nature biotechnology

Integrating single-cell transcriptomic data across different conditions, technologies, and species

Andrew Butler[1,2], Paul Hoffman[1], Peter Smibert[1], Efthymia Papalexi[1,2] & Rahul Satija[1,2]

Computational single-cell RNA-seq (scRNA-seq) methods have been successfully applied to experiments representing a single condition, technology, or species to discover and define cellular phenotypes. However, identifying subpopulations of cells that are present across multiple data sets remains challenging. Here, we introduce an analytical strategy for integrating scRNA-seq data sets based on common sources of variation, enabling the identification of shared populations across data sets and downstream comparative analysis. We apply this approach, implemented in our R toolkit Seurat (http://satijalab.org/seurat/), to align scRNA-seq data sets of peripheral blood mononuclear cells under resting and stimulated conditions, hematopoietic progenitors sequenced using two profiling technologies, and pancreatic cell 'atlases' generated from human and mouse islets. In each case, we learn distinct or transitional cell states jointly across data sets, while boosting statistical power through integrated analysis. Our approach facilitates general comparisons of scRNA-seq data sets, potentially deepening our understanding of how distinct cell states respond to perturbation, disease, and evolution.

With recent improvements in cost and throughput[1–3], and the availability of fully commercialized workflows[4], high-throughput single-cell transcriptomics has become an accessible and powerful tool for unbiased profiling of complex and heterogeneous systems. In concert with novel computational approaches, these data sets can be used for the discovery of cell types and states[5,6], the reconstruction of developmental trajectories and fate decisions[7,8], and to spatially model complex tissues[9,10]. Indeed, scRNA-seq is poised to transform our understanding of developmental biology and gene regulation[11–14], and enable systematic reconstruction of cellular taxonomies across the human body[6,15], although substantial computational obstacles remain. In particular, integrated analysis of different scRNA-seq data sets, consisting of multiple transcriptomic subpopulations, either to compare heterogeneous tissues across different conditions or to integrate measurements produced by different technologies, remains challenging.

Many powerful methods address individual components of this problem. For example, zero-inflated differential expression tests have been tailored to scRNA-seq data to identify changes within a single-cell type[16,17], and clustering approaches[18–23] can detect proportional shifts across conditions if cell types are conserved. However, comparative analysis for scRNA-seq poses a unique challenge, as it is difficult to distinguish between changes in the proportional composition of cell types in a sample and expression changes within a given cell type, and simultaneous analysis of multiple data sets will confound these two disparate effects. Therefore, new methods are needed that can jointly analyze multiple data sets and facilitate comparative analysis downstream. Progress toward this goal is essential for translating the oncoming wealth of single-cell sequencing data into biological insight. An integrated computational framework for joint learning between data sets would allow for robust and insightful comparisons of heterogeneous tissues in health and disease, integration of data from diverse technologies, and comparison of single-cell data from different species.

Here, we present a novel computational strategy for integrated analysis of scRNA-seq data sets, motivated by techniques in computer vision designed for the alignment and integration of imaging data sets[24,25]. We demonstrate that multivariate methods designed for 'manifold alignment'[26,27] can be successfully applied to scRNA-seq data to identify gene–gene correlation patterns that are conserved across data sets and can embed cells in a shared low-dimensional space. We identify and compare 13 aligned peripheral blood mononuclear cell (PBMC) subpopulations under resting and interferon β (IFN-β)-stimulated conditions, align scRNA-seq data sets of complex tissues produced across multiple technologies, and jointly discover shared cell types from droplet-based 'atlases' of human and mouse pancreatic tissue. These analyses pose distinct challenges for alignment, but in each case, we successfully integrate the data sets to obtain deeper biological insight than would be possible from independent analysis. Our approach can be applied to data sets ranging from hundreds to tens of thousands of cells, is compatible with diverse profiling technologies, and is implemented as part of Seurat, an open-source R toolkit for single-cell genomics.

RESULTS

Overview of Seurat alignment workflow

We aimed to develop a diverse integration strategy that could compare scRNA-seq data sets across different conditions, technologies, or species. To be successful in diverse settings, this computational strategy must fulfill the following requirements, as illustrated with a toy example where heterogeneous scRNA-seq data sets are generated in the presence or absence of a drug (**Fig. 1a**). First, subpopulations must be aligned even if each has a unique drug response. This

[1]New York Genome Center, New York, New York, USA. [2]Center for Genomics and Systems Biology, New York University, New York, New York, USA. Correspondence should be addressed to R.S. (rsatija@nygenome.org).

Received 4 July 2017; accepted 9 February 2018; published online 2 April 2018; doi:10.1038/nbt.4096

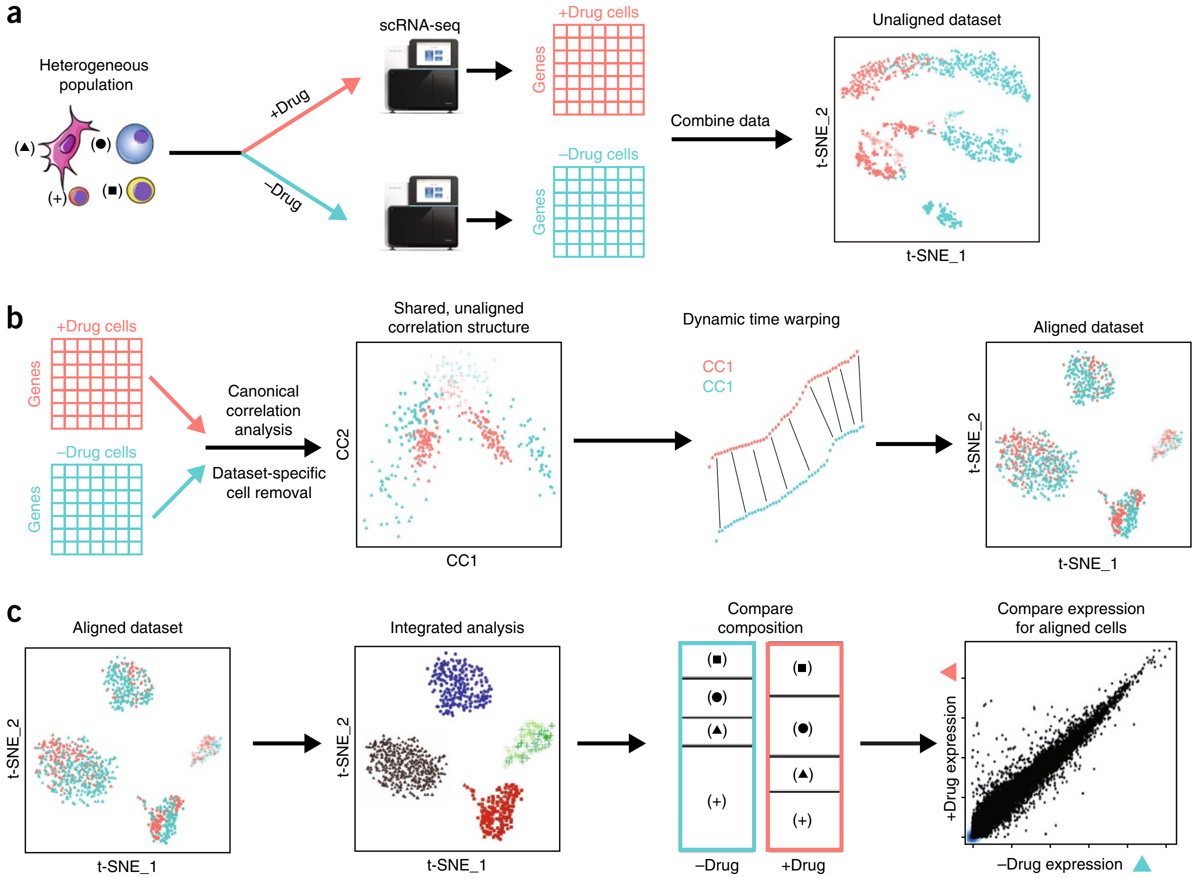

Figure 1 Overview of Seurat alignment of single-cell RNA-seq data sets. (**a**) Heterogeneous populations profiled in a case–control study after drug treatment. Four cell types are represented by different symbols, while drug treatment is encoded by color. In a standard workflow, cells often cluster both by cell type and drug treatment, creating challenges for downstream comparative analysis. (**b**) The Seurat alignment procedure uses canonical correlation analysis to identify shared correlation structures (i.e., canonical correlation vectors, CC) across data sets, and aligns these dimensions using dynamic time warping. After alignment, cells are embedded in a shared low-dimensional space (visualized here in 2D with t-SNE). (**c**) After alignment, a single integrated clustering can identify conserved cell types across conditions, allowing for comparative analysis to identify shifts in cell type proportion, as well as cell-type-specific transcriptional responses to drug treatment.

key challenge lies outside of the scope of batch correction methods developed for bulk assays, which assume that confounding variables have uniform effects on all cells in a data set. Second, the method must allow for changes in cellular density (shifts in subpopulation frequency) between conditions. Third, the method must be robust to changes in feature scale across conditions, allowing either global transcriptional shifts, or differences in normalization strategies between data sets produced with different technologies. Lastly, the process should not be targeted toward defined cell subsets, with no requirement for pre-established sets of markers that can be used to match subpopulations.

The Seurat alignment workflow takes as input a list of at least two scRNA-seq data sets, and briefly consists of the following steps (**Fig. 1b**,**c**). (i) It learns a shared gene correlation structure that is conserved between the data sets using canonical correlation analysis (CCA) (**Fig. 1b**). (ii) As an optional step, it identifies individual cells that cannot be well described by this shared structure. This can help to identify rare populations that may be non-overlapping between the data sets and can therefore be flagged for further analysis. (iii) It aligns the data sets into a conserved low-dimensional space, using nonlinear 'warping' algorithms to normalize for differences in feature scale, in a manner that is robust to shifts in populations density. (iv) It proceeds with an integrated downstream analysis, for example, identifying discrete subpopulations through clustering, or reconstructing continuous developmental processes (**Fig. 1c**). (v) It performs comparative analysis on aligned subpopulations between the data sets, to identify changes in population density or gene expression (**Fig. 1c**). We describe these steps briefly below, and then apply and validate this strategy on five sets of scRNA-seq experiments from the literature.

Identifying shared correlation structures across data sets

Machine-learning techniques for 'data fusion' aim to integrate information from multiple experiments into a consistent representation. For example, CCA aims to find linear combinations of features across data sets that are maximally correlated, identifying shared correlation structures across data sets[28,29]. CCA has been used for multimodal genomic analysis from bulk samples, for example, identifying relationships between gene expression and DNA copy number measurements based on the same set of samples[30]. Here, in contrast to its traditional use in multimodal analysis[31,32], we apply CCA to identify relationships between single cells from different data sets based on the same set of genes. Effectively, we treat the data sets as multiple measurements of a gene–gene covariance structure, and search for patterns that are common to the data sets. We use CCA for pairwise integration of two data sets, and extend this to multi-set CCA (multi-CCA)[33,34] for the integration of multiple data sets. In the description of all methods below, we refer only to CCA for simplicity, but note that each of the individual techniques can extend to multi-CCA when multiple data sets are included as input (Online Methods).

We employ a variant of CCA, diagonal CCA, to account for cases where there are more cells than genes and apply this using the single-cell RNA-seq data sets as input[30] (Online Methods). The procedure can consider any gene that is measured across all data sets, though we choose to focus only on genes that exhibit high single-cell variation in at least one data set (Online Methods). CCA identifies sets of canonical 'basis' vectors, embedding cells from each data set in a low-dimensional space, such that the variation along these vectors (gene-level projections) is highly correlated between data sets. We note that CCA is robust to affine transformations in the original data, and is unaffected by linear shifts in gene expression (e.g., due to different normalization strategies).

Aligning basis vectors from CCA

CCA returns vectors whose gene-level projections are correlated between data sets, but not necessarily aligned. While linear transformations may be required to correct for global shifts in feature scale or normalization strategy, nonlinear shifts may also be needed to correct for shifts in population density. We therefore align the CCA basis vectors between the data sets, resulting in a single, integrated low-dimensional space. Briefly, we represent each basis vector as a metagene, defined as a weighted expression average of the top genes whose expression exhibits robust correlation with the basis vector (Online Methods). We first linearly transform the metagenes to match their 95% reference range (Online Methods), correcting for global differences in feature scale. Next, we determine a mapping between the metagenes using dynamic time warping, which locally compresses or stretches the vectors during alignment to correct for changes in population density[35]. We apply this procedure to each pair (or set, for multiple alignment) of basis vectors individually, defining a single, aligned, low-dimensional space representing all data sets. This enables us to perform integrated downstream analyses, including unbiased clustering and the reconstruction of developmental trajectories, as demonstrated below.

Comparative analysis of stimulated and resting PBMCs

We first demonstrate our alignment strategy on a data set containing many distinct cell types in the presence and absence of perturbation. A recent study split 14,039 human PBMCs from eight patients into two groups: one stimulated with interferon-beta (IFN-β) and a culture-matched control[36], and performed droplet-based single-cell RNA-seq. Since all cells contain machinery to respond to IFN-β, stimulation results in a drastic but highly cell-type-specific response. Consequently, a standard analysis of both data sets together yielded confusing results, as cells tend to cluster both by cell type but also by stimulation condition (**Fig. 2a**). As an alternative to unbiased clustering, a supervised strategy to assign cells to classes based on known markers resulted in a final set of eight clusters[36].

In contrast, the Seurat alignment returned a set of canonical correlation vectors that separated PBMC subsets irrespective of stimulation condition. We chose to include 20 vector pairs for downstream analysis (Online Methods), but results for this and all examples in the manuscript were robust to the exact choice of this parameter (**Supplementary Fig. 1**). We performed joint graph-based clustering on these aligned vectors and visualized the results with t-distributed stochastic neighbor embedding (t-SNE) to verify that cells grouped entirely by cell type and were properly aligned across conditions (**Fig. 2b**). Our analysis revealed 13 cell clusters, which included the eight immune subsets described in the original publication, but separated additional populations as well (**Fig. 2c** and **Supplementary Data 1**). In particular, we were able to separate naive from memory CD4$^+$ T cells, plasmacytoid dendritic cells (pDCs) from conventional dendritic cells (DCs), and to identify an extremely rare (0.4%) population of contaminating erythroblasts. In addition, for T cells and B cells, we discovered activated subpopulations marked by a strong stress response expression signature that is likely an artifact of the culturing process in both conditions (**Supplementary Fig. 2a**,**b**). We verified the identity of our clusters by examining the expression of canonical cell-type markers (i.e., *CD3D* for T cells, *CD79A* for B cells), that were conserved across conditions (**Fig. 2d** and **Supplementary Fig. 3**).

Having aligned the data sets, we next sought to compare how PBMCs vary in response to IFN-β. As both conditions were drawn from the same pool of cells, we observed a strikingly similar proportional representation of all clusters in stimulated and control experiments (R = 0.997; **Fig. 2e**). However, each cell type exhibited significant gene expression changes upon IFN-β stimulation. Applying single-cell differential expression tests separately for each cluster, we were able to identify constitutive markers of the IFN-β response induced in all cells (*ISG15*, *IFIT1*), as well as components of the IFN-β response that varied across cell types (i.e., *CXCL10* was activated primarily in myeloid cells upon stimulation) (**Fig. 2d**). We noted that even canonical cell-type markers such as *CD14* were differentially expressed by monocytes (1.98-fold downregulation; **Supplementary Fig. 3**) in response to stimulation, highlighting the value of our unsupervised analyses in initially classifying cells.

Focusing on the novel subsets we were able to resolve, we compared the IFN-β response program between naive and memory CD4$^+$ T cells and observed nearly identical response signatures (**Supplementary Fig. 4a**). However, while we observed a general correlation between pDC and DC responses, we also saw stark differences that reproduced across patients (**Fig. 2f**). When comparing the IFN-β responses across all cell types, we observed that myeloid and lymphoid cells strongly clustered together, but pDC exhibited a distinct response to IFN-β and clustered separately (**Fig. 2g** and **Supplementary Fig. 4b**).

We validated these findings externally by replicating the setup and stimulation of the original experiment, sorting populations of pDC and DC using standard surface markers (Online Methods), and performing bulk RNA-seq experiments in triplicate on stimulated and control cells. These bulk experiments strongly confirmed our single-cell predictions: genes that were differentially regulated by IFN-β stimulation exhibited strikingly similar patterns in both the single-cell and bulk data sets, and the bulk samples clustered directly with *in silico*-averaged data from the same cell type (**Supplementary Fig. 5**). Therefore, in a single transcriptome-wide analysis, our alignment procedure sensitively identified shared cell states through integrated clustering, and allowed for the identification of cell-type-specific response modules that are likely to play important roles *in vivo* during immune response to infection.

Strategies to identify non-overlapping populations

In the previous example, identical cell populations were used as input for both populations, and the cell subpopulations should therefore be fully overlapping. We wished to assess how our integration procedure would perform when non-overlapping populations were present in only one of the data sets. This is an important concern for both abundant populations, where absence in one data set could throw off the

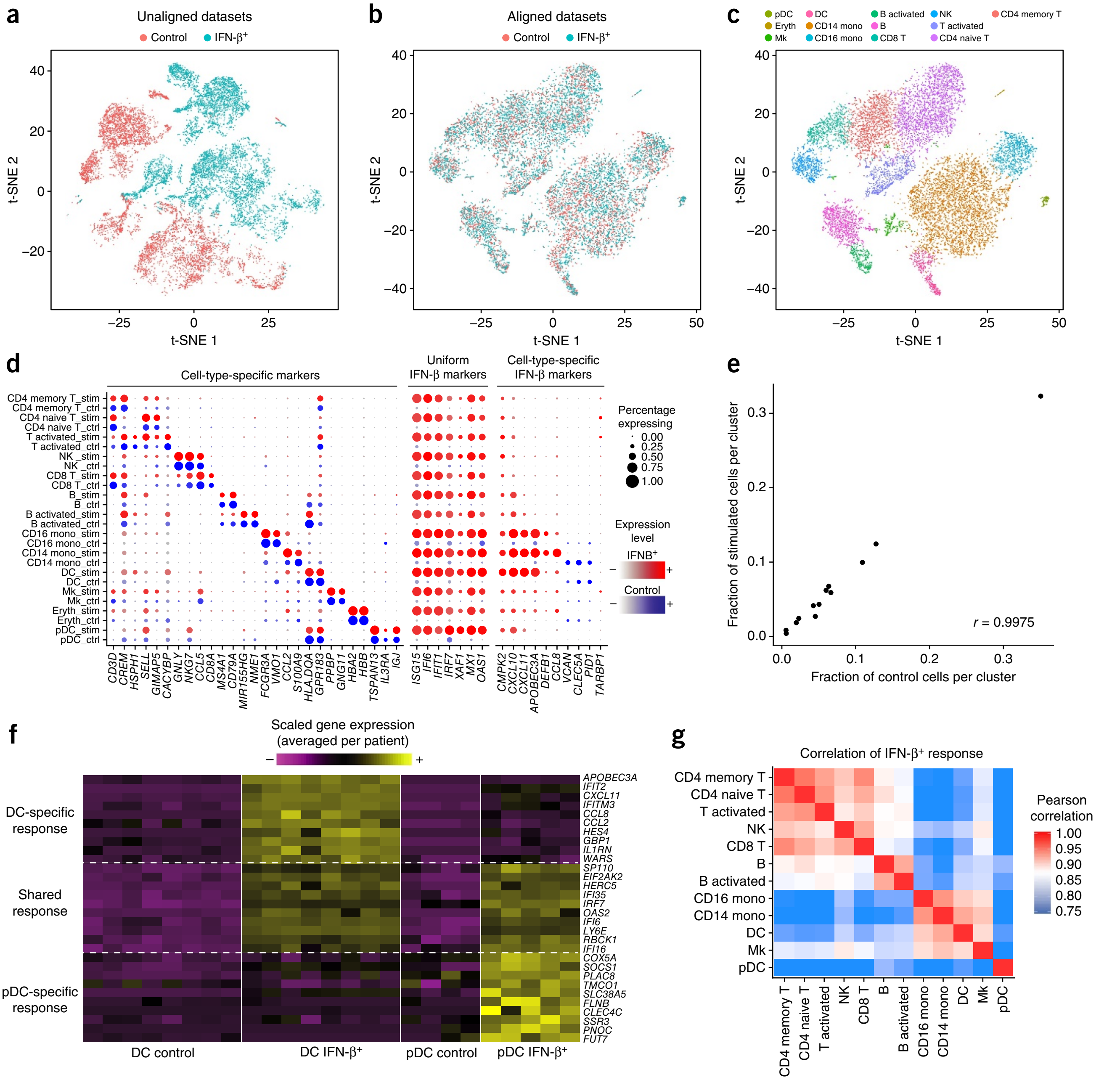

Figure 2 Integrated analysis of resting and stimulated PBMCs. (**a**–**c**) t-SNE plots of 14,039 human PBMCs split between control and IFN-β-stimulated conditions, before (**a**) and after (**b**) alignment. After alignment, cells across stimulation conditions group together based on shared cell type, allowing for a single joint clustering (**c**) to detect 13 immune populations. (**d**) Integrated analysis reveals markers of cell types (conserved across stimulation conditions), uniform markers of IFN-β response independent of cell type, and components of the IFN-β response that vary across cell types. The size of each circle reflects the percentage of cells in a cluster where the gene is detected, and the color intensity reflects the average expression level within each cluster. (**e**) The fraction of cells (median across eight donors) falling in each cluster (n = 13 clusters) for stimulated and unstimulated cells. (**f**) Examples of heterogeneous responses to IFN-β in conventional and plasmacytoid dendritic cells (global analysis shown in **Supplementary Fig. 4b**). Each column represents the average gene expression of single cells within a single patient. Only patient–cluster combinations with at least five cells are shown. (**g**) Correlation heatmap (n = 430 genes) of cell-type-specific responses to IFN-β (individual correlations for T cell and DC subsets shown in **Supplementary Fig. 4a**,**b**). Cells from myeloid and lymphoid lineages show highly correlated responses, but plasmacytoid dendritic cells (pDC) exhibit a unique IFN-β response. NK, natural killer cells; Mk, megakaryocytes.

integration, and rare populations, which could blend in with an abundant cluster if unmatched, but may have significant biological importance.

To address this, we performed two *in silico* experiments, where we artificially removed abundant (CD14$^+$ and CD16$^+$ monocytes; 38%), or rare (erythroblasts; 0.5%) cells from the stimulated data set only and repeated the alignment procedure. When we removed abundant populations, we observed negligible effects on the overall clustering and both CD14$^+$ and CD16$^+$ control monocytes were readily

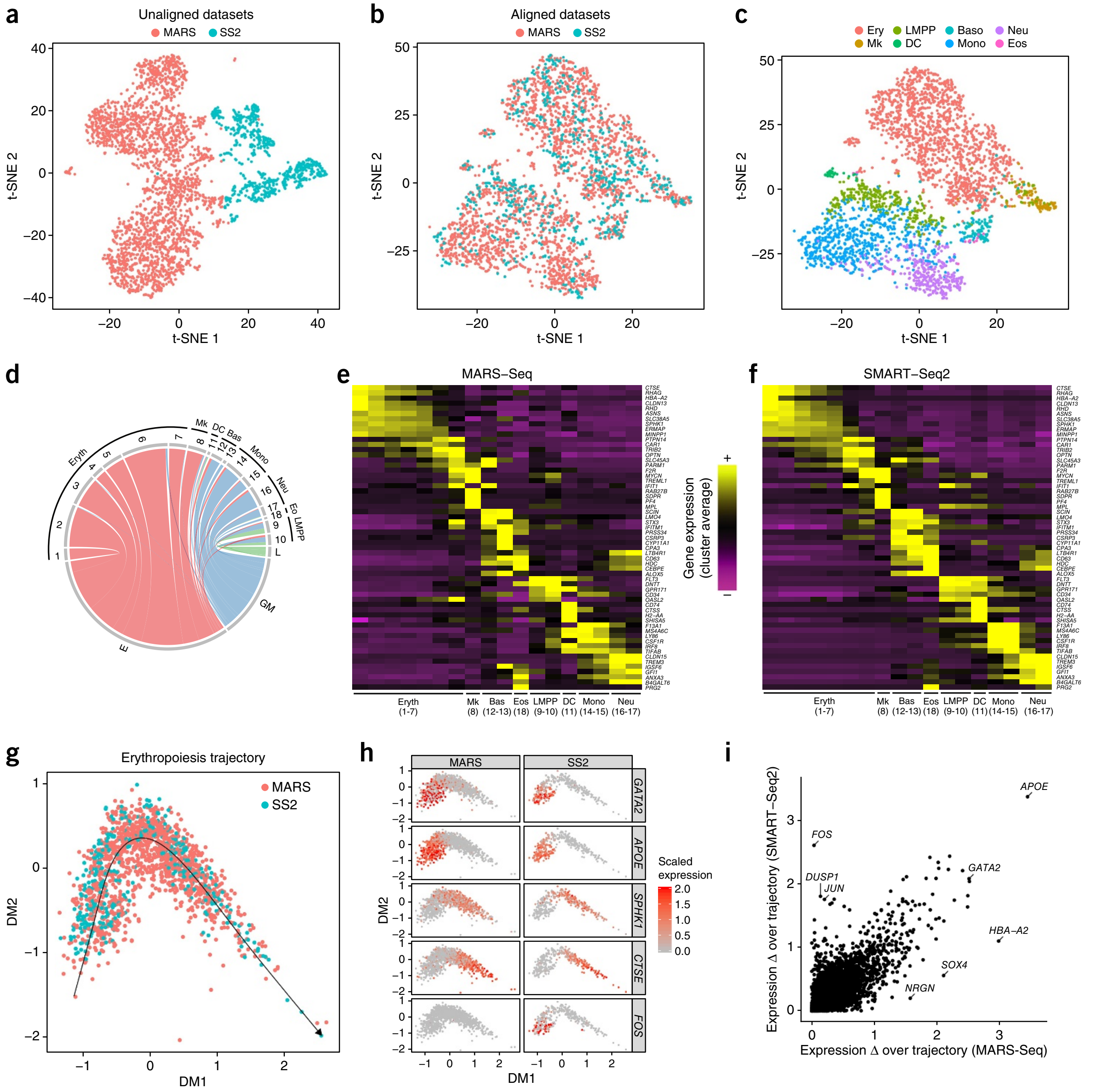

Figure 3 Comparative analysis of mouse hematopoietic progenitors across scRNA-seq technologies. (**a**–**c**) t-SNE plots of 3,451 hematopoietic progenitor cells from murine bone marrow sequenced using MARS-Seq (2,686) and SMART-Seq2 (SS2; 765), before (**a**) and after (**b**,**c**) alignment. After alignment, cells group together based on shared progenitor type irrespective of sequencing technology. (**c**,**d**) Cells from the SMART-Seq2 data set were mapped onto the closest MARS-Seq cluster and associated lineage (from Paul *et al.*[38]). t-SNE plot of cells colored by assigned lineage (**c**). Mapping correspondence between SMART-Seq2 lineage assignments (from Nestorawa *et al.*[37]) and MARS-Seq clusters (**d**). (**e**,**f**) Heatmaps showing lineage-specific gene expression patterns in MARS-Seq and SMART-Seq2 data sets. Each column represents average expression after cells are grouped by either the original MARS-Seq cluster assignments (**e**), or the MARS-Seq cluster they map to (**f**). (**g**,**h**) Integrated diffusion maps of erythroid-committed cells in both data sets reveal an aligned developmental trajectory (**g**), with conserved 'pseudo-temporal' dynamics (**h**). (**i**) Scatter plot comparing the range in expression (absolute value) over the developmental trajectory, for each gene, across both data sets.

identified by visualization and graph-based clustering despite being present in only one data set (**Supplementary Fig. 6a**–**c**).

After removing stimulated erythroblasts, we observed that control erythroblast cells no longer separated in the integrated analysis, while other populations were unaffected (**Supplementary Fig. 6d**). We therefore aimed to design a new test to identify these cells as non-overlapping, so they could be flagged for further exploration downstream. We reasoned that while CCA may struggle to identify canonical correlation vectors that define rare subpopulations present in only one data set, PCA may be able to separate these cells, as we

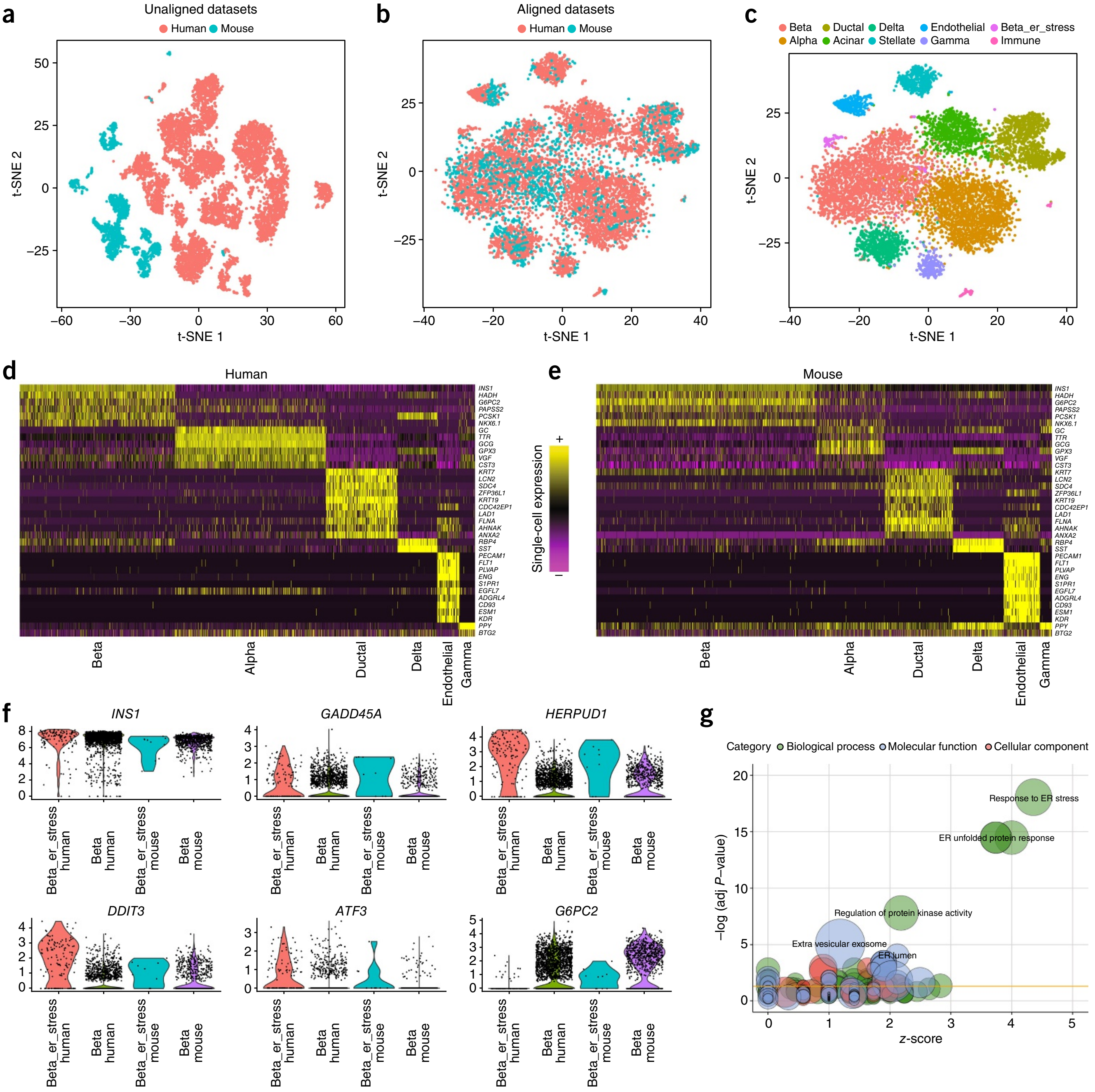

Figure 4 Joint identification of cell types across human and mouse islet scRNA-seq atlases. (**a**–**c**) t-SNE plots of 10,191 pancreatic islet cells from human (n = 8,424 cells) and mouse (n = 1,767 cells) donors, before (**a**) and after (**b**) alignment. After alignment, cells group across species based on shared cell type, allowing for a joint clustering (**c**) to detect ten cell populations. (**d**,**e**) Unsupervised identification of shared cell-type markers between human (**d**) and mouse (**e**). Single-cell expression heatmap for genes identified with joint DE testing across species. (**f**) Violin plots showing the distribution of gene expression of select genes in the beta cell cluster for human (n = 2,431 cells) and mouse (n = 762 cells) and the stressed beta cell clusters for human (n = 126 cells) and mouse (n = 10 cells). (**g**) Top n = 100 genes upregulated in the 'ER-stress' subpopulation of beta cells in both species are strongly enriched for components of the ER unfolded protein stress response. GO enrichment is visualized using the GOplot R package.

have previously shown[3]. Therefore, we quantify how well the low-dimensional space defined by CCA explains each cell's expression profile, and compare this to PCA, which is performed on each data set independently (Online Methods). Cells where the percent variance explained is reduced by a user-defined cutoff in CCA compared to PCA are therefore defined by strong sources of variance that are not shared between the data sets. We use a cutoff of 50% for all examples in this manuscript to identify these cells.

This procedure enabled us to sensitively identify a rare group of non-overlapping cells in the control population, all of which could be identified as expressing high levels of *HBA1* and *HBA2*, and corresponded to the erythroblast population whose signal was previously

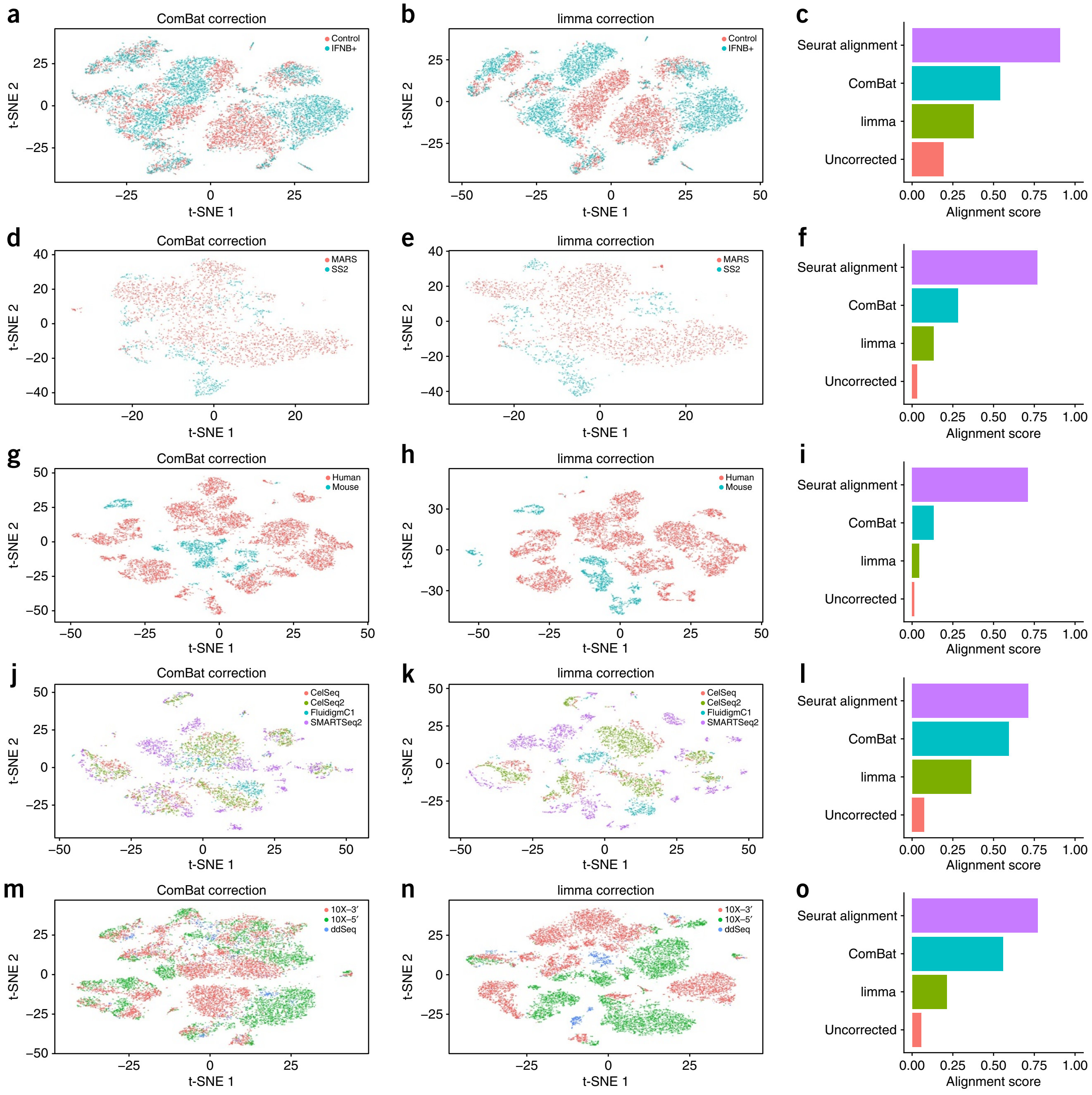

Figure 5 Benchmarking alignment and batch correction methods. (**a**,**d**,**g**,**j**,**m**) t-SNE plots for the PBMC data set (n = 14,039 cells) (**a**), hematopoietic progenitor cell data set (n = 3,451 cells) (**d**), human and mouse pancreatic islet cell data sets (n = 10,306 cells) (**g**), multiple human pancreatic islet cell data sets (n = 6,224 cells) (**j**), and multiple PBMC data sets (n = 16,653 cells) (**m**) after correction with ComBat, and (**b**,**e**,**h**,**k**,**n**) with limma. (**c**,**f**,**i**,**l**,**o**) Bar plots of the alignment score after correction using the Seurat alignment procedure, ComBat, limma, and after no correction. Seurat alignment outperforms other methods in all five examples. Additional examples of 'negative controls' where Seurat fails to align data sets from different tissues are shown in **Supplementary Figure 15**.

blended into the rest of the data (**Supplementary Fig. 6d**–**f**). Notably, this test correctly did not flag similar cells when applied to the original analysis of the full data sets (where the populations were fully overlapping), or in the *in silico* monocyte removal (where the aligned canonical correlation vectors enabled the identification of both rare and abundant cell states). Taken together, we conclude that our integration procedure is robust to abundant non-overlapping populations and can also identify rare populations that are present in a single data set, enabling further characterization.

Integrated analysis of scRNA-seq technologies

We next examined two scRNA-seq experiments that profiled the same tissue (hematopoietic progenitors from murine bone marrow), but with starkly different technologies. Nestorowa *et al.*[37] used the full-length SMART-Seq2 protocol with deep sequencing (6,558 genes/cell) to profile 765 progenitors, while Paul *et al.*[38] applied the 3′ MARS (massively parallel RNA single-cell)-Seq protocol with shallow sequencing (1,453 genes/cell) to examine 2,686 cells. The distinct differences in amplification, normalization, and coverage pose challenges to integrate these data sets. Additionally, independent analyses from both papers highlighted different aspects of the data; the SMART-Seq2 analysis focused on the broad and continuous trajectories of cells committing to lymphoid, myeloid, erythroid lineages, while the MARS-Seq data set identified 18 distinct clusters (and one contaminating group of NK cells), representing progenitors of eight distinct hematopoietic lineages. Despite these differences, we asked whether the same distinct progenitor subsets might be found in both data sets through integrated analysis.

Seurat alignment returned canonical correlation vectors that separated distinct progenitor subtypes, revealing populations committed to all eight distinct hematopoietic lineages in both data sets, and successfully identifying the contaminating NK population as ‘non-overlapping’ (**Fig. 3a**–**c** and **Supplementary Fig. 7**). After alignment, we mapped cells from the SMART-Seq2 data set onto their closest cluster in the MARS-Seq data set (**Fig. 3c**–**f** and **Supplementary Data 2**). We observed that early megakaryocyte-erythrocyte progenitor cells, identified in the original SMART-Seq2 publication, mapped exclusively onto erythroid and megakaryocytic progenitors in the MARS-Seq data (clusters C1–7). Similarly, SMART-Seq2 granulocyte-macrophage progenitors mapped onto basophil, eosinophil, dendritic cell, neutrophil, and monocyte progenitors (C11-18). While the MARS-Seq data specifically enriched for myeloid cells, the authors identified populations of very early progenitors that were FLT3$^+$ (C9-10). These cells represent lympho-myeloid component progenitors (lymphoid-primed multipotent progenitors)[39], and early lymphoid progenitors from the SMART-Seq2 data mapped exclusively to these clusters. Indeed, after mapping, we observed nearly identical segregation of gene expression markers between SMART-Seq2 and MARS-Seq data sets (**Fig. 3e**,**f** and **Supplementary Fig. 8**), demonstrating that the biological drivers of alignment were lineage-determining factors. Therefore, Seurat alignment demonstrated that distinct committed progenitor populations were present in the SMART-Seq2 data set, but were challenging to detect in the original analysis owing to reduced cell number.

Lastly, as both data sets identified developmentally heterogeneous populations during erythroid differentiation (broken into seven stages in the MARS-Seq analysis), we applied diffusion maps to erythroid-committed cells to reconstruct a joint developmental trajectory (**Fig. 3g**). We observed that this developmental path maintained the ‘pseudotemporal’ ordering of cells within both data sets (**Supplementary Fig. 9**) and also aligned the two together, exhibiting nearly identical expression dynamics for canonical differentiation markers (**Fig. 3h**). Extending this analysis globally, we observed that gene expression changes across the trajectory were largely conserved between data sets, particularly for well-characterized effectors of erythropoiesis, yet we also saw technology-specific effects—for example, a strong JUN/FOS response that has previously been associated with cellular stress during scRNA-seq[40] (**Fig. 3h**,**i**). Therefore, our procedure can successfully align both discrete and transitioning populations and enables the identification of gene-expression programs that are conserved or unique to individual data sets.

The ability to pool data sets of the same heterogeneous tissues has the potential to enable similar ‘meta-analyses’ for data sets produced across multiple laboratories and technologies. To further demonstrate this, we include two additional examples (**Supplementary Figs. 10** and **11**), demonstrating the integration of human pancreatic islets produced with four plate-based scRNA-seq technologies (CelSeq, CelSeq2, Fluidigm C1, SmartSeq2), and human PBMCs produced with three distinct technologies (10× Genomics 3′ assay, 10× Genomics 5′ assay, and the Illumina/BioRad ddSeq). In the first example, we identified eight populations of endocrine, exocrine, and stellate cells, clearly defined by cell-type-specific markers that were conserved across technologies (**Supplementary Fig. 10**). Notably, we also identified a rare population (1%) of endothelial cells which were present in all data sets, but whose rarity precluded their automated annotation in three of the four original analyses. As each sample was also from a different human donor, the proportion of cell types in each sample was highly variable but did not confound the integration procedure (**Supplementary Fig. 10e**).

In the second example, pooling the data sets yielded 16,653 PBMCs, allowing us to identify 16 immune populations, including 6 T-cell clusters (**Supplementary Fig. 11a**–**d**), and a rare subpopulation (0.5% frequency) of NK cells. This subpopulation lacked *FCGR3A* expression but was enriched for *XCL1* and *GZMK*, consistent with highly cytotoxic CD56$^{\mathrm{bright}}$ NK cells[41] (**Supplementary Fig. 11e**). As with previous examples, the rarity of these cells precludes their identification in any individual data set, and they were not identified in a previous analysis of 68,000 PBMCs[4]. These integrated data sets provide the opportunity to perform meta-analyses for differential expression across multiple technologies. As an example of this, we first performed individual ‘within data set’ differential expression tests, and then combined the results (Online Methods). Using this approach to identify differential gene expression (DE) between NK cell subsets, we were able to more than triple the number of DE genes detected between these two cell groups (**Supplementary Fig. 11f**), including chemokines (*CCL5*), transcriptional regulators (*RORA*), and surface receptors (*FCRL6*), which, although not highly expressed, are functionally important. Therefore, integrating different scRNA-seq technologies boosts the statistical power not only to discover rare cell phenotypes, but also to identify transcriptomic markers of cell state.

Joint learning of cell types across species

As a final example, we tested the ability of Seurat to align heterogeneous populations from the same tissue but originating from different species. We examined a recent single-cell study of both human and mouse pancreatic islets, performed with the inDrop technology[1], that identified islet cell types independently in both species[42]. The study found that cell-type transcriptomes were poorly conserved between human and mouse (average correlation between bulk transcriptomes of individual cell types: $R = 0.42$), often finding very few strongly expressed markers that were preserved between species. This widespread divergence poses significant challenges for integration, as structure in the data set was largely driven by species as well as by individual donor (**Fig. 4a** and **Supplementary Fig. 12**). However, we reasoned that a subset of gene–gene correlations should still be conserved, and therefore aligned all human cells against all mouse cells.

Indeed, Seurat alignment identified canonical correlation vectors that separated cell types, and flagged primarily small populations of immune cells (human mast cells and murine B cells) as non-overlapping (**Fig. 4b** and **Supplementary Fig. 12**). We next performed a single integrated clustering analysis, identifying ten clusters, corresponding to alpha, delta, gamma, acinar, stellate, ductal, epithelial, immune, and two subgroups of beta cells (**Fig. 4c**, **Supplementary Data 3** and **Supplementary Fig. 12**). Our clusters agreed overwhelmingly with the analyses from the independent data sets[42] (**Supplementary Fig. 13**), though we did observe a low rate (5.8%) of discordant calls, particularly for cells with low UMI counts (**Supplementary Fig. 14a**,**b**). We were also able to identify a subset of cell-type markers that were conserved between human and mouse (**Fig. 4d**,**e**).

Notably, our procedure identified a rare subpopulation of beta cells in both human and mouse. These cells expressed identical levels of *INS*, but upregulated the expression of endoplasmic reticulum (ER) stress genes (*HERPUD1*; *GADD45A*) in both species (**Fig. 4f**). A similar signal was observed in a semi-supervised analysis of the human beta cells in the original manuscript[42], but could not be detected in automated clustering, or an independent analysis of the murine data set. In contrast, our integrated analyses revealed a conserved set of markers that were strikingly enriched for regulators of ER stress response to unfolded proteins[43,44] (**Fig. 4g**), which has been shown to play an important role in the onset and progression of diabetes. Notably, expression of the transcription factors *ATF3* and *ATF4* was highly upregulated in both species, representing factors that have well-established roles in the initiation of stress responses in the pancreas[45,46]. Taken together, these results demonstrate that our alignment procedure can identify shared cell states even in the face of significant global transcriptional shifts, driven in this case by millions of years of evolution.

Benchmarking alignment and batch correction techniques

We next compared Seurat's performance to widely used batch correction tools that have been applied to both bulk[47] and single-cell genomics data[5]. To evaluate each technique, we designed an 'alignment score', which examines the local neighborhood of each cell after alignment (Online Methods). When data sets are well aligned, this local neighborhood will consist equally of cells from both data sets, enabling us to quantify the success of each procedure with a score ranging from 0 to 1.

On the five data sets presented here, we benchmarked Seurat's performance against ComBat[48] and limma[49] (**Fig. 5**). In each case, as can be visualized by t-SNE or quantified with our alignment score, Seurat's integration procedure yielded superior results. The differences between these procedures were particularly striking when the transcriptomic differences between data sets (i.e., batch effect) substantially outweighed differences between cell types ('biology'), as in cross-species integration. However, when we attempted to align data sets from different tissues as a negative control, we observed poor results and low alignment scores, even when cells were not automatically classified as non-overlapping (**Supplementary Fig. 15** and **Supplementary Data 4**).

DISCUSSION

We have developed a strategy to integrate scRNA-seq data sets by identifying shared sources of variation, corresponding to subpopulations present in multiple experiments. Implemented in the R toolkit Seurat, our procedure addresses several technical challenges, including the unbiased identification of shared gene–gene correlations across data sets, as well as the alignment of canonical correlation vectors using nonlinear 'warping' algorithms.

Data set integration represents a key step in a general framework for case-control studies performed with single-cell resolution. As new data sets are generated, we expect that similar computational analyses will not only be invaluable for characterizing the immune system's response to vaccination, inflammatory disease, and cancer, but also provide deeper insight into how genetic variation and manipulation affect heterogeneous populations. Similarly, we anticipate that these methods will enable consortia, such as the Human Cell Atlas[15], which aims to define all human cell types by integrating data generated across diverse single-cell omics approaches, to combine data sets produced across many laboratories and technologies[50]. Recent benchmarking studies of diverse scRNA-seq[51,52] technologies have consistently demonstrated that no single method is uniformly superior, but rather, that each has individual strengths and weaknesses, further highlighting the potential value of data integration.

We demonstrate the ability to align differentiated cell types between human and mouse pancreatic islets, identifying a shared population of beta cells responding to ER protein misfolding stress. These and similar analyses may provide valuable comparative tools for studies using mouse models of human disease, potentially enabling the identification of human correlates of pathogenic populations discovered in mouse (or vice versa). Furthermore, new data sets will enable the alignment and comparison of developmental trajectories across species, leading to a deeper understanding of how the gene regulatory networks generating cellular diversity are rewired across evolution. As comparative genomics has played a fundamental role in our understanding of the human genome, we believe that cross-species analyses may yield similar insights toward our understanding of cellular diversity.

Lastly, we note many challenges that future methods will address in extending this work. Although our procedure can jointly analyze multiple data sets with overlapping and non-overlapping populations, future data sets that consist of tens to hundreds of batches with dramatically varying sizes and non-overlapping populations will likely require new methods. We also note that examples in this manuscript, including data sets with tens of thousands of cells, run in less than half an hour on a standard laptop computer, but new data sets extending to millions of cells may require advanced computation, subsampling, or newly optimized techniques for integration. While we focus here on alignment of sequencing-based data sets, the recent invention of spatially resolved or *in situ* methods for transcriptomic profiling[53–55] raises the potential for integration with scRNA-seq data sets, extending previous efforts to spatially resolve scRNA-seq data[9,10] towards an unsupervised procedure generalizable to any tissue.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank members of the Satija laboratory, as well as P. Roelli, M. Stoeckius, G. Fishell, C. Desplan, R. Bonneau, E. Macosko, and A. Corvelo for their valuable feedback, and F. Hamey, HM Kang, and J. Ye for assistance with published data sets. This work was supported by an NIH New Innovator Award (1DP2HG009623-01) and R01 (5R01MH071679-12) to R.S. and an NSF Graduate Fellowship (DGE1342536) to A.B.

AUTHOR CONTRIBUTIONS

A.B. and R.S. conceived the research. A.B., P.H., and R.S. implemented the alignment procedure, performed all data analysis, and wrote the manuscript. E.P. performed the PBMC validation experiments, and P.S. performed the ddSeq experiments.

COMPETING INTERESTS

The authors declare no competing interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Klein, A.M. *et al.* Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187–1201 (2015).

2. Zilionis, R. *et al.* Single-cell barcoding and sequencing using droplet microfluidics. *Nat. Protoc.* **12**, 44–73 (2017).
3. Macosko, E.Z. *et al.* Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).
4. Zheng, G.X.Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).
5. Shekhar, K. *et al.* Comprehensive classification of retinal bipolar neurons by single-cell transcriptomics. *Cell* **166**, 1308–1323.e30 (2016).
6. Villani, A.-C. *et al.* Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science* **356**, eaah4573 (2017).
7. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**, 381–386 (2014).
8. Welch, J.D., Hartemink, A.J. & Prins, J.F. SLICER: inferring branched, nonlinear cellular trajectories from single cell RNA-seq data. *Genome Biol.* **17**, 106 (2016).
9. Satija, R., Farrell, J.A., Gennert, D., Schier, A.F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33**, 495–502 (2015).
10. Achim, K. *et al.* High-throughput spatial mapping of single-cell RNA-seq data to tissue of origin. *Nat. Biotechnol.* **33**, 503–509 (2015).
11. DeLaughter, D.M. *et al.* Single-cell resolution of temporal gene expression during heart development. *Dev. Cell* **39**, 480–490 (2016).
12. Bendall, S.C. *et al.* Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. *Cell* **157**, 714–725 (2014).
13. Blakeley, P. *et al.* Defining the three cell lineages of the human blastocyst by single-cell RNA-seq. *Development* **142**, 3613 (2015).
14. Johnson, M.B. *et al.* Single-cell analysis reveals transcriptional heterogeneity of neural progenitors in human cortex. *Nat. Neurosci.* **18**, 1–30 (2015).
15. Regev, A. *et al.* The Human Cell Atlas. *Elife* **6**, 1–30 (2017).
16. Kharchenko, P.V., Silberstein, L. & Scadden, D.T. Bayesian approach to single-cell differential expression analysis. *Nat. Methods* **11**, 740–742 (2014).
17. Finak, G. *et al.* MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* **16**, 278 (2015).
18. Wang, B., Zhu, J., Pierson, E., Ramazzotti, D. & Batzoglou, S. Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. *Nat. Methods* **14**, 414–416 (2017).
19. Kiselev, V.Y. *et al.* SC3: consensus clustering of single-cell RNA-seq data. *Nat. Methods* **14**, 483–486 (2017).
20. Lin, P., Troup, M. & Ho, J.W.K. CIDR: Ultrafast and accurate clustering through imputation for single-cell RNA-seq data. *Genome Biol.* **18**, 59 (2017).
21. Prabhakaran, S., Azizi, E. & Pe'er, D. Dirichlet process mixture model for correcting technical variation in single-cell gene expression data. *Proc. 33rd Int. Conf. Mach. Learn.* **48**, 1070–1079 (2016).
22. Ntranos, V., Kamath, G.M., Zhang, J.M., Pachter, L. & Tse, D.N. Fast and accurate single-cell RNA-seq analysis by clustering of transcript-compatibility counts. *Genome Biol.* **17**, 112 (2016).
23. Xu, C. & Su, Z. Identification of cell types from single-cell transcriptomes using a novel clustering method. *Bioinformatics* **31**, 1974–1980 (2015).
24. Lei, Z., Bai, Q., He, R. & Li, S.Z. Face shape recovery from a single image using CCA mapping between tensor spaces. *26th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR* doi:10.1109/CVPR.2008.4587341 (2008).
25. Zhou, F. & Torre, F. in *Advances in Neural Information Processing Systems 22; NIPS 2009* (eds. Y. Bengio, Y., Schuurmans, D., Lafferty, J.D., Williams, C.K.I. & Culotta, A.) https://papers.nips.cc/paper/3728-canonical-time-warping-for-alignment-of-human-behavior (Neural Information Processing Systems Foundation, Inc., 2009).
26. Wang, C. & Mahadevan, S. in *Proc. Twenty-Second International Joint Conference on Artificial Intelligence,* Vol. 2 (ed. Walsh, T.) 1541–1546 (AAAI, 2011).
27. Huang, H., He, H., Fan, X. & Zhang, J. Super-resolution of human face image using canonical correlation analysis. *Pattern Recognit.* **43**, 2532–2543 (2010).
28. Hotelling, H. Relations between two sets of variates. *Biometrika* **28**, 321–377 (1936).
29. Hardoon, D.R., Szedmak, S. & Shawe-Taylor, J. Canonical correlation analysis: an overview with application to learning methods. *Neural Comput.* **16**, 2639–2664 (2004).
30. Witten, D.M., Tibshirani, R. & Hastie, T. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* **10**, 515–534 (2009).
31. Lê Cao, K.-A., Martin, P.G., Robert-Granié, C. & Besse, P. Sparse canonical methods for biological data integration: application to a cross-platform study. *BMC Bioinformatics* **10**, 34 (2009).
32. Waaijenborg, S., Verselewel de Witt Hamer, P.C. & Zwinderman, A.H. Quantifying the association between gene expressions and DNA-markers by penalized canonical correlation analysis. *Stat. Appl. Genet. Mol. Biol.* **7**, e3 (2008).
33. Kettenring, J. Canonical analysis of several sets of variables. *Biometrika* **58**, 433–451 (1971).
34. Nielsen, A.A. Multiset canonical correlations analysis and multispectral, truly multitemporal remote sensing data. *IEEE Trans. Image Process.* **11**, 293–305 (2002).
35. Berndt, D. & Clifford, J. Using dynamic time warping to find patterns in time series. *Work. Knowl. Knowl. Discov. Databases* **398**, 359–370 (1994).
36. Kang, H.M. *et al.* Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat. Biotechnol.* **36**, 89–94 (2018).
37. Nestorowa, S. *et al.* A single-cell resolution map of mouse hematopoietic stem and progenitor cell differentiation. *Blood* **128**, e20–e31 (2016).
38. Paul, F. *et al.* Transcriptional heterogeneity and lineage commitment in myeloid progenitors. *Cell* **163**, 1663–1677 (2015).
39. Adolfsson, J. *et al.* Identification of Flt3+ lympho-myeloid stem cells lacking erythro-megakaryocytic potential a revised road map for adult blood lineage commitment. *Cell* **121**, 295–306 (2005).
40. Lacar, B. *et al.* Corrigendum: nuclear RNA-seq of single neurons reveals molecular signatures of activation. *Nat. Commun.* **8**, 15047 (2017).
41. Poli, A. *et al.* CD56bright natural killer (NK) cells: an important NK cell subset. *Immunology* **126**, 458–465 (2009).
42. Baron, M. *et al.* A single-cell transcriptomic map of the human and mouse pancreas reveals inter- and intra-cell population structure. *Cell Syst.* **3**, 346–360.e4 (2016).
43. Scheuner, D. & Kaufman, R.J. The unfolded protein response: a pathway that links insulin demand with β-cell failure and diabetes. *Endocr. Rev.* **29**, 317–333 (2008).
44. Walter, W., Sánchez-Cabo, F. & Ricote, M. GOplot: an R package for visually combining expression data with functional analysis. *Bioinformatics* **31**, 2912–2914 (2015).
45. Jiang, H.-Y. *et al.* Activating transcription factor 3 is integral to the eukaryotic initiation factor 2 kinase stress response. *Mol. Cell. Biol.* **24**, 1365–1377 (2004).
46. Papa, F.R. Endoplasmic reticulum stress, pancreatic β-cell degeneration, and diabetes. *Cold Spring Harb. Perspect. Med.* **2**, a007666 (2012).
47. Conesa, A. *et al.* A survey of best practices for RNA-seq data analysis. *Genome Biol.* **17**, 13 (2016).
48. Johnson, W.E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007).
49. Ritchie, M.E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
50. Lake, B.B. *et al.* Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. *Science* **352**, 1586–1590 (2016).
51. Ziegenhain, C. *et al.* Comparative analysis of single-cell RNA sequencing methods. *Mol. Cell* **65**, 631–643.e4 (2017).
52. Svensson, V. *et al.* Power analysis of single-cell RNA-sequencing experiments. *Nat. Methods* **14**, 381–387 (2017).
53. Junker, J.P. *et al.* Genome-wide RNA tomography in the zebrafish embryo. *Cell* **159**, 662–675 (2014).
54. Lee, J.H. *et al.* Highly multiplexed subcellular RNA sequencing in situ. *Science* **343**, 1360–1363 (2014).
55. Ståhl, P.L. *et al.* Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* **353**, 78–82 (2016).

ONLINE METHODS

The Seurat alignment procedure is designed to integrate single-cell RNA sequencing data (scRNA-seq) across distinct data sets. Following is an overview of the main steps comprising a typical workflow:

1. Data preprocessing and gene selection.
2. Define a shared correlation space with canonical correlation analysis.
3. Identify rare non-overlapping subpopulations.
4. Align correlated subspaces using dynamic time warping.
5. Integrated analysis across data sets (clustering, trajectory building, differential expression).

Below, we describe each of these steps in detail. Additionally, we provide full command lists for the integration of the stimulated and resting immune data sets and for the integration of the four scRNA-seq data sets of human pancreatic islet cells (produced with four different plate-based technologies CelSeq, CelSeq2, Fluidigm C1, SmartSeq2) as **Supplementary Software**.

Single-cell data set preprocessing. For all single-cell analysis, we performed the same initial normalization. Gene expression values for each cell were divided by the total number of transcripts and multiplied by 10,000. These values were then natural-log transformed using log1p before further downstream analyses. After normalization, we calculated scaled expression (*z*-scores for each gene) for downstream dimensional reduction.

Comparison of stimulated and resting immune cells. We obtained a unique molecular identifier (UMI) count matrix for the Kang *et al.*[36] study from GSE96583. The authors generously provided us with the output of their demuxlet algorithm, which computationally identifies doublets, and assigns individual single cells to one of eight patients. We removed cells with fewer than 500 genes detected, leaving 14,039 single cells in total.

Integrated analysis of scRNA-seq technologies. We obtained a read count matrix for the SMART-Seq2 data set (Nestorowa *et al.*)[37] under the GEO accession GSE81682, and considered 765 annotated progenitors cells expressing at least 4,000 genes. The authors generously provided lineage annotations for each cell (corresponding to **Fig. 4** in the original publication, used in our **Fig. 3**). We obtained a batch-corrected UMI count matrix for the MARS-Seq data set[38] from the authors' online resource (http://compgenomics.weizmann.ac.il/tanay/?page_id=649), where we also obtained the MARS-Seq cluster IDs for each cell. This data set had been previously filtered to remove cells with less than 500 detected UMI for a total of 2,686 single cells.

Both data sets contain cycling progenitors, and heterogeneity between cell cycle stages for these cells has previously been shown to confound developmental analyses. Therefore, independently for both data sets, we first assigned a cell cycle score to each cell using the PCA method[56] on a previously annotated list of cell cycle genes[57]. We then used the ScaleData function in Seurat (using the cell cycle score as latent variable in a linear regression framework) to mitigate this source of variation in the data set, before CCA.

Joint clustering across species. We obtained UMI count matrices for the human and mouse inDrops data sets (Baron *et al.*, 2016)[42] from GEO accession GSE84133. For both species, we removed cells with less than 500 detected genes to obtain 8,536 and 1,770 single cells, respectively. We also regressed out individual-specific effects using ScaleData before CCA. We considered all homologous genes with identical gene names between the human and mouse data sets, and allowed the *INS* human gene to map to the mouse *Ins1* and *Ins2* genes as in the original manuscript[42].

Alignment of multiple human pancreas data sets. For the four human pancreas data sets, we obtained count matrices from accession numbers GSE81076 (CelSeq), GSE85241 (CelSeq2), GSE86469 (Fluidigm C1), and E-MTAB-5061 (SMART-Seq2). We filtered cells that expressed less than 1,750 unique genes/cell (CelSeq), or 2,500 genes/cell (CelSeq2/Fluidigm C1/SMART-Seq2), leaving 6,224 cells in total.

Integrated analysis of multiple PBMC data sets across technologies. For the three human PBMC data sets, we obtained gene expression matrices from 10× genomics (https://support.10xgenomics.com/single-cell-gene-expression/datasets/2.1.0/vdj_v1_pbmc_5gex, https://support.10xgenomics.com/single-cell-gene-expression/datasets/2.1.0/pbmc8k). For the ddSeq experiment, PBMCs from a healthy donor were diluted to 2,500 cells/µl and run according to manufacturer's protocol through 8 ddSeq wells (two cartridges) with an expected yield of 2,400 cells in total. Sequencing libraries were prepared according to manufacturer's instructions and sequenced on two lanes of a HiSeq 2500 in rapid run mode, with 68 cycles for read 1 (cell barcodes + UMI), 8 cycle sample index and 75 cycle read 2 (transcript). We filtered out those cells with fewer than 750 unique genes, resulting in 16,653 cells in total. Additionally, we observed significant mitochondrial heterogeneity within each data set, in keeping with previous reports[58], and regressed out mitochondrial heterogeneity from each data set before running CCA.

Gene set selection. Although the alignment procedure can utilize any gene that is measured with nonzero variance in all data sets, we focused on genes that were highly variable in one or both data sets. We identified these genes by calculating the dispersion (variance to mean ratio) for all genes in each data set and selected 1,000 genes with the highest dispersion from each. We took the union of these two resulting gene lists as the input genes for CCA. For multi-CCA, we required that input genes be in the highly variable gene list for at least two data sets.

Calculation of canonical correlation vectors. Standard canonical correlation analysis was designed to find projections that maximize correlation between two vectors (data sets or groups). We first describe the two-set scenario and then extend this to multiple sets.

Two set canonical correlation. The first step in the alignment utilizes a variation on canonical correlation analysis (CCA) to find projections of both data sets such that the correlation between the two projections is maximized. Formally, CCA finds projection vectors *u* and *v* such that the correlation between the two indices u^TX and v^TY is maximized[28].

$$\max_{u,v} u^T X^T Y v \ \text{ subject to } u^T X^T X u \leq 1,\ v^T Y^T Y v \leq 1 \tag{1}$$

To apply this in the context of scRNA-seq, let $X_{g,c}$ be a gene expression matrix of genes $g_1, g_2, ..., g_n$ by cells $c_1, c_2, ..., c_m$ and $\mathrm{Y}_{\mathrm{g,d}}$ be a gene expression matrix of the same genes $g_1, g_2, ..., g_n$ by cells $d_1, d_2, ..., d_p$. In many scRNA-seq experiments, the number of genes of interest that are shared between the two data sets is often much smaller than the total number of cells that were measured ($n << m + p$). Consequently, the vectors *u* and *v* that are returned from CCA as described in equation (1) will not be unique.

One potential solution to this is to regularize or penalize the CCA procedure to promote sparsity. However, this would assign many cells zero loadings in the resulting projections and result in a complete loss of information for a significant proportion of cells. Therefore, we treat the covariance matrix within each data set as diagonal, a solution that has demonstrated promising results in other high-dimensional problems[59,60]. We substitute the identity matrix for X^TX and Y^TY to arrive at equation (2).

$$\max_{u,v} u^T X^T Y v \text{ subject to } \|u\|_2^2 \leq 1, \|v\|_2^2 \leq 1 \tag{2}$$

To construct our canonical correlation vectors, we standardized *X* and *Y* to have a mean of 0 and variance of 1.

$$\forall_c \sum X[,c]/n = 0,\ var(X[,c]) = 1 \text{ and } \forall_d \sum X[,d]/n = 0,\ var(X[,d]) = 1$$

We then are able to solve for the canonical correlation vectors *u* and *v* using singular value decomposition (SVD) as follows:

Let

$$K = X^T Y$$

K can be decomposed using SVD as

$$K = \Gamma \Lambda \Delta^T \tag{3}$$

where

$$\Gamma = (\gamma_1, ..., \gamma_k)$$

$$\Delta = (\delta_1, ..., \delta_k)$$

$$\Lambda = (\lambda_1^{1/2}, ..., \lambda_k^{1/2})$$

Since we have substituted the identity matrix for X^TX and Y^TY, we can obtain our canonical correlation vectors u and v as the left and right singular vectors from the SVD for $i = 1, \ldots, k$.

$$u_i = \gamma_i$$
$$v_i = \delta_i$$

Since we are interested in only a subset of the canonical correlation vectors, we approximated the singular value decomposition with a partial singular value decomposition using the augmented implicitly restarted Lanczos bidiagonalization algorithm implemented in the irlba R package[61]. This procedure returns a user-defined number (k) of left and right singular vectors, which approximate the canonical correlation vectors that project each expression matrix into the maximally correlated subspace.

Multi-set canonical correlation. Two-set canonical correlation analysis can be extended to multi-set analysis (i.e., Multi-CCA), aiming to identify projection vectors that maximize the overall correlation across all data sets. There are several options for how to exactly formulate this optimization problem, for which full descriptions can be found in Kettenring (1971)[33]. Here, we've chosen to use the same approach as described in Witten and Tibshirani (2009)[30].

Formally, if we have N data sets $X_1,\ldots,X_N$, the goal is to find projection vectors $W = w_1,\ldots,w_N$ that maximize:

$$\max_{w_1,\ldots,w_N} \sum_{i<j} w_i^T X_i^T X_j w_j \ \text{ subject to } \ w_n^T X_n^T X_n w_n = 1 \forall n$$

To address cases where there are more cells than genes, as discussed in the previous section, we again make the diagonalizing assumption for the covariance matrix of each data set. By substituting the identity matrix for each $X_n^T X_n$ we arrive at

$$\max_{w_1,\ldots,w_N} \sum_{i<j} w_i^T X_i^T X_j w_j \ \text{ subject to } \ \|w_n\|^2 \leq 1 \tag{4}$$

To construct the canonical correlation vectors, we first standardized each X_n such that $\forall_c \sum X[,c]/n = 0, \ \mathrm{var}(X[,c]) = 1$. We then are able to solve for the canonical correlation vectors w_n using the following iterative algorithm:

Algorithm 1 Multi-CCA

```
procedure MCCA
    for all $X_n$ do
        $[\Gamma,\Lambda,\Delta] \leftarrow \mathrm{SVD(X)}$
        initialize $w_n \leftarrow \Delta$
    for 1, …, # of CCs to compute do
        while $|o1 - o2| / |o1| <$ threshold do
            $o1 \leftarrow$ calculate objective for $w_n[cc]$    (4)
            for all $X_n$ do
                update $w_n[cc] \leftarrow \frac{(X_n^T \sum_{k \neq n} X_k w_k[cc])}{\|(X_n^T \sum_{k \neq n} X_k w_k[cc])\|_2}$
            $o2 \leftarrow$ calculate objective for $w_n[cc]$    (4)
    return $W$
```

For computational efficiency, we again use irlba for the initializing singular value decomposition and set a default convergence threshold of 10^{-3} with a maximum of 25 iterations in the while loop.

Identification of rare non-overlapping subpopulations. CCA returns vectors that capture sources of variance that are shared between data sets. Therefore, CCA will not pick up sources of variation that are unique to a single data set, for example, if there is a rare unique population in only one data set. In contrast, principal component analysis (PCA) would capture this signal when performed individually on each data set.

We therefore reasoned that we could compare the results of PCA and CCA to identify cells whose expression patterns were not well-explained by a shared correlation structure. In principle, this allows for unsupervised identification of non-overlapping subpopulations, which can be filtered out before continuing the alignment procedure. This is of particular importance for rare subpopulations, which, if not identified as non-overlapping, could blend into abundant cell states after alignment (**Supplementary Fig. 6d–f**).

Therefore, we quantified how well the low-dimensional subspace defined by CCA explains the variance in gene expression as compared to PCA run on each data set independently. By computing the ratio η of these two measures of variance, we are able to identify cells that have low values for η and may originate from non-overlapping states. In our demonstrated alignment examples, we chose to use the first 20 dimensions from our CCA and PCA calculations when computing η.

To compute this ratio, we first calculate the gene loading matrices A and B for X and Y respectively as

$$A = Xu$$
$$B = Yv$$

We then form orthonormal bases D and E via QR decomposition such that

$$A = DR$$
$$B = ER$$

and project the expression data onto D and E to get $\tilde{X}$ and $\tilde{Y}$.

$$\tilde{X} = X^T D$$
$$\tilde{Y} = Y^T E$$

Next, we reconstruct the data to get $\widehat{X}$ and $\widehat{Y}$

$$\widehat{X} = D\tilde{X}^T$$
$$\widehat{Y} = EY^T$$

We then calculate the variance in gene expression σ_{CCA} or every cell in $\widehat{X}$ and $\widehat{Y}$.

$$\sigma_{CCA} = \sum_g^n \mathrm{var}(\widehat{X}[g,]) \ \sigma_{CCA} = \sum_g^n \mathrm{var}(\widehat{Y}[g,]) \tag{5}$$

Then, we run a principle component analysis on X and Y to produce orthogonal gene loading matrices F and G. Similarly, we can project the expression data onto F and G and reconstruct the data to get $\widehat{X}$ and $\widehat{Y}$.

$$\tilde{X} = X^T F$$
$$\tilde{Y} = Y^T G$$
$$\widehat{X} = F\tilde{X}^T$$
$$\widehat{Y} = G\tilde{Y}^T$$

We calculate the variance in gene expression σ_{PCA} in $\widehat{X}$ and in $\widehat{Y}$ using equation (5). Finally, we define η as the ratio of σ_{CCA} to σ_{PCA} to serve as an indicator of how well each cell is defined by shared sources of variance (lower values indicating non-overlapping cells).

$$\eta = \frac{\sigma_{CCA}}{\sigma_{PCA}} \tag{6}$$

Empirically, we applied a threshold of 0.5 uniformly in all data sets, where cells with $\eta < 0.5$ were considered non-overlapping. We found that this unsupervised procedure robustly identified rare populations that were unique to only one data set. These included terminally differentiated NK cells in the MARS-Seq hematopoietic progenitors, B and T cells in the murine pancreatic islet data set, and mast cells in the human pancreatic islet data set (**Supplementary Figs. 7** and **12**).

However, identifying a single value for this threshold was challenging for abundant populations that were specific to a single data set, for example, in our negative control experiments where we aligned data sets that have negligible biological similarity. However, in each of these cases (**Supplementary Fig. 15**), even though we did not initially identify these cells as non-overlapping, our procedure did not artificially align cells from these data sets together. While we anticipate that new methods that can robustly identify non-overlapping subpopulations before alignment will be exciting avenues for further development, our examples demonstrate that our method does not artificially align either rare or abundant non-overlapping subpopulations together.

Aligning canonical correlation vectors. After CCA, CC vectors are by definition correlated, but not necessarily aligned between data sets. In particular, shifts in feature scale or population densities can drive global differences between CC loadings, and must be corrected for as part of the alignment procedure, as described below.

Gene selection for canonical correlation vector alignment. We first identify genes whose expression robustly correlates with each projection vector in both data sets, and therefore drive shared sources of variation. For this, we use the biweight midcorrelation (bicor), a median based similarity metric.

For each canonical correlation vector *i*,…, *k*

$$\varsigma = \forall_{g} \min(bicor\left(X\left[g,\right], u_i\right), bicor\left(Y\left[g,\right], v_i\right))$$

We take the genes with the highest ζ values (M) to construct a “metagene”, a weighted linear combination of genes, to use for alignment. In all examples here, we used the top 30 genes to construct the metagene average. However, we note that exact choice of this parameter is robust across a wide range of values. Across all examples in **Figures 2**–**4**, we varied this parameter across a range (20–100 genes), and assessed the final alignment score. We observed only minor differences (with an average of less than 2% shift compared to 30 genes).

However, when we continued to reduce this parameter we did begin to observe larger changes in the alignment score (>5% when using fewer than ten genes). This is likely due to the fact that when only small numbers of genes are considered, biological stochasticity or technical noise will play a larger role in the pooled metagene. Therefore, the minimum number of genes that are required to have conserved expression patterns between data sets, in order to be correctly aligned, will depend on the scale and sequencing depth of each data set. Data sets with larger cell number, or deeper sequencing, will be able to pick up on more subtle patterns (with fewer conserved genes), analogous to experimental design considerations for detecting subtle transcriptomic states using unsupervised clustering.

Alignment of two canonical correlation vectors. We define two vectors of metagenes Φ_i and Θ_i where for each cell *c* in *X* and each cell *d* in *Y*, the metagene is defined as

$$\Phi_{i,c} = u_i X[M,c] \, \Theta_{\mathrm{i,d}} = v_i Y[M,d] \tag{7}$$

Each vector of metagenes is then scaled from 0 and 1 to match its 95% reference range. To do this, we define Q as the quantile function that gives the inverse of the empirical distribution function.

$$\Phi_i' = \frac{\Phi_i - Q_{\Phi_i}(2.5)}{Q_{\Phi_i}(97.5) - Q_{\Phi_i}(2.5)} \quad \Theta_i' = \frac{\Theta_i - Q_{\Theta_i}(2.5)}{Q_{\Theta_i}(97.5) - Q_{\Theta_i}(2.5)}$$

We then look for systematic shifts that still remain after scaling which are largely driven by outliers and linearly shift the metagenes to correct for this. This procedure robustly corrects for differences in feature scale.

$$\Phi_i' = \Phi_i' + \min_{z=10,\ldots,90} \left| Q_{\Phi_i'}(z) - Q_{\Theta_i'}(z) \right|$$

Next, we determine an optimal mapping between the metagenes, using dynamic time warping (DTW) as implemented in the dtw R package[62] with default parameters. Traditionally used to find an alignment between two time series[35], DTW effectively aligns each cell in the smaller data set to the cell with the most similar metagene expression in the larger data set, while maintaining the relative ordering of cells within each data set. To do this, DTW computes a warping path *W* that maps elements of *X* and *Y* in order to minimize the distance between them.

$$W = w_1, w_2, \ldots, w_k$$

Each w_k corresponds to a point along the warping path that maps an element in *X* to an element in Y. The minimization problem can then be defined in terms of the cumulative warping distance. We chose to use Euclidean distance as the distance function δ.

$$\mathrm{DTW}(X,Y) = \min_{W} \left[\sum_{k=1}^{p} (w_k) \right] \tag{8}$$

A key feature of DTW in the alignment procedure is the nonlinear warping of each metagene vector. These compressions and stretches correspond to potential shifts in population density across data sets. We then apply an identical warping to the canonical correlation vectors, mapping the CC values from both data sets onto a common aligned scale. We apply this procedure to each pair of basis vectors individually to define a single, aligned, low-dimensional space representing both data sets.

Alignment of multiple canonical correlation vectors. Extending the alignment procedure to multiple data sets follows naturally from the two data-set case. We first choose a reference data set, which we set by default to be the data set with the largest number of cells. We then perform repeated pairwise alignments of the canonical correlation vectors to the reference exactly as described for the two-set case above. This procedure warps the canonical correlation vectors for each data set onto a common aligned space, defined by the “reference” data set.

CC selection for downstream analysis. Dimensionality reduction, such as PCA, is a commonly applied tool in scRNA-seq analysis to help overcome technical noise and summarize the data in a smaller number of features. Choosing the number of PCs to include for downstream analyses is often performed by plotting the variance explained as a function of the number of principal components, and examining this relationship for saturation. Similarly, here we must decide on the number of aligned canonical correlation vectors to include for downstream analysis. To help guide this parameter choice, we calculated a measure of the correlation strength for each CC vector. Specifically, for each data set X_n and each CC vector w_n, we examined all genes involved in the construction of the “metagene” (as described above, $M = m_1, \ldots, m_g$) and calculated the average biweight midcorrelation.

$$\tau = \frac{1}{g} \sum_{i=1}^{g} bicor\left(X_n\left[m_i,\right], w_n\right) \tag{9}$$

For the reference data set, we take the average of τ across all pairwise calculations. For each data set, we plot a LOESS curve of τ (“Shared correlation strength”) as a function of CC vector. Curves for all five examples in this manuscript are shown in **Supplementary Figure 1**. The saturation point on these curves provides a valuable guide for the number of CCs to include in downstream analyses. Importantly, while the exact saturation point can be subjective, we observe that the global structure of our integrated data set is robust to the exact choice of this parameter within 5CCs (**Supplementary Fig. 1**).

Calculating an alignment score. While our t-SNE plots provide a visual representation of the overlap between data sets after alignments, we sought to develop a quantitative metric to ask how well any group of data sets is aligned. We calculated an alignment score as follows. First, we randomly downsample the data sets to have the same number of cells as the smallest data set. Then, we construct a nearest-neighbor graph based on the cells′ embedding in some low-dimensional space (the aligned CC space after running the alignment procedure). For every cell, we then calculate how many of its *k* nearest-neighbors belong to the same data set and average this over all cells to obtain $\overline{x}$. If the data sets are well-aligned, we would expect that each cells’ nearest neighbors would be evenly shared across all data sets. For all of our examples, we chose *k* to be 1% of the total number of cells. We then normalize by the expected number of same data set cells and scale to range from 0 to 1.

$$\text{Alignment Score} = 1 - \frac{\overline{x} - \frac{k}{N}}{k - \frac{k}{N}}$$

Integrated analysis across data sets. The aligned canonical basis vectors form a shared low-dimensional space that can be used for integrated downstream analyses, for example, clustering or trajectory building. We describe our analyses individually for each data set below.

Modularity-based clustering to identify cell types. To partition cells into clusters, we used the smart local moving (SLM) algorithm for modularity-based clustering[63]. For each of the five data sets, we computed a cell–cell distance matrix constructed on selected aligned canonical correlation vectors. We constructed a shared-nearest neighbor (SNN) graph based on this distance matrix to use as input to the SLM algorithm, implemented through the FindClusters function in Seurat. To visualize the resulting clusters in two dimensions, we used Barnes–Hut implementation of the *t*-distributed stochastic neighbor embedding (t-SNE) algorithm[64].

Identification of PBMC subtypes. We chose to use the first 20 aligned canonical correlation vectors to calculate the cell–cell distance matrix and subsequent SNN. We ran FindClusters with a resolution parameter of 0.6, resulting in 13 distinct clusters of cells. These clusters corresponded to CD14$^+$ and CD16$^+$ monocytes, CD4$^+$ memory and naive T cells, CD8$^+$ T cells, B cells, NK cells, dendritic cells, and erythrocyte populations, which all showed significant enrichment for canonical cell-type markers after running a Wilcoxon rank-sum test for differential expression implemented in Seurat as FindAllMarkers. The same 20 CCs were used as input for visualization via t-SNE.

To understand global correlations between IFNβ responses for each cell type (**Fig. 2g**), we first placed cells into 26 bins (based on the 13 immune clusters, but also grouped stimulated and resting cells within each cluster separately), and calculated the average expression for each gene within each group. The difference between the average expression of stimulated and resting cells for each cluster represents its transcriptional response to IFNβ stimulation. We then calculated the Pearson correlation of these responses between all pairs of clusters, using 430 genes that exhibited at least a twofold change in response to stimulation for at least 1 of the 13 clusters.

Identification of hematopoietic progenitor populations. To identify the hematopoietic progenitor populations present in both the SMART-Seq2 data set and the MARS-Seq data set, we used the first ten aligned canonical correlation vectors as input to calculate the cell–cell distance matrix, SNN, and visualization via t-SNE.

We then mapped cells from the SMART-Seq2 data set to one of the clusters originally identified in the MARS-Seq data set. In principle, we could simply map each SMART-Seq2 cell to the cluster identity of its nearest neighbor in the MARS-Seq data set. However, in order to make sure that our mappings were also consistent with the overall structure of the data, we used a two-step procedure.

We first performed a joint clustering on the first ten aligned CC embeddings using the SLM algorithm via FindClusters with default parameters, revealing ten clusters (referred to below as 'joint clusters'). We then calculated the percentage of cells in each MARS-Seq cluster that fell into each of the joint clusters. Let $Freq_{xy}$ represent the proportion of cells in MARS-Seq cluster *x*, that fall into joint cluster *y*.

We next mapped each SMART-Seq cell to the cluster of its closest MARS-Seq neighbor, based on the SNN-defined distance matrix. However, we defined the mapping as discordant if the mapped cluster was present at less than 25% frequency in the joint clustering, i.e., $Freq_{xy} < 0.25$. In this case, we mapped the SMART-Seq2 cell to its next closest neighbor. Finally, once all cells had been mapped to MARS-Seq clusters, we assigned each cell a lineage identity based on **Figure 2** in Paul *et al.* 2015 (ref. 38). These cluster and lineage assignments were used in all downstream analyses.

Identification of pancreatic islet subtypes in human and mouse. We identified conserved populations of islet subtypes by running FindClusters using the first 20 aligned CC embeddings with a resolution parameter of 0.5. This resulted in ten clusters corresponding to alpha, normal beta, ER stressed beta, delta, gamma, ductal, acinar, stellate, endothelial, and immune cells. The same 20 aligned CCs were used for t-SNE visualization. We also observed an 11th cluster of 115 cells that was defined almost entirely by low complexity (median 1,088 genes), which we removed from further analysis (**Supplementary Fig. 13**).

Identification of pancreatic islet subtypes in the four human data sets. In order to identify populations of islet subtypes across the four human pancreas data sets, we ran FindClusters using the first ten CC embeddings with a resolution parameter of 0.4. This gave eight distinct clusters corresponding to alpha, beta, ductal, acinar, delta, gamma, stellate, and endothelial populations. The same ten aligned CCs were used for t-SNE visualization.

Identification of PBMC subtypes across three technologies. We chose to use the first 20 aligned canonical correlation vectors as input to FindClusters with a resolution parameter of 1.2. Sixteen clusters were identified, which correspond to CD4$^+$ memory, naive, and regulatory T cells, two CD14$^+$ monocyte populations (HLA low and HLA high), pre-B cells, B cells, CD8$^+$ naive T cells, two populations of CD8$^+$ effector T cells, CD16$^+$ monocytes, conventional dendritic cells, plasmacytoid dendritic cells, megakaryocytes, and two populations of NK cells (CD56$^{\mathrm{bright}}$ and CD56$^{\mathrm{dim}}$).

Construction of joint erythroid developmental trajectories. We first took a subset of the combined hematopoietic progenitor data set that included all cells originally assigned to the first seven clusters in the MARS-Seq data set and cells that mapped to one of those clusters from the SMART-Seq2 data set. We then built a diffusion map using the first ten aligned CC embeddings using the diffuse function from the diffusionMap R package (Richards 2014)[65] with epsilon parameter set to 9, corresponding to the median distance to the 0.05**n* nearest neighbor. Next, we fit a principal curve[66] through the first two diffusion map coordinates using the principal. curve function from the princurve R package with default parameters[67]. The order of a cell's projection onto this principal curve represents its predicted progression through erythropoiesis, or "pseudotime" value, as shown in **Supplementary Figure 9a**,**b**.

In order to determine the transcriptomic range for each gene across erythropoiesis for the SMART-Seq2 and MARS-Seq data sets, we first calculated the average expression within clusters C1–C7 for each gene, and calculated the range (max-min) of these values. We performed this procedure independently for cells in both data sets, and plotted the values in **Figure 3i**.

Differential expression testing to detect conserved cell-type markers. To identify cell-type markers that are conserved across data sets, we first performed a joint clustering of the data as described above. Then we conducted differential expression testing on each cell-type cluster for each data set independently using a Wilcoxon rank sum test, requiring a minimum 1.25-fold difference between the two groups of cells and expression in at least 10% of cells in both groups. We used the metap R package to combine *P*-values using the minimump method. For a detailed review of meta-analysis methods for differential expression, see Tseng, *et al.* 2012 (ref. 68). We visualize the top five markers, ranked by combined *P*-value, for each cluster in **Figure 4d**,**e**.

To identify markers differentially expressed between the beta cell populations, we used the same integrated differential expression procedure, but limited our analysis to only the two beta cell populations. We used the top 100 differentially expressed markers, ranked by integrated p-value, as input for gene ontology enrichment as performed using EnrichR[69].

Comparison to other batch correction methods. We compared our alignment method to both ComBat[48] and limma[49]. For each pair of data sets, we first combined the UMI count matrices and scaled and normalized the combined expression matrix. For the ComBat comparisons, we performed batch correction on the scaled and normalized gene expression data using the ComBat function from the sva R package, treating the data set as the batch. For the limma comparisons, we performed batch correction on the scaled and normalized gene expression data using the removeBatchEffect function from the limma R package, treating the data set as the batch. All other default parameters were left unchanged for both methods. We then performed a principle component analysis to identify sources of variation that accounted for a majority of the variation in the corrected data. For the PBMC, hematopoietic progenitor, and pancreas data sets we used the first 19, 18, and 21 PCs respectively to visualize with t-SNE and to calculate an alignment score. For the two multiple alignment examples of human pancreatic islet cells and PBMCs, we used the first 20 PCs.

Validation of pDC vs. DC response to IFNβ. Based on our analysis of the Kang *et al.* data set, we observed subsets of genes whose transcriptional response to IFNβ stimulation differed between plasmacytoid and conventional DCs. While these changes were observed at the single-cell level, we wished to validate these in bulk experiments. We therefore repeated the original experiment, where PBMCs from a healthy human donor (ALLCELLS) were cultured with RPMI medium supplemented with 10% FBS and stimulated for 6 h with IFNβ (100 U/ml, PBL Assay Science), with a subset of cells left unexposed to the stimulation as a control. After stimulation, we sorted pure populations of pDCs (20,000) and cDCs (60,000) based on the following panel of standard antibodies (from BioLegend and BD Pharmingen): *CD3* (*HIT3a*), *CD19* (*HIB19*), *CD56* (*HCD56*), *CD14* (*HCD14*), *HLA-DR* (*LN3*), *CD11c* (*Bu15*) and *CD123* (*7G3*). pDCs were defined as: $CD3^{-}CD19^{-}CD56^{-}CD14^{-}HLA^{-}DR^{+}CD11c^{-}CD123^{+}$.cDCs were defined as:$CD3^{-}CD19^{-}CD56^{-}CD14^{-}HLA\text{-}DR^{+}CD11c^{+}$. Following sorting, we extracted RNA using TRIzol (Invitrogen) and performed bulk RNA-seq (three technical replicates per sample) using a version of the SMART-Seq2 protocol as previously described[70,71].

Life Sciences Reporting Summary. Further information on experimental design is available in the **Life Sciences Reporting Summary**.

Software availability. Software used to generate all analyses in this manuscript is publicly available as an R package (https://cran.r-project.org/web/packages/Seurat/index.html) and included here as **Supplementary Software**.

Data availability. Full data sets and command lists to reproduce the integration of stimulated and control PBMCs, and four human pancreatic islet data sets are included as **Supplementary Data 1–3**. The published data used in this study can be accessed in the Gene Expression Omnibus under accession numbers GSE96583, GSE81682, GSE84133, GSE81076, GSE85241, GSE86469, and the ArrayExpress database under accession E-MTAB-5061.

56. Scialdone, A. *et al.* Resolving early mesoderm diversification through single-cell expression profiling. *Nature* **535**, 289–293 (2016).
57. Tirosh, I. *et al.* Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352**, 189–196 (2016).
58. Ilicic, T. *et al.* Classification of low quality cells from single-cell RNA-seq data. *Genome Biol.* **17**, 29 (2016).
59. Dudoit, S., Fridlyans, J. & Speed, T.P. Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Am. Stat. Assoc.* **97**, 77–87 (2002).
60. Tibshirani, R., Hastie, T., Narasimhan, B. & Chu, G. Class prediction by nearest shrunken centroids, with applications to DNA microarrays. *Stat. Sci.* **18**, 104–117 (2003).
61. Baglama, J. & Reichel, L. Augmented implicitly restarted lanczos bidiagonalization methods. *SIAM J. Sci. Comput.* (2005).
62. Giorgino, T. Computing and visualizing dynamic time warping alignments in R: the dtw package. *J. Stat. Softw.* **31**, 1–24 (2009).
63. Waltman, L. & Van Eck, N.J. A smart local moving algorithm for large-scale modularity-based community detection. *Eur. Phys. J. B* **86**, 1–33 (2013).
64. Van Der Maaten, L. Accelerating t-SNE using tree-based algorithms. *J. Mach. Learn. Res.* **15**, 1–21 (2014).
65. Richards, J. diffusionMap: diffusion map. (2014) at https://cran.r-project.org/package=diffusionMap.
66. Hastie, T. & Stuetzle, W. Principal curves. *J. Am. Stat. Assoc.* **84**, 502 (1989).
67. S original by Trevor Hastie R port by Andreas Weingessel. princurve: Fits a Principal Curve in Arbitrary Dimension. https://cran.r-project.org/package=princurve (2013).
68. Tseng, G.C., Ghosh, D. & Feingold, E. Comprehensive literature review and statistical considerations for microarray meta-analysis. *Nucleic Acids Res.* **40**, 3785–3799 (2012).
69. Kuleshov, M.V. *et al.* Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* **44**, W90–W97 (2016).
70. Mayer, C. *et al.* Developmental diversification of cortical inhibitory interneurons. *Nature* **555**, 457–462 (2018).
71. Picelli, S. *et al.* Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* **9**, 171–181 (2014).

nature research

Corresponding author(s): RAHUL SATIJA

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Please do not complete any field with "not applicable" or n/a. Refer to the help text for what text to use if an item is not relevant to your study. For final submission: please carefully check your responses for accuracy; you will not be able to make changes later.

▸ Experimental design

1. Sample size

Describe how sample size was determined.

Sample size was chosen based on the sample size in the publicly available datasets

2. Data exclusions

Describe any data exclusions.

We excluded single cells with low numbers of detected genes or molecules from downstream analysis. Cells expressing below 500 genes were removed from the PBMC (Figure 2) and inDrop (Figure 4) pancreatic islet datasets. Cells expressing less than 4,000 genes were removed from the SMART-Seq2 hematopoietic dataset (Figure 3).

3. Replication

Describe the measures taken to verify the reproducibility of the experimental findings.

As this was a computational study, we did not verify experimental reproducibility.

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

Sample randomization was based on the publicly available datasets.

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

Blinding was based on the publicly available datasets.

Note: all in vivo studies must report how sample size was determined and whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a | Confirmed

□ ☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)

□ ☒ A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly

□ ☒ A statement indicating how many times each experiment was replicated

□ ☒ The statistical test(s) used and whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.

□ ☒ A description of any assumptions or corrections, such as an adjustment for multiple comparisons

□ ☒ Test values indicating whether an effect is present
Provide confidence intervals or give results of significance tests (e.g. P values) as exact values whenever appropriate and with effect sizes noted.

□ ☒ A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)

□ ☒ Clearly defined error bars in all relevant figure captions (with explicit mention of central tendency and variation)

See the web collection on statistics for biologists for further resources and guidance.

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.	Software used to generate all analyses in this manuscript is publicly available as an R package (https://cran.r-project.org/web/packages/Seurat/index.html) and included here as Supplementary Software.

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a third party.	There are no restrictions

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).	No antibodies were used

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.	No cell lines were used
b. Describe the method of cell line authentication used.	No cell lines were used
c. Report whether the cell lines were tested for mycoplasma contamination.	No cell lines were used
d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.	No cell lines were used

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide all relevant details on animals and/or animal-derived materials used in the study.	No animals were used in this study

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.	No human participants were used in this study