nature biotechnology

Article

https://doi.org/10.1038/s41587-023-02097-9

# Characterizing the impacts of dataset imbalance on single-cell data integration

Received: 17 November 2022

Accepted: 13 December 2023

Published online: 1 March 2024

Check for updates

Hassaan Maan[1,2,3] ✉, Lin Zhang[1,4], Chengxin Yu[5,6], Michael J. Geuenich[5,6], Kieran R. Campbell[2,5,6,7,8,9] ✉ & Bo Wang[1,2,3,8,10] ✉

Computational methods for integrating single-cell transcriptomic data from multiple samples and conditions do not generally account for imbalances in the cell types measured in different datasets. In this study, we examined how differences in the cell types present, the number of cells per cell type and the cell type proportions across samples affect downstream analyses after integration. The Iniquitate pipeline assesses the robustness of integration results after perturbing the degree of imbalance between datasets. Benchmarking of five state-of-the-art single-cell RNA sequencing integration techniques in 2,600 integration experiments indicates that sample imbalance has substantial impacts on downstream analyses and the biological interpretation of integration results. Imbalance perturbation led to statistically significant variation in unsupervised clustering, cell type classification, differential expression and marker gene annotation, query-to-reference mapping and trajectory inference. We quantified the impacts of imbalance through newly introduced properties—aggregate cell type support and minimum cell type center distance. To better characterize and mitigate impacts of imbalance, we introduce balanced clustering metrics and imbalanced integration guidelines for integration method users.

Single-cell sequencing technologies developed over the past decade have led to major discoveries across fields[1–3]. Methods for removing batch effects from bulk RNA sequencing data have demonstrated poor performance in single-cell settings[4], and batch correction/integration techniques have been developed specifically for single-cell RNA sequencing (scRNA-seq) data[5]. Current single-cell integration methods underperform in settings where datasets are imbalanced based on differences in the cell types present, the number of cells per cell type and cell type proportions across samples[4,6]. Imbalanced datasets occur in many integration contexts, including developmental and cancer biology. As these contexts are common in single-cell data analysis, integration methods and analysis pipelines must be able to explicitly address dataset imbalance, or integration results may lead to inaccurate biological conclusions.

In comprehensive single-cell integration benchmarking studies by Tran et al.[4] and Luecken et al.[7], scRNA-seq integration methods were found to perform poorly, in terms of both batch correction and cell type identity conservation metrics, in large and imbalanced datasets. Ming et al.[6] performed simulation studies for imbalanced cell type compositions in scRNA-seq integration settings and demonstrated that

[1]Peter Munk Cardiac Centre, University Health Network, Toronto, Ontario, Canada. [2]Vector Institute, Toronto, Ontario, Canada. [3]Department of Medical Biophysics, University of Toronto, Toronto, Ontario, Canada. [4]Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, British Columbia, Canada. [5]Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada. [6]Lunenfeld-Tanenbaum Research Institute, Toronto, Ontario, Canada. [7]Department of Statistical Sciences, University of Toronto, Toronto, Ontario, Canada. [8]Department of Computer Science, University of Toronto, Toronto, Ontario, Canada. [9]Ontario Institute for Cancer Research, Toronto, Ontario, Canada. [10]Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, Ontario, Canada. ✉e-mail: hassaan.maan@mail.utoronto.ca; kierancampbell@lunenfeld.ca; bo.wang@uhnresearch.ca

https://doi.org/10.1038/s41587-023-02097-9

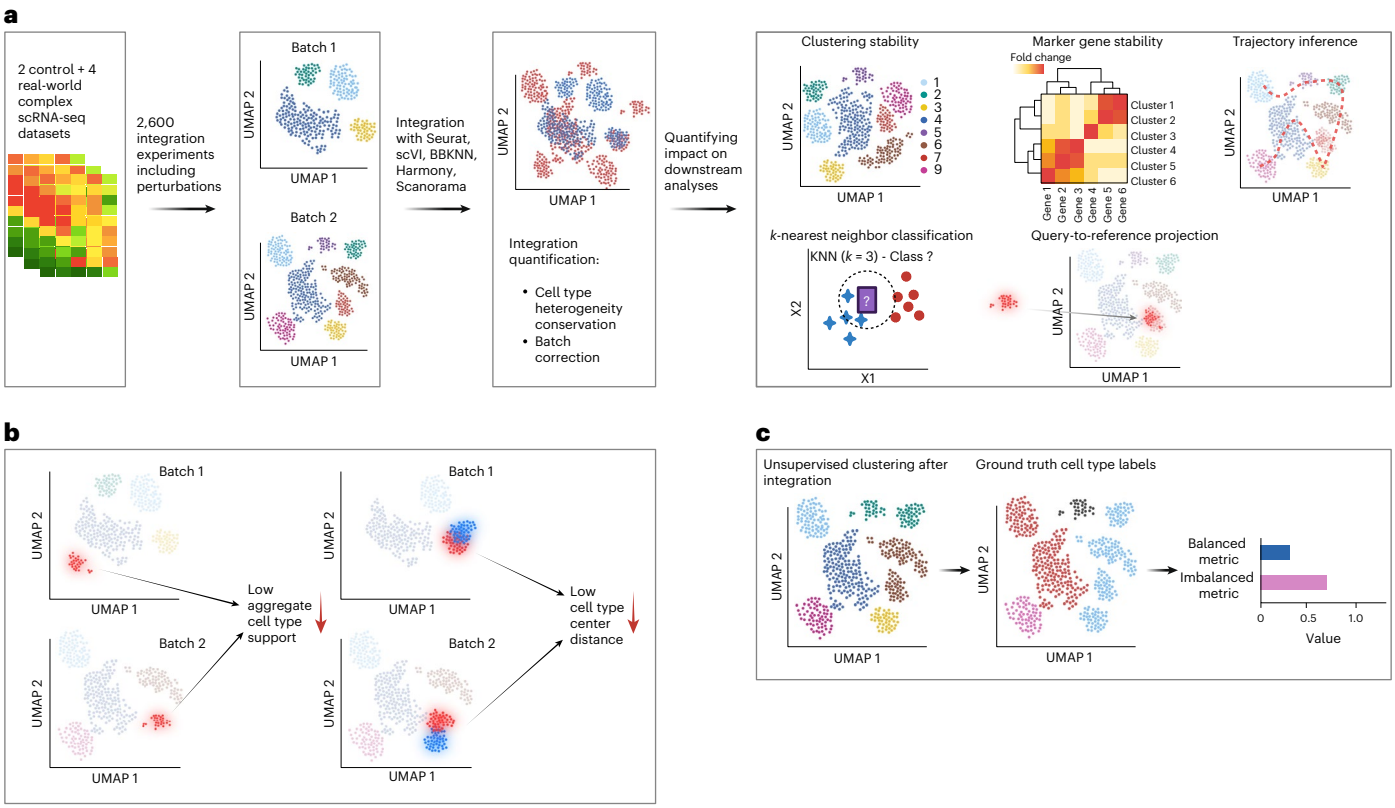



**Fig. 1 | Overview of the Iniquitate pipeline and analysis results. a**, To determine the effects of dataset imbalance in scRNA-seq integration, two control balanced datasets and four complex datasets with imbalance already present were integrated using current state-of-the-art scRNA-seq integration techniques. A total of 2,600 integration experiments involving downsampling across datasets were performed, and the effects of imbalance on integration results as well as downstream analyses (clustering, DGE, cell type classification, query-to-reference prediction and trajectory inference) were quantified. **b**, Two key data characteristics were found to contribute to altered downstream results in imbalanced settings: aggregate cell type support (cell type imbalance) and minimum cell type center distance (transcriptomic similarity). **c**, To account for imbalanced scRNA-seq integration scenarios in evaluation and benchmarking, typically used metrics and scores were reformulated to reweigh disproportionate cell types, which includes the bARI, bAMI, Balanced Homogeneity Score, Balanced Completeness and Balanced V-measure.

cell type proportion imbalance leads to skewed distributions in standardized gene expression values between datasets. This drives major changes in the dimensionality reduction step in scRNA-seq analysis and subsequently leads to inaccurate integration results[6]. Currently, no existing study has quantified the effects of dataset imbalance on both integration results and downstream biological conclusions. This aspect is highly relevant, as mechanisms to account for dataset imbalance do not readily exist in frequently used integration techniques[4,7].

Here we present an extensive analysis of the effects of dataset imbalance on scRNA-seq data integration. We begin by examining two balanced scRNA-seq batches of human peripheral blood mononuclear cells (PBMCs)[4,8,9] and two balanced batches of mouse mesenchymal organogenesis cells[10], as controlled settings. To determine the effects of dataset imbalance on integration results and downstream analyses, we performed integration experiments with perturbations to dataset balance. The downstream analyses tested include unsupervised clustering[5], differential expression to determine marker genes[5], nearest-neighbor-based cell type classification[11], query-to-reference cell type annotation[12] and trajectory inference[10,13]. To extend testing to more complex settings, we analyze datasets with prevalent imbalance, including imbalanced PBMC datasets[14], temporal mouse hindbrain developmental data[15] and pancreatic ductal adenocarcinoma (PDAC) samples from different patients[16].

Our analyses reveal that dataset imbalance has cell-type-specific effects on integration performance as well as the downstream results, and that these effects are largely method agnostic. We define two key properties of multi-sample single-cell data that act in concert to affect downstream results: 'aggregate cell type support' and 'minimum cell type center distance'. To address limitations in benchmarking single-cell integration of imbalanced data, we reformulated current integration metrics to consider imbalance explicitly. Finally, we provide a series of guidelines and recommendations to help mitigate the impacts of dataset imbalance in scRNA-seq integration settings.

## Results

### A perturbation pipeline to quantify impacts of imbalance

To assess the impacts of dataset imbalance in scRNA-seq integration, we developed a pipeline, termed Iniquitate, that tests the effects of downsampling perturbations on integration and downstream analysis results (Fig. 1a and Methods). Datasets used were annotated by experts in their respective studies, with the exception of the PDAC data, which were re-annotated to better identify malignant cells (Methods). We tested five state-of-the-art scRNA-seq integration methods, including BBKNN[17], Harmony[18], Scanorama[19], scVI[20] and Seurat[21]. A uniform integration pipeline embedded within Iniquitate was used to make analyses between methods and across datasets comparable (Methods).

After investigating factors that can quantifiably lead to distinct results after integration, we found the transcriptomic similarity between cell types (minimum cell type center distance) and the imbalance between cell types (aggregate cell type support) to be the most relevant and predictive in this regard (Fig. 1b). To account for gaps in benchmarking, we developed balanced clustering metrics, which

https://doi.org/10.1038/s41587-023-02097-9

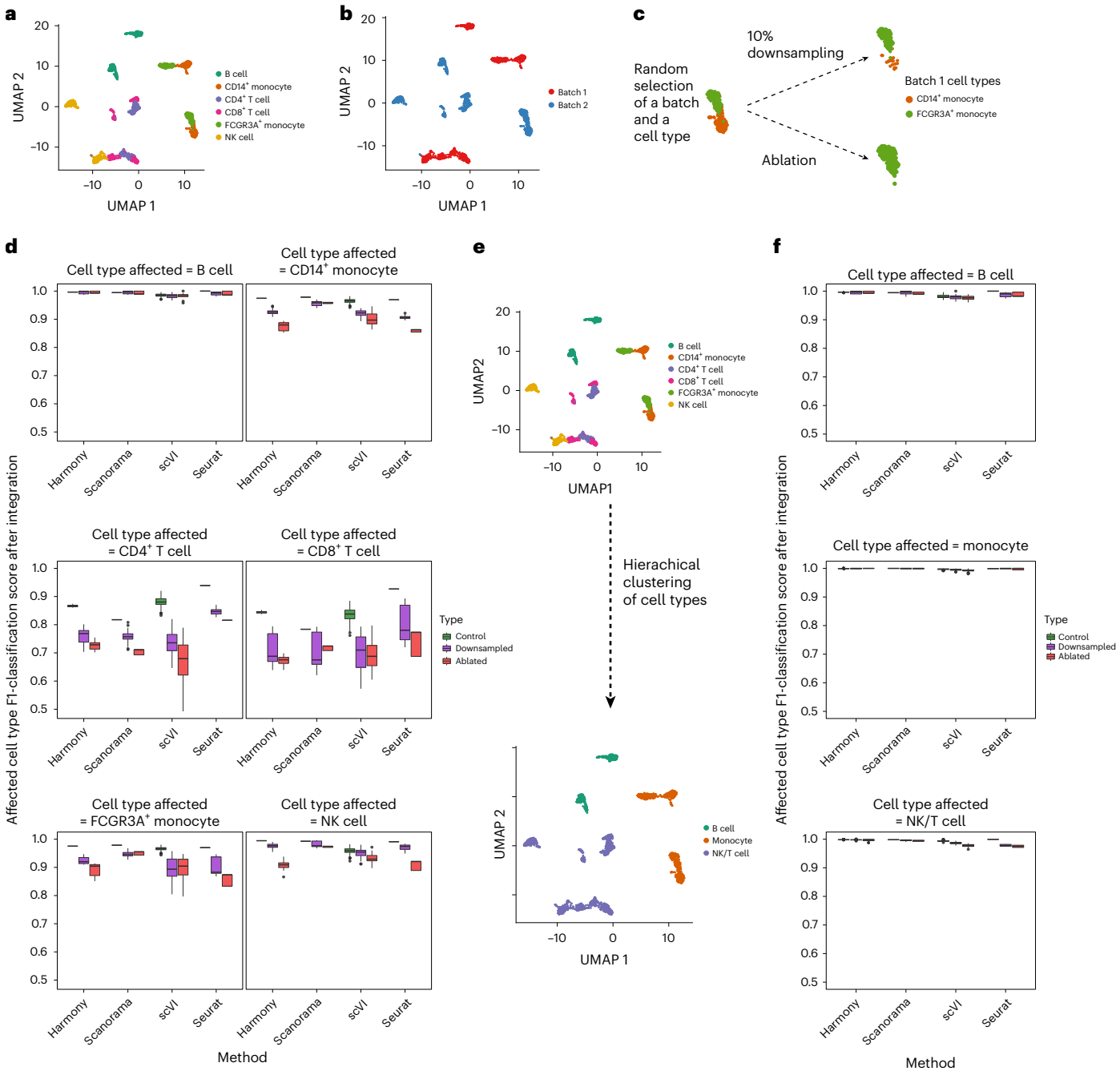



**Fig. 2 | Perturbation analysis of controlled PBMC dataset and effects on cell-type-specific integration. a**,**b**, The cell type and batch representations of the balanced two batch PBMC dataset. **c**, The perturbation setup for the balanced PBMC data. In each iteration, one batch and one cell type are randomly selected, and the cell type is randomly either downsampled to 10% of its original number or ablated. Control experiments are also performed where no downsampling occurs. **d**, KNN classification within the integrated embedding space in control, downsampling and ablation experiments ($n = 800$ independent integration experiments) and across methods. The F1 scores are indicated for the same cell type that was downsampled. **e**, Hierarchical clustering of similar cell types in the balanced two-batch PBMC data. **f**, Cell-type-specific integration results using a KNN classifier after hierarchical clustering across perturbation experiments ($n = 800$ independent integration experiments) with the same setup as **d**. The cell types here are based on the label after hierarchical clustering from **e**. Box plots (**d**,**f**) indicate median values across experiments; hinges are the 25th and 75th percentile values; and whiskers indicate the 1.5× interquartile range (IQR) values from the hinges.

reweigh the base scores such that each ground truth cell type's contribution to the score is considered equally (Fig. 1c).

## Imbalance leads to cell-type-specific integration effects

We began by analyzing a PBMC cohort of two batches/samples processed independently from two different healthy donors[4,8,9]. Downsampling was performed in each batch, leading to six major cell types and an equal number of cells within each cell type (400 cells for each cell type) (Fig. 2a and Methods). The cell types were selected such that they are equivalent between the batches. Therefore, the cell types present, the number of cells per cell type and cell type proportions between the batches are equal, and the integration scenario is balanced (Fig. 2a,b). A batch effect is present between the samples (Fig. 2b) (10x Genomics 3′ versus 5′ protocols; Methods)[8,9]. We aimed to

assess how the integration results of the two PBMC batches varied between the control balanced data and perturbed imbalanced data. For each perturbation experiment, we randomly selected one of the two batches and one cell type to either downsample to 10% of the original population or ablate/remove completely from the selected batch (Fig. 2c). Perturbations were repeated 200 times for both downsampling and ablation of a random batch/cell type, and control experiments with no perturbations to the balanced data were repeated 400 times (Methods).

Global metrics, such as the Adjusted Rand Index (ARI), inadequately captured cell-type-specific performance variation in imbalanced scenarios (Supplementary Note 1). Therefore, we examined integration performance at a cell-type-specific level through a *k*-nearest-neighbor (KNN) classifier[11,22] that was trained on 70% of the post-integration embeddings from each method independently, and the remaining 30% was used as a test set for cell type classification (Methods). The train/test split was stratified, ensuring equal proportions of cell types in both subsets (Methods). Overall, the classification results provide evidence for cell-type-specific effects of dataset imbalance, as downsampling or ablating a specific cell type led to a statistically significant decrease in the KNN classification F1 score[23] for the same cell type after integration (ANOVA $P < 0.05$, $F = 1{,}304.96$) (Fig. 2d). This result is method agnostic as the ANOVA test factored in method used and cell type downsampled (Methods). The only cell type that exhibited stability was B cells (Fig. 2d; standard deviation of median F1 score across methods and experiment types < 0.01).

We hypothesized that the integration performance for B cells was unaffected because they are transcriptionally distinct from the other cell types (Supplementary Fig. 24). As a test, we performed hierarchical clustering of the cell types into three higher-level subsets based on their similarity: B cells, monocytes and natural killer (NK)/T cells (Fig. 2e and Methods). Downsampling these subsets did not result in worsening performance to the same degree as the base cell types (Fig. 2d,f) (ANOVA $F = 374.46$ (hierarchical) < 1,304.96 (base); Supplementary Fig. 7 and Methods). This result indicates that the relative transcriptomic similarity of cell types can lead to changes in cell-type-specific performance of integration techniques in imbalanced scenarios. The 'transcriptomic similarity' property is formalized as the minimum cell type center distance, whereas the 'cell type imbalance' property is formalized as the aggregate cell type support (Fig. 1b and Supplementary Note 4). Minimum cell type center distance refers to the distance of the closest cell type in principal component (PC) space, and aggregate cell type support refers to the total number of cells for a given cell type across all batches (Supplementary Note 4 and Methods). For brevity, we subsequently refer to these properties as cell type center distance and cell type support, respectively. The experimental results indicate that these two properties act in tandem to lead to quantitation differences in scRNA-seq integration.

We further tested a method that was designed to handle imbalance in CIDER, but, despite the claims by CIDER and other tested methods, the same effects were observed (Supplementary Note 2).

## Balance and similarity affect downstream analyses

To further analyze the impact of the perturbation experiments on the balanced PBMC cohort, we quantified the effects of imbalance on downstream analyses typically performed after integration, including unsupervised clustering, differential gene expression (DGE), query-to-reference annotation and trajectory inference (Fig. 1a). We used the same dataset, perturbation setup and downsampling experiments as in Results section "Imbalance leads to cell-type-specific integration effects". Assessment of the impacts of imbalance on trajectory inference was done on a separate dataset of mammalian organogenesis[10].

**Stability of unsupervised clustering after integration.** We observed a significant variation in the inferred number of clusters after integration across all tested methods due to perturbation of cell type balance (ANOVA $P < 0.05$, $F = 10.189$) (Fig. 3a and Methods). After integration in both balanced and perturbed experiments, clustering was performed using the Leiden clustering algorithm with a method-specific resolution that best approximated the true number of cell types in the balanced data (Methods). Although all methods indicated at least some degree of variation in the number of clusters between control and downsampling/ablation experiments, there were also method-dependent effects present (Fig. 3a). For instance, whereas Scanorama exhibited variation in the number of clusters regardless of cell type downsampled/ablated, downsampling/ablation of NK cells and FCGR3A$^+$ monocytes specifically led to a much smaller number of clusters after integration (Fig. 3a). For BBKNN and Seurat, the greatest decrease in the number of clusters occurred after ablating CD8$^+$ T cells (Fig. 3a). Harmony exhibited an increase in the number of clusters after downsampling/ablation, regardless of cell type, and scVI diverged only when the monocyte subsets were affected (Fig. 3a). There was variation observed for the control experiments across methods as well, but stronger deviation after perturbation was present in all tested methods. This result indicates that differing levels of imbalance can cause deviations in cluster number, even though the number of clusters should be stable, as the number of cell types across all batches remains the same in both perturbed and unperturbed experiments.

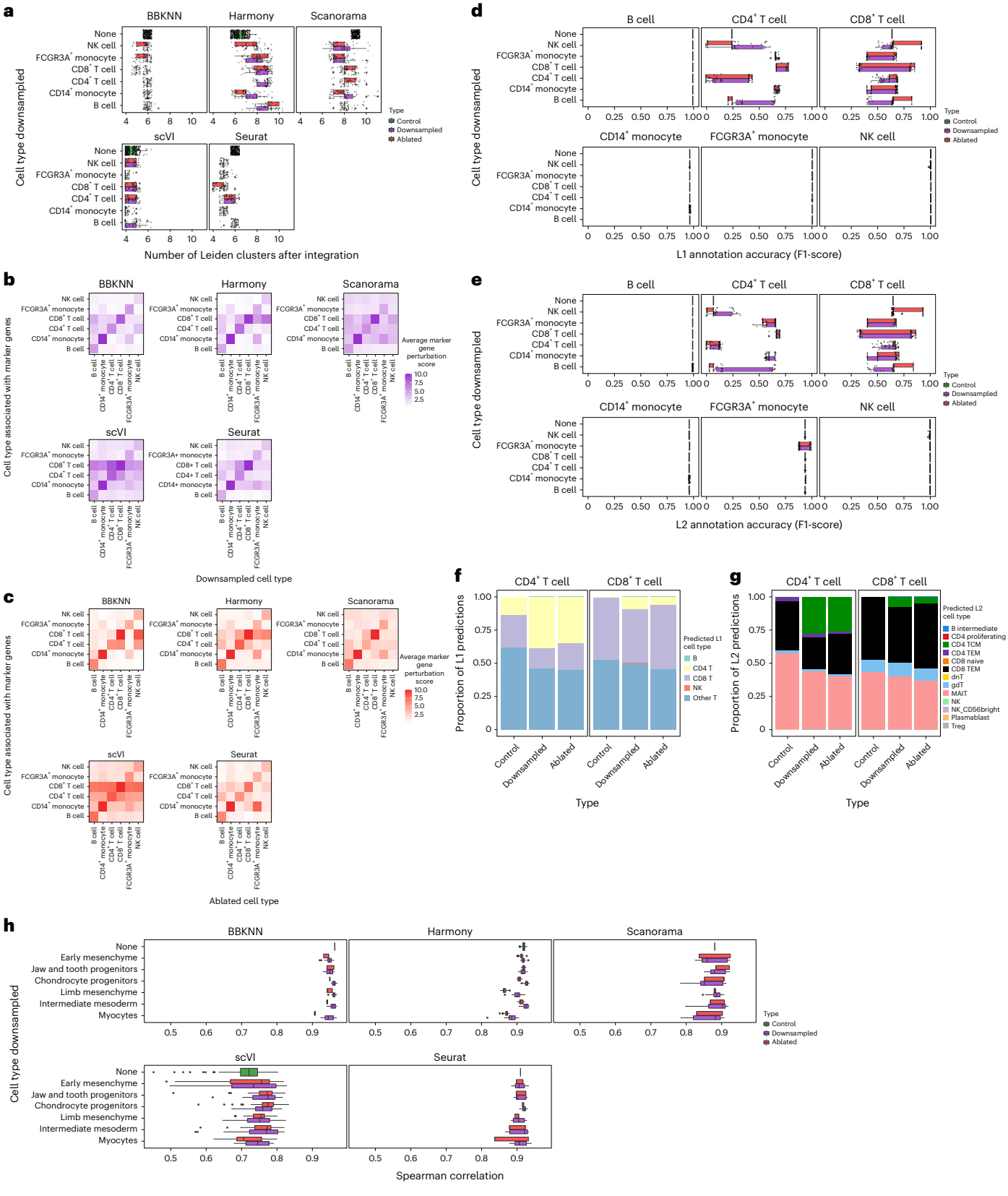
**DGE and marker gene stability.** Frequently, the next step after integration and unsupervised clustering in an scRNA-seq analysis workflow is DGE analysis[5,24]. A series of one-versus-all differential expression experiments, using statistical tests such as the non-parametric Wilcoxon rank-sum test or more RNA-seq-specific techniques such as DESeq2, are typically performed for each cluster to determine the top-ranking 'marker genes'[5,12,24]. These marker genes are used to annotate clusters into putative cell types[5,12,24]. One way to assess marker gene stability before and after perturbation is to constrain the number of clusters to be equivalent across experiments, but this would be unrealistic as variation in cluster number in both control and perturbed experiments was observed across methods (Fig. 3a and Supplementary Fig. 15). As the 'ranking' of marker genes is typically used for annotation[5,24], we considered deviation in ranking for genes with known cell type associations to be an important endpoint. We determined the

**Fig. 3 | Quantification of the effects of dataset imbalance on downstream analyses. a**, After integration of the balanced PBMC dataset in different perturbation scenarios (type) and based on the cell type downsampled, the number of unsupervised clusters from the results of each method based on Leiden clustering across experiments. **b**, The average marker gene ranking change in DGE (average marker gene perturbation score) for cell types downsampled in the balanced PBMC dataset, across methods. **c**, The average marker gene ranking change in DGE, for the 'ablation' experiment type in the balanced PBMC dataset. **d**,**e**, The cell-type-specific L1 annotation (coarse-grained) (**d**) and L2 annotation (fine-grained) (**e**) accuracy scores across experiments for query-to-reference results for individual batches in the balanced PBMC dataset, based on experiment type (control, downsampling and ablation) and cell type downsampled. **f**,**g**, The L1 predictions (**f**) and L2 predictions (**g**) by proportion across experiment types and experiments for CD4$^+$ T cells and CD8$^+$ T cells. **h**, The Spearman correlation between the estimated pseudotime for cells in the unintegrated data compared to the integrated data for the different methods in the balanced mesenchymal organogenesis dataset. A total of $n = 800$ integration experiments involving control, downsampling and ablation subsets were done for each analysis. Box plots (**a**,**d**,**e**,**h**) indicate median values across experiments; hinges are the 25th and 75th percentile values; and whiskers indicate the 1.5× interquartile range (IQR) values from the hinges. All values are overlaid on the box plots in **a**.

https://doi.org/10.1038/s41587-023-02097-9

top 10 marker genes for each cell type and assessed the stability of their ranking before and after perturbation (Methods). Changes in ranking of marker genes after perturbation were defined as the 'marker gene perturbation score' (Methods). In the case of examining all marker genes for a given cell type, the changes in ranking across marker genes was averaged ('average marker gene perturbation score') (Methods).

For most marker genes, we observed deviations in ranking after downsampling and ablation up to 10 ranks (Supplementary Fig. 12). This could lead to changes in the biological interpretation of results if the top marker genes are used as a heuristic for annotation (Supplementary Note 3). An ANOVA test factoring in the specific marker gene, method and downsampled cell types indicated that perturbation



led to statistically significant changes in ranking (ANOVA $P < 0.05$, $F = 57.174$—highest of all factors) (Methods). We examined whether downsampling or ablation of a specific cell type will change the ranking of marker genes for the same cell type, and we observed that this was the case across all methods (Fig. 3b,c). The strongest ranking change of marker genes occurred after downsampling or ablation of $CD8^+$ T cells and $CD14^+$ monocytes (Fig. 3b,c). As these two cell types are highly similar to $CD4^+$ T cells and $FCGR3A^+$ monocytes, respectively, downsampling induces a loss of their information in the integrated representation and subsequent clustering step, which the DGE step is reliant upon. This result is consistent with the synergy between the cell type center distance and cell type support properties. Changes in marker gene ranks were also observed for cell types that were not downsampled or ablated, such as in NK cells, which were pronounced for Harmony and scVI results (Fig. 3b,c). Once again, this is likely due to mixing of cell types within clusters after an imbalance is introduced, as NK cells are very transcriptionally similar to $CD4^+$ and $CD8^+$ T cell subsets (Supplementary Fig. 24).

**Query-to-reference projection and cell type annotation.** The accuracy of annotation in a query-to-reference projection setting depends on the quality of the integrated space. To examine the effects of imbalance in this setting, we used the Seurat 4.0 query-to-reference annotation pipeline and a large-scale multi-modal PBMC dataset of 211,000 cells as a ref. 25. In the Seurat 4.0 pipeline, each batch (query) is projected to the reference dataset, such that integration is performed individually for each batch[25]. Perturbations were done for the query batches (balanced PBMC two-batch data), and the reference was static (Methods). The annotation of most cell types was stable across control and downsampling/ablation experiments with near-perfect scores (Fig. 3d,e). However, the two T cell subsets had varying performance to a high degree, regardless of which cell type was downsampled or ablated (Fig. 3d,e). This result is indicative of the fact that the imbalance between the projected batch (which was perturbed) and the reference dataset (held constant across all experiments) is driving variance in integration and subsequent annotation results. This highlights a similar problem concomitant with previous results, in that perturbing the degree of balance for transcriptionally similar cell types (T cell subsets) can lead to biologically distinct results compared to the balanced scenario. After perturbing the degree of balance within a given batch, the tradeoff point is moved in favor of either T cell subset (Fig. 3d,e).

Examining the predicted cell type annotations more closely at two levels of resolution, we observed that both the $CD4^+$ and $CD8^+$ T cells were largely mis-annotated as mucosal-associated invariant T (MAIT) cells (Fig. 3f,g). After downsampling or ablation of a given cell type and subsequent analysis of annotation accuracy of the same cell type, we found that $CD4^+$ T cells were annotated more accurately, whereas $CD8^+$ T cells were further mis-annotated (Fig. 3f). The transcriptional similarity between not just the $CD4^+$/$CD8^+$ subsets but also the many subsets that fall under 'other T' is a challenging problem for integration and subsequent label transfer. This challenge is exacerbated when imbalance is present, as indicated by the perturbation experiments and their effects on the annotation results. Similar to previous experiments, this result highlights the combined effects of cell type center distance and cell type support.

**Trajectory inference.** As the balanced PBMC dataset does not have an inherent differentiation trajectory, a mouse organogenesis dataset comprising two batches of measurements on different days was used[10]. The mesenchymal trajectory from this dataset was isolated, and abundant cell types in both batches were downsampled to an equivalent number, leading to a balanced scenario with respect to the cell types, cell type numbers and proportions (Methods and Supplementary Figs. 17 and 18). Downsampling and ablation experiments were performed in the same manner as the balanced PBMC dataset, and the correlations of the estimated pseudotime values were compared before and after integration (Methods). The stability of these correlations was then examined among the control, downsampled and ablated cases (Methods). Even though the batch effects are subtle in this dataset, the results indicated that imbalance can still lead to instability in estimated pseudotime values (ANOVA $P < 0.05$, $F = 24.504$) (Fig. 3h). In particular, downsampling or ablation of myocytes led to the largest deviations in correlation across methods (Fig. 3h).

Overall, cell type imbalance affected all four aspects of downstream analysis that were tested, and we observed strong evidence of impact on biological interpretation of the results. This observation is likely even more relevant in complex datasets, as the balanced PBMC and mammalian organogenesis cohorts are not representative of the ever-increasing throughput of scRNA-seq protocols[26]. An extended analysis of all downstream analyses steps tested, with additional stability experiments, is outlined in Supplementary Note 3.

## Tumor compartment-specific effects of dataset imbalance

To further analyze the effects of dataset imbalance in complex scenarios, we considered a PDAC dataset of eight batches comprising tumor samples across eight different biopsies[16]. One major challenge in the analysis of PDAC data is accurate delineation of tumor cells from normal non-cancerous epithelial cells[27,28]. As both acinar and ductal epithelial cells have been proposed as cell-of-origin candidates in PDAC across numerous studies[29,30], reliably separating tumor cells from these normal epithelial cell types in scRNA-seq data remains a major computational challenge. We sought to determine if different levels of imbalance between epithelial normal and epithelial tumor compartments could influence the accuracy of PDAC sample integration and subsequent classification of tumor cells. We pre-processed and annotated cells in the PDAC samples through reference-based annotation and copy number analysis (Methods). We grouped epithelial normal cells (acinar and ductal) into the 'epithelial normal' compartment, tumor cells into the 'epithelial tumor' compartment and the remaining microenvironment cells into the 'microenvironment' compartment (Fig. 4a,b and Methods). Perturbation experiments included downsampling or ablation of a randomly selected compartment within four randomly selected batches out of eight (Fig. 4a,b), and, after including control runs, this led to a total of 200 integration experiments (Methods).

Batch mixing is a poor quantifier of integration performance in this highly imbalanced scenario (Fig. 4a,b). Therefore, we examined the KNN classification scores on a per-compartment downsampled, per-compartment assessed basis (Methods). The results indicated that downsampling or ablating the microenvironment compartment leads to stable compartment classification across all methods, with a slight decrease in performance observed for Seurat in the epithelial normal and tumor compartments (Fig. 4c). Acinar and ductal cells, which comprise the epithelial normal and epithelial tumor populations, are two of the most distant cell types from others in the data (Supplementary Fig. 26). Therefore, this result is concordant with the cell type center distance property. Downsampling the tumor and normal epithelial compartments led to the greatest decrease in the integration performance of the same compartment (Fig. 4c) ($F_{\text{Epithelial tumor,Epithelial normal}} > F_{\text{Microenvironment}}$; Supplementary Fig. 23 and Methods). These results demonstrate that the degree of imbalance between transcriptionally similar compartments across tumor tissue cohorts can substantially affect the downstream results and possibly subsequent analyses.

## Balanced clustering metrics for imbalanced integration

Metrics such as the ARI are agnostic to information on label proportions[31,32] and were found to be inadequate for assessing integration performance in imbalanced datasets (Supplementary Note 1). As such, we developed balanced versions of these scores, including the Balanced Adjusted Rand Index (bARI), Balanced Adjusted Mutual Information

https://doi.org/10.1038/s41587-023-02097-9

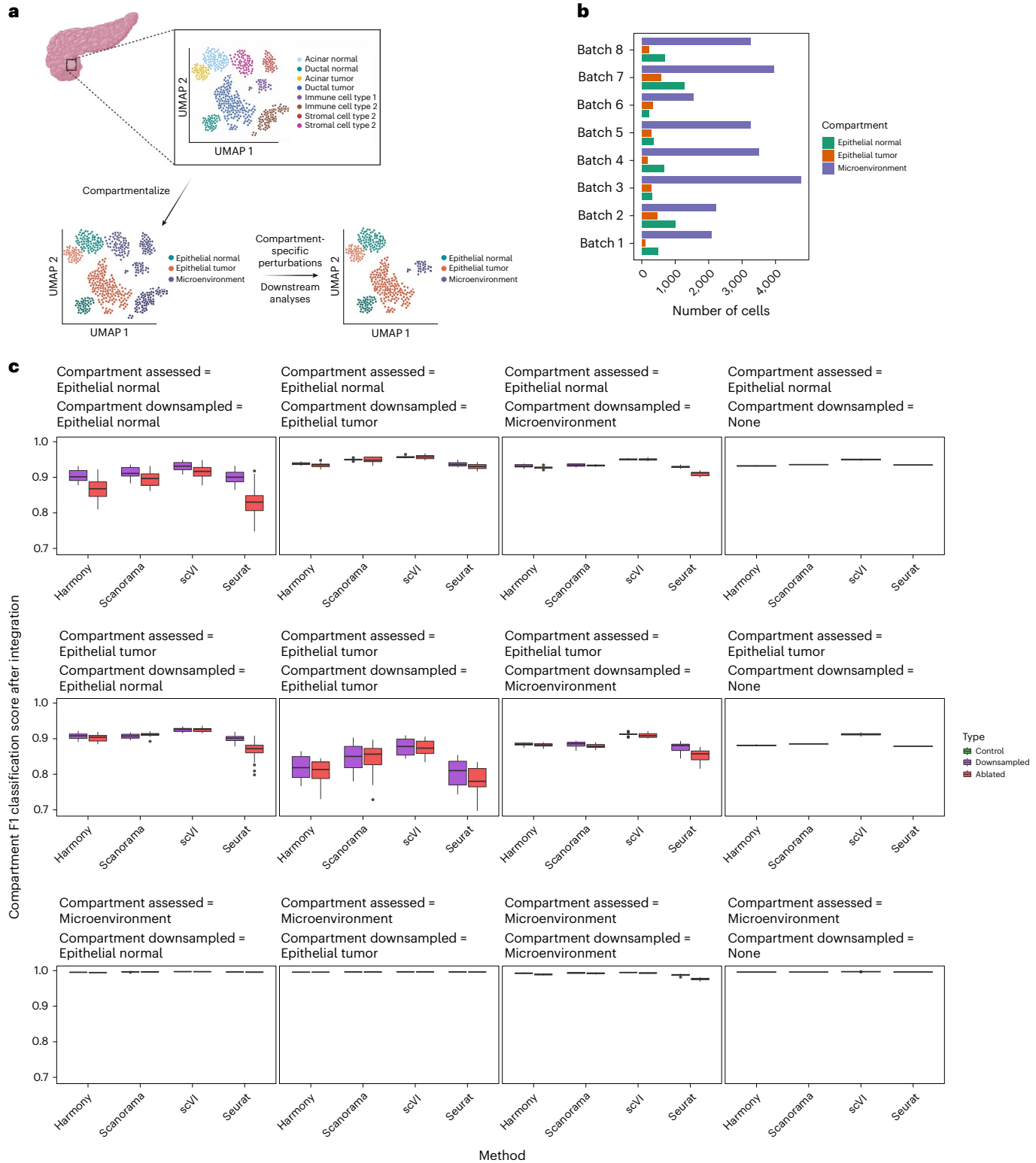



**Fig. 4 | Compartment-wise perturbation experiments for eight batches of PDAC biopsy samples. a**, Overview of the experimental setup. To determine the effects of dataset imbalance across epithelial cell compartments, various microenvironment cells were collapsed into the 'microenvironment' compartment, normal ductal and acinar cells into the 'epithelial normal' compartment and malignant ductal and acinar cells into the 'epithelial tumor' compartment. The perturbation experiments involved downsampling (10% of a compartment) and ablation (complete removal of a compartment) for four of eight randomly selected batches ($n = 200$ independent integration experiments). Note that all batches are integrated at once using each method. **b**, Number of cells in each compartment after cell type collapse, across batches/biopsy samples in the PDAC data. **c**, F1 classification score for KNN classification after integration, specific to each compartment when compared to the compartment that was downsampled or ablated, across experiments and methods used for integration. Box plots (**c**) indicate median values across experiments; hinges are the 25th and 75th percentile values; and whiskers indicate the 1.5× interquartile range (IQR) values from the hinges.

https://doi.org/10.1038/s41587-023-02097-9

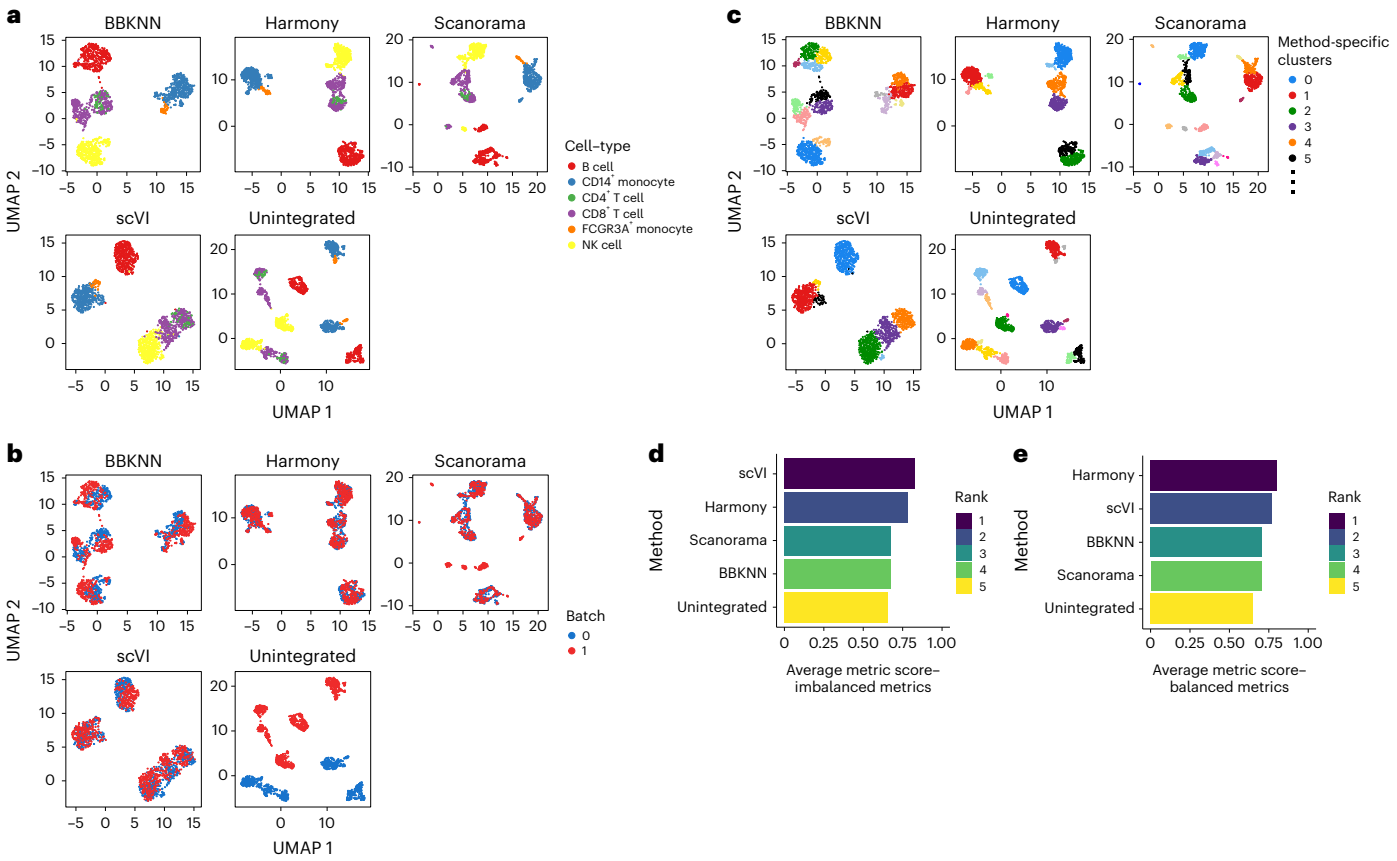



**Fig. 5 | Benchmarking single-cell data integration using balanced clustering metrics. a**, Cell-type-labeled UMAP plot for the balanced two-batch PBMC data with FCGR3A$^+$ monocytes and CD4$^+$ T cells downsampled to 10% of their original proportion in one batch, after integration with the tested methods as well as an unintegrated representation. **b**, Batch-labeled UMAP plot for the integrated and unintegrated downsampled two-batch PBMC data. **c**, Unsupervised clustering-labeled UMAP plot for Leiden clustering in the embedding space of the integrated and unintegrated results for the downsampled two-batch PBMC data. Note that each method/subset has its own unsupervised clusters, and they do not overlap. **d**,**e**, Scoring and ranking of integration results, when considering concordance of the unsupervised clustering labels and ground truth cell type labels for each integration method and the unintegrated subset, using the average results of the base (imbalanced) clustering metrics (**d**) (ARI, AMI, Completeness and Homogeneity) and average of the balanced clustering metrics (**e**) (bARI, bAMI, Balanced Completeness, Balanced Homogeneity and Balanced V-measure).

(bAMI), Balanced Homogeneity, Balanced Completeness and the Balanced V-measure (Fig. 1c and Methods). These metrics weigh each cell type present equally and are not driven by cell types present in high proportions (Methods). We first demonstrated the utility and stability of the proposed balanced clustering metrics on simulated data and a constructed single-cell example (Supplementary Note 5).

To determine if the balanced metrics can change the results of a benchmarking analysis, we downsampled CD4$^+$ T cells and FCGR3A$^+$ monocytes from one batch in the balanced two-batch PBMC dataset and performed integration using BBKNN, Harmony, Scanorama and scVI (Fig. 5a,b and Methods). After integration, unsupervised clustering was performed, and the clusters were compared to ground truth labels using the average base and balanced metric scores (Fig. 5c–e and Methods). When using the average base metric scores, scVI ranked the highest and BBKNN ranked the worst (Fig. 5d). The base metric scores for Scanorama and BBKNN were almost the same as using the unintegrated embedding (Fig. 5d). Scanorama and BBKNN have performed well in previous comprehensive integration benchmarking studies[4,7], which is not in concordance with this result. Considering the average balanced metric scores, the rankings changed, as Harmony became the top performer (switched with scVI), and BBKNN now performed better than Scanorama (Fig. 5e). There is a larger separation in scores between the unintegrated embedding and the results of BBKNN and Scanorama using the balanced metrics (Fig. 5d), which is a more valid and expected result. Overall, the results indicate that the balanced clustering metrics capture nuances in the data related to cell type imbalance that the base metrics cannot.

## End-user guidelines for imbalanced integration

To aid in the integration of imbalanced datasets, we introduce general guidelines for users of integration techniques (Fig. 6 and Supplementary Table 1) and provide a full vignette containing an extensive code implementation (see 'Code availability'). An important aspect to consider when using these guidelines is that prior knowledge on potential disparity and cell type similarity in the datasets can help guide the degree of desired batch mixing. For instance, in analyzing heterogeneous tumor samples from distinct patients with disparate cell types and proportions, biological heterogeneity conservation is likely to be poor if batch mixing is prioritized in integration[5]. However, this may be a desired result if the end analysis goal is only to assess common variation between the tumor samples[5]. Regardless, judging the degree of desired batch mixing is often very difficult in practice[5]. As such, we emphasize an iterative process where imbalance, degree of batch correction and conservation of biological heterogeneity are assessed at multiple steps in the scRNA-seq integration pipeline (Fig. 6).

Imbalance within datasets to be integrated can be assessed based on pre-integration tests using unsupervised clustering and/or query-to-reference annotation (Fig. 6 and Supplementary Table 1). If either of these outlined pre-integration tests reveals imbalance, the integration step itself can be altered by (1) picking an integration

https://doi.org/10.1038/s41587-023-02097-9

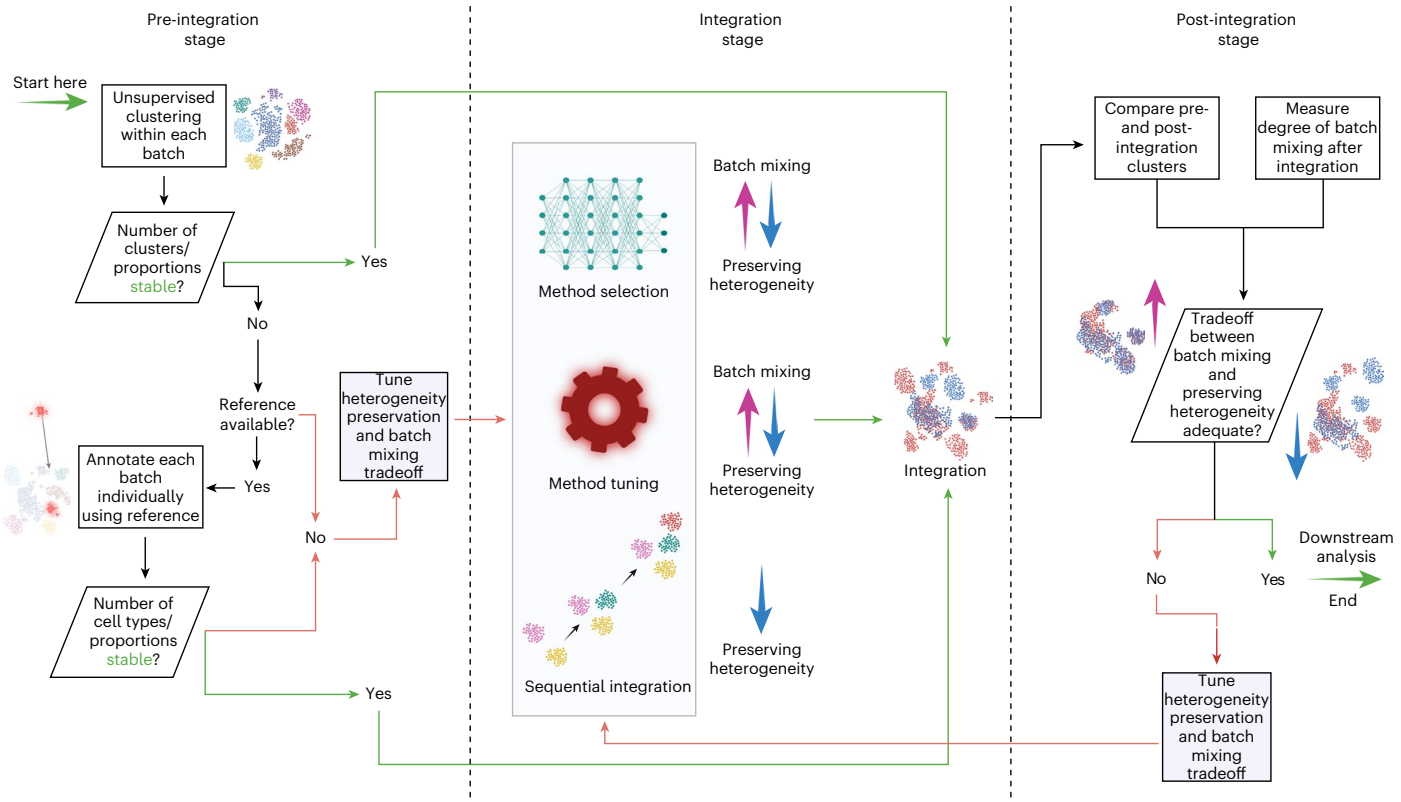



**Fig. 6 | Guidelines for single-cell data integration in imbalanced settings.** A stepwise procedure is outlined, starting with diagnostic tests in the pre-integration stage that dictate whether or not to tune integration methods at the integration stage or perform further steps in the pre-integration stage. After integration, the tradeoff between batch mixing and conservation of biological heterogeneity can be diagnosed, and, if determined to be inadequate, further tuning at the integration stage can be done.

method that is suitable for preserving biological heterogeneity; (2) tuning the integration method itself to better preserve biological heterogeneity; and (3) performing sequential integration if the datasets are known or suspected to have temporal structure[33] (for example, developmental data) (Fig. 6). There is also the possibility of integrating only shared putative cell types between batches if a reference dataset is available, as this would better ensure that imbalance is minimized in the integration step (Supplementary Table 1). After the integration step, post-integration techniques to assess preservation of biological heterogeneity and degree of batch mixing can be used to determine the current balance between the two desired outcomes[5], and integration and pre-integration steps can be further tuned to strike the desired balance (Fig. 6).

## Discussion

In this study, we analyzed the effects of dataset imbalance in scRNA-seq integration scenarios and its impacts on downstream analyses and overall biological conclusions. The observed effects were not method specific and, thus, have implications for single-cell data integration overall. Plausible biological conclusions under the same conditions may not be concordant if the pre-integration data distributions have variation in dataset balance. These results have ramifications for single-cell data integration, as most datasets being integrated will likely not have a high degree of shared variation with the increasing complexity of the tissues being analyzed and higher throughput of single-cell sequencing protocols[26,34].

Although single-cell data integration is ubiquitous in current computational analysis pipelines for both scRNA-seq and multi-modal single-cell sequencing data, analysis of the nuanced behavior and properties of integration techniques has lagged. Extensive benchmarking studies have been performed for scRNA-seq integration, but these studies have largely focused on performance in certain datasets, as determined by batch mixing and conservation of biological heterogeneity[4,7]. Some studies have raised concerns about the impacts of dataset imbalance, and a few methods have been developed specifically to address this challenge[6,35–38], but an extensive analysis on downstream effects had yet to be performed. Given the results of the present study, future approaches to integration benchmarking should factor in non-trivial cases with imbalance prevalent, such as tumor samples from multiple patients and cohorts. Further understanding of the properties of both the pre-integration and post-integration representation spaces will likely shed light on the situational tradeoffs between batch mixing and conservation of biological heterogeneity.

Our analysis is limited by the extent of datasets analyzed and methods tested. We included frequently used and best-performing scRNA-seq integration techniques based on previous benchmarking studies[4,7], but we tested only one method that focuses specifically on preserving biological heterogeneity when differing cell populations are present, in CIDER[37]. Future method-based benchmarking studies should exhaustively feature techniques that have sought to explicitly address the issue of dataset imbalance and several datasets with a high degree of imbalance present. Lastly, this analysis focused on scRNA-seq integration and did not incorporate multi-modal datasets and techniques, and, although extrapolation may be possible, this must be confirmed by future work addressing imbalanced integration in jointly and separately profiled multi-modal datasets.

## Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information,

acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41587-023-02097-9.

## References


1. Argelaguet, R. et al. Multi-omics profiling of mouse gastrulation at single-cell resolution. *Nature* **576**, 487–491 (2019).
2. Chiou, J. et al. Interpreting type 1 diabetes risk with genetics and single-cell epigenomics. *Nature* **594**, 398–402 (2021).
3. Pijuan-Sala, B. et al. A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature* **566**, 490–495 (2019).
4. Tran, H. T. N. et al. A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biol.* **21**, 12 (2020).
5. Amezquita, R. A. et al. Orchestrating single-cell analysis with Bioconductor. *Nat. Methods* **17**, 137–145 (2020).
6. Ming, J. et al. FIRM: flexible integration of single-cell RNA-sequencing data for large-scale multi-tissue cell atlas datasets. *Brief. Bioinform.* **23**, bbac167 (2022).
7. Luecken, M. D. et al. Benchmarking atlas-level data integration in single-cell genomics. *Nat. Methods* **19**, 41–50 (2022).
8. Zheng, G. X. et al. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).
9. 10x Genomics. 8k PBMCs from a healthy donor, single cell gene expression dataset by Cell Ranger 2.1.0. https://www.10xgenomics.com/resources/datasets/8-k-pbm-cs-from-a-healthy-donor-2-standard-2-1-0 (2017).
10. Cao, J. et al. The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566**, 496–502 (2019).
11. Abdelaal, T. et al. A comparison of automatic cell identification methods for single-cell RNA sequencing data. *Genome Biol.* **20**, 194 (2019).
12. Clarke, Z. A. et al. Tutorial: guidelines for annotating single-cell transcriptomic maps using automated and manual methods. *Nat. Protoc.* **16**, 2749–2764 (2021).
13. Saelens, W., Cannoodt, R., Todorov, H. & Saeys, Y. A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.* **37**, 547–554 (2019).
14. Ding, J. et al. Systematic comparison of single-cell and single-nucleus RNA-sequencing methods. *Nat. Biotechnol.* **38**, 737–746 (2020).
15. Vladoiu, M. C. et al. Childhood cerebellar tumours mirror conserved fetal transcriptional programs. *Nature* **572**, 67–73 (2019).
16. Peng, J. et al. Single-cell RNA-seq highlights intra-tumoral heterogeneity and malignant progression in pancreatic ductal adenocarcinoma. *Cell Res.* **29**, 725–738 (2019).
17. Polański, K. et al. BBKNN: fast batch alignment of single cell transcriptomes. *Bioinformatics* **36**, 964–965 (2020).
18. Korsunsky, I. et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).
19. Hie, B., Bryson, B. & Berger, B. Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. *Nat. Biotechnol.* **37**, 685–691 (2019).
20. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nat. Methods* **15**, 1053–1058 (2018).
21. Stuart, T. et al. Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902 (2019).
22. Buitinck, L. et al. API design for machine learning software: experiences from the scikit-learn project. Preprint at https://doi.org/10.48550/arXiv.1309.0238 (2013).
23. Goutte, C. & Gaussier, E. A probabilistic interpretation of precision, recall and *F*-score, with implication for evaluation. In *Advances in Information Retrieval* 345–359. https://doi.org/10.1007/978-3-540-31865-1_25 (Springer, 2005).
24. Luecken, M. D. & Theis, F. J. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol.* **15**, e8746 (2019).
25. Hao, Y. et al. Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587 (2021).
26. Svensson, V., da Veiga Beltrame, E. & Pachter, L. A curated database reveals trends in single-cell transcriptomics. *Database (Oxford)* **2020**, baaa073 (2020).
27. Dohmen, J. et al. Identifying tumor cells at the single-cell level using machine learning. *Genome Biol.* **23**, 123 (2022).
28. Trinh, M. K. et al. Precise identification of cancer cells from allelic imbalances in single cell transcriptomes. *Commun. Biol.* **5**, 884 (2022).
29. Xu, Y., Liu, J., Nipper, M. & Wang, P. Ductal vs. acinar? Recent insights into identifying cell lineage of pancreatic ductal adenocarcinoma. *Ann. Pancreat. Cancer* **2**, 11 (2019).
30. Backx, E. et al. On the origin of pancreatic cancer: molecular tumor subtypes in perspective of exocrine cell plasticity. *Cell Mol. Gastroenterol. Hepatol.* **13**, 1243–1253 (2022).
31. Hubert, L. & Arabie, P. Comparing partitions. *J. Classif.* **2**, 193–218 (1985).
32. Vinh, N. X., Epps, J. & Bailey, J. Information theoretic measures for clusterings comparison: variants, properties, normalization and correction for chance. *J. Mach. Learn. Res.* **11**, 2837–2854 (2010).
33. Argelaguet, R., Cuomo, A. S., Stegle, O. & Marioni, J. C. Computational principles and challenges in single-cell data integration. *Nat. Biotechnol.* **39**, 1202–1215 (2021).
34. Ogbeide, S., Giannese, F., Mincarelli, L. & Macaulay, I. C. Into the multiverse: advances in single-cell multiomic profiling. *Trends Genet.* **38**, 831–843 (2022).
35. Andreatta, M. & Carmona, S. J. STACAS: sub-type anchor correction for alignment in Seurat to integrate single-cell RNA-seq data. *Bioinformatics* **37**, 882–884 (2021).
36. Johansen, N. & Quon, G. ScAlign: a tool for alignment, integration, and rare cell identification from scRNA-seq data. *Genome Biol.* **20**, 166 (2019).
37. Hu, Z., Ahmed, A. A. & Yau, C. CIDER: an interpretable meta-clustering framework for single-cell RNA-seq data integration and evaluation. *Genome Biol.* **22**, 337 (2021).
38. Demetçi, P., Santorella, R., Sandstede, B. & Singh, R. Unsupervised integration of single-cell multi-omics datasets with disproportionate cell-type representation. Preprint at *bioRxiv* https://doi.org/10.1101/2021.11.09.467903 (2022).

# Methods

## Dataset pre-processing

**Pre-processing and normalization.** All datasets used in the study were pre-processed using a uniform pipeline. Datasets would only be further processed if it was clear that no filtering was done on the raw scRNA-seq data, such as removal of low-quality cells and genes[5]. However, these steps had been completed for all of the datasets tested, and, as such, no cell or gene-level filtering was performed. As normalization and log transformation need to be tuned specific to the method being used and are done in the integration pipeline where necessary, these were not done in the pre-processing steps.

Datasets were split and saved as individual batches in h5ad format, and the scanpy library (version 1.8.2)[39] was used for all further downstream processing within the integration pipeline, including total count per cell normalization, log1p transformation and highly variable gene selection[39]. These steps were carried out uniformly for each method tested, with the exception of scVI, as the technique must use the raw counts[20]. Therefore, total count normalization and log1p transformation were not done for scVI.

Most scRNA-seq datasets used were pre-processed in this manner, including the balanced two-batch PBMC dataset[4,8,9], imbalanced two-batch PBMC dataset[14], four-batch PBMC dataset[14], six-batch mouse hindbrain development dataset[15] and two-batch mammalian organogenesis dataset[10]. The eight-batch PDAC dataset[16] used a separate pipeline for quality control and annotation to ensure that the annotations were not biased to a specific integration method (see Methods section "Setting up the PDAC dataset").

The ground truth annotations for cell types in each dataset across batches were determined specific to the annotation protocol followed by each original study, with the exception of the eight-batch PDAC data, which were re-annotated (see Methods section "Setting up the PDAC dataset").

**Setting up the PBMC control dataset.** For testing in a scenario where the cell types and cell type proportions are perfectly balanced between batches and subsequent perturbation experiments, a dataset that was pre-processed by Tran et al.[4] was used. This dataset comprised two batches of PBMCs sequenced using two variants of 10x Genomics protocols: 5′ versus 3′ end. As these technologies capture different regions of mRNA, there is an expected batch effect present. To create a balanced dataset, the two batches were downsampled for cell types that had at least 200 cells in each batch, leaving B cells, $CD14^+$ monocytes, $CD4^+$ T cells, $CD8^+$ T cells, $FCGR3A^+$ monocytes and NK cells. Within each batch, these remaining cell types were randomly downsampled to 200 cells, leading to a perfectly balanced control setup for perturbation experiments.

**Setting up the mammalian organogenesis control dataset.** As the PBMC dataset does not have a differentiation trajectory, for testing the stability of trajectory inference in imbalanced scenarios, a dataset of mammalian organogenesis from Cao et al.[10] was used. The quantified gene-level counts and cell annotations were accessed from Gene Expression Omnibus record GSE119945, and cells from the mesenchymal trajectory (Main trajectory = 'Mesenchymal trajectory', from the study metadata) were selected. The nuclei extraction dates (day 1 or day 5) were used to denote the batch (batch 1 and batch 2, respectively). To create a balanced scenario, cell types that had at least 1,000 cells in both batches were selected, excluding stromal cells as they are not part of the differentiation trajectory. This left the following cell types in the mesenchymal trajectory: Early mesenchyme, Intermediate mesoderm, Limb mesenchyme, Myocytes, Jaw and tooth progenitors and Chondrocyte progenitors. From here, in both batches, the selected cell types were randomly downsampled such that there are 1,000 of each in both batches, creating a balanced scenario similar to the PBMC data.

**Setting up the PDAC dataset.** The PDAC dataset was taken from the Peng et al.[16] multi-patient study, which comprised 24 samples. Data for this cohort are available from the Genome Sequence Archive of the National Genomics Data Center (NGDC), accession number CRA001160. For these data, annotations for both tumor and normal cells were done in the following manner.

Cell Ranger outputs for each Peng et al. sample were downloaded from the Genome Sequence Archive of the NGDC. The barcode files, feature files and read matrices were read into R and used to create a SingleCellExperiment object[5]. Quality control was performed to remove low-quality cells with fewer than 500 detected genes, fewer than 1,000 total counts across genes and more than 20% mitochondrial genome content (per total RNA counts). After single-cell quality control, DoubletFinder[40] was used to find and remove doublet cells in the dataset.

SingleR[41] was used to perform cell type annotation for the filtered dataset. SingleR leverages a user-provided labeled reference to infer the origin of single cells in a query dataset based on similarity to the reference. SingleR labels each of the cells in the query dataset independently, providing a more fine-grained annotation compared to cluster-based methods and not biasing the results toward any integration techniques. A multi-cohort PDAC single-cell atlas[42] was used as the reference to annotate the Peng et al. dataset samples. Pancreatic acinar cells, pancreatic ductal cells and other microenvironment cells were labeled by SingleR in the dataset. The acinar cells and ductal cells were grouped and renamed 'pancreatic epithelial cells'.

InferCNV[43] was run to estimate copy number variation (CNV) mutations in the pancreatic epithelial cells, and those cells identified as having CNV alterations were annotated as tumor cells. Normal tissue samples from Peng et al.[16], Steele et al.[44] and Chen et al.[45] were processed, quality controlled and SingleR annotated. This process was the same as outlined above for the complete Peng et al. samples. In total, 2,000 epithelial cells from the normal tissue samples were randomly selected as the 'normal' reference for InferCNV. Pancreatic epithelial cells from each Peng et al. tumor sample were combined with the normal epithelial reference independently, such that InferCNV was run on each tumor sample independently. Denoising and hidden Markov model (HMM) subcluster mode were activated for InferCNV.

After annotation of the tumor cells for the Peng et al. samples, the tumor epithelial cells were collapsed into the 'Epithelial tumor' compartment and the non-tumor epithelial cells into the 'Epithelial normal' compartment. The rest of the cells were collapsed into the 'Microenvironment' compartment. Samples were filtered based on the presence of at least 50 cells in each of the three compartments (Epithelial normal, Epithelial tumor and Microenvironment). From these remaining samples, eight were randomly selected for analysis, and these corresponded to the following IDs from the NGDC archive: CRR034500, CRR034501, CRR034506, CRR034510, CRR034511, CRR034516, CRR034519 and CRR241798. Both the compartment-based and full cell type annotations were retained for experiments and analyses.

## scRNA-seq integration methods and parameters

Five state-of-the-art scRNA-seq integration methods were used, based on their performance in previous benchmarking studies[4,7], including BBKNN (version 1.5.1)[17], Harmony (Python implementation, version 0.0.5)[18], scVI (scvi-tools, version 0.14.4)[20], Scanorama (version 1.7.1)[19] and Seurat (version 4.0.6)[21]. LIGER (version 0.5.0)[46] was also originally tested but did not indicate strong performance and resulted in a high degree of variability due to the removal of seeding in different steps. Therefore, the results from LIGER were omitted from the findings, as the high variance of results even within the control experiments did not allow for a statistically sound comparison between control and perturbation groups.

To get a more clear sense of variability in control experiments, seeding mechanisms within each method were removed. This included removing any calls in the method source code to R-based seeding for

Seurat and any calls to seeding from the following libraries for BBKNN, Harmony, Scanorama and scVI: random, numpy and torch. This led to a more true estimation of the variability in performance of each method in control experiments, leading to a more reliable estimation of the effects of perturbation because the variability due to seeding differences will be captured in the control experiments before comparison.

With the exception of scVI, each method used the same processing pipeline for the data, where scanpy's functions[39] were used in the following manner:

1. Count normalization for each cell to a total value of $1 \times 10^4$
2. Transformation of counts using the $\log(1 + x)$ function
3. Highly variable gene (HVG) selection using the Seurat method for 2,500 genes
4. Integration at this step for Seurat—return-corrected HVG counts
5. Principal component analysis (PCA) reduction to top 50 PCs that explain the highest variance for HVG counts
6. Integration at this step for Harmony and Scanorama—return-corrected PCs
7. Creating neighborhood graph using the embedding with 20 dimensions (highest explained variance) and 15 nearest neighbors—integration at this step for BBKNN (replaces neighborhood graph step in scanpy pipeline)
8. Leiden clustering on the neighborhood graph using scanpy's default parameters
9. Uniform manifold approximation and projection (UMAP)[47] on the neighborhood graph using scanpy's default parameters

The only exception to this setup was scVI, which requires raw scRNA-seq expression counts[20] and used the entire set of genes for each dataset and the raw counts. For scVI, steps 1–6 outlined are omitted, and it simply integrates the raw data and returns a 10-dimensional embedding, which replaces the reduced dimensions of the other methods (input embedding for step 7). Ten dimensions were used in this case instead of 20, as this was the indicated default setting for scVI. The rest of the steps (7–9) are the same.

BBKNN performs integration on the embedding neighborhood representation[17], and, as a result, many of the downstream analyses that required embeddings did not have data for BBKNN, as it was untestable. Default parameters were used for all methods to ensure fairness in across-method comparisons as well as comparisons before and after perturbations.

**CIDER.** To assess the ability of an integration method designed specifically to address cell population imbalance, CIDER[37] was tested in a supplementary analysis. Integration with CIDER returns a list of corrected cluster labels instead of a low-dimensional representation or corrected counts matrix[37]. The high-level de novo (dn) CIDER pipeline was used to return corrected cluster labels for the control and perturbed PBMC two-batch balanced data (see Methods sections "Setting up the PBMC control dataset" and "Balanced two-batch PBMC data"). The exact steps indicated in the CIDER R implementation vignette for dnCIDER were followed (https://zhiyhu.github.io/CIDER/articles/dnCIDER_highlevel.html#perform-dncider-high-level) on the raw counts for the two-batch PBMC dataset perturbation experiments (see Methods section "Balanced two-batch PBMC data"). The corrected cluster labels were used in a supplementary analysis that modified the KNN classification setup to use cluster labels for all methods (see Methods section "Cluster-based KNN classification supplementary experiment").

## Perturbation experiments

Perturbation experiments were carried out in three settings: the balanced two-batch PBMC data, the imbalanced eight-batch PDAC data and the balanced two-batch mammalian organogenesis data. In all three instances, batches were randomly selected to be perturbed as well as given cell types/compartments. Three types of perturbation experiments were performed: control, downsampling and ablation. Control experiments did not downsample any data but allowed for replicates of integration runs across methods to get a sense of intra-method and inter-method variance on the data without perturbation. Downsampling experiments involved randomly selecting cells of a selected cell type across the indicated number of batches and downsampling to 10% of the original cell type population. Ablation experiments involved completely removing selected cell types from the indicated number of batches. Randomness of selection for the batches, cell types and cells within indicated cell types was ensured through randomly generated numbers for each perturbation experiment/run. To determine the effects of perturbation, results from the control experiments were compared to results from downsampling and ablation experiments, across all methods and selected datasets.

**Balanced two-batch PBMC data.** Within the balanced two-batch PBMC dataset, perturbation experiments were performed for one of two batches (randomly selected) at a time and for one cell type at a time. Four hundred runs were done for the control experiments, and 200 runs were done for the downsampling and ablation experiments (resulting in 400 total for perturbation and 800 total when adding perturbation and control experiments), ensuring that both batches ($n = 2$) and each cell type ($n = 6$) are sampled repeatedly and method performance variance within control experiments is taken into account adequately. Within the hierarchical setup, where similar cell types were hierarchically clustered into three groups (B cells, monocytes and NK/T cells), the same number of runs was done for the control, downsampling and ablation experiments, with the groups replacing the cell types. This resulted in 1,600 total integration experiments.

**Eight-batch PDAC data.** For the eight-batch PDAC dataset, where cells were grouped into three major compartments, four batches were randomly selected for downsampling or ablation, and one compartment was downsampled or ablated within each non-control replicate. In total, 100 control runs and 50 downsampling and ablation runs were performed, as the number of compartments is small ($n = 3$), and there will be adequate sampling and repetition within 50 runs. This resulted in 200 integration experiments.

**Balanced two-batch organogenesis data.** For the mesenchymal trajectory inference analysis in the mammalian organogenesis dataset, a similar setup to that of the PBMC data was used: perturbation experiments were performed for one randomly selected batch and randomly selected cell type at one time, and 400 runs were done for the control experiments and 200 for the downsampling and ablation experiments (leading to 400 total for the perturbation experiments). The stability of trajectory inference was assessed between the control and downsampling/ablation runs. This resulted in 800 integration experiments.

## Benchmarking integration performance—PBMC two-batch control dataset

Performance in integration and downstream tasks was assessed using the integrated embeddings and Leiden clustering[48] results from each integration technique. After integration, the neighborhood calculation step was done, and Leiden clustering was performed after this, both using the scanpy library[39] (see Methods section "scRNA-seq integration methods and parameters"). Only BBKNN did not result in embeddings to be used as it performs integration at the neighborhood level and, therefore, was not included in the KNN classification experiments, as these relied on integrated embeddings.

**Quantifying cell type conservation and batch mixing with clustering metrics.** The ARI[31] was calculated for all integration experiments. The sklearn implementation of the ARI was used. Details of the ARI can be found in the scikit learn documentation[22]. ARI values were

calculated by comparing the known annotated labels with the cluster labels obtained after integration for each technique. Both cell type and (1 − batch) values were calculated, where cell type metrics compared the known cell type annotations with the cluster labels to determine how well the integrated embeddings corresponded to known cell type labels, and the (1 − batch) values used the batch annotations and the cluster labels to determine how well the different batches co-aggregate in the embeddings. The assumption of the latter is that integration should lead to strong batch mixing, and the shadow of the $ARI_{batch}$ value was used ($1 - ARI_{batch}$) to reflect this desired property. The median value across all experiments for a given combination of {method, experiment type, downsampled cell type} was determined.

**z-score normalization of ARI metrics.** To determine intra-method variation based on the properties of perturbations versus the control experiments, the median values for $ARI_{cell\text{-}type}$ and ($1 - ARI_{batch}$), for all combinations of {method, experiment type, downsampled cell type}, were *z*-score normalized. For example, for $ARI_{cell\text{-}type}$ values for a specific subset:

$$\frac{\text{Median ARI}_{\{\text{method}_x,\text{type}_y,\text{cell-type}_z\}} - \mu\left(\text{Median ARI}_{\{\text{method}_\cdot,\text{type}_\cdot,\text{cell-type}_\cdot\}}\right)}{\sigma\left(\text{Median ARI}_{\{\text{method}_\cdot,\text{type}_\cdot,\text{cell-type}_\cdot\}}\right)} \quad (1)$$

The exact same procedure is followed for the ($1 - ARI_{batch}$) values.

**Downstream analysis—unsupervised clustering.** In this downstream analysis test after integration, the number of unsupervised Leiden clusters was determined and compared between the different perturbation experiments and control groups. The number of clusters was determined using Leiden clustering in the scanpy library[39]. As each method resulted in different Leiden clusters, these results were analyzed independently, and intra-method and experiment type comparisons were performed.

**Optimal Leiden resolution.** A variant of the analysis for the clustering stability experiments included a case where the unsupervised clustering resolution was changed based on the control experiments, such that the number of clusters corresponded to the number of ground truth cell types, specific to each method. For the unsupervised clustering analysis main result (Fig. 3a), this variant was used, and the unconstrained version that used the default Leiden resolution across experiments (resolution = 1) can be found in the supplementary analyses (Supplementary Fig. 15). Increasing the Leiden resolution resulted in a higher number of clusters, whereas decreasing it had the opposite effect. The following steps were done to determine the optimal resolution for each method:

1. A Leiden clustering resolution was sampled from the following values:

$$0.1, 0.25, 0.5, 0.75, 1.0, 2.0, 3.0, 4.0, 5.0$$

2. In the control data (no downsampling/ablation), integration was performed with each method, and the integrated space was used for unsupervised clustering with the sampled Leiden resolution.
3. Using the clustering results, for each method, the difference between the number of clusters and ground truth cell types was determined ($\Delta C$ = |number of clusters − number of cell types|).

The above procedure was done for each resolution and repeated 50 times (iterations) to factor in variance in the integration results for the methods. Within each iteration ($i$), the optimal resolution was determined for each method ($m$), across tested resolutions ($r$):

$$\text{Optimal resolution}_{im}(O^{im}) = \underset{r}{\text{argmin}}\ \Delta\mathbf{C}_{imr}$$

Then, using these values, the median optimal resolution was determined for each method across the iterations:

$$\text{Median optimal resolution}_m = \text{Med}(\mathbf{O}^{im})$$

This median optimal resolution was used for Leiden clustering in both the control and downsampling/ablation experiments, and the effects of perturbations (downsampling/ablation) on the number of clusters was assessed using this setup.

**Downstream analysis—KNN classification.** The goal of this downstream analysis test was to determine the performance of integration techniques at a per-cell-type level before and after perturbation. After obtaining the integrated embeddings for all methods, with the exception of BBKNN, a KNN classifier was trained on a 70/30 training/test split of the integrated embeddings to predict the cell type labels of the test data. Stratified sampling was used for the split to ensure that all cell types were represented in the same proportions between train and test sets. The sklearn (version ≥1.0.1) library was used for the data preparation, test/train split, stratified sampling, KNN classifier training and prediction[22]. The explicit formulation for prediction of a class on a test data point $x_i$ is:

$$\text{class } x_i = \max_{y \in \mathbb{Y}} \sum_{y_i \in \mathbb{N}_{x_i}}^{k} \delta(y_i, y) \quad (2)$$

$$\delta(y_i, y) = \begin{cases} 1, & \text{if } y_i = y \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

where $k$ indicates the number of neighbors used in the classifier, which was set to 15 for all runs. As different runs could possibly lead to different test/train splits of the integrated embedding, a seed was used to ensure that the same split occurs across experimental groups. Using the results of the predictions, the F1 classification scores were determined, and the F1 score specific to each cell type was used as the primary metric. For comparison, cases were examined where a specific cell type was downsampled or ablated, and the effects of performance on the same cell type based on the KNN classification F1 score were analyzed. The form of the score is given by ref. 23:

$$\text{F1 score per cell-type} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \quad (4)$$

where $TP$ is the number of true-positive calls, $FP$ is the number of false-positive calls, and $FN$ is the number of false-negative calls, all on a per-cell-type basis.

*Cluster-based KNN classification supplementary experiment*. For a supplementary analysis that involved testing the CIDER method[37], a modified version of the KNN classification approach was used, because, like BBKNN, CIDER does not return a corrected low-dimensional space or corrected counts. Instead, CIDER performs correction on the cluster labels obtained by clustering within each batch[37]. The exact same data split, KNN classifier and F1 score evaluation approach were used, and the only difference in this analysis was that the KNN graph and classifier were not trained and tested on the integrated embeddings but on the one-hot encoded representation of the cluster labels[22]. For methods aside from CIDER, the cluster labels from Leiden clustering in the scanpy pipeline after integration were used (see Methods section "scRNA-seq integration methods and parameters"). Given that the corrected cluster labels were used and not an integrated low-dimensional representation, BBKNN was included in this analysis. As stratification is done on cell type labels between the train and test sets, there is no guarantee that all cluster labels from test will be encoded in train and

vice versa. The one-hot encoder was fit on the train set, and the 'handle_unknown' parameter was set to 'ignore', such that labels in the test set that were not in train were encoded as an array of zeros.

**Downstream analysis—marker gene ranking.** DGE analysis is typically performed after clustering to determine marker genes specific to each cluster that are then used to annotate cells within those clusters[5]. The goal of this analysis was to determine to what extent can the results of DGE be altered after perturbation of dataset balance.

First, the top 10 marker genes corresponding to each cell type were determined in each batch for each dataset (for example, two-batch PBMC dataset) using the Wilcoxon rank-sum test and the scanpy package sc.tl.rank_genes_groups[39]. To ensure selection of relevant markers, ribosomal and mitochondrial genes were removed from the pool of tested genes. After this, to obtain a consensus on the marker genes across batches, the union of markers for each cell type across batches was determined. This will lead to an uneven number of markers for some cell types if completely distinct sets are called from different batches, but it leads to a more complete set as the integrated space across batches is being analyzed. This set of markers for each cell type in a dataset was deemed the 'master marker list'.

From here, after the integration step using each method in each control or perturbation experiment for a given dataset, unsupervised clustering using the Leiden clustering algorithm with the scanpy default parameters was done for the integrated embedding. Then, DGE using the Wilcoxon rank-sum test was performed for each of the obtained unsupervised clusters[39]. A challenge here is that there is no correspondence between the unsupervised clusters obtained in this integrated embedding and the cell types used for determining the master marker list. Therefore, we did DGE for each cluster and checked the 'maximum ranking' of a given marker gene across all clusters. Ranking is defined by how significant the DGE *P* value is for a given gene, where the highest rank is the most statistically significant differentially expressed gene for a given cluster. If a cluster still corresponds to a given cell type (which is the central assumption in unsupervised integration), then that cluster should return a high ranking for a given marker gene corresponding to that cell type in DGE. Therefore, for the markers in the master marker list, we analyzed their maximum ranking across unsupervised clusters in the integrated space, to see if biological information specific to that cell type and its markers is still being retained after integration.

The change in ranking for all of the marker genes corresponding to the different known cell types in datasets was quantified by their standard deviation in a given subset of experiments (control, downsampling or ablation) and termed the 'change in maximum ranking'. This change in maximum ranking within an experiment group (for example, the control group) was indicated as the 'marker gene perturbation score':

$$\text{Marker gene perturbation score} = \sigma(\text{max marker gene ranking}) \quad (5)$$

If this value is being averaged over many genes (for example, for a cell type), this is indicated as the 'average marker gene perturbation score'. If there are $m$ marker genes for a given cell type:

$$\text{Avg. marker perturbation score} = \frac{1}{m}\sum_{i=1}^{m}\text{Marker } i \text{ perturbation score} \quad (6)$$

A supplementary analysis variant of this procedure using the top 50 marker genes per cell type instead of the top 10 was also completed, with the aforementioned steps remaining the same.

In a different supplementary analysis, the Leiden clustering step before the DGE step was changed such that it corresponded to the optimal resolution based on concordance with number of ground truth cell types, as outlined in Methods section "Optimal Leiden resolution". The other aforementioned steps remained the same.

*Case study—CD4$^+$/CD8$^+$ T cell assignment based on marker genes.* To determine if the changes in marker gene ranking that were observed could realistically influence the results of a single-cell analysis, the same marker gene perturbation setup was used. In this case, however, each of the unsupervised clusters after integration were annotated as specific cell types based on the majority of cells present (for example, Cluster 1 − > Majority CD4$^+$ T cells − > CD4$^+$ T cells). In this setup, only clusters that contained a majority of either CD4$^+$ or CD8$^+$ T cells were used. For simplifying the case study, only integration results from the Seurat method were used.

After integration and selection of clusters with a majority of CD4$^+$ and CD8$^+$ T cells, differential expression analysis was performed as previously indicated, and a permissive threshold of the top 50 marker genes was used to select markers for the CD4$^+$ and CD8$^+$ majority clusters. From here, each of the CD4$^+$ and CD8$^+$ T cell majority clusters were predicted to be either CD4$^+$ or CD8$^+$ based on the presence of canonical marker genes: *IL7R* for CD4$^+$ T cells and *CD8A* for CD8$^+$ T cells[21].

Examining the top 50 marker genes for each cluster, the rules for predicting the cell types in each of the CD4$^+$ and CD8$^+$ T majority clusters were the following:

```
if IL7R and CD8A are present, then
  if Rank (IL7R) > Rank (CD8A), then
  Annotate as CD4+ T
  else if Rank (CD8A) > Rank (IL7R), then
  Annotate as CD8+ T
  end if
else if IL7R present, then
  Annotate as CD4+ T
else if CD8A present, then
  Annotate as CD8+ T
else
  Annotate as Undefined
end if
```

The fraction of unsupervised clusters that contained a majority of CD4$^+$ and CD8$^+$ T cells was predicted for their cell types based on differential expression in control and perturbation (downsampling and ablation) experiments. Only downsampling and ablation experiments that affected CD4$^+$ and CD8$^+$ T cells were analyzed, as downsampling/ablating these cell types was found to most likely affect the marker gene rankings of either cell type.

**Downstream analysis—query-to-reference annotation.** To test the robustness of query-to-reference annotation techniques across varying degrees of unshared variation, the Seurat 4.0 multi-modal projection technique was used[25]. Although the control PBMC two-batch dataset has only scRNA-seq information, a multi-modal reference can still be used as is, and only the RNA sequencing modality will be integrated. The reference dataset used is from Hao et al.[25], and the same parameters indicated in the vignette (https://satijalab.org/seurat/articles/multimodal_reference_mapping.html) were used.

It is important to note that integration was not performed before projection using the Seurat 4.0 method. Instead, each batch/sample was individually projected/integrated with the reference dataset and annotated, as per the guidelines for Seurat 4.0 (ref. 25). Therefore, there are no method-specific comparisons to be made in this analysis.

As the annotations in the reference will not exactly match the annotations from the PBMC two-batch data (due to a higher degree of granularity and different naming conventions)[4,25], a scoring guide was created to determine if the annotation correctly matches the ground truth cell type label by using 'fuzzy matching' of ground truth cell type

labels from the PBMC two-batch dataset and the labels in the reference data. The following table summarizes the guide for the PBMC two-batch data and acceptable annotations for L1 (coarse-grained label from Hao et al.[25]) and L2 (fine-grained label from Hao et al.[25]):

| Ground truth label | Acceptable L1 reference | Acceptable L2 reference |
|---|---|---|
| CD4 T cell | CD4 T | CD4 TCM, CD4 naive, CD4 CTL, CD4 proliferating, CD4 TEM |
| CD8 T cell | CD8 T | CD8 naive, CD8 TEM, CD8 TCM, CD8 proliferating |
| Monocyte_CD14 | Mono | CD14 mono |
| Monocyte_FCGR3A | Mono | CD16 mono |
| NK cell | NK | NK, NK proliferating, NK_CD56bright |

Using this annotation guide, the annotation accuracy through the F1 score (equation (4)) was determined for each experiment, across experiment types (control, downsampling and ablation). This value was calculated using both the L1 and L2 annotations.

### Trajectory inference in the balanced mammalian organogensis data

As the PBMC dataset did not contain a differentation trajectory, the Cao et al. organogenesis dataset[10] was used with a balanced setup for the mesenchymal trajectory comprising two batches (see Methods section "Setting up the mammalian organogenesis control dataset"). The batch effect in this data is minimal, and preserving the fine-grained differentation trajectory when doing trajectory inference is key. Given this, and the fact that a ground truth trajectory with pseudotime values for each cell is unknown, the following approach was taken to assess the stability of trajectory inference estimates before and after perturbation:

1. PAGA and diffusion pseudotime were used jointly to estimate pseudotime values for both the unintegrated and integrated representations.
2. The correlation between the pseudotime values for the cells between the unintegrated data and the integrated data (using each method) was determined.
3. The change in correlation values between the control and downsampled/ablated runs was determined, based on the cell type downsampled and method used.

Before running PAGA and diffusion pseudotime through scanpy, the pre-processing steps and subsequent integration for each method were the same as those outlined in the integration setup (See Methods section "scRNA-seq integration methods and parameters").

PAGA[49] and diffusion pseudotime[50] were used through scanpy[39], with default parameters and 'celltype' as the groups (sc.tl.paga). The diffusion pseudotime was estimated (sc.tl.dpt) with the known root cell type as the assumed root (Early mesenchyme).

For all experiment types (control, downsampling and ablation) and each iteration, the Spearman correlation for pseudotime values determined using the unintegrated representation and the integration representations was determined for each method. The rationale behind this comparison is that, because the batch effect is minimal and the unintegrated representation will return a reasonable pseudotime estimate, the pseudotime estimates for the integrated representations should be well correlated with the unintegrated results. Otherwise, this would be indicative of loss of biological information, likely due to overcorrection. This is also a key factor in these data because there are very fine-grained developmental gradients present, and preserving biological signal by not collapsing distinct cell types and cell states together is crucial.

### Visualization pipelines

UMAP visualizations were done for a number of datasets and experiments.

The main figure UMAPs for the balanced PBMC two-batch dataset and the supplementary UMAP figures for the Cao et al. organogenesis were created using Seurat and the following pipeline:

1. log normalization of the data
2. Highly variable gene selection with the 'vst' method and 2,000 features
3. Scaling and PCA reduction
4. Neighborhood construction using the first 20 PCA dimensions
5. Leiden clustering using the neighborhood graph using a resolution of 0.5
6. UMAP reduction of the first 20 PCA dimensions

The supplementary figures for the balanced PBMC two-batch perturbation experiments were generated using the exact pipeline outlined in Methods section "scRNA-seq integration methods and parameters" and the calculated UMAP coordinates.

### Complex imbalanced dataset analysis

After quantifying the effects of unshared variation in the control two-batch PBMC dataset through perturbation experiments, complex datasets that are multi-batch and already imbalanced were analyzed, including: imbalanced two-batch PBMC dataset, imbalanced four-batch PBMC dataset, six-batch mouse hindbrain development dataset and eight-batch PDAC dataset.

Experiments involving these datasets did not factor in perturbations, but replicates were performed to assess method-based variability. For each dataset, a total of 50 experiments (without any perturbation to dataset balance) were performed.

**Cell type center distances.** To determine the distance between cell types in the embedding space used for integration, across all batches to be integrated, the following pre-processmethoding steps were performed on the raw data for each batch in a dataset:

1. Count normalization for each cell to a total value of $1 \times 10^4$
2. Transformation of counts using the $\log(1+x)$ function
3. Highly variable gene selection using the Seurat method for 2,500 genes
4. PCA for the top 20 PCs on the counts data

After obtaining the PCs for a given dataset, the ground truth cell type labels are used to determine the cell type center distance between all cell types in the data in a pairwise manner. The cell type center distance is defined as the weighted cosine distance between the center (average) of the PCA representation of one cell type with another.

For each batch $b$ and cell type $a$ with $n$ cells and a PCA reduction of the data:

$$\mathrm{PC}_{a_b} \in \mathbb{R}^{n \times 20} \tag{7}$$

$$\text{Cell-type } a_b \text{ center} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{PC}_{a_b i} \in \mathbb{R}^{1 \times 20} \tag{8}$$

Then, for quantifying the distance between cell types $a$ and $c$ in batch $b$:

Let $\mathbf{v} \in \mathbb{R}^{1 \times 20}$ be the variance explained by each of the top 20 PCs

Let $\mathbf{CC}_{a_b}$ be the cell-type center for cell-type $a$ in batch $b$

Let $\mathbf{CC}_{c_b}$ be the cell-type center for cell-type $c$ in batch $b$

$$\text{Cell-type center distance } ac_b = 1 - \frac{(\mathbf{CC}_{a_b} \circ \mathbf{v}) \cdot (\mathbf{CC}_{c_b} \circ \mathbf{v})}{||(\mathbf{CC}_{a_b} \circ \mathbf{v})||||(\mathbf{CC}_{c_b} \circ \mathbf{v})||} \tag{9}$$

where $\mathbf{CC}_{a_b} \circ \mathbf{v}$ is the element-wise rescaling of the center of cell type $a$ based on the variance explained by the PCs.

The rationale behind a reweighted cosine distance is that the distance itself between cell types should be scaled according to the variance explained by each PC because the distance is being calculated in the joint PCA reduction of all cells, and not every PC axis will have equal contribution for the variance explained.

We can take the average of this cell type center distance across $p$ batches:

$$\text{Avg. cell-type center distance} = \frac{1}{p}\sum_{b=1}^{p} \text{Cell-type center distance}\, ac_b \tag{10}$$

The minimum cell type center distance, or the distance corresponding to the cell type closest to cell type $a$, is simply the minimum value across all batches ($p$ total). Assume there are $k$ total cell types across all batches, missing cell type pairs (for example, cell type present in batch 1 and not batch 2) have an imputed maximum cosine distance of 1. Values between the same cell types are also imputed as 1. Then, using the tensor of cell type center distances across batches $D$:

$$D \in \mathbb{R}^{k \times k \times p} \tag{11}$$

$$\text{Minimum cell-type center distance}\, a = \min(D_{a,:,:}) \tag{12}$$

where $D_a$ is the subset of the first axis for cell type $a$. The cell type and batch corresponding to this value can be found through the argmin.

The minimum distance in any batch is taken instead of averaging distances across all batches because this minimum distance will correspond to the most 'haphazard' scenario for a given batch being integrated. Two scenarios are possible here:

1. The distance between cell types is largely similar across batches, and the minimum value will correspond roughly to the average.
2. The distance between cell types can be very different across batches, due to scenarios/factors such as developmental data or treatment effects.

The first case is most readily applicable to the PBMC datasets, but the second scenario may be more applicable to the PDAC and hindbrain developmental data. However, even in these cases, taking the minimum may lead to a better approximation of proximity affecting integration results because it will factor in the worst possible scenario (across batches) for a given cell type.

**Aggregate cell type support.** The cell type support (or aggregate cell type support) is the $\log_2$ transformation of the number of cells for cell type $a$ across all batches $b$:

$$\text{Aggregate cell-type support}\, a = \log_2\left(\sum_{b=1}^{p} \text{Cell-type}\, a_b\right) \tag{13}$$

## Statistical testing

**One-way ANOVA tests.** To determine statistical significance for the effects of perturbations, the following generic one-way ANOVA setup was used[51]:

$$\text{response} \sim x_0 + x_1 + x_2 + \ldots + x_m + \text{condition} \tag{14}$$

$$\mathcal{H}_0 : \text{response} = x_0 + x_1 + x_2 + \ldots + x_m \tag{15}$$

where $\mathcal{H}_0$ is the null hypothesis; response can be an endpoint of interest in the analysis (for example, number of clusters after integration); $x_0$ is a constant (intercept/bias); $x_1, \ldots, x_m$ are factors that we would like to control before testing significance with respect to condition (for example, method and cell type that was downsampled); and "condition" is a binary covariate indicating the experiment type that was done:

$$\text{condition} = \begin{cases} 1, & \text{if } y_i = \text{downsampling, ablation} \\ 0, & \text{if } y_i = \text{control} \end{cases} \tag{16}$$

For simplification, downsampling and ablation are grouped into the 'perturbed' experiment type, and the control experiments are indicated as 'unperturbed'. This convention is followed in Supplementary Table 2.

After accounting for the various factors to control ($x_1, \ldots, x_m$), we can assess the statistical significance of perturbation of unshared variation (condition) with respect to the response covariate through the ANOVA $F$-statistic value and $P$ value associated with the condition covariate.

In situations where significance is achieved across various groups, the magnitudes of the $F$-statistic values were compared.

## Control PBMC two-batch dataset

**KNN classification per cell type.** For assessing the effects of perturbation on the F1 classification scores after integration on a per-cell-type level, the following ANOVA (Methods section "One-way ANOVA tests") setup was used":

$$\text{F1 classification score} \sim x_0 + \text{method} + \text{downsampled cell-type} + \text{condition} \tag{17}$$

$$\mathcal{H}_0 : \text{F1 classification score} = x_0 + \text{method} + \text{downsampled cell-type} \tag{18}$$

The F1 classification scores here are across all cell types in the integrated dataset. The cell type being analyzed (for the F1 classification score in each instance) is equivalent to the downsampled cell type in each sample included in the test.

**Unsupervised clustering.** For comparing the significance of perturbation on the number of unsupervised clusters obtained after integration using Leiden clustering, the following ANOVA setup was used:

$$n\,\text{clusters} \sim x_0 + \text{method} + \text{downsampled cell-type} + \text{condition} \tag{19}$$

$$\mathcal{H}_0 : n\,\text{clusters} = x_0 + \text{method} + \text{downsampled cell-type} \tag{20}$$

**Marker gene ranking.** To test the statistical significance of perturbations for each marker gene analyzed, the following ANOVA setup was used for each marker gene $g$:

$$\text{Marker}\, g\, \text{max rank} \sim x_0 + \text{method} + \text{downsampled cell-type} + \text{condition} \tag{21}$$

$$\mathcal{H}_0 : \text{Marker}\, g\, \text{max rank} = x_0 + \text{method} + \text{downsampled cell-type} \tag{22}$$

Then, to test the overall effects on marker gene ranking, considering all marker genes at once, the following test was done:

$$\text{Marker max rank} \sim x_0 + \text{gene} + \text{method} + \text{downsampled cell-type} + \text{condition} \tag{23}$$

$$\mathcal{H}_0 : \text{Marker max rank} = x_0 + \text{gene} + \text{method} + \text{downsampled cell-type} \tag{24}$$

### Control mammalian organogenesis dataset

**Trajectory inference.** To test the significance of perturbation effects on the correlation between the unintegrated pseudotime estimates and the integrated pseudotime estimates, the following ANOVA setup was used:

$$\text{Spearman correlation} \sim x_0 + \text{method} + \text{downsampled cell-type} + \text{condition} \quad (25)$$

$$\mathcal{H}_0 : \text{Spearman correlation} = x_0 + \text{method} + \text{downsampled cell-type} \quad (26)$$

### Complex imbalanced datasets

**Cell type support and cell type center distance.** To determine if the two key metrics that were determined in the complex dataset analysis—aggregate cell type support (Methods section "Aggregate cell type support") and cell type center distance (Methods section "Cell type center distances")—are in fact predictive of integration performance, the following ANOVA setups were used where the F1 classification score for each experiment, cell and associated ground truth cell type was tested:

$$\text{F1 classification score} \sim x_0 + \text{method} + \text{minimum cell-type center distance} \quad (27)$$

$$\mathcal{H}_0 : \text{F1 classification score} = x_0 + \text{method} \quad (28)$$

$$\text{F1 classification score} \sim x_0 + \text{method} + \text{aggregate cell-type support} \quad (29)$$

$$\mathcal{H}_0 : \text{F1 classification score} = x_0 + \text{method} \quad (30)$$

The 'cell type analyzed' was not included as a factor to control, because the 'minimum cell type center distance' and 'aggregate cell type support' metrics were calculated on a per-cell-type basis. Therefore, these metrics are perfectly collinear with cell type, and this factor would absorb the residuals that would be picked up by the key metrics.

**PDAC perturbation analysis.** Perturbations were performed for the compartmentalized PDAC data (Methods sections "Setting up the PDAC dataset" and "Eight-batch PDAC data") to determine the effects of downsampling/ablation on the classification scores of all compartments. Here, the following ANOVA setup was used to determine the effects on F1 scores for a specific compartment based on downsampling of the same compartment for each compartment $c$:

$$\text{F1 classification score } c \sim x_0 + \text{method} + \text{condition} \quad (31)$$

$$\mathcal{H}_0 : \text{F1 classification score } c = x_0 + \text{method} \quad (32)$$

These results were analyzed independently and jointly for all compartments downsampled, where joint comparison included comparison of $F$-statistic values for perturbation in each setup.

Results of all statistical tests can be found in Supplementary Table 2.

### Balanced clustering scores

None of the used clustering metrics in this analysis and other integration benchmarking/methods papers factor in class balance. These metrics include the ARI, AMI, Homogeneity Score and Completeness Score. The implementation details of these metrics can be found in the scikit learn documentation[22].

Strictly speaking, the Homogeneity Score and Completeness Scores are not metrics, because they are not symmetric. However, this symmetry is not necessary in the case of single-cell benchmarking, and in the general case of comparing clustering labels with ground truth annotations, because one set of labels is known to be ground truth. In fact, balancing the ARI and AMI will break their symmetry as well.

To introduce the procedure behind reweighing these metrics, we will begin with the balanced ARI. Then, we will extrapolate this procedure to the entropy-based metrics/scores (AMI, Homogeneity and Completeness), as this extrapolation involves only a slight modification to the ARI procedure.

Code notebooks on implementing the balanced clustering scores with usage demonstrations and relevant examples are available at https://github.com/hsmaan/balanced-clustering.

### The bARI

**The Rand Index and ARI.** For a set of $n$ objects, $S = \{O_1, O_2, O_3, \ldots, O_n\}$, the goal of clustering is to partition these objects into meaningful subsets, which we can call partitioning $V$. Assume we have access to either ground truth labels or clusters from another technique, which we can denote partitioning $U$. Both $U$ and $V$ contain subsets, which we call either classes or clusters: $U = \{u_1, u_2, \ldots, u_R\}$ and $V = \{v_1, v_2, \ldots, v_C\}$. These clustering results are subject to the following constraints to be valid for calculating the Rand Index (RI):

1. All $n$ objects within the set $S$ must be within sets $U$ and $V$:

$$U_{i=1}^{R} u_i = U_{j=1}^{C} v_j = S \quad (33)$$

2. No element from set $S$ can belong to more than one subset in either $U$ or $V$

$$1 \leq i \neq i' \leq R \quad (34)$$

$$1 \leq j \neq j' \leq C \quad (35)$$

$$u_i \cap u_{i'} = \varnothing = v_j \cap v_{j'} \quad (36)$$

To quantify the overlap between the partitions $U$ and $V$ (for example, in determining overlap between a set of ground truth labels and the results of clustering), we can start by creating a contingency table that indicates the overlap:

| | $v_1$ | $v_2$ | $v_3$ | ... | $v_c$ | |
|---|---|---|---|---|---|---|
| $u_1$ | $t_{11}$ | $t_{12}$ | $t_{13}$ | ... | $t_{1C}$ | $t_{1.}$ |
| $u_2$ | $t_{21}$ | $t_{22}$ | ... | ... | ... | ... |
| $u_3$ | $t_{31}$ | ... | ... | ... | ... | ... |
| ... | ... | ... | ... | ... | ... | ... |
| $u_R$ | $t_{R1}$ | ... | ... | ... | $t_{RC}$ | $t_{R.}$ |
| | $t_{.1}$ | ... | ... | ... | $t_{.C}$ | $t_{..}$ |

Each element of this table indicates overlapping elements. For example, $t_{11}$ indicates the number of samples that have the label $v_1$ in $V$ and $u_1$ in $U$. The total number of values in the matrix is $\binom{n}{2}$ if $n$ objects/samples are present. As we now have a table/matrix that represents the overlap of assignments to subsets in $U$ and $V$ for $n$ objects, we can determine the concordance of partitions $U$ and $V$ for these objects using the RI:

$$a = \sum_{r=1}^{R}\sum_{c=1}^{C}\binom{t_{rc}}{2}, \binom{x}{2} = 0 \text{ if } x = 0 \quad (37)$$

$$b = \left[\sum_{r=1}^{R}\binom{t_{r.}}{2}\right] - a \tag{38}$$

$$c = \left[\sum_{c=1}^{C}\binom{t_{.c}}{2}\right] - a \tag{39}$$

$$d = \binom{n}{2} - a - b - c \tag{40}$$

$$\mathrm{RI} = \frac{a+d}{a+b+c+d} \tag{41}$$

Intuitively, the RI aims to calculate how many pairs are concordantly in the same subsets in $V$ and $U$ ($a$), how many pairs are concordantly in different subsets in $V$ and $U$ and how many are discordant (in the same group in one partition and otherwise in the other). It is important to note that pairs here refer to all combinations of two different objects, not the same object being considered in the two partitions.

Although the RI is normalized (lower bound = 0, upper bound = 1), it is not adjusted for chance clustering. A correction can be made[31] to the RI formula that takes into account the expected value of the RI for two partitions of the objects $U$ and $V$, denoted by the ARI[31]:

$$\mathrm{ARI} = \frac{\binom{n}{2}(a+d) - [(a+b)(a+c)+(c+d)(b+d)]}{\binom{n}{2}^2 - [(a+b)(a+c)+(c+d)(b+d)]} \tag{42}$$

With this correction, the ARI is a metric (symmetric, positive-definite and the triangle inequality)[32] that has the properties of normalization and expectation[31].

**Balancing the ARI.** Rebalancing the ARI (as well as the other entropy-based scores/metrics) will amount to rescaling the total number of values in the subsets of the partition that we consider ground truth. In this case, assume $U$ is the partition with the ground truth information. We want each subset from $U$ to have an equal contribution to the ARI value; this is concomitant with each class from the ground truth data (which we have assumed to be $U$) having an equal weighing in the calculation of the score. This can be done in the following stepwise manner:

1. Determine contribution of each subset of $U(t_1, t_2, \ldots, t_R)$ to the score through the mean of marginals from the contingency table:

$$(t_{1.}, t_{2.}, t_{3.}, \ldots, t_{R.}) \tag{43}$$

2. Get the mean contributions of all subsets:

$$C = \frac{1}{R}\sum_{i=1}^{R} t_{i.} \tag{44}$$

3. For each subset ($t_i$) of $U$, normalize the contribution to be equal to the mean using a scaling factor:

$$S_i = \frac{C}{t_{i.}} \tag{45}$$

$$\forall t_i, (t_{i1}, t_{i2}, \ldots, t_{iC}) = S_i \times (t_{i1}, t_{i2}, \ldots, t_{iC}) \tag{46}$$

After these steps, we have essentially rescaled the contingency table such that the contribution from each subset in $U$ will be considered equally in calculations using the table results. To calculate the bARI, we can apply this normalization procedure, and use the same ARI formula as before (equation (42)), with values for $a$, $b$ and $c$ obtained through the rescaled contingency table:

$$\text{Balanced ARI} = \frac{\binom{n}{2}(a+d) - [(a+b)(a+c)+(c+d)(b+d)]}{\binom{n}{2}^2 - [(a+b)(a+c)+(c+d)(b+d)]} \tag{47}$$

Examining the values needed to calculate the RI and ARI (equations (37)–(42)), we can see that this normalization procedure will effectively rescale the calculations for $a$, $b$ and $c$. This procedure does this while still retaining the total counts ($n$), such that the calculation for $d$ will be unaffected. Because the RI and ARI calculations simply depend on these values in the contingency table that can be calculated independently, the application of this normalization procedure is straightforward and does not require any further steps.

## Balancing entropy-based scores

**Mutual information.** Central to the AMI, Homogeneity and Completeness scores is the calculation of mutual information between partitions $U$ and $V$[32,52]. For the contingency table previously defined in Methods section "The Rand Index and ARI", the mutual information between these two partitions is equal to the following[32]:

$$I(U,V) = \sum_{r=1}^{R}\sum_{c=1}^{C}\frac{t_{rc}}{n}\log\frac{t_{rc}/n}{t_{r.}t_{.c}/n^2} \tag{48}$$

We will follow the same normalization procedure that we did in Methods section "Balancing the ARI", as we are starting from the same contingency table of overlapping objects in subsets of partitions $U$ and $V$. From here, we can calculate the mutual information value and proceed with the rest of the calculations for the entropy-based scores.

**Entropy.** Aside from mutual information, the other important factor that is used by all of the entropy-based scores is the calculation of the entropy of the labeling—that is, how ordered/disordered are the objects in partitions $U$ and $V$. This can also be calculated from the contingency table from Methods section "The Rand Index and ARI", in the following manner (ref. 32):

$$H(U) = -\sum_{r=1}^{R}\frac{t_{r.}}{n}\log\frac{t_{r.}}{n} \tag{49}$$

$$H(V) = -\sum_{c=1}^{C}\frac{t_{.c}}{n}\log\frac{t_{.c}}{n} \tag{50}$$

The Homogeneity and Completeness scores also require the conditional entropy formulation[32,52]:

$$H(U|V) = -\sum_{r=1}^{R}\sum_{c=1}^{C}\frac{t_{rc}}{n}\log\frac{t_{rc}/n}{t_{.c}/n} \tag{51}$$

$$H(V|U) = -\sum_{r=1}^{R}\sum_{c=1}^{C}\frac{t_{rc}}{n}\log\frac{t_{rc}/n}{t_{r.}/n} \tag{52}$$

**Balanced entropy-based scores.** The calculation of the entropy and mutual information can proceed as is after the normalization procedure from Methods section "Balancing the ARI", and this will balance the contributions from a presumed ground truth partition $U$ in calculating the entropy and mutual information. From here, the bAMI, Balanced

Homogeneity and Balanced Completeness scores can be calculated using these terms based on the rescaled contingency table[32,52]:

$$\text{Balanced AMI} = \frac{I(U,V) - \mathbb{E}[I(U,V)]}{\frac{1}{2}[H(U) + H(V)] - \mathbb{E}[I(U,V)]} \quad (53)$$

$$\text{Balanced Homogeneity} = 1 - \frac{H(U|V)}{H(U)} \quad (54)$$

$$\text{Balanced Completeness} = 1 - \frac{H(V|U)}{H(V)} \quad (55)$$

The V-measure and Balanced V-measure are simply the harmonic mean of the Completeness (Compl.) and Homogeneity (Homog.) scores[52]:

$$\text{Balanced V-measure} = \frac{2 \times \text{Bal. Homog.} \times \text{Bal. Compl.}}{\text{Bal. Homog.} + \text{Bal. Compl.}} \quad (56)$$

**Balanced clustering evaluations.** The following section details the evaluations that were used for the balanced clustering metric analysis. Seeding was set for all of these cases to ensure reproducibility of the experiments, downsampling and integration results.

*Three imbalanced well-separated classes, two clusters*. In this scenario, three well-separated but imbalanced classes were used, and a mis-clustering of the smaller class was done with *k*-means clustering with $k = 2$. These data were simulated using two-dimensional (2D) Gaussian densities (with diagonal covariance and the same mean for both dimensions) with the following values for each class:

- Class A ~ $N(0, 0.5)$—500 samples
- Class B ~ $N(-2, 0.1)$—20 samples
- Class C ~ $N(3, 1)$—500 samples

*k*-means clustering with $k = 2$ led to class B overlapping with class A in the clustering result.

The balanced and imbalanced metrics were compared when calculating the concordance of the ground truth labels (class labels) and *k*-means clustering labels.

*Three imbalanced overlapping classes, three clusters*. In this case, three classes that are overlapping and imbalanced (two smaller classes on the edges of larger class) were analyzed, and *k*-means clustering with $k = 3$ was done. *k*-means correctly clustered most of the samples from the smaller classes, but, due to slicing of the larger class present because of overlap, it mis-clustered a large number of majority class samples.

These data were simulated using 2D Gaussian densities with the following values for each class:

- Class A ~ $N(0, 0.5)$—1,500 samples
- Class B ~ $N(1, 1)$—200 samples
- Class C ~ $N(-1, 1)$—200 samples

*Balanced two-batch PBMC—co-clustered CD4$^+$ T cells and CD8$^+$ T cells*. The balanced two-batch PBMC dataset was used here (see Methods section "Setting up the PBMC control dataset"). Batch 1 was kept as is, and batch 2 had all of the cells ablated except for CD4$^+$ T cells, which were downsampled to 10% of their original proportion.

The default Leiden clustering resolution of 1 in the scanpy implementation was changed to 0.1, as this value perfectly clusters all of the cell types with the exception of the CD4$^+$ T cells, which get collapsed into a cluster with CD8$^+$ T cells, simulating a case where a less prevalent cell type is co-clustered with a more prevalent cell type.

The resultant embedding with no integration was used, and the ground truth cell type labels and unsupervised clustering labels were used to compare the balanced and imbalanced scores, where the ARI and Homogeneity scores were shown.

*Balanced two-batch PBMC data—downsampled CD4$^+$ T cells and FCGR3A$^+$ monocytes*. In this evaluation, the two-batch balanced PBMC dataset was once again used. For the two batches, each one had either the CD4$^+$ T cells or FCGR3A$^+$ monocytes downsampled to 10% of their original population, creating an imbalanced scenario specific to these two cell types.

After this, integration was done using BBKNN, Harmony, Scanorama and scVI. The same integration pipeline from Methods section "scRNA-seq integration methods and parameters" was used. An 'unintegrated' control subset was used, where the pipeline from Methods section "scRNA-seq integration methods and parameters" was followed without integration with any method.

From here, the average value of the balanced and imbalanced metrics was used for comparison. For example, for imbalanced metrics:

$$\text{Avg imbalanced} = \frac{1}{5}\sum(\text{ARI,AMI,Homog.,Compl.,V-measure}).$$

### Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

## Data availability

The data necessary to reproduce the results of the study can be downloaded from Figshare: https://doi.org/10.6084/m9.figshare.24625302.v1. The raw data for the datasets analyzed in this study can be found in the following links/accessions: two-batch balanced PBMC data[8,9]: SRP073767 and https://www.10xgenomics.com/resources/datasets/8-k-pbm-cs-from-a-healthy-donor-2-standard-2-1-0; two-batch and four-batch imbalanced PBMC data[14]: GSE132044; PDAC tumor data[16]—Genome Sequencing Archive: CRA001160; mouse hindbrain developmental data[15]: GSE118068; and mammalian organogenesis data[10]: GSE119945.

## Code availability

All of the code necessary to reproduce the results of the Iniquitate pipeline is available at https://github.com/hsmaan/Iniquitate. The vignette for a walkthrough of the imbalanced integration guidelines outlined in Results section "End-user guidelines for imbalanced integration" can be found in the Iniquitate documentation (https://github.com/hsmaan/Iniquitate/tree/main/docs). The Python package for implementing the balanced clustering metrics can be found at https://github.com/hsmaan/balanced-clustering.

## References


39. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
40. McGinnis, C. S., Murrow, L. M. & Gartner, Z. J. DoubletFinder: doublet detection in single-cell RNA sequencing data using artificial nearest neighbors. *Cell Syst.* **8**, 329–337 (2019).
41. Aran, D. et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat. Immunol.* **20**, 163–172 (2019).
42. Chijimatsu, R. et al. Establishment of a reference single-cell RNA sequencing dataset for human pancreatic adenocarcinoma. *iScience* **25**, 104659 (2022).
43. Tickle, T., Tirosh, I., Georgescu, C., Brown, M. & Haas, B. Infer copy number variation from single-cell RNA-seq data. https://doi.org/doi:10.18129/B9.bioc.infercnv (2019).
44. Steele, N. G. et al. Multimodal mapping of the tumor and peripheral blood immune landscape in human pancreatic cancer. *Nat. Cancer* **1**, 1097–1112 (2020).

45. Chen, K. et al. Immune profiling and prognostic model of pancreatic cancer using quantitative pathology and single-cell RNA sequencing. *J. Transl. Med.* **21**, 210 (2023).
46. Welch, J. D. et al. Single-cell multi-omic integration compares and contrasts features of brain cell identity. *Cell* **177**, 1873–1887 (2019).
47. McInnes, L., Healy, J. & Melville, J. UMAP: uniform manifold approximation and projection for dimension reduction. Preprint at https://doi.org/10.48550/arXiv.1802.03426 (2018).
48. Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep.* **9**, 5233 (2019).
49. Wolf, F. A. et al. PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biol.* **20**, 59 (2019).
50. Haghverdi, L., Büttner, M., Wolf, F. A., Buettner, F. & Theis, F. J. Diffusion pseudotime robustly reconstructs lineage branching. *Nat. Methods* **13**, 845–848 (2016).
51. Winer, B. J., Brown, D. R. & Michels, K. M. *Statistical Principles in Experimental Design* 3rd edn (McGraw-Hill, 1991).
52. Rosenberg, A. & Hirschberg, J. V-Measure: a conditional entropy-based external cluster evaluation measure. *Proc. of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* 410–420 (Association for Computational Linguistics, 2007).

## Acknowledgements

We would like to thank members of the Bo Wang and Kieran Campbell laboratories for insightful discussion and feedback on this work. H.M. is supported by a doctoral fellowship from the Natural Sciences and Engineering Research Council of Canada (NSERC). M.G. is supported by the Princess Margaret Cancer Foundation and a Health Informatics and Data Science award from the Terry Fox Research Institute. C.Y. and M.G are supported by University of Toronto Data Science Institute doctoral fellowships. This work was supported by funding from the Canadian Institutes of Health Research (project grant PJT175270, to K.C.), funding from the NSERC (RGPIN-2020-04083, to K.C., and RGPIN-2020-06189 and DGECR-2020-00294, to B.W.), a Canada Foundation for Innovation John R. Evans Leaders Fund award (to K.C.), the CIFAR AI Chairs Program (to B.W.) and the Peter Munk Cardiac Centre AI Fund at the University Health Network (to B.W.). This research was undertaken, in part, thanks to funding from the Canada Research Chairs Program. Figures 1, 4a and 6 were created with BioRender.

## Author contributions

H.M., K.C. and B.W. conceptualized the ideas and experiments. H.M. performed the manuscript experiments and associated analysis. H.M. and L.Z. performed the statistical analyses accompanying the experiments. C.Y. and M.G. processed and annotated the PDAC data and helped with associated experiments and analysis. H.M. developed the balanced metrics and performed associated analysis. H.M. developed the guidelines for imbalanced integration. K.C. and B.W. funded the project and provided supervision. H.M. wrote the original manuscript, and all authors provided review and approved the final version.

## Competing interests

B.W. is on the Strategic Advisory Board of Vevo Therapeutics, Inc. The remaining authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s41587-023-02097-9.

**Correspondence and requests for materials** should be addressed to Hassaan Maan, Kieran R. Campbell or Bo Wang.

**Peer review information** *Nature Biotechnology* thanks the anonymous reviewers for their contribution to the peer review of this work.

nature portfolio

Corresponding author(s): Bo Wang, Kieran R Campbell

Last updated by author(s): Nov 29, 2023

# Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

## Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

| n/a | Confirmed |
|---|---|
| ☐ | ☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement |
| ☐ | ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| ☐ | ☒ The statistical test(s) used AND whether they are one- or two-sided<br>*Only common tests should be described solely by name; describe more complex techniques in the Methods section.* |
| ☐ | ☒ A description of all covariates tested |
| ☒ | ☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
| ☐ | ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| ☐ | ☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted<br>*Give P values as exact values whenever suitable.* |
| ☒ | ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| ☒ | ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
| ☐ | ☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated |

*Our web collection on statistics for biologists contains articles on many of the points above.*

## Software and code

Policy information about availability of computer code

Data collection: N/A - all datasets utilized were public.

| | |
|---|---|
| Data analysis | All of the code necessary to reproduce the results of the Iniquitate pipeline is available at https://github.com/hsmaan/Iniquitate.<br><br>The following two environments - Iniquitate and Analysis - were used in the benchmarking and analysis phases, where all integration experiments and downstream analysis tests were done with the Iniquitate environment, and all results analysis and plotting was done with the Analysis environment. The only exceptions were the balanced metric analyses and tests, which used the Analysis environment for generation and testing of the various scenarios outlined.<br><br>The Iniquitate environment is specified in YAML format at https://github.com/hsmaan/Iniquitate/blob/main/workflow/envs/integrate.yaml, and the Analysis environment is specified at https://github.com/hsmaan/Iniquitate/blob/main/workflow/envs/analysis.yaml. R and Python packages used, as well as version numbers are included in the YAML files.<br><br>The library for the balanced metrics was developed independently, and all of the information on dependency versions is available at https://github.com/hsmaan/balanced-clustering |

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

## Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The datasets utilized in this study, which are associated with the various configurations used in the Iniquitate GitHub repository, can be all be downloaded from FigShare:

https://doi.org/10.6084/m9.figshare.24625302.v1.

The raw data for the datasets analyzed in this study can be found in the following links/accessions: two batch balanced PBMC data- SRP073767 and https://www.10xgenomics.com/resources/datasets/8-k-pbm-cs-from-a-healthy-donor-2-standard-2-1-0, two batch and four batch imbalanced PBMC data - GSE132044, pancreatic ductal adenocarcinoma tumor data - GSA: CRA001160, mouse hindbrain developmental data - GSE118068, and mammalian organogenesis data - GSE119945.

## Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

| | |
|---|---|
| Reporting on sex and gender | N/A |
| Population characteristics | N/A |
| Recruitment | N/A |
| Ethics oversight | N/A |

Note that full information on the approval of the study protocol must also be provided in the manuscript.

# Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

# Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

| | |
|---|---|
| Sample size | Sample size in this case corresponds to the number of cells, cell-types and batches present in the datasets utilized in this study. To ensure a variety of datasets and possible single-cell integration scenarios were tested, both low-complexity controlled settings (e.g. the balanced 2 batch peripheral blood mononuclear cell [PBMC] dataset) and high-complexity scenarios (e.g. imbalanced PBMC datasets, mouse hindbrain developmental data) were tested. Perturbation experiments (where perturbation corresponds to altering the degree of dataset balance) were performed in controlled settings (balanced 2 batch PBMC data, balanced 2 batch mammalian organogenesis data) and a more complex setting with imbalance prevalent (8 batch PDAC dataset).<br><br>We note that all of the datasets used were public and none of the data was generated by the authors of the study, so a standard sample size calculation was not done.<br><br>Datasets, reflective of the different sample size properties indicated, were selected based on complexity of typical scRNA-seq datasets in different biological contexts. The lowest complexity settings had at least 400 cells per cell-type across batches, which is adequate for assessing variance explained by perturbations in performance of downstream tasks. |
| Data exclusions | No data was excluded from the analysis. |
| Replication | In the analysis, independent integration experiments were performed across experiment types (downsampling, ablation, control), and variance in performance across subsets was assessed. In other words, each analysis involved running perturbation and control experiments multiple times, and performance was assessed across multiple experiments.<br><br>Rerunning the Iniquitate pipeline leads to reproducible results, although derived statistics may not be exactly the same (e.g. p-values) due to different cells being affected in each experiment because of randomization.<br><br>Broader-scale results (e.g. drop in performance in KNN classification after perturbation) were reproducible across datasets (e.g. PBMC 2 batch balanced data perturbation and 8 batch PDAC data perturbation). |
| Randomization | In downsampling (perturbation) experiments, cell-types or compartments downsampled/ablated were selected at random, as were the batches that were affected. The cells in the specific cell-type/compartment and batch to downsample or ablate, were also selected at random. This was true for all perturbation experiments done. |
| Blinding | N/A - the perturbation experiments are completely randomized and the experiments and analyses are computational in nature. |

# Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

## Materials & experimental systems

| n/a | Involved in the study |
|---|---|
| ☒ | ☐ Antibodies |
| ☒ | ☐ Eukaryotic cell lines |
| ☒ | ☐ Palaeontology and archaeology |
| ☒ | ☐ Animals and other organisms |
| ☒ | ☐ Clinical data |
| ☒ | ☐ Dual use research of concern |

## Methods

| n/a | Involved in the study |
|---|---|
| ☒ | ☐ ChIP-seq |
| ☒ | ☐ Flow cytometry |
| ☒ | ☐ MRI-based neuroimaging |