Journal of Advanced Research 71 (2025) 141–153

Contents lists available at ScienceDirect

Journal of Advanced Research

journal homepage: www.elsevier.com/locate/jare

Original Manuscript

# Systematic evaluation of single-cell RNA-seq analyses performance based on long-read sequencing platforms

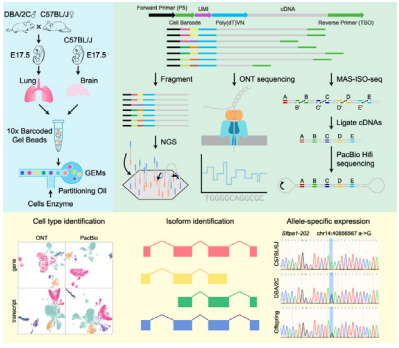
Enze Deng [a,b,1], Qingmei Shen [b,c,1], Jingna Zhang [b], Yaowei Fang [b], Lei Chang [c], Guanzheng Luo [a], Xiaoying Fan [b,c,*]

[a] MOE Key Laboratory of Gene Function and Regulation, Guangdong Province Key Laboratory of Pharmaceutical Functional Genes, State Key Laboratory of Biocontrol, School of Life Sciences, Sun Yat-sen University, Guangzhou 510275, China
[b] Guangzhou National Laboratory, No. 9 XingDaoHuanBei Road, Guangzhou International Bio Island, Guangzhou 510005, Guangdong Province, China
[c] GMU-GIBH Joint School of Life Sciences, The Fifth Affiliated Hospital of Guangzhou Medical University, Guangzhou Medical University, Guangzhou 510005, China

## HIGHLIGHTS

- The TGS-based scRNA-seq data could be independently used to generate singlecell gene/isoform expression matrices.
- Although the gene detection sensitivity is relatively low due to limited sequencing throughput, the TGS-based scRNA-seq accurately captures all cell types.
- PacBio demonstrates superior performance in discovering novel transcripts.
- Both TGS techniques were able to determine the allelic origins of the transcript reads, and PacBio could specify more allele-specific transcripts.

## GRAPHICAL ABSTRACT




## ARTICLE INFO

*Article history:*
Received 26 October 2023
Revised 23 April 2024
Accepted 20 May 2024
Available online 22 May 2024

*Keywords:*
Single cell RNA-sequencing
Third-generation sequencing
Cell type identification
Novel isoform
Allele-specific gene/isoform

## ABSTRACT

*Introduction:* The rapid development of next-generation sequencing (NGS)-based single-cell RNA sequencing (scRNA-seq) allows for detecting and quantifying gene expression in a high-throughput manner, providing a powerful tool for comprehensively understanding cellular function in various biological processes. However, the NGS-based scRNA-seq only quantifies gene expression and cannot reveal the exact transcript structures (isoforms) of each gene due to the limited read length. On the other hand, the long read length of third-generation sequencing (TGS) technologies, including Oxford Nanopore Technologies (ONT) and Pacific Biosciences (PacBio), enable direct reading of intact cDNA molecules.
*Objectives:* Both ONT and PacBio have been used in conjunction with scRNA-seq, but their performance in single-cell analyses has not been systematically evaluated.
*Methods:* To address this, we generated ONT and PacBio data from the same single-cell cDNA libraries containing different amount of cells.
*Results:* Using NGS as a control, we assessed the performance of each platform in cell type identification. Additionally, the reliability in identifying novel isoforms and allele-specific gene/isoform expression by

* Corresponding author at: Guangzhou National Laboratory, No. 9 XingDaoHuanBei Road, Guangzhou International Bio Island, Guangzhou 510005, Guangdong Province, China.
*E-mail address:* fan_xiaoying@gzlab.ac.cn (X. Fan).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.jare.2024.05.020
2090-1232/© 2024 The Authors. Published by Elsevier B.V. on behalf of Cairo University
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

both platforms was verified, providing a systematic evaluation to design the sequencing strategies in single-cell transcriptome studies.
*Conclusion:* Beyond gene expression analysis, which the NGS-based scRNA-seq only affords, TGS-based scRNA-seq achieved gene splicing analyses, identifying novel isoforms. Attribute to higher sequencing quality of PacBio, it outperforms ONT in accuracy of novel transcripts identification and allele-specific gene/isoform expression.
© 2024 The Authors. Published by Elsevier B.V. on behalf of Cairo University This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

## Introduction

The advent of single-cell sequencing technology has revolutionized the field of genomics and transcriptomics, facilitating the acquisition of genome and transcriptome sequencing data at the resolution of individual cells [1]. It has overcome the limitations of traditional sequencing methods in resolving cell populations, providing us with an unprecedented opportunity to investigate biological systems and understand cellular heterogeneity [2]. Single-cell sequencing has found extensive applications in the fields of development [3], aging [4], and disease [5], enabling us to unravel molecular mechanisms underlying cell differentiation and fate determination [6], unveil key signaling pathways and molecular markers involved in the aging process [7], and uncover critical mutations and pathways driving tumor development [8].

The development of single-cell sequencing technology has undergone pivotal milestones and breakthroughs. Significantly, notable progress has been achieved in single-cell isolation and capture techniques, such as microfluidic chips [9] and droplet-based [10,11] methods, facilitating efficient separation and capture of individual cells. However, the short-read based high-throughput scRNA-seq methodologies are limited in their ability to capture comprehensive information about transcript structure and diversity due to their focus on a small region near one end of the transcript [12]. This limitation hampers the analysis of splicing events, chimeric transcripts, and sequence variations across the entire molecule. In contrast, long-read sequencing offers a unique advantage by sequencing the entire RNA (cDNA) molecule without the need for fragmentation or read length constraints [13]. Consequently, full-length transcripts can be captured in single reads, enabling accurate characterization and quantification of transcript isoforms at the isoform level [14].

On the other hand, TGS technologies showed significant drawbacks in data throughput, sequencing accuracy, and prices. Thus, previous studies almost utilized the TGS in parallel with NGS to investigate the single-cell transcriptome. For instance, by combing NGS sequencing and PacBio ISO-Seq, cell-type specific isoforms have been identified in mouse cerebellum [15]. Additionally, researchers have utilized NGS to guide the correction of cell barcode (CB) sequences generated by ONT in mouse brain [16]. Dondi et al. revealed many novel transcripts in tumors, including gene fusions that had been misclassified in short-read transcriptome data [17]. Furthermore, TGS has been proven efficient in detecting haplotype expression, and provides a robust foundation for further investigations into various biological systems at single-cell resolution level [18].

Both ONT and PacBio have been applied in single-cell studies, but how they performed in different aspects of single-cell analyses is not assessed. In this study, we employed the two TGS technologies with 10x genomics scRNA-seq and compared them to the benchmark NGS using exact the same cDNA libraries of different sample sizes. The TGS platforms generated more consistent gene expression profiles compared with NGS. In detail, ONT generated more cDNA reads, but PacBio showed better performance in CB identification. Both TGS technologies performed better than NGS in cell annotation with a small cell sampling size. Nevertheless, they showed bias toward different gene isoforms. In addition, although more genes were captured in each individual cell when the input cells were fewer in a library, more cell type-specific molecules could be identified in a larger cell sample size with low gene detection sensitivity. PacBio identified a greater amount of cell type specific genes and isoforms, and is more facilitate in the analysis of haplotype expression. The novel isoforms identified using PacBio data were more accurate. Taken together, we conducted a methodical assessment to offer researchers guidance on the selection of suitable TGS platforms for single-cell studies.

## Materials and methods

### *Animals and single cell suspension preparation*

We used 6- to 8-week-old C57BL/6J female mice to mate the DBA/2C male mice to obtain the lung tissues from the offsprings at E17.5. The brains were collected from E17.5 C57BL/6J mice. The tissues were cut into pieces and digested in 2U/mL papain and 200U/mL DNase I at 37℃ for 20 min (lung) or 10 min (brain), and pipetted sufficiently with 1ml tips every 5 min. After digestion, cells were passed through a 40 μm cell strainer on ice and centrifuged at 500g for 5 min at 4℃. Resuspend cells with Red Blood Cell Lysis Buffer and incubated for 1 min at room temperature to remove the red blood cells. Resuspend cells with cold 1x PBS with 0.25% BSA after centrifugation and washed again. Trypan blue staining solution and cells mixed in equal proportions to count cell numbers and assess cell variability.

### *Single-cell transcriptome amplification and NGS library preparation*

The disassociated cells were captured on the 10x Genomics Chromium controller according to the Chromium Next GEM Single Cell 3′ Reagent Kits v3.1 User Guide (10x Genomics CG000204 Rev D) with the following modification: PCR cycles were increased, from the recommended ten cycles for recovery of 500 and 5000 cells, to 16 cycles to target a yield of cDNA enabling simultaneous NGS, ONT and PacBio library preparation. We used Qsep100 to assess size distribution of cDNA fragments and Libraries and Qubit 4.0 fluorometer for concentration. NGS library was sequenced on MGISEQ2000 sequencer using a 100 PE run mode with read1 28 cycles, read2 100 cycles, and i7 index 8 cycles.

### *ONT and PacBio library preparation*

ONT libraries were prepared with 350 ng of input cDNA using SQk-LSK114 kit and sequencing was performed on PromethION sequencer (Oxford Nanopore Technologies, Oxford, UK) for 72 h. PacBio libraries were prepared using MAS-ISO-seq (Multiplexed Arrays Sequencing) for 10x single cell 3′kit with 100 ng of input cDNA. According to the procedure for constructing single-cell MAS-ISO-seq libraries with the following modification. After 16 parallel cDNA amplification reactions, we used 0.4x Ampure XP

beads to generate >1kp DNA fragments. Sequencing was performed on the PacBio sequel II System with the Sequel Sequencing Kit 3.0 chemistry for 36 h to collect data.

### *NGS data preprocess and analyses*

The 10x CellRanger pipeline (v7.0.1) [12] was run on the raw sequencing data to obtain single-cell expression matrices. The obtained raw_feature_bc_matrix was loaded and scanpy [19] workflow was further used for downstream analyses. The quality of cells was assessed based on three metrics: (1) the number of total UMI count per cell (library size) was above 1,000; (2) the number of detected genes was above 500; (3) the percentage of mitochondrial genes was below 20. Next, we used a cluster-level approach to remove potential doublet cells. In brief, the doublet score was calculated for each cell using Scrublet [20] (v0.2.3). After quality control, a total of 5,699 cells comprising 425 cells from 5h samples and 5,274 cells from 5k samples were retained for downstream analysis. We applied the total-count correction (library-size correction) and normalized the data matrix to 10,000 reads per cell, so that such normalized counts became comparable among cells. The logarithmic data matrix was obtained using the *scanpy.pp.log1p* function, and the *scanpy.pp.highly_variable_genes* function was performed to select the highly variable genes (HVGs). Then, effects of the total counts per cell and the percentage of mitochondrial gene counts were regressed out by the *scanpy.pp.regress_out* function. A principal component analysis (PCA) matrix with 50 components was calculated to reveal the main axes of variation and denoise the data by using the *scanpy.tl.pca* function.

### *Generation of circular consensus reads*

Using the default SMRT-Link parameters, we performed circular consensus sequencing (CCS) with IsoSeq3 with the following modified parameters: -hifi-kinetics -min-passes 3 -min-rq 0.99 -max-length 100,000 -min-length 100 -chunk 10/30 -j 8.

### *Preprocessing of ONT and PacBio data*

Nanopore reads were identified, oriented, and trimmed by Pychopper (v2.7.2) (https://github.com/epi2me-labs/pychopper) using the following parameters: -m edlib -b primer.fa -c primer.txt -q 0.5 -p. We obtained the full-length transcripts according to the forward primer (CTACACGACGCTCTTCCGATCT) and reverse primer (CCCATGTACTCTGCGTTGATACCACTGCTT) on both ends of the reads. CBs and UMIs were extracted from the sequences following the forward primer, encompassing 16 and 12 base pairs respectively using a custom script. In essence, we searched for polyT sequences within the 200bp region following the forward primer. The identified polyT sequences were required to have a minimum length of 10bp and a minimum T content of 0.75. The sequence extending from the end of ployT motif to the reverse primer was considered as a full-length transcript. Full-length transcripts shorter than 50bp were filtered out from further analysis.

PacBio CCS reads were annotated using Longbow (v0.6.14) [21] (via Longbow annotate) with MAS-ISO-seq adapters, and reads with mis-ordered MAS-ISO-seq adapters were filtered out (via Longbow filter). Subsequently, each read was segmented (via Longbow segment) based on the model and the 10× Genomics adapter. The forward primer, BC, UMI, and polyT location were further refined using a custom script. Leading and trailing adapter sequences were eliminated (via Longbow extract), and the cDNA segment of each read were then transformed into fastq files using samtools (v1.17) [22].

### *Alignment of ONT and PacBio full length transcript fragments*

FLTs obtained from ONT sequencing were aligned to the reference genome (GRCm39) using minimap2 (v2.24) [23] with the following parameters: -ax splice -MD -eqx -secondary=no -t 20 -y -Y -L. Similarly, FLTs derived from PacBio sequencing were aligned with the parameters: -ax splice:hq -MD -eqx -secondary=no -t 20 -y -Y -L.

### *Cell barcode error correction*

Correcting for potential errors in cell barcodes is a pivotal stage in the analysis of single-cell data, and our approach for achieving this is outlined in below. Initially, each long read was annotated with its original cell barcode, utilizing the preprocessing methodology outlined in the previous description. Subsequently, starcode (v1.4) [24] was applied to cluster these barcodes with parameters: -s, -d *dist*, and -print-clusters. Notably, the parameter *dist* was set at 1 for PacBio and 2 for ONT. Barcodes were clustered by size in starcode, with centroids encapsulating all corresponding matches. These clustered cell barcodes were then rectified to their canonical forms, which were determined based on the highest counts. Any raw barcodes that correlated with a canonical barcode listed in the 10x whitelist file underwent correction.

### *UMI error correction*

Reads were first partitioned into distinct groups, wherein reads sharing identical CB and Ensembl gene ID were grouped together. UMI correction was then executed individually for each resulting read group. We formulated UMI correction process as a maximum in-degree problem within a directed graph denoted as G = (V, E), where V represented the set of UMIs and E denoted the set of edges. Given a set R encompassing reads within a specific group, the collection of vertices in V was defined as comprising all unique combinations of three-tuple (UMI, cDNA length, and GC content) derived from the reads in R. Consequently, directed edges $(s, t) \in E$ were created, linking source vertex s with target vertex t, subject to the satisfaction of three conditions: (1) the Levenshtein distance between the UMI of s and t was limited to no greater than 1 for PacBio reads and 2 for ONT reads, (2) the disparity in cDNA length between s and t was constrained to no more than 50 bp for PacBio reads and 100 bp for ONT reads, and (3) the discrepancy in GC content between s and t was capped at 0.05 for PacBio reads and 0.1 for ONT reads. The constraint parameters were established to reflect the occurrence rate of indel sequencing errors and the observed empirical distributions of cDNA lengths and GC content within intragroup reads that shared identical UMIs. Under these constraints, a directed edge extending from a source to a target vertex denoted the potential that they represented the same molecular entity. With the graph thus derived, an iterative greedy strategy was implemented to identify the maximum subset of target vertices in V that comprehensively covered all read vertices. Commencing with the initial assignment, each iteration involved selecting the target vertex in V with the highest degree, indicating the greatest number of supporting reads. Consequently, the UMIs of the reads within the group were corrected to align with the UMI exhibiting the highest in-degree. To address potential variations due to distinct PCR copies and sequencing errors associated with the same UMI, we implemented a strategy that designating the read with the longest length and the highest alignment score within the group as the primary UMI.

### *Transcript assignment using IsoQuant*

Read-to-isoform mapping was performed using IsoQuant (v3.2.0) [25], employing a general feature file (GFF) annotation

based on GENCODE M30. PacBio transcript assignment utilized the following parameters: -data_type pacbio, while Nanopore transcript assignment employed: -data_type ont. Only reads mapping unambiguously to a single gene or isoform were subsequently extracted and sorted for further gene and isoform analyses.

*Generation of the single-cell gene count matrix*

This procedure aligns with the CellRanger pipeline. Subsequent to the error correction of CBs and UMIs, reads that satisfy the following criteria are considered for UMI counting: (1) possession of a valid UMI, (2) possession of a valid 10x barcode, (3) having the longest length and the highest alignment score, and (4) being confidently assigned to one gene or one isoform. The cell barcode calling algorithm employs a threshold based on total UMI counts for each CB to identify cells. The order of magnitude algorithm (OrdMag) estimates the number of recovered cells such that barcodes are called cells if total UMI counts exceed m/10, where m represents the 50th percentile of top N barcodes based on total UMI counts.

*ScRNA-Seq data processing*

The analysis of single-cell gene and isoform expression matrices for each sample was conducted using Scanpy (v1.9.3) [19]. Specifically, the raw UMI matrix was processed to exclude cells with fewer than 100 genes, and genes/isoforms detected in less than 10 cells. We further quantified the gene and UMI counts for each cell and retained high quality cells with thresholds of 1,000 UMIs and less than 10% mitochondrial gene counts. Considering mitochondrial and hemoglobin genes, as well as *Malat1* genes, primarily contribute to technical aspects, we removed them from the dataset prior to subsequent analysis. Subsequently, Scrublet (v0.2.3) [20] was employed to eliminate potential doublets with the expected doublet rate of 6%. Normalized expression matrix was calculated by normalizing the raw UMI counts to the total counts per cell using a normalization factor of 1e4, followed by logarithmic transformation.

*Dimension reduction and unsupervised clustering*

Dimension reduction and unsupervised clustering followed a standard approach within the Scanpy [19] framework. Initially, the identification of highly variable genes (HVGs) was conducted using the dispersion-based methods, where the normalized dispersion is obtained by scaling with the mean and standard deviation of the dispersions for genes falling into a given bin [26]. Unwanted sources of variation including total counts and percentages of mitochondrial gene counts were further regressed out from the normalized expression matrices. Principal component analysis (PCA) was performed on the variable gene matrix to reduce noise, with the top 50 principal components retained for subsequent analyses. A neighborhood graph of observations was constructed to estimate the connections between data points [27]. Subsequently, Leiden algorithm was applied to these nearest neighbor graphs to detect communities and identify cell clusters [28]. Notably, the same principal components were also used for non-linear dimension reduction, generating the Uniform Manifold Approximation and Projection (UMAP) visualization. Following the initial round of unsupervised clustering, each cell cluster was assigned an annotation based on canonical cell markers, allowing for the identification of the major cell types, including epithelial cells, endothelial cells, mesenchymal cells, muscle cells and immune cells. The cell type identification of cluster in the mouse brain utilized the *cell annotation* function of the CellMarker 2.0 database [29], with the top 30 differentially expressed genes from each cluster serving as input, and the reference results serving as the primary subtypes. For cell type specific gene and isoform marker identification, as well as differential expression analyses, the *scanpy.tl.rank_genes_groups* function was employed, with the method parameter set to 'wilcoxon'.

*Validation of novel isoforms*

Novel isoforms were validated via RT-PCR. Total RNA was extracted from E17.5 lung using the RNeasy Mini Kit (QIAGEN. cat. nos. 74,104 and 74106). To wipe out gDNA, RT reaction was carried out with HiScript III RT SuperMix for qPCR (+gDNA wiper) kit (Vazyme. R323-01). Specific primers for the PCR template were designed using NCBI Primer-BLAST tool. The PCR amplified product was analyzed by agarose gel electrophoresis and the expected bands were further recovered and sequenced through Sanger sequencing.

*Detection of single nucleotide polymorphism (SNP) using WGS data*

We obtained approximately 30G and 27G of whole-genome sequencing data from DBA/2C and C57BL/J mice using MGISEQ2000 sequencer. To ensure data quality, we used fastp (v0.23.3) [30] with default parameters to filter out low-quality reads. The resulting high-quality reads were then aligned to the mouse genome (GENCODE, M30) using bwa mem (v0.7.17) [31]. After removing duplicate reads, GATK4 HaplotypeCaller (v4.4.0.0) [32] was utilized to call SNPs. We considered SNPs with a depth of $\geq$10 as strain-specific SNPs, resulting in a total of 3,928,849 and 3,513 homozygous SNPs for DBA2/C and C57BL/J.

*Calling SNPs from single-cell TGS data with DeepVariant*

Genomic variants were called from the ONT output bam file using DeepVariant (v1.5.0) [33] with the argument -model_type ONT_R104. Similarly, genomic variants were called from the PacBio output BAM file with the argument -model_type PACBIO. Variants with a read depth below 10 were excluded, ensuring a minimum coverage threshold for accurate variant calling. Heterozygous SNPs were then extracted for subsequent analysis.

*Identification of allelic-specific transcripts*

Classification of FLTs as DBA/2C or C57BL/6J allelic-specific was determined based on the following conditions: (1) if a read consisted of no SNP sites in the gene, it was classified as 'unclassifiable'; (2) if a read contained SNP sites in the gene but did not cover any of these SNP sites, it was classified as 'unknown'; (3) if a read contained only one strain-specific SNP, it was assigned to the corresponding strain; (4) if a read had multiple strain-specific SNPs, it was assigned to one strain only if the number of SNPs assigned to this strain was at least twice that assigned to the other strain; (5) in cases where the number of SNPs did not meet the above conditions, the read was classified as 'unassignable'.

*Ethics statement*

All animal experiments were performed according to the guidelines of the Guangzhou Institutes of Biomedicine and Health (Guangzhou, China).

## Results

### *Quality of scRNA-seq data generated from different sequencing platforms*

We collected E17.5 lung and brain samples for single-cell transcriptome amplification, yielding either 500 (5h) or 5,000 (5k) cells (see methods). Subsequently, cDNAs were divided into three groups, subjected to NGS, ONT, and PacBio sequencing platforms (Fig. 1a). The data processing steps are illustrated in Supplementary Fig. 1a. Generally, the reads quality of PacBio was comparable to that of NGS (around Q30), and ONT reads were predominantly below Q20 (Fig. 1b-d). Full-length transcript fragments (FLTs) in ONT reads, characterized by the two cDNA amplification primers (P5 and TSO in Fig. 1a) with correct orientation, were subsequently extracted (Supplementary Fig. 2a). FLTs in PacBio reads were extracted using the official software Longbow [21]. Only about half of the ONT reads were obtained with FLTs (Supplementary Fig. 2b). In contrast, majority of the PacBio reads can be split and derived FLTs (Fig. 2a).

For CB detection, PacBio outperformed NGS in perfect matching to the whitelist (Fig. 2b). In contrast, less than 60% FLTs captured by ONT contained CBs perfectly matched to the whitelist in both lung and brain samples. However, gene expression correlation coefficients of biological duplicates were not affected by sequencing quality of reads. All sequencing batches generated highly reproducible gene expression profiles for the same type of samples. The most important influencing factor of correlation coefficients between biological replicates turned to be sequencing depth, where the fewest mapped reads in the sample generated the lowest correlation values (Table 1, Supplementary Fig. 3a). For the same samples, although NGS with the highest sequencing coverage, the two TGS platforms generated more consistent gene expression profiles with each other (Supplementary Fig. 3b).

Next, we evaluated cell capture efficiency of different sequencing platforms. For the same cDNA libraries, all the platforms generated similar CB curves (Supplement Fig. 4a, b), and the number of detected cells was mainly related to the number of mapped cDNA reads. Meanwhile, it is worth mentioning that although PacBio generated the fewest cDNA reads for each library, the detected gene number in single cells is higher than that of ONT (Supplement Fig. 4c). For each sequencing technology, the detected gene numbers of single cell were positively correlated with the cell mapped counts, except for a subset of ONT_lung_5h, ONT_brain_5h and PacBio_brain_5h cells, which contained many mapped counts but detected with few genes (Supplement Fig. 4d). This might be par-

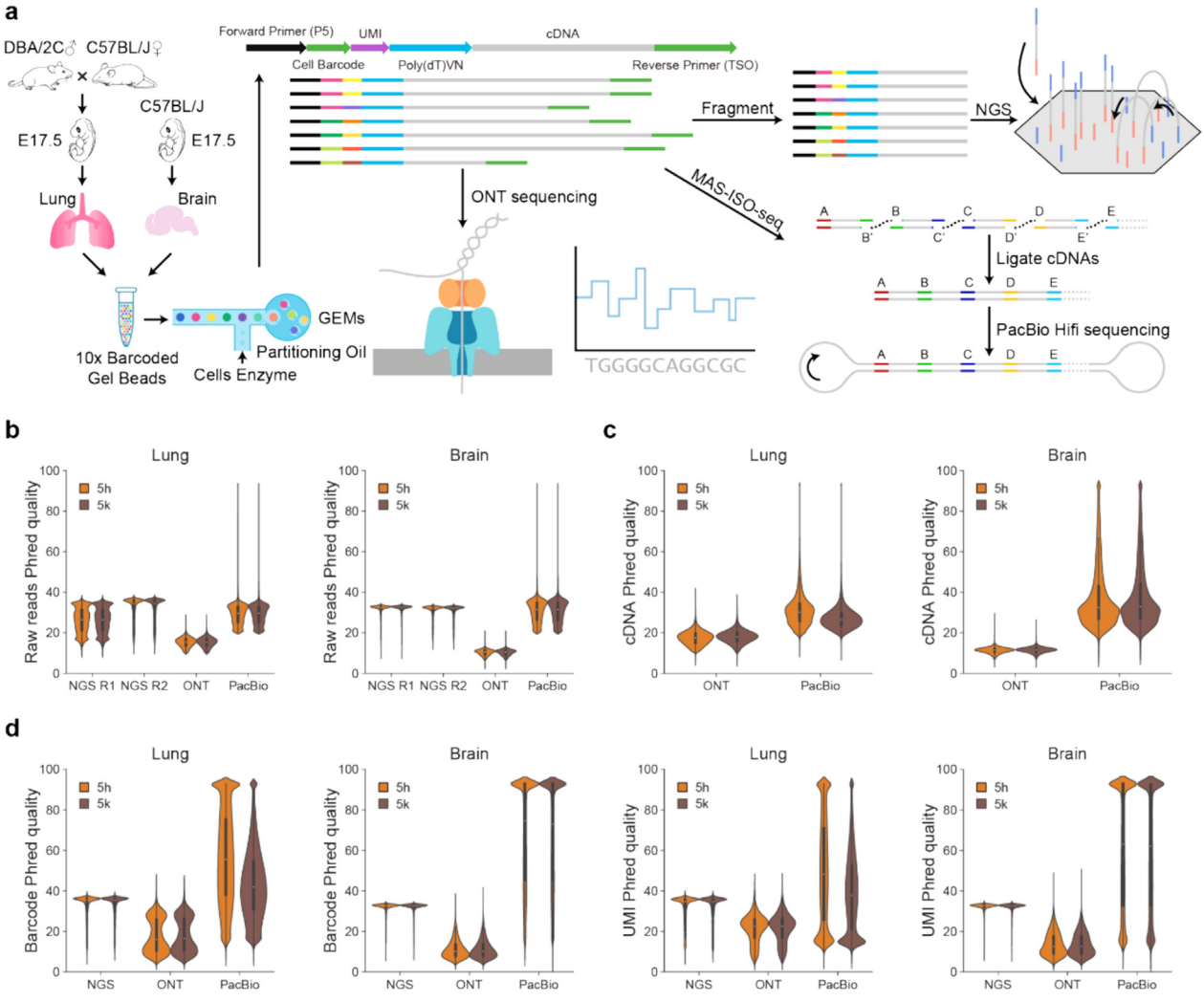



**Fig. 1.** Overview of the workflow and the quality of sequencing techniques. (a) Schematic diagram of experimental workflow for scRNA-seq of mouse lung samples. (b) Phred quality scores of raw reads from NGS, ONT, and PacBio platforms. (c) Phred quality scores of the whole cDNA reads from ONT and PacBio. (d) Phred quality scores of CB and UMI sequences from NGS, ONT and PacBio.

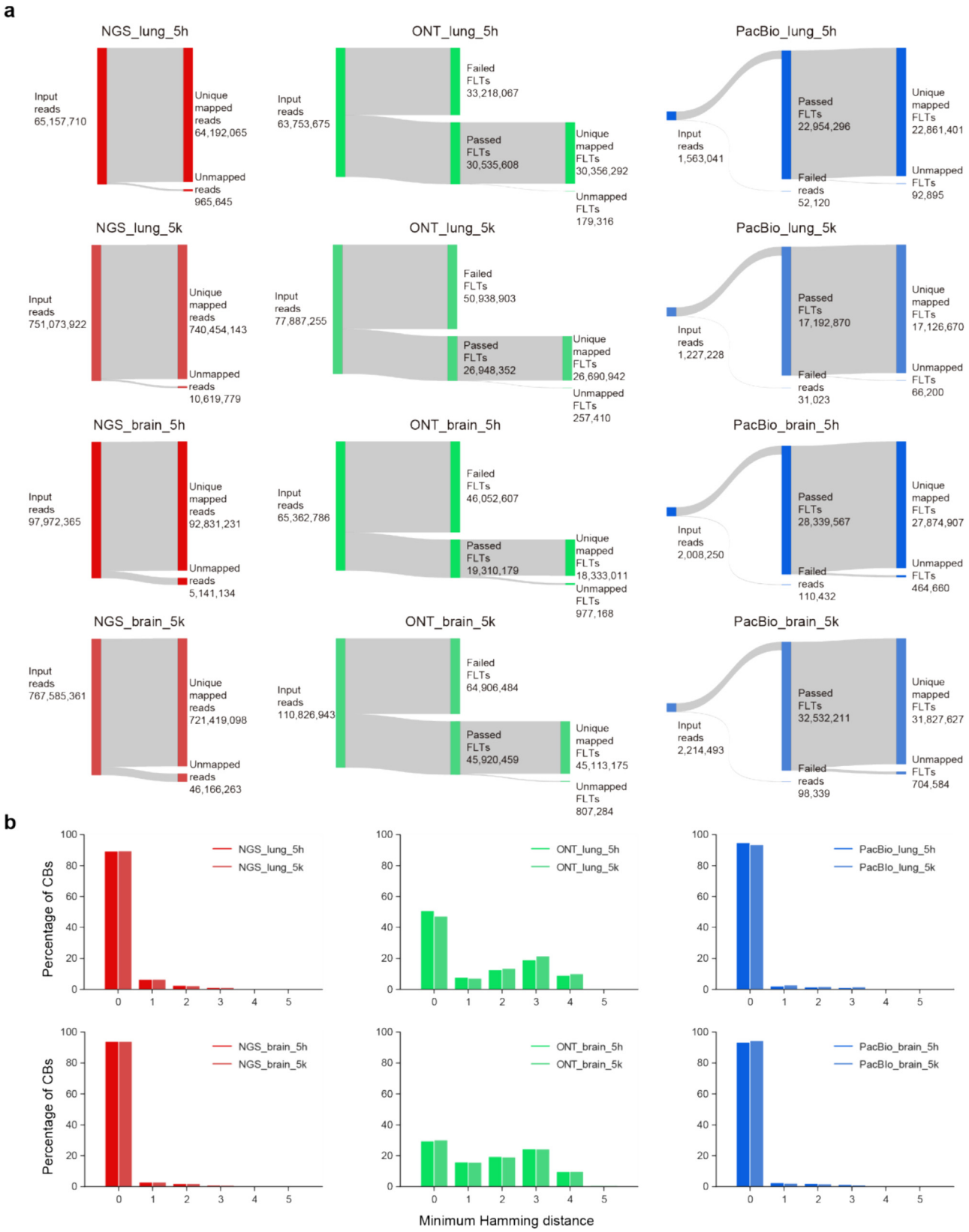



**Fig. 2.** Visualization and evaluation of data processing. (a) Sankey plot depicting the flow of reads at each processing step. (b) Distribution of minimum Hamming distance between detected CB sequences and those in whitelist across the three methods.

tially due to the miscalling of CBs in ONT data (Fig. 2b). For the lung_5k library, both TGS platforms exhibited relatively shallow sequencing depths, resulting in a median detection of only ~700 genes for ONT and ~900 genes for PacBio in individual cells (Supplement Fig. 4c). The correlation of gene expression profiles of the same cells measured by different sequencing technologies was

**Table 1**
Sequencing performance of NGS, ONT and PacBio for scRNA-seq samples.

| Batch | Tissue | Sample | Sequencing Platform | No. of Total Reads (Pairs, M) | Average Length of Reads | Yields (Gb) | Number of cDNA Reads (M) | Unique Mapping Ratio |
|---|---|---|---|---|---|---|---|---|
| NGS_lung_5h | Lung | 5 h | MGISEQ-2000 | 65.2 | 50 + 100 | 9.8 | – | 98.52 % |
| NGS_lung_5k | | 5 k | | 751.1 | 50 + 100 | 112.7 | – | 98.59 % |
| ONT_lung_5h | | 5 h | Nanopore | 63.8 | 1168 | 74.5 | 30.5 | 99.41 % |
| ONT_lung_5k | | 5 k | Promethion | 77.9 | 1205 | 93.8 | 26.9 | 99.04 % |
| PacBio_lung_5h | | 5 h | Pacbio Sequel lle | 1.6 | 21,380 | 33.4 | 23.0 | 99.60 % |
| PacBio_lung_5k | | 5 k | | 1.2 | 21,269 | 26.1 | 17.2 | 99.61 % |
| NGS_brain_5h | Brain | 5 h | MGISEQ-2000 | 98.0 | 28 + 100 | 12.5 | – | 94.75 % |
| NGS_brain_5k | | 5 k | | 767.6 | 28 + 100 | 98.2 | – | 93.99 % |
| ONT_brain_5h | | 5 h | Nanopore | 65.4 | 1022 | 66.8 | 19.3 | 94.94 % |
| ONT_brain_5k | | 5 k | Promethion | 110.8 | 1082 | 119.9 | 45.9 | 98.24 % |
| PacBio_brain_5h | | 5 h | Pacbio Sequel lle | 2.0 | 13,613 | 27.3 | 28.3 | 98.36 % |
| PacBio_brain_5k | | 5 k | | 2.2 | 14,748 | 32.6 | 32.5 | 97.83 % |

generally higher in the 5h samples than the 5k samples, suggesting that deeper sequencing depth per cell improved the consistency of gene expression measurement across sequencing technologies (Supplement Fig. 4e). Interestingly, ONT and PacBio exhibited a stronger correlation in 5h cells compared to NGS, while the 5k sample showed the opposite trend. Downsampling of ONT_5h and PacBio_5h data to 10% of their original levels revealed a significant decrease in pairwise single-cell gene expression correlation between the TGS platforms, falling below the correlation seen with NGS platforms (Supplement Fig. 4f).

### *Cell type identification with TGS-based scRNA-seq in mouse lung*

Unsupervised clustering and cell type identification was carried out using single-cell gene count matrices from all three sequencing methods and single-cell isoform count matrices from the two TGS methods. We identified five cell types in all the datasets, except that the muscle cells were not identified in the NGS_lung_5h sample (Fig. 3a). These cells identified in the TGS datasets (both gene and isoform level clustering) were clustered into mesenchymal cells in the NGS_lung_5h sample (Fig. 3b). Also, a few NGS defined epithelial cells in the 5h sample were assigned as mesenchymal cells in the two TGS datasets (Fig. 3b). This suggested the TGS preforms better in cell type identification when the sequenced cell number is limited. With increase sample size (as in the lung_5k sample), the NGS captured the cell types more efficiently (Fig. 3a). The same cells were more consistently assigned of the cell type identities between NGS and PacBio. Although the TGS showed low gene detection ability in single cells of the lung_5k sample, they did not cause effect on cell clustering and cell type identification at both gene expression and isoform expression level (Fig. 3a). In all the datasets, epithelial cells were highly detected with *Epcam* expression, endothelial cells specifically expressed *Pecam1*, mesenchymal cells expressed the marker gene *Col1a1*, and the muscle cells highly expressed *Tnnt2*. Interestingly, although three sequencing batches captured immune cells, the classical marker gene *Ptprc* (CD45) was more uniformly detected in the NGS datasets (Fig. 3c, Supplement Fig. 5). These results provide insights into the differences in gene expression profiles captured by the NGS and TGS technologies.

Although the expression levels of cell type marker genes for each cell type were nearly identical between the two TGS technologies, there were differences at the isoform expression level. For example, PacBio, but not ONT, detected more *Tnnt2-205* and *Tnnt2-209* transcripts in muscle cells (Fig. 3d, Supplement Fig. 6). The same cells were almost identical to the same cell type in both TGS platforms (Fig. 3e). We identified half more cell type differentially expressed genes (DEGs) (except the muscle cells) using NGS within a given sample. The PacBio data were obtained with more cell type DEGs than the ONT data (Fig. 3f). Such superiority of PacBio was also shown when analyzing the cell type specific isoforms (DEIs) (Fig. 3g), suggesting its advantages in capturing the molecular diversity between cell types. Therefore, ONT and PacBio scRNA-seq could successfully corroborate cell type identifications, aligning with NGS findings, based on both gene and isoform expression, but the DEGs and DEIs were not exactly the same. In addition, we also noticed that each cell type was obtained with more DEGs when more cells were captured in a dataset (Fig. 3f). But more cells did not provide more DEIs, probably due to limited depth of single cells in a larger sample size from TGS (Fig. 3g).

### *Cell type identification with TGS-based scRNA-seq in mouse brain*

On both gene and transcripts expression levels, six distinct cell types were identified in mouse brain samples: the most abundant excitatory neurons further specified into deep-layer (*Bcl11b*$^+$) and up-layer (*Satb2*$^+$), radial glias (*Fabp7*$^+$), intermediate progenitors (*Eomes*$^+$), inhibitor neurons (*Gad2*$^+$) and microglia cells (*Tmem119*$^+$) (Fig. 4a, c). Notably, microglia cells were absent in the brain_5h sample (Fig. 4a). This may be attributed to the limited proportion of microglia in the fetal mouse brain which can not be recovered in the reduced cell count libraries. The same to the observations in lung samples, a small number of NGS-defined cells were clustered as other cell types in both TGS platforms. For example, a few radial glias and inhibitory neurons were clustered as excitatory neurons respectively (Fig. 4b). On the other hand, the marker genes' expressions were more specific in the TGS datasets (Fig. 4c). ONT still displayed more pronounced differences compared to NGS than PacBio (Fig. 4b). There were no difference between the two TGS platforms in the isoform expression patterns of the brain cell type marker genes, although slight differences in cell clustering persisted (Fig. 4d-e). These findings highlight subtle differences between NGS and TGS technologies. Consistent with observations in mouse lung samples, more cell type DEGs and DEIs were obtained when more cells were captured in a dataset. NGS yielded the highest number of DEGs, followed by PacBio, with ONT exhibiting the lowest cell type DEGs and DEIs (Fig. 4f-g).

### *Accuracy of isoform identification by different TGS platforms*

To evaluate the ability of detecting isoforms by each TGS method, we sorted genes by the number of detected isoform types. In both TGS sequencing methods, some genes were detected with over 100 isoform types, and over 1000 genes were detected with at least 10 isoform types (Fig. 5a), indicating the importance of expression analyses at the isoform level. ONT datasets were assigned with more isoform types for the top genes ranked by the number of isoforms detected. Then, we analyzed the isoform

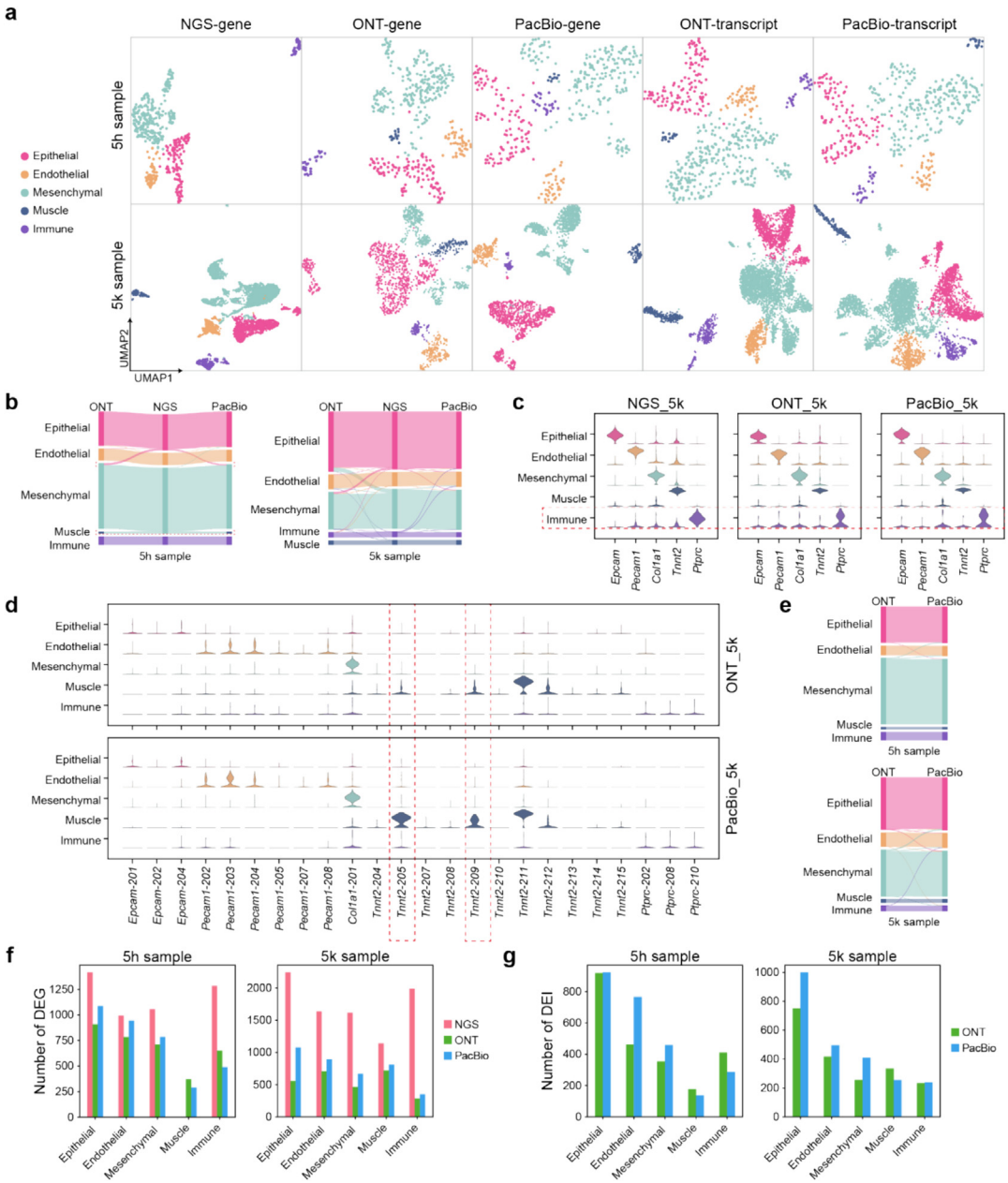



**Fig. 3.** Cell type identification of lung sample using different sequencing technologies. (a) UMAPs showing cell types in different samples and sequencing technologies. Top and bottom rows represent samples contain 500 and 5,000 cells, respectively. The five columns represent different sequencing technologies at gene level and isoform level. (b) Sankey diagram illustrating the consistency of cell identities identified by gene expression across three methods. (c) Violin plots showing expression level of cell type marker genes detected by each sequencing technology in the 5 k sample. Left: NGS; Center: ONT; Right: PacBio. The red dotted box highlights lower efficiency of TGS in detecting immune cell marker *Ptprc*. (d) Violin plots showing expression level of isoforms of cell type marker genes in the 5 k sample. Top: ONT; bottom: PacBio. The red dotted boxes highlight that the two isoforms were highly detected by PacBio but lowly detected by ONT. (e) Sankey diagram illustrating the consistency of cell identities identified by isoform expression between ONT and PacBio. (f) Bar plots depicting the number of cell type DEGs identified using different sequencing methods. Left: 5 h sample; right: 5 k sample. (g) Bar plots depicting the number of cell type DEIs identified using ONT and PacBio; Left: 5 h sample; right: 5 k sample.

types in each cell type using IsoQuant [25]. Reads were classed into consistent match, inconsistency and misalignment, based on their alignment with the reference sequences (Fig. 5b, Supplementary Table 1). More consistent transcripts were identified in ONT data, which could be attributed to the use of less stringent processing parameters as recommended. Furthermore, we compared the uniquely annotated transcripts and found slightly higher ratios of full splice matching (FSM) transcripts for each cell type in PacBio data (Fig. 5c, Supplementary Fig.7a), suggesting the possibility of more splicing noise or artificial errors in ONT. For novel transcript identification, we found that both methods captured similar types of novel transcripts (Supplementary Fig.7b-e). Generally, a higher number of novel transcripts were observed by PacBio, and there was greater concordance in the novel transcripts identified between different methods within the same sample (Fig. 5d).

We then evaluated the accuracy of novel isoforms captured by each TGS method. We randomly chose novel isoforms specifically identified in each technology for experimental validation using RT-PCR and Sanger sequencing (Supplementary Table 2). A higher ratio of PacBio-specific (11/15) than ONT-specific (3/14) novel transcripts was validated (Fig. 5e, Supplementary Table 2), suggesting higher accuracy of PacBio in identifying novel transcripts.

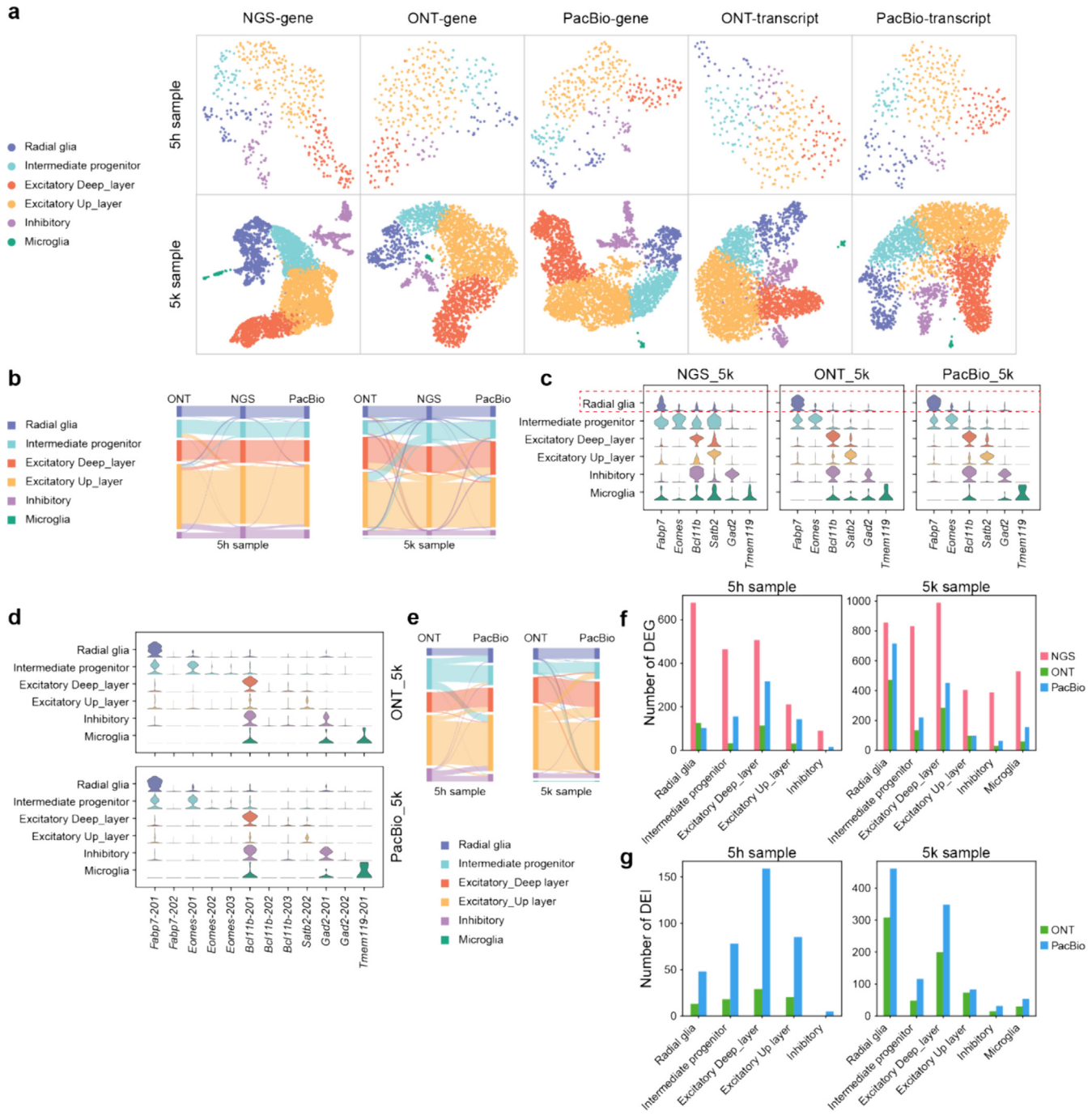



**Fig. 4.** Cell type identification of brain sample using different sequencing technologies. (a) UMAPs showing cell types in different samples and sequencing technologies. Top and bottom rows represent samples contain 500 and 5,000 cells, respectively. The five columns represent different sequencing technologies at gene level and isoform level. (b) Sankey diagram illustrating the consistency of cell identities identified by gene expression across three methods. (c) Violin plots showing expression level of cell type marker genes detected by each sequencing technology in the 5 k sample. Left: NGS; Center: ONT; Right: PacBio. (d) Violin plots showing expression level of isoforms of cell type marker genes in the 5 k sample. Top: ONT; bottom: PacBio. (e) Sankey diagram illustrating the consistency of cell identities identified by isoform expression between ONT and PacBio. (f) Bar plots depicting the number of cell type DEGs identified using different sequencing methods. Left: 5 h sample; right: 5 k sample. (g) Bar plots depicting the number of cell type DEIs identified using ONT and PacBio; Left: 5 h sample; right: 5 k sample.

These results support the utility of TGS technologies in unraveling the complexity of transcriptomic landscapes, highlighting the better performance of PacBio in novel transcript discovery.

*Exploring allele-specific expression with TGS-based scRNA-seq*

Studying allele-specific gene expression enables us to delve into the genetic regulatory mechanisms that govern gene expression, leading to a more profound comprehension of genetic variation, individual differences, gene regulatory networks, evolution, and adaptability [34–36]. The NGS-based scRNA-seq is difficult to achieve such analyses as only a few dozen bases could be captured for a transcript sequence. The FLTs directly read by TGS afford us the opportunity to do such investigation. As we crossed two different strains of mice (DBA/2C and C57BL/6J, Fig. 1a) for sampling, we could figure out the allele-specific reads according to the single

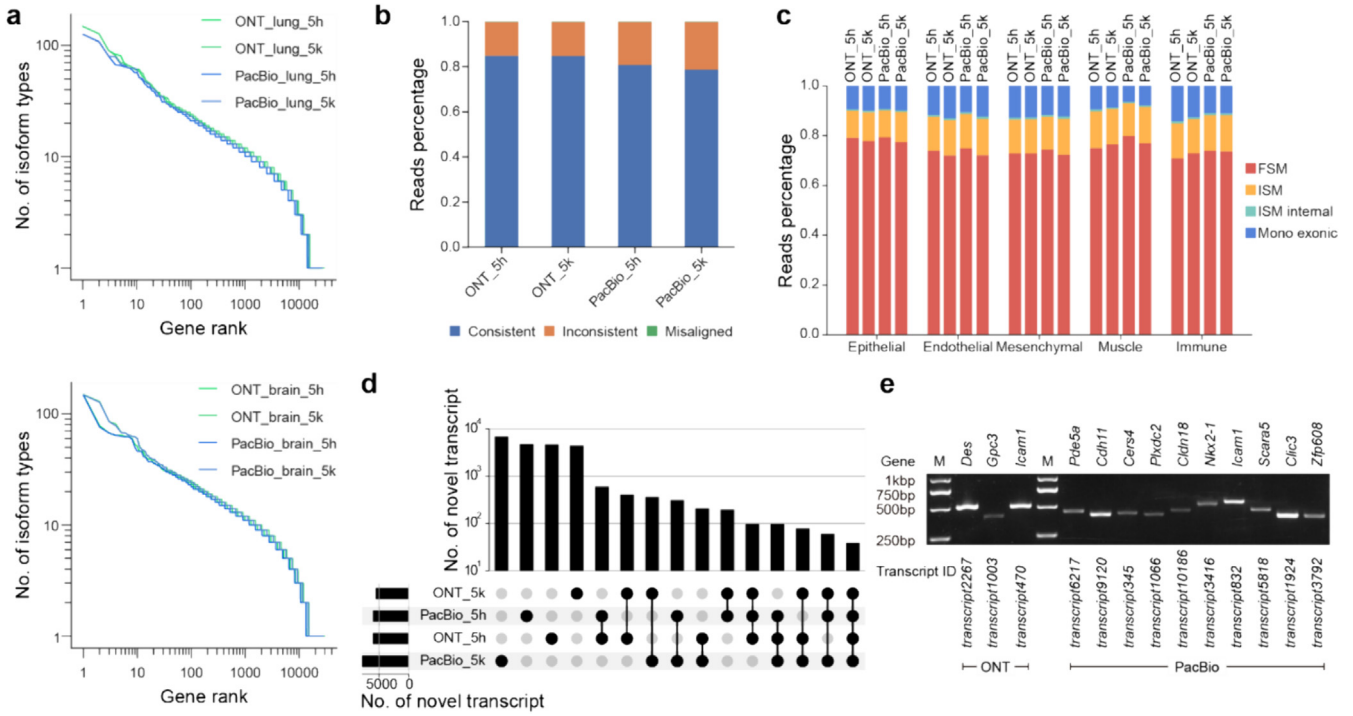



**Fig. 5.** Varied transcripts diversity captured by ONT and PacBio. (a) Genes ranked based on the count of isoforms in each sample. (b) Bar plots showing the proportions of different isoform types of isoforms in each sample. Consistent isoform schemes are shown in Fig. S7a, inconsistent isoform schemes are shown in Fig. S7b and misalignment isoform schemes are shown in Fig. S7d. (c) Bar plots showing the proportions of different isoform subtypes belonging to consistent classification in each sample. (d) Upset plot displaying the overlap of identical novel transcripts in each sample. (e) Gel image of the validated novel transcripts.

nucleotide polymorphisms (SNPs) between the two strains. A total of ~1.77 million SNP sites were identified in the gene region by comparing the whole genome sequencing data of the two mouse strains (see methods). Comparing to the reference genome (GRCm39), we identified over 120 thousand SNP sites in the PacBio samples. For the ONT samples, we obtained about 70 thousand SNP sites, 70% of which were shared with PacBio samples (Supplement Fig. 8a, b). Consequently, more reads (~45%) in PacBio data were confidently assigned to allelic origin than those in ONT data (less than 40%) for each cell type (Fig. 6a, Supplement Fig. 8c). Unknown reads were defined as those that did not cover any SNP sites, and unassigned reads included SNP sites but could not be attributed to a specific allele. The DBA/2C and C57BL/6J genomes transcribed equal amounts of total RNAs in each cell type (Fig. 6a, Supplement Fig. 8c), reflecting little bias in calculating allele-specific gene expression using both TGS methods.

Nevertheless, there were also DEGs and DEIs between the two parental alleles. PacBio detected more allelic DEGs and DEIs than ONT did (Fig. 6b, Supplementary Table 3). We further compared the allelic DEGs and DEIs in each sample, observing that PacBio detected larger differences (Supplementary Fig. 8d). Interestingly, there were 71 genes that showed no allelic expression difference at the gene level, while the corresponding isoforms were differentially expressed between the two alleles. For instance, Sftpa1, a classical marker gene of alveolar epithelial cells, contains two isoform types (Supplementary Fig. 8e). PacBio detected no differences in expression from both alleles for the whole gene, but the C57BL/6J allele expressed more Sftpa1-201 transcripts and the DBA/2C allele expressed more Sftpa1-202 transcripts (Fig. 6c). This indicated that the SNPs on the gene may affect the splicing of the gene, which was previously reported in TPO [37]. To validate whether there was isoform expression bias of Sftpa1 between the two alleles, we did RT-PCR using the lung tissues from the two strains of mice and the hybridized offspring. Indeed, Sanger sequencing revealed nested peaks in the offspring sample at SNP loci, confirming comparable amount of transcripts from both alleles (Fig. 6e). While for the Sftpa1-202 isoform, the DBA/2C allele (chr14:40856567 G) expressed a higher level than the C57BL/6J allele (chr14:40856567 A). Another alveolar epithelial marker gene, Sftpb, was detected with globally more transcripts from the DBA/2C allele (Fig. 6d). All three SNPs revealed absolutely higher expression of DBA/2C allele (Fig. 6f). Together, these results suggest that there are biased gene transcription and splicing between the two alleles of mouse genome, which can be confidently captured by TGS based scRNA-seq.

## Discussion

The two TGS platforms have different sequencing advantages and shortcomings. ONT showed much higher sequencing error rate, although the company has made effort to improve the sequencing quality by improving the sequencing chip to R10 version [38]. Thus, the ONT read quality is still far from accurately revealing CBs and UMIs in scRNA-seq. Such conclusion is consistent with a previous amplicon sequencing study [39]. In this way, we have to sacrifice the accuracy in determining cell belonged reads, which further affected the DEGs and DEIs. PacBio generates a limited amount of CCS data for cDNA libraries because only one molecule is read per sequencing pore. The company employed MAS-ISO-seq to conjunct 15 cDNA molecules in a sequencing molecule [21]. At least ten more times of cDNA reads could be obtained through such improvements. Although PacBio MAS-ISO-seq still gets fewer FLTs than ONT, the cell type analyses were better performed as more accurate cDNA reads were assigned. As sequencing throughput and experimental strategies have advanced to enhance the data output of TGS, an increasing number of studies have adopted TGS platforms to explore isoform-level regulations [13].

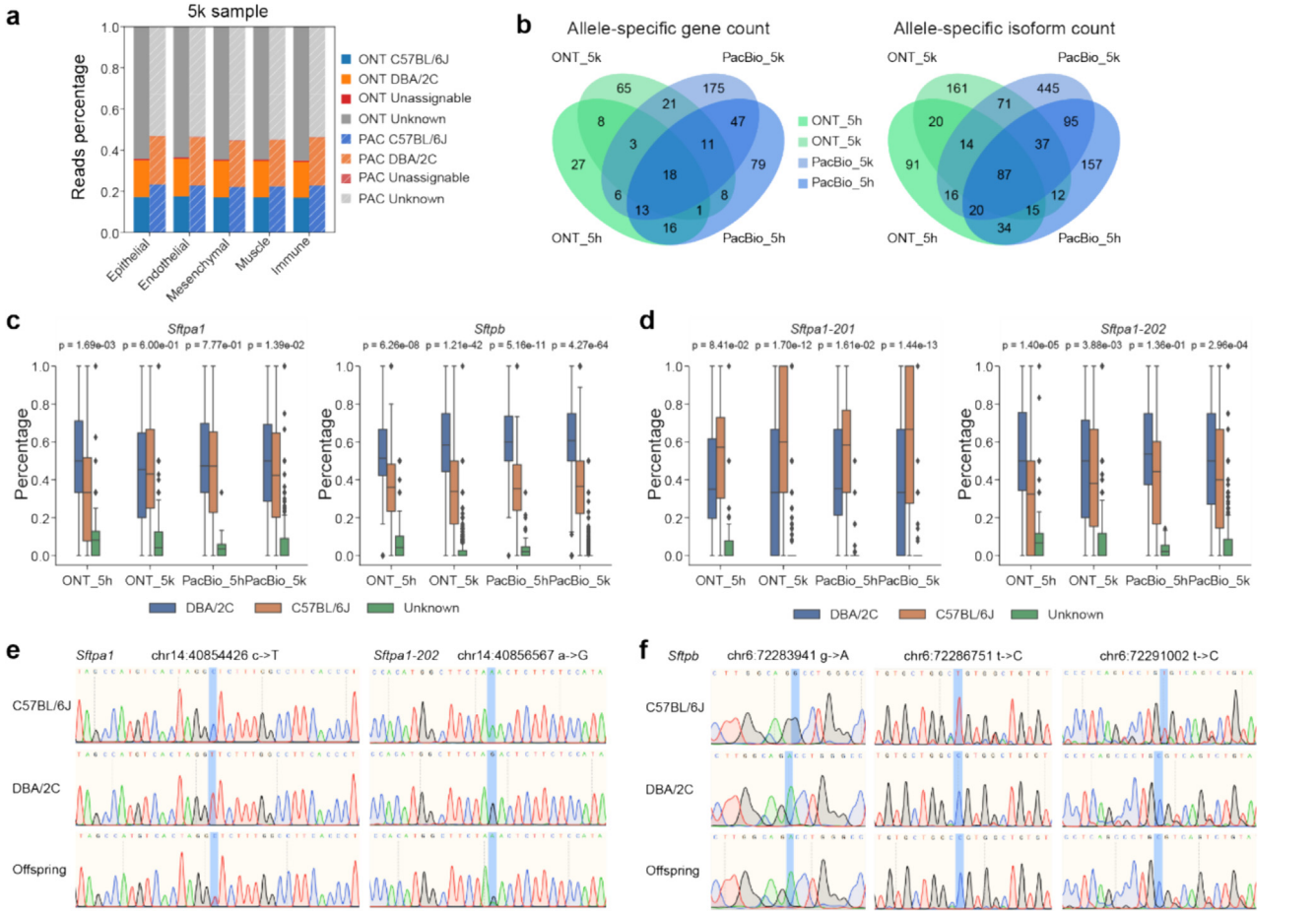



**Fig. 6.** Allele DEGs/DEIs identified by ONT and PacBio. (a) Proportions of detected allelic transcripts using ONT and PacBio across different cell types. (b) Venn diagram illustrating the count of allele DEGs/DEIs identified in each sample. (c) Box plots illustrating the ratios of Sftpa1 and Sftpb transcripts originated from each allele in each sample. (d) Box plots illustrating the ratios of Sftpa1-201 and Sftpa1-202 isoforms originated from each allele in each sample. (e, f) Sanger sequencing results of *Sftpa1*, *Sftpa1-202* and *Sftpb*. The shadow marks the distinguish SNPs. The genome loci of the SNPs are listed on the top and lowercase represents the base in C57BL/6J, the uppercase represents the base in DBA/2C.

**Table 2**
Cost comparison of three sequencing methods.

| Metric | NGS_5h | NGS_5k | ONT_5h | ONT_5k | PacBio_5h | PacBio_5k |
|---|---|---|---|---|---|---|
| Cost | 1705.97 | 1705.97 | 3067.08 | 3067.08 | 3717.08 | 3717.08 |
| Cell num | 710 | 6861 | 617 | 5864 | 569 | 5351 |
| Cost per cell | 2.40 | 0.25 | 4.97 | 0.52 | 6.53 | 0.69 |
| Times | 1 | 1 | 2.07 | 2.08 | 2.72 | 2.76 |

Since TGS has limited sequencing throughput, we have to determine the appropriate sample size in a scRNA-seq library. In our comparison, although fewer cells in a sample achieved more genes for each cell, the cell clustering and cell type identification were not affected by the low gene/isoform coverage in the 5k sample. Moreover, more cells with lower gene coverage could generate more cell type specific DEGs and DEIs than fewer cells with higher gene coverage. Thus, we recommend more cells in a library rather than more genes detected in each cell under a given sequencing yield. The cell number is not unlimited (better no more than 5000 cells in our sequencing systems).

Additionally, more cells in a library can decrease the cost for each cell. In our case, based on 10x genomics single-cell amplification of the 5k samples, the total cost for NGS was $0.25 per cell (Table 2). For ONT and PacBio, the cost was $0.52 and $0.69 per cell, respectively. Based on our evaluation, both ONT and PacBio achieved comparable results to NGS on cell type analyses, with only a twofold increase in cost. Moreover, they offered supplementary insights into isoforms and allelic expressions. The reduction in TGS-based scRNA-seq price would enable more researchers to afford relevant studies, thereby facilitating broader scientific exploration and discovery.

In summary, the future of TGS-based scRNA-seq is promising, with potential improvements in throughput, longer read lengths, enhanced accuracy, cost reduction, computational advancements, and expanded applications. These advancements will undoubtedly contribute to its application in cellular diversity, disease mechanisms, and many intricate biological processes to achieve wider scientific exploration and discovery in relevant fields.

## Conclusion

The TGS technologies offer significant advantages in isoform analyses because of their long read lengths. However, due to the limited sequencing throughput and accuracy, TGS was seldom

used for scRNA-seq analyses. According to direct comparison, both TGS platforms achieved comparable, even higher accuracy on cell type identification based on both gene expression and isoform expression. Moreover, beyond gene expression analysis, which the NGS-based scRNA-seq only affords, TGS-based scRNA-seq achieved gene splicing analyses, identifying novel isoforms. Possibly attribute to higher sequencing quality of PacBio, it outperforms ONT in accuracy of novel transcripts identification and allele-specific gene/isoform expression. These findings not only underline the potency of TGS-based scRNA-seq, but also emphasize its pivotal role in elucidating the intricacies of single-cell transcriptome regulation.

## Compliance with ethics requirements

All animal experiments were performed according to the guidelines of the Animal Ethics Committee of Guangzhou National Laboratory (GZLAB-AUCP-2022-08-A01).

## CRediT authorship contribution statement

**Enze Deng:** Data curation, Visualization, Writing – original draft. **Qingmei Shen:** Methodology, Validation. **Jingna Zhang:** Methodology. **Yaowei Fang:** Methodology. **Lei Chang:** Investigation. **Guanzheng Luo:** Investigation. **Xiaoying Fan:** Conceptualization, Resources, Supervision, Writing – review & editing.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgements

We would like to thank the Advanced Cell Technology Core Facility, Guangzhou National Laboratory for offering us the sequencing service in this study. X.F. was supported by grants from the National Key Research and Development Program of China (2020YFA0112201), the Guangdong Provincial Pearl River Talents Program (2021QN02Y747), the Major Project of Guangzhou National Laboratory (GZNL2024A03001) and the R&D Program of Guangzhou National Laboratory (SRPG21-001, ZL-SRPG2201704).

## Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jare.2024.05.020.

## References

[1] Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, et al. mRNA-Seq whole-transcriptome analysis of a single cell. Nat Methods 2009;6:377–82. doi: https://doi.org/10.1038/nmeth.1315.
[2] Svensson V, Vento-Tormo R, Teichmann SA. Exponential scaling of single-cell RNA-seq in the past decade. Nat Protoc 2018;13:599–604. doi: https://doi.org/10.1038/nprot.2017.149.
[3] Sun X, Perl A-K, Li R, Bell SM, Sajti E, Kalinichenko VV, et al. A census of the lung: Cell Cards from LungMAP. Dev Cell 2022;57:112–145.e2. doi: https://doi.org/10.1016/j.devcel.2021.11.007.
[4] Allen WE, Blosser TR, Sullivan ZA, Dulac C, Zhuang X. Molecular and spatial signatures of mouse brain aging at single-cell resolution. Cell 2023;186:194–208.e18. doi: https://doi.org/10.1016/j.cell.2022.12.010.
[5] Sauler M, McDonough JE, Adams TS, Kothapalli N, Barnthaler T, Werder RB, et al. Characterization of the COPD alveolar niche using single-cell RNA sequencing. Nat Commun 2022;13:494. doi: https://doi.org/10.1038/s41467-022-28062-9.
[6] Shiraishi K, Shah PP, Morley MP, Loebel C, Santini GT, Katzen J, et al. Biophysical forces mediated by respiration maintain lung alveolar epithelial cell fate. Cell 2023;186:1478–1492.e15. doi: https://doi.org/10.1016/j.cell.2023.02.010.
[7] Wang S, Yao X, Ma S, Ping Y, Fan Y, Sun S, et al. A single-cell transcriptomic landscape of the lungs of patients with COVID-19. Nat Cell Biol 2021;23:1314–28. doi: https://doi.org/10.1038/s41556-021-00796-6.
[8] Sinjab A, Han G, Treekitkarnmongkol W, Hara K, Brennan PM, Dang M, et al. Resolving the Spatial and Cellular Architecture of Lung Adenocarcinoma by Multiregion Single-Cell Sequencing. Cancer Discov 2021;11:2506–23. doi: https://doi.org/10.1158/2159-8290.CD-20-1285.
[9] Mazutis L, Gilbert J, Ung WL, Weitz DA, Griffiths AD, Heyman JA. Single-cell analysis and sorting using droplet-based microfluidics. Nat Protoc 2013;8:870–91. doi: https://doi.org/10.1038/nprot.2013.046.
[10] Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, et al. Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells. Cell 2015;161:1187–201. doi: https://doi.org/10.1016/j.cell.2015.04.044.
[11] Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell 2015;161:1202–14. doi: https://doi.org/10.1016/j.cell.2015.05.002.
[12] Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun 2017;8:14049. doi: https://doi.org/10.1038/ncomms14049.
[13] Amarasinghe SL, Su S, Dong X, Zappia L, Ritchie ME, Gouil Q. Opportunities and challenges in long-read sequencing data analysis. Genome Biol 2020;21:30. doi: https://doi.org/10.1186/s13059-020-1935-5.
[14] Zhu C, Wu J, Sun H, Briganti F, Meder B, Wei W, et al. Single-molecule, full-length transcript isoform sequencing reveals disease-associated RNA isoforms in cardiomyocytes. Nat Commun 2021;12:4203. doi: https://doi.org/10.1038/s41467-021-24484-z.
[15] Gupta I, Collier PG, Haase B, Mahfouz A, Joglekar A, Floyd T, et al. Single-cell isoform RNA sequencing characterizes isoforms in thousands of cerebellar cells. Nat Biotechnol 2018;36:1197–202. doi: https://doi.org/10.1038/nbt.4259.
[16] Lebrigand K, Magnone V, Barbry P, Waldmann R. High throughput error corrected Nanopore single cell transcriptome sequencing. Nat Commun 2020;11:4025. doi: https://doi.org/10.1038/s41467-020-17800-6.
[17] Dondi A, Lischetti U, Jacob F, Singer F, Borgsmüller N, Tumor Profiler Consortium, et al. Detection of isoforms and genomic alterations by high-throughput full-length single-cell RNA sequencing for personalized oncology. bioRxiv 2022. Doi: 10.1101/2022.12.12.520051.
[18] Fan X, Tang D, Liao Y, Li P, Zhang Y, Wang M, et al. Single-cell RNA-seq analysis of mouse preimplantation embryos by third-generation sequencing. PLoS Biol 2020;18:e3001017.
[19] Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. Genome Biol 2018;19:15. doi: https://doi.org/10.1186/s13059-017-1382-0.
[20] Wolock SL, Lopez R, Klein AM. Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. Cell Syst 2019;8:281–291.e9. doi: https://doi.org/10.1016/j.cels.2018.11.005.
[21] Al'Khafaji AM, Smith JT, Garimella KV, Babadi M, Popic V, Sade-Feldman M, et al. High-throughput RNA isoform sequencing using programmed cDNA concatenation. Nat Biotechnol 2023. doi: https://doi.org/10.1038/s41587-023-01815-7.
[22] Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. GigaScience 2021;10:giab008. doi: https://doi.org/10.1093/gigascience/giab008.
[23] Li H. New strategies to improve minimap2 alignment accuracy. Bioinformatics 2021;37:4572–4. doi: https://doi.org/10.1093/bioinformatics/btab705.
[24] Zorita E, Cuscó P, Filion GJ. Starcode: sequence clustering based on all-pairs search. Bioinformatics 2015;31:1913–9. doi: https://doi.org/10.1093/bioinformatics/btv053.
[25] Prjibelski AD, Mikheenko A, Joglekar A, Smetanin A, Jarroux J, Lapidus AL, et al. Accurate isoform discovery with IsoQuant using long reads. Nat Biotechnol 2023. doi: https://doi.org/10.1038/s41587-022-01565-y.
[26] Satija R, Farrell JA, Gennert D, Schier AF, Regev A. Spatial reconstruction of single-cell gene expression data. Nat Biotechnol 2015;33:495–502. doi: https://doi.org/10.1038/nbt.3192.
[27] McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction 2018. Doi: 10.48550/ARXIV.1802.03426.
[28] Traag VA, Waltman L, van Eck NJ. From Louvain to Leiden: guaranteeing well-connected communities. Sci Rep 2019;9:5233. doi: https://doi.org/10.1038/s41598-019-41695-z.
[29] Hu C, Li T, Xu Y, Zhang X, Li F, Bai J, et al. Cell Marker 2.0: an updated database of manually curated cell markers in human/mouse and web tools based on scRNA-seq data. Nucleic Acids Res 2023;51:D870–6. doi: https://doi.org/10.1093/nar/gkac947.
[30] Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 2018;34:i884–90. doi: https://doi.org/10.1093/bioinformatics/bty560.
[31] Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009;25:1754–60. doi: https://doi.org/10.1093/bioinformatics/btp324.

[32] Poplin R, Ruano-Rubio V, DePristo MA, Fennell TJ, Carneiro MO, Van Der Auwera GA, et al. Scaling accurate genetic variant discovery to tens of thousands of samples. bioRxiv 2017. Doi: 10.1101/201178.
[33] Poplin R, Chang P-C, Alexander D, Schwartz S, Colthurst T, Ku A, et al. A universal SNP and small-indel variant caller using deep neural networks. Nat Biotechnol 2018;36:983–7. doi: https://doi.org/10.1038/nbt.4235.
[34] Pai AA, Pritchard JK, Gilad Y. The Genetic and Mechanistic Basis for Variation in Gene Regulation. PLoS Genet 2015;11:e1004857.
[35] Crowley JJ, Zhabotynsky V, Sun W, Huang S, Pakatci IK, Kim Y, et al. Analyses of allele-specific gene expression in highly divergent mouse crosses identifies pervasive allelic imbalance. Nat Genet 2015;47:353–60. doi: https://doi.org/10.1038/ng.3222.
[36] Gaur U, Li K, Mei S, Liu G. Research progress in allele-specific expression and its regulatory mechanisms. J Appl Genetics 2013;54:271–83. doi: https://doi.org/10.1007/s13353-013-0148-y.
[37] Wiestner A, Schlemper RJ, Van Der Maas APC, Skoda RC. An activating splice donor mutation in the thrombopoietin gene causes hereditary thrombocythaemia. Nat Genet 1998;18:49–52. doi: https://doi.org/10.1038/ng0198-49.
[38] Sereika M, Kirkegaard RH, Karst SM, Michaelsen TY, Sørensen EA, Wollenberg RD, et al. Oxford Nanopore R10.4 long-read sequencing enables the generation of near-finished bacterial genomes from pure cultures and metagenomes without short-read or reference polishing. Nat Methods 2022;19:823–6. doi: https://doi.org/10.1038/s41592-022-01539-7.
[39] Karst SM, Ziels RM, Kirkegaard RH, Sørensen EA, McDonald D, Zhu Q, et al. High-accuracy long-read amplicon sequences using unique molecular identifiers with Nanopore or PacBio sequencing. Nat Methods 2021;18:165–9. doi: https://doi.org/10.1038/s41592-020-01041-y.