Downloaded from genome.cshlp.org on June 8, 2026 . Published by Cold Spring Harbor Laboratory Press

GENOME RESEARCH

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

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

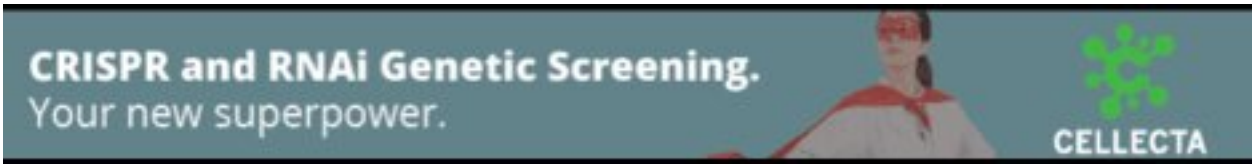
Genome Res. 2024 34: 1624-1635 originally published online September 25, 2024
Access the most recent version at doi:10.1101/gr.278659.123

References This article cites 41 articles, 10 of which can be accessed free at:
http://genome.cshlp.org/content/34/10/1624.full.html#ref-list-1

Open Access Freely available online through the *Genome Research* Open Access option.

Creative Commons License This article, published in *Genome Research*, is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at http://creativecommons.org/licenses/by-nc/4.0/.

Email Alerting Service Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or **click here.**



To subscribe to *Genome Research* go to:
https://genome.cshlp.org/subscriptions

© 2024 Han et al.; Published by Cold Spring Harbor Laboratory Press

Downloaded from genome.cshlp.org on June 8, 2026 . Published by Cold Spring Harbor Laboratory Press

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

Seong Woo Han,[1,5] San Jewell,[2,5] Andrei Thomas-Tikhonenko,[3,4] and Yoseph Barash[1,2]

[1]Department of Computer and Information Sciences, School of Engineering, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA; [2]Department of Genetics, [3]Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA; [4]Division of Cancer Pathobiology, Children's Hospital of Philadelphia, Philadelphia, Pennsylvania 19104, USA

Mapping transcriptomic variations using either short- or long-read RNA sequencing is a staple of genomic research. Long reads are able to capture entire isoforms and overcome repetitive regions, whereas short reads still provide improved coverage and error rates. Yet, open questions remain, such as how to quantitatively compare the technologies, can we combine them, and what is the benefit of such a combined view? We tackle these questions by first creating a pipeline to assess matched long- and short-read data using a variety of transcriptome statistics. We find that across data sets, algorithms, and technologies, matched short-read data detects ~30% more splice junctions, such that ~10%–30% of the splice junctions included at ≥20% 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 can be combined with short-read data for improved transcriptome analysis.

[Supplemental material is available for this article.]

Long-read sequencing technology has been revolutionizing genomic studies in recent years, leading to it being elected recently as the "method of the year 2022" (Foord et al. 2023; Lucas and Novoa 2023; Marx 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 (Gao et al. 2023), IsoQuant (Prjibelski et al. 2023), Bambu (Chen et al. 2023), TALON (Wyman et al. 2019), SQANTI (Tardaguila et al. 2018), StringTie (Kovaka et al. 2019; Shumate et al. 2022), 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-read 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 (LSVs). LSV, first introduced in MAJIQ (Vaquero-Garcia et al. 2016), denotes splits in a gene splicegraph 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. LSVs are typically quantified using junction-spanning reads in terms of percentage 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 they 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 Figure 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 splicegraphs 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 data sets involving both PacBio and ONT using four different

[5]These authors contributed equally to this work.
Corresponding author: yosephb@upenn.edu
Article published online before print. Article, supplemental material, and publication date are at https://www.genome.org/cgi/doi/10.1101/gr.278659.123. Freely available online through the *Genome Research* Open Access option.

© 2024 Han et al. This article, published in *Genome Research*, is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at http://creativecommons.org/licenses/by-nc/4.0/.

34:1624–1635 Published by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/24; www.genome.org

Downloaded from genome.cshlp.org on June 8, 2026 . Published by Cold Spring Harbor Laboratory Press

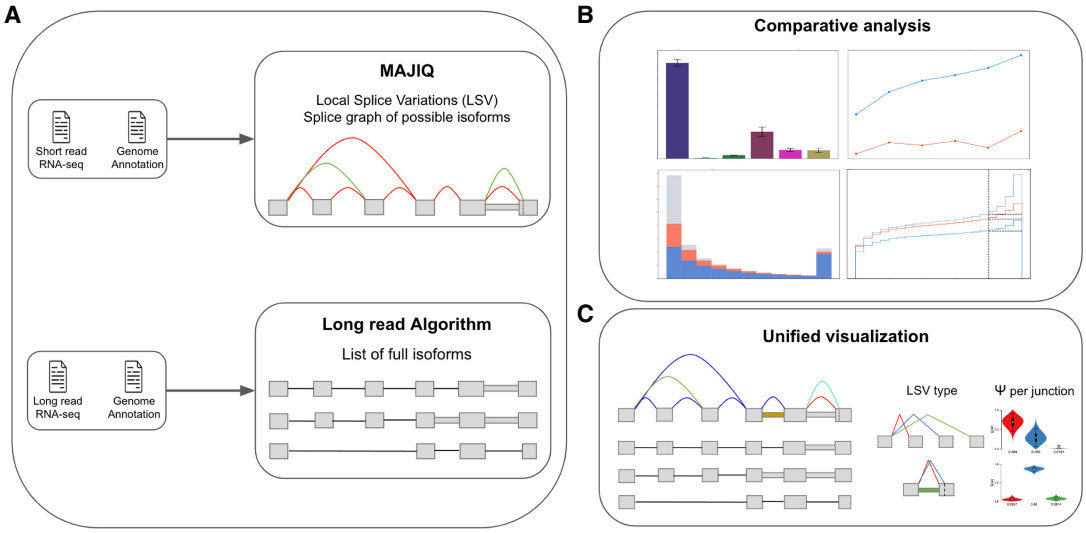


Figure 1. MAJIQ-L overview. (*A*) Input. Matched short- and long-read RNA sequencing data with genome annotation. Short reads are mapped with an aligner and then passed to MAJIQ, producing a splicegraph with a matching set of LSVs per gene. Long reads are processed with a long-read algorithm (e.g., IsoQuant) to produce isoform counts. (*B*) The outputs from the two RNA sequencing sources along with the annotation are compared. Each splice junction or intron retention (IR) is assigned to each of the possible six categories depending on which subset of the three sources supports 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.

long-read 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 the splice junctions contributes to their differential detection by short and long reads. Then, we turn to IR, 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-read 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 (for details, see Methods) (Tang et al. 2020; Chen et al. 2023; Gao et al. 2023; Prjibelski et al. 2023). 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 data sets: three replicates of human cell lines from the Long-read RNA-seq Genome Annotation Assessment Project (LRGASP) Consortium (Fig. 2B; Pardo-Palacios et al. 2024); 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 relapsed B cell acute lymphoblastic leukemia (B-ALL) (Supplemental Fig. S1; Bagashev et al. 2021; Schulz et al. 2021). To address coverage differences, we subsampled the files such that the total number of bases sequenced across the various platforms was similar (see Methods). Overall, our results indicate that across all three data sets, short reads detected 30% more splice junctions, with PacBio detecting ~10% more junctions than ONT. Some differences between long-read algorithms were also apparent. FLAIR detected the most numbers of long-read-only de novo splice junctions (~8%), whereas Bambu reported the least (~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 data set without subsampling, in which long reads have 1.3- to 2.9-fold more bases sequenced than short reads (LRGASP, B-ALL) and in which Illumina has 1.7-fold more coverage than ONT (GTEx). Similar trends were observed in this analysis as well (Supplemental Fig. S2; for details, see Supplemental Tables S1–S3).

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-read-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 samples, 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 data set revealed that short-read-only de

Downloaded from genome.cshlp.org on June 8, 2026 . Published by Cold Spring Harbor Laboratory Press

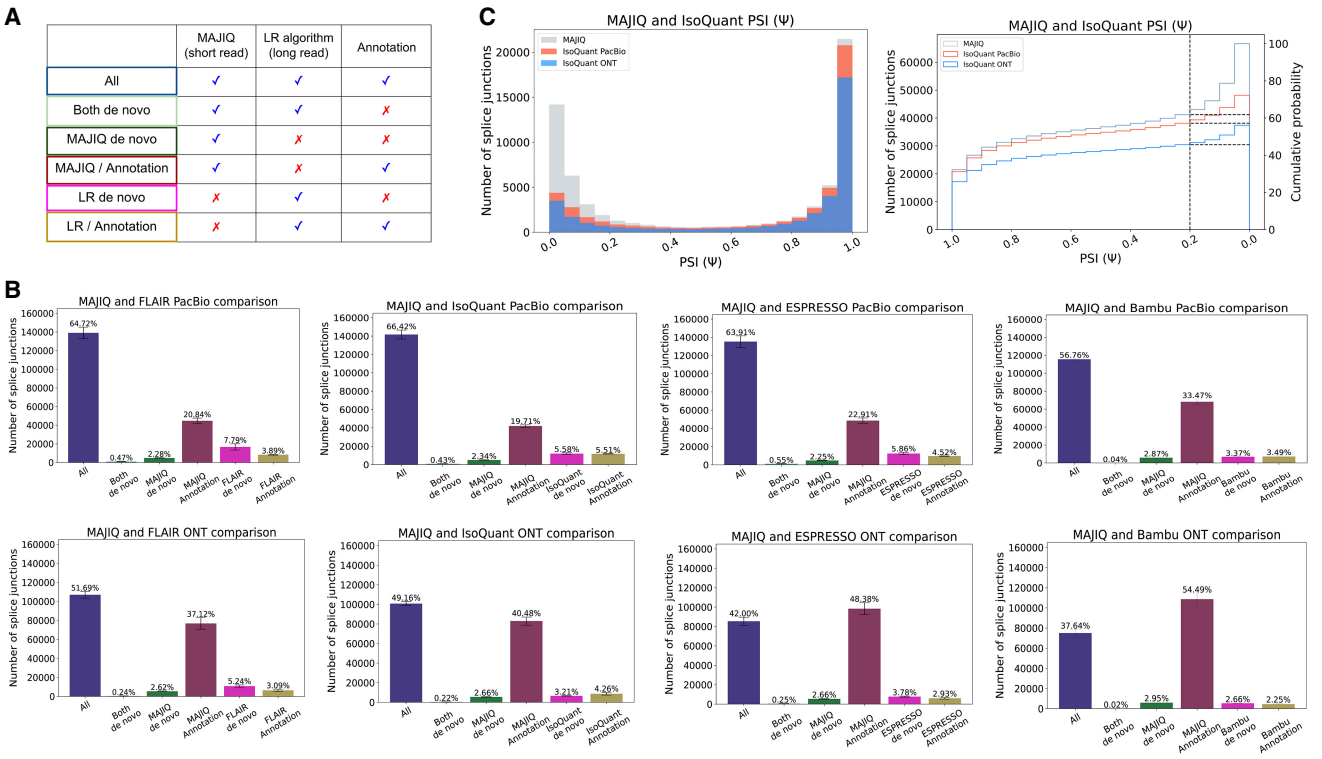


Figure 2. Splice junction comparative analysis. (*A*) Any detected splice junction can fall into one of six categories, each represented by a color, depending which of the three sources of information (short reads, long reads, annotation) support it. (*B*) Bar charts corresponding to the aforementioned six categories. Mean and standard error bars are computed using matched data sets from three replicates of human cell-line samples sequenced by LRGASP (Pardo-Palacios et al. 2024). These data include short reads processed by STAR and MAJIQ, long reads from PacBio and ONT assays, and four long-read algorithms used to process the long-read data. (*C*) Taking the splice junctions reported in *B* by MAJIQ (green) and assessing the number of those also identified when using PacBio (red) or ONT (blue) long reads, as a function of the PSI values. Here IsoQuant was used for long-read data. Note that if a junction appears in multiple LSVs, 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%.

novo junctions maintained high alignment at 86% ± 0.5%. Long-read-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, although this kind of overlap analysis with an existing splice junction database cannot be used to assess reliable junction detection directly, these results offer additional support for the assertion that the vast majority of short-read-only de novo junctions are real whereas their long-read counterpart may suffer from significantly higher false positives. This conclusion regarding the low false discovery of junctions following MAJIQ's processing of short-read data is further supported by additional analysis described in the Supplemental Information.

Nevertheless, the significant number of short-read-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 owing to its ease of use, as it can be efficiently executed with a single command line, which simplifies the workflow compared with other tools that require a couple of steps. We ensured transparency in our methodology by demonstrating in both the main and Supplemental Texts that consistent trends are observable across all analyses, regardless of the algorithms. The results shown in Figure 2C for IsoQuant indicate most short-read-only de novo splice junctions are, as expected, relatively lowly included ($\Psi < 10\%$). Nonetheless, the cumulative distribution function plot shows IsoQuant + PacBio misses 8% of junctions with significant inclusion levels ($\Psi > 20\%$), and IsoQuant + ONT misses 35% of junctions detected by MAJIQ (Fig. 2C, right). When comparing the different algorithms, IsoQuant misses the least number of junctions with $\Psi > 20\%$, and Bambu misses the most with the same inclusion level (Supplemental Fig. S8B).

Patterns of novel splice variant differences 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 LSVs 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

Downloaded from genome.cshlp.org on June 8, 2026 . Published by Cold Spring Harbor Laboratory Press

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 ~90% involve novel splice sites whereas the rest is novel combinations. A similar trend is observed when compared with different algorithms (Supplemental Fig. S3).

In long-read de novo junctions, we observe another type of transcript change, termed putative transcript start or end site (pTSS/pTES). pTSSs/pTESs 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 pTSSs/pTESs 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 bona fide TSS/TES missing in the annotation. All possible combinations of novel splice junctions with pTSSs/pTESs are shown in Figure 3B, along with their matching Roman numerals in the associated Venn diagram. Notably, short-read-based splicing algorithms such as MAJIQ are generally unable to call TSSs/TESs, so comparison of such cases is not feasible. For this reason, we do not analyze here long-read-based transcripts that only involve pTSSs/pTESs (category VII in Fig. 3B). Nonetheless, we see that when analyzing long-read transcripts with novel splice junctions, most of those involve novel splice sites, and almost all of them also include pTSSs/pTESs (Fig. 3C).

We examined long-read novel splice variants and further categorized those into splice sites that involve alternative 5′/3′ splice sites, IRs, or new exons (Fig. 3D). We find most of the novel splice variants reported by long reads involve IRs, with PacBio and ONT reads yielding 3255 and 1754 such cases, respectively (Fig. 3E). Novel alternative 5′/3′ splice sites or mixes of those were significantly less common, and novel exons were quite rare, only ~10% of the number of IR cases. These results, shown in Figure 3, C and E, are the average values of the three replicates of PacBio and ONT shown in Figure 2B. Results from different algorithms with different data sets have similar trends except for Bambu (Supplemental Figs. 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

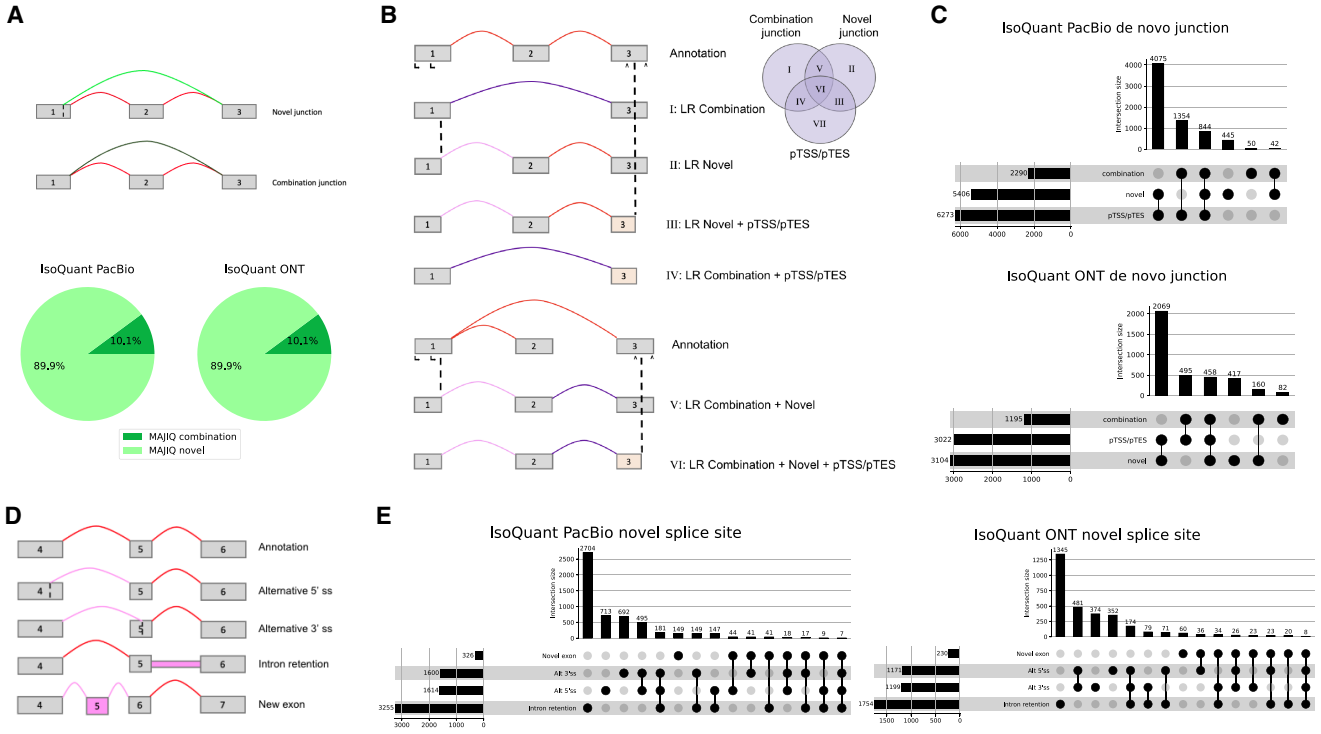


Figure 3. Analysis of de novo elements. (*A*) Short-read-only de novo splice junctions reported by MAJIQ (green junctions between exon 1 and 3 in the splicegraphs) can be classified as those involving novel splice sites (light green) or a novel combination of known splice sites (dark green). Red junctions correspond to annotated ones. The pie chart shows that compared with long reads processed with IsoQuant, ~90% of MAJIQ de novo splice junctions involve novel splice sites. (*B*) Representative cartoon examples for six different categories of long-read de novo transcript variations. Novel combination junctions (dark purple), junctions involving novel splice sites (light purple), and junctions supported by annotation (red) are the same as in *A*. Putative starts or ends (pTSSs/pTESs; light yellow), or partial exons, represent cases when the transcript start sites 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* pTSSs/pTESs (class VII in the Venn diagram) are not included in downstream analysis as those are not handled by MAJIQ or similar short-read-based splicing algorithms, so they cannot 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 pTSSs/pTESs. (*D*) Representative cartoon examples for the types of novel splice variations (pink) that a novel splice variant in long reads can introduce compared with the annotation (*top* graph, red junctions). (*E*) Breakdown of long-read 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.

Downloaded from genome.cshlp.org on June 8, 2026 . Published by Cold Spring Harbor Laboratory Press

score are annotated, providing an intuitive precision estimation. For completion, we also include pie chart and upset plot results for the original data when long reads have significantly more coverage than short-read data, exhibiting similar trends as well (Supplemental Figs. S4, S6).

Coverage, 3′ bias, and GC content lead to differences between short- and long-read transcriptome views

Splice-site disagreement between PacBio and ONT was previously shown to frequently represent small shifts in splice-site calls, possibly owing to technical artifacts (Mikheenko et al. 2022). Thus, we hypothesized that this phenomenon 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-read 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 three to eight, whereas the most significant decrease is in splice sites that are only found by long reads (magenta bars). These long-read-only splice junctions that were "fuzzy matched" may represent an error in short reads or situations in which 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 six 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, although FLAIR de novo cases decreased more compared with 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 data sets and tend to fall below cutoffs in long-read analysis. Although matched short and long nuclear poly(A)+ and cytosolic poly(A)+ data sets 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 with 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 IR events compared with 446,905 splice junctions and 12,325 IR events in the cytosol. Thus, cytosolic poly(A)+ samples exhibit a marginal 3.82% decrease in the detection of junctions compared with nuclear poly(A)+ samples. However, nuclear poly(A)+ samples demonstrate a 381.53% increase in IR detection compared with 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 are termed "retained" or "detained" introns (Boutz et al. 2015; Barutcu et al. 2022). This "sequencing budget spending" on introns, as well as the fact some splicing events have not finished processing, leads to a slight decrease in the number of junctions detected in the cytosol. Thus, although we do not have matched long reads, these results do not point to lowly abundant nuclear junctions as the main driver for the observed discrepancy between the 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 (more than 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-read technology. This raises the question of how many of such AS events, or LSVs 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 the LSV minimal read number to be quantifiable by MAJIQ. Figure 4B shows that using the PacBio data, IsoQuant is unable to quantify 36% of the LSVs quantified by MAJIQ when having 10 to 20 short reads, whereas using the ONT reads, IsoQuant cannot quantify >65% of the LSVs in the same bin. Although the nonquantifiability decreases as the number of short reads per LSV increase, even for LSVs with more than 100 short reads, PacBio is unable to quantify 8% of the LSVs, a fraction similar to the one observed for splice junction detection above. The fraction of nonquantifiable LSVs by ONT is higher, at 18% of LSVs 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-read-based splicing variations is the inherent 3′-to-5′ bias of long-read 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 data set shown in Figure 4C (right panel) actually span ≥3000 bp, which is roughly the median length of the human transcripts (Lopes et al. 2021). Furthermore, as noted above, long reads report more novel IR. This means that any splice junctions downstream from those IR events are captured further away from the poly(A) tail and, hence, are less likely to be detected. To assess

Downloaded from genome.cshlp.org on June 8, 2026 . Published by Cold Spring Harbor Laboratory Press

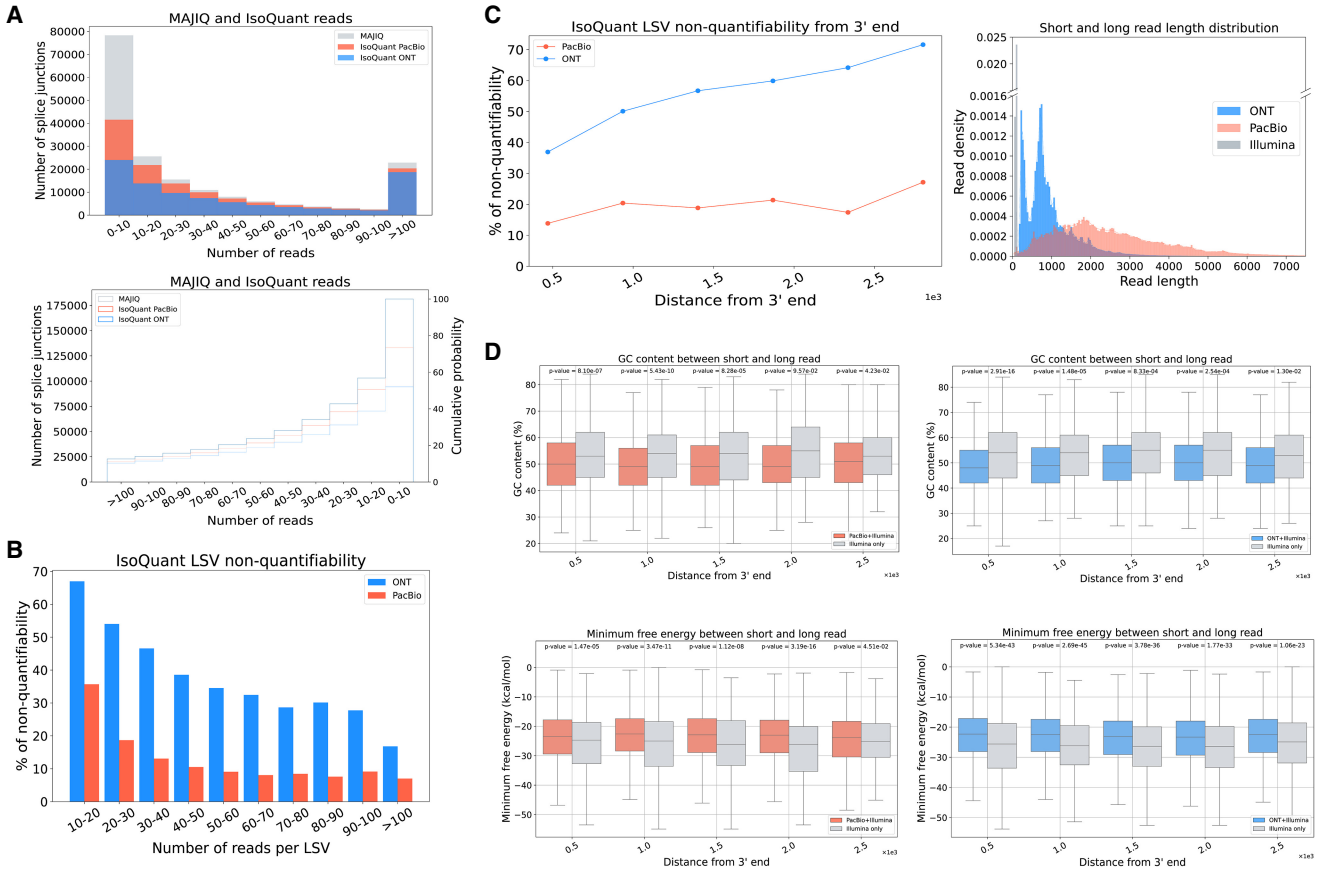


Figure 4. Analysis of sources of discrepancy between short- and long-read-based transcriptome variations. (*A*) The number of MAJIQ's splice junctions (gray) identified by IsoQuant using PacBio (red) or ONT (blue) as a function of the number of short reads covering the junctions. The histogram on the *top* and CDF on the *bottom* 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-read analysis, which were "nonquantifiable" by IsoQuant using PacBio-matched (orange) and ONT-matched (light blue) long-read 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 (more than 100 reads). (*C*) Same plot as in *B* for the fraction of nonquantifiable LSVs by long-read 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 data set was used in all the above subfigures. (*D*) Boxplots showing GC content and minimum free energy (MFE) across various distances from the transcript 3′-end for junctions with more than 40 Illumina reads in *A* that are only detected by short reads (gray) or are 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.

the effect of the 3′ bias in long compared to short reads, we repeated the analysis of Figure 4B but with LSVs binned by their distance from the long-read 3′ end. Figure 4C (left panel) shows that with ONT IsoQuant cannot quantify ~78% of LSVs when these are >2500 bp away from the 3′ end. However, ~38% of LSVs close to the 3′-end are also nonquantifiable. For PacBio data, which offered longer reads, the distribution of nonquantifiable LSVs as a function of 3′-end distance is much more flat: 29% of the LSV were nonquantifiable by PacBio reads when those were >2500 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 Figure 4A (top) and computed GC content within a 100 bp window, 50 bp flanking each side of the junction for two groups of junctions: those detected in long reads versus 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. Figure 4D (top) indicates that short-read-only junctions have 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 Figure 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 IR events but fewer long introns

Previous sections investigated the differences between short and long reads in terms of splice junctions. However, long reads have

Downloaded from genome.cshlp.org on June 8, 2026 . Published by Cold Spring Harbor Laboratory Press

a natural advantage in detecting IRs because a single molecule may be sufficient to call such events. In contrast, IRs 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).

Although 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-read IR events reported by different algorithms (for details, see Methods). 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.

Figure 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 about 10,000 unique IR events compared

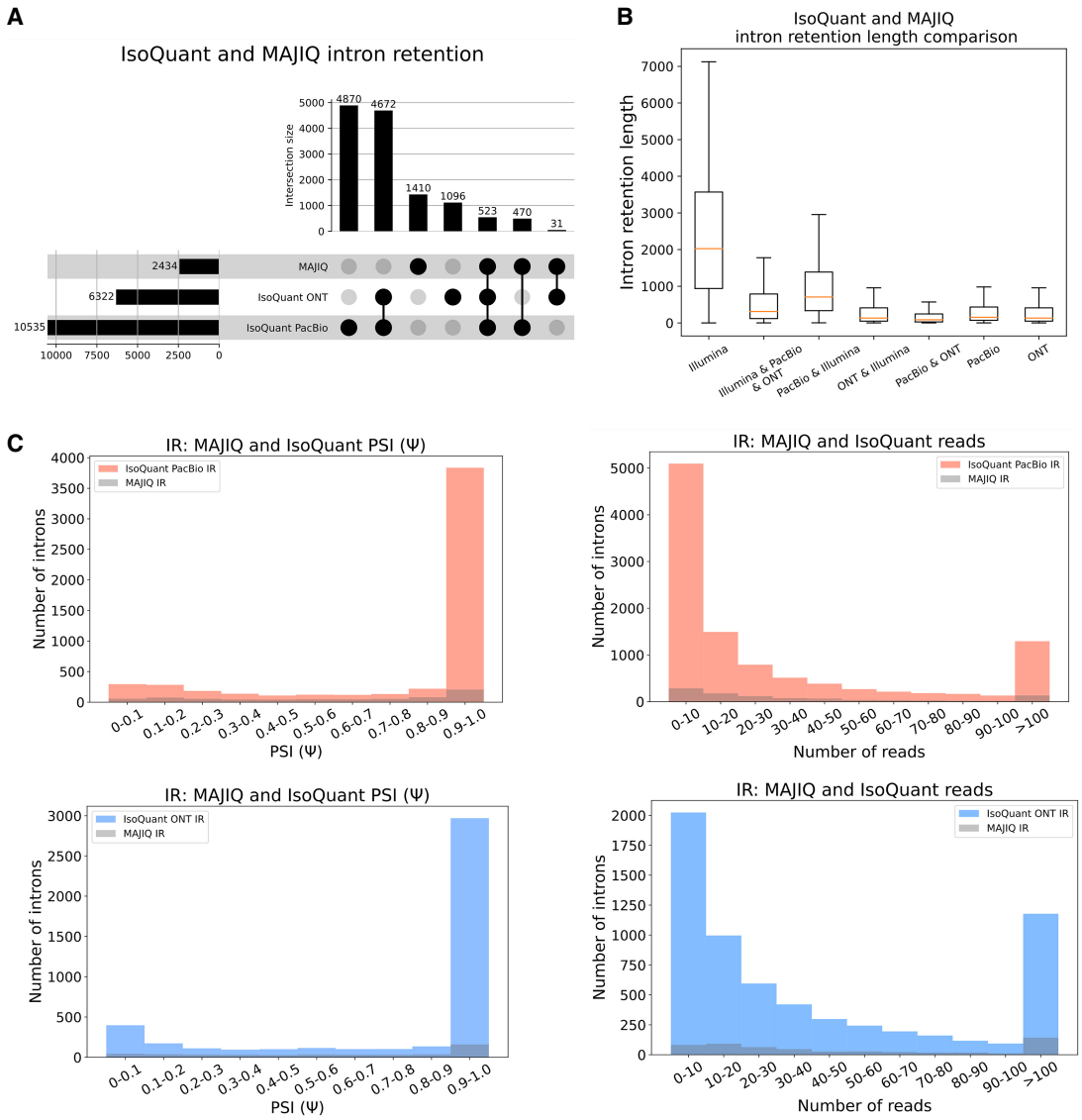


Figure 5. Comparison of IR events. (*A*) Upset plot showing overlap and total IR events reported by MAJIQ from short reads and IsoQuant using PacBio- or ONT-matched long reads (LRGASP data set). (*B*) Boxplots showing IR length distribution across seven categories in *A*. Each boxplot represents the 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 (red) or ONT (blue) and how many of those were also identified by MAJIQ (gray) 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.

Downloaded from genome.cshlp.org on June 8, 2026 . Published by Cold Spring Harbor Laboratory Press

with about 6300 with ONT and 2400 unique MAJIQ short-read-based IR events. Investigating these IR event sets, we find MAJIQ's unique set includes longer introns (Fig. 5B). This result is to be expected given the limitation of long-read 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 Figure 5, A–C, are the average values of three replicates in LRGASP data, and similar trends were observed when we used different long-read algorithms and GTEx data (Supplemental Figs. 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-read RNA sequencing. 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-read splicing analysis with GTF output files from any long-read algorithm. An illustrative example is shown in Figure 6, in which we ran MAJIQ with short reads and IsoQuant with PacBio long reads on

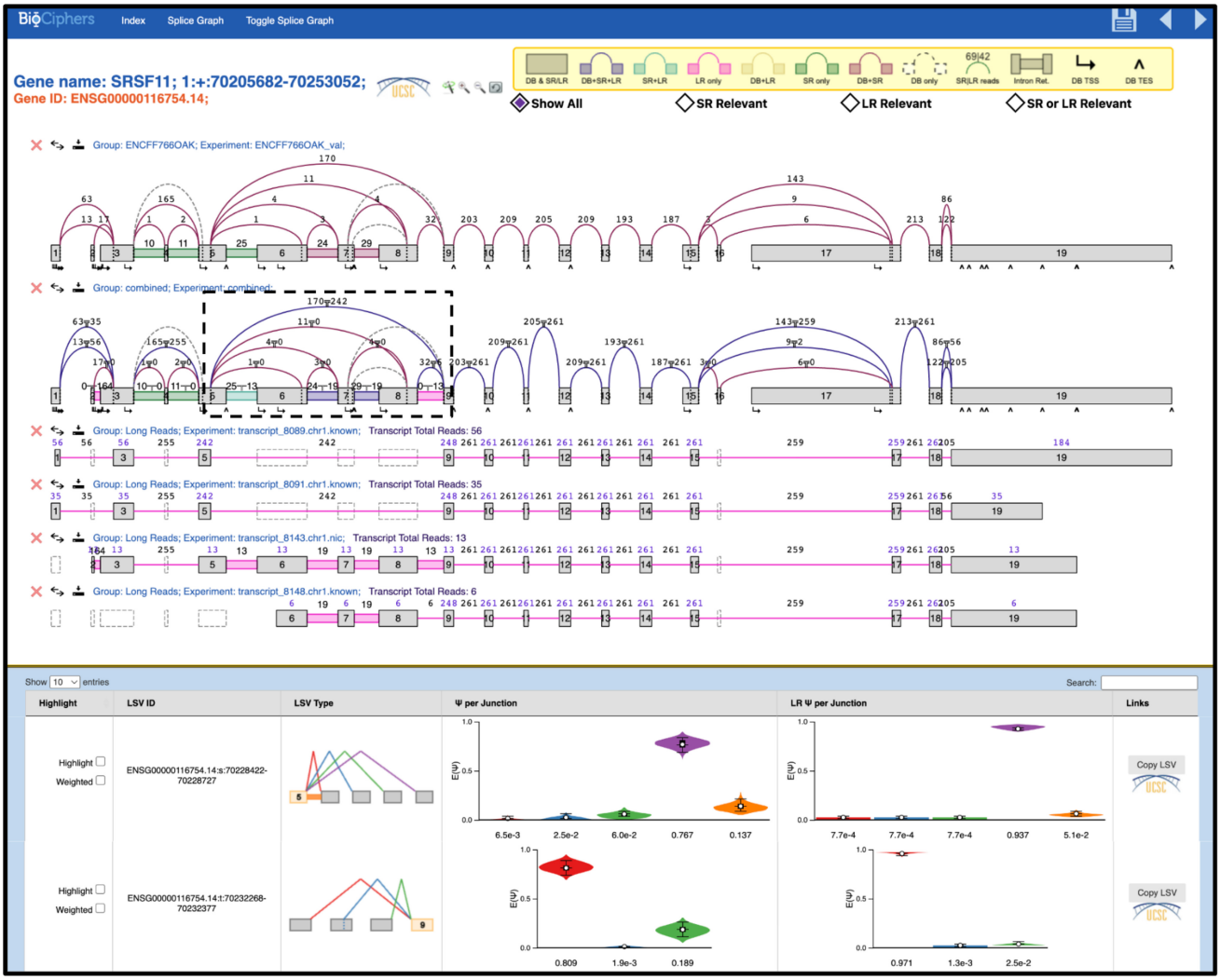


Figure 6. MAJIQ-L integrative analysis of splicing variations using short and long reads. Snapshot of the VOILA v3 interactive visualization of MAJIQ-L output for the *SRSF11* splice factor using the LRGASP data, including short reads processed by MAJIQ and PacBio reads processed by IsoQuant. The *top* portion shows gene information and filtering criteria between short and long reads as well as the short-read 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 element (junctions, introns, and exons) are included. In the unified splicegraph view, read counts for junctions and introns are shown for both short (*left*) and long (*right*) reads, separated by a *T* sign. Note that pTSSs/pTESs 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.

Downloaded from genome.cshlp.org on June 8, 2026 . Published by Cold Spring Harbor Laboratory Press

one of the human cell-line samples in LRGASP (Pardo-Palacios et al. 2024). VOILA v3 can show a short-read-based gene splicegraph (first row), a unified short- and long-read gene splicegraph (second row), and a list of transcripts found by only long reads (row three and below). Each color for splice junction and IR shows what source supports it (short, long, and annotation), with colors matching those in Figure 2B. Users can filter their data by several criteria, including which source the splicegraph 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 TSSs/TESs 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 ultraconserved 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 (Lareau et al. 2007; Ni 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 in line with our more general transcriptome-wide analysis as long reads are more likely to have a bias toward short isoforms (skipping the exons), yet can more easily detect IR 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 AS 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 laboratories, 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 splicegraph 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. 2022). Similarly, the ability to unequivocally and with high sensitivity detect IR and overcome mappability limitations over repetitive regions are significant advantages of long-read technologies. However, the qualitative limitations of low coverage and higher error rates, combined with the fact that extensive short-read data already exist, led us to compare and contrast the short and long reads based on transcriptome variation 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-read technology or algorithm can be compared with MAJIQ's short-read-based splicing analysis. We showed there is a significant gap in both detection and quantification of splicing variations. Short-read-based analysis of matched data sets revealed 30% more splice junctions, with PacBio detecting ~10% more than ONT. Eleven percent to 18% junctions with high short-read coverage (>100) were missed by PacBio and ONT, and ~10%–30% of splice junctions with PSI > 20% were missed by PacBio and ONT, respectively. As for the ability to quantify LSVs, we found a clear 3′-to-5′ bias; 12%–29% (PacBio) and 38%–78% (ONT) of the LSVs quantified by MAJIQ from short reads were unquantifiable by IsoQuant as a function of the distance from the poly(A) tail. This phenomenon is not unique to IsoQuant and seem to reflect the limited length of the long reads, especially those in the ONT data sets 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, 10,000 and 6300, compared with about 2400 for MAJIQ short-read IR. Finally, trying to relax the default parameters of the long-read 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-read analysis. Second, we are painfully aware that long-read 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 by 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-read data subsume short-read 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-read analysis.

Our analysis points to several key conclusions. First, to answer many scientific questions, researchers may be best served by combined short- and long-read analysis, possibly utilizing existing short-read data, and a two-stage approach: initial discovery with short reads and 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

Downloaded from genome.cshlp.org on June 8, 2026 . Published by Cold Spring Harbor Laboratory Press

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 versus a template switching one, which is the standard long-read cDNA protocol. Although 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.

Although 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, although error rates provide valuable insight into sequencing quality, they may not directly correlate with the efficacy of transcriptome analysis. Evaluating the effectiveness of the long-read RNA-seq algorithms' 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-read results discussed above, and assessing the ability to perform transcript 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). UMIs 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-read 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 subsampling from either short- or long-read data sets 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 postsubsampled number of bases for each data set can be found in Supplemental Tables S1–S3.

MAJIQ's short-read 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, IR, and exons. The resulting splicegraphs are used as MAJIQ-L's input along with the long-read tools' output in GTF file format for the downstream comparative analysis.

GTF 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 the "*–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 one 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-read model, we set a prior distribution on PSI that favors either high or low PSI values, which can be generalized by the Jeffrey's prior for an LSV with J junctions:

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

Because this prior is conjugate to the binomial distribution, our posterior distribution of Ψ_j given the observed number of reads is 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 sets

The matched human cell-line data set can be accessed through the LRGASP (https://www.gencodegenes.org/pages/LRGASP/). The GTEx v9 heart atrial appendage is available at the GTEx website (https://www.gtexportal.org/home/datasets).

Software availability

MAJIQ 2.5 provides the new VOILA visualization features of MAJIQ-L, which is available at https://majiq.biociphers.org/

Downloaded from genome.cshlp.org on June 8, 2026 . Published by Cold Spring Harbor Laboratory Press

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/ and as Supplemental Code.

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 laboratory including Y.B., S.W.H., and S.J. Otherwise, all authors declare they have no competing interests.

Acknowledgments

We thank Danielle Gutman, Matthew Gazzara, and Nathaniel Islas for helpful comments on the manuscript. This work was supported by National Institutes of Health U01 CA232563 (Y.B. and A.T.-T.), R01 LM013437 (Y.B.), and a CureBRCA grant (Y.B.).

Author contributions: Y.B. conceived the project. S.W.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. S.W.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. doi:10.1186/s13059-020-1935-5

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. doi:10.1182/blood-2021-148671

Barutcu AR, Wu M, Braunschweig U, Dyakov BJ, 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.e9. doi:10.1016/j.molcel.2021.12.010

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. doi:10.1038/nmeth.4106

Boutz PL, Bhutkar A, Sharp PA. 2015. Detained introns are a novel, widespread class of post-transcriptionally spliced introns. *Genes Dev* **29:** 63–80. doi:10.1101/gad.247361.114

Calvo-Roitberg E, Daniels RF, Pai AA. 2023. Challenges in identifying mRNA transcript starts and ends from long-read sequencing data. bioRxiv doi:10.1101/2023.07.26.550536

Chen Y, Sim A, Wan YK, Yeo K, Lee JJX, Ling MH, Love MI, Göke J. 2023. Context-aware transcript quantification from long-read RNA-seq data with Bambu. *Nat Methods* **20:** 1187–1195. doi:10.1038/s41592-023-01908-w

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. doi:10.1038/nature11233

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. doi:10.1093/bioinformatics/bts635

Engström PG, Steijger T, Sipos B, Grant GR, Kahles A, Rätsch G, Goldman N, Hubbard TJ, Harrow J, et al. 2013. Systematic evaluation of spliced alignment programs for RNA-seq data. *Nat Methods* **10:** 1185–1191. doi:10.1038/nmeth.2722

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. doi:10.1038/s41592-022-01715-9

Galante PAF, Sakabe NJ, Kirschbaum-Slager N, De Souza SJ. 2004. Detection and evaluation of intron retention events in the human transcriptome. *RNA* **10:** 757–765. doi:10.1261/rna.5123504

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. doi:10.1126/sciadv.abq5072

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. doi:10.1038/s41586-022-05035-y

Gruber AR, Lorenz R, Bernhart SH, Neuböck R, Hofacker IL. 2008. The Vienna RNA Websuite. *Nucleic Acids Res* **36:** W70–W74. doi:10.1093/nar/gkn188

Hardwick SA, Joglekar A, Flicek P, Frankish A, Tilgner HU. 2019. Getting the entire message: progress in isoform sequencing. *Front Genet* **10:** 709. doi:10.3389/fgene.2019.00709

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. doi:10.1186/s13059-019-1910-1

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. doi:10.1038/s41592-022-01716-8

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. doi:10.1038/nature05676

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34:** 3094–3100. doi:10.1093/bioinformatics/bty191

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. doi:10.1038/s41588-017-0004-9

Lopes I, Altab G, Raina P, de Magalhaães JP. 2021. Gene size matters: an analysis of gene length in the human genome. *Front Genet* **12:** 559998. doi:10.3389/fgene.2021.559998

Lucas MC, Novoa EM. 2023. Long-read sequencing in the era of epigenomics and epitranscriptomics. *Nat Methods* **20:** 25–29. doi:10.1038/s41592-022-01724-8

Marx V. 2023. Method of the year: long-read sequencing. *Nat Methods* **20:** 6–11. doi:10.1038/s41592-022-01730-w

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. doi:10.1101/gr.276405.121

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:** 266. doi:10.1186/s13059-016-1118-6

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. doi:10.1101/gad.1525507

Pardo-Palacios F, Reese F, Carbonell-Sala S, Diekhans M, Liang C, Wang D, Williams B, Adams M, Behera A, Lagarde J, et al. 2024. Systematic assessment of long-read RNA-seq methods for transcript identification and quantification. *Nat Methods* **21:** 1349–1363. doi:10.1038/s41592-024-02298-3

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* **41:** 915–918. doi:10.1038/s41587-022-01565-y

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. doi:10.1186/s13059-021-02411-1

Sharon D, Tilgner H, Grubert F, Snyder M. 2013. A single-molecule long-read survey of the human transcriptome. *Nat Biotechnol* **31:** 1009–1014. doi:10.1038/nbt.2705

Shen S, Park JW, Zx L, 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. doi:10.1073/pnas.1419161111

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. doi:10.1371/journal.pcbi.1009730

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. doi:10.1038/nmeth.2714

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. doi:10.1038/s41467-020-15171-6

Downloaded from genome.cshlp.org on June 8, 2026 . Published by Cold Spring Harbor Laboratory Press

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. doi:10.1101/gr.222976.117

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. doi:10.7554/eLife.11752

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. doi:10.1038/s41467-023-36585-y

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 doi:10.1101/672931

Zheng S, Gillespie E, Naqvi AS, Hayer KE, Ang Z, Torres-Diz M, Quesnel-Vallieres M, Hottman DA, Bagashev A, Chukinas J, et al. 2022. Modulation of CD22 protein expression in childhood leukemia by pervasive splicing aberrations: implications for CD22-directed immunotherapies. *Blood Cancer Discov* **3:** 103–115. doi:10.1158/2643-3230.BCD-21-0087

Received October 24, 2023; accepted in revised form September 11, 2024.