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.

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

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

73 We previously applied VAE and designed SCALE (Single-Cell ATAC-seq
74 Analysis via Latent feature Extraction) to model and analyze single-cell ATAC-seq
75 data[20]. We found that the VAE framework in SCALE can disentangle cell-type-related
76 and batch-related features in a low-dimensional embedding space. Here, having
77 redesigned the VAE framework, we introduce SCALEX as a method for integration of
78 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

105 the original data from *z* by incorporating batch information back during data
106 reconstruction.

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

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

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

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

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

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

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

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

179 **SCALEX integrates partially-overlapping datasets**

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

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

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

202 **Projection of unseen data into an existing cell-embedding space**

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

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

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

222 **Expanding an existing cell space by including new data**

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

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

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

244 SCALEX supports construction of expandable single-cell atlases

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

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

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

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

285 **An integrative SCALEX COVID-19 PBMC atlas**

286 Many single-cell studies have been conducted to analyze COVID-19 patient immune
287 responses[57-64]. However, these studies often suffer from small sample size and/or
288 limited sampling of various disease states[58,64]. For a comprehensive study, we collected
289 data from multiple COVID-19 PBMC studies, involving 860,746 single cells, and 10
290 batches from 9 studies[57-63] (Fig. 5a, Fig. S9a), and used SCALEX to generate a
291 COVID-19 PBMC atlas, identifying 22 cell-types, each of which were supported by
292 canonical marker gene expression (Fig. 5b,c, Fig. S9b,c. Methods). Cells across
293 different studies were integrated accurately with the same cell-types aligned together,
294 confirming integration performance of SCALEX (Fig. 5c, Fig. S9d).

295 We observed that some cell subpopulations were differentially associated with
296 patient status (Fig 5d). A subpopulation of CD14 monocytes (CD14-ISG15-Mono),
297 specifically associated with COIVD-19 patients, was characterized by its high
298 expression of Type I interferon-stimulated genes (ISGs) and genes associated with
299 immune-response-related GO terms (Fig 5e,f). The frequency of CD14-ISG15-Mono
300 cells increased significantly from healthy donors to mild/moderate and severe patients
301 (Fig. 6g, Fig. S9e. Methods). Within the COVID-19 patients, we observed a significant
302 decrease in ISG gene expression in CD14-ISG15-Mono cells between the
303 mild/moderate and severe cases, indicating apparently dysfunctional anti-viral immune
304 response in severe COVID-19 patients (Fig. 5e). Specifically enriched in severe verse
305 mild/moderate patients, a neutrophil subpopulation (NCF1-Immature_Neutrophil)
306 lacked expression of the genes responsible for neutrophil activation but showed
307 elevated expression of genes associated with viral-process-related GO terms (Fig.
308 S10a,b). Also enriched in severe patients, a plasma cell subpopulation (MZB1-Plasma)
309 cells displayed decreased expression for antibody production and were enriched for GO
310 terms of immune and inflammatory responses (Fig. S10c,d). Thus, the SCALEX
311 COVID-19 PBMC atlas, generated by integrating a highly diverse collection of
312 single-cell data from individual studies, identified multiple immune cells-types
313 showing dysregulations during COVID-19 disease progression. Note that these trends

314 could not have been detected in the small-scale, individual studies that served as the
315 basis for our SCALEX COVID-19 PBMC atlas.

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

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

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

334 Integration of the SC4 data further substantially improved both the scope and
335 resolution of the SCALEX COVID-19 PBMC atlas. First, this data added macrophages
336 and epithelial cells to the cell space, enabling investigation of their potential
337 involvement in COVID-19. The integration also supported more precise
338 characterization of specific cell subpopulations. For example, the megakaryocyte
339 population, not distinguished in either single atlas, could be divided into two

340 subpopulations in the combined atlas (Fig. 5h). An exploratory functional analysis of
341 the differentially expressed genes in these two newly delineated megakaryocyte
342 subpopulations (TUBA8-Mega and IGKC-Mega, Fig. S11d,e) revealed enrichment for
343 the GO terms "humoral immune response" for IGKC-Mega cells yet enrichment for
344 "negative regulation of platelet activation" for TUBA8-Mega cells (Fig. 5k). These
345 results illustrate how the continuously expandable single-cell atlases generated using
346 SCALEX capitalize on existing large-scale data resources and also facilitate discovery
347 of biological and biomedical insights.

348 **DISCUSSION**

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

360 These two features make SCALEX particularly useful for construction and
361 integrative analysis of large-scale single-cell atlases based on very heterogenous data
362 (*i.e.*, datasets acquired by different labs and using different single-cell analysis
363 platforms). Our construction of human, mouse, and COVID-19 patient single-cell
364 atlases—which aligned well with previously reported atlases generated from
365 coordinated large-scale consortium efforts—demonstrates the particular ability of

366 SCALEX to producing large-scale atlases from extant small-scale datasets. SCALEX
367 achieves data integration by projecting all single cells into a generalized
368 cell-embedding space using a universal data projector (*i.e.*, the encoder). This data
369 projector only needs to be trained once, and then can be used without retraining to
370 continuously integrate new incoming data into an existing single-cell atlas. This
371 continuous growth ability makes a SCALEX atlas an elastic resource, allowing the
372 integration of many single-cell studies to support ongoing, very large-scale research
373 programs throughout the life sciences and biomedicine.

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

386

387 **REFERENCES**

388 1 Kolodziejczyk, A. A., Kim, J. K., Svensson, V., Marioni, J. C. & Teichmann,
389 S. A. The technology and biology of single-cell RNA sequencing. *Mol Cell* **58**,
390 610-620, doi:10.1016/j.molcel.2015.04.005 (2015).
391 2 Papalexi, E. & Satija, R. Single-cell RNA sequencing to explore immune cell
392 heterogeneity. *Nat Rev Immunol* **18**, 35-45, doi:10.1038/nri.2017.76 (2018).
393 3 Tanay, A. & Regev, A. Scaling single-cell genomics from phenomenology to
394 mechanism. *Nature* **541**, 331-338, doi:10.1038/nature21350 (2017).

395 4 Satpathy, A. T. *et al.* Massively parallel single-cell chromatin landscapes of
396 human immune cell development and intratumoral T cell exhaustion. *Nature*
397 *Biotechnology* **37**, 925-936, doi:10.1038/s41587-019-0206-z (2019).
398 5 Regev, A. *et al.* The Human Cell Atlas. *Elife* **6**, doi:10.7554/eLife.27041
399 (2017).
400 6 Tabula Muris, C. *et al.* Single-cell transcriptomics of 20 mouse organs creates
401 a Tabula Muris. *Nature* **562**, 367-372, doi:10.1038/s41586-018-0590-4 (2018).
402 7 Lahnemann, D. *et al.* Eleven grand challenges in single-cell data science.
403 *Genome Biol* **21**, 31, doi:10.1186/s13059-020-1926-6 (2020).
404 8 Leek, J. T. *et al.* Tackling the widespread and critical impact of batch effects
405 in high-throughput data. *Nat Rev Genet* **11**, 733-739, doi:10.1038/nrg2825
406 (2010).
407 9 Hicks, S. C., Townes, F. W., Teng, M. & Irizarry, R. A. Missing data and
408 technical variability in single-cell RNA-sequencing experiments. *Biostatistics*
409 **19**, 562-578, doi:10.1093/biostatistics/kxx053 (2018).
410 10 Haghverdi, L., Lun, A. T. L., Morgan, M. D. & Marioni, J. C. Batch effects in
411 single-cell RNA-sequencing data are corrected by matching mutual nearest
412 neighbors. *Nat Biotechnol* **36**, 421-427, doi:10.1038/nbt.4091 (2018).
413 11 Hie, B., Bryson, B. & Berger, B. Efficient integration of heterogeneous
414 single-cell transcriptomes using Scanorama. *Nat Biotechnol*,
415 doi:10.1038/s41587-019-0113-3 (2019).
416 12 Polanski, K. *et al.* BBKNN: fast batch alignment of single cell transcriptomes.
417 *Bioinformatics* **36**, 964-965, doi:10.1093/bioinformatics/btz625 (2020).
418 13 Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating
419 single-cell transcriptomic data across different conditions, technologies, and
420 species. *Nat Biotechnol* **36**, 411-420, doi:10.1038/nbt.4096 (2018).
421 14 Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**,
422 1888-1902 e1821, doi:10.1016/j.cell.2019.05.031 (2019).
423 15 Barkas, N. *et al.* Joint analysis of heterogeneous single-cell RNA-seq dataset
424 collections. *Nat Methods* **16**, 695-698, doi:10.1038/s41592-019-0466-z
425 (2019).
426 16 Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data
427 with Harmony. *Nat Methods*, doi:10.1038/s41592-019-0619-0 (2019).
428 17 Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative
429 modeling for single-cell transcriptomics. *Nat Methods* **15**, 1053-1058,
430 doi:10.1038/s41592-018-0229-2 (2018).
431 18 Kingma, D. P. & Welling, M. Auto-Encoding Variational Bayes.
432 *arXiv:1312.6114* (2013).
433 19 Rezende, D. J., Mohamed, S. & Wierstra, D. Stochastic Backpropagation and
434 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).

475 35 Kim, N. *et al.* Single-cell RNA sequencing demonstrates the molecular and
476 cellular reprogramming of metastatic lung adenocarcinoma. *Nat Commun* **11**,
477 2285, doi:10.1038/s41467-020-16164-1 (2020).
478 36 McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold
479 Approximation and Projection for Dimension Reduction. *arXiv:1802.03426*
480 (2018).
481 37 Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and
482 validation of cluster analysis. *Journal of Computational and Applied*
483 *Mathematics* **20**, 53-65, doi:https://doi.org/10.1016/0377-0427(87)90125-7
484 (1987).
485 38 Cao, J. *et al.* A human cell atlas of fetal gene expression. *Science* **370**,
486 doi:10.1126/science.aba7721 (2020).
487 39 Han, X. *et al.* Construction of a human cell landscape at single-cell level.
488 *Nature* **581**, 303-309, doi:10.1038/s41586-020-2157-4 (2020).
489 40 Fang, R. *et al.* Comprehensive analysis of single cell ATAC-seq data with
490 SnapATAC. *Nat Commun* **12**, 1337, doi:10.1038/s41467-021-21583-9 (2021).
491 41 Genomics, X. 10k Peripheral blood mononuclear cells (PBMCs) from a
492 healthy donor, Single Cell ATAC Dataset by Cell Ranger 1.0.1. (2018).
493 42 Genomics, X. 10k PBMCs from a Healthy Donor (v3 chemistry), Single Cell
494 Gene Expression Dataset by Cell Ranger 3.0.0. (2018).
495 43 Wang, Y. J. *et al.* Single-Cell Transcriptomics of the Human Endocrine
496 Pancreas. *Diabetes* **65**, 3028-3038, doi:10.2337/db16-0405 (2016).
497 44 Enge, M. *et al.* Single-Cell Analysis of Human Pancreas Reveals
498 Transcriptional Signatures of Aging and Somatic Mutation Patterns. *Cell* **171**,
499 321-330 e314, doi:10.1016/j.cell.2017.09.004 (2017).
500 45 Xin, Y. *et al.* Pseudotime Ordering of Single Human beta-Cells Reveals States
501 of Insulin Production and Unfolded Protein Response. *Diabetes* **67**, 1783-1794,
502 doi:10.2337/db18-0365 (2018).
503 46 Hubert, L. & Arabie, P. Comparing partitions. *Journal of Classification* **2**,
504 193-218, doi:10.1007/BF01908075 (1985).
505 47 Amelio, A. & Pizzuti, C. in *Proceedings of the 2015 IEEE/ACM International*
506 *Conference on Advances in Social Networks Analysis and Mining 2015*
507 1584–1585 (Association for Computing Machinery, Paris, France, 2015).
508 48 Tirosh, I. *et al.* Dissecting the multicellular ecosystem of metastatic melanoma
509 by single-cell RNA-seq. *Science* **352**, 189-196, doi:10.1126/science.aad0501
510 (2016).
511 49 Li, H. *et al.* Dysfunctional CD8 T Cells Form a Proliferative, Dynamically
512 Regulated Compartment within Human Melanoma. *Cell* **176**, 775-789.e718,
513 doi:https://doi.org/10.1016/j.cell.2018.11.043 (2019).
514 50 Plasschaert, L. W. *et al.* A single-cell atlas of the airway epithelium reveals
515 the CFTR-rich pulmonary ionocyte. *Nature* **560**, 377-381,
516 doi:10.1038/s41586-018-0394-6 (2018).

517 51 Han, X. *et al.* Mapping the Mouse Cell Atlas by Microwell-Seq. *Cell* **172**,
518 1091-1107 e1017, doi:10.1016/j.cell.2018.02.001 (2018).
519 52 He, S. *et al.* Single-cell transcriptome profiling of an adult human cell atlas of
520 15 major organs. *Genome Biol* **21**, 294, doi:10.1186/s13059-020-02210-0
521 (2020).
522 53 Tabula Muris, C. A single-cell transcriptomic atlas characterizes ageing
523 tissues in the mouse. *Nature* **583**, 590-595, doi:10.1038/s41586-020-2496-1
524 (2020).
525 54 Kimmel, J. C. *et al.* Murine single-cell RNA-seq reveals cell-identity- and
526 tissue-specific trajectories of aging. *Genome Res* **29**, 2088-2103,
527 doi:10.1101/gr.253880.119 (2019).
528 55 Sole-Boldo, L. *et al.* Single-cell transcriptomes of the human skin reveal
529 age-related loss of fibroblast priming. *Commun Biol* **3**, 188,
530 doi:10.1038/s42003-020-0922-4 (2020).
531 56 He, H. *et al.* Single-cell transcriptome analysis of human skin identifies novel
532 fibroblast subpopulation and enrichment of immune subsets in atopic
533 dermatitis. *J Allergy Clin Immunol* **145**, 1615-1628,
534 doi:10.1016/j.jaci.2020.01.042 (2020).
535 57 Schulte-Schrepping, J. *et al.* Severe COVID-19 Is Marked by a Dysregulated
536 Myeloid Cell Compartment. *Cell* **182**, 1419-1440.e1423,
537 doi:10.1016/j.cell.2020.08.001 (2020).
538 58 Lee, J. S. *et al.* Immunophenotyping of COVID-19 and influenza highlights
539 the role of type I interferons in development of severe COVID-19. *Sci
540 Immunol* **5**, doi:10.1126/sciimmunol.abd1554 (2020).
541 59 Wilk, A. J. *et al.* A single-cell atlas of the peripheral immune response in
542 patients with severe COVID-19. *Nat Med* **26**, 1070-1076,
543 doi:10.1038/s41591-020-0944-y (2020).
544 60 Guo, C. *et al.* Single-cell analysis of two severe COVID-19 patients reveals a
545 monocyte-associated and tocilizumab-responding cytokine storm. *Nat
546 Commun* **11**, 3924, doi:10.1038/s41467-020-17834-w (2020).
547 61 Yao, C. *et al.* Cell-Type-Specific Immune Dysregulation in Severely Ill
548 COVID-19 Patients. *Cell Rep* **34**, 108590, doi:10.1016/j.celrep.2020.108590
549 (2021).
550 62 Zhang, J. Y. *et al.* Single-cell landscape of immunological responses in
551 patients with COVID-19. *Nat Immunol* **21**, 1107-1118,
552 doi:10.1038/s41590-020-0762-x (2020).
553 63 Ballestar, E. *et al.* Single cell profiling of COVID-19 patients: an international
554 data resource from multiple tissues. *medRxiv*, 2020.2011.2020.20227355,
555 doi:10.1101/2020.11.20.20227355 (2020).
556 64 Bernardes, J. P. *et al.* Longitudinal Multi-omics Analyses Identify Responses
557 of Megakaryocytes, Erythroid Cells, and Plasmablasts as Hallmarks of Severe

558 COVID-19. *Immunity* **53**, 1296-1314 e1299,
559 doi:10.1016/j.immuni.2020.11.017 (2020).
560 65 Ren, X. *et al.* COVID-19 immune features revealed by a large-scale single-cell
561 transcriptome atlas. *Cell*, doi:10.1016/j.cell.2021.01.053 (2021).
562 66 Chen, G. *et al.* Clinical and immunological features of severe and moderate
563 coronavirus disease 2019. *J Clin Invest* **130**, 2620-2629,
564 doi:10.1172/JCI137244 (2020).
565 67 Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization.
566 *arXiv:1412.6980* (2014).
567 68 Danese, A., Richter, M. L., Fischer, D. S., Theis, F. J. & Colomé-Tatché, M.
568 EpiScanpy: integrated single-cell epigenomic analysis. *bioRxiv*,
569 doi:10.1101/648097 (2019).
570 69 Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene
571 expression data analysis. *Genome Biol* **19**, 15,
572 doi:10.1186/s13059-017-1382-0 (2018).
573 70 Stuart, T., Srivastava, A., Lareau, C. & Satija, R. Multimodal single-cell
574 chromatin analysis with Signac. *bioRxiv*, doi:10.1101/2020.11.09.373613
575 (2020).
576 71 Abdi, H. & Williams, L. J. Principal component analysis. *WIREs Comput. Stat.*
577 **2**, 433-459, doi:10.1002/wics.101 (2010).
578 72 Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden:
579 guaranteeing well-connected communities. *Scientific Reports* **9**, 5233,
580 doi:10.1038/s41598-019-41695-z (2019).
581
582

843 **Figures**

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

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

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

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

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

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

896 across healthy and influenza controls, and among mild/moderate, severe, and
897 convalescent COVID-19 patients. Dirichlet-multinomial regression was used for
898 pairwise comparisons, ***p<0.001, **p<0.01, *p<0.05. **h**, United UMAP
899 embeddings of the SCALEX COVID-19 PBMC atlas and the SS4 atlas (from the
900 Single Cell Consortium for COVID-19 in China, projected onto the cell space of the
901 SCALEX COVID-19 PBMC atlas). Left: the SCALEX COVID-19 PBMC atlas,
902 middle: SC4 colored by cell clusters in the original study, right: Expanded atlas
903 combining the SCALEX COVID-19 PBMC atlas and the SC4 atlas. **i**, Similarity
904 matrix of meta-cell representations for cell-types between the SCALEX COVID-19
905 PBMC atlas and SC4 in the generalized cell-embedding space after SCALEX
906 integration. Color bar represents the Pearson correlation coefficient between the
907 average meta-cell representation of two cell-types from a respective data batch. **j**,
908 UMAP embeddings of the SCALEX COVID-19 PBMC atlas colored by the cytokine
909 score and the inflammatory score. **k**, GO terms enriched in the
910 differentially-expressed genes for TUBA8-Mega and IGKC-Mega cells.

911 **Supplementary figures**

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

915 **Fig. S2 | The human fetal atlas. a,** UMAP embeddings of the *human fetal atlas*
916 dataset colored by batch before integration. **b,** Similarity matrix of meta-cell
917 representations for different cell-types in the two data batches in the generalized
918 cell-embedding space. Color bar represents the Pearson correlation coefficient
919 between the average meta-cell representation of two cell-types from a respective data
920 batch. **c,** Comparison of computation efficiency on datasets of different sizes

921 (sampled from the whole *human fetal atlas* dataset), including runtime (left) and
922 memory usage (right), in log scale.

923 **Fig. S3 | Canonical marker genes of different cell-types and UMAP embeddings**
924 **of the *liver* dataset. a,** Dotplot of canonical marker genes for each cell-type. Dot
925 color represents average expression level, while dot size represents the proportion of
926 cells in the group expressing the marker. **b,** UMAP embeddings of the *liver* dataset,
927 colored by batch (left) and cell-type (right) after SCALEX integration. **c**, Normalized
928 marker gene expression on the UMAP embeddings of the five hepatocyte subtypes.
929 Color bar represents the expression level.

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

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

944 **Fig. S6 | The *pancreas* dataset and the additional data batches. a**, UMAP
945 embeddings of the *pancreas* dataset, the three additional pancreas data batches and
946 the bronchial epithelium data batches (data from three donors), grouped by batch. **b**,

947 Dot plot of canonical markers of cell-types of reference *pancreas* dataset; dot color
948 represents average expression level, while dot size represents the proportion of cells
949 in the group expressing the marker.

950 **Fig. S7 | The SCALEX mouse atlas. a**, UMAP embeddings of the mouse atlas data
951 before integration, colored by batch. **b**, UMAP embeddings of three mouse atlas data
952 batches (Smart-seq2, 10X, and Microwell-seq) after integration, colored by cell-type;
953 the light gray shadows represent the original *mouse atlas* dataset. **c**, Dotplot of the top
954 5 cell-type-specific genes for each cell-type in the *mouse atlas* dataset. Dot color
955 represents average expression level, while dot size represents the proportion of cells
956 in the group expressing the marker.

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

969 **Fig. S9 | COVID-19 immune landscape. a**, UMAP embeddings of the raw
970 COVID-19 PBMC dataset before integration. **b**, UMAP embeddings of the
971 COVID-19 PBMC atlas colored by condition and Leiden clustering after SCALEX
972 integration. **c**, Dotplot of canonical marker genes for each cell-type. Dot color

973 represents average expression level, while dot size represents the proportion of cells
974 in the group expressing the marker. **d**, UMAP embeddings of the COVID-19 PBMC
975 atlas in individual batches after SCALEX integration, colored by cell-type; the light
976 gray shadows represent the other batches of COVID-19 PBMC atlas. **e**, Frequency of
977 cell distributions across healthy people and influenza patient controls, and among
978 mild/moderate, severe, and convalescent COVID-19 patients. Dirichlet-multinomial
979 regression was used for pairwise comparisons, ***$p<0.001$, **$p<0.01$, *$p<0.05$.

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

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

994

Fig 1

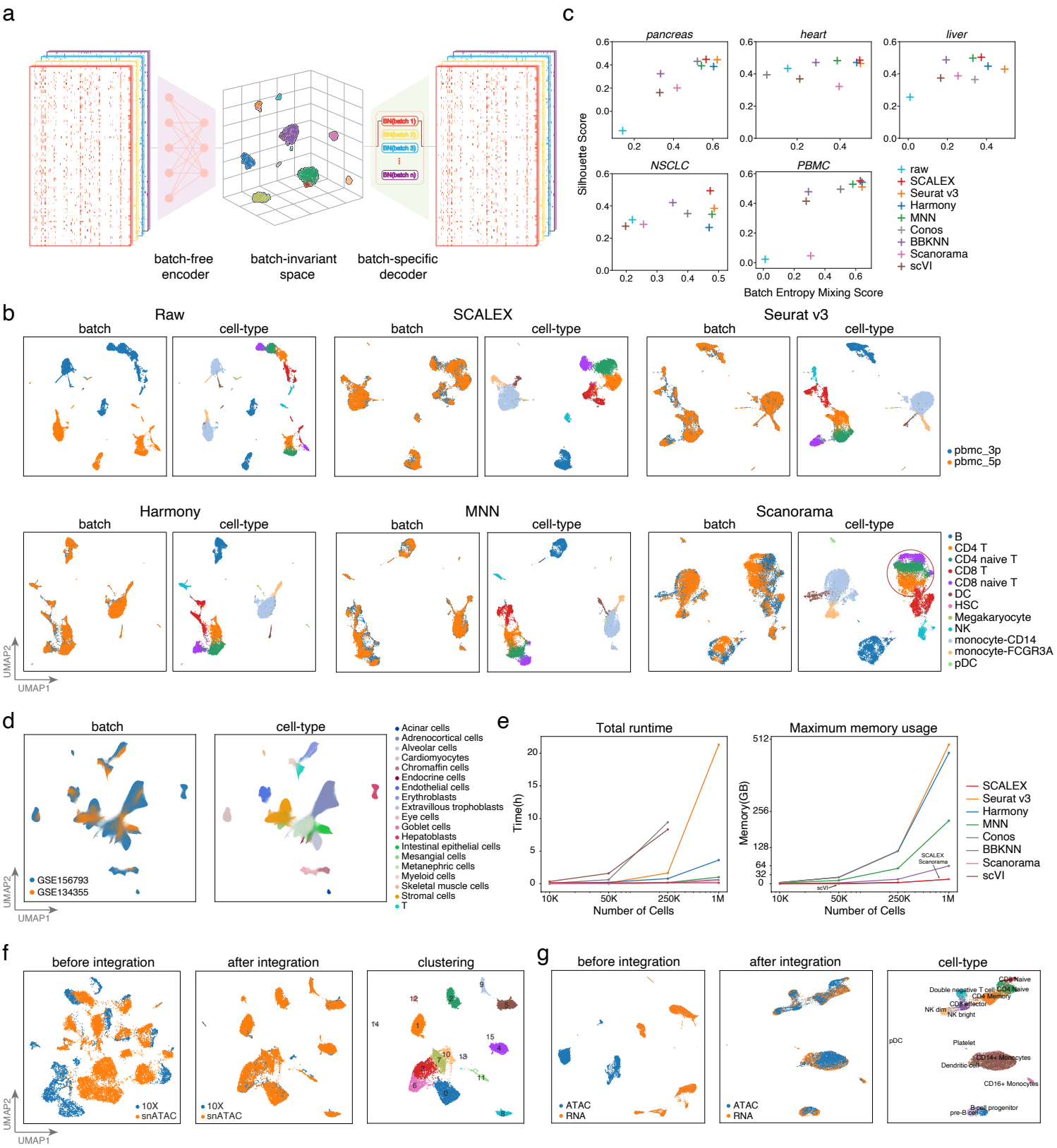

a
batch-free encoder
batch-invariant space
batch-specific decoder
BN(batch 1)
BN(batch 2)
BN(batch 3)
BN(batch n)
c
pancreas
heart
liver
NSCLC
PBMC
Silhouette Score
Batch Entropy Mixing Score
raw
SCALEX
Seurat v3
Harmony
MNN
Conos
BBKNN
Scanorama
scVI
b
Raw
SCALEX
Seurat v3
Harmony
MNN
Scanorama
batch
cell-type
pbmc_3p
pbmc_5p
B
CD4 T
CD4 naive T
CD8 T
CD8 naive T
DC
HSC
Megakaryocyte
NK
monocyte-CD14
monocyte-FCGR3A
pDC
UMAP2
UMAP1
d
batch
cell-type
GSE156793
GSE134355
Acinar cells
Adrenocortical cells
Alveolar cells
Cardiomyocytes
Chromaffin cells
Endocrine cells
Endothelial cells
Erythroblasts
Extravillous trophoblasts
Eye cells
Goblet cells
Hepatoblasts
Intestinal epithelial cells
Mesangial cells
Metanephric cells
Myeloid cells
Skeletal muscle cells
Stromal cells
T
e
Total runtime
Maximum memory usage
Time(h)
Memory(GB)
Number of Cells
10K
50K
250K
1M
SCALEX
Seurat v3
Harmony
MNN
Conos
BBKNN
Scanorama
scVI
f
before integration
after integration
clustering
10X
snATAC
g
before integration
after integration
cell-type
ATAC
RNA
CD8 Naive
Double negative T cell
CD4 Naive
CD4 Memory
CD8 effector
NK dim
NK bright
pDC
Platelet
CD14+ Monocytes
Dendritic cell
CD16+ Monocytes
B cell progenitor
pre-B cell

Fig 2

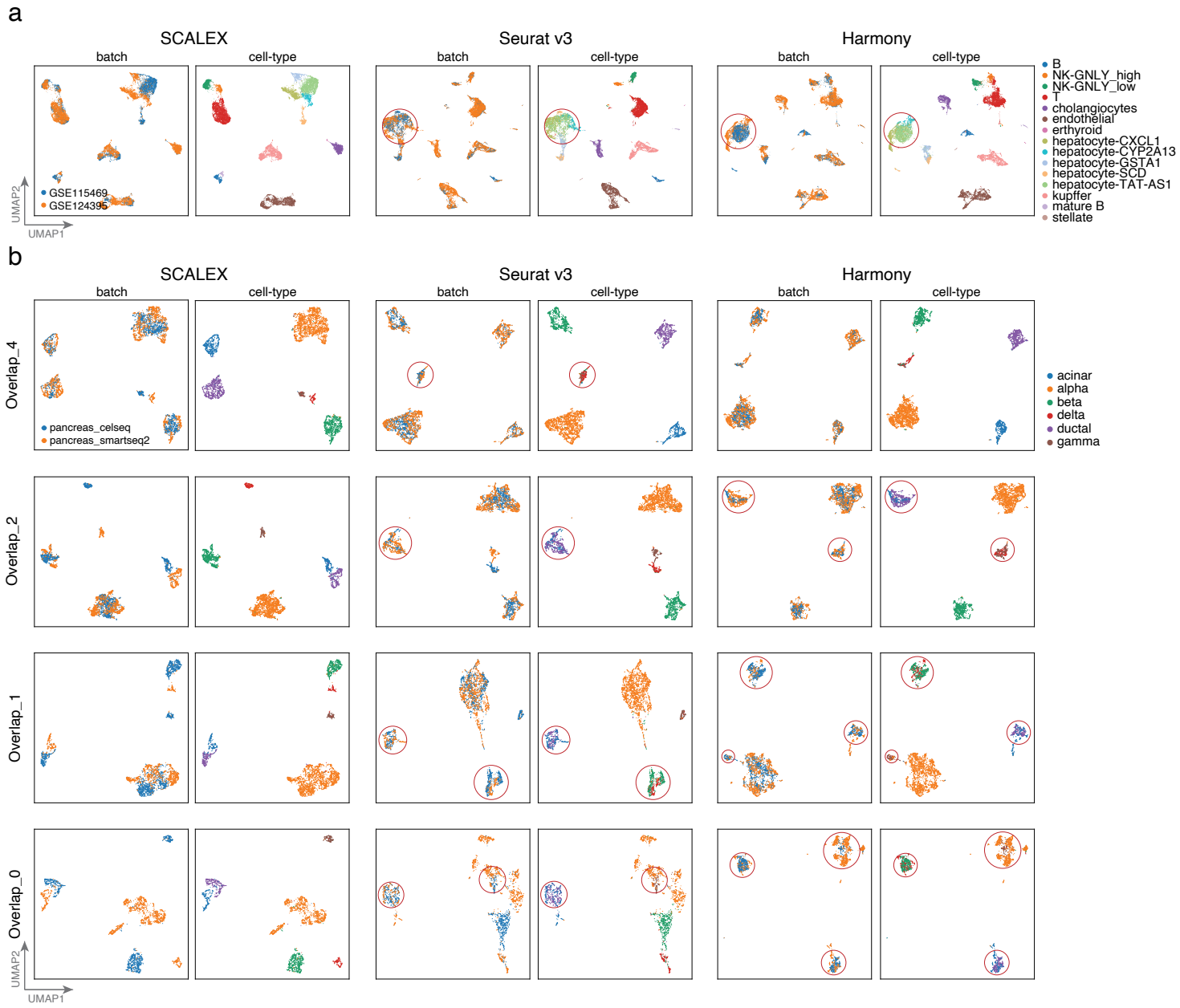

a
SCALEX
Seurat v3
Harmony
batch
cell-type
GSE115469
GSE124395
UMAP2
UMAP1
B
NK-GNLY_high
NK-GNLY_low
T
cholangiocytes
endothelial
erthyroid
hepatocyte-CXCL1
hepatocyte-CYP2A13
hepatocyte-GSTA1
hepatocyte-SCD
hepatocyte-TAT-AS1
kupffer
mature B
stellate
b
Overlap_4
Overlap_2
Overlap_1
Overlap_0
pancreas_celseq
pancreas_smartseq2
acinar
alpha
beta
delta
ductal
gamma

Fig 3

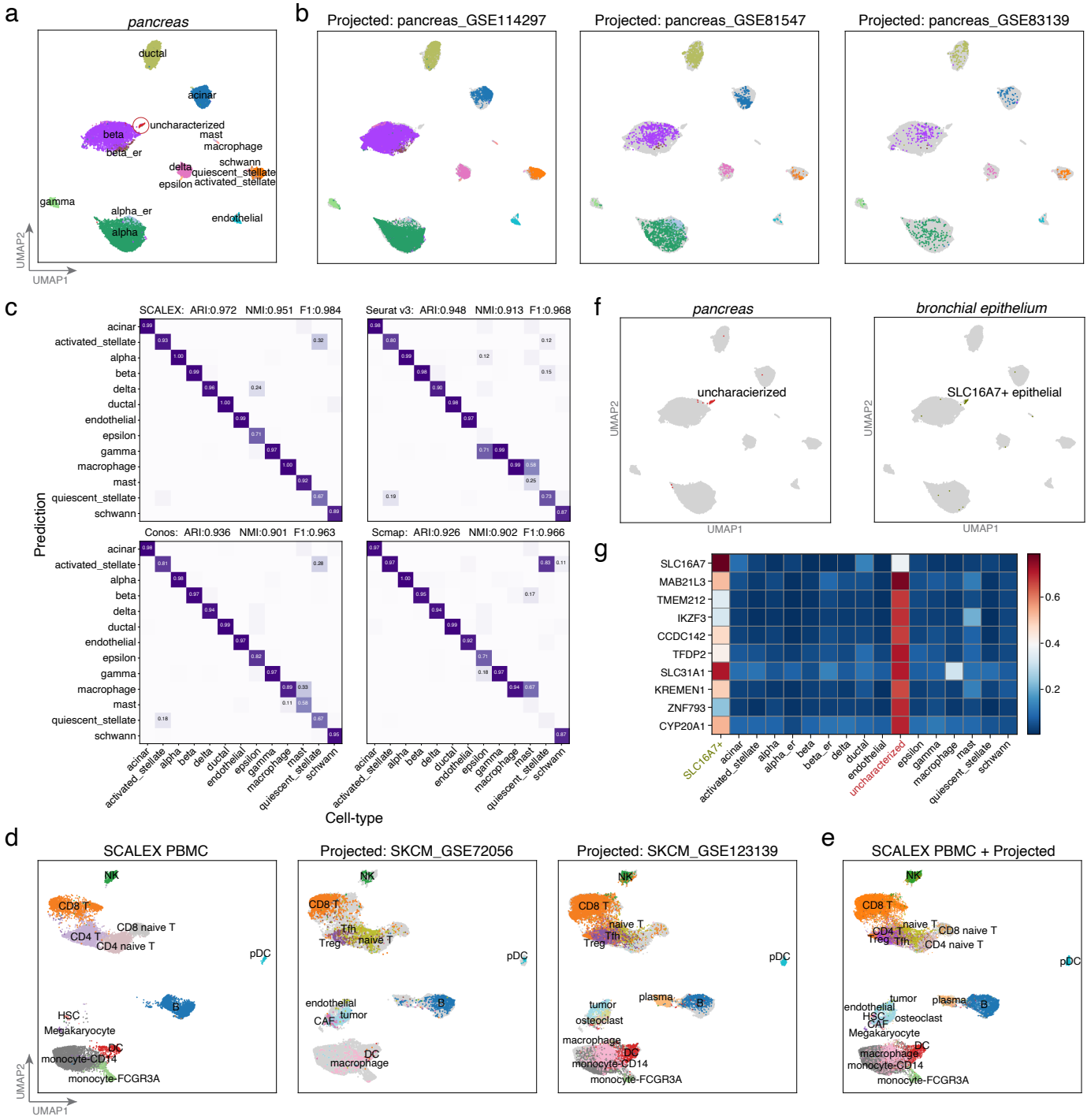

a
pancreas
ductal
acinar
uncharacterized
beta
mast
macrophage
beta_er
schwann
delta
quiescent_stellate
epsilon
activated_stellate
gamma
alpha_er
endothelial
alpha
UMAP2
UMAP1
b
Projected: pancreas_GSE114297
Projected: pancreas_GSE81547
Projected: pancreas_GSE83139
c
SCALEX: ARI:0.972 NMI:0.951 F1:0.984
Seurat v3: ARI:0.948 NMI:0.913 F1:0.968
Conos: ARI:0.936 NMI:0.901 F1:0.963
Scmap: ARI:0.926 NMI:0.902 F1:0.966
Prediction
Cell-type
acinar
activated_stellate
alpha
beta
delta
ductal
endothelial
epsilon
gamma
macrophage
mast
quiescent_stellate
schwann
f
pancreas
bronchial epithelium
uncharacierized
SLC16A7+ epithelial
UMAP2
UMAP1
g
SLC16A7
MAB21L3
TMEM212
IKZF3
CCDC142
TFDP2
SLC31A1
KREMEN1
ZNF793
CYP20A1
SLC16A7+
uncharacterized
0.6
0.4
0.2
d
SCALEX PBMC
Projected: SKCM_GSE72056
Projected: SKCM_GSE123139
NK
CD8 T
CD8 naive T
CD4 T
CD4 naive T
pDC
B
HSC
Megakaryocyte
DC
monocyte-CD14
monocyte-FCGR3A
Tfh
Treg
naive T
endothelial
CAF
tumor
macrophage
plasma
osteoclast
e
SCALEX PBMC + Projected

Fig 4

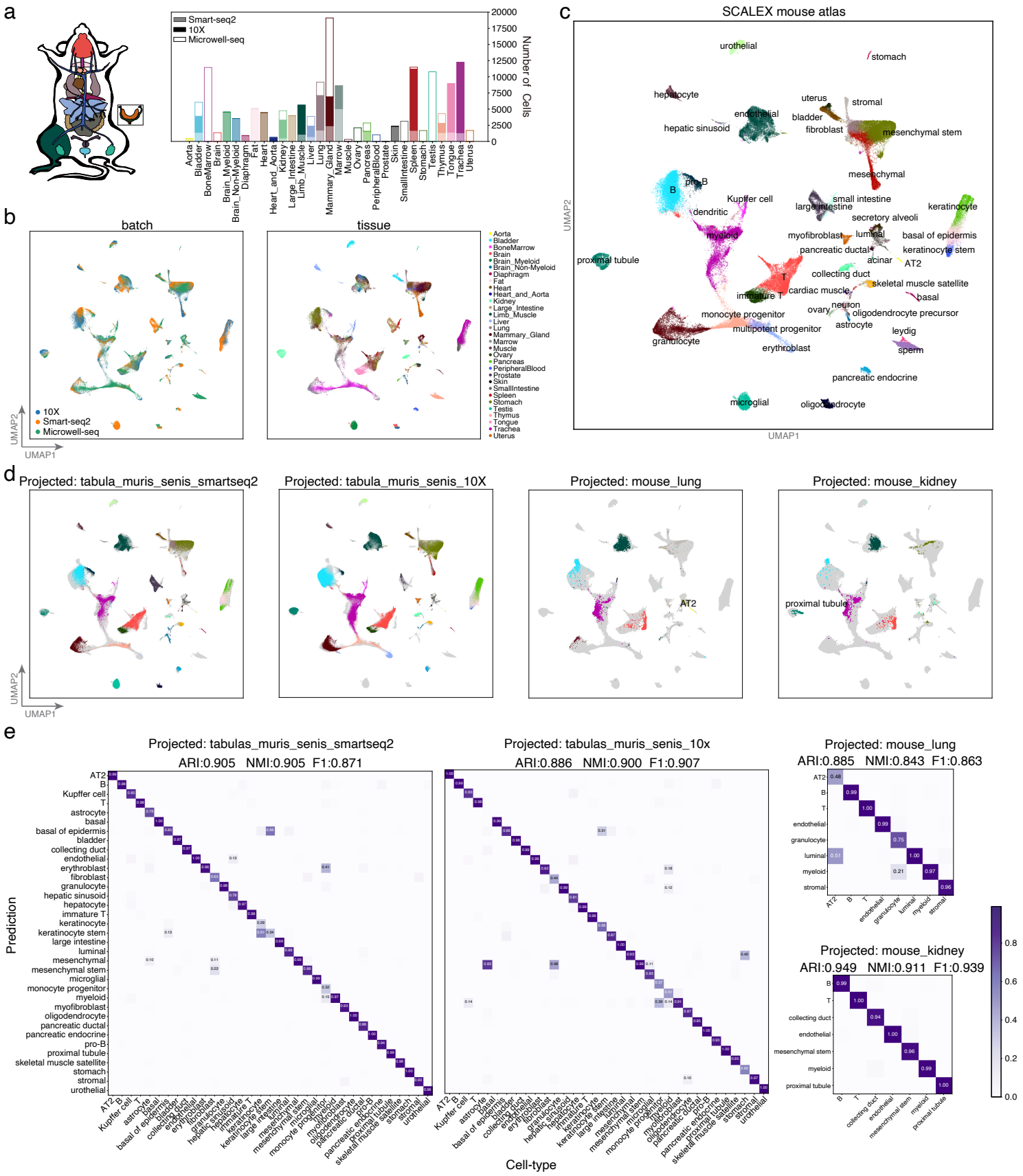
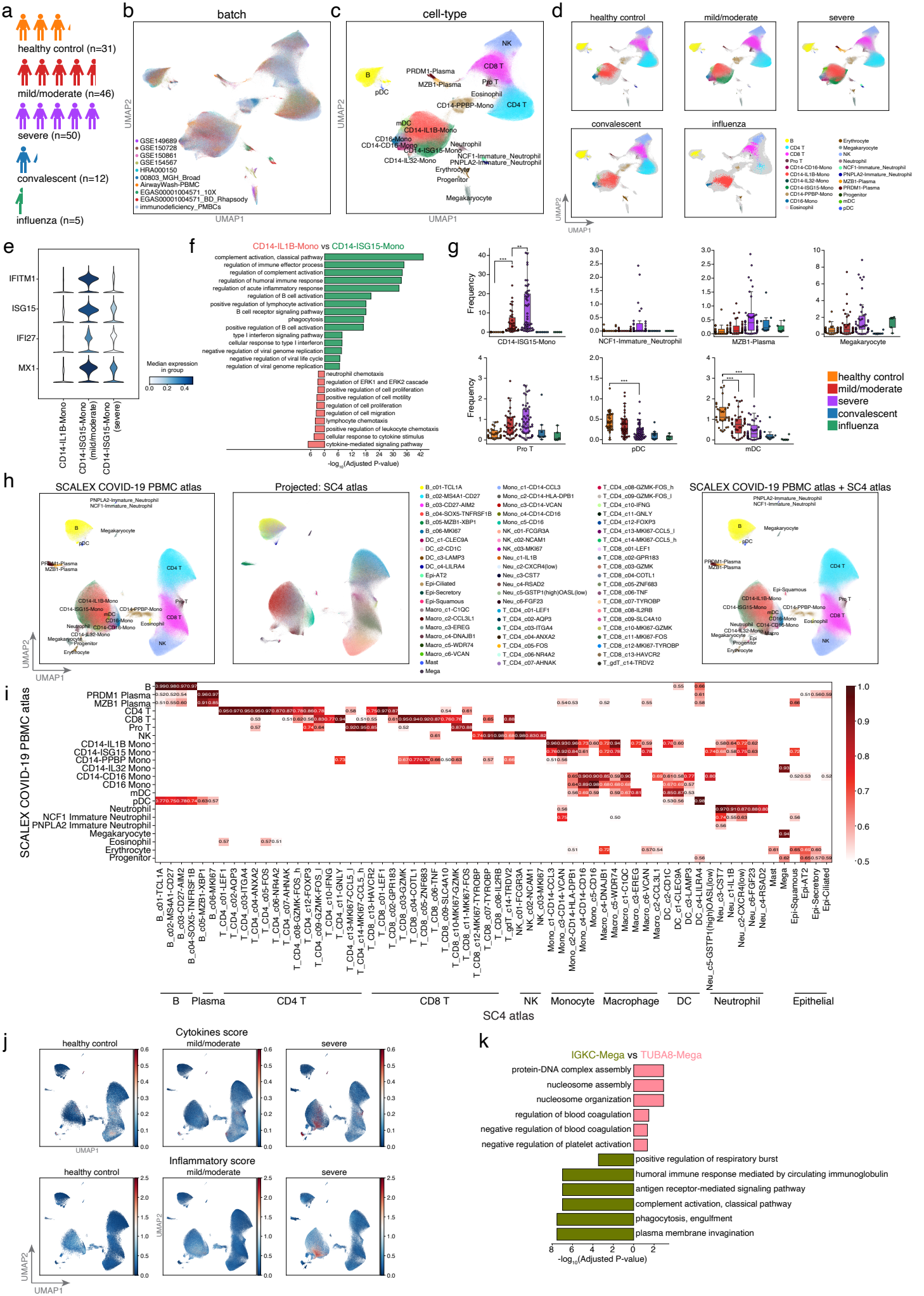
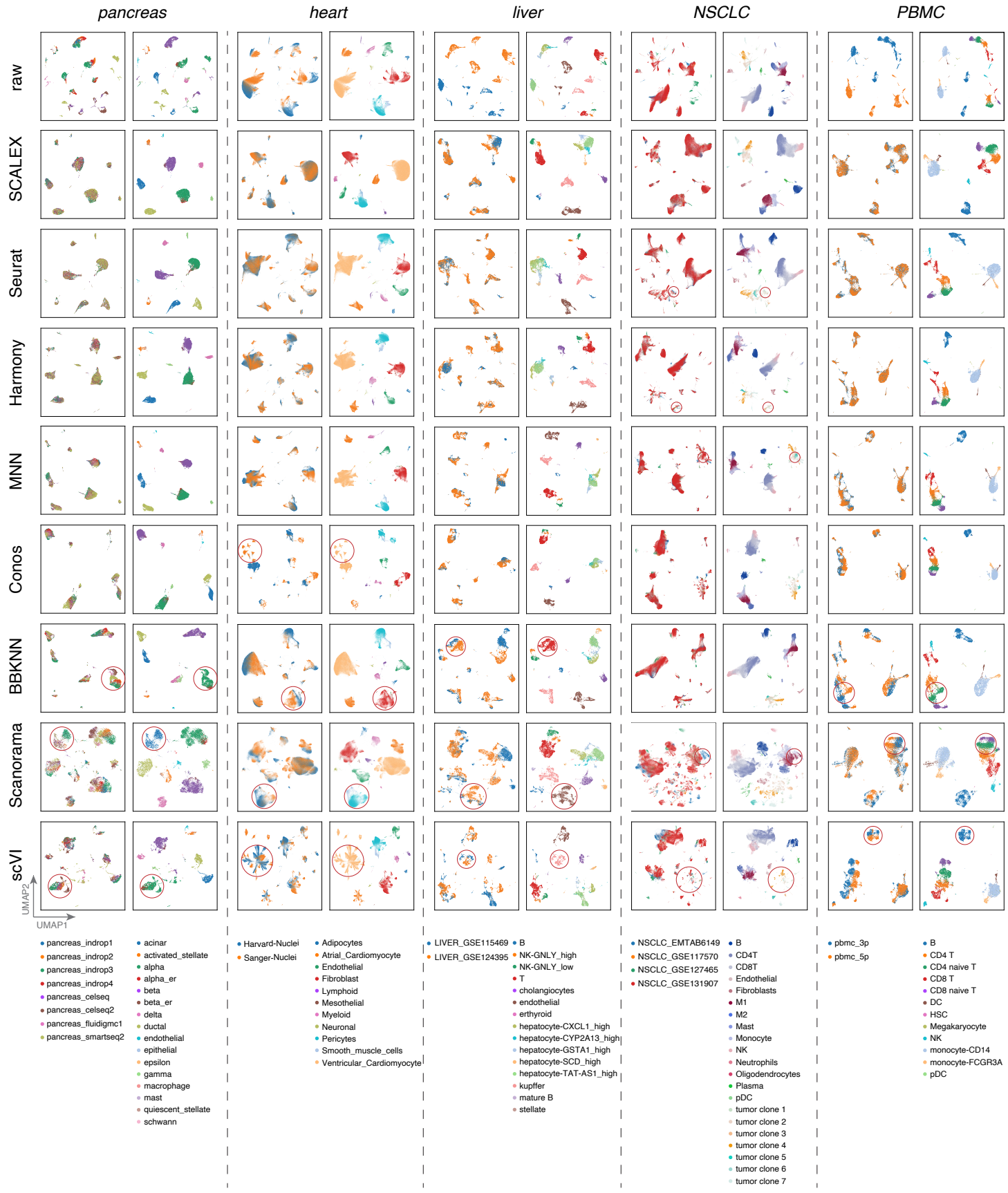

a
Smart-seq2
10X
Microwell-seq
Number of Cells
Aorta
Bladder
BoneMarrow
Brain
Brain_Myeloid
Brain_Non-Myeloid
Diaphragm
Fat
Heart
Heart_and_Aorta
Kidney
Large_Intestine
Limb_Muscle
Liver
Lung
Mammary_Gland
Marrow
Muscle
Ovary
Pancreas
PeripheralBlood
Prostate
Skin
SmallIntestine
Spleen
Stomach
Testis
Thymus
Tongue
Trachea
Uterus
b
batch
tissue
10X
Smart-seq2
Microwell-seq
UMAP1
UMAP2
c
SCALEX mouse atlas
urothelial
stomach
hepatocyte
uterus
stromal
endothelial
bladder
hepatic sinusoid
fibroblast
mesenchymal stem
mesenchymal
pro-B
B
Kupffer cell
small intestine
large intestine
keratinocyte
dendritic
secretory alveoli
myeloid
myofibroblast
luminal
basal of epidermis
pancreatic ductal
keratinocyte stem
proximal tubule
acinar
AT2
T
collecting duct
skeletal muscle satellite
cardiac muscle
immature T
basal
ovary
neuron
oligodendrocyte precursor
monocyte progenitor
astrocyte
multipotent progenitor
leydig
granulocyte
erythroblast
sperm
pancreatic endocrine
microglial
oligodendrocyte
UMAP1
UMAP2
d
Projected: tabula_muris_senis_smartseq2
Projected: tabula_muris_senis_10X
Projected: mouse_lung
Projected: mouse_kidney
AT2
proximal tubule
UMAP1
UMAP2
e
Projected: tabulas_muris_senis_smartseq2
ARI:0.905 NMI:0.905 F1:0.871
Projected: tabulas_muris_senis_10x
ARI:0.886 NMI:0.900 F1:0.907
Projected: mouse_lung
ARI:0.885 NMI:0.843 F1:0.863
Projected: mouse_kidney
ARI:0.949 NMI:0.911 F1:0.939
Prediction
Cell-type

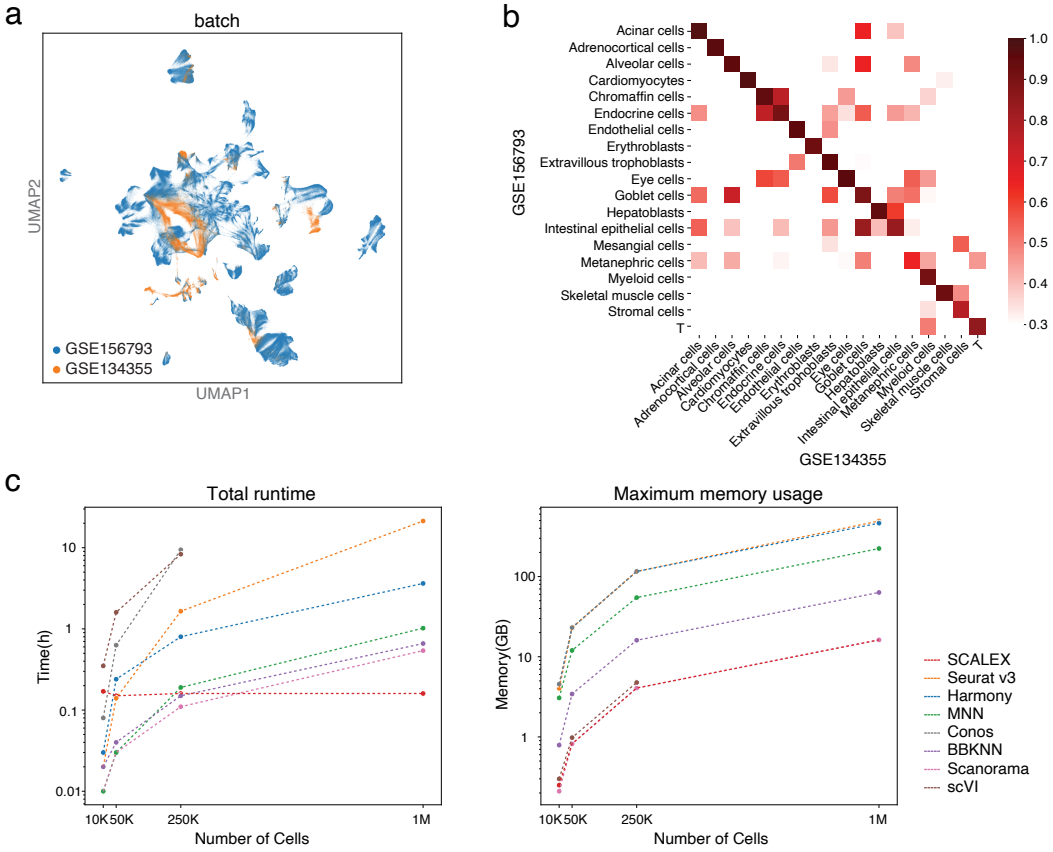
Fig 5


a
healthy control (n=31)
mild/moderate (n=46)
severe (n=50)
convalescent (n=12)
influenza (n=5)
b
batch
GSE149689
GSE150728
GSE150861
GSE154567
HRA000150
00803_MGH_Broad
AirwayWash-PBMC
EGAS00001004571_10X
EGAS00001004571_BD_Rhapsody
immunodeficiency_PMBCs
UMAP1
UMAP2
c
cell-type
NK
CD8 T
Pro T
CD4 T
B
pDC
PRDM1-Plasma
MZB1-Plasma
Eosinophil
CD14-PPBP-Mono
mDC
CD14-IL1B-Mono
CD16-Mono
CD14-CD16-Mono
CD14-ISG15-Mono
Neutrophil
CD14-IL32-Mono
NCF1-Immature_Neutrophil
PNPLA2-Immature_Neutrophil
Erythrocyte
Progenitor
Megakaryocyte
d
healthy control
mild/moderate
severe
convalescent
influenza
e
IFITM1
ISG15
IFI27
MX1
Median expression in group
CD14-IL1B-Mono
CD14-ISG15-Mono (mild/moderate)
CD14-ISG15-Mono (severe)
f
CD14-IL1B-Mono vs CD14-ISG15-Mono
complement activation, classical pathway
regulation of immune effector process
regulation of complement activation
regulation of humoral immune response
regulation of acute inflammatory response
regulation of B cell activation
positive regulation of lymphocyte activation
B cell receptor signaling pathway
phagocytosis
positive regulation of B cell activation
type I interferon signaling pathway
cellular response to type I interferon
negative regulation of viral genome replication
negative regulation of viral life cycle
regulation of viral genome replication
neutrophil chemotaxis
regulation of ERK1 and ERK2 cascade
positive regulation of cell proliferation
positive regulation of cell motility
regulation of cell proliferation
regulation of cell migration
lymphocyte chemotaxis
positive regulation of leukocyte chemotaxis
cellular response to cytokine stimulus
cytokine-mediated signaling pathway
-log10(Adjusted P-value)
g
Frequency
CD14-ISG15-Mono
NCF1-Immature_Neutrophil
MZB1-Plasma
Megakaryocyte
Pro T
pDC
mDC
healthy control
mild/moderate
severe
convalescent
influenza
h
SCALEX COVID-19 PBMC atlas
Projected: SC4 atlas
SCALEX COVID-19 PBMC atlas + SC4 atlas
UMAP1
UMAP2
i
SCALEX COVID-19 PBMC atlas
B
PRDM1 Plasma
MZB1 Plasma
CD4 T
CD8 T
Pro T
NK
CD14-IL1B Mono
CD14-ISG15 Mono
CD14-PPBP Mono
CD14-IL32 Mono
CD14-CD16 Mono
CD16 Mono
mDC
pDC
Neutrophil
NCF1 Immature Neutrophil
PNPLA2 Immature Neutrophil
Megakaryocyte
Eosinophil
Erythrocyte
Progenitor
B
Plasma
CD4 T
CD8 T
NK
Monocyte
Macrophage
DC
Neutrophil
Epithelial
SC4 atlas
j
Cytokines score
healthy control
mild/moderate
severe
Inflammatory score
healthy control
mild/moderate
severe
UMAP1
UMAP2
k
IGKC-Mega vs TUBA8-Mega
protein-DNA complex assembly
nucleosome assembly
nucleosome organization
regulation of blood coagulation
negative regulation of blood coagulation
negative regulation of platelet activation
positive regulation of respiratory burst
humoral immune response mediated by circulating immunoglobulin
antigen receptor-mediated signaling pathway
complement activation, classical pathway
phagocytosis, engulfment
plasma membrane invagination
-log10(Adjusted P-value)

Supplementary Fig 1

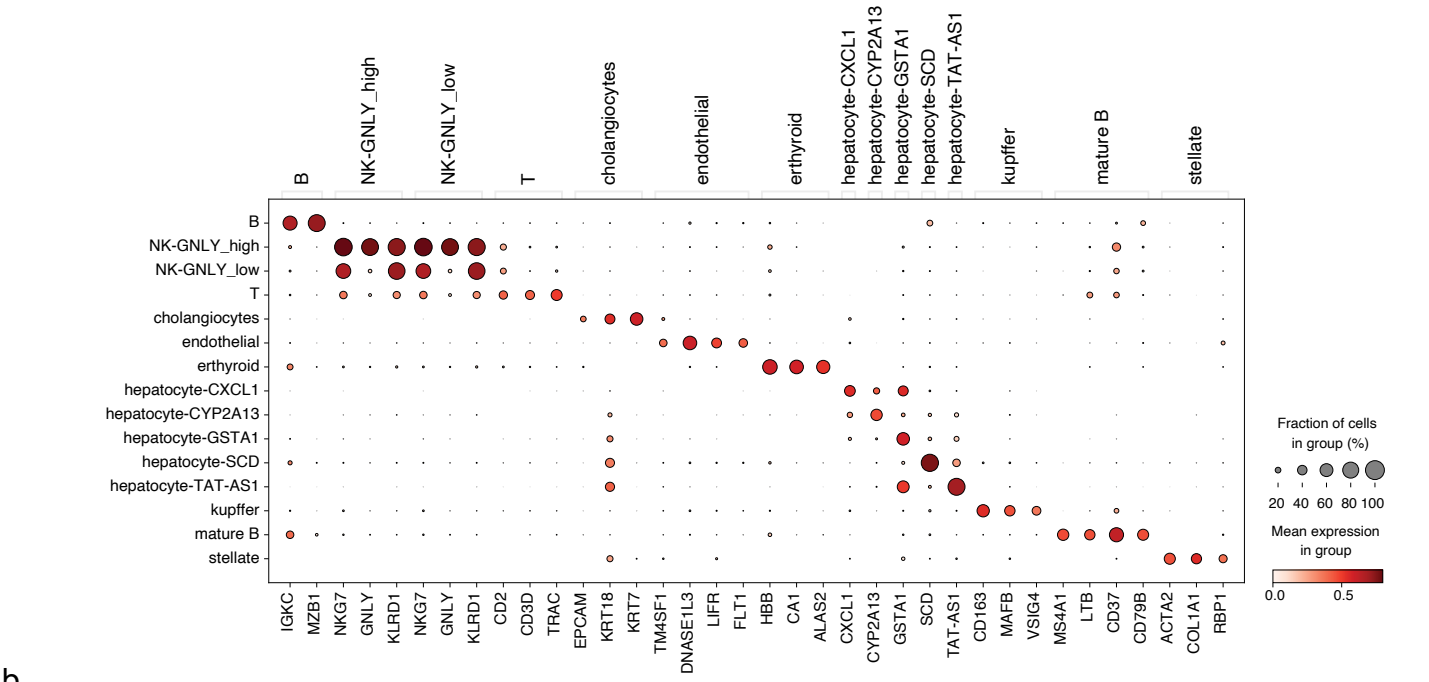


Supplementary Fig 2

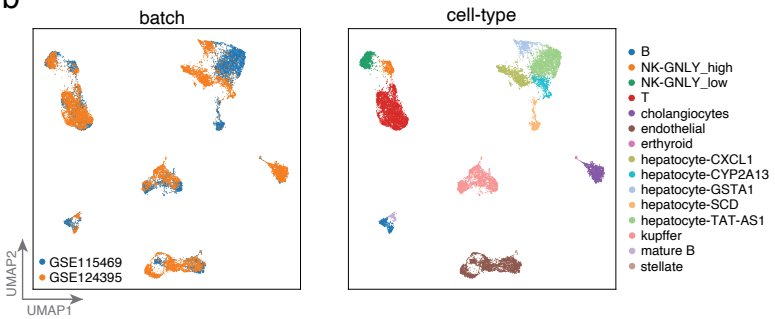


Supplementary Fig 3

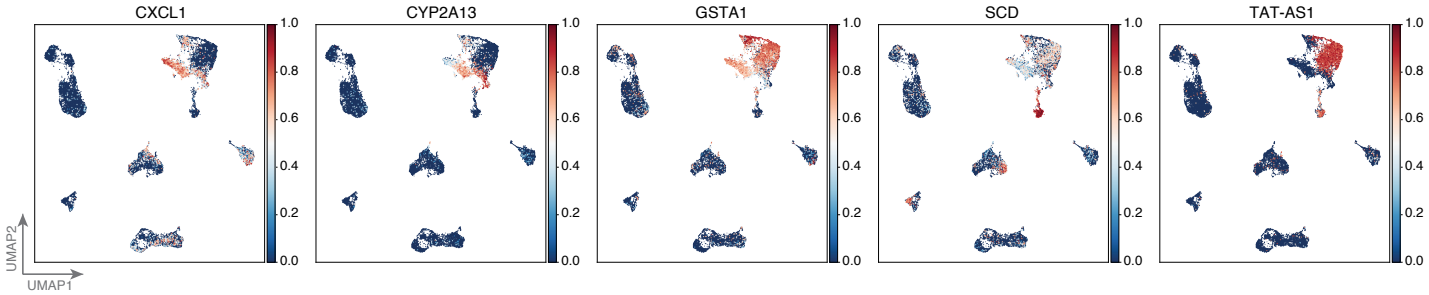
a



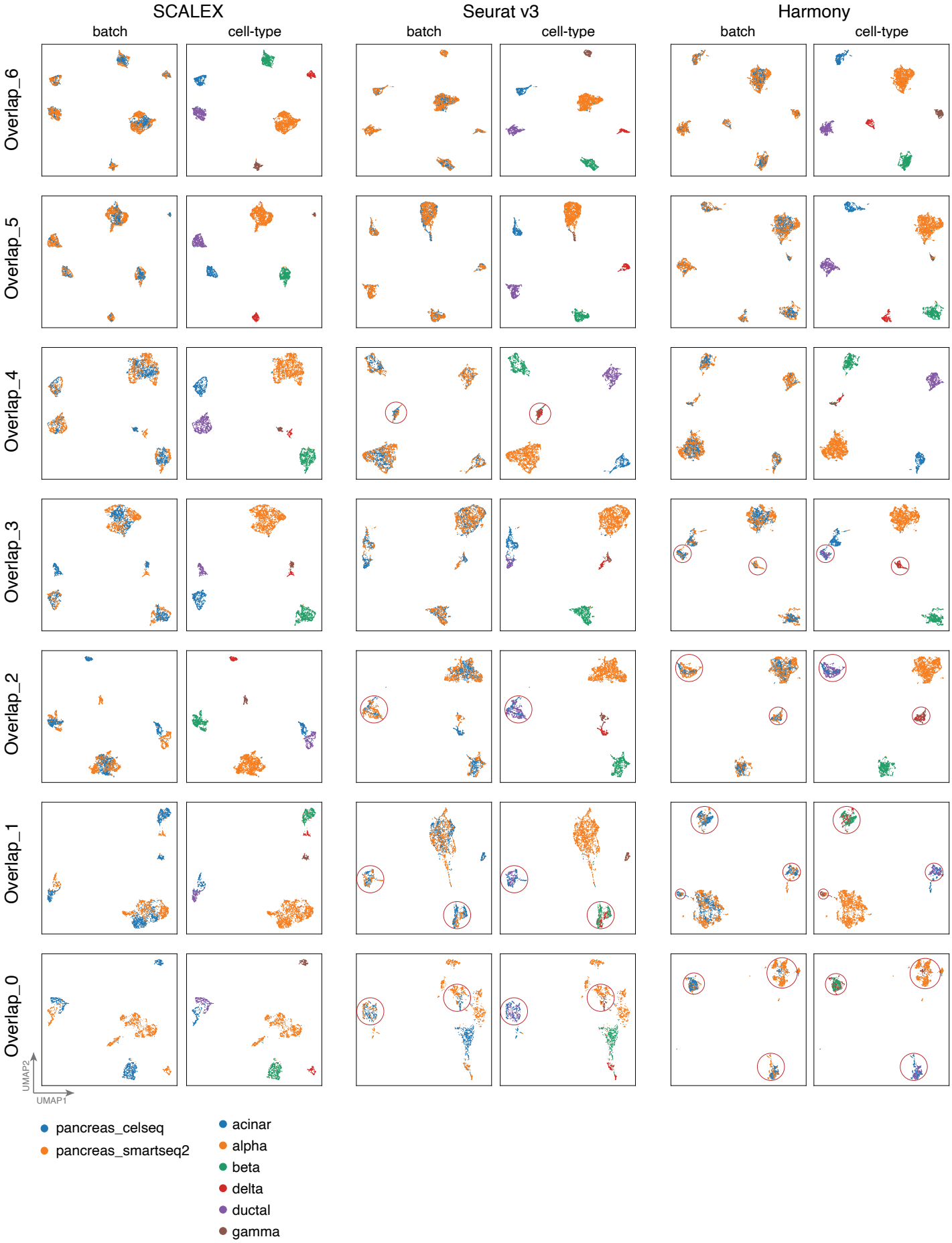
b



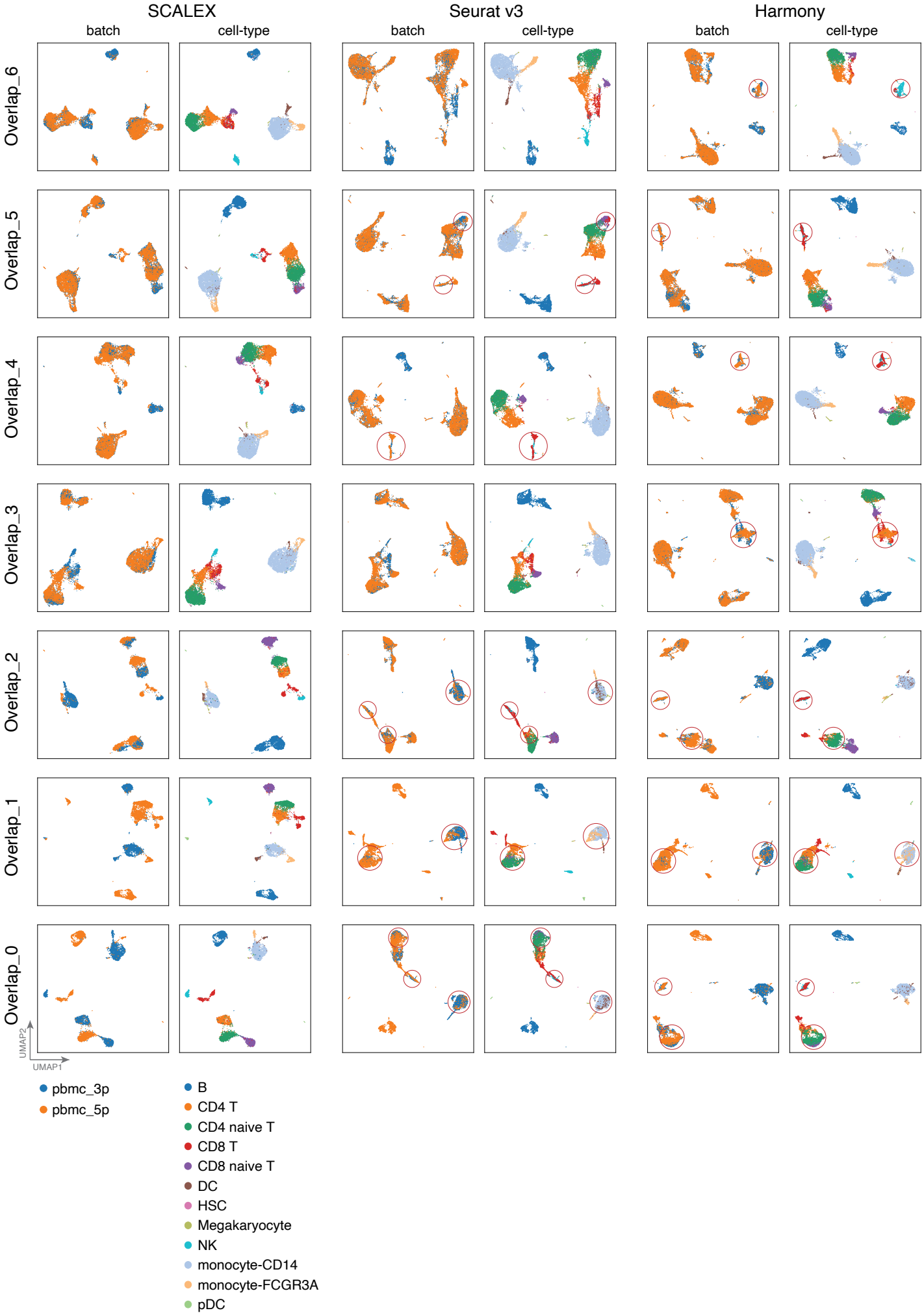
c



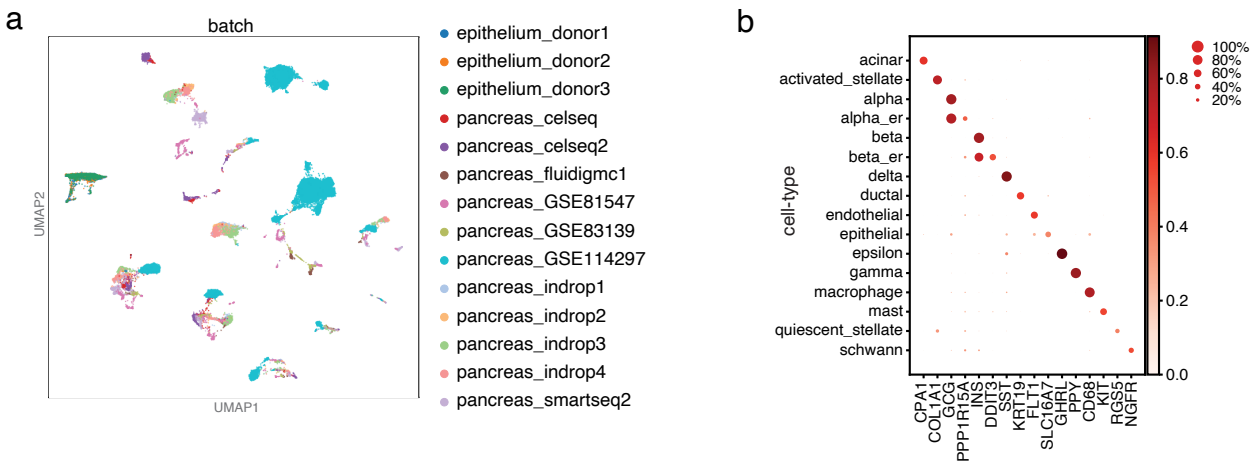
Supplementary Fig 4



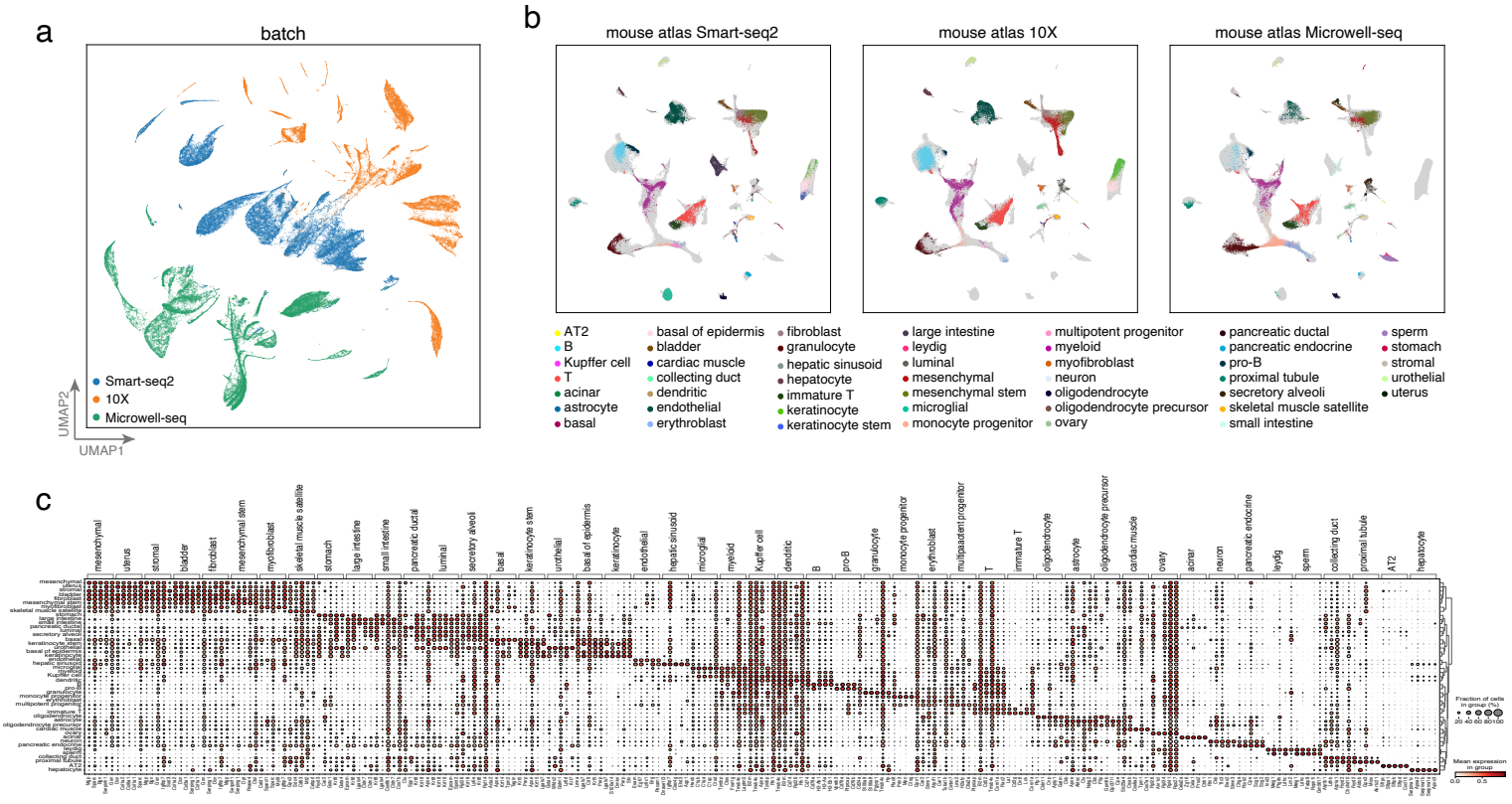
Supplementary Fig 5



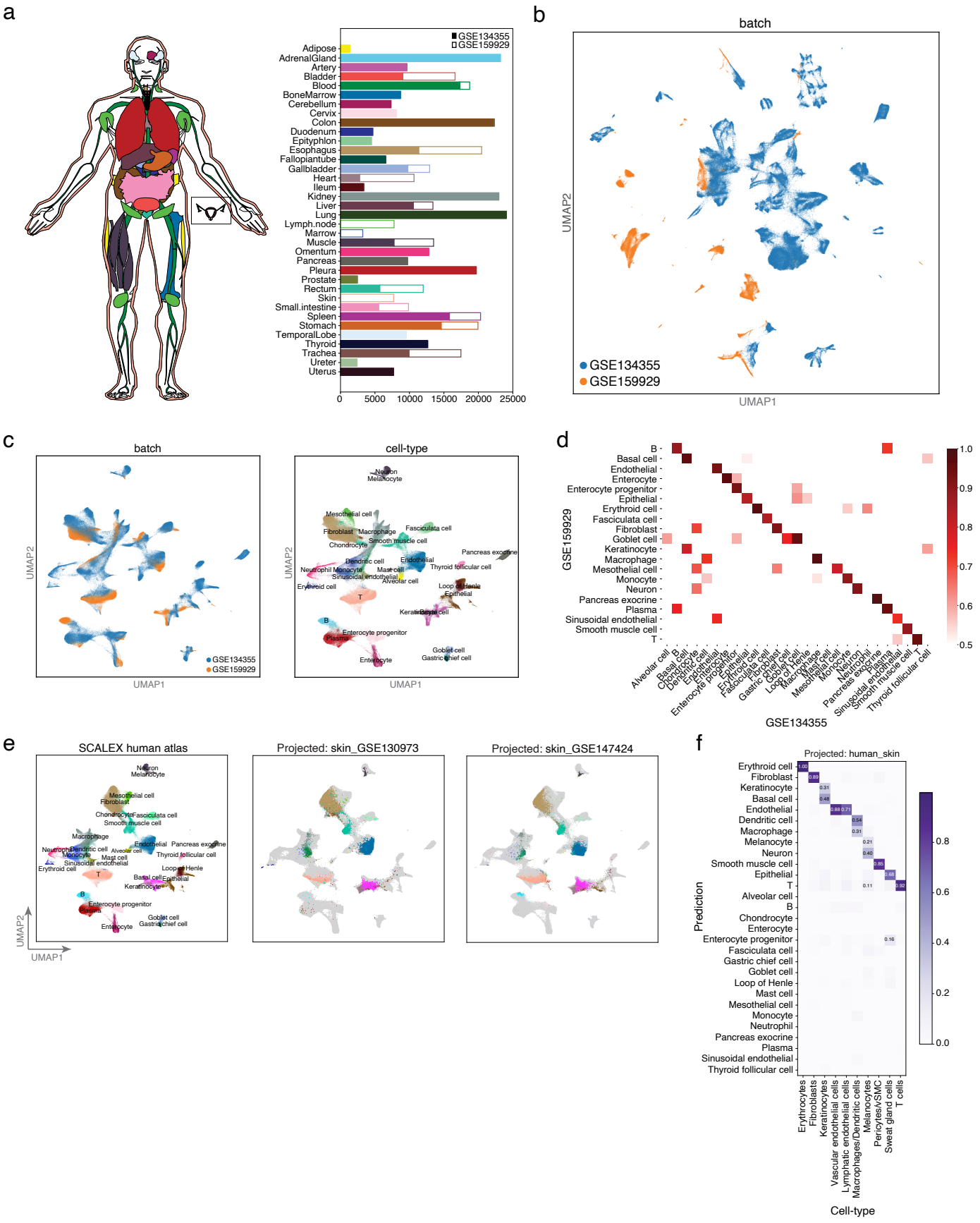
Supplementary Fig 6



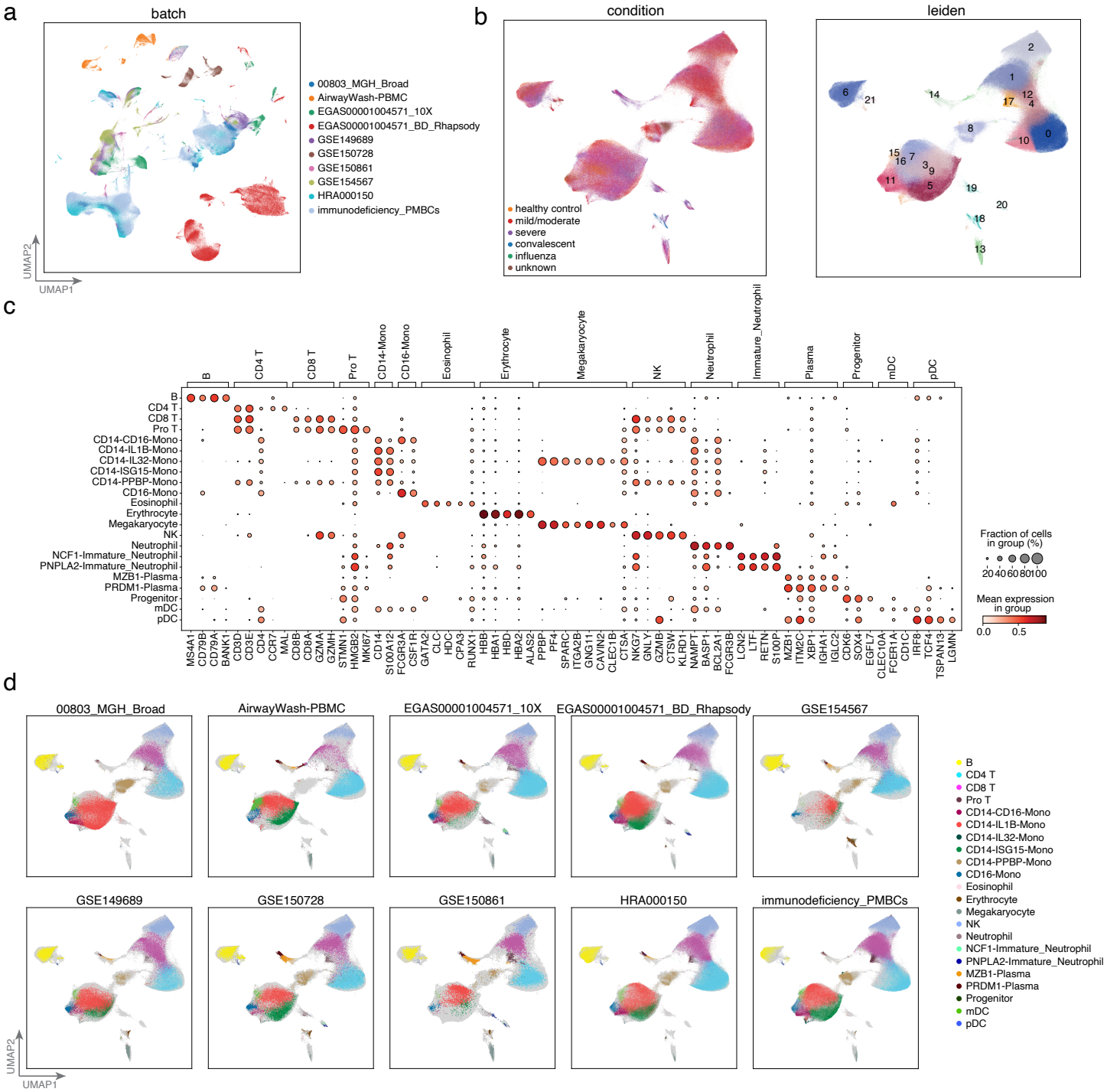
Supplementary Fig 7



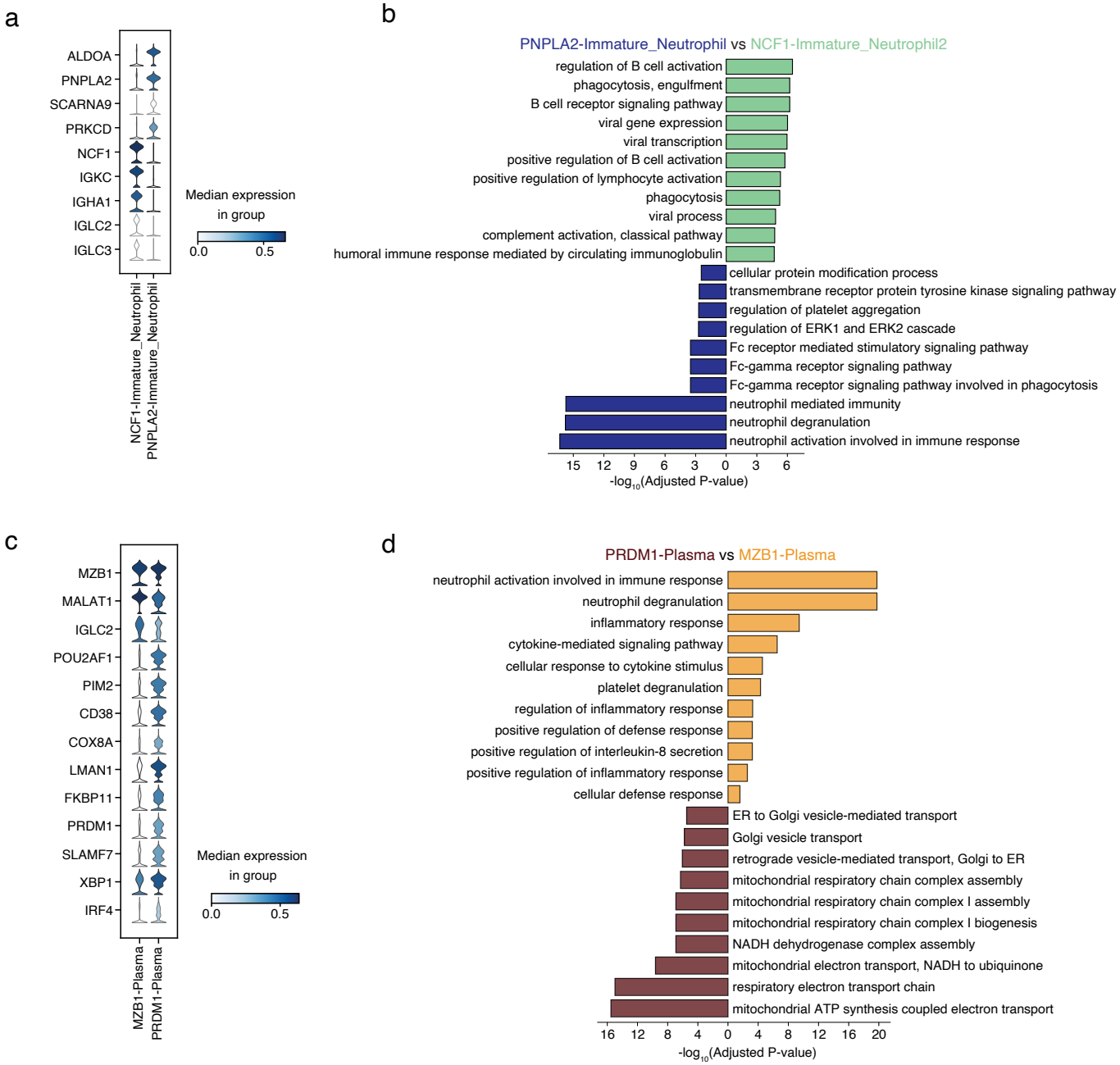
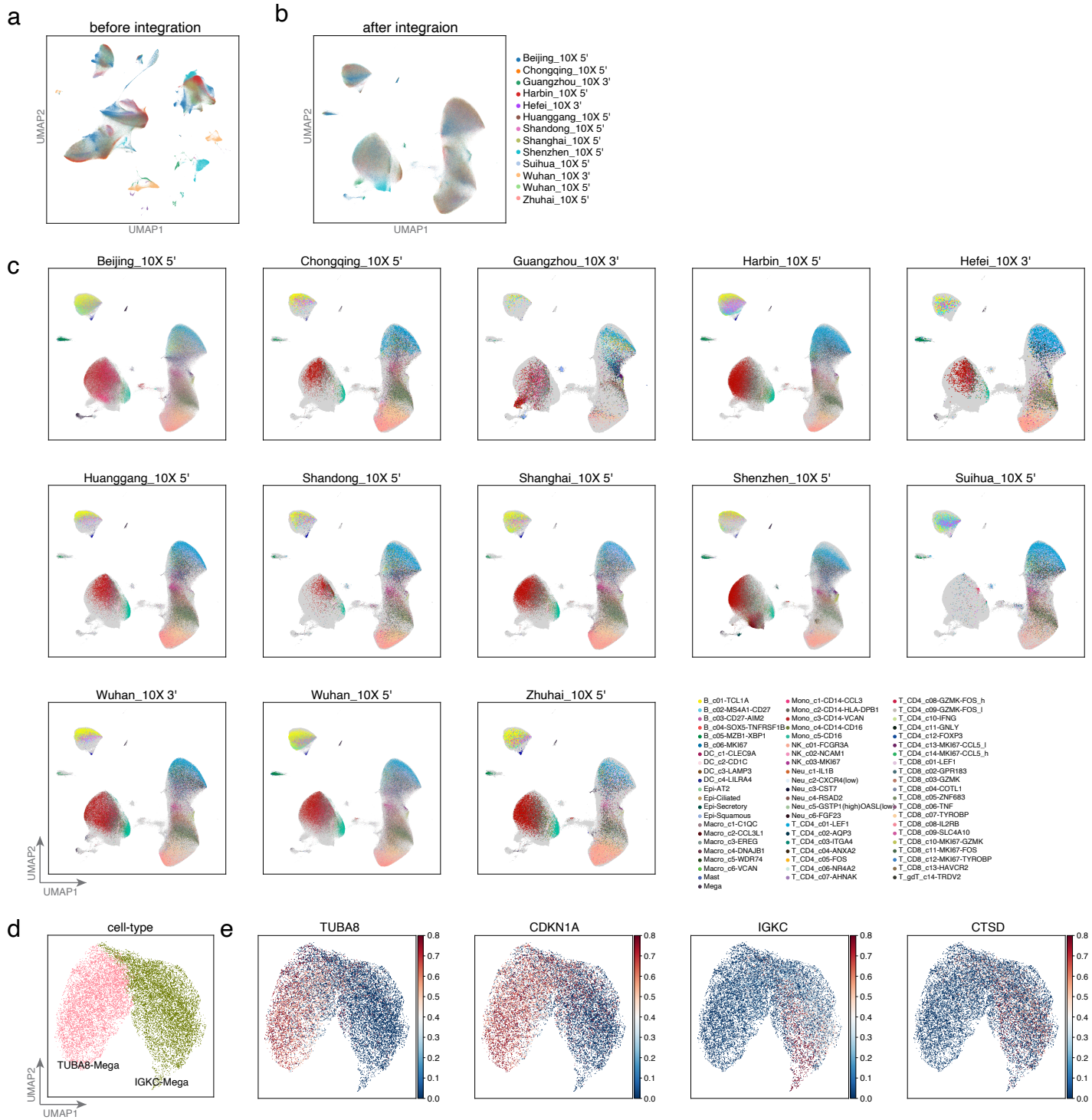
Supplementary Fig 8



Supplementary Fig 9



Supplementary Fig 10



Supplementary Fig 11