Published in final edited form as:

Nat Biotechnol. 2018 June ; 36(5): 411–420. doi:10.1038/nbt.4096.

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,#]

[1]New York Genome Center, New York, NY 10013, USA

[2]Center for Genomics and Systems Biology, New York University, New York, NY 10003-6688, USA

Abstract

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 datasets remains challenging. Here, we introduce an analytical strategy for integrating scRNA-seq datasets based on common sources of variation, enabling the identification of shared populations across datasets and downstream comparative analysis. Implemented in our R toolkit Seurat (http://satijalab.org/seurat/), we use our approach to align scRNA-seq datasets of peripheral blood monocytes (PBMCs) 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 datasets, while boosting statistical power through integrated analysis. Our approach facilitates general comparisons of scRNA-seq datasets, potentially deepening our understanding of how distinct cell states respond to perturbation, disease, and evolution.

INTRODUCTION

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 datasets can be used for the discovery

[#]To whom correspondence should be addressed: rsatija@nygenome.org.

AUTHOR CONTRIBUTIONS

AB and RS conceived the research. AB, PH, and RS implemented the alignment procedure, performed all data analysis, and wrote the manuscript. EP performed the PBMC validation experiments, and PS performed the ddSeq experiments.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

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 datasets and command lists to reproduce the integration of stimulated and control PBMCs, and four human pancreatic islet datasets, 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.

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 datasets 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 datasets will confound these two disparate effects. Therefore, new methods are needed that can learn jointly between multiple datasets and facilitate comparative analysis downstream. Progress towards 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 datasets 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 datasets, motivated by techniques in computer vision designed for the alignment and integration of imaging datasets[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 datasets and can embed cells in a shared low-dimensional space. We identify and compare 13 aligned PBMC subpopulations under resting and interferon β (IFN-β)–stimulated conditions, align scRNA-seq datasets of complex tissues produced across multiple technologies, and jointly learn 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 datasets and learn deeper biological insight than would be possible from individual analysis. Our approach can be applied to datasets 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 datasets 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 datasets are generated in the presence or absence of a drug (Figure 1A). First, subpopulations must be aligned even if each has a unique drug

response. This 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 dataset. 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 datasets produced with different technologies (i.e. UMI vs. FPKM). Lastly, the process should not be targeted towards 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 datasets, and briefly consists of the following steps (Figure 1B–C; Online Methods). (i) It learns a shared gene correlation structure that is conserved between the datasets using canonical correlation analysis (CCA) (Figure 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 datasets, and can therefore be flagged for further analysis. (iii) It aligns the datasets into a conserved low-dimensional space, using non-linear ‘warping’ algorithms to normalize for differences in feature scale, in a manner that is robust to shifts in population density. (iv) It proceeds with an integrated downstream analysis, for example, identifying discrete subpopulations through clustering, or reconstructing continuous developmental processes (Figure 1C). (v) It performs comparative analysis on aligned subpopulations between the datasets, to identify changes in population density or gene expression (Figure 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 datasets—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 datasets that are maximally correlated, identifying shared correlation structures across datasets[28,29]. CCA has been used for multi-modal 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 multi-modal analysis[31,32], we apply CCA to identify relationships between single cells from different datasets based on the same set of genes. Effectively, we treat the datasets as multiple measurements of a gene-gene covariance structure, and search for patterns that are common to the datasets. We use CCA for pairwise integration of two datasets, and extend this to multi-set CCA (i.e. Multi-CCA)[33,34] for the integration of multiple datasets. 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 datasets 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 datasets as input. The procedure can consider any gene that is measured across all datasets, though we choose to focus only on genes that exhibit high single cell variation in at least one dataset (Online Methods). CCA identifies sets of canonical ‘basis’ vectors, embedding cells from each dataset in a low-dimensional space, such that the variation along these vectors (gene-level projections) is

highly correlated between datasets. We note that CCA is robust to affine transformations in the original data, and is unaffected by linear shifts in gene expression (for example, due to different normalization strategies).

Aligning basis vectors from CCA—CCA returns vectors whose gene-level projections are correlated between datasets, but not necessarily aligned. While linear transformations may be required to correct for global shifts in feature scale or normalization strategy, non-linear shifts may also be needed to correct for shifts in population density. We therefore align the CCA basis vectors between the datasets, 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 correlations 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 datasets. 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 dataset containing many distinct cell types in the presence and absence of perturbation. For example, a recent study examining the effects of interferon stimulation split 14,039 human PBMCs from 8 patients into two groups: one stimulated with interferon-beta (IFN-β) and a culture matched control[36]. Since all cells contain machinery to respond to IFN-β, stimulation results in a drastic but highly cell-type specific response. Consequently, a traditional joint analysis yields confusing results, as cells tend to cluster both by cell type but also by stimulation condition (Figure 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 note that results for this and all examples in the manuscript were robust to the exact choice of this parameter (Supplementary Figure 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 (Figure 2B). Our analysis revealed 13 cell clusters, which included the eight immune subsets described in the original publication, but separated additional populations as well (Figure 2C; Supplementary Data 1). In particular, we were able to separate naïve from memory T cells, plasmacytoid dendritic cells (pDCs) from conventional dendritic cells, and 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 Figure 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 (Figure 2D; Supplementary Figure 3).

Having aligned the datasets, 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; Figure 2E). However, each cell type exhibited significant 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) (Figure 2D). We noted that even canonical cell-type markers such as *CD14* were differentially expressed by monocytes (1.98-fold down-regulation; Supplementary Figure 3) in response to stimulation, highlighting the value of our non-targeted analyses in initially classifying cells.

Focusing in on the novel subsets we were able to resolve, we compared the IFN-β response program between naïve and memory CD4+ T cells, and observed nearly identical response signatures (Supplementary Figure 4A). However, while we observed a general correlation between pDC and DC responses, we also saw stark differences that reproduced across patients (Figure 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 (Figure 2G; Supplementary Figure 4B).

We externally validated these findings 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 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 datasets, and the bulk samples clustered directly with *in silico* averaged data from the same cell type (Supplementary Figure 5; Online Methods). 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 datasets. This is an important concern both for abundant populations, where absence in one dataset could throw off the integration, but also 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 dataset 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 identified by visualization and graph-based clustering despite being present in only one dataset (Supplementary Figure 6 A–C).

After removing stimulated erythroblasts, we observed that control erythroblast cells no longer separated in the integrated analysis, while other populations were unaffected (Supplementary Figure 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 dataset, PCA may be able to separate these cells, as we 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 dataset 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 datasets. 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 blended into the rest of the data (Supplementary Figure 6E). Notably, this test correctly did not flag similar cells when applied to the original analysis of the full datasets (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 dataset, enabling further characterization.

Integrated analysis of scRNA-seq technologies—We next examined two recent single cell RNA-seq profiles of hematopoietic progenitors from murine bone marrow, but produced with starkly different technologies. Nestorowa et al.[37] used the full-length SMART-Seq2 with deep sequencing (6,558 genes/cell) to profile 765 progenitors, while Paul et al.[38] applied the 3’ MARS-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 datasets. 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 dataset 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 datasets 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 datasets, but successfully identifying the contaminating NK population as 'non-overlapping' (Figure 3A–C; Supplementary Figure 7). After alignment, we mapped cells from the SMART-Seq2 dataset onto their closest cluster in the MARS-Seq dataset (Figure 3C–F; Online Methods; 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 multi-potent progenitors (LMPP))[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 datasets (Figure 3E–F, Supplementary Figure 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 dataset, but were challenging to detect in the original analysis due to reduced cell number.

Lastly, as both datasets identified developmentally heterogeneous populations during erythroid differentiation (broken into 7 stages in the MARS-Seq analysis), we applied diffusion maps to erythroid-committed cells to reconstruct a joint developmental trajectory (Figure 3G). We observed that this developmental path maintained the 'pseudotemporal' ordering of cells within both datasets (Supplementary Figure 9) and also aligned the two together, exhibiting nearly identical expression dynamics for canonical differentiation markers (Figure 3H). Extending this analysis globally, we observed that gene expression changes across the trajectory were largely conserved between datasets, 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] (Figure 3H–I). Therefore, our procedure can successfully align both discrete and transitioning populations, and enable the identification of gene-expression programs that are conserved or unique to individual datasets.

The ability to pool datasets of the same heterogeneous tissues has the potential to enable similar 'meta-analyses' for datasets produced across multiple labs and technologies. To further demonstrate this we include two additional examples (Supplementary Figure 10, 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 (10X Genomics 3' assay, 10X Genomics 5' assay, and the Illumina/BioRad ddSeq). In the first example, we identify 8 populations of endocrine, exocrine, and stellate cells, clearly defined by cell-type specific markers that were conserved across technologies (Supplementary Figure 10). Notably, we also identified a rare population (1%) of endothelial cells which were present in all datasets, 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 Figure 10E).

In the second example, pooling the datasets yielded 16,653 PBMCs, allowing us to identify 16 immune populations, including 6 T cell clusters (Supplementary Figure 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^{bright} NK cells[41] (Supplementary Figure 11E). As with previous examples, the rarity of these cells precludes their identification in any individual dataset, and they were not identified in a previous analysis of 68,000 PBMCs[4]. These integrated datasets provide the opportunity to perform 'meta-analyses' for differential expression across multiple technologies. As an example of this, we first performed individual 'within-dataset' 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 Figure 11F), including lowly expressed markers but functionally important chemokines (*CCL5*), transcriptional regulators (*RORA*), and surface receptors (*FCRL6*). Therefore, pooling cells produced by integrating different scRNA-seq technologies not only boosts the statistical power 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 (avg. 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 dataset was largely driven by species as well as by individual donor (Figure 4A, Supplementary Figure 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, identifying primarily small populations of immune cells (human mast cells and murine B cells) as 'non-overlapping' (Figure 4B, Supplementary Figure 12). We next performed a single integrated clustering analysis, identifying 10 clusters, corresponding to alpha, delta, gamma, acinar, stellate, ductal, epithelial, immune, and two subgroups of beta cells (Figure 4C; Supplementary Data 3, Supplementary Figure 9). Our clusters agreed overwhelmingly with the analyses from the independent datasets[42] (Supplementary Figure 13), though we did observe a low rate (5.8%) of discordant calls, particularly for cells with low UMI counts (Supplementary Figure 14A–B). We were also able to identify a subset of cell type markers that were conserved between human and mouse (Figure 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 up-regulated the expression of endoplasmic reticulum (ER) stress genes (*HERPUD1*; *GADD45A*) in both species (Figure 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 dataset. In contrast, our integrated analyses reveal a conserved set of markers that are strikingly enriched for regulators of ER stress response to unfolded proteins[43,44] (Figure 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 up-regulated 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[48]. To evaluate each technique, we designed an 'alignment score', which examines the local neighborhood of each cell after alignment (Online Methods). When datasets are well aligned, this local neighborhood will consist equally of cells from both datasets, enabling us to quantify the success of each procedure with a score ranging from 0 to 1.

On the five datasets presented here, we benchmarked Seurat's performance against ComBat[49] and limma[50] (Figure 5). In each case, as can be visualized by tSNE 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 datasets (i.e 'batch effect') significantly outweighed differences between cell types ('biology'), as in cross-species integration. However, when we attempted to align datasets from different tissues as a negative control, we observed poor results and low alignment scores, even when cells are not automatically classified as 'non-overlapping' (Supplementary Figure 15).

DISCUSSION

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

Dataset integration represents a key step in a general framework for case/control studies performed with single-cell resolution. As new datasets 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[15,42,51], which aims to define all human cell types by integrating data generated across diverse single-cell 'omics' approaches, to combine datasets produced across many labs and technologies. Recent benchmarking studies of diverse scRNA-seq[52,53] 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 invaluable comparative tools for studies utilizing mouse models of human disease, potentially enabling the identification of human correlates of pathogenic populations discovered in mouse (or vice versa). Furthermore, new datasets 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 to similar insights towards our understanding of cellular diversity.

Lastly, we note many challenges that future methods will address in extending this work. Although our procedure can robustly integrate multiple datasets with overlapping and non-overlapping populations, future datasets that consist of tens to hundreds batches with dramatically varying sizes and non-overlapping populations will likely require new methods. We also note that examples in this manuscript, including datasets with tens of thousands of cells, run in less than half an hour on a standard laptop computer, but new datasets extending to millions of cells may require advanced computation, subsampling, or newly optimized techniques for integration. Lastly, whereas we focus here on alignment of sequencing-based datasets, the recent invention of spatially resolved or *in-situ* methods for transcriptomic profiling[54–56] raise the potential for integration with scRNA-seq datasets, including the ability to extend previous efforts to spatially resolve scRNA-seq data[9,10] towards an unsupervised procedure generalizable to any tissue.

ONLINE METHODS

The Seurat alignment procedure is designed to integrate single-cell RNA sequencing data (scRNA-seq) across distinct datasets. 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 datasets (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 datasets and for the integration of the four scRNA-seq datasets of human pancreatic islet cells (produced with four different plate-based technologies CelSeq, CelSeq2, Fluidigm C1, SmartSeq2) as supplementary code.

1 Single cell dataset pre-processing

For all single cell analysis, we performed the same initial normalization. Gene expression values for each cell were scaled by the total number of transcripts and multiplied by 10,000. These scaled expression data 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 dataset (Nestorawa 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 Figure 4 in the original publication, used in our Figure 3). We obtained a batch-corrected UMI count matrix for the MARS-Seq dataset[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 dataset had been previously filtered to remove cells with less than 500 detected UMI for a total of 2,686 single cells.

Both datasets contain cycling progenitors, and heterogeneity between cell cycle stages for these cells has been previously been shown to confound developmental analyses. Therefore, independently for both datasets, we first assigned a cell cycle score to each cell using the PCA method[57] on a previously annotated list of cell cycle genes[58]. 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 dataset, prior to CCA.

Joint clustering across species—We obtained UMI count matrices for the human and mouse inDrops datasets (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 prior to CCA. We considered all homologous genes with identical gene names between the human and mouse datasets, 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 datasets—For the four human pancreas datasets, 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 datasets across technologies—For the three human PBMC datasets, we obtained gene expression matrices from 10X 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 (2 cartridges) with an expected yield of 2,400 cells in total. Sequencing libraries were prepared according to manufacturer's instructions and sequenced on 2 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 dataset, in keeping with previous reports[59], and regressed out mitochondrial heterogeneity from each dataset prior to running CCA.

Gene set selection—Although the alignment procedure can utilize any gene that is measured with non-zero variance in all datasets, we focused on genes that were highly variable in one or both datasets. We identified these genes by calculating the dispersion (variance to mean ratio) for all genes in each dataset 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 datasets.

2 Calculation of canonical correlation vectors

Standard canonical correlation analysis was designed to find projections that maximize correlation between two vectors (datasets 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 datasets 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 $u^T X$ and $v^T Y$ is maximized[28].

$$max_{u,v} u^T X^T Y v \qquad \text{subject to} \quad u^T X^T X u \leq 1, \qquad v^T Y^T Y v \leq 1 \qquad (1)$$

To apply this in the context of scRNA-seq, let $X_{g,c}$ be a gene expression matrix of genes $g_1, g_2, \ldots, g_n$ by cells $c_1, c_2, \ldots, c_n$ and $Y_{g,d}$ be a gene expression matrix of the same genes $g_1,$

$g_2, \ldots, g_n$ by cells $d_1, d_2, \ldots, d_p$ In many scRNA-seq experiments, the number of genes of interest that are shared between the two datasets is often much smaller than the total number of cells that were measured ($g \ll c + d$). 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 dataset as diagonal, a solution that has demonstrated promising results in other high-dimensional problems[60,61]. We substitute the identity matrix for X^TX and Y^TY to arrive at equation 2.

$$max_{u,v} u^T X^T Y v \qquad \text{subject to} \quad \left\|u\right\|^2 \leq 1, \quad \left\|v\right\|^2 \leq 1 \qquad (2)$$

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

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

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

$$K = \Gamma \Lambda \Delta^T \qquad (3)$$

where

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

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

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

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[62]. 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 datasets. 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 datasets $X_1, \ldots, X_N$, the goal is to find projection vectors $W = w_1, \ldots, w_N$ that maximize:

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

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 dataset. By substituting the identity matrix for each $X_n^T X_n$ we arrive at

$$\max_{w_1, \ldots, w_N} \sum_{i<j} w_i^T X_i^T X_j w_j \quad \text{subject to } \left\| w_n \right\|^2 \leq 1 \quad \forall n \quad (4)$$

To construct the canonical correlation vectors, we first standardized each X_n such that $\forall_c$ $\sum X[,c]/g = 0$, $var(X[,c]) = 1$. 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**
    [T, Λ, Δ]←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**

$$\text{update } w_n[cc] \leftarrow \frac{\left(X_n^T \Sigma_{k \neq n} X_k w_k[cc]\right)}{\left\|\left(X_n^T \Sigma_{k \neq n} X_k w_k[cc]\right)\right\|_2}$$

  $o2 \leftarrow$ 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.

3 Identification of rare non-overlapping subpopulations

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

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 prior to 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 Figure 6D–E).

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 dataset 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 $\widetilde{X}$ and $\widetilde{Y}$.

$$\widetilde{X} = X^T D$$

$$\widetilde{Y} = Y^T E$$

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

$$\widehat{X} = D\widetilde{X}^T$$

$$\widehat{Y} = EY^T$$

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

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

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 $\widehat{Y}$.

$$\widetilde{X} = X^T F$$

$$\widetilde{Y} = Y^T G$$

$$\widehat{X} = F\widetilde{X}^T$$

$$\widehat{Y} = G\widetilde{Y}^T$$

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).

$$\eta = \frac{\sigma_{CCA}}{\sigma_{PCA}} \quad (6)$$

Empirically, we applied a threshold of 0.5 uniformly in all datasets, 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 dataset. These included terminally differentiated NK cells in the MARS-Seq hematopoietic progenitors, B and T cells in the murine pancreatic islet dataset and mast cells in the human pancreatic islet dataset (Supplementary Figures 7 and 12).

However, identifying a single value for this threshold was challenging for abundant populations that were specific to a single dataset, for example, in our negative control experiments where we aligned datasets that have negligible biological similarity. However, in each of these cases (Supplementary Figure 15), even though we did not initially identify these cells as non-overlapping, our procedure did not artificially align cells from these datasets together. While we anticipate that new methods that can robustly identify non-overlapping subpopulations prior to 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.

4 Aligning canonical correlation vectors

After CCA, CC vectors are by definition correlated, but not necessarily aligned between datasets. 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 that drive the shared sources of variation in both datasets by finding those genes whose expression robustly correlates with each projection vector in both datasets. For this, we use the biweight midcorrelation (bicor), a median based similarity metric.

For each canonical correlation vector $i, \ldots, k$

$$\zeta = \forall_g \min\left(bicor\left(X[g,], u_i\right), bicor\left(Y[g,], v_i\right)\right)$$

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 less than 10 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 datasets, in order to be correctly aligned, will depend on the scale and sequencing depth of each dataset. Datasets 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

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

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.

$$\Phi'_i = \frac{\Phi_i - Q_{\Phi_i}\left(2.5\right)}{Q_{\Phi_i}(97.5) - Q_{\Phi_i}\left(2.5\right)} \qquad \Theta'_i = \frac{\Theta_i - Q_{\Theta_i}\left(2.5\right)}{Q_{\Theta_i}(97.5) - Q_{\Theta_i}\left(2.5\right)}$$

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[63] with default parameters. Traditionally used to find an alignment between two time series[35], DTW effectively aligns each cell in the smaller dataset to the cell with the most similar metagene expression in the larger dataset, while maintaining the relative ordering of cells within each dataset. 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.

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

A key feature of DTW in the alignment procedure is the non-linear warping of each metagene vector. These compressions and stretches correspond to potential shifts in population density across datasets. We then apply an identical warping to the canonical correlation vectors, mapping the CC values from both datasets 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 datasets.

Alignment of multiple canonical correlation vectors—Extending the alignment procedure to multiple datasets follows naturally from the two dataset case. We first choose a reference dataset, which we set by default to be the dataset 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 dataset onto a common aligned space, defined by the "reference" dataset.

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 dataset 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.

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

For the reference dataset, we take the average of τ across all pairwise calculations. For each dataset, 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 dataset is robust to the exact choice of this parameter ∓ 5CCs (Supplementary Figure 1).

Calculating an alignment score—While our tSNE plots provide a visual representation of the overlap between datasets after alignments, we sought to develop a quantitative metric to ask how well any group of datasets are aligned. We calculated an alignment score as follows. First, we randomly downsample the datasets to have the same number of cells as the smallest dataset. 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 dataset and average this over all cells to obtain $\bar{x}$. If the datasets are well-aligned, we would expect that each cells' nearest neighbors would be evenly shared across all datasets. 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 dataset cells and scale to range from 0 to 1.

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

5 Integrated Analysis Across Datasets

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 dataset 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[64]. For each of the five of datasets, 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 (tSNE) algorithm[65].

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 likelihood-ratio test for differential expression implemented in Seurat as FindAllMarkers. The same 20 CCs were used as input for visualization via tSNE.

To understand global correlations between IFNβ responses for each cell type (Figure 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 431 genes that exhibited at least a two-fold change in response to stimulation for at least one of the 13 clusters.

Identification of hematopoietic progenitor populations—To identify the hematopoietic progenitor populations present in both the SMART-Seq2 dataset and the MARS-Seq dataset, we used the first 10 aligned canonical correlation vectors as input to calculate the cell-cell distance matrix, SNN, and visualization via tSNE.

We then mapped cells from the SMART-Seq2 dataset to one of the clusters originally identified in the MARS-Seq dataset. In principle, we could simply map each SMART-Seq2 cell to the cluster identity of its nearest neighbor in the MARS-Seq dataset. 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 10 aligned CC embeddings using the SLM algorithm via FindClusters with default parameters, revealing 10 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 percent 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[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 10 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 tSNE 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 Figure 13).

Identification of pancreatic islet subtypes in the four human datasets—In order to identify populations of islet subtypes across the four human pancreas datasets, we ran FindClusters using the first 10 CC embeddings with a resolution parameter of 0.4. This gave 8 distinct clusters corresponding to alpha, beta, ductal, acinar, delta, gamma, stellate, and endothelial populations. The same 10 aligned CCs were used for tSNE 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. 16 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 ($CD56^{bright}$ and $CD56^{dim}$).

Construction of joint erythroid developmental trajectories—We first took a subset of the combined hematopoietic progenitor dataset that included all cells originally assigned to the first seven clusters in the MARS-Seq dataset and cells that mapped to one of those clusters from the SMART-Seq2 dataset. We then built a diffusion map using the first 10 aligned CC embeddings using the diffuse function from the diffusionMap R package (Richards, 2014)[66] with epsilon parameter set to 9, corresponding to the median distance to the 0.05*n nearest neighbor. Next, we fit a principal curve[67] through the first two diffusion map coordinates using the principal.curve function from the princurve R package with default parameters[68]. 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 datasets, 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 datasets, 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 datasets, 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 dataset 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[69]. 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[70].

Comparison to other batch correction methods—We compared our alignment method to both ComBat[49] and limma[50]. For each pair of datasets, 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 dataset 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 dataset 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 datasets we used the first 19, 18 and 21 PCs respectively to visualize with tSNE 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. dataset, 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 a healthy human donor (ALLCELLS) were cultured with RPMI medium supplemented with 10% FBS and stimulated for 6 hours with IFNβ (100U/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−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[71,72].

A Life Sciences Reporting Summary is available

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

We thank members of the Satija lab, 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 datasets. This work was supported by an NIH New Innovator Award (1DP2HG009623–01) to RS and an NSF Graduate Fellowship (DGE1342536) to AB.

REFERENCES

1. Klein AM et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell 161, 1187–1201 (2015). [PubMed: 26000487]
2. Zilionis R et al. Single-cell barcoding and sequencing using droplet microfluidics. Nat. Protoc 12, 44–73 (2017). [PubMed: 27929523]
3. Macosko EZ et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell 161, 1202–1214 (2015). [PubMed: 26000488]
4. Zheng GXY et al. Massively parallel digital transcriptional profiling of single cells. Nat. Commun 8, 14049 (2017). [PubMed: 28091601]
5. Shekhar K et al. Comprehensive Classification of Retinal Bipolar Neurons by Single-Cell Transcriptomics. Cell 166, 1308–1323 (2016). [PubMed: 27565351]
6. Villani A-C et al. Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. Science (80-.). 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, 381–6 (2014). [PubMed: 24658644]
8. Welch J, Hartemink A & Prins JF SLICER: Inferring Branched, Nonlinear Cellular Trajectories from Single Cell RNA-seq Data. Genome Biol. 17, 1–15 (2016). [PubMed: 26753840]
9. Satija R, Farrell J. a, Gennert D, Schier AF & Regev A Spatial reconstruction of single-cell gene expression data. Nat. Biotechnol 33, (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). [PubMed: 25867922]
11. DeLaughter DM et al. Single-Cell Resolution of Temporal Gene Expression during Heart Development. Dev. Cell 39, 480–490 (2016). [PubMed: 27840107]
12. Bendall SC et al. Single-cell Trajectory Detection Uncovers Progression and Regulatory Coordination in Human B Cell Development. Cell 157, 714–725 (2014). [PubMed: 24766814]
13. Blakeley P et al. Defining the three cell lineages of the human blastocyst by single-cell RNA-seq. Development 142, 3613–3613 (2015). [PubMed: 26487783]
14. Johnson MB et al. Single-cell analysis reveals transcriptional heterogeneity of neural progenitors in human cortex. Nat. Neurosci 18, 637–646 (2015). [PubMed: 25734491]
15. Regev A et al. The Human Cell Atlas. bioRxiv (2017). doi:10.1101/121202
16. Kharchenko PV, Silberstein L & Scadden DT Bayesian approach to single-cell differential expression analysis. Nat. Methods 11, 740–742 (2014). [PubMed: 24836921]
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). [PubMed: 26653891]
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). [PubMed: 28263960]
19. Kiselev VY et al. SC3: consensus clustering of single-cell RNA-seq data. Nat. Methods 14, (2017).
20. Lin P, Troup M & Ho JWK CIDR: Ultrafast and accurate clustering through imputation for single-cell RNA-seq data. Genome Biol. 18, 59 (2017). [PubMed: 28351406]
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 GM, Zhang J, Pachter L & Tse DN Fast and accurate single-cell RNA-Seq analysis by clustering of transcript-compatibility counts. Genome Biol. 17, 1–14 (2016). [PubMed: 26753840]
23. Xu C & Su Z Identification of cell types from single-cell transcriptomes using a novel clustering method. Bioinformatics 1–8 (2015).
24. Lei Z, Bai Q, He R & Li SZ Face shape recovery from a single image using CCA mapping between tensor spaces. 26th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR (2008). doi: 10.1109/CVPR.2008.4587341
25. Zhou F & de la Torre F Canonical time warping for alignment of human behavior. Adv. Neural Inf. Process. Syst. 22 (Proceedings NIPS) 1–9 (2009). doi:10.1103/PhysRevB.72.205311
26. Wang C & Mahadevan S Heterogeneous Domain Adaptation Using Manifold Alignment. in International Joint Conference on Artificial Intelligence 1541–1546 (2010). doi:10.3901/JME. 2016.09.065
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 DR, Szedmak S & Shawe-Taylor J Canonical Correlation Analysis: An Overview with Application to Learning Methods. Neural Comput. 16, 2639–2664 (2004). [PubMed: 15516276]
30. Witten DM, Tibshirani R & Hastie T A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics 10, 515–534 (2009). [PubMed: 19377034]

31. Lê Cao K-A, Martin PG, Robert-Granié C & Besse P Sparse canonical methods for biological data integration: application to a cross-platform study. BMC Bioinformatics 10, 34 (2009). [PubMed: 19171069]
32. Waaijenborg S, Verselewel de Witt Hamer PC & Zwinderman AH Quantifying the Association between Gene Expressions and DNA-Markers by Penalized Canonical Correlation Analysis. Stat. Appl. Genet. Mol. Biol 7, (2008).
33. Kettenring J Canonical analysis of several sets of variables. Biometrika 58, 433–451 (1971).
34. Nielsen AA Multiset canonical correlations analysis and multispectral, truly multitemporal remote sensing data. IEEE Trans. Image Process. 11, 293–305 (2002). [PubMed: 18244632]
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 HM et al. Multiplexing droplet-based single cell RNA-sequencing using natural genetic barcodes. bioRxiv (2017).
37. Nestorowa S et al. A single-cell resolution map of mouse hematopoietic stem and progenitor cell differentiation. Blood 128, 20–32 (2016).
38. Paul F et al. Transcriptional Heterogeneity and Lineage Commitment in Myeloid Progenitors. Cell 163, 1663–1677 (2015). [PubMed: 26627738]
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). [PubMed: 15851035]
40. Lacar B et al. Corrigendum: Nuclear RNA-seq of single neurons reveals molecular signatures of activation. Nat. Commun 8, 15047 (2017). [PubMed: 28303884]
41. Poli A et al. CD56bright natural killer (NK) cells: An important NK cell subset. Immunology 126, 458–465 (2009). [PubMed: 19278419]
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 (2016). [PubMed: 27667365]
43. Scheuner D & Kaufman RJ The unfolded protein response: A pathway that links insulin demand with β-cell failure and diabetes. Endocr. Rev 29, 317–333 (2008). [PubMed: 18436705]
44. Walter W, Sanchez-Cabo F & Ricote M GOplot: An R package for visually combining expression data with functional analysis. Bioinformatics 31, 2912–2914 (2015). [PubMed: 25964631]
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, 1365–77 (2004). [PubMed: 14729979]
46. Papa FR Endoplasmic reticulum stress, pancreatic β-cell degeneration, and diabetes. Cold Spring Harb. Perspect. Med 2, 1–17 (2012).
47. Conesa A et al. A survey of best practices for RNA-seq data analysis. Genome Biol. 17, 13 (2016). [PubMed: 26813401]
48. Shekhar K et al. Comprehensive Classification of Retinal Bipolar Neurons by Single-Cell Transcriptomics. Cell 166, 1308–1323.e30 (2016). [PubMed: 27565351]
49. Johnson WE, Li C & Rabinovic A Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 8, 118–127 (2007). [PubMed: 16632515]
50. Ritchie ME et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 43, e47 (2015). [PubMed: 25605792]
51. Lake BB et al. Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. Science. 357, 352–357 (2015).
52. Ziegenhain C et al. Comparative Analysis of Single-Cell RNA Sequencing Methods. Mol. Cell 65, 631–643.e4 (2017). [PubMed: 28212749]
53. Svensson V et al. Power analysis of single-cell RNA-sequencing experiments. Nat. Methods 14, 381–387 (2017). [PubMed: 28263961]
54. Junker JP et al. Genome-wide RNA Tomography in the Zebrafish Embryo. Cell 159, 662–675 (2014). [PubMed: 25417113]
55. Lee JH et al. Highly Multiplexed Subcellular RNA Sequencing in Situ. Science. 343, 1360–1363 (2014). [PubMed: 24578530]

56. Stahl PL et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. Science. 353, 78–82 (2016). [PubMed: 27365449]
57. Scialdone A et al. Resolving early mesoderm diversification through single-cell expression profiling. Nature 535, 289–293 (2016). [PubMed: 27383781]
58. Tirosh I et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science 352, 189–196 (2016). [PubMed: 27124452]
59. Ilicic T et al. Classification of low quality cells from single-cell RNA-seq data. Genome Biol. 17, 29 (2016). [PubMed: 26887813]
60. Dudoit S, Fridlyans J & Speed TP Comparison of discrimination methods for the classification of tumors using gene expression data. J. Am. Stat. Assoc 97, 77–87 (2002).
61. 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).
62. Baglama J & Reichel L Augmented Implicitly Restarted Lanczos Bidiagonalization Methods. SIAM J. Sci. Comput (2005).
63. Giorgino T Computing and Visualizing Dynamic Time Warping Alignments in R: The dtw Package. J. Stat. Softw 31, 1–24 (2009).
64. Waltman L & Van Eck NJ A smart local moving algorithm for large-scale modularity-based community detection. Eur. Phys. J. B 86, 1–33 (2013).
65. Van Der Maaten, L.. Accelerating t-SNE using Tree-Based Algorithms. J. Mach. Learn. Res 15, 1–21 (2014).
66. Richards J diffusionMap: Diffusion map. (2014). at <https://cran.r-project.org/package=diffusionMap>
67. Hastie T & Stuetzle W Principal Curves. J. Am. Stat. Assoc 84, 502 (1989).
68. S original by Trevor Hastie R port by Andreas Weingessel<Andreas.Weingessel@ci.tuwien.ac.at>. princurve: Fits a Principal Curve in Arbitrary Dimension. (2013). at <https://cran.r-project.org/package=princurve>
69. Tseng GC, Ghosh D & Feingold E Comprehensive literature review and statistical considerations for microarray meta-analysis. Nucleic Acids Res. 40, 3785–3799 (2012). [PubMed: 22262733]
70. Kuleshov MV et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res. 44, W90–W97 (2016). [PubMed: 27141961]
71. Mayer C et al. Developmental diversification of cortical inhibitory interneurons. bioRxiv (2017).
72. Picelli S et al. Full-length RNA-seq from single cells using Smart-seq2. Nat. Protoc 9, 171–181 (2014). [PubMed: 24385147]

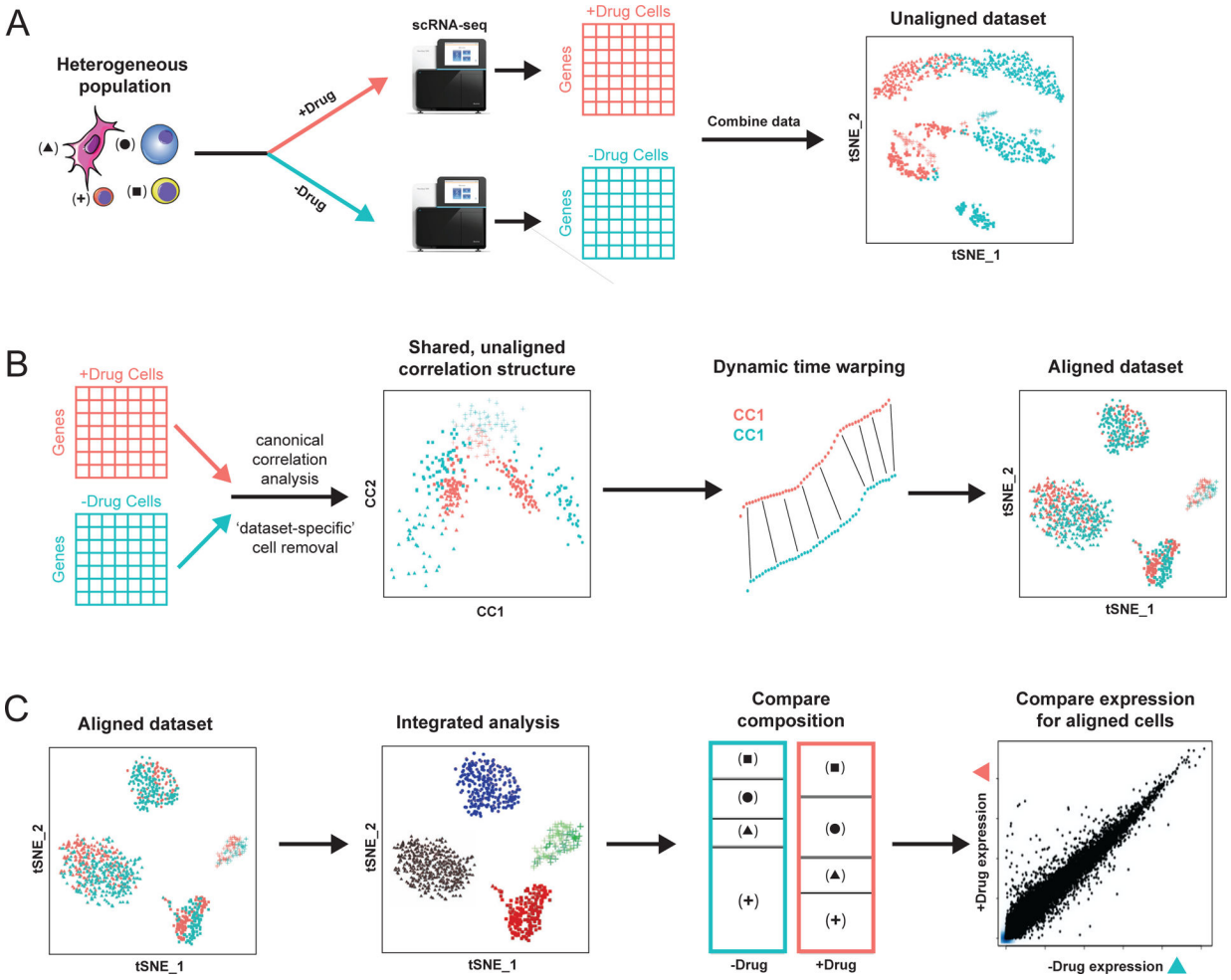

Figure 1. Overview of Seurat alignment of single cell RNA-seq datasets

(A) Toy example of heterogeneous populations profiled in a case/control study after drug treatment. Cells across four types are plotted with different symbols, while stimulation condition is encoded by color. In a standard workflow, cells often cluster both by cell type and stimulation condition, creating challenges for downstream comparative analysis. **(B)** The Seurat alignment procedure uses canonical correlation analysis to identify shared correlation structures across datasets, and aligns these dimensions using dynamic time warping. After alignment, cells are embedded in a shared low-dimensional space (visualized here in 2D with tSNE). **(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.

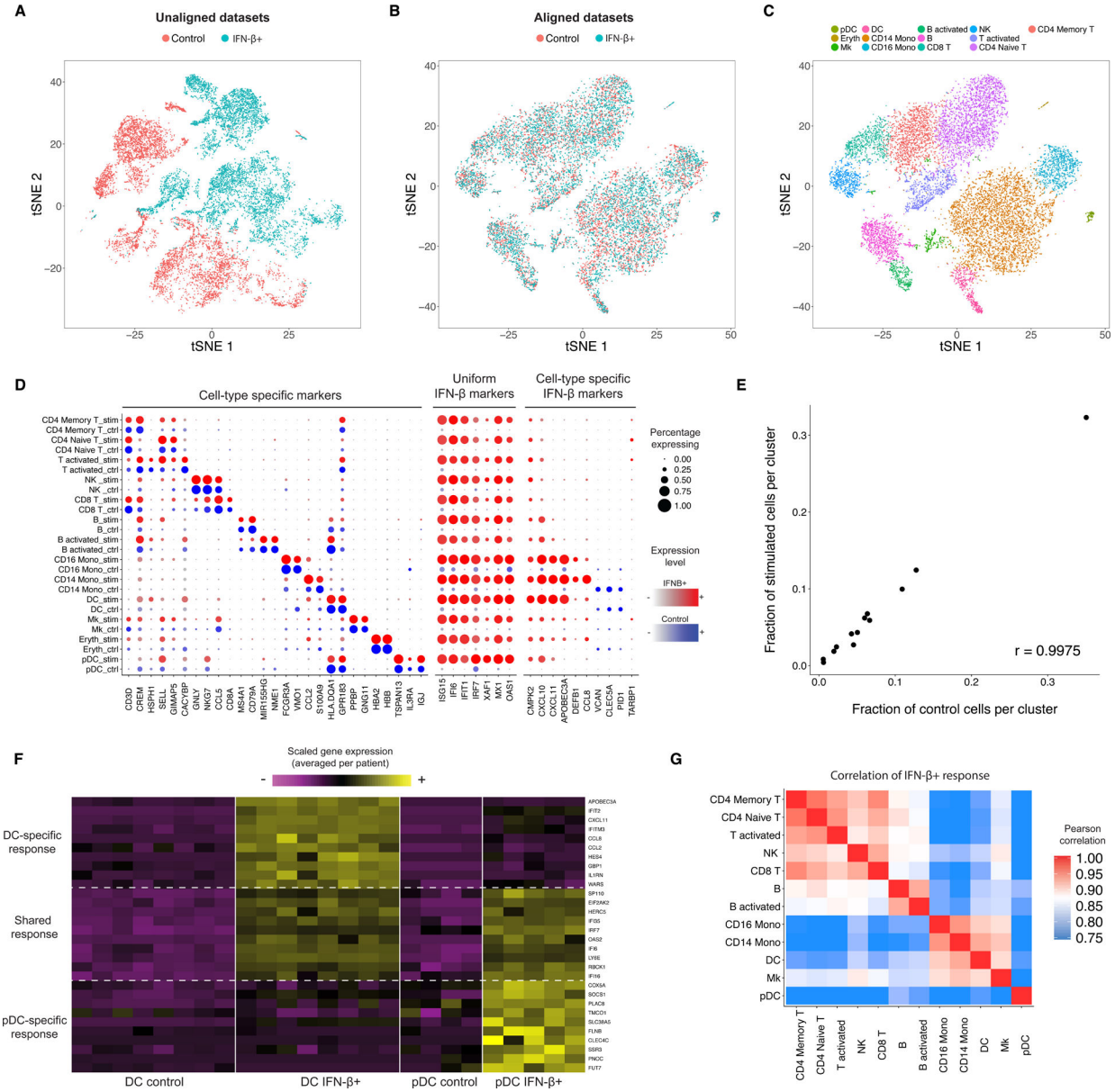

Figure 2. Integrated analysis of resting and stimulated PBMC

(A-C) tSNE plots of 14,039 human PBMCs split between control and IFN-β-stimulated conditions, prior to (A) and post (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 reflects the average expression level within each cluster. **(E)** The fraction of cells (median across 8 donors) falling in each cluster (n = 13 clusters) for stimulated and unstimulated cells. **(F)** Examples of heterogeneous responses to IFN-β between conventional and plasmacytoid dendritic cells (global analysis shown in Supplementary Figure 4B). Each

column represents the average 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 with difference > ln(2) between resting and stimulated) of cell-type specific responses to IFN-β (individual correlations for T and DC subsets shown in Supplementary Figure 4A–B). Cells from myeloid and lymphoid lineages show highly correlated responses, but plasmacytoid dendritic cells exhibit a unique IFN-β response.

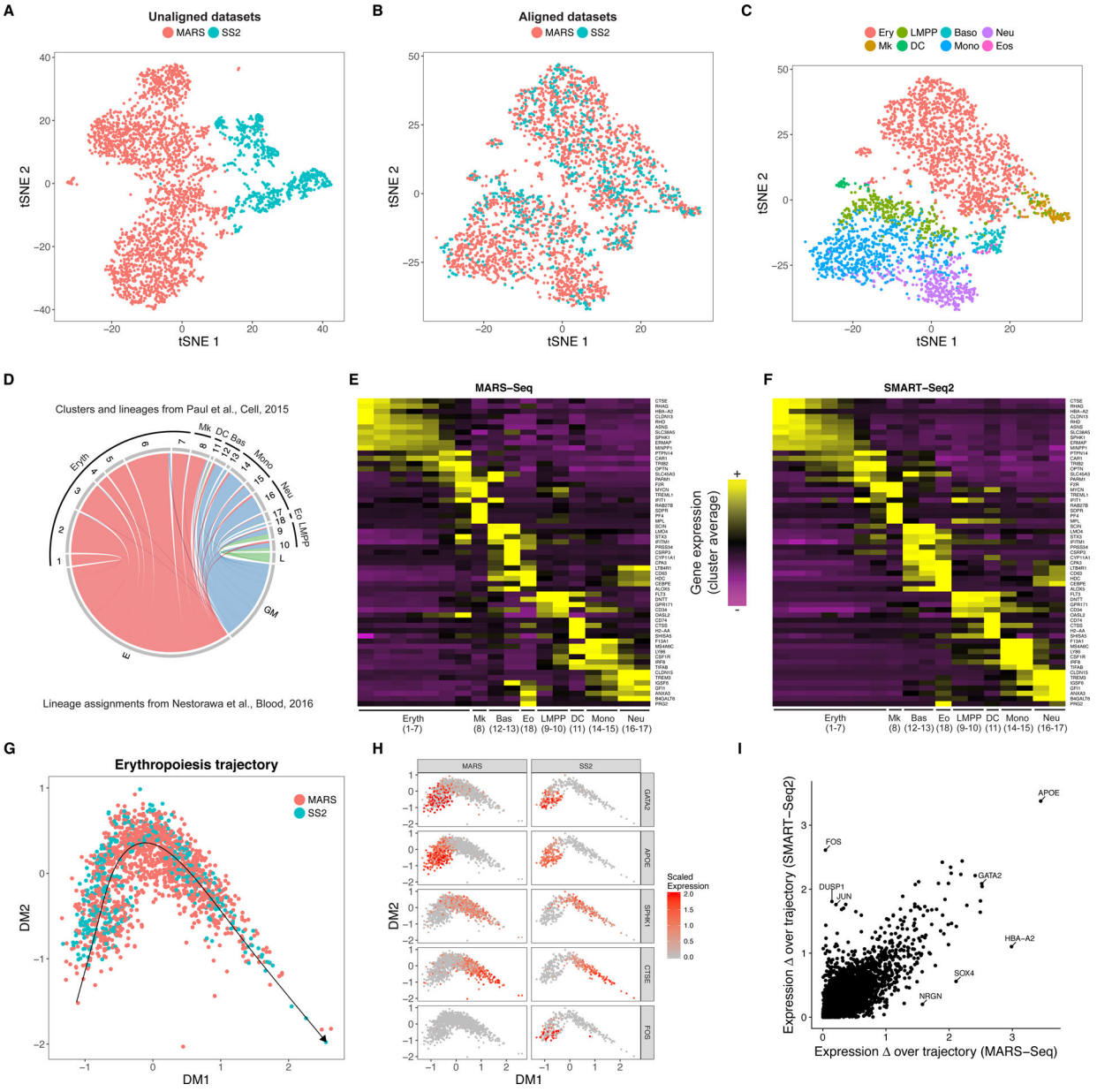

Figure 3. Comparative analysis of mouse hematopoietic progenitors across scRNA-seq technologies

(A-C) tSNE plots of 3,451 hematopoietic progenitor cells from murine bone marrow sequenced using MARS-seq (2,686) and SMART-Seq2 (765), prior to (A) and post (B-C) alignment. After alignment, cells group together based on shared progenitor type irrespective of sequencing technology. **(C-D)** Cells from the SMART-Seq2 dataset were mapped onto the closest MARS-Seq cluster and associated lineage (from Paul et al.). (C) tSNE plot of cells colored by assigned lineage. (D) Mapping correspondence between SMART-Seq2 lineage assignments (from Nestorawa et al.) and MARS-Seq clusters. **(E-F)** Heatmaps showing lineage- specific gene expression patterns in MARS-Seq and SMART-Seq2 datasets. Each column represents average expression after cells are grouped either by 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 datasets reveals 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 datasets.

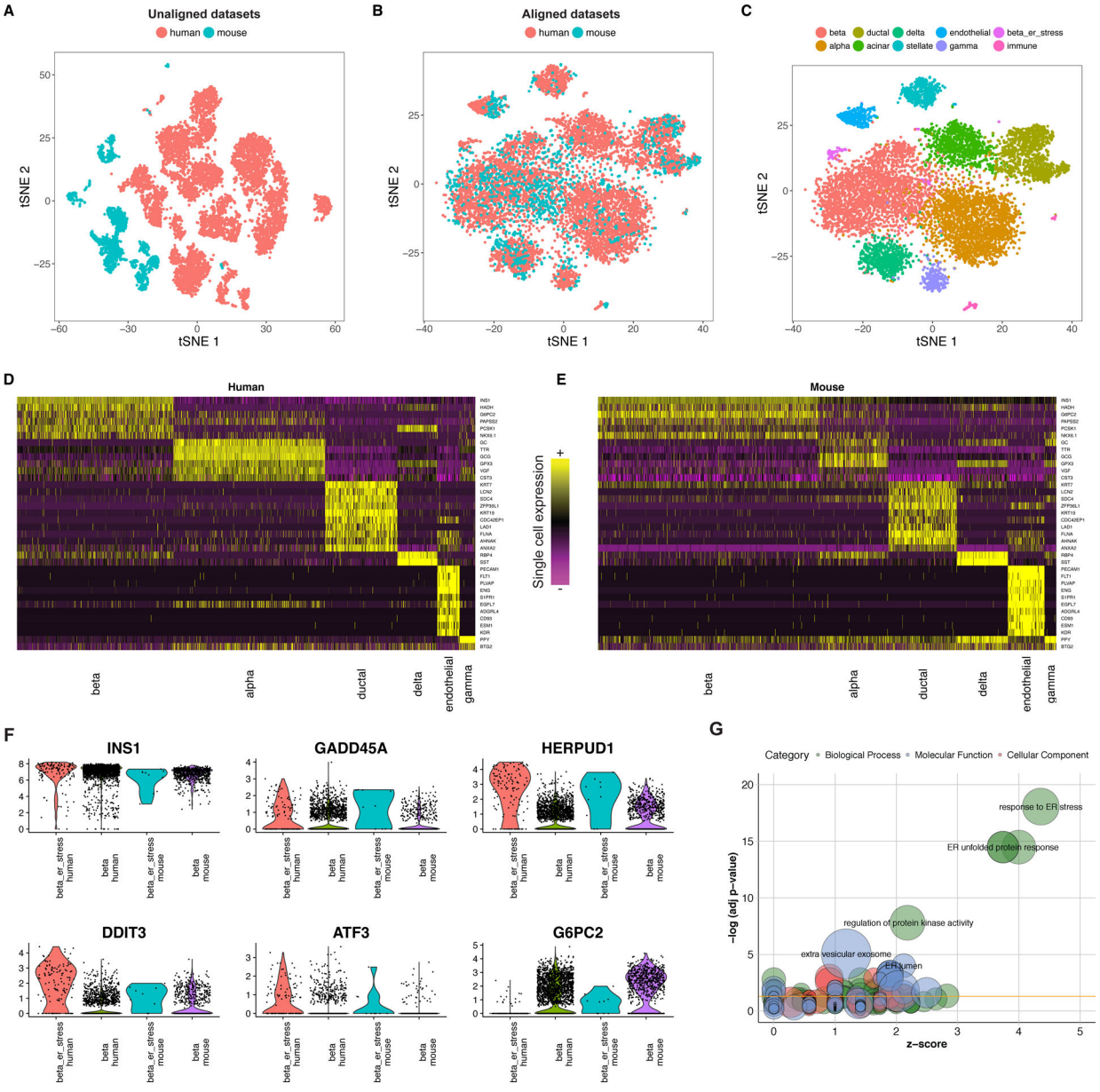

Figure 4. Joint identification of cell types across human and mouse islet scRNA-seq atlases

(A-C) tSNE plots of 10,191 pancreatic islet cells from human (n = 8,424 cells) and mouse (n = 1,767 cells) donors, prior to (A) and post (B) alignment. After alignment, cells group across species based on shared cell type, allowing for a joint clustering (C) to detect 10 cell populations. **(D-E)** Unsupervised identification of shared cell-type markers between human and mouse. 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 up-regulated 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.

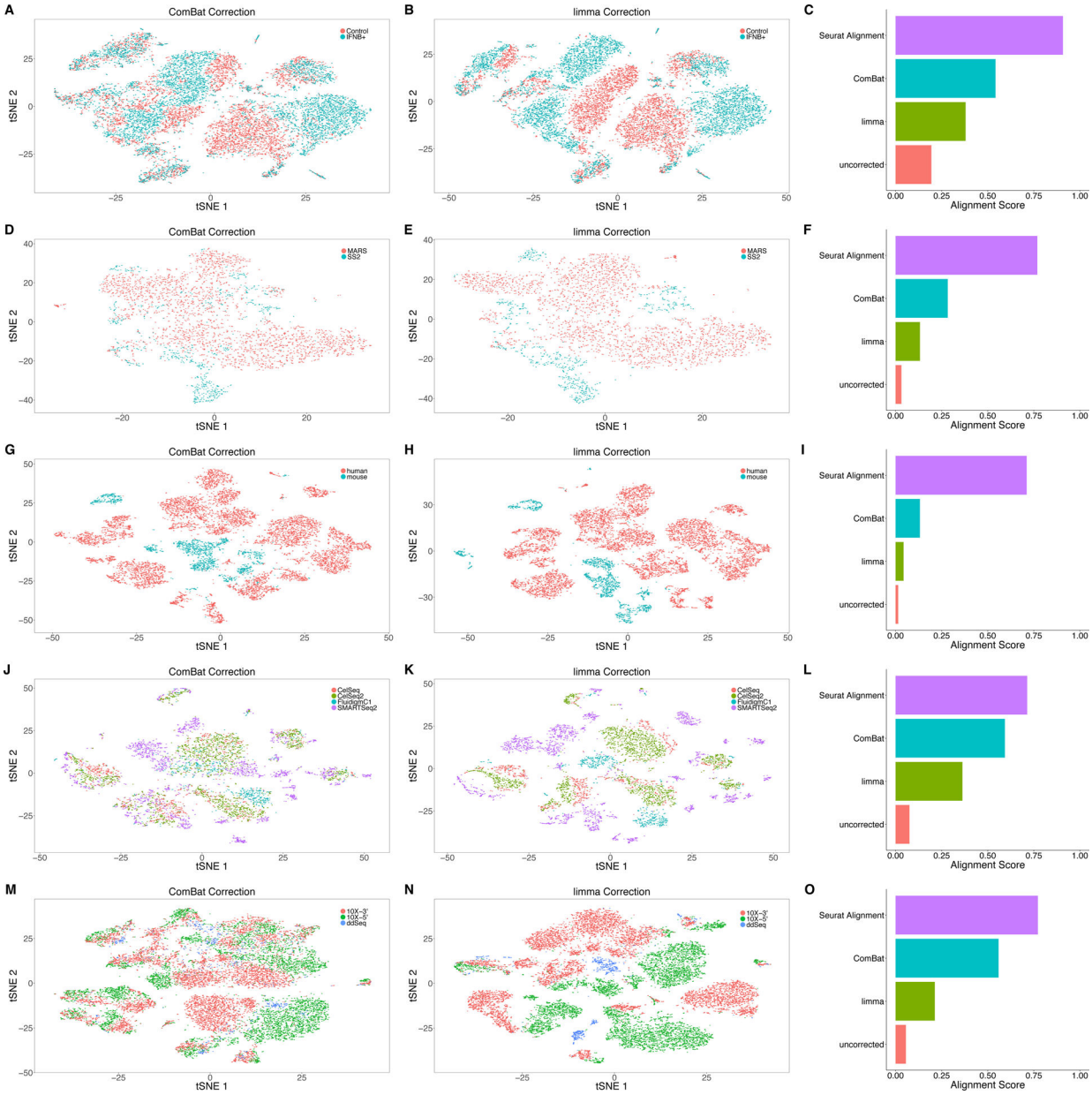

Figure 5. Benchmarking alignment and batch correction methods

(A, D, G, J, M) tSNE plots for the PBMC dataset (n = 14,039 cells) (A), hematopoietic progenitor cell dataset (n = 3,451 cells) (B), pancreatic islet cell dataset (n = 10,306 cells) (C), multiple human pancreatic islet cell datasets (n = 6,224 cells) (J), and multiple PBMC datasets (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 datasets from different tissues are shown in Supplementary Figure 15.