Toward uncharted territory of cellular heterogeneity: advances and applications of single-cell RNA-seq

Brandon Lieberman1,#, Meena Kusi1,#, Chia-Nung Hung1,#, Chih-Wei Chou1, Ning He2, Yen-Yi Ho3, Josephine A. Taverna4,5, Tim H. M. Huang1,5, Chun-Liang Chen1,5

1Department of Molecular Medicine, University of Texas Health Science Center at San Antonio, San Antonio, TX 78229, USA. 
2Department of Nursing, University of Texas Health Science Center at San Antonio, San Antonio, TX 78229, USA. 
3Department of Statistics, University of South Carolina, Columbia, SC 29208, USA. 
4Department of Medicine, University of Texas Health Science Center at San Antonio, San Antonio, TX 78229, USA. 
5Mays Cancer Center, University of Texas Health Science Center at San Antonio, San Antonio, TX 78229, USA. 
#Authors contributed equally.

Correspondence to: Dr. Chun-Liang Chen and Dr. Tim H. M. Huang, Department of Molecular Medicine, University of Texas Health Science Center at San Antonio, San Antonio, Texas 78229, USA. E-mails: chenc4@uthscsa.edu; thuang3@uthscsa.edu


Received: 16 Oct 2020 First Decision: 10 Nov 2020 Revised: 21 Nov 2020 Accepted: 30 Nov 2020 Available online: 1 Jan 2021

Abstract
Among single-cell analysis technologies, single-cell RNA-seq (scRNA-seq) has been one of the front runners in technical inventions. Since its induction, scRNA-seq has been well received and undergone many fast-paced technical improvements in cDNA synthesis and amplification, processing and alignment of next generation sequencing reads, differentially expressed gene calling, cell clustering, subpopulation identification, and developmental trajectory prediction. scRNA-seq has been exponentially applied to study global transcriptional profiles in all cell types in humans and animal models, healthy or with diseases, including cancer. Accumulative novel subtypes and rare subpopulations have been discovered as potential underlying mechanisms of stochasticity, differentiation, proliferation, tumorigenesis, and aging. scRNA-seq has gradually revealed the uncharted territory of cellular heterogeneity in transcriptomes and developed novel therapeutic approaches for biomedical applications. This review of the advancement of scRNA-seq methods provides an exploratory guide of the quickly evolving technical landscape and insights of focused features and strengths in each prominent area of progress.

Keywords: Single-cell RNA-seq, transcriptome, heterogeneity, multiplexing, high throughput, dimensional reduction, cancer, diseases
INTRODUCTION

Homogeneity and heterogeneity are proportionally co-existent in all phenotypical and genetic levels of humans. The extent of heterogeneity is increased from individuals down to molecules, whereas homogeneity is decreased [Figure 1A]. The diversity is observed in individuals, organs, tissues, cells, organelles, and molecules, and even more abundant in protein, DNA, and RNA molecules. The long-standing paradigm that cells of the same tissue origin are homogeneous based on bulk cell studies has lately been challenged by single-cell studies [1-5]. New data show that cells arising from the same tissue origins are far more heterogeneous than they seemingly appear [Figure 1B] [6-8]. Even genetically identical cells cultured in the same conditions have shown variations in gene expression [9,10].

In the new paradigm, the diverse properties of cells are mainly reflected in the heterogeneous gene expression, genomic alterations, epigenomic modifications, and proteomic fluctuations [7,11-16]. Cellular transcriptomic heterogeneity helped to establish a new paradigm of cellular heterogeneity with the invention of scRNA-seq [17-19]. The cellular transcriptomic heterogeneity arises from stochasticity, differentiation, environmental stimuli, diseases, aging, and other factors [5,8,11,20-23]. The development of single-cell analysis was overshadowed by traditional bulk cell approaches and technically limited by the absence of high throughput single-cell isolation and minute initiation materials (picogram DNA and mRNA per cell) [5]. Combined technological advances in cell isolation, high throughput multiplexing, amplification, and next generation sequencing facilitated scRNA-seq and uncovered cellular heterogeneity [24]. Mapping transcriptomic changes at single-cell level has since revealed global gene expression profiles and exposed stochasticity, differentiation, cell fate plasticity, and diseases [25]. In this review, we highlight the novel scRNA-seq platforms and conduct a comparative analysis of each technology and their future applications in translational science.

SINGLE-CELL ISOLATION TECHNOLOGIES

Cell purity is paramount for scRNA-seq and other single-cell analysis methods. Tissues, organoids, and 2D and 3D cultured cells are multi-cellular, and the first step to dissociate aggregated cells into individual cells risks potential contamination with cell doublets, DNA, and RNA by incomplete enzymatic digestion and cell lysis. The impurity of single cells distorts the scRNA-seq data and leads to false interpretations. To ensure the purity and integrity of single cells, several instrumental technologies have been adopted to overcome these technical challenges.

Manual cell picking (micromanipulation)

Manual cell picking presents a simple and cost-effective method for single-cell isolation. This technique involves an inverted fluorescent microscope, manipulator, and microinjector for precise cell location and picking after cells are labeled with markers [26]. The instrumental efficiency of picking individual cells exceeds old-fashioned mouth-pipetting [25]. Specifically, cells are maintained in suspension and manually isolated by capillary pipettes connected with a microinjector. Cellular integrity is maintained for further analysis and is particularly workable with rare cells. However, high operator skills are required through training and practice. Additionally, the throughput is relatively low compared with the other methods [27].

Flow activated cell sorting

This high throughput method relies on antibody affinity to cell surface markers and has become the most common strategy for single-cell isolation. Cells are labeled with fluorescent or conjugated antibodies and run through flow cytometry, sensed by laser detectors or a magnetic field, and sorted with surface specific markers [27,28]. With advanced fluorochrome and microscope techniques, 18 fluorescent, inorganic semiconductor nanocrystals (Quantum Dots) are used to label antigens on cells, which increases the specificity and sensitivity of single-cell isolation from a bulk sample [29,30]. However, greater than 10,000
cells are required for this method and signal overlap may affect the purity of the target cells. Moreover, this method cannot perform single-cell analysis with rare cells.

Microfluidic technology

This technology for single-cell isolation can be divided into three main approaches: droplet-based microfluidics, channel-based microfluidics, and hydrodynamic traps. These methods rely on cell adhesion, hydrodynamics, physical characteristics (e.g., size and shape), cellular density, and elasticity. Microfluidic technology platforms can actively or passively recognize and sort single cells from a heterogenic population. In droplet-based microfluidics, each single cell is embedded in a hydrophilic droplet which suspends hydrophobic channels. The advantages of this approach are high throughput and yield, making it feasible to isolate rare cell types. Besides, genetic barcodes can be added within the cell droplet that record the cell origin, allowing profiling of cells from simultaneous preparation of thousands of single-cell libraries. In channel-based microfluidics, single cell is controlled and confined by pneumatic membrane valves according to the biological requirements. This selection approach increases the accuracy of cell isolation and the flexibility of experimental design. However, it is limited by the low throughput compared with droplet-based microfluidics. Hydrodynamic traps such as Fluidigm C1 passively isolate and trap single cells based on cell size. Both channel-based microfluidics and hydrodynamic traps enable long-term cell culture and high-resolution observation for further biological experiments such as drug treatment or cDNA library preparation.
Laser capture microdissection

This microscopy-based technology carries out isolation of specific single cells on a microscope slide without cell dissociation from solid samples. Tissue sections are either top-covered by or laid on a thermoplastic polymer film, which is heat-activated by infrared or ultraviolet laser. The boundaries of a single cell on a tissue section are recognized and severed precisely by laser, and the dissected cell is captured \[36,37\]. Laser capture microdissection is a rapid and precise isolation technology that maintains versatility for further analysis including scRNA-seq \[38,39\].

TECHNICAL ADVANCES FOR CDNA SYNTHESIS AND AMPLIFICATION

The uniform and full coverage of cDNA synthesis from single-cell mRNA/RNA is a crucial step for the success of scRNA-seq because the limited starting materials are as little as 5-30 pg and need to be amplified for next generation sequencing. The cDNA synthesis from single cells has been attempted for qRT-PCR and microarrays, and the technical predecessors were adopted and modified for scRNA-seq \[40,41\]. The protocols of scRNA-seq have fruitfully advanced in a decade \[19,25,42-44\]. The technical variations have strengths and weaknesses in linear amplification, length coverage, low copy RNA species detection, multiplexing, high throughputs, and cost reduction. Tang’s protocol was the first scRNA-seq modality and was based on single-cell RNA amplification from RNA microarray assays \[45,46\]. Tang’s protocol uses poly(T) primers to generate full-length cDNA of transcripts less than 3 kb and can detect ~13,000 genes, 65% of microarray genes \[42\]. Two years later, Single-cell Tagged Reverse Transcription sequencing (STRT-seq) introduced template switching to incorporate bead-linked barcoded primers for strand-specific amplification of 3’ ends and high throughput 96-cell multiplexing with 2000-4000 genes detected in individual cells \[18,47,48\]. In 2012, a significant advancement for full-length cDNA synthesis of 40% transcripts was made with Smart-seq, and it was updated with the Smart-seq2 in 2013 \[49-51\]. Smart-seq has laid the foundation for future scRNA-seq methods, employing more stable template switching ribo(guanosine), oligos and having more unique mapping reads, higher recovery rates of low expression genes, and a two-fold increase in spliced forms discovered. At about the same time, Cell Expression by Linear amplification and sequencing (CEL-seq) utilized linear strand-specific in vitro transcript amplification mainly at 3’ ends, and an improved CEL-seq2 version reduced mRNA molecule counting biases with the introduction of unique molecular identifiers (UMIs) \[52-55\]. Single Cell RNA Barcoding and sequencing (SCRB-seq) is a protocol for high throughput of 12,000 cells at a low cost and one of the first scRNA-seq protocols to include UMIs \[56\]. Previous scRNA-seq platforms utilized relative measures such as reads per kilobase per million reads (RPKM), which masked differences in total mRNA content. As an example, a gene may be “upregulated” in terms of RPKM and have a decrease in absolute expression level. UMIs are short unique sequences integrated in cDNAs before PCR amplification to allow for unique identification of amplified DNAs carrying the same UMI sharing the same mRNA/RNA molecule origin and reduce nonlinear PCR amplification bias. For full length transcript coverage and analysis of noncoding RNA, Multiple Annealing and dC-Tailing-based Quantitative single-cell RNA-seq (MATQ-seq) and Random Displacement Amplification sequencing (RamDA-seq) can be employed, which allow for poly(A)+ and non-poly(A) scRNA-seq, useful for characterization of lncRNA or circRNA \[57-59\]. RamDA-seq also detects enhancer RNAs differentially expressed in a cell-type specific manner. Quartz-Seq builds upon both CEL-seq2 and STRT-seq to perform 3’ coverage scRNA-seq, vastly improving poly(A) tailing and initial read UMI conversion and augmenting sequencing depth and accuracy \[60\].

Many methods employ cDNA amplification strategies, as mentioned above, but they each have unique methods of single-cell sorting and significant high throughput improvement. MAssively parallel RNA Single-cell sequencing (MARS-seq) uses fluorescence-activated cell sorting sorting to separate and sort 100-1000 cells into individual wells \[61\]. Drop-seq and InDrop are two similar methods that employ droplet capture microfluidic methods to isolate cells. The main differences are that Drop-seq uses reagent containing beads, while InDrop uses reagent carrying hydrogel microspheres. Both platforms can quickly
process tens of thousands of cells daily [33,34]. Chromium is similar to both previous methods, employing a gel bead in emulsion [62] microfluidic capture method, but has the advantage of being able to process eight samples at once or a single sample more quickly due to its eight-channel microfluidic chip [63]. Seq-Well is a low-cost alternative, not requiring any expensive microfluidic devices and instead utilizing semi-permeable membranes on a picowell plate with wells that contain one barcoded capture bead and space for one single cell per well [64]. Seq-Well plates have ~86,000 wells, but the actual capture efficiency varies. Similar to the other platforms (CEL-seq, STRT-seq, and SCRB-seq), CytoSeq employs a microfluidic method automating high throughput cell settling in 1/10 of 100,000-well plate by gravity [65]. It employs a similar plate system to Seq-Well with 30-µm well sizes only allowing one cell per well and one magnetic bead with a universal primer plus 10^6 diverse UMIs created by a split-pool synthesis process. It can easily reach up to 10,000 cells with detection of ~100 genes per cell.

For harder to work with tissues, such as frozen samples, DroNC-seq is able to salvage samples and produce high quality data, employing single nucleus RNA-seq with 3' coverage [66]. sci-RNA-seq performs

| Table 1. cDNA synthesis and amplification techniques for scRNAseq |
|-----------------|----------------|----------------|----------------|----------------|
| Methods         | coverage       | UMI            | Strand specific | cDNA synthesis  | Detected genes |
| Tang's          | Nearly full-length | No  | No  | poly(T) primer | 13K          |
| STRT-seq and STRT/C1 | 3' and 5' only | Yes | Yes | tailed oligo-DT primer; a barcoded r(G), helper oligo primer | -2-4K |
| Smart-seq       | Full-length   | No  | No  | tailed oligo(dT) priming using the CDS primer | -8K |
| CEL-seq (CEL-seq2) | 3' only    | Yes | Yes | 8bp-barcoded poly(T) primer | -5K |
| Smart-seq2      | Full-length   | No  | No  | tailed oligo(dT) priming using the CDS primer | -10K |
| Quartz-Seq      | Full-length   | No  | No  | poly(T) primer | 5.8-6.3K |
| DP-seq          | 3'-only       | No  | No  | hexamer | 11K transcripts |
| SCRB-seq        | 3' only       | Yes | Yes | cell-barcoded UMI-Poly(T) primer | 3k transcripts |
| MARS-seq        | 3'-only       | Yes | Yes | barcoded Poly(T) primer | ~200-1500 transcripts |
| Drop-seq        | 3'-only       | Yes | Yes | bead-based barcoded UMI-poly(T) primer | 6-7K genes |
| InDrop          | 3'-only       | Yes | Yes | hydrogel sphere encapped cell barcoded UMI-poly(T) | 29K UMIFM |
| SUPER-seq       | Full-length   | No  | No  | random (AnchorX-T15N6) primers | -10K |
| CytoSeq         | 3'-only       | Yes | Yes | Illumina universal PCR primer & cell UMI-Poly(T) | -100 |
| SC3-seq         | 3' only       | No  | No  | V1(dT)24 | 4-6K |
| MATQ-seq        | Full-length   | Yes | Yes | GATdT primers; MALBAC primers | -14K |
| Chromium        | 3'-only       | Yes | Yes | Gel bead based 14x GEM index-10x barcoded-poly(T) primer | -500 |
| SPLiT-seq       | 3'-only       | Yes | Yes | random hexamer and anchored poly(dt)15 barcoded RT primers | 4.5-5.5K |
| sci-RNA-seq     | 3'-only       | Yes | Yes | 10bp barcoded-Bbp UMI-Poly (T30) primer | 4-5.5K |
| Seq-Well        | 3'-only       | Yes | Yes | bead-based 12bp barcoded 8bp UMI- Poly(T)30 primer | 6-7K |
| DroNC-seq       | 3'-only       | Yes | Yes | bead-based barcoded UMI-poly(T) primer | 1.7-3.3K |
| Quartz-Seq2     | 3'-only       | Yes | Yes | cell-barcoded UMI-poly(T) primer (v3.1: 73-mer) | 8K |

STRT-seq: single-cell tagged reverse transcription sequencing; CEL-seq: cell expression by Linear amplification and sequencing; DP-seq: designed primer-based RNA-sequencing; SCRB-seq: single cell RNA barcoding and sequencing; MARS-seq: Massively parallel RNA single-cell sequencing; MATQ-seq: Multiple annealing and dC-tailing-based quantitative single-cell RNA-seq; SPLiT-seq: split-pool ligation-based transcriptome sequencing
the analysis of single cells or nuclei isolated from methanol-fixed whole organisms (~50,000 cells), with 3’ coverage and high depth sequencing employing double UMI barcoding. Split-Pool Ligation-based Transcriptome sequencing (SPLiT-seq) is an extremely high throughput 3’ coverage method distinct from other methods, offering scRNA-seq analysis without single-cell isolation from 1.33% formaldehyde-fixed tissues. It employs combinatorial indexing to identify single cells without isolation, by performing three successive barcoding steps through in situ reverse transcription on groups of cells and mixing after each time, leaving each cell with a unique identifier totaling up to 21 million for downstream data analysis. This invention posits potential cost reduction per cell and time effectiveness.

Not all methods employ UMIs. Designed Primer-based RNA-sequencing (DP-seq) is a 3’ coverage method useful for small samples, analyzes at least 50 pg of RNA, and employs random hexamer-based amplification and sequencing. The protocol requires knowledge of the intended target’s genome prior to use. SC3-seq provides 3’ coverage method for only 3’ end characterization of mRNA, allowing for higher reproducibility and reduced noise, thus it is useful for projects not requiring deep sequencing with a low budget.

Attempts have been made to evaluate diverse scRNA-seq protocols using systematic comparisons. Six commonly used methods - CEL-seq2, Drop-seq, MARS-seq, SCRIB-seq, Smart-seq, and Smart-seq2 - were performed side by side using mouse embryonic stem cells. Overall, Smart-seq2 detected the most genes per cell, while the other methods displayed less amplification noise due to the use of UMIs. Power analysis showed that Drop-seq provides a less costly option for a larger number of cells analyzed, whereas MARS-seq, SCRIB-seq, and Smart-seq2 were suitable for fewer cells. To evaluate sensitivity and accuracy as performance metrics for 15 different methods, in silico power analysis of External RNA Controls Consortium spike-in standards was performed in those scRNA-seq studies. The vast majority of methods displayed high accuracy. The exceptions were CEL-seq and MARS-seq data, which show more variations among cells. Better sensitivity appeared in SMARTer (C1), CEL-Seq2 (C1), STRT-Seq, and InDrop-seq that could detect digital copies of spike-ins; however, the sensitivity was sequencing depth-dependent. For the droplet-based high-throughput platforms (InDrop, Drop-seq, and 10x Genomics Chromium), a thorough comparative study revealed insights regarding their efficacy and limits. The 10X Genomics Chromium protocol was maturely developed with a higher cost and delivered high degree of sensitivity and accuracy with less technical noise. Drop-seq provided similar data quality with fewer cells, but a more affordable cost. InDrop is also less expensive with high compatibility with other protocols, such as Smart-seq2. In a large-scale validation of 13 protocols by multi-centered collaboration effort using mixed human and murine cells, CEL-seq2, Quartz-seq2, SMART-seq2, and Chromium platforms were superior in producing high-resolution transcriptomic profiles.

**PROCESSING OF NEXT GENERATION SEQUENCING DATA OF SCRNA-SEQ**

Synthesized and amplified cDNAs are subsequently subject to library preparation and NGS to generate massive sequencing short reads, as depicted in Figure 2. The current state-of-the-art computational tools and algorithms widely used in bulk RNA-seq analysis can be extended for processing scRNA-seq data. However, the transcriptome at single-cell resolution presents specific analytical challenges, which requires dedicated analytical power and specific packages. Some of the key challenges encountered during single-cell transcriptional data analysis includes greater dimensionality, high level of noise, absence of biological replicates, and data sparsity. However, major efforts in the development of advanced algorithms and computational strategies as well as adaptations of existing workflows have shown great promise for comprehensive and detailed analysis of scRNA-seq data. Several programming- (R- or Python-based) and web interface-based toolkits have been proposed to facilitate systematic analysis that can be scaled up as per requirements. Seurat and Single Cell ANalysis in PYthon are the two most comprehensive packages that can, respectively, integrate scRNA-seq data with other single-cell data and enable scaling-up to simultaneously analyze millions of cells at the same time. Of note, the core analysis
pipelines show higher resemblance with the bulk RNA-seq and can be broadly categorized in the following: (1) quality control; (2) read alignment and generation of counts; (3) removal of confounding factors; and (4) normalization and annotation of cell types and cellular states. The quality of individual single-cell libraries needs to be carefully assessed to abolish the underlying noise as downstream interpretation relies
Table 2. NGS data analysis tools and software for scRNA-seq

<table>
<thead>
<tr>
<th>Category</th>
<th>Tools</th>
<th>Software</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quality control</td>
<td>MultiQC</td>
<td><a href="http://multiqc.info">http://multiqc.info</a></td>
<td>Ewels et al. [80], 2016</td>
</tr>
<tr>
<td></td>
<td>SinQC</td>
<td><a href="http://www.morgridge.net/SinQ.html">http://www.morgridge.net/SinQ.html</a></td>
<td>Jiang et al. [82], 2016</td>
</tr>
<tr>
<td></td>
<td>SCell</td>
<td><a href="https://github.com/diazlab/SCell">https://github.com/diazlab/SCell</a></td>
<td>Diaz et al. [93], 2016</td>
</tr>
<tr>
<td></td>
<td>Celloline</td>
<td><a href="https://github.com/Teichlab/celloline">https://github.com/Teichlab/celloline</a></td>
<td>Ilici et al. [94], 2016</td>
</tr>
<tr>
<td></td>
<td>Kraken</td>
<td><a href="http://ccb.jhu.edu/software/kraken/">http://ccb.jhu.edu/software/kraken/</a></td>
<td>Wood and Salzberg [81], 2014</td>
</tr>
<tr>
<td></td>
<td>HTQC</td>
<td><a href="https://sourceforge.net/projects/htqc/">https://sourceforge.net/projects/htqc/</a></td>
<td>Yang et al. [83], 2013</td>
</tr>
<tr>
<td>Alignment</td>
<td>Kallisto</td>
<td><a href="https://github.com/pachterlab/kallisto">https://github.com/pachterlab/kallisto</a></td>
<td>Bray et al. [88], 2016</td>
</tr>
<tr>
<td></td>
<td>HISAT</td>
<td><a href="https://github.com/infphilo/hisat">https://github.com/infphilo/hisat</a></td>
<td>Kim et al. [87], 2015</td>
</tr>
<tr>
<td></td>
<td>TopHat2</td>
<td><a href="https://github.com/infphilo/tophat">https://github.com/infphilo/tophat</a></td>
<td>Kim et al. [89], 2013</td>
</tr>
<tr>
<td></td>
<td>STAR</td>
<td><a href="https://code.google.com/archive/p/rna-star/">https://code.google.com/archive/p/rna-star/</a></td>
<td>Dobin et al. [80], 2013</td>
</tr>
<tr>
<td></td>
<td>GSNAP</td>
<td><a href="https://bioinformaticshome.com/tools/rna-seq/descriptions/GSNAP.html">https://bioinformaticshome.com/tools/rna-seq/descriptions/GSNAP.html</a></td>
<td>Wu et al. [92], 2010</td>
</tr>
<tr>
<td></td>
<td>MapSplice</td>
<td><a href="http://www.netlab.uke.de/p/bioinfo/MapSplice">http://www.netlab.uke.de/p/bioinfo/MapSplice</a></td>
<td>Wang et al. [90], 2010</td>
</tr>
<tr>
<td>Quantification</td>
<td>StringTie</td>
<td><a href="http://ccb.jhu.edu/software/STRINGTIE/">http://ccb.jhu.edu/software/STRINGTIE/</a></td>
<td>Pertea et al. [93], 2015</td>
</tr>
<tr>
<td></td>
<td>HTSeq</td>
<td><a href="https://htseq.readthedocs.io/en/master/">https://htseq.readthedocs.io/en/master/</a></td>
<td>Anders et al. [90], 2014</td>
</tr>
<tr>
<td></td>
<td>FeatureCounts</td>
<td><a href="http://subread.sourceforge.net">http://subread.sourceforge.net</a></td>
<td>Liao et al. [94], 2013</td>
</tr>
<tr>
<td></td>
<td>RSEM</td>
<td><a href="http://deweylab.github.io/RSEM/">http://deweylab.github.io/RSEM/</a></td>
<td>Li and Dewey [95], 2011</td>
</tr>
<tr>
<td></td>
<td>Cufflinks</td>
<td><a href="http://cole-trapnell-lab.github.io/cufflinks/">http://cole-trapnell-lab.github.io/cufflinks/</a></td>
<td>Trappnell et al. [96], 2010</td>
</tr>
<tr>
<td>Normalization</td>
<td>sctransform</td>
<td><a href="https://github.com/ChristophR/sctransform">https://github.com/ChristophR/sctransform</a></td>
<td>Hafemeister and Satija [92], 2019</td>
</tr>
<tr>
<td>Analysis</td>
<td>SCnorm</td>
<td><a href="https://github.com/rhodabacher/SCnorm">https://github.com/rhodabacher/SCnorm</a></td>
<td>Batcher et al. [97], 2017</td>
</tr>
<tr>
<td>pipeline</td>
<td>Linnorm</td>
<td><a href="http://www.jjwanglab.org/linnorm">http://www.jjwanglab.org/linnorm</a></td>
<td>Yip et al. [99], 2017</td>
</tr>
<tr>
<td></td>
<td>SCran</td>
<td><a href="https://rdr.io/bioc/scran/">https://rdr.io/bioc/scran/</a></td>
<td>Lun et al. [96], 2016</td>
</tr>
<tr>
<td></td>
<td>BASICS</td>
<td><a href="https://github.com/catavallejos/BASICS">https://github.com/catavallejos/BASICS</a></td>
<td>Vallejos et al. [94], 2015</td>
</tr>
<tr>
<td></td>
<td>GRM</td>
<td><a href="http://wanglab.ucsd.edu/star/GRM/">http://wanglab.ucsd.edu/star/GRM/</a></td>
<td>Ding et al. [91], 2015</td>
</tr>
<tr>
<td></td>
<td>SAMstrt</td>
<td><a href="https://github.com/shka/R-SAMstrt">https://github.com/shka/R-SAMstrt</a></td>
<td>Katayama et al. [90], 2013</td>
</tr>
<tr>
<td></td>
<td>Seurat</td>
<td><a href="https://github.com/satijalab/seurat">https://github.com/satijalab/seurat</a></td>
<td>Butler et al. [97], 2018</td>
</tr>
<tr>
<td></td>
<td>SCANPY</td>
<td><a href="https://github.com/theislab/Scanpy">https://github.com/theislab/Scanpy</a></td>
<td>Wolf et al. [79], 2018</td>
</tr>
<tr>
<td></td>
<td>scater</td>
<td><a href="https://rdr.io/github/daviscm/scater/">https://rdr.io/github/daviscm/scater/</a></td>
<td>Zhu et al. [198], 2017</td>
</tr>
<tr>
<td></td>
<td>Granatum</td>
<td><a href="https://github.com/lanagarmire/Granatum">https://github.com/lanagarmire/Granatum</a></td>
<td>Gardeux et al. [199], 2017</td>
</tr>
<tr>
<td></td>
<td>ASAP</td>
<td><a href="https://github.com/DeplanckeLab/ASAP">https://github.com/DeplanckeLab/ASAP</a></td>
<td>Lun et al. [202], 2016</td>
</tr>
<tr>
<td></td>
<td>scRat</td>
<td><a href="https://rdr.io/bioc/scrat/">https://rdr.io/bioc/scrat/</a></td>
<td>Guo et al. [93], 2015</td>
</tr>
<tr>
<td></td>
<td>Harmony</td>
<td><a href="https://github.com/immunogenomics/harmony">https://github.com/immunogenomics/harmony</a></td>
<td>Stuart et al. [100], 2019</td>
</tr>
<tr>
<td></td>
<td>scGEN</td>
<td><a href="https://github.com/theislab/scgen">https://github.com/theislab/scgen</a></td>
<td>Korsunsky et al. [101], 2019</td>
</tr>
<tr>
<td></td>
<td>scMerge</td>
<td><a href="https://sydneybiox.github.io/scMerge/">https://sydneybiox.github.io/scMerge/</a></td>
<td>Lotfollahi et al. [200], 2019</td>
</tr>
<tr>
<td></td>
<td>MNN Correct</td>
<td><a href="https://github.com/MarionLab/MNN2017/">https://github.com/MarionLab/MNN2017/</a></td>
<td>Lin et al. [198], 2019</td>
</tr>
<tr>
<td>Batch correction</td>
<td>Expedition</td>
<td><a href="https://github.com/YeoLab/Expedition">https://github.com/YeoLab/Expedition</a></td>
<td>Haghverdi et al. [199], 2018</td>
</tr>
<tr>
<td></td>
<td>BRIE</td>
<td><a href="https://github.com/huanghy09/brie">https://github.com/huanghy09/brie</a></td>
<td>Song et al. [200], 2017</td>
</tr>
<tr>
<td>Alternative splicing</td>
<td>Census</td>
<td><a href="https://github.com/cole-trapnell-lab/monocle-release">https://github.com/cole-trapnell-lab/monocle-release</a></td>
<td>Huang and Sanguinetti [202], 2017</td>
</tr>
<tr>
<td></td>
<td>SingleSplice</td>
<td><a href="https://github.com/jw156605/SingleSplice">https://github.com/jw156605/SingleSplice</a></td>
<td>Qiu et al. [203], 2017</td>
</tr>
<tr>
<td></td>
<td>ccRemover</td>
<td><a href="https://cran.r-project.org/web/packages/ccRemover/index.html">https://cran.r-project.org/web/packages/ccRemover/index.html</a></td>
<td>Welch et al. [204], 2016</td>
</tr>
<tr>
<td></td>
<td>scLVM</td>
<td><a href="https://github.com/PMBio/scLVM">https://github.com/PMBio/scLVM</a></td>
<td>Barron and Li [199], 2016</td>
</tr>
<tr>
<td></td>
<td>COMBAT</td>
<td><a href="https://github.com/Jfortin1/CombatHarmonization">https://github.com/Jfortin1/CombatHarmonization</a></td>
<td>Buetner et al. [200], 2015</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Johnson et al. [202], 2007</td>
</tr>
</tbody>
</table>

HTQC: high-throughput quality control; HISAT: hierarchical indexing for spliced alignment of transcripts; BASICS: bayesian analysis of single-cell sequencing; GRM: gamma regression model; SCANPY: single cell analysis in python; SINCERA: SINgle cell RNA-seq profiling analysis; MNN: mutual nearest neighbors; sclVM: single-cell latent variable mode

heavily on the preprocessing steps. Generic quality control (QC) metrics including FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/), High-Throughput Quality Control [80], or Kraken [81] provide insights into overall quality of raw sequence files. Characterization of heterogeneity is one of the primary purposes of performing single-cell analysis, however not all outliers contribute to unique cell populations. Single-cell specific quality evaluation tools such as SinQC [82], SCell [83], and Celloline [84] enable identification of technical artifacts that interfere with gene expression patterns. Mapping sequencing reads to a reference
genome or transcriptome allows identification of the specific location from which the transcripts originate and subsequent quantification. Although dedicated mappers for scRNA-seq are not available, existing aligners such as TopHat [85], STAR [86], and Hierarchical Indexing for Spliced Alignment of Transcripts (HiSat) [87] have shown considerable precision and accuracy. Recently, two pseudo-alignment tools, Kallisto [88] and Salmon [89], have been proposed which pseudo-align splicing isoforms to a reference transcriptome and overcome the requirement of significant amount of computational power and time to process the reads. An improved Salmon with Selective Alignment and expanded decay sequences was introduced recently and significantly reduced false mappings [90]. However, careful consideration should be made while implementing pseudo-alignment with scRNA-seq data since the data themselves can have lower depth in the first place and 3' coverage bias. Another important step in the analysis pipeline is normalization of the expression data, which is particularly important in single-cell analysis as many technical parameters including cell capture efficiency, drop out events, read depth, and coverage bias can induce variation [91,92]. Tagging individual RNA molecules using UMI’s enables absolute quantification of the transcripts from each cell. In the cases without the use of UMI’s, external spike-in RNAs (i.e., ERCCs) can be used as internal controls. Additionally, several single-cell specific normalization approaches including SAMstr [93], Gamma Regression Model (Gamma Regression Model) [91], sctransform [92], Scran [93], SCnorm [94], Linnorm [95], and Single-Cell Analysis Via Expression Recovery (SAVER) [96,97] can be utilized, of which the last three do not require incorporation of additional spike-ins. Eight commonly applied normalization methods (trimmed mean of M-values [98], count-per-million [100], and DESeq2 [101], as well as others tailored for scRNA-seq, namely scnee, BASICS, SCnorm, Linnorm, and scran) were subject to benchmarking [102]. Data show that scRNA-seq normalization methods outperformed bulk RNA-seq counterparts. However, a bulk RNA-seq normalization method using Differentially Expressed Genes Elimination Strategy is competitive with scRNA-seq normalization methods [103,104]. Three scRNA-seq imputation methods, namely K-Nearest Neighbor smoothing (kNN-smoothing) [105], DrImpute [106], and Single-cell Analysis Via Expression Recovery (SAVER) [107], were evaluated for their capacity to tackle the zero-inflation issue either derived from a technical contribution or a normal distribution [102,108,109] [Table 2]. Both kNN-smoothing and DrImpute analysis gave more reliable results compared with SAVER.

Single-cell RNA isolation at different time points or in different laboratories can induce systemic variations and batch effects which may compromise biologically meaningful interpretation of signals [101,102]. Batch correcting algorithms such as Mutual Nearest Neighbors Correct [111], Seurat 3 [112], Harmony [113], scGen, and scMerge [114], among many others, can compensate the discrepancy. In unsynchronized cells, cell-cycle variation can also mask other important physiological variations, which can be overcome by eliminating cell-cycle factors using packages such as single-cell Latent Variable Model (scLVM) [115] and ccRemover [116]. Drop-out, high number of zero counts, sparsity, and multimodality are some of the unique events encountered in single-cell expression analysis which demand more sophisticated algorithms for identifying differentially expressed genes (DEGs) [117].

Algorithms for scRNA-seq data analysis

Algorithms for scRNA-seq data analysis have been developed recently in different computer languages, such as R program or Python, and few are designed as a website interface or software package [Table 3]. scRNA-seq data are high-dimensional datasets among a great number of cells. Therefore, application of appropriate algorithms is necessary to have better analysis and visualization of scRNA-seq data. After QC and normalization, scRNA-seq data can be processed using diverse algorithms according to variant purposes, such as investigation of DEGs, identification of cell subpopulations, and cell fate trajectories (pseudotime analysis), which are the most common methods to process scRNA-seq data. Visualizations of scRNA-seq data are also diverse. The heatmap is the most common method to present DEGs between groups or within different cell types. Heatmaps are generated by most algorithms for DEGs analysis. T-distributed stochastic neighbor embedding (tSNE), scatter plot, and uniform manifold approximation, and projection (UMAP) are used for visualization of dimension reduction results in cell clustering or cell subpopulations [118-120].
CELL CLUSTERING AND SUBPOPULATION IDENTIFICATION

Cell clustering and cell type identification are critical features of scRNA-seq and, unlike bulk cell RNA-seq, can reveal heterogeneous cell types using entire transcriptomes from an enormous quantity of cells [25,44,121]. Recently, many software algorithms have been developed to achieve cell clustering and cell type identification [Table 3] through unsupervised dimensionality reduction based on principal component analysis (PCA), tSNE, or diffusion maps [28,122]. Based on an unsupervised clustering method, such as Seurat or Monocle 3, novel cell types or populations might be revealed with scRNA-seq data [77,78,123]. Recently, cell type identification of scRNA-seq data has been exponentially applied to studies in developmental biology, neurology, cancer biology, and immunology and can provide the type, quantity, and gene signature of different cell populations [124-129].

DIFFERENTIAL EXPRESSION ANALYSIS

Differential expression analysis can reveal significant DEGs to identify novel pathways or biological functions in different cell types or treatments. Identification of DEGs can be performed by comparing gene expression between two predetermined groups or treatments. For example, Horning et al. [130] identified a group of cell-cycle genes upregulated in a subpopulation which had an attenuated androgen response using Single Cell Differential Expression (SCDE) algorithm in R program. DEGs can also be identified among different cell types in a tissue or organ based on unsupervised algorithms, such as Seurat or Monocle 3. Wang et al. [131] identified single-cell transcriptome profiling of cardiopharyngeal lineages and characterized their cell fate using Seurat package in R program. SCDE [132], PAthway and Gene set OverDispersion Analysis (PAGODA), Model-based Analysis of Single-cell Transcriptomics, Monocle, and SigEMD [134] algorithms and SINgle CEll RNA-seq profiling Analysis [135] workflows have addressed some common challenges to some extent and improved sensitivity in calling DEGs. Since each single cell has the potential to behave as a unique entity, oftentimes the curse of high dimensionality can impose restrictions in clustering and data visualization [122]. UMAP, Zero Inflated Factor Analysis, Single-cell Interpretation via Multikernel LeaRning, and scvis belong to recent dimensionality reduction techniques that can address the underlying confounding factors and enable proper visualization of diverse expression patterns over conventional PCA analysis [126,136,137]. With the advent of automated advanced tools and packages including

### Table 3. Software/packages for single-cell RNA-seq analysis: differential expression, subpopulation identification, clustering, and pseudotime projection

<table>
<thead>
<tr>
<th>Software/package</th>
<th>Differential expression</th>
<th>Clustering cell type</th>
<th>Cell fate trajectories</th>
<th>Language</th>
<th>Programming skill</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>PAGODA</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>R</td>
<td>+++</td>
<td>[133]</td>
</tr>
<tr>
<td>SCDE</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>R</td>
<td>+++</td>
<td>[132]</td>
</tr>
<tr>
<td>Seurat</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>R or R</td>
<td>+++</td>
<td>[77,112]</td>
</tr>
<tr>
<td>SCENIC</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>R or Python</td>
<td>+++</td>
<td>[132]</td>
</tr>
<tr>
<td>Destiny</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>R</td>
<td>++</td>
<td>[206]</td>
</tr>
<tr>
<td>TSCAN</td>
<td>Yes</td>
<td>no</td>
<td>Yes</td>
<td>R or website interface</td>
<td>+</td>
<td>[147]</td>
</tr>
<tr>
<td>Monocle 3</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>R</td>
<td>+++</td>
<td>[123]</td>
</tr>
<tr>
<td>Waterfall</td>
<td>Yes</td>
<td>no</td>
<td>Yes</td>
<td>R</td>
<td>+++</td>
<td>[150]</td>
</tr>
<tr>
<td>Wishbone</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Python</td>
<td>+++</td>
<td>[29]</td>
</tr>
<tr>
<td>GrandPrix</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Python</td>
<td>+++</td>
<td>[207]</td>
</tr>
<tr>
<td>DPT</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>R or Python</td>
<td>+++</td>
<td>[144]</td>
</tr>
<tr>
<td>SCUBA</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>MATLAB</td>
<td>+</td>
<td>[151]</td>
</tr>
<tr>
<td>STREAM</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Python</td>
<td>+++</td>
<td>[145]</td>
</tr>
<tr>
<td>Slingshot</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>R</td>
<td>+++</td>
<td>[148]</td>
</tr>
<tr>
<td>CellRouter</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>R</td>
<td>+++</td>
<td>[208]</td>
</tr>
</tbody>
</table>

PAGODA: pathway and gene set overdispersion analysis; SCDE: single cell differential expression; TSCAN: tools for single cell analysis; DPT: diffusion pseudotime; SCUBA: single-cell clustering using bifurcation analysis; STREAM: single-cell trajectories reconstruction, exploration and mapping.
SingleR and scMatch, cell-type annotations have significantly improved, leading to the identification of rare events or specific cell populations with the ability to scale-up\cite{138-140}. Cell BLAST is a cell type query algorithm for the analysis of new scRNA-seq data. It utilizes a neural network-based generative model to extract low-dimensional cell-to-cell relationships from high-dimensional transcriptomic data and predicts cell types via batch correction with a large-scale curated reference cell type database\cite{141}.

**CELL LINEAGE AND CELL FATE RECONSTRUCTION**

Following cell type identification, cell fate trajectory is the next step to uncover how different cell types coordinate in many aspects of biology including the developmental process or cancer progression\cite{142,143}. Based on transcriptome information, some algorithms provide a pseudotime scale and cell fate branches within all the cells to reveal potential progression or direction of cell types based on cell phenotypic clusters\cite{123,144}. Cell fate trajectory analysis provides an opportunity to investigate the dynamic processes of large-scale cells in developmental processes, cellular differentiation, or drug responses\cite{145,146}. Several software packages can perform trajectory inference. Monocle 3\cite{123}, Diffusion PseudoTime\cite{144}, and Single-cell Trajectories Reconstruction, Exploration And Mapping\cite{145} are well-developed algorithms to perform cell fate trajectory prediction but require a mastery of computer programming skills. Tools for Single Cell ANalysis (TSCAN) provides a friendly webpage interface to access and perform cell fate trajectory\cite{147}.

Of note, transcriptional dynamics represent an important feature of single-cell analysis which enables the analysis of gene expression in given time series such that the output can generate biological signals inferring potential cell lineages. Seurat\cite{77,78}, Slingshot\cite{148}, Monocle 2\cite{149}, Waterfall\cite{150}, Single-cell Clustering Using Bifurcation Analysis\cite{151}, and TSCAN\cite{147} can allow construction of pseudotime trajectory and assessment of expression kinetics that can provide novel insights into cellular differentiation of stem cells as well as oncogenic progression during tumor development. Tian et al.\cite{102} evaluated five trajectory analysis methods in a thorough combination of normalization and imputation of four independent scRNA-seq datasets and found that Slingshot and Monocle 2 led to more robust results. Integration of single-cell transcriptome profiling with other single-cell or bulk analysis and spatial measurements can significantly enhance our understanding of molecular basis of cellular heterogeneity\cite{77} and crosstalk among cellular populations in *in-vivo* studies\cite{152,153}.

**APPLICATIONS OF SCRNA-SEQ**

**Heterogeneity of cell fate determination in embryogenesis**

In the last decade, cumulative efforts have been undertaken to explore the uncharted territory of cellular heterogeneity in different species, organs, tissues, developmental stages, and microenvironments. The first attempt was to interrogate the transitional transcription profiles in the formation of pluripotent embryonic stem cells (ESCs) in early embryogenesis\cite{17}. The differentiational modulation from the inner cell mass to blastocysts and ESCs was not fully understood based on bulk cell studies, but they provided a good initial model for scRNA-seq analysis\cite{17,154}. DEGs showed self-renewal and pluripotency signals with high gene expression variations, particularly for genes with medium expression levels. While epigenetic repressor expression was increased, a suppressive transcription became apparent during the development. A group of miRNAs targeting early differentiation genes and pluripotency genes plays a role in transcriptional alterations. Meanwhile, many spliced forms were discovered for the first time. The same approach was also applied to profile the transcriptional dynamics of earlier embryonic stages, preimplantation human embryos, using 124 cells from oocyte to blastocyst stage\cite{155}. Previous bulk studies have shown that the expressions of ~1900 genes were mainly transcriptionally suppressed during the stages. In this study, about 2,495 and 2,675 genes were significantly up- and downregulated between the four- and eight-cell stages. Splicing isoforms of 4,822 transcripts were enriched in different stages and 20% of transcripts displayed more than two splicing variants. FOXP1 with exon 18b transcripts, ESC-specific splicing species, are 25-fold
more abundant than those with exon 18a in undifferentiated ESCs. From the 90 single embryonic cells analyzed, 64% of the total known human lncRNAs (28,640) were found to be expressed. Another study using Smart-seq method with deep sequencing analyzed the cell fate determination between two- and four-cell embryos and later stage blastomeres. However, this scRNA-seq study discovered that dozens of protein-encoding genes, including Gadd45a, showed significant differential bimodal expression between blastomeres at two- and four-cell embryonic stages. Differential monoallelic expression in 24% genes was clearly observed to be independently regulated in early embryonic development using scRNA-seq. In later embryonic developmental stages within 5-7 days, X-chromosome dose compensation was found in single-cell transcriptomes of 1,529 individual cells from 88 human preimplantation embryos. The cell lineage expression patterns were concurrent as an intermediate status before the establishment of the trophectoderm, epiblast, and primitive endoderm lineages that are contemporary with blastocyst formation. Linnarson’s group studied the differential expression between mouse embryonic stem cells and fibroblasts with a high throughput scRNA-seq method. Nematode embryogenesis between two- and eight-cell stage was dissected with CEL-seq with RNA linear amplification by scoring single-cell transcriptomes. Seventeen genes had significant two-fold mean difference between AB and P1 cells. EMS had more genes expressed compared to P2 cells, while P3 had fewer new genes expressed than C lineage. Taken together, the single-cell transcriptome data map the cell fates in early embryonic differentiation and ESC pluripotency establishment.

**Heterogeneity in complex differentiated tissues and systems**

With multiplexing and high throughputs improvements, scRNA-seq has served as a molecular scalpel directed at the heterogeneity of cells in much more complex tissues, systems, and organisms. The immune system is a complex of bone-marrow-derived differentiated cells. More and more scRNA-seq technologies are adopted for exploring transcriptomes and functional relevance in this biological system. Dendritic cells (DCs) are a group of highly heterogeneous antigen-presenting cells and important for pathogen recognition and immune defense. The bulk RNA-seq of marker-sorted subpopulations did not sufficiently capture their complex functions and led to great controversy. An unbiased global transcriptomic mapping of 18 bone-marrow-derived DCs exposed to lipopolysacharrides (LPS) using Smart-seq revealed hundreds of genes expressed in high variability and unique bimodal profiles that were similarly observed during early mammalian embryogenesis. Among them, 137 genes are anti-virus genes. The spleen is the largest lymphatic organ in the human body. The heterogeneity of 1,536 splenic cells was explored using massively parallel MART-seq with low-depth RNA sampling. From them, the method coupled with a probabilistic mixture model demonstrated sensitive cell classification for distinct identification of B cells, natural killer cells, macrophages, monocytes, and plasmacytoid DCs. In DCs, four subpopulations were found either significantly linked or supported by internal combinatorial marker gene expressions. After exposure to LPS, 1,536 spleen cells’ scRNA-seq displayed the heterogeneity in DCs with enriched CD11c expression and their response to LPS. In the adaptive immune system, differentiation of naïve T cells into T helper 2 (Th2) cells is a feedback loop to restrain immune overreaction. From 91 single Th2 cells acquired post infection of naïve T cells, scRNA-seq revealed unique subpopulations with transcriptional profiles and changes in transcription factors, cytokines, surface receptors, and other pathways.

Mining efforts on heterogeneity of other tissues are ongoing in muscle, lung, intestine, testis, pancreas, and the nervous system. The sci-RNA-seq was applied to profile nearly 50,000 cells from nematodes (Caenorhabditis elegans) with more than 50-fold somatic cellular coverage at the L2 larval stage. From the data, consensus expression profiles for 27 cell types were defined and rare neuronal cell types with one or two cells were sensitively recovered. The global view of regulatory networks for human skeletal muscle myoblast differentiation has been masked by the low resolution of bulk genomic data. ScRNA-seq coupled with a nonlinear MONOCLE pseudotime trajectory prediction model discovered dynamic
expression in 1,061 genes that clustered in gene regulatory groups responsible for activation and suppression at three time points after differentiation initiation. During Embryonic Days 16.5-18.5, murine lung cell lineages at respiratory airway tips are developed from columnar epithelial progenitor cells into flat alveolar type 1 (AT1) or cuboidal type 2 (AT2) cells for gas exchange or surfactant secretion, respectively. A few markers have been identified for four cell types but the global transcriptomic dynamics during the transition is unknown. Microfluidic scRNA-seq of 196 cells have delineated transcriptional signatures for an intermediate bipotential progenitor cells that precede AT1 and AT2 cells, in addition to Clara and ciliated cells. CEL-seq of 238 randomly selected cells from intestinal organoids composed of major intestinal cell lineages brought a better understanding of diversity in intestinal differentiation. Hierarchical clustering of gene expression correlation and rare cell identification method identified the major intestinal cell lineages and 10 clusters as novel diverse subtypes of cells. Spermatogenesis in testes is a complicated and highly orchestrated process including the differentiation of diploid spermatogonia into haploid sperm. The whole picture of spermatogenesis is still far from complete. Two research groups have run scRNA-seq on thousands of dissociated cells from testis samples using Drop-seq and STAR. A conserved continuous temporal trajectory of transcriptional dynamics was consistent in both murine and monkey reproductive models. Novel subpopulations were identified in several time points of differentiation and displayed unique transcriptional regulators and signatures. Based on CEL-seq2 data of pancreatic islet cells from four deceased patients, cell clusters by t-distributed Stochastic Neighbor Embedding (t-SNE) analysis showed the classical pancreatic cell types with marker genes and additional novel markers that have not been reported previously.

The central nervous system is composed of large amounts of neuronal and glial cells with numerous types, and the classical methods to identify them with some molecular markers were limited and not definitive. Single-cell transcripts of ~3000 cells from mouse somatosensory S1 cortex and hippocampus Cornu Ammonis (CA) were analyzed by STRT/C1. Cell type classification identified nine major classes and 47 molecularly distinct subclasses. scRNA-seq of 30,000 nuclei from mouse and human archived brain tissues from hippocampus and prefrontal cortex was carried out by DroNc-seq. With fewer genes detected, cell clustering analysis still identified novel cell types along with well-known cell types.

In other independent studies, there were more than 100 subclasses of cells found in mouse brain and spinal cord. Ribosomes And Intact Single Nucleus (RAISIN) RNA-seq and Mining RAre Cells sequencingMIRACL-seq processed transcriptomes of thousands of neurons in mouse and human enteric nervous system for species-specific transcription signatures and dozens of neuronal subtypes. From 44,808 mouse retinal cells, 39 transcriptionally distinct cell populations were identified, creating an atlas of gene expression for the classification of retinal cells and novel rare subtypes.

**Heterogeneity in cancers**

The transcriptomic heterogeneity of tumors evolves temporospatially during tumor progression with genetic, epigenetic, and tumor immune microenvironmental fluctuations. ScRNA-seq is a powerful tool to address the tumoral heterogeneity, particularly for rare cells and previously unrecognizable subpopulations. Smart-seq was applied to stratify heterogenous cell subpopulations in 672 cells from five glioblastoma tumors. Despite apparent cell-to-cell variability, unbiased cell hierarchical clustering showed four meta-signatures comprised of cell-cycle, hypoxia, complement/immune response, and oligodendritic function. Gene expression profiling of 4,347 cells from six Isocitrate dehydrogenase 1(IDH1) or IDH2 mutant human oligodendrogliomas displayed distinct expression signatures. With bulk exome sequencing and copy number variation estimation, a hierarchical cell lineage map with variant stem/progenitor cell components was delineated in each tumor. Noncanonical WNT activation signaling was noted in retrospective analysis of 77 circulating tumor cells from 13 prostate cancer (PCa) patients following tumor progression compared with stable counterparts undergoing androgen deprivation therapy. This study indicated a potential novel therapeutic target and predictive biomarker for PCa.
From multicellular ecosystem of metastatic melanoma, 4,645 single cells isolated from 19 patients were subject to analysis for profiling malignant, immune, stromal, and endothelial cells\[^{180}\]. The principle component analysis of scRNA-seq data showed that the transcriptomic expression could discern malignant cells from tumor and nonmalignant cells (immune cells, stromal cells, endothelial cells, and fibroblasts) independent of biopsy sites. The transcriptional signatures for malignant cells consist of a core set of cell-cycle genes and a set of immediate early-activation transcription factors that displayed spatial difference. Meanwhile, a drug-resistant subpopulation with high AXL or MITF signals was present in treatment-naive tumors. Treatment-naive tumors are usually sensitive to initial therapy and generally respond to first-line therapy. However, most advanced tumors acquire drug resistance and lead to poor survival outcomes. Androgen deprivation therapy\[^{8}\] is effective for the majority of PCa but biochemical recurrence occurs in 30% of patients subject to treatment, and there is a limited understanding of the underlying mechanisms. From 144 cells treated or untreated with androgen, subpopulations of heterogeneous LNCaP cells were revealed and exhibited high levels of ten cell-cycle-related genes using Smart-seq2 analysis\[^{130}\]. The subpopulations of cells showed cancer stemness phenotype and became resistant to cell-cycle targeting agents. ScRNA-seq and imaging found transcriptional variation and a pre-adapted subpopulation that exhibited resistance to endocrine therapy\[^{181}\]. ScRNA-seq identified a stem-like subpopulation of PCa cells from monolayer and organoid culture\[^{182}\].

Smart-seq2 was deployed to sequence single cells derived from treatment naïve, residual disease, and progressive disease following tyrosine kinase inhibitor (TKI)-based therapies in tumor derived from non-small cell lung cancer patients for mapping transcriptional alterations unique to drug-sensitive and drug-resistant tumor cell populations\[^{183}\]. The scRNA-seq data of 23,261 cells from 49 samples show high-power resolution of high cellular heterogeneity and that residual disease tumors have fewer proliferative markers and increased alveolar cell markers. In TKI-resistant tumors, the upregulated genes were related to oncogenesis and inflammation. Moreover, progressive disease had increased infiltration of immune cells, predominant MF2 macrophages, and suppressive T cells in tumor microenvironments.

Melanoma-associated immune and stromal cells were isolated and analyzed by Smart-seq2 at three time points during tumor development\[^{184}\]. The three temporal subpopulations of stromal cells displayed unique functional signatures. The lymphocytes from lymph nodes underwent activation and clonal expansion in tumors. To map the heterogeneity in the immune cells within hepatocellular carcinoma tumors, scRNA-seq methods were used to study CD45\(^+\) cells isolated from tumors and four immune-relevant sites of 16 treatment-naïve liver cancer patients\[^{129}\]. It was found that LAMP3\(^+\) dendric cells contain unique transcriptional features affecting other immune cell types and show the ability to migrate to lymph nodes. Exhibiting distinct transcriptional states, tumor-associated macrophages were associated with poor prognosis\[^{185}\]. The inflammatory roles of SLC40A1 and GPNMB were clearly demonstrated in these cells.

**CONCLUSION**

Cell heterogeneity has been more appreciated under the light of a new paradigm due to the advances of scRNA-seq and other single-cell analysis technologies. Since its induction, scRNA-seq has been well received and undergone fast-paced technical advances in uniform cDNA amplification, length coverage, rare copy detection, multiplexing, high throughput, processing of metadata, DEG calling, cell clustering, subpopulation identification, and cell fate trajectory predictions. Along with the new technology progress with higher sensitivity and accuracy, our understanding about the extent of cellular heterogeneity has been swiftly updated and repeatedly brought to another level. The discovery of new cell subpopulations and rare cell types with transcriptomic signatures posit new mechanisms for cell functions and defects that lead to novel biomedical applications and rising therapeutic venues.
DECLARATIONS

Authors' contributions

Literature search and manuscript draft: Lieberman B, Kusi M, Hung CN, Chou CW, Chen CL
Tables: Lieberman B, Kusi M, Hung CN, Chen CL
Graphs: He N, Chen CL
Review and revision: Ho YY, Taverna JA, Huang THM, Chen CL

Availability of data and materials

Not applicable.

Financial support and sponsorship

This work was supported by NIH grants U54CA217297 (Huang THM and Chen CL), Cancer Prevention & Research Institute of Texas (CPRIT) grants (RP150600) (Huang THM and Chen CL). Lieberman B and Kusi M were supported by U54 Summer Undergraduate Scholar grant and the CPRIT predoctoral training grant (RP170345), respectively.

Conflicts of interest

All authors declared that there are no conflicts of interest.

Ethical approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Copyright

© The Author(s) 2021.

REFERENCES

53. Kivioja
51. Picelli
50.
49.
48. Islam
46. Kurimoto
45. Kurimoto
44.
42.
41.
37.
36.
35.
34.
33. Macosko
30. Bendall
27.
24.
22.
21.
Page 16                          Lieberman
Biotechnol
2015;161:1187-201.
2010;5:516-35.
single cells.
cDNA library construction.
microarray analysis.
EMT-related genes in metastatic prostate cancer.
applications.
Fu GK, Hu J, Wang PH, Fodor SP . Counting individual DNA molecules by the stochastic attachment of diverse labels.
92. Katayama S, Töhönen V, Linnares K, Kere J. SAMstr: statistical test for differential expression in single-cell transcriptome with spike-
96. Lun AT, Bach K, Marioni JC. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. Genome Biol 2016;17:75.


161. Hume DA. Applications of myeloid-specific promoters in transgenic mice support in vivo imaging and functional genomics but do not support the concept of distinct macrophage and dendritic cell lineages or roles in immunity. *J Leukoc Biol* 2011;89:525-38.


169. Hammoud SS, Low DH, Yi C, et al. Transcription and imprinting dynamics in developing postnatal male germline stem cells. *Genes Dev*


