A Manually Curated Gene Model Set for an Ascidian, Ciona robusta (Ciona intestinalis Type A)

Authors: Satou, Yutaka, Tokuoka, Miki, Oda-Ishii, Izumi, Tokuhiro, Sinichi, Ishida, Tasuku, et al.

Source: Zoological Science, 39(3) : 253-260

Published By: Zoological Society of Japan

URL: https://doi.org/10.2108/zs210102

ZOOLOGICAL SCIENCE **39**: 253–260 (2022)

A Manually Curated Gene Model Set for an Ascidian, *Ciona robusta* (*Ciona intestinalis* type A)

Yutaka Satou*, Miki Tokuoka, Izumi Oda-Ishii, Sinichi Tokuhiro, Tasuku Ishida, Boqi Liu, and Yuri Iwamura

*Department of Zoology, Graduate School of Science, Kyoto University,
Sakyo, Kyoto 606-8502, Japan*

Gene/transcript model sets predicted from decoded genome sequences are an important resource for a wide range of biological studies. Accuracy of gene models is therefore critical for deducing accurate conclusions. Computationally predicted models are sometimes inconsistent with experimental data from cDNA clones and RNA-sequencing. In an ascidian, *Ciona robusta* (*Ciona intestinalis* type A), a manually curated gene/transcript model set, which was constructed using an assembly in which 68% of decoded sequences were associated with chromosomes, had been used during the last decade. Recently a new genome assembly was published, in which over 95% of decoded sequences are associated with chromosomes. In the present study, we provide a high-quality version of the gene/transcript model set for the latest assembly. Because the *Ciona* genome has been used in a variety of studies such as developmental biological studies, evolutionary studies, and physiological studies, the current gene/transcript model set provides a fundamental biological resource.

Key words: *Ciona robusta*, *Ciona intestinalis* type A, gene model, transcript model, operon

INTRODUCTION

Gene/transcript model sets predicted from decoded genome sequences are an important resource for a wide range of biological studies. Accurate models are important for genome-wide studies, including genome-wide expression studies (RNA-sequencing), studies that examine transcription binding sites by chromatin-immunoprecipitation-sequencing, studies that examine chromatin dynamics including ATAC-sequencing and Hi-C analysis, and computational and experimental inter-species comparisons of gene contents. Although many computational programs have been developed to predict genes or transcripts and have provided useful models (Yeh et al., 2001; Solovyev et al., 2006; Stanke et al., 2008; Lomsadze et al., 2014), these models sometimes show inconsistencies with experimental evidence, such as cDNA sequences. Therefore, in many cases, manual inspection and revision of models are necessary.

An ascidian, *Ciona robusta* (or *Ciona intestinalis* type A), has been used as a model organism for developmental biology, evolutionary biology, and physiology (Satoh, 2003; Lemaire, 2011; Satake et al., 2019). The initial draft genome sequence was determined in 2002 (Dehal et al., 2002). In a major update in 2008, 68% of decoded sequences were associated with chromosomes, and a manually inspected gene/transcript model set, the KH models, was concomitantly published (Satou et al., 2008). In the second major update in 2019, over 95% of decoded sequences were associated with chromosomes (Satou et al., 2019). Because the genome sequence became more continuous, some KH models were expected to become more continuous. Indeed, a gene/transcript model set, the KY models, which were constructed with Augustus (Stanke et al., 2008), was provided with the second major updated version of the genome sequence (Satou et al., 2019). Meanwhile, more cDNA sequences, including RNA-sequencing data, have recently accumulated, and these data provide insights into more accurate gene prediction; not all gene/transcript models are perfectly consistent with cDNA sequences.

The *Ciona* genome contains operons, in which multiple genes are encoded (Satou et al., 2006). Genes encoded in an operon are transcribed as one polycistronic RNA molecule, which is resolved into monocistronic RNA molecules by spliced leader (SL)-*trans*-splicing. Operons are found in many organisms and are thought to have independently evolved multiple times (Hastings, 2005). *Ciona* operons are unique, because operonic genes are encoded with no intergenic sequences (Satou et al., 2006). Because of this feature, accurate computational prediction of gene structures within operons is often difficult. This is an additional reason that manual inspection of gene/transcript models is necessary for *Ciona*.

MATERIALS AND METHODS

Construction of gene/transcript models

The genome sequence of *Ciona robusta* (*Ciona intestinalis* type A) and the original KY gene/transcript models were published in a previous study (Satou et al., 2019). We inspected all KY

* Corresponding author. E-mail: yutaka@ascidian.zool.kyoto-u.ac.jp
doi:10.2108/zs210102

gene/transcript models by eye in the genome browser of the Ghost database (Satou et al., 2005). When models were not consistent with expressed sequence tags (ESTs) or RNA-sequencing data, such models were revised by rewriting the corresponding entry in the annotation file in the General Feature Format (version 3; GFF3) format. When different genes incorrectly shared the same gene name, their names were revised to clarify that they are distinct genes.

We used ESTs for this species deposited in the EST division of the DDBJ/EMBL/GenBank database (accession numbers: AV671213–AV681457, AV837126–AV908849, AV947525–AV999999, BP000001–BP018879, BW000001–BW034962, BW034964–BW509978, BW648671–BW650748, FF685517–FF836289, FF848360–FF999999, FK041677–FK250481, FG000001–FG007279) (Nishikata et al., 2001; Satou et al., 2001; Fujiwara et al., 2002; Inaba et al., 2002; Kusakabe et al., 2002; Ogasawara et al., 2002; Shida et al., 2003; Satou et al., 2005; Tassy et al., 2010), and RNA-sequencing data (accession numbers: DRR033088–DRR033090, DRR075500–DRR075505, DRR110784–DRR110793, DRR146815–DRR146818, SRR6283055–SRR6283068, SRR6479421–SRR6479424, and SRR8587679–SRR8587684) (Waki et al., 2015; Tokuhiro et al., 2017; Brozovic et al., 2018; Kobayashi et al., 2018; Tokuoka et al., 2018; Racioppi et al., 2019; Imai et al., 2020).

Nomenclature of gene/transcript names followed that used in a previous study (Satou et al., 2019). Transcript names consist of five fields delimited by dots (e.g., KY21.Chr1.1.v1.nonSL1-1). The first field represents the gene/transcript model version; therefore, all models have the same tag (KY21 stands for the Kyoto version developed in 2021). The second field represents the chromosome (or unassembled contig) name. The third field represents the serial number for gene loci on individual chromosomes. The fourth field specifies alternative transcript variants by a number preceded by the letter “v.” The fifth field includes information for the 5′- and 3′-ends of models, and consists of two subfields delimited by hyphens. The first subfield refers to the evidence identifying the 5′-end: SL means an experimentally defined *trans*-splice acceptor site, nonSL means an experimentally determined non-*trans*-spliced mRNA 5′-end, and ND means a 5′-end that was computationally predicted (not determined by experimental evidence). The number concatenated to the 5′-end code identifies individual alternative 5′-ends within each locus. The second subfield refers to the 3′-end and consists of numbers identifying individual alternative 3′-ends within each locus. Gene models are defined with the first three fields (KY.Chr1.1). Note that gene numbers are not compatible between the KY and KY21 model sets.

To determine the 5′-ends of models, we used sequences derived from cDNA libraries constructed with a method that captured the 5′-end of mRNAs (Satou et al., 2006; Yokomori et al., 2016), and sequences from PCR products for 5′-ends of mRNAs with an SL (Matsumoto et al., 2010). When two or more sequences with the SL sequence (Vandenberghe et al., 2001) were mapped next to the acceptor ‘AG’, we considered that position to be a potential SL-*trans*-splice acceptor site (SLTSAS). When two or more sequences without the spliced-leader were mapped to a single position, we regarded it as a potential transcription start site (TSS).

Gene/transcript models were written in the GFF3 format (see Supplementary File S1). The gene/transcript models were visualized with jbrowse (Skinner et al., 2009; Buels et al., 2016) in Ghost database (http://ghost.zool.kyoto-u.ac.jp/default_ht.html) (Satou et al., 2005).

Evaluation of transcript models

To calculate coverage for transcript models by ESTs and RNA-sequencing data, we used the same ESTs and RNA-sequencing dataset that we used to inspect models. ESTs were mapped to transcript models using pblat (version 35) (Kent, 2002; Wang and Kong, 2019) with default parameters. ESTs were used to calculate coverage of individual models only when 80% or more of the entire length was successfully mapped. RNA-sequencing data were mapped to transcript models using Bowtie2 (version 2.2.2) (Langmead and Salzberg, 2012) with the “-a” option, and the resultant SAM files were used to calculate coverage of individual models.

To evaluate gene model completeness, we used BUSCO (version 3.1.0) with the metazoan gene model set (Simao et al., 2015). For comparison, we evaluated the Ensembl (Howe et al., 2021) (the file Ciona_intestinalis.KH.pep.all.fa included in release 105 was used) and RefSeq model sets (O’Leary et al., 2016) (file name: GCF_000224145.3_KH_protein.faa), which were constructed using the KH assembly (Satou et al., 2008).

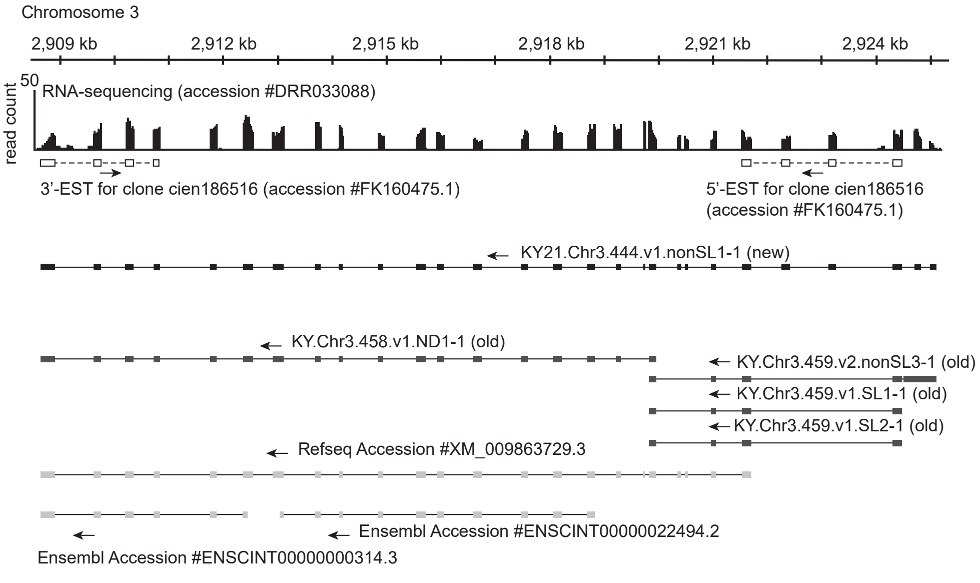

Fig. 1. A genomic region that contains a gene locus encoding a protein similar to fibronectin. In this genomic region, one or two genes have been predicted in gene models constructed in previous studies (Howe et al., 2021; O’Leary et al., 2016; Satou et al., 2019) (bottom; dark and light gray boxes indicate exons and dark and light gray lines indicate introns). However, paired ESTs derived from a single cDNA clone indicate that this genomic region contains one gene (white boxes), and RNA-sequencing data (shown in the top) indicates that not all exons are accurately predicted. Therefore, we fused two previous KY models and added missing exons to construct a new transcript model, KY21.Chr3.444.v1.nonSL1-1 (middle; exons are shown by black boxes and introns are shown by black lines). The first two exons of this model overlapped the first exon of an old model (KY.Chr3.459.v2.nonSL3-1). The RNA-sequencing data agree with the current model. The entire length of this locus encodes a polypeptide similar to fibronectin, and the polypeptide encoded by the first and second exons is also similar to fibronectin, which also supports the current model.

RESULTS AND DISCUSSION

Construction of a new gene/transcript model set

Using the Ghost database genome browser, we inspected all KY gene models (14,072 gene models), because the KY gene/transcript model set is the only gene/transcript model set built using the latest version of the *Ciona* genome assembly (Satou et al., 2019). We also inspected 4119 *ab initio* models, which were excluded from the core set of the KY models because of lack of cDNA evidence. In particular, we paid attention to whether each model was consistent with cDNA sequences.

ESTs were useful, because they more clearly gave information on gene boundaries than RNA-sequencing data. On the other hand, RNA-sequencing data were especially useful for inspecting exons of long transcripts, although transcript orientations were not deduced from RNA-sequencing data. We also used TSS and SLTSAS data to determine 5′-ends of genes (see Materials and Methods). One example is shown in Fig. 1. In this genomic region, two KY genes were predicted in the previous KY set (Satou et al., 2019), and one RefSeq model (O'Leary et al., 2016) and two Ensemble models (Howe et al., 2021) were mapped to this region. A pair of 5′- and 3′-ESTs for a single clone indicated that this region encoded one gene, but none of the above models represented the actual transcript. Therefore, we constructed one new gene model (KY21.Chr3.444) by combining two KY gene models and adding new exons. RNA-sequencing data also supported this new transcript model.

In this way, we updated transcript models for 6265 genes and added transcript models for 503 new genes (transcript models that were not updated are listed in Supplementary Table S1). This new transcript model set corresponded to 22,083 splicing variants of 18,788 genes (Table 1). When multiple 5′-ends (TSSs or SLTSASs) or multiple 3′-ends were indicated by experimental data, multiple transcript models for a single splicing variant were built. Consequently, 55,505 transcript models were built. We named this set Kyoto 2021 model set (KY21). In the KY21 model set, 17,295 genes (92%) were found in chromosomes and the remaining 1493 were found in contigs unassociated with chromosomes.

Ideally, each model should start with a TSS or SLTSAS, and its coding sequence should end with a stop codon and be followed by a 3′-untranslated region (UTR). However, among the 55,505 transcript models, 27,888 started with TSSs and 17,951 started with SLTSASs (Table 1). Because the 5′-ends of the remaining 9666 models did not have solid experimental evidence, these models might lack exons of actual transcripts. Such models were found more frequently in contigs unassociated with chromosomes (1515/1792 transcripts; 84.5%) than in chromosomes (8151/53,713 transcripts; 15.2%). Meanwhile, 1080 transcript models did not contain 3′-UTRs, which indicated that these models did not correctly represent 3′-regions of actual transcripts. Such models with no 3′-UTR were found more frequently in contigs unassociated with chromosomes (193/1792 transcripts; 10.8%) than in chromosomes (887/53,713 transcripts; 1.65%).

Distribution of lengths of the longest open reading frames (ORFs) of 18,788 gene models is shown in Fig. 2.

Table 1. Basic statistics of the present (KY21) gene model set.

	KY21[*1]	KY[*1, *2]
Genes	18,788 (17,295)	18,191 (16,701)
Splicing variants	22,083 (20,495)	24,110 (22,332)
Transcripts	55,505 (53,713)	68,338 (66,059)
transcripts that begins with a TSS	27,888 (27,709)	32,335 (32,098)
transcripts that begins with a SL-*trans*-splicing-acceptor site	17,951 (17,853)	22,666 (22,535)
transcripts whose 5'-end is not experimentally determined	9,666 (8,151)	13,337 (11,426)

*1 Numbers of genes and transcripts encoded in contigs associated with chromosomes are shown in parentheses.
*2 "*ab initio*" models are included

Table 2. Evaluation of the present (KY21) gene model set using BUSCO.

	Found			
	Total	Complete	Fragmented	Missing
KY21 models	97.1%	96.2%	0.9%	2.9%
KY models[*1]	96.9%	95.6%	1.3%	3.1%
Ensembl[*2]	89.7%	84.5%	5.2%	10.3%
Refseq[*3]	94.8%	93.8%	1.0%	5.2%
KH	94.2%	93.0%	1.2%	5.8%

*1 Values for KY models were adopted from a previous study (Satou et al., 2019).
*2 We used protein sequences contained in the file Ciona_intestinalis.KH.pep.all.fa
*3 We used protein sequences contained in the file GCF_000224145.3_KH_protein.faa

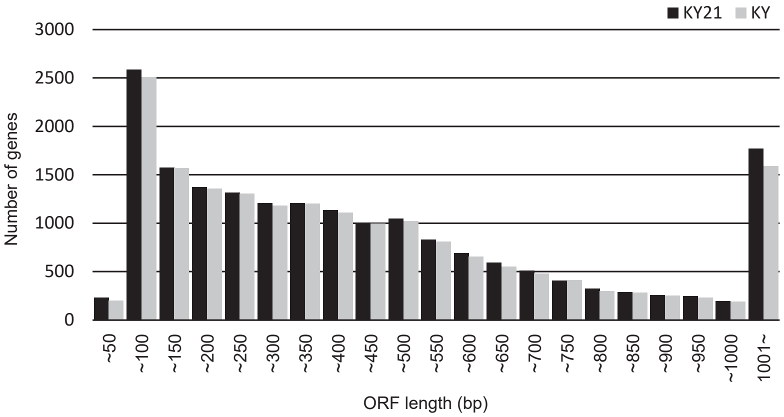

Fig. 2. A histogram that shows the number of genes with respect to ORF lengths. The amino acid length of the longest ORF for each gene was calculated (x-axis). The y-axis indicates the number of genes in each bin. For comparisons, histograms for the KY21 and KY model sets are shown by black and gray bars, respectively. Note that *ab initio* models were included in the KY model set we analyzed.

Among these gene models, 2816 models had ORFs shorter than or equal to 100 base-pairs long, and 4396 models had an ORF shorter than or equal to 150 base-pairs long. In the KY21 version, we included gene models with such short ORFs. Therefore, some of these models with short ORFs may represent non-coding RNAs. Consequently, the number of gene models was greater than that of previous ones (Dehal et al., 2002; Satou et al., 2008, 2019).

Evaluation of KY21 models

We first evaluated the KY21 model set using BUSCO, a software for assessing completeness of gene models with single-copy orthologs (Simao et al., 2015). Among the "metazoan" genes, 97.1% were found in the KY21 model set, whereas 96.9% were found in the original KY model set (Table 2). The proportion of fragmented genes was improved from 1.3% in the original KY set to 0.9% in the revised KY21 set. For further comparison, we tested gene model sets from Ensembl and RefSeq projects (O'Leary et al., 2016; Howe et al., 2021), which were constructed using the earlier KH version of the genome assembly and our gene model set built on the KH version of the genome assembly (Satou et al., 2008). As shown in Table 2, the KY21 set produced better values than these previous sets.

Next, we mapped publicly available ESTs onto the transcript models by pblat (Wang and Kong, 2019) to examine how many nucleotides of each model were covered by ESTs. ESTs were mapped to over 95% of the entire length of 42,949 (77.4%) models, but were mapped to less than 10% of the entire length of 3789 (6.8%) models (Fig. 3A). A similar analysis was also conducted using publicly available RNA-sequencing data and the Bowtie2 program (Langmead and Salzberg, 2012). ESTs and RNA-sequencing data were mapped to over 95% of the entire length of 53,842 (97.0%) models, but were mapped to less than 10% of the entire length of 50 (0.09%) models (Fig. 3B). Thus, the majority of the present models had cDNA evidence over their entire lengths. Finally, to see how we improved the models, we performed a similar analysis only for the updated models (Fig. 3C). ESTs and RNA-sequencing data were mapped to over 95% of the entire length of 98.6% of the updated models in the KY21 set; this percentage was better than that of the KY model set (94.7%).

Operons

The *Ciona* genome contains operons, in which multiple genes are encoded with no intergenic nucleotides and downstream genes were always subject to SL-*trans*-splicing to be resolved into monocistronic mRNA molecules (Satou et

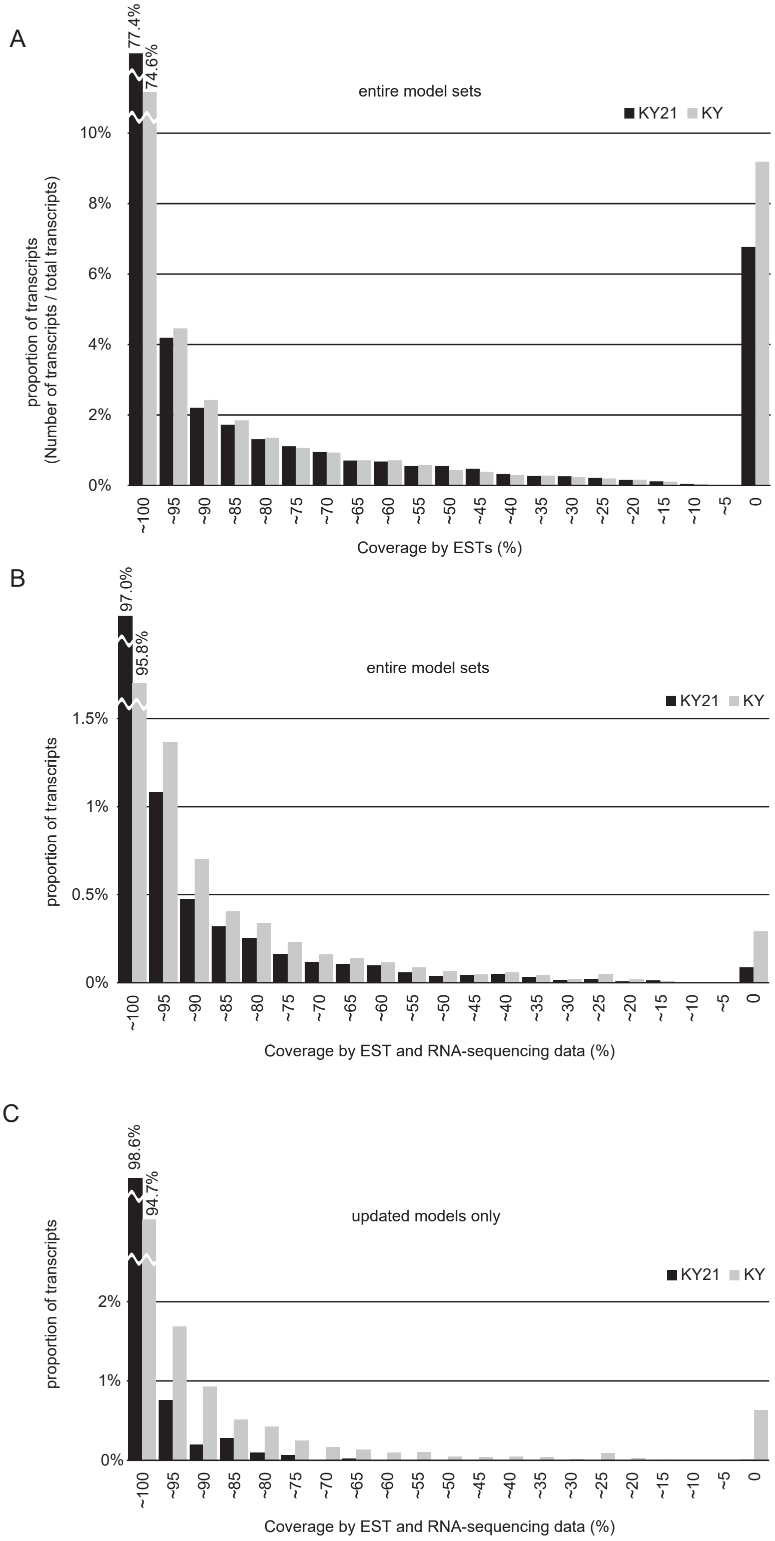

Fig. 3. Histograms that show proportions of transcripts with cDNA evidence. **(A)** We aligned ESTs to each transcript model and calculated how many bases of each model were aligned to ESTs (x-axis). The y-axis indicates the proportion of transcripts in each bin. **(B, C)** We also aligned RNA-sequencing data and calculated how many bases of each model were aligned to ESTs and RNA-sequencing data for the entire model set **(B)** and for only the updated models **(C)**. For comparisons, histograms for the KY21 and KY model sets are shown by black and gray bars, respectively. Note that *ab initio* models were included in the KY model set we analyzed.

al., 2006). In the KY21 set, there were 1163 candidate operons that satisfied the above two criteria (Table 3; all operonic genes are listed in Supplementary Table S2). Most of the operons (981) contained two genes, whereas the remaining 182 operons contained three or more genes. The largest operons contained seven genes (Fig. 4A). A total of 2561 genes were encoded in operons, which represented 13.6% of the entire gene set.

As previously shown (Satou et al., 2008), single-exon

Table 3. Operons in the *C. robusta* (*C. intestinalis* type A) genome.

Number of genes in each operon	Number of operons
2	981
3	137
4	40
5	3
6	1
7	1

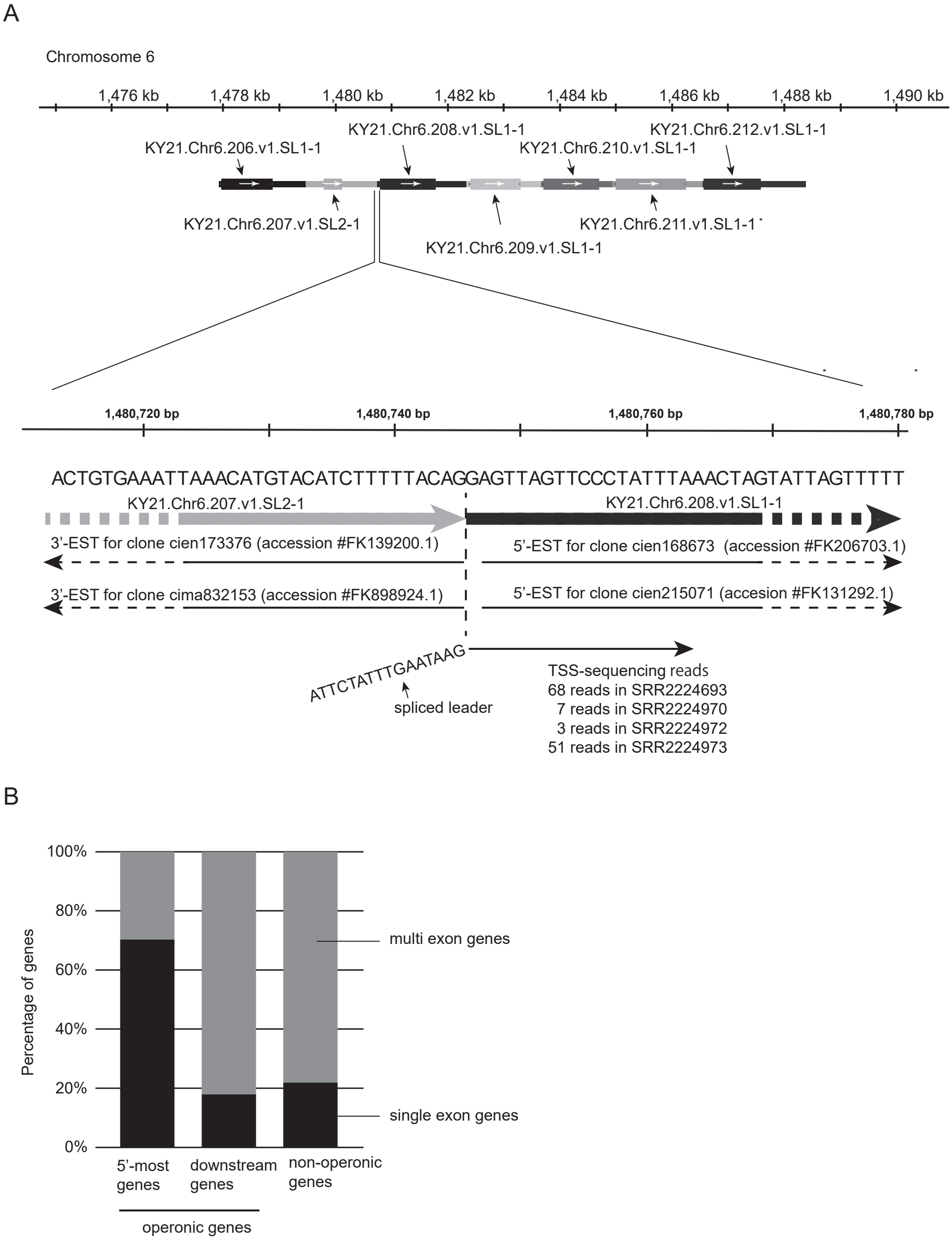

Fig. 4. A chromosomal region encoding an operon. **(A)** Depiction of the largest operon, which consists of seven genes (KY21.Chr6.206–KY21.Chr6.212). These genes are all single exon genes (top: protein-coding regions are indicated by thick lines, and untranslated regions are indicated by thin lines). The boundary between the second and third genes is shown at the bottom. Expressed sequence tags indicated that two different genes are tandemly encoded in this region in the same orientation. Transcription start site (TSS)-sequencing reads (Yokomori et al., 2016) indicate that the third gene is spliced leader (SL)-*trans*-spliced and pinpoint the SL-*trans*-splicing acceptor site. The SL sequence is the same as the one previously identified SL sequence (Vandenberghe et al., 2001). **(B)** Proportions of single-exon genes in the most upstream genes within operons, downstream genes within operons, and in non-operonic genes.

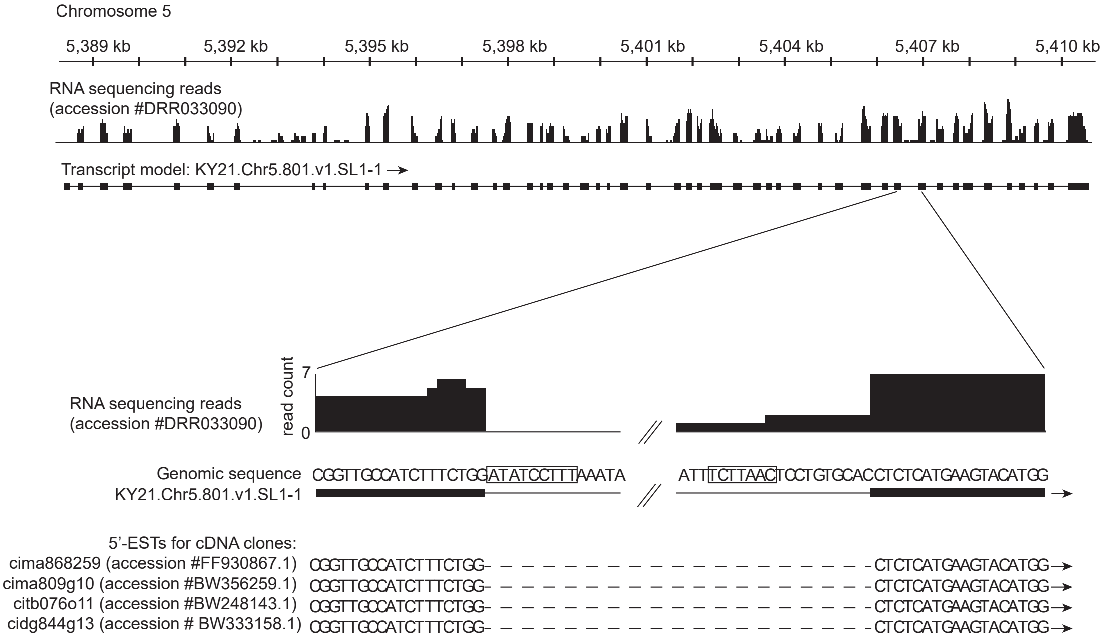
Chromosome 5
5,389 kb
5,392 kb
5,395 kb
5,398 kb
5,401 kb
5,404 kb
5,407 kb
5,410 kb
RNA sequencing reads
(accession #DRR033090)
Transcript model: KY21.Chr5.801.v1.SL1-1 →
read count
7
0
RNA sequencing reads
(accession #DRR033090)
Genomic sequence CGGTTGCCATCTTTCTGGATATCCTTTAAATA
ATTTCTTAACTCCTGTGCACCTCTCATGAAGTACATGG
KY21.Chr5.801.v1.SL1-1
5'-ESTs for cDNA clones:
cima868259 (accession #FF930867.1) CGGTTGCCATCTTTCTGG- CTCTCATGAAGTACATGG→
cima809g10 (accession #BW356259.1) CGGTTGCCATCTTTCTGG- CTCTCATGAAGTACATGG→
citb076o11 (accession #BW248143.1) CGGTTGCCATCTTTCTGG- CTCTCATGAAGTACATGG→
cidg844g13 (accession # BW333158.1) CGGTTGCCATCTTTCTGG- CTCTCATGAAGTACATGG→

Fig. 5. A non-canonical intron that begins with an AT dinucleotide and ends with an AC dinucleotide. A genomic region encoding KY21. Chr5.801.v1.SL1-1 is shown. This transcript model is supported by RNA-sequencing data shown on the top. Among its 49 introns, the 40-th intron begins with AT and ends with AC. The boundaries between the 40-th exon and 40-th intron and between the 40-th intron and 41-st exon are supported by experimental evidence (RNA-sequencing reads and four expressed sequence tags [ESTs] shown on the bottom). The nucleotide sequences typical of U12-type introns are found in the 5′-splice site and putative branch point of this intron (enclosed by boxes).

genes were overrepresented in the most-upstream genes in operons (Fig. 4B). Most (or possibly all) of these single-exon most-upstream genes appeared to encode proteins, but not to be "outrons" removed by SL-*trans*-splicing. First, 82% of the proteins encoded by these genes showed a significant similarity to human proteins in BLASTP (threshold E-value, 1E-5). Although this percentage was slightly lower than that of proteins encoded by downstream genes within operons (91%), it was much higher than that of proteins encoded by non-operonic genes (59%). Although we do not understand what this bias means, it might be related to operon occurrence and evolution. Second, the most-upstream genes in operons were represented by ESTs. Because these ESTs are derived from oligo(dT)-primed cDNAs, it is likely that polyadenylated mRNAs are produced from these most-upstream genes.

Non-canonical AT-AC introns

Among 131,567 introns of the KY21 model set, the majority (131,514) began with GT or GC dinucleotides and ended with AG dinucleotides. However, we found that 53 introns began with AT dinucleotides and ended with AC dinucleotides. The majority of these introns are removed by a minor spliceosome containing U12 RNA (Patel and Steitz, 2003). These AT–AC introns included all previously predicted AT–AC introns (Alioto, 2007; Moyer et al., 2020) and 10 new AT–AC introns (see Supplementary Table S3). One example of the newly found exons is shown in Fig. 5, in which the boundaries of the intron were supported by cDNA evidence. Introns spliced with the minor spliceosome have a longer and more conserved sequences in their 5′-ends and branch points, which are typically 5′-ATATCCTTT-3′ and 5′-CCTTAAC-3′ (Turunen et al., 2013). The example shown in Fig. 5 had sequences that were almost identical to these sequences. On the other hand, we did not find nucleotide sequences similar to the branch point consensus sequence in five of these newly found AT–AC introns, and the 5′-splice sites of four of these five introns were not similar to the typical 5′-splice site sequence of U12-type introns (see Supplementary Table S3). Therefore, these five introns may be spliced by a major spliceosome that contains U2 RNA but not by the minor spliceosome.

CONCLUSIONS

Accurate gene/transcript models and genomic sequences are important for various modern biological studies. The KY21 gene/transcript model set better represents cDNA data; therefore, it will outperform the previously published gene/transcript model sets (Satou et al., 2008, 2019; O'Leary et al., 2016; Howe et al., 2021), and provide a resource to perform more accurate analyses using genomic information. This resource will also be useful for evolutionary studies that use genomic information to compare genes among different species. In particular, because *Ciona* belongs to the sister group of vertebrates (Delsuc et al., 2006; Putnam et al., 2008), the current gene model set will be useful for understanding the origin of vertebrates and evolution of chordates.

These models will also be useful for annotating the genome of *Ciona intestinalis* (or *Ciona intestinalis* type B), because the genome sequences of *Ciona robusta* and *Ciona intestinalis* are highly similar, especially in exons (Satou et al., 2021).

It is possible that accumulation of experimental data in future studies will reveal new genes and new variants that are not included in the KY21 set. In particular, 5′-ends of

17% of the transcript models in the KY21 set have not been determined by experimental evidence. In these loci, actual transcripts may be longer than the current models. Therefore, the KY21 model set will need to be updated in the future again. Nevertheless, the KY21 set is undoubtedly the best among the existing models. With the nearly completed decoded genomic sequence (Satou et al., 2019), the KY21 gene/transcript model set will make *Ciona* an ideal animal for a wide range of genomic studies.

ACKNOWLEDGMENTS

The research was supported by grants from the Japan Society for the Promotion of Science (21H02486 and 21H05239 to Y.S., 19J40136 to I.O., and 20J40280 to M.T.). A draft of the manuscript was edited by a language editing service provided by Edanz Group.

COMPETING INTERESTS

The authors declare that they have no competing interests.

AUTHOR CONTRIBUTIONS

Conceptualization: YS; investigation: YS, MT, IO, ST, TI, BL, YI; writing — original draft: YS; writing — review and editing: MT, IO, ST, TI, BL.

SUPPLEMENTARY MATERIALS

Supplementary materials for this article are available online. (URL : https://doi.org/10.2108/zs210102)

Supplementary Table S1. KY21 transcript models whose structures were not changed from the original KY model set.

Supplementary Table S2. Operons in the *C. robusta* (*C. intestinalis* type A) genome.

Supplementary Table S3. AT-AC introns in the *C. robusta* (*C. intestinalis* type A) genome.

Supplementary File S1. Gene and transcript models represented in the GFF3 format (gz compressed).

REFERENCES

Alioto TS (2007) U12DB: a database of orthologous U12-type spliceosomal introns. Nucleic Acids Res 35: D110–D115

Brozovic M, Dantec C, Dardaillon J, Dauga D, Faure E, Gineste M, et al. (2018) ANISEED 2017: extending the integrated ascidian database to the exploration and evolutionary comparison of genome-scale datasets. Nucleic Acids Res 46: D718–D725

Buels R, Yao E, Diesh CM, Hayes RD, Munoz-Torres M, Helt G, et al. (2016) JBrowse: a dynamic web platform for genome visualization and analysis. Genome Biology 17: 66

Dehal P, Satou Y, Campbell RK, Chapman J, Degnan B, De Tomaso A, et al. (2002) The draft genome of *Ciona intestinalis*: insights into chordate and vertebrate origins. Science 298: 2157–2167

Delsuc F, Brinkmann H, Chourrout D, Philippe H (2006) Tunicates and not cephalochordates are the closest living relatives of vertebrates. Nature 439: 965–968

Fujiwara S, Maeda Y, Shin-I T, Kohara Y, Takatori N, Satou Y, et al. (2002) Gene expression profiles in *Ciona intestinalis* cleavage-stage embryos. Mech Dev 112: 115–127

Hastings KE (2005) SL trans-splicing: easy come or easy go? Trends Genet 21: 240–247

Howe KL, Achuthan P, Allen J, Allen J, Alvarez-Jarreta J, Amode MR, et al. (2021) Ensembl 2021. Nucleic Acids Res 49: D884–D891

Imai KS, Kobayashi K, Kari W, Rothbacher U, Ookubo N, Oda-Ishii I, et al. (2020) Gata is ubiquitously required for the earliest zygotic gene transcription in the ascidian embryo. Dev Biol 458: 215–227

Inaba K, Padma P, Satouh Y, Shin-I T, Kohara Y, Satoh N, et al. (2002) EST analysis of gene expression in testis of the ascidian *Ciona intestinalis*. Mol Reprod Dev 62: 431–445

Kent WJ (2002) BLAT—the BLAST-like alignment tool. Genome Res 12: 656–664

Kobayashi K, Maeda K, Tokuoka M, Mochizuki A, Satou Y (2018) Controlling cell fate specification system by key genes determined from network structure. iScience 4: 281–293

Kusakabe T, Yoshida R, Kawakami I, Kusakabe R, Mochizuki Y, Yamada L, et al. (2002) Gene expression profiles in tadpole larvae of *Ciona intestinalis*. Dev Biol 242: 188–203

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nature Methods 9: 357–359

Lemaire P (2011) Evolutionary crossroads in developmental biology: the tunicates. Development 138: 2143–2152

Lomsadze A, Burns PD, Borodovsky M (2014) Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. Nucleic Acids Res 42: e119

Matsumoto J, Dewar K, Wasserscheid J, Wiley GB, Macmil SL, Roe BA, et al. (2010) High-throughput sequence analysis of *Ciona intestinalis* SL trans-spliced mRNAs: Alternative expression modes and gene function correlates. Genome Res 20: 636–645

Moyer DC, Larue GE, Hershberger CE, Roy SW, Padgett RA (2020) Comprehensive database and evolutionary dynamics of U12-type introns. Nucleic Acids Res 48: 7066–7078

Nishikata T, Yamada L, Mochizuki Y, Satou Y, Shin-i T, Kohara Y, et al. (2001) Profiles of maternally expressed genes in fertilized eggs of *Ciona intestinalis*. Dev Biol 238: 315–331

Ogasawara M, Sasaki A, Metoki H, Shin-i T, Kohara Y, Satoh N, et al. (2002) Gene expression profiles in young adult *Ciona intestinalis*. Dev Genes Evol 212: 173–185

O'Leary NA, Wright MW, Brister JR, Ciufo S, McVeigh DHR, Rajput B, et al. (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res 44: D733–D745

Patel AA, Steitz JA (2003) Splicing double: Insights from the second spliceosome. Nat Rev Mol Cell Bio 4: 960–970

Putnam NH, Butts T, Ferrier DE, Furlong RF, Hellsten U, Kawashima T, et al. (2008) The amphioxus genome and the evolution of the chordate karyotype. Nature 453: 1064–1071

Racioppi C, Wiechecki KA, Christiaen L (2019) Combinatorial chromatin dynamics foster accurate cardiopharyngeal fate choices. eLife 8: e49921

Satake H, Matsubara S, Shiraishi A, Yamamoto T, Osugi T, Sakai T, et al. (2019) Peptide receptors and immune-related proteins expressed in the digestive system of a urochordate, *Ciona intestinalis*. Cell Tissue Res 377: 293–308

Satoh N (2003) The ascidian tadpole larva: comparative molecular development and genomics. Nat Rev Genet 4: 285–295

Satou Y, Takatori N, Yamada L, Mochizuki Y, Hamaguchi M, Ishikawa H, et al. (2001) Gene expression profiles in *Ciona intestinalis* tailbud embryos. Development 128: 2893–2904

Satou Y, Kawashima T, Shoguchi E, Nakayama A, Satoh N (2005) An integrated database of the ascidian, *Ciona intestinalis*: Towards functional genomics. Zool Sci 22: 837–843

Satou Y, Hamaguchi M, Takeuchi K, Hastings KEM, Satoh N (2006) Genomic overview of mRNA 5′-leader trans-splicing in the ascidian *Ciona intestinalis*. Nucleic Acids Res 34: 3378–3388

Satou Y, Mineta K, Ogasawara M, Sasakura Y, Shoguchi E, Ueno 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 Biol 9: R152

Satou Y, Nakamura R, Yu D, Yoshida R, Hamada M, Fujie M, et al. (2019) A nearly complete genome of *Ciona intestinalis* type A (*C. robusta*) reveals the contribution of inversion to chromosomal evolution in the genus *Ciona*. Genome Biol Evol 11:

3144–3157

Satou Y, Sato A, Yasuo H, Mihirogi Y, Bishop J, Fujie M, et al. (2021) Chromosomal inversion polymorphisms in two sympatric ascidian lineages. Genome Biol Evol 13: evab068

Shida K, Terajima D, Uchino R, Ikawa S, Ikeda M, Asano K, et al. (2003) Hemocytes of *Ciona intestinalis* express multiple genes involved in innate immune host defense. Biochem Biophys Res Commun 302: 207–218

Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31: 3210–3212

Skinner ME, Uzilov AV, Stein LD, Mungall CJ, Holmes IH (2009) JBrowse: a next-generation genome browser. Genome Res 19: 1630–1638

Solovyev V, Kosarev P, Seledsov I, Vorobyev D (2006) Automatic annotation of eukaryotic genes, pseudogenes and promoters. Genome Biol 7: S10

Stanke M, Diekhans M, Baertsch R, Haussler D (2008) Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics 24: 637–644

Tassy O, Dauga D, Daian F, Sobral D, Robin F, Khoueiry P, et al. (2010) The ANISEED database: Digital representation, formalization, and elucidation of a chordate developmental program. Genome Res 20: 1459–1468

Tokuhiro S, Tokuoka M, Kobayashi K, Kubo A, Oda-Ishii I, Satou Y (2017) Differential gene expression along the animal-vegetal axis in the ascidian embryo is maintained by a dual functional protein Foxd. PLoS Genet 13: e1006741

Tokuoka M, Kobayashi K, Satou Y (2018) Distinct regulation of *Snail* in two muscle lineages of the ascidian embryo achieves temporal coordination of muscle development. Development 145: dev163915

Turunen JJ, Niemela EH, Verma B, Frilander MJ (2013) The significant other: splicing by the minor spliceosome. WIRESs RNA 4: 61–76

Vandenberghe AE, Meedel TH, Hastings KE (2001) mRNA 5′-leader trans-splicing in the chordates. Genes Dev 15: 294–303

Waki K, Imai KS, Satou Y (2015) Genetic pathways for differentiation of the peripheral nervous system in ascidians. Nat Commun 6: 8719

Wang M, Kong L (2019) pblat: a multithread blat algorithm speeding up aligning sequences to genomes. BMC Bioinform 20: 28

Yeh RF, Lim LP, Burge CB (2001) Computational inference of homologous gene structures in the human genome. Genome Res 11: 803–816

Yokomori R, Shimai K, Nishitsuji K, Suzuki Y, Kusakabe TG, Nakai K (2016) Genome-wide identification and characterization of transcription start sites and promoters in the tunicate *Ciona intestinalis*. Genome Res 26: 140–150

(Received November 14, 2021 / Accepted January 15, 2022 /
Published online March 1, 2022)