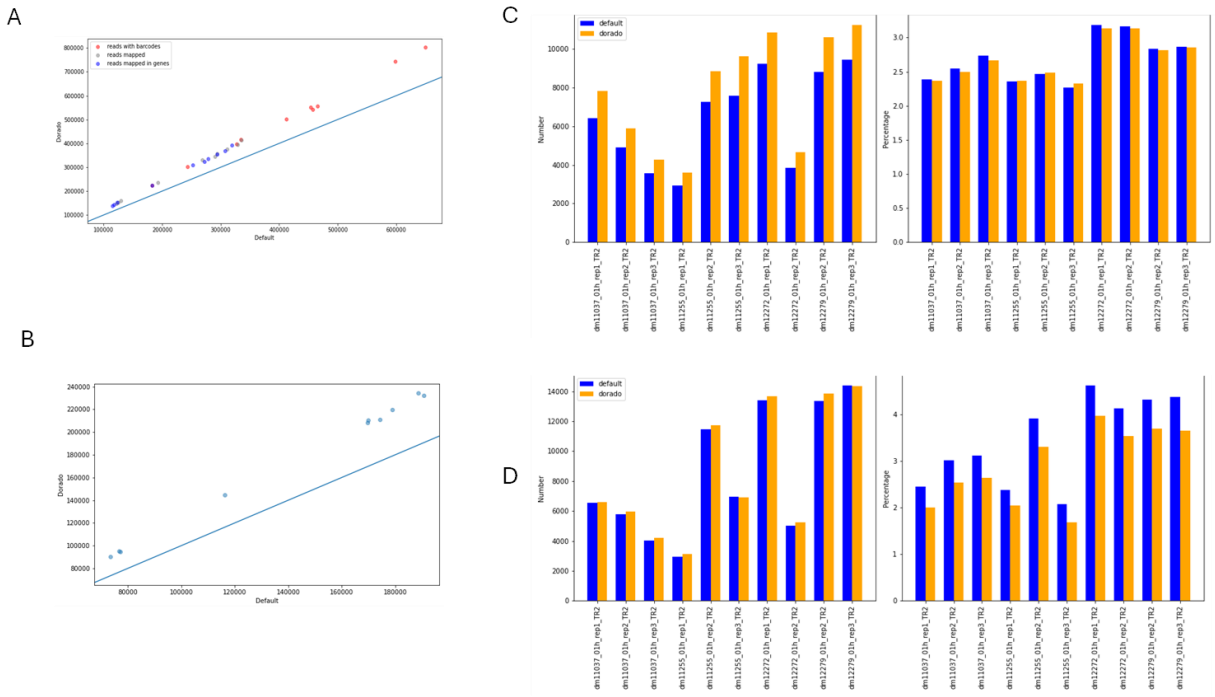
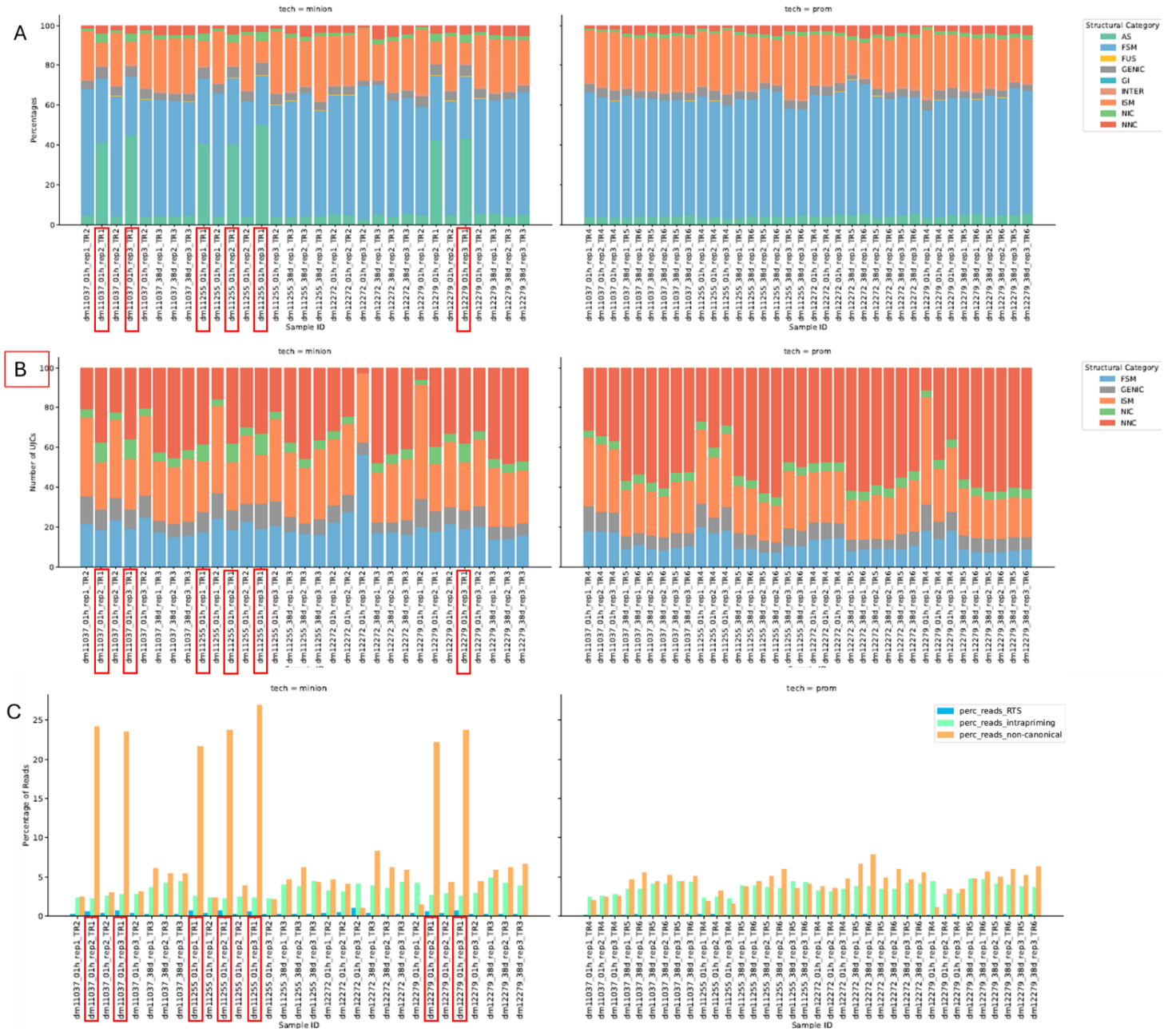
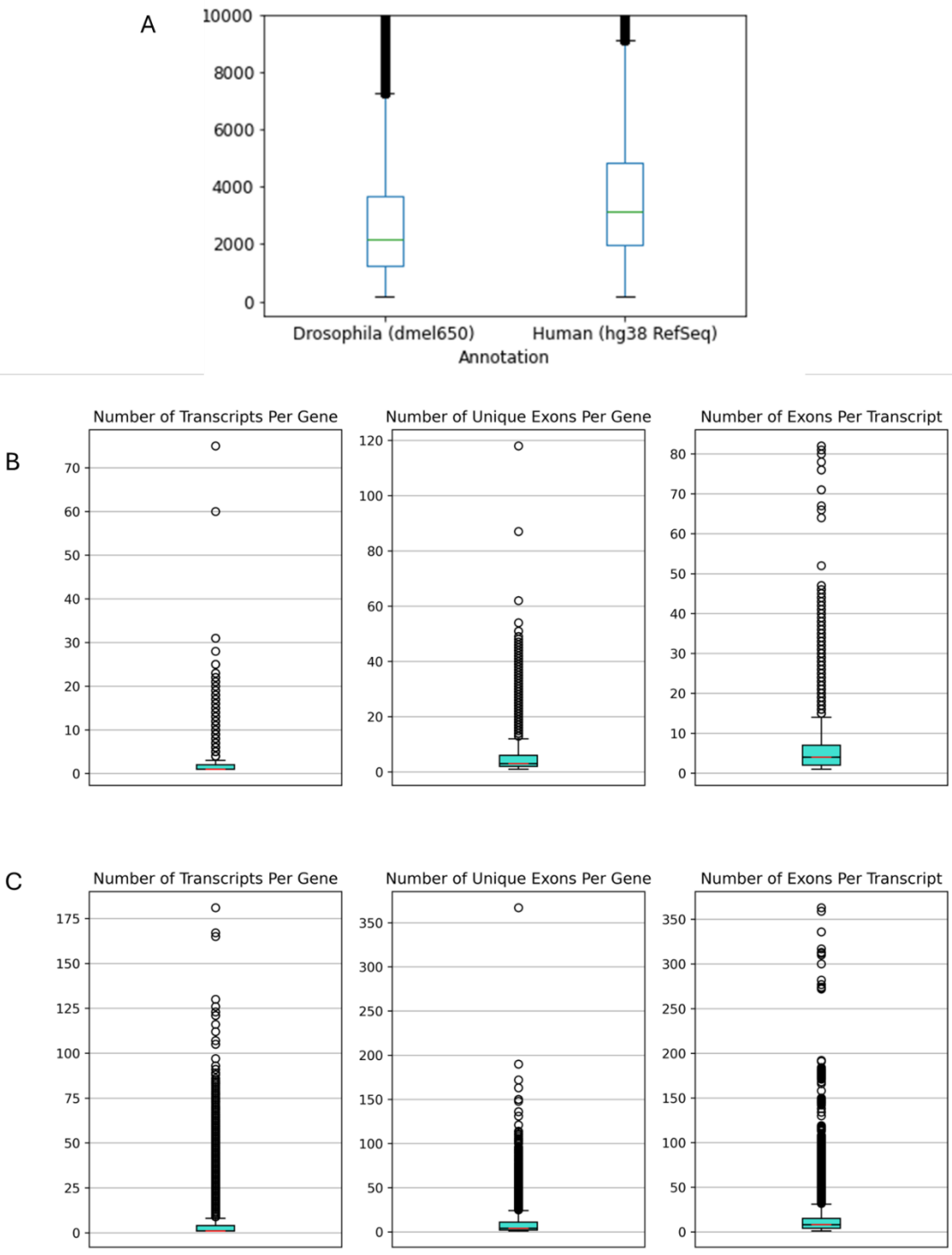
Table of Contents



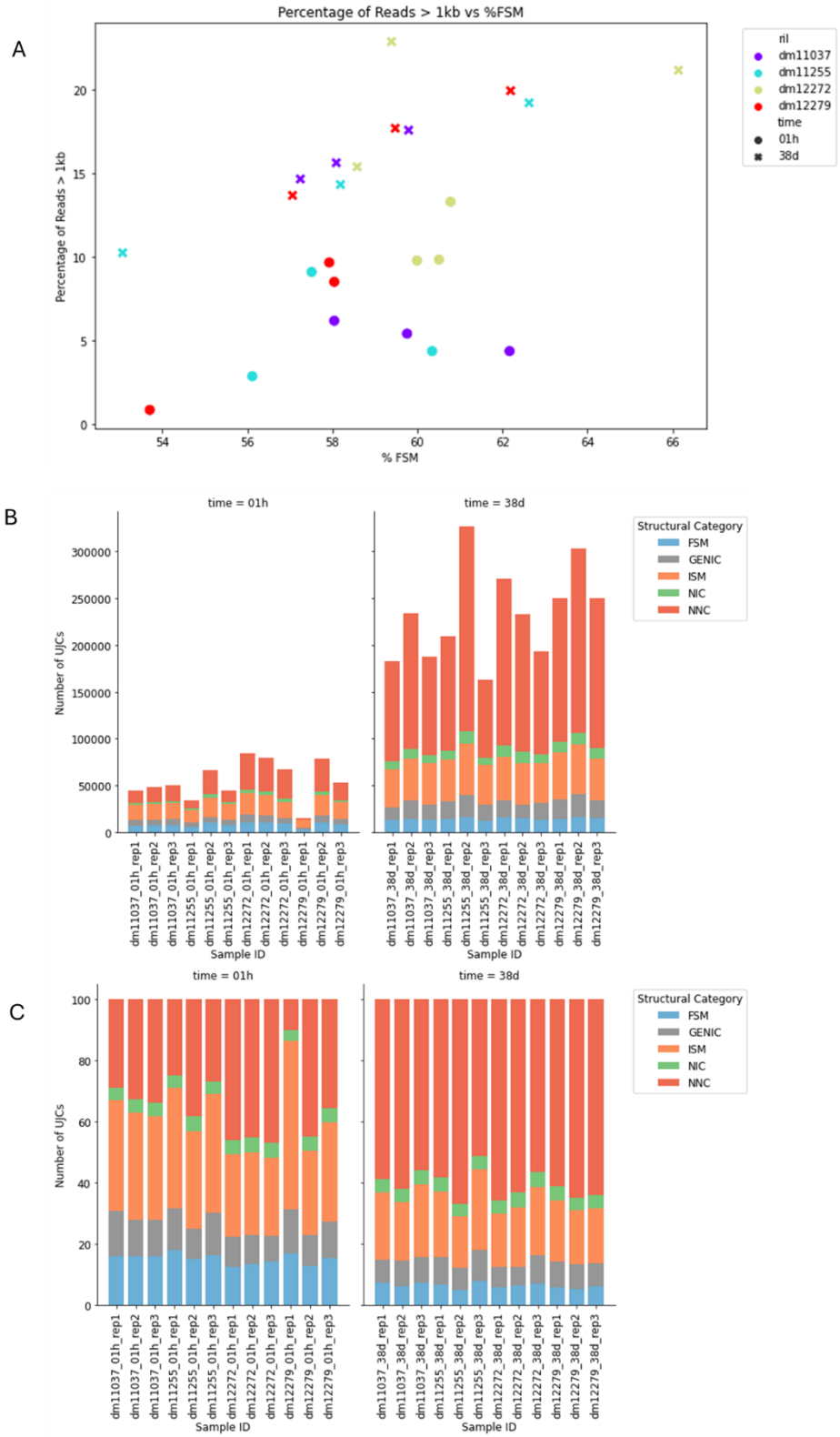
Supplementary Figure 1: Comparison of Guppy (default) and Dorado basecalling for technical replicate 2. A) Number of reads with assignable barcodes (red), number of mapped reads (grey) and number of reads mapped in genes (blue) in Dorado vs Guppy. B) Number of reads >2kb in Dorado vs Guppy C) Number and percentage of reads with evidence of intrapriming in Dorado and Guppy D) Number and percentage of reads with non-canonical junctions in Dorado and Guppy.



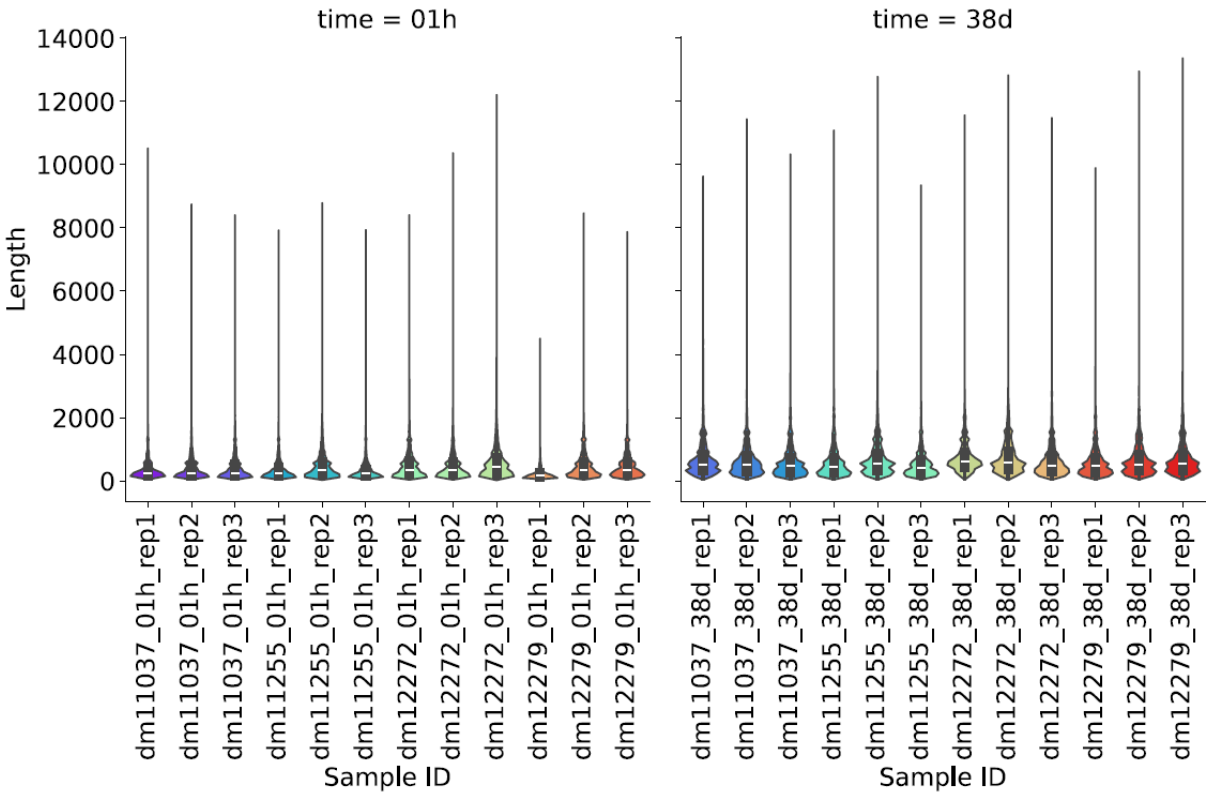
Supplementary Figure 2: *Drosophila* technical replicates basecalled with Dorado. A) Proportion of reads in each structural category. B) Proportion of unique junction chains (UJCs) in each structural category. C) Proportion of reads with evidence of RT-switching (blue), intrapriming (green) and non-canonical junctions (orange). Samples from technical replicate 1 are highlighted with red boxes.



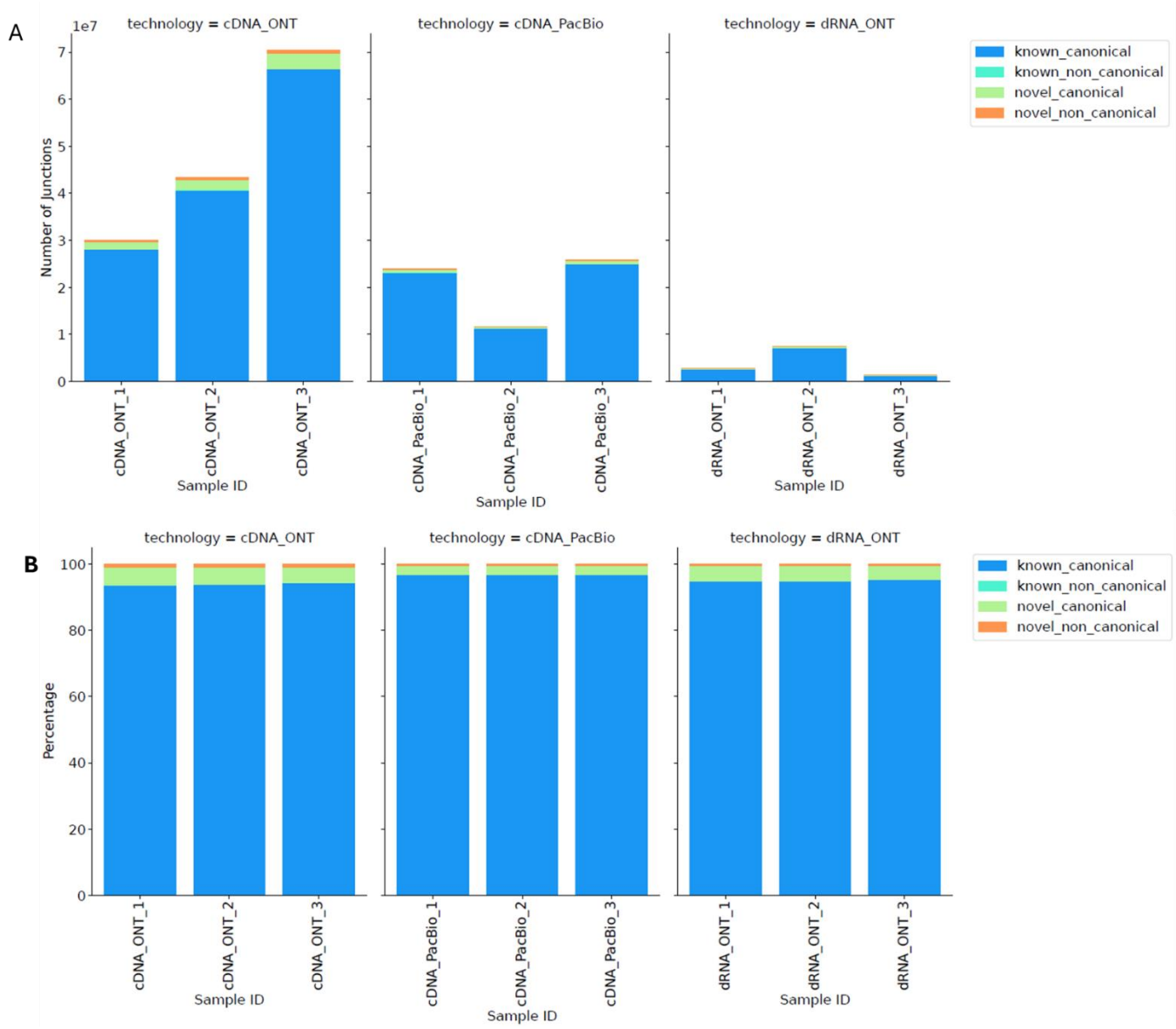
Supplementary Figure 3: Complexity Metrics of Drosophila (dmel650) and Human (hg38 Refseq) Annotations. A) Distribution of protein-coding transcript lengths in dmel650 and hg38 Refseq. B) Transcriptome Complexity Metrics for dmel650 C) Transcriptome complexity metrics for hg38 Refseq



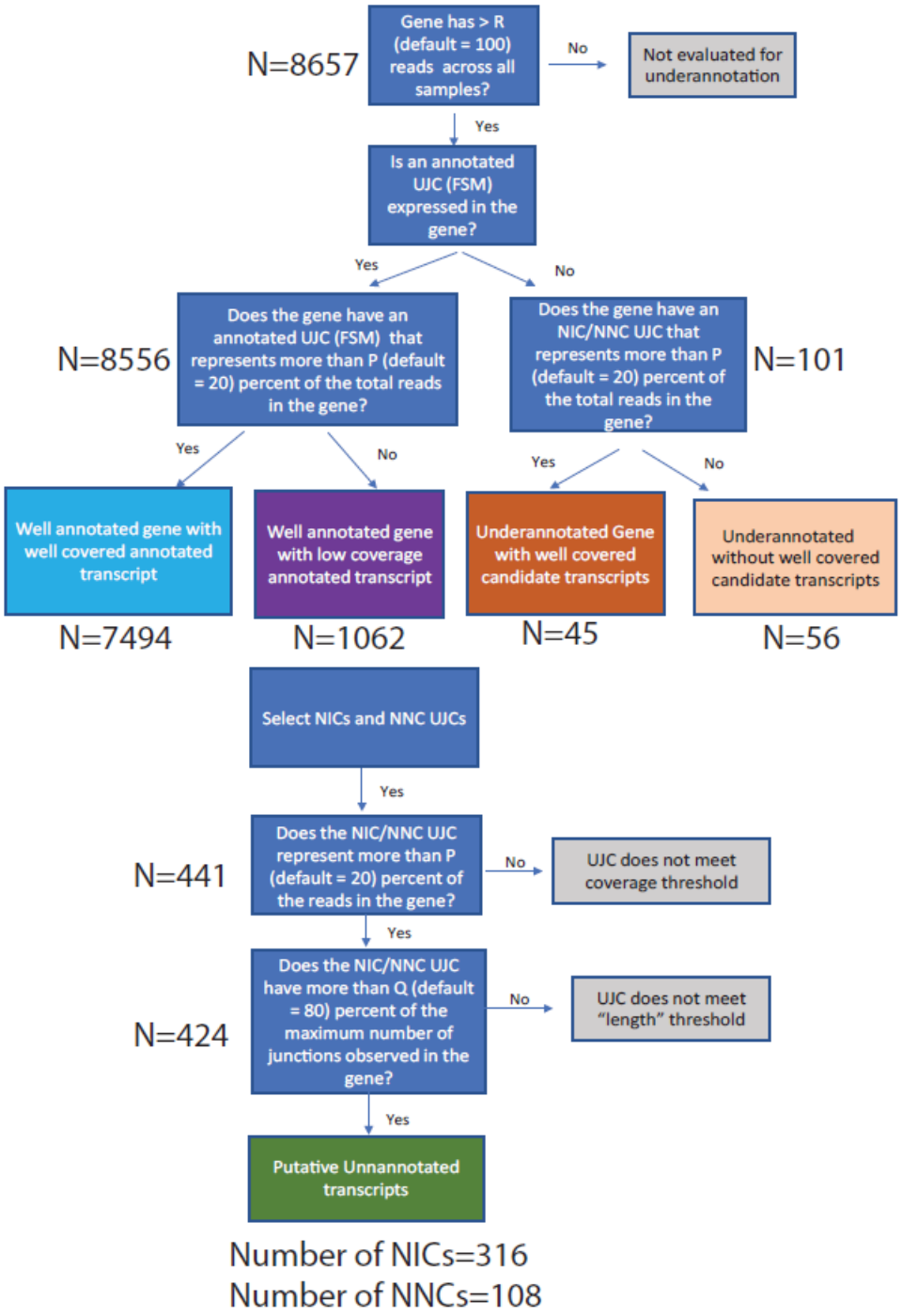
Supplementary Figure 4: SQANTI-reads plots for *Drosophila* samples basecalled with Dorado A) Percentage of reads greater than 1kb vs Percentage of reads classified as full-splice match (FSM) B) Number of UJCs detected coloured by structural category of the UJC. C) Proportion of UJCs coloured by structural category of the UJC



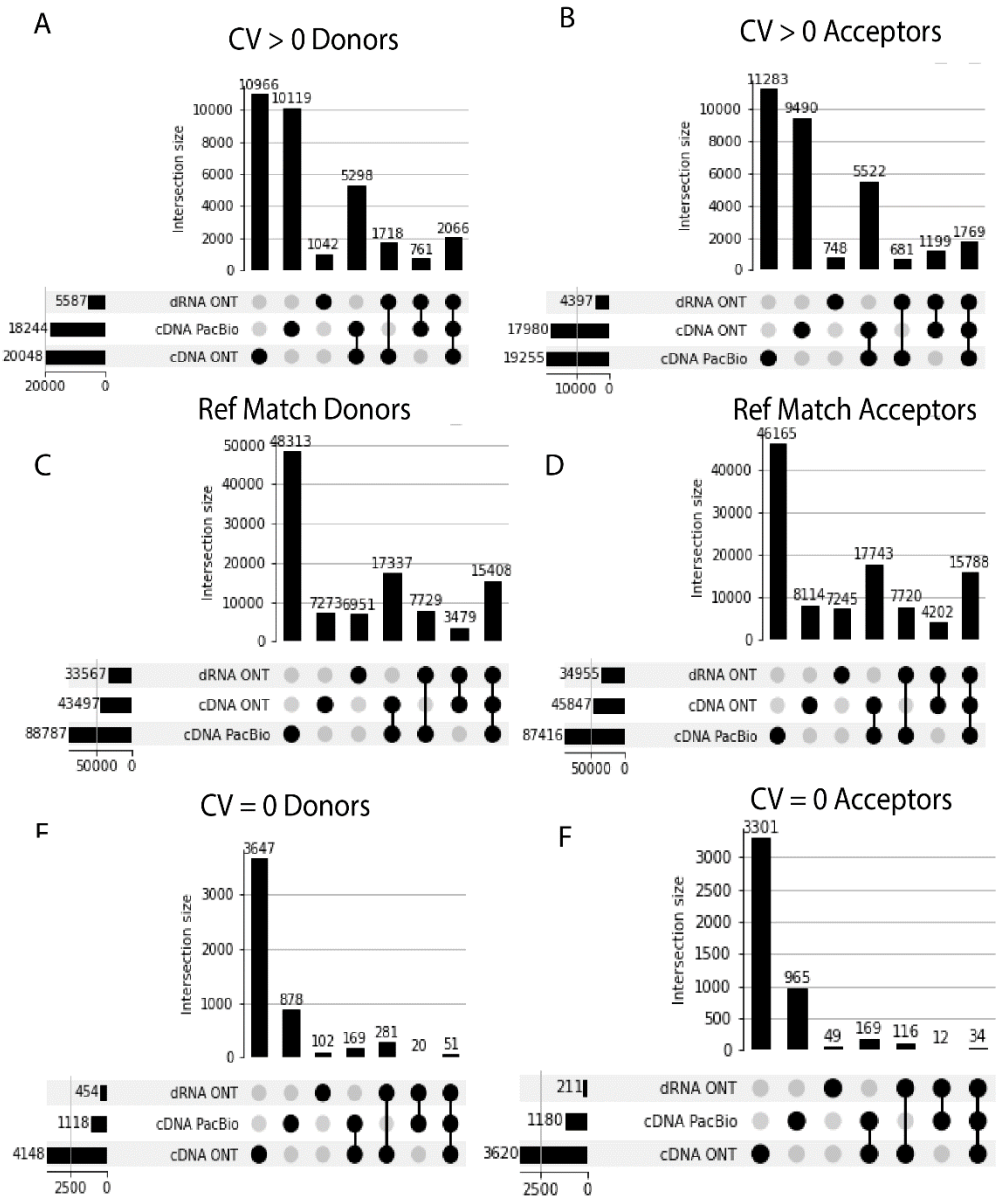
Supplementary Figure 5: SQANTI-reads violin plots of read length for *Drosophila* samples



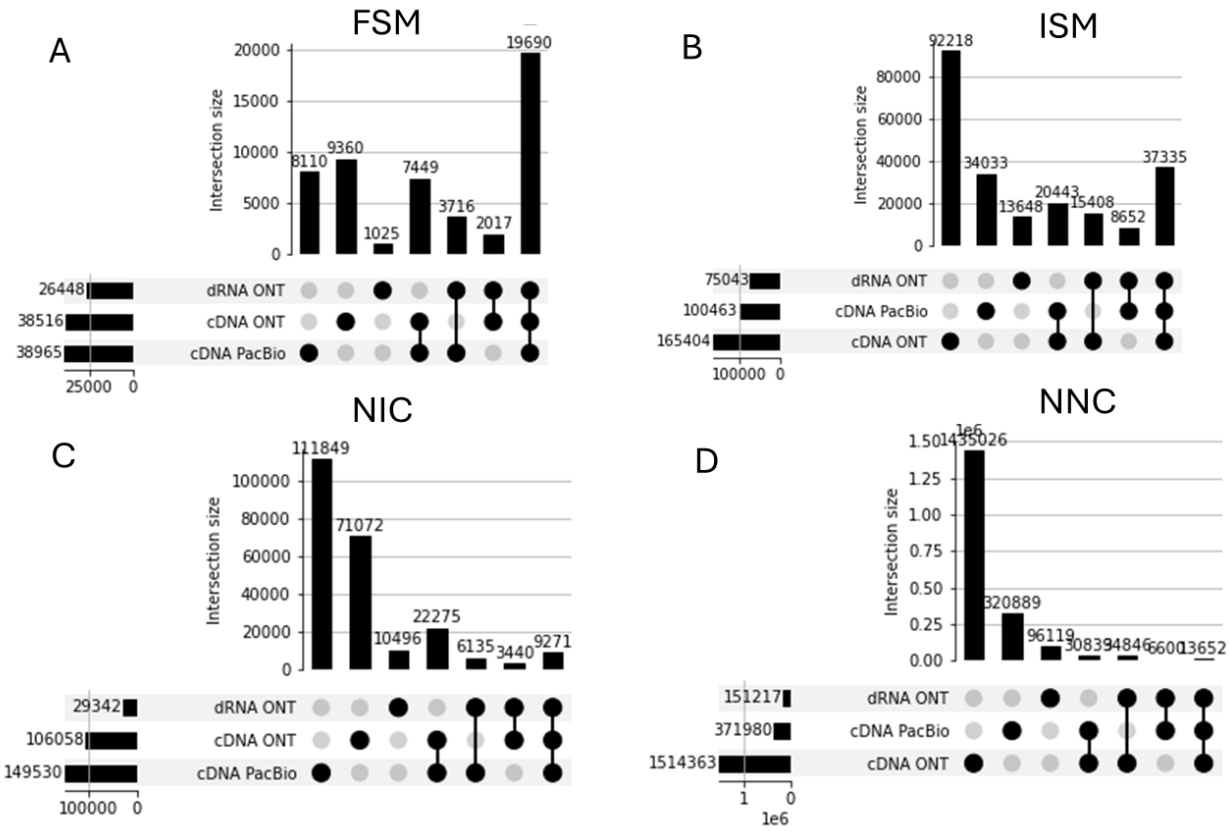
Supplementary Figure 6: Junction plots for WTC11 samples A) Number of junctions in reads coloured by junction categories (known-canonical, known-non-canonical, novel-canonical and novel-non-canonical) B) Proportion of junctions in reads in each junction category



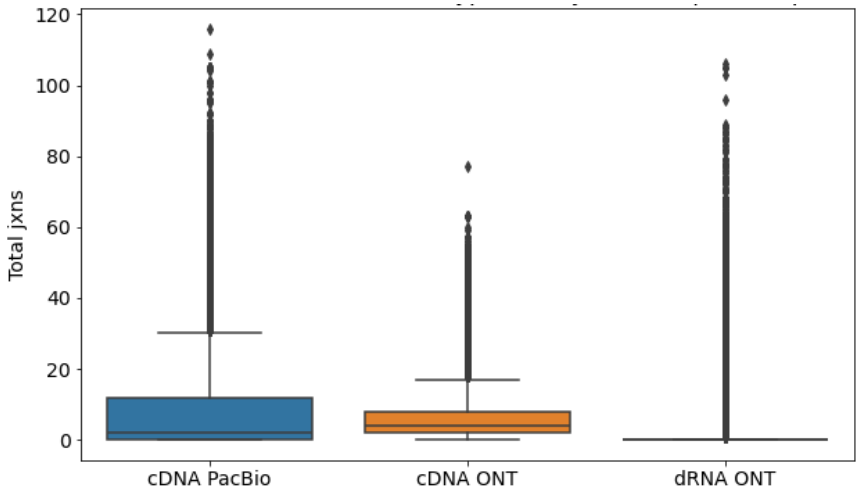
Supplementary Figure 7: Decision tree for classifying genes as well annotated or under-annotated and classifying transcripts as putative novel candidate transcripts annotated with the number of genes/transcripts classified at each step.



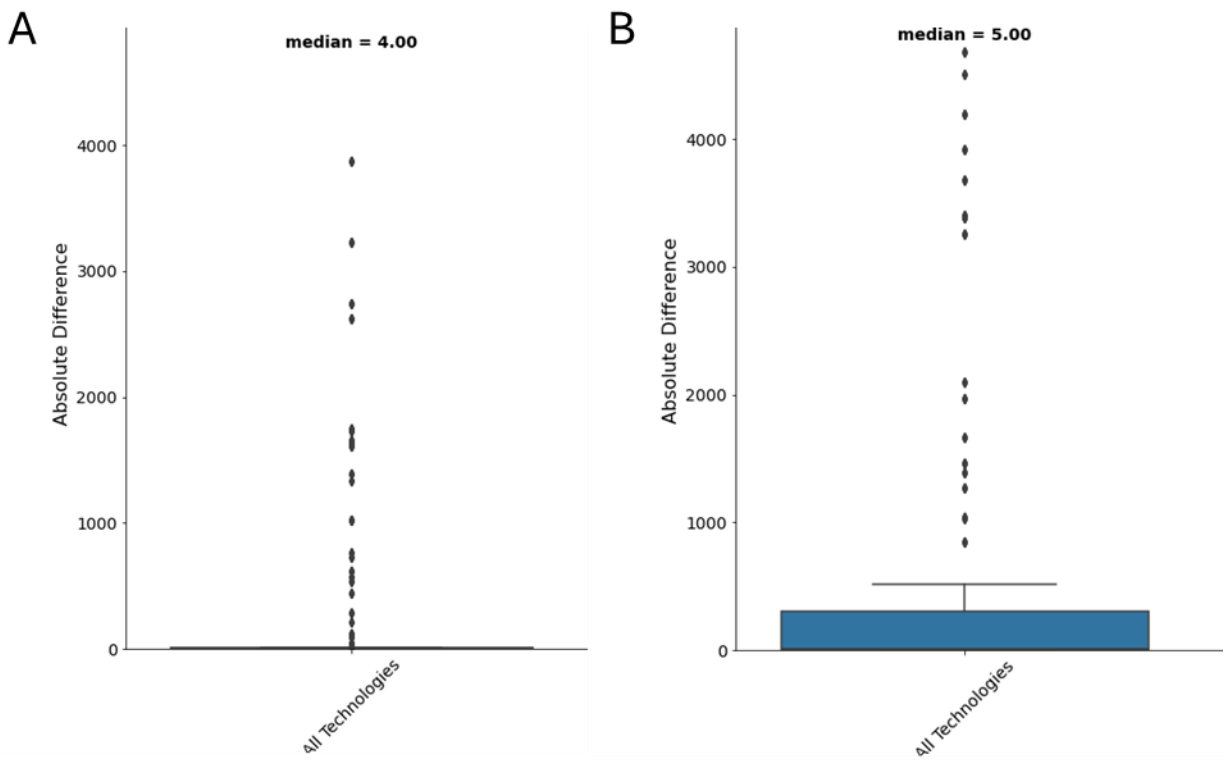
Supplementary Figure 8: Upset plot of donors and acceptors detected in replicate 1 of the WTC11 samples of each technology by variation category. CV > 0 donors (A) and acceptors (B) . Reference match donors (C) and acceptors (D). CV = 0 donors (E) and acceptors (F)



Supplementary Figure 9: Upset plot of multi-exon UJCs detected in each technology by structural category in the WTC11 samples. A) FSM UJCs B) ISM UJCs C) NIC UJCs and D) NNC UJCs

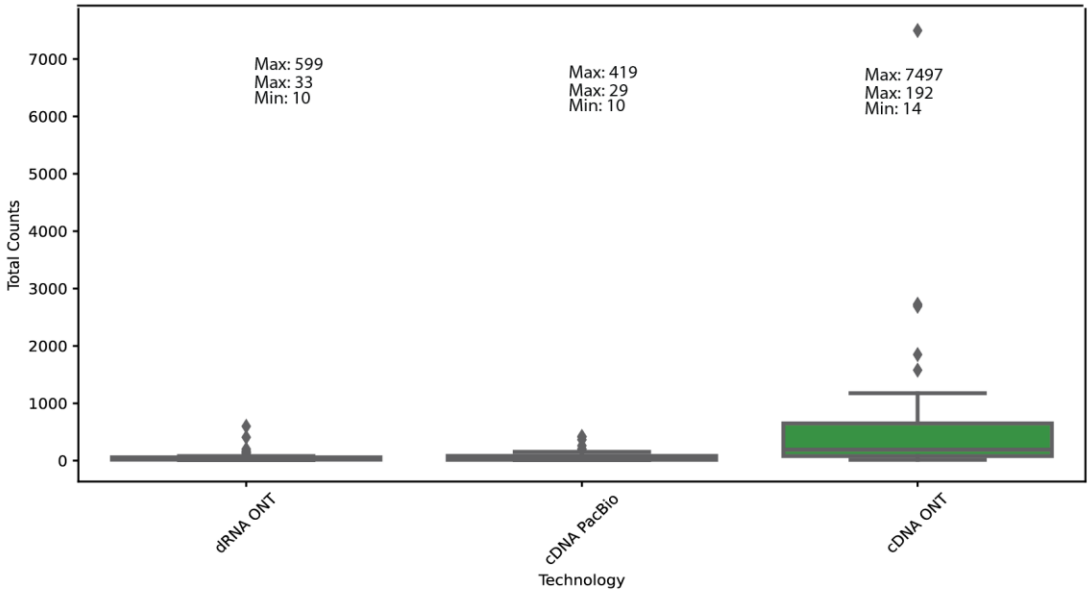


Supplementary Figure 10: Distribution of number of junctions in FSM UJCs detected exclusively in one technology for the WTC11 samples

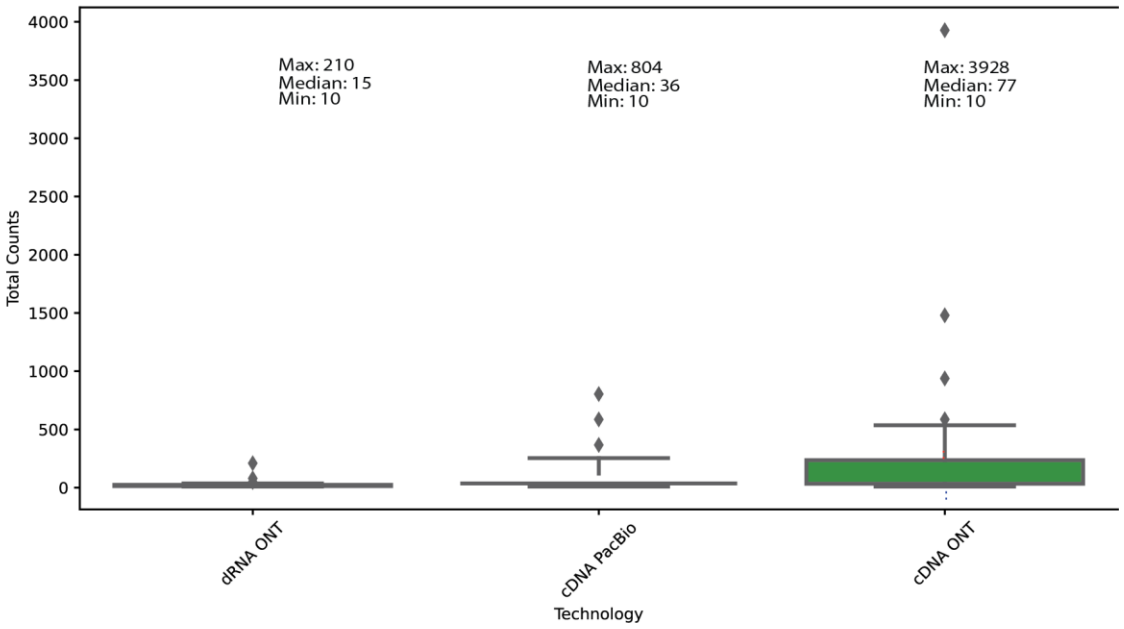


Supplementary Figure 11: For donors (A) and acceptors (B) with cv=0 in replicate 1 of the WTC11 samples, boxplot of the distribution of the distance of the read donors and acceptors to the annotated donors and acceptors in replicate 1.

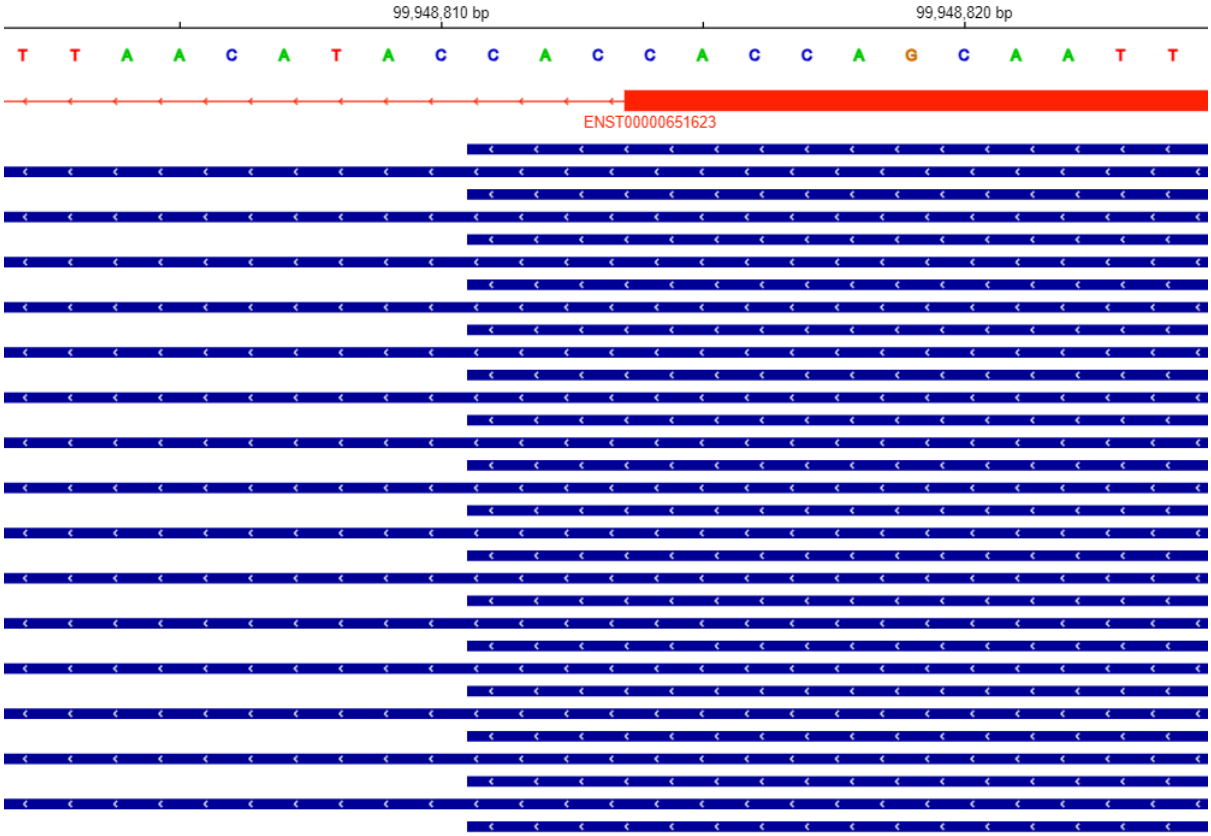
A



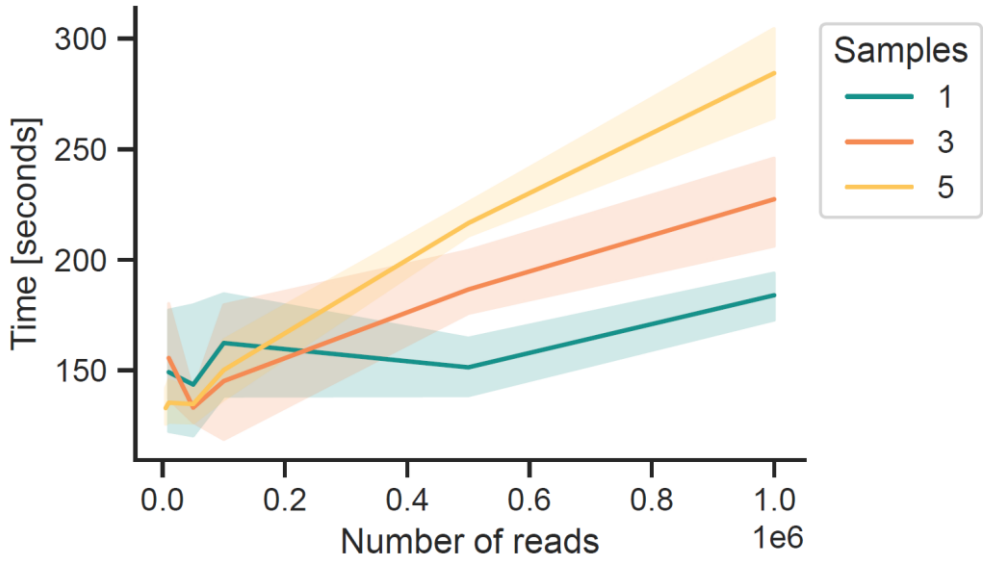
B



Supplementary Figure 12: For donors (A) and acceptors (B) with cv=0 in replicate 1 of the WTC11 samples, boxplot of the number of reads associated with the reference junction with cv=0.



Supplementary Figure 13: IGV visualization of a reference donor with CV = 0 in the H2AZ1 gene at position 99,948,814 on Chromosome 4. All of the reads associated with this donor map to position 99,948,811 which is 3 nts away from the annotated donor position. The annotated transcript is shown in red and the reads are shown in blue.



Supplementary Figure 14: Computational time for running SQANTI-reads with 1 (green), 3 (orange) and 5 (samples)

Supplementary Table 1: Metadata for Drosophila Experiment

Sample	Biological Replicate	Genotype	Stage	Number of Technical Replicates	Minion Runs	Promethion Runs
S1	1	dm11037	01h	2	1	1
S1	2	dm11037	01h	3	2	1
S1	3	dm11037	01h	3	2	1
S2	2	dm11255	01h	3	2	1
S2	2	dm11255	01h	3	2	1
S2	3	dm11255	01h	3	2	1
S3	1	dm12272	01h	2	1	1
S3	2	dm12272	01h	2	1	1
S3	3	dm12272	01h	2	1	1
S4	1	dm12279	01h	2	1	1
S4	2	dm12279	01h	3	2	1
S4	3	dm12279	01h	3	2	1
S5	1	dn11037	38d	3	1	2
S5	2	dn11037	38d	3	1	2
S5	3	dm11037	38d	3	1	2
S6	1	dm11255	38d	3	1	2
S6	2	dm11255	38d	3	1	2
S6	3	dm11255	38d	3	1	2
S7	1	dm12272	38d	3	1	2
S7	2	dm12272	38d	3	1	2
S7	3	dm12272	38d	3	1	2
S8	1	dm12279	38d	3	1	2
S8	2	dm12279	38d	3	1	2
S8	3	dm12279	38d	3	1	2

Supplementary Table 2: Metadata for WTC11 experiment

Sample	Library Prep	Platform	Run Accession	Rep	BioSample	File Accession
WTC11	dRNA	ONT	ENCSR392BGY	1	ENCBS944CBA	ENCFF155CFF
WTC11	dRNA	ONT	ENCSR392BGY	2	ENCBS593PKA	ENCFF771DIX
WTC11	dRNA	ONT	ENCSR392BGY	3	ENCBS474NOC	ENCFF600LIU
WTC11	cDNA	PacBio	ENCSR507JOF	1	ENCBS944CBA	ENCFF563QZR
WTC11	cDNA	PacBio	ENCSR507JOF	2	ENCBS593PKA	ENCFF370NFS
WTC11	cDNA	PacBio	ENCSR507JOF	3	ENCBS474NOC	ENCFF245IPA
WTC11	cDNA	ONT	ENCSR539ZXJ	1	ENCBS944CBA	ENCFF263YFG
WTC11	cDNA	ONT	ENCSR539ZXJ	2	ENCBS593PKA	ENCFF023EXJ
WTC11	cDNA	ONT	ENCSR539ZXJ	3	ENCBS474NOC	ENCFF961HLO

Supplementary Methods

Comparing Putative Novel Transcripts with Annotated Transcripts

SQANTI-reads was run on all the WTC11 samples using the option –factor-level cDNA PacBio which allowed us to evaluate under-annotation on the cDNA PacBio samples only. Using the ujc_counts.csv and putative_novel_transcripts.csv out of SQANTI-reads, we created a list of all putative novel transcripts (conditions in Figure 3A) and the most expressed annotated (FSM) UJC in the same gene. We ran idUJC.py (Nanni et al. 2024) on all the WTC11 cDNA PacBio samples to generate GTF files of the representative UJCs in each WTC11 cDNA PacBio sample. From the UJC GTF files we created one GTF containing only putative novel transcripts and one GTF containing the most expressed annotated (FSM) UJC in the corresponding gene. We then ran TranD (Nanni et al. 2024) in two GTF mode to calculate the nucleotide difference metrics between each putative novel transcript and its corresponding most expressed annotated (FSM) UJC. This process was done separately for genes with a well covered FSM transcript (FSM > 20% of total gene coverage) (Figure 3D) and genes with a low coverage FSM (FSM <= 20% of the total gene coverage) (Figure 3E).

References

Nanni A, Titus-McQuillan J, Bankole KS, Pardo-Palacios F, Signor S, Vlaho S, Moskalenko O, Morse Alison M, Rogers RL, Conesa A et al. 2024. Nucleotide-level distance metrics to quantify alternative splicing implemented in TranD. Nucleic Acids Research 52: e28-e28.