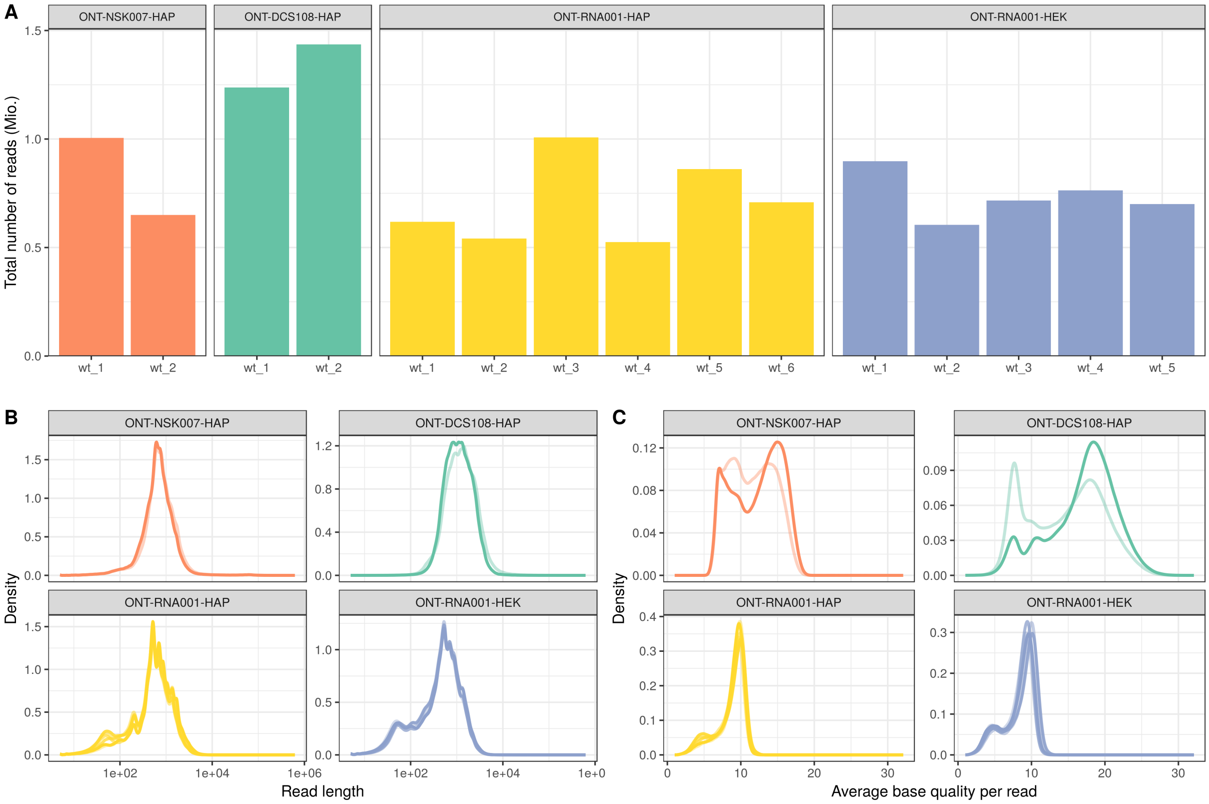
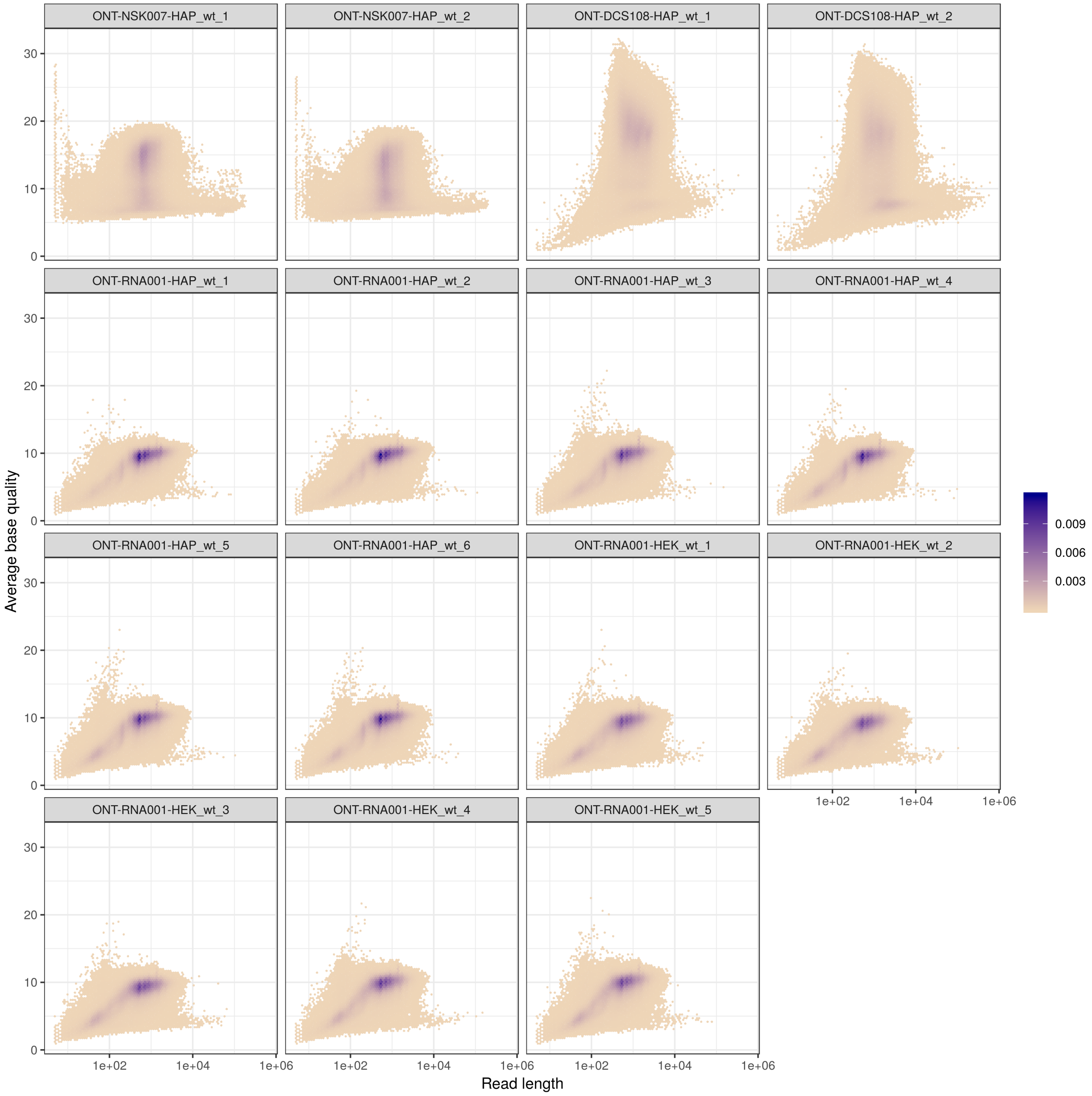
A comprehensive examination of Nanopore native RNA sequencing for characterization of complex transcriptomes

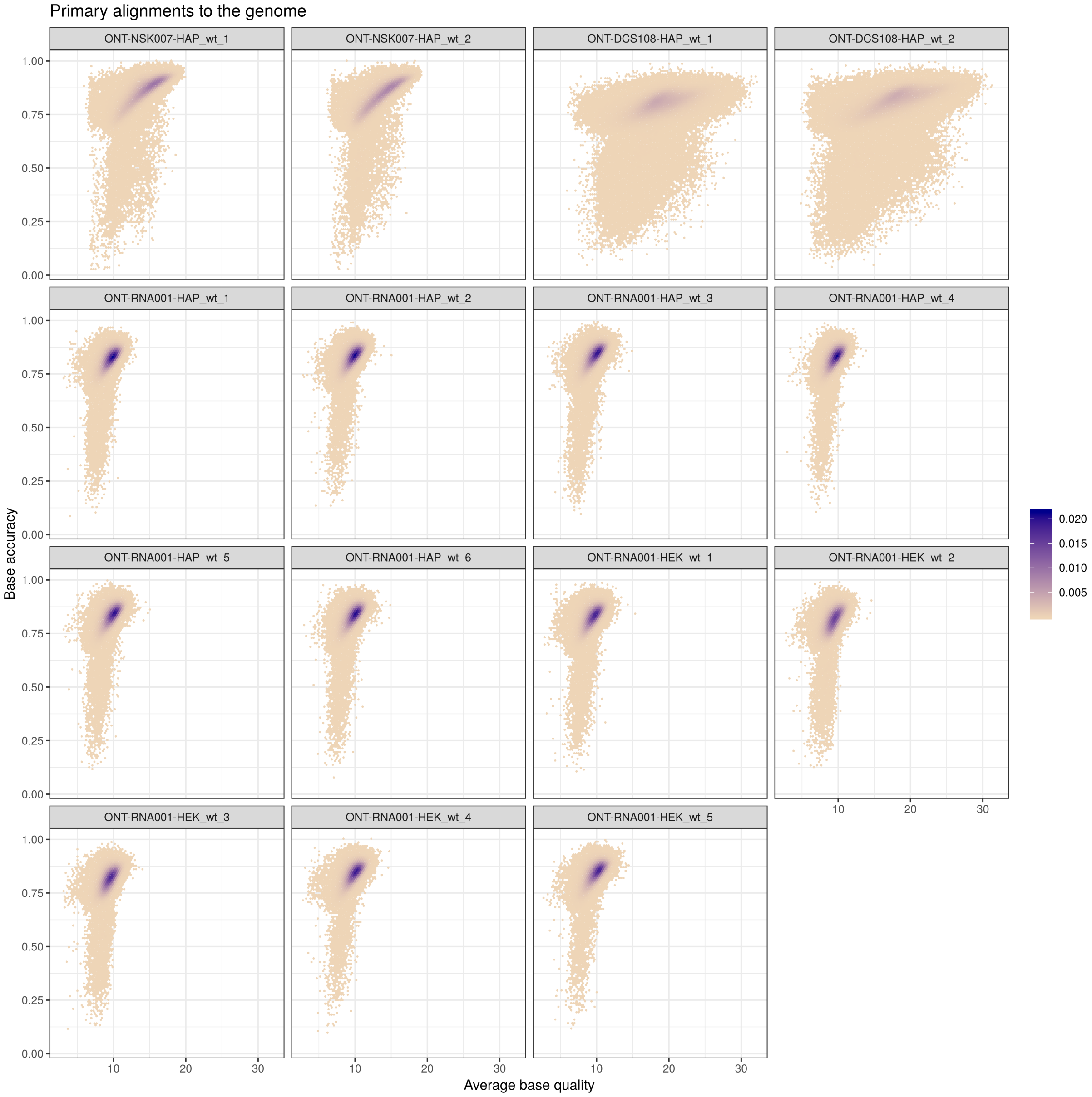
Soneson et al.



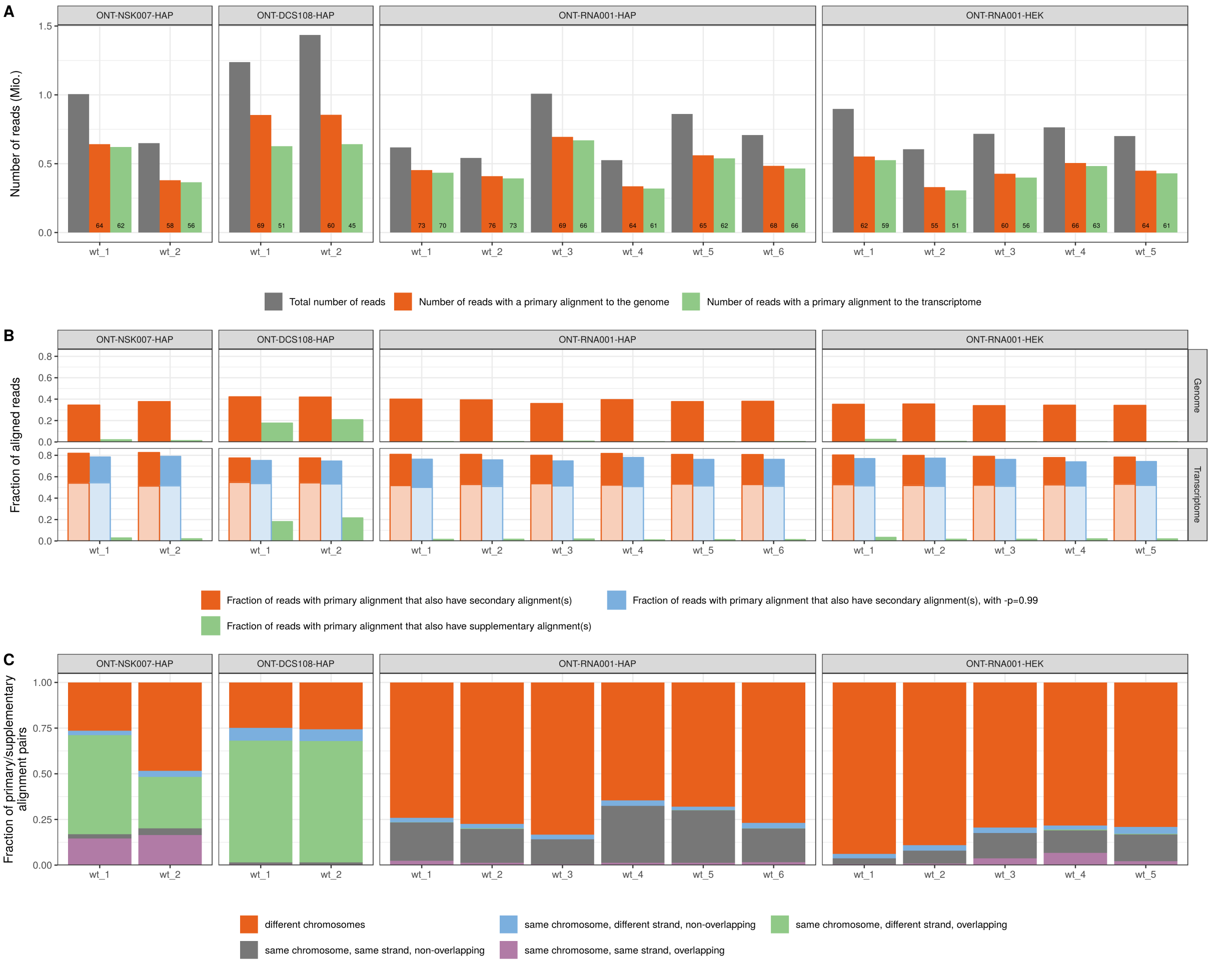
Supplementary Figure 1: Summary of the full set of reads in each ONT library, stratified by data set. A. Total number of reads. B. Read length distribution. C. Average base quality distribution. In B-C, each line corresponds to one library. Source data for panel A are provided as a Source Data file.



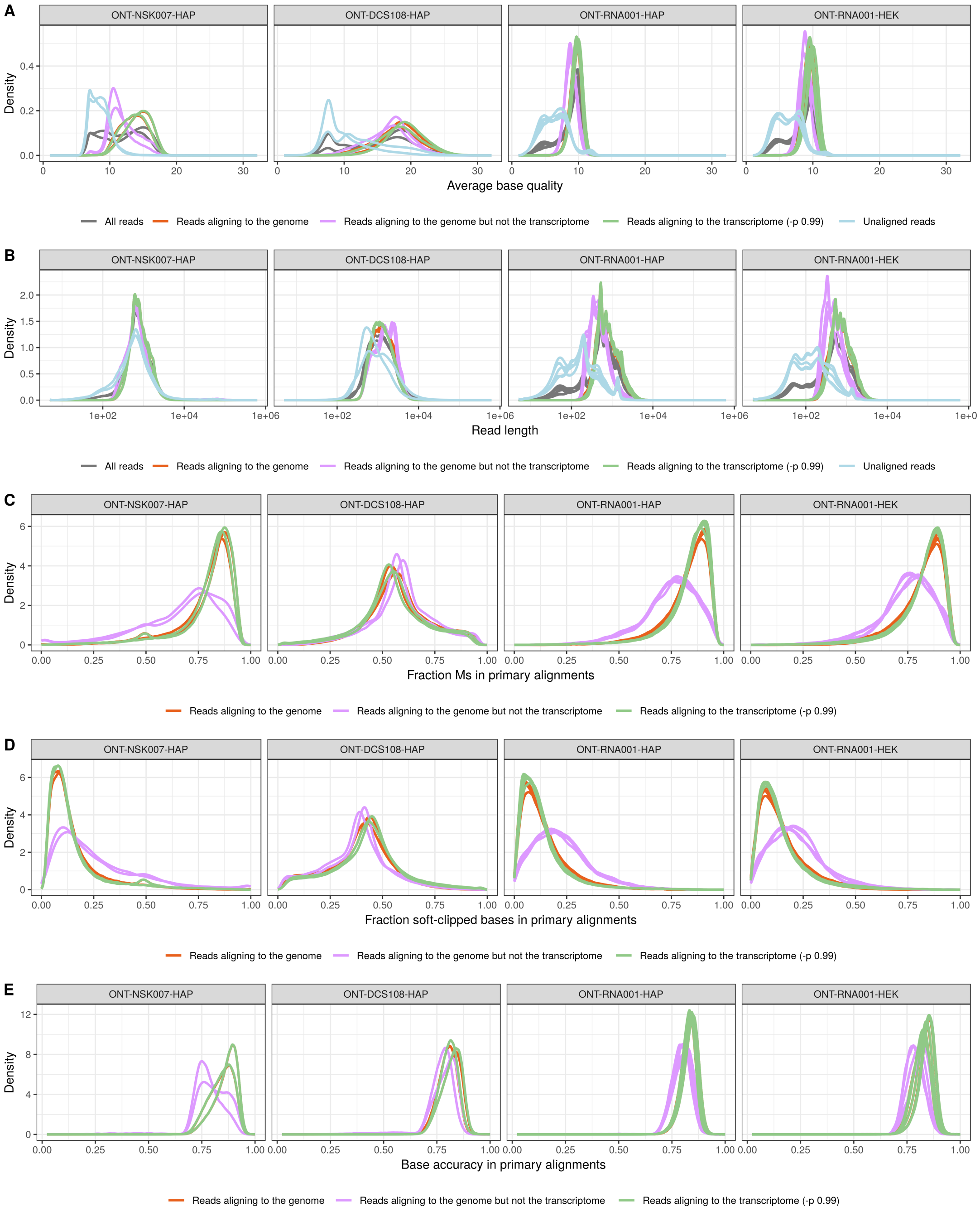
Supplementary Figure 2: Total read length (x) vs average base quality (y, as reported in the basecalled FASTQ file), for all reads in each of the ONT libraries. The colour indicates point density.



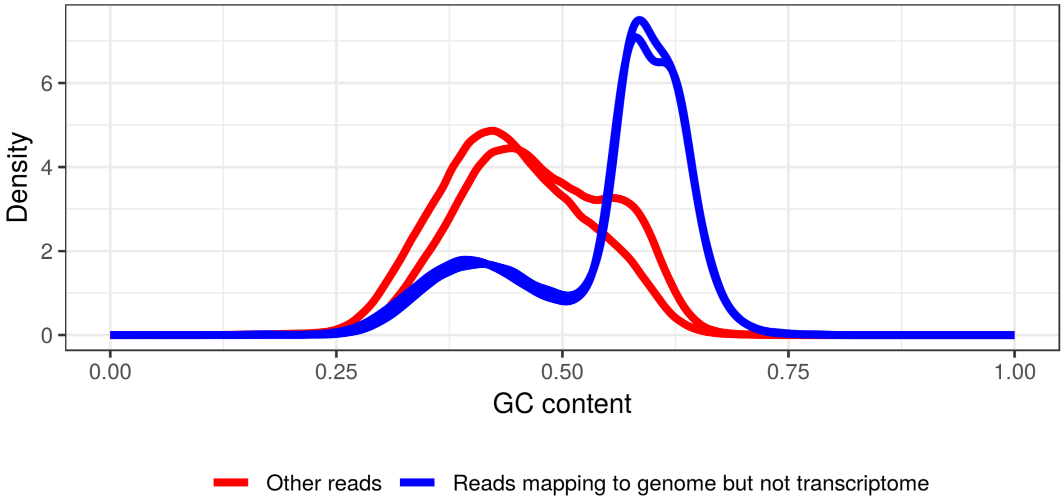
Supplementary Figure 3: Average base quality (x, as reported in the basecalled FASTQ file) and base-level accuracy (y, calculated as $(nbrM + nbrI + nbrD - NM)/(nbrM + nbrI + nbrD)$ from the part of the read involved in the primary genome alignment), for all aligned reads in each of the ONT libraries. Here, $nbrM$, $nbrI$ and $nbrD$ are the reported number of M, I and D characters in the CIGAR string of the alignment, and NM is the reported edit distance. The colour indicates point density. The mean accuracies were 85% (ONT-NSK007-HAP), 80% (ONT-DCS108-HAP), 83% (ONT-RNA001-HAP) and 83% (ONT-RNA001-HEK), respectively.



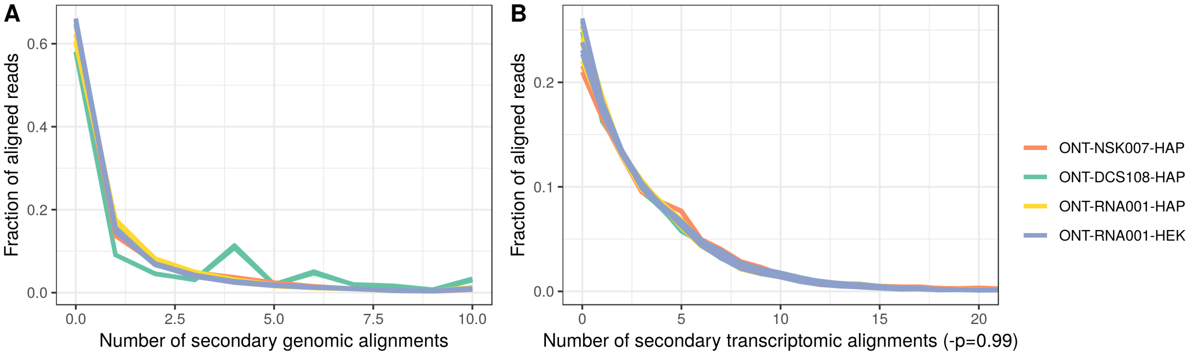
Supplementary Figure 4: Characterization of aligned reads in each of the ONT libraries. A. Total number of reads and the number of reads with a primary alignment to the genome or transcriptome, respectively. The number displayed in each bar represents the alignment rate in % (the fraction of the total number of reads for which minimap2 reports a primary alignment). B. Fraction of the reads with a primary alignment to the genome or transcriptome, respectively, that also have at least one reported secondary or supplementary alignment. The lighter shaded parts of the secondary transcriptome alignment bars correspond to reads where all primary and secondary alignments are to isoforms of the same gene, while the darker shaded parts correspond to reads with reported alignments to transcripts from different genes. C. Investigation of supplementary genome alignments. Each supplementary alignment is categorized based on whether it is on the same chromosome and strand as the primary alignment, and if the alignment positions of the primary and supplementary alignments overlap. Source data are provided as a Source Data file.



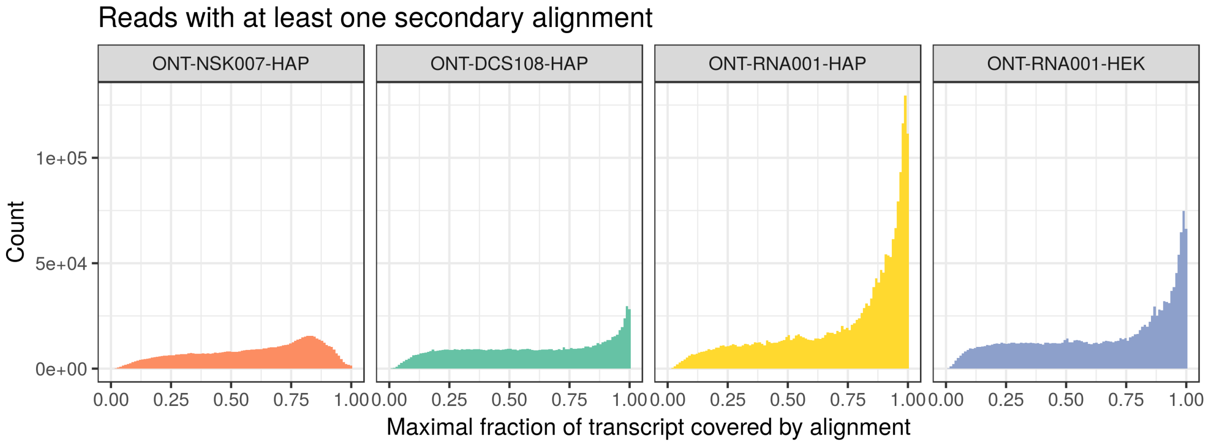
Supplementary Figure 5: Summary of read properties by alignment status. Each density corresponds to one ONT library. A. Average base quality distribution. Unaligned reads are predominantly found among reads with low base quality. B. Read length distribution. C. Number of M characters in the CIGAR string of the primary alignment to the genome and transcriptome, respectively, divided by the read length. D. Number of soft-clipped (S) bases in the CIGAR string of the primary alignment to the genome and transcriptome, respectively, divided by the read length. E. Base accuracy, calculated as $(nbrM + nbrI + nbrD - NM)/(nbrM + nbrI + nbrD)$ from the part of the read involved in the primary alignment.



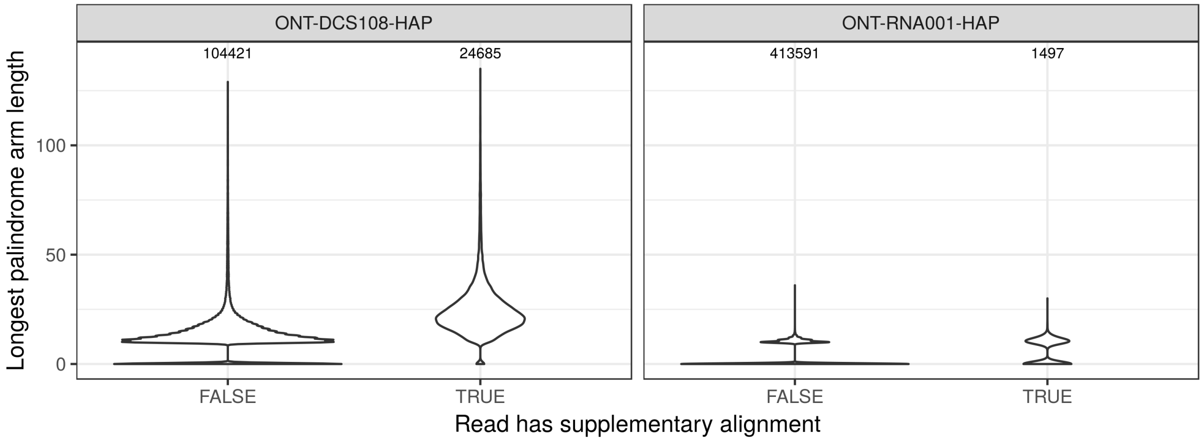
Supplementary Figure 6: GC content distribution for reads mapping to the genome but not the transcriptome, and all other reads, in the two ONT-DCS108-HAP libraries.



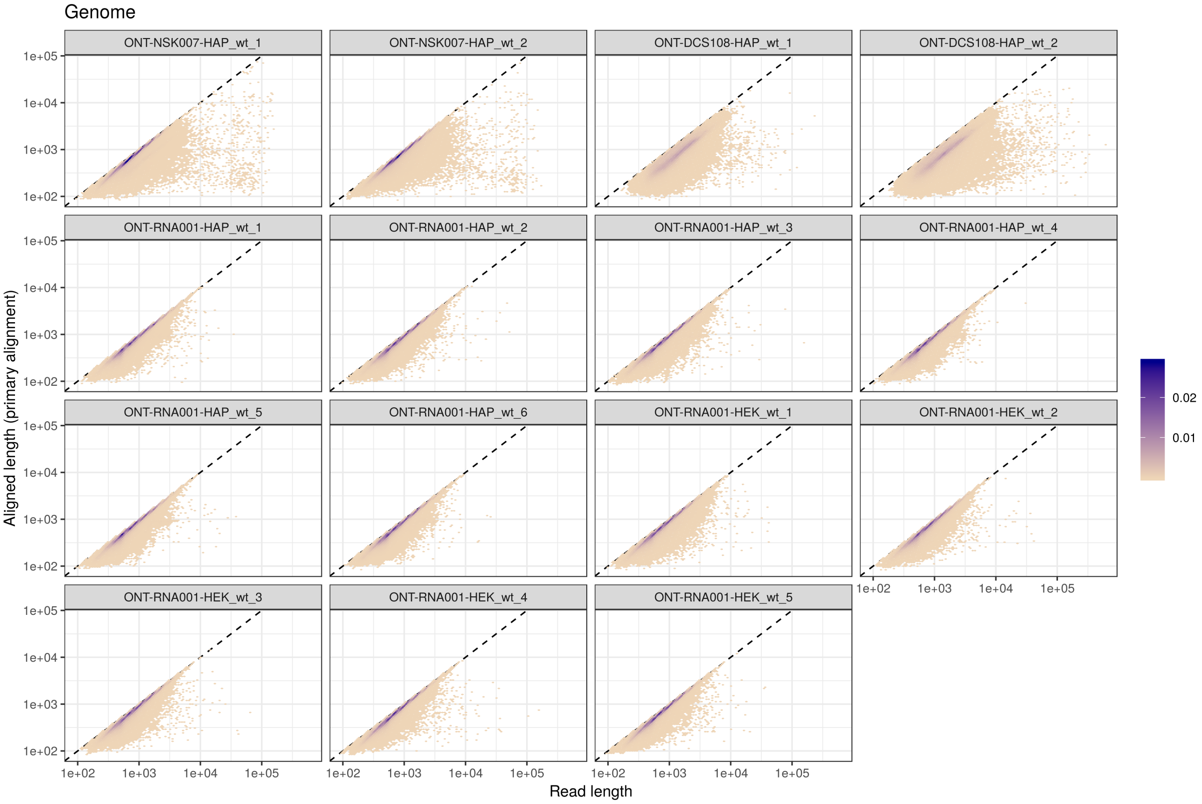
Supplementary Figure 7: Distribution of the number of secondary genome and transcriptome (with the `-p 0.99` setting) minimap2 alignments, for each ONT library. A. For the genome alignments, up to 10 secondary alignments were allowed. B. For the transcriptome alignments, up to 100 secondary alignments were allowed, in order to allow for the high similarity among isoforms. However, very few reads had more than 20 secondary alignments, and thus the *x*-axis is truncated to the range [0-20]. Source data are provided as a Source Data file.



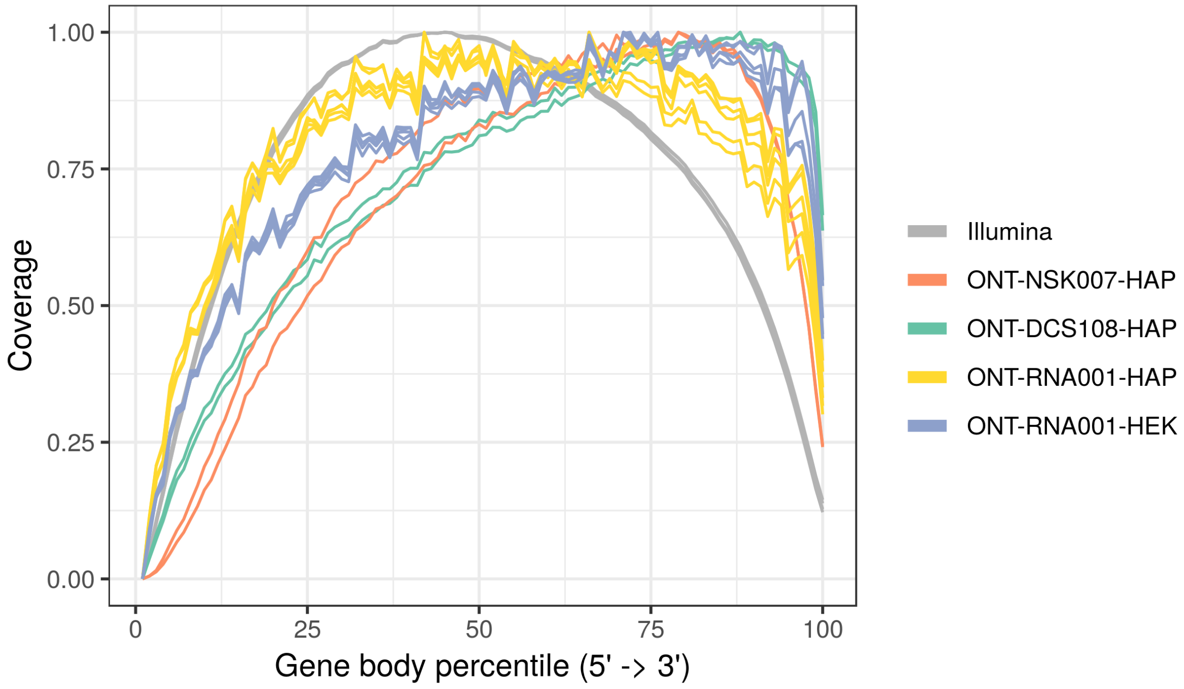
Supplementary Figure 8: The highest degree of coverage of a reference transcript by single ONT reads, for each read with at least one secondary alignment.



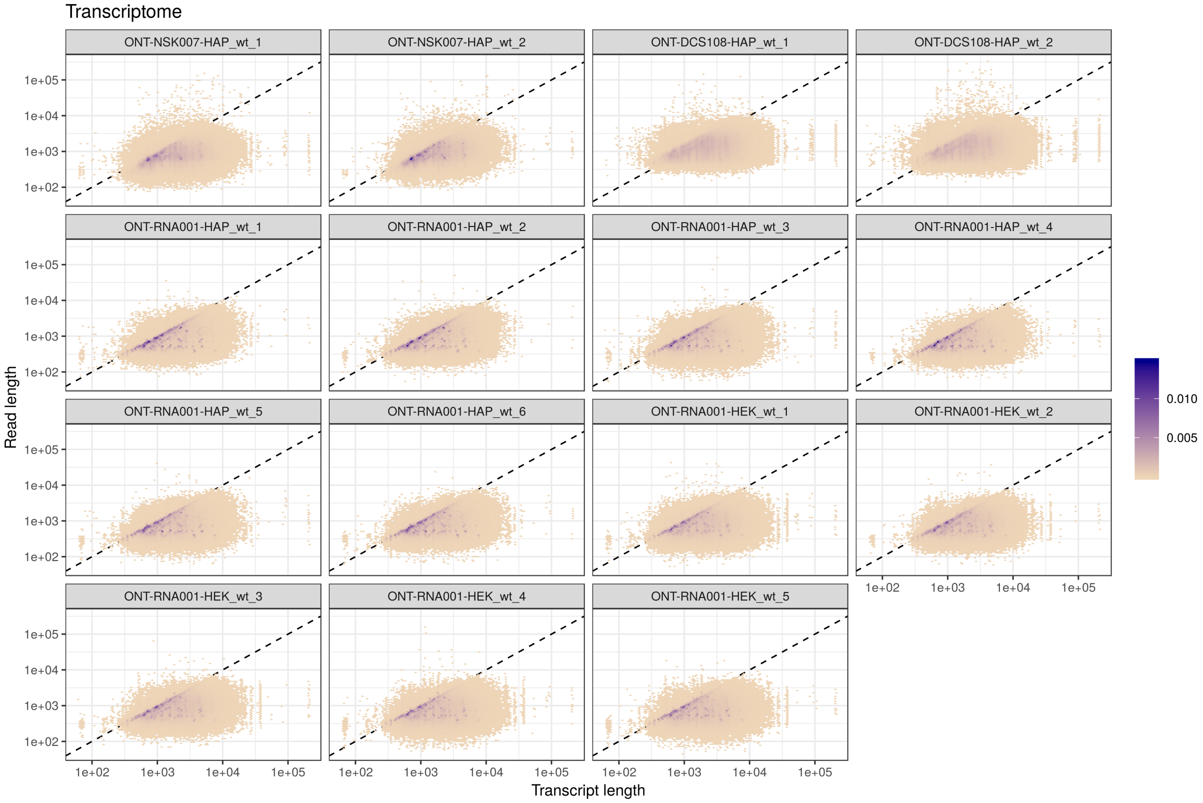
Supplementary Figure 9: Distribution of the length of the longest palindrome found in aligned reads with and without supplementary alignments, in the ONT-DCS108-HAP and ONT-RNA001-HAP libraries, aggregated by data set. For each library, 100,000 reads were randomly selected; only the ones with reported alignments are considered here. Only palindromes with arm length exceeding 10 bases were considered, reads without such palindromes were assigned a maximal palindrome arm length of 0.



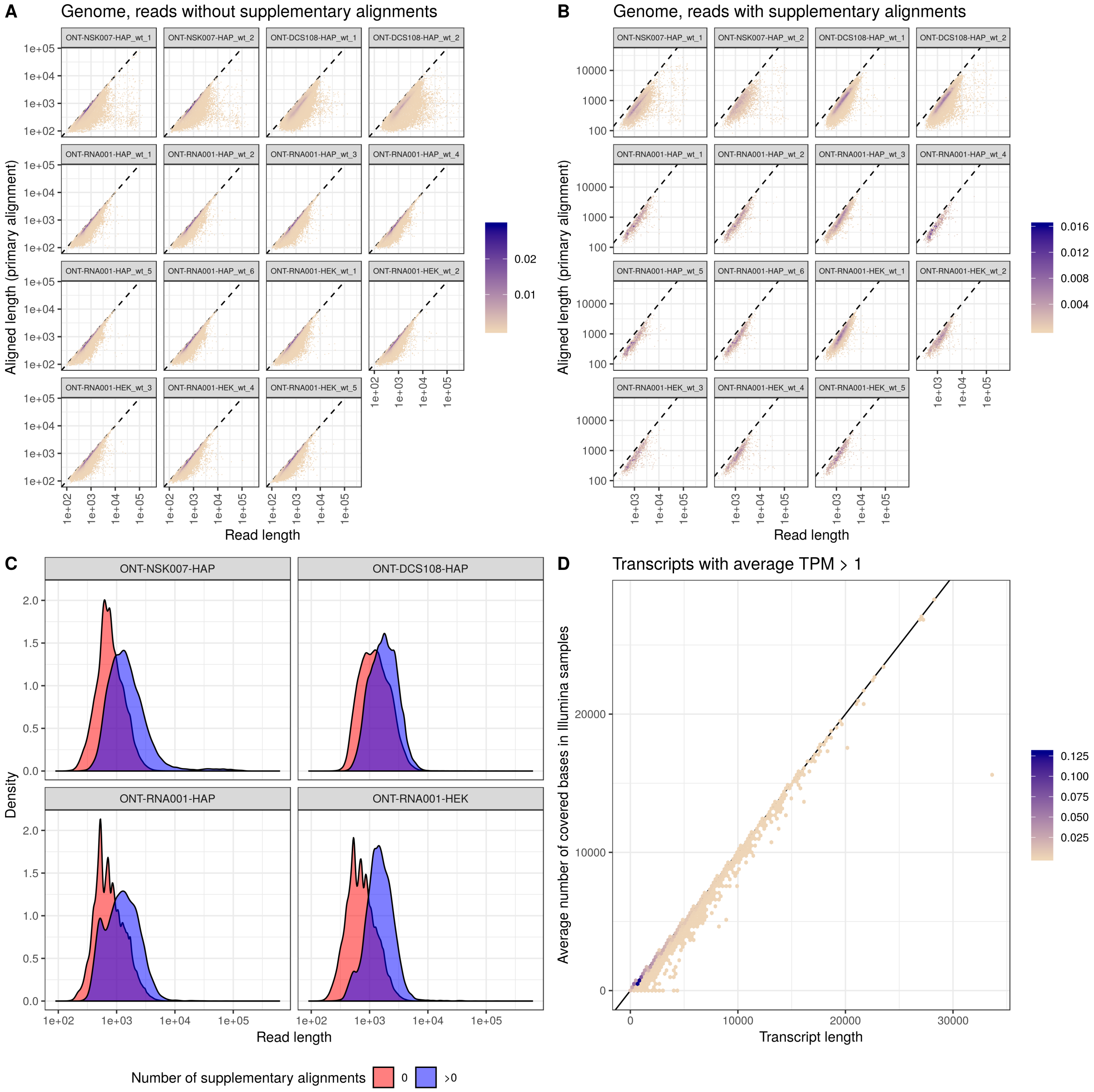
Supplementary Figure 10: Total read length (x) vs aligned length (y, the sum of the number of M and I characters in the CIGAR string) for the primary genome alignment of each read in each of the ONT libraries. The colour indicates point density.



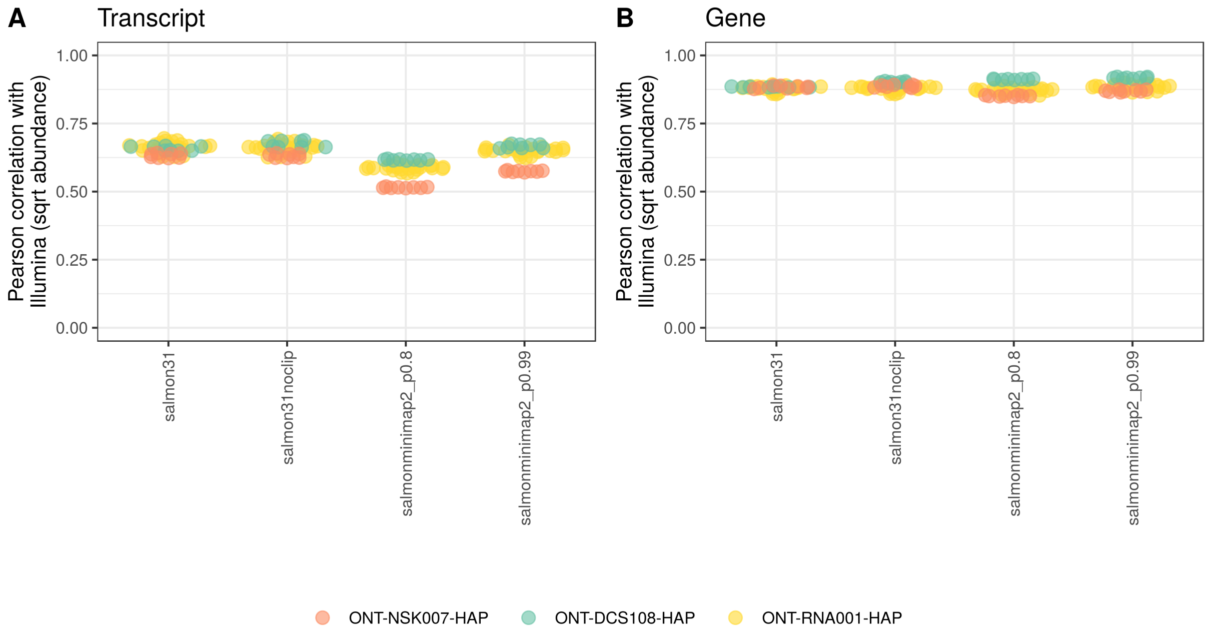
Supplementary Figure 11: Gene body coverage, estimated by RSeQC, in the ONT and Illumina libraries.



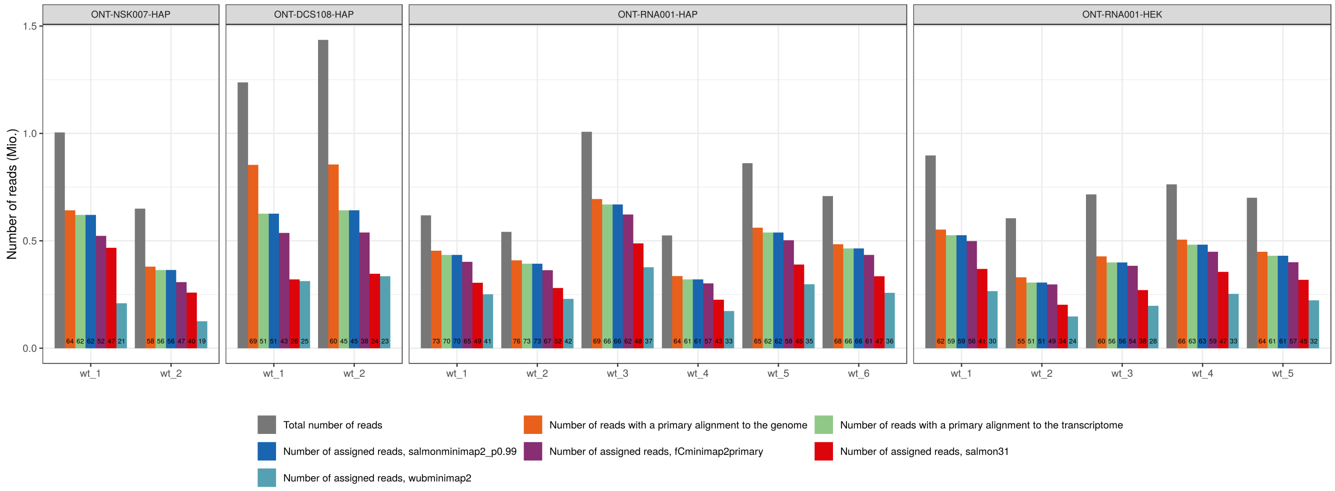
Supplementary Figure 12: Annotated length of the target transcript (x) vs total read length (y) for all primary transcriptome alignments in each of the ONT libraries. Reads aligning to the shortest transcripts are often longer than the transcript, and are thus soft clipped in the alignment. The colour indicates point density.



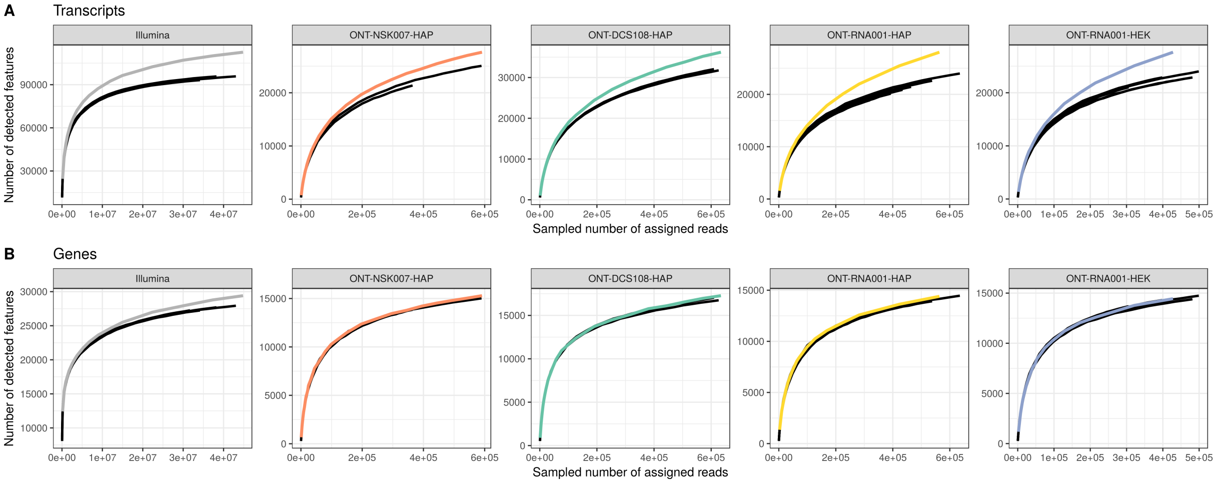
Supplementary Figure 13: A-B. Total read length (x) vs length of the primary genome alignment (y) for reads without (A) and with (B) any reported supplementary alignment. The colour indicates point density. C. Read length distribution for reads without (red) and with (blue) any reported supplementary alignment, in the four ONT data sets. D. The average number of nucleotides annotated to the transcript that are covered by reads in the Illumina samples *vs* the transcript length, for transcripts with an average estimated TPM exceeding 1 across the Illumina samples.



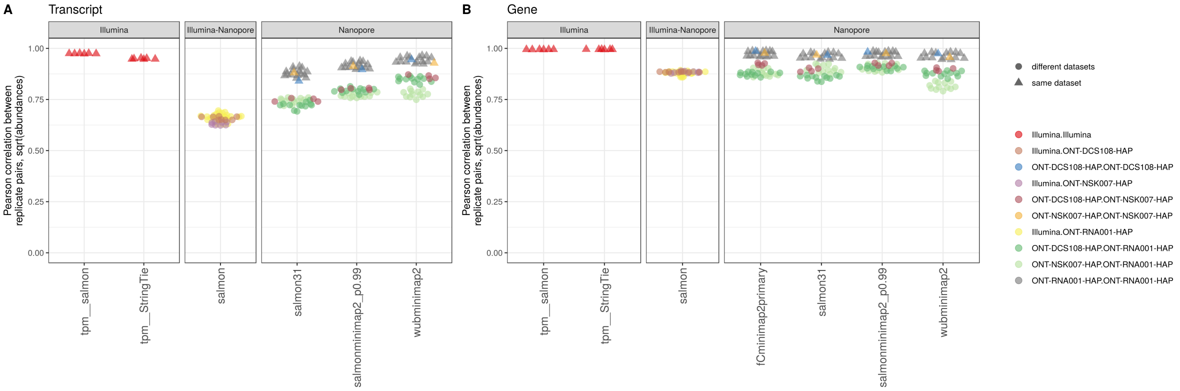
Supplementary Figure 14: Correlation between estimated abundances in the Illumina samples and those obtained from the ONT data, running Salmon in different configurations. ONT abundances are estimated counts, and Illumina abundances are estimated TPMs from Salmon. Each point corresponds to a pair of one Illumina sample and one ONT sample, and all such pairwise combinations were considered. Source data are provided as a Source Data file.



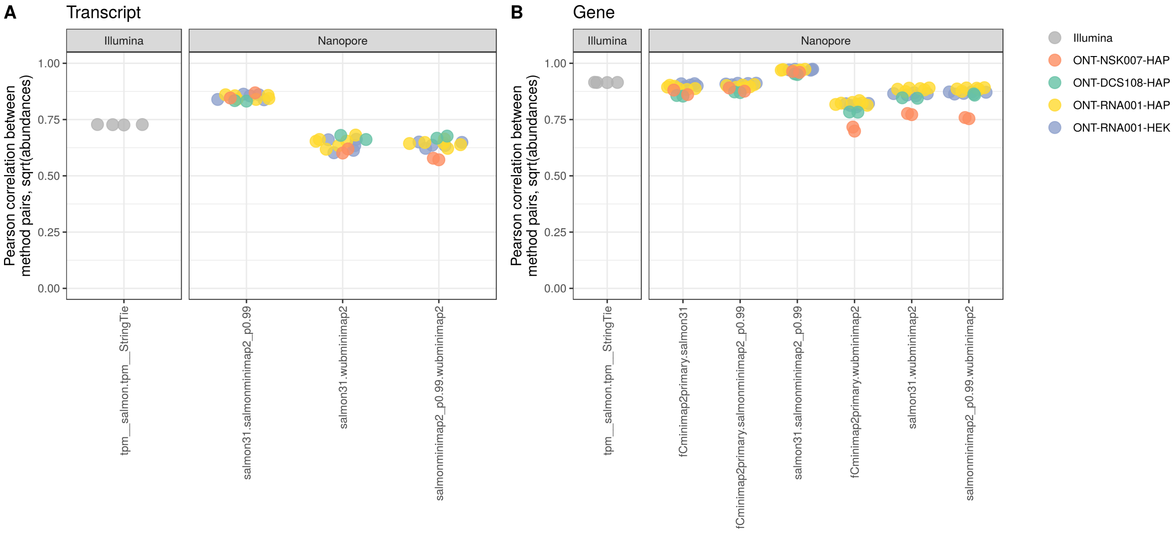
Supplementary Figure 15: Number of reads assigned to features (genes or transcripts) by each of the abundance estimation methods, for each ONT library. The number written in each bar indicates the percentage of the total number of reads that are assigned to features by the corresponding method. Source data are provided as a Source Data file.



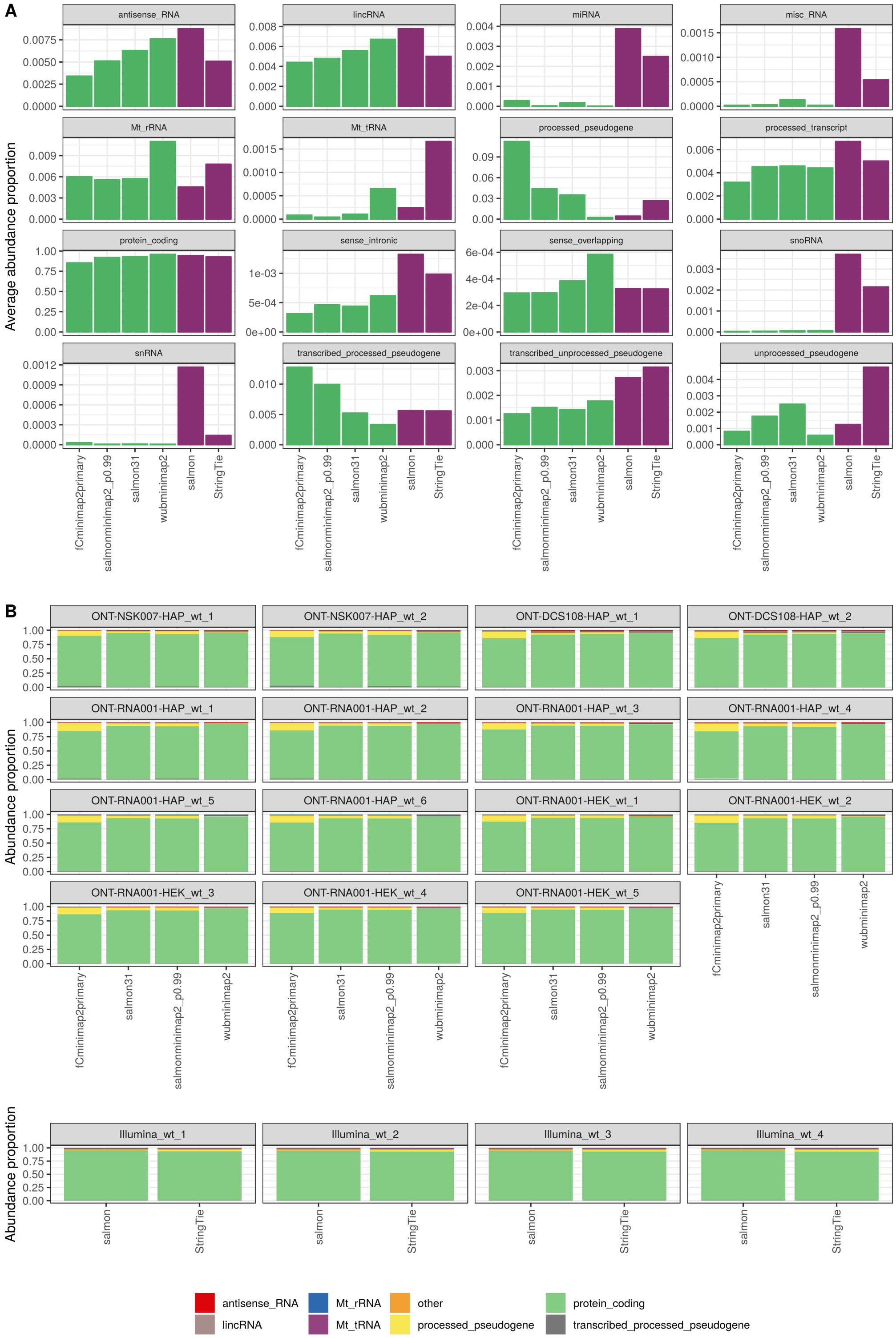
Supplementary Figure 16: Saturation curves for transcript (A) and gene (B) detection. Black curves represent individual libraries, and colored curves are obtained by pooling the reads across all samples in the data set before subsampling. These curves are truncated to the same range as the individual sample curves to facilitate comparison. A feature is considered detected if it has an estimated salmonminimap2 count (ONT libraries) or Salmon count (Illumina libraries) ≥ 1.



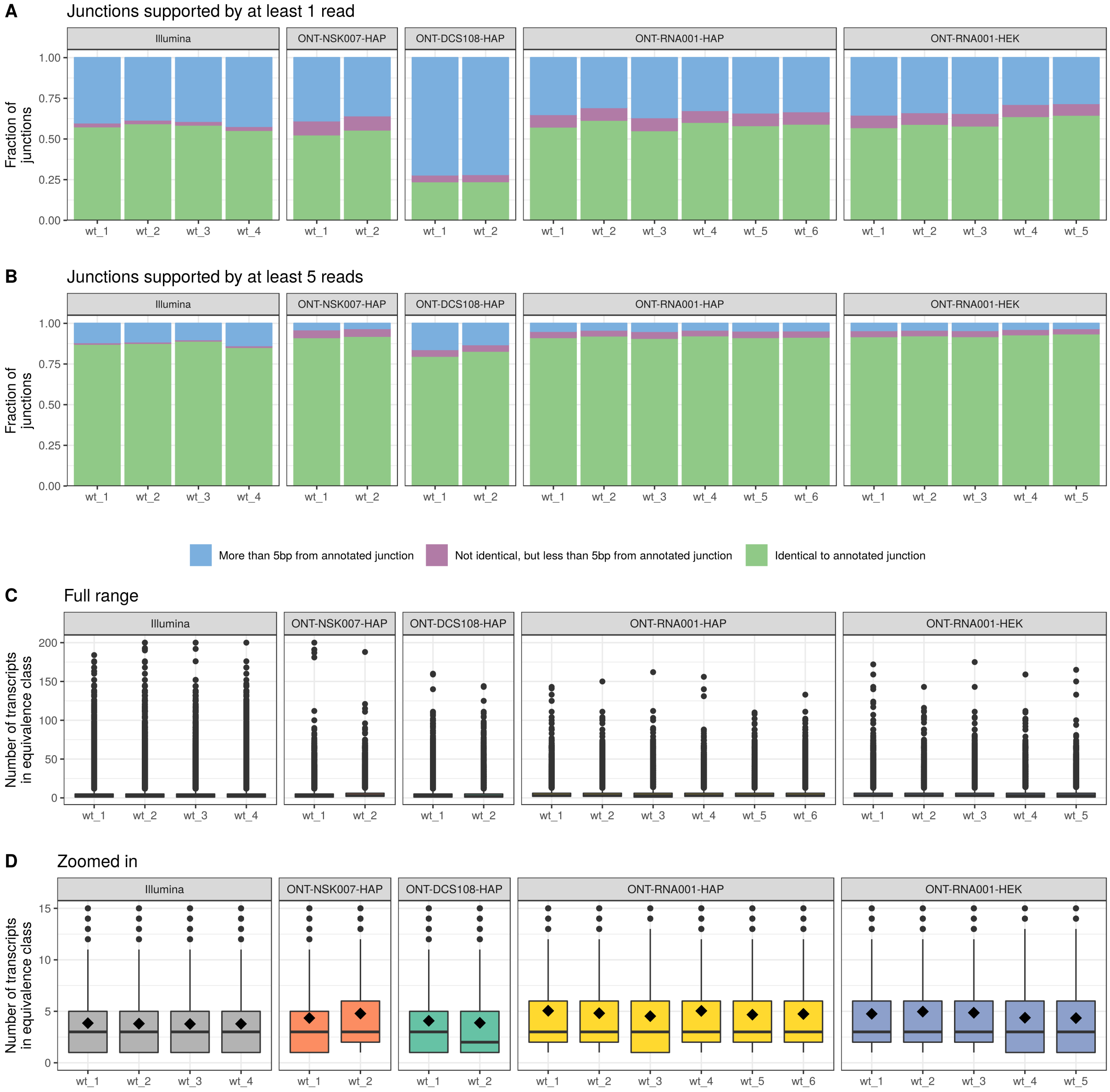
Supplementary Figure 17: Pearson correlations between transcript (A) or gene (B) abundances for each pair of samples, within and between data sets, for each abundance estimation method. Triangles correspond to pairs of samples within the same data set, and circles to pairs from different data sets. Correlations were calculated between square root-transformed estimated counts from the respective ONT methods, and for square root-transformed estimated TPMs for the Illumina libraries. Source data are provided as a Source Data file.



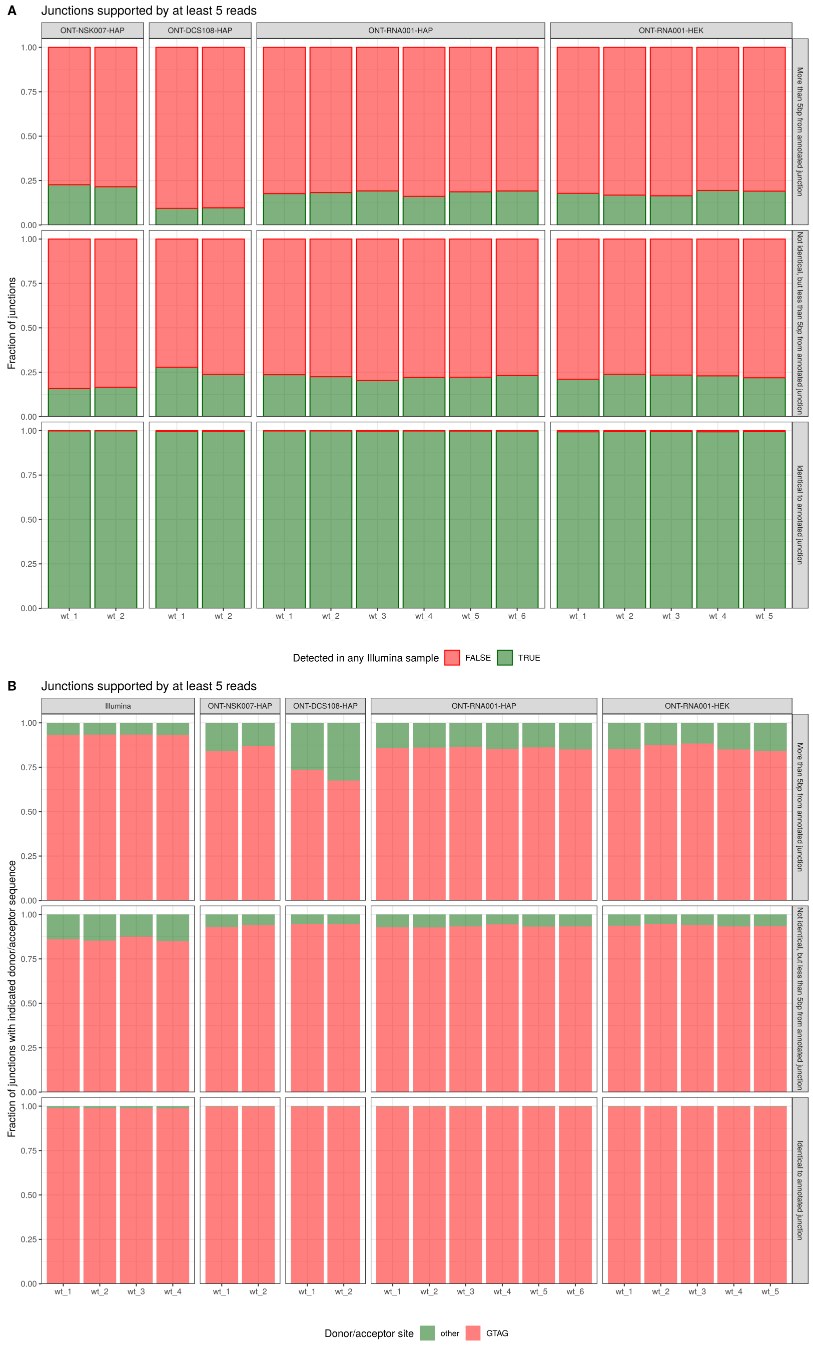
Supplementary Figure 18: Pearson correlations between transcript (A) or gene (B) abundances for each pair of abundance estimation methods, for each library. Correlations were calculated between square root-transformed estimated counts from the respective ONT methods, and for square root-transformed estimated TPMs for the Illumina libraries. Source data are provided as a Source Data file.



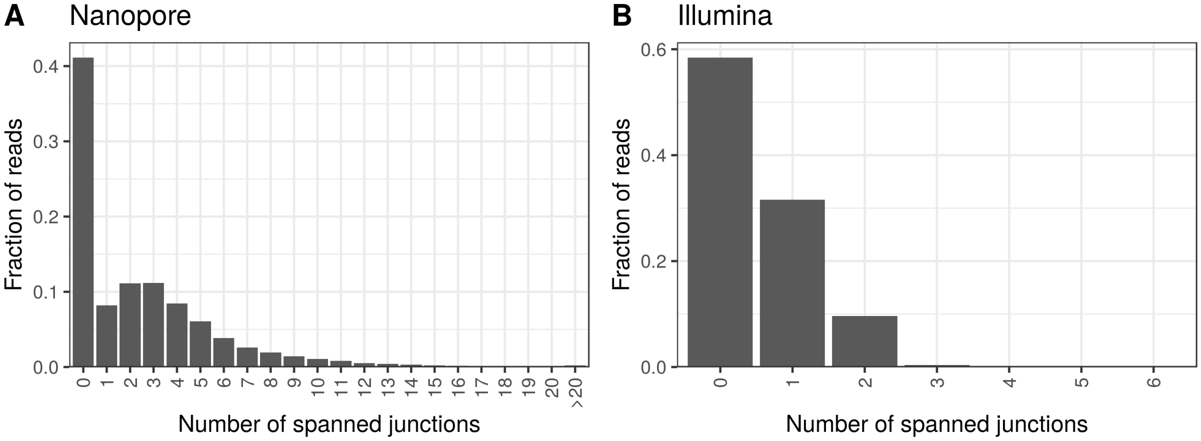
Supplementary Figure 19: A. Relative abundance (proportion of total abundance) assigned to genes of different biotypes by the different quantification methods, averaged across all libraries. Green bars represent abundances (counts) for ONT data, while purple bars represent abundances (TPMs) for the Illumina libraries. B. Relative abundance (proportion of total count) assigned to genes of different biotypes by the different quantification methods, for all libraries. Only the most abundant biotypes are represented separately, all others are collapsed into the ’other’ category. Source data are provided as a Source Data file.



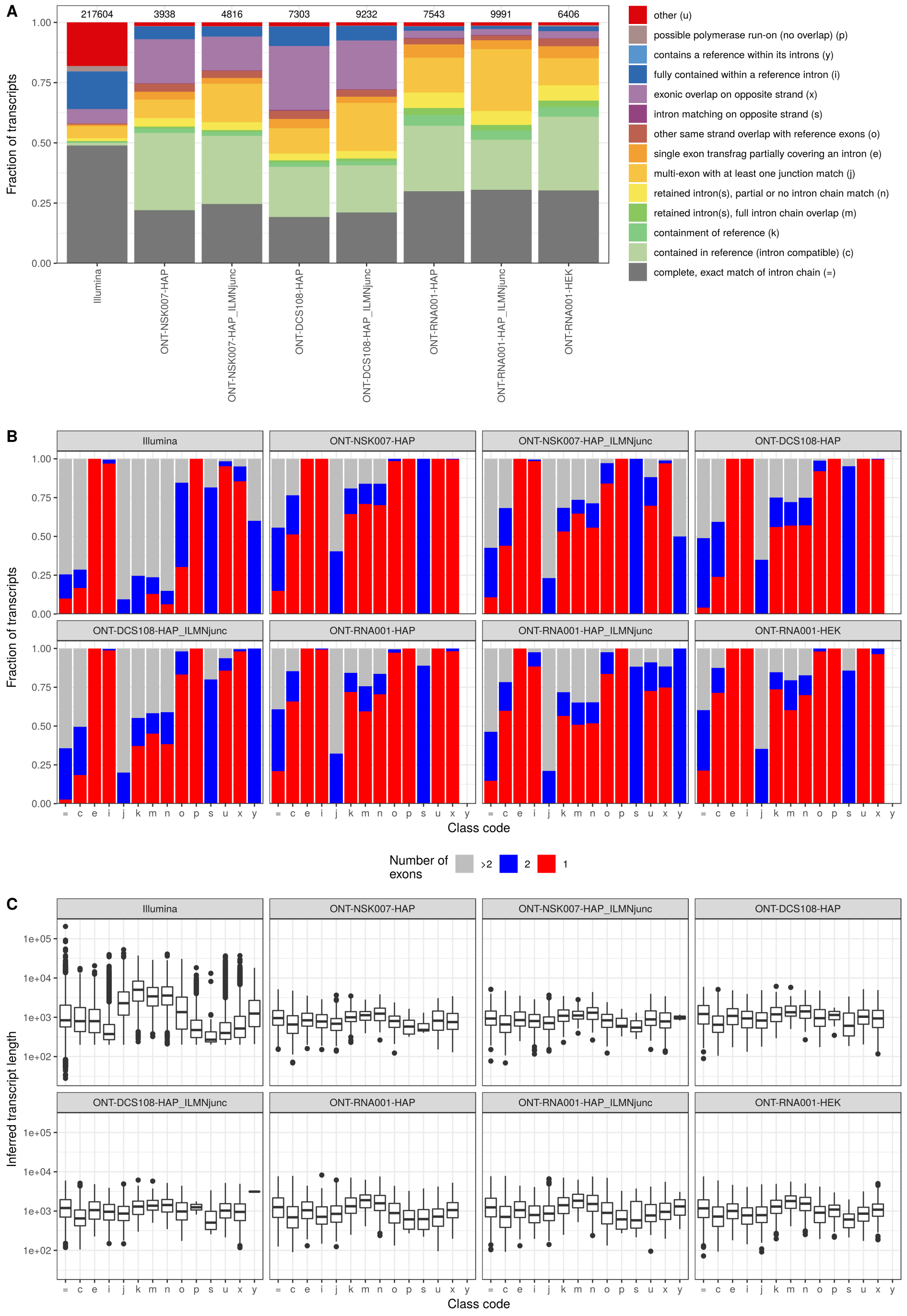
Supplementary Figure 20: A-B. Annotation status of junctions observed in each ONT and Illumina library. A junction is considered observed if it is supported by at least 1 (A) or 5 (B) reads. For each observed junction, the distance to each annotated junction was defined as the absolute difference between the start positions plus the absolute difference between the end positions. This distance was used to find the closest annotated junction. C. Distribution of the number of transcripts contained in the Salmon equivalence class that a read is assigned to, across all reads, for each ONT and Illumina library. D. As C, but zoomed in to the range [0, 15]. The black diamond shape indicates the mean. The center line represents the median; hinges represent first and third quartiles; whiskers the most extreme values within 1.5 interquartile range from the box. Source data are provided as a Source Data file.



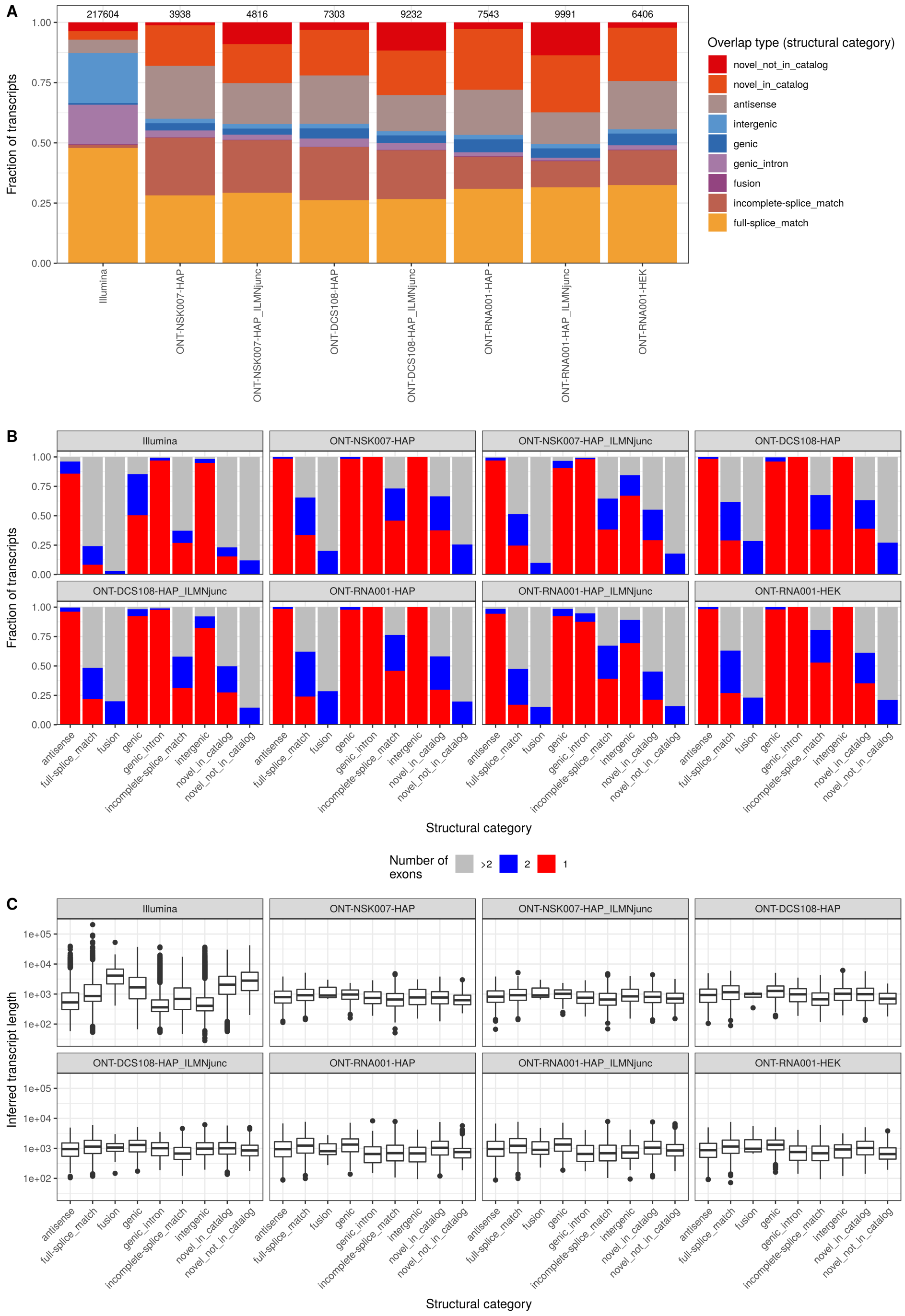
Supplementary Figure 21: A. Fraction of the junctions covered by at least 5 ONT reads that are also detected in at least one of the Illumina samples (supported by at least 1 read), for each category. B. Fraction of the junctions that contain a canonical (GT-AG) or non-canonical splicing motif, respectively, for each category. Source data are provided as a Source Data file.



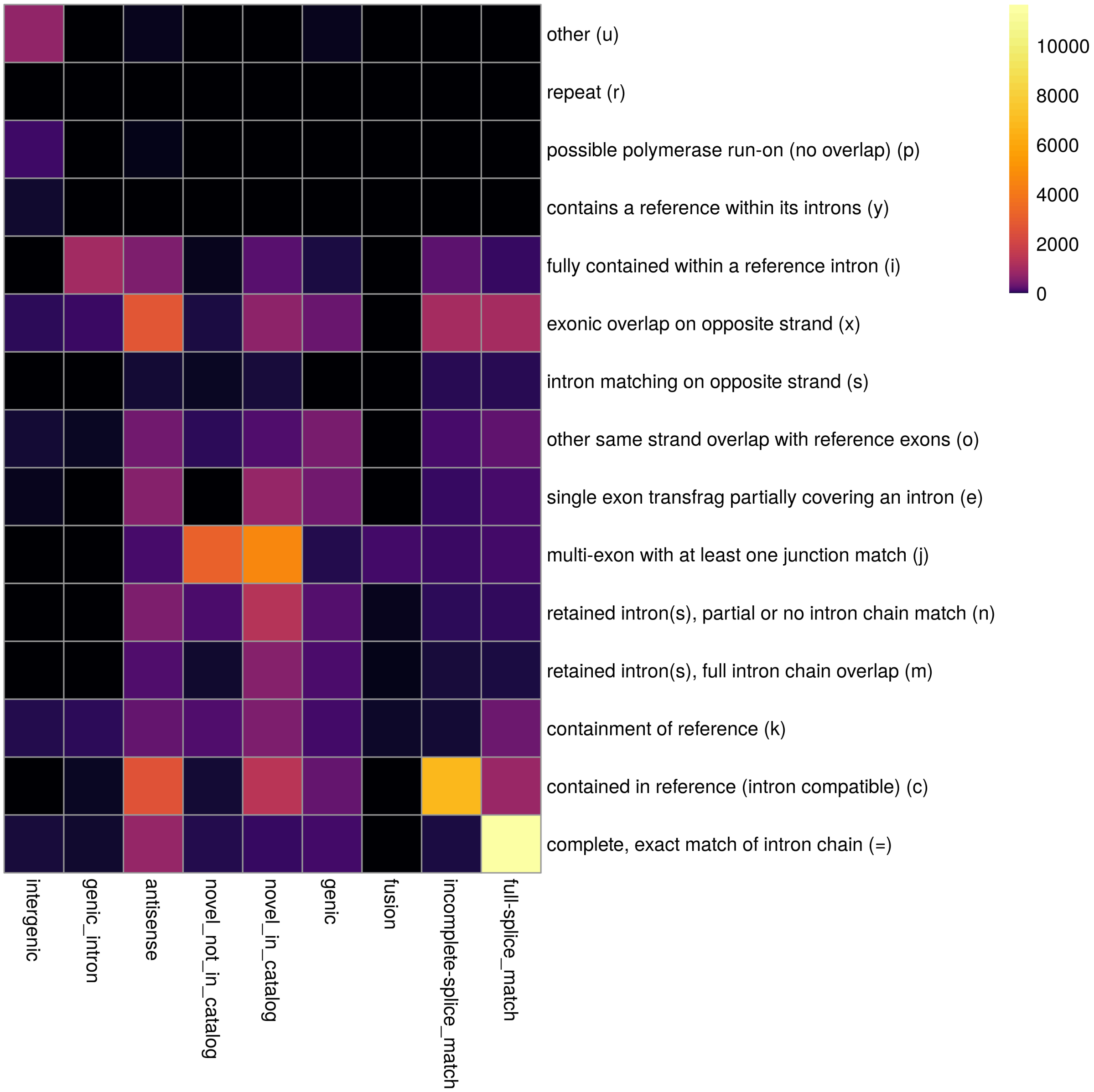
Supplementary Figure 22: Distribution of the number of junctions spanned by individual reads in the ONT (A) and Illumina (B) libraries. For each Illumina library, the number of spanned junctions were counted for a subset of 5 million reads, and only reads being the first in a properly mapped pair were contributing to the summary shown in this plot. Source data are provided as a Source Data file.



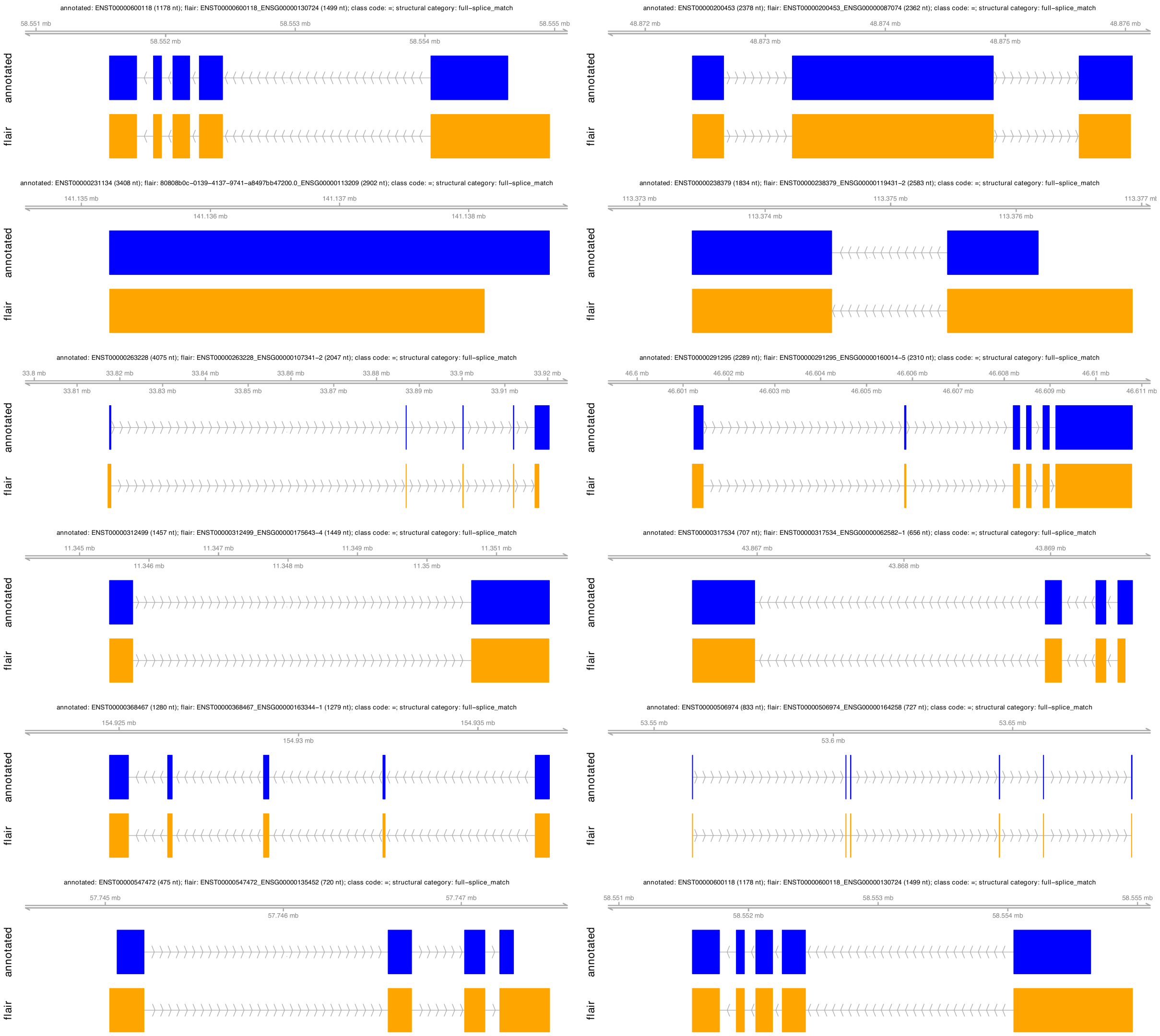
Supplementary Figure 23: Characterization of transcripts identified by FLAIR. A. Class code distribution for de novo identified transcripts from FLAIR (for ONT libraries) or StringTie (for Illumina libraries), compared to the set of annotated transcripts using gffcompare. The number above each bar represents the number of assembled transcripts. The class code of a transcript indicates its relation to the closest annotated transcript. B. Number of exons in each transcript identified by FLAIR/StringTie, stratified by the relation to the annotated transcripts (represented by the assigned class code). C. Length distribution of transcripts identified by FLAIR/StringTie, stratified by the relation to the annotated transcripts (represented by the assigned class code). The *_ILMNjunc* suffix indicates that FLAIR was supplied with the junctions observed in the Illumina samples. The center line represents the median; hinges represent first and third quartiles; whiskers the most extreme values within 1.5 interquartile range from the box. Source data for panels A-B are provided as a Source Data file.



Supplementary Figure 24: Characterization of transcripts identified by FLAIR. A. Structural category distribution for de novo identified transcripts from FLAIR (for ONT libraries) or StringTie (for Illumina libraries), compared to the set of annotated transcripts using SQANTI. The number above each bar represents the number of assembled transcripts. The structural category of a transcript indicates its relation to the closest annotated transcript. B. Number of exons in each transcript identified by FLAIR/StringTie, stratified by the relation to the annotated transcripts (represented by the assigned structural category). C. Length distribution of transcripts identified by FLAIR/StringTie, stratified by the relation to the annotated transcripts (represented by the assigned structural category). The *_ILMNjunc* suffix indicates that FLAIR was supplied with the junctions observed in the Illumina samples. The center line represents the median; hinges represent first and third quartiles; whiskers the most extreme values within 1.5 interquartile range from the box. Source data for panel B are provided as a Source Data file.



Supplementary Figure 25: Comparison of classification of FLAIR transcripts obtained by SQANTI structural categories (columns) and gffcompare class codes (rows), across all ONT data sets. The color of a square corresponds to the number of FLAIR transcripts annotated to each combination of structural category/class code. Source data are provided as a Source Data file.


annotated: ENST00000600118 (1178 nt); flair: ENST00000600118_ENSG00000130724 (1499 nt); class code: =; structural category: full−splice_match
58.551 mb
58.552 mb
58.553 mb
58.554 mb
58.555 mb
annotated
flair
annotated: ENST00000200453 (2378 nt); flair: ENST00000200453_ENSG00000087074 (2362 nt); class code: =; structural category: full−splice_match
48.872 mb
48.873 mb
48.874 mb
48.875 mb
48.876 mb
annotated
flair
annotated: ENST00000231134 (3408 nt); flair: 80808b0c−0139−4137−9741−a8497bb47200.0_ENSG00000113209 (2902 nt); class code: =; structural category: full−splice_match
141.135 mb
141.136 mb
141.137 mb
141.138 mb
annotated
flair
annotated: ENST00000238379 (1834 nt); flair: ENST00000238379_ENSG00000119431−2 (2583 nt); class code: =; structural category: full−splice_match
113.373 mb
113.374 mb
113.375 mb
113.376 mb
113.377 mb
annotated
flair
annotated: ENST00000263228 (4075 nt); flair: ENST00000263228_ENSG00000107341−2 (2047 nt); class code: =; structural category: full−splice_match
33.8 mb
33.81 mb
33.82 mb
33.83 mb
33.84 mb
33.85 mb
33.86 mb
33.87 mb
33.88 mb
33.89 mb
33.9 mb
33.91 mb
33.92 mb
annotated
flair
annotated: ENST00000291295 (2289 nt); flair: ENST00000291295_ENSG00000160014−5 (2310 nt); class code: =; structural category: full−splice_match
46.6 mb
46.601 mb
46.602 mb
46.603 mb
46.604 mb
46.605 mb
46.606 mb
46.607 mb
46.608 mb
46.609 mb
46.61 mb
46.611 mb
annotated
flair
annotated: ENST00000312499 (1457 nt); flair: ENST00000312499_ENSG00000175643−4 (1449 nt); class code: =; structural category: full−splice_match
11.345 mb
11.346 mb
11.347 mb
11.348 mb
11.349 mb
11.35 mb
11.351 mb
annotated
flair
annotated: ENST00000317534 (707 nt); flair: ENST00000317534_ENSG00000062582−1 (656 nt); class code: =; structural category: full−splice_match
43.867 mb
43.868 mb
43.869 mb
annotated
flair
annotated: ENST00000368467 (1280 nt); flair: ENST00000368467_ENSG00000163344−1 (1279 nt); class code: =; structural category: full−splice_match
154.925 mb
154.93 mb
154.935 mb
annotated
flair
annotated: ENST00000506974 (833 nt); flair: ENST00000506974_ENSG00000164258 (727 nt); class code: =; structural category: full−splice_match
53.55 mb
53.6 mb
53.65 mb
annotated
flair
annotated: ENST00000547472 (475 nt); flair: ENST00000547472_ENSG00000135452 (720 nt); class code: =; structural category: full−splice_match
57.745 mb
57.746 mb
57.747 mb
annotated
flair
annotated: ENST00000600118 (1178 nt); flair: ENST00000600118_ENSG00000130724 (1499 nt); class code: =; structural category: full−splice_match
58.551 mb
58.552 mb
58.553 mb
58.554 mb
58.555 mb
annotated
flair

Supplementary Figure 26: Examples of transcripts identified by FLAIR in the ONT-RNA001-HAP data set, with a complete, exact match to the intron chain of an annotated transcript (class code '=' from gffcompare, structural category 'full-splice_match' from SQANTI).