A Multi-center Cross-platform Single-cell RNA Sequencing Reference Dataset

Xin Chen^{1*}, Zhaowei Yang^{2,1*}, Wanqiu Chen^{1*}, Yongmei Zhao³, Andrew Farmer⁴, Bao Tran⁵,
Vyacheslav Furtak⁶, Malcolm Moos Jr.⁶, Wenming Xiao⁷, Charles Wang^{1#}

¹Center for Genomics, School of Medicine, Loma Linda University, Loma Linda, CA, 92350, USA.

²Department of Allergy and Clinical Immunology, State Key Laboratory of Respiratory Disease, Guangzhou Institute of Respiratory Health, the First Affiliated Hospital of Guangzhou Medical University, Guangzhou, Guangdong, 510182, P.R. China.

³CCR-SF Bioinformatics Group, Advanced Biomedical and Computational Sciences, Biomedical Informatics and Data Science Directorate, Frederick National Laboratory for Cancer Research, 8560 Progress Drive, Frederick, MD 21701.

⁴Takara Bio USA, Inc., Mountain View, CA 94043.

⁵Sequencing Facility, Frederick National Laboratory for Cancer Research, 8560 Progress Drive, Frederick, MD 21701.

⁶Center for Biologics Evaluation and Research & Division of Cellular and Gene Therapies, U.S. Food and Drug Administration, 10903 New Hampshire Avenue, Silver Spring, MD 20993.

⁷The Center for Devices and Radiological Health, U.S. Food and Drug Administration, FDA, Silver Spring, MD, 20993.

*Co-1st author and equal contribution.

#Correspondence shall be addressed to: Charles Wang at oxwang@gmail.com or chwang@llu.edu

Abstract

Single-cell RNA sequencing (scRNA-seq) is developing rapidly, and investigators seeking to use this technology are left with a variety of options for both experimental platform and bioinformatics methods. There is an urgent need for scRNA-seq reference datasets for benchmarking of different scRNA-seq platforms and bioinformatics methods. To be broadly applicable, these should be generated from renewable, well characterized reference samples and processed in multiple centers across different platforms. Here we present a benchmarking scRNA-seq dataset that includes 20 scRNA-seq datasets acquired either as a mixtures or as individual samples from two biologically distinct cell lines for which a large amount of multi-platform whole genome sequencing data are also available. These scRNA-seq datasets were generated from multiple popular platforms across four sequencing centers. Our benchmark datasets provide a resource that we believe will have great value for the single-cell community by serving as a reference dataset for evaluating various bioinformatics methods for scRNA-seq analyses, including but not limited to data preprocessing, imputation, normalization, clustering, batch correction, and differential analysis.

Background and summary

A variety of scRNA-seq technologies and protocols have been developed for biomedical research in the single-cell community¹⁻⁷. These technologies can be divided into two broad categories: full-length and 3' end counting-based. The 3' end counting-based methods allow the incorporation of unique molecular identifiers (UMIs) to improve quantification of mRNA molecules; whereas full-length methods generally provide greater sensitivity of gene detection and ability to identify changes across the length of a transcript, such as alternative splicing, novel transcripts, and mutations, etc. Large differences exist across different protocols in specificity, sensitivity, throughput, chemistry of library construction, bioinformatics⁸⁻¹², as well as cost. As described in

our associated NBT paper, prior studies have attempted to address various aspects relating to RNA-seq benchmarking^{9, 11, 12}. However, currently there is no systematic multi-center study that evaluates the influence of technology platform, sample composition, and bioinformatic methods (including preprocessing, normalization, and batch-effect correction) using publicly available standard reference samples and datasets consisting of both mixed and non-mixed biologically distinct samples.

Recently, we benchmarked scRNA-seq performance across several popular instrumentation platforms at multiple centers, also focusing on the effects of bioinformatic processing; including preprocessing, normalization, and batch-effect correction (see companion manuscript under *Nature Biotechnology* consideration, supplementary files 1-2). As stated in our associated NBT paper, our benchmark study has produced well-characterized reference materials (reference samples, datasets) and methods, which will have similar value and utility for the single-cell sequencing community as the Zook et al. study¹³, carried out by the Genome in a Bottle Consortium (GIAB), did for genome sequencing. The findings from our study offer practical guidance for optimizing and benchmarking a platform or experimental protocol, and for selecting appropriate bioinformatics methods when designing scRNA-seq experiments. We analyzed two well-characterized, but biologically distinct reference cell lines, i.e., a human breast cancer cell line (HCC1395; Sample A) and a B lymphocytes cell line (HCC1395BL; Sample B) derived from the same donor. A total of 20 scRNA-seq datasets derived from the two cell lines, processed either separately or as mixtures of different ratio of both cell lines, were generated using four scRNA-seq platforms (10X Genomics Chromium, Fluidigm C1, Fluidigm C1 HT, and Takara Bio's ICELL8 system) across four centers. We evaluated seven preprocessing pipelines for raw scRNA-seq fastq data, eight normalization methods¹⁴⁻¹⁹, and seven batch correction methods²⁰⁻²⁴. Our study showed that although pre-processing and normalization contributed to variability in gene detection and cell classification, batch effects were quite large, and the ability to assign cell types

correctly across platforms and sites was dependent on the bioinformatic pipelines, particularly the batch correction algorithms used. In many scenarios, Seurat v3²⁵, Harmony²⁴, BBKNN²³, and fastMNN²⁰ all corrected the batch effects fairly well for scRNA-seq data derived from either biologically identical or dissimilar samples across platforms and sites. However, when samples containing large fractions of biologically distinct cell types were compared, Seurat v3 over-corrected the batch-effect and misclassified the cell types (i.e., breast cancer cells and B lymphocytes clustered together), while limma and ComBat failed to remove batch effects. The datasets we present here can help researchers select the scRNA-seq protocol and bioinformatic method best suited to the samples to be analyzed. In addition, they can be used to benchmark current or newly developed scRNA-seq protocols and evaluate various existing and emerging bioinformatics methods for scRNA-seq data analysis.

Methods

Detailed methods were described in our associated paper (under NBT consideration). The following is a brief summary adapted from the Online Methods.

Study design

Fig. 1 shows our overall study design. A total of 20 scRNA-seq datasets were generated, including fourteen 3' end counting-based and six full-length datasets. The fourteen 3' end counting-based datasets were generated at four different centers and referred to as follows: 10X_LLU, 10X_NCI, 10X_NCI_M (modified shorter sequencing protocol), and C1_FDA_HT. The six full-length datasets were generated by two centers and referred to as: C1_LLU and ICELL8 (includes both single-end/SE and paired-end/PE). In the case of the 10X data sets, mixtures of samples A and B were processed in addition to individual sample processed separately. All other data sets were generated from samples A and B separately. For simplicity, we will use the labels in **Figure 2** to represent the 20 datasets throughout our analysis.

Cell culture

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

Full-length Single-cell RNA-seq using the C1 Fluidigm system

Single cell suspensions were loaded on a medium-sized (10-17 μ m) RNA-seq integrated fluidic circuit (IFC) at a concentration of 200 cells/ μ l. Full-length cDNAs were generated using the Fluidigm C1 system at the LLU Center for Genomics using the SMART-Seq v4 Ultra Low Input RNA kit (Takara Bio) according to the manufacturer's protocol. Libraries were prepared using the modified Illumina Nextera XT DNA library preparation protocol. 80 libraries were generated from HCC1395 cells (sample A) and 66 libraries were generated from HCC1395 BL cells (sample B). Library pools were sequenced at the LLU Center for Genomics on an Illumina HiSeq 4000 sequencer using 150x2 bp paired-end sequencing.

3' End Single-cell RNA-seq using C1 Fluidigm high-throughput (HT) system

High-throughput single cell 3' end cDNA libraries were generated according to the manufacturer's instructions at the FDA's Center for Biologics Evaluation and Research. Briefly, single cells were loaded on an HT IFC at a concentration of 400 cells/ μ l (Nexcelom Cellometer Auto T4). After cell lysis, the captured mRNA was barcoded during the reverse transcription step with a barcoded primer, and the tagmentation step was done following the Nextera XT DNA library preparation guide. Lastly, sequencing adapters and Nextera indices were applied during library preparation. Only the 3' end of the transcript was enriched following PCR amplification. 203 libraries were

generated from HCC1395 cells (sample A) and 241 libraries were generated from HCC1395 BL cells (sample B). Library pools were sequenced at the FDA/CBER Core Facility on an Illumina HiSeq 2500 using 75x2 bp, paired-end.

Single-cell RNA-seq using Takara Bio ICELL8 platform

A bulk cell suspension of either cancer or B cells was fluorescently labeled and diluted to ~1 cell in 35 nl. Each cell type solution was dispensed from a 384-well source plate into individually addressable wells in a 5,184 nano-well, 250 nl volume ICELL8 chip (SMARTer™ ICELL8® 250v Chip, Takara Bio USA, CA, USA) using a SMARTer™ ICELL8® Single-Cell System (Takara). Wells containing individual live cells were identified by imaging using CellSelect software to generate a well-selection map (filter file), which was then used to enable individual addressing of the chosen wells for addition of cDNA synthesis and library preparation reagents as detailed in the following sections. All on-chip liquid handling was performed with the SMARTer™ ICELL8® Single-Cell System. Full-length cDNA synthesis, P5/P7 index addition and tagmentation were done on-chip. Following round 1 PCR, amplicons were collected and pooled by centrifugation of the chip and purified using Ampure beads. This was followed by round 2 PCR amplification off-chip. All steps were performed per manufacturer's instructions. The library quality was determined using a Qubit fluorometer (Thermo Fisher), a 2100 Bioanalyzer and a corresponding High Sensitivity DNA Kit (Agilent). The ICELL scRNA-seq libraries were sequenced both at the Takara Bio USA site on an Illumina NextSeq 550 using 75x2 bp, paired-end and at the LLU Center for Genomics on a HiSeq 4000 using 150x1 bp, single-end sequencing.

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

After filtering with a 30-micron MACS SmartStrainer (Miltenyi Biotec), single cells were resuspended in PBS (calcium and magnesium free) containing 0.04% weight/volume BSA (400 µg/ml), and further diluted to 300 cells/µl after cell count (Countess II FL, Life Technologies). For

the 5% spike-in and 10% spike-in cell mixtures, 5% or 10% of HCC1395 breast cancer cells were mixed with either 95% or 90% of HCC1395BL cells. Library preparation was performed following the 3' scRNA-seq 10X Genomics platform protocol using v2 chemistry.

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 al²⁶. The fixed samples underwent two different treatments. For the first sample (spikein_5%_Fixed_1), the normal and tumor cells were harvested, washed, counted, and a 5% spike-in of breast cancer cells plus 95% normal B cells were prepared and mixed as described above. Approximately 130,000 cells were then processed for fixation. The cells were washed twice with 1X DPBS at 4 °C and resuspended gently in 100µl 1X DPBS (ThermoFisher Scientific). 900 µl chilled methanol (100%) was then added drop by drop to the cells with gentle vortexing. Cells were then fixed on ice for 15 min, following which they were stored at 4 °C for 6 days. For rehydration, the fixed cells were pelleted by centrifugation at 3000g for 10 min at 4 °C and washed twice with 1X DPBS containing 1% BSA and 0.4U/µl RNase inhibitor (Sigma Aldrich). The cells were then counted and the concentration was adjusted to be close to 1000 cells/µl. Approximately 8000 cells were loaded onto a single-cell chip for GEM generation using the 10X Genomics Chromium controller. 3' mRNA-seq gene expression libraries were prepared using the Chromium Single Cell 3' Library & Gel Bead Kit v2 (10X Genomics) according to the manufacturer's guidelines.

For the second sample (spikein_5%_Fixed_2), breast cancer cells and normal B cells (approximately 4 million each) were harvested and fixed as described above. The cells were initially washed with 1X DPBS and resuspended in 10% 1X DPBS and 90% chilled methanol, as described above. Cells were then fixed on ice for 15 mins, following which they were stored at 4 °C for 24 hrs. For rehydration, the fixed cells were washed with 1X DPBS containing 1% BSA and

0.4U/μl RNase inhibitor and counted. Approximately 8000 cells were loaded onto a single-cell chip for GEM generation using the 10X Genomics Chromium controller. 3'mRNA-seq gene expression libraries were prepared using the Chromium Single Cell 3' Library & Gel Bead Kit v2 (10X Genomics) according to the manufacturer's guidelines.

All the 10X Genomics scRNA-seq libraries constructed at the LLU were sequenced on the NextSeq 550 and HiSeq 4000 with the standard sequencing protocol of 26x98 bp read length at the LLU Center for Genomics, whereas the libraries constructed at the NCI site were either sequenced on the NextSeq 500 with a modified sequencing protocol of 26x57 bp read length at the NCI Sequencing Facility or on the HiSeq 4000 using the standard sequencing protocol of 26x98 bp read length at the LLU Center for Genomics.

Bulk cell RNA-seq

We isolated total RNA from bulk HCC1395 (cancer) and HCC1395 (B cells) using miRNeasy Mini kit (QIAGEN), and constructed RNA-seq libraries using the NuGEN Ovation universal RNA-seq kit at LLU according to the manufacturer's instructions. All the libraries were quantified using Qubit 3.0 (Life Technologies) and quality was checked on a TapeStation 2200 (Agilent Technologies). The bulk-cell RNA-seq libraries were sequenced both on a NextSeq 550, 75x2 bp paired-end; and on a HiSeq 4000, 100x2 bp paired-end at the LLU Center for Genomics.

Reference genome

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

Preprocessing of 10X scRNA-seq data

For UMI based 10X samples, four pre-processing pipelines, Cell Ranger¹ (v2.0.0), Cell Ranger (v3.1.0), UMI-tools²⁷ (v1.0.0), and zUMIs²⁸ (v2.4.5) were used to process the raw fastq data and generate gene count matrices. In the Cell Ranger pipeline, 'cellranger count' was used with all default parameter settings. In umi-tools and zUMIs pipelines, reads were filtered out if phred sequence quality of either the cell barcode or UMI bases were < 10. In UMI-tools, 'umi_tools whitelist' with default parameter settings was used to generate a list of cell barcodes for downstream analysis. 'umi_tools extract' was used to extract the cell barcodes and filter the reads (options: --quality-filter-threshold=10 --filter-cell-barcode). STAR²⁹ (v2.5.4b) was used for alignment to generate BAM files containing the unique mapped reads (option: outFilterMultimapNmax 1) for gene counting. featureCounts³⁰ (v1.6.1) was used to assign reads to genes and generate a BAM file (option: -R BAM). 'samtools (v1.3) 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 scRNA-seq data from C1 and Takara Bio ICELL8 platforms

For non-UMI based samples, three pre-processing pipelines were used to process the raw fastq data and generate gene count matrices. The pipelines included trimming and filtering, alignment, and gene counting. In the trimming and filtering process, one of the three tools [Trimmomatic³¹ (v0.35), trim_galore (v0.4.1), or cutadapt³² (v1.9.1)] was used to process the raw fastq data. Bases with quality less than 10 were trimmed from 5' and 3' ends of reads. Reads less 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, RSEM³³ (v1.3.0), and kallisto³⁴ (v0.43.1) were used to generate gene counts per cell. All default parameter settings were used except the following: In RSEM, option '--single-cell-prior' was used to estimate gene expression

levels for scRNA-seq data; Option of '--paired-end' was used if the data were paired-end fastqs; In kallisto, options '-l 500' and '-s 120' were used to represent estimated average fragment length and standard deviation of fragment length if the data were single-end fastqs. For simplicity, we will use featureCounts, RSEM, and kallisto to refer to the three pre-processing pipelines later.

Cell filtering and quality control metrics

We used gene count per cell matrices from Cell Ranger v3.1 and featureCounts pipelines in the downstream analyses for 10X and non-10X data, respectively. The following strategies were used to filter dead cells and doublets. 1) Cells were removed from analysis if they expressed less than 200 genes. We also removed genes expressed in less than 3 cells. 2) The total number of UMIs and genes for each cell were counted. The upper bound was calculated as mean plus two standard deviations (SD) and the lower bound as mean minus two SD for both the total UMIs and genes, respectively. Cells with total UMIs or genes outside of the upper and lower bounds were removed. 3) Cells were removed if greater than 10% reads mapped to mitochondrial genes.

scRNA-seq data batch effect correction

We used the filtered gene count per cell matrices from the Cell Ranger v3.1 and featureCounts pipelines as input to perform batch correction. Seurat (v3.0.3) based data processing was applied to each data set. The data sets were then log transformed and scaled. The top 2,000 highly variable genes (HVGs) were selected in each dataset with function *FindVariableGenes*. The processed data and HVGs were used as input to perform batch correction by Harmony, BBKNN, and Seurat V3. The Uniform Manifold Approximation and Projection (UMAP)³⁵ plots were generated from the batch-corrected low-dimensional embedding matrices. The uncorrected 20 scRNA-seq datasets showed strong batch effects in UMAP plots (**Fig. 2a**), clustering by individual dataset instead of by cell type. When the Harmony method was applied to the combined data, two clusters corresponding to the different cell types were clearly seen (**Fig. 2b**). BBKNN batch

correction also generated two clusters representing each cell type (**Fig. 2c**). However, Seuratv3 over-corrected and did not generate separate clusters for the two cell lines.

Estimation of copy number variation and clustering analysis

To estimate copy number variation (CNV) of Sample A cells (tumor cell line), we performed CNV analysis by inferCNV (<https://github.com/broadinstitute/inferCNV>, v1.1.3) on 13 datasets. These datasets included different batches of Sample A cells either alone or mixed with cells from Sample B. The 10X datasets were down sampled to 1,000 cells per dataset for the CNV analysis, which generated a total of 10,353 cells. In addition, the 10X_LLUI_B dataset was down sampled to 1,000 cells and used as a control. In the CNV analysis, a de-noising filtering step and a Hidden Markov Model (HMM)-based CNV prediction step were enabled with a cut-off of 0.1 for gene selection. For hierarchical clustering-based tree partitioning, parameters including 'tumor_subcluster_partition_method='qnorm'', 'hclust_method='ward.D2'' and 'tumor_subcluster_pval=0.05' were used. The CNV analysis showed clear separation by cell type differences instead of protocol differences. The results indicated the consistency of different protocols, which all captured similar CNVs in the tumor cells (**Fig. 3a**). Meanwhile, we applied Harmony batch correction and Seurat clustering to the same data and detected the two major clusters as well (**Fig. 3b**). The CNV analysis and expression-based clustering analysis were highly consistent (**Fig. 3c**).

Data Records

All sequence data have been deposited in the National Center for Biotechnology Information (NCBI) Sequence Reads Archive (SRA) with accession ID: PRJNA504037³⁶. The processed gene count matrices have been uploaded to Figshare³⁷.

Technical Validation

QC assessment of the effect of preprocess pipelines for 10X and none 10X data

For the 10X scRNA-seq data, we evaluated four pipelines to pre-process the data: Cell Ranger 2.0, Cell Ranger 3.1 (10X Genomics), UMI-tools, and zUMIs, and examined the consistency between the four pipelines regarding the number of cells identified, the number of genes detected per cell, and percentage of Sample A cells in spike-in mixtures (**Fig. 4a-c**). In most of the datasets (except 10X_LLUI_A), Cell Ranger 3.1 and zUMIs always called the largest and second largest number of cells. For most of the datasets from Sample A only or Sample B only, Cell Ranger 2.0 and UMI-tools called consistent numbers of cells. For spike-in datasets, especially for cells fixed in methanol, the numbers of cells called were variable. The percentages of Sample A cells in **Fig. 4c** were also inconsistent in the samples with fixed cells. These observations suggest that methanol fixation may reduce the quality of data, causing inconsistent cell calling by different pipelines, which may further affect downstream analysis. For the 10X single-cell protocol, we compared the standard sequencing protocol (98bp) with the modified sequencing protocol (57bp) using the same scRNA-seq libraries. The two protocols yielded consistent cell calling and percentage of spike-in A cells, suggesting that sequence length and sequencing instrument do not significantly impact the data quality. **Fig. 4b** shows the number of genes expressed per cell. Because Cell Ranger 3.1 modified the algorithm to call more cells with low RNA content, it always generated a lower number of genes expressed per cell because it called the largest number of cells compared with other pipelines.

For the non-droplet scRNA-seq data, consistent numbers of genes expressed per cell were observed using the three different pre-processing pipelines (**Fig. 4d**). Kallisto identified more genes per cell in the full-length protocols (C1_LLUI and ICELL8) and fewer genes per cell in the 3'-end counting-based protocol (C1-FDA_HT). For the ICELL8 protocol, we compared the PE

samples with SE samples using the same scRNA-seq libraries with similar results; sequence length and sequencing instrument did not affect the quality of the same library significantly.

Cell quality assessment across 20 scRNA-seq datasets

We used three metrics: 1) average number of genes per cell, 2) average number of UMIs/read counts per cell, 3) average mitochondrial gene percentage per cell to benchmark the consistency of the 20 scRNA-seq datasets by comparing high quality cells with filtered low quality cells (**Fig. 5a-c**, see cell filtering in method section for details). In high quality cells, the average number of genes detected per cell was greater than 3,000 genes; only the C1 platform (C1_LLUI, C1_FDA_HT) detected genes above this value in low quality cells (**Fig. 5a**). This may be due to small numbers of cells captured with the C1 platform, leading to high read counts per cell. Samples analyzed using the same method showed similar average numbers of genes per cell and average numbers of UMIs/read counts per cell and were clustered together (**Fig. 5a-b**). Due to the sequencing depth differences between full-length and 3' end counting-based methods, C1_LLUI and ICELL8 showed higher values of metrics 1 and 2 in high quality cells than those in 10X and C1_FDA_HT. Furthermore, we found that no more than 3.85% of the cells had mitochondrial gene percentages greater than 10% across the combined 20 datasets. In high quality cells, the mitochondrial gene percentages were below 10% in all 20 datasets. However, only 7 datasets (35%) had less than 10% mitochondrial genes in the cells with low quality (**Fig. 5c**). These results showed increased quality of the cells when cell filtering was applied to the raw data.

Consistency of gene expression across 20 scRNA-seq datasets

For the 20 scRNA-seq data sets, we performed correlation analyses to evaluate the consistency of gene expression data using bulk RNA-seq data as a reference. We first labelled cells in the spike-in datasets as Sample A and Sample B cells, respectively. The data subsets containing

either A or B cells in the spike-in datasets were then used to perform correlation analyses (a total of 13 datasets for A and B cells, respectively). The top 2,000 HVGs of the 13 scRNA-seq datasets were used in the analyses. The gene expression mean and variance were calculated across all cells in each scRNA-seq dataset and across three replicates in bulk RNA-seq dataset. The results showed that bulk RNA-seq data sets had good correlation ($r \geq 0.78$) of gene expression mean with all 13 datasets in both A and B cells (**Fig. 6a, d**). The correlation of gene expression variance between bulk RNA-seq dataset and scRNA-seq datasets, suggesting cell-to-cell diversity in scRNA-seq (**Fig. 6b, e**). High consistency was observed across the 10X datasets and non-10X datasets in both gene expression mean and variance. C1_LLU datasets showed the highest correlation of gene expression mean with bulk RNA-seq data in both A and B cells, likely because this dataset had the highest sequencing depth (An average of over 4 million reads per cell). We also found that samples with standard sequencing and modified sequencing protocols were highly correlated ($r \approx 1$). The results showed high consistency of gene expression across 20 scRNA-seq datasets.

Code Availability

All code used in processing the scRNA-seq data and in drawing the figures are available on Github at the following link: https://github.com/oxwang/SciData_scRNAseq

Acknowledgements

The authors would like to thank Ms. Diana Ho of the LLU Center for Genomics for her great administrative support, particularly in coordinating the weekly Zoom conference calls and assistance in preparation of meeting minutes for the FDA SEQC-2 single-cell sequencing project. The authors would like to thank Dr. Zhong Chen at LLU and Wells Wu of the FDA/CBER core facility. The genomic work carried out at the LLU Center for Genomics was funded in part by the

National Institutes of Health (NIH) grant S10OD019960 (CW), the Ardmore Institute of Health grant 2150141 (CW) and Dr. Charles A. Sims' gift to LLU Center for Genomics.

Conflict of interests and disclaimer: Andrew Farmer is an employee of Takara Bio USA, Inc. 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.

Author contributions

CW conceived and designed the study. XC, ZY and WC drafted the manuscript. MMJ, AF, WX and CW edited the manuscript. WC, AF, BT, MMJ, VF performed single cell captures, scRNA-seq library construction and sequencing. XC and ZWY performed bioinformatics data analyses. All authors reviewed the manuscript. CW finalized and submitted the manuscript.

Figure legends

Figure 1. Experimental design.

Figure 2. UMAP projections before (a) and after batch correction using (b) Harmony, (c) BBKNN, and (d) Seurat v3.

Figure 3. InferCNV analysis compared with expression-based clustering. (a) Estimation of copy number variants by inferCNV in 7 data sets containing tumor cells (HCC1395) and 6 spike-in data sets containing both HCC1395 and HCC1395BL (B lymphocytes). 10X_LLUI_B was used as a control (heatmap not shown). All data sets were down sampled to 1,000 cells. Top color bar indicates different chromosome regions. Left color bar indicates different data sets. The color intensities of the heatmap correspond to the residual expression values by inferCNV, with red or blue indicating higher or lower values compared with those of control cells. (b) UMAP projection of Harmony-corrected expression data. Dark red and blue indicate the true identity of HCC1395 and HCC1395BL, respectively. (c) A dendrogram of inferCNV clusters (left) and heatmap comparison (right) of cell labels generated by inferCNV and Seurat. The order of dendrogram leaves was generated by inferCNV as indicated in panel (a). Heatmap colors indicate the top two groups of cells in the dendrogram tree and cell clusters identified by Seurat.

Figure 4. Effect of preprocessing pipelines. (a) Barplot of the number of cells identified, (b) Boxplot of the number of genes detected per cell, and (c) Barplot of percentage of breast cancer cells detected in spike-in mixtures processed by the four pipelines using the droplet (10X Genomics) pre-processing algorithms. (d) Boxplot of the number of expressed genes per cell processed by three pipelines on non-droplet pre-processing algorithms.

Figure 5. Evaluation of cell quality. (a) Average number of genes detected per cell, (b) Average number of UMIs/read counts (droplet/non-droplet) per cell, (c) Average mitochondrial gene percentage per cell. Auxiliary lines represented by grey dash lines were applied for (a)-(c) and several datasets are labeled for better visualization. # indicates the number of genes or read counts. % indicates percentage of mitochondrial genes.

Figure 6. Platform consistency. Pairwise Pearson correlation of the expression mean and variance of the top 2,000 highly variable genes (HVG) between bulk RNA-seq and scRNA-seq datasets for (a-b) HCC1395 and (d-e) HCC1395BL. (c) and (f) represent the cell composition of each scRNA-seq dataset. The HVG expression of bulk RNA-seq represents the average gene expression across 3 biological replicates. The HVG expression of scRNA-seq represents the average gene expression across all cells in each dataset.

Reference

- 1 Zheng, G. X. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nature communications* **8**, 1-12 (2017).
- 2 Hashimshony, T. *et al.* CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome biology* **17**, 77 (2016).
- 3 Macosko, E. Z. *et al.* Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202-1214 (2015).
- 4 Gao, R. *et al.* Nanogrid single-nucleus RNA sequencing reveals phenotypic diversity in breast cancer. *Nature communications* **8**, 1-12 (2017).
- 5 Hashimshony, T., Wagner, F., Sher, N. & Yanai, I. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell reports* **2**, 666-673 (2012).
- 6 Ramsköld, D. *et al.* Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nature biotechnology* **30**, 777 (2012).
- 7 Picelli, S. *et al.* Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nature methods* **10**, 1096-1098 (2013).
- 8 Ziegenhain, C. *et al.* Comparative analysis of single-cell RNA sequencing methods. *Molecular cell* **65**, 631-643. e634 (2017).
- 9 Tian, L. *et al.* Benchmarking single cell RNA-sequencing analysis pipelines using mixture control experiments. *Nature methods* **16**, 479-487 (2019).
- 10 Zhang, X. *et al.* Comparative analysis of droplet-based ultra-high-throughput single-cell RNA-seq systems. *Molecular cell* **73**, 130-142. e135 (2019).
- 11 Tran, H. T. N. *et al.* A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome biology* **21**, 1-32 (2020).
- 12 Mereu, E. *et al.* Benchmarking single-cell RNA-sequencing protocols for cell atlas projects. *Nature Biotechnology*, 1-9 (2020).
- 13 Zook, J. M. *et al.* Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nature biotechnology* **32**, 246-251 (2014).
- 14 Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome biology* **11**, R25 (2010).
- 15 Risso, D., Ngai, J., Speed, T. P. & Dudoit, S. Normalization of RNA-seq data using factor analysis of control genes or samples. *Nature biotechnology* **32**, 896 (2014).
- 16 Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biology* **20**, 1-15 (2019).
- 17 Lun, A. T., Bach, K. & Marioni, J. C. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome biology* **17**, 75 (2016).
- 18 Bacher, R. *et al.* SCnorm: robust normalization of single-cell RNA-seq data. *Nature methods* **14**, 584 (2017).
- 19 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 research* **45**, e179-e179 (2017).
- 20 Haghverdi, L., Lun, A. T., Morgan, M. D. & Marioni, J. C. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nature biotechnology* **36**, 421-427 (2018).
- 21 Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature biotechnology* **36**, 411-420 (2018).
- 22 Hie, B., Bryson, B. & Berger, B. Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. *Nature biotechnology* **37**, 685-691 (2019).

- 23 Polański, K. *et al.* BBKNN: fast batch alignment of single cell transcriptomes. *Bioinformatics* **36**, 964-965 (2020).
- 24 Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature methods*, 1-8 (2019).
- 25 Stuart, T. *et al.* Comprehensive integration of single-cell data. *Cell* **177**, 1888-1902. e1821 (2019).
- 26 Alles, J. *et al.* Cell fixation and preservation for droplet-based single-cell transcriptomics. *BMC Biol* **15**, 44, doi:10.1186/s12915-017-0383-5 (2017).
- 27 Smith, T., Heger, A. & Sudbery, I. UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome research* **27**, 491-499 (2017).
- 28 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).
- 29 Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2013).
- 30 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).
- 31 Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120 (2014).
- 32 Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal* **17**, 10-12 (2011).
- 33 Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC bioinformatics* **12**, 323 (2011).
- 34 Weijers, S. *et al.* KALLISTO: cost effective and integrated optimization of the urban wastewater system Eindhoven. *Water Practice and Technology* **7** (2012).
- 35 McInnes, L., Healy, J. & Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).
- 36 Yang, Z. & Wang, C. NCBI. Sequence Read Archive <https://www.ncbi.nlm.nih.gov/bioproject/PRJNA504037> (2020).
- 37 Chen, X. *et al.* Gene count matrices from a scRNA-seq benchmark study. *figshare* <https://figshare.com/s/7eaff863913678a4166d> (2020).

Derived from same donor

A (HCC1395)
Breast Cancer Cell Line

B (HCC1395BL)
Normal B Lymphocyte Cell Line

Multi-platform

20 datasets

- 10X_LLU_A
- 10X_LLU_B
- 10X_LLU_Mix10
- 10X_NCI_A
- 10X_NCI_B
- 10X_NCI_M_A
- 10X_NCI_M_B
- 10X_NCI_M_Mix5
- 10X_NCI_M_Mix5_F
- 10X_NCI_M_Mix5_F2
- 10X_NCI_Mix5
- 10X_NCI_Mix5_F
- C1_FDA_HT_A
- C1_FDA_HT_B
- C1_LLU_A
- C1_LLU_B
- ICELL8_PE_A
- ICELL8_PE_B
- ICELL8_SE_A
- ICELL8_SE_B

a **b** **c**

a Correlation between expression mean of HVG (A, HCC1395) **b** Correlation between expression variance of HVG (A, HCC1395) **c** HCC1395 proportion in 13 data sets **d** Correlation between expression mean of HVG (B, HCC1395BL) **e** Correlation between expression variance of HVG (B, HCC1395BL) **f** HCC1395BL proportion in 13 data sets