Journal of Biomolecular Techniques •

Comparative Analysis of Single-Cell RNA Sequencing Platforms and Methods

John M. Ashton[1], Hubert Rehrauer[2], Jason Myers[1], Jacqueline Myers[1], Michelle Zanche[1], Malene Balys[1], Jonathan Foox[3], Chistopher E. Mason[3], Robert Steen[4], Marcy Kuentzel[5], Catharine Aquino[2], Natàlia Garcia-Reyero[5], Sridar V. Chittur[6]

[1]University of Rochester Medical Center, University of Rochester, West Henrietta, New York 14642, USA,

[2]Functional Genomics Center Zurich, ETH and University of Zurich, CH-8057 Zurich, Switzerland,

[3]Department of Physiology and Biophysics, Weill Cornell Medicine, New York, NY 10065, USA,

[4]Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA,

[5]Environmental Laboratory, US Army Engineer Research and Development Center, Vicksburg, Mississippi 39180, USA,

[6]Center for Functional Genomics, University at Albany-SUNY, Rensselaer, New York 12144, USA

Published on: Dec 06, 2021

ABSTRACT

Single-cell RNA sequencing (scRNA-seq) offers great new opportunities for increasing our understanding of complex biological processes. In particular, development of an accurate Human Cell Atlas is largely dependent on the rapidly advancing technologies and molecular chemistries employed in scRNA-seq. These advances have already allowed an increase in throughput for scRNA-seq from 96 to 80,000 cells on a single instrument run by capturing cells within nanoliter droplets. Although this increase in throughput is critical for many experimental questions, a thorough comparison between microfluidic-based, plate-based, and droplet-based technologies or between multiple available platforms utilizing these technologies is largely lacking. Here, we report scRNA-seq data from SUM149PT cells treated with the histone deacetylase inhibitor trichostatin A versus untreated controls across several scRNA-seq platforms (Fluidigm C1, WaferGen iCell8, 10x Genomics Chromium Controller, and Illumina/BioRad ddSEQ). The primary goal of this project was to demonstrate RNA sequencing methods for profiling the ultra-low amounts of RNA present in individual cells, and this report discusses the results of the study, as well as technical challenges and lessons learned and present general guidelines for best practices in sample preparation and analysis.

ADDRESS CORRESPONDENCE TO: John M. Ashton, University of Rochester Medical Center, University of Rochester, West Henrietta, NY 14642, USA (E-mail: John_ashton@urmc.rochester.edu).

ADDRESS CORRESPONDENCE TO: Hubert Rehrauer, Functional Genomics Center Zurich, ETH and University of Zurich, CH-8057 Zurich, Switzerland (E-mail: Hubert.Rehrauer@fgcz.ethz.ch)

Conflict of Interest Disclosures: The authors declare no conflicts of interest.

Keywords: platforms, RNA-seq, single cell

INTRODUCTION

Cells are the most fundamental building blocks of all living organisms. From a unicellular bacteria to complex living forms composed of many different cell types, the transcriptional signatures of individual cells reflect the physiological and pathological state of a biological being. Although analyses at the single-cell level have long been practiced in biology and medicine through microscopic observations, protein

immunohistochemistry, *in situ* hybridization, and flow cytometry, the ability to perform whole transcriptome profiling [RNA sequencing (RNA-seq)] by next generation sequencing (NGS) at the resolution of single cells is a relatively recent and, due to its ability to do parallel profiling of thousands of cells, very powerful approach. These new techniques enable novel discoveries and provide great insights into cell identity, cell function, cellular composition of different organs, and cell origin, evolution, and heterogeneity in many different cancer types.[1]

In the past decade, different strategies have been explored for transcriptome profiling of single cells. Early on, single-cell analyses by NGS were carried out using cell sorting to plate single cells into individual wells, which was followed by cell lysis, cDNA synthesis, barcoded library generation, pooling, and sequencing.[2] This method allows for individual cell assessment, such as viability and individual cell capture, prior to downstream processing. However, it has low throughput and is labor intensive and expensive. Utilizing integrated fluidic circuits, Fluidigm C1 autoprep system isolates single cells into individual nanochannels for visual examination, which is followed by cell lysis, cDNA conversion, preamplification, and retrieval for library construction and sequencing.[3] The C1 system significantly simplified the individual cell isolation while still enabling whole transcript sequencing analysis; however, cell partitioning is size restricted based on the nanochannel tolerance of the nanofluidic plate. DropSeq technology encapsulates single cells of different sizes in an oil-based droplet and barcoded beads attached to unique oligomers in individual aqueous droplets for gene expression profiling of many hundreds to thousands of cells in parallel.[4] Although DropSeq is highly efficient with commercially available platforms that are relatively easy to operate, only 5′- or 3′-tag profiling is possible at present, thereby not allowing for full-length transcript analysis. Furthermore, no intermediate assessment of cell-capture quality and quantity is possible until after the completion of the NGS sequencing. The Icell8 platform, in contrast, isolates 1000–1800 cells in a 5184-nanowell chip to allow for analysis of cell capture and viability.[5] Gene expression profiling by RNA-seq can be done for both 3′ profiling and full-length transcriptome. Currently, the most commonly employed microfluidics-based platform is the single-cell Chromium controller from 10x Genomics.[6] The 10x protocol is based on a 5′- or 3′-tag sequencing method. The Illumina/BioRad ddSEQ uses disposable microfluidic cartridges to coencapsulate single cells and barcodes into subnanoliter droplets. Cell lysis and barcoding occur in the droplets, and libraries can then be subsequently prepared and sequenced.

Considering each platform employs different strategies for single-cell transcriptome profiling, the capacity, sensitivity, and reproducibility of each approach can differ greatly.[7] To better understand those strategies, the Association of Biomolecular Resource Facilities Genomics Research Group developed a study to compare different single-cell RNA sequencing (scRNA-seq) platforms and methods. In order to reduce sample heterogeneity associated with dissociated cells from a given tissue type and to enable direct comparison of the platforms, a single cancer cell line was selected, SUM149 P/T, and gene expression signature changes were tested after cells were treated with or without a histone deacetylase inhibitor, trichostatin A (TSA) for this multiplatform comparison study.[8] Bulk RNA-seq data were used as reference to assess the performance of 5 platforms, including Fluidigm C1 and HT, 10x Chromium, BioRad ddSEQ, and ICELL8.

MATERIALS AND METHODS

Cell culture and drug treatment

SUM149 P/T cells (gift from Dr. Martin Tenniswood, SUNY Albany) were grown at the University of Rochester Medical Center in Ham's F-12 medium (Invitrogen) supplemented with 5% fetal bovine serum (Sigma-Aldrich), 5 µg/mL insulin (Sigma-Aldrich), 1 µg/mL hydrocortisone (Sigma-Aldrich), 10 mM 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (Fisher Scientific, Pittsburgh, PA, USA), and antimycotic/antibiotic (Sigma-Aldrich). After the cells attached to the culture dishes, cells were grown and maintained in the above media with no fetal bovine serum. Cells were maintained at 37°C in a humidified atmosphere of 95% air/5% CO_2. SUM149PT cells were plated at a density of 1.5×10^6 cells per 150-cm^2 dish for 48 h prior to treatment with 10 nM TSA (in DMSO) or an equivalent volume of DMSO. After 48 h of treatment, cells were harvested by trypsinization, washed with PBS, and shipped overnight in media for the different laboratories to perform the scRNA-seq experiments.

Fluidigm C1 96 scRNA-seq

Single SUM149 P/T cells treated with vehicle or TSA were captured on a 10–17-µm cell diameter integrated microfluidic chip (IFC) using the Fluidigm C1 Single-cell AutoPrep system (Fluidigm Corporation). Cells were prestained with Calcien AM/EthD-1 LIVE/DEAD cell viability assay (Life Technologies, Carlsbad, CA, USA) and loaded onto the IFC at a concentration of 500–700 cells/µL. Viable single-cell confirmation was performed with phase-contrast fluorescence microscopy to assess the number and

viability of cells per capture site. Only single, live cells were included in the analysis. For RNA-seq analysis, cDNAs were prepared “on-IFC” using the SMARTer Ultra Low RNA kit for Illumina (Clontech, Mountain View, CA, USA) following Fluidigm recommendations. Single-cell cDNA size distribution and concentration was assessed with PicoGreen (Life Technologies) and Agilent Bioanalyzer 2100 analysis (Agilent Technologies, Santa Clara, CA, USA). Illumina libraries were constructed in 96-well plates using Illumina’s NexteraXT DNA Sample Preparation kit following the protocol supplied by Fluidigm. For each C1 experiment, a bulk RNA control and negative control were processed in parallel, using the same reagent mixes as used on the chip. Libraries were quantified by Agilent Bioanalyzer, using the High Sensitivity DNA analysis kit, and were also quantified fluorometrically, using Qubit dsDNA HS Assay kits and a Qubit 2.0 Fluorometer (Life Technologies). Single-end reads of 100 nucleotiedes (nt) were generated for each sample using Illumina’s HiSeq2500v4.

Fluidigm C1 HT scRNA-seq

Single SUM149 P/T cells treated with vehicle or TSA were captured on a high-throughput 10–17-μm cell diameter integrated microfluidic chip (IFC) using the Fluidigm C1 Single-cell AutoPrep system (Fluidigm Corporation). Cells were prestained with Calcien AM/EthD-1 LIVE/DEAD cell viability assay (Life Technologies) and loaded onto the IFC at a concentration of 400 cells/μL. Viable single-cell confirmation was performed with phase-contrast fluorescence microscopy to assess the number and viability of cells across 10% of the capture wells. For RNA-seq analysis, cDNAs were prepared “on-IFC” using the SMARTer Ultra Low RNA kit for Illumina (Clontech) following Fluidigm recommendations. Single-cell cDNA size distribution and concentration were assessed with PicoGreen (Life Technologies) and Agilent Bioanalyzer 2100 analysis. Illumina libraries were constructed in 96-well plates using Illumina’s NexteraXT DNA Sample Preparation kit following the protocol supplied by Fluidigm. For each C1 experiment, a bulk RNA control and a negative control were processed in parallel, using the same reagent mixes as used on the chip. Libraries were quantified by Agilent Bioanalyzer, using the High Sensitivity DNA analysis kit, and were also quantified fluorometrically, using Qubit dsDNA HS Assay kits and a Qubit 2.0 Fluorometer (Life Technologies). Single-end reads of 100 nt were generated for each sample using Illumina’s HiSeq2500v4.

WaferGen iCell8 scRNA-seq

Cells were stained with Hoechst 33324 and Propidium Iodide (Thermo Fisher Scientific) for 20 min. The cell viability and density were checked with Moxi Cell

counter (VWR International, LLC.), and cells were diluted to achieve a density of 1 cell per 50 nL in a final dispensing mix, which contained a diluent, Murine RNase inhibitor (New England Biolabs), and 0.35X PBS (without Ca^{++} and Mg^{++}, pH 7.4, Thermo Fisher Scientific). A 384-well source plate with 8 designated wells containing cell suspensions, 1 well for positive control, 1 well for negative control, and 1 well for fiducial mix (fluorescent dye permitting image alignment confirmation) was placed in the MultiSample NanoDispenser (MSND; WaferGen). Each of the 8 sample source wells in the 384-source plate was sampled by 1 of the 8 channels in MSND. Cells, positive controls, negative controls, and fiducial mix were dispensed onto 1 chip within 16 min. Each well received 50 nL of cell mix, positive control, negative control, or fiducial mix. Two ICELL8 chips were used to collect single cells, after which the chips were sealed and centrifuged at 300 *g* for 5 min at room temperature. The chips were subsequently imaged using the imaging station and frozen at −80°C overnight. Microchip images were analyzed using CellSelect software (WaferGen) to determine the viability and number of cells present in each nanowell. From each chip, we selected 425 TSA-treated single cells, 425 DMSO-treated single cells, 4 positive controls, and 4 negative controls.

Microchips were removed from −80°C and left at room temperature for 10 min. Cells were lysed by freeze-thaw at this step. Cells were lazed by freeze-thaw at this step and microchips centrifuged at 3800 *g* for 5 min at 4°C, transferred to a thermocycler with a program of 72°C for 3 min followed by 4°C hold to anneal preprinted oligonucleotides to poly(A) mRNAs. The microchips were centrifuged as previously described and were placed back into the MSND. A separate 384-well source plate containing reverse-transcription (RT) reagents [GC melting reagent (5 M), deoxyribonucleotide triphosphate Mix (25 mM each), MgCl2 (1 M), DTT (100 mM), 5X First-Strand Buffer, Triton X-100 (10%), SMARTer ICELL8 3′ DE Oligo Mix, SMARTScribe Reverse Transcriptase (100 U/μL), and SeqAmp DNA Polymerase] in 4 wells was used in the MSND, which delivered 50 nL of RT mix to selected nanowells. The microchips were spun down and transferred to a chip cycler to perform RT-PCR using the preinstalled program (500°C for 3 min, 4°C for 5 min, 42°C for 90 min, and 2 cycles of 50°C for 2min and 42^{o}°C for 2 min, which was followed by heating at 70°C for 15 min, 95°C for 1 min, and 24 cycles of 98°C for 10 s, 65°C for 30 s, 68°C for 3 min, and finally 72°C for 10 min, which was followed by ramping down to a 4°C hold).

After the reaction, chips were inverted and centrifuged (3800 *g* for 10 min at 4°C) to simultaneously collect and pool well contents into a single microcentrifuge collection tube. Double-stranded cDNA was cleaned by the DNA Clean & Concentrator-5 kit

(Zymo Research) purified using a 0.6X proportion of AMPure XP beads (Beckman Coulter). cDNA quality was assessed using a Bioanalyzer High Sensitivity DNA chip (Agilent Technologies), and cDNA quantity was determined by a Qubit High Sensitivity kit (Thermo Fisher Scientific). Sequencing libraries were made using the Nextera XT DNA Library Preparation Kit (Illumina, San Diego, CA, USA) that added Illumina adapters and indexes to the purified cDNA via a tagmentation reaction followed by PCR. The sequenced-ready libraries were checked for quality and quantity using an Agilent High Sensitivity Bioanalyzer assay, Qubit dsDNA HS Assay, and NEBNext library quantitation assay (New England Biolabs) before sequencing on Illumina's HiSeq2500v4 (version 2, HiSeq reagents).

10x Genomics scRNA-seq

Cellular suspensions were loaded on a Chromium Single-Cell Instrument (10x Genomics, Pleasanton, CA, USA) to generate single-cell Gel Bead–in–Emulsions (GEMs). Single-cell RNA-seq libraries were prepared using Chromium Single-Cell 3′ Library & Gel Bead Kit (version 1.1; 10x Genomics). The beads were dissolved, and cells were lysed per the manufacturer's recommendations. GEM-RT was performed to produce a barcoded, full-length cDNA from polyadenylated mRNA. After incubation, GEMs were broken, the pooled post–GEM-RT reaction mixtures were recovered, and cDNA was purified with silane magnetic beads (DynaBeads MyOne Silane Beads, PN37002D; Thermo Fisher Scientific). The entire purified post–GEM-RT product was amplified by PCR. This amplification reaction generated sufficient material to construct a 3′ cDNA library. Enzymatic fragmentation and size selection was used to optimize the cDNA amplicon size, and indexed sequencing libraries were constructed by end repair, A-tailing, adaptor ligation, and PCR. Final libraries contained the P5 and P7 priming sites used in Illumina bridge amplification. Sequence data were generated using Illumina's HiSeq400 (version 1, HiSeq reagents).

Illumina SureCell/ddSEQ scRNA-seq

Single SUM149 P/T cells treated with vehicle or TSA were filtered with 22-µm filter (Millipore) and kept in cold PBS–bovine serum albumin solution (1X PBS + 0.1% bovine serum albumin). Cell viability was determined by trypan blue staining and analyzed using a BioRad TC20 Automated Cell Counter. Cell suspensions were loaded on the ddSEQ droplet generator at a final dilution of 2500 cells/µL and containing no less than 90% live cells. Illumina SureCell WTA reagents for encapsulation, cDNA synthesis, and library construction were used. Cell suspensions were loaded into a ddSEQ cartridge in a ratio of 21.5 µL of cell enzyme mix to every 4.5 µL of cell

suspension. A total of 60 µL of 3′ barcode suspension mix and encapsulation oil were also loaded on the cartridge. The chip was then processed on a ddSEQ Single-Cell Isolator to generate single-cell droplets. Droplet-encapsulated cells were transferred to a 96-well plate, and RT mix was added and followed by RT on a thermal cycler. After RT, droplets were broken to release cDNA and followed by the first-strand purification using magnetic beads. The beads were immobilized using a magnetic peg stand, supernatant removed and discarded, beads washed twice with 80% ethyl alcohol, cDNA eluted with resuspension buffer and transferred to a fresh microcentrifuge tube. The purified first strand was used to synthesize the second strand to generate the double-strand cDNA. Amplified cDNA was simultaneously fragmented and barcoded by tagmentation using Nextera Tn5 on a thermal cycler. Illumina DNA adapters were added, which was followed by the indexing. Final libraries were cleaned up using AmpureXP beads. Purified libraries were eluted with 20 µL of resuspension buffer. Final libraries were assessed for quality and quantity with Qubit 2.0 flourometer and an Agilent Bioanalyzer 2100. Libraries were sequenced using an Illumina HiSeq2500v4 sequencer with custom SureCell sequencing primers.

TruSeq bulk RNA scRNA-seq

The total RNA was isolated using the RNeasy Plus Micro Kit (Qiagen, Valencia, CA, USA) and concentration was determined with the NanoDrop 1000 Spectrophotometer (Wilmington, DE, USA), and RNA quality was assessed with the Agilent Bioanalyzer (. The TruSeq Stranded mRNA Sample Preparation Kit (Illumina) was used for NGS library construction per the manufacturer’s protocols. Briefly, mRNA was purified from 200 ng of total RNA with oligo(dT) magnetic beads and fragmented. First-strand cDNA synthesis was performed with random hexamer priming followed by second-strand cDNA synthesis using dUTP incorporation for strand marking. End repair and 3` adenylation was then performed on the double-stranded cDNA. Illumina adaptors were ligated to both ends of the cDNA, purified by gel electrophoresis, and amplified with PCR primers specific to the adaptor sequences to generate cDNA amplicons of approximately 200–500 bp in size. The amplified libraries were hybridized to the Illumina single-end flow cell and amplified using the cBot (Illumina). Single-end reads of 100 nt were generated for each sample using Illumina’s HiSeq2500v4.

Low-input RNA scRNA-seq

Total RNA was isolated using the RNeasy Plus Micro Kit (Qiagen, Valencia, CA, USA) or Arcturus Pico Pure kit (Life Technnologies) per the manufacturer’s recommendations. RNA concentration was determined with the NanoDrop 1000

Spectrophotometer), and RNA quality was assessed with the Agilent Bioanalyzer 2100. One nanogram of total RNA was preamplified with the SMARTer Ultra Low Input kit v4 (Clontech) per manufacturer's recommendations. The quantity and quality of the subsequent cDNA was determined using the Qubit Flourometer (Life Technologies) and the Agilent Bioanalyzer 2100. One hundred fifty picograms of cDNA were used to generate Illumina-compatible sequencing libraries with the NexteraXT library preparation kit (Illumina) per the manufacturer's protocols. The amplified libraries were hybridized to the Illumina single-end flow cell and amplified using the cBot (Illumina). Single-end reads of 100 nt were generated for each sample using Illumina's HiSeq2500v4.

scRNA-seq data processing

All single-cell data were processed per vendor-recommended or provided software tools to assess performance under ideal conditions.

ddSEQ/SureCell WTA

Data processing was performed with Illumina the SureSelect WTA BaseSpace application. Briefly, raw data were demultiplexed using bcl2fastq version 1.8.4. Quality filtering and adapter removal were performed using Trimmomatic version 0.36 with the following parameters: "TRAILING:13 LEADING:13 ILLUMINACLIP:adapters.fasta:2:30:10 SLIDINGWINDOW:4:20 MINLEN:15." Cell debarcoding and unique molecular identifiers (UMIs) identification was performed with the SureCell BaseSpace application using vendor-recommended parameters. Reads that had missing pairs and cells with barcodes that could not be assigned to known barcode combinations were removed from downstream analysis. Reads assigned to each cell were aligned to Genome Reference Consortium Human Build 38 (GRCh38.7) library (Gencode-25) using the Spliced Transcript Alignment to a Reference (STAR) algorithm (version 2.5.2b), and duplicated reads with the same UMI aligning to the same genomic position were filtered to minimize amplification bias. Gene annotation was performed using featureCounts. Gene expression levels were calculated by counting the number of distinct UMIs of all gene-specific reads ("UMIs per gene"), as determined by featureCounts.

Fluidigm C1 and HT

Raw data were demultiplexed using bcl2fastq version 1.8.4. Quality filtering and adapter removal were performed using Trimmomatic version 0.36 with the following parameters: "TRAILING:13 LEADING:13 ILLUMINACLIP:adapters.fasta:2:30:10

SLIDINGWINDOW:4:20 MINLEN:15." Sequencing data were cleaned using Trimmomatic and aligned to Genome Reference Consortium Human Build 38 (GRCh38.7) library (Gencode-25) using the STAR algorithm (version 2.5.2b). Uniquely aligned reads were quantified and quality assessed using Picard Tools (version 1.114). Outlier cells were removed based on the number of genes expressed (<2000 or >10,000) and percent mitochondrial (0–0.2) or ribosomal (0–0.1) reads. Gene annotation was performed using featureCounts.

WaferGen iCell8

Raw data were demultiplexed using bcl2fastq version 1.8.4. Quality filtering and adapter removal were performed using Trimmomatic version 0.36 with the following parameters: "TRAILING:13 LEADING:13 ILLUMINACLIP: adapters.fasta:2:30:10 SLIDINGWINDOW:4:20 MINLEN:15." Sequencing data were cleaned using Trimmomatic and aligned to Genome Reference Consortium Human Build 38 (GRCh38.7) library (Gencode-25) using the STAR algorithm (version 2.5.2b). Uniquely aligned reads were quantified and quality assessed using Picard Tools (version 1.114). Outlier cells were removed based on the number of genes expressed (<2000 or >10,000) and percent mitochondrial (0–0.2) or ribosomal (0–0.1) reads. Gene annotation was performed using featureCounts. Gene expression levels were calculated by counting the number of distinct UMIs of all gene-specific reads ("UMIs per gene"), as determined by featureCounts.

10x Genomics

Raw data were demultiplexed using bcl2fastq version 1.8.4. Quality filtering and adapter removal were performed using Trimmomatic version 0.36 with the following parameters: "TRAILING:13 LEADING:13 ILLUMINACLIP:adapters.fasta:2:30:10 SLIDINGWINDOW:4:20 MINLEN:15." Quality-filtered data samples are demultiplexed using 10x Genomics–recommended workflow via Cell Ranger mkfastq. The data were then aligned and counted using cellranger count.

Bulk RNA-seq data processing

Raw reads generated from the Illumina HiSeq2500 sequencer were demultiplexed using bcl2fastq version 2.19.0. Quality filtering and adapter removal are performed using Trimmomatic-0.36 with the following parameters: "TRAILING:13 LEADING:13 ILLUMINACLIP:adapters.fasta:2:30:10 SLIDINGWINDOW:4:20 MINLEN:35." Processed/cleaned reads were then mapped to the human reference sequence (GRCh38.7) with STAR-2.6.0c with the following parameters: "--twopassMode Basic --

runMode alignReads --genomeDir ${GENOME} --readFilesIn ${SAMPLE} --outSAMtype BAM SortedByCoordinate --outSAMstrandField intronMotif --outFilterIntronMotifs RemoveNoncanonical." The subread-1.6.1 package (featureCounts) was used to derive gene counts given the following parameters: "-s 2 -t exon -g gene_name." Differential expression analysis and data normalization were performed using DESeq2-1.16.1 with an adjusted *P* value threshold of 0.05 within an R-3.4.1 environment.

Expression analysis

All analyses were performed using R/Bioconductor. The package SingleCellExperiment was used to load and represent the scRNA-seq data and the bulk RNA-seq data. Computation of quality metrics was performed using the *scater* package. The hard thresholds of 3 reads per gene was used to call a gene detected, and 500 detected genes were required per cell to accept a cell. Additionally, at least 20,000 reads on protein-coding genes were required per cell. For computation of GC content, genes were stratified according to their GC content and defined as *low-GC* (genes with a GC content below 0.4) and *high-GC* (genes with a GC content above 0.6). For reference, the genes with a GC content in the range of 0.5–0.55 were used. The fraction of detected genes in the *low-GC* group was then computed and compared to the fraction of detected genes in the reference group. The GC bias plots show the log-ratio of the 2 fractions per cell. For the length bias, the same strategy was applied by using the thresholds <800 nt for short genes, >2500 nt for long genes, and 1200–1800 nt for the reference gene sets.

Correlation of expression values was computed as Spearman's rank correlation. The edgeR package was used to compute differential expressions. Therefore, the Trimmed Mean of M Values normalization was used to estimate the between sample normalization factors. Highly variable genes were computed using the Seurat package.

RESULTS

Generation of scRNA-seq data sets

Much of the variation observed in scRNA-seq gene expression data is because of technical variation of the various methods utilized and basic biological variability.[9], [10] In order to reduce biological variation, a well-established, triple negative breast cancer cell line (SUM149 P/T) was used to limit cell heterogeneity confounding crossplatform results. A well-defined and robust gene signature of SUM149 P/T exposed to TSA (histone deacetylase inhibitor) is already published.[8] SUM149 P/T

cells were treated with vehicle (DMSO) or TSA for 16 h prior to harvesting cells for scRNA-seq across the various platforms tested. Importantly, bulk total RNA was sampled from each experimental batch of cells across platforms to evaluate both robustness and reproducibility of the treatment and to control for technical treatment batch effects. For all 5 platforms tested, duplicate cell captures were performed to control for and assess intraplatform technical variability (Fig. 1). Additionally, all SUM149 P/T cells utilized in this study were derived from the same stock cell-culture passage using a defined standard operating procedure to reduce technical batch effects from cell culture. This controlled experimental study was designed with intent to minimize as much variation as possible in order to most accurately assess scRNA-seq technologies.

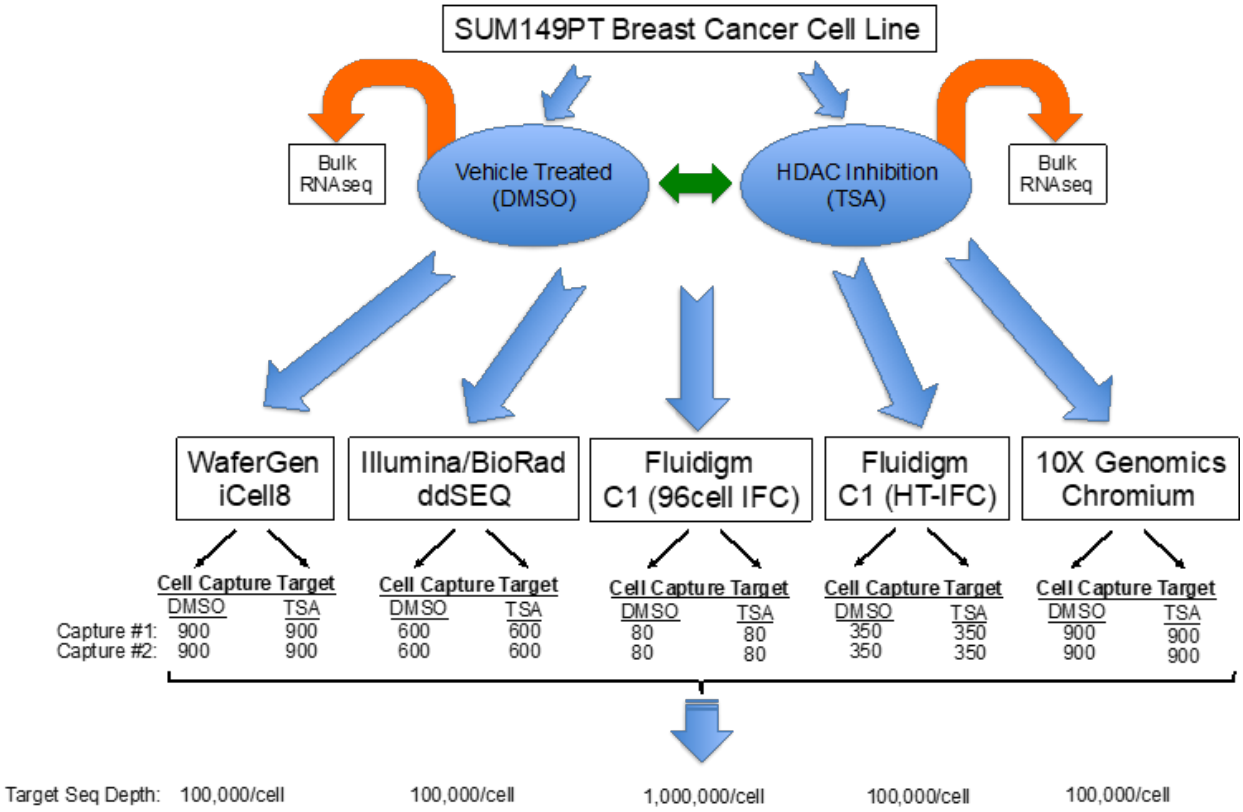

FIGURE 1. Experimental setup. Cells from the sum149PT breast cancer cell line were subjected to DMSO and TSA treatment and subsequently sequenced using different technologies.

Throughput and sensitivity across platforms

Although most platforms are capable of transcriptome-wide analysis, they vary in single-cell throughput from as little as 96 to upwards of 80,000 single cells. Although each vendor claims a specific cell throughput, this is theoretical and influenced by

many factors, such as cell viability, cell type, and quality of cell suspension (i.e., presence of cellular debris). Given this, we sought to evaluate both cell-capture throughput and gene-detection sensitivity across platforms. In order to be as fair as possible to each technology, an attempt was made to keep read depth (sequence reads) per cell at vendor recommendations across technologies (Fig. 2*A*). The number of unique genes detectable for each platform tested was then detected at the recommended sequencing depth (Fig. 2*B*). Unsurprisingly, technologies differ in the number of cells that can be processed in a single experiment (throughput) and in the number of genes detectable, but interestingly, they also differ in the amount of "usable" data, in terms of cells that pass quality-control metrics and in terms of reads that actually do contribute to gene expression counts (see Fig. 2*C*). Successful cell assessment can be impacted at the initial cell-capture step because of the stochastic nature of cell capturing between the different platform-specific protocols leading to the absence of data for those events. In addition to inefficient cell capture, each platform protocol may fail to generate sufficient reads from protein-coding genes that are necessary to generate reliable gene expression profiles. Given these potential issues, a fixed absolute threshold was used to assess the suitability of an expression profile. Specifically, a minimum of 20,000 reads/cell and greater than 500 genes detected per cell were required, and a gene was counted as detected if at least 3 reads were assigned to that specific gene locus. Implementation of these absolute thresholds imposed a minimum information requirement per cell for each technology that allowed a fairer assessment of each platform through use of only highly stringent cell and gene criteria. In summary, these observations suggest that some platform-specific bias is present and that careful consideration should be given to this during experimental design to limit such confounding factors.

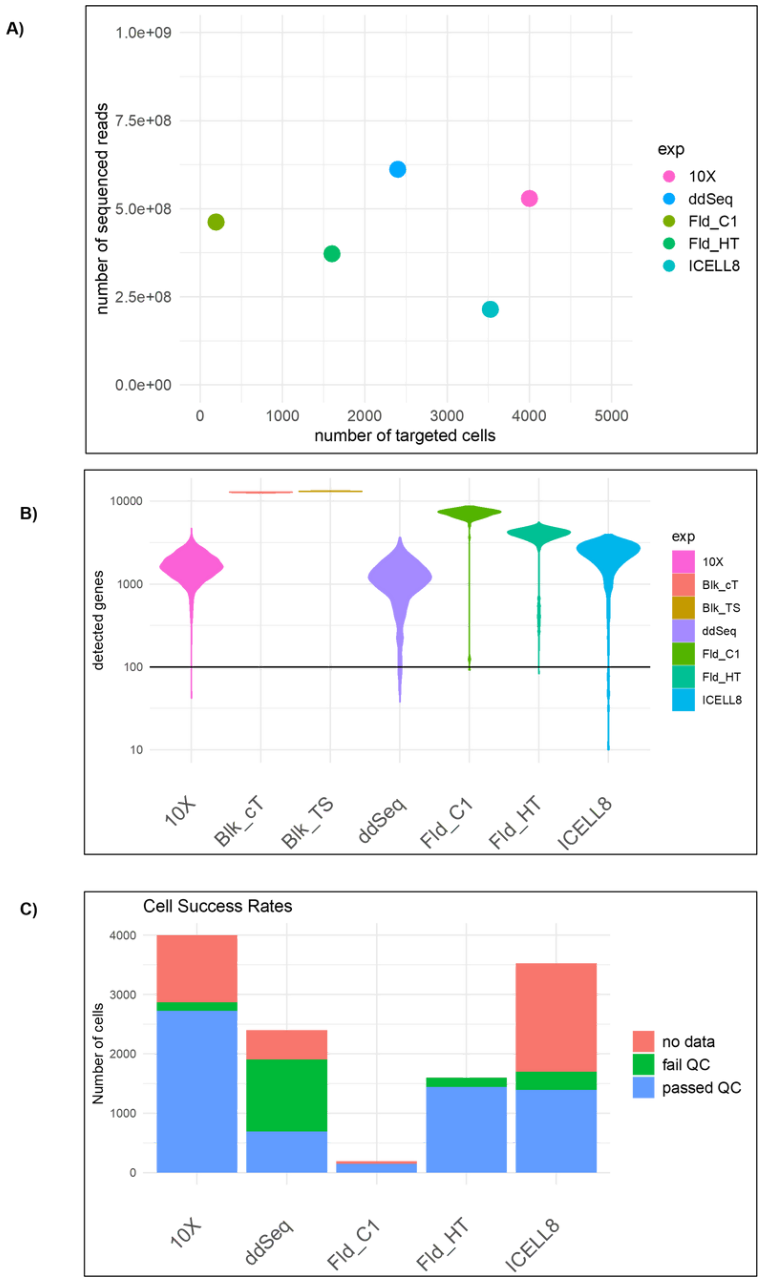

FIGURE 2. We show basic statistics of the single-cell data. (*A*) The plot shows the number of reads sequenced versus the number of cells targeted for each technology. 10x Genomics targeted the most cells. (*B*) Violin plots showing the number of genes detected in all cells. The plot shows all cells without quality filtering. (*C*) The bar plot shows whether the targeted cells were actually detected and passed the QC filtering for each technology.

The total number of genes detected depends on the number of expressed genes, the transcript abundance in each cell type, the sensitivity of the RNA capture and library

preparation, and the sequencing depth. Estimates of the relative sensitivity of the different technologies were obtained by relating the number of genes detected in a cell to the number of UMIs/reads sequenced for that cell. This provides information on the gene-detection sensitivity and the saturation behavior. Our experiments show that, for the given cell type and respective sequencing depth, saturation was not yet reached for any of the platform technologies (Fig. 3*A, B*). These data indicate that the number of detected genes per read depth, aligned and counted, is similar for the different technologies. As shown in Fig. 2*B*, the Fluidigm C1 methodology outperforms the other technologies at the recommended sequencing depth, whereas ddSEQ has the lowest number of detected genes. However, these numbers do depend on the sequencing depth, which was chosen according to what is typically recommended for each scRNA-seq technology. As an alternative comparison, Fig. 3*C* shows the number of genes detected when considering only 20,000 reads/cell that were randomly sampled. Using the down-sampled dataset, all platforms performed comparably well with gene-detection rates around 1500, except for 10x that detected less than 1000 genes. These results suggest that in our study, the diversity of the detected reads is lower for the 10x platform compared with the other platforms.

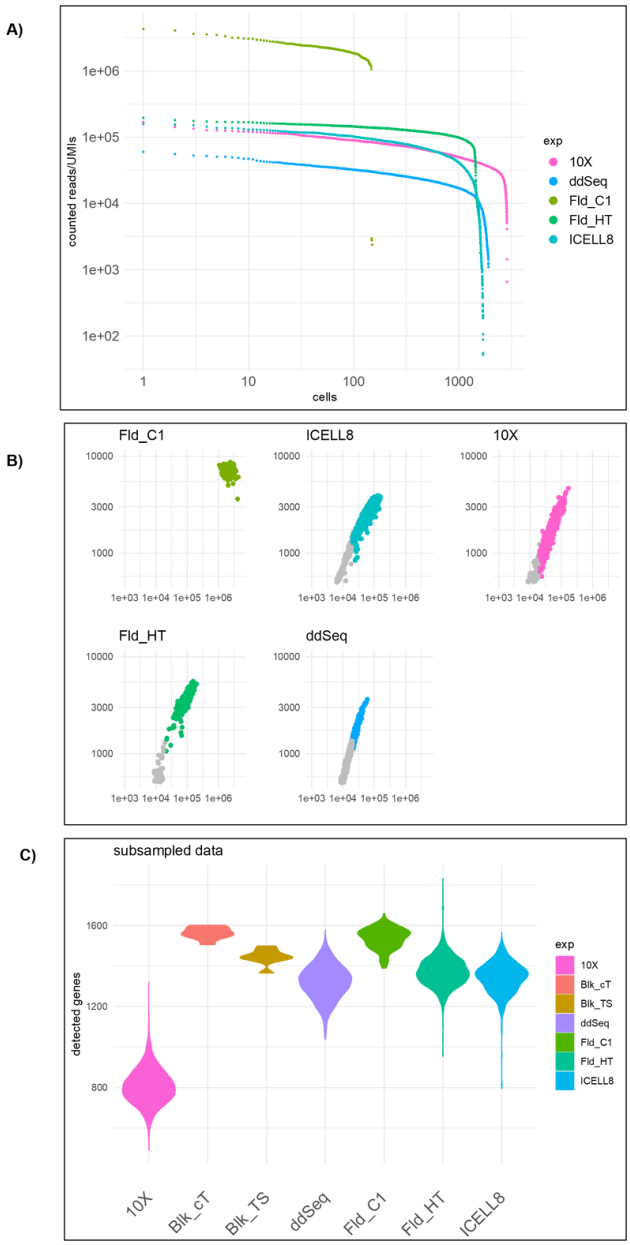

FIGURE 3. Sequencing saturation analysis. (*A*) Cells were sorted by decreasing number of reads, and the number of reads contributing to expression values is plotted. The Fluidigm C1 technology has more reads per cell but less cells than the other technologies. (*B*) Relationship of the number of genes detected in a cell versus the number of reads contributing to the expression values. (*C*) Violin plots show the number of genes detected when the sequencing depth is subsampled to 20,000 reads for each cell.

Physical gene properties impact capture efficiency of scRNA-seq methods

Previous reports suggest that RNA-seq methods are biased with respect to gene length and GC content of the genes[11], [12], [13], [14], [15]; for example, genes with high-GC content tend to be underrepresented. We investigated whether certain gene properties (e.g., GC content and gene length) impact the efficiency with which the single-cell technologies capture genes, thereby imparting platform-specific bias in these data. As shown in Fig. 4*A, B*, single-cell platforms Fluidigm HT and ICELL8 show a lower capture efficiency for high-GC content genes than for reference genes, which we defined as genes with a GC content in the interval of 0.5–0.55. Moreover, these data suggest that the 10x platform has a much lower bias for high-GC content genes relative to the other technologies, more similar to bulk RNA-seq data. Interestingly, the ddSEQ platform has reduced capture efficiency for both high-GC and low-GC content genes, whereas Fluidigm and ICELL8 have higher capture efficiency for low-GC genes. The relative capture efficiency for gene-length bias (short/long gene) was also visualized. These data suggest that 3′-tagging technologies tend to over-represent short genes, with Fluidigm HT having the smallest length bias and 10x having the strongest observed bias (Fig. 4*C, D*). Although all single-cell technologies tend to under-represent long genes, Fluidigm platforms (Fluidigm C1 and Fluidigm HT) exhibit the least bias in that respect. Interestingly, in our experiment, the Fluidigm HT platform performs better across genes of varying length than Fluidigm C1. This is despite the fact that Fluidigm HT detects reads next to the 3′-end only (3′-end sequencing), whereas Fluidigm C1 detects genes along the entire gene body.

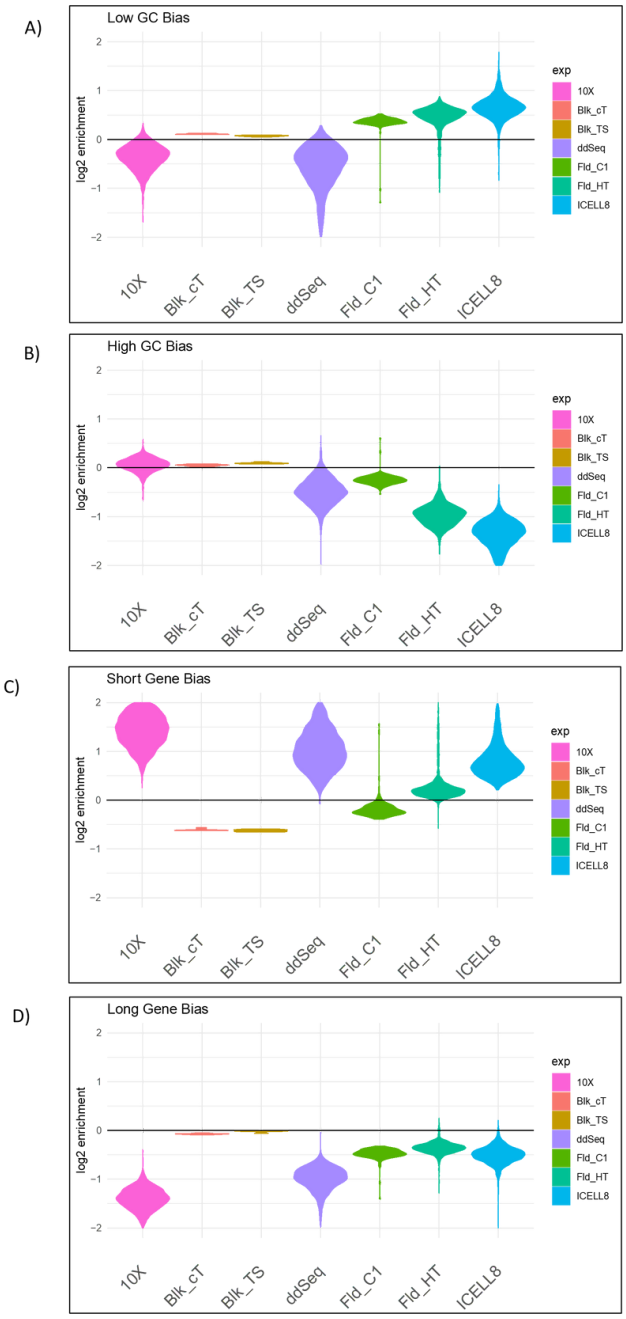

FIGURE 4. Detection bias of low-GC, high-GC, short, and long genes. The violin plots show the bias present in the individual cells. For reference, we also added the bulk protocols. Interestingly, short genes are underrepresented in bulk RNA-seq and Fluidigm C1 but not in the other scRNA-seq technologies. Overall, the bulk technologies show the smallest bias.

All technology platforms have a high degree of specificity on protein-coding regions of transcripts

A commonly utilized metric indicating the quality of RNA-seq data, is the percentage of data aligned to protein-coding regions of transcripts. The basic concept is that in poor-quality RNA-seq data, more data align to regions outside of protein-coding regions, such as intergenic regions of the genome. To assess the quality of data produced by each scRNA-seq platform, the proportion of data identified as various gene classes was determined (Fig. 5). All technologies exhibited good specificity with over 90% of the data assigned to protein-coding genes (Fig. 5*A*). Figure 5*B, C* demonstrates the percentages of the data across platforms that are assigned to other gene types, such as long noncoding RNA and small RNA species. Surprisingly, the 10x platform contained more antisense and miscellaneous RNA–assigned data than the other platforms (Fig. 5*E–G*). This observation correlates with the high capture efficiency of short protein-coding genes of the 10x platform (Fig. 4). These results suggest that although all technologies perform well with regard to detecting and quantifying protein-coding transcripts, some are better suited to identify polyadenylated, long noncoding RNAs that may be of interest for specific experimental goals.

FIGURE 5. Distribution of gene types detected by the different technologies.

Normalization and filtering of high-quality cells

In the subsequent sections, the crosstechnology concordance of the gene expression matrices were evaluated. In particular, we computed the correlation of the gene expression profiles between the different single-cell technologies and the bulk RNA protocols. Additionally, the consistence of the expression ratios computed between the 2 treatment conditions TSA and DMSO was evaluated. To this end, the expression data using the scater package was filtered and normalized. Importantly, this analysis was done using the protein-coding genes only. For the concordance analysis we used the Illumina TruSeq data as reference.

Concordance of gene expression

To begin evaluating concordance across scRNA-seq platforms, first the level of gene expression correlation within the DMSO control samples relative to the mean expression of the Illumina bulk TruSeq data was determined. As expected, the 3 replicate DMSO samples processed with Illumina TruSeq data were highly concordant with their mean, with correlation values close to 1.0 (Fig. 6). Additionally, the bulk samples generated with the Clontech low-input protocols also showed high correlation, around 0.96 (Fig. 6). However, the correlation across the individual cells from each scRNA-seq platform with bulk data was considerably lower, with values between 0.5 and 0.8 (Fig. 6). The highest correlation was found for Fld_C1, and the lowest correlation was found for ICELL8. 10x, ddSEQ and Fld_HT performed similarly with respect to this measure (Fig. 6). Similar results were found when stratifying the capture rates of the individual technologies against the bulk RNA-seq expression (Fig. 6). For any given expression level in bulk, Fld_C1 had the highest capture rate, whereas ICELL8 had the lowest capture rate.

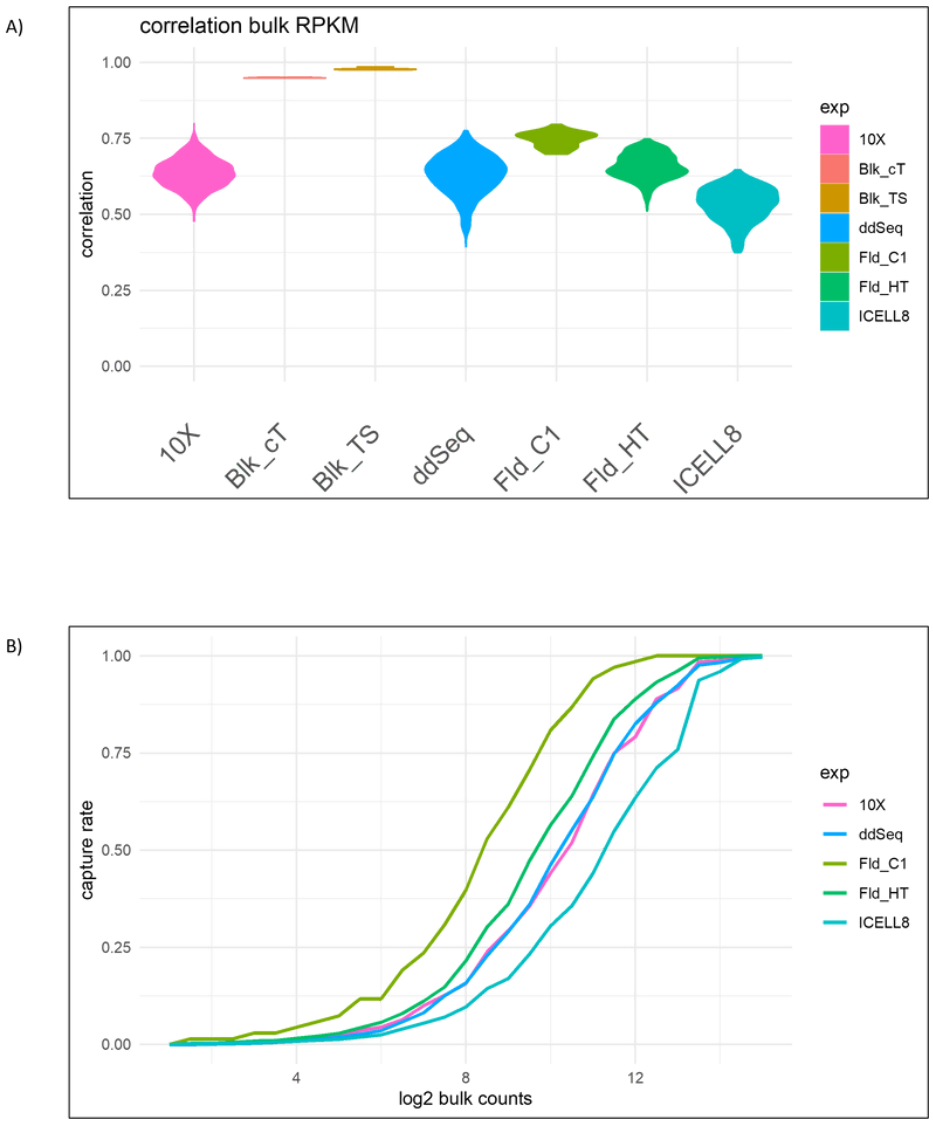

FIGURE 6. (*A*) Correlation of single-cell expression profiles with the bulk RNA-seq profile. (B) Gene-detection rate stratified by bulk RNA-seq expression. The plot shows the median detection rate in each strata. The *x* axis shows the log2 read counts in the bulk RNA-seq data.

In order to assess how well expression differences between populations of cells can be quantified by the various single-cell technologies, the expression difference between TSA and DMSO samples was computed. Again, the Illumina bulk TruSeq data were used as the gold standard against which the other technologies were compared. Figure 7*A* shows that the 2 bulk RNA-seq technologies, TruSeq versus Clontech, not only correlate well with respect to their absolute expression levels but are also consistent with respect to the expression profile differences (Fig. 7*B*). When looking at the single-cell technologies, the correlation values of the expression changes are markedly lower

(Fig. 7*B*). Here, the expression differences of Fld_HT were most in line with expression differences observed with the bulk RNA-seq methods (Figure 7*B*). As a noteworthy result, it was found that for the Fld_C1 technology, the expression differences between TSA-treated and DMSO-treated samples had a low correlation with the remaining single-cell technologies and the bulk RNA-seq samples.

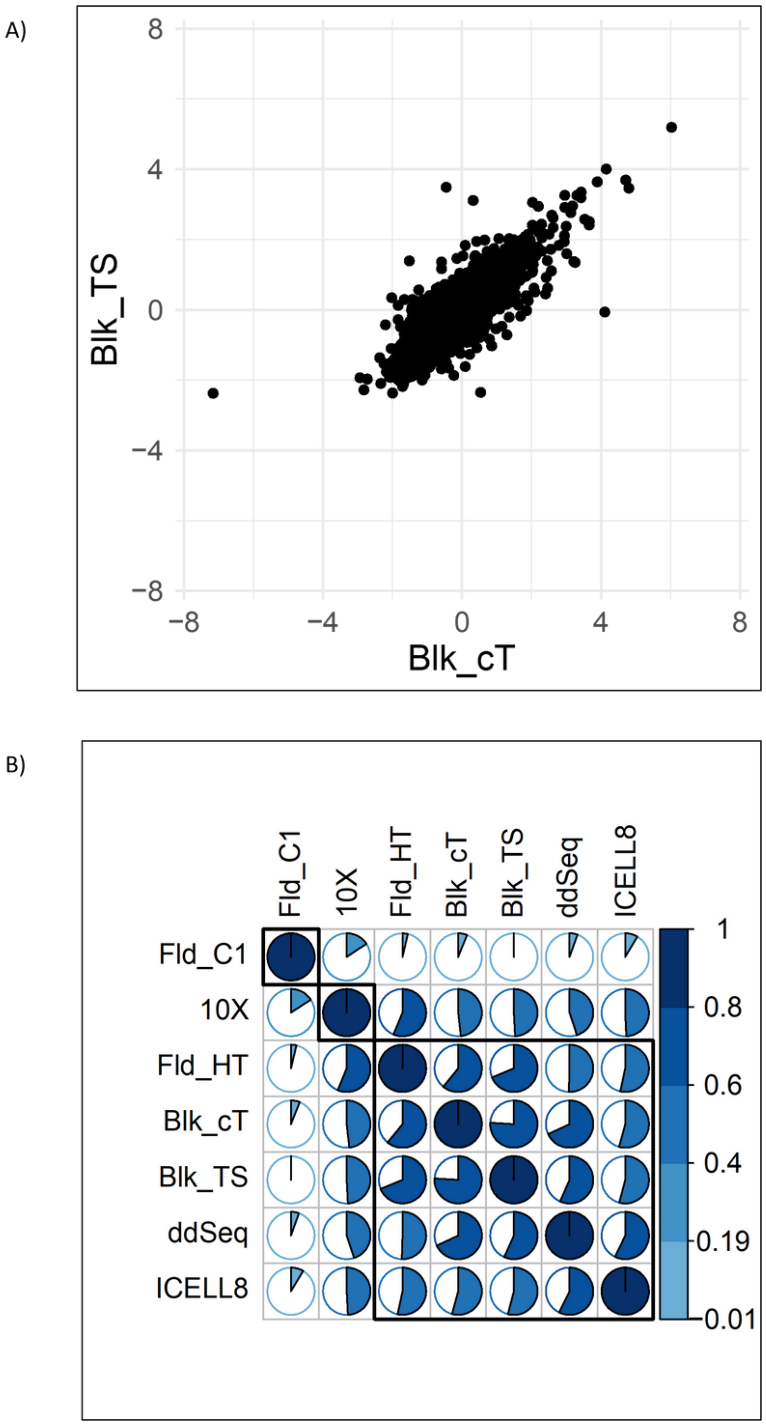

FIGURE 7. Correlation of expression ratios TSA versus DMSO. (*A*) Scatter plot comparing the 2 bulk RNA-seq technologies. (*B*) Correlation plot showing the pairwise correlation values as pie charts.

Highly variable genes significantly overlap across platforms

A key feature of single-cell technologies is the ability to capture expression heterogeneity among cells, and in fact, for this type of analysis, the most informative genes are those that are highly variable across cells. Following a recent comparative study on the effectiveness of existing methods,[10] the method implemented in Seurat[16] was used to identify highly variable genes whose expression values have variability that is higher than expected given their average expression. Figure 8 shows that pairwise overlaps, in terms of their Jaccard coefficient, have been used to cluster the sets of highly variable genes from different platforms. Unsurprisingly, the highest overlap is found between ddSEQ and 10x, the 2 droplet-based platforms, which form a cluster. The 2 Fluidigm technologies also form a cluster (Fig. 8). Only the ICELL8 technology shows a very small overlap with the other methods (Fig. 8). Taken together, these results support the crossplatform differences inherent to each scRNA-seq technology evaluated in this study. Importantly, this highlights the potential need to tightly control technology utilized for large, multicenter studies to allow for more accurate data integration.

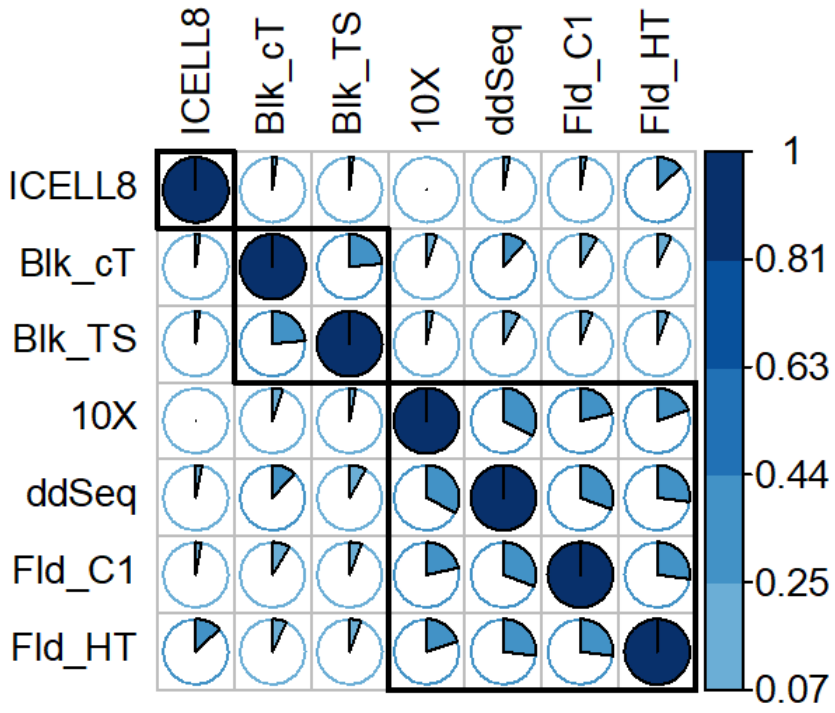

FIGURE 8. Comparing the highly variable genes across technologies. The plot shows the Jaccard Index of the pairwise overlap of the highly variable genes.

Availability

Data generated by all technologies have been uploaded to the Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/) and are available under the accession GSE142652.

DISCUSSION

Single-cell RNA-seq technologies offer unprecedented resolution toward dissecting the cellular architecture of complex biological systems. The single-cell genomics landscape is rapidly evolving with various chemistries, platforms, and technologies currently in practice. Regardless of these differences, single-cell transcriptome profiling is a multistage process involving cell capture, efficient cell lysis, transcript capture, RT, cDNA preamplification, and final cDNA library construction suitable for high-

throughput DNA sequencing. Understanding the intricacies and limitations of the various scRNA-seq modalities available is paramount to the success of the study. At the time of this study, there were 4 commercial scRNA-seq platforms available for evaluation: Fluidigm C1, WaferGen/Takara iCell8, 10x Genomics Chromium Controller, and Illumina/BioRad ddSEQ.

The goal of this comparative study was to demonstrate the level of reproducibility across several commercially available scRNA-seq technologies using a well-characterized experimental model system for the purpose of highlighting the key characteristics of each platform. Our results suggest that although each scRNA-seq platform can provide detailed transcript information on the individual cell level, some biases exist that should be carefully considered during experimental design. Overall, the technologies do capture well the protein-coding genes within the limits of systematic detection biases; biases are present in terms of preferential sequencing high-GC and short genes. Despite these biases, the experiments still provide valuable biological insight because low-GC and long genes are still being detected. This bias does not hinder the comparison of individual gene expression levels across cells. Such a bias would, however, become relevant when directly comparing read counts expression across technologies or when relating read counts of 2 different genes to each other.

A key question is how compatible and comparable are the expression values obtained from individual cells with the traditional RNA-seq analysis of bulk tissue. In our analysis of SUM149PT cells, both approaches correlate and give overall consistent results. Bulk RNA-seq was considered the gold standard of the single-cell technologies, and 2 well-established protocols, the Illumina TruSeq protocol and the Clontech protocol, were performed. Both protocols gave highly consistent results, which supports the idea of using bulk RNA-seq as reference. When reporting correlation values with single-cell protocols, the TruSeq protocol was used as reference, and the Clontech protocol was not used because it shares some library preparation steps with the Fld_C1 protocol, which might favor the Fld_C1 protocol in the performance evaluation.

Our study involved 2 different treatments of the cells, TSA and DMSO, and permitted the computation of expression differences and the evaluation of their consistency. Overall, the bulk RNA-seq expression differences between the 2 treatments DMSO and TSA are also reflected in the single-cell data.

A key characteristic of single-cell technologies is the throughput in terms of the number of cells and the associated cell success rate. The capacity of the microfluidic technology is lower than that of droplet-based and plate-based technologies, with the later technologies providing the possibilities to scale up the experiments. The higher capacity, however, comes with a compromise regarding the cell success rate. Although the microfluidic chips reached up to a 90% success rate in our study, this is not true for the other systems. For them, we observed only a success rate that ranged between 50% and 80%. These numbers certainly depend on the cell types and cell viability. As a consequence that cannot be generalized to any cell type and any condition. An important lesson is that researchers should factor in failing cells when designing single-cell studies and plan the number of cells entering the study accordingly.

Of equal importance is the number of genes detected per cell, which represents a characteristic complementary to the number of cells. Only when transcripts are detected with a high sensitivity in a cell, the full potential of scRNA-seq unfolds. In terms of gene-detection sensitivity, the microfluidics platforms perform very well. They have the highest number of detected genes. The other platforms that measure many more cells perform less well regarding the gene-detection sensitivity per cell. We are aware that the detection sensitivity also depends on the sequencing depth, since transcripts may not only be lost at cDNA and subsequent library generation but also in the sequencing step that acts as a random sampling process. So, by putting more sequencing efforts into the experiment, gene detection can be improved until the limits imposed by the library generation are reached. The downsides are the increasing costs and inefficiencies caused by sequencing duplicated reads. Our general finding that the sequencing is not yet saturated is consistent with Ziegenhain et al.,[7] which shows an increase in gene-detection sensitivity for sequencing depths up to approximately 2 million reads/cell for protocols doing full transcript sequencing.

Our summary of the single-cell data demonstrates the capabilities of the technologies and sheds light on the limitations. We abstain from a performance rating of the technologies since we have evaluated them based only on a single experiment using one cell type. Our analyses highlight the considerations to be taken into account when planning, executing, and especially analyzing single-cell data.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Martin Tenniswood for the generous donation of the SUM149PT cells. They would also like to acknowledge the generous donation of

kits and reagents by 10x Genomics, BioRad, Fluidigm, Illumina, and WaferGen Biosystems.

Citations

1. Karaayvaz M, Cristea S, Gillespie SM, et al. Unravelling subclonal heterogeneity and aggressive disease states in TNBC through single-cell RNA-seq. *Nat Commun* 2018;9:3588. ↩
2. Navin N, Kendall J, Troge J, et al. Tumour evolution inferred by single-cell sequencing. *Nature* 2011;472:90–94. ↩
3. Sen R, Dolgalev I, Bayin NS, Heguy A, Tsirigos A, Placantonakis DG. Single-cell RNA sequencing of glioblastoma cells. *Methods Mol Biol* 2018;1741:151–170. ↩
4. Saunders A, Macosko EZ, Wysoker A, et al. Molecular diversity and specializations among the cells of the adult mouse brain. *Cell* 2018;174:1015–1030.e16. ↩
5. Goldstein LD, Chen Y-JJ, Dunne J, et al. Massively parallel nanowell-based single-cell gene expression profiling. *BMC Genomics* 2017;18:519. ↩
6. Baran-Gale J, Chandra T, Kirschner K. Experimental design for single-cell RNA sequencing. *Brief Funct Genomics* 2018;17:233–239. ↩
7. Ziegenhain C, Vieth B, Parekh S, et al. Comparative analysis of single-cell RNA sequencing methods. *Mol Cell* 2017;65:631–634.e4. ↩
8. Chatterjee N, Wang W-L, Conklin T, Chittur S, Tenniswood M. Histone deacetylase inhibitors modulate miRNA and mRNA expression, block metaphase, and induce apoptosis in inflammatory breast cancer cells. *Cancer Biol Ther* 2013;14:658–671. ↩
9. Hicks SC, Townes FW, Teng M, Irizarry RA. Missing data and technical variability in single-cell RNA-sequencing experiments. *Biostatistics* 2018;19:562–578. ↩
10. Hwang B, Lee JH, Bang D. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp Mol Med* 2018;50:1–14. ↩
11. Zheng W, Chung LM, Zhao H. Bias detection and correction in RNA-sequencing data. *BMC Bioinformatics* 2011;12:290. ↩

12. Tuerk A, Wiktorin G, Güler S. Mixture models reveal multiple positional bias types in RNA-seq data and lead to accurate transcript concentration estimates. *PLoS Comput Biol* 2017;13:e1005515. ↩

13. Young MD, Wakefield MJ, Smyth GK, Oshlack A. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol* 2010;11:R14. ↩

14. Roberts A, Trapnell C, Donaghey J, Rinn JL, Pachter L. Improving RNA-seq expression estimates by correcting for fragment bias. *Genome Biol* 2011;12:R22. ↩

15. Oshlack A, Wakefield MJ. Transcript length bias in RNA-seq data confounds systems biology. *Biol Direct* 2009;4:14. ↩

16. Satija R, Farrell JA, Gennert D, Schier AF, Regev A. Spatial reconstruction of single-cell gene expression data. *Nat Biotechnol* 2015;33:495–502. ↩