Integration and transfer learning of single-cell transcriptomes via cFIT

Minshi Peng[a], Yue Li[a], Brie Wamsley[b], Yuting Wei[a], and Kathryn Roeder[a,c,1]

[a]Department of Statistics and Data Science, Carnegie Mellon University, Pittsburgh, PA 15213; [b]Neurogenetics Program, University of California, Los Angeles, CA 90095; and [c]Computational Biology Department, Carnegie Mellon University, Pittsburgh, PA 15213

Contributed by Kathryn Roeder, December 13, 2020 (sent for review November 25, 2020; reviewed by Eric Courchesne and Yun Li)

Large, comprehensive collections of single-cell RNA sequencing (scRNA-seq) datasets have been generated that allow for the full transcriptional characterization of cell types across a wide variety of biological and clinical conditions. As new methods arise to measure distinct cellular modalities, a key analytical challenge is to integrate these datasets or transfer knowledge from one to the other to better understand cellular identity and functions. Here, we present a simple yet surprisingly effective method named common factor integration and transfer learning (cFIT) for capturing various batch effects across experiments, technologies, subjects, and even species. The proposed method models the shared information between various datasets by a common factor space while allowing for unique distortions and shifts in genewise expression in each batch. The model parameters are learned under an iterative nonnegative matrix factorization (NMF) framework and then used for synchronized integration from across-domain assays. In addition, the model enables transferring via low-rank matrix from more informative data to allow for precise identification in data of lower quality. Compared with existing approaches, our method imposes weaker assumptions on the cell composition of each individual dataset; however, it is shown to be more reliable in preserving biological variations. We apply cFIT to multiple scRNA-seq datasets of developing brain from human and mouse, varying by technologies and developmental stages. The successful integration and transfer uncover the transcriptional resemblance across systems. The study helps establish a comprehensive landscape of brain cell-type diversity and provides insights into brain development.

single-cell RNA-seq | data integration | transfer learning | brain cells

Individual single-cell RNA sequencing (scRNA-seq) experiments have been used to discover new cell states and reconstruct cellular differentiation trajectories. Recent studies have shown that cellular features can be preserved across experimental systems from related biological contexts (1). The information learned from different data sources can improve the analysis and interpretation of diverse biological systems. However, the advantages of integrated data can be compromised by differences due to experimental batch, sampling (sample acquisition and handling, sample composition, reagents or media, and sampling time), or technology (sequencing depth, sequencing lanes, read length, plates or flow cells, protocol) (2). The challenge is exacerbated when technical differences in data sources are confounded with biological heterogeneity. Many methods have been established to integrate scRNA-seq studies across multiple experiments. Some methods employ supervised cross-domain transfer learning (3–6) to remove domain effects with models learned from labeled datasets. These methods rely heavily on data labeling, thus failing to capture novel cell types and continuous trajectories. In contrast, unsupervised methods are less restrictive and therefore, more widely applied to integrate data from multiple resources (7–11). However, many of these methods tend to prioritize uniformity of mixing across different batches over preserving biological variation (2). Such a principle can lead to a loss in biological heterogeneity and interpretability, especially when integrating collections of datasets with considerable differences in cellular composition. Also, most existing methods work on the assumption that all datasets share most cell types or that the within-domain biological variance defining distinct cell types dominates the cross-domain effects (10, 12); such assumptions do not hold when integrating biologically heterogeneous datasets or data consisting of continuously transitioning cell types or refined subtypes.

Here, we present an effective unsupervised integration and transfer learning model, called cFIT (common factor integration and transfer learning). The model assumes a shared common factor space across datasets but with location-scale shifts on genewise expression unique by domain. Our model is motivated from the machine-learning subdomain of transfer learning (13–17), assuming that information is shared across different tasks, and common data representations can be learned and generalized to other unseen tasks. In this framework, the shared latent space represents the underlying biological processes across systems, such as common cell-type compositions and developmental trajectories across measurements, samples, or even species. After the robustness of a biological process is established, these learned latent spaces enable varied learning tasks across data platforms, modalities, and studies, through transfer learning.

The proposed model is capable of capturing various batch effects and integrating across various domains by employing a linear model that is more parsimonious than existing methods

Significance

Overcorrection has been one of the main concerns in employing various data integration methods, which risk removing the biological distinction and are harmful for cell-type identification. Here, we present a simple yet surprisingly effective model named common factor integration and transfer learning for capturing various batch effects across experiments, technologies, subjects, and even species. The method generates robust results when batch effects are confounded with the variability of cell-type compositions and when the population exhibits continuous developing patterns. The successful integration and transfer uncover the transcriptional resemblance described by the proposed location-scale shift model across systems. In addition, the model enables transferring via low-rank matrix from more informative data to allow for precise identification in data of lower quality.

Author contributions: M.P., B.W., Y.W., and K.R. designed research; M.P. and Yue Li performed research; M.P., Yue Li, Y.W., and K.R. contributed new reagents/analytic tools; M.P. and Yue Li analyzed data; and M.P., Yue Li, Y.W., and K.R. wrote the paper.

Reviewers: E.C., University of California San Diego; and Yun Li, University of North Carolina at Chapel Hill.

The authors declare no competing interest.

[1]To whom correspondence may be addressed. Email: roeder@andrew.cmu.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2024383118/-/DCSupplemental.

Published March 3, 2021.

such as Linked Inference of Genomic Experimental Relationships (LIGER) (11). cFIT is powerful as it corrects for technical variation but does not remove biological heterogeneity, providing both flexibility and interpretability. For implementation, we derived an algorithm for inferring model parameters under an iterative nonnegative matrix factorization (NMF) framework. The algorithm is also applicable for synchronized integration of cross-domain assays. Finally, the learned biological signatures can apply transfer learning to allow for precise inference for data with lower quality or smaller sample size.

cFIT enables successfully integration of two independent datasets derived from the developing human cortex (18, 19). The integration disentangled the domain-specific technical effects with the biological processes unique to each dataset, where the latter is preserved and depicted in the recovered developing trajectories. The learned latent biological signatures were then transferred to several previously published datasets from fetal brain (20–23)—allowing for finer characterizations of cell identities and the biological process they are involved in, which would not be feasible otherwise. In addition, the resilience to overcorrection facilitates the detection of possible contaminations in the data source. By integration of data across species, we identified transcriptomic heterogeneity between mouse and human cells during the embryonic stage of interneuron (IN) development. The findings shed light on similarities and differences in IN fate between species. In aggregate, these analyses highlight the utility of the proposed method to borrow strength across multiple datasets and to transfer information between related datasets.

Results

Methods Overview. cFIT models the scRNA-seq expression of individual cells using its cellular identity and domain-specific factors. Here, domain refers to any standalone dataset profiled at a single laboratory using a single technology from one batch (Fig. 1). Specifically, we are given a total number of N cells, and each cell i is associated with a p-dimensional feature vector $\boldsymbol{x}_i$ corresponding to its gene expression values (*SI Appendix* has data preprocessing steps). Each cell comes from a specific domain with a known domain identification. Given M different domains, we use $m_i \in \{1, 2, \ldots, M\}$ to denote the domain identification of cell i. We model the scRNA-seq data via a high-dimensional linear model with a latent low-dimensional structure. Let $\boldsymbol{x}_i$ be the observation of cell i that is generated from

$$\boldsymbol{x}_i = \boldsymbol{\Lambda}_{m_i} \boldsymbol{W} \boldsymbol{h}_i + \boldsymbol{b}_{m_i} + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, N. \qquad [1]$$

Here, we use the diagonal $(p \times p)$ matrix $\boldsymbol{\Lambda}_{m_i} = diag(\boldsymbol{\lambda}_{m_i})$ to control the discrepancy of individual gene expression resulting from domain-specific technical effects. The matrix $\boldsymbol{W}$ denotes a $(p \times r)$ nonnegative factor matrix shared across all samples and domains representing the gene expression profiles (signatures) associated with the cells; each vector $\boldsymbol{h}_i$ is a cell-specific nonnegative vector of length r representing the factor loading vector of cell i. The domain-specific vector $\boldsymbol{b}_{m_i}$ is a nonnegative vector of length p that captures the domain-associated shift. The noise terms $\{\boldsymbol{\epsilon}_i\}_{i=1}^N$ are modeled as independent, normally distributed random vectors with mean $\boldsymbol{0}$ and variance $\sigma^2 \mathbf{I}_p$ to account for measurement error from various sources.

Let n_j denote the number of cells from batch j and $N = \sum_{j=1}^M n_j$ the total number of cells. Concatenating the scRNA-seq expressions of all cells from each domain j as an $(n_j \times p)$ matrix $\boldsymbol{X}_j$, then model Eq. **1** in matrix form is

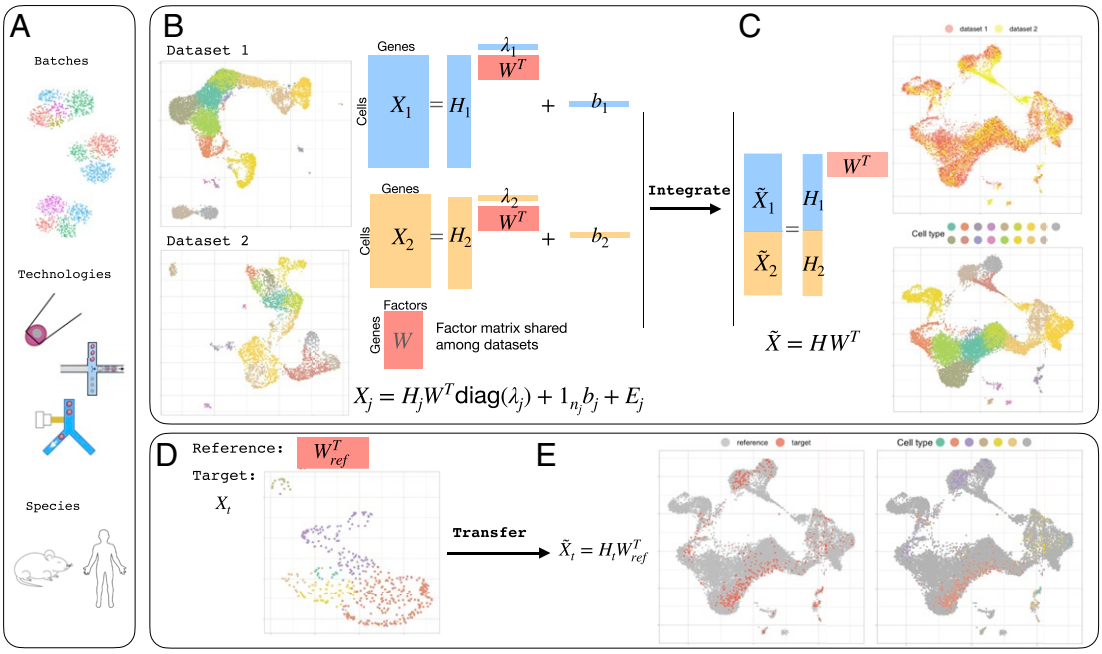

Fig. 1. cFIT integration and transfer approach overview. (*A*) cFIT performs integration or transfer among scRNA-seq datasets from different batches, technologies, and across species. (*B*) Data integration takes in two or more datasets from different domains, where some cell-level biological processes are shared. Each dataset is modeled by a low-dimensional latent space corresponding to gene-level features (gene expression signatures), W, shared across domains, domain-specific factor loading H_j characterizing cell composition, and domain-unique scaling, λ_j, and shift, b_j, capturing the technical distinction. (*C*) The integration algorithm estimates the set of parameters through iterative NMF. The integrated data can be obtained by eliminating the technical distinctions and projected onto a common subspace, where downstream analysis can be performed, such as clustering and trajectory inference. (*D*) The transfer process takes a reference factor matrix representing the gene-level signature profiles and a target dataset sharing the signature space. (*E*) The transfer algorithm estimates the target-specific parameters to project the target data onto the same low-dimensional space as inferred from reference data. Cell labels can be assigned directly with unsupervised learning in low-dimensional space or querying reference data.

$$\boldsymbol{X}_j = \boldsymbol{H}_j \boldsymbol{W}^\top \boldsymbol{\Lambda}_j + \boldsymbol{1}_{n_j} \boldsymbol{b}_j^\top + \boldsymbol{E}_j, \quad j = 1, \ldots, M. \qquad [2]$$

Here, $\boldsymbol{H}_j$ is a nonnegative factor loading matrix, and $\boldsymbol{E}_j$ is the noise matrix (Fig. 1*B*).

Note that $\boldsymbol{H}_j$ captures the biological heterogeneity originating from disparate cell-type compositions, and $\{\boldsymbol{\Lambda}_j, \boldsymbol{b}_j\}$ are domain-specific parameters that accommodate domain differences such as batch effects from samples and libraries, different sequencing technologies, and even species, whereas $\boldsymbol{W}$ is the common factor space—the information shall be extracted and shared among datasets. The above model (Eq. **2**) hinges upon the rationale that all of the samples belong to, after proper shift and rescaling, the same lower-dimensional linear subspace, which makes it possible to leverage information from diverse datasets.

The set of unknown parameters is estimated by minimizing the objective function provided in *Methods*. Then, one can use the integrated, low-dimensional representation of each dataset $\boldsymbol{H}_i \hat{\boldsymbol{W}}^\top$ for downstream analysis, as well as leverage the common factor matrix $\hat{\boldsymbol{W}}$ for efficient transfer to a target dataset as detailed in *Methods* and *SI Appendix*.

Comparisons with other methods. LIGER (11) is a popular linear method that employs a similar matrix factorization: it factorizes each batch expression matrix into a shared factor matrix W and a batch-unique factor matrix with $p \times r$ parameters to describe the domain effects. By contrast, we consider a different modeling approach using only $2p$ parameters based on patterns of technical variations. By imposing a structural constraint on the dataset-specific effect in cFIT, we restrict the domain-specific effects to the deviation by location and scale but rely on the shared factor matrix to model the relative signature matrix. The remaining orthogonal effects are preserved as biological distinctions; thus, the corrected batch effects are less likely to confound with biological effects. Although our model is comparatively conservative, we shall show in our data analyses that the structured model is often sufficient to capture and remove the domain effects from various sources. Nonlinear methods, on the other hand, typically allow higher flexibility and have shown superior power and performance in blending multiple sources of data sharing similar compositions. These methods typically involve the identification of mappings between datasets and remove the differences accordingly. It can be achieved by finding the mutual nearest neighbors (MNNs), as used in MNNcorrect (12) and Seurat v3 (10), where the target data are mapped to the query data, guided by the pairwise points identified by MNN. Other types of nonlinear method leverage the alignment between clusters (distinct cell types) (8) to eliminate the batch effects within the clusters and/or between mapped clusters. The success of MNN relies heavily on the assumption that the batch effect is almost orthogonal to the biological subspace, and there is a substantial overlap of the cell compositions between the source and target data. As such, cluster-based methods require relatively well-separated clusters mapped across datasets. Apart from that, both types of approaches assume that the biological differences dominate the domain effects; however, this assumption is likely violated when cells are obtained along a continuous developmental trajectory with transitioning subtypes. These nonlinear methods eliminate the pairwise distinctions regardless of the source of distinctions. Therefore, they tend to overcorrect the biological variance. Alternatively, our method imposes less assumption on the distinctions of biological and domain effects by parsimoniously modeling the most likely generative source of domain effects. When our model assumptions are violated, it is likely that cFIT does not remove the batch effects that are unexplained by the model rather than falsely erasing signals. In this sense, cFIT is less prone to overcorrection.

Simulation Studies: Differential Expression Analysis and Robustness. The performance of cFIT was first evaluated on differential expression gene (DEG) discovery, a key downstream analysis. The comparison assessed the integration methods' ability to remove domain-specific factors while preserving biological signals. We compared with two widely used single-cell integration methods Seurat v3 (10) and MNNcorrect (12), which performed well in a benchmarking study (24). We do not compare with LIGER because this method does not produce a reconstructed scRNA expression matrix. We performed simulations with single-cell simulator under five settings with a combination of balanced/unbalanced batches, regular/high-dropout rates, and two/multiple biological groups. After obtaining the reconstructed expression matrices, we used the two-sided Wilcox rank-sum test for DEG detection and reported the false discovery rate (FDR) among top 50 and 100 discoveries. All three methods are effective in capturing the biological signals in some settings. Seurat v3 has advantages in the unbalanced batch/multiple group settings, while MNNcorrect shows better performances in the balanced batch scenarios. Overall, cFIT has the smallest FDR for most settings and is most robust across different parameter settings (*SI Appendix*, Fig. S1*A*).

We then evaluated the robustness of cFIT with respect to tuning parameters. In *SI Appendix*, we show that, with our proposed penalty term and parameter constraints, cFIT is identifiable and guaranteed to converge; furthermore, performance is enhanced using our initialization process. Hence, the only key tuning parameter remaining is r, the number of latent factors. Here, we demonstrate that our results are robust when r is chosen from a reasonably wide range. By construction, the latent dimension r corresponds to the number of biological groups. We considered two simulation designs such that the expression matrices follow model Eq. **2** with $r = 5$, and 1) all five cell groups presented in both domains, 2) three cell groups presented in both domains, and each domain has a unique cell group. A Uniform Manifold Approximation and Projection (UMAP) visualization revealed the impact of severe batch effects (*SI Appendix*, Fig. S1 *B* and *C*). Provided $r \geq 5$, cFIT performed well, successfully integrating cells in each biological group into one cluster. Note that the second setting was challenging for most current integration methods, but cFIT was still able to recover the "unique cell groups." In practice, when we do not have prior knowledge about the number of biological groups, taking into consideration the complexity of real datasets and computation cost, we recommend a choice of r between 10 and 50.

Applications to Single-Cell Data.

Integration of scRNA-seq datasets from multiple technologies. Next, we evaluated our approach on human pancreatic islet cell datasets produced across five technologies, CelSeq, CelSeq2, Fluidigm C1, SMART-Seq2, and InDrop. After applying the cFIT integration procedure, technical distinctions that originally grouped cells by batch were effectively removed, so that cells belonging to the same cell type, regardless of data sources, were well mixed (*SI Appendix*, Fig. S2). In addition to detecting all major cell classes (alpha, beta, delta, gamma, acinar, and stellar), we also identified some rare cell types (schawann and mast) with the integrated data that could not be reliably detected with a smaller number of cells through individual clustering analysis. We benchmarked the performance of cFIT against two popular methods: LIGER (11) and Seurat v3 (10). The integration results were compared using the alignment score, a measurement of how well different datasets mix (*Methods* and *SI Appendix*), and the accuracy for preserving the cell-type structure measured by the Adjusted Rand Index (ARI) on clustering. Methods that perform well in both metrics effectively matched populations across datasets without blending distinct populations. Results show that cFIT achieved comparable high alignment scores and

clustering accuracy as Seurat, indicating the effectiveness of our method relying on a simpler linear model compared with the nonlinear multistep procedure employed by Seurat. LIGER produced lower clustering accuracy, due to overcorrection of the domain-specific effects that falsely removed the biological gene-level distinctions between cells from different cell types. LIGER promoted the use of a postquantile normalization step to further align the quantiles of the batches within obtained clusters. Similarly, we show that this step can also be coupled with cFIT and produces higher alignment scores, despite the increased risk of mixing distinct cell types.

Integration of datasets consisting of distinct cell types. To examine the robustness of the proposed method on integrating populations with different cell-type compositions, we removed all cells of one prechosen type from each pancreas islet cells dataset. The cell type was randomly chosen as the largest or second-largest major cell type for each dataset (*SI Appendix*). The integration procedure was applied to these perturbed datasets to compare performance. cFIT successfully integrated the datasets without blending distinct cell types. As with the original data, cFIT achieved high clustering accuracy (ARI ≈ 0.9) and alignment score (Fig. 2 and *SI Appendix*, Fig. S3). In contrast, LIGER and Seurat integration failed to preserve the cell-type structure, and cells from different cell types were mixed together. Specifically, in both LIGER and Seurat results, a fraction of ductal cells was clustered with acinar cells, and beta, delta, alpha, and gamma cells became entangled. The clustering accuracy (ARI) dropped to approximately 0.6 (*SI Appendix*, Fig. S3*C*). In summary, cFIT successfully characterized the domain-specific effects and was robust to the perturbation in relative cell-type compositions.

To further ensure that cFIT was able to distinguish between the technical and biological effects when they were confounded, we jointly analyzed profiles of hippocampal oligodendrocytes and interneurons (INs) (25). The two cell classes share a common origin in mouse development, but they are born in distinct time frames and have very different functions in the mature brain. Therefore, we expected the two sets of mature cells to share few, if any, common cell populations. Compared with LIGER and Seurat, cFIT generated the minimum false alignment, which is apparent visually in the UMAP display of integrated data and by comparing the alignment scores (*SI Appendix*, Fig. S4).

Integration of human fetal cortical scRNA-seq data. Several studies, all following a unified set of protocols, have dissected multiple regions of human fetal brain and sequenced single cells (18–23). These studies enable the study of cellular programs in early development, comparisons of regional differences in cell-type compositions and expression profiles, and mining associations between brain cell types and neurological disorders (19, 26). However, challenges remain to integrate these datasets, which have been produced using disparate technologies, providing different coverage of developmental stage and brain regions. As a first step, we integrated the two largest scRNA-seq datasets from fetal cortex: 1) the Polioudakis scRNA-seq data (19) (Drop-seq) of 33,976 cells from cortical anlage at gestation week 17 or 18 and 2) the Nowakowski data (18) (Fluidigm C1) of 4,261 cells from multiple brain areas and ranging in age from 5.85 to 37 post-conception weeks (PCW). We aimed to borrow the advantages from both datasets via their integration, thus providing a more comprehensive characterization of human neocortical development. By applying our proposed algorithm on these datasets, we obtained integrated data with good overlap from initially divergent data sources, as indicated via UMAP visualization and alignment score (Fig. 3 *A*–*C*). Similar cell types matched perfectly, including oligodendrocyte progenitor cells (OPCs), radial glia, progenitor cells, and excitatory neurons (ENs), while others were unique to Nowakowski data, including IN progenitors (medial ganglionic eminences radial glia [MGE-RG], medial ganglionic eminences intermediate progenitor cell [MGE-IPC], newborn IN [nIN]) and unlabeled types (U1, U2). The IN progenitors were sampled from a brain region not included in the Polioudakis dataset; the unknown types could be other unrepresented cell types or artifacts. It was notable that cFIT did not force these cells to overlap with closely related cell types present in the larger Polioudakis dataset. Moreover, two maturation trajectories of ENs were revealed in the Nowakowski data. One ends at prefrontal cortex (PFC) maturing excitatory cells, and the other ends at primary visual cortex (V1) maturing excitatory cells; while the mature cells sampled from the two different brain regions were clearly differentiated, the immature cells were not. The Polioudakis data spanned this same space, but this study did not distinguish cells by region. Thus, by aligning the two datasets and noting the two trajectories apparent in the Nowakowski data, we can infer the regional origin of a portion of the Polioudakis cells (Fig. 3*D*). In addition, we were able to assign labels to previously unannotated cells and likely misidentified cells. For instance, a group of unlabeled cells (NA) in the Nowakowski data was identified via nearest neighbor matching as ENs due to their substantial overlap with migrating and maturing ENs in the Polioudakis data. A subset of 39 OPCs in Polioudakis data are likely astrocytes (ASTs) given they are matched to ASTs from the Nowakowski data and expressed ASTs markers *GFAP*, *SOX9*, and *EGFR*. Compared with the initial attempt to integrate the same datasets (19) using Seurat, cFIT results exhibited superior alignments, especially among ENs, while preserving the batch unique profiles in medial ganglionic eminences (MGE) lineage that were eliminated by Seurat integration.

Transfer the learned signature from a dataset with a larger number of human cortical cells to a dataset with a smaller number of human cortical cells. The learned factor matrix from the integration of the Nowakowski and Polioudakis data can serve as a comprehensive reference that characterizes the cellular processes in the fetal brain, covering a wide range of developmental stages (6 to 37 PCW) and major and more specialized cell types. Thus, we applied our proposed transfer learning methods and used this comprehensive reference to enable cell-type identification for 2,309 PFC cells from the Zhong dataset (23), which

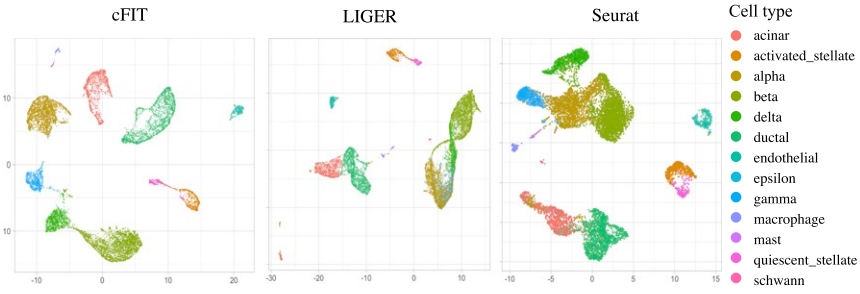

Fig. 2. UMAP plots of integrated data from perturbed pancreas islet datasets, created by moving one major cell type from each dataset, comparing results from three methods: cFIT (*Left*), LIGER (*Center*), and Seurat (*Right*). Cells are colored by cell types.

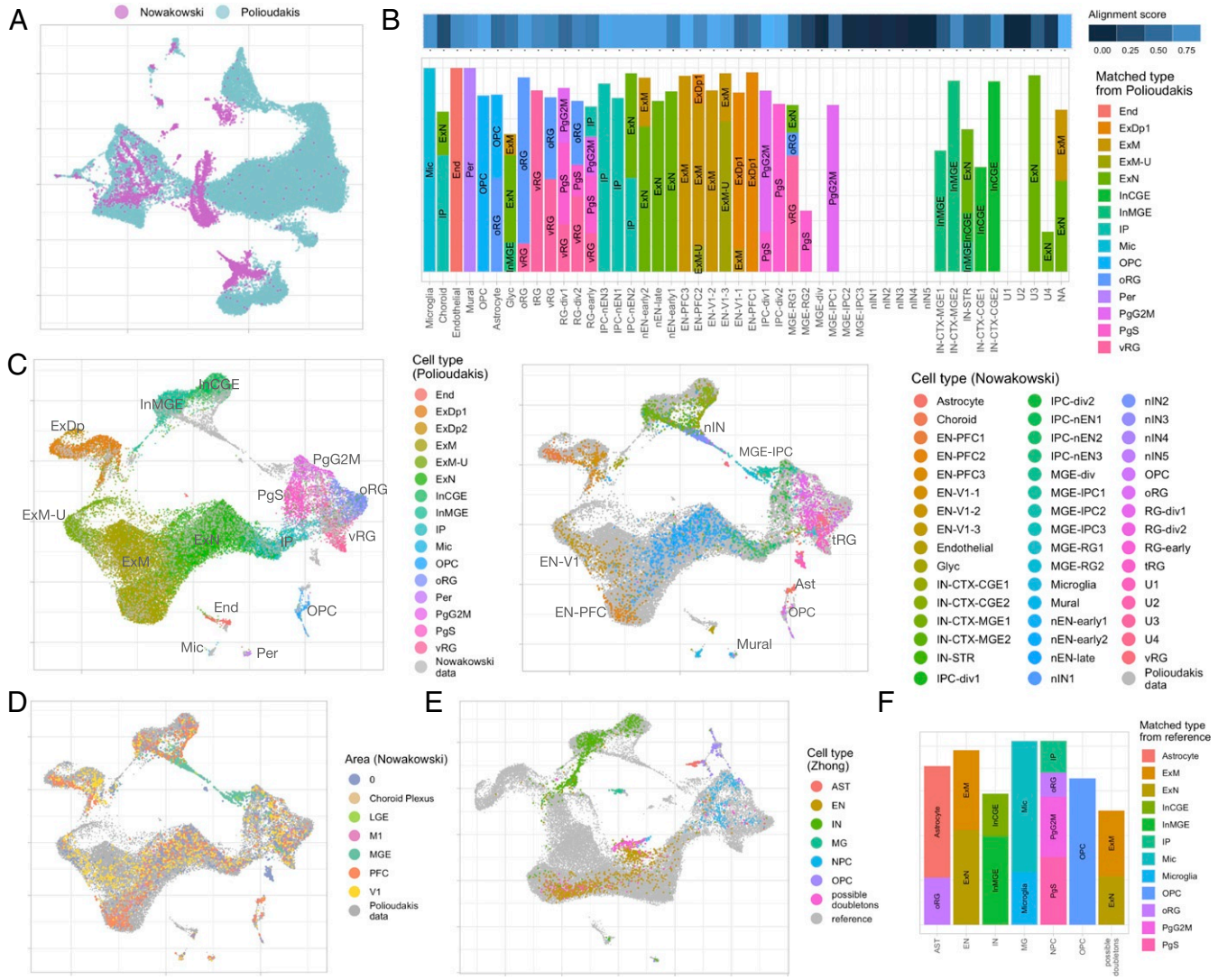

Fig. 3. (*A–D*) Integration of two scRNA-seq datasets from the fetal brain (Nowakowski and Polioudakis data). (*A*) UMAP plot of two datasets before integration. (*B*) Match of the identified clusters in Nowakowski data with major types identified from the Polioudakis data. The alignment score demonstrates how well each cluster in the Nowakowski data are aligned with cells from the Polioudakis data visualized in the color bar above. A higher score means the cells are well matched with cells from the other dataset. The bar plot shows the top matched cluster from the Polioudakis data for each cluster of the Nowakowski data. (*C*) UMAP of scaled factor loadings obtained from data integration. In *Left*, only cells from the Polioudakis data are colored into 16 major cell types. Similarly, in *Right*, only cells from Nowakowski data are colored according to the 48-cell-type label (see *SI Appendix*, Tabel S2 for detailed cell-type annotations from respective studies). (*D*) UMAP of integrated Polioudakis and Nowakowski data, colored by brain area annotated for the Nowakowski data, including prefrontal cortex (PFC), primary visual cortex (V1), medial ganglionic eminence (MGE), and lateral ganglionic eminence (LGE). (*E* and *F*) Transfer results on 2,309 cells from the Zhong data. (*E*) UMAP of 2,309 cells (colored by Zhong labels, *SI Appendix*, Table S2) overlaying on cells from the Polioudakis and Nowakowski (reference) datasets, among them a group of cells outside the range of reference cells and previously identified as possible doubletons (27). (*F*) The composition of cells based on matched cell types in reference datasets for each major group, with the alignment score computed for each major group measuring how well it matches with reference data.

contains 2,309 single cells from the human embryonic PFC from 8 to 26 PCW. Using the Seurat package, the authors identified six major clusters: neural progenitor cells (NPCs), ENs, INs, ASTs, OPCs, and microglia (MIC); these will be referred to as Zhong labels. Through the transfer procedure, we obtained the following results. First, the 2,309 cells, represented in low-dimensional space using estimated factor loadings, overlay substantially with cells from the Polioudakis and Nowakowski datasets. We were able to identify finer structure within each major group by matching group labels from the reference datasets. Among them, NPC consisted of cycling progenitors, intermediate progenitors, and radial glia. EN cells contain both migrating and maturing ENs. IN matched perfectly with MGE and caudal ganglionic eminence (CGE)-derived INs, OPC and MIC aligned perfectly with OPC and MIC from reference data, and AST aligned partially with ASTs and partially with outer radial glia (oRG). Most groups showed high alignment scores, except for the AST clusters, which while laying adjacent to each other, were not well mixed. This analysis showed that by transferring rich information, we were capable of delineating finer structures within each major type in another dataset. Second, we were able to identify cells from rare types that were likely misclustered in the Zhong analysis. For instance, previously labeled as INs, five cells were relabeled as endothelial, and another five cells were relabeled as pericytes. This finding was supported by the UMAP visualization, where these cells lie in clusters corresponding to endothelial and pericytes that were well separated from other types. Third, the migrating and maturing ENs align well with the

PFC branch as identified from the Nowakowski data (the branch extends downward in the UMAP visualization), while the upper branch displays maturing ENs from the visual cortex. Fourth, we observed a fraction of EN cells that align poorly with the reference data (outside the gray range in the UMAP plot); among them, some were labeled as NPC. This was indicative of the underrepresentation of these cells in the reference. It was not clear whether these were unknown neuronal signatures characteristic of these cells or were mischaracterized cells. The latter is more likely because previous work (27) discovered a cluster of cells as possible doubletons (i.e., transcripts captured from two cells rather than one).

Transfer learning is particularly valuable when applied to much smaller datasets. We conducted similar transfer analysis on a variety of scRNA-seq datasets of human cortical cells (20–22), covering different technologies and populations. The Li data (22) contain 762 cells collected from nine brains ranging in age from 5 to 20 PCW. Applying cFIT, the cells were transferred to the reference data in the low-dimensional factor space and showed good alignment of most cell labels (*SI Appendix*, Fig. S5 *A* and *C*). The alignment was further validated by the incremental age along the trajectory observed (*SI Appendix*, Fig. S5*B*). By transferring onto well-characterized developmental trajectories, we were also able to identify some likely mislabeled cells, which likely arose due to having too few cells per type available originally: for instance, the rainbow points along the maturation trajectory of ENs and CGE/MGE INs. We next examined an even smaller scRNA-seq dataset from the fetal brain, with 220 cells sampled between 12 and 13 PCW (20). Guided by marker genes, these single cells were previously labeled into seven types in the original paper: two subtypes of apical progenitors (AP1, AP2), two subtypes of basal progenitors (BP1, BP2), and three subtypes of neurons (N1, N2, N3). By transferring these cells onto the reference data, we were able to validate the cell labels (*SI Appendix*, Fig. S5*D*). As a further step, we identified the finer structure within each cell group (different types of radial glia within apical progenitor cells and different progenitor cells within the basal progenitor group) (*SI Appendix*, Fig. S5*E*). We also revisited a widely studied dataset composed of both fetal and adult brain cells (21) (134 fetal cells and 334 adult cells from cortical tissues at 16 to 18 PCW). The transfer procedure evenly distributed the fetal cells along the reference trajectory, where labels were easily inferred (*SI Appendix*, Fig. S5*F*). By contrast, the transfer process projected all adult neurons outside of the fetal developing trajectory (*SI Appendix*, Fig. S5*G*).

Integrate cell expressions across species. We further examined the performance of the proposed algorithm for integrating mouse and human brain cells. Although we expect differences between the two species, similarities have also been noted in some transcriptomic patterns (28), and once understood, shared features are likely to provide a deeper insight into the fundamental architecture underlying cellular development and physiology. Diverse subsets of cortical INs have vital roles in higher-order brain functions. Due to the limited number of profiled early inhibitory precursors from human, we leveraged several mouse scRNA-seq datasets collected along a developmental time course using multiple technologies (29, 30) and integrated them with human MGE progenitors and INs (18). We first examined the heterogeneity within the early development in MGE (mainly composed of mitotic progenitors within MGE before migrating to cortex). By integrating the 733 Nowakowski cells collected from MGE (age between 5 and 21 PCW) with 5,622 mouse cells from Drop-seq data (embryonic day 13.5), we identified a common mitotic developmental trajectory shared between the two species (*SI Appendix*, Fig. S6 *A* and *B*). In their prior work (29, 30), each mouse cell was assigned a maturation score (a continuous value quantifying the extent of cell development), allowing the cells to be divided into mitotic and postmitotic stages. We mapped this maturation score to human cells by averaging over the scores of neighboring mouse cells (among the 30 nearest neighbors). We observed a match between the maturation score and the reference labels of human cells, starting from the two subtypes of MGE-RG, followed by the dividing MGE progenitors, subtypes of MGE-IPC, and ending with newborn INs (*SI Appendix*, Fig. S6*C*). Seurat clustering analysis was performed, which identified eight major clusters, each composed of both mouse and human cells (*SI Appendix*, Fig. S6*D*). Clusters A to F are composed of mitotic cells concordant with the reference labels (*SI Appendix*, Fig. S6*E*). Meanwhile, the two clusters of postmitotic cells aligned with the reference branch labels. The newborn INs fell in branch 1 cluster, which was conjectured to give rise of cortical INs (29). The alignment score calculated per cluster demonstrates that the human cells were evenly distributed in each cluster, except for F (*SI Appendix*, Fig. S6*E*). Cluster F contained a fraction of newborn INs from a later developmental stage (>20 PCW), beyond the range covered by mouse cells.

Next, we investigated the full developmental process starting from mitotic progenitors in MGE, which differentiate and migrate to the cortex to become mouse mature INs, and examined whether the human MGE-derived INs developmental trajectory could be aligned. We integrated six datasets, a human dataset containing 733 cells from MGE and 271 MGE-derived cortical INs (18) (5 to 22 PCW) and five from mouse at different ages, and sequenced by different technologies (29, 30) (*SI Appendix*). We observed a continuous developmental trajectory from the integrated data (*SI Appendix*, Fig. S6*F*) starting from the mitotic cells, followed by those transitioning into the postmitotic stage where progenitors diverge and differentiated by distinct transcriptional states. The states aligned with the previously identified three branches (*SI Appendix*, Fig. S6*H*). Along branch 1, new subbranches emerged and ultimately arrived at distinct types of mature INs (*Sst*, *Pvalb*, *Nos1*, and *Th*) (*SI Appendix*, Fig. S6*I*). The human cell development largely resembled the mouse cells, where the human cells ranged from early mitotic to relatively mature INs along branch 1. In another trajectory endpoint, striatal INs (IN-STR) were blended with mouse cells from branch 2 (*SI Appendix*, Fig. S6*J*). This analysis identified 12 major groups, including three clusters of mitotic cells (1, 4, 11) and branch 1 clusters of four different subbranches: an *Sst* branch (cluster 3, 9); a *Pvalb* (cluster 2, 10); an *Nos1* branch (cluster 7); and an unknown branch (cluster 8; expressing markers such as *BmT7*, *Col2a1*, *Notch2*, *Cd9*) (*SI Appendix*, Fig. S6 *K* and *L*). The human cells were grouped with mouse cells in corresponding clusters (*SI Appendix*, Fig. S7*B*). Particularly, a group of human cells in cluster 3 was identified as relatively mature precursors of *Sst* INs (*SI Appendix*, Fig. S7*B*). Our analyses revealed that transcriptional profiles underlying INal fate specification are largely conserved between mouse and human. It also highlighted the power of cFIT to capture and integrate the shared biological processes across different systems effectively. These are important steps toward the goal of harnessing information across species to understand mammalian neurodevelopment and its relevant physiology.

Discussion

We develop a method for integrating single-cell datasets and transferring knowledge to target settings that is robust and interpretable. cFIT assumes a shared common factor space across datasets, but it models distortions and shifts on genewise expression that are unique to each source. In doing so, it captures the advantages without the disadvantages of existing methods. Similar to LIGER, the nonnegativity constraint of NMF yields interpretable factors that can be biologically meaningful. While the shift in cFIT corresponds to the batch-specific factor in

LIGER, their scaling is applied to all of the shared factors concurrently. Therefore, our model takes a different approach to characterize the impact of domain effects originated from different experimental tools and measurements employed. Nonlinear methods typically allow for higher flexibility and superior power to correct the varying sources of batch effects, while it comes with a much higher risk of false correction of biological effects. This can cause reduced power of downstream analysis such as identifying DEG. Seurat v3 intrinsically assumes batch effects being orthogonal to biological effects and a substantial overlap of cell-type compositions between target and source datasets and produces results dependent on the order of pairwise integration. By contrast, our method makes no assumptions on cell populations in individual data and can perform integration simultaneously across all data sources. Our method also does not depend on the degree of domain effects relative to biological differences (e.g., between cell types), as required to ensure the success of methods such as MNNcorrect (12). Notably, far fewer parameters are required by our model, and yet, it retains the power to capture domain effects. It maintains identifiability and robustness through the choice of tuning parameters. Like Seurat v3, our method also provides an estimate of a corrected expression matrix, which can be used as input for downstream analyses such as pseudotime or differential gene expression analysis. There are notable advantages to having access to corrected versions of both gene expression and lower-dimensional factor loadings, which can be used to reveal interesting biological features. We would also like to point out the possibility of using cFIT as a first step to eliminate major batch effects and then applying alternative methods to remove more subtle effects in a subsequent iteration.

Unlike many competing approaches, cFIT is less prone to removing biological heterogeneity, which facilitates combining datasets with strong biological heterogeneity and capturing the advantages of each source into a single dataset. We show this in simulations and several scRNA-seq datasets. Consequentially, in our analysis of fetal brain development, we were able to combine datasets sampled from widely divergent protocols, spanning different developmental epochs with both mouse and human cells. The resulting integrated analysis delineated closely related neuronal subtypes, drew inferences about developmental trajectories, and correctly classified rare cells from datasets with small numbers of cells sampled. These insights could not be obtained from any single dataset, nor with an integration approach that minimized biological differences between datasets.

Methods

To integrate multiple datasets, we start with selecting those informative genes and then normalize the expression of each cell by the library size and a log transformation (*SI Appendix* has more details). The set of unknown parameters $\{\boldsymbol{W}, \{\boldsymbol{H}_j, \boldsymbol{\Lambda}_j, \boldsymbol{b}_j\}_{j=1}^M\}$ is estimated by minimizing the following objective:

$$\frac{1}{N}\sum_{j=1}^{M}\left\|\boldsymbol{X}_j-(\boldsymbol{H}_j\boldsymbol{W}^\top\boldsymbol{\Lambda}_j+\boldsymbol{1}_{n_j}\boldsymbol{b}_j^\top)\right\|^2 +\gamma\sum_{l=1}^{p}\left(\sum_{j=1}^{M}\frac{n_j}{N}\boldsymbol{\Lambda}_j(l,l)-1\right)^2, \quad [3]$$

subjecting to the nonnegative constraints for parameters $\boldsymbol{W}, \boldsymbol{H}_j, \boldsymbol{\Lambda}_j, \boldsymbol{b}_j$. Here, we use $\boldsymbol{X}(i,j)$ to denote the (i,j)-th entry of matrix $\boldsymbol{X}$. The positive parameter γ determines how much penalization is imposed on the batch-specific parameter $\boldsymbol{\Lambda}_j$, which ensures the identifiability of the learned model.

To optimize the above nonconvex objective function, we adopt the widely used block coordinate descent approach (31, 32), with the details deferred to *SI Appendix*. Having eliminated the domain-related factors, the integrated scRNA-seq expression matrices are ready for downstreaming analysis, such as clustering and trajectory inference with the low-dimensional representation $\boldsymbol{H}_j$ and inference of differentially expressed genes with the scaled cell by gene matrix $\boldsymbol{H}_j\boldsymbol{W}^\top$. With the estimated common factor matrix $\boldsymbol{W}$, one can rapidly transfer the learned pattern to a target dataset through a transfer algorithm (detailed below). The dataset-specific factor loading $\boldsymbol{H}_{\text{target}}$, scaling $\boldsymbol{\Lambda}_{\text{target}}$, and shift $\boldsymbol{b}_{\text{target}}$ are estimated with fixed reference factor matrix through a similarly block coordinate descent procedure. Ultimately, the proposed algorithm projects the target data onto the low-dimensional space estimated from the reference dataset. This procedure capitalizes on available information gleaned from prior analyses to optimize the value of small and low-quality datasets. Below, we briefly describe the steps for data integration and transfer tasks, with details deferred to *SI Appendix*.

Data Integration. Given the solution $\left(\hat{\boldsymbol{W}}, \{\hat{\boldsymbol{H}}_j, \hat{\boldsymbol{\Lambda}}_j, \hat{\boldsymbol{b}}_j\}_{j=1}^M\right)$ to the aforementioned optimization problem, one can obtain the estimated data matrix in the shared lower-dimensional factor space as

$$\hat{\boldsymbol{X}}=\begin{bmatrix}\hat{\boldsymbol{X}}_1\\ \vdots\\ \hat{\boldsymbol{X}}_M\end{bmatrix}=\begin{bmatrix}\hat{\boldsymbol{H}}_1\hat{\boldsymbol{W}}^\top\\ \vdots\\ \hat{\boldsymbol{H}}_M\hat{\boldsymbol{W}}^\top\end{bmatrix}=\begin{bmatrix}\widehat{\boldsymbol{H}}_1\\ \vdots\\ \widehat{\boldsymbol{H}}_M\end{bmatrix}\widehat{\boldsymbol{W}}^\top=\widehat{\boldsymbol{H}}\widehat{\boldsymbol{W}}^\top,$$

and one can obtain the intrinsic low-dimensional representation as the estimated factor loadings $\widehat{\boldsymbol{H}}$. As is seen in our real data examples, these batch-adjusted observations are particularly useful for downstreaming analysis, such as cell-type identification.

Data Transfer. We transfer the knowledge to a target dataset that is potentially of lower quality or contains smaller numbers of cells. Specifically, given the target expression matrix X_{target} and the factor matrix $\boldsymbol{W}_{\text{ref}}$ obtained from reference datasets, the goal is to recover parameters $(\boldsymbol{H}_{\text{target}}, \boldsymbol{\Lambda}_{\text{target}}, \boldsymbol{b}_{\text{target}})$ for the target dataset. Similarly, these parameters can be estimated through the following optimization problem:

$$\min_{\boldsymbol{H},\boldsymbol{\Lambda},\boldsymbol{b}} G(\boldsymbol{H},\boldsymbol{\Lambda},\boldsymbol{b};\boldsymbol{W}_{\text{ref}}):=\left\|\boldsymbol{X}_{\text{target}}-(\boldsymbol{H}\boldsymbol{W}_{\text{ref}}^\top\boldsymbol{\Lambda}+\boldsymbol{1}\,\boldsymbol{b}^\top)\right\|_F^2, \quad [4]$$

subject to the nonnegative constraints $\boldsymbol{H}, \boldsymbol{\lambda}, \boldsymbol{b} \geq 0$ and row stochastic constraint $\boldsymbol{H}\boldsymbol{1}=\boldsymbol{1}$. The low-dimensional factor loadings $\hat{\boldsymbol{H}}_{\text{target}}$ can be directly used for further data analysis.

Scalability and Algorithm Speedup. The computation complexity of cFIT scales linearly with the number of cells (*SI Appendix*, Fig. S8*A*). It can accommodate tens of thousands of cells within reasonable run time on a standard personal computer (depending on the number of genes p and the number of factors r). To speed up the processing of millions of cells, we implemented a fast version of the algorithm employing the idea of random sketching (33) and stochastic proximal point (SPP) method (34). In each iteration, a random projection matrix $\boldsymbol{S}_j \in \mathbb{R}^{\tilde{n}_j \times n_j}$ is generated for each batch independently. Then, the parameter sets are updated solving the sketched problem

$$J\left(\boldsymbol{\Theta};\{\boldsymbol{S}_j\}_{j=1}^M\right)=\frac{1}{N}\sum_{j=1}^{M}\left\|\boldsymbol{S}_j\boldsymbol{X}_j-\boldsymbol{S}_j(\boldsymbol{H}_j\boldsymbol{W}^\top\boldsymbol{\Lambda}_j+\boldsymbol{1}_{n_j}\boldsymbol{b}_j^\top)\right\|^2 \quad [5]$$

with the same constraints and penalty. For each coordinate descent subproblem, SPP is employed to ensure the solution stays close to the previous updates, thus encouraging consistency across different sketched problem. This is achieved by adding additional penalty in updating each block of parameters

$$\boldsymbol{\theta}^{(t+1)} = \underset{\boldsymbol{\theta}\geq 0}{\arg\min} \left[J\left(\boldsymbol{\theta}; \boldsymbol{S}^{(t)}, (\boldsymbol{\Theta}\backslash\boldsymbol{\theta})^{(t)}\right) + \frac{1}{2\mu_t} \left\| \boldsymbol{\theta} - \boldsymbol{\theta}^{(t)} \right\|^2 \right], \quad [6]$$

where μ_t controls the diminishing step size. This procedure enables reduced run time roughly proportional to the fraction of subsample size (*SI Appendix*, Fig. S8*B*). More details can be found in *SI Appendix*.

Data Availability. There are no data underlying this work.

ACKNOWLEDGMENTS. We thank Kevin Lin for helpful comments. This work was supported in part by National Institute of Mental Health Grants R01MH123184 (to K.R.) and R37MH057881 (to K.R.). Y.W. was partially supported by NSF Grants CCF-2007911 and DMS-2015447.

1. M. D. Luecken, F. J. Theis, Current best practices in single-cell RNA-seq analysis: A tutorial. *Mol. Syst. Biol.* **15**, e8746 (2019).
2. M. D. Luecken *et al.*, Benchmarking atlas-level data integration in single-cell genomics. https://doi.org/10.1101/2020.05.22.111161 (23 May 2020).
3. V. Y. Kiselev, A. Yiu, M. Hemberg, Scmap: Projection of single-cell RNA-seq data across data sets. *Nat. Methods* **15**, 359–362 (2018).
4. H. A. Pliner, J. Shendure, C. Trapnell, Supervised classification enables rapid annotation of cell atlases. *Nat. Methods* **16**, 983–986 (2019).
5. S. Ge, H. Wang, A. Alavi, E. Xing, Z. Bar-Joseph, Supervised adversarial alignment of single-cell RNA-seq data. *J. Comput. Biol.* http://doi.org/10.1089/cmb.2020.0439 (19 January 2021).
6. Y. Yang *et al.*, SMNN: Batch effect correction for single-cell RNA-seq data via supervised mutual nearest neighbor detection. https://doi.org/10.1101/672261 (18 April 2020).
7. A. Butler, P. Hoffman, P. Smibert, E. Papalexi, R. Satija, Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).
8. E. A. DePasquale *et al.*, CellHarmony: Cell-level matching and holistic comparison of single-cell transcriptomes. *Nucleic Acids Res.* **47**, e138 (2019).
9. B. Hie, B. Bryson, B. Berger, Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. *Nat. Biotechnol.* **37**, 685–691 (2019).
10. T. Stuart *et al.*, Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902 (2019).
11. J. D. Welch *et al.*, Single-cell multi-omic integration compares and contrasts features of brain cell identity. *Cell* **177**, 1873–1887 (2019).
12. L. Haghverdi, A. T. L. Lun, M. D. Morgan, J. C. Marioni, Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.* **36**, 421–427 (2018).
13. R. Raina, A. Battle, H. Lee, B. Packer, A. Y. Ng, "Self-taught learning: Transfer learning from unlabeled data" in *Proceedings of the 24th International Conference on Machine Learning* (Association for Computing Machinery, New York, NY, 2007), pp. 759–766.
14. S. J. Pan, Q. Yang, A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* **22**, 1345–1359 (2009).
15. R. Vilalta, Y. Drissi, A perspective view and survey of meta-learning. *Artif. Intell. Rev.* **18**, 77–95 (2002).
16. J. Donahue et al., "DeCAF: A deep convolutional activation feature for generic visual recognition" in *Proceedings of the 31st International Conference on Machine Learning*, E. P. Xing, T. Jebara, Eds. (Proceedings of Machine Learning Research, Bejing, China, 2014), vol. 32, pp. 647–655.
17. N. Tripuraneni, C. Jin, M. I. Jordan, Provable meta-learning of linear representations. arXiv:2002.11684 (26 February 2020).
18. T. J. Nowakowski *et al.*, Spatiotemporal gene expression trajectories reveal developmental hierarchies of the human cortex. *Science* **358**, 1318–1323 (2017).
19. D. Polioudakis *et al.*, A single-cell transcriptomic atlas of human neocortical development during mid-gestation. *Neuron* **103**, 785–801 (2019).
20. J. G. Camp *et al.*, Human cerebral organoids recapitulate gene expression programs of fetal neocortex development. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 15672–15677 (2015).
21. S. Darmanis *et al.*, A survey of human brain transcriptome diversity at the single cell level. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 7285–7290 (2015).
22. M. Li *et al.*, Integrative functional genomic analysis of human brain development and neuropsychiatric risks. *Science* **362**, eaat7615 (2018).
23. S. Zhong *et al.*, A single-cell RNA-seq survey of the developmental landscape of the human prefrontal cortex. *Nature* **555**, 524–528 (2018).
24. H. T. N. Tran *et al.*, A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biol.* **21**, 1–32 (2020).
25. A. Saunders *et al.*, Molecular diversity and specializations among the cells of the adult mouse brain. *Cell* **174**, 1015–1030 (2018).
26. F. K. Satterstrom *et al.*, Large-scale exome sequencing study implicates both developmental and functional changes in the neurobiology of autism. *Cell* **180**, 568–584 (2020).
27. L. Zhu, J. Lei, L. Klei, B. Devlin, K. Roeder, Semisoft clustering of single-cell data. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 466–471 (2019).
28. J. Wang *et al.*, Transfer learning in single-cell transcriptomics improves data denoising and pattern discovery. *Nat. Methods* **16**, 875–878 (2019).
29. C. Mayer *et al.*, Developmental diversification of cortical inhibitory interneurons. *Nature* **555**, 457–462 (2018).
30. B. Tasic *et al.*, Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nat. Neurosci.* **19**, 335–346 (2016).
31. D. P. Bertsekas, On the Goldstein-Levitin-Polyak gradient projection method. *IEEE Trans. Automat. Contr.* **21**, 174–184 (1976).
32. S. J. Wright, Coordinate descent algorithms. *Math. Program.* **151**, 3–34 (2015).
33. D. P. Woodruff, Sketching as a tool for numerical linear algebra. arXiv:1411.4357 (17 November 2014).
34. P. Bianchi, Ergodic convergence of a stochastic proximal point algorithm. *SIAM J. Optim.* **26**, 2235–2260 (2016).