Supplementary Inforamation: 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

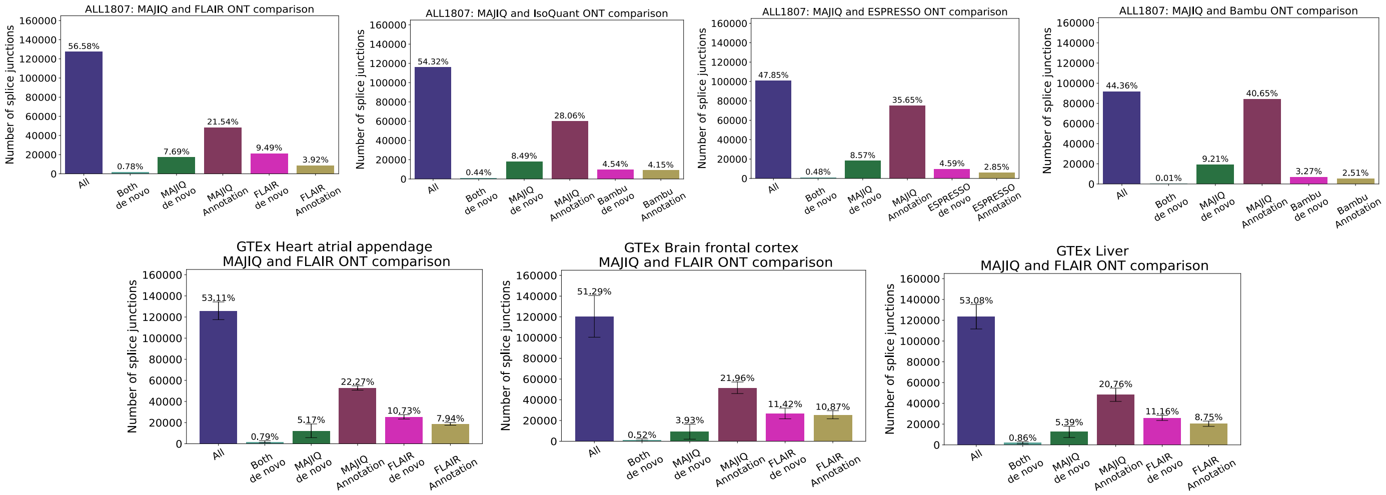
[*]Equal contribution
To whom correspondence should be addressed; E-mail: yosephb@upenn.edu.

Supplementary Analysis

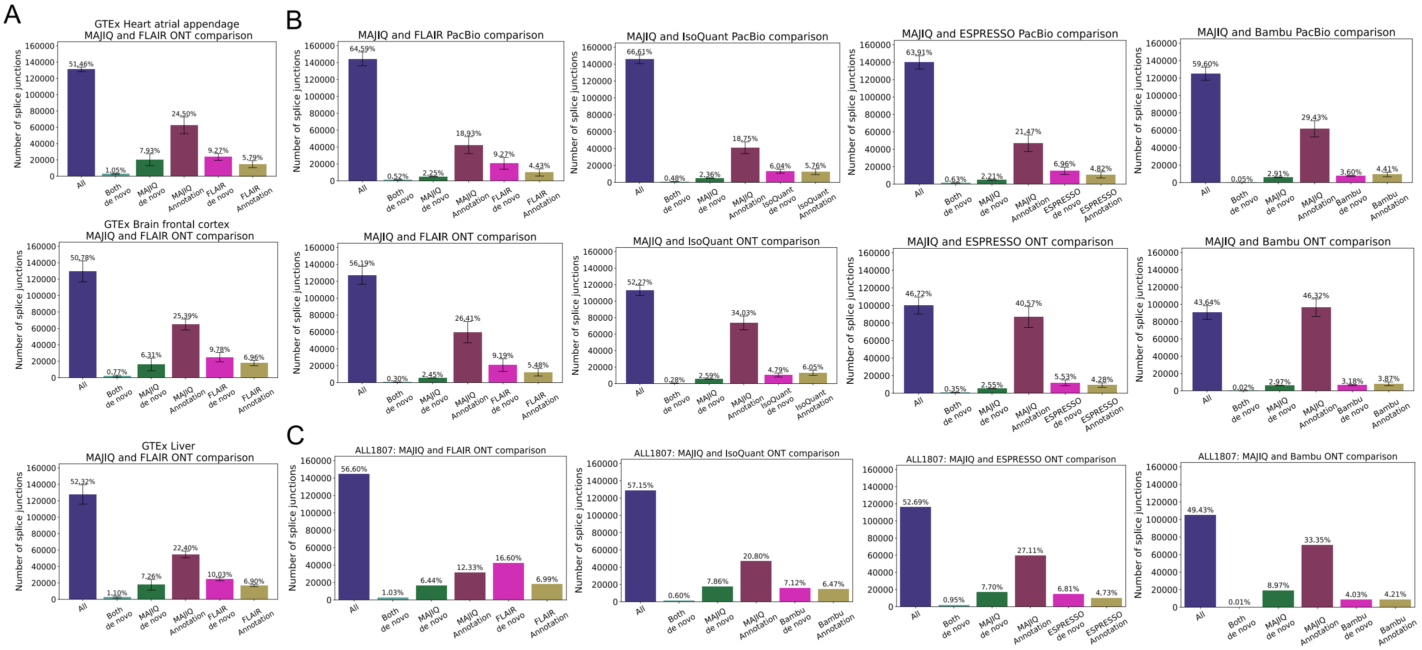
Assessing reliability of MAJIQ's short reads junction detection

To further strengthen the reliability of our junction analysis, we elaborate on MAJIQ's short-read junction detection process. We stress that the initial detection of junction spanning reads is done by the mapper (STAR in this case) and not MAJIQ. Nonetheless, to be considered as 'reliably detected' in a sample, MAJIQ sets a threshold not only on the number of reads (default = 3 for annotated junctions and = 5 for de novo) but also on the minimum number of read positions that must each contain at least one read (default = 2). This criterion ignores soft-clipped bases, which are not aligned segments of reads and is in addition to the typical overhang requirements (8 bases) on both ends of the junction. This default usage prevents detection based on a single PCR duplication. Elaborate work based on simulated data indicates a false discovery rate (FDR) of $\approx 2\%$ in a sample when two different positions are used[1]. Furthermore, only if enough samples in a sample group (e.g., tissue/cell type) have the junction reliably detected will the junction be considered and added to the splicing graph and subsequent analysis. We used the default setting of 51% of the group's samples, which translates to 2 of the 3 samples in the consortium's replicates. Assuming mapping errors in different samples are independent and using the above estimate for a single sample leads to an estimate of 0.04% falsely detected junctions. To assess the robustness of our detection criteria, we also tested the effect of taking the 50% most trustworthy junctions by requiring a minimum of 4 different read positions, rather than the default 2, for each junction. As expected, we observed a drop in the number of identified junctions. In the PDX cell line sample derived from a patient with relapsed B-Cell Acute Lymphoblastic Leukemia (B-ALL), referenced in Supplemental Fig. S1, the total number of junctions decreased from 428,265 (with 15,563 being de novo) to 426,784 (with 14,082 being de novo). Thus, the overall effect was only a 0.35% drop in detected junctions, though, for the $\approx 3.6\%$ subset of those that are de novo (15.5K/428K), there was a higher relative drop of $\approx 9.5\%$. This higher ratio is to be expected, given that de novo junctions are generally less frequent/abundant. While the observed drop should not necessarily be attributed to false positives, it indicates that short-read false positive junctions are not a significant component of our results.

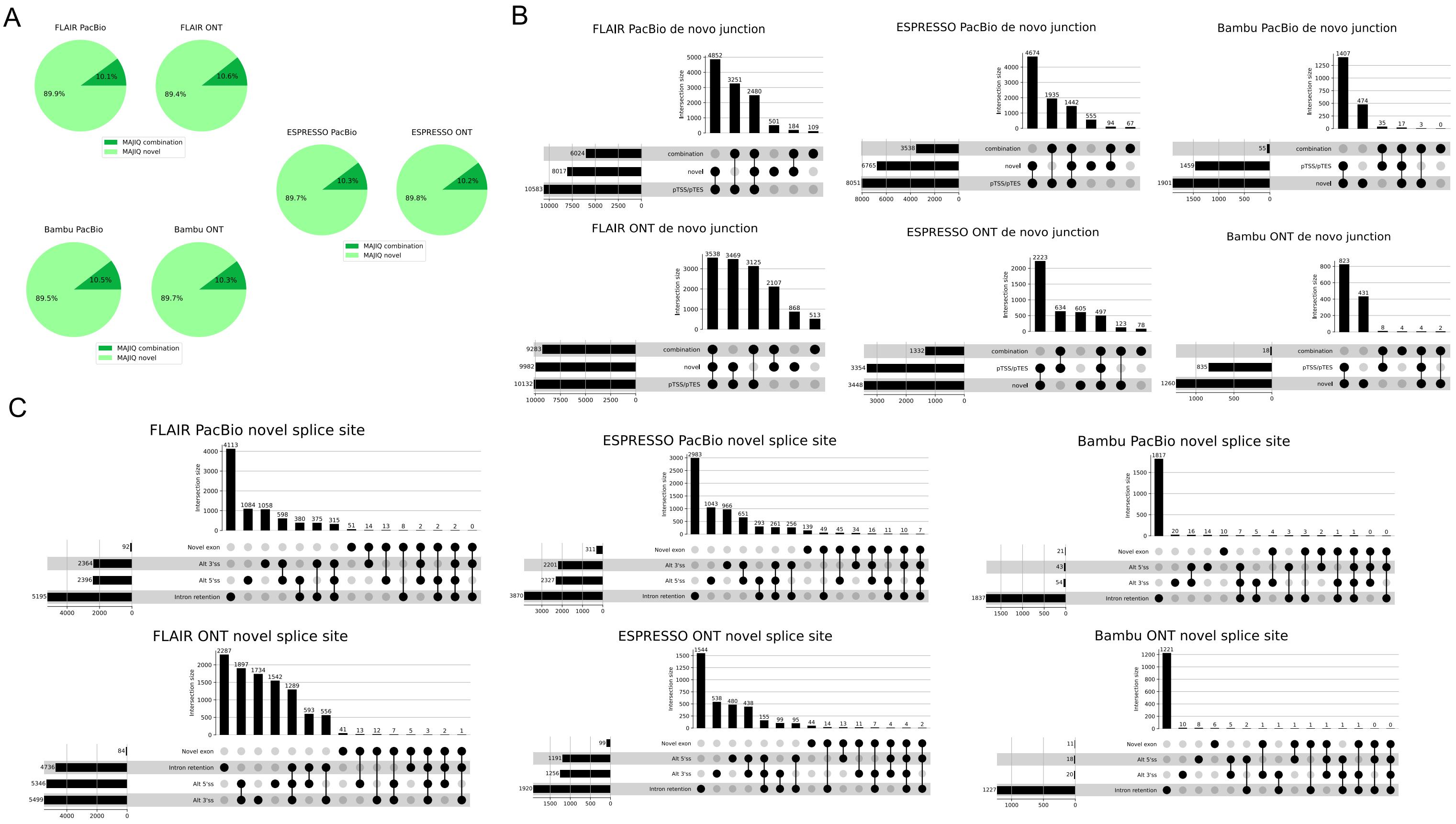
Supplementary Figures



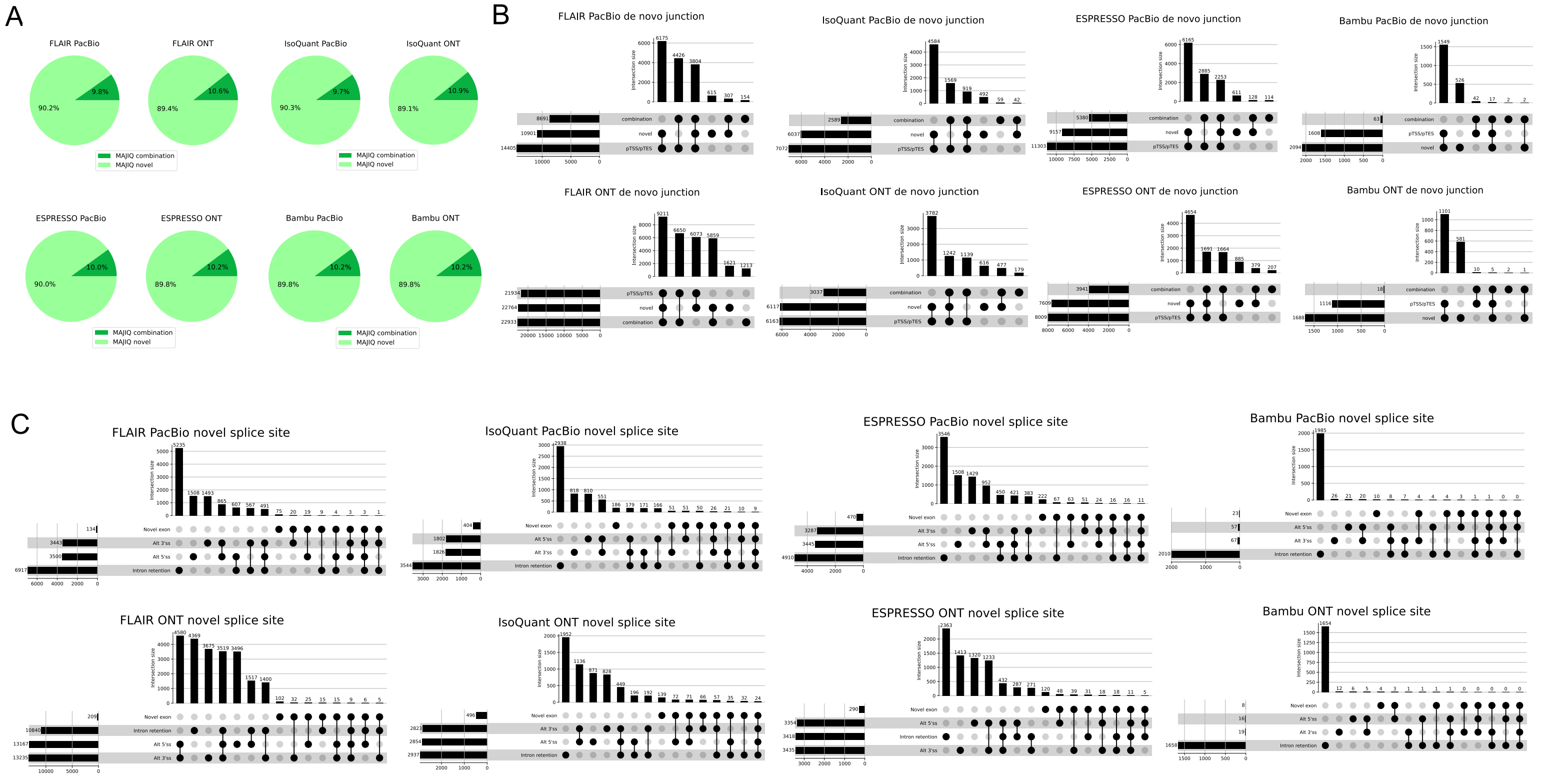
Supplemental Fig. S1. Splice junctions comparative analysis in GTEx and PDX cell line samples. Bar charts corresponding to the aforementioned six categories. Mean and standard error bars are computed using matched datasets from three samples of human heart atrial appendage, brain frontal cortex, and liver sequenced by GTEx. PDX cell line derived from a patient with a relapsed B-ALL contains only one sample. This data includes short reads processed by STAR and MAJIQ, Long reads from ONT assays, and four long read algorithms used to process the long reads data. Note that heart atrial appendage, brain frontal cortex, and liver samples are only processed with FLAIR.



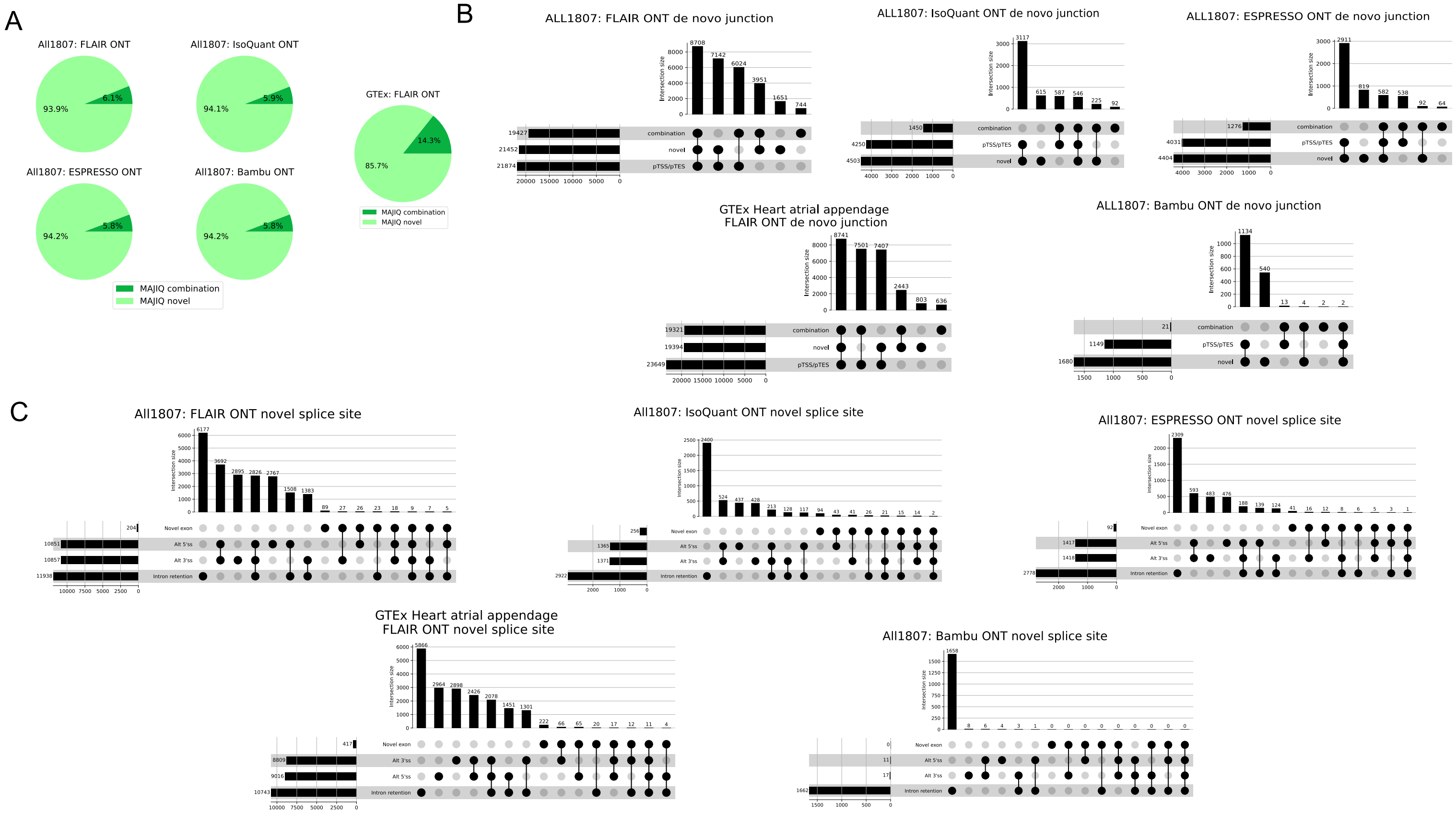
Supplemental Fig. S2. Additional splice junctions comparative analysis when there are coverage differences between short and long reads. (A) Bar charts corresponding to the aforementioned six categories. Mean and standard error bars are computed using matched datasets from three samples of human heart atrial appendage, brain frontal cortex, and liver sequenced by GTEx. This data includes short reads processed by STAR and MAJIQ, Long reads from ONT assays. GTEx samples are only processed with FLAIR. Note that Illumina has 1.7-fold more coverage than ONT in three GTEx tissues. **(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 the LRGASP14. 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. Note that PacBio has 1.3-fold and ONT has 2.4-fold more coverage than Illumina in this figure. **(C)** Bar charts corresponding to the aforementioned six categories. Mean and standard error bars are computed using matched datasets from one PDX cell line sample derived from a patient with a relapsed B-ALL. This data includes short reads processed by STAR and MAJIQ, Long reads from ONT assays, and four long read algorithms used to process the long reads data. Note that ONT has 2.9-fold more coverage than Illumina in PDX cell line sample.



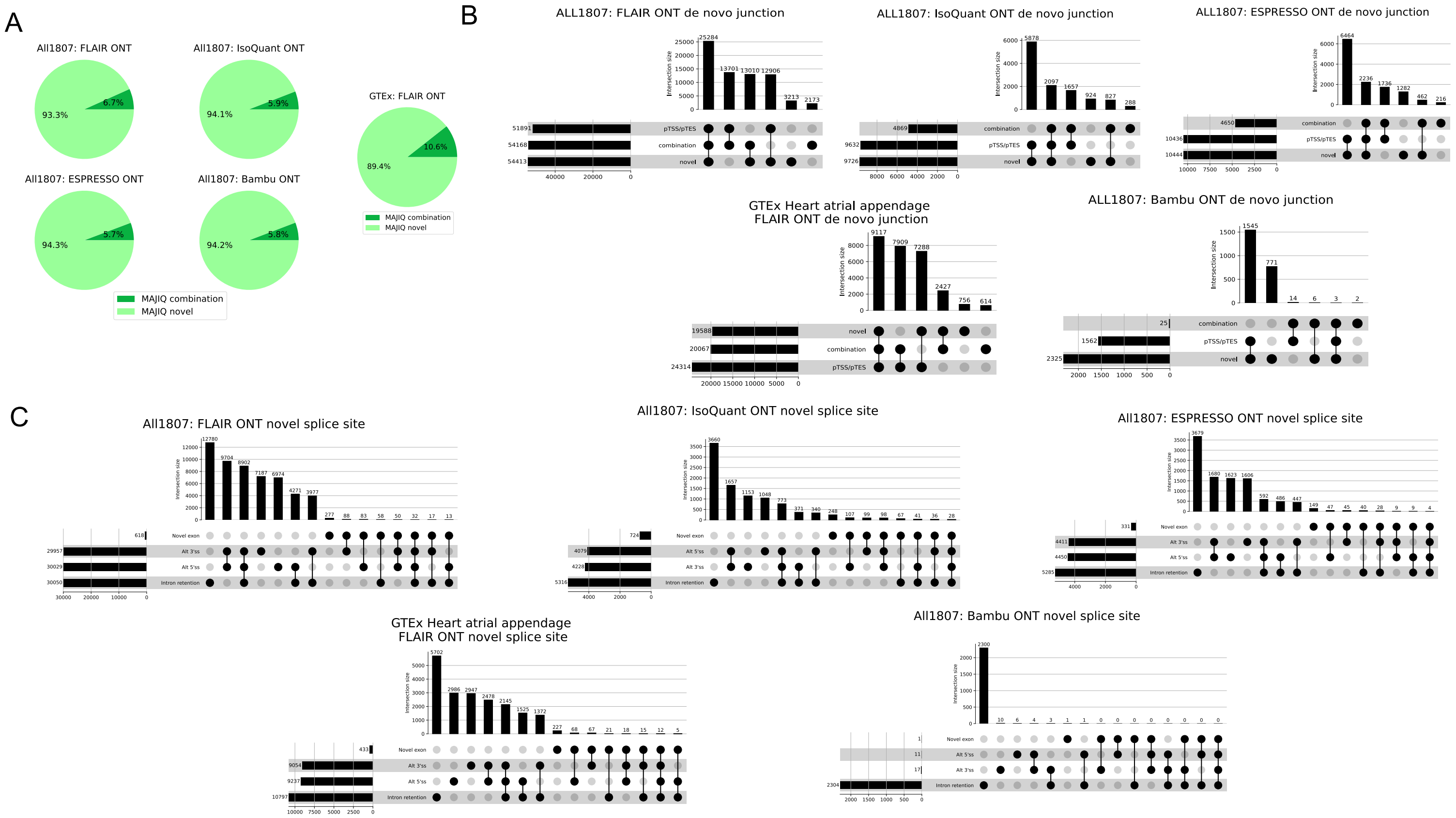
Supplemental Fig. S3. Additional analysis of *de novo* elements in LRGASP dataset. (A) Short reads *de novo* splice junctions reported by MAJIQ (green junctions) can be classified as those involving novel splice sites (light green) or a novel combination of known splice sites (dark green). For clarity, de novo junctions extend from exon 1 to exon 3, whereas the annotated junctions are constitutive. The pie chart shows that compared to long reads processed with FLAIR, ESPRESSO, and Bambu, ∼ 90% of MAJIQ *de novo* splice junctions involve novel splice sites. **(B)** Breakdown of all cases involving *de novo* junctions reported by FLAIR, ESPRESSO, and Bambu using either PacBio (top) or ONT (bottom) long reads. Notably, almost all of those cases also include pTSS/pTES. **(C)** Breakdown of long reads novel splice junctions (light purple in Fig 3-b) into the four different categories shown in Fig 3-d when using FLAIR, ESPRESSO, and Bambu to analyze PacBio (top) and ONT (bottom) matched reads.



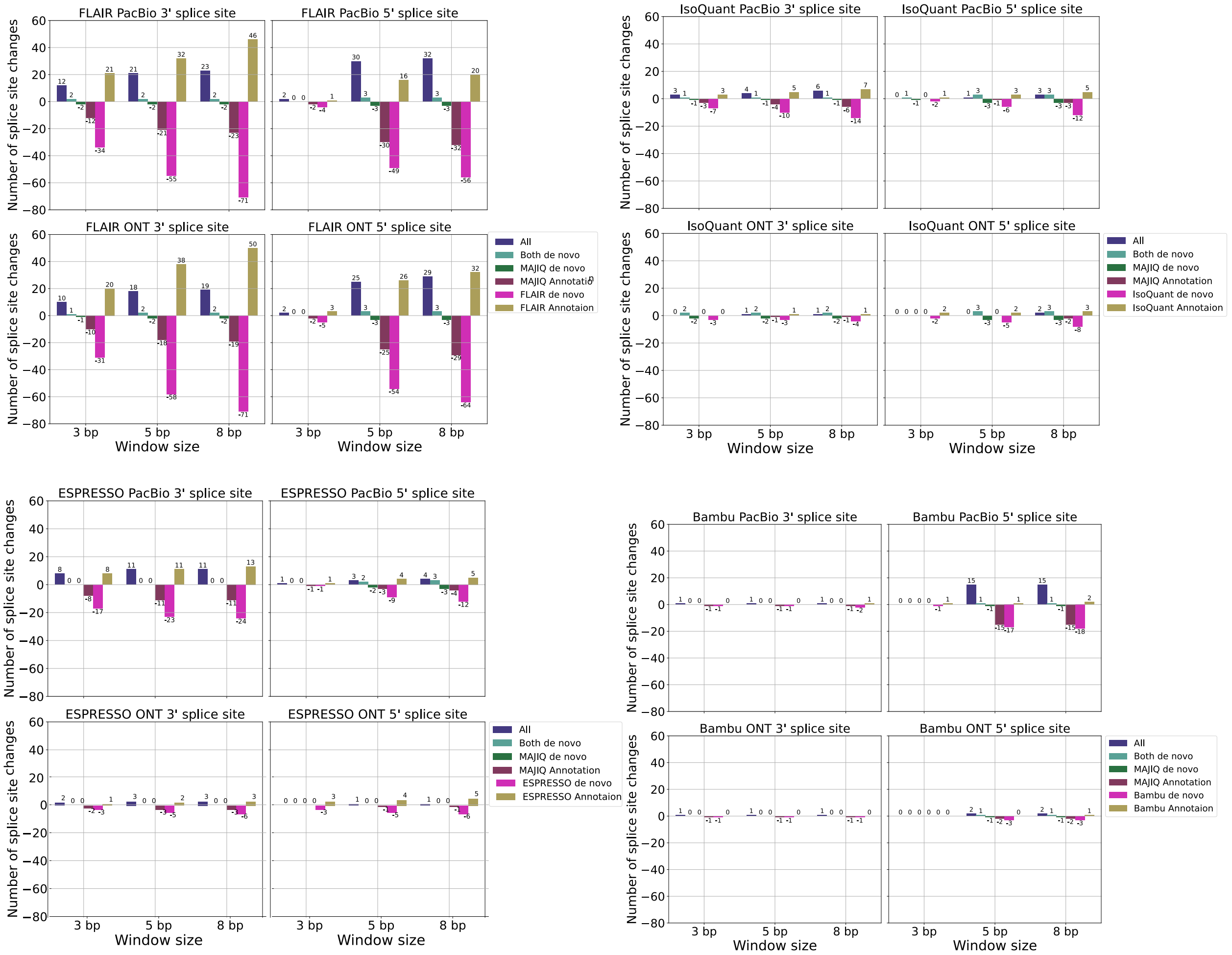
Supplemental Fig. S4. Additional analysis of *de novo* elements in LRGASP when there are coverage differences between short and long reads. (A) Short reads *de novo* splice junctions reported by MAJIQ (green junctions) can be classified as those involving novel splice sites (light green) or a novel combination of known splice sites (dark green). The pie chart shows that compared to long reads processed with IsoQuant, FLAIR, ESPRESSO, and Bambu, ∼ 90% of MAJIQ *de novo* splice junctions involve novel splice sites. **(B)** Breakdown of all cases involving *de novo* junctions reported by FLAIR, ESPRESSO, and Bambu using either PacBio (top) or ONT (bottom) long reads. Notably, almost all of those cases also include pTSS/pTES. **(C)** Breakdown of long reads novel splice junctions (light purple in Fig 3-b) into the four different categories shown in Fig 3-d when using FLAIR, ESPRESSO, and Bambu to analyze PacBio (top) and ONT (bottom) matched reads. Note that PacBio has 1.3-fold and ONT has 2.4-fold more coverage than Illumina in figures (a)-(c).



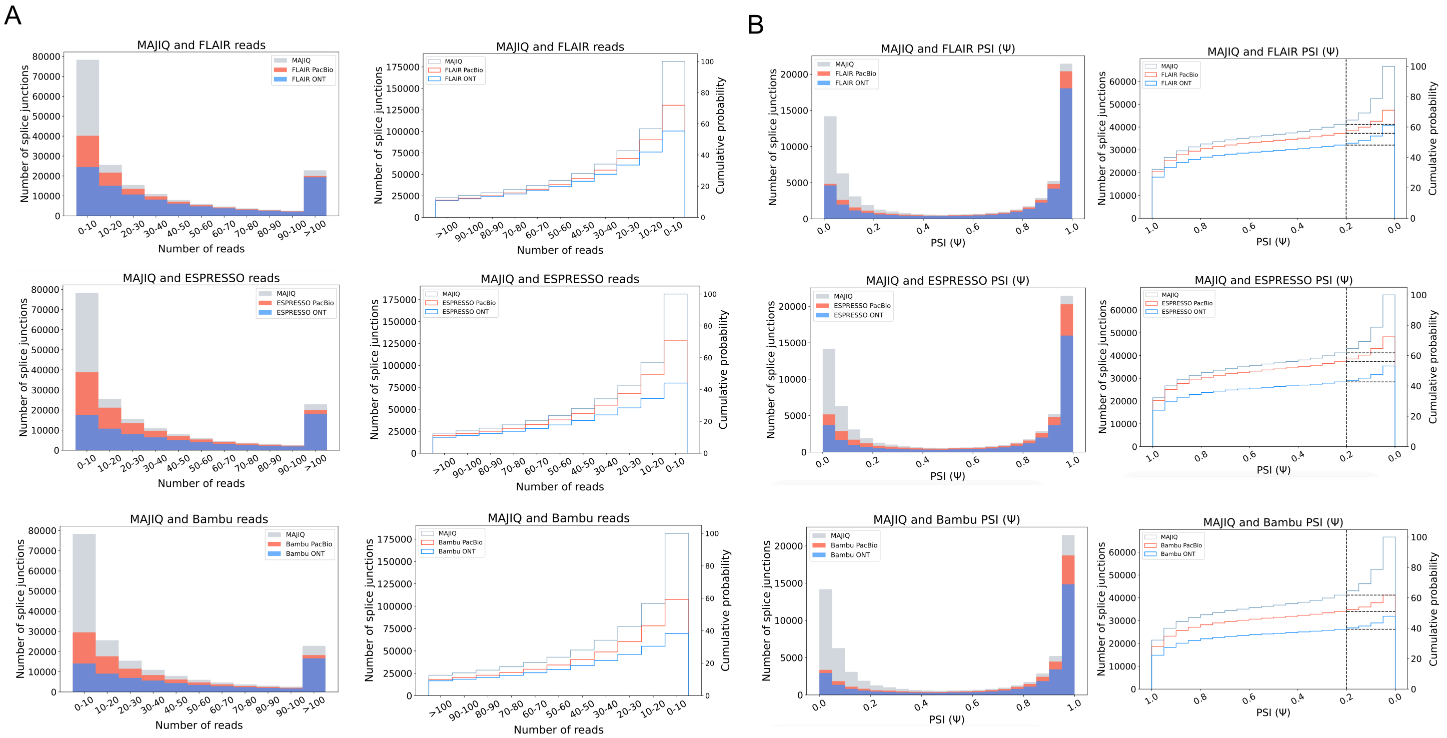
Supplemental Fig. S5. Additional analysis of *de novo* elements in GTEx heart atrial appendage and PDX cell line samples (A) Short reads *de novo* splice junctions reported by MAJIQ (green junctions) can be classified as those involving novel splice sites (light green) or a novel combination of known splice sites (dark green). The pie chart shows that compared to long reads processed with IsoQuant, FLAIR, ESPRESSO, and Bambu, ∼ 94% of MAJIQ *de novo* splice junctions involve novel splice sites in PDX cell line sample. Compared to long reads processed with FLAIR, ∼ 85% of MAJIQ *de novo* splice junctions involve novel splice sites in heart atrial appendage samples. **(B)** Breakdown of all cases involving *de novo* junctions reported by IsoQuant, FLAIR, ESPRESSO, and Bambu using ONT long reads. Notably, almost all of those cases also include pTSS/pTES. **(C)** Breakdown of long reads novel splice junctions (light purple in Fig 3-b) into the four different categories shown in Fig 3-d when using IsoQuant, FLAIR, ESPRESSO, and Bambu to analyze ONT matched reads. Note that heart atrial appendage samples are only processed with FLAIR.



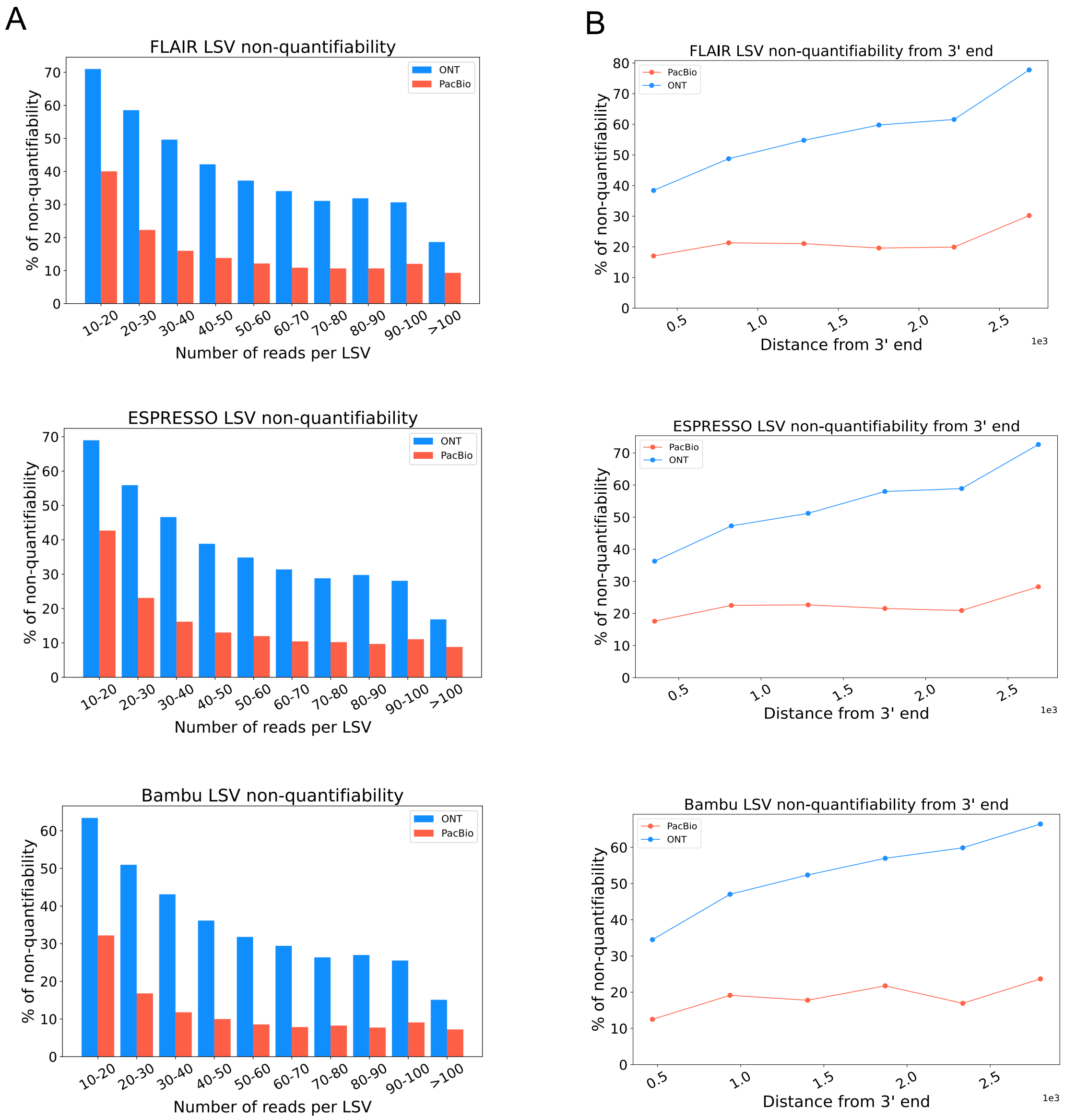
Supplemental Fig. S6. Additional analysis of *de novo* elements in GTEx heart atrial appendage and PDX cell line samples when there are coverage differences between short and long reads. (A) Short reads *de novo* splice junctions reported by MAJIQ (green junctions) can be classified as those involving novel splice sites (light green) or a novel combination of known splice sites (dark green). The pie chart shows that compared to long reads processed with IsoQuant, FLAIR, ESPRESSO, and Bambu, ∼ 94% of MAJIQ *de novo* splice junctions involve novel splice sites in PDX cell line sample. Compared to long reads processed with FLAIR, ∼ 90% of MAJIQ *de novo* splice junctions involve novel splice sites in heart atrial appendage samples. **(B)** Breakdown of all cases involving *de novo* junctions reported by IsoQuant, FLAIR, ESPRESSO, and Bambu using ONT long reads. Notably, almost all of those cases also include pTSS/pTES. **(C)** Breakdown of long reads novel splice junctions (light purple in Fig 3-b) into the four different categories shown in Fig 3-d when using IsoQuant, FLAIR, ESPRESSO, and Bambu to analyze ONT matched reads. Heart atrial appendage samples are only processed with FLAIR. Note that ONT has 2.9-fold more coverage than Illumina in PDX cell line sample, and Illumina has 1.7-fold more coverage than ONT in heart atrial appendage samples.



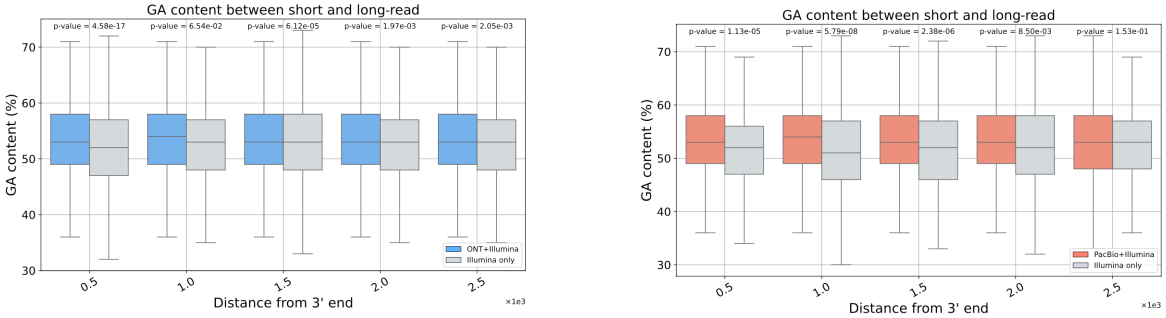
Supplementary Fig. 7: Analysis of splice site changes. Changes in the number of splice sites associated with each category (color) as a function of ’fuzzy’ matching window size. Here the window size (x-axis) represents the distance between long reads based splice sites and those reported by MAJIQ or the annotation at which they are still considered to match. As the window size increases the number of splice junctions in the categories All (blue), Both de novo (light green), or long reads and annotation (olive) increases, while the categories for junctions short reads and annotation (green) or detected only by long reads (magenta) drop. However the total number of splice junctions switching their categories remains small. From all four long reads algorithms FLAIR (top left) was the most affected by the ’fuzzy’ matching.



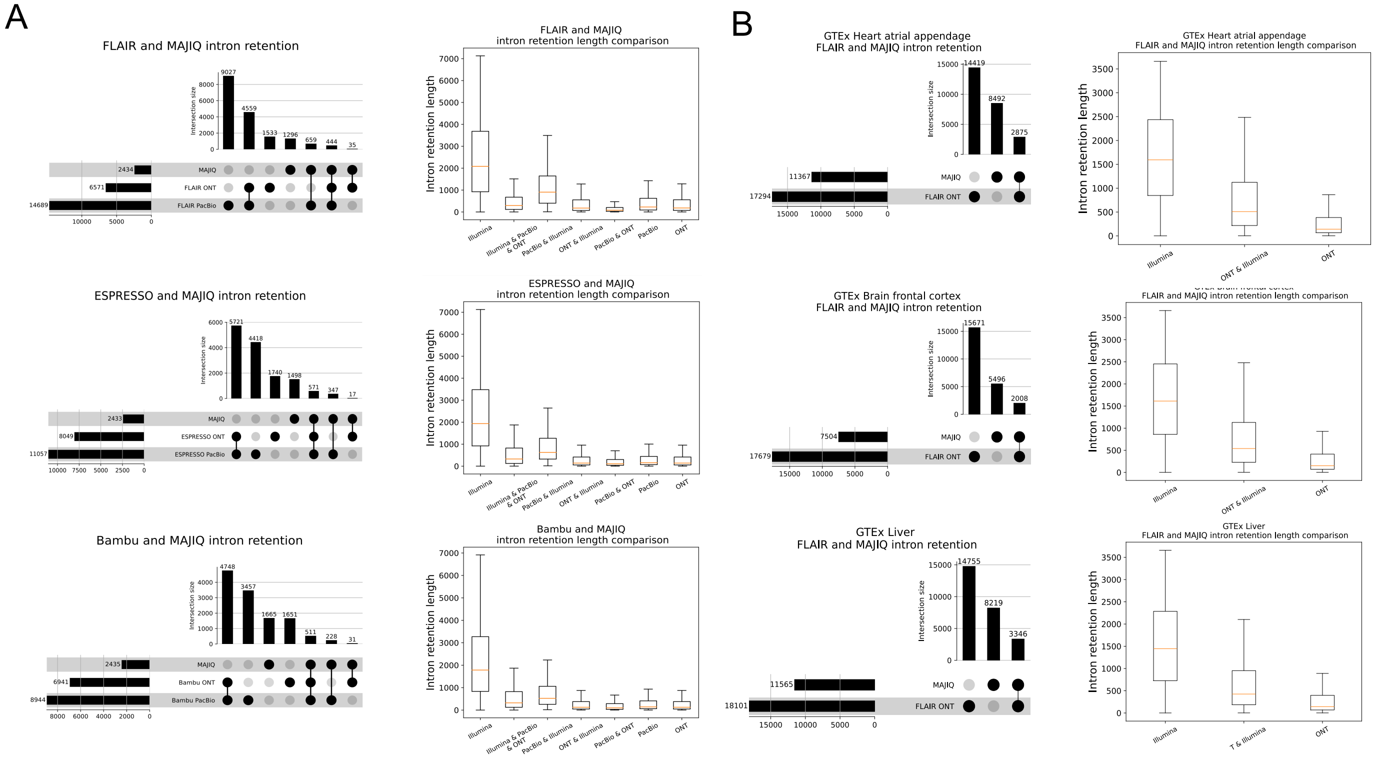
Supplemental Fig. S8. Number of junctions identified by FLAIR, ESPRESSO, and Bambu from the junctions that MAJIQ finds. **(A)** Bar plots showing the fraction of LSV reported by MAJIQ's short reads analysis, which were 'non-quantifiable' by FLAIR, ESPRESSO, and Bambu using PacBio (orange) and ONT (light 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 unjustifiable by long reads even for those with extremely high short read coverage (>100 reads). **(B)** 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 FLAIR, ESPRESSO, Bambu were 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.



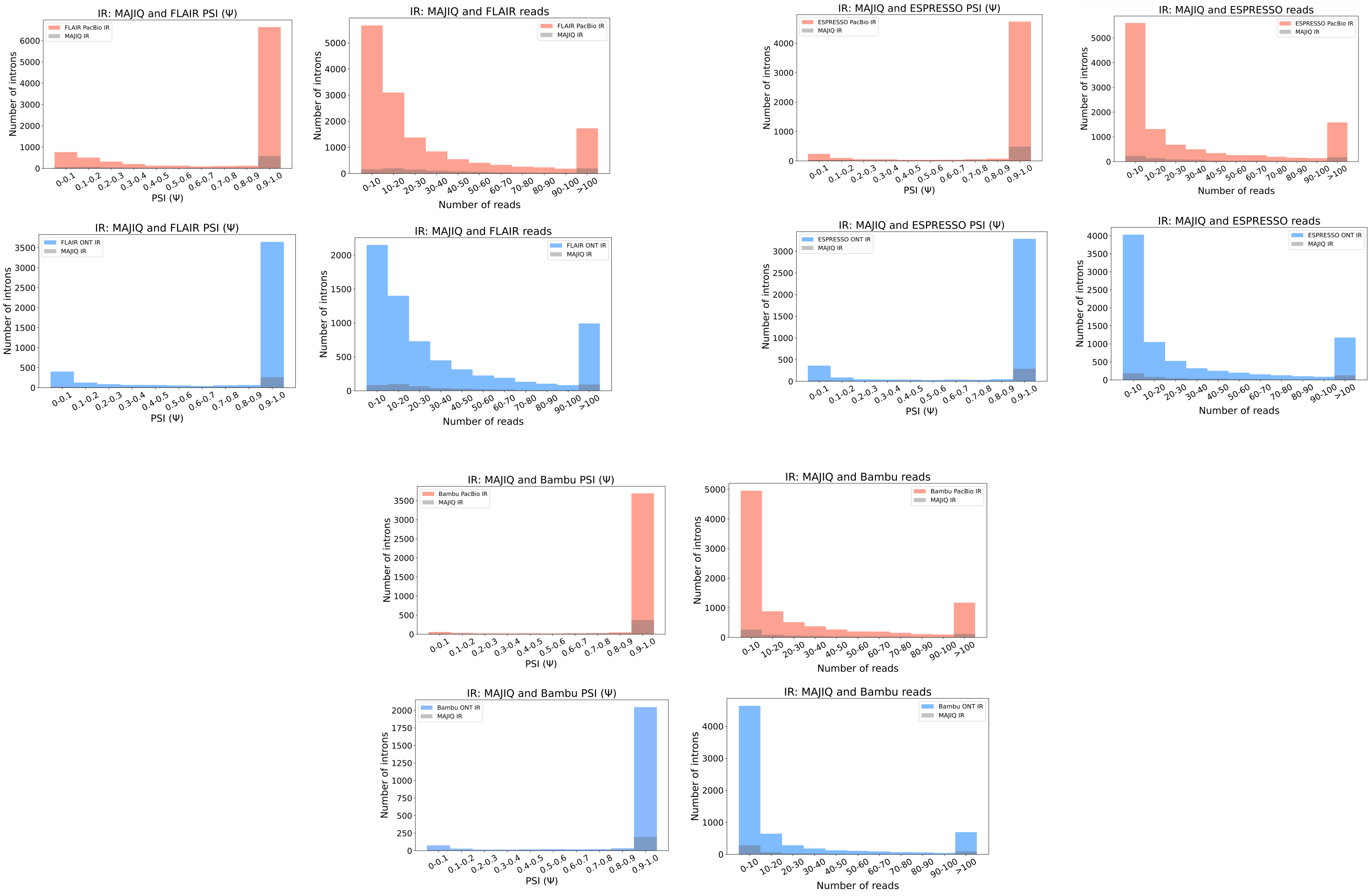
Supplemental Fig. S9. LSV non-quantifiability and 3' to 5' bias analysis in FLAIR, ESPRESSO, and Bambu. **(A)** Bar plots showing the fraction of LSV reported by MAJIQ's short reads analysis, which were 'non-quantifiable' by FLAIR, ESPRESSO, Bambu using PacBio (orange) and ONT (light blue) matched long reads data. Here a 'quantifiable' LSV requires at least 10 reads covering its respective junctions. Of note, a substantial fraction of LSV remain unjustifiable by long reads even for those with extremely high short read coverage (>100 reads). **(B)** Same plot as in (A) 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.



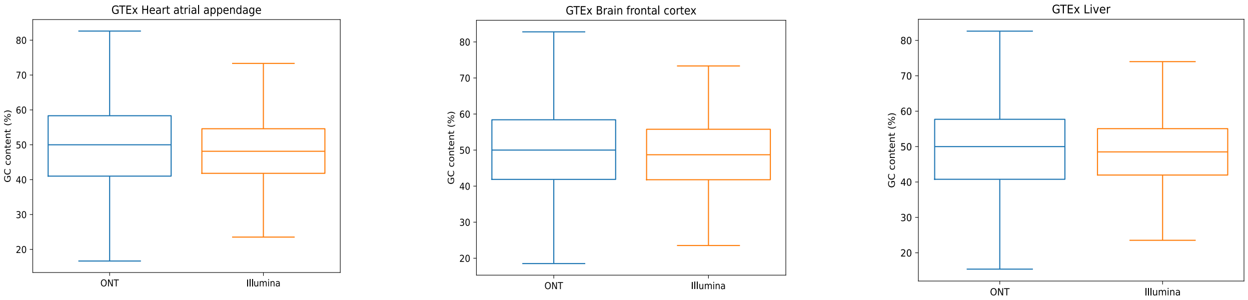
Supplemental Fig. S10. GA content at splice junctions between short and long reads. Boxplots showing GA content across various distances from the transcript 3' end. Each boxplot represents the GA content in distance from 3' end (x-axis). The median is denoted by the horizontal line in each box, the upper 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.



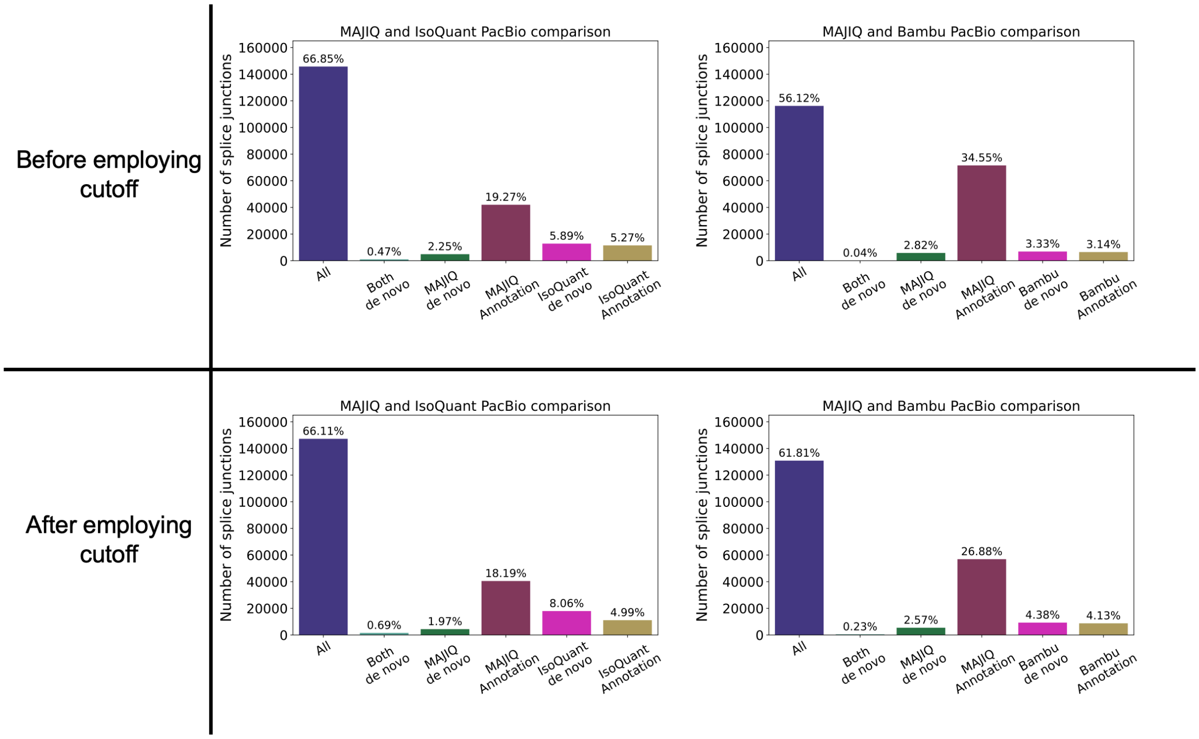
Supplemental Fig. S11. Intron Rention (IR) events found between MAJIQ and FLAIR, ESPRESSO, Bambu in GTEx and LRGASP samples. (A) Upset plot (left) showing overlap and total IR events reported by MAJIQ from short reads and MAJIQ and FLAIR, ESPRESSO, Bambu using PacBio or ONT matched long reads (LRGASP dataset). Boxplots (right) showing IR length distribution across seven categories in upsetplot. 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 upsetplot. **(B)** Upset plot (left) showing overlap and total IR events reported by MAJIQ from short reads and MAJIQ and FLAIR using ONT matched long reads (GTEx). Boxplots (right) showing IR length distribution across three categories in upsetplot. 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 upsetplot.



Supplemental Fig. S12. Number of introns identified by MAJIQ from the introns that FLAIR, ESPRESSO, Bambu find. Taking the introns by FLAIR, ESPRESSO, Bambu PacBio (tomato) or ONT (blue) and assessing the number of those also identified by MAJIQ (grey) as a function of the PSI values (left). Note that if an intron appears multiple times, the lowest PSI values are chosen. The histogram shows the number of introns (y-axis) in each PSI value (x-axis). The number of FLAIR, ESPRESSO, Bambu's introns using PacBio or ONT identified by MAJIQ as a function of the number of long reads covering the introns (right). The histogram shows the number of introns (y-axis) as a function of read number (x-axis).



Supplemental Fig. S13. GC content of Intron Retention (IR) in GTEx samples. Boxplots showing GC content of IR between short and long reads in GTEx tissues. Each boxplot represents GC content (y-axis) in each technology (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.



Supplemental Fig. S14. Comparative Analysis of long read algorithms before and after enforcing default cutoff thresholds. We experimented with relaxing the cutoffs for isoform detection across four different algorithms on a human cell line from LRGASP PacBio dataset. In general, transcript discovery involves a balance between sensitivity and precision, and each algorithm uses different parameters to control for that:

- Bambu uses Novel Discovery Rate (NDR) threshold for isoform detection. The NDR threshold approximates the proportion of novel candidates output by bambu, relative to the number of known transcripts it found, i.e., an NDR of 0.1 would mean that 10% of all transcripts passing the threshold are classified as novel. We increased NDR from 0.1 (default setting) to 1.0.

- IsoQuant can discover more transcripts at a cost of precision using model_construction_strategy parameter. We included this parameter in our new run.

- ESPRESSO sets the minimum perfect read count for de novo detected candidate splice junctions to be 2. We lowered the threshold to 1.

- FLAIR can correct long-read alignments using matched short-read junctions to improve transcript discovery.

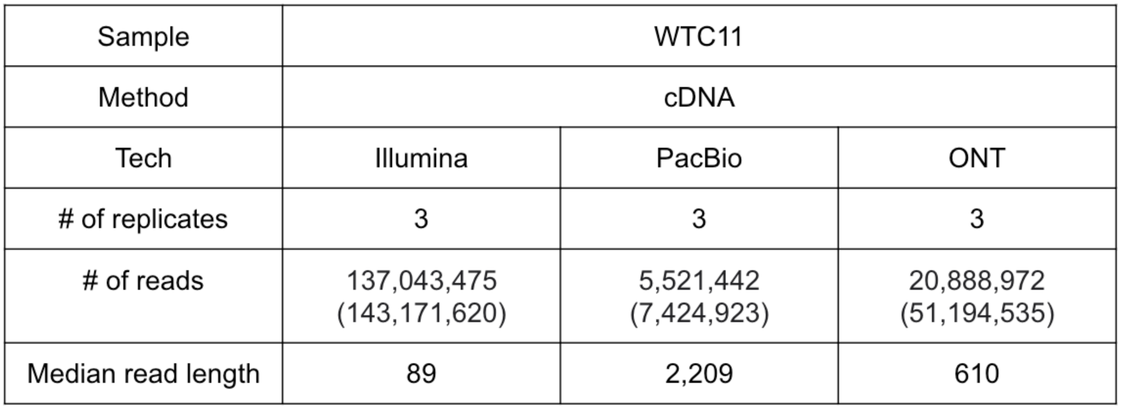
We tried varying the above parameters. For FLAIR, incorporating short-read junctions did not lead to any improvement in transcript discovery. We received feedback from the FLAIR author suggesting that including short-read junctions might not enhance the correction step if both annotated and Illumina-derived splice junctions are already being used. As for ESPRESSO, it encountered a memory overload and was subsequently terminated on several occasions. Thus, we focus on the results we got for the two remaining algorithms.

Before relaxing the cutoff, Bambu identifies 6,500 Bambu & Annotation junctions and 6,900 de novo junctions. After the relaxation, the numbers increased to 8,750 and 9,278 junctions. While relatively these numbers mark a significant increase of $\sim$34% in detected junction, the overall picture compared to the annotation and short reads remains quite similar with detection by all increasing from 56% to close to 62%, MAJIQ and Annotation (no LR) dropping from $\sim$34% to $\sim$27%, and Bamboo only increasing from $\sim$3.5% to $\sim$4.5%.

For IsoQuant, prior to relaxing the cutoff, IsoQuant detects 11,491 IsoQuant & Annotation junctions and 12,839 IsoQuant only de novo junctions (no short reads). Upon easing the cutoff criteria, the count of IsoQuant & Annotation junctions marginally decreased by approximately 11,105 while IsoQuant only de novo junctions (no short reads) experienced a significant increase of roughly 40% to 17,949. Yet again, if we look at the overall picture with short reads, the picture remains quite similar: The fraction of junctions supported by all stays almost the same ($\sim$66.2%), SR+Annotation drops from 19.3% to 18.2% and LR only increases from $\sim$5.9% to $\sim$8.1%.

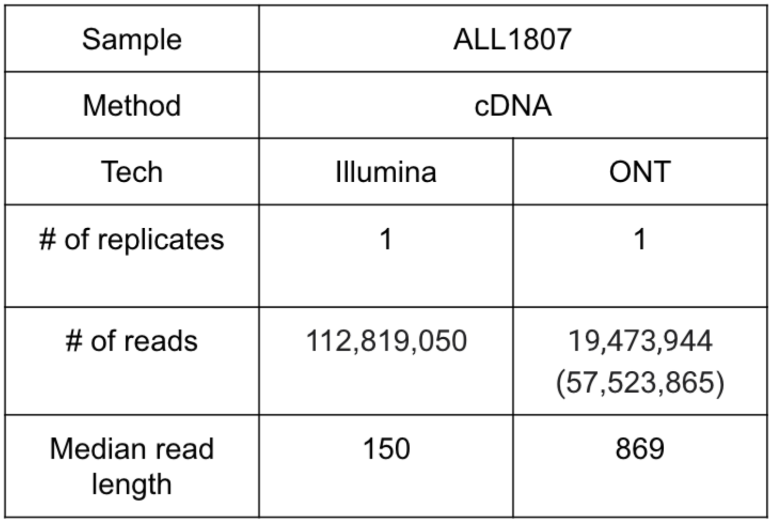
Overall, these results indicate that relaxing the default execution parameters of LR algorithms does make the expected change of increased sensitivity at a likely price of an increase in false positives, yet the overall picture/conclusions regarding the relation between Annotation/SR/LR remains quite similar.

Supplementary Table



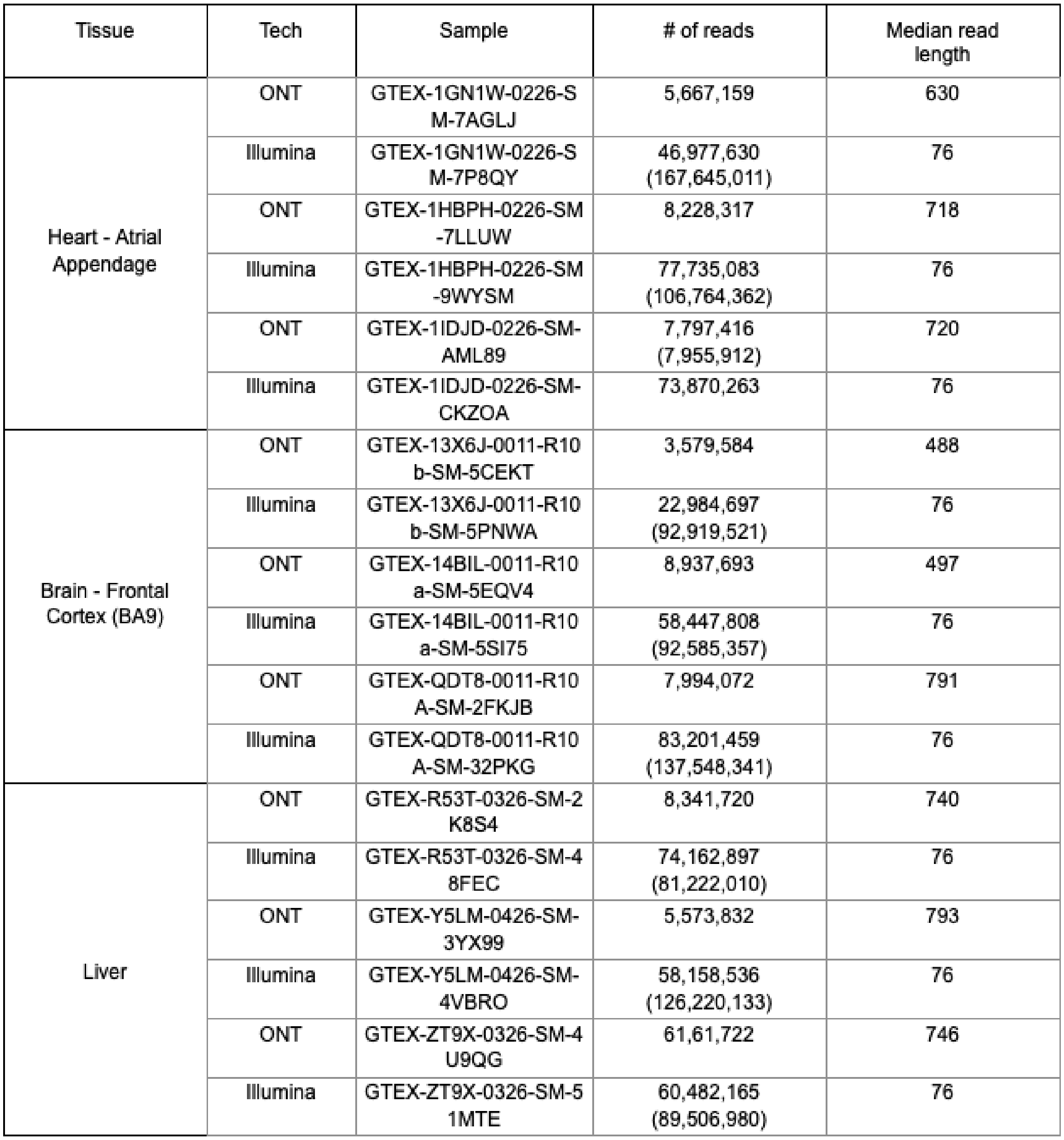
Sample	WTC11		
Method	cDNA		
Tech	Illumina	PacBio	ONT
# of replicates	3	3	3
# of reads	137,043,475 (143,171,620)	5,521,442 (7,424,923)	20,888,972 (51,194,535)
Median read length	89	2,209	610

Supplemental Table S1. Coverage summary statistics of human cell line datasets in LARGASP. For each sample, replicates were combined when reporting statistics. Note that the number inside the parentheses denotes the number of reads before subsampling. PacBio has 1.3-fold and ONT has 2.4-fold more coverage than Illumina.



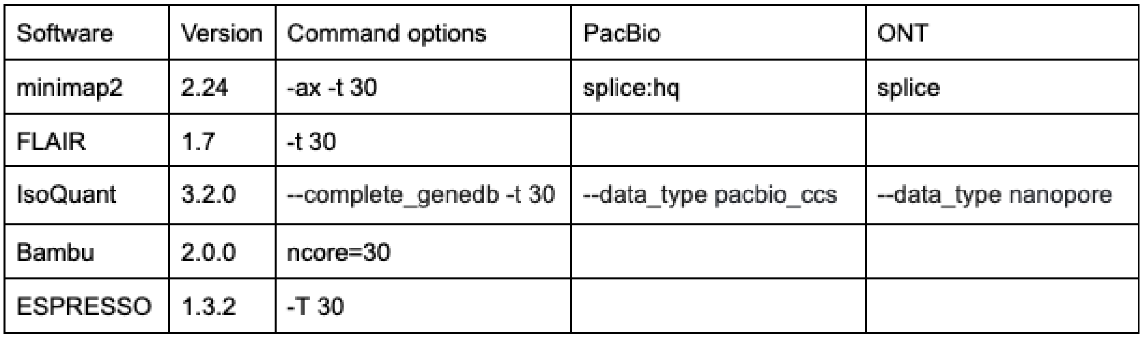
Sample	ALL1807	
Method	cDNA	
Tech	Illumina	ONT
# of replicates	1	1
# of reads	112,819,050	19,473,944 (57,523,865)
Median read length	150	869

Supplemental Table S2. Coverage summary statistics in B-ALL. Note that the number inside the parentheses denotes the number of reads before sub-sampling. ONT has 2.9-fold more coverage than Illumina.



Tissue	Tech	Sample	# of reads	Median read length
Heart - Atrial Appendage	ONT	GTEX-1GN1W-0226-SM-7AGLJ	5,667,159	630
	Illumina	GTEX-1GN1W-0226-SM-7P8QY	46,977,630 (167,645,011)	76
	ONT	GTEX-1HBPH-0226-SM-7LLUW	8,228,317	718
	Illumina	GTEX-1HBPH-0226-SM-9WYSM	77,735,083 (106,764,362)	76
	ONT	GTEX-1IDJD-0226-SM-AML89	7,797,416 (7,955,912)	720
	Illumina	GTEX-1IDJD-0226-SM-CKZOA	73,870,263	76
Brain - Frontal Cortex (BA9)	ONT	GTEX-13X6J-0011-R10b-SM-5CEKT	3,579,584	488
	Illumina	GTEX-13X6J-0011-R10b-SM-5PNWA	22,984,697 (92,919,521)	76
	ONT	GTEX-14BIL-0011-R10a-SM-5EQV4	8,937,693	497
	Illumina	GTEX-14BIL-0011-R10a-SM-5SI75	58,447,808 (92,585,357)	76
	ONT	GTEX-QDT8-0011-R10A-SM-2FKJB	7,994,072	791
	Illumina	GTEX-QDT8-0011-R10A-SM-32PKG	83,201,459 (137,548,341)	76
Liver	ONT	GTEX-R53T-0326-SM-2K8S4	8,341,720	740
	Illumina	GTEX-R53T-0326-SM-48FEC	74,162,897 (81,222,010)	76
	ONT	GTEX-Y5LM-0426-SM-3YX99	5,573,832	793
	Illumina	GTEX-Y5LM-0426-SM-4VBRO	58,158,536 (126,220,133)	76
	ONT	GTEX-ZT9X-0326-SM-4U9QG	61,61,722	746
	Illumina	GTEX-ZT9X-0326-SM-51MTE	60,482,165 (89,506,980)	76

Supplemental Table S3. Coverage summary statistics in GTEx. Note that the number inside the parentheses denotes the number of reads before sub-sampling. Illumina has 1.7-fold more coverage than ONT.



Software	Version	Command options	PacBio	ONT
minimap2	2.24	-ax -t 30	splice:hq	splice
FLAIR	1.7	-t 30		
IsoQuant	3.2.0	--complete_genedb -t 30	--data_type pacbio_ccs	--data_type nanopore
Bambu	2.0.0	ncore=30		
ESPRESSO	1.3.2	-T 30		

Supplemental Table S4. Long read tools Command line options and software versions FLAIR, IsoQuant, Bambu, and ESPRESSO were run using the same BAM file, reference annotation, and reference genome as input.

References

[1] 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.