

FEBS  
Lettersjournal homepage: [www.FEBSLetters.org](http://www.FEBSLetters.org)

# Identification and evolution of the orphan genes in the domestic silkworm, *Bombyx mori*



Wei Sun, Xin-Wei Zhao, Ze Zhang\*

Laboratory of Evolutionary and Functional Genomics, School of Life Sciences, Chongqing University, Chongqing 400044, China

## ARTICLE INFO

### Article history:

Received 1 July 2015

Revised 24 July 2015

Accepted 1 August 2015

Available online 18 August 2015

Edited by Takashi Gojobori

### Keywords:

Orphan gene

Silkworm

Expression

Evolution

## ABSTRACT

**Orphan genes (OGs) which have no recognizable homology to any sequences in other species could contribute to the species specific adaptations. In this study, we identified 738 OGs in the silkworm genome. About 31% of the silkworm OGs is derived from transposable elements, and 5.1% of the silkworm OGs emerged from gene duplication followed by divergence of paralogs. Five *de novo* silkworm OGs originated from non-coding regions. Microarray data suggested that most of the silkworm OGs were expressed in limited tissues. RNA interference experiments suggested that five *de novo* OGs are not essential to the silkworm, implying that they may contribute to genetic redundancy or species-specific adaptation. Our results provide some new insights into the evolutionary significance of the silkworm OGs.**

© 2015 Federation of European Biochemical Societies. Published by Elsevier B.V. All rights reserved.

## 1. Introduction

The birth of new genes makes contribution to the variation of the gene numbers in different organisms. It is well known that new genes with novel functions may supply wealthy genetic resources to promote evolution of a genome and morphological diversity among different species. Susumu Ohno considered gene duplication as a major mechanism to generate new genes [1]. The genes formed by gene duplication can always be grouped into gene families. Though different copies may have divergent functions, they retain significant sequence similarity.

Moreover, many studies showed that genomes contain another kind of genes, orphan genes (OGs) [2,3]. OGs are defined as the genes that have no recognizable homology to any sequences in other species. Therefore, OGs are always present in a restricted phylogenetic lineage. Since the release of more and more genome sequences, OGs are widely identified in all domains of life and viruses [3]. The percentage of OGs varies enormously between species [4]. In addition, the majority of OGs are single copy in one

genome. They can also contain multiple copies which may have lineage-specific functions. Although the functions of most OGs are still unclear, it is thought that they play very important roles in species specific developmental adaptations. In *Hydra* sp., OGs can regulate tentacle formation [5]. One *Arabidopsis* OG (Qua-Quine Starch, QQS) is thought to be a regulator of starch biosynthesis [6]. A yeast OG (BSC4) may be involved in the DNA repair pathway [7]. In addition, OGs can also take part in the interactions with environments. For example, previous studies showed that the OGs in *Daphnia pulex* become specifically activated in response to environmentally stimuli [8]. And the *Arabidopsis thaliana* OGs are enriched for responses to a wide range of abiotic stresses [9]. Previous studies have shown that several mechanisms could explain the emergence of OGs, including gene duplication, frame-shift mutations, gene fusion and fission, exon shuffling and domestication from transposable elements (TEs) [9,10]. All the OGs by these mechanisms shown above are derived from present parental genes. Besides, the genes could also originate *de novo* from intergenic regions [3]. Despite several *de novo* originated OGs have been discovered in different species, the question that how they emerged from ancestral non-coding sequences is still obscure [3].

As mentioned above, all sequenced genomes contain OGs. However, the function and evolution of the OGs in the domestic silkworm, *Bombyx mori*, remain unknown. As more and more genomic and transcriptomic resources are available, the domestic silkworm is used as a model for the genetics study of the Lepidoptera. To date, five lepidopteran genomes have been

**Abbreviations:** OGs, orphan genes; TE, transposable element; pI, isoelectric point; RNAi, RNA interference

**Author contributions:** W.S. designed the study, performed bioinformatics analysis and experiments, and drafted the manuscript. X.W.Z. did partial data analysis. Z.Z. supervised the study and revised the manuscript.

\* Corresponding author.

E-mail address: [zezhang@cqu.edu.cn](mailto:zezhang@cqu.edu.cn) (Z. Zhang).

<http://dx.doi.org/10.1016/j.febslet.2015.08.008>

0014-5793/© 2015 Federation of European Biochemical Societies. Published by Elsevier B.V. All rights reserved.

released which may help us to systematically identify the OGs in the silkworm [11–14]. Thus, in this study, we firstly performed comparative genomic analyses to identify the OGs in the silkworm genome. Then, we further analyze the origin of the OGs and perform a comprehensive analysis of expression patterns in different silkworm tissues and in various developmental stages using microarray data and real-time polymerase chain reaction (PCR). All these results will provide important information for understanding the evolution and functions of the silkworm OGs.

## 2. Materials and methods

### 2.1. Identification of the silkworm OGs

In this study, 47 arthropod species (44 insects and 3 non-insecta species) and National Center for Biotechnology Information (NCBI) non-redundant (nr) protein sequences were used to do the comparative genomic analyses (Supplementary Table S1). The silkworm predicted proteins were collected from silkworm genome database (SilkDB) [15] and NCBI. The method to identify the silkworm OGs is the same as shown in previous studies [10,16]. Briefly, all silkworm proteins were firstly searched using BLASTP against the protein sequences of other four lepidopteran insects, i.e. *Manduca sexta*, *Danaus plexippus* and *Heliconius melpomene* (<http://agripestbase.org/manduca>) [11,13,14]. If the silkworm protein sequence has BLAST hit in other species with an expectation value smaller than  $10^{-3}$ , as previously described [10,16], it was discarded. Then, the remaining sequences (silkworm protein dataset1 in Fig. 1) were used as queries to do tBlastN against lepidopteran EST sequences (excluding *B. mori*) from NCBI expressed sequence tags (EST) databases, WildSilkBase (<http://www.cdfd.org.in/wild-silkbase/>) [17], ButterflyBase (<http://butterflybase.ice.mpg.de/>) [18], and SPODOBASE (<http://bioweb.ensam.inra.fr/spodobase/>) [19]. The homologous sequences were also discarded for the next analyses. Furthermore, the silkworm protein dataset2 was searched against other insect proteins and the proteins from NCBI nr database using BLASTP. Finally, the silkworm orphan genes which we could not find homologs in any other species were identified. It should be pointed out that the domestic silkworm OGs may be also present in its wild relative *Bombyx mandarina* due to the very close relationship between them (the divergence time is only about 5000 years) [20]. Therefore, here, the silkworm OGs represent the specific genes in the domestic and wild silkworms.

In addition, we also used the same method as shown in Fig. 1 to identify OGs in other four lepidopteran insects, respectively.

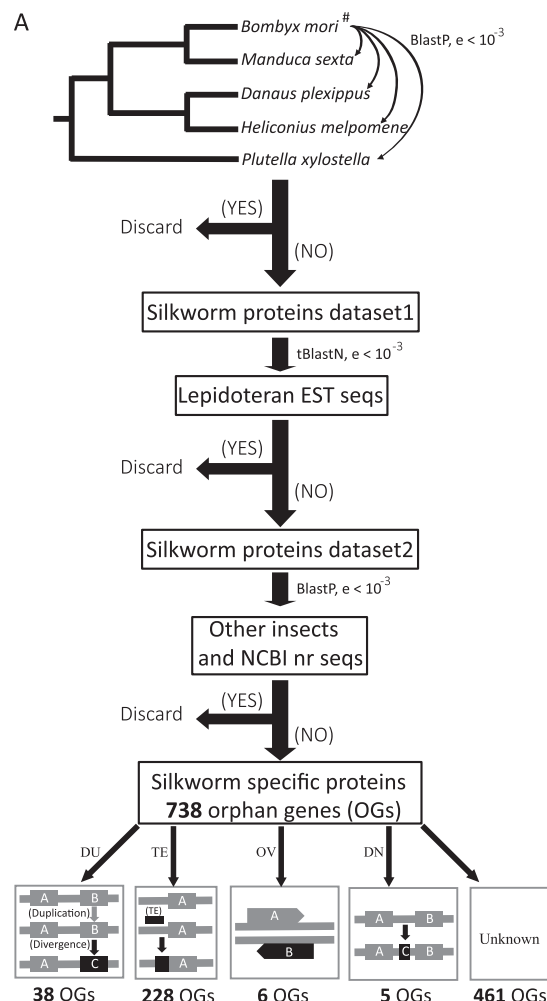
### 2.2. Origin of the silkworm OGs

Previous studies showed that at least seven different mechanisms could explain how OGs emerged [10,21,22]. In this study, we used similar methods to identify the origin of the silkworm OGs.

Paralogs of the silkworm OGs were inferred using BLASTP searches against all silkworm proteins. As shown in previous study, BLASTP cutoff E value was 0.001, and the hit sequences are defined as gene duplicates [21].

To identify the OGs overlapped with transposable elements (TEs), we firstly downloaded all silkworm conserved TE sequences from the silkworm TE database (BmTEdb) [23]. Then, the nucleotide sequences of the silkworm OGs were used as queries to do BLASTN searches against the silkworm TE sequences with an E value cutoff of  $10^{-5}$ .

For the *de novo* originated genes, the silkworm orphan proteins were used to do TBLASTN against other four lepidopteran insect genomes to identify orthologous sequences. Silkworm orphan proteins which have orthologous regions with over 50% of sequence



**Fig. 1.** Method for the identification of OGs in the silkworm genome. OG represents orphan genes; DU represents duplication originated OGs; TE represents transposable elements derived OGs; OV represents overlapping gene models; DN represents *de novo* originated OGs.

identity and covering at least 50% of the length of the gene in the other lepidopteran genomes were kept for further analysis. Then the hits were manually checked one by one. According to the criterion to identify the *de novo* genes in a previous study, the candidate silkworm *de novo* genes must have been disrupted in all other lepidopteran genomes [22]. For example, other lepidopteran orthologous sequences lack translation start codon or have frame shift mutations or indels that result in a premature stop codon. In addition, the protein lengths of other lepidopteran genes that have premature stop codons should be shorter than 50% of the length of the silkworm candidates. Based on the above filter methods, the remaining silkworm OGs are thought to be the silkworm *de novo* originated genes. Furthermore, MUSCLE was used to align the nucleotide sequences of each silkworm *de novo* genes with the homologous regions in other lepidopteran insects [24].

For the other mechanisms, such as overlapping gene models, non-deleterious frame shift, alternative reading frames and horizontal gene transfer, the methods are the same as the previous study [21].

### 2.3. Data analysis

GC contents and exon numbers of the silkworm OGs as well as numbers of amino acids and isoelectric points of the silkworm OG

encoding proteins were calculated and compared with those of the silkworm non-OGs. DAMBE software was used to measure the GC content, number of amino acids and isoelectric point [25]. Chromosomal distribution of the silkworm OGs was also analyzed.

#### 2.4. Microarray analysis

A previous study has customized a genome-wide oligonucleotide microarray with 22987 probes of 9 tissues on day 3 of the fifth instar in the domestic silkworm [26]. The nucleotide sequences of the silkworm OGs were used for BLAST searching against the silkworm probe database (SilkDB) [15] to identify a specific probe for each OG. In order to compare the spatial expression profiles between silkworm OGs and non-OGs, the expression breadth (the number of tissues in which a gene is transcribed) was calculated for each gene. According to the previous analysis, a gene was considered to be expressed in a tissue if its signal intensity exceeded a value of 400 [26]. Hierarchical clustering of gene expression patterns was performed using DNA-Chip Analyzer (dChip) [27]. All the statistical analyses in this study were performed in the statistical R package.

#### 2.5. Gene expression analysis

The DaZao strain of silkworm was used to survey the expression profiles of the silkworm *de novo* genes. For temporal expression analysis: egg, larvae, pupae or adults were collected at different developmental time points. For every time point, at least five individuals were pooled together and then frozen immediately in liquid nitrogen. For spatial expression analysis: nine main tissues were dissected from Day 3 of the fifth instar larvae, and frozen immediately in liquid nitrogen. Every tissue sample was collected from more than five larvae. The samples were homogenized in liquid nitrogen to powders. Total RNA of every sample was extracted by EasyPure RNA purification kit (Transgen Biotech, China) and treated with DNase I (Promega, USA) to remove the genomic DNA contamination. RNA was quantified by UV spectrum absorbance and reverse-transcribed into the first strand cDNA by an M-MLV Reverse Transcriptase Kit (Invitrogen). For expression analysis, the specific amplification primers for the *de novo* genes are shown in [Supplementary Table S2](#). The PCR products were sequenced to confirm the specificity of the primers.

#### 2.6. RNA interference

Based on the cDNA sequences of the silkworm *de novo* genes, we designed specific primers containing T7 promoter sequence. The primers were listed in [Supplementary Table S2](#). Double-stranded RNAs (dsRNAs) were generated based on our previous paper [28]. According to the temporal expression patterns of the silkworm *de novo* genes, 50 µg dsRNAs of these genes in ten microliter solutions were injected into the hemocoel of the silkworm at different developmental points, respectively. Then the phenotypes were surveyed. The mRNA levels of the targeted genes were investigated 12 h after dsRNAs injection using reverse transcription PCR as described above.

### 3. Results and discussion

#### 3.1. Identification of the silkworm OGs

OGs are always considered as important genes to understand the species-specific adaptive process [3]. Up to now, although OGs have been characterized in many organisms, no analysis has ever been performed in the domestic silkworm. Therefore, combined with other 46 arthropod genomes, we firstly identified the

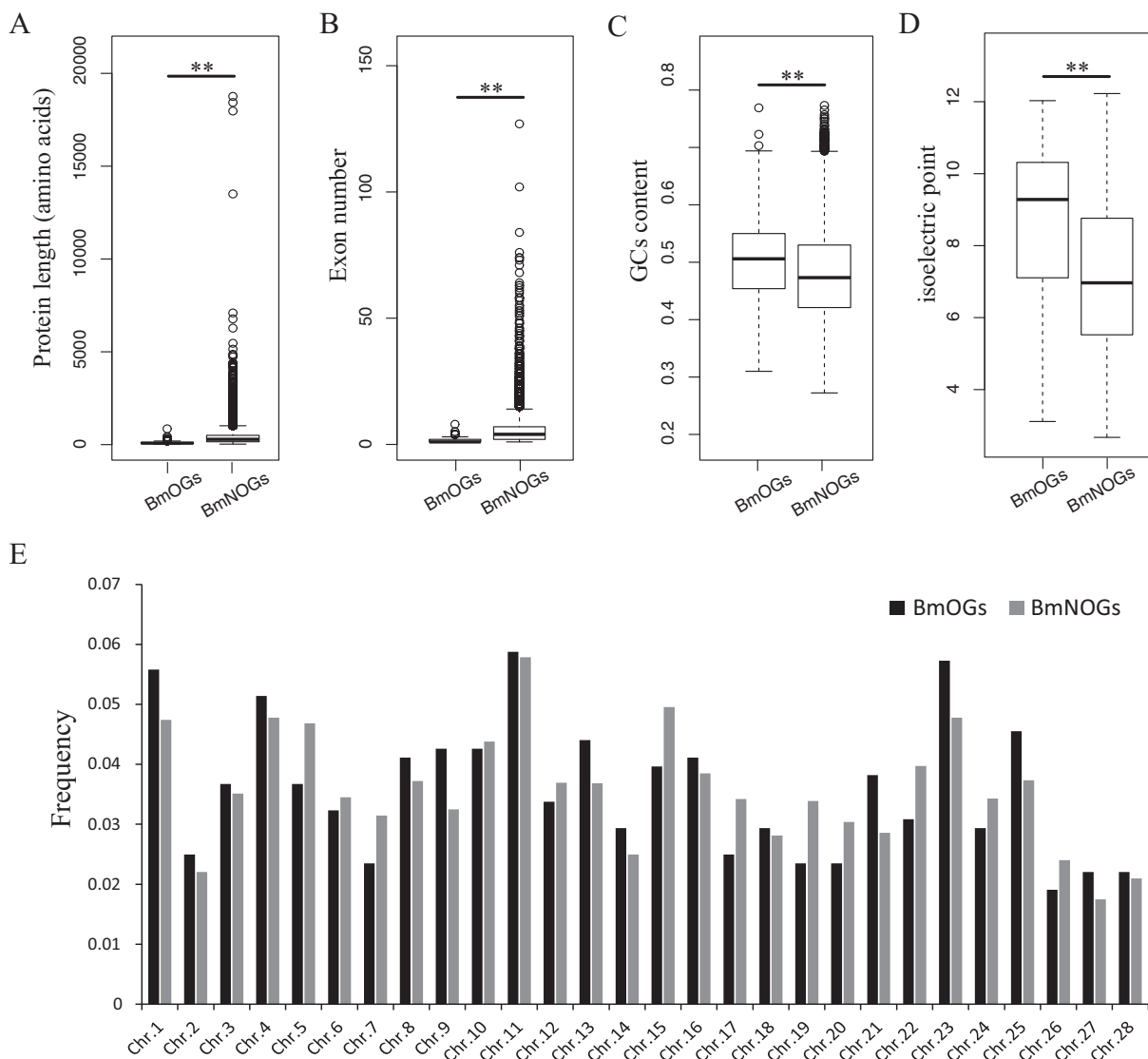
silkworm OGs using BLAST which is widely used in previous studies [3,10,16]. Following the procedure described in [Fig. 1](#), we finally identified 738 silkworm OGs ([Fig. 1](#); [Supplementary Table S3](#)). Recently, one previous study identified the OGs in Diptera and Hymenoptera by comparing 30 arthropod genomes [21]. Meanwhile, they also uncovered 2701 OGs in the domestic silkworm. Tautz and Domazet-Lošo (2011) suggested that the reference genomes from closely related species are very important to identification of OGs [3]. However, no other lepidopteran sequences were used to do the comparative analyses in that study, which may result in much more false-positives. Hence, more reference genomes and lepidopteran sequences used in this study may obtain a more exact annotation.

In addition, using the same approach, we also identified 649, 328, 528 and 1188 OGs in *M. sexta*, *H. melpomene*, *D. plexippus* and *P. xylostella*, respectively ([Supplementary Table S3](#)). For lepidopteran insects, OGs account for about 4.3% of all the genes, and this proportion is similar to *Drosophila* (4.4%), but less than hymenopteran insects (10.2%) which experienced high rates of OG gains [21]. Among lepidopteran insects, *P. xylostella* has the largest number of OGs. *P. xylostella* is a basal lepidopteran species [13,29], and the distant relationship between *P. xylostella* and other four species (125 Mya) may affect the accuracy of the identification of the OGs [3].

#### 3.2. Features of the silkworm OGs

Compared with the conserved genes present in all species, the origination time of OGs is likely to be shorter on average. Some features of OGs may be different from the non-OGs, such as protein length, exon number and GC content. To determine whether these features exist, we characterized the genic properties of the silkworm OGs. Compared with the non-orphan proteins, orphan proteins in silkworm have shorter protein length ([Fig. 2A](#)). The average protein length of non-orphan proteins is 406.84 amino acids which is 4.16 times longer than the orphan proteins (97.87) (Mann–Whitney *U* test,  $P < 0.001$ ). This difference may be due to less exon numbers in OGs (Mann–Whitney *U* test,  $P < 0.001$ ) ([Fig. 2B](#)). In the silkworm, 50.38% of OGs only contain one exon, while the percentage of non-OGs with one exon is 13.20%. The shorter length and less exon number were also observed in other four lepidopteran OGs ([Supplemental Fig. S1](#)). These results are consistent with the previous studies in primates [10], plants [9,30] and other insects [16], suggesting that the two characteristics are common for OGs in all eukaryotic species. However, GCs content may vary among different species. Compared with the non-OGs, *B. mori* and *D. plexippus* OGs have significantly higher GC contents as shown in *Poaceae* [30], while *H. melpomene* and *P. xylostella* have significantly less GC contents, which is similar to the observations in fruit fly and zebrafish (Mann–Whitney *U* test, *P* values are much smaller than 0.001 in all tests) ([Fig. 2C](#); [Supplemental Fig. S1](#)) [16,31]. OGs always have unusual GC contents. However, the GC contents of the *M. sexta* OGs are not significantly different from the genome level. The reason for this is still unknown.

The theoretical pI (isoelectric point) of a protein is important for solubility, subcellular localization and interaction [32,33]. Thus, the shift of the value is always considered as the changes of the protein functions [34]. We found that silkworm orphan proteins have significantly higher average pI value ( $9.18 \pm 2.41$ ) than non-orphan proteins ( $7.62 \pm 2.09$ ) (Mann–Whitney *U* test,  $P < 0.001$ ) ([Fig. 2D](#)). Previous studies have shown that some factors, such as selection, can drive the shift of the pI [35]. For example, the changes in the pI of prokaryotic proteins may be due to adaptation to various environments [33,36]. Though the functions are still poorly characterized, OGs are thought to play very important roles



**Fig. 2.** Box-plot comparisons of protein length (A), exon number (B), GCs content (C), isoelectric point (D) and chromosomal distribution (E) for the OGs and non-OGs in the silkworm. BmOGs mean silkworm OGs; BmNOGs mean the silkworm non-OGs; Chr means chromosome. Mann-Whitney *U* test was used for the statistical analyses. Statistical significance: \**P* < 0.001.

in the split of the species. Therefore, the elevated *pi* values observed in the silkworm orphan proteins may partly reflect their newly evolved functions which may be important for the species-specific adaptation and interaction with the environments.

### 3.3. Emergence of the silkworm OGs

Because of lacking protein sequence similarity in other species, the origin of OGs is a very interesting question and has been investigated in several model organisms [3]. Several emergence mechanisms of OGs were proposed [10,21,31]. In this study, we investigated the emergence of the silkworm OGs based on the available silkworm genome data.

Firstly, we surveyed the chromosome distribution of the silkworm OGs. About 92% of the silkworm OGs (681 OGs) can be located on the silkworm chromosomes. And all 28 chromosomes harbored OGs (Fig. 2E). The number of OGs varies among different chromosomes. However, the percentage of OGs on each chromosome is consistent with the non-OGs, suggesting the silkworm OGs are randomly dispersed on different chromosome without preference.

Then, we further investigated the possible origin of the silkworm OGs and found at least 4 scenarios covering 37.5% of silkworm OGs (277 OGs) (Fig. 1; Supplementary Tables S4–S7). Among them, 38 OGs were from gene duplication followed by divergence of paralogs beyond the threshold of detectable similarity. Twenty-eight paralogs have the functional annotations, while all the duplicated OGs are uncharacterized, further demonstrating the rapid divergence of the OGs (Supplementary Table S4). Gene duplication is thought to be a major mechanism to generate OGs. In zebrafish, 36.4% of the OGs had paralogs [31]. The percentage of duplicated OGs in primate and plant is about 20% [9,10]. However, for insects, the percentage of OGs originated from gene duplication and divergence is much less than the other species (silkworm: 5.1%; ant OGs: 9.9%) [21]. Gene duplication and divergence has long been considered as a major source of genetic novelty and adaptation [37]. Thus, different numbers of duplicated OGs between species may correlate with the adaptations to various environments.

One striking result in this study is a high percentage of the silkworm OGs that contain TE-like sequences. Overall, 30.9% of the silkworm OGs (228 OGs) were generated by overlapping between



TE related regions in their coding sequences (Supplementary Table S5). This proportion is higher than that in *A. thaliana* (9.73%), but less than that in primate (53%) (Table 1) [9,10]. These differences may correlate with the levels of TE contents in whole genomes. Toll-Riera et al. (2009) showed that 93% of TE-related human OGs are exonized from *Alu* elements, a major type of short interspersed nuclear elements (SINEs) [10]. We also detected similar results in the silkworm OGs. The TE derived sequences in 54.39% of the silkworm TE-related OGs belong to the long terminal repeats (LTR retrotransposons) (Table 1). In human, *Alus* account for 10% of the genome and contributed to most of TE exonization events [38,39]. However, though the TEs constitute about 40% of the silkworm genome, the proportion of the LTR retrotransposons in all TEs is very small (1.7%) [40]. Why there is a high percentage of LTR-derived OGs in the silkworm is an interesting question. One possible reason may be the special characteristics of these transposons. LTR transposons always have long length and contain several signals for gene expression, such as promoter, alternative splice site, enhancer and transcription regulatory signals [41]. These features may provide an opportunity to be a part of coding genes. Indeed, 256 human protein coding regions are derived from LTR [42]. Future studies will be required to ascertain whether these TE-derived OGs have the functions and to reveal their relevance in the generation of adaptive evolutionary novelty in the silkworm.

Besides originating from known old genes, OGs can emerge from non-coding DNA region, and this kind of genes are also called *de novo* genes. Several studies have systematically identified *de novo* genes in different species. Nevertheless, the *de novo* originated protein coding genes from non-coding DNA region in the genome are still rare. Levine et al. (2006) showed only 5 *de novo* genes in *D. melanogaster* [43]. Wu et al. (2011) identified 60 *de novo* genes in human [22]. Moreover, *de novo* genes could play essential roles during the evolution of organisms [44,45]. In this study, we also identified five *de novo* originated OGs in the silkworm (Supplementary Table S6). The silkworm and its closest related species with the genome – *M. sexta* shared a common ancestor about 35 million years ago [46]. Using this as a calibration, we estimated the rate of origin of the silkworm *de novo* genes being approximately 0.14 gene per million years. This estimate is much less than the previous reports in fruit fly (1.79 genes per million years) [43] and human (9.83–11.8 genes per million years) [22]. The reason to explain the extremely lower rate of the origin of *de novo* genes in silkworm is still unclear. To further investigate the origin of the silkworm *de novo* genes, we performed the sequence comparison with their homologous non-coding regions from other lepidopteran insects (Fig. 3). Syntenic analyses showed the conserved gene order in the flanking regions of the all five *de novo* genes across the lepidopteran genomes, suggesting that the non-coding regions are vertically inherited from a common

ancestor. Moreover, sequence comparison indicated that similarity between the silkworm *de novo* genes and the corresponding regions in other species is very high (Supplementary Figs. S2–S6). Furthermore, we found that several silkworm specific mutations or indels resulted in the generation of these *de novo* genes. For instance, BGIBMGA004015 and BGIBMGA008845 obtained potential start codon from ATA or ATT in other lepidopteran insects to ATG (Supplementary Figs. S4 and S5). In BGIBMGA001421, a mutation from T to A transition removed a stop codon that is present in the other species (Supplementary Fig. S3). In addition, the orthologous DNA of non-silkworm species also harbors several other mutations and indels. These results further suggested that the ancestral sequences were non-coding.

In conclusion, gene duplication and TEs contributed to the origin of 36% of the silkworm OGs. The proportion is much higher than the previous study in ants (about 17%–22.3%) [21], implying that the origin mechanisms of the OGs are different among species. A plausible explanation may be that different genomes have different genomic structures and compositions (e.g., the TE contents).

### 3.4. Expression and functional analysis of the silkworm OGs

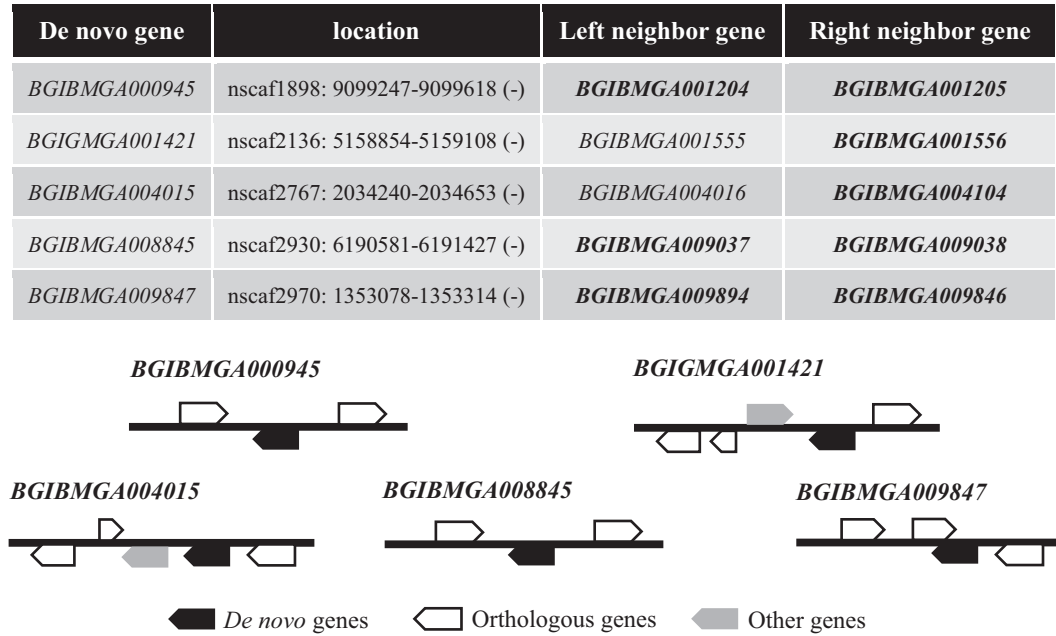
Though a large number of OGs have been identified, their functions are largely still uncharacterized. To reveal the functions of the silkworm OGs, we performed real-time PCR and analyzed previous microarray data to investigate the expression patterns of the genes.

The expression pattern of the gene is effective information to understand its biological function. First, all the silkworm OGs were used as queries to search the EST database and only 374 genes had EST evidence. This database is constructed by 36 cDNA libraries from 17 tissues (excluding cell lines) and developmental stages (embryo, larva, pupa and adult) [47]. We found that silkworm OGs were expressed in significantly less tissues than silkworm NOGs (Mann–Whitney *U* test,  $P < 0.001$ ) (Fig. 4A). For example, 28.2% silkworm OGs expressed in only one tissue, and the proportion is only 11.1% for silkworm NOGs. In addition, microarray data from different silkworm tissues were released [26]. We then used this dataset to survey the spatial expression pattern of the silkworm OGs (Supplementary Fig. S7). About 52% of the silkworm OGs (382) had the probes in the microarray data, and 141 of which had the EST evidence shown above. In total, at least 615 silkworm OGs have expression evidence (EST or microarray data), indicating that these genes may be functional. For the remaining genes, they may be expressed in other developmental stages or tissues that cannot be detected in EST and microarray data. We further surveyed the transcriptional features of the silkworm OGs by the microarray data. In Supplementary Fig. S7, most of the genes (332) were preferentially expressed in some special tissues. The expression breadth of the silkworm OGs is narrower than that of the silkworm NOGs (Mann–Whitney *U* test,  $P < 0.001$ ) (Fig. 4B). For instance, 43 genes showed a mid-gut biased expression pattern. Seventy genes were mainly expressed in testis. Previous studies suggested that OGs tended to express in limited tissues [10,31]. Moreover, we did not find any significant tissue expression bias between duplicated OGs and TE originated OGs (Mann–Whitney *U* test,  $P = 0.921$ ) (Fig. 4C). Nevertheless, all the results shown above suggest that the majority of OGs tend to be expressed in restricted tissues. Previous studies speculated that some special tissues, such as testis and human brain, may be the birthplace of novel genes because their ‘permissive’ environments facilitated the transcription of new genes [48,49].

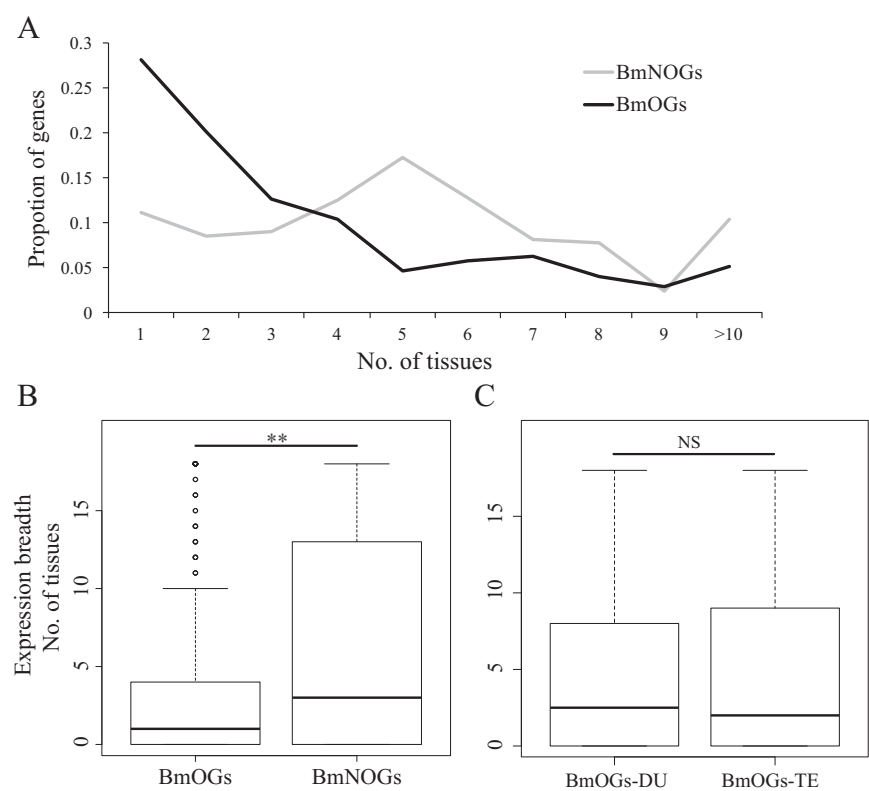
It is generally accepted that *de novo* genes are considered as important resource for the species-specific evolution, though the functions of most genes are still unknown. Thus, we knocked down the expressions of the silkworm *de novo* genes with RNA interference (RNAi) to reveal their importance to possible organismal

**Table 1**  
The proportion of transposable elements in the OGs.

	<i>Bombyx mori</i> (%)	<i>Arabidopsis thaliana</i> (%)	<i>Homo sapiens</i>
TE in genome	40.00	10	46.00
TE related OGs		9.73	53.00
1	LTR 54.39	22.46%	
	LINE 17.54		
	SINE 2.19		93.00
2	TIR 18.42		
	MITE 0.44		
	Helitron 0.44	40.64	
	Other	23.53 (DNA/MuDR)	
Unknown	6.58		



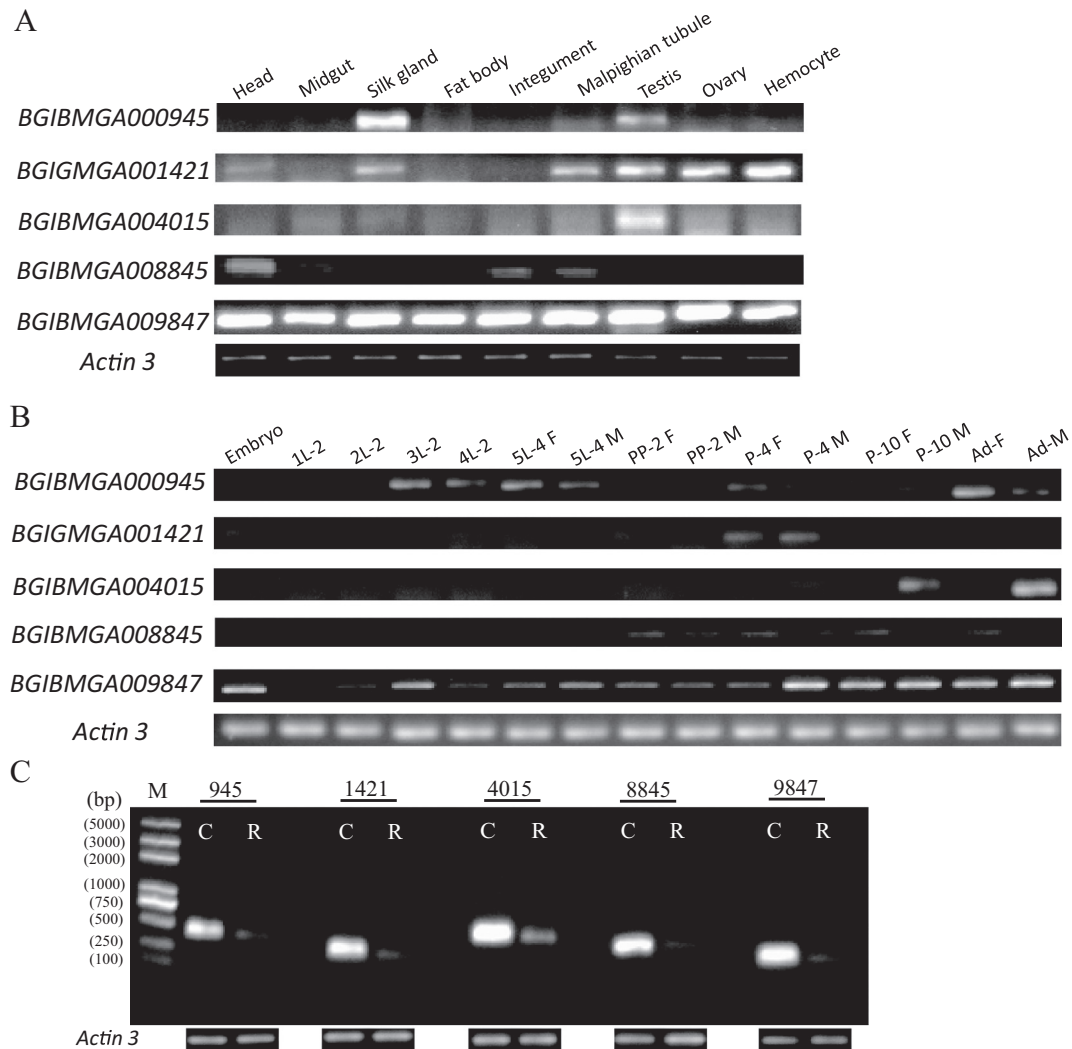
**Fig. 3.** Syntenic analyses of the silkworm *de novo* genes. Table shows the location and neighbor genes of the silkworm *de novo* genes. The bold neighbor genes mean these genes have orthologous gene conserved in other lepidopteran species. The bottom figure represents the location of silkworm *de novo* genes in the syteny region. The black boxes mean the silkworm *de novo* genes; the white boxes mean the neighbor genes with the homologous genes in other lepidopteran insects; the gray boxes mean the neighbor genes without similarity in other lepidopteran insects.



**Fig. 4.** Comparison analyses of the expression pattern between the silkworm OGs and NOGs from EST data (A); distribution of the expression breadth of the silkworm OGs and NOGs (B); distribution of the expression breadth of the duplicated OGs and TE originated OGs (C). Expression breadth is measured as the number of tissues in which the gene is present. Mann–Whitney *U* test was used for the statistical analyses. Statistical significance: \*\**P* < 0.001.

fitness. Firstly, we surveyed the spatial and temporal expression patterns of these genes by RT-PCR (Fig. 5). We found that all five genes have transcriptional signals. Consistent with the *de novo* genes in fruit fly, BGIBMGA004015 was obviously expressed in

the testis, and the temporal analysis also showed that it was only expressed in the male pupa and adult (Fig. 5A and B). For other genes, they were expressed in at least two different tissues and developmental stages. Especially for BGIBMGA009847, the



**Fig. 5.** RT-PCR analyses of the silkworm *de novo* genes. (A) Spatial expression pattern of the silkworm *de novo* genes; (B) temporal expression pattern of the silkworm *de novo* genes; (C) the effect of RNAi was analyzed by RT-PCR. For (B), 1L to 5L represents the 1st to 5th instar larvae; PP represents pre-pupa; P-*n* represents the *n* days after pupa; Ad represents adult. For (C), C: control; R: RNAi; 945: BGIBMGA000945; 1421: BGIBMGA001421; 4015: BGIBMGA004015; 8845: BGIBMGA008845; 9847: BGIBMGA009847.

transcriptional signals could be detected in all tested tissues and developmental stages (Fig. 5A and B). In general, the *de novo* genes are always expressed in limited tissues or developmental stages. For example, the *de novo* genes in *D. melanogaster* showed predominantly testis-biased expression [43]. By contrast, one previous study showed that all three identified *de novo* genes are broadly expressed in different human tissues [50]. Taken together, the wide expression of the silkworm *de novo* genes indicated that these genes may rapidly obtain several active regulatory motifs [51]. However, RNAi of the five genes did not produce any visible phenotype though the transcript levels significantly decreased (Fig. 5C). Contrary to previous study in *D. melanogaster*, *de novo* genes are not essential for the silkworm [52]. This suggests that the silkworm OGs may contribute to genetic redundancy or species-specific adaptation. Of course, the low efficiency of RNAi in lepidopteran insects may also affect the results [53]. In the future, further study is needed to reveal the exact functions of the silkworm *de novo* genes and to help us understand their evolutionary significance.

In conclusion, we systematically identified orphan genes in the silkworm and other four lepidopteran genomes by comparative genomic analyses. Compared with non-orphan genes, the silkworm orphan genes have some special features. About half of the silkworm OGs were derived from duplication and transposable

elements; this proportion is much higher than other insects. We further found that most of the silkworm OGs were preferentially expressed in limited tissues. Finally, by RNA interference experiments, we found that the *de novo* genes are not essential to the silkworm, implying that they may contribute to genetic redundancy or species-specific adaptation. Taken together, our results provide some valuable information to understand the evolutionary significance and the functions of the orphan genes in the domestic silkworm.

#### Acknowledgements

This work was supported by the Hi-Tech Research and Development (863) Program of China (2013AA102507-2), National Natural Science Foundation of China (Nos. 31272363 and 31402014), and Chongqing Postdoctoral Science special Foundation (Xm2014077).

#### Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at <http://dx.doi.org/10.1016/j.febslet.2015.08.008>.

## References

- [1] Ohno, S. (1970) Evolution by Gene Duplication, Springer, New York.
- [2] Long, M., Betran, E., Thornton, K. and Wang, W. (2003) The origin of new genes: glimpses from the young and old. *Nat. Rev. Genet.* 4, 865–875.
- [3] Tautz, D. and Domazet-Lošo, T. (2011) The evolutionary origin of orphan genes. *Nat. Rev. Genet.* 12, 692–702.
- [4] Khalturin, K., Hemmrich, G., Fraune, S., Augustin, R. and Bosch, T.C. (2009) More than just orphans: are taxonomically-restricted genes important in evolution? *Trends Genet.* 25, 404–413.
- [5] Khalturin, K., Anton-Erxleben, F., Sassmann, S., Wittlieb, J., Hemmrich, G. and Bosch, T.C. (2008) A novel gene family controls species-specific morphological traits in *Hydra*. *PLoS Biol.* 6, e278.
- [6] Li, L., Foster, C.M., Gan, Q.L., Nettleton, D., James, M.G., Myers, A.M. and Wurtele, E.S. (2009) Identification of the novel protein QQS as a component of the starch metabolic network in Arabidopsis leaves. *Plant J.* 58, 485–498.
- [7] Cai, J.J., Zhao, R., Jiang, H. and Wang, W. (2008) *De novo* origination of a new protein coding gene in *Saccharomyces cerevisiae*. *Genetics* 179, 487–496.
- [8] Colbourne, J.K. et al. (2011) The ecoresponsive genome of *Daphnia pulex*. *Science* 331, 555–561.
- [9] Donoghue, M.T., Keshavaiah, C., Swamidatta, S.H. and Spillane, C. (2011) Evolutionary origins of Brassicaceae specific genes in *Arabidopsis thaliana*. *BMC Evol. Biol.* 11, 47.
- [10] Toll-Riera, M., Bosch, N., Bellora, N., Castelo, R., Armengol, L., Estivill, X. and Alba, M.M. (2009) Origin of primate orphan genes: a comparative genomics approach. *Mol. Biol. Evol.* 26, 603–612.
- [11] Heliconius Genome Consortium (2012) Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature* 487, 94–98.
- [12] International Silkworm Genome Consortium (2008) The genome of a lepidopteran model insect, the silkworm *Bombyx mori*. *Insect Biochem. Mol. Biol.* 38, 1036–1045.
- [13] You, M. et al. (2013) A heterozygous moth genome provides insights into herbivory and detoxification. *Nat. Genet.* 45, 220–225.
- [14] Zhan, S., Merlin, C., Boore, J.L. and Reppert, S.M. (2011) The monarch butterfly genome yields insights into long-distance migration. *Cell* 147, 1171–1185.
- [15] Duan, J. et al. (2010) SilkDB v2.0: a platform for silkworm (*Bombyx mori*) genome biology. *Nucleic Acids Res.* 38, D453–D456.
- [16] Domazet-Lošo, T. and Tautz, D. (2003) An evolutionary analysis of orphan genes in *Drosophila*. *Genome Res.* 13, 2213–2219.
- [17] Arunkumar, K.P., Tomar, A., Daimon, T., Shimada, T. and Nagaraju, J. (2008) WildSilkbase: an EST database of wild silkworms. *BMC Genomics* 9, 338.
- [18] Papanicolaou, A., Gebauer-Jung, S., Blaxter, M.L., Owen McMillan, W. and Jiggins, C.D. (2008) ButterflyBase: a platform for lepidopteran genomics. *Nucleic Acids Res.* 36, D582–D587.
- [19] Negre, V. et al. (2006) SPODOBASE: an EST database for the lepidopteran crop pest *Spodoptera*. *BMC Bioinformatics* 7, 322.
- [20] Sun, W., Yu, H., Shen, Y., Banno, Y., Xiang, Z. and Zhang, Z. (2012) Phylogeny and evolutionary history of the silkworm. *Sci. China Life. Sci.* 55, 483–496.
- [21] Wissler, L., Gadau, J., Simola, D.F., Helmkampf, M. and Bornberg-Bauer, E. (2013) Mechanisms and dynamics of orphan gene emergence in insect genomes. *Genome Biol. Evol.* 5, 439–455.
- [22] Wu, D.D., Irwin, D.M. and Zhang, Y.P. (2011) *De novo* origin of human protein-coding genes. *PLoS Genet.* 7, e1002379.
- [23] 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.
- [24] Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32, 1792–1797.
- [25] Xia, X. and Xie, Z. (2001) DAMBE: software package for data analysis in molecular biology and evolution. *J. Hered.* 92, 371–373.
- [26] Xia, Q.Y. et al. (2007) Microarray-based gene expression profiles in multiple tissues of the domesticated silkworm, *Bombyx mori*. *Genome Biol.* 8, R162.
- [27] Li, C. and Wong, W.H. (2003) DNA-chip analyzer (dChip) in: *The Analysis of Gene Expression Data: Methods and Software* (Parmigiani, G., Garrett, E.S., Irizarry, R. and Zeger, S.L., Eds.), pp. 120–141, Springer, New York.
- [28] Sun, W., Shen, Y.H., Yang, W.J., Cao, Y.F., Xiang, Z.H. and Zhang, Z. (2012) Expansion of the silkworm GMC oxidoreductase genes is associated with immunity. *Insect Biochem. Mol. Biol.* 42, 935–945.
- [29] Mutanen, M., Wahlberg, N. and Kaila, L. (2010) Comprehensive gene and taxon coverage elucidates radiation patterns in moths and butterflies. *Proc. R. Soc. B: Biol. Sci.* rspb20100392.
- [30] Campbell, M.A. et al. (2007) Identification and characterization of lineage-specific genes within the Poaceae. *Plant Physiol.* 145, 1311–1322.
- [31] Yang, L., Zou, M., Fu, B. and He, S. (2013) Genome-wide identification, characterization, and expression analysis of lineage-specific genes within zebrafish. *BMC Genomics* 14, 65.
- [32] Kiraga, J. et al. (2007) The relationships between the isoelectric point and: length of proteins, taxonomy and ecology of organisms. *BMC Genomics* 8, 163.
- [33] Nandi, S., Mehra, N., Lynn, A.M. and Bhattacharya, A. (2005) Comparison of theoretical proteomes: identification of COGs with conserved and variable pI within the multimodal pI distribution. *BMC Genomics* 6, 116.
- [34] Khaldi, N. and Shields, D.C. (2011) Shift in the isoelectric-point of milk proteins as a consequence of adaptive divergence between the milks of mammalian species. *Biol. Direct.* 6, 40.
- [35] Alende, N., Nielsen, J.E., Shields, D.C. and Khaldi, N. (2011) Evolution of the isoelectric point of mammalian proteins as a consequence of indels and adaptive evolution. *Proteins* 79, 1635–1648.
- [36] Seshadri, R. et al. (2003) Complete genome sequence of the Q-fever pathogen *Coxiella burnetii*. *Proc. Natl. Acad. Sci. U.S.A.* 100, 5455–5460.
- [37] Hittinger, C.T. and Carroll, S.B. (2007) Gene duplication and the adaptive evolution of a classic genetic switch. *Nature* 449, 677–681.
- [38] Batzer, M.A. and Deininger, P.L. (2002) Alu repeats and human genomic diversity. *Nat. Rev. Genet.* 3, 370–379.
- [39] Lander, E.S. et al. (2001) Initial sequencing and analysis of the human genome. *Nature* 409, 860–921.
- [40] 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.
- [41] Cohen, C.J., Lock, W.M. and Mager, D.L. (2009) Endogenous retroviral LTRs as promoters for human genes: a critical assessment. *Gene* 448, 105–114.
- [42] Piriyaopongsa, J., Polavarapu, N., Borodovsky, M. and McDonald, J. (2007) Exonization of the LTR transposable elements in human genome. *BMC Genomics* 8, 291.
- [43] Levine, M.T., Jones, C.D., Kern, A.D., Lindfors, H.A. and Begun, D.J. (2006) Novel genes derived from noncoding DNA in *Drosophila melanogaster* are frequently X-linked and exhibit testis-biased expression. *Proc. Natl. Acad. Sci. U.S.A.* 103, 9935–9939.
- [44] Heinen, T.J., Staubach, F., Haming, D. and Tautz, D. (2009) Emergence of a new gene from an intergenic region. *Curr. Biol.* 19, 1527–1531.
- [45] Samusik, N., Krukovskaya, L., Meln, I., Shilov, E. and Kozlov, A.P. (2013) PBOV1 is a human *de novo* gene with tumor-specific expression that is associated with a positive clinical outcome of cancer. *PLoS One* 8, e56162.
- [46] Misof, B. et al. (2014) Phylogenomics resolves the timing and pattern of insect evolution. *Science* 346, 763–767.
- [47] Mita, K. et al. (2003) The construction of an EST database for *Bombyx mori* and its application. *Proc. Natl. Acad. Sci. U.S.A.* 100, 14121–14126.
- [48] Kaessmann, H. (2010) Origins, evolution, and phenotypic impact of new genes. *Genome Res.* 20, 459.
- [49] Xie, C. et al. (2012) Hominoid-specific *de novo* protein-coding genes originating from long non-coding RNAs. *PLoS Genet.* 8, e1002942.
- [50] Knowles, D.G. and McLysaght, A. (2009) Recent *de novo* origin of human protein-coding genes. *Genome Res.* 19, 1752–1759.
- [51] Bellora, N., Farre, D. and Alba, M.M. (2007) Positional bias of general and tissue-specific regulatory motifs in mouse gene promoters. *BMC Genomics* 8, 459.
- [52] Reinhardt, J.A., Wanjiu, B.M., Brant, A.T., Saelao, P., Begun, D.J. and Jones, C.D. (2013) *De novo* ORFs in *Drosophila* are important to organismal fitness and evolved rapidly from previously non-coding sequences. *PLoS Genet.* 9, e1003860.
- [53] Terenius, O. et al. (2011) RNA interference in Lepidoptera: an overview of successful and unsuccessful studies and implications for experimental design. *J. Insect Physiol.* 57, 231–245.