Full-length transcriptome sequencing of melanoma cell line complements long-read assessment of genomic rearrangements

Elizabeth Tseng[1], Brendan Galvin[1], Ting Hon[1], Wigard P. Kloosterman[2], Meredith Ashby[1]
[1]Pacific Biosciences, Menlo Park, CA; [2]UMC Utrecht, Utrecht, Netherlands

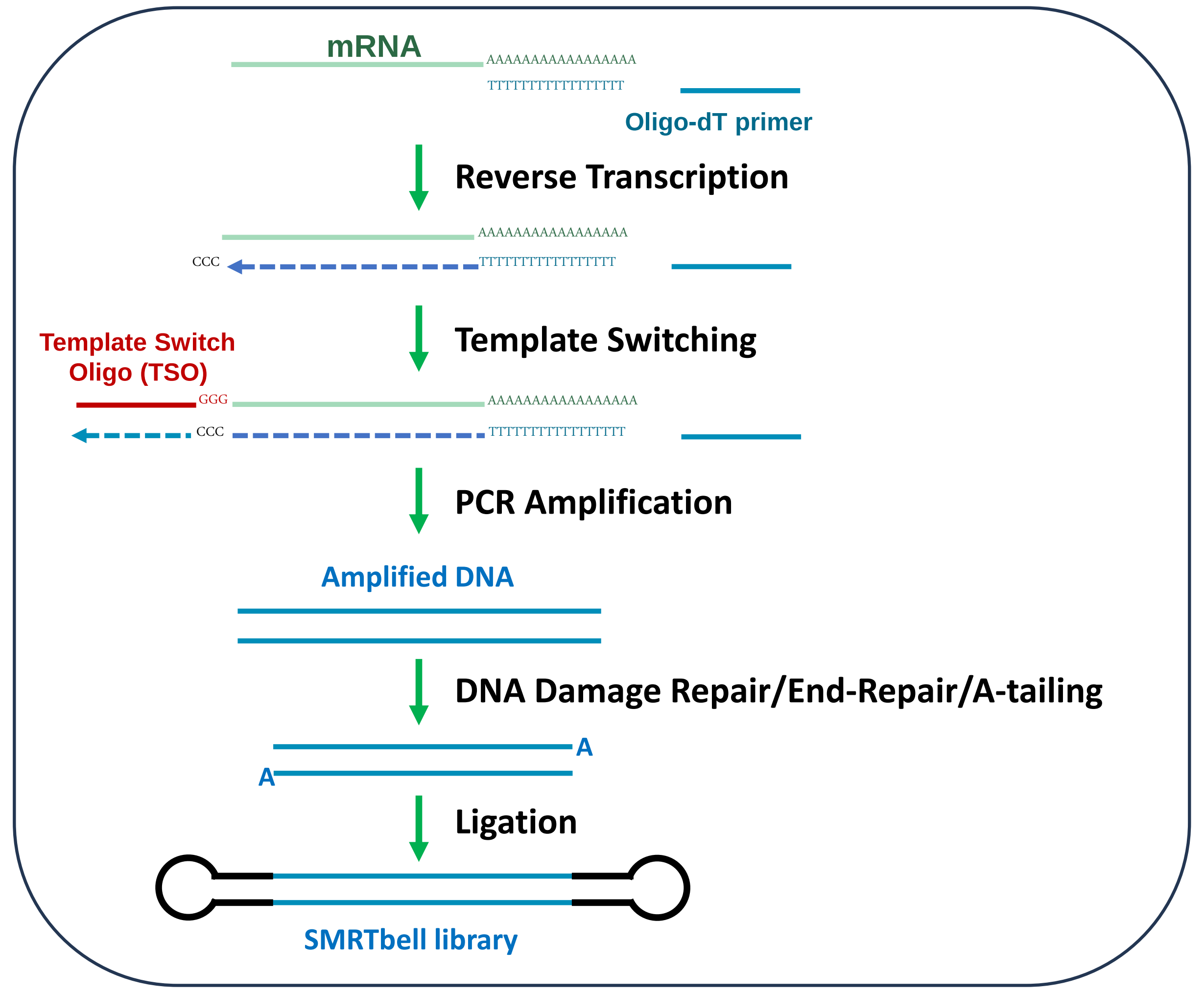
Abstract #: 1669

Abstract

Transcriptome sequencing has proven to be an important tool for understanding the biological changes in cancer genomes including the consequences of structural rearrangements. Short-read sequencing has been the method of choice, as the high throughput at low cost allows for transcript quantitation and the detection of even rare transcripts. However, the reads are generally too short to reconstruct complete isoforms. Conversely, long-read sequencing can provide unambiguous full-length isoforms, but lower throughput has complicated quantitation and high RNA input requirements has made working with cancer samples challenging.

Recently, the COLO 829 cell line was sequenced to 50-fold coverage with PacBio Single Molecule, Real-Time (SMRT) Sequencing. To validate and extend the findings from this effort, we have generated long-read transcriptome data using an updated PacBio Iso-Seq method, the results of which will be shared at the AACR 2019 General Meeting. With this complimentary transcriptome data, we demonstrate how recent innovations in the PacBio Iso-Seq method sample preparation and sequencing chemistry have made long-read sequencing of cancer transcriptomes more practical. In particular, library preparation has been simplified and throughput has increased. The improved protocol has reduced sample prep time from several days to one day while reducing the sample input requirements ten-fold. In addition, the incorporation of unique molecular identifier (UMI) tags into the workflow has improved the bioinformatics analysis. Yield has also increased, with 3.0 sequencing chemistry typically delivering >30 Gb per SMRT Cell 1M. By integrating long and short read data, we demonstrate that the Iso-Seq method is a practical tool for annotating cancer genomes with high-quality transcript information.

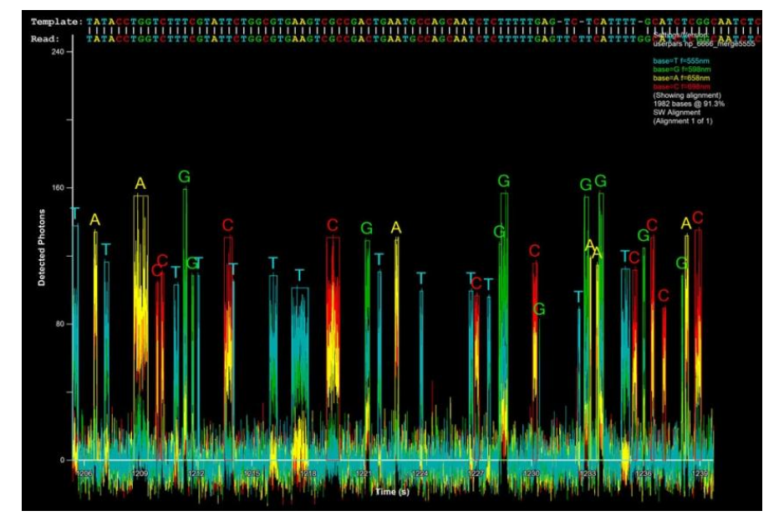
Iso-Seq Express Workflow



Iso-Seq Express Library Workflow Features:

- RNA to SMRTbell library in 1 day
- 300 ng total RNA input requirement
- Single RT and amplification reactions
- Increased full-length transcript recovery

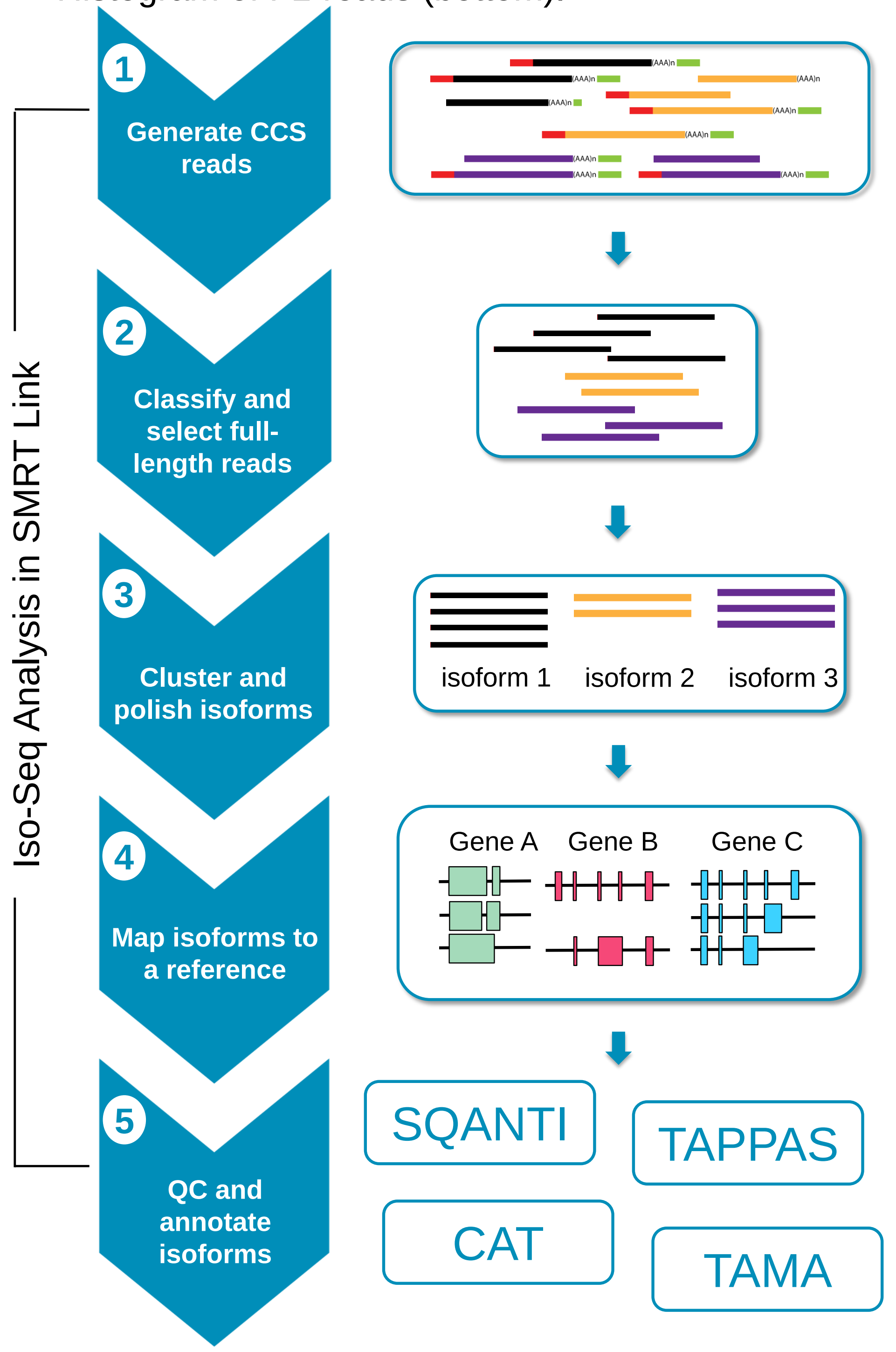
Figure 1. Iso-Seq Express Workflow. Full-length mRNA is converted into cDNA using the NEBNext Single Cell/Low Input RNA Library Prep Kit followed by PCR amplification. The amplified cDNA is converted into SMRTbell templates using the PacBio SMRTbell Express Template Prep Kit 2.0 for sequencing on the Sequel System.



	Iso-Seq Current
# of SMRT Cells 1M	6
CCS Reads	2,059,062
FL Reads	1,888,322 (92%)
FL Read Length	3,076 bp

Iso-Seq 3 Bioinformatics Workflow

Figure 2. Iso-Seq 3 Bioinformatics Workflow. Analysis pipeline outlined conceptually (left) and graphically (right) to demonstrate FL-read inputs and isoform outcomes of the improved workflow. Histogram of FL-reads (bottom).


Iso-Seq Analysis in SMRT Link
1
Generate CCS reads
2
Classify and select full-length reads
3
Cluster and polish isoforms
4
Map isoforms to a reference
5
QC and annotate isoforms
(AAA)n
isoform 1
isoform 2
isoform 3
Gene A
Gene B
Gene C
SQANTI
TAPPAS
CAT
TAMA

The Iso-Seq analysis workflow begins with the generation of high accuracy long reads using the circular consensus sequencing (CCS) method on a per molecule basis. Then, full-length reads are selected and trimmed of 5' and 3' primers and poly-A tails. The trimmed full-length reads are clustered at the isoform level and consensus is called. Lastly, the consensus isoforms can be optionally mapped back to the reference genome and/or used in downstream analysis.

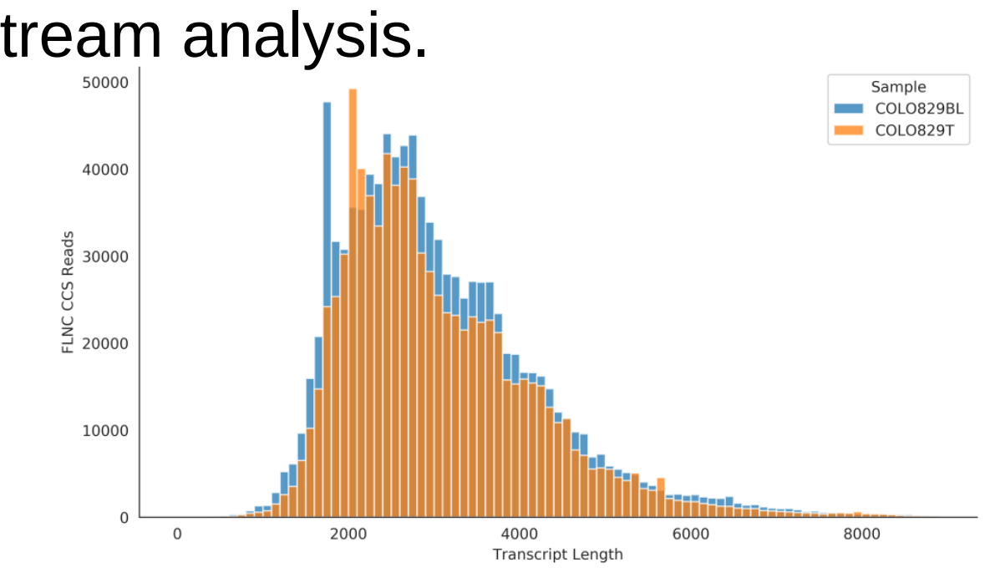
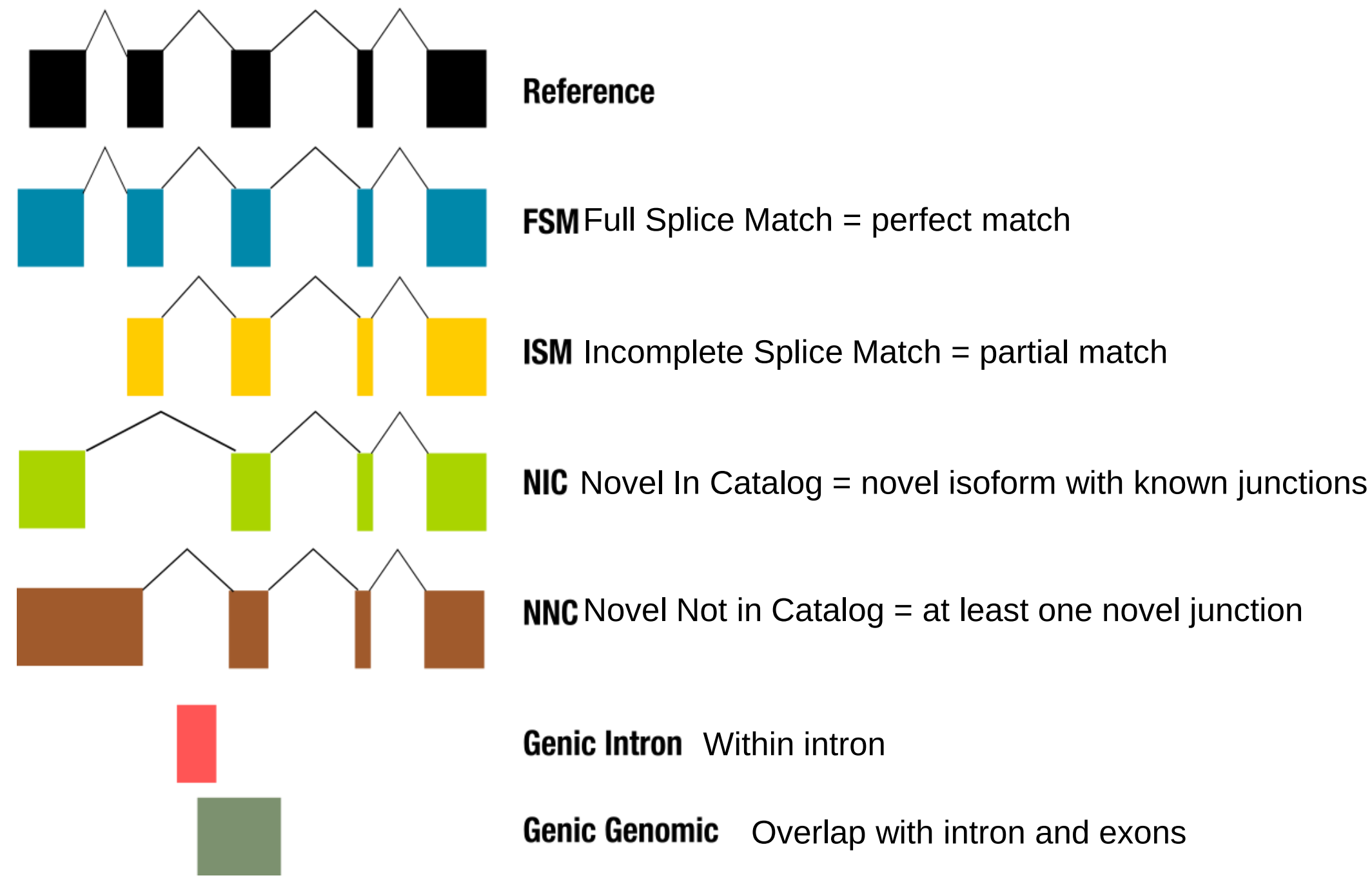
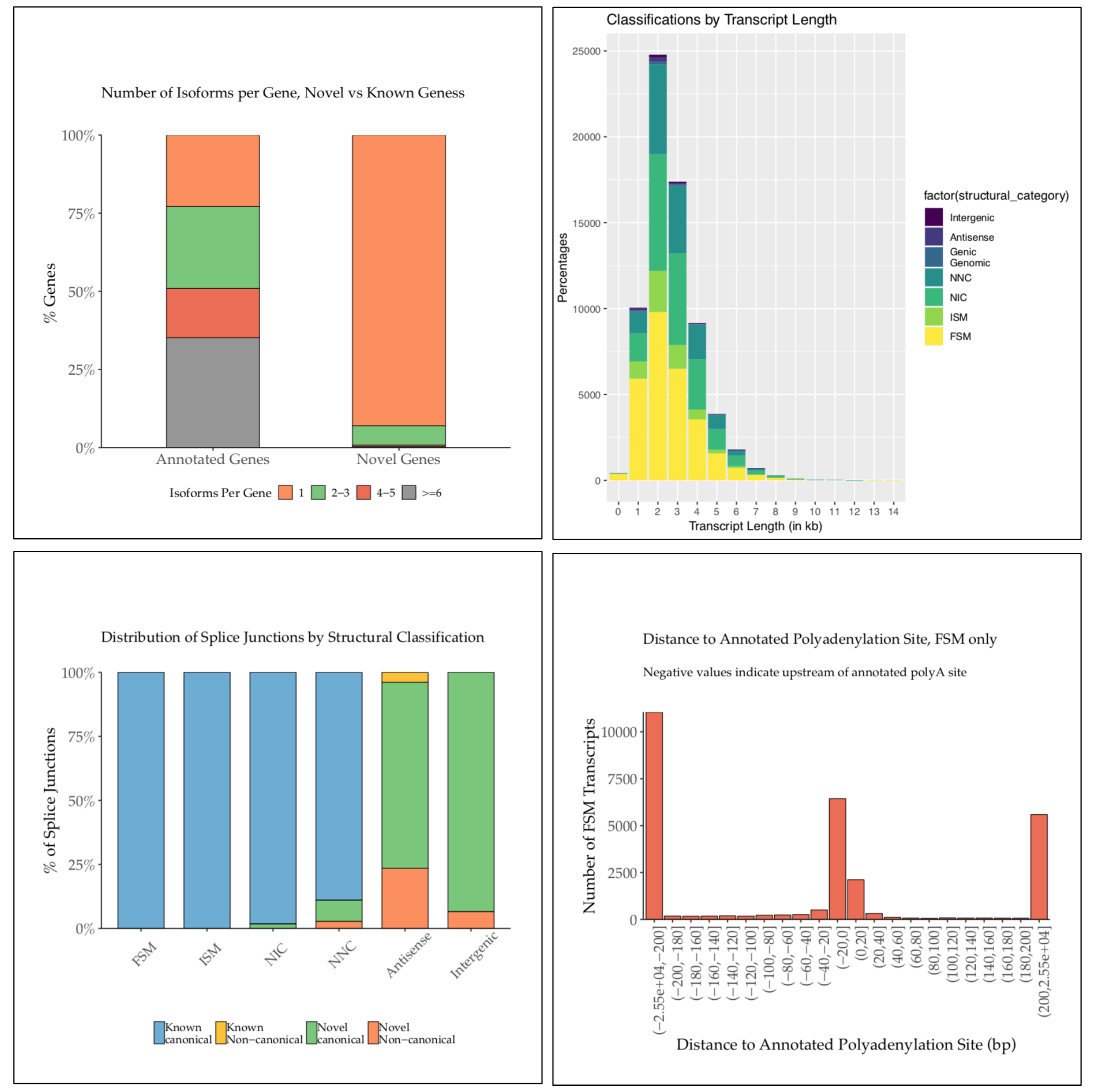


Figure 3. SQANTI Software compares the Iso-Seq analysis output against a reference annotation.

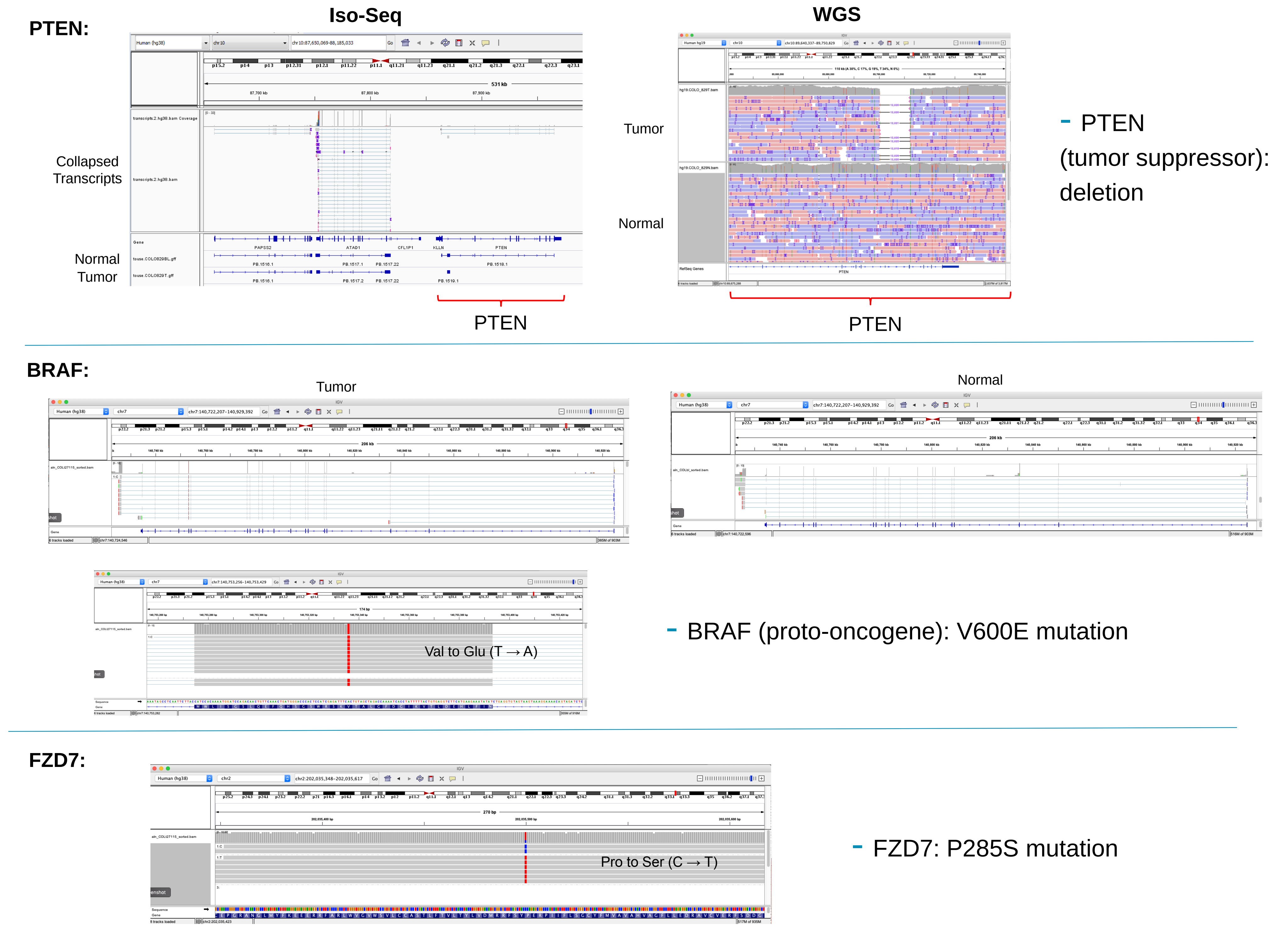
SQANTI is a pipeline for the in-depth characterization of isoforms obtained by full-length transcript sequencing without information about gene/transcript annotation or attribute description. SQANTI provides a wide range of descriptors of transcript quality and generates a graphical report to aid in the interpretation of the sequencing results[4].


Reference
FSM Full Splice Match = perfect match
ISM Incomplete Splice Match = partial match
NIC Novel In Catalog = novel isoform with known junctions
NNC Novel Not in Catalog = at least one novel junction
Genic Intron Within intron
Genic Genomic Overlap with intron and exons

Cells	Pol RL (mean)	Pol Base	CCS Reads	FL Reads	Unique Genes	Unique Isoforms
6	59 kb	145 Gb	2,059,062	188,322	12,806	68,553



COLO 829 - Iso-Seq Results



- PTEN (tumor suppressor): deletion
- BRAF (proto-oncogene): V600E mutation
- FZD7: P285S mutation

Figure 4. Identification of altered transcripts. Elimination of the PTEN transcript and mutations in BRAF and FZD7 can be identified in the Iso-Seq libraries.

Conclusions

- The consequences of many somatic events such as point mutations, deletions, and loss of heterozygosity, can be detected in the Iso-Seq data
- The PacBio Iso-Seq method is a valuable tool to characterize full-length transcripts and identify altered transcript sequences or abundances
- The improved workflow of the new Iso-Seq Express protocol makes the generation of Iso-Seq libraries easy

References

1. https://www.pacb.com/software
2. https://github.com/PacificBiosciences/pbbioconda
3. https://github.com/PacificBiosciences/IsoSeq_SA3nUP
4. https://bitbucket.org/ConesaLab/sqanti/