Construction of continuously expandable single-cell atlases through integration of heterogeneous datasets in a generalized cell-embedding space

Lei Xiong
Tsinghua University https://orcid.org/0000-0002-2392-114X
Kang Tian
Tsinghua University
Yuzhe Li
Peking University
Qiangfeng Zhang (qczhang@tsinghua.edu.cn)
Tsinghua University https://orcid.org/0000-0002-4913-0338

Article

Keywords: COVID-19, SCALEX, disease severity, immune subtypes

Posted Date: April 28th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-398163/v1

Construction of continuously expandable single-cell atlases through integration of heterogeneous datasets in a generalized cell-embedding space

Lei Xiong[1,2,4], Kang Tian[1,2,4], Yuzhe Li[1,3], Qiangfeng Cliff Zhang[1,2,*]

[1] MOE Key Laboratory of Bioinformatics, Beijing Advanced Innovation Center for Structural Biology & Frontier Research Center for Biological Structure, Center for Synthetic and Systems Biology, School of Life Sciences, Tsinghua University, Beijing, China 100084

[2] Tsinghua-Peking Center for Life Sciences, Beijing, China 100084

[3] Academy for Advanced Interdisciplinary Studies, Peking University, Beijing, China 100871

[4] Co-first authorship

[*] Correspondence: qczhang@tsinghua.edu.cn (Q.C.Z.)

ABSTRACT

Single-cell RNA-seq and ATAC-seq analyses have been widely applied to decipher cell-type and regulation complexities. However, experimental conditions often confound biological variations when comparing data from different samples. For integrative single-cell data analysis, we have developed SCALEX, a deep generative framework that maps cells into a generalized, batch-invariant cell-embedding space. We demonstrate that SCALEX accurately and efficiently integrates heterogenous single-cell data using multiple benchmarks. It outperforms competing methods, especially for datasets with partial overlaps, accurately aligning similar cell populations while retaining true biological differences. We demonstrate the advantages of SCALEX by constructing continuously expandable single-cell atlases for human, mouse, and COVID-19, which were assembled from multiple data sources and can keep growing through the inclusion of new incoming data. Analyses based on these atlases revealed the complex cellular

landscapes of human and mouse tissues and identified multiple peripheral immune subtypes associated with COVID-19 disease severity.

INTRODUCTION

Single-cell RNA sequencing (scRNA-seq) and assay for transposase-accessible chromatin using sequencing (scATAC-seq) technologies enable decomposition of diverse cell-types and states to elucidate their function and regulation in tissues and heterogeneous systems[1-4]. Efforts like the Human Cell Atlas project[5] and Tabula Muris Consortium[6] are constructing a single-cell reference landscape for a new era of highly resolved cell research. With the explosive accumulation of single-cell studies, integrative analysis of data from experiments of different contexts is essential for characterizing heterogenous cell populations[7]. However, potentially informative biological insights are often confounded by batch effects that reflect different donors, conditions, and/or analytical platforms[8,9].

Integration methods have been developed to remove batch effects in single-cell datasets[10-16]. One common strategy is to identify similar cells or cell populations across batches. This includes the mutual nearest neighborhood (MNN) method[10] which identifies correspondent pairs of cells between two batches by searching for mutual nearest neighbors in gene expression. Scanorama[11] generalizes the process of neighbor searching from within two batches to a multiple-batch manner. Seurat v2[13] applies canonical correlation analysis (CCA) to identify common cell populations in low-dimensional embeddings across data batches, while Seurat v3[14] introduces "cell anchors" to mitigate the problem of mixing non-overlapping populations, an issue experienced in Seurat v2. Harmony[16] also applies population matching across batches, specifically through a fuzzy clustering algorithm.

It is notable that all of these cell similarity-based methods are local-based, wherein cell-correspondence across batches are identified through the similarity of individual cells or cell anchors/clusters. Accordingly, these methods all suffer from two common limitations. First, they are prone to mixing cell populations that only exist in some batches. This becomes a severe problem for the integration of datasets that contain non-overlapping cell populations in each batch (*i.e.*, partially-overlapping data). Second, these methods can only remove batch effects from the current batches being assessed but cannot manage batch effects from additional, subsequently obtained batches. So each time a new batch is added, it requires an entirely new integration process that again examines the previous batches. This severely limits the capacity to integrate new single-cell sequencing datasets.

As an alternative to the cell similarity-based local methods, scVI[17] applies a conditional variational autoencoder (VAE)[18] framework to model the inherent distribution/structure of the input single-cell data. VAE is a deep generative method that comprises an encoder and a decoder, wherein the encoder projects all high-dimensional input data into a low-dimensional embedding, and the decoder recovers them back to the original data space. The VAE framework can maintain the same global internal data structure between the high- and low-dimensional spaces[19]. However, scVI includes a set of batch-conditioned parameters into its encoder that restrains the encoder from learning a batch-invariant embedding space, limiting its generalizability with new batches.

We previously applied VAE and designed SCALE (Single-Cell ATAC-seq Analysis via Latent feature Extraction) to model and analyze single-cell ATAC-seq data[20]. We found that the VAE framework in SCALE can disentangle cell-type-related and batch-related features in a low-dimensional embedding space. Here, having redesigned the VAE framework, we introduce SCALEX as a method for integration of heterogeneous single-cell data. We demonstrate that SCALEX integration is accurate,

scalable, and computationally efficient for multiple benchmark datasets from scRNA-seq and scATAC-seq studies. As a specific advantage, SCALEX accomplishes data integration through projecting all single-cell data into a generalized cell-embedding space using a batch-free encoder and a batch-specific decoder. Since the encoder is trained to only preserve batch-invariant biological variations, the resulting cell-embedding space is a generalized one, *i.e.*, common to all projected data. SCALEX is therefore able to accurately integrate partially-overlapping datasets without mixing of non-overlapping cell populations. By design, SCALEX runs very efficiently on huge datasets. These two advantages make SCALEX especially useful for the construction and research utilization of large-scale single-cell atlas studies, based on integrating data from heterogeneous sources. New data can be projected to augment an existing atlas, enabling continuous expansion and improvement of an atlas. We demonstrated these functionalities of SCALEX in the construction and analyses of atlases for human, mouse, and COVID-19 PBMCs.

RESULTS

Projecting single-cell data into a generalized cell-embedding space

The central goal of single-cell data integration is to identify and align similar cells across different batches, while retaining true biological variations within and across cell-types. The fundamental concept underlying SCALEX is disentangling batch-related components away from batch-invariant components of single-cell data and projecting the batch-invariant components into a generalized, batch-invariant cell-embedding space. To accomplish this, SCALEX implements a batch-free encoder and a batch-specific decoder in an asymmetric VAE framework[18] (Fig. 1a. Methods). While the batch-free encoder extracts only biological-related latent features (z) from input

single-cell data (x), the batch-specific decoder is responsible for reconstructing the original data from z by incorporating batch information back during data reconstruction.

Supplying batch information to the decoder in data reconstruction allows the encoder to learn a batch-invariant data representation for each individual cell during model training, which, as a whole, defines a generalized low-dimensional cell-embedding space. This learning is also facilitated by random slicing of all input single cells from different batches into mini-batches. Each mini-batch is forced into alignment with the same data distribution under the restriction of KL-divergence in the same cell-embedding space[21]. SCALEX also implements Domain-Specific Batch Normalization (DSBN)[22] (Methods), a multi-branch Batch Normalization[23], in its decoder to support incorporation of batch-specific variations to reconstruct single-cell data.

The design underlying SCALEX renders the encoder to function as a data projector that projects single cells of different batches into a generalized, batch-invariant cell-embedding space. SCALEX thus removes batch-related variations present in single-cell data while preserving batch-invariant biological signals in cell-embedding, making it an enabling tool for integration analyses of diverse single cell datasets, without relying on searching for cell similarities.

SCALEX integration is accurate, scalable, and accommodates diverse data types

We first evaluated the data integration performance of SCALEX on multiple well-curated scRNA-seq datasets, including human *pancreas* (eight batches of five studies)[24-28], *heart* (two batches of one study)[29] and *liver* (two studies)[30,31]; as well as human non-small-cell lung cancer (*NSCLC*, four studies)[32-35] and peripheral blood mononuclear cell (*PBMC*; two batches assayed by two different protocols)[13]. For comparison, we included several other methods in the analyses, including Seurat v3, Harmony, Conos, BBKNN, MNN, Scanorama, and scVI (Methods).

We used Uniform Manifold Approximation and Projection (UMAP)[36] embeddings to visualize the integration performance of all methods (Methods). Note that all of the raw datasets displayed strong batch effects: cell-types that were common in different batches were separately distributed. Overall, SCALEX, Seurat v3, and Harmony achieved the best integration performance for most of the datasets by merging common cell-types across batches while keeping disparate cell-types apart (Fig. S1). MNN and Conos integrated many datasets but left some common cell populations not well aligned. BBKNN, Scanorama, and scVI often had unmerged common cell-types, and sometimes incorrectly mixed distinct cell-types together. For example, in the *PMBC* dataset (Fig. 1b), considering the T cell populations between the two batches, while SCALEX, Seurat v3, Harmony, and MMN integrations were effective, Scanorama showed both a larger misalignment and mixed all cell-types together without maintaining clear boundaries.

We quantified single-cell data integration performance using a silhouette score[37] and a batch entropy mixing score[10] (Methods). Briefly, the silhouette score assesses the separation of biological distinctions, and the batch entropy mixing score evaluates the extent of mixing of cells across batches. Overall, SCALEX outperformed all of the other methods as assessed by the silhouette score, and tied with Seurat and Harmony as the best-performing methods based on the batch entropy mixing score (Fig. 1c). We note that SCALEX obtained a slightly lower batch entropy mixing score, compared to Seurat v3 and Harmony on the *liver* dataset, which contains batch-specific cell-types and thus is a partially-overlapping dataset. However, Seurat v3 and Harmony may have obtained a high batch entropy mixing score because of misaligning different cell-types together. Indeed, by only considering the degree of batch mixing but ignoring cell-type differences, the batch entropy mixing score is not ideally suited for assessing batch mixing for partially-overlapping datasets.

We also tested the scalability and computation efficiency of SCALEX on large-scale datasets by applying it to 1,369,619 cells from the *human fetal atlas* dataset (two data batches, Methods)[38,39]. SCALEX accurately integrated these two batches, showing good alignment of the same cell-types (Fig. S2, Fig. 1d). We then compared the computational efficiency of different methods using down-sampled datasets (of 10 K, 50 K, 250 K, 1 M) from the *human fetal atlas* dataset. SCALEX consumed almost constant runtime and memory that increased only linearly with data size, whereas MNN, Seurat v3, and Conos consumed runtime and memory that increased exponentially, thus did not scale well beyond 250 K cells. Harmony consumed over 400 gigabytes (GB) of memory in analyzing the 1 M dataset, rendering it unsuitable for integration of datasets at this scale (Fig. 1e). Notably, the deep learning framework of SCALEX enables it to run very efficiently on GPU devices, requiring much reduced runtime (took about 10 minutes and 16 GB of memory on the 1 M dataset).

Finally, SCALEX can be used to integrate scATAC-seq data as well as cross-modality data (*e.g.* scRNA-seq and scATAC-seq) (Methods). For example, SCALEX integrated the mouse brain scATAC-seq dataset (two batches assayed by snATAC and 10X)[40] very well, aligning common cell subpopulations and separate distinct ones (Fig. 1f). We also integrated the cross-modality PBMC data between scRNA-seq and scATAC-seq[41,42], and found that SCALEX could correctly integrate the two types of data, and could distinguish rare cells that are specific to scRNA-seq data, including pDC and platelet cells (Fig. 1g). Thus, SCALEX has broad integration capacity across various types of single-cell data.

SCALEX integrates partially-overlapping datasets

Partially-overlapping datasets present a major challenge for single-cell data integration for local cell similarity-based methods[13,14], often leading to over-correction (*i.e.*, mixing of distinct cell-types). As a global integration method that project cells into a

generalized cell-embedding space, SCALEX is expected to be immune to this problem. For example, the *liver* dataset is a partially-overlapping dataset where the hepatocyte population contains multiple subtypes specific to different batches: three subtypes are specific to LIVER_GSE124395, and two other subtypes only appear in LIVER_GSE115469 (Fig. S3). We noticed that SCALEX maintained the five hepatocyte subtypes apart, whereas Seurat v3 mixed all five and Harmony mixed the hepatocyte-SCD and hepatocyte-TAT-AS1 cells (Fig. 2a).

To characterize the performance of SCALEX on partially-overlapping datasets, we constructed test datasets with a range of common cell-types, down-sampled from the six major cell-types in the *pancreas* dataset (Methods). SCALEX integration was accurate for all cases, aligning the same cell-types without over-correction, whereas both Seurat v3 and Harmony frequently mixed the cell-types, particularly for the low-overlapping cases (Fig. 2b, Fig. S4). When there was none common cell-type, both Seurat v3 and Harmony collapsed the six cell-types to three, mixing alpha with gamma cells, beta with delta cells, and acinar with ductal cells in various extent. We repeated the cell-type down-sampling analysis from the 12 cell-types in the *PBMC* dataset as a more complex partial-overlapping example and observed similar results (Fig. S5), demonstrating that SCALEX is robust in retaining informative biological variations for partially-overlapping datasets.

Projection of unseen data into an existing cell-embedding space

The accurate, scalable, and efficient integration performance of SCALEX depends on its encoder's capacity to project cells from various sources into a generalized, batch-invariant cell-embedding space. We speculate that once a cell-embedding space has been constructed after integration of existing data, SCALEX should be able to use the same encoder to project additional (*i.e.*, previously unseen) data onto the same embedding space. To test this hypothesis, we used the *pancreas* dataset. SCALEX

integration removed the strong batch effect in the raw data and aligned the same cell-types together and kept different cell-types were clearly distinguished (Fig. 3a, Fig. S6a). Cell-types were validated by the expression of their canonical markers, including rare cells such as Schwann cells, epsilon cells (Fig. S6b).

We projected three new batches[43-45] for pancreas tissues (Fig. 3b) into this "pancreas cell space" using the same encoder trained on the *pancreas* dataset. After projection, most of the cells in the new batches were accurately aligned to the correct cell-types in the pancreas cell space, enabling their accurate annotation by cell-type label transfer (Fig. 3c, Method). We benchmarked annotation accuracy by calculating the adjusted Rand Index (ARI)[46], the Normalized Mutual Information (NMI)[47], and the F1 score using the cell-type information in the original studies as a gold standard (Methods). The SCALEX annotations achieved the highest accuracy in comparisons with annotations using three other methods (Seurat v3, Conos, and scmap).

Expanding an existing cell space by including new data

The ability to project new single-cell data into a generalized cell-embedding space allows SCALEX to readily extend this cell space. To verify this, we projected two additional melanoma data batches (SKCM_GSE72056, SKCM_GSE123139)[48,49] onto the previously constructed PBMC space. The common cell-types were correctly projected onto the same locations in the PBMC cell space (Fig. 3d). For the tumor and plasma cells only present in the melanoma data batches, SCALEX did not project these cells onto any existing cell populations in the PBMC space; rather, it projected them onto new locations close to similar cells, with the plasma cells projected to a location near B cells, and the tumor cells projected to a location near HSC cells (Fig. 3e).

SCALEX projection enables *post hoc* annotation of unknown cell-types in the existing cell space using new data. We noted a group of cells previously uncharacterized in the *pancreas* dataset (Fig. 3a). We found that these cells displayed

high expression levels for known epithelial genes (Methods). We therefore assembled a collection of epithelial cells from the *bronchial epithelium* dataset[50]. We then projected these epithelial cells onto the pancreas cell space and found that a group of antigen-presenting airway epithelial (SLC16A7+ epithelial) cells were projected onto the same location of the uncharacterized cells (Fig. 3f). This, together with the observation that both cell populations showed similar marker gene expression (Fig. 3g), indicates that these uncharacterized cells are also SLC16A7+ epithelial cells. SCALEX thus enables discovery science in cell biology by supporting exploratory analysis with large numbers of diverse datasets.

SCALEX supports construction of expandable single-cell atlases

The ability to combine partially-overlapping data onto a generalized cell-embedding space makes SCALEX a powerful tool to construct a single-cell atlas from a collection of diverse and large datasets. We applied SCALEX integration to two large and complex datasets—the *mouse atlas* dataset (comprising multiple organs from two studies assayed by 10X, Smart-seq2, and Microwell-seq[6,51]) (Fig. 4a) and the *human atlas* dataset (comprising multiple organs from two studies assayed by 10X and Microwell-seq[39,52]).

Despite the strong batch effects in the raw data, SCALEX integrated the three batches of the *mouse atlas* dataset into a unified cell-embedding space (Fig. 4b,c, Fig. S7a). Common cell-types (including both B, T, and endothelial cells in all tissues and proximal tubule, urothelial, and hepatocytic cells in certain tissues) were well-aligned together at the same position in the cell space. Non-overlapping cell-types (such as sperm, Leydig, and small intestine cells from the Microwell-seq data, keratinocyte stem cells and large intestine cells in the Smart-seq2 data, and oligodendrocytes in the Smart-seq2 and Microwell-seq data) were located separately in the space, indicating that biological variations were preserved well (Fig S7b).

Importantly, atlases generated with SCALEX can be used and further expanded by projecting new single-cell data to support comparative studies of cells both in the original atlas and in the new data. Illustrating this, we projected two additional data batches of aged mouse tissues from *Tabula Muris Senis* (Smart-seq2 and 10X) [53] and two single tissue datasets (lung and kidney)[54] onto the SCALEX mouse atlas space. We found that the same cell-types in the new data batches were correctly projected onto the same locations on the cell-embedding space of the initial mouse atlas (Fig. 4d), which was also confirmed by the accurate cell-type annotations for the new data by label transfer from the corresponding cell-types in the initial atlas (Fig. 4e. Methods). On one way, this mouse atlas then can be used to accurately identify/characterize the cells in the new data based on their projected locations in the cell space; and on the other way, projection of new data enables ongoing (and informative) expansion of an existing atlas.

Following the same strategy, we also constructed a human atlas by SCALEX integration of multiple tissues from two studies (GSE134255, GSE159929) (Fig. S8a,b). SCALEX, effectively eliminated the batch effects in the original data and integrated the two datasets in a unified cell-embedding space (Fig. S8c,d). Again, we were able to correctly project two additional human skin datasets (GSE130973, GSE147424)[55,56] onto the human atlas cell-embedding space (Fig. S8e), and again accurately annotated these projected skin cells (Fig. S8f. Methods). These results illustrate that: i) SCALEX enables researchers to evaluate their project-specific single cell datasets by leveraging existing information in large-scale (and ostensibly well annotated) cell atlases; and ii) it also enables atlas creators to informatively integrate new datasets and attendant biological insights from many research programs.

An integrative SCALEX COVID-19 PBMC atlas

Many single-cell studies have been conducted to analyze COVID-19 patient immune responses[57-64]. However, these studies often suffer from small sample size and/or

limited sampling of various disease states[58,64]. For a comprehensive study, we collected data from multiple COVID-19 PBMC studies, involving 860,746 single cells, and 10 batches from 9 studies[57-63] (Fig. 5a, Fig. S9a), and used SCALEX to generate a COVID-19 PBMC atlas, identifying 22 cell-types, each of which were supported by canonical marker gene expression (Fig. 5b,c, Fig. S9b,c. Methods). Cells across different studies were integrated accurately with the same cell-types aligned together, confirming integration performance of SCALEX (Fig. 5c, Fig. S9d).

We observed that some cell subpopulations were differentially associated with patient status (Fig 5d). A subpopulation of CD14 monocytes (CD14-ISG15-Mono), specifically associated with COIVD-19 patients, was characterized by its high expression of Type I interferon-stimulated genes (ISGs) and genes associated with immune-response-related GO terms (Fig 5e,f). The frequency of CD14-ISG15-Mono cells increased significantly from healthy donors to mild/moderate and severe patients (Fig. 6g, Fig. S9e. Methods). Within the COVID-19 patients, we observed a significant decrease in ISG gene expression in CD14-ISG15-Mono cells between the mild/moderate and severe cases, indicating apparently dysfunctional anti-viral immune response in severe COVID-19 patients (Fig. 5e). Specifically enriched in severe verse mild/moderate patients, a neutrophil subpopulation (NCF1-Immature_Neutrophil) lacked expression of the genes responsible for neutrophil activation but showed elevated expression of genes associated with viral-process-related GO terms (Fig. S10a,b). Also enriched in severe patients, a plasma cell subpopulation (MZB1-Plasma) cells displayed decreased expression for antibody production and were enriched for GO terms of immune and inflammatory responses (Fig. S10c,d). Thus, the SCALEX COVID-19 PBMC atlas, generated by integrating a highly diverse collection of single-cell data from individual studies, identified multiple immune cells-types showing dysregulations during COVID-19 disease progression. Note that these trends could not have been detected in the small-scale, individual studies that served as the basis for our SCALEX COVID-19 PBMC atlas.

Comparative analysis of the SCALEX COVID-19 PBMC atlas and the SC4 consortium study

Recently, a large-scale effort of the Single Cell Consortium for COVID-19 in China (SC4) has generated a single-cell atlas that contains over 1 million cells (including PBMCs and other tissues) from 171 COVID-19 patients and 25 healthy controls[65] (Fig. S11a). We projected the consortium dataset into the cell-embedding space of the SCALEX COVID-19 PBMC atlas, and found that the cell-types of two atlases were well-aligned in the embedding space (Fig. 5h,i, Fig. S11b,c).

Our analysis, based on the SCALEX COVID-19 PBMC atlas, yielded findings consistent with two conclusions from the SC4 study[65]. First, in both analyses diverse immune subpopulations displayed differential associations with COVID-19 severity. The proportions of CD14 monocytes, megakaryocytes, plasma cells, and pro T cells were elevated with increasing disease severity, while the proportion of pDC and mDC cells decreased (Fig. 5g). Second, we confirmed that the megakaryocytes and monocyte populations are associated with cytokine storms triggered by SARS-Cov2 infection and are further elevated in severe patients[66], based on calculating the same cytokine score and inflammatory score (defined in the SC4 study) for the cells of our SCALEX COVID-19 PBMC atlas (Fig. 5j. Methods).

Integration of the SC4 data further substantially improved both the scope and resolution of the SCALEX COVID-19 PBMC atlas. First, this data added macrophages and epithelial cells to the cell space, enabling investigation of their potential involvement in COVID-19. The integration also supported more precise characterization of specific cell subpopulations. For example, the megakaryocyte population, not distinguished in either single atlas, could be divided into two subpopulations in the combined atlas (Fig. 5h). An exploratory functional analysis of the differentially expressed genes in these two newly delineated megakaryocyte

subpopulations (TUBA8-Mega and IGKC-Mega, Fig. S11d,e) revealed enrichment for the GO terms “humoral immune response” for IGKC-Mega cells yet enrichment for “negative regulation of platelet activation” for TUBA8-Mega cells (Fig. 5k). These results illustrate how the continuously expandable single-cell atlases generated using SCALEX capitalize on existing large-scale data resources and also facilitate discovery of biological and biomedical insights.

DISCUSSION

SCALEX provides a VAE framework for integration of heterogeneous single-cell data by disentangling batch-invariant components from batch-related variations and projecting the batch-invariant components into a generalized, low-dimensional cell-embedding space. By design, SCALEX models the inherent batch-invariant patterns of single-cell data, distinguishing it from previously reported integration methods based on cell similarities. SCALEX does not rely on the identification of common cell-types across batches, and therefore avoids the problem of cell-type over-correction, a severe problem for partially-overlapping datasets. SCALEX thus also overcomes issues of computational complexity in cell similarity-based methods; that is, the computational time required to identify similar cells may increase exponentially as the cell number increases.

These two features make SCALEX particularly useful for construction and integrative analysis of large-scale single-cell atlases based on very heterogenous data (*i.e.*, datasets acquired by different labs and using different single-cell analysis platforms). Our construction of human, mouse, and COVID-19 patient single-cell atlases—which aligned well with previously reported atlases generated from coordinated large-scale consortium efforts—demonstrates the particular ability of SCALEX to producing large-scale atlases from extant small-scale datasets. SCALEX achieves data integration by projecting all single cells into a generalized cell-embedding

space using a universal data projector (*i.e.*, the encoder). This data projector only needs to be trained once, and then can be used without retraining to continuously integrate new incoming data into an existing single-cell atlas. This continuous growth ability makes a SCALEX atlas an elastic resource, allowing the integration of many single-cell studies to support ongoing, very large-scale research programs throughout the life sciences and biomedicine.

While the number of single-cell studies is increasing enormously each year, best practices for experimental design and sample processing are not established, and there is no obviously dominant data-acquisition platform. SCALEX's ability to informatively combine data from heterogenous studies and platforms makes it particularly suitable for the current era of single-cell biological research. Finally, the ability to conduct exploratory analysis within a generalized cell space supports that SCALEX should be particularly useful for large-scale integrative (*e.g.*, pan-cancer) studies. We speculate that use of SCALEX to project single-cell datasets (including for example scATAC-seq and scRNA-seq) from highly diverse cancer types to construct a pan-cancer single-cell atlas may lead to the discovery of previously unknown cell types that are common to divergent carcinomas and that function in pathogenesis, malignant progression, and/or metastasis.

REFERENCES

1 Kolodziejczyk, A. A., Kim, J. K., Svensson, V., Marioni, J. C. & Teichmann, S. A. The technology and biology of single-cell RNA sequencing. *Mol Cell* **58**, 610-620, doi:10.1016/j.molcel.2015.04.005 (2015).

2 Papalexi, E. & Satija, R. Single-cell RNA sequencing to explore immune cell heterogeneity. *Nat Rev Immunol* **18**, 35-45, doi:10.1038/nri.2017.76 (2018).

3 Tanay, A. & Regev, A. Scaling single-cell genomics from phenomenology to mechanism. *Nature* **541**, 331-338, doi:10.1038/nature21350 (2017).

4 Satpathy, A. T. *et al.* Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. *Nature Biotechnology* **37**, 925-936, doi:10.1038/s41587-019-0206-z (2019).

5 Regev, A. *et al.* The Human Cell Atlas. *Elife* **6**, doi:10.7554/eLife.27041 (2017).
6 Tabula Muris, C. *et al.* Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* **562**, 367-372, doi:10.1038/s41586-018-0590-4 (2018).
7 Lahnemann, D. *et al.* Eleven grand challenges in single-cell data science. *Genome Biol* **21**, 31, doi:10.1186/s13059-020-1926-6 (2020).
8 Leek, J. T. *et al.* Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet* **11**, 733-739, doi:10.1038/nrg2825 (2010).
9 Hicks, S. C., Townes, F. W., Teng, M. & Irizarry, R. A. Missing data and technical variability in single-cell RNA-sequencing experiments. *Biostatistics* **19**, 562-578, doi:10.1093/biostatistics/kxx053 (2018).
10 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, doi:10.1038/nbt.4091 (2018).
11 Hie, B., Bryson, B. & Berger, B. Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. *Nat Biotechnol*, doi:10.1038/s41587-019-0113-3 (2019).
12 Polanski, K. *et al.* BBKNN: fast batch alignment of single cell transcriptomes. *Bioinformatics* **36**, 964-965, doi:10.1093/bioinformatics/btz625 (2020).
13 Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol* **36**, 411-420, doi:10.1038/nbt.4096 (2018).
14 Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902 e1821, doi:10.1016/j.cell.2019.05.031 (2019).
15 Barkas, N. *et al.* Joint analysis of heterogeneous single-cell RNA-seq dataset collections. *Nat Methods* **16**, 695-698, doi:10.1038/s41592-019-0466-z (2019).
16 Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat Methods*, doi:10.1038/s41592-019-0619-0 (2019).
17 Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nat Methods* **15**, 1053-1058, doi:10.1038/s41592-018-0229-2 (2018).
18 Kingma, D. P. & Welling, M. Auto-Encoding Variational Bayes. *arXiv:1312.6114* (2013).
19 Rezende, D. J., Mohamed, S. & Wierstra, D. Stochastic Backpropagation and Approximate Inference in Deep Generative Models. doi:abs/ (2014).
20 Xiong, L. *et al.* SCALE method for single-cell ATAC-seq analysis via latent feature extraction. *Nat Commun* **10**, 4576, doi:10.1038/s41467-019-12630-7 (2019).
21 Kullback, S. & Leibler, R. A. On Information and Sufficiency. *Ann. Math. Statist.* **22**, 79-86, doi:10.1214/aoms/1177729694 (1951).

22 Chang, W.-G., You, T., Seo, S., Kwak, S. & Han, B. Domain-Specific Batch Normalization for Unsupervised Domain Adaptation. *arXiv:1906.03950* (2019).

23 Ioffe, S. & Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *arXiv:1502.03167* (2015).

24 Lawlor, N. *et al.* Single-cell transcriptomes identify human islet cell signatures and reveal cell-type-specific expression changes in type 2 diabetes. *Genome Res* **27**, 208-222, doi:10.1101/gr.212720.116 (2017).

25 Segerstolpe, A. *et al.* Single-Cell Transcriptome Profiling of Human Pancreatic Islets in Health and Type 2 Diabetes. *Cell Metab* **24**, 593-607, doi:10.1016/j.cmet.2016.08.020 (2016).

26 Muraro, M. J. *et al.* A Single-Cell Transcriptome Atlas of the Human Pancreas. *Cell Syst* **3**, 385-394 e383, doi:10.1016/j.cels.2016.09.002 (2016).

27 Grun, D. *et al.* De Novo Prediction of Stem Cell Identity using Single-Cell Transcriptome Data. *Cell Stem Cell* **19**, 266-277, doi:10.1016/j.stem.2016.05.010 (2016).

28 Baron, M. *et al.* A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. *Cell Syst* **3**, 346-360 e344, doi:10.1016/j.cels.2016.08.011 (2016).

29 Litvinukova, M. *et al.* Cells of the adult human heart. *Nature* **588**, 466-472, doi:10.1038/s41586-020-2797-4 (2020).

30 Aizarani, N. *et al.* A human liver cell atlas reveals heterogeneity and epithelial progenitors. *Nature* **572**, 199-204, doi:10.1038/s41586-019-1373-2 (2019).

31 MacParland, S. A. *et al.* Single cell RNA sequencing of human liver reveals distinct intrahepatic macrophage populations. *Nat Commun* **9**, 4383, doi:10.1038/s41467-018-06318-7 (2018).

32 Lambrechts, D. *et al.* Phenotype molding of stromal cells in the lung tumor microenvironment. *Nat Med* **24**, 1277-1289, doi:10.1038/s41591-018-0096-5 (2018).

33 Song, Q. *et al.* Dissecting intratumoral myeloid cell plasticity by single cell RNA-seq. *Cancer Med* **8**, 3072-3085, doi:10.1002/cam4.2113 (2019).

34 Zilionis, R. *et al.* Single-Cell Transcriptomics of Human and Mouse Lung Cancers Reveals Conserved Myeloid Populations across Individuals and Species. *Immunity* **50**, 1317-1334 e1310, doi:10.1016/j.immuni.2019.03.009 (2019).

35 Kim, N. *et al.* Single-cell RNA sequencing demonstrates the molecular and cellular reprogramming of metastatic lung adenocarcinoma. *Nat Commun* **11**, 2285, doi:10.1038/s41467-020-16164-1 (2020).

36 McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv:1802.03426* (2018).

37 Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* **20**, 53-65, doi:https://doi.org/10.1016/0377-0427(87)90125-7 (1987).
38 Cao, J. *et al.* A human cell atlas of fetal gene expression. *Science* **370**, doi:10.1126/science.aba7721 (2020).
39 Han, X. *et al.* Construction of a human cell landscape at single-cell level. *Nature* **581**, 303-309, doi:10.1038/s41586-020-2157-4 (2020).
40 Fang, R. *et al.* Comprehensive analysis of single cell ATAC-seq data with SnapATAC. *Nat Commun* **12**, 1337, doi:10.1038/s41467-021-21583-9 (2021).
41 Genomics, X. 10k Peripheral blood mononuclear cells (PBMCs) from a healthy donor, Single Cell ATAC Dataset by Cell Ranger 1.0.1. (2018).
42 Genomics, X. 10k PBMCs from a Healthy Donor (v3 chemistry), Single Cell Gene Expression Dataset by Cell Ranger 3.0.0. (2018).
43 Wang, Y. J. *et al.* Single-Cell Transcriptomics of the Human Endocrine Pancreas. *Diabetes* **65**, 3028-3038, doi:10.2337/db16-0405 (2016).
44 Enge, M. *et al.* Single-Cell Analysis of Human Pancreas Reveals Transcriptional Signatures of Aging and Somatic Mutation Patterns. *Cell* **171**, 321-330 e314, doi:10.1016/j.cell.2017.09.004 (2017).
45 Xin, Y. *et al.* Pseudotime Ordering of Single Human beta-Cells Reveals States of Insulin Production and Unfolded Protein Response. *Diabetes* **67**, 1783-1794, doi:10.2337/db18-0365 (2018).
46 Hubert, L. & Arabie, P. Comparing partitions. *Journal of Classification* **2**, 193-218, doi:10.1007/BF01908075 (1985).
47 Amelio, A. & Pizzuti, C. in *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015* 1584–1585 (Association for Computing Machinery, Paris, France, 2015).
48 Tirosh, I. *et al.* Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352**, 189-196, doi:10.1126/science.aad0501 (2016).
49 Li, H. *et al.* Dysfunctional CD8 T Cells Form a Proliferative, Dynamically Regulated Compartment within Human Melanoma. *Cell* **176**, 775-789.e718, doi:https://doi.org/10.1016/j.cell.2018.11.043 (2019).
50 Plasschaert, L. W. *et al.* A single-cell atlas of the airway epithelium reveals the CFTR-rich pulmonary ionocyte. *Nature* **560**, 377-381, doi:10.1038/s41586-018-0394-6 (2018).
51 Han, X. *et al.* Mapping the Mouse Cell Atlas by Microwell-Seq. *Cell* **172**, 1091-1107 e1017, doi:10.1016/j.cell.2018.02.001 (2018).
52 He, S. *et al.* Single-cell transcriptome profiling of an adult human cell atlas of 15 major organs. *Genome Biol* **21**, 294, doi:10.1186/s13059-020-02210-0 (2020).

53 Tabula Muris, C. A single-cell transcriptomic atlas characterizes ageing tissues in the mouse. *Nature* **583**, 590-595, doi:10.1038/s41586-020-2496-1 (2020).
54 Kimmel, J. C. *et al.* Murine single-cell RNA-seq reveals cell-identity- and tissue-specific trajectories of aging. *Genome Res* **29**, 2088-2103, doi:10.1101/gr.253880.119 (2019).
55 Sole-Boldo, L. *et al.* Single-cell transcriptomes of the human skin reveal age-related loss of fibroblast priming. *Commun Biol* **3**, 188, doi:10.1038/s42003-020-0922-4 (2020).
56 He, H. *et al.* Single-cell transcriptome analysis of human skin identifies novel fibroblast subpopulation and enrichment of immune subsets in atopic dermatitis. *J Allergy Clin Immunol* **145**, 1615-1628, doi:10.1016/j.jaci.2020.01.042 (2020).
57 Schulte-Schrepping, J. *et al.* Severe COVID-19 Is Marked by a Dysregulated Myeloid Cell Compartment. *Cell* **182**, 1419-1440.e1423, doi:10.1016/j.cell.2020.08.001 (2020).
58 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**, doi:10.1126/sciimmunol.abd1554 (2020).
59 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, doi:10.1038/s41591-020-0944-y (2020).
60 Guo, C. *et al.* Single-cell analysis of two severe COVID-19 patients reveals a monocyte-associated and tocilizumab-responding cytokine storm. *Nat Commun* **11**, 3924, doi:10.1038/s41467-020-17834-w (2020).
61 Yao, C. *et al.* Cell-Type-Specific Immune Dysregulation in Severely Ill COVID-19 Patients. *Cell Rep* **34**, 108590, doi:10.1016/j.celrep.2020.108590 (2021).
62 Zhang, J. Y. *et al.* Single-cell landscape of immunological responses in patients with COVID-19. *Nat Immunol* **21**, 1107-1118, doi:10.1038/s41590-020-0762-x (2020).
63 Ballestar, E. *et al.* Single cell profiling of COVID-19 patients: an international data resource from multiple tissues. *medRxiv*, 2020.2011.2020.20227355, doi:10.1101/2020.11.20.20227355 (2020).
64 Bernardes, J. P. *et al.* Longitudinal Multi-omics Analyses Identify Responses of Megakaryocytes, Erythroid Cells, and Plasmablasts as Hallmarks of Severe COVID-19. *Immunity* **53**, 1296-1314 e1299, doi:10.1016/j.immuni.2020.11.017 (2020).
65 Ren, X. *et al.* COVID-19 immune features revealed by a large-scale single-cell transcriptome atlas. *Cell*, doi:10.1016/j.cell.2021.01.053 (2021).

66 Chen, G. *et al.* Clinical and immunological features of severe and moderate coronavirus disease 2019. *J Clin Invest* **130**, 2620-2629, doi:10.1172/JCI137244 (2020).

67 Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980* (2014).

68 Danese, A., Richter, M. L., Fischer, D. S., Theis, F. J. & Colomé-Tatché, M. EpiScanpy: integrated single-cell epigenomic analysis. *bioRxiv*, doi:10.1101/648097 (2019).

69 Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol* **19**, 15, doi:10.1186/s13059-017-1382-0 (2018).

70 Stuart, T., Srivastava, A., Lareau, C. & Satija, R. Multimodal single-cell chromatin analysis with Signac. *bioRxiv*, doi:10.1101/2020.11.09.373613 (2020).

71 Abdi, H. & Williams, L. J. Principal component analysis. *WIREs Comput. Stat.* **2**, 433-459, doi:10.1002/wics.101 (2010).

72 Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports* **9**, 5233, doi:10.1038/s41598-019-41695-z (2019).

Methods

Overview of the SCALEX model. SCALE applies a variational autoencoder (VAE) to project the different batches of datasets into the same batch-invariant low-dimensional embeddings by learning a batch-free encoder and a batch-specific decoder simultaneously. Since the encoder and decoder are coupled to learn a batch-free encoder, a batch label is only exposed to the decoder within the domain-specific batch normalization, thus the decoder captures the batch information while the encoder learns the domain-invariant features. SCALEX takes the input expression profile **x** across all the batches as a whole mixture distribution to learn a global data structure. In the training process, SCALEX randomly samples data from all batches within each mini-batch and trains on them together. Once trained, the encoder of SCALEX is generalized to any batches and serves as a universal function for mapping different batches of datasets into the same batch-invariant space.

Training SCALEX is to maximize the log-likelihood of the observed single-cell sequencing data x:

$$\begin{aligned} logp(x) &= log \int_z p(x,z)dz \\ &\geq E_{q(z|x)}\left[log\frac{p(x,z)}{q(z|x)}\right] \\ &= \mathcal{L}_{ELBO}(x) \end{aligned}$$

Then the loss function is transformed into the evidence lower bound (ELBO). While the ELBO can be further decomposed into two terms:

$$\mathcal{L}_{ELBO}(x) = E_{q(z|x)}[logp(x|z)] - D_{KL}(q(z|x)||p(z))]$$

The first term is the reconstruction term, which minimizes the distance between the generated output data and the original input data. The second term is the regularization term, which minimizes the Kullback-Leibeler divergence between posterior

distribution and prior distribution of latent variable z. To enable a more flexible alignment under the latent space, we adjusted the coefficient of the second term to 0.5, thus, the final loss function is:

$$\mathcal{L}_{ELBO}(x) = E_{q(Z|X)}[logp(x|z)] - 0.5 * D_{KL}(q(z|x)||p(z))]$$

The overall network architecture of SCALEX consists of an encoder and a decoder. The encoder is a two-layer neural network (fc[1024]-BN-ReLU-fc[10]), and the decoder has only one layer (no hidden layer), directly connecting latent feature z to the output x (fc-DSBN-Sigmoid) with domain-specific batch normalization and a Sigmoid activation function. We use Adam[67] optimizer with a 5e-4 weight decay to optimize the model. The mini-batch size for training input is 64. The maximum number of training iterations is 30,000 and an early stopping is triggered when there has been no improvement for 10 epochs.

Domain-specific batch normalization (DSBN). Batch normalization (BN)[23] is a widely used training technique in deep neural networks to reduce internal covariate shifting. A BN layer whitens activations within a mini-batch of samples followed by scaling and shifting with learned affine parameters γ and β. For a mini-batch of samples: $\mathcal{B} = \{x_{1\ldots m}\}$;

$$\mu_{\mathcal{B}} = \frac{1}{m}\sum_{i=1}^{m} x_i$$

$$\sigma_{\mathcal{B}}^2 = \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_{\mathcal{B}})^2$$

$$\hat{x}_i = \frac{x_i - \mu_{\mathcal{B}}}{\sqrt{\sigma_{\mathcal{B}}^2 + \epsilon}}$$

$$y_i = \gamma\hat{x}_i + \beta \equiv \mathrm{BN}_{\gamma,\beta}(x_i)$$

Where $\mu_{\mathcal{B}}$ is the mini-batch mean, $\sigma_{\mathcal{B}}^2$ is the mini-batch variance, $\hat{x}_i$ is the normalized output by $\mu_{\mathcal{B}}$ and $\sigma_{\mathcal{B}}^2$, y_i is the BN output by scaling and shifting $\hat{x}_i$ with parameters γ and β, and ϵ is a constant added to the mini-batch variance for numerical stability.

Domain specific batch normalization (DSBN)[22] is a combination of multiple sets of BN specific to each domain. DSBN learns domain-specific affine parameters γ_d and β_d for each domain, d is the domain label; here, domain represents different batches. In the neural network, DSBN serves like multi-channel BN and switches to the corresponding BN given the domain label d. DSBN could capture the domain-specific information by estimating mini-batch statistics by learning affine parameters for each domain separately, thus enabling the network to learn the domain-invariant features.

Preprocessing for scRNA-seq. We downloaded gene expression matrices and preprocessed them using the following procedure: i). Cells with fewer than 600 genes and genes present in fewer than 3 cells were filtered out. ii). Total counts of each cell were normalized to 10,000. iii). Values of each gene were subjected to log transformation with an offset of 1. iv). The top 2,000 highly variable genes were identified. v). Values of each gene were normalized to the range of 0-1 within each batch by the *MaxAbsScaler* function in the *scikit-learn* package in Python. The processed matrix was used as input for the SCALEX model for downstream differential gene expression analysis.

For the *human fetal atlas* dataset, we collected two batches (batch GSE156793, which contains 4,062,980 cells by sciRNA-seq3, and batch GSE124355, which contains 254,266 cells by Microwell-seq). We then selected the cells from the common tissues (1,369,619 cells) for integration and computational efficiency benchmarking (down-sampled from different data sizes including 10K, 50K, 250K, and 1M).

Preprocessing for scATAC-seq. We downloaded open chromatin profile matrices (peaks or bins), merged them by peaks (or bins), and processed them using the following procedure: i). The combined matrix was binarized and filter bins with fewer than 3 cells. ii). The top 30,000 most variable peaks (or bins) were selected using the *select_var_feature* function in the *EpiScanpy*[68] package. iii). Total counts of each cell were normalized to the median of the total counts of all cells by using the *normalize_total* function, with parameters *target_sum*="None" in the *Scanpy*[69] package. iv). Values of each peak (or bin) were normalized to the range of 0-1 within each batch by the *MaxAbsScaler* function in the *scikit-learn* package in Python. The processed matrix was used as input for the SCALEX model.

Preprocessing for cross-modality data (scRNA-seq and scATAC-seq). We first created a gene activity matrix by the *GeneActivity* function in the *Signac*[70] R package to quantify the activity of each gene from scATAC-seq data. We then combined gene activity score matrix with scRNA-seq data matrix as two individual "batches" for integration. The subsequent preprocessing followed the same preprocessing used for the scRNA-seq data (above).

Visualization. UMAP algorithm[36] was used for visualization. We applied the *neighbors* function from the Python package *Scanpy* with the parameters *n_neighbors*=30 and *metric*="Euclidean" for computing the neighbor graph, followed by *umap* function with *min_dist*=0.1 to visualize cells in a two-dimensional space.

Silhouette score. We used the silhouette score to assess the separation of biological populations with the function *silhouette_score* in the *scikit-learn* package in Python. The silhouette score was computed by combining the average intra-cluster distance (a) and the average nearest-cluster (b) for each cell.

$$silhouette\ score = (b - a)/\max(a, b)$$

Here, we took UMAP embeddings as input to calculate silhouette score.

Batch entropy mixing score. Batch entropy mixing score (adapted from "entropy of batch mixing"[10]) was used to access the regional mixing of cells from different batches, with a high score suggesting that cells from different batches are well mixed together.

The batch entropy mixing score was computed as follows:

(1) Calculated the proportion Pi of cell numbers in each batch to the total cell numbers.
(2) Randomly chose 100 cells from all batches.
(3) Calculated the 100 nearest neighbors for each randomly chosen cell.
(4) The regional mixing entropies for each cell were defined as:

$$pi' = pi/Pi$$

$$\mathrm{E} = \sum_{i=0}^{n} pi' log(pi')$$

where *pi* is the proportion of cells from batch *i* in a given region, such that $\sum_{i=0}^{n} pi = 1$, *pi*' is a correction item to eliminate the deviation caused by the different cell numbers in different batches. The total mixing entropy was then calculated as the sum of the regional mixing entropies.

(5) Repeated (2)-(4) for 100 iterations with different randomly chosen cells and calculated the average, E, as the final batch entropy mixing score.

Comparison with other integration methods. We compared SCALEX to seven other batch effect removal methods (see below for specific details of each method). For each dataset as input for all methods, we performed the same filtration, followed by method-specific normalization, batch correction and visualization. Note that for visual comparison, we also included the embeddings of the raw input data, wherein we performed dimensionality reduction by Principal Component Analysis (PCA)[71]

followed by UMAP visualization to see the batch effects. No correction function was used. All parameters were kept as default values.

Scanorama (v1.6). We performed the preprocessing pipelines as stated above (as the same below), and used the *Scanpy* and *scanorama* Python packages for integration. For the *highly_variable_genes* function, we set flavor="seurat", batch_key="batch", and n_top_genes=2,000. After extracting highly variable genes, we divided the datasets according to the batch labels and formed a new list of datasets as the input for the *correct_scanpy* function. The integration matrix was kept for downstream analysis. All other parameters were kept their default values.

BBKNN (v1.3.12). We used *Scanpy* and *bbknn* Python packages and followed the suggested pipelines for integration. For the *highly_variable_genes* function, we set flavor="seurat", batch_key="batch", and n_top_genes=2,000. After selecting cell neighbors at the low-dimensional space from the PCA analysis, we performed the *bbknn* function with neighbors_within_batch=5, n_pcs=20, and trim=0. All other parameters were default.

scVI (v0.6.5): We used the *scvi* Python package and followed the suggested pipelines. Batch information was added to the VAE model by setting n_batch.

Seurat v3 (v3.2.3): We used the *Seurat* R package and followed the standard integration workflow. We normalized different batches of a dataset separately. For the *FindVariableFeatures* function, we set selection.method = "vst" and nfeatures = 2,000 to select 2,000 highly variable genes for each batch of a dataset. For the *FindIntegrationAnchors* function, we set k.filter=100. All other parameters were kept at default values. If the number of input cells in a dataset exceeded 50,000, we employed the reciprocal PCA and reference-based integration to improve computational efficiency.

Harmony (v1.0): We used the *harmony* R package. We created a Seurat object with all cells and performed the standard workflow. After PCA, we used the *RunHarmony* function for integration. All parameters were default.

Conos (v1.3.1): We used the *Conos* R package. For each batch of dataset, we used the *basicSeuratProc* and *RunTSNE* functions for preprocessing. After that, we built a joint graph using the *buildGraph* function with k=30 and k.self=5. All other parameters were default.

MNN (FastMNN, v0.3.0): We used the *SeuratWrappers* R package. We created a Seurat object with all cells and performed the standard workflow. Then we used the *RunFastMNN* function with default parameters for integration.

Cell-type annotation by clustering. This type of annotation was used for *de novo* annotation of a single-cell dataset. We used a Leiden clustering[72] method for cell clustering (specifically employing the *leiden* function from the Python package *Scanpy* with default parameters). Then for each cluster, we annotate its cell-type based on: i) cell-type annotations of each cell in the original study, if available, or ii) expression levels of canonical marker genes in each cell. A majority vote strategy was used when needed. Similar to Ren et al. 2021, we also employed a hierarchical annotation strategy, *i.e.*, we first clustered all cells in a dataset into several major clusters, then for some big clusters, we further clustered them into minor clusters respectively.

Cell-type annotation by label transfer. This type of annotation was used for annotation of a new single-cell data batch using the annotations in a large single-cell dataset as a reference, or for *post hoc* annotations of unknown cell population(s) in a large dataset using new batches of data of known cell-types. Both scenarios require "single cell projection" (see details below).

The basic idea of cell-type annotation by label transfer is based-on that the same cell-types will occupy the same locations in the generalized SCALEX cell-embedding space, thus cell-type annotation in one data batch can be transferred to another data batch, for the cells positioned at the same locations. Technically, we used the *KNeighborsClassifier* function from the *scikit-learn* package to train a prediction model, using the representations (in the low-dimensional cell-embedding space) of the single-cell data with known cell-type labels as input. We then used this model to make cell-type predictions for cells without annotations using their representations (in the low-dimensional cell-embedding space) as input.

Single cell projection. We defined single cell projection as the operation to convert high-dimensional single-cell data (*e.g.*, gene expression profiles in scRNA-seq or open chromatin profiles in scATAC-seq) to low-dimensional representations in the generalized SCALEX cell-embedding space using the trained encoder.

Similarity matrix and confusion matrix. We used similarity matrix to evaluate the congruence of two different batches for the same cell-types in the generalized cell-embedding space. Technically, we merged all cells with the same cell-type label and calculated an average representation (in the low-dimensional cell-embedding space) for the cell-type. This was repeated for all cell-types. We then calculated the similarity matrix $S=[S_{ij}]$ for the cell-type similarities of the two batches, where S_{ij} is the Pearson correlation coefficient between the average representation of cell-type *i* in data_batch_1 and the average representation of cell-type *j* in data_batch_2.

We used the confusion matrix to evaluate the accuracy of cell-type annotations (prediction) when a gold-standard annotation is available, which is typical for "cell-type annotation by label transfer" (see above). In cell-type annotation by label transfer, we predict the cell-types for a single-cell data_batch_1, using the annotations in another data_batch_2. When data_batch_1 was already annotated with cell-types, we can

calculate the confusion matrix $C=[C_{ij}]$ to compare the cell-type predictions with the existing cell-type annotations, where C_{ij} equals the percentage of cells known to be in cell-type *i* and predicted to be in cell-type *j*.

Adjusted Rand Index. The Rand Index (RI) computes a similarity score between two clustering assignments by considering matched and unmatched assignment pairs, independent of the number of clusters. The Adjusted Rand Index (ARI) score is calculated by "adjust for chance" with RI as follows:

$$ARI = \frac{RI - Expected\_RI}{\max(RI) - Expected\_RI}$$

If given the contingency table, then ARI can also be represented by:

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right]/\binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right]/\binom{n}{2}}$$

The ARI score is 0 for random prediction and 1 for perfectly matching.

Normalized mutual information.

$$NMI = \frac{I(P;T)}{\sqrt{\mathrm{H(P)H(T)}}}$$

Where P and T are categorical distributions for the predicted and real clustering, *I* is the mutual entropy, and H is the Shannon entropy.

F1 score.

$$score = 2 * (\mathrm{precision} * \mathrm{recall})/(\mathrm{precision} + \mathrm{recall})$$

Generation of partially-overlapping datasets. To simulate partially-overlapping datasets from the *pancreas* dataset, we used the pancreas_celseq2 and pancreas_smartseq2 data batches, and worked with only six cell-types (alpha, beta, ductal, acinar, delta, gamma). For each simulated partially-overlapping dataset, we randomly selected three to six cell-types from each batch, and counted the number of the common cell-types, which was used as the indicator for the overlapping level

(whole integers, 0 to 6). We required the union of cell-types in the newly simulated partially-overlapping dataset to cover all six cell-types.

For the *PBMC* dataset, we used both of the two data batches and worked with twelve cell-types (B, CD4 T, CD4 naive T, CD8 T, CD8 naive T, DC, HSC, Megakaryocyte, NK, monocyte-CD14, monocyte-FCGR3A, pDC). We used the same down-sampling strategy as for the *pancreas* dataset (above).

Analysis of changes in cell-type frequency across multiple conditions. To identify differences in cell-type frequency among the scRNA-seq data from the mild/moderate, severe, convalescent COVID-19 patients, as well as the healthy and influenza patient controls, we applied a Dirichlet-multinomial regression model. This model accounts for the constraint that the cell frequencies in a scRNA-seq data are not independent of each other. In detail, we normalized the regression coefficients to a standard normal distribution and calculated a z score, and then conducted significance testing based on the regression model generated by the *DirichReg* function in the R package *DirichletReg* (v 0.7).

Differential gene expression analysis and Gene Ontology term enrichment analysis. Differential gene expression analysis was performed on all expressed genes using the *rank_genes_groups* function with method="t-test" in the *Scanpy* package, for two certain cell-types in a COVID-19 single-cell atlas. A gene was considered differentially expressed when a log2-fold change was >1 in the two conditions in comparison, and the Benjamini-Hochberg adjusted P-value was < 0.01. The top 200 highly expressed genes sorted by scores (implemented in *Scanpy*) of each cell-type were used as the input for GO analysis, and enriched GO terms were acquired for each group of cells of the "GO_Biological_Process_2018" dataset using the Python package *GSEApy*.

Inflammatory and cytokine score analysis. We defined the inflammatory score and the cytokine score for each cell following Ren et al. 2021, based on the expression of a defined collection of cytokine genes and inflammatory-response-related genes, which

were from Liberzon et al. 2015 (Supplementary Table 1). We then calculated the cytokine and inflammatory scores from the raw gene expression profile using the *score_genes* function implemented in the *Scanpy*.

Software availability.

SCALEX is available at https://github.com/jsxlei/SCALEX.

Data availability.

All data analyzed in this study are publicly available; the data sources are detailed in Supplementary Table 2.

Acknowledgements

We thank Jianbin Wang, Jin Gu and Fuchou Tang for helpful comments and advice. This work is supported by the State Key Research Development Program of China (Grant No. 2018YFA0107603 and 2019YFA0110002); the National Natural Science Foundation of China (Grants No. 91740204, 91940306, and 31761163007); the Beijing Advanced Innovation Center for Structural Biology; and the Tsinghua-Peking Joint Center for Life Sciences.

Author contributions

Q.C.Z. conceived and supervised the project. L.X. designed and implemented the SCALEX model. L.X. and K.T. validated the SCALEX model. L.X., K.T., and Y.L. analyzed the results. L.X. and Q.C.Z. wrote the manuscript, with inputs from all the authors.

Competing interests

The authors declare no competing interests.

Figures

Fig. 1 | The design and performance of SCALEX for single-cell data integration. **a**, SCALEX models the global structure of single-cell data using a variational autoencoder (VAE) framework. **b**, UMAP embeddings of the *PBMC* dataset before and after integration using SCALEX, Seurat v3, Harmony, Conos, or Scanorama integration, colored by batch and cell-type. **c**, Scatter plot showing a quantitative comparison of the silhouette score (y-axis) and the batch entropy mixing score (x-axis) on the benchmark datasets. **d**, UMAP embeddings of the SCALEX integration of the *human fetal atlas* dataset, colored by batch and cell-type. **e**, Comparison of computation efficiency on datasets of different sizes sampled from the whole *human fetal atlas* dataset) including runtime (left) and memory usage (right). **f**, UMAP embeddings of the mouse brain scATAC-seq dataset before (left) and after integration (middle, right); colored by data batch or Leiden clustering. **g**, UMAP embeddings of the PBMC cross-modality dataset before (left) and after integration (middle, right); colored by batch or cell-type.

Fig. 2 | Comparison of integration performance over partially-overlapping datasets by different methods. a, Comparison over the *liver* dataset. **b,** Comparison over simulated datasets with different numbers of common cell-types (obtained by down-sampling the *pancreas* dataset). Misalignments are highlighted with red circles.

Fig. 3 | Projecting heterogenous data into a generalized cell-embedding space. a, UMAP embeddings of the *pancreas* dataset after integration by SCALEX, colored by cell-type. **b**, UMAP embeddings of three projected pancreas data batches projected onto the pancreas space, colored by cell-types; the light gray shadows represent the original *pancreas* dataset. **c**, Confusion matrix between ground truth cell-types and those annotated by different methods. ARI, NMI and F1 scores (top) measure the annotation accuracy. **d**, UMAP embeddings of the *PBMC* dataset after integration and the two projected melanoma data batches onto the PBMC space, colored by cell-types with light

gray shadows represent the original *PBMC* dataset. **e**, The PBMC space that includes the original *PBMC* dataset and the two projected melanoma data batches. **f**, Annotating an uncharacterized small cell population in the *pancreas* dataset by projection of the bronchial epithelium data batches into the pancreas cell space. Only the uncharacterized cells in the *pancreas* dataset (left) and the SLC16A7+ epithelial cells in the bronchial epithelium data batches (right) are colored. **g**, Heatmap showing the normalized expression of the top-10 ranking specific genes for the uncharacterized cell population in different cell-types.

Fig. 4 | Construction of an expandable mouse single-cell atlas. a, Datasets acquired using different technologies (Smart-seq2, 10X, and Microwell-seq) covering various tissues used for construction of the mouse atlas. **b**, UMAP embeddings of the *mouse atlas* dataset colored by batch and tissue. **c**, UMAP embeddings of the *mouse atlas* after SCLAEX integration, labeld with and colored by cell-type. **d**, Two *Tabula Muris Senis* data batches and two mouse tissues (lung and kidney) data are projected onto the cell space of the mouse atlas, with the same cell-type color as in **c**. **e**, Confusion matrix of the cell-type annotations by SCALEX and those in the original studies. Color bar represents the percentage of cells in confusion matrix C_{ij} known to be cell-type i and predicted to be cell-type j.

Fig. 5 | Construction and expansion of a COVID-19 single-cell atlas. a, COVID-19 dataset composition, including healthy controls and influenza patients, as well as mild/moderate, severe, and convalescent COVID-19 patients. **b**,**c** UMAP embeddings of COVID-19 PBMC atlas after SCLAEX integration colored by batch (**b**), and by cell-types (**c**). **d**, UMAP embeddings of the COVID-19 PBMC atlas separated by disease state. **e**, Stacked violinplot of differentially-expressed ISGs among CD14 monocytes across disease states. **f**, GO terms enriched in the differentially-expressed genes for CD14-IL1B-Mono and CD14-ISG15-Mono cells. **g**, Cell-type frequency across healthy and influenza controls, and among mild/moderate, severe, and convalescent COVID-

19 patients. Dirichlet-multinomial regression was used for pairwise comparisons, ***p<0.001, **p<0.01, *p<0.05. **h**, United UMAP embeddings of the SCALEX COVID-19 PBMC atlas and the SS4 atlas (from the Single Cell Consortium for COVID-19 in China, projected onto the cell space of the SCALEX COVID-19 PBMC atlas). Left: the SCALEX COVID-19 PBMC atlas, middle: SC4 colored by cell clusters in the original study, right: Expanded atlas combining the SCALEX COVID-19 PBMC atlas and the SC4 atlas. **i**, Similarity matrix of meta-cell representations for cell-types between the SCALEX COVID-19 PBMC atlas and SC4 in the generalized cell-embedding space after SCALEX integration. Color bar represents the Pearson correlation coefficient between the average meta-cell representation of two cell-types from a respective data batch. **j**, UMAP embeddings of the SCALEX COVID-19 PBMC atlas colored by the cytokine score and the inflammatory score. **k**, GO terms enriched in the differentially-expressed genes for TUBA8-Mega and IGKC-Mega cells.

Supplementary figures

Fig. S1 | Comparison of integration performance on benchmark datasets. UMAP embeddings for benchmark datasets grouped by batches and cell-types, before and after integration by different methods. Misalignments are highlighted with red circles.

Fig. S2 | The human fetal atlas. a, UMAP embeddings of the *human fetal atlas* dataset colored by batch before integration. **b,** Similarity matrix of meta-cell representations for different cell-types in the two data batches in the generalized cell-embedding space. Color bar represents the Pearson correlation coefficient between the average meta-cell representation of two cell-types from a respective data batch. **c,** Comparison of computation efficiency on datasets of different sizes (sampled from the whole *human fetal atlas* dataset), including runtime (left) and memory usage (right), in log scale.

Fig. S3 | Canonical marker genes of different cell-types and UMAP embeddings of the *liver* dataset. a, Dotplot of canonical marker genes for each cell-type. Dot color

represents average expression level, while dot size represents the proportion of cells in the group expressing the marker. **b,** UMAP embeddings of the *liver* dataset, colored by batch (left) and cell-type (right) after SCALEX integration. **c**, Normalized marker gene expression on the UMAP embeddings of the five hepatocyte subtypes. Color bar represents the expression level.

Fig. S4 | Integration over partially-overlapping datasets down-sampled from the *pancreas* dataset. Partially-overlapping datasets were generated by down-sampling the *pancreas* dataset, consisted of common cell-types with a decreased overlapping number (ranging from 0 to 6). Integration results for SCALEX, Seurat, and Harmony are shown in the UMAP embeddings colored by batches (left) and cell-types (right) respectively (overlapping number decreases from 6 to 0). Misalignments are highlighted with red circles.

Fig. S5 | Integration over partially-overlapping datasets down-sampled from the *PBMC* dataset. Partially-overlapping datasets were generated by down-sampling the *PBMC* dataset, consisted of common cell-types with a decreased overlapping number (ranging from 0 to 6). Integration results for SCALEX, Seurat and Harmony are shown in the UMAP embeddings colored by batches (left) and cell-types (right) respectively (overlapping number decreases from 6 to 0). Misalignments are highlighted with red circles.

Fig. S6 | The *pancreas* dataset and the additional data batches. a, UMAP embeddings of the *pancreas* dataset, the three additional pancreas data batches and the bronchial epithelium data batches (data from three donors), grouped by batch. **b**, Dot plot of canonical markers of cell-types of reference *pancreas* dataset; dot color represents average expression level, while dot size represents the proportion of cells in the group expressing the marker.

Fig. S7 | The SCALEX mouse atlas. a, UMAP embeddings of the mouse atlas data before integration, colored by batch. **b**, UMAP embeddings of three mouse atlas data batches (Smart-seq2, 10X, and Microwell-seq) after integration, colored by cell-type; the light gray shadows represent the original *mouse atlas* dataset. **c**, Dotplot of the top 5 cell-type-specific genes for each cell-type in the *mouse atlas* dataset. Dot color represents average expression level, while dot size represents the proportion of cells in the group expressing the marker.

Fig. S8 | The SCALEX human atlas. a, The *human atlas* dataset acquired using different technologies (Smart-seq2, 10X, and Microwell-seq) covering various tissues used for construction of the human atlas. **b-c**, UMAP embeddings of the *human atlas* dataset colored by batch and cell-type, before (**b**) and after integration (**c**). **d**, Similarity matrix of meta-cell representations for cell-types in the two data batches in the generalized cell-embedding space after SCALEX integration between two batches. Color bar represents the Pearson correlation coefficient between the average meta-cell representation of two cell-types from a respective data batch. **e**, UMAP embeddings of the human atlas and two additional projected data batches colored by cell-type. **f**, Confusion matrix of the cell-type annotations by SCALEX and those in the original study. Color bar represents the percentage of cells in confusion matrix C_{ij} known to be in cell-type *i* and predicted to be in cell-type j.

Fig. S9 | COVID-19 immune landscape. a, UMAP embeddings of the raw COVID-19 PBMC dataset before integration. **b**, UMAP embeddings of the COVID-19 PBMC atlas colored by condition and Leiden clustering after SCALEX integration. **c**, Dotplot of canonical marker genes for each cell-type. Dot color represents average expression level, while dot size represents the proportion of cells in the group expressing the marker. **d**, UMAP embeddings of the COVID-19 PBMC atlas in individual batches after SCALEX integration, colored by cell-type; the light gray shadows represent the other batches of COVID-19 PBMC atlas. **e**, Frequency of cell distributions across

healthy people and influenza patient controls, and among mild/moderate, severe, and convalescent COVID-19 patients. Dirichlet-multinomial regression was used for pairwise comparisons, ***p<0.001, **p<0.01, *p<0.05.

Fig. S10 | COVID-19 heterogeneous dysfunctional immune response. a, Stacked violin plot of differentially-expressed genes between PNPLA2-Immature_Neutrophil and NCF1-Immature_Neutrophil cells. **b**, GO terms enriched in the differentially-expressed genes for PNPLA2-Immature_Neutrophil and NCF1-Immature_Neutrophil cells. **c**, Stacked violinplot of differentially-expressed genes between PRDM1-Plasma and MZB1-Plasma. **d**, GO terms enriched in the differentially-expressed genes for PRDM1-Plasma and MZB1-Plasma cells.

Fig S11 | Projection of the SC4 dataset onto the SCLAEX COVID-19 PBMC atlas. **a-b**, UMAP embeddings of the SC4 dataset before integration (**a**) and after projection onto the SCLAEX COVID-19 PBMC space (**b**). **c**, Separate UMAP embeddings of each SC4 data batch, after being projected onto the SCALEX COVID-19 PBMC space, colored by cell-type. **d**, UMAP embeddings of the TUBA8-Mega and IGKC-Mega cells. **e**, UMAP embeddings of the differentially-expressed genes of TUBA8-Mega and IGKC-Mega cells.

Figures

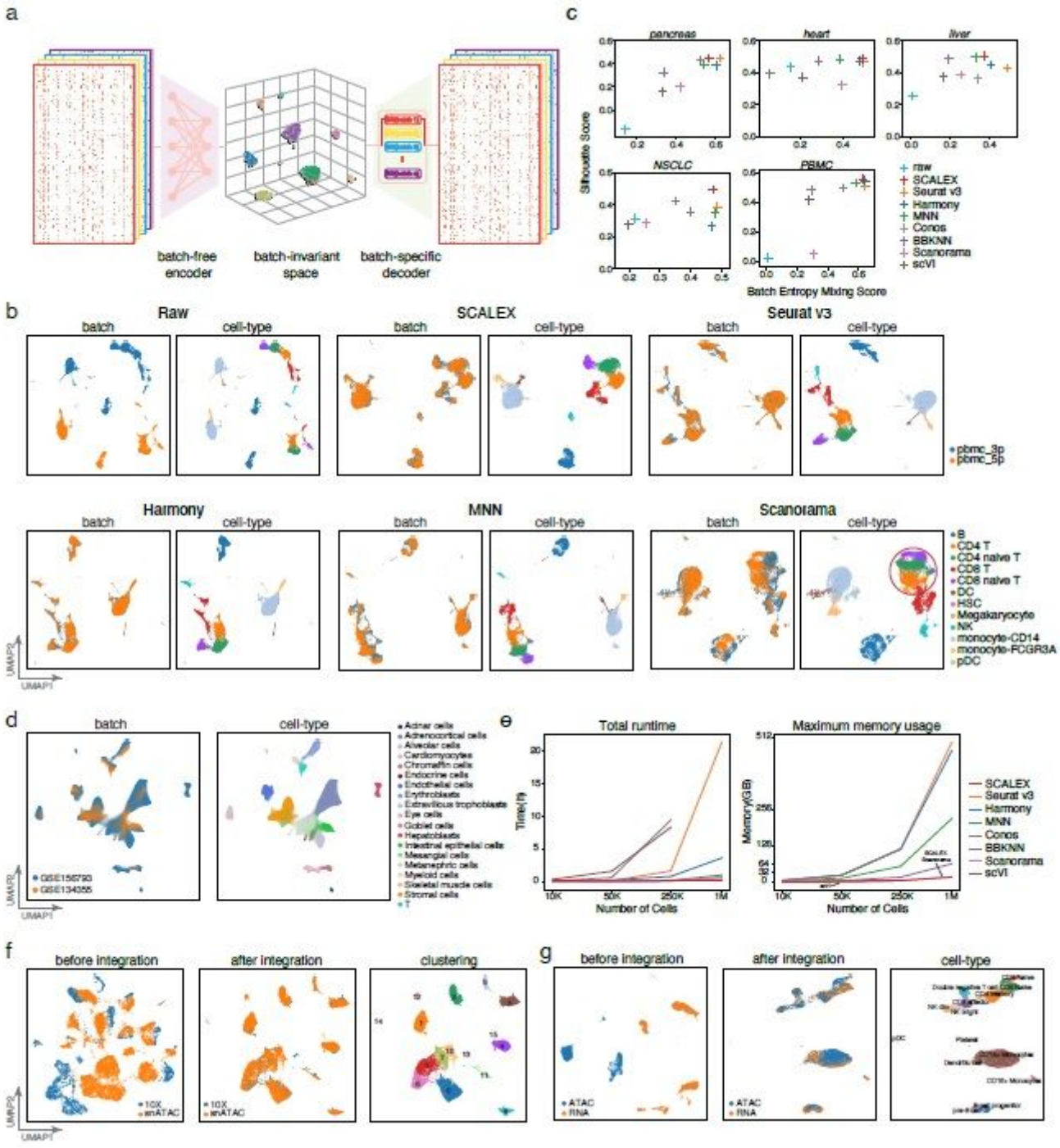

Figure 1

The design and performance of SCALEX for single-cell data integration. a, SCALEX models the global structure of single-cell data using a variational autoencoder (VAE) framework. b, UMAP embeddings of the PBMC dataset before and after integration using SCALEX, Seurat v3, Harmony, Conos, or Scanorama integration, colored by batch and cell-type. c, Scatter plot showing a quantitative comparison of the silhouette score (y-axis) and the batch entropy mixing score (x-axis) on the benchmark datasets. d, UMAP

embeddings of the SCALEX integration of the human fetal atlas dataset, colored by batch and cell-type. e, Comparison of computation efficiency on datasets of different sizes sampled from the whole human fetal atlas dataset) including runtime (left) and memory usage (right). f, UMAP embeddings of the mouse brain scATAC-seq dataset before (left) and after integration (middle, right); colored by data batch or Leiden clustering. g, UMAP embeddings of the PBMC cross-modality dataset before (left) and after integration (middle, right); colored by batch or cell-type.

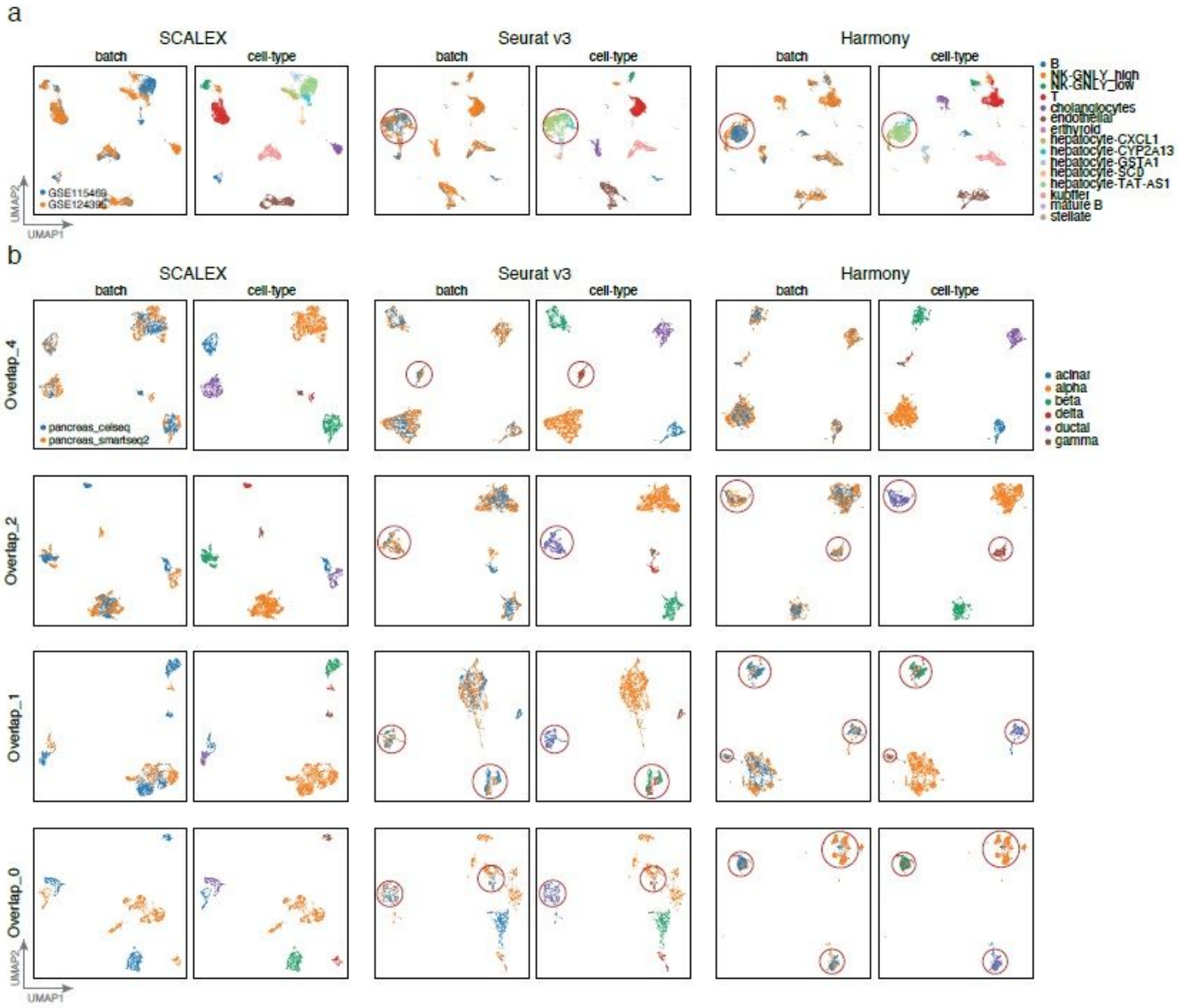

Figure 2

Comparison of integration performance over partially-overlapping datasets by different methods. a, Comparison over the liver dataset. b, Comparison over simulated datasets with different numbers of common cell-types (obtained by down-sampling the pancreas dataset). Misalignments are highlighted with red circles.

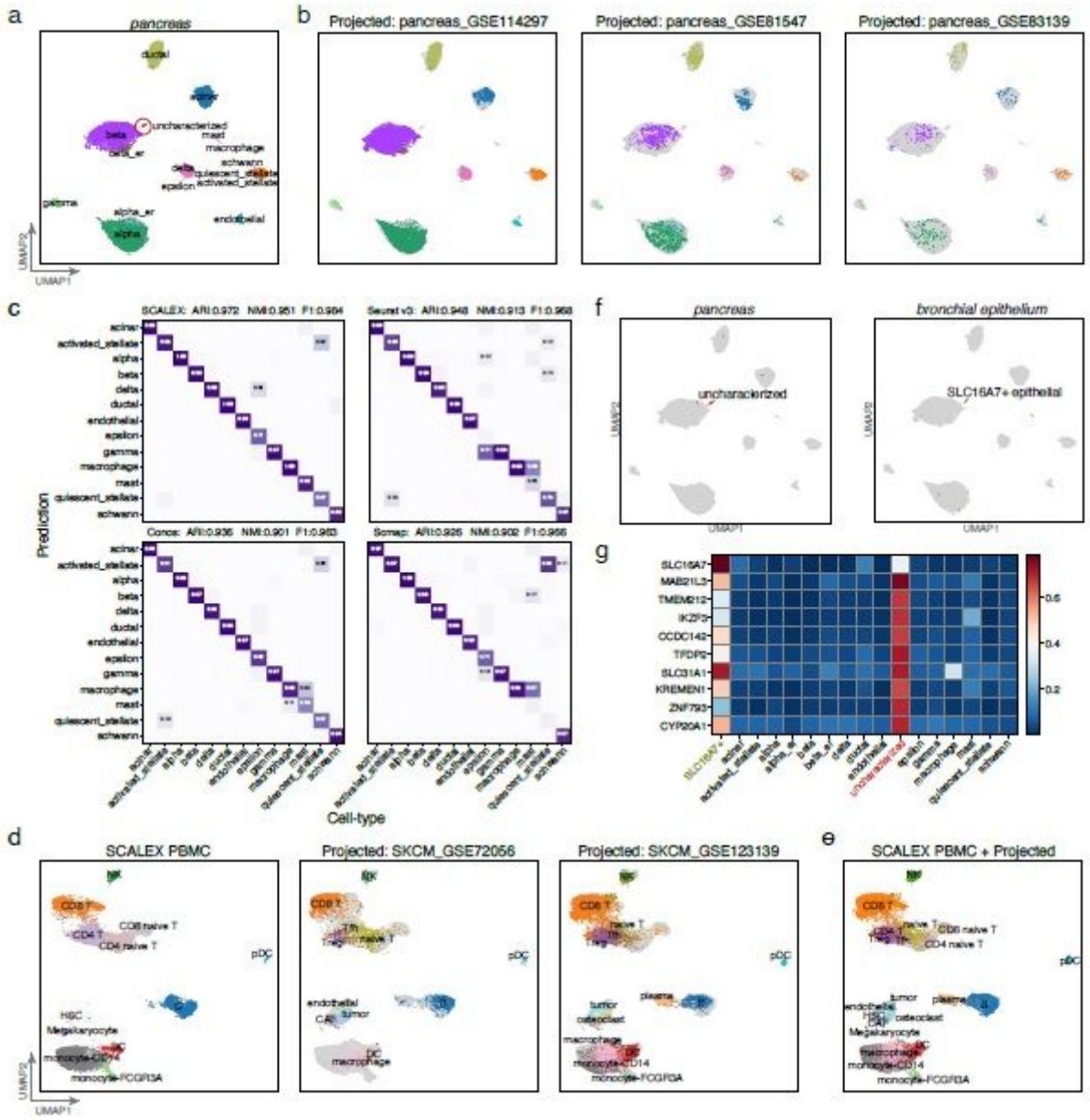

Figure 3

Projecting heterogenous data into a generalized cell-embedding space. a, UMAP embeddings of the pancreas dataset after integration by SCALEX, colored by cell-type. b, UMAP embeddings of three projected pancreas data batches projected onto the pancreas space, colored by cell-types; the light gray shadows represent the original pancreas dataset. c, Confusion matrix between ground truth cell-types and those annotated by different methods. ARI, NMI and F1 scores (top) measure the annotation accuracy. d, UMAP embeddings of the PBMC dataset after integration and the two projected melanoma data batches onto the PBMC space, colored by cell-types with light gray shadows represent the original PBMC dataset. e, The PBMC space that includes the original PBMC dataset and the two projected melanoma data batches. f, Annotating an uncharacterized small cell population in the pancreas dataset by projection of the bronchial epithelium data batches into the pancreas cell space. Only the uncharacterized cells in the

pancreas dataset (left) and the SLC16A7+ epithelial cells in the bronchial epithelium data batches (right) are colored. g, Heatmap showing the normalized expression of the top-10 ranking specific genes for the uncharacterized cell population in different cell-types.

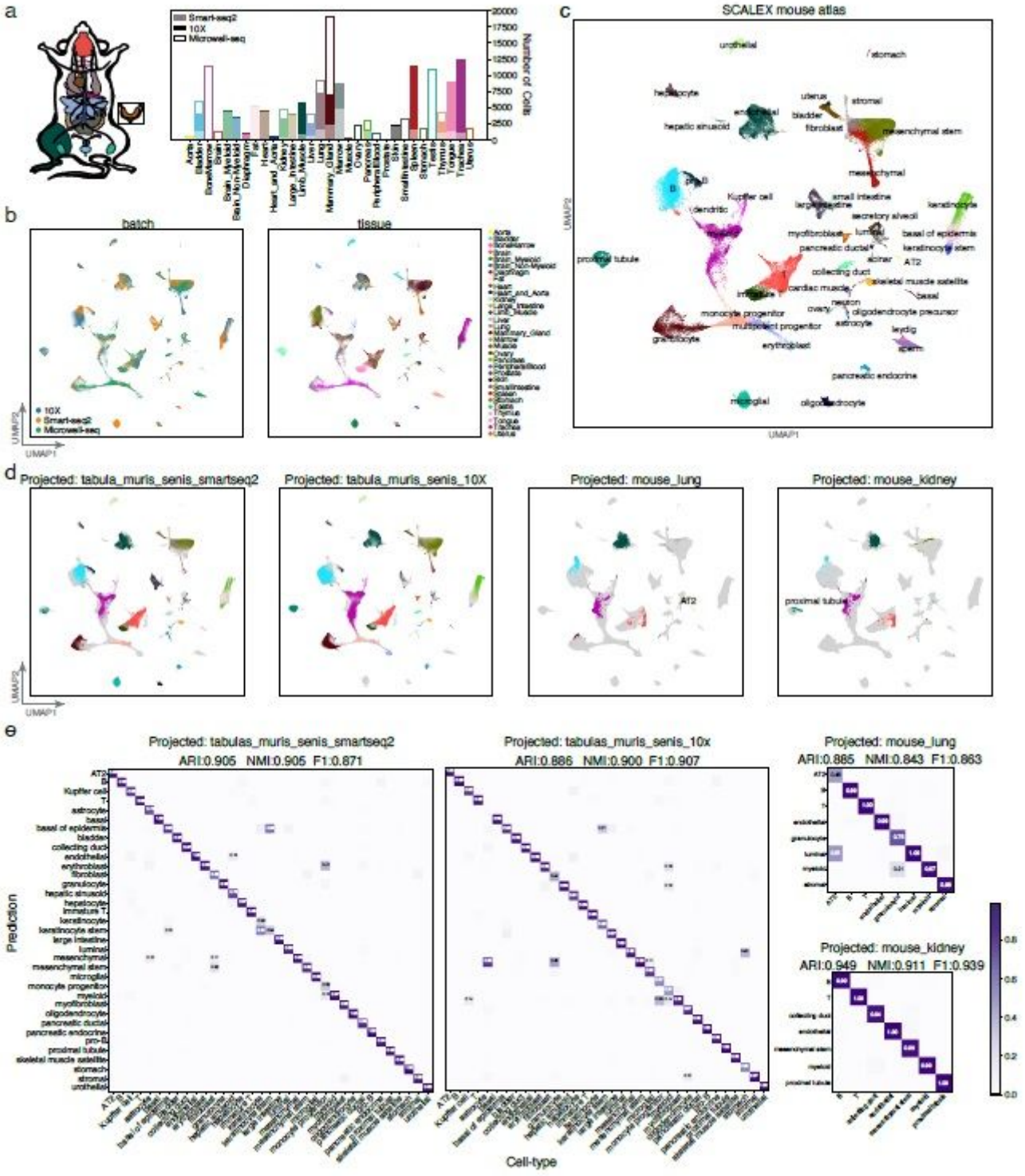

Figure 4

Construction of an expandable mouse single-cell atlas. a, Datasets acquired using different technologies (Smart-seq2, 10X, and Microwell-seq) covering various tissues used for construction of the mouse atlas. b, UMAP embeddings of the mouse atlas dataset colored by batch and tissue. c, UMAP embeddings of

the mouse atlas after SCLAEX integration, labeld with and colored by cell-type. d, Two Tabula Muris Senis data batches and two mouse tissues (lung and kidney) data are projected onto the cell space of the mouse atlas, with the same cell-type color as in c. e, Confusion matrix of the cell-type annotations by SCALEX and those in the original studies. Color bar represents the percentage of cells in confusion matrix Cij known to be cell-type i and predicted to be cell-type j.

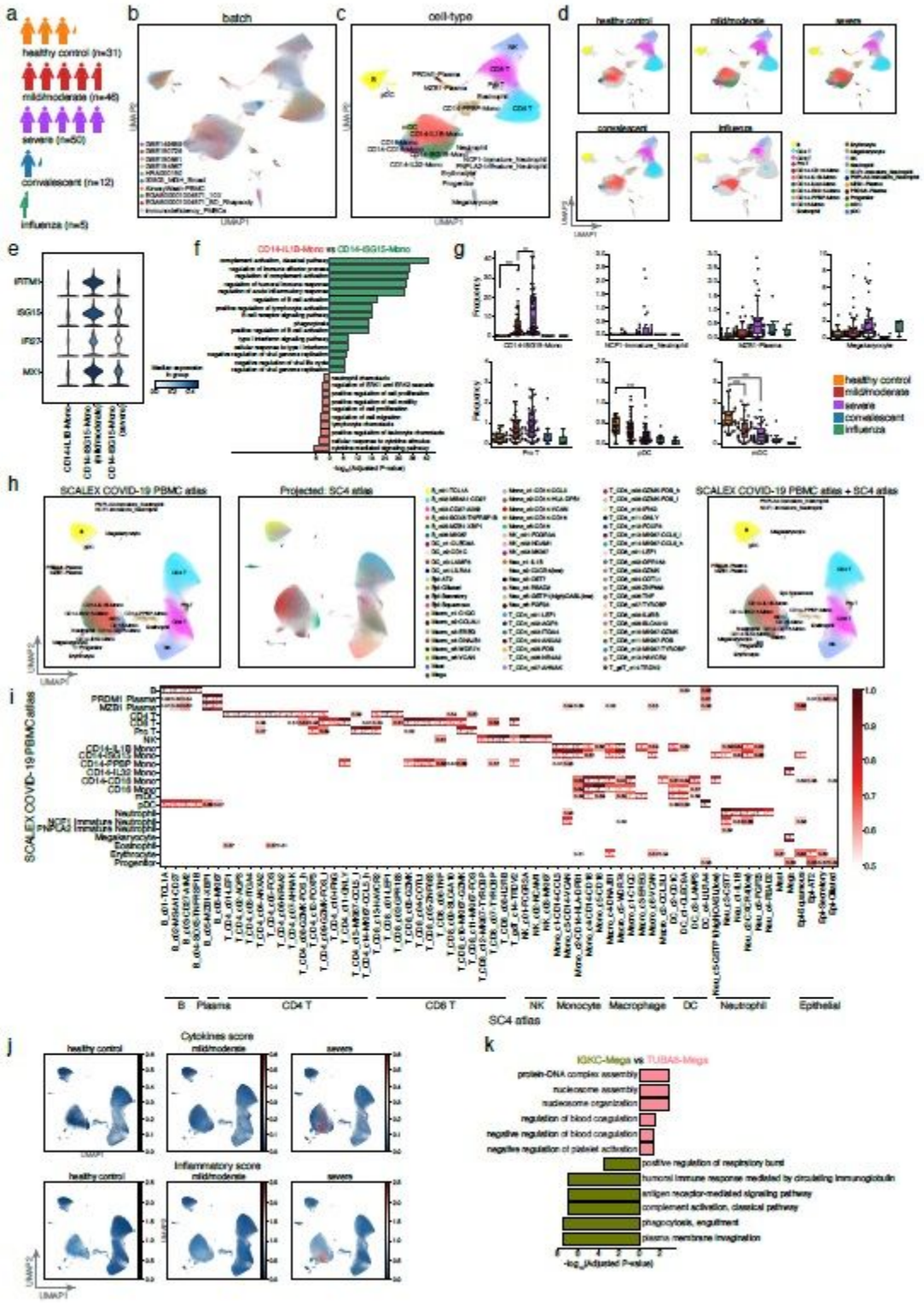

Figure 5

Construction and expansion of a COVID-19 single-cell atlas. a, COVID-19 dataset composition, including healthy controls and influenza patients, as well as mild/moderate, severe, and convalescent COVID-19 patients. b,c UMAP embeddings of COVID-19 PBMC atlas after SCLAEX integration colored by batch (b), and by cell-types (c). d, UMAP embeddings of the COVID-19 PBMC atlas separated by disease state. e, Stacked violinplot of differentially-expressed ISGs among CD14 monocytes across disease states. f, GO terms enriched in the differentially-expressed genes for CD14-IL1B-Mono and CD14-ISG15-Mono cells. g, Cell-type frequency across healthy and influenza controls, and among mild/moderate, severe, and convalescent COVID-19 patients. Dirichlet-multinomial regression was used for pairwise comparisons, ***p<0.001, **p<0.01, *p<0.05. h, United UMAP embeddings of the SCALEX COVID-19 PBMC atlas and the SS4 atlas (from the Single Cell Consortium for COVID-19 in China, projected onto the cell space of the SCALEX COVID-19 PBMC atlas). Left: the SCALEX COVID-19 PBMC atlas, middle: SC4 colored by cell clusters in the original study, right: Expanded atlas combining the SCALEX COVID-19 PBMC atlas and the SC4 atlas. i, Similarity matrix of meta-cell representations for cell-types between the SCALEX COVID-19 PBMC atlas and SC4 in the generalized cell-embedding space after SCALEX integration. Color bar represents the Pearson correlation coefficient between the average meta-cell representation of two cell-types from a respective data batch. j, UMAP embeddings of the SCALEX COVID-19 PBMC atlas colored by the cytokine score and the inflammatory score. k, GO terms enriched in the differentially-expressed genes for TUBA8-Mega and IGKC-Mega cells.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTable1Inflammatoryandcytokinegenes.xlsx
- SupplementaryTable2datasets.xlsx
- SuppFigure.pdf