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

Yingxin Lin[1,2,3,4], Yue Cao[1,2,3,4], Elijah Willie[1], Ellis Patrick[1,3,4,5], and Jean Y.H. Yang[1,2,3,4*]

[1]Sydney Precision Data Science Centre, The University of Sydney, NSW, Australia.

[2]Charles Perkins Centre, The University of Sydney, NSW, Australia.

[3]School of Mathematics and Statistics, The University of Sydney, NSW, Australia.

[4]Laboratory of Data Discovery for Health Limited (D24H), Science Park, Hong Kong SAR, China

[5]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

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

Introduction

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

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

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

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

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

Results

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

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

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

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

scMerge2 outperforms existing integration methods in detecting differential expression.

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

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

Next, we investigate the performance of the adjusted matrix in identifying genes that are differentially expressed between two conditions (termed as differential state (DS) analysis by [23]) through a simulation study. We generated synthetic single-cell datasets with two batches and multiple samples from two conditions using a simulation framework that extended from scDesign3 model [24], with known ground truth DS genes (**Supp Fig. S1-S2**) (See Methods). Cell-type-specific DS analysis was performed using the limma-trend algorithm [25] on the sample-wise aggregated data by taking the mean of the log-transformed or adjusted data. By simulating data with different log fold change ($1.1 \sim 2$) and proportions of DS genes (5% and 10%), we find that scMerge2 substantially outperforms the other two data integration methods that also return adjusted matrices in detecting DS genes (**Fig. 2c** and **Supp Fig. S3**). scMerge2 has much lower FDR than fastMNN and Seurat, and higher TPR compared to the unadjusted data (**Supp Fig. S4-S5**), illustrating that scMerge2 outputs an adjusted matrix with less unwanted variation for single-cell downstream analysis.

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

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

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

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

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

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

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

(**Supp Fig. S12**).

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

scMerge2 is versatile to other single-cell platforms.

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

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

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

Discussion

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

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

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

batches within one study or studies with similar size first. In contrast, a priori information such as sequencing platforms or cell extraction techniques can be used in knowledge-guided order of integration. Noted that the hierarchical data integration design can be broadly classified into two strategies [33], balanced trees and concatenating approaches. The balanced tree approach integrates between pairs of datasets at different levels of the tree, and the procedure is continued until all data is merged. The concatenating approach sequentially integrates datasets, therefore for n data sets, this will need $n - 1$ steps of integration. Previous studies have found that normalisation results are very similar between the two types of integration tree structures [33]. The key difference between the approach is computational burden with the concatenating approach being more computational intensive. Currently, the scMerge2 approach is closer to the balance approaches allowing for many datasets to be added simultaneously at each level.

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

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

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

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

Methods

scMerge2

Single-cell grouping within one batch

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

- Clustering-based grouping: This is performed by default when no cell type label is used as input. Firstly, the top 2000 highly variables genes (HVG) are selected using *getTopHVGs* in the *scran* R Package, using batch information as block information. For data like CyTOF and ADT from CITE-seq data, this step will be skipped and all features will be used in the next step. Next, within each batch, instead of using k-means clustering as in the previous version, we construct a shared nearest neighbour graph on the gene expression of the HVGs, with a default number of neighbours of 10, followed by louvain clustering. This therefore relieves the need of predefining the number of clusters that is required in our previous version.

- Reference-based grouping: This refers to the use of supervised cell type classification to predict or annotate the cell types using one or more reference datasets. This ensures the cell-type annotations are consistent among datasets. Cell type classification algorithms (e.g. scClassify [34] and SingleR [35]) can also be used and the reference dataset can be external datasets with similar cell types to the data to be integrated. This approach unifies cell type annotation across all datasets and eliminates the need for clustering and cell type annotation after data integration. It is noted that this approach is used in the COVID-19 case study to integrate the data collection of 20 datasets.

Pseudo-bulk construction

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

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

- when count data are not available for all datasets, for each cell type grouping, we randomly assign the cells into k subsets and take the gene-wise average of each subset as one pseudo-bulk. This therefore results with k pseudo-bulk for one cell type grouping.

- when counts data are available for all data, we can perform a similar pool-and-divide strategy that is proposed in RUVIII-NB [36]. Here, we can have two strategies in pooling the cells: (1) assign the cells based on library size; (2) randomly assign the cells into k subsets. Then we gene-wisely take the sum of the counts for each subset and generate the counts data following a negative binomial distribution. While the pseudobulk matrix generated by this strategy is able to maintain the gene mean-variance relationship [36], we find that this approach does not improve the quality of data integration in scMerge2 (Supp Fig. S6).

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

Pseudo-replicates identification across batches in scMerge2

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

- The pseudo-replicates identification is based on the pseudo-bulk matrix to reduce the computational time;

- For data with multiple conditions (or other observed biological factors), scMerge2 allows the MNC graph to be constructed within each condition to preserve the biological variation. Note that this strategy can only be used when the batches to be merged have at least one common condition and can only be performed in the condition with multiple batches.

Estimation of RUVIII model using pseudo-bulk

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

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

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

- Step i: estimate α via the first k right singular vectors of Singular Value Decomposition (SVD) on $R_M Z$, where $R_M = 1 - M(M^T M)^{-1} M^T$, with the replicate matrix $M \in R^{C\times N}$, N indicates the number of types of pseudo-replicates;

- 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 submatrix of α, which columns include only the genes that belongs to single-cell stably expressed genes (SEG) with number of genes as G_s (SEG selection and evaluation can be found in [38]);

- Step iii: adjust the matrix by subtracting the estimated unwanted variation component: $\hat{Z}_{C\times G} = Z_{C\times G} - \hat{W}_{C\times k}\hat{\alpha}_{k\times G}$.

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

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

Hierarchical merging

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

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

Data collection and preprocessing

COVID-19 scRNA-seq data collection

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

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

COVID-19 200k CITE-Seq data collection (COVID-19 200k)

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

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

COVID-19 60k data collection (COVID-19 60k)

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

COVID-19 CyTOF data collection

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

COVID-19 IMC data collection

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

Evaluation

Part I - Simulation

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

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

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

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

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

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

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

Part II - Real data comparison

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

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

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

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

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

Evaluation on IMC data collection. We applied scMerge2 to perform data integration of the 23 samples. This is achieve by first filtering and selecting the data using the 38 markers specified in the original publication [29] and removing all undefined cell types (i.e. cells having cell type annotation as "nan"). Next, considering sample labels as batch information, we applied scMerge2 with settings $k_{\text{RUV}} = 2$, $k_{\text{pseudoBulk}} = 5$, $k_{\text{celltype}} = 20$, using all markers as negative control genes and highly variable genes. Thirdly, unsupervised clustering was performed on both the unnormalised and scMerge2 normalised datasets using the FlowSOM [30] algorithm with 17 clusters. The Adjusted Rand Index (ARI) was used to compared the concordance between this unsupervised clustering with the manually curated cell types in the original manuscript [29]. The results are visualised using heatmaps showing the average marker abundance in the cell types. Average marker abundance were generated after scaling the marker expression by computing the ratio of the mean of each marker and its standard deviation.

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

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

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

Evaluation metrics

We used three metrics to assess the performance of data integration results from different methods. Details of the evaluation metrics are described as follows:

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

Benchmarking methods

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

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

(ii) SeuratRPCA. Similar to Seurat (CCA), within each batch, we first found the variable features, with an addition PCA step performed. After integration features were selected, *FindIntegrationAnchors()* was performed with reduction set as "rpca". Lastly, *IntegrateData()* was performed to obtain the integrated data.

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

(iv) Liger. R package *rliger v1.0.0* [44] was used. Online integrative nonnegative matrix factorization was performed to obtain the batch corrected embedding following the tutorial (`https://github.com/welch-lab/liger/blob/master/vignettes/online_iNMF_tutorial.html`), where we first ran *selectGenes()* to select the features, *scaleNotCenter()* to scale the features, and *online_iNMF()* with *miniBatch_size = 5000* and *max.epochs = 5*.

(v) Harmony. R package *Harmony v0.1.0* [18] was used. The PCA space returned by *runPCA()* of R package *scater* was used as input, and then *HarmonyMatrix()* was performed with *do_pca = FALSE* to retain the batch corrected embedding.

COVID-19 downstream analysis

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

Differential abundance (DA) analysis was performed on the cells from mild/moderate and severe/critical samples using DA-seq [26]. The top 30 PCs derived from the adjusted expression data were used as input for the algorithm to calculate the DA scores. A range of k values from 50 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.

Data availability

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

Code availability

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

Acknowledgments

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

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

Author contributions

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

Conflict of interest

The authors declare that they have no conflict of interest.

References

[1] Byungjin Hwang, Ji Hyun Lee, and Duhee Bang. "Single-cell RNA sequencing technologies and bioinformatics pipelines". In: *Experimental & molecular medicine* 50.8 (2018), pp. 1–14.

[2] Ting Zhang et al. "Progress and applications of mass cytometry in sketching immune landscapes". In: *Clinical and Translational Medicine* 10.6 (2020), e206.

[3] Felix J Hartmann and Sean C Bendall. "Immune monitoring using mass cytometry and related high-dimensional imaging approaches". In: *Nature Reviews Rheumatology* 16.2 (2020), pp. 87–99.

[4] Aviv Regev et al. "The Human Cell Atlas". en. In: *Elife* 6 (Dec. 2017).

[5] Junyue Cao et al. "A human cell atlas of fetal gene expression". en. In: *Science* 370.6518 (Nov. 2020).

[6] Silvia Domcke et al. "A human cell atlas of fetal chromatin accessibility". en. In: *Science* 370.6518 (Nov. 2020).

[7] Orit Rozenblatt-Rosen et al. "The Human Tumor Atlas Network: Charting Tumor Transitions across Space and Time at Single-Cell Resolution". en. In: *Cell* 181.2 (Apr. 2020), pp. 236–249.

[8] Mengwei Li et al. "DISCO: a database of Deeply Integrated human Single-Cell Omics data". In: *Nucleic Acids Research* 50.D1 (2022), pp. D596–D602.

[9] Yuan Tian et al. "Single-cell immunology of SARS-CoV-2 infection". en. In: *Nat. Biotechnol.* 40.1 (Jan. 2022), pp. 30–41.

[10] Rachel L Belote et al. "Human melanocyte development and melanoma dedifferentiation at single-cell resolution". en. In: *Nat. Cell Biol.* 23.9 (Sept. 2021), pp. 1035–1047.

[11] Viktor Petukhov et al. "Case-control analysis of single-cell RNA-seq studies". en. Mar. 2022.

[12] Malte D Luecken et al. "Benchmarking atlas-level data integration in single-cell genomics". In: *Nature methods* 19.1 (2022), pp. 41–50.

[13] Yuhan Hao et al. "Integrated analysis of multimodal single-cell data". In: *Cell* 184.13 (2021), pp. 3573–3587.

[14] Matthew Amodio et al. "Exploring single-cell data with deep multitasking neural networks". In: *Nature methods* 16.11 (2019), pp. 1139–1145.

[15] Brian Hie, Bryan Bryson, and Bonnie Berger. "Efficient integration of heterogeneous single-cell transcriptomes using Scanorama". In: *Nature biotechnology* 37.6 (2019), pp. 685–691.

[16] Bin Zou et al. "deepMNN: deep learning-based single-cell RNA sequencing data batch correction using mutual nearest neighbors". In: *Frontiers in Genetics* (2021), p. 1441.

[17] Krzysztof Polański et al. "BBKNN: fast batch alignment of single cell transcriptomes". In: *Bioinformatics* 36.3 (2020), pp. 964–965.

[18] Ilya Korsunsky et al. "Fast, sensitive and accurate integration of single-cell data with Harmony". In: *Nature methods* 16.12 (2019), pp. 1289–1296.

[19] Romain Lopez et al. "Deep generative modeling for single-cell transcriptomics". In: *Nature methods* 15.12 (2018), pp. 1053–1058.

[20] Chenling Xu et al. "Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models". In: *Molecular systems biology* 17.1 (2021), e9620.

[21] Xiangjie Li et al. "Deep learning enables accurate clustering with batch effect removal in single-cell RNA-seq analysis". In: *Nature communications* 11.1 (2020), pp. 1–14.

[22] Yingxin Lin et al. "scMerge leverages factor analysis, stable expression, and pseudoreplication to merge multiple single-cell RNA-seq datasets". In: *Proceedings of the National Academy of Sciences* 116.20 (2019), pp. 9775–9784.

[23] Helena L Crowell et al. "Muscat detects subpopulation-specific state transitions from multi-sample multi-condition single-cell transcriptomics data". In: *Nature communications* 11.1 (2020), pp. 1–12.

[24] Dongyuan Song et al. "A unified framework of realistic in silico data generation and statistical model inference for single-cell and spatial omics". In: *bioRxiv* (2022). DOI: 10.1101/2022.09.20.508796. eprint: https://www.biorxiv.org/content/early/2022/09/22/2022.09.20.508796.full.pdf. URL: https://www.biorxiv.org/content/early/2022/09/22/2022.09.20.508796.

[25] Matthew E Ritchie et al. "limma powers differential expression analyses for RNA-sequencing and microarray studies". In: *Nucleic acids research* 43.7 (2015), e47–e47.

[26] Jun Zhao et al. "Detection of differentially abundant cell subpopulations in scRNA-seq data". In: *Proceedings of the National Academy of Sciences* 118.22 (2021), e2100293118.

[27] Matthew L Meizlish et al. "A neutrophil activation signature predicts critical illness and mortality in COVID-19". In: *Blood advances* 5.5 (2021), pp. 1164–1177.

[28] Can Liu et al. "Time-resolved systems immunology reveals a late juncture linked to fatal COVID-19". In: *Cell* 184.7 (2021), pp. 1836–1857.

[29] André F Rendeiro et al. "The spatial landscape of lung pathology during COVID-19 progression". In: *Nature* 593.7860 (2021), pp. 564–569.

[30] Sofie Van Gassen et al. "FlowSOM: Using self-organizing maps for visualization and interpretation of cytometry data". In: *Cytometry Part A* 87.7 (2015), pp. 636–645.

[31] Ricard Argelaguet et al. "MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data". In: *Genome biology* 21.1 (2020), pp. 1–17.

[32] Van Hoan Do and Stefan Canzar. "A generalization of t-SNE and UMAP to single-cell multimodal omics". In: *Genome Biology* 22.1 (2021), pp. 1–9.

[33] Taiyun Kim et al. "hRUV: Hierarchical approach to removal of unwanted variation for large-scale metabolomics data". In: *bioRxiv* (2020).

[34] Yingxin Lin et al. "scClassify: sample size estimation and multiscale classification of cells using single and multiple reference". In: *Molecular systems biology* 16.6 (2020), e9389.

[35] Dvir Aran et al. "Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage". In: *Nature immunology* 20.2 (2019), pp. 163–172.

[36] Agus Salim et al. "RUV-III-NB: Normalization of single cell RNA-seq Data". In: *bioRxiv* (2021).

[37] Ramyar Molania et al. "A new normalization for Nanostring nCounter gene expression data". In: *Nucleic acids research* 47.12 (2019), pp. 6073–6083.

[38] Yingxin Lin et al. "Evaluating stably expressed genes in single cells". In: *GigaScience* 8.9 (2019), giz106.

[39] Davis J McCarthy et al. "Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R". In: *Bioinformatics* 33.8 (2017), pp. 1179–1186.

[40] Daniel Geanon et al. "A streamlined whole blood CyTOF workflow defines a circulating immune cell signature of COVID-19". In: *Cytometry Part A* 99.5 (2021), pp. 446–461.

[41] David J. Ahern et al. "A blood atlas of COVID-19 defines hallmarks of disease severity 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.

[42] Tim Stuart et al. "Comprehensive integration of single-cell data". In: *Cell* 177.7 (2019), pp. 1888–1902.

[43] Laleh Haghverdi et al. "Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors". In: *Nature biotechnology* 36.5 (2018), pp. 421–427.

[44] Chao Gao et al. "Iterative single-cell multi-omic integration using online learning". In: *Nature biotechnology* 39.8 (2021), pp. 1000–1007.

[45] Gennady Korotkevich et al. "Fast gene set enrichment analysis". In: *BioRxiv* (2021), p. 060012.

744 # Figure legends

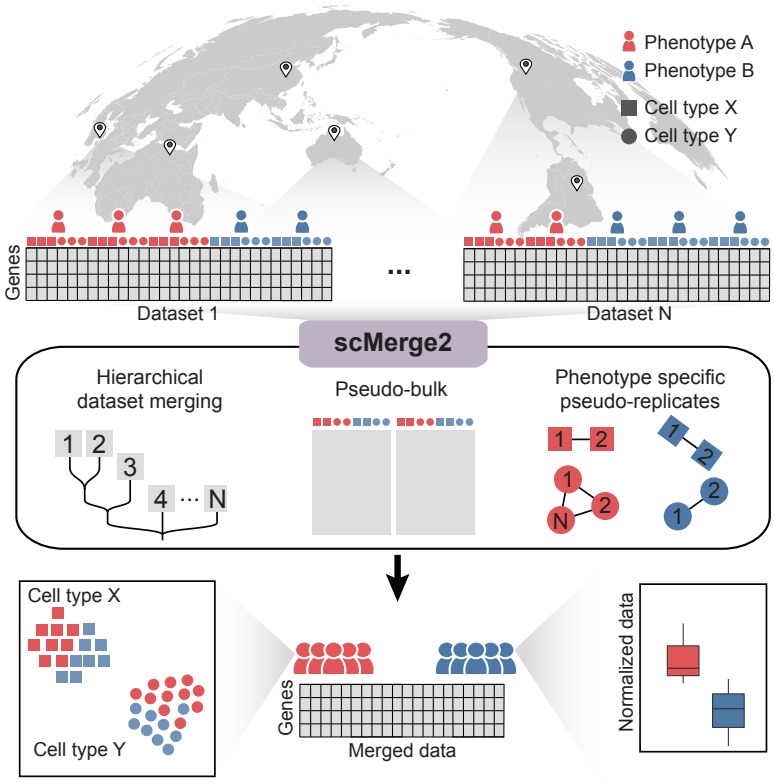


Figure 1: Overview of scMerge2: This new scalable algorithm uses (i) hierarchical integration to capture both local and global variation; (ii) pseudo-bulk construction to reduce computational load; and (iii) phenotype specific pseuduo-replicate, and outputs adjusted expression matrix for millions of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

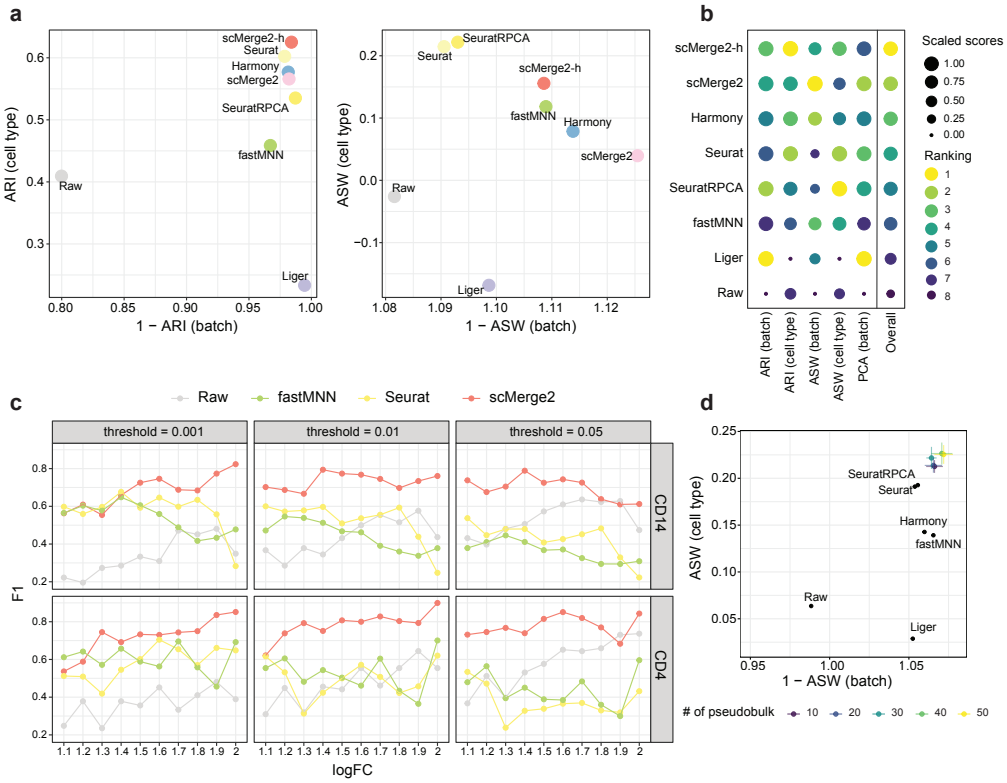


Figure 2: (a) Scatter plots of evaluation metrics of data integration of a 200k cells subset of two COVID-19 studies (Liu and Stephenson) for scMerge2, scMerge2-h (data merged in a hierarchical manner), Seurat, Seurat (RPCA), Harmony, fastMNN, Liger and Raw: Adjusted rand index (ARI) (left panel), where 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.

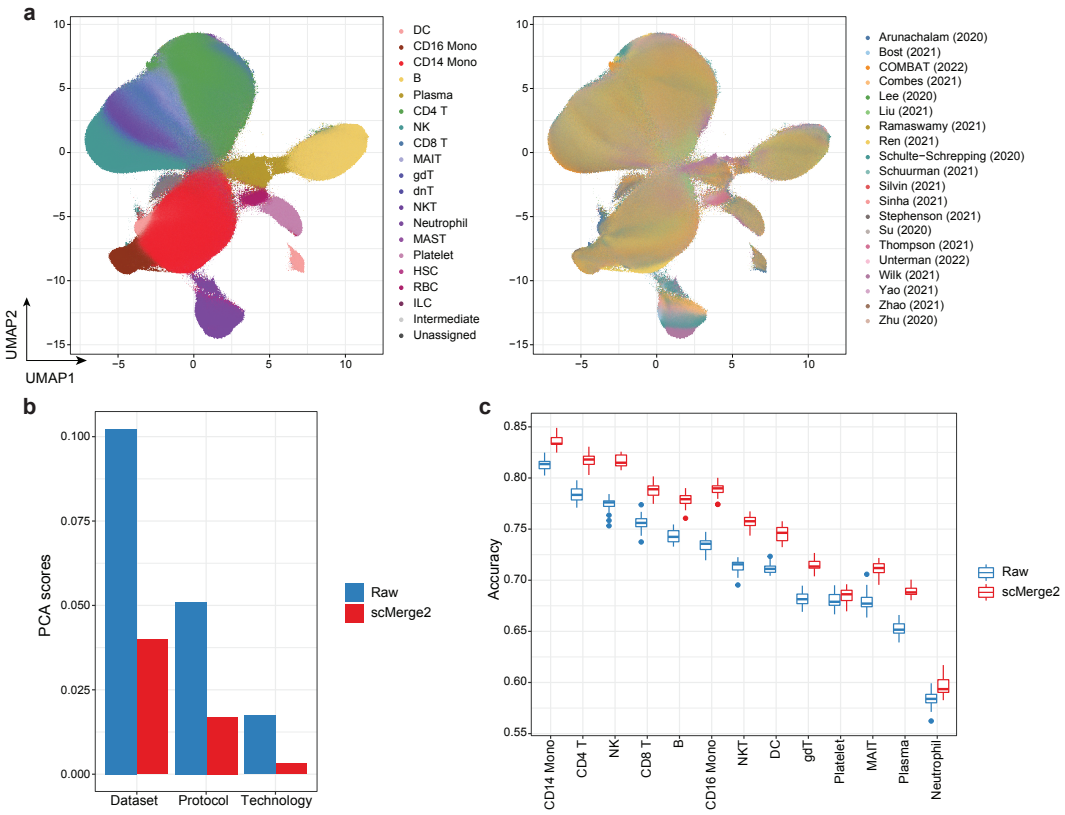


Figure 3: (a) UMAP of integration of COVID-19 data collection by scMerge2, colored by cell type (left) and studies (right). (b) Evaluation metrics of PCA scores using dataset, protocol and technology as labels, comparing raw logcounts (blue) and scMerge2 normalised results. A lower score indicates better 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).

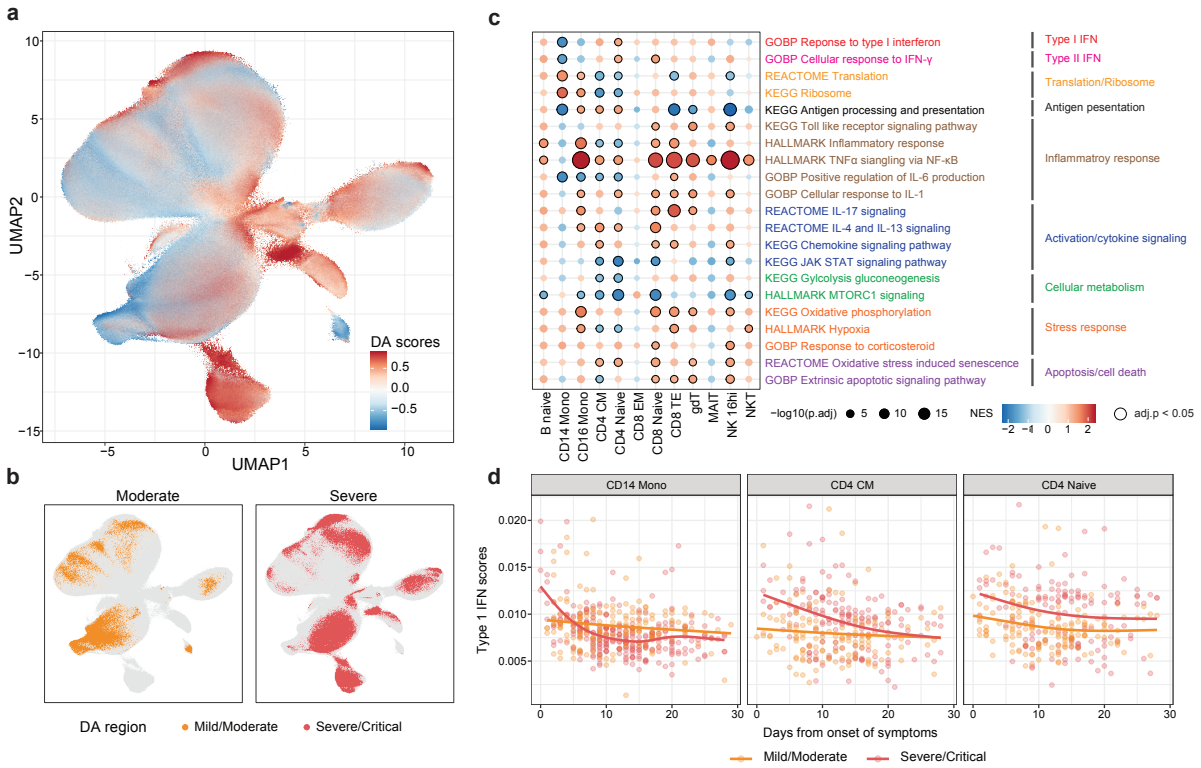


Figure 4: (a-b) UMAP plot of integrated COVID-19 data coloured by (a) differential abundance (DA) probability scores calculated by DA-seq between the moderate and severe patients, where higher scores indicated the cells are more related to severe states; (b) DA region associated with disease severity 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 -log10 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.

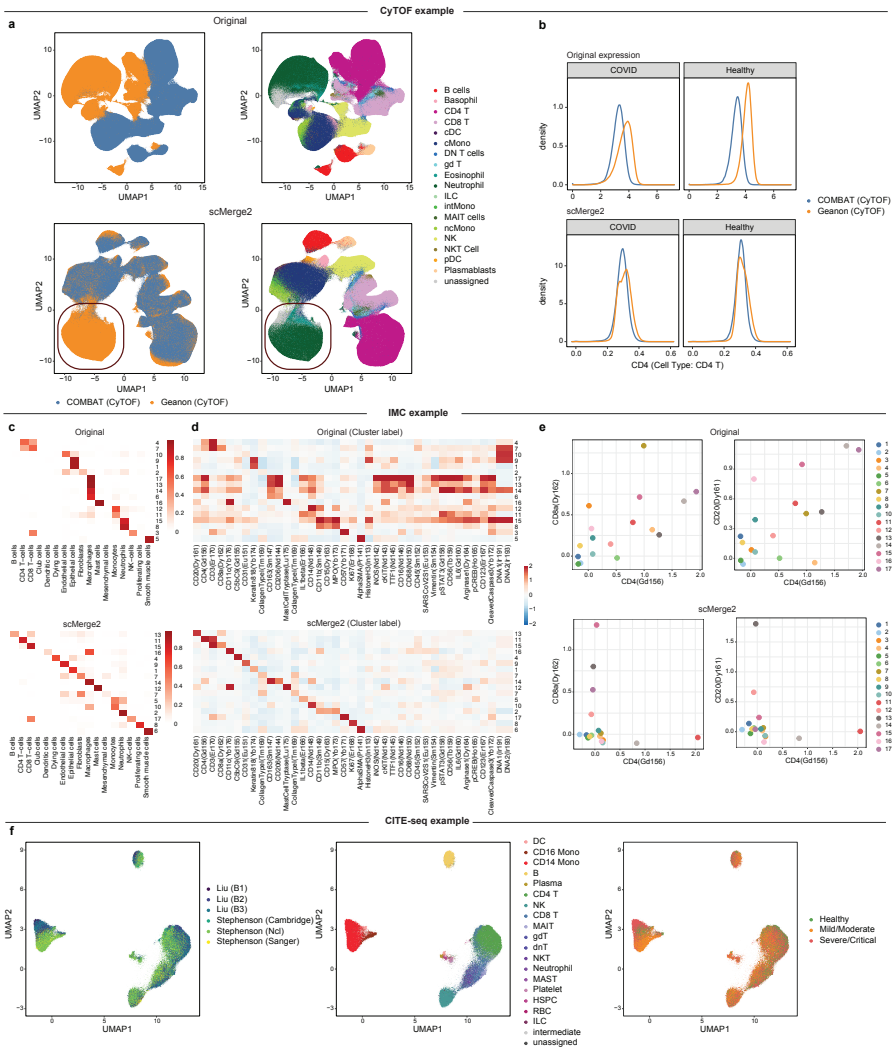
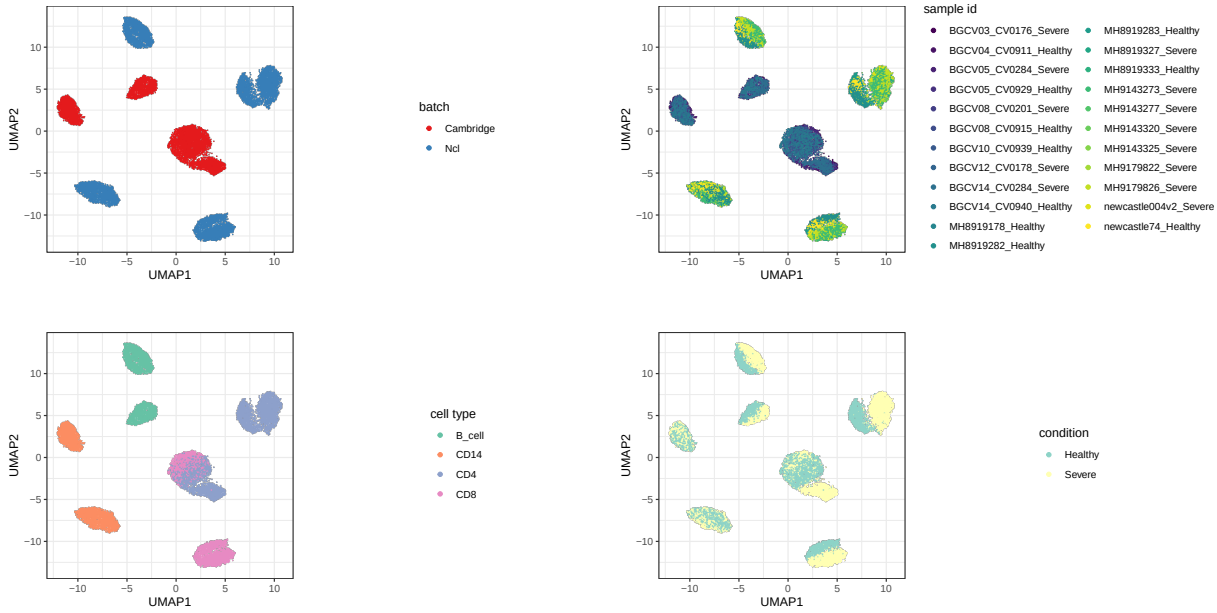
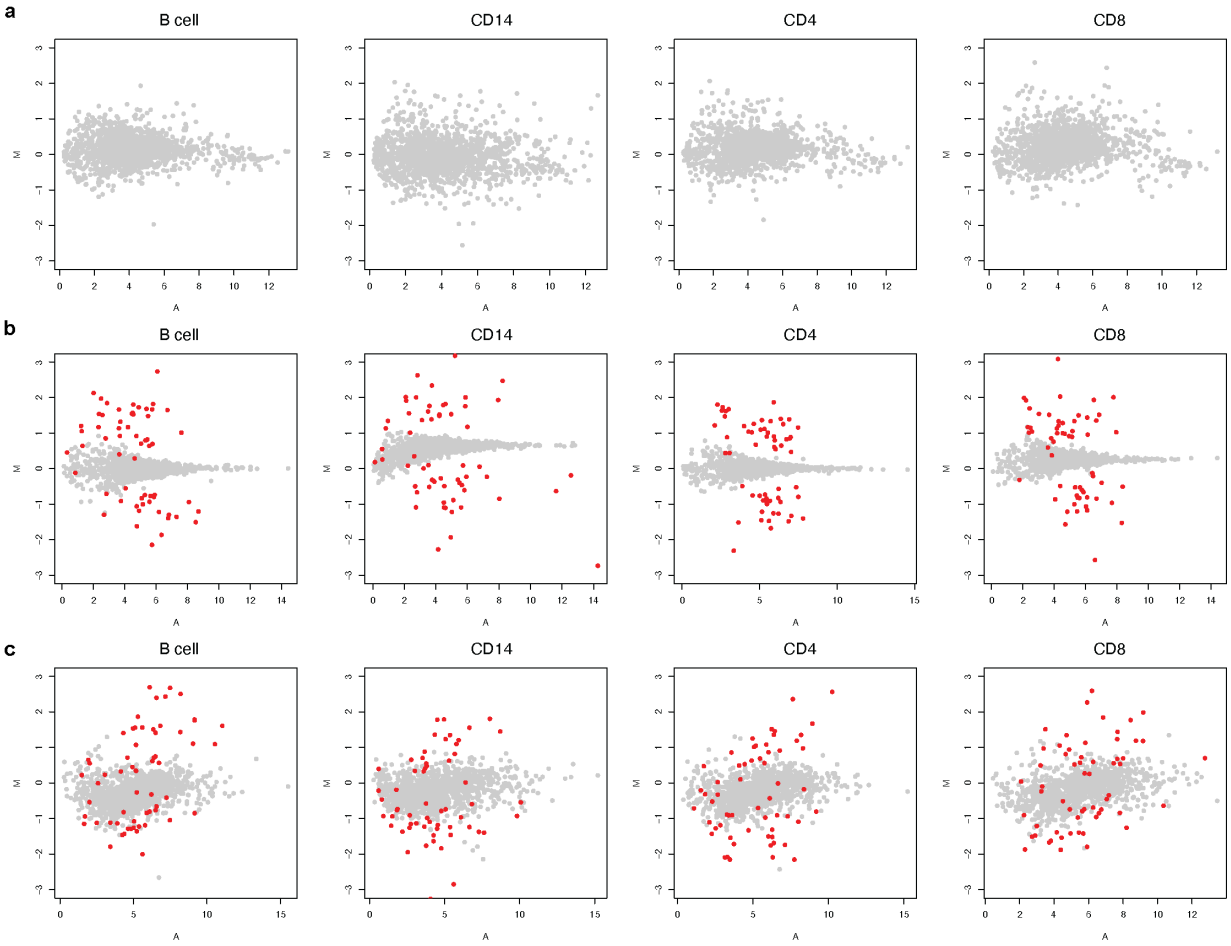


Figure 5: (a) UMAP plots of CyTOF data colored by dataset (left) and cell type (right), for original (first row) and scMerge2 (second row). The red circles highlight the cell types (Neutrophils and Eosinophils) that are unique to Geanon (CyTOF). (b) Density plot of selected markers in specific cell types (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).

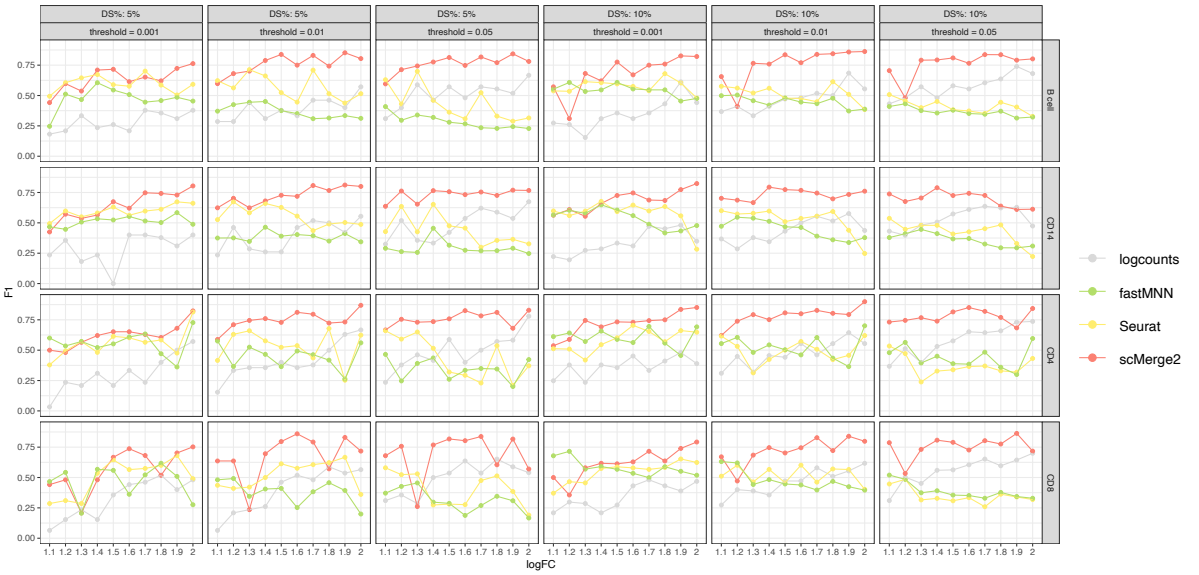
Supplementary Figures



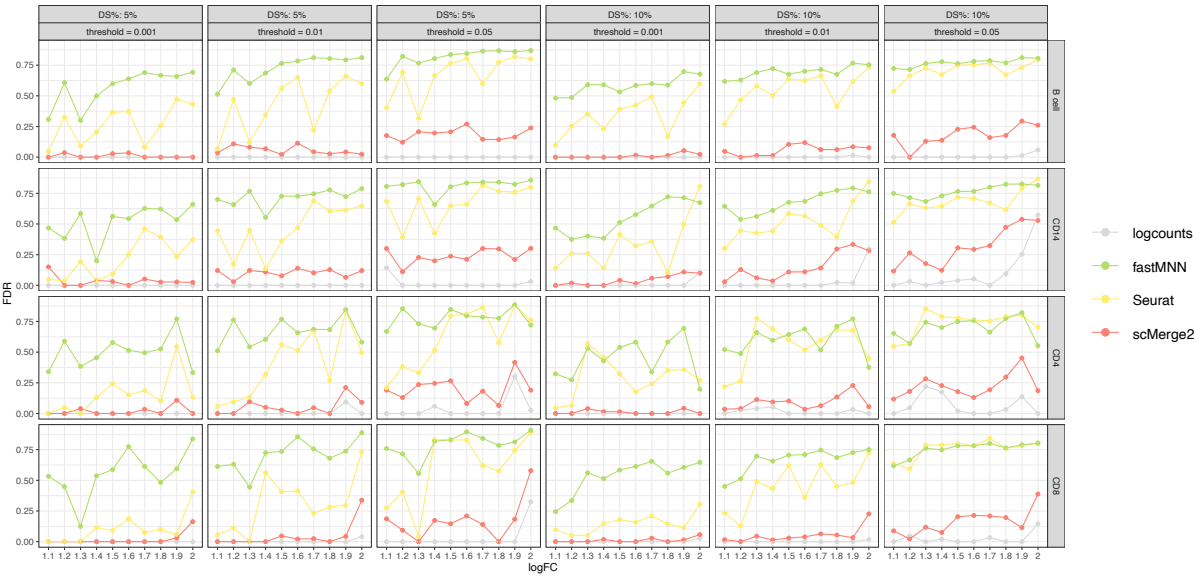
Supplementary Figure S1: UMAP plots of an example of simulated data (logFC = 1.2, DS% = 5%), coloured by batch, sample id, cell type and condition.



Supplementary Figure S2: MA plots of the real and simulated data, where x-axis is the average of gene expression and y-axis is the difference of the gene expression between two condition: (a) Real data; (b) Simulated data using mu formula ∼`cell type`, estimated from data with one condition; (c) Simulated data using mu formula ∼`cell type + sample ID + condition`, estimated from data from two conditions but with condition label permuted. The red dots indicates the simulated ground truth DS genes. The simulation strategy (c) exhibits a more similar pattern with the real data, which therefore is used in this study.

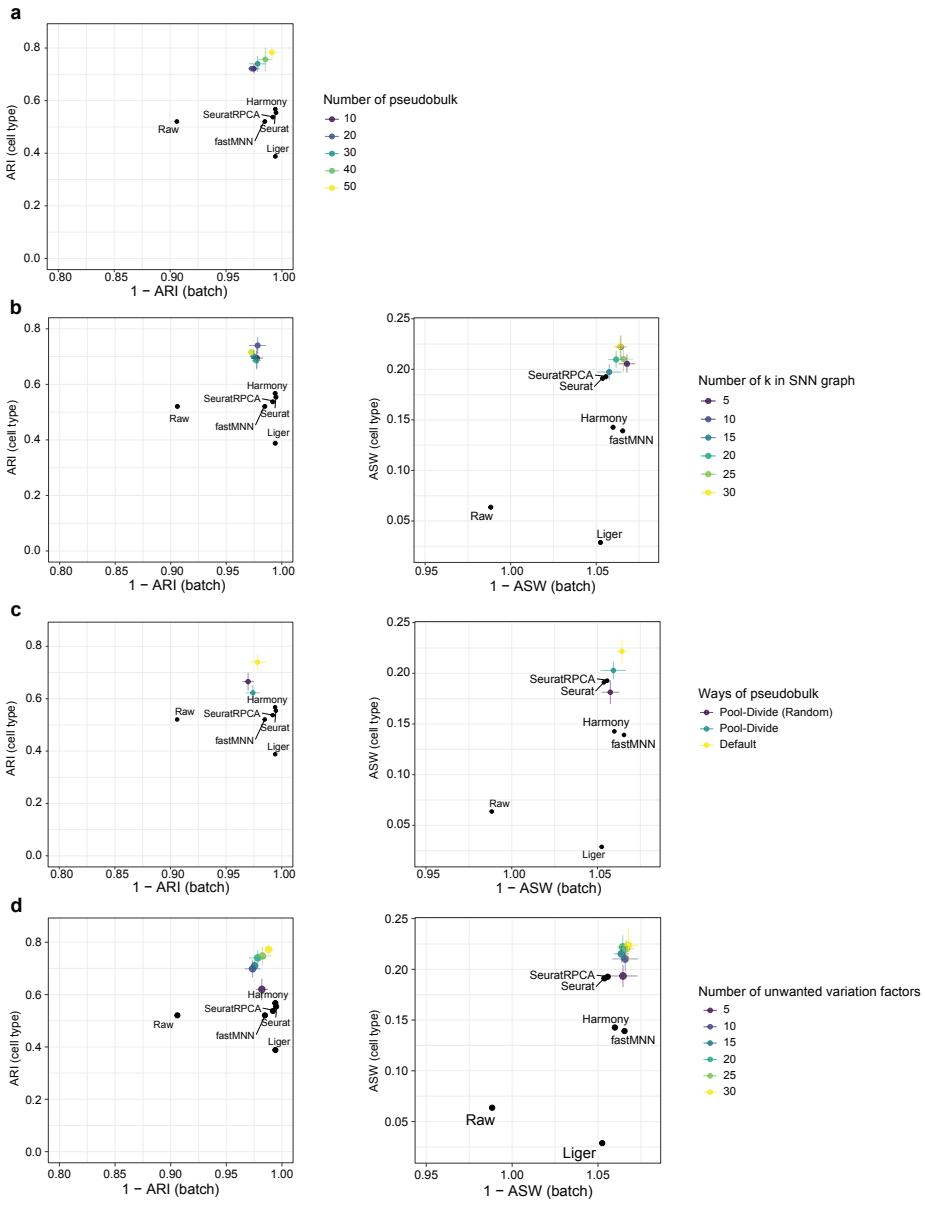


Supplementary Figure S3: F1-score of the differential state (DS) results of four cell types (B cell, CD14, CD4 and CD8) (row) of simulated data, with 5% (1st - 3rd column) and 10% DS genes (4th - 6th column) within each cell type, for scMerge2, Seurat, fastMNN and raw, varying simulated log fold change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

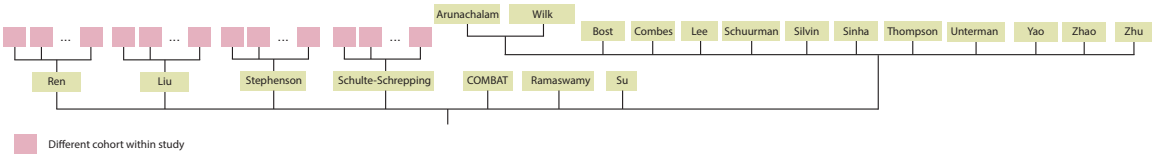


Supplementary Figure S4: FDR of the differential state (DS) results of four cell types (B cell, CD14, CD4 and CD8) (row) of simulated data, with 5% (1st - 3rd column) and 10% DS genes (4th - 6th column) within each cell type, for scMerge2, Seurat, fastMNN and raw, varying simulated log fold change (logFC)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

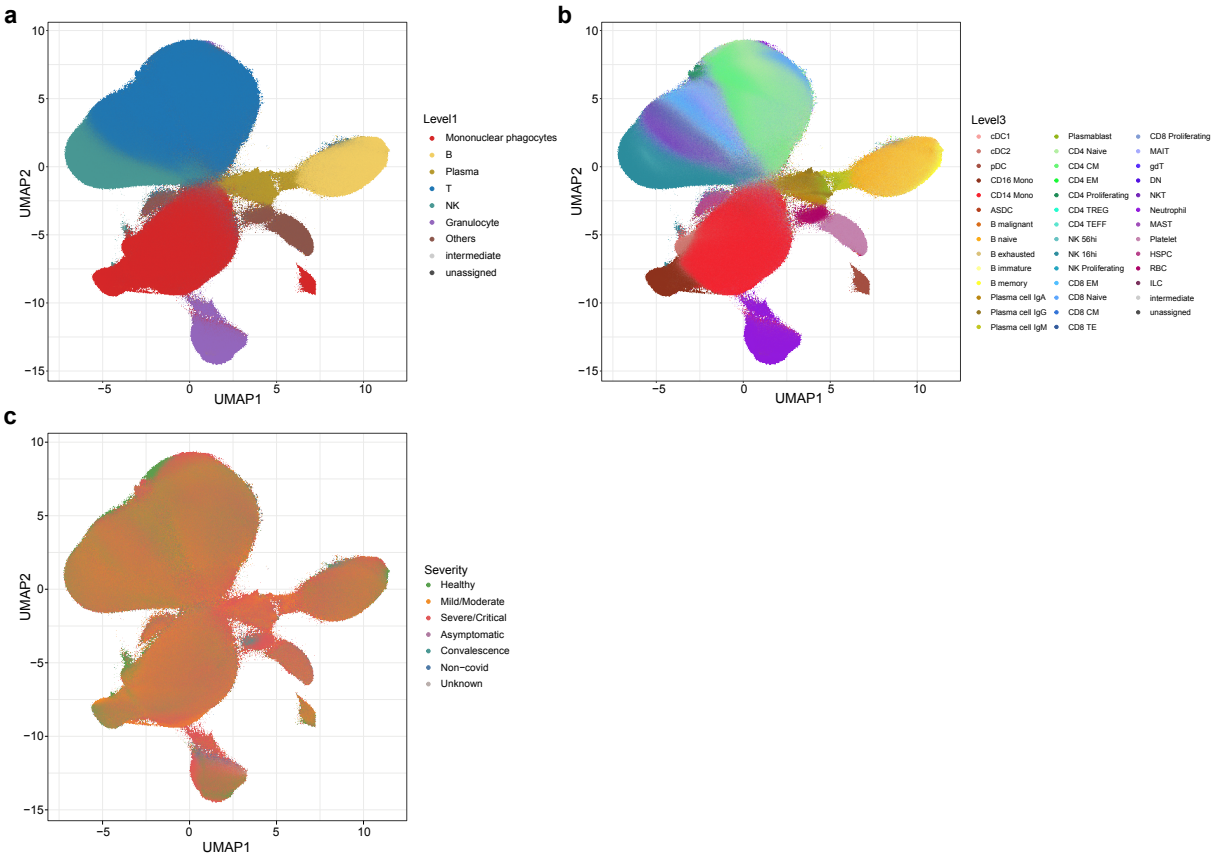
Supplementary Figure S5: TPR of the differential state (DS) results of four cell types (B cell, CD14, CD4 and CD8) (row) of simulated data, with 5% (1st - 3rd column) and 10% DS genes (4th - 6th column) within each cell type, for scMerge2, Seurat, fastMNN and raw, varying simulated log fold change (logFC)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



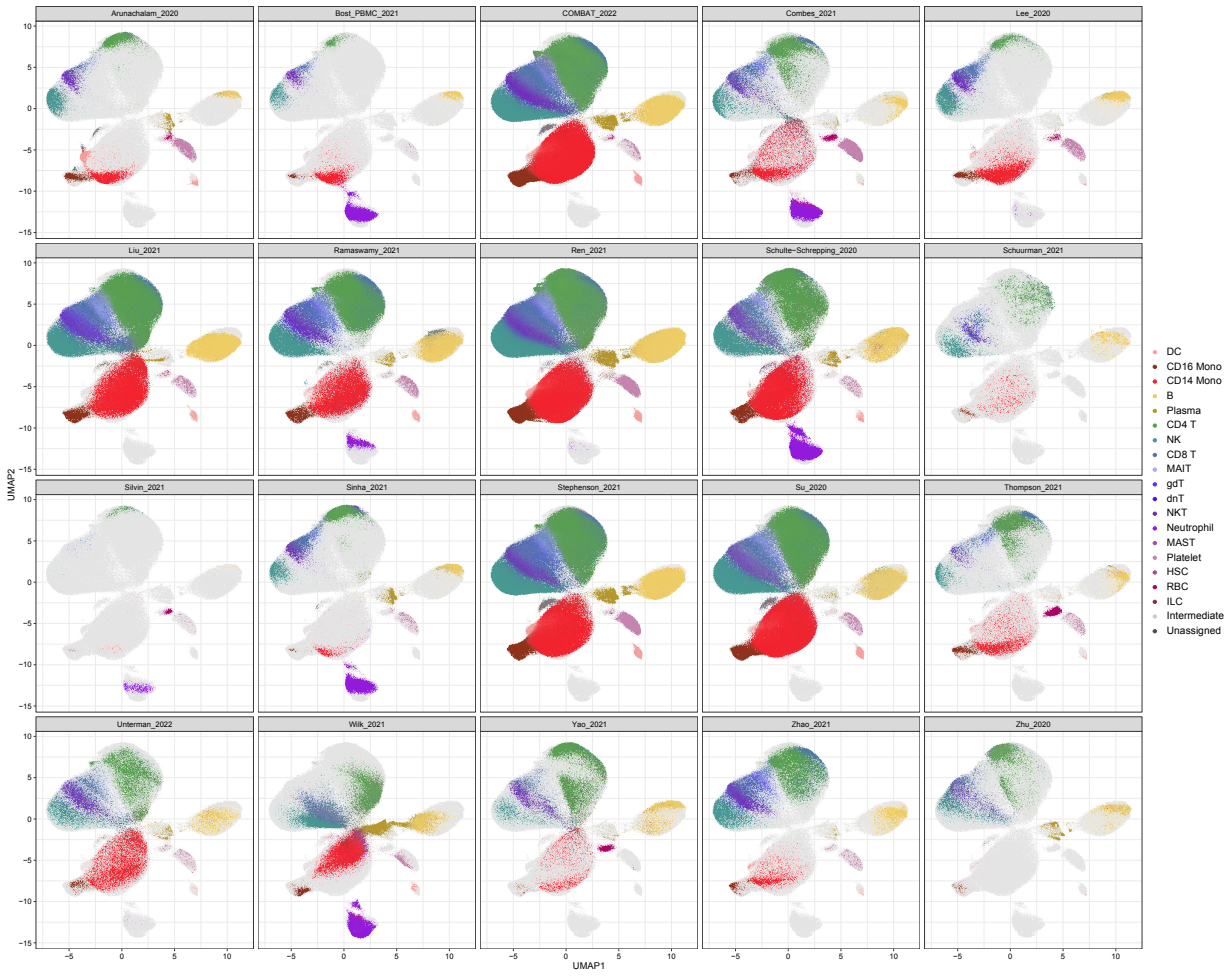
Supplementary Figure S6: Robustness analysis of the tuning parameters of scMerge2 using COVID-19 60k data: Adjusted rand index (ARI) (left panel), where x-axis indicates 1 minus batch ARI and y-axis indicates cell type ARI; Average silhouette width (ASW), where x-axis indicates 1 minus batch ASW and 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 unwatned variation factors (5, 10, 15, 20 (default), 25, 30).



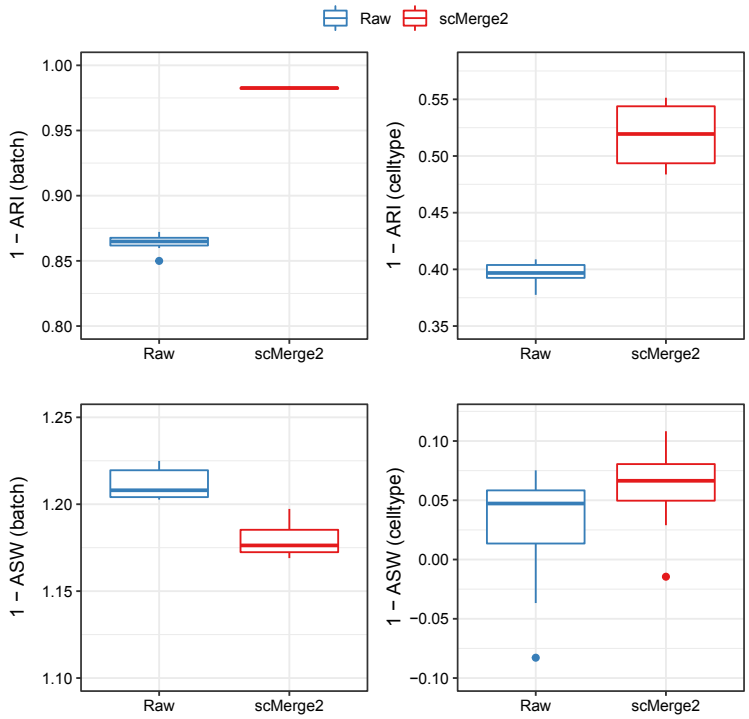
Supplementary Figure S7: Hierarchical merging strategy for COVID-19 scRNA-seq data collection.



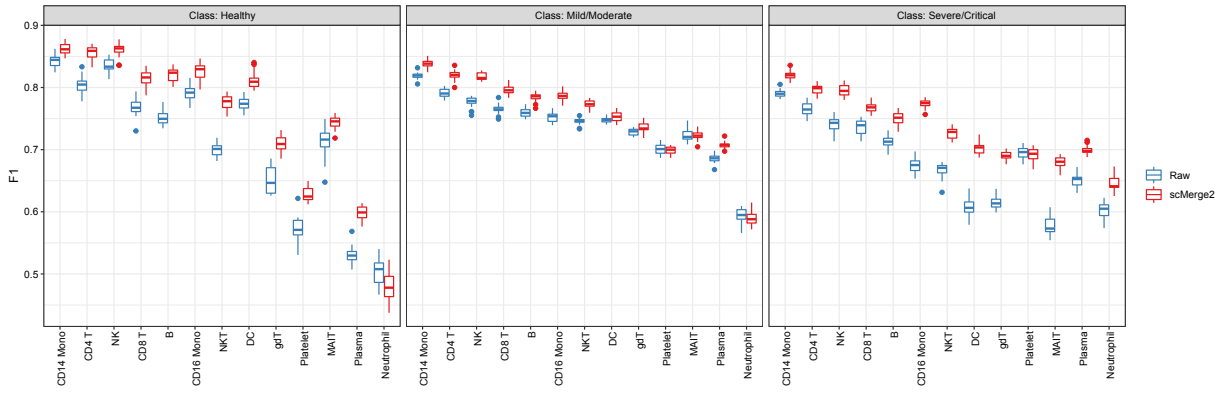
Supplementary Figure S8: UMAP of integration of COVID-19 data collection after scMerge2 integration, coloured by (a) level 1 cell type annotation; (b) level 3 cell type annotation and (c) severity.



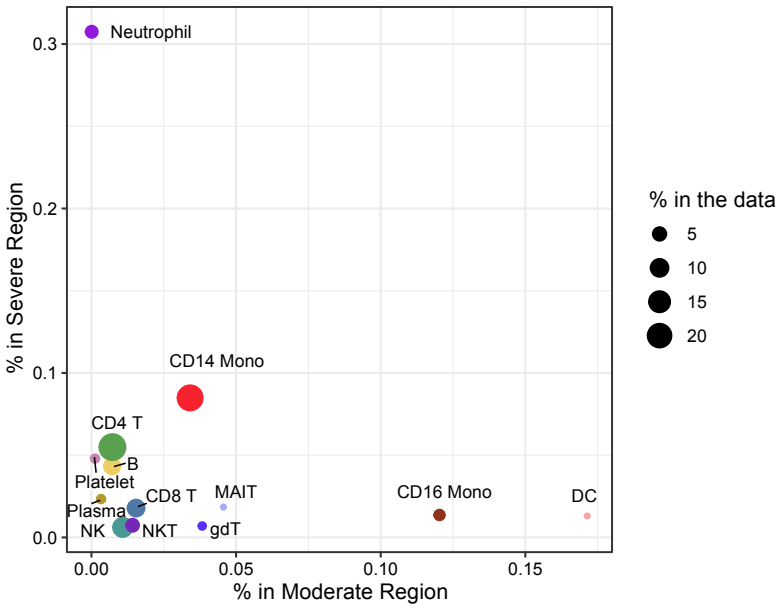
Supplementary Figure S9: UMAP of integration of COVID-19 data collection after scMerge2 integration, coloured by cell type (level 2) and faceted by dataset.



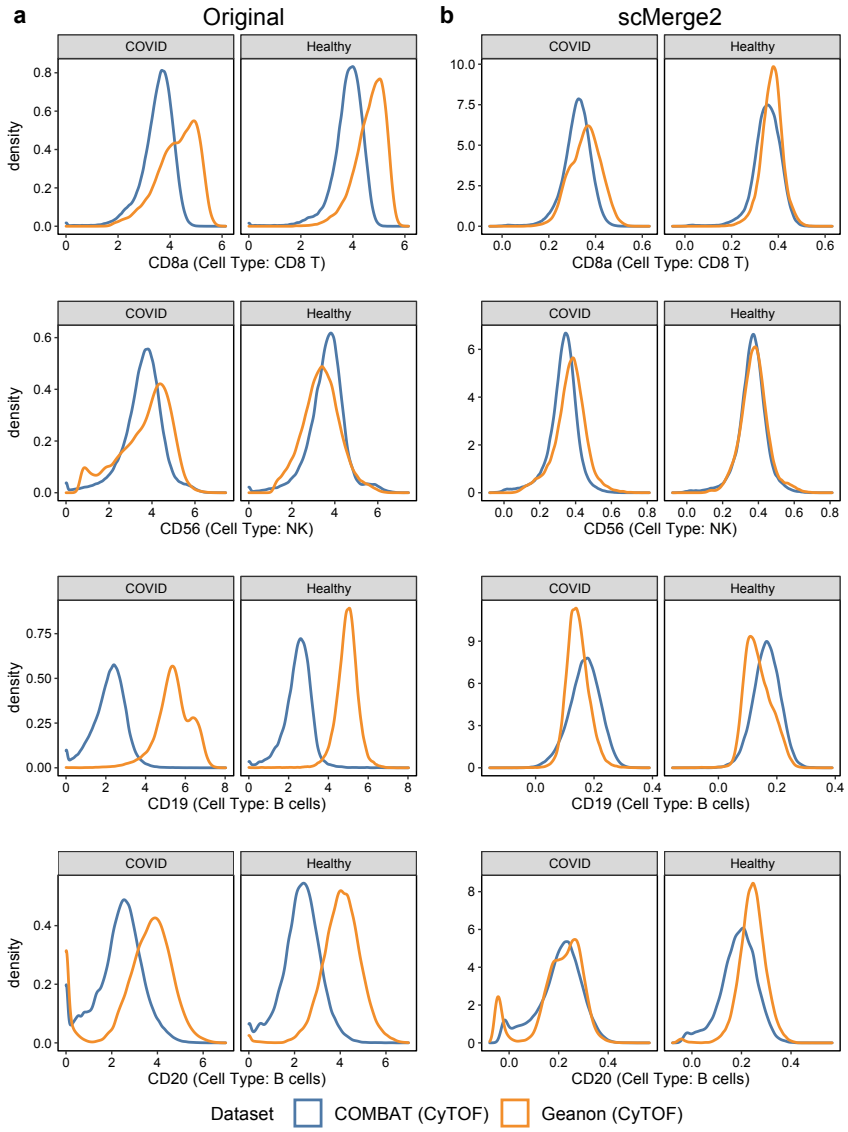
Supplementary Figure S10: Boxplots of evaluation metrics of COVID-19 scRNA-seq data collection for scMerge2-h (data merged in a hierarchical manner) and Raw, where the first row indicates the results of adjusted rand index (ARI): 1 minus batch ARI (left) and cell type ARI (right); the second row indicates 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.



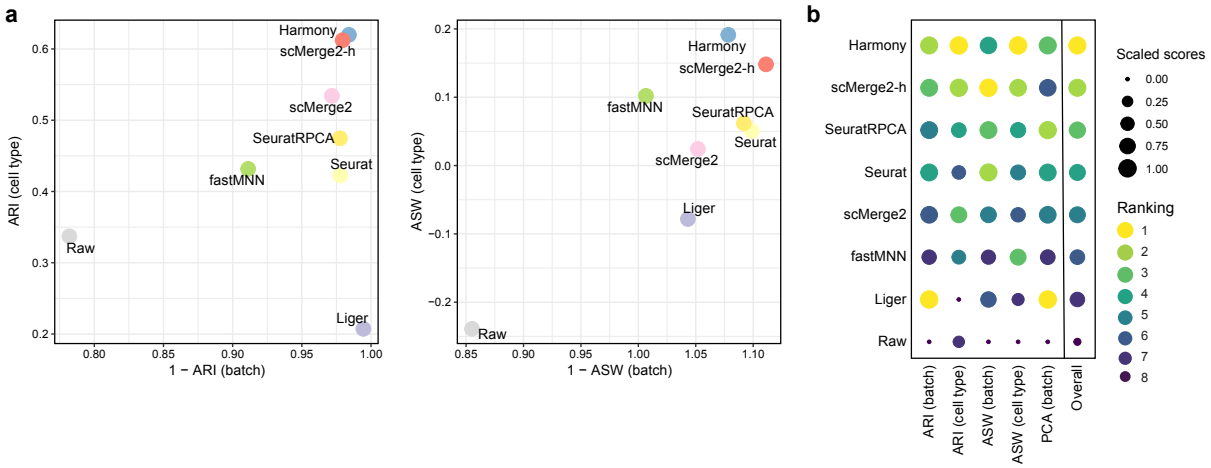
Supplementary Figure S11: Prediction results of disease severity using cell type-specific aggregated expression calculated from raw logcounts (blue) and scMerge2 adjusted results (red), evaluated by class-specific F1 scores.



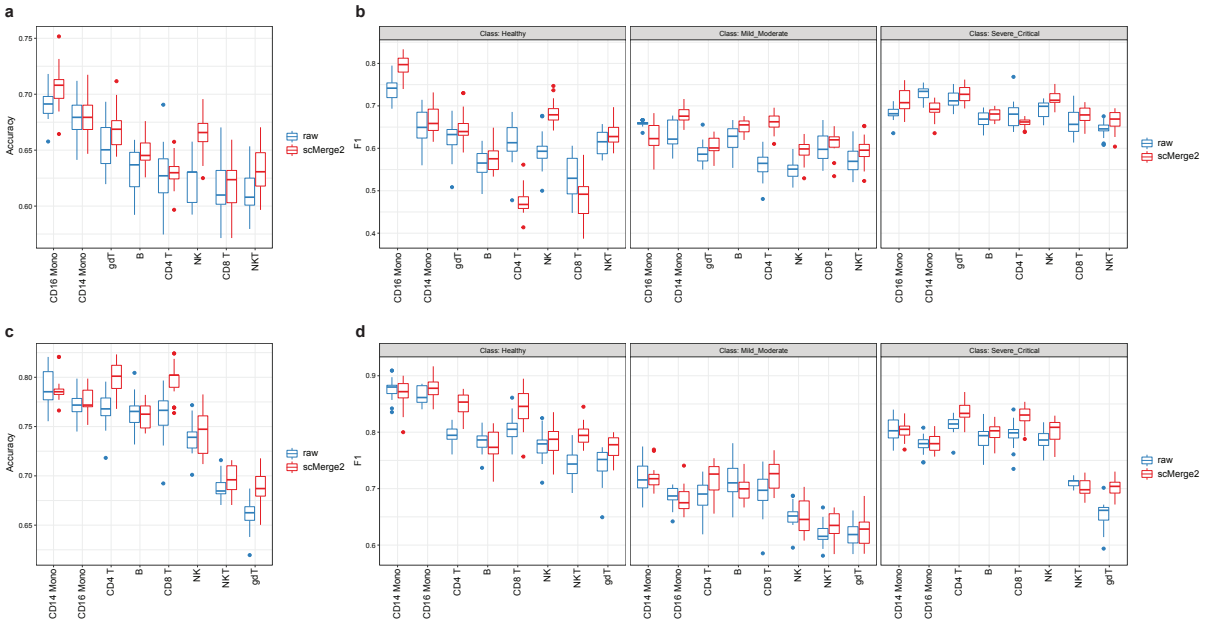
Supplementary Figure S12: Scatter plot shows the proportion of cells in Moderate region (x-axis) vs the proportion of cells in Severe region, determined by DAseq. The size of each point indicates the cell type proportion in the all data (Only cell types that have more than 1% in the data are shown).



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.