



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

Wanqiu Chen^{1,12}, Yongmei Zhao^{2,3,12}, Xin Chen^{1,4,12}, Zhaowei Yang^{1,5,12}, Xiaojiang Xu⁶, Yingtao Bi⁷, Vicky Chen^{2,3}, Jing Li^{4,5}, Hannah Choi¹, Ben Ernest⁸, Bao Tran³, Monika Mehta^{1,3}, Parimal Kumar³, Andrew Farmer⁹, Alain Mir⁹, Urvashi Ann Mehra⁸, Jian-Liang Li^{1,6}, Malcolm Moos Jr.¹⁰, Wenming Xiao¹¹✉ and Charles Wang^{1,4}✉

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 preprocessing, normalization and batch-effect correction methods at multiple centers. Although preprocessing 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 (for example, sample and 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 an scRNA-seq study.

cRNA-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 bioinformatic methods, each with its own set of capabilities, limitations and costs^{1–11}.

Various aspects of this problem were recently examined^{12–14}. Investigators from the Human Cell Atlas consortium performed a comprehensive multicenter study that compared 13 different scRNA-seq protocols using a reference sample containing cells from humans, mice and dogs¹⁵. Consistent with previous reports^{12–14}, this group not only observed striking differences among protocols for quantifying gene expression and identifying cell type markers but also found large cross-protocol differences in their capacity to be integrated into reference tissue atlases¹⁵. 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 method 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 (that is, 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 interlaboratory 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 the biological classification of the cells analyzed.

As part of the second phase of the Sequencing Quality Control (SEQC-2) consortium, we designed a comprehensive multicenter study to evaluate the influence of technology platforms, sample composition and bioinformatic methods (including preprocessing, normalization and batch-effect correction). As samples, we used two well-characterized, biologically distinct, commercially available

¹Center for Genomics, School of Medicine, Loma Linda University, Loma Linda, CA, USA. ²CCR-SF Bioinformatics Group, Advanced Biomedical and Computational Sciences, Biomedical Informatics and Data Science Directorate, Frederick National Laboratory for Cancer Research, Frederick, MD, USA.

³Sequencing Facility, Frederick National Laboratory for Cancer Research, Frederick, MD, USA. ⁴Department of Basic Sciences, School of Medicine, Loma Linda University, Loma Linda, CA, USA. ⁵Department 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, People's Republic of China. ⁶Integrative Bioinformatics Support Group, National Institute of Environment Health Sciences, Research Triangle Park, NC, USA. ⁷AbbVie Cambridge Research Center, Cambridge, MA, USA. ⁸Digicon Corporation, McLean, VA, USA. ⁹Takara Bio USA, Inc., Mountain View, CA, USA. ¹⁰Center for Biologics Evaluation and Research & Division of Cellular and Gene Therapies, U.S. Food and Drug Administration, Silver Spring, MD, USA. ¹¹The Center for Devices and Radiological Health, U.S. Food and Drug Administration, Silver Spring, MD, USA. ¹²These authors contributed equally: Wanqiu Chen, Yongmei Zhao, Xin Chen, Zhaowei Yang. ✉e-mail: wenming.xiao@fda.hhs.gov; oxwang@gmail.com

reference cell lines¹⁸ that have whole-genome and whole-exome sequences (WGS and WES) available, obtained from multiple technology platforms¹⁹. 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 therapies²⁰. 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), the National Cancer Institute (NCI), the 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, cross-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 bioinformatic pipelines for the single-cell sequencing community and for integrating diverse datasets contributed to large collaborative projects, such as the Human Cell Atlas. 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, with two well-characterized reference cell lines¹⁸, 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 (PE; 75 × 2 bp, ICELL8_PE) with single-end (SE; 150 × 1 bp, ICELL8_SE) analysis. We applied three different preprocessing 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, counts per million (CPM), logCPM, trimmed mean of *M* values (TMM), DESeq, quantile and Linnorm) and seven different batch-effect correction algorithms (Seurat version 3, fastMNN, Scanorama, batch-balanced *k*-nearest neighbors (BBKNN), Harmony, limma, and ComBat), as well as the consistency of scRNA-seq data across sites and platforms with bulk cell RNA-seq (BK RNA-seq) on the two cell lines (each in triplicate, 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 Fig. 1 or reported in our companion paper²¹. 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 , *P* value ≤ 0.01 and false discovery rate = 0.05). The BK RNA-seq sequencing depth and mapping QC are shown in Supplementary Table 4 and Supplementary Fig. 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 downsampled 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 100,000 reads per cell for both cancer cells (sample A) and B lymphocytes (sample B), particularly for the Fluidigm C1 platform, before a read depth of 50,000. With increasing depth, gene counts ultimately plateaued, as expected. However, the rate of saturation was slower after 100,000 reads for 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 and Supplementary Fig. 3). One interpretation of this is that full-length libraries have higher complexity 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 3'-based technologies, few cells within the population obtain this read depth (Extended Data Fig. 1), so that on a population level, 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 do 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 was previously reported²².

Effects of data preprocessing. For the unique molecular identifier (UMI)-based scRNA-seq data, we compared three pipelines for preprocessing, Cell Ranger 3.1 (10x Genomics)²³, UMI-tools²⁴ and zUMIs²⁵, and examined the consistency between the three pipelines with regard to the number of barcoded cells identified and the number of genes detected per cell (Fig. 2a,b and Supplementary Tables 2 and 5). For the non-UMI-based scRNA-seq data, we examined three additional preprocessing pipelines, featureCounts²⁶, kallisto²⁷ and RSEM²⁸ (Fig. 2d and Supplementary Tables 3 and 5), which include trimming processes (cutadapt²⁹ or trimmomatic³⁰), alignment (STAR³¹ 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 in the number of genes detected per cell (Fig. 2a,b). Cell Ranger version 3 was the most sensitive method for cell barcode identification. UMI-tools and zUMIs filtered most low gene- or 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 preprocessing 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 preprocessing pipelines (Fig. 2d). We found that kallisto identified a substantially 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 remarkably 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 underperform when preprocessing 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 preprocessing pipelines, both for UMI- and non-UMI-based scRNA-seq datasets (Fig. 2c,e). To simplify our comparison, in all subsequent analyses we used Cell Ranger, one of

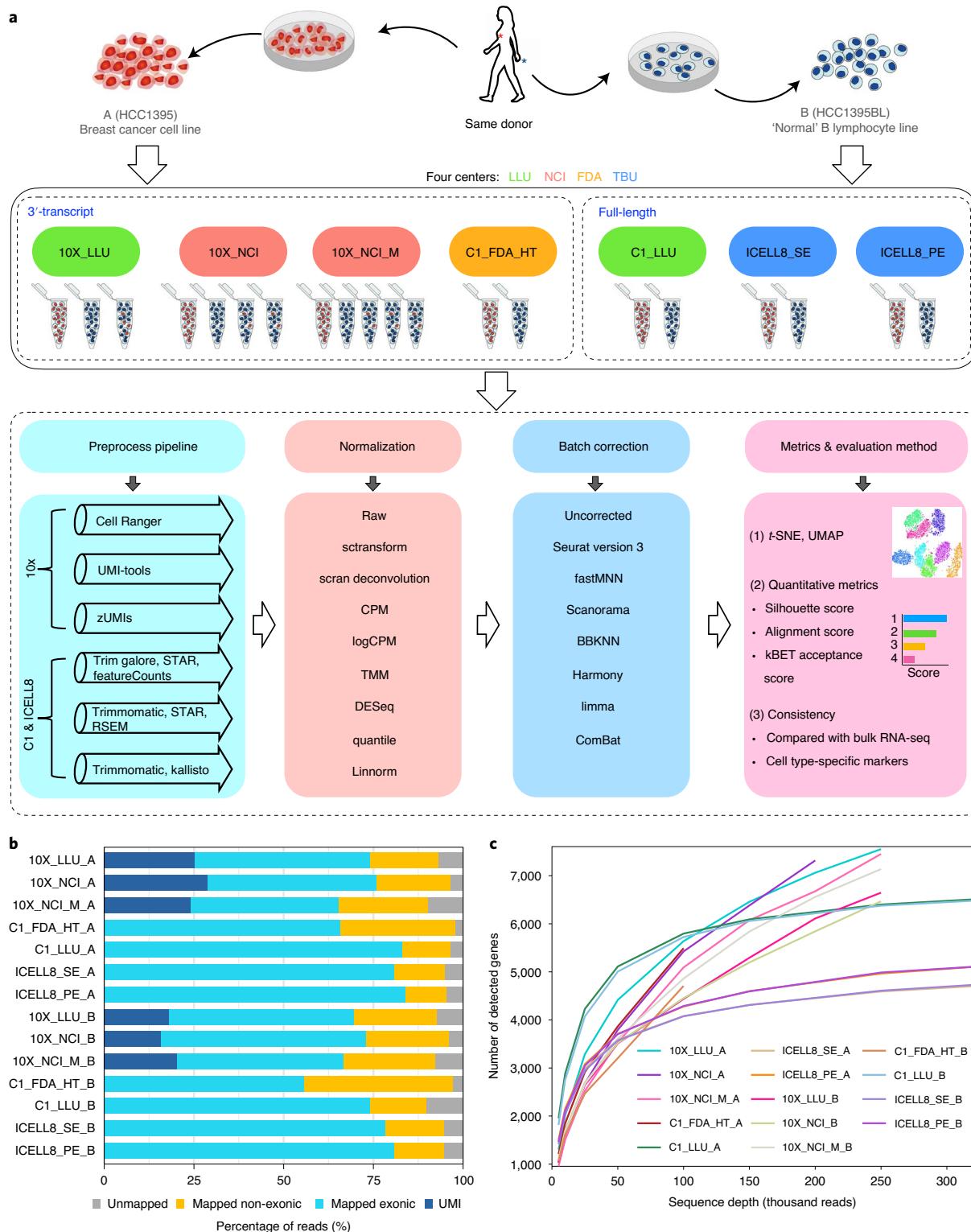


Fig. 1 | Overall study design, scRNA-seq mapping and numbers of genes detected across datasets. **a**, Schematic overview of the study design (see detailed descriptions and notations in the Methods). Two reference cell lines (sample A, HCC1395; sample B, HCC1395BL) were used to generate scRNA-seq data across four platforms (10x Genomics, Fluidigm C1 HT, Fluidigm C1 and Takara Bio ICELL8) and 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 two methanol-fixed cell mixtures (5% cancer cells spiked into B lymphocytes, termed fixed_1 and fixed_2). One set of 10x scRNA libraries from the NCI was also sequenced using a shorter modified sequencing method. BK RNA-seq data were 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 pairwise datasets, percentages are shown of reads that mapped to the exonic region (blue) or the non-exonic region (orange) or did not map to the human genome (gray). For UMI methods (10x), dark blue indicates the exonic reads with UMIs. **c**, Median number of genes detected per cell at different sequencing read depths.

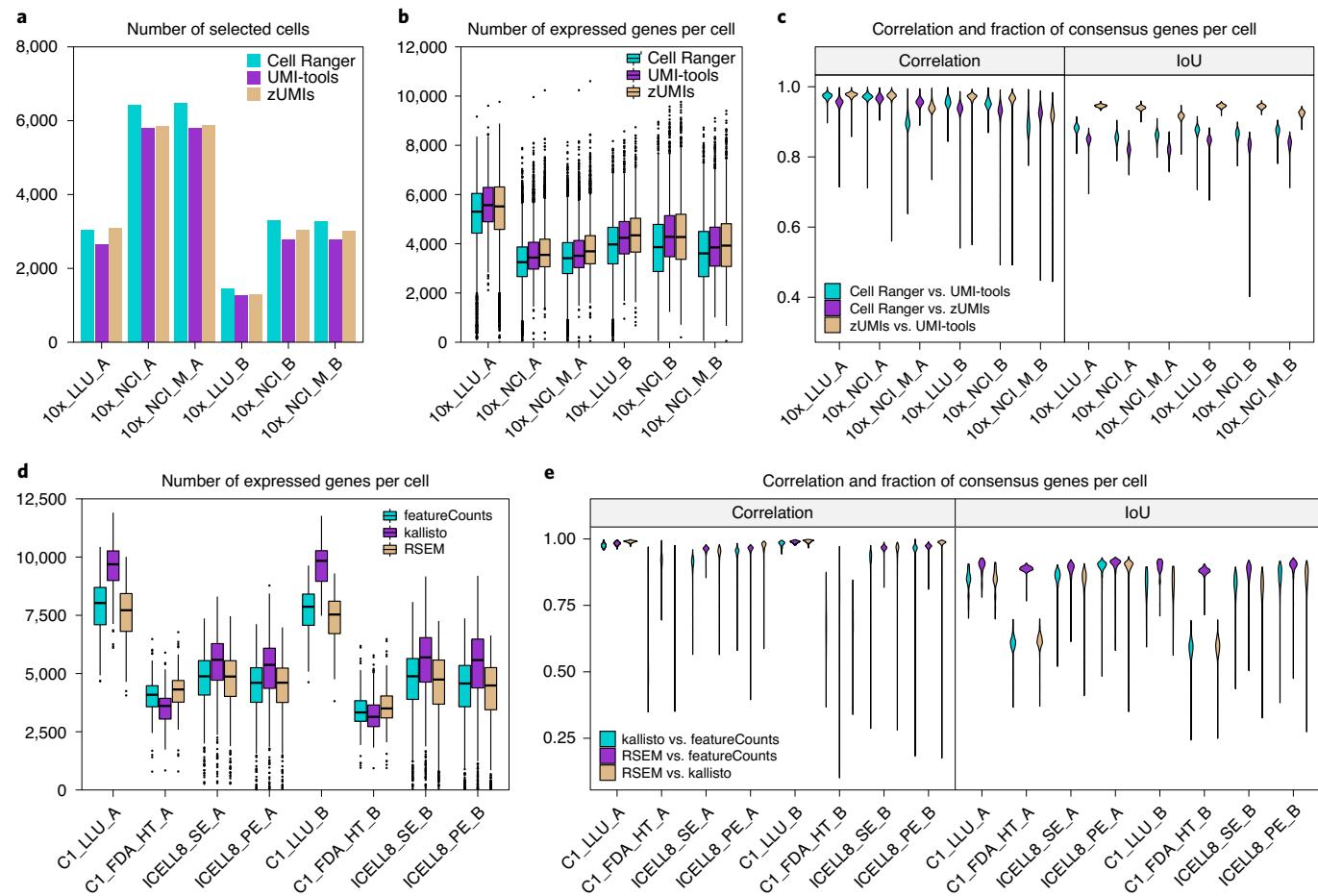


Fig. 2 | Effect of preprocessing pipeline on the number of genes detected with UMI- and non-UMI-based scRNA-seq datasets. **a–c**, Evaluation of the UMI-based (10x) data with Cell Ranger, UMI-tools or zUMIs. **d,e**, Evaluation of data from non-UMI-based technologies, C1, C1 HT or iCELL8 full-length transcript, using featureCounts, kallisto or RSEM. **a**, Bar plot showing the number of cells captured with UMI-based technology. **b,d**, Box plot showing the number of genes detected per cell with UMI-based and non-UMI-based technologies, respectively. **c,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 as follows: **b**, 10X_LLU_A, $n=3,045$ cells; 10X_NCI_A, $n=6,425$ cells; 10X_NCI_M_A, $n=6,483$ cells; 10X_LLU_B, $n=1,439$ cells; 10X_NCI_B, $n=3,296$ cells; 10X_NCI_M_B, $n=3,273$ 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. Box-and-whisker plots shows the first and third quartiles as well as the median value. The upper and lower whisker extends from the hinge to the largest or smallest value no further than $1.5 \times \text{IQR}$ from the hinge (where IQR is the inter-quartile range). For detailed statistics regarding minima, maxima, center, bounds of box and whiskers and percentiles related to this figure, please refer to Supplementary Table 5.

the most popular UMI-based methods for 3'-based count technologies, and featureCounts for non-UMI-based technologies, because it was more consistent with RSEM than with kallisto.

Effects of normalization. scRNA-seq data characteristically demonstrate substantial numbers of zero read counts³². This can be due to both biological (for example, bistable gene regulation) and technical reasons (for example, ‘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 BK RNA-seq data have been used fairly often for scRNA-seq data, including CPM, upper quantile (UQ), TMM and DESeq³³. However they were never systematically evaluated using a standard reference scRNA-seq dataset until the computational integration analysis of Tian et al., which involved two batches of mixed sample sets but no unmixed samples captured independently¹⁶. Regression-based methods were also proposed to remove known nuisance factors in scRNA-seq data.

There are also methods specifically tailored to scRNA-seq datasets, such as sctransform³⁴, scran³⁵, SCnorm³⁶ and Linnorm³⁷. sctransform was developed most recently and was integrated within Seurat version 3 (ref. 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 deconvolution³⁹, CPM, logCPM, TMM, DESeq, quantile and Linnorm, using the silhouette width metric, which evaluates how well two samples from the same cell type group with each other (Methods). We noticed that TMM and quantile failed to normalize either the breast cancer cells (sample A) or B lymphocytes (sample B), with silhouette scores that were similar to the unnormalized raw data (Fig. 3a–g). The other methods provided similar normalization, as measured by silhouette scores. sctransform seemed to perform slightly better than scran deconvolution, logCPM or Linnorm, in that it had the least variation among all normalization

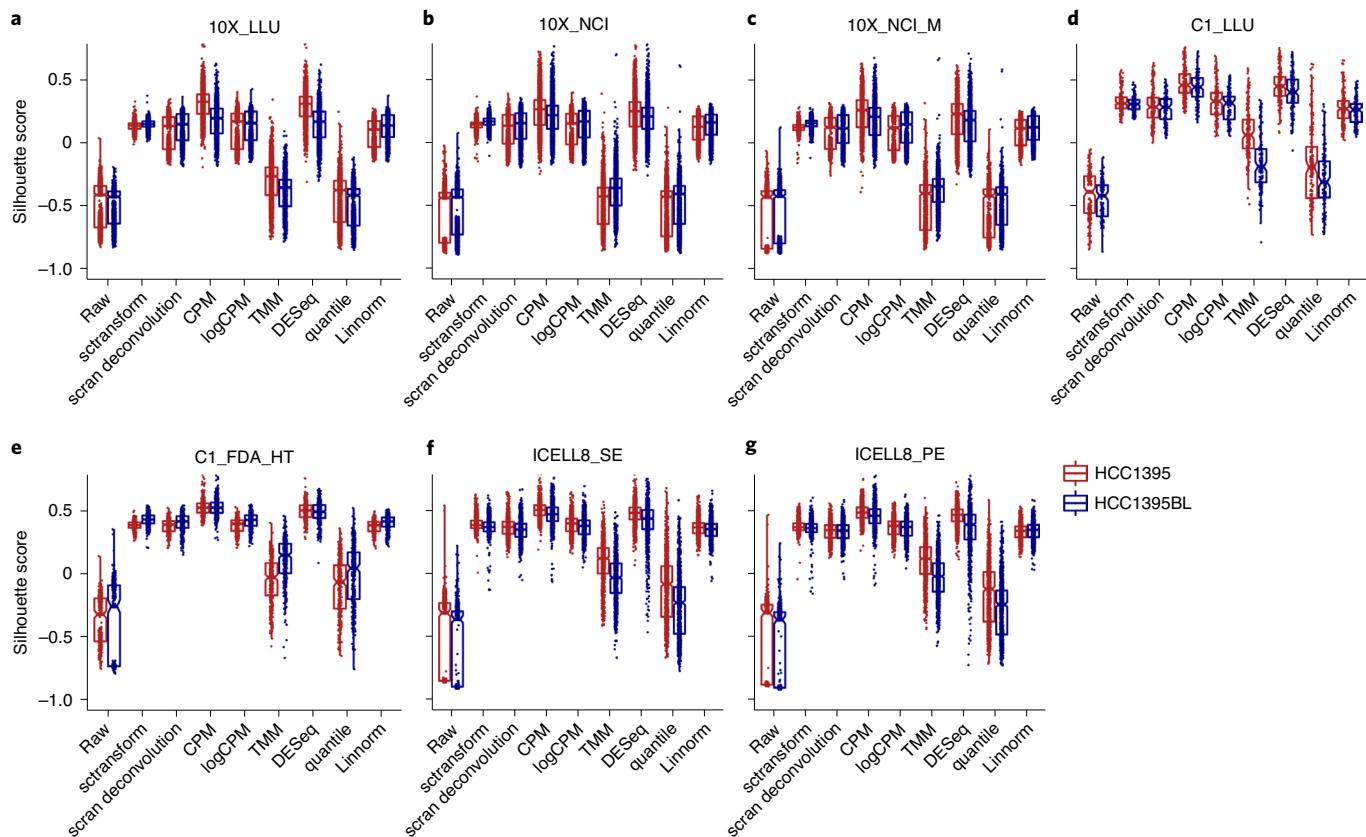


Fig. 3 | Silhouette score box plot comparing eight normalization methods. Box plot of silhouette values stratified by eight normalization methods across 14 datasets, including 10X_LLU (a), 10X_NCI (b), 10X_NCI_M (c), C1_LLU (d), C1_FDA_HT (e), ICELL8_SE (f) and ICELL8_PE (g) in breast cancer cells (HCC1395, 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 downsampled to two different read depths (10,000 and 100,000 reads 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 BK RNA-seq (TMM and quantile) had the lowest scores. The sample sizes (n) used to derive statistics were 10X_LLU_A, $n=3,560$ cells; 10X_LLU_B, $n=1,770$ cells; 10X_NCI_A, $n=4,284$ cells; 10X_NCI_B, $n=4,136$ cells; 10X_NCI_M_A, $n=1,372$ cells; 10X_NCI_M_B, $n=2,082$ 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, $n=1,134$ cells; ICELL8_SE_B, $n=1,078$ cells; ICELL8_PE_A, $n=980$ cells; ICELL8_PE_B, $n=954$ cells). For detailed statistics regarding minima, maxima, center, bounds of box and whiskers and percentiles related to this figure, please refer to Supplementary Table 6.

methods (Fig. 3a–g, Supplementary Fig. 4a–g and Supplementary Tables 6 and 7). When comparing silhouette scores across different scRNA-seq platforms and datasets, we noticed that the 10x scRNA-seq data gave consistently lower scores than did 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 preprocess scRNA-seq data to remove uninteresting differences, such as those from 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 scLVM⁴⁰, to remove variations before applying subsequent analysis methods⁴⁰. 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).

Because log transformation has a large impact on downstream feature selection and clustering analysis, and because our analysis

showed that it performed similarly to sctransform, scran deconvolution and Linnorm, we used logCPM in our subsequent batch-effect and benchmarking evaluations, except when 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 version 3 (ref. ³⁸), fastMNN or mutual nearest neighbors (MNN)⁶, Scanorama⁸, BBKNN⁹, Harmony¹⁰, limma⁴² and ComBat¹³. We visualized clustering projections with both *t*-distributed stochastic neighbor embedding (*t*-SNE) and uniform manifold approximation and projection (UMAP)⁴⁴ and applied quantitative metrics, including silhouette width, kBET⁴⁵ and alignment score⁷, to evaluate batch-effect removal and cross-platform, cross-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 unmixed datasets, 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 (Figs. 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 (for example, *sctransform* for Seurat version 3). 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 and 5a and 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 version 3 was one of the best methods for grouping similar cells from different batches together, particularly for breast cancer cells, but overcorrected and clustered B lymphocytes and breast cancer cells, two highly dissimilar cell types, together, which was a misclassification (Fig. 4a,e,g, Extended Data Figs. 3a, 4 and 5a and Supplementary Figs. 6a and 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 and 5a and Supplementary Figs. 6a and 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 and 6 and Supplementary Fig. 6d). For fastMNN, a spike-in sample was required to provide a subpopulation of cells common to all samples analyzed. We also found that the order of dataset loading into fastMNN was critical for correcting batch effects; specifically, the mixed 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 (scenario 4; Fig. 4d), *t*-SNE and UMAP showed that the cells 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 and Supplementary Fig. 6b–d). When we applied batch-effect correction methods, we observed that Harmony, Seurat version 3 and fastMNN had the highest kBET score, with respect to their ability to group similar cells together, for breast cancer cells and B cells in scenarios 2 and 3 (Fig. 4b,c,g, Extended Data Figs. 3b,c and 5b,c and Supplementary Figs. 6b,c and 7c,d). Consistent with our findings in scenario 1, BBKNN analysis yielded 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 for cellular mixability (Fig. 4b,c,g, Extended Data Figs. 3b,c and 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 and 5c and Supplementary Figs. 6c and 7d). In the datasets with spike-ins from 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 version 3 (notwithstanding that the latter failed in scenario 1), whereas Harmony, Seurat version 3, BBKNN and Scanorama all performed well in grouping similar cells together (Fig. 4d,g, Extended Data Figs. 3d and 5d and Supplementary Figs. 6d and 7e,f).

We also compared mnnCorrect and fastMNN using the 20 scRNA-seq datasets and found that both versions performed similarly, as evaluated by *t*-SNE and UMAP. However, fastMNN took much less computation time; for example, 51.1 s for fastMNN versus 62,781.8 s for MNN in scenario 1 (1,229-fold faster, Supplementary Fig. 8).

Cell Ranger 3.1 allows some cells with extremely low gene expression to be identified. As a result, substantially more cells were detected using Cell Ranger 3.1 than with Cell Ranger 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 Cell Ranger 3.1 and Cell Ranger 2.0 preprocessed data, we noticed that the batch corrections performed better using Cell Ranger 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 three or five lung cancer cell lines in two batches. Consistent with findings from our own datasets, our analysis of the Tian et al. datasets showed that fastMNN, Seurat version 3 and Harmony performed well (Supplementary Figs. 11 and 12), whereas both Seurat version 2 CCA (canonical correlation analysis) and MNN failed to separate cells from the five different cell lines in the dataset from Tian et al.¹⁶.

Consistency of global gene expression across platforms and sites. We first evaluated global gene expression cross-platform consistency using scatterplots and the common transcripts detected across seven scRNA-seq datasets for either sample A or sample B (Extended Data Fig. 9a,b). The bar chart plots clearly showed that the BK RNA-seq had a much wider range of genes with high levels of expression than any of the scRNA-seq platforms, in which most genes had very low UMI counts. Overall, our scatterplot analyses showed cross-platform correlation coefficients between single-cell datasets of 0.8–0.98; correlation coefficients between scRNA-seq and BK 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 correlation for both breast cancer cells (sample A) and B lymphocytes (sample B) than the interplatform correlation. 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 (Methods) that expressed 500 abundant, 500 intermediate and 500 scarce genes, as defined by BK RNA-seq data (Supplementary Fig. 13a–f). To account for variable sequencing depth across the different datasets, we performed downsampling to 100,000 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 technology (that is, 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 technologies but not C1 technologies (Supplementary

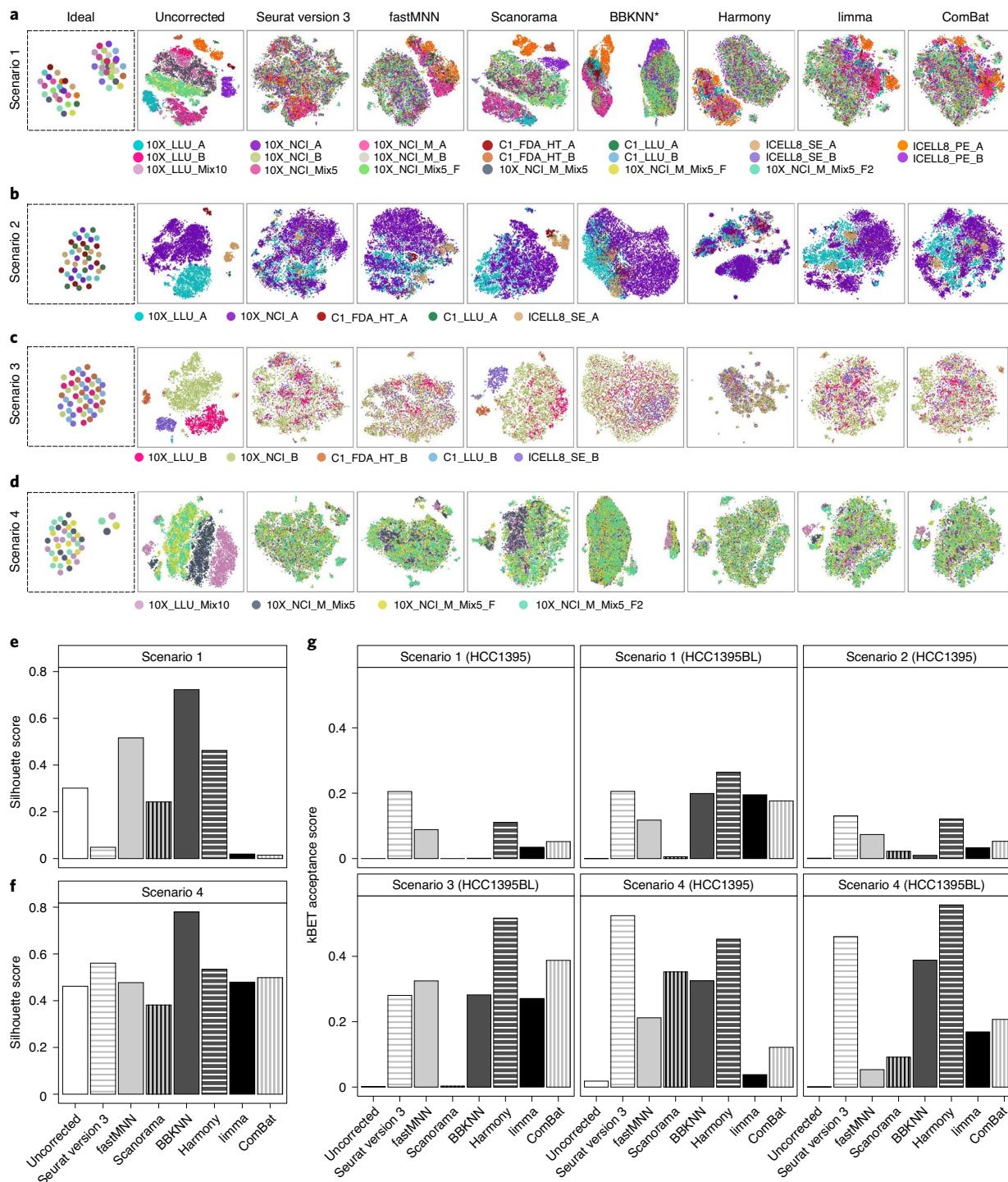


Fig. 4 | Batch-effect corrections evaluated in four different sample composition scenarios. **a**, Batch-effect correction in scenario 1, in which all 20 scRNA-seq datasets were combined, including mixed and unmixed datasets, with large proportions of two dissimilar types of cells (sample A, breast cancer cell line HCC1395; sample B, B lymphocyte line HCC1395BL). Datasets from the 10x platform were downsampled to 1,200 cells per dataset. **b**, Batch-effect correction in scenario 2, in which five scRNA-seq datasets (10X_LLU_A, 10X_NCI_A, C1_FDA_HT_A, C1_LLU_A and ICCELL8_SE_A) from breast cancer cells were generated separately at four centers (LLU, NCI, FDA and TBU) on four platforms (10x, Fluidigm C1, Fluidigm C1_HT and TBU ICCELL8). **c**, Batch-effect correction in scenario 3, in which five scRNA-seq datasets (10X_LLU_B, 10X_NCI_B, C1_FDA_HT_B, C1_LLU_B and ICCELL8_SE_B) from B lymphocytes were generated separately at the four centers on the same four platforms. **d**, Batch-effect correction in scenario 4, in which 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% 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-d**. Idealized projections of cells for the four different scenarios are presented on the left. *Note that for the BBKNN analysis, only data generated by UMAP are available and shown; all others are t-SNE plots. Silhouette width score quantifying the clusterability for scenario 1 (**e**) or scenario 4 (**f**), corresponding to **a** and **d**, respectively. **g**, kBET acceptance score quantifying mixability, calculated using the cross-platform, cross-center scRNA-seq data acquired either from breast cancer cells only or from B lymphocytes only from each of the four scenarios (**a-d**, also labeled as scenarios 1-4).

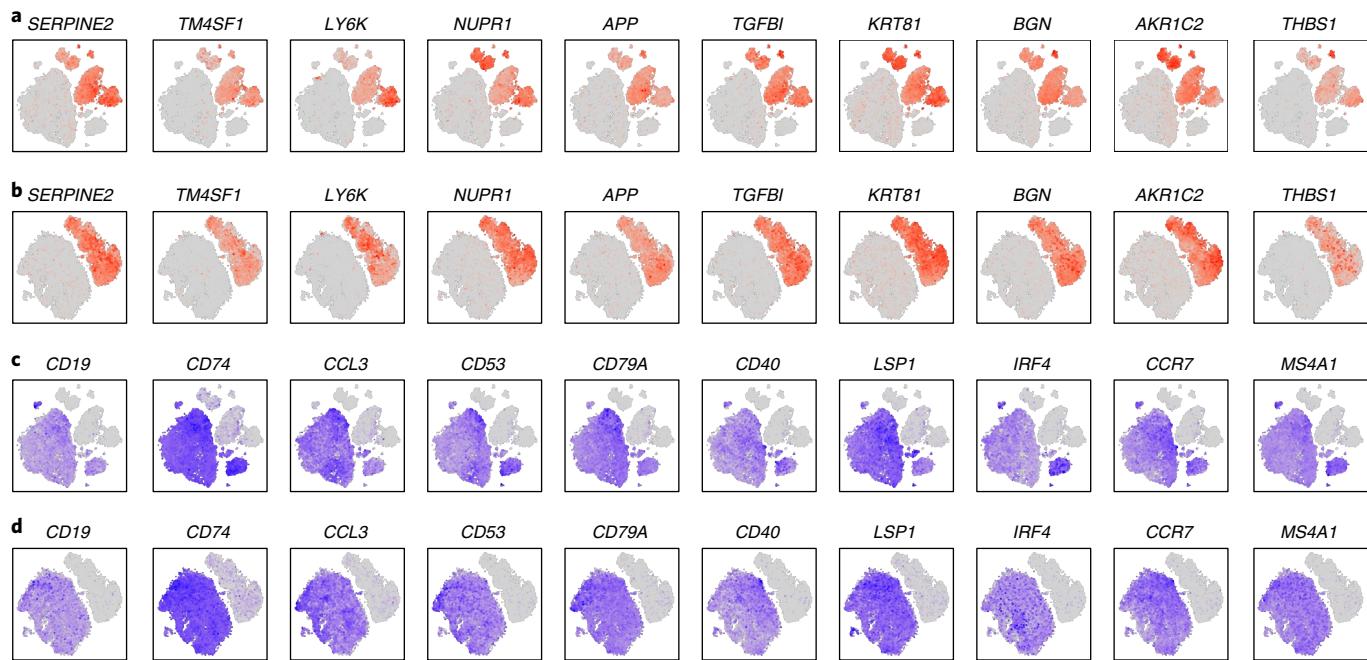


Fig. 5 | Feature plots showing cell type clustering based on cell type-specific marker genes across 20 scRNA-seq datasets. Feature plots generated across 20 scRNA-seq datasets using the top ten DEGs specific for breast cancer cells before batch-effect correction (**a**), breast cancer cells after fastMNN batch-effect correction (**b**), B lymphocytes before batch correction (**c**) and B lymphocytes after fastMNN batch-effect correction (**d**). Datasets from the 10x platform were downsampled to 1,200 cells per dataset. In feature plots, genes with relatively high levels of expression in each cell are highlighted in brick red (corresponding to breast cancer cells, sample A) or blue (corresponding to B cells, sample B).

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 cell numbers were very different across platforms (that is, 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 (linc)RNA and miscellaneous RNA (Supplementary Fig. 14 and Supplementary Table 10). As a comparison, the BK 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 RNA transcripts 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 ten cancer-specific DEGs and top ten B cell-specific DEGs^{46,47}, based on the DEGs derived from BK 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, breast cancer cells (sample A) and

B lymphocytes (sample B) were neither clustered together, nor clearly separated (Fig. 5a,c). 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,d).

We further compared the consistency of the single-cell gene expression profiles across platforms for *CD40*, *CD74* and *TPM1* with a subsampling at 100,000 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, which were 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 in the C1_FDA_HT dataset to as many as 53% in the C1_LLU dataset. A substantial percentage of cells (44–44.6% for the 10x platform and 23.2–28.1% for C1 and ICELL8 platforms) 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 the C1_FDA_HT dataset, in which 5% of the B cells had an intermediate level of this transcript ($1 \leq CPM < 10$; 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 3'-scRNA-seq technologies (Supplementary 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* expression, there was little or no detection of this transcript ($CPM < 1$) (Supplementary Table 13).

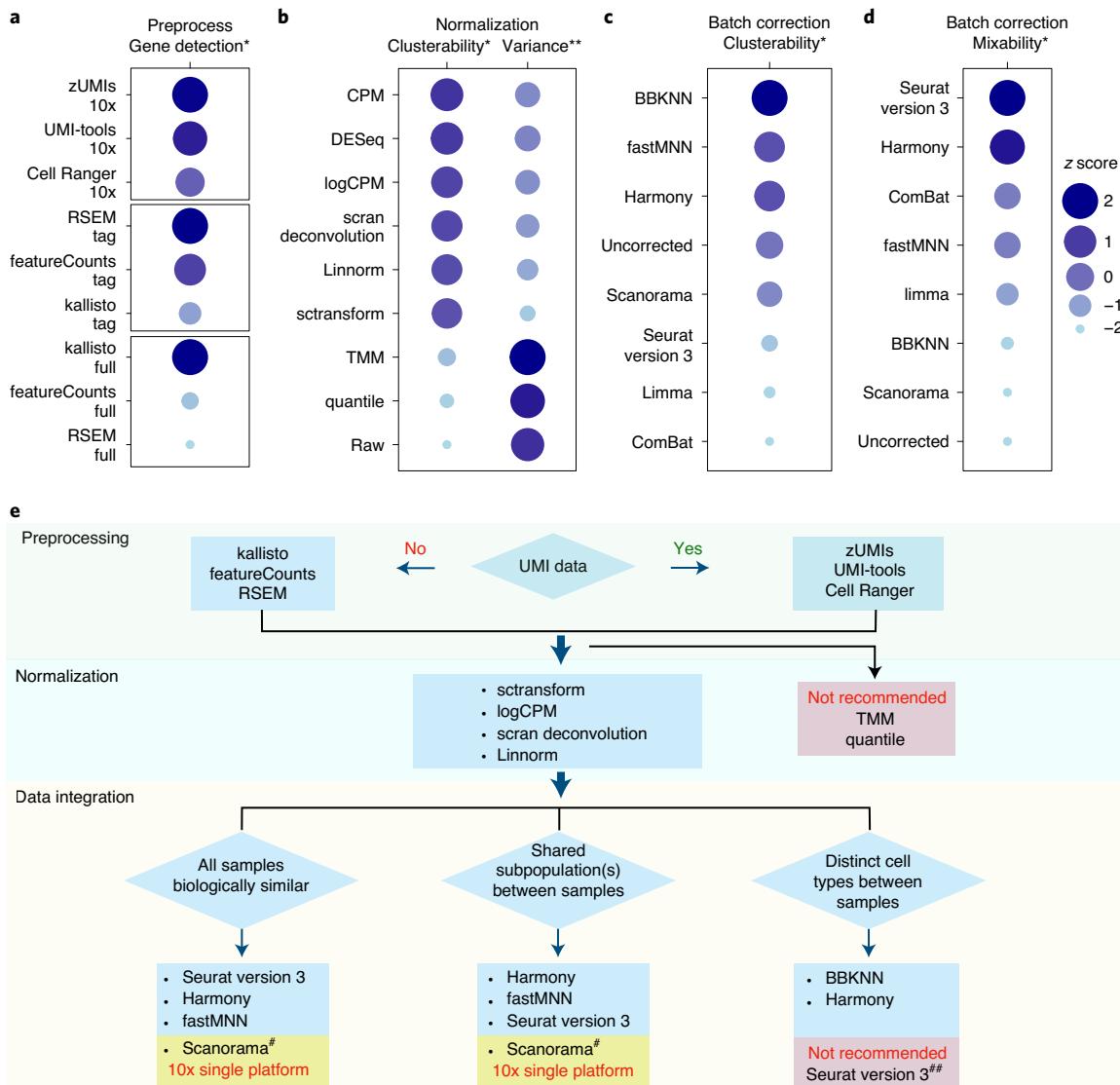


Fig. 6 | Performance ranking of bioinformatic metrics and best practice recommendations. **a**, Gene detection sensitivity measured separately for each of the three scRNA-seq protocol classes, 10x, non-10x-based 3' tagging and full-length. **b**, Normalization methods ranked by their clusterability, 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 scores from the harmonic mean of the silhouette scores (scenarios 1 and 4). **d**, Batch-correction methods ranked by their mixability, as measured by z scores from the harmonic mean of kBET acceptance scores (scenarios 1–4). z scores are plotted as circles with their sizes and color shades scaled to z-score values from highest to lowest 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 scRNA-seq analysis. #The current version of Scanorama did not correct batch effects for data from multiple platforms; however, it performed well when only 10x Genomics data were analyzed. ##Seurat version 3 was suitable for biologically similar samples but overcorrected batch effects and misclassified cell types if large fractions of distinct cell types were present in different batches.

Discussion

Here, we assessed scRNA-seq performance across four 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 lines¹⁸, either separately or as mixtures, for which a large amount of multiplatform WGS and WES data are available¹⁹. 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 et al. study⁴⁸, carried out by the Genome in a Bottle Consortium, which aimed to develop reference materials, data and methods to enable the

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 preprocessing and normalization contributed to variability in gene detection and cell classification, batch effects were large and the ability to assign cell types correctly across platforms and sites was dependent on bioinformatic pipelines, particularly the batch-correction algorithms used. In many scenarios, Seurat version 3, Harmony, BBKNN and fastMNN allowed for correct classification of the two cell types. However, when samples containing large fractions of biologically

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 batch effects.
4. Different normalization strategies performed differently across datasets and platforms; *sctransform*, *scran*, *logCPM* and *Linnorm* performed well for either 3'- or full-length-transcript scRNA-seq platforms, but *TMM* and *quantile* performed poorly and are not recommended.
5. Seurat version 3, Harmony, BBKNN, fastMNN and Scanorama all could correct and remove batch variations in specific sample and dataset scenarios; we recommend users apply appropriate batch-effect correction methods depending on the characteristics of their datasets (for example, cellular and sample heterogeneity and composition, platforms used; Fig. 6e).
6. BBKNN, fastMNN and Harmony ranked best for clusterability/cell type classification, whereas Seurat version 3, Harmony and fastMNN performed best for mixability.
7. fastMNN, BBKNN and Harmony removed batch variations effectively across different platforms, including both mixed

and unmixed distinct samples, but the order of importing the datasets into the pipeline and the requirement for a mixed sample was critical for MNN and fastMNN, whereas BBKNN and Harmony performed well regardless of the inclusion of mixed heterogeneous biologically distinct samples across platforms and batches; thus, for MNN and fastMNN, we recommend including a mixed sample and importing the mixed data into the pipeline first.

8. CCA/Seurat version 3 had superior mixability for biologically similar samples but overcorrected batch effects and misclassified cells (that is, poor clusterability/cell type classification) when large proportions of distinct cell types were present. However, Seurat version 3 performed well both for clusterability and mixability for datasets when only a small fraction of dissimilar cells (for example, 5–10%) was present. Thus, we do not recommend using CCA Seurat version 3 for scenarios containing large fractions of biologically distinct cell types.
9. BBKNN performed best in clusterability and cell type classification, but it ranked low in mixability, particularly in heterogeneous cell samples.
10. The current version of Scanorama performed well only for the 10x Genomics data and did not work for non-10x platforms; thus, we do not recommend it for data from non-10x platforms.
11. We observed good consistency between Cell Ranger 3.1 and 2.0 preprocessed data; however, Cell Ranger 3.1 can detect some extra cells with very few transcripts; this may affect batch-effect corrections in certain scenarios.

distinct cell types were compared, Seurat version 3 overcorrected the batch effect and misclassified the cell types (that is, breast cancer cells and B lymphocytes clustered together), whereas limma and ComBat failed to remove batch effects. However, we also showed that cross-center, cross-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 bioinformatic 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 kind of experimental variability likely to be encountered in real-world collaborations (in contrast to the situation with the Genome in a Bottle Consortium⁴⁸ or our companion paper¹⁹, in which identical aliquots of gDNA reference material were distributed to the study sites). As expected, site-to-site and platform-to-platform variabilities were 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 an 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 unmixed 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 subcultured for a few rounds without materially affecting the accuracy of cell clustering or identification, as demonstrated by our study. We also recommend obtaining BK 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 11–13 and Fig. 5).

The acquired scRNA-seq data can be preprocessed using any of the methods ranked in Fig. 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 (Fig. 6e and Box 1). Moreover, our scRNA-seq reference datasets and results from the bioinformatic 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 preprocessing and normalization contributed to variability in gene detection and cell classification. For UMI-based datasets, Cell Ranger version 3 detected the most cells, whereas zUMIs detected the most genes per cell. For 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. Of the eight normalization methods evaluated, *sctransform*, *logCPM*, *scran* deconvolution and *Linnorm* all performed well, with *sctransform* yielding the lowest variance. In contrast, *TMM* and *quantile* performed poorly across all datasets (Figs. 3 and 6b). Moreover, we found that regressing mitochondrial genes and normalizing UMI counts could not remove the batch effects (Extended Data Fig. 2 and Supplementary Fig. 5).

An ideal workflow would remove variability due purely to technical factors or sampling ambiguity without obscuring meaningful biological differences that are important to accurate classification of diverse cell types¹⁵. 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 version 3, Harmony, BBKNN, fastMNN and Scanorama performed well in removing batch effects and classifying cells in some scenarios. However, sample heterogeneity, dataset composition and the platform used influenced the outcome of the analyses (Fig. 4a–g, Extended Data Figs. 3a–d, 4, 5a–d, 6a–d, 7 and 8a–f and Supplementary Figs. 6a–d and 7a–f). For example, despite its high mixability for data from the same cell line (Fig. 4b,c,g), Seurat version 3 overcorrected batch effects and misclassified cells in datasets containing large fractions of highly dissimilar cells (Fig. 4a,e and Extended Data Figs. 3a and 4). Scanorama performed 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 algorithm⁸; we found that this algorithm failed to remove batch effects both from our own non-10x platform data and when we reanalyzed the data from Hie et al.⁸ (Extended Data Fig. 10a–f).

BBKNN performed the best, whereas limma and ComBat performed poorest in the cross-platform, cross-center separation of two types of cells from each other, particularly when there were large proportions of dissimilar cells in the datasets (Fig. 4a,e and 6c). In contrast, Seurat version 3, despite failing to separate distinct cell types in the multiplatform datasets consisting of large proportions of each cell type (Fig. 4a,e), performed well with 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 in which 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 version 3, fastMNN and Harmony all performed well in mixability for the scRNA-seq data derived from biologically identical or similar samples across platforms and sites, whereas limma and ComBat could mix B cells well across platforms and sites, perhaps because these cells are more homogeneous (Fig. 4b,c,g and Extended Data Figs. 3b,c and 4). Indeed, for all algorithms, we found that the cross-platform, cross-center mixability was better in B cells than in the more heterogeneous cancer cells (Fig. 4b,c,g). However, for fastMNN and 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 requirement for the presence of mixed cell types to provide a shared subpopulation across the batches to be corrected is consistent with the logic of the MNN algorithm⁶. Thus, our findings clearly highlight that the choice of appropriate batch-effect method depends on the characteristics of the samples (for example, 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 perform 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 (Supplementary Fig. 8d in their paper)¹⁶. However, when we applied our preferred preprocessing method to the datasets of Tian et al., fastMNN, Seurat version 3 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 bioinformatic methods, particularly with regard to batch-effect correction algorithms¹⁶. One example is the requirement for shared subpopulations

(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). Mereu et al.¹⁵ used mixed cells of three different species but focused primarily on the analytical platform, with limited investigation of bioinformatic methods. Nevertheless, as in the study of Tian et al., no non-mixture cells or 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—indpendently—assess the ability of a method to separate dissimilar cells correctly. We found that analyzing both mixtures of dissimilar cells in various proportions and unmixed samples of the two distinct cell lines provided important additional insights. For example, the widely used Seurat version 3 method³⁸ excelled at grouping similar cells together, but it overcorrected—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 cell 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 samples and 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 data¹⁹, which will be a valuable resource for benchmarking future single-cell WGS or proteomic 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 that this will provide a useful resource for the community to benchmark additional scRNA-seq protocols and bioinformatic 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. The capabilities and limitations of the methods we evaluated are displayed graphically in Fig. 6. Best practice recommendations based on our findings are presented in Box 1 and in flow-chart format in Fig. 6e.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41587-020-00748-9>.

Received: 10 May 2019; Accepted: 22 October 2020;

Published online: 21 December 2020

References

- Klein, A. M. et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187–1201 (2015).
- Macosko, E. Z. et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).

3. Gierahn, T. M. et al. Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput. *Nat. Methods* **14**, 395–398 (2017).
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).
5. Wu, H., Wang, C. & Wu, S. Single-cell sequencing for drug discovery and drug development. *Curr. Top. Med. Chem.* **17**, 1769–1777 (2017).
6. Haghverdi, L., Lun, A. T. L., Morgan, M. D. & Marioni, J. C. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.* **36**, 421–427 (2018).
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).
8. Hie, B., Bryson, B. & Berger, B. Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. *Nat. Biotechnol.* **37**, 685–691 (2019).
9. Polanski, K. et al. BBKNN: fast batch alignment of single cell transcriptomes. *Bioinformatics* **36**, 964–965 (2019).
10. Korsunsky, I. et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).
11. Saelens, W., Cannoodt, R., Todorov, H. & Saeys, Y. A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.* **37**, 547–554 (2019).
12. Ziegenhain, C. et al. Comparative analysis of single-cell RNA sequencing methods. *Mol. Cell* **65**, 631–643 (2017).
13. Zhang, X. et al. Comparative analysis of droplet-based ultra-high-throughput single-cell RNA-seq systems. *Mol. Cell* **73**, 130–142 (2019).
14. Svensson, V. et al. Power analysis of single-cell RNA-sequencing experiments. *Nat. Methods* **14**, 381–387 (2017).
15. Mereu, E. et al. Benchmarking single-cell RNA-sequencing protocols for cell atlas projects. *Nat. Biotechnol.* **38**, 747–755 (2020).
16. Tian, L. et al. Benchmarking single cell RNA-sequencing analysis pipelines using mixture control experiments. *Nat. Methods* **16**, 479–487 (2019).
17. Tran, H. T. N. et al. A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biol.* **21**, 12 (2020).
18. Gazdar, A. F. et al. Characterization of paired tumor and non-tumor cell lines established from patients with breast cancer. *Int. J. Cancer* **78**, 766–774 (1998).
19. Xiao, W. et al. Towards best practice in cancer mutation detection with whole-genome and whole-exome sequencing. *Nat. Biotechnol.* (in the press).
20. Zhang, J., Spath, S. S., Marjani, S. L., 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).
21. Chen, X. et al. A multi-center cross-platform single-cell RNA sequencing reference dataset. Preprint at bioRxiv <https://doi.org/10.1101/2020.09.20.305474> (2020).
22. Zhang, M. J., Ntranos, V. & Tse, D. Determining sequencing depth in a single-cell RNA-seq experiment. *Nat. Commun.* **11**, 774 (2020).
23. Li, B. et al. Cumulus provides cloud-based data analysis for large-scale single-cell and single-nucleus RNA-seq. *Nat. Methods* **17**, 793–798 (2020).
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).
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**, giy059 (2018).
26. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
27. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).
28. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
29. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* **17**, 10–12 (2011).
30. Bolger, A. M., Lohse, M. & Usadel, B. J. B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
31. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
32. Hicks, S. C., Townes, F. W., Teng, M. & Irizarry, R. A. Missing data and technical variability in single-cell RNA-sequencing experiments. *Biostatistics* **19**, 562–578 (2017).
33. Risso, D., Ngai, J., Speed, T. P. & Dudoit, S. Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat. Biotechnol.* **32**, 896–902 (2014).
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).
35. Lun, A. T., Bach, K. & Marioni, J. C. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol.* **17**, 75 (2016).
36. Bacher, R. et al. SCnorm: robust normalization of single-cell RNA-seq data. *Nat. Methods* **14**, 584–586 (2017).
37. Yip, S. H., Wang, P., Kocher, J.-P. A., Sham, P. C. & Wang, J. Linnorm: improved statistical analysis for single cell RNA-seq expression data. *Nucleic Acids Res.* **45**, e179 (2017).
38. Stuart, T. et al. Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902 (2019).
39. Yip, S. H., Sham, P. C. & Wang, J. Evaluation of tools for highly variable gene discovery from single-cell RNA-seq data. *Brief Bioinform.* **20**, 1583–1589 (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).
41. Kang, H. M. et al. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat. Biotechnol.* **36**, 89–94 (2018).
42. Ritchie, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
43. Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E. & Storey, J. D. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* **28**, 882–883 (2012).
44. Becht, E. et al. Dimensionality reduction for visualizing single-cell data using tSNE. *Nat. Biotechnol.* **37**, 38–44 (2019).
45. Buttner, M., Miao, Z., Wolf, F. A., Teichmann, S. A. & Theis, F. J. A test metric for assessing single-cell RNA-seq batch correction. *Nat. Methods* **16**, 43–49 (2019).
46. Kaminski, D. A., Wei, C., Qian, Y., Rosenberg, A. F. & Sanz, I. Advances in human B cell phenotypic profiling. *Front. Immunol.* **3**, 302 (2012).
47. Starlets, D. et al. Cell-surface CD74 initiates a signaling cascade leading to cell proliferation and survival. *Blood* **107**, 4807–4816 (2006).
48. Zook, J. M. et al. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nat. Biotechnol.* **32**, 246–251 (2014).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2020

Methods

Study design. A schematic overview of the study design is illustrated in Fig. 1a. Briefly, two well-characterized reference cell lines (sample A, breast cancer cell line and sample B, a matched control B lymphocyte line) were used to generate scRNA-seq data on 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 protocols. At the LLU and NCI sites, 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 cell spiked into B lymphocytes, named as fixed_1 and fixed_2 in two independent sample captures). The 10x scRNA-seq libraries constructed at the 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). BK RNA-seq samples were also prepared from these cell lines in triplicate. All scRNA-seq data were subjected to three different preprocessing pipelines, for the 10x, C1 or ICELL8 technology, respectively. We evaluated eight normalization methods, sctransform, scran deconvolution, CPM, logCPM, TMM, DESeq, quantile and Linnorm and seven batch-effect correction algorithms, including Seurat version 3, 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 with scatterplots and feature plots. scRNA-seq data were also compared with population-average RNA-seq data.

Abbreviations and notations for Fig. 1a are as follows: 10X_LLU, single cells were captured using a 10x Genomics Chromium controller and scRNA-seq was performed 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, ~(4–4.77) × 10⁶ reads per cell); ICELL8_PE, single cells were captured using an ICELL8 chip at TBU, and scRNA-seq libraries were sequenced (75 × 2 bp, PE) at TBU; ICELL8_SE, the same scRNA-seq libraries generated at the TBU site were also sequenced at LLU (150 × 1 bp, SE, ~1 × 10⁶ reads per 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 (HCC1395BL, sample B) from the American Type Culture Collection (ATCC). The two cell lines were derived from the same human patient (43 years old, female). HCC1395 cells were cultured in RPMI 1640 medium supplemented with 10% FBS. HCC1395BL cells were cultured in 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 (HCC1395BL). We passed all cells through a 30-μm 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 per μl. Capture occupancy and live and 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 cDNA was generated using the Fluidigm C1 Single-Cell System at LLU with the SMART-Seq version 4 Ultra Low Input RNA kit (Takara Bio) according to the manufacturer's protocol. Only cDNA generated from live single cells was used for further library construction.

Libraries were prepared using the modified Illumina Nextera XT DNA library preparation protocol. Briefly, concentrations of cDNA 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⁻¹. Then, 1.25 μl of diluted cDNA was incubated with 1.25 μl tagmentation mix and 2.5 μl tagment DNA buffer for 10 min at 55 °C. Tagmentation was terminated by adding 1.25 μl NT buffer, and samples were 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 volumes. Eighty libraries were generated from HCC1395 cells (sample A) and 66 libraries were generated from HCC1395BL cells (sample B). Library pools were sequenced on an Illumina HiSeq 4000 sequencer (150 × 2 bp, PE) at the LLU Center for Genomics.

Single-cell 3' end RNA-seq using the C1 Fluidigm high-throughput system. High-throughput (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 an HT IFC at a concentration of 400 cells per μl (Nexcelom Cellometer Auto T4). Capture occupancy and live and dead cells at the capture site were recorded using a fluorescence microscope after staining with the 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 performed by following the Nextera XT DNA library preparation guide. Only polyadenylated RNA containing the preamplification adaptor sequence at both ends was amplified. Lastly, sequencing adaptors and Nextera indices were applied during library preparation. Only the 3' ends of the transcripts were enriched following PCR amplification.

Two hundred and three libraries were generated from HCC1395 cells (sample A) and 241 libraries were generated from HCC1395BL cells (sample B). Library pools were sequenced on an Illumina HiSeq 2500 (75 × 2 bp, PE) at the FDA's Genomics Facility.

Single-cell RNA-seq using the 10x Genomics platform. After filtering with a 30-μm MACS SmartStrainer (Miltenyi Biotec), single cells were resuspended in PBS (calcium- and magnesium-free) containing 0.04% (wt/vol) BSA (400 μg ml⁻¹) and further diluted to 300 cells per μl after cell counting (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.

scRNA-seq library preparation was performed by following the 3' scRNA-seq 10x Genomics platform protocol using version 2 chemistry. Briefly, based on the cell suspension volume calculator table, 3,000 cells (17.4 μl of a suspension of 300 cells per μl) and barcode beads, as well as RT reagents were loaded into the Chromium Controller to generate single Gel Bead-in-Emulsions (GEMs). cDNA was 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 the SPRIselect reagent, cDNA was incubated for fragmentation, end repair, A tailing and adaptor ligation. Lastly, sequencing library amplification was performed using sample index primers for ten cycles.

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 spikein_5%_Fixed_1 sample, the normal and tumor cells were harvested, washed, counted, and 5% spike-ins of breast cancer cells with 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 1× Dulbecco's PBS (DPBS) at 4 °C and resuspended gently in 100 μl 1× DPBS (Thermo Fisher Scientific). Chilled methanol (900 μl, 100%) was then added, drop by drop, to the cells with gentle vortexing. The cells were then fixed on ice for 15 min and were stored at 4 °C for 6 d. For rehydration, fixed cells were pelleted by centrifugation at 3,000g for 10 min at 4 °C and washed twice with 1× DPBS containing 1% BSA and 0.4 μl⁻¹ RNase inhibitor (Sigma Aldrich). The cells were then counted, and the concentration was adjusted to be close to 1,000 cells per μl. Approximately 8,000 cells were loaded onto a single-cell chip for GEM generation using the 10x Genomics Chromium controller. mRNA-seq (3') gene expression libraries were prepared using the Chromium Single Cell 3' Library & Gel Bead kit, version 2 (10x Genomics), according to the manufacturer's 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 1× DPBS and resuspended in 10% 1× DPBS and 90% chilled methanol, as described above. The cells were then fixed on ice for 15 min and stored at 4 °C for 24 h. For rehydration, fixed cells were washed with 1× DPBS containing 1% BSA and 0.4 μl⁻¹ RNase inhibitor and counted. Approximately 8,000 cells were loaded onto a single-cell chip for GEM generation using the 10x Genomics Chromium controller. mRNA-seq (3') gene expression libraries were prepared using the Chromium Single Cell 3' Library & Gel Bead kit version 2 (10x Genomics) according to the manufacturer's guidelines.

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

Single-cell RNA-seq using the Takara Bio ICELL8 platform. We also constructed ICELL8 scRNA-seq libraries at TBU. For ICELL8 cell preparation and single-cell selection, a bulk cell suspension of either cancer or B cells (~1 × 10⁶ cells each) was fluorescently labeled with a premade mix of Hoechst 33324 and propidium iodide (ReadyProbes Cell Viability Imaging kit, Thermo Fisher Scientific) in the 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 according to the manufacturer's instructions (Thermo Fisher Scientific) to dissociate cells from the flask surface. The cells were washed in 1× PBS (no Ca²⁺, Mg²⁺, phenol red or serum, pH 7.4; Thermo Fisher Scientific) and centrifuged (100g, 3 min) and resuspended in 1 ml 1× DPBS. Cell counts were determined using a Moxi Flow cell counter (ORFLO Technologies) and diluted to ~1 cell in 35 nl (~28,600 cells per ml) in 1× DPBS (no Ca²⁺, Mg²⁺, phenol red or serum, pH 7.4; Thermo Fisher Scientific) containing Second Diluent (1×), RNase inhibitor (0.4 U) and 1.92 μM of the 3' oligo(dT) terminating primer (SMART-Seq ICELL8 CDS (Takara Bio)).

Each cell type solution was dispensed from a 384-well source plate into individually addressable wells in a 5,184-nanolwell, 250-nl-volume ICELL8 chip (SMARTer ICELL8 250v Chip, Takara Bio) using a MultiSample NanoDispenser (MSND, SMARTer ICELL8 Single-Cell System, Takara). Chip wells were sealed using SmartChip Optical Imaging Film (Takara Bio) and centrifuged at 300g for 5 min at 22 °C. All nanowell in the chip were imaged with a ×4 objective using Hoechst and Texas Red excitation and emission filters. Images (TIFF format) were analyzed using CellSelect automated microscopy image analysis software (Takara Bio). The chip was stored in a chip holder at -80 °C overnight. Image analysis confirmed that cell deposition followed a Poisson distribution. Six hundred individual nanowells, each bearing microscopically 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 the 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 3,220g (3 min). All on-chip thermal cycling was performed using a SMARTer ICELL8 Thermal Cycler (Takara Bio).

For in-chip, full-length cDNA synthesis, the ICELL8 chip (containing dispensed samples) was thawed at room temperature for 10 min and centrifuged at 3,220g 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 nanowells (identified as bearing a single cell by the ICELL8 filter file), and the reactions were thermal cycled in-chip as follows: 45.6 °C, 5 s; 41 °C, 90 min; 99 °C, 9 s; 95.5 °C, 1 min; 100 °C, 5 s; 99 °C, 7 s; 9 °C, 5 s; 64 °C, 30 s; 69.5 °C, 5 s; 67.5 °C, 3 min; go to step 5 and repeat 7×; 4 °C hold.

For in-chip P5 index addition and fragmentation, 72 primer sequences bearing P5 indices (SMART-Seq ICELL8 Forward Indexing Primer Set A (5'-AATGATACGGCGACCCAGGAGTCTACAC(i5)TCGTCGGCAGCGTC-3'), where i5 refers to 1 of 72 unique, eight-nucleotide indices (Hamming distance between P5 indices, 3), were dispensed from a prealiquoted 384-well plate in 35 nl aliquots into 72 filter file-identified, nanowell 'rows'. The chip was sealed with Microseal A film and centrifuged at 3,220g (3 min) at 4 °C before returning to the MSND, permitting addition of the fragmentation master mix containing MgCl₂, Nextera Amplicon Tagment Mix (Illumina), Terra PCR Direct Polymerase Mix and TRH (Takara Bio). The chip was sealed with Microseal A film and centrifuged again as above. Fragmentation was performed in-chip at the following temperatures: 42 °C, 4 s; 37 °C, 30 min; 4 °C hold.

For in-chip P7 index and PCR reagent addition, the first PCR generated 5,184 unique indices. A reagent mix containing 72 primers bearing P7 indices (SMART-Seq ICELL8 Reverse Indexing Primer Set A (5'-CAAGCAGAACGCGCA TACGAGAT(i7)GTCTCGTGGGCTCGG-3'; i7 refers to 1 of 72 unique, eight-nucleotide indices (Hamming distance between P7 indices, 3)) was dispensed from the same prealiquoted 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 nanowells. The chip was sealed with SmartChip Sealing Film and centrifuged at 3,220g for 3 min at 4 °C. PCR cycling was performed as follows: 77 °C, 12 s; 72 °C, 3 min; 99 °C, 11 s; 95.5 °C, 1 min; 100 °C, 20 s; 99 °C, 10 s; 53.3 °C, 5 s; 58 °C, 15 s; 71 °C, 5 s; 67.5 °C, 2 min; go to step 5 and repeat 7×; 4 °C hold.

For 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 according to the manufacturer's instructions (Takara Bio). Fifty percent of the extracted library was purified twice using a 1× proportion of AMPure XP beads (Beckman Coulter) at a final volume of 14 μl in the elution buffer provided with the SMART-Seq ICELL8 Reagent kit.

For off-chip library amplification (second PCR), the first round amplicon (14 μl, from above, doubly purified with AMPure beads) was PCR amplified in a 50 μl volume of second PCR mixture containing SeqAmp CB PCR Buffer (25 μl), 5× primer mix (P5 and P7 primers) and Terra PCR Direct Polymerase Mix (0.05 U μl⁻¹ final concentration (Takara Bio)) via a thermal protocol as follows: 98 °C, 2 min 1×; followed by eight thermal cycles of 98 °C, 10 s; 60 °C, 15 s; 68 °C, 2 min. This sequencing-ready library was purified using one round of a 1× proportion of AMPure XP beads (Beckman Coulter). The final elution volume was 17 μl in elution buffer.

For ICELL8 scRNA-seq library QC and sequencing, the scRNA-seq library concentration (ng μl⁻¹) was determined using a Qubit fluorometer (Thermo

Fisher). Based on Qubit readings, 1–2 ng μl⁻¹ was examined with a Bioanalyzer 2100 and a corresponding High Sensitivity DNA kit (Agilent) to determine the size of selected libraries. The Bioanalyzer amplicon sizes ranged from 200 to 3,000 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. Bulk cell total RNA was isolated from HCC1395 and HCC1395BL cells using the 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 bp using a Covaris S220 and then underwent end repairing to blunt the ends, followed by barcoded adaptor ligation. The remainder of the library preparation proceeded according to 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 BK 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 at all four sites. A total of 30,693 single cells were captured, from which 20 different scRNA-seq datasets were obtained (Supplementary Table 1 and Extended Data Fig. 1). Five libraries from the NCI site (10X_NCI_M) were also resequenced using a modified sequencing protocol (26+57 bp). Across all platforms and datasets, over 93.6% of the reads mapped to 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 data generated by the 10x platform showed that 35.3% of the mapped reads (or 57.0% of the exonic reads) were derived from deduplicated UMIs in breast cancer cells (sample A) and 22.2% of the mapped reads (or 35.2% of the exonic reads) were derived from deduplicated 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 nonspecific 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 protocols for the 10x Genomics scRNA-seq (Supplementary Fig. 1).

Bioinformatic methods. 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 the Ensembl version 84 transcriptome. All following bioinformatic 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 preprocessing pipelines, Cell Ranger (version 3.1.0), UMI-tools²⁴ (version 1.0.0) and zUMIs²⁵ (version 2.4.5) 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 UMI-tools and zUMIs pipelines, reads were filtered out if the phred quality scores of cell barcode bases or UMI bases was < Q10. In the zUMIs pipeline, option '-d' was used to perform downsampling analyses to eight fixed depths (5,000, 10,000, 25,000, 50,000, 100,000, 150,000, 200,000 and 250,000) to generate gene count tables. With UMI-tools, the umi_tools whitelist with default parameter settings was used to generate a list of cell barcodes for downstream analysis. The umi_tools extract was used to extract cell barcodes and filter reads (options: quality-filter-threshold, 10; filter-cell-barcode). STAR (version 2.5.4b)³¹ was used for alignment to generate BAM files containing the unique mapped reads (option, outFilterMultimapNmax 1) for gene counting. featureCounts (version 1.6.1)²⁶ was used to assign reads to genes and generate a BAM file (option, -R BAM). samtools (version 1.3)³⁰ sort and samtools index were used to generate sorted and indexed BAM files. Finally, umi-tools 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 the C1 and TBU ICELL platforms. For non-UMI-based samples, three preprocessing 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 (trimmmomatic (version 0.35)³⁰, trim_galore (version 0.4.1)³¹ or cutadapt (version 1.9.1)²⁹) was used to

process the raw fastq data. Bases with quality less than ten were trimmed from the 5' and 3' ends of reads. Reads with 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 (version 1.3.0)²⁸ or kallisto (version 0.43.1)²⁷ were used to generate gene counts per cell. All default parameter settings were used except the following: in RSEM, the 'single-cell-prior' option was used to estimate gene expression levels for scRNA-seq data; the 'paired-end' option was used if the data were PE fastqs; in kallisto, options '-l 1500' and '-s 120' were used to represent the estimated average fragment length and s.d. of fragment length if the data were SE fastqs.

Preprocessing and differential gene expression analysis of bulk RNA-seq data. The preprocessing pipeline of bulk RNA-seq data included QC (FastQC version 0.11.4), 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 (version 1.24.0) was used to perform the differential expression analysis between breast cancer samples (sample A) and B lymphocyte samples (sample B) with default parameters.

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

Normalization methods for all datasets. We investigated some existing bulk RNA-seq normalization procedures including CPM, TMM, UQ, 'DESeq' normalization, implemented in the DESeq Bioconductor package, and 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 downsampling of each cell to two different read depths (10,000 and 100,000 reads per cell) for each dataset and evaluated the performance of the normalization methods at these depths. Similar to the method used in the 'scOne' paper⁵², 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 such 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 is 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. A Monocle 2 (refs. ^{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 (version 0.99.5) acceptance score, modified alignment score and silhouette score.

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

In step 1, all 10x single-cell datasets were processed by the Monocle 2 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 + 2 s.d. and the lower bound as mean - 2 s.d. for both the total UMIs and genes. Cells with total UMIs or genes outside of the upper or lower bounds were removed.

In step 2, Seurat (version 3.0.3)-based data processing was applied to each dataset. Genes detected in fewer than three cells and cells containing less than 200 genes were removed from the datasets before 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 version 1.12.1 and SeuratWrappers version 0.1.0), Harmony (version 0.99.9), limma (version 3.40.4) and ComBat (version 3.32.1). For the Python-based batch-correction methods Scanorama (version 1.4) and BBKNN (version 1.3.5), a detailed description of data processing is provided in the 'Scanorama pipeline' and 'BBKNN pipeline' 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, t-SNE plots and UMAPs and 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 version 3, 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 datasets and perform batch correction. The script 'process.py' with default parameters was used to perform cell filtering and normalization. Two thousand HVGs were used in the function 'correct' to perform batch correction and generate Scanorama-corrected gene expression matrices.

BBKNN pipeline. The Seurat-inspired Scanpy (version 1.4.4) 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 the functions 'scanpy.api.pp.filter_cells' and 'scanpy.api.pp.filter_genes' with the same parameter settings as in step 1. The processed data matrices were merged to generate a master gene expression matrix and further log transformed and normalized by the 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 in 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 Fig. 8, the performance of fastMNN and MNN was compared. The steps to perform MNN correction are the same as for fastMNN, except for the batch correction (step 3). We used the function 'mnnCorrect' (scran package version 1.8.4) with default parameters to perform the batch correction.

Preprocessing and batch-effect correction on data from Tian et al. We preprocessed the data from Tian et al.¹⁶ using the Cell Ranger pipeline for 10x data and the UMI-tools 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.

Bioinformatic pipelines validated and performed in BioGenLink. We carried out some bioinformatic pipelines in BGL to cross-validate some of our bioinformatic data analyses. Bioinformatic tools were created in BGL for performing batch correction of scRNA-seq data using Seurat version 3, 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) with the function 'silhouette' from the R package cluster (version 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 pipeline from Buttner et al.⁴⁵ 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 as described in step 4 of section “Batch-effect correction pipeline.” The kBET acceptance 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 cell 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 ($i = 1 \dots N$, N is the total number of datasets).
2. For each cell j ($j = 1 \dots N_j$) 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 obtain x'_j .
3. Alignment score = $\sum_{i=1}^N w_i \left(1 - \frac{x'_j - w_i k}{k - w_i k} \right)$
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 and sites using all scRNA-seq data. To investigate the consistency of global gene expression across different platforms and sites and scRNA-seq datasets, we selected benchmarking genes according to the average gene expression ($\log_2(\text{TPM} + 1)$) 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 robust genes, we further filtered out genes with s.d. of gene expression greater than 1 across three replicates. 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 1,500 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 obtain comparable cell percentages, we considered only gene count matrices from the downsampling results (100,000 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.

Scatterplots. To assess the variation of gene expression across different datasets, we generated scatterplot matrices. For all platform-specific datasets, which include seven single-cell datasets and three BK 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 as follows.

$$L_m^{(i)} = \frac{1}{N} \sum_{j=1}^N G_{mj}^{(i)}, (m = 1, \dots, M) \text{ where } G_{mj}^{(i)} = \log(CPM_{mj}^{(i)} + 1)$$

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

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

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

The final normalized matrix A was used as input to generate scatterplot figures with 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 and 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 plots. To assess the scRNA-seq gene expression profiles across different platforms based on four different RNA groups including protein-coding RNA, antisense RNA, lincRNA and miscellaneous RNA, 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. See scatterplot

normalized count matrix computing method for details. The genes that 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 and the ggplot2_3.2.0 and dplyr_0.8.3 packages.

Feature plots. For the uncorrected data, Seurat objects from t-SNE dimensional reduction were used as the data source for generating feature plots. A total of 20 genes (ten for sample A, cancer cells; ten for sample B, B cells) were selected as cell type-specific markers for samples A and B based on both the BK RNA-seq DEG ranking and the 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.

Bioinformatic 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 and datasets, we used the normalized gene expression data (CPM values) from the downsampling results (100,000 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 $CPM > 0$, $0 < CPM < 1$, $1 \leq CPM < 10$ and $CPM > 10$.

Methods used to generate summary benchmarking statistics for the various bioinformatic pipelines. The performance of the various pipelines regarding preprocessing, normalization and batch-effect correction is summarized in Fig. 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 14 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 (six datasets), those processed with 3' end counting using the Fluidigm HT platform (two datasets) and those processed with the two full-length-based platforms (six 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 14 paired cancer cell datasets compared to the 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; data from spike-in experiments). To assess the 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).

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The datasets generated and analyzed in the current study are available in the SRA repository under accession code no. PRJNA504037. The data from Tian et al. are available at GEO accession code GSE118767. The data from Hie et al. are available at the following URL: <http://scanorama.csail.mit.edu/data.tar.gz>.

Code availability

We used many algorithms and code sets for batch correction that were previously published. All of our code is provided in GitHub and Code Ocean at the following links: https://github.com/oxwang/fda_scRNA-seq and <https://codeocean.com/capsule/0497386> or <https://doi.org/10.24433/CO.1559060.v1>.

References

49. Alles, J. et al. Cell fixation and preservation for droplet-based single-cell transcriptomics. *BMC Biol.* **15**, 44 (2017).
50. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
51. Krueger, F. Trim Galore! http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/ (2015).
52. Cole, M. B. et al. Performance assessment and selection of normalization procedures for single-cell RNA-seq. *Cell Syst.* **8**, 315–328 (2019).
53. Qiu, X. et al. Single-cell mRNA quantification and differential analysis with Census. *Nat. Methods* **14**, 309–315 (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).

Acknowledgements

We thank D. Ho of the LLU Center for Genomics for her great administrative support, particularly in coordinating the weekly Zoom conference calls and assistance in preparing meeting minutes for the SEQC-2 single-cell sequencing project. We thank ATCC, and particularly L. Kerrigan, for providing the two cell lines, that is, HCC1395 and HCC1395BL, for our study. We thank W. Jones at EA Genomics, Q² Solutions for critical review and helpful comments. We thank Z. Chen at LLU and J. Shetty at the NCI for technical assistance in performing sequencing. J. Bettridge at the NCI for technical assistance in 10x Genomics scRNA-seq library preparation, V. Furtak at the FDA for library preparation and W. Wu at the FDA/CBER Core Facility for Illumina sequencing. We also thank S. 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 (C.W.), the Ardmore Institute of Health grant 2150141 (C.W.) and C.A. Sims' gift to LLU Center for Genomics.

Author contributions

C.W. and W.X. conceived and designed the study. C.W. managed the project and directed bioinformatic data analyses. C.W. drafted the manuscript and annotated all the results.

M.M. Jr. and A.F. helped edit the manuscript. W.C., B.T., M.M., P.K., M.M. Jr., A.F. and A.M. performed single-cell culturing, single-cell captures and scRNA-seq library synthesis and sequencing. X.C., Z.Y., Y.Z., X.X., V.C., Y.B., B.E., W.X., U.A.M., J.L., J.-L.L. and C.W. performed bioinformatic data analyses. W.C., X.C., Z.Y., Y.Z., Y.B., X.X., V.C., M.M., A.M., M.M. Jr. and J.-L.L. prepared the methods for the manuscript. Z.Y. prepared all figures; W.C., C.W. and H.C. prepared the tables. C.W., M.M. Jr., W.C., A.F. and W.X. revised the manuscript. All authors reviewed the manuscript. C.W. finalized and submitted the manuscript.

Competing interests

A.F. and A.M. are employees of Takara Bio USA, Inc., and B.E. and U.A.M. 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.

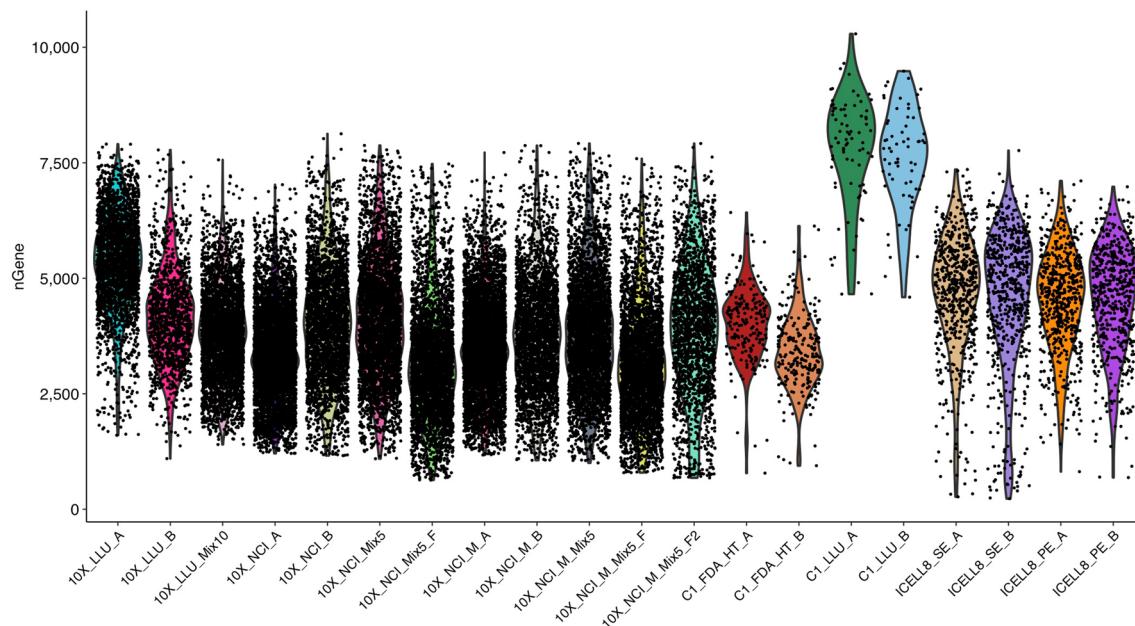
Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41587-020-00748-9>.

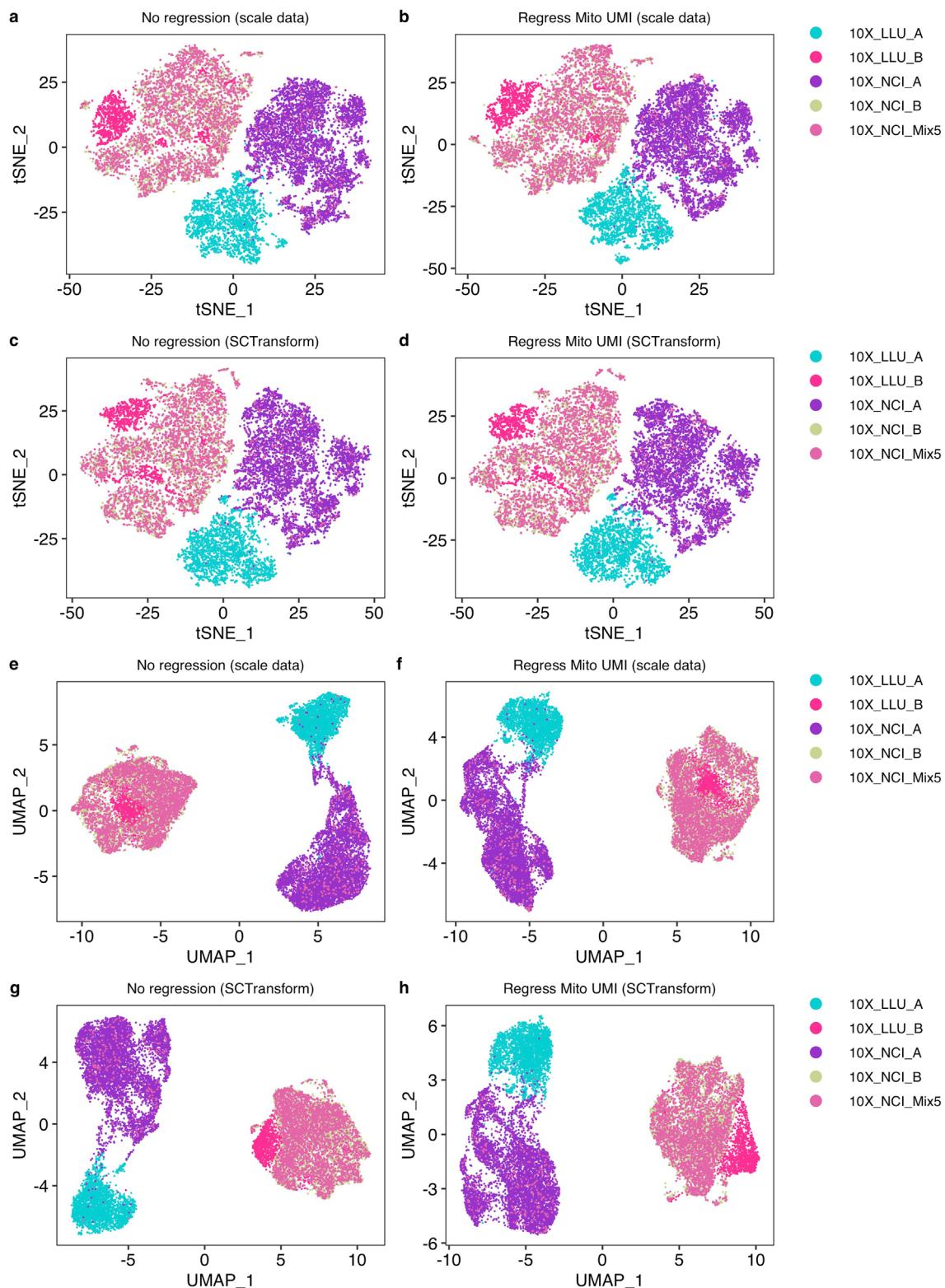
Supplementary information is available for this paper at <https://doi.org/10.1038/s41587-020-00748-9>.

Correspondence and requests for materials should be addressed to W.X. or C.W.

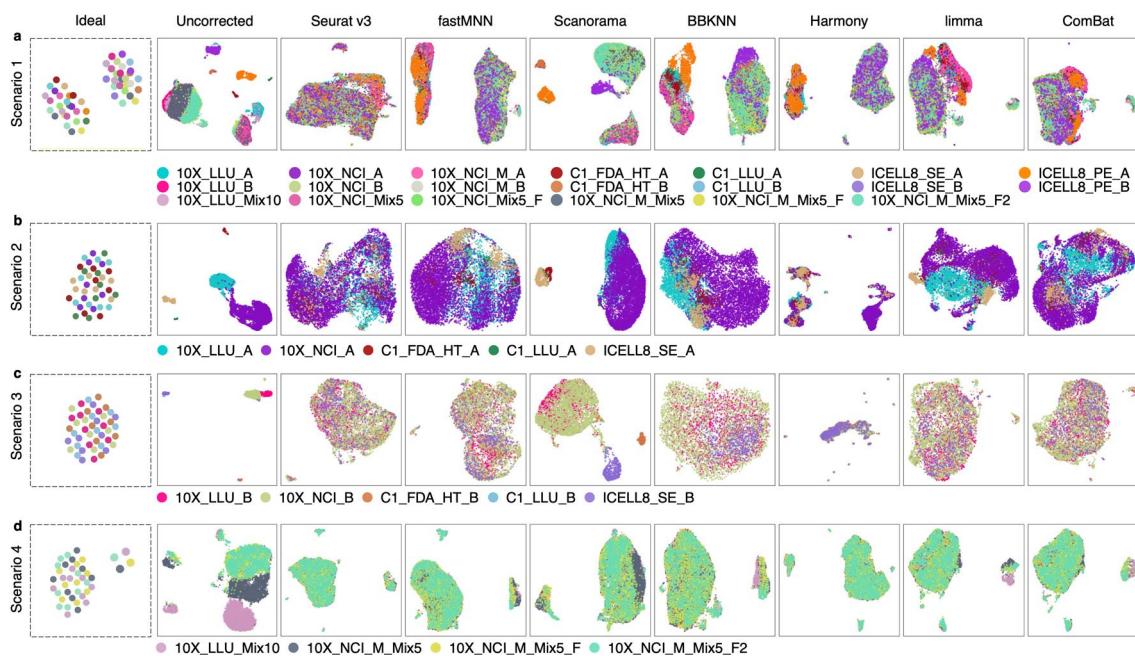
Reprints and permissions information is available at www.nature.com/reprints.



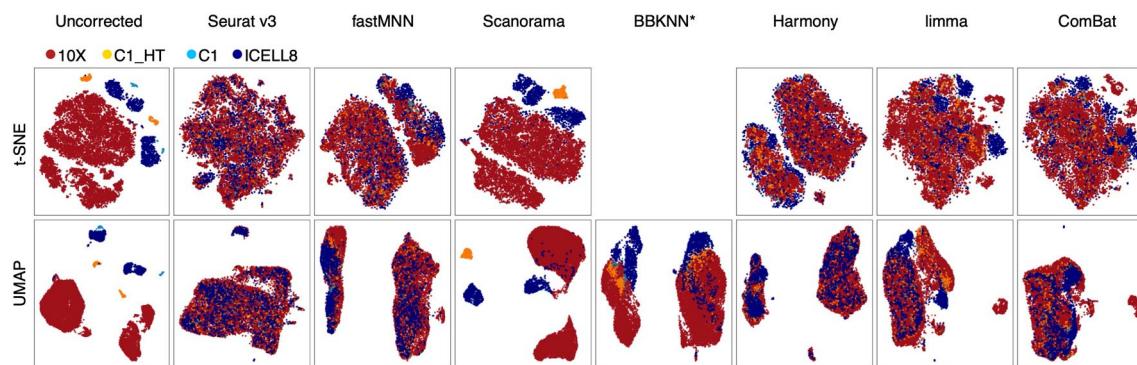
Extended Data Fig. 1 | An overview of the number of genes detected in each cell across all datasets. The violin plot shows the number of genes detected in each cell across 20 scRNA-seq datasets. The plot was generated using Seurat (version 3.1). Each dot represents a single cell. The violin shapes summarize 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 Cell Ranger version 3.1.



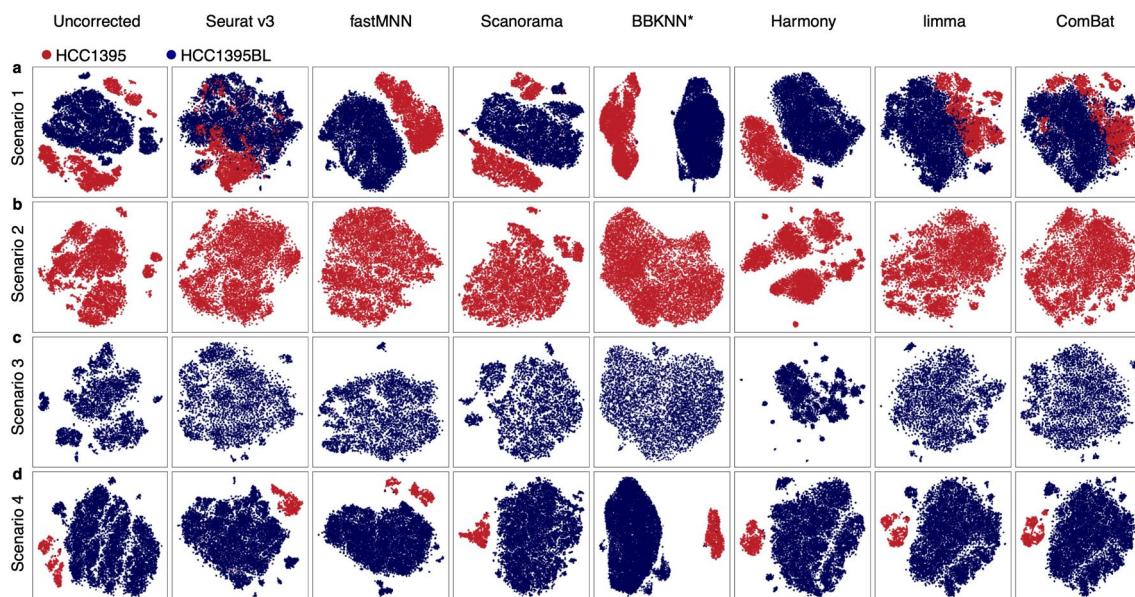
Extended Data Fig. 2 | Regressing mitochondrial genes and normalizing UMI did not remove batch effects. Five different batches of scRNA-seq data (10X_LLU_A, 10X_LLU_B, 10X_NCI_A, 10X_NCI_B, and 10X_NCI_Mix5) generated at two sites (LLU and NCI) are shown either as t-SNE plots (panels a-d) or as UMAPs (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.



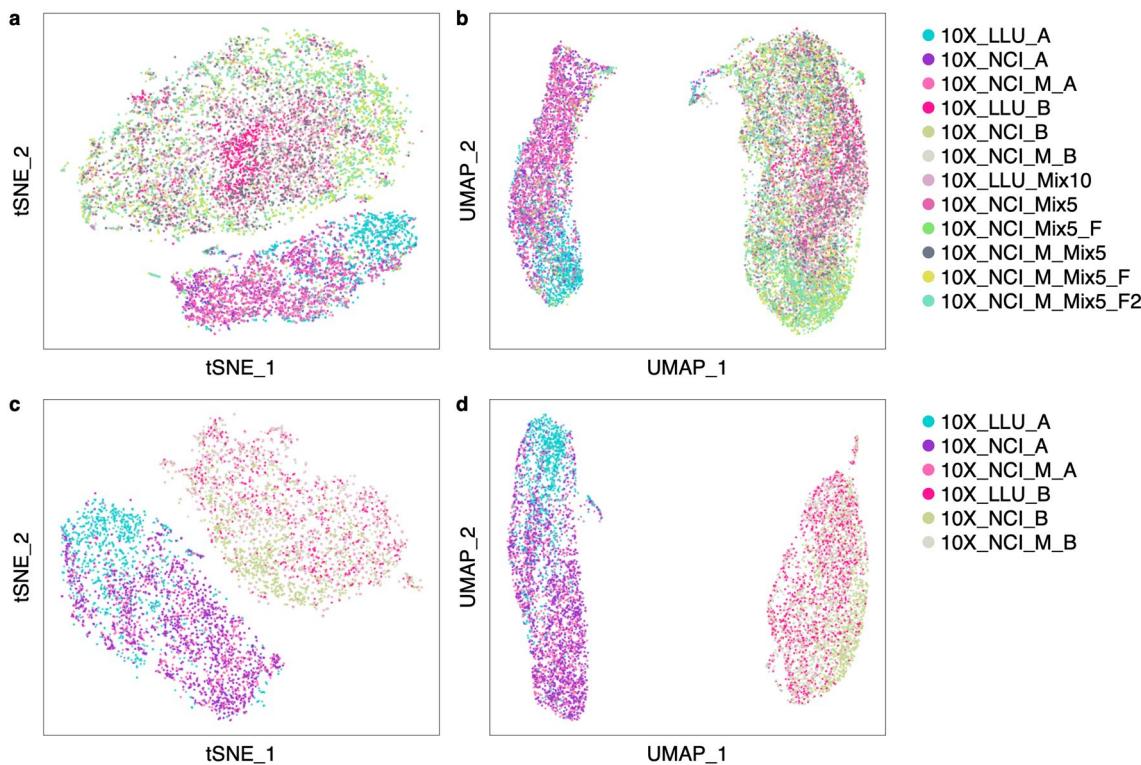
Extended Data Fig. 3 | UMAP showing batch-effect correction by mixability and clusterability using scRNA-seq datasets in four different sample scenarios. Batch-effect corrections were performed for the following four scenarios: (a) Scenario 1, where all 20 scRNA-seq datasets were combined, including 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 Cell Ranger version 3.1.



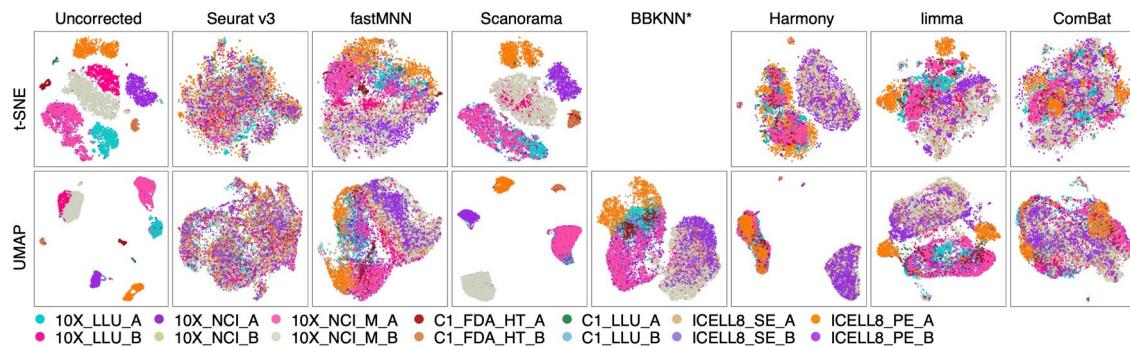
Extended Data Fig. 4 | t-SNE plots and UMAPs showing batch-effect corrections by mixability and clusterability across four scRNA-seq platforms.
t-SNE plots and UMAPs showing the batch-effect corrections performed by seven methods using 20 scRNA-seq datasets across different platforms. Datasets from 10x 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 Cell Ranger version 3.1.



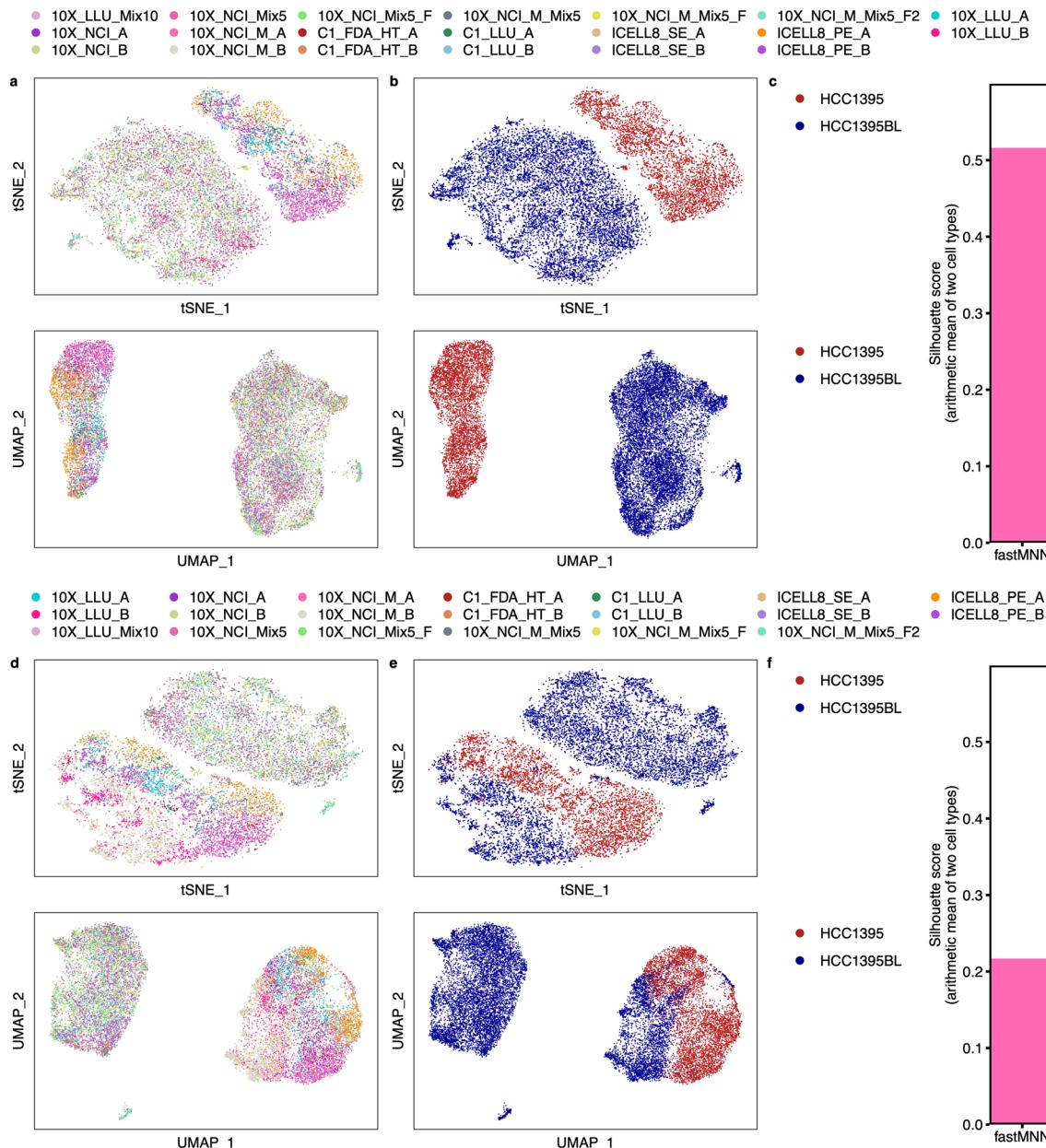
Extended Data Fig. 5 | Batch-effect correction displayed by cell type identity. Batch-effect corrections were performed for the following four scenarios: (a) Scenario 1, where all 20 scRNA-seq datasets were combined, including mixed and non-mixed, with large proportions of two dissimilar types of cells (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); all others are t-SNE plots. 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 Cell Ranger version 3.1.



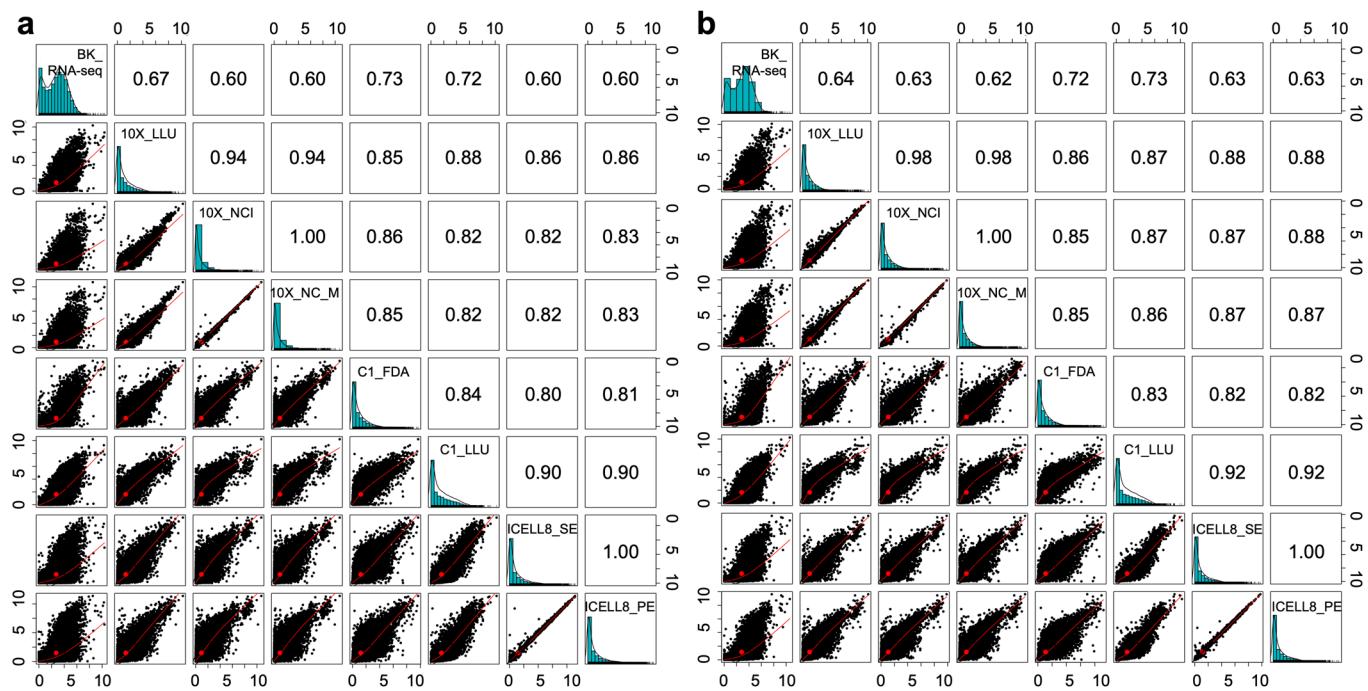
Extended Data Fig. 6 | Scanorama worked well for 10x Genomics scRNA-seq datasets regardless of the presence of shared cell types across batches.
(a) t-SNE plot and (b) UMAP showing batch-effect corrections using twelve 10x Genomics scRNA-seq datasets consisting of both mixed and non-mixed samples from 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 Cell Ranger version 3.1.



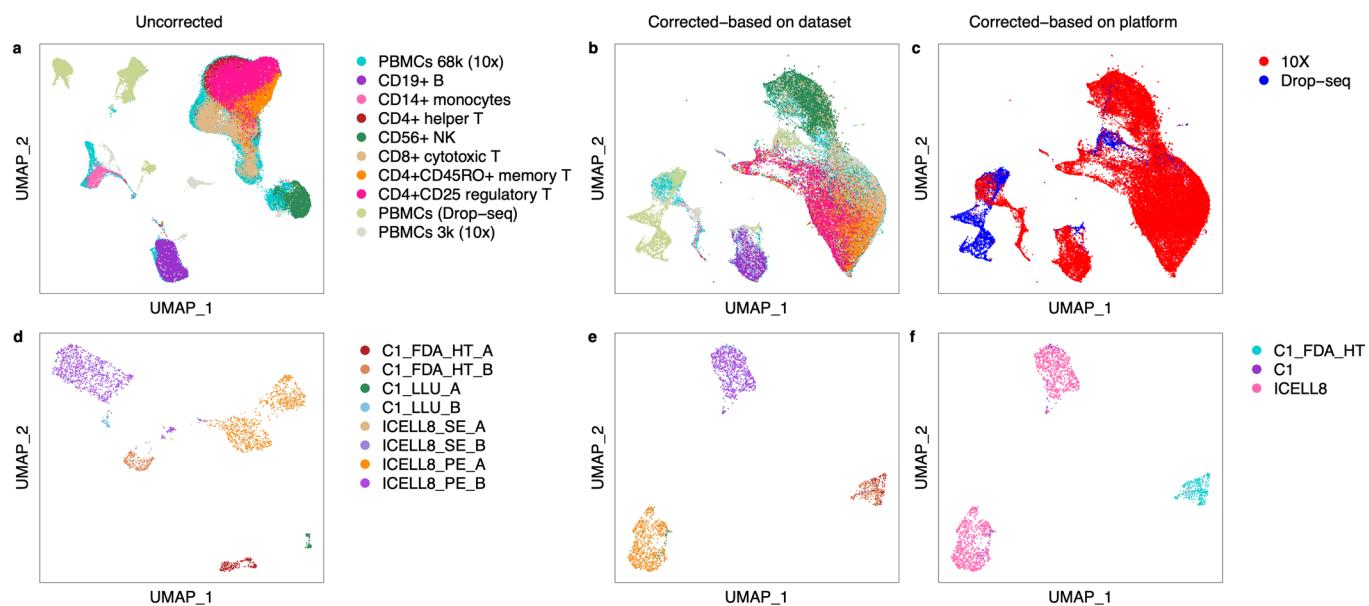
Extended Data Fig. 7 | Batch-effect correction evaluating clusterability using 14 scRNA-seq datasets without spiked-in mixtures. t-SNE plots and UMAPs showing batch-effect corrections performed by seven methods using 14 non-mixture scRNA-seq datasets across different platforms and sites. Six spiked-in 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 Cell Ranger version 3.1.



Extended Data Fig. 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 (that is, 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) t-SNE 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) t-SNE vs. UMAP with color-coding by datasets or (e) t-SNE 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 Fig. 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 [$\log(\text{CPM} + 1)$ 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 Fig. 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.

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection

We used many algorithms and code 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>

Data analysis

The details of methods used were described in the Methods section.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The datasets generated during and/or analyzed during the current study are available in the SRA repository with the access code # (PRJNA504037) and the data are already available to the public.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

- Life sciences Behavioural & social sciences Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample size was reported explicitly in the manuscript.
Data exclusions	Cells not passing the sequencing QC were excluded as described in the manuscript.
Replication	Our study design is very special and the details of the replications were described in the manuscript.
Randomization	N/A
Blinding	N/A

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems		Methods	
n/a	Involved in the study	n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies	<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input type="checkbox"/>	<input checked="" type="checkbox"/> Eukaryotic cell lines	<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology	<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms		
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants		
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data		

Eukaryotic cell lines

Policy information about [cell lines](#)

Cell line source(s)	Two cells were provided by ATCC (American Type Culture Collection)
Authentication	Per ATCC, we analyzed multiple cell specific markers for each cell line.
Mycoplasma contamination	No contamination found.
Commonly misidentified lines (See ICLAC register)	N/A