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).

PacBio RNA sequencing has been shown to offer the longest read lengths and highest quality reads that are required for isoform discovery and isoform quantification (Pardo-Palacios et al., 2024). The *Kinnex™ full-length RNA kit* takes full-length cDNA generated by the *Iso-Seq express 2.0 kit* 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 *Read Segmentation & 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 REV03 NOV2024

PacBio

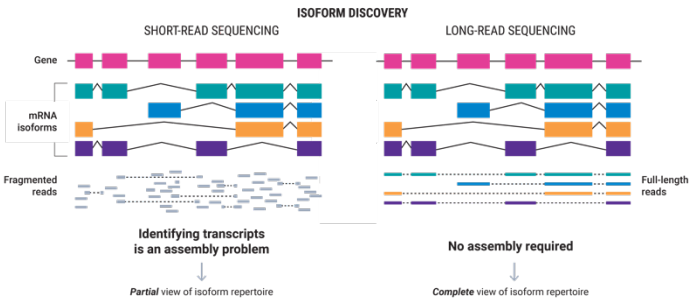



Figure 1. Long-read RNA sequencing eliminates the need for transcript assembly, which cannot accurately resolve the isoform structure. Long-read RNA-Seq using PacBio (the Iso-Seq method) sequences the entire full-length cDNA to provide an unambiguous view of the transcriptome.

## 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.
- Discover novel genes and isoforms.
- Detect variants and allele-specific isoform expressions.
- 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. It 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 full-length cDNA generated by the *Iso-Seq express 2.0 kit* as input and produces a sequencing-ready library in a two-day workflow.

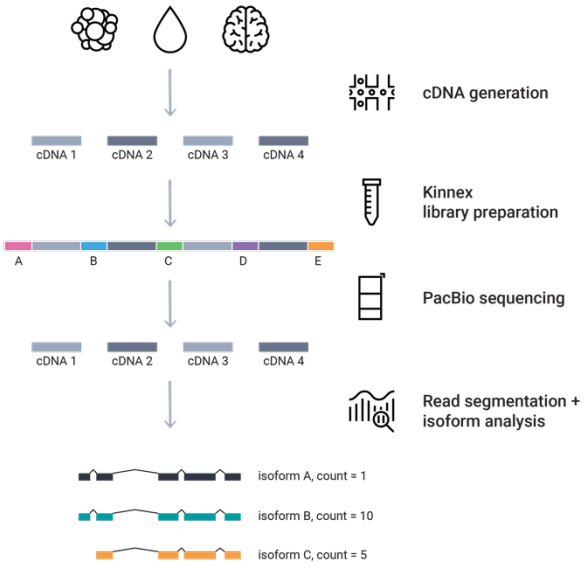



Figure 2. Kinnex full-length RNA sequencing. Full-length cDNA molecules are concatenated into a large-insert library, sequenced, then processed using the PacBio software.

## Kinnex RNA library workflow

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

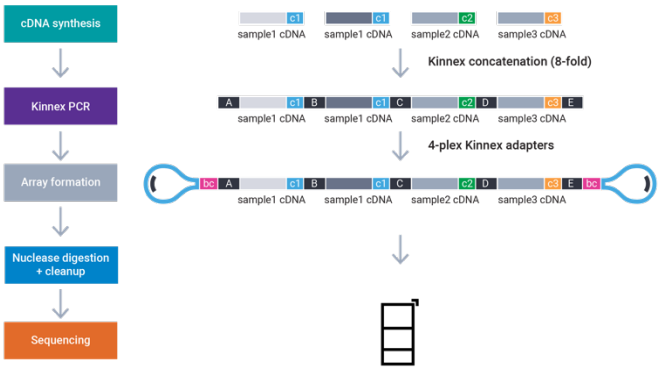



Figure 3. Kinnex full-length RNA library workflow.

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 Vega™ and Revio® systems are expected to achieve 20–30 million and 50–60 million cDNA sequences, respectively (Table 1).

| Metric | Performance |
|---|---|
| Sample preparation time | 2 days |
| Expected library size | 11,000–18,000 bp |
| Target P1 loading | 60–80% |
| Expected HiFi yield | 2.5–3.5 million HiFi reads (Vega)<br>6–7.5 million HiFi reads (Revio system with SPRQ™ chemistry) |
| Expected full array % | 80–90% |
| Expected read yield | 20–30 million reads (Vega)<br>50–60 million reads (Revio SPRQ) |

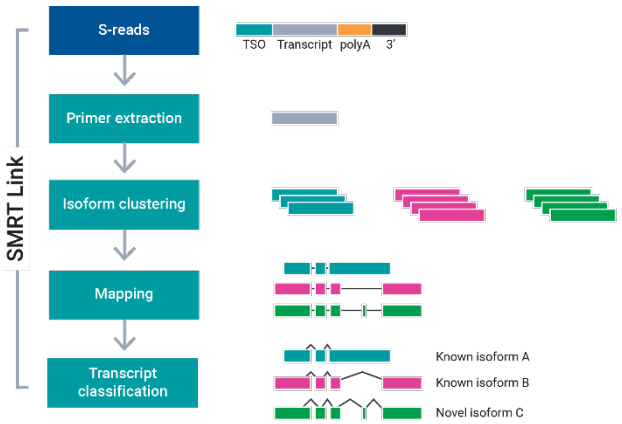
Table 1. Target Kinnex full-length RNA library performance.

## 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).

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 of the Universal Human Reference RNA (UHRR), HG002 cell line, human heart and cerebellum. 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).

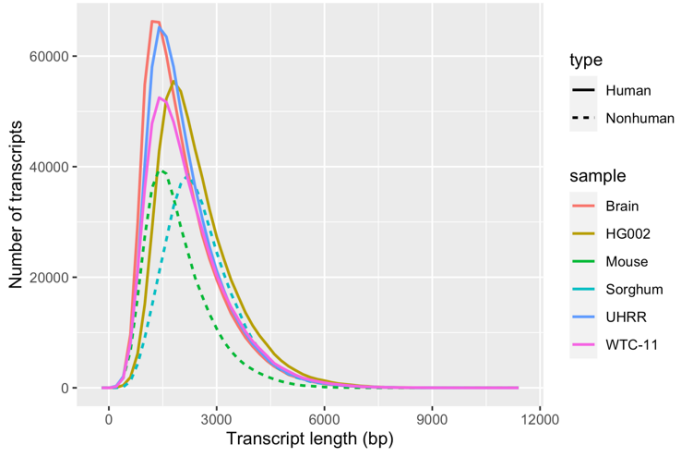



Figure 5. Transcript length of different Kinnex libraries shows variations that appear sample- and species-specific but are largely within the same size ranges.

| 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 | 3-plex cDNA for 1 Vega SMRT® Cell or 6-plex cDNA for 1 Revio SMRT Cell | 6-plex cDNA for 1 Vega SMRT Cell, or 12-plex cDNA for 1 Revio SMRT Cell | |
| 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 (6-plex cDNA) | Kinnex-Vega | 4,642,616 | 35,917,435 (~5 million per sample) | 1,759 | ~18,600 per sample | ~1180 per sample | ~51,600 per sample | ~82,200 per sample |
| | Kinnex-Revio | 8,247,394 | 63,611,648 (~10 million per sample) | 1,753 | ~20,800 per sample | ~2300 per sample | ~60,400 per sample | ~136,800 per sample |
| | Kinnex-Revio SPRQ | 10,448,251 | 81,377,012 (~13 million per sample) | 1,758 | ~21,760 per sample | ~3050 per sample | ~64,400 per sample | ~170,200 per sample |
| 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 and above. All samples were Kinnex libraries and sequenced on one SMRT Cell on Vega or Revio systems. Novel genes and isoforms are determined against Gencode v39 annotation using pigeon. More Kinnex full-length datasets can be found at https://pacb.com/datasets.

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 consistent across sequencing platforms (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 8, top). Alternatively, by looking at the number of additional known isoforms detected, the discovery rate drops below 0.1% at ~10 million reads, i.e., for every additional 1,000 reads sequenced, one more known isoform is detected (Figure 8, bottom). Increasing sequencing depth also increases the number of novel isoforms detected, though most newly discovered isoforms would have lower abundances.

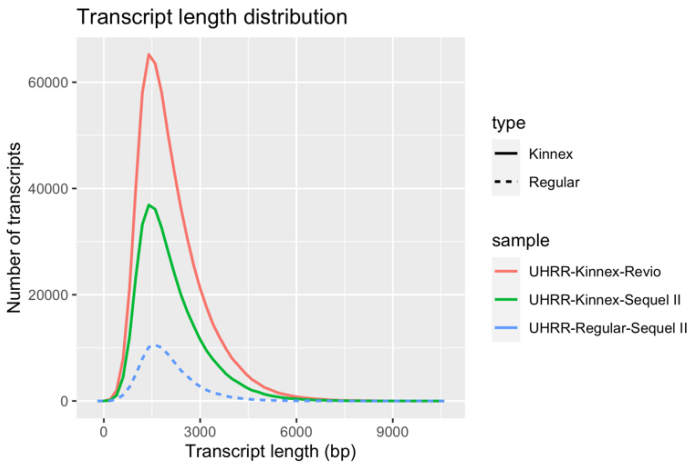



Figure 6. Transcript length distributions did not differ between Kinnex (concatenated) and unconcatenated libraries for the same UHRR sample.

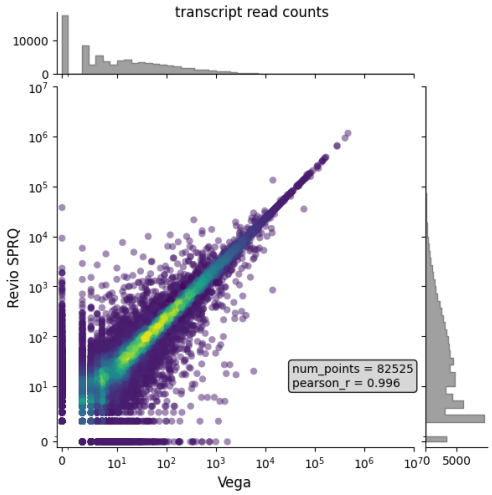



Figure 7. High isoform abundance correlation across different sequencing platforms. The same UHRR library run on Vega and the Revio (with SPRQ chemistry) system shows high isoform count concordance.

| | WTC-11 day 0 | | | WTC-11 day 5 | | |
|---|---|---|---|---|---|---|
| | Rep 1 | Rep 2 | Rep 3 | Rep 1 | Rep 2 | Rep 3 |
| Rep 1 | 1.00 | 0.80 | 0.79 | 1.00 | 0.80 | 0.80 |
| Rep 2 | 0.80 | 1.00 | 0.81 | 0.80 | 1.00 | 0.79 |
| Rep 3 | 0.79 | 0.81 | 1.00 | 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).

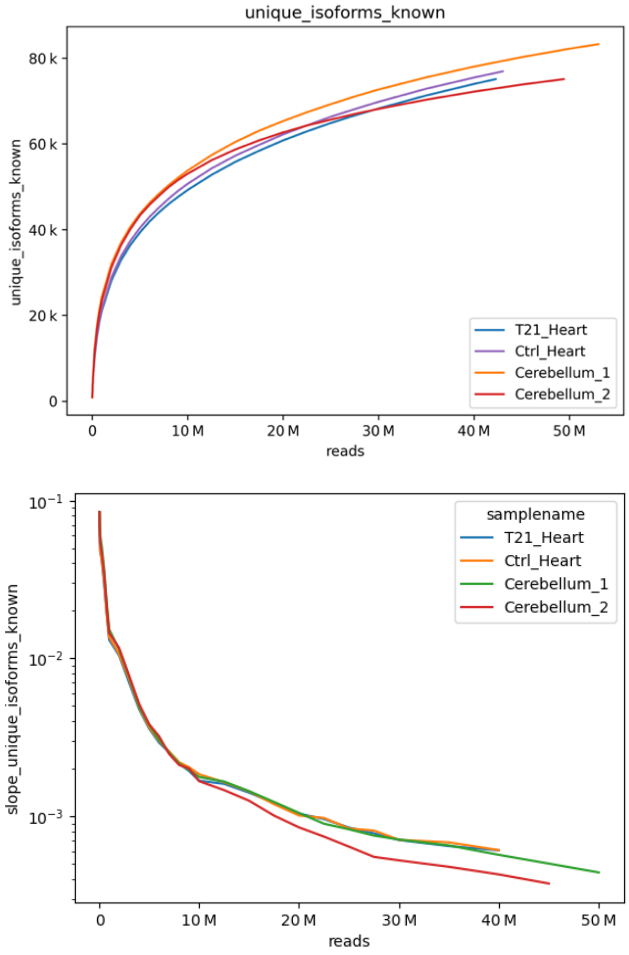



Figure 8. Saturation and discovery curves for known isoforms of human heart and cerebellum samples. The saturation curve (top) for known isoforms detected at different read depths shows that 80% of the known isoforms are detected at ~10M reads. The discovery rate for known isoforms at different read depths (bottom) drops below 0.1% at ~10M reads.

Analysis of the Kinnex RNA datasets generated on various human samples demonstrated:

- Kinnex full-length RNA kit is compatible with all PacBio long-read sequencing platforms including Sequel II/IIe, Vega, and Revio systems.
- 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 website

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. (2024). An isoform-resolution transcriptomic atlas of colorectal cancer from long-read single-cell sequencing. *Cell Genomics* https://doi.org/10.1016/j.xgen.2024.100641

Pardo-Palacios, F., et al. (2024). Systematic assessment of long-read RNA-seq methods for transcript identification and quantification. *Nature Biotech* https://doi.org/10.1038/s41592-024-02298-3

Pardo-Palacios, F., et al., (2024). SQANTI3: curation of long-read transcriptomes for accurate identification of known and novel isoforms. *Nature Methods* https://doi.org/10.1038/s41592-024-02229-2

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 REV03 NOV2024