

Dictionary learning for integrative, multimodal and scalable single-cell analysis

Received: 8 March 2022

Accepted: 28 March 2023

Published online: 25 May 2023

 Check for updates

Yuhan Hao^{1,2}, Tim Stuart^{1,2}, Madeline H. Kowalski^{2,3}, Saket Choudhary^{1,2}, Paul Hoffman^{1,2}, Austin Hartman¹, Avi Srivastava^{1,2}, Gesmira Molla², Shaista Madad^{1,2}, Carlos Fernandez-Granda^{4,5} & Rahul Satija^{1,2}✉

Mapping single-cell sequencing profiles to comprehensive reference datasets provides a powerful alternative to unsupervised analysis. However, most reference datasets are constructed from single-cell RNA-sequencing data and cannot be used to annotate datasets that do not measure gene expression. Here we introduce ‘bridge integration’, a method to integrate single-cell datasets across modalities using a multiomic dataset as a molecular bridge. Each cell in the multiomic dataset constitutes an element in a ‘dictionary’, which is used to reconstruct unimodal datasets and transform them into a shared space. Our procedure accurately integrates transcriptomic data with independent single-cell measurements of chromatin accessibility, histone modifications, DNA methylation and protein levels. Moreover, we demonstrate how dictionary learning can be combined with sketching techniques to improve computational scalability and harmonize 8.6 million human immune cell profiles from sequencing and mass cytometry experiments. Our approach, implemented in version 5 of our Seurat toolkit (<http://www.satijalab.org/seurat>), broadens the utility of single-cell reference datasets and facilitates comparisons across diverse molecular modalities.

In the same way that read-mapping tools have transformed genome sequence analysis^{1–3}, the ability to map new datasets to established references represents an exciting opportunity for the field of single-cell genomics. As an alternative to fully unsupervised clustering, supervised mapping approaches leverage large and well-curated references to interpret and annotate query profiles. This strategy is enabled by the curation and public release of reference datasets as well as the development of new computational tools, including statistical learning^{4–7} and deep learning-based approaches^{8,9}, that have been successfully applied toward this goal.

A current limitation of existing approaches is the primary focus on single-cell RNA-sequencing (scRNA-seq) data. Single-cell transcriptomics is well suited for the assembly and annotation of reference datasets, particularly as differentially expressed (DE) gene markers

can typically be interpreted to help annotate cell clusters. This has led to the development of high-quality, carefully curated and expertly annotated references, particularly from consortia including the Human Cell Atlas¹⁰, the Human Biomolecular Atlas Project (HuBMAP¹¹) and the Chan Zuckerberg Biohub¹². Mapping to these references facilitates data harmonization, standardization of cell ontologies and naming schemes and comparison of scRNA-seq datasets across experimental conditions and disease states.

A crucial challenge is to extend reference mapping to additional molecular modalities, including single-cell measurements of chromatin accessibility (for example, single-cell assay for transposase-accessible chromatin with sequencing (scATAC-seq^{13,14})), DNA methylation (single-cell bisulfite sequencing¹⁵), histone modifications (single-cell cleavage under targets and tagmentation (scCUT&Tag^{16,17})) and protein

¹Center for Genomics and Systems Biology, New York University, New York, NY, USA. ²New York Genome Center, New York, NY, USA. ³Institute for System Genetics, NYU Langone Medical Center, New York, NY, USA. ⁴Center for Data Science, New York University, New York, NY, USA. ⁵Courant Institute of Mathematical Sciences, New York University, New York, NY, USA. ✉e-mail: rsatija@nygenome.org

levels (cytometry by time of flight (CyTOF¹⁸)), each of which measures a different set of features than scRNA-seq. The lack of transcriptome-wide measurements creates challenges for unsupervised annotation. Ideally, datasets from different modalities could be mapped onto scRNA-seq references, ensuring that established cell labels and ontologies would be preserved. We and others have proposed methods to map datasets across modalities^{19–21}, for example, taking the gene body sum of ATAC-seq signal (or the inverse of the DNA methylation levels) as a proxy for transcriptional output. These make strict biological assumptions (for example, that accessible chromatin is associated with active transcription) that may not always hold true, particularly when analyzing cellular transitions or developmental trajectories²².

Here, we introduce ‘bridge integration’, which integrates single-cell datasets measuring different modalities by leveraging a separate dataset where both modalities are simultaneously measured as a molecular ‘bridge’. The multiomic bridge dataset, which can be generated by a diverse set of technologies^{23–32}, helps to translate information between disparate measurements, resulting in robust integration without requiring any limiting biological assumptions. We illustrate the broad applicability of our approach, demonstrating its performance across five different molecular modalities (Fig. 1a). Moreover, we introduce ‘atomic sketch integration’, which combines dictionary learning and dataset sketching to improve the computational efficiency of large-scale single-cell analysis and enables rapid integration of dozens of datasets spanning millions of cells.

Results

Using multiomic dictionaries for bridge integration

We aimed to develop a flexible and robust integration strategy to integrate data from single-cell sequencing experiments where different modalities are measured (‘single-modality datasets’). The fundamental challenge is that different single-modality datasets measure different sets of features. We reasoned that an approach would be to leverage a multiomic dataset as a bridge that can help to translate between disparate modalities. To perform this translation, we were inspired by the field of dictionary learning, a form of representation learning that is commonly used in image analysis and genomics^{33–37}. The goal of dictionary learning is to represent input data in terms of individual elements that are called atoms and together comprise a dictionary. Reconstructing input data as a weighted linear combination of these atoms is an effective tool for denoising and represents a transformation of the input data into a dictionary-defined space.

We find that dictionary learning enables cross-modality bridge integration at single-cell resolution. Our key insight is to treat a multiomic dataset as a dictionary, with each individual cell’s multiomic profile representing an atom. We learn a ‘dictionary representation’ of each unimodal dataset based on these atoms. For clarity, we emphasize that in contrast to the original applications of dictionary learning where the atoms represent a set of features^{33,37}, we use individual instances (cells) as dictionary elements. This transformation takes datasets in which completely different sets of features were measured and represents them each in a space where the defining features represent the same set of atoms (Fig. 1b). Once different modalities can be represented using the same set of features, they can be readily aligned in a final step.

Our bridge integration is illustrated in Fig. 1b and is described fully in the Supplementary Methods, and we note a few key points below. First, our procedure makes no assumptions about the relationships

between modalities, as these are learned automatically from the multiomic dataset. Second, the key advance we present here is a transformation to project datasets profiling different modalities to be represented by a shared set of atoms. Once transformed, the final alignment step is compatible with a wide diversity of single-cell integration techniques, including Harmony³⁸, mnnCorrect³⁹, Seurat¹⁹, Scanorama⁴⁰ or scVI⁴¹. In this manuscript, we perform this step with an implementation of the mnnCorrect algorithm³⁹.

Third, we found that when working with sizable bridge datasets, the large number of atoms (single cells in the bridge dataset) created a substantial computational burden. Motivated by a similar problem addressed by Laplacian Eigenmaps⁴², we compute an eigen decomposition of the graph Laplacian for the multiomic dataset to reduce the dimensionality from the number of atoms to the number of selected eigenvectors (Supplementary Methods). We then use these eigenvectors to transform the learned dictionary representations into the same lower-dimensional space, substantially increasing the efficiency of our bridge integration procedure.

Mapping scATAC-seq data onto scRNA-seq references

We first demonstrate our bridge integration strategy by performing cross-modality mapping on scATAC-seq and scRNA-seq samples of human bone marrow mononuclear cells (BMMCs). These samples consist of cells representing the full spectrum of hematopoietic differentiation, including hematopoietic stem cells (HSCs), multipotent and oligopotent progenitors and fully differentiated cells. As part of HuBMAP, we have leveraged public datasets to construct a comprehensive scRNA-seq reference (‘Azimuth reference’; 297,627 cells) of human BMMCs, carefully annotating 10 progenitor and 25 differentiated cell states (Fig. 2a). We aimed to map scATAC-seq ‘query’ datasets of human BMMCs⁴³ (16,266 whole bone marrow profiles and 9,893 CD34⁺-enriched profiles) to this reference (Fig. 2b). We used a 10x multiome dataset⁴⁴ (32,368 cells paired single-nucleus RNA-seq + scATAC-seq) that was publicly released as part of NeurIPS 2021 as a bridge.

Our bridge procedure successfully mapped the scATAC-seq dataset on our Azimuth reference, enabling joint visualization of scATAC-seq and scRNA-seq data (Fig. 2c) and automated annotation of scATAC-seq profiles with accompanying prediction scores. Reference mapping also aligned shared cell populations across multiple samples, mitigating sample-specific batch effects. Query samples representing CD34⁺ BMMC fractions mapped exclusively to the HSC and progenitor components in the reference dataset, demonstrating that bridge integration can robustly handle cases where the query dataset represents a subset of the reference, while whole fractions mapped to all 35 cell states (Supplementary Fig. 1a).

Our reference-derived annotations were concordant with the annotations accompanying the query dataset produced by the original authors (Supplementary Fig. 1b), but we found that bridge integration annotated additional rare and high-resolution subpopulations. For example, our annotations separated monocytes into CD14⁺ and CD16⁺ fractions, natural killer cells into CD56^{bright} and CD56^{dim} subgroups and cytotoxic T cells into CD8⁺ and mucosal-associated invariant T (MAIT) cell subpopulations. While these subdivisions were not identified in the unsupervised scATAC-seq analysis, we confirmed these predictions by observing differential accessibility at canonical loci after grouping by reference-derived annotations (Fig. 2d,e and Supplementary Fig. 1c).

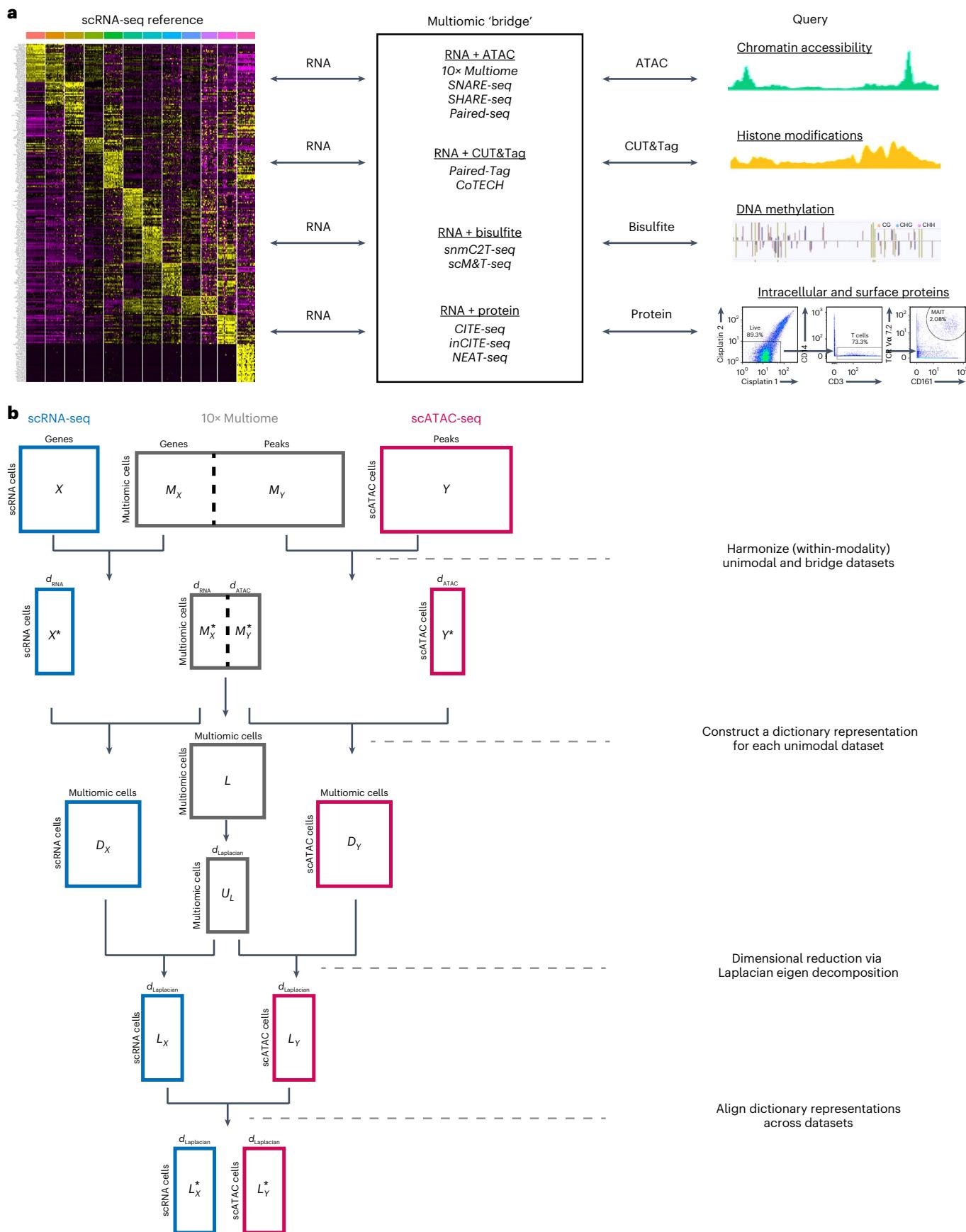
Fig. 1 | Integrating across modalities with molecular bridges. a, Broad schematic of the bridge integration workflow. Two datasets where different modalities are measured (for example, scRNA-seq and scATAC-seq) can be harmonized via a third dataset where both modalities are simultaneously measured (for example, 10x multiome). We demonstrate bridge integration using a variety of multiomic technologies that can be used as bridges, including 10x multiome, Paired-Tag, snmC2T and CITE-seq, each of which facilitates

integration with a different molecular modality. The middle box lists alternative multiomic technologies that can be used to generate bridge datasets.

b, Mathematical schematic of each of the steps in the bridge integration procedure. A full description is provided in the Supplementary Methods. For clarity, the matrix names illustrated in this schematic are the same as the matrix names defined in the Supplementary Methods.

We validated these chromatin patterns using independent multiome datasets, where cell identity was assigned based on concurrent RNA measurements (Supplementary Fig. 1d,e). Similarly, bridge integration

identified extremely rare groups of innate lymphoid cells (ILCs; 0.15%) and recently discovered AXL⁺SIGLEC6⁺ dendritic cells (ASDCs^{45,46}; 0.10%; Fig. 2f and Supplementary Fig. 1c). To our knowledge, these



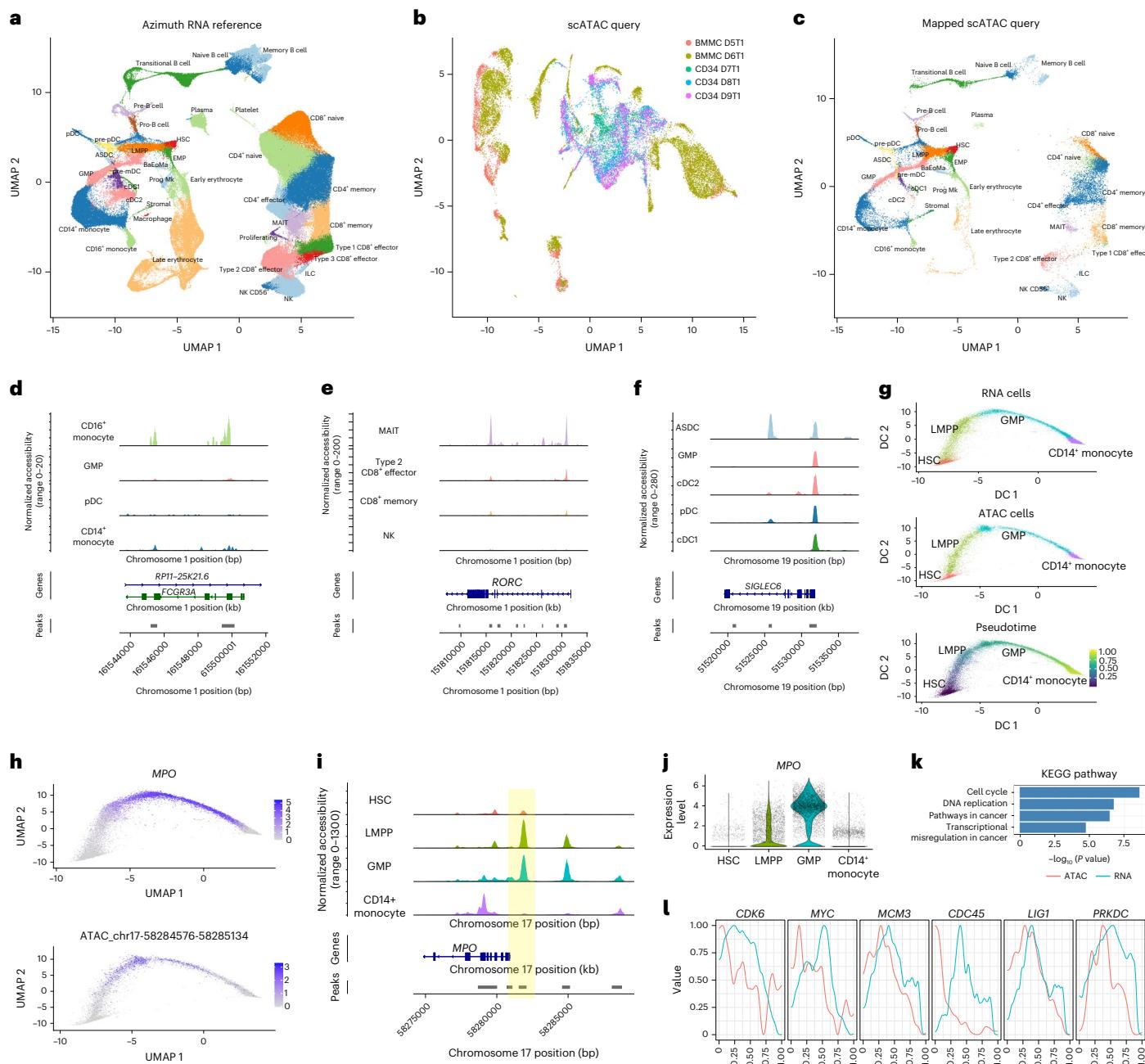


Fig. 2 | Mapping scATAC-seq data onto scRNA-seq references. **a**, Uniform manifold approximation and projection (UMAP) visualization scRNA-seq reference dataset of human bone marrow, representing 297,627 annotated scRNA-seq profiles; mDC, myeloid DC; EMP, erythro-myeloid progenitor; BaEoMa, basophil, eosinophil, mast progenitor; cDC1, conventional type 1 DC; cDC2, conventional type 2 DC; NK, natural killer; Prog Mk, progenitor megakaryocyte. **b**, UMAP visualization of an scATAC-seq query dataset from Granja et al.⁴³, representing 26,159 profiles spanning five batches, three of which are enriched for CD34-expressing cells. **c**, After bridge integration, query cells are annotated based on the scRNA-seq-defined cell ontology and can be visualized on the same embedding. **d–f**, Coverage plots showing chromatin accessibility at selected loci after grouping query cells by their predicted annotations. In each case, the predicted cell labels agree with the expected accessibility patterns;

bp, base pairs; kb, kilobases. **g**, We constructed a differentiation trajectory and pseudotime ordering of cells undergoing myeloid differentiation. The pseudotime ordering in diffusion map coordinates (DC) encompasses both scRNA-seq and scATAC-seq cells. **h**, Example locus where we observe a 'lag' between the gene expression dynamics for *MPO* and the accessibility dynamics for an upstream regulatory region (denoted by a yellow box in **i**). **i**, chromatin accessibility at the *MPO* regulatory locus. The highlighted region becomes accessible at the multipotent LMPP stage. **j**, *MPO* becomes highly expressed at the RNA level at the myeloid-committed GMP stage. **k**, KEGG pathway enrichment for 236 genes where we identified a lag between accessibility and transcriptional dynamics. *P* value is calculated by a Fisher's exact test. **l**, Smoothed chromatin accessibility levels (red) and lagging expression of associated genes (blue) as a function of pseudotime for six cell cycle-associated genes.

cell populations have not been previously identified in scATAC-seq data. Again, we found that differentially accessible sites, such as an ASDC-specific peak in the *SIGLEC6* gene (Fig. 2f), fully supported the accuracy of our mapping procedure;

By projecting datasets from multiple modalities into a common space, our reference-mapping procedure not only enables the transfer of discrete annotations but also allows us to explore how variation in one corresponds to variation in another. For example, after integration,

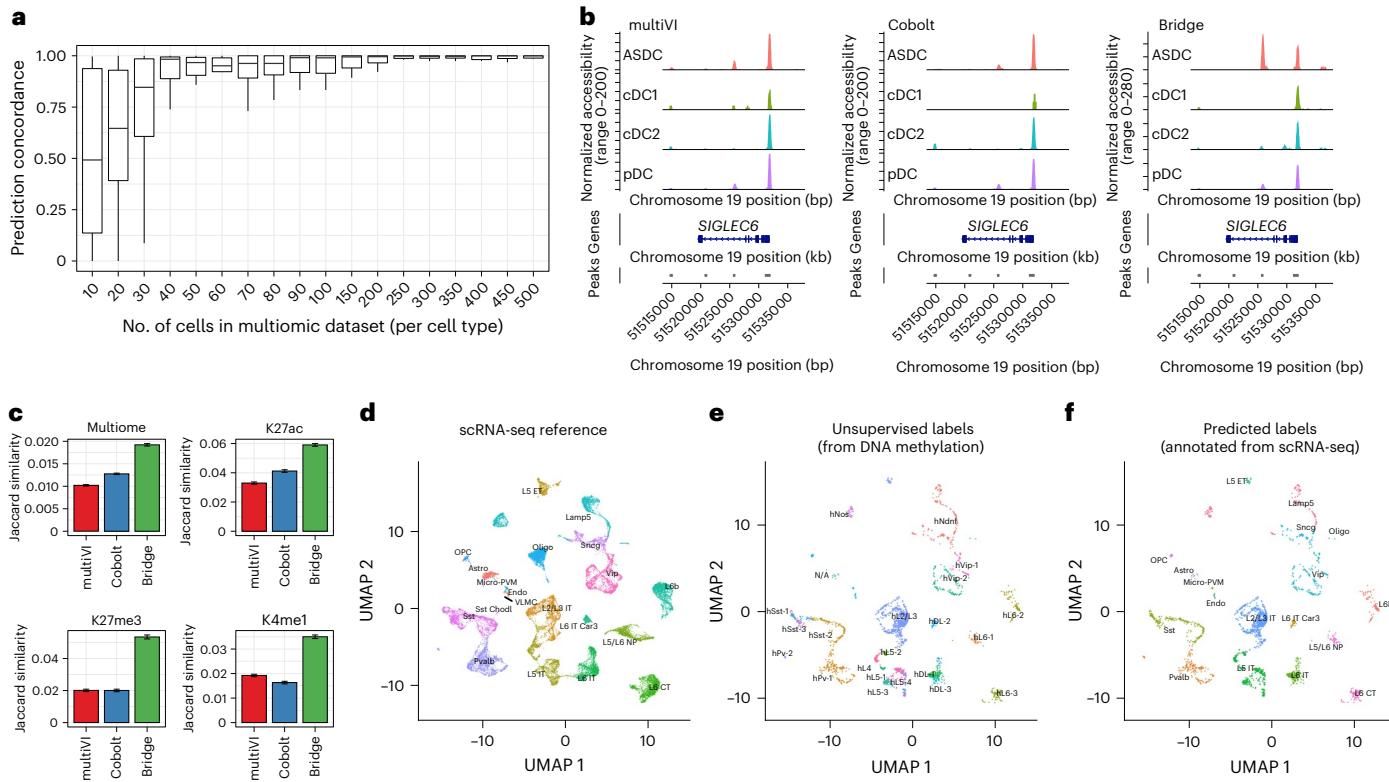


Fig. 3 | Robustness and benchmarking analysis for bridge integration

a, Per cell-type prediction concordance of bridge integration based on the number of cells representing each cell type in the multiomic dataset. Concordance results were obtained by serially downsampling the multiomic dataset, repeating bridge integration and comparing resulting query annotations with those derived from the full dataset. Box plots represent the observed range of values across 21 cell types. Box plots exhibit the median at the center, with the 25% quantile and 75% quantile represented by the lower and upper edges of the boxes, respectively. The whiskers extend from the edge to $1.5 \times$ the interquartile range. **b**, Coverage plots for the *SIGLEC6* locus after performing cross-modality annotation with bridge integration, multiVI and Cobolt. Only cells classified as ASCDs by bridge integration exhibit cell-type-specific accessibility at this locus. Additional loci are shown in Supplementary Fig. 2e,f. **c**, Ground truth benchmarking analysis. RNA and ATAC profiles from a 10x multiome dataset were unpaired and integrated. Bar plots show the average Jaccard similarity value \pm s.d. between each scATAC-seq cell and its matched scRNA-seq cell ($n = 30,253$ cell pairs). Results are split by individual cell types in Supplementary Fig. 3. Results are also shown for Paired-Tag datasets for three histone modification profiles: H3K27ac ($n = 10,906$ cells), H3K27me3 ($n = 6,280$ cells) and H3K4me1 ($n = 12,638$ cells). In each case,

we applied diffusion maps to the harmonized measurements to construct a joint differentiation trajectory spanning multiple progenitor states during myeloid differentiation (Fig. 2g). Because this trajectory represents both reference and query cells, we can explore how pseudotemporal variation in chromatin accessibility correlates with gene expression, even though the two modalities were measured in separate experiments.

Consistent with previous findings, we identified cases where gene expression changes ‘lagged’ behind variation in chromatin accessibility. For example, while myeloperoxidase (encoded by *MPO*) is expressed in granulocyte–macrophage progenitors (GMPs) and is associated with myeloid fate commitment^{47,48}, the regulatory region immediately upstream acquired accessibility in lymphoid-primed multipotent progenitors (LMPPs; Fig. 2h–j). We used a cross-correlation-based metric to systematically identify 236 ‘lagging’ loci (Supplementary Methods) across this trajectory. KEGG pathway enrichment analysis revealed a strong enrichment for genes involved in the cell cycle and

bridge integration achieves the highest Jaccard similarity. **d**, scRNA-seq reference of the human motor cortex; Astro, astrocyte; Endo, endothelial cell; L2/3 IT, layer 2-3 glutamatergic neuron, intratelencephalon-projecting; L5 ET, layer 5 glutamatergic neuron, extratelencephalon-projecting; L5 IT, layer 5 glutamatergic neuron, intratelencephalon-projecting; L5/6 NP, layer 5-6 glutamatergic neuron, near-projecting; L6 CT, layer 6 glutamatergic neuron, corticothalamic-projecting; L6 IT, layer 6 glutamatergic neuron, intratelencephalon-projecting; L6 IT Car3, layer 6 Car3+ glutamatergic neuron, intratelencephalon-projecting; L6b, layer 6b glutamatergic neuron; Lamp5, Lamp5+ GABAergic neuron; Micro-PVM, microglia / perivascular macrophage; Oligo, oligodendrocyte; OPC, oligodendrocyte precursor cell; PvAlb, PvAlb+ GABAergic neuron; Sncg, Sncg+ GABAergic neuron; Sst, Sst+ GABAergic neuron; Sst Chodl, Sst+ Chodl+ GABAergic neuron; Vip, Vip+ GABAergic neuron; VLMC, vascular leptotmeningeal cell. **e,f**, Mapping of single-cell DNA methylation profiles of human cortical cells onto the reference using an snmC2T-seq multiomic dataset as a bridge. Cells are colored by the methylation-derived annotations from the original study (**e**) or the scRNA-seq-derived labels from bridge integration (**f**); near projecting; L6b, deep neocortical laminar 6b. Reference-derived labels at higher levels of granularity are shown in Supplementary Fig. 3.

DNA replication (Fig. 2k). These loci were characterized by accessible chromatin at the earliest stages of differentiation (HSCs), but there is a delay before the associated genes become transcriptionally active (Fig. 2l). The accessible state of these loci in the earliest progenitors may represent a form of priming to enable rapid cell cycle entry once the decision to differentiate has been made and may represent the type of discovery that can be enabled through integrative analysis across modalities.

Robustness and benchmarking analysis

As our strategy relies on the ability for the dictionary to represent and reconstruct individual datasets, we explored how the size and composition of the multiomic dataset affected the accuracy of integration. We sequentially downsampled the multiomic dataset, repeated bridge integration and compared the results to our original findings. Downsampling the bridge generally returned results that were concordant with the full analysis but, as expected, could affect annotation concordance.

for rare cell types, which are most sensitive to downsampling (Fig. 3a). We found that if a bridge dataset contained at least 50 cells ('atoms') representing a given cell type, this was sufficient for robust integration. We note that this threshold is not a strict requirement; we found that integration can be successful for rare cell types, such as ASCDs, even when fewer than ten cells are present in the bridge, but we also observed failure modes in this regimen. We note that generating bridge datasets consisting of more than 50 cells per subpopulation is quite feasible for many multiomic technologies and that our findings represent guidelines to assist in experimental design when performing multiomic experiments. Notably, we found that substantially altering the relative composition of cell types in the bridge dataset (while maintaining the minimum threshold) did not negatively affect performance, demonstrating that bridge integration can be successful even in cases where there are substantial compositional differences in the sample used to generate the multiomic bridge (Supplementary Fig. 2a,b).

We next compared the performance of bridge integration against two recently proposed methods for integrated analysis of multimodal and single-modality datasets. Both multiVI⁴⁹ and Cobolt⁵⁰ use variational autoencoders for integration, and while they do not explicitly treat multiomic datasets as a bridge, they aim to integrate datasets across technologies and modalities into a shared space. When applied to the previously described datasets, both methods were broadly successful in integrating scRNA-seq and scATAC-seq data but did not identify matches at the same level of resolution (for example, neither method successfully matched ASCDs in scATAC-seq data to the ASCDs in the Azimuth reference; Fig. 3b and Supplementary Fig. 2d-f). We also found that the latent space and neighbor relationships learned by bridge integration were most consistent with the labels originally assigned in the ATAC-seq analysis (Supplementary Fig. 2c). When comparing computational efficiency, bridge integration (0.8 h, not including 1.2 h of preprocessing time) and Cobolt (3.3 h) were the most efficient methods, while multiVI required more computational resources (15.7 h).

We next performed quantitative benchmarking of multiomic integration methods (bridge integration, Cobolt and multiVI) and also evaluated 'bridge-free' methods (Canonical Correlation-based Integration and LIGER), which perform integration on the basis of gene activity scores (Supplementary Methods). We found that our bridge integration most consistently and effectively matched cells in the same biological state across modalities (Fig. 3c and Supplementary Fig. 3a). Consistent with our previous results, we found that the strongest improvements were observed when mapping rare cell types, including plasma cells and DCs (Supplementary Fig. 3b). As our procedure is compatible with multiple integration techniques, we compared the performance of bridge integration when using either mnnCorrect³⁹ or Seurat v3 (ref. 19) for the final alignment step and observed very similar results (Supplementary Fig. 3a,b). We also computed additional metrics based on the cluster labels originally assigned based on the scRNA-seq measurements⁴⁴ (Supplementary Table 1). In all cases, we consistently found that bridge integration exhibited superior performance.

As a second quantitative benchmark with ground truth data, we pursued a similar strategy using a recently published Paired-Tag dataset²⁶, where individual histone modification binding profiles via

scCUT&Tag were simultaneously measured with RNA transcriptomes. We performed cross-modality integration between scRNA-seq and scCUT&Tag for active histone marks (H3K27ac), repressive histone marks (H3K27me3) and enhancer histone marks (H3K4me1). In each case, bridge integration successfully integrated cells across modalities and returned the highest Jaccard similarity and classification metrics between matched scRNA-seq and scCUT&Tag profiles (Fig. 3c, Supplementary Fig. 3d,e and Supplementary Table 1).

To further demonstrate the flexibility of our approach, we used bridge integration to map and annotate an snmC-seq dataset, which measures DNA methylation profiles in single cells from the human cortex⁵¹. As a reference, we used a dataset from the Allen Brain Atlas, which defines an expertly curated and multilevel cell ontology⁵² in the human cortex. Using an snmC2T-seq dataset, which simultaneously measures methylation and gene expression as a bridge²⁸, we were able to annotate the snmC-seq profiles with high confidence (Supplementary Fig. 3f). Even when our reference-derived annotations did not augment the resolution to unsupervised clustering of snmC-seq data, they did add substantial interpretability (Fig. 3d-f). For example, unsupervised clustering identified multiple populations of layer 6 (L6) neurons (labeled as L6-1, L6-2 and L6-3), but RNA-assisted annotation clearly labeled these clusters as either 'near projecting' or deep neocortical laminar 6b excitatory neurons (Fig. 3f).

Last, we aimed to characterize the performance of our method specifically in cases where the bridge dataset was missing specific cell populations or exhibited low data quality. Using the BMMC multiome benchmark dataset, we removed all plasmacytoid DCs (pDCs) from the multiomic dataset and repeated bridge integration. We found that this modification did not alter the annotations or confidence scores of non-pDCs in the query but that pDC query cells did exhibit a drop in annotation performance (94.4% annotated as pDCs using the full bridge and 83.5% annotated as pDCs using the depleted bridge dataset). However, we found that these query cells also exhibited a specific and sharp drop in prediction confidence (average prediction scores of 0.907 using the full bridge and 0.514 using the depleted bridge), demonstrating that our procedure correctly reduced the confidence of prediction when the underlying assumptions were not met. We repeated this analysis after separately depleting three additional cell populations (B cells, CD8⁺ T cells and CD14⁺ monocytes) and observed similar results (Supplementary Fig. 4a). Moreover, we found that substantially reducing bridge data quality by discarding unique molecular identifiers (UMIs; 86% downsampling to 750 RNA UMIs per cell or 70% downsampling to 2,500 ATAC fragments per cell) did not adversely affect integration, although we did observe performance reductions after further downsampling (Supplementary Fig. 4b,c).

Taken together, these results demonstrate the accuracy, robustness and flexibility of our bridge integration procedure. We demonstrate applications on multiple modalities and data types as well as best-in-class performance via quantitative and ground truth benchmark comparisons.

Using dictionary learning for massively scalable integration

The recent increase in publicly available single-cell datasets poses a challenge for integrative analysis. For example, multiple tissues

Fig. 4 | Using dictionary learning for massively scalable integration.

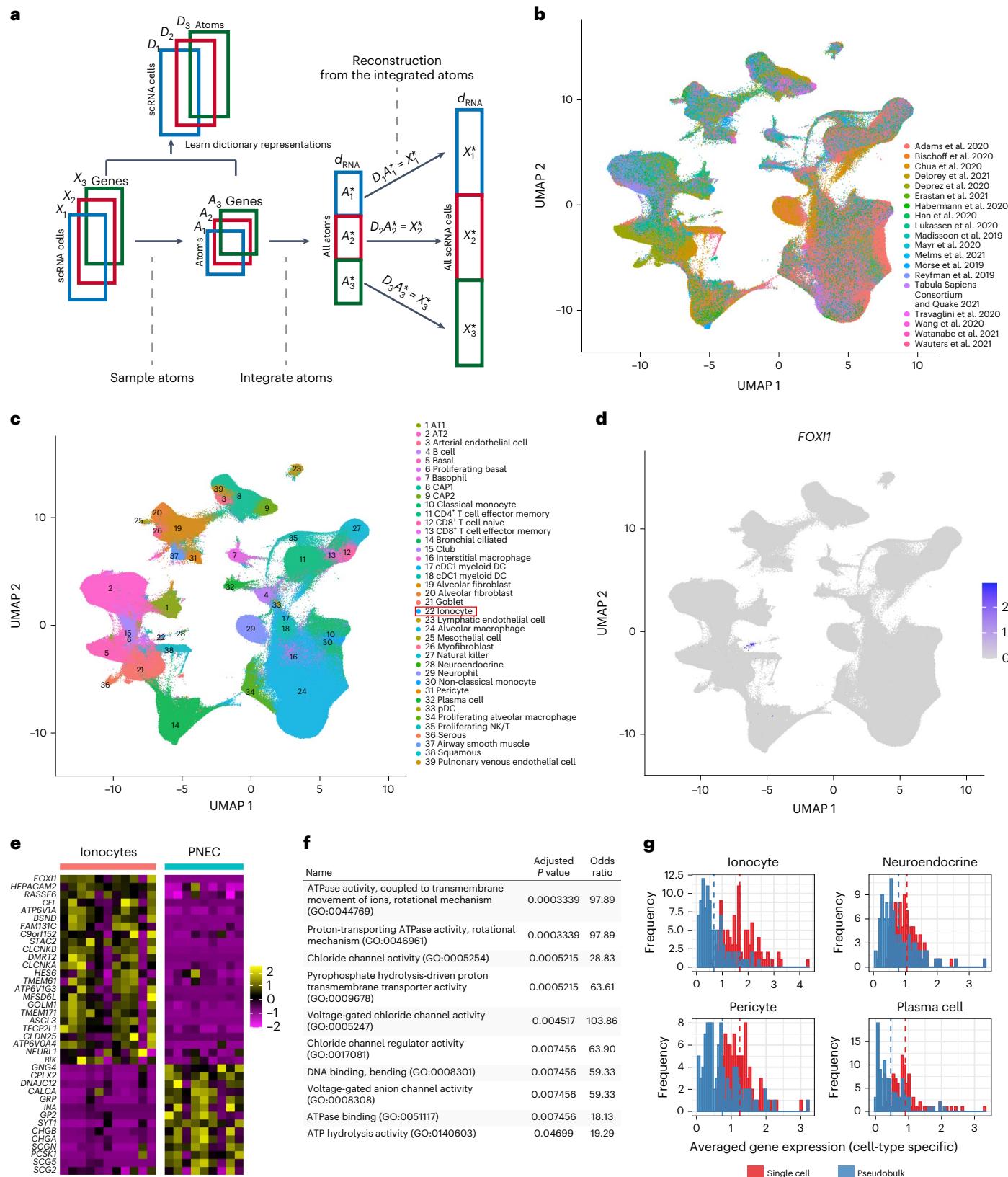
a, Schematic of the atomic sketch integration procedure. After selecting a representative set of cells from each dataset, these cells are integrated and used to reconstruct harmonized profiles for all cells. Matrix notation is consistent with the full mathematical description in the Supplementary Methods. **b,c**, UMAP visualization of 1,525,710 scRNA-seq profiles spanning 19 studies from the lung and upper airways, which were harmonized using atomic sketch integration in 55 min. Cells are colored by their study of origin (**b**) or annotated cell type after integration (**c**); ATI, alveolar type 1; AT2, alveolar type 2. **d**, Expression of *FOXJ1*, a transcriptional marker of pulmonary ionocytes, in the integrated dataset.

e, Heat map showing the top transcriptional markers of pulmonary ionocytes that are consistent across multiple studies. Pulmonary neuroendocrine cells (PNECs), the most transcriptionally similar cell type, are shown for contrast. Each column represents a pseudobulk average of all cells from a single cell type and single study. Top transcriptional markers for all cell types are shown in Supplementary Fig. 3. **f**, Gene ontology (GO) enrichment terms for ionocyte markers. *P* values were calculated by Fisher's exact test and were adjusted by the Benjamini–Hochberg test. **g**, Expression distributions of top transcriptional markers recovered from single-cell differential expression analysis (red) or pseudobulk analysis (blue).

have now been profiled across dozens of studies, representing hundreds of individuals and millions of cells. We refer to the challenge of harmonizing a broad swath (or the entirety) of publicly available single-cell datasets from a single organ as ‘community-wide’ integration. While a rich diversity of analytical methods can harmonize datasets of hundreds of thousands of cells, performing unsupervised

‘community-wide’ integration remains challenging, even when analyzing a single modality.

We were inspired by previous work on ‘geometric sketching’, which first selects a representative subset of cells (a ‘sketch’) across all datasets, integrates them and then propagates the integrated result back to the full dataset⁵³. This pioneering approach substantially improves



the scalability of integration, as the heaviest computational steps are focused on subsets of the data. However, this approach is dependent on the results of principal-component analysis (PCA) that must first be performed on the full dataset. As datasets continue to grow in scale, performing dimensional reduction can become a limiting step. We aimed to devise a strategy that could integrate large compendiums of datasets, without ever needing to simultaneously analyze or perform intensive computation on the full set of cells.

We reasoned that dictionary learning could also enable efficient and large-scale integrative analysis. We first selected a representative sketch of cells (that is, 5,000 cells) from each dataset and treated these cells as atoms in a dictionary (Fig. 4a and Supplementary Methods). We next learned a dictionary representation, a weighted linear combination of atoms that can reconstruct the full dataset. These steps can occur for each dataset independently, allowing for efficient parallel processing. We then performed integration on the atoms from each dataset. This is the only step that simultaneously analyzes cells from multiple datasets, but because only the atoms are considered, this does not impose scalability challenges. Finally, we applied our previously learned dictionary representations to the harmonized atoms from each dataset individually and reconstructed harmonized profiles for the full dataset. We refer to this procedure as ‘atomic sketch integration’. We highlight that for this application, the ‘atoms’ used to reconstruct a dataset represent a subset of cells from the dataset itself. By contrast, in bridge integration, the atoms refer to cells from a different (multiomic) dataset.

The success of atomic sketch integration rests on identifying a representative subset of cells for each dataset. Sketching techniques for single-cell analyses aim to find subsamples that preserve the overall geometry of these datasets^{53–55}. These methods do not require a preclustering of the data but aim to ensure that the sketched dataset represents both rare and abundant cell states even after downsampling. Here, we perform sketching using a leverage score sampling-based strategy that has been proposed for large-scale information retrieval problems⁵⁶ and can be rapidly and efficiently computed on sparse datasets. Leverage score-based sampling does not require performing PCA but maintains the ability to efficiently identify cells from rare subpopulations compared to geometric sketching techniques⁵³ (Supplementary Fig. 5a,b). We emphasize that atomic sketch integration represents a general strategy for improving scalability that can be broadly coupled with existing methods. For example, a wide variety of integration techniques, including Harmony³⁸, Scanorama⁴⁰, mnnCorrect³⁹, scVI⁴¹ and Seurat¹⁹, can be used to integrate the atom elements in each dictionary, with our procedure then enabling these results to be extended to full datasets.

Community-scale integration for human lung scRNA-seq

To demonstrate the potential of atomic sketch integration to perform ‘community-wide’ analysis, we first considered scRNA-seq datasets of the human lung. During the coronavirus disease 2019 (COVID-19) pandemic, there has been widespread scRNA-seq data collection from respiratory tissues, particularly by the Human Cell Atlas Lung Biological Network⁵⁷. Leveraging a recently published ‘database’ of scRNA-seq studies⁵⁸ and a collection of openly released lung and upper airway datasets from the Human Cell Atlas (<https://www.covid19cellatlas.net>,

org/index.healthy.html), we assembled a group of 19 datasets spanning a total of 1,525,710 individual cells. We created an atomic dictionary consisting of 5,000 cells from each dataset (95,000 total atoms), integrated these cells and reconstructed the full datasets. Our atomic sketch integration procedure performed all these steps (including preprocessing) in 55 min using a single computational core. We found that the integrated latent space preserved the neighbor relationships between cell types independently assigned in each dataset but also mixed cells across datasets (Supplementary Fig. 5c–e).

Our results exhibit the advantages of community-scale integration compared to individual analysis. First, by matching biological states across datasets and technologies, the integrated reference can help to standardize cell ontologies and naming schemes (Fig. 4b,c). When observing previously assigned annotations derived from each study, we found that matched cell populations were often assigned slightly different names (Supplementary Fig. 5f). We also identified cases where integrated annotations exhibited increased resolution compared to the original labels and verified that our higher-resolution annotations were supported by the expression patterns of reproducible gene expression markers (Supplementary Fig. 5g).

As a second benefit, we found that community-scale integration enabled consistent identification of ultra-rare populations and, in particular, a population of *FoxI1*-expressing ‘pulmonary ionocytes’ that were recently discovered in both human and mouse lungs⁵⁹ (Fig. 4d). While these cells were only independently annotated in 6 of 19 studies, our integrated analysis discovered at least one pulmonary ionocyte in 17 of 19 studies. The identified ionocytes were extremely rare (0.047%) but exhibited clear expression of canonical markers (Fig. 4c), highlighting the potential value for pooling multiple datasets to characterize these cells. We note that selection of dictionary atoms by sketching or leverage score sampling is essential for optimal performance (Supplementary Fig. 5h,i); repeating the analysis using a set of atoms determined by random downsampling successfully integrated abundant cell types but failed to integrate ionocytes, as they were not sufficiently represented in the dictionary.

Finally, we found that community-scale integration can substantially improve the identification of DE cell-type markers. The use of 19 study replicates specifically enables us to identify genes that show consistent patterns across laboratories and technologies, representing robust and reproducible markers. We grouped cells by both sample replicate and cell-type identity and performed differential expression on the resulting pseudobulk profiles (Fig. 4e and Supplementary Fig. 6). For example, we identified 116 positive markers for pulmonary ionocytes, representing one of the deepest transcriptional characterizations of this cell type. These markers included canonical markers, such as the transcription factor *FOXII*, but also revealed clear ontology enrichments for ATPases (for example, *ATP6V1G3* and *ATP6VOA4*) and chloride channels (for example, *CLCNKA*, *CLCNKB* and *CFTR*), supporting the role of these cells in regulating chemical concentrations in the lung (Fig. 4f). One advantage of working with pseudobulk values is increased quantification accuracy for genes expressed at low levels. Indeed, we repeatedly found that the top DE markers found using this strategy tended to capture more genes at a lower range of average expression values (Fig. 4g).

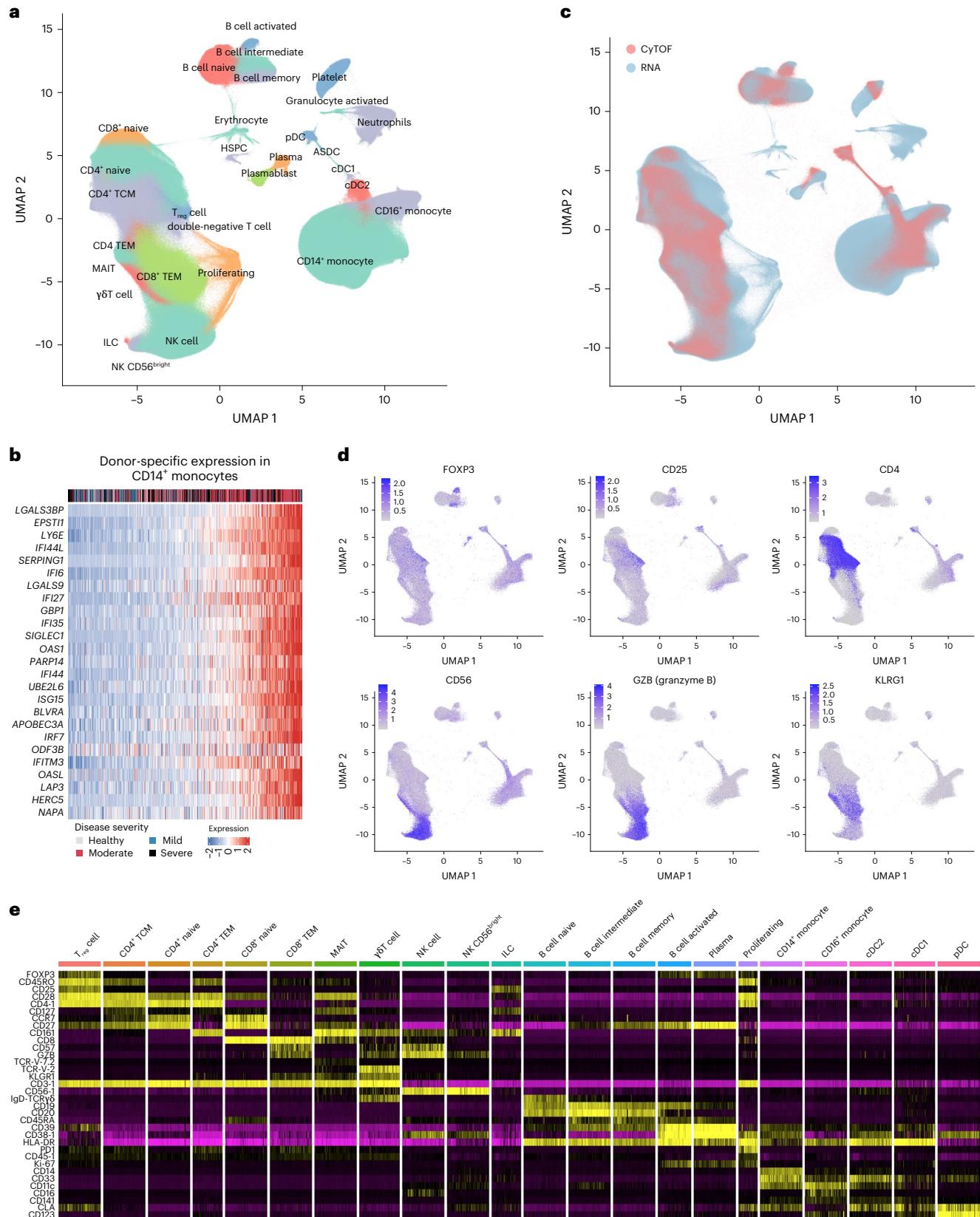
Fig. 5 | ‘Community-scale’ integration of sequencing and cytometry immune datasets. **a**, UMAP visualization of 3,461,171 human PBMC scRNA-seq profiles spanning 14 studies and 639 individuals after performing atomic sketch integration; HSPC, hematopoietic stem and progenitor cell; T_{reg} cell, regulatory T; TCM, central memory T; TEM, effector memory T cell. **b**, Expression of a COVID-19 response module in CD14⁺ monocytes. Each column represents a pseudobulk average of CD14⁺ monocytes from 1 of 506 individuals. Expression of the module is correlated with disease severity within the individual, which is indicated by the color scale above the heat map. Responses for additional cell states are shown in Supplementary Fig. 5b. **c**, Mapping of 5,170,249 additional

CyTOF profiles spanning 119 individuals using a published CITE-seq dataset (Hao et al.⁴) as a multiomic bridge. Each CyTOF profile is annotated with one of the scRNA-seq-defined cell types. **d**, Cross-modality integration enables the exploration of cell surface and intracellular protein markers on cell landscapes defined by scRNA-seq. As an example, intracellular FOXP3 levels are highly enriched in annotated regulatory T cells, validating the accuracy of our mapping. Two hundred thousand cells are shown in each visualization to alleviate overplotting. **e**, Heat map showing the expression of 34 protein markers in the CyTOF dataset. Each column represents a pseudobulk average after grouping cells by individual and reference-derived annotation.

Community-scale integration of scRNA-seq and CyTOF

As a final demonstration, we considered a similar problem of community-wide integration for circulating human peripheral blood cells, which is one of the most widely profiled systems with diverse single-cell technologies. Exploring publicly available studies of either COVID-19 samples or healthy controls, we accumulated a collection

of 14 studies with scRNA-seq measurements, representing a total of 3.46 million cells from 639 individuals. Data from 11 of the studies were obtained from a recently published collection of standardized single-cell sequencing datasets⁶⁰. We performed unsupervised atomic sketch integration, yielding a harmonized collection in which we annotated 30 cell states (Fig. 5a). We identified specific populations



of activated granulocytes and B cells that were specific to COVID-19 samples (Supplementary Fig. 7a). Consistent with previous reports, monocytes in COVID-19 samples sharply upregulated the expression of interferon response genes^{61,62} but were correctly harmonized with healthy monocytes (Fig. 5b and Supplementary Fig. 7b). By matching shared cell types across disease states (while still allowing for the possibility of disease-specific subpopulations), this collection represents a valuable resource for identifying cell-type-specific transcriptional changes that reproduce across multiple studies. We characterized cell-type-specific responses for eight additional cell types, each of which exhibited a conserved interferon-driven response alongside the activation of cell-type-specific response genes (Supplementary Fig. 8).

While single-cell sequencing technologies are capable of measuring RNA transcripts and surface proteins in thousands of single cells, cytometry-based techniques can measure both extracellular and intracellular proteins in millions of cells. As our bridge integration procedure should enable the mapping of CyTOF profiles onto scRNA-seq datasets, we obtained a collection of CyTOF datasets spanning 119 individuals and a total of 5,170,249 cells⁶³. We used our previously collected CITE-seq dataset of 161,764 peripheral blood mononuclear cells (PBMCs) from healthy donors as a multiomic bridge⁴. The CyTOF and CITE-seq dataset both shared 30 cell surface protein features, while the CyTOF dataset also measured 17 unique proteins, which included intracellular targets that cannot be measured via CITE-seq.

Bridge integration annotated each CyTOF dataset with cluster labels derived from our scRNA-seq collection of 3.46 million cells and allowed us to infer intracellular protein levels for each of these clusters (Fig. 5c). Predicted regulatory CD4⁺ T cells expressed high levels of the transcription factor FOXP3 (ref. 64), and effector T cells exhibited enriched KLRG1 levels⁶⁵ (Fig. 5d). We also found that among cytotoxic lymphocyte populations, MAIT cells were uniquely depleted for expression of the cytotoxic protease granzyme B, consistent with previous reports⁶⁶. Each of these patterns supports the accuracy of our cross-modality mapping. Finally, we successfully annotated a rare population of ILCs (0.024%), which were not independently identified in the CyTOF dataset but correctly exhibited a CD25⁺CD127⁺CD161⁺CD56⁺ immunophenotype^{4,67} (Fig. 5d,e). Taken together, we conclude that dictionary learning enhances the scalability of integration and the ability to integrate and compare diverse molecular modalities.

Discussion

To map datasets measuring a diverse set of modalities to scRNA-seq reference datasets, we developed bridge integration, an approach for cross-modality alignment that leverages a multiomic dataset as a bridge. We characterize specific requirements for the bridge dataset and demonstrate the broad applicability of our method to a wide variety of technologies and modalities. Finally, we demonstrate how to use atomic sketch integration to extend the scalability of our approach to harmonize dozens of datasets spanning millions of cells.

We anticipate that our methods will be valuable to individual labs but also larger consortia that have already invested in constructing and annotating comprehensive scRNA-seq references. For example, the Human Cell Atlas, Human Biomolecular Atlas Project, Tabula Sapiens⁶⁸ and Human Cell Landscape⁶⁹ have all released scRNA-seq references spanning hundreds of thousands of cells for multiple human tissues. Similar efforts are present in model organisms as well, including the Fly Cell Atlas⁷⁰ and Plant Cell Atlas projects⁷¹. In each case, these efforts involve careful, collaborative and expert-driven cell annotation alongside the curation of reference cell ontologies. While repeating this manual effort for each modality is not feasible, bridge integration enables the mapping of new modalities without having to modify the reference. As additional multiomic datasets become available, we expect that tools such as Azimuth will also begin to map additional modalities.

We note that bridge integration is particularly well suited for experimental designs where multiomic technologies can be applied

to a subset of, rather than all, experimental samples due to its increased cost, lower throughput and reduced data quality. In particular, combinatorial indexing approaches can be readily applied to profile a single modality in hundreds of thousands of cells^{72,73} but not for multiomic technologies. We propose that the collection of large single-modality datasets, harmonized via a smaller but representative multiomic bridge, may represent an efficient and robust strategy to explore cross-modality relationships across millions of cells.

We note that future extensions of our work can further broaden the applicability of bridge integration or demonstrate its potential in new contexts. For example, performing bridge integration on spatially resolved unimodal datasets (for example, CODEX⁷⁴) could help to better characterize the spatial localization of scRNA-seq-defined cell types in large tissue sections. New multiomic technologies that couple high-resolution mass spectrometry imaging to single-cell or spatial transcriptomics could serve as a bridge to harmonize lipidomic and metabolic profiles^{75,76} with sequencing-based references. In addition, future computational improvements will further lower the requirements of the bridge dataset, enabling robust integration with an even smaller number of multiomic cells.

We emphasize the ability for bridge and atomic sketch integration to identify and characterize rare cell populations, including ASDCs and pulmonary ionocytes. Single-cell transcriptome profiling played an essential role in the initial discovery of these cell types, but a deeper understanding of their biological role and function will benefit from multimodal characterization. The goal of moving beyond an initial taxonomic classification of cell types toward a complete multimodal reference will not be accomplished with a single experiment or technology. We envision that computational tools for cross-modality integration will have key contributions to the construction of this map.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41587-023-01767-y>.

References

1. Kent, W. J. BLAT—the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
2. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
3. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
4. Hao, Y. et al. Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587 (2021).
5. Kang, J. B. et al. Efficient and precise single-cell reference atlas mapping with Symphony. *Nat. Commun.* **12**, 5890 (2021).
6. Kiselev, V. Y., Yiu, A. & Hemberg, M. scmap: projection of single-cell RNA-seq data across data sets. *Nat. Methods* **15**, 359–362 (2018).
7. Domínguez Conde, C. et al. Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science* **376**, eabl5197 (2022).
8. Xu, C. et al. Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. *Mol. Syst. Biol.* **17**, e9620 (2021).
9. Lotfollahi, M. et al. Mapping single-cell data to reference atlases by transfer learning. *Nat. Biotechnol.* **40**, 121–130 (2022).
10. Regev, A. et al. The human cell atlas. *eLife* **6**, e27041 (2017).
11. Hu, B. C. The human body at cellular resolution: the NIH Human Biomolecular Atlas Program. *Nature* **574**, 187–192 (2019).

12. Tabula Muris Consortium et al. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* **562**, 367–372 (2018).
13. Buenrostro, J. D. et al. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486–490 (2015).
14. Cusanovich, D. A. et al. Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* **348**, 910–914 (2015).
15. Clark, S. J. et al. Genome-wide base-resolution mapping of DNA methylation in single cells using single-cell bisulfite sequencing (scBS-seq). *Nat. Protoc.* **12**, 534–547 (2017).
16. Wu, S. J. et al. Single-cell CUT&Tag analysis of chromatin modifications in differentiation and tumor progression. *Nat. Biotechnol.* **39**, 819–824 (2021).
17. Bartosovic, M., Kabbe, M. & Castelo-Branco, G. Single-cell CUT&Tag profiles histone modifications and transcription factors in complex tissues. *Nat. Biotechnol.* **39**, 825–835 (2021).
18. Bendall, S. C. et al. Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. *Science* **332**, 687–696 (2011).
19. Stuart, T. et al. Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902 (2019).
20. Barkas, N. et al. Joint analysis of heterogeneous single-cell RNA-seq dataset collections. *Nat. Methods* **16**, 695–698 (2019).
21. Welch, J. D. et al. Single-cell multi-omic integration compares and contrasts features of brain cell identity. *Cell* **177**, 1873–1887 (2019).
22. Lara-Astiaso, D. et al. Immunogenetics. Chromatin state dynamics during blood formation. *Science* **345**, 943–949 (2014).
23. Chen, S., Lake, B. B. & Zhang, K. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nat. Biotechnol.* **37**, 1452–1457 (2019).
24. Ma, S. et al. Chromatin potential identified by shared single-cell profiling of RNA and chromatin. *Cell* **183**, 1103–1116 (2020).
25. Zhu, C. et al. An ultra high-throughput method for single-cell joint analysis of open chromatin and transcriptome. *Nat. Struct. Mol. Biol.* **26**, 1063–1070 (2019).
26. Zhu, C. et al. Joint profiling of histone modifications and transcriptome in single cells from mouse brain. *Nat. Methods* **18**, 283–292 (2021).
27. Xiong, H., Luo, Y., Wang, Q., Yu, X. & He, A. Single-cell joint detection of chromatin occupancy and transcriptome enables higher-dimensional epigenomic reconstructions. *Nat. Methods* **18**, 652–660 (2021).
28. Luo, C. et al. Single nucleus multi-omics identifies human cortical cell regulatory genome diversity. *Cell Genomics* **2**, 100107 (2022).
29. Clark, S. J. et al. scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. *Nat. Commun.* **9**, 781 (2018).
30. Stoeckius, M. et al. Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods* **14**, 865–868 (2017).
31. Chung, H. et al. Joint single-cell measurements of nuclear proteins and RNA in vivo. *Nat. Methods* **18**, 1204–1212 (2021).
32. Chen, A.F. et al. NEAT-seq: simultaneous profiling of intra-nuclear proteins, chromatin accessibility and gene expression in single cells. *Nat. Methods* **19**, 547–553 (2022).
33. Elad, M. & Aharon, M. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Trans. Image Process.* **15**, 3736–3745 (2006).
34. Rams, M. & Conrad, T. O. F. Dictionary learning allows model-free pseudotime estimation of transcriptomic data. *BMC Genomics* **23**, 56 (2022).
35. Ramirez, I., Sprechmann, P. & Sapiro, G. in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 3501–3508 (IEEE, 2010).
36. Zhang, Q. & Li, B. in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 2691–2698 (IEEE, 2010).
37. Aharon, M., Elad, M. & Bruckstein, A. K-SVD: an algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Trans. Signal Process.* **54**, 4311–4322 (2006).
38. Korsunsky, I. et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).
39. Haghverdi, L., Lun, A. T. L., Morgan, M. D. & Marioni, J. C. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.* **36**, 421–427 (2018).
40. Hie, B., Bryson, B. & Berger, B. Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. *Nat. Biotechnol.* **37**, 685–691 (2019).
41. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nat. Methods* **15**, 1053–1058 (2018).
42. Belkin, M. & Niyogi, P. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput.* **15**, 1373–1396 (2003).
43. Granja, J. M. et al. Single-cell multiomic analysis identifies regulatory programs in mixed-phenotype acute leukemia. *Nat. Biotechnol.* **37**, 1458–1465 (2019).
44. Luecken, M. D. et al. in *35th Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)* (NeurIPS, 2021).
45. Villani, A. C. et al. Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science* **356**, eaah4573 (2017).
46. See, P. et al. Mapping the human DC lineage through the integration of high-dimensional techniques. *Science* **356**, eaag3009 (2017).
47. Paul, F. et al. Transcriptional heterogeneity and lineage commitment in myeloid progenitors. *Cell* **163**, 1663–1677 (2015).
48. Zheng, S., Papalexci, E., Butler, A., Stephenson, W. & Satija, R. Molecular transitions in early progenitors during human cord blood hematopoiesis. *Mol. Syst. Biol.* **14**, e8041 (2018).
49. Ashuach, T., Gabitto, M. I., Jordan, M. I. & Yosef, N. MultiVI: deep generative model for the integration of multi-modal data. Preprint at [bioRxiv https://doi.org/10.1101/2021.08.20.457057](https://doi.org/10.1101/2021.08.20.457057) (2021).
50. Gong, B., Zhou, Y. & Purdom, E. Cobolt: integrative analysis of multimodal single-cell sequencing data. *Genome Biol.* **22**, 351 (2021).
51. Luo, C. et al. Single-cell methylomes identify neuronal subtypes and regulatory elements in mammalian cortex. *Science* **357**, 600–604 (2017).
52. Bakken, T. E. et al. Comparative cellular analysis of motor cortex in human, marmoset and mouse. *Nature* **598**, 111–119 (2021).
53. Hie, B., Cho, H., DeMeo, B., Bryson, B. & Berger, B. Geometric sketching compactly summarizes the single-cell transcriptomic landscape. *Cell Syst.* **8**, 483–493 (2019).
54. DeMeo, B. & Berger, B. Hopper: a mathematically optimal algorithm for sketching biological data. *Bioinformatics* **36**, i236–i241 (2020).
55. Hicks, S. C., Liu, R., Ni, Y., Purdom, E. & Risso, D. mbkmeans: fast clustering for single cell data using mini-batch k-means. *PLoS Comput. Biol.* **17**, e1008625 (2021).
56. Clarkson, K. L. & Woodruff, D. P. Low-rank approximation and regression in input sparsity time. *JACM* **63**, 1–45 (2017).
57. Schiller, H. B. et al. The Human Lung Cell Atlas: a high-resolution reference map of the human lung in health and disease. *Am. J. Respir. Cell Mol. Biol.* **61**, 31–41 (2019).
58. Svensson, V., da Veiga Beltrame, E. & Pachter, L. A curated database reveals trends in single-cell transcriptomics. *Database* **2020**, baaa073 (2020).
59. Plasschaert, L. W. et al. A single-cell atlas of the airway epithelium reveals the CFTR-rich pulmonary ionocyte. *Nature* **560**, 377–381 (2018).

60. Tian, Y. et al. Single-cell immunology of SARS-CoV-2 infection. *Nat. Biotechnol.* **40**, 30–41 (2022).
61. Lee, J. S. & Shin, E. C. The type I interferon response in COVID-19: implications for treatment. *Nat. Rev. Immunol.* **20**, 585–586 (2020).
62. Wilk, A. J. et al. A single-cell atlas of the peripheral immune response in patients with severe COVID-19. *Nat. Med.* **26**, 1070–1076 (2020).
63. COvid-19 Multi-omics Blood ATlas (COMBAT) Consortium. A blood atlas of COVID-19 defines hallmarks of disease severity and specificity. *Cell* **185**, 916–938.e58 (2022).
64. Rudensky, A. Y. Regulatory T cells and Foxp3. *Immunol. Rev.* **241**, 260–268 (2011).
65. Thimme, R. et al. Increased expression of the NK cell receptor KLRG1 by virus-specific CD8 T cells during persistent antigen stimulation. *J. Virol.* **79**, 12112–12116 (2005).
66. Kurioka, A. et al. MAIT cells are licensed through granzyme exchange to kill bacterially sensitized targets. *Mucosal Immunol.* **8**, 429–440 (2015).
67. Bjorklund, A. K. et al. The heterogeneity of human CD127⁺ innate lymphoid cells revealed by single-cell RNA sequencing. *Nat. Immunol.* **17**, 451–460 (2016).
68. Tabula Sapiens Consortium. The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. *Science* **376**, eabl4896 (2022).
69. Han, X. et al. Construction of a human cell landscape at single-cell level. *Nature* **581**, 303–309 (2020).
70. Li, H. et al. Fly Cell Atlas: A single-nucleus transcriptomic atlas of the adult fruit fly. *Science* **375**, eabk2432 (2022).
71. Plant Cell Atlas Consortium et al. Vision, challenges and opportunities for a Plant Cell Atlas. *eLife* **10**, e66877 (2021).
72. Lareau, C. A. et al. Droplet-based combinatorial indexing for massive-scale single-cell chromatin accessibility. *Nat. Biotechnol.* **37**, 916–924 (2019).
73. Datlinger, P. et al. Ultra-high-throughput single-cell RNA sequencing and perturbation screening with combinatorial fluidic indexing. *Nat. Methods* **18**, 635–642 (2021).
74. Goltsev, Y. et al. Deep profiling of mouse splenic architecture with CODEX multiplexed imaging. *Cell* **174**, 968–981 (2018).
75. Li, Z. et al. Single-cell lipidomics with high structural specificity by mass spectrometry. *Nat. Commun.* **12**, 2869 (2021).
76. Capolupo, L. et al. Sphingolipid control of fibroblast heterogeneity revealed by single-cell lipidomics. Preprint at bioRxiv <https://doi.org/10.1101/2021.02.23.432420> (2021).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2023

Methods

Bridge integration procedure

Our bridge integration procedure is designed to perform integration of single-cell datasets profiling different modalities by leveraging a separate multiomic dataset as a molecular bridge. The individual multiomic profiles each represent individual atoms, which together comprise a multiomic dictionary (that is, each cell in the bridge dataset represents an atom, and the entire bridge dataset represents a dictionary). This dictionary is used to transform both unimodal datasets into a shared space defined by the same set of features, facilitating cross-modality integration. Our approach consists of the following four broad steps described in detail below: (1) within-modality harmonization of unimodal and bridge datasets, (2) construction of a dictionary representation for each unimodal dataset, (3) dimensional reduction via Laplacian eigen decomposition and (4) alignment of dictionary representations across datasets. We illustrate each step of the method in Fig. 1b using the same mathematical notations that we introduce below.

All methods are implemented in our open-source R package Seurat (www.satijalab.org/seurat and www.github.com/satijalab/seurat).

Within-modality harmonization of unimodal and bridge datasets. The first step in our procedure is to harmonize the unimodal and bridge datasets based on shared modalities. For example, when performing bridge integration to map an scATAC-seq dataset onto an scRNA-seq reference (via a 10x multiome bridge), we first harmonize the gene expression measurements from the scRNA-seq and multiome experiments and the chromatin accessibility measurements from the scATAC-seq and multiome experiments. Specifically, we define the following:

$X \in \mathbb{R}^{n_{\text{scRNA-seq}} \times d_{\text{genes}}}$ is the scRNA-seq expression counts matrix,

$Y \in \mathbb{R}^{n_{\text{scATAC-seq}} \times d_{\text{peaks}}}$ is the scATAC-seq accessibility counts matrix,

$M = [M_X M_Y]$ is the multiomic expression + accessibility counts matrix, where

$M_X \in \mathbb{R}^{n_{\text{multiomic}} \times d_{\text{genes}}}$ is the scRNA-seq subset of the multiomic matrix and

$M_Y \in \mathbb{R}^{n_{\text{multiomic}} \times d_{\text{peaks}}}$ is the scATAC-seq subset of the multiomic matrix.

Our goal is to harmonize X and M_X and Y and M_Y . This can be performed with a wide variety of existing tools for the harmonization of single-cell datasets. For example, Seurat, Harmony, LIGER, scVI, Scanorama, fastMNN, scVI and scArches all learn a shared low-dimensional space that jointly represents the datasets and aligns cells in a matched biological state. Our goal is therefore to learn

$X^* \in \mathbb{R}^{n_{\text{scRNA-seq}} \times d_{\text{RNA}}}$, harmonized space for scRNA-seq data,

$Y^* \in \mathbb{R}^{n_{\text{scATAC-seq}} \times d_{\text{ATAC}}}$, harmonized space for scATAC-seq data, and

$$M^* = [M_X^* M_Y^*],$$

where

$M_X^* \in \mathbb{R}^{n_{\text{multiomic}} \times d_{\text{RNA}}}$ is the harmonized space for the scRNA-seq subset of the multiomic dataset and

$M_Y^* \in \mathbb{R}^{n_{\text{multiomic}} \times d_{\text{ATAC}}}$ is the harmonized space for the scATAC-seq subset of the multiomic dataset.

In this work, we treat the scRNA-seq dataset X as a reference and map the multiomic gene expression profiles (M_X) onto this reference using the FindTransferAnchors and MapQuery functions in Seurat to obtain X and M_X^* . An example workflow is provided at https://satijalab.org/seurat/articles/integration_mapping.html ('Mapping and Annotating Query Datasets').

The same functionality has been implemented in the Signac package for the mapping and harmonization of scATAC-seq datasets (https://satijalab.org/signac/articles/integrate_atac.html). However, we emphasize that our approach is compatible with a wide variety of preexisting approaches for within-modality harmonization, including all the methods listed above.

We also note that when finding anchors between the bridge and query datasets, we can leverage the multimodal nature of the bridge dataset to perform 'supervised' dimensional reduction, which uses both modalities when calculating a low-dimensional representation during harmonization. For example, we have previously described the use of 'supervised PCA' to learn optimized transformations from CITE-seq data^{4,77}. When working with bridge datasets that measure ATAC-seq or CUT&Tag chromatin features (for example, Paired-Tag and 10x multiome), we use an analogous procedure for supervising the latent semantic indexing reduction.

Construction of a dictionary representation for each unimodal dataset.

The goal of dictionary learning is to reconstruct individual data points as a weighted linear combination of atoms in a dictionary. We treat M as a dictionary, with each row of this matrix representing an atom. We aim to learn reconstructions of X and Y based on the atoms of M while minimizing the error between the original and reconstructed values. Specifically, we aim to identify the matrices D_X and D_Y , where

$D_X \in \mathbb{R}^{n_{\text{scRNA-seq}} \times n_{\text{multiomic}}}$ is the dictionary representation of the scRNA-seq dataset, and

$D_Y \in \mathbb{R}^{n_{\text{scATAC-seq}} \times n_{\text{multiomic}}}$ is the dictionary representation of the scATAC-seq dataset, such that

$$\arg \min_{D_X} (\|D_X(M_X^*) - X^*\|_F^2 + \|D_X\|_F^2)$$

and

$$\arg \min_{D_Y} (\|D_Y(M_Y^*) - Y^*\|_F^2 + \|D_Y\|_F^2).$$

As described in refs. 56,78, this optimization problem is analogous to matrix regression and has a closed-form solution for calculating D_X and D_Y ,

$$D_X = X^*(M_X^*)^\dagger$$

$$D_Y = Y^*(M_Y^*)^\dagger,$$

where \dagger represents the pseudoinverse of the matrix.

We note that D_X and D_Y represent transformations of the original scRNA-seq and scATAC-seq datasets. While the two experiments originally measured different sets of features, after the transformation, they now are represented by the same set of features, namely, the atoms of the multiomic experiment.

Dimensional reduction via Laplacian eigen decomposition. After the datasets have been transformed in the previous step, it is possible to integrate them directly. The dimensionality of the datasets is based on the number of cells in the multiomic dataset. Unlike the original measurements, the dictionary representations are not sparse. As multiomic datasets often consist of thousands of cells, working with high-dimensional and non-sparse dictionary representations is computationally inefficient. We therefore aimed to reduce the dimensionality of the dictionary representation. Motivated by a similar problem addressed by Laplacian eigenmaps⁴², a non-linear dimensionality reduction technique, we perform dimensionality reduction by computing an eigen decomposition of the graph Laplacian matrix. Unlike a PCA, which aims to identify low-dimensional representations that preserve data variance, Laplacian eigenmaps represent a low-dimensional reduction that optimally preserves the graph-defined local neighbor relationships⁴².

We first compute a graph representation of the multiomic dataset M . We use a 'shared nearest neighbor' graph representation, as proposed by Levine et al.⁷⁹ for clustering single-cell datasets. We note that the matrix representation of this graph is symmetric, which is a

requirement for downstream eigen decomposition. Our approach is compatible with any user-defined distance metric when constructing this graph, although we recommend using either the Euclidean distance based on harmonized gene expression measurements (that is, M_X^*) or, alternately, a weighted combination of modalities using the ‘weighted nearest neighbor’ distance metric that we have previously introduced⁴. We define

$G \in \mathbb{R}^{n_{\text{multiomic}} \times n_{\text{multiomic}}}$ as the symmetric graph representation of the multiomic dataset and

$$L = I - D^{-\frac{1}{2}} G D^{-\frac{1}{2}}$$

as the graph Laplacian matrix.

We next perform an eigen decomposition of the graph Laplacian matrix:

$$L = U \Lambda U^T$$

$$0 = \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_n.$$

Here, U_L is the leftmost $n_{\text{Laplacian}}$ eigenvectors of L , where n specifies the reduced dimensionality of the dataset. We select $n_{\text{Laplacian}} = 50$ for all examples in this work.

We now multiply the learned dictionary representations for the scRNA-seq and scATAC-seq datasets by this truncated set of eigenvectors. Doing so transforms these representations into the same lower-dimensional space ($n_{\text{Laplacian}}$). We define

$L_X \in \mathbb{R}^{n_{\text{scRNA-seq}} \times n_{\text{Laplacian}}}$ as the reduced dictionary representation for the scRNA-seq data,

$L_Y \in \mathbb{R}^{n_{\text{scATAC-seq}} \times n_{\text{Laplacian}}}$ as the reduced dictionary representation for the scATAC-seq data and

$L_M \in \mathbb{R}^{n_{\text{multiomic}} \times n_{\text{Laplacian}}}$ as the reduced dictionary representation for the multiomic dataset and calculate the following matrices:

$$L_X = D_X U_L = X^* ((M_X^*)^\dagger U_L),$$

$$L_Y = D_Y U_L = Y^* ((M_Y^*)^\dagger U_L)$$

and

$$L_M = U_L.$$

Alignment of dictionary representations across datasets. Both the scRNA-seq and scATAC-seq datasets have now been transformed into a low-dimensional space defined by the same set of features. They can now be directly harmonized using existing methods. As in step 1, multiple published methods can accomplish this goal. In this work, we use our internal implementation of the mnnCorrect integration technique to perform this harmonization³⁹. We choose mnnCorrect, as we find that after performing the steps described above, any remaining sample-specific differences are minor and are typically far less than the differences we observe when aligning scRNA-seq datasets across different technologies. To demonstrate the compatibility of our approach with alternative methods, we also repeat our quantitative benchmarking experiments using our previously developed integration workflow in Seurat v3 (ref. 19) and observe very similar results (Supplementary Fig. 3).

Specifically, the final output of our procedure represents

$L_X^* \in \mathbb{R}^{n_{\text{scRNA-seq}} \times n_{\text{Laplacian}}}$ as the harmonized reduced dictionary representation for the scRNA-seq data,

$L_Y^* \in \mathbb{R}^{n_{\text{scATAC-seq}} \times n_{\text{Laplacian}}}$ as the harmonized reduced dictionary representation for the scATAC-seq data and

$L_M^* \in \mathbb{R}^{n_{\text{multiomic}} \times n_{\text{Laplacian}}}$ as the harmonized reduced dictionary representation for the multiomic dataset.

These representations can be used as input for common downstream analytical tasks, including t -distributed stochastic neighbor embedding (t-SNE) or UMAP visualization, graph-based clustering and the identification of developmental trajectories.

Atomic sketch integration

Our approach consists of four steps: (1) for each dataset, sample a representative subset of cells (atoms) that span both rare and abundant populations; (2) for each dataset, learn a dictionary representation to reconstruct each cell based on the atoms; (3) integrate the atoms from each dataset and (4) for each dataset, reconstruct each cell from the integrated atoms. Each step is described in detail below. We note that steps 1, 2 and 4 are performed on each dataset individually, and step 3 only requires performing joint computation on the downsampled set of atoms. Therefore, our procedure never requires loading or processing the entirety of the datasets at one time. Our approaches should therefore successfully extend to and beyond the analysis of 100,000,000 cells, which is now an achievable scale for combinatorial barcoding technologies.

All methods are implemented in our open-source R package Seurat (www.satijalab.org/seurat, www.github.com/satijalab/seurat).

Sample a representative subset of cells (‘atoms’) from each dataset. Our first step is to selectively downsample the cells in each dataset, aiming to identify a reduced set of cells that are representative of the full dataset. In particular, we aim to ensure that rare populations continue to be represented even after downsampling. We also aim to identify cell subsets in a computationally efficient manner and to minimize any computation that must be performed on the full dataset before downsampling. We aim to select a subset of k cells from each dataset, each of which is referred to as an atom. In this manuscript, we use $k = 5,000$, unless otherwise noted.

We define

$X \in \mathbb{R}^{n_{\text{scRNA-seq}} \times d_{\text{genes}}}$ as the count matrix for scRNA-seq and

$S \in \mathbb{R}^{k \times n_{\text{scRNA-seq}}}$ as the sampling matrix for the dataset; each row is one-hot row vector matrix indicating which cells are selected (that is, $S_{i,j} = 1$ if cell i is the j th cell to be selected; $i = 1, 2, \dots, n_{\text{cells}}$ and $j = 1, 2, \dots, k$).

$SX \in \mathbb{R}^{k \times d_{\text{genes}}}$ is the scRNA-seq matrix after downsampling to the k cells selected. We also call this matrix A, as it represents the ‘atoms’ selected from the original dataset.

We can use a variety of techniques to define the sketching matrix S. These include geometric sketching techniques, such as geosketch⁵³ or Hopper⁵⁴, or fast clustering procedures, such as mini-batch k -means⁵⁵ followed by cluster-informed downsampling.

In this work, we select cells based on their statistical leverage scores, a method for selecting influential data points in a dataset. In the context of linear regression, statistical leverage represents the influence of an individual data point in determining the best least-squares fit. In this context, cells with high leverage scores will tend to make the largest contribution to the gene covariance matrix and, therefore, reflect the importance of the cell’s profile. The exact statistical leverage score for a cell can be computed via an eigen decomposition of the X matrix, but this is computationally inefficient. As an alternative, Clarkson and Woodruff⁵⁶ propose a randomized algorithm that efficiently computes a fast approximation of statistical leverage⁵⁶. This algorithm is attractive for single-cell sequencing analysis as it is highly scalable and runs efficiently on sparse datasets. Briefly, the algorithm amounts to constructing a ‘randomized’ sketch of the input matrix based on the Johnson–Lindenstrauss lemma and computing the Euclidean norms of the rows of that sketch. The algorithm is fully described in Clarkson and Woodruff⁵⁶, but we note the key mathematical steps below.

For the randomized sketching matrix, we use the sparse random CountSketch matrix C, which consists of 0, 1 and -1 elements and is defined in ref. 80.

$C \in \mathbb{R}^{C \times n_{\text{scRNA-seq}}}$ is the sparse randomized CountSketch matrix.

We then perform a QR decomposition

$$CX = [Q, R].$$

We then apply a fast Johnson–Lindenstrauss transformation using a very sparse random projection matrix Π ⁸¹. We calculate this matrix using the RandPro package⁸² in R ('li' projection function),

$$Z = X \times (R^{-1} \times \Pi).$$

We can now calculate the leverage score for each cell, which are the Euclidean norms of the rows of the Z matrix. We can also calculate a sampling probability for selecting each cell i as an atom based on the leverage scores.

$$l_i = \|Z[i,]\|_2^2 \text{ is the leverage score for cell } i, \text{ and}$$

$$p_i = \frac{l_i}{\sum_{j=1}^n l_j} \text{ is the probability of selecting cell } i \text{ as an atom.}$$

Finally, we sample k cells as atoms based on these probabilities. As described above, this procedure results in a downsampled dataset in which only the atoms remain, which we name A .

Learn a dictionary representation to reconstruct each cell based on the atoms. We aim to learn reconstructions of X based on the atoms of A while minimizing the error between the original and reconstructed values. Specifically, we aim to identify the matrix D , where

$D \in \mathbb{R}^{n_{\text{scRNA-seq}} \times k}$ is the dictionary representation of the scRNA-seq dataset such that

$$\arg \min_D (\|DA - X\|_F^2 + \|D\|_F^2).$$

As described previously, this optimization problem is analogous to matrix regression and has a closed-form solution for calculating D

$$D = XA^\dagger,$$

where \dagger represents the pseudoinverse of the matrix.

Integrate the atoms from each dataset. Let $i = 1, 2, \dots, n_{\text{dataset}}$ represent the datasets to be integrated, and let A_i represent the matrix of atoms that result from downsampling dataset i . Our goal is to harmonize the set of matrices $[A_1, A_2, \dots, A_{n_{\text{dataset}}}]$.

This can be performed with a wide variety of existing tools for the harmonization of single-cell datasets. For example, Seurat, Harmony, LIGER, scVI, Scanorama, fastMNN, scVI and scArches all learn a shared low-dimensional space that jointly represents the datasets and aligns cells in a matched biological state together. Our goal is therefore to learn

$$[A_1^*, A_2^*, \dots, A_{n_{\text{dataset}}}^*],$$

where $A_i^* \in \mathbb{R}^{n_{\text{scRNA-seq}} \times d_{\text{RNA}}}$ is the harmonized space for scRNA-seq dataset i .

In this manuscript, we use our previously developed anchor-based workflow to integrate the datasets using reciprocal PCA, which is optimized for integration tasks with large numbers of samples and cells ('fast integration using reciprocal PCA' at https://satjalab.org/seurat/articles/integration_rpca.html). The integration procedure returns a low-dimensional space that jointly represents atoms from all datasets.

Reconstruct each cell from the integrated atoms. The last step is performed individually for each dataset. Let $i = 1, 2, \dots, n_{\text{dataset}}$ represent the datasets to be integrated, and let X_i represent the full scRNA-seq count matrix representing dataset i .

We reconstruct integrated values for each cell in dataset i using the previously computed dictionary representation for the dataset along with the harmonized space A_i^* ,

$$X_i^* = D_i A_i^* = X_i (A_i^\dagger A_i^*).$$

The collection of matrices $[X_1^*, X_2^*, \dots, X_{n_{\text{dataset}}}^*]$ now represents a low-dimensional space that jointly represents all cells from all datasets. Because these matrices are low dimensional, each of them can be simultaneously loaded into memory. These representations can be used as input for common downstream analytical tasks, including t-SNE or UMAP visualization, graph-based clustering and the identification of developmental trajectories.

Preprocessing details for each dataset

Adult mouse frontal cortex and hippocampus Paired-Tag dataset. The datasets from Zhu et al.²⁶ were generated with Paired-Tag, which performs simultaneous profiling of histone modifications and cellular transcriptomes and contains a total of 64,849 nuclei. We extracted three datasets for the histone modifications H3K27ac, H3K4me1 and H3K27me3. We used the gene expression matrices as quantified in the original experiment. For each epigenetic modification, the original manuscript quantified read densities in 5,000 bins. These were aggregated into larger peaks using the CombineTiles function in Signac, and aggregated peaks less than 1 megabase in size were retained. We retained cells with total RNA counts between 500 and 10,000. We applied SCTransform to normalize the gene expression data and TF-IDF to normalize the histone modification data. We used PCA (dimensions 1:30) and TF-IDF (dimensions 2:30, excluding the first dimension, as this is typically correlated with technical metrics in ATAC-seq or scCUT&Tag data) to reduce the dimensionality of the RNA and histone modification modalities and construct the weighted nearest neighbor (WNN) graph.

Data acquisition source: Gene Expression Omnibus (GEO), accession number [GSE15200](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE15200).

Human frontal cortex snmC-seq data. This human frontal cortex dataset is an snmC-seq dataset from Luo et al.⁵¹ and contains 2,784 nuclei. We used the non-CG methylation 100,000-kb bin count matrices as quantified in the original experiment. We applied SCTransform⁸³ to normalize the gene expression data and log normalization to normalize the methylation data. Because this dataset was used as a query dataset in this manuscript, we did not perform unsupervised dimensionality reduction on the methylation data.

Data acquisition source: GEO, accession number [GSE97179](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE97179) (https://brainome.ucsd.edu/annoj/brain_single_nuclei/).

Human frontal cortex snmC2T-seq data. This human frontal cortex dataset is an snmC2T-seq dataset from Luo et al.²⁸ and contains 4,357 nuclei. We used the non-CG methylation 100,000-kb bin count matrices as quantified in the original experiment. We applied SCTransform to normalize the gene expression data and log normalization to normalize the methylation data. We used PCA to reduce the dimensionality to 30 for both datasets and construct the WNN graph.

Data acquisition source: GEO, accession number [GSE140493](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE140493).

BMMC multiome. We collected a total of ten 10x multiome datasets from the NeurIPS Multimodal Single-Cell Data Integration challenge website, representing 32,368 paired single-nucleus profiles of transcriptome and chromatin accessibility. We retained cells with total RNA counts between 1,000 and 10,000 and with total ATAC peak counts between 2,000 and 30,000. We applied SCTransform to normalize the gene expression data and TF-IDF to normalize ATAC peak counts. We used PCA (dimensions 1:40) and TF-IDF (dimensions 2:40) to reduce the dimensionality of each modality and construct the WNN graph.

Data acquisition source: https://openproblems.bio/competitions/neurips_2021/.

Human BMMC ATAC-seq. This human bone marrow dataset is an snATAC-seq dataset from Granda et al.⁴³ As the reads were originally mapped to hg19, we used cellranger-atac v2 to remap fastq files to hg38. In each cell, we quantified the same set of peaks that were detected in

the BMMC multiome dataset. After removing low-quality cells, 26,159 cells were retained, with total ATAC peaks of <50,000 and >2,000. We applied TF-IDF to normalize the ATAC-seq data. As this dataset was used as a query dataset in this manuscript, we did not perform unsupervised dimensionality reduction on the ATAC-seq data.

Data acquisition source: GEO, accession number [GSE139369](#).

Human PBMC scRNA. This human PBMC scRNA dataset was obtained from the 10x Genomics website (<https://www.10xgenomics.com/resources/datasets/>) and consists of 33,015 cells. We retained cells with total RNA counts between 400 and 10,000. We applied log normalization for the gene expression matrix. We annotated these cells by mapping them to the Azimuth PBMC reference with the Seurat4 reference-mapping framework and refined the annotations by de novo clustering. These data were used for sketching benchmark analysis (Supplementary Fig. 5).

Data acquisition source: <https://support.10xgenomics.com/single-cell-gene-expression/datasets/1.1.0/pbmcs3k>.

Human PBMC multome. This human PBMC multiome (RNA + ATAC) dataset was obtained from the 10x Genomics website (<https://www.10xgenomics.com/resources/datasets/>) and consists of 10,970 cells. We retained cells with total RNA counts between 500 and 10,000 and total ATAC peak counts between 2,000 and 100,000. We applied SCTransform to normalize the gene expression data and TF-IDF to normalize ATAC peak counts. We annotated these cells by mapping the RNA profile to the Azimuth PBMC RNA reference with the Seurat4 reference-mapping framework. We then used these annotations to create ATAC-seq tracks, shown in Supplementary Fig. 2d.

Data acquisition source: <https://www.10xgenomics.com/resources/datasets/10-k-human-pbm-cs-multome-v1-0-chromium-x-1-standard-2-0-0>.

Human CD34⁺ bone marrow multiome. This human CD34⁺ bone marrow multiome (RNA + ATAC) dataset was obtained from Persad et al.⁸⁴ and consists of 13,398 cells from two replicates. We retained cells with total RNA counts between 500 and 30,000 and total ATAC peak counts between 1,000 and 100,000. We used the same normalization method used for the human PBMC multiome. We annotated these cells by mapping the RNA profile to the Azimuth RNA BMMC reference with the Seurat4 reference-mapping framework. When using the human PBMC multiome dataset, we did not observe sufficient numbers of ASCDs to create a chromatin track for this dataset. However, we identified 12 cells annotated as ASCDs in these CD34⁺ bone marrow data. We used these cells to generate a chromatin track for the *SIGLEC6* locus (Supplementary Fig. 2e), which validates our predicted ASCD identified via bridge integration.

Data acquisition source: <https://zenodo.org/record/6383269>.

Human PBMC CyTOF dataset. This human PBMC CyTOF dataset was generated by the COVID-19 Multi-omics Blood Atlas COMBAT consortium and consists of 7.11 million cells with a panel of 47 antibodies. We removed cells from individuals with sepsis, yielding a remainder of 5.17 million cells. We used the normalized expression matrices as quantified in the original study. As this dataset was used as a query dataset in this manuscript, we did not perform unsupervised dimensionality reduction on the protein data.

Data acquisition source: <https://zenodo.org/record/5139561>.

Azimuth reference. Azimuth scRNA-seq references for human bone marrow (297,627 cells) and the human motor cortex (159,738 cells) were downloaded from the HuBMAP portal. The portal includes descriptions of each public data source used when compiling the reference dataset and a link to a GitHub repository and Docker Hub to reproduce the construction of the reference.

Data acquisition source: <https://azimuth.hubmapconsortium.org>.

Lung scRNA-seq dataset atlas. Nineteen datasets profiling human lung samples using scRNA-seq were downloaded from publicly available sources (links for each source dataset are provided in Supplementary Table 2). Low-quality cells were filtered using uniform quality control thresholds; cells with RNA counts between 300 and 100,000 and with mitochondrial read percentages below 20% were retained. Normalization was performed using log normalization implemented in Seurat. We used PCA (dimensions 1:40) to reduce the dimensionality of each dataset.

Data acquisition source: Supplementary Table 2 and lung scRNA datasets^{68,69,85–101}.

PBMC COVID scRNA-seq dataset atlas. Fourteen datasets profiling human PBMC samples using scRNA-seq were downloaded from publicly available sources (links for each source dataset are provided in Supplementary Table 2). Eleven of these datasets had been previously organized in Tian et al.⁶⁰. Low-quality cells were filtered using uniform quality control thresholds; cells with RNA counts between 150 and 150,000 and with mitochondrial read percentages below 15% were retained. Normalization was performed using log normalization implemented in Seurat. We used PCA (dimensions 1:40) to reduce the dimensionality of each dataset.

Data acquisition source: Supplementary Table 2 and PBMC scRNA datasets^{4,62,63,102–112}.

Differentiation trajectory and pseudotime analysis

In Fig. 2, we identify a myeloid differentiation trajectory and pseudotime ordering of cells that describes both reference (scRNA-seq) and query (scATAC-seq) cells. We extracted reference cells belonging to HSC, LMPP, GMP and CD14⁺ monocyte populations and query cells that mapped to any of these subsets after bridge integration. We next constructed a *k*-nearest neighbor (KNN) graph representing cells from both modalities using the latent space learned during the bridge integration procedure. This graph was used as input to the destiny package, which reduces the dimensionality of the data using diffusion maps¹¹³. We note that as we manually selected cell populations that are known to encompass monocytic differentiation, we did not expect or observe branching events. We used the first two diffusion map coordinates as input to monocle3 (ref. 114) to infer a pseudotemporal ordering.

We next aimed to identify cases where dynamic gene expression patterns ‘lag’ behind the accessibility dynamics of nearby regulatory regions. We can perform this analysis because our pseudotemporal ordering encompasses both scATAC-seq and scRNA-seq cells. We first associated each scATAC-seq peak with a gene using the ClosestFeature function in Signac. For each gene, we next smoothed the expression profile along the learned trajectory using the ksmooth function (‘stats’ package in R¹¹⁵) using 1,000 intervals and a bandwidth of 0.01. We repeated the same process for the accessibility of each peak linked to this gene (bandwidth of 0.05). We next calculated the cross-correlation of the smoothed expression and accessibility values, which measures the similarity for the two time series and calculates the optimal displacement of one relative to the other. We used the ccf function (‘stats’ package in R¹¹⁵) and identified a total of 574 gene–peak pairs with a cross-correlation of >0.6. Of these, we identified 236 cases exhibiting an optimal displacement of >0.01 (we illustrate 6 such cases in Fig. 2l).

Bridge cell downsampling analysis

To explore how the size and composition of the multiomic dataset affected the robustness of bridge integration, we performed 25 serial downsampleings of the entire BMMC multiomic dataset (200, 300, 400, 500, 600, 700, 800, 900, 1,000, 2,000, 3,000, 4,000, 5,000, 6,000, 7,000, 8,000, 9,000, 10,000, 11,000, 12,000, 13,000, 14,000, 15,000, 20,000 and 30,000). We used one batch of the scATAC dataset (12,256 cells) as a query, repeated bridge integration and compared the resulting predictions with our original findings. As expected, we

found that the degree of agreement after downsampling was cell-type dependent, as cells from abundant cell types were more robust to downsampling. We therefore expressed our results as a function of the number of cell types present in the bridge dataset for each cell type. For example, the 7,000-cell downsampled dataset contained 144 CD16⁺ monocytes (prediction concordance of 1.00) and 22 pro-B cells (prediction concordance of 0.66). The 2,000-cell downsampled dataset contained 41 CD16⁺ monocytes (prediction concordance of 0.94) and 6 pro-B cells (prediction concordance of 0.55). We aggregated all these results across downsamples and displayed the results in Fig. 2a. For visual clarity, we only showed an x axis range of 10 to 500 in Fig. 2a.

Bridge cell-type composition resampling analysis

To assess the robustness of our bridge integration procedure to the relative proportion of cell types in the bridge dataset, we scrambled the proportions and then repeated the bridge integration procedure. To accomplish this, we sampled 10,000 cells from the original bridge dataset without replacement. We set each cell's sampling probability inversely to the proportion of cell types in the original dataset, ensuring that we would substantially alter cell-type composition. We then repeated bridge integration with the resampled dataset, mapping the same query dataset, and then compared the results in Supplementary Fig. 2.

Benchmark analysis with multiVI and Cobolt

To assess the performance of our bridge integration method alongside other recently proposed integration tools, we compared our results with multiVI⁴⁹ from scvi-tools v 0.14.5 and Cobolt⁵⁰ (v1.0.0). As both Cobolt and multiVI use variational autoencoders, both methods are run on a server with a discrete NVIDIA A100GPU with 40 gigabytes of memory and pyTorch-lightning v1.3.8 installed. Seurat analyses are run on an Intel Xeon Platinum 8280L server and use a single computational core.

For multiVI, we used the scRNA-seq, scATAC-seq and multiomic RNA–ATAC paired counts matrices as input. We used the multiome_anndatas function to generate one anndata object for integration. We set batch information in categorical_covariate_keys, using the setup_anndata function. We then integrated the datasets by running the multiVI function, as outlined in the multiVI tutorial (https://docs.scvi-tools.org/en/stable/tutorials/notebooks/MultiVI_tutorial.html). We used 500 epochs for model training, as suggested in the multiVI tutorial. All other parameters were set to default settings. multiVI learns a latent space, which jointly represents cells across the scRNA-seq and scATAC-seq datasets. We extracted this space and performed nearest neighbor calculations and UMAP visualization in Seurat.

For Cobolt, we used the scRNA-seq, scATAC-seq and multiomic RNA–ATAC paired counts matrices as input. We used the SingleData function from cobolt_utils to generate three Cobolt objects and trained the model using 20 latent variables, a 0.001 learning rate and 100 iterations, as recommended in the Cobolt tutorial (<https://github.com/epurdom/cobolt/blob/master/docs/tutorial.ipynb>). All other parameters were set to default or were the recommended settings in the tutorial. Cobolt learns a latent space that jointly represents cells across the scRNA-seq and scATAC-seq datasets. We extracted this space and performed nearest neighbor calculations and UMAP visualization in Seurat.

We performed comparative benchmarking in three contexts. First, we ran all three approaches on the datasets from Fig. 2, aiming to map an scATAC-seq query dataset onto an scRNA-seq-defined reference. We did not have ground truth information for this dataset, so we did not calculate quantitative benchmarks, although we visualized the performance of all methods in Fig. 3b and Supplementary Fig. 2. As multiVI and Cobolt do not provide methods to explicitly label query scATAC-seq cells using scRNA-seq references, we used a commonly used heuristic for label transfer; for each scATAC-seq cell, we identified the closest five neighbors in scRNA-seq cells and transferred the most common cell annotation among the neighbors. In Fig. 3b, we visualized chromatin

accessibility at the *SIGLEC6* locus for cells predicted as ASDCs by all methods, and additional loci are shown in Supplementary Fig. 2.

Second, we performed quantitative benchmarking in a context where we had a ground truth dataset to establish the accuracy of scATAC-seq/scRNA-seq integration. We split the BMMC multiomic dataset into two groups. The first group consists of a randomly sampled subset of 2,115 cells representing at most 100 cells per author-defined cell type. This group of cells was used as the multiomic bridge dataset for benchmarking. The remaining cells were placed in the second group and were split into separate scRNA-seq and scATAC-seq profiles (that is, the multiomic pairing information was temporarily discarded). We then integrated the datasets using bridge integration (using both Seurat v3 and mnnCorrect for the final alignment step), multiVI or Cobolt. After integration, all methods return a latent space that jointly represents cells from both the scATAC-seq and scRNA-seq datasets. For each scATAC-seq cell, we know its matched scRNA-seq profile, as they were originally measured simultaneously. Successful integration techniques will place matched profiles close together in this latent space. For each scATAC-seq cell, we therefore calculated the Jaccard similarity metric to its matched scRNA-seq profile (we note that this similarity metric is symmetric). We report these results in Fig. 3c and Supplementary Fig. 3, either averaged together across all cells or averaged within author-defined cell types.

Third, we repeated the ground truth benchmarking analysis on a second multiomic technology. Paired-Tag enables simultaneous CUT&Tag and transcriptomic profiling in single cells. We used data for three histone modifications: H3K27ac, H3K27me3 and H3K4me1. As each dataset consists of multiple replicates, we used replicate 1 as the multiomic dataset and split the CUT&Tag and RNA modalities from the second replicate for benchmarking. We ran multiVI, Cobolt and bridge integration (using both Seurat v3 and mnnCorrect for the final alignment step) as before, substituting the CUT&Tag counts matrix for the scATAC-seq matrix, as previously described.

Bridge-free benchmark analysis with Seurat-CCA and Liger

To benchmark bridge-free integration methods for the cross-modality integration of scATAC and scRNA data, we initially transformed ATAC peaks into gene activity scores using the function GeneActivity in the Signac package. This generates a gene activity score matrix by summing peak counts per cell in the gene body and promoter region. scRNA gene expression and scATAC gene activity score count matrices were used as input for integration. We performed integration by following the procedure from publicly available vignettes, https://satijalab.org/seurat/articles/atacseq_integration_vignette.html and http://htmlpreview.github.io/?https://github.com/welch-lab/liger/blob/master/vignettes/Integrating_scRNA_and_scATAC_data.html.

Classification metrics for benchmark analysis

In addition to Jaccard similarity, we also calculated additional quantitative benchmarking metrics (discussed below) that leverage predefined cellular annotations.

Multiclass classification area under the receiver operating characteristic (ROC) curve (AUC). We performed a one versus rest multiclass ROC analysis because cell-type annotations include multiple classes. We assessed multiclass predictions by iteratively contrasting each class with all the others. For each iteration, we designated one class as the ‘positive’ class and the remaining classes as the ‘negative’ classes. Combining the prediction score for the ‘positive’ class, we calculated the AUC for each. We report the average AUC value across all cell types.

Query annotation KNN purity. This metric quantifies the consistency between cell-type labels and neighbor relationships in the latent space. For each query cell, it measures the fraction of neighbors that receive the same annotation as the query cell itself. Using the mapped query

dataset, we calculated a KNN graph in the integrated space to find $k = 30$ nearest neighbors for each query cell and calculated the fraction of cells receiving the same annotation.

Multiclass binary cross entropy. Multiclass binary cross entropy is a commonly used metric in classification and machine learning tasks¹¹⁶ and considers both the accuracy of prediction and the associated prediction score,

$$L(y, \hat{y}) = - \sum_i^K y^{(i)} \log \hat{y}^{(i)},$$

where K is the number of potential classes (cell types), $y^{(i)}$ is an indicator variable that denotes a correct (1) or incorrect (0) prediction, and $\hat{y}^{(i)}$ is the prediction score (probability associated with predicting class i).

Bridge cell-type remove analysis. We removed certain cell types from the bridge and reperformed bridge integration to characterize the performance of our method in situations where cell populations were missing from the bridge dataset. We separately deleted the CD8⁺ T cell, pDC, CD14⁺ monocyte and B cell subpopulations from the BMMC multiome benchmark dataset, respectively. We then repeated the bridge integration procedure using the modified bridge dataset. We then compared the predicted labels (and prediction scores) assigned to query cells based on the full and modified bridge datasets.

Bridge cell-type RNA and ATAC UMI downsample analysis. To simulate reduced data quality for RNA or ATAC modalities in the multiome dataset, we downsampled RNA or ATAC UMI counts in the multiome dataset by 1, 5, 10, 20, 30, 40, 50, 60, 70, 80 and 90 using downsampleMatrix from the scuttle package¹¹⁷. We renormalized RNA or ATAC data after downsampling and repeated the bridge integration procedure. We assessed the prediction results using a number of evaluation methods, including classification AUC, query KNN purity, multiclass binary cross entropy and Jaccard similarity.

Sketching benchmark analysis and evaluation metrics

We applied the three assessment metrics listed below to evaluate the performance of sketching algorithms. These metrics assess the ability of sketching algorithms to identify a compact subset of cells that is maximally representative of the full dataset and, in particular, to retain cells from rare subpopulations in the dataset sketch. We computed metrics on two datasets: a 33,000-cell dataset of human PBMCs publicly available from 10x Genomics (<https://support.10xgenomics.com/single-cell-gene-expression/datasets/1.0/pbm33k>) and a 66,000-cell lung scRNA-seq dataset⁹⁸. Using two sketching algorithms (leverage score-based sketching and geosketch⁵³), we sketched 100, 300, 500, 1,000, 1,500, 2,000, 2,500, 3,000, 3,500, 4,000, 4,500, 5,000, 6,000, 7,000, 8,000, 9,000 and 10,000 cells from the entire dataset individually. We also performed a uniform downsampling method as a negative control. To measure computing time, we altered the size of the original dataset and fixed the number of sampled cells to 5,000.

Change of gene covariance matrix. Optimal cell sketches should modify cellular density but preserve information in the dataset gene-gene covariance matrix. We therefore calculated the magnitude of the difference between the gene covariance matrix calculated on the full dataset and the gene covariance matrix calculated on the sketched dataset. Before calculating the covariance matrix, we first performed a PCA on the dataset and reconstructed a gene expression matrix using the top 50 principal components. We then calculated the Frobenius norm of changes in the gene covariance matrix using the following equation:

$$X_r = U_r \sum_r V_r^T$$

$$X_S = SX,$$

$$\left\| \frac{X_r^T X_r - X_S^T X_S}{n-1} \right\|_F^2,$$

where $X_r \in \mathbb{R}^{n \times d}$ is the PCA denoised full gene expression matrix, S is the sampling matrix for the dataset to sample k cells from the full matrix, and $X_S \in \mathbb{R}^{k \times d}$ is the denoised gene expression matrix for the k sampled cells.

Cell-type entropy. We assessed the evenness of cell types in the sketched data based on the original annotations from the dataset. Cell-type entropy will increase when the sketched data effectively represent rare cell types. Cell-type entropy will decrease when abundant cell types dominate the sketched data and rare cell types are not represented.

Hausdorff distance. We also evaluated the performance of sketching using Hausdorff distance, a metric fully described in the geosketch manuscript⁵³. The Hausdorff distance measures the largest closest distance between the full and sketched datasets. A low Hausdorff distance indicates that all cells in the full dataset are represented by the sketched cells.

Community-wide integration analyses

To facilitate the harmonization and subsequent meta-analysis of a diversity of publicly available scRNA-seq datasets, we applied our atomic sketch integration approach to 1,525,710 scRNA-seq profiles spanning 19 publicly available human lung scRNA-seq datasets. As described above, we calculated a leverage score for each cell in each dataset and used this to sample 5,000 cells as atoms. We found that these 5,000 cells retain rare cell types, despite downsampling (Supplementary Fig. 5). We learned a dictionary representation that reconstructs cells from each dataset based on the selected atoms using the methods described above. We used our previously developed reciprocal PCA-based integration workflow (https://satijalab.org/seurat/articles/integration_rpca.html) to integrate the 95,000 atoms originating from these 19 datasets. Finally, the learned dictionary representations can be used to reconstruct harmonized profiles (in low-dimensional space) for all 1,525,710 scRNA-seq profiles. This space was used as input for UMAP to generate the visualization in Fig. 4b,c.

The harmonized space for all 1,525,710 scRNA-seq profiles can also be used as input to graph-based clustering approaches. However, because annotation is an iterative and manual process, we chose to first perform clustering on the harmonized dataset of 95,000 atoms. We constructed a shared nearest neighbor graph and partitioned this into clusters using the graph-based smart local moving algorithm¹¹⁸. We initially clustered cells at a high resolution (resolution = 5) and performed differential expression analysis on all pairs of clusters for RNA markers. We merged clusters that did not exhibit clear evidence of separation. We removed clusters that showed clear evidence of expressing markers for two different cell types as likely doublets. To assign names to individual clusters, we used the recently published anatomical structures, cell types and biomarker tables¹¹⁹, except for five clusters (adventitial fibroblast, alveolar fibroblast, myofibroblast and proliferating NK/T, squamous), where our desired annotation was not present in the most recently available version of the table (v1)¹²⁰. For each cell in the full dataset, we found its ten nearest neighbors among the annotated atoms and transferred the most commonly observed annotation.

Lung integration evaluation. We computed two evaluation metrics to assess the performance of the integration of 19 lung datasets. Local inverse Simpson index (LISI) is used to evaluate for batch effect correction, and KNN purity is used to evaluate preservation of the original

labels in the integrated space. We merged the raw RNA expression from 19 lung datasets, normalized them and performed a PCA as a control for the integration result.

To compare the batch effects, we first computed the LISI score using the top 50 dimensions of cell embeddings in the RNA PCA and integrated latent space individually. We used the same integrated latent space to determine KNN purity.

In Fig. 5, we performed ‘community-wide’ integration on 3.46 million cells spanning 639 individuals and 14 studies. As these studies varied widely in the number of cells present in each dataset, we selected at least 5,000 and at most 10% of the cells in each dataset as atoms based on their leverage score. This enabled the larger and more comprehensive datasets to contribute additional weight to the integrated reference. We performed integration, reconstruction and annotation using the same steps as described for the lung.

Identifying DE genes across cell types and conditions

In the lung and PBMC community-wide integration, we identified DE genes on the ‘pseudobulk’ expression values calculated from each individual study. We performed a logistic regression-based method to identify DE genes. For space considerations, we typically reported only the top 10 markers in each heat map and sorted genes first by adjusted *P* value and next by log (fold change) to determine the top markers. To compare the results of single-cell and bulk analyses, we used the wilcoxauc method from presto¹²¹ to identify DE genes using either the single-cell or pseudobulk profiles as input and sorted by the AUC statistic. In Fig. 4g, we compared the distribution of average expression values (within a cell type) for the top 100 markers identified by either single-cell or pseudobulk analysis.

To identify COVID-19 response signatures that are consistent across multiple individuals, we first calculated a pseudobulk average for CD14⁺ monocytes for each of the 506 donors who were either healthy or whose metadata indicated mild, moderate or severe COVID-19 (ref. 60). We performed DE analysis at the pseudobulk level to identify markers of CD14⁺ monocytes expressed in severe COVID samples compared to healthy samples. In Fig. 5b, we ordered each pseudobulk profile by the expression levels of these genes, which are enriched for interferon response genes, for visualization. We repeated this process for eight additional cell states in Supplementary Fig. 7b.

Reporting summary

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

Data availability

We used publicly available datasets in this work. Download locations for each dataset are listed in the Supplementary Methods and Supplementary Tables. Azimuth references are available for download at <http://azimuth.hubmapconsortium.org>.

Code availability

Bridge integration and atomic sketch integration are implemented as part of the Seurat R package. In this work, we also make use of the Signac and Azimuth packages. All are freely available as open-source software at the following websites: <https://github.com/satijalab/seurat>, <https://github.com/timoast/signac> and <https://github.com/satijalab/azimuth>. We include two vignettes describing the ‘bridge integration’ and ‘atomic sketch integration’ procedures as Supplementary Notes with this manuscript.

References

77. Barshan, E., Ghodsi, A., Azimifar, Z. & Jahromi, M. Z. Supervised principal component analysis: visualization, classification and regression on subspaces and submanifolds. *Pattern Recognit.* **44**, 1357–1371 (2011).
78. Woodruff, D. P. Sketching as a tool for numerical linear algebra. Preprint at <https://doi.org/10.48550/arXiv.1411.4357> (2014).
79. Levine, J. H. et al. Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. *Cell* **162**, 184–197 (2015).
80. Charikar, M., Chen, K. & Farach-Colton, M. in *International Colloquium on Automata, Languages, and Programming* 693–703 (Springer, 2002).
81. Li, P., Hastie, T. J. & Church, K. W. in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 287–296 (Association for Computing Machinery, 2006).
82. Siddharth, R. & Aghila, G. RandPro—a practical implementation of random projection-based feature extraction for high dimensional multivariate data analysis in R. *SoftwareX* **12**, 100629 (2020).
83. Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol.* **20**, 296 (2019).
84. Persad, S. et al. SEACells infers transcriptional and epigenomic cellular states from single-cell genomics data. *Nat. Biotechnol.*, 1–12 (2023).
85. Adams, T. S. et al. Single-cell RNA-seq reveals ectopic and aberrant lung-resident cell populations in idiopathic pulmonary fibrosis. *Sci. Adv.* **6**, eaba1983 (2020).
86. Bischoff, P. et al. Single-cell RNA sequencing reveals distinct tumor microenvironmental patterns in lung adenocarcinoma. *Oncogene* **40**, 6748–6758 (2021).
87. Chua, R. L. et al. COVID-19 severity correlates with airway epithelium–immune cell interactions identified by single-cell analysis. *Nat. Biotechnol.* **38**, 970–979 (2020).
88. Delorey, T.M. et al. COVID-19 tissue atlases reveal SARS-CoV-2 pathology and cellular targets. *Nature* **595**, 107–113 (2021).
89. Deprez, M. et al. A single-cell atlas of the human healthy airways. *Am. J. Respir. Crit. Care Med.* **202**, 1636–1645 (2020).
90. Eraslan, G. et al. Single-nucleus cross-tissue molecular reference maps toward understanding disease gene function. *Science* **376**, eabl4290 (2022).
91. Habermann, A. C. et al. Single-cell RNA sequencing reveals profibrotic roles of distinct epithelial and mesenchymal lineages in pulmonary fibrosis. *Sci. Adv.* **6**, eaba1972 (2020).
92. Lukassen, S. et al. SARS-CoV-2 receptor ACE2 and TMPRSS2 are primarily expressed in bronchial transient secretory cells. *EMBO J.* **39**, e105114 (2020).
93. Madissoon, E. et al. scRNA-seq assessment of the human lung, spleen, and esophagus tissue stability after cold preservation. *Genome Biol.* **21**, 1 (2019).
94. Mayr, C.H. et al. Integrative analysis of cell state changes in lung fibrosis with peripheral protein biomarkers. *EMBO Mol. Med.* **13**, e12871 (2021).
95. Melms, J. C. et al. A molecular single-cell lung atlas of lethal COVID-19. *Nature* **595**, 114–119 (2021).
96. Morse, C. et al. Proliferating SPP1/MERTK-expressing macrophages in idiopathic pulmonary fibrosis. *Eur. Respir. J.* **54**, 1802441 (2019).
97. Reyfman, P. A. et al. Single-cell transcriptomic analysis of human lung provides insights into the pathobiology of pulmonary fibrosis. *Am. J. Respir. Crit. Care Med.* **199**, 1517–1536 (2019).
98. Travaglini, K. J. et al. A molecular cell atlas of the human lung from single-cell RNA sequencing. *Nature* **587**, 619–625 (2020).
99. Wang, A. et al. Single-cell multiomic profiling of human lungs reveals cell-type-specific and age-dynamic control of SARS-CoV2 host genes. *eLife* **9**, e62522 (2020).
100. Watanabe, N. et al. Anomalous epithelial variations and ectopic inflammatory response in chronic obstructive pulmonary disease. *Am. J. Respir. Cell Mol. Biol.* **67**, 708–719 (2022).

101. Wauters, E. et al. Discriminating mild from critical COVID-19 by innate and adaptive immune single-cell profiling of bronchoalveolar lavages. *Cell Res.* **31**, 272–290 (2021).
102. Arunachalam, P. S. et al. Systems biological assessment of immunity to mild versus severe COVID-19 infection in humans. *Science* **369**, 1210–1220 (2020).
103. Combes, A. J. et al. Global absence and targeting of protective immune states in severe COVID-19. *Nature* **591**, 124–130 (2021).
104. Lee, J. S. et al. Immunophenotyping of COVID-19 and influenza highlights the role of type I interferons in development of severe COVID-19. *Sci. Immunol.* **5**, eabd1554 (2020).
105. Ren, X. et al. COVID-19 immune features revealed by a large-scale single-cell transcriptome atlas. *Cell* **184**, 1895–1913 (2021).
106. Schulte-Schrepping, J. et al. Severe COVID-19 is marked by a dysregulated myeloid cell compartment. *Cell* **182**, 1419–1440 (2020).
107. Silvin, A. et al. Elevated calprotectin and abnormal myeloid cell subsets discriminate severe from mild COVID-19. *Cell* **182**, 1401–1418 (2020).
108. Stephenson, E. et al. Single-cell multi-omics analysis of the immune response in COVID-19. *Nat. Med.* **27**, 904–916 (2021).
109. Su, Y. et al. Multi-omics resolves a sharp disease-state shift between mild and moderate COVID-19. *Cell* **183**, 1479–1495 (2020).
110. Yao, C. et al. Cell-type-specific immune dysregulation in severely ill COVID-19 patients. *Cell Rep.* **34**, 108943 (2021).
111. Yu, K. et al. Dysregulated adaptive immune response contributes to severe COVID-19. *Cell Res.* **30**, 814–816 (2020).
112. Zhu, L. et al. Single-cell sequencing of peripheral mononuclear cells reveals distinct immune response landscapes of COVID-19 and influenza patients. *Immunity* **53**, 685–696 (2020).
113. Haghverdi, L., Buettner, F. & Theis, F. J. Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics* **31**, 2989–2998 (2015).
114. Qiu, X. et al. Reversed graph embedding resolves complex single-cell trajectories. *Nat. Methods* **14**, 979–982 (2017).
115. R Core Team. R: a language and environment for statistical computing (R Foundation for Statistical Computing, 2013).
116. Bishop, C. M. & Nasrabadi, N. M. *Pattern Recognition and Machine Learning*, Vol. 4 (Springer, 2006).
117. McCarthy, D. J., Campbell, K. R., Lun, A. T. & Wills, Q. F. Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics* **33**, 1179–1186 (2017).
118. Waltman, L. & Van Eck, N. J. A smart local moving algorithm for large-scale modularity-based community detection. *Eur. Phys. J. B* **86**, 471 (2013).
119. Borner, K. et al. Anatomical structures, cell types and biomarkers of the Human Reference Atlas. *Nat. Cell Biol.* **23**, 1117–1128 (2021).
120. Gloria Pryhuber, X.S. HuBMAP ASCT+B Tables. Lung v1.1 <https://doi.org/10.48539/HBM323.SGDF.945> (2021).
121. Korsunsky, I., Nathan, A., Millard, N. & Raychaudhuri, S. Presto scales Wilcoxon and auROC analyses to millions of observations. Preprint at bioRxiv <https://doi.org/10.1101/653253> (2019).

Acknowledgements

We thank all members of the Satija Lab for thoughtful discussions related to this work. We thank A. Butler and H. Srivastava for assistance in identifying and locating scRNA-seq datasets from human lung and PBMCs. We acknowledge the Gottardo and Newell labs for publicly releasing a standardized compendium of human PBMC scRNA-seq datasets. This work was supported by the Chan Zuckerberg Initiative (EOSS-0000000082 and HCA-A-1704-01895 to R.S.) and the NIH (K99HG011489-01 to T.S.; K99CA267677 to A.S.; RM1HG011014-02, 1OT2OD026673-01, DP2HG009623-01, R01HD096770 and R35NS097404 to R.S.).

Author contributions

T.S., Y.H. and R.S. conceived the research. Y.H., T.S., M.H.K., S.C., P.H., A.H., A.S., G.M. and S.M. performed the computational analyses, supervised by C.F.-G. and R.S. Y.H., T.S. and R.S. wrote the manuscript, with input and assistance from all authors.

Competing interests

In the past 3 years, R.S. has worked as a consultant for Bristol-Myers Squibb, Regeneron and Kallyope and served as an SAB member for ImmunAI, Resolve Biosciences, Nanostring and the NYC Pandemic Response Lab. The other authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41587-023-01767-y>.

Correspondence and requests for materials should be addressed to Rahul Satija.

Peer review information *Nature Biotechnology* thanks Rhonda Bacher and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Reporting Summary

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

Statistics

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

n/a Confirmed

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

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection

No specific software was used for data collection as data were collected solely from preexisting publicly available data sets.

Data analysis

Code Availability

Bridge Integration and Atomic Sketch Integration are implemented as part of the Seurat R package. In this work, we also make use of the Signac, and Azimuth packages. All are freely available as open-source software:
[https://github.com/satijalab/seurat \(v4.1\)](https://github.com/satijalab/seurat)
[https://github.com/timoast/signac \(v1.5\)](https://github.com/timoast/signac)
[https://github.com/satijalab/azimuth \(v0.4.3\)](https://github.com/satijalab/azimuth)

Additional software packages used for analyses in this study are detailed in Methods and include: destiny(v3.10.0), monocle3(v1.2.4), stats(4.1.3), scVI-tools (v0.14.5), Cobolt(v1.0.0), pytorch(v1.3.8), Liger(v1.0.0), geosketch(v1.2)

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

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Data Availability

We used publicly available datasets in this work. Download locations for each dataset are listed in the Supplementary Methods, and also Supplementary Table 1. Azimuth references are available for download at <http://azimuth.hubmapconsortium.org>

Field-specific reporting

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

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

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

Life sciences study design

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

Sample size	No sample-size estimates were performed to ensure adequate power to detect a pre-specified effect size. Sample size selection for downsampling analyses is described in the Supplementary Methods
Data exclusions	Some single cells were excluded using quality control metrics described fully in Methods. Data QC criteria are described in the Supplementary Methods
Replication	Use of replicate samples for quantitative benchmarking is described in the Supplementary Methods
Randomization	Randomized downsampling analyses is described in the Supplementary Methods
Blinding	We did not perform blinding analyses in this study. The mapping procedure within Bridge integration was blinded to cell type labels

Reporting for specific materials, systems and methods

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

Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology and archaeology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data
<input checked="" type="checkbox"/>	<input type="checkbox"/> Dual use research of concern

Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging