doi:10.1111/evo.12155

SUCCESSIVE GAIN OF INSULATOR PROTEINS IN ARTHROPOD EVOLUTION

Peter Heger,[1,2] Rebecca George,[1] and Thomas Wiehe[1,3]

[1] Cologne Biocenter, Institute for Genetics, University of Cologne, Zülpicher Straße 47a, 50674 Köln, Germany

[2] E-mail: peter.heger@uni-koeln.de

[3] E-mail: twiehe@uni-koeln.de

Received July 22, 2012
Accepted April 8, 2013

Alteration of regulatory DNA elements or their binding proteins may have drastic consequences for morphological evolution. Chromatin insulators are one example of such proteins and play a fundamental role in organizing gene expression. While a single insulator protein, CTCF (CCCTC-binding factor), is known in vertebrates, *Drosophila melanogaster* utilizes six additional factors. We studied the evolution of these proteins and show here that—in contrast to the bilaterian-wide distribution of CTCF—all other *D. melanogaster* insulators are restricted to arthropods. The full set is present exclusively in the genus *Drosophila* whereas only two insulators, Su(Hw) and CTCF, existed at the base of the arthropod clade and all additional factors have been acquired successively at later stages. Secondary loss of factors in some lineages further led to the presence of different insulator subsets in arthropods. Thus, the evolution of insulator proteins within arthropods is an ongoing and dynamic process that reshapes and supplements the ancient CTCF-based system common to bilaterians. Expansion of insulator systems may therefore be a general strategy to increase an organism's gene regulatory repertoire and its potential for morphological plasticity.

KEY WORDS: Adaptive evolution, barrier element, BEAF-32, CP190, GAGA factor, gene loss, lineage-specific genes, Mod(mdg4), Su(Hw), Zw5.

Since more than a century, the molecular causes of morphological change are being examined in the fruit fly *Drosophila melanogaster* (Morgan 1911; Altenburg and Muller 1920). A well-studied example of morphological change are homeotic mutations. They alter the identity of particular body parts by transforming them into other parts (for review, see Lewis 1978). Genetic analysis of these mutations revealed that they often affect regulatory elements controlling the expression of an associated homeotic (Hox) gene (see Pfeifer et al. 1987, for examples). Misexpression of genes by regulatory mutations can therefore contribute to morphological change.

Comparative studies in additional arthropods demonstrated that differences in Hox gene expression are correlated with morphological differences across the phylum (e.g., Warren et al. 1994; Averof and Akam 1995; Averof and Patel 1997; Abzhanov and Kaufman 2000; Hughes and Kaufman 2002). Expression of the Hox gene *ubx*, for example, is linked to abdominal limb number (Warren et al. 1994; Averof and Patel 1997) and functional studies suggest that altered Hox gene expression is indeed a cause of morphological diversity (Lewis et al. 1999; Liubicich et al. 2009; Pavlopoulos et al. 2009). Also in other contexts, there is ample evidence that mutations in regulatory elements play an important role for the evolution of morphological traits (Jeong et al. 2008; Peter and Davidson 2011).

Regulatory elements, however, are abundant and can exert their function over a wide range of physical distances (Miele and Dekker 2008; Lieberman-Aiden et al. 2009). To protect genes from the inappropriate influence of these sequences, a process called chromatin insulation participates in the creation of independent chromatin domains (for review, see Wallace and Felsenfeld 2007; Yang and Corces 2012). As mediators of this kind of regulation, insulator proteins and the mechanisms by which they act have been studied intensely (see Van Bortle and Corces 2013, for a recent review].

On the basis of its ability to protect genes from position effects in transgenic flies, Suppressor of Hairy Wing [Su(Hw)] was the first insulator protein to be identified (Spana et al. 1988; Geyer and Corces 1992). To date, several additional proteins with insulator activity are known in *D. melanogaster*: Boundary Element Associated Factors (BEAF-32A and B), Zeste-white 5 (Zw5), GAGA Associated Factor (GAF), Modifier of mdg4 [Mod(mdg4)], Centrosomal Protein 190 (CP190), and dCTCF, the *D. melanogaster* ortholog of mammalian CCCTC-binding factor. Although insulator proteins have been described originally as enhancer blockers when positioned between a promoter and an enhancer, there is evidence for additional and more complex functions. An emerging role is their involvement in the spatiotemporal control of gene expression by modifying long-range chromosomal interactions, suggesting that they are key players in establishing an appropriate three-dimensional chromosome structure during cell differentiation and development (reviewed in Van Bortle and Corces 2013).

In agreement with this view, knockdown or mutation of insulator proteins and the sequences they bind has severe consequences. Impairment of dCTCF and the deletion of CTCF-binding sites, for example, eliminates boundary elements required for proper Hox gene expression and leads to homeotic transformations (Mohan et al. 2007; Iampietro et al. 2010); expressing dominant-negative BEAF-32 during embryogenesis is lethal (Gilbert et al. 2006) and disturbs Hox gene expression (Roy et al. 2011); mutations in *trithorax-like*, the gene encoding GAGA factor, display a maternal effect lethal phenotype and abnormalities in the expression of homeotic genes (Biggin and Tjian 1988; Bhat et al. 1996; Ohtsuki and Levine 1998; Belozerov et al. 2003). Recently, comparative ChIP-seq analysis in several *Drosophila* species revealed that CTCF and BEAF-32 are directly involved in the evolution of gene expression and genome organization through adaptive changes in their respective binding sites (Ni et al. 2012; Yang et al. 2012). Thus, insulator proteins regulate fundamental processes during *Drosophila* development, and evolutionary changes in the binding pattern of these factors have direct consequences for gene expression and phenotype.

The conservation of many *D. melanogaster* insulator binding sequences (Holohan et al. 2007; Negre et al. 2010; Ni et al. 2012) together with studies in other animals (e.g., Heger et al. 2012; Schmidt et al. 2012) indicates further that this fundamental aspect of genetic regulation may be relevant across different phyla. One would expect therefore that most eukaryots possess orthologous genes to implement insulator mechanisms. However, the phylogenetic distribution of only one factor, CTCF, has been investigated in detail (Heger et al. 2012). In this study, we examine the origin of the other known *D. melanogaster* insulator proteins.

Table 1. Number of collected candidate sequences. Using BLAST searches with a specified cutoff (threshold), the number of arthropod candidate sequences retrieved in total, the number of unique sequences, and the number of sequences retained after clustering are shown. Candidates for a vertebrate GAF were collected from *Danio rerio*, *Homo sapiens*, and *Strongylocentrotus purpuratus*.

Insulator	Threshold	Total	Unique	After clustering
CTCF, Su(Hw), Zw5	10^{-05}	8929	4245	587
CP190, GAF, Mod(mdg4)	10^{-14}	5166	1501	n. d./727
BEAF-32	∞	130	64	n. d.
Vertebrate GAF	10^{-05}	1135	161	n. d.

n. d., not determined.

Results

Seven proteins with insulator function have been described in *D. melanogaster*. In contrast, our previous work showed that orthologs of only one chromatin insulator, CTCF, can be found in nematodes (Heger et al. 2009) echoing the situation in vertebrates (Phillips and Corces 2009). Thus, the possession of additional insulator systems might be an arthropod-, insect-, or *Drosophila*-specific property. To investigate this idea, we searched the sequence databases at NCBI for putative orthologs of the seven *D. melanogaster* insulator proteins CTCF, Su(Hw), Zw5, CP190, Mod(mdg4), GAF, and BEAF-32. Using the respective *D. melanogaster* sequences as query, we performed within the arthropod phylum separate searches for each protein and retrieved > 14,000 candidates in total (Table 1). As many of these sequences (59.2%) were collected multiple times, we removed redundancy and retained 5810 unique sequences (Table 1). Subsequently, we performed in two parallel workflows clustering and phylogenetic analysis of the ZF (zinc finger) and of the BTB (broad complex, tramtrack, and bric-à-brac) domain containing subsets of insulator proteins.

THE ZF DOMAIN CONTAINING INSULATORS: CTCF, Su(Hw), AND Zw5

The insulator proteins Su(Hw), CTCF, and Zw5 are poly-ZF proteins with 12, 11, and eight C2H2 ZF domains, respectively (Fig. 1). The ZF domain constitutes an ancient DNA-binding motif present in all eukaryotes and also in some Archaea (Bouhouche et al. 2000) and the C2H2 ZF in particular is the most common DNA-binding motif of eukaryotic transcription factors (Clarke and Berg 1998; Tadepally et al. 2008). Thus, it is not surprising that our search recovered more than 4200 candidates (Table 1) belonging to 227 different arthropod species (Table S1).

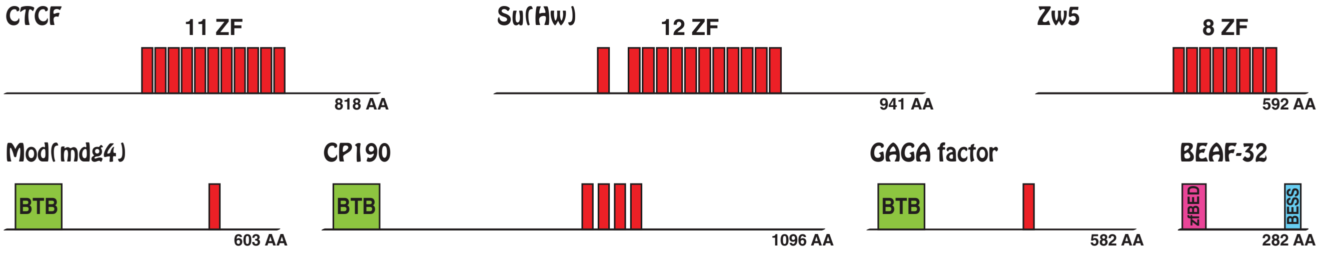


Figure 1. Domain structure of *D. melanogaster* insulator proteins. The domain composition of known *D. melanogaster* insulator proteins is drawn approximately to scale. Upper row: Factors with poly-zinc finger (ZF) domain. Individual ZFs (red) comprise about 29 AA and are regularly spaced, except for Su(Hw)'s first ZF. Lower row: Three insulator proteins with BTB (~105 AA, green) + ZF domain (red) and BEAF-32 (with zf-BED and BESS domain).

To extract potential insulators from the 4245 candidates, we clustered the sequences according to their similarity to known insulators and obtained a set with 587 proteins. We next aligned these sequences and determined their orthology to given ZF insulators with phylogenetic methods. The resulting maximum likelihood tree displayed well-supported clusters for CTCF and Zw5, but low support for a Su(Hw) cluster (not shown). To prevent this problem, we extracted according to the maximum likelihood tree the members of potential insulator protein clusters and evaluated them separately in new experiments, thereby omitting the bulk of nonorthologous sequences. Using this strategy, we obtained high support for all three groups of ZF insulator proteins (CTCF, Su(Hw), and Zw5; Fig. S1).

To visualize the phylogenetic composition of these clusters, we mapped the source organism of the respective sequences to a consensus arthropod phylogeny (Figs. 2, S1). We found that CTCF orthologs are present in all arthropod groups with a sequenced genome and in many unsequenced species of the three arthropod subgroups (Fig. S1). These results emphasize the importance of this factor for arthropod biology and agree with a more general role of CTCF in bilaterians (Heger et al. 2012). In addition, they indicate that our strategy is able to detect orthologous sequences in arthropods with confidence.

The phylogenetic distribution of Su(Hw) was similar to that of CTCF, with all three arthropod subgroups being represented in the respective cluster (Figs. 2, S1). As we could not find orthologs of Su(Hw) in nematodes, a sister phylum to arthropods, in a previous study (Heger et al. 2009), it is likely that this protein evolved at the base of arthropods or close to this base. Although this conclusion is derived from a modest number of sequences (two chelicerate, two crustacean) and the absence of a detectable ortholog in fully sequenced nematode genomes, the branching pattern and lineage-specific synapomorphies of the chelicerate and crustacean candidates argue for a common ancestry with Su(Hw) orthologs from insects.

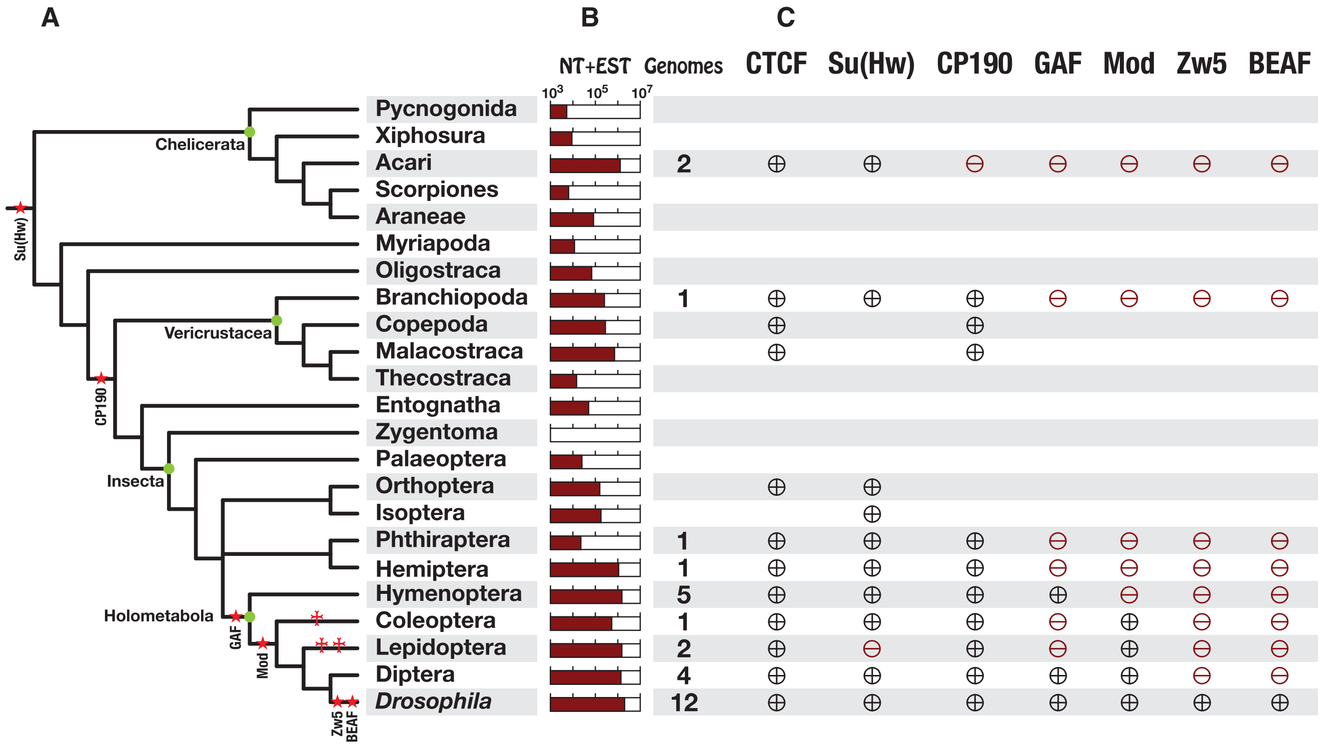
We observed a misplacement of crustacean and chelicerate Su(Hw) and a split of dipteran sequences in a few experiments (Fig. S1 and data not shown). These inconsistencies with accepted arthropod relationships are likely a consequence of our dataset (single gene phylogeny with many short and incomplete sequences) and of the fast evolution in *Drosophila* (Savard et al. 2006) and also affect the CTCF and YY1 clusters (Fig. S1). Indeed, it is well known that gene trees and species trees do not necessarily agree if the number of analyzed loci is small (Pamilo and Nei 1998.; Degnan and Rosenberg 2006). Even if our data are not sufficient to reconstruct the correct species genealogy in all detail, they offer substantial support for the existence of a distinct clade of Su(Hw) orthologs with representatives from all three arthropod subphyla.

Despite its broad distribution, Su(Hw) is not indispensable for arthropods. While orthologs to other ZF proteins such as CTCF or YY1 can be found in all arthropod lineages, there is no evidence for the presence of Su(Hw) in Lepidoptera (butterflies) although large amounts of expressed sequence tags (ESTs) and several genome sequences are available in this group (Figs. 2, S1).

When we analyzed the Zw5 cluster in preliminary experiments, we noticed that it seemed to contain exclusively sequences from the 12 *Drosophila* species we used to define this cluster. Occasionally however, some other arthropod sequences were positioned nearby. When we probed the reliability of this association in smaller datasets, the 12 *Drosophila* Zw5 orthologs alone gave rise to a distinct and highly supported cluster in all cases (Fig. S1). Thus, no sequences from other dipterans or more distantly related arthropods are orthologous to *Drosophila* Zw5 although, for example, more than 2000 ZF sequences from 57 non-*Drosophila* dipterans were present in the "unique" dataset (Table S1). These results indicate that the insulator protein Zw5 is specific for the genus *Drosophila* (Fig. 2) as it has been suggested in a previous study (Schoborg and Labrador 2010). To investigate whether Zw5 is a true synapomorphy of drosophilids, sequences from closely related brachyceran outgroups need to be analyzed.

THE BTB DOMAIN CONTAINING INSULATORS: GAGA FACTOR, Mod(mdg4), AND CP190

Like the C2H2 ZF domain, the BTB domain is ancient and found in all eukaryotes (Perez-Torrado et al. 2006). It consists of an



	Genomes	CTCF	Su(Hw)	CP190	GAF	Mod	Zw5	BEAF
Pycnogonida								
Xiphosura								
Acari	2	⊕	⊕	⊖	⊖	⊖	⊖	⊖
Scorpiones								
Araneae								
Myriapoda								
Oligostraca								
Branchiopoda	1	⊕	⊕	⊕	⊖	⊖	⊖	⊖
Copepoda		⊕		⊕				
Malacostraca		⊕		⊕				
Thecostraca								
Entognatha								
Zygentoma								
Palaeoptera								
Orthoptera		⊕	⊕					
Isoptera			⊕					
Phthiraptera	1	⊕	⊕	⊕	⊖	⊖	⊖	⊖
Hemiptera	1	⊕	⊕	⊕	⊖	⊖	⊖	⊖
Hymenoptera	5	⊕	⊕	⊕	⊕	⊖	⊖	⊖
Coleoptera	1	⊕	⊕	⊕	⊖	⊕	⊖	⊖
Lepidoptera	2	⊕	⊖	⊕	⊖	⊕	⊖	⊖
Diptera	4	⊕	⊕	⊕	⊕	⊕	⊖	⊖
Drosophila	12	⊕	⊕	⊕	⊕	⊕	⊕	⊕

Figure 2. **Successive gain of insulator proteins during arthropod evolution. (A) Schematic representation of arthropod relationships (after Regier et al. (2010); Simon et al. (2009); Wiegmann et al. (2009); Trautwein et al. (2012)). Green dots highlight the position of major arthropod groups. The birth (*) and loss (†) of each insulator protein is indicated in red. Several clades/orders are omitted for clarity. Ephemeroptera (mayflies) and Odonata (dragonflies) are combined in Palaeoptera. (B) For the taxa shown in A, the number of ESTs and nucleotide sequences deposited at NCBI (http://www.ncbi.nlm.nih.gov; reference day is May 7, 2012) is indicated as logarithmic bar graph, along with the number of available genome sequences (right column). Sequence data are biased toward holometabolous insects and particular crustacean and chelicerate lineages. Zygentoma have less than 1000 sequences (empty bar graph). (C) Phylogenetic mapping of insulator proteins. Presence of a particular insulator protein is indicated by ⊕, absence by ⊖. Absence is only indicated in orders with sufficient EST and genomic data. Results refer to the BLAST/phylogeny-based strategy and are identical for the HMM/OrthoMCL-based strategy, except some minor modifications (Table S4).**

N-terminal 105 amino acid motif that mediates homo- and heteromeric dimerization in a number of *Drosophila* transcriptional regulators, for example, *broad complex*, *tramtrack*, and *bric-à-brac* (Zollman et al. 1994). Searching for BTB domain containing insulators, we collected more than 1500 candidate sequences (Table 1) that belonged to 111 arthropod species (Table S2). To determine which of these sequences could be orthologous to a *D. melanogaster* insulator, we clustered them and obtained a dataset with 727 sequences. A maximum likelihood analysis of this dataset indicated high support for distinct CP190 and GAGA factor clusters. In contrast, the Mod(mdg4) cluster, containing more than 200 nearly identical sequences from Lepidoptera, was supported less well (not shown). We therefore extracted from the previous sequence set all potential orthologs of CP190, Mod(mdg4) (9/220 sequences), and GAF and determined their orthology to the *Drosophila* insulators in additional experiments.

These new analyses revealed that only two insect orders were represented within the GAF cluster: Hymenoptera and Diptera (Figs. 2, S2). We could not identify GAF orthologs in other arthropods despite the availability of substantial genomic and EST data (Fig. 2). This suggests that GAGA factor evolved in the last common ancestor of Hymenoptera and Diptera and has been lost at least twice during evolution of holometabolous insects, in Coleoptera (beetles) and Lepidoptera (butterflies).

In previous studies, putative vertebrate homologs of the GAGA factor have been reported (Matharu et al. 2010; Kumar 2011). As vertebrates and *Drosophila* share a common ancestor with all other bilaterians, these conclusions imply that GAGA factors should be present in other protostomes and deuterostomes. To investigate this assumption, we collected with a relaxed E-value of 10^{-05} more than 150 candidate sequences from three different deuterostome lineages (vertebrates: *Danio rerio* and *Homo sapiens*; echinoderms: *Strongylocentrotus purpuratus*), including four proposed vertebrate GAFs, and analyzed their relation to the insect GAF cluster. However, none of these sequences localized to the highly supported cluster (97% bootstrap support; Fig. S3).

Rather, the proposed vertebrate GAFs formed a separate cluster, arguing for a common ancestry of these proteins in the vertebrate lineage.

To test whether this result is a consequence of insufficient BLAST sensitivity, we generated from the members of the insect GAGA cluster two representative HMM profiles (full length and BTB domain only) and performed a more sensitive profile–profile search on the HHpred server (http://toolkit.tuebingen.mpg.de/hhpred). We retrieved from these searches 45 additional candidates from humans and mice and examined their relationship to the insect counterparts. As in the previous case, the resulting trees did not place any of the new candidates to the insect GAGA cluster (Fig. S4). Thus, none of the 200 analyzed deuterostome candidates is closer related to the insect GAGA cluster than any other candidate, strongly limiting the possibility that there is among them a GAGA ortholog, that is, a sequence that originated from the same last common ancestor than insect GAFs. Given that we could not identify GAF orthologs in arthropods preceding the Hymenoptera–Diptera split (Fig. S2) and in nematodes (Heger et al. 2009), these results indicate that the GAGA factor is unique to particular lineages of holometabolous insects and is related to the proposed homologs in other phyla by the presence of an ancient BTB domain. This conclusion contradicts Matharu et al. (2010) and Kumar (2011). However, Kumar (2011) presented his result on the basis of an HHpred search initiated with a single sequence and without verification in a phylogenetic context whereas Matharu et al. (2010) carried out a phylogenetic analysis without bootstrapping, also lacking phylogenetic implications. Our comprehensive survey of candidates within the whole range of arthropods, nematodes (Heger et al. 2009), and several deuterostomes, including those proposed by Matharu et al. (2010) and Kumar (2011), argues that GAGA factor originated in the ancestor of Hymenoptera and Diptera rather than in the ancestor of the Bilateria. A functional similarity that has been attributed to the proposed vertebrate GAGA factors (Matharu et al. 2010) therefore likely involves convergent evolution.

After removing most lepidopteran sequences (231/240) that clustered to *D. melanogaster* Mod(mdg4) in the 727 candidate set, we newly analyzed the candidates positioned within each supposed BTB cluster (108 sequences). Our results showed high bootstrap support (100%) for the presence of Mod(mdg4) orthologs in *Drosophila*, other Diptera, Lepidoptera (butterflies), and Coleoptera (beetles; Figs. 2, S2). Although we collected in the original dataset 277 sequences from hymenopterans, the next possible sister group, none of these were orthologous to Mod(mdg4). Taking into account the availability of five genome sequences, this indicates that Mod(mdg4) does not belong to the gene repertoire of hymenopterans (Fig. 2). These findings confirm previous reports of a *mod(mdg4)* locus in Lepidoptera (Dorn and Krauss 2003; Krauss and Dorn 2004) and establish the origin of this locus even earlier, in the common ancestor of Coleoptera, Lepidoptera, and Diptera.

In *D. melanogaster*, the *mod(mdg4)* locus gives rise to > 20 different isoforms that share an N-terminal 405 AA region containing the BTB domain (Buchner et al. 2000). A similarly complex organization was reported for the *mod(mdg4)* locus of lepidopterans (Shao et al. 2012). To find out whether this feature is also present in Coleoptera, we searched for *Tribolium castaneum* ESTs that share their 5′ region (with BTB domain), but have different 3′ ends. However, we could not find evidence for different isoforms in 12 ESTs that mapped in *T. castaneum* to the BTB domain region of the *mod(mdg4)* locus, as it is indicated by EST GI:189241700 (ChLG10:8,779,000–8,780,000; Fig. S2). Alignment of 18 sequences from the beetle *Dendroctonus ponderosae* that belonged to the *mod(mdg4)* cluster did not give evidence for the existence of separate isoformes either. Thus, currently available data from Coleoptera are not able to resolve the origins of the complex *mod(mdg4)* locus.

The *D. melanogaster* CP190 protein is an essential component of many insulator complexes organized by CTCF, Su(Hw), and BEAF-32 (Pai et al. 2004; Mohan et al. 2007; Bartkuhn et al. 2009; Negre et al. 2010; Van Bortle et al. 2012). Its N-terminal BTB domain is indispensable in providing this activity (Oliver et al. 2010). Our search for CP190 orthologs in arthropods revealed a distinct set of sequences clustering to *D. melanogaster* CP190 with high confidence. These sequences covered three crustacean branches and all insects with a sequenced genome, but not the Chelicerata although two genome sequences exist and 71 candidate sequences from different chelicerate lineages were present in the "unique" dataset (Figs. 2, S2). Thus, it is likely that CP190 originated in the ancestor of hexapods and crustaceans. As we could not observe a loss in any of the well-sampled taxa, the interaction between CP190 and CTCF/Su(Hw) insulator complexes could be a critical feature of all pancrustaceans (Crustacea plus Hexapoda; Regier et al. 2010).

BEAF-32

The insulator protein BEAF-32 exists in two isoforms and is associated with chromosomal domains (Zhao et al. 1995) and transcriptionally active regions in *D. melanogaster* (Jiang et al. 2009). It has an unusual N-terminal ZF, the BED finger (58 AA), and a C-terminal BESS domain (35 AA; Fig. 1). An extensive analysis of BEAF-32 has been performed by Schoborg and Labrador (2010), which suggested, on the basis of BLAST experiments, that BEAF-32 is specific to the *Drosophila* genus. We wanted to challenge this finding with a more powerful phylogenetic approach. Despite the relaxed threshold (E-value $= \infty$), we obtained with our search only 64 candidate sequences from the whole arthropod phylum (Table 1). None of these sequences was a reasonable BEAF-32 candidate (data not shown). In agreement with the previous study

(Schoborg and Labrador 2010), we assume therefore that BEAF-32, like Zw5, is a *Drosophila*-specific insulator protein absent from other insects (Fig. 2).

AN HMM/OrthoMCL-BASED PIPELINE TO VALIDATE OUR CONCLUSIONS

Our results suggest that individual insulator proteins have been acquired at different stages of arthropod evolution. To rule out that these observations suffer from a lack of sensitivity and do not reflect the underlying evolutionary history, we complemented the BLAST-based analysis with a more sensitive approach, a combination of HMM scans and OrthoMCL-clustering of candidates into orthologous groups, performed on all accessible sequence data from 26 species in all groups of protostomes available (Table S3). To achieve the best possible coverage, we translated every genome assembly and the unplaced reads data into six open reading frames (ORFs) and combined the resulting ORFeomes with their corresponding downloaded protein sets. With this approach, the potential failure to detect an ortholog cannot be attributed to the often incorrect or incomplete annotation of proteomes. To find insulator orthologs in this wealth of sequence data, we prepared from the clusters obtained in the previous approach HMM profiles specific for each insulator protein (except CTCF). Scanning the 26 datasets with all six insulator profiles yielded a total of 39,573 unique candidate sequences below the default threshold. We analyzed the orthology of these sequences to a given insulator protein family using a custom implementation of the OrthoMCL clustering pipeline (Li et al. 2003). All findings of the previous approach, which employed BLAST search and phylogenetic reconstruction, could be confirmed or further refined in this workflow (Table S4). Some important additional aspects shall be mentioned shortly: (1) The six insulator proteins Su(Hw), CP190, Mod(mdg4), Zw5, BEAF-32, and—notably—GAGA factor are specific for arthropods. In none of the seven protostome outgroup species could we find sequences orthologous to these proteins whereas the previously reported pattern of CTCF occurrence in annelids, molluscs, platyhelminthes, and nematodes (Heger et al. 2009, 2012) could be reproduced accurately. (2) Su(Hw) could not be found in an additional butterfly genome, confirming that it may have been lost in this insect order. (3) If GAGA factor disappeared secondarily in Lepidoptera and Coleoptera (Fig. 2 and Table S4), it should also be missing in Strepsiptera. Our results from the genome scan of *Mengenilla moldrzyki* are consistent with this assumption. (4) A sequence orthologous to Zw5 is present in *Glossina morsitans*, a dipteran closer to *Drosophila* than the nematoceran flies *Aedes* and *Anopheles*. As this sequence was generated during translation of the genome, it could not be detected with the BLAST-based strategy. This finding demonstrates the power of our methodology and slightly modifies the previous conclusion that Zw5 is *Drosophila* specific.

On the other hand, negative results for particular insulator proteins in some genomes where we expected them may indicate insufficient assembly quality or that the evolution of insulator proteins is more dynamic than anticipated (e.g., Su(Hw) not detected in *Glossina* and *Lepeophtheirus*; no CP190 in *Heliconius*; no GAGA factor in *Acyrthosiphon*; no Mod(mdg4) in *Mengenilla*). Further work is necessary to resolve these issues.

It is possible that a fragment of a particular ortholog is present in our dataset, but was not recognized as ortholog by OrthoMCL, for example, because of its shortness. Indeed, we observed in the HMM-derived candidate set some short open reading frames that belonged to insulator orthologs, but did not appear in the final clusters. For two reasons we think that this shortcoming does not confound our conclusions. First, the "twilight zone" is confined to ORFs in the range of approximately 30–75 AA (ORF minimum length—shortest ORF in an observed orthologous cluster). The majority of ORFs (84.5%) is larger. To completely miss an ortholog in our genome scan requires that all ORFs corresponding to that ortholog are within this range, a possible, but unlikely event. Second and more important, these limitations apply to the 26 genomes likewise. It is therefore implausible that deficiencies of our methodology generate the evolutionary patterns we report.

Finally, our results suggest that there is a single BEAF-32 ortholog outside *Drosophila* in the distantly related insect *Pediculus humanus* (Table S4). To explain this unexpected result, we analyzed the domain composition of the 95 BEAF-32 candidates from Table S4 and observed that a duplicated zfBED domain is present exclusively in this protein (ID: PHUM580690-PA). As the best BLAST hit of this sequence in *D. melanogaster* is not BEAF-32 either, we conclude that the domain duplication led to a false-positive signal and erroneously triggered its inclusion into the BEAF orthology group.

Discussion

Insulator proteins confer activity to insulators sequences, a class of functional elements with important roles in the regulation of chromosomal organization. In this study, we investigated the origin of the seven known proteins associated with insulator activity in *D. melanogaster*. To find potential orthologs of these proteins in other organisms, we followed two independent strategies. On the one hand, we performed BLAST searches and analyzed a large number of candidate sequences in phylogenetic experiments (Table 1; Figs. S1–S4). On the other hand, we combined HMM searches and MCL clustering to examine which of the candidates share orthology with a known *Drosophila* insulator (Table S4). With both methods we found robust evidence that the known *Drosophila* insulators (except CTCF) are restricted to arthropods and have been acquired successively during arthropod evolution

(Fig. 2; Table S4). We draw these conclusions on the basis of several observations.

First, we did not find evidence for the existence of known *Drosophila* insulator proteins outside the arthropods by analyzing seven genomes from four protostome phyla (Annelida, Mollusca, Platyhelminthes, Nematoda). It is unlikely that a lack of sensitivity is responsible for this result as we find CTCF orthologs in the annelid *Capitella* spI, in the mollusc *Lottia gigantea*, and in the nematode *Trichinella spiralis*, faithfully replicating the results of previous work with an independent method (Heger et al. 2012). The detection of CTCF in all 19 analyzed arthropod genomes, including so far unknown orthologs (e.g., from the myriapod *Strigamia maritima* and the strepsipteran *M. moldrzyki*), further confirms the specificity and sensitivity of our approach.

Second, the ZF proteins CTCF and Su(Hw) are consistently present in arthropods from which genome sequences are available plus in some additional orders with a significant number of ESTs. In both cases, the presence of orthologous sequences in the three subgroups of arthropods indicates that the common ancestor of arthropods already had these proteins. Such an assumption is well supported for the CTCF protein that has been found in all bilaterians (Heger et al. 2012). The origin of the Su(Hw) protein is less clear. As it is absent in nematodes (Table S4; Heger et al. 2009), it could have evolved in the ancestor of arthropods or in the common ancestor of arthropods and a closely related ecdysozoan sister group, for example, tardigrades or onychophorans. The resolution of our study is not sufficient to answer this question.

Third, in contrast to CTCF and Su(Hw), the proteins Zw5 and BEAF-32 are restricted to a remarkably limited subset of arthropods. We obtained with both methods only few BEAF-32 candidates outside the *Drosophila* genus (64 and 95, respectively), and none of them could be placed into the *Drosophila* BEAF-32 orthology group. Although the domain composition of Zw5 issued a much higher number of candidates (28,431) across the 26 genomes (Table S4), we could only recover Zw5 orthologs in *Drosophila* and *G. morsitans*, another brachyceran fly. A Zw5 ortholog could not be found in nematoceran dipteres and other arthropods despite the existence of several sequenced genomes and large amounts of ESTs. These results are based on the most comprehensive study undertaken so far and provide consistent evidence that Zw5 and BEAF-32 likely emerged in or close to the common ancestor of the genus *Drosophila*. They are therefore the most recent additions to the group of insulator proteins.

Finally, our results with respect to the BTB domain containing insulators suggest that at least some of them have evolved at intermediate stages when compared with the "ubiquitous" proteins CTCF/Su(Hw) and the "restricted" factors Zw5/BEAF-32. The mapping of CP190, for example, shows that it is present in all insects with a sequenced genome and in three crustacean orders. The inability to detect this protein in a myriapod and two mite genomes and in a large amount of ESTs from chelicerates suggests that CP190 evolved in the ancestor of Pancrustacea.

GAGA factor, on the other hand, formed a highly supported cluster containing exclusively sequences from Diptera and Hymenoptera in phylogenetic experiments (Fig. S2). The HMM-based approach confirmed this result, but assigned in addition sequences from the hemipteran *Rhodnius prolixus* to the GAGA factor orthology group (Table S4). In all further insect, crustacean, and chelicerate genomes, orthologs were not detectable, indicating that this protein likely emerged in the common ancestor of Hemiptera, Hymenoptera, and Diptera. Importantly, these findings imply that GAGA factors do not exist outside the arthropod phylum. To elucidate the conflict of these findings with the proposed existence of vertebrate GAGA factors (Matharu et al. 2010; Kumar 2011), we performed phylogenetic analysis with two sets of candidates, acquired by BLAST and HMM searches. However, neither the four proposed vertebrate GAGA factors nor our additional candidates were placed to the insect orthology group (Figs. S3, S4), emphasizing the consistency of our results.

Although our study provides evidence for a consecutive gain of insulator proteins in arthropods, it also suggests that insulator evolution is dynamic in terms of losses. Although it is difficult to prove the absence of a gene in potentially inaccurate genome assemblies, our results indicate that inference of at least some secondary losses is reasonable. The two best examples are the repetitive loss of GAGA factor in Coleoptera/Strepsiptera and in Lepidoptera and the loss of Su(Hw) in Lepidoptera, each supported by the analysis of two genomes with both methods (Fig. 2, Table S4). In five additional cases, we were not able to discover an expected ortholog in a single genome (Table S4). This may be a consequence of incomplete genome assemblies, but could alternatively reflect a dynamic nature of insulator evolution in arthropods. Importantly, these patterns of change are confined to the arthropod-specific insulators. We did not observe a loss of the more ancient CTCF insulator in any arthropod genome. This is compatible with the idea that CTCF function is needed for fundamental processes in the Bilateria (Heger et al. 2012) and might have been supplemented and modified by additional components in the arthropod lineage.

Although our results have been established by two fairly independent methods, we cannot formally prove the absence of orthologous proteins from certain clades. With growing databases and dependent on the sensitivity of homology detection tools, the exact placement of the origin of some proteins may still change.

However, irrespective of uncertainties in the exact time of gain and loss, our observations reveal a consistent model for the evolution of the known *D. melanogaster* chromatin insulators through a series of successive acquisitions.

This result has several implications. It has been confirmed repeatedly that insulator proteins colocalize and interact with each other (Gerasimova et al. 1995; Melnikova et al. 2004; Pai et al. 2004; Gerasimova et al. 2007; Bartkuhn et al. 2009; Negre et al. 2010; Van Bortle et al. 2012), thereby creating a network of dependencies that is thought to contribute to cell-specific differences in nuclear organization and gene expression (Yang and Corces 2011). Moreover, there is recent evidence that each *D. melanogaster* insulator subclass, determined by the presence of Su(Hw), CTCF, BEAF, or GAF, shares the CP190 protein and possibly also Mod(mdg4) (Van Bortle et al. 2012; Yang and Corces 2012). Our findings point out that the mechanisms, interactions, and components of insulator complexes must be considerably different from *Drosophila* not only in the great majority of arthropods (that share two of the seven factors), but also in other bilaterian phyla that only have CTCF in common. According to our findings, the presence of different insulator systems and the complex interactions between their components seen in *D. melanogaster* today are a result of ongoing evolution and diversification. These processes started more than ∼ 600 million years ago in the ancestor of bilaterians (www.timetree.org) with CTCF, the oldest known chromatin insulator of multicellular animals (Heger et al. 2012). At or close to the root of arthropods, Su(Hw) emerged. These two basal systems experienced subsequent additions and modifications, with the gain of BEAF-32 in the genus *Drosophila*, ∼ 60 million years ago, being the most recent acquisition. To what extent the successive acquisition of new insulator genes is involved in adaptive processes, shall be the topic of future investigations.

Although we can describe in detail an expansion of insulator proteins only in the *Drosophila* history, this process is not necessarily confined to a single species. Millions of arthropod species trace back to the same common ancestor that was equipped with the CTCF and Su(Hw) insulators. It is therefore possible that other arthropods simultaneously increased their initial repertoire of insulator proteins during the past 600 million years, giving rise to a plethora of unexplored territories. Thus, the expansion of insulator mechanisms that happened in *D. melanogaster* history might in fact be a general characteristic of arthropods and other animals. The description of non-CTCF insulators in echinoderms exemplifies that such an expansion could indeed apply to a greater variety of animals (Yajima et al. 2012). If so, this characteristic could provide a mechanism to modulate an ancient insulator system and fine-tune gene expression in a lineage-specific way.

Methods Summary

BLAST-BASED SEARCH FOR CANDIDATE INSULATOR PROTEINS

With the known *Drosophila* insulator proteins as query (dCTCF, GI:21356747; Su(Hw), GI:33860216; CP190, GI:23171337; GAF, GI:83287912; Mod(mdg4), GI:158030328; Zw5, GI:45549097; BEAF, GI:17647187), standard BLASTX, BLASTP, or TBLASTN (Altschul et al. 1997) searches were conducted in publicly available sequence databases at NCBI (http://blast.ncbi.nlm.nih.gov/). To minimize the chance of missing an ortholog, we performed parallel searches in different databases (nucleotide, EST, and protein) and subdivided a search into smaller entities if a given taxonomic range reported > 500 hits below threshold (= download restriction at the NCBI BLAST web interface). We collected all ZF domain candidates below a relaxed BLAST expectation value of 10^{-05}. We set the threshold for BTB domain protein candidates to 10^{-14} because in preliminary experiments this value effectively incorporated BTB domain proteins distinct from BTB domain containing insulators, for example, *tramtrack* or *broad complex*. Collected nucleotide sequences were translated to the appropriate reading frame using EMBOSS (Rice et al. 2000).

CLUSTERING AND MULTIPLE SEQUENCE ALIGNMENT

We obtained from the initial dataset a collection of unique sequences using the EMBOSS tool “skipredundant” (Rice et al. 2000). We added to this collection a set of reference sequences for each insulator protein. The references served as guide for the clustering step. All other candidates from the initial dataset were passed to the SiLiX clustering algorithm (Miele et al. 2011) as “partial” sequences (command line parameter “-p”). We ran SiLiX with default parameters (35% min. identity; 80% min. overlap; 100 nt min. length; 50% min. overlap of partial sequences). For subsequent analysis, we took all reference sequences plus candidates that clustered to one of the references. Multiple sequence alignments were performed using the Clustal Ω algorithm (Sievers et al. 2011). Alignments were viewed and manually edited using SeaView (Galtier et al. 1996).

PHYLOGENETIC ANALYSIS

For the Zw5, Su(Hw), and CTCF proteins, the described *Drosophila* orthologs served as positive control for cluster identification. Similarly, we included the known *Drosophila* orthologs of the BTB-domain proteins GAGA factor, CP190, and Mod(mdg4) to specify these clusters. Outgroup for the analysis of ZF insulators was the widely distributed ZF transcription factor YY1. For BTB domain insulators, we used as outgroup a set of *lola-like* orthologs from diverse arthropods. Phylogenetic trees resulting from the alignments were computed under the maximum likelihood criterion using parallel RAxML version 7.2.6 (Stamatakis 2006) with 100 bootstrap resamplings. As optimal models of sequence evolution we used the WAG+Γ or DCmut+Γ+F model (ZF domain proteins), the JTT+Γ+F model (BTB domain proteins), and the WAG+Γ+F model (vertebrate GAFs) as selected by ProtTest3 (Darriba et al. 2011).

Likelihood trees were visualized and arranged with FigTree (http://tree.bio.ed.ac.uk/software/figtree/) and then graphically edited with Adobe Illustrator software.

HMM-BASED CANDIDATE SEARCH AND MCL-CLUSTERING OF ORTHOLOGOUS GROUPS

To verify our results with an independent method, we combined a search on the basis of insulator-specific hidden Markov models (http://hmmer.org/) with the Markov cluster algorithm (van Dongen 2000). We downloaded from various sites (Table S3) genome and, if available, proteome and unplaced reads data of 19 different arthropod and seven outgroup species. The 26 selected species represent maximal diversity while avoiding over-representation of well-sampled groups such as dipterans. As subsequent comparisons relied on protein sequences, we translated the 26 genomes and the respective unplaced reads data into all six reading frames (>90 nucleotides) and combined the resulting open reading frames of each species with the corresponding protein set as offered by the sequencing center. To obtain specific HMMer profiles, we selected—with attention to maximal diversity and length—for each insulator protein 8–15 previously verified orthologs as representatives of a cluster. We then calculated and manually refined multiple alignments of these sequences using the MAFFT "einsi" algorithm (Katoh et al. 2005) and derived a representative full length HMMer profile for each insulator protein. All 26 ORF sets were scanned with the six custom made HMMer profiles and all sequences below the default inclusion threshold of HMMSEARCH (E-value < 0.01; 39,573 unique sequences) were fed to a dedicated OrthoMCL pipeline (Li et al. 2003) as described elsewhere (http://orthomcl.org/common/downloads/software/v2.0/UserGuide.txt). We used the recommended inflation parameter "1.5" for the MCL step. Before clustering, we removed duplicates and supplemented each of the 26 sequence collections with reference insulator sequences of that species (verified by phylogenetic experiments) to facilitate cluster recognition. Orthologous clusters that contained at least one reference sequence were analyzed further. For CTCF, Mod(mdg4), Zw5, and BEAF, we detected a single orthology group whereas Su(Hw) (2), CP190 (4), and GAF (3) orthologs split to more than one group.

HHpred-BASED SEARCH FOR GAGA FACTOR ORTHOLOGS

Like described above, we constructed two multiple sequence alignments from representative members of the GAGA cluster (full length and BTB domain only) and uploaded these alignments to HHpred (http://toolkit.tuebingen.mpg.de/hhpred) for highly sensitive profile–profile searches. We searched with default parameters in the proteomes of *H. sapiens* and *Mus musculus*, the only deuterostome datasets available. After removing duplicates, we analyzed the remaining 45 candidates with phylogenetic methods (see above).

ACKNOWLEDGMENTS

This research was supported by grants from the German Research Foundation to TW (DFG-SFB680). Phylogenies and parallel BLAST were computed on CHEOPS, the Cologne High Efficient Operating Platform for Science. The genome sequences of *Capitella teleta* and *Lottia gigantea* were produced by the U.S. Department of Energy Joint Genome Institute, http://www.jgi.doe.gov/, in collaboration with the user community. Preliminary sequence data for *Schmidtea mediterranea* were obtained from the Genome Institute at Washington University, St. Louis (http://genome.wustl.edu/pub/organism/Invertebrates/). Preliminary sequence data for *Strigamia maritima* were obtained from the Baylor College of Medicine Human Genome Sequencing Center website at http://www.hgsc.bcm.edu/. Preliminary sequence data for *Ixodes scapularis*, *Rhodnius prolixus*, and *Glossina morsitans* were obtained from http://www.vectorbase.org. We thank the *Daphnia* Genomics consortium for providing preliminary sequence data for *Daphnia magna*. These sequence data were produced by The Center for Genomics and Bioinformatics at Indiana University and distributed via wFleaBase in collaboration with the *Daphnia* Genomics Consortium http://daphnia.cgb.indiana.edu. We further wish to thank E. Schierenberg for support in the early stages of this work and B. Marin for helpful discussion.

LITERATURE CITED

Abzhanov, A., and T. C. Kaufman. 2000. Crustacean (malacostracan) Hox genes and the evolution of the arthropod trunk. Development 127:2239–2249.

Altenburg, E., and H. J. Muller. 1920. The genetic basis of truncate wing—an inconstant and modifiable character in Drosophila. Genetics 5:1–59.

Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389–3402.

Averof, M., and M. Akam. 1995. Hox genes and the diversification of insect and crustacean body plans. Nature 376:420–423.

Averof, M., and N. H. Patel. 1997. Crustacean appendage evolution associated with changes in Hox gene expression. Nature 388:682–686.

Bartkuhn, M., T. Straub, M. Herold, M. Herrmann, C. Rathke, H. Saumweber, G. D. Gilfillan, P. B. Becker, and R. Renkawitz. 2009. Active promoters and insulators are marked by the centrosomal protein 190. EMBO J. 28:877–888.

Belozerov, V. E., P. Majumder, P. Shen, and H. N. Cai. 2003. A novel boundary element may facilitate independent gene regulation in the Antennapedia complex of Drosophila. EMBO J. 22:3113–3121.

Bhat, K. M., G. Farkas, F. Karch, H. Gyurkovics, J. Gausz, and P. Schedl. 1996. The GAGA factor is required in the early Drosophila embryo not only for transcriptional regulation but also for nuclear division. Development 122:1113–1124.

Biggin, M. D., and R. Tjian. 1988. Transcription factors that activate the Ultrabithorax promoter in developmentally staged extracts. Cell 53:699–711.

Bouhouche, N., M. Syvanen, and C. I. Kado. 2000. The origin of prokaryotic C2H2 zinc finger regulators. Trends Microbiol. 8:77–81.

Buchner, K., P. Roth, G. Schotta, V. Krauss, H. Saumweber, G. Reuter, and R. Dorn. 2000. Genetic and molecular complexity of the position effect variegation modifier mod(mdg4) in Drosophila. Genetics 155:141–157.

Clarke, N. D., and J. M. Berg. 1998. Zinc fingers in *Caenorhabditis elegans*: finding families and probing pathways. Science 282:2018–2022.

Darriba, D., G. L. Taboada, R. Doallo, and D. Posada. 2011. ProtTest 3: fast selection of best-fit models of protein evolution. Bioinformatics 27:1164–1165.

Degnan, J. H., and N. A. Rosenberg. 2006. Discordance of species trees with their most likely gene trees. PLoS Genet. 2:e68.

Dorn, R., and V. Krauss. 2003. The modifier of mdg4 locus in Drosophila: functional complexity is resolved by trans splicing. Genetica 117:165–177.

Galtier, N., M. Gouy, and C. Gautier. 1996. SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. Comput. Appl. Biosci. 12:543–548.

Gerasimova, T. I., D. A. Gdula, D. V. Gerasimov, O. Simonova, and V. G. Corces. 1995. A Drosophila protein that imparts directionality on a chromatin insulator is an enhancer of position-effect variegation. Cell 82:587–597.

Gerasimova, T. I., E. P. Lei, A. M. Bushey, and V. G. Corces. 2007. Coordinated control of dCTCF and gypsy chromatin insulators in Drosophila. Mol. Cell 28:761–772.

Geyer, P. K., and V. G. Corces. 1992. DNA position-specific repression of transcription by a Drosophila zinc finger protein. Genes Dev. 6:1865–1873.

Gilbert, M. K., Y. Y. Tan, and C. M. Hart. 2006. The Drosophila boundary element-associated factors BEAF-32A and BEAF-32B affect chromatin structure. Genetics 173:1365–1375.

Heger, P., B. Marin, and E. Schierenberg. 2009. Loss of the insulator protein CTCF during nematode evolution. BMC Mol. Biol. 10:84.

Heger, P., B. Marin, M. Bartkuhn, E. Schierenberg, and T. Wiehe. 2012. The chromatin insulator CTCF and the emergence of metazoan diversity. Proc. Natl. Acad. Sci. USA 109:17507–17512.

Holohan, E. E., C. Kwong, B. Adryan, M. Bartkuhn, M. Herold, R. Renkawitz, S. Russell, and R. White. 2007. CTCF genomic binding sites in Drosophila and the organisation of the bithorax complex. PLoS Genet. 3:e112.

Hughes, C. L., and T. C. Kaufman. 2002. Exploring the myriapod body plan: expression patterns of the ten Hox genes in a centipede. Development 129:1225–1238.

Iampietro, C., M. Gummalla, A. Mutero, F. Karch, and R. K. Maeda. 2010. Initiator elements function to determine the activity state of BX-C enhancers. PLoS Genet. 6:e1001260.

Jeong, S., M. Rebeiz, P. Andolfatto, T. Werner, J. True, and S. B. Carroll. 2008. The evolution of gene regulation underlies a morphological difference between two Drosophila sister species. Cell 132:783–793.

Jiang, N., E. Emberly, O. Cuvier, and C. M. Hart. 2009. Genome-wide mapping of boundary element-associated factor (BEAF) binding sites in *Drosophila melanogaster* links BEAF to transcription. Mol. Cell Biol. 29:3556–3568.

Katoh, K., K. Kuma, H. Toh, and T. Miyata. 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res. 33:511–518.

Krauss, V., and R. Dorn. 2004. Evolution of the trans-splicing Drosophila locus mod(mdg4) in several species of Diptera and Lepidoptera. Gene 331:165–176.

Kumar, S. 2011. Remote homologue identification of Drosophila GAGA factor in mouse. Bioinformation 7:29–32.

Lewis, E. B. 1978. A gene complex controlling segmentation in Drosophila. Nature 276:565–570.

Lewis, D. L., M. A. DeCamillis, C. R. Brunetti, G. Halder, V. A. Kassner, J. E. Selegue, S. Higgs, and S. B. Carroll. 1999. Ectopic gene expression and homeotic transformations in arthropods using recombinant Sindbis viruses. Curr. Biol. 9:1279–1287.

Li, L., C. J. Stoeckert, Jr., and D. S. Roos. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 13:2178–2189.

Lieberman-Aiden, E., N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, et al. 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science 326:289–293.

Liubicich, D. M., J. M. Serano, A. Pavlopoulos, Z. Kontarakis, M. E. Protas, E. Kwan, S. Chatterjee, K. D. Tran, M. Averof, and N. H. Patel. 2009. Knockdown of Parhyale Ultrabithorax recapitulates evolutionary changes in crustacean appendage morphology. Proc. Natl. Acad. Sci. USA 106:13892–13896.

Matharu, N. K., T. Hussain, R. Sankaranarayanan, and R. K. Mishra. 2010. Vertebrate homologue of Drosophila GAGA factor. J. Mol. Biol. 400:434–447.

Melnikova, L., F. Juge, N. Gruzdeva, A. Mazur, G. Cavalli, and P. Georgiev. 2004. Interaction between the GAGA factor and Mod(mdg4) proteins promotes insulator bypass in Drosophila. Proc. Natl. Acad. Sci. USA 101:14806–14811.

Miele, A., and J. Dekker 2008. Long-range chromosomal interactions and gene regulation. Mol. Biosyst. 4:1046–1057.

Miele, V., S. Penel, and L. Duret. 2011. Ultra-fast sequence clustering from similarity networks with SiLiX. BMC Bioinformatics 12:116.

Mohan, M., M. Bartkuhn, M. Herold, A. Philippen, N. Heinl, I. Bardenhagen, J. Leers, R. A. White, R. Renkawitz-Pohl, H. Saumweber, et al. 2007. The Drosophila insulator proteins CTCF and CP190 link enhancer blocking to body patterning. EMBO J. 26:4203–4214.

Morgan, T. H. 1911. The origin of nine wing mutations in Drosophila. Science 33:496–499.

Negre, N., C. D. Brown, P. K. Shah, P. Kheradpour, C. A. Morrison, J. G. Henikoff, X. Feng, K. Ahmad, S. Russell, R. A. White, et al. 2010. A comprehensive map of insulator elements for the Drosophila genome. PLoS Genet. 6:e1000814.

Ni, X., Y. E. Zhang, N. Negre, S. Chen, M. Long, and K. P. White. 2012. Adaptive evolution and the birth of CTCF binding sites in the Drosophila genome. PLoS Biol. 10:e1001420.

Niehuis, O., G. Hartig, S. Grath, H. Pohl, J. Lehmann, H. Tafer, A. Donath, V. Krauss, C. Eisenhardt, J. Hertel, et al. 2012. Genomic and morphological evidence converge to resolve the enigma of Strepsiptera. Curr. Biol. 22:1309–1313.

Ohtsuki, S., and M. Levine. 1998. GAGA mediates the enhancer blocking activity of the eve promoter in the Drosophila embryo. Genes Dev. 12:3325–3330.

Oliver, D., B. Sheehan, H. South, O. Akbari, and C. Y. Pai. 2010. The chromosomal association/dissociation of the chromatin insulator protein Cp190 of *Drosophila melanogaster* is mediated by the BTB/POZ domain and two acidic regions. BMC Cell Biol. 11:101.

Pai, C. Y., E. P. Lei, D. Ghosh, and V. G. Corces. 2004. The centrosomal protein CP190 is a component of the gypsy chromatin insulator. Mol. Cell 16:737–748.

Pamilo, P., and M. Nei. 1998. Relationships between gene trees and species trees. Mol. Biol. Evol. 5:568–583.

Pavlopoulos, A., Z. Kontarakis, D. M. Liubicich, J. M. Serano, M. Akam, N. H. Patel, and M. Averof. 2009. Probing the evolution of appendage specialization by Hox gene misexpression in an emerging model crustacean. Proc. Natl. Acad. Sci. USA 106:13897–13902.

Perez-Torrado, R., D. Yamada, and P. A. Defossez. 2006. Born to bind: the BTB protein-protein interaction domain. Bioessays 28:1194–1202.

Peter, I. S., and E. H. Davidson. 2011. Evolution of gene regulatory networks controlling body plan development. Cell 144:970–985.

Pfeifer, M., F. Karch, and W. Bender. 1987. The bithorax complex: control of segmental identity. Genes Dev. 1:891–898.

Phillips, J. E., and V. G. Corces. 2009. CTCF: master weaver of the genome. Cell 137:1194–1211.

Regier, J. C., J. W. Shultz, A. Zwick, A. Hussey, B. Ball, R. Wetzer, J. W. Martin, and C. W. Cunningham. 2010. Arthropod relationships revealed by phylogenomic analysis of nuclear protein-coding sequences. Nature 463:1079–1083.

Rice, P., I. Longden, and A. Bleasby. 2000. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. 16:276–277.

Roy, S., N. Jiang, and C. M. Hart. 2011. Lack of the Drosophila BEAF insulator proteins alters regulation of genes in the Antennapedia complex. Mol. Genet. Genomics 285:113–123.

Savard, J., D. Tautz, and M. J. Lercher. 2006. Genome-wide acceleration of protein evolution in flies (Diptera). BMC Evol. Biol. 6:7.

Schmidt, D., P. C. Schwalie, M. D. Wilson, B. Ballester, A. Goncalves, C. Kutter, G. D. Brown, A. Marshall, P. Flicek, and D. T. Odom. 2012. Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. Cell 148:335–348.

Schoborg, T. A., and M. Labrador. 2010. The phylogenetic distribution of non-CTCF insulator proteins is limited to insects and reveals that BEAF-32 is Drosophila lineage specific. J. Mol. Evol. 70:74–84.

Shao, W., Q. Y. Zhao, X. Y. Wang, X. Y. Xu, Q. Tang, M. Li, X. Li, and Y. Z. Xu. 2012. Alternative splicing and trans-splicing events revealed by analysis of the *Bombyx mori* transcriptome. RNA. 18:1395–1407.

Sievers, F., A. Wilm, D. Dineen, T. J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Soding, et al. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol. Syst. Biol. 7:539.

Simon, S., S. Strauss, A. von Haeseler, and H. Hadrys. 2009. A phylogenomic approach to resolve the basal pterygote divergence. Mol. Biol. Evol. 26:2719–2730.

Spana, C., D. A. Harrison, and V. G. Corces. 1988. The *Drosophila melanogaster* suppressor of Hairy-wing protein binds to specific sequences of the gypsy retrotransposon. Genes Dev. 2:1414–1423.

Stamatakis, A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22:2688–2690.

Tadepally, H. D., G. Burger, and M. Aubry. 2008. Evolution of C2H2-zinc finger genes and subfamilies in mammals: species-specific duplication and loss of clusters, genes and effector domains. BMC Evol. Biol. 8:176.

Trautwein, M. D., B. M. Wiegmann, R. Beutel, K. M. Kjer, and D. K. Yeates. 2012. Advances in insect phylogeny at the dawn of the postgenomic era. Annu. Rev. Entomol. 57:449–468.

Van Bortle, K., and V. G. Corces. 2013. The role of chromatin insulators in nuclear architecture and genome function. Curr. Opin. Genet. Dev 23:212–218.

Van Bortle, K., E. Ramos, N. Takenaka, J. Yang, J. E. Wahi, and V. G. Corces. 2012. Drosophila CTCF tandemly aligns with other insulator proteins at the borders of H3K27me3 domains. Genome Res. 22:2176–2187.

van Dongen, S. 2000. Graph clustering by flow simulation. Ph.D. thesis, University of Utrecht, Utrecht, Netherlands.

Wallace, J. A., and G. Felsenfeld. 2007. We gather together: insulators and genome organization. Curr. Opin. Genet. Dev. 17:400–407.

Warren, R. W., L. Nagy, J. Selegue, J. Gates, and S. Carroll. 1994. Evolution of homeotic gene regulation and function in flies and butterflies. Nature 372:458–461.

Wiegmann, B. M., M. D. Trautwein, J. W. Kim, B. K. Cassel, M. A. Bertone, S. L. Winterton, and D. K. Yeates. 2009. Single-copy nuclear genes resolve the phylogeny of the holometabolous insects. BMC Biol. 7: 34.

Yajima, M., W. G. Fairbrother, and G. M. Wessel. 2012. ISWI contributes to ArsI insulator function in development of the sea urchin. Development 139:3613–3622.

Yang, J., and V. G. Corces. 2012. Insulators, long-range interactions, and genome function. Curr. Opin. Genet. Dev. 22:86–92.

———. 2011. Chromatin insulators: a role in nuclear organization and gene expression. Adv. Cancer Res. 110:43–76.

Yang, J., E. Ramos, and V. G. Corces. 2012. The BEAF-32 insulator coordinates genome organization and function during the evolution of Drosophila species. Genome Res. 22:2199–2207.

Zhao, K., C. M. Hart, and U. K. Laemmli. 1995. Visualization of chromosomal domains with boundary element-associated factor BEAF-32. Cell 81:879–889.

Zollman, S., D. Godt, G. G. Prive, J. L. Couderc, and F. A. Laski. 1994. The BTB domain, found primarily in zinc finger proteins, defines an evolutionarily conserved family that includes several developmentally regulated genes in Drosophila. Proc. Natl. Acad. Sci. USA 91:10717–10721.

Associate Editor: E. Abouheif

Supporting Information

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

Figure S1. Phylogeny of ZF domain insulator proteins.
Figure S2. Phylogeny of BTB domain insulator proteins.
Figure S3. Absence of GAGA factor orthologs from vertebrates I.
Figure S4. Absence of GAGA factor orthologs from vertebrates II.
File S1. Multiple sequence alignment of ZF domain proteins.
File S2. Multiple sequence alignment of BTB domain proteins.
File S3. Multiple sequence alignment of vertebrate GAGA factor candidates (BLAST).
File S4. Multiple sequence alignment of vertebrate GAGA factor candidates (HMM).
Table S1. The top 40 arthropod species providing ZF candidate sequences.
Table S2. The top 40 arthropod species providing BTB candidate sequences.
Table S3. Download location of 26 genomes and proteomes used for a HMM/OrthoMCL-based search for insulator proteins.
Table S4. Summary of the HMM/OrthoMCL-based search for insulator proteins.