Supplemental Materials

Contents

1 Supplemental Methods

1.1 Data preprocessing

The reads were filtered in a four-step process. First, the reads passing the Illumina quality filtering were selected. Second, the reads beginning with GG were selected because the first 2 nt must be GG in successful TSS-seq. The first GG of the selected read was trimmed before mapping. Third, we searched for an adapter sequence (TCGTATGCCGTCTTCT) and a primer sequence (AATGATACGGCGACCA) to remove contaminant reads. In this step, if the Levenshtein edit distance between the strings was ≤ 3, the string was considered to be the adapter or the primer. The reads that contained the adapter or primer sequence were removed. Fourth, rRNA-derived reads were removed by using BLAST (Zhang et al., 2000) against rRNA sequences. If reads had hits with E-value $< 10^{-4}$, they were considered to be rRNA-derived reads. The rRNA sequences (SSUParc_115.fasta and LSUParc_115.fasta) were downloaded from Silva (Quast et al., 2013). Finally, the SL sequence (ATTCTATTTGAATAAG), which is added to the $5'$ end of a pre-mRNA by SL *trans*-splicing, was searched in each read. In this step, if the Levenshtein edit distance between the strings was ≤ 3, the string was considered to be the SL sequence. Then, reads were classified into the following three classes: (i) reads without SL sequences, (ii) reads with the SL sequence at the $5'$ end, and (iii) reads with the SL sequence at any position except the $5'$ end. We refer to the first class as SL($-$) reads and the second class as SL(+) reads. The reads in the third class were not used because whether they were derived from *trans*-spliced or non-*trans*-spliced mRNAs was ambiguous.

1.2 Identification of TSSs and TASs

In order to identify TSSs and TASs, SL($-$) reads and SL(+) reads were mapped on a reference genome using NovoAlign (V2.07.11; http://www.novocraft.com) and MapSplice (version 1.15.2) (Wang et al., 2010). The KH assembly (Satou et al., 2008) was used as the reference genome of *C. intestinalis*. First, the reads were mapped to the genome using NovoAlign. Second, multiply mapped reads and unmapped reads were re-mapped to the genome using MapSplice (see Table S12 for the command options used in the mappings). The $5'$ end positions of uniquely mapped SL($-$) reads and SL(+) reads were considered candidate TSSs and TASs, respectively.

In order to determine candidate TSSs in human, we obtained mapping data of human TSS-seq reads from 15 samples (adipose, brain, breast, colon, heart, kidney, liver, lung, lymph, muscle, ovary, prostate, testis, thyroid, and adrenal gland) from DBTSS, which provides the mapped positions and the counts for each read at the base-pair level of the human genome (hg19).

1.3 Identification of TSCs and TACs

TSCs, which are high-density regions of candidate TSSs, were identified by clustering. The clustering consists of two different iterations of clustering. In the first round of clustering, we merged the candidate TSSs in all samples and clustered them using a 35-bp sliding window to identify initial TSCs. Although clustering using a sliding window was used in a previous study using TSS-seq data (Yamashita et al., 2011), the drawback of this simple method is that it only considers the positions of TSSs. This drawback could cause unreliable wide and low-density clusters if there are noise TSSs with a low read frequency between true TSSs. Thus, to address this problem, we performed a second round of clustering that considered the frequency of TSSs.

The second clustering was applied to each initial TSC as follows. Each TSC consisted of up to 5 (in *C. intestinalis*) and 15 (in human) TSCs because we used 5 and 15 samples from *C. intestinalis* and human, respectively. First, we searched for frequent TSSs and peak TSSs of the TSC in each sample. The frequent TSSs were defined as the candidate TSSs with a frequency that was more than one-tenth of that of the most frequent candidate TSS in the TSC. The relatively high frequency indicates that they are likely to be non-noise TSSs. Also, of the frequent TSSs, the candidate TSSs with a frequency that was more than half of that of the most frequent candidate TSS were defined as peak TSSs. In this step, only TSCs with a sufficient number of tags (≥ 100 tags) were used because, if the number of tags is small, the TSS distribution may not represent the true distribution, leading to incorrect classifications of frequent TSSs and peak TSSs. Second, the frequent TSSs obtained from all samples were merged and clustered using the sliding window, yielding sub TSCs. Then, the TSC was split at the middle positions between adjacent sub TSCs. However, if the sub TSCs did not contain peak TSSs, they were considered ambiguous TSCs and were not used in promoter analysis because of the absence of clear candidate TSSs with a high frequency. Finally, we repeated the first and second steps for all TSCs until none of them were split. In *C. intestinalis*, we also identified TAS clusters (TACs) by using the first clustering procedure above with a 4-bp sliding window.

After clustering, extreme outliers were removed to avoid clusters of an unreasonable size because of noise tags with a low frequency on both sides of the cluster. For each cluster of each sample, the interquartile range (IQR), the range between the 25th (Q1) and 75th percentile (Q3), was determined. Then, the tags below $Q1 - 3 \times IQR$ or above $Q3 + 3 \times IQR$ were defined as extreme outliers.

1.4 Locations of TSCs and TACs in *C. intestinalis*

In *C. intestinalis*, the locations of TSCs and TACs were examined based on the KH gene model (version 2013). The KH gene model contains three types of transcript models: non-SL, SL, and not determined (ND). The non-SL transcripts are non-*trans*-spliced transcripts, and their $5'$ end represents a TSS. On the other hand, the SL transcripts are SL *trans*-spliced transcripts, and their $5'$ end represents a TAS. Finally, the ND transcripts are the ambiguous transcripts: we do not know whether the $5'$ ends represent TSSs or TASs. The TSCs and TACs were classified into the following categories: TSS, TAS, $5'$ UTR, CDS, $3'$ UTR, intron, and intergenic. The intergenic category was defined as the regions where transcript models do not exist on the same strand. The location of each cluster (TSC and TAC) was determined based on the most frequent position in a core region, which was defined as the IQR of the representative cluster. For example, if the most frequent position was located in CDSs, the cluster was placed in the CDS category. The clusters that overlapped with TSSs or TASs of transcript models were placed in the TSS or TAS category unless they overlapped with coding regions. The TSCs and TACs that overlapped with the $5'$ ends of ND transcripts represent the TSSs and TASs of the transcripts, respectively, and therefore were placed in the TSS and TAS category. In addition, to avoid discrepancy between the positions of TSCs and annotated TSSs due to the incompleteness of the $5'$ end annotation, TSCs within ± 60 nt of $5'$ ends of nonSL or ND transcripts were placed in the TSS category unless the following two conditions were met: (i) other TSCs existed at the $5'$ ends and (ii) the TSCs were located within 60 nt upstream of the $5'$ ends, but there were ATG trinucleotides, which can be start codons, between the TSCs and the downstream $5'$ ends. If clusters could be classified into multiple categories because of overlapping transcripts, they were placed in one category according to the following priorities: (1) TSS, (2) TAS, (3) $5'$ UTR, (4) CDS, (5) $3'$ UTR, (6) intron, and (7) intergenic. For TACs, the priorities of the top 2 categories changed to (1) TAS and (2) TSS.

1.5 Locations of TSCs in human

In human, the locations of TSCs were classified into the following categories based on the RefSeq annotations: TSS, $5'$ UTR, CDS, $3'$ UTR, exon of non-coding RNA (ncRNA), intron, and intergenic. However, the TSCs located in $5'$ UTRs of 1st exons < 150 nt downstream of annotated TSSs were placed in the TSS category because many TSCs were located in $5'$ UTRs of 1st exons near annotated TSSs (Fig. S36). This discrepancy may be because the annotated TSSs do not represent major TSSs; instead, they are the most upstream minor TSSs of transcripts. We therefore placed them in the TSS category if other TSCs did not exist on the annotated TSSs of the transcripts. For the same reason, the TSCs < 30 nt upstream of annotated TSSs were also placed in the TSS category unless there were ATG trinucleotides between the TSCs and the downstream $5'$ ends. If TSCs could be placed in multiple categories because of overlapping transcripts, they were placed in one category according to the following priorities: (1) TSS, (2) $5'$ UTR, (3) CDS, (4) $3'$ UTR, (5) exon(ncRNA), (6) intron, and (7) intergenic.

1.6 Removing CTGG TSCs

TSS-seq may erroneously generate artifacts by unexpected hybridization of $5'$ oligos to CCAG as shown in Fig. S9. The reads derived from the erroneous hybridization result in false TSSs in which the 4-nt sequence immediately upstream of them is CTGG. If the mechanism shown in Fig. S9 is true, TSCs with the CTGG sequence are expected to be located on the antisense strand of exon regions. Indeed, many 1-bp width TSCs were among the TSCs that were located on the antisense strand of 5′ UTRs, CDSs, and 3′ UTRs (Fig. S37), and they exhibited clear CTGG motifs as expected (Fig. S38). We therefore removed TSCs that were located on the antisense strand of exons if the upstream 4-nt sequence was CTGG or a variant that was significantly overrepresented in TSCs (Table S13). In addition, even if not located on the antisense strand of exons, TSCs with the 4-nt CTGG sequence were removed. The number of removed TSCs is shown in Table S6 and S7.

1.7 Removing 1-bp width TSCs

The A+T content of 15-nt downstream regions of 1-bp width TSCs in introns and intergenic regions were examined in both *C. intestinalis* and human. We found that there were many A+T-rich TSCs with ≥ 0.80 and ≥ 0.66 A+T content in *C. intestinalis* and human, respectively (Fig. S11). Therefore, the 1-bp TSCs with ≥ 0.80 and ≥ 0.66 A+T content were removed in *C. intestinalis* and human, respectively. In addition, 1-bp TSCs within -3 to $+2$ relative to splice donor sites on the reverse strand were removed in both *C. intestinalis* and human.

1.8 Removing TSCs in CDSs and 3′ UTRs

1.8.1 Removing possibly truncated RNA-derived TSCs

Many TSCs were found in exons, such as CDS and 3′ UTRs. These TSCs may be false TSCs derived from truncated RNAs. Assuming that transcribed RNAs are broadly truncated in exons independent of the nucleotide preference, unlike transcription, which preferentially starts from PyPu sites, tags derived from the 5′ ends of truncated RNAs should be widely mapped to the exons downstream of their TSSs, producing TSCs in the exon regions. In addition, TSCs in exon regions often had their peak positions, the most frequent positions, near *cis*- or *trans*-splice acceptor sites (Fig. S13D), and many (> 50%) of them with peaks within -5 to $+4$ relative to splice acceptor sites exhibited a right-skewed TSS distribution (see Supplementary Methods for the definition of right-skewed TSCs, Fig. S39). This result may suggest that the regions near splice acceptor sites are more easily truncated than other regions.

The simplest method to avoid false TSCs derived from truncated RNAs is to remove all TSCs in exons from a data set. However, according to the KH model, annotated TSSs exist in exons, due to overlapping transcripts. This simplest method has a possibility to remove true TSCs at annotated TSSs in exons, and therefore we need a different method to keep them as many as possible.

To find dubious TSCs that are likely derived from truncated RNAs, we searched for exon features in which tags are widely distributed. For each exon feature, we calculated the proportion of the region covered by TSCs. If more than half of a region was covered by TSCs, the exon feature was defined as a dubious exon. Then, all the TSCs overlapping with exons of transcripts with at least one dubious exon were defined as dubious TSCs.

Each dubious TSC in a given transcript model was removed in the following steps. First, we examined whether there was a TSC at the annotated TSS of the transcript model because there should be not only false TSCs on exons but also a true TSC at the annotated TSS when a portion of but not all transcribed RNAs are truncated. In this case, the true TSC at the annotated TSS would have higher TSS peaks than the false TSCs on exons. Thus, when there was a TSC at the annotated TSS, the dubious TSC was regarded as a false TSC and was removed if its peak height was lower than one-fifth of that of the TSC at the annotated TSS in all samples. Second, when there was no TSC at the annotated TSS, but there were other dubious TSCs on the same, previous, or next exon, the dubious TSC was removed if, in all samples, its peak was not greater than five times higher than the other dubious TSCs. However, when the transcript model has only one exon, the dubious TSC was removed if it covered > 50% of the entire exon. Third, dubious TSCs with right-skewed TSS distributions were removed if their peaks were located near *cis*- or *trans*-splice acceptor sites (-5 to $+4$, where the 0 position represents the first position of the exon). The TSCs that were removed in the above steps were regarded as false TSCs that were likely derived from truncated RNAs. Finally, TSCs on exons were removed if, in all samples, they did not have peaks that were more than five times higher than those of the false TSCs in the same transcript model.

1.8.2 Removing remaining TSCs in CDSs and $3'$ UTRs

By using the method described above, we removed possibly truncated RNA-derived TSCs overlapping with exon regions (Table S6 and S7). However, remaining TSCs in CDSs and $3'$ UTRs (TSCs that were not located at annotated TSSs and were placed into CDS and $3'$ UTR category by the classification described in 1.4) did not exhibit clear PyPu motifs (Fig. S40), suggesting that they still contain many false TSCs. To increase the reliability of a final set of TSCs as much as possible, we removed all the remaining TSCs in CDSs and $3'$ UTRs. The total number of removed TSCs is shown in Table S6 and S7.

1.9 Right-skewed TSCs

A right-skewed TSC was defined as a TSC with a TSS distribution that met two of the following conditions [(i and ii) or (i and iii)]. (i) The 5th percentile is not equal to the 95th percentile, and the mode is less than or equal to the 5th percentile. (ii) There are no peak TSSs on the right side from the mode. The peak TSSs were defined as TSSs with a frequency that was more than half of that of the most frequent TSS. (iii) The skewness is more than zero. The skewness (y) was defined by the equation below.

$$y = \sqrt{n}\frac{\sum(x_i - \overline{x})^3}{\{\sum(x_i - \overline{x})^2\}^{3/2}}$$

where n, x, and $\overline{x}$ represent the total number of tags in the TSC, the position of the tags, and the mean position, respectively.

1.10 TATA box searching

TATA boxes were searched using TRANSFAC’s MATCH program (Kel et al., 2003). The vertebrate position weight matrices (PWMs) in the TRANSFAC database (release 2014.2) (Matys et al., 2006) (V$TATA_C and V$TATA_01) were used with the following search conditions: plus strand; cut-off values of minSUM.prf. We used the first position of the core sequence of the PWM as the position of the TATA box.

1.11 Identification of orthologous RP genes

In order to find orthologous RP genes, we used protein-protein BLAST (Zhang et al., 2000). Protein sequences in the KH model were used as *C. intestinalis* protein sequences. All the protein sequences were searched against human protein sequences in the RefSeq database using BLAST with the E-value threshold 10^{-4}. Then, for each KH transcript, the human protein with the lowest E-value was considered the orthologous protein.

1.12 TSS distribution types

Promoters were classified into three types (sharp, broad, and other) based on their TSS distribution in the following steps. First, if more than 90% of tags in the TSS distribution were within 10 bp, promoters were defined as narrow (NR) promoters because most tags were focused in a narrow region (10 bp). The 10 bp was derived from the distribution of TSC width; it exhibited a sudden decrease from 1 to 10 bp (Fig. S41). Second, we searched for peak TSSs in the NR promoters to define a sharp peak. The peak TSSs were defined as the TSSs with a frequency that is more than half of that of the most frequent TSS. If the distance from the first to the last peak TSS was less than 5 bp, NR promoters were regarded as narrow and sharp peak (NSP) promoters. Third, the remaining promoters were classified into three groups according to the number of sharp peaks. For each promoter, peaks were identified by clustering the peak TSSs using a 10-bp sliding window. To evaluate whether the identified peaks are sharp or broad, we calculated the standard deviation (SD) of the TSS-tag distribution in the peaks. Then, if the SDs were less than a threshold, they were regarded as sharp peaks. The 90th percentile of the distribution of SDs derived from peaks in the NSP promoters was used as the threshold. If the promoter had one sharp peak, it was classified as a wide and sharp peak (WSP) promoter because there was one sharp peak, but its TSSs were not focused within the narrow region. If the promoter had multiple peaks and all of them were sharp peaks, it was classified as a multiple peak (MP) promoter because it had multiple sharp peaks. If the promoter was not classified as NR, WSP, or MP, it was classified as a broad (BR) promoter. We referred to the NSP promoters and the BR promoters as "sharp" promoters and "broad" promoters, respectively. The other types (NR, WSP, and MP) were considered ambiguous types and were integrated as "other".

1.13 Relative entropy

Expression specificity was evaluated by using Kullback-Leibler (KL) divergence, which is also known as relative entropy (Ponjavic et al., 2006; Zhao et al., 2011). The KL divergence (KLD) of a given cluster (TSC or TAC) was calculated by the following equation.

$$\mathrm{KLD} = \sum_{i=1}^{N} p_i log \frac{p_i}{q_i}$$

where i represents a sample; N is 5 in *C. intestinalis* and 15 in human; p and q represent a discrete probability distribution of tags in the cluster and in total clusters, respectively; and p_i and q_i represent the probability of sample i in p and q, respectively.

1.14 Hypergeometric test

Hypergeometric test was used to statistically evaluate the high expression of clusters (TSCs and TACs) in each sample. First, in each sample, a normalized expression level of each cluster, called the ppm value, was calculated by the following: number of tags in the cluster / total number of tags in the sample × 1,000,000 (Yamashita et al., 2011). Because the hypergeometric test is available for count data, the ppm value was rounded to the nearest whole number. However, we set the ppm value equal to 1 when the ppm value was more than 0 and less than 0.5. Then, for each cluster, we evaluated whether it was significantly highly expressed in each sample using the hypergeometric test. The p-value of a cluster in a given sample was calculated by the following equation.

$$p = \sum_{i=x}^{min(m,k)} \frac{\binom{m}{i}\binom{N-m}{k-i}}{\binom{N}{k}}$$

where N, m, k, and x represent the total expression levels of all clusters in all samples, the total expression levels of all clusters in the sample, the total expression levels of the cluster in all samples, and the expression level of the cluster in the sample, respectively. Because we performed this test for each cluster and for each sample, each p-value was multiplied by the number of clusters and the number of samples (Bonferroni correction).

To determine in which sample a given cluster is significantly highly expressed, we used the p-value of each sample that was calculated by the hypergeometric test and expression specificity evaluated by the KL divergence. First, clusters with low expression specificity (KL divergence < 0.7) were regarded as non-specific clusters. This threshold was chosen based on the distribution of KL divergence of clusters associated with RP genes; about 90% of the clusters showed a KL divergence less than this value. Then, if the p-value of the cluster in a given sample was less than 0.01, the cluster was considered to be significantly highly expressed in the sample. The clusters that were significantly highly expressed in only one sample (ovary, heart, body wall muscle, neural complex, or larva) were regarded as tissue- or larva- specific clusters.

1.15 Classification of pairs of TSCs and TACs

Predicted pairs of clusters were classified into two types: unannotated-operon-type and non-operon-type. If pairs of TSCs and TACs were located at the 5′ ends of different genes that do not overlap each other and do not constitute an operon in the KH model, they were classified into unannotated-operon-type. The other pairs were classified into non-operon-type. The regions between the non-operon-type pairs of TSCs and TACs were considered putative outrons.

2 Supplemental Figures

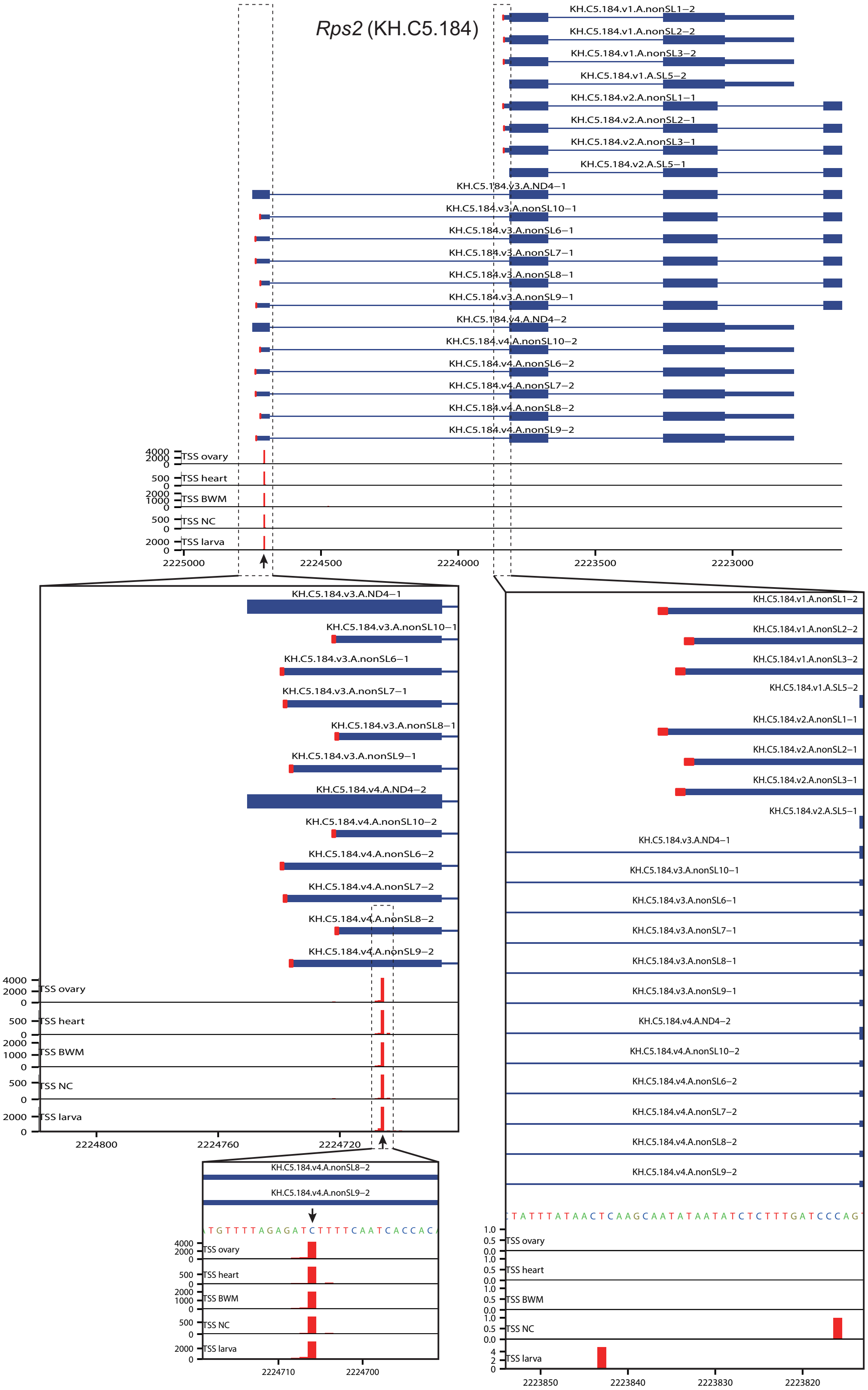


Figure S1: Distribution of mapped TSS-seq tags around the representative TSS and the annotated TSSs of the *Rps2* gene (KH.C5.184). The representative TSS is marked by a black arrow. The annotated TSSs, which are the 5′ ends of nonSL-type transcript models, are highlighted by red rectangles. The x and y axes represent the genome position and the number of mapped tags, respectively. The red bars show the distribution of mapped TSS-seq tags. BWM, body wall muscle; NC, neural complex.

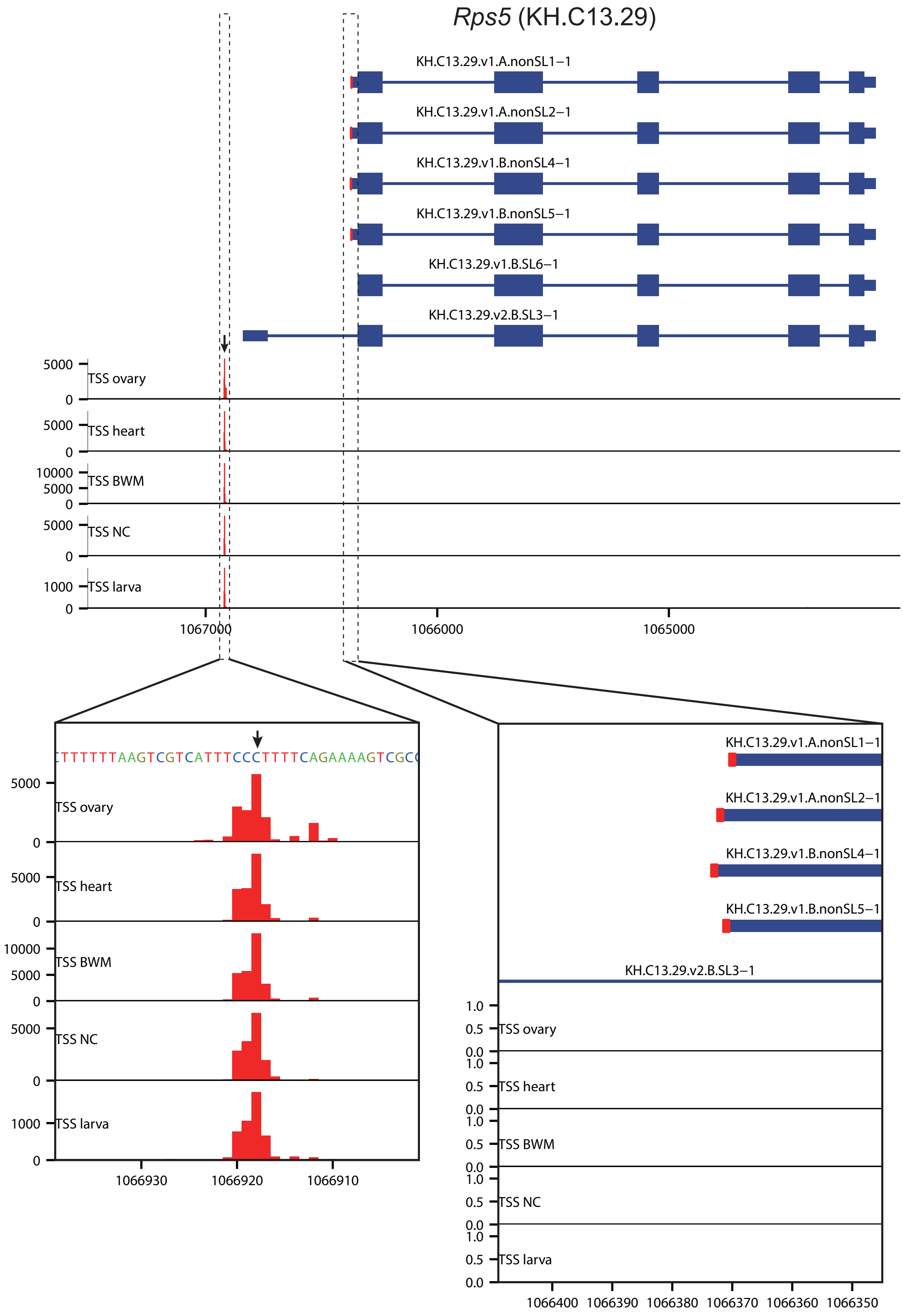


Figure S2: Distribution of mapped TSS-seq tags around the representative TSS and the annotated TSSs of the *Rps5* gene (KH.C13.29). The representative TSS is marked by a black arrow. The annotated TSSs, which are the 5′ ends of nonSL-type transcript models, are highlighted by red rectangles. The x and y axes represent the genome position and the number of mapped tags, respectively. The red bars show the distribution of mapped TSS-seq tags. BWM, body wall muscle; NC, neural complex.

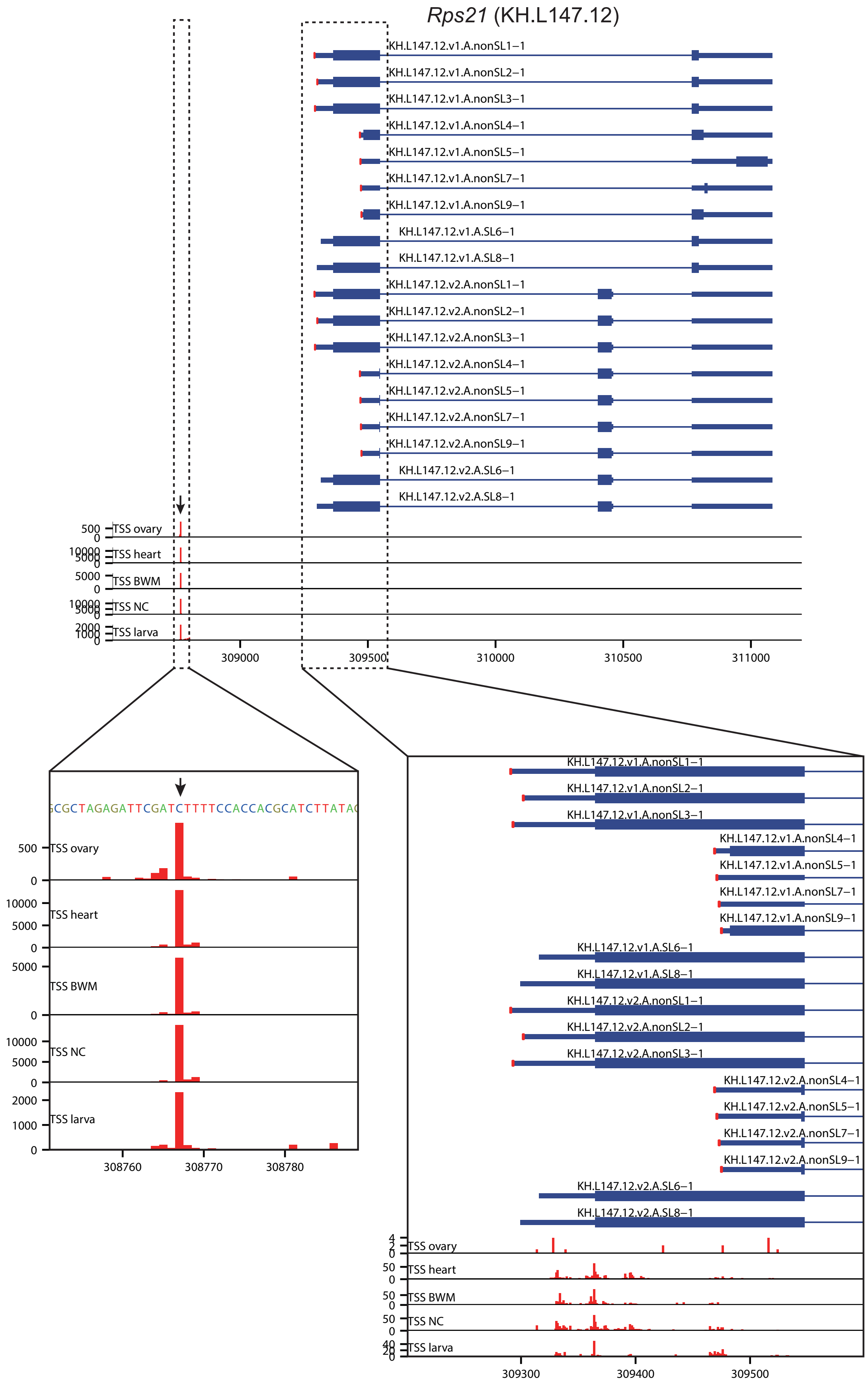


Figure S3: Distribution of mapped TSS-seq tags around the representative TSS and the annotated TSSs of the *Rps21* gene (KH.L147.12). The representative TSS is marked by a black arrow. The annotated TSSs, which are the 5′ ends of nonSL-type transcript models, are highlighted by red rectangles. The x and y axes represent the genome position and the number of mapped tags, respectively. The red bars show the distribution of mapped TSS-seq tags. BWM, body wall muscle; NC, neural complex.

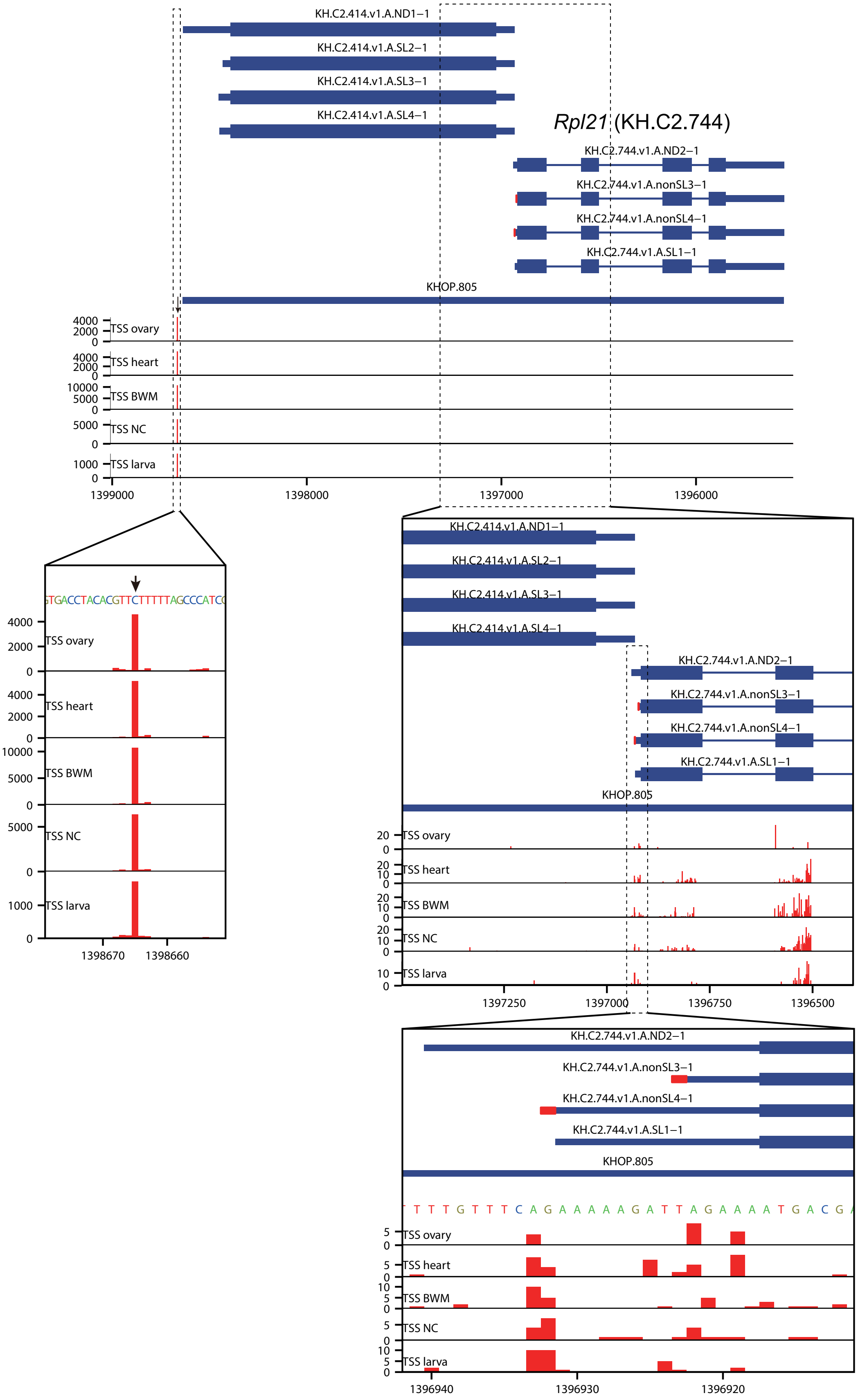


Figure S4: Distribution of mapped TSS-seq tags around the representative TSS and the annotated TSSs of the *Rpl21* gene (KH.C2.744). The representative TSS is marked by a black arrow. The annotated TSSs, which are the 5′ ends of nonSL-type transcript models, are highlighted by red rectangles. The x and y axes represent the genome position and the number of mapped tags, respectively. The red bars show the distribution of mapped TSS-seq tags. BWM, body wall muscle; NC, neural complex.

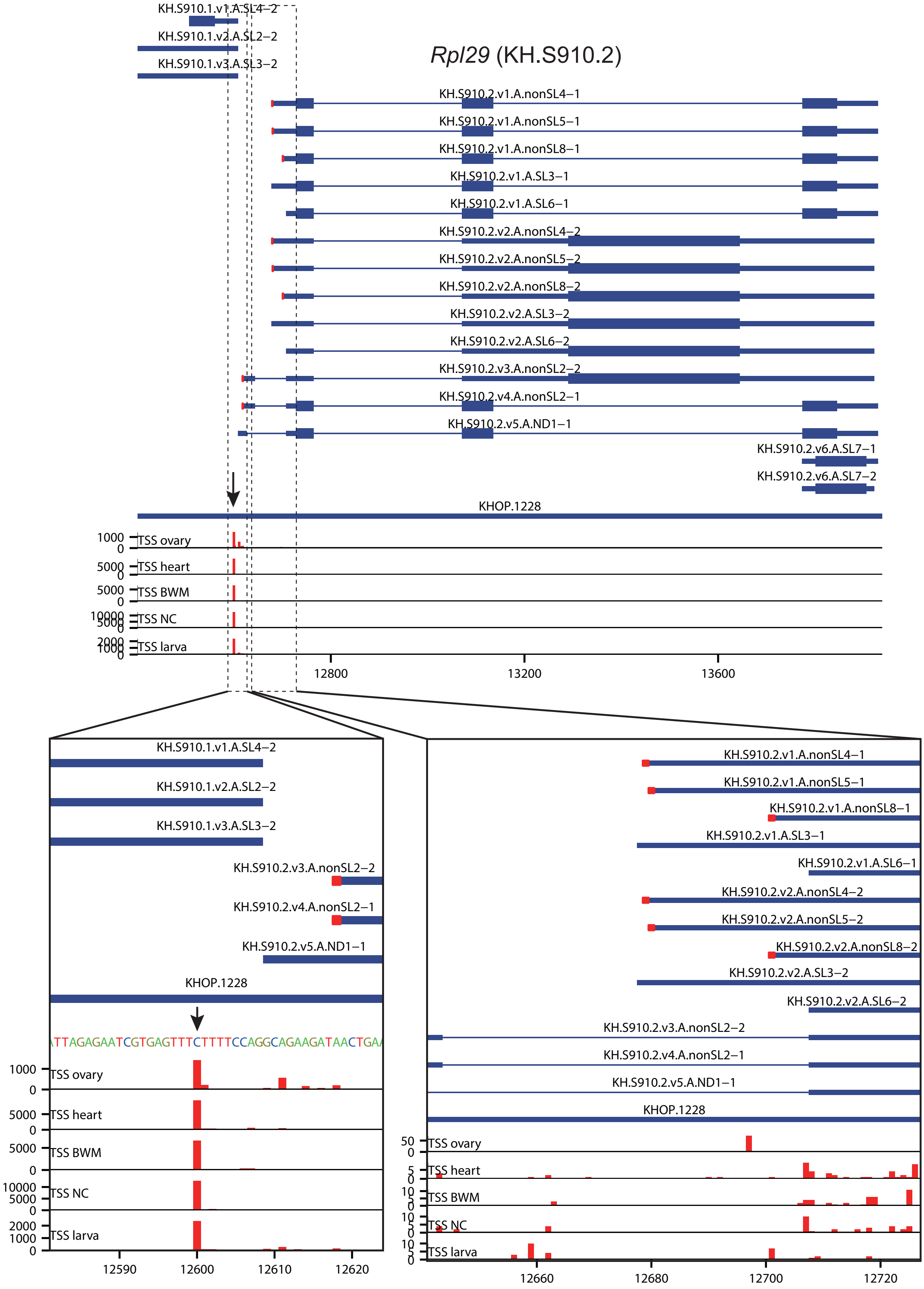


Figure S5: Distribution of mapped TSS-seq tags around the representative TSS and the annotated TSSs of the *Rpl29* gene (KH.S910.2). The representative TSS is marked by a black arrow. The annotated TSSs, which are the 5′ ends of nonSL-type transcript models, are highlighted by red rectangles. The x and y axes represent the genome position and the number of mapped tags, respectively. The red bars show the distribution of mapped TSS-seq tags. BWM, body wall muscle; NC, neural complex.

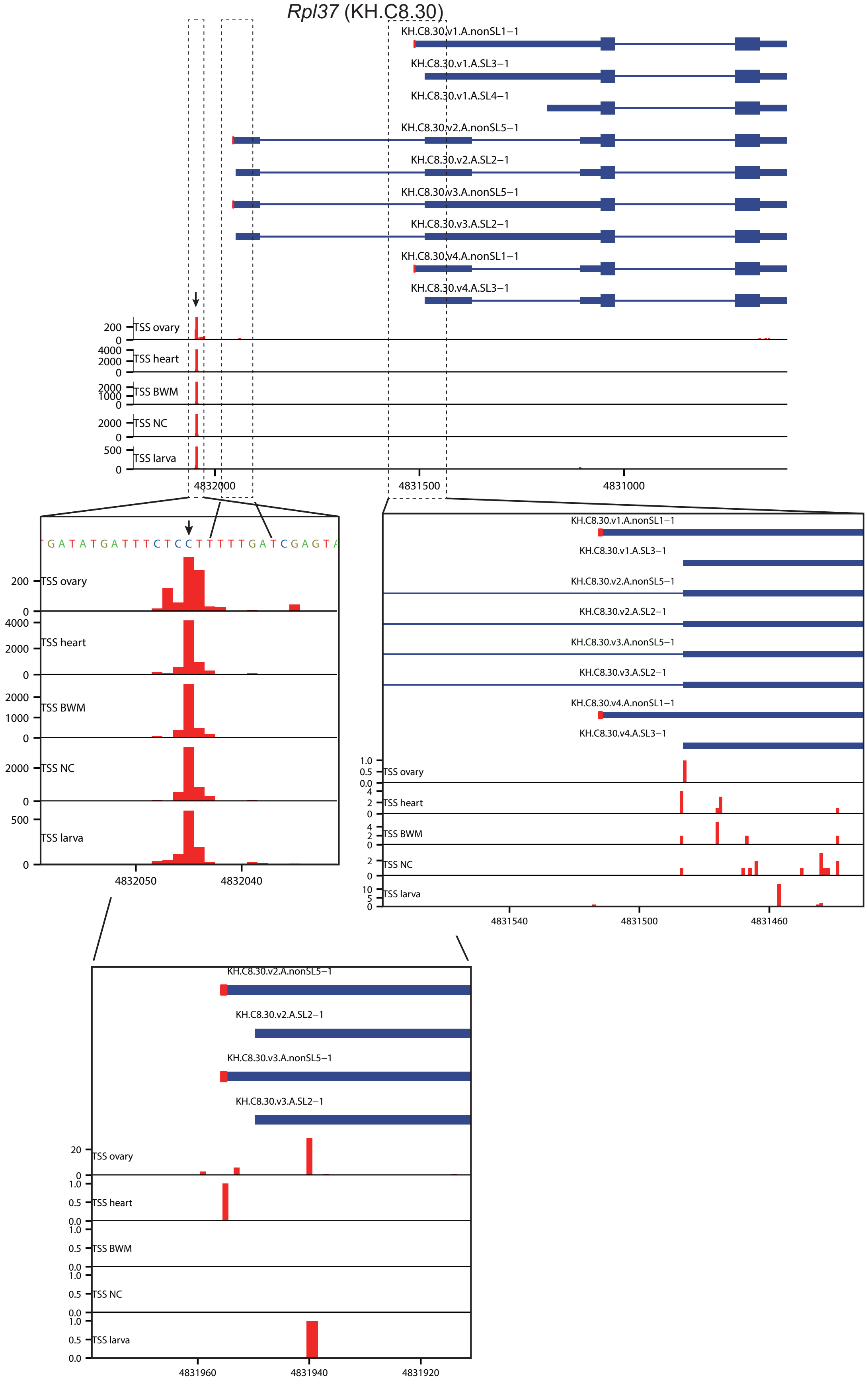


Figure S6: Distribution of mapped TSS-seq tags around the representative TSS and the annotated TSSs of the *Rpl37* gene (KH.C8.30). The representative TSS is marked by a black arrow. The annotated TSSs, which are the 5′ ends of nonSL-type transcript models, are highlighted by red rectangles. The x and y axes represent the genome position and the number of mapped tags, respectively. The red bars show the distribution of mapped TSS-seq tags. BWM, body wall muscle; NC, neural complex.

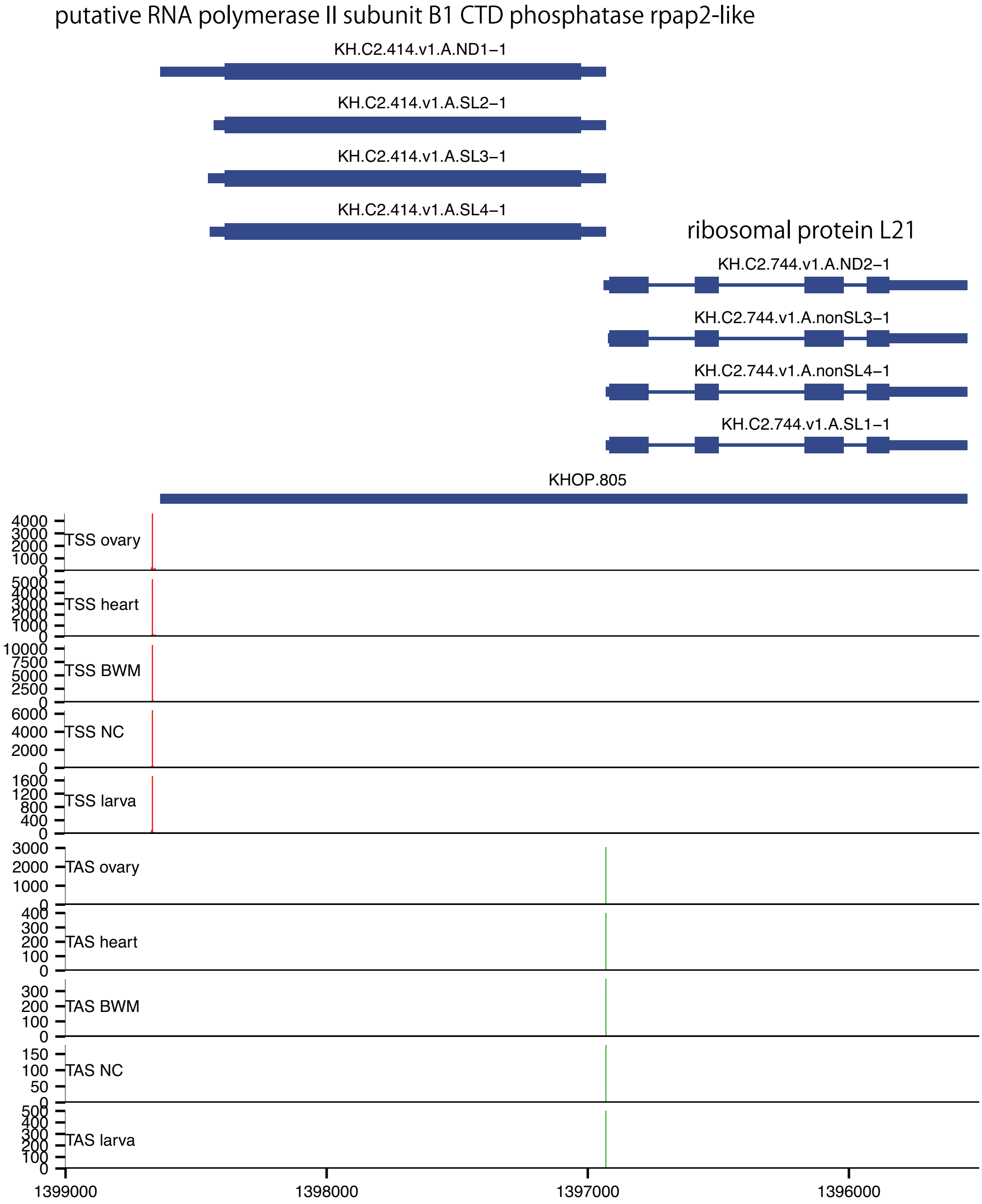


Figure S7: TSS of the ribosomal protein L21 gene. The ribosomal protein L21 gene is the dowstream gene of the operon (KHOP.805) that consists of two genes. The upstream gene encodes putative RNA polymerase II subunit B1 CTD phosphatase rpap2-like. The red and green bars represent the TSSs and TASs, respectively. The y axis represents the number of tags. BWM, body wall muscle; NC, neural complex.

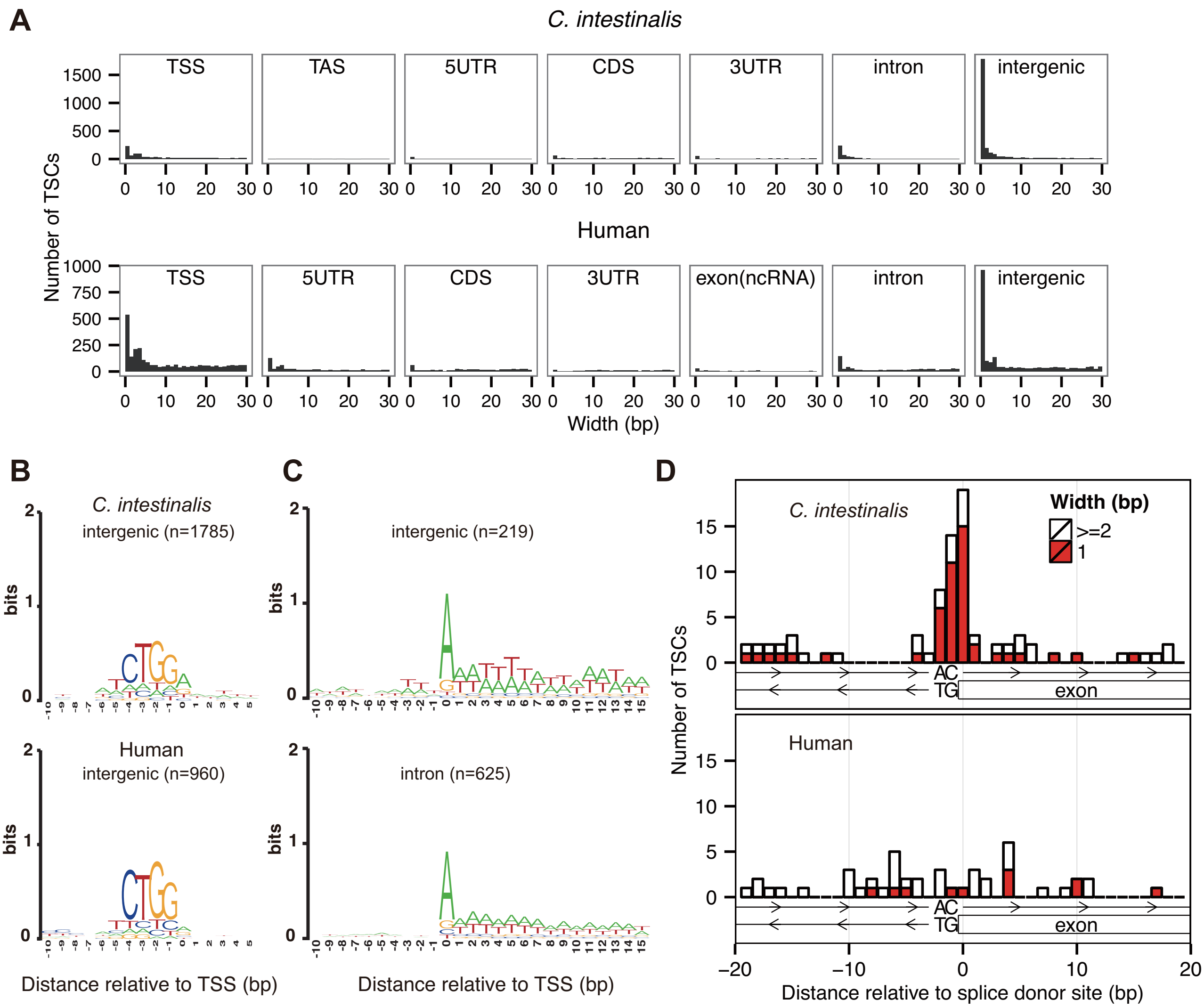


Figure S8: One-bp width TSCs. (A) Width of transcription start site clusters (TSCs). Only TSCs that are less than 30 bp in width are shown. Many 1-bp width TSCs were found in intergenic regions in both *Ciona intestinalis* and human. (B) Sequence logos of 1-bp width TSCs in intergenic regions. The sequences of the 1-bp width TSCs in intergenic regions were aligned relative to the representative TSS. The number in parentheses represents the number of TSCs. The 1-bp width TSCs in intergenic regions exhibited a clear CTGG motif immediately upstream of the TSS. The logos were created by sequence logo (Crooks et al., 2004). (C) Sequence logos of 1-bp width TSCs in introns and intergenic regions in *C. intestinalis*. The sequences of the 1-bp width TSCs in introns and intergenic regions were aligned relative to the TSS. The number in parentheses represents the number of TSCs. The region immediately downstream of the TSS was A+T-rich. (D) One-bp width TSCs near splice donor sites on the reverse strand. The number of TSCs near splice donor sites on the reverse strand was examined. The red and white bars indicate the number of 1-bp width TSCs and other TSCs, respectively. In *C. intestinalis*, many 1-bp width TSCs were found near splice donor sites on the reverse strand.

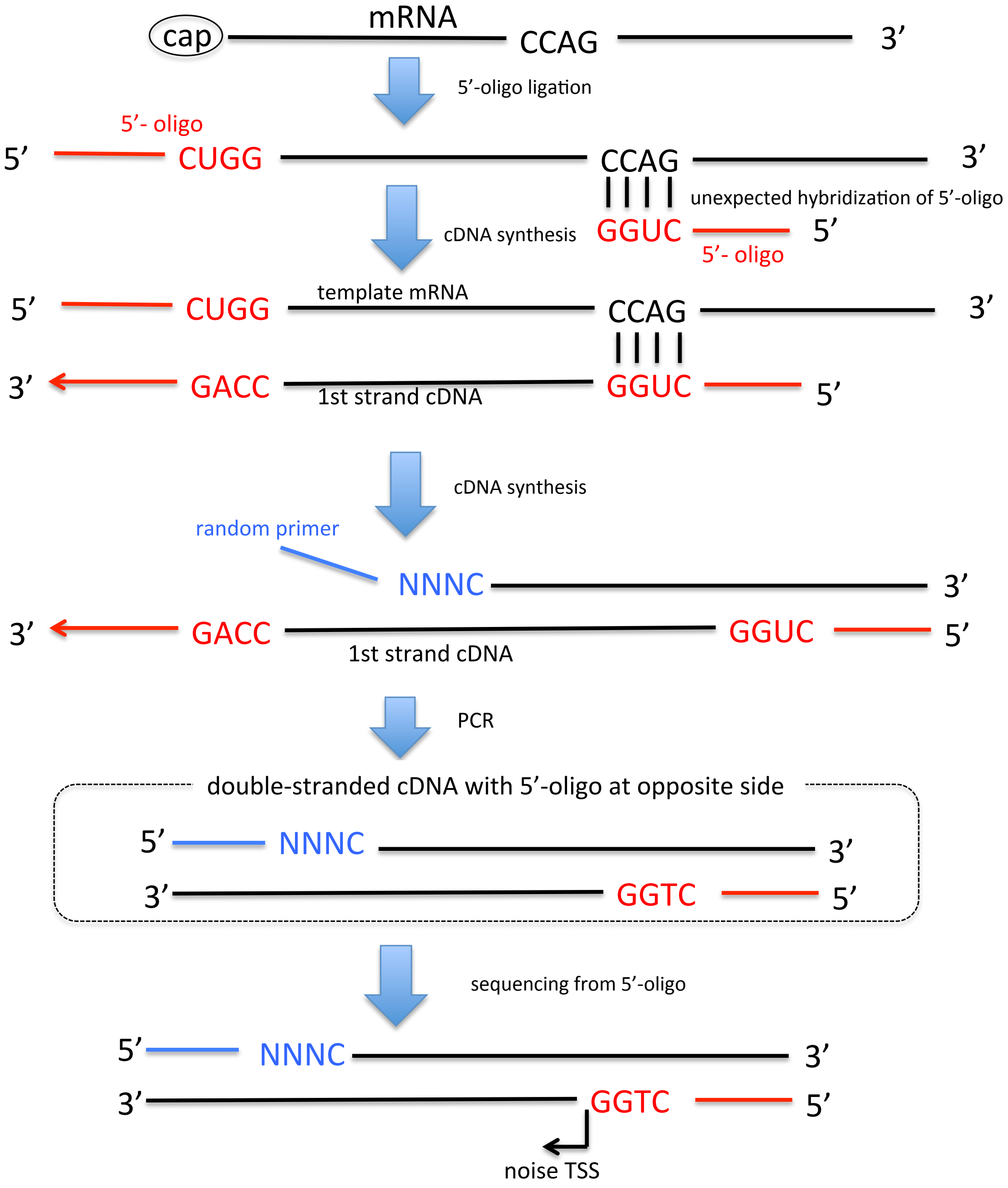


Figure S9: Possible mechanism of generation of CTGG TSCs. In 5′-oligo capping, 5′-oligo sequences (shown in red) can hybridized to CCAG on mRNAs. After cDNA synthesis, this unexpected hybridization generates the double-stranded cDNA that has the 5′-oligo sequence on the opposite side of the true 5′ end. The subsequent sequencing can produce reads from the opposite side. The reads sequenced from the erroneous oligo are mapped to the antisense strand of the mRNA. The 4 nt immediately upstream of the mapped position should be CTGG.

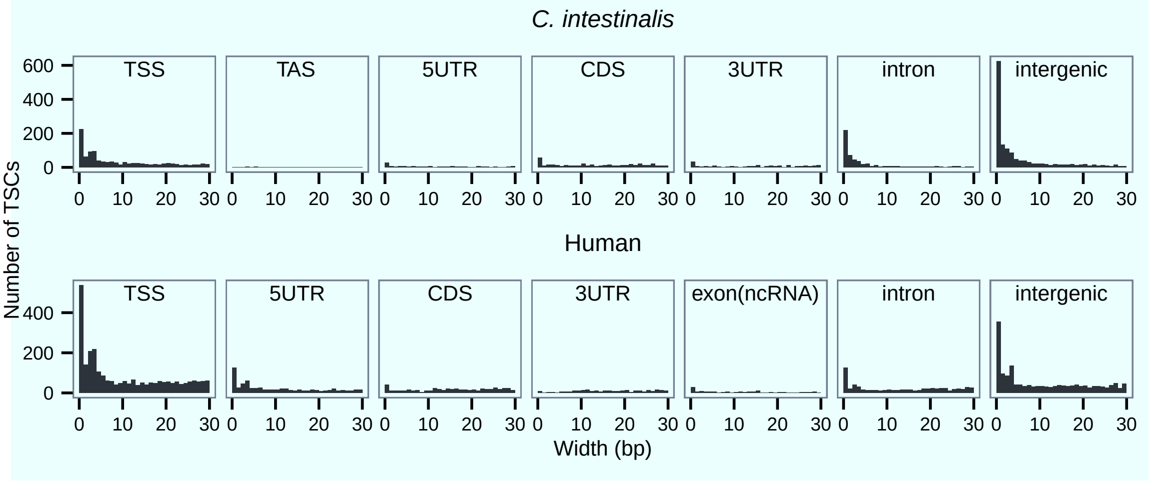


Figure S10: Width of TSCs after removing CTGG TSCs. The x and y axes represent the width of TSCs and the number of TSCs, respectively. Only the TSCs that were less than 30 bp in width are shown.

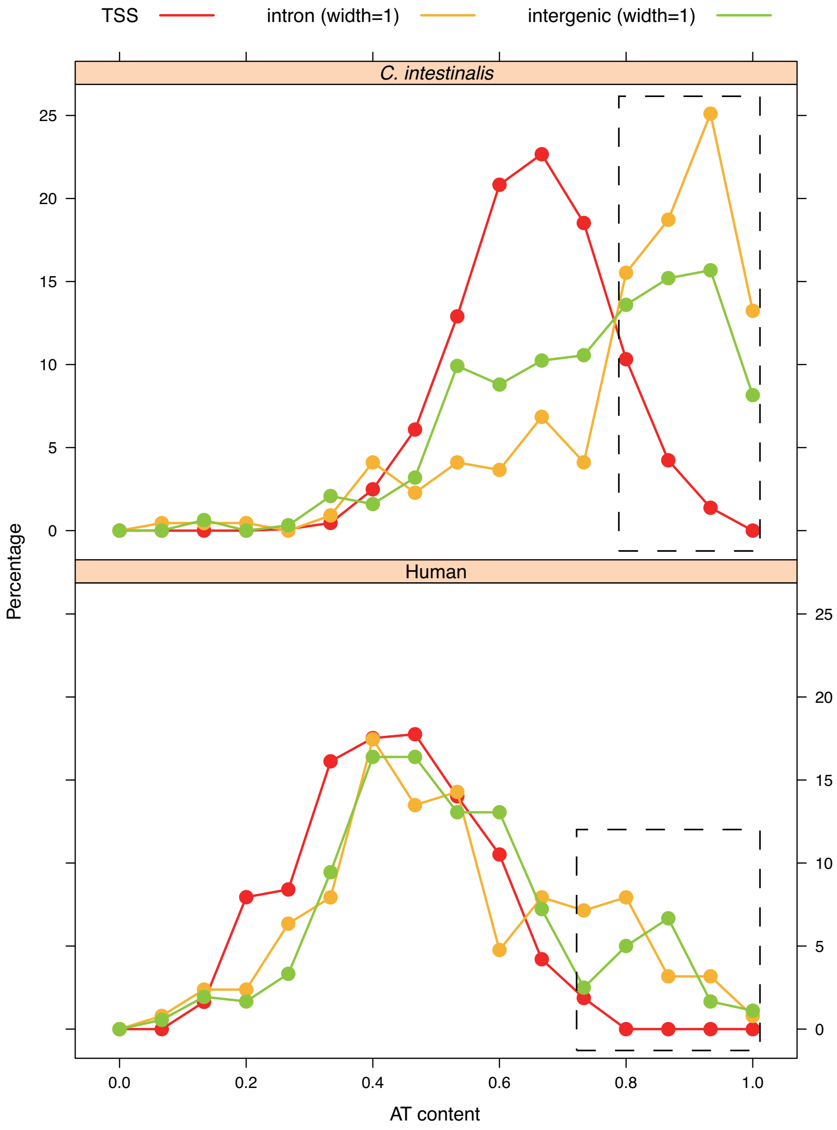


Figure S11: A+T content of the region 15 bp downstream of 1-bp width TSCs in introns and intergenic regions. We examined the A+T content of regions 15 bp downstream (+1 to +15), where position 0 represents the representative TSS. The A+T content of the 1-bp width TSCs located in introns and intergenic regions was compared to that of the TSCs located at known TSSs. The x and y axes represent the A+T content and the percentage of the TSCs with a given A+T content, respectively. We found many A+T-rich TSCs with ≥ 0.80 and ≥ 0.66 A+T content in *C. intestinalis* and human, respectively.

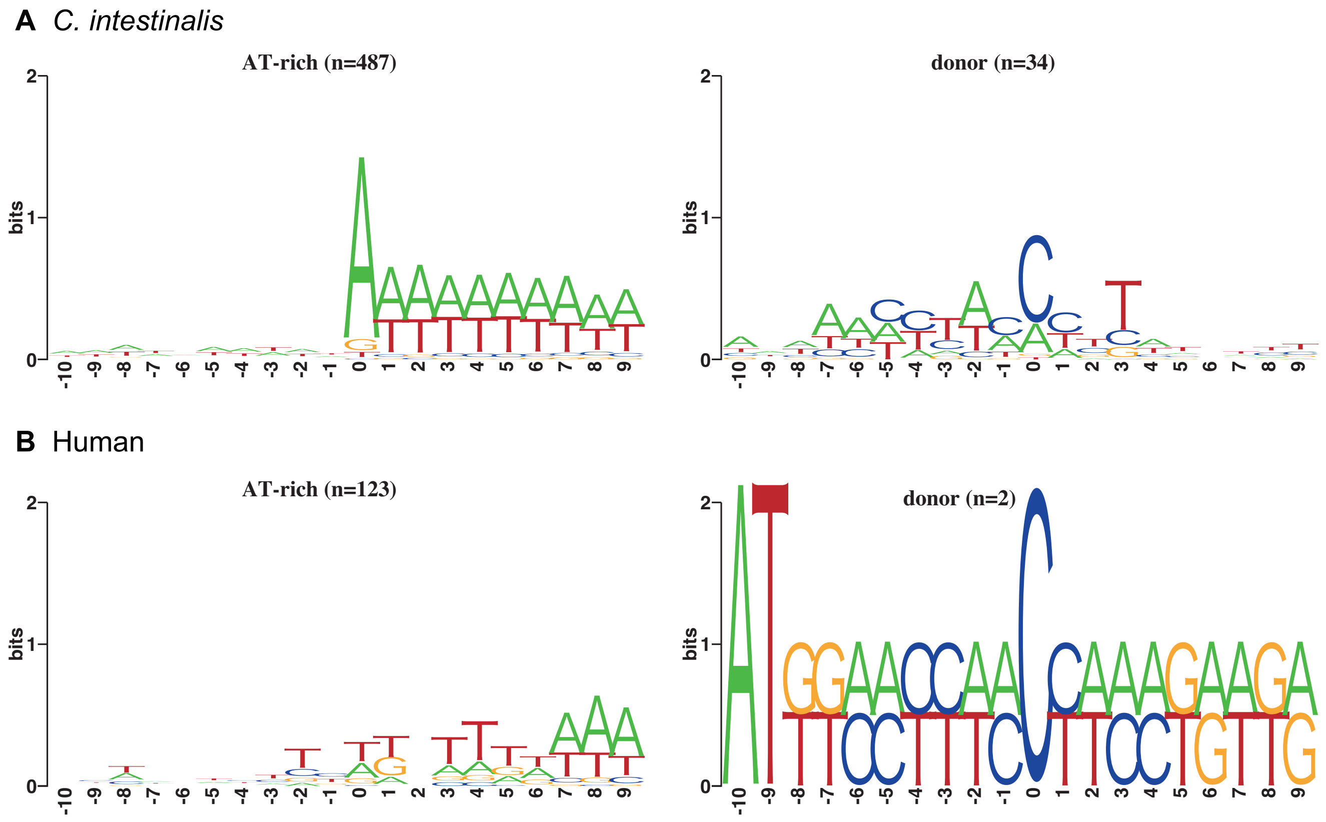


Figure S12: Sequence logos of two types of 1-bp width TSCs. "A+T-rich" represents the 1-bp width TSCs with ≥ 0.80 and ≥ 0.66 A+T content in 15-nt downstream regions in *C. intestinalis* (A) and human (B), respectively. "donor" represents the 1-bp width TSCs within -3 to $+2$ relative to splice donor sites on the reverse strand. The x-axis represents the distance relative to the representative TSS. The number in parentheses represents the number of TSCs.

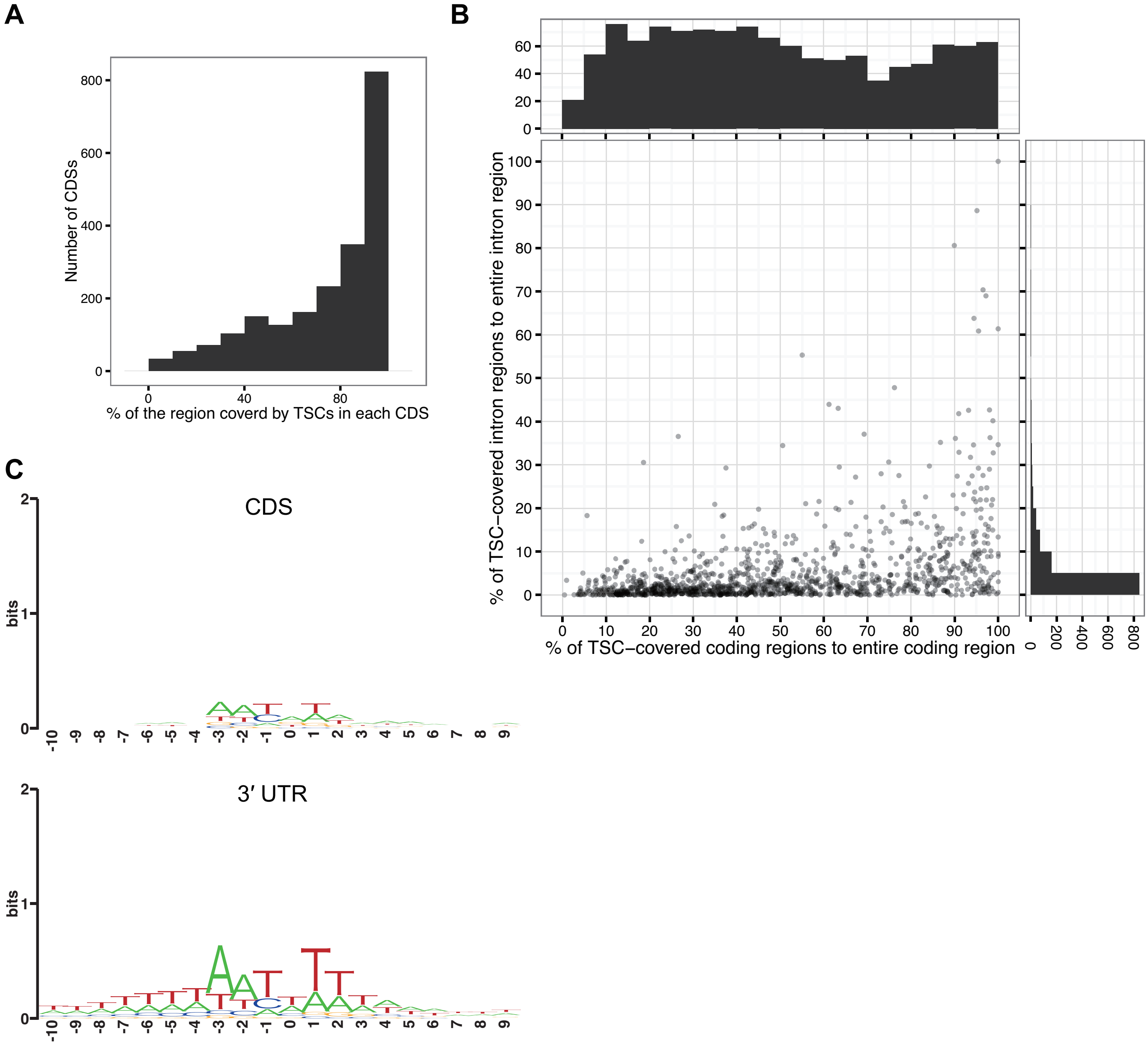


Figure S13: Analysis of TSCs in coding DNA sequences (CDSs) and 3′ untranslated regions (UTRs) in *C. intestinalis.* (A) Proportion of the region covered by TSCs. For each CDS in which at least one TSC was located, we examined the proportion of the region covered by TSCs. In about 80% of the CDSs, more than half of the region was covered by TSCs. (B) Proportion of the TSC-covered region relative to the entire coding region and the entire intron. For each transcript model with at least one TSC in its coding regions, we calculated the proportion of the TSC-covered region relative to the entire coding region and the entire intron. Each dot represents the transcript model with at least one TSC in the coding regions. Because there can be multiple transcript models with the same entire coding region in a gene locus, the transcript model with the highest proportion of the TSC-covered region was selected as the representative transcript model for each gene locus. The transcript models with no introns were not included in this figure. We found many transcript models in which the entire coding region was well covered by TSCs while the entire intron was not covered. (C) Sequence logos of TSCs in CDSs and 3′ UTRs. The x-axis represents the distance relative to the representative TSS.

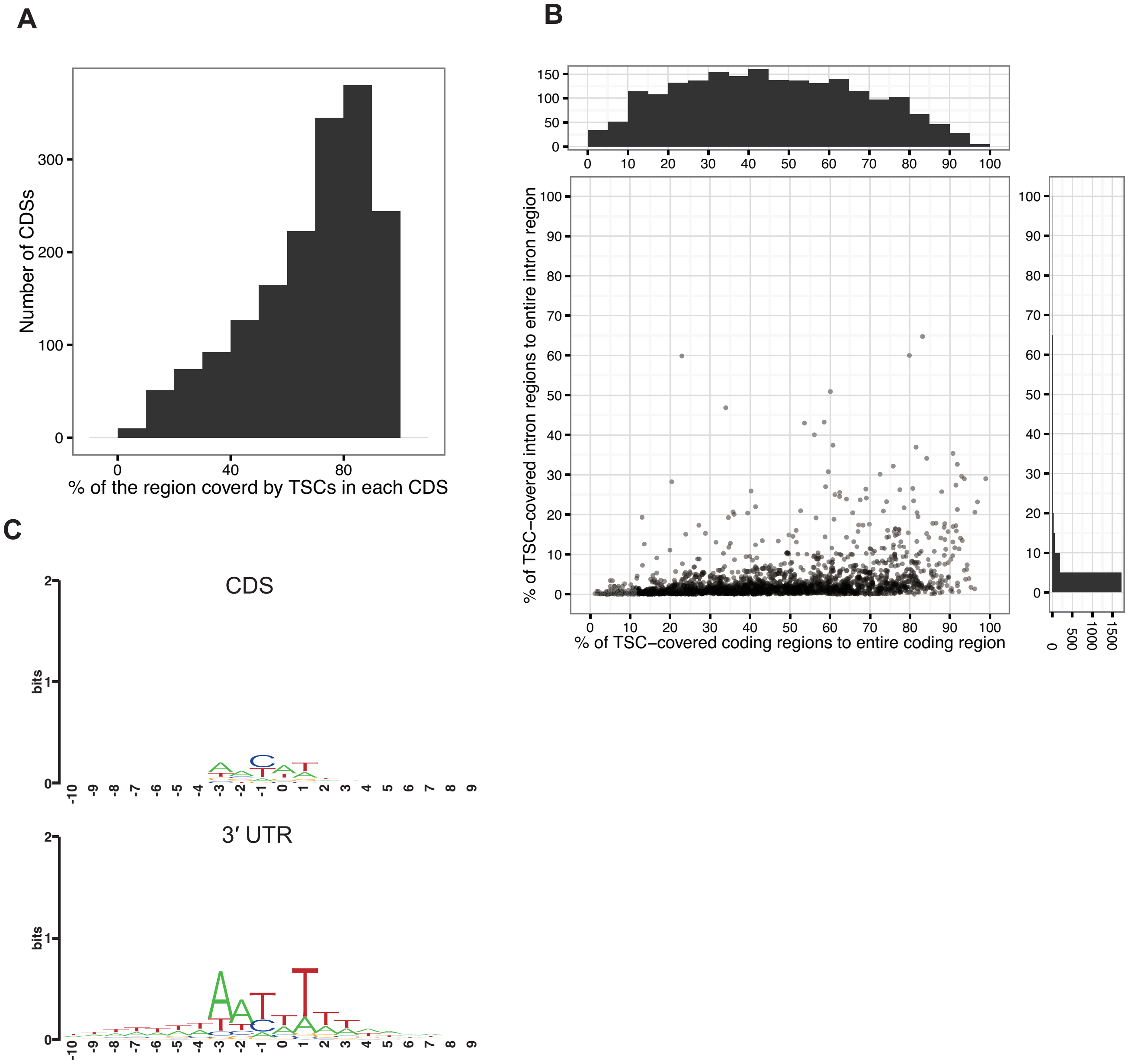


Figure S14: Analysis of TSCs in coding DNA sequences (CDSs) and 3′ untranslated regions (UTRs) in human. (A) Proportion of the region covered by TSCs. For each CDS in which at least one TSC was located, we examined the proportion of the region covered by TSCs. (B) Proportion of the TSC-covered region relative to the entire coding region and the entire intron. For each transcript model with at least one TSC in its coding regions, we calculated the proportion of the TSC-covered region relative to the entire coding region and the entire intron. Each dot represents the transcript model with at least one TSC in its coding regions. Because there can be multiple transcript models with the same entire coding region in a gene locus, the transcript model with the highest proportion of the TSC-covered region was selected as the representative transcript model for each gene locus. The transcript models with no introns were not included in this figure. (C) Sequence logos of TSCs in CDSs and 3′ UTRs. The x-axis represents the distance relative to the representative TSS.

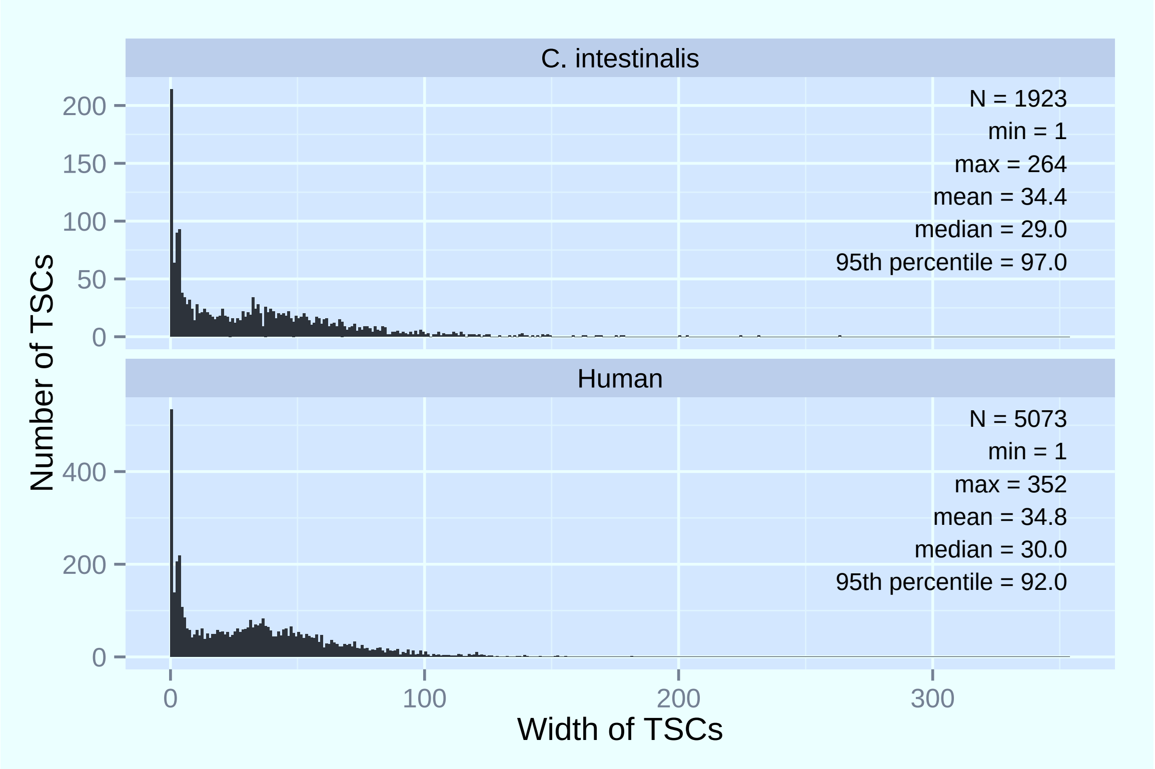


Figure S15: Distribution of width of TSCs located at known TSSs. Distance from the 5th percentile to the 95th percentile of the TSS distribution was used as a measure of width of TSCs. N represents the number of TSCs located at known TSSs.

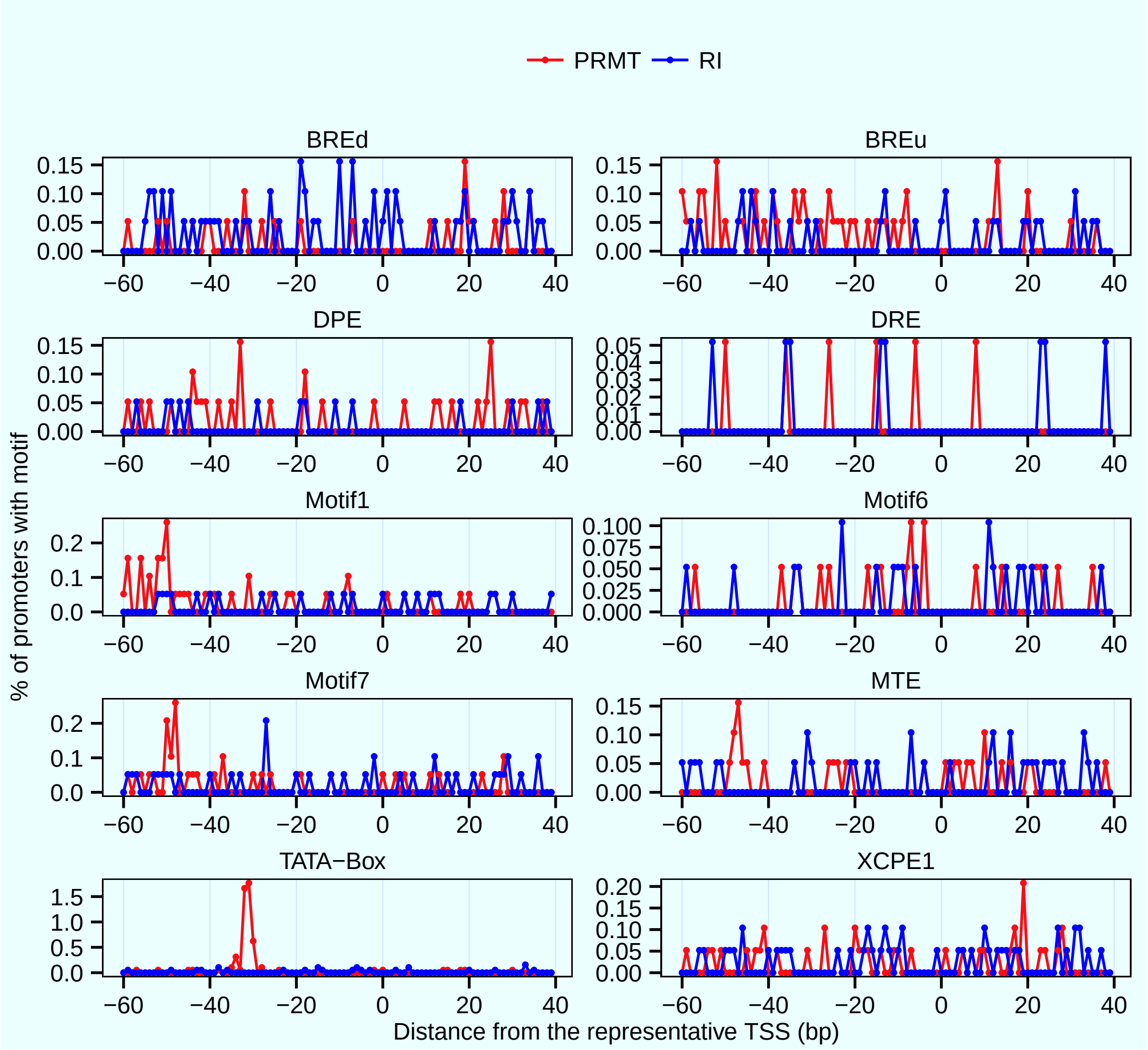


Figure S16: Distributions of core promoter elements. Core promoter elements were searched in *C. intestinalis* core promoter regions (−60 to +39) using FIMO (Grant et al., 2011) with the default threshold. The position weight matrix (PWM) of the core promoter elements (TATA-box, BRE^{d}, BRE^{u}, DPE, DCE S$_{I}$, DCE S$_{II}$, DCE S$_{III}$, MTE, XCPE1) were obtained from JASPAR database (Sandelin et al., 2004). The PWM of the DRE and motif 1, 6 and 7 were obtained from (Ni et al., 2010). The distribution of each core promoter element in random intergenic regions of the same number and length as the core promoter regions are also shown (blue line). The results for the three DCEs were not shown because they were not predicted in the core promoter sequences. The x and y axes represent the distance from the representative TSS and the percentage of promoters with predicted elements at a given position, respectively. PRMT, promoter regions; RI, random intergenic regions.

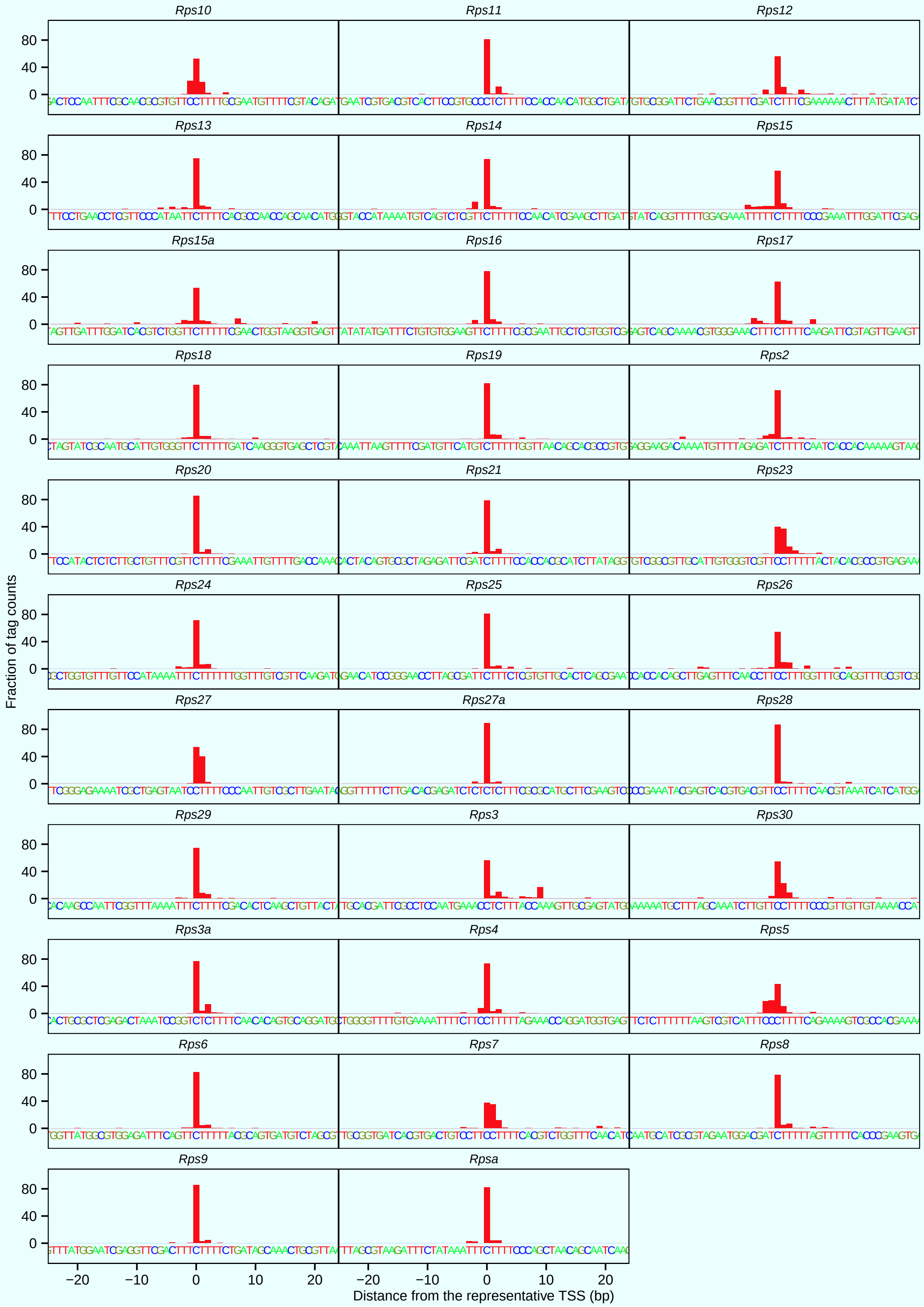


Figure S17: TSS distributions of small subunit RP genes. Each panel shows the TSS distribution of each small subunit RP gene. The nucleotides at the corresponding positions including polypyrimidine tracts are also shown.

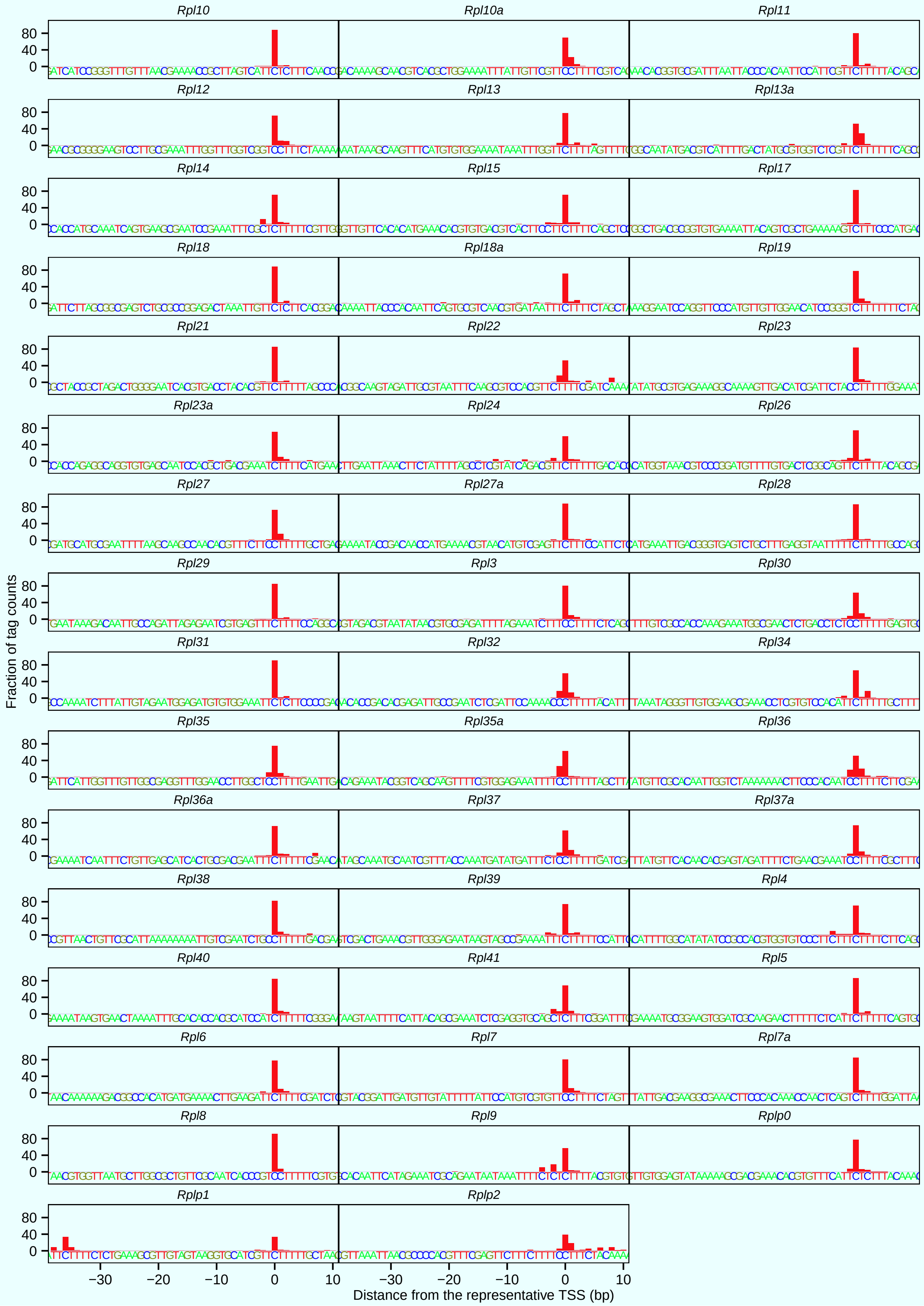


Figure S18: TSS distributions of large subunit RP genes. Each panel shows the TSS distribution of each large subunit RP gene. The nucleotides at the corresponding positions including polypyrimidine tracts are also shown.

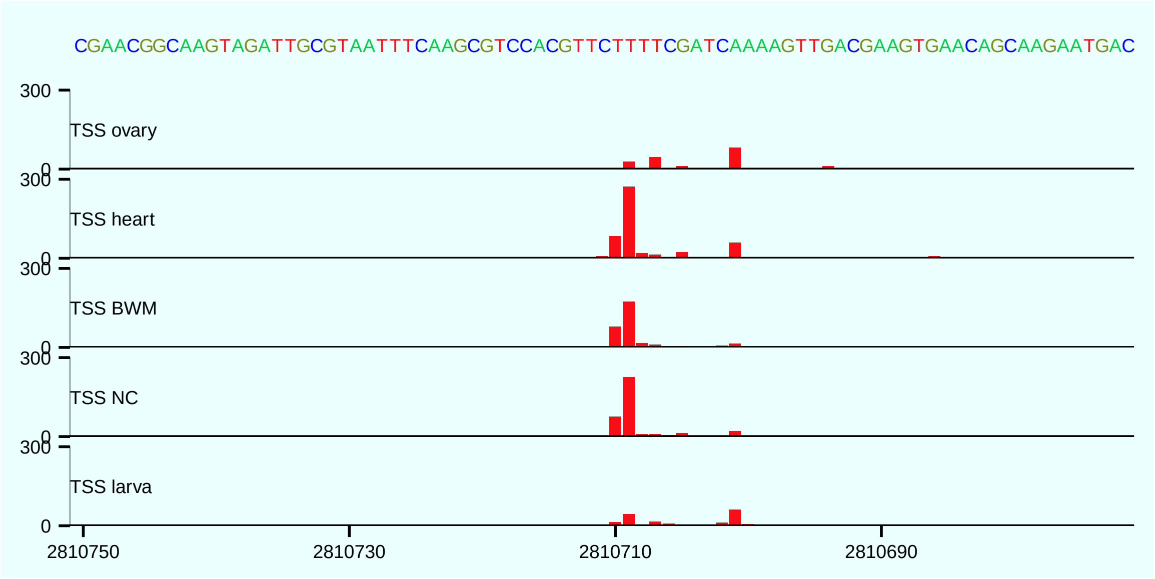


Figure S19: TSS of the ribosomal protein L22 gene. The red bars represent the TSSs. The y-axis represents the number of tags. BWM, body wall muscle; NC, neural complex.

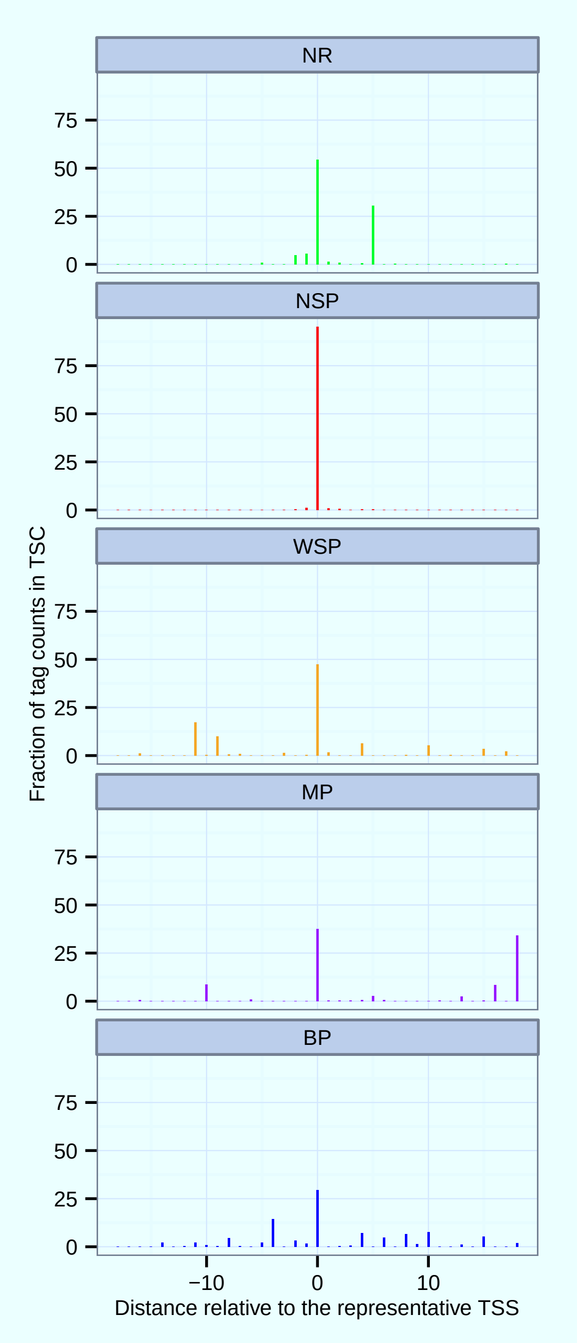


Figure S20: Examples of TSS distribution types. Promoters were classified into five types (NR, NSP, WSP, MP, and BP) based on their TSS distribution. Each panel shows an example of each TSS distribution type. The NSP and BP promoters were referred to as "sharp-type" and "broad-type" promoters, respectively. The other promoters were integrated and referred to as "other" promoters. The sharp-type promoter is a promoter with a sharp TSS distribution where transcription starts within a narrow region. On the other hand, the broad-type promoter is a promoter with a broad TSS distribution where transcription starts in a wide range and there is no clear sharp peak.

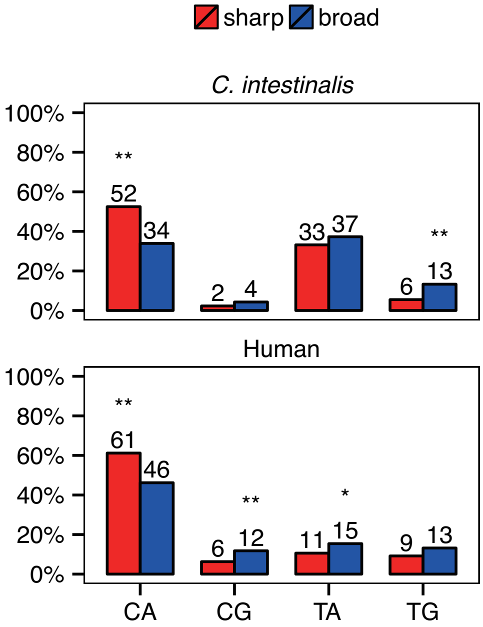


Figure S21: Difference in TSS usage between sharp and broad promoters. The use of four pyrimidine-purine (PyPu) dinucleotides was compared in sharp and broad promoters. The red and blue bars represent the usage (%) of each PyPu in sharp and broad promoters, respectively. The difference was evaluated by Fisher's exact test. The asterisks (* and **) denote Bonferroni-corrected $P < 0.05$ and $P < 0.01$, respectively.

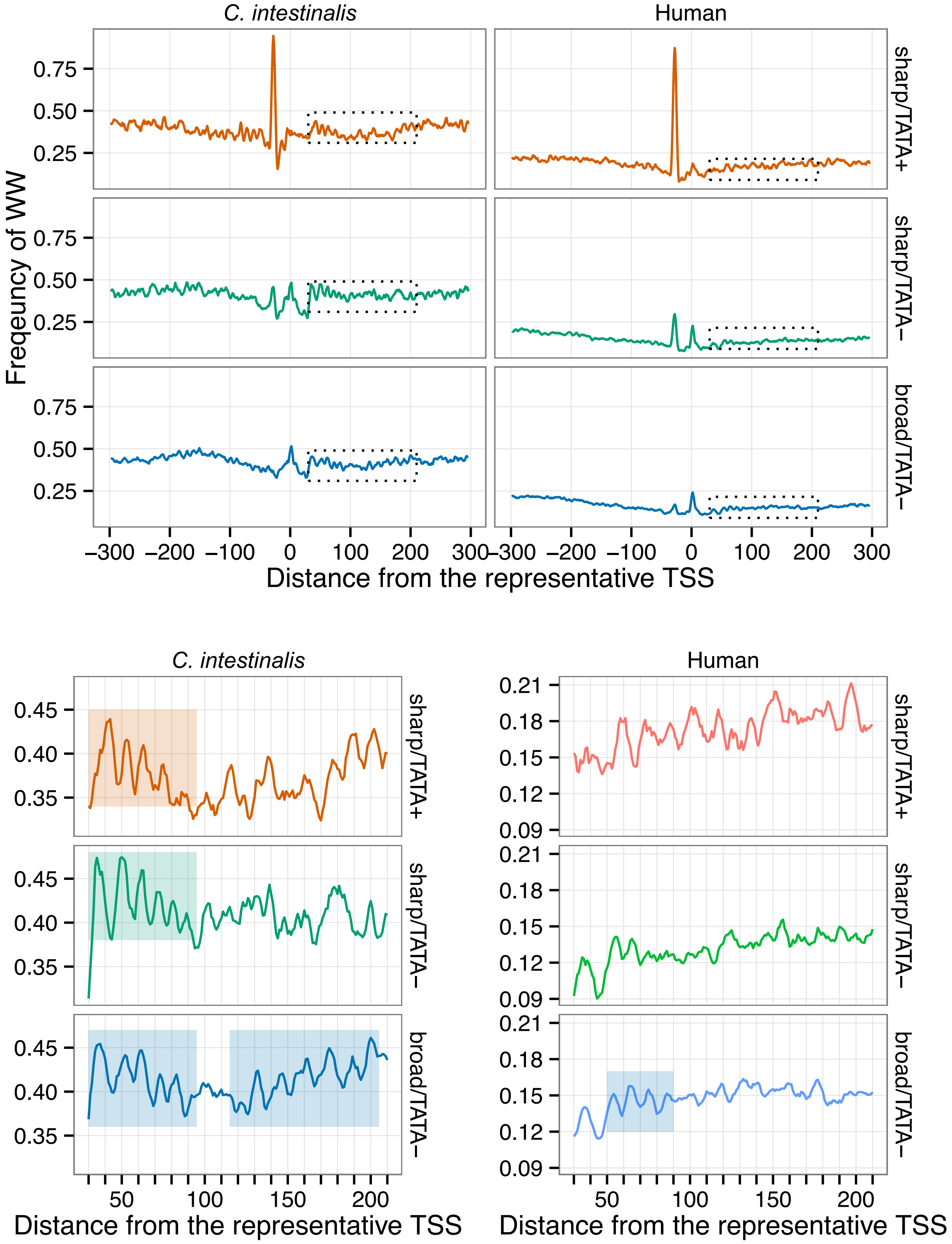


Figure S22: Frequency distribution of WW motifs. Top figures (A) show the frequency of WW (where W is A or T) motifs in each type of promoters (sharp/TATA+, sharp/TATA−, and broad/TATA−). Each distribution was smoothed by 5-point sliding average. Bottom figures (B) show the magnified view of the regions between +30 to +210 (dotted rectangles in top figures). The smooth distribution with about 10-bp periodicity is shown in corresponding semi-transparent colors.

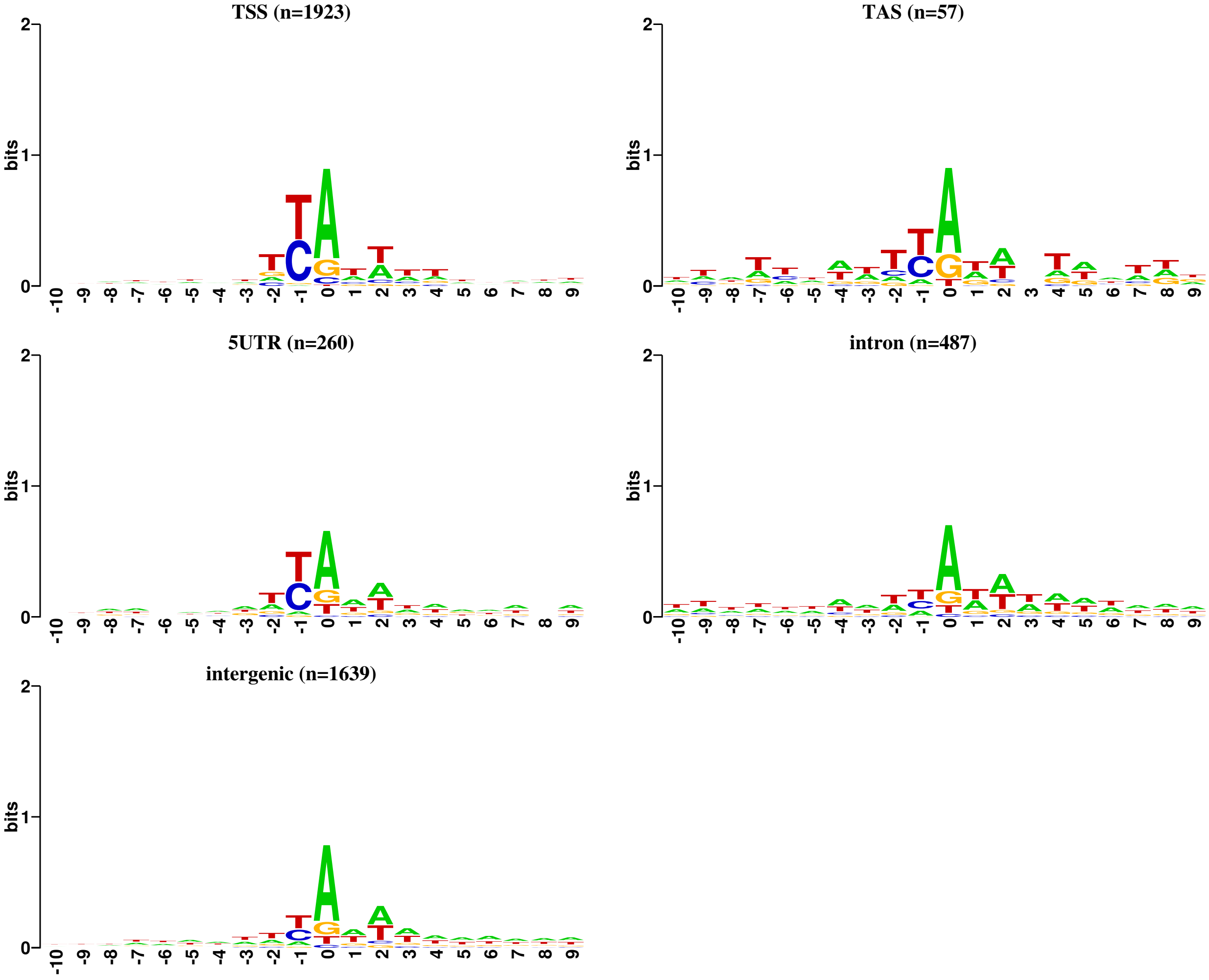


Figure S23: Initiator motifs. The initiator motif of TSCs in each location is shown. The x-axis represents the distance from the representative TSS. The number in parentheses represents the number of TSCs in each location.

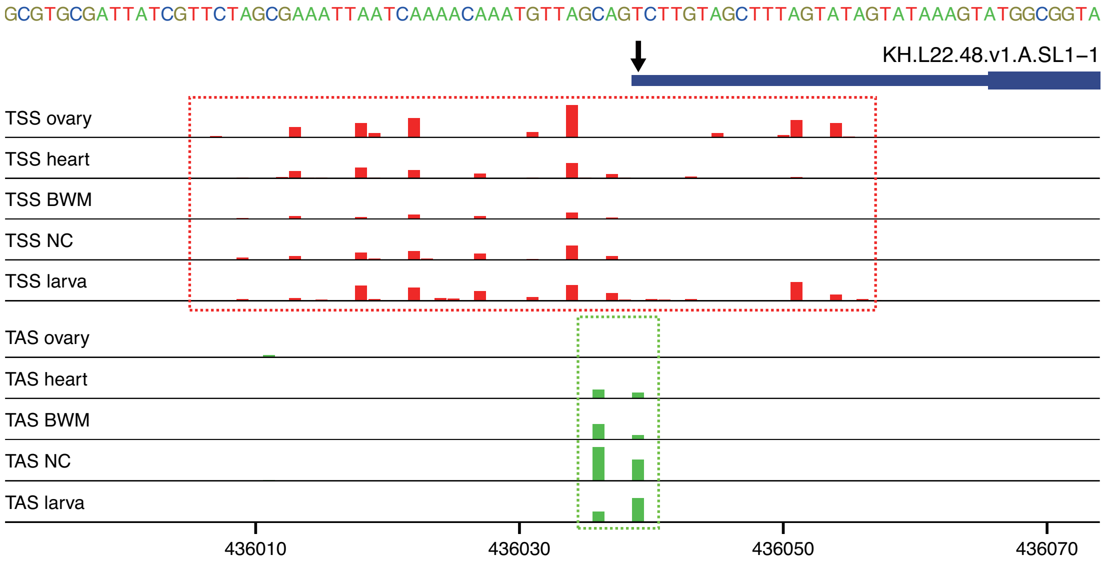


Figure S24: Example of a TSS cluster overlapping with *trans*-splice acceptor sites (TASs). The figure shows the TSS cluster (TSC) and TAS cluster (TAC) that overlap with the annotated TAS of the transcript (KH.L22.48.v1.A.SL1-1) encoding a protein similar to piggyBac transposable element derived 4. The TSC and TAC are marked by red and green dotted rectangles, respectively. The distribution of TSSs and TASs is shown by red and green bars, respectively. The arrow indicates the annotated TAS of the transcript. BWM, body wall muscle; NC, neural complex.

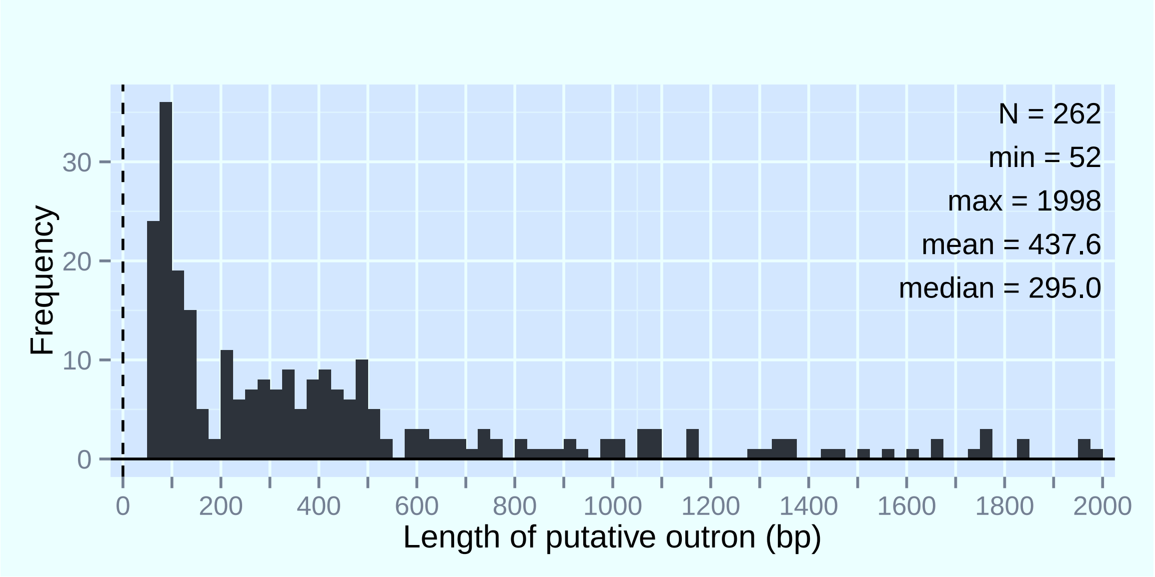


Figure S25: Distribution of putative outron lengths. For each pair of clusters [*trans*-splice acceptor site cluster (TAC) and upstream TSC], we examined the distance between the clusters, which indicates the putative outron length. N represents the number of pairs of clusters.

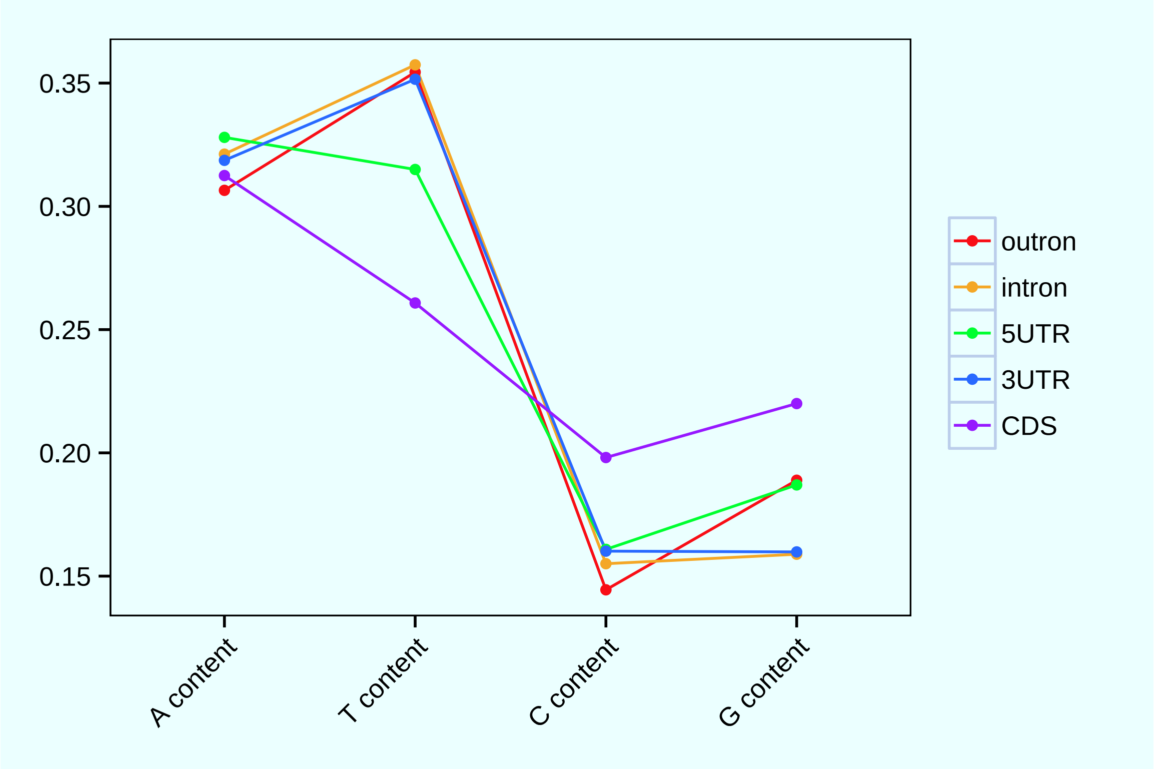


Figure S26: A, C, G, and T content of putative outrons. The A, C, G, and T content were calculated for the following five classes: putative outrons, introns, $5'$ UTRs, CDSs, and $3'$ UTRs. The x-axis shows the median of A, C, G, and T content in each class. Considering the range of lengths of putative outrons (51 to 2000 nt), only introns, $5'$ UTRs, CDSs, and $3'$ UTRs of length 51 nt to 2000 nt were used to create this figure.

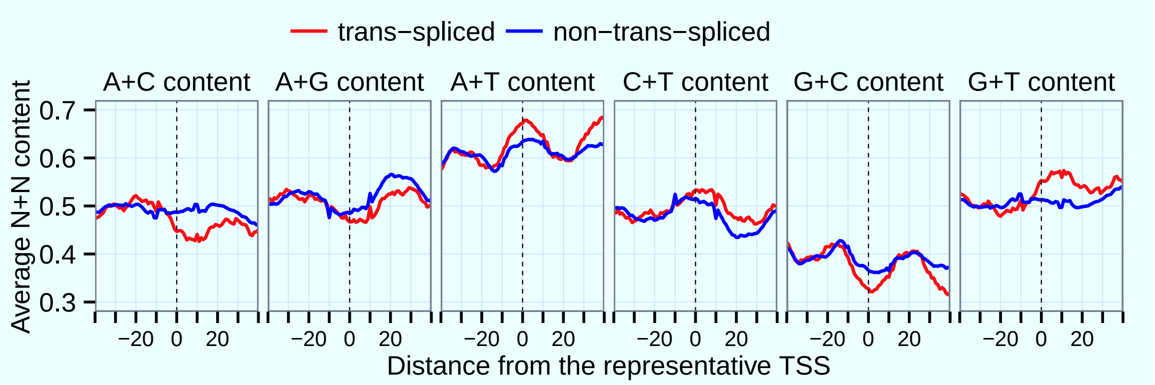


Figure S27: Distribution of N_1+N_2 content in non-RP *trans*-spliced gene promoters and non-RP non-*trans*-spliced gene promoters. Average N_1+N_2 content was calculated using a 20-bp sliding window in non-RP *trans*-spliced gene promoters and non-RP non-*trans*-spliced gene promoters, where N_1 and N_2 are different nucleotides.

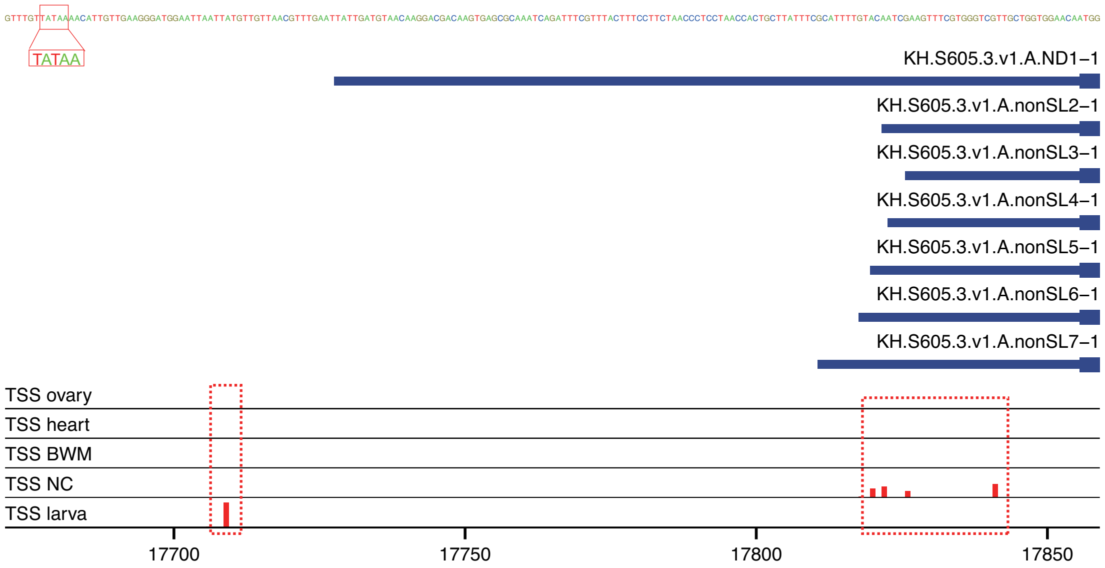


Figure S28: Alternative promoters of the betagamma crystallin gene. TSCs are marked by red dotted rectangles, and their TSS distribution is shown by red bars. BWM, body wall muscle; NC, neural complex.

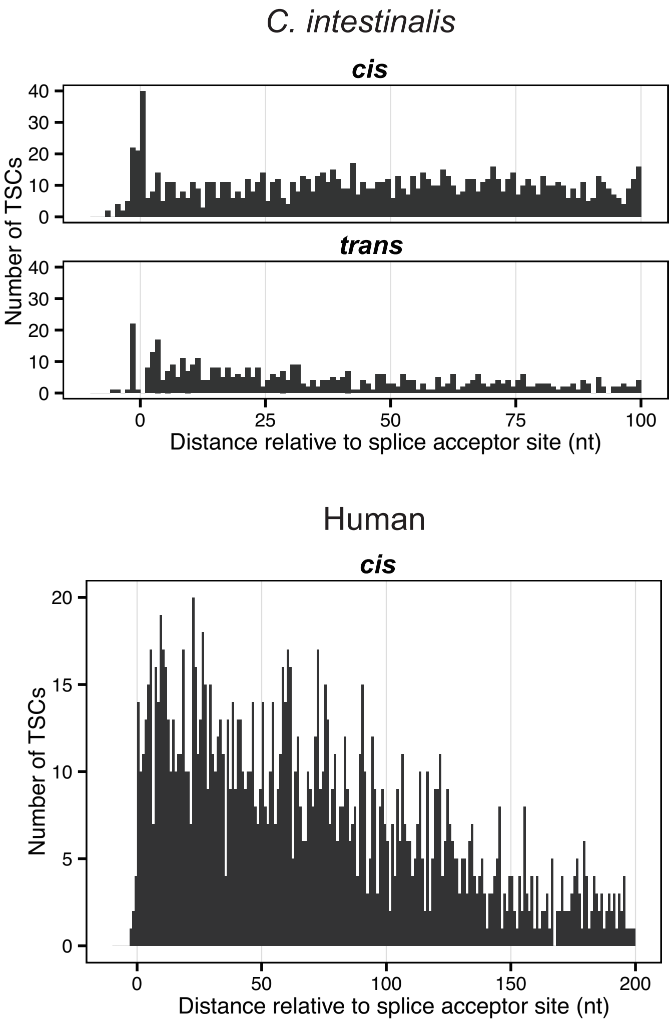


Figure S29: Peak positions of TSCs in exons. For each TSC located in exons, we examined the position of its peak (that is, the position with the highest number of mapped tags). The x and y axes represent the distance relative to *cis*- or *trans*-splice acceptor sites and the number of TSCs, respectively.

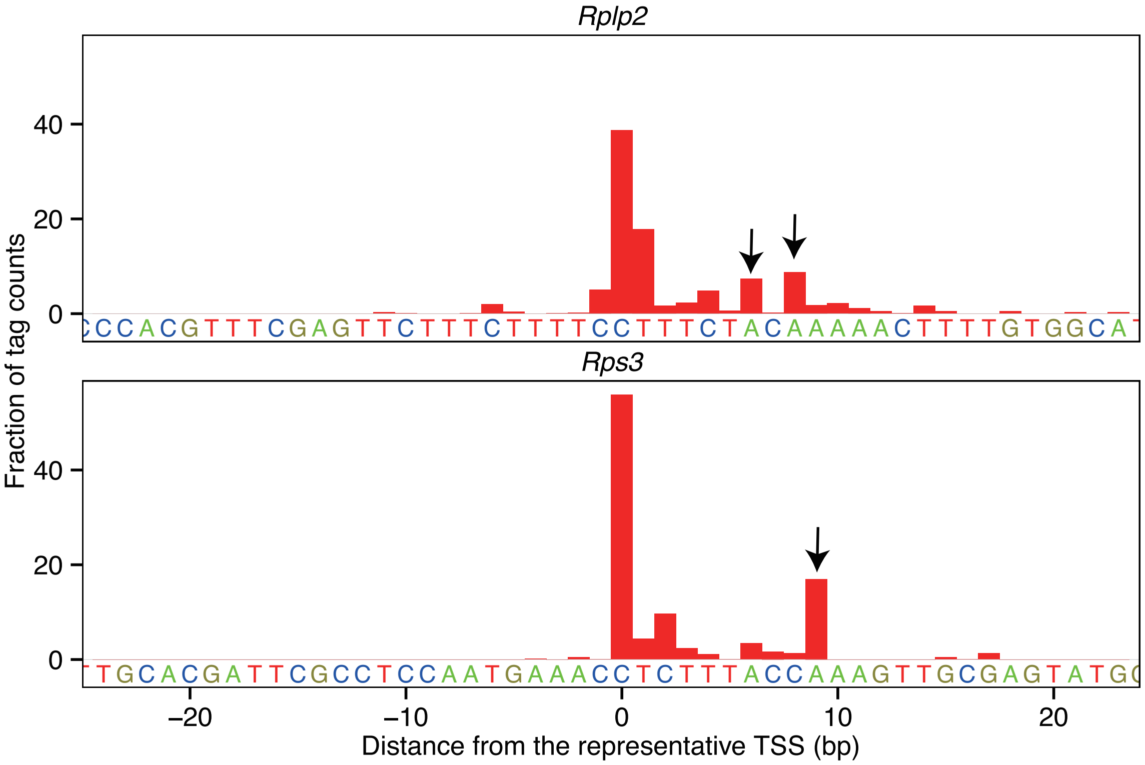


Figure S30: Examples of minor TSSs of RP genes. The figure shows the TSS distribution of the *Rplp2* and the *Rps3* gene. Examples of minor TSSs near the downstream of polypyrimidine tracts are marked by black arrows.

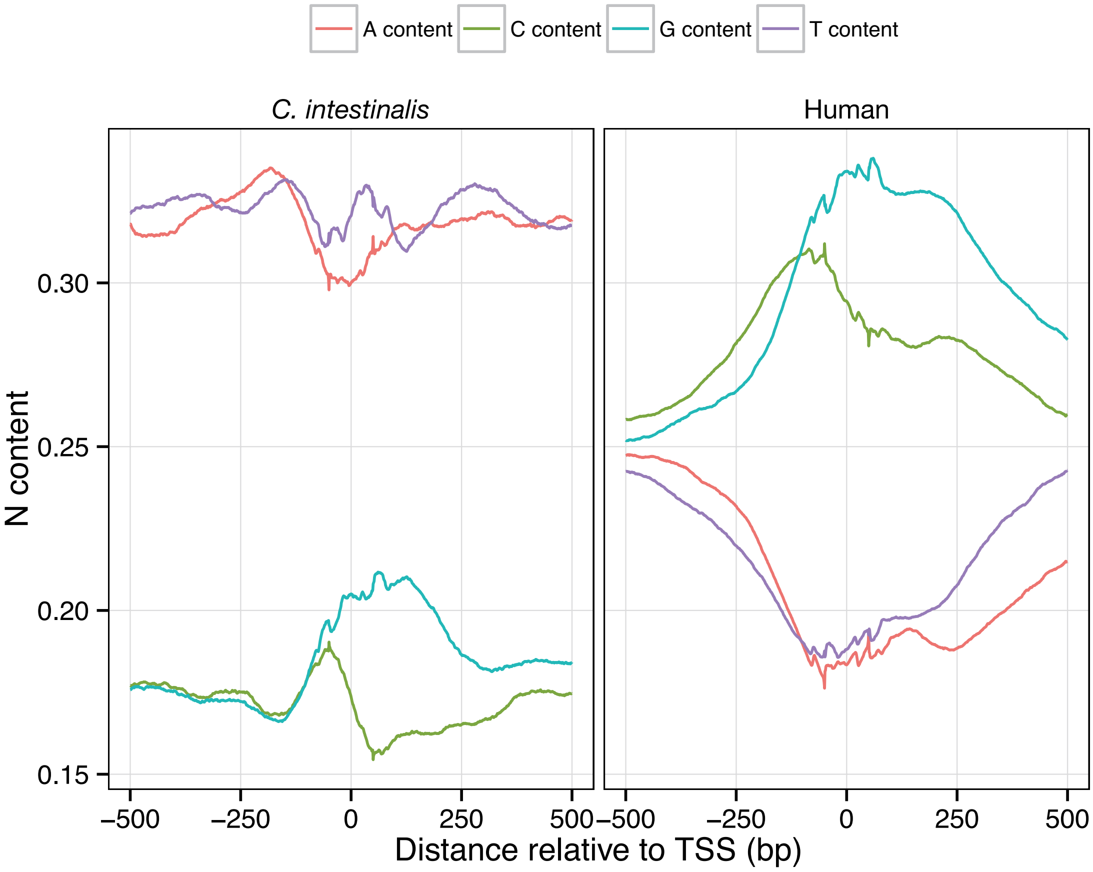


Figure S31: Distribution of A, C, G, and T content. The distribution of A, C, G, and T content was examined in non-RP gene promoters using a 100-bp sliding window. The N content was defined as the number of N in the window divided by 100, where N is A, C, G, or T. The x and y axes represent the distance relative to the representative TSS and average N content, respectively.

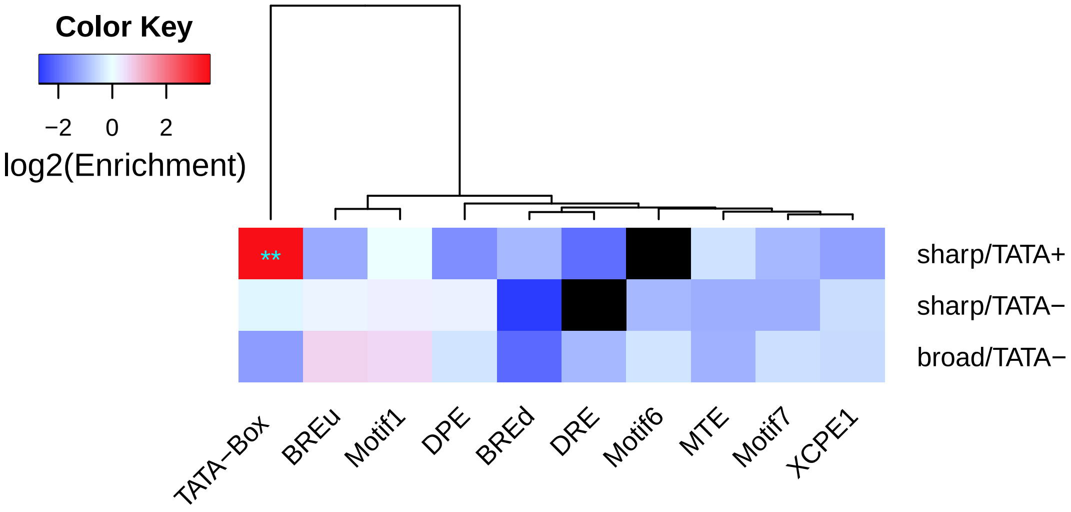
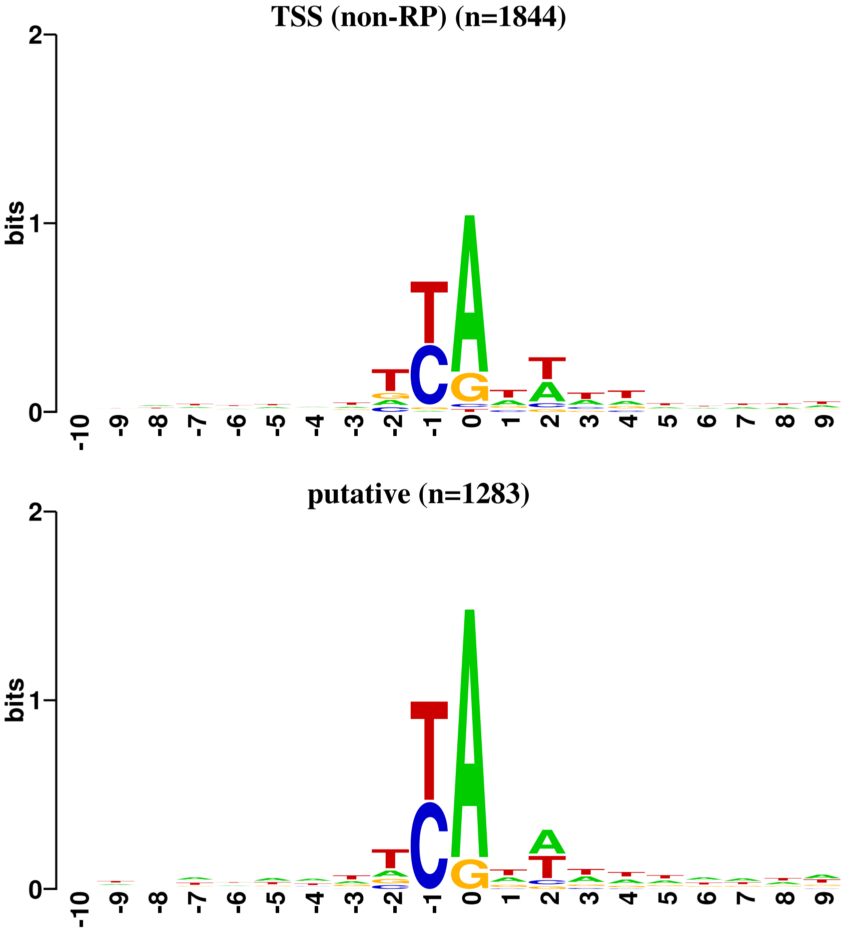


Figure S32: Enrichment of RNA polymerase II promoter elements. In each promoter class, we examined the enrichment of known polymerase II promoter elements and *Drosophila* promoter elements (DRE, motif 1, 6 and 7) in core promoter regions (−60 to +39). Promoter elements associated with polymerase II promoter in the JASPAR database (TATA-box, BRE^d, BRE^u, DPE, DCE S$_I$, DCE S$_{II}$, DCE S$_{III}$, MTE, MED-1, XCPE1, GC-box, CCAAT-box) (Sandelin et al., 2004) were used as known polymerase II promoter elements. The PWM of the DRE, motif 1, 6 and 7 were obtained from (Ni et al., 2010). The presence of each element in the core promoter regions was predicted using FIMO (Grant et al., 2011) with the default threshold. The enrichment of each element was defined as the number of core promoter sequences with the element divided by the total number of core promoter sequences. To evaluate background enrichment of each element, we created ten sets of random intergenic sequences of the same number and length as the core promoter regions. The enrichment of each element was calculated in each set of random intergenic sequences, and the max value of the enrichment values was used as the background enrichment of the element. The heatmap shows enrichment versus background of each element in each promoter class. We also examined the statistical difference of enrichment between core promoter regions and background by binomial test. The asterisks (* and **) denote Bonferroni-corrected $P < 0.05$ and $P < 0.01$, respectively. Black cells mean that elements were not predicted in the core promoter regions. The results for the three DCEs were not shown because they were not predicted in all the promoter classes.


TSS (non-RP) (n=1844)

putative (n=1283)

Figure S33: PyPu motifs of non-RP promoters and putative promoters. “TSS (non-RP)” and “putative” represent the TSCs at known TSSs of non-RP genes and the TSCs in regions other than known TSSs. The x-axis represents the distance from the representative TSS. The number in parentheses represents the number of TSCs.

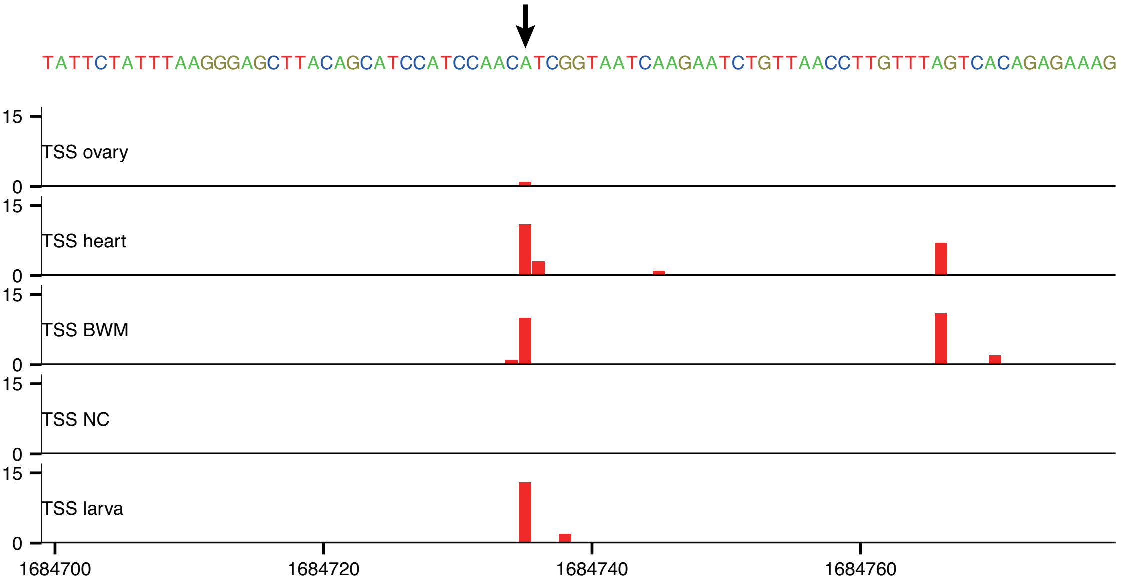


Figure S34: TSS of the troponin I gene. The TSS of the troponin I gene is indicated by an arrow. The red bars represent the TSSs identified by our data. The y axis represents the number of tags. BWM, body wall muscle; NC, neural complex.

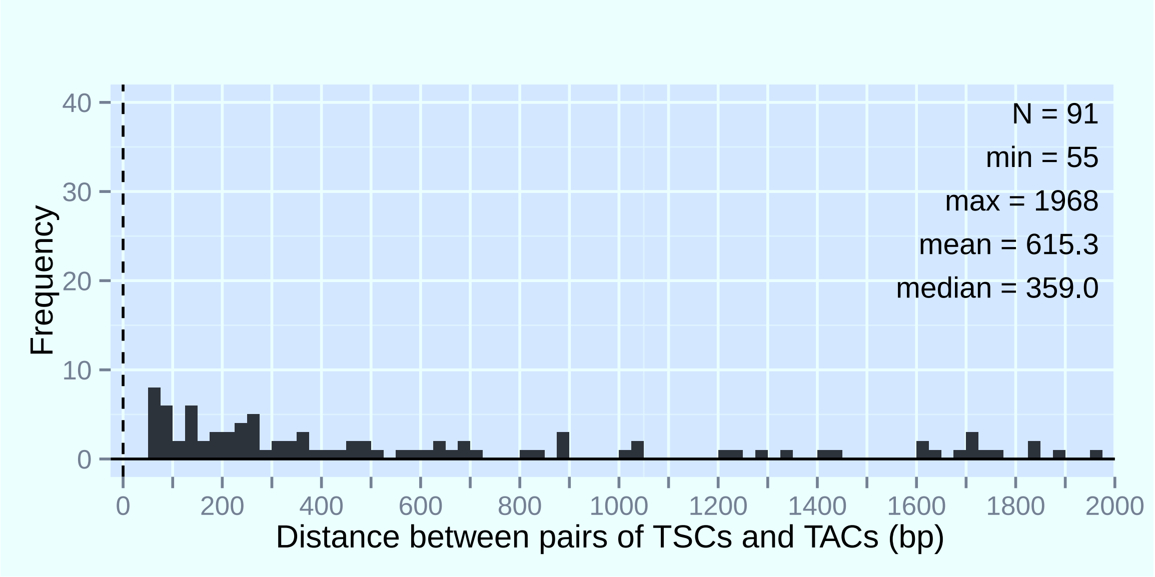


Figure S35: Distribution of distance between pairs of clusters that do not have the same expression pattern.

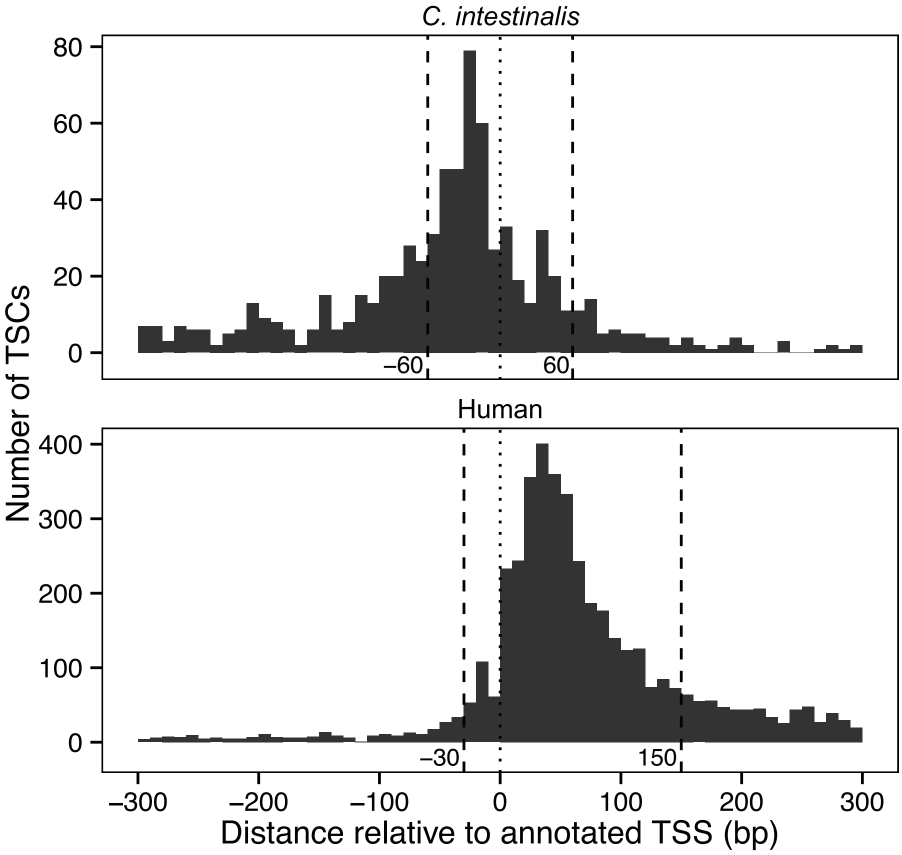


Figure S36: Distance relative to annotated TSSs. The locations of TSCs were determined based on the KH gene model of *C. intestinalis* and the RefSeq gene model of humans. For the TSCs located in 5′ UTRs of 1st exons and intergenic regions, we examined the distance relative to annotated TSSs. Only TSCs with ≥ 100 tags were used to create this figure. Many TSCs were located near annotated TSSs.

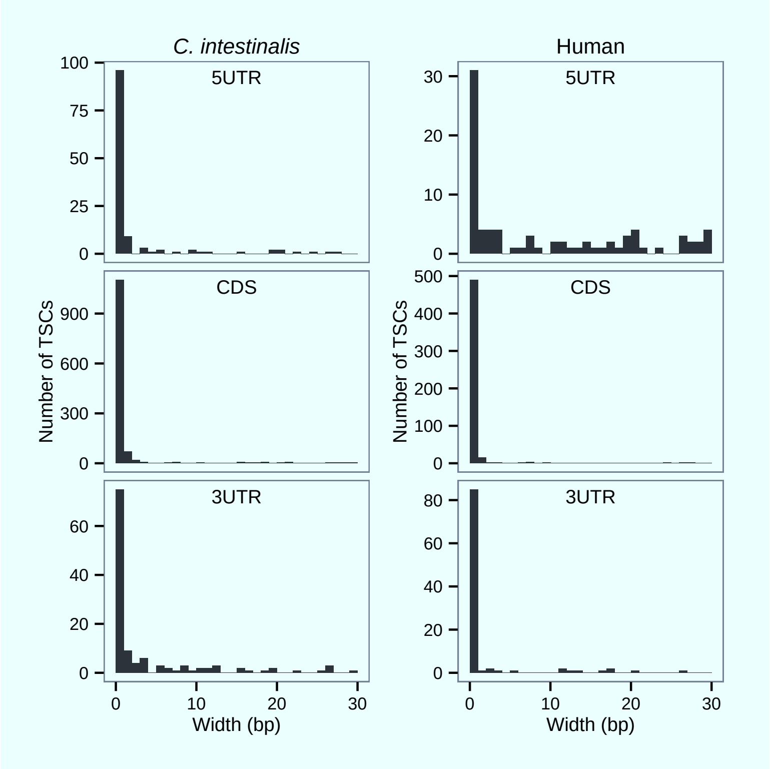


Figure S37: Width of TSCs on the antisense strand of exon regions. The TSCs were classified according to their location on the reverse strand. The x and y axes represent the width of TSCs and the number of TSCs, respectively. Only the TSCs on the antisense strand of exons that were less than 30 bp in width are shown. There were many 1-bp width TSCs on the antisense strand of exon regions.

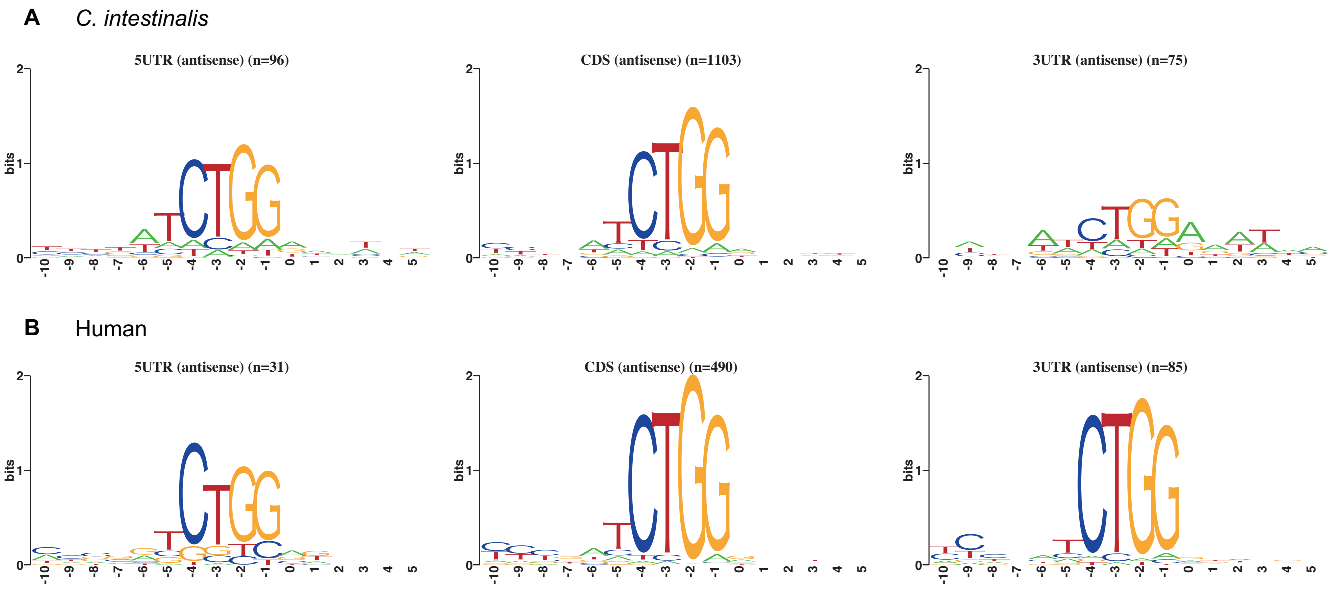


Figure S38: CTGG motifs of 1-bp width TSCs on the antisense strand of 5′ UTRs, CDSs, and 3′ UTRs. The x axis represents the position relative to the representative TSS. The number in parentheses represents the number of TSCs.

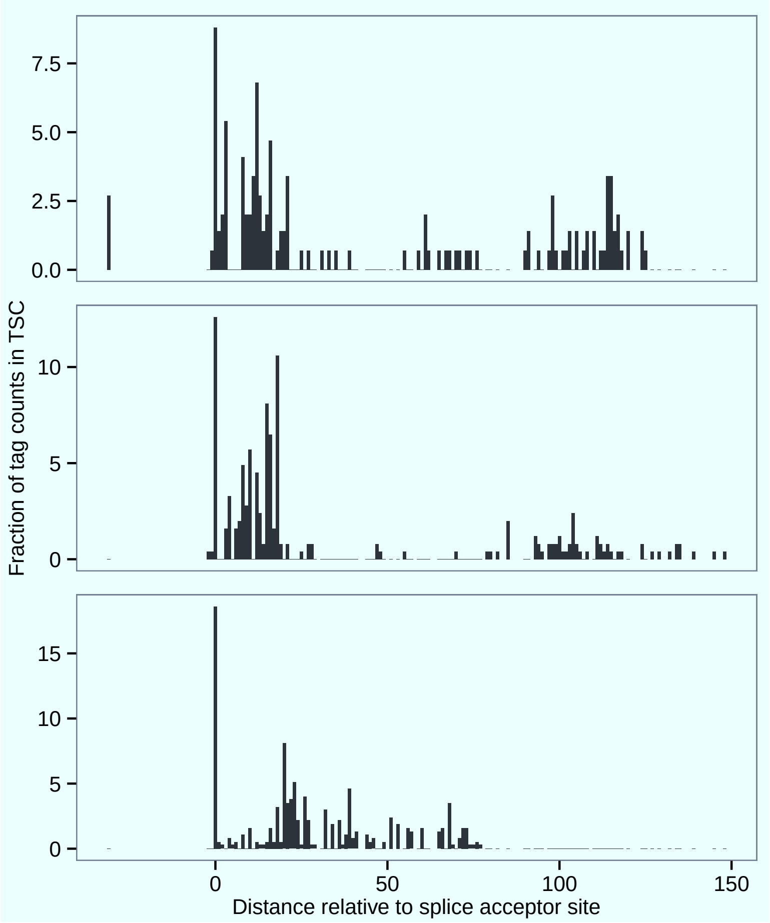


Figure S39: Examples of right-skewed TSCs with peaks near splice acceptor sites. The x and y axes represent the distance relative to splice acceptor sites and the fraction of tags in the TSC, respectively.

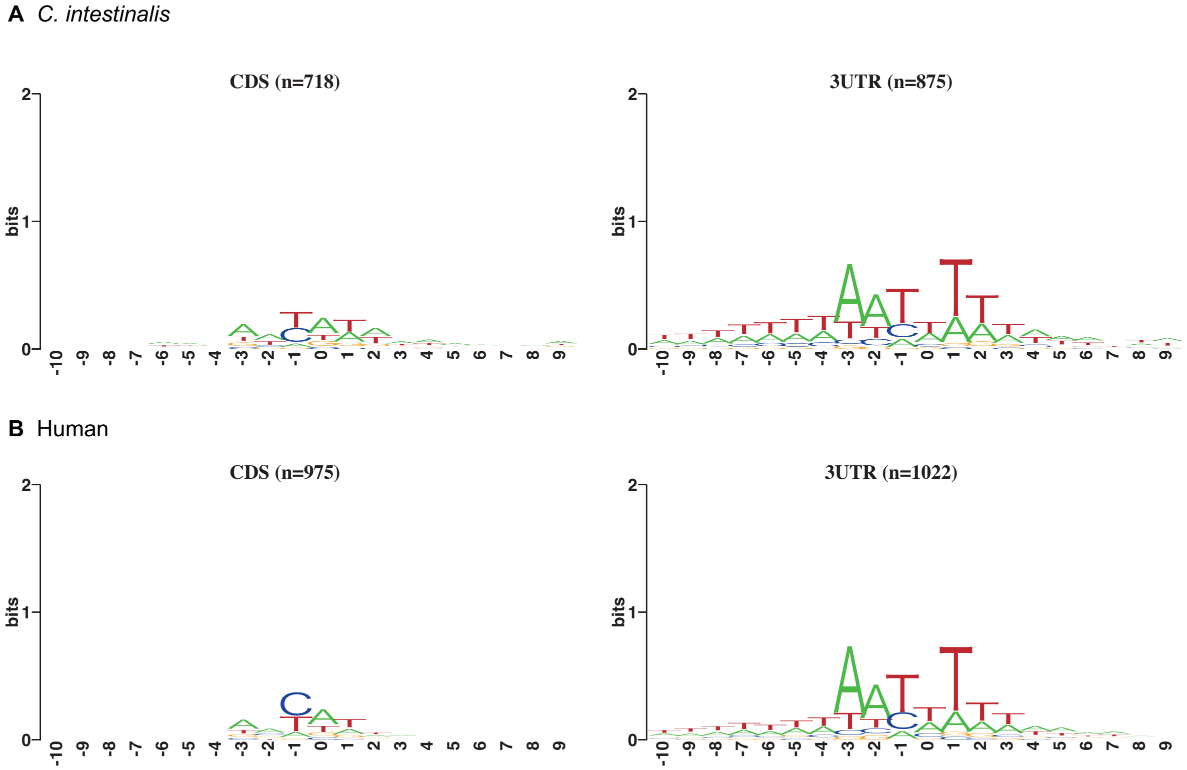


Figure S40: Sequence logos of TSCs in CDSs and 3′ UTRs. The top and bottom figures show the sequence logos of TSCs in CDSs and 3′ UTRs after removing CTGG TSCs, two types of 1-bp width TSCs, and possibly truncated-RNA derived TSCs in *C. intestinalis* (A) and human (B), respectively. The x-axis represents the distance from the representative TSS. The number in parentheses represents the number of TSCs.

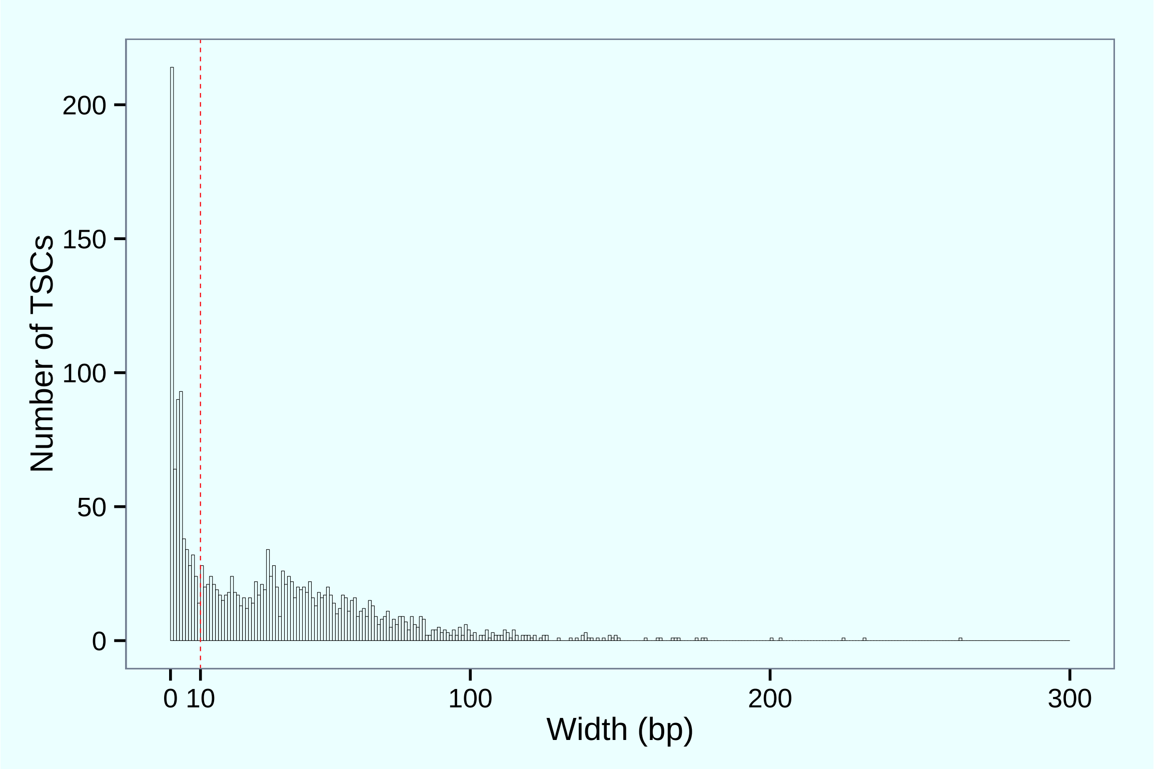


Figure S41: Distribution of width of TSCs located at known TSSs.

3 Supplemental Tables

Table S1: Preprocessing statistics. The raw reads were preprocessed before mapping them to the reference genome. The table shows the number of reads after each preprocessing step. “quality1” and “with GG” represent the reads that passed the Illumina quality filtering and the reads with GG at the 5′ ends, respectively. BWM, body wall muscle; NC, neural complex.

Sample	raw reads	quality1	with GG	not contaminants	unmapped to rRNA	SL(−) reads	SL(+) reads
ovary	34913433	23951726	21597614	19723497	13628650	7148409	6161390
heart	29518199	22147530	20538561	19233112	18214384	14094602	3919239
BWM	32564064	23437074	21679278	20341295	18080682	13887188	4027858
NC	31628569	23377130	21921787	20966915	19802857	16612659	3018631
larva	26967445	20501993	18482078	17079141	13838575	8859559	4775063

Table S2: Mapping statistics. The SL(−) reads and the SL(+) reads were mapped to the reference genome. The table shows the number of reads and the percentage of reads relative to total reads in each sample. BWM, body wall muscle; NC, neural complex.

SL(−) reads

Sample	reads		lowquality or homopolymer		unmapped		mapped		multiply mapped		uniquely mapped	
ovary	7,148,409	100%	89314	1.20%	437,486	6.10%	6,621,609	92.60%	269,960	3.80%	6,351,649	88.90%
heart	14,094,602	100%	111684	0.80%	2,756,007	19.60%	11,226,911	79.70%	510,169	3.60%	10,716,742	76.00%
BWM	13,887,188	100%	140469	1.00%	429,160	3.10%	13,317,559	95.90%	635,947	4.60%	12,681,612	91.30%
NC	16,612,659	100%	135680	0.80%	5,549,562	33.40%	10,927,417	65.80%	395,750	2.40%	10,531,667	63.40%
larva	8,859,559	100%	97619	1.10%	372,453	4.20%	8,389,487	94.70%	540,936	6.10%	7,848,551	88.60%

SL(+) reads

Sample	reads		lowquality or homopolymer		unmapped		mapped		multiply mapped		uniquely mapped	
ovary	6,161,390	100%	280667	4.60%	26	0.00%	5,880,697	95.40%	1,966,452	31.90%	3,914,245	63.50%
heart	3,919,239	100%	118889	3.00%	366	0.00%	3,799,984	97.00%	929,679	23.70%	2,870,305	73.20%
BWM	4,027,858	100%	167763	4.20%	17	0.00%	3,860,078	95.80%	910,913	22.60%	2,949,165	73.20%
NC	3,018,631	100%	89912	3.00%	28	0.00%	2,928,691	97.00%	696,718	23.10%	2,231,973	73.90%
larva	4,775,063	100%	156292	3.30%	106	0.00%	4,618,665	96.70%	1,037,245	21.70%	3,581,420	75.00%

Table S3: Number of identified TACs in *C. intestinalis*. The TACs were classified into seven categories according to their location (see Supplementary Method 1.4).

Location	TACs
TSS	1 (0.0%)
TAS	4748 (88.4%)
$5'$ UTR	70 (1.3%)
CDS	40 (0.7%)
$3'$ UTR	8 (0.1%)
intron	104 (1.9%)
intergenic	402 (7.5%)
total	5373 (100%)

Table S4: Number of identified TSCs in *C. intestinalis*. The TSCs were classified into seven categories according to their location (see Supplementary Method 1.4). The 79 RP gene TSCs that were manually identified were included in the TSS category even if they were not located at annotated TSSs.

Location	TSCs
TSS	2097 (21.4%)
TAS	122 (1.2%)
$5'$ UTR	420 (4.3%)
CDS	1623 (16.6%)
$3'$ UTR	1459 (14.9%)
intron	721 (7.4%)
intergenic	3350 (34.2%)
total	9792 (100%)

Table S5: Number of identified TSCs in human. The TSCs were classified into seven categories according to their location (see Supplementary Method 1.5).

Location	TSCs
TSS	5207 (33.6%)
$5'$ UTR	1452 (9.4%)
CDS	1622 (10.5%)
$3'$ UTR	1312 (8.5%)
exon(ncRNA)	273 (1.8%)
intron	1650 (10.6%)
intergenic	3982 (25.7%)
total	15498 (100%)

Table S6: Number of removed TSCs in *C. intestinalis*. The columns show, respectively, Location: location of TSCs, initial: initial set of TSCs shown in Table S4, CTGG: CTGG TSCs, A+T-rich: A+T-rich, 1-bp width TSCs, donor: 1-bp width TSCs near splice donor sites on the reverse strand, CDS+$3'$ UTR: possible noise TSCs that overlap with CDSs and $3'$ UTRs, non-PyPu: TSCs without PyPu motifs, where PyPu is TA, CA or TG, removed: total of removed TSCs, final: final set of TSCs (see Supplemental Methods for the details of each filtering). The numbers in parentheses represent the percentage of the TSCs to the initial TSCs.

Location	initial	CTGG	A+T-rich	donor	CDS+$3'$ UTR	non-PyPu	removed	final
TSS	2097 (100)	11 (0.5)	-	0 (0)	163 (7.8)	-	174 (8.3)	1923 (91.7)
TAS	122 (100)	0 (0)	-	0 (0)	65 (53.3)	25 (20.5)	90 (73.8)	32 (26.2)
$5'$ UTR	420 (100)	16 (3.8)	-	0 (0)	144 (34.3)	117 (27.9)	277 (66.0)	143 (34.0)
CDS	1623 (100)	18 (1.1)	-	0 (0)	1605 (98.9)	-	1623 (100)	0 (0.0)
$3'$ UTR	1459 (100)	25 (1.7)	-	0 (0)	1434 (98.3)	-	1459 (100)	0 (0.0)
intron	721 (100)	24 (3.3)	159 (22.1)	0 (0)	51 (7.1)	248 (34.4)	482 (66.9)	239 (33.1)
intergenic	3350 (100)	1339 (40.0)	328 (9.8)	34 (1.0)	10 (0.3)	770 (23.0)	2481 (74.1)	869 (25.9)
total	9792 (100)	1433 (14.6)	487 (5.0)	34 (0.3)	3472 (35.5)	1160 (11.8)	6586 (67.3)	3206 (32.7)

Table S7: Number of removed TSCs in human. The columns show, respectively, Location: location of TSCs, initial: initial set of TSCs shown in Table S5, CTGG: CTGG TSCs, A+T-rich: A+T-rich, 1-bp width TSCs, donor: 1-bp width TSCs near splice donor sites on the reverse strand, CDS+$3'$ UTR: possible noise TSCs that overlap with CDSs and $3'$ UTRs, removed: total of removed TSCs, final: final set of TSCs (see Supplemental Methods for the details of each filtering). The numbers in parentheses represent the percentage of the TSCs to the initial TSCs.

Location	initial	CTGG	A+T-rich	donor	CDS+$3'$ UTR	removed	final
TSS	5207 (100)	29 (0.6)	-	0 (0)	105 (2.0)	134 (2.6)	5073 (97.4)
$5'$ UTR	1452 (100)	4 (0.3)	-	0 (0)	112 (7.7)	116 (8.0)	1336 (92.0)
CDS	1622 (100)	28 (1.7)	-	0 (0)	1594 (98.3)	1622 (100)	0 (0.0)
$3'$ UTR	1312 (100)	17 (1.3)	-	0 (0)	1295 (98.7)	1312 (100)	0 (0.0)
exon(ncRNA)	273 (100)	3 (1.1)	-	0 (0)	26 (9.5)	29 (10.6)	244 (89.4)
intron	1650 (100)	29 (1.8)	38 (2.3)	0 (0)	34 (2.1)	101 (6.1)	1549 (93.9)
intergenic	3982 (100)	648 (16.3)	85 (2.1)	2 (0.1)	0 (0)	735 (18.5)	3247 (81.5)
total	15498 (100)	758 (4.9)	123 (0.8)	2 (0.0)	3166 (20.4)	4049 (26.1)	11449 (73.9)

Table S8: Number of TSCs after removing three types of 1-bp width TSCs and TSCs in CDSs and 3′ UTRs in *C. intestinalis*. The 79 RP gene TSCs that were manually identified were included in the TSS category.

Location	TSCs
TSS	1923 (44.0%)
TAS	57 (1.3%)
5′ UTR	260 (6.0%)
CDS	0 (0.0%)
3′ UTR	0 (0.0%)
intron	487 (11.2%)
intergenic	1639 (37.5%)
total	4366 (100%)

Table S9: Number of TSCs after removing three types of 1-bp width TSCs and TSCs in CDSs and 3′ UTRs in human

Location	TSCs
TSS	5073 (44.3%)
5′ UTR	1336 (11.7%)
CDS	0 (0.0%)
3′ UTR	0 (0.0%)
exon(ncRNA)	244 (2.1%)
intron	1549 (13.5%)
intergenic	3247 (28.4%)
total	11449 (100%)

Table S10: Number of promoters in each class. The 1844 non-RP promoters and 79 RP promoters were classified into four classes: predicted *trans*-spliced gene promoters, non-*trans*-spliced gene promoters, annotated operon gene promoters, predicted operon gene promoters and not-determined (ND) promoters. Also, to examine whether the *trans*-spliced gene promoters have different characteristics depending on their outron length, the predicted *trans*-spliced gene promoters were further divided into two classes: those with a short ($>$ 200 bp) outron and those with a long ($\geq$ 200 bp) outron. If predicted *trans*-spliced gene promoters had $\geq$ two candidate outrons (that is, TSCs paired with $\geq$ two TACs), the TAC with the highest number of tags was selected. The 200 bp cutoff was selected based on the distribution of outron length. The non-*trans*-spliced genes were defined as the genes without TACs and annotated TASs between their TSCs and annotated translational start sites. The annotated operon gene promoters were defined as the promoters of genes that are organized as operons. The ND promoters represents the promoters that were not classified into any of the other classes.

Class	non-RP	RP	total
ND	947	45	992
non-*trans*-spliced	525	13	538
annotated operon	305	14	319
predicted operon	2	0	2
trans-spliced(<200bp)	42	4	46
trans-spliced(>=200bp)	23	3	26
total	1844	79	1923

Table S11: Frequency of genes with multiple TSCs. The final set of TSCs was assigned to genes. The table shows the frequency of genes with a given number of TSCs. Each TSC represents a known promoter or a putative promoter.

Number of TSCs	Frequency of genes
1	2466
2	109
3	5
4	1
total	2581

Table S12: Command options used in mappings

Reads	NovoAlign	MapSplice
SL(−) reads	-s -o SAM -l 17 –3Prime	-L 17 -E 2 -m 2
SL(+) reads	-s -o SAM -l 9 –3Prime	-L 7 -E 0 -m 0

Table S13: Overrepresented CTGG variants. We examined whether CTGG variants, which were defined as the CTGG sequences with up to two mismatches, were overrepresented in the TSCs on the antisense strand of exons compared to the TSCs in the TSS category using Fisher's exact test. The table shows the CTGG variants that were found in at least one TSC on the antisense strand of exons in *C. intestinalis.* The asterisks (* and **) mean that the CTGG variants were significantly overrepresented in the TSCs on the antisense strand of exons (* for q-value $<$ 0.01 and ** for q-value $<$ 0.05). The q-values were calculated by the R package qvalue (Dabney et al., 2013). "#" and "%" represent the number of TSCs with the CTGG variant and the percentage relative to the total number of TSCs in each location, respectively.

CTGG variant	# (on antisense exon)	%	# (in TSS)	%	q-value
TTGG**	90	9.5	0	0	1.77E-30
CTGA**	81	8.5	0	0	1.34E-27
CCGG**	71	7.5	0	0	2.88E-24
TCGG**	46	4.8	0	0	9.63E-16
CAGG**	45	4.7	0	0	1.70E-15
ATGG**	37	3.9	0	0	7.61E-13
GTGG**	27	2.8	0	0	1.60E-09
CTGC**	24	2.5	8	0.7	1.51E-03
CTCG**	23	2.4	0	0	3.13E-08
CCAG**	17	1.8	0	0	2.89E-06
CGGG**	13	1.4	2	0.2	1.82E-03
CTTG**	13	1.4	2	0.2	1.82E-03
CTAG**	11	1.2	0	0	2.20E-04
CCGA**	10	1.1	0	0	4.07E-04
ACGG**	10	1.1	0	0	4.07E-04
AAGG**	10	1.1	1	0.1	2.53E-03
CAGA**	8	0.8	0	0	1.65E-03
TGGG**	7	0.7	0	0	2.85E-03
CTTT	6	0.6	36	3.3	1.51E-05
ATGC	6	0.6	8	0.7	0.24
CTGT	6	0.6	2	0.2	0.07
GAGG*	6	0.6	1	0.1	0.03
TTGC	4	0.4	11	1	0.08
CTTC	4	0.4	30	2.8	1.91E-05
TAGG	3	0.3	1	0.1	0.11
GCGG	3	0.3	1	0.1	0.11
GTGT	3	0.3	7	0.6	0.11
CGAG	3	0.3	1	0.1	0.11
TTGA	3	0.3	0	0	0.05
GTGA	3	0.3	0	0	0.05
TTGT	3	0.3	4	0.4	0.24
CAGT	2	0.2	7	0.6	0.08
CAAG	2	0.2	1	0.1	0.16
CTCC	2	0.2	3	0.3	0.24
TTAG	2	0.2	0	0	0.08
GGGG	2	0.2	0	0	0.08
ATGA	2	0.2	0	0	0.08
CTAA	2	0.2	0	0	0.08
GTAG	2	0.2	0	0	0.08
TTTG	2	0.2	0	0	0.08
CTCT	1	0.1	5	0.5	0.08
CCGC	1	0.1	6	0.6	0.06
CGGT	1	0.1	7	0.6	0.04
CATG	1	0.1	0	0	0.13
CTAT	1	0.1	2	0.2	0.24
CGTG	1	0.1	0	0	0.13
ATGT	1	0.1	3	0.3	0.17
CACG	1	0.1	0	0	0.13
CGGA	1	0.1	0	0	0.13

4 Supplemental Files

4.1 Supplemental File 1 - List of ribosomal protein (RP) gene promoters

Supplemental File 1 contains a table listing 79 RP genes and their transcription start sites (TSSs) in *Ciona intestinalis*. The meaning of each column is as follows. "RP", "Gene", and "TSS" represent an RP name, the gene locus of the RP in the Kyoto Hoya (KH) model, and the representative TSS of the RP gene, respectively. "Nearest RP transcript" represents the non-spliced leader (nonSL) *trans*-spliced RP transcript model in which the 5′ end is nearest to the representative TSS. "Distance" represents the distance between the 5′ end of the nonSL RP transcript model and the representative TSS. "TATA(C)" and "TATA(H)" represent the presence (+) or absence (−) of TATA boxes based on our definition in *C. intestinalis* and human, respectively. "TSSD" represents the TSS distribution type of RP gene promoters in *C. intestinalis*.

4.2 Supplemental File 2 - List of candidate promoters of spliced leader (SL) *trans*-spliced genes

Supplemental File 2 contains a table listing pairs of clusters [*trans*-splice acceptor site cluster (TAC) and upstream TSS cluster (TSC)] with the same expression specificity. The meaning of each column is as follows. "TAC" indicates the SL or not determined (ND) transcript; the TAC was located at the 5′ end of this transcript. If the TAC was not located at the 5′ ends of the SL and not determined (ND) transcript models, it was considered to be a putative *trans*-splice acceptor site (TAS). "TSC" indicates the nonSL or ND transcript model in which the TSC was located at the 5′ end of the transcript. If the TSC was not located at the 5′ ends of nonSL and ND transcript models, it was considered a putative promoter. "Location(TSC)" represents the location of the TSC. "Type" indicates the type of pairs of clusters. "TAS", "TSS", and "Distance" indicate the representative TAS, the representative TSS, and the distance between the TSS and TAS, respectively. "Expression" represents in which samples the TSC is significantly highly expressed. NC: neural complex, BWM: body wall muscle.

4.3 Supplemental File 3 - List of putative alternative promoters

Supplemental File 3 contains a table listing putative alternative promoters. The meaning of each column is as follows. "Gene", "TSS", and "Location" represent a gene locus in the KH model, the representative TSS, and the location of the TSC based on the KH model, respectively. "Status" indicates whether the TSC is a known or a putative promoter. "Transcript" represents the transcript model in which the TSC was located at the 5′ end. "NA" means that transcript models were not available because the TSC did not represent a known promoter. "Expression" represents samples in which the TSC is significantly highly expressed. NC: neural complex, BWM: body wall muscle. "Homolog" represents a human homologous protein identified by BLAST. "No hits" means that significant hits were not found. "Unknown" means that the TSC represented a putative promoter.

References

Crooks, G. E., Hon, G., Chandonia, J. M., and Brenner, S. E. (2004). WebLogo: A sequence logo generator. *Genome Research* **14**:1188–1190.

Dabney, A., Storey, J. D., and with assistance from Gregory R. Warnes (2013). *qvalue: Q-value estimation for false discovery rate control.* R package version 1.36.0.

Grant, C. E., Bailey, T. L., and Noble, W. S. (2011). FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**:1017–8.

Kel, A. E., Gossling, E., Reuter, I., Cheremushkin, E., Kel-Margoulis, O. V., and Wingender, E. (2003). MATCH (TM): a tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Research* **31**:3576–3579.

Matys, V., Kel-Margoulis, O. V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K., et al. (2006). TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res* **34**:D108–10.

Ni, T., Corcoran, D. L., Rach, E. A., Song, S., Spana, E. P., Gao, Y., Ohler, U., and Zhu, J. (2010). A paired-end sequencing strategy to map the complex landscape of transcription initiation. *Nature Methods* **7**:521–U57.

Ponjavic, J., Lenhard, B., Kai, C., Kawai, J., Carninci, P., Hayashizaki, Y., and Sandelin, A. (2006). Transcriptional and structural impact of TATA-initiation site spacing in mammalian core promoters. *Genome Biol* **7**:R78.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., and GlÃűckner, F. O. (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* **41**:D590–6.

Sandelin, A., Alkema, W., EngstrÃűm, P., Wasserman, W. W., and Lenhard, B. (2004). JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res* **32**:D91–4.

Satou, Y., Mineta, K., Ogasawara, M., Sasakura, Y., Shoguchi, E., Ueno, K., Yamada, L., Matsumoto, J., Wasserscheid, J., Dewar, K., et al. (2008). Improved genome assembly and evidence-based global gene model set for the chordate Ciona intestinalis: new insight into intron and operon populations. *Genome Biology* **9**.

Wang, K., Singh, D., Zeng, Z., Coleman, S. J., Huang, Y., Savich, G. L., He, X., Mieczkowski, P., Grimm, S. A., Perou, C. M., et al. (2010). MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res* **38**:e178.

Yamashita, R., Sathira, N. P., Kanai, A., Tanimoto, K., Arauchi, T., Tanaka, Y., Hashimoto, S., Sugano, S., Nakai, K., and Suzuki, Y. (2011). Genome-wide characterization of transcriptional start sites in humans by integrative transcriptome analysis. *Genome Res* **21**:775–89.

Zhang, Z., Schwartz, S., Wagner, L., and Miller, W. (2000). A greedy algorithm for aligning DNA sequences. *J Comput Biol* **7**:203–14.

Zhao, X., Valen, E., Parker, B. J., and Sandelin, A. (2011). Systematic Clustering of Transcription Start Site Landscapes. *Plos One* **6**.