


ORIGINAL ARTICLE

Genomic data reveal high conservation but divergent evolutionary pattern of Polycomb/Trithorax group genes in arthropods

Feng Jiang^{1,*}, Qing Liu^{2,*}, Xiang Liu³, Xian-Hui Wang³ and Le Kang^{1,3} 

¹Beijing Institutes of Life Science, Chinese Academy of Sciences, Beijing, China; ²Sino-Danish College, University of Chinese Academy of Sciences, Beijing, China and ³State Key Laboratory of Integrated Management of Pest Insects and Rodents, Institute of Zoology, Chinese Academy of Sciences, Beijing, China

Abstract Epigenetic gene control is maintained by chromatin-associated Polycomb group (PcG) and Trithorax group (TrxG) genes, which act antagonistically via the interplay between PcG and TrxG regulation to generate silenced or permissive transcriptional states. In this study, we searched for PcG/TrxG genes in 180 arthropod genomes, covering all the sequenced arthropod genomes at the time of conducting this study, to perform a global investigation of PcG/TrxG genes in a phylogenetic frame. Results of ancestral state reconstruction analysis revealed that the ancestor of arthropod species has an almost complete repertoire of PcG/TrxG genes, and most of these genes were seldom lost above order level. The domain diversity analysis indicated that the PcG/TrxG genes show variable extent of domain structure changes; some of these changes could be associated with lineage-specific events. The likelihood ratio tests for selection pressure detected a number of PcG/TrxG genes which underwent episodic positive selection on the branch leading to the insects with holometabolous development. These results suggest that, despite their high conservation across arthropod species, different members of PcG/TrxG genes showed considerable differences in domain structure and sequence divergence in arthropod evolution. Our cross species comparisons using large-scale genomic data provide insights into divergent evolutionary pattern on highly conserved genes in arthropods.

Key words chromatin; epigenetics; expression regulation; gene evolution; histone modification; insects

Correspondence: Le Kang, State Key Laboratory of Integrated Management of Pest Insects and Rodents, Institute of Zoology, Chinese Academy of Sciences, Beijing 100101, China. Tel: +86 10 64807219; fax: +86 10 64807099; email: lkang@ioz.ac.cn

Xian-Hui Wang, State Key Laboratory of Integrated Management of Pest Insects and Rodents, Institute of Zoology, Chinese Academy of Sciences, Beijing, 100101, China. Tel: +86 10 64807220; fax: +86 10 64807099; email: wangxh@ioz.ac.cn

*These authors contributed equally to this study.

[Correction added on 17 December 2018, after first online publication: the article title has been amended to ‘Genomic data

Introduction

The genomic DNA of metazoan species is organized in chromatin fibers, which are tightly wrapped around an octamer of four highly evolutionarily conserved histone proteins: H2A, H2B, H3 and H4. The post-translational modifications of histones restrict chromatin accessibility to chromatin-associated factors involved in gene transcription (Kharchenko *et al.*, 2011). Therefore, gene transcription activities are controlled by modulating the structure of chromatin by reversible modifications of

reveal high conservation but divergent evolutionary pattern of Polycomb/Trithorax group genes in arthropods.’]

histone proteins. Epigenetics has emerged as an important mechanism by which chromatin-associated genes regulate the post-translational modifications of histones and gene expression (Schwartz *et al.*, 2010). Gene expression is controlled by chromatin-associated Polycomb group (PcG) and Trithorax group (TrxG) proteins, which act antagonistically by the Polycomb/Trithorax group response (PRE/TRE) elements to generate silenced or permissive transcriptional states (Geisler & Paro, 2015). PcG genes generally maintain transcription silencing of their target genes, whereas TrxG genes maintain the active transcriptional states. The dynamic changes of gene expression can be achieved by the antagonistic interplay between PcG and TrxG regulation.

PcG genes constitute an epigenetic silencing system with key regulators in stable and heritable transcriptional repression in metazoan species (Beisel & Paro, 2011; Aranda *et al.*, 2015). These PcG proteins are a conserved family of transcriptional regulatory factors; they are functionally diverse and are part of Polycomb repressive complexes (PRCs). In *Drosophila melanogaster*, PcG proteins form large multimeric complexes of the two distinct families: the PRC1 and PRC2 complexes (Di Croce & Helin, 2013). PRC1 consists of five core PcG proteins: sex combs extra (*Sce*), Polycomb (*Pc*), posterior sex combs (*Psc*), sex comb on midleg (*Scm*), and Polyhomeotic (*ph-p*). PRC2 contains six proteins: enhancer of zeste [*E(z)*], extra sex combs (*esc*), alternatively Esc-like (*escl*), suppressor of zeste 12 [*Su(z)12*], chromatin assembly factor 1 subunit (*Caf1-55*), and Polycomb-like (*Pcl*). PRC1 and PRC2 establish a repressive chromatin state through distinct mechanisms: PRC1 ubiquitylates histone H2A lysine 118 and PRC2 monomethylates, dimethylates, and trimethylates histone H3 at lysine 27 (Lee *et al.*, 2015). PRC1 can recognize the H3K27me3 repressive marks set by PRC2 through the chromodomain of *Pc* gene. In addition to these two PRC complexes, PcG proteins are also present in the three distinct accessory complexes, as follows: the Polycomb repressive deubiquitinase (PR-DUB), dRING-associated factors (dRAF) and Pho repressive complex (PhoRC) complexes (Calvo-Martin *et al.*, 2016). The PR-DUB complex consists of an ubiquitin carboxy-terminal hydrolase, *calypso*, and an additional sex comb gene, *Asx*, which encodes a chromatin protein that regulates the balance of Polycomb and Trithorax function. The dRAF complex, containing *Sce*, *Psc* and the histone H3K36 demethylase *Kdm2*, mediates the monoubiquitylation of histone H2A (Schwartz & Pirrotta, 2013). Two YY1-related DNA-binding proteins, Pleiohomeotic (Pho) and Pleiohomeotic-like (phol), and Sfmtb (Scm-like with four MBT domain proteins) are involved in the PhoRC complex, which is hypothesized

to play a role in recruitment of PcG complexes to Polycomb response elements (Schwartz & Pirrotta, 2013).

PcG genes regulate a large number of target genes, and the same is true for TrxG genes; TrxG genes, through their activities in histone methylation and chromatin remodeling, are required for maintaining the “on” state of PcG target genes (Kingston & Tamkun, 2014). The DNA-binding proteins of the TrxG genes recruit chromatin-remodeling and histone-modifying complexes to regulate transcription (Schuettengruber *et al.*, 2011). Based on their molecular function, TrxG genes can be divided into two distinct families: histone-modifying complexes and adenosine triphosphate (ATP)-dependent chromatin remodeling complexes. The former includes the trithorax acetyltransferase complex 1 (TAC1), absent small or homeotic discs 1 (ASH1), complex proteins associated with Set1 (COMPASS), and COMPASS-like complexes, whereas the latter includes the switch/sucrose nonfermentable (SWI/SNF), imitation switch (ISWI) and chromodomain helicase DNA-binding (CHD) complexes (Schuettengruber *et al.*, 2011; Geisler & Paro, 2015). As for histone-modifying complexes, only a few TrxG genes have been identified in the TAC1 (Trithorax, *nejire* and *Sbf*) and ASH1 (*ash1* and *nejire*) complexes. All the COMPASS complexes contain a TrxG protein named absent, small or homeotic discs 2 (*ash2*), which is required for the stability of COMPASS (Mohan *et al.*, 2011). Three H3K4 methyltransferases, Trithorax (*Trx*), SET domain containing 1 (*Set1*), and Trithorax-related (*Trr*), co-purify with *ash2*, which indicates the existence of one COMPASS complex and two COMPASS-like complexes in *Drosophila*. The COMPASS and COMPASS-like complexes contain four other common TrxG proteins, Retinoblastoma binding protein 5 (*Rbbp5*), will die slowly (*wds*), Dpy-30-like 1 (*Dpy-30L1*) and Host cell factor (Hcf). In addition to these common TrxG genes, CXXC finger protein 1 (*Cfp1*), WD repeat domain 82 (*Wdr82*), Menin 1 (*Mnn1*), Utx histone demethylase (*Utx*), Nuclear receptor coactivator 6 (*Ncoa6*), PAX transcription activation domain interacting protein (Ptip) and Ptip associated 1 (*Pa1*) are specifically involved in each complex. As for ATP-dependent chromatin remodeling complexes, chromatin-remodeling reactions coupling with ATP hydrolysis are catalyzed to regulate chromatin structure and gene expression. Based on the structure of the ATPase subunit, the TrxG genes in ATP-dependent chromatin remodeling complexes can be subdivided into three different families: the SWI/SNF, ISWI and CHD complexes (Schuettengruber *et al.*, 2011). The SWI/SNF complex is a large protein complex consisting of at least 12 TrxG proteins, including brahma (*brm*, a bromodomain-containing

protein), *osa* (an ARID DNA-binding domain protein), *moira* (*mor*), *brahma* associated protein 60 kD (*Bap60*), *polybromo*, *Bap55*, *Bap111*, *Bap170*, *enhancer of yellow 3* [*e(y)3*], *Snf5-related 1* (*Snr1*), *Actin 5C* (*Act5C*) and *Actin 42A* (*Act42A*). Four TrxG proteins, *Enhancer of bithorax* [*E(bx)*], *Iswi*, *Nucleosome remodeling factor 38kD* (*Nurf-38*) and *Chromatin assembly factor 1, p55 subunit* (*Caf1-55*) in the ISWI complex form the nucleosome-remodeling factor (NURF) complex (Ho & Rabtree, 2010). The CHD complexes of chromatin modifiers are defined by the presence of chromodomains, which couple chromatin remodeling and histone deacetylation to function as repressors of transcription. The TrxG proteins in the CHD complexes are *Mi-2*, *Histone deacetylase 1* (*HDAC1*), *Metastasis associated 1-like* (*MTA1-like*), *Methyl-CpG binding domain protein-like* (*MBD-like*), *simjang* (*simj*), *Caf1-55*, *Chromodomain-helicase-DNA-binding protein 1* (*Chd1*) and *kismet* (*kis*).

Arthropods, as the most successful animal group in terms of their diverse habitats, demonstrate countless evolutionary adaptations that help them adapt to their environment or a particular lifestyle. Gene expression regulation by chromatin-associated PcG/TrxG genes has been recognized as an important component in these evolutionary adaptations (Simola et al., 2013; Yan et al., 2014; Simola et al., 2016). Although the fundamental aspect of this regulation mechanism is relevant across different phyla, the evolutionary differences of PcG/TrxG genes between insects and mammals have been identified using a few representative species from plant and bilateral animal kingdoms (Whitcomb et al., 2007). However, most previous studies for arthropod PcG/TrxG genes focused on a few insect species, including *D. melanogaster* and *Bombyx mori*, and a comprehensive study providing a global view of PcG/TrxG genes in arthropods has not yet been reported (Schuettengruber et al., 2011; Li et al., 2012; Calvo-Martin et al., 2016; Calvo-Martin et al., 2017). Arthropod genome sequencing initiatives have greatly accelerated the accumulation of genomic resource data of arthropods, which are obtained by using samples from different lineages to explore arthropod genome diversity. The availability of a large amount of arthropod genomes offers an opportunity to performing a global investigation of arthropod PcG/TrxG genes in a phylogenetic frame. In this study, we searched for PcG/TrxG genes in 174 panarthropod genomes covering all the sequenced arthropod genomes at the time of conducting this study. We performed the ancestral state reconstruction analysis to trace the evolutionary history of PcG/TrxG genes across arthropod phylogeny. Finally, we performed the protein domain diversity and selection testing analyses to explore

the gene structure diversity and sequence divergence of PcG/TrxG genes in arthropods.

Materials and methods

Completeness assessment of official protein-coding gene sets

We screened the genome assemblies of 182 panarthropod species, including 159 insect species, 11 chelicerates, seven crustaceans, two non-insect hexapods, two tardigrades (non-arthropod outgroup) and one myriapod (Table S1). These 182 analyzed panarthropod species represent 30 orders, 79 families and 118 genera. To assess the completeness of official protein-coding gene sets, we used a set of 2675 near-universal single copy orthologs of arthropod genomes in Benchmarking Universal Single-Copy Orthologs (BUSCO) v1.22 (Simao et al., 2015). These 2675 single copy genes are identified in nearly 95% of arthropod genomes and are considered as a benchmark for gene annotation completeness. The target gene sets could be classified as complete and single copy, complete and duplicated, fragmented and missing genes using a combined approach of BLAST and HMMER (version 3.1b) searches (Altschul et al., 1997; Finn et al., 2011).

Gene identification of Polycomb/Trithorax group genes

The 17 PcG and 40 TrxG genes identified in *D. melanogaster* that confer transcriptional repression and activation activity (*Sce*, *Psc*, *Pc*, *ph-p*, *Scm*, *E(z)*, *esc*, *escl*, *Su(z)12*, *Caf1-55*, *Pcl*, *pho*, *phol*, *Sfmbt*, *Kdm2*, *Asx*, and *calypso*; and *Set1*, *Cfp1*, *Wdr82*, *ash2*, *Dpy-30L1*, *Hcf*, *Rbbp5*, *wds*, *trx*, *Mnn1*, *trr*, *Utx*, *Ncoa6*, *Pa1*, *Ptip*, *nejire*, *Sbf*, *ash1*, *brm*, *Act5C*, *Act42A*, *Bap60*, *polybromo*, *Bap55*, *Bap111*, *Bap170*, *osa*, *mor*, *e(y)3*, *Snr1*, *E(bx)*, *Iswi*, *Nurf-38*, *Caf1-55*, *Chd*, *Mi-2*, *HDAC1*, *MBD-like*, *MTA1-like*, *simj*, *Rm62*, and *kis*) might also function in other arthropod taxa. We searched for PcG and TrxG orthologous genes in a wide variety of arthropod genomes in which genome sequences and protein-coding gene annotation were available. Putative orthologous genes were identified using the amino acid sequence of each *D. melanogaster* PcG and TrxG protein as a query. For cases in which multiple isoforms of the orthologous gene were present, only the isoform with the longest protein sequences was used. The reciprocal BLAST hit method was used to determine the significant hits which were further manually verified (Altschul et al., 1997). One PcG/TrxG gene from the *D. melanogaster* genome and another gene from an arthropod genome were considered as orthologous genes if these two genes returned

the highest scoring match in reciprocal BLAST searches (excluding paralogous genes). To prevent overestimating the gene number for each orthologous gene, a length threshold criterion (more than 60% of the corresponding gene length in *D. melanogaster*) was utilized for short gene fragment filtering in case of duplicates. To exclude the potential PcG/TrxG genes from the endosymbiont genomes or contamination of microbial genomes, all the putative orthologous genes were BLAST searched against the non-redundant National Center for Biotechnology Information (NCBI) protein database to verify that the top hit genes were not from endosymbionts or microorganisms or plants. The hidden Markov model-based HMMER program version 3.1b was used to determine domain architecture in the Pfam protein family database with a conditional *E*-value cutoff of $1E-5$ (Finn *et al.*, 2010; Finn *et al.*, 2011).

Ancestral state reconstruction and phylogenetic analyses

A character matrix that represents the existence states for each PcG/TrxG gene was constructed to reconstruct the ancestral states of interior clades in arthropod evolution. Only the binary state was considered. The copy number variation for each PcG/TrxG gene was not considered. Ancestral reconstruction were performed in Mesquite version 3.2 (<http://mesquiteproject.org/>) using the Markov k-state 1-parameter model, which gives equal probability for changes between any two character states. The emergence events of each PcG/TrxG gene along each branch in the phylogenetic tree were inferred based on the criterion: the PcG/TrxG gene was absent at the ancestral nodes of a given node and either of the outgroups. A phylogenetic tree for the species involved is required for the ancestral reconstruction process. As the capability of parallelizing computation is required for the phylogenetic inference of large genomic data to obtain accurate results within reasonable computing time, the IQ-TREE program version 1.5 was used to construct the species tree using the single-copy complete BUSCO genes based on the maximum likelihood principle (Nguyen *et al.*, 2015). The best protein substitution model was selected by the build-in model-selection function of the IQ-TREE program, and bootstrap support values from 1000 replicates were assessed with the ultrafast bootstrap approximation.

Tests for selection using likelihood ratio tests

The protein sequences of PcG/TrxG genes were aligned with the MAFFT alignment program version 7.215 and subsequently back-translated into the corresponding nucleotide sequences (Katoch *et al.*, 2009). The PcG/TrxG

genes showing signals of gene conversion were filtered from the selection analysis using the GENECONV program version 1.81a. To assess the contribution of natural selection during the diversification of PcG/TrxG genes in arthropod species, the ratios of nonsynonymous substitution per nonsynonymous site to synonymous substitution per synonymous site (ω) across the phylogenetic tree of the species examined were determined using the program codeml of the PAML package version 4.48a (Yang, 2007). To test for positive selection, the null model (one-ratio model) assumes that the ω ratios are invariable among all branches examined, whereas the alternative model allows the ω ratio to vary along specific branches. Likelihood ratio tests with one degree of freedom were used to compare the null and alternative models (Yang, 1998). Compared with the score of the null model, a significantly higher likelihood score of the alternative model implied a better fit to the tested data, indicating a variation of selective pressures in specific branches. Bonferroni corrected significance threshold was set to be 0.005, as followed by the previous study (Pauli *et al.*, 2016).

Results

Assessment of protein-coding gene annotation in arthropod genomes

We searched for the orthologous genes of 17 PcG genes and 40 TrxG genes from *D. melanogaster* in genome datum sets from 180 arthropod species, including 159 insect species, 11 chelicerates, seven crustaceans, two non-insect hexapods and a myriapod. We also included two tardigrades as outgroup species. These 182 species represent 30 orders, 79 families and 118 genera (Table S1). Due to different genome assembly completeness, protein-coding gene annotation quality resulted in gene repertoire variations. Assessing protein-coding gene annotation based on total predicted gene number in official gene sets is not feasible. For example, a relatively small gene set might have been explained by low quality of protein-coding gene annotation rather than lower gene numbers. Therefore, the gene identification of target genes in a given genome was influenced by gene annotation quality in official gene sets. We assessed the completeness of official gene sets using a set of 2675 near-universal single copy orthologs of arthropod genomes in BUSCO. The BUSCO assessment results are summarized in Figure 1, and complete assessment values for all studied arthropod genomes are provided in Table S2. The gene annotation completeness as calculated by the percentage of detected BUSCO genes ranges from 100% (*D. melanogaster*, Insecta, Diptera) to 27%

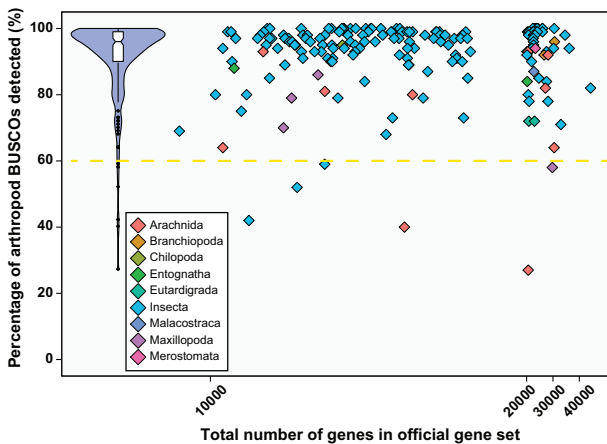


Fig. 1 Assessments of annotation completeness with single-copy orthologs. The 2675 near-universal single-copy orthologs of arthropods from BUSCO database were used in the annotation completeness assessments for the official gene sets of 182 genomes. The violin plot in the left panel shows the distribution of the percentage of BUSCO annotation completeness.

(*Loxosceles reclusa*, Arachnida, Araneae). The high-quality gene annotation completeness for a large portion (median, 96%; 25th–75th percentile, 90%–99%) of the arthropod genomes studied suggested that the protein-coding gene annotation quality for most genome data is sufficient for further analysis. However, due to their low coverage of BUSCO gene sets, we excluded the genome data for six arthropods, including *Pachypsylla venusta* (59%, Insecta), *Eurytemora affinis* (58%, Maxillopoda), *Limnephilus lunatus* (52%, Insecta), *Megaselia scalaris* (42%, Insecta), *Latrodectus hesperus* (40%, Chelicerata) and *Loxosceles reclusa* (27%, Chelicerata), for further analysis due to low coverage of BUSCO gene sets. Therefore, the gene sets of 174 arthropod genomes and two non-arthropod genomes were used in further analysis.

Gene identification of Polycomb/Trithorax group genes

The gene identification of 17 genes encoding the subunits of five PRCs and 40 genes encoding subunits of TrxG complexes has been analyzed. The studied 17 PcG genes included *Sce*, *Psc*, *Pc*, *ph-p* and *Scm* in PRC1 complex; *E(z)*, *esc*, *escl*, *Su(z)12*, *Caf1-55*, and *Pcl* in PRC2 complex; *pho*, *phol*, and *Sfmbt* in PhoRC; *Kdm2* in dRAF complex (*Sce* and *Psc* are also involved in dRAF complex formation); and *Asx* and *calypso* in PR-DUB complex. The studied 40 TrxG genes included the following: *Set1*, *Cfp1*, *Wdr82*, *ash2*, *Dpy-30L1*, *Hcf*, *Rbbp5*, and *wds* in COMPASS complex; *trx*, *Mnn1*, *trr*, *Utx*, *Ncoa6*, *Pa1*, and *Ptip* in COMPASS-like complex; *nej*, *Sbf*, *ash1* in

TAC1 (also including *trx*) and ASH1 complexes; *brm*, *Bap60*, *polybromo*, *Bap55*, *Bap111*, *Bap170*, *osa*, *mor*, *e(y)3*, and *Snr1* in SWI/SNF complex; *E(bx)*, *Iswi*, *Nurf-38* and *Caf1-55* in ISWI complex; and *Chd*, *Mi-2*, *HDAC1*, *MBD-like*, *MTA1-like*, *simj*, *Rm62* and *kis* in CHD complex. Five genes, *ash2*, *Dpy-30L1*, *Hcf*, *Rbbp5* and *wds*, were also involved in COMPASS-like complex formation. We searched for PcG and TrxG orthologous genes using the protein sequence of *D. melanogaster* PcG and TrxG protein as a BLAST query. Searching the official gene sets of 176 species allowed us to identify 12 466 putative PcG/TrxG genes. Genome assembly or annotation artifacts may result in two gene fragments, which are split from one complete canonical PcG/TrxG gene. In case of fragmentation for each orthologous gene, a total of 2569 short gene fragments (less than 60% of the corresponding gene length in *D. melanogaster*) were filtered to exclude the possibility of gene copy number overestimating. As DNA contamination from other species, including microorganisms, plants and endosymbionts, is an important challenge of arthropod genome sequencing project, we filtered the putative PcG/TrxG genes which are not from arthropod genomes (Alkan *et al.*, 2011). The BLAST searches against the non-redundant NCBI protein database revealed that 37 TrxG/PcG genes were from non-arthropod species. All of these non-arthropod genes are bacterial in origin, and a majority of them (57%, 21 in 37) are from Proteobacteria (Fig. S1). Gene annotation in the *Blattella germanica* genome contains the most abundant non-arthropod TrxG/PcG genes, including three copies of *Rm62* (BGER006221-PA, BGER026712-PA, and BGER004656-PA), three copies of *Nurf-38* (BGER009429-PA, BGER011989-PA, and BGER024170-PA), and two copies of *Iswi* (BGER000971-PA, and BGER019004-PA). The non-arthropod PcG/TrxG genes in the 176 gene sets are listed in Table S3. A previous study showed that BGIBMGA006325 is *Esc* in *B. mori* and that *Escl* is not identified in *B. mori* and three other insects, *Aedes aegypti*, *Tribolium castaneum* and *Apis mellifera* (Li *et al.*, 2012). In fact, our reciprocal BLAST hit results indicated that, consistent with the ortholog relationships in FlyBase, BGIBMGA006325 is the ortholog of *Escl* gene in *D. melanogaster*.

Phylogenetic distribution of PcG/TrxG genes across arthropod phylogeny

To visualize the phylogenetic distribution of PcG/TrxG genes, we mapped the respective PcG/TrxG genes (Table S4) to the arthropod phylogeny, which was constructed by the 593 single-copy complete BUSCO

genes using the parallelizing program IQ-TREE under maximum likelihood principle. The arthropod phylogenetic tree is consistent with the phylogenomic tree inferred from transcriptome data (Misof *et al.*, 2014). The PcG genes could be divided into two different categories based on their phylogenetic distribution (Fig. S2): (1) a large portion of the 17 PcG genes, including *Kdm2*, *Caf1-55*, *escl*, *Sce*, *Su(z)12*, *E(z)*, *Psc*, *Sfmbt*, *Asx*, *calypso*, *Scm*, *Pcl*, *Pc*, and *Pho*, is particularly widespread across arthropod species; and (2) the three PcG genes, *ph-p*, *phol*, and *esc*, show patchy phylogenetic distribution and were only found in a limited number of arthropod species. Similar to the broad distribution of PcG genes in arthropods, most TrxG genes are found in all major arthropod lineages. Multiple gene copies, which were possibly derived by gene duplication from an ancestral gene, could be detected in both PcG and TrxG genes. The duplication events generally occurred in the species, which were closely related on the phylogenetic tree. This finding suggests that lineage-specific duplication events might be correlated with a diversification and functional specialization of PcG and TrxG genes in specific lineages of arthropods. The lysine-specific demethylase 2, *Kdm2*, was subject to the most frequent duplication events among the 17 PcG genes. As shown in Figure S2, the multiple gene duplicates of *Kdm2* (number of gene copies, mean: 2.16, standard deviation: 1.07) were observed in the major insect orders, including Lepidoptera, Coleoptera, Thysanoptera, Hymenoptera, Thysanoptera, Blattodea, Ephemeroptera, Odonata, and Orthoptera. The duplication events of *Kdm2* were also observed in non-insect arthropod orders, including Diplura, Branchiopoda, Malacostraca, Maxillopoda, Merostomata, Arachnida, and the outgroup Eutardigrada. Besides *Kdm2*, *Psc* (number of gene copies, mean: 1.50, standard deviation: 0.72) and *Sfmbt* (number of gene copies, mean: 1.38, standard deviation: 0.60) also had duplicated copies. As shown in Figures S3–S5, the most prominent duplicated gene in the TrxG genes is *wds* (number of gene copies, mean: 4.13, standard deviation: 1.94), an essential gene coding for a WD-repeat protein. Although the *wds* duplication events could be detected in all the arthropod orders studied, the duplication frequencies might vary among different arthropod orders. Most of the *wds* genes outside Diptera have several copies ranging from two to nine, whereas only one copy of *wds* was detected in the 75% of Dipteran species, suggesting that the family size of TrxG gene can not only grow but also shrink. In addition, *Bap55*, *brm*, and *Rm62* also have multiple copies. However, their duplication events could only be observed in a few species, implying a limited phylogenetic distribution of these duplication events across arthropod evolution.

A character matrix that shows the present/absent states for each PcG/TrxG gene was used in the ancestral state inference of interior nodes along with the arthropod phylogeny. The ancestral states at different nodes could infer the emergences/losses of the PcG/TrxG genes that occurred at and above the level of arthropod orders (Fig. 2). The putative ancestral state was composed of 50 PcG/TrxG genes present in the last common ancestor of the species involved. The numbers of PcG/TrxG genes range from 44 to 56 in all the arthropod species, suggesting that the number of PcG/TrxG genes in a given species are relatively invariable ($P > 0.05$, Mann–Whitney *U*-tests) during arthropod evolution. Consistent with the relative invariability of gene number, the ancestral state inference results showed that, along with arthropod evolution, no emergence of either PcG or TrxG genes occurred above the order level. Compared with those of the outgroup tardigrades, the gene emergence events for few genes (*Su(z)12*, *Pcl*, *Nco66*, *Pa1*, *Chd1*, and *Rm62*) were only observed at the basal branches of the arthropod phylogenetic tree, which indicated that the ancestor of arthropod species has an almost complete repertoire of PcG/TrxG genes. The PcG/TrxG genes are seldom lost above orders during arthropod evolution, possibly indicating a strong selection pressure to maintain PcG/TrxG genes in arthropod species. The exception includes *esc*, *Dpy-30L1*, *Pa1*, *Rm62*, *phol*, *Asx*, *ph-p*, *Nco66*, and *trx*. However, we did detect the losses of *esc* and *Rm64* in more than three specific lineages, taking account of the ancestral states of these genes in the early arthropod phylogeny. The previous study in Diptera showed that the *esc/escl* duplication took place ~130 million years ago, after the split of the Psychodidae family from other dipteran species (Calvo-Martin *et al.*, 2017). But, the presence of *esc/escl* in Tardigrada and in a wide range of arthropod species suggested that *esc* and *escl* are ancient duplicates, which were already present in the last common ancestor of arthropod species.

Domain diversity analysis of PcG/TrxG genes

The putative protein domains for all the PcG/TrxG genes were identified in the Pfam signatures using a cut-off *E*-value of $1E-5$. The *D. melanogaster* PcG/TrxG genes as domain structure examples are shown in Figure 3. Different PcG/TrxG genes showed distinct domain structure, which indicated that the domain structure pattern reflected the gene function. In general, the protein domains in PcG (average: 3.11 domains per gene) genes are less abundant than those in TrxG (average: 4.34 domains per gene) genes, but the differences were not significant ($P = 0.144$, Mann–Whitney *U*-tests). The most frequent

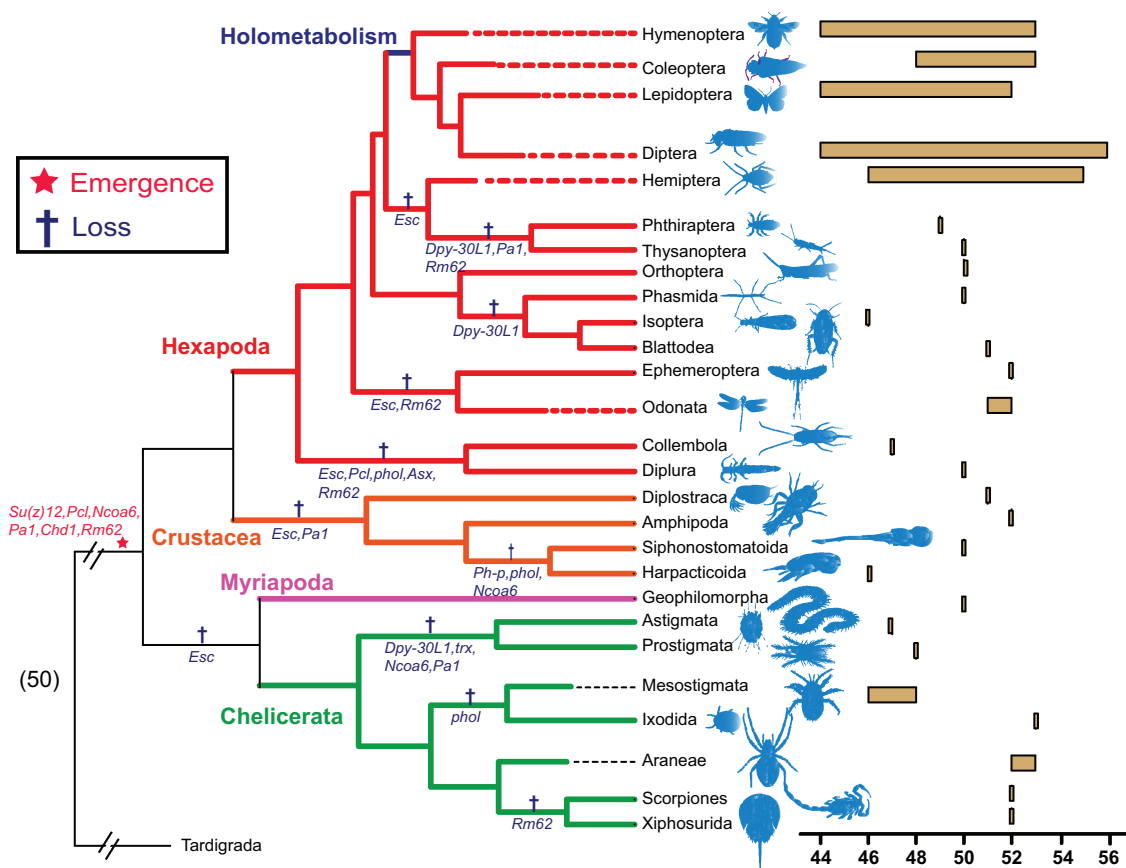


Fig. 2 Inference of ancestral state of PcG and TrxG genes along arthropod evolution. The emergence and loss of PcG and TrxG genes are indicated in the internal nodes of the phylogenetic tree. The numbers in each order are as follows: Amphipoda, 1; Araneae, 2; Astigmata, 1; Blattodea, 1; Coleoptera, 9; Collembola, 1; Diplostraca, 1; Diptera, 16; Ephemeroptera, 1; Geophilomorpha, 1; Harpacticoida, 1; Hemiptera, 13; Hymenoptera, 38; Isoptera, 1; Ixodida, 1; Lepidoptera, 9; Mesostigmata, 2; Odonata, 2; Orthoptera, 1; Parachela, 2; Phasmida, 1; Phthiraptera, 1; Prostigmata, 1; Scorpiones, 1; Siphonostomatoida, 1; Thysanoptera, 1; Xiphosurida, 1.

domains in PcG genes are the WD40, MBT (PF02820), Zinc finger C2H2 type (zf-C2H2, PF00096), and Zinc-finger double (zf-H2C2.2, PF13465), and sterile alpha motif (SAM, PF07647) domains. Meanwhile, the most frequent domains in TrxG genes are the WD40, PHD, Bromodomain, Chromo and ANAPC4.WD40 domains. The Pfam signature results indicated that five PcG/TrxG genes had enzymatic activity. Four TrxG genes, *Set1*, *ash1*, *trr* and *trx*, and one PcG gene, *E(z)*, are histone methyltransferases. Their enzyme activities are mediated by Su(var)3-9, Enhancer-of-zeste, Trithorax (SET, PF00856) domain. Protein domain gain and loss are important for gene function innovation which results in the creation of molecular biodiversity and functional molecular changes (Bornberg-Bauer & Alba, 2013). To better understand the domain diversity of PcG/TrxG genes, we generated domain structure maps for the coding region of each

PcG/TrxG gene for all the studied arthropod genomes. The detailed results of the domain structure analysis are present in Table S5. The conserved domains of each PcG/TrxG gene are combined with other variable domains in different modular arrangements. For example, *ash1* in *D. melanogaster* is composed of three functional domains, including SET domain (SET, PF00856), bromo-adjacent homology domain (BAH, PF01426), and PHD-finger domain (PHD, PF00628), as shown in Figure 4A. However, *ash1* in most arthropod species contains an additional conserved domain, Bromodomain (PF00439), except in the Dipteran and Lepidopteran species. A fragmentary Bromodomain domain could be identified in a small portion of the Dipteran and Lepidopteran species, thereby indicating that the Bromodomain domain is exclusively decayed and lost in Diptera and Lepidoptera. In a global view, the PcG/TrxG genes show variable extent

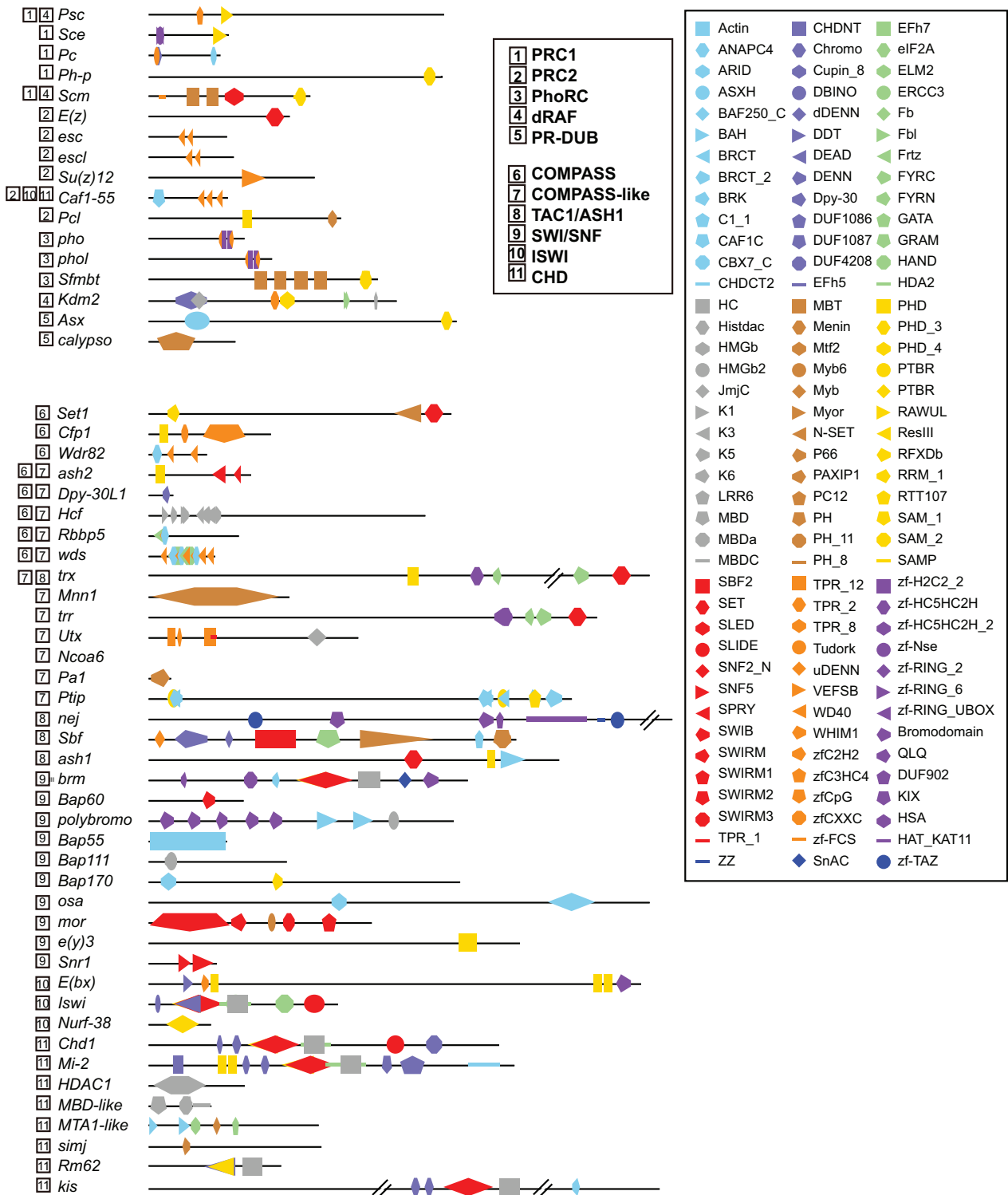


Fig. 3 Schematic diagram of protein domain structure of PcG/TrxG genes in *Drosophila melanogaster*. The hidden Markov model-based HMMER program was used to determine domain architecture in Pfam protein family database with a conditional *E*-value cutoff of $1E-5$. Distinct protein domains were labeled in different colors/shapes. The length of the black long line is directly proportional to the length of the corresponding gene.

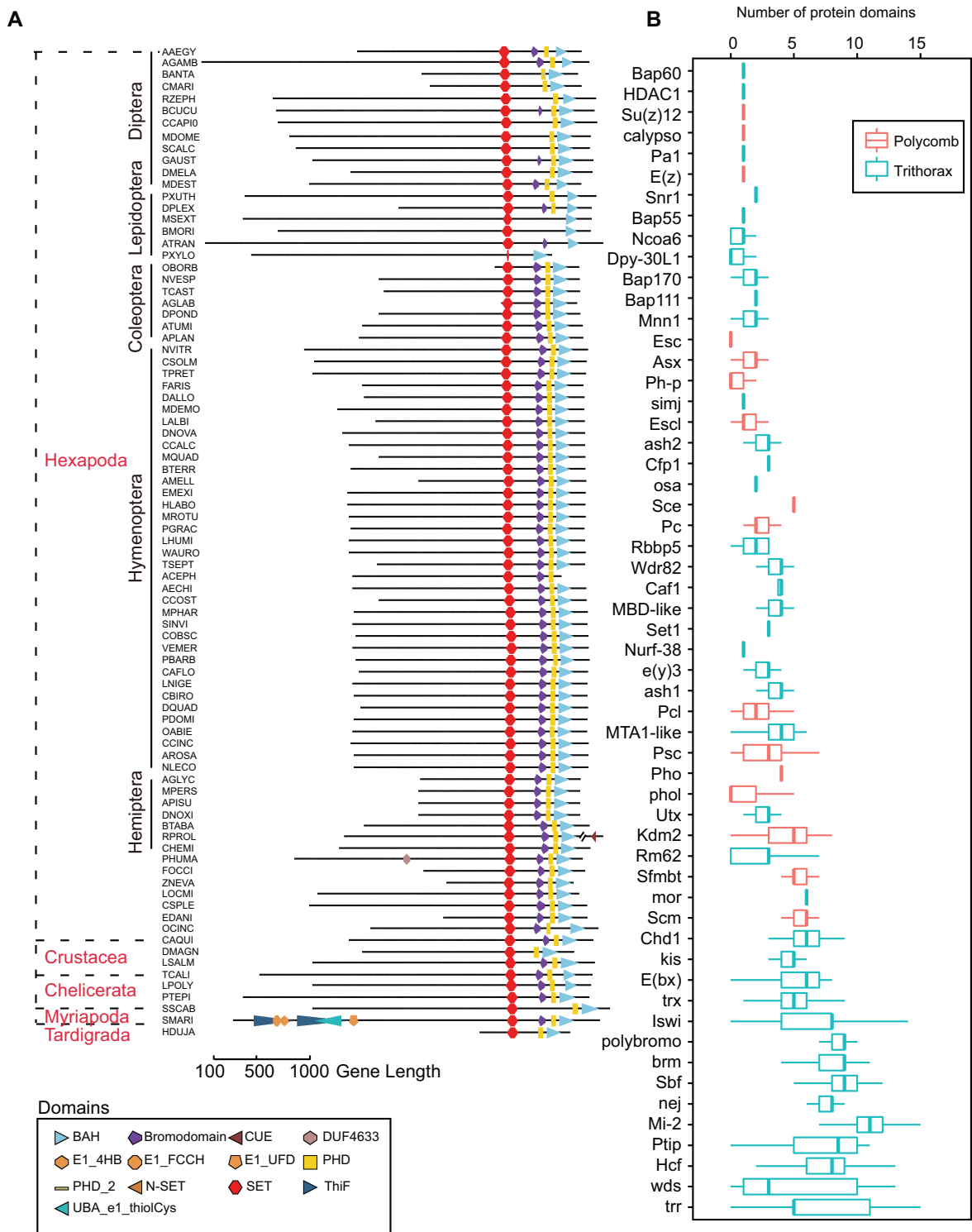


Fig. 4 Domain structure analysis of PcG/TrxG genes in arthropods. (A) Cross species comparison of protein domain structure of *ash1* in arthropods. The hidden Markov model-based HMMER program was used to determine domain architecture in Pfam protein family database with a conditional *E*-value cutoff of $1E-5$. Distinct protein domains were labeled in different colors/shapes. The length of the black long line is directly proportional to the length of the corresponding gene. (B) Boxplot showing variations of domain number of PcG/TrxG genes in arthropods.

of domain structure changes (Fig. 4B). The top 14 most variable genes in domain number are *trr*, *wds*, *Hcf*, *Ptip*, *Mi-2*, *nej*, *Sbf*, *brm*, *polybramo*, *Iswi*, *trx*, *E(bx)*, *kis*, and *Chd1*. All of these genes are TrxG genes. Compared with other PcG/TrxG genes, five TrxG genes (*Bap60*, *HDAC1*, *Pal*, *Snr1*, and *Bap55*) and three PcG genes [*Su(z)12*, *calypso* and *E(z)*] have fewer fluctuations in domain structure composition.

Tests of selection acting on PcG/TrxG gene

To assess the contribution of natural selection during the diversification of PcG/TrxG genes in arthropod species, we measured the pairwise ω values. All the ω values in PcG/TrxG genes are less than 1, indicating a purifying (or negative) selection on these genes (Fig. 5). Although the ω values for each PcG/TrxG gene are not greater than 1, the distribution of the ω values showed the heterogeneity pattern of evolutionary rates acting on arthropod PcG/TrxG genes. *Psc* in PcG genes and *Ptip* in TrxG showed a relatively high ω value in their corresponding gene group. However, high sequence divergence deduced by high ω values is not equated with function divergence of *Psc*, due to their shared physical properties (Beh *et al.*, 2012). Holometabolism, which is the most distinctive characteristic of insects, is a highly successful biological adaptation (Truman & Riddiford, 1999). We performed the branch-model selection analysis to determine whether positive selection acting on PcG/TrxG genes plays a role in the novelty of holometabolism. Specifically, we tested the hypothesis that the selection pressure acting on the specific branch leading to holometabolous insects was significantly different from the average over the other branches (Fig. 2). Under the homogeneous one ratio model (M0) assuming the invariable ω values among sites and branches, ω ranges from 0.007 (*Wdr82*) to 0.169 [*e(y)3*] among different PcG/TrxG genes. The likelihood ratio tests provide evidence for four genes, *brm*, *kis*, *nej*, and *trr*, showed signals of positive selection on the specific branch leading to holometabolous insects (Table 1).

Discussion

In this study, all the sequenced arthropod genomes that are publicly accessible at the time of conducting this study were assessed for the annotation completeness of official gene sets. A total of 174 arthropod genomes were preserved for further analysis due to their high-quality annotated gene sets. The PcG/TrxG genes were identified by BLAST searching under the best reciprocal hit criterion. The results of ancestral state reconstruction analysis

indicated that the ancestor of arthropod species had an almost complete repertoire of PcG/TrxG genes and that most of these genes are seldom lost above order level in arthropod evolution. Finally, the domain diversity and selection test analyses revealed considerable differences of domain structure and sequence divergence among PcG/TrxG genes. These data suggested that, in spite of their high conservation, the different members of PcG/TrxG genes have undergone divergent evolutionary patterns in arthropod evolution.

We did not search for the PcG/TrxG gene fragments in the genome sequences by TBLASTN searches; a complementary search method is usually used in gene identification in limited species or for limited genes. The annotation authenticity based on the TBLASTN searches was considered to be impaired by the presence of processed pseudogenes: nonfunctional, fragmentary, and intronless copies of authentic genes found elsewhere in eukaryotic genomes (van Baren & Brent, 2006). The gene identification strategy using the TBLASTN searches frequently mistake processed pseudogenes for authentic genes, which leads to biologically irrelevant gene identification. In most cases, the official gene sets are predicted by the standard genome-wide gene annotation programs. Although different genome-wide gene annotation programs differ in their process details, they share a core set of features (Yandell & Ence, 2012). In general, these annotation programs identify genomic repeat sequences, align RNA transcripts and homology proteins to a genome using splice-site-aware alignment algorithms, generate *ab initio* and/or evidence-driven gene predictions, and automatically combine these computed data into final gene sets. In addition, the information about expression evidence, exon boundaries, and splice sites is fully taken into account in gene structure determination (Cantarel *et al.*, 2008). Therefore, the official gene set, which is usually predicted by standard genome-wide gene annotation programs, is clearly a more reliable gene repertoire than those predicted by the TBLASTN searches. In this study, we only identified the PcG/TrxG genes in the official gene sets and abandoned the gene identification methods based on TBLASTN searches. Alternatively, we filtered the potential low-quality gene sets, according to the gene annotation assessment results using a set of 2675 near-universal single copy orthologs of arthropod genomes in BUSCO. This filtering approach by BUSCO assessment is not influenced by species-specific biological concerns. For example, the body louse (*Pediculus humanus*) has the smallest known insect genome and retains a limited gene repertoire of 10 773 protein-coding genes (Kirkness *et al.*, 2010), which is remarkably less than the gene number in most arthropod species studied, due to its obligate

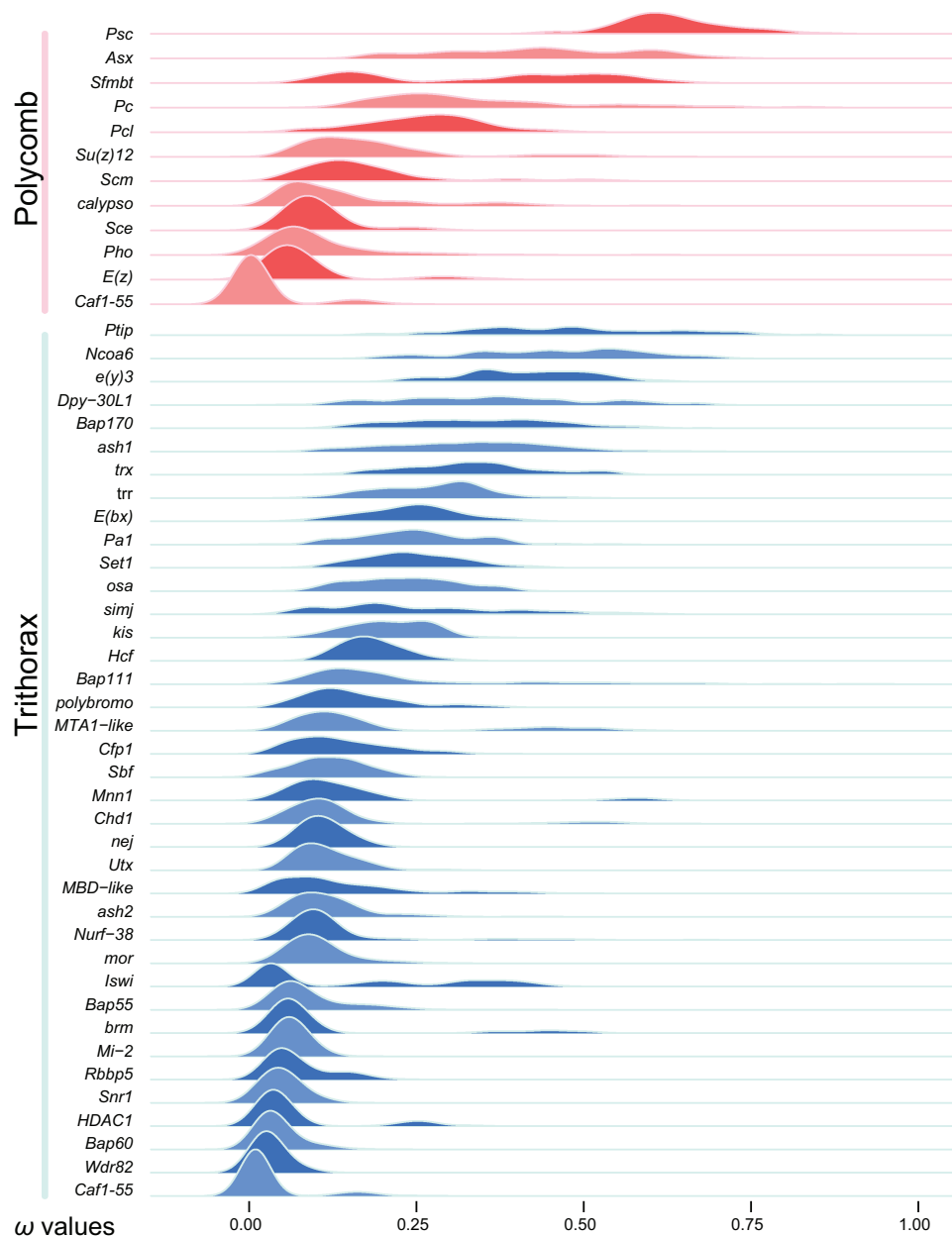


Fig. 5 Distribution of ω values in pairwise comparisons of PcG and TrxG genes in arthropods. The ratios of nonsynonymous substitution per nonsynonymous site to synonymous substitution per synonymous site (ω) across the phylogenetic tree of the species examined were determined using the program KaKs_Calculator version 2.0 under MA method. The ω ratios summarize the evolutionary rates of genes, and thus their distribution reflects the variable extent of strength of natural selection in a global view. In general, the distribution comparison of ω ratios should be cautious with the phylogenetic distances among involved species. Because the PcG and TrxG genes are highly conserved in most of arthropod species studied, the ω ratios were determined in a pairwise manner which is not required to consider the phylogenetic relationships of the species being studied.

parasitic lifestyle. However, more than 99% of BUSCO genes could be identified in the *P. humanus* genome, indicating a complete representation of gene repertoire in the official gene set of the *P. humanus* genome. Further-

more, this filtering approach is practicable in gene identification for large-scale genomic data due to the avoidance of the laborious and time-consuming TBLASTN searching.

Table 1 Tests of rate heterogeneity acting on PcG/TrxG genes in arthropods. The evidence of heterogeneity in evolutionary rate was tested by using the branch model within PAML. The Bonferroni correction for multiple testing was applied in significance determination. The degree of freedom (df) was 1 for all tests.

Gene	$-2(\ln L1 - \ln L0)$	<i>P</i> -value	<i>P</i> _{adj}
<i>ash1</i>	-5.782	0.016	0.059
<i>ash2</i>	-1.427	0.232	0.423
<i>Asx</i>	-1.058	0.304	0.500
<i>Bap111</i>	-5.129	0.024	0.071
<i>Bap170</i>	-7.572	0.006	0.034
<i>Bap55</i>	-0.023	0.880	0.898
<i>Bap60</i>	-7.146	0.008	0.038
<i>brm</i>	-17.957	<0.001	<0.001*
<i>Caf1</i>	-5.841	0.016	0.059
<i>calypso</i>	-1.914	0.167	0.315
<i>Cfp1</i>	-0.191	0.662	0.750
<i>Chd1</i>	-8.542	0.003	0.029
<i>Dpy-30L1</i>	-0.322	0.571	0.693
<i>E(bx)</i>	-7.591	0.006	0.034
<i>e(y)3</i>	0.000	0.997	0.997
<i>E(z)</i>	-2.263	0.133	0.261
<i>Hcf</i>	-5.300	0.021	0.071
<i>HDAC1</i>	-0.051	0.821	0.855
<i>Iswi</i>	-0.519	0.471	0.608
<i>Kdm2</i>	-0.520	0.471	0.608
<i>kis</i>	-14.601	<0.001	0.002*
<i>MBD-like</i>	-1.339	0.247	0.434
<i>Mi-2</i>	-6.911	0.009	0.040
<i>Mnm1</i>	-3.315	0.069	0.159
<i>mor</i>	-8.215	0.004	0.030
<i>MTA1-like</i>	-11.318	0.001	0.008
<i>Ncoa6</i>	-0.155	0.694	0.769
<i>nej</i>	-19.300	<0.001	<0.001*
<i>Nurf-38</i>	-4.925	0.026	0.071
<i>osa</i>	-4.948	0.026	0.071
<i>Pa1</i>	-0.536	0.464	0.608
<i>Pc</i>	-0.806	0.369	0.538
<i>Pcl</i>	-0.500	0.480	0.608
<i>Pho</i>	-0.223	0.637	0.738
<i>polybromo</i>	-0.854	0.356	0.538
<i>Psc</i>	-0.817	0.366	0.538
<i>Ptip</i>	-4.784	0.029	0.073
<i>Rbbp5</i>	-0.055	0.814	0.855
<i>Rm62</i>	-0.725	0.395	0.560
<i>Sbf</i>	-3.819	0.051	0.123
<i>Sce</i>	-0.924	0.336	0.536
<i>Scm</i>	-3.164	0.075	0.167

(to be continued)

Table 1 Continue.

Gene	$-2(\ln L1 - \ln L0)$	<i>P</i> -value	<i>P</i> _{adj}
<i>Set1</i>	-6.540	0.011	0.045
<i>Sfmbt</i>	-0.090	0.764	0.829
<i>simj</i>	-0.479	0.489	0.608
<i>Snr1</i>	-5.212	0.022	0.071
<i>Su(z)12</i>	-3.054	0.081	0.171
<i>trr</i>	-19.199	<0.001	<0.001*
<i>trx</i>	-2.480	0.115	0.235
<i>Utx</i>	-1.277	0.259	0.440
<i>Wdr82</i>	-0.278	0.598	0.709

$\ln L0$, likelihood value for the null model; $\ln L1$, likelihood value for the alternative model; *p*_{adj}, corrected *P*-value using Bonferroni correction. *Bonferroni corrected significance threshold was set to be 0.005.

The arthropods showing phenotypic plasticity contain the same genome information, but differential spatial-temporal gene regulation provides each arthropod individual its own developmental tactics. Therefore, the developmental tactic choice requires the establishment of a controlling system for gene expression. In arthropod species, PcG/TrxG genes play important roles in promoting the repression and activation of gene expression, which are correlated with phenotypic plasticity (Geisler & Paro, 2015). For example, caste-specific foraging and scouting behaviors are regulated epigenetically by the balance between the two TrxG genes, *HDAC1* and *nej* in carpenter ant *Camponotus floridanus* (Simola *et al.*, 2016). Insulator protein, preventing gene activation, is another important regulator of gene expression (Kim *et al.*, 2015). The PcG/TrxG-dependent and insulator-dependent expression controlling systems are two important regulatory mechanisms which are both conserved between vertebrates and insects (Heger *et al.*, 2012). However, a recent study showed that several members of insulator genes were absent in a majority of insect clades, or even were only detected in *D. melanogaster* (Pauli *et al.*, 2016). The distinct difference was observed in the component conservation of these two systems. The components of PcG/TrxG-dependent system are highly conserved among arthropods, while those of insulator-dependent system show a patchy distribution pattern. This suggests that, in contrast to the sophisticated establishment of PcG/TrxG-dependent expression controlling system pre-dating the arthropod evolution, the insulator-dependent expression controlling systems were gradually established along arthropod evolution. The distinct evolutionary fates between these two important classes of gene expression regulators suggest a complex expression controlling system in arthropods.

Proteins are composed of a combination of discrete functional domains, associated with specific roles that have arisen at different times during evolution (Toll-Riera & Alba, 2013). In contrast to high conservation of PcG/TrxG genes, the comparison of domain structure and selection pressure showed a variability of domain number and selection strength, which indicated diversified evolutionary patterns among different PcG/TrxG genes in arthropod species. In the previous study, the domain diversity analysis in 20 arthropod species of the pancrustacean clade provided strong evidence that domain emergence is foremost associated with environmental adaptation (Moore & Bornberg-Bauer, 2012). *Ash1* is likely to interact on chromatin with target genes (Nakamura et al., 2000). The presence of the Bromodomain domain of *ash1* could be detected in a large portion of arthropod species studied but not in Diptera and Lepidoptera. Bromodomain domain, which is present in many transcription and chromatin regulators, can interact specifically with acetylated lysine residues in histones and non-histone proteins (Yang, 2004). Acetylated lysine residues in histones or non-histone proteins would target Bromodomain-containing proteins and their associated complexes, acting in transcriptional control (Nakamura et al., 2000). These results suggested that additional protein–protein interaction of *ash1* in a wide range of insects might be lost in Diptera and Lepidoptera. The loss of this protein–protein interaction in the species was from a broad phylogenetic distribution of Diptera and Lepidoptera, a lineage-specific adaptation or functional deprivation in these two insect orders. Although the specific explanation is unknown yet, the absence of Bromodomain domain would possibly result in fewer target genes or in lower connectivity of *ash1* node in the regulatory network (Di Roberto & Peisajovich, 2014). Therefore, protein domains that emerged or are lost in particular arthropod lineages might be of special interest in helping understand the mechanism of lineage-specific functional adaptation with the help of experimental evidence. The variable domain structure is in concordance with the variable selection strength among different PcG/TrxG genes. Although the ω values for each PcG/TrxG gene are not greater than 1 (positive selection), the distributions of the ω values in pairwise comparisons are quite different from each other. The likelihood ratio tests for selection pressure showed that four genes, *brm*, *kis*, *nej*, and *trr*, showed signatures of positive selection on the branch leading to the insects with holometabolous development. *kis* and *brm* are from the CHD and SWI/SNF complexes, respectively. *trr* and *nej* are from the COMPASS-like and TAC1/ASH1 complexes, respectively. The episodic adaptive selection usually results in an increase of ω values in a pre-assumed branch

leading to unique specific biological relevance (Messier & Stewart, 1997; Kosiol et al., 2008). Therefore, that all the four positive selected genes are TrxG genes implied a critical role of expression activation in holometabolism emergence. Taken together, our results showed that different members of PcG/TrxG genes exhibited considerable differences in domain structure and sequence divergence, revealing a divergent evolutionary pattern on highly conserved PcG/TrxG genes in arthropods.

Acknowledgments

This research was supported by grants from the National Natural Science Foundation of China (Nos. 31672353 and 31472047) and The State Key Laboratory of Integrated Management of Pest Insects and Rodents (Grant No. ChineseIPM1708). The funders had no role in study design, data collection or analysis, decision to publish, or preparation of the manuscript. The computational resources were provided by the Research Network of Computational Biology and the Supercomputing Center at Beijing Institutes of Life Science, Chinese Academy of Sciences.

Disclosure

The authors declare no conflicts of interest.

References

- Alkan, C., Sajjadian, S. and Eichler, E.E. (2011) Limitations of next-generation genome sequence assembly. *Nature Methods*, 8, 61–65.
- Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25, 3389–3402.
- Aranda, S., Mas, G. and Di Croce, L. (2015) Regulation of gene transcription by Polycomb proteins. *Science Advances*, 1, e1500737.
- Beh, L.Y., Colwell, L.J. and Francis, N.J. (2012) A core subunit of Polycomb repressive complex 1 is broadly conserved in function but not primary sequence. *Proceedings of the National Academy of Sciences USA*, 109, E1063–E1071.
- Beisel, C. and Paro, R. (2011) Silencing chromatin: comparing modes and mechanisms. *Nature Reviews Genetics*, 12, 123–135.
- Bornberg-Bauer, E. and Alba, M.M. (2013) Dynamics and adaptive benefits of modular protein evolution. *Current Opinion in Structural Biology*, 23, 459–466.
- Calvo-Martin, J.M., Librado, P., Aguade, M., Papacit, M. and Segarra, C. (2016) Adaptive selection and coevolution at the

- proteins of the Polycomb repressive complexes in *Drosophila*. *Heredity*, 116, 213–223.
- Calvo-Martin, J.M., Papaceit, M. and Segarra, C. (2017) Evidence of neofunctionalization after the duplication of the highly conserved Polycomb group gene *Caf1-55* in the obscure group of *Drosophila*. *Scientific Reports*, 7, 40536.
- Cantarel, B.L., Korf, I., Robb, S.M., Parra, G., Ross, E., Moore, B. *et al.* (2008) MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Research*, 18, 188–196.
- Di Croce, L. and Helin, K. (2013) Transcriptional regulation by Polycomb group proteins. *Nature Structural & Molecular Biology*, 20, 1147–1155.
- Di Roberto, R.B. and Peisajovich, S.G. (2014) The role of domain shuffling in the evolution of signaling networks. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, 322, 65–72.
- Finn, R.D., Clements, J. and Eddy, S.R. (2011) HMMER web server: interactive sequence similarity searching. *Nucleic Acids Research*, 39, W29–W37.
- Finn, R.D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J.E. *et al.* (2010) The Pfam protein families database. *Nucleic Acids Research*, 38, D211–D222.
- Geisler, S.J. and Paro, R. (2015) Trithorax and Polycomb group-dependent regulation: a tale of opposing activities. *Development*, 142, 2876–2887.
- Heger, P., Marin, B., Bartkuhn, M., Schierenberg, E. and Wiehe, T. (2012) The chromatin insulator CTCF and the emergence of metazoan diversity. *Proceedings of the National Academy of Sciences USA*, 109, 17507–17512.
- Ho, L. and Crabtree, G.R. (2010) Chromatin remodelling during development. *Nature*, 463, 474–484.
- Katoh, K., Asimenos, G. and Toh, H. (2009) Multiple alignment of DNA sequences with MAFFT. *Methods in Molecular Biology*, 537, 39–64.
- Kharchenko, P.V., Alekseyenko, A.A., Schwartz, Y.B., Minoda, A., Riddle, N.C., Ernst, J. *et al.* (2011) Comprehensive analysis of the chromatin landscape in *Drosophila melanogaster*. *Nature*, 471, 480–485.
- Kim, S., Yu, N.K. and Kaang, B.K. (2015) CTCF as a multifunctional protein in genome regulation and gene expression. *Experimental & Molecular Medicine*, 47, e166.
- Kingston, R.E. and Tamkun, J.W. (2014) Transcriptional regulation by trithorax-group proteins. *Cold Spring Harbor Perspectives in Biology*, 6, a019349.
- Kirkness, E.F., Haas, B.J., Sun, W., Braig, H.R., Perotti, M.A., Clark, J.M. *et al.* (2010) Genome sequences of the human body louse and its primary endosymbiont provide insights into the permanent parasitic lifestyle. *Proceedings of the National Academy of Sciences USA*, 107, 12168–12173.
- Kosiol, C., Vinar, T., da Fonseca, R.R., Hubisz, M.J., Bustamante, C.D., Nielsen, R. *et al.* (2008) Patterns of positive selection in six mammalian genomes. *PLoS Genetics*, 4, e1000144.
- Lee, H.G., Kahn, T.G., Simcox, A., Schwartz, Y.B. and Pirrotta, V. (2015) Genome-wide activities of Polycomb complexes control pervasive transcription. *Genome Research*, 25, 1170–1181.
- Li, Z., Tatsuke, T., Sakashita, K., Zhu, L., Xu, J., Mon, H. *et al.* (2012) Identification and characterization of Polycomb group genes in the silkworm, *Bombyx mori*. *Molecular Biology Reports*, 39, 5575–5588.
- Messier, W. and Stewart, C.B. (1997) Episodic adaptive evolution of primate lysozymes. *Nature*, 385, 151–154.
- Misof, B., Liu, S., Meusemann, K., Peters, R.S., Donath, A., Mayer, C. *et al.* (2014) Phylogenomics resolves the timing and pattern of insect evolution. *Science*, 346, 763–767.
- Mohan, M., Herz, H.M., Smith, E.R., Zhang, Y., Jackson, J., Washburn, M.P. *et al.* (2011) The COMPASS family of H3K4 methylases in *Drosophila*. *Molecular and Cellular Biology*, 31, 4310–4318.
- Moore, A.D. and Bornberg-Bauer, E. (2012) The dynamics and evolutionary potential of domain loss and emergence. *Molecular Biology and Evolution*, 29, 787–796.
- Nakamura, T., Blechman, J., Tada, S., Rozovskaia, T., Itoyama, T., Bullrich, F. *et al.* (2000) huASH1 protein, a putative transcription factor encoded by a human homologue of the *Drosophila* ash1 gene, localizes to both nuclei and cell-cell tight junctions. *Proceedings of the National Academy of Sciences USA*, 97, 7284–7289.
- Nguyen, L.T., Schmidt, H.A., von Haeseler, A. and Minh, B.Q. (2015) IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, 32, 268–274.
- Pauli, T., Vedder, L., Dowling, D., Petersen, M., Meusemann, K., Donath, A. *et al.* (2016) Transcriptomic data from panarthropods shed new light on the evolution of insulator binding proteins in insects: insect insulator proteins. *BMC Genomics*, 17, 861.
- Schuettengruber, B., Martinez, A.M., Iovino, N. and Cavalli, G. (2011) Trithorax group proteins: switching genes on and keeping them active. *Nature Reviews Molecular Cell Biology*, 12, 799–814.
- Schwartz, Y.B., Kahn, T.G., Stenberg, P., Ohno, K., Bourgon, R. and Pirrotta, V. (2010) Alternative epigenetic chromatin states of polycomb target genes. *PLoS Genetics*, 6, e1000805.
- Schwartz, Y.B. and Pirrotta, V. (2013) A new world of Polycombs: unexpected partnerships and emerging functions. *Nature Reviews Genetics*, 14, 853–864.
- Simao, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., and Zdobnov, E.M. (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, 31, 3210–3212.

- Simola, D.F., Graham, R.J., Brady, C.M., Enzmann, B.L., Deplan, C., Ray, A. *et al.* (2016) Epigenetic (re)programming of caste-specific behavior in the ant *Camponotus floridanus*. *Science*, 351, aac6633.
- Simola, D.F., Ye, C., Mutti, N.S., Dolezal, K., Bonasio, R., Liebig, J. *et al.* (2013) A chromatin link to caste identity in the carpenter ant *Camponotus floridanus*. *Genome Research*, 23, 486–496.
- Toll-Riera, M. and Alba, M.M. (2013) Emergence of novel domains in proteins. *BMC Evolutionary Biology*, 13, 47.
- Truman, J.W. and Riddiford, L.M. (1999) The origins of insect metamorphosis. *Nature*, 401, 447–452.
- van Baren, M.J. and Brent, M.R. (2006) Iterative gene prediction and pseudogene removal improves genome annotation. *Genome Research*, 16, 678–685.
- Whitcomb, S.J., Basu, A., Allis, C.D. and Bernstein, E. (2007) Polycomb Group proteins: an evolutionary perspective. *Trends in Genetics*, 23, 494–502.
- Yan, H., Simola, D.F., Bonasio, R., Liebig, J., Berger, S.L. and Reinberg, D. (2014) Eusocial insects as emerging models for behavioural epigenetics. *Nature Reviews Genetics*, 15, 677–688.
- Yandell, M. and Ence, D. (2012) A beginner's guide to eukaryotic genome annotation. *Nature Reviews Genetics*, 13, 329–342.
- Yang, X.J. (2004) Lysine acetylation and the bromodomain: a new partnership for signaling. *BioEssays*, 26, 1076–1087.
- Yang, Z. (1998) Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Molecular Biology and Evolution*, 15, 568–573.
- Yang, Z. (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, 24, 1586–1591.

Manuscript received September 10, 2017

Final version received November 4, 2017

Accepted November 5, 2017

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's web-site:

Fig. S1. Summary of non-arthropod PcG and TrxG genes in 176 genomes studied. The BLAST searches against the NCBI nonredundant (nr) protein database were used to determine the origin of non-arthropod genes. As the best hit in the nr protein database was in the corresponding genome, it was excluded in origin determination.

Fig. S2. Phylogenetic distribution of PcG genes in arthropod phylogeny. Blue and transparent circles indicate the presence and absence of a given PcG gene in genomes, respectively. Gray circles indicate the presence of a po-

tential fragmentary PcG gene with a length of less than 60% of its corresponding homolog in *D. melanogaster*. Multiple blue circles indicate the copy number of identified PcG genes. One representative was elected for each genus, and 112 species were included. Table S2 presents the abbreviations of insect species.

Fig. S3. Phylogenetic distribution of TrxG genes in arthropod phylogeny. The TrxG genes in this figure include *E(bx)*, *Iswi*, *polybromo*, *Rbbp5*, *wds*, *HDAC1*, *Utx*, *ash1*, *ash2*, *Bap170*, *Mi-2*, *mor*, *Nurf-38*, *Sbf*, *Snr1*, *Bap55*, and *Caf1-55*. Blue and transparent circles indicate the presence and absence of a given TrxG gene in genomes, respectively. Gray circles indicate the presence of a potential fragmentary TrxG gene with a length of less than 60% of its corresponding homolog in *D. melanogaster*. Multiple blue circles indicate the copy number of identified PcG genes. One representative was selected for each genus, and 112 species were included. Table S2 presents the abbreviations of insect species.

Fig. S4. Phylogenetic distribution of TrxG genes in arthropod phylogeny. The TrxG genes in this figure include *Cfp1*, *Hcf*, *MBD-like*, *nej*, *brm*, *trr*, *kis*, *Mnn1*, *osa*, *Set1*, *Wdr82*, *Bap60*, *MTA1-like*, *Ptip*, *Chd1*, *Bap111*, and *e(y)3*. Blue and transparent circles indicate the presence and absence of a given TrxG gene in genomes, respectively. Gray circles indicate the presence of a potential fragmentary TrxG gene with a length of less than 60% of its corresponding homolog in *D. melanogaster*. Multiple blue circles indicate the copy number of identified PcG genes. One representative was selected for each genus, and 112 species were included. Table S2 presents the abbreviations of insect species.

Fig. S5. Phylogenetic distribution of TrxG genes in arthropod phylogeny. The TrxG genes in this figure include *simj*, *trx*, *Pa1*, *Ncoa6*, *Rm62*, and *Dpy-30L1*. Blue and transparent circles indicate the presence and absence of a given TrxG gene in genomes, respectively. Gray circles indicate the presence of a potential fragmentary TrxG gene with a length of less than 60% of its corresponding homolog in *D. melanogaster*. One representative was elected for each genus, and 112 species were included. Multiple blue circles indicate the copy number of identified PcG genes. Table S2 presents the abbreviations of insect species.

Table S1. 180 arthropod species and 2 non-arthropod outgroup species involved in this study.

Table S2. BUSCO assessment of official gene sets.

Table S3. Non-arthropod PcG and TrxG genes potentially from DNA contamination in genome sequencing.

Table S4. PcG and TrxG genes identified in this study.

Table S5. Domain structures of the PcG and TrxG genes identified in this study.