# Benchmarking long-read RNA-sequencing technologies with *LongBench:* a cross-platform reference dataset profiling cancer cell lines with bulk and single-cell approaches

**Yupei You[1,2✉*], Ashleigh Solano[1,2*], James Lancaster[1,2*], Margaux David[3*], Changqing Wang[1,2*], Shian Su[1,2], Kathleen Zeglinski[1,2], Reza Ghamsari[4], Manveer Chauhan[5], Josie Gleeson[5,6], Yair D. J. Prawer[5], Jin Ng[1,2], Bénédicte Dubois[3], Isabelle Cleynen[7], Marie-Liesse Asselin-Labat[1,2], Kate D. Sutherland[1,2], Michael B. Clark[5], Quentin Gouil[8,9†✉], and Matthew E. Ritchie[1,2†✉]**

[1]The Walter and Eliza Hall Institute of Medical Research, Parkville, Victoria, Australia
[2]Department of Medical Biology, The University of Melbourne, Parkville, Victoria, Australia
[3]Laboratory for Neuroimmunology, Department of Neurosciences, Leuven Brain Institute, KU Leuven, Leuven, Belgium
[4]Graduate School of Biomedical Engineering, University of New South Wales, Kensington, New South Wales, Australia
[5]Department of Anatomy and Physiology, The University of Melbourne, Parkville, Victoria, Australia
[6]Science for Life Laboratory, Department of Gene Technology, KTH Royal Institute of Technology, Solna, Sweden
[7]Laboratory for Complex Genetics, Department of Human Genetics, Faculty of Medicine, KU Leuven, Leuven, Belgium
[8]Olivia Newton-John Cancer Research Institute, Heidelberg, Victoria, Australia
[9]School of Cancer Medicine, La Trobe University, Bundoora, Victoria, Australia
[*]These authors contributed equally as co-first
[†]These authors contributed equally as co-last

**Long-read RNA sequencing enables full-length transcript profiling and improved isoform resolution, but variable platforms and evolving chemistries demand careful benchmarking for reliable application. We present *LongBench*, a matched, multi-platform reference dataset spanning bulk, single-cell, and single-nucleus transcriptomics across eight human lung cancer cell lines with synthetic spike-in controls. *LongBench* incorporates three state-of-the-art long-read protocols alongside Illumina short reads: Oxford Nanopore Technologies (ONT) PCR-cDNA, ONT direct RNA, and PacBio Kinnex. We systematically evaluate transcript capture, quantification accuracy, differential expression, isoform usage, variant detection, and allele-specific analyses. Our results show high concordance in gene-level differential analyses across protocols, but reduced consistency for transcript-level and isoform analyses due to length- and platform-dependent biases. Single-cell long-read data are highly concordant with bulk for high-confidence features, though single-nuclei data show reduced feature detection. *LongBench* provides one of the largest publicly available long-read benchmarking resources, enabling rigorous cross-platform evaluation and guiding technology selection for transcriptomic research.**

**single-cell | bulk | gene expression | long-read sequencing | ONT | PacBio | methods benchmarking | differential expression**

**Correspondence:** ***you.yu@wehi.edu.au, Quentin.Gouil@onjcri.org.au, mritchie@wehi.edu.au***

Long-read RNA sequencing (RNA-seq) provides access to the still largely untapped complexity of transcriptomes, facilitating isoform discovery and quantification. These capabilities have successfully revealed transcript diversity and highlighted biologically and clinically relevant isoforms in human diseases, including cancer[1–5] and neurological disorders[6–9]. Pacific Biosciences (PacBio)[10] and Oxford Nanopore Technologies (ONT)[11] are leading long-read platforms with distinct strategies. PacBio's recently released Kinnex protocol[12], a commercialised version of MAS-seq[13], boosts throughput by concatenating cDNAs into extended circular inserts. ONT has introduced Kit 14 and RNA004 chemistry for cDNA and direct RNA sequencing respectively, which, along with improved basecalling models, have substantially increased read accuracy. Advances in instrumentation (e.g., PacBio Revio, ONT PromethION) have pushed cDNA yields to over 100 million reads per flow cell[14], enabling high-throughput transcriptomics. ONT direct RNA sequencing also allows simultaneous detection of RNA base modifications, while avoiding reverse transcription and PCR biases[15–18]. Both technological and bioinformatics landscapes continue to evolve rapidly, highlighting the need for systematic benchmarking.

The rapid development and growth of both long-read RNA-seq and single-cell transcriptomics catalysed interest in developing long-read single-cell approaches, which have revealed the isoform information and genetic variants missed by standard short-read single-cell RNA-seq[1,19,20]. Early attempts to apply ONT and PacBio to single-cell RNA-seq faced key limitations in read accuracy and sequencing yield. These necessitated the creation of custom protocols to improve accuracy, such as R2C2[21], or the use of matched short-read scRNA-seq to provide key information such as cell barcodes, cell expression profiles or cell-type identification[1,19,22–27]. With recent improvements in software, throughput and sequencing accuracy, ONT–only protocols without short-reads have become feasible[20,28]. Concurrently, the standardisation of PacBio "HiFi" sequencing and developments in PacBio single-cell RNA-seq protocols, including targeted enrichment of barcoded cDNA[29] and improved read yield in Kinnex, have enhanced isoform recovery and sensitivity[19,30,31]. As single-cell long-read protocols

evolve, an expanding ecosystem of computational tools has emerged to support tasks such as cell barcode demultiplexing, isoform annotation, and transcript quantification[1,20,28,32–36]. Nevertheless, knowledge of the comparative performance of different methods remains limited, highlighting the importance of comprehensive benchmarking to properly evaluate different methods.

Several foundational benchmarking studies have provided key insights into isoform detection, transcript quantification and platform-specific biases, yet are often constrained by study designs that limit their relevance to current workflows. Weirather et al.[37] found fewer false positives for PacBio and improved RNA modification detection with ONT. Soneson et al.[38] revealed protocol-specific transcript coverage biases with ONT cDNA and direct RNA. Dong et al.[39] compared ONT PromethION and Illumina, highlighting differences in isoform quantification and differential expression analyses. Larger studies, including LRGASP[40] and SG-NEx[41], demonstrated variability in isoform discovery, improved major isoform quantification, and enhanced transcript diversity with long reads. More recently, Wissel et al.[42] reported enhanced transcript-level precision using PacBio Kinnex, and Zajac et al.[43] and Scoones et al.[44] highlighted advances in single-cell long-read analyses. Together, these studies have established essential groundwork for long-read RNA-seq benchmarking. However, notable limitations persist: (1) most benchmarks evaluate only a subset of platforms (Supplementary Fig.1A); (2) many rely on synthetic or homogeneous inputs with minimal biological replication (Supplementary Fig.1B), restricting generalisability; and (3) due to the fast-moving nature of this field, most feature obsolete or outdated sequencing chemistries or instrumentation. To support future transcriptomic research, comprehensive benchmarking resources that incorporate current platforms, biological diversity, and multiple data modalities are needed.

Here we present *LongBench*, a long-read transcriptomic dataset for benchmarking sample preparation workflows, sequencing technologies and analysis methods. *LongBench* includes 3 modalities (bulk, single-cell and single-nucleus transcriptomics) across 3 long-read sequencing technologies (ONT PCR-cDNA and PacBio Kinnex for all modalities and ONT direct RNA for bulk). The experimental design included samples from 8 lung cancer cell lines representing 3 distinct disease groups and spiked-in transcripts. A diverse panel of cancer cell lines was included to capture realistic biological variation with all samples sequenced deeply. These design features facilitated systematic assessment of transcript capture efficiency and quantification accuracy across transcripts of varying lengths and abundances, as well as differential expression testing between disease groups. The single-cell and single-nucleus components enable transcript-level analyses at cell resolution, while also allowing assessment of the impact of additional library preparation and computational steps. By providing a comprehensive resource for comparing sequencing platforms and analysis tools across diverse conditions, this dataset aims to guide researchers in making informed decisions and advancing the use of long-read RNA-seq in transcriptomic research.

## Results

**Experimental design with realistic levels of biological variation.** *LongBench* includes eight lung cancer cell lines derived from patients representing three distinct cancer types (Fig.1A): Lung Adenocarcinoma (LUAD) and two Small Cell Lung Cancer (SCLC) subtypes, SCLC-A and SCLC-P, distinguished by transcriptional regulators, ASCL1 and POU2F3[45]. Synthetic RNA spike-ins, namely Sequins[46] (mix A or B) and Lexogen SIRV set 4, were added to each bulk RNA sample to provide ground truth internal standards (Supplementary Table 1). A total of 32 bulk RNA-seq libraries were generated using four sequencing protocols: ONT PCR-cDNA, ONT direct RNA (dRNA), PacBio Kinnex cDNA, and Illumina short-read sequencing. In parallel, a pooled mixture of cells from all eight cell lines was prepared for single-cell and single-nucleus RNA-seq. cDNA libraries were generated with 10X Genomics Chromium and sequenced with ONT PCR-cDNA, PacBio Kinnex cDNA, and Illumina, yielding three scRNA-seq and three snRNA-seq datasets. Together, these bulk, single-cell, and single-nucleus datasets enable a comprehensive cross-platform comparison. Our short-read sequencing data validated the design and cell line selection by demonstrating clustering by cancer type while preserving realistic within-group biological variability (Fig.1B-C), supporting the reliability and biological relevance of subsequent differential analysis. With *LongBench*, we present an overview of multiple downstream analyses (Fig.1D), detailed in subsequent sections.

***LongBench* features high-quality, deeply sequenced RNA-seq data.** *LongBench* comprises 2.17 billion total reads, including 1.45 billion long reads—making it one of the largest publicly available long-read datasets to date (Fig.1E, Supplementary Fig.1C,D). We ensured deep sequencing with at least 23.6 million reads per bulk dataset and 102 million reads per single-cell or single-nucleus dataset, enabling robust gene and isoform detection, accurate quantification, and increased power for downstream analyses, providing a fair basis for cross-platform comparison.

Overall, *LongBench* achieved high-quality data across platforms, with read length and accuracy reflecting inherent technological differences (Fig.1F). The PacBio bulk datasets yielded the largest median read lengths overall (median length 1.53–1.75 kb), followed by ONT direct RNA (dRNA, median length 0.92–1.20 kb) and ONT PCR cDNA datasets (median length 0.59–0.67 kb). However, ONT dRNA produced the longest individual reads (Supplementary Fig.2A). Notably, 10X library preparation yielded similar read lengths between PacBio and ONT (median length 0.74–0.85 kb). These length differences can be linked to library preparation: bulk PacBio Kinnex includes size selection that depletes shorter fragments (0.86–0.9X bead wash after cDNA amplification[47]); this step is not present in the single-cell/single-nuclei protocols. Bias against longer transcripts can arise

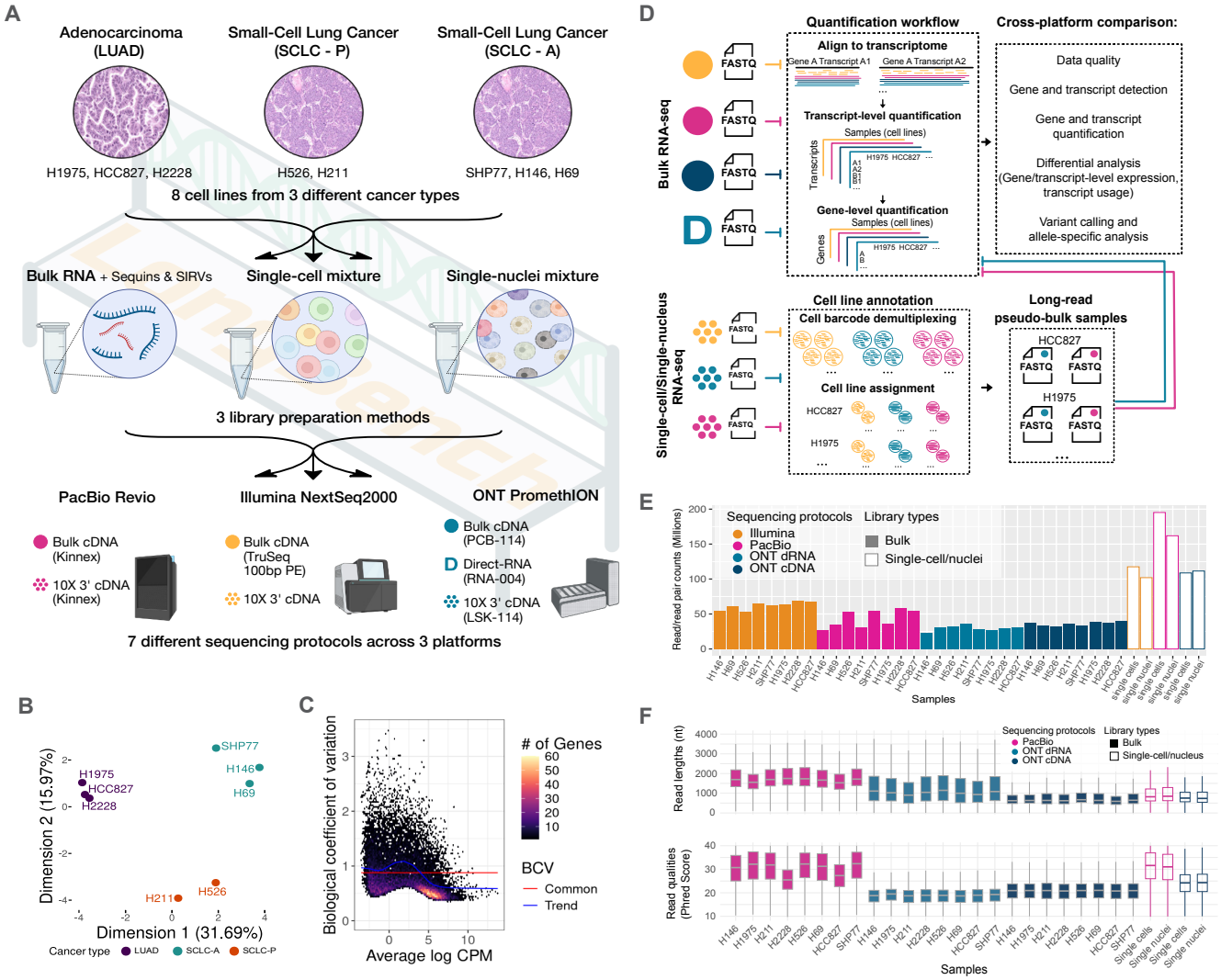


**Fig. 1.** *LongBench* **study design and data overview. A.** Design of the *LongBench* study, which included bulk, single-cell and single-nucleus samples derived from 8 lung cancer cell lines, profiled across 3 long-read sequencing technologies, including Oxford Nanopore Technologies PCR-cDNA and PacBio Kinnex for all modalities and Oxford Nanopore Technologies direct RNA sequencing for bulk, as well as Illumina short-read cDNA sequencing across all sample types. **B.** Multidimensional scaling (MDS) plot of samples based on the gene quantification matrix from short-read RNA-seq dataset in *LongBench*. **C.** Biological coefficient of variation (BCV) generated from the short-read RNA-seq dataset using data from the 8 cancer cell lines. The BCV was estimated using edgeR. **D.** Summary of the data available and the main analyses performed in this study. Colours indicate the different sequencing technologies, consistent with those in A. **E.** The number of reads generated for each sample. **F.** A summary of read lengths and qualities from each library.

from incomplete reverse transcription (all protocols), premature template-switching (all cDNA protocols) and PCR bias (PCR-based protocols)[48,49]. PacBio achieved the highest per-base accuracy and the largest proportion of reads aligned to the genome across the bulk datasets. However, it yielded more intronic reads than the ONT datasets (Fig.2A), likely due to a higher chance of reverse transcriptase priming on internal A-rich regions within pre-mRNA instead of the true poly-A tails (Fig.2B, Supplementary Fig.3). In contrast, ONT dRNA datasets had the lowest proportion of non-exonic reads (Fig.2A), and importantly, it also uniquely enables reliable detection of RNA modifications (Supplementary Fig.4). All long-read platforms produced reads that span nearly the full length of transcripts shorter than 1 kb, which represent $\sim 60\%$ of annotated human transcripts. However, the coverage decreased with increasing transcript length, with ONT PCR cDNA showing the most pronounced decline among the three long-read platforms (Fig.2C).

**Long-read platforms consistently reduce transcript assignment ambiguity.** Accurate transcript-level quantification remains a fundamental challenge in RNA-seq analysis, particularly for multi-isoform genes. To evaluate whether long-read platforms consistently reduce transcript assignment ambiguity, we compared the distributions of read-to-transcript ambiguity across protocols. For short reads, ambiguity greatly increased with isoform diversity (Fig.2D), while long reads showed comparatively stable transcript assignment across all protocols, underscoring their advantage in resolving transcript-level expression.

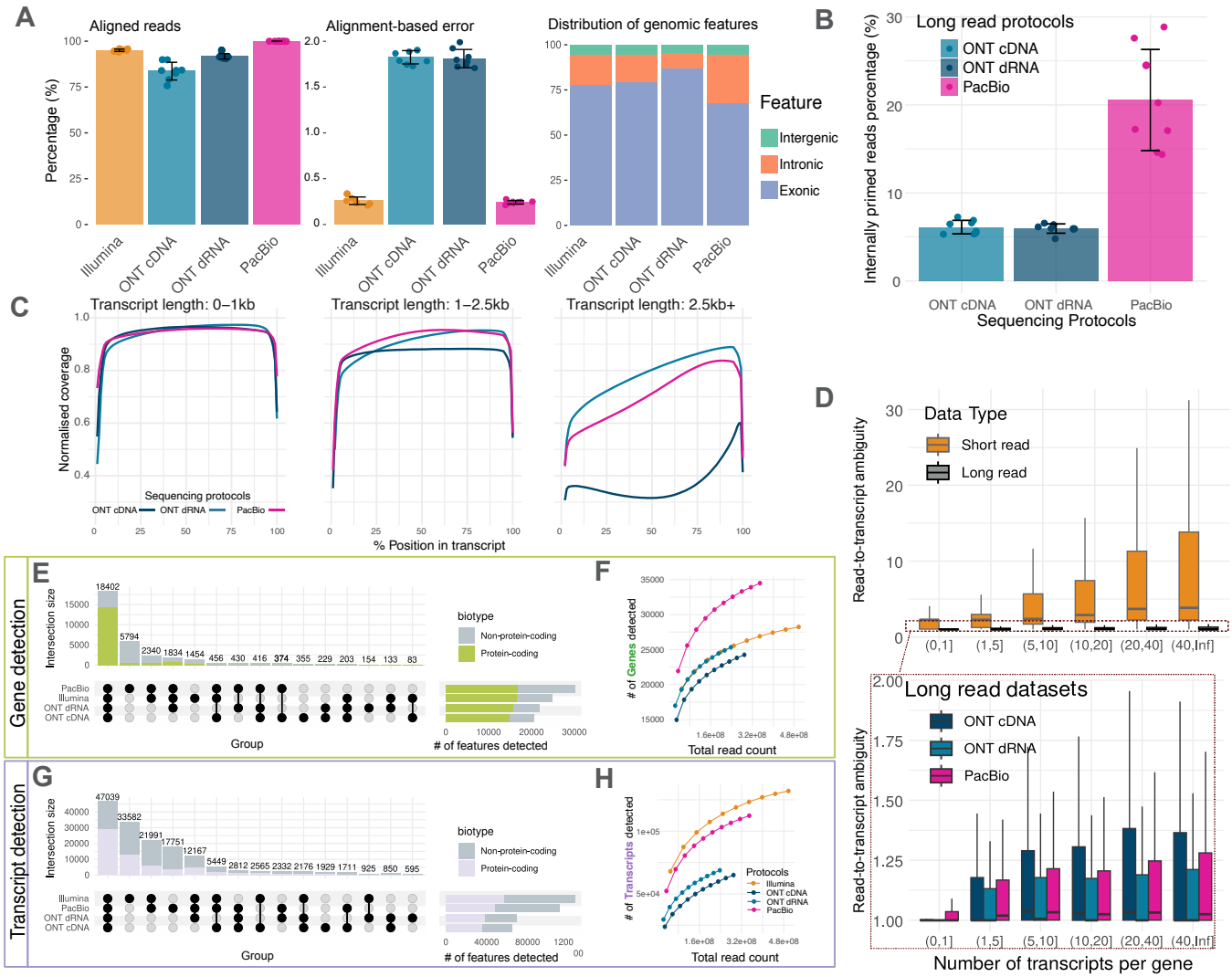


**Fig. 2. Alignment-based read quality, read-to-gene and read-to-transcript metrics for bulk RNA-seq datasets. A.** Alignment-based quality summary, including the percentage of aligned reads, rates of mismatches, insertions, and deletions, as well as the distribution of reads across genomic features. **B.** Percentage of internally primed reads in long-read RNA-seq datasets. The proportions of the internally primed reads were estimated using a custom tool, PrimeSpotter (`https://github.com/youyupei/PrimeSpotter`), to identify reads likely resulting from internal priming at genomically encoded A stretches within transcripts, rather than at genuine polyA tails. **C.** Normalised transcript coverage plot. Transcripts were grouped into length bins, and for each group, the normalised coverage (y-axis) is shown across transcript percentiles (x-axis), where 0% corresponds to the 5 end and 100% to the 3 end of the transcripts. Normalised coverage at a given position percentile was calculated as the number of reads covering that percentile position divided by the total number of reads mapped to the transcripts in the group. **D.** Read-to-transcript ambiguity estimation following the approach of Baldoni et al. [50]. **E, G.** Overlaps of detected genes (E) and transcripts (G) from the GENCODE annotation **F, H.** Rarefaction curve showing the number of genes (F) and transcripts (H) detected when down-sampling the data to account for the effect of different sequencing depths across the datasets.

**Differences in detection of expressed genes and transcripts across platforms.** We next evaluated how different bulk RNA-seq protocols detect comparable or distinct gene and transcript sets (Fig.2E-H). Among the long-read datasets, PacBio detected the most genes and transcripts. Detected protein-coding features were largely consistent across platforms, whereas non-protein-coding features were more platform-specific.

Within long-read datasets, PacBio detected twice as many non-protein-coding features (Fig.2E,G). However, the PacBio reads do not cover significantly more splice junctions with matched depth (Supplementary Fig.2B), and many PacBio-specific genes are single-exon genes that overlapped with introns of other genes (Supplementary Fig.5A). Given the higher proportion of intronic reads and propensity for internal priming in PacBio, a possible explanation is that some reads assigned to non-coding genes originate from intronic regions amplified from incompletely spliced overlapping coding transcripts, making false positive detections (Supplementary Fig.5A,B). In contrast, ONT cDNA missed many longer genes and transcripts that were detected by all other platforms (Supplementary Fig.5C,D), consistent with the finding of limited coverage for long transcripts (Fig.2C). Notably, short-read data reported significantly more transcripts than long-read data at matched read depth (Fig.2H), likely due to the higher ambiguity in isoform assignment (Fig.2D), leading to the assignment of reads to reference isoforms not actually expressed in the sample. Taken together, the different limitations demonstrated by PacBio (intronic reads, internal priming and possible false positive detections), ONT cDNA (poor coverage of longer transcripts, false negative detections), and Illumina (ambiguous isoform assign-

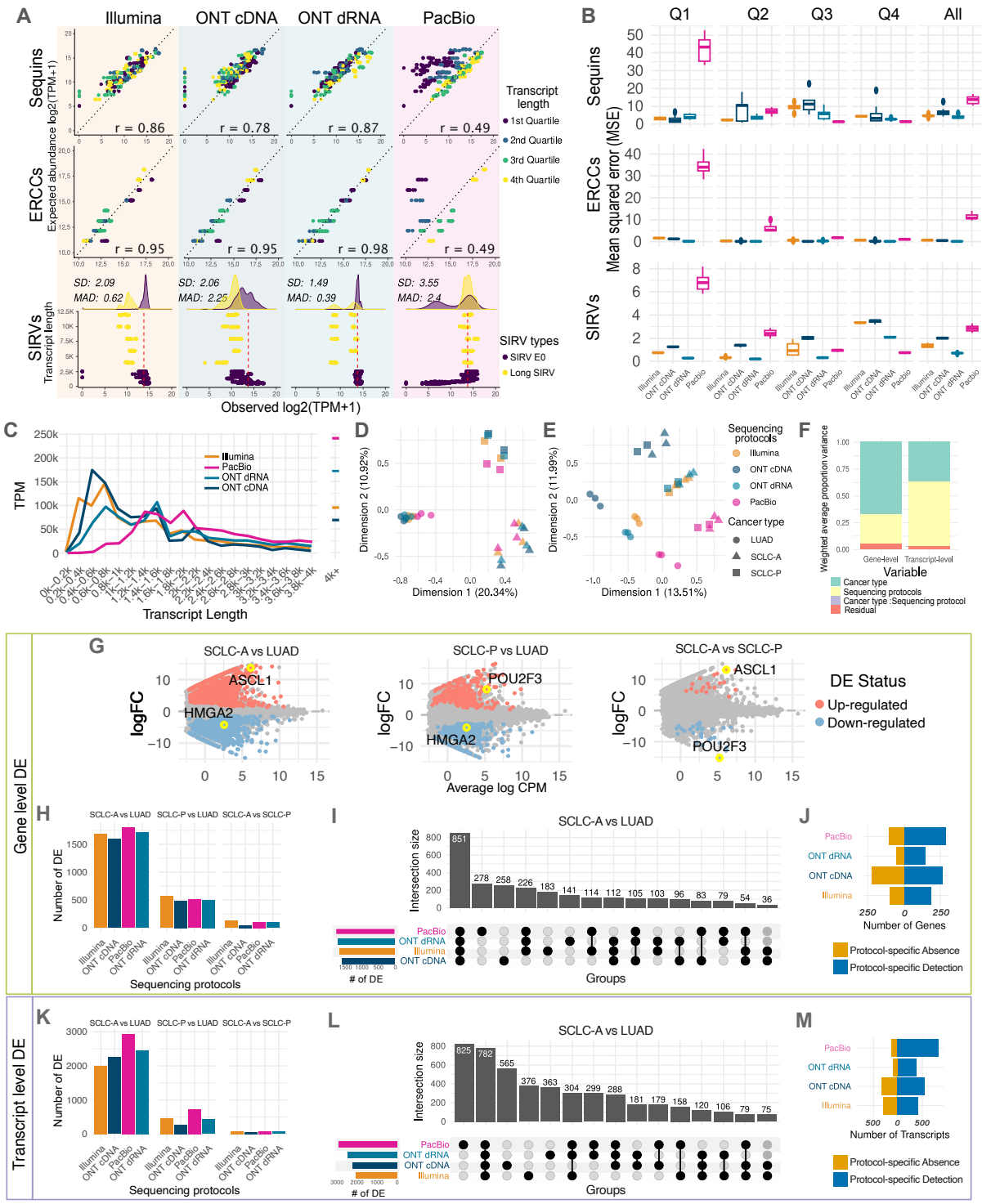


**Fig. 3. Quantification and differential analysis on bulk RNA-seq datasets. A.** Comparison of observed $\log_2(\text{TPM})$ of spike-in transcripts to expected abundance. Each dot represents one transcript in one cell line, with all eight cell lines shown in each panel. For Sequins and ERCC, each transcript has a distinct expected abundance (y-axis) and is coloured by transcript length. Pearson correlation (r) is calculated to assess the consistency between observed and expected values. For SIRVs, all transcripts are expected to have equal abundance (indicated by red lines), with transcript length plotted on the y-axis. Standard deviation (SD) and Mean Absolute Error (MAD) are calculated to quantify the biases in quantification. **B.** Mean squared errors for spike-in transcript quantification in different length quartiles. *Q1* to *Q4* represent the first to the fourth length quartiles in each type of spike-ins. **C.** Count distribution across human transcripts with different lengths. Transcript-level counts were summed across eight cell lines and then converted to TPM. **D, E.** Multidimensional scaling (MDS) plot of samples based on gene-level (E) and transcript-level (E) quantification. **F.** Principal Variance Component Analysis (PVCA) showing the proportion of total variance in gene- and transcript-level quantification explained by each factor and their interactions. **G.** Mean-Average (MA) plot with marker genes of each cancer type highlighted. The plot was generated using the bulk ONT cDNA dataset. The same markers are successfully detected from the other datasets (Supplementary Fig.9C). **H, K.** Number of gene-level (H) and transcript-level (K) DE identified across three cancer-type comparisons. **I, L.** Upset plots showing the overlap of differentially expressed genes (I) and transcripts (L) identified from each sequencing protocol in the comparison between LUAD and SCLC-A. **J, M.** Number of differentially expressed genes (J) and transcripts (M) that are specifically identified or missed by each sequencing protocol in the LUAD vs SCLC-A comparison.

ment) suggest ONT dRNA may provide a more accurate view of transcribed genes and isoforms.

**Large variation in transcript quantification across platforms.** We next examined how protocol choice affected transcript quantification. We first compared quantification accuracy using spike-in controls with known concentrations (Fig.3A-B, Supplementary Fig.6, Supplementary Information Section 2.1). All platforms showed good agreement between observed and expected transcript abundances, with ONT dRNA consistently achieving the highest correlation. However, length-dependent platform-specific biases were evident: shorter spike-ins were underestimated in PacBio, whereas longer ones, like long SIRVs, were underestimated in other datasets, particularly in the ONT cDNA datasets.

A similar pattern was observed for human transcripts, with PacBio undercounting transcripts below 1 kb and ONT cDNA overestimating them relative to other protocols (Fig.3C). Gene-level correlations were generally high, with the strongest concordance observed between ONT dRNA and Illumina datasets (Supplementary Fig. 7). In contrast, transcript-level correlations exhibited substantial variability, with ONT direct RNA and PacBio datasets showing the highest agreement. To assess whether variation reflected underlying biology, we applied MDS and PVCA, which showed that gene-level variation was primarily explained by cancer type, whereas transcript-level variation was dominated by protocol choice (Fig.3D–F).

**Bulk RNA-seq differential expression results are consistent at the gene level but protocol-dependent at the transcript level.** Building on these observations, we next evaluated how platform-specific differences in transcript quantification impact differential expression (DE) analysis. As an initial validation, all protocols accurately recovered the expected fold changes of spike-in transcripts (Supplementary Fig.8), with ONT dRNA and Illumina showing slightly higher accuracy than the others.

With the 18,402 and 47,039 commonly captured human genes and transcripts, we performed DE analysis between each pair of cancer types. As expected from the MDS plots, the most DE genes and transcripts were found between SCLC-A and LUAD (Fig.3H,K), representing distinct cancer types with three samples each. Fewer DE features were seen in SCLC-P vs LUAD, partly due to lower power (only 2 SCLC-P samples), and the least between the SCLC-A and SCLC-P subtypes.

Overlapping DE genes were highly consistent across sequencing protocols (Fig.3I, Supplementary Fig.9A,B, Supplementary Table 2). Most DE genes were detected by all four protocols and included well-established cancer-associated genes. For example, HMGA2, known to be highly expressed in non–small cell lung cancer[51], was significantly upregulated in LUAD across all protocols. Key transcription factors ASCL1 and POU3F3, markers for SCLC-A and SCLC-P, were also consistently upregulated in the expected samples (Fig.3G, Supplementary Fig.9C). Among the sequencing protocols, ONT cDNA showed relatively lower sensitivity, missing the highest number of DE features consistently identified by the other three protocols (Fig.3J, Supplementary Fig.9D,E).

The transcript-level DE results were less consistent across protocols, with many DE transcripts being protocol-specific (Fig.3L, M, Supplementary Fig.9F, Supplementary Table 3). In particular, log-fold change correlations were markedly lower at the transcript level, and in many cases, platforms reported opposite directions for the same transcript (Supplementary Fig.9G,H). These results highlight the vulnerability of transcript-level analysis to platform-specific biases.

**Transcript usage analysis shows moderate agreement in major isoform detection but divergent differential results.** Beyond expression level, isoforms can be analysed in terms of relative isoform usage. We found that the three long-read protocols (ONT cDNA, ONT dRNA, and PacBio) agreed on the major isoform for 67–69% of the multi-isoform genes, whereas short-read results were notably more divergent (Supplementary Fig.10D), likely reflecting the inherent limitations of short-read sequencing in isoform quantification. Moreover, previously described length biases can compromise isoform proportion estimates, especially when isoforms differ substantially in length. When examining differential transcript usage (DTU), the PacBio dataset identified the most DTU events, while ONT cDNA detected the fewest (Supplementary Fig.10E). All four protocols reliably detected known events from the Sequins transcripts (Supplementary Fig.10A–C). This confirms the technical ability of each method to detect isoform proportion changes in a controlled setting. However, similar to previous studies[39,52], substantial inconsistencies were observed in human genes (Supplementary Fig.10F,G), indicating that biological complexity and limited sequencing depth continue to make accurate DTU detection challenging in real samples.

**Consistency of long-read platforms on single-cell and single-nucleus RNA-seq analysis.** Single-cell (SC) and single-nucleus (SN) cDNA libraries were prepared from a mixture of all eight cancer cell lines using the 10X Genomics 3' assay. The same cDNA was then subjected to short-read (Illumina) and long-read (PacBio Kinnex and ONT cDNA) sequencing to compare performance in gene expression analysis tasks. ONT and PacBio protocols assigned comparable proportions of reads to cells (Fig.4A, Supplementary Fig.11). However, a smaller fraction of reads were assigned to nuclei due to elevated ambient RNA introduced during nuclei isolation[53–56]. In addition, a higher doublet rate was observed in SN compared to SC data (Fig.4A).

The identification and SNP-based cell line annotation of single cells and nuclei was consistent across Illumina, PacBio and ONT sequencing platforms (Fig.4B) and validated by expression-based clustering (Fig.4C), with clear separation of all cell lines across all protocols. Notably, H1975 formed two distinct clusters, consistent with previous observations[57]. Higher ambient RNA proportion in the SN data again decreased SNP-based annotation success compared to SC (Fig.4B,C). Together with the variation of cell size across

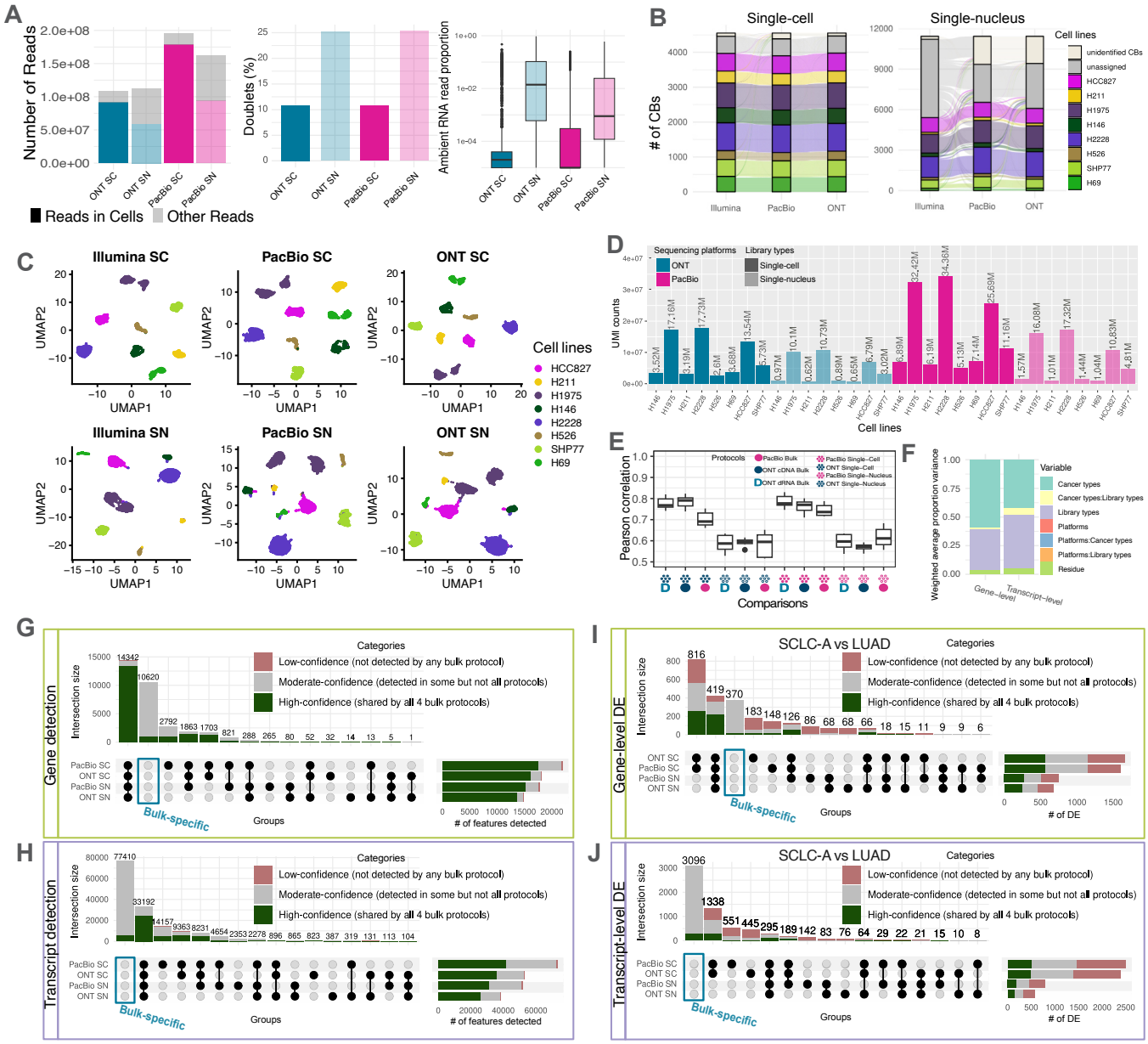


**Fig. 4. Single-cell and single-nucleus RNA-seq data analysis. A.** Summary of library quality and read demultiplexing metrics, including the number of reads in cells, doublet rate and proportion of ambient RNA reads. **B.** Genotype-based cell line annotation for single cells and nuclei. Each line connecting the bars represents a matching cell barcode across sequencing platforms. Unassigned cells were excluded from downstream analyses. **C.** UMAP of gene expression profiles from *LongBench* single-cell and single-nucleus datasets, with cells colored by genotype-based cell line labels. **D.** Total UMI counts per cell line among successfully annotated cells. **E.** Pearson correlation of transcript-level quantification between pseudo-bulk and bulk samples was calculated for each cell line, comparing the pseudo-bulk and bulk samples from the same cell line. Only transcripts with non-zero counts across all protocols were included in the calculation. Correlation values were then summarised by the protocol pairs compared, which are indicated on the x-axis. **G, I.** Overlap of detected genes (G) and transcripts (I) in pseudo-bulk samples. Features are classified as low-, moderate-, or high-confidence based on the number of bulk RNA-seq datasets in which they were also detected. **H, J.** Overlap of gene-level (H) and transcript-level (J) DE results in pseudo-bulk samples. DE features are categorised by the number of bulk protocols in which they were also identified as DE.

the cell lines, cell lines with larger cells/nuclei generated on the order of 10-times more reads, and pseudo-bulk libraries by cell line ranged from 0.65 to 34M reads (Fig.4D).

Pseudo-bulk quantification showed moderate to high correlation to matched bulk datasets (Fig.4E), with SC samples correlating better than SN samples. Across 32 pseudo-bulk samples (Fig.4D), library type (SC vs SN) and cancer type explained most variation, while sequencing platform (PacBio vs ONT) had minimal impact (Fig.4F, Supplementary Fig.11C-F), contrasting with bulk RNA-seq protocols. While individual single cells exhibited sparse feature detection and quantification (Supplementary Fig.11A), pseudo-bulk aggregation recovered most high-confidence features commonly detected in bulk datasets, although many features seen in bulk remained absent from pseudo-bulk datasets (Fig.4G,H). Downsampling the bulk data to a similar depth demonstrated the potential for a substantial increase in feature detection with deeper sequencing (Supplementary Fig.11B). Notably, SC and SN datasets detected only a few low-confidence features, highlighting low levels of spurious feature detection.

We then performed gene- and transcript-level differential expression (DE) analysis on high-confidence features com-

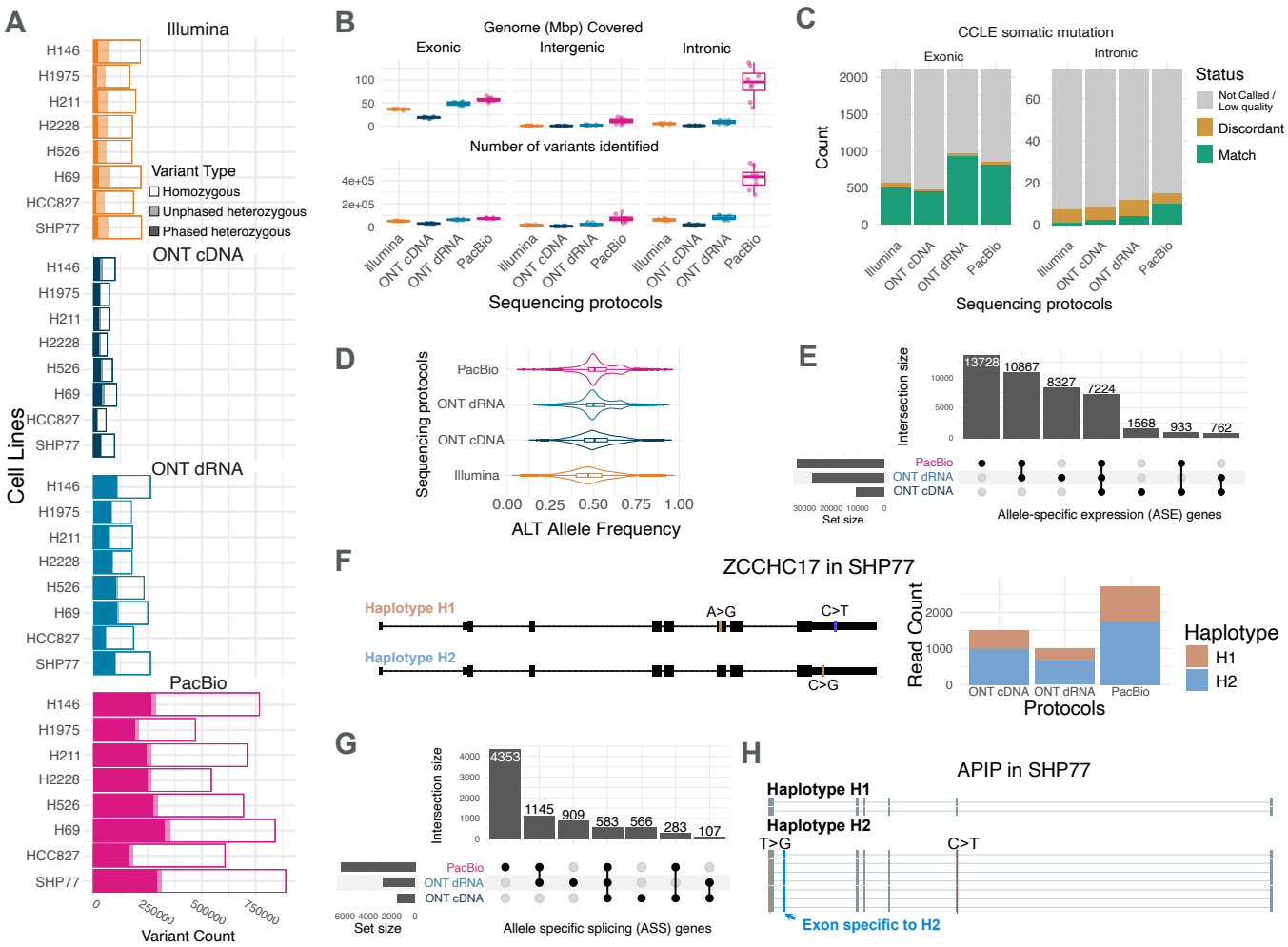


**Fig. 5. Variant calling and allele-specific analyses: A.** Number of variants called by Clair3-RNA (long reads) and Clair3-Illumina (short reads), phased using WhatsHap. **B**. Coverage and number of variants detected across different genomic regions. **C.** Detection of known somatic mutations annotated in the Cancer Cell Line Encyclopedia (CCLE). **D.** Alternative allele frequency distributions from Clair3-RNA and Clair3-Illumina across datasets. **E.** Allele-specific expression (ASE) genes from different bulk long-read datasets identified using LongcallR. The analysis is performed independently on each cell line, and the sum of the results is presented. A gene detected in multiple cell lines is counted once per cell line. **F.** Example of the ASE gene *ZCCHC17* in SHP77, consistently identified across all long-read protocols; haplotype-distinguishing variants are highlighted. **G.** Overlaps of allele-specific splicing (ASS) genes from different bulk long-read datasets identified using LongcallR. The analysis is performed independently on each cell line, and the sum of the results is presented. A gene detected in multiple cell lines is counted once per cell line. **H.** Example of the ASS gene *APIP* in SHP77, consistently identified across all long-read protocols. An exon is specifically observed in one haplotype.

monly detected in SC, SN and bulk datasets (13,376 genes and 24,499 transcripts; Fig.4I, J). DE features were largely concordant within SC and SN datasets, regardless of sequencing platform. SC datasets identified a comparable number of DE genes as bulk (Fig.3H, 4I and Supplementary Fig.11G) and recovered the majority of high-confidence DE genes, whereas the SN datasets showed significantly lower sensitivity. Notably, both SC and SN datasets reported many DE genes and transcripts not detected in any bulk datasets (Fig.4I,J). Given the higher read depth and statistical power in bulk, a substantial portion of these additional SC/SN-specific DE features may represent false positives.

**Variants and allele-specific analyses.** RNA-seq also enables the detection of genetic variants and the investigation of allelic differences in gene expression. Using the *LongBench* bulk datasets, we evaluated variant calling and phasing across four sequencing protocols (Fig.5A). PacBio detected the largest number of variants, predominantly due to greater coverage of intronic regions, identifying more intronic variants (Fig.5B). ONT dRNA sequencing identified the highest proportion of CCLE-reported mutations, predominantly exonic, followed by PacBio (Fig. 5C).

Beyond variant detection, RNA-seq enables investigation of allelic imbalance analysis, which measures deviations from the expected 50/50 expression ratio at heterozygous loci. This can arise from *cis*-acting regulatory variation, imprinting, or post-transcriptional effects[58,59]. Examining alternative allele frequencies, all long-read datasets showed medians near 0.5, indicating balanced expression and unbiased mapping, whereas Illumina displayed a slight reference-allele skew (Fig.5D), and more allelic imbalance events were detected in Illumina at matched depth (Supplementary Fig.12C–E), likely due to reduced mappability of alternative-allele reads. To mitigate the mapping bias, a common issue in short-read data, conventional methods include masking variant positions in the reference genome, making personalised references. However, such approaches may be less critical for

long-read sequencing, which spans larger genomic regions and provides a simpler workflow.

While conventional allele-specific analysis focuses on individual heterozygous sites, haplotype-aware analysis links full-length transcripts to specific haplotypes, revealing coordinated regulation, allele-specific splicing, and cis-regulatory effects missed by site-level approaches[60]. Short-read RNA-seq is limited in haplotype reconstruction, phasing only a small proportion of heterozygous variants (Fig.5A). In contrast, long reads span multiple variants per transcript, enabling phasing of the majority of sites and increasing confidence in variant calls, as spurious errors are less likely to be phased with real variants (Supplementary Fig.12A,B).

Phasing variants with long reads enables the detection of allele-specific isoform usage and improves robustness by aggregating signals across multiple phased variants. Using the long-read datasets, we performed haplotype-resolved allele-specific expression (ASE) and allele-specific splicing (ASS) analyses to identify haplotype-dependent differences in gene expression and isoform structure (Fig.5E–H). PacBio detected the highest number of ASE and ASS events, followed by the ONT direct RNA dataset, with many events consistently identified across all three long-read platforms (Fig.5E, G). For example, the ASE gene *ZCCHC17* in SHP77 was detected across all protocols, showing a clear difference in expression levels between haplotypes (Fig.5F), and the *APIP* gene showed haplotype-specific exon usage (Fig.5H). An additional example is provided in Supplementary Fig.12F,G.

## Discussion

Despite the promise of long-read RNA-seq to deliver quantitative, full-length transcript sequencing, *LongBench* demonstrates that all protocols show specific length and quantitation biases. PacBio Kinnex's improved throughput enables differential analysis of bulk, SC and SN transcriptomes with highly accurate reads, though cost remains higher than other platforms and transcript lengths are constrained at both ends of the spectrum, particularly the under representation of transcripts below 1 kb with the standard bulk protocol[42]. ONT cDNA sequencing is less costly but yields shorter and less accurate reads, with reduced sensitivity for gene, transcript, splice junction detection and differential expression analysis. However, these differences may mainly arise from differences in manufacturer-recommended library preparation protocols; in single-cell and single-nucleus libraries, where cDNA is prepared uniformly, the choice of sequencer has only subtle effects in these analyses. ONT direct RNA sequencing had the lowest bias of bulk transcriptomic methods, while the accuracy and throughput improvements of the RNA004 chemistry enable confident isoform assignment and SNP analyses to complement its unique ability to detect RNA modifications. However, direct RNA requires relatively large RNA inputs, is slightly more costly than ONT cDNA and has lower accuracy. It also lacks compatibility with single-cell applications and provides limited barcoding options.

Hybrid approaches may allow sequencing platforms to complement each other for more accurate transcript quantification. While methods exist for isoform identification[33,35,61,62], few studies have explored combining long and short reads for transcript quantification[63–65]. As sequencing costs decline, integrating multiple long-read platforms could also become a practical way to overcome individual limitations. However, no established framework yet exists for joint analysis across long-read platforms. Resources like *LongBench* provide valuable benchmarks, but further method development is needed to fully exploit the complementary strengths of these technologies.

We showed limited concordance for transcript-level differential analyses and the increased complexity and potential technical variability involved in transcript quantification. These findings emphasise the need for caution when interpreting transcript-level changes. Although several previous studies have noted high variability in transcript quantification[39–41], *LongBench* uniquely incorporates realistic biological variability, enabling realistic differential analyses. In addition, unlike benchmarks limited to either bulk or SC RNA-seq, *LongBench* extends these analyses to matched bulk, SC, and SN datasets, increasing relevance across diverse experimental contexts.

Variant analysis highlighted another major advantage of long reads: the ability to detect and phase variants directly, enabling allele-specific expression and splicing analyses at high confidence. This provides critical insights into regulatory mechanisms such as imprinting, *cis*-regulatory effects, and isoform-level consequences of genetic variation. Furthermore, it lays the foundation for long-read-based expression QTL and splicing QTL analysis[66,67], offering improved resolution in linking genetic variants to transcript-level regulatory effects. At present, most commonly used variant callers are not yet optimised for single-cell RNA-seq. Therefore, in this manuscript, we focused on bulk RNA-seq. However, *LongBench* provides a valuable resource for future method development and validation in that direction.

Our deeply sequenced, multi-platform dataset provides a valuable resource for lung cancer cell lines, particularly at the transcript level, which is often underrepresented in large-scale resources such as CCLE. By integrating complementary technologies, our design helps mitigate platform-specific biases and reduce false discoveries. In this study, although we focused on benchmarking sequencing technologies using pseudo-bulk analyses, the single-cell long-read data could enable high-resolution studies of expression heterogeneity. Technical noise and sparse coverage at the single-cell level may further limit the reliability of differential expression analyses[23,44], making the matched bulk datasets a critical reference to control for such artefacts.

We anticipate that *LongBench* will serve as a foundational dataset for the development, validation, and refinement of RNA-seq analysis tools, including those that integrate multi-platform data, capture transcript diversity more accurately, or advance single-cell and allele-specific transcriptomics.

## Methods

**Cell Culture.** Our study included adherent cell lines (H1975, HCC827, H2228) and suspension cell lines (H211, H526, H69, H146, SHP77). Cells were maintained in RPMI 1640 medium supplemented with GlutaMAX™ (Thermo Fisher Scientific), 10% heat-inactivated fetal bovine serum (FBS; Gibco), and 1% penicillin-streptomycin (Thermo Fisher Scientific). Cultures were incubated at 37°C in a humidified atmosphere containing 5% $CO_2$. Cells were routinely passaged at 70–80% confluency using standard culture procedures.

**RNA extraction for bulk RNA sequencing.** For consistency with the single-cell experiment, adherent cell lines were first detached with TrypLE Express (Gibco). For all cell lines, 1.5–2 M cells were collected for TRIzol-mediated RNA extraction. RNA quantity and quality was assessed with a Nanodrop spectrophotometer, Qubit and TapeStation. The yield was 20–80 µg of RNA per cell line with an RNA Integrity Index (RIN) of 10. Total RNA was supplemented with spike-in controls from Sequins mixtures A and B[46], as well as SIRV Set 4 (Lexogen), which includes Iso Mix E0, ERCC, and Long SIRVs. Sequins mix A was added to RNA from SCLC-P and LUAD cell lines, while mix B was used for SCLC-A. Spike-in preparations followed the manufacturers' protocols. An estimated mRNA content of 2% in the cell lines was used to calculate the spike-in input, targeting approximately 1% of sequenced reads for Sequins and SIRVs. For Oxford Nanopore direct-RNA libraries sequenced on secondary flow cells, the input amount of Sequins prior to library preparation was increased tenfold due to observed throughput issues specific to Sequins on the ONT platform.

**Single-cell and single-nucleus preparation.** Suspension cell lines (H211, H526, H146, H69, SHP-77) were prepared for fluorescence-activated cell sorting (FACS). As these cell lines grow in sheets and aggregates, they were difficult to manually separate into singlets and required sorting for input to 10X Capture. Each cell line was collected from culture, washed twice with PBS, and re-suspended in 500 µL FACS buffer (PBS with 2% FBS). Cells were then passed through a 35 µm strainer cap into FACS tubes and were subsequently stained with 7-amino-actinomycin D for live/dead sorting. Cells were sorted on the BD Aria III to collect 270,000 cells per sample. The remaining 3 adherent lines (H1975, H2228, HCC827) were detached using TrypLE Express (Gibco), washed twice with PBS, and resuspended in FACS buffer. Cell counts were performed in triplicate per cell line on the Countess 3 (Thermo Fisher Scientific) to ensure accuracy, and then 270,000 cells per line were combined into a single mixture. The cell mixture was loaded onto the 10X Genomics Chromium according to the manufacturer's protocol using the 10X Chromium NEXT GEM single-cell 3' kit (v3.1) with a target capture of 5,000 cells.

Nuclei isolation was performed using a custom protocol based on the Nuclei EZ Prep kit (Sigma-Aldrich) with the remaining 1.7 M cell mixture. Cells were pelleted at 300 x g for 5 minutes at 4°C, then gently resuspended in 300 µL of cold cell lysis buffer (EZ lysis buffer supplemented with 0.2 U/µL RNaseOUT™). Lysis was carried out on ice for 2.5 minutes to release nuclei, followed by centrifugation at 300 x g for 5 minutes at 4°C. The nuclei pellet was resuspended in 500 µL of cold wash buffer(1X PBS, 1% BSA, 0.2 U/µL RNaseOUT™), incubated on ice for 1 minute, then supplemented with a second 500 µL of wash buffer. Nuclei were gently resuspended by pipetting and pelleted at 200 x g for 5 minutes at 4°C. The pellet was resuspended in 200 µL of wash buffer, topped up to 1 mL, and centrifuged at 130 x g for 5 minutes. Nuclei were finally resuspended in 500 µL of wash buffer for quality assessment via DAPI staining and microscopy using Countess 3 system. The suspension was filtered twice using FlowMi 40 µM filter tips (Sigma-Aldrich) to remove any present clumps and re-counted prior to 10X capture. The nuclei mixture was loaded onto the 10X Genomics Chromium following the same protocol as described for single-cell capture, with a target capture of 5,000 nuclei.

**Library preparation and RNA sequencing.**

***Bulk RNA sequencing libraries.*** Illumina libraries we constructed using the Illumina Truseq v2 Stranded kit with polyA selection. Superscript III was used instead of SuperScript II, and the reverse transcription incubation step was accordingly adjusted to 50°C. The pool of eight indexed libraries was sequenced on one NextSeq2000 Mid-Output flow cell (400M 100PE reads). Nanopore direct-RNA sequencing libraries were constructed using the ONT SQK-RNA004 kit, with the following modifications: input was increased to 2 µg total RNA, we omitted the RNA CS (calibrating strand), and used the NEB Induro Reverse Transcriptase rather than the SuperScript III. Four libraries were prepared for each cell line, and sequenced over two PromethION RNA flow cells (FLO-PRO004RA), including one wash (custom wash mix based on the ONT SQK-WSH004 kit of 2 µL WMX, 1 µL RNase Cocktail, 1 µL RNase H, 396 µL DIL) and reload per flow cell (16 flow cells total). Nanopore PCR-cDNA libraries were constructed using the ONT SQK-PCB114.24 kit, with the following modifications: reverse transcription was performed for 90 minutes at 42°C, followed by 14 cycles of PCR with an extension time of 4 minutes. Barcoded samples were pooled and sequenced on three Promethion R10.4.1 flow cells (FLO-PRO114M, 3 flow cells total). PacBio Bulk Kinnex libraries were prepared with the Iso-Seq express 2.0 kit and Kinnex PCR 8-fold kit at the Australian Genomics Research Facility (AGRF). Individual libraries were sequenced on their own PacBio Revio flow cell (non-SPRQ, 8 flow cells total).

***Single-cell RNA sequencing libraries.*** From the cDNA generated by the 10X Genomics single-cell 3' kit (v3.1, 11 cycles of PCR), libraries were prepared for Illumina, Nanopore and PacBio sequencing. Illumina scRNA-seq libraries were prepared with the 10X Library Construction Kit (PN-1000190) and sequenced on one lane of a NextSeq2000 High-Output flow cell (1.8B 100PE reads). The nanopore library was prepared from 10 ng of input cDNA, following the ONT single-cell protocol including the artifact removal

step (v9198_v114_revE_06Dec2023). 125 ng of the post-pulldown PCR product were taken to adapter ligation with the ONT SQK-LSK114 kit. The final wash was performed with a 0.5X bead ratio (adjusted from 0.4X) and the short fragment buffer. 50 ng of the final library were loaded onto one Promethion R10.4.1 flow cell (FLO-PRO114M, 1 flow cell total). The PacBio single-cell Kinnex library was prepared with the Kinnex single-cell 12-plex concatenation kit according to the manufacturer's instructions, which also include an artifact removal step. The final library was loaded onto two PacBio Revio flow cells (2 flow cells total).

***Single-nuclei RNA sequencing libraries.*** From the cDNA generated by the 10X Genomics single-nuclei (12 cycles of PCR), libraries were prepared for Illumina, Nanopore and PacBio sequencing. Illumina snRNA-seq libraries were prepared following the same protocol as the single-cell library and sequenced on one lane of a High-Output flowcell (1.8B 100PE reads). A nanopore single-nuclei library was prepared following the same protocol as the single-cell library, and sequenced on one PromethION R10.4.1 flow cell (FLO-PRO114M, 1 flow cell total). A PacBio single-nuclei Kinnex library was prepared identically to the single-cell protocol. The final library was loaded onto two PacBio Revio flow cells (non-SPRQ, 2 flow cells total).

**Data pre-processing and quality control.** The ONT datasets were basecalled with Dorado using *rna004_130bps_sup@v5.1.0* for direct RNA and *dna_r10.4.1_e8.2_400bps_sup@v5.0.0* for the cDNA datasets; reads with mean quality scores $<10$ were discarded. The read length and quality distribution were summarised using Nanoplot. The reads were mapped to the reference genome GRCh38 and GENCODE v44 transcriptome using minimap2[68] (v2.28). The resulting BAM files of genome alignment were subsampled to 3 million reads per sample to estimate the alignment rate, alignment-based error and the distribution of the genomic feature using AlignQC in Fig.2A.

**Gene and transcript quantification.** Transcript-level quantification was performed using Salmon[69] (v1.10.2) for short-read datasets and oarfish[34] (v0.6.2) for long-read datasets, using the GENCODE v44 transcriptome supplemented with spike-in transcripts for bulk datasets (not included in pseudo-bulk quantification). Before quantification, read-to-transcript alignment is essential; Salmon incorporates an in-built alignment process, while minimap2[68] (v2.28) was used for long-read alignment with oarfish as mentioned above. Both salmon and oarfish can compute bootstrapped abundance estimates. Gene-level quantification was obtained by aggregating transcript-level quantification using tximport[70] (v1.34.0).

**Read-to-transcript ambiguity.** During quantification, 50 bootstrap samples were generated per run to assess the technical variance introduced during the quantification step. Following the approach of Baldoni et al.[50], we used transcript-level bootstrap counts to estimate the extra-Poisson overdispersion parameters, which capture the uncertainty arising from the probabilistic assignment of reads to transcripts. In line with Baldoni et al.[50], we refer to this overdispersion as read-to-transcript ambiguity throughout the manuscript. The method is described in detail in their work and is implemented in the edgeR package.

**Gene and transcript detection.** Gene and transcript detection analyses were based on whether a feature had sufficient counts from the quantification step. Specifically, we used the *filterByExpr* function[71] from edgeR[72], which retained features with counts-per-million (CPM) above a threshold determined from a minimum count and the median library size, and required that this CPM threshold be met in at least two of the eight cell lines in our case. We set the minimum count to 10 for genes and 5 for transcripts. The *filterByExpr* function is typically used to define features suitable for differential expression analysis. Although *filterByExpr* is typically used to define features suitable for differential expression analysis, we applied it here to define "detected features," ensuring that the criteria for feature presence in the detection analysis are directly comparable to those used in DE analysis.

**Analysis of spike-in transcripts.** The spike-in transcripts were quantified together with human transcripts in bulk datasets. SIRV Set 4 includes ERCC transcripts, which are present at half the molar concentration of the SIRV Set 4. However, we observed that the different types of spike-ins are captured with varying efficiency, due to differences in their design and synthesis. Especially, both ONT PCR cDNA and dRNA protocols did not efficiently capture Sequins and ERCC transcripts (Supplementary Fig.6). To avoid introducing this artificial bias in the comparison, we analysed ERCC transcripts separately from the other SIRV transcripts. Unless otherwise specified, "SIRV" refers only to the standard SIRV mix (i.e., SIRV Mix 0) and the long SIRV transcripts, which were added at equal molar concentrations.

**Filtering and normalisation for quantification comparison.** Because the differences in feature detection could obscure meaningful quantitative differences, we restricted the quantification comparisons to the genes and transcripts detected by all platforms. The estimated counts were then normalised to "Transcript per million (TPM)" and the TPM for short-read data were adjusted by the effective length of each transcript. Specifically:

$$\mathrm{TPM}_i = \begin{cases} \dfrac{\mathrm{Counts}_i}{\sum_{j\in T}\mathrm{Counts}_j} \times 10^6 & \text{(Long-read data)} \\[2ex] \dfrac{\dfrac{\mathrm{Counts}_i}{\mathrm{Length}_i}}{\sum_{j\in T}\dfrac{\mathrm{Counts}_j}{\mathrm{Length}_j}} \times 10^6 & \text{(Short-read data)} \end{cases}$$

where $T$ denotes the set of all transcripts, $\mathrm{Counts}_i$ is the count for transcript $i$, and $\mathrm{Length}_j$ is the effective length of transcript $j$. Note that this TPM normalisation is used

only for quantification comparison across sequencing protocols and was performed separately for human transcripts, Sequins, ERCC, and SIRVs. For differential expression analysis across cancer types, raw gene or transcript-level counts were normalised using the strategy below.

**Differential Expression Analysis.** Similar to the quantification comparison, we tested the genes and transcripts detected across all platforms for differential expression (DE). The gene and transcript counts were normalised using trimmed mean of M values (TMM)[73]. The transcript counts were further scaled by the overdispersion following the workflow suggested in Baldoni et al.[50]. We used edgeR[72] quasi-likelihood (QL) pipeline to perform gene- and transcript-level DE testing between the 3 groups LUAD, SCLC-A and SCLC-P.

**Different Transcript Usage Analysis.** Differential transcript usage (DTU) analysis was performed using DTUrtle[74] (v1.0.3), which implements the DRIMSeq[75] framework. Transcript-level counts were generated using tximport (v1.34.0), and our analysis was limited to the set of transcripts deemed "detected features" as described in the gene and transcript detection section. We then applied DTUrtle's *sparse_filter* function (all `min_*` parameters = 0), to remove genes with only one transcript. The filtered transcript lists were then intersected across Illumina, ONT cDNA, ONT dRNA, and PacBio so that DTU testing used the same feature set for every protocol across all pairwise Sequins (Sequin A vs Sequin B ) and human (LUAD vs SCLC-A, LUAD vs SCLC-P, SCLC-A vs SCLC-P) comparisons. Significant DTU events at a gene and transcript level were identified by applying DTUrtle's *posthoc_and_stager* function (`ofdr=0.05`, `posthoc=0.1`) to the DRIMSeq fitted models for each pairwise comparison.

**Single-cell and single-nucleus RNA-seq data processing.** The short-read single-cell (SC) and single-nucleus (SN) datasets were processed using Cell Ranger (v8.0.1) with GENCODE v44 comprehensive annotations and the GRCh38 primary assembly reference genome. Long-read SC and SN reads were first demultiplexed using BLAZE (v2.5), and the resulting FASTQ files were processed with the FLAMES (v2.0.1) single-cell pipeline to obtain gene-level quantification. Low-quality cells in each SC and SN dataset were filtered out by applying a median absolute deviation (MAD) threshold of 3 on the log-transformed number of detected features (nFeature) and total counts (nCount) to remove extreme outliers. Additionally, cells with mitochondrial gene percentages exceeding 3 MADs above the median were excluded. Cells failing any of these criteria were removed to ensure data quality. After filtering, cells and nuclei commonly detected across Illumina, PacBio, and ONT datasets were retained. Those successfully assigned to a cell line were pooled by cell line within each SC or SN dataset to create pseudobulk FASTQ files. These pseudobulk files were processed identically to bulk datasets to obtain gene- and transcript-level quantification and to perform differential expression analysis.

**Cell line annotation and detection of doublets and ambient RNA in single-cell and single-nucleus data.** Cells and nuclei were annotated based on genotypes at common human SNPs from the 1000 Genomes Project[76]. To establish genotypes for the eight cell lines, we first ran CellSNP-lite[77] on the Illumina bulk RNA-seq data. For each single-cell and single-nucleus dataset, we then ran CellSNP-lite separately to genotype each barcode. Both the bulk and single-cell VCF files were provided to Vireo[78], which assigned each cell or nucleus to the most likely cell line based on its SNP profile. Cell line annotations were obtained independently from the Illumina, ONT, and PacBio datasets. We then merged these annotations by assigning each cell or nucleus to the cell line reported by the majority of platforms. Cells with an equal number of conflicting assignments were discarded from the final merged annotation. Unannotated calls were ignored in the comparison and did not count as disagreement. In each Vireo run, doublets were reported by default and ambient RNA estimation was enabled via the `--callAmbientRNAs` option.

**Variant calling and allele-specific analyses.** Variant calling was performed using tools from the Clair3 series[79,80] (v1.0.10), with technology-specific presets. Clair3-Illumina was used with the preset `ilmn` for short-read RNA-seq, and Clair3-RNA[80] was used for the bulk long-read datasets with the presets `hifi_mas_minimap2`, `ont_r10_dorado_cdna`, and `ont_dorado_drna004`. Variants called with QUAL < 10 were considered low quality and discarded. These calls were compared with somatic mutation data for each cell line obtained from the Cancer Cell Line Encyclopedia (CCLE) using BCFtools[81] (v1.21). Next, phasing of heterozygous variants was performed using WhatsHap[82] (v2.8.0). Finally, allele-specific expression (ASE) and allele-specific splicing (ASS) events were identified with the LongcallR pipeline[60], which is tailored for SNP detection, phasing and allele-specific analyses based on long-read RNA-seq data.

**Benchmarking Analysis.** In this study, we focused on comparing different sequencing protocols. To ensure that observed differences arise from the data rather than from analysis variability, we standardised the computational workflow as much as possible. For transcript quantification, we used Salmon and Oarfish—tools developed by the same group with similar models. For mutation analysis, we applied Clair3-Illumina for short reads and Clair3-RNA for long reads. Single-cell and single-nucleus datasets were aggregated into pseudobulk profiles and processed through the same pipeline as bulk data (Fig.1D).

**Data Availability.** The *LongBench* data are available at https://github.com/mritchielab/LongBench.io. Sequencing data are available from Gene Expression Omnibus under accession number GSE303762.

**Code Availability.** The code for the analysis can be found at https://github.com/youyupei/LongBench_workflow.

**Author Contributions.** Y.Y., A.S., J.L., M.D. and C.W. conducted data analysis, generated the figures and wrote the manuscript with input from all authors. K.Z., S.S., R.G., M.C., J.G. and Y.D.J.P. conducted data analysis. J.L. and Q.G. generated benchmarking data. B.D. and I.C. supervised research. M.L., J.N. and K.D.S. designed the study design, provided cell lines and culture protocols and interpreted results. M.B.C., Q.G. and M.E.R. obtained research funding, designed the study and supervised the research. All authors read and approved the final manuscript.

**Funding.** This work was supported by the Australian National Health and Medical Research Council (NHMRC) Investigator Grants (GNT1196841 to M.B.C, GNT2007996 to Q.G. and GNT2017257 to M.E.R), Medical Research Future Fund Researcher Exchange and Development in Industry Fellowship (REDIF249 to M.E.R), PhD Fellowship of the Research Foundation-Flanders (FWO-Vlaanderen, 11F9623N/11F9625N to M.D.), long-stay abroad funding (FWO-Vlaanderen, V439923N to M.D.), Dutch MS Research Foundation (Monique Blom - de Wagt grant, 23-1187 MS to M.D.), research mandate of KU Leuven (BOF-FKO, Bijzonder Onderzoeksfonds – Fundamenteel Klinisch Onderzoeker to B.D.), Victorian State Government Operational Infrastructure Support, Australian Government NHMRC IRIISS and support from the Australian Cancer Research Foundation.

**Declaration of interests.** The authors acknowledge the support of both PacBio and Oxford Nanopore Technologies who provided sequencing reagents used in the single-cell / single-nuclei arm of this study. Y.Y, K.Z, M.B.C. and Q.G. received travel support from Oxford Nanopore Technologies to attend conferences. Q.G. received travel support from PacBio to attend a conference. The authors have no other competing interests and the Funders had no involvement in study design, data analysis, interpretation and writing of the article, or the decision to submit the work for publication.

**ACKNOWLEDGEMENTS**
We are grateful to the WEHI Genomics Lab (Casey Anttila, Johannes Wichmann and Daniela Zalcenstein) for processing the single-cell mixtures using the 10x platform, the Australian Genome Research Facility (Cath Moore, Peter Lau, David Hawkes, Angus Li and Trent Peters) for performing PacBio long-read sequencing, Ricardo De Paoli-Iseppi for sharing direct RNA sequencing advice, Paul Gooding (Millenium Science), James Miller and Elizabeth Tseng (PacBio) for advice on sequencing protocols and data analysis and Qingwang Chen for discussions on data analysis. We thank AWS for providing access to the *LongBench* data through the AWS Open Data Sponsorship Program.

## References

1. Luyi Tian, Jafar S. Jabbari, Rachel Thijssen, Quentin Gouil, Shanika L. Amarasinghe, Oliver Voogd, Hasaru Kariyawasam, Mei R. M. Du, Jakob Schuster, Changqing Wang, Shian Su, Xueyi Dong, Charity W. Law, Alexis Lucattini, Yair David Joseph Prawer, Coralina Collar-Fernández, Jin D. Chung, Timur Naim, Audrey Chan, Chi Hai Ly, Gordon S. Lynch, James G. Ryall, Casey J. A. Anttila, Hongke Peng, Mary Ann Anderson, Christoffer Flensburg, Ian Majewski, Andrew W. Roberts, David C. S. Huang, Michael B. Clark, and Matthew E. Ritchie. Comprehensive characterization of single-cell full-length isoforms in human and mouse with long-read sequencing. *Genome Biology*, 22(1):310, November 2021. ISSN 1474-760X. doi: 10.1186/s13059-021-02525-6.
2. Alison D. Tang, Cameron M. Soulette, Marijke J. van Baren, Kevyn Hart, Eva Hrabeta-Robinson, Catherine J. Wu, and Angela N. Brooks. Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. *Nature Communications*, 11(1):1438, March 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-15171-6.
3. Ying Jing, Yi Zhang, Hui Zhu, Ke Zhang, Mei-Chun Cai, Pengfei Ma, Peiye Shen, Zhenfeng Zhang, Minghui Shao, Jing Wang, Minhua Yu, Xia Yin, Meiying Zhang, Yuan Hu, Danni Chen, Wen Di, Xiaojie Wang, and Guanglei Zhuang. Hybrid sequencing-based personal full-length transcriptomic analysis implicates proteostatic stress in metastatic ovarian cancer. *Oncogene*, 38(16):3047–3060, April 2019. ISSN 1476-5594. doi: 10.1038/s41388-018-0644-y.
4. Kie Kyon Huang, Jiawen Huang, Jeanie Kar Leng Wu, Minghui Lee, Su Ting Tay, Vikrant Kumar, Kalpana Ramnarayanan, Nisha Padmanabhan, Chang Xu, Angie Lay Keng Tan, Charlene Chan, Dennis Kappei, Jonathan Göke, and Patrick Tan. Long-read transcriptome sequencing reveals abundant promoter diversity in distinct molecular subtypes of gastric cancer. *Genome Biology*, 22(1):44, January 2021. ISSN 1474-760X. doi: 10.1186/s13059-021-02261-x.
5. Miho Oka, Liu Xu, Toshihiro Suzuki, Toshiaki Yoshikawa, Hiromi Sakamoto, Hayato Uemura, Akiyasu C. Yoshizawa, Yutaka Suzuki, Tetsuya Nakatsura, Yasushi Ishihama, Ayako Suzuki, and Masahide Seki. Aberrant splicing isoforms detected by full-length transcriptome sequencing as transcripts of potential neoantigens in non-small cell lung cancer. *Genome Biology*, 22(1):9, January 2021. ISSN 1474-760X. doi: 10.1186/s13059-020-02240-8.
6. Michael B. Clark, Tomasz Wrzesinski, Aintzane B. Garcia, Nicola A. L. Hall, Joel E. Kleinman, Thomas Hyde, Daniel R. Weinberger, Paul J. Harrison, Wilfried Haerty, and Elizabeth M. Tunbridge. Long-read sequencing reveals the complex splicing profile of the psychiatric risk gene CACNA1C in human brain. *Molecular Psychiatry*, 25(1):37–47, January 2020. ISSN 1476-5578. doi: 10.1038/s41380-019-0583-1.
7. Bernardo Aguzzoli Heberle, J. Anthony Brandon, Madeline L. Page, Kayla A. Nations, Ketsile I. Dikobe, Brendan J. White, Lacey A. Gordon, Grant A. Fox, Mark E. Wadsworth, Patricia H. Doyle, Brittney A. Williams, Edward J. Fox, Anantharaman Shantaraman, Mina Ryten, Sara Goodwin, Elena Ghiban, Robert Wappel, Senem Mavruk-Eskipehlivan, Justin B. Miller, Nicholas T. Seyfried, Peter T. Nelson, John D. Fryer, and Mark T. W. Ebbert. Mapping medically relevant RNA isoform diversity in the aged human frontal cortex with deep long-read RNA-seq. *Nature Biotechnology*, 43(4):635–646, April 2025. ISSN 1546-1696. doi: 10.1038/s41587-024-02245-9.
8. Liang Ma, Stephen A. Semick, Qiang Chen, Chao Li, Ran Tao, Amanda J. Price, Joo Heon Shin, Yankai Jia, Nicholas J. Brandon, Alan J. Cross, Thomas M. Hyde, Joel E. Kleinman, Andrew E. Jaffe, Daniel R. Weinberger, Richard E. Straub, and The BrainSeq Consortium. Schizophrenia risk variants influence multiple classes of transcripts of sorting nexin 19 (SNX19). *Molecular Psychiatry*, 25(4):831–843, April 2020. ISSN 1476-5578. doi: 10.1038/s41380-018-0293-0.
9. Ricardo De Paoli-Iseppi, Shweta Joshi, Josie Gleeson, Yair David Joseph Prawer, Yupei You, Ria Agarwal, Anran Li, Anthea Hull, Eloise Marie Whitehead, Yoonji Seo, Rhea Kujawa, Raphael Chang, Mriga Dutt, Catriona McLean, Benjamin Leo Parker, and Michael Ben Clark. Long-read sequencing reveals the RNA isoform repertoire of neuropsychiatric risk genes in human brain. *medRxiv*, 2024. doi: 10.1101/2024.02.22.24303189.
10. John Eid, Adrian Fehr, Jeremy Gray, Khai Luong, John Lyle, Geoff Otto, Paul Peluso, David Rank, Primo Baybayan, Brad Bettman, et al. Real-time DNA sequencing from single polymerase molecules. *Science*, 323(5910):133–138, 2009.
11. James Clarke, Hai-Chen Wu, Lakmal Jayasinghe, Alpesh Patel, Stuart Reid, and Hagan Bayley. Continuous base identification for single-molecule nanopore DNA sequencing. *Nature Nanotechnology*, 4(4):265–270, 2009.
12. Pacific Biosciences. Application note: Kinnex full-length RNA kit for isoform sequencing. Online: https://www.pacb.com/wp-content/uploads/Application-note-Kinnex-full-length-RNA-kit-for-isoform-sequencing.pdf, 2023.
13. Aziz M. Al'Khafaji, Jonathan T. Smith, Kiran V. Garimella, Mehrtash Babadi, Victoria Popic, Moshe Sade-Feldman, Michael Gatzen, Siranush Sarkizova, Marc A. Schwartz, Emily M. Blaum, Allyson Day, Maura Costello, Tera Bowers, Stacey Gabriel, Eric Banks, Anthony A. Philippakis, Genevieve M. Boland, Paul C. Blainey, and Nir Hacohen. High-throughput RNA isoform sequencing using programmed cDNA concatenation. *Nature Biotechnology*, 42(4): 582–586, April 2024. ISSN 1087-0156, 1546-1696. doi: 10.1038/s41587-023-01815-7.
14. Carolina Monzó, Tianyuan Liu, and Ana Conesa. Transcriptomics in the era of long-read sequencing. *Nature Reviews Genetics*, pages 1–21, 2025.
15. Daniel R Garalde, Elizabeth A Snell, Daniel Jachimowicz, Botond Sipos, Joseph H Lloyd, Mark Bruce, Nadia Pantic, Tigist Admassu, Phillip James, Anthony Warland, et al. Highly parallel direct RNA sequencing on an array of nanopores. *Nature Methods*, 15(3):201–206, 2018.
16. Daniel P. Depledge, Kalanghad Puthankalam Srinivas, Tomohiko Sadaoka, Devin Bready, Yasuko Mori, Dimitris G. Placantonakis, Ian Mohr, and Angus C. Wilson. Direct RNA sequencing on nanopore arrays redefines the transcriptional complexity of a viral pathogen. *Nature Communications*, 10(1):754, February 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-08734-9.
17. Rachael E Workman, Alison D Tang, Paul S Tang, Miten Jain, John R Tyson, Roham Razaghi, Philip C Zuzarte, Timothy Gilpatrick, Alexander Payne, Joshua Quick, et al. Nanopore native RNA sequencing of a human poly (A) transcriptome. *Nature Methods*, 16(12):1297–1305, 2019.
18. Mattia Furlan, Anna Delgado-Tejedor, Logan Mulroney, Mattia Pelizzola, Eva Maria Novoa, and Tommaso Leonardi. Computational methods for RNA modification detection from nanopore direct RNA sequencing data. *RNA Biology*, 18(sup1):31–40, October 2021. ISSN 1555-8584. doi: 10.1080/15476286.2021.1978215.
19. Ishaan Gupta, Paul G Collier, Bettina Haase, Ahmed Mahfouz, Anoushka Joglekar, Taylor Floyd, Frank Koopmans, Ben Barres, August B Smit, Steven A Sloan, et al. Single-cell isoform RNA sequencing characterizes isoforms in thousands of cerebellar cells. *Nature Biotechnology*, 36(12):1197–1202, 2018.
20. Cheng-Kai Shiau, Lina Lu, Rachel Kieser, Kazutaka Fukumura, Timothy Pan, Hsiao-Yun Lin, Jie Yang, Eric L Tong, GaHyun Lee, Yuanqing Yan, et al. High throughput single cell

long-read sequencing analyses of same-cell genotypes and phenotypes in human tumors. *Nature Communications*, 14(1):4124, 2023.

21. Roger Volden, Theron Palmer, Ashley Byrne, Charles Cole, Robert J Schmitz, Richard E Green, and Christopher Vollmers. Improving nanopore read accuracy with the R2C2 method enables the sequencing of highly multiplexed full-length single-cell cDNA. *Proceedings of the National Academy of Sciences*, 115(39):9726–9731, 2018.
22. Kevin Lebrigand, Virginie Magnone, Pascal Barbry, and Rainer Waldmann. High throughput error corrected Nanopore single cell transcriptome sequencing. *Nature Communications*, 11(1):4025, August 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-17800-6.
23. Mandeep Singh, Ghamdan Al-Eryani, Shaun Carswell, James M Ferguson, James Blackburn, Kirston Barton, Daniel Roden, Fabio Luciani, Tri Giang Phan, Simon Junankar, et al. High-throughput targeted long-read single cell sequencing reveals the clonal and transcriptional landscape of lymphocytes. *Nature Communications*, 10(1):3120, 2019.
24. Hongke Peng, Jafar S Jabbari, Luyi Tian, Changqing Wang, Yupei You, Chong Chyn Chua, Natasha S Anstee, Noorul Amin, Andrew H Wei, Nadia M Davidson, et al. Single-cell Rapid Capture Hybridization sequencing reliably detects isoform usage and coding mutations in targeted genes. *Genome Research*, 35(4):942–955, 2025.
25. Yanping Long, Zhijian Liu, Jinbu Jia, Weipeng Mo, Liang Fang, Dongdong Lu, Bo Liu, Hong Zhang, Wei Chen, and Jixian Zhai. FlsnRNA-seq: protoplasting-free full-length single-nucleus RNA profiling in plants. *Genome Biology*, 22(1):66, 2021.
26. Elisabeth Rebboah, Fairlie Reese, Katherine Williams, Gabriela Balderrama-Gutierrez, Cassandra McGill, Diane Trout, Isaryhia Rodriguez, Heidi Liang, Barbara J Wold, and Ali Mortazavi. Mapping and modeling the genomic basis of differential RNA isoform expression at single-cell resolution with LR-Split-seq. *Genome Biology*, 22(1):286, 2021.
27. Qi Wang, Sven Bönigk, Volker Böhm, Niels Gehring, Janine Altmüller, and Christoph Dieterich. Single-cell transcriptome sequencing on the Nanopore platform with ScNapBar. *RNA*, 27(7):763–770, 2021.
28. Yupei You, Yair DJ Prawer, Ricardo De Paoli-Iseppi, Cameron PJ Hunt, Clare L Parish, Heejung Shim, and Michael B Clark. Identification of cell barcodes from long-read single-cell RNA-seq with BLAZE. *Genome Biology*, 24(1):66, 2023.
29. Shuyao Zhang, Yuhua Xiao, Xinzhi Mo, Xu Chen, Jiawei Zhong, Zheyao Chen, Xu Liu, Yuanhui Qiu, Wangxuan Dai, Jia Chen, et al. Simultaneous profiling of RNA isoforms and chromatin accessibility of single cells of human retinal organoids. *Nature Communications*, 15(1):8022, 2024.
30. Anoushka Joglekar, Careen Foord, Julien Jarroux, Shaun Pollard, and Hagen U Tilgner. From words to complete phrases: insight into single-cell isoforms using short and long reads. *Transcription*, 14(3-5):92–104, 2023.
31. Pallavi Gupta, Hannah O'Neill, Ernst J. Wolvetang, Aniruddha Chatterjee, and Ishaan Gupta. Advances in single-cell long-read sequencing technologies. *NAR Genomics and Bioinformatics*, 6(2):lqae047, June 2024. ISSN 2631-9268. doi: 10.1093/nargab/lqae047.
32. Matthias Lienhard, Twan van den Beucken, Bernd Timmermann, Myriam Hochradel, Stefan Boerno, Florian Caiment, Martin Vingron, and Ralf Herwig. IsoTools: a flexible workflow for long-read transcriptome sequencing analysis. *Bioinformatics*, 39(6):btad364, 2023.
33. Ying Chen, Andre Sim, Yuk Kei Wan, Keith Yeo, Joseph Jing Xian Lee, Min Hao Ling, Michael I Love, and Jonathan Göke. Context-aware transcript quantification from long-read RNA-seq data with Bambu. *Nature Methods*, 20(8):1187–1195, 2023.
34. Zahra Zare Jousheghani, Noor Pratap Singh, and Rob Patro. Oarfish: Enhanced probabilistic modeling leads to improved accuracy in long read transcriptome quantification. *Bioinformatics*, 41(Supplement_1):i304–i313, 2025.
35. Andrey D Prjibelski, Alla Mikheenko, Anoushka Joglekar, Alexander Smetanin, Julien Jarroux, Alla L Lapidus, and Hagen U Tilgner. Accurate isoform discovery with IsoQuant using long reads. *Nature Biotechnology*, 41(7):915–918, 2023.
36. Oliver Cheng, Min Hao Ling, Changqing Wang, Shuyi Wu, Matthew E Ritchie, Jonathan Göke, Noorul Amin, and Nadia M Davidson. Flexiplex: a versatile demultiplexer and search tool for omics data. *Bioinformatics*, 40(3):btae102, 2024.
37. Jason L Weirather, Mariateresa de Cesare, Yunhao Wang, Paolo Piazza, Vittorio Sebastiano, Xiu-Jie Wang, David Buck, and Kin Fai Au. Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. *F1000Research*, 6:100, 2017.
38. Charlotte Soneson, Yao Yao, Anna Bratus-Neuenschwander, Andrea Patrignani, Mark D. Robinson, and Shobbir Hussain. A comprehensive examination of Nanopore native RNA sequencing for characterization of complex transcriptomes. *Nature Communications*, 10(1): 3359, July 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-11272-z.
39. Xueyi Dong, Mei RM Du, Quentin Gouil, Luyi Tian, Jafar S Jabbari, Rory Bowden, Pedro L Baldoni, Yunshun Chen, Gordon K Smyth, Shanika L Amarasinghe, et al. Benchmarking long-read RNA-sequencing analysis tools using in silico mixtures. *Nature Methods*, 20(11): 1810–1821, 2023.
40. Francisco J Pardo-Palacios, Dingjie Wang, Fairlie Reese, Mark Diekhans, Sílvia Carbonell-Sala, Brian Williams, Jane E Loveland, Maite De María, Matthew S Adams, Gabriela Balderrama-Gutierrez, et al. Systematic assessment of long-read RNA-seq methods for transcript identification and quantification. *Nature Methods*, 21(7):1349–1363, 2024.
41. Ying Chen, Nadia M Davidson, Yuk Kei Wan, Fei Yao, Yan Su, Hasindu Gamaarachchi, Andre Sim, Harshil Patel, Hwee Meng Low, Christopher Hendra, et al. A systematic benchmark of nanopore long-read RNA sequencing for transcript-level analysis in human cell lines. *Nature Methods*, pages 1–12, 2025.
42. David Wissel, Madison M. Mehlferber, Khue M. Nguyen, Vasilii Pavelko, Elizabeth Tseng, Mark D. Robinson, and Gloria M. Sheynkman. A systematic benchmark of high-accuracy PacBio long-read RNA sequencing for transcript-level quantification. *bioRxiv*, 2025. doi: 10.1101/2025.05.30.656561.
43. Natalia Zajac, Qin Zhang, Anna Bratus-Neuenschwander, Weihong Qi, Hella Anna Bolck, Tülay Karakulak, Tamara Carrasco Oltra, Holger Moch, Abdullah Kahraman, and Hubert Rehrauer. Comparison of single-cell long-read and short-read transcriptome sequencing via cDNA molecule matching: quality evaluation of the MAS-ISO-seq approach. *NAR Genomics and Bioinformatics*, 7(3):lqaf089, 2025.
44. Anita L Ahlert Scoones, Yuxuan Lan, Charlotte Utting, Lydia Pouncey, Ashleigh Lister, Sofia Kudasheva, Neelam Mehta, Naomi Irish, David Swarbreck, Karim Gharbi, et al. A comparison of long-read single-cell transcriptomic approaches. *bioRxiv*, page 2025.07.03.662955, 2025. doi: 10.1101/2025.07.03.662955.
45. Charles M Rudin, John T Poirier, Lauren Averett Byers, Caroline Dive, Afshin Dowlati, Julie George, John V Heymach, Jane E Johnson, Jonathan M Lehman, David MacPherson, et al. Molecular subtypes of small cell lung cancer: a synthesis of human and mouse model data. *Nature Reviews Cancer*, 19(5):289–297, 2019.
46. Simon A Hardwick, Wendy Y Chen, Ted Wong, Ira W Deveson, James Blackburn, Stacey B Andersen, Lars K Nielsen, John S Mattick, and Tim R Mercer. Spliced synthetic genes as internal controls in RNA sequencing experiments. *Nature Methods*, 13(9):792–798, 2016.
47. PacBio. Procedure checklist: Preparing Kinnex libraries using the Kinnex full-length RNA kit. https://www.pacb.com/wp-content/uploads/Procedure-checklist-Preparing-Kinnex-libraries-using-the-Kinnex-full-length-RNA-kit.pdf, 2023. Accessed: 2025-07-25.
48. Madalee G Wulf, Sean Maguire, Paul Humbert, Nan Dai, Yanxia Bei, Nicole M Nichols, Ivan R Corrêa, and Shengxi Guan. Non-templated addition and template switching by Moloney murine leukemia virus (MMLV)-based reverse transcriptases co-occur and compete with each other. *Journal of Biological Chemistry*, 294(48):18220–18231, 2019.
49. Jamie Auxillos, Arnaud Stigliani, Christian Skov Vaagensø, William Garland, Adnan Muhammed Niazi, Eivind Valen, Torben Heick Jensen, and Albin Sandelin. True length of diverse capped RNA sequencing (TLDR-seq): 5–3-end sequencing of capped RNAs regardless of 3-end status. *Nucleic Acids Research*, 53(6):gkaf240, 2025.
50. Pedro L Baldoni, Yunshun Chen, Soroor Hediyeh-Zadeh, Yang Liao, Xueyi Dong, Matthew E Ritchie, Wei Shi, and Gordon K Smyth. Dividing out quantification uncertainty allows efficient assessment of differential transcript expression with edgeR. *Nucleic Acids Research*, 52(3):e13–e13, 2024.
51. Francescopaolo Di Cello, Joelle Hillion, Alexandra Hristov, Lisa J Wood, Mita Mukherjee, Andrew Schuldenfrei, Jeanne Kowalski, Raka Bhattacharya, Raheela Ashfaq, and Linda MS Resar. HMGA2 participates in transformation in human lung cancer. *Molecular Cancer Research*, 6(5):743–750, 2008.
52. Aliaksei Z Holik, Charity W Law, Ruijie Liu, Zeya Wang, Wenyi Wang, Jaeil Ahn, Marie-Liesse Asselin-Labat, Gordon K Smyth, and Matthew E Ritchie. RNA-seq mixology: designing realistic control experiments to compare protocols and analysis methods. *Nucleic Acids Research*, 45(5):e30–e30, 2017.
53. Terence Li, Marcus Alvarez, Cuining Liu, Kevin Abuhanna, Yu Sun, Jason Ernst, Kathrin Plath, Brunilda Balliu, Chongyuan Luo, and Noah Zaitlen. The impact of ambient contamination on demultiplexing methods for single-nucleus multiome experiments, February 2025. Pages: 2025.02.06.636969 Section: New Results.
54. Emre Caglayan, Yuxiang Liu, and Genevieve Konopka. Neuronal ambient RNA contamination causes misinterpreted and masked cell types in brain single-nuclei datasets. *Neuron*, 110(24):4043–4056, 2022.
55. Aaron TL Lun, Samantha Riesenfeld, Tallulah Andrews, The Phuong Dao, Tomas Gomes, Participants in the 1st Human Cell Atlas Jamboree, and John C Marioni. Emptydrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. *Genome Biology*, 20(1):63, 2019.
56. Matthew D Young and Sam Behjati. Soupx removes ambient RNA contamination from droplet-based single-cell RNA sequencing data. *Gigascience*, 9(12):giaa151, 2020.
57. Leonardo Morelli, Valentina Giansanti, and Davide Cittaro. Nested stochastic block models applied to the analysis of single cell data. *BMC Bioinformatics*, 22:1–19, 2021.
58. Stephane E Castel, Ami Levy-Moonshine, Pejman Mohammadi, Eric Banks, and Tuuli Lappalainen. Tools and best practices for data processing in allelic expression analysis. *Genome Biology*, 16:1–12, 2015.
59. Daan van Beek, Job Verdonschot, Kasper Derks, Han Brunner, Theo M de Kok, Ilja CW Arts, Stephane Heymans, Martina Kutmon, and Michiel Adriaens. Allele-specific expression analysis for complex genetic phenotypes applied to a unique dilated cardiomyopathy cohort. *Scientific Reports*, 13(1):564, 2023.
60. Neng Huang, Human Pangenome Reference Consortium, and Heng Li. SNP calling, haplotype phasing and allele-specific analysis with long RNA-seq reads. *bioRxiv*, 2025. doi: 10.1101/2025.05.26.656191.
61. Alison D Tang, Cameron M Soulette, Marijke J van Baren, Kevyn Hart, Eva Hrabeta-Robinson, Catherine J Wu, and Angela N Brooks. Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. *Nature Communications*, 11(1):1438, 2020.
62. Sam Kovaka, Aleksey V Zimin, Geo M Pertea, Roham Razaghi, Steven L Salzberg, and Mihaela Pertea. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biology*, 20(1):278, 2019.
63. Haoran Li, Dingjie Wang, Qi Gao, Puwen Tan, Yunhao Wang, Xiaoyu Cai, Aifu Li, Yue Zhao, Andrew L. Thurman, Seyed Amir Malekpour, Ying Zhang, Roberta Sala, Andrea Cipriano, Chia-Lin Wei, Vittorio Sebastiano, Chi Song, Nancy R. Zhang, and Kin Fai Au. Improving gene isoform quantification with miniQuant. *Nature Biotechnology*, June 2025. ISSN 1546-1696. doi: 10.1038/s41587-025-02633-9.
64. Sojung Lee, Jayne A Barbour, Yee Man Tam, Haocheng Yang, Yuanhua Huang, and Jason WH Wong. LocusMasterTE: integrating long-read RNA sequencing improves locus-specific quantification of transposable element expression. *Genome Biology*, 26(1):72, 2025.
65. Michael Apostolides, Benedict Choi, Albertas Navickas, Ali Saberi, Larisa M Soto, Hani Goodarzi, and Hamed S Najafabadi. Accurate isoform quantification by joint short-and long-read RNA-sequencing. *bioRxiv*, 2024. doi: 10.1101/2024.07.11.603067.
66. Isabelle Heifetz Ament, Nicole DeBruyne, Feng Wang, and Lan Lin. Long-read RNA sequencing: A transformative technology for exploring transcriptome complexity in human diseases. *Molecular Therapy*, 33(3):883–894, March 2025. ISSN 1525-0016. doi: 10.1016/j.ymthe.2024.11.025.
67. Kensuke Yamaguchi, Kazuyoshi Ishigaki, Akari Suzuki, Yumi Tsuchida, Haruka Tsuchiya, Shuji Sumitomo, Yasuo Nagafuchi, Fuyuki Miya, Tatsuhiko Tsunoda, Hirofumi Shoda, Keishi Fujio, Kazuhiko Yamamoto, and Yuta Kochi. Splicing QTL analysis focusing on coding sequences reveals mechanisms for disease susceptibility loci. *Nature Communications*, 13: 4659, August 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-32358-1.
68. Heng Li. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18):

3094–3100, 2018.
69. Rob Patro, Geet Duggal, Michael I Love, Rafael A Irizarry, and Carl Kingsford. Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, 14(4): 417–419, mar 2017. doi: 10.1038/nmeth.4197.
70. Charlotte Soneson, Michael I. Love, and Mark D. Robinson. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Research*, 4, 2015. doi: 10.12688/f1000research.7563.1.
71. Yunshun Chen, Aaron T. L. Lun, and Gordon K. Smyth. From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline. *F1000Research*, 5:1438, 2016. ISSN 2046-1402. doi: 10.12688/f1000research.8987.2.
72. Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1): 139–140, 2010.
73. Mark D Robinson and Alicia Oshlack. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, 11:1–9, 2010.
74. Tobias Tekath and Martin Dugas. Differential transcript usage analysis of bulk and single-cell RNA-seq data with DTUrtle. *Bioinformatics*, 37(21):3781–3787, 2021.
75. Malgorzata Nowicka and Mark D Robinson. DRIMSeq: a Dirichlet-multinomial framework for multivariate count outcomes in genomics. *F1000Research*, 5:1356, 2016.
76. 1000 Genomes Project Consortium et al. A global reference for human genetic variation. *Nature*, 526(7571):68, 2015.
77. Xianjie Huang and Yuanhua Huang. Cellsnp-lite: an efficient tool for genotyping single cells. *Bioinformatics*, 37(23):4569–4571, 2021.
78. Yuanhua Huang, Davis J McCarthy, and Oliver Stegle. Vireo: Bayesian demultiplexing of pooled single-cell RNA-seq data without genotype reference. *Genome Biology*, 20(1):273, 2019.
79. Zhenxian Zheng, Shumin Li, Junhao Su, Amy Wing-Sze Leung, Tak-Wah Lam, and Ruibang Luo. Symphonizing pileup and full-alignment for deep learning-based long-read variant calling. *Nature Computational Science*, 2(12):797–803, 2022.
80. Zhenxian Zheng, Xian Yu, Lei Chen, Yan-Lam Lee, Cheng Xin, Angel On Ki Wong, Miten Jain, Rupesh K Kesharwani, Fritz J Sedlazeck, and Ruibang Luo. Clair3-RNA: a deep learning-based small variant caller for long-read RNA sequencing data. *bioRxiv*, pages 2024–11, 2025. doi: 10.1101/2024.11.17.624050.
81. Petr Danecek, James K Bonfield, Jennifer Liddle, John Marshall, Valeriu Ohan, Martin O Pollard, Andrew Whitwham, Thomas Keane, Shane A McCarthy, Robert M Davies, et al. Twelve years of SAMtools and BCFtools. *Gigascience*, 10(2):giab008, 2021.
82. Marcel Martin, Murray Patterson, Shilpa Garg, Sarah O Fischer, Nadia Pisanti, Gunnar W Klau, Alexander Schöenhuth, and Tobias Marschall. WhatsHap: fast and accurate read-based phasing. *bioRxiv*, page 085050, 2016. doi: 10.1101/085050.