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

Alla Mikheenko[1,*], Andrey D Prjibelski[1,*], Anoushka Joglekar[2], and Hagen U Tilgner[2,+]

Author Affiliations

1. Center for Algorithmic Biotechnology, Institute of Translational Biomedicine, St. Petersburg State University, St. Petersburg, Russia.
2. Brain and Mind Research Institute and Center for Neurogenetics, Weill Cornell Medicine, New York, New York, USA.

* Contributed equally

\+ Corresponding author: hut2006@med.cornell.edu

Running title: Comparing ONT and PacBio with single-molecule RNA-seq

Abstract

Long-read transcriptomics requires understanding error sources inherent to technologies. Current approaches cannot compare methods for an individual RNA molecule. Here, we present a novel platform-comparison method that combines barcoding strategies and long-read sequencing to sequence cDNA copies representing an individual RNA molecule on both Pacific Biosciences ("PacBio) and Oxford Nanopore ("ONT"). We compare these long-read pairs in terms of sequence content and isoform patterns. Although individual read pairs show high similarity, we find differences in (i) aligned length, (ii) transcription start site (TSS) and (iii) polyadenylation site (poly(A)-site) assignment, and (iv) exon-intron structures. Overall 25% of read pairs disagree on either TSS, poly(A)-site, or a splice site. Intron-chain disagreement typically arises from alignment errors of microexons and complicated splice sites. Our single-molecule technology comparison reveals that inconsistencies are often caused by sequencing-error induced inaccurate ONT alignments, especially to downstream GUNNGU donor motifs. However, annotation-disagreeing upstream shifts in NAGNAG acceptors in ONT are often confirmed by PacBio and are thus likely real. In both barcoded and non-barcoded ONT reads, we find that number of introns and proximity of GU/AGs better predict inconsistencies with the annotation than read quality alone. We

summarize these findings in an annotation-based algorithm for spliced alignment correction that improves subsequent transcript construction with ONT reads.

Introduction

Long-read sequencing is being increasingly used in transcriptomics, particularly for barcoded unique molecules (Gupta et al. 2018; Singh et al. 2019), which yields single-cell and spatially resolved long-read transcriptomes. Various platforms such as Pacific Biosciences (Eid et al. 2009; Koren et al. 2012; Sharon et al. 2013; Au et al. 2013; Tilgner et al. 2014; Weirather et al. 2015), Oxford Nanopore (Oikonomopoulos et al. 2016; Byrne et al. 2017), as well as linked-read technologies. Linked-read technologies for RNA were originally represented either by synthetic long reads (SLRs) (Tilgner et al. 2015) or usually more sparsely covered 10x Genomics linked-reads (Tilgner et al. 2018), although more recently other linked-read technologies have emerged (Wu et al. 2019; Chen et al. 2020). Furthermore, for all these platforms a variety of protocols either exists or can be imagined. Comparing the accuracy of these distinct approaches is therefore fundamental in modern transcriptomics just as it has been fundamental for short-read sequence analysis (Engström et al. 2013; Steijger et al. 2013; Li et al. 2014b, 2014a).

A drawback of commonly used strategies is their lack of single-molecule resolution. For example, percent-spliced-in (PSI)-values (Wang et al. 2008) of splice sites, or transcript-per-million (TPM) (Wagner et al. 2012) values can easily be compared between multiple strategies. However, these approaches do not allow for the comparison of the accuracy of different strategies for a single molecule. Usually, platforms are compared by the estimated percentages of molecules that behave in a similar way. High concordance of such percentages theoretically suggest that both platforms would behave identically on an individual molecule. However, this theoretical suggestion has so far been impossible to verify because of the impossibility to assess whether two platforms sequence a representation of the same molecule. Our single-RNA-molecule reasoning provides a framework where an alignment is either correct

or false. This is not the case for groups of molecules, in which an alignment can be correct for one molecule and false for another molecule.

Single-cell and spatial barcoding of cDNAs have revolutionized the investigation of complex organs (Macosko et al. 2015; Zeisel et al. 2015). In most single-cell approaches, cDNAs are generated with a polydeoxythymidylic acid (poly(dT)) primer carrying added sequences. In single-cell barcoding, a portion of this added sequence (the "barcode", here 16 bases) is identical for all cDNAs from the same individual cell, but distinguishes one cell from others. Another portion (the unique molecular identifier, "UMI", here 12 bases) is random for each reverse transcription event and thus informs on whether two sequenced reads represent two distinct reverse transcription events or polymerase chain reaction (PCR) duplicates of only one reverse transcription event. Similarly, spatial approaches use barcodes and UMIs to distinguish spatial locations and reverse transcription events. Our advances in long-read sequencing of single-cell (Gupta et al. 2018; Joglekar et al. 2021; Hardwick et al. 2021) and spatially (Joglekar et al. 2021) barcoded cDNAs allow the identification of full-length isoforms for barcoded molecules.

Here, we aim to compare the Pacific Biosciences ("PacBio") and Oxford Nanopore ("ONT") platforms in terms of error profiles, spliced alignments, their discrepancies with the reference, and potential pitfalls in downstream analysis. For this purpose, we use barcoded cDNA copies corresponding to the same RNA molecule sequenced on both platforms.

Results

Identification of RT pairs sequenced on PacBio and ONT platforms. Here we compare cDNAs that are barcoded by their single cell (ScISOr-Seq) of origin or their spatial location (Sl-ISO-Seq), sequenced on both the PacBio Sequel II (circular consensus reads) and ONT systems (Hardwick et al. 2021; Joglekar et al. 2021). A reverse-transcription event is identified by the combination of (i) single-cell/spatial location barcode, (ii) a unique molecular identifier (UMI), and (iii) the gene the molecule is mapped to

(Fig. 1A). PCR copies of the same reverse-transcription event are sequenced on both platforms ("RT read pairs"), which enables their comparison for individual RNA molecules. To have the highest confidence barcodes and ONT-PacBio read pair correspondences, we only use a perfect matching strategy for barcode and UMI detection (Supplemental Note "Experimental details"). Thus, we compare sequencing errors and intron structure in identified RT read pairs.

Molecule identification discards lower quality ONT reads. We first compare Phred scores of all individual reads sequenced on both platforms independently of molecule detection. The Phred quality score of a base indicates the probability of a base being called correctly. The Phred quality score of a read is computed as the average across Phred scores of all bases in a read. PacBio shows much higher Phred scores than ONT, both in single-cell and in spatial data (Fig. 1B). Since we conservatively analyze only perfectly matched barcodes, ONT reads with barcodes have significantly higher Phred scores than those without (Fig. 1D, two-sided Wilcoxon-rank-sum p-value=2.4×10^{-9}), while no significant difference is detected for PacBio (Fig. 1C, two-sided Wilcoxon-rank-sum p-value=0.2). Similar observations are made on the single-cell dataset, although the difference for ONT reads is even more prominent (Supplemental Fig. S1A-B). Overall, barcode detection in spatial data yields 2,873,455 PacBio reads (of 3,371,331 - 85%) and 12,153,599 ONT reads (of 73,181,790 - 16%). Although ONT essentially has a deeper sequencing depth than PacBio, the difference between the number of usable reads is not so high when only considering reads with detected barcodes. It is likely that allowing for barcode mismatches could lead to more ONT reads being retained, although that comes at the risk of introducing more inaccurate barcodes.

Sequence comparison using reference-based and reference-free alignments. Our downstream analysis mostly relies on read mappings, so we first aligned reads using different tools: the widely used minimap2 (Li 2018) and specialized transcriptome aligners: deSALT (Liu et al. 2019), GraphMap2 (Marić et al.

2019), and uLTRA (Sahlin and Mäkinen 2021). Of note, we did not use STARlong (Dobin et al. 2013) since it has strong performance for PacBio reads, but is not optimized for error-prone ONT data.

Although all three aligners yield largely similar results, GraphMap2 and uLTRA produce slightly shorter alignments than deSALT and minimap2, and uLTRA generates alignments with more prominent differences between aligned lengths of PacBio and ONT reads (Supplemental Table S3). In addition, GraphMap2 produces the least number of RT read pairs because a PacBio read and an ONT read sharing the same barcode and UMI are mapped to different genes more frequently.

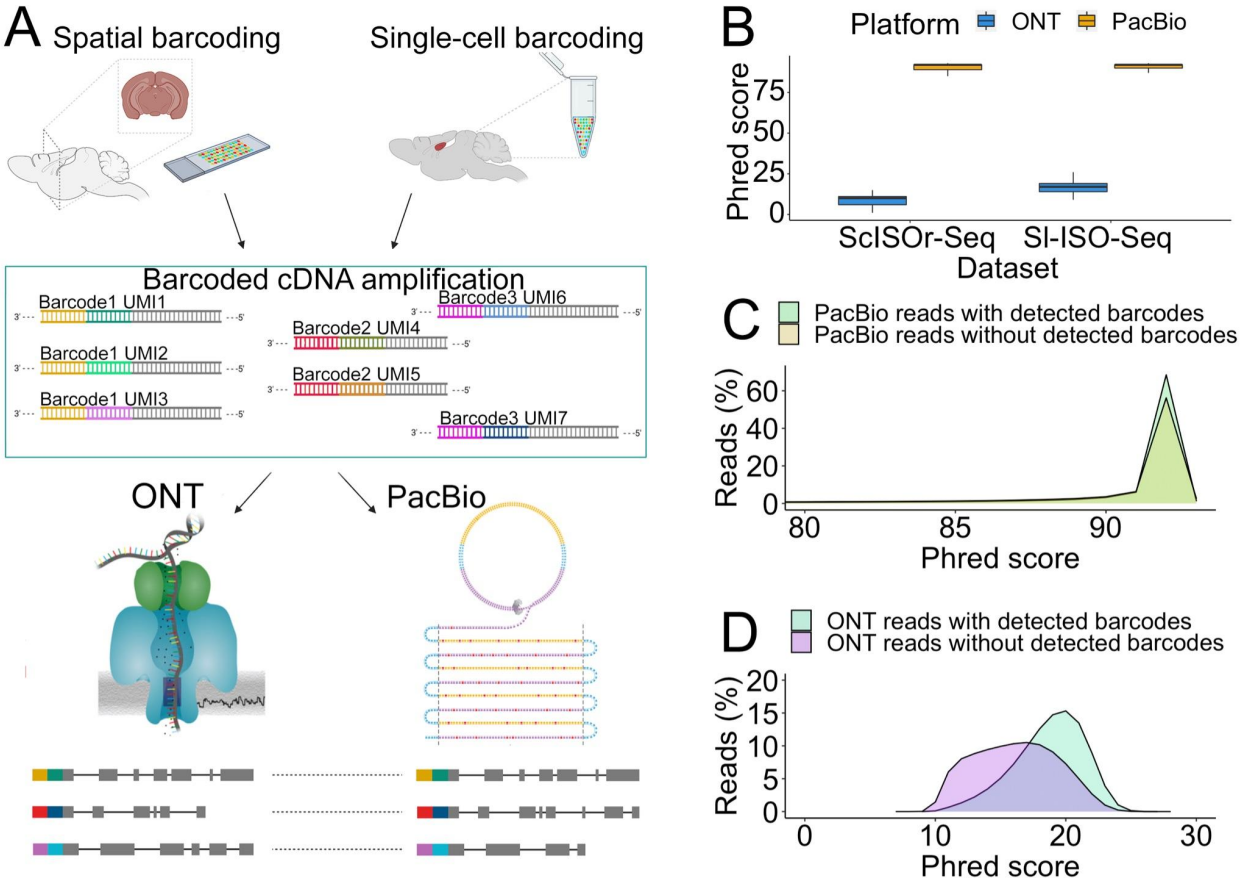

Fig. 1. Outline and primary read characteristics. (*A*) Individual reverse transcription events turn an individual RNA molecule into a barcoded RNA-cDNA hybrid, which is amplified into many cDNA molecules that carry the same barcode and UMI. We previously performed this process in two distinct ways - by single-cell 10xGenomics barcoding as well as by spatial 10xGenomics Visium barcoding. Aliquots of these cDNAs are then sequenced on PacBio and ONT. Using the identity of barcode and UMI, we can detect individual RNA molecules whose cDNA copies have been sequenced on both ONT and PacBio. We refer to these read pairs as RT read pairs. (*B*) Comparison

between Phred scores of PacBio CCS and ONT reads from both datasets. (*C*) Phred score distribution for PacBio CCS reads from the Sl-ISO-Seq dataset with (light green) and without (yellow) detected barcodes. (*D*) Phred score distribution for ONT reads from the Sl-ISO-Seq dataset with (light blue) and without (purple) detected barcodes.

Alignments generated by deSALT and especially Graphmap2 show more discordance between PacBio and ONT reads in terms of splice sites than alignments produced by uLTRA and minimap2 (Supplemental Fig. S6). Moreover, minimap2 performs significantly better in terms of isoform detection (Section "Splice site correction improves transcript discovery precision", Table 1). To enforce the same strategy for PacBio and ONT, here we use minimap2 for alignment.

Mapping RT read pairs (54,752 in the Sl-ISO-Seq and 274,287 in the ScISOr-Seq data) to the genome with minimap2 (Li 2018) showed highly correlated alignment lengths (Fig. 2A and Supplemental Fig. S2A). PacBio reads were significantly longer than their ONT counterparts in ScISOr-Seq and, although less pronounced, in Sl-ISO-Seq data, but differences were small compared to the entire read (Fig. 2B). However, in terms of absolute numbers, the difference is notable: 70% of read pairs have a longer aligned portion (median: 15bp) in the PacBio and 27% in the ONT read (median 9bp, Fig. 2C). In comparison to read length, the differences are minor. However, 9-15bp sequences can harbor important elements such as polyadenylation (poly(A)) signals, protein/miRNA-binding sites, microexons, or G-quadruplexes (Lee et al. 2020).

To delineate common and diverging sequences in reads from RT read pairs, we aligned them to each other using the Smith-Waterman algorithm. Of note, unlike modern mapping algorithms based on heuristics, Smith-Waterman is an exact solution and not subject to future improvements. We divided each alignment pair into an unaligned 5' part, an aligned portion, and an unaligned 3' part. In all three compartments within the spatial data, PacBio showed much higher read-wise Phred scores than ONT, but PacBio qualities slightly drop in the non-aligned 5' portion. However, for the single-cell data, PacBio qualities in

the 5’ unaligned portion remained constant and comparable to the aligned portion. In both datasets, ONT qualities deteriorate in the 5’ unaligned portion and gradually decrease in the unaligned 3’ portion (Fig. 2D and Supplemental Fig. S2B).

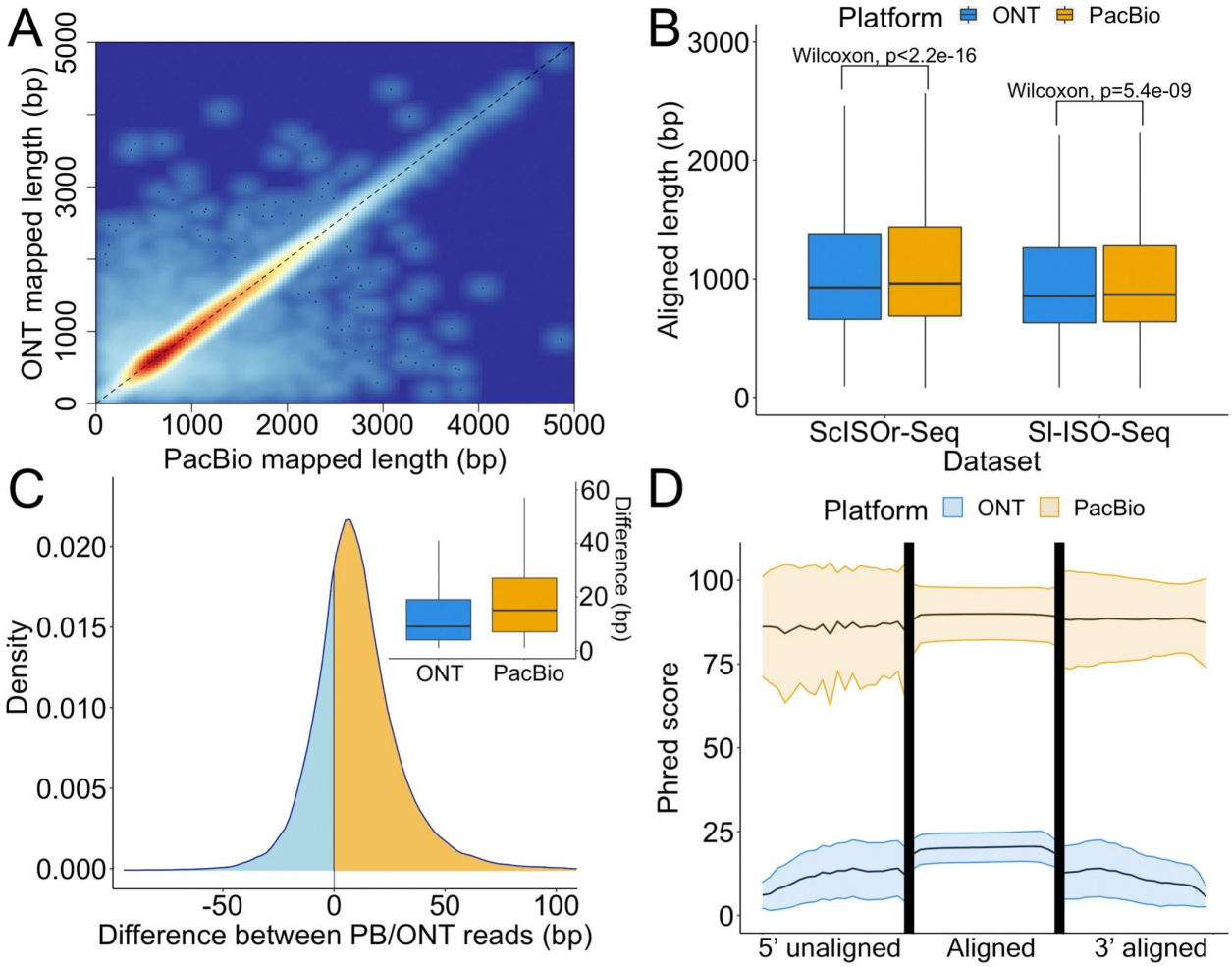

Fig. 2. Alignment characteristics of RT read pairs. (*A*) Heatscatter plot showing aligned lengths of respective PacBio read (X-axis) and ONT read (Y-axis) from the RT read pair after mapping to the genome using minimap2 (Sl-ISO-Seq dataset, Spearman’s Rho: 0.96, $p<2.2\times10^{-16}$). (*B*) Comparison between aligned lengths of PacBio and ONT reads for both datasets after mapping to the genome using minimap2. (*C*) Density plot showing the difference between aligned lengths of PacBio read and ONT read from the RT read pair after mapping to the genome using minimap2 (Sl-ISO-Seq dataset) and box plot showing distribution of the differences. In the density plot, cases when PacBio alignment is longer correspond to the yellow area under the curve; the opposite being represented by the blue area. (*D*) Mean Phred score distribution along the read 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 ONT reads from the RT read pair (Sl-ISO-Seq dataset). Lower and upper bounds represent the standard deviation of the Phred score distribution.

Sequencing error rates and *k*-mer identity in alignments to the genome. Deducing sequencing errors from alignments is difficult because in addition to sequencing and alignment errors, PCR errors, single nucleotide variants (SNVs), and mutations cause a divergence between reads and the genome. We used a three-way comparison between paired PacBio and ONT reads as well as the genome to define a "ground truth" using a majority call among all three sources, and to delineate error patterns as a divergence from this ground truth (Methods). ONT has more errors than PacBio (Supplemental Fig. S3A). For PacBio, deletions or mismatches are ~3-fold less abundant than insertions, whose frequency increases towards read ends (Supplemental Fig. S3B). ONT behaves very differently: deletions dominate over insertions and mismatches, and all error types decrease towards alignment ends (Supplemental Fig. S3C). 41% of PacBio errors and only 23% of ONT errors occur in homopolymers (Supplemental Fig. S3D). For PacBio, indels within homopolymers are more prevalent than mismatches, and slightly more errors occur in homopolymers towards the 3'end (Supplemental Fig. S3E). For ONT, similar trends were observed, although the insertions are less biased towards homopolymers than in PacBio (Supplemental Fig. S3F). In summary, PacBio has fewer errors than ONT, but a large fraction of PacBio errors occur in homopolymers, while ONT errors mostly arise from other areas. Similar observations were obtained for the older ScISOr-Seq dataset, although the overall error rate in ONT reads was higher (Supplemental Fig. S4A-D).

Alignments to a genome often employ seeding through matching *k*-mers. We considered 14-mer identity, a commonly used *k*-mer, for example in minimap2 (Li 2018), and analyzed each exon alignment separately. As expected, 14-mer identity was lower than single-base identity and specifically affected ONT reads (Supplemental Fig. S3G). For ONT data, we found lower 14-mer identities for slightly longer

exons (two-sided Wilcoxon-rank-sum test p=0.005, 1-20bp vs. 21-50bp exons). The most reasonable explanation is that short exons may become unmappable with few sequencing errors, which thus excludes them from this analysis and creates a bias for *k*-mer identity values. However, for both PacBio and ONT reads, the interquartile range of *k*-mer identities in short exons is higher compared to longer exons, as a single sequencing error may disrupt all *k*-mers and lead to 0% *k*-mer identity (Supplemental Fig. S3g). These effects are noticeably stronger for ONT alignments than for PacBio alignments. After homopolymer compression (Methods), 14-mer identity reached almost 100% for PacBio regardless of exon length. However, for ONT, compression caused high variability in short exons (<21bp): while the median increased to 100%, the 1st quartile decreased to 0%, since a single error in a short exon can affect all 14-mers (Supplemental Fig. S3H). Broadly similar observations were made for ScISOr-Seq data (Supplemental Fig. S4E-F). Thus, homopolymer compression should be applied to ONT reads with care.

Three-way comparison of annotation, ONT, and PacBio shows differences at exon and splice-site calling. Long-read experiments regularly uncover many isoforms that are inconsistent with annotations (Sharon et al. 2013; Au et al. 2013; Tilgner et al. 2014, 2015; Oikonomopoulos et al. 2016; Tardaguila et al. 2018; Tung et al. 2019; Kovaka et al. 2019; Tang et al. 2020; Wyman et al. 2020). While for short-read experiments, splice site identification and the following splice site quantification has been addressed with large success (Vaquero-Garcia et al. 2016), long-read based annotation-inconsistent isoforms can be truly novel or simply false. This question can only be conclusively answered for a single molecule as a correct alignment for one molecule can be false for another molecule. Here, we exploit RT read pairs to evaluate inconsistencies between alignments and the annotation using a "three-way comparison" where the ground truth is defined as a variant supported by the reference and at least one long-read technology for an RT read pair.

We considered 22,600 and 48,993 RT read pairs, where at least one of the PacBio/ONT reads is assigned to a known transcription start site (TSS) (Lizio et al. 2015) and poly(A) site (Herrmann et al. 2019)

respectively (Methods). Indeed, all barcoded reads have a poly(A) tail, which creates a bias towards 3'completeness in RT read pairs, and, thus, a known poly(A) site is assigned for a significantly larger portion of reads than a TSS. Moreover, in 95% of RT read pairs, both the PacBio and ONT read are assigned to the same annotated poly(A) site, while agreement on TSS is lower (87%, Fig. 3A-B). For TSS assignment, a significant portion of disagreeing pairs arises from unassigned ONT reads (8% of all RT read pairs), which suggests that 5' truncation of PacBio reads is less frequent. We noticed that the results differ for spliced and unspliced reads. While spliced reads are assigned to TSS more often, the percentage of RT read pairs with both reads spliced and assigned to different TSS is also higher (3.2% for spliced reads vs 1.7% for unspliced). This may potentially be explained by short starting exons misaligned in ONT reads. For poly(A) sites, however, the difference between spliced and unspliced reads is marginal (Fig. 3c).

We then considered for each read mapping the list of all its introns, which we refer to as an intron chain. In 81.8% of RT read pairs (of 28,330 total pairs in which both reads are spliced), the PacBio and the ONT read had identical intron chains. In 2% of cases, the intron chains were not contradictory, but the PacBio read had extra intron(s) at its extremities, while ONT reads had longer intron chains only in 0.6% of pairs. In the remaining 15.7%, PacBio and ONT intron chains disagreed with each other. When compared to the annotation, 17% of the disagreeing RT read pairs had both intron chains that corresponded to different known transcripts (green in Fig. 3D, top left), while in 15% of cases both were inconsistent with the annotation (dark blue in Fig. 3D, top left). In these cases, it is not easy to ascertain which mapping is true. However, in a large fraction (65%) of the disagreeing pairs, the PacBio mappings were consistent with the annotation, while ONT mappings were not (yellow in Fig. 3D, top left). In this case, assuming that the PacBio intron chain is in fact correct appears more parsimonious than the contrary.

The above statements are based on a single-base interpretation of PacBio and ONT splice sites (delta=0bp, Methods). To account for slight shifts in splice-site mapping, we explored inexact intron chain

comparison, in which junctions are considered equal if the distance between them does not exceed 6bp (delta=6bp). This reduced disagreements between paired PacBio and ONT reads by 48%. Among the remaining disagreements, in nearly half of the cases the PacBio mapping corresponded to an annotated transcript, while the ONT read did not (Fig. 3D middle and right). Overall, 43% of ONT and 15% of PacBio mappings inconsistent with the annotation at delta=0bp were reclassified as annotated with delta=6bp. Notably, further increasing delta to 10 bp affects only a small portion of reads: specifically, 78 PacBio and 468 ONT reads (Supplemental Table S6).

In addition, we compared intron chains of PCR duplicated read pairs sequenced on the same platforms for intra-molecular consistency (Supplemental Fig. S8). Indeed, when two reads that correspond to one original RNA molecule disagree in terms of alignment, only one can be correct. As expected, PacBio read pairs originating from PCR duplicates have significantly lower intron chain disagreement (1.8% with delta=0bp, 0.4% with delta=6bp) compared to ONT (8.3% with delta=0bp, 4.7% with delta=6bp), thus confirming the observations stated above.

We further hypothesized that the fraction of disagreeing RT read pairs would increase with the number of splice sites per read. Indeed, reads with ≥8 introns disagreed with its pair ~3-fold more often than reads with 2 introns. However, read pairs with ≥8 introns still agreed in 70% of cases (Fig. 3E). Using delta=6bp reduces disagreements but roughly preserves the trend (Fig. 3F). Broadly similar observations were also made for ScISOr-Seq data (Supplemental Fig. S5). These observations suggest that other factors beyond the intron chain length influence disagreements between PacBio and ONT reads. We therefore investigated sequence characteristics of disagreeing introns.

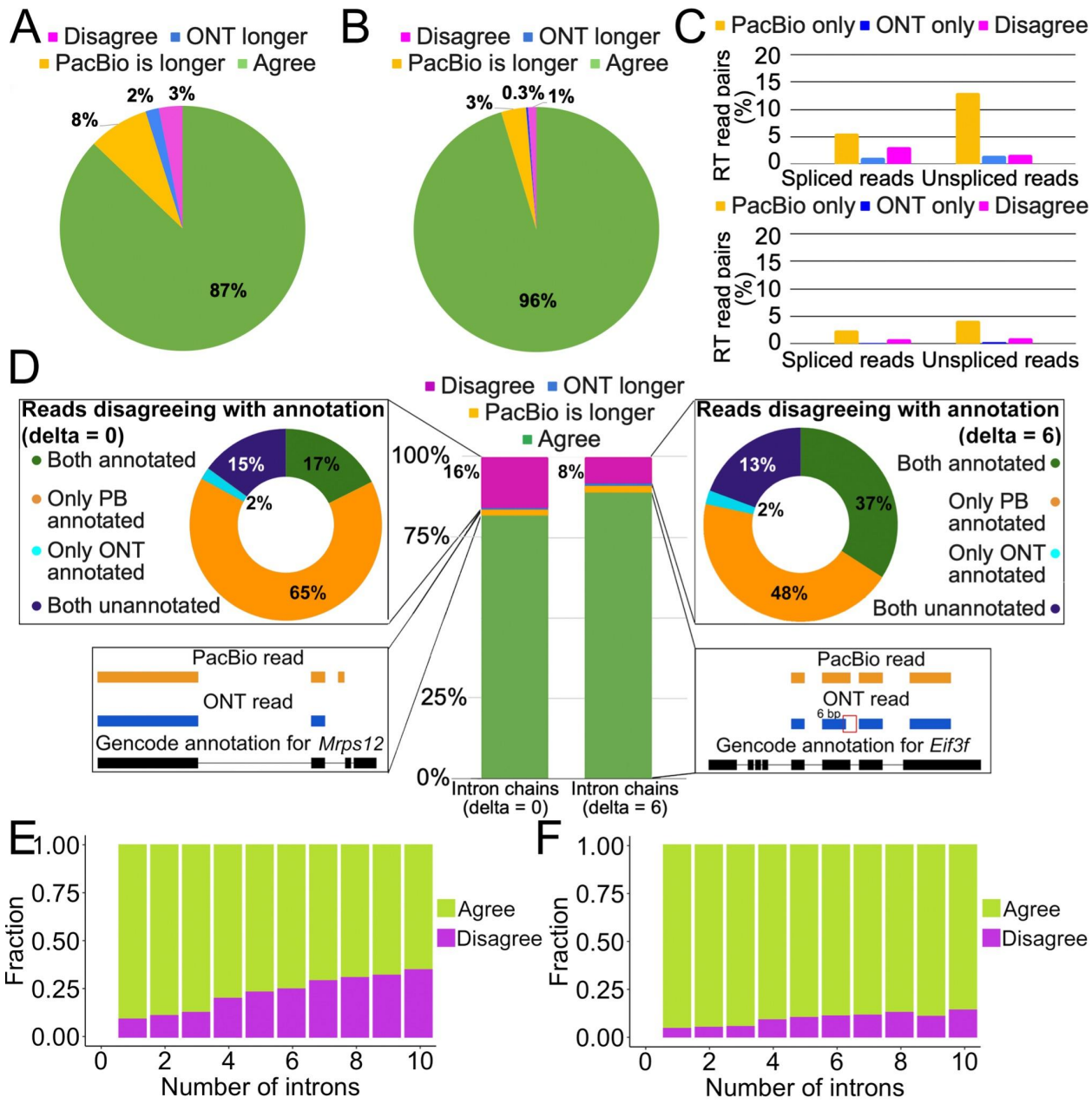

Fig. 3. Agreement in RT read pairs of Sl-ISO-Seq data. (*A*) Fractions of TSS assignments that agree (green), disagree (magenta), or are found only in one read (blue for ONT, yellow for PacBio) from the RT read pair. (*B*) Same as Fig. 3A, but for poly(A) sites. (*C*) Percentage of RT read pairs that disagree on the assigned TSS (top) and poly(A) site (bottom): only PacBio read assigned (yellow), only ONT (blue), both assigned but to different sites (magenta). (*D*) *Middle*: percentage of 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 is not (yellow) and vise versa (light blue). *Top right:* classification of disagreeing intron chains with respect to the reference annotation using inexact comparison (delta=6bp). *Bottom left:* an example of agreeing intron chains from an RT read pair in which PacBio intron chain is longer (*Mrps12* gene). *Bottom right:* An example of intron chains from an RT read pair that have a 6bp difference in the donor site of the second intron. Comparing intron chains with delta=6bp classifies them as agreeing (*Eif3f* gene). (*E*) Fraction of agreeing (green) and disagreeing (magenta) intron chains with respect to intron chain length when compared precisely (delta=0bp). (*F*) Same as Fig. 3E, but with delta=6bp.

Sequence characteristics underlying disagreements within RT read pairs. We then analyzed alternative splicing events in 23,356 pairs of PacBio and ONT reads where both reads in a pair were spliced, uniquely mapped, and unambiguously assigned to an isoform (Methods). Exon skipping with respect to the identified isoform was observed 461 times (1.9%) in ONT data and only 45 times (0.2%) in PacBio data (Fig. 4A). Exon skipping in ONT data is typically observed in exons <= 40 bp, however in PacBio data, skipped is rarely observed in exons > 15bp (Fig. 4B). Minimap2 employs exact matching of certain *k*-mers with default k=15 (Li 2018) followed by dynamic programming. However, sequencing errors, especially in ONT data, may cause the above short exons to be missed.

Alternative acceptors (4.2% of ONT reads; 2% of PacBio reads), and comparatively more often, alternative donors (8.3% for ONT and 1.3% for PacBio) were more commonly observed than exon skipping (Fig. 4A). Thus, overall 11.6% of ONT reads and 3.7% of PacBio reads showed one or more discrepancies to the annotation. We found similar trends in the single-cell data, albeit with higher inconsistencies for ONT data (Supplemental Fig. S7A-B). Discrepancies between an alignment and annotation found in both PacBio and ONT likely are novel isoforms. Such cases generally exceed discrepancies supported by PacBio only (Fig. 4A).

From here on, we only consider canonical introns with GU/AG splice sites. We found that inconsistent acceptors were usually shifted by 3-5bp downstream in ONT only, while 80% upstream 3bp-shift ("NAGNAG") acceptors were supported by both PacBio and ONT. Thus, downstream ONT acceptor shifts are questionable, while upstream NAGNAG shifts appear often true (Fig. 4C). For inconsistent donors, a downstream shift of 4bp predominantly occurred for ONT data and with a much smaller overlap between both technologies than for the acceptor sites (Fig. 4D). Such shifts are caused by misalignment at the commonly known GUNNGU donor motif (Wang and Ruvinsky 2010). In summary, downstream 4bp shifts from an annotated donor observed in ONT are doubtful, while 3bp shifts from an annotated acceptor harbor a significant number of trustworthy novel splice sites. Broadly similar observations were made with ScISOr-Seq data (Supplemental Fig. S7C-D).

It is worth noting that modern transcriptome aligners can use annotated splice junctions. This highly reduced the discrepancies between ONT and the annotation, but had a marginal effect on PacBio and cases for which ONT and PacBio agreed. Using the annotation makes PacBio seem to have more inconsistencies than ONT, possibly because novel ONT splice sites are overcorrected to the annotation (Supplemental Fig. S7E-H).

Extrapolating characteristics observed in barcoded read pairs to non-barcoded ONT reads. The observations described above are based on RT read pairs, which require a detected barcode and UMI in both the PacBio and the ONT read. However, as opposed to PacBio data, ONT reads with barcodes have higher Phred scores than those without (Fig. 1D). To understand quality effects on read characteristics, we analyzed the entire Sl-ISO-seq ONT data, which mimics non-barcoded transcriptomic experiments. We observed that the aligned length increased from 532bp for read-wise Phred-score=10 reads to 815bp for Phred-score=20 (Fig. 5A). Similarly, high-quality reads had one more detected intron on average: 2.7 and 3.7 introns for spliced reads with Phred score 10 and 20 respectively (Fig. 5B).

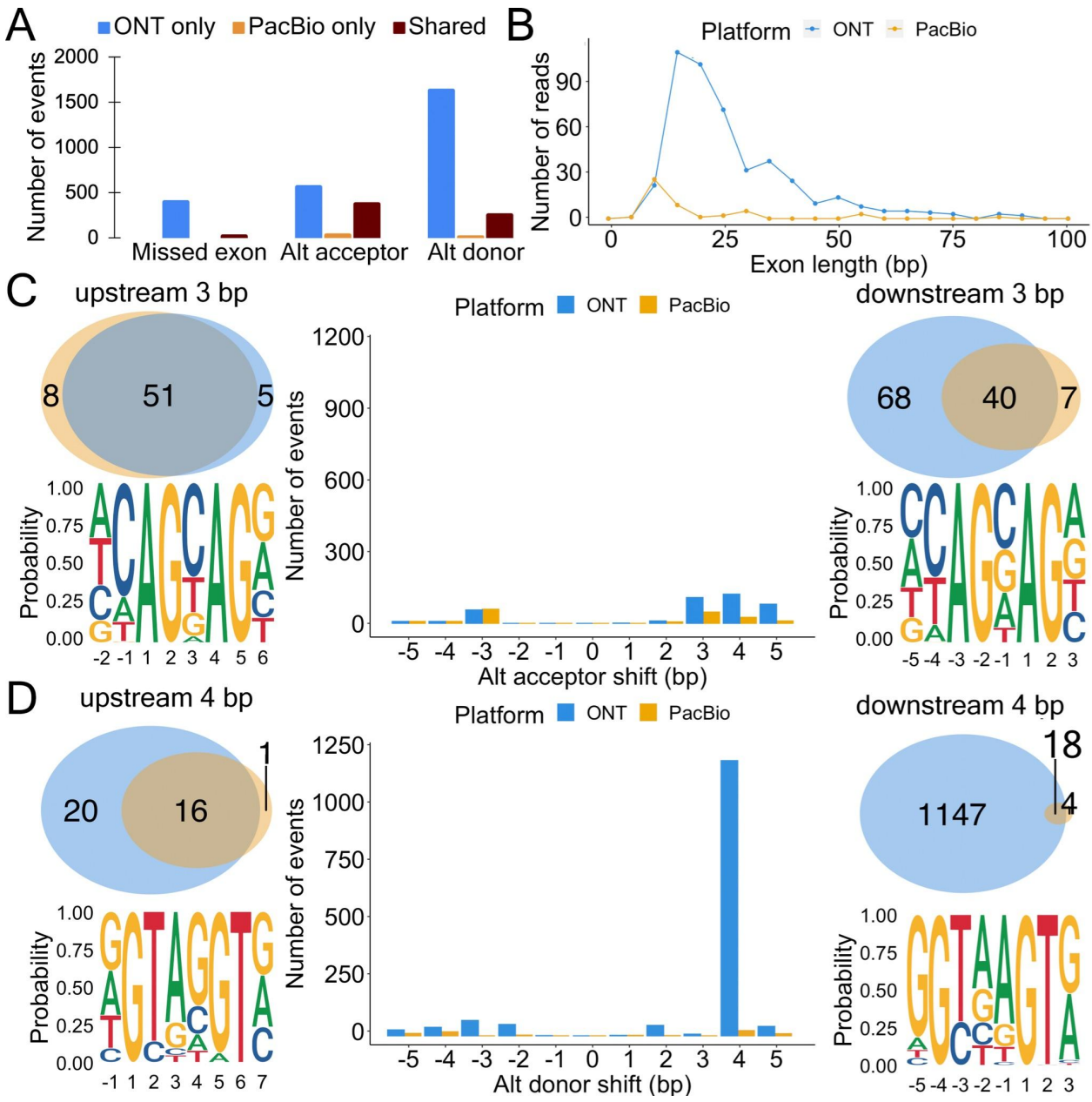

Fig. 4. Exon and splice site characteristics underlying disagreements between PacBio and Nanopore in Sl-ISO-Seq data. (*A*) Number of missed exons (left), alternative acceptors (middle), and donors (right) with respect to the reference annotation that occur only in ONT read (blue), only in PacBio read (yellow), and in both reads from an 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 (yellow) and ONT reads (blue) with respect to the distance from the annotated acceptor site. *Top left:* Venn diagram for 3bp upstream alternative acceptor sites in PacBio (yellow) and ONT reads (blue) from an 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 an RT read pair. *Bottom right:* Nucleotide frequency for loci where 3bp downstream acceptor sites occur. (*D*) *Middle:* number of alternative donor sites in PacBio (yellow) and ONT reads (blue) with respect to the distance from the annotated donor site. *Top left:* Venn diagram for 4bp upstream alternative donor sites in PacBio (yellow) and ONT reads (blue) from an 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 an RT read pair. *Bottom right:* Nucleotide frequency for loci where 4bp downstream donor sites occur.

Furthermore, we compared each read's intron chain against annotated transcripts (Methods). The inconsistency rate goes down with higher Phred score: 28% of reads are inconsistent with the annotation for Phred-score=10 (n=206,675), but only 15% are inconsistent for reads with Phred-score=20 (n=2,667,759, Fig. 5C). Moreover, as found previously (Tilgner et al. 2013; Sharon et al. 2013; Tilgner et al. 2015; Au et al. 2013), inconsistency increases with longer intron chains, i.e., 40% of reads with ≥7 introns are inconsistent, which is ~2.7-fold more than for reads containing ≤3 introns (~15%). Similar trends were observed when intron chains were compared using delta=6bp, although overall inconsistency rates dropped (Fig. 5D). However, we noticed that the lowest inconsistency rate is observed for reads with 2 introns (11%), rather than for single-intron reads (17%). While this observation was previously reported, no explanation was found (Tilgner et al. 2013; Sharon et al. 2013; Tilgner et al. 2015). We hypothesized that aligners can arbitrarily split a read into two exons, while such splits into 3 or more exons are less likely. Additional analysis (Methods) showed that 19% of inconsistent single-intron alignments have their intron entirely within an annotated exon, ~5-fold more than for alignments with 2 introns (3.5%, Fig. 5E). Moreover, only 12% of such mappings are supported by PacBio reads. Thus, rather than representing true alternative isoforms, some inconsistent single-intron mappings likely occur due to misalignments.

We further examined individual splice sites by considering all canonical introns in ONT alignments that matched annotated canonical introns with a loose threshold of 10bp. Similar to barcoded ONT reads, mapped acceptors are rarely shifted upstream and are more consistent with the annotation than donors, which have a dominant 4bp downstream shift caused by the GUNNGU motif (Fig. 5F). To understand the major source of these shifts, we computed distances from all annotated splice sites to the nearest GU/AG (Fig. 5G). As these distances strongly resemble the distribution of ONT shifts, we hypothesized that shifts may depend on the proximity of GU/AG in general, rather than a particular motif.

Thus, we analyzed the inconsistency rate of annotated splice sites with respect to (i) the nearest GU/AG and (ii) read quality (Methods). Donor inconsistency strongly depends on both read quality and distance to the nearest GU (Fig. 5I). For example, the probability of a downstream 4bp shift caused by a GUNNGU motif is ~2 fold lower than of a downstream 2bp shift near a GUGU (3.06% and 6.48% respectively), despite GUNNGU being vastly more frequent than GUGU overall (36,663,585 and 1,066,886 of total detected splice sites near respective motif). Acceptor sites on the other hand, show visible quality dependency only for downstream shifts, most likely due to rare occurrences of upstream AGs. The upstream shifts are dominated by NAGNAG acceptors, which show only a 1.5-fold decrease in inconsistency rate between reads with Phred-score 10 and 20 (4.25% vs. 2.86%). As this difference is noticeably smaller than that for all acceptors (~4.4-fold, 3.13% vs 0.71%), in conjunction with our RT read pair analysis, it suggests that a portion of the upstream NAGNAG shifts may be real.

In summary, due to the elevated number of sequencing errors, ONT read alignments obtained with minimap2 may not provide exact splice site coordinates, especially for the cases when (i) a read has low quality, (ii) a read spans multiple introns, and (iii) canonical dinucleotides are located near splice sites. To avoid a potential misinterpretation of reads disagreeing with the annotated transcripts, one should treat such alignments with additional care or use inexact intron comparison when matching the annotation.

Splice site correction improves transcript discovery precision. Based on the observations made for ONT data we implemented an algorithm for correcting splice junctions in individual reads with the aid of the annotation. This algorithm works with aligned reads and is capable of restoring (i) skipped short exons and (ii) incorrectly detected splice sites (Methods). To evaluate how the designed algorithm affects transcript discovery we simulated ONT reads with NanoSim (Hafezqorani et al. 2020), since the ground truth is unknown for the real data used in this study. Although we could use PacBio reads from the same RT read pair for verification, the fraction of ONT reads having an RT read pair is comparably low and not suitable for transcript model construction.

Transcript discovery was performed using StringTie2 (Kovaka et al. 2019). To mimic real-life situations we removed some transcripts from the annotation before running our correction algorithm and StringTie2. The generated transcripts were matched against the set of "expressed" transcripts (known) and ones that were removed from the annotation (novel) using gffcompare (Pertea and Pertea 2020) to compute precision and recall of different approaches.

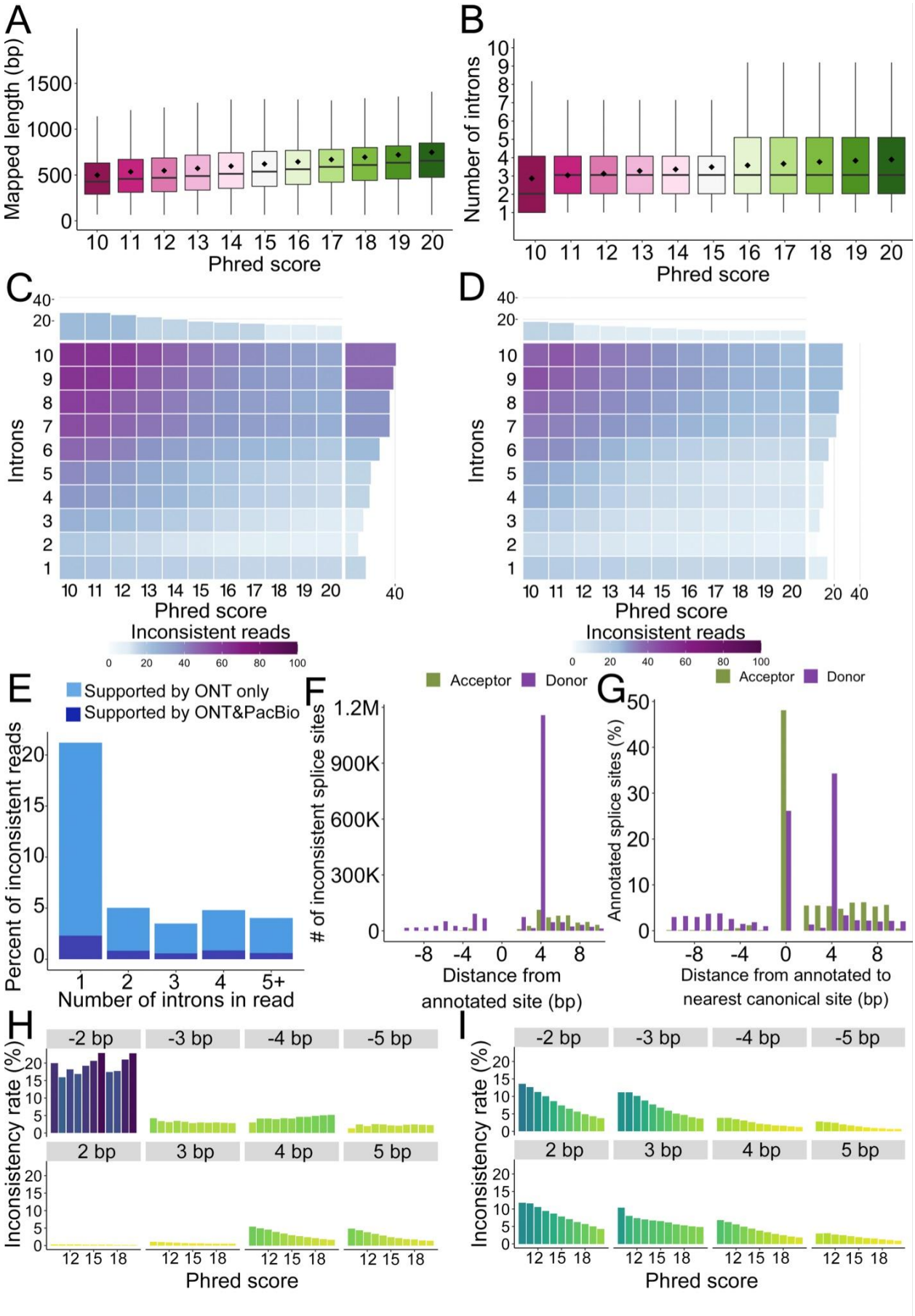
A
Mapped length (bp)
Phred score
B
Number of introns
Phred score
C
Introns
Phred score
Inconsistent reads
D
Introns
Phred score
Inconsistent reads
E
Supported by ONT only
Supported by ONT&PacBio
Percent of inconsistent reads
Number of introns in read
F
Acceptor
Donor
# of inconsistent splice sites
Distance from annotated site (bp)
G
Acceptor
Donor
Annotated splice sites (%)
Distance from annotated to nearest canonical site (bp)
H
-2 bp
-3 bp
-4 bp
-5 bp
2 bp
3 bp
4 bp
5 bp
Inconsistency rate (%)
Phred score
I
-2 bp
-3 bp
-4 bp
-5 bp
2 bp
3 bp
4 bp
5 bp
Inconsistency rate (%)
Phred score

Fig. 5. Characteristics of all ONT reads including non-barcoded ones for Sl-ISO-Seq data. (*A*) Aligned read length with respect to read Phred score (average across all read bases) for all ONT reads from the Sl-ISO-Seq dataset. (*B*) Read intron chain length with respect to read Phred score for all ONT reads from the Sl-ISO-Seq dataset. (*C*) Heatmap showing average inconsistency rate between read intron chains and annotated intron chains (exact comparison, delta=0bp) with respect to read Phred score (X-axis) and intron chain length (Y-axis) for all ONT reads from the Sl-ISO-Seq dataset. Barplot at the top (on the right side) summarizes the inconsistency rate with respect to only the Phred score (only intron chain length). Purple corresponds to a higher inconsistency rate, while light blue indicates a lower inconsistency. (*D*) Same as Fig. 5C, but using inexact intron chain comparison (delta=6bp). (*E*) Histogram showing a fraction of inconsistent ONT reads that have at least one intron entirely contained inside an annotated exon. Dark blue represents reads for which the contained intron is supported by at least one PacBio read, while light blue corresponds to the rest of ONT reads. (*F*) Number of inconsistent donor (purple) and acceptor (green) splice sites in ONT reads from the Sl-ISO-Seq dataset with respect to the distance from the annotated splice site. (*G*) Percentage of annotated canonical donor (green) and acceptor (purple) splice sites with respect to distance to the nearest canonical dinucleotides (GU for donors, AG for acceptors). 0 corresponds to the case when no canonical dinucleotides were detected within 10bp. (*H*) Inconsistency rates of individual acceptor splice sites in ONT reads from the Sl-ISO-Seq dataset with respect to reads' Phred scores. Each histogram represents splice sites with a certain distance from the annotated splice site (grey bars on top). (*I)* Same as Fig. 5H, but for donor splice sites.

To understand the effect of splice site correction we generated read alignments using: (i) deSALT with default options; (ii) minimap2 with default options and without correction; (iii) minimap2 with the annotation and without additional correction; (vi) minimap2 with the annotation and FLAIR (Tang et al. 2020) correction; (v) minimap2 with the annotation and the additional correction by 2passtools (Parker et al. 2021); (vi) minimap2 with the annotation and the additional correction with our algorithm.

Table 1 demonstrates that transcripts generated by StringTie2 using deSALT alignments have significantly lower quality compared to minimap2. Further analysis of deSALT alignments showed that the tool often

incorrectly reports the transcript's strand (for 1.3 out of 2.6M alignments), which can substantially affect downstream analysis.

Running minimap2 with the annotation greatly improves results for novel transcripts. Using FLAIR for additional correction slightly improved precision, but significantly reduced recall (a reduction from 86.4% to 75.7% overall; from 64.2% to 31.0% for novel isoforms). At the same time, the additional correction with our algorithm makes substantial improvements: overall recall and precision increase by 0.9% and 8.8% respectively, while the precision of novel transcripts improves by ~50% (from 21.9% to 32.5%) as the number of false positives among novel transcripts decreases 1.6 fold (from 12,135 down to 7,368). 2passtools performed similarly well, although it was slightly worse in terms of precision (a reduction from 76.5% to 71.4% overall; from 32.5% to 26.2% for novel isoforms). Although minimap2 and the correction algorithms only used information about known transcripts, this information aids the more precise detection of novel isoforms since they often share one or more exons with known isoforms.

	All isoforms		**Known isoforms**		**Novel isoforms**		
	Recall %	Precision %	Recall %	Precision %	Recall %	Precision %	# false isoforms
deSALT	36	24	39	81	17	2	38,209
minimap2	84	58	86	88	58	14	18,683
minimap2 + annotation	86	68	88	89	64	22	12,135
minimap2 + annotation + FLAIR	75	70	80	**93**	31	14	11,767
minimap2 + annotation + 2passtools	85	71	86	91	**67**	26	9,979
minimap2 + annotation + our correction	**87**	**76**	**89**	91	**67**	**33**	**7,368**

Table 1. Precision, recall, and false positive rates of StringTie2 results on the simulated ONT dataset aligned with different strategies. The best values across different methods are highlighted in bold.

Discussion

Different long-read RNA approaches are being increasingly used for isoform analysis (Koren et al. 2012; Au et al. 2013; Sharon et al. 2013; Tilgner et al. 2014, 2015; Oikonomopoulos et al. 2016; Tilgner et al. 2018; Garalde et al. 2018; Gupta et al. 2018; Depledge et al. 2019; Wang et al. 2019; Tardaguila et al. 2018; Volden et al. 2018; Tang et al. 2020; Sun et al. 2021; Joglekar et al. 2021; Hardwick et al. 2021). Therefore, understanding how each approach fares in detecting RNA traits is fundamental. Previous and current comparisons (Li et al. 2014b, 2014a; Weirather et al. 2017; Cui et al. 2020; Pardo-Palacios et al. 2021) of long-read technologies focused on such important data properties as overall read mappability, sequencing error rates, quantification analysis, isoform reconstruction and alternative splicing detection. Indeed, broad conclusions of these studies correlate with our results: ONT does have noticeably higher yield compared to PacBio, but also contains significantly more sequencing errors that complicates spliced alignment and consecutive transcript discovery. However, while being highly useful, none of these studies examines individual reads and compares multiple technologies for an individual RNA molecule. Moreover, previous approaches do not analyze spliced alignment error patterns and their dependency on isoform complexity.

Here, we employed single-molecule barcoding technologies (Gupta et al. 2018; Joglekar et al. 2021) to sequence cDNA copies of single reverse transcription events on PacBio and ONT. Using perfectly matching barcodes and UMIs, we established the correspondence of a pair of ONT and PacBio reads to an individual RNA molecule. This procedure is highly specific while discarding doubtful PacBio-ONT read pairs, but causes a selection for higher-quality reads in ONT but not in PacBio (see Fig. 1C-D). We found important differences that can avoid misinterpretation of data and guide researchers in their choice of technology.

PacBio and ONT reads from an RT read pair frequently differ in length, usually with up to 50 extra nucleotides in the PacBio read, which is small compared to the entire read. However, extra nucleotides in

ONT also exist, although less frequently and fewer. These differences are small but can harbor poly(A) signals, Kozak sequences, or splicing factor binding sites. Additionally, we observed that PacBio reads extended more often to a known TSS and poly(A) site than ONT - a criterion important to defining complete isoforms. Of note, despite an overall low error rate, a significant fraction of PacBio errors arises from homopolymers (up to 40%), while ONT shows more errors but with less bias towards homopolymers.

With respect to exon-intron structures, PacBio-ONT inconsistencies mostly come from splice site shifts and skipped short exons due to alignment errors. Such errors appear more often in ONT, although PacBio may also miss exons shorter than 15bp. These inconsistencies increase as intron chains become longer.

The probability of a donor shift in an alignment primarily depends on the distance from the donor site to the nearest GU dinucleotide. However, the GTNNGT donor motif is very common in mammalian genomes and thus, more donor shifts are explained by this double GT arrangement than with fewer or more separating nucleotides. Donor shifts in ONT reads are usually not confirmed by PacBio reads from the same RNA molecule. For acceptors, the most commonly observed shift is NAGNAG acceptors and such shifts in ONT are often confirmed in the corresponding PacBio reads. Thus, GUNNGU donors in ONT that diverged from the annotation are most often not real while such diverging NAGNAG acceptors are often likely real.

Since using only barcoded ONT reads creates a bias towards high-quality reads, we also analyzed all ONT reads. Low-quality reads are shorter, cover fewer introns, and disagree with the annotation more frequently than reads with high Phred scores. Other trends detected in RT read pairs comparison, such as higher inconsistency for long intron chains and intron shifts in the proximity of canonical dinucleotides, are generally preserved and therefore useful to non-barcoded approaches.

We leveraged the above observations in an algorithm for correcting individual read alignments based on gene annotation. Using simulated Nanopore reads we demonstrate that correcting splice site coordinates and misaligned microexons with our method has a noticeable positive effect on subsequent transcript detection using StringTie2. Moreover, annotation-based correction improves discovery of novel transcripts as they often share exons with known isoforms. Thus, the described findings can be of use for other researchers developing novel algorithms for long-read transcriptome analysis.

While various sources provide different cost estimates for ONT and PacBio RNA sequencing, it seems that a single PromethION flow cell yield is about 3-10 times larger compared to a single Sequel II SMRT cell. Taking into account significantly higher accuracy of PacBio reads in terms of both per-base quality and the ability to correctly detect splice site positions, we believe that PacBio reads can be especially useful for creating de novo annotations and detecting novel isoforms in annotated genomes. As modern transcript discovery tools such as StringTie2 may generate a significant amount of false positive isoforms when using ONT data, Nanopore sequencing should be used for automatic annotation with care. However, as Nanopore sequencing generates noticeably more data, it may be applied for estimating expression levels of annotated transcripts and further differential isoforms expression analysis, as well as detecting isoforms with subsequent manual validation. We believe that more studies and benchmarks, such as LRGASP (Pardo-Palacios et al. 2021), will shed additional light on this. Moreover, as the field of long-read transcriptomics continues its rapid expansion, novel protocols and computational methods will ensure more accurate usage of both PacBio and ONT technologies for all kinds of research projects.

Overall, the single-reverse-transcription event approach provides a powerful instrument for platform comparisons. In contrast to the comparisons of distinct molecules, this method offers tertium-non-datur reasoning, where disagreements are known to be caused by errors of one of the platforms.

Methods

Experimental details. No new samples or sequencing data was generated for this study, however, we provide a brief description of the samples used and experimental protocols followed in the Supplemental Material. The description is taken from Joglekar et al., 2021. Sequenced data was previously submitted to the NCBI Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/) under accession numbers GSE158450 and GSE178175. Read statistics are shown in Supplemental Tables S1 and S2.

Alignment to the reference genome using minimap2. PacBio reads were mapped to GRCm38 mouse reference genome with minimap2 v2.17 (Li 2018) using -t 16 -a -x splice:hq --secondary=no options. ONT reads were mapped with -t 16 -a -x splice -k 14 --secondary=no options. When GENCODE M21 mouse annotation was used for read mapping, it was converted to BED format using paftools.js gff2bed command (included in minimap2 package) and provided to minimap2 using --junc-bed option.

Alignment to the genome using deSALT. PacBio reads were mapped to GRCm38 mouse reference genome with deSALT v1.5.6 (Liu et al. 2019) using -t 16 -x ccs options. ONT reads were mapped with -t 16 -x ont1d options.

Alignment to the genome using GraphMap2. Reads were mapped to GRCm38 mouse genome with GraphMap2 v0.6.5 (Marić et al. 2019) using -t 16 -x rnaseq options. In addition, GraphMap2 was run with GENCODE M21 mouse annotation provided using --gtf option, however, the run failed with an error.

Alignment to the genome using uLTRA. PacBio reads were mapped to GRCm38 mouse reference genome with uLTRA v0.0.4.1 (Sahlin and Mäkinen 2021) using --isoseq --t 24 options. ONT reads were mapped with --ont --t 24 options.

Pairwise read alignment. Pairwise read alignment was performed using the Smith-Waterman local alignment algorithm implemented in SSW Python library (Zhao et al. 2013) with default options.

Sequencing error rate. Sequencing error rates were computed based on minimap2' alignments using a 3-way comparison between the reference genome and RT read pairs. An error at a certain position in a read from RT read pair was reported only when the alignment shows a difference from the genome (i.e. insertion, deletion, or substitution), while the second read from the pair either matches the genome or

contains an alternative discrepancy at this position (e.g., another base is inserted). Identical differences from the reference genome (same position and nucleotide) detected in both PacBio and ONT reads from an RT read pair were not classified as sequencing errors. An error is deemed to occur within a homopolymer region if any 3bp window in the genome that contains an error position consists of the same nucleotides.

***k*-mer identity and homopolymer compression.** The *k*-mer identity (k=14bp) with the reference genome was calculated using minimap2 alignments for each exon individually. We first extracted all genomic *k*-mers from the respective mapped region (of the exon) and then calculated the fraction of the *k*-mers that occurred within this exon in the read. Homopolymer compression (Au et al. 2012) was performed by substituting all stretches of identical nucleotides (2bp and longer) with a single nucleotide of the same kind in both read sequence and reference sequence from the respective mapping region. The *k*-mer identity was then computed in the same way as for non-compressed sequences.

TSS / poly(A) analysis. TSS and poly(A) sites were assigned to each read as previously done (Joglekar et al. 2021). For published TSS we used high-quality calls from the FANTOM5 consortium (Lizio et al. 2015). Among all published TSS calls, within 100 bp of the 5' end of the read mapping we assigned the closest TSS to the read and none if there are no such calls within 100 bp. Using very recent poly(A) site calls (Herrmann et al. 2019), we applied a similar procedure to assign poly(A) sites to the read (within 100 bp of the 3' end of the read mapping).

An RT read pair is considered as "agreeing" on TSS/poly(A) site assignment if both reads have an assigned TSS/poly(A) site and the two assigned sites are identical.

Intron chain comparison and inconsistency detection. Intron chains were compared against each other as ordered lists of coordinate pairs. In the precise intron chain comparison, two introns are considered equal if their splice site coordinates are identical (delta=0bp), while in the inexact comparison each splice site is allowed to differ by delta=6bp at most. Intron chains are considered as agreeing if they are equal or one chain is a sub-chain of another with respect to the given delta value and disagreeing otherwise.

To detect inconsistencies between reads in RT read pairs, intron chains for both reads were extracted from the BAM files obtained with minimap2 and compared against each other as described above. Similarly, to detect agreement between a read and the annotation, a read intron chain extracted from the BAM file was compared against intron chains of known transcripts from GENCODE M21 comprehensive mouse annotation. Read is deemed to be consistent if its intron chain agrees with at least one annotated transcript and inconsistent otherwise. Reads that do not overlap with annotated exons (i.e. entirely map to intergenic or intronic regions) are considered as non-informative and are ignored in the analysis.

Classification of splicing modifications. To classify splice site inconsistencies with the gene annotation reads were assigned to known transcripts using a custom script assign_reads.py available at https://github.com/ablab/platform_comparison, which assigns mapped reads to known isoforms based on intron chains and nucleotide identity. For PacBio reads the script was run with --data_type pacbio_ccs option, for ONT reads --data_type nanopore was used. Benchmarking of the method on the simulated data is presented in Supplemental Note "Benchmarking of the read-to-isoform assignment algorithm" (Supplemental Table S4).

For further analysis, we selected unambiguous assignments with respective reported splicing modifications (skipped exons, alternative donor or acceptor site). To investigate alternative donors and acceptors (i.e. shift frequency, nucleotide content) only introns with canonical splice sites were used (GT-AG).

The output of the script was also used to track the origin of inconsistent non-barcoded reads. To detect reads having at least one intron entirely located within an annotated exon we selected uniquely assigned reads having this specific type of inconsistency (additional novel intron according to our categories).

Splice site analysis for non-barcoded reads. To analyze splice site consistency in non-barcoded reads we assigned each read intron separately to an annotated intron (rather than the entire intron chain) with a loose threshold delta=10bp. Such an approach allows to maximize the number of investigated splice sites and consider individual introns even from inconsistent chains. For the analysis, we selected only cases when both read and annotation intron have canonical splice sites (GU/AG). We say that an assigned read

intron correctly detects a splice site if its position is equal to the annotated splice site (0bp difference), and incorrectly otherwise. The inconsistency rate of an annotated splice site is defined as the number of incorrect calls divided by the total number of read introns assigned. Each annotated splice site was classified according to the distance to the nearest GU (for donors) or AG (for acceptors) in the vicinity of 10bp. It allowed us to compute overall inconsistency rates for splice sites with respect to this distance.

Splice site correction algorithm. The correction algorithm takes aligned reads and genome annotation as input. Each read is processed individually, as opposed to the classic transcript construction method that relies on clustering and splice site consensus (Kovaka et al. 2019; Wyman et al. 2020; Tang et al. 2020; Sahlin and Medvedev 2021). An aligned read is first assigned to a reference isoform based on inexact intron chain matching and exon similarity as described above. Further, each read is examined with respect to the accuracy of the detected intron structure. Coordinates of corrected alignments are output in BED12 format. The algorithm is available at https://github.com/ablab/platform_comparison (correct_splice_sites.py).

Splice site correction algorithm: restoring skipped exons from neighboring splice sites. A reference exon is considered to be skipped during the alignment if it is: (i) shorter than 50bp; (ii) spanned by a read intron; (iii) adjacent exons in the alignment contain extra sequences reaching into the annotated introns surrounding the reference exon and these two extra sequences are of a similar total length as the reference exon (Supplemental Fig. S9).

Splice site correction algorithm: correcting individual splice sites. With the same considerations as above, an individual spice site in the read is to be corrected if (1) it is no further than delta=6bp apart from a known splice site and (2) the read alignment has indels close to this position.

Simulating ONT data. To simulate ONT data we employed the NanoSim software in transcriptome mode (Hafezqorani et al. 2020) using the pre-trained ONT cDNA error model. However, examining the code, we found that NanoSim randomly selects a starting position of a read in an mRNA to simulate truncation. This is performed using a uniform distribution, thus assuming that 5' and 3' are identical, which is not the case for the real data. To avoid this pitfall, we mapped raw ONT reads to the reference

transcripts using minimap2 (with *-x map-ont* option) and estimated the probabilities of the initial sequence being truncated on each side by N% of its length. We thus modified the NanoSim truncation procedure so that reference sequences are clipped according to empirically derived probabilities (Supplemental Fig. S10). In addition, we turned off the simulation of random decoy reads, which represent the background noise of the sequencing experiment. We simulated 30M ONT reads using transcripts from the Mouse GENCODE v26 basic annotation (Frankish et al. 2021). In addition, a 30bp poly(A) tail was attached to every transcript prior to simulation. Each transcript with at least one generated read was considered as "expressed" and then represented the ground truth.

Evaluating transcript model construction. We first generated a reduced genome annotation by removing 20% of expressed spliced transcripts. Removed transcripts are considered novel, while expressed transcripts kept in the annotation represent the set of known models. Using these two sets allowed to independently evaluate the ability of the algorithm to report known and discover novel isoforms. StringTie2 results were similarly split into novel and known models based on the information provided in the GTF file and gffcompare (Pertea and Pertea 2020) was further launched to estimate precision and recall.

In addition, we performed the series of experiments with different fractions of excluded expressed isoforms to analyze how this parameter affects the results (Supplemental Note “Novel isoform discovery with different fractions of unknown transcripts”, Supplemental Table S5).

Running StringTie2. StringTie2 (Kovaka et al. 2019) was run with -L option for long reads and the reduced genome annotation.

Running FLAIR. Correction module of FLAIR (Tang et al. 2020) was run with minimap2’ alignments (converted to BED file using bam2Bed12.py) and the reduced genome annotation as an input.

Running 2passtools. First, 2passtools (Parker et al. 2021) ‘score’ command was run on minimap2’ alignments. Results were filtered with ‘filter --exprs 'decision_tree_2_pred'’ command and provided to minimap2 using --junc-bed option.

Data access

Simulated data generated in this study is available at Zenodo: https://doi.org/10.5281/zenodo.6325107. The barcode detection tool is available as the "GetBarcodes" function in the scicorseqr R-package (github.com/noush-joglekar/scisorseqr). All scripts used for data analysis and spliced alignment correction are available as Supplemental Code and at http://www.github.com/ablab/platform_comparison.

Competing interests

A.J., A.D.P., A.M., and H.U.T. declare no competing interests.

Acknowledgments

We thank Alyona Sidorova and Alexandra Bazarova for their aid with the analysis. This work was supported by NIGMS (grant 1R01GM135247-01 to H.U.T), St. Petersburg State University, Russia (grant ID PURE 93023187 to A.M. and A.D.P.). Scientific research was performed at the Research park of St. Petersburg State University «Computing Center».

References

Au KF, Sebastiano V, Afshar PT, Durruthy JD, Lee L, Williams BA, van Bakel H, Schadt EE, Reijo-Pera RA, Underwood JG, et al. 2013. Characterization of the human ESC transcriptome by hybrid sequencing. *Proc Natl Acad Sci U S A* **110**: E4821–30.

Au KF, Underwood JG, Lee L, Wong WH. 2012. Improving PacBio long read accuracy by short read alignment. *PLoS One* **7**: e46679.

Byrne A, Beaudin AE, Olsen HE, Jain M, Cole C, Palmer T, DuBois RM, Forsberg EC, Akeson M, Vollmers C. 2017. Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells. *Nat Commun* **8**: 16027.

Chen Z, Pham L, Wu TC, Mo G, Xia Y, Chang PL, Porter D, Phan T, Che H, Tran H, et al. 2020. Ultralow-input single-tube linked-read library method enables short-read second-generation sequencing systems to routinely generate highly accurate and economical long-range sequencing information. *Genome Res* **30**: 898–909.

Cui J, Shen N, Lu Z, Xu G, Wang Y, Jin B. 2020. Analysis and comprehensive comparison of PacBio and nanopore-based RNA sequencing of the Arabidopsis transcriptome. *Plant Methods* **16**: 85.

Depledge DP, Srinivas KP, Sadaoka T, Bready D, Mori Y, Placantonakis DG, Mohr I, Wilson AC. 2019. Direct RNA sequencing on nanopore arrays redefines the transcriptional complexity of a viral pathogen. *Nat Commun* **10**: 754.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**: 15–21.

Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, et al. 2009. Real-time DNA sequencing from single polymerase molecules. *Science* **323**: 133–138.

Engström PG, Steijger T, Sipos B, Grant GR, Kahles A, Rätsch G, Goldman N, Hubbard TJ, Harrow J, Guigó R, et al. 2013. Systematic evaluation of spliced alignment programs for RNA-seq data. *Nat Methods* **10**: 1185–1191.

Frankish A, Diekhans M, Jungreis I, Lagarde J, Loveland JE, Mudge JM, Sisu C, Wright JC, Armstrong J, Barnes I, et al. 2021. GENCODE 2021. *Nucleic Acids Res* **49**: D916–D923.

Garalde DR, Snell EA, Jachimowicz D, Sipos B, Lloyd JH, Bruce M, Pantic N, Admassu T, James P, Warland A, et al. 2018. Highly parallel direct RNA sequencing on an array of nanopores. *Nat Methods* **15**: 201–206.

Gupta I, Collier PG, Haase B, Mahfouz A, Joglekar A, Floyd T, Koopmans F, Barres B, Smit AB, Sloan SA, et al. 2018. Single-cell isoform RNA sequencing characterizes isoforms in thousands of cerebellar cells. *Nat Biotechnol*. http://dx.doi.org/10.1038/nbt.4259.

Hafezqorani S, Yang C, Lo T, Nip KM, Warren RL, Birol I. 2020. Trans-NanoSim characterizes and simulates nanopore RNA-sequencing data. *Gigascience* **9**. http://dx.doi.org/10.1093/gigascience/giaa061.

Hardwick SA, Hu W, Joglekar A, Fan L, Collier PG, Foord C, Balacco J, Belchikov N, Jarroux J, Prjibelski A, et al. 2021. Single-nuclei isoform RNA sequencing reveals combination patterns of transcript elements across human brain cell types. *bioRxiv* 2021.12.29.474385. https://www.biorxiv.org/content/10.1101/2021.12.29.474385 (Accessed January 17, 2022).

Herrmann CJ, Schmidt R, Kanitz A, Artimo P, Gruber AJ, Zavolan M. 2019. PolyASite 2.0: a consolidated atlas of polyadenylation sites from 3′ end sequencing. *Nucleic Acids Res* **48**: D174–D179.

Joglekar A, Prjibelski A, Mahfouz A, Collier P, Lin S, Schlusche AK, Marrocco J, Williams SR, Haase B, Hayes A, et al. 2021. A spatially resolved brain region- and cell type-specific isoform atlas of the postnatal mouse brain. *Nat Commun* **12**: 463.

Koren S, Schatz MC, Walenz BP, Martin J, Howard JT, Ganapathy G, Wang Z, Rasko DA, McCombie WR, Jarvis ED, et al. 2012. Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nat Biotechnol* **30**: 693–700.

Kovaka S, Zimin AV, Pertea GM, Razaghi R, Salzberg SL, Pertea M. 2019. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol* **20**: 278.

Lee DSM, Ghanem LR, Barash Y. 2020. Integrative analysis reveals RNA G-quadruplexes in UTRs are selectively constrained and enriched for functional associations. *Nat Commun* **11**: 527.

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**: 3094–3100.

Li S, Łabaj PP, Zumbo P, Sykacek P, Shi W, Shi L, Phan J, Wu P-Y, Wang M, Wang C, et al. 2014a. Detecting and correcting systematic variation in large-scale RNA sequencing data. *Nat Biotechnol* **32**: 888–895.

Li S, Tighe SW, Nicolet CM, Grove D, Levy S, Farmerie W, Viale A, Wright C, Schweitzer PA, Gao Y, et al. 2014b. Multi-platform assessment of transcriptome profiling using RNA-seq in the ABRF next-generation sequencing study. *Nat Biotechnol* **32**: 915–925.

Liu B, Liu Y, Li J, Guo H, Zang T, Wang Y. 2019. deSALT: fast and accurate long transcriptomic read alignment with de Bruijn graph-based index. *Genome Biol* **20**: 274.

Lizio M, Harshbarger J, Shimoji H, Severin J, Kasukawa T, Sahin S, Abugessaisa I, Fukuda S, Hori F, Ishikawa-Kato S, et al. 2015. Gateways to the FANTOM5 promoter level mammalian expression atlas. *Genome Biol* **16**: 22.

Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, Tirosh I, Bialas AR, Kamitaki N, Martersteck EM, et al. 2015. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**: 1202–1214.

Marić J, Sović I, Križanović K, Nagarajan N, Šikić M. 2019. Graphmap2-splice-aware RNA-seq mapper for long reads. *bioRxiv*. https://www.biorxiv.org/content/10.1101/720458v1.abstract.

Oikonomopoulos S, Wang YC, Djambazian H, Badescu D, Ragoussis J. 2016. Benchmarking of the Oxford Nanopore MinION sequencing for quantitative and qualitative assessment of cDNA populations. *Sci Rep* **6**: 31602.

Pardo-Palacios F, Reese F, Carbonell-Sala S, Diekhans M, Liang C, Wang D, Williams B, Adams M, Behera A, Lagarde J, et al. 2021. Systematic assessment of long-read RNA-seq methods for transcript identification and quantification. *Research Square*. https://www.researchsquare.com/article/rs-777702/latest (Accessed January 17, 2022).

Parker MT, Knop K, Barton GJ, Simpson GG. 2021. 2passtools: two-pass alignment using machine-learning-filtered splice junctions increases the accuracy of intron detection in long-read RNA sequencing. *Genome Biol* **22**: 72.

Pertea G, Pertea M. 2020. GFF Utilities: GffRead and GffCompare. *F1000Res* **9**. http://dx.doi.org/10.12688/f1000research.23297.2.

Sahlin K, Mäkinen V. 2021. Accurate spliced alignment of long RNA sequencing reads. *Bioinformatics*. http://dx.doi.org/10.1093/bioinformatics/btab540.

Sahlin K, Medvedev P. 2021. Author Correction: Error correction enables use of Oxford Nanopore technology for reference-free transcriptome analysis. *Nat Commun* **12**: 992.

Sharon D, Tilgner H, Grubert F, Snyder M. 2013. A single-molecule long-read survey of the human transcriptome. *Nat Biotechnol* **31**: 1009–1014.

Singh M, Al-Eryani G, Carswell S, Ferguson JM, Blackburn J, Barton K, Roden D, Luciani F, Giang Phan T, Junankar S, et al. 2019. High-throughput targeted long-read single cell sequencing reveals the clonal and transcriptional landscape of lymphocytes. *Nat Commun* **10**: 3120.

Steijger T, Abril JF, Engström PG, Kokocinski F, RGASP Consortium, Hubbard TJ, Guigó R, Harrow J, Bertone P. 2013. Assessment of transcript reconstruction methods for RNA-seq. *Nat Methods* **10**:

1177–1184.

Sun YH, Wang A, Song C, Shankar G, Srivastava RK, Au KF, Li XZ. 2021. Single-molecule long-read sequencing reveals a conserved intact long RNA profile in sperm. *Nat Commun* **12**: 1361.

Tang AD, Soulette CM, van Baren MJ, Hart K, Hrabeta-Robinson E, Wu CJ, Brooks AN. 2020. Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. *Nat Commun* **11**: 1438.

Tardaguila M, de la Fuente L, Marti C, Pereira C, Pardo-Palacios FJ, Del Risco H, Ferrell M, Mellado M, Macchietto M, Verheggen K, et al. 2018. SQANTI: extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification. *Genome Res*. http://dx.doi.org/10.1101/gr.222976.117.

Tilgner H, Grubert F, Sharon D, Snyder MP. 2014. Defining a personal, allele-specific, and single-molecule long-read transcriptome. *Proc Natl Acad Sci U S A* **111**: 9869–9874.

Tilgner H, Jahanbani F, Blauwkamp T, Moshrefi A, Jaeger E, Chen F, Harel I, Bustamante CD, Rasmussen M, Snyder MP. 2015. Comprehensive transcriptome analysis using synthetic long-read sequencing reveals molecular co-association of distant splicing events. *Nat Biotechnol* **33**: 736–742.

Tilgner H, Jahanbani F, Gupta I, Collier P, Wei E, Rasmussen M, Snyder M. 2018. Microfluidic isoform sequencing shows widespread splicing coordination in the human transcriptome. *Genome Res* **28**: 231–242.

Tilgner H, Raha D, Habegger L, Mohiuddin M, Gerstein M, Snyder M. 2013. Accurate identification and analysis of human mRNA isoforms using deep long read sequencing. *G3* **3**: 387–397.

Tung LH, Shao M, Kingsford C. 2019. Quantifying the benefit offered by transcript assembly with Scallop-LR on single-molecule long reads. *Genome Biology* **20**. http://dx.doi.org/10.1186/s13059-019-1883-0.

Vaquero-Garcia J, Barrera A, Gazzara MR, González-Vallinas J, Lahens NF, Hogenesch JB, Lynch KW, Barash Y. 2016. A new view of transcriptome complexity and regulation through the lens of local splicing variations. *Elife* **5**: e11752.

Volden R, Palmer T, Byrne A, Cole C, Schmitz RJ, Green RE, Vollmers C. 2018. Improving nanopore read accuracy with the R2C2 method enables the sequencing of highly multiplexed full-length single-cell cDNA. *Proc Natl Acad Sci U S A* **115**: 9726–9731.

Wagner GP, Kin K, Lynch VJ. 2012. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory Biosci* **131**: 281–285.

Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB. 2008. Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**: 470–476.

Wang PP-S, Ruvinsky I. 2010. Computational prediction of Caenorhabditis box H/ACA snoRNAs using genomic properties of their host genes. *RNA* **16**: 290–298.

Wang Y, Wang A, Liu Z, Thurman AL, Powers LS, Zou M, Zhao Y, Hefel A, Li Y, Zabner J, et al. 2019. Single-molecule long-read sequencing reveals the chromatin basis of gene expression. *Genome Res* **29**: 1329–1342.

Weirather JL, Afshar PT, Clark TA, Tseng E, Powers LS, Underwood JG, Zabner J, Korlach J, Wong WH, Au KF. 2015. Characterization of fusion genes and the significantly expressed fusion isoforms in breast cancer by hybrid sequencing. *Nucleic Acids Res* **43**: e116.

Weirather JL, de Cesare M, Wang Y, Piazza P, Sebastiano V, Wang X-J, Buck D, Au KF. 2017. Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. *F1000Res* **6**: 100.

Wu I, Kim HS, Ben-Yehezkel T. 2019. A Single-Molecule Long-Read Survey of Human Transcriptomes using LoopSeq Synthetic Long Read Sequencing. *bioRxiv* 532135. https://www.biorxiv.org/content/10.1101/532135v2 (Accessed April 26, 2021).

Wyman D, Balderrama-Gutierrez G, Reese F, Jiang S. 2020. A technology-agnostic long-read analysis pipeline for transcriptome discovery and quantification. *BioRxiv*. https://www.biorxiv.org/content/10.1101/672931v2.abstract.

Zeisel A, Munoz-Manchado AB, Codeluppi S, Lonnerberg P, La Manno G, Jureus A, Marques S, Munguba H, He L, Betsholtz C, et al. 2015. Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* **347**: 1138–1142.

Zhao M, Lee W-P, Garrison EP, Marth GT. 2013. SSW library: an SIMD Smith-Waterman C/C++ library for use in genomic applications. *PLoS One* **8**: e82138.