PacBio

Application note

Kinnex full-length RNA kit for isoform sequencing

Introduction

Alternative splicing (AS) in eukaryotic species generates functional diversity by expressing different combinations of exons in the same gene. Accurate characterization of full-length transcript isoforms generated by AS is critical for biological and disease studies. Bulk RNA-Seq using short reads cannot fully resolve isoform structures, as the complex nature of AS prohibits unambiguous transcript assembly with even the most sophisticated computational tools (Stark et al., 2019). Long-read RNA-Seq using PacBio® technology (the Iso-Seq® method) eliminates the need for transcript assembly by sequencing full-length cDNAs and enables new discoveries across many applications (Figure 1).

The *Kinnex™ full-length RNA kit* takes total RNA as input and outputs a sequencing-ready library that results in an 8-fold throughput increase compared to typical Iso-Seq libraries. Combined with the Iso-Seq analysis in SMRT® Link software, PacBio offers cost-effective isoform sequencing that does not require orthogonal sequencing methods. SMRT Link software produces an isoform classification report with abundance information that can be used by tertiary analysis tools.

102-326-591 REV02 NOV2023

PacBio

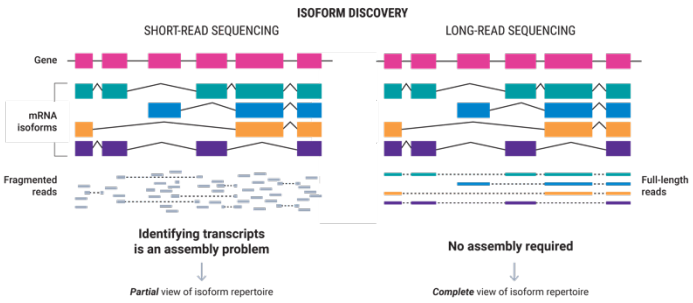


Figure 1. Long-read RNA sequencing eliminates the need for transcript assembly, which cannot accurately resolve the isoform structure. Long-read RNA-Seq using PacBio (the Iso-Seq method) sequences the entire full-length cDNA to provide an unambiguous view of the transcriptome.

Full-length RNA sequencing with the Iso-Seq method

Traditional RNA-Seq fragments cDNA for short-read sequencing (100–200 bp) must be followed by computational methods to infer the original transcript isoforms. However, given the complexity of alternative splicing, many isoforms share highly similar structures, and the inferred transcripts are often inaccurate. PacBio HiFi reads sequence full-length RNA isoforms without the need for cDNA fragmentation and transcript assembly (Figure 1), allowing for unambiguous full-length isoform detection.

HiFi sequencing advantages for full-length RNA sequencing

- Sequence full-length isoforms from 5′ to 3′ ends.
- Accurately characterize splice sites.
- Detect novel genes and isoforms.
- Obtain isoform read count information.

The Iso-Seq method sequences full-length transcripts using PacBio HiFi sequencing and has been applied to many areas in biology and disease. The Iso-Seq method has been used in human disease research to identify aberrant splicing linked to rare diseases, phenotypic traits, and neurological diseases. In cancer research, the Iso-Seq method has been used to discover cancer-driving mutations, fusion genes, and neoepitopes that could potentially be used as cancer vaccine candidates (Li et al., 2023).

The Iso-Seq method has also been used in plant and animal research to create high-quality genome annotations (Zhang et al., 2023) as well as identify parental-specific isoform expressions (Wang et al., 2020).

Kinnex full-length RNA kit

The *Kinnex full-length RNA kit* utilizes the MAS-Seq method to increase throughput on PacBio long-read sequencers. MAS-Seq is a concatenation method for joining cDNA molecules into longer fragments (Al'Khafaji et al., 2023). HiFi reads generated from sequencing the concatenated molecules can then be broken up bioinformatically to retrieve the original cDNA sequences. The result is higher throughput and reduced sequencing needs for cost-effective isoform sequencing.

The PacBio Iso-Seq workflow in SMRT Link processes the full-length cDNA sequences to classify them against a reference annotation (e.g., GENCODE) to identify novel genes and isoforms. The output consists of classified full-length isoforms with read counts that are compatible with tertiary analysis software.

The *Kinnex full-length RNA kit* takes total RNA (300 ng) as input and produces a sequencing-ready library in a two-day workflow.

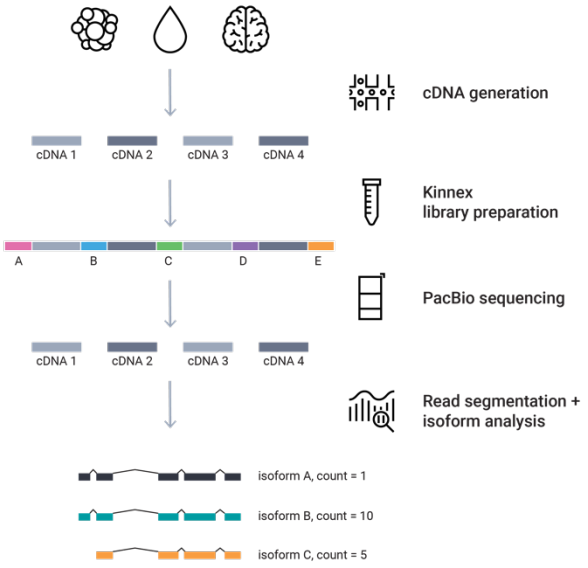


Figure 2. Kinnex full-length RNA sequencing. Full-length cDNA molecules are concatenated into a large-insert library, sequenced, then processed using the PacBio software.

Kinnex RNA library workflow

The Kinnex full-length RNA workflow (Figure 3) begins with total RNA and produces a sequencing-ready library.

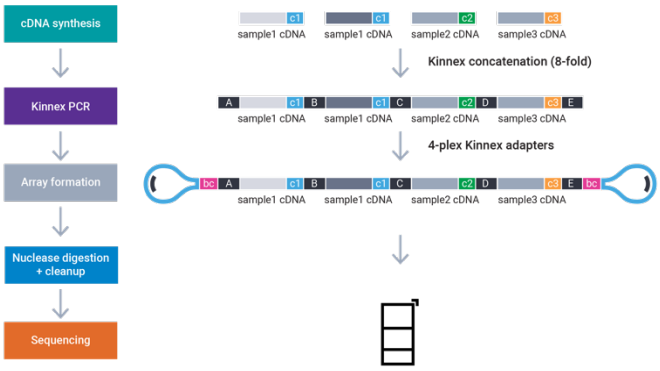


Figure 3. Kinnex full-length RNA library workflow.

Full-length cDNA molecules are synthesized with handles (using *Iso-Seq express 2.0 kit*) that are compatible with Kinnex array formation. cDNA barcodes are added as part of the cDNA amplification on the 5′ end. As partial arrays are removed during nuclease digestion, Kinnex adapter ligation ensures full arrays are enriched prior to HiFi sequencing. The barcoded cDNAs support up to 12-plex, while the Kinnex adapters support 4-plex at the library level.

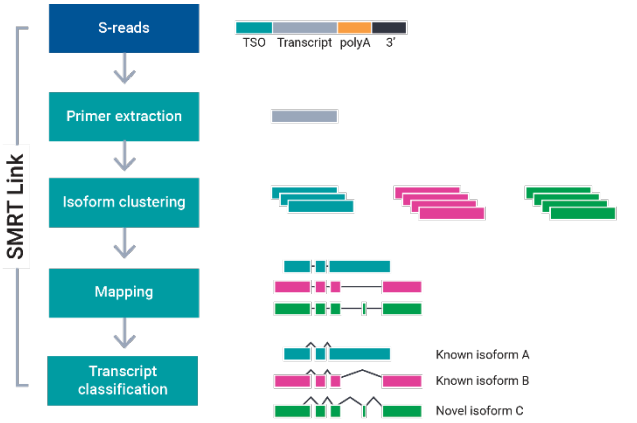
With proper full array formation and adequate sequencing, one SMRT® Cell on the Sequel® II/IIe and Revio™ systems are expected to achieve ~15 million and ~40 million cDNA sequences, respectively (Table 1).

Kinnex RNA bioinformatics workflow

The SMRT Link *Read segmentation and Iso-Seq* workflow (Figure 4) processes the HiFi reads generated from the Kinnex full-length RNA library to produce classified isoforms with read counts that are compatible with tertiary analysis tools.

Isoform clustering

FLNC reads are clustered by their sequencing similarity to produce isoform consensus sequences. This step is the last step of Iso-Seq analysis if no genome is provided.



ID	Gene	Transcript	Sample 1 TPM	Sample 2 TPM
PB.1.1	ARGN	Novel	100	200
PB.1.2	ARGN	ENST00000379370.7	10	1.4
		...		
PB.10.1	ACTB	ENST00000646664.1	8132	8000

Figure 4. Kinnex full-length RNA analysis using the *Read segmentation and Iso-Seq* workflow.

Mapping

If a genome is provided, isoform consensus sequences from the previous step are mapped and further collapsed by their exonic structures to produce isoforms as GFF files for visualization.

Transcript classification

If an annotation (e.g., Gencode) is provided, isoforms are classified against it using pigeon (the PacBio implementation of SQANTI3) to identify known and novel genes/isoforms. The Iso-Seq workflow can jointly analyze pooled sample reads to produce a unified isoform annotation with per-sample read counts, both raw and normalized as counts per million (CPM).

Metric	Performance
Sample preparation time	2 days
Expected library size	11,000–18,000 bp
Target P1 loading	60–80%
Expected HiFi yield	1.5–2.5 million HiFi reads (Sequel II/IIe) 4–6 million HiFi reads (Revio)
Expected full array %	80–90%
Expected read yield	~15 million reads (Sequel II/IIe) ~40 million reads (Revio)

Table 1. Target Kinnex full-length RNA library performance

Currently, SMRT Link only supports transcript classification for human and mouse samples. Non-human/mouse samples will require customized annotation GTF files to be run via the command line

SMRT Link considerations

Below are some common considerations in running the Iso-Seq workflow and recommendations.

Currently the SMRT Link *Read segmentation and Iso-Seq* workflow support human and mouse reference genomes and annotations to produce classified isoforms with read counts. If working with other organisms, see Table 2 for analysis recommendations.

Reference/annotation	Analysis recommendation
Human or mouse	Use the Iso-Seq workflow with pre-loaded human/mouse annotation to get mapped, unique isoforms with classifications and read count information (FASTA, GFF, TXT).
Model organism with good annotation	Run Iso-Seq workflow with uploaded reference genome to get mapped, unique isoforms (FASTA, GFF). Generate pigeon-compliant annotation and use the command line for isoform classification with read count information (TXT).
Non-model organism with genome	Run Iso-Seq workflow with uploaded reference genome to get mapped, unique isoforms (FASTA, GFF).
No genome	Run Iso-Seq workflow without reference genome to get unique isoforms (FASTA).

Table 2. Analysis recommendations for Iso-Seq data based on reference genome and annotation availability.

While sequencing depth varies depending on the experimental goal and sample, Table 3 offers some general recommendations.

Kinnex public dataset release

The Kinnex RNA dataset release consists predominantly of HG002 cell line and the Universal Human Reference RNA (UHRR). Additional samples from WTC-11 cell line, human brain, sorghum and mouse (to be published) are included as comparison. After the *Read segmentation and Iso-Seq* workflow, each sample obtained >20,000 unique genes (Table 4). Transcript lengths ranged from 100 bp to 11,000 bp with minor differences that appear sample- and species-dependent (Figure 5).

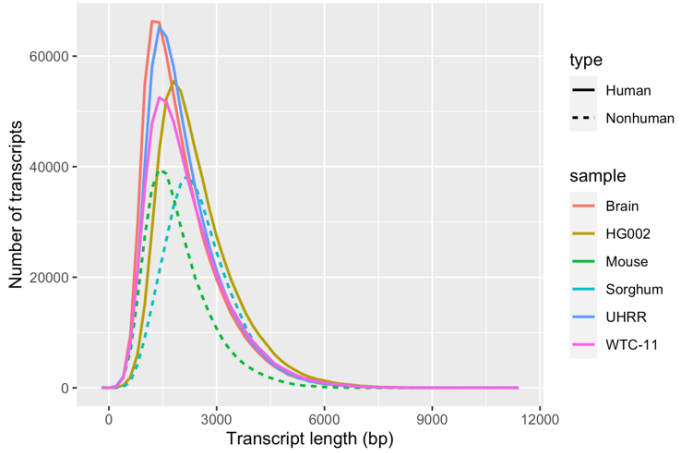


Figure 5. Transcript length of different Kinnex libraries shows variations that appear sample- and species-specific but are largely within the same size ranges.

Study goal	Isoform discovery and quantification of moderate-to-rare transcripts	Isoform discovery of high expressed transcripts	Comprehensive transcript annotation in a species
Example	Disease vs. normal tissues with multiple replicates	Disease cohort with >20+ samples	Plant or animal with multiple tissue types
Target depth	10M reads per sample	5M reads per sample	5M reads per sample
Library	4-plex cDNA for 1 Revio SMRT Cell or 2-plex cDNA for 1 SMRT Cell 8M	8-plex cDNA for 1 Revio SMRT Cell, or 3-plex cDNA for 1 SMRT Cell 8M	
Analysis	*Read segmentation and Iso-Seq* workflow with option to "pool reads and cluster together" to get a master isoform classification file with per-sample full-length read counts		

Table 3. Example sequencing and analysis recommendations based on different study goals.

Sample	Library	HiFi reads	S-reads	Mean S-read length	Known genes	Novel genes	Known isoforms	Novel isoforms
UHRR	Non-Kinnex – Sequel II/IIe	3,194,311	n/a	n/a	12,921	121	27,821	16,323
	Kinnex – Sequel II/IIe	2,720,033	20,453,853	1,918	18,903	1369	53,623	102,059
	Kinnex-Revio	6,546,645	47,250,258	1,914	22,365	4,223	68,087	231,467
HG002	HG002	5,984,046	38,740,671	2,227	18,230	8,448	55,689	281,460
WTC-11	Day 0 – rep 1	6,920,750	54,110,504	1,856	19,905	3,636	58,651	212,523
	Day 0 – rep 2	8,611,025	67,547,611	1,764	21,170	4,932	63,553	277,189
	Day 0 – rep 3	8,124,744	63,251,235	1,864	20,744	4,822	62,349	257,665
	Day 1 – rep 1	6,430,958	49,897,067	1,743	20,204	4,629	60,451	213,429
	Day 2 – rep 1	7,353,759	58,217,895	1,201	21,169	6,570	64,066	165,496
	Day 3 – rep 1	5,483,994	42,173,159	1,844	19,436	2,692	56,533	185,650
	Day 3 – rep 2	6,687,580	52,317,384	1,705	21,270	4,482	63,430	241,726
	Day 4 – rep 1	7,295,962	57,061,795	1,727	21,751	3,594	63,466	225,636
	Day 5 – rep 1	6,645,009	51,741,094	1,751	21,754	3,195	62,217	185,807
	Day 5 – rep 2	7,542,604	59,092,202	1,792	21,721	3,613	65,369	228,584
	Day 5 – rep 3	6,358,300	49,466,302	1,803	21,389	3,652	59,638	187,394

Table 4. Kinnex full-length RNA dataset release for the UHRR, HG002, and additional collaborator WTC-11 cell lines. HiFi reads were analyzed using the *Read segmentation and Iso-Seq* worfklow in SMRT Link v13.0. All samples were Kinnex libraries and sequenced on one Revio SMRT Cell with the exception of Non-Kinnex and Kinnex which were sequenced on the Sequel II/IIe system. Novel genes and isoforms are determined against Gencode v39 annotation using pigeon.

Comparing Kinnex against non-Kinnex libraries showed that the transcript lengths did not shift when using Kinnex concatenation or different sequencing platforms (Figure 6). Further, isoform abundances remained largely consistent (Figure 7).

Kinnex libraries showed high technical reproducibility (Table 5), consistent with what is shown for technical reproducibility in matching Illumina data (not shown).

Saturation curves showed that at 10 million reads, most of the known genes and isoforms could be detected (Figure 9). To corroborate the simulated saturation curves, we subsampled WTC-11 samples at 5 and 10 million reads each and pooled them at different plexity to simulate the total yields obtainable on Sequel II/IIe and Revio systems. As expected, a 3-plex 5M (for a total of 15M reads per SMRT Cell 8M on a Sequel II/IIe system) would detect fewer isoforms than an 8-plex 5M (for a total of 40M reads per Revio SMRT Cell) or a 4-plex 10M library (Figure 10).

Together, the saturation and subsampling data shows that at 10M reads per sample, ~80% of the known isoforms could be detected. Increasing sequencing depth also increases the number of novel isoforms detected, though most newly discovered isoforms would have lower abundances.

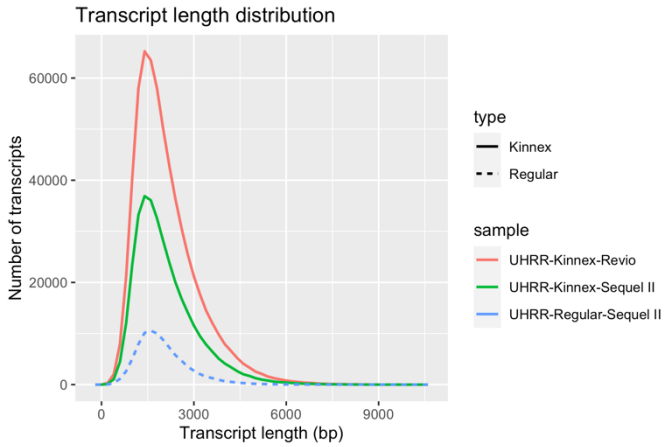


Figure 6. Transcript length distributions did not differ between Kinnex (concatenated) and unconcatenated libraries for the same UHRR sample.

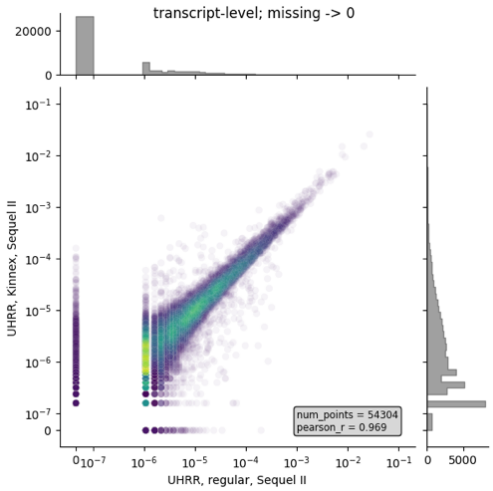


Figure 7. Kinnex concatenation did not skew isoform abundances. There was high isoform abundance correlation between Kinnex and non-Kinnex UHRR data on Sequel II/e system. Only known Gencode isoforms were compared. Similar correlation (0.964) was observed against Kinnex-Revio data (not shown).

WTC-11 day 0			
	Rep 1	Rep 2	Rep 3
Rep 1	1.00	0.80	0.79
Rep 2	0.80	1.00	0.81
Rep 3	0.79	0.81	1.00

WTC-11 day 5			
	Rep 1	Rep 2	Rep 3
Rep 1	1.00	0.80	0.80
Rep 2	0.80	1.00	0.79
Rep 3	0.80	0.79	1.00

Table 5. Good technical reproducibility in Kinnex libraries. Three technical replicates each from WTC-11 sample show high isoform abundance correlation for both day 0 and day 5, similar to observed correlation values for matching Illumina technical replicates (0.78–0.82, data not shown).

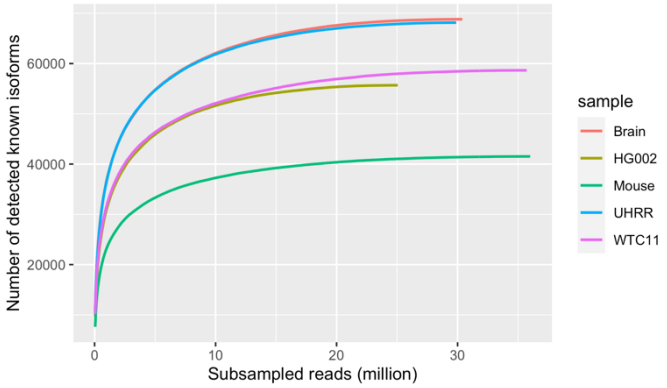


Figure 9. Saturation curves for WTC-11 Kinnex samples at the isoform level. At 10 million reads, most known isoforms were detected.

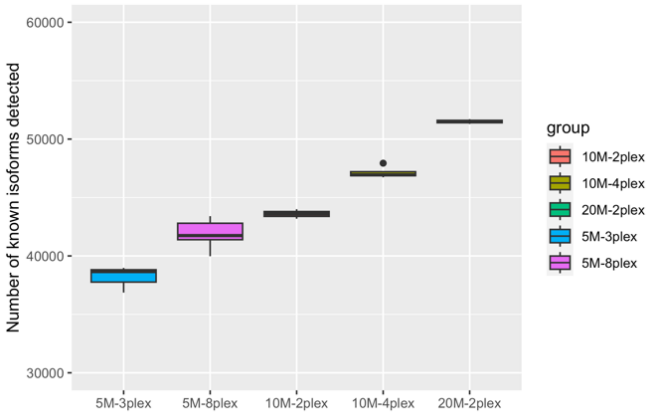


Figure 10. Number of detected known isoforms per sample at either 5 or 10 million read depth simulation. The higher the total pooled sequencing depth, as well as the higher number of per-sample reads, the more number of known isoforms are detected. The number of novel isoforms detected ranged from 30–70 k (for 5M-3-plex, 10M-2-plex, 5M-8-plex) to 90–130 k (for 10M-4-plex, 20M-2-plex).

Analysis of the Kinnex RNA datasets generated on the WTC-11 and UHRR samples demonstrated:

- Kinnex libraries did not alter detected transcript sizes or abundances compared to regular Iso-Seq libraries.
- Differences in detected transcript sizes varied by species and sample type, but largely remained the same in Kinnex libraries.
- High technical reproducibility across library replicates and PacBio long-read sequencing platforms.
- At 10 million reads, most known genes and isoforms were detected. The number of novel isoforms detected increases with sequencing depth but become increasingly rare.

The *Kinnex full-length RNA kit* together with SMRT Link analysis deliver high-quality, full-length that promises to deliver extraordinary insight into disease and biology.

Conclusion

The PacBio Iso-Seq method sequences full-length transcripts with high accuracy, enabling unambiguous isoform characterization, allele-specific isoform expression, differential expression analysis, and others.

The *Kinnex full-length RNA kit* increases throughput by 8-fold using the MAS-Seq concatenation technology. Coupled with flexible multiplexing strategies and a SMRT Link bioinformatics workflow, users can now achieve comprehensive isoform sequencing in a cost-effective manner.

Resources & references

Resources

Application brief – A more complete cancer transcriptome with the Iso-Seq method – single-cell and bulk RNA sequencing.

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

Application note – Bioinformatics tools for full-length isoform sequencing.

Kinnex full-length RNA dataset: https://pacb.com/datasets

Iso-Seq documentation: https://isoseq.how/

pigeon documentation: https://isoseq.how/classification/

References

Al'Khafaji, A. M., et al. (2023). High-throughput RNA isoform sequencing using programmed cDNA concatenation. *Nature Biotechnology*, 1-5. https://doi.org/10.1038/s41587-023-01815-7

Li, Z., et al. (2023). An isoform-resolution transcriptomic atlas of colorectal cancer from long-read single-cell sequencing. *bioRxiv*, 2023-04. https://doi.org/10.1101/2023.04.21.536771

Pardo-Palacios, F., et al., (2023). SQANTI3: curation of long-read transcriptomes for accurate identification of known and novel isoforms. *bioRxiv*, 2023-05. https://doi.org/10.1101/2023.05.17.541248

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

Wang, B., et al. (2020). Variant phasing and haplotypic expression from long-read sequencing in maize. *Communications Biology*, *3*(1), 78. https://doi.org/10.1038/s42003-020-0805-8

Zhang, R., et al. (2022). A high-resolution single-molecule sequencing-based Arabidopsis transcriptome using novel methods of Iso-Seq analysis. *Genome Biology*, *23*(1), 149. https://doi.org/10.1186/s13059-022-02711-0

102-326-591 REV02 NOV2023