Direct full-length RNA sequencing reveals unexpected transcriptome complexity during *C. elegans* development

Runsheng Li[1#], Xiaoliang Ren[1#], Qiutao Ding[1], Yu Bi[1], Dongying Xie[1], Zhongying Zhao[1,2*]

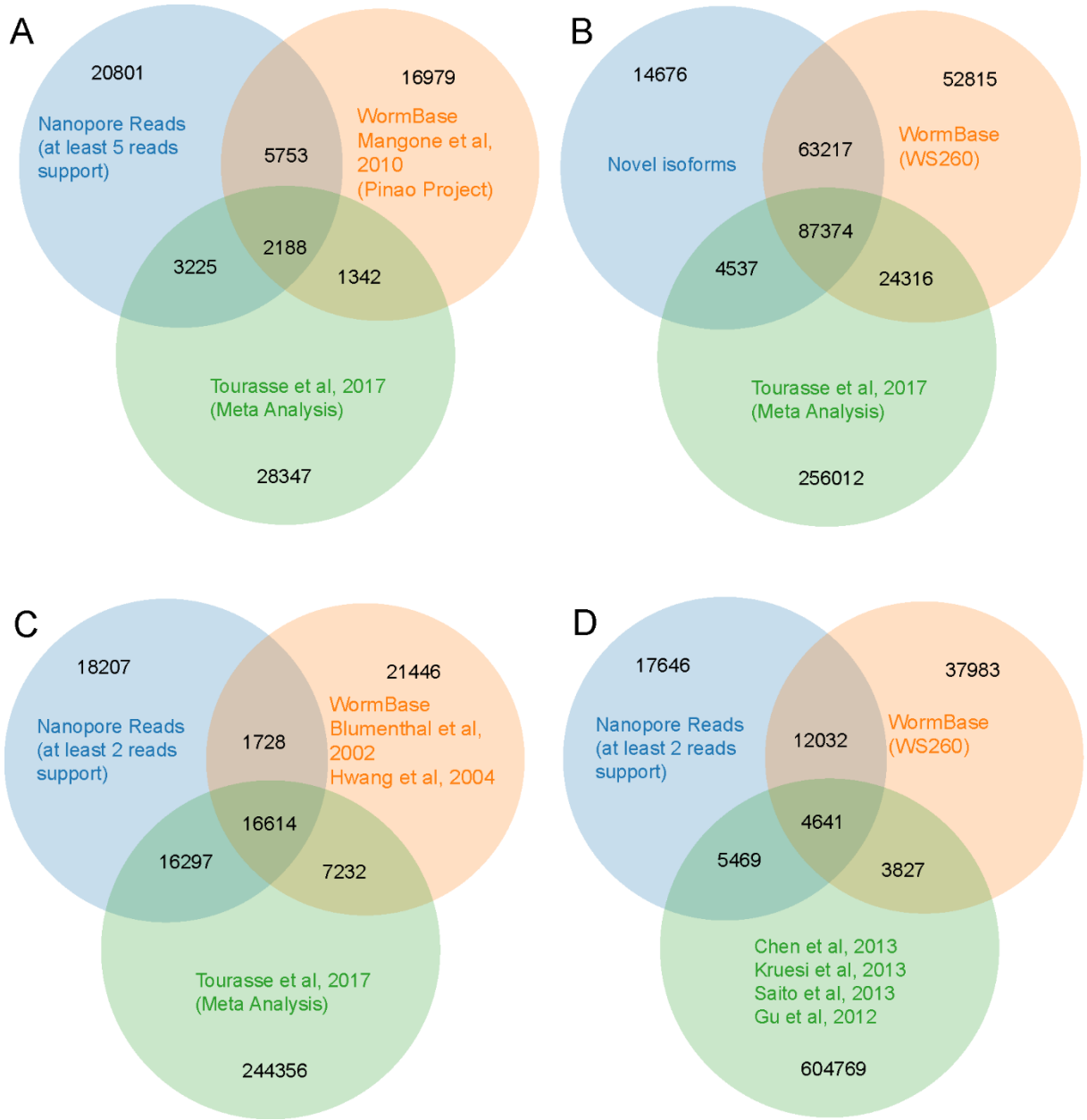
[1]Department of Biology, Hong Kong Baptist University, Hong Kong, China, [2]State Key Laboratory of Environmental and Biological Analysis, Hong Kong Baptist University, Hong Kong, China

#These authors contributed to the manuscript equally.

[*]Corresponding author

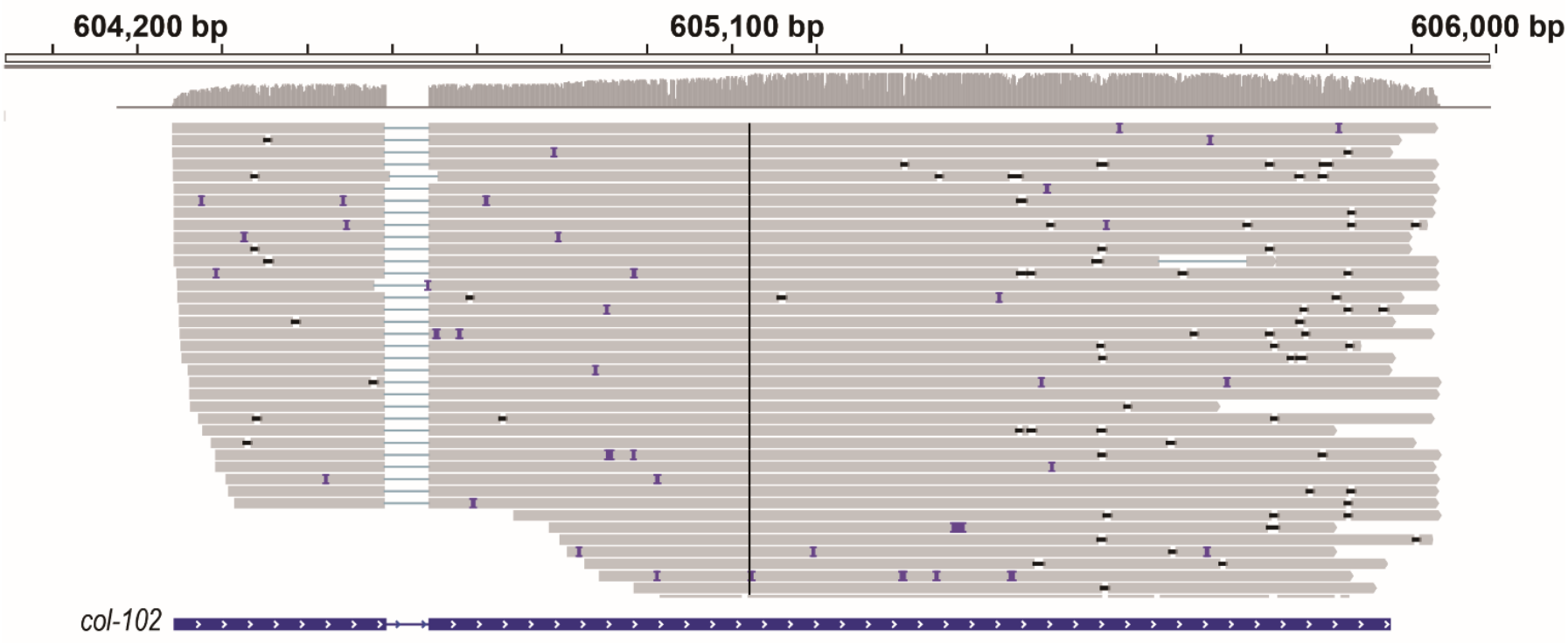
Email: zyzhao@hkbu.edu.hk

This PDF includes supplemental Figures S1-S13, Supplemental Methods and References.

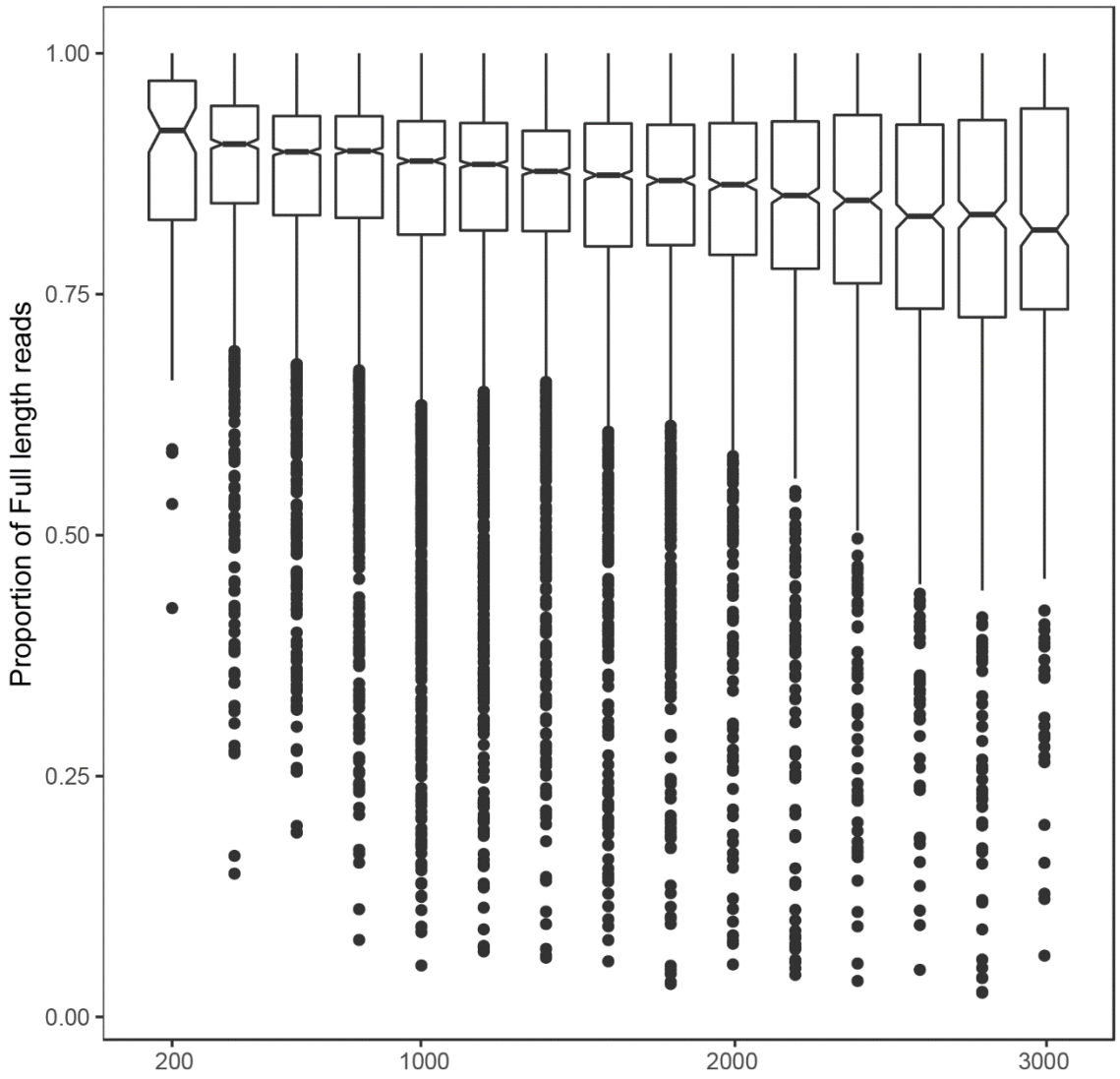


Supplemental Figure S1.

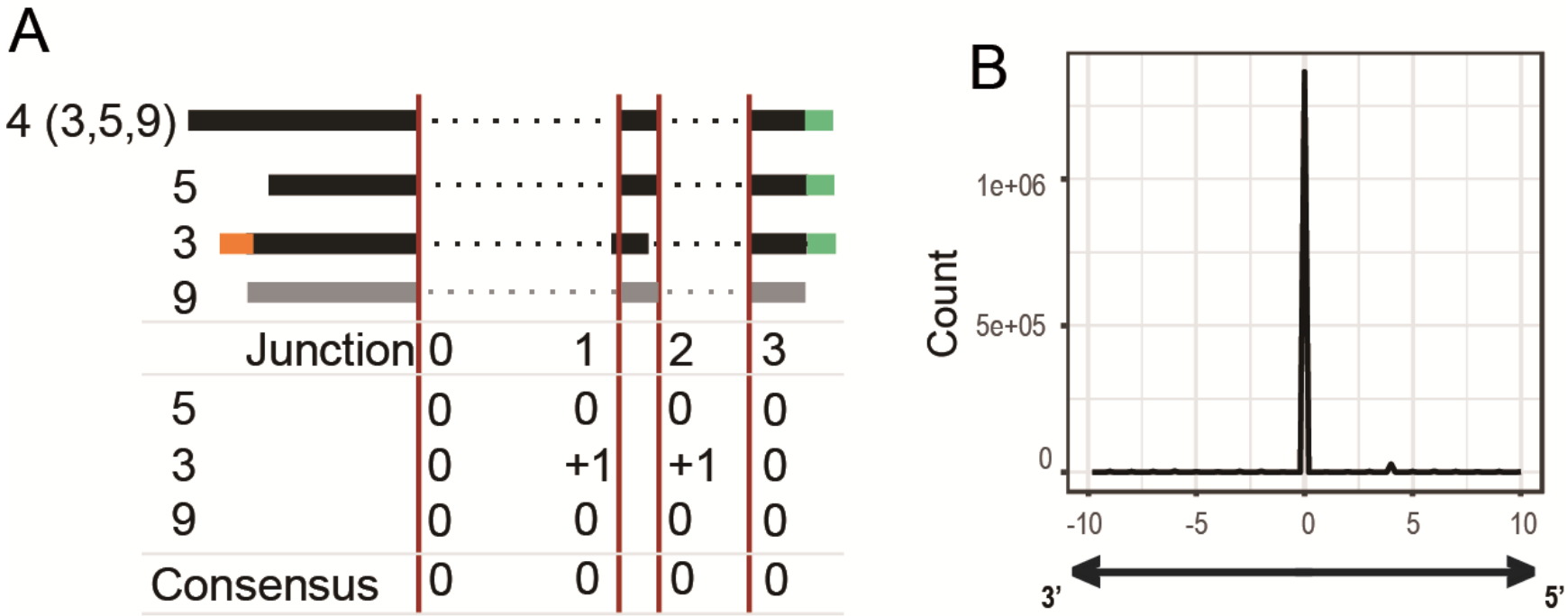
Comparisons of isoform characteristics between those derived from the long reads and those annotated in WormBase and those from Mangone et al, 2010 or those from meta-analysis of NGS RNA-seq data (Tourasse et al, 2017). (A) Venn diagram showing the intersection of polyadenylation sites derived from at least five long reads and those from Mangone et al, 2010 or Tourasse et al, 2017. (B) Venn diagram showing the intersection of splicing junctions derived from TrackCluster results and those from WormBase annotation (WS260) or from Tourasse et al, 2017. (C) Venn diagram showing the intersection of trans-splicing junctions derived from at least two Nanopore reads and those from Blumenthal et al, 2002 and Hwang et al, 2004 (downloaded from WormBase WS260) or from Tourasse et al, 2017. (D) Venn diagram showing the intersection of transcriptional start sites identified with Nanopore long reads (this study) and those annotated in WormBase (WS260) or those identified with Cap-seq studies by four groups.



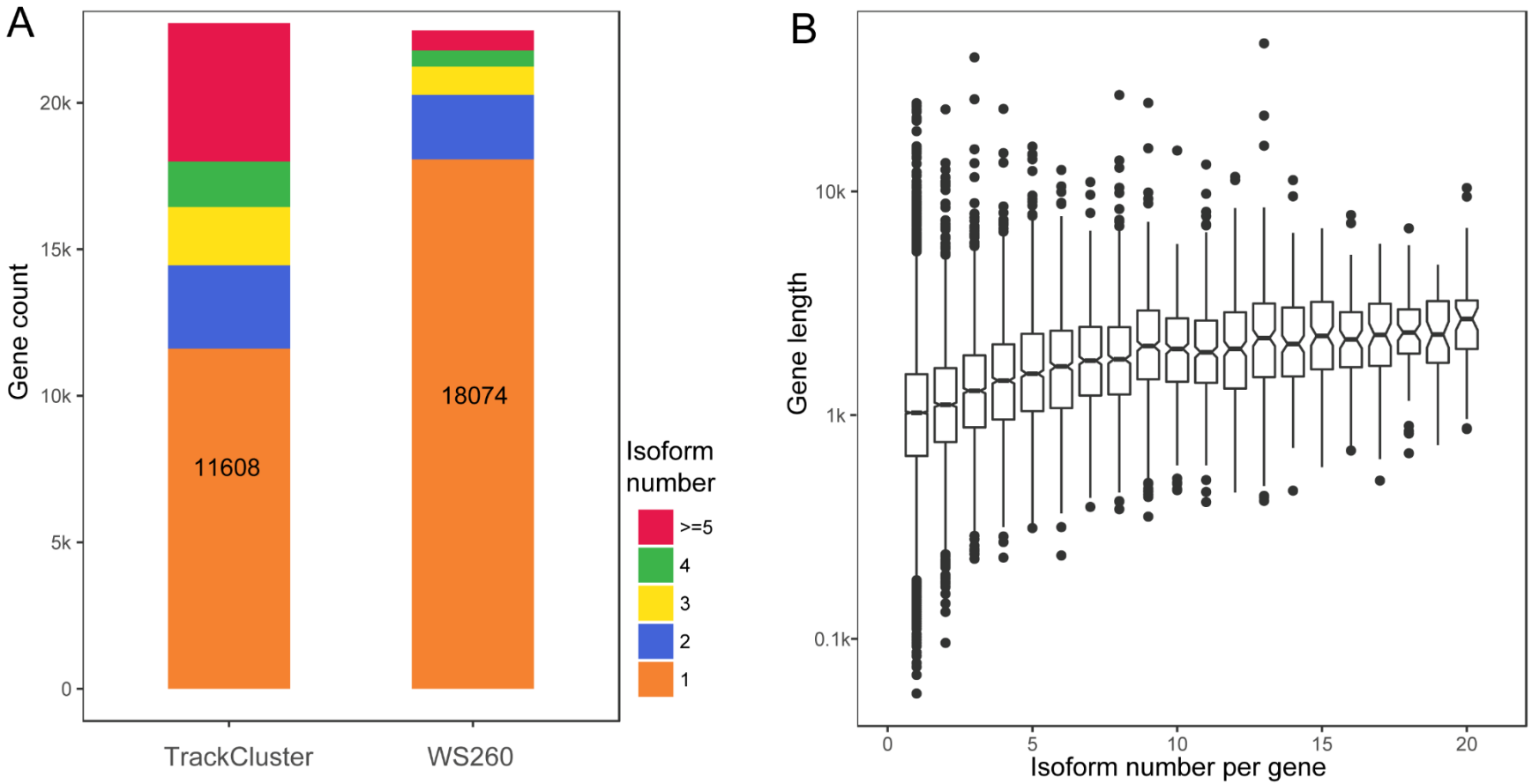
Supplemental Figure S2. An IGV track view of the reads mapped to the gene *col-102*. Notably, the 5' end of the reads demonstrate decreasing length compared with a full-length annotated transcript shown on the bottom. Insertion and deletion are shown as "I" and a gap in the track, respectively.



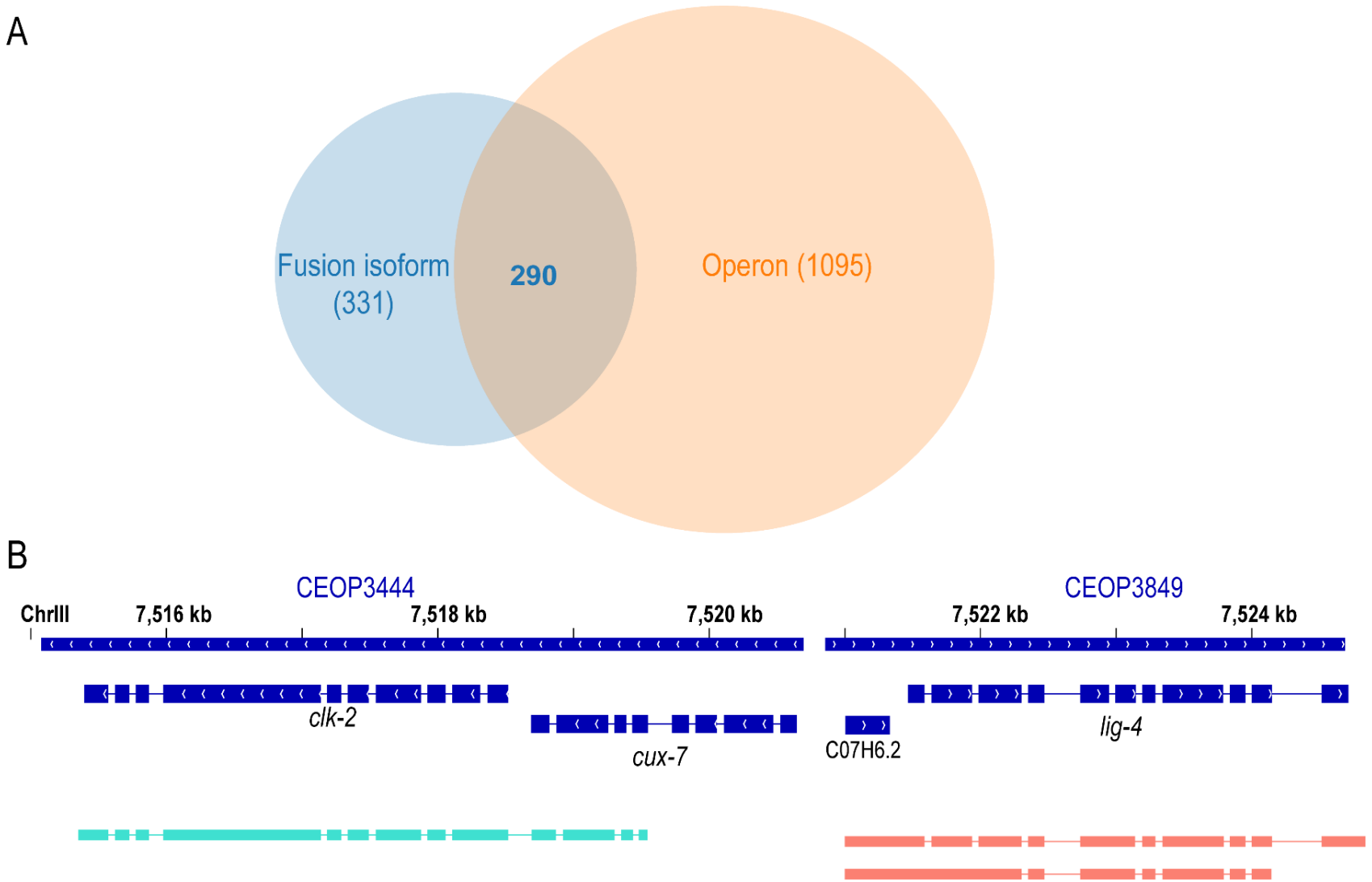
Supplemental Figure S3. Boxplots of the proportion of full-length reads out of all reads (x axis) against the isoforms of various lengths in nt (y axis). The median proportion decreases from approximately 95% to 85% for transcripts ranging from 200 to 3000 nt.



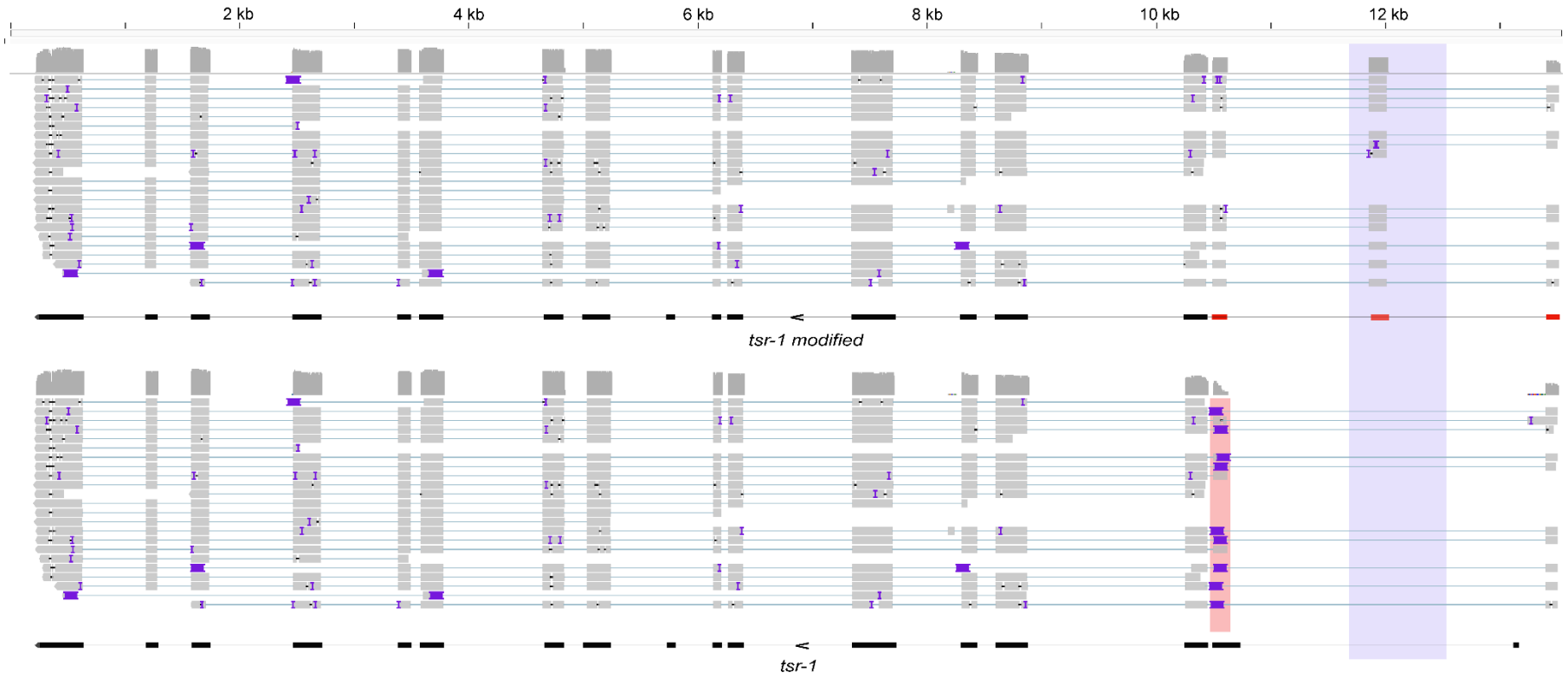
Supplemental Figure S4. Correction of intron-exon junction. (A) Schematic diagram for the splicing junction correction. Shown here is a representative isoform 4 in Fig. 2, which are associated with 3 subreads (3, 5, 9). The junctions from the "subreads" are aligned against the junctions defined in the isoform for the longest full-length long read. The shifted nt between the isoform and each "subread" are counted (shown at the bottom). A consensus junction was defined as the one with the highest frequency of support by all the "subreads". (B) Distribution of junction shifting after correction. "0" indicates that the junction remains unchanged. Shifting to the 5' and 3' in nt relative to the isoform junction is indicated with minus and plus number, respectively.



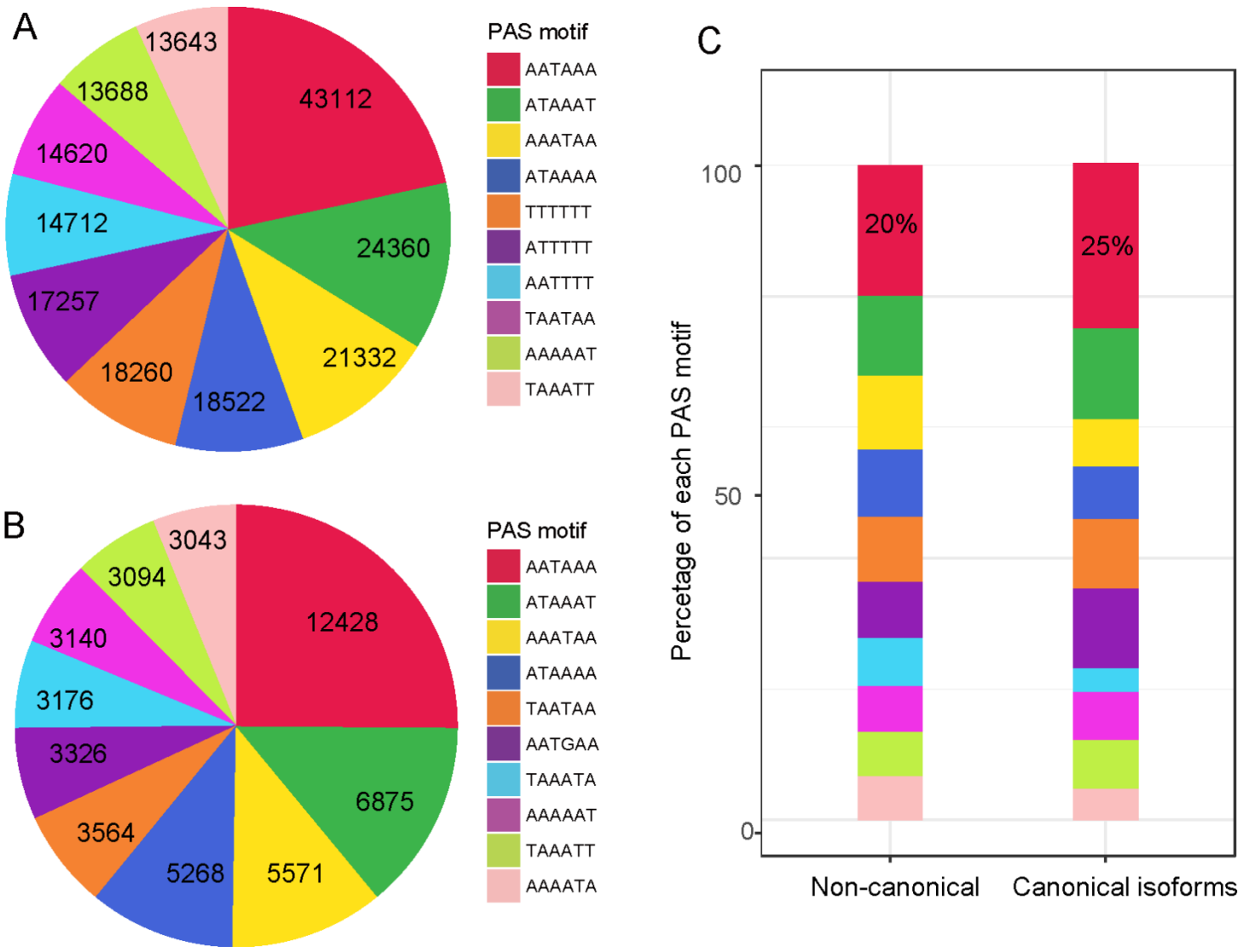
Supplemental Figure S5. Distribution of isoform number per gene. (A) Bar plots showing count of genes with various number of isoform. Isoforms output by TrackCluster or WormBase are shown on the left and right, respectively. (B) Correlation between gene size and isoform. Boxplots show gene length in nt versus isoform number per gene.



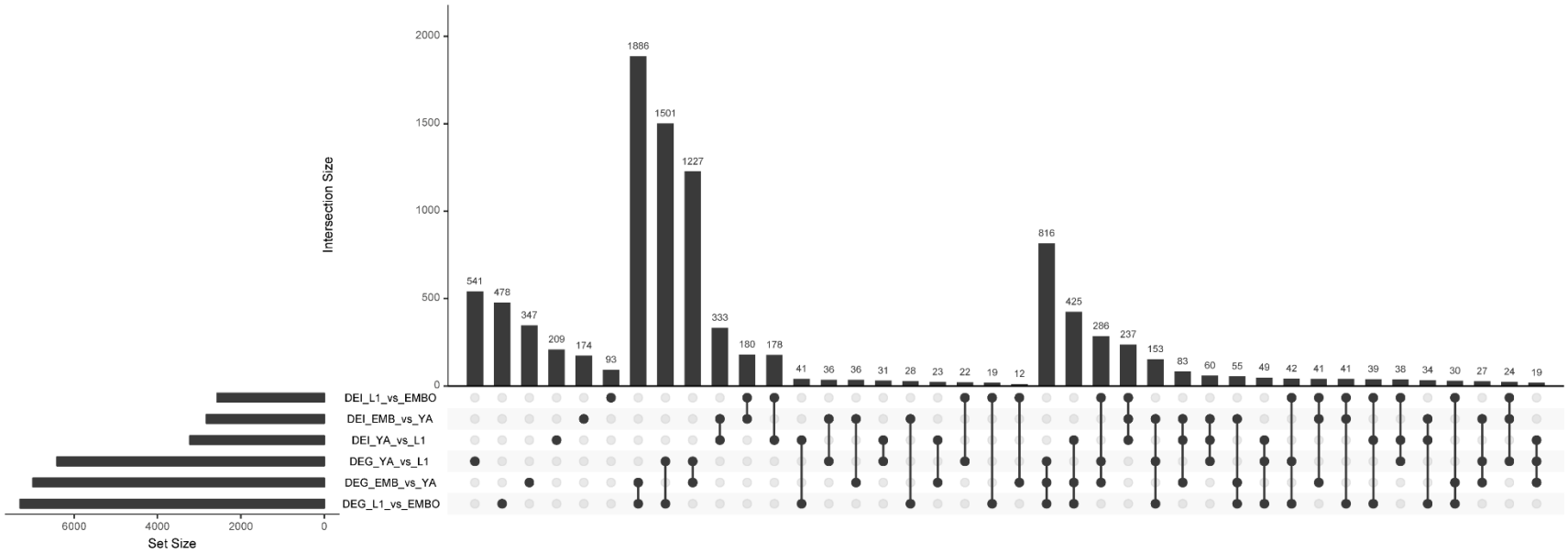
Supplemental Figure S6. Fusion isoforms. (A)Venn diagram showing the intersection between the fusion isoforms output by TrackCluster and the operons annotated in WS260. Around 46 % of the fusion isoforms are explained by operon. (B) An example of two head-to-head fusion isoforms each from two adjacent transcripts that are known to be operonic ones.



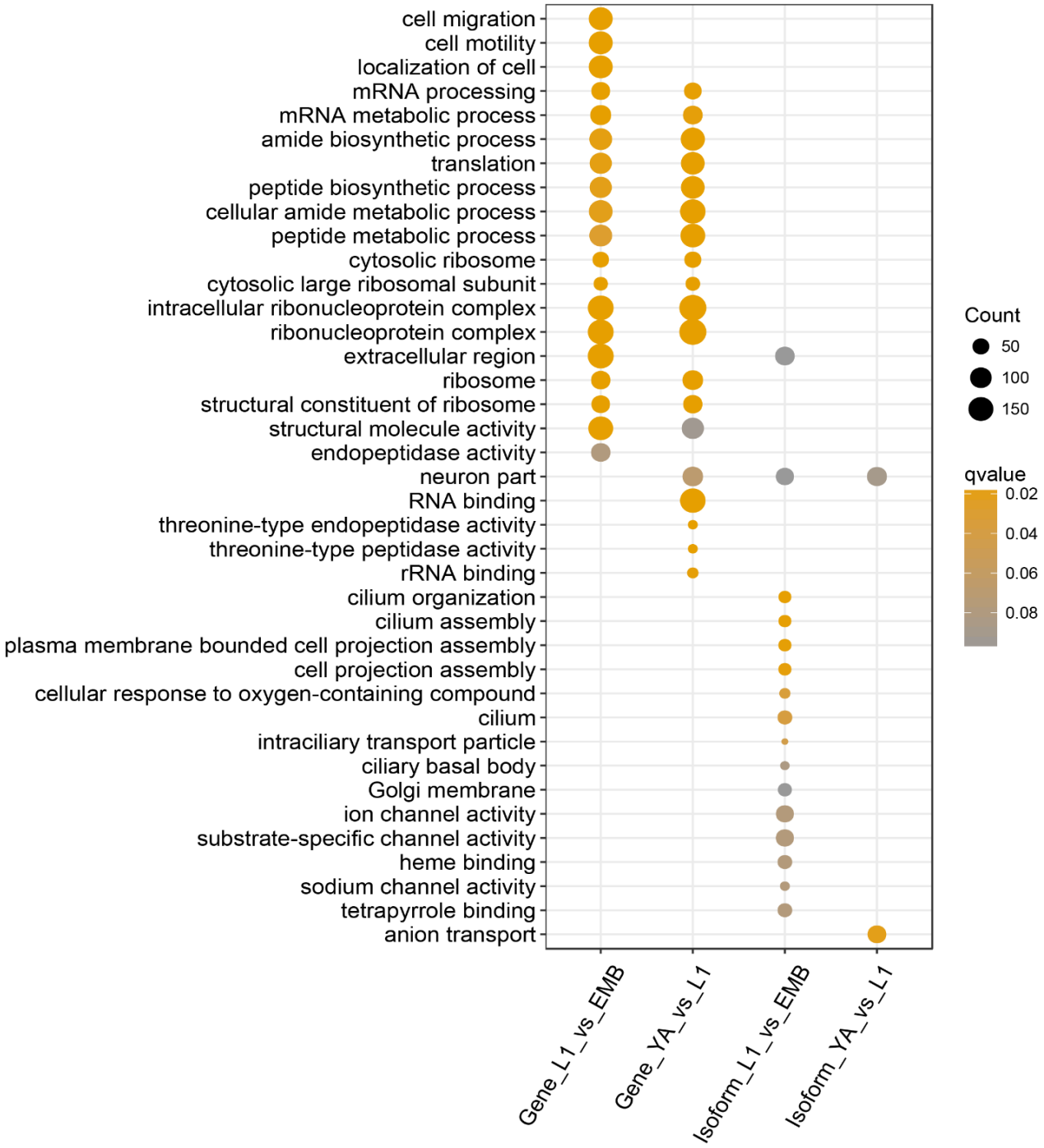
Supplemental Figure S7. A missing exon sequence in the N2 genome is recovered by long reads. Shown are mapping results of long reads against genomic region of gene *tsr-1*. Each mapped read is shown in a single line with accumulated coverage on the top. Corrected and existing gene models are shown in the middle and bottom, respectively. Existing and recovered exons are shown in black and red, respectively. Genomic region recovered by long reads is shaded in light blue. Alignment of long read with a small insertion (> 5nt but < 20 nt) relative to the reference genome is indicated with "I". Alignment of long read with a relative big insertion (> 100 nt) relative to the reference genome is shaded in brown.



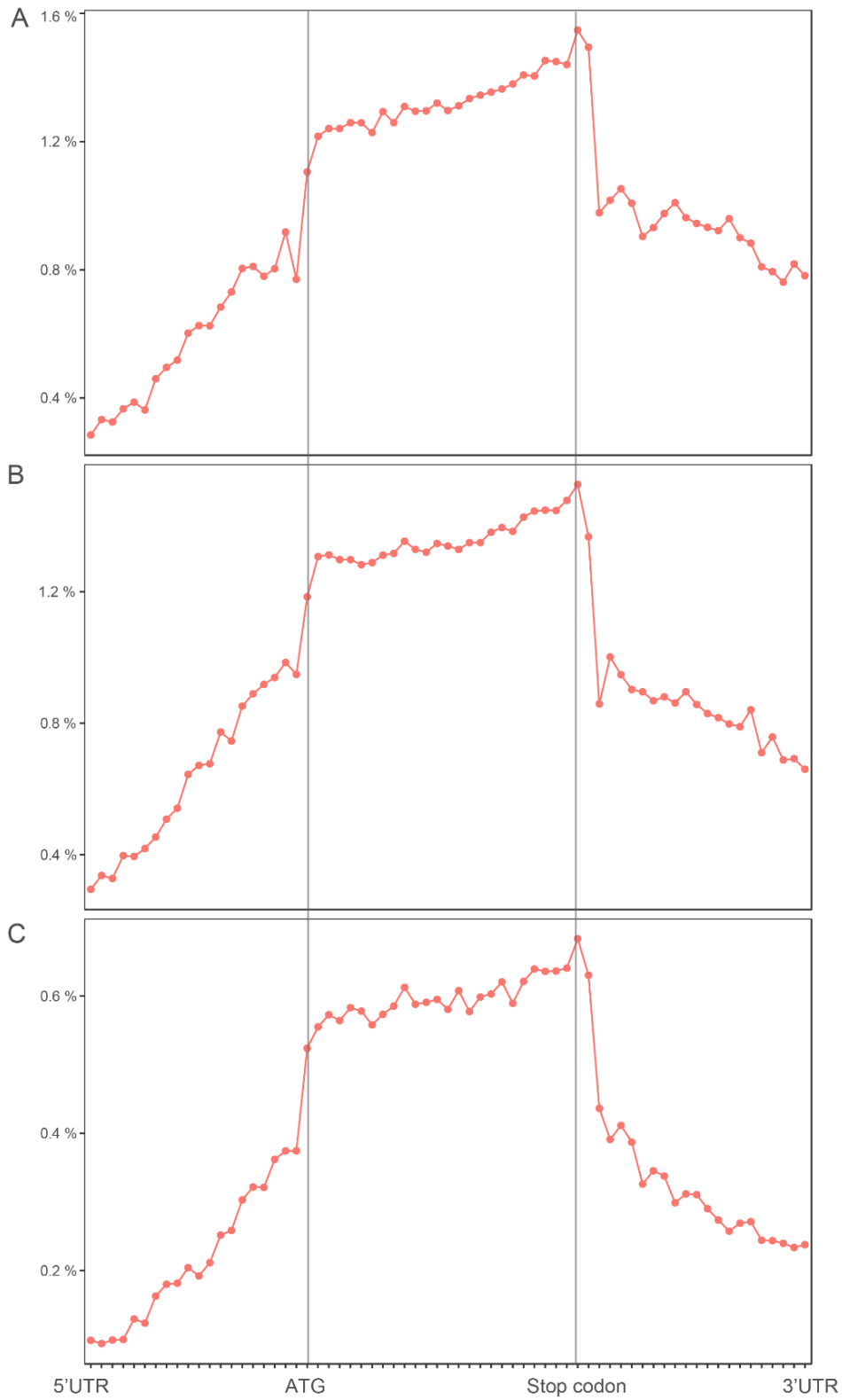
Supplemental Figure S8. Top 10 motifs for alternative PAS identified. (A) Distribution of PAS motifs identified using 351,437 PAS defined with at least two independent long reads (see Materials and Methods). (B) Distribution of PAS motifs identified using 44,700 isoforms output by TrackCluster. (C) Occurrence of the top PAS motifs in canonical and non-canonical isoforms (see text for details).



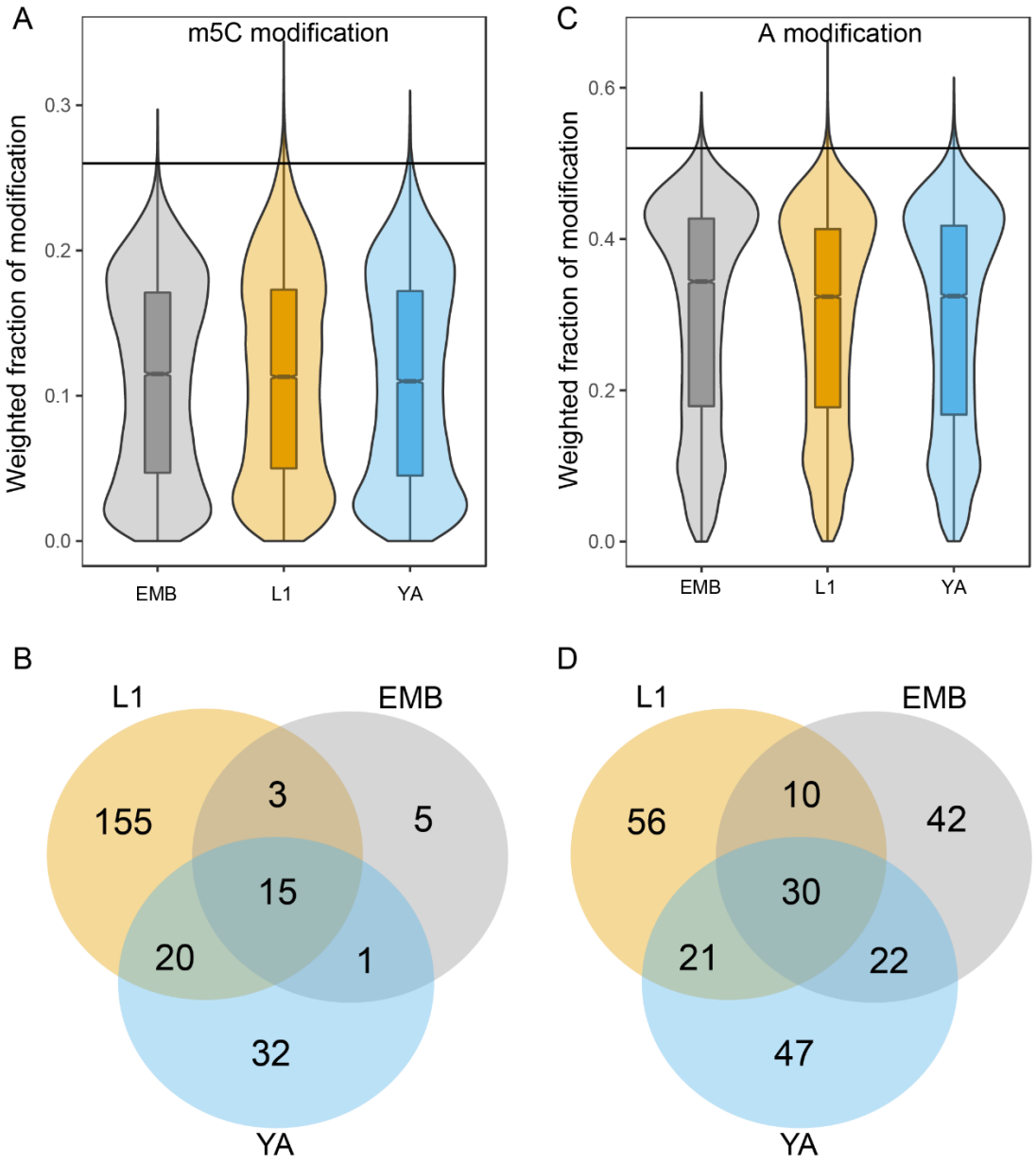
Supplemental Figure S9. Intersection of stage-specific expression between genes and isoforms. Upset plot shows the intersection between stage-specific expressed genes and isoforms. Six pairwise comparisons (shown on the left) are defined as individual group. The dot not associated with any line represents genes unique in this group. The dots linked with lines indicate groups used for intersection finding. Horizontal bars indicate gene number of each pairwise comparison. Vertical bars denote the numbers of intersection between groups.



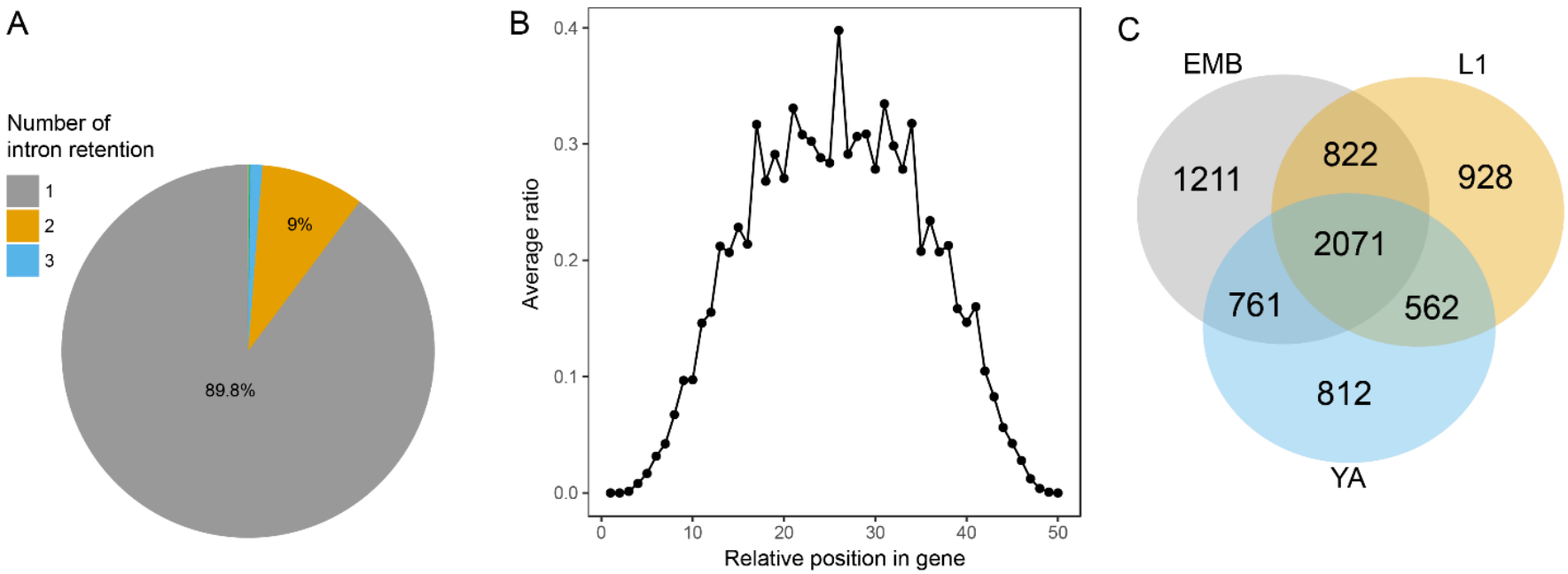
Supplemental Figure S10. Enriched GO terms for genes or isoforms with stage-specific expression. Shown are GO terms with a q value <0.1. Note that the sets of genes with differential expression show obvious overlap among various GO terms during development (comparison between the first two columns). However, sets of genes with differential isoform usage show little overlap among various GO terms during development (comparison between the first and the third or between second and fourth column). L1, L1 larvae; EMB, embryo; YA, young adult.



Supplemental Figure S11. Fraction of accumulative modification of $5^{m}C$ along the gene body in three developmental stages. (A) Embryo; (B) L1 larvae; (C) Young adult. Modifications of the nucleotide are predicted with Tombo (version 1.4) with "$5^{m}C$" mode



Supplemental Figure S12. RNA modifications over development. (A)Violin plots showing the distribution of weighted fraction of m5C (5-methylcytosine) type of modification in EMB, L1 and YA stage. The black horizontal line indicates the 99% confidential interval of the distribution. The genes with weighted fraction higher than this level was defined as significantly modified genes. (B) Venn diagram showing the intersection of the genes with significantly modified m5C between three different developmental stages. (C) Violin plots showing the distribution of weighted fraction of modification of adenine "A" in EMB, L1 and YA stage. The black horizonal line indicates the 99% confidential interval of the distribution. (D) Venn diagram showing the intersection of the genes with significantly modified "A" between three different developmental stages.



Supplemental Figure S13. Retained introns did not show obvious bias toward either end of gene body. (A) Relative ratio of the isoforms involving retention of one (1), two (2) or three or more (3) introns among all "intron retention" isoforms. (B) The average ratio of the retained introns alongside the relative position in gene body. The bin size for each isoform is 50 nt. (C) Venn diagram showing the intersection of gene numbers with intron retention supported by at least 2 reads in three developmental stages.

Supplemental Methods

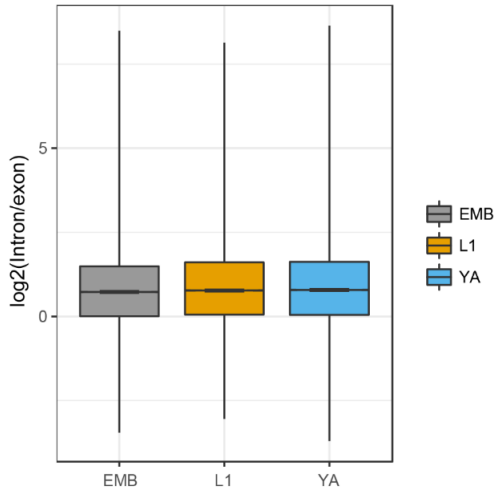
Purification and sequencing of mRNAs with MinION

Animals were as described (Ho et al. 2015) and synchronized as embryo, L1 and young adult following a previous study (Grün et al. 2014). Briefly, *C. elegans* (N2) was cultured on NGM plates with OP50 E. coli at 20°C. Gravid adult worms were treated with bleach to isolate embryos. The embryos were incubated in M9 buffer without food at room temperature for 12h to hatch and arrest at the L1 stage for harvesting. Part of the starved, synchronized L1 larvae were fed with OP50 and cultivated at 20°C until adulthood to be harvested for RNA preparation. Animals of different stages, i.e., embryo, L1 larva and young adult, were collected and total RNAs were extracted using TRIzol (Invitrogen) following the manufacturer's instructions. Approximately 100 µg total RNAs were extracted for each sample. Around 900 ng of poly(A) tailed mRNAs was purified using Dynabeads™ mRNA Purification Kit (Invitrogen) based on the user's manual for each library preparation. Nanopore sequencing libraries were constructed using Direct RNA sequencing kit (cat# SQK-RNA001). The libraries were loaded onto Nanopore R9.4.1 flow cell (cat# FLO-MIN106) and sequenced on MinION acquired from Oxford Nanopore Technologies. The software used for sequencing was MINKNOW 2.1 with base-caller, Albacore (v2.0.1).

Defining novel isoforms with TrackCluster using the long reads

We devised the pipeline "TrackCluster" to identify novel isoforms by clustering of isoforms based on their similarity in track structure, i.e., combination of intron/exon and their positions. Briefly, after mapping, each read mapped to the reference genome was converted to a read track in bigGenePred format. Tracks from each locus were subjected to three rounds of filtering steps to generate novel isoforms (Fig. 2A). First, exons defined by read tracks were compared against all the existing isoforms within a given locus. Any track that did not overlap with any existing isoform for over 50 nt was excluded from being used for novel isoform calling. Second, the following formula were used for calculating pairwise similarity score between two tracks for clustering the tracks into distinct groups. For example, in track A and B, the amount of shared sequence between

exons from each track was calculated in nt ($A^{\text{exon}} \cap B^{\text{exon}}$). The total number of nt were also calculated as the pooled exon sequence in nt between the same two exons ($A^{\text{exon}} \cup B^{\text{exon}}$). The number of shared and pooled intron sequences were similarly calculated as in exon sequence, and was normalized with a weight of 0.5, assuming average intron length is about twice of that of exon length in *C. elegans* (WormBase WS260) which was also supported by our long reads (See the supporting figure below).



Boxplots showing the ratio of summed intron length over summed exon length derived from long reads in three developmental stages. Only the long reads with intron longer than 1000 nt was used.

If the similarity score was higher than 95% between tracks, the one with a shorter length in summed exons was treated as a "subread" and merged to the one with a longer length. This step served to cluster the tracks with similar structures into distinct groups.

$$score\ 1 = \frac{(A^{\text{exon}} \cap B^{\text{exon}})/(A^{\text{exon}} \cup B^{\text{exon}}) + weight * (A^{\text{intron}} \cap B^{\text{intron}})/(A^{\text{intron}} \cup B^{\text{intron}})}{1 + \text{weight}}$$

Third, for the remaining tracks that showed a similarity score lower than 95% between each other, a pairwise similarity score 2 was calculated as follows.

$$score\ 2 = \frac{(A^{\text{exon}} \cap B^{\text{exon}})/\min(A^{\text{exon}}, B^{\text{exon}}) + weight * (A^{\text{intron}} \cap B^{\text{intron}})/\min(A^{\text{intron}}, B^{\text{intron}})}{1 + \text{weight}}$$

If the similarity score 2 was higher than 99% between tracks, the one with a shorter length in summed exons was treated as a "subread" and merged to the one with a longer length. This step served to merge the tracks from partial-length long reads with the one defined by a full-length read. Representative isoform(s) for a locus were/was generated as a result.

TrackCluster was first run using full-length long reads, and then with the remaining reads. We did this for three purposes. First was to reduce data processing time. Second was to determine the

expression level of isoforms using all long reads as described below. Third was to recover the isoforms that were potentially missed by our cutoff. We recovered a fraction of isoforms with a "truncated" 5' end ("5' missing" or "UTR truncations" (Supplemental Figs. 2 & 3)) relative to all existing transcripts using the remaining reads. For most of these isoforms, we defined them based on presence of an SL, but those newly recovered "truncated" ones were determined without an SL, which were judged by deviations in their remaining part from any existing transcript.

To quantify isoform for each locus, the subreads for each representative isoform was counted. If one subread showed 99% identity to with multiple representative isoforms (number = N), the count for each of these isoforms was counted as 1/N.

Supplemental References

Grün D, Kirchner M, Thierfelder N, Stoeckius M, Selbach M, Rajewsky N. 2014. Conservation of mRNA and protein expression during development of C. elegans. *Cell Rep* **6**: 565–77. http://linkinghub.elsevier.com/retrieve/pii/S2211124714000023 (Accessed February 9, 2017).

Ho VW, Wong MK, An X, Guan D, Shao J, Ng HC, Ren X, He K, Liao J, Ang Y, et al. 2015. Systems-level quantification of division timing reveals a common genetic architecture controlling asynchrony and fate asymmetry. *Mol Syst Biol* **11**: 814.

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

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078–2079. http://www.ncbi.nlm.nih.gov/pubmed/19505943 (Accessed July 20, 2018).

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841–2. https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btq033 (Accessed April 3, 2019).