BMC Genomics

RESEARCH ARTICLE Open Access

Transcriptomic data from panarthropods shed new light on the evolution of insulator binding proteins in insects

Insect insulator proteins

Thomas Pauli[1*], Lucia Vedder[2], Daniel Dowling[3], Malte Petersen[1], Karen Meusemann[1,4,5], Alexander Donath[1], Ralph S. Peters[6], Lars Podsiadlowski[7], Christoph Mayer[1], Shanlin Liu[8,9], Xin Zhou[10,11], Peter Heger[12], Thomas Wiehe[12], Lars Hering[13], Georg Mayer[13], Bernhard Misof[1] and Oliver Niehuis[1*]

Abstract

Background: Body plan development in multi-cellular organisms is largely determined by homeotic genes. Expression of homeotic genes, in turn, is partially regulated by insulator binding proteins (IBPs). While only a few enhancer blocking IBPs have been identified in vertebrates, the common fruit fly *Drosophila melanogaster* harbors at least twelve different enhancer blocking IBPs. We screened recently compiled insect transcriptomes from the 1KITE project and genomic and transcriptomic data from public databases, aiming to trace the origin of IBPs in insects and other arthropods.

Results: Our study shows that the last common ancestor of insects (Hexapoda) already possessed a substantial number of IBPs. Specifically, of the known twelve insect IBPs, at least three (*i.e.*, CP190, Su(Hw), and CTCF) already existed prior to the evolution of insects. Furthermore we found GAF orthologs in early branching insect orders, including Zygentoma (silverfish and firebrats) and Diplura (two-pronged bristletails). Mod(mdg4) is most likely a derived feature of Neoptera, while Pita is likely an evolutionary novelty of holometabolous insects. Zw5 appears to be restricted to schizophoran flies, whereas BEAF-32, ZIPIC and the Elba complex, are probably unique to the genus *Drosophila*. Selection models indicate that insect IBPs evolved under neutral or purifying selection.

Conclusions: Our results suggest that a substantial number of IBPs either pre-date the evolution of insects or evolved early during insect evolution. This suggests an evolutionary history of insulator binding proteins in insects different to that previously thought. Moreover, our study demonstrates the versatility of the 1KITE transcriptomic data for comparative analyses in insects and other arthropods.

Keywords: Insulator binding proteins, Comparative transcriptomic analyses, Gene evolution, Arthropod evolution

Background

Chromatin insulation accounts for the formation of independent transcriptional units on eukaryote chromosomes [1–3]. Chromatin insulation is mediated by insulator binding proteins (IBPs), which insulate transcriptional units either by acting as chromatin barriers (preventing the formation of heterochromatin and thus the silencing of active genes) or as enhancer blockers (preventing enhancers from binding to off-target promoters). Due to their large-scale effects on transcription and on the regulation of fundamental developmental processes, IBPs can significantly impact body plan formation [4–6]. Consequently, IBPs may play an important role in the evolution of body plans and biological diversity. Following this line of reasoning, studying the evolution of IBPs in insects[1] appears rewarding. In the common fruit fly, *Drosophila melanogaster*, twelve different IBPs have been identified (Table 1). However,

* Correspondence: s6thpaul@uni-bonn.de; oliver.niehuis@gmail.com
[1] Center of Molecular Biodiversity Research, Zoological Research Museum Alexander Koenig, Adenauerallee 160, 53113 Bonn, Germany
Full list of author information is available at the end of the article

Table 1 Summary of all currently known insulator binding proteins (IBPs) in *Drosophila melanogaster*, with information on the Pfam symbol of the conserved protein domain families found in the respective proteins with the corresponding references

Insulator binding protein	Conserved domains	Reference
CTCF	zf-C2H2 [11]	[24]
Su(Hw)	zf-C2H2 [12]	[22, 23]
Pita	zf-AD [1], zf-C2H2 [10]	[43]
ZIPIC	zf-C2H2 [7]	[43]
Zw5	zf-C2H2 [8]	[67]
CP190	BTB [1], zf-C2H2 [4]	[32, 68]
GAF	BTB [1], GAGA [1]	[69, 70]
Mod (mdg4)	BTB [1], FLYWCH [1]	[71, 72]
BEAF-32	zf-BED [1], BESS [1]	[34]
Ibf1	zf-BED [1]	[44]
Ibf2	zf-BED [1]	[44]
Elba-complex (Elba 1,2,3)	BEN [1] (Elba 1,2), none (Elba 3)	[30]

The number of repeats of each conserved domain in the respective protein is given in square brackets

the taxonomic distribution of IBPs in insects and the IBPs' possible correlation with biological diversity has only been studied in a small number of species [7, 8]. In the present investigation, we therefore exploit information in recently published transcriptome and genome sequence data to trace the evolution of IBPs in insects and show that the evolution of IBPs in 100 insect species is more complex than previously anticipated.

Transcriptional units comprise groups of genes and associated regulatory elements, such as enhancers, silencers, and promoters, that can be brought into close spatial proximity to each other by folding of chromatin fibers [9]. It has been shown that transcriptionally active units can be immediately adjacent to inactive genomic regions [10]. Such a spatial arrangement can result in inadvertent genic interactions. Experiments show that IBPs are capable of effectively impeding such interactions [11, 12]. In *D. melanogaster*, the protein Cut acts as a chromatin barrier insulator, like the homologous protein CDP of humans that binds to a similar target region [13]. As chromatin barriers, Cut and CDP inhibit interactions between heterochromatin and actively transcribed euchromatin [14]. In general, when heterochromatin comes into spatial proximity of transcribed euchromatin, it can spread along the chromatin fiber into adjacent euchromatin regions and repress transcription. Chromatin barrier IBPs seem to be ancient proteins in eukaryotes since it has also been demonstrated by the interaction between TFIIIC and tRNA genes found in yeast and humans [15–18]. The taxonomically wide distribution of chromatin-barring IBPs (*e.g.,* Cut in *D. melanogaster* and CDP and TFIIIC in humans and yeast) implies that chromatin barring is essential for chromosomal organization in eukaryotes [19].

Enhancer blocking IBPs apparently evolved later than chromatin barrier IBPs and are possibly restricted to bilaterians [20]. Enhancers are regulatory elements that can bind to a promoter and thereby enhance transcription of the associated gene. The switch between a euchromatic and a heterochromatic state of adjacent chromosome regions can result in unfavorable alignments of enhancers in spatial proximity of otherwise distant promoters. Consequently, enhancers could interact with off-target promoters. Such interactions can be prevented by enhancer-blocking IBPs [21]. Su(Hw) (suppressor of hairy wing) was the first enhancer blocker to be functionally characterized in *D. melanogaster.* Su(Hw) was discovered due to its ability to protect DNA of transgenic flies from the phenotypic effect of the transposable element *gypsy*, which induces mutations affecting transcription by inserting itself into splice sites and sequences necessary for initiating transcription [22, 23]. Su(Hw) seems to be restricted to arthropods [7, 8]. Bell and colleagues [24] described a second enhancer blocker, called CTCF (CCCTC binding factor), in birds and mammals. In contrast to Su(Hw), CTCF was shown to be taxonomically widespread and has been found in all bilaterian lineages studied [7, 20].

As of yet CTCF is the only enhancer-blocking IBP known in vertebrates. However, B1 and B2 type *SINEs* (Short Interspersed Nuclear Elements), which are transposable elements, can also encode for enhancer blocking peptides [25, 26]. Additionally, tRNA genes have been shown to exhibit enhancer-blocking or chromatin barring properties [18, 27]. Furthermore, a homolog of the GAGA factor (GAF) has been identified in vertebrates, where it might function as an enhancer blocking IBP [28]. So far, twelve IBPs with enhancer-blocking properties have been identified in *D. melanogaster,* including CTCF and Su(Hw) (Table 1). All IBPs contain DNA-binding domains. The most common are zinc-finger domains, or domains with a zinc-finger core, such as zf-C2H2, zf-BED, GAGA and FLYWCH. The Elba (Early boundary activity) protein complex and a specific isoform of Mod(mdg4) (modifier of mdg4) use BEN domains to bind DNA instead [29, 30]. Three IBPs, CP190 (Centrosomal protein 190 kD), GAF, and Mod(mdg4), additionally have a BTB domain (bric-a-brac, ttk and broad complex), which is assumed to mediate DNA binding and protein binding [31]. Mod(mdg4) and CP190 often interact with CTCF [5] and Su(Hw) [32] and are shown to form complexes in *D. melanogaster*. These interactions might possibly be mediated through the BTB domain. Other domains are a zf-AD (zinc-finger associated domain) found in Pita and a BESS domain (named after the three proteins in which it was found: BEAF-32 (Boundary element associated factor of 32 kD), Suvar(3)7, and Stonewall [33–35]) found in BEAF-32.

In *D. melanogaster*, IBPs exhibiting enhancer-blocking function actively regulate larval development. For example, individual deletion of *CTCF*, *CP190*, *BEAF-32*, and *GAF* alters the expression of hox genes, resulting in lethal homeotic transformations [4–6]. Deletion of *Su(Hw)* induces sterility in female *D. melanogaster* due to changes in the expression of oogenesis-related genes [36]. These experiments demonstrate the importance of IBP-mediated transcriptional regulation for proper larval development and oogenesis in *D. melanogaster* and raise the intriguing question of when and how these important IBPs evolved in arthropods.

Schoborg and Labrador [7] as well as Heger and colleagues [8, 20] screened publicly available transcriptomes as well as draft genomes of insects for genes orthologous to *D. melanogaster* IBPs. They inferred that *CTCF* likely evolved in the stem lineage of Bilateria. *Su(Hw)* possibly evolved in the stem lineage of arthropods and *CP190* possibly evolved in the stem lineage of the Pancrustacea (insects plus crustaceans). The IBP *GAF* likely evolved in the last common ancestor of Holometabola and Hemiptera, and Mod(mdg4) likely emerged in the last common ancestor of Aparaglossata (all holometabolan insects except Hymenoptera, see [37]). Finally, *Zw5* and *BEAF-32* are possibly unique to the dipteran family Drosophilidae. Because *GAF* and *Mod(mdg4)* apparently emerged during the diversification of Holometabola, we suggest that IBPs may have played a key role for the tremendous diversification of holometabolous insects.

We therefore analyzed whole-body transcriptomes sampled across all described insect orders, which were compiled in the international 1KITE project [38]. We additionally considered sequence data of other panarthropod lineages, including RNAseq data of onychophorans and a tardigrade. Additionally, we screened the genome of a nematode (*Trichinella spiralis*). We screened for all twelve enhancer-blocking IBPs that have previously been identified in insects (Hexapoda). We assessed the orthology of all identified candidate transcripts of IBPs by using the best reciprocal hit criterion, inferred the phylogeny of each gene from the assembled transcripts and studied selective forces that might have acted on these genes. Our data and results furthermore set the stage for future comparative and experimental studies on this intriguing group of proteins.

Results

We used profile Hidden Markov Models (pHMMs) in order to search for orthologous sequences of twelve enhancer-blocking IBPs known from *D. melanogaster* in transcriptome data sets from 100 insect species and in transcriptomes and genomes of ten outgroup species, including crustaceans, chelicerates, myriapods, onychophorans (velvet worms), a tardigrade, and a nematode (Fig. 1). We found that three IBPs are particularly widespread across insect orders and outgroups: (i) *CTCF* was found in the transcript libraries of 105 species, including the nematode, *Trichinella spiralis*; (ii) *Su(Hw)* occurs in the transcript libraries of 86 species, including crustaceans, chelicerates, and myriapods (iii) *CP190* was found in the transcript libraries of 81 species, including crustaceans. Ancestral state reconstruction corroborates the idea that *CTCF* was already present in the last common ancestor of Panarthropoda (Onychophora + Tardigrada + Arthropoda; Additional file 1: Figure S1), Su(Hw) was already present in the last common ancestor of Arthropoda (Additional file 1: Figure S2), and *CP190* in the last common ancestor of Pancrustacea (Additional file 1: Figure S3).

In contrast, we detected *GAF* exclusively in insects, including coneheads (Protura), but not in all species studied. In fact, only 38 screened insect transcriptome assemblies included putative transcripts of *GAF*. We did not find any *GAF* transcripts in the screened transcriptomes of butterflies and moths (Lepidoptera), caddisflies (Trichoptera), scorpionflies (Mecoptera), fleas (Siphonaptera), and springtails (Collembola). In addition, we did not find *GAF* in the draft genomes of *Bombyx mori* (Lepidoptera), *Limnephilus lunatus* (Trichoptera), *Machilis hrabei* (Archaeognatha), and *Catajapyx aquilonaris* (Diplura). Ancestral state reconstruction for GAF reveals multiple losses of this protein (Additional file 1: Figure S4). A search for the vertebrate *GAF* homolog in the insect transcriptomes yielded several positive hits, which however did not fulfill the best reciprocal hit criterion.

Transcripts of *Mod(mdg4)* were exclusively detected in species of neopteran insects (*i.e.*, insects with the ability to flex their wings above their abdomen; 57 species of all extant neopteran insect orders, except for ground lice, Zoraptera, and earwigs, Dermaptera). We also searched an early draft genome of a bristletail (*Machilis hrabei*; Archaeognatha), a mayfly (*Ephemera danica*; Ephemeroptera), and a dragonfly (*Ladona fulva*; Odonata) for possible orthologs of *Mod(mdg4)*. We identified a FLYWCH zinc finger domain (domain orthology was confirmed by the best reciprocal hit criterion; see the Methods section) when searching the *M. hrabei* genome. However, since other proteins, such as Su(Kpn) (Suppressor of Killer of prune) [39], are known to also contain FLYWCH domains, we deem these hits as insufficient evidence for the occurrence of *Mod(mdg4)* in bristletails.

We found orthologs of *Pita* only in transcript assemblies of holometabolous insects (30 species, covering 11 orders), and ancestral state reconstruction of *Pita* suggests that this IBP was present in the last common ancestor of Holometabola (Additional file 1: Figure S5).

We identified transcripts encoding the IBP *Zw5* only in two species of Diptera (*i.e.*, *Lipara lucens* and *Triarthria setipennis*).

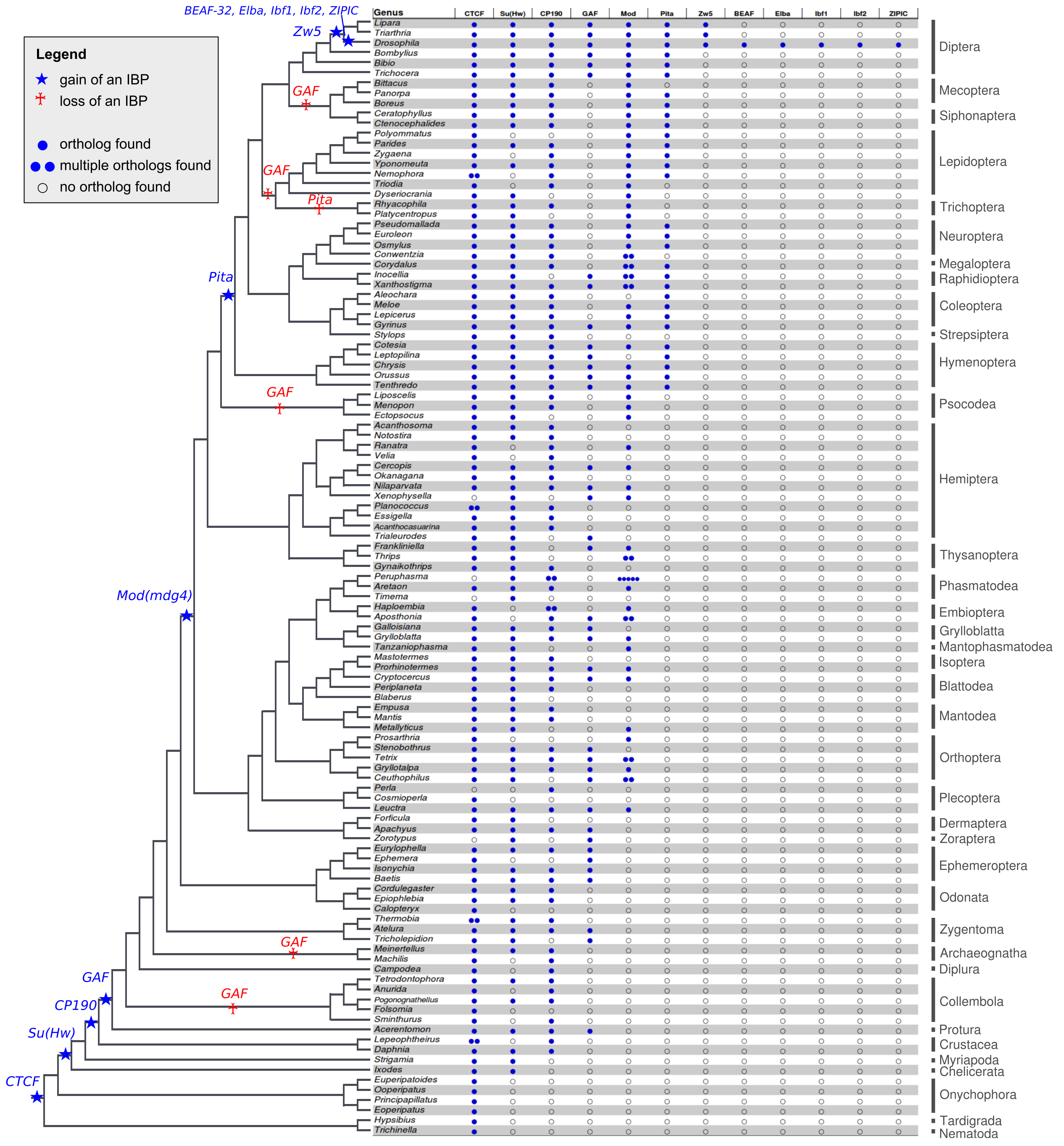

Fig. 1 Evolution of enhancer-blocking insulator proteins (IBP) in arthropods. Gains of IBPs are indicated by blue stars, potential losses by red crosses. Blue circles in the table indicate the presence of transcripts of a given IBP in transcriptomes. Multiple blue circles indicate the number of copies found. Transparent circles indicate the putative absence of transcripts of an IBP. The backbone tree topology is adopted from [38]. The phylogenetic relationships among outgroup taxa (*i.e.*, Onychophora, Tardigrada, and Nematoda) are according to [62]

We could not find evidence for the presence of orthologs of *ZIPIC* (zinc-finger protein interacting with CP190), *BEAF-32*, *Ibf1* (Insulator binding factor 1), *Ibf2*, (Insulator binding factor 1) and the genes encoding the Elba complex in any of the investigated species when searching all available transcriptomes. We did find such evidence, however, in the genome of *D. willistoni* (Drosophilidae). Note that *Ibf1*, *Ibf2*, *ZIPIC*, *BEAF-32*, and the proteins of the Elba complex have only been identified in Drosophila to date.

Finally, we conducted a branch-specific analysis of d_N/d_S-ratios to test for positive selective pressure (Table 2).

Table 2 Results from analyzing d_N/d_S ratios in genes encoding insulator proteins in insects

Gene	Branch	lnL0	lnL1	LRT	*p*-value
CP190	Crustacea	−5495.527	−5495.388	0.278	0.598
CP190	Holometabola	−5495.284	−5495.260	0.047	0.828
CTCF	Onychophora	−1314.281	−1310.626	7.308	0.007
CTCF	Holometabola	−1311.997	−1312.166	0.338	0.561
GAF	Acerentomon	−2006.810	−2006.810	0.0	1.000
GAF	Holometabola	−2006.810	−2006.810	0.0	1.000
Mod (mdg4)	Polyneoptera	−15377.060	−15377.060	4.000 10–6	0.998
Mod (mdg4)	Holometabola	−15374.903	−15373.888	2.032	0.154
Pita	Hymenoptera	−1054.403	−1046.840	15.13	<0.001*
Su (Hw)	Holometabola	−11052.401	−11052.210	0.383	0.536

Shown are the gene name and the branch, along which the respective selection model was tested, the log-likelihood for the neutral model (lnL0) and for positive selection (lnL1), the likelihood ratio test statistic (LRT), and the associated *p*-value. Branches on which the positive selection model fits significantly better than the neutral selection model are indicated by *. Bonferroni corrected significance threshold was $\alpha = 0.005$. The degree of freedom (df) was 1 for all tests

We found no statistically significant evidence for positive selection in *CTCF* in Onychophora ($p = 0.007$; Bonferroni corrected $\alpha = 0.005$). *Pita* showed evidence for positive selection in Hymenoptera ($p < 0.001$; Bonferroni corrected $\alpha = 0.005$).

Completeness of the transcriptomes was assessed by using the BUSCO (Benchmarking Universal Single-Copy Orthologs) pipeline [40]. The transcriptome completeness ranges from 15.2 % (*Bittacus pilicornis,* Mecoptera) to 81.2 % (*Lipara lucens*, Diptera). Results of the analysis are summarised in Table 3, absolute values for all used 1KITE transcriptomes can be found in Additional file 2: Table S1.

None of the phylogenetic analyses of the transcripts of the above genes and proteins provided evidence for gene duplication events (Additional file 1: Figures S8–S14).

Discussion

We traced the evolutionary origin of all twelve enhancer-blocking insulator proteins (IBPs) known from *D. melanogaster.* We searched for transcripts of these IBPs in 110 different species of panarthropods by applying profile hidden Markov models (pHMMs) and the best reciprocal hit criterion. This procedure proved necessary to account for the fact that some IBPs are comprised of multiple zinc finger domains. These domains are found in various chromatin binding proteins [41, 42] and are not specific to IBPs.

Table 3 BUSCO assessment for completeness of the 100 1KITE transcriptomes

	Complete [%]	Fragmented [%]	Missing [%]
Min	15.3	3.8	14.7
1st Qu.	49.0	9.3	22.4
Median	57.9	11.0	30.7
Mean	57.3	11.0	31.8
3rd Qu.	68.6	12.5	37.9
Max	81.2	19.0	72.5

Given are the proportions of complete, fragmented and missing BUSCO genes

Since our pHMMs were constructed from IBP amino acid sequences of primarily dipteran species, we can expect a taxonomic bias in the analysis. However, this caveat was unavoidable, since many of these proteins have not been detected in other insect species yet.

Since the IBP CTCF is expected to occur in all Bilateria, we used it to assess the sensitivity of our search strategy and the quality of the analyzed transcript libraries. As expected, we identified transcripts of *CTCF* in almost all analyzed transcript assemblies, confirming the ubiquitous occurrence of this IBP in arthropods. We also found the zinc finger protein Su(Hw) in all major investigated arthropod lineages. Ancestral state reconstruction suggests that Su(Hw) evolved in the last common ancestor of Euarthropoda. We further inferred that the BTB domain protein CP190 evolved either in the last common ancestor, or during the early radiation of Pancrustacea. Consequently, the sequences encoding for CTCF, Su(Hw), and CP190 must have been part of the ancestral gene repertoire of insects, which is in accordance with the current knowledge on the evolution of IBPs [8].

The BTB domain protein GAF was assumed to be unique to holometabolous insects and Hemiptera and was lost secondarily in moths and butterflies [8]. In contrast, we recovered GAF orthologs in nearly all insect orders, except for moths and butterflies (Lepidoptera), caddisflies (Trichoptera), scorpionflies (Mecoptera), fleas (Siphonaptera), twisted wing parasites (Strepsiptera), bark lice and true lice (Psocodea), two-pronged bristletails (Diplura), jumping bristletails (Archaeognatha) and springtails (Collembola). Thus, this pattern suggests that GAF most likely evolved in the last common ancestor of insects and was secondarily lost in some insect lineages. Since GAF was found to play an important role in early embryonic development of *D. melanogaster* [4], it is possible that its expression is down-regulated in adult individuals of the above lineages (*i.e.,* Lepidoptera, Trichoptera, Mecoptera, Siphonaptera, and Collembola). However, we confirmed the absence of *GAF* in the publicly available draft genome assemblies of *B. mori (Lepidoptera), L. lunatus (Trichoptera), M. hrabei (Archaeognatha),* and *C. aquilonaris (Diplura)* (see Fig. 1). Therefore the absence of *GAF* in the transcriptomes of the aforementioned insect orders corroborates the likely secondary loss of *GAF* in these insect orders. The IBP GAF must have evolved during the Ordovician (509–452 million years ago (mya); [38]), between 106–220 million years earlier than previously thought [8]. While ancestral state reconstruction inferred separate gains of GAF within insects, we deem this scenario highly unlikely. We furthermore investigated the transcriptomes for the vertebrate GAF sequence, but were unable to infer an orthologous relationship between the best hits in insects and the vertebrate sequences.

The occurrence of the zinc finger protein Pita in holometabolous insects, previously only known from *D. melanogaster,* suggests that it was already present in the last common ancestor of Holometabola. Since Pita has previously been investigated only in Diptera [43], our data represent the first evidence for a much older evolutionary origin (Carboniferous, 372–317 mya) and a wider taxonomic distribution of this gene in insects.

Mod(mdg4) is another example of an IBP that shows a much wider taxonomic distribution than previously thought. The data available to Heger and colleagues [8] led the authors to the conclusion that *Mod(mdg4)* likely evolved in the last common ancestor of Aparaglossata (all Holometabola, excluding Hymenoptera). The presence of *Mod(mdg4)* transcripts in various polyneopteran insect lineages suggests, however, that *Mod(mdg4)* must have evolved in the stem lineage of Neoptera (see Fig. 1), whose origin was in the Devonian (413–360 mya) [38]. The occurrence of the FLYWCH domain in sections of coding sequences in the early draft genome of the bristletail *M. hrabei* (Archaeognatha) suggests that *Mod(mdg4)* might have evolved even earlier, within primarily apterygote insects. However, the presence of the FLYWCH domain alone is insufficient to draw solid conclusions, as the domain has also been found in other proteins, such as Su(Kpn) [39].

While most previously discussed IBPs, except for Pita, have already been found in species other than *D. melanogaster,* Zw5 and the proteins discussed in the following section are only known from *D. melanogaster* [7, 8, 43, 44]. Our search for Zw5 in the 1KITE data revealed orthologous transcripts in two additional species of Diptera, *Lipara lucens* (Chloropidae) and *Triarthria setipennis* (Tachinidae). Both belong to the lineage Schizophora, which uses an eversible front pouch to escape from their puparium. This lineage comprises one-third of all extant dipteran species, including those of the genus *Drosophila.* Schizophora diverged from the remaining Diptera in the early Tertiary (65–40 mya; [45]). This distribution is in accordance with the results obtained by Heger and colleagues [8], who found *Zw5* already in another schizophoran fly, *Glossina morsitans.* When searching for *Zw5* transcripts in the 1KITE transcriptome assemblies, we consistently received also transcripts of the protein "meiotic central spindle" (Meics) as promising hits. Both proteins share a similar domain configuration, with Zw5 differing from Meics by having one fewer zinc finger domain. This led us to speculate that *Zw5* could be a paralog of the *meics* gene that evolved within Diptera. We tested this hypothesis by inferring a gene tree from amino acid sequences of Zw5 and Meics, including representatives of Diptera and holometabolous insects. However, in the inferred gene tree (see Additional file 1: Figure S15), Zw5 does not group with the Meics protein subtree. We therefore conclude that *Zw5* is unlikely to be the result of a duplication of *meics* in Diptera.

The IBPs BEAF-32, ZIPIC, Ibf1, Ibf2 as well as the proteins of the Elba protein complex are known only from *D. melanogaster.* We were unable to identify transcripts of these IBPs in any of the analyzed transcriptomes. Since BEAF-32 contains the BESS domain only known from *Drosophila* [33–35], chances of finding the gene in non-dipterans seem to be low, and previous reports already concluded that BEAF-32 is likely being restricted to species of the genus Drosophila [7, 8]. Elba1 and Elba2 of the tripartite protein complex Elba, each contain a chromatin-binding BEN domain, which is known to occur in invertebrates, vertebrates, and viral proteins [29]. In *D. melanogaster,* expression of genes of the Elba complex is restricted to embryonic development [30]. Thus, the transcriptomes from the 1KITE project, which primarily represent tissue samples from adult insects, may be unsuitable to trace back the evolution of this gene, since they do not cover the appropriate developmental stages. The same might hold true for the zinc finger IBPs ZIPIC, Ibf1, and Ibf2, since our searches for the corresponding coding sequences in the draft genomes of *D. willistoni, Aedes aegypti* and *Anopheles gambiae* (Diptera) only revealed significant hits in *D. wilistoni.* This finding corroborates the idea that the absence of transcripts of these IBPs in the screened 1KITE transcriptomes indeed reflects the actual distribution of these proteins in insect transcriptomes.

We found possible evidence for positive selection in the genes encoding for CTCF and Pita. *CTCF* was seemingly underlying positive selection in the onychophoran branch. This might be an artifact of the d_N/d_S.ratio test however. Long divergence times lead to a saturation of d_S [46, 47]. This results in an increase of ω (*i.e.* the ratio of the nonsynonymous substitution rate and the synonymous substitution rate), which means that positive selection is more likely to be erroneously detected, as could be the case for *CTCF,* for which we analyzed sequence data spanning the entire range of Arthropoda. Evidence for positive selection in Pita corresponds with the branch lengths in the Pita gene tree (Additional file 1: Figure S5) and suggests that the gene is rapidly evolving. Identification of Pita orthologs consequently proved to be difficult. This opens the possibility that the gene could have evolved even earlier and occurs also in hemimetabolous insects. We might have been unable to identify it properly due to its high amino acid sequence divergence.

The occurrence of IBPs in a wide range of species, or restricted to particular taxa, may provide clues about evolutionarily conserved and evolutionarily labile autonomous transcriptional units. Both phylogenetically older and younger IBPs have been shown to actively insulate regions of the same gene complex. The bithorax complex in *D. melanogaster,* for example, contains binding sites of CTCF, GAF and also of Elba [30, 48]. It is possible that the presence of CTCF, Su(Hw), CP190, and GAF across insects most likely ensures proper transcription of genes in rather conserved units and regions (*e.g.,* genes that share an evolutionary conserved gene neighborhood and/or that are in close spatial proximity to, at least temporarily, heterochromatic regions). Likewise, we hypothesize that the restricted occurrence of Mod(mdg4), Pita and, in particular, of Zw5, BEAF-32, ZIPIC and the Elba complex may be the result of recent evolutionary changes in the architecture or transcription of genomic regions in the respective insect lineages.

Conclusions

The exceptionally broad taxonomic sampling of whole-body transcriptomes and the sequencing depth of the analyzed transcriptomes of insects from the 1KITE project proved to be useful for screening and delineating the occurrence of IBPs in arthropods. Our search for and identification of IBPs in all currently recognized extant insect orders implies that the enhancer-blocking IBPs CTCF, Su(Hw), CP190, and GAF were already present in the last common ancestor of insects. The evolution of two insect-specific IBPs is associated with the origin of two major insect lineages: Mod(mdg4) with evolution of Neoptera (413-360 mya) and Pita with the evolution of Holometabola (372-317 mya). Finally, the IBPs Zw5, BEAF-32, and ZIPIC as well as the IBPs of the Elba complex are apparently restricted to Diptera, with BEAF-32, ZIPIC, and Elba possibly being unique to drosophilids. Considering the likely fundamental importance of IBPs for maintaining proper transcription of genes in a frequently altering genomic environment, the currently known diversity of IBPs in *D. melanogaster* likely still represents a lower estimate of the actual diversity of IBPs in flies. The large number of IBPs that are seemingly unique to drosophilids furthermore implies that, if IBP diversity in drosophilids is representative for a given insect lineage with a given age, a plethora of IBPs is yet to be discovered in other insect lineages.

Methods

Transcript libraries and draft genomes

We screened the transcriptomic assemblies of 100 insect (Hexapoda) species sequenced by Misof and colleagues [38] in the 1KITE project for potential transcripts orthologous to IBP genes known from *D. melanogaster* (accession and version numbers are provided in Additional file 3: Table S2). The 100 analyzed species comprise all currently recognized insect orders. We also studied sequence data of species previously analyzed by Heger and colleagues [8]: two crustaceans (*Daphnia pulex* and *Lepeophtheirus salmonis*), one myriapod (*Strigamia maritima*), one chelicerate (*Ixodes scapularis*), and one nematode (*Trichinella spiralis*). We furthermore analyzed the transcript sequences of one tardigrade (*Hypsibius dujardini*) [49], and four species of onychophorans (*Euperipatoides rowelli*, *Ooperipatus hispidus*, *Principapillatus hitoyensis*, and *Eoperipatus* sp.) [50]. We additionally screened genomes of the following species for IBP-coding genes (see Additional file 2: Table S1 for accession numbers): *Drosophila wilistoni* [51], *Aedes aegypti* [52], *Anopheles gambiae* (Diptera) [53], *Bombyx mori* (Lepidoptera) [54], *Limnephilus lunatus* (Trichoptera), *Machilis hrabei* (Archaeognatha), *Catajapyx aquilonaris* (Diplura), *Ephemera danica* (Ephemeroptera), and *Ladona fulva* (Odonata) [55].

Identification of insulator proteins (IBPs)

We searched the transcriptome assemblies for IBP candidate transcripts using profile hidden Markov models (pHMMs) specific to each IBP. The pHMMs were obtained by first aligning all published amino acid sequences that are orthologous to a given *D. melanogaster* IBP with the program MAFFT using the L-INS-i algorithm (v7.164b) [56]. Specifically, we used the IBP amino acid sequences identified and published by Heger and colleagues [8] for building multiple sequence alignments of CTCF, Su(Hw), Mod(mdg4), GAF, CP190, and Zw5. We additionally retrieved the amino acid sequences of all remaining IBPs from NCBI: BEAF-32 (AFH08082.1), Elba1 (AAF50991.2), Elba2 (AAF51239.1), Elba3 (AAF50989.1), Pita (AAF47025.2), ZIPIC/CG7928 (AAF56994.1), Ibf1 (NP_649875), Ibf2 (NP_649874.1). We subsequently built pHMMs from each multiple sequence alignment with the program hmmbuild of the HMMER software package (version 3.1b) [57]. We then screened each transcriptome assembly with the program hmmsearch (also part of the HMMER package) after translating the transcripts into all six possible reading frames with the program fastatranslate (part of the Exonerate software package version 2.2.0) [58]. Only hits with a global *e*-value $\leq 10^{-14}$ were considered as promising IBP transcript candidates. All IBP candidate transcripts were then reciprocally searched against the non-redundant protein (nr) databases entries of *D. melanogaster* (Diptera)*,* *Bombyx mori* (Lepidoptera*), Camponotus* *floridanus* (Hymenoptera), and *Zootermopsis nevadensis* (Isoptera) available at NCBI between January and March 2016 using BLASTP [59] in order to identify best reciprocal genome/transcriptome-wide hits. We considered those identified transcripts orthologous to a specific IBP for which the reciprocal search found the same IBP as best reciprocal database-wide hit. The identified IBP transcripts were subsequently aligned at the transcriptional level with the MAFFT L-INS-i algorithm. If the absence of transcripts suggested a possible IBP-coding gene loss, we searched (draft) genomes with TBLASTN (part of the BLAST+ program suite version 2.2.31) for possible coding sequences of the target proteins.

Domain identification

To annotate the domains within amino acid sequences, we used pHMMs of protein family domains compiled in the Pfam-A database (Release 29.0) [60]. All candidate transcripts of IBPs were searched for protein domains with the program hmmscan (part of the HMMER package) [57] employing the above pHMMs.

Transcriptome completeness assessment

To assess transcriptome assembly completeness, we used BUSCO [40] to search for a set of 2675 conserved genes that are near-universal single copy orthologs in arthropods. These genes are present in single-copy in 95 % of the arthropod species in the OrthoDB database and serve as a benchmark for genome or transcriptome completeness. BUSCO uses a combination of BLAST, pHMMs and a gene model refinement procedure to identify and discriminate present, duplicated, fragmented and missing genes in the searched nucleotide sequence database.

Ancestral state reconstruction

Ancestral state reconstruction was applied in order to infer a hypothesis about the evolutionary gains, or losses, of all IBPs. We compiled a matrix, in which we coded the presence and absence of transcripts of each IBP in each species studied. We used Mesquite (version 3.03; http://mesquite-project.org) [61] to map the gains and losses of insulator proteins on the phylogenetic tree of insects and added the phylogenetic relationships among outgroup taxa (*i.e.*, Onychophora, Tardigrada, and Nematoda) according to Meusemann and colleagues [38, 62] under the Maximum Parsimony optimality criterion. Note that Mesquite does not allow Ancestral state reconstruction under Dollo's parsimony criterion.

Phylogenetic analyses

To better assess the possible occurrence of gene duplication events, we inferred gene trees from the identified putative transcripts of each IBP. For this purpose, we inferred for each IBP a Maximum Likelihood phylogenetic tree based on the corresponding multiple sequence alignment with the program PhyML (version 3.0) [63], using the WAG + Γ substitution model with default settings. Tree robustness was assessed from 1000 bootstrap replicates. We applied the same method when testing whether or not *Zw5* could be a Diptera-specific paralog of the gene *meics*. Specifically, we aligned all available amino acid sequences of Zw5 to the amino acid sequences of Meics of holometabolous insects. We retrieved the latter sequences from OrthoDB (version 8) [64]. Phylogenetic analysis was done as described in the preceding paragraph.

Modes of selection

To search for evidence of positive or negative selection on insulator protein genes, we used the program codeML of the PAML package (version 4.8) [65] to measure the ratio of non-synonymous (amino acid replacing) to synonymous (silent) substitutions (ω). For this purpose, we compiled corresponding nucleotide multiple sequence alignments of the identified transcripts for each IBP separately with Pal2Nal (version 14) [66] by using the multiple sequence alignments of the translated transcripts as blueprints. We used a branch site model, in which ω is allowed to vary along specific branches of the phylogenetic tree, to test for positive selection along these branches. We specifically tested for changes of ω along branches that immediately followed nodes at which we inferred the evolutionary origin of a specific IBP. We used a likelihood ratio test with one degree of freedom to test models, in which ω was allowed to vary along a specific branch, against the null model, in which ω was kept at 1 in all branches of the phylogenetic tree. For each gene, we used the same tree topology as in Fig. 1. Species in which we did not find orthologs of the respective gene were pruned from the tree.

Endnotes

[1] We are using the term insects in a broad sense, including all Hexapoda, equivalent to the nomenclature used in [46].

Additional files

Additional file 1: Figure S1. Tracing the evolutionary origin of CTCF with ancestral state reconstruction. Figure S2. Tracing the evolutionary origin of Su(Hw) with ancestral state reconstruction. Figure S3. Tracing the evolutionary origin of CP190 with ancestral state reconstruction. Figure S4. Tracing the evolutionary origin of GAF with ancestral state reconstruction. Figure S5. Tracing the evolutionary origin of Pita with ancestral state reconstruction. Figure S6. Tracing the evolutionary origin of Mod(mdg4) with ancestral state reconstruction. Figure S7. Tracing the evolutionary origin of Zw5 with ancestral state reconstruction. Figure S8. Phylogenetic gene tree of *CTCF* orthologs. Figure S9. Phylogenetic gene tree of *Su(Hw)* orthologs. Figure S10. Phylogenetic gene tree of *CP190* orthologs. Figure S11. Phylogenetic gene tree of *GAF* orthologs. Figure S12. Phylogenetic gene tree of *Pita* orthologs. Figure S13. Phylogenetic gene tree of *Mod(mdg4)* orthologs. Figure S14. Phylogenetic gene tree of *Zw5* orthologs. Figure S15. Phylogenetic analysis of Zw5 and *meiotic central spindle* (Meics). (PDF 474 kb)

Additional file 2: Table S1. BUSCO assessment of the 1KITE transcriptomes. (XLS 21 kb)

Additional file 3: Table S2. NCBI accession numbers of transcriptome and genome data. (XLS 35 kb)

Abbreviations
BEAF-32: Boundary element associated factor of 32 kD; BUSCO: Benchmarking universal single-copy orthologs; CP190: Centrosomal protein 190 kD; CTCF: CCCTC binding factor; Elba: Early boundary activity; GAF: GAGA-Factor; Ibf1: Insulator binding factor 1; Ibf2: Insulator binding factor 2; IBP: Insulator binding protein; Mod(mdg4): Modifier of mdg4; pHMM: Profile Hidden Markov Model; Su(Hw): Suppressor of hairy wing; ZIPIC: Zinc-finger protein interacting with CP190; Zw5: Zeste white 5

Acknowledgments
This manuscript has been enabled by the 1KITE consortium and the i5K initiative. Alexander Donath, Christoph Mayer, Bernhard Misof, Oliver Niehuis, and Ralph S. Peters furthermore acknowledge the Leibniz association for installing the graduate school for Genomic Biodiversity Research, in which the present study arose. We especially thank Stephen Richards and Richard Gibbs of the Baylor College of Medicine Human Genome Sequencing Center for granting access to i5K pilot data prior to their official publication.

Funding
The sequencing and assembly of the 1KITE transcriptomes were funded by BGI through support to the China National GeneBank. TW and PH were supported by grants from the German Research Foundation (Sonderforschungsbereich 680).

Availability of data and materials
The accession numbers of sequence data used in this study is given in Additional file 3: Table S2. In the rare cases in which no accession number is given, a download link or contact information of a responsible person is given instead. Additionally we made the following data publicly available: (1) The amino acid alignments we based the inference of our gene trees on, (2) the gene trees, (3) the nucleotide alignments we based our inference of d_N/d_S-ratios on. The data is available from the Dryad digital repository: http://dx.doi.org/10.5061/dryad.f4r38.

Authors' contributions
Participated in the design of the study: BM, ON, TP. Contributed data: AD, BM, CM, GM, KM, ON LH, MP, LP, PH, RSP, SL, TW, XZ. Performed data analysis: AD, CM, DD, KM, MP, LP, LV, TP. Manuscript preparation: all authors contributed to the writing of the manuscript, with BM, ON, and TP taking the lead. All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Consent for publication
This section is not applicable to the present study.

Ethics approval and consent to participate
This section is not applicable to the present study.

Author details
[1]Center of Molecular Biodiversity Research, Zoological Research Museum Alexander Koenig, Adenauerallee 160, 51113 Bonn, Germany. [2]University of Tübingen, Geschwister-Scholl-Platz, 72074 Tübingen, Germany. [3]Johannes Gutenberg University Mainz, Institute of Molecular Biology (IMB), Ackermannweg 4, 55128 Mainz, Germany. [4]Department for Evolutionary Biology and Ecology (Institut for Biology I, Zoology), University of Freiburg, Hauptstr. 1, 79104 Freiburg, Germany. [5]Australian National Insect Collection, CSIRO National Research Collections Australia, Clunies Ross Street, Acton, ACT 2601, Australia. [6]Zoological Research Museum Alexander Koenig, Arthropod Department, Adenauerallee 160, 53113 Bonn, Germany. [7]University of Bonn, Institute of Evolutionary Biology and Ecology, An der Immenburg 1, 53121 Bonn, Germany. [8]China National GeneBank-Shenzhen, BGI-Shenzhen, Shenzhen, Guangdong Province 518083, China. [9]Centre for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, Øster Voldgade 5-7, 1350 Copenhagen, Denmark. [10]Beijing Advanced Innovation Center for Food Nutrition and Human Health, China Agricultural University, Beijing 100193, China. [11]College of Food Science and Nutritional Engineering, China Agricultural University, Beijing 100083, China. [12]University of Cologne, Cologne Biocenter, Institute for Genetics, Zülpicher Straße 47a, 50674 Köln, Germany. [13]Department of Zoology, University of Kassel, Heinrich-Plett-Str. 40, 34132 Kassel, Germany.

Received: 7 June 2016 Accepted: 25 October 2016
Published online: 03 November 2016

References
1. Wallace JA, Felsenfeld G. We gather together: insulators and genome organization. Curr Opin Genetics Dev. 2007;17(5):400–7.
2. Hou C, Li L, Qin ZS, Corces VG. Gene Density, Transcription and Insulators contribute to the partition of the *Drosophila* genome into physical domains. Mol Cell. 2012;3:471–84.
3. Yang J, Corces VG. Insulators, long-range interactions, and genome function. Curr Opin Genet Dev. 2012;22(2):86–92.
4. Bhat KM, Farkas G, Karch F, Gyurkovics H, Gausz J, et al. The GAGA factor is required in the early *Drosophila* embryo not only for transcriptional regulation but also for nuclear division. Development. 1996;122:1113–24.
5. Mohan M, Bartkuhn M, Herold M, Philippen A, Heinl N, et al. The *Drosophila* insulator proteins CTCF and CP190 link enhancer blocking to body patterning. EMBO J. 2007;26:4203–14.
6. Roy S, Jiang N, Hart CM. Lack of the *Drosophila BEAF* insulator proteins alters regulation of genes in the antennapedia complex. Mol Genet Genomics. 2011;285:113–23.
7. Schoborg TA, Labrador M. The phylogenetic distribution of non-CTCF insulator proteins is limited to insects and reveals that BEAF-32 is Drosophila lineage specific. J Mol Evol. 2010;70:74–84.
8. Heger P, George R, Wiehe T. Successive gain of insulator proteins in arthropod evolution. Evolution(N Y). 2013;67:2945–56.
9. Blackwood EM, Kadonaga JT. Going the distance: a current view of enhancer action. Science. 1998;291:60–3.
10. Kellum R, Schedl P. A position-effect assay for boundaries of higher order chromosomal domains. Cell. 1991;64(5):941–50.
11. Gaszner M, Felsenfeld G. Insulators: exploiting transcriptional and epigenetic mechanisms. Nat Rev Genet. 2006;7:703–13.
12. Burgesse-Beusse B, Farrell C, Gaszner M, Litt M, Mutskov V, et al. The insulation of genes from external enhancers and silencing chromatin. Proc Natl Acad Sci USA. 2002;99:16433–7.
13. Neufeld EJ, Skalnik DG, Lievens PMJ, Orkin SH. Human CCAAT displacement protein is homologous to the Drosophila homeoprotein, *cut*. Nature Genetics. 1992;1:50–5.
14. Lin N, Li X, Cui K, Chepelev I, Tie F, et al. A barrier-only boundary element delimits the formation of facultative heterochromatin in Drosophila melanogaster and vertebrates. Mol Cell Biol. 2011;31:2729–41.
15. Donze D, Adams CR, Rine J, Kamakaka RT. The boundaries of the silenced HMR domain in *Saccharomyces cerevisiae*. Genes Dev. 1999;13:698–708.
16. Donze D, Kamakaka RT. RNA polymerase III and RNA polymerase II promoter complexes are heterochromatin barriers in *Saccharomyces cerevisiae*. EMBO J. 2001;20:520–31.
17. Noma KI, Cam HP, Maraia RJ, Grewal SIS. A role for *TFIIIC* transcription factor complex in genome organization. Cell. 2006;125:859–72.
18. Raab JR, Chiu J, Zhu J, Katzman S, Kurukuti S, et al. Human tRNA genes function as chromatin insulators. EMBO J. 2012;31:330–50.
19. Heger P, Wiehe T. New tools in the box: An evolutionary synopsis of chromatin insulators. Trends Genet. 2014;30:161–70.
20. Heger P, Marin B, Bartkuhn M, Schierenberg E, Wiehe T. The chromatin insulator CTCF and the emergence of metazoan diversity. Proc Natl Acad Sci. 2012;109:17507–12.
21. Van Bortle K, Corces VG. The role of chromatin insulators in nuclear architecture and genome function. Curr Opin Genet Dev. 2013;23:212–8.
22. Parkhurst SM, Harrison DA, Remington MP, Spana C, Kelley RL, et al. The *Drosophila* su(Hw) gene, which controls the phenotypic effect of the gypsy transposable element, encodes a putative DNA-binding protein. Genes Dev. 1988;2:1205–15.
23. Spana C, Harrison DA, Corces VG. The *Drosophila melanogaster* suppressor of Hairy-wing protein binds to specific sequences of the gypsy retrotransposon. Genes Dev. 1999;2:1414–23.

24. Bell AC, West AG, Felsenfeld G. The protein CTCF is required for the enhancer blocking activity of vertebrate insulators. Cell. 1999;98(3):387–96.
25. Lunyak VV, Prefontaine GG, Núñez E, Cramer T, Ju BG. Developmentally regulated activation of a SINE B2 repeat as a domain boundary in organogenesis. Science. 2007;317:248–51.
26. Román AC, Gonzáles-Rico FJ, Moltó E, Hernando H, Neto A. Dioxin receptor and SLUG transcription factors regulate the insulator activity of B1 SINE retrotransposons via an RNA polymerase switch. Genome Res. 2011;21:422–32.
27. Van Bortle K, Nichols MH, Li L, Ong CT, Takenaka N, et al. Insulator function and topological domain border strength scale with architectural protein occupancy. Genome Biol. 2014;15(6):R82.
28. Matharu NK, Hussain T, Sankaranarayanan R, Mishra RK. Vertebrate homologue of *Drosophila* GAGA factor. J Mol Biol. 2010;400(3):434–47.
29. Abhiman S, Iyer LM, Aravind L. BEN: a novel domain in chromatin factors and DNA viral proteins. Bioinformatics. 2008;24:458–61.
30. Aoki T, Sarkeshik A, Yates J, Schedl P. Elba, a novel developmentally regulated chromatin boundary factor is a hetero-tripartite DNA binding complex. Elife. 2012;2012:1–24.
31. Zollman S, Godt D, Privé GG, Couderc JL, Laski FA. The BTB domain, found primarily in zinc finger proteins, defines an evolutionarily conserved family that includes several developmentally regulated genes in *Drosophila*. P Natl Acad Sci USA. 1994;91:10717–21.
32. Pai CY, Lei EP, Ghosh D, Corces VG. The Centrosomal Protein CP190 is a component of the *gypsy* chromatin insulator. Mol Cell. 2004;16(5):737–48.
33. Reuter G, Giarre M, Farah J, Gausz J, Spierer A, et al. Dependence of position-effect variegation in *Drosophila* on dose of a gene encoding an unusual zinc-finger protein. Nature. 1990;344:219–23.
34. Zhao K, Hart CM, Laemmli UK. Visualization of chromosomal domains with boundary element-associated factor BEAF-32. Cell. 1995;81:879–89.
35. Clark KA, McKearin DM. The *Drosophila* stonewall gene encodes a putative transcription factor essential for germ cell development. Development. 1996;122:937–50.
36. Hsu S-J, Plata MP, Ernest B, Asgarifar S, Labrador M. The insulator protein *Suppressor of Hairy wing* is required for proper ring canal development during oogenesis in *Drosophila*. Dev Biol. 2015;403(1):57–68.
37. Peters RS, Meusemann K, Petersen M, Mayer C, Wilbrandt J, et al. The evolutionary history of holometabolous insects inferred from transcriptome-based phylogeny and comprehensive morphological data. BMC Evol Biol. 2014;14:52.
38. Misof B, Liu S, Meusemann K, Peters RS, Donath A, et al. Phylogenomics resolves the timing and pattern of insect evolution. Science. 2014;346:763–7.
39. Provost E, Shearn A. The suppressor of killer of prune, a unique glutathione S-transferase. J Bioenerg Biomembr. 2006;38:189–95.
40. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31:3210–2.
41. Klug A, Rhodes D. Zinc fingers: a novel protein fold for nucleic acid recognition. Cold Spring Harb Symp Quant Biol. 1987;52:473–82.
42. Klug A. The discovery of zinc fingers and their applications in genome manipulation. Annu Rev Biochem. 2010;79:213–31.
43. Maksimenko O, Bartkuhn M, Stakhov V, Herold M, Zolotarev N, et al. Two new insulator proteins, Pita and *ZIPIC*, target CP190 to chromatin. Genome Res. 2015;25:89–99.
44. Cuartero S, Fresán A, Reina O, Planet E, Espinàs ML. Ibf1 and Ibf2 are novel CP190-interacting proteins required for insulator function. EMBO J. 2014; 33(6):637–47.
45. Wiegmann BM, Trautwein MD, Winkler IS, Barr NB, Kim J, et al. Episodic radiations in the fly tree of life. Proc Natl Acad Sci U S A. 2011;108:5690–5.
46. Gojobori T. Codon substitution in evolution and the "saturation" of synonymous changes. Genetics. 1983;105:1011–27.
47. Smith JM, Smith NH. Synonymous nucleotide divergence: what is "saturation"? Genetics. 1996;142:1033–6.
48. Holohan EE, Kwong C, Adryan B, Bartkuhn M, Herold M, et al. CTCF Genomic Binding Sites in *Drosophila* and the organisation of the bithorax complex. PLOS Genet. 2007;3(7):e112.
49. Hering L, Meyer G. Analysis of the opsin repertoire in the tardigrade Hypsibius dujardini provides insights into the evolution of opsin genes in Panarthropoda. Genome Biol Evol. 2014;6(9):2380–91. *Bioinformatics* 14(9):755-763.
50. Hering L, Henze MJ, Kohler M, Kelber A, Bleidorn C, et al. Opsins in Onychophora (velvet worms) suggest a single origin and subsequent diversification of visual pigments in arthropods. Mol Biol Evol. 2012;29:3451–8.
51. Clark AG, Eisen MB, Smith DR, Bergman CM, Oliver B, et al. Evolution of genes and genomes on the *Drosophila* phylogeny. Nature. 2007;450:203–18.
52. Nene V, Wortman JR, Lawson D, Haas B, Kodira C, et al. Genome sequence of *Aedes aegypti*, a major arbovirus vector. Science. 2007;316:1718–23.
53. Holt RA, Subramanian GM, Halpern A, Sutton GG, Charlab R, et al. The genome sequence of the malaria mosquito anopheles gambiae. Science. 2002;298:129–49.
54. Goldsmith MR, Shimada T, Abe H. The genetics and genomics of the silkworm, *Bombyx mori*. Annu Rev Entomol. 2005;50:71–100.
55. Evans JD, Brown SJ, Hackett KJJ, Robinson G, Richards S, et al. The i5K initiative: advancing arthropod genomics for knowledge, human health, agriculture, and the environment. J Hered. 2013;104:595–600.
56. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Mol Biol Evol. 2013;30:772–80.
57. Eddy SR. Profile hidden Markov models. Bioinformatics. 1998;14(9):755–63.
58. Slater GSC, Birney E. Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics. 2005;6:31.
59. Altschul S, Gish W, Miller W. Basic local alignment search tool. J Mol Biol. 1990;215:403–10.
60. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, et al. Pfam: the protein families database. Nucleic Acids Res. 2014;42:222–30.
61. Maddison WP, Maddison DR. Mesquite: a modular system for evolutionary analysis. Version 2.75.2011. 2015. http://mesquiteproject.org. Accessed Feb 2016.
62. Meusemann K, von Reumont BM, Simon S, Roeding F, Strauss S, et al. A phylogenomic approach to resolve the arthropod tree of life. Mol Biol Evol. 2010;27(11):2451–64.
63. Guindon S, Gascuel O, Dufayard J-F, Lefort V, Anisimova M, et al. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol. 2010;59(3):1–37.
64. Kriventseva EV, Tegenfeldt F, Petty TJ, Waterhouse RM, Simao F, et al. OrthoDB v8: update of the hierarchical catalog of orthologs and the underlying free software. Nucleic Acids Res. 2014;43:D250–6.
65. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 2007;24:1586–91.
66. Suyama M, Torrents D, Bork P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. 2006;34:609–12.
67. Gaszner M, Vazquez J, Schedl P. The Zw5 protein, a component of the *scs* chromatin domain boundary, is able to block enhancer–promoter interaction. Genes Dev. 1999;13(16):2098–107.
68. Whitfield WG, Chaplin MA, Oegema K, Parry H, Glover DM. The 190 kDa centrosome-associated protein of Drosophila melanogaster contains four zinc finger motifs and binds to specific sites on polytene chromosomes. J Cell Sci. 1995;108:3377–87.
69. Omichinski JG, Pedone PV, Felsenfeld G, Gronenborn AM, Clore GM. The solution structure of a specific GAGA factor-DNA complex reveals a modular binding mode. Nat Struct Biol. 1997;4:122–32.
70. Ohtsuki S, Levine M. GAGA mediates the enhancer blocking activity of the eve promoter in the Drosophila embryo. Genes Dev. 1998;12(21):3325–30.
71. Gerasimova TI, Gdula D a, Gerasimov DV, Simonova O, Corces VG. A *Drosophila* protein that imparts directionality on a chromatin insulator is an enhancer of position-effect variegation. Cell. 1995;82:587–97.
72. Dorn R, Krauss V. The modifier of mdg4 locus in *Drosophila*: functional complexity is resolved by trans splicing. Genetica. 2003;117:165–77.