HHS Public Access

Author manuscript

Nat Biotechnol. Author manuscript; available in PMC 2024 July 12.

Published in final edited form as:

Nat Biotechnol. 2021 September ; 39(9): 1103–1114. doi:10.1038/s41587-020-00748-9.

A multi-center study benchmarking single-cell RNA sequencing technologies using reference samples

Wanqiu Chen[1,#], **Yongmei Zhao**[2,8,#], **Xin Chen**[1,3,#], **Zhaowei Yang**[4,1,#], **Xiaojiang Xu**[5], **Yingtao Bi**[6], **Vicky Chen**[2,8], **Jing Li**[4,3], **Hannah Choi**[1], **Ben Ernest**[7], **Bao Tran**[8], **Monika Mehta**[8], **Parimal Kumar**[8], **Andrew Farmer**[9], **Alain Mir**[9], **Urvashi Ann Mehra**[7], **Jian-Liang Li**[5], **Malcolm Moos Jr.**[10], **Wenming Xiao**[11,*], **Charles Wang**[3,1,*]

[1]Center for Genomics, School of Medicine, Loma Linda University, Loma Linda, California, United States of America.

[2]CCR-SF Bioinformatics Group, Advanced Biomedical and Computational Sciences, Biomedical Informatics and Data Science Directorate, Frederick National Laboratory for Cancer Research, Frederick, Maryland, United States of America.

[3]Department of Basic Sciences, School of Medicine, Loma Linda University, Loma Linda, California, United States of America.

[4]Department of Allergy and Clinical Immunology, State Key Laboratory of Respiratory Disease, Guangzhou Institute of Respiratory Health, the First Affiliated Hospital of Guangzhou Medical University, Guangzhou, Guangdong Province, Peoples Republic of China.

[5]Integrative Bioinformatics Support Group, National Institute of Environment Health Sciences, Research Triangle Park, North Carolina, United States of America.

[6]Abbvie Cambridge Research Center, Cambridge, Massachusetts, United States of America.

[7]Digicon Corporation, McLean, Virginia, United States of American.

[8]Sequencing Facility, Frederick National Laboratory for Cancer Research, Frederick, Maryland, United States of America.

[*]Correspondence should be addressed to: CW (oxwang@gmail.com or chwang@llu.edu) and WX (wenming.xiao@fda.hhs.gov).
[#]Equal contribution (Co-1st author)

Authors' contributions

CW and WX conceived and designed the study. CW managed the project and directed bioinformatics data analyses. CW drafted the manuscript and annotated all the results. MMJ and AF helped edit the manuscript. WC, BT, MM, PK, MMJ, AF, and AM performed single-cell culturing, single cell captures, scRNA-seq libraries and sequencing. XC, ZWY, YMZ, XJX, VC, YTB, BE, WX, UAM, JL, JLL, and CW performed bioinformatics data analyses. WC, XC, ZWY, YMZ, YTB, XJX, VC, MM, AM, MMJ, and JLL prepared the methods for the manuscript. ZWY drew all the figures, WC, CW and HC prepared the tables. CW, MMJ, WC, AF, and WX revised the manuscript. All authors reviewed the manuscript. CW finalized and submitted the manuscript.

Competing interests statement

Andrew Farmer and Alain Mir are employees of Takara Bio USA, Inc., and Ben Ernest and Urvashi Mehra were employees of Digicon Corporation. All other authors claim no conflicts of interest. The views presented in this article do not necessarily reflect current or future opinion or policy of the US Food and Drug Administration. Any mention of commercial products is for clarification and not intended as an endorsement.

Software and code availability statement

We used many algorithms and code sets for batch correction which have been published previously. All of our code is provided in GitHub and Code Ocean at the following links.

https://github.com/oxwang/fda_scRNA-seq

https://codeocean.com/capsule/0497386

[9]Takara Bio USA, Inc., Mountain View, California, United States of America.

[10]Center for Biologics Evaluation and Research & Division of Cellular and Gene Therapies, U.S. Food and Drug Administration, Silver Spring, Maryland, United States of American.

[11]The Center for Devices and Radiological Health, U.S. Food and Drug Administration, Silver Spring, Maryland, United States of America.

Abstract

Comparing diverse single-cell RNA sequencing (scRNA-seq) datasets generated by different technologies and in different laboratories remains a major challenge. Here we address the need for guidance in choosing algorithms leading to accurate biological interpretations of varied data types acquired with different platforms. Using two well-characterized cellular reference samples (breast cancer cells and B-cells), captured either separately or in mixtures, we compared different scRNA-seq platforms and several pre-processing, normalization, and batch-effect correction methods at multiple centers. Although pre-processing and normalization contributed to variability in gene detection and cell classification, batch-effect correction was by far the most important factor in correctly classifying the cells. Moreover, scRNA-seq dataset characteristics (e.g., sample/cellular heterogeneity and platform used) were critical in determining the optimal bioinformatic method. However, reproducibility across centers and platforms was high when appropriate bioinformatic methods were applied. Our findings offer practical guidance for optimizing platform and software selection when designing a scRNA-seq study.

Introduction

Single-cell RNA sequencing (scRNA-seq) allows transcriptomic profiling of individual cells in unprecedented detail[1–5], prompting increasingly widespread application of this technology. However, investigators seeking to adopt this technology are presented with a bewildering choice of analytical platforms and bioinformatics methods, each with its own set of capabilities, limitations, and costs[1–3, 6–11].

Various aspects of this problem have been examined recently[12–14]. Investigators from the Human Cell Atlas (HCA) consortium performed a comprehensive multi-center study that compared 13 different scRNA-seq protocols using a reference sample containing cells from human, mouse, and dog[15]. Consistent with the previous reports[12–14], this group not only observed striking differences among protocols in quantifying gene expression and identifying cell-type markers, but also found large cross-protocol differences in their capacity to be integrated into reference tissue atlases[15]. The large number of methods compared and the diversity of cells analyzed will likely establish this work as an important milestone for the field; however, comparison of data analysis methods was not a major emphasis of this work. Tian et al.[16] conducted a single-laboratory comparison of three batch-correction methods applied to four scRNA-seq datasets using mixtures of five lung cancer cell lines as reference material. These investigators found significant differences between the methods tested; however, the best-performing of the methods evaluated nevertheless identified six clusters of cells in a mixture consisting of five cell lines. More recently, Tran et al.[17] assessed 14 bioinformatic methods using datasets from several public

domain sources to simulate five different cellular input scenarios; this group nevertheless did not compare various data preprocessing or normalization procedures and evaluated comparatively simple sample composition scenarios (i.e., most samples consisted of only two batches).

The above studies used mixtures of cells exclusively, making it difficult to distinguish biological variability among heterogeneous cell types (cell classification) from purely technical factors (analytical technology platform, institutional or other inter-laboratory differences in cell handling, library protocols, and data-processing methods). This ambiguity makes it difficult to identify the various factors that affect the accuracy of biological classification of the cells analyzed.

As part of the 2nd phase of the Sequencing Quality Control (SEQC-2) consortium, we designed a comprehensive multi-center study to evaluate the influence of technology platform, sample composition, and bioinformatic methods (including preprocessing, normalization, and batch-effect correction). As samples, we used two well-characterized, biologically distinct, commercially available reference cell lines[18] that have whole-genome and whole-exome sequences (WGS and WES) available obtained from multiple technology platforms[19]. Moreover, we analyzed the two cell samples both independently and as mixtures. By comparing a breast cancer cell line versus a ‘normal’ B lymphocyte line, our study models practical, realistic situations in which malignant and normal tissues are analyzed in parallel for diagnostic purposes or for designing personalized medicine therapies[20]. In total, 20 scRNA-seq datasets derived from the two cell lines, analyzed either separately or in mixtures, were generated using four scRNA-seq platforms across four centers: Loma Linda University (LLU), National Cancer Institute (NCI), US Food and Drug Administration (FDA), and Takara Bio USA (TBU).

We compared six scRNA-seq data preprocessing pipelines, eight normalization methods, and seven batch-correction algorithms. These large cross-platform/site scRNA-seq datasets from well-characterized and banked cell lines not only provide a valuable resource for biomedical researchers, but also represent a useful reference for benchmarking single-cell technologies and bioinformatics pipelines for the single-cell sequencing community, and for integrating diverse datasets contributed to large collaborative projects, such as the HCA. Our findings offer practical guidance for selecting the combination of technology platform and bioinformatic methods best suited to the scientific question addressed.

Results

Study design, single cells sequenced, and scRNA-seq data generated.

We used four scRNA-seq platforms: 10X Genomics Chromium, Fluidigm C1, Fluidigm C1 HT, and Takara Bio ICELL8, across four sites, using two well-characterized reference cell lines[18], a human breast cancer cell line (Sample A) and a matched control ‘normal’ B lymphocyte line (Sample B) derived from the same donor (Fig. 1a). Overall, we generated 20 scRNA-seq datasets, including 3′-transcript and full-length transcript scRNA-seq datasets (Supplementary Table 1). For the 10X platform, we compared the standard sequencing protocol (26+98 bp, 10X_NCI) with a modified sequencing method (26+57

bp, 10X_NCI_M) using the same scRNA-seq libraries (Supplementary Fig. 1). For the ICELL8 platform, we also compared paired-end (75×2 bp, ICELL8_PE) with single-end (150×1 bp, ICELL8_SE) analysis. We applied three different pre-processing pipelines either for the 3′-transcript scRNA-seq or for the full-length scRNA-seq (Supplementary Tables 2 and 3). We also evaluated eight different normalization methods (SCTransform, Scran Deconvolution, CPM, LogCPM, TMM, DESeq, Quantile, and Linnorm) and seven different batch-effect correction algorithms (Seurat v3, fastMNN, Scanorama, BBKNN, Harmony, limma, and ComBat) as well as the consistency of scRNA-seq across sites/platforms with bulk cell RNA-seq (BK RNA-seq) on the two cell lines (triplicates each, six RNA-seq datasets).

The overall assessments of the data generation and data quality control (QC) are in the Online Methods, Fig. 1b, Supplementary Tables 1–3, Extended Data Figure 1, or reported in our companion paper[21]. We sequenced a total of 30,693 single cells with either 3′ or full-length scRNA-seq methods (Supplementary Table 1 and Extended Data Fig. 1).

For benchmarking scRNA-seq data, we also identified a large number of differentially expressed genes (DEGs) between the two cell lines at the population level [Supplementary Data 1, using fold-change ≥ 2 plus P-value ≤ 0.01, and false discovery rate (FDR) = 0.05]. The bulk cell RNA-seq sequencing depth and mapping QC are shown in Supplementary Table 4 and Supplementary Figure 2.

To investigate the effect of sequence depth on the number of genes detected and saturation rates across all platforms and scRNA-seq datasets, we down-sampled the different datasets to varying read depths. We observed that the number of genes detected per cell increased rapidly with sequencing depth per single cell up to 100k reads/cell for both cancer cells (Sample A) and B-lymphocytes (Sample B), particularly for the Fluidigm C1 before 50k read-depth. With increasing depth, gene counts ultimately plateau, as expected. However, the rate of saturation was slower after 100k reads for the full-length technologies (C1_LLU and ICELL8), with fewer additional genes being detected for the same increase in sequencing depth when compared with 3′-based scRNA-seq technologies (Fig. 1c;Supplementary Fig. 3). One interpretation of this is the higher complexity of full-length libraries due to sampling fragments across a gene’s entire transcript compared with 3′-based technologies, which by design are biased by sampling only the end of the gene. A caveat here is that in the case of the 3′-based technologies, few cells within the population obtain this read depth (Extended Data Fig. 1), so that on a population level, the full-length technologies provide greater sensitivity. Our data confirm this observation and libraries from full-length technologies have higher library complexity and provide better representations of the captured transcripts with lower sequencing depth than 3′-based technologies. However, the continuous increase in the number of genes detected with deeper sequencing observed for the 10X scRNA-seq data may be dependent on the transcript content of the cell-type; dependence of saturation rate on cell RNA content has been reported previously[22].

Effects of data pre-processing.

For the unique molecular identifier (UMI)-based scRNA-seq data, we compared three pipelines for pre-processing the data: Cell Ranger 3.1 (10X Genomics)[23], UMI-tools[24],

and zUMIs[25], and examined the consistency between the three pipelines regarding the number of barcoded cells identified and the number of genes detected per cell (Fig. 2a,b; Supplementary Tables 2 and 5). For the non-UMI based scRNA-seq data, we examined three additional pre-processing pipelines: FeatureCounts[26], Kallisto[27], and RSEM[28] (Fig. 2d; Supplementary Tables 3 and 5), which included trimming processes (cutadapt[29] or trimmomatic[30]), alignment (STAR[31] and Kallisto), and gene counting (FeatureCounts, Kallisto, and RSEM). For simplicity, we used FeatureCounts, Kallisto, and RSEM for the non-UMI-based platforms. We observed that, for the UMI-based scRNA-seq data, there were variations across the three pipelines both in the number of cells identified and number of genes detected per cell (Fig. 2a & 2b). Cell Ranger v3 was the most sensitive method for cell barcode identification. Umitools and zUMIs filtered most low gene/transcript expressing cells, but detected more genes per cell. Nonetheless, the gene expression level and the consensus genes per cell were highly correlated between any two of the UMI-based pre-processing pipelines; with Umi-tools and zUMIs showing the highest concordance (Fig. 2c).

For non-UMI based scRNA-seq data, much larger variation was observed in the number of genes detected across the three different pre-processing pipelines (Fig. 2d). We found that Kallisto identified a significantly higher number of genes per cell in the full-length transcript scRNA-seq datasets (C1_LLU and ICELL8) and the fewest genes per cell in the C1-FDA_HT (3′-counting) datasets (Fig. 2d). In addition, the consensus genes per cell from the Kallisto pipeline differed significantly from the gene list generated by the other two pipelines for the Fluidigm C1-HT 3′-method. This suggests that the performance of genome alignment-based (RSEM and FeatureCounts) and pseudo-aligner tools (Kallisto), which are most commonly used for full-length isoform analysis algorithms, might under-perform when pre-processing scRNA-seq data from 3′-based technologies. Overall, we found that the gene expression (counts) and the fraction of consensus genes per cell were highly variable across three pre-processing pipelines, both for UMI- and non-UMI based scRNA-seq datasets (Fig. 2c, e). To simplify our comparison, in all of our subsequent analyses we used Cell Ranger—one of the most popular UMI-based methods for 3′-based count technologies—and FeaturesCounts for non-UMI based technologies because it was more consistent with RSEM than Kallisto.

Effects of normalization.

Single-cell RNA-seq data characteristically demonstrate substantial numbers of zero read counts[32]. This can be due to both biological (e.g., bi-stable gene regulation) and technical reasons (e.g., ‘drop out’ due to Poisson sampling limitations or limited efficiency of reverse transcription), making the normalization of scRNA-seq data very challenging. So far, global scaling normalization methods developed for bulk cell RNA-seq data have been used fairly often for scRNA-seq data including CPM (counts per million), UQ (upper quantile), TMM (trimmed mean of M-value), and DESeq[33]. However they were never systemically evaluated using a standard reference scRNA-seq dataset until Tian’s computational integration analysis, which involved two batches of mixed sample sets, but no non-mixed samples captured independently[16]. Regression-based methods have also been proposed to remove known nuisance factors in scRNA-seq data. There are also methods specifically tailored to scRNA-seq datasets, such as SCTransform[34], scran[35], SCnorm[36],

and Linnorm[37]. SCTransform was developed most recently and has been integrated within Seurat v3[38]. Nevertheless, a systematic, thorough evaluation of these methods using scRNA-seq datasets derived from standard reference samples analyzed at multiple centers using multiple platforms is very much needed by the community.

We evaluated eight different normalization methods including SCTransform, Scran deconvolution[39], CPM, LogCPM, TMM, DESeq, Quantile, and Linnorm, using the silhouette width metric, which evaluates how well the two samples from the same cell type are grouped with each other (**see** Methods). We noticed that TMM and Quantile failed to normalize either the breast cancer (Sample A) or B-lymphocytes (Sample B), with silhouette scores that were similar to the un-normalized raw data (Fig. 3a–g). The other methods provided similar normalization as measured by silhouette score. SCTransform seemed to perform slightly better than Scran deconvolution, LogCPM, and Linnorm in that it had the least variation among all normalization methods (Fig. 3a–g; Supplementary Fig. 4a–g; Supplementary Table 6). When comparing silhouette scores across different scRNA-seq platforms and datasets, we noticed that the 10X scRNA-seq data gave consistently lower scores than either the C1 data (both full-length and 3′ across two sites) or the ICELL8 data (Supplementary Fig. 4a–g and Supplementary Table 7).

Distinguishing between irrelevant variations and biological changes of interest can be challenging. A common approach is to pre-process single-cell RNA-seq data to remove uninteresting differences, such as high mitochondrial gene levels in subpopulations or different cell cycle stages. When multiple weak and non-independent factors are present in a dataset, it is also common to apply computational strategies, such as linear regression with read depth normalization, or tailored methods, such as scLVM[40] to remove variations before applying subsequent analysis methods[40]. We found that regressing out mitochondrial genes did not improve the downstream clustering results (Extended Data Fig. 2a–h) and that regressing out the number of genes detected or using a sequencing depth approach did not improve silhouette scores (Supplementary Fig. 5 and Supplementary Table 8).

Since log transformation has a high impact on downstream feature selection and clustering analysis and since our analysis showed it performed similarly to SCTransform, Scran deconvolution and Linnorm, we used logCPM in our subsequent batch-effect and benchmarking evaluations, except where specific normalization methods were embedded in the pipelines.

Batch effects and batch effect corrections.

As noted above, variability between datasets can result from both technical and biological factors[32, 41]. We benchmarked seven algorithms for batch-effect correction: Seurat v3[38], fastMNN/Mutual Nearest Neighbors (MNN)[6], Scanorama[8], Batch-Balanced k-Nearest Neighbors (BBKNN)[9], Harmony[10], limma[42], and ComBat[43]. We visualized clustering projections with both t-SNE and UMAP[44] and applied quantitative metrics, including silhouette width, kBET[45], and alignment score[7] to evaluate the batch-effect removal and cross-platform/center dataset integration as measured by clusterability (ability to separate dissimilar cell types) and mixability (ability to group similar cell types). Both kBET and

alignment score quantify mixability, whereas silhouette width score quantifies how well two different types of cells are separated from each other.

Four different sample scenarios were investigated: in Scenario #1 (Fig. 4a), all 20 scRNA-seq datasets were combined, including mixed and non-mixed, with large proportions of two dissimilar types of cells; in Scenario #2 (Fig. 4b), the breast cancer cell line data (Sample A) were evaluated separately; in Scenario #3 (Fig. 4c), the B cell line data (Sample B) were evaluated separately; and in Scenario #4 (Fig. 4d), 5% or 10% of breast cancer cells (Sample A) were spiked into the B lymphocytes (Sample B), and analyzed with the 10X Genomics platform across two centers in four different batches. Clustering projections were visualized in all four different sample scenarios; silhouette width score was applied to sample scenarios #1 and #4 to assess the separation of cell types.

For Scenario #1, we took gene counts based on the preprocessing pipelines selected as above using either logCPM or the normalization method embedded in the pipelines (e.g., SCTransform for Seurat v3). We sought to determine: 1) which of the algorithms could remove the batch-effects and also separate the two cell types correctly (clusterability/cell classification); and 2) how well cells of the same type from different batches were grouped together (mixability).

The uncorrected data from Scenario #1 showed large variations across platforms and centers (Fig. 4a; Extended Data Figs. 3a, 4, 5a; Supplementary Fig. 6a). In terms of removing batch effects and separating breast cancer cells from B lymphocytes, BBKNN (which ranked the best in clusterability; Fig. 4e), fastMNN, and Harmony were most effective. In contrast, Scanorama, limma, and ComBat did not separate cell types discretely.

In terms of mixability, BBKNN performed well in grouping B cells together from different batches, but was the worst of the methods tested for breast cancer cells (Fig. 4a–b,e,g and Supplementary Fig. 7); Seurat v3 was one of the best at grouping similar cells from different batches together particularly for breast cancer cells, but over-corrected and clustered B lymphocytes and breast cancer cells, two highly dissimilar cell types, together—a misclassification (Fig. 4a,e,g; Extended Data Figs. 3a, 4, 5a; Supplementary Figs. 6a & 7a,b). Scanorama, limma, and ComBat also failed to separate breast cancer cells from B cells (Fig. 4a/e/g; Extended Data Figs. 3a, 4, 5a; Supplementary Figs. 6a & 7a,b).

However, when only data from the 10X platform were analyzed, Scanorama both separated dissimilar cells clearly and grouped similar cells together very well, regardless of center (Fig. 4d; Extended Data Figs. 3d, 5d, 6; & Supplementary Fig. 6d). For fastMNN, a spiked in sample was required to provide a subpopulation of cells common to all samples analyzed. We also found that the order of loading the datasets into fastMNN was critical for correcting batch effects; specifically, the mixed, or most heterogeneous, data should be loaded into the pipeline first (Extended Data Figs. 7 and 8).

For the scRNA-seq datasets derived from five batches of each cell line separately (Scenarios #2 and #3; Fig. 4b,c) or four batches of B lymphocytes spiked with 5% or 10% of cancer cells (Fig. 4d), t-SNE and UMAP showed that the cells were clustered separately by batch and that similar cells were not evenly mixed, indicating large variations and/or strong batch

effects (Fig. 4b–d; Extended Data Figs. 3b–d, 5b–d; Supplementary Fig. 6b–d). When we applied batch-effect correction methods, we observed that Harmony, Seurat v3, and fastMNN were ranked on the top in grouping similar cells together based on kBET score for the breast cancer cells and B cells in Scenarios #2 and #3 (Fig. 4b,c,g; Extended Data Figs. 3b,c, 5b,c; and Supplementary Figs. 6b,c and 7c,d). Consistent with our findings in Scenario #1, BBKNN had the poorest mixability for breast cancer cells as measured by kBET, despite its good performance in grouping B cells together (Fig. 4g). In contrast, Scanorama performed worst in cellular mixability (Fig. 4b,c,g; Extended Data Figs. 3b,c, 5b,c; and Supplementary Fig. 7c,d). For B cells (Scenario #3), which are relatively more homogeneous than breast cancer cells, limma and ComBat also seemed to perform well in grouping similar cells together (Fig. 4c,g; Extended Data Figs. 3c, 5c; and Supplementary Figs. 6c and 7d). In the spike-in datasets of Scenario #4, all methods were able to remove batch-effects and separate the spiked in cancer cells from B cells discretely, with BBKNN performing the best (Fig. 4d, f), followed by Harmony and Seurat v3 (notwithstanding that the latter failed in Scenario #1), whereas Harmony, Seurat v3, BBKNN, and Scanorama all performed well in grouping similar cells together (Fig. 4d, g; Extended Data Figs. 3d, 5d; Supplementary Figs. 6d and 7e,f).

We also compared mnnCorrect versus fastMNN using the 20 scRNA-seq datasets and found both versions performed similarly as evaluated by the t-SNE and UMAP. However, fastMNN took much less computation time; for example, 51.1 seconds for fastMNN versus 62781.8 seconds for MNN in Scenario #1 (1,229-fold faster, Supplementary Fig. 8).

CellRanger 3.1 allows some cells with extremely low gene expression to be identified. As a result, significantly more cells were detected using CellRanger 3.1 than with CellRanger 2.0 (Supplementary Table 9). We therefore compared batch correction results obtained with the two versions for all four sample combination scenarios. Overall, even though there was substantial consistency between CellRanger 3.1 and CellRanger 2.0 preprocessed data; we noticed that the batch corrections performed better using CellRanger 2.0 (Supplementary Figs. 9 and 10). As part of our cross validation, we also performed batch-correction analysis on the four scRNA-seq datasets from Tian et al.[16], which were generated from mixtures of either 3 or 5 lung cancer cell lines in two batches. Consistent with the findings from our own datasets, our analysis of the Tian et al. datasets showed that fastMNN, Seurat v3, and Harmony performed well (Supplementary Figs 11 and 12), whereas both Seurat v2 CCA (Canonical Correlation Analysis) and MNN failed to separate cells from the five different cell lines in the Tian et al. analysis[16].

Consistency of global gene expression across platforms and sites.

We first evaluated global gene expression cross-platform consistency using scatter plots and the common transcripts detected across seven scRNA-seq datasets for either Sample A or Sample B (Extended Fig. 9a, b). The bar chart plots clearly showed that the bulk cell RNA-seq had a much wider range of genes with high expression than any of the scRNA-seq platforms where most genes had very low UMI counts. Overall, our scatter plot analyses showed cross-platform correlation coefficients between single-cell datasets of 0.8–0.98; correlation coefficients between scRNA-seq and bulk cell RNA-seq were lower, which could

be due to the large differences in gene expression counting between bulk and scRNA-seq. Pearson correlation coefficient analysis indicated a higher intra-platform than inter-platform correlation for both breast cancer cells (Sample A) and B lymphocytes (Sample B). For example, 10X scRNA-seq data had high correlations across centers, ≥ 0.94 for Sample A and 0.98 for Sample B.

We then evaluated global gene expression consistency across different platforms and sites by calculating a pairwise Pearson correlation (R) based on the percentage of cells (see Methods) that expressed 500 abundant, 500 intermediate, and 500 scarce genes, as defined by bulk cell RNA-seq data (Supplementary Fig. 13a–f). To account for variable sequencing depth across the different datasets, we performed down sampling to 100K reads for each dataset. At this read depth, we observed a much higher Pearson correlation when using the 500 highly expressed genes than when using 500 intermediate or 500 scarce genes in both cell types (Supplementary Fig. 13a–f). We also observed higher consistency among the sites using the same platform or type of technologies (i.e., 10X, ICELL8). Even with the 500 low-abundance genes, we observed good Pearson correlations (Sample A: 0.7–0.99; Sample B: 0.8–1.0) between sites within either 10X 3′ or ICELL8 except C1 technologies (Supplementary Fig. 13c, f). However, the consistency (Pearson correlation) within 3′ technologies (10X and C1_FDA_HT) or within full-length (C1_LLU, ICELL8_SE, ICELL8_PE) platforms was not always better than that between 3′ and full-length platforms. Nevertheless, we caution that there might be some biases in this analysis because the cell numbers were very different across platforms (i.e., only 66 or 80 single cells for the Fluidigm C1 full-length versus up to a few thousand cells for the 10X platform). Thus, the influence of variation due simply to sampling must be considered.

We further compared the single-cell gene expression profiles [log(CPM+1) with normalized counts] across four different classes of RNA, including protein-encoding RNA, antisense RNA, long intergenic non-coding RNA (lincRNA), and miscellaneous RNA (miscRNA; Supplementary Fig. 14 and Supplementary Table 10). As a comparison, the bulk cell RNA-seq gene expression profile was also plotted side-by-side. We noticed that ICELL8_SE gene expression profiles showed relatively higher detection sensitivity for the lower abundance transcripts. The 10X technology also seemed to show good detection sensitivity for lower abundance protein coding, antisense RNA, and lincRNA transcripts, and there was high consistency across three 10X scRNA-seq datasets (10X_LLU versus 10X_NCI versus 10X_NCI_M). Gene counts across all scRNA-seq platforms and datasets for the protein coding RNAs were comparable. For the C1 platforms (full-length and 3′), the detection range was compressed, with much lower log(CPM+1) values for antisense RNA and lincRNA.

Consistency of cell-type specific markers across scRNA-seq platforms and sites.

We exploited feature plotting using the top 10 cancer-specific DEGs and top 10 B cell-specific DEGs[46, 47], based on the DEGs derived from the bulk-cell RNA-seq to further evaluate single-cell gene expression consistency before and after fastMNN correction across all scRNA-seq datasets and platforms. Clearly, before fastMNN batch-effect correction, the breast cancer cells (Sample A) and B lymphocytes (Sample B) were neither clustered

together, nor clearly separated (Fig. 5a and 5c). However, after applying fastMNN, cells expressing breast-cancer specific versus B-cell specific marker genes were clustered together and there was clear separation between the two cell types (Fig. 5b and 5d).

We further compared the consistency of the single-cell gene expression profiles across platforms for CD40, CD74, and TPM1 with a subsampling at 100k reads for each dataset. The rationale for selecting these three markers to benchmark the transcript detection consistency across platforms was based on both their cell-type specificity and their expression levels, either intermediate or highly abundant. The B-cell specific marker gene, CD40, was most often expressed at an intermediate level ($1 \leq CPM < 10$) per cell, and it was detected in as few as 24.9% of cells with the C1_FDA_HT to as many as 53% with the C1_LLU. A significant percentage of cells (44–44.6% for 10X and 23.2–28.1% for C1 and ICELL8) expressed this gene at levels close to the limit of detection ($CPM < 1$). In contrast, the CD40 transcript was detected at either low or near noise levels ($CPM < 1$) in breast cancer cells (Supplementary Table 11). However, CD74, also a B-cell specific marker gene, was much more abundant ($CPM \geq 10$) in almost all single B cells (98.9–100%) with excellent consistency across all platforms, except for C1_FDA_HT, where 5% of the B cells had an intermediate level ($1 \leq CPM < 10$; Supplementary Table 12). In contrast, CD74 was present at low or near noise levels ($CPM < 1$) in breast cancer cells. For this marker, full-length transcript technologies were more sensitive than the 3′-scRNA-seq technologies (Suppl. Table 12). With some variation across platforms, a high percentage of single cells expressed TPM1 in breast cancer cells, but the detection level fell mostly within an intermediate level ($1 \leq CPM < 10$). In B cells, which are not associated with TPM1, there was little or no detection ($CPM < 1$) (Supplementary Table 13).

Discussion

Here, we assessed scRNA-seq performance across four sequencing platforms at four centers, focusing on the effects of bioinformatic processing, including preprocessing, normalization, and batch-effect correction. We analyzed two biologically distinct reference cell lines[18], either separately or as mixtures, for which a large amount of multi-platform WGS and WES data are available[19]. Our benchmark study has produced well-characterized reference materials (reference samples A and B), 20 openly available scRNA-seq datasets, and detailed methods. In this regard, it will have similar resource value and utility for the single-cell sequencing community as the Zook at al. study[48], carried out by the Genome in a Bottle Consortium (GIAB), which aimed at developing reference materials, data, and methods to enable translation of genome sequencing to clinical practice. The availability of scRNA-seq datasets based on sustainable, well-characterized reference samples that have been processed across multiple platforms and centers is critical for benchmarking single-cell technologies and bioinformatic methods.

Our analyses indicated that although pre-processing and normalization contributed to variability in gene detection and cell classification, batch effects were large, and the ability to assign the cell types correctly across platforms and sites was dependent on the bioinformatic pipelines, particularly the batch-correction algorithms used. In many scenarios, Seurat v3, Harmony, BBKNN, and fastMNN allowed correct classification of

the two cell types. However, when samples containing large fractions of biologically distinct cell types were compared, Seurat v3 over-corrected the batch-effect and misclassified the cell types (i.e., breast cancer cells and B lymphocytes clustered together), whereas limma and ComBat failed to remove batch effects. However, we also showed cross-center/platform consistency was high when appropriate bioinformatic methods were applied.

The findings from our study offer practical guidance for optimizing and benchmarking a platform or a protocol, and for selecting appropriate bioinformatics methods when designing scRNA-seq experiments. In our study, samples of both lines were distributed to different centers and grown out separately at these locations to reflect the sort of experimental variability likely to be encountered in real-world collaborations (in contrast to the situation with GIAB[48] or our companion paper[19], in which identical aliquots of gDNA reference material were distributed to the study sites). As expected, site-to-site and platform-to-platform variability was large, but when an appropriate combination of computational methods was chosen, these effects could be corrected.

For benchmarking a newly developed scRNA-seq platform or protocol, or for quality control while starting a scRNA-seq experiment regardless of platform, we recommend including a mixed sample with 5–10% of the reference breast cancer cells spiked into the reference B-cell line sample. Single cells of the mixed samples may be processed in different batches, while non-mixed samples of the breast cancer cells and B cells should be processed in the same batch. The cells obtained from the provider should first be expanded for cryopreservation in multiple aliquots. Any aliquot of cells may then be sub-cultured for a few rounds without significantly affecting the accuracy of cell clustering or identification, as demonstrated in our study. We also recommend obtaining bulk cell RNA-seq in triplicate for both reference lines. Gene detection for breast cancer (TMP1) or B-cell (CD40 and CD74) specific markers can be compared with our reference data (Supplementary Tables 11, 12 and 13; Fig. 5).

The acquired scRNA-seq data can be pre-processed using any of the methods ranked in Figure 6a, as appropriate to the scRNA-seq technology employed, and any of the normalization methods except for TMM and Quantile, which we do not recommend. If desired, the scRNA-seq data obtained can be merged with our reference datasets from any of the four data composition scenarios (Fig. 4) and analyzed using our benchmarked reference methods with different batch-correction algorithms (see Fig. 6e and Supplementary File 1). Moreover, our scRNA-seq reference datasets and results from the bioinformatics methods we used can be a valuable resource for developing or benchmarking new methods, and we recommend using all four different sample/data composition scenarios (Fig. 4) to gain a thorough performance evaluation.

We found that pre-processing and normalization contributed to variability in gene detection and cell classification. For the UMI-based datasets, Cell Ranger v3 detected the most cells, whereas zUMIs detected the most genes per cell. For the non-UMI-based datasets, Kallisto identified the highest number of genes per cell in the full-length scRNA-seq datasets, but the fewest genes per cell in the C1_HT dataset. Out of eight normalization methods evaluated, SCTransform, LogCPM, Scran deconvolution, and Linnorm all performed well, with the

SCTransform displaying the lowest variance. In contrast, TMM and Quantile performed poorly across all datasets (Figs. 3, 6b). Moreover, we found that regressing mitochondrial genes and normalizing UMI counts could not remove the batch effects (Extended Data Fig. 2; Supplementary Fig. 5).

An ideal workflow would remove variability due purely to technical factors or sampling ambiguity without obscuring meaningful biological differences important to accurate classification of diverse cell types[15]. We reasoned that not all algorithms would perform equally well at these tasks, so we compared seven algorithms with respect to performance on two aspects of cross-platform data integration in addition to removing batch variations: clusterability and mixability. Seurat v3, Harmony, BBKNN, fastMNN, and Scanorama worked well in removing batch effects and classifying cells in some scenarios. However, sample heterogeneity, dataset composition, and platform used influenced the outcome of the analyses (Fig. 4a–g, Extended Data Figs. 3a–d, 4, 5a–d, 6a–d, 7, 8a–f; and Supplementary Figs. 6a–d, 7a–f). For example, despite its high mixability for the data from the same cell line (Fig. 4b,c,g), Seurat v3 over-corrected batch effects and misclassified cells in datasets containing large fractions of highly dissimilar cells (Fig. 4a,e; Extended Figs. 3a, 4). Scanorama worked well only for datasets generated entirely with the 10X platform (Extended Data Fig. 6a–d), an issue not identified by the original developers of the algorithm[8]; we found that this algorithm failed to remove batch effects both from our own non-10X platform data and when we re-analyzed the data from Hie et al. (Extended Data Fig. 10a–f).

BBKNN performed the best whereas limma and ComBat were poorest in cross-platform/center separation of two types of cells from each other, particularly when there were large proportions of dissimilar cells in the datasets (Figs 4a, e and 6c). In contrast, Seurat v3—despite failing to separate distinct cell types in the multi-platform datasets consisting of large proportions of each cell-type (Fig. 4a/e)—worked well in the cross-center datasets comprising a large proportion of one cell type spiked with a small portion of different cells (Fig. 4d,f). This method may be best suited to situations where the primary objective is to remove differences between datasets due mainly to technical sources of variations rather than to integrate data from biologically dissimilar cell populations. Consistent with this notion, Seurat v3, fastMNN, and Harmony all performed well in mixability for the scRNA-seq data derived from biologically identical or similar samples across platforms/sites, whereas limma and ComBat could mix B-cells well across platforms/sites, perhaps because these cells are more homogeneous (Fig. 4b,c,g; Extended Data Figs. 3b,c and 4). Indeed, for all algorithms, we found that the cross-platform/center mixability was better in the B cells than in the more heterogeneous cancer cells (Fig. 4b, c,4g). However, for fastMNN/MNN, both the requirement for mixed samples and the order of importing data into the pipeline were critical for effective batch-effect correction: the mixed samples should be imported first (Extended Data Figs. 7 and 8). The presence of mixed cell types to provide a shared sub-population across the batches to be corrected is consistent with the logic of MNN algorithm[6]. Thus, our findings clearly highlight that the choice of appropriate batch-effect method depends on the characteristics of the samples (e.g., heterogeneity and cell composition) and datasets (multiple versus fewer platforms used).

Some combinations of different preprocessing methods with different batch-effect algorithms might not always work well. For example, Tian et al.[16] examined imputation methods by combining Linnorm and Drimpute with MNN; TMM and Drimpute with Scanorama; and SAVER with Seurat using data from a mixture of five cell lines. However, this approach failed to correct batch effects not only for Seurat and Scanorama, but also for MNN – the top-performing method of those evaluated – which identified six clusters in the samples, instead of five cell lines (see Supplementary Fig. 8d in their paper)[16]. However, when we applied our preferred preprocessing method to Tian's datasets, fastMNN, Seurat v3, and Harmony all separated the five different cell lines into discrete clusters (Supplementary Figs. 11 and 12).

Previous studies using heterogeneous mixtures containing different cell types have provided useful insights into bioinformatics methods, particularly with regards to batch-effect correction algorithms[16]. One example is the requirement for shared sub-populations (anchor cells) between different datasets for Seurat and fastMNN to integrate data from biologically dissimilar cell populations; the proportion of the anchor cells present was critical (Fig. 4a,d—g). The study of Mereu et al.[15] used mixed cells of three different species, but focused primarily on the analytical platform, with limited investigation on bioinformatics methods. Nevertheless, as in Tian's study, no non-mixture cells/data were captured. However, studies restricted to mixtures of cell types cannot examine the ability of a method to eliminate variability due to technical factors, important in its capacity to group similar cells together, and also—independently—assess the ability of a method to separate dissimilar cells correctly. We found that analyzing both mixtures of dissimilar cells in various proportions and un-mixed samples of the two distinct cell lines provided important additional insights. For example, the widely used Seurat v3 method[38] excelled at grouping similar cells together, but it over-corrected—completely failing to separate B cells from breast cancer cells-when large proportions of two dissimilar cell types were analyzed (Fig. 4a–g). Also, as noted previously, it may be difficult to replicate the cellular sample used in Mereu et al. for confirmatory or further exploratory evaluations. In contrast, our study was not only empowered by a mixology design but was also further strengthened by the inclusion of non-mixture sample/data captured separately. Moreover, our reference cell lines are commercially available for future benchmarking studies. Another unique advantage of our standard reference samples is the availability of massive cross-platform deep WGS and WES data[19], which will be valuable resource for benchmarking future single-cell WGS or proteomics technologies.

In summary, we assessed scRNA-seq data generated with multiple platforms across several centers using samples derived from two well-characterized, biologically distinct cell lines, analyzed either separately or as mixtures. This experimental design allowed us to benchmark the effect of sample cell composition and evaluate variations due to both platform-specific effects and each element of the bioinformatic analysis pipelines, particularly the batch-effect correction algorithms. We believe this will provide a useful resource for the community to benchmark additional scRNA-seq protocols and bioinformatics algorithms. Overall, our study shows that while batch effects are large, the variations across sites and platforms can be corrected by appropriate computational methods. Critically, our study highlights the importance of choosing computational methods appropriate to both the technology platform

used and the composition of the samples analyzed. An important conclusion we draw is that the capabilities and limitations of the various elements of the bioinformatic analyses should be chosen to match the experimental situation. A more detailed enumeration of our conclusions is presented in the Online Methods and in Supplementary File 1. The capabilities and limitations of the methods we evaluated are displayed graphically in Figure 6. Best practice recommendations based on our findings are presented in Box 1 and in flow-chart format in Figure 6e.

METHODS

Study design

A schematic overview of the study design is illustrated in Figure. 1a. Briefly, two well-characterized reference cell lines (sample A, breast cancer cell line vs. sample B, a matched control B lymphocyte line) were used to generate scRNA-seq data across four platforms (10X Genomics Chromium, Fluidigm C1, Fluidigm C1 HT, and Takara Bio ICELL8), at four testing sites (LLU, NCI, FDA, and TBU) using standard manufacturer's protocols. At the LLU and NCI sites, mixed singe-cell captures and library constructions were also prepared with either 10% or 5% cancer cells spiked into the B lymphocytes. At the NCI site, single-cell captures and library constructions were also performed with methanol-fixed cell mixtures (5% cancer cell spiked into B lymphocytes, named as fixed_1 and fixed_2 in two independent sample captures). The 10X scRNA-seq libraries constructed at NCI were sequenced using a shorter modified sequencing method (26+57 bp) at the NCI site. One set of 10X scRNA-seq libraries constructed at the NCI site was also sequenced at LLU using the standard sequencing method (26+98 bp) (Supplementary Table 1). Bulk cell RNA-seq was also prepared from these cell lines, each in triplicate. All scRNA-seq data were subject to 3 different pre-processing pipelines for either 10X or C1/ICELL8 technologies, respectively. We evaluated eight normalization methods, SCTranform, Scran Deconvolution, CPM, LogCPM, TMM, DESeq, Quantile, and Linnorm and seven batch effect correction algorithms including Seurat v3, fastMNN, Scanorama, BBKNN, Harmony, limma, and ComBat. The cross-platform and cross-center performances were further evaluated by t-SNE, UMAP, and three quantitative metrics (silhouette score, modified alignment score, and kBET), as well as scatter plotting and feature plotting. scRNA-seq data were also compared with population-average RNA-seq data.

Abbreviations and notations for Fig. 1a: **10X_LLU**, single cells were captured using a 10X Genomics Chromium controller; scRNA-seq was done at the LLU Center for Genomics using the standard 10X Genomics protocol (26+98 bp); **10X_NCI_M**, 10X Genomics scRNA-seq libraries were prepared and sequenced at the NCI sequencing facility using a modified 10X sequencing protocol (26+56 bp); **10X_NCI**, the same 10X Genomics scRNA-seq libraries prepared at the NCI sequencing facility were also sequenced at LLU using the standard 10X sequencing protocol (26+98 bp); **C1_FDA_HT**, single cells were captured using a Fluidigm C1 HT IFC and the scRNA-seq libraries were sequenced at the FDA sequencing facility (75×2 bp, PE); **C1_LLU**, single cells were captured using a Fluidigm C1 IFC chip and the scRNA-seq libraries were sequenced at the LLU Center for Genomics (150×2 bp, PE, ~4–4.77M reads/cell); **ICELL8_PE**, single cells were captured using an

ICELL8 chip at TBU and scRNA-seq libraries were paired end sequenced (75×2 bp, PE) at TBU; **ICELL8_SE**, the same scRNA-seq libraries generated at TBU site were also sequenced, single-end (SE), at LLU (150×1 bp, SE, ~1M reads/cell). See Supplementary. Table 1 for details on the numbers of single cells captured and sequencing read depths for each platform and each site.

Cell culture and single cell preparation

We obtained the human breast cancer cell line (HCC1395, Sample A) and the matched normal B lymphocyte line (HCC1395 BL, Sample B) from ATCC (American Type Culture Collection, Manassas, VA, USA). The two cell lines were derived from the same human subject (43 years old, female). HCC1395 cells were cultured in RPMI-1640 medium supplemented with 10% fetal bovine serum (FBS). HCC1395BL cells were cultured in Iscove's Modified Dulbecco's Medium (IMDM) supplemented with 20% FBS.

Single cell suspensions were generated by dissociating adherent cells (HCC1395) with Accutase (Innovative Cell Technologies, AT104) or by harvesting suspension cells (HCC1395 BL). We passed all cells through a 30-micron MACS SmartStrainer (Miltenyi Biotec, 130–098-458) to remove cell aggregates.

Single-cell full-length cDNA generation and RNA-seq using the C1 Fluidigm system

Single cells were loaded on a medium-sized (10–17 μm) RNA-seq integrated fluidic circuit (IFC) at a concentration of 200 cells/μl. Capture occupancy and live/dead cells at the capture site were recorded using a fluorescence microscope after staining with the live/dead viability/cytotoxicity kit (Life Technologies, L3224). Full-length cDNAs were generated using the Fluidigm C1 Single-cell System at LLU using the SMART-Seq v4 Ultra Low Input RNA kit (Takara Bio) according to the manufacturer's protocol. Only cDNAs generated from live single cells were used for further library construction.

Libraries were prepared using the modified Illumina Nextera XT DNA library preparation protocol. Briefly, the concentrations of cDNAs harvested from the IFC were quantified using the Quant-iT PicoGreen dsDNA Assay (Life Technologies) and then further diluted to 0.1–0.3 ng/μl. Then, 1.25 μl diluted cDNA was incubated with 1.25 μl tagmentation mix and 2.5 μl tagment DNA buffer for 10 minutes (min) at 55 °C. Tagmentation was terminated by adding 1.25 μl of NT buffer and centrifuged at 2,000g for 5 min. Sequencing library amplification was performed using 1.25-μl Nextera XT Index primers (Illumina) and 3.75 μl Nextera PCR Master Mix with 12 PCR cycles. Barcoded libraries were purified and pooled at equal volume. Eighty libraries were generated from HCC1395 cells (Sample A) and 66 libraries were generated from HCC1395 BL cells (Sample B). Library pools were sequenced on an Illumina HiSeq 4000 sequencer for 150×2 bp, paired-end sequencing at the LLU Center for Genomics.

Single-cell 3′ end RNA-seq using the C1 Fluidigm high-throughput (HT) system

HT single cell 3′-end cDNA libraries were generated according to the manufacturer's instructions at the FDA site. Briefly, single cells were loaded on a HT IFC at a concentration of 400 cells/μl (Nexcelom Cellometer Auto T4). Capture occupancy and live/dead cell

at the capture site were recorded using a fluorescence microscope after staining with live/dead viability/cytotoxicity kit (Life Technologies). After cell lysis, the captured mRNA was barcoded during the reverse transcription step with a barcoded primer, and the tagmentation step was done following the Nextera XT DNA library preparation guide. Only polyadenylated RNAs containing the preamplification adapter sequence at both ends were amplified. Lastly, sequencing adapters and Nextera indices were applied during library preparation. Only the 3'-ends of the transcripts were enriched following PCR amplification.

203 libraries were generated from HCC1395 cells (Sample A) and 241 libraries were generated from HCC1395 BL cells (Sample B). Library pools were sequenced on an Illumina HiSeq 2500, 75×2 bp, paired-end at the FDA's Genomics Facility.

Single-cell RNA-seq using the 10X Genomics platform

After filtering with a 30-micron MACS SmartStrainer (Miltenyi Biotec), single cells were resuspended in PBS (calcium and magnesium free) containing 0.04% weight/volume BSA (400 μg/ml), and further diluted to 300 cells/μl after cell count (Countess II FL, Life Technologies). For the 5% spike-in and 10% spike-in cell mixtures, 5% or 10% of HCC1395 breast cancer cells were mixed with either 95% or 90% of HCC1395BL cells.

Single-cell RNA-seq library preparation was performed following the 3' scRNA-seq 10X Genomics platform protocol using v2 chemistry. Briefly, based on the cell suspension volume calculator table, 3000 cells (17.4 μl of 300 cells/μl suspension) and barcode-beads as well as RT reagents were loaded into the Chromium Controller to generate single Gel Bead-in-Emulsions (GEMs). cDNAs were generated after GEM-RT incubation at 53 °C for 45 min and 85 °C for 5 min. cDNA amplification was performed for 12 PCR cycles following GEM cleanup. After cleanup with SPRIselect Reagent, cDNA was incubated for fragmentation, end-repairing, A-tailing, and adapter ligation. Lastly, sequencing library amplification was performed using sample index primer for 10 cycles.

The methods for other single-cell captures, scRNA-seq library constructions, and sequencing data generation can be found in the Online Methods section.

ONLINE METHODS

Methods for other single cell captures, library construction, and sequencing data generation 10X Genomics scRNA-seq library construction using fixed cells

We also constructed 10X scRNA-seq libraries using fixed cells at the NCI site. Briefly, for delayed captures, cells were fixed in methanol using a method described by Alles et al[49]. The fixed samples underwent two different treatments. For the sample of spikein_5%_Fixed_1, the normal and tumor cells were harvested, washed, counted, and 5% spike-ins of breast cancer cells plus 95% normal B cells were prepared and mixed as described above. Approximately 130,000 cells were then processed for fixation. The cells were washed twice with 1X DPBS at 4 °C and resuspended gently in 100 μl 1X Dulbecco's phosphate-buffered saline (DPBS, ThermoFisher Scientific). 900 μl chilled methanol (100%) was then added, drop by drop, to the cells with gentle vortexing. The cells were then fixed on ice for 15 mins, and were stored at 4 °C for 6 days. For rehydration, the fixed

cells were pelleted by centrifugation at 3000 g for 10 mins at 4 °C and washed twice with 1X DPBS containing 1% BSA and 0.4 U/μl RNase inhibitor (Sigma Aldrich). The cells were then counted, and the concentration was adjusted to be close to 1000 cells/μl. Approximately 8000 cells were loaded onto a single-cell chip for GEM generation using the 10X Genomics Chromium controller. 3'mRNA-seq gene expression libraries were prepared using the Chromium Single Cell 3′ Library & Gel Bead Kit v2 (10X Genomics) according to the manufacturer guidelines.

For the sample spikein_5%_Fixed_2, breast cancer cells and B cells (~4 million each) were harvested and fixed. The cells were initially washed with 1X DPBS and resuspended in 10% 1X DPBS and 90% chilled methanol, as described above. The cells were then fixed on ice for 15 mins, and were then stored at 4 °C for 24 hrs. For rehydration, the fixed cells were washed with 1X DPBS containing 1% BSA and 0.4 U/μl RNase inhibitor and counted. Approximately 8000 cells were loaded onto a single-cell chip for GEM generation using the 10X Genomics Chromium controller. 3'mRNA-seq gene expression libraries were prepared using the Chromium Single Cell 3′ Library & Gel Bead Kit v2 (10X Genomics) according to the manufacturer's guidelines.

All the 10X Genomics scRNA-seq libraries constructed at LLU were sequenced on a NextSeq 550 and a HiSeq 4000 at LLU Center for Genomics with the standard sequencing protocol of 26+98 bp read length, whereas the libraries constructed at the NCI site were either sequenced on a NextSeq 550 with a modified sequencing protocol of 26+57 bp read length at NCI or on a HiSeq 4000 and a NextSeq 550 using the standard sequencing protocol of 26+98 bp read length at LLU.

Single-cell RNA-seq using the Takara Bio ICELL8 platform

We also constructed ICELL8 scRNA-seq libraries at TBU. ICELL8 Cell preparation and single cell selection: A bulk cell suspension of either cancer or B cells (~ 1×10^6 each) was fluorescently labeled with a premade mix of Hoechst 33324 and Propidium Iodide (Ready Probes Cell Viability Imaging Kit, Thermo Fisher Scientific) in appropriate complete medium (RPMI-1640 medium supplemented with 10% FBS for HCC1395 cancer cells; IMDM supplemented with 20% FBS for HCC1395BL B cells) for 20 min at 37 °C. Adherent cells were first treated with Accutase as per manufacturer's instructions (Thermo Fisher Scientific) to dissociate cells from the flask surface. The cells were washed in 1X PBS (no Ca^{2+}, Mg^{2+}, Phenol Red, or serum, pH 7.4; Thermo Fisher Scientific) and centrifuged (100 g, 3 min) and resuspended in 1 ml of 1X DPBS. Cell counts were determined using a Moxie Flow cell counter (ORFLO Technologies, ID, USA) and diluted to ~1 cell in 35 nl (~ 28,600 cells/ml) in 1X DPBS (1X DPBS (no Ca^{2+}, Mg^{2+}, Phenol Red, or serum, pH 7.4; Thermo Fisher Scientific) containing Second Diluent (1X), RNase Inhibitor (0.4 U), and 1.92 μM of the 3' oligo dT terminating primer: SMART-Seq ICELL8 CDS (Takara Bio USA, CA, USA).

Each cell type solution was dispensed from a 384-well source plate into individually addressable wells in a 5,184 nano-well, 250 nl volume ICELL8 chip (SMARTer ICELL8 250v Chip, Takara Bio USA, CA, USA) using a Multi Sample Nano Dispenser (MSND, SMARTer ICELL8 Single-Cell System, Takara). Chip wells were sealed using SmartChip

Optical Imaging Film (Takara Bio USA) and centrifuged at 300 g for 5 min at 22 °C. All nano-wells in the chip were imaged with a 4X objective using Hoechst and Texas Red excitation and emission filters. Images (TIFF format) were analyzed using automated microscopy image analysis software CellSelect (Takara Bio USA). The chip was stored in a chip holder at −80 °C overnight. Image analysis confirmed cell deposition followed a Poisson distribution. 600 individual nano-wells, each bearing microscopy-identified single live cells, were chosen from each cell type. A well-selection map (filter file) was then autogenerated by CellSelect software to enable individual addressing of the chosen wells for addition of cDNA synthesis and library preparation reagents as detailed in the following sections. All on-chip liquid handling was performed with the MSND. After all dispensing and sealing steps, chips were centrifuged at 3,220 g (3 min). All on-chip thermal cycling was performed using a SMARTer ICELL8 Thermal Cycler (Takara Bio USA).

In-chip, full-length cDNA synthesis: The ICELL8 chip (containing dispensed samples) was thawed at room temperature for 10 min and centrifuged at 3,220 *g* for 3 min at 4 °C. The chip was subsequently incubated at 72 °C (3 min) and immediately placed at 4 °C. RT-PCR mix (35 nl total) was added to each of the previously selected nano-wells (identified as bearing a single cell via the ICELL8 filter file), and the reactions were thermally cycled in-chip as follows: 45.6 °C, 5 sec; 41 °C, 90 min; 99 °C, 9 sec; 95.5 °C, 1 min; 100 °C, 5 sec; 99 °C, 7 sec; 9 °C, 5 sec; 64 °C, 30 sec; 69.5 °C, 5 sec; 67.5 °C, 3 min; GoTo step 5 and repeat 7X; 4 °C hold.

In-chip, P5 index addition and tagmentation: 72 primer sequences bearing P5 indices (SMART-Seq ICELL8 Forward Indexing Primer Set A (5′-AATGATACGGCGACCACCGAGATCTACAC(*i5*)TCGTCGGCAGCGTC-3′); *i5* refers to 1-of-72 unique, 8 nucleotide indices (Hamming distance between P5 indices = 3), were dispensed from a pre-aliquoted 384-well plate in 35 nl aliquots into 72 filter-file identified, nano-well ‘rows’. The chip was sealed with Microseal A film and centrifuged at 3,220 *g* (3 min) at 4 °C before returning to the MSND, permitting addition of Tagmentation Master Mix containing: $MgCl_2$, Nextera Amplicon Tagment Mix (Illumina); Terra™ PCR Direct Polymerase Mix, and TRH (Takara Bio USA). The chip was sealed with Microseal A film and recentrifuged as above. Tagmentation was performed in-chip at the following temperatures: 42 °C, 4 sec; 37 °C for 30 min; 4 °C hold.

In-chip, P7 index and PCR reagent addition: first PCR generating 5,184 unique indices. A reagent mix containing 72 primer sequences bearing P7 indices (SMART-Seq ICELL8 Reverse Indexing Primer Set A, 5′-CAAGCAGAAGACGGCATACGAGAT(*i7*)GTCTCGTGGG CTCGG-3′)—*i7* refers to 1-of-72 unique, 8 nucleotide indices (Hamming distance between P7 indices = 3)—was dispensed from the same pre-aliquoted 384-well index plate (separate location for P7 indices) in 35 nL aliquots, into 72 filter file identified “columns” of the chip. As a consequence of adding separate P5 and P7 indices to rows or columns, a 72 × 72 $m \times n$ matrix of combinatorial P5 and P7 pairs was generated, uniquely identifying each of the 5,184 nano-wells. The chip was sealed with SmartChip Sealing Film and centrifuged at 3,220 *g* for 3 minutes at 4 °C. PCR cycling was performed as follows: 77 °C, 12 sec; 72 °C,

3 min; 99 °C, 11 sec; 95.5 °C, 1 min; 100 °C, 20 sec; 99 °C, 10 sec; 53.3 °C, 5 sec; 58 °C, 15 sec; 71 °C, 5 sec; 67.5 °C, 2 min; Go To step 5 and repeat 7X; 4 °C, hold.

Off-chip, sample extraction, and purification of round 1 PCR amplicons: Round 1 PCR amplicons were collected from the ICELL8 chip using the SMARTer ICELL8 Collection Kit: Collection Fixture, Collection Tube, and Collection Film into a collection and storage tube as per manufacturer's instructions (Takara Bio USA). 50% of the extracted library was purified twice using a 1X proportion of AMPure XP beads (Beckman Coulter) to a final volume of 14 μl in Elution Buffer, provided with the SMART-Seq ICELL8 Reagent Kit.

Off-chip library amplification (2nd PCR): Double-AMPure bead-purified, first round amplicon (14 μl, from above) was PCR amplified in a 50 μl volume of 2nd PCR Mixture containing SeqAmp™ CB PCR Buffer (25 μl), 5X Primer Mix (P5 and P7 primers), and Terra™ PCR Direct Polymerase Mix 0.05 U/ μl final concentration (Takara Bio USA) via a thermal protocol: 98 °C, 2 min x1; followed by 8 thermal cycles: 98 °C, 10 sec; 60 °C, 15 sec; 68 °C, 2 min. This sequencing-ready library was purified using 1 round of a 1X proportion of AMPure XP beads (Beckman Coulter). The final elution volume was 17 μl in Elution Buffer.

ICELL8 scRNA-seq library QC and sequencing: The scRNA-seq library concentration (ng/μl) was determined using a Qubit fluorometer (Thermo Fisher). Based on Qubit readings, 1–2 ng/μl was examined using a Bioanalyzer 2100 and a corresponding High Sensitivity DNA Kit (Agilent) to determine the size of selected libraries. The Bioanalyzer amplicon sizes ranged between 200 to 3000 bp, with an average size of 550 bp. The ICELL8 scRNA-seq libraries were sequenced both at TBU on an Illumina NextSeq 550, 75×2 bp, PE, and at LLU on a HiSeq 4000, 150×1 bp, SE.

Bulk cell RNA-seq

We isolated bulk-cell total RNA from the HCC1395 and HCC1395BL cells using miRNeasy Mini kit (QIAGEN), and constructed RNA-seq libraries using the NuGEN Ovation universal RNA-seq kit at LLU. Briefly, 100 ng of total RNA was reverse transcribed and then converted into double stranded cDNA (ds-cDNA) by addition of a DNA polymerase. The ds-cDNA was fragmented to ~200 bps using a Covaris S220, and then underwent end-repairing to blunt the ends followed by barcoded adapter ligation. The remainder of the library preparation followed the manufacturer's protocol. All the libraries were quantified using a Qubit 3.0 (Life Technologies) and quality checked on a TapeStation 2200 (Agilent Technologies). The bulk-cell RNA-seq libraries were sequenced both on a NextSeq 550, 75×2 bp, PE; and on a HiSeq 4000, 100×2 bp, PE, at LLU.

Overall data generated, and data QC assessments

Supplementary Table 1 summarizes the overall cell numbers and sequencing reads of single cells captured across all four sites. A total of 30,693 single cells were captured, from which twenty different scRNA-seq datasets were obtained (Supplementary Table 1 **&** Extended Data Fig. 1). Five libraries from the NCI site (10X_NCI_M) were also re-sequenced using a modified sequencing protocol (26+57 bp). Across all the platforms and datasets, over 93.6%

of the reads were mapped to the exonic and non-exonic regions, except for Sample A of 10X_NCI_M (modified shorter sequencing), which had a mapping rate of 87% (Sample A) and 90.3% (Sample B) (Fig. 1b). However, there were variations in the mapping rates to exonic regions across platforms and sites, with ICELL8 and Fluidigm C1 full-length transcript methods showing a higher mapping rate than 3′-transcript scRNA data in breast cancer cells (Sample A: C1_LLU_A, 83.1%; ICELL8_SE_A, 80.7%; ICELL8_PE_A, 84.0%; versus 10X_LLU_A, 65.3%; 10X_NCI_A, 66.3%). The UMI (unique molecular identifier) data generated by the 10X platform showed that 35.3% of the mapped reads (or 57.0% of the exonic reads) were derived from de-duplicated UMIs in breast cancer cells (Sample A), and 22.2% of the mapped reads (or 35.2% of the exonic reads) were derived from de-duplicated UMIs in normal B cells (Sample B). We also noticed that the exonic mapping rates were slightly lower for the 10X Genomics technologies when using the modified shorter sequencing protocol compared with the standard sequencing protocol (26+57 bp versus 26+98 bp). The modified protocol used a shorter sequencing read length, which could cause a higher percentage of non-specific mapping reads. Nevertheless, many overlapping genes were detected (96.6–97.3%) with a high correlation (R=0.997–0.998) between the standard and modified sequencing protocol for the 10X Genomics scRNA-seq (Supplementary Fig. 1).

Bioinformatics methods

Reference genome: The reference genome and transcriptome were downloaded from the 10X website as refdata-cellranger-GRCh38–1.2.0.tar.gz, which corresponds to the GRCh38 genome and Ensembl v84 transcriptome. All the following bioinformatics data analyses are based on the above reference genome and transcriptome.

Preprocessing of UMI based scRNA-seq data from the 10X platform

For UMI based 10X samples, three pre-processing pipelines, Cell Ranger (v3.1.0), umitools[24] (v1.0.0), and zUMIs[25](v2.4.5) were used to process the raw fastq data and generate gene count matrices. In the Cell Ranger pipeline, cellranger count was used with all default parameter settings. In the umitools and zUMIs pipelines, reads were filtered out if phred sequence quality of cell barcode bases was < 10 or UMI bases < 10. In the zUMIs pipeline, option -d was used to perform down-sampling analyses to 8 fixed depths (5k, 10k, 25k, 50k, 100k, 150k, 200k, and 250k) to generate gene count tables. With umitools, umi_tools whitelist with default parameter settings was used to generate a list of cell barcodes for downstream analysis. umi_tools extract was used to extract the cell barcodes and filter the reads (options: --quality-filter-threshold=10 --filter-cell-barcode). STAR (v2.5.4b)[31] was used for alignment to generate bam files containing the unique mapped reads (option: outFilterMultimapNmax 1) for gene counting. featureCounts (v1.6.1)[26] was used to assign reads to genes and generate a BAM file (option: -R BAM). samtools (v1.3)[50] sort and samtools index were used to generate sorted and indexed BAM files. Finally, umitools count (options: --per-gene --gene-tag=XT --per-cell --wide-format-cell-counts) was used for the sorted BAM files to generate gene count per cell matrices

Preprocessing of non-UMI based scRNA-seq data from C1 and TBU ICELL8 platforms

For non-UMI based samples, three pre-processing pipelines were compared for processing the raw fastq data and generating gene count matrices. The pipelines included trimming and filtering, alignment, and gene counting. In the trimming and filtering process, one of the three tools [Trimmomatic (v0.35)[30], trim_galore (v0.4.1)[51], or cutadapt (v1.9.1)[29]] was used to process the raw fastq data. Bases with quality less than 10 were trimmed from 5′ and 3′ ends of reads. Reads fewer than 20 bases were excluded from further analysis. STAR with default parameter settings was used for alignment to generate bam files. Three gene counting tools: featureCounts, RSEM (v1.3.0)[28], or kallisto (v0.43.1)[27] were used to generate gene counts per cell. All default parameter settings were used except the following: In RSEM, option --single-cell-prior was used to estimate gene expression levels for scRNA-seq data; Option --paired-end was used if the data were paired-end fastqs; In kallisto, options -l 500 and -s 120 were used to represent estimated average fragment length and standard deviation of fragment length if the data were single-end fastqs.

Preprocessing and differential gene expression analysis of bulk RNA-seq data

The preprocessing pipeline of bulk RNA-seq data included QC (FastQC v0.11.4), trimming and filtering (Trimmomatic), alignment (STAR), and gene counting (RSEM). The parameter settings in the pipeline were the same as the preprocessing pipelines used for non-UMI scRNA-seq data. In RSEM, the option --single-cell-prior was turned off for estimation of gene expression levels in bulk RNA-seq data. DESeq2 (v1.24.0) was used to perform the differential expression analysis between breast cancer samples (Sample A) and B lymphocyte samples (Sample B) with default parameters.

BGL and data sharing within the team

Working under the FDA single-cell sequencing consortium, to streamline fast pre-processed data sharing, access, and analysis, we used the BioGenLink (BGL) platform from Digicon as a central repository to host the pre-processed data as described above. All data including the single-cell RNA-seq data were pre-processed at LLU and then the pre-processed data were either uploaded into BGL from LLU servers or by using tools within BGL that utilized Globus, file transfer protocol (FTP), and secure copy protocol (SCP). Detailed data annotation files about all genomics data were also uploaded into the BGL.

Normalization methods across all datasets

We investigated some existing bulk RNA-seq normalization procedures including ‘Counts per Million (CPM)’, ‘Trimmed Mean of M values (TMM)’, ‘Upper Quantiles’, ‘DESeq’ normalization implemented in the DESeq Bioconductor package, and ‘Trimmed Mean of M values (TMM)’ implemented in edgeR. There were also methods that were specifically tailored to scRNA-seq datasets, such as SCTransform, scran, and Linnorm. Both scran and Linnorm were run using default parameters. SCTransform was run without regressing out any variables with default settings.

We performed down-sampling of each cell to two different read depths (10K and 100K per cell) for each dataset and evaluated the performance of the normalization methods at these depths. Similar to the method used in the *scone* paper[52], the metric we used to assess

normalization methods was based on how well the two samples from the same cell were grouped with each other. Specifically, we used silhouette width, which is defined as:

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}}$$

For each cell i, let $a(i)$ be the average distance between I and all other cells within the same cluster. Let $b(i)$ be the lowest average distance of I to all points in any other cluster, of which I is not a member. Here we defined the clustering structure that the same cells from two different sequencing runs form a single cluster, thus we have a total number of n/2 clusters if the total number of samples is n.

We calculated the silhouette width values for each dataset. The larger the silhouette width values, the better the performance of the normalization method.

scRNA-seq data batch effects and batch-effect correction pipelines

We used the gene count matrix from the Cell Ranger pipeline (10X Genomics data) and the STAR-featureCounts pipeline (non-10X Genomics data) as input to evaluate batch correction methods. In the Cell Ranger pipeline, both Cell Ranger 2.0 and Cell Ranger 3.1 were applied to 10X Genomics data. The batch correction evaluation of the data processed by Cell Ranger 2.0 are shown as supplementary figures. For the evaluation, three different conditions were considered: (1) all datasets; (2) datasets with biologically similar cells; and (3) datasets with biologically different cells. The evaluation procedure included the following four major steps:

1. Monocle2[53, 54] strategy to filter dead cells and doublets for 10X Genomics single cell data
2. Single-cell data processing and highly variable gene (HVG) selection
3. Batch correction by seven different methods
4. Evaluation by t-SNE or UMAP, kBET (kBET v0.99.5) acceptance score, modified alignment score, and silhouette score.

A detailed description and functions used for batch correction are summarized in Supplementary Table 14.

In step 1, all 10X single cell datasets were processed by the Monocle2 strategy to filter dead cells and doublets. In this strategy, the total numbers of UMIs and genes for each cell were counted. The upper bound was calculated as mean plus two standard deviations (SD) and the lower bound as mean minus two SD for both the total UMIs and genes. Cells with total UMIs or genes outside of the upper and lower bounds were removed.

In step 2, Seurat (v3.0.3) based data processing was applied to each dataset. Genes detected in fewer than 3 cells and cells containing less than 200 genes were removed from the datasets prior to further analysis. The datasets were then log transformed and scaled. The top 2,000 HVGs were selected in each dataset with the function *FindVariableGenes* for the five R-based batch correction methods Seurat, fastMNN (scran v1.12.1 and SeuratWrappers

v0.1.0), Harmony (v0.99.9), limma (v3.40.4), and ComBat (v3.32.1). For the Python-based batch correction methods Scanorama (v1.4) and BBKNN (v1.3.5), the detailed description of data processing is provided in the 'Scanorama processing' and 'BBKNN processing' sections.

In step 3, the processed data and HVGs in step 2 were used as input to perform batch correction. The main functions and parameter settings of the seven batch correction methods are summarized in Supplementary Table 14.

In step 4, the t-SNE plots and UMAPs and the calculations of the kBET acceptance scores, modified alignment scores, and silhouette scores were based on the low-dimensional embedding matrices of each batch correction method. For Seurat v3, fastMNN, and Harmony, Seurat and SeuratWrappers were used to generate low-dimensional embedding matrices. Seurat-based principal component analysis (PCA) reduction was applied to batch corrected matrices by Scanorama, limma, and ComBat to generate low-dimensional matrices, whereas for BBKNN, the UMAP coordinate matrices were used as the low-dimensional embedding matrices. The functions used to generate t-SNE plots and UMAPs can be found in Supplementary Table 14.

Scanorama pipeline

The Scanorama Python package was used to process the datasets and perform batch correction. The script *process.py* with default parameters was used to perform cell filtering and normalization. 2000 HVGs were used in the function *correct* to perform batch correction and generate Scanorama-corrected gene expression matrices.

BBKNN pipeline

The Seurat-inspired Scanpy (v1.4.4) Python workflow was applied to process the datasets. All datasets were input using the function *pd.read_csv* in the pandas package, transferred into annotated data matrices, and appended into a list using the function *anndata.AnnData* from the package anndata. Cells and genes were filtered using functions *scanpy.api.pp.filter_cells* and *scanpy.api.pp.filter_genes* with the same parameter settings as at Step 1. The processed data matrices were merged to generate a master gene expression matrix and further log transformed and normalized by functions *scanpy.api.pp.log1p* and *scanpy.api.pp.normalize_per_cell*. The top 2,000 HVGs were selected from the merged gene expression matrices by the function *filter_genes_dispersion* with the same parameter settings as at Step 2. Further log transformation (function *scanpy.api.pp.log1p*) and scaling (function *scanpy.api.pp.scale*) were performed for the newly generated gene expression matrices. The function *bbknn* with default parameters was carried out for the batch correction.

FastMNN versus MNN

In Supplementary Figure 8, we compared the performance of fastMNN and MNN. The steps to perform MNN correction are the same as fastMNN except the batch correction (step 3). We used the function *mnnCorrect* (scran package v1.8.4) with default parameters to perform the batch correction.

Preprocessing and batch-effect correction on Tian et al. data[16]

We preprocessed Tian's data using the Cell Ranger pipeline for 10X data and the umitools pipeline for their non-10X data. The same procedures for the seven batch correction methods described previously were applied to the preprocessed data to perform batch correction evaluation.

Bioinformatics pipelines validated and performed in BGL

We carried out some bioinformatics pipelines in BGL to cross-validate some of our bioinformatic data analyses. Bioinformatics tools were created in BGL for performing batch correction of single-cell RNA-seq data using Seurat v3, fastMNN, Scanorama, BBKNN, Harmony, limma, and Combat and for visualizing the results of each procedure using t-SNE and UMAP for Scenario # 1. For each procedure, a tool was created in BGL that allows a user to point and click to select input data and parameters for running methods from one or more packages. For each tool, BGL ran a script on the back end to execute the steps described below. Unless otherwise stated, all functions and procedures used default settings.

Silhouette width to quantify batch-effect correction

The silhouette width score of each cell was calculated based on the two cell types, HCC1395 and HCC1395BL, for Scenarios #1 and #4 (Fig. 4a/d) by the function *silhouette* from the R package cluster (v.2.0.8). We further calculated the average silhouette width scores of the cells in each cluster. Finally, the mean of the average silhouette width score was used to represent the performance of the seven batch correction methods.

kBET acceptance score to quantify batch-effect correction

kBET acceptance scores were calculated using the Buttner et al.[45] pipeline for four different sample combination scenarios (Fig. 4a–d) to assess the batch correction performance. This metric was calculated using the low-dimensional embedding matrices of each batch correction method. For Seurat v3, fastMNN, and Harmony, both Seurat and SeuratWrappers were used to generate low-dimensional embedding matrices. Seurat-based PCA reduction was applied to batch corrected matrices by Scanorama, limma, and ComBat to generate low-dimensional matrices, whereas for BBKNN, UMAP coordinate matrices were used as low-dimensional embedding matrices for the evaluation. The score was calculated for either breast cancer cells or B lymphocytes across different batches.

Modified alignment score to quantify batch-effect correction

We adopted the idea of alignment score from Butler et al.[7] to calculate alignment scores based on the cells' embedding in two-dimensional space constructed by t-SNE or UMAP. Like kBET, this metric was also calculated with the low-dimensional embedding matrices of each batch correction method as described above. The score was calculated for either breast cancer cells or B lymphocytes across different batches of scRNA-seq datasets for each of four sample combination scenarios (Fig. 4a–d). However, due to the difference in cell numbers across different datasets in our study, we developed a modified alignment score calculation algorithm as follows:

1. Calculate the percentage of cells in each dataset i as w_i ($i = 1 \ldots N$, N is the total number of datasets).
2. For each cell $j (j = 1 \ldots N_j)$ of dataset i, calculate how many of its k nearest-neighbors belong to the same dataset as x_{ij} and then take an average of x_{ij} in dataset i to get $\acute{x}_i$.
3. Alignment score = $\sum_{i=1}^{N} w_i \left(1 - \frac{\acute{x}_i - w_i k}{k - w_i k}\right)$
4. We chose k to be 1% of the total number of cells, as recommended by Butler et al[7].

Evaluation of global and cell-type specific gene expression consistency across platforms/sites using all scRNA-seq data

To investigate the consistency of global gene expression across different platforms/sites and scRNA-seq datasets, we selected benchmarking genes according to the average gene expression (log2(TPM+1)) determined by bulk RNA-seq (three biological replicates) from samples A and B. We excluded the top 0.1% highly expressed genes to avoid abnormally expressed genes. To obtain the robust genes, we further filtered out genes with standard deviation of gene expression greater than 1 across three replicates to obtain robust genes. The remaining genes were used to define three different expression groups by selecting the top 500 most highly expressed, 500 intermediately expressed, and 500 rarely expressed genes based on the ranking of average gene expression levels. For the 1500 genes selected, we calculated the percentage of cells per gene by defining the percentage of cells with the expressed gene (gene counts >= 1) for different scRNA-seq datasets. To get comparable cell percentages, we considered only gene count matrices from the down-sampling results (100K reads per cell) of the zUMIs (10X datasets) and featureCounts (non-10X datasets) pipelines. The Pearson correlations for the percentages of cells between any two scRNA-seq datasets were calculated for each of the three expression groups to evaluate the consistency of gene expression of cell-type specific genes across different platforms and datasets.

Scatter plotting

To assess the variation of gene expression across different datasets, we generated scatterplot matrices. For all platform-specific datasets, which include 7 single-cell datasets and 3 bulk cell RNA-seq datasets for each cell line, the raw gene count matrices were converted to normalized gene lists $L^{(i)}$ by computing the average gene expression count $G_{mj}^{(i)}$ of all cells N:

$$L_m^{(i)} = \frac{1}{N} \sum_{j=1}^{N} G_{mj}^{(i)}, \quad (m = 1, \ldots, M) \text{ Where } G_{mj}^{(i)} = log\left(CPM_{mj}^{(i)} + 1\right).$$

$$L^{(i)} = \begin{pmatrix} L_1^{(i)} \\ \vdots \\ L_M^{(i)} \end{pmatrix}, \quad (i = 1, \ldots, 8)$$

For $i = 1, \ldots, 8$ of gene list $L^{(i)}$, this gives 8 columns which can be grouped to an M x 8 matrix as

$$A = \begin{pmatrix} L_1^{(1)} & \ldots & L_1^{(8)} \\ \vdots & \vdots & \vdots \\ L_M^{(1)} & \ldots & L_M^{(8)} \end{pmatrix}.$$

The final normalized matrix A was used as input to generate scatterplot figures using R packages ggplot2 and psych. Scatterplots display read count distributions across all genes and all datasets. Each gene is represented as a point in each scatterplot; x, y values represent the gene expression variation in a pair of datasets compared. In addition, each sample's gene expression distributions were computed and displayed in a bar chart. Pearson correlation coefficients between any pair of datasets were calculated to show the consistency of gene expression across different datasets.

Violin plotting

To assess the scRNA-seq gene expression profiles across different platforms based on 4 different RNA groups including protein coding RNAs, antisense RNAs, lincRNAs, and miscRNAs, we took the raw gene count matrices for each dataset and converted them to a normalized gene list $L^{(i)}$ by computing the average gene expression count $G_{mj}^{(i)}$ of all cells. Please see scatter plotting normalized count matrix computing method for details. The genes which had zero expression were removed from the comparison; the filtered gene expression lists were used to extract the specific RNA group genes to generate violin plots using R version 3.6.0, ggplot2_3.2.0, and dplyr_0.8.3 packages.

Feature plotting

For the un-corrected data, Seurat objects from t-SNE dimensional reduction were used as the data source for generating feature plots. A total of 20 genes (10 for Sample A, cancer cells; 10 for Sample B, B cells) were selected as cell-type specific markers for Sample A and B based on both the bulk-cell RNA-seq DEG ranking and literature. Each cell was assigned a "CellType" (A if the cell came from cancer cells, B if the cell came from B cells). The Seurat function *FeaturePlot* with default parameters was run to generate the gene expression feature plots, in which each cell was colored based on the expression level of the selected gene.

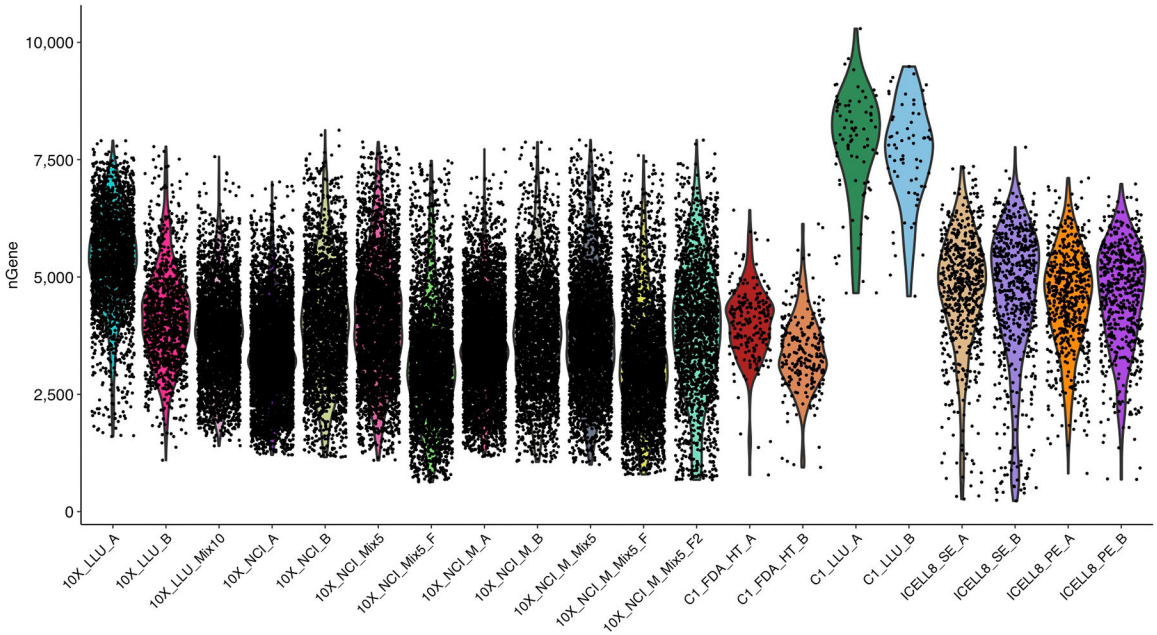
Bioinformatics methods for single-cell detection consistency of cell-type specific markers CD40, CD74, and TPM1

To examine the consistency of three marker genes across different single-cell platforms/datasets, we used the normalized gene expression data (CPM value) from the down-sampling results (100K reads per cell) of the zUMIs (10X datasets) and featureCounts (non-10X datasets) pipelines. The expression matrices of three marker genes per cell were generated. The cell percentages with detectable, low, intermediate, or high expression were defined by the percentage of cells with CPM > 0, 0 < CPM < 1, 1 ≤ CPM < 10, and CPM ≥10.

Methods used to generate summary benchmarking statistics for the various bioinformatics pipelines

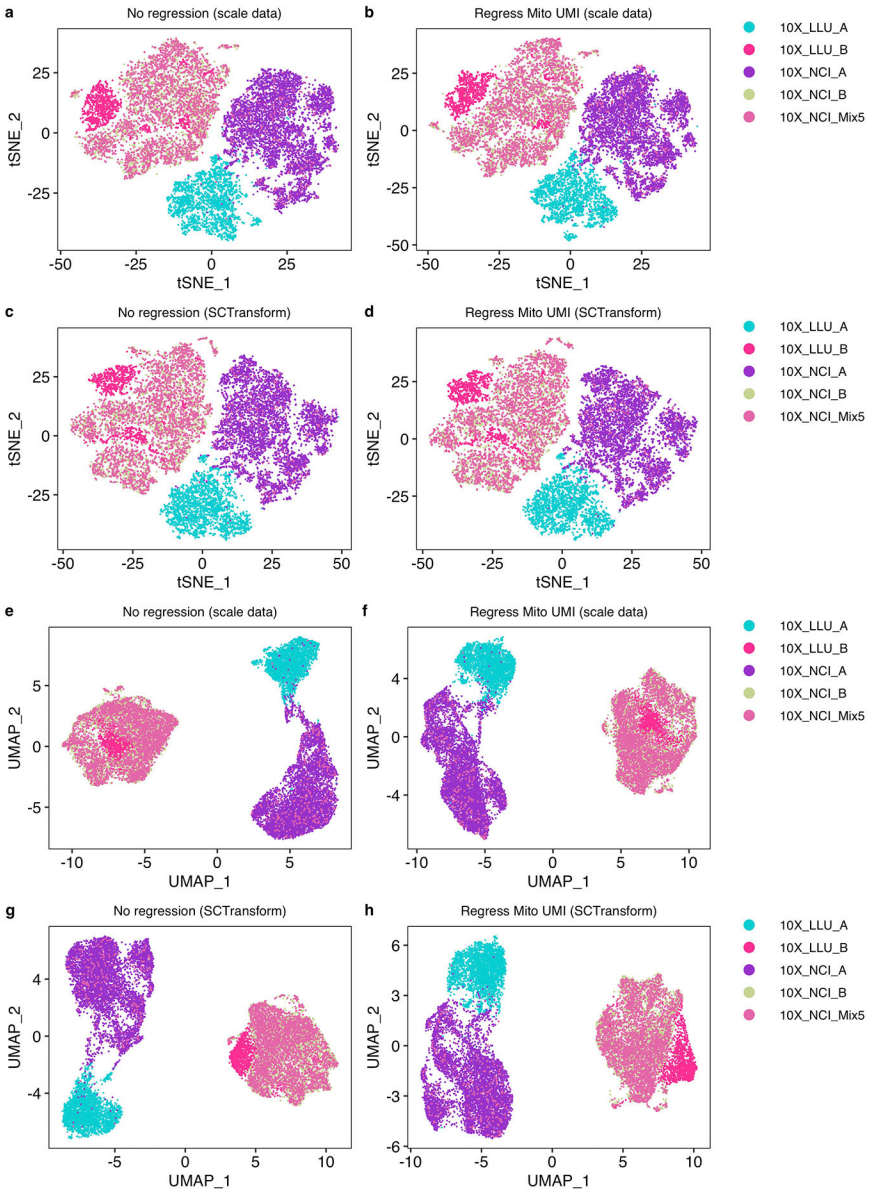
The performance of the various pipelines regarding preprocessing, normalization and batch-effect correction is summarized in Figure 6a–d based on a Z-score statistic calculated for each metric as detailed below. To benchmark preprocessing methods in terms of gene detection, we first grouped the fourteen pairwise datasets representing either the B cell line or the breast cancer cell line (Fig. 1b) into three categories: those processed using the 10X platform (6 datasets), 3′ end counting using the Fluidigm high-throughput platform (2 datasets), and the two full-length-based platforms (6 datasets). For each dataset, we calculated the proportion of the number of genes detected per pipeline compared with the maximum number of genes detected for that group. Then, for each pipeline, the average scaled ratios of the detected genes within the three categories were calculated. Finally, a Z-score was calculated based on the average scaled ratios per category per preprocessing pipeline. To assess clusterability of normalization, we determined Z-scores for both the median and variance of the calculated silhouette width scores of the fourteen paired cancer cell vs. B cell datasets as depicted in (Fig. 3). Batch-effect correction performance was assessed in terms of clusterability (ability to separate different cell types from each other) and mixability (ability to group similar cells together across datasets). To assess clusterability of batch-effect correction, a Z-score was derived from the harmonic means calculated for the silhouette width scores obtained from the datasets combining both Scenario #1 (Fig. 4a; combination of all 20 datasets in a single analysis) and Scenario #4 (Fig. 4d; spiked-in data). To assess mixability of batch-effect correction, a Z-score was derived from the harmonic means of the kBET acceptance scores obtained from datasets of all four tested scenarios (Fig. 4a–d/g).

Extended Data



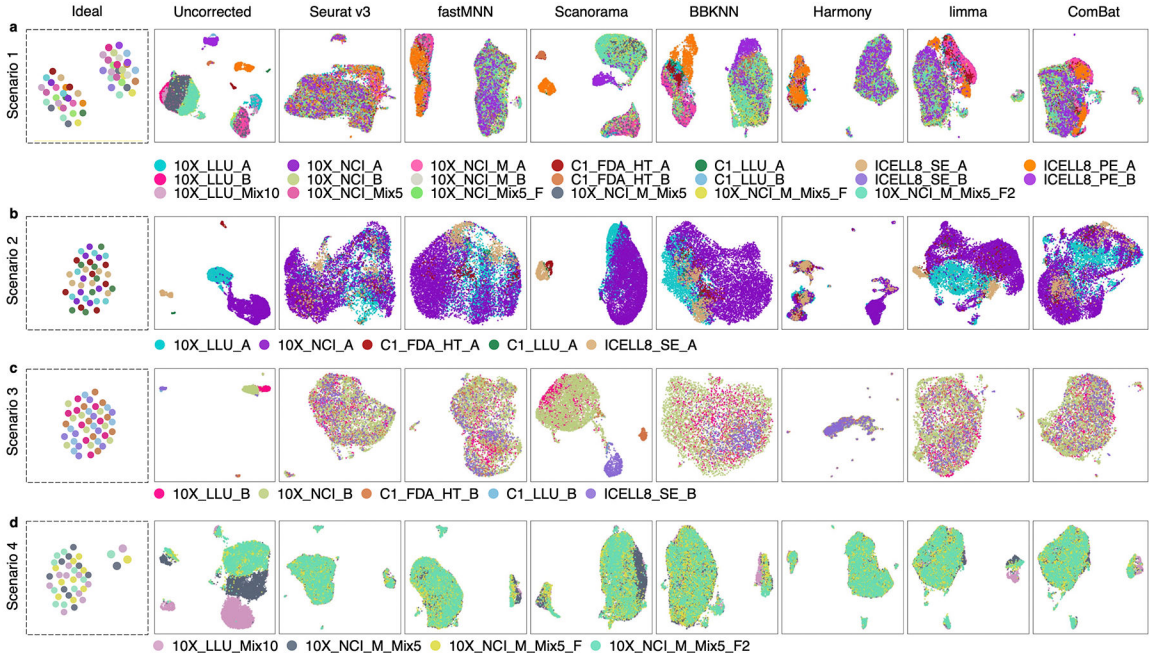
Extended Data Figure 1. An overview of the number of genes detected in each cell across all datasets.

The violin plot shows the number of genes detected in each cell across 20 scRNA-seq datasets. The plot was generated using Seurat (version 3.1). Each dot represents a single cell. The violin shapes summarize the data distributions, which are colored in the background to signify each of the 20 different scRNA seq datasets. Each scRNA-seq dataset is plotted on the X-axis; the Y-axis shows the corresponding number of genes detected in a cell (nGene) for that dataset. The average number of genes detected in each cell was about 4000 and most of the cells had 2500–7500 genes, except for samples C1_LLU_A and C1_LLU_B. The 10X Genomics scRNA datasets were preprocessed using CellRanger 3.1.



Extended Data Figure 2. Regressing mitochondrial genes and normalizing UMI did not remove batch effects.

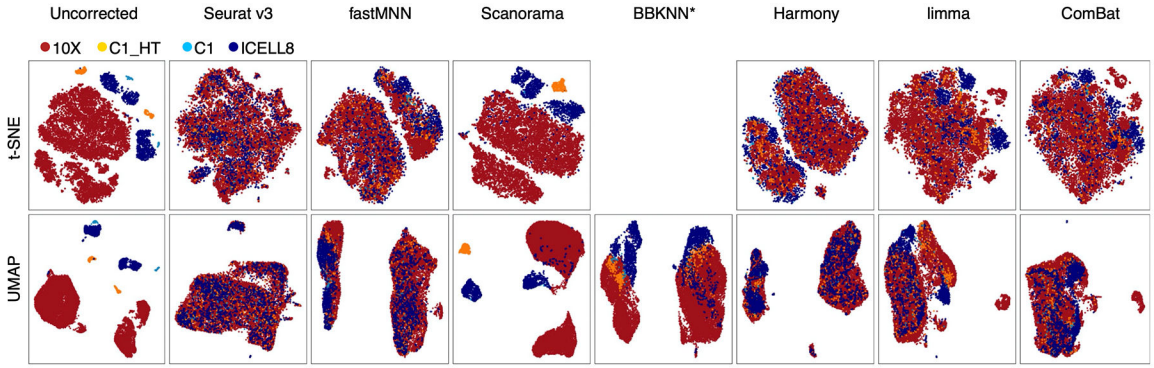
Five different batches of scRNA-seq data (10X_LLU_A, 10X_LLU_B, 10X_NCI_A, 10X_NCI_B, and 10X_NCI_Mix5) generated at two sites (LLU and NCI) are shown either as t-SNE plots (panels a-d) or as UMAPs (panels e-h). (**a**) LogNormalized, scaled data with no regression; (**b**) LogNormalized, scaled data filtered with mitochondrial (Mito) gene regression >5% and UMI normalization by Seurat v3; (**c**) ScTransform with no regression; (**d**) SCTransform with mitochondrial gene regression and UMI normalization; (**e**) LogNormalized, scaled data with no regression; (**f**) scaled data with mitochondrial gene regression and UMI normalization; (**g**) SCTransform with no regression; and (**h**) SCTransform with mitochondrial gene regression and UMI normalization.



Extended Data Figure 3. UMAP showing batch effect correction by mixability and clusterability using scRNA-seq datasets in four different sample scenarios.

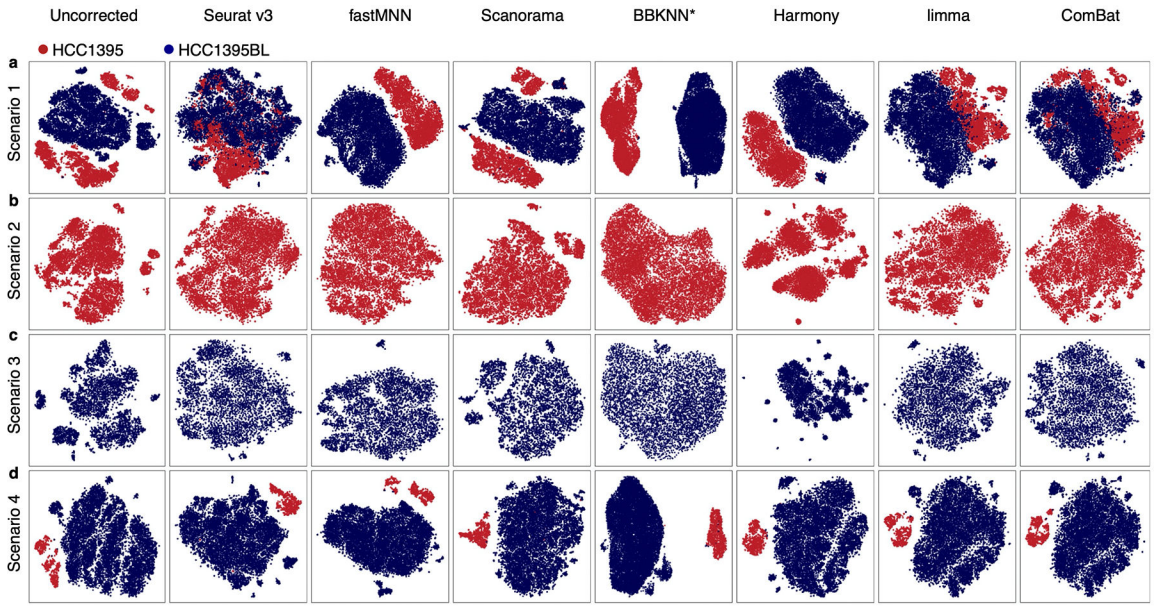
Batch-effect corrections were performed for the following four scenarios: **(a)** Scenario 1, where all 20 scRNA-seq datasets were combined, including mixed and non-mixed, with large proportions of two dissimilar types of cells (Sample A, breast cancer cell line HCC1395 and Sample B, B-lymphocyte line HCC1395BL); Datasets from 10X were down-sampled to 1200 cells per dataset. (**b**) Scenario 2, where five datasets (10X_LLU_A, 10X_NCI_A, C1_FDA_HT_A, C1_LLU_A, and ICELL8_SE_A) from the breast cancer cells (Sample A, HCC1395) were generated separately at four centers (LLU, NCI, FDA, and TBU) on four platforms (10X, Fluidigm C1, Fluidigm C1_HT, and TBU ICELL8); (**c**) Scenario 3, where five datasets (10X_LLU_B, 10X_NCI_B, C1_FDA_HT_B, C1_LLU_B, and ICELL8_SE_B) from B-lymphocytes (Sample B, HCC1395BL) were generated separately at four centers (LLU, NCI, FDA, and TBU) on four platforms (10X, Fluidigm C1, Fluidigm C1_HT, and TBU ICELL8); and (**d**) Scenario 4, where four datasets (10X_LLU_Mix10, 10X_NCI_M_Mix5, 10X_NCI_M_Mix5_F, and 10X_NCI_M_Mix5_F2) were generated from 5% or 10% of breast cancer cells (Sample A, HCC1395), spiked into the B-lymphocytes (Sample B, HCC1395BL), and analyzed with

the 10X Genomics platform at two centers (LLU and NCI) in four different batches. Batch correction methods included Seurat v3.1, fastMNN (SeuratWrappers v0.1.0), Scanorama V1.4, BBKNN V1.3.5, Harmony V0.99.9, limma V3.40.4, and Combat (sva V3.32.1). The top 2000 highly variable genes (HVGs) of these datasets were used as the gene set for batch correction. All the 10X data were preprocessed using CellRanger 3.1.



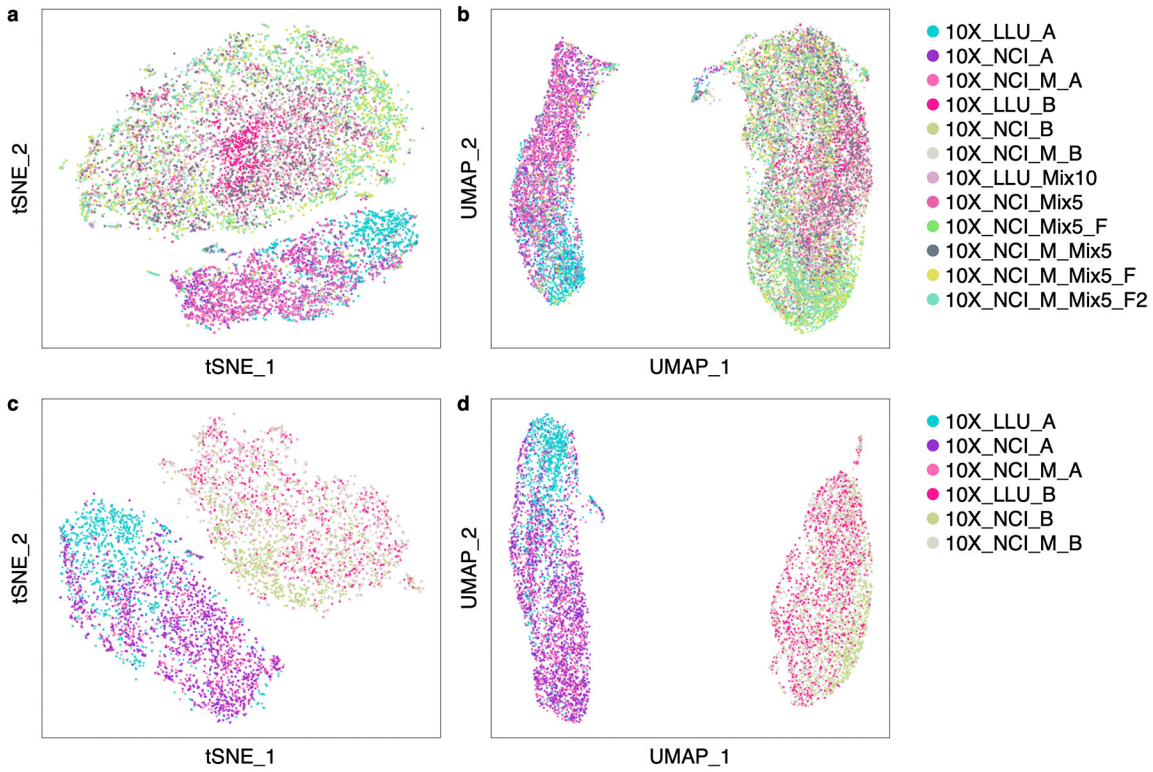
Extended Data Figure 4. t-SNE plots and UMAPs showing batch-effect corrections by mixability and clusterability across four scRNA-seq platforms.

t-SNE plots and UMAPs showing the batch-effect corrections performed by seven methods using 20 scRNA-seq datasets across different platforms. Datasets from 10X were down-sampled to 1200 cells per dataset. *Note, for BBKNN, only UMAP was available and shown. The scRNA-seq datasets are colored to identify the four different platforms: 10X 3′ scRNA-seq platform (red), C1 3′ HT scRNA-seq platform (yellow), C1 full-length scRNA-seq platform (light blue), and ICELL8 full-length scRNA-seq platform (dark blue). Batch correction methods included: Seurat v3.1, fastMNN (SeuratWrappers v0.1.0), Scanorama V1.4, BBKNN V1.3.5, Harmony V0.99.9, limma V3.40.4, and Combat (sva V3.32.1). Scanorama failed to separate two cell types into discrete clusters when non-10X platforms were included in the analysis. The top 2000 HVGs across all datasets were used as the gene set for batch correction. All the 10X data were preprocessed using CellRanger 3.1.



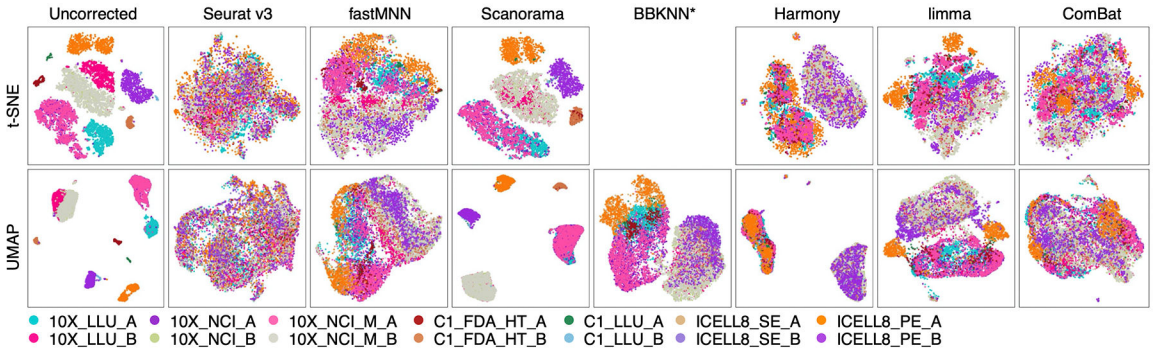
Extended Data Figure 5. Batch effect correction displayed by cell type identity.

Batch-effect corrections were performed for the following four scenarios: **(a)** Scenario 1, where all 20 scRNA-seq datasets were combined, including mixed and non-mixed, with large proportions of two dissimilar types of cells (Sample A, breast cancer cell line HCC1395 and Sample B, B-lymphocyte line HCC1395BL); Datasets from 10X were down-sampled to 1200 cells per dataset. (**b**) Scenario 2, where five datasets (10X_LLU_A, 10X_NCI_A, C1_FDA_HT_A, C1_LLU_A, and ICELL8_SE_A) from the breast cancer cells (Sample A, HCC1395) were generated separately at four centers (LLU, NCI, FDA, and TBU) on four platforms (10X, Fluidigm C1, Fluidigm C1_HT, and TBU ICELL8); (**c**) Scenario 3, where five datasets (10X_LLU_B, 10X_NCI_B, C1_FDA_HT_B, C1_LLU_B, and ICELL8_SE_B) from B-lymphocytes (Sample B, HCC1395BL) were generated separately at four centers (LLU, NCI, FDA, and TBU) on four platforms (10X, Fluidigm C1, Fluidigm C1_HT, and TBU ICELL8); and (**d**) Scenario 4, where four datasets (10X_LLU_Mix10, 10X_NCI_M_Mix5, 10X_NCI_M_Mix5_F, 10X_NCI_M_Mix5_F2) were generated from 5% or 10% of breast cancer cells (Sample A, HCC1395) spiked into the B-lymphocytes (Sample B, HCC1395BL) and analyzed with the 10X Genomics platform at two centers (LLU and NCI) in four different batches. *For BBKNN, only UMAPs were available and shown in (**a-d)**. The HCC1395 breast cancer cells (Sample A) were labeled in red and the HCC1395BL B lymphocytes (Sample B) were labeled in blue. Batch correction methods included Seurat v3.1, fastMNN (SeuratWrappers v0.1.0), Scanorama V1.4, BBKNN V1.3.5, Harmony V0.99.9, limma V3.40.4, and Combat (sva V3.32.1). The top 2000 HVGs were used as the gene set for batch correction. All the 10X data were preprocessed using CellRanger 3.1.



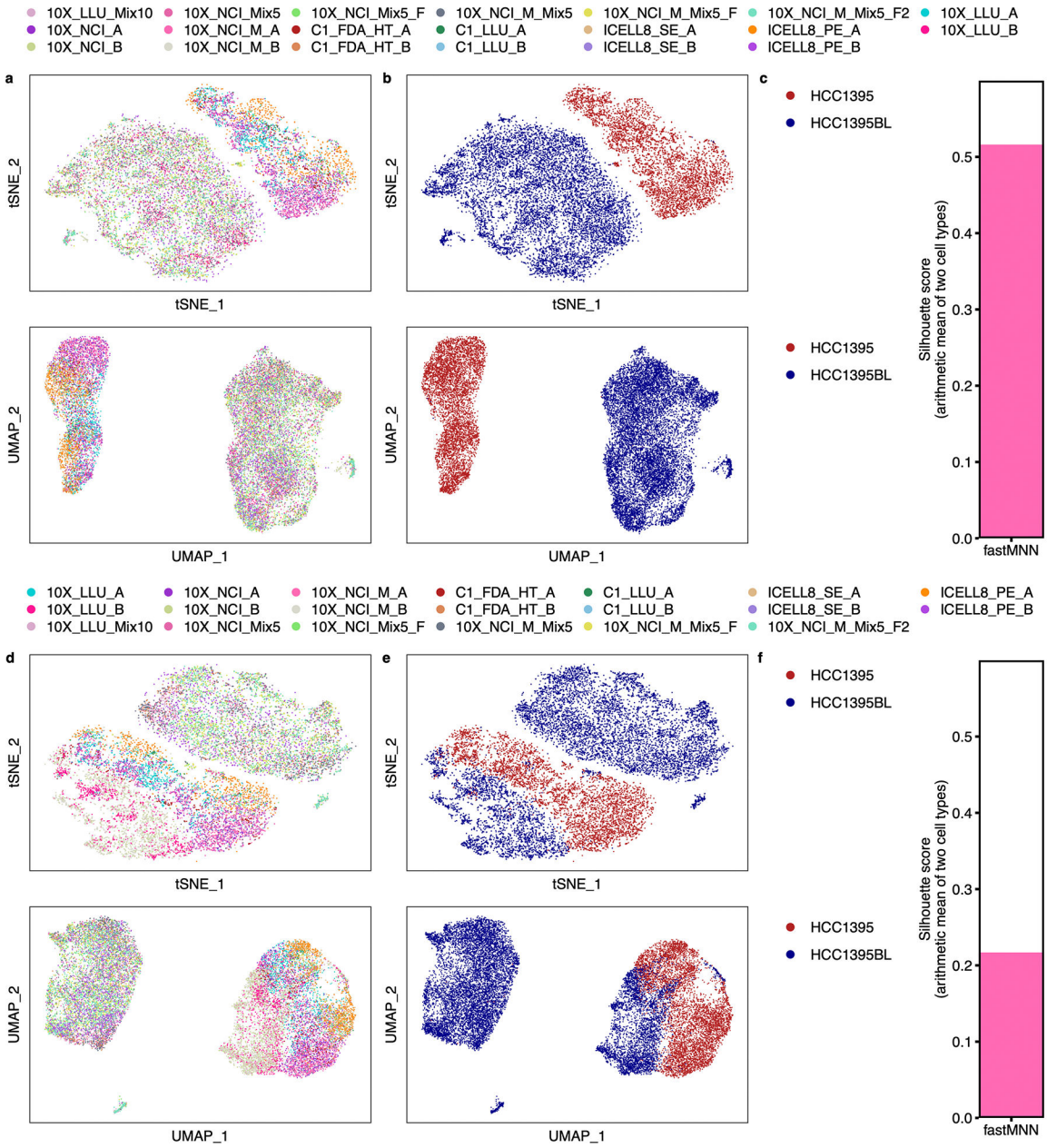
Extended Data Figure 6. Scanorama worked well for 10X Genomics scRNA-seq datasets regardless of the presence of shared cell types across batches.

(a) t-SNE plot and **(b)** UMAP showing batch-effect corrections using twelve 10X Genomics scRNA-seq datasets consisting of both mixed and non-mixed samples from two sites (LLU and NCI) in different batches after Scanorama (version 1.4.) batch correction. **(c)** t-SNE plot and **(d)** UMAP showing projections of batch-effect corrections using six 10X scRNA-seq datasets consisting of only non-mixed samples from two sites (LLU and NCI) in different batches after Scanorama (version 1.4.) batch correction. Different colors represent different datasets. All the datasets were down-sampled to 1200 cells per dataset. After the batch correction, cells from the same cell line type clustered together and mixed adequately within the same cell types. All the data were preprocessed using CellRanger 3.1.



Extended Data Figure 7. Batch-effect correction evaluating clusterability using 14 scRNA-seq datasets without spiked-in mixtures.

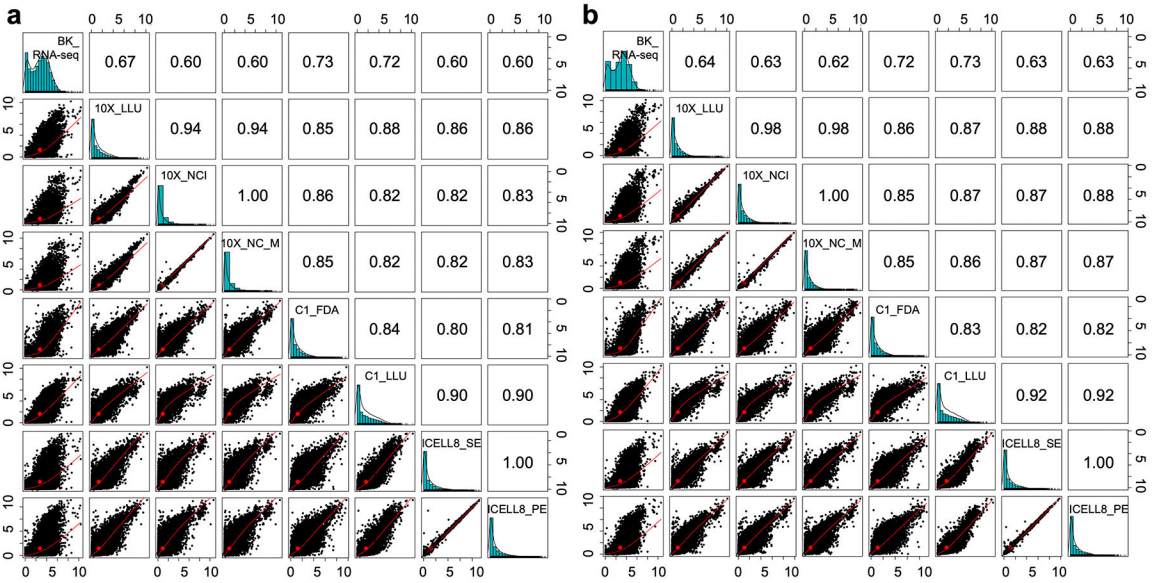
t-SNE plots and UMAPs showing batch-effect corrections performed by seven methods using 14 non-mixture scRNA-seq datasets across different platforms and sites. Six spiked-in mixture scRNA-seq datasets (10X_LLU_Mix10, 10X_NCI_Mix5, 10X_NCI_Mix5_F, 10X_NCI_M_Mix5, 10X_NCI_M_Mix5_F, and 10X_NCI_M_Mix5_F2) were removed from the 20 datasets in Scenario 1 for batch-effect correction evaluation. The fourteen non-mixture scRNA-seq datasets are from both breast cancer cells (10X_LLU_A, 10X_NCI_A, 10X_NCI_M_A, C1_FDA_HT_A, C1_LLU_A, ICELL8_SE_A, and ICELL8_PE_A) and B-lymphocytes (10X_LLU_B, 10X_NCI_B, 10X_NCI_M_B, C1_FDA_HT_B, C1_LLU_B, ICELL8_SE_B, and ICELL8_PE_B). Datasets from 10X were down-sampled to 1200 cells per dataset. *Note, for BBKNN, only UMAP was available and shown. Batch correction methods included Seurat v3.1, fastMNN (SeuratWrappers v0.1.0), Scanorama V1.4, BBKNN V1.3.5, Harmony V0.99.9, limma V3.40.4, and Combat (sva V3.32.1). All the 10X data were preprocessed using CellRanger 3.1.



Extended Data Figure 8. fastMNN batch-effect correction depends on the order of importing scRNA-seq data into the pipeline.

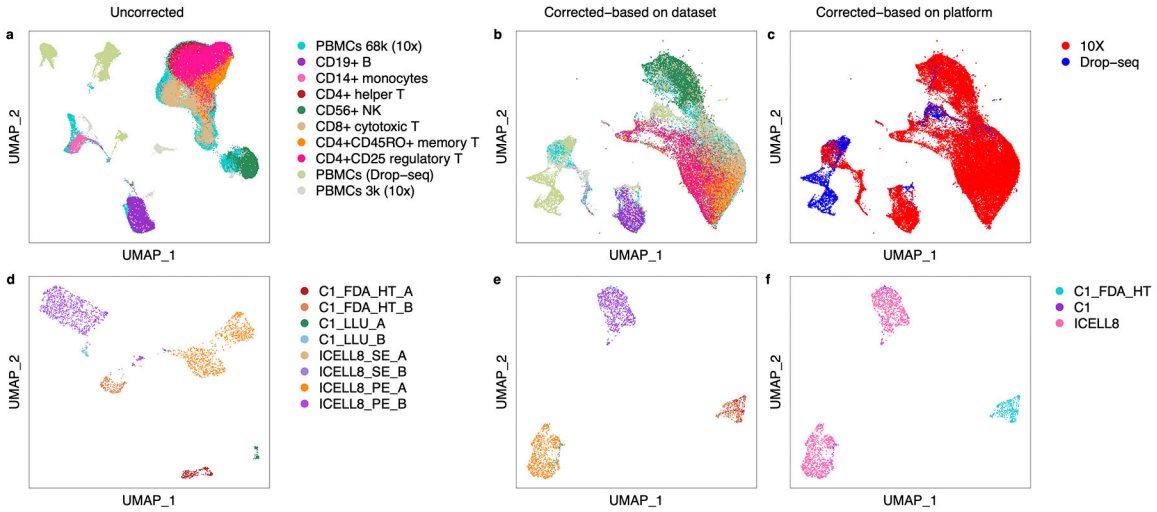
Panels (**a-c**) show results obtained using fastMNN when the spiked-in (mixed) datasets (i.e., 10X_LLU_Mix10, 10X_NCI_Mix5, 10X_NCI_Mix5_F, 10X_NCI_M_Mix5, 10X_NCI_M_Mix5_F, and 10X_NCI_M_Mix5_F2) were imported into the pipeline before other non-mixed scRNA-seq datasets from the 20 scRNA-seq datasets of Scenario 1. **(a)** t-SNE vs. UMAP with color-coding by dataset; **(b)** tSNE vs. UMAP, colored by cell types (HCC1395, red; HCC1395BL, blue); and **(c)** A *silhouette* score = 0.52 showing that fastMNN correctly separated the two cell types into two clusters representing breast cancer cells and B lymphocytes. Panels (**d-f**) show results obtained using fastMNN when the non-mixed datasets were imported into the pipeline before the mixture datasets. **(d)** tSNE

vs. UMAP with color-coding by datasets or (**e**) tSNE vs. UMAP colored by cell types; and **(f)** A low *silhouette* score of 0.22 showing that fastMNN had difficulty correctly separating the two cell types in this case. Batch-effect corrections were performed using fastMNN (SeuratWrappers v0.1.0) and *silhouette* width scores were calculated using the *silhouette* function from the R package *cluster* (v.2.0.8). Datasets from 10X were down-sampled to 1200 cells per dataset. The order of dataset input is shown on the top of the Figures **(a, b, c** or **d, e, f)**.



Extended Data Figure 9. Correlations of gene expression profiles across datasets.

Scatter plots displaying the gene expression profile correlations between each of seven scRNA-seq datasets (10X_LLU, 10X_NCI, 10X_NCI_M, C1_FDA, C1_LLU, ICELL8_SE, and ICELL8_PE) vs. their corresponding bulk RNA-seq dataset (BK_RNA-seq) for either (**a**) breast cancer cells or (**b**) B lymphocytes. The commonly detected transcripts [(log(CPM +1) normalized] across all datasets were used (15,553 genes for breast cancer cells and 15,201 genes for B lymphocytes) to generate the scatter plots. Each dot represents each gene as a point in each scatterplot; x, y values represent the gene expression variation in a pair of compared datasets. The middle diagonal bar charts display the distribution of the most abundant or rare genes in each dataset and also provide the labels for the respective datasets. The Pearson correlation coefficient R between each of the datasets compared is shown to display the consistency of the different RNA-seq datasets.



Extended Data Figure 10. Scanorama batch correction using 10X and non-10X scRNA-seq datasets from two different studies.

(**a**, un-corrected) UMAP of 10 datasets (10X: PBMCs 68K, PBMCs 3K, CD19+ B cells, CD14+ monocytes, CD4+ helper T cells, CD56+ NK cells, CD8+ cytotoxic T cells, CD4+CD45RO+ memory T cells, CD4+CD25+ regulatory T cells; Drop-seq: PBMCs) out of 26 datasets from Hie et al.[8] before batch correction by Scanorama. (**b**, corrected-based on dataset) UMAP of 10 different datasets shown in (**a**) from Hie et al. after batch correction by Scanorama, colored to identify the datasets. (**c,** corrected-based on platform) UMAP of 10 different datasets shown in (**a**) from Hie et al. colored to identify the two different platforms used (10X Genomics and Drop-seq); note poor results using Drop-seq. (**d,** un-corrected) UMAP of 8 datasets (breast cancer cells: C1_FDA_HT_A, C1_LLU_A, ICELL8_SE_A, and ICELL8_PE_A; and B lymphocytes: C1_FDA_HT_B, C1_LLU_B, ICELL8_SE_B, and ICELL8_PE_B) out of 20 datasets in our study analyzed using three different non-10X sequencing platforms before batch correction by Scanorama. (**e,** corrected-based on dataset) UMAP of 8 datasets shown in (**d**) after batch correction by Scanorama, colored to identify the datasets. Note lack of discrimination between different cell types. (**f,** corrected-based on platform) UMAP of 8 datasets shown in (**d**) after batch correction by Scanorama, colored to identify the platforms (C1_FDA_HT, blue; C1, purple; ICELL8, pink). The PBMC datasets were downloaded from http://scanorama.csail.mit.edu/data_light.tar.gz. Our eight datasets were preprocessed using the featureCounts pipeline and batch-effect correction was performed using Scanorama V1.4.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The authors would like to thank Ms. Diana Ho of the LLU Center for Genomics for her great administrative support, particularly in coordinating the weekly Zoom conference calls and assistance in preparation of meeting minutes for the SEQC-2 single-cell sequencing project. The authors would like to thank ATCC, and particularly Liz Kerrigan for providing the two cell lines, i.e., HCC1395 and HCC1395BL for our study. The authors would like to thank Dr. Wendell Jones at Q^2 Solutions | EA Genomics for critical review and helpful comments.

The authors would like to thank Dr. Zhong Chen at LLU and Jyoti Shetty at NCI for technical assistance in performing sequencing; John Bettridge at NCI for technical assistance in 10X Genomics scRNA-seq library preparation; Vyacheslav Furtak at FDA for library preparation; Wells Wu at the FDA/CBER Core Facility for Illumina sequencing. The authors also would like to thank Sangeetha Anandakrishnan of Takara Bio USA, Inc. for technical assistance with TBU ICELL8 single cell capture and library preparation. The genomic work carried out at the LLU Center for Genomics was funded in part by the National Institutes of Health (NIH) grant S10OD019960 (CW), the Ardmore Institute of Health grant 2150141 (CW) and Dr. Charles A. Sims' gift to LLU Center for Genomics.

Data availability statement

The datasets generated and analyzed in the current study are available in the SRA repository with the access code # (PRJNA504037), and the following link:

https://www.ncbi.nlm.nih.gov/bioproject/PRJNA504037

The data from Tian et al. are available at the following URL:

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE118767

The data from Hie et al. are available at the following URL:

http://scanorama.csail.mit.edu/data.tar.gz

References

1. Klein AM et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell 161, 1187–1201 (2015). [PubMed: 26000487]
2. Macosko EZ et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell 161, 1202–1214 (2015). [PubMed: 26000488]
3. Gierahn TM et al. Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput. Nat Methods 14, 395–398 (2017). [PubMed: 28192419]
4. Liu T, Wu H, Wu S & Wang C Single-Cell Sequencing Technologies for Cardiac Stem Cell Studies. Stem Cells Dev 26, 1540–1551 (2017). [PubMed: 28859577]
5. Wu H, Wang C & Wu S Single-Cell Sequencing for Drug Discovery and Drug Development. Curr Top Med Chem 17, 1769–1777 (2017). [PubMed: 27848892]
6. Haghverdi L, Lun ATL, Morgan MD & Marioni JC Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. Nat Biotechnol 36, 421–427 (2018). [PubMed: 29608177]
7. Butler A, Hoffman P, Smibert P, Papalexi E & Satija R Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol 36, 411–420 (2018). [PubMed: 29608179]
8. Hie B, Bryson B & Berger B Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. Nat Biotechnol 37, 685–691 (2019). [PubMed: 31061482]
9. Polanski K et al. BBKNN: Fast Batch Alignment of Single Cell Transcriptomes. Bioinformatics (2019).
10. Korsunsky I et al. Fast, sensitive and accurate integration of single-cell data with Harmony. Nat Methods 16, 1289–1296 (2019). [PubMed: 31740819]
11. Saelens W, Cannoodt R, Todorov H & Saeys Y A comparison of single-cell trajectory inference methods. Nat Biotechnol 37, 547–554 (2019). [PubMed: 30936559]
12. Ziegenhain C et al. Comparative Analysis of Single-Cell RNA Sequencing Methods. Mol Cell 65, 631–643 e634 (2017). [PubMed: 28212749]
13. Zhang X et al. Comparative Analysis of Droplet-Based Ultra-High-Throughput Single-Cell RNA-Seq Systems. Mol Cell 73, 130–142 e135 (2019). [PubMed: 30472192]

14. Svensson V et al. Power analysis of single-cell RNA-sequencing experiments. Nature Methods 14, 381-+ (2017). [PubMed: 28263961]
15. Mereu E et al. Benchmarking single-cell RNA-sequencing protocols for cell atlas projects. Nature Biotechnology (2020).
16. Tian L et al. Benchmarking single cell RNA-sequencing analysis pipelines using mixture control experiments. Nat Methods 16, 479–487 (2019). [PubMed: 31133762]
17. Tran HTN et al. A benchmark of batch-effect correction methods for single-cell RNA sequencing data. Genome Biol 21, 12 (2020). [PubMed: 31948481]
18. Gazdar AF et al. Characterization of paired tumor and non-tumor cell lines established from patients with breast cancer. Int J Cancer 78, 766–774 (1998). [PubMed: 9833771]
19. Xiao W Towards best practice in cancer mutation detection with whole-genome and whole-exome esequencing. Nature Biotechnology XX,XXX(2020).
20. Zhang J, Spath SS, Marjani SL, Zhang W & Pan X Characterization of cancer genomic heterogeneity by next-generation sequencing advances precision medicine in cancer treatment. Precis Clin Med 1, 29–48 (2018). [PubMed: 30687561]
21. Chen X et al. A Multi-center Cross-platform Single-cell RNA Sequencing Reference Dataset. bioRxiv, 2020.2009.2020.305474 (2020).
22. https://kb.10xgenomics.com/hc/en-us/articles/115005062366-What-issequencing-saturation.
23. https://support.10xgenomics.com/single-cell-gene-expression/software/.
24. Smith T, Heger A & Sudbery I UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. Genome Res 27, 491–499 (2017). [PubMed: 28100584]
25. Parekh S, Ziegenhain C, Vieth B, Enard W & Hellmann I zUMIs - A fast and flexible pipeline to process RNA sequencing data with UMIs. Gigascience 7 (2018).
26. Liao Y, Smyth GK & Shi W featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 30, 923–930 (2014). [PubMed: 24227677]
27. Bray NL, Pimentel H, Melsted P & Pachter L Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol 34, 525–527 (2016). [PubMed: 27043002]
28. Li B & Dewey CN RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 12, 323 (2011). [PubMed: 21816040]
29. Martin M Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet. journal 17, 10–12 (2011).
30. Bolger AM, Lohse M & Usadel BJB Trimmomatic: a flexible trimmer for Illumina sequence data. 30, 2114–2120 (2014).
31. Dobin A et al. STAR: ultrafast universal RNA-seq aligner. 29, 15–21 (2013).
32. Hicks SC, Townes FW, Teng M & Irizarry RA Missing data and technical variability in single-cell RNA-sequencing experiments. Biostatistics (2017).
33. Risso D, Ngai J, Speed TP & Dudoit S Normalization of RNA-seq data using factor analysis of control genes or samples. Nat Biotechnol 32, 896–902 (2014). [PubMed: 25150836]
34. Hafemeister C & Satija R Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. Genome Biol 20, 296 (2019). [PubMed: 31870423]
35. Lun AT, Bach K & Marioni JC Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. Genome Biol 17, 75 (2016). [PubMed: 27122128]
36. Bacher R et al. SCnorm: robust normalization of single-cell RNA-seq data. Nat Methods 14, 584–586 (2017). [PubMed: 28418000]
37. Yip SH, Wang P, Kocher J-PA, Sham PC & Wang J Linnorm: improved statistical analysis for single cell RNA-seq expression data. Nucleic acids research 45, e179–e179 (2017). [PubMed: 28981748]
38. Stuart T et al. Comprehensive Integration of Single-Cell Data. Cell 177, 1888–1902 e1821 (2019). [PubMed: 31178118]
39. Yip SH, Sham PC & Wang J Evaluation of tools for highly variable gene discovery from single-cell RNA-seq data. Brief Bioinform, bby011 (2018).

40. Buettner F et al. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. Nat Biotechnol 33, 155–160 (2015). [PubMed: 25599176]
41. Kang HM et al. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. Nature Biotechnology 36, 89-+ (2018).
42. Ritchie ME et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 43, e47 (2015). [PubMed: 25605792]
43. Leek JT, Johnson WE, Parker HS, Jaffe AE & Storey JD The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics 28, 882–883 (2012). [PubMed: 22257669]
44. Becht E et al. Dimensionality reduction for visualizing single-cell data using UMAP. Nat Biotechnol (2018).
45. Buttner M, Miao Z, Wolf FA, Teichmann SA & Theis FJ A test metric for assessing single-cell RNA-seq batch correction. Nat Methods 16, 43–49 (2019). [PubMed: 30573817]
46. Kaminski DA, Wei C, Qian Y, Rosenberg AF & Sanz I Advances in human B cell phenotypic profiling. Front Immunol 3, 302 (2012). [PubMed: 23087687]
47. Starlets D et al. Cell-surface CD74 initiates a signaling cascade leading to cell proliferation and survival. Blood 107, 4807–4816 (2006). [PubMed: 16484589]
48. Zook JM et al. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. Nat Biotechnol 32, 246–251 (2014). [PubMed: 24531798]
49. Alles J et al. Cell fixation and preservation for droplet-based single-cell transcriptomics. BMC Biol 15, 44 (2017). [PubMed: 28526029]
50. Li H et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079 (2009). [PubMed: 19505943]
51. Krueger F & Galore T (2015).
52. Cole MB et al. Performance Assessment and Selection of Normalization Procedures for Single-Cell RNA-Seq. Cell Syst 8, 315–328 e318 (2019). [PubMed: 31022373]
53. Qiu X et al. Single-cell mRNA quantification and differential analysis with Census. 14, 309 (2017).
54. Trapnell C et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat Biotechnol 32, 381–386 (2014). [PubMed: 24658644]

Box 1.

Best practice recommendations

We summarize below 11 best practice recommendations for the community based on our analysis:

1. There were large variations across different scRNA-seq platforms and centers.
2. While most of the genes and cells detected were consistent between the different methods, we observed variations for low expression genes and cells with low mRNA content across different methods. However, these differences did not affect our analyses of cell classification or mixability.
3. Normalization algorithms alone could not remove the batch effects.
4. Different normalization strategies performed differently across datasets and platforms; most performed well for either 3'- or full-length-transcript scRNA-seq platforms such as SCTransform, scran, logCPM, and Linnorm, but TMM and Quantile performed poorly, and are not recommended.
5. Seurat v3, Harmony, BBKNN, fastMNN, and Scanorama all could correct and remove the batch variations in specific sample and dataset scenarios; we recommend users apply appropriate batch-effect correction methods depending on the characteristics of their datasets (e.g., cellular/sample heterogeneity and composition, platforms used, see Fig. 6e).
6. BBKNN, fastMNN, and Harmony ranked best for clusterability/cell type classification, whereas Seurat v3, Harmony, and fastMNN performed best for mixability.
7. fastMNN, BBKNN, and Harmony removed batch variations well across different platforms, including both mixed and non-mixed distinct samples, but the order of importing the datasets into the pipeline and the requirement for a mixed sample was critical for MNN/fastMNN; whereas BBKNN and Harmony worked well regardless of the inclusion of mixed heterogeneous biologically distinct samples across platforms and batches; thus, for MNN/fastMMN, we recommend including a mixed sample and importing the mixed data into the pipeline first.
8. CCA/Seurat v3, had superior mixability for biologically similar samples, but over-corrected batch effects and misclassified cells (i.e., poor clusterability/cell type classification) if large proportions of distinct cell types were present. However, Seurat v3 worked well both for clusterability and mixability for datasets when only a small fraction of dissimilar cells (e.g., 5–10%) was present. Thus, we do not recommend using CCA/Seurat v3 for scenarios containing large fractions of biologically distinct cell types.
9. BBKNN performed best in clusterability/cell type classification, but it ranked low in mixability, particularly in heterogeneous cellular samples.

10. The current version of Scanorama worked well for the 10X Genomics data only, but did not work for non-10X platforms, thus we do not recommend it for non-10X data.

11. We observed good consistency between CellRanger 3.1 and 2.0 pre-processed data; however, Cell Ranger 3.1 can detect some extra cells with very few transcripts; this may affect batch-effect corrections in certain scenarios.

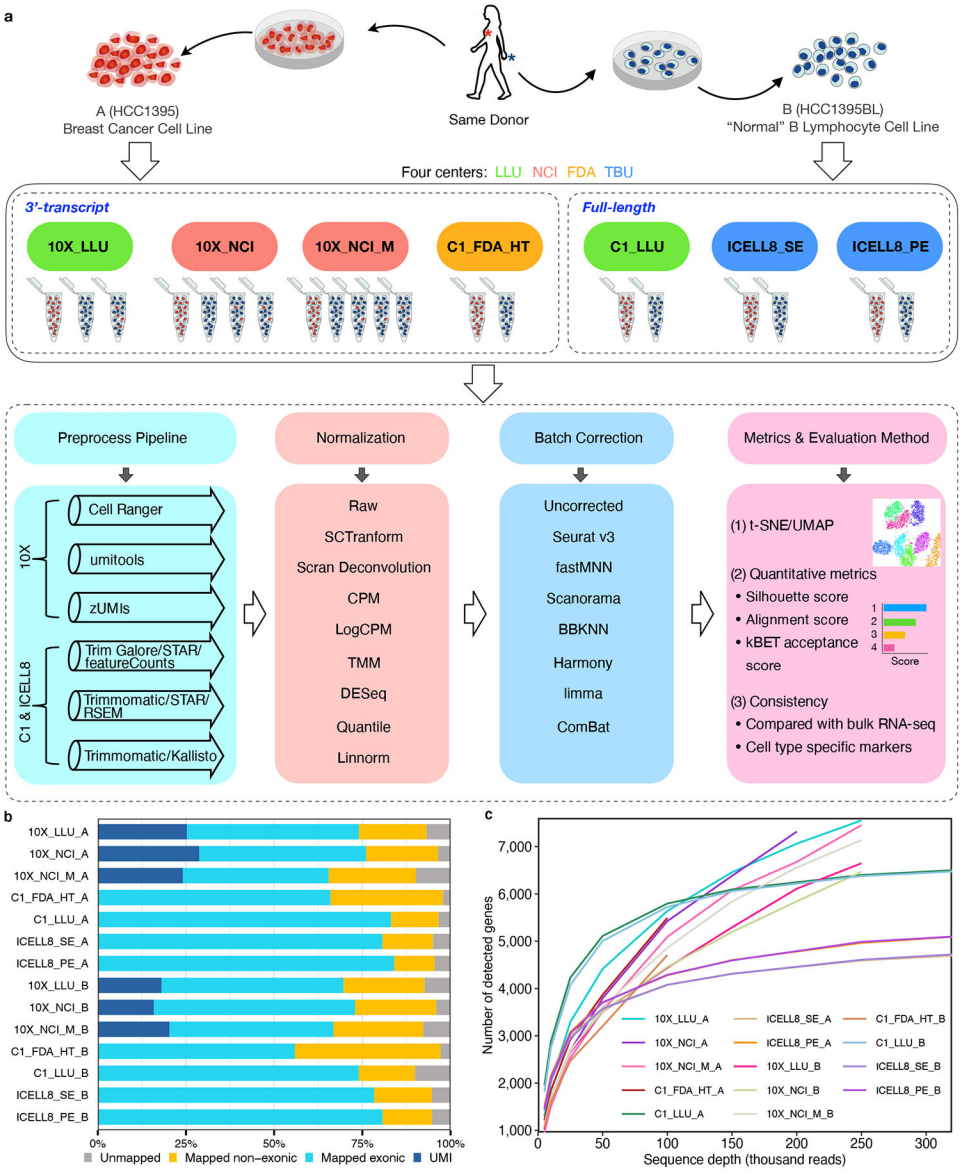


Figure 1. Overall study design, scRNA-seq mapping, and numbers of genes detected across datasets.

(a) Schematic overview of the study design (see detailed descriptions and notations in the Methods). Two reference cell lines (Sample A, HCC1395; and Sample B, HCC1395BL) were used to generate scRNA-seq data across four platforms (10X Genomics, Fluidigm C1, Fluidigm C1 HT, and Takara Bio ICELL8), four testing sites (LLU, NCI, FDA, and TBU). At the LLU and NCI sites (10X), mixed single-cell captures and library constructions were also prepared with either 10% or 5% cancer cells spiked into the B lymphocytes. At the NCI site, single-cell captures and library constructions were also performed with methanol-fixed cell mixtures (5% cancer cells spiked into B lymphocytes, Fixed 1 & 2). One set of 10X scRNA libraries from NCI was also sequenced using a shorter modified sequencing method. Bulk cell RNA-seq was also obtained from these cell lines, each in triplicate. See Methods for details about study design. **(b)** For both the breast cancer cell line (Sample A) and the

B lymphocyte line (Sample B) across 14 pair-wise datasets, percentage of reads mapped to the exonic region (blue), non-exonic region (orange), or not mapped to the human genome (gray). For unique molecular identifier (UMI) methods (10X), dark blue indicates the exonic reads with UMIs. **(c)** Median number of genes detected per cell at different sequencing read depths.

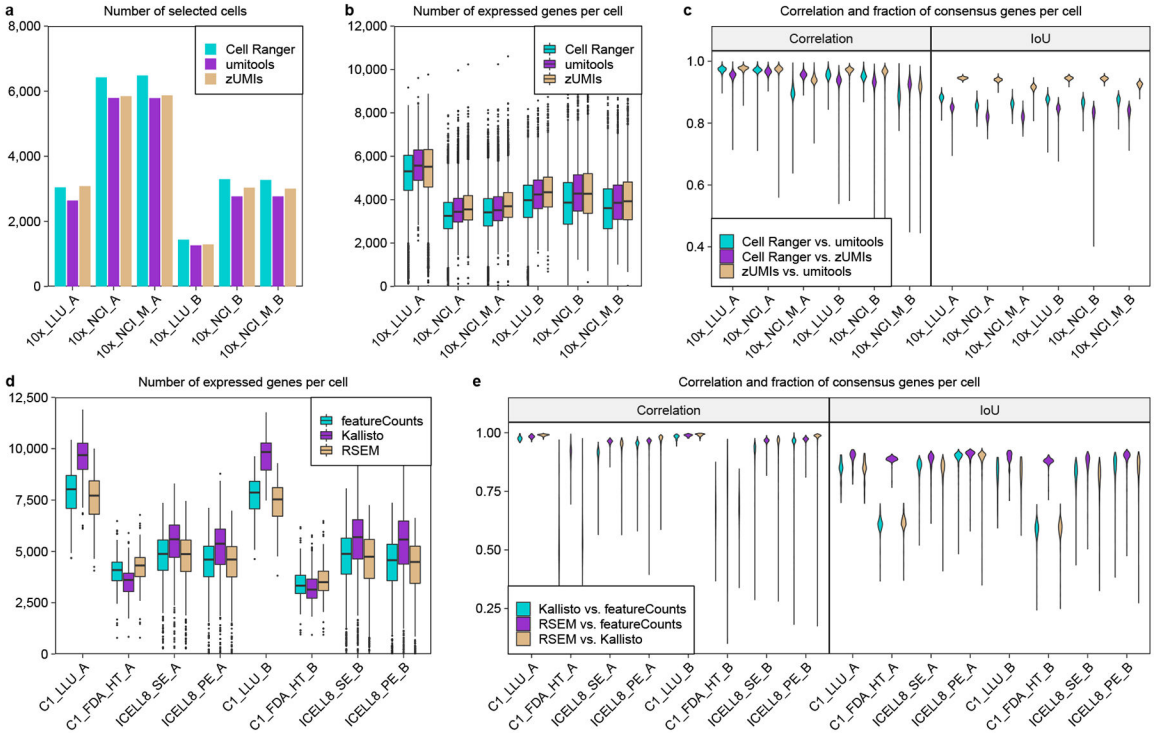


Figure 2. Effect of pre-processing pipeline on the number of genes detected with UMI- and non-UMI-based scRNA-seq datasets.

(a–c) Evaluation of the UMI-based (10X) data with Cell Ranger, UMI-Tools, or zUMIs. **(d–e)** Evaluation of data from non-UMI based technologies C1 full-length transcript, C1 HT, and ICELL8 full-length transcript using FeatureCounts, Kallisto, or RSEM. (**a**) Bar plot showing the number of cells captured with UMI-based technology; (**b**) and (**d**) Box plot showing the number of genes detected per cell in UMI-based and non-UMI based technologies, respectively; (**c**) and (**e**) Violin plots showing the gene expression correlation and consensus genes [represented by IoU (Intersection over Union)] per cell between any two pipelines in UMI-based and non-UMI based technologies, respectively. The sample sizes (n) used to derive statistics in (**b**) and (**d**) were: (**b**) 10X_LLU_A, n= 3045 cells; 10X_NCI_A, n=6425 cells; 10X_NCI_M_A, n=6483 cells; 10X_LLU_B, n=1439 cells; 10X_NCI_B, n=3296 cells; 10X_NCI_M_B, n=3273 cells; (**d**) C1_LLU_A, n=80 cells; C1_FDA_HT_A, n=203 cells; ICELL8_SE_A, n=600 cells; ICELL8_PE_A, n=598 cells; C1_LLU_B, n=66 cells; C1_FDA_B, n=241 cells; ICELL8_SE_B, n=600 cells; ICELL8_PE_B, n=596 cells. For detailed statistics regarding minima, maxima, centre, bounds of box and whiskers and percentile related to the figure, please refer to Supplementary Table 5.

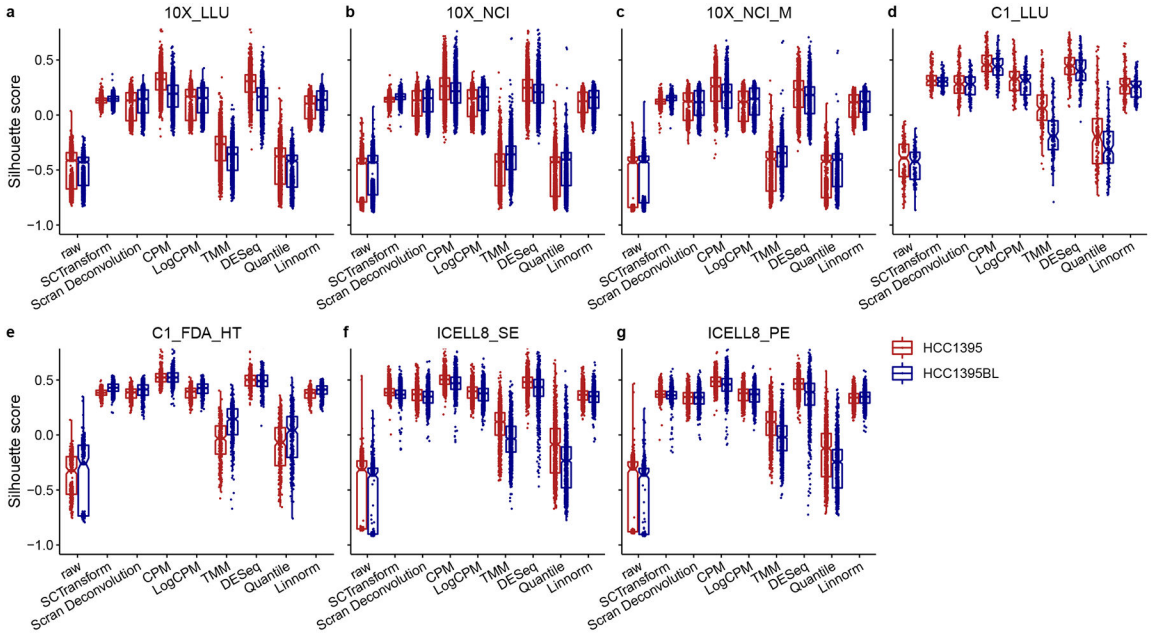


Figure 3. Silhouette score boxplot comparing eight normalization methods.

Boxplot of silhouette values stratified by eight normalization methods across 14 datasets, including (**a**) 10X_LLU, (**b**) 10X_NCI, (**c**) 10X_NCI_M, (**d**) C1_FDA_HT, (**e**) C1_LLU, (**f**) ICELL8_PE, and (**g**) ICELL8_SE in breast cancer cells (HCC1395; Sample A) and B lymphocytes (HCC1395BL; Sample B). Eight normalization methods included SCTransform, Scran Deconvolution, CPM, LogCPM, TMM, DESeq, Quantile, and Linnorm. For each dataset, reads of each cell were down-sampled to two different read depths (10K and 100K per cell) before calculating the silhouette width values. LogCPM normalization performed fairly well and was used as the default normalization for our subsequent batch-effect correction analyses. Two normalization methods developed for bulk cell RNA-seq (TMM and Quantile) had the lowest scores. The sample sizes (n) used to derive statistics were: 10X_LLU_A, n= 3560 cells, 10X_LLU_B, n=1770 cells; 10X_NCI_A, n=4284 cells, 10X_NCI_B, n=4136 cells; 10X_NCI_M_A, n=1372 cells, 10X_NCI_M_B, n=2082 cells; C1_LLU_A, n=160 cells, C1_LLU_B, n=132 cells; C1_FDA_HT_A, n=318 cells, C1_FDA_HT_B, n=374 cells; ICELL8_SE_A, n=1134 cells, ICELL8_SE_B, n=1078 cells; ICELL8_PE_A, n=980 cells, ICELL8_PE_B, n=954 cells). For detailed statistics regarding minima, maxima, centre, bounds of box and whiskers and percentile related to the figure, please refer to Supplementary Table 6.

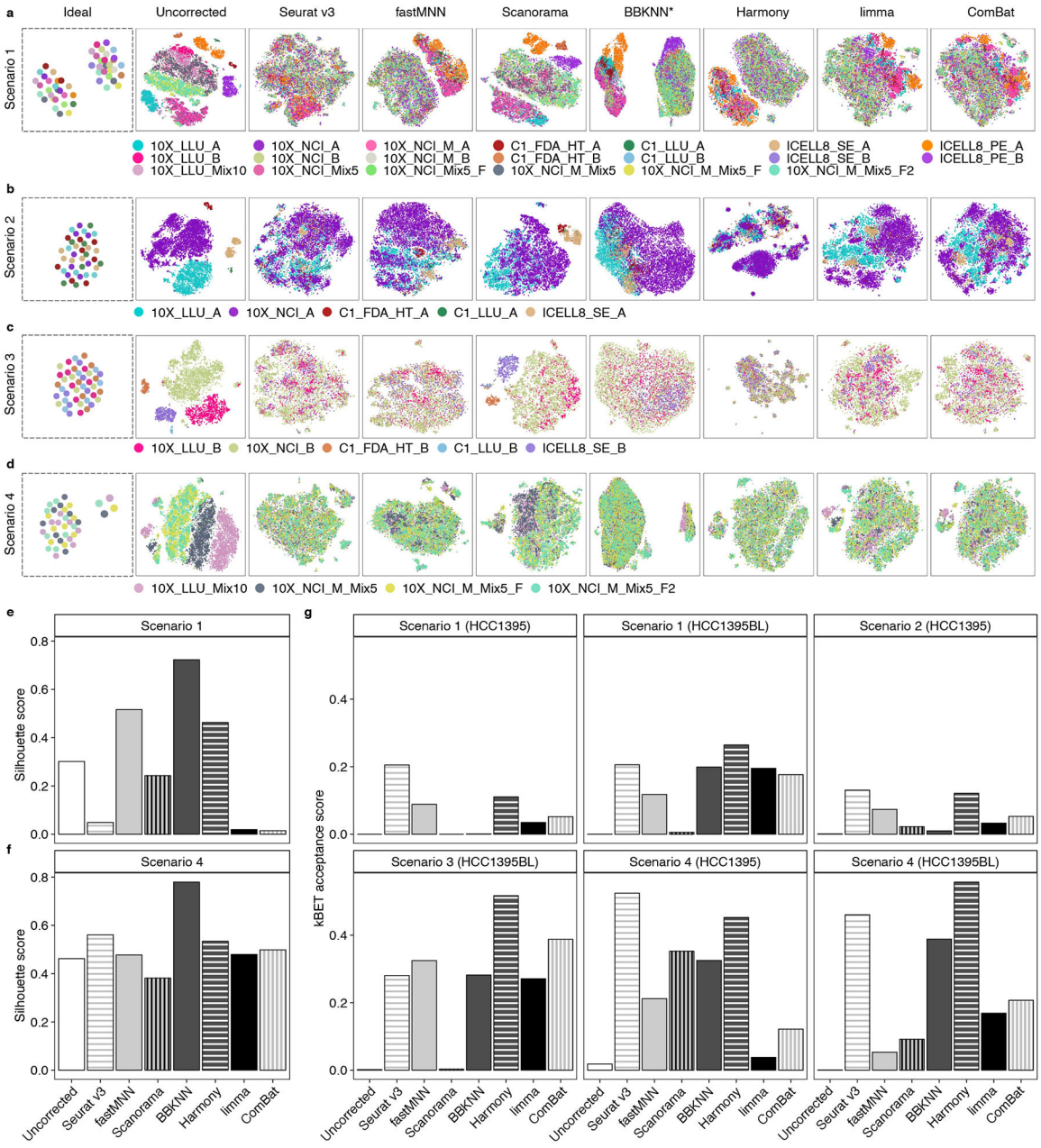


Figure 4. Batch-effect corrections evaluated in four different sample composition scenarios.

(a) Batch-effect correction in Scenario #1, where all 20 scRNA-seq datasets were combined, including mixed and non-mixed, with large proportions of two dissimilar types of cells (Sample A, breast cancer cell line HCC1395; and Sample B, B-lymphocyte line HCC1395BL). Datasets from 10X were down-sampled to 1200 cells per dataset. (**b**) Batch-effect correction in Scenario #2, where five scRNA-seq datasets (10X_LLU_A, 10X_NCI_A, C1_FDA_HT_A, C1_LLU_A, and ICELL8_SE_A) from the breast cancer cells were generated separately at the four centers (LLU, NCI, FDA, and TBU) on four platforms (10X, Fluidigm C1, Fluidigm C1_HT, and TBU ICELL8); (**c**) Batch-effect correction in Scenario #3, where five scRNA-seq datasets (10X_LLU_B,

10X_NCI_B, C1_FDA_HT_B, C1_LLU_B, and ICELL8_SE_B) from the B lymphocytes were generated separately at the four centers on the same four platforms; (**d**) Batch-effect correction in Scenario #4, where four datasets (10X_LLU_Mix10, 10X_NCI_M_Mix5, 10X_NCI_M_Mix5_F, 10X_NCI_M_Mix5_F2) were generated from 5% or 10% breast cancer cells spiked into B lymphocytes, and analyzed with the 10X Genomics platform at two centers in four different batches. Each dataset is indicated by a unique color in panels (**a**) to (**d**). Idealized projection of cells for the four different scenarios is presented on the left. *****Note for BBKNN, only UMAP is available and shown. Silhouette width score quantifying the clusterability for (**e**) Scenario #1 or (**f**) Scenario #4, corresponding to panels (**a**) and (**d**), respectively. (**g**) kBET acceptance score quantifying the mixability, calculated using the cross-platform/center scRNA-seq data acquired either from breast cancer cells only or from B-lymphocytes only for all four scenarios (**a-d**, also labeled as Scenarios #1–#4).

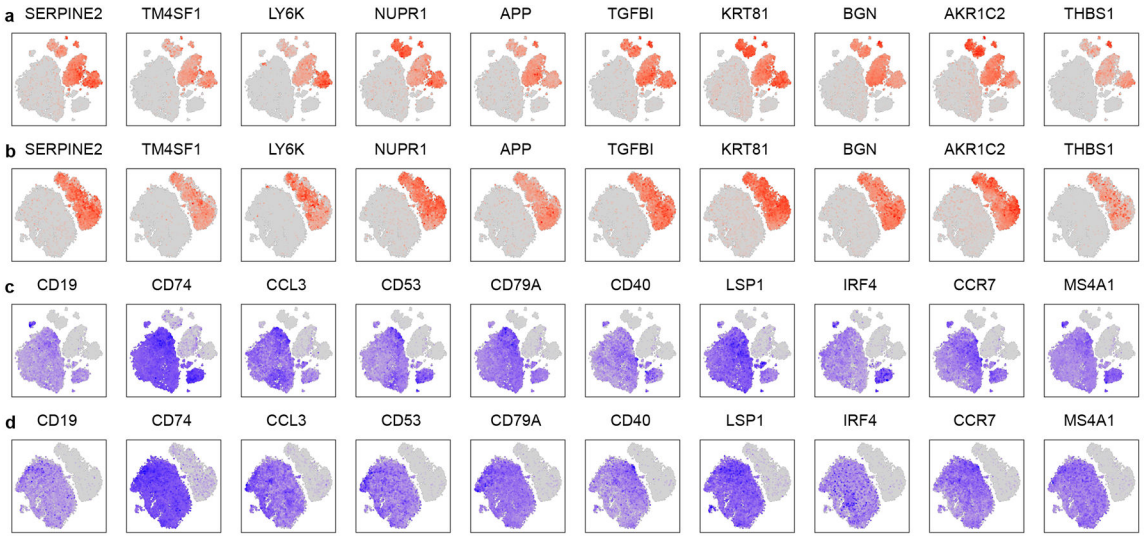


Figure 5. Feature plots showing cell type clustering based on cell type-specific marker genes across 20 scRNA-seq datasets.

Feature plots generated across 20 scRNA-seq datasets using the top 10 DEGs specific for (**a**) breast cancer cells before batch-effect correction; (**b**) breast cancer cells after fastMNN batch-effect correction; (**c**) B lymphocytes before batch correction; and (**d**) B lymphocytes after fastMNN batch-effect correction. Datasets from 10X were down-sampled to 1200 cells per dataset. In feature plots, genes with relatively high expression in each cell are highlighted in brick red (corresponding to breast cancer cells; Sample A) or blue (corresponding to B cells; Sample B).

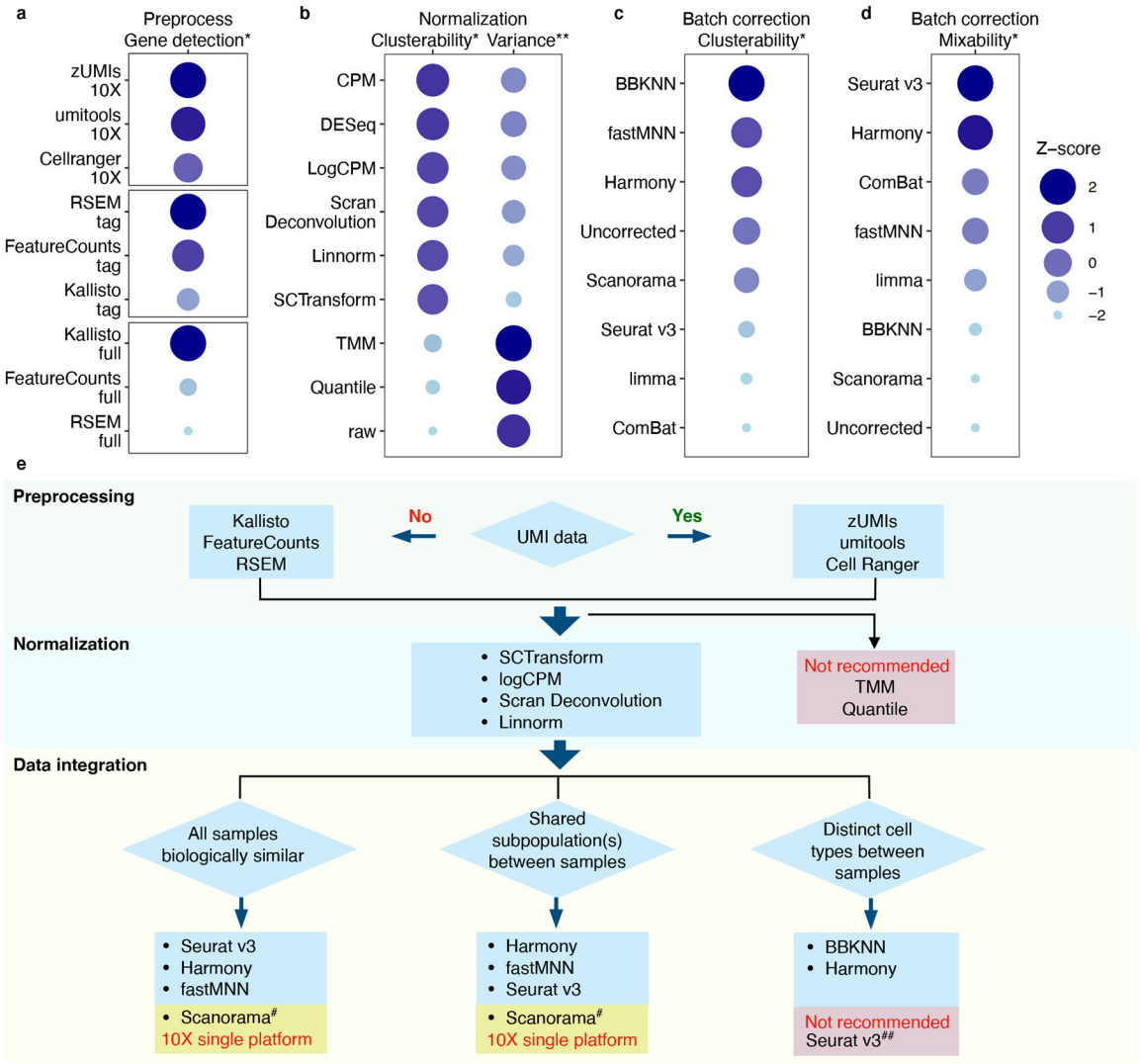


Figure 6. Performance ranking of bioinformatics metrics and best-practice recommendations.

(a) Gene detection sensitivity measured separately for each of the three classes of scRNA-seq protocol: 10X-, non-10X-based 3′ tagging, and full-length. **(b)** Normalization methods ranked by their clusterability as measured by Z-scores (either the median or the variance of the silhouette width across the 14 datasets). **(c)** Batch-correction methods ranked by their clusterability as measured by Z-score from the harmonic mean of the silhouette scores (Scenarios #1 and #4). **(d)** Batch-correction methods ranked by their mixability as measured by Z-score from the harmonic mean of kBET acceptance scores (Scenarios #1–#4). Z-scores are plotted as circles with their size and color shade scaled to the Z-score value from large to small, and dark blue to light blue. Note that larger Z-score values imply better performance, except for clusterability variance, where a smaller value is preferred: *Larger is better; **Smaller is better. (**e**) Best practice recommendations for single-cell RNA-seq analysis. #The current version of Scanorama did not correct batch effects for data from multiple platforms; however, it worked well when only 10X Genomics data were analyzed. ##Seurat v.3 was suitable for biologically similar samples, but over-corrected batch effects

and misclassified cell types if large fractions of distinct cell types were present in different batches.