frontiers in Genetics

BRIEF RESEARCH REPORT
published: 14 September 2021
doi: 10.3389/fgene.2021.704162

The Landscapes of Full-Length Transcripts and Splice Isoforms as Well as Transposons Exonization in the Lepidopteran Model System, *Bombyx mori*

Zongrui Dai[1,2], *Jianyu Ren*[1], *Xiaoling Tong*[1], *Hai Hu*[1], *Kunpeng Lu*[1], *Fangyin Dai*[1]* and *Min-Jin Han*[1]*

[1]State Key Laboratory of Silkworm Genome Biology, Key Laboratory of Sericultural Biology and Genetic Breeding, Ministry of Agriculture and Rural Affairs, College of Sericulture, Textile and Biomass Science, Southwest University, Chongqing, China, [2]WESTA College, Southwest University, Chongqing, China

OPEN ACCESS

Edited by:
Fei Li,
Zhejiang University, China

Reviewed by:
Yu Zhou,
Wuhan University, China
Tsukasa Fukunaga,
Waseda University, Japan

***Correspondence:**
Fangyin Dai
fydai@swu.edu.cn
Min-Jin Han
minjinhan@126.com

Specialty section:
This article was submitted to RNA, a section of the journal Frontiers in Genetics

Received: 01 May 2021
Accepted: 01 September 2021
Published: 14 September 2021

Citation:
Dai Z, Ren J, Tong X, Hu H, Lu K, Dai F and Han M-J (2021) The Landscapes of Full-Length Transcripts and Splice Isoforms as Well as Transposons Exonization in the Lepidopteran Model System, Bombyx mori. Front. Genet. 12:704162. doi: 10.3389/fgene.2021.704162

The domesticated silkworm, *Bombyx mori*, is an important model system for the order Lepidoptera. Currently, based on third-generation sequencing, the chromosome-level genome of *Bombyx mori* has been released. However, its transcripts were mainly assembled by using short reads of second-generation sequencing and expressed sequence tags which cannot explain the transcript profile accurately. Here, we used PacBio Iso-Seq technology to investigate the transcripts from 45 developmental stages of *Bombyx mori*. We obtained 25,970 non-redundant high-quality consensus isoforms capturing ~60% of previous reported RNAs, 15,431 (~47%) novel transcripts, and identified 7,253 long non-coding RNA (lncRNA) with a large proportion of novel lncRNA (~56%). In addition, we found that transposable elements (TEs) exonization account for 11,671 (~45%) transcripts including 5,980 protein-coding transcripts (~32%) and 5,691 lncRNAs (~79%). Overall, our results expand the silkworm transcripts and have general implications to understand the interaction between TEs and their host genes. These transcripts resource will promote functional studies of genes and lncRNAs as well as TEs in the silkworm.

Keywords: full-length transcripts, long noncoding RNA, transposable elements, exonization, *Bombyx mori*

INTRODUCTION

As a lepidopteran model organism, the draft genome of *Bombyx mori* has been released 17 years ago (Mita, et al., 2004; Xia, et al., 2004). Subsequently, population genomes and chromosome-level reference genome of the silkworm have been completed one after another (Xia, et al., 2008; Xia, et al., 2009; Xiang, et al., 2018; Kawamoto, et al., 2019). These genome resources play an important role in silkworm domestication history and functional genomics studies (Yang, et al., 2014; Xiang, et al., 2018; Zhu, et al., 2019; Li, et al., 2020; Wang, et al., 2020). Compared with the high-quality genome of the silkworm, the quality of transcripts is poor. The transcripts of *B. mori* so far were mainly assembled by using short reads of second-generation sequencing and expressed sequence tags (ESTs) (Shao, et al., 2012; Suetsugu, et al., 2013).

LncRNAs play important roles in most forms of life (Mercer, et al., 2009; Rinn and Chang 2020). For instance, Locasta migratoria PAHAL lncRNA positively regulates phenylalanine hydroxylase resulting in dopamine production in brain and modulates locust behavioral aggregation (Zhang, et al., 2020). Mouse

Braveheart lncRNA contributes to mesoderm and cardiac differentiation (Xue, et al., 2016). Two lncRNAs (roX1 and roX2) of *Drosophila melanogaster* take part in X-chromosome dosage compensation (Ilik, et al., 2013). Human NORAD lncRNA is required for the assembly of topoisomerase complex NARC1, which involve in maintaining genomic stability (Munschauer, et al., 2018). In *B. mori*, systematic characterizations of its lncRNA were studied based on next-generation sequencing (Wu et al., 2016; Zhou et al., 2016; Zhou et al., 2018). Moreover, some studies indicate that some lncRNA may play a role in 20E-induced autophagy (Qiao et al., 2021). However, prior lncRNA based on limited organs and short-read data which cannot fully explain the landscape of lncRNA in silkworms. Additionally, the functions *B. mori* lncRNAs so far remain poorly understood.

Transposable elements (TEs) are the largest component of most eukaryotic genomes, and function in the evolution of genome architecture and gene regulatory network (Feschotte 2008; Kapusta, et al., 2017; Cosby, et al., 2021). Recently, a study in tetrapod showed that a vast majority of transposase DNA binding domains fused to host regulatory domains through exon shuffling (Cosby, et al., 2021). A study in locusts revealed that TEs occupied ~20% of the locust transcriptome via its exonization (Jiang, et al., 2019). Past studies discovered that ~40% *B. mori* genome is composed of the known TEs (Osanai-Futahashi, et al., 2008; Xu, et al., 2013). Helitron families of *B. mori* genome contributed to 123 full-length cDNAs (Han, et al., 2013). Nevertheless, in *B. mori*, the contribution of the whole genome TEs to the transcriptome remains unclear.

In this work, we use PacBio Iso-Seq sequencing technology to generate high-quality full-length transcripts from 45 developmental stages of the silkworm. These transcripts are further divided into protein-coding genes and lncRNA based on protein-coding potential and lncRNA characteristics. Finally, the contribution of TEs to the transcripts is investigated.

MATERIALS AND METHODS

Sample Source and RNA Extraction

The silkworm (*Bombyx mori*) strain DaZao in this study was obtained from the Silkworm Gene Bank, Southwest University, China. This strain has been used to generate the reference genome (Kawamoto, et al., 2019). The silkworm was reared on fresh mulberry leaves at 25°C under 12°hours-light/12°hours-dark photoperiod. To obtain as many transcript isoforms as possible, we sampled almost all developmental stages of the silkworm (**Supplementary Table S1**). Each individual at the larval stage was dissected and then removed food residues in the intestinal to reduce the contamination of mulberry leaves and intestinal microorganisms. Total RNA was extracted using TRIzon Reagent kit No. CW0580S (CoWin Bioscience) and then treated with DNase I (TaKaRa) to remove genomic DNA.

RNA Library Preparation for SMRT Sequencing

The isolated total RNA (5 μg RNA, equally mixed from each developmental stage sample, **Supplementary Figure S1**) was used to synthesize cDNA by SMARTer cDNA Synthesis kit (Clontech). Three libraries (1–2 kb, 2–3 kb, and 3–6 kb) were constructed by using Pacific Biosciences DNA Template Prep Kit 2.0. Using the Pacific Bioscience RS II platform, we sequenced 8 SMRT cells including 3 cells for 1–2 kb, 3 cells for 2–3 kb libraries and 2 cells for 3–6 kb libraries.

RNA Polishing and Non-redundant Transcripts Identification

SMRT Analysis was used to obtain full-length transcripts (v2.3.0, https://www.pacb.com/). Where the polymerase reads with lengths small than 50°bp or quality less than 0.75 were discarded. The full-length reads were defined as containing 5′ and 3′ primers and ployA tail. While the other reads were defined as non-full length reads. The full-length non-chimeric (FLNC) transcripts were defined as full-length ROIs without any additional cDNA sequence. The consensus isoforms were obtained by using ICE (Iterative Clustering for Error Correction) with default parameters. The consensus isoforms were polished using Quiver (parameters: -hq_quiver_min_accuracy 0.99). High-quality consensus isoforms were classified with the criteria post-correction accuracy above 99%. Then the high-quality consensus isoforms were mapped to the reference genome using GMAP (Wu and Watanabe, 2005) (2017–11–15, parameters: -direction sense_force--cross_species--allow_close_indels 0). The sequences with identity less than 0.9 or coverage less than 0.85 were filtered using the pbtranscript-ToFU package (parameters: -min-trimmed-coverage = 0.85°min-identity = 0.9). The non-redundant high-quality consensus isoforms were obtained by merging sequences that differ only at the 5' terminal exon and the other exons were identical.

Alternative Splicing Analysis and Verification

Alternative splicing (AS) events including Intron Retention (IR), Exon skipping (ES), Alternative 5′ splice site (A5S), Alternative 3' splice site (A3S), and Mutually exclusive exon (MEE) were identified by the AStalavista tool (Foissac and Sammeth 2007). To validate alternative splicing events, five AS events were randomly selected to perform RT-PCR and sanger sequencing. The gene-specific primer was designed based on flanking regions of each splicing site using NCBI primer-Blast tool (https://www.ncbi.nlm.nih.gov/tools/primer-blast/). The five pairs of primers were listed in **Supplementary Table S2**.

Long Non-Coding RNA Identification, Open Reading Frames Identification, and Functional Annotation

All non-redundant transcripts were used to identify Long non-coding RNA (lncRNAs) and open reading frames (ORFs). The LncRNAs were identified by using Coding Potential Calculator (CPC2) (Kang, et al., 2017), Coding-Non-Coding Index (CNCI) (Sun, et al., 2013), Coding-Potential Assessment Tool (CPAT) (Wang, et al., 2013), and an ab initio lncRNA identification tool (LncAdeep) (Yang, et al., 2018). The ORFs were identified by using TransDecoder software (http://transdecoder.sourceforge.net/). All ORFs were annotated by

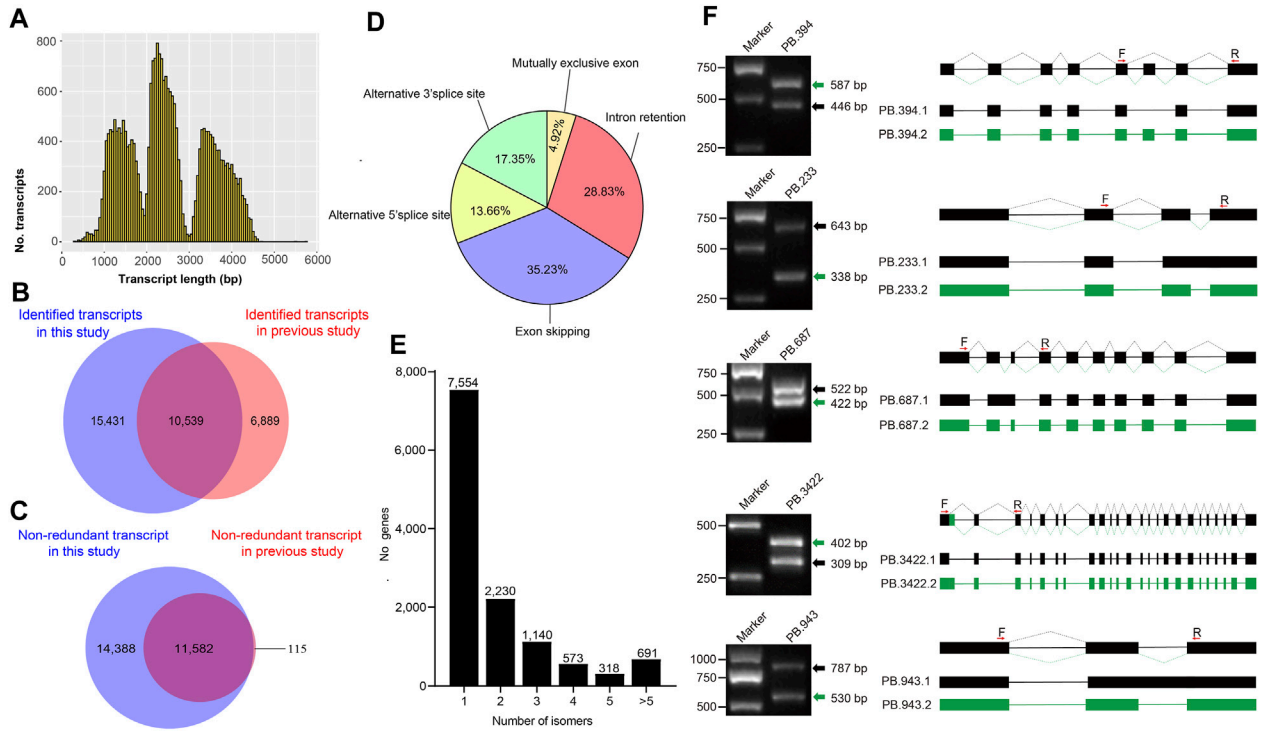

FIGURE 1 | Non-redundant transcripts and alternative splicing events (**A**) The distribution of all non-redundant transcripts length (**B**) The Venn diagram shows the comparative results of the transcript identified in this study and the previously reported transcript (**C**) The Venn diagram shows the comparative result between the non-redundant transcripts identified in this study and the previously reported transcripts in silkworm silk gland (**D**) Proportion of each alternative splicing type (**E**) Distribution of splice isoforms of genes (**F**) The random selected alternative splicing events were verified by Polymerase Chain Reaction (PCR) and Sanger sequencing. Primers (F represent forward primer and R represent reverse primer) were designed based on flanking regions of alternative splicing sites. The PCR product was run on 1% agarose gel and observed in UV light. For each gel, the green and black arrows indicate different splicing isoforms. The structures of the isoforms are showed on the right side of the figure. The filled boxes represent Exons and lines represent introns. Different isoforms are showed with different color box connected by lines. The red arrows upon the boxes represent the locations sites of the primers.

WEGO (Ye, et al., 2018), an online gene ontology website (http://wego.genomics.cn/), and based on COG (clusters of Orthologous Groups) database (Tatusov, et al., 2000).

Identification of Transcripts With Transposable Elements Exonization

The silkworm repeat library (Xu, et al., 2013) was used to identify transposon sequences in all non-redundant transcripts by RepeatMasker v4.1.0 (http://repeatmasker.org/) with RMBlast v2.9.0 (http://www.repeatmasker.org/RMBlast.html).

RESULTS

Transcriptome Sequencing Using SMRT

To obtain more transcripts of *Bombyx mori*, the total RNAs of whole-body samples from 45 developmental stages were extracted for long-reads sequencing (PacBio RS II platform) (**Supplementary Figure S1**). We obtained 777,701 polymerase reads and 7,100,417 subreads (15.88Gb clean data) by PacBio sequencing (**Supplementary Table S3**). Based on the number of full passes >0 and accuracy >0.75, a total of 516,326 reads of insert were generated (**Supplementary Table S4**). The average length of ROI from 1 to 2kb, 2–3kb, and 3–6kb cDNA size library were 1,736bp, 2,628bp, and 3,819bp, respectively (**Supplementary Table S4**). The observed distribution of ROI length for each cDNA size library was consistent with the expected (**Supplementary Figure S2A-C**). After filtering short-length ROIs (<300bp), there were 286,153 full-length ROI (containing 5′primer, 3′ primer, and polyA tail) and 196,776 non-full-length ROI (**Supplementary Table S5**). After filtering chimeric ROIs, 285,496 ROIs were identified as full-length non-chimeric (FLNC) reads constituted 55.3% of all ROIs (**Supplementary Figure S3A**). The distribution of FLNC length for each cDNA size library was shown in **Supplementary Figure S2D–E**. The average length of FLNC reads from 1 to 2kb, 2–3kb, and 3–6kb cDNA size library were 1,340bp, 2,364bp, and 3620bp, respectively (**Supplementary Table S5**).

Isoforms Clustering, Error Correction, and Alternative Splicing Analysis

To obtain consensus isoforms, all ROIs were clustered by iterative isoform-clustering (ICE) algorithm. We obtained 69,427 ICE

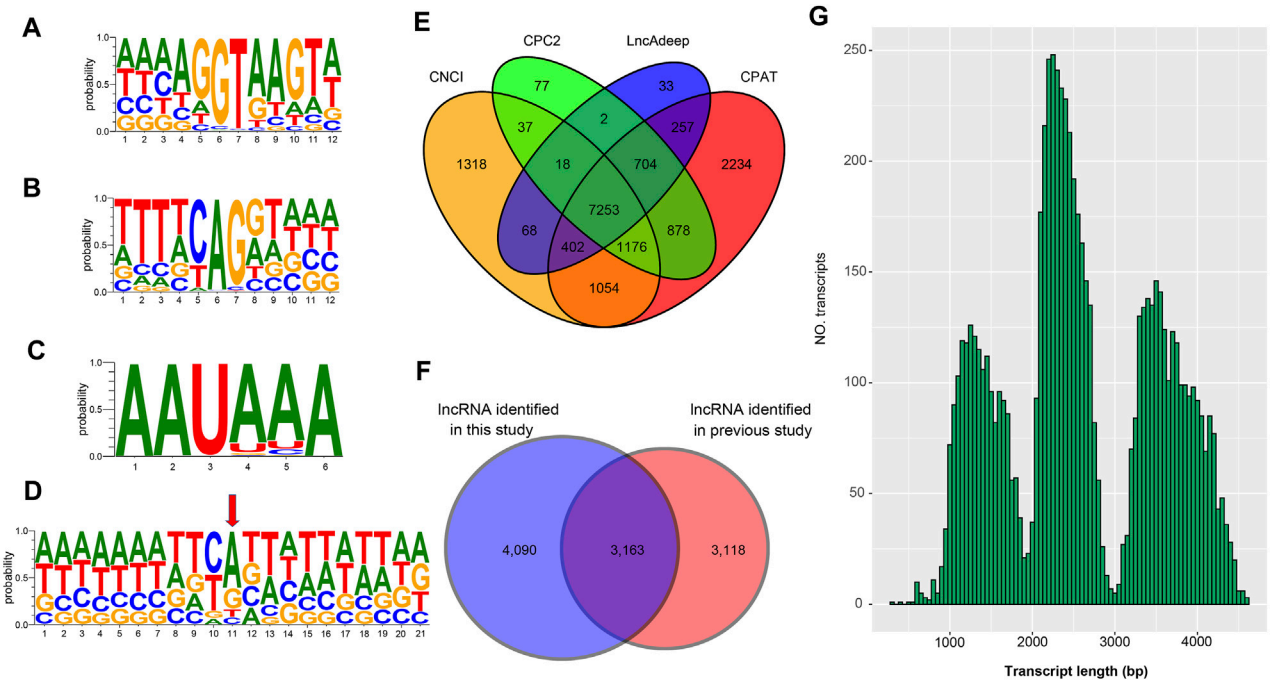

FIGURE 2 | Analysis of splice junctions, poly(A) signal sites, TSS site, and lncRNA (**A**) The mononucleotide variant distributions of the donor site (**B**) The mononucleotide variant distributions of the acceptor sites (**C**) The mononucleotide variant distributions of the poly(A) site (**D**) The mononucleotide variant distributions of the TSS sites **(E)** Venn diagram of lncRNA identification by four methods including CNCI, CPC2, LncAdeep and CPAT (**F**) The Venn diagram shows the comparative result of the lncRNA identified in this study and the previously reported LncRNAs (**G**) The length distribution of the lncRNAs identified in this study.

consensus isoforms. Among these isoforms, the number of isoforms from the sequence length of <1kb, 1–2kb, 2–3kb, 3–6kb, and >6kb were 2,249, 24,180, 26,438, 15,913, and 647, respectively (**Supplementary Table S6** and **Supplementary Figure S3B–F**). To filter low-quality isoforms, the consistent sequences of each cluster were corrected and evaluated by quiver program. We obtained 53,508 high-quality isoforms (HQs) with accuracy >99% and 15,919 low-quality isoforms (LQs) (**Supplementary Table S6**). The HQs constitute 25,970 non-redundant high-quality consensus isoforms (nrHQCIs) by merging high-quality isoforms that differ only at the 5' terminal exon and the other exons were identical (**Supplementary Table S7**). The distribution of nrHQCIs length was consistent with the expected size of the three libraries (**Figure 1A**). The average length of nrHQCIs was 2,504 bp, and there were 25,332 nrHQCIs (~97.5%) with sequence length of more than 1,000 bp. Compared with the previous transcripts assembled by short reads of second-generation sequencing and expressed sequence stags, 15,431 transcripts out of 25,970 nrHQCIs were novel (**Figure 1B** and **Supplementary Table S8**). Meanwhile, 6,889 transcripts were found in previous studies but not in this study. This phenomenon could be caused by that our sequencing depth was not enough, and some low-expressed genes may not be detected. We further compared our nrHQCIs with prior identified 11,697 nrHQCIs based on silk gland long-read transcriptome (Chen et al., 2020), 11,582 (~91%) 11,697 or prior nrHQCIs were overlapped in our 25,970 nrHQCIs (**Figure 1C**) and there were 14,388 new transcripts in our nrHQCIs.

To detect alternative splicing (AS) events, the AStalavista tool was used to identify Intron Retention (IR), Exon skipping (ES), Alternative 5′ splice site (A5S), Alternative 3′ splice site (A3S), and mutually exclusive exon (MEE) events. We detected a total of 18,416 AS events, and the majority of AS events being Exon skipping (**Figure 1D**). The distribution of isoform numbers of genes was shown in **Figure 1E**. Only one isoform was detected for 7,554 genes, and 4,952 genes produced two or more transcripts. To validate the accuracy of the detected AS events, five genes were randomly selected to perform RT-PCR and Sanger sequencing. The size of the gel band and the results of Sanger sequencing were consistent with the detected AS isoforms (**Figure 1F**).

To verify the 25,970 nrHQCIs, splicing sites, TSS sites, and AS events were verified by next-generation sequencing (NGS) data. The results revealed that 105,468 (91%) out of 116,399 splicing sites, 16,021 (~87%) of 18,416 AS events of 25,970 nrHQCIs were verified by NGS data. However, only 8,851 (~34%) transcription start sites (TSSs) were verified by NGS data (**Supplementary Table S8**). This phenomenon may be caused by the construction method of the PacBio library that did not take into account the 5′CAP integrity of RNA. Besides, the TSS sites identified by NGS data may not be accurate. Furthermore, we analyzed the consensus motif around donor and acceptor sites, polyadenylation sites, and (TSS). We found that the splicing donor and acceptor sites have conserved GT-AG consensus motif (**Figures 2A,B**), which is consistent with the other organism such as perennial ryegrass (Xie, et al., 2020), *B. malayi* (Nicolas J Wheeler, et al., 2020), and Medicago sativa L. (Chao, et al., 2019).

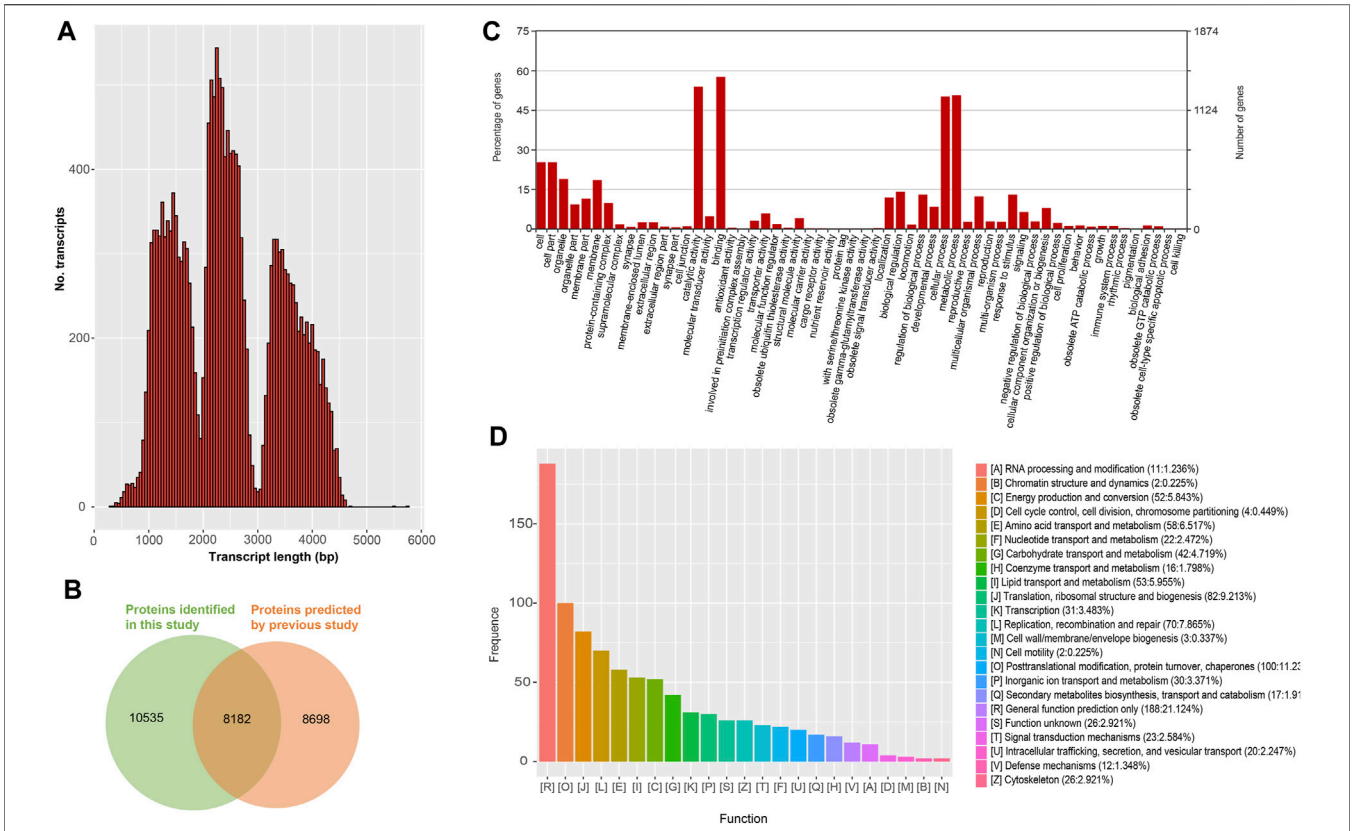

FIGURE 3 | Protein coding transcripts **(A)** The distribution of all protein coding transcripts length (**B**) Venn diagram of the protein coding transcripts identified in this study and previously reported protein coding transcripts (**C**) Gene Ontology (GO) terms distribution of all protein coding transcripts (**D**) Clusters of Orthologous Groups (COGs) of all protein coding transcripts.

In addition, the result of polyadenylation sites analysis indicates that "AAUAAA" has the highest frequency (**Figure 2C**). This result has highly corresponded with other studies (Wheeler, et al., 2020; Zhao et al., 2019).

Long Non-Coding RNA Identification and Open Reading Frames Prediction as Well as Protein-Coding Genes Annotation

To detect lncRNA in the 25,970 nrHQCIs, we used CNCI, CPC2, CPAT, and LncAdeep programs to identify lncRNAs. We took RNAs that all four software considered as lncRNAs as lncRNAs. A total of 7,253 nrHQCIs were detected as lncRNAs (**Figure 2E**). Compared with the previously reported lncRNAs identified by next-generation high-throughput sequencing technology (NGST), 4,090 lncRNA were novel (**Figure 2F** and **Supplementary Table S9**). The average length of identified lncRNAs was 2,585 bp, and the length distribution of lncRNAs was showed in **Figure 2G**.

The length distribution of the remained 18,717 nrHQCIs was showed in **Figure 3A**, and the protein-coding sequences were predicted by TransDecoder. Compared with the previously predicted protein, 10,535 predicted proteins were novel (**Figure 3B** and **Supplementary Table S10**). The 18,717 nrHQCIs were functionally annotated by searching the databases of COG, GO, KEGG, KOG, Pfam, Swissprot, eggnog, and NR. A total of 9,942 nrHQCIs were annotated (**Supplementary Table S11**). For instance, 5,569 nrHQCIs were annotated by GO analysis (**Figure 3C**). For COG annotation, 2,111 nrHQCIs were annotated, and the largest category was "General function prediction only" (**Figure 3D**). This finding is consistent with the observation in Medicago sativa L (Chao, et al., 2019).

Transposon Exonization

Transposable elements (TEs) constitute a significant component (~40%) of the *Bombyx mori* genome (Osanai-Futahashi, et al., 2008; Xu, et al., 2013). However, TEs contribute to the transcripts of *Bombyx mori* remain unclear. Here, transcripts with TEs were identified by homology searching. We found the number of transcripts with TEs was 11,671 (~45% of all identified nrHQCIs) containing 5,980 (~32% of all protein-coding transcripts) protein-coding transcripts and 5,691 (~79% of all identified lncRNAs) lncRNAs (**Figure 4A**). The proportion of TEs exonization in lncRNA was significantly higher (two-sample test of proportions, $p < 0.01$) than the proportion of TE

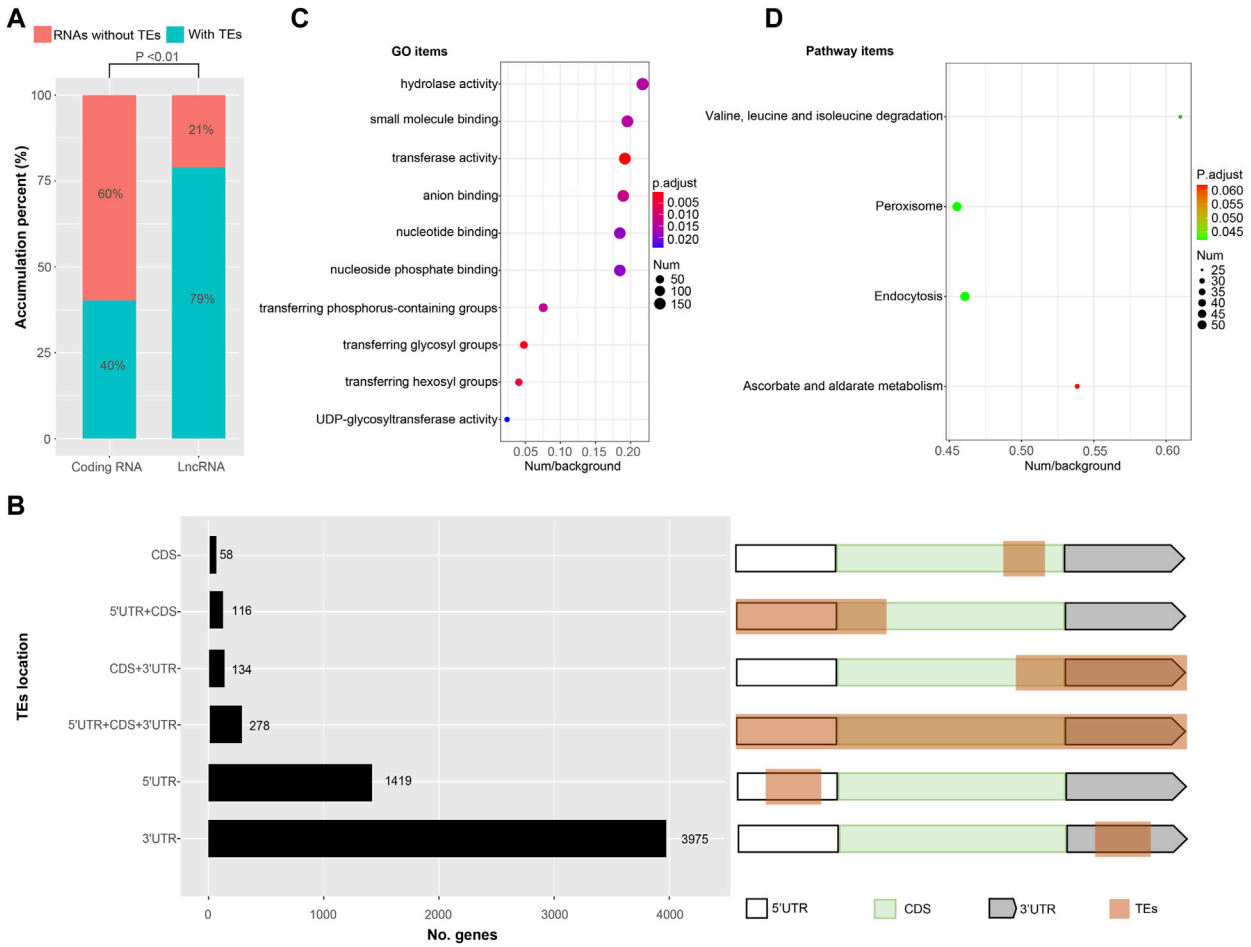

FIGURE 4 | Transcripts with transposable elements (TEs) exonization **(A)** The proportion of TEs exonization transcripts in protein coding transcripts and lncRNAs, respectively (**B**) The location distribution of TEs derived sequences in protein coding transcripts. White boxes represent 5′UTR, green boxes represent coding DNA sequences (CDS), gray boxes represent 3′UTR and brown box represent TEs derived sequences (**C**) GO enrichment results of TEs exonization transcripts (**D**) KEGG enrichment results of TEs exonization transcripts.

exonization in the protein-coding transcripts (**Figure 4A**). For 5,980 protein-coding transcripts with TEs, there were 278 transposase coding transcripts, and the remain 5,702 protein-coding transcripts composed of TEs fragments plus other non-transposase coding genes (**Figure 4B**). Where TEs contributed to the 5′UTRs of 1,535 protein-coding transcripts discarded transposase transcripts, to the 3′UTRs of 4,109, and to the ORFs of 308 (**Figure 4B** and **Supplementary Table S12**).

For 5,702 protein-coding transcripts with TEs, GO enrichment analysis showed that these genes enriched in 10 GO items including hydrolase activity, small molecule binding, transferase activity, anion binding, nucleotide binding, nucleoside phosphate binding, transferring phosphorus-containing groups, transferring glycosyl groups, transferring hexosyl groups and UDP-glycosyltransferase activity (**Figure 4C** and **Supplementary Figure S5**). KEGG enrichment analysis showed that the genes enriched in four pathway items including valine/leucine/isoleucine degradation, peroxisome, endocytosis, and ascorbate/alternate metabolism (**Figure 4D**).

DISCUSSION

B.mori was an important lepidopteran model system. Although the high-quality reference genome has been released in 2019 (Kawamoto et al., 2019), the transcriptome so far was obtained through assembling short reads resulting in poor-quality transcripts and incorrect genome annotation (Shao et al., 2012; Suetsugu et al., 2013). Recently, with the development of sequencing technology, PacBio sequencing, which has a profound advantage in long reads, has been applied widely to generate high-quality full-length transcripts in eukaryotes (Sharon et al., 2013; Wang et al., 2016; Jiang et al., 2019; Wang et al., 2019; Yang et al., 2020; Xu et al., 2021). In this study, we identified 25,970 high-quality transcripts in the silkworm by using PacBio sequencing technology. Compared with prior identified full-length transcripts based on EST and silk gland long-read transcriptom (Chen et al., 2020; Suetsugu et al., 2013), we identified 15,431 and 14,388 new transcripts, respectively,

which will improve the genome annotation, and promote functional genomic studies.

B.mori is the only truly domesticated insect. Long-term artificial selection has resulted in significant differences between the domesticated silkworm and its ancestors (Bombyx mandarina) in traits such as silk yield, behavior, body color and so on (Li, et al., 2020; Lu, et al., 2020; Wang, et al., 2020). Besides, more than 3,000 silkworm strains and >600 mutations with diverse phenotypes are available worldwide which are generated through spontaneous mutation, artificial mutagenesis, or breeding (Nagaraju et al., 2000; Goldsmith, et al., 2005; Furdui, et al., 2014). As far as we know, more than 60 mutations so far have been deciphered. For instance, the twin-spot markings on *B. mori* larval are caused by periodic Wnt1 expression (Yamaguchi, et al., 2013). The Toll ligand Spz-3 controls the black stripe of each segment of the silkworm (Kondo, et al., 2017). The BmGlcNase1 gene is involved in the synthesis of sericin (Li, et al., 2020). However, a great number of control genes of the traits of domestication and mutation remain unclear. The high-quality reference transcriptome of the silkworm obtained in this study will facilitate the deciphering of these traits.

LncRNAs are widespread in eukaryotes and play an important role in gene-regulatory networks (Kopp and Mendell, 2018). Prior studies of *B. mori* identified 6,281 possibly lncRNAs through second-generation sequencing technology and found two lncRNAs that could be related to silk protein translation (Zhou et al., 2016; Zhou et al., 2018). However, the function of the vast majority of *B. mori* lncRNAs remains unknown. Moreover, these lncRNAs maybe not accurate due to the limitation of short reads. Here, we identified 7,253 high-quality lncRNAs with 4,090 completely novel lncRNAs through long-read sequencing technology. The studies of the function and biological relevance of these lncRNAs are interesting topics in the future.

Transposable elements (TEs) are mobile elements and are powerful mutagens that play important roles in eukaryotic genome evolution and adaptation as well as disease (Lisch 2013; Chuong et al., 2017; Payer and Burns, 2019). TEs occupied ~40% of the *B. mori* genome (Osanai-Futahashi et al., 2008; Xu et al., 2013). Moreover, past studies revealed that a large number of traits of domestication and mutation are caused by TEs transposition. For example, the trait of developmental uniformity of *B. mori* is attributed to a non-LTR transposon (Taguchi) inserted in upstream of the silkworm ecdysone oxidase (Sun et al., 2014). A Tc1-mariner transposon inserted in the upstream of tyrosine hydroxylase is responsible for the sex-linked chocolate (sch) mutant of *B. mori* (Liu et al., 2010). A transposon-associated genomic deletion is involved in the trait of white cocoon (Sakudoh et al., 2007). Howbeit, the mechanism of the large majority of silkworm traits are unknown. In this work, we identified 11,671 transcripts with TEs exonization. Which has general implication for understanding the evolution of genes. Furthermore, whether transposon can alter gene structure, function or expression through its exonization to control the trait of *B. mori* is another interesting question in future.

DATA AVAILABILITY STATEMENT

The PacBio sequencing data has already been submitted to CNGBdb (China National GeneBank DataBase. The sequencing ID is CNP0001781. The information of experimental samples is accessible with the sample ID CNS0360521). Details for the project can be searched through the project ID CNP0=001781. The SRA data has also been released in NCBI (National Center for Biotechnology Information). The Biosample and Bioproject ID are SAMN20348872 and PRJNA748960.

AUTHOR CONTRIBUTIONS

MH and FD designed this study. MH and ZD analyzed the data. MH and ZD wrote and edited the manuscript. JR and ZD conducted lab work and conducted silkworm rearing with HL assisted with bioinformatics. FD supervised and guided the research.

FUNDING

This work was supported by grants from the National Natural Science Foundation of China (No. U20A2058, No. 31830094 TO and No. 31401106) for samples preparation fees, the Natural Science Foundation of Chongqing, China (No. cstc2020jcyj-msxmX0450) for publication fees, the Fundamental Research Funds for the Central Universities in China (No. XDJK 2019C009) and the Funds of China Agriculture Research System (No. CARS-18-ZJ0102) for sequencing fees.

ACKNOWLEDGMENTS

We thank all members of Dai's group for their useful comments.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.704162/full#supplementary-material

REFERENCES

Chao, Y., Yuan, J., Guo, T., Xu, L., Mu, Z., and Han, L. (2019). Analysis of Transcripts and Splice Isoforms in Medicago Sativa L. By Single-Molecule Long-Read Sequencing. *Plant Mol. Biol.* 99, 219–235. doi:10.1007/s11103-018-0813-y

Chen, T., Sun, Q., Ma, Y., Zeng, W., Liu, R., Qu, D., et al. (2020). A Transcriptome Atlas of Silkworm Silk Glands Revealed by PacBio Single-Molecule Long-Read Sequencing. *Mol. Genet. Genomics* 295, 1227–1237. doi:10.1007/s00438-020-01691-9

Chuong, E. B., Elde, N. C., and Feschotte, C. (2017). Regulatory Activities of Transposable Elements: from Conflicts to Benefits. *Nat. Rev. Genet.* 18, 71–86. doi:10.1038/nrg.2016.139

Cosby, R. L., Judd, J., Zhang, R., Zhong, A., Garry, N., Pritham, E. J., et al. (2021). Recurrent Evolution of Vertebrate Transcription Factors by Transposase Capture. *Science* 371, eabc6405. doi:10.1126/science.abc6405

Feschotte, C. (2008). Transposable Elements and the Evolution of Regulatory Networks. *Nat. Rev. Genet.* 9, 397–405. doi:10.1038/nrg2337

Foissac, S., and Sammeth, M. (2007). ASTALAVISTA: Dynamic and Flexible Analysis of Alternative Splicing Events in Custom Gene Datasets. *Nucleic Acids Res.* 35, W297–W299. doi:10.1093/nar/gkm311

Furdui, E. M., Mărghitaş, L. A., Dezmirean, D. S., Paşca, I., Pop, I. F., Erler, S., et al. (2014). Genetic Characterization of *Bombyx mori* (Lepidoptera: Bombycidae) Breeding and Hybrid Lines with Different Geographic Origins. *J. Insect Sci.* 14, 211. doi:10.1093/jisesa/ieu073

Goldsmith, M. R., Shimada, T., and Abe, H. (2005). The Genetics and Genomics of the Silkworm,bombyx Mori. *Annu. Rev. Entomol.* 50, 71–100. doi:10.1146/annurev.ento.50.071803.130456

Han, M.-J., Shen, Y.-H., Xu, M.-S., Liang, H.-Y., Zhang, H.-H., and Zhang, Z. (2013). Identification and Evolution of the Silkworm Helitrons and Their Contribution to Transcripts. *DNA Res.* 20, 471–484. doi:10.1093/dnares/dst024

Ilik, I. A., Quinn, J. J., Georgiev, P., Tavares-Cadete, F., Maticzka, D., Toscano, S., et al. (2013). Tandem Stem-Loops in roX RNAs Act Together to Mediate X Chromosome Dosage Compensation in Drosophila. *Mol. Cel* 51, 156–173. doi:10.1016/j.molcel.2013.07.001

Jiang, F., Zhang, J., Liu, Q., Liu, X., Wang, H., He, J., et al. (2019). Long-read Direct RNA Sequencing by 5'-Cap Capturing Reveals the Impact of Piwi on the Widespread Exonization of Transposable Elements in Locusts. *Rna Biol.* 16, 950–959. doi:10.1080/15476286.2019.1602437

Kang, Y.-J., Yang, D.-C., Kong, L., Hou, M., Meng, Y.-Q., Wei, L., et al. (2017). CPC2: a Fast and Accurate Coding Potential Calculator Based on Sequence Intrinsic Features. *Nucleic Acids Res.* 45, W12–W16. doi:10.1093/nar/gkx428

Kapusta, A., Suh, A., and Feschotte, C. (2017). Dynamics of Genome Size Evolution in Birds and Mammals. *Proc. Natl. Acad. Sci. USA* 114, E1460–E1469. doi:10.1073/pnas.1616702114

Kawamoto, M., Jouraku, A., Toyoda, A., Yokoi, K., Minakuchi, Y., Katsuma, S., et al. (2019). High-quality Genome Assembly of the Silkworm, *Bombyx mori*. *Insect Biochem. Mol. Biol.* 107, 53–62. doi:10.1016/j.ibmb.2019.02.002

Kondo, Y., Yoda, S., Mizoguchi, T., Ando, T., Yamaguchi, J., Yamamoto, K., et al. (2017). Toll Ligand Spätzle3 Controls Melanization in the Stripe Pattern Formation in Caterpillars. *Proc. Natl. Acad. Sci. USA* 114, 8336–8341. doi:10.1073/pnas.1707896114

Kopp, F., and Mendell, J. T. (2018). Functional Classification and Experimental Dissection of Long Noncoding RNAs. *Cell* 172, 393–407. doi:10.1016/j.cell.2018.01.011

Li, C., Tong, X., Zuo, W., Hu, H., Xiong, G., Han, M., et al. (2020). The Beta-1, 4-N-Acetylglucosaminidase 1 gene, Selected by Domestication and Breeding, Is Involved in Cocoon Construction of *Bombyx mori*. *Plos Genet.* 16, e1008907. doi:10.1371/journal.pgen.1008907

Lisch, D. (2013). How Important Are Transposons for Plant Evolution? *Nat. Rev. Genet.* 14, 49–61. doi:10.1038/nrg3374

Liu, C., Yamamoto, K., Cheng, T.-C., Kadono-Okuda, K., Narukawa, J., Liu, S.-P., et al. (2010). Repression of Tyrosine Hydroxylase Is Responsible for the Sex-Linked Chocolate Mutation of the Silkworm, *Bombyx mori*. *Proc. Natl. Acad. Sci.* 107, 12980–12985. doi:10.1073/pnas.1001725107

Lu, K., Liang, S., Han, M., Wu, C., Song, J., Li, C., et al. (2020). Flight Muscle and Wing Mechanical Properties Are Involved in Flightlessness of the Domestic Silkmoth, *Bombyx mori*. *Insects* 11, 220. doi:10.3390/insects11040220

Mercer, T. R., Dinger, M. E., and Mattick, J. S. (2009). Long Non-coding RNAs: Insights into Functions. *Nat. Rev. Genet.* 10, 155–159. doi:10.1038/nrg2521

Mita, K., Kasahara, M., Sasaki, S., Nagayasu, Y., Yamada, T., Kanamori, H., et al. (2004). The Genome Sequence of Silkworm, *Bombyx mori*. *DNA Res.* 11, 27–35. doi:10.1093/dnares/11.1.27

Munschauer, M., Nguyen, C. T., Sirokman, K., Hartigan, C. R., Hogstrom, L., Engreitz, J. M., et al. 2018. The NORAD lncRNA Assembles a Topoisomerase Complex Critical for Genome Stability. *Nature* 561:132, 136. doi:10.1038/s41586-018-0453-z

Nagaraju, J. G., Klimenko, V., and Couble, P.. 2000. *The Silkworm* Bombyx mori: *A Model Genetic System*. Editor E. Reeves London, United Kingdom: Encyclopedia of Genetics. 219–239.

Osanai-Futahashi, M., Suetsugu, Y., Mita, K., and Fujiwara, H. (2008). Genome-wide Screening and Characterization of Transposable Elements and Their Distribution Analysis in the Silkworm, *Bombyx mori*. *Insect Biochem. Mol. Biol.* 38, 1046–1057. doi:10.1016/j.ibmb.2008.05.012

Payer, L. M., and Burns, K. H. (2019). Transposable Elements in Human Genetic Disease. *Nat. Rev. Genet.* 20, 760–772. doi:10.1038/s41576-019-0165-8

Qiao, H., Wang, J., Wang, Y., Yang, J., Wei, B., Li, M., et al. (2021). Transcriptome Analysis Reveals Potential Function of Long Non-coding RNAs in 20-hydroxyecdysone Regulated Autophagy in *Bombyx mori*. *BMC Genomics*. doi:10.1186/s12864-021-07692-1

Rinn, J. L., and Chang, H. Y. (2020). Long Noncoding RNAs: Molecular Modalities to Organismal Functions. *Annu. Rev. Biochem.* 89, 283–308. doi:10.1146/annurev-biochem-062917-012708

Sakudoh, T., Sezutsu, H., Nakashima, T., Kobayashi, I., Fujimoto, H., Uchino, K., et al. (2007). Carotenoid Silk Coloration Is Controlled by a Carotenoid-Binding Protein, a Product of the Yellow Blood Gene. *Proc. Natl. Acad. Sci.* 104, 8941–8946. doi:10.1073/pnas.0702860104

Shao, W., Zhao, Q.-Y., Wang, X.-Y., Xu, X.-Y., Tang, Q., Li, M., et al. (2012). Alternative Splicing and Trans-splicing Events Revealed by Analysis of the *Bombyx mori* Transcriptome. *RNA* 18, 1395–1407. doi:10.1261/rna.029751.111

Sharon, D., Tilgner, H., Grubert, F., and Snyder, M. (2013). A Single-Molecule Long-Read Survey of the Human Transcriptome. *Nat. Biotechnol.* 31, 1009–1014. doi:10.1038/nbt.2705

Suetsugu, Y., Futahashi, R., Kanamori, H., Kadono-Okuda, K., Sasanuma, S.-i., Narukawa, J., et al. (2013). Large Scale Full-Length cDNA Sequencing Reveals a Unique Genomic Landscape in a Lepidopteran Model Insect, *Bombyx mori*. *G3-Genes Genomes Genet.* 3, 1481–1492. doi:10.1534/g3.113.006239

Sun, L., Luo, H., Bu, D., Zhao, G., Yu, K., Zhang, C., et al. (2013). Utilizing Sequence Intrinsic Composition to Classify Protein-Coding and Long Non-coding Transcripts. *Nucleic Acids Res.* 41, e166. doi:10.1093/nar/gkt646

Sun, W., Shen, Y.-H., Han, M.-J., Cao, Y.-F., and Zhang, Z. (2014). An Adaptive Transposable Element Insertion in the Regulatory Region of the EO Gene in the Domesticated Silkworm, *Bombyx mori*. *Mol. Biol. Evol.* 31, 3302–3313. doi:10.1093/molbev/msu261

Tatusov, R. L., Galperin, M. Y., Natale, D. A., and Koonin, E. V. (2000). The COG Database: a Tool for Genome-Scale Analysis of Protein Functions and Evolution. *Nucleic Acids Res.* 28, 33–36. doi:10.1093/nar/28.1.33

Wang, B., Tseng, E., Regulski, M., Clark, T. A., Hon, T., Jiao, Y., et al. (2016). Unveiling the Complexity of the maize Transcriptome by Single-Molecule Long-Read Sequencing. *Nat. Commun.* 7, 11708. doi:10.1038/ncomms11708

Wang, K., Wang, D., Zheng, X., Qin, A., Zhou, J., Guo, B., et al. (2019). Multi-strategic RNA-Seq Analysis Reveals a High-Resolution Transcriptional Landscape in Cotton. *Nat. Commun.* 10, 4714. doi:10.1038/s41467-019-12575-x

Wang, L., Park, H. J., Dasari, S., Wang, S., Kocher, J. P., and Li, W. (2013). CPAT: Coding-Potential Assessment Tool Using an Alignment-free Logistic Regression Model. *Nucleic Acids Res.* 41, e74. doi:10.1093/nar/gkt006

Wang, M., Lin, Y. J., Zhou, S. Y., Cui, Y., Feng, Q. L., Yan, W., et al. (2020). Genetic Mapping of Climbing and Mimicry: Two Behavioral Traits Degraded during Silkworm Domestication. *Front. Genet.* 11, 566961. doi:10.3389/fgene.2020.566961

Wheeler, N. J., Airs, P. M., and Zamanian, M. (2020). Long-read RNA Sequencing of Human and Animal Filarial Parasites Improves Gene Models and Discovers

Operons. *Plos Negl. Trop. Dis.* 14 (11), e0008869. doi:10.1371/journal.pntd.0008869

Wu, T. D., and Watanabe, C. K. (2005). GMAP: a Genomic Mapping and Alignment Program for mRNA and EST Sequences. *Bioinformatics* 21, 1859–1875. doi:10.1093/bioinformatics/bti310

Wu, Y., Cheng, T., Liu, C., Liu, D., Zhang, Q., Long, R., et al. (2016). Systematic Identification and Characterization of Long Non-coding RNAs in the Silkworm, *Bombyx mori*. *PLOS ONE* 11 (1), e0147147. doi:10.1371/journal.pone.0147147

Xia, Q., Guo, Y., Zhang, Z., Li, D., Xuan, Z., Li, Z., et al. (2009). Complete Resequencing of 40 Genomes Reveals Domestication Events and Genes in Silkworm (Bombyx). *Science* 326, 433–436. doi:10.1126/science.1176620

Xia, Q. Y., Wang, J., Zhou, Z. Y., Li, R. Q., Fan, W., Cheng, D. J., et al. (2008). The Genome of a Lepidopteran Model Insect, the Silkworm *Bombyx mori*. *Insect Biochem. Mol. Biol.* 38, 1036–1045. doi:10.1016/j.ibmb.2008.11.004

Xia, Q. Y., Zhou, Z. Y., Lu, C., Cheng, D. J., Dai, F. Y., Li, B., et al. (2004). A Draft Sequence for the Genome of the Domesticated Silkworm (*Bombyx mori*). *Science* 306, 1937–1940. doi:10.1126/science.1102210

Xiang, H., Liu, X., Li, M., Zhu, Y. n., Wang, L., Cui, Y., et al. (2018). The Evolutionary Road from Wild Moth to Domestic Silkworm. *Nat. Ecol. Evol.* 2, 1268–1279. doi:10.1038/s41559-018-0593-4

Xie, L., Teng, K., Tan, P., Chao, Y., Li, Y., Guo, W., et al. (2020). PacBio Single-Molecule Long-Read Sequencing Shed New Light on the Transcripts and Splice Isoforms of the Perennial Ryegrass. *Mol. Genet. Genomics* 295, 475–489. doi:10.1007/s00438-019-01635-y

Xu, D., Yang, H., Zhuo, Z., Lu, B., Hu, J., and Yang, F. (2021). Characterization and Analysis of the Transcriptome in Opisina Arenosella from Different Developmental Stages Using Single-Molecule Real-Time Transcript Sequencing and RNA-Seq. *Int. J. Biol. Macromolecules* 169, 216–227. doi:10.1016/j.ijbiomac.2020.12.098

Xu, H. E., Zhang, H. H., Xia, T., Han, M. J., Shen, Y. H., and Zhang, Z. (2013). BmTEdb: A Collective Database of Transposable Elements in the Silkworm Genome. *Database (Oxford)* 2013, bat055. doi:10.1093/database/bat055

Xue, Z., Hennelly, S., Doyle, B., Gulati, A. A., Novikova, I. V., Sanbonmatsu, K. Y., et al. (2016). A G-Rich Motif in the lncRNA Braveheart Interacts with a Zinc-Finger Transcription Factor to Specify the Cardiovascular Lineage. *Mol. Cel* 64, 37–50. doi:10.1016/j.molcel.2016.08.010

Yamaguchi, J., Banno, Y., Mita, K., Yamamoto, K., Ando, T., and Fujiwara, H. (2013). Periodic Wnt1 Expression in Response to Ecdysteroid Generates Twin-Spot Markings on Caterpillars. *Nat. Commun.* 4, 1857. doi:10.1038/ncomms2778

Yang, C., Yang, L., Zhou, M., Xie, H., Zhang, C., Wang, M. D., et al. (2018). LncADeep: Anab initiolncRNA Identification and Functional Annotation Tool Based on Deep Learning. *Bioinformatics* 34, 3825–3834. doi:10.1093/bioinformatics/bty428

Yang, H., Xu, D., Zhuo, Z., Hu, J., and Lu, B. (2020). SMRT Sequencing of the Full-Length Transcriptome of the Rhynchophorus Ferrugineus (Coleoptera: Curculionidae). *Peerj* 8, e9133. doi:10.7717/peerj.9133

Yang, S. Y., Han, M. J., Kang, L. F., Li, Z. W., Shen, Y. H., and Zhang, Z. (2014). Demographic History and Gene Flow during Silkworm Domestication. *BMC Evol. Biol.* 14, 185. doi:10.1186/s12862-014-0185-0

Ye, J., Zhang, Y., Cui, H., Liu, J., Wu, Y., Cheng, Y., et al. (2018). WEGO 2.0: a Web Tool for Analyzing and Plotting GO Annotations, 2018 Update. *Nucleic Acids Res.* 46, W71–W75. doi:10.1093/nar/gky400

Zhang, X., Xu, Y., Chen, B., and Kang, L. (2020). Long Noncoding RNA PAHAL Modulates Locust Behavioural Plasticity through the Feedback Regulation of Dopamine Biosynthesis. *Plos Genet.* 16, e1008771. doi:10.1371/journal.pgen.1008771

Zhao, Z., Wu, X., Ji, G., Liang, C., and Li, Q. Q. (2019). Genome-Wide Comparative Analyses of Polyadenylation Signals in Eukaryotes Suggest a Possible Origin of the AAUAAA Signal. *Int. J. Mol. Sci.* 20 (4), 958. doi:10.3390/ijms20040958

Zhou, Q.-Z., Fang, S.-M., Zhang, Q., Yu, Q.-Y., and Zhang, Z. (2018). Identification and Comparison of Long Non-coding RNAs in the Silk Gland between Domestic and Wild Silkworms. *Insect Sci.* 25, 604–616. doi:10.1111/1744-7917.12443

Zhou, Q. Z., Zhang, B., Yu, Q. Y., and Zhang, Z. (2016). BmncRNAdb: a Comprehensive Database of Non-coding RNAs in the Silkworm, *Bombyx mori*. *BMC Bioinformatics* 17, 370. doi:10.1186/s12859-016-1251-y

Zhu, Y. N., Wang, L. Z., Li, C. C., Cui, Y., Wang, M., Lin, Y. J., et al. (2019). Artificial Selection on Storage Protein 1 Possibly Contributes to Increase of Hatchability during Silkworm Domestication. *Plos Genet.* 15, e1007616. doi:10.1371/journal.pgen.1007616

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.