PacBio

Whitepaper

Comparative studies show PacBio full-length RNA sequencing advantages over other long-read technology

102-326-592 REV01 NOV2023

PacBio

Long reads for full-length RNA sequencing

RNA sequencing (RNA-Seq) has become a ubiquitous tool in molecular biology for our understanding of genomic function. Knowing which transcripts are expressed in which tissues, conditions, or cells, allows assessment of the functional impact of transcriptional changes, since many mRNA transcripts are then translated into proteins. Meanwhile, other non-coding RNAs serve a myriad of regulatory roles.

Alternative splicing (AS) in eukaryotic species generates functional diversity by using a different combination of exons in the same gene. In humans, up to 95% of genes undergo alternative splicing to encode proteins with different functions (Pan et al., 2008). Aberrant AS has been shown to lead to neurological diseases, cancer, and is responsible for aging, infection, inflammation, immune disorders, etc. (Ren et al., 2021). In plants, 40–60% of genes exhibit alternative splicing, which is thought to be related to environmental fitness (Shang et al., 2017).

RNA-Seq using short reads cannot fully resolve isoform structures, as the complex nature of alternative splicing prohibits unambiguous transcript assembly with even the most sophisticated computational tools (Stark et al., 2019). Long-read RNA-Seq eliminates the need for transcript assembly by sequencing full-length cDNAs and has enabled new discoveries across many disease and biological applications (Figure 1).

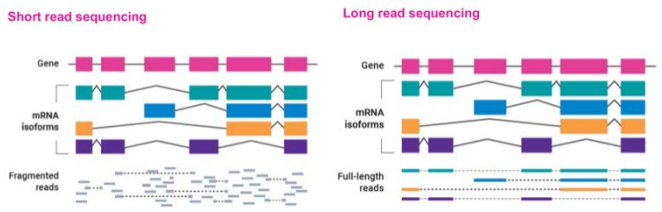

Figure 1. Full-length RNA sequencing using the PacBio® Iso-Seq® method. Long-read RNA sequencing eliminates the need for transcript assembly, which cannot accurately resolve the isoform structure. The Iso-Seq method sequences the entire full-length cDNA to provide an unambiguous view of the transcriptome.

Goal	Typical Size	Requirement
Full isoform structure	1–3 kb, up to 10+ kb for some	Long reads
Isoform quantification	Same as above	Long reads
Allele-specific isoform expression	Same as above	Long reads & high accuracy
Fusion isoform	Same as above	Long reads & high accuracy
Splice junction (SJ) detection	1–3 bp	High accuracy
Small variant detection	1–10 bp	High accuracy

Table 1. Full-length RNA sequencing analysis goals and requirements.

Third-party comparative studies evaluating PacBio® and Oxford Nanopore for full-length RNA sequencing

Several studies have compared PacBio and Oxford Nanopore (ONT) for use in whole transcriptome analyses (Pardo-Palacio et al., 2023; Calvo-Roitberg et al., 2023; Prjibelski et al., 2023; Leung et al., 2023). These studies assessed the ability of long-read platforms to:

- Retain usable reads after quality filtering.
- Sequence full-length isoforms from 5’ to 3’ ends.
- Accurately characterize splice sites.
- Detect novel genes and isoforms.
- Quantify isoform-level expression.

Across the following studies, authors were able to detect rare, longer, and more complete full-length transcript isoforms using PacBio systems compared to ONT. Additionally, while some analyses were conducted on earlier PacBio long-read systems with lower throughput, these studies were able to identify differentially expressed isoforms comparable to ONT cDNA data at higher read depth.

Systematic assessment of PacBio + ONT for isoform identification and quantification

Summary of Pardo-Palacio et al. (2023)

- The PacBio Iso-Seq method detected the greatest number of genes and longer transcripts, with higher abundance resolution compared to ONT.
- On matched samples, more ONT reads did not consistently lead to improved transcript identification.

In Pardo-Palacios et al. (2023), the Long-read RNA-Seq Genome Annotation Assessment Project (LRGASP) consortium generated over 400 million long reads using PacBio and ONT systems. SQANTI3, a long-read isoform classification and QC tool was used to uniformly assess all sequencing data and software output.

The consortium found that **PacBio sequencing detected the greatest number of genes.** Using SQANTI3's isoform classification nomenclature, the PacBio Iso-Seq method frequently had the most FSM, NIC, and NNC isoforms, while ONT data more frequently included anti-sense and -genic genomics transcripts, which are likely to be artifacts.

In addition, while more ONT data was collected – sometimes up to 10-times as much as matching PacBio data – the **Iso-Seq method more exclusively detected transcripts that were longer and had lower expression.** The consortium observed that "**more [ONT] reads did not consistently lead to more transcripts**, indicating that read quality and length are important factors for transcript identification" (Figure 2).

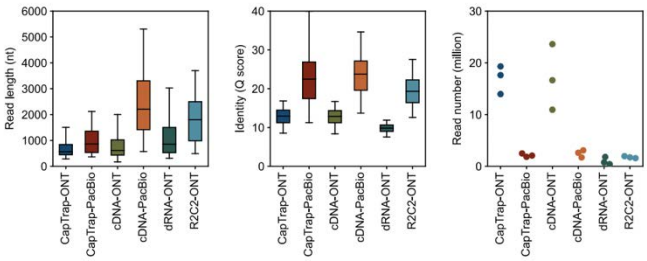

Figure 2. Comparison of different library preparation methods and sequencing platforms in the LRGASP study. Despite having fewer number of reads, PacBio Iso-Seq (cDNA-PacBio) method detected the most genes and found longer and rarer isoforms.

For isoform quantification, while both technologies showed good reproducibility and consistency across the replicates, the consortium found the **Iso-Seq method to have 2-fold higher abundance resolution (ability to quantify isoforms) compared to ONT cDNA data.**

Assessment of transcript completeness in long-read data

Summary of Calvo-Roitberg et al. (2023)

- ONT data is likely to have more mono-exonic artifacts.
- The PacBio Iso-Seq method captures true transcript start and end sites more accurately than ONT, with ONT showing more severe truncation.
- Orthogonal evidence (CAGE peak and polyA sites) show that most Iso-Seq reads have 5′ and 3′ evidence support.

Calvo-Roitberg et al. (2023) reanalyzed published ONT (direct RNA and cDNA) data against PacBio Iso-Seq data to assess the 5′ and 3′ completeness of the sequenced transcripts.

Both ONT data types generated shorter reads (mean ~1kb) compared to PacBio Iso-Seq data (mean ~2kb). The authors found that PacBio data had significantly less (~10%) mono-exonic reads compared to ONT data (~25% for direct RNA and ~35% for direct cDNA). While mono-exonic transcripts do genuinely exist, the authors conclude these higher proportions of mono-exonic reads in ONT data are a result of protocol and sequencing artifacts.

When comparing to annotated start and end sites, PacBio had a higher proportion of reads starting in the first annotated exon (~75%) compared to ONT (~30% for direct RNA and ~50% for direct cDNA), where the same trend is seen in reads ending the last annotated exon for PacBio (~90%) compared to ONT (~50% and ~75% respectively).Within the first/last exons, PacBio also showed reads starting and ending closer to the transcript start/end sites. The authors conclude that while both long-read systems can fail to capture the true 5′ ends, due to natural 5′ degradation and cDNA synthesis protocols, **ONT data showed more severe truncation, especially at the 5′ ends.**

Finally, using orthogonal 5′ start site (CAGE) and 3′ end site (polyA site) information, the **authors found PacBio reads to have significantly more 5′ and 3′ evidence support than ONT data.**

Novel isoform discovery with long reads

Summary of Prjibelski et al. (2023)

- For discovery of known and novel transcripts, PacBio data yielded more consistent performance across a variety of software tools compared to ONT, for both reference-based and reference-free models.
- IsoQuant correctly reconstructed more confirmed novel isoforms with PacBio data, compared to ONT data on the same samples.

In Prjibelski et al. (2023), authors describe a graph-based software called IsoQuant that can be used for *de novo* transcript discovery with or without a reference genome annotation. IsoQuant was trained and tested on numerous datasets, including real and simulated mouse and human PacBio and ONT data. To test the discovery power of the tool, 15% of expressed GENCODE isoforms were arbitrarily hidden during analysis, to be used as ground truth for assessing novel transcript discovery.

Compared to other tools for reference-based and reference-free transcript discovery, **IsoQuant applied to PacBio data yielded highly accurate transcript models compared to ONT, which varied substantially depending on software.** On real human data, "IsoQuant's potential false predictions are below 25% on ONT cDNA data and below 10% on PacBio." On mouse brain data, the authors found that IsoQuant correctly reconstructs 71% of the confirmed novel transcripts with PacBio data, compared to 49% of these 76 novel isoforms on ONT. On simulated mouse data, **ONT data resulted in poorer performance in all categories, with thousands of false positives identified.**

Differential transcript usage and isoform-level expression with long-read sequencing

Summary of Leung et al. (2023)

- Using the Iso-Seq method, isoform-level expression can be quantified, with results that highly correlate with short-read RNA-seq data on the same samples.
- Differential transcript usage can be observed, even in cases where differential gene-level expression is not present, revealing underlying mechanisms of biology that may be missed by short-read RNA-seq.
- Results from ONT differential transcript expression analyses can be replicated with PacBio at a lower read depth.

In Leung et al. (2023), the authors leveraged Illumina short-read RNA-seq, the PacBio Iso-Seq method, and ONT cDNA sequencing to profile transcript diversity in wild-type and transgenic mice harboring a mutant form of human *tau*, which serves as a model organism for Alzheimer's disease (AD). Using the Iso-Seq method for whole transcriptome profiling of the entorhinal cortex, the authors showed that **differential transcript usage can explain underlying pathology, even in cases where gene-level differential expression is absent.** At the whole transcriptome level, **Iso-Seq reads showed a strong correlation of effect sizes and directionality compared to short-read differentially expressed genes and could be used to find specific transcripts driving gene-level differences.**

The authors performed ultra-deep targeted transcriptome enrichment of 20 AD-associated genes using the Iso-Seq method on PacBio and cDNA capture on ONT. All mice had RNA isolated at multiple time points, and all mice used for ONT sequencing also had PacBio Iso-Seq data. An analysis of merged reads yielded 7,162 isoforms for the 20 genes, 86% of which were novel.

For differential isoform expression analysis, authors used DESseq 2 with full-length read counts as proxies of gene and transcript expression. Using ONT data, they found 12 novel differentially expressed transcripts in 6 major target genes (Figure 3 of Leung et al., 2023), all of which were replicated in the PacBio dataset, despite the lower depth of PacBio sequencing.

Note: A 1M SMRT® Cell was used on a PacBio Sequel® in this study; much higher coverage can now be achieved with a 25M SMRT Cell on the Revio™ system.

Conclusion

Long read lengths and high accuracy are essential components to successful whole transcriptome analysis. PacBio full-length RNA sequencing demonstrates a comparative advantage against ONT in its ability to:

- Recover rare and longer isoforms.
- Detect more genes.
- Quantify isoform expression.
- Generate more usable, high-quality reads.

With increased throughput on the latest PacBio sequencers, full-length RNA sequencing is now scalable and cost-effective for many applications. Discover the power of a more complete and accurate transcriptome to uncover biological insights with PacBio sequencing.

102-326-592 REV01 NOV2023

References + resources

References

Calvo-Roitberg, E., Daniels, R. F., & Pai, A. A. (2023). Challenges in identifying mRNA transcript starts and ends from long-read sequencing data. *bioRxiv*, 2023-07. https://doi.org/10.1101/2023.07.26.550536.

Leung, S. K., et al., (2023). Long-read transcript sequencing identifies differential isoform expression in the entorhinal cortex in a transgenic model of tau pathology. *bioRxiv*, 2023-09. https://doi.org/10.1101/2023.09.20.558220

Pan, Q., Shai, O., Lee, L. J., Frey, B. J., & Blencowe, B. J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature Genetics*, *40*(12), 1413-1415. https://doi.org/10.1038/ng.259.

Pardo-Palacios, F. J., et al., (2023). Systematic assessment of long-read RNA-seq methods for transcript identification and quantification. *bioRxiv*, 2023-07. https://doi.org/10.1101/2023.07.25.550582.

Prjibelski, A. D., et al., (2023). Accurate isoform discovery with IsoQuant using long reads. *Nature Biotechnology*, 1-4. https://doi.org/10.1038/s41587-022-01565-y

Ren, P., Lu, L., Cai, S., Chen, J., Lin, W., & Han, F. (2021). Alternative splicing: a new cause and potential therapeutic target in autoimmune disease. *Frontiers in Immunology*, *12*, 713540. https://doi.org/10.3389/fimmu.2021.713540.

Shang, X., Cao, Y., & Ma, L. (2017). Alternative splicing in plant genes: a means of regulating the environmental fitness of plants. *International Journal of Molecular Sciences*, *18*(2), 432. https://doi.org/10.3390/ijms18020432

Stark, R., Grzelak, M., & Hadfield, J. (2019). RNA sequencing: the teenage years. *Nature Reviews Genetics*, *20*(11), 631-656. https://doi.org/10.1038/s41576-019-0150-2

Resources

Application note – Kinnex full-length RNA kit for isoform sequencing

Whitepaper – Bulk and single-cell isoform sequencing for human disease research

Procedure & checklist – Preparing Iso-Seq libraries using SMRTbell® prep kit 3.0

LRGASP – Long-read RNA-seq Genome Annotation Assessment Project Consortium

Publication – SQANTI3: curation of long-read transcriptomes for accurate identification of known and novel isoforms