

# Atlas-scale single-cell multi-sample multi-condition data integration using scMerge2

Yingxin Lin<sup>1,2,3,4</sup>, Yue Cao<sup>1,2,3,4</sup>, Elijah Willie<sup>1</sup>, Ellis Patrick<sup>1,3,4,5</sup>, and Jean Y.H.  
Yang<sup>1,2,3,4\*</sup>

<sup>1</sup>Sydney Precision Data Science Centre, The University of Sydney, NSW, Australia.

<sup>2</sup>Charles Perkins Centre, The University of Sydney, NSW, Australia.

<sup>3</sup>School of Mathematics and Statistics, The University of Sydney, NSW, Australia.

<sup>4</sup>Laboratory of Data Discovery for Health Limited (D24H), Science Park, Hong Kong SAR,  
China

<sup>5</sup>The Westmead Institute for Medical Research, The University of Sydney, NSW 2006,  
Australia.

\*To whom correspondence should be addressed. Email: Jean Y.H.Yang,  
jean.yang@sydney.edu.au.

## Abstract

The recent emergence of multi-sample multi-condition single-cell multi-cohort studies allow researchers to investigate different cell states. The effective integration of multiple large-cohort studies promises biological insights into cells under different conditions that individual studies cannot provide. Here, we present scMerge2, a scalable algorithm that allows data integration of atlas-scale multi-sample multi-condition single-cell studies. We have generalised scMerge2 to enable the merging of millions of cells from single-cell studies generated by various single-cell technologies. Using a large COVID-19 data collection with over five million cells from 1000+ individuals, we demonstrate that scMerge2 enables multi-sample multi-condition scRNA-seq data integration from multiple cohorts and reveals signatures derived from cell-type expression that are more accurate in discriminating disease progression. Further, we demonstrate that scMerge2

25 can remove dataset variability in CyTOF, imaging mass cytometry and CITE-seq experiments,  
26 demonstrating its applicability to a broad spectrum of single-cell profiling technologies.

## 27 **Introduction**

28 Technological advances of large-scale single-cell profiling of genes and proteins, such as single-  
29 cell RNA-seq (scRNA-seq) [1], Cytometry by Time-Of-Flight (CyTOF) [2] and imaging mass  
30 cytometry [3] have exploded in recent years and enabled unprecedented insight into the identity  
31 and function of individual cells. This has enabled the discovery of cell-type-specific knowledge  
32 and has transformed our understanding of biological systems. This myriad of single-cell data  
33 has prompted the recent creation of data atlases that collate single-cell omics data from multi-  
34 ple studies. Examples of large-scale atlases containing over two millions cells are the Human  
35 Cell Atlas which aims to map every cell type in the human body [4]; atlas of gene expression  
36 and chromatin accessibility of 4 million human fetal cells across 15 organs [5, 6]; the Human  
37 Tumor Atlas Network [7] and DISCO [8], which provides integrated human single-cell omics  
38 data across 107 tissues/cell lines/organoids and 158 diseases. These atlases serve as valuable  
39 references for the exploration of healthy and diseased cells.

40  
41 As single-cell technologies advance, there are an increasing number of studies around the  
42 globe that perform multi-condition and multi-sample large-cohort single-cell profiling to exam-  
43 ine persisting questions associated with human health. These datasets enable researchers to delve  
44 into biological insights of cells under multiple treatment conditions across multiple individuals.  
45 For example, to investigate the cell-type-specific cellular mechanism underlying COVID-19 dis-  
46 ease severity [9] and to predict treatment response to cancer [10]. Such data and studies are  
47 expected to rise in the coming years [11] in the continuing quest to improve human health. This  
48 expected increase necessitates the effective access and joint interpretation of multiple datasets to  
49 unleash the power of meta-analysis at single-cell resolution.

50  
51 Last year, benchmarking studies [12] began to investigate atlas-scale integration. Luecken  
52 and colleagues investigated 16 popular data integration technologies on 13 data integration tasks  
53 with up to 1 million cells. While significant progress has been achieved in batch correction and  
54 data integration over the years (including our research), the increasing scale of cohort sizes and  
55 the number of related studies for integration has introduced additional scalability challenges. The

56 new challenge for atlas-scale integration is to have a scalable algorithm that can handle a large  
57 number of studies, consisting a large collection samples (thousands) and millions of cells. With  
58 the exception of Seurat [13], SAUCIE [14] and Scanorama [15], several of these rapid procedures  
59 (deepMNN [16], BBKNN [17], Harmony [18], scVI [19], scANVI [20] and DESC [21]) focus  
60 on extracting the joint embedding and do not return adjusted gene expression matrices. With the  
61 growing need for sample level analysis, the lack of adjusted expression matrices restricts the util-  
62 isation of such integrative results and diminishes their potency and generalizability. As a result,  
63 the next generation of atlas-scale integration algorithms should be capable of integrating a large  
64 number of studies and producing consensus cell type maps as well as adjusted expression matrix  
65 for further downstream analysis. In particular, these methods need to overcome the computa-  
66 tional challenge of integrating over a million cells and create adjusted gene expression matrix for  
67 all genes for downstream analysis.

68

69 To this end, we present scMerge2, a scalable, high-capacity algorithm that allows data inte-  
70 gration of atlas-scale multi-sample multi-condition single-cell studies. We achieve this through  
71 three key innovations in (i) hierarchical integration to capture both local and global variation be-  
72 tween studies; (ii) pseudo-bulk construction to ensure computational scalability; and (iii) pseudo-  
73 replication inside each condition to capture signals from multiple conditions. Our new scMerge2  
74 algorithm is able to integrate many millions of cells from single-cell studies generated from  
75 various single-cell technologies, including scRNA-seq, CyTOF, and imaging mass cytometry.  
76 Leveraging pseudo-bulk to perform factor analysis of stably expressed genes and pseudorepli-  
77 cates, scMerge2 is able to integrate five million cells from a large COVID-19 data collection with  
78 over 1000 samples from 20 studies globally within a day. We further demonstrate that the integra-  
79 tion using scMerge2 improves the performance of discriminating distinct cell states in COVID-19  
80 patients with varying degrees of severity and facilitates diverse single-cell downstream analyses.

## 81 **Results**

### 82 **scMerge2 effectively integrates single-cell multi-sample, multi-condition data.**

83 scMerge2 provides a scalable data integration method for the rapid growth of multi-sample,  
84 multi-condition single-cell studies. This new extension of scMerge is specifically designed to  
85 address unwanted intra- and inter-dataset variation that can overshadow true biological signals

86 between conditions. In our previous study, we introduced scMerge, a novel algorithm that inte-  
87 grates multiple single-cell RNA-seq data by factor analysis of stably expressed genes and pseudo-  
88 replicates across datasets and enhances biological discovery, including inferring cell development  
89 trajectories [22]. The integration approach supports diverse integration settings, enabling cross-  
90 batch, cross-dataset, and cross-species discoveries. In particular, the semi-supervised aspect of  
91 scMerge allows incorporation of prior knowledge facilitated by experimental design.

92 With the rapid emergence of multi-sample multi-condition single-cell studies and the in-  
93 creased number of datasets for integration, our proposed scMerge2 addresses challenges asso-  
94 ciated with scalability of cells and studies as well as producing analytically ready data (i.e. ad-  
95 justed expression matrix). This is achieved via three key innovations as illustrated in **Fig. 1**.  
96 First, hierarchical integration is used to capture both local and global variation. This is a clear  
97 contrast to the conventional data integration that involves estimating unwanted variation across  
98 all datasets as a whole. When integrating across a large collection (over 10) of datasets with  
99 different pairwise differences, sequential integration better captures the difference in pairwise  
100 variations. Second, pseudo-bulk construction is used to reduce computing load, allowing for the  
101 analysis of datasets containing millions of cells. Third, pseudo-replication inside each condition  
102 is built, allowing for the modelling of numerous conditions. Details of these components are  
103 included in Methods. In essence, scMerge2 takes gene expression matrices from a collection  
104 of datasets and integrates them in a hierarchical manner. The final output of scMerge2 is a sin-  
105 gle adjusted expression matrix with all input data matrices merged and ready for downstream  
106 analysis.

## 107 **scMerge2 outperforms existing integration methods in detecting differential** 108 **expression.**

109 We demonstrate the performance of scMerge2 in removing multi-level unwanted variation of  
110 multiple scRNA-seq datasets from three aspects. Firstly, to illustrate the effectiveness of the hier-  
111 archical integration strategy, we applied scMerge2 to a 200k subset of cells from two COVID-19  
112 studies (Liu and Stephenson) that contain three cohorts/batches within each dataset. We com-  
113 pared the performance of two different scMerge2 settings: scMerge2-h, where we performed  
114 intra-study correction before inter-study correction; and scMerge2, where we integrated two  
115 datasets (6 batches) in one go. We find that integrating the two studies in a hierarchical man-  
116 ner improves the performance of data integration, especially in terms of revealing the cell type

117 signals (**Fig. 2a-b**). Compared to the other data integration methods (Seurat, SeuratRPCA,  
118 fastMNN, Liger and Harmony), both settings of scMerge2 (scMerge2-h and scMerge2) have  
119 overall better performance in achieving the balance of batch effect removal and biological signal  
120 preservation, based on the five evaluation metrics that quantify the data integration performance  
121 (**Fig. 2a-b**).

122

123 Next, we investigate the performance of the adjusted matrix in identifying genes that are dif-  
124 ferentially expressed between two conditions (termed as differential state (DS) analysis by [23])  
125 through a simulation study. We generated synthetic single-cell datasets with two batches and  
126 multiple samples from two conditions using a simulation framework that extended from scDe-  
127 sign3 model [24], with known ground truth DS genes (**Supp Fig. S1-S2**) (See Methods). Cell-  
128 type-specific DS analysis was performed using the limma-trend algorithm [25] on the sample-  
129 wise aggregated data by taking the mean of the log-transformed or adjusted data. By simulating  
130 data with different log fold change ( $1.1 \sim 2$ ) and proportions of DS genes (5% and 10%), we  
131 find that scMerge2 substantially outperforms the other two data integration methods that also  
132 return adjusted matrices in detecting DS genes (**Fig. 2c** and **Supp Fig. S3**). scMerge2 has much  
133 lower FDR than fastMNN and Seurat, and higher TPR compared to the unadjusted data (**Supp**  
134 **Fig. S4-S5**), illustrating that scMerge2 outputs an adjusted matrix with less unwanted variation  
135 for single-cell downstream analysis.

136

137 Finally, we illustrate the robustness of scMerge2 by varying the key tuning parameters of the  
138 algorithm, including the number of unwanted variation factors, the number of pseudo-bulk, the  
139 ways of pseudo-bulk construction and the number of nearest neighbours. As shown in **Fig. 2d**  
140 and **Supp Fig. S6**, despite varying the settings in the algorithm, scMerge2 has consistently better  
141 performance than the other methods. Together, these results demonstrate the effectiveness and  
142 utility of scMerge2 in data integration of scRNA-seq data.

### 143 **scMerge2 is scalable to integrate five millions COVID-19 PMBC cells.**

144 To demonstrate the scalability of scMerge2 in integrating multi-sample multi-condition single-  
145 cell data, we performed scMerge2 on a COVID-19 data collection of consisting of  $\sim 5$ m cells  
146 from 1298 samples (963 individuals) PBMC samples from 20 studies worldwide (See Methods).  
147 We considered the cell type annotation refined by scClassify as pseudo-replicates information.

148 We also used a hierarchical integration strategy, where we first performed integration of different  
149 cohorts within one study respectively (e.g. Ren, Stephenson, Liu and Schulte-Schrepping) and  
150 also two studies with distinguished sequencing depth, followed by the integration of 13 studies  
151 with small number of cells (hierarchical integration strategy shown in **Supp Fig. S7**). We then in-  
152 tegrated all the data in the next step. An inspection of UMAP visualisations shows that scMerge2  
153 effectively integrates the 20 studies, while preserving the multi-level cell type information (**Fig.**  
154 **3a, Supp Fig. S8**). A UMAP plot faceted by dataset further illustrates the successful removal of  
155 dataset induced unwanted variation (**Supp Fig. S9**). The quantitative evaluation metrics further  
156 confirm this observation, where we find that scMerge2 reduces the technical variation caused by  
157 dataset, protocol and technology, resulting in improved cell type identification (**Fig. 3b, Supp**  
158 **Fig. S10**).

159

160 To further illustrate the utility of scMerge2, we demonstrate that it improves the prediction of  
161 disease severity in the COVID-19 dataset using cell-type-specific expression. Comparing to the  
162 original raw log-normalised data, identifying cell types with scMerge2 substantially improves the  
163 prediction accuracy rate of disease severity for all cell types that have more than 1% abundance  
164 in the data, with a 3.2% increase in accuracy on average (**Fig. 3c** and **Supp Fig. S11**). Notably,  
165 we find that CD14 Monocytes have the highest discriminative power for disease severity among  
166 all cell types, and scMerge2 is able to further improve the accuracy rate from 81.3% to 83.6%.

167

## 168 **scMerge2 enables differential cell state detection for multi-conditions data.**

169 We next illustrate how the adjusted expression matrix output from scMerge2 facilitates several  
170 downstream analysis of single-cell multi-condition multi-sample studies, including differential  
171 abundance analysis and differential expression analysis. As a case study, we focus on the analysis  
172 of identification and characterisation of cell states that are distinguished between the moderate  
173 and severe patients using COVID-19 data collection. We first calculated the differential abun-  
174 dance score for each cell to quantify the difference between the moderate and severe patients  
175 using DASEq [26]. As shown in **Fig. 4a-b**, we are able to identify regions on the UMAP plots  
176 that are associated with the disease severity. As expected, when mapping these regions to cell  
177 types, we find that neutrophils have the highest proportion of cells that are associated with se-  
178 vere disease outcome as their accumulation marks the critical illness of COVID-19 patients [27]

179 **(Supp Fig. S12).**

180

181 Next, we investigate the cell-type-specific underlying biological process pathways that are  
182 associated with the disease severity and time for each cell type. We performed the differential  
183 expression analysis on the cell-type specific pseudo-bulk by considering both disease severity and  
184 days from onset of symptoms as covariates, followed by gene set enrichment analysis (GSEA).  
185 The pathways enriched with disease severity include hallmark TNF $\alpha$  signaling and hallmark in-  
186 flammatory response (**Fig. 4c**) and are upregulated in severe patients in most of the cell types,  
187 while GO IL6 positive production and Hallmark MTORC1 signalling are upregulated in moder-  
188 ate patients. Notably, we observe that a few pathways reveal distinct enrichment patterns between  
189 different cell types, including GO response to type-I IFN. We find that for CD14 Monocytes (**Fig.**  
190 **4c-d**), the type-I IFN signatures is negatively associated disease severity and also decrease over  
191 time, consistent with the previous findings [28] (**Fig. 4d**). While other cell types such as CD4  
192 CM and CD4 Naive have an enrichment of type-I IFN in severe patients, this enrichment is also  
193 decreased over time. Together, these analysis demonstrate that the integration of multiple stud-  
194 ies using scMerge2 enables a variety of data analysis approaches that address a wide range of  
195 biological questions.

### 196 **scMerge2 is versatile to other single-cell platforms.**

197 One of the key strengths of scMerge2 is its generalizability to data from multiple biotechnology  
198 platforms. We illustrate that scMerge2 is generalizable to other single cell modalities including  
199 spatially resolved modality and multi-modalities. We start by illustrating that our algorithm is di-  
200 rectly applicable to other single-cell single-modal data, using two mass cytometry time-of-flight  
201 (CyTOF) datasets as an example. The two datasets (COMBAT (CyTOF) and Geanon (CyTOF))  
202 contain more than 11 million cells in total collected from healthy controls, COVID-19 and sepsis  
203 patients, with 18 immune cell populations and activation states. The UMAP plots constructed  
204 after integration (**Fig. 5a**) reveal that the two datasets are successfully integrated compared to  
205 the raw data. Notably, we find that Granulocytes (Neutrophils and Eosinophils), cell types that  
206 are only present in Geanon (CyTOF) but not COMBAT (CyTOF), are represented as a discrete  
207 and distinct cluster, suggesting that scMerge2 is able to reveal the unique cell types existing only  
208 in specific batches. An inspection of the cell-type-specific marker expression distribution further  
209 confirms the effective dataset effect removal (**Fig. 5b** and **Supp Fig. S13**).

210

211 Next, we show that scMerge2 enables normalisation of spatially resolved single-cell data for  
212 better cell type identification with specific cluster markers. We applied scMerge2 to a COVID-  
213 19 Imaging Mass Cytometry (IMC) dataset [29], followed by clustering using FlowSOM [30],  
214 with the number of clusters set equal to the manually annotated cell types in the original study.  
215 We find that compared to the original data, the scMerge2 adjusted matrix provides better clus-  
216 tering results that are more consistent with the manual cell type annotation (**Fig. 5c**), with ARI  
217 increasing from 0.13 to 0.58. These clusters are also marked by more specific enrichment of  
218 protein markers (**Fig. 5d**). For example, scMerge2 is able to reveal a cluster of T cells that  
219 uniquely expressed CD8a but not CD4 and a cluster that expressed of CD4 but not CD8a. Sim-  
220 ilarly, scMerge2 identifies the B cell cluster that has high expression in CD20, while clustering  
221 directly on the unadjusted matrix results in several clusters with qualitatively similar enrichment  
222 of markers, lacking the ability to identify distinguished cell types (**Fig. 5e**).

223

224 Lastly, we demonstrate scMerge2 can efficiently remove the unwanted variation of multi-  
225 modal data, such as Cellular Indexing of Transcriptomes and Epitopes by Sequencing (CITE-  
226 seq) data that concurrently measure RNA and cell-surface proteins of the same cell. In this case,  
227 we can remove the unwanted variation for each of the two modalities separately using scMerge2.  
228 We first examined the quality of data integration using two CITE-seq datasets with six batches  
229 and 87 common surface proteins measured (The same data used in **Fig. 2a-b**). We find that  
230 scMerge2 utilising the hierarchical merging strategies achieves a better balance between batch  
231 effect removal and cell type signal preservation than most of the other methods, with comparable  
232 performance with Harmony (**Supp Fig. S14**). Similar to the findings in scRNA-seq, using  
233 surface protein expression adjusted by scMerge2 improves the severity prediction, compared  
234 to the raw data (**Supp Fig. S15**). With the adjusted expression matrix of each modality, one can  
235 perform any multi-modal integration approach to obtain the joint latent space and visualisation  
236 of cells with batch effect removal [13, 31, 32]. As an example, we used j-UMAP that generates  
237 joint visualisation of the adjusted multi-modal data [32], which further confirms the effective  
238 integration of the six batches from the two CITE-seq datasets (**Fig. 5f**).



## 239 Discussion

240 We have presented scMerge2, a scalable approach for integrating data from large-scale multi-  
241 sample multi-condition single-cell studies. This was achieved via the use of three essential in-  
242 novations with hierarchical integration, pseudo-bulk building to minimise processing demand,  
243 and pseudo-replication that accounts for circumstances with phenotypes. Our algorithm enabled  
244 the atlas-scale integration of 20 global COVID-19 studies with around 5 million cells from 963  
245 donors, 1298 samples. We illustrated that scMerge2 data integration enabled the detection of  
246 distinct cell states in COVID-19 patients of variable severity. Finally, scMerge2 merged millions  
247 of cells from a number of single-cell technologies, including as CITE-seq, CyTOF, and image  
248 mass cytometry.

249

250 The type of output extracted from atlas-scale data integration has an important impact on  
251 the analytical question of interest. To date, there are three standard types of output from re-  
252 cent atlas-scale data integration (defined as over millions of cells). These are (i) an adjusted  
253 gene expression matrix, (ii) a low-dimensional projection of the data, known in machine learning  
254 as “embeddings”; and (iii) a unified graph representation. Various methodological approaches  
255 may provide one or more of these types of outputs. In general, there are a number of existing  
256 approaches that use modern deep learning-based algorithms to achieve fast, atlas-scale integra-  
257 tion. Given that single-cell data are ultra sparse high-dimensional datasets, “embeddings” are a  
258 natural output since they are effective for joint data visualisation and reduce memory load. How-  
259 ever, an embedding output by itself increases interpretability challenges since a low-dimensional  
260 representation does not naturally lend itself to the development of interpretable features such  
261 as cell-cell interactions or pathway information, which is crucial for downstream case-control  
262 studies or multi-treatment analysis. One step towards achieving a balance between generating  
263 adjusted expression matrices and appropriate memory usage is to enable selective adjusted out-  
264 put. For example, scMerge2 enables the extraction of a subset of genes (such as the top  $n$  highly  
265 variable genes) of the adjusted matrix for all 5 million cells in the COVID-19 data sets as well  
266 as outputting the adjusted matrix by batches, allowing users to effectively balance computational  
267 burden with specific downstream analytical strategies.

268

269 The order of integration is an important factor in hierarchical merging, which can be knowledge-  
270 guided or data-guided. Our current method is based on a data-guided order, in which we integrate

271 batches within one study or studies with similar size first. In contrast, a priori information such  
272 as sequencing platforms or cell extraction techniques can be used in knowledge-guided order  
273 of integration. Noted that the hierarchical data integration design can be broadly classified into  
274 two strategies [33], balanced trees and concatenating approaches. The balanced tree approach  
275 integrates between pairs of datasets at different levels of the tree, and the procedure is continued  
276 until all data is merged. The concatenating approach sequentially integrates datasets, therefore  
277 for  $n$  data sets, this will need  $n - 1$  steps of integration. Previous studies have found that nor-  
278 malisation results are very similar between the two types of integration tree structures [33]. The  
279 key difference between the approach is computational burden with the concatenating approach  
280 being more computational intensive. Currently, the scMerge2 approach is closer to the balance  
281 approaches allowing for many datasets to be added simultaneously at each level.

282

283 We demonstrated that our curation and effective integration of the COVID-19 gene expres-  
284 sion data with over 1000 individual samples facilitates flexible downstream meta-analysis, offer-  
285 ing the opportunity to examine particular sub-populations that cannot be adequately addressed  
286 with individual datasets. Scientists, for example, may investigate the molecular differences un-  
287 derlying mild and severe outcomes for a given age group (e.g., middle-aged individuals between  
288 41 - 50). Such analyses are difficult to perform in individual studies due to the limited sample  
289 sizes. This challenge can be overcome by merging several datasets.

290

291 Recent technological advancements substantially extend beyond scRNA-seq, enabling other  
292 data modalities (e.g. DNA, proteins) to be profiled in individual cells providing a more com-  
293 prehensive molecular view of the cellular regulation. For the datasets with multi-modal profiles  
294 measured for the same cell (paired data), such as CITE-seq and ASAP-seq, scMerge2 can be  
295 applied to integrate data from different batches by either considering each each modality as a  
296 separated matrix, or concatenating the data into a single matrix. Currently, the integration il-  
297 lustrated in this paper was done within each modality. In the future, we can incorporate the  
298 multi-modal information to better identify the pseudo-replicates of the paired data as well as  
299 utilise the higher-order relationship of features to improve the integration performance.

300

301 In summary, scMerge2 enables atlas-scale integrative analysis of large collections of single-  
302 cell data. As the availability of public multi-sample multi-conditional single-cell studies con-  
303 tinues to surge, scMerge2 demonstrates its ability to integrate over 5 million cells for further

304 downstream analysis, thereby enabling effective downstream meta-analysis. Notability, when  
305 compared to the raw log-normalised data from the outset, we demonstrated that scMerge2 offers  
306 a significant improvement in the prediction accuracy rate across all of the main cell types. The  
307 merge of large collections of scRNA-seq datasets from several cohorts further enables identifi-  
308 cation of distinct cell states in COVID-19 patients whose symptoms are of varying degrees of  
309 severity. Finally, scMerge2 has the ability to combine the data from millions of cells obtained  
310 from a variety of single-cell technologies, such as CITE-seq, CyTOF, and image mass cytometry.

## 311 **Methods**

### 312 **scMerge2**

#### 313 **Single-cell grouping within one batch**

314 Following the same principals as scMerge, the new scMerge2 approach begins by grouping the  
315 cells that share similar biological signals within each dataset or batch. We can approach this in  
316 two ways: one way is to perform unsupervised clustering; the other way is using results from  
317 supervised cell type classification.

- 318 • **Clustering-based grouping:** This is performed by default when no cell type label is used as  
319 input. Firstly, the top 2000 highly variables genes (HVG) are selected using *getTopHVGs*  
320 in the *scran* R Package, using batch information as block information. For data like CyTOF  
321 and ADT from CITE-seq data, this step will be skipped and all features will be used in the  
322 next step. Next, within each batch, instead of using k-means clustering as in the previ-  
323 ous version, we construct a shared nearest neighbour graph on the gene expression of the  
324 HVGs, with a default number of neighbours of 10, followed by louvain clustering. This  
325 therefore relieves the need of predefining the number of clusters that is required in our  
326 previous version.
- 327 • **Reference-based grouping:** This refers to the use of supervised cell type classification to  
328 predict or annotate the cell types using one or more reference datasets. This ensures the  
329 cell-type annotations are consistent among datasets. Cell type classification algorithms  
330 (e.g. *scClassify* [34] and *SingleR* [35]) can also be used and the reference dataset can be  
331 external datasets with similar cell types to the data to be integrated. This approach unifies  
332 cell type annotation across all datasets and eliminates the need for clustering and cell type  
333 annotation after data integration. It is noted that this approach is used in the COVID-19  
334 case study to integrate the data collection of 20 datasets.

#### 335 **Pseudo-bulk construction**

336 With the cell type grouping of each batch determined, scMerge2 next constructs multiple pseudo-  
337 bulk within each cell type. The pseudo-bulk construction significantly reduces the computational  
338 time in two main steps of the original version of scMerge [22]: identification of pseudo-replicates

339 and RUVIII model estimation. scMerge2 provides two approaches to calculate cell-type-specific  
340 pseudo-bulk for each batch:

- 341 • when count data are not available for all datasets, for each cell type grouping, we randomly  
342 assign the cells into  $k$  subsets and take the gene-wise average of each subset as one pseudo-  
343 bulk. This therefore results with  $k$  pseudo-bulk for one cell type grouping.
- 344 • when counts data are available for all data, we can perform a similar pool-and-divide strat-  
345 egy that is proposed in RUVIII-NB [36]. Here, we can have two strategies in pooling the  
346 cells: (1) assign the cells based on library size; (2) randomly assign the cells into  $k$  subsets.  
347 Then we gene-wisely take the sum of the counts for each subset and generate the counts  
348 data following a negative binomial distribution. While the pseudobulk matrix generated by  
349 this strategy is able to maintain the gene mean-variance relationship [36], we find that this  
350 approach does not improve the quality of data integration in scMerge2 (Supp Fig. S6).

351 Noted that  $k$  is set as 30 by default for cell type group with more than  $k$  number of cells, and  
352 pseudo-bulk are not constructed for cell types with less than  $k$  cells, i.e., all the cells from these  
353 cell types will be retain for the next steps of scMerge2.

## 354 **Pseudo-replicates identification across batches in scMerge2**

355 Replicates are considered as the samples with similar biological variation across batches. Con-  
356 struction of pseudo-replicates is one of the key steps in scMerge which later are utilised to es-  
357 timate the unwanted variation from the data. In scMerge, we proposed a five-step procedure to  
358 identify pseudo-replicates by clustering on a mutual nearest cluster (MNC) graph, where each  
359 node of the MNC graph indicates a group of cells in a batch. scMerge2 follows similar steps as  
360 the previous version, but with two major improvements:

- 361 • The pseudo-replicates identification is based on the pseudo-bulk matrix to reduce the com-  
362 putational time;
- 363 • For data with multiple conditions (or other observed biological factors), scMerge2 allows  
364 the MNC graph to be constructed within each condition to preserve the biological variation.  
365 Note that this strategy can only be used when the batches to be merged have at least one  
366 common condition and can only be performed in the condition with multiple batches.

## 367 Estimation of RUVIII model using pseudo-bulk

368 The underlying model of scMerge2 is the fastRUVIII model that takes the gene-wise standard-  
369 ized gene expression matrix that is log-transformed and cosine normalised as input. Let  $Z_{cg}$   
370 be the standardized data, where  $c = 1, \dots, C$ , with  $C$  indicates the number of cells from all  
371 batches/datasets in total;  $g = 1, \dots, G$ , with  $G$  indicates the number of genes. Following the same  
372 annotation in scMerge, we formulate  $Z_{C \times G}$  using RUVIII model as

$$Z_{C \times G} = X_{C \times p} \beta_{p \times G} + W_{C \times k} \alpha_{k \times G} + \epsilon_{C \times G},$$

373 where  $X$  denotes the matrix of observed factors of interest;  $p$  denotes the number of factors  
374 of interest;  $W$  denotes the matrix of unobserved factors of unwanted variation;  $\alpha$  denotes the  
375 coefficient of  $W$ ;  $k$  denotes the number of unwanted factors, which is unknown (set as 20 by  
376 default for scRNA-seq data, and 10 for ADT from CITE-seq data and CyTOF data);  $\epsilon$  denotes the  
377 random error. Following the RUVIII model estimation proposed in [37, 22], the model removes  
378 the unwanted variation from  $Z_{C \times G}$ . In summary, it follows the three steps:

- 379 • Step i: estimate  $\alpha$  via the first  $k$  right singular vectors of Singular Value Decomposition  
380 (SVD) on  $R_M Z$ , where  $R_M = 1 - M(M^T M)^{-1} M^T$ , with the replicate matrix  $M \in R^{C \times N}$ ,  
381  $N$  indicates the number of types of pseudo-replicates;
- 382 • Step ii: estimate  $W$  by  $W_{C \times k} = Z_s \hat{\alpha}_s^T (\hat{\alpha}_s \hat{\alpha}_s^T)^{-1}$ , where  $\hat{\alpha}_s \in R^{k \times G_s}$  indicates the the  
383 submatrix of  $\alpha$ , which columns include only the genes that belongs to single-cell stably  
384 expressed genes (SEG) with number of genes as  $G_s$  (SEG selection and evaluation can be  
385 found in [38]);
- 386 • Step iii: adjust the matrix by subtracting the estimated unwanted variation component:  
387  $\hat{Z}_{C \times G} = Z_{C \times G} - \hat{W}_{C \times k} \hat{\alpha}_{k \times G}$ .

388 SVD is a computationally intensive algorithm, especially for large matrices like single-cell data.  
389 We argue that for Step 1, we do not need the full single-cell data to estimate  $\alpha$ . Instead, we can  
390 subsample the data or construct cell-type-specific pseudo-bulk which are informative enough  
391 to approximate the full single-cell matrix to reduce the computational burden in estimation of  
392  $\alpha$ . Let  $Z_{C_b \times G}$  denote the the “sketch” of the full single-cell matrix derived from pseudo-bulk  
393 construction step, where the column denotes the number of the genes, with the same dimension  
394 as the full data  $Z$ ; the row now indicates the number of pseudo-bulk, with dimension  $C_b$ . We then

395 construct pseudo-replicates based on the pseudo-bulk matrix  $Z_b$  to obtain the replicate matrix  
396  $M_b \in R^{C_b \times N_b}$  (See Section *Pseudo-replicates identification across batches in scMerge2* for more  
397 details). We estimate  $\hat{\alpha}^b$  using the first  $k$  right singular vectors of SVD on  $R_{M_b}Z_b$ . By treating  
398  $\hat{\alpha}^b$  as the approximation of  $\hat{\alpha}$ , we then next bring back the full single-cell matrix  $Z$  to estimate  
399  $W$  and adjusted  $\hat{Z}$  following the same Steps 2-3 above.

## 400 **Hierarchical merging**

401 When we integrate data from different studies, the unwanted variation can come from multiple  
402 levels, such as batch effect of samples within each study but also between studies. In this case,  
403 a hierarchical integration strategy would be useful to first adjust intra-study unwanted variation  
404 effect, and then perform the inter-study data integration. On the other hand, when we integrate a  
405 large number of studies, such as the COVID-19 data collection in this paper, starting from cor-  
406 recting the data of a smaller set of studies can be a more efficient way to estimate the parameters  
407 of the model to harmonise the data [33].

408 scMerge2 allows users to input a hierarchical tree strategy to perform the data adjustment in  
409 a multi-level manner. The data adjusted on the current level will be used as input on the next  
410 level. For the COVID-19 200k data collection, we first integrated the the 3 batches within each  
411 dataset before integrating the two datasets. For the COVID-19 scRNA-seq data collection, we  
412 first performed the adjustment on four datasets that have multiple cohorts (Ren, Stephenson, Liu  
413 and Schulte-Schrepping) to correct the intra-study unwanted variation (where the cohort label is  
414 used as batch label) as well as between the two datasets that have very different sequencing depth  
415 (Arunachalam and Wilk). Next, we performed the adjustment of the 13 datasets with less than  
416 200,000 cells. We finally integrated all the 20 studies together, where the study label is used as  
417 batch label.

## 418 **Data collection and preprocessing**

### 419 **COVID-19 scRNA-seq data collection**

420 We collected 20 public COVID-19 PBMC and whole blood scRNA-seq datasets (**Supplementary**  
421 **Table 1**). The raw count matrix of each dataset is size-factor standardized and log-transformed  
422 using *logNormCount* function from *scater* [39] R package. To unify the cell types from differ-  
423 ent studies, we performed scClassify to reannotate the cell types based on a 3-level hierarchical  
424 cell type tree [34], using three distinct reference datasets that were either generated from whole

425 blood (Wilk) or generated by CITE-seq protocol that contains multi-level annotations (Liu and  
426 Stephenson).

#### 427 **COVID-19 200k CITE-Seq data collection (COVID-19 200k)**

428 To benchmark scMerge2 with other methods, we subset 200k cells from the two COVID-19 stud-  
429 ies (Liu and Stephenson) as a benchmarking dataset that with 17446 genes, 87 proteins and 184  
430 samples from 3 conditions (Healthy, Mild/Moderate, Severe/Critical) to assess the concordance  
431 performance of the adjusted gene expression matrix after data integration. Both of these two  
432 studies have three batches within the studies, which allows us to evaluate the hierarchical merg-  
433 ing strategy in scMerge2 (i.e., scMerge2-h), where we first integrated the three batches within  
434 each batch, with  $k_{\text{RUV}} = 10$  ( $k_{\text{RUV}}$  denotes the number of unwanted variation) and then performed  
435 the integration across two datasets, with  $k_{\text{RUV}} = 10$ .

436 The raw antibody derived tag (ADT) counts matrix of each dataset is size-factor standardized  
437 and log-transformed using the *logNormCount* function from *scater* [39]. In scMerge2, we used  
438 all features as negative controls and used  $k_{\text{RUV}} = 3$  in both levels in scMerge2-h.

#### 439 **COVID-19 60k data collection (COVID-19 60k)**

440 To evaluate the robustness of the parameters in scMerge2, we further created a smaller subset  
441 of data, which is derived from selecting the cells from moderate/mild patients of the Stephenson  
442 data from the COVID-19 200k data. The selected subset has 66967 cells from 58 samples and  
443 17446 genes where the aim is to integrate three different batches in the Stephenson data.

#### 444 **COVID-19 CyTOF data collection**

445 Two public COVID-19 PBMC CyTOF datasets (**Supplementary Table 1**) were downloaded  
446 from FlowRepository with ID FR-FCM-Z2XA for Geanon data [40] (4,747,543 cells from 21  
447 samples) and zenodo <https://doi.org/10.5281/zenodo.6120249> for data from gran-  
448 ulocyte depleted whole blood in COMBAT study [41] (7,118,158 cells from 160 samples), which  
449 both contain the expression matrix and cell type annotations. To combine the two studies, we  
450 manually unified antibody names and the cell type annotations to 18 cell types. The expres-  
451 sion matrices were then used as input for scMerge2. Noted that we used all features as negative  
452 controls in scMerge2.



## 453 **COVID-19 IMC data collection**

454 The COVID-19 IMC dataset generated by [29] aims to assess the pathology of lungs across  
455 Covid-19 disease progression. The dataset, including cell intensities and metadata, was ob-  
456 tained from the repository [https://zenodo.org/record/4139443#.Yw\\_gk9LMKXI](https://zenodo.org/record/4139443#.Yw_gk9LMKXI)  
457 provided in the publication and contained 237 images generated from 23 samples across 43 mark-  
458 ers. In the original manuscript [29], the cell types were annotated by first clustering using the  
459 Leiden algorithm and then manually curated into 17 meta-clusters based on marker expression,  
460 phenotype, and proximity to lung structures.

## 461 **Evaluation**

### 462 **Part I - Simulation**

463 Simulation framework. We adopted a simulation framework to generate single-cell multi-condition  
464 and multi-sample data with batch effect based on scDesign3 [24]. This framework is able to sim-  
465 ulate single-cell count data that preserve the gene-wise correlation structure. Similar to many  
466 other simulators, scDesign3 required a real training scRNA-seq data to estimate the required  
467 parameters. Here, we have taken a subset of Stephenson data that contains four cell types (B cell,  
468 CD14 Monocytes, CD4 T and CD8 T) and 23 samples from two conditions (Healthy and Severe)  
469 as training data. From each sample, we randomly subsampled 400 cells. Only genes that were  
470 in the top 2000 highly variable genes and expressed in more than 2% of the cells were included.  
471 We further excluded any genes that were originally considered as differentially expressed (with  
472 adjusted p-value  $< 0.2$ ). This resulted in the training data with 9200 cells and 1196 genes from  
473 23 samples. Our simulation framework includes three main steps.

474

475 *Step 1:* Construct a null dataset with no differentially expressed genes by first permuting the  
476 condition labels in the training data. We then estimate both cell-type and sample variation in  
477 the data using the function *fit\_marginal()* in scDesign3 that fits the marginal distribution of each  
478 gene using a negative binomial distribution with the mu formula  $\sim \text{cell type} + \text{sample}$   
479  $\text{ID} + \text{condition}$  and the sigma formula  $\sim 1$ . Then we used a vine copula to estimate the  
480 gene correlation from the real training data.

481

482 *Step 2:* Introduce the batch effect to the simulated data. Assuming all genes are affected by

483 the batch variation, we drew a vector with length equal to the number of genes from a log-normal  
484 distribution with mean  $\log(2)$  and standard deviation 0.43 as batch effect on the mean of the gene  
485 distribution. The direction of the batch effect is randomly assigned to each gene.

486

487 *Step 3:* Introduce the ground truth differential state genes to the simulated data. For each  
488 cell type, we randomly select  $p\%$  of genes to be differentially expressed between two conditions  
489 ( $p = 5, 10$  in our study). The log fold changes (logFC) vector is simulated from a log-normal  
490 distribution, with the mean  $\mu_{lfc}$  and the standard deviation  $\sigma_{lfc}$ . In our evaluation setting, we  
491 consider a range of logFC values from  $\mu_{lfc} = 1.1$  to 2 in 0.1 increment and  $\sigma_{lfc} = 0.43$ . The  
492 direction of the regulation is randomly assigned to each DS genes using a binomial distribution  
493 with probability 0.5.

494

495 Lastly, with the fold change of both batch effect and condition effect combined with the  
496 parameters estimated in *Step 1*, the simulated single-cell data is generated from the negative bi-  
497 nomial distribution using strategies implemented in *simu\_new()* of *scDesign3*. For each value of  
498 logFC, we simulated 18,400 cells (23 samples, each sample with 800 cells), with 5% or 10%  
499 differential states genes within each cell types.

500

501 Evaluation metrics and settings - Differential states analysis. To assess the impact of data inte-  
502 gration on downstream analytics, we considered the performance of the differential states anal-  
503 ysis results on the simulated data. Our evaluation is based on three metrics; false discovery rate  
504 (FDR), true positive rate (TPR) and F1 scores. For each log-transformed simulated matrix with  
505 dimension  $G \times C$ , with  $S$  samples and  $T$  cell types, we took the gene-wise average of each sample  
506 within each cell type, resulting in a  $G \times S$  matrix for each cell type. We then performed a differ-  
507 ential state analysis using the *limma-trend* algorithm [25] on the cell-type specific sample-wise  
508 aggregated data using the default parameters.

## 509 **Part II - Real data comparison**

510 Evaluation setting for scRNA-seq and CITE-seq data collection.

- 511 1. *Signal to noise ratio:* We used ARI and ASW (see evaluation metrics below) to evaluate  
512 the concordance of clustering results with respect to the cell type labels and the datasets.  
513 A desirable data integration method will show a high concordance between the clustering

514 result and known cell type information (signal refers to cell types) and a low concordance  
515 between the clustering results and known datasets information (noise refers to batch effect).

516 2. *Severity prediction:* We aggregated cell-type-specific average expression of each sample to  
517 a gene by sample matrix for each cell type. We then used each cell-type specific matrix to  
518 predict the sample condition (Healthy, Mild/Moderate and Severe/Critical) using support  
519 vector machine (SVM) with radial basis function kernel. The prediction performance was  
520 evaluated using repeated 5-fold cross validation with 20 repeats. We evaluate the prediction  
521 performance using  $F_1$  score.

522 3. *Visualisation plot:* For scRNA-seq data, we used Uniform Manifold Approximation and  
523 Projection (UMAP) to visualise and evaluate the results of the adjusted expression matrix.  
524 For CITE-seq case study, we used j-UMAP to jointly visualise the two modalities [32],  
525 where we first performed PCA within each modality, and then j-UMAP was performed to  
526 obtain the joint UMAP embeddings of the two modalities.

527 Evaluation on IMC data collection. We applied scMerge2 to perform data integration of the 23  
528 samples. This is achieved by first filtering and selecting the data using the 38 markers specified in  
529 the original publication [29] and removing all undefined cell types (i.e. cells having cell type an-  
530 notation as “nan”). Next, considering sample labels as batch information, we applied scMerge2  
531 with settings  $k_{RUV} = 2$ ,  $k_{pseudoBulk} = 5$ ,  $k_{celltype} = 20$ , using all markers as negative control genes  
532 and highly variable genes. Thirdly, unsupervised clustering was performed on both the unnor-  
533 malised and scMerge2 normalised datasets using the FlowSOM [30] algorithm with 17 clusters.  
534 The Adjusted Rand Index (ARI) was used to compare the concordance between this unsuper-  
535 vised clustering with the manually curated cell types in the original manuscript [29]. The results  
536 are visualised using heatmaps showing the average marker abundance in the cell types. Average  
537 marker abundance were generated after scaling the marker expression by computing the ratio of  
538 the mean of each marker and its standard deviation.

539

540 Sensitivity analysis of scMerge2. We examined the robustness of the following parameters in  
541 scMerge2: the number of pseudobulk constructed; the number of neighbours in SNN graph;  
542 the pseudobulk construction strategy and the number of unwanted variation. We performed our  
543 sensitivity analysis on the COVID-19 60k data on a number of settings for each of the four  
544 parameters as below:

- 545 • Number of pseudobulk constructed within each group: 10, 20, 30, 40 and 50
- 546 • Number of neighbours in SNN graph: 5, 10, 15, 20, 25 and 30
- 547 • Ways of pseudobulk construction: Default, Pool-Divide, Pool-Divide (Random)
- 548 • Number of factors of unwanted variation to be removed: 10, 15, 20, 25 and 30

549 For each setting, we repeat the analysis 10 times with a different seed and assess the concordance  
550 performance of the signal to noise ratio using ASW and ARI as evaluation metrics as describe in  
551 the Section *Evaluation metrics*. We compared against benchmarking methods described in the  
552 Section *Benchmarking methods*.

### 553 **Evaluation metrics**

554 We used three metrics to assess the performance of data integration results from different meth-  
555 ods. Details of the evaluation metrics are described as follows:

- 556 • *Adjusted Rand Index (ARI) - Clustering analysis*: We used ARI to quantify the concordance  
557 of the clustering results with respect to the cell type (ARI (cell type)) and batch labels  
558 (ARI (batch)). The clustering results for all methods were derived from first building  
559 a shared nearest neighbour from the batch corrected embeddings with a default number  
560 of neighbours of 10, followed by louvain clustering. For scMerge2, the batch corrected  
561 embeddings were derived from the top 20 PCs of the adjusted gene expression matrix.
- 562 • *Average silhouette width (ASW) - Embedding visualisation*: We calculated the average of  
563 silhouette coefficients for each cell (ASW) by considering two different groupings: cell  
564 type (ASW (cell type)) and batch label (ASW (batch)), based on the Euclidean distance  
565 obtained from the UMAP embeddings generated from the batch corrected embeddings.
- 566 • *PCA scores*: We calculated the coefficient of determination ( $R^2$ ) for a linear regression  
567 model that fitted each of the first 20 principal component with technical variation labels,  
568 such as batch, technology and protocol labels. We then calculated the product of the vari-  
569 ance explained by each principal component and the corresponding  $R^2$ . The final PCA  
570 score was calculated by summing across the products, which quantify how much the PCs  
571 explained the unwanted technical variation.

## 572 **Benchmarking methods**

573 We benchmarked the performance of scMerge2 against five other methods that are designed  
574 for data integration of scRNA-seq datasets in terms of the batch corrected embeddings in the  
575 COVID-19 200k data. Detailed settings used in each method are as follows:

576 **(i) Seurat.** Applying Seurat with canonical correlation analysis set as the reduction method.  
577 Version 4.1.1. of the *Seurat*[42] R package was used. We first identified the variable features  
578 within each batch using *FindVariableFeatures()* and then selected the integration features using  
579 *SelectIntegrationFeatures()*. The integration anchors were then identified using *FindIntegratio-*  
580 *nAnchors()* with reduction set as “cca”, followed by *IntegrateData()* to obtain the integrated data.

581 **(ii) SeuratRPCA.** Similar to Seurat (CCA), within each batch, we first found the variable  
582 features, with an addition PCA step performed. After integration features were selected, *Find-*  
583 *IntegrationAnchors()* was performed with reduction set as “rpca”. Lastly, *IntegrateData()* was  
584 performed to obtain the integrated data.

585 **(iii) fastMNN.** This is a fast version of the mutual nearest neighbors (MNN) method [43]. R  
586 package *batchelor v1.12.3* was used. We ran *fastMNN()* with default parameters to derived both  
587 the batch corrected embeddings and adjusted expression matrix.

588 **(iv) Liger.** R package *rliger v1.0.0* [44] was used. Online integrative nonnegative ma-  
589 trix factorization was performed to obtain the batch corrected embedding following the tutorial  
590 ([https://github.com/welch-lab/liger/blob/master/vignettes/online\\_](https://github.com/welch-lab/liger/blob/master/vignettes/online_inMF_tutorial.html)  
591 [inMF\\_tutorial.html](https://github.com/welch-lab/liger/blob/master/vignettes/online_inMF_tutorial.html)), where we first ran *selectGenes()* to select the features, *scaleNotCen-*  
592 *ter()* to scale the features, and *online\_inMF()* with *miniBatch\_size = 5000* and *max.epochs = 5*.

593 **(v) Harmony.** R package *Harmony v0.1.0* [18] was used. The PCA space returned by *run-*  
594 *PCA()* of R package *scater* was used as input, and then *HarmonyMatrix()* was performed with  
595 *do\_pca = FALSE* to retain the batch corrected embedding.

## 596 **COVID-19 downstream analysis**

### 597 **Differential abundance analysis on the cells from mild/moderate and severe/critical samples**

598 Differential abundance (DA) analysis was performed on the cells from mild/moderate and se-  
599 vere/critical samples using DA-seq [26]. The top 30 PCs derived from the adjusted expression  
600 data were used as input for the algorithm to calculate the DA scores. A range of  $k$  values from 50  
601 to 500 was used for the calculation of DA score vector with kNN. We define salient differential

602 abundance (DA) cells as cells with absolute abundance scores greater than 0.8.

### 603 **Differential states analysis of DA cells**

604 For all DA cells, we aggregated cell-type-specific abundance scores (or values) of each sample  
605 to a gene by sample matrix for each cell type. Next, we model the aggregated cell-type-specific  
606 abundance values across using a linear model with severity and the days since symptom onset  
607 as covariates. We account for sample level variability using the limma-trend implementation in  
608 the R package *limma* [25]. We then ranked the genes based on the test statistics. The preranked  
609 based gene set enrichment analysis (GSEA) of the selected pathways that are related COVID-  
610 19 disease mechanism [28] (as listed in **Fig. 4c**) is measured using the *fgsea* function in the  
611 R package *fgsea v1.22.0* [45]. Significant pathways are defined with adjusted p-value less than  
612 0.05.

## 613 **Data availability**

614 All data used in this study are included in Supplementary Data 1. All analysis was done in R  
615 version 4.1.2.

## 616 **Code availability**

617 The code to run scMerge2 is part of the scMerge package (Github: <https://github.com/SydneyBioX/scMerge>)  
618 and is available under the GPL-3 license.

## 619 **Acknowledgments**

620 The authors thank all their colleagues, particularly at The University of Sydney, Sydney Precision  
621 Data Science Centre and Charles Perkins Centre for their support and intellectual engagement.  
622 A special thank to Mr. Dongyuan Song and A/Prof. Jingyi Jessica Li from UCLA for providing  
623 the R code for scDesign3.

624 The following sources of funding for each author are gratefully acknowledged: the AIR@innoHK  
625 programme of the Innovation and Technology Commission of Hong Kong to JYHY, EP, YC and  
626 YL. Australian Research Council Discovery Early Career Researcher Award (DE200100944)  
627 funded by the Australian Government to EP; Research Training Program Tuition Fee Offset  
628 and Stipend Scholarship and Chen Family Research Scholarship to YL; Australian Government  
629 Research Training Program (RTP) Scholarship to YC and YL; and the University of Sydney  
630 Postgraduate Excellence Award for EW. The funding sources had no impact on the study design;  
631 in the collection, analysis, and interpretation of data, in the writing of the manuscript, and in the  
632 decision to submit the manuscript for publication.

## 633 **Author contributions**

634 JYHY and YL conceived and designed the study with input from EP. YL and JYHY led the  
635 method development and guided the evaluation data analysis. YL and YC jointly curated the  
636 scRNA-seq data and YL implemented all data analytics and developed the corresponding R code.  
637 EW and EP curated the IMC data and performed the data analytics related to imaging data. All  
638 authors wrote, read, reviewed the manuscript and approved the final version.

## 639 **Conflict of interest**

640 The authors declare that they have no conflict of interest.



## 641 **References**

- 642 [1] Byungjin Hwang, Ji Hyun Lee, and Duhee Bang. “Single-cell RNA sequencing technolo-  
643 gies and bioinformatics pipelines”. In: *Experimental & molecular medicine* 50.8 (2018),  
644 pp. 1–14.
- 645 [2] Ting Zhang et al. “Progress and applications of mass cytometry in sketching immune  
646 landscapes”. In: *Clinical and Translational Medicine* 10.6 (2020), e206.
- 647 [3] Felix J Hartmann and Sean C Bendall. “Immune monitoring using mass cytometry and  
648 related high-dimensional imaging approaches”. In: *Nature Reviews Rheumatology* 16.2  
649 (2020), pp. 87–99.
- 650 [4] Aviv Regev et al. “The Human Cell Atlas”. en. In: *Elife* 6 (Dec. 2017).
- 651 [5] Junyue Cao et al. “A human cell atlas of fetal gene expression”. en. In: *Science* 370.6518  
652 (Nov. 2020).
- 653 [6] Silvia Domcke et al. “A human cell atlas of fetal chromatin accessibility”. en. In: *Science*  
654 370.6518 (Nov. 2020).
- 655 [7] Orit Rozenblatt-Rosen et al. “The Human Tumor Atlas Network: Charting Tumor Tran-  
656 sitions across Space and Time at Single-Cell Resolution”. en. In: *Cell* 181.2 (Apr. 2020),  
657 pp. 236–249.
- 658 [8] Mengwei Li et al. “DISCO: a database of Deeply Integrated human Single-Cell Omics  
659 data”. In: *Nucleic Acids Research* 50.D1 (2022), pp. D596–D602.
- 660 [9] Yuan Tian et al. “Single-cell immunology of SARS-CoV-2 infection”. en. In: *Nat. Biotech-*  
661 *nol.* 40.1 (Jan. 2022), pp. 30–41.
- 662 [10] Rachel L Belote et al. “Human melanocyte development and melanoma dedifferentiation  
663 at single-cell resolution”. en. In: *Nat. Cell Biol.* 23.9 (Sept. 2021), pp. 1035–1047.
- 664 [11] Viktor Petukhov et al. “Case-control analysis of single-cell RNA-seq studies”. en. Mar.  
665 2022.
- 666 [12] Malte D Luecken et al. “Benchmarking atlas-level data integration in single-cell genomics”.  
667 In: *Nature methods* 19.1 (2022), pp. 41–50.
- 668 [13] Yuhan Hao et al. “Integrated analysis of multimodal single-cell data”. In: *Cell* 184.13  
669 (2021), pp. 3573–3587.

- 670 [14] Matthew Amodio et al. “Exploring single-cell data with deep multitasking neural net-  
671 works”. In: *Nature methods* 16.11 (2019), pp. 1139–1145.
- 672 [15] Brian Hie, Bryan Bryson, and Bonnie Berger. “Efficient integration of heterogeneous  
673 single-cell transcriptomes using Scanorama”. In: *Nature biotechnology* 37.6 (2019), pp. 685–  
674 691.
- 675 [16] Bin Zou et al. “deepMNN: deep learning-based single-cell RNA sequencing data batch  
676 correction using mutual nearest neighbors”. In: *Frontiers in Genetics* (2021), p. 1441.
- 677 [17] Krzysztof Polański et al. “BBKNN: fast batch alignment of single cell transcriptomes”.  
678 In: *Bioinformatics* 36.3 (2020), pp. 964–965.
- 679 [18] Ilya Korsunsky et al. “Fast, sensitive and accurate integration of single-cell data with Har-  
680 mony”. In: *Nature methods* 16.12 (2019), pp. 1289–1296.
- 681 [19] Romain Lopez et al. “Deep generative modeling for single-cell transcriptomics”. In: *Na-  
682 ture methods* 15.12 (2018), pp. 1053–1058.
- 683 [20] Chenling Xu et al. “Probabilistic harmonization and annotation of single-cell transcrip-  
684 tomics data with deep generative models”. In: *Molecular systems biology* 17.1 (2021),  
685 e9620.
- 686 [21] Xiangjie Li et al. “Deep learning enables accurate clustering with batch effect removal in  
687 single-cell RNA-seq analysis”. In: *Nature communications* 11.1 (2020), pp. 1–14.
- 688 [22] Yingxin Lin et al. “scMerge leverages factor analysis, stable expression, and pseudorepli-  
689 cation to merge multiple single-cell RNA-seq datasets”. In: *Proceedings of the National  
690 Academy of Sciences* 116.20 (2019), pp. 9775–9784.
- 691 [23] Helena L Crowell et al. “Muscat detects subpopulation-specific state transitions from  
692 multi-sample multi-condition single-cell transcriptomics data”. In: *Nature communica-  
693 tions* 11.1 (2020), pp. 1–12.
- 694 [24] Dongyuan Song et al. “A unified framework of realistic in silico data generation and sta-  
695 tistical model inference for single-cell and spatial omics”. In: *bioRxiv* (2022). DOI: 10 .  
696 1101/2022.09.20.508796. eprint: [https://www.biorxiv.org/content/  
697 early/2022/09/22/2022.09.20.508796.full.pdf](https://www.biorxiv.org/content/early/2022/09/22/2022.09.20.508796.full.pdf). URL: [https://www.  
698 biorxiv.org/content/early/2022/09/22/2022.09.20.508796](https://www.biorxiv.org/content/early/2022/09/22/2022.09.20.508796).

- 699 [25] Matthew E Ritchie et al. “limma powers differential expression analyses for RNA-sequencing  
700 and microarray studies”. In: *Nucleic acids research* 43.7 (2015), e47–e47.
- 701 [26] Jun Zhao et al. “Detection of differentially abundant cell subpopulations in scRNA-seq  
702 data”. In: *Proceedings of the National Academy of Sciences* 118.22 (2021), e2100293118.
- 703 [27] Matthew L Meizlish et al. “A neutrophil activation signature predicts critical illness and  
704 mortality in COVID-19”. In: *Blood advances* 5.5 (2021), pp. 1164–1177.
- 705 [28] Can Liu et al. “Time-resolved systems immunology reveals a late juncture linked to fatal  
706 COVID-19”. In: *Cell* 184.7 (2021), pp. 1836–1857.
- 707 [29] André F Rendeiro et al. “The spatial landscape of lung pathology during COVID-19 pro-  
708 gression”. In: *Nature* 593.7860 (2021), pp. 564–569.
- 709 [30] Sofie Van Gassen et al. “FlowSOM: Using self-organizing maps for visualization and  
710 interpretation of cytometry data”. In: *Cytometry Part A* 87.7 (2015), pp. 636–645.
- 711 [31] Ricard Argelaguet et al. “MOFA+: a statistical framework for comprehensive integration  
712 of multi-modal single-cell data”. In: *Genome biology* 21.1 (2020), pp. 1–17.
- 713 [32] Van Hoan Do and Stefan Canzar. “A generalization of t-SNE and UMAP to single-cell  
714 multimodal omics”. In: *Genome Biology* 22.1 (2021), pp. 1–9.
- 715 [33] Taiyun Kim et al. “hRUV: Hierarchical approach to removal of unwanted variation for  
716 large-scale metabolomics data”. In: *bioRxiv* (2020).
- 717 [34] Yingxin Lin et al. “scClassify: sample size estimation and multiscale classification of cells  
718 using single and multiple reference”. In: *Molecular systems biology* 16.6 (2020), e9389.
- 719 [35] Dvir Aran et al. “Reference-based analysis of lung single-cell sequencing reveals a transi-  
720 tional profibrotic macrophage”. In: *Nature immunology* 20.2 (2019), pp. 163–172.
- 721 [36] Agus Salim et al. “RUV-III-NB: Normalization of single cell RNA-seq Data”. In: *bioRxiv*  
722 (2021).
- 723 [37] Ramyar Molania et al. “A new normalization for Nanostring nCounter gene expression  
724 data”. In: *Nucleic acids research* 47.12 (2019), pp. 6073–6083.
- 725 [38] Yingxin Lin et al. “Evaluating stably expressed genes in single cells”. In: *GigaScience* 8.9  
726 (2019), giz106.

- 727 [39] Davis J McCarthy et al. “Scater: pre-processing, quality control, normalization and visu-  
728 lization of single-cell RNA-seq data in R”. In: *Bioinformatics* 33.8 (2017), pp. 1179–  
729 1186.
- 730 [40] Daniel Geanon et al. “A streamlined whole blood CyTOF workflow defines a circulating  
731 immune cell signature of COVID-19”. In: *Cytometry Part A* 99.5 (2021), pp. 446–461.
- 732 [41] David J. Ahern et al. “A blood atlas of COVID-19 defines hallmarks of disease severity  
733 and specificity”. In: *Cell* 185.5 (2022), 916–938.e58. ISSN: 0092-8674. DOI: <https://doi.org/10.1016/j.cell.2022.01.012>. URL: <https://www.sciencedirect.com/science/article/pii/S0092867422000708>.  
734  
735
- 736 [42] Tim Stuart et al. “Comprehensive integration of single-cell data”. In: *Cell* 177.7 (2019),  
737 pp. 1888–1902.
- 738 [43] Laleh Haghverdi et al. “Batch effects in single-cell RNA-sequencing data are corrected by  
739 matching mutual nearest neighbors”. In: *Nature biotechnology* 36.5 (2018), pp. 421–427.
- 740 [44] Chao Gao et al. “Iterative single-cell multi-omic integration using online learning”. In:  
741 *Nature biotechnology* 39.8 (2021), pp. 1000–1007.
- 742 [45] Gennady Korotkevich et al. “Fast gene set enrichment analysis”. In: *BioRxiv* (2021),  
743 p. 060012.

744 **Figure legends**

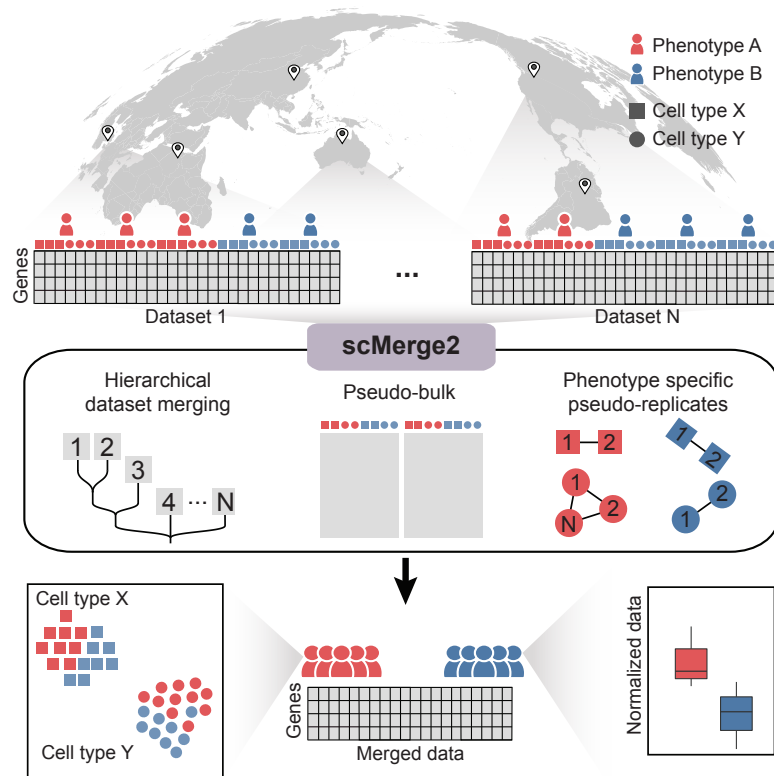


Figure 1: Overview of scMerge2: This new scalable algorithm uses (i) hierarchical integration to capture both local and global variation; (ii) pseudo-bulk construction to reduce computational load; and (iii) phenotype specific pseudo-replicate, and outputs adjusted expression matrix for millions of cells ready for downstream analysis.

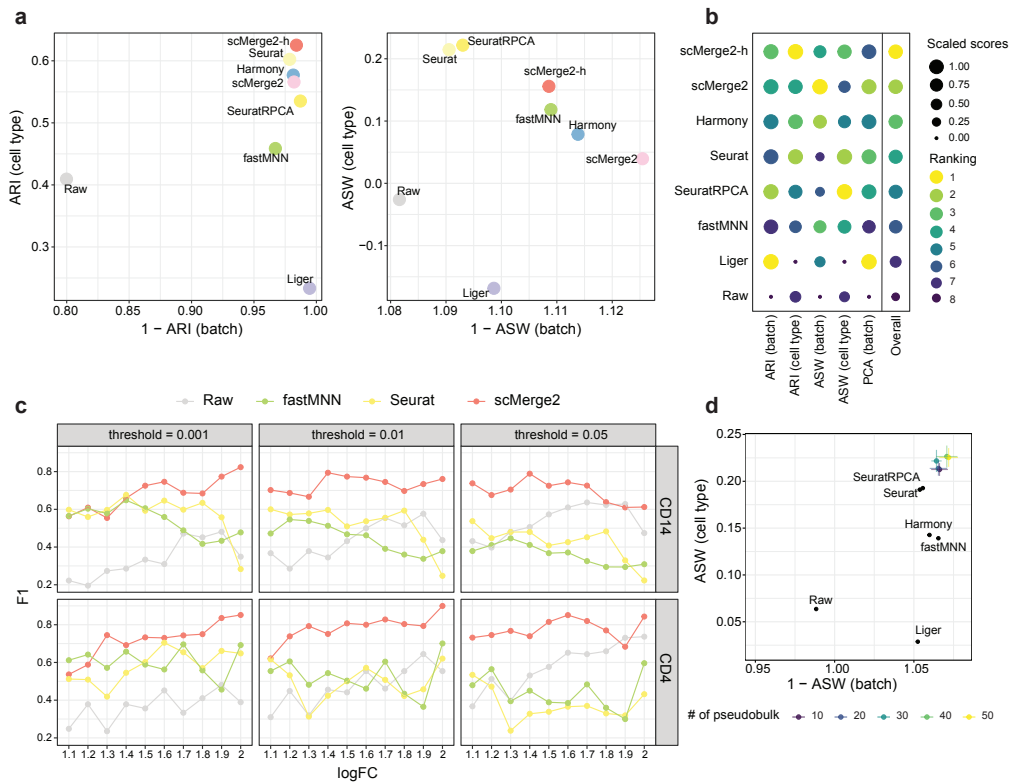


Figure 2: (a) Scatter plots of evaluation metrics of data integration of a 200k cells subset of two COVID-19 studies (Liu and Stephenson) for scMerge2, scMerge2-h (data merged in a hierarchical manner), Seurat, Seurat (RPCA), Harmony, fastMNN, Liger and Raw: Adjusted rand index (ARI) (left panel), where x-axis indicates 1 minus batch ARI and y-axis indicates cell type ARI; Average silhouette width (ASW), where the x-axis is 1 minus batch ASW and y-axis is the cell type ASW (right panel). (b) Dot plots indicate the ranking of the data integration methods in terms of five different evaluation metrics. The size of the dot indicates the scaled scores, which are obtained from the min-max scaling of the original values. The overall ranking is ranked based on the average ranking of the five evaluation metrics. (c) F1-score of the differential state (DS) results of two selected cell types (CD14 and CD4) (row) of simulated data, with 10% DS genes within each cell type, for scMerge2, Seurat, fastMNN and raw, varying simulated log fold change (logFC) of DS genes (x-axis) and different threshold of adjusted p-value (column). (d) Scatter plots of evaluation metrics of robustness analysis when varying the number of pseudobulk constructed within each cell type of each batch, where the x-axis is 1 minus batch ASW and y-axis is the cell type ASW.

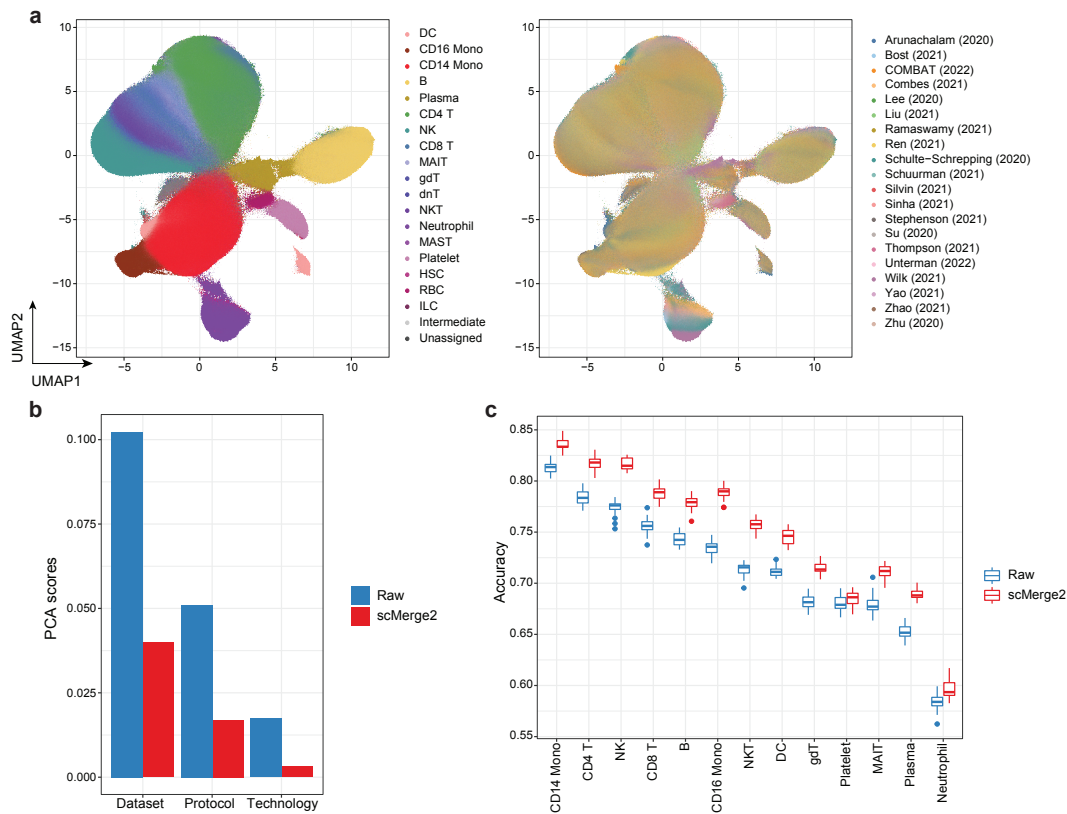


Figure 3: (a) UMAP of integration of COVID-19 data collection by scMerge2, colored by cell type (left) and studies (right). (b) Evaluation metrics of PCA scores using dataset, protocol and technology as labels, comparing raw logcounts (blue) and scMerge2 normalised results. A lower score indicates better unwanted technical variation removal. (c) Prediction results of disease severity using cell type-specific aggregated expression calculated from raw logcounts (blue) and scMerge2 normalised results (red).

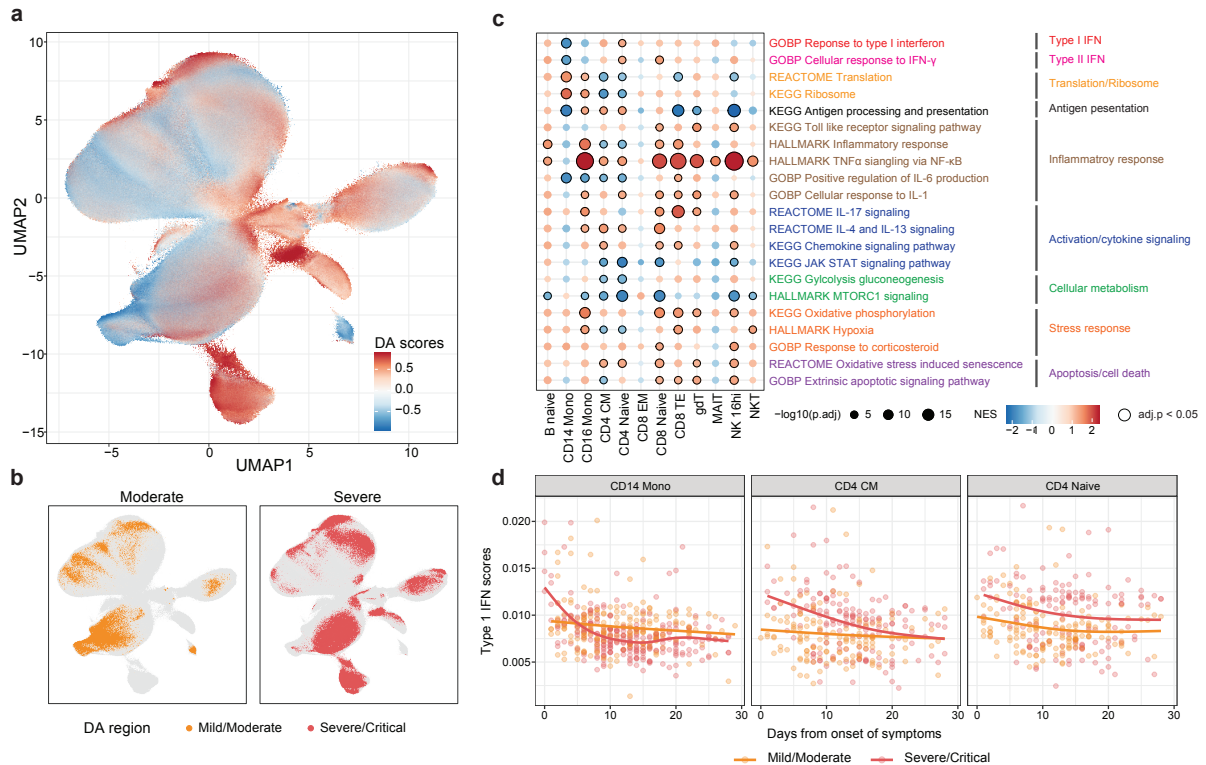


Figure 4: (a-b) UMAP plot of integrated COVID-19 data coloured by (a) differential abundance (DA) probability scores calculated by DA-seq between the moderate and severe patients, where higher scores indicated the cells are more related to severe states; (b) DA region associated with disease severity identified by DA-seq. (c) Enrichment scores of selected pathways for cell-type-specific differential expressed genes distinguished the severity, where a higher score indicates a higher enrichment associated with severe states. The size of the dot indicates the  $-\log_{10}$  adjusted p-value, where black circles indicate statistical significance (adjusted p-value < 0.05); and the colour indicates the normalised enrichment scores of the pathways. (d) Scatter plots showing per-sample gene set signatures (Type-1 IFN) calculated from the scMerge2 normalised data along the days since symptom onset, coloured by disease severity of the patient. CD14 Monocytes, CD4 CM and CD4 Naive are shown as examples.



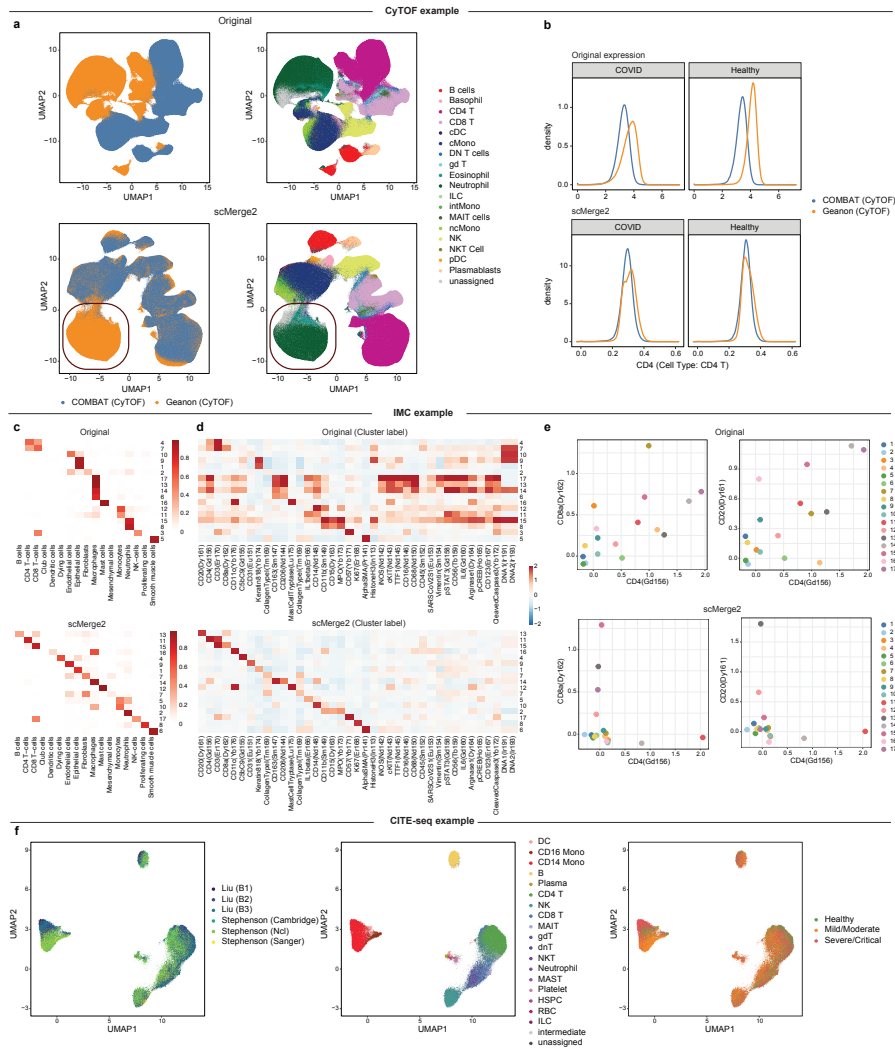
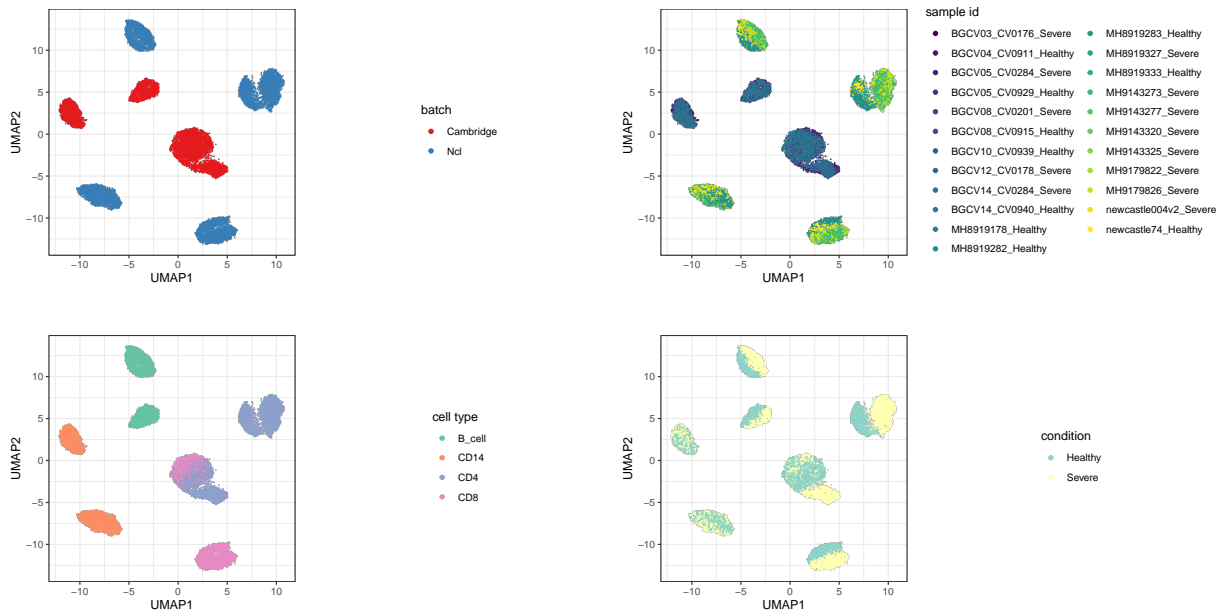
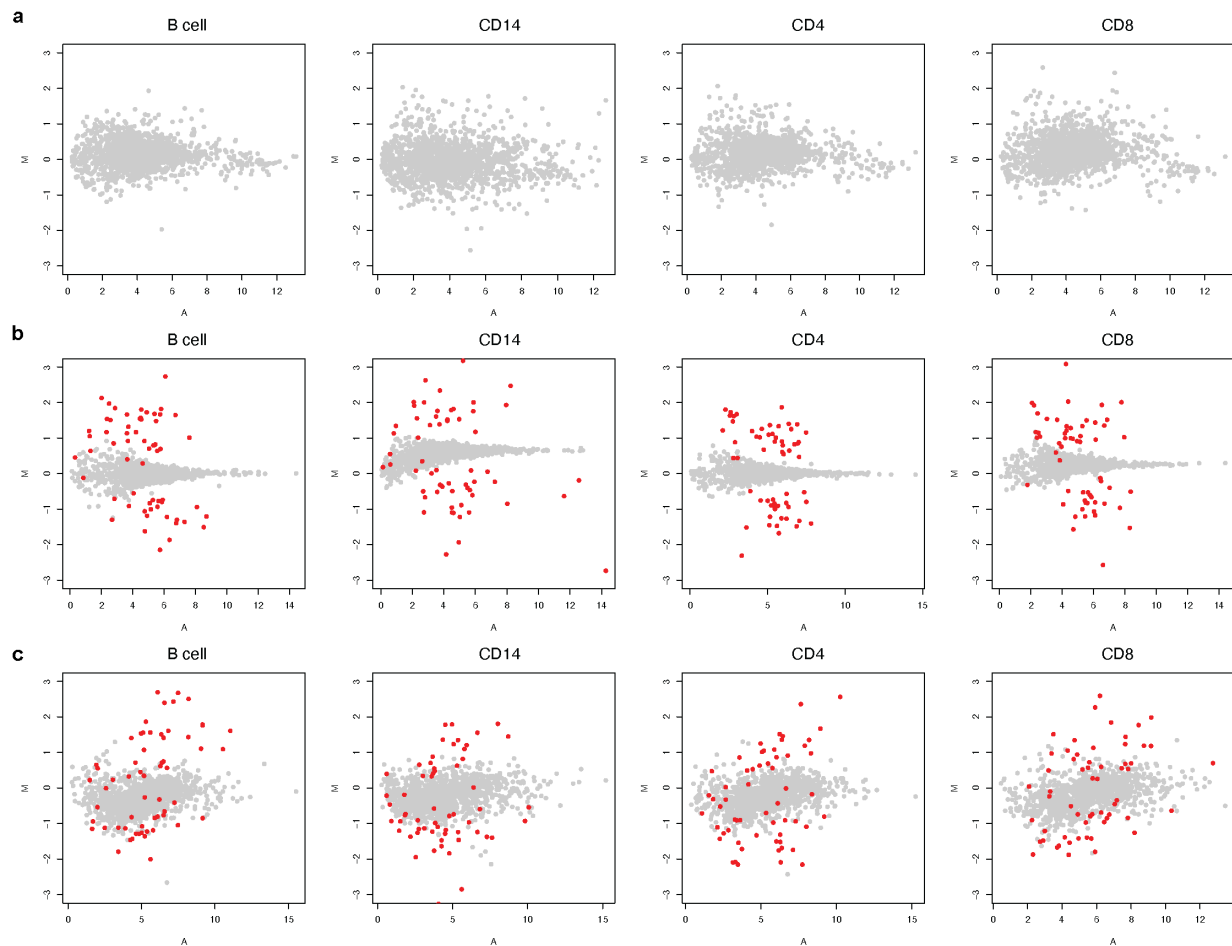


Figure 5: (a) UMAP plots of CyTOF data colored by dataset (left) and cell type (right), for original (first row) and scMerge2 (second row). The red circles highlight the cell types (Neutrophils and Eosinophils) that are unique to Geanon (CyTOF). (b) Density plot of selected markers in specific cell types (CD4 in CD4 T cells), using original expression (first row) and scMerge2 adjusted expression (second row). Within a specific cell type, the distribution of the cell type markers are expected to be similar between two datasets. (c) Heatmaps indicate the clustering results and their fractions of concordance with the original cell type annotation given in [29] for Original (first row) and scMerge2 (second row). Clearer diagonal structure illustrates better concordance. (d) Heatmaps indicate the average marker expression, calculated from cells aggregated by clusters for Original (first row) and scMerge2 (second row). More specific markers for each column and row indicates more distinguished clusters being identified. (e) Scatter plot indicates the average marker expression for each cluster, calculated using Original data (first row) and scMerge2 adjusted data (second row), for two pairs of protein markers: CD4 vs CD8 (first column); and CD4 vs CD20 (second column). Low concordance between the two markers is expected to reveal cluster with specific markers. (f) J-UMAP plot of integrated CITE-seq data colored by dataset (left) and cell type (middle) and severity (right).

## 745 Supplementary Figures



Supplementary Figure S1: UMAP plots of an example of simulated data ( $\logFC = 1.2$ ,  $DS\% = 5\%$ ), coloured by batch, sample id, cell type and condition.





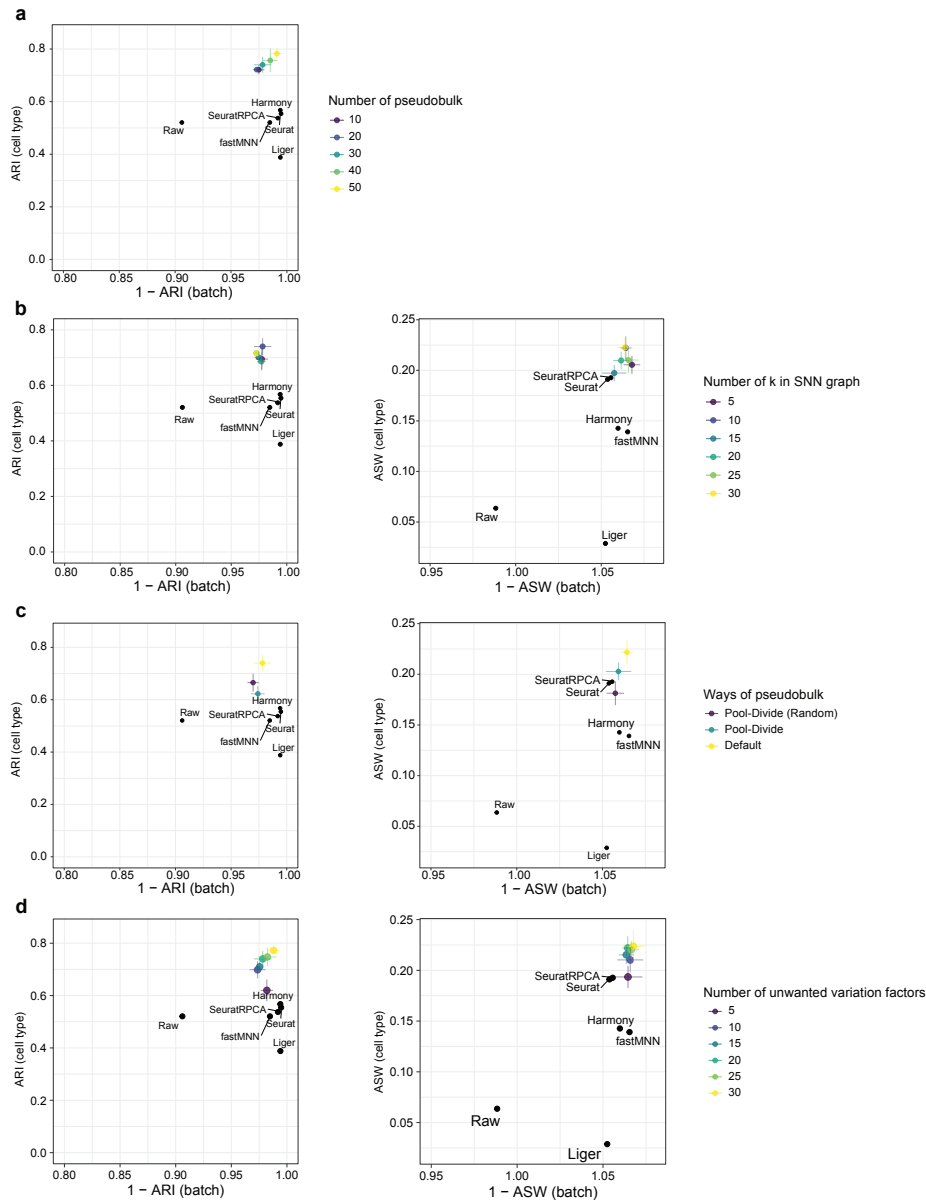
Supplementary Figure S3: F1-score of the differential state (DS) results of four cell types (B cell, CD14, CD4 and CD8) (row) of simulated data, with 5% (1st - 3rd column) and 10% DS genes (4th - 6th column) within each cell type, for scMerge2, Seurat, fastMNN and raw, varying simulated log fold change (logFC) of DS genes (x-axis) and different threshold of adjusted p-value (column).



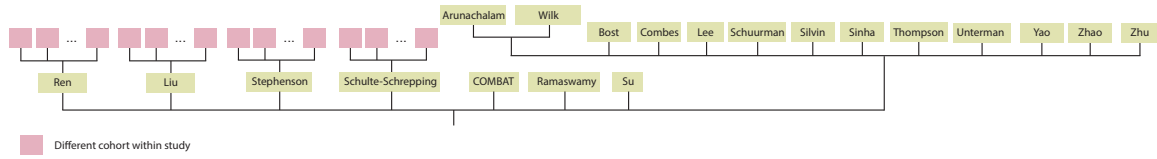
Supplementary Figure S4: FDR of the differential state (DS) results of four cell types (B cell, CD14, CD4 and CD8) (row) of simulated data, with 5% (1st - 3rd column) and 10% DS genes (4th - 6th column) within each cell type, for scMerge2, Seurat, fastMNN and raw, varying simulated log fold change (logFC) of DS genes (x-axis) and different threshold of adjusted p-value (column).



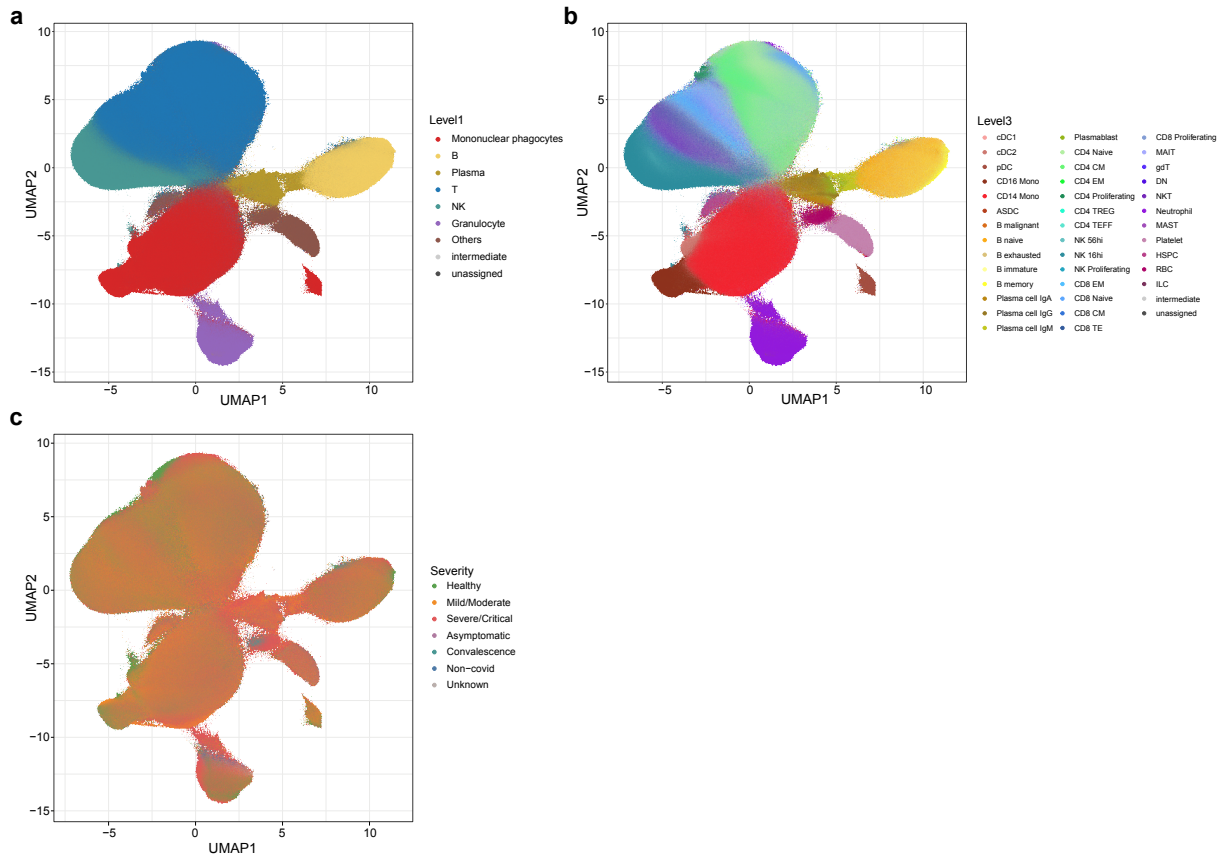
Supplementary Figure S5: TPR of the differential state (DS) results of four cell types (B cell, CD14, CD4 and CD8) (row) of simulated data, with 5% (1st - 3rd column) and 10% DS genes (4th - 6th column) within each cell type, for scMerge2, Seurat, fastMNN and raw, varying simulated log fold change (logFC) of DS genes (x-axis) and different threshold of adjusted p-value (column).



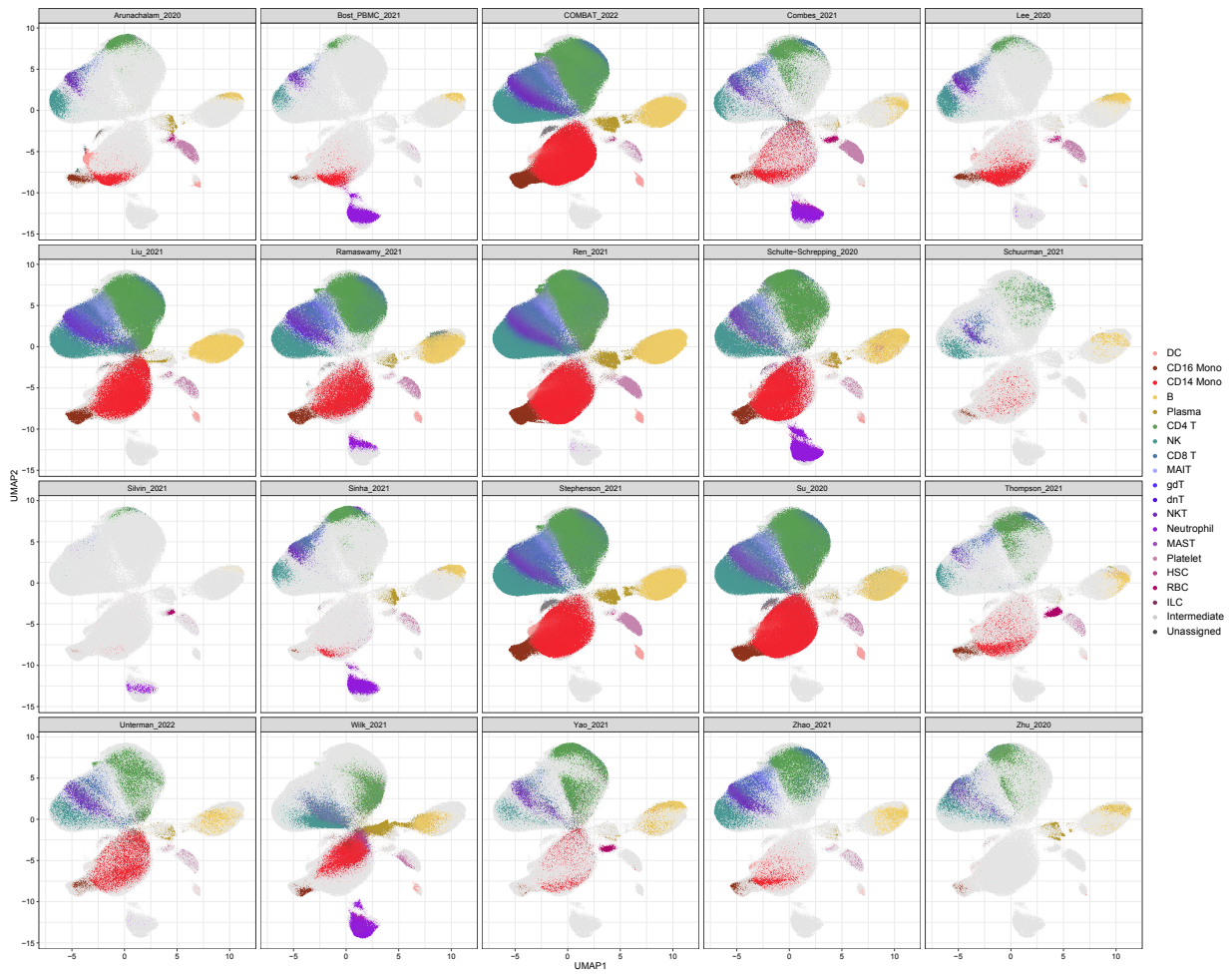
Supplementary Figure S6: Robustness analysis of the tuning parameters of scMerge2 using COVID-19 60k data: Adjusted rand index (ARI) (left panel), where x-axis indicates 1 minus batch ARI and y-axis indicates cell type ARI; Average silhouette width (ASW), where x-axis indicates 1 minus batch ASW and y-axis indicates cell type ASW (right panel), when varying (a) the number of pseudobulk constructed (10, 20, 30 (default), 40, 50); (b) the number of k used in SNN graph (5, 10 (default), 15, 20, 25, 30); (c) different methods to construct pseudobulk. (d) Number of unwanted variation factors (5, 10, 15, 20 (default), 25, 30).



Supplementary Figure S7: Hierarchical merging strategy for COVID-19 scRNA-seq data collection.

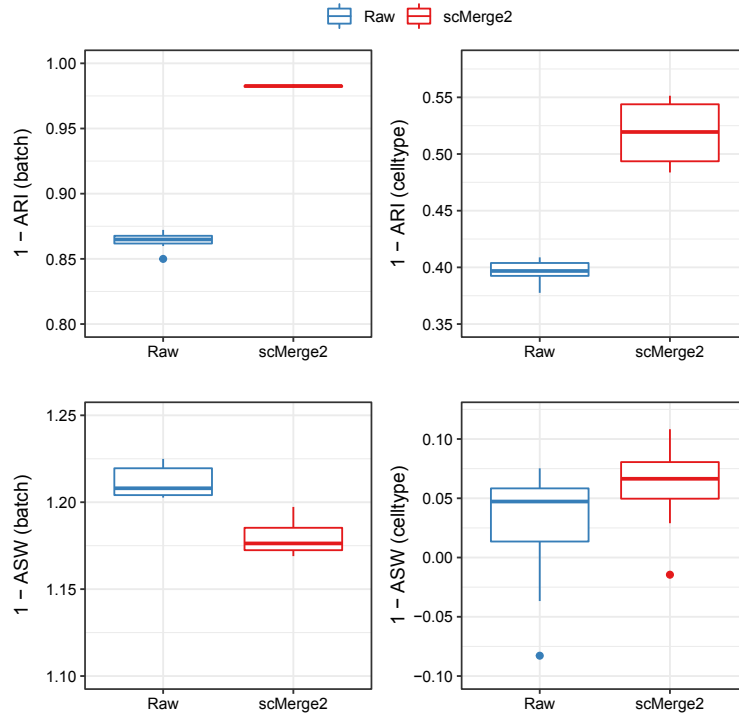


Supplementary Figure S8: UMAP of integration of COVID-19 data collection after scMerge2 integration, coloured by (a) level 1 cell type annotation; (b) level 3 cell type annotation and (c) severity.

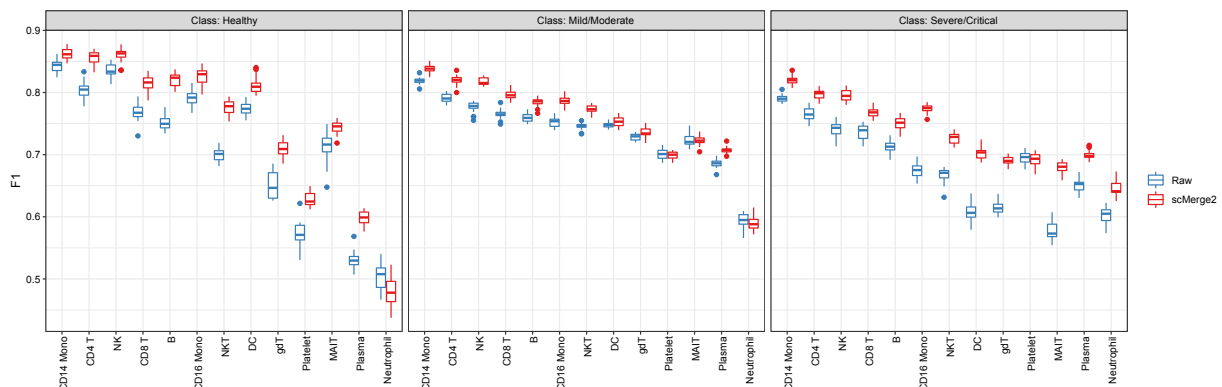


Supplementary Figure S9: UMAP of integration of COVID-19 data collection after scMerge2 integration, coloured by cell type (level 2) and faceted by dataset.

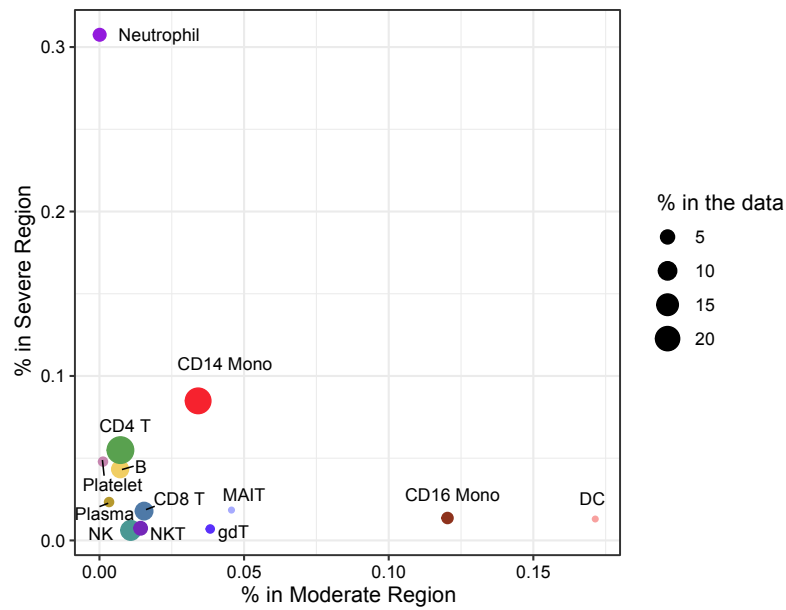




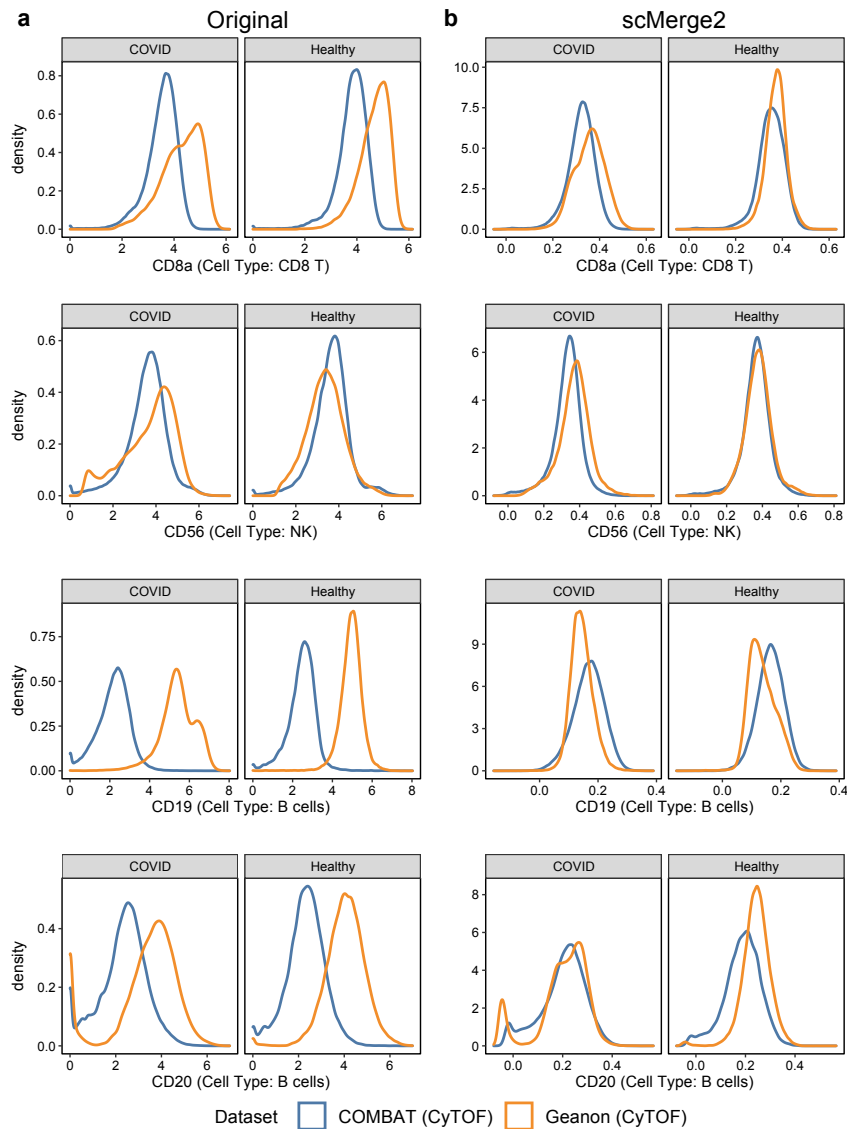
Supplementary Figure S10: Boxplots of evaluation metrics of COVID-19 scRNA-seq data collection for scMerge2-h (data merged in a hierarchical manner) and Raw, where the first row indicates the results of adjusted rand index (ARI): 1 minus batch ARI (left) and cell type ARI (right); the second row indicates the results of Average silhouette width (ASW): 1 minus batch ASW (left) and cell type ASW (right). For all of the four metrics, higher value indicates better performance. Since the size of this data collection is large, we subsampled 1% of the cells to calculate the metrics, and repeated this procedure 10 times.



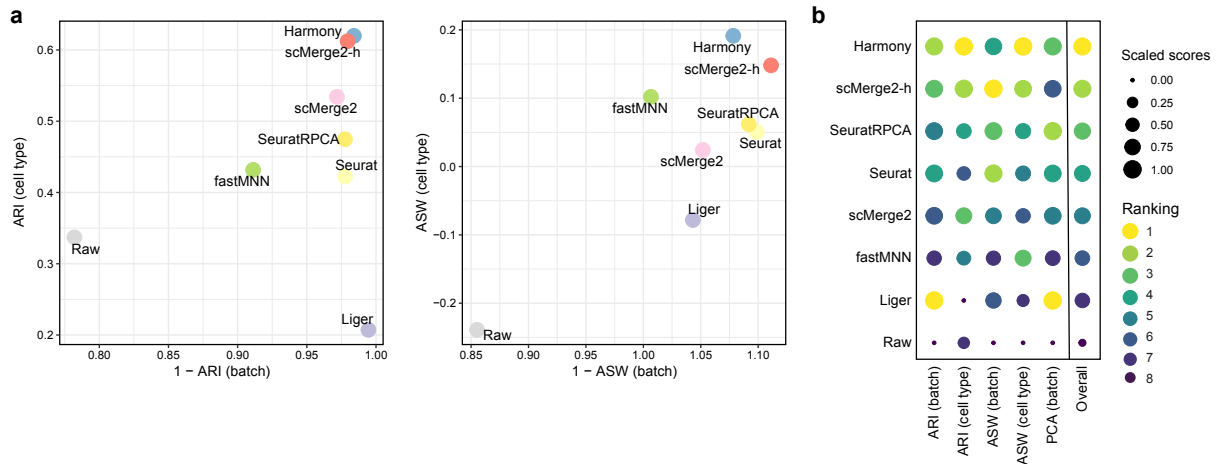
Supplementary Figure S11: Prediction results of disease severity using cell type-specific aggregated expression calculated from raw logcounts (blue) and scMerge2 adjusted results (red), evaluated by class-specific F1 scores.



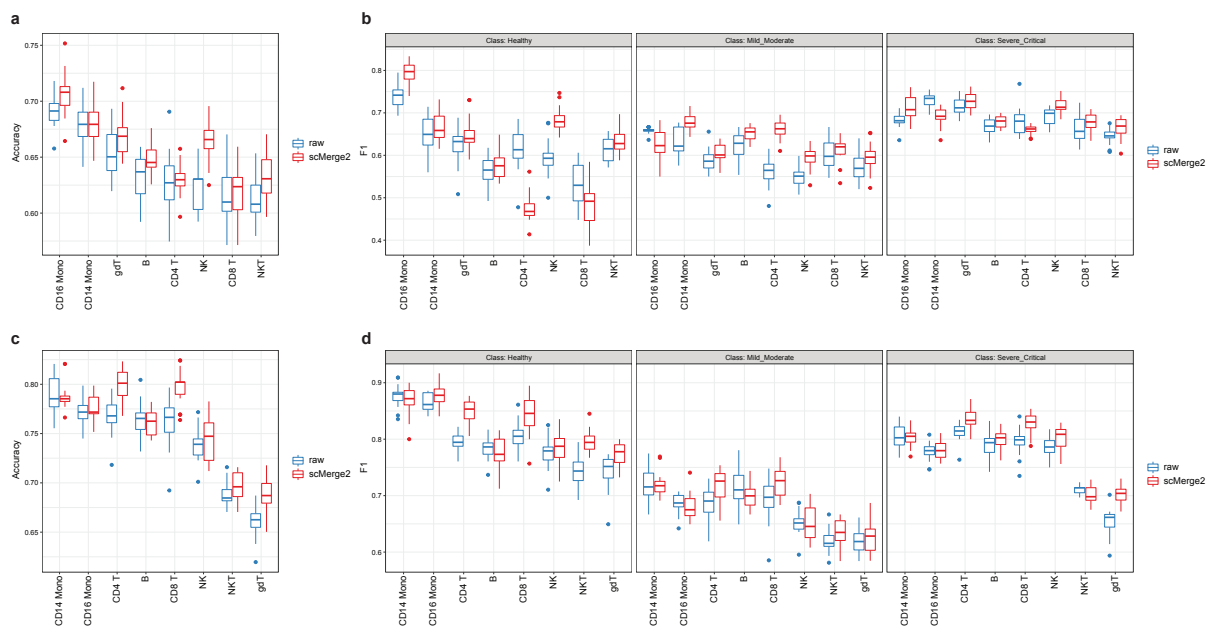
Supplementary Figure S12: Scatter plot shows the proportion of cells in Moderate region (x-axis) vs the proportion of cells in Severe region, determined by Daseq. The size of each point indicates the cell type proportion in the all data (Only cell types that have more than 1% in the data are shown).



Supplementary Figure S13: Density plot of selected marker in specific cell type: CD8a in CD8 T cells; CD56 in NK cells; CD19 in B cells and CD20 in B cells, using (a) original expression and (b) scMerge2 adjusted expression. Within a specific cell type, the distribution of the cell type marker is expected to be similar between two datasets.



Supplementary Figure S14: CITE-seq data example: (a) Scatter plots of evaluation metrics of ADT data integration of a 200k cells subset of two COVID-19 studies (Liu and Stephenson) for scMerge2, scMerge2-h (data merged in a hierarchical manner), Seurat, Seurat (RPCA), Harmony, fastMNN, Liger and Raw: Adjusted rand index (ARI) (left panel), where x-axis indicates 1 minus batch ARI and y-axis indicates cell type ARI; Average silhouette width (ASW), where x-axis indicates 1 minus batch ASW and y-axis indicates cell type ASW (right panel). (b) Dot plots indicates the ranking of the data integration methods in terms of 5 different evaluation metrics. The size of the dot indicates the scaled scores, which are obtained from the min-max scaling of the original values. The overall ranking is ranked based on the average ranking of the five evaluation metrics.



Supplementary Figure S15: CITE-seq data example: Prediction results of disease severity using cell type-specific aggregated expression calculated from raw logcounts (blue) and scMerge2 normalised results (red), using (a-b) ADT expression and (c-d) RNA expression.