PROCEEDINGS B

royalsocietypublishing.org/journal/rspb

Research

Cite this article: Luo M, Finet C, Cong H, Wei H-y, Chung H. 2020 The evolution of insect metallothioneins. *Proc. R. Soc. B* **287**: 20202189.
http://dx.doi.org/10.1098/rspb.2020.2189

Received: 5 September 2020
Accepted: 5 October 2020

Subject Category:
Evolution

Subject Areas:
bioinformatics, evolution, genomics

Keywords:
metallothionein, heavy metal, insect, genome evolution, tandem duplication, MtnF

Author for correspondence:
Henry Chung
e-mail: hwchung@msu.edu

The evolution of insect metallothioneins

Mei Luo[1,3,†], Cédric Finet[4,†], Haosu Cong[1], Hong-yi Wei[3] and Henry Chung[1,2]

[1]Department of Entomology, and [2]Ecology, Evolution, and Behavior Program, Michigan State University, East Lansing, MI 48824, USA
[3]College of Agronomy, Jiangxi Agricultural University, Nanchang 330045, People's Republic of China
[4]Yale-NUS College, 16 College Avenue West, Singapore 138527, Republic of Singapore

ML, 0000-0003-1596-418X; HCo, 0000-0001-6505-4075; HCh, 0000-0001-5056-2755

Metallothioneins (MTs) are a family of cysteine-rich metal-binding proteins that are important in the chelating and detoxification of toxic heavy metals. Until now, the short length and the low sequence complexity of MTs have hindered the inference of robust phylogenies, hampering the study of their evolution. To address this longstanding question, we applied an iterative BLAST search pipeline that allowed us to build a unique dataset of more than 300 MT sequences in insects. By combining phylogenetics and synteny analysis, we reconstructed the evolutionary history of MTs in insects. We show that the MT content in insects has been shaped by lineage-specific tandem duplications from a single ancestral MT. Strikingly, we also uncovered a sixth MT, MtnF, in the model organism *Drosophila melanogaster*. MtnF evolves faster than other MTs and is characterized by a non-canonical length and higher cysteine content. Our methodological framework not only paves the way for future studies on heavy metal detoxification but can also allow us to identify other previously unidentified genes and other low complexity genomic features.

1. Introduction

Heavy metals like copper (Cu) and zinc (Zn) have been co-opted over time as essential components of numerous transcriptional factors and catalytic enzymes [1]. However, high concentrations of heavy metals can be cytotoxic, and organisms have evolved intricate strategies to detoxify and excrete heavy metals. Over time, these detoxification strategies have also been used to detoxify other non-essential heavy metals such as cadmium (Cd) [2]. Detoxification strategies are diverse and can often be taxon-specific [3,4]. However, a common mechanism among many organisms is the use of low molecular weight, cysteine-rich peptides known as metallothioneins (MTs) to chelate and regulate the concentration of heavy metals in the cell [1].

Since their seminal discovery in the horse kidney [5], MTs have been discovered in both eukaryotes and prokaryotes. However, because of their short sequences (approx. 60 amino acids) and low sequence complexity due to their high cysteine content, it has been claimed that obtaining robust phylogenies is very difficult [6,7]. Instead, MTs have been classified into 15 families according to the organism they are isolated from [7] or classified by their function as either Zn- or Cu-thioneins [8]. While most organisms have Cu-thioneins, Zn-thioneins have only been found in higher organisms such as the Metazoa [9,10]. Another reason that makes phylogenetic analyses of MTs difficult is the scarcity of annotated MT sequences in sequenced genomes. Indeed, *in silico* gene predictions often fail to identify MT-encoding genes that contain very small exons separated by large introns [11,12].

MTs have been intensively studied in the fruit fly *Drosophila melanogaster*. *MtnA* was the first member of this gene family cloned from copper-fed larvae [13], followed by *MtnB* cloned from a cadmium-resistant *Drosophila* cell line [14]. More than a decade later, the release of the *D. melanogaster* genome [15] allowed the identification of *MtnC* and *MtnD* by sequence similarity [16]. The four *MtnA–D* genes are inducible by copper and cadmium through the binding of the transcription factor MTF-1 [16]. Further work showed that *MtnA*

[†]These authors contributed equally.

Electronic supplementary material is available online at https://doi.org/10.6084/m9.figshare.c.5178251.

THE ROYAL SOCIETY
PUBLISHING

knockout flies are sensitive to copper, *MtnB* knockout flies are sensitive to cadmium, whereas *MtnC* and *MtnD* knockout flies do not show any differences in copper or cadmium resistance, suggesting that these MTs have distinct roles in heavy metal detoxification [17]. In 2011, a fifth member, *MtnE*, was discovered in the *D. melanogaster* genome through bioinformatic analysis [18]. *MtnE* is also inducible by heavy metals such as copper [18], and it is classified as a Cu-thionein like the other four *Drosophila* MTs [4,19]. A striking feature is the absence of Zn-thioneins in *D. melanogaster*, and more generally in insects, whereas they are found in all other metazoans [20]. Outside the *Drosophila* genus [21], only 14 MT genes have been published in insects (electronic supplementary material, table S1). There is only one MT in the honeybee *Apis mellifera* [12], two MTs in the Chinese grasshopper *Oxya chinensis* [22], but five MTs in sequenced *Drosophila* species genomes [21]. These studies suggest that the number of MTs in species is dynamic and evolves rapidly across insects. Furthermore, insects live in diverse environments with different heavy metal challenges [4,23,24], offering a good model to study how heavy metal detoxification evolves.

To investigate how insect MTs evolve, we built a large dataset of MT sequences encompassing the main insect orders. We took advantage of the recent release of a large amount of genomic and transcriptomic data in insects [25,26]. To avoid dealing with large introns in genomes, we applied an iterative BLAST search pipeline on available insect transcriptomes on the InsectBase [26], NCBI NR and NT databases. In total, we identified and annotated more than 300 insect MTs from about 100 insect species based on available sequenced genomes and transcriptomes. Using a combination of phylogenetic and synteny analyses, we showed that the insect MTs evolved from one single ancestral MT gene prior to the diversification of insects. We also discovered MtnF, a putative sixth metallothionein in the Diptera, including *D. melanogaster* in spite of previous intensive work in this model species. MtnF possesses non-canonical features compared with other insect MTs, suggesting putative different binding specificities.

2. Results and discussion

(a) Combined phylogenetic and synteny analyses reveal a single ancestral metallothionein in insects

To facilitate comparative studies of MTs in insects, we reasoned that a large sampling should be of help. We applied an exhaustive BLAST search to around a hundred insect available transcriptomes. We identified more than 300 MT sequences encompassing 13 orders across different insect species (table 1), whereas the number of published MTs was 19 prior to this study (electronic supplementary material, table S1). One caveat of this approach is that using data mainly from transcriptomics may miss lowly expressed MTs as well as remnants of MT pseudogenes that are not expressed. The coverage of the 135 insect transcriptomes available from InsectBase is biased towards the orders Diptera, Hymenoptera, Lepidoptera, Coleoptera and Hemiptera (electronic supplementary material, figure S1A), which might explain the overrepresentation of these orders in our dataset. However, we found that the InsectBase coverage does reflect the species richness across extant insect orders (electronic supplementary material, figure S1B) according to the 'Catalogue of life' [27].

Table 1. Summary of MT content in sampled insect species.

order	number of sampled species	number of MTs
Blattodea	2	3
Coleoptera	11	13
Dermaptera	1	2
Diptera	46	195
Hemiptera	7	8
Hymenoptera	31	35
Lepidoptera	27	73
Neuroptera	1	2
Odonata	1	2
Orthoptera	2	3
Phasmatodea	4	6
Plecoptera	1	3
Siphonaptera	1	1
total	135	346

After preliminary phylogenetic analyses, we removed some overrepresented dipteran sequences to avoid sampling bias, and highly divergent sequences to avoid long branch attraction. We performed maximum-likelihood phylogenetic reconstruction on the resulting 202-taxon MT dataset. The tree obtained clarifies the number of main MT lineages in insects. In this tree, the coleopteran MTs are clustered in a single clade, like the hymenopteran MTs, suggesting they derive from a single MT in the last common ancestor of the extant Coleoptera and Hymenoptera, respectively. The lepidopteran MTs split into three main clades, which we named clades α, β and γ (figure 1*a*). This result suggests that at least three MT genes were present in the last common ancestor of the extant Lepidoptera. Similarly, the dipteran MTs split into several clades, suggesting multiple ancestral MTs in the Diptera. The obtained tree also showed that the lepidopteran clade γ groups with dipteran clades (figure 1*a*). The branch leading to this Lepidoptera–Diptera superclade is well supported with a bootstrap value of 73 (figure 1*a*; electronic supplementary material, figure S2A). Because Lepidoptera and Diptera are two closely-related orders [28], our data support a MT gene duplication prior to the divergence between Lepidoptera and Diptera.

Owing to their short size and high divergence, MTs are markers with limited phylogenetic information. To provide further evidence for our aforementioned observations, we also investigated the conservation of synteny at MT loci as an independent and supplementary approach [29,30]. We found that the MT sequences are flanked by the same genes in the Isoptera (*Zootermopsis nevadensis* and *Cryptotermes secundus*), in the Hymenoptera (*Apis mellifera* and *Orussus abietinus*), in the Coleoptera (*Aethina tumida* and *Tribolium castaneum*) and in the Lepidoptera (*Danaus plexippus* and *Bombyx mori*) (figure 1*b*). No evidence of such conservation was found in available hemipteran genomes. Along with the presence of orthologues between Lepidoptera and Diptera, this deep conservation of microsynteny across the insects suggests the existence of a single MT gene in the last common ancestor of extant insects.

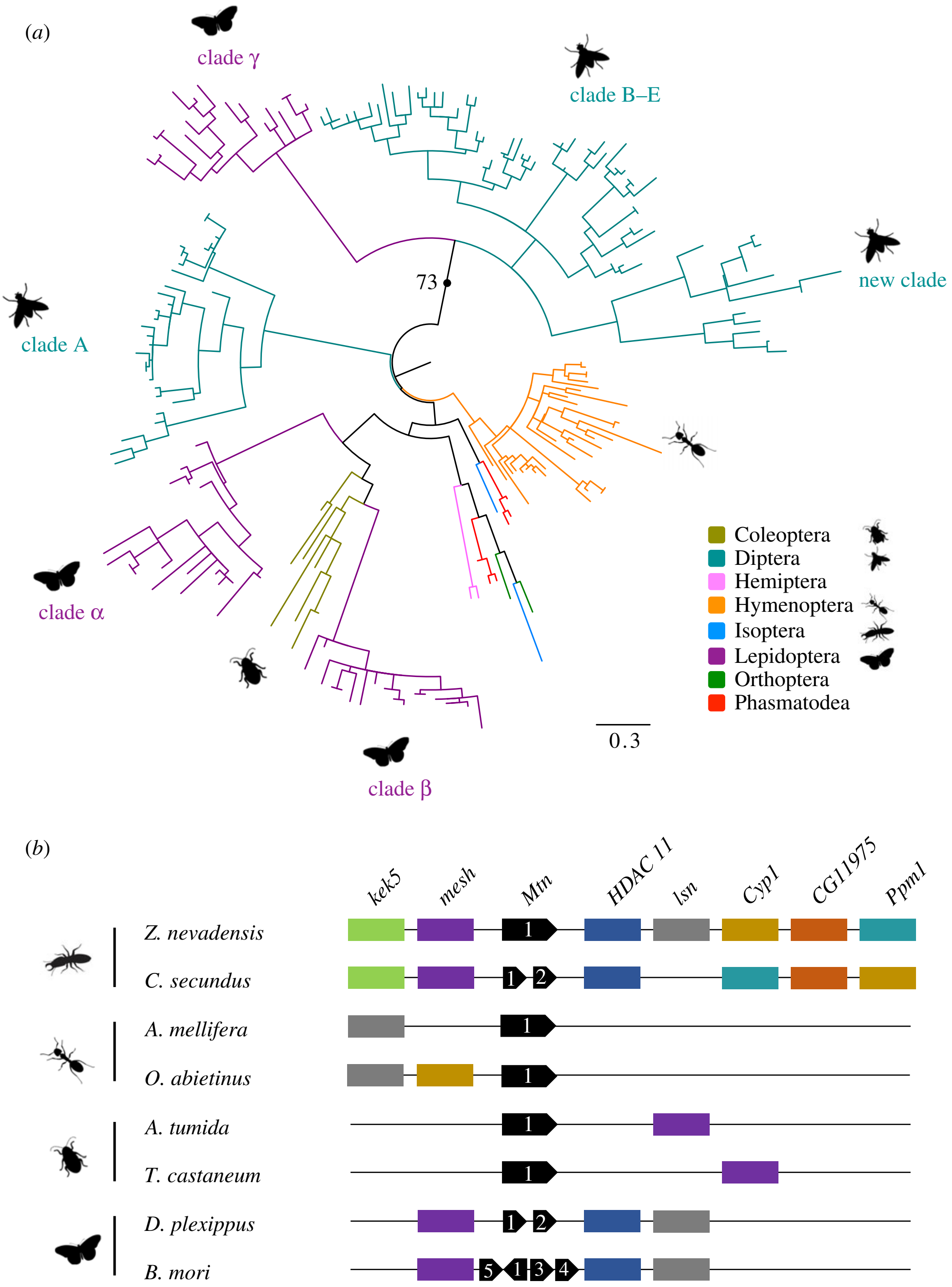

Figure 1. Phylogeny and synteny of MT genes in Insecta. (*a*) Phylogram of the 202-taxon analyses obtained through PhyML maximum-likelihood reconstruction. Analyses were conducted using the LG + *Γ* model. Support value is shown for selected branches (all branch supports are shown in electronic supplementary material, figure S2A). Scale bar indicates number of changes per site. (*b*) The conservation of synteny was found for several MT genes in the Coleoptera, Hymenoptera, Lepidoptera and Isoptera, suggesting a single MT locus in the last common ancestor of extant insects. No evidence of synteny conservation was found in the Hemiptera. Orthologous genes have the same colour.

It is noteworthy that dipteran MT genes are not located at the same genomic locus shared with other insects. Nevertheless, the shared flanking genes *mesh*, *HDAC11*, *lsn* and *CG11975* have been identified on the chromosome 3R in *D. melanogaster*, where the MTs are located. The high rate of chromosomal rearrangement in the Diptera [31–34] might explain the breakdown of the ancestral MT locus in the Diptera. In particular, the genomic locus surrounding *MtnF* is deeply rearranged in *Drosophila buzzatii* [35]. On the contrary, the conservation of microsynteny at the MT locus in the Lepidoptera is in line with low rates of change in gene order that have been found in this order based on coarse-scale mapping data [36–38]. Whereas conservation of microsynteny between Hemiptera and other insects has been reported [39,40], we did not find clear evidence of microsynteny at the MT locus in Hemiptera. Future increase of the currently scarce genomic data in Hemiptera should help to conclude whether the MT locus has been translocated.

(b) Insect metallothionein repertoire originated through tandem duplication

A widespread feature of the insect MT gene family is the presence of several copies in tandem in sequenced insect genomes. For example, the moth *B. mori* has four MT genes in tandem (figure 1*b*) and the housefly *Musca domestica* has four clustered copies of *MntA* (see below). This is also true for earliest-diverging lineages of insects like *C. secundus*, for which we identified two MT copies in tandem (figure 1*b*). Similar results were obtained for the genes of the clade B–E in *Drosophila* [16,21]. Those observations suggest that the MT content of insect genomes is mainly shaped by many lineage-specific duplication events.

Among the insects we investigated, lepidopteran insects show a high number of duplicated MTs, with an average of 2.7 copies in tandem per species. Whereas an ancient whole-genome duplication in Lepidoptera might explain the extant MT content [41], recent studies support the occurrence of segmental duplications rather than a whole-genome duplication in Lepidoptera [42,43]. Tandem duplications are associated with evolutionarily relevant traits in the butterflies and moths such as host plant detoxification [44], gustatory and odorant receptor diversification [45,46], vision [47], and wing colour variation [46,48]. Extensive tandem duplication and retention of MT genes in the Lepidoptera reinforces the putative adaptive role of these proteins. Our findings raise the question about the specific high number of MT genes in lepidopteran insects. Caterpillars feed on plant leaves, but also on seeds and flowers of a large or restricted range of host species according to their lifestyle. After metamorphosis, butterflies and moths feed primarily from floral organs and rewards, in which heavy metals can accumulate [49]. In line with the increase in MT copies that correlates with augmented metal tolerance in *Drosophila* [50–52], having a richer repertoire of MT proteins might be an advantage for the fitness of Lepidoptera which occupy and feed on a wide variety of host species/tissues during their life cycle.

(c) Evolution of the metallothionein repertoire in the Diptera

As our large-scale analysis suggested substantial changes in the Diptera, we focused on the evolution of MTs in this particular insect order. We performed maximum-likelihood phylogenetic reconstruction using an alignment of 141 MT proteins restricted to the Diptera, including more dipteran sequences than the 202-taxon dataset. The tree obtained clarifies the number of main MT clades and their relationships (figure 2*a*; electronic supplementary material, figure S2B). In this tree, MT sequences split into three main clades: clade A contains orthologues of the *D. melanogaster* gene *MtnA* [13], clade B–E contains homologues of the *D. melanogaster MtnB*-like cluster genes [16,18] and clade F contains orthologues of the uncharacterized *D. melanogaster* open reading frame (ORF) *CG43222*. Clade F, which represents a new clade of MT previously unidentified, is discussed in detail in the next section. Clade B–E contains sequences from the early-diverging Culicidae (mosquitoes), indicating that clade B–E can be traced back to the origin of Diptera. Clades A and F contain sequences that encompass several families of Diptera, suggesting that clades A and F could also be traced back to the origin of Diptera. According to the tree topology, it is reasonable to propose that the last common ancestor of extant Diptera possessed at least three MT genes. As noted, several MT copies have been generated through tandem gene duplication events, and such copies may be subject to regular gene conversion events. For example, the clustering of *D. yakubaB* and *D. yakubaC* likely results from gene conversion (figure 2*a*). The cases of gene conversion remain, however, rare in our dataset, and should not deeply obscure true evolutionary histories.

We also investigated the conservation of synteny at the MT loci. Genomic data were available for eight sequences that group into clade A in the species *D. melanogaster*, *M. domestica*, *Lucilia cuprina* and *Zeugodacus cucurbitae*. All these MT sequences show a high level of synteny conservation with a recurrent handful of flanking genes (figure 2*b–d*). Our results confirm that sequences clustered in clade A correspond to orthologues of *MtnA*. We found such a pattern of conserved microsynteny for sequences within clade F (figure 2*c*). The synteny of the genes of the *MtnB*-like cluster is less conserved at microscale, but well conserved at a larger genomic scale (figure 2*d*). Our finding suggests that the numerous MT genes are clustered on the same part of the chromosome in at least *L. cuprina* (Calliphoridae) and *M. domestica* (Muscidae), as is the case in Drosophilidae [21]. Our syntenic approach also confirms that the last common ancestor of extant Diptera possessed at least three MT genes.

(d) Identification of MtnF, a new metallothionein member in the Diptera

Our phylogenetic analysis reveals a new clade of MT in the extant Diptera (figures 1*a* and 2*a*). With regard to the MT repertoire already known in *D. melanogaster*, we subsequently named it *MtnF* in this species, in which it corresponds to the uncharacterized ORF *CG43222*. This finding is even more remarkable in the model species *D. melanogaster*, where the MT family has been studied since the 1980s. Following the cloning of its first MT [13], the *D. melanogaster* genome had been reported to contain four MT genes (*MtnA*, *MtnB*, *MtnC* and *MtnD*) that are clustered on the right arm of the third chromosome [16]. In 2011, the fifth MT gene, *MtnE*, was identified and shown to locate inside the MtnB-like cluster in *D. melanogaster* [18].

The alignment of the N-terminal end of MtnF with its *Drosophila* homologues highlights the conservation between the different MT members (figure 3*a*), suggesting that MtnF is a member of the MT family. To better characterize this new MT member, we use I-TASSER to predict the 3D structure of the protein MtnF in *D. melanogaster*. The best model (*C*-score = −0.63) predicts two possible β-sheet secondary structures (figure 3*b*) and fits well (TM-score = 0.538) to the crystal structure of the rat Mt2 (purple structure, figures 3*c*,*d*), supporting the hypothesis that MtnF might function as a MT. Another piece of evidence supporting *MtnF* as a MT-encoding gene is the presence of two putative MTF-1 binding sites in its upstream region (electronic supplementary material, figure S3). We found one core consensus sequence of metal response element TGCRCNCG [53,54] in the non-coding 5′ region of the *MtnF* locus, and one in the coding region of the 5′ flanking gene *puffyeye*.

(e) Might MtnF be a Zn-thionein in the Diptera?

As previously mentioned, a striking feature is the absence of Zn-thioneins in insects whereas they are found in all other metazoans [20]. The identification of a sixth MT member in *D. melanogaster* naturally raises the question of whether MtnF could act as a Zn-thionein. We found that *MtnF* orthologues evolve faster than *MtnA* and *MtnB–E* orthologues. On average, *MtnF* orthologues have longer branch lengths

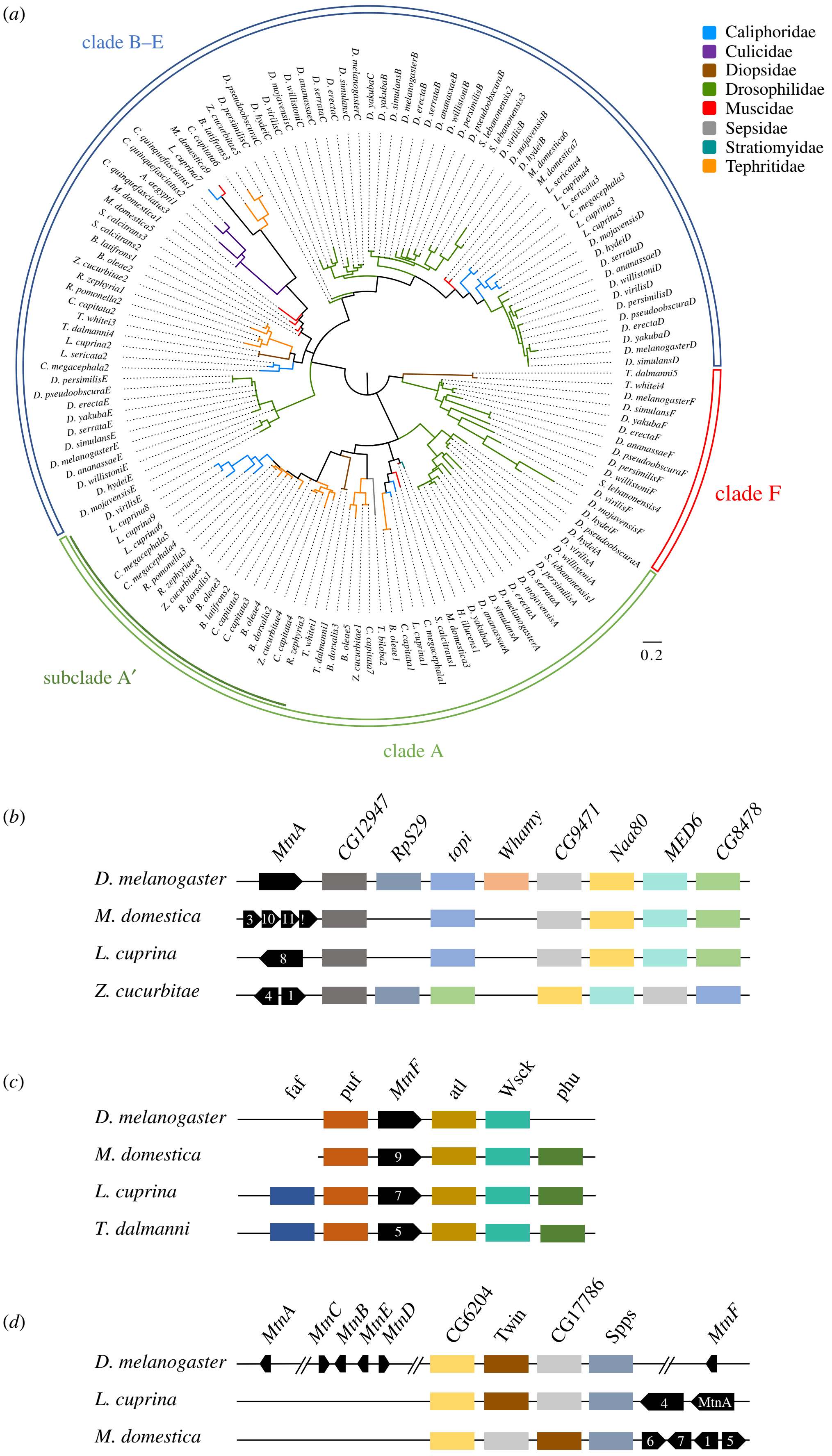

Figure 2. Synteny and phylogeny of MT genes in Diptera. (*a*) The phylogram of the 141-taxon analyses was obtained through PhyML maximum-likelihood reconstruction. Subclade A′ is a subclade with different features like longer length and higher cysteine content in clade A. Analyses were conducted using the LG + Γ model. Support values obtained after 1000 bootstrap replicates are shown in electronic supplementary material, figure S2B. Scale bar indicates number of changes per site. (*b*) Deep conservation across the Diptera at the *MtnA* locus. The copie *M. domestica10* and *M. domestica11* have been identified in the genome of *M. domestica*, but not in the transcriptomes. The truncated copy *M. domesticaψ* is likely a pseudogene. (*c*) Deep conservation across the Diptera at the *MtnF* locus. (*d*) Conserved chromosomal clustering of the *MtnA*, *MtnB–E* and *MtnF* loci. Broken lines indicate non-contiguous genomic positions. In the *D. melanogaster* genome, the distances are: [*MtnA*, *MtnC*] ≈ 10 500 kb, [*MtnD*-CG6204] ≈ 3600 kb and [Spps-*MtnF*] ≈ 400 kb.

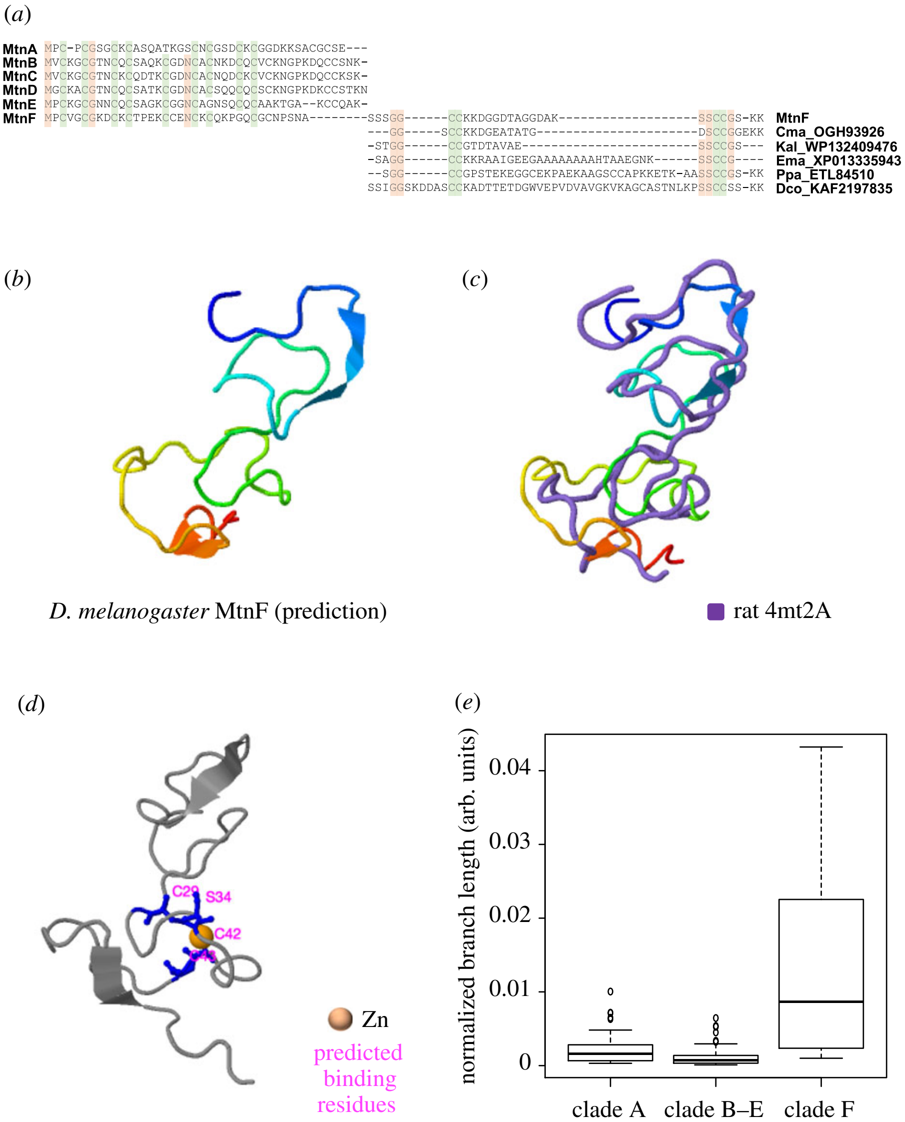

Figure 3. Evolution and structure of the newly identified MtnF of *D. melanogaster*. (*a*) Alignment of the six MT proteins of *D. melanogaster*. MtnF has an extra motif in the C-terminus that can be aligned with non-insect proteins. (*b*) Predicted 3D conformation of the protein MtnF of *D. melanogaster* (RasMol colours). (*c*) Alignment of the predicted 3D structure of *D. melanogaster* MtnF with that the rat MT 4mt2A obtained by crystallography. (*d*) Position of the atom of Zn predicted as ligand and predicted associated binding residues. (*e*) Comparison of branch lengths between clade A, clade B–E and clade F. Cma: *Candidatus magasanikbacteria*, Kal: *Kribbella albertanoniae*, Ema: *Eimeria maxima*, Ppa: *Phytophthora parasitica*, Dco: *Delitschia confertaspora*.

than *MtnA* (*t*-test: d.f. = 23, $p = 0.0003$) and *MtnB–E* (*t*-test: d.f. = 23, $p = 0.0001$) orthologues (figure 3*e*). Because paralogues are present in our dataset, the clades with more paralogues are expected to show overall shorter branch lengths. We performed the same analysis on an alternative dataset without paralogues. We obtained very similar results. *MtnF* sequences have longer branch lengths than *MtnA* (*t*-test: d.f. = 18, $p = 0.0005$) and *MtnB-E* (*t*-test: d.f. = 18, $p = 0.0002$) sequences (electronic supplementary material, figure S4). This higher divergence might explain why previous searches failed to identify *CG43222* as a MT-coding gene.

With an average length of 62 amino acids, the proteins of the clade F are the longest MTs in insects (figure 4*a*). The alignment of the MtnF protein sequence with the other five MTs in *D. melanogaster* shows the longer length is due to extra amino acid residues in the C-terminus (figure 3*a*). This observation raises the question of the origin of these residues. BLASTP and PSI-BLAST searches identified similar motifs in bacteria (Candidatus magasanikbacteria, *Kribbella albertanoniae*), apicomplexan protozoans (*Eimeria maxima*), fungus-like oomycota (*Phytophthora parasitica*) and fungi (*Delitschia confertaspora*). Interestingly, the proteins Ema_XP013335943 and Ppa_ETL84510 are a putative copper transporter and a putative heavy-metal-translocating P-type ATPase, respectively. However, the conservation of the N-terminus across *Drosophila* MTs does not really support an acquisition of the C-terminal peptide by horizontal transfer.

The augmented length of the clade F proteins goes along with a richer cysteine content. We counted the number of cysteines per MT protein for each clade (figure 4*b*). On average, proteins of the MtnF clade contain 13.4 cysteines, which is statistically higher than the average number of 12 cysteines found in insects in our whole dataset (*t*-test: d.f. = 13, $p < 1.4 \times 10^{-4}$). We also unravelled a bimodal distribution of the number of cysteines within clade A. Whereas most of the clade A proteins contain an average number of 10 cysteines (figure 4*b*), a subgroup of sequences, subclade A′, contains an average number of 16 cysteines (figure 4*b*), which is statistically higher than the cysteine average content in clade A (*t*-test: d.f. = 17, $p < 1.0 \times 10^{-5}$) and in insects (*t*-test:

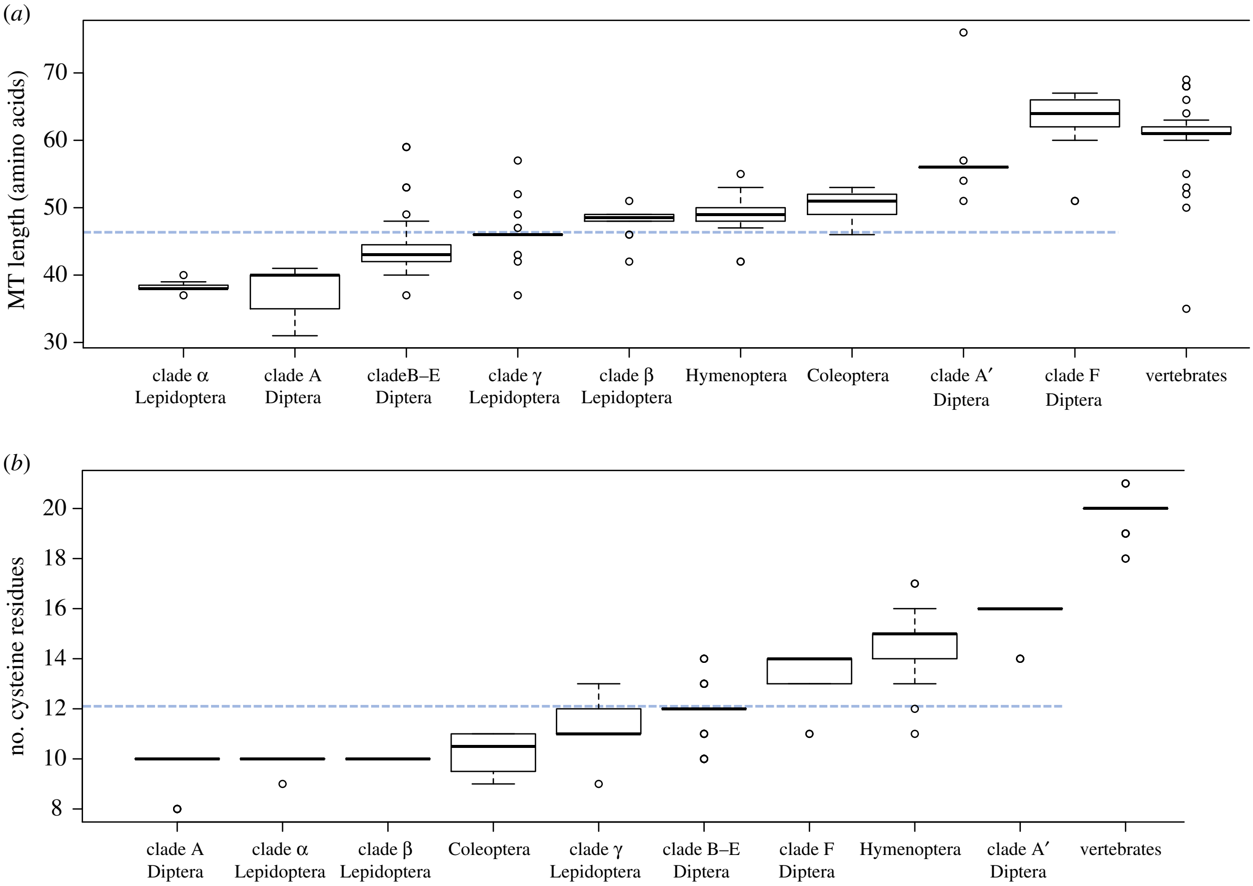

Figure 4. Comparative characterization of two MT features. (*a*) Length of MT proteins in insects and vertebrates, (*b*) Number of cysteine residues per MT in insects and vertebrates. The blue dashed line indicates the mean of the plotted variable.

d.f. = 17, $p < 1.0 \times 10^{-5}$). The embedded position of subclade A′ within clade A suggests that the average gain of six cysteines results from a unique event during the evolution of the clade A. Similarly, we observed a cysteine enrichment in the hymenopteran MTs (*t*-test: d.f. = 33, $p < 1.0 \times 10^{-5}$) compared with the average cysteine content in insects. Our study suggests that the enrichment in cysteine is a derived MT feature that appeared several times independently during the course of evolution.

What could be the biological function(s) of MtnF? Determining the specificity of metal binding properties of MTs has been a prevalent question for many years. Because of their longer size and higher cysteine content, the MTs of clade F are reminiscent of the vertebrate MTs (figure 4*a*,*b*). Vertebrate MTs contain two metal-thiolate clusters, named the β-domain, with three binding sites for divalent ions involving nine cysteinyl sulfurs, and α-domain, capable of binding four divalent metal ions involving 11 cysteinyl sulfurs [55,56], and are able to bind divalent Zn(II) in contrast to insect MTs. According to its structural and evolutionary peculiarities, we propose that MtnF might be a putative Zn-thionein in the Diptera.

3. Conclusion

In this paper, we have identified more than 300 new insect MT sequences and studied their evolutionary relationship using a combination of phylogenetics and synteny analysis. This combined approach has allowed us to leverage the conserved microsynteny in insects to reconstruct the evolution of sequences with limited informative sites. Our data suggest the presence of a single MT in the last common ancestor of extant insects (figure 5). The MT content was subsequently shaped by lineage-specific tandem duplications. We suggest that the dynamic changes in the MT content across insects may reflect differences in environment, diet, and past contact with heavy metals. We also expect that the newly identified MT sequences will allow future biomarkers of heavy metal contamination to be developed. More importantly, our combined phylogenetics and synteny analysis can allow us to identify other previously unidentified genes and other low complexity genomic features.

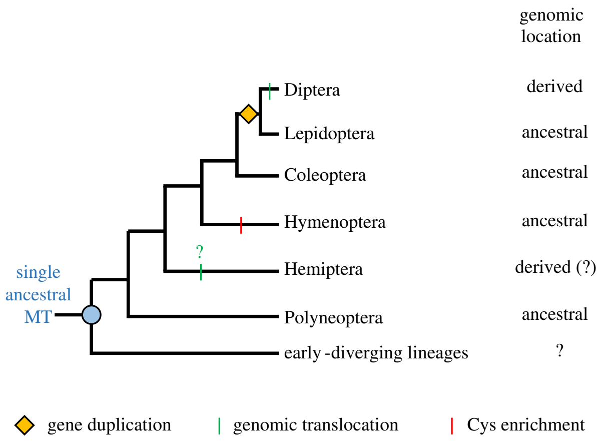

Figure 5. Summary of main molecular events during insect evolution. Gene duplication events are depicted by a yellow diamond, genomic translocations by a vertical green bar and cysteine-enrichment by a vertical red line.

4. Material and methods

(a) Data collection

MT sequences were identified in 116 insect transcriptomes by TBLASTN using 20 published MT sequences as queries (electronic supplementary material, table S1). The transcriptomes were retrieved from the InsectBase website (http://www.insect-genome.com/) [26] and used to build a local database with BioEdit 5.0.6 [57]. The putative identified MTs were added to the set of probes, and TBLASTN searches were iteratively repeated until no new MTs were found. Subsequently, we used all these newly identified MTs as probes to search for additional MT sequences in NCBI NR (protein) and NT (nucleotide) databases using BLASTP and TBLASTN, respectively. In order to remove false positive MTs, we predicted ORFs of all putative MT nucleotide sequences using the EMBOSS program getorf (http://www.bioinformatics.nl/cgi-bin/emboss/getorf). Putative MT sequences for which we failed to identify ORFs were discarded. Finally, repeated sequences and unlikely isoforms from InsectBase and NCBI NR and NT databases were manually removed from our dataset based on protein sequences and conserved cysteine motifs of MT. Species used to build the tree are found in electronic supplementary material, table S2.

(b) Phylogenetic analysis

Amino acid sequences were aligned with MUSCLE [58] and manually adjusted, and conserved blocks were used for phylogenetic reconstruction. Maximum-likelihood searches were performed using PhyML 3.0 [59] under the LG substitution matrix with final likelihood evaluation using a gamma distribution. One-thousand bootstrap replicates were conducted for support estimation. Molluscan MT sequences and lepidopteran MT sequences were used to root the Insecta tree (figure 1*a*) and the Diptera tree (figure 2*a*), respectively.

(c) Synteny analysis

Genomic sequences were retrieved from NCBI (http://www.ncbi.nlm.nih.gov), InsectBase [26] and 5000 Insect Genome Project (i5 k) [60] databases by BLASTN using the corresponding coding sequences as queries. Gene and order content of the genomic scaffolds or contigs were assessed by BLASTX against the annotated proteins of *D. melanogaster* (release 6.31) [61].

(d) Statistical analysis

Branch lengths (BL) were obtained as outputs of PhyML software [59]. To consider differences in number of sequences per clade, we calculated the normalized BL, that is the value of BL/number of MT sequences per clade. First, we used the whole dipteran 141-taxon dataset (figure 2*a*). Second, we tested for the effect of recent paralogues with shorter branch. We generated an alternative dataset by retaining copies with longer branch and removing the following copies with shorter branch: clade A (*C. megacephala4*, *L. cuprina6*, *L. cuprina9*), clade B–E (*C. quinquefasciatus1*, *C. quinquefasciatus3*, *L. cuprina3*, *M. domestica5*, *M. domestica6*, *S. calcitrans2*, *S. lebanonensis2*). We compared the BL means between clade A, clade B–E and clade F using a *t*-test as the BL followed a normal distribution. Statistical tests and graphics were performed using R statistics package v. 3.5.0 (the R Project for Statistical Computing, www.r-project.org, last accessed 17 December 2019).

(e) Prediction of protein structure

The 3D structure of the newly identified metallothionein MtnF in *D. melanogaster* was predicted using the I-TASSER server [62]. I-TASSER simulations generate a large ensemble of structural conformations, called decoys. To select the final models, I-TASSER uses the SPICKER program to cluster all the decoys based on the pairwise structure similarity, and reports up to five models that correspond to the five largest structure clusters. The confidence of each model is quantitatively measured by *C*-score calculated based on the significance of threading template alignments and the convergence parameters of the structure assembly simulations. *C*-score is typically in the range of [−5, 2], where a *C*-score of a higher value signifies a model with a higher confidence. The crystal structure of the rat Mt2 [63] (PDB code 4mt2) was used to model the MtnF protein. Ligand binding sites were predicted using the COFACTOR software [64].

Data accessibility. Datasets, sequence alignments and tree files are available from the Dryad Digital Repository: https://doi.org/10.5061/dryad.pc866t1mh [65]. This manuscript was originally uploaded as a preprint on *bioRxiv* [66].

Authors' contributions. M.L., C.F. and H.Ch. designed research. M.L. and H.Co. collected the data. C.F., M.L. and H.Ch. analysed the data. C.F and H.Ch. wrote the first draft of the manuscript with input from M.L. All authors contributed to various drafts of the manuscript.

Competing interests. We declare we have no competing interests.

Funding. This work is supported by USDA NIFA via Michigan State University AgBioresearch (Umbrella project MICL02522 to H.Ch.) and by the National Natural Science Foundation of China (nos 31760637, 31640064 to H.-y.W.). M.L. is supported by the China Scholarship Council (CSC) (file no. 201708360095).

References

1. Hamer DH. 1986 Metallothionein. *Annu. Rev. Biochem.* **55**, 913–951. (doi:10.1146/annurev.bi.55.070186.004405)
2. Klaassen CD, Liu J, Choudhuri S. 1999 Metallothionein: an intracellular protein to protect against cadmium toxicity. *Annu. Rev. Pharmacol. Toxicol.* **39**, 267–294. (doi:10.1146/annurev.pharmtox.39.1.267)
3. Cobbett C, Goldsbrough P. 2002 Phytochelatins and metallothioneins: roles in heavy metal detoxification and homeostasis. *Annu. Rev. Plant Biol.* **53**, 159–182. (doi:10.1146/annurev.arplant.53.100301.135154)
4. Navarro JA, Schneuwly S. 2017 Copper and zinc homeostasis: lessons from *Drosophila melanogaster*. *Front. Genet.* **8**, 223. (doi:10.3389/fgene.2017.00223)
5. Margoshes M, Vallee BL. 1957 A cadmium protein from equine kidney cortex. *J. Am. Chem. Soc.* **79**, 4813–4814. (doi:10.1021/ja01574a064)
6. Ziller A, Fraissinet-Tachet L. 2018 Metallothionein diversity and distribution in the tree of life: a multifunctional protein. *Metallomics* **10**, 1549–1559. (doi:10.1039/C8MT00165K)
7. Binz P-A, Kägi JHR. 1999 Metallothionein: molecular evolution and classification. In *Metallothionein IV. Advances in Life Sciences* (ed. CD Klaassen), pp. 7–13. Basel, Switzerland: Birkhäuser. (doi:10.1007/978-3-0348-8847-9_2)
8. Palacios Ò, Atrian S, Capdevila M. 2011 Zn- and Cu-thioneins: a functional classification for metallothioneins? *J. Biol. Inorg. Chem.* **16**, 991. (doi:10.1007/s00775-011-0827-2)
9. Capdevila M, Atrian S. 2011 Metallothionein protein evolution: a miniassay. *J. Biol. Inorg. Chem.* **16**, 977–989. (doi:10.1007/s00775-011-0798-3)
10. Wang W-C, Mao H, Ma D-D, Yang W-X. 2014 Characteristics, functions, and applications of metallothionein in aquatic vertebrates. *Front. Mar. Sci.* **1**, 34. (doi:10.3389/fmars.2014.00034)
11. Ragusa MA, Nicosia A, Costa S, Cuttitta A, Gianguzza F. 2017 Metallothionein gene family in the sea urchin *Paracentrotus lividus*: gene structure,

differential expression and phylogenetic analysis. *Int. J. Mol. Sci.* **18**, 812. (doi:10.3390/ijms18040812)

12. Purać J, Nikolić TV, Kojić D, Ćelić AS, Plavša JJ, Blagojević DP, Petri ET. 2019 Identification of a metallothionein gene in honey bee *Apis mellifera* and its expression profile in response to Cd, Cu and Pb exposure. *Mol. Ecol.* **28**, 731–745. (doi:10.1111/mec.14984)
13. Lastowski-Perry D, Otto E, Maroni G. 1985 Nucleotide sequence and expression of a *Drosophila* metallothionein. *J. Biol. Chem.* **260**, 1527–1530.
14. Mokdad R, Debec A, Wegnez M. 1987 Metallothionein genes in *Drosophila melanogaster* constitute a dual system. *Proc. Natl Acad. Sci. USA* **84**, 2658–2662. (doi:10.1073/pnas.84.9.2658)
15. Adams MD *et al.* 2000 The genome sequence of *Drosophila melanogaster*. *Science* **287**, 2185–2195. (doi:10.1126/science.287.5461.2185)
16. Egli D, Selvaraj A, Yepiskoposyan H, Zhang B, Hafen E, Georgiev O, Schaffner W. 2003 Knockout of 'metal-responsive transcription factor' MTF-1 in *Drosophila* by homologous recombination reveals its central role in heavy metal homeostasis. *EMBO J.* **22**, 100–108. (doi:10.1093/emboj/cdg012)
17. Egli D, Domènech J, Selvaraj A, Balamurugan K, Hua H, Capdevila M, Georgiev O, Schaffner W, Atrian S. 2006 The four members of the *Drosophila* metallothionein family exhibit distinct yet overlapping roles in heavy metal homeostasis and detoxification. *Genes Cells* **11**, 647–658. (doi:10.1111/j.1365-2443.2006.00971.x)
18. Atanesyan L, Günther V, Celniker SE, Georgiev O, Schaffner W. 2011 Characterization of MtnE, the fifth metallothionein member in *Drosophila*. *J. Biol. Inorg. Chem.* **16**, 1047. (doi:10.1007/s00775-011-0825-4)
19. Pérez-Rafael S, Kurz A, Guirola M, Capdevila M, Palacios Ò, Atrian S. 2012 Is MtnE, the fifth *Drosophila* metallothionein, functionally distinct from the other members of this polymorphic protein family? *Metallomics* **4**, 342–349. (doi:10.1039/c2mt00182a)
20. Atrian S. 2009 Metallothioneins in Diptera. In *Metallothioneins and related chelators: metal ions in life sciences* (eds A Sigel, H Sigel, RKO Sigel), pp. 155–181. Cambridge, UK: Royal Society of Chemistry Publishing. (doi:10.1039/9781847559531)
21. Guirola M, Naranjo Y, Capdevila M, Atrian S. 2011 Comparative genomics analysis of metallothioneins in twelve *Drosophila* species. *J. Inorg. Biochem.* **105**, 1050–1059. (doi:10.1016/j.jinorgbio.2011.05.004)
22. Liu Y, Wu H, Kou L, Liu X, Zhang J, Guo Y, Ma E. 2014 Two metallothionein genes in *Oxya chinensis*: molecular characteristics, expression patterns and roles in heavy metal stress. *PLoS ONE* **9**, e112759. (doi:10.1371/journal.pone.0112759)
23. Merritt TJ, Bewick AJ. 2017 Genetic diversity in insect metal tolerance. *Front. Genet.* **8**, 172. (doi:10.3389/fgene.2017.00172)
24. Janssens TK, Roelofs D, Van Straalen NM. 2009 Molecular mechanisms of heavy metal tolerance and evolution in invertebrates. *Insect Sci.* **16**, 3–18. (doi:10.1111/j.1744-7917.2009.00249.x)
25. Thomas GW *et al.* 2020 Gene content evolution in the arthropods. *Genome Biol.* **21**, 15. (doi:10.1186/s13059-019-1925-7)
26. Yin C *et al.* 2015 InsectBase: a resource for insect genomes and transcriptomes. *Nucleic Acids Res.* **44**, D801–D807. (doi:10.1093/nar/gkv1204)
27. Roskov Y *et al.* (eds). 2019 Species 2000 & ITIS Catalogue of Life, 2019 Annual Checklist. Leiden, The Netherlands: Naturalis. See www.catalogueoflife.org/annual-checklist/2019.
28. Misof B *et al.* 2014 Phylogenomics resolves the timing and pattern of insect evolution. *Science* **346**, 763–767. (doi:10.1126/science.1257570)
29. Finet C, Slavik K, Pu J, Carroll SB, Chung H. 2019 Birth-and-death evolution of the fatty acyl-CoA reductase (FAR) gene family and diversification of cuticular hydrocarbon synthesis in *Drosophila*. *Genome Biol. Evol.* **11**, 1541–1551. (doi:10.1093/gbe/evz094)
30. Vakirlis N, Carvunis A-R, McLysaght A. 2020 Synteny-based analyses indicate that sequence divergence is not the main source of orphan genes. *eLife* **9**, e53500. (doi:10.7554/eLife.53500)
31. Ranz JM, Casals F, Ruiz A. 2001 How malleable is the eukaryotic genome? Extreme rate of chromosomal rearrangement in the genus *Drosophila*. *Genome Res.* **11**, 230–239. (doi:10.1101/gr.162901)
32. Stewart NB, Rogers RL. 2019 Chromosomal rearrangements as a source of new gene formation in *Drosophila yakuba*. *PLoS Genet.* **15**, e1008314. (doi:10.1371/journal.pgen.1008314)
33. Adler PH, Yadamsuren O, Procunier WS. 2016 Chromosomal translocations in black flies (Diptera: Simuliidae)—facilitators of adaptive radiation? *PLoS ONE* **11**, e0158272. (doi:10.1371/journal.pone.0158272)
34. Artemov GN, Peery AN, Jiang X, Tu Z, Stegniy VN, Sharakhova MV, Sharakhov IV. 2017 The physical genome mapping of *Anopheles albimanus* corrected scaffold misassemblies and identified interarm rearrangements in genus *Anopheles*. *G3* **7**, 155–164. (doi:10.1534/g3.116.034959)
35. Calvete O, González J, Betrán E, Ruiz A. 2012 Segmental duplication, microinversion, and gene loss associated with a complex inversion breakpoint region in *Drosophila*. *Mol. Biol. Evol.* **29**, 1875–1889. (doi:10.1093/molbev/mss067)
36. Pringle EG, Baxter SW, Webster CL, Papanicolaou A, Lee SF, Jiggins CD. 2007 Synteny and chromosome evolution in the Lepidoptera: evidence from mapping in *Heliconius melpomene*. *Genetics* **177**, 417–426. (doi:10.1534/genetics.107.073122)
37. d'Alencon E *et al.* 2010 Extensive synteny conservation of holocentric chromosomes in Lepidoptera despite high rates of local genome rearrangements. *Proc. Natl Acad. Sci. USA* **107**, 7680–7685. (doi:10.1073/pnas.0910413107)
38. Kanost MR *et al.* 2016 Multifaceted biological insights from a draft genome sequence of the tobacco hornworm moth, *Manduca sexta*. *Insect Biochem. Mol. Biol.* **76**, 118–147. (doi:10.1016/j.ibmb.2016.07.005)
39. Finet C, Decaras A, Armisen D, Khila A. 2018 The *achaete–scute* complex contains a single gene that controls bristle development in the semi-aquatic bugs. *Proc. R. Soc. B* **285**, 20182387. (doi:10.1098/rspb.2018.2387)
40. Mandrioli M, Melchiori G, Panini M, Chiesa O, Giordano R, Mazzoni E, Manicardi GC. 2019 Analysis of the extent of synteny and conservation in the gene order in aphids: a first glimpse from the *Aphis glycines* genome. *Insect Biochem. Mol. Biol.* **113**, 103228. (doi:10.1016/j.ibmb.2019.103228)
41. Li Z, Tiley GP, Galuska SR, Reardon CR, Kidder TI, Rundell RJ, Barker MS. 2018 Multiple large-scale gene and genome duplications during the evolution of hexapods. *Proc. Natl Acad. Sci. USA* **115**, 4713–4718. (doi:10.1073/pnas.1710791115)
42. Nakatani Y, McLysaght A. 2019 Macrosynteny analysis shows the absence of ancient whole-genome duplication in lepidopteran insects. *Proc. Natl Acad. Sci. USA* **116**, 1816–1818. (doi:10.1073/pnas.1817937116)
43. Roelofs D, Zwaenepoel A, Sistermans T, Nap J, Kampfraath AA, Van de Peer Y, Ellers J, Kraaijeveld K. 2020 Multi-faceted analysis provides little evidence for recurrent whole-genome duplications during hexapod evolution. *BMC Biol.* **18**, 57. (doi:10.1186/s12915-020-00789-1)
44. Fischer HM, Wheat CW, Heckel DG, Vogel H. 2008 Evolutionary origins of a novel host plant detoxification gene in butterflies. *Mol. Biol. Evol.* **25**, 809–820. (doi:10.1093/molbev/msn014)
45. Briscoe AD *et al.* 2013 Female behaviour drives expression and evolution of gustatory receptors in butterflies. *PLoS Genet.* **9**, e1003620. (doi:10.1371/journal.pgen.1003620)
46. Pinharanda A, Martin S, Barker S, Davey J, Jiggins C. 2017 The comparative landscape of duplications in *Heliconius melpomene* and *Heliconius cydno*. *Heredity* **118**, 78–87. (doi:10.1038/hdy.2016.107)
47. Smith G, Briscoe AD. 2015 Molecular evolution and expression of the CRAL_TRIO protein family in insects. *Insect Biochem. Mol. Biol.* **62**, 168–173. (doi:10.1016/j.ibmb.2015.02.003)
48. Westerman EL *et al.* 2018 *Aristaless* controls butterfly wing color variation used in mimicry and mate choice. *Curr. Biol.* **28**, 3469–3474.e4. (doi:10.1016/j.cub.2018.08.051)
49. Milošević T, Đurić M, Milošević N. 2014 Accumulation of heavy metals in flowers of fruit species. *Water Air Soil Pollut.* **225**, 2019. (doi:10.1007/s11270-014-2019-5)
50. Maroni G, Wise J, Young J, Otto E. 1987 Metallothionein gene duplications and metal tolerance in natural populations of *Drosophila melanogaster*. *Genetics* **117**, 739–744.
51. Meyer JL, Hoy MA, Jeyaprakash A. 2006 Insertion of a yeast metallothionein gene into the model insect *Drosophila melanogaster* (Diptera: Drosophilidae) to assess the potential for its use in genetic improvement programs with natural enemies. *Biol. Control* **36**, 129–138. (doi:10.1016/j.biocontrol.2005.08.007)

52. Otto E, Allen J, Young J, Palmiter R, Maroni G. 1987 A DNA segment controlling metal-regulated expression of the *Drosophila melanogaster* metallothionein gene *Mtn*. *Mol. Cell. Biol.* **7**, 1710–1715. (doi:10.1128/MCB.7.5.1710)
53. Günther V, Lindert U, Schaffner W. 2012 The taste of heavy metals: gene regulation by MTF-1. *Biochim. Biophys. Acta Mol. Cell Res.* **1823**, 1416–1425. (doi:10.1016/j.bbamcr.2012.01.005)
54. Sims HI, Chirn G-W, Marr MT. 2012 Single nucleotide in the MTF-1 binding site can determine metal-specific transcription activation. *Proc. Natl Acad. Sci. USA* **109**, 16 516–16 521. (doi:10.1073/pnas.1207737109)
55. Robbins A, McRee D, Williamson M, Collett S, Xuong N, Furey W, Wang B, Stout C. 1991 Refined crystal structure of Cd, Zn metallothionein at 2.0 Å resolution. *J. Mol. Biol.* **221**, 1269–1293. (doi:10.1016/0022-2836(91)90933-w)
56. Rigby Duncan KE, Stillman MJ. 2007 Evidence for noncooperative metal binding to the α domain of human metallothionein. *FEBS J.* **274**, 2253–2261. (doi:10.1111/j.1742-4658.2007.05762.x)
57. Hall TA. 1999 BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. In *Nucleic Acids Symp. Ser.* **41**, 95–98.
58. Edgar RC. 2004 MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinform.* **5**, 113. (doi:10.1186/1471-2105-5-113)
59. Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010 New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307–321. (doi:10.1093/sysbio/syq010)
60. Poelchau MF, Chen MM, Lin YY, Childers CP. 2018 Navigating the i5k Workspace@NAL: a resource for arthropod genomes. *Methods Mol. Biol.* **1757**, 557–577. (doi:10.1007/978-1-4939-7737-6_18)
61. Thurmond J *et al.* 2018 FlyBase 2.0: the next generation. *Nucleic Acids Res.* **47**, D759–D765. (doi:10.1093/nar/gky1003)
62. Yang J, Zhang Y. 2015 Protein structure and function prediction using I-TASSER. *Curr. Protoc. Bioinform.* **52**, 5.8.1–5.8.15. (doi:10.1002/0471250953.bi0508s52)
63. Braun W, Vasak M, Robbins A, Stout C, Wagner G, Kägi J, Wüthrich K. 1992 Comparison of the NMR solution structure and the X-ray crystal structure of rat metallothionein-2. *Proc. Natl Acad. Sci. USA* **89**, 10 124–10 128. (doi:10.1073/pnas.89.21.10124)
64. Zhang C, Freddolino PL, Zhang Y. 2017 COFACTOR: improved protein function prediction by combining structure, sequence and protein–protein interaction information. *Nucleic Acids Res.* **45**, W291–W299. (doi:10.1093/nar/gkx366)
65. Luo M, Finet C, Cong H, Wei H-y, Chung H. 2020 Data from: The evolution of insect metallothioneins. Dryad Diigital Repository. (doi:10.5061/dryad.pc866t1mh)
66. Luo M, Finet C, Cong H, Wei H-Y, Chung H. 2020 The Evolution of Insect Metallothioneins. *bioRxiv*. (https://doi.org/10.1101/2020.06.25.172213)