"Sequencing of individual barcoded cDNAs using Pacific Biosciences and Oxford Nanopore technologies reveals platform-specific error patterns" Supplemental Material

Supplemental Note “Experimental details”

Here we provide a brief description of the samples used and experimental protocols followed (taken from Joglekar et al., 2021). C57BL/6NTac (n=3) female pups were used in generating the data. For the single-cell experiments (P7; n=1), the brains were removed and placed on a stainless-steel brain matrix for mouse (coronal repeatable sections, 1mm spacing), and the hippocampus was dissected in cold PBS solution. After dissection, tissues were snap frozen in dry ice until processing. For 10X Visium experiments (P8; n=1) brains were fresh-frozen and embedded in OCT. Mice were housed in an air-conditioned room at 65-75°F (~18-23°C) with 40-60% humidity and a 12-h-light, 12-h-dark cycle. Animal experiments were approved by the institutional animal care and use committee (IACUC) of Weill Cornell Medicine

Tissue dissociation and 10x Genomics single-cell capture.

The dissected hippocampal tissue was dissociated into a single-cell suspension according to the manufacturer's protocol and the disassociated cells were captured on the 10x Genomics Chromium controller according to the Chromium Single Cell 3′ Reagent Kits V2 User Guide with the following modification. PCR cycles were increased, from the recommended ten cycles for recovery of 8,000 cells to 16 cycles to target a yield of cDNA enabling simultaneous Illumina and PacBio library preparation.

Illumina and Pacific Biosystems library preparation.

Illumina library preparation was performed using 100 ng of amplified cDNA following the Chromium Single Cell 3′ Reagent Kits V2 User Guide reducing final indexing PCR cycles to ten cycles from the recommended 14 cycles to increase library complexity. Sequencing was performed on HiSeq4000 according to 10x Genomics run mode. PacBio library preparation was performed with 500 ng of amplified cDNA using SMRTbell Express Template Prep Kit V2.0 to obtain Sequel II compatible library complex and was sequenced on a total 16 Sequel I SMRTcells with a run time of 10 hours and 10 Sequel II SMRTcells with a run time of 30 hours across samples and replicates.

PromethION library preparation and sequencing of cDNA.

Oxford Nanopore compatible library was produced using 350ng of either Sage Blue Pippin size-selected cDNA or non-selected cDNA derived from 10x Genomics following the Genomic DNA by Ligation protocol from Oxford Nanopore. Loading inputs on the PromethION was increased to 150fmol and sequenced using a R9.4.1 flow cell for 20 hours, and base-calling was done using Guppy (3.2.10).

Generation of circular consensus reads.

Using the default SMRT-Link parameters, we performed circular consensus sequencing (CCS) with IsoSeq3 with the following modified parameters: maximum subread length 14,000bp, minimum subread length 10bp, and minimum number of passes 3.

Barcode detection.

Post filtering for quality control, a list of 16mer barcodes corresponding to single cells or spots was obtained from 10X sequencing output. Assuming an error rate of 10%, poly(A) tails were identified by obtaining the position of 9 consecutive A's or T's in the last 100nt of PacBio CCS and ONT reads respectively. Using the position of the poly(A) tail, a sliding window was used to identify perfect matching barcodes in the 36 bp window preceding the poly(A) tail, and UMIs were identified as the 10 or 12 bp (depending on the dataset) following the barcode. This method is implemented in the scisorseqr package under the GetBarcodes function. Read statistics are shown in Supplemental Tables 1 and 2.

	PacBio	**ONT**
# reads	3,371,331	73,181,790
# spliced reads	1,266,498	28,961,005
# reads with poly(A) tails	3,321,897	38,986,771
# of unique barcodes detected	3,024	3,023
# reads with detected barcodes	2,873,455	12,153,599
# common molecules	54,752	

Supplemental Table S1. Read statistics on the entire Sl-ISO-Seq dataset. For barcode detection, we used a list of 3,024 16mer barcodes obtained from 283,600,728 10X reads.

	PacBio	ONT
# reads	38,290,265	29,212,434
# spliced reads	18,223,425	9,799,668
# reads with poly(A) tails	23,020,473	7,830,696
# of unique barcodes detected	6,921	6,921
# reads with detected barcodes	15,969,422	1,741,924
# common molecules	274,287	

Supplemental Table S2. Read statistics on the entire sc-ISOr-Seq dataset. For barcode detection, we used a list of 6,922 16mer barcodes obtained from 357,009,685 10X reads.

	minimap2	GraphMap2	deSALT	uLTRA
Total running time (sec)	433	7,331	500	2,942
# RT read pairs	54,752	52,927	57,040	54,616
Median alignment length (PacBio)	888	870	892	875
Median alignment length (ONT)	877	856	882	871
Median difference (PacBio read is longer)	14	15	14	16
Median difference (ONT read is longer)	8	9	8	15

Supplemental Table S3. Different aligners statistics on the SI-ISO-Seq dataset. An RT read pair is a pair where ONT and PacBio read have the same barcode and UMI and align to the same gene. The median difference between aligned lengths of PacBio read and ONT read from the RT read pair was calculated separately for cases when PacBio (ONT) alignment was longer.

Supplemental Note “Benchmarking of the read-to-isoform assignment algorithm”

To assess the developed read-to-isoform assignment algorithm, we ran the algorithm on reference transcripts and simulated reads aligned to the reference genome using minimap2 (Li, 2018) in spliced mode. We further computed precision and recall values based on the number of sequences/reads that were uniquely assigned to its original isoform without any splicing modifications. Since in the paper we focused on the analysis of RT read pairs where PacBio and ONT reads are mapped to the same gene, here we only considered reads that are correctly mapped to their genes of origin.

For this experiment we used reference transcript sequences from the Human GENCODE v36 and Mouse GENCODE v27 basic annotation, and PacBio CCS and ONT simulated data. PacBio CCS reads were simulated using IsoSeqSim (https://github.com/yunhaowang/IsoSeqSim) with an average error rate 1.4%. ONT reads were simulated using the Trans-NanoSim (Hafezqorani et al., 2020) tool with several modifications (see details in “Simulating ONT data” paragraph in Methods). Both datasets were simulated with uniform coverage across all transcripts in the annotation to ensure that the results are not biased due to highly-expressed transcripts.

As Supplemental Table 4 indicates, our assignment method yields almost perfect precision for reference transcripts and PacBio data and fairly high precision for ONT data.

	Human transcripts	Mouse transcripts	PacBio simulated	ONT simulated
Precision, %	99.7	99.6	99.5	97.5
Recall, %	97.8	97.6	97.5	75.5

Supplemental Table S4. Benchmarking of read assignment algorithm. Precision and recall of the read-to-isoform assignment algorithm on reference transcript sequences and simulated PacBio and ONT reads. Note, that precision and recall were computed only for the sequences mapped to its original gene to exclude effect of incorrect alignments.

Supplemental Note “Novel isoform discovery with different fractions of unknown transcripts”

To evaluate transcript model construction we mimic real-life situations by removing a part of the annotation (Methods). While it is not clear how many unknown transcripts can be present in a real sample, various researchers suggested values ~15-20% for a normal sample (suggestions were given informally during oral discussions). For this work we decided to select a rather arbitrary reasonable fraction of 20% to demonstrate the ability of StringTie2 to discover unknown isoforms.

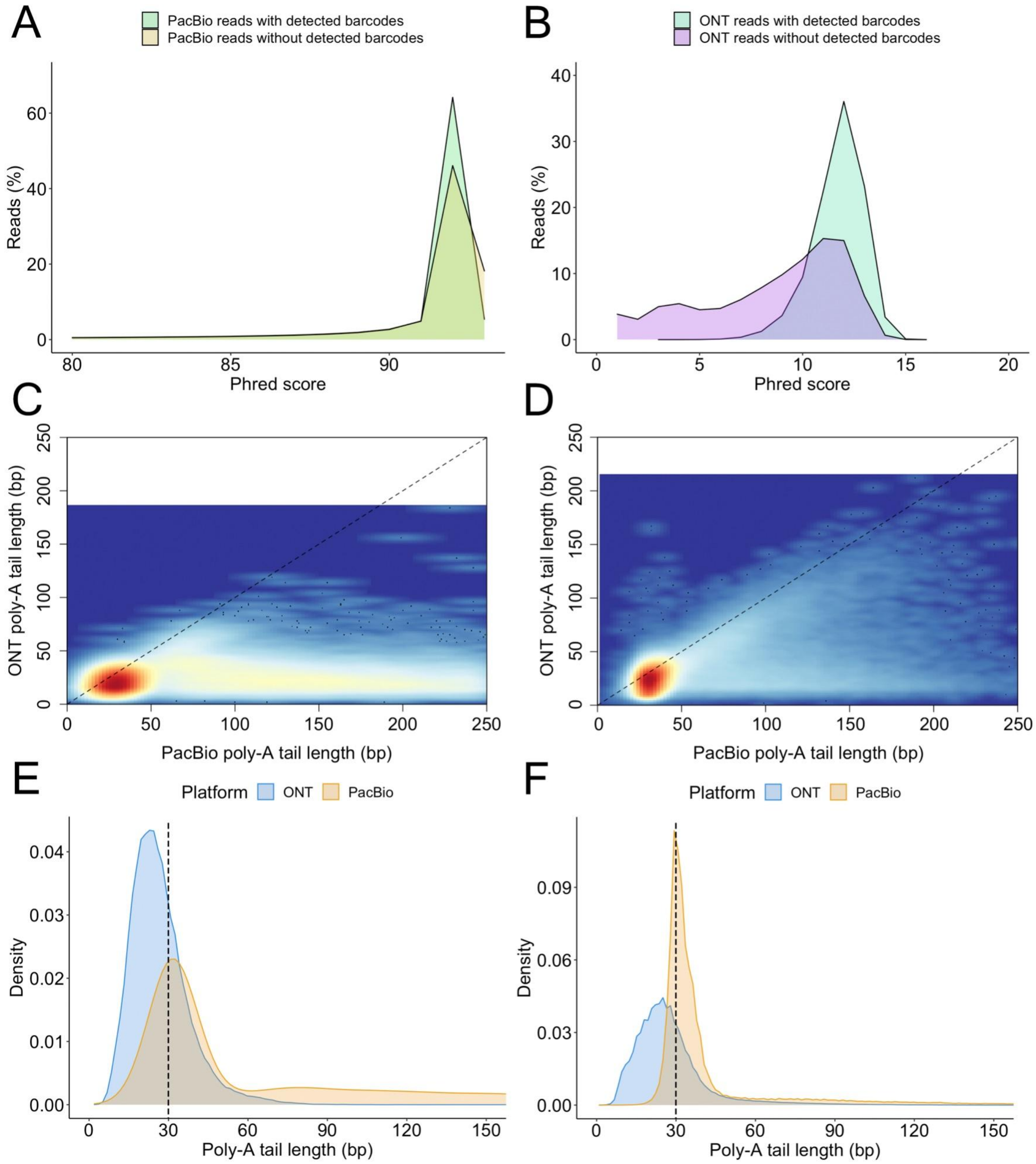
To inspect how this parameter affects the results we decided to run StringTie2 on the similarly obtained simulated human data using the Human GENCODE v36 annotation with different fractions of excluded expressed isoforms (10%, 15%, 20% and 25%). As the Supplemental Table 5 shows, overall precision and recall expectedly decrease for the larger fraction of excluded isoforms. However, for novel transcripts, recall remains virtually the same, while precision noticeably grows as more expressed isoforms are hidden from the annotation. While such behavior is not entirely obvious, it seems that novel isoforms are reported more accurately when the larger portion of reads belong to unannotated transcripts.

Fraction of excluded isoforms	All transcripts		Novel transcripts	
	Recall, %	Precision, %	Recall, %	Precision, %
10%	83.7	73.7	61.3	17.1
15%	81.9	72.6	59.9	24.1
20%	80.4	71.8	59.9	29.8
25%	78.8	70.7	59.9	34.2

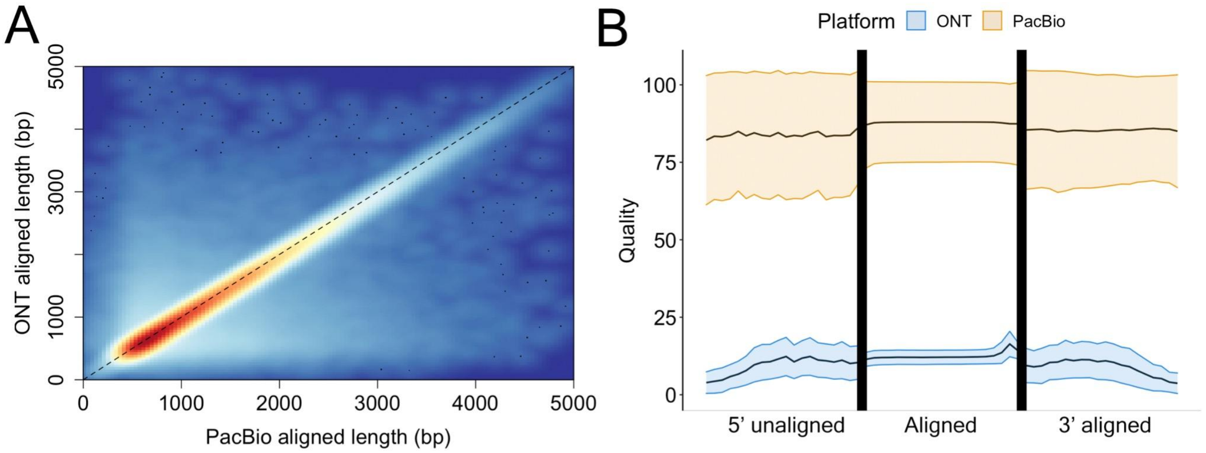
Supplemental Table S5. Benchmarking of StringTie2 on the datasets with different numbers of excluded isoforms. Precision and recall of the StringTie2 results on the simulated human data with 10%, 15%, 20% and 25% of isoforms removed from the annotation. Removed transcripts are considered novel, while expressed transcripts kept in the annotation represent the set of known models.

delta (bp)	PacBio reads agreeing with the annotation (%)	ONT reads agreeing with the annotation (%)
0	93.3	83.1
6	94.3	90.2
10	94.6	91.9

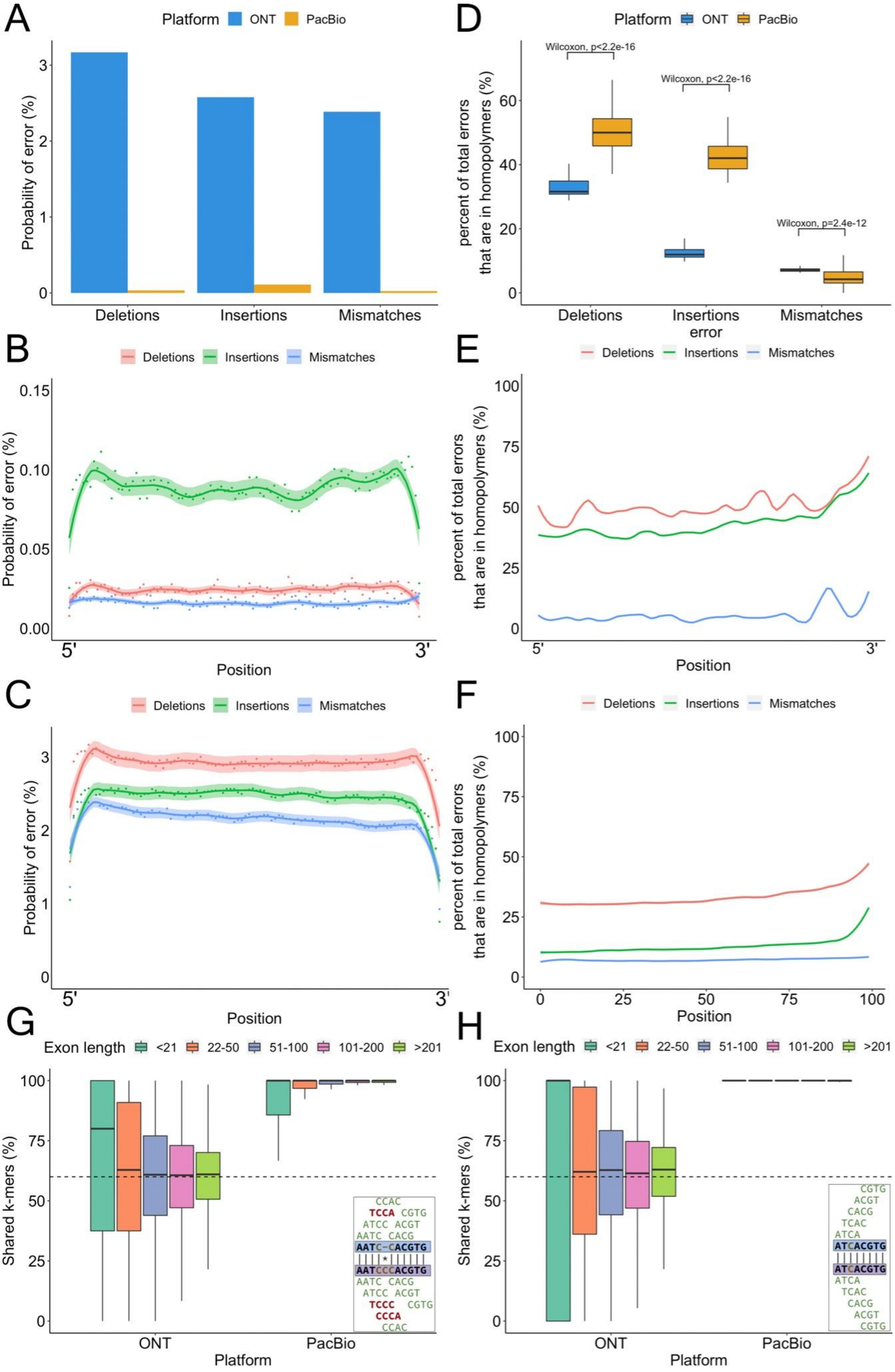
Supplemental Table S6. Agreement in RT read pairs of Sl-ISO-Seq data using different delta parameters. Numbers of PacBio and ONT reads agreeing with the annotation using different values of delta. Two introns are considered equal if their splice site coordinates differ by not more than delta.



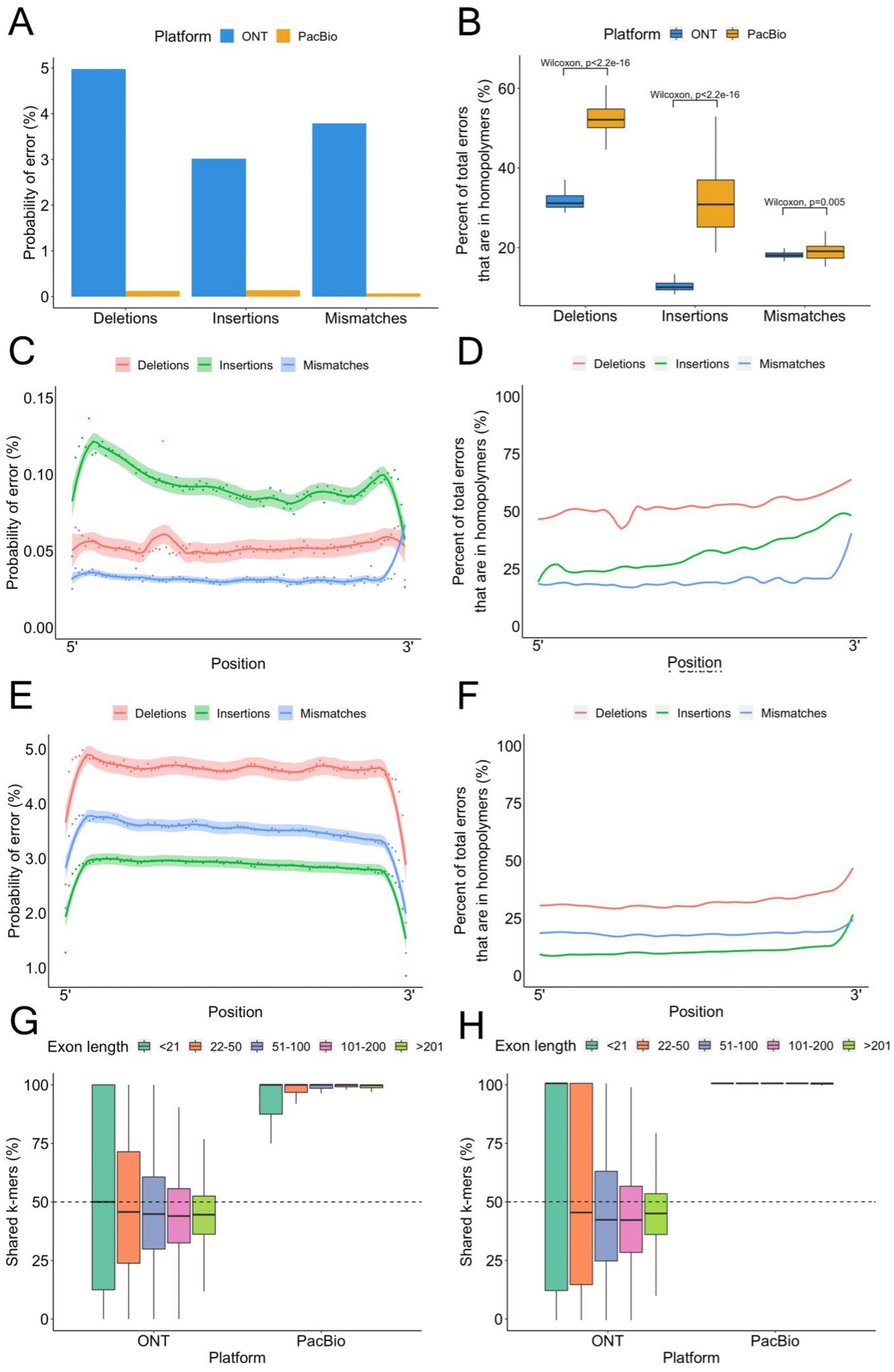
Supplemental Fig. S1. Primary read characteristics of ScISOr-Seq and Sl-ISO-Seq data. *(A)* Phred score distribution for Oxford Nanopore reads from ScISOr-Seq dataset with (light blue) and without (purple) detected barcodes. *(B)* Phred score distribution for PacBio CCS reads from ScISOr-Seq dataset with (light green) and without (yellow) detected barcodes. *(C)* Heatscatter plot showing the distribution of RT read pairs according to the length of a poly(A) tail detected in the respective PacBio read (X axis) and ONT read (Y axis) for Sl-ISO-Seq dataset. *(D)* Heatscatter plot showing the distribution of RT read pairs according to the length of a poly(A) tail detected in the respective PacBio read (X axis) and ONT read (Y axis) for ScISOr-Seq dataset. *(E)* Density distribution of poly(A) tail lengths in PacBio (yellow) and ONT (blue) reads for Sl-ISO-Seq dataset. The dashed line represents the expected length of poly(A) tails (30 bp). *(F)* Density distribution of poly(A) tail lengths in PacBio (yellow) and ONT (blue) reads for ScISOr-Seq dataset. The dashed line represents the expected length of poly(A) tails (30 bp).



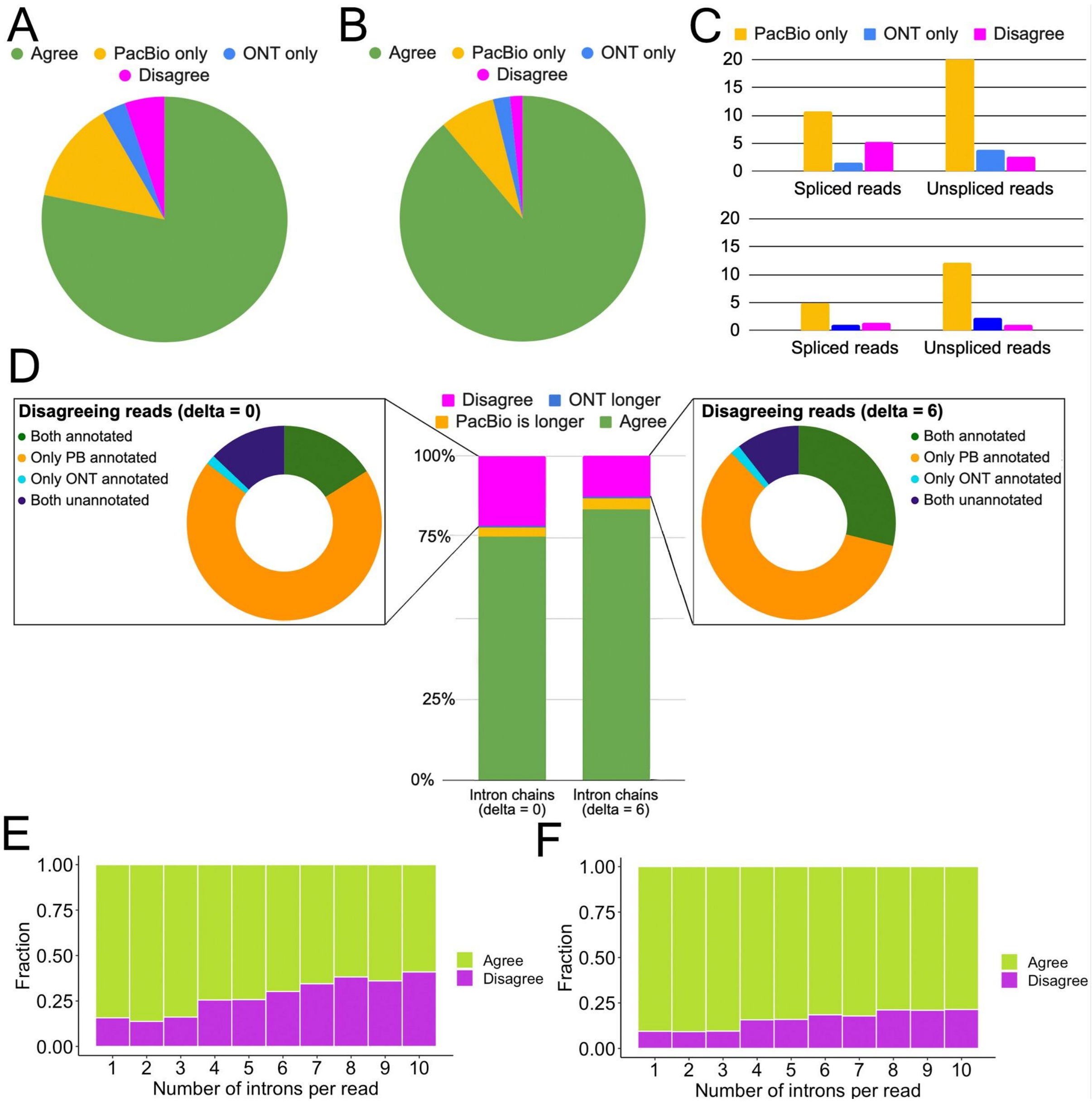
Supplemental Fig. S2. Alignment characteristics of RT read pairs of ScISOr-Seq data. *(A)* Heatscatter plot showing aligned lengths of respective PacBio read (X axis) and ONT read (Y axis) from the RT read pair after mapping to the genome using minimap2 for ScISOr-Seq dataset (Spearman's Rho: 0.90, $p<2.2*10^{-16}$). *(B)* Mean phred score distribution along the reads for aligned (middle) and unaligned (left and right) parts of PacBio (yellow) and ONT (blue) reads based on a (reference-free) pairwise Smith-Waterman alignment of the PacBio and Nanopore reads from the RT read pair for ScISOr-Seq dataset. Lower and upper bounds represent the standard deviation of the Phred score distribution.


A
Platform
ONT
PacBio
Probability of error (%)
0
1
2
3
Deletions
Insertions
Mismatches
D
Platform
ONT
PacBio
Wilcoxon, p<2.2e-16
Wilcoxon, p<2.2e-16
Wilcoxon, p=2.4e-12
percent of total errors
that are in homopolymers (%)
0
20
40
60
Deletions
Insertions
Mismatches
error
B
Deletions
Insertions
Mismatches
Probability of error (%)
0.00
0.05
0.10
0.15
5'
3'
Position
E
Deletions
Insertions
Mismatches
percent of total errors
that are in homopolymers (%)
0
25
50
75
100
5'
3'
Position
C
Deletions
Insertions
Mismatches
Probability of error (%)
0
1
2
3
5'
3'
Position
F
Deletions
Insertions
Mismatches
percent of total errors
that are in homopolymers (%)
0
25
50
75
100
0
25
50
75
100
Position
G
Exon length
<21
22-50
51-100
101-200
>201
Shared k-mers (%)
0
25
50
75
100
CCAC
TCCA CGTG
ATCC ACGT
AATC CACG
AATC-CACGTG
||||*||||||
AATCCCACGTG
AATC CACG
ATCC ACGT
TCCC CGTG
CCCA
CCAC
ONT
PacBio
Platform
H
Exon length
<21
22-50
51-100
101-200
>201
Shared k-mers (%)
0
25
50
75
100
CGTG
ACGT
CACG
TCAC
ATCA
ATCACGTG
||||||||
ATCACGTG
ATCA
TCAC
CACG
ACGT
CGTG
ONT
PacBio
Platform

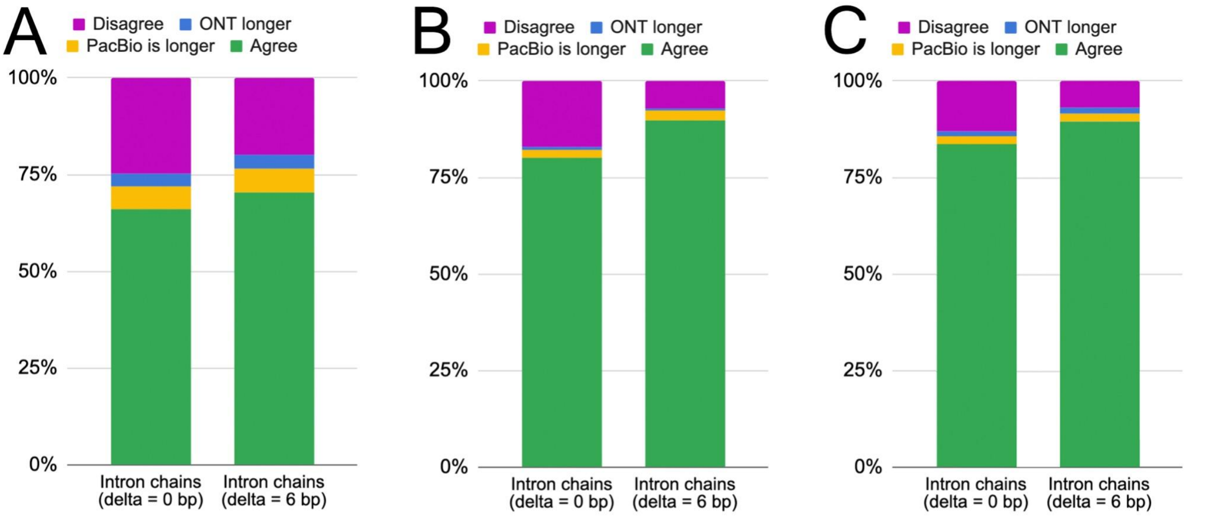
Supplemental Fig. S3. Sequencing error rates and *k*-mer identity for PacBio and Nanopore from Sl-ISO-Seq dataset. *(A)* Frequency of each sequencing error type in PacBio (yellow) and ONT reads (blue). *(B)* Sequencing error probability with respect to the position in read for PacBio data. *(C)* Sequencing error probability with respect to the position in read for ONT data. *(D)* Percent of sequencing errors located in homopolymers for each error type in PacBio (yellow) and ONT reads (blue). *(E)* Fraction of sequencing errors located in homopolymers with respect to position in the read for PacBio data. *(F)* Fraction of sequencing errors located in homopolymers with respect to position in the read for ONT data. (*G*) *k*-mer identity (k=14) of separate exons for ONT (left) and PacBio reads (right) with respect to the reference exon length. *Bottom right:* An example of two sequences and their shared (green) and distinct *k*-mers (red) before homopolymer compression. (*H*) *k*-mer identity (k=14) of separate exons after homopolymer compression for ONT (left) and PacBio reads (right) with respect to the reference exon length. *Bottom right:* An example of the same two sequences from Fig. 3G and their shared (green) and distinct *k*-mers (red) after homopolymer compression.


A
Platform ONT PacBio
Probability of error (%)
0 1 2 3 4 5
Deletions Insertions Mismatches
B
Platform ONT PacBio
Percent of total errors that are in homopolymers (%)
20 40 60
Wilcoxon, p<2.2e-16
Wilcoxon, p<2.2e-16
Wilcoxon, p=0.005
Deletions Insertions Mismatches
C
Deletions Insertions Mismatches
Probability of error (%)
0.00 0.05 0.10 0.15
5' 3'
Position
D
Deletions Insertions Mismatches
Percent of total errors that are in homopolymers (%)
0 25 50 75 100
5' 3'
Position
E
Deletions Insertions Mismatches
Probability of error (%)
1.0 2.0 3.0 4.0 5.0
5' 3'
Position
F
Deletions Insertions Mismatches
Percent of total errors that are in homopolymers (%)
0 25 50 75 100
5' 3'
Position
G
Exon length <21 22-50 51-100 101-200 >201
Shared k-mers (%)
0 25 50 75 100
ONT PacBio
Platform
H
Exon length <21 22-50 51-100 101-200 >201
Shared k-mers (%)
0 25 50 75 100
ONT PacBio
Platform

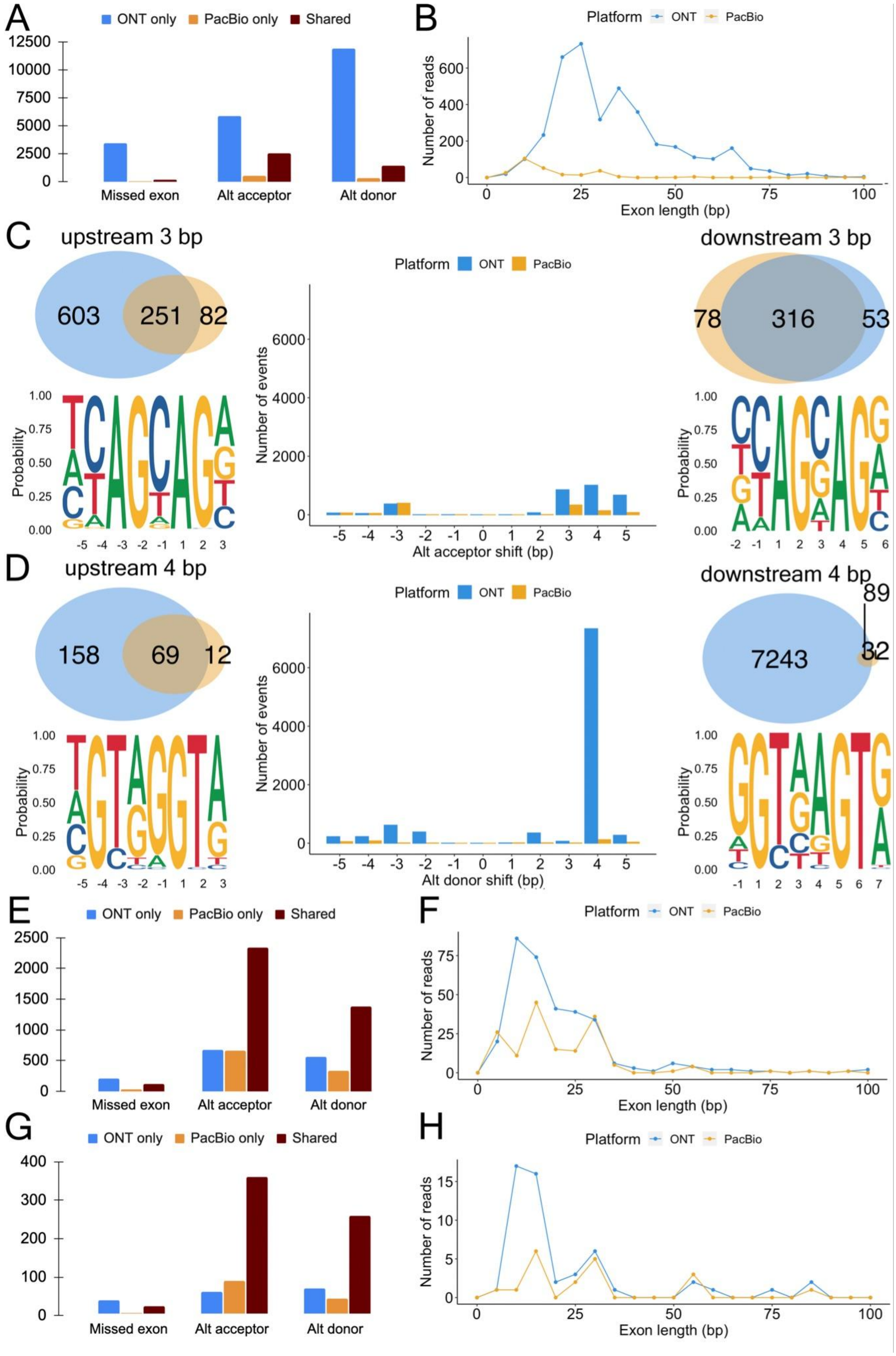
Supplemental Fig. S4. Sequencing error rates and *k*-mer identity for PacBio and Nanopore from ScISOr-Seq dataset. *(A)* Frequency of each sequencing error type in PacBio (yellow) and ONT reads (blue). *(B)* Percent of sequencing errors located in homopolymers for each error type in PacBio (yellow) and ONT reads (blue). *(C)* Sequencing error probability with respect to the position in read for PacBio data. *(D)* Fraction of sequencing errors located in homopolymers with respect to position in the read for PacBio data. *(E)* Sequencing error probability with respect to the position in read for ONT data. *(F)* Fraction of sequencing errors located in homopolymers with respect to position in the read for ONT data. (*G*) *k*-mer identity (k=14) of separate exons for ONT (left) and PacBio reads (right) with respect to the reference exon length. (*H*) *k*-mer identity (k=14) of separate exons after homopolymer compression for ONT (left) and PacBio reads (right) with respect to the reference exon length.



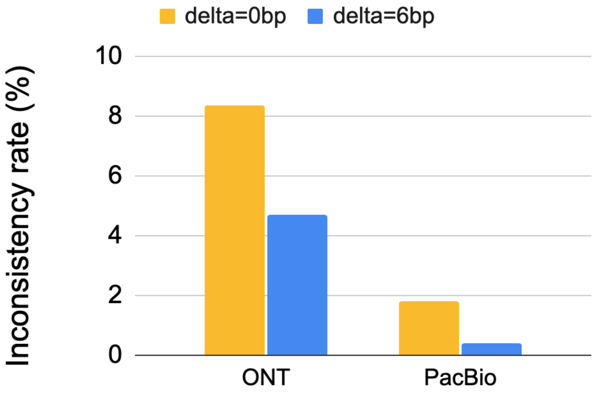
Supplemental Fig. S5. Agreement in RT read pairs of ScISOr-Seq data. *(A)* Fractions of TSS assignments that agree (green), disagree (magenta), or found only in one read (blue for ONT, yellow for PacBio) from the RT read pair. *(B)* Same as Fig. S5A, but for poly(A) sites. *(C)* Percentage of RT read pairs that disagree on the assigned TSS (top) and poly(A) site (bottom): only PacBio read was assigned (yellow), only ONT (blue), both assigned but to different sites (magenta). *(D) Middle*: fraction of intron chains from the RT read pairs that agree (green), disagree (magenta), or one chain being longer (blue for ONT, yellow for PacBio) when splice junctions are compared precisely (left, delta=0bp) or inexactly (right, delta=6bp). *Top left:* classification of disagreeing intron chains from RT read pairs with respect to the reference annotation (delta=0bp): both are inconsistent with the annotation (dark blue), both correspond to known (different) transcripts despite the disagreement (green), PacBio is consistent with the annotation while ONT chain is not (yellow) and vise versa (light blue). *Top right:* classification of disagreeing intron chains from RT read pairs with respect to the reference annotation using inexact comparison (delta=6bp). *(E)* Fraction of agreeing (green) and disagreeing (magenta) intron chains with respect to intron chain length when compared precisely (delta=0bp). *(F)* Fraction of agreeing (green) and disagreeing (magenta) intron chains with respect to intron chain length when compared with delta=6bp.



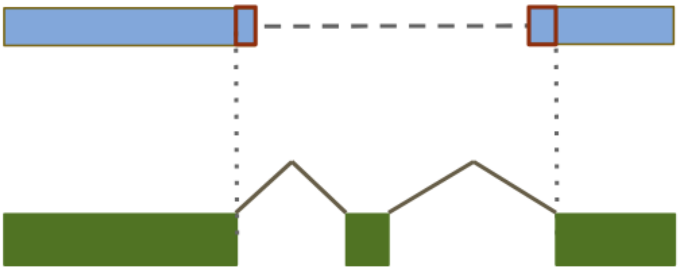
Supplemental Fig. S6. Agreement in RT read pairs of Sl-ISO-Seq data using alignments produced by alternative aligners. The corresponding plot for minimap2' alignments is shown in Fig. 3D, where in 81.8% (89%) of RT read pairs the PacBio and the ONT read had identical intron chains with delta=0bp (6bp). *(A)* Fraction of intron chains from the RT read pairs that agree (green), disagree (magenta), or one chain being longer (blue for ONT, yellow for PacBio) when splice junctions are compared precisely (left, delta=0bp) or inexactly (right, delta=6bp). Read alignments were produced by GraphMap2. In 66.1% (69.2%) of RT read pairs the PacBio and the ONT read had identical intron chains with delta=0bp (6bp). *(B)* Same as Fig. S6A, but for alignments produced by deSALT. In 79.8% (89.8%) of RT read pairs the PacBio and the ONT read had identical intron chains with delta=0bp (6bp). *(C)* Same as Fig. S6A, but for alignments produced by uLTRA. In 83.7% (89.5%) of RT read pairs the PacBio and the ONT read had identical intron chains with delta=0bp (6bp).


A
ONT only
PacBio only
Shared
12500
10000
7500
5000
2500
0
Missed exon
Alt acceptor
Alt donor
B
Platform
ONT
PacBio
Number of reads
600
400
200
0
0
25
50
75
100
Exon length (bp)
C
upstream 3 bp
603
251
82
Probability
1.00
0.75
0.50
0.25
0.00
-5 -4 -3 -2 -1 1 2 3
Platform
ONT
PacBio
Number of events
6000
4000
2000
0
-5 -4 -3 -2 -1 0 1 2 3 4 5
Alt acceptor shift (bp)
downstream 3 bp
78
316
53
Probability
1.00
0.75
0.50
0.25
0.00
-2 -1 1 2 3 4 5 6
D
upstream 4 bp
158
69
12
Probability
1.00
0.75
0.50
0.25
0.00
-5 -4 -3 -2 -1 1 2 3
Platform
ONT
PacBio
Number of events
6000
4000
2000
0
-5 -4 -3 -2 -1 0 1 2 3 4 5
Alt donor shift (bp)
downstream 4 bp
89
7243
32
Probability
1.00
0.75
0.50
0.25
0.00
-1 1 2 3 4 5 6 7
E
ONT only
PacBio only
Shared
2500
2000
1500
1000
500
0
Missed exon
Alt acceptor
Alt donor
F
Platform
ONT
PacBio
Number of reads
75
50
25
0
0
25
50
75
100
Exon length (bp)
G
ONT only
PacBio only
Shared
400
300
200
100
0
Missed exon
Alt acceptor
Alt donor
H
Platform
ONT
PacBio
Number of reads
15
10
5
0
0
25
50
75
100
Exon length (bp)

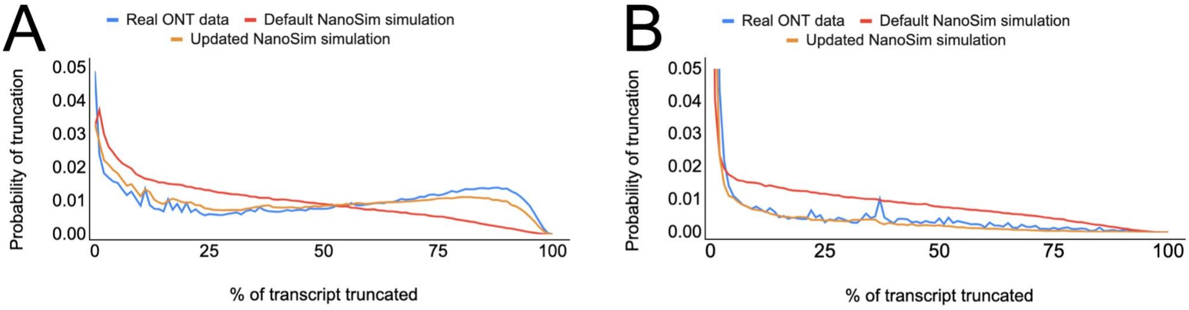
Supplemental Fig. S7. Exon and splice site characteristics underlying disagreements between PacBio and Nanopore in ScISOr-Seq data (A-F) and Sl-ISO-Seq data (G-H). *(A)* Number of missed exons (left), alternative acceptors (middle), and donors (right) with respect to the reference annotation that occur only in ONT read (blue), only in PacBio read (yellow), and in both reads from RT read pair (brown). *(B)* Length distribution for skipped exons in PacBio reads (yellow) and ONT reads (blue). *(C) Middle:* number of alternative acceptor sites in PacBio reads (yellow) and ONT reads (blue) with respect to distance from the annotated acceptor site. *Top left:* Venn diagram for 3bp upstream alternative acceptor sites in PacBio (yellow) and ONT reads (blue) from RT read pair. *Bottom left:* Nucleotide frequency for loci where 3bp upstream acceptor sites occur. *Top right:* Venn diagram for 3bp downstream alternative acceptor sites in PacBio (yellow) and ONT reads (blue) from RT read pair. *Bottom right:* Nucleotide frequency for loci where 3bp downstream acceptor sites occur. *(D) Middle:* number of alternative donor sites in PacBio reads (yellow) and ONT reads (blue) with respect to distance from the annotated donor site. *Top left:* Venn diagram for 4bp upstream alternative donor sites in PacBio (yellow) and ONT reads (blue) from RT read pair. *Bottom left:* Nucleotide frequency for loci where 4bp upstream donor sites occur. *Top right:* Venn diagram for 4bp downstream alternative donor sites in PacBio (yellow) and ONT reads (blue) from RT read pair. *Bottom right:* Nucleotide frequency for loci where 4bp downstream donor sites occur. *(E)* Number of missed exons (left), alternative acceptors (middle), and donors (right) with respect to the reference annotation that occur only in ONT read (blue), only in PacBio read (yellow), and in both reads from RT read pair (brown) in ScISOr-Seq data, the genome annotation was used during the alignment step. *(F)* Length distribution for skipped exons in PacBio reads (yellow) and ONT reads (blue) in ScISOr-Seq data, the genome annotation was used during the alignment step. (*G*) Number of missed exons (left), alternative acceptors (middle), and donors (right) with respect to the reference annotation that occur only in ONT read (blue), only in PacBio read (yellow), and in both reads from RT read pair (brown) in Sl-ISO-Seq data, the genome annotation was used during the alignment step. (*H*) Length distribution for skipped exons in PacBio reads (yellow) and ONT reads (blue) in Sl-ISO-Seq data, the genome annotation was used during the alignment step.



Supplemental Fig. S8. Inconsistency rate of PCR duplicates in PacBio and Nanopore sequencing (Sl-ISO-Seq data). Only PCR duplicated read pairs for which both reads map to the same gene were considered. Inconsistency rate was calculated as a percentage of PCR duplicated read pairs for which PCR duplicates were mapped to the same gene but disagreed on intron chains when splice junctions are compared precisely (yellow, delta=0bp) or inexactly (blue, delta=6bp). Total number of PacBio (ONT) pairs: 14,632 (187,279). In ONT (left), in 91.7% (95.3%) of read pairs PCR duplicates had identical intron chains with delta=0bp (6bp). In PacBio (right), in 98.2% (99.6%) of PacBio read pairs PCR duplicates had identical intron chains with delta=0bp (6bp).



Supplemental Fig. S9. Misaligned short exon. An ONT read alignment (blue) misses a short reference exon (green) by extending an alignment of the neighboring exons (red).



Supplemental Fig. S10. Truncation probabilities for real and simulated ONT data. *(A)* Probability of an ONT read being truncated by X% on the 5' end for real Sl-ISO-seq data (red), simulated using NanoSim with default settings (blue), and simulated using NanoSim with improved truncation procedure (orange). Truncated fractions were estimated by mapping reads onto reference transcriptome using minimap2 with -x map-ont option. *(B)* Same as S9A but for the 3' end.