

Advances in single-cell long-read sequencing technologies

Pallavi Gupta^{1,2,3,*}, Hannah O'Neill⁴, Ernst J. Wolvetang², Aniruddha Chatterjee^{2,4,*} and Ishaan Gupta^{3,*}

¹University of Queensland – IIT Delhi Research Academy, Hauz Khas, New Delhi 110016, India

²Australian Institute of Bioengineering and Nanotechnology (AIBN), The University of Queensland, St Lucia, QLD 4072, Australia

³Department of Biochemical Engineering and Biotechnology, Indian Institute of Technology Delhi, Hauz Khas, New Delhi 110016, India

⁴Department of Pathology, Dunedin School of Medicine, University of Otago, 58 Hanover Street, Dunedin 9054, New Zealand

*To whom correspondence should be addressed. Tel: +91 11 26591008; Email: ishaan@iitd.ac.in

Correspondence may also be addressed to Aniruddha Chatterjee. Email: aniruddha.chatterjee@otago.ac.nz

Correspondence may also be addressed to Pallavi Gupta. Email: pgupta3005@gmail.com

Abstract

With an increase in accuracy and throughput of long-read sequencing technologies, they are rapidly being assimilated into the single-cell sequencing pipelines. For transcriptome sequencing, these techniques provide RNA isoform-level information in addition to the gene expression profiles. Long-read sequencing technologies not only help in uncovering complex patterns of cell-type specific splicing, but also offer unprecedented insights into the origin of cellular complexity and thus potentially new avenues for drug development. Additionally, single-cell long-read DNA sequencing enables high-quality assemblies, structural variant detection, haplotype phasing, resolving high-complexity regions, and characterization of epigenetic modifications. Given that significant progress has primarily occurred in single-cell RNA isoform sequencing (scRiso-seq), this review will delve into these advancements in depth and highlight the practical considerations and operational challenges, particularly pertaining to downstream analysis. We also aim to offer a concise introduction to complementary technologies for single-cell sequencing of the genome, epigenome and epitranscriptome. We conclude by identifying certain key areas of innovation that may drive these technologies further and foster more widespread application in biomedical science.

Introduction

Sequencing the genome, transcriptome and epigenome provide complementary information about a cell's nucleic acids, including their abundance, intracellular distribution and chemical state. Conventional DNA sequencing assumes that all the cells, taken as a bulk sample, contain no underlying mutation (i.e. their genomes are identical). However, cells may harbor single nucleotide variants (SNVs), structural variants (SVs), and copy number variants (CNVs) and carry variable epigenetic signatures. Similarly, traditional bulk RNA sequencing also involves pooling RNA from thousands of cells for sequencing; yielding averaged transcriptome information that ignores variable cell states such as cell cycle phases (1). Furthermore, research into epigenetic alterations has historically been done in bulk cell populations. However, no two cells of an individual are identical in their DNA, RNA, or epigenetic profiles across tissues (2), developmental time points, and disease status (3,4). This necessitates a deeper look into individual cells to understand cellular variation in different contexts. The emergence of single-cell genomic, transcriptomic, and epigenomic sequencing technologies over the years allows for unprecedented resolution of omics profiles at the single-cell level (5).

Single-cell resolution is achieved by attaching short nucleotide tags called cell barcodes (CBs) to the molecules of interest. All such molecules within a cell, typically DNA or RNA, will contain identical barcodes. These allow *in-silico* demultiplexing of sequenced reads to assign them to individual cells. Single-cell sequencing protocols involve PCR ampli-

fication to increase the available starting material, which can introduce bias when quantifying the reads. This is especially a concern when determining gene expression or identifying CNVs. To circumvent this issue, unique molecular identifiers (UMIs) are attached to molecules prior to amplification, implying sequenced reads with the same UMI can be removed for a more accurate quantification (6). While UMIs ensure the removal of PCR-generated chimeric fragments that contain more than one UMI, they also allow consensus consequence generation as well as variant calling using all the reads with the same UMI (7). Currently, most of the single-cell sequencing is performed on short-read next-generation sequencing (NGS) platforms which offer cost-effective and high accuracy outputs, which is essential for reliable demultiplexing of cellular barcodes.

Upon performing single-cell RNA sequencing (scRNA-seq), gene expression profiles of individual cells are generally used to group them into clusters that correspond to cell types and/or cell states (8). In addition to differences at the level of gene expression, cells also exhibit alternative splicing to produce different transcript isoforms from the same gene. These include alternative transcription start or termination sites (TSS and TTS), differential exon usage, differential transcript usage, alternative 3' and/or 5' splice sites, and intron retention (IR). However, the current scRNA-seq technologies elicit a heavy 3'/5' bias because only one end of the transcript, generally up to 150 nucleotides, is sequenced (dependent on the library preparation protocol). This prevents accurate determination of true chimeric transcripts generated from

Received: February 1, 2024. Revised: April 18, 2024. Editorial Decision: April 22, 2024. Accepted: April 29, 2024

© The Author(s) 2024. Published by Oxford University Press on behalf of NAR Genomics and Bioinformatics.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

chromosomal rearrangements - such as immunoglobulins (Igs) and chimeric RNAs produced in cancers (9,10). Moreover, some scRNA-seq library preparation methods (discussed further in the next section) are dependent on the completeness of 3' UTR annotation to calculate gene counts. For species where the annotation is incomplete, a significant proportion of reads that map to such loci might be discarded (11).

Genome and epigenome sequencing techniques such as single-cell whole-genome amplification and sequencing (scWGA, scWGS), single-cell assay for transposase-accessible chromatin using sequencing (scATAC-seq), single-cell whole-genome bisulfite sequencing (scWGBS) and single-cell reduced-representation bisulfite sequencing (scRRBS-seq) are constrained by the available starting material that is one diploid DNA molecule per cell. This limitation is further exacerbated in single-cell epigenome profiling, where harsh bisulfite conversion leads to significant loss of DNA (12). Moreover, the short reads obtained have a high mapping uncertainty in repetitive regions, one of the key regions examined in DNA methylation sequencing (13). These may also yield incomplete information when complex structural variations, i.e. a combination of SVs, are prevalent in the genomes (14).

While these technologies have enhanced our understanding of the genome, epigenome, and transcriptome of individual cells, the information is incomplete at many levels. These capability gaps have driven innovations in both library preparation protocols and sequencing technologies that can more accurately identify genomic variants, determine methylation profiles at repetitive elements, capture known as well as novel full-length (FL) isoforms, and determine the extent of alternative splicing at a single-cell level. This has been made possible through third-generation long-read sequencing (TGS) platforms that ideally capture the FL molecules. Figure 1 shows the key advancements of long-read sequencing of various -omes at a single-cell level.

As most of these advances have been made in the field of single-cell RNA isoform profiling, the current review will discuss these in detail while providing a brief overview of the complementary technologies for genome, epigenome, and epitranscriptome sequencing at a single-cell level. We discuss how the long-read sequencing technologies were integrated into existing scRNA-seq pipelines and the challenges pertaining to library preparation, sequencing, and downstream data analysis that these involve.

Advances in single-cell full-length RNA capture

The library preparation protocols currently used for scRNA-seq are either plate-based, well-based, or droplet-based, depending on the cell separation method (15). Plate-based methods rely on physically isolating cells into 96- or 384-well plates and have low scalability. Microwell-based techniques have higher throughput, but the procedure is lengthy and involves long hands-on time, which can potentially introduce variability due to human error. Conversely, encapsulating cells into individual droplets containing all the required reagents, including barcoded beads, allows parallel sequencing of thousands of cells but demands equal capture rates for all cell types examined in a sample. The emerging combinatorial indexing methods such as sci-RNA-seq and SPLiT-seq, commercialized as SCALE Biosciences and Parse Biosciences respectively, alleviate the requirement of microfluidic devices while ensuring high barcoding sensitivity and throughput (16–18).

The 10X Genomics Chromium assay is a droplet-based library preparation protocol that can then be sequenced on the Illumina platform (19). This method uses a polydT oligo to capture the polyA tail of mRNAs; therefore, only polyadenylated transcripts are captured. Consequently, this method has a lower RNA capture efficiency, more dropout events, and increased technological noise for lowly expressed RNAs. Paired-end sequencing is performed for these samples, where one of the read pairs provides the sequence of the CB and UMI while the other pair captures the transcript (3' or 5' end depending on the chemistry used). This technique effectively examines cells from a heterogeneous sample and enables cell clustering based on their gene expression profiles and is widely used. However, to inspect variations in transcript isoforms across cells, the entire length of the isoform must be captured.

An alternative technique is Smart-seq, a plate-based scRNA-seq method that uses tagmentation to capture the transcriptome from short reads while retaining cell information through sample indices analogous to CBs (20–22). Though this technique can delineate relative changes in exon usage, TSS, TTS, and splice sites, it cannot accurately reconstruct FL transcripts, especially if novel transcripts are also expected in the cells. The UMIs are only introduced in Smart-seq3, although most of the fragments along the gene, called internal reads, do not contain these tags as they are only retained in the 5' fragments (22). This leads to UMI-containing reads having a 5' bias whereas the others show a 3' bias, which prevents accurate isoform-level quantification (23).

A bulk RNA-seq technique called LoopSeq tags each fragment of the FL molecule with the same UMI, ensuring more accurate transcript assembly even when performing short-read sequencing (24). It holds the potential to be incorporated into single-cell sequencing workflows, but it remains to be adopted. An innovation called scRCAT-seq (single-cell RNA cap and tail sequencing) and its improvement scRCAT-seq2 were developed to capture variation in transcripts at a single-cell level using short-read sequencing but were limited to the TSS and TTS sequencing (25,26). Another technology, called VASA-seq (vast transcriptome analysis of single cells by dA-tailing), alleviates the issue of bias and single-ended sequencing as observed in Smart-seq and Illumina, respectively, by polyadenylating all fragments to be sequenced (27). However, this approach also suffers from inaccurate assembly of transcripts when multiple isoforms from the same gene are produced. The resulting workflow, hereby called single-cell RNA isoform sequencing (scRiso-seq), can elucidate differences across cells at the isoform-level, including capturing novel isoforms. This has been made possible by using one of the prevalent technologies, such as 10X chromium (droplet-based) and Smart-seq3 (plate-based), to prepare barcoded libraries to multiplex cells while using long-read sequencing platforms instead of the conventional short-read sequencing.

Choice of long-read sequencing platform for single-cell libraries

Availability of long-read sequencing platforms, the read lengths they offer, their throughput, multiplexing options, and error rates are important considerations for single-cell sequencing. PacBio single-molecule real-time (SMRT) system uses the kinetics of DNA polymerase to incorporate the base and move onto the next—pulse width and interpulse duration, respectively (28). It currently provides

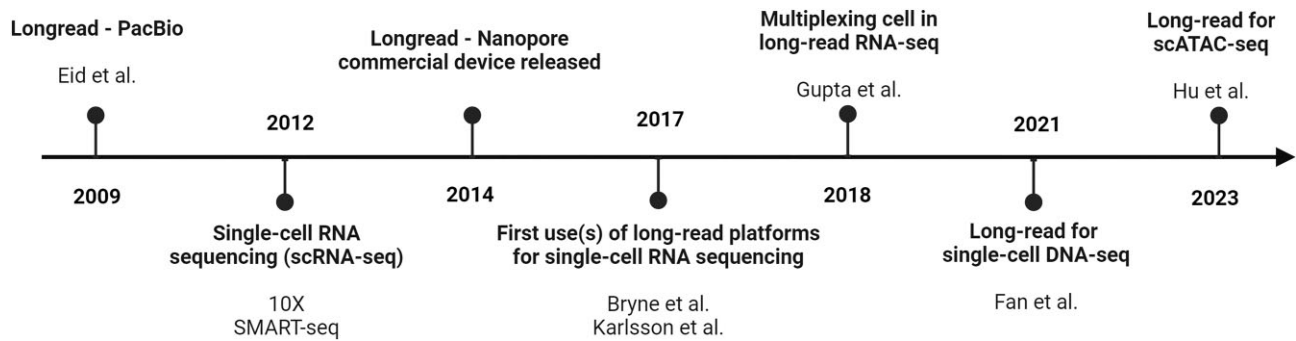


Figure 1. Timeline of technological advancements in single-cell long-read sequencing of the genome, transcriptome and epigenome.

Sequel, Sequel II, Sequel IIe and Revio systems that may be used with 1M and 8M SMRT Cells that generate up to 500 000 and 2 000 000 HiFi reads, respectively. The newly launched Revio system offers sequencing in up to four SMRT Cells in parallel. ONT uses the fluctuation in electrical currents to identify the k-mer sequence within a fragment (29). For nanopore sequencing, the MinION, GridION and PromethION devices can be used with either MinION, Flongle, or PromethION flow cells. Illumina also recently introduced synthetic long-read sequencing for the human genome (<https://sapac.illumina.com/science/technology/next-generation-sequencing/long-read-sequencing.html>).

The early chemistries of PacBio and Nanopore technologies suffered from higher error rates as compared to short-read sequencing platforms, preventing SNV identification and splice-site sequence annotation. Concerning scRiso-seq, this posed a unique challenge—if the error is incorporated in the CB or the UMI, the read(s) will not be accurately assigned. While the accuracies of both long-read platforms have increased over the years, strategies to perform error correction include increasing the coverage of reads, i.e. sequence more fragments corresponding to a region as is performed using consensus circular sequencing (CCS) for PacBio, or using corresponding short reads to perform hybrid error correction (also see subsequent sections).

Conventionally, for bulk RNA-seq, the library preparation protocol for both long-read sequencing platforms does not include a PCR step. However, to ensure that the few molecules from single cells are captured and sequenced sufficiently, the library prepared for scRiso-seq must be PCR amplified. To assist error-correction in downstream analysis, the single-cell libraries prepared through either of the discussed protocols is split into two - one for long-read and the other for short-read sequencing. For the former, the number of starting cells (tens, hundreds, or thousands depending on the single-cell library preparation method used and its corresponding multiplexing extent), the coverage desired (in terms of FL counts per cell), and the throughput offered by the flow cells from existing platforms affect the choice of the platform used as shown in Table 1. The number of FL reads obtained per cell varies depending on the machine throughput and the input number of cells (Figure 2 provides a graphical representation of this concept).

Analysis of single-cell long-read RNA-seq data

Once the long-read sequencing run has been performed with or without complementary short-read sequencing, the down-

stream computational pipeline for an scRiso-seq workflow is outlined in Figure 3. While the initial steps, such as QC and adapter trimming, are performed independently for either technology, data may be merged in the subsequent steps when short-read data is used to correct errors in the long-read set.

Identifying and correcting CBs and UMIs

Owing to the high error rate of long-read sequencing platforms, one of the essential steps that is different from a standard scRNA-seq workflow is accurately assigning cell barcodes (CBs) and unique molecular identifiers (UMIs) to each read which also harbor these errors. Several bioinformatics tools, combined with various library preparation methods, have been developed to correct and/or assign CBs and UMIs. They are categorized based on whether the algorithm depends on a complementary set of short-read sequences to use as the whitelist for CBs and UMIs.

One of the first publicly-available tools to correctly assign CBs and UMIs using the corresponding short-read data, forming the ScNaUmi-seq (Single-cell Nanopore sequencing with UMIs) workflow, was a Java companion toolkit called siceLore (Single Cell Long Read) (38). Short stretches from the long read between a valid adapter sequence and a threshold number of poly-As contain the CB and UMI. Both the long-read and short-read data are aligned to the genome, and the CBs for each gene or genomic location from both platforms are compared to assign barcodes (see yellow box in Figure 3). Once CBs are correctly assigned to Nanopore reads, UMI sequences for long reads aligning to each gene or genomic location from the same cell are compared to corresponding short-read data. The accuracy of this strategy is over 97% for both CB and UMI assignment and correctly assigned CBs and UMIs for ~70% of Nanopore reads. As an improvement on SiCeLoRe (38), the authors removed the dependency of the algorithm on the polyA tail to determine the region containing the CB and UMI to develop snuupy (Single Nucleus Utility in Python) (42). In this approach, the unmapped region of the long-read sequences is searched against CB and UMI combinations identified from the short-read sequences. Furthermore, another tool, scNapBar (Single-cell Nanopore Barcode Demultiplexer), also builds on the algorithm of siceLore with an additional aim to reduce the dependency of the algorithm on the depth of short-read sequencing and incomplete genome annotation (39). Especially for low sequencing throughput, instead of using UMI assignment via genome mapping, a Naïve Bayes model is employed to predict the likelihood of the correctness of CB assignment.

Table 1. A tabular summary of all the peer-reviewed studies till date that perform single-cell long-read isoform sequencing (scRiso-seq)

S. No.	Author year	Total # of cells LR sequenced	Library prep protocol	Sequencing device configuration	FL counts per cell	Total FL counts	Ref	In Figure 2
1	Byrne 2017	7	Smart-seq2	3 cells on 3 R7.3 (2D) + 4 cells on 1 R9.4 (2D) MinION flowcells	35 222 + 93 300	-	(30)	1
2	Karlsson 2017	6	STRT-C1	7 SMRT cells on RSII	NA	NA	(31)	0
3	Gupta 2018	6627	10X 3' v2 chromium	11 1M SMRT cells on Sequel	-	5.2×10^6	(32)	1
4	Volden 2018	96	Tn5Prime	1 R9.5 (1D) MinION flowcell	-	1 132 707	(33)	1
5	Ranum 2019	12	Smart-seq2	4 R9.4 (1D) MinION flowcells	NA	NA	(34)	0
6	Russel 2019	1614 + 50	10X chromium → PCR enrichment	1 SMRT cell on RSII + 1 SMRT cell on Sequel	NA	NA	(35)	0
7	Singh 2019	3743	10X 3' v2 chromium	21 R9.4.1 (1D) + 1 R9.5.1 (1D2) MinION flowcells	-	20 346 396	(10)	1
8	van Galen 2019	255	Seq-well WTA	4 R9.4 (1D) / R9.5 (1D2) MinION flowcells	-	0.97×10^6 (3 genes)	(36)	1
9	Fan 2020	183	Smart-seq2	1 PromethION flowcell	-	44×10^6	(37)	1
10	Lebrigand 2020	190	10X 3' v2 chromium	8 PromethION flowcells	1	32	(38)	1
11	Wang 2021	2000	10X 3' v3 chromium	2 R9.4 MinION flowcells	-	12.5×10^6	(39)	1
12	Schmidt 2021*	1000	10X 3' v3 chromium	1 SMRT cell on Sequel II	-	838 363	(40)	1
13	Palmer 2021	>170 000	10X 3' v3 chromium	16 SMRT cells on Sequel II	-	98×10^6	(41)	1
14	Long 2021	5000	10X 3' v2 chromium	2 R9.4.1 MinION flowcells	729 UMI 362 UMI	NA	(42)	1
15	Joglekar 2021	NA	10X 3' v2 chromium 10X visium	24 SMRT cells on Sequel I + 20 SMRT cells on Sequel II 1 PromethION flowcell	-	5 764 421	(43)	0
16	Rebboah 2021	1000	Split-seq	2 SMRT cells on Sequel II	-	NA	(44)	1
17	Philpott 2021	1200	Drop-seq	1 PromethION flowcell	NA	NA	(45)	0
18	Tian 2021	2737	10X 3' v2/v3 chromium + 10X 5' v1 chromium	4 PromethION flowcells	NA	NA	(46)	0
19	Boileau 2022	12000# spots	10X visium	4 GridION flowcells	-	25.5×10^6	(47)	1
20	Ebrahimi 2022	15375	10X 3' v3 chromium	NA PromethION flowcells	-	17.5×10^6	(48)	0
21	Hardwick 2022	>13 800	10X 3' chromium	5 PromethION flowcells + 15 SMRT cells	-	290×10^6	(49)	1
22	Healey 2022	10 000	10X 3' v3 chromium	1 8M SMRT cell Sequel II	-	152×10^6	(11)	1
23	Volden 2022	3000	10X 3' chromium	3 MinION + 3 PromethION flowcells	-	12×10^6	(50)	1
24	Oguchi 2022 #	40	SMART-seq v4	8 R9.4 MinION flowcells	-	17.6×10^6	(51)	1
25	Hazzard 2022	12 000	10X 3' chromium	4 SMRT cells on Sequel II	-	11.5×10^6	(52)	1
26	Liao 2023	3168	Smart-seq3	5 PromethION 48 flowcells	-	241×10^6	(53)	1
27	Shi 2023	2000	10X 3' v3 chromium	1 8M SMRT cell on Sequel II	-	3.83×10^6	(54)	1
28	Lebrigand 2023	5978	10X visium	13 PromethION	-	535×10^6	(55)	1
29	Mincarelli 2023	>8000	10X 3' v2 chromium	6 SMRT cells on Sequel/ Sequel II	-	17.9×10^6	(56)	1
30	You 2023	1000	10X 3' v3 chromium	1 PromethION + 2 GridION flowcells	-	115×10^6	(57)	1
31	Liu 2023	9845	10X 3' v3 chromium	NA SMRT cells on Sequel II	-	4.96×10^6	(58)	0
32	Yang 2023	28 290	10X 3' v3 chromium	7 8M SMRT cells on Sequel II	-	12.47×10^6	(59)	1

It lists the number of cells (or spots in spatial transcriptomics) sequenced and the total or per-cell full-length (FL) read counts prior to any processing or quality check. NA: Not available. * indicates the study has not yet been peer-reviewed, # means 60% spot coverage assumed. Note: Since 2020, all ONT sequencing is 1D, therefore not explicitly mentioned.



Figure 2. Pictorial representation of studies from Table 1 (24 out of 32) depicting the number of cells sequenced and the average number of FL reads obtained per cell. The number of flow cells from different long-read sequencing technologies used have also been indicated.

Kim et al. also used a similar principle to assign only the CBs at the 5' end of the transcripts (60). Here too, the soft-clipped or unaligned portion of reads was compared with the corresponding list of CBs from short-read data using cosine similarity scores. The short-read CB with the highest score and the minimum edit distance to the 5' end is considered the valid match. However, the sample set of CBs in the given study was limited to one amplified gene, and the applicability of this technique to transcriptome-level barcode matching has not been assessed. On the other hand, scTagger is designed to suit both 3' and 5' scRNA-seq chemistries to correct CBs (48). It filters out the lowest-frequency barcodes determined from cell ranger (19) and uses the remaining set to assign CBs to long reads. As the library design is known apriori, the regions downstream (or upstream) in forward (or reverse) strands of long reads are compared to the subset of CBs. The CB can be assigned if the computed Levenshtein edit distance is lesser than the user-specific cut-off. Neither of these studies (green box in Figure 3) attempt to correct UMIs; instead perform UMI counting directly.

Two tools developed to identify CBs and/or UMIs when performing only long-read sequencing are Sockeye (<https://github.com/nanoporetech/sockeye>) and BLAZE (Barcode identification from Long-reads for AnalyZing single-cell gene Expression). Sockeye is used to identify UMIs and CBs in Nanopore reads generated using one of 10X scRNA-seq protocols. However, it does not perform isoform-based single-cell analysis and only returns a gene $\times$ cell expression matrix. Benchmarking study also showed that it retains cluster(s) with low UMI count cells and/or non-cell associated barcodes (57). The new implementation of this approach

is a Nextflow workflow (<https://github.com/epi2me-labs/wf-single-cell>), which provides an isoform $\times$ cell expression matrix as well. BLAZE uses a three-step approach to identify CBs using only long-read data, making it more conservative than Sockeye (57). Once the putative location of barcodes is identified as immediately downstream of the adapter sequence, those not appearing in the 10X whitelist or with low-quality scores are discarded. Then a quantile-based system is used to retain the high-count CBs.

While the tools mentioned above use the same CBs and UMIs as in the short-read scRNA-seq library preparation protocols (blue box in Figure 3), irrespective of whether short-read sequencing is performed, it is possible to use custom-designed oligos instead when short-read sequencing is not done. This was adopted in scCOLOR-seq - Single-cell Corrected Long-read sequencing (45), where both the oligos are homodimer sequences. This enables the identification and correction of CBs and UMIs using a directional protocol and Levenshtein distances, respectively (61).

Single-cell long-read RNA-seq analysis pipelines

Once CBs and UMIs have been reliably assigned to FL reads, downstream steps such as aligning to the reference and error correction of the reads, especially of the splice junctions, followed by transcriptome assembly, quantification, and differential transcript analyses can proceed. Most tools for analyzing bulk long-read RNA (or cDNA) sequencing data can be used directly for these purposes. These have been collated in a review elsewhere (62). Analytical toolkits for single-cell data, such as Seurat and Scanpy, can be used to assign cell types or

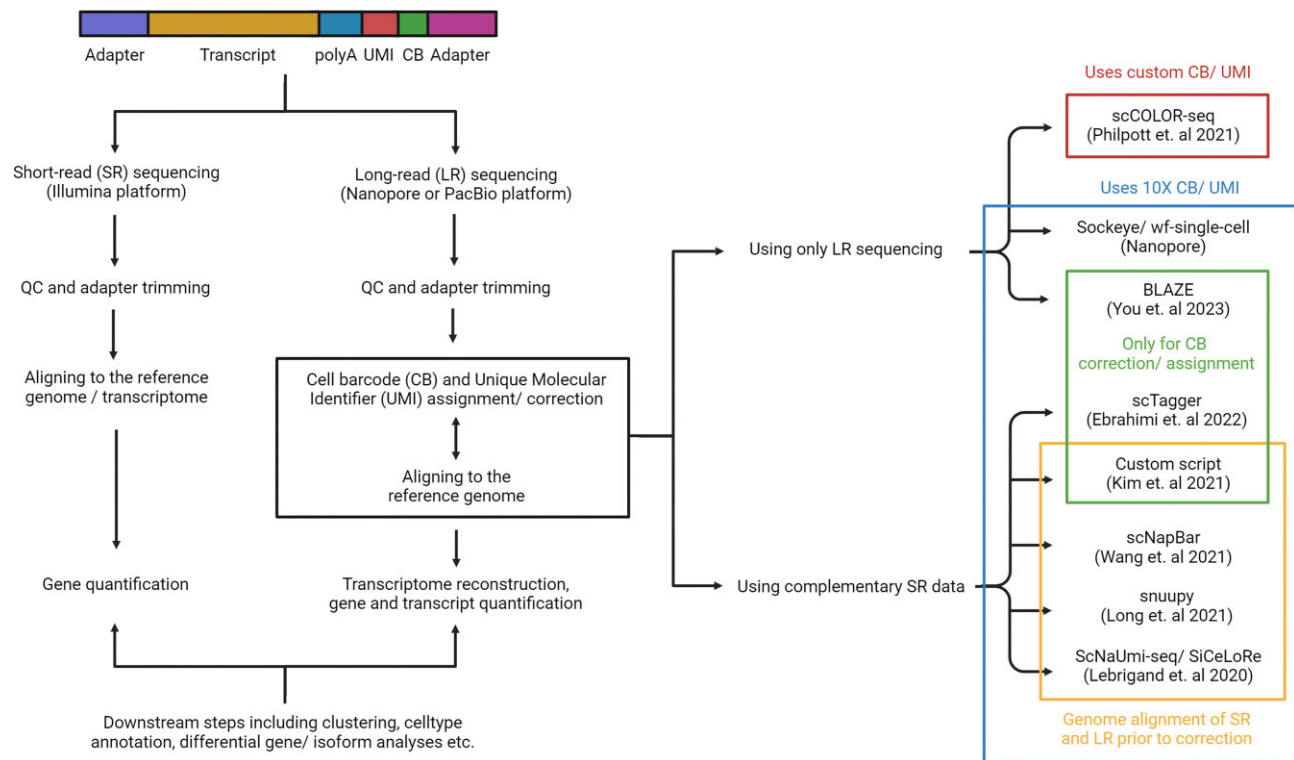


Figure 3. A typical long-read single-cell RNA sequencing (aka scRiso-seq) workflow. The additional step of UMI and CB correction and assignment and the available tools for the same is highlighted.

cell states at both gene- and transcript-level (63,64). Tools are also available to visualize (65) interesting transcripts identified through long-read data, including novel fusion transcripts discovered at a single-cell level (66).

Software suites that integrate some of these steps have been designed to aid scRiso-seq processing. Mandalorian was a Python-based pipeline that could perform steps including basic QC, alignment, isoform identification and quantification, and differential isoform usage analysis on ONT 2D reads, which were discontinued in 2017 (30). An updated version could be used for R2C2 reads in combination with the C3POa (Concatemeric Consensus Caller using partial order alignments) consensi-calling tool (33). The FLAMES workflow (Full-Length Analysis of Mutations and Splicing) was designed for CB and UMI assignment in long-read data prior to alignment to the reference genome, using complementary short-read data (46). It subsequently performs downstream steps such as read alignment, transcript assembly, and quantification. The authors reported improved quality of transcript identification over tools such as FLAIR, Stringtie2, and TALON, originally designed for bulk RNA-seq; however, no comparison of CB and UMI assignment accuracy was made. The output from other tools, such as BLAZE (discussed in the previous section), can also be merged with FLAMES to perform only the downstream steps (57). The scISA-tools pipeline was originally developed to prevent the existing scRNA-seq tools from classifying the true full-length non-chimeric (FLNC) reads from HIT-scISOseq as chimeric due to its underlying concatenation principle (54). However, the steps—including reference-guided mapping, UMI and CB error correction (using only long reads), count matrix generation, and cell type annotation—can also be applied to any other scRNA-seq dataset. The scNaST set of tools builds up

on scNapBar (39), using a hybrid approach to assign spatial barcodes to long reads, followed by spatial spot deconvolution, spatial gene expression, transcript classification, and differential transcript usage (47). Only recently, two toolkits, sc-NanoGPS (single-cell Nanopore sequencing analysis of Genotypes and Phenotypes Simultaneously) and Scywalker (not peer-reviewed yet), were developed to assign CBs and UMIs as well as calculate gene and transcript-wise expression profiles without complementary single-cell short reads (67,68). sc-NanoGPS features CB correction that does not rely on a user-provided barcode whitelist; instead, it is generated within the algorithm. It subsequently identifies duplicate UMIs for reads that map to the same genomic region.

Single-cell long-read sequencing of alternative -omes

So far, this review has focused on single-cell long-read transcriptome sequencing. However, long-read sequencing of other -omes, such as the genome, epigenome and epitranscriptome at a single-cell level is an emerging area, warrants discussion.

Single-cell long-read genome sequencing

scWGS or single-cell whole-genome sequencing has allowed small variants in the DNA, such as SNVs, short insertion-deletions (indels), and CNVs, to be identified between cells (69). The existing methods were developed using NGS, where short albeit accurate reads are produced, which are, however, insufficient to detect simple or complex SVs, transposable elements, and extrachromosomal circular DNA (ecDNA) (70). The first report of a single-cell long-read genome sequenc-

ing technique was in 2021, termed SMOOTH-seq (single-molecule real-time sequencing of long fragments amplified through transposon insertion) (70). This technology adopted a Tn5 transposition method, previously utilized for scWGS (71), to generate long fragments for third-generation sequencing platforms. Here, instead of two different adapter sequences, the Tn5 transposase with one adapter sequence was used to ensure retrieval of all original DNA fragments. The researchers optimized the reaction conditions to capture and amplify long fragments suitable for TGS platforms efficiently. Briefly, it included the use of Tks Gflex DNA Polymerase for amplification and reduction in the concentration of Tn5 transposase to prevent self-looping. The resulting fragments from each cell are then barcoded and pooled before sequencing.

More recently, a second protocol has emerged that utilizes a droplet-based multiple displacement amplification method (dMDA) (72). Upon the lysis of a single cell, the genomic fragments are encapsulated into droplets such that only one or few DNA molecules are present within each, thereby mitigating the chances of forming inter-molecular chimeras. The small volume and limited reagents in the droplet also reduce over-amplification. Once the dMDA occurs, typical library preparation for long-read and short-read sequencing platforms can proceed. While dMDA and similar techniques have been previously described for scWGS or to amplify limited DNA (73,74), this paper is the first to describe its use in long-read sequencing technology.

Pre-processing and genome assembly for either technology proceeds similarly to the corresponding bulk sequencing reads with an extra barcode-based deconvolution step. Downstream applications, such as SNV, CNV or SV identification, can also be performed through the tools available for a typical bulk long-read DNA sequencing analysis. The dMDA method yielded more data per cell, longer read lengths and greater genomic coverage than SMOOTH-seq. The first application of both protocols sequenced a limited number of cell(s) per sequencing run. While SMOOTH-seq had 16 cells per library, producing ~1 Gb of data per cell, the dMDA method had only one cell per library, producing 20 Gb of data from it. While both the sequencing runs produced HiFi reads, the average read lengths were significantly different—6 kb for the former and 10–12 kb for the latter. The average genomic coverages for the two protocols also differed—19.5% and 40%, respectively. The cost per cell was not stated for dMDA; however, for SMOOTH-seq, the cost of ~1 Gb per cell was ~\$260. An improved SMOOTH-seq version utilized ONT with the same sequencing depth rather than PacBio, leading to a substantially lower cost of ~\$14 per Gb per cell (75).

Single-cell long-read epigenome sequencing

Over the years, several single-cell epigenomic methods have been described, exploring DNA methylation (76,77), chromatin accessibility (78), and histone modifications (79). While identifying histone modifications is outside the purview of the technology in discussion, advances in single-cell long-read profiling can boost DNA methylation and chromatin accessibility profiling. However, most epigenomic long-read sequencing techniques have been applied to bulk populations (80–82). PacBio and ONT sequencers rely on the ability of methylated bases to alter the polymerase kinetics and electrical charge pattern, respectively, thus enabling analysis of these DNA modifications (80,83). Currently, the required min-

imum DNA input for PacBio and ONT is relatively high, with low inputs requiring PCR amplification prior to sequencing (84), while the DNA available from a single cell is ~6 pg. Unlike single-cell long-read genome sequencing, amplification is not feasible for sequencing the DNA methylome since it would remove the native methylation marks. While bisulfite conversion could still be performed prior to library preparation, it fragments DNA, rendering long-read sequencing futile. Alternative enzymatic reactions for the detection of 5mC and 5hmC using long reads, such as those employed in LR-EM-seq, have been shown to alleviate this shortcoming; however, it is yet to be used for single cells (82). Hence, though there are no publications on single-cell long-read sequencing of DNA methylation to our knowledge yet, as long-read sequencing technologies progress and the technical processes are streamlined, the input requirements may be lowered, making it a more feasible option for single-cell methylome studies.

Only one single-cell long-read epigenomic technique has been described, a single-cell assay for transposase-accessible chromatin on the Nanopore sequencing platform (scNanoATAC-seq) (85). A combination of the plate-based scATAC-seq method (86) with the Tn5 transposition technique was adapted to produce longer fragments for TGS. This allows for the simultaneous analysis of chromatin accessibility and genetic variants such as SVs, SNVs and CNVs within a single cell. This also enables the identification of allele-specific and co-accessible neighboring peaks through haplotype phasing, which is impossible through short-read sequencing. While both SMOOTH-seq and scNanoATAC use Tn5 transposition methods, scNanoATAC-seq has slightly shorter fragments (median 4–5 kb), focusing more on chromatin-accessible regions (85). Up to 960 cells could be multiplexed for a single sequencing run of scNanoATAC-seq on one PromethION flow cell, which reduced the cost to \$2.50 per cell. The pre-processing of scNanoATAC-seq reads follows the same steps as bulk ONT. For downstream analyses, the ArchR pipeline (87), originally developed for short-read scATAC-seq, was modified such that only the ends of the long reads were extracted for chromatin accessibility signal analysis.

Single-cell long-read epitranscriptome sequencing

Similar to epigenomics, epitranscriptomics focuses on the study of functionally relevant RNA modifications that do not alter the underlying ribonucleotide sequence, such as *N*⁶-methyladenosine (m⁶A), 5-methylcytidine, amongst others. These modifications can affect RNA folding, stability, and nuclear export and regulate interaction with other cellular molecules (88). Moreover, different positions of these epitranscriptomic marks regulate transcription differently (89). The ONT sequencing platform can now simultaneously identify isoforms as well as modifications to transcripts, unraveling the multi-layered regulation surrounding the transcriptome. However, like genomic and epigenomic long-read sequencing, publications have focused on bulk populations of cells for epitranscriptome long-read sequencing (90). Theoretically PacBio's kinetics-based sequencing has the capacity to detect RNA base modifications, but the reliable extraction of this information and sequencing of native RNA remains challenging tasks at present (91,92). As with single-cell long-read DNA methylation sequencing, it is not yet possible to carry this out in single cells as the modifications to the RNA are removed

during PCR amplification. As the input requirements continue to be lowered for these TGS platforms, this is likely possible in future studies.

Specific applications of single-cell long-read sequencing

The underlying heterogeneity of ribonucleic acids in nuclear (49), cytoplasmic fractions (51), or total cells (44) can be elucidated through single-cell RNA isoform sequencing. This includes determining the proportion and potential role of compartment-specific transcripts. Some single-cell long-read sequencing strategies aim to enrich target sequences by designing appropriate primers (36) or probe hybridization as used in RAGE-seq (10) and RaCH-seq (93). These methods lead to sequencing cost reduction while investigating informative genes and full-length transcripts. Though scRNA-seq is extensively used to calculate RNA velocity, allowing pseudotime and trajectory analysis, the inability to accurately determine spliced versus unspliced transcripts necessitates several underlying assumptions that can potentially be alleviated by scRiso-seq (62,94). Different developmental stages of the blood parasites, such as *Plasmodium vivax*, have been delineated using FL transcripts from individual cells (52). While the heterogeneity of cells within and across cancer patients can be compared well using this technology, the subpopulations can also be classified based on mutational signatures identified using single-molecule sequencing of transcripts (36). In addition to applying long-read sequencing technologies at a single-cell level, these can also be combined with spatial transcriptomics libraries such as those from 10X visium to detect changes in patterns such as isoform switching across spatial regions (43,47,55). SMOOTH-seq has also demonstrated its effectiveness in detecting structural variants, including duplication events longer than 5 kb, in single cells in a cancer context (70). The group found an enrichment of such events near the telomeres in both colorectal cancer patient samples and K562 cells. This discovery is significant as genomic structural variations are known to drive malignant phenotypes (95,96).

While most of the studies discussed in this review have employed one of the two prevalent technologies, namely 10X chromium or Smart-seq (or its improvements), for library preparation, some authors have also altered these protocols to use in-house tagging (45) and amplification methods instead. Examples include a modification developed for PacBio sequencing called HIT-scISOseq which involves the use of biotinylated primers and concatenation of cDNA inserts to reduce TSO (template switching oligos) artifacts and improve consensus accuracy, respectively (54). Strategies such as SCAN-seq and SCAN-seq2 use identical or different ONT-compatible 24-nucleotide barcodes on 3' and 5' ends to tag RNA molecules of each cell to be sequenced (37,53). Another scheme called Rolling Circle Amplification to Concatemeric Consensus (R2C2) circularizes a single-cell library which is then amplified and debranched before sequencing on a Nanopore flow cell (33,50). Furthermore, the improved SPLiT-seq protocol is now also being used in conjunction with long-read sequencing platforms for whole transcriptome sequencing of single cells (44).

Challenges and future directions

Long-read genome, transcriptome, and epigenome sequencing at a single-cell scale hold immense potential to elucidate in-

dividual cell function, regulation, and cellular heterogeneity. However, there are three major challenges that the field currently faces, namely a lack of streamlined library preparation protocols for long-read DNA sequencing of single cells, failure to detect rarer isoforms and cell populations, and the absence of comparative analysis of the existing wet-lab techniques and data analysis pipelines across all omics data. As is evident from the proportion within this review, the field of scRiso-seq has expanded tremendously over the past years, while the exploration of genome, epigenome, and epitranscriptome is either limited or absent. Less than a handful of technologies exist that have not been compared across similar samples. The degree of multiplexing of cells is also limited both at the library preparation step (no droplet-based protocol) and for sequencing (small number of cells per flowcell). Unlike scRNA-seq and scRiso-seq, UMIs are not incorporated into single-cell short-read or long-read DNA sequencing workflows. Although this does not affect SNV identification, it prevents accurate quantification of CNVs. Furthermore, careful consideration must be taken for the genome assembly algorithm(s) used as it can significantly impact the continuity of the assembly (75). Moreover, the sequencing depth in the case of single-cell accessibility profiling (currently using CCS) must be chosen as it offers a trade-off between a greater number of fragments and longer reads, which give stronger epigenetic signals vs. allowing the detection of genomic features, respectively (85).

While capturing every isoform in a cell is of pivotal importance, this largely hinges on the abundances of isoforms themselves, implying that those with low abundance may go undetected entirely. This notion extends to rarer cell types, which are represented by a fewer number of cells and consequently remain uncharacterized or dismissed as spurious signals in single-cell sequencing data. This problem may potentially be alleviated by either increasing the depth of sequencing or sequencing a greater number of cells. Enrichment of certain cells and genes (or isoforms) may also be a potential solution, although it requires prior knowledge about the intended target(s) and this cannot be applied if novel cell types and isoforms are desired in discovery studies.

The third challenge is the unavailability of comparative benchmarking of the available bioinformatics tools and pipelines, especially for CB and UMI assignments in scRiso-seq. Testing these on both simulated and real datasets is necessary for researchers to make an informed decision for their use (97). As the error rates of long-read sequencing platforms is decreasing, such a comparison would enable the users to choose whether to skip corresponding short-read data, without compromising on accuracy, thereby saving both time and money. This demands the development of read simulators that provide the ground truth, which needs improvement (57). Moreover, complications such as chimeric and truncated reads, as well as the high error rates of sequencing, continue to prevent the full utilization of scRiso-seq at single-cell and spatial levels (38,55). Differential gene expression analysis on single-cell data (short-read) poses significant challenges owing to low expression and variability, which requires judicious selection of the best-suited method (98). Such recommendations for method selection must be expanded to long-read single-cell data. Therefore, newer and better library preparation protocols and bioinformatics data analysis pipelines must be developed in the future to address the existing knowledge and capability gaps. These would help advance the field of long-read sequencing of various -omes at a single-cell scale and lead to more widespread deployment of the technologies.

Data availability

No new data were generated or analysed in support of this research.

Acknowledgements

We are grateful to The New Zealand Centre at IIT Delhi (URL: <https://international.iitd.ac.in/new-zealand-center-at-iitd/>) for supporting this work and a seed grant to AC (NZ PI) and IG (Indian PI). A.C. and H.N. would like to thank the Department of Pathology, Dunedin School of Medicine, University of Otago for the support. A.C. is thankful to Royal Society Te Apārangi for the Rutherford Discovery Fellowship and Health Research Council (HRC, New Zealand) for a Sir Charles Hercus Fellowship to support his position.

Funding

Department of Biotechnology, Government of India through the Ramalingaswami fellowship [ST/HRD/35/02/200 to I.G.]; Ministry of Human Resource Development, Government of India sponsored Prime Minister's Research Fellowship (PMRF) [IITD/Admission/PhD/PMRF/2020–21/380485 to P.G.].

Conflict of interest statement

None declared.

References

- Trapnell, C. (2015) Defining cell types and states with single-cell genomics. *Genome Res.*, **25**, 1491–1498.
- Goldman, S.L., MacKay, M., Afshinnekoo, E., Melnick, A.M., Wu, S. and Mason, C.E. (2019) The impact of heterogeneity on single-cell sequencing. *Front. Genet.*, **10**, 8.
- Hou, Y., Guo, H., Cao, C., Li, X., Hu, B., Zhu, P., Wu, X., Wen, L., Tang, F., Huang, Y., *et al.* (2016) Single-cell triple omics sequencing reveals genetic, epigenetic, and transcriptomic heterogeneity in hepatocellular carcinomas. *Cell Res.*, **26**, 304–319.
- Gupta, R.K. and Kuznicki, J. (2020) Biological and medical importance of cellular heterogeneity deciphered by single-cell RNA sequencing. *Cells*, **9**, 1751.
- Kashima, Y., Sakamoto, Y., Kaneko, K., Seki, M., Suzuki, Y. and Suzuki, A. (2020) Single-cell sequencing techniques from individual to multiomics analyses. *Exp. Mol. Med.*, **52**, 1419–1427.
- Zhang, Y., Wang, D., Peng, M., Tang, L., Ouyang, J., Xiong, F., Guo, C., Tang, Y., Zhou, Y., Liao, Q., *et al.* (2021) Single-cell RNA sequencing in cancer research. *J. Exp. Clin. Cancer Res.*, **40**, 81.
- Österlund, T., Filges, S., Johansson, G. and Ståhlberg, A. (2022) UMIErrorCorrect and UMIAlyzer: software for consensus read generation, error correction, and visualization using unique molecular identifiers. *Clin. Chem.*, **68**, 1425–1435.
- Bacher, R. and Kendziorski, C. (2016) Design and computational analysis of single-cell RNA-sequencing experiments. *Genome Biol.*, **17**, 63.
- Li, Z., Qin, F. and Li, H. (2018) Chimeric RNAs and their implications in cancer. *Curr. Opin. Genet. Dev.*, **48**, 36–43.
- Singh, M., Al-Eryani, G., Carswell, S., Ferguson, J.M., Blackburn, J., Barton, K., Roden, D., Luciani, F., Giang Phan, T., Junankar, S., *et al.* (2019) High-throughput targeted long-read single cell sequencing reveals the clonal and transcriptional landscape of lymphocytes. *Nat. Commun.*, **10**, 3120.
- Healey, H.M., Bassham, S. and Cresko, W.A. (2022) Single-cell Iso-Sequencing enables rapid genome annotation for scRNAseq analysis. *Genetics*, **220**, iyac017.
- O'Neill, H., Lee, H., Gupta, I., Rodger, E.J. and Chatterjee, A. (2022) Single-cell DNA methylation analysis in cancer. *Cancers*, **14**, 6171.
- Slotkin, R.K. and Martienssen, R. (2007) Transposable elements and the epigenetic regulation of the genome. *Nat. Rev. Genet.*, **8**, 272–285.
- Huddleston, J., Ranade, S., Malig, M., Antonacci, F., Chaisson, M., Hon, L., Sudmant, P.H., Graves, T.A., Alkan, C., Dennis, M.Y., *et al.* (2014) Reconstructing complex regions of genomes using long-read sequencing technology. *Genome Res.*, **24**, 688–696.
- Basu, A., Budhraj, A., Juwayria, Abhilash, D. and Gupta, I. (2021) In: Kumar, D. (ed.) *Advances in Genetics*. Academic Press, Vol. 108, pp. 81–145.
- Martin, B.K., Qiu, C., Nichols, E., Phung, M., Green-Gladden, R., Srivatsan, S., Blecher-Gonen, R., Beliveau, B.J., Trapnell, C., Cao, J., *et al.* (2023) Optimized single-nucleus transcriptional profiling by combinatorial indexing. *Nat. Protoc.*, **18**, 188–207.
- Rosenberg, A.B., Roco, C.M., Muscat, R.A., Kuchina, A., Sample, P., Yao, Z., Graybuck, L.T., Peeler, D.J., Mukherjee, S. and Chen, W. (2018) Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science*, **360**, 176–182.
- Cao, J., Packer, J.S., Ramani, V., Cusanovich, D.A., Huynh, C., Daza, R., Qiu, X., Lee, C., Furlan, S.N. and Steemers, F.J. (2017) Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science*, **357**, 661–667.
- Zheng, G.X.Y., Terry, J.M., Belgrader, P., Ryvkin, P., Bent, Z.W., Wilson, R., Ziraldo, S.B., Wheeler, T.D., McDermott, G.P., Zhu, J., *et al.* (2017) Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.*, **8**, 14049.
- Ramsköld, D., Luo, S., Wang, Y.-C., Li, R., Deng, Q., Faridani, O.R., Daniels, G.A., Khrebtkova, I., Loring, J.F., Laurent, L.C., *et al.* (2012) Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat. Biotechnol.*, **30**, 777–782.
- Picelli, S., Faridani, O.R., Björklund, A.K., Winberg, G., Sagasser, S. and Sandberg, R. (2014) Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.*, **9**, 171–181.
- Hagemann-Jensen, M., Ziegenhain, C., Chen, P., Ramsköld, D., Hendriks, G.-J., Larsson, A.J.M., Faridani, O.R. and Sandberg, R. (2020) Single-cell RNA counting at allele and isoform resolution using Smart-seq3. *Nat. Biotechnol.*, **38**, 708–714.
- Pan, L., Dinh, H.Q., Pawitan, Y. and Vu, T.N. (2022) Isoform-level quantification for single-cell RNA sequencing. *Bioinformatics*, **38**, 1287–1294.
- Liu, S., Wu, J., Yu, Y.-P., Balamotis, M., Ren, B., Ben Yehezkel, T. and Luo, J.-H. (2021) Targeted transcriptome analysis using synthetic long read sequencing uncovers isoform reprogramming in the progression of colon cancer. *Commun. Biol.*, **4**, 506.
- Hu, Y., Zhang, S., Mo, X., Xiao, Y., Qiu, Y., Zhong, J., Chen, Z., Liu, X., Chen, X. and Dai, W. (2023) Simultaneous profiling of full-length RNA transcripts and chromatin accessibility within single cells of human retinal organoids. bioRxiv doi: <https://doi.org/10.21203/rs.3.rs-3640234/v1>, 20 December 2014, preprint: not peer reviewed.
- Hu, Y., Zhong, J., Xiao, Y., Xing, Z., Sheu, K., Fan, S., An, Q., Qiu, Y., Zheng, Y., Liu, X., *et al.* (2020) Single-cell RNA cap and tail sequencing (scRCAT-seq) reveals subtype-specific isoforms differing in transcript demarcation. *Nat. Commun.*, **11**, 5148.
- Salmen, F., De Jonghe, J., Kaminski, T.S., Alemany, A., Parada, G.E., Verity-Legg, J., Yanagida, A., Kohler, T.N., Battich, N., van den Brekel, F., *et al.* (2022) High-throughput total RNA sequencing in single cells using VASA-seq. *Nat. Biotechnol.*, **40**, 1780–1793.
- Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., *et al.* (2009) Real-time DNA sequencing from single polymerase molecules. *Science*, **323**, 133–138.
- Kasianowicz, J.J., Brandin, E., Branton, D. and Deamer, D.W. (1996) Characterization of individual polynucleotide molecules using a membrane channel. *Proc. Natl. Acad. Sci. U.S.A.*, **93**, 13770–13773.

30. Byrne, A., Beaudin, A.E., Olsen, H.E., Jain, M., Cole, C., Palmer, T., DuBois, R.M., Forsberg, E.C., Akesson, M. and Vollmers, C. (2017) Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells. *Nat. Commun.*, **8**, 16027.
31. Karlsson, K. and Linnarsson, S. (2017) Single-cell mRNA isoform diversity in the mouse brain. *BMC Genomics*, **18**, 126.
32. Gupta, J., Collier, P.G., Haase, B., Mahfouz, A., Joglekar, A., Floyd, T., Koopmans, F., Barres, B., Smit, A.B., Sloan, S.A., et al. (2018) Single-cell isoform RNA sequencing characterizes isoforms in thousands of cerebellar cells. *Nat. Biotechnol.*, **36**, 1197–1202.
33. Volden, R., Palmer, T., Byrne, A., Cole, C., Schmitz, R.J., Green, R.E. and Vollmers, C. (2018) Improving nanopore read accuracy with the R2C2 method enables the sequencing of highly multiplexed full-length single-cell cDNA. *Proc. Natl. Acad. Sci. U.S.A.*, **115**, 9726–9731.
34. Ranum, P.T., Goodwin, A.T., Yoshimura, H., Kolbe, D.L., Walls, W.D., Koh, J.-Y., He, D.Z.Z. and Smith, R.J.H. (2019) Insights into the biology of hearing and deafness revealed by single-cell RNA sequencing. *Cell Rep.*, **26**, 3160–3171.
35. Russell, A.B., Elshina, E., Kowalsky, J.R., Te Velhuis, A.J.W. and Bloom, J.D. (2019) Single-cell virus sequencing of influenza infections that trigger innate immunity. *J. Virol.*, **93**, e00500–19.
36. van Galen, P., Hovestadt, V., Wadsworth, M.H.I., Hughes, T.K., Griffin, G.K., Battaglia, S., Verga, J.A., Stephansky, J., Pastika, T.J., Lombardi, S., et al. (2019) Single-cell RNA-Seq reveals AML hierarchies relevant to disease progression and immunity. *Cell*, **176**, 1265–1281.
37. Fan, X., Tang, D., Liao, Y., Li, P., Zhang, Y., Wang, M., Liang, F., Wang, X., Gao, Y., Wen, L., et al. (2020) Single-cell RNA-seq analysis of mouse preimplantation embryos by third-generation sequencing. *PLoS Biol.*, **18**, e3001017.
38. Lebrigand, K., Magnone, V., Barbry, P. and Waldmann, R. (2020) High throughput error corrected Nanopore single cell transcriptome sequencing. *Nat. Commun.*, **11**, 4025.
39. Wang, Q., Boenigk, S., Boehm, V., Gehring, N.H., Altmueller, J. and Dieterich, C. (2021) Single cell transcriptome sequencing on the Nanopore platform with ScNapBar. *RNA*, **27**, 763–770.
40. Schmidt, M.O., Mays, A.D., Barefoot, M.E., Riegel, A.T. and Wellstein, A. (2021) Full-length, single-cell RNA-sequencing of human bone marrow subpopulations reveals hidden complexity. *bioRxiv* doi: <https://doi.org/10.1101/2021.07.28.454226>, 03 August 2021, preprint: not peer reviewed.
41. Palmer, C.R., Liu, C.S., Romanow, W.J., Lee, M.-H. and Chun, J. (2021) Altered cell and RNA isoform diversity in aging Down syndrome brains. *Proc. Natl. Acad. Sci. U.S.A.*, **118**, e2114326118.
42. Long, Y., Liu, Z., Jia, J., Mo, W., Fang, L., Lu, D., Liu, B., Zhang, H., Chen, W. and Zhai, J. (2021) FlsRNA-seq: protoplasting-free full-length single-nucleus RNA profiling in plants. *Genome Biol.*, **22**, 66.
43. Joglekar, A., Prjibelski, A., Mahfouz, A., Collier, P., Lin, S., Schlusche, A.K., Marrocco, J., Williams, S.R., Haase, B., Hayes, A., et al. (2021) A spatially resolved brain region- and cell type-specific isoform atlas of the postnatal mouse brain. *Nat. Commun.*, **12**, 463.
44. Rebboah, E., Reese, F., Williams, K., Balderrama-Gutierrez, G., McGill, C., Trout, D., Rodriguez, I., Liang, H., Wold, B.J. and Mortazavi, A. (2021) Mapping and modeling the genomic basis of differential RNA isoform expression at single-cell resolution with LR-Split-seq. *Genome Biol.*, **22**, 286.
45. Philpott, M., Watson, J., Thakurta, A., Brown, T. Jr, Brown, T. Sr, Oppermann, U. and Cribbs, A.P. (2021) Nanopore sequencing of single-cell transcriptomes with scCOLOR-seq. *Nat. Biotechnol.*, **39**, 1517–1520.
46. Tian, L., Jabbari, J.S., Thijssen, R., Gouil, Q., Amarasinghe, S.L., Voogd, O., Kariyawasam, H., Du, M.R.M., Schuster, J., Wang, C., et al. (2021) Comprehensive characterization of single-cell full-length isoforms in human and mouse with long-read sequencing. *Genome Biol.*, **22**, 310.
47. Boileau, E., Li, X., Naarmann-de Vries, I.S., Becker, C., Casper, R., Altmüller, J., Leuschner, F. and Dieterich, C. (2022) Full-length spatial transcriptomics reveals the unexplored isoform diversity of the myocardium post-MI. *Front. Genet.*, **13**, 912572.
48. Ebrahimi, G., Orabi, B., Robinson, M., Chauve, C., Flannigan, R. and Hach, F. (2022) Fast and accurate matching of cellular barcodes across short-reads and long-reads of single-cell RNA-seq experiments. *iScience*, **25**, 104530.
49. Hardwick, S.A., Hu, W., Joglekar, A., Fan, L., Collier, P.G., Foord, C., Balacco, J., Lanjewar, S., Sampson, M.M., Koopmans, F., et al. (2022) Single-nuclei isoform RNA sequencing unlocks barcoded exon connectivity in frozen brain tissue. *Nat. Biotechnol.*, **40**, 1082–1092.
50. Volden, R. and Vollmers, C. (2022) Single-cell isoform analysis in human immune cells. *Genome Biol.*, **23**, 47.
51. Oguchi, Y., Ozaki, Y., Abdelmoez, M.N. and Shintaku, H. (2021) NanoSINC-seq dissects the isoform diversity in subcellular compartments of single cells. *Sci. Adv.*, **7**, eabe0317.
52. Hazzard, B., Sá, J.M., Ellis, A.C., Pascini, T.V., Amin, S., Welles, T.E. and Serre, D. (2022) Long read single cell RNA sequencing reveals the isoform diversity of Plasmodium vivax transcripts. *PLoS Negl. Trop. Dis.*, **16**, e0010991.
53. Liao, Y., Liu, Z., Zhang, Y., Lu, P., Wen, L. and Tang, F. (2023) High-throughput and high-sensitivity full-length single-cell RNA-seq analysis on third-generation sequencing platform. *Cell Discov.*, **9**, 5.
54. Shi, Z.-X., Chen, Z.-C., Zhong, J.-Y., Hu, K.-H., Zheng, Y.-F., Chen, Y., Xie, S.-Q., Bo, X.-C., Luo, F., Tang, C., et al. (2023) High-throughput and high-accuracy single-cell RNA isoform analysis using PacBio circular consensus sequencing. *Nat. Commun.*, **14**, 2631.
55. Lebrigand, K., Bergensträhle, J., Thrane, K., Mollbrink, A., Meletis, K., Barbry, P., Waldmann, R. and Lundberg, J. (2023) The spatial landscape of gene expression isoforms in tissue sections. *Nucleic Acids Res.*, **51**, e47.
56. Mincarelli, L., Uzun, V., Wright, D., Scoones, A., Rushworth, S.A., Haerty, W. and Macaulay, I.C. (2023) Single-cell gene and isoform expression analysis reveals signatures of ageing in haematopoietic stem and progenitor cells. *Commun. Biol.*, **6**, 558.
57. You, Y., Prawer, Y.D.J., De Paoli-Iseppi, R., Hunt, C.P.J., Parish, C.L., Shim, H. and Clark, M.B. (2023) Identification of cell barcodes from long-read single-cell RNA-seq with BLAZE. *Genome Biol.*, **24**, 66.
58. Liu, H., Liu, H., Wang, L., Song, L., Jiang, G., Lu, Q., Yang, T., Peng, H., Cai, R., Zhao, X., et al. (2023) Cochlear transcript diversity and its role in auditory functions implied by an otoferlin short isoform. *Nat. Commun.*, **14**, 3085.
59. Yang, Y., Yang, R., Kang, B., Qian, S., He, X. and Zhang, X. (2023) Single-cell long-read sequencing in human cerebral organoids uncovers cell-type-specific and autism-associated exons. *Cell Rep.*, **42**, 113335.
60. Kim, H.S., Grimes, S.M., Hooker, A.C., Lau, B.T. and Ji, H.P. (2021) Single-cell characterization of CRISPR-modified transcript isoforms with nanopore sequencing. *Genome Biol.*, **22**, 331.
61. Smith, T., Heger, A. and Sudbery, I. (2017) UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res.*, **27**, 491–499.
62. Wu, S. and Schmitz, U. (2023) Single-cell and long-read sequencing to enhance modelling of splicing and cell-fate determination. *Comput. Struct. Biotechnol. J.*, **21**, 2373–2380.
63. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. and Satija, R. (2018) Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.*, **36**, 411–420.
64. Wolf, F.A., Angerer, P. and Theis, F.J. (2018) SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.*, **19**, 15.
65. Stein, A.N., Joglekar, A., Poon, C.-L. and Tilgner, H.U. (2022) ScisrWiz: visualizing differential isoform expression in single-cell long-read data. *Bioinformatics*, **38**, 3474–3476.

66. Davidson, N.M., Chen, Y., Sadras, T., Ryland, G.L., Blombery, P., Ekert, P.G., Göke, J. and Oshlack, A. (2022) JAFFAL: detecting fusion genes with long-read transcriptome sequencing. *Genome Biol.*, **23**, 10.
67. Shiao, C.-K., Lu, L., Kieser, R., Fukumura, K., Pan, T., Lin, H.-Y., Yang, J., Tong, E.L., Lee, G., Yan, Y., et al. (2023) High throughput single cell long-read sequencing analyses of same-cell genotypes and phenotypes in human tumors. *Nat. Commun.*, **14**, 4124.
68. Rijk, P.D., Watzeels, T., Küçükali, F., Dongen, J.V., Faura, J., Willems, P., Deyn, L.D., Duchateau, L., Grönes, C., Eekhout, T., et al. (2024) Scywalker: scalable end-to-end data analysis workflow for nanopore single-cell transcriptome sequencing. bioRxiv doi: <https://doi.org/10.1101/2024.02.22.581508>, 24 February 2024, preprint: not peer reviewed.
69. Huang, L., Ma, F., Chapman, A., Lu, S. and Xie, X.S. (2015) Single-cell whole-genome amplification and sequencing: methodology and applications. *Annu. Rev. Genomics Hum. Genet.*, **16**, 79–102.
70. Fan, X., Yang, C., Li, W., Bai, X., Zhou, X., Xie, H., Wen, L. and Tang, F. (2021) SMOOTH-seq: single-cell genome sequencing of human cells on a third-generation sequencing platform. *Genome Biol.*, **22**, 195.
71. Chen, C., Xing, D., Tan, L., Li, H., Zhou, G., Huang, L. and Xie, X.S. (2017) Single-cell whole-genome analyses by Linear Amplification via Transposon Insertion (LIANTI). *Science*, **356**, 189–194.
72. Hard, J., Mold, J.E., Eisefeldt, J., Tellgren-Roth, C., Haggqvist, S., Bunikis, I., Contreras-Lopez, O., Chin, C.S., Nordlund, J., Rubin, C.J., et al. (2023) Long-read whole-genome analysis of human single cells. *Nat. Commun.*, **14**, 5164.
73. Marcy, Y., Ishoe, Y., Lasken, R.S., Stockwell, T.B., Walenz, B.P., Halpern, A.L., Beeson, K.Y., Goldberg, S.M.D. and Quake, S.R. (2007) Nanoliter reactors improve multiple displacement amplification of genomes from single cells. *PLoS Genet.*, **3**, 1702–1708.
74. Rhee, M., Light, Y.K., Meagher, R.J. and Singh, A.K. (2016) Digital droplet multiple displacement amplification (ddMDA) for whole genome sequencing of limited DNA samples. *PLoS One*, **11**, e0153699.
75. Xie, H., Li, W., Hu, Y., Yang, C., Lu, J., Guo, Y., Wen, L. and Tang, F. (2022) De novo assembly of human genome at single-cell levels. *Nucleic Acids Res.*, **50**, 7479–7492.
76. Smallwood, S.A., Lee, H.J., Angermueller, C., Krueger, F., Saadeh, H., Peat, J., Andrews, S.R., Stegle, O., Reik, W. and Kelsey, G. (2014) Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nat. Methods*, **11**, 817–820.
77. Gu, H., Raman, A.T., Wang, X., Gaiti, F., Chaligne, R., Mohammad, A.W., Arczewski, A., Smith, Z.D., Landau, D.A., Aryee, M.J., et al. (2021) Smart-RRBS for single-cell methylome and transcriptome analysis. *Nat. Protoc.*, **16**, 4004–4030.
78. Cusanovich, D.A., Daza, R., Adey, A., Pliner, H.A., Christiansen, L., Gunderson, K.L., Steemers, F.J., Trapnell, C. and Shendure, J. (2015) Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science*, **348**, 910–914.
79. Rotem, A., Ram, O., Shores, N., Sperling, R.A., Goren, A., Weitz, D.A. and Bernstein, B.E. (2015) Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. *Nat. Biotechnol.*, **33**, 1165–1172.
80. Simpson, J.T., Workman, R.E., Zuzarte, P.C., David, M., Dursi, L.J. and Timp, W. (2017) Detecting DNA cytosine methylation using nanopore sequencing. *Nat. Methods*, **14**, 407–410.
81. Liu, Y., Cheng, J., Siejka-Zielińska, P., Weldon, C., Roberts, H., Lopopolo, M., Magri, A., D'Arienzo, V., Harris, J.M., McKeating, J.A., et al. (2020) Accurate targeted long-read DNA methylation and hydroxymethylation sequencing with TAPS. *Genome Biol.*, **21**, 54.
82. Sun, Z., Vaisvila, R., Hussong, L.M., Yan, B., Baum, C., Saleh, L., Samaranyake, M., Guan, S., Dai, N., Correa, I.R. Jr, et al. (2021) Nondestructive enzymatic deamination enables single-molecule long-read amplicon sequencing for the determination of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. *Genome Res.*, **31**, 291–300.
83. Feng, Z., Fang, G., Korlach, J., Clark, T., Luong, K., Zhang, X., Wong, W. and Schadt, E. (2013) Detecting DNA modifications from SMRT sequencing data by modeling sequence context dependence of polymerase kinetic. *PLoS Comput. Biol.*, **9**, e1002935.
84. Schneider, C., Woehle, C., Greve, C., D'Haese, C.A., Wolf, M., Hiller, M., Janke, A., Bálint, M. and Huettel, B. (2021) Two high-quality de novo genomes from single ethanol-preserved specimens of tiny metazoans (Collembola). *Gigascience*, **10**, giab035.
85. Hu, Y., Jiang, Z., Chen, K., Zhou, Z., Zhou, X., Wang, Y., Yang, J., Zhang, B., Wen, L. and Tang, F. (2023) scNanoATAC-seq: a long-read single-cell ATAC sequencing method to detect chromatin accessibility and genetic variants simultaneously within an individual cell. *Cell Res.*, **33**, 83–86.
86. Chen, X., Miragaia, R.J., Natarajan, K.N. and Teichmann, S.A. (2018) A rapid and robust method for single cell chromatin accessibility profiling. *Nat. Commun.*, **9**, 5345.
87. Granja, J.M., Corces, M.R., Pierce, S.E., Bagdatli, S.T., Choudhry, H., Chang, H.Y. and Greenleaf, W.J. (2021) ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nat. Genet.*, **53**, 403–411.
88. Kumar, S. and Mohapatra, T. (2021) Deciphering epitranscriptome: Modification of mRNA bases provides a new perspective for post-transcriptional regulation of gene expression. *Front. Cell Dev. Biol.*, **9**, 628415.
89. Shi, H., Chai, P., Jia, R. and Fan, X. (2020) Novel insight into the regulatory roles of diverse RNA modifications: Re-defining the bridge between transcription and translation. *Mol. Cancer*, **19**, 78.
90. Zhao, L., Zhang, H., Kohnen, M.V., Prasad, K.V.S.K., Gu, L. and Reddy, A.S.N. (2019) Analysis of transcriptome and epitranscriptome in plants using PacBio Iso-Seq and Nanopore-based direct RNA sequencing. *Front. Genet.*, **10**, 253.
91. Vilfan, I.D., Tsai, Y.-C., Clark, T.A., Wegener, J., Dai, Q., Yi, C., Pan, T., Turner, S.W. and Korlach, J. (2013) Analysis of RNA base modification and structural rearrangement by single-molecule real-time detection of reverse transcription. *J. Nanobiotechnol.*, **11**, 8.
92. Helm, M. and Motorin, Y. (2017) Detecting RNA modifications in the epitranscriptome: predict and validate. *Nat. Rev. Genet.*, **18**, 275–291.
93. Thijssen, R., Tian, L., Anderson, M.A., Flensburg, C., Jarratt, A., Garnham, A.L., Jabbari, J.S., Peng, H., Lew, T.E., Teh, C.E., et al. (2022) Single-cell multiomics reveal the scale of multilayered adaptations enabling CLL relapse during venetoclax therapy. *Blood*, **140**, 2127–2141.
94. Zhang, C., Fang, Y., Chen, W., Zhang, Y., Xie, Y., Chen, W., Xie, Z., Guo, M., Wang, J., et al. (2023) Improving the RNA velocity approach with single-cell RNA lifecycle (nascent, mature and degrading RNAs) sequencing technologies. *Nucleic Acids Res.*, **51**, e112–e112.
95. Quigley, D.A., Dang, H.X., Zhao, S.G., Lloyd, P., Aggarwal, R., Alunkal, J.J., Foye, A., Kothari, V., Perry, M.D., Bailey, A.M., et al. (2018) Genomic hallmarks and structural variation in metastatic prostate cancer. *Cell*, **174**, 758–769.
96. Zhang, Y., Chen, F., Fonseca, N.A., He, Y., Fujita, M., Nakagawa, H., Zhang, Z., Brazma, A., Group, P.T.W., Group, P.S.V.W., et al. (2020) High-coverage whole-genome analysis of 1220 cancers reveals hundreds of genes deregulated by rearrangement-mediated cis-regulatory alterations. *Nat. Commun.*, **11**, 736.
97. Sandve, G.K. and Greiff, V. (2022) Access to ground truth at unconstrained size makes simulated data as indispensable as experimental data for bioinformatics methods development and benchmarking. *Bioinformatics*, **38**, 4994–4996.
98. Squair, J.W., Gautier, M., Kathe, C., Anderson, M.A., James, N.D., Hutson, T.H., Hudelle, R., Kaiser, T., Matson, K.J.E., Barraud, Q., et al. (2021) Confronting false discoveries in single-cell differential expression. *Nat. Commun.*, **12**, 5692.

Received: February 1, 2024. Revised: April 18, 2024. Editorial Decision: April 22, 2024. Accepted: April 29, 2024

© The Author(s) 2024. Published by Oxford University Press on behalf of NAR Genomics and Bioinformatics.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.