PLOS | ONE

RESEARCH ARTICLE

Systematic Identification and Characterization of Long Non-Coding RNAs in the Silkworm, *Bombyx mori*

Yuqian Wu[1,2], Tingcai Cheng[2], Chun Liu[2], Duolian Liu[2], Quan Zhang[2], Renwen Long[2], Ping Zhao[2], Qingyou Xia[2]*

1 School of Life Sciences, Chongqing University, Chongqing 400044, China, **2** State Key Laboratory of Silkworm Genome Biology, Southwest University, Chongqing 400715, China

* xiaqy@swu.edu.cn

Citation: 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

Editor: Erjun Ling, Institute of Plant Physiology and Ecology, CHINA

Received: October 1, 2015

Accepted: November 22, 2015

Published: January 15, 2016

Data Availability Statement: All raw data files are available from the NCBI Short Read Archive database (accession number PRJNA284192).

Funding: The work was supported by National Basic Research Program of China (2012CB114600), National Natural Science Foundation of China (31372380), key program of the National Natural Science Foundation of China (31530071) and Chongqing Fundamental and Advanced Research Projects (CSTC2014JCYJA80004). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

Long noncoding RNAs (lncRNAs) are emerging as important regulators in various biological processes. However, to date, no systematic characterization of lncRNAs has been reported in the silkworm *Bombyx mori*. In the present study, we generated eighteen RNA-seq datasets with relatively high depth. Using an in-house designed lncRNA identification pipeline, 11,810 lncRNAs were identified for 5,556 loci. Among these lncRNAs, 474 transcripts were intronic lncRNAs (ilncRNAs), 6,250 transcripts were intergenic lncRNAs (lincRNAs), and 5,086 were natural antisense lncRNAs (lncNATs). Compared with protein-coding mRNAs, silkworm lncRNAs are shorter in terms of full length but longer in terms of exon and intron length. In addition, lncRNAs exhibit a lower level of sequence conservation, more repeat sequences overlapped and higher tissue specificity than protein-coding mRNAs in the silkworm. We found that 69 lncRNA transcripts from 33 gene loci may function as miRNA precursors, and 104 lncRNA transcripts from 72 gene loci may act as competing endogenous RNAs (ceRNAs). In total, 49.47% of all gene loci (2,749/5,556) for which lncRNAs were identified showed sex-biased expression. Co-expression network analysis resulted in 19 modules, 12 of which revealed relatively high tissue specificity. The highlighted darkgoldenrod module was specifically associated with middle and posterior silk glands, and the hub lncRNAs within this module were co-expressed with proteins involved in translation, translocation, and secretory processes, suggesting that these hub lncRNAs may function as regulators of the biosynthesis, translocation, and secretion of silk proteins. This study presents the first comprehensive genome-wide analysis of silkworm lncRNAs and provides an invaluable resource for genetic, evolutionary, and genomic studies of *B. mori*.

Introduction

Long non-coding RNAs (lncRNAs) have been arbitrarily defined as non-coding RNAs greater than 200 nucleotides in length. A generally used criterion for distinguishing from non-coding

Competing Interests: The authors have declared that no competing interests exist.

RNAs and protein-coding RNAs, is that the former do not encode an open reading frame (ORF) of more than 100 amino acids (aa) [1]. Similar to mRNAs, lncRNAs are subject to post-transcriptional modifications such as capping, polyadenylation, and splicing [2]. Putting protein-coding genes as reference, lncRNAs are transcribed from intronic or intergenic regions of the genome in a sense or antisense orientation. During the last decade, lncRNAs have attracted much attention due to their important roles in regulating complex biological processes. LncRNAs are capable of interacting with DNA and/or proteins to generate modular scaffolds for transcriptional gene silencing, alternative pre-mRNA splicing, direct modification of chromatin and chromosome architecture, and protein degradation [3, 4].

Unlike microRNAs, lncRNAs are generally less conserved in terms of nucleotide sequence across phylogenetically related species, making it difficult to detect lncRNAs by sequence similarity searching [5]. Next-generation sequencing technologies have emerged as powerful tools for exploring whole-genome lncRNA. A human transcriptome analysis of thousands of tumors, normal tissues, and cell lines yielded 90,013 expressed genes, of which 68% (58,648) were classified as lncRNAs [6]. In addition, more than 8,000 lncRNAs have been identified in mouse testis during postnatal testis development [7]. Although less well characterized than vertebrates and plants, to our knowledge, thousands of lncRNAs have been identified in three insect species [6, 8–17]. In the fruit fly (*Drosophila*), up to 4,000 candidate lncRNA genes were identified, resulting in a catalog of about 1,875 lncRNAs producing 3,085 transcripts [18]. Approximately 3,008 genic and 6,855 intergenic lncRNAs (lincRNAs) were identified by deep midgut transcriptome annotation [14]. In *Anopheles gambiae*, 2,949 lncRNAs have been identified in samples representing multiple life stages using deep RNA-seq technology [19]. More recently, Jayakodi identified 1,514 lincRNAs in *Apis mellifera* and 2,470 lincRNAs in *Apis cerana*, and investigated their response to viral infection [15]. Functionally, several lncRNAs have been experimentally validated as important regulators of gene regulation, dosage compensation, development, and behavior in the fruit fly. For instance, lncRNA hswω-n transcript forms perinuclear omega-speckles in nuclei in response to heat shock [20]. Two male-specific lncRNAs, *roX1* and *roX2*, present in the male-specific lethal (MSL) protein complex play pivotal roles in targeting chromosome-wide modification for dosage compensation in *Drosophila* [21]. *Yellow-achaete intergenic RNA* (*yar*) lncRNA serves as a regulator of yellow and achaete gene transcription to alter sleep regulation in the context of a normal circadian rhythm in the fruit fly [22]. The neural-specific *Drosophila* lncRNA *CRG* (CASK regulatory gene) participates in locomotion and climbing by enhancing its neighboring *CASK* expression via the recruitment of RNA polymerase II to the *CASK* promoter regions [23]. Another example of a behavior-related *Drosophila* lncRNA is *Sphinx*, whose 5'-flanking 300-bp sequence is conserved across *Drosophila* species. The *Sphinx* lncRNA is involved in regulating courtship behavior [24]. In *Apis mellifera*, only six lncRNAs have been experimentally confirmed to date, of which four (*Nb-1*, *Ks-1*, *AncR-1*, and *kakusei*) are preferentially expressed in the brain and related to behavior [25–29] and the other two (*lncov1* and *lncov2*) are expressed in the ovaries. *lncov1*, which is overexpressed in the ovaries of worker bees, is associated with transgressive ovary size [30].

The silkworm, which is a lepidopteron model insect of economic importance, has huge value for studying the fundamental mechanisms of non-coding gene regulation [31]. To date, efforts have been made to study non-coding RNA of silkworm. Genome-wide analysis has revealed a landscape of microRNAs [32, 33], snoRNA [34], and PIWI-interacting RNAs [35]. However, silkworm lncRNAs remain poorly characterized. To the best of our knowledge, only one silkworm lncRNA (*Fben-1*), which is preferentially expressed in the female brain, has been reported to date [36]. The systematic screening of potential lncRNAs in the silkworm genome has not yet been reported. In this study, we performed deep transcriptome sequencing of 18 tissue samples collected from fifth instar silkworm larvae. By combining our data with 2

additional public silkworm RNA-seq datasets (which represented 3 tissue samples), we systematically identified lncRNAs at the whole-genome level. Our results indicate that a large number of silkworm lncRNAs show relatively low expression levels, high spatial specificity, and low levels of sequence conservation compared with silkworm protein-coding mRNAs. These lncRNAs may serve as miRNA precursors or ceRNAs, and are suspected to be involved in miRNA regulatory pathways. In addition, our results reveal that a proportion of lncRNAs in the silk gland gene co-expression network core may participate in the biosynthesis, translocation, and secretion of silk proteins.

Materials and Methods

Silkworm rearing and tissue collection

The silkworm strain *Dazao* were obtained from the Silkworm Gene Bank of Southwest University, Chongqing, China. All larvae were reared at 25°C, 60% relative humidity, with a 16:8 h light-dark regimen, and fed with mulberry leaves. Sexed tissues including the anterior silk gland (ASG), the anterior section of middle silk gland (AMSG), the middle section of middle silk gland (MMSG), the posterior section of middle silk gland (PMSG), the posterior silk gland (PSG), gonad (testis/ovary), fat body, Malpighian tubule (MpT), and brain were dissected from day 3 fifth instar male and female larvae, respectively. All samples were frozen immediately in liquid nitrogen and stored at -80°C until use.

RNA extraction, library construction, and sequencing

Total RNA was extracted from silkworm tissues using the TRIzol reagent (Invitrogen) and further purified with the RNeasy kit (Qiagen). The integrity and quality of RNA were assessed using the Agilent 2100 Bioanalyzer (Agilent technologies). For non-strand-specific libraries, mRNAs were selected using oligo(dT) magnetic beads (Invitrogen), fragmented, and used to synthesize cDNA according to the TruSeq RNA Sample Preparation v2 Guide (Illumina). For strand-specific libraries, ribosomal RNA was depleted using Ribo-Zero rRNA removal beads. Then, the total RNA was purified and fragmented in fragmentation buffer. Next, the strand-specific sequencing libraries were constructed using TruSeq Stranded Total RNA Sample Preparation kits (Illumina, San Diego, CA). Libraries were sequenced on the Hiseq2000 system (Illumina, San Diego, CA). All RNA sequencing data produced in present study have been deposited in NCBI Short Read Archive (http://www.ncbi.nlm.nih.gov/sra/) and can be accessed under the SRA accession number: PRJNA284192.

Public available RNA-seq data

RNA-seq data from early-sexed embryonic stages of silkworm were obtained from a previously published study [37] and downloaded from the NCBI SRA website under the accession number DRA001104. RNA-seq data for the integument (GenBank accession numbers PRJNA215013 and PRJNA238971), previously reported by our group, were also included in this study [38].

Mapping of RNA-seq reads

The quality of raw reads was evaluated using FastQC [39]. Raw reads were filtered and trimmed using Trimmomatic 0.32 (parameters: ILLUMINACLIP: TruSeq3-PE.fa:2: 30:10; HEADCROP:10; TRAILING:3; SLIDINGWINDOW:4:20; MINLEN:75) [40]. Remaining reads were mapped against silkworm rRNA, tRNA, and mtDNA sequences collected in-house, using bowtie2 (version 2.2.3, parameters:–N 1;–L 20;–k 20), and matching reads were discarded [41]. The remaining high-quality clean reads were mapped to the silkworm genome (SilkDB 2.0

release) [42] using the spliced read aligner TopHat (version 2.09) [43]. In order to maximize the usage of splice junction information derived from all tissues, the previously described two rounds of TopHat mapping strategy was adopted [44]. In brief, reads from each samples were mapped with TopHat using the default parameters except 'min-anchor = 5' and 'min-isoform-fraction = 0'. All splice junctions detected by initial mapping were pooled and used as raw junctions for the second round of mapping, with the following parameters: 'raw-juncs', 'no-novel-juncs'. In order to facilitate transcript assembly and quantification, all mapped reads from the same tissue were merged into a single BAM file.

Transcriptome assembly

The transcriptome of each tissue was assembled from the TopHat mapped reads separately by Cufflinks [45], Scripture [46], and StringTie [47]. Cufflinks (version 2.02) was run with default parameters (and 'min-frags-per-transfrag = 0'), Scripture (VPaperR 3) was run with default parameters (and omission of the '-pairedEnd' option), StringTie (version 1.0.1) was run with the parameters (-f 0.01 -a 10 -j 1 -c 0.01), which from the slightly alter default parameters. The transcripts which was supported by at least two assembly programs or occurred in at least two tissues, was extracted as stringent transcripts. Stringent transcripts were merged into a unique transcript set using Cuffmerge. Then, the read coverage and fragments per kilobase of transcript per million mapped reads (FPKM) values for the 21 tissue types were estimated using Cufflinks.

LncRNA identification pipeline

We developed an analysis pipeline to identify bona fide lncRNAs from the newly generated silkworm transcriptome (Fig 1). (1) Transcripts that overlapped with any protein-coding exon in the sense orientation were removed; (2) transcripts with $<$ 200 bp, single-exon, read coverage $<$ 0.8, and FPKM $<$ 0.1 were eliminated; (3) transcripts with predicted large ORFs ($>$ 100 aa) were filtered out; (4) transcripts with predicted protein-coding potential were removed (protein-coding potential criteria: CPC score $>$ 0, CPAT score $>$ 0.345, and CNIC score $>$ 0) [48–50]; (5) transcripts with similarity to known protein sequences in the Swiss-Prot database (E-value $<$ 1e-6) [51] and known protein-coding domains in the Pfam (AB) database (E-value $<$ 1e-6) [52] were discarded; (6) transcripts within the $<$ 2k scaffold-end range were excluded; (7) finally, transcripts with class code 'i','u','x' subsets were retained as bona fide silkworm lncRNAs.

Analysis of sequence conservation of silkworm lncRNAs

The sequence conservation of silkworm lncRNAs was evaluated based on sequence similarity search, using the previously described method [2]. The genomes of *Caenorhabditis elegans*, *Acyrthosiphon pisum*, *A. mellifera*, *A. gambiae*, *Drosophila melanogaster*, *Tribolium castaneum*, *Heliconius melpomene*, *Danaus plexippus*, *Melitaea cinxia*, *Mnemiopsis leidyi*, *Solenopsis invicta*, and *Tetranychus urticae* were downloaded from ENSEMBL database [53]. The genomes of *Plutella xylostella* and *Manduca sexta* were obtained from the Diamondback moth genome database (http://iae.fafu.edu.cn/DBM/) and Agripest Base (http://agripestbase.org/manduca/), respectively. The genomes of *Papilio polytes* and *Papilio xuthus* were obtained from PapilioBase (http://papilio.nig.ac.jp/index.php). The genome of *Papilio glaucus* was downloaded from the official website for the tiger swallowtail genome (http://prodata.swmed.edu/LepDB/). The lncRNA sequences were searched against these 18 genomes using BLASTN (with E-value $<$ 1e-10); the best hit for each query and for each genome was retained, and a matrix of lncRNA/homolog pairwise similarity was obtained. The similarity matrix was visualized using the pheatmap R package [54].

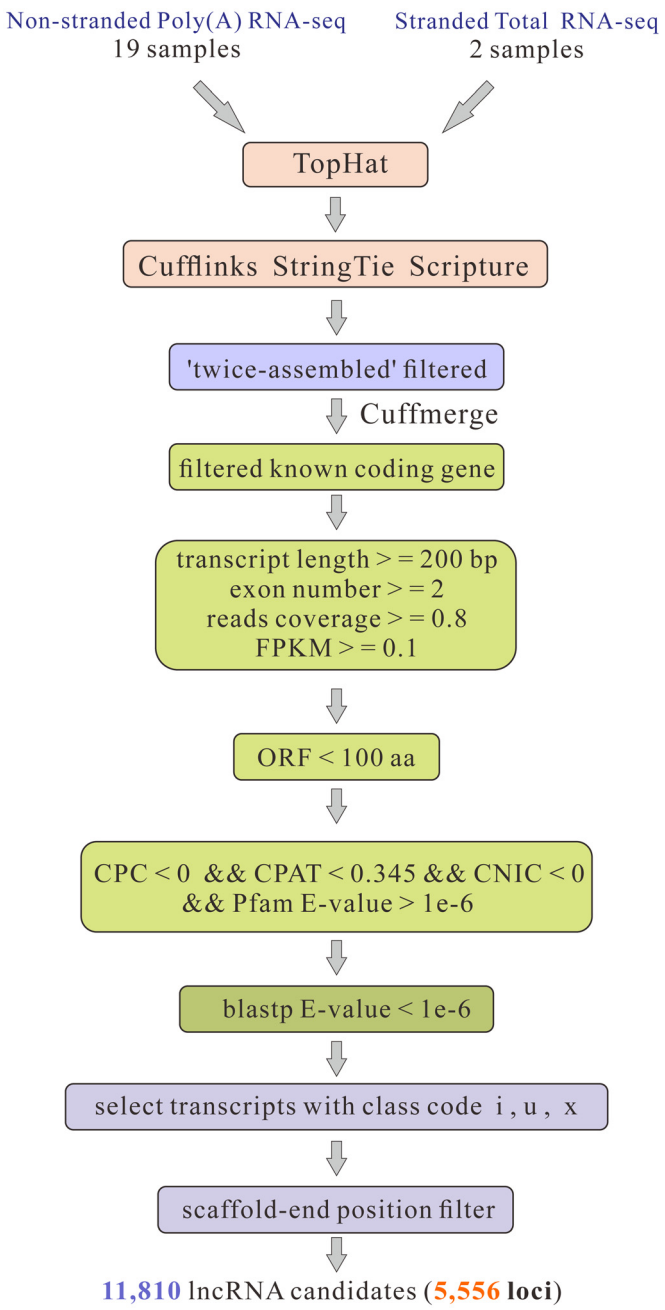

Fig 1. Integrative computational pipeline for the systematic identification of lncRNAs in silkworm. FPKM, Fragments per kilobase of transcript per million mapped reads; ORF, open reading frame; CPC, coding potential calculator; CPAT, RNA coding potential assessment tool; CNIC, coding non-coding index.

doi:10.1371/journal.pone.0147147.g001

Tissue specificity score

The tissue-specific score (JS score) has been previously defined by Cabili et al. [44]. In the current study, we calculated the tissue-specific score for each transcript using the csSpecificity() function provided by the CummeRbund R package [55].

LncRNAs as precursors of miRNAs

In order to identify lncRNAs as precursors of miRNAs, we intersected the GFF (Generic Feature Format) file containing silkworm lncRNA genomic positions with the GFF file containing mature miRNA sequences downloaded from miRBase (Release 21) [56]. LncRNA loci that overlapped with miRNA loci on the same strand were considered as the precursors of these miRNAs.

Prediction of competing endogenous RNAs (ceRNAs)

CeRNAs may be identified by traditional miRNA target prediction methods [57–59]. In the present study, we inferred the conserved regions of silkworm lncRNAs that may harbor miRNA response elements (MREs) for ceRNA network. MREs in the conserved regions of lncRNAs were predicted using miRanda [60], PITA [61], and RNAhybrid [60].

Weighted gene co-expression network analysis

All transcripts, including protein-coding mRNAs and lncRNAs, expressed in at least 2 samples were used for constructing the weighted gene co-expression network (WGCNA). WGCNA construction and module detection were performed using the "WGCNA" R package v. 1.4.6 [62]. The overall procedure involved the generation of a Pearson correlation matrix between all transcript pairs, followed by the transformation of this correlation matrix into an adjacency matrix with a soft-thresholding power of 9, using an adjacency function that implement in WGCNA package. Then, the adjacency matrix was transformed into a topological matrix (TOM). Primary modules were identified via linkage hierarchical clustering with the topological overlap dissimilarity matrix (1-TOM), those with high correlation (module eigengene correlation > 0.70) were merged, and the module membership (kME) of each gene was calculated. Cytoscape v.3.2.1 software was used for network visualization [63].

Gene ontology and pathway analysis

Gene Ontology (GO) enrichment analysis of each module was performed using GOseq [64]. KEGG metabolic pathway enrichment analysis was carried out using KOBAS 2.0 [65], with the *Bombyx mori* database as background. All data were statistically analyzed using the hypergeometric test and Benjamini-Hochberg FDR (false discovery rate) correction, and only GO terms or KEGG pathways with corrected *p*-values of less than 0.05 were considered enriched.

Identification of sex-biased transcripts

In order to identify sex-biased transcripts among the ten sex-sampled tissues, the raw read counts of each transcript were re-estimated using RSEM [66]. Fold change (FC) and FDR between females and males were calculated for each tissue type using DEGseq [67]. For a given tissue, transcripts with a $|\log2FC| > 1$ and FDR < 0.05 were considered sex-biased.

Results and Discussion

Genome-wide identification of lncRNAs in silkworm

In order to systematically identify lncRNAs in the silkworm genome, we sequenced 18 libraries. Totally, 2.15 billion raw reads were generated and 1.71 billion clean reads were retained after stringent filtering (S1 Table). In addition, two public datasets from silkworm embryo and integument were also included in this study (S1 Table). In order to obtain a comprehensive silkworm transcriptome, reads from each tissue were assembled using the three most widely

used assemblers (Cufflinks, Scripture, and StringTie). A total of 6,524,370 transcripts were generated, of which 3,511,465 transcripts were assembled at least twice (i.e. these transcript were assembled by at least two assemblers, or assembled in at least two tissues). We defined these 'twice-assembled' transcripts as stringent transcripts. The stringent transcripts were merged into a unique transcript set, composed of 29,416 gene loci and 553,658 transcripts, using Cuffmerge.

An lncRNA identification pipeline was developed as shown in Fig 1. Briefly, we filtered out transcripts that overlapped with coding gene exons in the sense orientation, retaining 55,739 transcripts for 17,553 gene loci. In order to obtain long, oriented, and expressed transcripts, we filtered out the transcripts shorter than 200 bp, those that possessed only a single exon, as well as transcripts with single base read coverage < 0.8 and FPKM < 0.1. In addition, transcripts with ORFs > 100 aa were discarded. Then, the protein coding potential of each transcript was accessed by CPC, CPAT, and CNIC, respectively. Transcripts with CPC score > 0, CPAT score > 0.345, or CNIC score > 0 were excluded. The remaining transcripts were subjected to protein domain filtering using HMMER (version 3.0) against known protein domains documented in the Pfam (version 27.0) database, in order to evaluate whether they contained a known protein-coding domain. In order to rule out incompletely assembled transcripts due to the effects of scaffold-end boundaries, transcripts within $<$ 2k scaffold-end range were excluded. Finally, only transcripts with class codes 'i', 'u', 'x', representing intronic, lncRNAs (ilncRNAs), lincRNAs, and natural antisense lncRNAs (lncNATs), respectively, were retained. This resulted in a final set of 11,810 silkworm lncRNA transcripts from 5556 loci, of which 474 were ilncRNAs, 6,250 were lincRNAs, and 5,086 were lncNATs. The genomic coordinates of the identified lncRNA transcripts (GTF format) are provided in S2 Table.

In the present study, we report the generation of a relatively robust list of silkworm lncRNAs. As most of the RNA-seq libraries in this study were prepared from day 3 fifth instar larvae by the non-strand-specific poly(A) selection method, it was expected that use of a broad variety of tissues would result in identification of a larger number of lncRNAs. As most of the RNA-seq libraries were non strand-specific and poly(A)-selected, several limitations should be addressed: first, a large proportion of non-poly(A) silkworm lncRNAs could not be detected; second, single-exon transcripts were excluded due to lack of strand information; third, the number of ilncRNAs and lncNATs were underestimated. The "twice-assembled" filter strategy, that has been adopted in several studies, was utilized to prevent mis-assembly of transcripts [10, 68]; however, some bonafide transcripts may have been lost. The combination of protein-coding potential filtering and protein-domain filtering steps, has been shown efficiently reduce false negative and false positive rate for distinguishing non-coding transcripts from protein-coding transcripts [10, 44]. In addition, some transcripts assembled in the scaffold-end region may have been incompletely assembled. Scaffold-end boundary effects should be avoided for unfinished genome with a large number of scaffolds, e.g. the silkworm genome. In summary, our approach, which was comparable to previously reported methods [8–12, 68, 69], resulted in the reliable identification of lncRNAs; however, a proportion of bona fide lncRNAs may have been filtered out.

The genomic features of silkworm lncRNAs

In order to characterize their genomic features, potential silkworm lncRNAs were compared with known protein-coding mRNAs. Overall, silkworm lncRNAs (median of length 1,459 bp for lncRNAs; 955 bp for lincRNAs) were found to be significantly shorter than protein-coding mRNAs (2,741 bp for mRNAs, Kolmogorov-Smirnov test (KS-test) p-value $<$ 2.2e-16), whereas, lncNATs (median of length 2,602 bp) were similar in length to protein-coding

transcripts (KS-test p-value = 0.03179) (Fig 2A). In contrast to the overall length of lncRNAs, their exon lengths (median length of 285 bp for lncRNAs, 239 bp for lincRNAs, and 405 bp for lncNATs) are significantly longer than those of protein-coding mRNAs (median length of 168 bp; KS test p-value < 2.2e-16). A similar pattern was observed for introns (Fig 2B); silkworm lncRNAs have fewer exons than mRNAs (2.73 vs. 5.17 on average; 2.66 for lincRNAs, 2 for lncNATs; KS test p-value < 2.2e-16) (Fig 2C). This finding may explain the longer exon length and shorted overall length of lncRNAs relative to mRNAs. LncRNA loci possess fewer transcript isoforms than protein-coding mRNA loci (2.03 vs. 3.18 on average per gene locus, 1.97 for lincRNA, 2.11 for lncRNAs, KS test p-value < 2.2e-16) (Fig 2D), suggesting that lncRNAs are less complex than protein-coding mRNAs. The median sizes of the max-ORF of lncRNAs (138 bp for lincRNAs and 189 bp for lncNATs) are significantly shorter than those of mRNAs (732 bp for mRNAs, KS test p-value < 2.2e-16) (Fig 2E). Analysis using the Wilcoxon rank sum test show that silkworm lncRNAs have lower protein-coding potential than well-annotated KAIKObase gene models and NM annotations downloaded from NCBI (Fig 2F). Similar to mammalian lncRNAs, silkworm lncRNAs contain more repeat sequences than mRNAs (18.7% vs. 4.54%) (Fig 2G). The predominant repeat sequences within lncRNAs are LINEs (7.9%) and SINEs (6.5%). For the four main classes of repeats (LINE, SINE, DNA, and LTR), except LTR, both lincRNAs and lncNATs show a greater preference overlapped with repeat elements than mRNA (Fig 2G). Interestingly, the GC content in silkworm lncRNAs is lower than in coding sequences (CDS) but slightly higher than in untranslated regions (UTRs) (Fig 2H). The expression profiles of silkworm lincRNAs did not show stronger correlation with the adjacent protein-coding genes. However, like the nearby coding gene pairs, lincRNAs tend to correlate with their nearest protein-coding neighbors compared with randomly selected counterparts (Fig 2I). In about 80% of lncNATs, only a short fraction of the length (less than 35% of the sequence) is overlapped by protein-coding exons (S1D Fig), whereas lncNATs show relatively high correlation with antisense protein-coding genes (Fig 2I).

Silkworm lncRNAs exhibit numerous features that are distinct from those of coding mRNAs; however, the vast majority of lncRNAs are spliced by canonical splices sites (GT/AG), and no differences in splicing signal usage are found compared with protein-coding mRNAs (S2B Fig). In addition, the distribution of lncRNAs in silkworm chromosomes was examined. Silkworm lncRNAs were unevenly distributed across the 28 silkworm chromosomes (Chi-Square Goodness of Fit Test, p-value < 2.2e-16) (S1A Fig, S3 Table). Intriguingly, chromosome Z contained the largest number of lncRNAs with the highest gene density (722 transcripts, 304 gene loci, 14.9 genes/Mb), whereas the smallest chromosome (chromosome 2) presented the least number and lowest density of lncRNAs (108 transcripts, 64 gene loci, 8.1 genes/Mb).

The levels of conservation and polymorphism of lncRNAs were also investigated, as shown in Fig S2C. We found that 58.3% lncRNAs (6,885) were specific to silkworm, in contrast with 17.9% of coding mRNAs (2,990). Homologous fragment sequences of 41.44% lncRNAs (4,894/11,810) can be found in other Lepidoptera species. In contrast, only 12.01% of lncRNAs (1,419/11,810) were found to be conserved when comparing with the slightly more distantly related species *A. pisum*, and only 1.74% of lncRNAs (206/11,810) were found to be conserved when comparing with even more distantly related species. For coding mRNAs, 63.14% (10,521/16,664) were conserved among the representative eighteen species. LincRNAs possessed a larger number of polymorphism sites than lncNATs and ilncRNAs (Fig S2E). These results suggest that silkworm lncRNAs are highly species-specific and conserved to a small extent among Lepidopterans. In addition, silkworm lncRNAs are considered to have undergone more rapid evolution than protein-coding mRNAs do.

Collectively, silkworm lncRNAs share similar patterns with those of other species such as flies, humans, and zebrafish. In particular, silkworm lncRNAs possess short exons, long

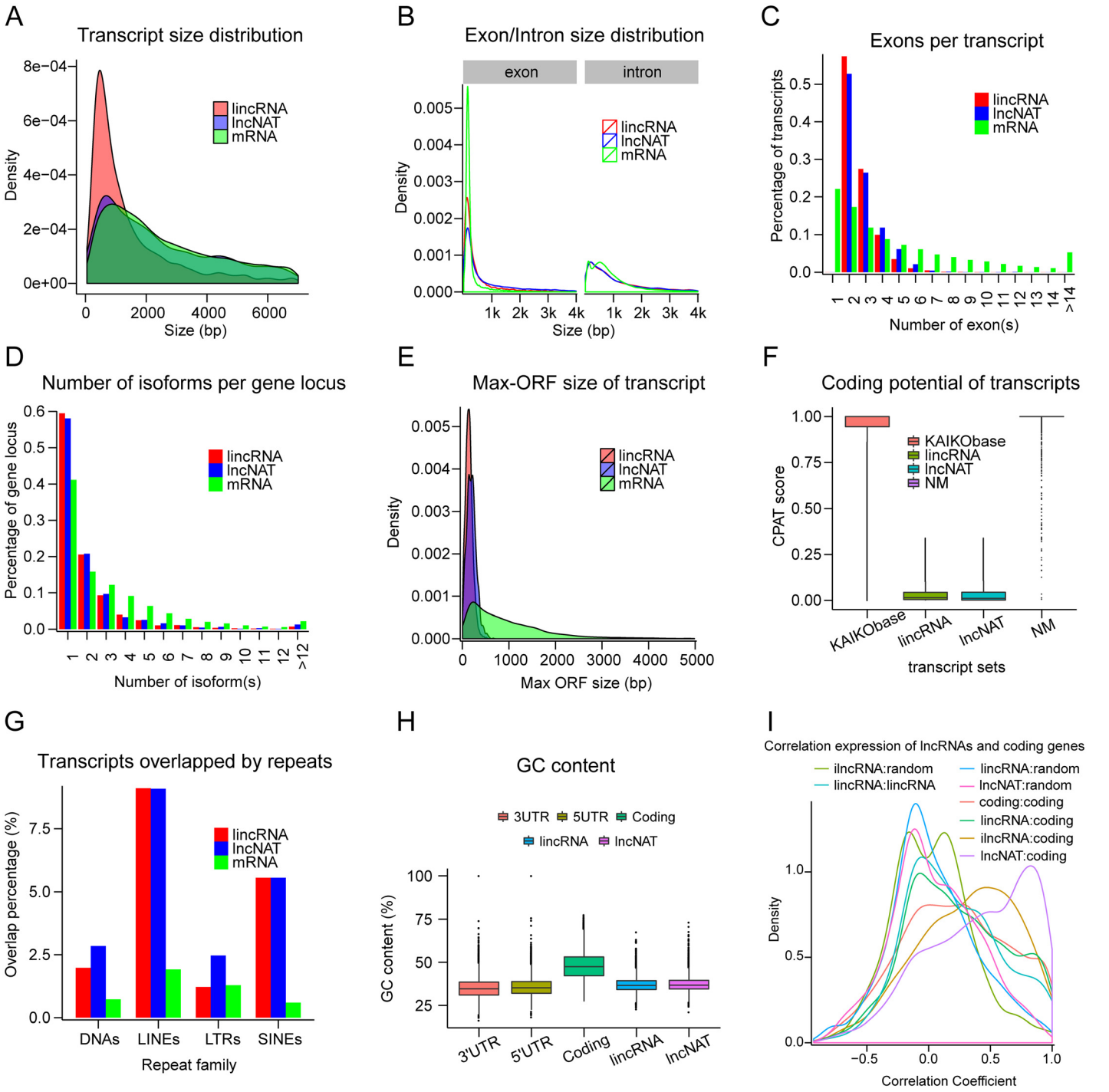

Fig 2. Features of silkworm lncRNAs. (A) Transcript size distribution for lincRNAs, lncNATs, and mRNAs. (B) Exon (left) and intron (right) size distributions for lincRNA, lncNATs, and mRNAs (C) Number of exons per transcript for all lincRNAs, lncNATs, and mRNAs. (D) Distribution of the number of isoforms for each lincRNA, lncNAT, and mRNA gene locus. (E) Maximum ORF size distribution for lincRNAs, lncNATs, and mRNAs. (F) CPAT score distribution for KAIKObase gene model, 'NM_' reference sequence of silkworm, lincRNAs, and lncNATs. (G) Proportion of lincRNAs, lncNATs, and mRNAs transcripts covered by main repeat classes annotated by RepeatMasker. (H) GC content of lincRNAs, lncNATs, and mRNAs. (I) Pearson correlation coefficient distribution for neighboring transcript pairs from different datasets. Coding, protein-coding mRNAs; Random, random shuffle of mRNA positions.

doi:10.1371/journal.pone.0147147.g002

introns, low levels of conservation, low GC content, and a large degree of overlap of repeat sequences. Additionally, silkworm lncRNAs are slightly related to their closest protein-coding neighbors. In addition, we found that the transcript lengths of lncNATs were similar to those of mRNAs compared with those of lincRNAs, although both types of non-coding RNAs (lncNATs and lincRNAs) shared several common features.

However, due to most of our RNA-seq libraries were non-strand specific, single-exon lncRNAs were excluded from this study, even though the silkworm, like other species, is considered to possess a large proportion of single-exon lncRNAs. It must be noted that the transcript lengths and exon numbers per transcript of lncRNAs may have been overestimated, and that the number of lncNATs and exon sizes of lncRNAs may have been underestimated in the present study. Previous studies have shown that lncRNAs exhibit more positional conservation than sequence conservation across species, and revealed that the positional conserved lncRNAs have important biological functions [70, 71]. The present study focused on the analysis of sequence conservation, rather than positional conservation, in silkworm lncRNAs, which is not suitable for inferring the functions of the conserved lncRNAs. Therefore, in order to elaborate the functional role of conserved lncRNAs, the further studies based on phylogenomic approach are needed.

Silkworm lncRNAs are more tissue-specific than mRNAs

Based on the FPKM values of genes estimated using Cufflinks, tissue-specific lncRNAs were investigated in the 21 silkworm tissues by determining the tissue specificity score, also termed the JS (Jensen-Shannon) score [44]. JS scores range from zero for genes expressed ubiquitously in all tissues, to one for genes expressed in a single tissue (Fig 3 and S2 Fig). Using JS score = 0.25 as a cutoff, the majority of lincRNAs (73.5%) were found to be tissue-specific, compared with 34.6% mRNAs. In contrast, 39.1% of lncNATs were tissue-specific (Fig 3A). Moreover, more than a third of all lincRNAs were specific to the testis and approximately one fifth were specifically expressed in the brain. These findings are consistent with previous reports [5, 44]. KS test revealed that lincRNAs show much higher tissue specificity than lncNATs (p-value $<$ 2.2e-16). Furthermore, both lincRNAs and lncNATs show greater tissue specificity than mRNAs (p-value = 2.23e-13) (Fig 3B).

The expression levels of silkworm lncRNAs (both lincRNAs and lncNATs) were significantly lower than those of protein-coding genes in the 21 tissues (KS test p-value $<$ 2.2e-16 for lincRNAs vs. mRNAs, p-value $<$ 2.2e-16 for lncNATs vs. mRNAs), although lncRNAs show slightly higher levels of expression in the brain and testis than in other tissues (S2A Fig). The median maximal expression levels of lincRNAs and lncNATs are ~29-fold and ~8-fold lower than those of mRNAs (median maximal FPKM is 1.0 for lincRNAs, 3.1 for lncNATs, 1.6 for lncRNA, and 24.9 for mRNAs) (S2B Fig). Although lncRNAs are expressed at relatively lower levels than protein-coding genes, their high tissue specificity suggests that they may perform *ad hoc* biological functions in specific tissues, rather than simply contributing to transcriptional noise.

Classification of silkworm lncRNAs as miRNA precursors and potential competing endogenous RNAs

Certain lncRNAs may function as precursor molecules that are processed into smaller regulatory RNAs such as miRNAs [9, 10, 72]. In order to determine whether silkworm lncRNAs are actually precursors of miRNAs, we compared their genomic coordinates with corresponding genomic locations on the same strand of miRNAs downloaded from miRBase (Release 21). In all, 69 lncRNAs from 33 gene loci, were identified as known precursors and found to be

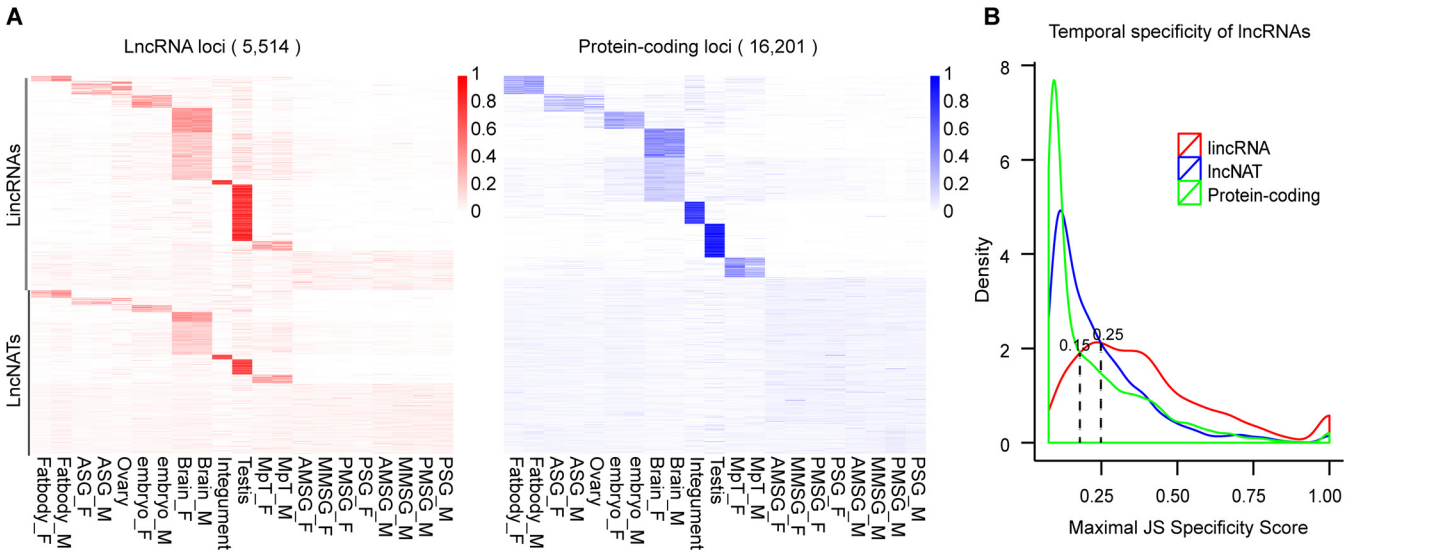

Fig 3. Tissue specificity of lncRNAs and protein-coding genes. (A) Heatmaps of 5,514 lncRNA loci (red; left) and 16,201 protein-coding loci (blue; right) based on normalized expression values (the sum of expression values across all tissues per locus is set to one, using the method described by Cabili, M. N., et al [44]). (B) The distribution of maximal tissue specificity scores for each transcript across 21 tissues.

doi:10.1371/journal.pone.0147147.g003

distributed among 25 silkworm miRNA families (S4 Table). For example, tow gene loci XLOC_010603 and XLOC_010759 residing in the *Hox* gene cluster were determined to be precursors of miR-iab-4 and miR-10, respectively (Fig 4). miR-iab-4 is analogous to miR-196 in vertebrate Hox clusters. In vivo experiments in *Drosophila* showed miR-iab-4-5p directly inhibits *Ubx* activity and regulates ectopic expression of mir-iab-4-5p, resulting in the transformation of halters into wings [73]. mirR-10 was predicted to target the *Scr* gene. The 3'UTR of *Scr* genes has been conserved over hundreds of millions of years of evolution, suggesting that this region is likely the functional target site for miR-10 [74]. The transcriptional levels of bmo-miR-10b-3p/5p were significantly increased during metamorphosis [75]. Moreover, the transcript TCONS_00253471 of the XLOC_010759 loci was highly expressed in embryos and body walls. Taken together, these data suggest that the XLOC_010603 and XLOC_010759 transcripts overlapping with miR-10 and miR-iab-4 may function as miRNA precursors, playing an important role in the regulation of *Hox* gene expression.

LncRNAs have undergone rapid sequence evolution; however, some lncRNAs still possess short functional elements that have remained conserved in different species [71]. LncRNAs may bind miRNAs as competing endogenous RNAs (ceRNAs), thereby functioning as miRNA sponges [76]. The lncRNA-miRNA interaction can be examined using traditional miRNA target prediction methods [57–59]. In the current study, we inferred the conservation elements region of silkworm lncRNAs that may harbor miRNA response elements (MREs) for the ceRNA network. In total, 104 lncRNAs from 72 gene loci were predicted as 'decoys' for 101 known miRNA families (S5 Table). For example, the transcripts TCONS_00111202 and TCONS_00111199 from the XLOC_004695 gene locus harbor the bmo-miR-184-3p, bmo-miR-3378-5p, and bmo-miR-745-5p response elements (S3A Fig). MiR-184, a single-copy gene that is evolutionarily conserved from insects to primates, is expressed ubiquitously in *Drosophila* embryos, larvae, and adults, and shows dynamic expression pattern in the central nervous system during embryonic development [77]. Loss of function of miRNA-184 results in reduced motility in adult *Drosophila* and complete loss of egg production in the female [78, 79]. More recently, a study found that miR-184, which plays an important role in energy

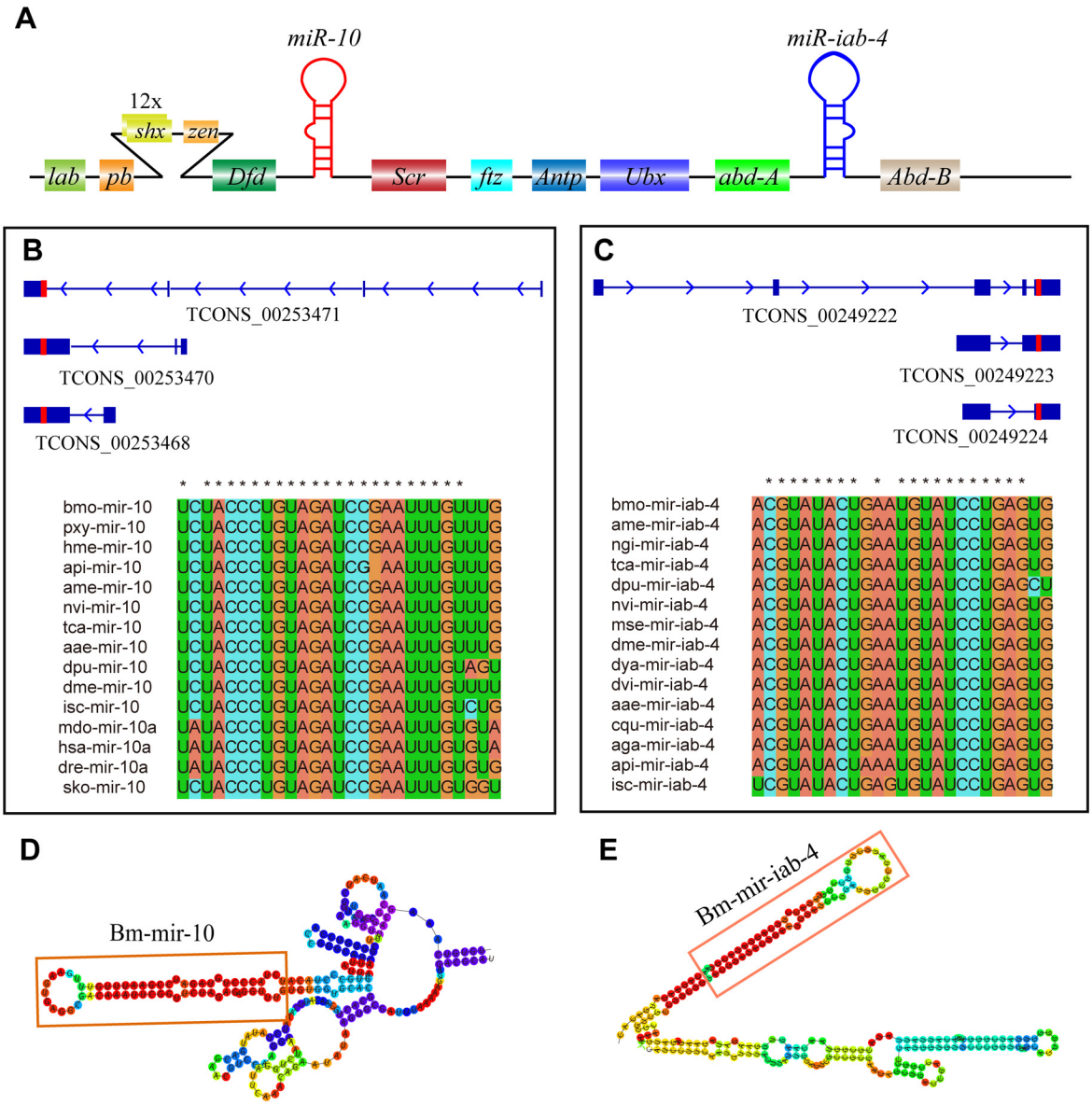

Fig 4. Potential miRNA precursors in the silkworm *Hox* cluster. (A) Schematic representation of the silkworm Hox cluster with potential miRNA precursors for miR-iab-4 (blue) and miR-10 (red). (B) Schematic of bmo-mir-10 precursors and alignment of bmo-miR-10 orthologues from selected bilaterians. Boxed regions indicate miR-10 primary sequences. Asterisk indicates sequences that are highly conserved with those of silkworm. (C) Schematic of bmo-mir-iab-4 precursors and alignment of bmo-mir-iab-4 orthologues from selected bilaterians. Boxed regions indicate bmo-mir-iab-4 primary sequences. Asterisk indicates sequences that are highly conserved with those of silkworm. (D) Secondary structures of bmo-mir-10 represent precursor (TCONS_). (E) Secondary structures of bmo-mir-iab-4 represent precursor (TCONS_). All secondary structures were predicted using RNAfold.

doi:10.1371/journal.pone.0147147.g004

homeostasis, was negatively regulated upon administration of a sucrose-rich diet [80]. TCONS_00111202, expressed ubiquitously in the 21 silkworm tissues studied, exhibits slightly higher expression levels in the ASG, fat body, and embryo compared with other tissues (S3B Fig). The miRNA response elements in TCONS_00111202 are conserved from *B. mori* to *T. castaneum*, implying that TCONS_00111202 may function as a ceRNA, with important roles in the silkworm. These results reveal that lncNAs may function as miRNA precursors or ceRNAs, and play important roles in numerous regulatory pathways.

Sex-biased expression of silkworm lncRNAs

The examination of sex-biased expression of silkworm coding genes, using microarray technology, reveals that male-biased genes are enriched on the Z chromosomes [81]. It is of great interest to investigate the sex-biased expression of silkworm lncRNAs. Firstly, in order to confirm the samples as sex-specific, the W chromosome-specific gene (*fem*) was adopted as a marker gene, as the male library may be derived from incorrectly sexed embryos or RNA produced by polar bodies [37], whereas the read counts of the *fem* gene from the male library are far fewer than those of the female library (Fig 5A). Therefore, it was ensured that the embryo samples were suitable for sex-biased analysis, whereas the integument sample yielded only one mix-sex sampled library [38]. Finally, all 20 samples were correctly sex-sampled and retained for further analysis. According to the criteria ($|\log2FC| > 1$ and $FDR < 0.05$), significantly more genes were found to be upregulated in male (male-biased expressed) gonads, Malpighian tubes, brains, and PSG, whereas the reverse was observed for PMSG, AMSG and ASG (S6 Table). A similar pattern was observed for the protein-coding genes (S6 Table). Notably, lncRNAs with female-biased expression were vastly outnumbered by those with male-biased expression in the gonad (female vs. male: 774 vs. 3,176). Among these male-biased lncRNAs, 1,772 transcripts were specifically expressed in the testis and found to be enriched on Chromosome Z, 13, and 22, suggesting that the male-biased lncRNAs contribute to spermatogenesis and other male-specific biological processes.

Among the 5,556 lncRNA gene loci, 49.47% (2,749) showed sex-biased expression. In detail, 1,029 single-isoform gene loci (32.33%) showed sex-biased expression, whereas 1,720 multi-isoform gene loci, with at least one isoform, exhibited sex-biased expression. In this study, we defined a sex-biased ratio for multi-isoform gene loci in order to represent the rate of sex-biased isoforms. Gene loci with a sex-biased ratio of over 0.75 were considered sex-biased. Applying this sex-biased ratio criterion, the numbers of sex-biased gene loci were found to be 1,029 and 523 for single- and multi-isoform gene loci, respectively. Taking gene locus XLOC_012091 as an example, 14 isoforms were annotated in this study, of which 11 isoforms showed sex-biased expression and were antisense with respect to *yellow-d* (KAIKOBASE ID: BMgn007254) (S4A Fig). The *yellow*-like gene, which has only been identified in insect and bacterial species, has been reported to be involved in the melanin biosynthetic pathway and associated with movement and mating behavior in *Drosophila* [82–84]. Ten isoforms of XLOC_012091, with the exception of TCONS_00285772, TCONS_00285767, TCONS_00285768, and TCONS_00285757 exhibited testis-specific expression (S4B Fig), suggesting that the XLOC_012091 gene loci, which shows a strong male-biased expression signal, may be involved in silkworm mating behavior.

LncRNAs that were nearing, or intersecting with, primary sex determination pathway genes (*Fem*, *Masc*, *Imp*, *Psi*, *Dsx*) were searched, and two lncRNA isoforms located in the *Psi* intron region, six lncRNA isoforms in the *Dsx* intron region, and one isoform antisense to Dsx were found. Notably, TCONS_00200625 isoform, an antisense transcript, was found to overlap with the 4th exon of Dsx, suggesting that TCONS_00200625 may interfere with sex-specific splicing in Dsx (Fig 5).

In summary, numerous silkworm lncRNAs showed sex-differential expression, with some gene loci displaying very high sex-biased ratios and sex-limited expression. A few lncRNAs were identified to play an important role in sex determination pathways. Although more evidences are needed to prove these findings, our results demonstrate sex-biased expression in silkworm lncRNAs and provide supplement account for sexual dimorphism in the silkworm.

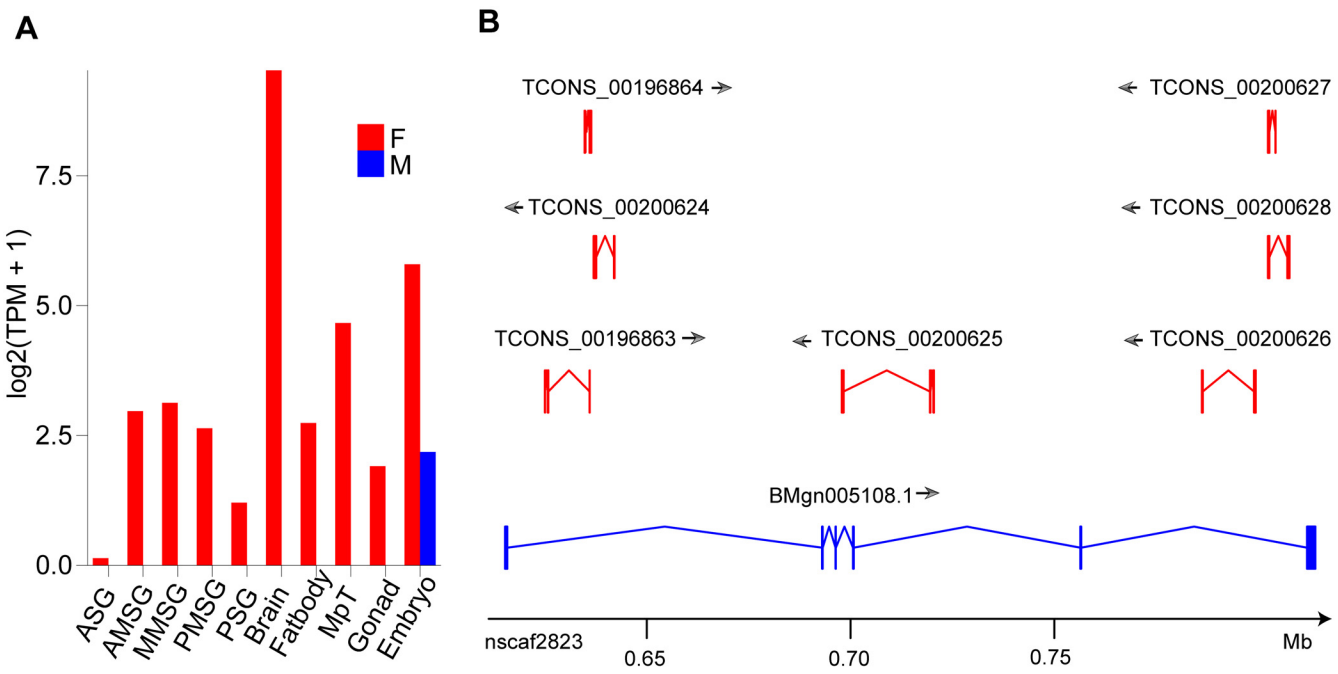

Fig 5. Quality of sex-sampled and sex-biased lncRNAs at the Bmdsx gene locus. (A) Expression pattern of the W chromosome-specific gene, fem. (B) Sex-biased lncRNAs at the Bmdsx gene locus.

doi:10.1371/journal.pone.0147147.g005

Functional annotation of silkworm lncRNAs

Due to the lack of annotated features, it is still challenging to predict putative function of lncRNAs merely from their sequence features. Fortunately, co-expression network-based "guilt-by-association" analysis methods have been successfully applied to the prediction of lncRNA function [85]. Therefore, we used WGCNA [62], an R package for weighted correlation network analysis, to associate lncRNAs with functional annotated mRNAs and predict their functions using a module-based method. The analysis resulted in 19 distinct modules with module sizes ranging from 17 to 4,248 (mean, 732; median, 123). Intriguingly, 12 of these 19 modules were strongly associated with tissue type (correlation > 0.65, p-value $<= 0.03$, Fig 6A, S5 and S6 Figs). Functional annotation and enrichment revealed that tissue-associated modules are biologically meaningful and related to tissue-specific biological processes (Fig 6, S7 Table).

The largest module (the blue module), which was specifically associated with the embryo (tissue correlation = 0.94, p-value = 2e-10), contained 1,586 lncRNAs and 2,662 mRNAs. In this module, genes related to "DNA binding" (GO:0003677), "DNA replication" (GO:0006260), "regulation of transcription, DNA-templated" (GO:0006355) were overrepresented and "Ubiquitin-mediated proteolysis" (bmor04120), "Spliceosome" (bmor03040), "Wnt signaling pathway" (bmor04310) and "FoxO signaling pathway" (bmor04068) were enriched, suggesting that lncRNAs play important roles in early embryonic development in the silkworm (S7 Table).

In the aquamarine module (tissue correlation = 0.99, p-value = 8e-20), which was highly correlated with the brain, genes related to "hormone activity" (GO:0005179), "signal transduction" (GO:0007165), "neuropeptide signaling pathway" (GO:0007218), "synaptic transmission" (GO:0007268) and "Neuroactive ligand-receptor interaction" (bmor04080) were overrepresented, indicating that the functional enrichment results were consistent with brain attributes,

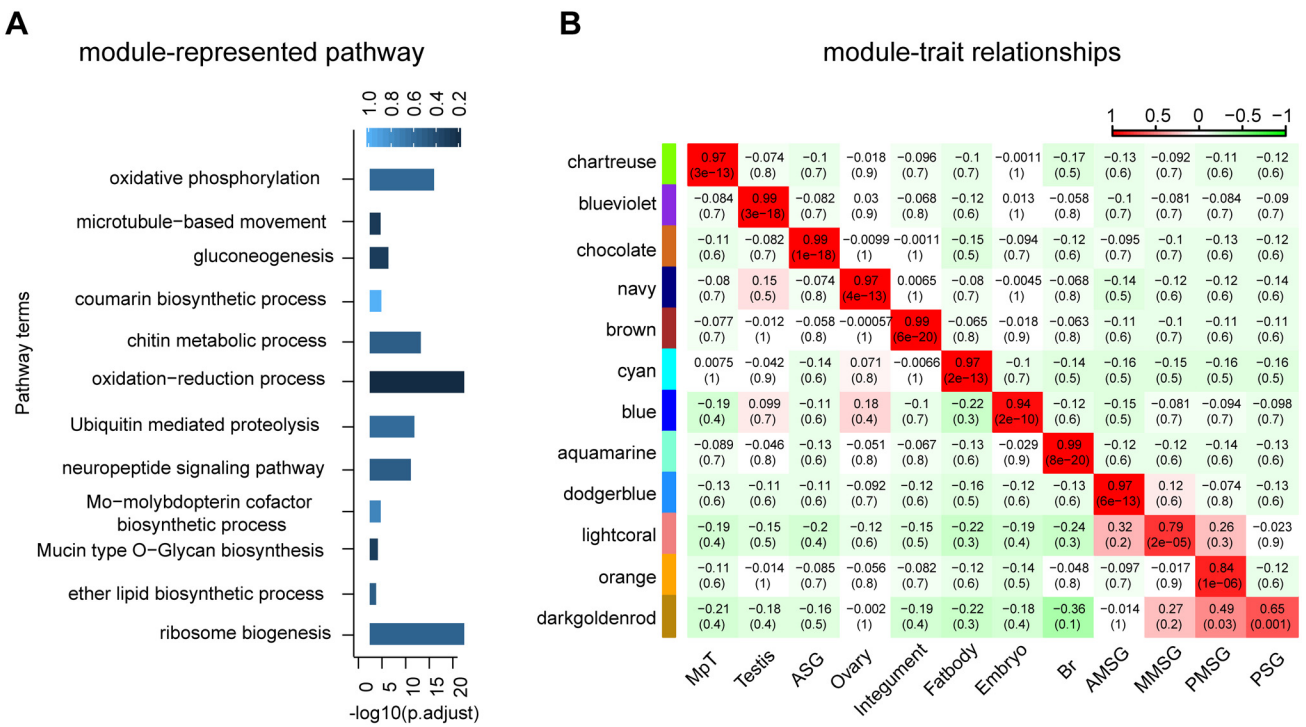

Fig 6. Functional enrichment of protein-coding genes in network modules and module-tissue correlation, and the corresponding *p*-values. (A) Functional enrichment of protein-coding genes in network modules. For each module, representative enrichment GO terms are shown, with bar plot of -log10 (p.adjust). Light to dark blue represent increasing enrichment factors (from 0 to 1). (B) Module-tissue correlations and corresponding *p*-values. Boxes contain Pearson correlation coefficients and their associated *p*-values. Positive correlation (red) indicates that the module is positively correlated with the specific tissue, whereas negative correlation (green) indicates the reverse. MpT, Malpighian tubule; ASG, anterior silk gland; AMSG, anterior-middle silk gland; MMSG, middle-middle silk gland; PMSG, posterior middle silk gland; PSG, posterior silk gland; Br, brain.

doi:10.1371/journal.pone.0147147.g006

and that the lncRNAs expressed in this module are involved in functional processes in the brain (S7 Table). The blueviolet module, which was enriched in "microtubule" (GO:0005874), "microtubule motor activity" (GO:0003777), "cilium morphogenesis" (GO:0060271), "microtubule-based movement" (GO:0007018), was highly correlated with testis (tissue correlation = 0.99, *p*-value = 3e-18), suggesting that lncRNAs in this module may play important roles in the spermatogenesis and the development of the testis (S7 Table). In addition, chartreuse, brown, cyan, chocolate, navy, and lightcoral, were specifically associated with the Malpighian tubules, integument, fatbody, ASG, ovary, and MSG, respectively (Fig 6, S5A Fig).

Notably, eight modules were related to the silk gland. Among these modules, the chocolate module was specific to ASG, and the dodgerblue module, lightcoral module, and orange module were highly correlated to AMSG, MMSG, and PMSG, respectively (S5B and S6 Figs). Given that the darkgoldenrod module is associated with both PMSG (tissue correlation = 0.49, *p*-value = 0.03) and PSG (tissue correlation = 0.65, *p*-value = 0.001), it is expected that the darkgoldenrod module is highly correlated with MPSG (tissue correlation = 0.84, *p*-value = 2e-06). The chocolate module, which is associated with ASG, contained 829 transcripts (470 mRNAs and 359 lncRNAs) and was found to be enriched in the "gluconeogenesis" (GO:0006094), "glycolytic process" (GO:0006096) biological processes, and the "Glycolysis/Gluconeogenesis"(bmor00010), and "Carbon metabolism" (bmor01200) pathways. Among these modules, darkgoldenrod module was mainly enriched in translation and protein export. Specifically, KEGG enrichment results showed the darkgoldenrod module enriched in "Ribosome"(bmor03010) and "Aminoacyl-tRNA biosynthesis"(bmor00970) pathways which were highly involved in translation process, and

enriched in "Protein export"(bmor03060) and "Protein processing in endoplasmic reticulum"(bmor04141) pathways which represented protein export processing. In the dodgerblue module, which consists of 224 transcripts (118 mRNAs and 106 lncRNAs) and is specific to AMSG, genes involved in the "Folate biosynthesis" (bmor00790) pathway, "oxidation-reduction process" (GO:0055114), and "Mo-molybdopterin cofactor biosynthetic process"(GO:0006777) are over-represented. The lightcoral module was enriched in the "Mucin type O-Glycan biosynthesis" (bmor00512), "Biosynthesis of unsaturated fatty acids" (bmor01040), and "Fatty acid metabolism" (bmor01212) pathways, suggesting that MSG not only participates in sericin biosynthesis but is also involved in the production of other components of the cocoon, e.g. fatty acids (S7 Table).

Silkworm lncRNAs function as regulators of silk protein biosynthesis and secretion

As described above, eight modules was found to be associated with the silk gland. Since the darkgoldenrod module was the only module mainly enriched in translation and protein export, we highlighted this module to deeply investigate the functional role of lncRNAs in the silk gland. Based on the knowledge of coding gene annotation, Gene Ontology enrichment, KEGG pathway enrichment and cell biology, we manually split the darkgoldenrod gene network into translation-, translocation-, secretory-, cellular-, protein protection-, and unknown sub-function-related networks. All the sub-networks, except for the unknown sub-network, were selected for further analysis.

The translation sub-network was the largest, consisted of 128 coding genes, which were mainly involved in ribosome biogenesis, translation, formation of the translation pre-initiation complex, translation initiation, translational elongation, and aminoacyl-tRNA ligase. The secretory network, which was the second largest sub-network, consisted of 64 coding genes that were mainly involved in protein export, endoplasmic reticulum organization, endoplasmic reticulum unfolded protein response, and transmembrane transport. The translocation sub-network, consisting of the signal recognition particle (SRP), signal peptidase complex, translocon, and translocon-associated proteins, was ranked as the 3rd sub-network. The fourth sub-network, namely the cellular and protein protection network, was composed of three negative regulators of macroautophagy proteins and seven serine protease inhibitors (Fig 7, S8 Table).

For each sub-network, the first neighboring lncRNAs of the well-annotated proteins were selected, and the top 5 lncRNAs in degree values from each four sub-networks were defined as hub genes, leading to the identification of 13 hub lncRNAs (7 intra-network hub lncRNAs and 6 inter-network hub lncRNAs) (Fig 7, S8 Table). Six of the hub lncRNAs were involved in at least two sub-networks. TCONS_00454328, with the highest degree, was in the module network core and participated in translation, translocation and secretion-related processes. TCONS_00149264 and TCONS_00518124 function as hub genes of the translation and translocation sub-network. TCONS_00427691 functions as a hub gene of the translation and secretory sub-network. In addition, seven lncRNAs were specific to each sub-network. For example, TCONS_00319007, with degrees of 122, was specific to the translation sub-network.

In addition to sub-network analysis, 17 lncRNAs of the darkgoldenrod module were found in the GROSS (genomic regions of selective signals) regions, and 10 lncRNAs were differentially expressed between domestic and wild silkworm strains. However, all of the above lncRNAs did not overlap with the 13 selected hub genes, suggesting that the 13 selected hub genes play baseline functional roles in silk protein synthesis and secretion. Additionally, several genes involved in the juvenile hormone (JH) pathway were identified. More recently, our group demonstrated that JH is involved in silk protein synthesis [86], suggesting that lncRNAs may interact with the JH pathway and participate in regulating silk protein synthesis.

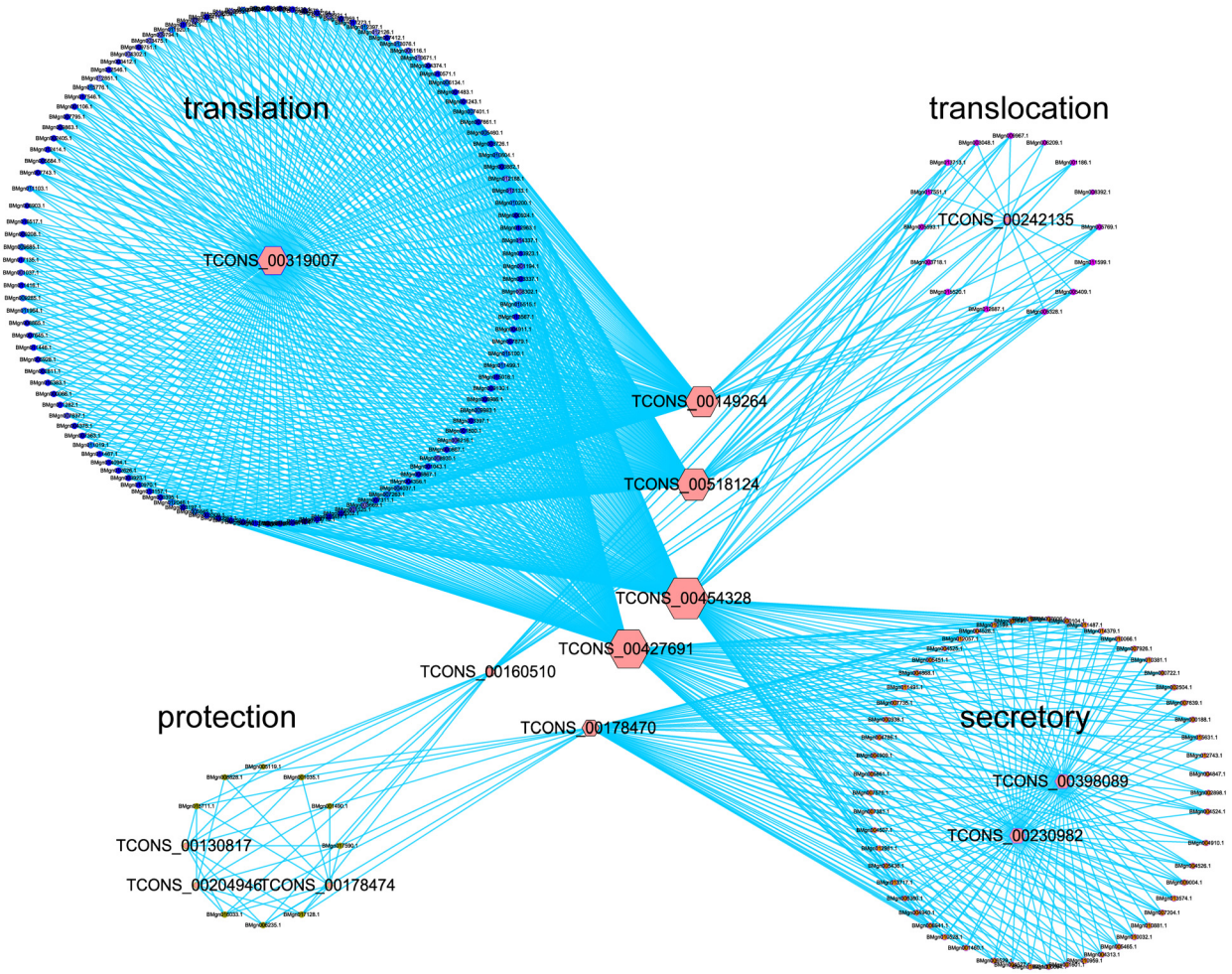

Fig 7. Network visualization of sub-networks derived from the darkgoldenrod module. Blue circular nodes represent protein-coding mRNAs and red hexagon nodes represent lncRNA hub genes. The network was grouped into four sub-networks (translation, translocation, secretory, and protection), displayed as four large circles. The node size represents the degree of connectivity of a particular gene. This image was created using Cytoscape software [63].

doi:10.1371/journal.pone.0147147.g007

Silk proteins are mainly synthesized in MSG and PSG. In 48~96-hour-old fifth instar larvae, an increase in the numbers of rough ER and Golgi vacuoles were observed [87]. Based on comparative proteomic analysis, the Zhong group showed that the aminoacyl-tRNA biosynthesis, ribosome, and secretory pathways are significantly enriched in MSG and PSG at the third day of the fifth instar larval stage [88]. Previously, using SAGE-aided transcriptomic analysis, the Pierre group found highly abundant transcripts, in both MSG and PSG cells, which encoded ribosomal proteins and translation factors [89]. Moreover, small RNA-seq of PSG showed that some miRNAs may be involved in the synthesis of silk protein [33]. Collectively, the 13 hub lncRNAs, especially the six inter-network hub lncRNAs, may function as regulators of silk protein biosynthesis and secretion.

Conclusion

In the present study, we identified 11,810 lncRNAs in the silkworm, including 6, 250 lincRNAs, 474 ilncRNAs, and 5,086 lncNATs, by integrative analysis of 21 relatively high-depth and

high-quality RNA-seq libraries. The genomic features of the identified lncRNAs were examined. We found that silkworm lncRNAs were shorter in terms of overall length, with longer exons and introns, smaller exon pre-transcripts, harbored more transposons, higher SNP density, and relatively low levels of expression compared with silkworm protein-coding mRNAs. Several limitations were existed in the present study. As most of the RNA-seq libraries generated were non strand-specific and poly(A)-selected, a large proportion of non-poly(A) silkworm lncRNAs were not detected. In addition, single-exon transcripts were excluded due to lack of availability of strand information. The number of ilncRNAs and lncNATs may have been underestimated, and the transcript lengths and exon numbers per transcript of lncRNAs overestimated. Additionally, some bona fide transcripts may have been lost. In the current study, we analyzed sequence conservation rather than positional conservation. Therefore, further investigation based on phylogenomics approaches are warranted for elucidation of the functional roles of conserved lncRNAs.

Like lncRNAs in other species, silkworm lncRNAs tend to be expressed in a tissue-specific manner, and may function as miRNA precursors or ceRNAs. Sexual dimorphism was also investigated, and 49.47% of lncRNA loci (2,749) were found to be expressed in a sex-biased manner. Co-expression network analysis showed that 12 out of 19 modules exhibited relatively high association with specific tissue types. Moreover, functional enrichment results suggested that the tissue-associated modules are biologically meaningful and related to tissue-specific biological processes. In-depth analysis of the highlighted darkgoldenrod module, which is specifically associated with the middle and posterior silk gland where main places of silk protein biosynthesis is, suggested that the hub lncRNAs of this module may function as regulators of silk protein biosynthesis, translocation, and secretion. This study presents the first comprehensive genome-wide analysis of silkworm lncRNAs and provides an invaluable resource for genetic, evolutionary, and genomic studies of the silkworm. Moreover, our findings are expected to provide new insights into the mechanisms underlying the biosynthesis of silk protein.

Supporting Information

S1 Fig. Genomic characterization of silkworm lncRNAs. (A) Distribution of lncRNAs along each chromosome: (a) percentage of repetitive sequences in 200-kb windows; (b) number of miRNAs in 200-kb windows; (c) number of protein-coding mRNAs in 200-kb windows; (d) number of lincRNAs in 200-kb windows; (e) number of lncNATs in 200-kb windows; (f) number of ilncRNAs in 200-kb windows. (B) Seqlogo of nucleotide frequencies at donor and acceptor sites of lncRNAs and protein-coding mRNAs. (C) Conservation of silkworm lncRNAs and protein-coding mRNAs. (a) The heatmap presents the homolog sequence fragment identified across 17 other selected genomes for lncRNAs. (b) The heatmap presents the homolog sequence fragment identified across 17 other selected genomes for protein-coding mRNAs. (c) The number of homolog sequences discovered for each lncRNA and protein-coding mRNAs. (D) The cumulative density of lncNAT sequences overlapped by mRNAs. (E) SNP density of different types of transcripts.
(PDF)

S2 Fig. Characteristics of lncRNA expression in silkworm tissues. (A) Distribution of expression log10 (FPKM+1) of lincRNA (red), lncNAT (blue) and protein-coding (green) mRNAs in 21 silkworm tissues. (B) Density distribution of maximum expression levels for lincRNAs, lncNATs, and protein-coding mRNAs in 21 different silkworm tissues.
(PDF)

S3 Fig. LncRNAs as potential competing endogenous RNAs. (A) Schematic of conservation elements region of XLOC_004695 gene locus that harbor bmo-miR-184 response elements. The purple box shows the region contains conservation sequence elements. The consensus logo highlights the 240-bp conserved sequence, which was identified from the 8 insect genome alignments. The sequences alignments represent for bmo-miR-184 response elements, the above vertical lines indicating Watson—Crick base pairs. (B) The expression pattern of potential competing endogenous RNA (TCONS_00111202) in the 21 silkworm tissues.
(PDF)

S4 Fig. Sex-biased alternatively spliced variant expression of lncRNA locus XLOC_012091 in silkworm. (A) Gene structure of XLOC_012091 locus and its antisense overlap protein-coding gene BMgn007254.1. (B) Expression pattern of transcript isoforms of the XLOC_012091 gene locus.
(PDF)

S5 Fig. WGCNA modules and relationship between module eigengenes (MEs) and the traits. (A) Relationships between module eigengenes (MEs) and traits. Horizontally, MEs are named according to module color. Vertically, traits of interest are listed (Sex, Female, Male, segments of silk gland (ASG, AMSG, MMSG, PMSG, and PSG), combination of adjacent silk gland parts, and other type of tissues). The correlation coefficients between the respective ME and the trait of interest, and the corresponding *p*-values (in parentheses), are shown in the boxes. The deeper the red color of the box, the more positive the correlation with the trait. Inversely, a deeper shade of green indicates a more negative correlation with the trait. (B) Relationships between silk gland-specific module eigengenes (MEs) and traits. Traits of interest are listed (Female, Male, combination of adjacent silk gland parts). MpPSG, PMSG+PSG; MmpPSG, (MMSG+PMSG+PSG); MmpSG, (MMSG+PMSG); AMaSG, (ASG_AMSG).
(PDF)

S6 Fig. Expression pattern of all genes in 12 selected modules across all 21 tissues. Heatmap in the upper panel showing the expression pattern of all genes in this module across all 21 tissues. Red, representing increased expression; black, representing neutral expression; green, representing decreased expression. Barplot in the middle panel showing the values of the module eigengene versus each tissues. Pie charts in the bottom panel indicating the number of mRNAs and lncRNAs within this module.
(PDF)

S1 Table. RNA-seq datasets.
(XLSX)

S2 Table. Genomic information for the identified silkworm lncRNAs (GTF format).
(XLSX)

S3 Table. The chromosome distribution of silkworm lncRNAs.
(XLSX)

S4 Table. Summary of lncRNAs function as putative miRNA precursors.
(XLSX)

S5 Table. Summary of lncRNAs with putative miRNA response regions.
(XLSX)

S6 Table. Summary of sex-biased expression of lncRNA and mRNA.
(XLSX)

S7 Table. Module members and functional enrichment of protein-coding genes in each modules.
(XLSX)

S8 Table. Sub-networks and gene annotation of darkgoldenrod module is shown in Fig 7.
(XLSX)

Acknowledgments

We thank the anonymous reviewers for their many insightful comments and suggestions. We thank Dr. Erjun Ling for his editorial assistance. We greatly thank Dr. Ying Lin for her kind supports in the experiments of this work. We thank Dr. Xingtan Zhang for his thoughtful discussions.

Author Contributions

Conceived and designed the experiments: YW TC QX. Performed the experiments: CL QZ RL. Analyzed the data: YW DL TC. Contributed reagents/materials/analysis tools: CL PZ QX. Wrote the paper: YW TC QX.

References

1. Kapranov P, Cheng J, Dike S, Nix DA, Duttagupta R, Willingham AT, et al. RNA maps reveal new RNA classes and a possible function for pervasive transcription. Science. 2007; 316(5830):1484–8. Epub 2007/05/19. doi: 10.1126/science.1138341 PMID: 17510325.
2. Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, et al. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome research. 2012; 22(9):1775–89. doi: 10.1101/gr.132159.111 PMID: 22955988; PubMed Central PMCID: PMCPMC3431493.
3. Ponting CP, Oliver PL, Reik W. Evolution and functions of long noncoding RNAs. Cell. 2009; 136 (4):629–41. Epub 2009/02/26. doi: 10.1016/j.cell.2009.02.006 PMID: 19239885.
4. Bonasio R, Shiekhattar R. Regulation of transcription by long noncoding RNAs. Annual review of genetics. 2014; 48:433–55. doi: 10.1146/annurev-genet-120213-092323 PMID: 25251851; PubMed Central PMCID: PMCPMC4285387.
5. Necsulea A, Soumillon M, Warnefors M, Liechti A, Daish T, Zeller U, et al. The evolution of lncRNA repertoires and expression patterns in tetrapods. Nature. 2014; 505(7485):635–40. PMID: 24463510.
6. Iyer MK, Niknafs YS, Malik R, Singhal U, Sahu A, Hosono Y, et al. The landscape of long noncoding RNAs in the human transcriptome. Nature genetics. 2015; 47(3):199–208. PMID: 25599403; PubMed Central PMCID: PMCPMC4417758.
7. Sun J, Lin Y, Wu J. Long non-coding RNA expression profiling of mouse testis during postnatal development. PloS one. 2013; 8(10):e75750. doi: 10.1371/journal.pone.0075750 PMID: 24130740; PubMed Central PMCID: PMCPMC3794988.
8. Li L, Eichten SR, Shimizu R, Petsch K, Yeh CT, Wu W, et al. Genome-wide discovery and characterization of maize long non-coding RNAs. Genome biology. 2014; 15(2):R40. doi: 10.1186/gb-2014-15-2-r40 PMID: 24576388; PubMed Central PMCID: PMCPMC4053991.
9. Zhang YC, Liao JY, Li ZY, Yu Y, Zhang JP, Li QF, et al. Genome-wide screening and functional analysis identify a large number of long noncoding RNAs involved in the sexual reproduction of rice. Genome biology. 2014; 15(12):512. doi: 10.1186/s13059-014-0512-1 PMID: 25517485; PubMed Central PMCID: PMCPMC4253996.
10. Pauli A, Valen E, Lin MF, Garber M, Vastenhouw NL, Levin JZ, et al. Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. Genome research. 2012; 22(3):577–91. doi: 10.1101/gr.133009.111 PMID: 22110045; PubMed Central PMCID: PMCPMC3290793.
11. Wang M, Yuan D, Tu L, Gao W, He Y, Hu H, et al. Long noncoding RNAs and their proposed functions in fibre development of cotton (Gossypium spp.). The New phytologist. 2015; 207(4):1181–97. PMID: 25919642.

12. Hao Z, Fan C, Cheng T, Su Y, Wei Q, Li G. Genome-Wide Identification, Characterization and Evolutionary Analysis of Long Intergenic Noncoding RNAs in Cucumber. PloS one. 2015; 10(3):e0121800. doi: 10.1371/journal.pone.0121800 PMID: 25799544; PubMed Central PMCID: PMC4370693.

13. Young RS, Marques AC, Tibbit C, Haerty W, Bassett AR, Liu JL, et al. Identification and properties of 1,119 candidate lincRNA loci in the *Drosophila melanogaster* genome. Genome biology and evolution. 2012; 4(4):427–42. doi: 10.1093/gbe/evs020 PMID: 22403033; PubMed Central PMCID: PMCPMC3342871.

14. Padron A, Molina-Cruz A, Quinones M, Ribeiro JM, Ramphul U, Rodrigues J, et al. In depth annotation of the *Anopheles gambiae* mosquito midgut transcriptome. BMC genomics. 2014; 15:636. doi: 10.1186/1471-2164-15-636 PMID: 25073905; PubMed Central PMCID: PMCPMC4131051.

15. Jayakodi M, Jung JW, Park D, Ahn YJ, Lee SC, Shin SY, et al. Genome-wide characterization of long intergenic non-coding RNAs (lincRNAs) provides new insight into viral diseases in honey bees Apis cerana and Apis mellifera. BMC genomics. 2015; 16(1):680. PMID: 26341079; PubMed Central PMCID: PMCPMC4559890.

16. Jenkins AM, Waterhouse RM, Muskavitch MA. Long non-coding RNA discovery across the genus anopheles reveals conserved secondary structures within and beyond the Gambiae complex. BMC genomics. 2015; 16(1):337. doi: 10.1186/s12864-015-1507-3 PMID: 25903279; PubMed Central PMCID: PMC4409983.

17. Liu J, Jung C, Xu J, Wang H, Deng S, Bernad L, et al. Genome-wide analysis uncovers regulation of long intergenic noncoding RNAs in Arabidopsis. The Plant cell. 2012; 24(11):4333–45. doi: 10.1105/tpc.112.102855 PMID: 23136377; PubMed Central PMCID: PMC3531837.

18. Brown JB, Boley N, Eisman R, May GE, Stoiber MH, Duff MO, et al. Diversity and dynamics of the *Drosophila transcriptome*. Nature. 2014; 512(7515):393–9. doi: 10.1038/nature12962 PMID: 24670639; PubMed Central PMCID: PMCPMC4152413.

19. Jenkins AM, Waterhouse RM, Muskavitch MA. Long non-coding RNA discovery across the genus anopheles reveals conserved secondary structures within and beyond the Gambiae complex. BMC genomics. 2015; 16:337. doi: 10.1186/s12864-015-1507-3 PMID: 25903279; PubMed Central PMCID: PMCPMC4409983.

20. Lakhotia SC, Mallik M, Singh AK, Ray M. The large noncoding hsromega-n transcripts are essential for thermotolerance and remobilization of hnRNPs, HP1 and RNA polymerase II during recovery from heat shock in Drosophila. Chromosoma. 2012; 121(1):49–70. doi: 10.1007/s00412-011-0341-x PMID: 21913129.

21. Deng X, Meller VH. Non-coding RNA in fly dosage compensation. Trends in biochemical sciences. 2006; 31(9):526–32. doi: 10.1016/j.tibs.2006.07.007 PMID: 16890440.

22. Soshnev AA, Ishimoto H, McAllister BF, Li X, Wehling MD, Kitamoto T, et al. A conserved long noncoding RNA affects sleep behavior in *Drosophila*. Genetics. 2011; 189(2):455–68. Epub 2011/07/22. doi: 10.1534/genetics.111.131706 PMID: 21775470; PubMed Central PMCID: PMCPMC3189806.

23. Li M, Wen S, Guo X, Bai B, Gong Z, Liu X, et al. The novel long non-coding RNA *CRG* regulates Drosophila locomotor behavior. Nucleic acids research. 2012; 40(22):11714–27. doi: 10.1093/nar/gks943 PMID: 23074190; PubMed Central PMCID: PMCPMC3526303.

24. Chen Y, Dai H, Chen S, Zhang L, Long M. Highly tissue specific expression of Sphinx supports its male courtship related role in Drosophila melanogaster. PloS one. 2011; 6(4):e18853. doi: 10.1371/journal.pone.0018853 PMID: 21541324; PubMed Central PMCID: PMCPMC3082539.

25. Kiya T, Ugajin A, Kunieda T, Kubo T. Identification of kakusei, a nuclear non-coding RNA, as an immediate early gene from the honeybee, and its application for neuroethological study. International journal of molecular sciences. 2012; 13(12):15496–509. doi: 10.3390/ijms131215496 PMID: 23443077; PubMed Central PMCID: PMCPMC3546645.

26. Tadano H, Yamazaki Y, Takeuchi H, Kubo T. Age- and division-of-labour-dependent differential expression of a novel non-coding RNA, *Nb-1*, in the brain of worker honeybees, Apis mellifera L. Insect molecular biology. 2009; 18(6):715–26. PMID: 19817910.

27. Kiya T, Kunieda T, Kubo T. Inducible- and constitutive-type transcript variants of *kakusei*, a novel non-coding immediate early gene, in the honeybee brain. Insect molecular biology. 2008; 17(5):531–6. PMID: 18691230.

28. Sawata M, Takeuchi H, Kubo T. Identification and analysis of the minimal promoter activity of a novel noncoding nuclear RNA gene, AncR-1, from the honeybee (Apis mellifera L.). Rna. 2004; 10(7):1047–58. doi: 10.1261/rna.5231504 PMID: 15208441; PubMed Central PMCID: PMCPMC1370596.

29. Sawata M, Yoshino D, Takeuchi H, Kamikouchi A, Ohashi K, Kubo T. Identification and punctate nuclear localization of a novel noncoding RNA, Ks-1, from the honeybee brain. Rna. 2002; 8(6):772–85. Epub 2002/06/29. PMID: 12088150; PubMed Central PMCID: PMCPMC1370296.

30. Humann FC, Tiberio GJ, Hartfelder K. Sequence and expression characteristics of long noncoding RNAs in honey bee caste development—potential novel regulators for transgressive ovary size. PloS one. 2013; 8(10):e78915. doi: 10.1371/journal.pone.0078915 PMID: 24205350; PubMed Central PMCID: PMCPMC3814967.

31. Xia Q, Li S, Feng Q. Advances in silkworm studies accelerated by the genome sequencing of *Bombyx mori*. Annu Rev Entomol. 2014; 59:513–36. doi: 10.1146/annurev-ento-011613-161940 PMID: 24160415.

32. Liu S, Li D, Li Q, Zhao P, Xiang Z, Xia Q. MicroRNAs of *Bombyx mori* identified by Solexa sequencing. BMC genomics. 2010; 11:148. doi: 10.1186/1471-2164-11-148 PMID: 20199675; PubMed Central PMCID: PMCPMC2838851.

33. Li J, Cai Y, Ye L, Wang S, Che J, You Z, et al. MicroRNA expression profiling of the fifth-instar posterior silk gland of *Bombyx mori*. BMC genomics. 2014; 15:410. doi: 10.1186/1471-2164-15-410 PMID: 24885170; PubMed Central PMCID: PMCPMC4045974.

34. Li D, Wang Y, Zhang K, Jiao Z, Zhu X, Skogerboe G, et al. Experimental RNomics and genomic comparative analysis reveal a large group of species-specific small non-message RNAs in the silkworm *Bombyx mori*. Nucleic acids research. 2011; 39(9):3792–805. Epub 2011/01/14. doi: 10.1093/nar/gkq1317 PMID: 21227919; PubMed Central PMCID: PMCPMC3089462.

35. Kawaoka S, Kadota K, Arai Y, Suzuki Y, Fujii T, Abe H, et al. The silkworm W chromosome is a source of female-enriched piRNAs. Rna. 2011; 17(12):2144–51. doi: 10.1261/rna.027565.111 PMID: 22020973; PubMed Central PMCID: PMCPMC3222127.

36. Taguchi S, Iwami M, Kiya T. Identification and characterization of a novel nuclear noncoding RNA, Fben-1, which is preferentially expressed in the higher brain center of the female silkworm moth, Bombyx mori. Neuroscience letters. 2011; 496(3):176–80. Epub 2011/04/26. doi: 10.1016/j.neulet.2011.04.011 PMID: 21514361.

37. Kiuchi T, Koga H, Kawamoto M, Shoji K, Sakai H, Arai Y, et al. A single female-specific piRNA is the primary determiner of sex in the silkworm. Nature. 2014; 509(7502):633–6. doi: 10.1038/nature13315 PMID: 24828047.

38. Nie H, Liu C, Cheng T, Li Q, Wu Y, Zhou M, et al. Transcriptome analysis of integument differentially expressed genes in the pigment mutant (quail) during molting of silkworm, Bombyx mori. PloS one. 2014; 9(4):e94185. doi: 10.1371/journal.pone.0094185 PMID: 24718369; PubMed Central PMCID: PMCPMC3981777.

39. Andrews S. A quality control tool for high throughput sequence data. Available: www.bioinformaticsbbsrcacuk/projects/fastqc. 2010.

40. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014; 30(15):2114–20. doi: 10.1093/bioinformatics/btu170 PMID: 24695404; PubMed Central PMCID: PMCPMC4103590.

41. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nature methods. 2012; 9 (4):357–9. PMID: 22388286; PubMed Central PMCID: PMCPMC3322381.

42. Duan J, Li R, Cheng D, Fan W, Zha X, Cheng T, et al. SilkDB v2.0: a platform for silkworm (Bombyx mori) genome biology. Nucleic acids research. 2010; 38(Database issue):D453–6. doi: 10.1093/nar/gkp801 PMID: 19793867; PubMed Central PMCID: PMCPMC2808975.

43. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome biology. 2013; 14(4): R36. doi: 10.1186/gb-2013-14-4-r36 PMID: 23618408; PubMed Central PMCID: PMCPMC4053844.

44. Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A, et al. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes & development. 2011; 25(18):1915–27. Epub 2011/09/06. doi: 10.1101/gad.17446611 PMID: 21890647; PubMed Central PMCID: PMCPMC3185964.

45. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature biotechnology. 2010; 28(5):511–5. PMID: 20436464; PubMed Central PMCID: PMCPMC3146043.

46. Guttman M, Garber M, Levin JZ, Donaghey J, Robinson J, Adiconis X, et al. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. Nature biotechnology. 2010; 28(5):503–10. doi: 10.1038/nbt.1633 PMID: 20436462; PubMed Central PMCID: PMCPMC2868100.

47. Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nature biotechnology. 2015; 33 (3):290–5. PMID: 25690850.

48. Wang L, Park HJ, Dasari S, Wang S, Kocher JP, Li W. CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. Nucleic acids research. 2013; 41(6):e74. doi: 10.1093/nar/gkt006 PMID: 23335781; PubMed Central PMCID: PMCPMC3616698.

49. Kong L, Zhang Y, Ye ZQ, Liu XQ, Zhao SQ, Wei L, et al. CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. Nucleic acids research. 2007; 35 (Web Server issue):W345–9. 0.1093/nar/gkm391. PMID: 17631615; PubMed Central PMCID: PMCPMC1933232.

50. Sun L, Luo H, Bu D, Zhao G, Yu K, Zhang C, et al. Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts. Nucleic acids research. 2013; 41(17):e166. doi: 10.1093/nar/gkt646 PMID: 23892401; PubMed Central PMCID: PMCPMC3783192.

51. UniProt C. UniProt: a hub for protein information. Nucleic acids research. 2015; 43(Database issue): D204–12. doi: 10.1093/nar/gku989 PMID: 25348405; PubMed Central PMCID: PMCPMC4384041.

52. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, et al. Pfam: the protein families database. Nucleic acids research. 2014; 42(Database issue):D222–30. doi: 10.1093/nar/gkt1223 PMID: 24288371; PubMed Central PMCID: PMCPMC3965110.

53. Cunningham F, Amode MR, Barrell D, Beal K, Billis K, Brent S, et al. Ensembl 2015. Nucleic acids research. 2015; 43(Database issue):D662–9. doi: 10.1093/nar/gku1010 PMID: 25352552; PubMed Central PMCID: PMCPMC4383879.

54. Kolde R. pheatmap: Pretty Heatmaps. R package version 102 ed 2015. Available: https://cran.r-project.org/web/packages/pheatmap/index.html.

55. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nature protocols. 2012; 7(3):562–78. PMID: 22383036; PubMed Central PMCID: PMCPMC3334321.

56. Kozomara A, Griffiths-Jones S. miRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic acids research. 2014; 42(Database issue):D68–73. doi: 10.1093/nar/gkt1181 PMID: 24275495; PubMed Central PMCID: PMCPMC3965103.

57. Gong J, Liu W, Zhang J, Miao X, Guo AY. lncRNASNP: a database of SNPs in lncRNAs and their potential functions in human and mouse. Nucleic acids research. 2015; 43(Database issue):D181–6. doi: 10.1093/nar/gku1000 PMID: 25332392; PubMed Central PMCID: PMCPMC4383871.

58. Das S, Ghosal S, Sen R, Chakrabarti J. lnCeDB: database of human long noncoding RNA acting as competing endogenous RNA. PloS one. 2014; 9(6):e98965. doi: 10.1371/journal.pone.0098965 PMID: 24926662; PubMed Central PMCID: PMCPMC4057149.

59. Liu K, Yan Z, Li Y, Sun Z. Linc2GO: a human LincRNA function annotation resource based on ceRNA hypothesis. Bioinformatics. 2013; 29(17):2221–2. doi: 10.1093/bioinformatics/btt361 PMID: 23793747.

60. Betel D, Wilson M, Gabow A, Marks DS, Sander C. The microRNA.org resource: targets and expression. Nucleic acids research. 2008; 36(Database issue):D149–53. doi: 10.1093/nar/gkm995 PMID: 18158296; PubMed Central PMCID: PMCPMC2238905.

61. Kertesz M, Iovino N, Unnerstall U, Gaul U, Segal E. The role of site accessibility in microRNA target recognition. Nature genetics. 2007; 39(10):1278–84. doi: 10.1038/ng2135 PMID: 17893677.

62. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC bioinformatics. 2008; 9:559. doi: 10.1186/1471-2105-9-559 PMID: 19114008; PubMed Central PMCID: PMCPMC2631488.

63. Smoot ME, Ono K, Ruscheinski J, Wang PL, Ideker T. Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics. 2011; 27(3):431–2. doi: 10.1093/bioinformatics/btq675 PMID: 21149340; PubMed Central PMCID: PMCPMC3031041.

64. Young MD, Wakefield MJ, Smyth GK, Oshlack A. Gene ontology analysis for RNA-seq: accounting for selection bias. Genome biology. 2010; 11(2):R14. doi: 10.1186/gb-2010-11-2-r14 PMID: 20132535; PubMed Central PMCID: PMCPMC2872874.

65. Xie C, Mao X, Huang J, Ding Y, Wu J, Dong S, et al. KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases. Nucleic acids research. 2011; 39(Web Server issue): W316–22. doi: 10.1093/nar/gkr483 PMID: 21715386; PubMed Central PMCID: PMCPMC3125809.

66. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC bioinformatics. 2011; 12:323. doi: 10.1186/1471-2105-12-323 PMID: 21816040; PubMed Central PMCID: PMCPMC3163565.

67. Wang L, Feng Z, Wang X, Wang X, Zhang X. DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. Bioinformatics. 2010; 26(1):136–8. doi: 10.1093/bioinformatics/btp612 PMID: 19855105.

68. Zhang K, Huang K, Luo Y, Li S. Identification and functional analysis of long non-coding RNAs in mouse cleavage stage embryonic development based on single cell transcriptome data. BMC

genomics. 2014; 15:845. doi: 10.1186/1471-2164-15-845 PMID: 25277336; PubMed Central PMCID: PMC4200203.

69. Zhou ZY, Li AM, Adeola AC, Liu YH, Irwin DM, Xie HB, et al. Genome-wide identification of long intergenic noncoding RNA genes and their potential association with domestication in pigs. Genome biology and evolution. 2014; 6(6):1387–92. doi: 10.1093/gbe/evu113 PMID: 24891613; PubMed Central PMCID: PMC4079208.

70. Mohammadin S, Edger PP, Pires JC, Schranz ME. Positionally-conserved but sequence-diverged: identification of long non-coding RNAs in the Brassicaceae and Cleomaceae. BMC plant biology. 2015; 15(1):217. doi: 10.1186/s12870-015-0603-5 PMID: 26362138; PubMed Central PMCID: PMC4566204.

71. Ulitsky I, Shkumatava A, Jan CH, Sive H, Bartel DP. Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. Cell. 2011; 147(7):1537–50. Epub 2011/12/27. doi: 10.1016/j.cell.2011.11.055 PMID: 22196729; PubMed Central PMCID: PMCPMC3376356.

72. Cai X, Cullen BR. The imprinted H19 noncoding RNA is a primary microRNA precursor. Rna. 2007; 13 (3):313–6. doi: 10.1261/rna.351707 PMID: 17237358; PubMed Central PMCID: PMCPMC1800509.

73. Ronshaugen M, Biemar F, Piel J, Levine M, Lai EC. The *Drosophila* microRNA *iab-4* causes a dominant homeotic transformation of halteres to wings. Genes & development. 2005; 19(24):2947–52. doi: 10.1101/gad.1372505 PMID: 16357215; PubMed Central PMCID: PMCPMC1315399.

74. Brennecke J, Stark A, Russell RB, Cohen SM. Principles of microRNA-target recognition. PLoS biology. 2005; 3(3):e85. doi: 10.1371/journal.pbio.0030085 PMID: 15723116; PubMed Central PMCID: PMCPMC1043860.

75. Liu S, Gao S, Zhang D, Yin J, Xiang Z, Xia Q. MicroRNAs show diverse and dynamic expression patterns in multiple tissues of *Bombyx mori*. BMC genomics. 2010; 11:85. doi: 10.1186/1471-2164-11-85 PMID: 20122259; PubMed Central PMCID: PMCPMC2835664.

76. Cesana M, Cacchiarelli D, Legnini I, Santini T, Sthandier O, Chinappi M, et al. A long noncoding RNA controls muscle differentiation by functioning as a competing endogenous RNA. Cell. 2011; 147 (2):358–69. Epub 2011/10/18. doi: 10.1016/j.cell.2011.09.028 PMID: 22000014; PubMed Central PMCID: PMCPMC3234495.

77. Parker BJ, Moltke I, Roth A, Washietl S, Wen J, Kellis M, et al. New families of human regulatory RNA structures identified by comparative analysis of vertebrate genomes. Genome Res. 2011; 21 (11):1929–43. Epub 2011/10/14. doi: 10.1101/gr.112516.110 PMID: 21994249; PubMed Central PMCID: PMC3205577.

78. Peng J, Wang C, Wan C, Zhang D, Li W, Li P, et al. miR-184 is Critical for the motility-related PNS development in *Drosophila*. International journal of developmental neuroscience: the official journal of the International Society for Developmental Neuroscience. 2015; 46:100–7. doi: 10.1016/j.ijdevneu.2015.07.006 PMID: 26306777.

79. Iovino N, Pane A, Gaul U. miR-184 has multiple roles in *Drosophila* female germline development. Developmental cell. 2009; 17(1):123–33. doi: 10.1016/j.devcel.2009.06.008 PMID: 19619497.

80. Tattikota SG, Rathjen T, Hausser J, Khedkar A, Kabra UD, Pandey V, et al. miR-184 Regulates Pancreatic beta-Cell Function According to Glucose Metabolism. The Journal of biological chemistry. 2015; 290(33):20284–94. doi: 10.1074/jbc.M115.658625 PMID: 26152724; PubMed Central PMCID: PMCPMC4536436.

81. Xia Q, Cheng D, Duan J, Wang G, Cheng T, Zha X, et al. Microarray-based gene expression profiles in multiple tissues of the domesticated silkworm, Bombyx mori. Genome biology. 2007; 8(8):R162. doi: 10.1186/gb-2007-8-8-r162 PMID: 17683582; PubMed Central PMCID: PMCPMC2374993.

82. Ferguson LC, Green J, Surridge A, Jiggins CD. Evolution of the insect *yellow* gene family. Molecular biology and evolution. 2011; 28(1):257–72. doi: 10.1093/molbev/msq192 PMID: 20656794.

83. Xia AH, Zhou QX, Yu LL, Li WG, Yi YZ, Zhang YZ, et al. Identification and analysis of YELLOW protein family genes in the silkworm, *Bombyx mori*. BMC genomics. 2006; 7:195. doi: 10.1186/1471-2164-7-195 PMID: 16884544; PubMed Central PMCID: PMCPMC1553450.

84. Noh MY, Kramer KJ, Muthukrishnan S, Beeman RW, Kanost MR, Arakane Y. Loss of function of the *yellow-e* gene causes dehydration-induced mortality of adult Tribolium castaneum. Developmental biology. 2015; 399(2):315–24. doi: 10.1016/j.ydbio.2015.01.009 PMID: 25614237.

85. Liao Q, Liu C, Yuan X, Kang S, Miao R, Xiao H, et al. Large-scale prediction of long non-coding RNA functions in a coding-non-coding gene co-expression network. Nucleic acids research. 2011; 39 (9):3864–78. doi: 10.1093/nar/gkq1348 PMID: 21247874; PubMed Central PMCID: PMCPMC3089475.

86. Zhao XM, Liu C, Jiang LJ, Li QY, Zhou MT, Cheng TC, et al. A juvenile hormone transcription factor Bmdimm-fibroin H chain pathway is involved in the synthesis of silk protein in silkworm, *Bombyx mori*. The Journal of biological chemistry. 2015; 290(2):972–86. doi: 10.1074/jbc.M114.606921 PMID: 25371208; PubMed Central PMCID: PMCPMC4294524.

87. Tashiro Y, Morimoto T, Matsuura S, Nagata S. Studies on the posterior silk gland of the silkworm, *Bombyx mori*. I. Growth of posterior silk gland cells and biosynthesis of fibroin during the fifth larval instar. The Journal of cell biology. 1968; 38(3):574–88. PMID: 5664226; PubMed Central PMCID: PMCPMC2108375.

88. Li JY, Ye LP, Che JQ, Song J, You ZY, Yun KC, et al. Comparative proteomic analysis of the silkworm middle silk gland reveals the importance of ribosome biogenesis in silk protein production. Journal of proteomics. 2015; 126:109–20. doi: 10.1016/j.jprot.2015.06.001 PMID: 26051239.

89. Royer C, Briolay J, Garel A, Brouilly P, Sasanuma S, Sasanuma M, et al. Novel genes differentially expressed between posterior and median silk gland identified by SAGE-aided transcriptome analysis. Insect biochemistry and molecular biology. 2011; 41(2):118–24. doi: 10.1016/j.ibmb.2010.11.003 PMID: 21078388.