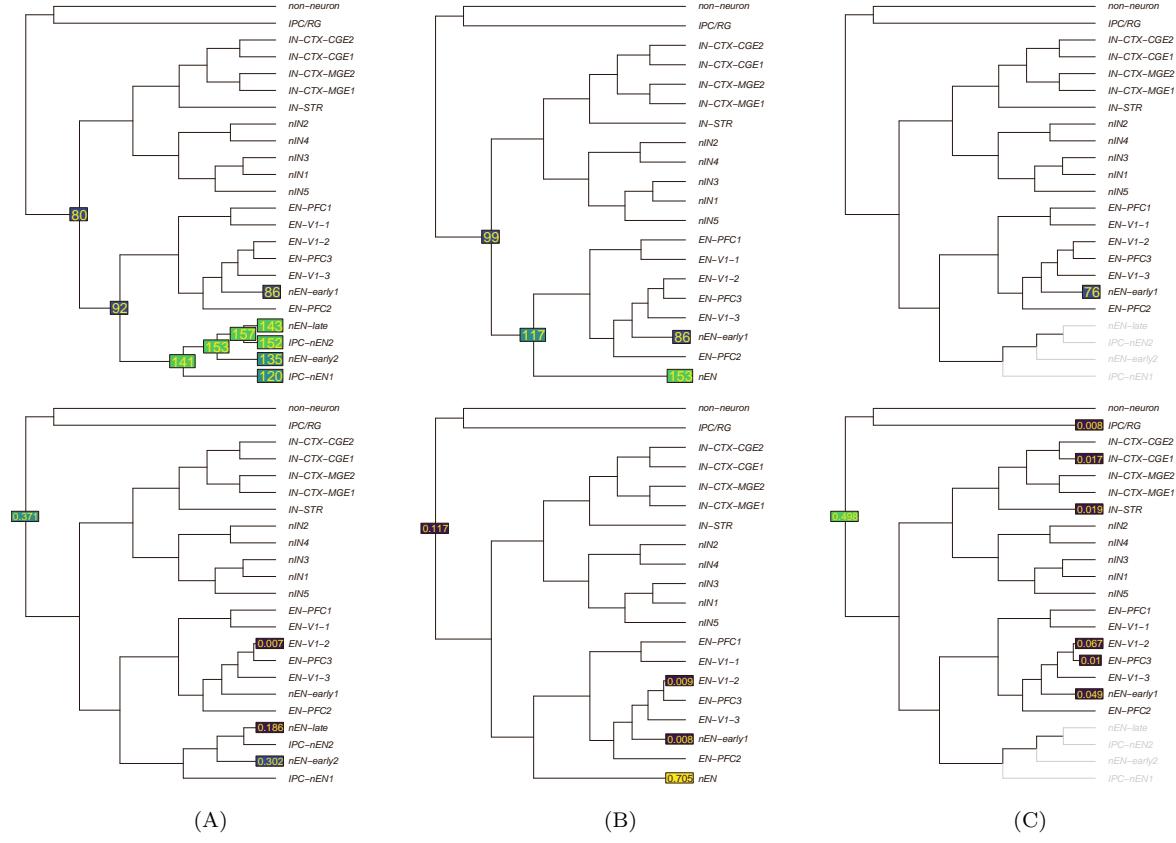


Figure S20: Mapping migrating excitatory (ExN) in Polioudakis et al. [S6] onto the hierarchical cell type ontology of Nowakowski et al. [S3] with newborn excitatory neurons subtree combined or removed, related to Figure 5A. (A) Original (B) Combine all cell types in the newborn excitatory neurons subtree (C) Remove all cell types in the newborn excitatory neurons subtree. treeArches and CellWalker2's cell type mapping results are shown across scenarios. Top: Z-scores from CellWalker2; Bottom: mapping probabilities from treeArches. The ranking of Z-scores is robust to either removing or combining related cell types, but treeArches' results can vary depending on which cell types are included in the dataset. In (C), treeArches showed increased scores to other related cell types, such as early born newborn excitatory neuron (nEN-early1) and early and late born excitatory Neuron V1 (EN-V1-2). This is consistent with the simulation of a ‘convergent cell type’, in which treeArches does not recover the similarity to the cell type from a different subtree and assigns large probability to the root. On the other hand, the Z-score mapping to ‘nEN-early1’ is similar or smaller when the subtree is excluded. The robustness of CellWalker2 to cell type composition might be due to using marker genes as well as the cells from each dataset. Furthermore, the marker genes of the remaining cell type labels stay the same even when cell type composition is changed. The rejection probabilities for treeArches are 0.10, 0.14 and 0.30 for these three scenarios respectively. Z-score cutoff is 75 and treeArches cutoff is 0.005. Only the top 10 nodes and above the cutoff values are shown.

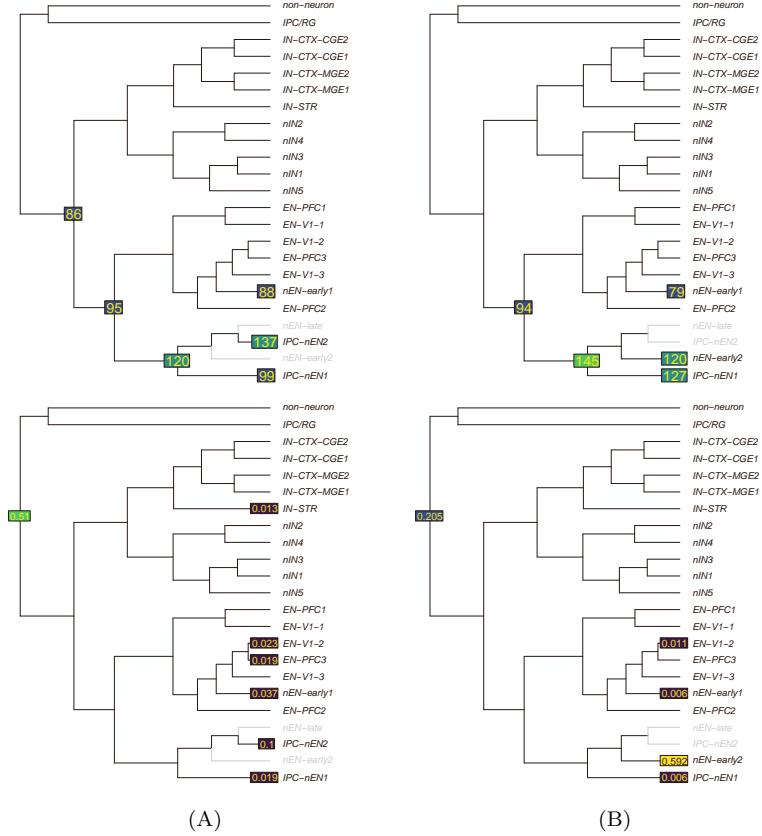


Figure S21: Mapping migrating excitatory (ExN) in Polioudakis et al. [S6] onto the hierarchical cell type ontology of Nowakowski et al. [S3] with top-mapped cell types removed, related to Figure 5A. (A) Remove nEN-late and nEN-early2, top cell types that ExN was mapped to by treeArches, treeArches assigned larger scores to ‘IPC-nEN2’ and ‘nEN-early1’ as well as the root node (or being rejected), while CellWalker2 shows similar Z-scores mapping to the remaining cell types. (B) Remove nEN-late and IPC-nEN2, top cell types that ExN was mapped to by CellWalker2. CellWalker2 showed similar or smaller Z-scores to the remaining cell types; while treeArches assigned larger scores to early born newborn excitatory neuron (nEN-early2). treeArches and CellWalker2’s cell type mapping results are shown across scenarios. Top: Z-scores from CellWalker2; Bottom: mapping probabilities from treeArches. The rejection probabilities for treeArches are 0.15 and 0.25 for these two scenarios respectively. Z-score cutoff is 75 and treeArches cutoff is 0.005. Only the top 10 nodes and above the cutoff values are shown.

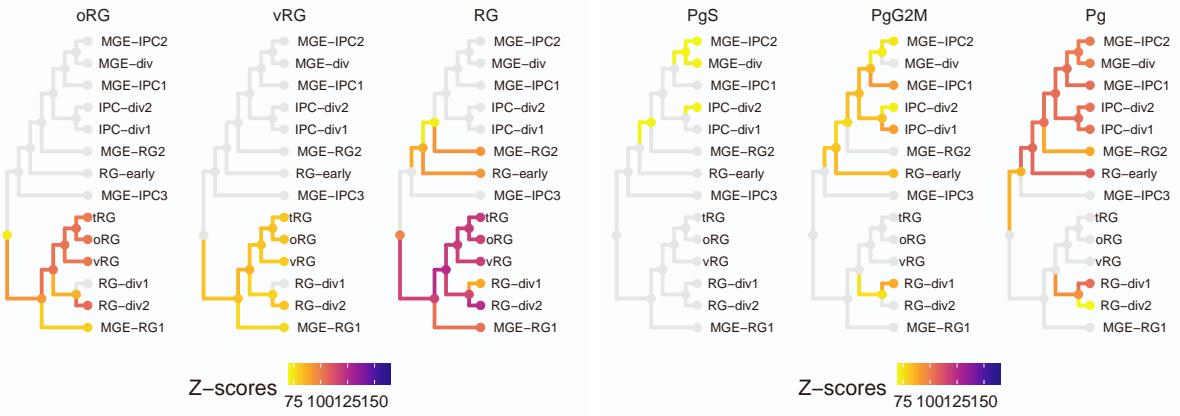


Figure S22: Mapping radial glias and cycling progenitors in [S6] onto the cell type hierarchy in [S3] by Z-scores (> 70 are shown), related to Figure 5C. RG represents the parent node of oRG and vRG, and Pg represents the parent node of PgS and PgG2M. oRG and vRG map to the radial glia subtree, but show no one-to-one correspondence with specific cell types. The RG node shows larger Z-scores mapping to the subtree of radial glia, with the largest Z-score at the internal node vRG:RG-div1. Cycling progenitors in G2M phase (PgG2M) maps to dividing radial glia in G2/M-phase (RG-div1), dividing intermediate progenitor cells RG-like (IPC-div1) and MGE Progenitors(MGE-IPC1). Cycling progenitors in S phase (PgS) maps to dividing intermediate progenitor Cells RG-like (IPC-div2) and dividing MGE Progenitors (MGE-div) weakly. The Pg node maps to dividing progenitors, MGE progenitors and early/dividing RG subtree with larger Z-scores, and the internal nodes of this subtree have the top Z-scores (e.g. RG-early:MGE-IPC2 and IPC-div1:MGE-IPC2).

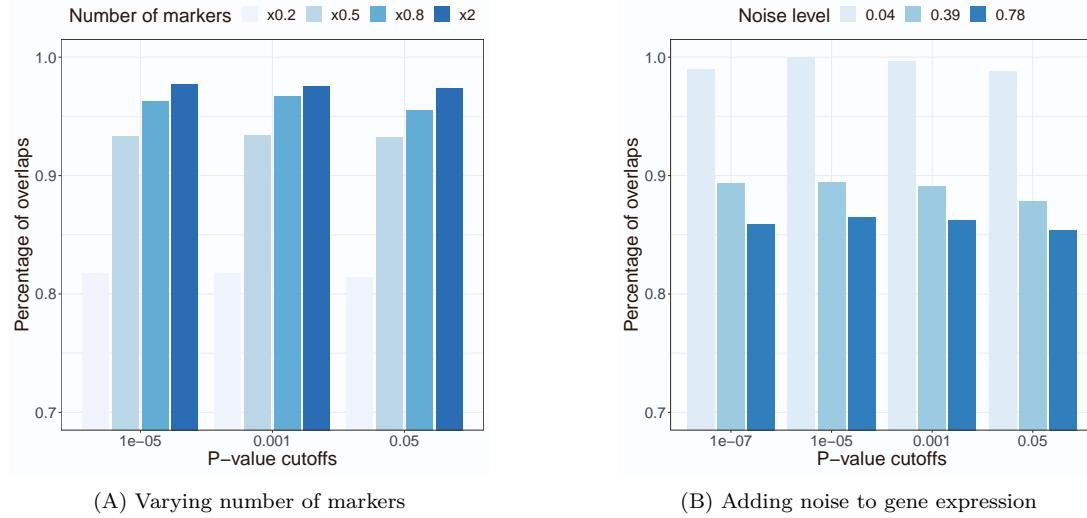
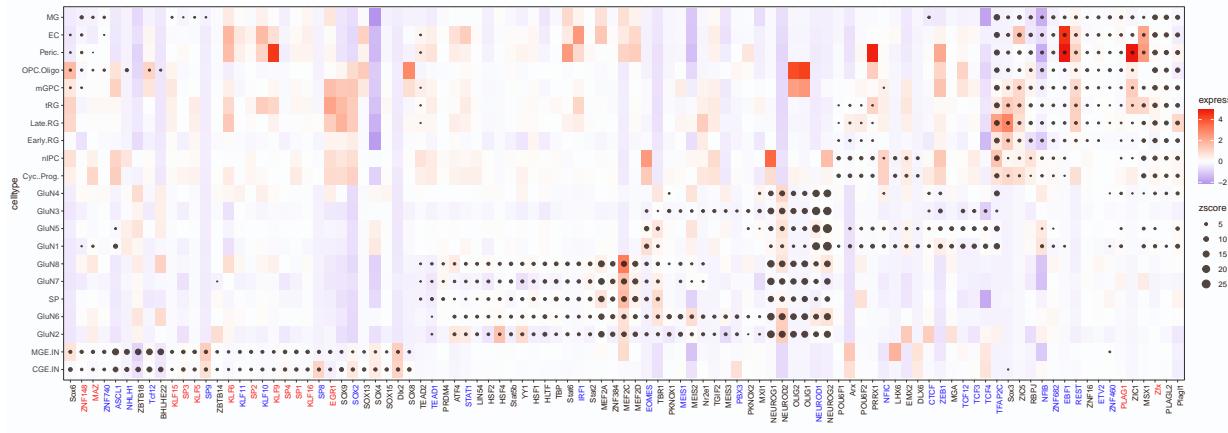
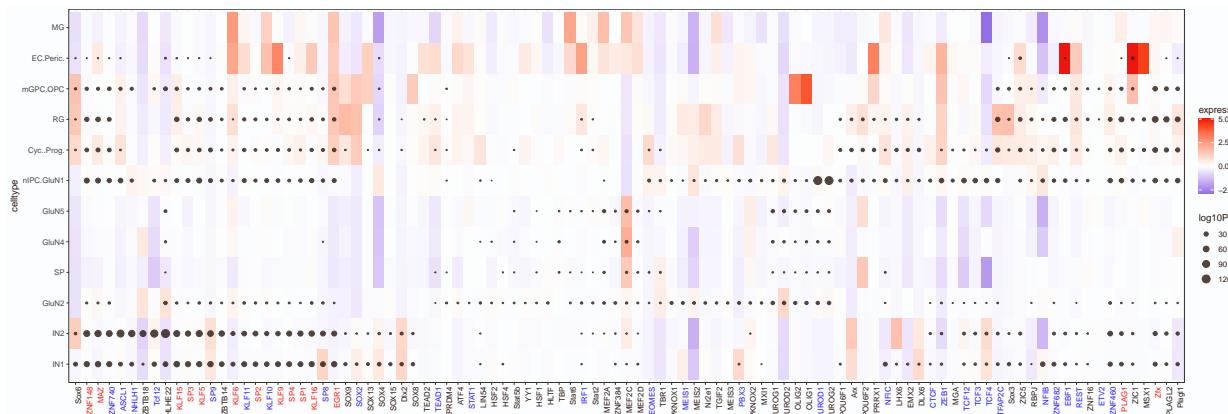


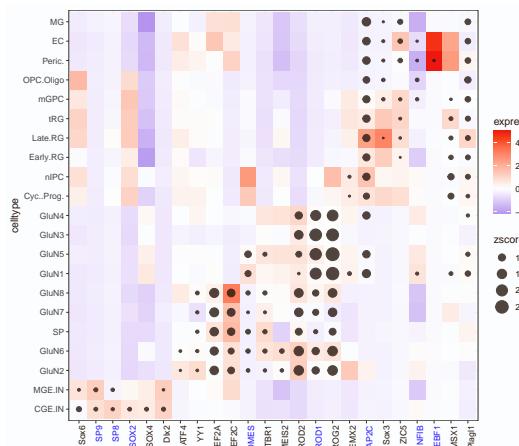
Figure S23: Robustness of CellWalker2's results for mapping cell types between Polioudakis et al. [S6] and Nowakowski et al. [S3], related to Figure 5A. (A) Varying the number of markers of each cell type from top 20%, 50%, 80% and 200% of the original markers in the order of log fold change. With fewer marker genes, the percentage of overlaps decreases. However, there are still above 80% overlap even with only 20% of markers. On the contrary, with twice the number of markers, the percentage of overlaps is above 95% at various significance levels. (B) Adding different levels of Gaussian noise to each PC coordinate of cells derived from gene expression matrix and reconstructing cell-cell graph. The noise level is the ratio of the standard deviation of noise and average of each coordinate. As the noise level increases, the percentage overlapping decreases. At noise level 0.78, which means that adding noise about 80% to the overall variability of the data, the percentage of overlaps is still above 85%. Interestingly, the percentage overlapping increases as we increase the noise level above the overall variation of the data (data not shown). We speculated that as cell-cell similarity is overwhelmed by noise, the label-to-cell information flow dominates the noisy cell-to-cell graph during CellWalker2's random walks. Label-to-cell edges can still guide towards the correct cell type mappings. Y-axis: the percentage of significant entries overlaps with original result. Suppose A and B are lists of entries above certain significance level under different experiments, we computed $|A \cap B|/|A \cup B|$ as the "percentage of overlaps". X-axis: different cutoffs for P-values (converted from Z-scores), where p-values are converted from Z-scores using the standard Normal distribution.



(A) CellWalker2's result for all motifs



(B) Signac's result for all motifs



(C) CellWalker2's result for selected motifs

Figure S24: Comparison between CellWalker2 and Signac using TF motifs in human developing cortex, related to Figure 5D. (A) CellWalker2 shows many TFs motifs (column) enriched in pREs that exhibit increased accessibility in particular cell types (row). (B) TF motif enrichment in differentially accessible regions across cell types that overlaps with pREs by Signac. (C) CellWalker2's results for selected TFs with Z-score >5 and normalized expression > 0.5 for at least one cell type from (A). The size of the dot is Z-score or $-\log_{10}P\text{value}$, and only Z-scores > 3 or adjusted p-value < 0.01 with BH procedure are shown. The color of the box is the normalized gene expression of a TF in a particular cell type. TFs colored as red/blue are universal/strip factors.

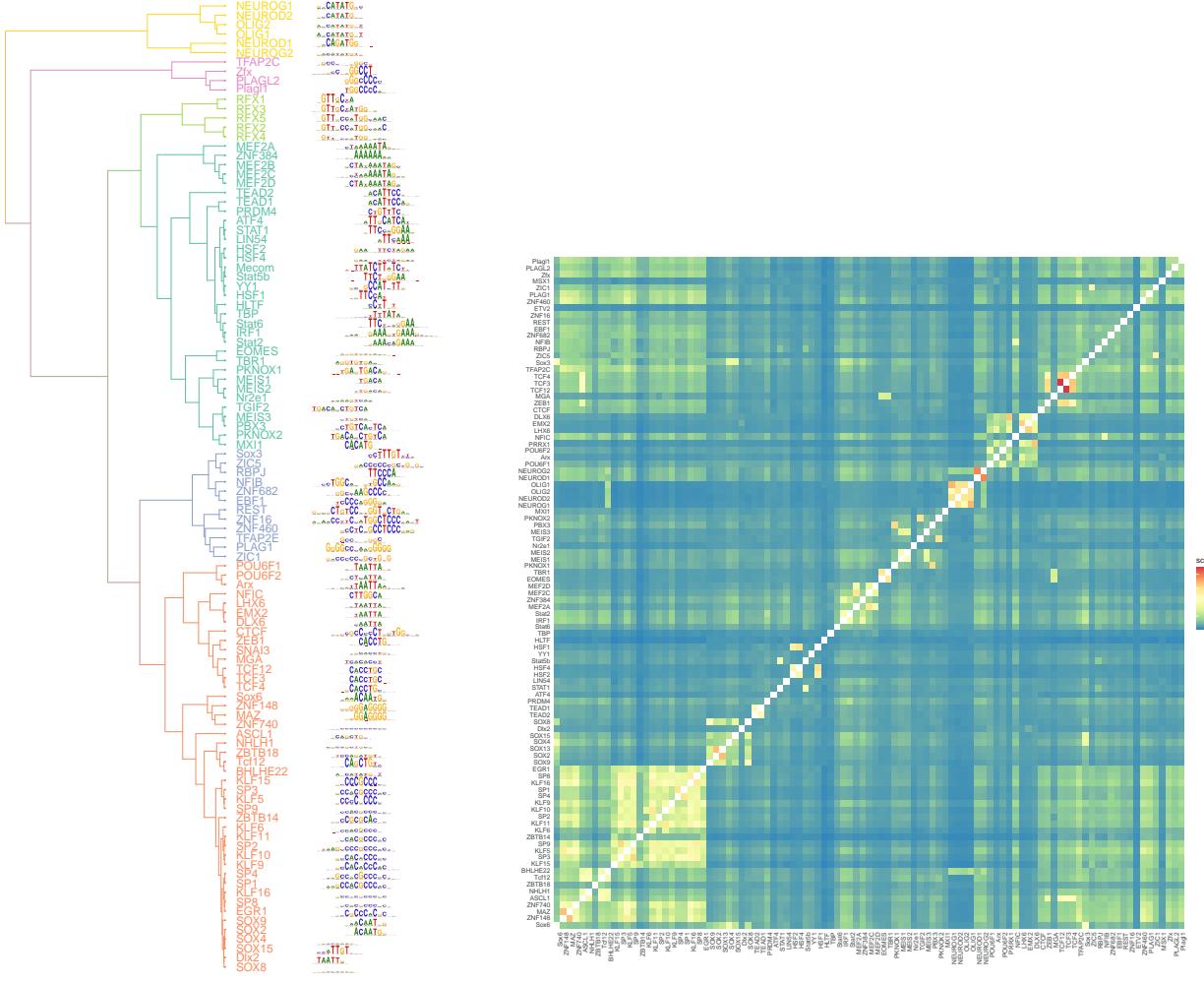
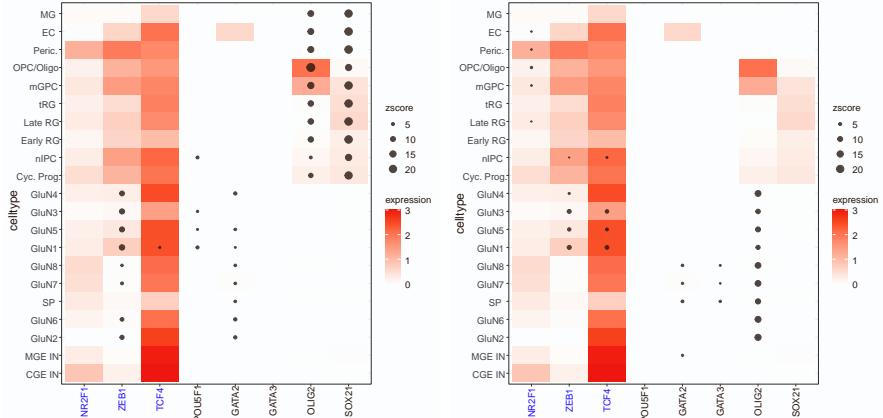


Figure S25: TFs that are clustered together by Z-score show high motif similarity and co-occurrence score, related to Figure 5D. (A) TFs with similar motifs have similar Z-scores. TFs are clustered based on the Z-score in Figure S24A. (B) TF motifs co-occurrence score within pREs. The co-occurrence score is computed as Jaccard distance of occurrence of TF motifs in pREs.



(A) CellWalker2's result using ChIP-Seq (B) CellWalker2's result using motifs (for TFs with ChIP-Seq data)

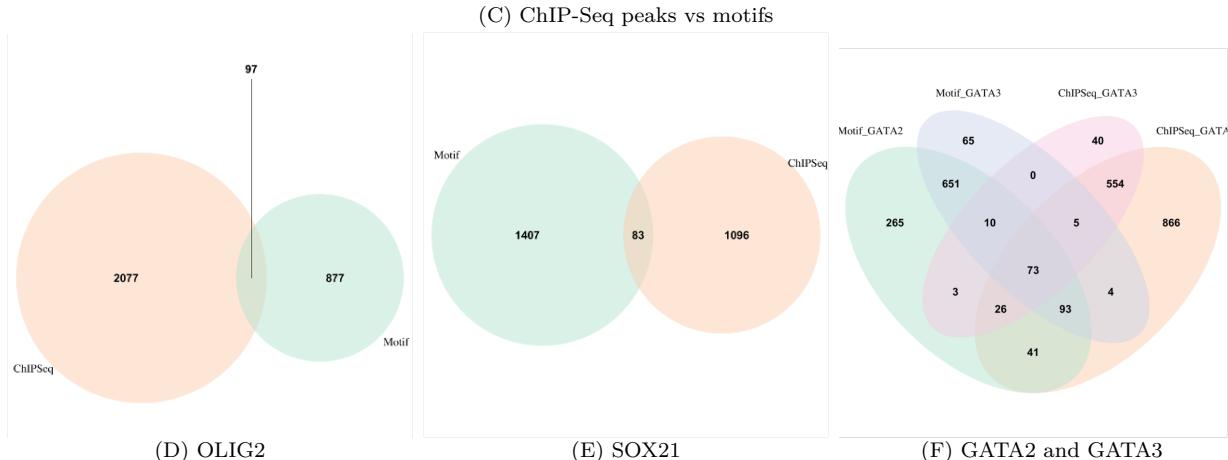
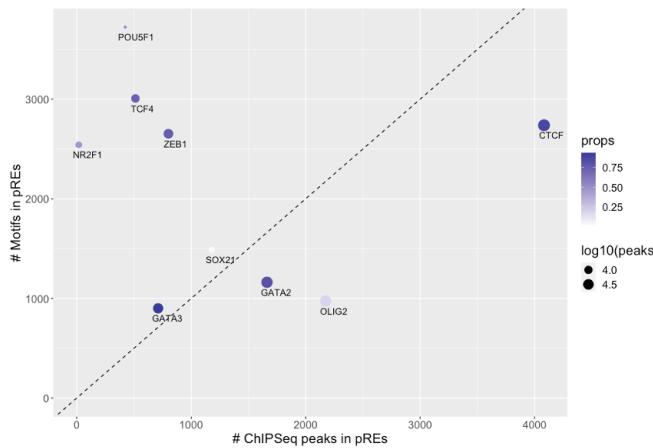


Figure S26: Comparison between using TF motifs with ChIP-Seq peaks in human developing cortex, related to Figure 5D and E. (A-B) Z-scores using either peaks or motifs for TFs with ChIP-Seq data in brain. Color represents log expression value. Using ChIP-Seq peaks identifies OLIG2 peaks enriched in OPC/Oligo cell types but not using motifs. As OLIG2s motif is similar to NEUROD2, which has higher expression in GluN, CellWalker2 using motifs fails to distinguish these two TFs. Using motifs, GATA2 and GATA3 has similar Z-scores in GluN; but using ChIP-Seq, GATA2 has higher Z-score. (C) The number of pREs overlapping with either ChIP-Seq peaks (x-axis) versus or TF motifs (y-axis). The size of the dot is $\log_{10}(\text{number of peaks})$ in the ChIP-Seq experiment and the color is the proportion of TF ChIP-Seq peaks having the motif. (D-F) Venn diagrams show the number of pREs overlapping with ChIP-Seq peaks or motifs or both for different TFs. For OLIG2 and SOX21, pREs with motifs and with ChIP-Seq peaks do not have many overlaps. pREs having both GATA2 and GATA3 motifs are more than that of ChIP-Seq peaks, and GATA2 has more peaks in pREs than GATA3.

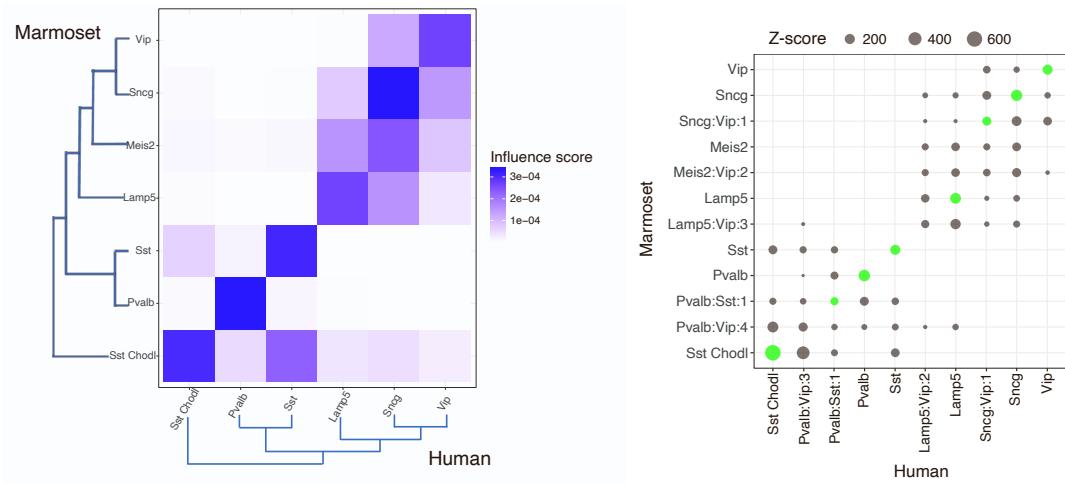


Figure S27: Comparing inhibitory neurons in motor cortex between human and marmoset at the subclass level, related to Figure 6A. CellWalker2's influence scores (left) and Z-scores (right) reflect the corresponding inhibitory neurons subclasses between human and marmoset. Subclasses cluster into two groups based on their Z-scores (Lamp5, Sncg, Vip and Meis2 versus Sst, Pvalb and Sst Chodl), which correspond with CGE- and MGE-derived inhibitory neurons. The only exception is marmoset Meis2 cells showing weak similarity to human Sncg cells but no strong mapping to any human subclass, consistent with the lack of Meis2 cells in human motor cortex. Each row is a marmoset cell subclass and each column is a human cell subclass, including ancestral nodes for Z-scores. The ancestral nodes are labeled with the two tips within its descendants and the depth of the node. The one-to-one relationship of the nodes on the human and marmoset tree is shown as green circle. Only Z-scores > 3 are shown.

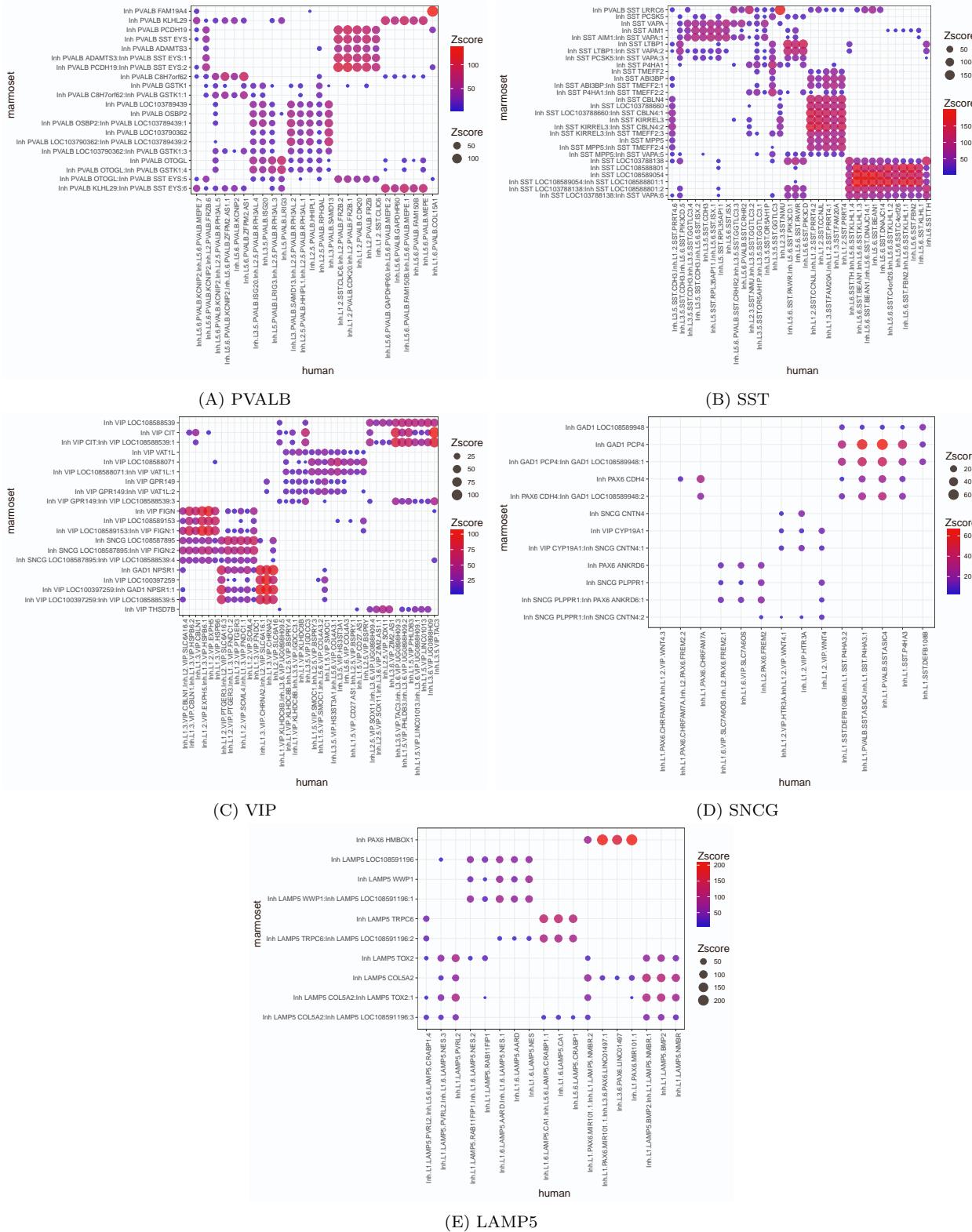


Figure S28: Comparing cell types within each GABAergic neuron subclass between human and marmoset using CellWalker2, related to Figure 6B. Each row is a marmoset cell type, and each column is a human cell type. Z-score is represented by the size and color of the dot. Each panel is a subclass.

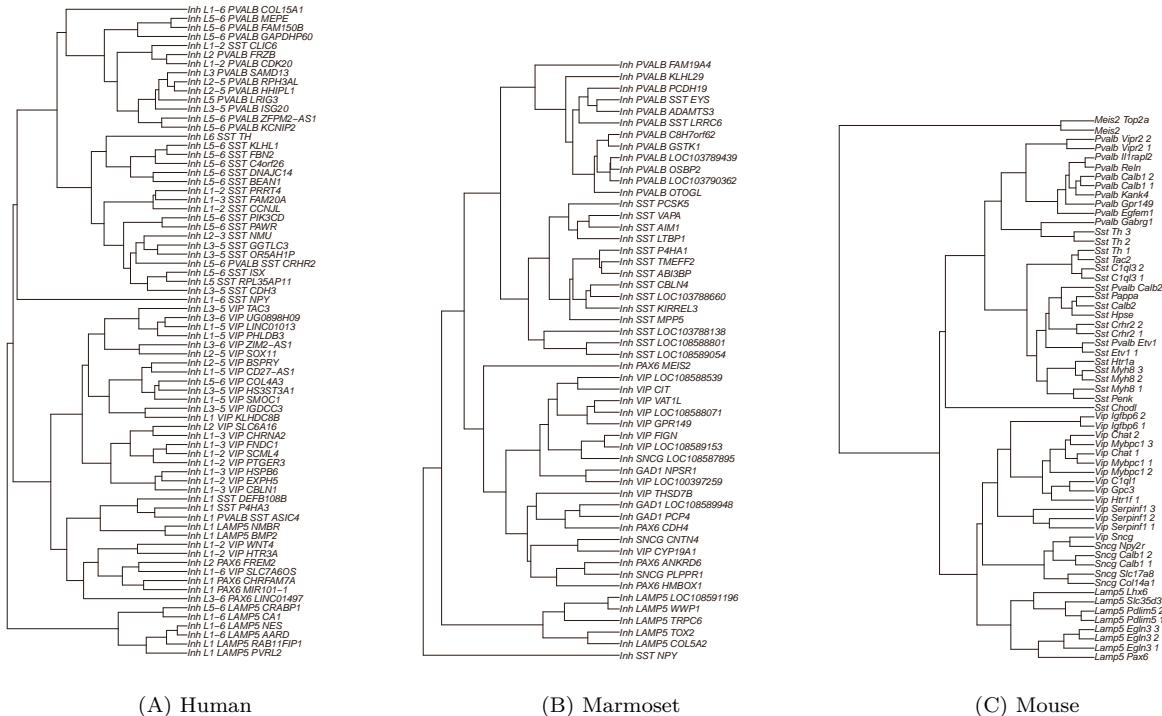


Figure S29: Cell type trees in human, marmoset and mouse, related to Figure 6A-C. Adapted from [S7].

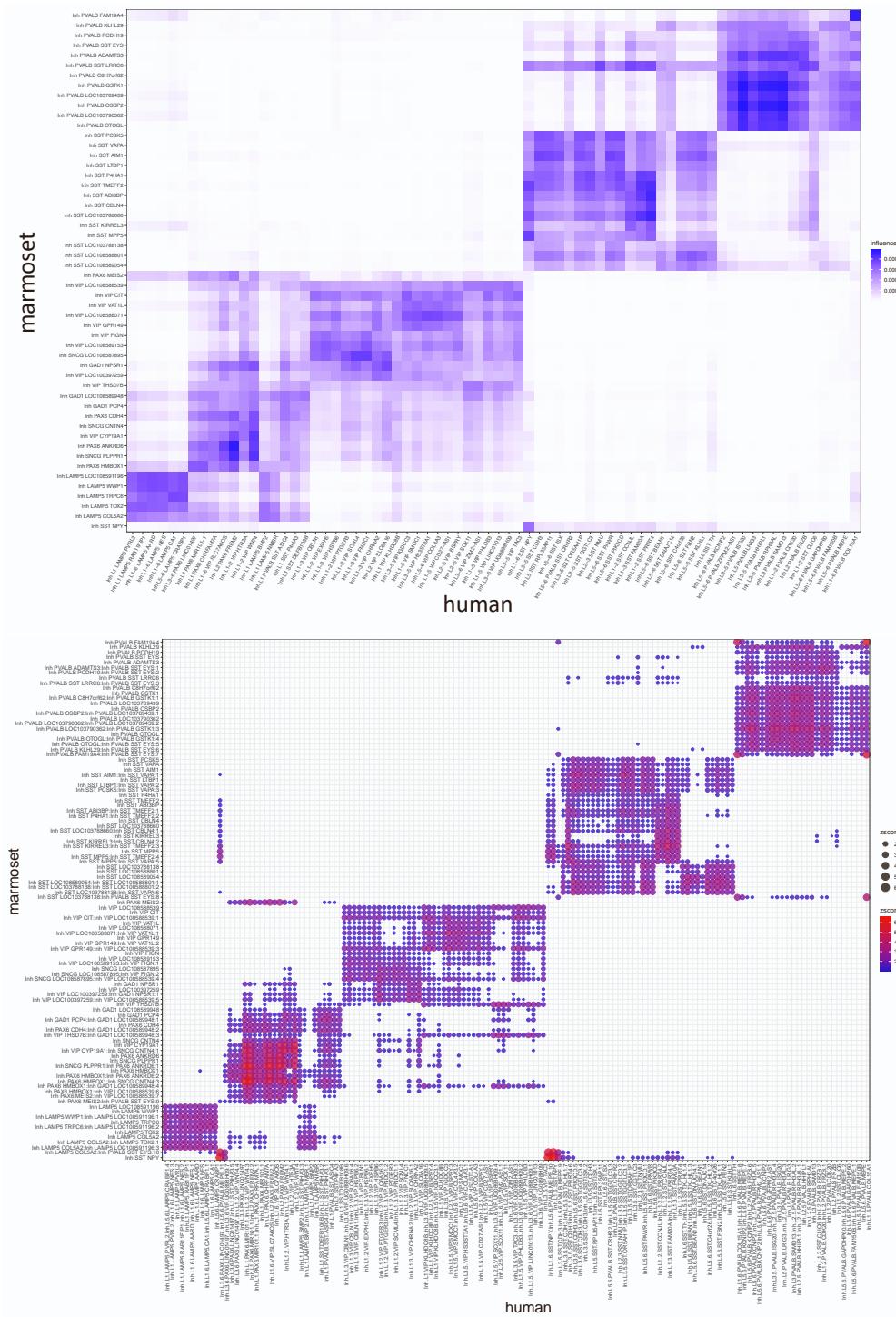
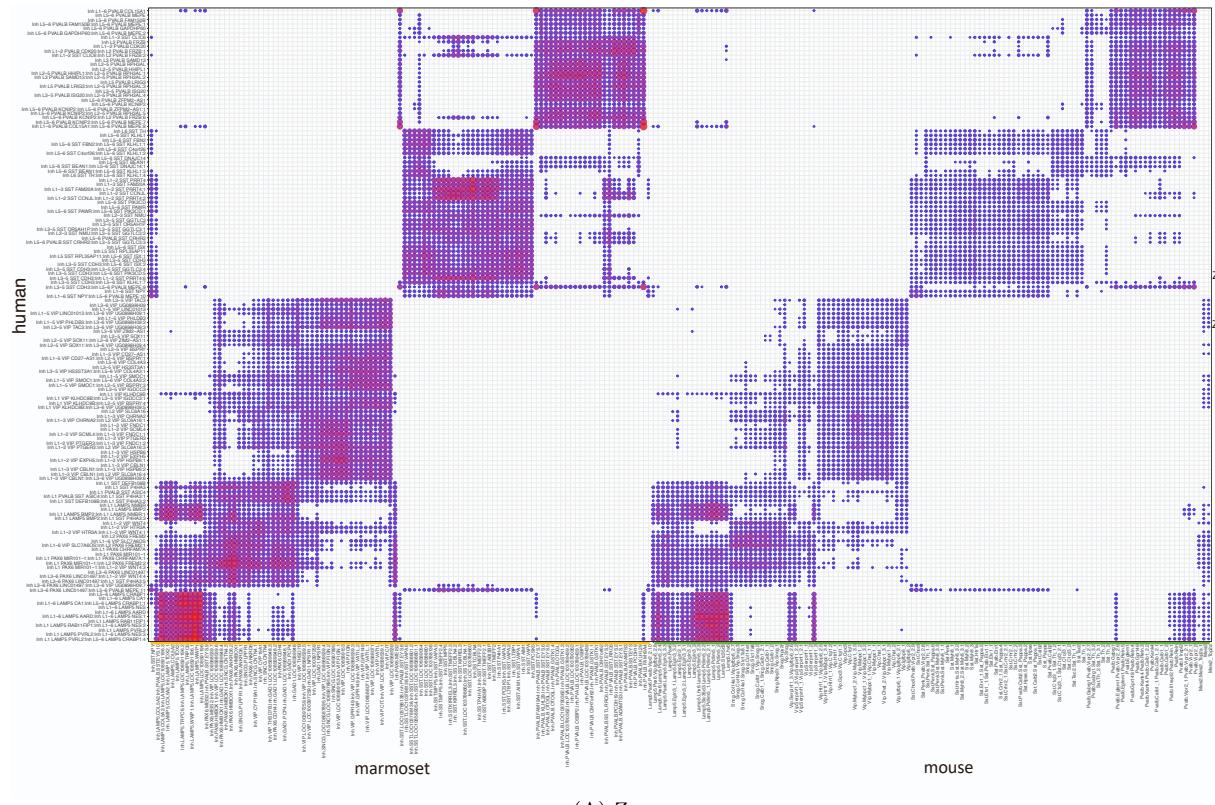
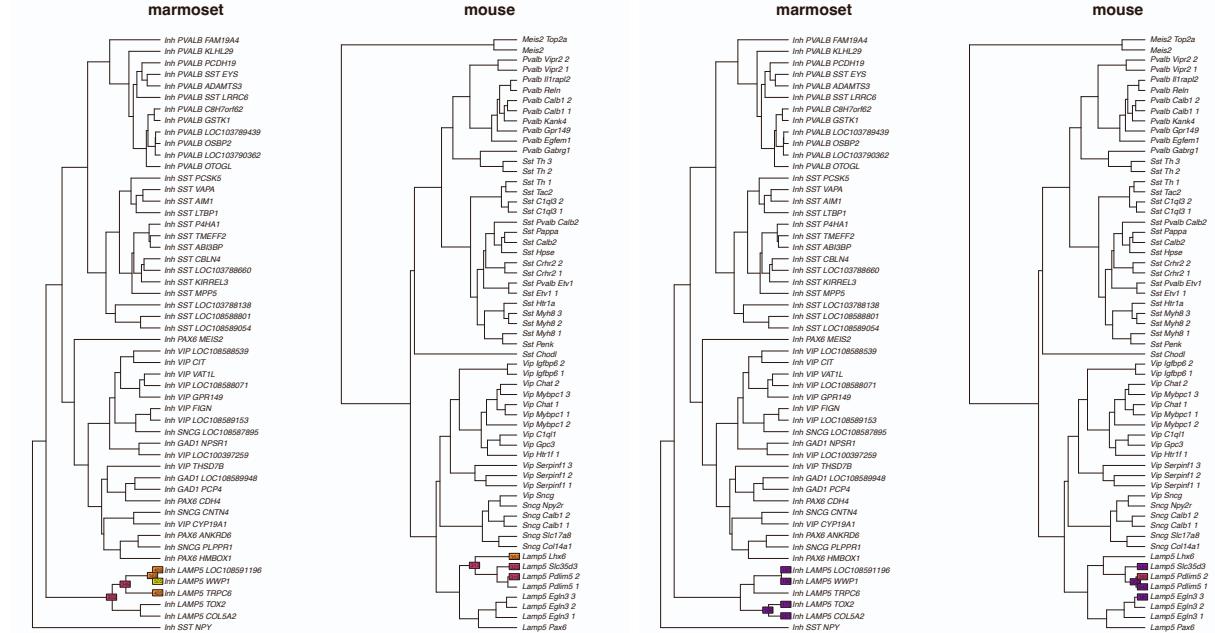


Figure S30: Comparing different cell types of inhibitory neurons in motor cortex between human and marmoset, related to Figure 6A-C. Top: influence scores from mapping marmoset to human cell types. Bottom: Z-scores from mapping the human and marmoset cell type trees (including ancestral nodes). The marmoset Sst chodl cell types (Inh SST NPY) is separated from other Sst cell types, while this cell type is grouped together with other Sst cell types in the human ontology. Although human and marmoset Sst chodl cell types have high Z-scores for each other, human Sst chodl cells also map to another subtree of marmoset Sst cell types, in which Inh SST MPP5 is also labeled as Sst chodl in the consensus taxonomy [S7]. A marmoset Sst cell type (Inh PVALB SST LRRC6) that express both *PVALB* and *SST* and shows similarity not only to the human Sst subtree but even higher similarity to the Pvalb subtree, consistent with this cell type having features of both Pval and Sst cells. Conversely, on the human cell ontology, a subtree within Sst subclass (Inh L5-6 SST DNAJC14 and Inh L5-6 SST BEAN1) not only maps to the marmoset Sst subtree but also to some Pvalb cell types. Only Z-scores > 100 are shown.



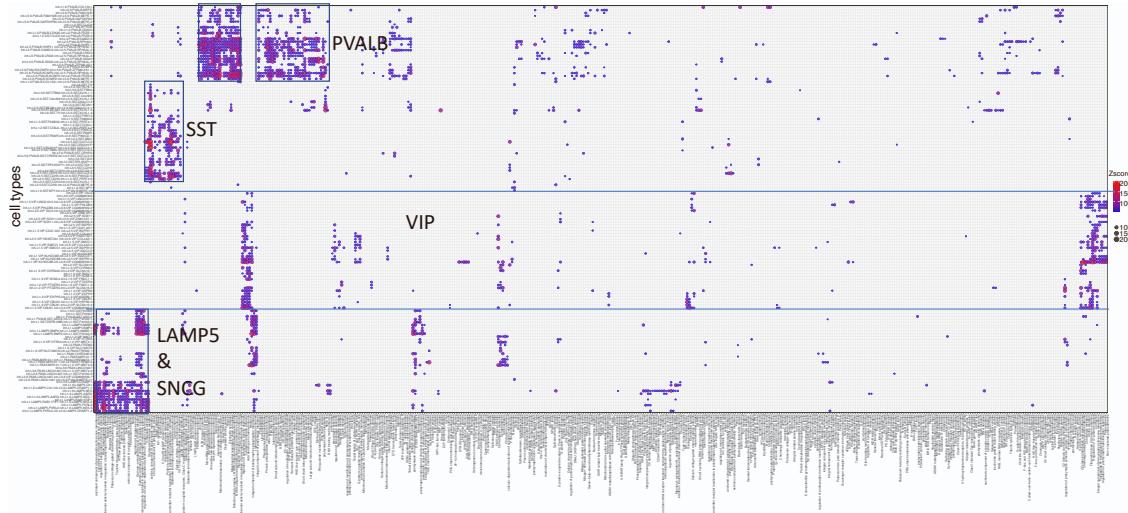
(A) Z-score



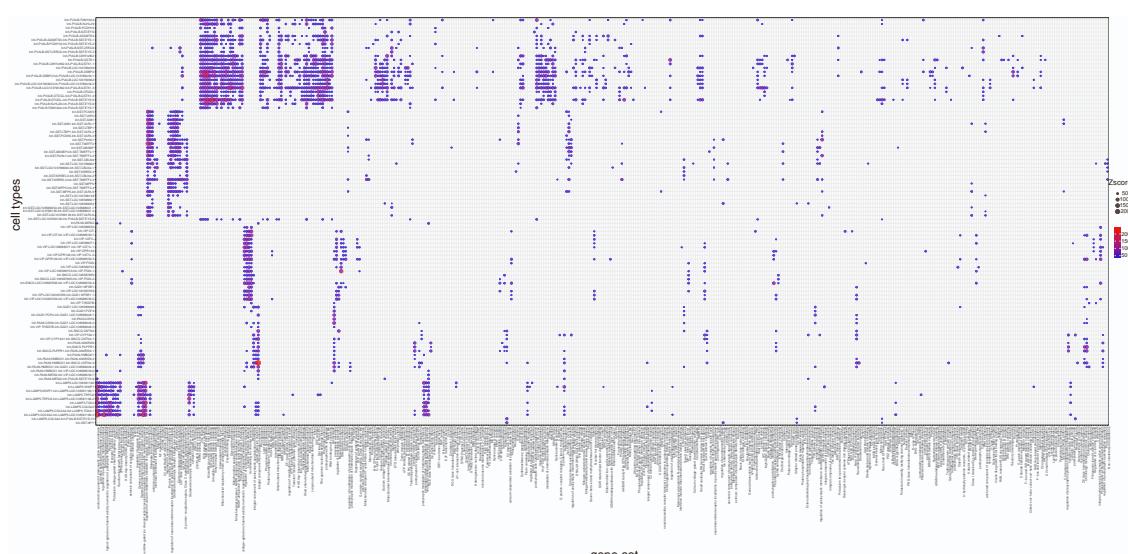
(B) LAMP5 AARD

(C) LAMP5 NMBR

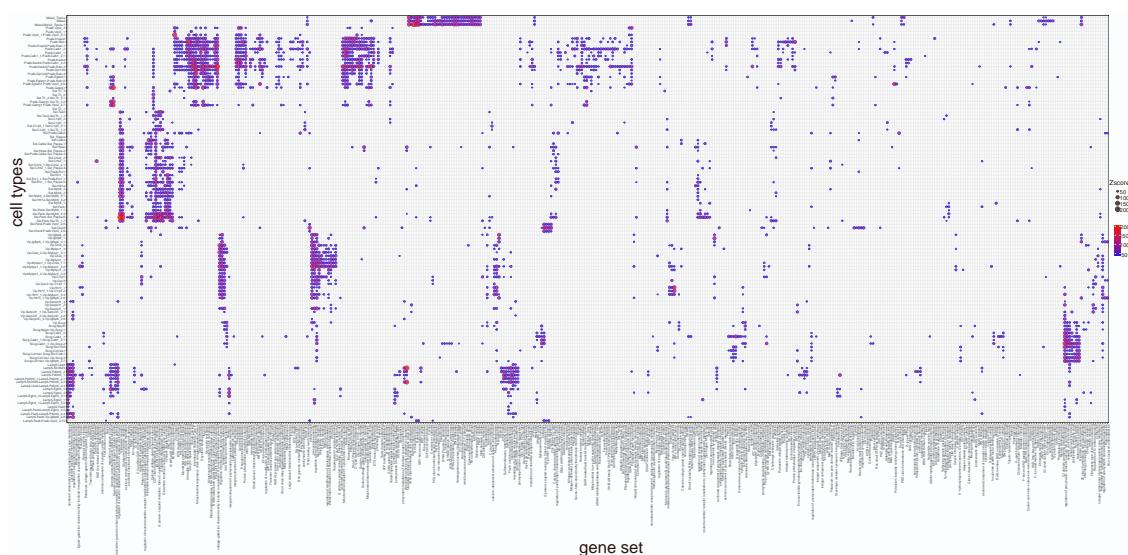
Figure S31: Comparing human cell types to marmoset or mouse cell types by CellWalker2's Z-score, related to Figure 6A-C. (A) The mapping between human and marmoset cell type trees (including ancestor nodes). Each row is a human cell type, and each column is a cell type from marmoset or mouse. Only Z-scores > 3 are shown. (b-c) Top cell types on marmoset or mouse cell tree that are mapped to human LAMP5 AARD (B) or LAMP5 NMBR (C). Top 10 nodes with largest scores are shown. The color of the node represents the magnitude of the Z-score and the number on each mapped nodes is the value of the Z-score.



(A) human



(B) marmoset



(C) mouse

Figure S32: Mapping active gene sets to cell types for different species, related to Figure 6D. (A)-(C) The heatmap shows the Z-scores from CellWalker2, where each row is a cell type (including ancestral nodes), and each column is a gene set. The color and size of the dot represents Z-score. TALE class homeobox transcription factors, which include Meis2, PBAF/BAF complex, SRY-box, and E2F, map to Meis2 and Top2a cell types that belong to the Meis2 subclass that is absent from the human ontology. Only top 5% of the entries are shown.

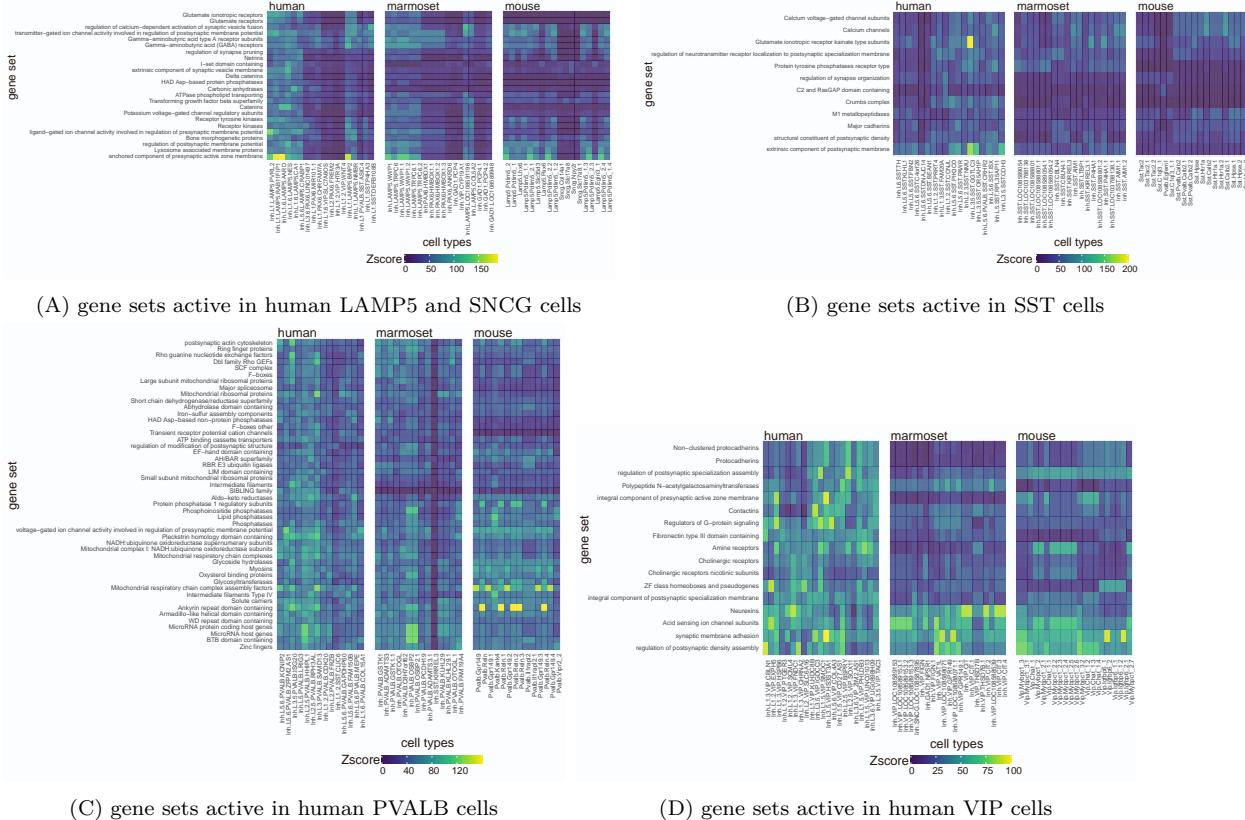


Figure S33: Gene sets that are activated in different subclasses of human cells show different cell type specificity across species, related to Figure 6D. The three panels of each heatmap show the Z-scores between gene sets and cell types in human, marmoset and mouse. In each panel, each row is a gene set, each column is a cell type and the color represents Z-score. For each human cell type, the matched cell type by CellWalker2 in marmoset or mouse is shown. Glutamate receptors and Catenins, which are important for normal cognitive development, have higher expression specificity for several cell types within the Lamp5/Sncg subclass in human versus mouse cells, and to a lesser degree compared to marmoset cells (A). Similarly, Glutamate ionotropic receptor kainate type subunits have a very high Z-score in the human SST GGTLC3 cell type, but lower Z-scores in marmoset and mouse (B). Several genes in this set (*GRIK1* and *GRIK3*) are marker genes of the human SST GGTLC3 cell type, and the autism spectrum disorder associated gene *GGTLC3* is involved in gamma-glutamyltransferase activity [S8]. Moreover, some pathways related to postsynaptic membrane have higher activity in human Sst cells than marmoset and mouse.

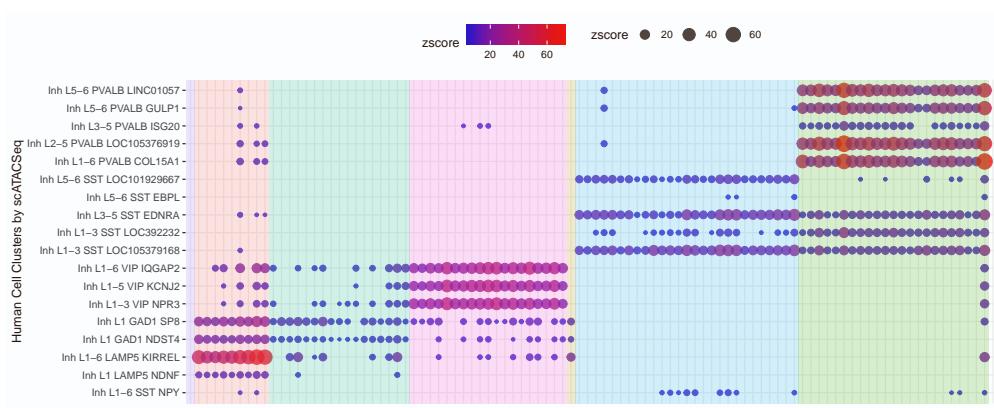


Figure S34: Labeling human cell type-specific DARs by marmoset cell types of inhibitory neurons using Z-score, related to Figure 6E. Each row is a human cell type cluster identified by scATAC-Seq data and each column is a node on the marmoset cell type tree. Only Z-scores > 3 are shown. The color and size of the dot represent the Z-score. The background color indicates the cell subclass of marmoset.

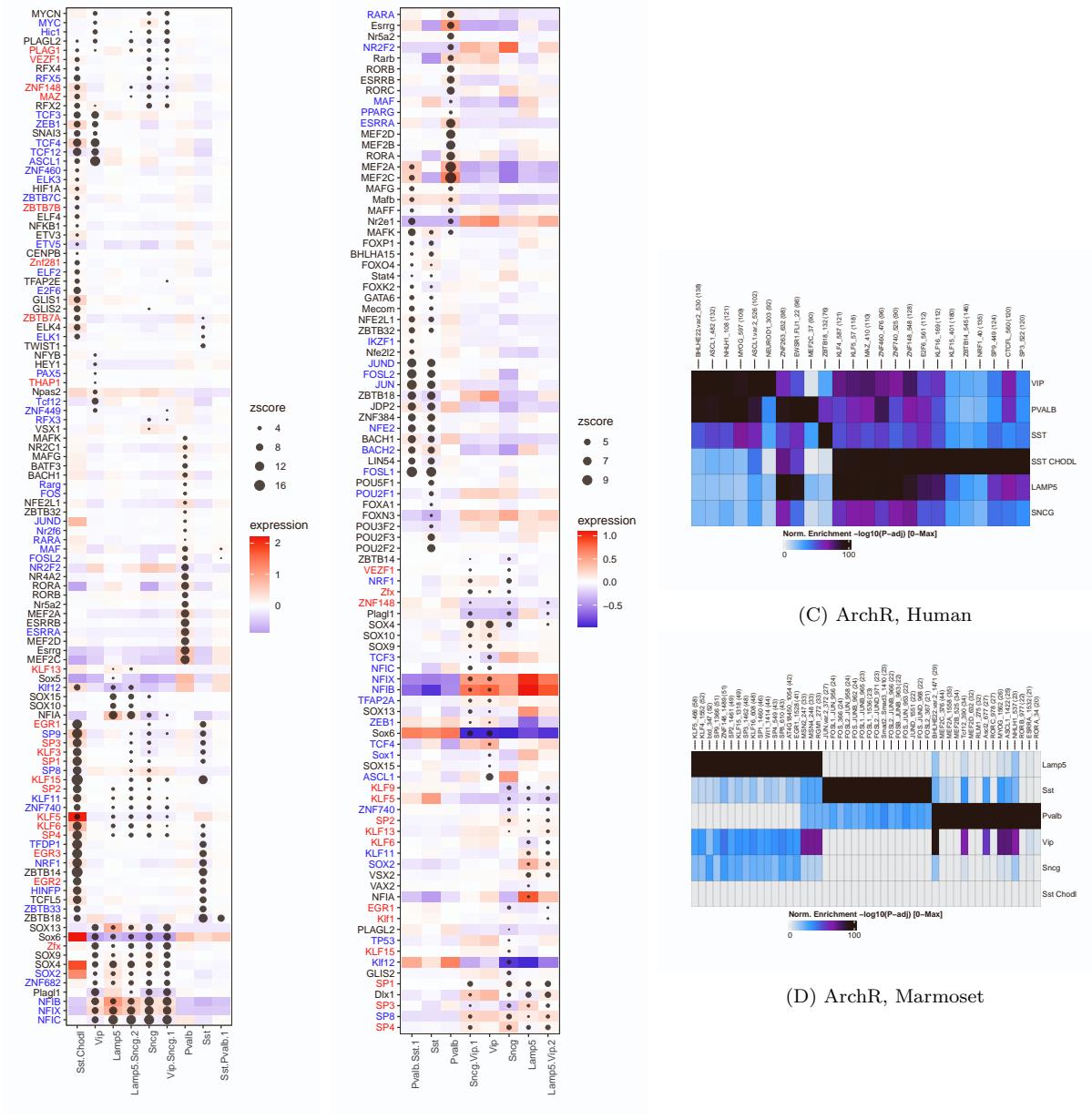
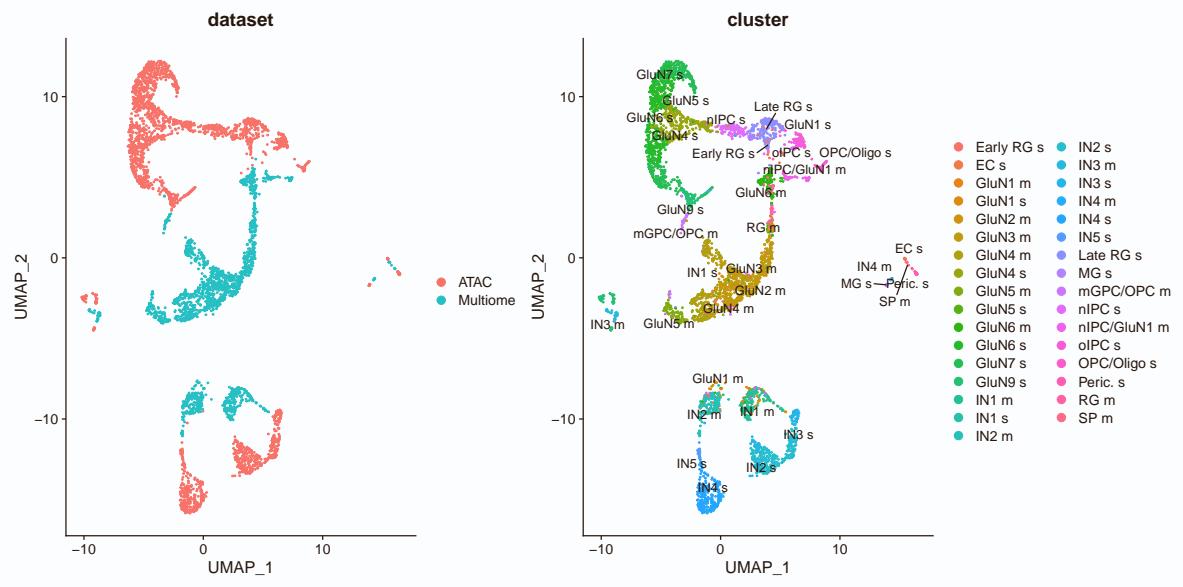
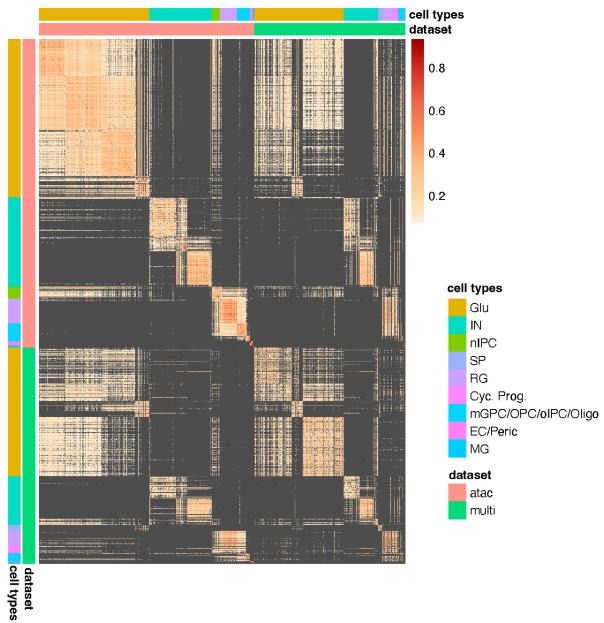


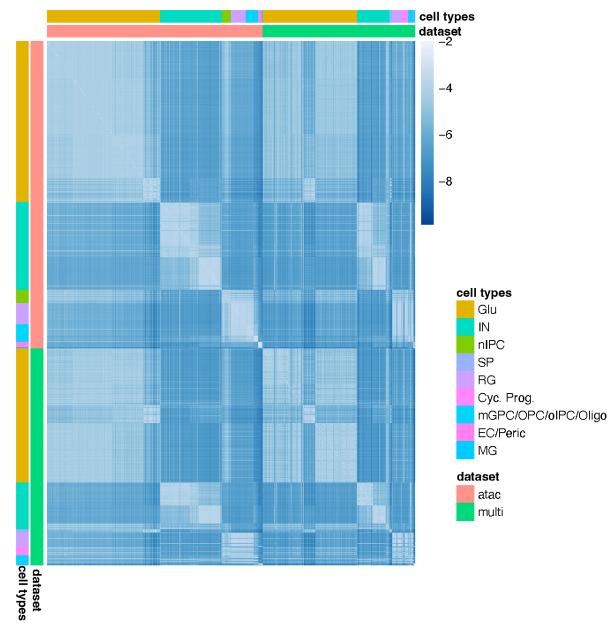
Figure S35: Mapping TFs to marmoset or human cell subclasses for inhibitory neurons using CellWalker2 and ArchR, related to Figure 6E. For CellWalker2, each row is a transcription factor (TF), and each column is a node on the cell subclass tree. Only Z-scores > 3 are shown. The size of the dot represents the Z-score, and the color represents the TF's standardized expression level. Sst chidl is excluded for Marmoset because it has too few cells. For ArchR, each row is a cell subclass, and each column is a TF. The color represents $-\log_{10}$ adjusted Pvalues. Top 7 enriched TF motifs are shown for each cell subclass. Minimum $-\log_{10}$ P-adj included in the plot is 20. We used CellWalker2's direct output, without filtering for expression, because ArchR does not do so. CellWalker2 identifies more cell subclass-specific TFs compared to ArchR and also some TFs shared among similar cell subclasses. Z-scores separate TFs into two major groups, one corresponding to CGE lineages (Vip, Sneg, and Lamp5 subclasses plus their ancestral nodes) and the other to MGE lineages (Pvalb and Sst subclasses plus their ancestors). Moreover, while CellWalker2 identified several TFs unique to Pvalb and Vip cells, which originate from distinct lineages (MGE- versus CGE- derived), while ArchR mostly identified TFs that are shared between Pvalb and Vip cells. In addition, CellWalker2 identified a group of TFs having higher Z-scores in both Sst chidl and Sst cells, probably due to the similarity between Sst chidl and Sst cells and the small sample size of Sst chidl cells.



(A) UMAP



(B) Cell-to-cell distance



(C) Influence matrix

Figure S36: CellWalker2's cell-to-cell graph integrates single cell datasets with batch effect, related to Figure 5D. (A) UMAP shows batch effect between ATAC-Seq data and multiomic data (ATAC-Seq part) from pcw21. UMAP is generated using Seurat. We combined the peaks from both ATAC-Seq data, ran SVD on the TF-IDF of count matrix, and obtained UMAP based on the top 30 components. Left: colored by data sources; right: colored by cell type from the original annotation in [S2]. Suffix 's' is cell annotations from the ATAC-Seq data and 'm' is from multiomic data. Cells from different datasets are separated. (B) and (C) Cell-to-cell distance and influence matrix for ATAC-Seq and multiomic cells (ATAC-Seq part) by CellWalker2. Cells are grouped by dataset and cell types, indicated by color bars on the margins of the heatmap. Subsampled 3000 cells are shown. Similar cell types within the same batch show higher connectivity, however, similar cell types between batches still show connectivities at a lesser extent. (B) cell-to-cell distance is measured by Jaccard distance on KNN graph, where the KNN graph is built using cosine similarity between peak profiles of each pair of cells. (C) shows the log10 influence score computed using cell-to-cell distance in (B).

Node Types	SubTypes	Cells			Labels	Annotations	
		RNA-Seq	ATAC-Seq	Multimodal		Gene sets	Genomic coordinates (include TF binding sites)
Cells	RNA-Seq	Correlation of gene expression		Correlation of gene expression	Weighted average of cell type markers' standardized expression	Average standardized expression of genes in the set	
	ATAC-Seq		Cosine similarity of peaks presence or absence	Cosine similarity of peaks presence or absence			Accessibility of the genomic regions (# reads in the overlapping ATAC-Seq peaks)
	Multimodal			Weighted average of gene expression and peaks' similarities	Weighted average of cell type markers' standardized expression		Accessibility of the genomic regions (# reads in the overlapping ATAC-Seq peaks)
Labels	Cell Type labels				(Hierarchical) relationships between cell types		

Figure S37: Possible type of nodes and edges in the CellWalker2 graph, related to STAR Methods. Each cell explains how CellWalker2 computes the edge weight between pairs of nodes (the row and the column of the cell). Blank white cell means that there are no edges between those two node types.

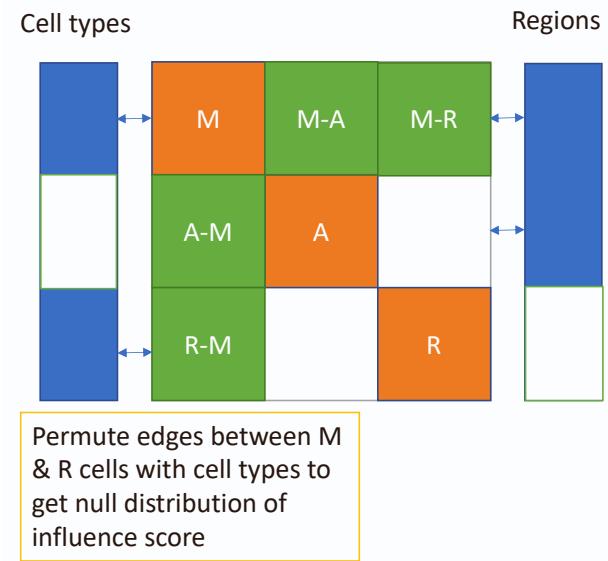


Figure S38: Illustration of integrating multiome data with scRNA-Seq and scATAC-Seq using CellWalker2, related to STAR Methods. The left column represents cell types, the middle matrix is the cell-to-cell similarity matrix, and the right column represents annotations (i.e., different sets of genomic coordinates). A is the submatrix between cells with ATAC-Seq data only, whose distance is the Jaccard or Cosine distance computed from open chromatin peaks. R is the submatrix between cells with RNA-Seq data only, in which scaled Euclidean distance of gene expression in PCA space is used. M is the submatrix between cells with multiome data, whose distance is a weighted combination of RNA-Seq and ATAC-Seq distance. M-A and A-M are the cell similarity between cells with multiome and with scATAC-Seq data, in which Jaccard or Cosine similarity is computed from chromatin peaks. M-R and R-M are the cell similarity between cells with multiome and cells with scRNA-Seq data by computing Euclidean distance of gene expression in PCA space. M & R cells are connected to cell types by expression of markers. M & A cells are connected to regulatory regions by reads in each region. Blank areas have no edges. CellWalker2 computes a KNN graph based on the similarity matrix and uses shared KNN as edge weight. To estimate Z-scores for mapping from annotations to cell types, CellWalker2 permutes the edges between M & R cells and cell type labels to get the null distribution of the influence score.

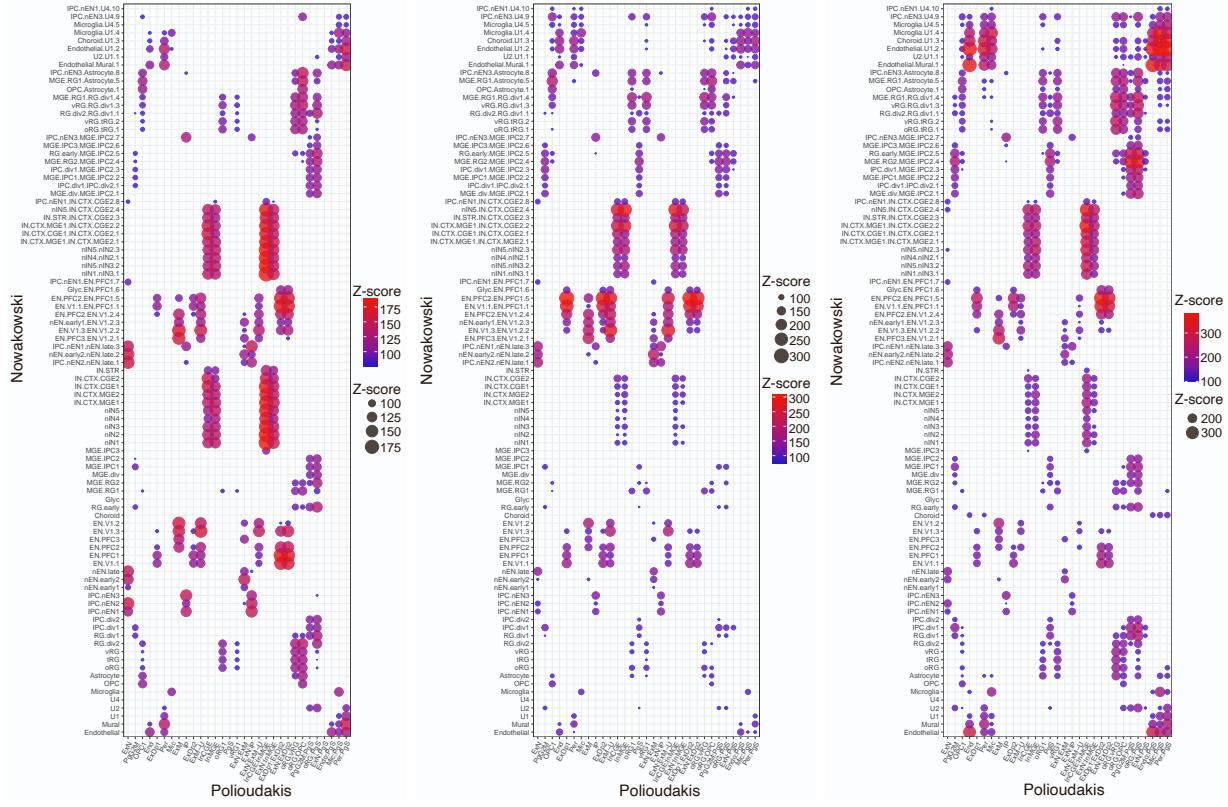
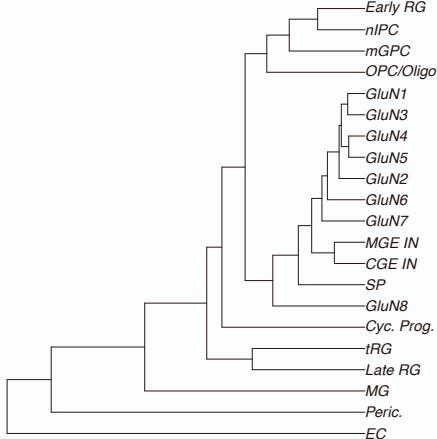
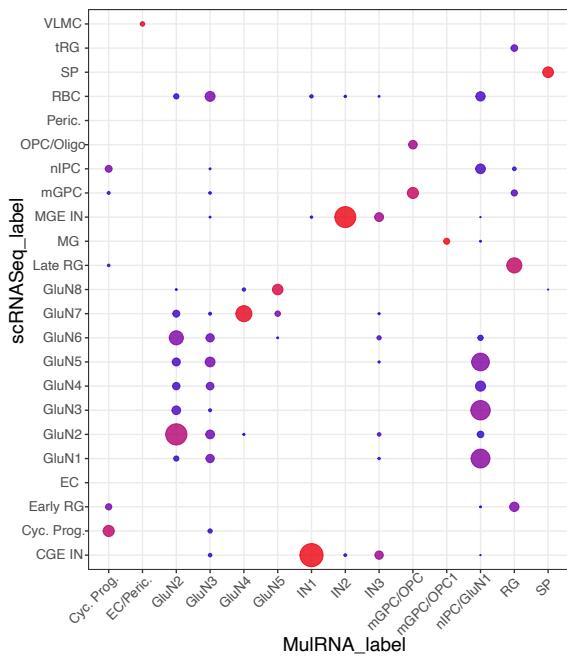


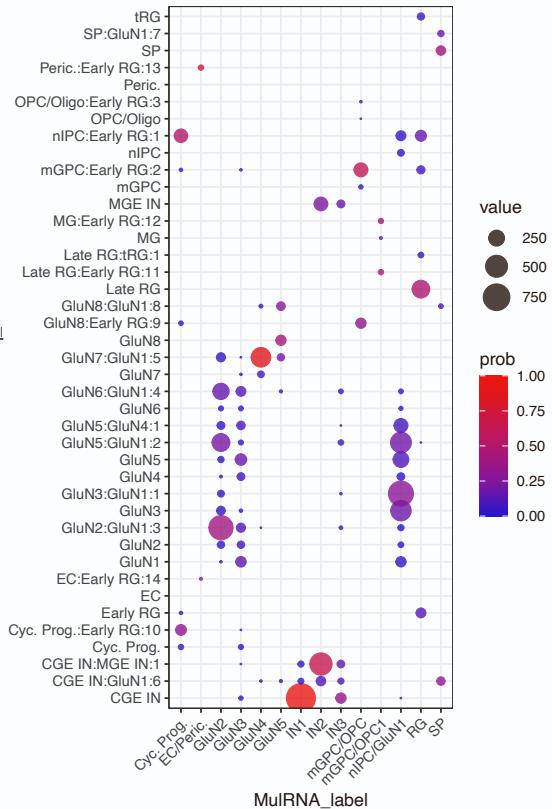
Figure S39: Different permutation schema for computing Z-scores for mapping cell types between the Polioudakis et al. [S6] and Nowakowski et al. [S3] datasets, related to STAR Methods. We permute edge weights between cells and cell type ontology of Nowakowski et al. (left), between cells and cell ontology of Polioudakis et al. (middle), or between cells and both cell type ontology (right).



(A) Cell type tree from scRNA-Seq

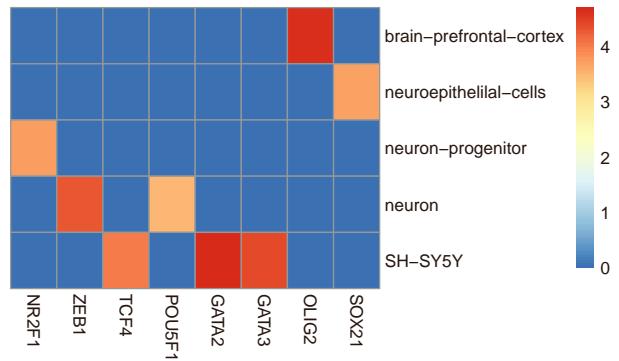


(B) tip cell types

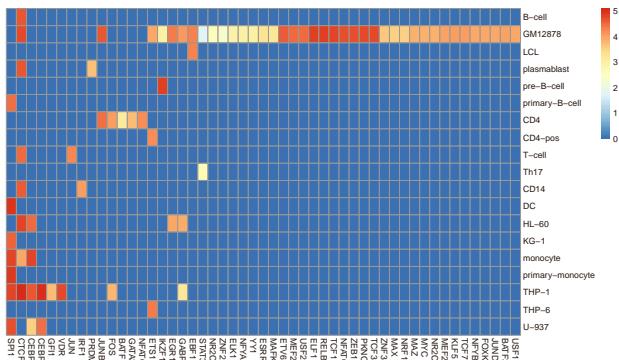


(C) all nodes on the cell type tree

Figure S40: Comparing cell type annotation in multiome data with scRNA-Seq in pcw 21 using CellWalker2, related to STAR Methods. The dotplot shows the assignment of cells of each cluster in multiome data by cell type hierarchy from scRNA-Seq data. The size and color of the circle represents the number and the proportion of cells, respectively, in each cluster in the multiome data assigned to each cell type label (B) or each node on the cell type hierarchy (C) in scRNA-Seq data. The cell type tree is shown in (A). We excluded VLMC and RBC which have less than 5 cells. Because of different resolutions and developmental stages of the two datasets, the clusters within glutamateric neurons do not have one-to-one correspondence with different groups of glutamateric neurons in the scRNA-Seq data, and some of the cells in the multiome data are mapped to internal nodes on the cell tree, even though the major cell types are matched. The IN1 cluster from the multiome data is mapped to CGE IN in the scRNA-Seq ontology, while IN2 is mapped to both MGE IN and the parent node of IN. The multiome RG cluster is mostly mapped to late RG, but also early RG. The EC/peric cluster is mapped to the parent node of Peric and EC. The mGPC/OPC cluster is mapped to the ancestor node of both mGPC and OPC/Oligo. nIPC/GluN1 is mapped to multiple glutamateric neuron types and nIPC.



(A) Brain ChIP-Seq experiments



(B) Blood ChIP-Seq experiments

Figure S41: ReMap ChIP-Seq data in the brain and the blood, related to STAR Methods. The number of peaks (log10) of each TF (columns) in each ChIP-Seq experiment (rows).

References

- S1. Markenscoff-Papadimitriou, E., Whalen, S., Przytycki, P., Thomas, R., Binyameen, F., Nowakowski, T.J., Kriegstein, A.R., Sanders, S.J., State, M.W., Pollard, K.S., and Rubenstein, J.L. (2020). A chromatin accessibility atlas of the developing human telencephalon. *Cell* *182*, 754–769.e18. doi: [10.1016/j.cell.2020.06.002](https://doi.org/10.1016/j.cell.2020.06.002).
- S2. Trevino, A.E., Müller, F., Andersen, J., Sundaram, L., Kathiria, A., Shcherbina, A., Farh, K., Chang, H.Y., Pasca, A.M., Kundaje, A., Pasca, S.P., and Greenleaf, W.J. (2021). Chromatin and gene-regulatory dynamics of the developing human cerebral cortex at single-cell resolution. *Cell* *184*, 5053–5069.e23. doi: [10.1016/j.cell.2021.07.039](https://doi.org/10.1016/j.cell.2021.07.039).
- S3. Nowakowski, T.J., Bhaduri, A., Pollen, A.A., Alvarado, B., Mostajo-Radji, M.A., Lullo, E.D., Haeussler, M., Sandoval-Espinosa, C., Liu, S.J., Velmeshev, D., Ounadjela, J.R., Shuga, J., Wang, X., Lim, D.A., West, J.A., Leyrat, A.A., Kent, W.J., and Kriegstein, A.R. (2017). Spatiotemporal gene expression trajectories reveal developmental hierarchies of the human cortex. *Science* *358*, 1318–1323. doi: [10.1126/science.aap8809](https://doi.org/10.1126/science.aap8809).
- S4. Yoshida, M., Worlock, K.B., Huang, N., Lindeboom, R.G.H., Butler, C.R., Kumashaka, N., Dominguez Conde, C., Mamanova, L., Bolt, L., Richardson, L., Polanski, K., Madison, E., Barnes, J.L., Allen-Hyttinen, J., Kilich, E., Jones, B.C., de Wilton, A., Wilbrey-Clark, A., Sungnak, W., Pett, J.P., Weller, J., Prigmore, E., Yung, H., Mehta, P., Saleh, A., Saigal, A., Chu, V., Cohen, J.M., Cane, C., Iordanidou, A., Shibuya, S., Reuschl, A.K., Herczeg, I.T., Argento, A.C., Wunderink, R.G., Smith, S.B., Poor, T.A., Gao, C.A., Dematte, J.E., NU SCRIPT Study Investigators, Reynolds, G., Haniffa, M., Bowyer, G.S., Coates, M., Clatworthy, M.R., Calero-Nieto, F.J., Göttgens, B., O'Callaghan, C., Sebire, N.J., Jolly, C., De Coppi, P., Smith, C.M., Misharin, A.V., Janes, S.M., Teichmann, S.A., Nikolić, M.Z., and Meyer, K.B. (2022). Local and systemic responses to SARS-CoV-2 infection in children and adults. *Nature* *602*, 321–327. doi: [10.1038/s41586-021-04345-x](https://doi.org/10.1038/s41586-021-04345-x).
- S5. Ren, X., Wen, W., Fan, X., Hou, W., Su, B., Cai, P., Li, J., Liu, Y., Tang, F., Zhang, F., Yang, Y., He, J., Ma, W., He, J., Wang, P., Cao, Q., Chen, F., Chen, Y., Cheng, X., Deng, G., Deng, X., Ding, W., Feng, Y., Gan, R., Guo, C., Guo, W., He, S., Jiang, C., Liang, J., Li, Y.M., Lin, J., Ling, Y., Liu, H., Liu, J., Liu, N., Liu, S.Q., Luo, M., Ma, Q., Song, Q., Sun, W., Wang, G., Wang, F., Wang, Y., Wen, X., Wu, Q., Xu, G., Xie, X., Xiong, X., Xing, X., Xu, H., Yin, C., Yu, D., Yu, K., Yuan, J., Zhang, B., Zhang, P., Zhang, T., Zhao, J., Zhao, P., Zhou, J., Zhou, W., Zhong, S., Zhong, X., Zhang, S., Zhu, L., Zhu, P., Zou, B., Zou, J., Zuo, Z., Bai, F., Huang, X., Zhou, P., Jiang, Q., Huang, Z., Bei, J.X., Wei, L., Bian, X.W., Liu, X., Cheng, T., Li, X., Zhao, P., Wang, F.S., Wang, H., Su, B., Zhang, Z., Qu, K., Wang, X., Chen, J., Jin, R., and Zhang, Z. (2021). COVID-19 immune features revealed by a large-scale single-cell transcriptome atlas. *Cell* *184*, 5838. doi: [10.1016/j.cell.2021.01.053](https://doi.org/10.1016/j.cell.2021.01.053).

- S6. Polioudakis, D., de la Torre-Ubieta, L., Langerman, J., Elkins, A.G., Shi, X., Stein, J.L., Vuong, C.K., Nicterwitz, S., Gevorgian, M., Opland, C.K., Lu, D., Connell, W., Ruzzo, E.K., Lowe, J.K., Hadzic, T., Hinz, F.I., Sabri, S., Lowry, W.E., Gerstein, M.B., Plath, K., and Geschwind, D.H. (2019). A single-cell transcriptomic atlas of human neocortical development during mid-gestation. *Neuron* *103*, 785–801.e8. URL: <https://www.sciencedirect.com/science/article/pii/S0896627319305616>. doi: <https://doi.org/10.1016/j.neuron.2019.06.011>.
- S7. Bakken, T.E., Jorstad, N.L., Hu, Q., Lake, B.B., Tian, W., Kalmbach, B.E., Crow, M., Hodge, R.D., Krienen, F.M., Sorensen, S.A., Eggermont, J., Yao, Z., Avermann, B.D., Aldridge, A.I., Bartlett, A., Bertagnolli, D., Casper, T., Castanon, R.G., Crichton, K., Daigle, T.L., Dalley, R., Dee, N., Dembrow, N., Diep, D., Ding, S.L., Dong, W., Fang, R., Fischer, S., Goldman, M., Goldy, J., Graybuck, L.T., Herb, B.R., Hou, X., Kancherla, J., Kroll, M., Lathia, K., van Lew, B., Li, Y.E., Liu, C.S., Liu, H., Lucero, J.D., Mahurkar, A., McMillen, D., Miller, J.A., Moussa, M., Nery, J.R., Nicovich, P.R., Niu, S.Y., Orvis, J., Osteen, J.K., Owen, S., Palmer, C.R., Pham, T., Plongthongkum, N., Poirion, O., Reed, N.M., Rimorin, C., Rivkin, A., Romanow, W.J., Sedeño-Cortés, A.E., Siletti, K., Somasundaram, S., Sulc, J., Tieu, M., Torkelson, A., Tung, H., Wang, X., Xie, F., Yanny, A.M., Zhang, R., Ament, S.A., Behrens, M.M., Bravo, H.C., Chun, J., Dobin, A., Gillis, J., Hertzano, R., Hof, P.R., Höllt, T., Horwitz, G.D., Keene, C.D., Kharchenko, P.V., Ko, A.L., Lelieveldt, B.P., Luo, C., Mukamel, E.A., Pinto-Duarte, A., Preissl, S., Regev, A., Ren, B., Scheuermann, R.H., Smith, K., Spain, W.J., White, O.R., Koch, C., Hawrylycz, M., Tasic, B., Macosko, E.Z., McCarroll, S.A. et al. (2021). Comparative cellular analysis of motor cortex in human, marmoset and mouse. *Nature* *598*, 111–119. URL: <https://doi.org/10.1038/s41586-021-03465-8>. doi: [10.1038/s41586-021-03465-8](https://doi.org/10.1038/s41586-021-03465-8).
- S8. Kushima, I., Aleksic, B., Nakatochi, M., Shimamura, T., Okada, T., Uno, Y., Morikawa, M., Ishizuka, K., Shiino, T., Kimura, H., Arioka, Y., Yoshimi, A., Takasaki, Y., Yu, Y., Nakamura, Y., Yamamoto, M., Iidaka, T., Iritani, S., Inada, T., Ogawa, N., Shishido, E., Torii, Y., Kawano, N., Omura, Y., Yoshikawa, T., Uchiyama, T., Yamamoto, T., Ikeda, M., Hashimoto, R., Yamamori, H., Yasuda, Y., Someya, T., Watanabe, Y., Egawa, J., Nunokawa, A., Itokawa, M., Arai, M., Miyashita, M., Kobori, A., Suzuki, M., Takahashi, T., Usami, M., Kodaira, M., Watanabe, K., Sasaki, T., Kuwabara, H., Tochigi, M., Nishimura, F., Yamasue, H., Eriguchi, Y., Benner, S., Kojima, M., Yassin, W., Munesue, T., Yokoyama, S., Kimura, R., Funabiki, Y., Kosaka, H., Ishitobi, M., Ohmori, T., Numata, S., Yoshikawa, T., Toyota, T., Yamakawa, K., Suzuki, T., Inoue, Y., Nakaoka, K., Goto, Y.I., Inagaki, M., Hashimoto, N., Kusumi, I., Son, S., Murai, T., Ikegame, T., Okada, N., Kasai, K., Kunimoto, S., Mori, D., Iwata, N., and Ozaki, N. (2018). Comparative analyses of copy-number variation in autism spectrum disorder and schizophrenia reveal etiological overlap and biological insights. *Cell Rep.* *24*, 2838–2856. doi: [10.1016/j.celrep.2018.08.022](https://doi.org/10.1016/j.celrep.2018.08.022).