Contrasting and combining transcriptome complexity captured by short and long RNA sequencing reads

Seong W Han, San Jewell, Andrei Thomas-Tikhonenko, et al.

Genome Res. published online September 25, 2024
Access the most recent version at doi:10.1101/gr.278659.123

Contrasting and combining transcriptome complexity captured by short and long RNA sequencing reads

Seong Woo Han,[1*] San Jewell,[2*] Andrei Thomas-Tikhonenko,[3,4] Yoseph Barash[1,2]

[1]Department of Computer and Information Sciences, School of Engineering, University of Pennsylvania
[2]Department of Genetics, Perelman School of Medicine, University of Pennsylvania
[3]Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania
[4]Division of Cancer Pathobiology, Children's Hospital of Philadelphia

[*]Equal contribution
To whom correspondence should be addressed; E-mail: yosephb@upenn.edu.

Mapping transcriptomic variations using either short or long reads RNA sequencing is a staple of genomic research. Long reads are able to capture entire isoforms and overcome repetitive regions, while short reads still provide improved coverage and error rates. Yet how to quantitatively compare the technologies, can we combine those, and what may be the benefit of such a combined view remain open questions. We tackle these questions by first creating a pipeline to assess matched long and short reads data using a variety of transcriptome statistics. We find that across datasets, algorithms, and technologies, matched short reads data detects roughly 30% more splice junctions such that 10-30% of the splice junctions included at 20% or more by short reads are missed by long reads. In contrast, long reads detect many more intron retention events and can detect full isoforms, pointing to the benefit of combining the technologies. We introduce MAJIQ-L, an extension of the MAJIQ software to enable a unified view of transcriptome variations from both technologies and demonstrate its benefits. Our software can be used to assess any future long-read technology or algorithm, and combine it with short reads data for improved transcriptome analysis.

Introduction

Long reads sequencing technology has been revolutionizing genomic studies in recent years, leading to it being elected recently as the "method of the year 2022" (Marx, 2023; Lucas and Novoa, 2023; Foord et al., 2023). The most commonly used platforms, Pacific Biosciences

(PacBio) and Oxford Nanopore Technologies (ONT) offer RNA sequencing with read lengths typically varying between a few hundred to a few thousand bases long, depending on the technology and the protocol used. Consequently, many algorithms have been developed for transcript discovery and quantification from long reads, such as FLAIR (Tang et al., 2020), ESPRESSO (ESPRESSO et al., 2023), IsoQuant (Prjibelski et al., 2023), Bambu (Chen et al., 2023), TALON (Wyman et al., 2019), SQANTI (Tardaguila et al., 2018), StringTie (Shumate et al., 2022; Kovaka et al., 2019), single-molecule long-read sequencing technology (Sharon et al., 2013), and IDP (Au et al., 2013). Although both the technology and associated algorithms move at a fast pace, long reads RNA sequencing still suffers from several key limitations (Kovaka et al., 2023). Specifically, many reads are still not long enough to capture entire transcripts, the high error rate makes it hard to detect exact isoforms and associated splice sites, and low coverage leads to limited isoform detection and quantification. In contrast, Illumina RNA short reads are typically only 100-150 bp long and, therefore, harder to assign to a specific isoform. Nonetheless, short reads still allow researchers to detect and quantify alternative splicing (AS) 'events' or, more generally, local splicing variations (LSV). LSV, first introduced in MAJIQ (Vaquero-Garcia et al., 2016), denotes splits in a gene splice graph coming out of or into a reference exon. As such, LSV captures 'classical' AS events (e.g., cassette exons) but also more complex events involving multiple junctions or exons, including *de novo* (unannotated) junctions, exons, and introns. LSV are typically quantified using junction spanning reads in terms of percent spliced in (PSI, denoted Ψ), representing the relative percentage or ratio of isoforms with a specific splicing junction or intron retention (IR).

The availability of short and long-read technologies raises the natural question of how these compare and whether they can be effectively combined. Yet previous work involving long reads has focused mainly on the benefits it may offer, lacking a comprehensive comparative evaluation of the resulting transcriptome maps. Similarly, tools that combine short and long reads to aid researchers in downstream splicing analysis are still underdeveloped.

To address these needs, we developed an analysis pipeline and accompanying software, MAJIQ-L. The analysis pipeline shown in Fig. 1A takes as input three sources of information: Transcriptome annotation; short reads processed by MAJIQ V2 (Vaquero-Garcia et al., 2023); and long reads in GTF format, processed by the user's algorithm of choice. It then computes and displays an extensive set of statistics that contrast the available annotation and the two sequencing sources in terms of novel junctions, introns, coverage, inclusion levels, etc., such that existing gaps between the three sources can be captured (Fig. 1B). Using the three input sources, MAJIQ-L constructs unified gene splice graphs with all isoforms and all LSVs visible for analysis. This unified view is implemented in a new visualization package (VOILA v3), allowing users to inspect each gene of interest where the three sources agree or differ (Fig. 1C).

We apply MAJIQ-L to matched short and long reads from several datasets involving both PacBio and ONT using four different long reads transcriptome mapping algorithms. First, we contrast short and long reads by statistics reflecting splice junction detection and quantification. Next, we inspect the coverage difference between short and long reads, 3' to 5' bias in long-read sequencing, and whether GC content at splice junctions contributes to their differential detection

by short and long reads. Then, we turn to intron retention, showing that, as expected, long reads detect many more introns than short reads, but short reads detect longer introns. Finally, we demonstrate the usefulness of a combined long and short reads analysis using VOILA v3 for splicing variations in the *SRSF11* gene.

Results

Short reads detect 30% more splice junctions at the same coverage level

To contrast the observed transcriptome complexity by short and long reads, we compared splice junctions detected by short reads processed by STAR (Dobin et al., 2013) followed by MAJIQ's LSV analysis (Vaquero-Garcia et al., 2016) and four different long-read algorithms (Tang et al., 2020; ESPRESSO et al., 2023; Prjibelski et al., 2023; Chen et al., 2023) (see Methods for details). Each detected splice junction was assigned to one of six categories, represented by distinct colors, based on the source of support from either short reads, long reads, or annotation (Fig. 2A). This analysis was performed using three different datasets: Three replicates of human cell lines from the Long-read RNA-seq Genome Annotation Assessment Project (LRGASP) Consortium (Pardo-Palacios et al., 2021) (Fig. 2B); Three heart atrial appendage, brain frontal cortex, and liver samples from GTEx v9 (Glinos et al., 2022) and a PDX cell line sample derived from a patient with a relapsed B Cell Acute Lymphoblastic Leukemia (B-ALL) (Bagashev et al., 2021; Schulz et al., 2021) (Supplemental Fig. S1). To address coverage differences, we sub-sampled the files such that the total number of bases sequenced across the various platforms was similar (see Materials and Methods). Overall, our results indicate that across all 3 datasets, short reads detected 30% more splice junctions, with PacBio detecting about 10% more junctions than ONT. Some differences between long-read algorithms were also apparent. FLAIR detected the most amount of long reads only de novo splice junctions ($\approx 8\%$), while Bambu reported the least ($\approx 3.3\%$). These differences may reflect lower precision and recall, respectively (Prjibelski et al., 2023). We also performed a comparative analysis when using the original dataset without sub-sampling, where long reads have 1.3-2.9 fold more bases sequenced than short reads (LRGASP, B-ALL) and where Illumina has 1.7-fold more coverage than ONT (GTEx). Similar trends were observed in this analysis as well (Supplemental Fig. S2, and Supplemental Table S1, S3, S3 for details).

The significant difference in splice junction detection naturally raises the question of whether the junctions uniquely detected by either technology are real. In the case of short reads, MAJIQ only reports splice junctions when they are supported by multiple unique reads that map distinctly to multiple positions, which has been shown to lead to very few false positives (Engström et al., 2013; Baruzzo et al., 2017). In contrast, long reads only de novo splice junctions are more likely to involve false positives given their known high error rates and limited coverage (Hardwick et al., 2019; Amarasinghe et al., 2020). To further assess the reliability of the unannotated (de-novo) splice junctions uniquely detected by short and long reads, we compared those detected junctions against those in intropolis (Nellore et al., 2016). For GTEx tissue sam-

ples, short reads displayed high overlap with Intropolis with heart atrial appendage, brain frontal cortex, and liver exhibiting overlaps of 89% ± 1%, 88% ± 1%, and 86% ± 1%, respectively. In contrast, the FLAIR processed GTEx long reads yielded significantly lower overlap with intropolis, showing 43% ± 2%, 41% ± 3%, and 43% ± 2%, for those tissue samples. Further assessment with the LRGASP dataset revealed that short reads only de novo junctions maintained high alignment at 86% ± 0.5%. Long reads only de novo junctions from four different algorithms demonstrated variable and generally lower alignment scores. FLAIR achieved 28% ± 0.7% on PacBio and 47% ± 0.8% on ONT; IsoQuant recorded 30% ± 0.3% on PacBio and 23% ± 1% on ONT; ESPRESSO reached 31% ± 3% on PacBio and 17% ± 3% on ONT; and Bambu showed markedly lower alignment of 3% ± 0.5% on PacBio and 2% ± 0.2% on ONT. In summary, while this kind of overlap analysis with existing splice junctions database can not be used to assess reliable junction detection directly, these results offer additional support for the assertion that the vast majority of short reads only de novo junctions are real while their long reads counterpart may suffer from significantly higher false positives. This conclusion regarding the low false discovery of junctions following MAJIQ's processing of short reads data is further supported by additional analysis described in the Supplementary Information.

Nevertheless, the significant number of short reads only de novo splice junctions begs the question of whether those additional junctions are meaningful. To address this and similar questions regarding the observed differences between transcriptomic variations detected by short and long reads, we performed the analysis below using IsoQuant. We choose IsoQuant primarily due to its ease of use, as it can be efficiently executed with a single command line, which simplifies the workflow compared to other tools that require a couple of steps. We ensured transparency in our methodology by demonstrating in both the main and supplementary texts that consistent trends are observable across all analyses, regardless of the algorithms. The results shown in Fig. 2C for IsoQuant indicate most short reads only de novo splice junctions are, as expected, relatively lowly included ($\Psi < 10\%$). Nonetheless, the cumulative distribution function plot shows IsoQuant+PacBio misses 10% of junctions with significant inclusion levels ($\Psi > 20\%$), and IsoQuant+ONT misses 30% of junctions detected by MAJIQ (Fig. 2C right). When comparing the different algorithms, IsoQuant misses the least amount of junctions with $\Psi > 20\%$, and Bambu misses the most with the same inclusion level (Supplemental Fig. S8B).

Patterns of novel splice variants difference in short and long reads

Next, we utilized the VOILA modulizer (Vaquero-Garcia et al., 2023) to more precisely characterize splice variants that are unique to either short-read or long-read technologies. Essentially, a module within this context represents a unique segment of a gene's splicegraph, encompassing overlapping LSV that are contained between a single source and a single target exon. By focusing on splice variants within these modules, we were able to assess distinct splice variants and how these relate to each other, as we describe below. For short reads, we classified *de novo* junctions reported by MAJIQ into two categories: A junction involving novel splice sites and a junction creating a novel combination of known splice sites (Fig. 3A, top). Compared with

IsoQuant, the distribution of the two categories is the same for PacBio and ONT, showing about 90% involve novel splice sites while the rest is novel combination. A similar trend is observed when compared with different algorithms (Supplemental Fig. S3).

In long reads *de novo* junctions, we observe another type of transcript change, termed putative transcript start or end site (pTSS/pTES). pTSS/pTES occur when only a partial exon is output at the edge of the transcript processed by the long read algorithm (*e.g.*, IsoQuant). We term those pTSS/pTES as these do not match the annotated transcript start site (TSS) or transcript end site (TES) and can thus be either a technical artifact or a bonafide TSS/TES missing in the annotation. All possible combinations of novel splice junctions with pTSS/pTES are shown in Fig. 3B, along with their matching Roman numerals in the associated Venn diagram. Notably, short reads based splicing algorithms such as MAJIQ are generally unable to call transcript start/end sites, so comparison of such cases is not feasible. For this reason, we do not analyze here long reads based transcripts that only involve pTSS/pTES (category VII in Fig. 3B). Nonetheless, we see that when analyzing long reads transcripts with novel splice junctions, most of those involve novel splice sites, and almost all of them also include pTSS/pTES (Fig. 3C).

We examined long reads novel splice variants and further categorized those into splice sites that involve alternative 5'/3' splice sites, intron retention (IR), or new exon (Fig. 3D). We find most of the novel splice variants reported by long reads involve intron retentions, with PacBio and ONT reads yielding 3,255 and 1,754 such cases respectively (Fig. 3E). Novel alternative 5'/3' splice sites or a mix of those were significantly less common, and novel exons were quite rare, only about 10% of the number of IR cases. These results, shown in Fig. 3C, 3E are the average values of three replicates of PacBio and ONT shown in Fig. 2B. Results from different algorithms with different datasets have similar trends except for Bambu (Supplemental Fig. S3, S5). Bambu employs a precision-focused threshold called novel discovery rate (NDR) to approximate the proportion of novel candidates relative to the known transcripts found. Here, NDR of 0.1, the default value, was chosen, which means 10% of all transcripts passing the threshold are novel candidates. Increasing the rate of NDR would increase the intersection sizes in the Upset plots, but it will increase false positive cases as well (Chen et al., 2023). We chose 0.1 because this indicates at least 90% of transcripts with a similar score are annotated, providing an intuitive precision estimation. For completion, we also include pie chart and upset plot results for the original data where long reads have significantly more coverage than short reads data, exhibiting similar trends as well (Supplemental Fig. S4, S6).

Coverage, 3' bias, and GC content lead to differences between short and long reads transcriptome views

Splice site disagreement between PacBio and ONT was previously shown to frequently represent small shifts in splice site calls, possibly due to technical artifacts (Mikheenko et al., 2022). Thus, we hypothesized that this phenomena may also explain some of the significant differences in splice sites detected by short and long reads. To test this hypothesis, we defined 'fuzzy

matching' such that splice sites found by long read algorithms are matched with splice sites reported by MAJIQ from STAR short read alignment if those are within a certain window size apart. Then, if a splice site found by long read algorithms still cannot be matched to short reads within the given window sizes, it is compared to any additional annotated splice sites. Using this 'fuzzy matching' approach, we increased the window size from 3 bp to 8 bp in both 5'/3' splice sites for both long reads technologies and documented the resulting changes in splice sites reported by each technology and how these relate to the annotation (Supplemental Fig. S7). As expected, the number of splice sites matching the annotation and captured by both long read algorithms and MAJIQ ('All', blue bars) increase as the window size grows from 3 to 8, while the most significant decrease is in splice sites that are only found by long reads (magenta bars). These long reads only splice junctions that were 'fuzzy matched' may represent an error in short reads or situations where short reads accurately call splice sites slightly different from the annotation (*e.g.*, NAGNAG). Regardless, the overall effect of applying the 'fuzziness' matching was minute. For example, only 6 splice sites changed in the window size of 8 bp among the 140,000 cases in All of IsoQuant PacBio. Similar patterns are observed for different algorithms, though FLAIR *de novo* cases decreased more compared to the other three algorithms.

As small discrepancies in junction mapping did not offer a significant explanation for the observed differences between short and long-read junction detection, we considered the potential role of nuclear junctions not being exported to the cytosol. Specifically, we hypothesized that nuclear junctions could be lowly abundant in datasets and tend to fall below cutoffs in long reads analysis. While matched short and long nuclear poly(A)+ and cytosolic poly(A)+ datasets are not available at the moment, we were able to examine short read derived junctions of "nuclear poly(A)+" and "cytosolic poly(A)+" from the ENCODE project (Djebali et al., 2012). This analysis offered potential insights into observed long-read patterns. Specifically, we computed junctions and introns for both nuclear and cytosolic poly(A)+ from brain tissue. This analysis revealed a notably higher number of introns in the nuclear fraction compared to the cytosolic fraction and only a small difference in the number of junctions with similar sequencing depth. Specifically, in the nucleus, we identified 464,656 splice junctions and 59,349 intron retention events compared to 446,905 splice junctions and 12,325 intron retention events in the cytosol. Thus, cytosolic poly(A)+ samples exhibit a marginal 3.82% decrease in the detection of junctions compared to nuclear poly(A)+ samples. However, nuclear poly(A)+ samples demonstrate a 381.53% increase in intron retention detection compared to cytosolic poly(A)+ samples. Taken together, and keeping in mind the "zero-sum game" of a fixed sequencing depth, these results suggest that when we sequence nuclear poly(A)+ RNA, we end up with many more of what is termed "retained" or "detained" introns (Barutcu et al., 2022; Boutz et al., 2015). This "sequencing budget spending" on introns and/or the fact some splicing events have not finished processing leads to a slight decrease in number of junctions detected in the cytosol. Thus, while we don't have matched long reads, these results do not point to lowly abundant nuclear junctions as the main driver for the observed discrepancy between short and long-read data.

Next, we turned to assess the effect of coverage. For this we plotted the fraction of splice sites detected by IsoQuant as a function of the number of short reads covering the junction

reported by MAJIQ. For both PacBio and ONT detection was significantly worse for lowly covered splice sites with up to 10 short reads (Fig 4A). Still, even for splice sites with high short read coverage ($>$ 100 reads), PacBio and ONT missed 11% and 18% of the splice sites, respectively, with IsoQuant. Overall, the cumulative distribution function plot shows IsoQuant PacBio detects 74%, and ONT detects 52% of MAJIQ's total amount of junctions (Fig 4A bottom). Among the four long read algorithms we tested, IsoQuant recovers the most junctions, and Bambu recovers the least (Supplemental Fig. S8A). In summary, much of the difference in splice site detection between short and long reads can be attributed to junctions with lower short read coverage, but a significant fraction of highly covered junctions are still not detected by long reads.

The combination of several splice junctions that include or exclude segments in a specific pre-mRNA region form AS 'events' (*e.g.*, cassette exons), and those 'events' in turn serve as the base for both detecting and quantifying transcriptome variations when using short reads technology. This raises the question of how many of such AS events, or LSV in MAJIQ's formulation, that can be quantified by MAJIQ using short reads can also be quantified by matched long reads. We defined an LSV to be quantified by long reads when it had at least 10 reads spanning across the LSV's junctions. This definition matches the default filter on LSV minimal read number to be quantifiable by MAJIQ. Fig 4B shows that using the PacBio data, IsoQuant is unable to quantify 36% of the LSV quantified by MAJIQ when having 10 to 20 short reads, while using the ONT reads, IsoQuant cannot quantify over 65% of the LSV in the same bin. Although the non-quantifiability decreases as the number of short reads per LSV increase, even for LSV with over 100 short reads, PacBio is unable to quantify 8% of the LSV, a fraction similar to the one observed for splice junctions detection above. The fraction of non-quantifiable LSV by ONT is higher, at 18% of LSV with more than 100 reads. These observations are not unique to IsoQuant and were consistent across all four long read algorithms (Supplemental Fig. S9a).

The results described above led us to hypothesize that an important contributing factor for the observed gaps between long and short reads based splicing variations is the inherent 3' to 5' bias of long reads technologies. Poly(A) selected long reads naturally begin from the 3' end. Their length distribution is such that only 5.4% of ONT and 51.6% of PacBio reads in the LRGASP dataset shown in Fig. 4C (right panel) actually span 3000bp or more, which is roughly the median length of human transcripts (Lopes et al., 2021). Furthermore, as noted above, long reads report more novel IR. This means that any splice junctions downstream of those IR events are captured further away from the poly(A) tail and, hence, less likely to be detected. To assess the effect of the 3' bias in long compared to short reads we repeated the analysis of Fig. 4B but with LSV binned by their distance from the long reads 3' end. Fig. 4C (left panel) shows that with ONT IsoQuant can not quantify about 78% of LSV when these are more than 2,500 bp away from the 3' end. However, approximately 38% of LSV close to the 3' end are also non-quantifiable. For PacBio data, which offered longer reads, the distribution of non-quantifiable LSV as a function of 3' end distance is much more flat: 29% of the LSV were non-quantifiable by PacBio reads when those were more than 2,500 bp away and 12% when

close to the 3' end. This 3' to 5' bias trend was consistent in all four algorithms (Supplemental Fig. S9B).

Finally, we explored whether GC content around splice junctions may help explain their differential detection by short and long reads. We selected junctions supported by more than 40 short reads from Fig. 4A (top) and computed GC content within a 100bp window, 50bp flanking each side of the junction for two groups of junctions: Those detected in long reads vs those absent. Given that the distance from the 3' end was a major confounder, we controlled for it in this analysis and compared the two distributions using a Mann–Whitney *U* test. Fig. 4D (top) indicates that short read only junctions have a higher median GC content, and the difference between the groups is more pronounced (lower p-value) for junctions that are missed by long reads closer to the 3' end and for junctions missed by ONT. Accordingly, we see similar trends for minimum free energy (MFE) computed using RNAfold (Gruber et al., 2008) in Fig. 4D (bottom). In contrast, the means are more similar for GA content (Supplemental Fig. S10). These results suggest that local structure associated with higher GC content and lower MFE may be a contributing factor to the differences between short and long-read junction detection, especially for ONT data closer to the 3' end.

Long reads detect many more intron retention events but fewer long introns

Previous sections investigated the differences between short and long reads in terms of splice junctions. However, long reads have a natural advantage in detecting Intron Retention (IR) since a single molecule may be sufficient to call such events. In contrast, IR are not directly detected by short reads and many commonly used short read algorithms such as LeafCutter (Li et al., 2018) do not detect IR, or do not allow *de novo* IR events (*e.g.*, rMATS (Shen et al., 2014)). Short-read algorithms that do detect IR events rely on various filters and thresholds over reads that cross the splice junction into the intron or read coverage across the body of the intron. This makes IR detection from short reads highly dependent on those filtering criteria. In the analysis below, we used MAJIQ's default parameters, which are quite conservative for IR detection (Vaquero-Garcia et al., 2023).

While long reads can give direct evidence for IR events, these detected IR events still raise the question of whether these are reliably detected and biologically significant. To address this, we computed PSI values for long reads IR events reported by different algorithms (see Methods for details). We also plotted the fraction of introns detected by MAJIQ as a function of the number of long reads covering the intron reported by IsoQuant.

Fig. 5A shows the counts of unique IR events reported by MAJIQ's short reads and matching PacBio and ONT long reads by IsoQuant using the LRGASP data. Here, IsoQuant PacBio reports a staggering number of $\sim$10K unique IR events, compared to about 6.3K with ONT, and 2.4K unique MAJIQ short read based IR events. Investigating these IR event sets, we find MAJIQ's unique set to include longer introns (Fig. 5B). This result is to be expected given the limitation of long reads overall length and the 3' bias discussed above. Finally, when we assess

the relation between IR events detected by IsoQuant using either PacBio or ONT reads and IR events detected by MAJIQ from short reads, we find no direct relation between detection and IR PSI value or number of reads. This result agrees with previous reports regarding the limitations in IR detection from short reads (Steijger et al., 2013). As retained introns have been reported to have certain characteristics, including higher GC content (Galante et al., 2004), we assessed the GC content of retained introns and did not find a specific enrichment for low/high values in short and long reads (Supplemental Fig. S13). Of note, the results shown in Fig. 5A-C, are the average values of three replicates in LRGASP data, and similar trends were observed when we used different long reads algorithms and GTEx data (Supplemental Fig. S11, S12).

A Unified Visualization of Short and Long Read-Seq with VOILA

The comparative analysis of transcriptome variations clearly pointed to the complementarity between short and long reads RNA sequencing. In order to facilitate unified visualization and downstream analysis of short and long reads, we developed the VOILA V3 package. VOILA V3 is able to combine MAJIQ's short reads splicing analysis with GTF output files from any long read algorithm. An illustrative example is shown in Fig. 6, where we ran MAJIQ with short reads and IsoQuant with PacBio long reads on one of the human cell line samples in LRGASP (Pardo-Palacios et al., 2021). VOILA v3 can show a short reads based gene splicegraph (first row), a unified short and long reads gene splicegraph (second row), and a list of transcripts found by only long reads (row three and below). Each color for splice junction and intron retention shows what source supports it (short, long, and annotation), with colors matching those in Fig. 2B. Users can filter their data by several criteria, including which source the splice graph element came from, read coverage over junctions, LSV types, and complexity. Distributions over PSI for both short and long reads are represented using violin plots as the dotted black boxes in the first and second rows (see "Methods"). For long reads, the read number per transcript, junction/IR, and exon are displayed. Also, the visualization displays the TSS/TES of each transcript. For a unified splicegraph visualization, junction/IR read numbers for both short and long reads are stated, and TSS/TES are not represented.

To demonstrate the usage of VOILA v3, we show the splicing analysis for the splicing factor *SRSF11*. The black dotted box of the *SRSF11* gene in the unified splicegraph (second row) highlights alternative exons that introduce ultra-conserved premature termination codon (PTC) that induce nonsense mediated decay (NMD). Such PTC introducing exons are known regulatory feature controlling many RBPs, especially those in the serine/arginine protein family (Ni et al., 2007; Lareau et al., 2007). Both short and long reads support the identification of this important regulatory mechanism, but some differences can be observed. Short reads generally detect more diverse splicing patterns with more splice junctions coming out of exon 5, resulting in more dispersed violin plots. Both short and long reads detect multiple unannotated IR events in this region. Notably, only long reads capture an extra IR event and the associated full isoform. These results are inline with our more general transcriptome wide analysis as long reads are more likely to have a bias towards short isoforms (skipping the exons) yet can more

easily detect intron retention events. Regardless of these differences, it is important to note the complex splicing patterns that emerge from this analysis. Specifically, in the context of NMD triggering alternative splicing events concerning serine/arginine proteins, the prevailing notion is that poison exon inclusion is the primary contributor to NMD. However, evidence from both short and long reads suggests that IR in this region may also play a role in controlling *SRSF11* expression.

Discussion

The work presented here was motivated by the rapid adaptation of long-read RNA sequencing. Our labs, as many others, found unique advantages to using long reads. Specifically, long reads allow researchers to resolve the relation between separate AS events along a gene splice graph and assess which isoforms include specific splice junctions of interest (Sharon et al., 2013). Resolving full isoforms is of particular importance when trying to detect, for example, novel immunotherapy targets (Zheng et al., 2021). Similarly, the ability to unequivocally and with high sensitivity detect intron retention and overcome mappability limitations over repetitive regions are significant advantages of long reads technologies. However, the qualitative limitations of low coverage and higher error rates, combined with the fact that extensive short reads data already exists, led us to compare and contrast the short and long reads based on transcriptome variations detection and quantification. Specifically, we formulated three questions as the base for this study: How to compare/contrast transcriptome variations detected by short/long reads, is there utility in combining those, and if so can we develop a method for such a combined analysis?

To address the first question, we formulated a set of metrics by which any long reads technology or algorithm can be compared to MAJIQ's short reads based splicing analysis. We showed there is a significant gap in both detection and quantification of splicing variations. Short reads based analysis of matched datasets revealed 30% more splice junctions, with PacBio detecting approximately 10% more than ONT. 11-18% junctions with high short reads coverage (> 100) were missed by PacBio and ONT, and 10% to 30% of splice junctions with PSI $> 20\%$ were missed by PacBio and ONT respectively. As for the ability to quantify local splicing variations, we found a clear 3' to 5' bias with 12-29% (PacBio) and 38-78% (ONT) of the LSV quantified by MAJIQ from short reads were unquantifiable by IsoQuant as a function of the distance from the poly(A) tail. This phenomena is not unique to IsoQuant and seem to reflect the limited length of the long reads, especially those in the ONT datasets we analyzed here. We also showed that local structure associated with higher GC content and lower MFE may be a contributing factor to the differences between short and long reads. On the other hand, IsoQuant PacBio and ONT detected significantly more unique IR events, respectively 10K and 6.3K, compared to $\sim$2.4K for MAJIQ short reads IR. Finally, trying to relax the default parameters of the long reads algorithms had the expected outcome: It led to increased junction detection at the expense of a suspected increase in false positives but had little effect on the overall comparative results

with respect to matched short reads (Supplemental Fig. S14). These results point to the complementary nature of currently available short and long-read assays, and we developed a software package, MAJIQ-L, to enable such a combined analysis.

We believe the significance of the work presented here stems from several factors. First, our results clearly demonstrate the benefit of a combined short and long reads analysis. Second, we are painfully aware that long reads RNA sequencing is a fast evolving technology with more algorithms and improved protocols or assays frequently announced by researchers or companies. Thus, a second significant component of this work is that the pipeline we developed here can be used to independently assess any newly released long-read algorithm or technology either by the developers themselves or interested users. Naturally, the introduction of new results or new technology tends to suffer from a strong confirmation bias, focusing on what is new or improved. For transcriptomics, researchers may consequently conclude long reads subsume short reads data. However, the picture we draw here is more complex. Thus, we view the ability to easily and independently compare technologies' output using our pipeline and the results we already provide as key for genomics researchers to make informed decisions. A third significant component of this work is MAJIQ-L with the VOILA V3 visualization package, which allows researchers to perform integrated long and short reads analysis.

Our analysis points to several key conclusions. First, to answer many scientific questions, researchers may be best served by short and long reads combined analysis, possibly utilizing existing short reads data, and a two-stage approach - initial discovery with short reads, then focused targeted sequencing with long reads of specific genes of interest. The higher costs of long-read sequencing further point to the utility of such an approach. Moreover, our results clearly show that costly deeper long-read sequencing alone may not be an effective solution. Rather, researchers who want a more complete transcriptomic view from long reads should aim to extend the length of the long reads. We note that this result is directly coupled with the library preparation protocols typically used in the application of each sequencing technology: Random primed, dUTP stranded protocol for Illumina, vs a template switching one, which is the standard long reads cDNA protocol. While direct RNA sequencing protocols are available and have been shown to avoid RT artifacts such as 'falsitrons' (Schulz et al., 2021), the length of the consequent reads remains similar. Thus, addressing the biases introduced by different library preparation protocols to improve length distribution and coverage across the entire body of transcripts remains an important direction for future technology improvement.

While our results highlight read length and coverage across the full body of transcripts as important for future improvement, much work has been dedicated to reducing the error rate in long reads. However, the effect of sequencing errors on transcriptome analysis can be complex and context-dependent. Correction algorithms like IsoQuant do not alter individual bases but focus on improving spliced alignment accuracy, making it challenging to measure error rates after correction (Prjibelski et al., 2023). Additionally, the performance of long-read technologies in transcript discovery remains comparable despite differences in error rates. Thus, while error rates provide valuable insight into sequencing quality, they may not directly correlate with the efficacy of transcriptome analysis. Evaluating the effectiveness of long-read RNA-seq al-

gorithms' error correction strategies, therefore, remains a challenge that is not addressed in this current work.

There are several limitations and possible extensions to this work. First, our comparative analysis focused solely on transcriptome variations reported by short and long reads. As such, several important questions remain open. These include the comparative assessment of gene expression estimates, the effect of error corrections on long reads results discussed above, and assessing the ability to perform transcripts reconstruction, including dedicated methods that can use both short and long reads such as StringTie (Shumate et al., 2022). Some of those questions were addressed in recent studies, such as expression estimation (Chen et al., 2023) and highlighting issues with the many pTES/TSS sites reported by long reads (Calvo-Roitberg et al., 2023). We also acknowledge that the data used in this study did not include Unique Molecular Identifiers (UMIs). UMI would potentially offer a more definitive solution to differentiate between true unique fragments and PCR duplicates.

With respect to the tool we developed, MAJIQ-L with VOILA V3 allows for an integrated splicing analysis of short and long reads but does not include a unified probabilistic model for those. Such a unified model could potentially further improve isoform-level quantification. Future extensions can also include allele-specific splicing and detection of variants directly from the long reads data. We are excited to explore these directions in the future and hope the combined comparative analysis pipeline and results, along with the MAJIQ-L package, would be highly useful for Genomics researchers focused on transcriptome variations.

Methods

Processing coverage differences between short and long reads

As the total number of bases sequenced across short and long-read technologies are different, we used Seqtk (`https://github.com/lh3/seqtk`) by sub-sampling from either short or long reads datasets before providing them as inputs to MAJIQ and long read tools. This step allows all platforms to have similar coverage for a fair comparative analysis. The coverage summary of pre and post-sub-sampled number of bases for each dataset can be found in Supplementary Table 1-3.

MAJIQ's short reads splicing analysis

We used STAR (v2.7.10b) (Dobin et al., 2013) to align short RNA-seq reads, performing a two-step gapped alignment to GRCh38, GENCODE release 42. MAJIQ then combined the annotation and aligned reads to build a splicegraph for each gene, including *de novo* elements such as junctions, intron retention, and exons. The resulting splicegraphs are used as MAJIQ-L's input along with long read tools' output in GTF file format for the downstream comparative analysis.

GTF file output files from long read tools

Long RNA-seq reads were mapped to GRCh38, GENCODE release 42, using minimap2 (v2.24) (Li, 2018) in splice mode. All long-read tools were provided with the same BAM file, reference genome, and reference annotation. IsoQuant was run with the default parameters with the appropriate data type using '$-data\_type$' option. ESPRESSO and FLAIR were launched with the default parameters in 30 threads. As Bambu outputs all reference transcripts, including unexpressed ones, we filtered out all transcripts with read count values smaller than 1 as the authors recommended. Software versions and command line options are in Supplemental Table S4.

Inferring posterior distribution in long reads PSI per junction

The likelihood function over PSI Ψ_j for a junction j is modeled as a binomial distribution, where r_j denotes long reads aligned to each junction j in the LSV:

$$r_j \sim Binomial\left(\sum_{j \in LSV} r_j, \Psi_j\right) \tag{1}$$

As in MAJIQ's short reads model, we set a prior distribution on PSI which favors either high or low PSI values, which can be generalized by the Jeffrey's prior for an LSV with j junctions:

$$\Psi_j \sim Beta\left(\frac{1}{J}, 1 - \frac{1}{J}\right) \tag{2}$$

Since this prior is conjugate to the binomial distribution, our posterior distribution of Ψ_j given the observed number of reads are the following:

$$\Psi_j | \{r_j^{'} : j^{'} \in LSV\} \sim Beta\left(\frac{1}{J} + r_j, 1 - \frac{1}{J} + \sum_{j^{'} \neq j} r_j^{'}\right) \tag{3}$$

The resulting distributions over PSI for both short and long reads are shown as violin plots by the VOILA visualization package.

Data access or Software availability

The matched human cell line dataset can be accessed through the LRGASP `https://www.gencodegenes.org/pages/LRGASP/`. The GTEx v9 heart atrial appendage is available at GTEx website `https://www.gtexportal.org/home/datasets`. MAJIQ 2.5 provides the new VOILA visualization features of MAJIQ-L, which is available at `https://majiq.biociphers.org/app_download/` with a matching user support group at `https://groups.google.com/g/majiq_voila?pli=1`. All scripts used for data analysis are available at `https://bitbucket.org/biociphers/majiq-l/src/main/`.

Competing interest statement

The MAJIQ software used in this study is available for licensing for free for academics or for a fee for commercial usage. Some of the licensing revenue by the University of Pennsylvania goes to members of the Barash lab including Y.B., SW.H., and S.J. Otherwise, all authors declare they have no competing interests.

Acknowledgements

We thank Danielle Gutman, Matthew Gazzara, and Nathaniel Islas for helpful comments on the manuscript. This work was supported by NIH U01 CA232563 (Y.B. and A.T.T), R01 LM013437 (Y.B), and a CureBRCA grant (Y.B). Y.B. conceived the project. SW.H., S.J, and Y.B. developed and tested the methodology for MAJIQ-L. A.T.T. provided continuous feedback on the development of the project. SW.H. and Y.B. wrote the final manuscript. All authors read and approved the final manuscript.

References

Amarasinghe SL, Su S, Dong X, Zappia L, Ritchie ME, Gouil Q. 2020. Opportunities and challenges in long-read sequencing data analysis. Genome Biol 21: 30.

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 110: E4821–E4830.

Bagashev A, Loftus JP, Wakefield C, Wertheim G, Hurtz C, Carroll MP, Stegmaier K, Pikman Y, Tasian SK. 2021. Alisertib synergistically strengthens the anti-leukemia activity of Venetoclax in TCF3-Hlf B-ALL. Blood 138: 705.

Barutcu AR, Wu M, Braunschweig U, Dyakov BJA, Luo Z, Turner KM, Durbic T, Lin ZY, Weatheritt RJ, Maass PG, et al. 2022. Systematic mapping of nuclear domain-associated transcripts reveals speckles and lamina as hubs of functionally distinct retained introns. Mol Cell 82: 1035–1052.

Baruzzo G, Hayer KE, Kim EJ, Di Camillo B, FitzGerald GA, Grant GR. 2017. Simulation-based comprehensive benchmarking of RNA-seq aligners. Nat Methods 14: 135–139.

Boutz PL, Bhutkar A, Sharp PA. 2015. Detained introns are a novel, widespread class of post-transcriptionally spliced introns. Genes Dev 29: 63–80.

Calvo-Roitberg E, Daniels RF, Pai AA. 2023. Challenges in identifying mRNA transcript starts and ends from long-read sequencing data. bioRxiv 2023–07.

Chen Y, Sim A, Wan YK, Yeo K, Lee JGX, Ling MH, Love MI, Göke J. 2023. Context-aware transcript quantification from long-read RNA-seq data with Bambu. Nat Methods 20: 1187–1195.

David JK, Maden SK, Wood MA, Thompson RF, Nellore A. 2022. Retained introns in long RNA-seq reads are not reliably detected in sample-matched short reads. Genome Biol 23: 240.

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.

Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F, et al. 2012. Landscape of transcription in human cells. Nature 489: 101–108.

Dvinge H, Kim E, Abdel-Wahab O, Bradley RK. 2016. RNA splicing factors as oncoproteins and tumour suppressors. Nat Rev Cancer 16: 413–430.

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.

Foord C, Hsu J, Jarroux J, Hu W, Belchikov N, Pollard S, He Y, Joglekar A, Tilgner HU. 2023. The variables on RNA molecules: concert or cacophony? Answers in long-read sequencing. Nat Methods 20: 20–24.

Gao Y, Wang F, Wang R, Kutschera E, Xu Y, Xie S, Wang Y, Kadash-Edmondson KE, Lin L, Xing Y. 2023. ESPRESSO: Robust discovery and quantification of transcript isoforms from error-prone long-read RNA-seq data. Sci Adv 9: eabq5072.

Galante PA, Sakabe NJ, Kirschbaum-Slager N, De Souza SJ. 2004. Detection and evaluation of intron retention events in the human transcriptome. RNA 10: 757–765.

Glinos DA, Garborcauskas G, Hoffman P, Ehsan N, Jiang L, Gokden A, Dai X, Aguet F, Brown KL, Garimella K, et al. 2022. Transcriptome variation in human tissues revealed by long-read sequencing. Nature 608: 353–359.

Gruber AR, Lorenz R, Bernhart SH, Neuböck R, Hofacker IL. 2008. The Vienna RNA websuite. Nucleic Acids Res 36: W70–W74.

Hardwick SA, Joglekar A, Flicek P, Frankish A, Tilgner HU. 2019. Getting the entire message: progress in isoform sequencing. Front Genet 10: 709.

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: 1–13.

Kovaka S, Ou S, Jenike KM, Schatz MC. 2023. Approaching complete genomes, transcriptomes and epi-omes with accurate long-read sequencing. Nat Methods 20: 12–16.

Lareau LF, Inada M, Green RE, Wengrod JC, Brenner SE. 2007. Unproductive splicing of SR genes associated with highly conserved and ultraconserved DNA elements. Nature 446: 926–929.

Li YI, Knowles DA, Humphrey J, Barbeira AN, Dickinson SP, Im HK, Pritchard JK. 2018. Annotation-free quantification of RNA splicing using LeafCutter. Nat Genet 50: 151–158.

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

Lopes I, Altab G, Raina P, De Magalhães JP. 2021. Gene size matters: an analysis of gene length in the human genome. Front Genet 12: 559998.

Lucas MC, Novoa EM. 2023. Long-read sequencing in the era of epigenomics and epitranscriptomics. Nat Methods 20: 25–29.

Marx V. 2023. Method of the year: long-read sequencing. Nat Methods 20: 6–11.

Mikheenko A, Prjibelski AD, Joglekar A, Tilgner HU. 2022. Sequencing of individual barcoded cDNAs using Pacific Biosciences and Oxford Nanopore Technologies reveals platform-specific error patterns. Genome Res 32: 726–737.

Nellore A, Jaffe AE, Fortin JP, Alquicira-Hernández J, Collado-Torres L, Wang S, Phillips RA III, Karbhari N, Hansen KD, Langmead B, et al. 2016. Human splicing diversity and the extent of unannotated splice junctions across human RNA-seq samples on the Sequence Read Archive. Genome Biol 17: 1–14.

Ni JZ, Grate L, Donohue JP, Preston C, Nobida N, O'Brien G, Shiue L, Clark TA, Blume JE, Ares M. 2007. Ultraconserved elements are associated with homeostatic control of splicing regulators by alternative splicing and nonsense-mediated decay. Genes Dev 21: 708–718.

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.

Prjibelski AD, Mikheenko A, Joglekar A, Smetanin A, Jarroux J, Lapidus AL, Tilgner HU. 2023. Accurate isoform discovery with IsoQuant using long reads. Nat Biotechnol 1–4.

Rivera OD, Mallory MJ, Quesnel-Vallières M, Chatrikhi R, Schultz DC, Carroll M, Barash Y, Cherry S, Lynch KW. 2021. Alternative splicing redefines landscape of commonly mutated genes in acute myeloid leukemia. Proc Natl Acad Sci 118: e2014967118.

Schulz L, Torres-Diz M, Cortés-López M, Hayer KE, Asnani M, Tasian SK, Barash Y, Sotillo E, Zarnack K, König J, et al. 2021. Direct long-read RNA sequencing identifies a subset of questionable exitrons likely arising from reverse transcription artifacts. Genome Biol 22: 190.

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

Shen S, Park JW, Lu ZX, Lin L, Henry MD, Wu YN, Zhou Q, Xing Y. 2014. rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. Proc Natl Acad Sci 111: E5593–E5601.

Shumate A, Wong B, Pertea G, Pertea M. 2022. Improved transcriptome assembly using a hybrid of long and short reads with StringTie. PLoS Comput Biol 18: e1009730.

Steijger T, Abril JF, Engström PG, Kokocinski F, Hubbard TJ, Guigó R, Harrow J, Bertone P. 2013. Assessment of transcript reconstruction methods for RNA-seq. Nat Methods 10: 1177–1184.

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 28: 396–411.

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.

Vaquero-Garcia J, Barrera A, Gazzara MR, Gonzalez-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.

Vaquero-Garcia J, Aicher JK, Jewell S, Gazzara MR, Radens CM, Jha A, Norton SS, Lahens NF, Grant GR, Barash Y. 2023. RNA splicing analysis using heterogeneous and large RNA-seq datasets. Nat Commun 14: 1230.

Wyman D, Balderrama-Gutierrez G, Reese F, Jiang S, Rahmanian S, Forner S, Matheos D, Zeng W, Williams B, Trout D, et al. 2019. A technology-agnostic long-read analysis pipeline for transcriptome discovery and quantification. bioRxiv 672931.

Yuan G, Wang F, Wang R, Kutschera E, Xu Y, Xie S, Wang Y, Kadash-Edmondson KE, Lin L, Xing Y. 2023. ESPRESSO: Robust discovery and quantification of transcript isoforms from error-prone long-read RNA-seq data. Sci Adv 9: eabq5072.

Zheng S, Gillespie E, Naqvi AS, Hayer KE, Ang Z, Torres-Diz M, Quesnel-Vallieres M, Hottman DA, Bagashev A, Chukinas J, et al. 2021. Modulation of CD22 protein expression in childhood leukemia by pervasive splicing aberrations: implications for CD22-directed immunotherapies. Blood Cancer Discov 1–14.

Figures

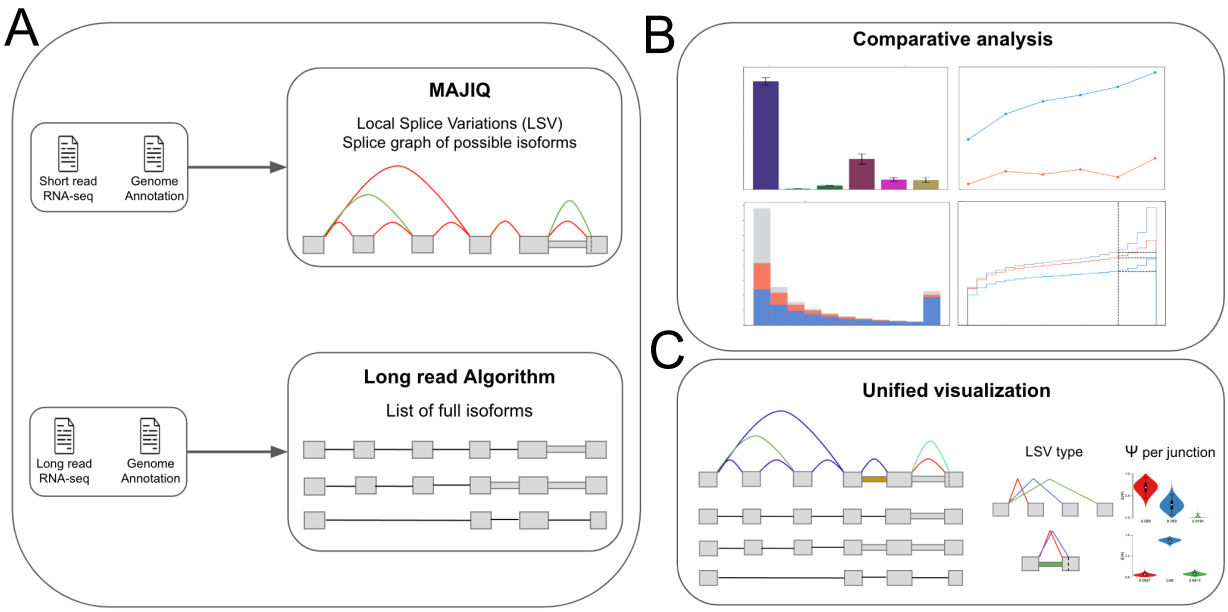


Fig. 1: MAJIQ-L overview. **(A)** Input: Matched short and long reads RNA-sequencing data with genome annotation. Short reads are mapped with an aligner, then passed to MAJIQ, producing a splice graph with a matching set of LSV per gene. Long reads are processed with a long read algorithm (*e.g.*, IsoQuant) to produce isoforms counts. **(B)** The outputs from the two RNA sequencing sources along with the annotation are compared. Each splice junction or intron retention is assigned to each of the possible six categories depending on which subset of the three sources support it. Various statistics regarding the location, coverage, and overlap of those elements are computed to compare the three sources and explore the source of discrepancy (see main text). **(C)** MAJIQ-L includes a unified visualization package, VOILA v3, for downstream splicing analysis. It allows users to see which splice junction maps to which isoforms and where the different sources of information agree or disagree in detection and quantification.

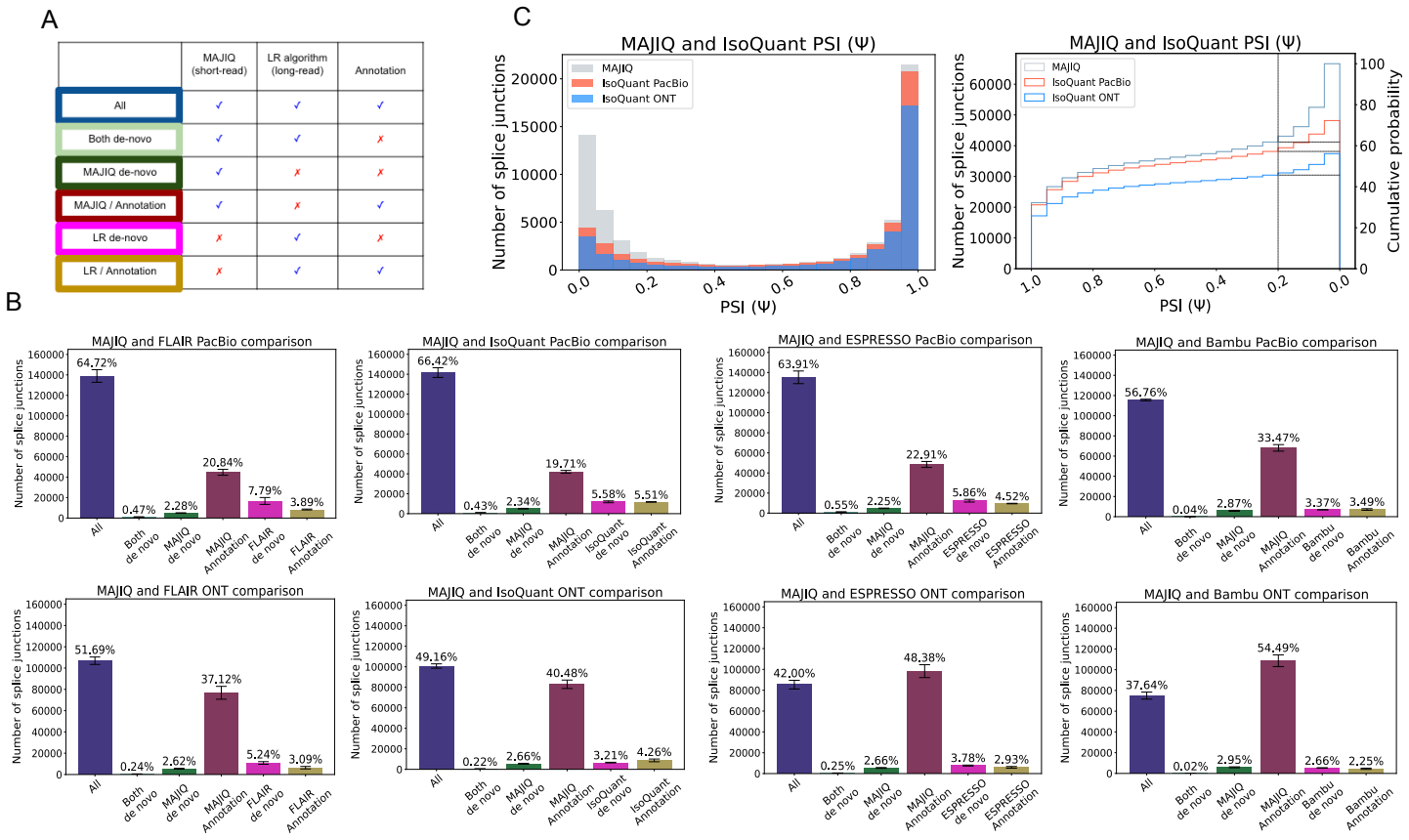


Fig. 2: Splice junctions comparative analysis. **(A)** Any detected splice junction can fall into one of six categories, each represented by a color, depending which of the three sources of information (short reads, long reads, annotation) support it. **(B)** Bar charts corresponding to the aforementioned six categories. Mean and standard error bars are computed using matched datasets from three replicates of human cell-line sequenced by LRGASP (Pardo-Palacios et al., 2021). This data includes short reads processed by STAR and MAJIQ, long reads from PacBio and ONT assays, and four long read algorithms used to process the long reads data. **(C)** Taking the splice junctions reported in (B) by MAJIQ (green) and assessing the number of those also identified when using PacBio (tomato) or ONT (blue) long reads, as a function of the PSI values. Here IsoQuant was used for long reads data. Note that if a junction appears in multiple LSV, the lowest PSI values are chosen (x-axis). The graph on the right is the CDF for the histogram shown on the left. Dashed lines denote splice junctions with a PSI of 20% or more.

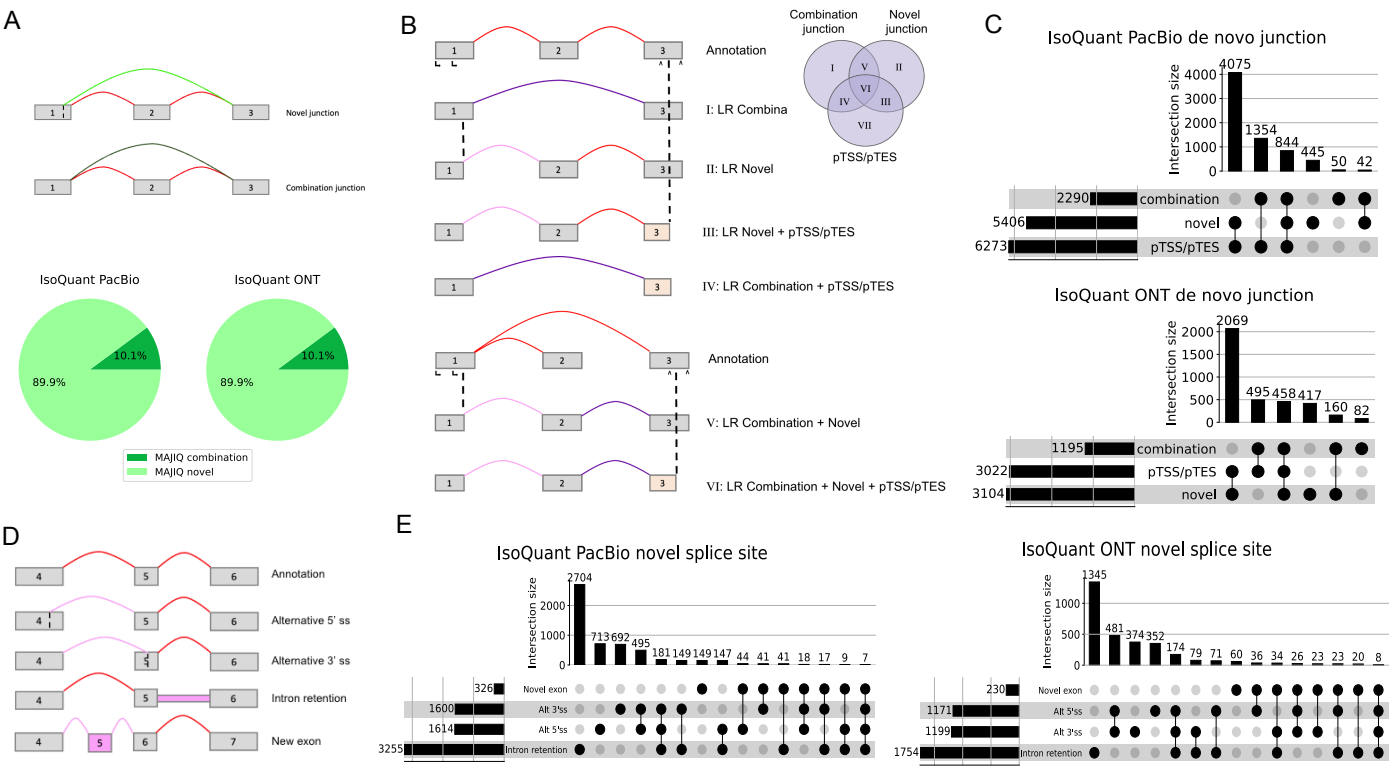


Fig. 3: Analysis of *de novo* elements. **(A)** Short reads only *de novo* splice junctions reported by MAJIQ (green junctions between exon 1 and 3 in the splice graphs) can be classified as those involving novel splice sites (light green) or a novel combination of known splice sites (dark green). Red junctions correspond to annotated ones. The pie chart shows that compared to long reads processed with IsoQuant, $\sim 90\%$ of MAJIQ *de novo* splice junctions involve novel splice sites. **(B)** Representative cartoon examples for six different categories of long reads *de novo* transcript variations. Novel combination junction (dark purple), junctions involving novel splice sites (light purple), and junctions supported by annotation (red) are the same as in (a). putative start or end (pTSS/pTES) (light yellow), or partial exons, represent cases when the transcript start site or transcript end sites do not match those in the annotation, which happens in the first or last exon of the transcript. We note that cases involving *only* pTSS/pTES (class VII in the Ven Diagram) are not included in downstream analysis as those are not handled by MAJIQ or similar short reads based splicing algorithms so can not be directly compared. **(C)** Breakdown of all cases involving *de novo* junctions reported by IsoQuant using either PacBio (top) or ONT (bottom) long reads. Notably, almost all of those cases also include pTSS/pTES. **(D)** Representative cartoon examples for the types of novel splice variations (pink) that a novel splice variant in long reads can introduce compared to the annotation (top graph, red junctions). **(E)** Breakdown of long reads novel splice junctions (light purple in (B)) into the four different categories shown in (D) when using IsoQuant to analyze PacBio (left) and ONT (right) matched reads.

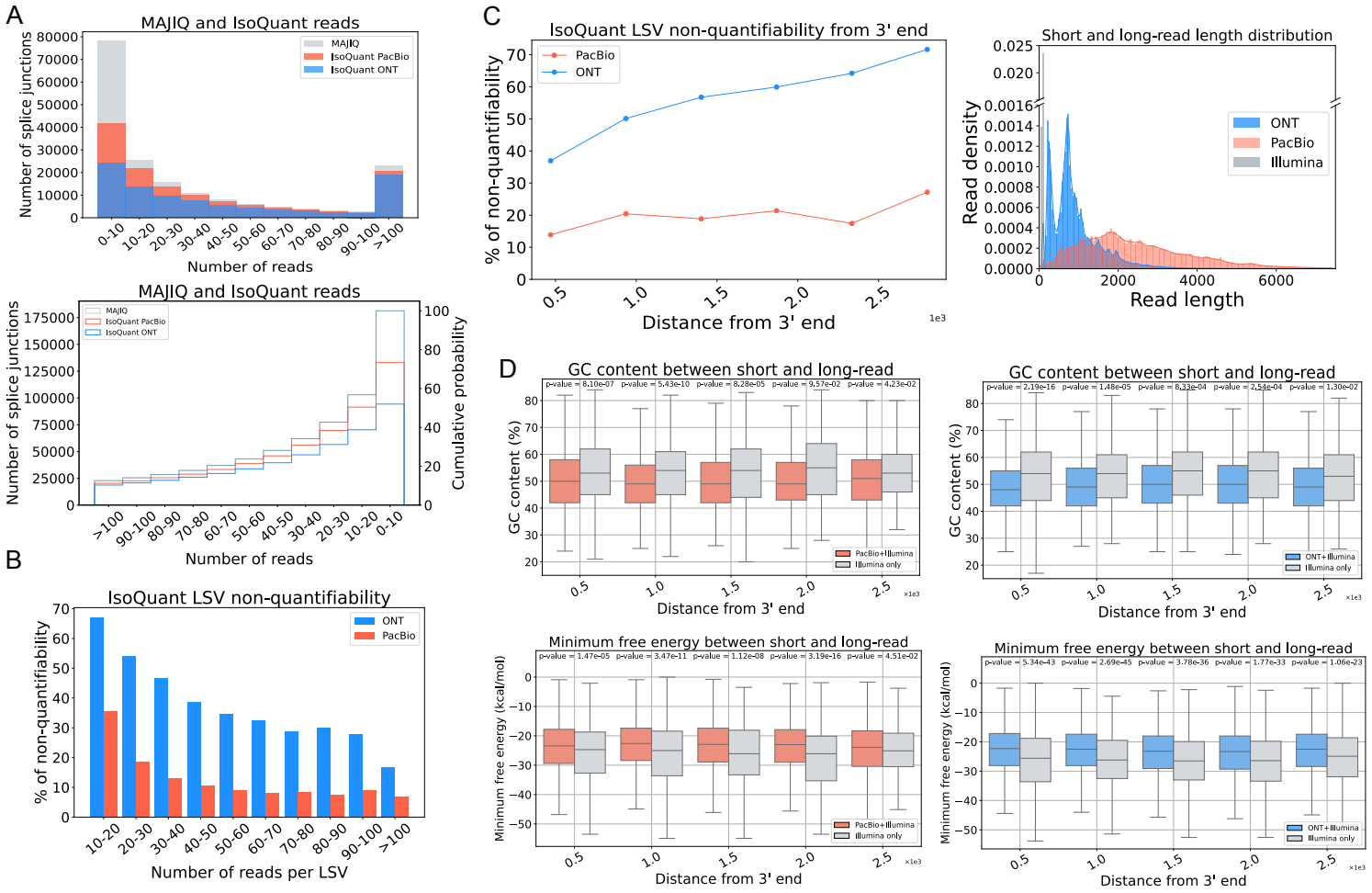


Fig. 4: Analysis of sources of discrepancy between short and long reads based transcriptome variations. **(A)** The number of MAJIQ's splice junctions (grey) identified by IsoQuant using PacBio (tomato) or ONT (blue) as a function of the number of short reads covering the junctions. The histogram on the left and CDF on the right show the number of splice junctions (y-axis) as a function of read number (x-axis). **(B)** Bar plots showing the fraction of LSV reported by MAJIQ's short reads analysis, which were 'non-quantifiable' by IsoQuant using PacBio (orange) and ONT (light blue) matched long reads data. Here a 'quantifiable' LSV require at least 10 reads covering its respective junctions. Of note, a substantial fraction of LSV remain unquantifiable by long reads even for those with extremely high short read coverage (>100 reads). **(C)** Same plot as in (B) for the fraction of non-quantifiable LSV by long reads data, but here as a function of distance from transcript 3' end. When LSV involved transcripts with multiple 3' ends, the shortest distance was used as a conservative estimate. The length distribution of short and long reads in the LRGASP dataset used in all the above sub-figures. **(D)** Boxplots showing GC content and Minimum free energy (MFE) across various distances from the transcript 3' end for junctions with over 40 illumina reads in (a) that are only detected

by short reads (grey) or also detected by long reads (red - PacBio, blue - ONT). Each boxplot represents the GC content and MFE (y-axis) as a function of the distance from 3' end (x-axis). The median is denoted by the horizontal line in each box, the upper and lower quartiles are denoted by the box, and the whiskers show points that lie within 1.5 IQRs of the lower and upper quartiles. P-values were calculated using the Mann–Whitney *U* test. Note that (a)-(d) are averaged across the three LRGASP data replicates.

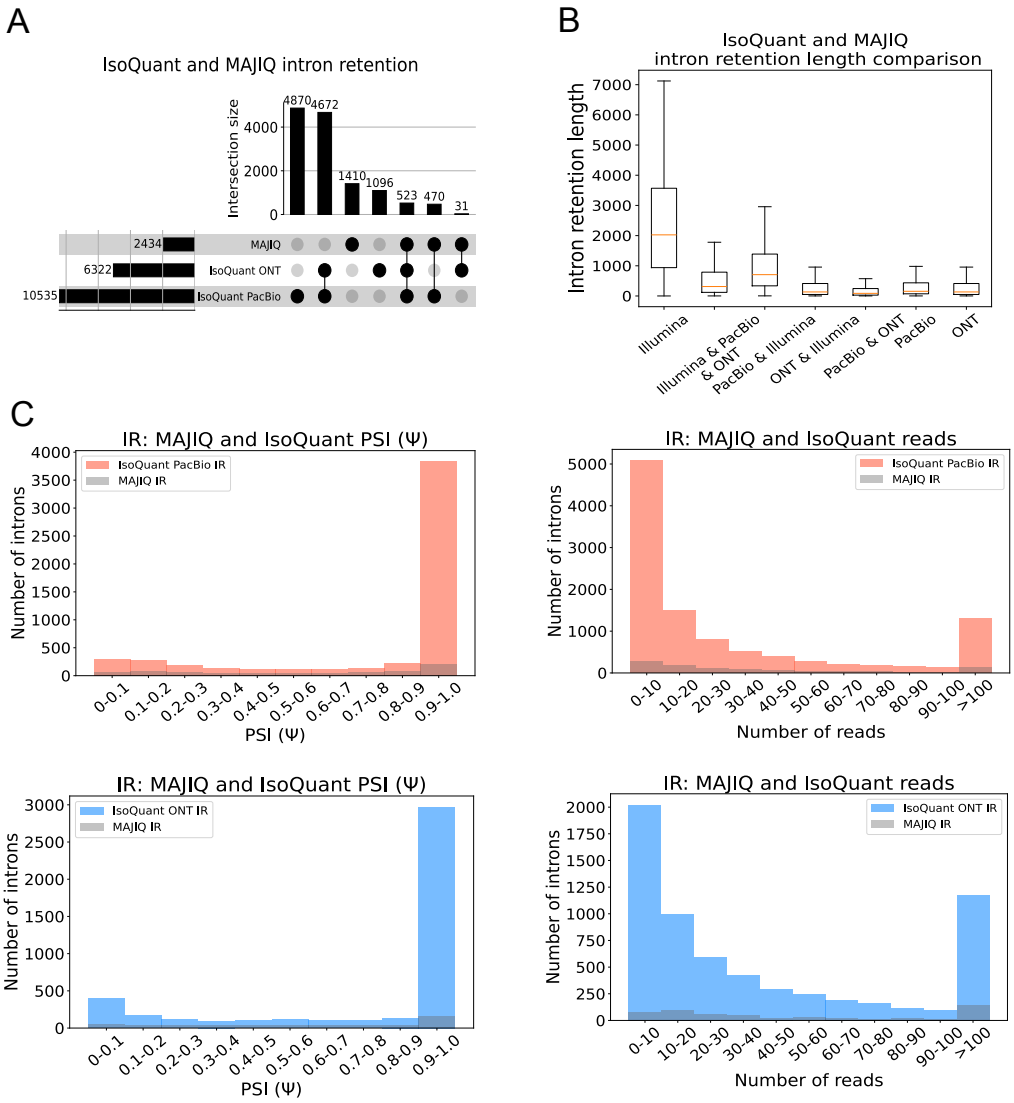


Fig. 5: Comparison of intron rention (IR) events. **(A)** Upset plot showing overlap and total IR events reported by MAJIQ from short reads and IsoQuant using PacBio or ONT matched long reads (LRGASP dataset). **(B)** Boxplots showing IR length distribution across seven categories in Fig 5a. Each boxplot represents the IR length (y-axis) in each category (x-axis). The median is denoted by the yellow line, the upper and lower quartiles are denoted by the box, and the whiskers show points that lie within 1.5 IQRs of the lower and upper quartiles. The number of events in each category corresponds to those in (A). **(C)** Introns reported by IsoQuant using PacBio (tomato) or ONT (blue) and how many of those were also identified by MAJIQ (grey) as a function of the PSI values (left) or the number of reads covering the intron (right). For PSI, only IR events with at least 10 reads were considered and if an intron appears multiple times, the lowest PSI value is chosen for it.

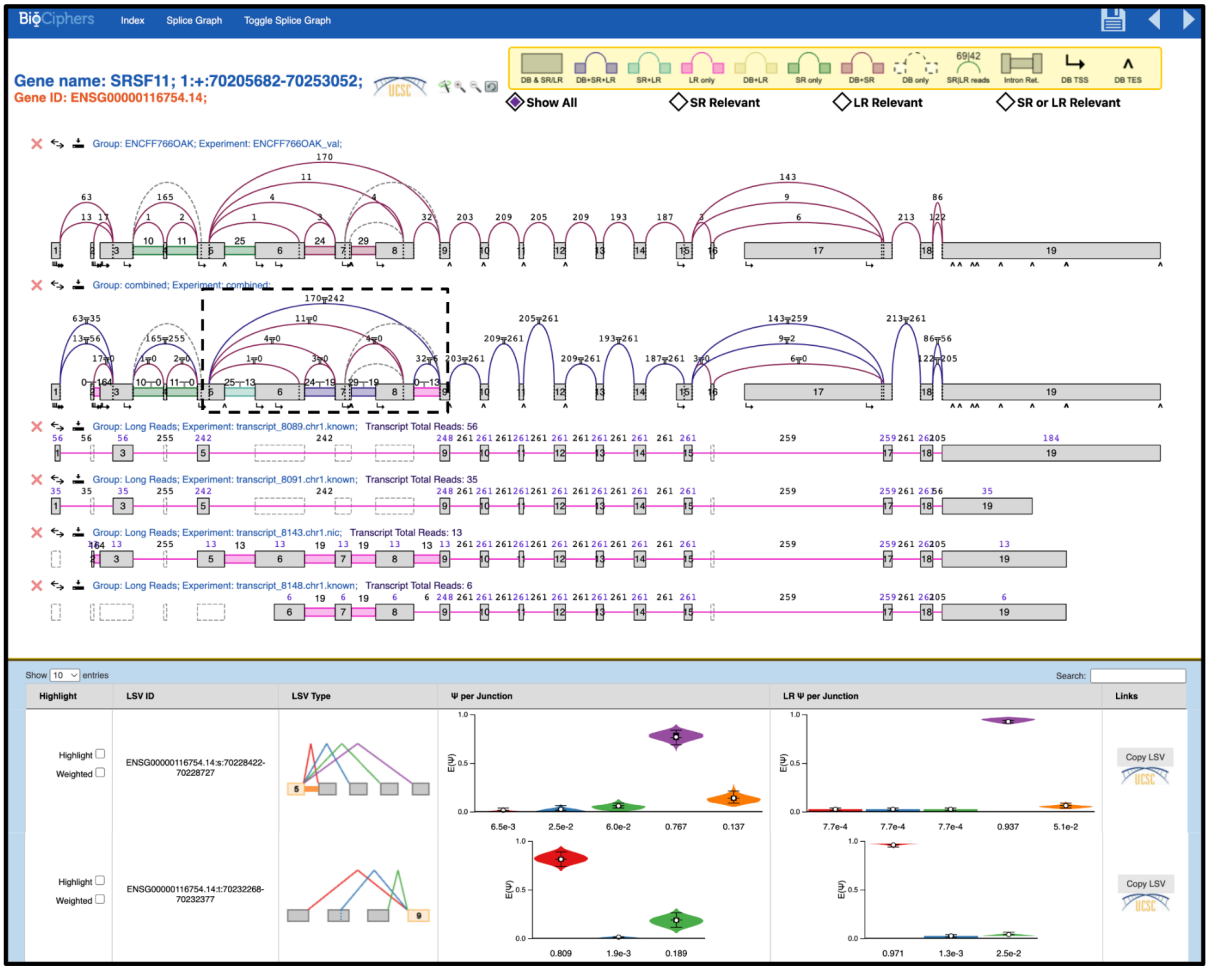
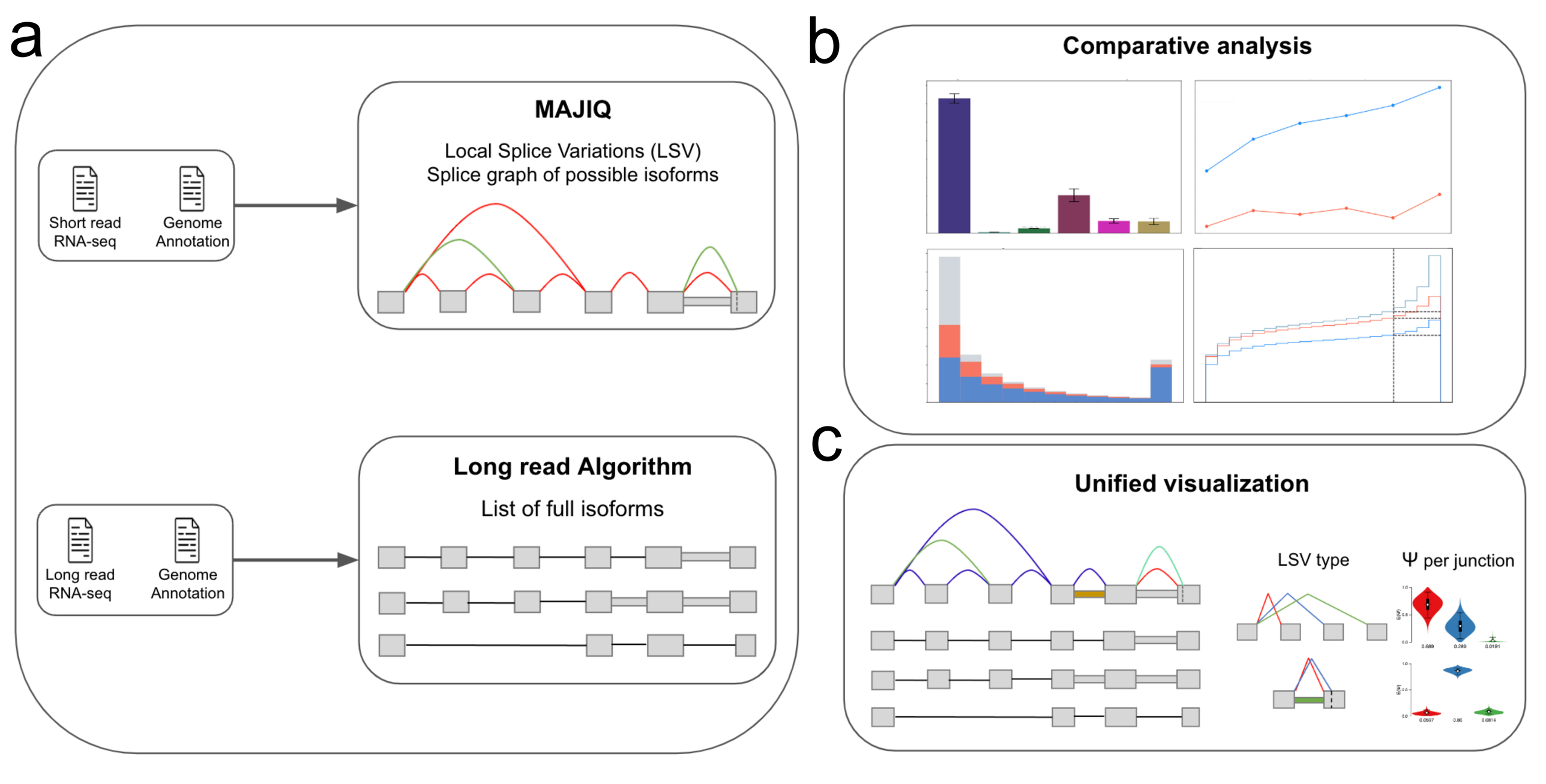


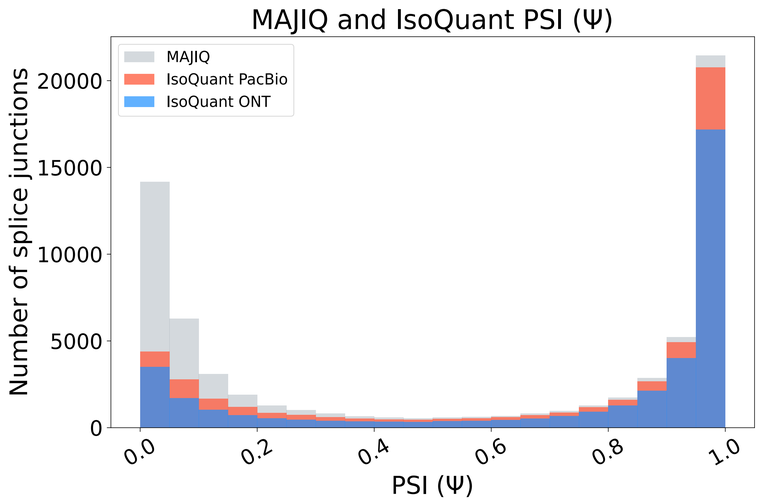
Fig. 6: MAJIQ-L integrative analysis of splicing variations using short and long reads. Snapshot of the VOILA v3 interactive visualization of MAJIQ-L output for the *SRSF11* splice factor using the LRGASP data, including short reads processed by MAJIQ and PacBio reads processed by IsoQuant. The top portion shows gene information and filtering criteria between short and long reads as well as the short reads splicegraph, a unified splicegraph, and a list of transcripts reported by IsoQuant for *SRSF11*. Read numbers for each transcript and the marginal count for each specific elements (junctions, introns, and exons) are included. In the unified splice graph view read count for junctions and introns are shown for both short (left) and long (right) reads, separated by a *T* sign. Note that pTSS/pTES are only shown in transcripts found by long reads. The bottom portion shows distributions of E(Ψ) values of LSV that both short and long reads find, displayed as a violin plot for the exon 5 source LSV and exon 9 target LSV in the black dotted box. The source of the individual junction can be highlighted by hovering the cursor over the junction and multiple filters can be applied interactively.

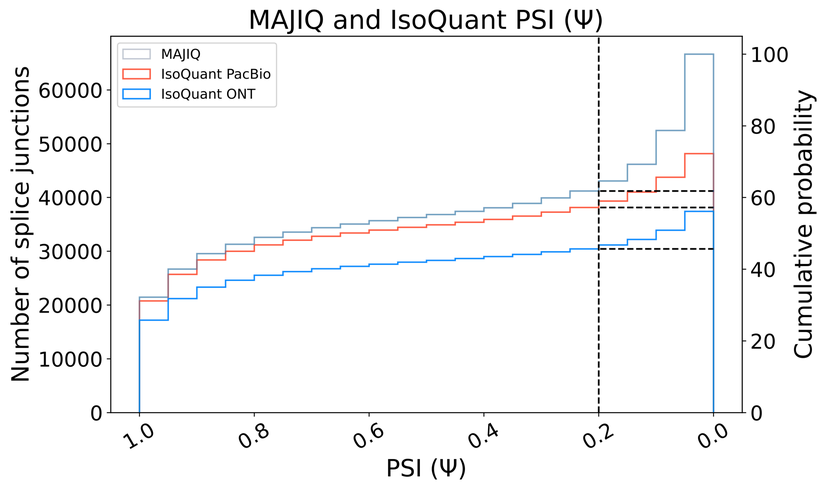

a
MAJIQ
Local Splice Variations (LSV)
Splice graph of possible isoforms
Short read RNA-seq
Genome Annotation
Long read Algorithm
List of full isoforms
Long read RNA-seq
Genome Annotation
b
Comparative analysis
c
Unified visualization
LSV type
Ψ per junction

a

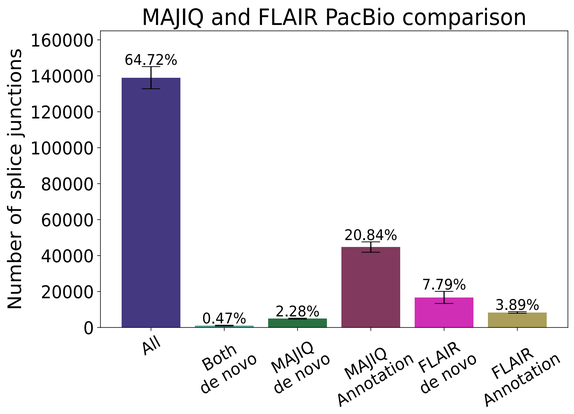
	MAJIQ (short read)	LR algorithm (long read)	Annotation
All	✓	✓	✓
Both de novo	✓	✓	✗
MAJIQ de novo	✓	✗	✗
MAJIQ / Annotation	✓	✗	✓
LR de novo	✗	✓	✗
LR / Annotation	✗	✓	✓

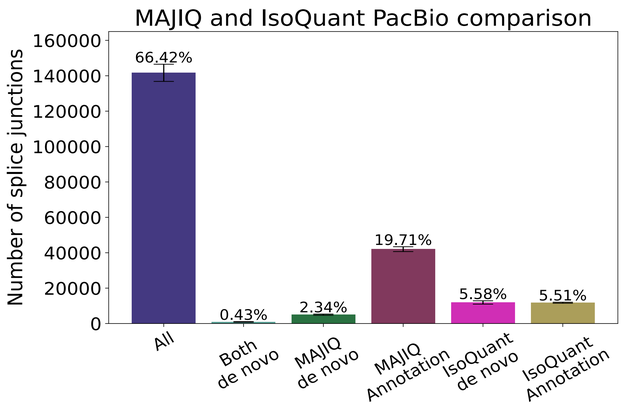
c

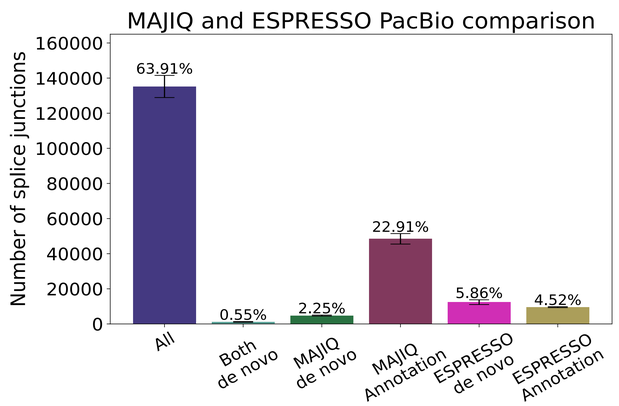


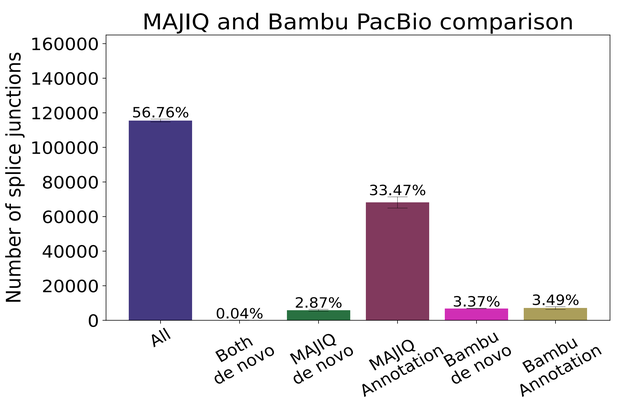


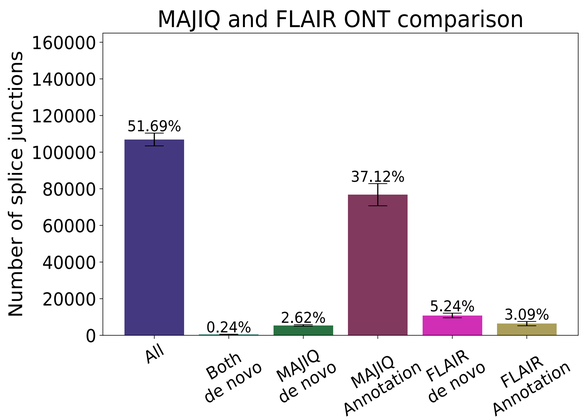
b

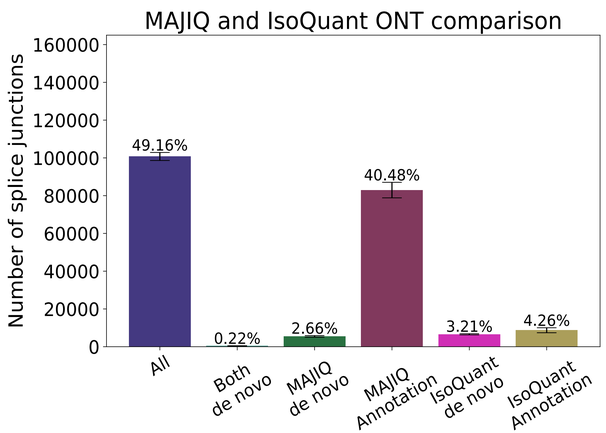


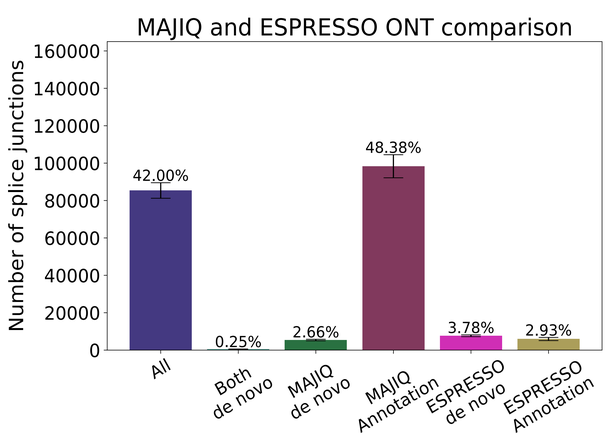


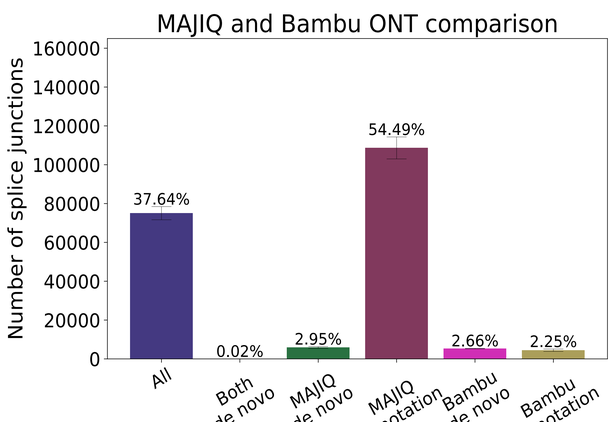


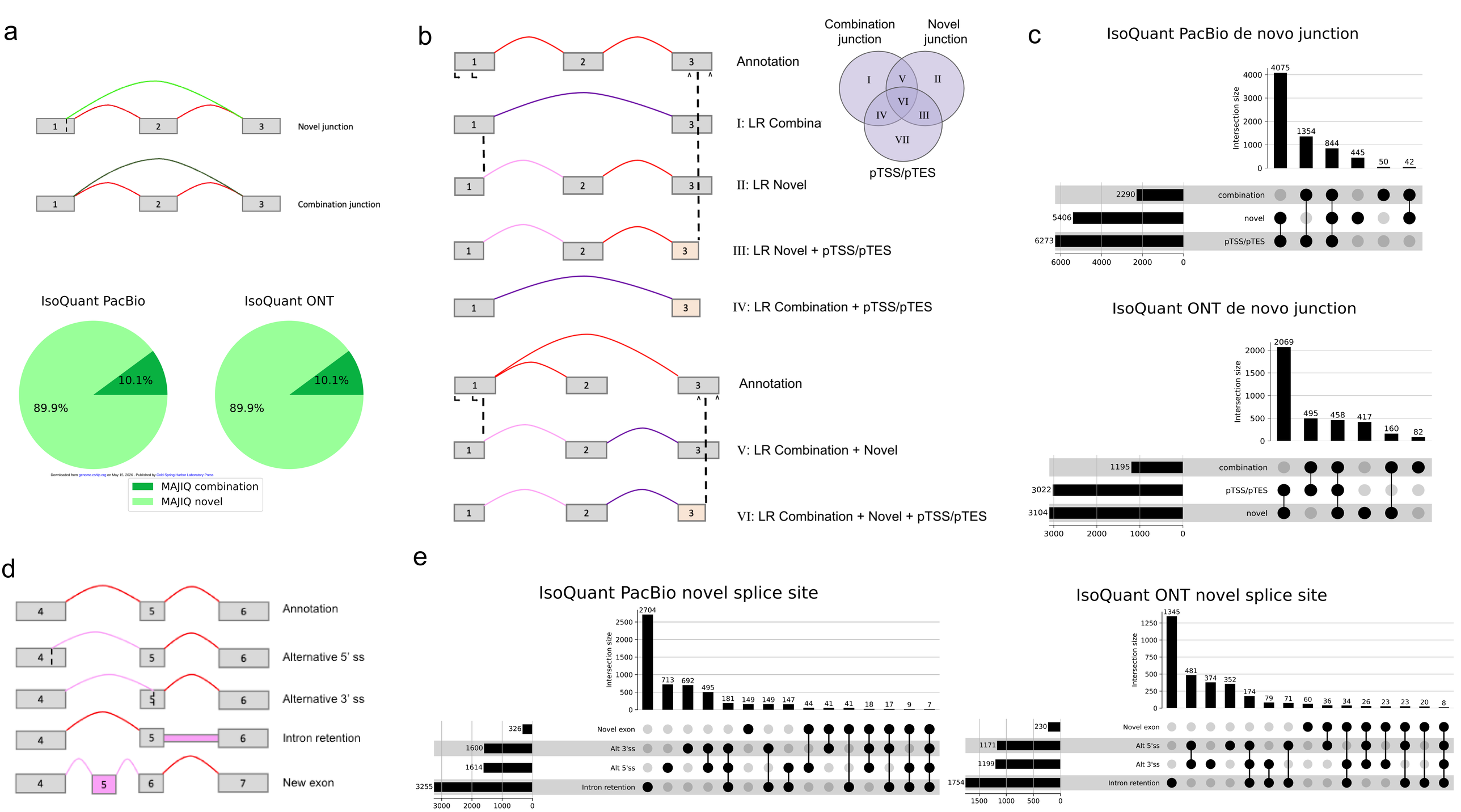




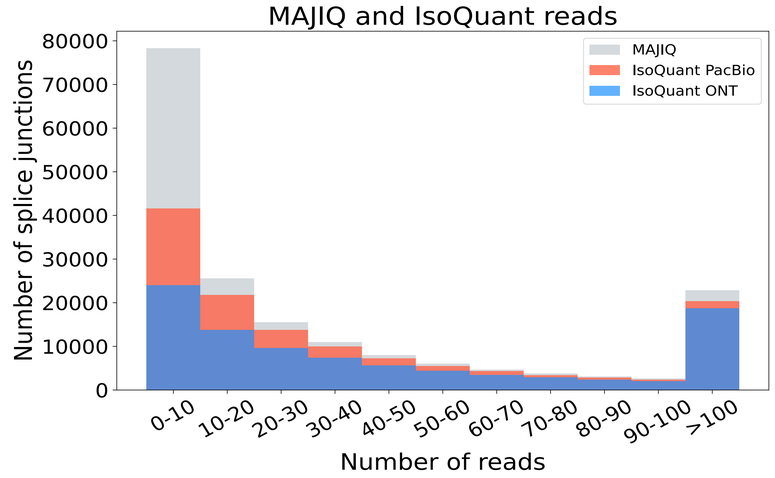


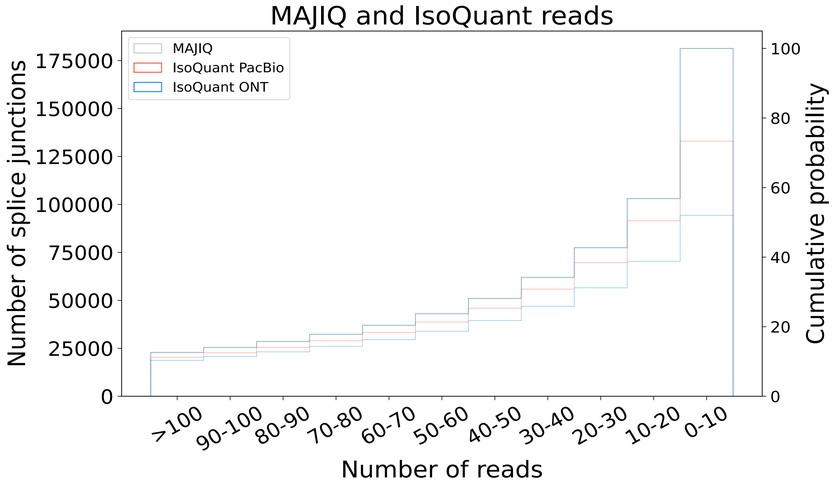




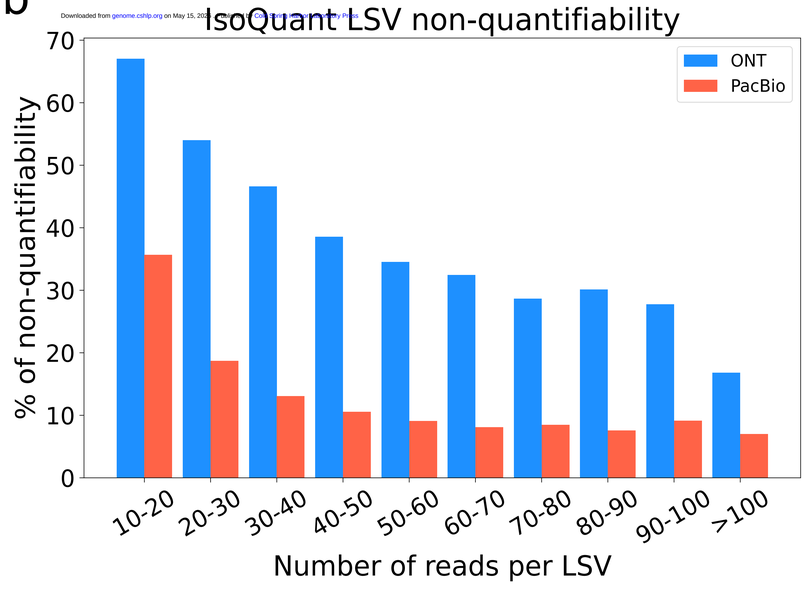

a
Novel junction
Combination junction
IsoQuant PacBio
IsoQuant ONT
10.1%
89.9%
MAJIQ combination
MAJIQ novel
b
Annotation
I: LR Combina
II: LR Novel
III: LR Novel + pTSS/pTES
IV: LR Combination + pTSS/pTES
V: LR Combination + Novel
VI: LR Combination + Novel + pTSS/pTES
Combination junction
Novel junction
pTSS/pTES
c
IsoQuant PacBio de novo junction
Intersection size
combination
novel
pTSS/pTES
IsoQuant ONT de novo junction
d
Annotation
Alternative 5' ss
Alternative 3' ss
Intron retention
New exon
e
IsoQuant PacBio novel splice site
Novel exon
Alt 3'ss
Alt 5'ss
Intron retention
IsoQuant ONT novel splice site

a

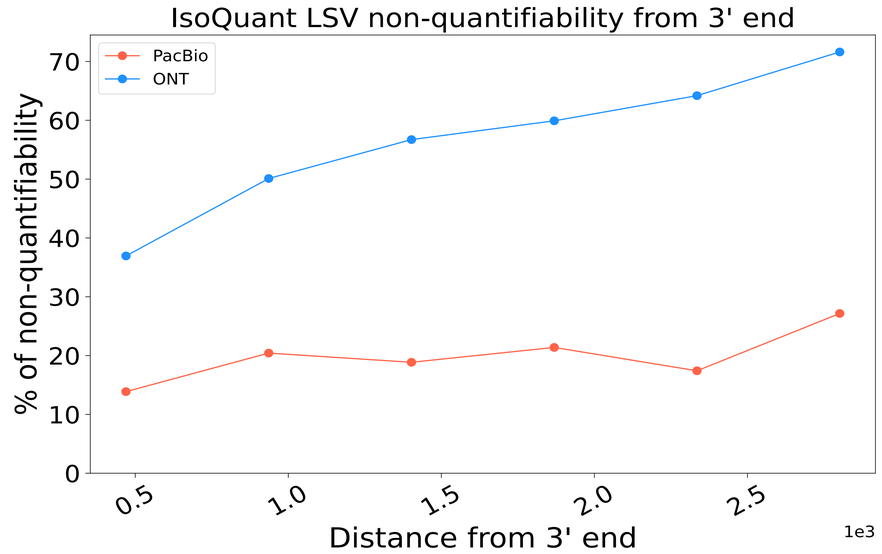


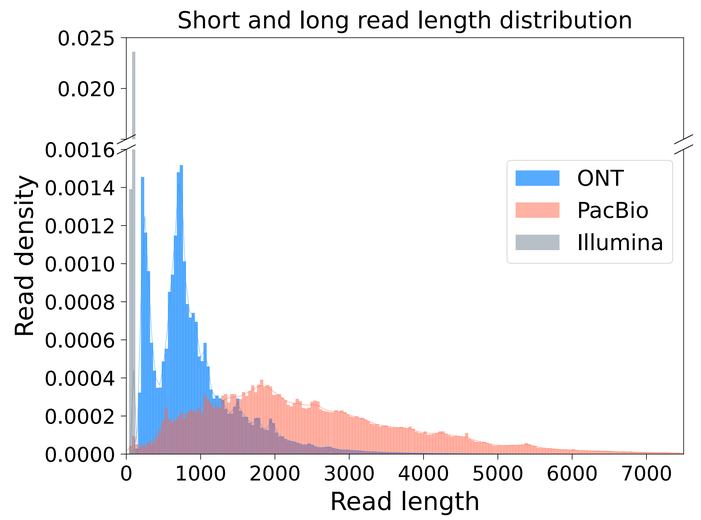


b

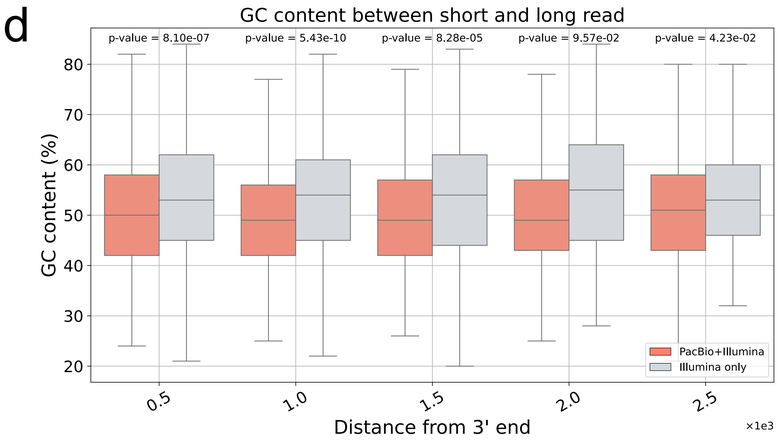


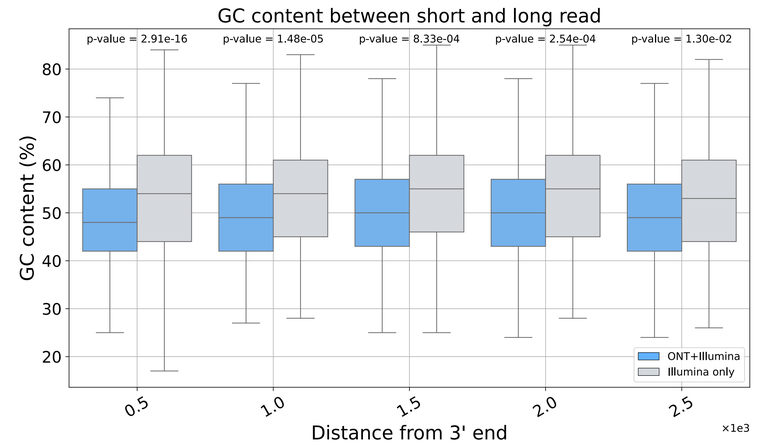
c

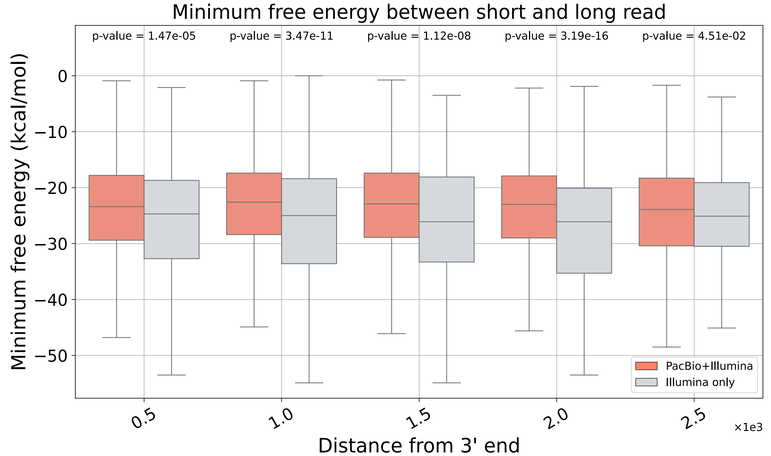


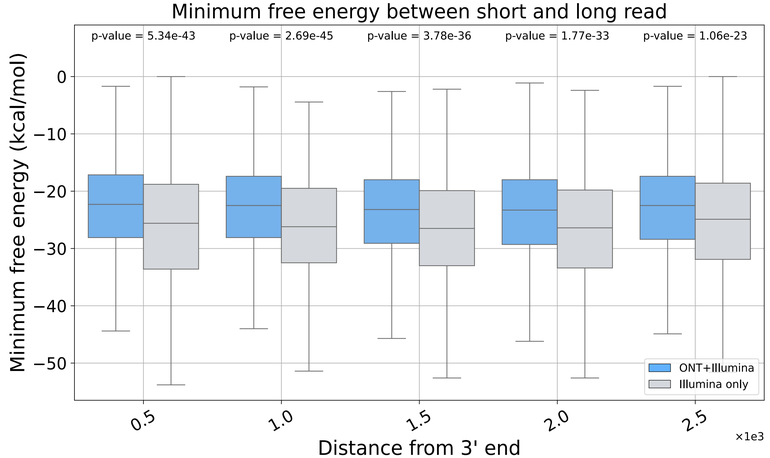


d

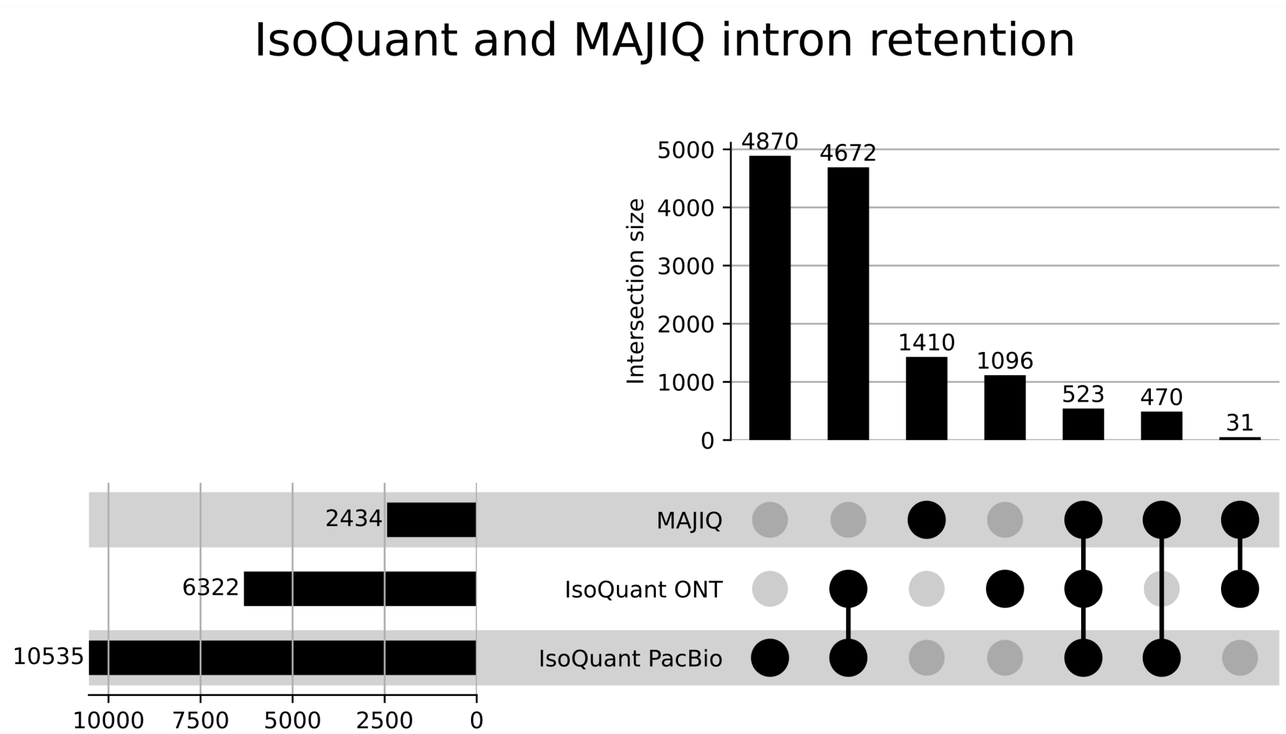




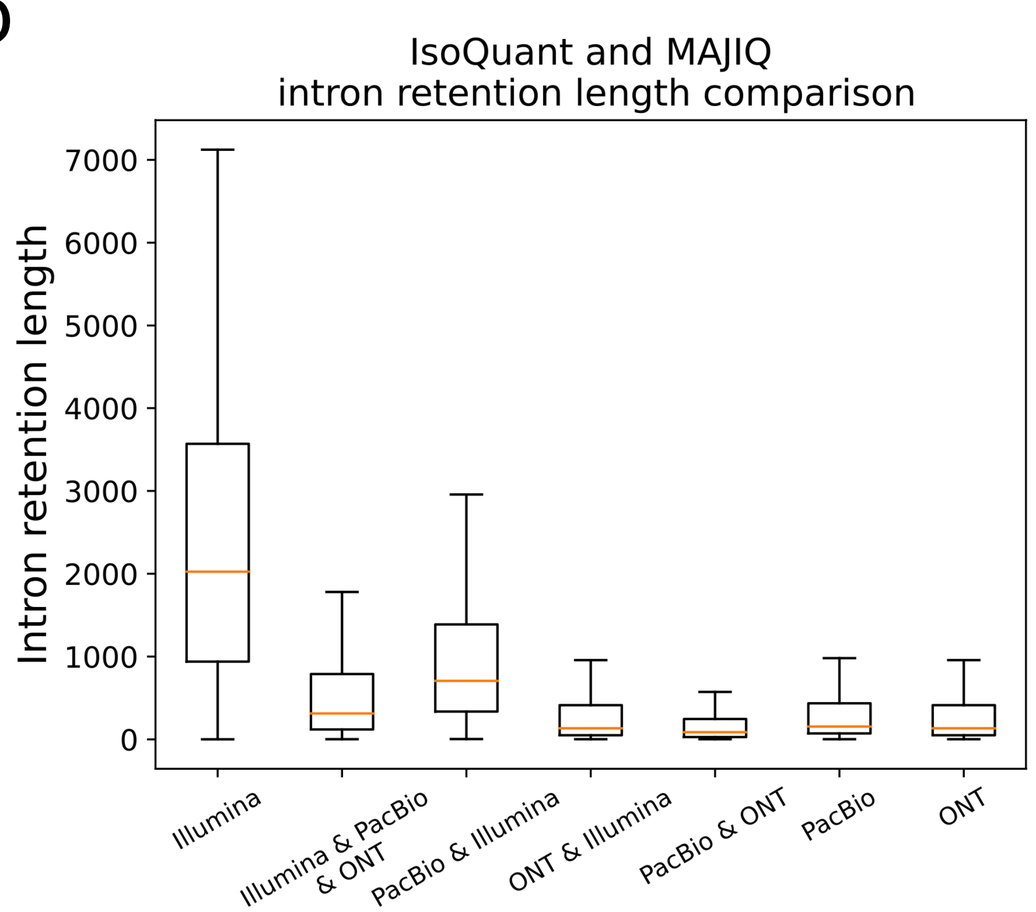




a



b



c