scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Long-read RNA sequencing dataset of human pancreatic cancer cell lines

Shengnan Luo[1,3], Jiling Feng[1,3], Yu Zeng[1,3], Hao Wu[1], Yingxia Zheng[2] ✉ & Shengli Li[1] ✉

Long-read RNA sequencing (RNA-seq) technologies have revolutionized transcriptomic research by enabling the sequencing of full-length RNA molecules, thus providing a more accurate characterization of complex transcript isoforms than traditional short-read approaches. In this study, we present a high-coverage long-read transcriptome dataset generated using Oxford Nanopore Technologies' PromethION platform from ten human pancreatic cancer cell lines, with two biological replicates per line. The dataset comprises approximately 189.8 million reads across 20 samples, providing a valuable resource for studying transcript structures in pancreatic cancer. We perform systematic quality assessments, including read length, base quality, and gene body coverage, and report high reproducibility between replicates. Processed files, including transcript annotations in GTF, FASTA, and BED formats, are publicly available to facilitate reuse. This resource supports a wide range of downstream applications such as isoform discovery, transcriptome annotation, and integration with other omics data, offering a foundation for further exploration of transcriptomic complexity in cancer biology.

Background & Summary

RNA sequencing (RNA-seq) has become a cornerstone of transcriptomic research, enabling comprehensive analysis of gene expression, transcript isoforms, splicing events, and alternative poly(A) adenylation[1–4]. Traditionally, short-read RNA-seq technologies, such as those provided by Illumina, have dominated transcriptomic studies[1]. These methods rely on sequencing short fragments of cDNA, which are then aligned to a reference genome[2]. While highly accurate, short-read technologies face challenges in capturing full-length transcript information and resolving complex splicing events due to the limited read length[5]. In contrast, long-read RNA sequencing technologies, such as those developed by Oxford Nanopore Technologies (ONT)[6] and Pacific Biosciences (PacBio)[7], offer significant advantages in overcoming these limitations[8]. Long-read RNA-seq generates reads that can span entire transcripts, allowing for the direct detection of full-length isoforms and more accurate identification of splicing events[9]. This capability has made long-read RNA-seq an invaluable tool for transcriptome analysis, particularly in the study of complex or poorly annotated genomes, as well as in the discovery of novel isoforms and regulatory elements[10]. Compared to short-read RNA-seq, long-read technologies provide higher sensitivity for detecting rare or complex transcript variants, improve isoform quantification, and enhance the resolution of splicing events. Although the accuracy of long-read sequencing has historically lagged behind short-read approaches, recent advancements in error correction algorithms and sequencing chemistry have significantly improved read quality, making long-read RNA-seq increasingly feasible for large-scale studies[11,12].

In cancer research, long-read RNA-seq has emerged as a powerful technique for understanding the transcriptomic alterations that drive tumorigenesis[13–15]. It enables the detection of splicing events, alternative poly(A) adenylation, and open reading frame (ORF) that are often identified inefficiently or missed by short-read RNA-seq. Furthermore, it offers insights into transcriptome-wide changes that may have implications for drug resistance, tumor progression, and metastasis. In this study, we conducted nanopore long-read RNA-seq on ten human pancreatic cancer cell lines, each with two biological replicates (Table 1). This dataset enables

[1]Precision Research Center for Refractory Diseases, Shanghai Jiao Tong University Pioneer Research Institute for Molecular and Cell Therapies, Shanghai General Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, 201620, China. [2]Department of Laboratory Medicine, Xin Hua Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, 200025, China. [3]These authors contributed equally: Shengnan Luo, Jiling Feng, Yu Zeng. ✉e-mail: zhengyingxia@xinhuamed.com.cn; shengli.li@sjtu.edu.cn

Cell line	Replicates #	Description	Method	Data accession
PANC0203	2	Pancreatic adenocarcinoma (epithelial)	Nanopore long-read RNA-seq	GSM8887897, GSM8887898
PANC1005	2	Pancreatic adenocarcinoma (epithelial)	Nanopore long-read RNA-seq	GSM8887901, GSM8887902
Panc0327	2	Pancreatic adenocarcinoma (epithelial)	Nanopore long-read RNA-seq	GSM8887899, GSM8887900
AsPC-1	2	Pancreatic adenocarcinoma (epithelial)	Nanopore long-read RNA-seq	GSM8887885, GSM8887886
BXPC3	2	Pancreatic adenocarcinoma (epithelial)	Nanopore long-read RNA-seq	GSM8887887, GSM8887888
Capan-1	2	Pancreatic adenocarcinoma (epithelial)	Nanopore long-read RNA-seq	GSM8887889, GSM8887890
Capan-2	2	Pancreatic adenocarcinoma (epithelial)	Nanopore long-read RNA-seq	GSM8887891, GSM8887892
HuP-T4	2	Pancreatic adenocarcinoma (epithelial)	Nanopore long-read RNA-seq	GSM8887893, GSM8887894
Mia-PaCa-2	2	Pancreatic adenocarcinoma (epithelial)	Nanopore long-read RNA-seq	GSM8887895, GSM8887896
SW1990	2	Pancreatic adenocarcinoma (epithelial)	Nanopore long-read RNA-seq	GSM8887903, GSM8887904

Table 1. Summary of the samples and nanopore RNA sequencing datasets.

comprehensive characterization of transcriptional variations, including alternative splicing events and alternative polyadenylation, as well as the identification of protein isoforms pertinent to pancreatic cancer.

Methods

Cell culture. The pancreatic cancer cell lines AsPC-1 (Cat# SCSP-5080), Capan-2 (Cat# SCSP-5310), Mia-PaCa-2 (Cat# TCHu271), and SW1990 (Cat# TCHu201) were obtained from the cell bank of the Chinese Academy of Sciences. PANC0203 (Cat# CRL-2553), PANC1005 (Cat# CRL-2547), Panc0327 (Cat# CRL-2549), BXPC3 (Cat# CRL-1687), Capan-1 (Cat# HTB-79) were obtained from American Type Culture Collection (ATCC). HuP-T4 (Cat# CBP60540) was obtained from Cobioer. These cell lines were cultured under recommended conditions. All cell lines were maintained in a humidified incubator at 37 °C with 5% CO_2. BXPC3 and Capan-2 were cultured in RPMI1640 medium (Bioagrio) supplemented with 10% fetal bovine serum (FBS, Vazyme). Capan-1 was cultured in Iscove's Modified Dulbecco's Medium (IMDM, Servicebio) supplemented with 20% FBS. PANC0203, PANC1005, and Panc0327 were cultured in RPMI1640 with 15% fetal bovine serum and 1‰ insulin. Mia-PaCa-2 was maintained in Dulbecco's Modified Eagle Medium (DMEM) (Bioagrio) supplemented with 10% FBS, and 2.5% horse serum (Beyotime). SW1990 and AsPC-1 were cultured in DMEM supplemented with 10% FBS. HuP-T4 was cultured in Modified Eagle Medium (MEM, Bioagrio) with 20% FBS and 1% Non Essential Amino Acids (Thermo Fisher) with 1 mM Sodium Pyruvate (Thermo Fisher).

All culture medium was supplemented with penicillin-streptomycin (P/S, Bioagrio). All cell lines were routinely tested for Mycoplasma contamination using a Quick Cell Mycoplasma Rapid Detection Kit and authenticated via short tandem repeat (STR) profiling to confirm their identity. Cells were passaged at 70-80% confluency using 0.25% Trypsin-EDTA (Gibco).

RNA extraction, library construction, and nanopore long-read RNA sequencing. Total RNA was extracted from the samples, and poly(A) + mRNA was isolated using the NEBNext Poly(A) mRNA Magnetic Isolation Module (New England BioLabs), following the manufacturer's protocol to selectively enrich for poly(A) RNA. This method ensured the removal of rRNA and most other non-coding RNA species thus providing high-quality mRNA for downstream applications. For long-read RNA sequencing, cDNA synthesis was performed using the strand-switching protocol provided by Oxford Nanopore Technologies (ONT). In this protocol, the first-strand synthesis incorporates a unique sequence at the 3′ end of the cDNA to enable strand-specific information. Full-length cDNA libraries were then prepared from the poly(A)-selected mRNA using the ONT cDNA-PCR Sequencing Kit (SQK-PCS109). The cDNA was amplified through PCR for 13-14 cycles, using specific barcoded adapters from the Oxford Nanopore PCR Barcoding Kit (SQKPBK004) to facilitate sample multiplexing. After PCR amplification, a 1D sequencing adapter was ligated to the cDNA to prepare the samples for sequencing. These prepared libraries were subsequently loaded onto a FLOPRO002 R9.4.1 flow cell, which was then inserted into a PromethION sequencer for high-throughput sequencing. The sequencing process was managed using MinKNOW (version 23.07.12), ONT's proprietary software, which allowed real-time monitoring of run performance and quality control. Basecalling was conducted with Dorado (version 7.1.4, https://github.com/nanoporetech/dorado) using the high-accuracy model. All sequencing were carried out by Wuhan Benagen Technology Co., Ltd. (Wuhan, China), ensuring high-quality output for subsequent analysis.

Long-read RNA-seq data processing. Basecalled ONT long-read RNA-seq FASTQ files were first processed using Porechop[16] (version 0.2.4, https://github.com/rrwick/Porechop) to remove adapter sequences. The trimmed reads were then subjected to pre-alignment quality control and filtration to remove reads with Phred score < 7 or length < 200 bp. Filtered reads were aligned to the human reference genome (GRCh38) using the FLAIR pipeline[13] (version 1.4, https://github.com/BrooksLabUCSC/flair). For alignment, minimap2[17] (version 2.17-r941, https://github.com/lh3/minimap2) was employed with the following parameters: *minimap2 -ax splice ./reference/ref-human-ont.mmi $i/out.pass.fq -t 10 -o $i/aln.sam*. Indels were removed from the alignments, and splicing sites were refined using the FLAIR-specific function. Only splice junctions annotated in the GENCODE v38 gene set, supported by a minimum of three uniquely mapped reads, were considered valid. Additionally, unmatched splicing sites were corrected by mapping them to the nearest valid site within a 10-nt window. This process ensured that only accurate and validated splicing events were included in the final analysis.

Sample	Total bases (GB)	Total reads	Mean length	Median length	N50 length	Mean quality
PANC0203_rep1	12.10	8984301	1347	1004	1696	12.87
PANC0203_rep2	12.86	9834031	1308	961	1646	12.82
PANC1005_rep1	12.26	9234118	1327	967	1743	12.95
PANC1005_rep2	12.06	9372961	1286	909	1714	12.97
Panc0327_rep1	12.58	9619700	1307	946	1673	12.37
Panc0327_rep2	12.12	9379692	1292	938	1640	12.37
AsPC-1_rep1	12.45	11406233	1091	772	1472	12.53
AsPC-1_rep2	12.69	11223560	1131	818	1490	12.38
BXPC3_rep1	12.57	8925558	1408	987	1854	13.04
BXPC3_rep2	12.44	9108138	1366	950	1815	12.96
Capan-1_rep1	12.89	8635492	1493	1059	1935	12.36
Capan-1_rep2	12.74	8313849	1533	1115	1929	12.35
Capan-2_rep1	14.42	9996224	1443	1061	1821	12.98
Capan-2_rep2	14.21	9327643	1524	1151	1879	12.93
HuP-T4_rep1	12.93	8966596	1442	1029	1853	12.40
HuP-T4_rep2	13.36	9103175	1467	1061	1859	12.27
Mia-PaCa-2_rep1	14.61	9554448	1529	1149	1859	12.82
Mia-PaCa-2_rep2	13.95	9849365	1417	1022	1784	12.87
SW1990_rep1	12.63	9972004	1267	908	1617	12.42
SW1990_rep2	12.42	8961562	1386	1039	1711	12.31

Table 2. Summary statistics of nanopore long-read RNA sequencing reads.

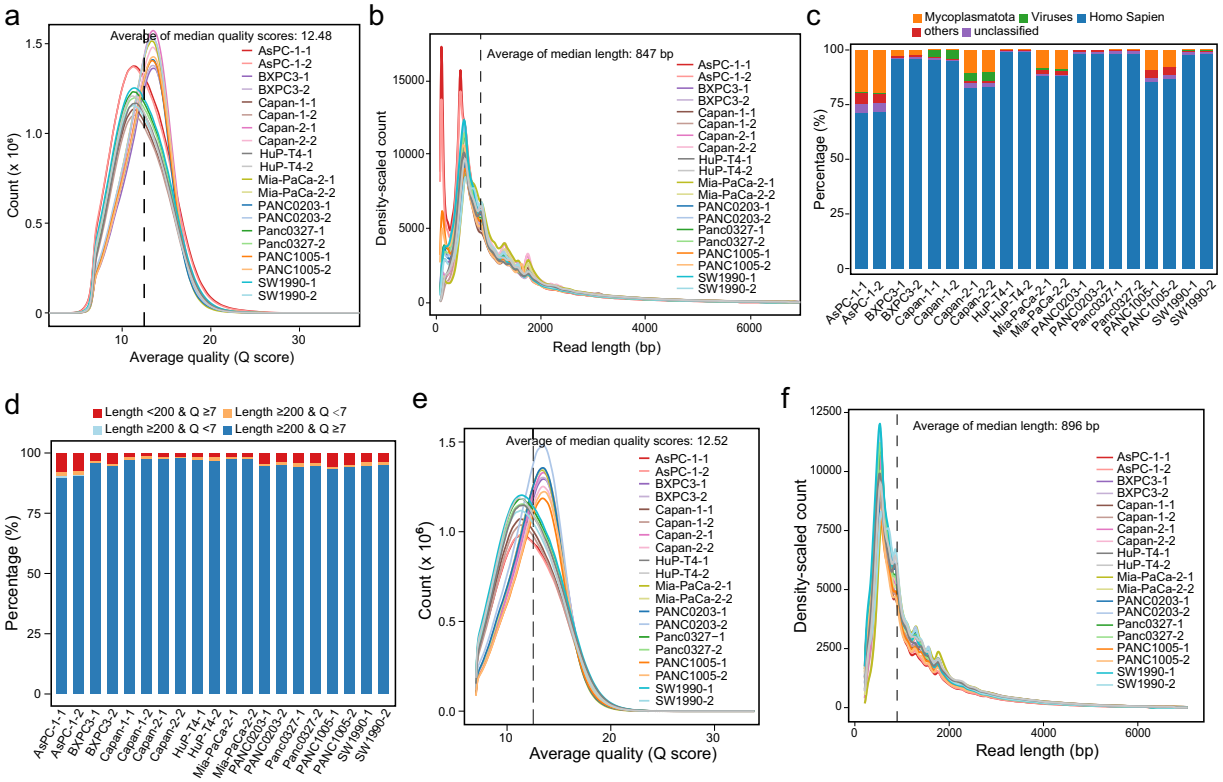

Fig. 1 Pre-alignment quality control of nanopore long-read sequencing reads. (**a**) Density plot illustrating the distribution of average quality scores for pre-filtered sequencing reads across each sample. (**b**) Density plot depicting the distribution of read length for pre-filtered sequencing reads in each sample. (**c**) Stacked bar plots representing the percentages of reads mapped to different species. (**d**) Stacked bar plots showing the percentages of reads with distinct lengths and quality scores after excluding those mapped to non-human genomes. (**e**) Density plot illustrating the distribution of average quality scores for post-filtered sequencing reads across each sample. (**f**) Density plot depicting the distribution of read lengths for post-filtered sequencing reads in each sample.

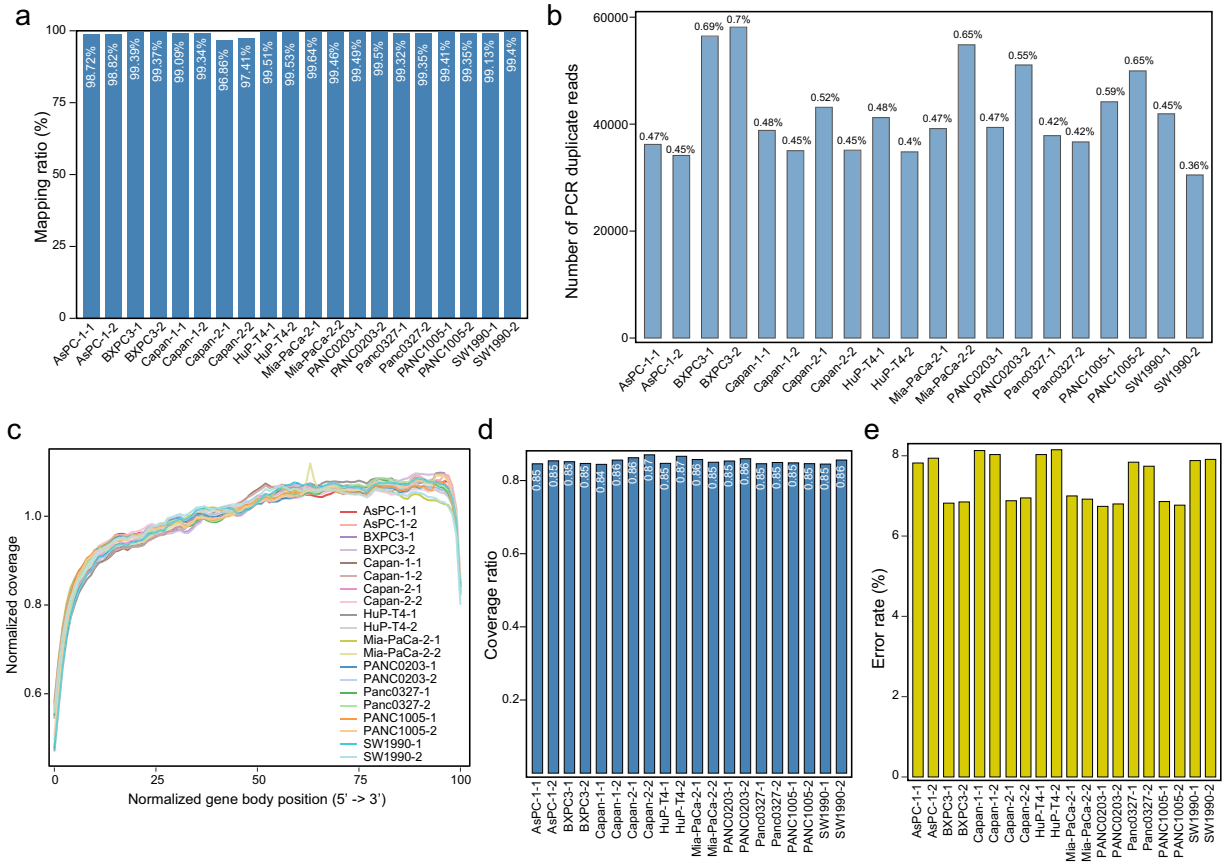

Fig. 2 Post-alignment quality control of nanopore long-read sequencing data. (**a**) Bar plots illustrating the proportion of sequencing reads successfully mapped to the human reference genome (GRCh38) for each sample. (**b**) Bar plots depicting the number of PCR duplicate reads identified per sample. (**c**) Line charts representing the distribution of normalized read coverage across transcript normalized positions in each sample, calculating with picard toolkit. (**d**) Bar plots displaying the percentage of reads covering more than 80% of their aligned transcript length in each sample. (**e**) Bar plots showing the estimated sequencing error rates, calculated as the proportion of mismatches, insertions, and deletions relative to total aligned bases, for each sample.

Basic statistics of the basecalled nanopore long-read RNA sequencing data were summarized in Table 2. Most sequencing reads exhibited high average quality scores (12.48) across all samples (Fig. 1a). The median read length across samples was approximately 847 bp (Fig. 1b). The majority of reads from most samples mapped successfully to the human genome, but some samples, notably AsPC-1-1 and AsPC-1-2, showed a notable proportion of non-human reads (Fig. 1c). Upon taxonomic classification of these reads using Kraken2 (version 2.1.3, https://github.com/DerrickWood/kraken2), we found that the majority aligned to the *Mycoplasmatota*, indicating the presence of mycoplasma contamination rather than contamination from other bacterial species. Mycoplasma contamination is a recognized issue in cell culture that can affect cell physiology, gene expression, and experimental outcomes[18,19]. While mycoplasma was not detected in all cell lines, we acknowledge it as a limitation in the affected samples. Importantly, we confirmed that human transcriptome profiles remained highly correlated between biological replicates, suggesting that the contamination did not substantially disrupt the overall transcriptomic landscape captured in this dataset. Nevertheless, we advise users to be aware of potential non-human reads and recommend applying read filtering strategies to exclude them when conducting expression analyses or downstream computational studies. Reads from all samples predominantly demonstrated good sequencing quality, with lengths of $\geq$200 bp and quality scores of $\geq$7 (Fig. 1d). We subsequently filtered out reads shorter than 200 bp, reads with quality score below 7, and reads mapped to non-human genomes. This filtering step slightly improved the overall average quality scores (Fig. 1e) and median read lengths (Fig. 1f).

The filtered reads were aligned to the human reference genome (GRCh38), resulting in high mapping ratios across all samples (Fig. 2a). Aligned reads were clustered according to genomic coordinates ($\pm$20 bp), chromosome locations, and splice junction patterns. Within each cluster, PCR duplicates were identified by calculating pairwise sequence similarity using the Levenshtein distance, marking reads with $\geq$95% identity to a representative sequence as duplicates. This strategy accounts for Nanopore-specific alignment variability and sequencing errors without inadvertently collapsing biologically distinct isoforms. PCR duplicates were flagged using the SAM flag (0x400), representing approximately 0.5% of total reads on average across samples (Fig. 2b). PCR duplicate reads were excluded from subsequent transcript quantification analyses. We then calculated normalized coverage across

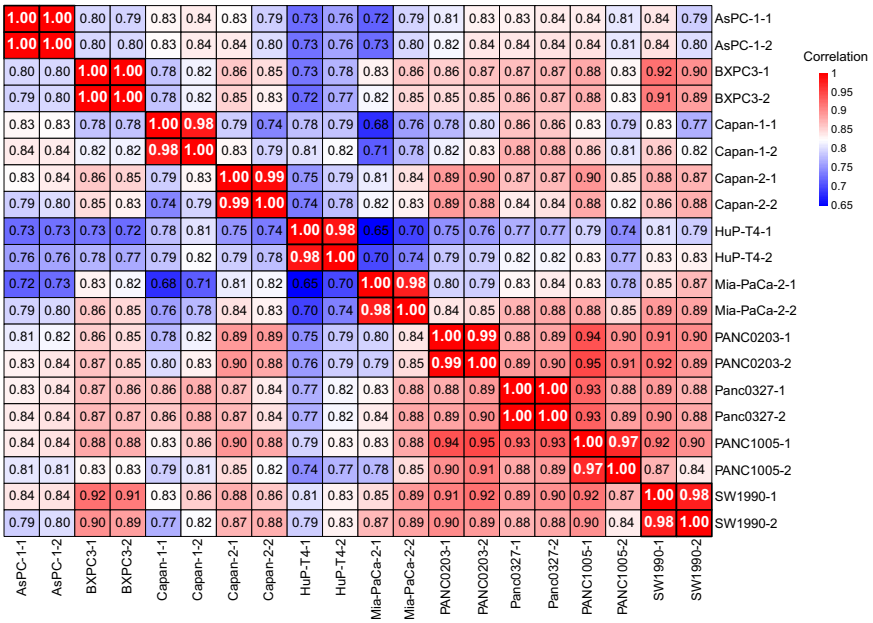

Fig. 3 Sample-wise correlation of transcript expression profiles.

transcript positions by picard toolkit (version 2.20.4-SNAPSHOT, https://github.com/broadinstitute/picard), which showed a slightly decline from the 3′ end to 5′ end (Fig. 2c). In each sample, approximately 85% of reads covered at least 80% of their aligned transcript length (Fig. 2d). Finally, we estimated sequencing error rates by computing the proportions of mismatches, insertions, and deletions relative to total aligned bases. The resulting average error across samples was approximately 7% (Fig. 2e). The PCR duplicates and sequencing errors were assessed by custom scripts. In the analysis of internal priming, we found that approximately 7.7% of reads had homopolymeric A stretches ($\geq$6 A) near their start sites, suggesting a moderate degree of internal priming[20]. While not dominant, this artifact should be considered when interpreting transcript boundary features. Most reads remain unaffected, and transcript coverage and correlation between replicates indicate that core transcriptome profiling is robust.

Data Records

The basecalled ONT long-read RNA-seq FASTQ files and expression files have been deposited in the Gene Expression Omnibus (GEO) database under the accession number GSE293661[21] (https://identifiers.org/geo/GSE293661) and the Sequence Read Archive (SRA) database with accession number of SRP576072 (https://identifiers.org/ncbi/insdc.sra:SRP576072)[22]. Annotations of all identified full-length transcripts are deposited under the primary GSE293661 dataset in GTF, FASTA, and BED format files.

Technical Validation

To ensure the integrity and accuracy of our RNA sequencing data, we implemented rigorous quality control measures throughout the sample preparation and sequencing processes. The concentrations of isolated RNA and synthesized cDNA were measured using a Qubit fluorometer (Life Technologies) to ensure sufficient input material for library preparation. The quality and size distribution of the prepared libraries were evaluated using an Agilent 2100 Bioanalyzer with the High Sensitivity DNA Kit. This step confirmed the presence of appropriately sized fragments and the absence of undesirable artifacts. Optimal conditions for adapter ligation and motor protein binding were determined using ONT's recommended protocols and tools, such as the Library Preparation Guide and the Flow Cell Priming Kit instructions. These guidelines ensured efficient library preparation and sequencing performance. Control RNA samples with known sequences were included to monitor sequencing accuracy and consistency. By integrating these validation steps, we ensured the reliability and reproducibility of our ONT RNA sequencing data, providing a solid foundation for subsequent analyses.

We assessed transcriptome correlations among samples by calculating pairwise Spearman correlation coefficients of transcript profiles. Replicate samples exhibited the highest correlations within the same cell lines (Fig. 3), indicating the robustness and reliability of our ONT RNA sequencing data. High correlation among biological replicates is essential for ensuring the reproducibility of RNA-seq experiments, as it reflects consistent gene expression measurements across replicates[2].

While our dataset provides high-coverage, full-length cDNA transcript information across ten pancreatic cancer cell lines, we recognize there is room for further improvement in read length and completeness. In particular, future library preparations may benefit from optimizing reverse transcription conditions to reduce premature termination events, using updated Oxford Nanopore Technologies (ONT) kits that incorporate enhanced enzymes and chemistries designed to improve cDNA integrity, and applying enzymatic strategies to minimize fragmentation during library preparation. Additionally, the adoption of newer ONT flow cells

(such as R10.4.1) and basecalling models can further increase read accuracy and yield. These optimizations would contribute to capturing a higher proportion of full-length transcripts, supporting even more comprehensive analyses of transcript isoform and RNA regulatory mechanisms in cancer transcriptomes.

Usage Notes

In this study, we included two biological replicates for each of the ten pancreatic cancer cell lines. The high correlation observed between replicates (Fig. 3) indicates strong technical reproducibility and supports the reliability of the transcriptomic profiles generated using long-read sequencing. While two replicates provide a solid foundation for descriptive and resource-generation studies, we acknowledge that three or more biological replicates are generally considered optimal for studies involving statistical inference, differential expression, or analysis of biological variability. Therefore, although our data are suitable for transcript discovery, isoform annotation, and other exploratory analyses, we recognize that the use of only two replicates per cell line may limit the power of certain downstream analyses.

Users should be aware of an important limitation of this dataset. Although all cell lines used in this study were confirmed free of mycoplasma at the time of routine testing, subsequent quality checks revealed evidence of mycoplasma contamination in some samples. Mycoplasma infection is a well-recognized issue in cell culture and can influence cell physiology, gene expression, and other molecular readouts, potentially confounding downstream analyses[18,19]. We therefore explicitly note this contamination as a limitation, and recommend that users interpret transcriptomic features and derived analyses with caution, particularly when comparing these data to other mycoplasma-free cell line datasets.

Data availability

The dataset has been deposited to GEO (GSE293661).

Code availability

Custom code used for data analysis in this study is deposited in GitHub: https://github.com/lishenglilab/Pancreatic-cancer-long-read-RNA-seq.

Received: 5 May 2025; Accepted: 4 September 2025;
Published online: 20 October 2025

References

1. Stark, R., Grzelak, M. & Hadfield, J. RNA sequencing: the teenage years. *Nat Rev Genet* **20**, 631–656 (2019).
2. Conesa, A. *et al*. A survey of best practices for RNA-seq data analysis. *Genome Biol* **17**, 13 (2016).
3. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* **10**, 57–63 (2009).
4. Hu, W. *et al*. Systematic characterization of cancer transcriptome at transcript resolution. *Nat Commun* **13**, 6803 (2022).
5. Byrne, A., Cole, C., Volden, R. & Vollmers, C. Realizing the potential of full-length transcriptome sequencing. *Philos Trans R Soc Lond B Biol Sci* **374**, 20190097 (2019).
6. Wang, Y., Zhao, Y., Bollas, A., Wang, Y. & Au, K. F. Nanopore sequencing technology, bioinformatics and applications. *Nat Biotechnol* **39**, 1348–1365 (2021).
7. Rhoads, A. & Au, K. F. PacBio Sequencing and Its Applications. *Genomics Proteomics Bioinformatics* **13**, 278–289 (2015).
8. Wu, J., Hu, W. & Li, S. Long-read transcriptome sequencing reveals allele-specific variants at high resolution. *Trends Genet* **39**, 31–33 (2023).
9. Monzo, C., Liu, T. & Conesa, A. Transcriptomics in the era of long-read sequencing. *Nat Rev Genet*, (2025).
10. Amarasinghe, S. L. *et al*. Opportunities and challenges in long-read sequencing data analysis. *Genome Biol* **21**, 30 (2020).
11. Ament, I. H., DeBruyne, N., Wang, F. & Lin, L. Long-read RNA sequencing: A transformative technology for exploring transcriptome complexity in human diseases. *Mol Ther* **33**, 883–894 (2025).
12. Rang, F. J., Kloosterman, W. P. & de Ridder, J. From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. *Genome Biol* **19**, 90 (2018).
13. Tang, A. D. *et al*. Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. *Nat Commun* **11**, 1438 (2020).
14. Chen, Z. *et al*. Long-read transcriptome landscapes of primary and metastatic liver cancers at transcript resolution. *Biomark Res* **12**, 4 (2024).
15. Li, Q. *et al*. Unraveling the hidden complexity of cancer through long-read sequencing. *Genome Res*, (2025).
16. Wick, R. R., Judd, L. M., Gorrie, C. L. & Holt, K. E. Completing bacterial genome assemblies with multiplex MinION sequencing. *Microb Genom* **3**, e000132 (2017).
17. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
18. Drexler, H. G. & Uphoff, C. C. Mycoplasma contamination of cell cultures: Incidence, sources, effects, detection, elimination, prevention. *Cytotechnology* **39**, 75–90 (2002).
19. Olarerin-George, A. O. & Hogenesch, J. B. Assessing the prevalence of mycoplasma contamination in cell culture via a survey of NCBI's RNA-seq archive. *Nucleic Acids Res* **43**, 2535–2542 (2015).
20. Svoboda, M., Frost, H. R. & Bosco, G. Internal oligo(dT) priming introduces systematic bias in bulk and single-cell RNA sequencing count data. *NAR Genom Bioinform* **4**, lqac035 (2022).
21. Luo, S. N. *et al*. https://identifiers.org/geo/GSE293661. *GEO* (2025).
22. Luo, S. N. *et al*. https://identifiers.org/ncbi/insdc.sra:SRP576072. *NCBI Sequence Read Archive* (2025).

Acknowledgements

This work was supported by Shanghai Rising-Star Program (23QA1407800 to S.L.) and National Key R&D Program of China (2021YFA1102300 to S.L.).

Author contributions

S.Li conceived and designed the project; S.Li and Y.Zheng supervised the project; J.F. conducted cell culture and prepared nanopore long-read RNA-seq libraries; S.Li, S.Luo and Y.Zeng performed data analysis; S.Luo conducted the data visualization; H.W. and Y.Zheng interpreted the results; S.Li drafted the manuscript with input from all other authors; all listed authors read and approved the final manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to Y.Z. or S.L.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.