Diversity, duplication, and genomic organization of homeobox genes in Lepidoptera

Peter Mulhair, Liam Crowley, Douglas H Boyes, et al.

Genome Res. published online December 8, 2022
Access the most recent version at doi:10.1101/gr.277118.122

Diversity, duplication and genomic organisation of homeobox genes in Lepidoptera

Peter O. Mulhair[1], Liam Crowley[1], Douglas H. Boyes[1,2], Amber Harper[1,3], Owen T. Lewis[1], Darwin Tree of Life Consortium[4], Peter W.H. Holland[1*]

[1]Department of Biology, University of Oxford, 11a Mansfield Road, Oxford OX1 3SZ, UK
[2]UK Centre for Ecology & Hydrology, Wallingford, OX10 8BB, UK
[3]Current address: Department of Biological and Medical Sciences, Faculty of Health and Life Sciences, Oxford Brookes University, Oxford, UK
**Corresponding author: Peter W.H. Holland, peter.holland@biology.ox.ac.uk*

Running title: Evolution of homeobox genes in Lepidoptera

[4] The list of Darwin Tree of Life Consortium members and affiliations is listed at the end of this paper.

Abstract

Homeobox genes encode transcription factors with essential roles in patterning and cell fate in developing animal embryos. Many homeobox genes, including Hox and NK genes, are arranged in gene clusters, a feature likely related to transcriptional control. Sparse taxon sampling and fragmentary genome assemblies mean that little is known about dynamics of homeobox gene evolution across Lepidoptera, or how changes in homeobox gene number and organisation relate to diversity in this large order of insects. Here we analyse an extensive dataset of high-quality genomes to characterise the number and organisation of all homeobox genes in 123 species of Lepidoptera from 23 taxonomic families. We find most Lepidoptera have around 100 homeobox loci, including an unusual Hox gene cluster in which the *lab* gene is repositioned and the *ro* gene is next to *pb*. A topologically associating domain spans much of the gene cluster, suggesting deep regulatory conservation of the Hox cluster arrangement in this insect order. Most Lepidoptera have four Shx genes, divergent *zen*-derived loci, but these loci underwent dramatic duplication in several lineages with some moths having over 165 homeobox loci in the Hox gene cluster; this expansion is associated with local LINE element density. In contrast, the NK gene cluster content is more stable, although there are differences in organisation compared to other insects, and major rearrangements within butterflies. Our analysis represents the first description of homeobox gene content across the order Lepidoptera, exemplifying the potential of newly generated genome assemblies for understanding genome and gene family evolution.

Introduction

Lepidoptera (moths and butterflies) are one of the four mega-diverse insect orders, with over 150,000 described species. Lepidoptera belong within the Endopterygota, meaning they undergo complete metamorphosis with development proceeding from a motile phytophagous larva to a pupal stage to a reproductive imago (adult). The imaginal stage is easily recognisable and typically follows a characteristic body plan: two pairs of scale-covered

membranous wings, six walking legs, filamentous antennae and a tube-like proboscis. There are, however, variations and exceptions. For example, in some moths the females are flightless with reduced wings; in many butterflies four legs rather than six are used for walking; anntenal morphology varies with clubbed, lamellate or plumose structure; in several moth species the larvae are fully aquatic; and in the family Micropterygidae the adults have biting rather than sucking mouthparts. Many variations and adaptations are hypothesised to have been driven by co-evolution with plants, driving novelties in egg laying behaviour, larval phenotype, and feeding strategies in both larvae and adults (Wiens et al. 2015; Mitter et al. 2017; Kawahara et al. 2019).

Associating evolutionary change in form or behaviour to changes in underlying loci is not straightforward, but insights can come from correlations between patterns in molecular evolution and changes in phenotype. Homeobox genes are candidates for loci in which molecular change may cause or facilitate evolutionary change to form and structure of animals, because most homeobox genes play regulatory roles in development. For example, the Hox genes, a subset of homeobox genes, encode transcription factors that control spatial identity along the anteroposterior axis in embryonic development and their number differs between animal lineages. There was an increase in Hox gene number on the stem lineage of bilaterian animals, when a head-to-tail axis evolved to dominate the body plan (Finnerty and Martindale 1998; Nong et al. 2020; Holland 2015); there was also an increase in the early evolution of vertebrates, traceable to genome duplication (Soshnikova et al. 2013; Aase-Remedios and Ferrier 2021). Hox genes are usually arranged in gene clusters, but these clusters have been secondarily broken or dispersed in some evolutionary lineages concomitant with changes to developmental pathways (Ferrier and Holland 2002); conversely, clusters have been further compacted in vertebrates in association with additional gene regulatory controls and the emergence of fins and limbs (Duboule 2007). We wished to address if changes to the homeobox complement in Lepidoptera were associated with phenotypic change.

In Lepidoptera, lack of high contiguity, chromosomal-scale genome assemblies have hampered studies into the structure and evolution of the Hox gene cluster, so the extent of gene cluster compaction, cluster integrity and the precise gene order remains unclear. One discovery was the presence of at least 11 divergent homeobox loci within the Hox gene cluster of the Silkworm, *Bombyx mori* (Chai et al. 2008), all located between the *zen* and *pb* Hox genes. This presence of unusual 'Special homeobox' (Shx) genes within the Hox gene cluster was later confirmed in several other Lepidoptera, most of which were found to possess four Shx genes, *ShxA*, *ShxB*, *ShxC* and *ShxD*, derived by tandem duplication and divergence from *zen* (Ferguson et al. 2014). These studies also highlighted *B. mori* as an aberrant outlier to the usual pattern, with the larger number of Shx genes reflecting further tandem duplication of *ShxD*. *Triodia sylvina* (Orange Swift Moth, family Hepialidae) was also noted as unusual, as it seemed to lack Shx genes altogether, although tentative evidence for *zen* duplication was found (Ferguson et al. 2014). We wished to refine when Shx genes arose, and also test if Shx expansion in *Bombyx* is unique.

Although the roles of Shx genes are not yet fully understood, studies in *Pararge aegeria* (Speckled Wood Butterfly) have shown expression in the extraembryonic serosa and suggested functions in extraembryonic membrane patterning (Ferguson et al. 2014). It should also be noted that relatively few species were compared in these initial surveys due to lack of genomic data, hence patterns of Shx gene evolution were poorly resolved. Outside of the Hox gene cluster, even less is known about evolution of homeobox genes across Lepidoptera. For example, the NK genes are members of the ANTP class, like Hox genes, and are arranged in a compact gene cluster in Diptera and Coleoptera (Jagla et al. 2001; Garcia-Fernàndez 2005; Butts et al. 2008); these genes are implicated in mesoderm development but their evolution has not been analysed comprehensively in Lepidoptera (Ranz et al. 2022). The same can be said for the many dispersed homeobox genes that are not arranged in gene clusters and are implicated in a wide diversity of developmental roles (Ferrier 2016). We aimed to assess the extent of homeobox gene clustering in Lepidoptera, beyond the Hox cluster.

Until recently, analysis of the copy number, organisation and molecular evolution of homeobox genes across a whole insect order has not been feasible due to limited sampling of species and, for study of clustered homeobox genes, the highly fragmented nature of many genome assemblies. Dense sampling of lepidopteran species in the Darwin Tree of Life Project (The Darwin Tree of Life Project Consortium 2022), has generated chromosome-level genome assemblies across a wide phylogenetic coverage. Analysing these data, we present an order-wide description of the homeobox gene content in Lepidoptera. Using chromosome-level genome assemblies for 123 lepidopteran genomes from 23 taxonomic families, we identified all homeobox genes from their characteristic homeodomain, determined their genomic organisation into gene clusters and traced their patterns and pathways of duplication and loss.

Results

Classification of all Lepidoptera homeobox genes

We identified all homeobox gene loci in the genomes of 123 lepidopteran species, including 87 moths and 36 butterfly species (Supplementary Table S1) (Mulhair et al. 2022). To place our analyses in an evolutionary context, we also constructed a phylogenetic tree of the species analysed using 2,262 BUSCO genes (Methods; Figure 1A, Supplementary Figure S1). Homeobox sequences were then classified using the characteristic homeodomain and a combination of reciprocal best BLAST and molecular phylogenetic analysis: this ‘total’ collection of homeobox loci could include functional genes, partial genes and pseudogenes.

We find that the catalogue of homeobox loci is relatively stable across Lepidoptera (Figure 1B and C), with most species possessing around 100 homeobox sequences. However, certain lineages and species showed marked increases in homeobox counts, resulting mainly from duplications within individual homeobox gene classes. The main contributors to these increases are large expansions within the Hox gene cluster in some clades or smaller scale duplications of PRD class genes. Homeobox gene loss has also

occurred. For example, the *HHEX (Hhex)* gene of the ANTP class is absent from the genomes of all three *Pieris* species sequenced, consistent with a loss in this clade (Figure 1B). We deduce that the *ShxD* gene was lost in the genus *Melitaea*, as it is absent in both *Melitaea cinxia* (Glanville Fritillary butterfly), consistent with an earlier report (Ferguson et al. 2014), and *Melitaea athalia* (Heath Fritillary Butterfly). Similarly, we do not find the *ShxD* gene in any of the eight Lycaenidae species in our dataset (*Lycaena phlaeas, Celastrina argiolus, Glaucopsyche alexis, Plebejus argus, Cyaniris semiargus, Aricia agestis, Lysandra bellargus* and *Lysandra coridon*), implying that this gene was also lost early in the evolution of the family Lycaenidae. Some homeobox genes, such as *Mkx* (orthologous to Dmel\CG11617) of the TALE class, were lost many times independently across Lepidoptera (Figure 1B).

Using a representative set of seven species, we examined expression levels for each homeobox gene using female whole body RNA-seq (Supplementary Figure S2). We find clear evidence for expression of Hox genes and Shx genes, with particularly strong expression of ShxC, consistent expression of homeobox genes in the SINE, TALE, CUT, PROS, ZF and CERS classes, and variable expression of PRD class and NK homeobox genes.

Rearrangement of the Hox gene cluster

Within insects, the Hox gene cluster generally comprises 10 homeobox genes arranged in a specific order reflecting their evolutionary origin by tandem gene duplication: *lab, pb, zen, Dfd, Scr, Antp, ftz, Ubx, abd-A, Abd-B*. The cluster may be split, as in many *Drosophila* species (Negre and Ruiz 2007; Duboule 2007), individual genes may be inverted, and the *zen* gene may be duplicated (e.g. *zen*, *zen2* and *bcd* in *D. melanogaster*), but radical gene order changes are rare, documented only within individual species or close relatives (Negre et al. 2005). A difficulty in studying gene order is that intergenic distances may be large and many genome assemblies do not provide long-range linkage information. Using chromosome-level gene assemblies (Supplementary Table S1), we have determined the

178 structure of the Hox gene cluster in 123 Lepidoptera genomes, providing the first
179 comprehensive description of the cluster evolution across this order.

180 We found all Hox genes on a single scaffold for 115/123 genomes (Supplementary
181 Figure S3). In all Lepidoptera we analysed we found the canonical *lab*, *pb*, *Dfd*, *Scr*, *Antp*,
182 *Ubx*, *abd-A* and *Abd-B* homeotic genes, plus the divergent Hox-derived genes *zen* and *ftz*,
183 along with gene order, orientation, and intergenic distances. In most Lepidoptera, excepting
184 some 'basal' lineages, we also found four distinct Shx genes (*ShxA* to *ShxD*) between *zen*
185 and *pb* (Figure 1 and 2, Supplementary Figure S3), as previously noted for a smaller sample
186 of species (Ferguson et al. 2014). The structure of the Hox cluster for *Autographa gamma*
187 (Silver Y Moth; Boyes et al. 2022c) shown in Figure 2B reflects the general structure found in
188 most lepidopteran species.

189 When compared to other insect orders, two rearrangements are apparent. First, we
190 consistently find a non-Hox homeobox gene, *ro* (rough), in close association with the gene
191 cluster. Across almost all Lepidoptera species, the *ro* gene is adjacent to *pb*, in the genomic
192 location where *lab* or its orthologue is found in most species (shown for *A. gamma* in Figure
193 2). Second, the *lab* gene has been translocated to a distant genomic location beyond *Abd-B*.
194 This dissociation of *lab* is consistent with a split between *lab* and other Hox genes previously
195 reported in *B. mori,* although the position of *lab* was unresolved in this earlier work
196 (Yasukochi et al. 2004; Chai et al. 2008). In *A. gamma* the *lab* gene is approximately 7 Mb
197 from *Abd-B*, while the main part of the Hox cluster spans 1.22 Mb from *Abd-B* to *ro*. We find
198 the Hox gene cluster (excluding *lab*) in Lepidoptera ranges from 1Mb in *Papilio machaon*
199 (Swallowtail Butterfly) to 6.8Mb in *Euproctis similis* (Yellow-tail Moth). A further inversion of
200 the *ro* gene occurred within the *Pieris* clade resulting in relocation of *ro* to between *lab* and
201 *Abd-B* (Supplementary Figure S3).

202 **An evolutionarily conserved topologically associating domain around the Hox cluster**

203 To assess whether the rearrangements in gene order could be associated with changes in
204 regulation of Hox genes, we used Hi-C data to annotate topologically associating domains

(TADs) across the genome. These data can reveal the 3D organisation of the chromatin and, at least in some cases, highlight regions of the genome under common regulatory constraints (Szabo et al. 2019; Schoenfelder and Fraser 2019). Given that the purpose of the Hi-C sequencing of these species was to assist in genome assembly (Lawniczak et al. 2022), the depth of Hi-C is lower than in some other studies (Liao et al. 2021), ranging from ~35 to 52 million paired-end reads (Supplementary Table S2). Nonetheless, we found this sequencing depth sufficient for the analysis, revealing TADs in lepidopteran genomes which were visualised at 5kb resolution using HiCExplorer (Ramírez et al. 2018). To our knowledge, this is the first such analysis for Lepidoptera genomes, and one of the few assessments of chromatin accessibility around invertebrate Hox gene clusters (Acemel et al. 2017). In a sample of nine species representing diverse families across Lepidoptera (Figure 3A), we observe strong evidence of an evolutionarily conserved, prominent TAD covering most of the Hox gene cluster from *pb* to *Abd-B* (Figure 3A & B). This TAD was also observed in species with a large increase in copy number within the Hox cluster (see “Independent tandem duplication of Shx genes” below) (Supplementary Figure S4A). In all species analysed *lab* and *ro* are located outside the distinct TAD (Figure 3A). Assessing the wider chromosomal organisation in *Pheosia gnoma* (Lesser Swallow Prominent), it is clear that there is a high degree of contact within the Hox-containing TAD relative to the rest of the chromosome (Figure 3B). While genome-wide conservation of TADs between species has been questioned (Eres and Gilad 2021), we argue that the strong and consistent signal for physical contacts across the Hox cluster in diverse moths and butterflies is evidence for a conserved TAD around a cluster of developmentally important genes. We make no assessment of possible conservation of other TADs in lepidopteran genomes. Outside the Hox gene cluster, the general structure of TADs across this chromosome appears similar to the pattern observed in *Drosophila*, with TADs representing condensed internal interactions and larger compartments showing long-range interactions between these domains (Figure 3B & C) (Sexton et al. 2012; Ulianov et al. 2016; Szabo et al. 2018; Liao et al. 2021).

Origin, duplication and loss of Shx genes

As in most Lepidoptera, four distinct Shx genes were identified between *zen* and *pb* in the Hox gene cluster of *A. gamma* (Figure 2, Supplementary Figure S3). Phylogenetic analysis supports the derivation of Shx genes by tandem duplication and sequence divergence from *zen* (see Ferguson et al. 2014; Figure 4A). We also find rapid sequence divergence within the homeodomain of these genes following duplication from *zen*, as previously described (Figure 4A; Ferguson et al 2014). Shx genes are found in representatives of the Erebidae, Nymphalidae, Sphingidae, Noctuidae, Lycaenidae, Pieridae, Papilionidae, Notodontidae, Drepanidae, Hesperiidae, Tortricidae, Geometridae, Sesiidae, Blastobasidae, Depressariidae, Crambidae, Pterophoridae, Pyralidae, Tineidae, Ypsolophidae, Cossidae and Zygaenidae families, but we do not identify these genes in Micropterigidae (Figure 2A, Supplementary Figure S3). Instead, extra copies of *zen* (four in addition to the original *zen*) were found in a species from the family Micropterigidae (*Neomicropteryx cornuta*). These loci, which are located outside the Hox cluster beyond the location of *Abd-B*, group outside of the Shx genes in a molecular phylogenetic analysis: two with the *zen* clade and three closer to the *lab* clade (Figure 4A). We suggest all are derived from *zen*. They display higher rates of substitution in the homeodomain compared to the other *zen* genes analysed, which may underlie erroneous placement of some genes closer to *lab*.

The *ShxD* gene was lost several times across the lepidopteran phylogeny. One loss event is shared by the Lycaenidae species (the ‘blue’ butterflies), suggesting gene loss along the ancestral branch of this diverse family. Loss of *ShxD* in these species is associated with longer branch lengths in the remaining Shx genes (*ShxA-C*) in a phylogenetic analysis (Figure 4A). The significantly increased rate of substitution in the homeodomain of the three remaining Shx genes (*ShxA*, *ShxB* and *ShxC*) following loss of the ShxD gene was confirmed by assessing pairwise sequence identity between Lycaenidae species and non-Lycaenidae species (Figure 4B).

Independent tandem duplication of Shx genes

As noted above, the number of Shx genes in ditryisian lepidopterans is usually four, or three in those taxa that have lost *ShxD*. However, there are some notable examples of Shx gene duplication. In earlier work using limited sampling and fragmentary genome assemblies, the large number of Shx loci in *Bombyx mori* was considered an exception to the normal pattern (Chai et al. 2008; Ferguson et al. 2014). The expanded sampling generated by the Darwin Tree of Life project reveals a more complex pattern of evolution. While presence of four Shx genes is still the norm for Lepidoptera, we find multiple independent examples of dramatic Shx gene number expansion (Figure 1B, Supplementary Figure S3). In 18 species of moth (*Zeuzera pyrina, Blastobasis lacticolella, Blastobasis adustella, Parapoynx stratiotata, Idaea aversata, Phalera bucephala, Euproctis similis, Schrankia costaestrigalis, Spilarctia lutea, Spilosoma lubricipeda, Eilema depressum, Eilema sororculum, Mythimna ferrago, Mythimna impura, Noctua pronuba, Noctua janthe, Noctua fimbriata* and *Apamea monoglypha*) and one butterfly species (*Aporia crataegi*) a large number of homeobox loci were found between *zen* and *pb*, each representing extensive tandem duplication of Shx genes (Supplementary Figure S2, Supplementary Figure S5). In these species the copy number ranges from 9 copies of Shx genes in *Noctua pronuba*, to 51 Shx copies in *Mythimna impura*, up to 165 Shx loci in *Apamea monoglypha*, the largest number observed. These species have a mean of 32 copies of *zen*/Shx and a median of 20 copies. The rate of sequence divergence of the Shx genes following tandem duplication varies between species, with tandemly duplicated copies in three species showing significantly lower pairwise identity (larger sequence change compared to the distribution of pairwise identity in non-duplicated orthologues), duplicated copies from three species showing higher pairwise identity (possibly reflecting recent duplication), and four species showing no significant difference.

This demonstrates that the Shx expansion phenomenon is more widespread across Lepidoptera than previously recognized. In some cases, we observe tandem duplication of Shx genes in closely related species, for example two *Blastobasis* species, three *Noctua*

species and two *Mythimna* species, suggesting these events occurred in the common ancestor of each of these lineages, or that these lineages are prone to Shx duplication. In total, we detect at least 11 cases of independent expansion of the Shx genes, in addition to the previously recognized *B. mori* expansion. We rarely see clearly intermediate cases: we detect either a conservative pattern of 3 to 6 Shx genes, or a dramatically expanded set of Shx genes.

We investigated whether retrotransposon activity may have impacted the copy number variation observed. Retrotransposons, particularly LINE elements, can facilitate non-allelic homologous recombination, resulting in segmental duplications and gene cluster expansions (Startek et al. 2015; Janoušek et al. 2016; Thybert et al. 2018). Repeat content across the whole genomes of 66 representative species was estimated using a combination of RepeatModeler and RepeatMasker pipelines (see Material and Methods, Supplementary Figure S6). To test the relation between transposon activity and the Hox gene cluster, transposable element (TE) density was annotated in windows of 5,000 bases within the Hox cluster (*lab* was excluded from this analysis due to its distant position). Density of the major classes of TEs (LINEs, SINEs, LTR and DNA) were compared between the region containing the Shx genes and the remaining Hox cluster. Significantly increased density of LINE elements was observed within the Shx gene region relative to the rest of the Hox gene cluster in 14 of 19 species with large tandem duplications (Wilcoxon rank-sum test; $p<0.05$, Bonferroni correction) (Figure 5A). These 14 species were: *Zeuzera pyrina, Blastobasis lacticolella, Blastobasis adustella*, *Euproctis similis*, *Spilarctia lutea, Spilosoma lubricipeda, Eilema depressum, Eilema sororculum, Mythimna ferrago, Mythimna impura, Noctua janthe, Noctua fimbriata* and *Apamea monoglypha* (Figure 5B). Further examining the correlation between Shx expansion and LINE proliferation in the species *Zeuzera pyrina*, which has 25 copies of *ShxA* (Supplementary Figure S3, Supplementary Figure S7A), we see that there is clear evidence for tandem duplication of specific LINE elements (LINE/CR1), which are all in the same orientation and evenly interspersed between the *ShxA* copies (Supplementary Figure S7B). The five species with large tandem duplications but no LINE enrichment were:

Noctua pronuba, *Parapoynx stratiotata*, *Idaea inversata*, *Phalera bucephala* and *Aporia crataegi*. For several species there is clear evidence that repeat elements were tandemly duplicated along with Shx loci. For example, *Parapoynx stratiotata*, with 16 ShxD copies, has a repeated array of Low_complexity, Simple_repeat and LINE/L2 elements between each ShxD. Quite different patterns are seen in *Phalera bucephala* and *Zeuzera pyrina*. There is no association, beyond the presence of LINEs, between repeat type, repeat number and which Shx gene is duplicated.

Other homeobox gene clusters

The Hox genes are the best-studied clustered homeobox genes, but other examples also occur. A cluster of three neuronally-expressed homeobox genes from the PRD class - Homeobrain (*hbn*), Retinal Homeobox (*Rx*) and Orthopedia (*otp*) - has been conserved in most animal lineages since the cnidarian-bilaterian ancestor (Mazza et al. 2010). The gene cluster has also been found in *Drosophila*, and representatives of Hymenoptera and Coleoptera, with a conserved gene order and comparable intergenic distances (Walldorf et al. 2000; Mazza et al. 2010). Across the lepidopteran species in this study, we also find that the cluster is conserved with the same gene order (Supplementary Figure S8). Genomic distances between genes are larger in lepidopteran species than in other insects studied to date, with an average overall cluster length of 348 kb. While gene order is conserved, transcriptional orientation varies between species.

In *Drosophila*, several 'NK' genes form a compact homeobox gene cluster comprising *tin* (*NK4*), *bap* (*NK3*), two '*Lbx'* genes (*lbl*, *lbe*), *C15* (*Tlx*) and *slou* (*NK1*) (Jagla et al. 2001; Luke et al. 2003; Garcia-Fernàndez 2005). Other NK-related genes are found more distantly and may have been translocated away, including *Dr* (*Msx)*, *ems* (*Emx*) and *Hmx* (*NK5*). Other groupings of NK genes are found in other animal genomes (Jagla et al. 2001; Luke et al. 2003; Garcia-Fernàndez 2005). In contrast to Hox gene clusters, we find the NK gene cluster has undergone extensive gene order changes during insect evolution (Figure 6A). Across all insect orders we find tight linkage between *tin*, *bap*, and *Lbx*; we also find *Dr* is

closely linked in several orders, but not Diptera represented by *Drosophila*. Outside these genes, there is considerable variation between orders.

The organisation of the NK gene cluster in *A. gamma* (Silver Y Moth) is typical for Lepidoptera (Figure 6B). We find a 'core' of five homeobox genes (two *Msx*, *tin*, *bap* and *Lbx*) spanning ~370kb, plus linkage to *C15* on one side and *slou*, *Hmx* and *ems* on the other (Figure 6B). The arrangement of these genes is generally conserved across most Lepidoptera species (Figure 6, Supplementary Figure S9). However, rearrangements within the cluster are observed in some butterfly lineages. For example, in the three *Pieris* species, the order of the *tin*/bap/*Lbx*/*Dr* core cluster is inverted in all species, *C15* is found on a separate chromosome, and *Abox* and *Bari* homeobox genes are located close to each end of the cluster (Figure 6C, Supplementary Figure S9). Rearrangements are also found in both Lycaenidae and Nymphalidae, with different gene orders suggesting independent rearrangements (Figure 6C). We infer that a series of translocation and inversion events have occurred independently. In lineages such as the *Pieris* butterflies, these changes in the structure of the NK gene cluster reflect general trends of genome remodelling (Hill et al. 2019). The changes within the NK cluster within butterflies represent at least seven likely rearrangement events, contrasting to the general stability in gene order observed in the Hox cluster. Rearrangements were also found outside the butterflies, with independent changes seen in *Ypsolopha scabrella*, *Emmelina monodactyla*, *Carcina quercana*, *Clostera curtula*, *Laspeyria flexula*, *Abrostola tripartita* and *Neomicropteryx cornuta*. These NK cluster rearrangements in moths include translocation of one, two or three of the *slou*/*Hmx*/*ems* genes to the opposite end of the cluster, and relocation of *C15* to the opposite end of the cluster in five of the seven species (Supplementary Figure S9). In contrast to the Hox gene cluster, the Hi-C contact data do not provide evidence for a strong TAD spanning the NK homeobox gene cluster (Supplementary Figure S4B).

Discussion

Overall stability of homeobox gene numbers

Although the expression, function and evolution of homeobox genes has been extensively studied in insects, few studies have made comparisons across an entire insect order. In addition, most studies have focussed on Hox genes, with less attention paid to the many other types of homeobox gene or to genomic organisation. To a large degree, this is a consequence of the limited number of high quality chromosomal-level genome assemblies available until very recently. With advances in DNA sequencing technology, coupled with scaffolding using Hi-C, this limitation is being overcome (The Darwin Tree of Life Project Consortium 2022). To better understand homeobox gene evolution in Lepidoptera, we annotated genes from all homeobox classes in 123 well-assembled Lepidoptera genomes.

We found general stability in homeobox gene numbers across the order, with most species having ~100 homeobox loci from all classes. This overall consistency in homeobox gene content may relate to overall body plan stability across Lepidoptera. There are some notable variations in gene content between species and families; most of these concern the Hox genes, including the Shx genes, discussed below. Otherwise, we see a degree of consistency, in gene number if not in gene organisation. Leaving Hox genes aside, most homeobox genes are dispersed in these genomes, and linkages are not conserved. The NK homeobox genes and the Homeobrain, Retinal Homeobox and Orthopedia genes from the PRD class are an exception, with both sets of genes having a conserved cluster arrangement in Lepidoptera. The NK cluster usually contains nine genes and spans 2.4Mb to 10Mb. In these genes, we find tight clustering across insects of *Msx* (*Dr*), *NK4* (*tin*), *NK3* (*bap*) and *Lbx* (*lbe*), suggestive of a functional constraint or common regulation, whereas the remaining genes *Tlx* (*C15*), *NK1* (*slou*), *Hmx* (*NK5*) and *Emx* (*ems*) have more variation in their gene order. The Homeobrain, Retinal Homeobox and Orthopedia cluster is a compact

cluster, with an average length of ~300 kb. The genes are highly conserved in order, but vary in gene orientation across Lepidoptera.

The unusual lepidopteran Hox gene cluster

Hox genes are arranged in genomic clusters as a result of tandem gene duplication, followed by selective pressure that has kept Hox genes together as neighbours for hundreds of millions of years. The nature of the selective pressure is not fully understood, but may in part be related to long range regulatory elements important for spatial colinearity of gene expression (McGinnis and Krumlauf 1992; Duboule and Morata 1994; Lemons and McGinnis 2006). Some changes to the structure of the Hox gene cluster have been found in insects (Lewis 1978; Duncan 1987; Ferrier and Akam 1996; Powers et al. 2000; Brown et al. 2002; Negre and Ruiz 2007), and some larger rearrangements observed in non-insect arthropods (Cook et al. 2001; Grbić et al. 2011; Chipman et al. 2014; Pace et al. 2016; Leite et al. 2018), but we have a fragmentary picture of insect Hox cluster evolution thus far. Indeed, within Lepidoptera the complete structure of a Hox gene cluster has not been reported; even in the pioneering studies on *Bombyx mori* Hox genes, the precise location of the labial gene could not be resolved (Yasukochi et al. 2004; Chai et al. 2008). With the availability of chromosomal level genome assemblies, this picture is changing. This study attempts to characterise Hox gene cluster evolution in an insect order on a large scale. Among the findings were (a) determining that the labial gene is located at a distant position beyond *Abd-B*, likely relocated by an inversion event, and (b) the finding that the non-Hox gene *ro* is very closely linked to *pb*, in the position where labial is found in other insects. These two features are seen in all the ditrysian Lepidoptera we analysed, with an intermediate situation found in *Neomicropteryx cornuta*, a member of the Micropterygidae. This basal moth has a gene order of *pb*, *ro* and *lab*, suggesting that movement of *ro* into the Hox gene cluster occurred in an ancestor of extant Lepidoptera, while the inversion that moved the *lab* gene was a later event. However, even in *Neomicropteryx cornuta* the *lab* gene is 3.8Mb from the end of the

cluster suggesting that it had already 'escaped' from common control in the earliest Lepidoptera.

What could have allowed these rearrangements in Lepidoptera? One hypothesis is that all functional reasons for maintaining Hox gene clustering have been lost in Lepidoptera, and random rearrangements have been permitted in evolution. An alternative hypothesis is that it is just the *lab* gene that has been permitted to 'escape', perhaps due to loss of common regulatory control. Our analysis of topologically associated domains (TADs), and comparison to the NK gene cluster, suggests the second hypothesis is most likely. We found a pattern of physical association of chromatin containing the Hox gene cluster, but only from *pb* to *Abd-B*. We find that *lab* and *ro* are located outside of this TAD across all species sampled. This suggests that it is the *lab* gene specifically that has escaped from any common regulation or control; there is evidence that the remaining Hox genes maintain physical association in three dimensions and are thus under conserved regulation (Krefting et al. 2018). Similarly, although the *ro* gene has moved to be adjacent to the rest of the Hox cluster, it has not been encompassed within the same TAD. Consistent with this conclusion, the *ro* gene has moved secondarily to the *Abd-B* end of the Hox cluster in four closely related Pieridae species (*Aporia crataegi*, *Pieris rapae*, *Pieris brassicae* and *Pieris napi*) (Supplementary Figure S3).

Moths take the record for the most Hox loci

The number of Hox genes is variable within insects, with most variation due to duplications of non-canonical Hox genes, especially the *zen* gene (the derived orthologue of the paralogy group 3 Hox gene; (Falciani et al. 1996)). For example, fruitfly *Drosophila melanogaster* has three loci derived *zen* duplication: *zen*, *zen2* and *bcd*, while *Tribolium castaneum* has two (*Tczen1*, *Tczen2*; (Brown et al. 2002). Several Lepidoptera have five zen-derived genes (*zen*, *ShxA*, *ShxB*, *ShxC*, *ShxD*; (Ferguson et al. 2014), with *Bombyx mori* having around 15 (Chai et al. 2008; Ferguson et al. 2014). In contrast, during chordate evolution tandem duplication of canonical Hox genes gave rise to 15 Hox genes in amphioxus, and 14 in the

common ancestor of vertebrates (Powers and Amemiya 2004; Holland et al. 2008). Genome duplications during vertebrate evolution increased the total number of Hox genes; for example, human and mouse have 39 Hox genes, African Butterfly Fish *Pantodon buchholzi* has 45 Hox genes, Atlantic Eel *Anguilla anguilla* has 73 Hox genes, and Atlantic Salmon *Salmo salar* has 118 Hox genes and pseudogenes (Mungpakdee et al. 2008; Henkel et al. 2012; Martin and Holland 2014). Our analysis of Lepidoptera genomes has uncovered many cases of Hox gene duplication, including enormous arrays of Hox-derived loci. We find some moths have the highest number of Hox loci known to date.

We found two rare cases of single gene tandem duplications in Lepidoptera, *ftz* in *Spilarctia lutea* (Buff Ermine Moth) and *Dfd* in *Acronicta aceris* (Sycamore Moth), but otherwise all variation in gene number was due to gains and losses of zen-derived genes, including the Shx genes. Consistent with Ferguson et al. (2014), it is true to a first approximation to say that most Lepidoptera have four Shx genes, plus *zen*, such that the full complement of Hox-derived genes is usually 14 (*lab, pb, zen, ShxA, ShxB, ShxC, ShxD, Dfd, Scr, Antp, ftz, Ubx, abd-A, Abd-B*). The minor exceptions we find to this rule include (a) a moth in the basal family Micropterygidae which has multiple zen-derived genes, although these lack the distinctive amino acid signatures of Shx genes and are likely an independent duplication; (b) the Six-Spotted Burnet Moth, *Zygaena filipendulae*, with only two Shx genes, annotated as *ShxB* and *ShxC* (Supplementary Figure S3); (c) butterflies in family Lycaenidae and the genus *Melitaea* which have each independently lost *ShxD* (although *Melitaea cinxia* has four copies of Shx due to a subsequent duplication of *ShxA*).

However, the biggest exceptions to the ‘four Shx’ rule are the cases we find of independent, very extensive tandem duplication of Shx genes in several evolutionary lineages of moths. These expansions ranged from the 7 copies in *Schrankia costaestrigalis* (Pinion-streaked Snout Moth) to an astonishing 165 loci found in *Apamea monoglypha* (Dark Arches Moth). Other examples include 58 and 66 copies in *Blastobasis lacticolella* and *Blastobasis adustella* respectively, 19 copies in *Parapoynx stratiotata* (Boyes et al. 2022a), 24 copies in *Phalera bucephala* (Buff-tip Moth; Boyes et al. 2022b), 20 copies in *Spilarctia*

lutea (Buff Ermine Moth) and 34 copies in *Noctua fimbriata* (Broad-bordered Yellow Underwing; Holland et al. 2021). The particular Shx genes which underwent tandem duplication differed between species, with some showing duplication of single genes (eg. *ShxD* in *Parapoynx stratiotata*, *Noctua pronuba* and *Idaea aversata*, and *ShxA* in *Zeuzera pyrina*), and others having multiple copies of several of the four Shx genes (Supplementary Figure S5). It is currently unclear whether these large gene arrays are adaptive, having been driven by selection, or whether they are neutral and a consequence of a genomic region prone to duplication. In other gene families, large changes in copy number have been found to be adaptive and related to certain environments or behaviours (Briscoe et al. 2013; Cheng et al. 2017; Rane et al. 2019; Chakraborty et al. 2021). The Shx genes are expressed in the serosa during development, an extraembryonic tissue implicated in innate immunity and desiccation resistance in insects (Panfilio 2008; Jacobs et al. 2013, 2014, 2022). It is therefore possible that Shx duplication is an adaptation associated with modifications to the egg, and indeed many of the highly duplicated genes show increased rates of sequence evolution (Figure 4A). One possibility is that specialisation of mulitple Shx genes permitted evolutionary refinement of serosal function, which may be important to survival of lepidopteran eggs laid on exposed surfaces of vegetation or in other challenging niches (Holland et al. 2017). However, although some of the moth species with large Shx expansions do have unusual ecology (such as aquatic eggs in *Parapoynx stratiotata*), we have not found a common developmental pattern, environmental link, or egg laying behaviour among all species with large tandem duplications of Shx genes.

The alternative hypothesis, that extensive tandem duplication of Shx genes is neutral, would demand an explanation for why the number of Shx genes is stable at four (or three) in most lepidopteran lineages, yet undergoes dramatic expansion in others. We do not find a pattern consistent with a widespread stochastic gain and loss: the pattern is one of either stability or expansion. We propose that such a pattern is indicative of an underlying mutational mechanism driving duplication in some species and not others. One possible mutational mechanism relates to transposable element content. In almost all species where

large tandem duplication occurs (14/19), we find significantly increased density of LINE elements in the region containing the Shx genes, relative to the rest of the Hox cluster (Figure 5). Generally, transposon activity is highly regulated and reduced within the Hox cluster, due to the importance of the order and structure of the genes for proper development (Fried et al. 2004). However, if LINE elements successfully invade the Hox gene cluster, they could potentially promote tandem gene duplication through non-homologous pairing at meiosis. Thus, a neutral explanation could be that LINE elements invaded in some species, and caused an increased rate of duplication mutations, without phenotypic effect.

The adaptive and the neutral hypotheses can be reconciled, since even if initial duplication is neutral the new loci could be substrates for later adaptive evolution which the TEs themselves could alter gene regulation. By analogy, enrichment of TEs within the Hox gene clusters of Anolis lizards correlates with rates of speciation and affects the expression of Hox genes during development (Feiner 2016, 2019). It is interesting to note that invasion of TEs into *Anolis* lizard Hox clusters is not associated with gene duplication. This is possibly because all vertebrate Hox genes have anteroposterior expression domains that could be disrupted by tandem duplication; in Lepidoptera, the *zen* gene has lost ancestral regional expression and gained tissue-specific expression.

Material and Methods

Data acquisition

The genome assemblies used in this analysis were produced by the Darwin Tree of Life project (The Darwin Tree of Life Project Consortium 2022) and can be found under the EBI and ENA bioproject number PRJEB40665 and on the DToL portal page: portal.darwintreeoflife.org. The genome for a non-ditrysian species was obtained from the recent sequencing of the Micropterigidae species, *Neomicropteryx cornuta* (Li et al. 2021).

Sequences for all homeodomains from three insects (*Drosophila melanogaster*, *Tribolium castaneum*, *Apis mellifera*) were downloaded from homeodb (http://homeodb.zoo.ox.ac.uk) (Zhong et al. 2008; Zhong and Holland 2011). Sequences for the lepidopteran specific special homeobox genes (Shx) were obtained from Ferguson et al (2014). Summary of genomes used, their shortened names, family membership, GenBank accession IDs and Project IDs are found in Supplementary Table S1.

Homeobox gene identification

To identify homeobox genes in the assembled genomes, the homeodomain protein sequences were used as queries in a TBLASTN search against the lepidopteran genomes (e-value threshold of 1 × 10-5). Overlapping hits from the lepidopteran genomes were filtered to retain a single sequence per homeobox gene with the longest sequence match. The resulting sequences from the lepidopteran genomes were then subsequently used in a reciprocal BLASTX search against the homeodomain protein dataset. For hits with significant percent identity (over 70%) the reciprocal BLAST search allowed for initial identification of the given homeobox gene. A second round of sequence similarity searches was carried out using MMSeqs2 (Steinegger and Söding 2017) and 1kb either side of the homeobox genes annotated from the initial BLAST search. The scripts for each step are at https://github.com/PeterMulhair/HbxFinder. For divergent sequences, identification was carried out using phylogenetic analysis (see Molecular analysis of homeobox evolution). Visualisation of the Hox gene clusters and gene tree employed R 4.0.3 (R Development Core Team, 2021) using gggenes (github.com/wilkox/gggenes) and ggtree (Yu et al. 2017), respectively. The newly-identified homeodomain nucleotide sequences were then translated into amino acid format using the sixpack package from EMBOSS (Madeira et al. 2019); amino acid sequences with the highest identity to known homeodomain sequences were retained.

Homeobox gene expression

Expression of all homeobox genes identified in our dataset of 123 species was assessed using whole body RNA-seq data from a representative set of seven species (*Biston betularia*, *Limenitis camilla*, *Nymphalis urticae*, *Pararge aegeria*, *Pieris rapae*, *Vanessa atalanta*, *Vanessa cardui*). RNA-seq data was downloaded from the DToL portal page: portal.darwintreeoflife.org. Transcriptomes assembly was carried out for each species using Trinity v2.8.5 (Grabherr et al. 2011). Next, for each transcriptome assembly, transcript abundance was calculated using kallisto v0.44 (Bray et al. 2016). Homeobox gene identification and expression quantification was then performed in each species using a reciprocal BLAST approach.

Species tree inference

A species tree for the 123 lepidopteran species in our dataset was generated using gene sets obtained from BUSCO v5.1.2 (Manni et al. 2021). First, genes were annotated using the Lepidoptera BUSCO gene sets. Next, the busco2phylo-nf pipeline (https://github.com/lstevens17/busco2phylo-nf) was used to extract FASTA files for each annotated gene, ensuring 100% species coverage in each one. Each gene was aligned using MAFFT v7.467 (Katoh et al. 2005) and gene trees were inferred using IQ-Tree v2.0 (Minh et al. 2020), using ModelFinder to find the model of best fit (Kalyaanamoorthy et al. 2017). Finally, a species tree was inferred using the supertree approach in ASTRAL v5.7.7 (Zhang et al. 2018).

Molecular analysis of homeobox evolution

Phylogenetic reconstruction was carried out using the homeodomain amino acid sequences Homeodomain sequences were aligned using MAFFT v7.467 (Katoh et al. 2005) and maximum likelihood trees were built using IQ-Tree v2.0 (Nguyen et al. 2015) and LG+G model of sequence evolution. Tree visualisation was carried out using ggtree (Yu et al.

2017). To test for changes in rates of homeodomain sequence evolution of the Shx genes between the Lycaenidae species (which lost ShxD) and all other lepidopteran species with a normal set of Shx genes, we measure pairwise identity between species as a proxy for evolutionary rate. This analysis was carried out using PhyKIT using the phykit pairwise_identity command (Steenwyk et al. 2021). To measure whether selection was relaxed or intensified in any of the three remaining Shx genes on any of the Lycaenidae branches, we used the RELAX model (Wertheim et al. 2015) implemented in HyPhy (Kosakovsky Pond et al. 2020).

Hi-C data processing and TAD identification

Hi-C reads were mapped to the genomes using BWA 0.7.5a-r405 (Li 2013). HiCExplorer was then used to process the Hi-C data to form interaction maps, annotate the TADs and visualise the results (Ramírez et al. 2018).

Repeat annotation and TE density analysis

TEs were annotated using both RepeatModeler and RepeatMasker pipelines. For each genome tested, a de novo repeat library was generated from the genome assemblies using RepeatModeler2 (Flynn et al. 2020). This library was combined with the RepeatMasker Insecta library (Bao et al. 2015) and the SINE database (Vassetzky and Kramerov 2013), and filtered for any protein coding genes and repeat elements below 50 bases in length. Repeats were classified using RepeatMasker v4.1.0 (Smit et al., 2013), and regions containing LINE, SINE, LTR and DNA elements were extracted for subsequent analysis. Next, for each of the four broad TE classes, densities in 5kb windows were calculated first for the regions containing the Shx genes and second for the full Hox gene cluster minus the Shx gene region and *lab*. Enrichment for TE density in the Shx gene region compared to the remaining Hox cluster was carried out for each TE class using Wilcoxon rank-sum test with Bonferroni correction in the SciPy python package (Virtanen et al. 2020). TE density enrichment across the Lepidoptera phylogeny was visualised using the Toytree python

600 package (Eaton 2020). These analyses were not intended as exhaustive, but to give insight
601 into TE density within the Hox gene cluster.

602 Data access

603 All data and code required to reproduce analyses and figures can be found in
604 Supplementary Materials and at the GitHub repository
605 github.com/PeterMulhair/Lepidoptera_homeobox and doi.org/10.5281/zenodo.7274111.

606 Competing interest statement

607 The authors declare no conflict of interest.

608 Acknowledgements

609 We thank Yi-Jyun Luo, Tom Lewin, Sarah Bannister, Jo Blagrove, Lewis Stevens, Charlotte
610 Wright, Emmelien Vancaester, Claudia Weber, Shane McCarthy, Marcela Uliano-Silva, Mark
611 Blaxter and Ignacio Maeso for helpful discussions and advice. We also acknowledge the
612 huge effort at each stage in the generation of the genomes by the Darwin Tree of Life
613 project, including species sampling and processing, DNA extraction and sequencing,
614 genome assembly and curation, and database construction. This research was funded by
615 the Wellcome Trust Darwin Tree of Life Discretionary Award [218328] and the John Fell
616 OUP Research Fund. We dedicate this work to the memory of Douglas Boyes who was
617 pivotal to this work, sampling the majority of species presented here and providing incredible
618 knowledge on lepidopteran biology.

619 Author contributions

620 PWHH and POM conceived the study and PWHH and OTL oversaw research. DHB, LC,
621 POM, PWHH and OTL were involved in species sampling and processing for genome
622 sequencing. POM, AH and PWH designed analyses. AH carried out initial analyses; POM

carried out bioinformatic analyses presented. POM, AH, LC and PWWH interpreted results. POM and PWHH wrote the initial draft of the manuscript. All authors read and approved the final manuscript.

Darwin Tree of Life Consortium

Mark Blaxter[5], Nova Mieszkowska[6,7], Federica Di Palma[8], Peter Holland[1], Richard Durbin[5,9], Thomas Richards[1], Matthew Berriman[5], Paul Kersey[10], Peter Hollingsworth[11], Willie Wilson[6,12], AlexTwyford[10,13], Ester Gaya[10], Mara Lawniczak[5], Owen Lewis[1], Gavin Broad[14], Kevin Howe[15], Michelle Hart[11], Paul Flicek[15], Ian Barnes[14]

[5] Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridgeshire CB10 1SA, UK

[6] Marine Biological Association of the United Kingdom, Citadel Hill, Plymouth PL1 2PB, UK

[7] University of Liverpool, Liverpool L69 3BX, UK

[8] University of East Anglia, Norwich Research Park, Norwich NR4 7TJ, UK

[9] Department of Genetics, University of Cambridge, Cambridge CB2 3EH, UK

[10] Royal Botanic Gardens, Kew, Richmond, London TW9 3AE, UK

[11] Royal Botanic Garden Edinburgh, Edinburgh EH3 5LR

[12] University of Plymouth, Drake Circus, Plymouth PL4 8AA, UK

[13] Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh,Edinburgh EH8 9YL

[14] Natural History Museum, Cromwell Road, London SW7 5BD, UK12EMBL-EBI, Wellcome Genome Campus, Hinxton, Cambridgeshire CB10 1SD,UK

[15] EMBL-EBI, Wellcome Genome Campus, Hinxton, Cambridgeshire CB10 1SD,UK

Figure Legends

Figure 1: Numbers of homeobox sequences across Lepidoptera. (A) Species tree of Lepidoptera analysed constructed using BUSCO gene set. Coloured boxes spanning tips of the tree represent distinct Lepidoptera families with the family names shown. Species in the tree are listed in the same order as in Supplementary Figure S1. **(B)** Heatmap showing numbers of homeobox loci in each gene class and subclass (from left to right: lab, *Abd-B*, *abd-A*, *Ubx*, *Antp*, *ftz*, *Scr*, *Dfd*, *zen*, *Shx*, *pb*, *ind*, *cad*, *exex, eve, unpg, btn, Tlx, Msx, NK4, NK3, Lbx, NK1, Hmx, Emx, Hhex, NK7, NK6, Nedx, Dlx, En, NK2.1, Msxlx, Hlx, NK2.2, Barhl, Bari, Bsx, Dbx, Abox, Noto, Ro, Uncx, Gsc, Pitx, Otp, Rx, Hbn, Repo, Prrx, Shox, Arx, Pax4/6, Phox, Prop, Vsx, CG11294, Pax3/7, Drgx, Otx, Lhx6/8, Lmx, Lhx2/9, Lhx3/4, Lhx1/5, Isl, Pou2, Pou3, Pou4, Pou6, Six3/6, Six1/2, Six4/5, Meis, Irx, Mkx, Pbx, Tgif, Onecut, Cux, Cmp, Prox, Zfhx, Cers*). **(C)** Total counts of homeobox loci in each genome.

Figure 2: Hox gene cluster evolution across Insecta. **(A)** Comparison of the general structure of the Hox gene cluster between representative species for Hymenoptera (*Bombus terrestris*), Coleoptera (*Tribolium castaneum*), Diptera (*Drosophila melanogaster*) and Lepidoptera. Lepidoptera are shaded in an orange box and split between non-ditrysia species (*Neomicropteryx cornuta*) and Ditrysia (represented by 122 species in our dataset). Lepidoptera specific Shx genes are coloured orange (ShxA), red (ShxB), green (ShxC) and blue (ShxD) in this figure and throughout the manuscript. **(B)** Genomic location of Hox genes in *Autographa gamma* with corresponding exon structures and genomic distances annotated below. Silhouette images of *Bombus terrestris*, *Tribolium castaneum* and *Drosophila melanogaster* were taken from PhyloPic (phylopic.org).

Figure 3: Evidence for a conserved topologically associated domain (TAD) spanning the Hox gene cluster across Lepidoptera. **(A)** Species tree of nine representative lepidopteran species on left, HiC matrix showing 1Mb either side of the Hox gene cluster (excluding lab). The location of the Hox gene cluster from (Abd-B to pb) is annotated by a blue bar, along with its orientation. The position of ro is annotated with a short vertical black dash. The intensity of chromatin compaction is represented by a blue (low) to red (high) colour gradient. Across the core Hox cluster (pb to Abd-B) in each species a TAD is represented by a region of strong contact, by the more yellow shaded regions. Black lines represent TADs or sub-TADs predicted by HiCExplorer. **(B)** Above shows the arrangement of the Hox gene cluster (excluding ro and lab) surrounded by a TAD (orange) in *Pheosia gnoma* (Lesser Swallow Prominent). Below displays the HiC matrix of Chromosome 10 showing the location of the Hox cluster, represented by a blue bar, along with ro and lab, represented by short vertical black dashes. **(C)** Schematic showing tologically folded domains in Chromosome 10 (red) interspersed by chromosome regions with less consistent topology (blue) based on the above HiC matrix. Shaded grey region shows the location of the condensed TAD containing the Hox cluster.

Figure 4: Sequence evolution of Hox genes across Lepidoptera (A) Phylogenetic tree of Hox and Hox-derived homeodomains across 46 Lepidoptera species. Shx gene clades are coloured orange, red, green and blue; canonical Hox genes are coloured yellow. The names of the Hox genes are placed alongside their clade in the tree. **(B) Shx genes show elevated sequence evolution following loss of ShxD in Lycaenidae.** Results of pairwise identity of Shx genes between Lycaenidae species and non-Lycaenidae species. For each gene (*ShxA-C*) pairwise identity between Lycaenidae and all other Lepidoptera species with normal Shx gene count (darker shade boxplot) is compared with pairwise identity between all Lepidoptera species with normal Shx gene count (lighter shade boxplot). Each pair of boxplots (light shade and dark shade) are coloured according to the colour code for each of the Shx genes. Wilcoxon rank-sum test was carried out between pairwise identity for Lycaenidae and non-Lycaenidae species (* = P-value < 0.05).

Figure 5: Association between increased LINE density and extensive tandem duplication of Shx genes. (A) Left shows the species tree of 122 Lepidoptera species. Bar chart in yellow

corresponds to the length of the Hox cluster (excluding labial) for each species in the tree measured in Mb. The first column in blue indicates those species with large tandem duplications of Shx genes in the Hox cluster (dark blue) or those with a 'normal' number of Shx genes (light blue). The second column in red indicates species with significantly enriched density of LINE elements (dark red) within the region containing the Shx genes. **(B)** LINE density plot across the Hox cluster plus 3Mb either side; this is shown for 14 species with enriched LINE density in the region containing Shx genes. The outer black dashed lines represent the edges of the Hox cluster (Abd-B to ro), while the inner red dashed lines represent the edges of the Shx genes (ShxD to ShxA).

Figure 6: NK gene cluster evolution across Insecta. **(A)** Comparison of the general structure of the NK gene cluster between representative species for Hymenoptera (*Bombus terrestris*), Coleoptera (*Tribolium castaneum*), Diptera (*Drosophila melanogaster*) and Lepidoptera. Lepidoptera are shaded in an orange box and split between non-ditrysia species (*Neomicropteryx cornuta*) and Ditrysia (represented by 122 species in our dataset). **(B)** Genomic location of NK genes in *Autographa gamma* with corresponding exon structures and genomic distances annotated below. Silhouette images of *Bombus terrestris*, *Tribolium castaneum* and *Drosophila melanogaster* were taken from PhyloPic (phylopic.org). **(C)** Left shows the species topology for the 36 butterflies in the dataset, along with an outgroup representative. Rearrangements in the NK cluster are annotated on the branches of the tree where they were estimated to have occurred (represented by yellow stars). Black lines spanning tips on the tree group species which show the same structure and order in the NK gene cluster. The NK gene cluster is represented by coloured boxes, in the 'canonical' order of *Tlx* (*C15*), *Msx* (*Dr*), *NK4* (*tin*), *NK3* (*bap*), *Lbx* (*lbe*), *NK1* (*slou*), *Hmx* (*NK5*) and *Emx* (*ems*). Species with the NK genes in this order are shadowed by a blue box. Synteny between the closely linked genes of both *Msx* (*Dr*) genes, *NK4* (*tin*), *NK3* (*bap*) and *Lbx* (*lbe*) is represented by shaded blocks, to show changes in the order and structure of the NK cluster.

References

Aase-Remedios ME, Ferrier DEK. 2021. Improved Understanding of the Role of Gene and Genome Duplications in Chordate Evolution With New Genome and Transcriptome Sequences. *frontiers in Ecology and Evolution*. http://dx.doi.org/10.3389/fevo.2021.703163.

Acemel RD, Maeso I, Gómez-Skarmeta JL. 2017. Topologically associated domains: a successful scaffold for the evolution of gene regulation in animals. *Wiley Interdiscip Rev Dev Biol* **6**. http://dx.doi.org/10.1002/wdev.265.

Bao W, Kojima KK, Kohany O. 2015. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob DNA* **6**: 11.

Boyes D, Chadd R, Mulhair P, University of Oxford and Wytham Woods Genome Acquisition Lab, Natural History Museum Genome Acquisition Lab, Darwin Tree of Life Barcoding collective, Wellcome Sanger Institute Tree of Life programme, Wellcome Sanger Institute Scientific Operations: DNA Pipelines collective, Tree of Life Core Informatics collective, Darwin Tree of Life Consortium. 2022a. The genome sequence of the ringed china-mark, *Parapoynx stratiotata* (Linnaeus, 1758). *Wellcome Open Res* **7**: 121.

Boyes D, Holland PWH, University of Oxford and Wytham Woods Genome Acquisition Lab, Darwin Tree of Life Barcoding collective, Wellcome Sanger Institute Tree of Life programme, Wellcome Sanger Institute Scientific Operations: DNA Pipelines collective, Tree of Life Core Informatics collective, Darwin Tree of Life Consortium. 2022b. The

742 genome sequence of the buff-tip, Phalera bucephala (Linnaeus, 1758). *Wellcome Open*
743 *Res* **7**: 28.

744 Boyes D, Holland PWH, University of Oxford and Wytham Woods Genome Acquisition Lab,
745 Darwin Tree of Life Barcoding collective, Wellcome Sanger Institute Tree of Life
746 programme, Wellcome Sanger Institute Scientific Operations: DNA Pipelines collective,
747 Tree of Life Core Informatics collective, Darwin Tree of Life Consortium. 2022c. The
748 genome sequence of the silver Y moth, *Autographa gamma* (Linnaeus, 1758).
749 *Wellcome Open Res* **7**: 100.

750 Bray NL, Pimentel H, Melsted P, Pachter L. 2016. Near-optimal probabilistic RNA-seq
751 quantification. *Nat Biotechnol* **34**: 525–527.

752 Briscoe AD, Macias-Muñoz A, Kozak KM, Walters JR, Yuan F, Jamie GA, Martin SH,
753 Dasmahapatra KK, Ferguson LC, Mallet J, et al. 2013. Female behaviour drives
754 expression and evolution of gustatory receptors in butterflies. *PLoS Genet* **9**: e1003620.

755 Brown SJ, Fellers JP, Shippy TD, Richardson EA, Maxwell M, Stuart JJ, Denell RE. 2002.
756 Sequence of the Tribolium castaneum homeotic complex: the region corresponding to
757 the Drosophila melanogaster antennapedia complex. *Genetics* **160**: 1067–1074.

758 Butts T, Holland PWH, Ferrier DEK. 2008. The urbilaterian Super-Hox cluster. *Trends Genet*
759 **24**: 259–262.

760 Chai C-L, Zhang Z, Huang F-F, Wang X-Y, Yu Q-Y, Liu B-B, Tian T, Xia Q-Y, Lu C, Xiang Z-
761 H. 2008. A genomewide survey of homeobox genes and identification of novel structure
762 of the Hox cluster in the silkworm, Bombyx mori. *Insect Biochem Mol Biol* **38**: 1111–
763 1120.

764 Chakraborty M, Ramaiah A, Adolfi A, Halas P, Kaduskar B, Ngo LT, Jayaprasad S, Paul K,
765 Whadgar S, Srinivasan S, et al. 2021. Hidden genomic features of an invasive malaria
766 vector, Anopheles stephensi, revealed by a chromosome-level genome assembly. *BMC*
767 *Biol* **19**: 28.

768 Cheng T, Wu J, Wu Y, Chilukuri RV, Huang L, Yamamoto K, Feng L, Li W, Chen Z, Guo H,
769 et al. 2017. Genomic adaptation to polyphagy and insecticides in a major East Asian
770 noctuid pest. *Nat Ecol Evol* **1**: 1747–1756.

771 Chipman AD, Ferrier DEK, Brena C, Qu J, Hughes DST, Schröder R, Torres-Oliva M, Znassi
772 N, Jiang H, Almeida FC, et al. 2014. The first myriapod genome sequence reveals
773 conservative arthropod gene content and genome organisation in the centipede
774 Strigamia maritima. *PLoS Biol* **12**: e1002005.

775 Cook CE, Smith ML, Telford MJ, Bastianello A, Akam M. 2001. Hox genes and the
776 phylogeny of the arthropods. *Curr Biol* **11**: 759–763.

777 Duboule D. 2007. The rise and fall of Hox gene clusters. *Development* **134**: 2549–2560.

778 Duboule D, Morata G. 1994. Colinearity and functional hierarchy among genes of the
779 homeotic complexes. *Trends Genet* **10**: 358–364.

780 Duncan I. 1987. The bithorax complex. *Annu Rev Genet* **21**: 285–319.

781 Eaton DAR. 2020. Toytree: A minimalist tree visualization and manipulation library for
782 Python. *Methods Ecol Evol*.
783 https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.13313.

784 Eres IE, Gilad Y. 2021. A TAD Skeptic: Is 3D Genome Topology Conserved? *Trends Genet*
785 **37**: 216–223.

786 Falciani F, Hausdorf B, Schröder R, Akam M, Tautz D, Denell R, Brown S. 1996. Class 3
787 Hox genes in insects and the origin of zen. *Proc Natl Acad Sci U S A* **93**: 8479–8484.

788 Feiner N. 2016. Accumulation of transposable elements in Hox gene clusters during
789 adaptive radiation of Anolis lizards. *Proc Biol Sci* **283**.
790 http://dx.doi.org/10.1098/rspb.2016.1555.

791 Feiner N. 2019. Evolutionary lability in Hox cluster structure and gene expression in Anolis
792 lizards. *Evol Lett* **3**: 474–484.

793 Ferguson L, Marlétaz F, Carter J-M, Taylor WR, Gibbs M, Breuker CJ, Holland PWH. 2014.
794 Ancient expansion of the hox cluster in lepidoptera generated four homeobox genes
795 implicated in extra-embryonic tissue formation. *PLoS Genet* **10**: e1004698.

796 Ferrier DE, Akam M. 1996. Organization of the Hox gene cluster in the grasshopper,
797 Schistocerca gregaria. *Proc Natl Acad Sci U S A* **93**: 13024–13029.

798 Ferrier DEK. 2016. Evolution of Homeobox Gene Clusters in Animals: The Giga-Cluster and
799 Primary vs. Secondary Clustering. *Frontiers in Ecology and Evolution* **4**: 36.

800 Ferrier DEK, Holland PWH. 2002. Ciona intestinalis ParaHox genes: evolution of
801 Hox/ParaHox cluster integrity, developmental mode, and temporal colinearity. *Mol*
802 *Phylogenet Evol* **24**: 412–417.

803 Finnerty JR, Martindale MQ. 1998. The evolution of the Hox cluster: insights from outgroups.
804 *Curr Opin Genet Dev* **8**: 681–687.

805 Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, Smit AF. 2020.
806 RepeatModeler2 for automated genomic discovery of transposable element families.
807 *Proc Natl Acad Sci U S A* **117**: 9451–9457.

808 Fried C, Prohaska SJ, Stadler PF. 2004. Exclusion of repetitive DNA elements from
809 gnathostome Hox clusters. *J Exp Zool B Mol Dev Evol* **302**: 165–173.

810 Garcia-Fernàndez J. 2005. The genesis and evolution of homeobox gene clusters. *Nat Rev*
811 *Genet* **6**: 881–892.

812 Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L,
813 Raychowdhury R, Zeng Q, et al. 2011. Full-length transcriptome assembly from RNA-
814 Seq data without a reference genome. *Nat Biotechnol* **29**: 644–652.

815 Grbić M, Van Leeuwen T, Clark RM, Rombauts S, Rouzé P, Grbić V, Osborne EJ, Dermauw
816 W, Ngoc PCT, Ortego F, et al. 2011. The genome of Tetranychus urticae reveals
817 herbivorous pest adaptations. *Nature* **479**: 487–492.

818 Henkel CV, Burgerhout E, de Wijze DL, Dirks RP, Minegishi Y, Jansen HJ, Spaink HP,
819 Dufour S, Weltzien F-A, Tsukamoto K, et al. 2012. Primitive duplicate Hox clusters in
820 the European eel's genome. *PLoS One* **7**: e32231.

821 Hill J, Rastas P, Hornett EA, Neethiraj R, Clark N, Morehouse N, de la Paz Celorio-Mancera
822 M, Cols JC, Dircksen H, Meslin C, et al. 2019. Unprecedented reorganization of
823 holocentric chromosomes provides insights into the enigma of lepidopteran
824 chromosome evolution. *Sci Adv* **5**: eaau3648.

825 Holland LZ, Albalat R, Azumi K, Benito-Gutiérrez E, Blow MJ, Bronner-Fraser M, Brunet F,
826 Butts T, Candiani S, Dishaw LJ, et al. 2008. The amphioxus genome illuminates
827 vertebrate origins and cephalochordate biology. *Genome Res* **18**: 1100–1111.

828 Holland PWH. 2015. Did homeobox gene duplications contribute to the Cambrian explosion?
829 *Zoological Lett* **1**: 1.

830 Holland PWH, Marlétaz F, Maeso I, Dunwell TL, Paps J. 2017. New genes from old:
831 asymmetric divergence of gene duplicates and the evolution of development. *Philos*
832 *Trans R Soc Lond B Biol Sci* **372**. http://dx.doi.org/10.1098/rstb.2015.0480.

833 Holland PWH, University of Oxford and Wytham Woods Genome Acquisition Lab, Darwin
834 Tree of Life Barcoding collective, Wellcome Sanger Institute Tree of Life programme,
835 Wellcome Sanger Institute Scientific Operations: DNA Pipelines collective, Tree of Life
836 Core Informatics collective, Darwin Tree of Life Consortium. 2021. The genome
837 sequence of the broad-bordered yellow underwing, Noctua fimbriata (Schreber, 1759).
838 *Wellcome Open Res* **6**: 345.

839 Jacobs CGC, Rezende GL, Lamers GEM, van der Zee M. 2013. The extraembryonic serosa
840 protects the insect egg against desiccation. *Proc Biol Sci* **280**: 20131082.

841 Jacobs CGC, Spaink HP, van der Zee M. 2014. The extraembryonic serosa is a frontier
842 epithelium providing the insect egg with a full-range innate immune response. *Elife* **3**.
843 http://dx.doi.org/10.7554/eLife.04111.

844 Jacobs CGC, van der Hulst R, Chen Y-T, Williamson RP, Roth S, van der Zee M. 2022.
845 Immune function of the serosa in hemimetabolous insect eggs. *Philos Trans R Soc*
846 *Lond B Biol Sci* **377**: 20210266.

847 Jagla K, Bellard M, Frasch M. 2001. A cluster of Drosophila homeobox genes involved in
848 mesoderm differentiation programs. *Bioessays* **23**: 125–133.

849 Janoušek V, Laukaitis CM, Yanchukov A, Karn RC. 2016. The Role of Retrotransposons in
850 Gene Family Expansions in the Human and Mouse Genomes. *Genome Biol Evol* **8**:
851 2632–2650.

852 Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. ModelFinder:
853 fast model selection for accurate phylogenetic estimates. *Nat Methods* **14**: 587–589.

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

856 Kawahara AY, Plotkin D, Espeland M, Meusemann K, Toussaint EFA, Donath A, Gimnich F,
857 Frandsen PB, Zwick A, Dos Reis M, et al. 2019. Phylogenomics reveals the evolutionary
858 timing and pattern of butterflies and moths. *Proc Natl Acad Sci U S A* **116**: 22657–
859 22663.

860 Kosakovsky Pond SL, Poon AFY, Velazquez R, Weaver S, Hepler NL, Murrell B, Shank SD,
861 Magalis BR, Bouvier D, Nekrutenko A, et al. 2020. HyPhy 2.5-A Customizable Platform
862 for Evolutionary Hypothesis Testing Using Phylogenies. *Mol Biol Evol* **37**: 295–299.

863 Krefting J, Andrade-Navarro MA, Ibn-Salem J. 2018. Evolutionary stability of topologically
864 associating domains is associated with conserved gene regulation. *BMC Biol* **16**: 87.

865 Lawniczak MKN, Durbin R, Flicek P, Lindblad-Toh K, Wei X, Archibald JM, Baker WJ, Belov
866 K, Blaxter ML, Marques Bonet T, et al. 2022. Standards recommendations for the Earth

867 BioGenome Project. *Proc Natl Acad Sci U S A* **119**.
868 http://dx.doi.org/10.1073/pnas.2115639118.

869 Leite DJ, Baudouin-Gonzalez L, Iwasaki-Yokozawa S, Lozano-Fernandez J, Turetzek N,
870 Akiyama-Oda Y, Prpic N-M, Pisani D, Oda H, Sharma PP, et al. 2018. Homeobox Gene
871 Duplication and Divergence in Arachnids. *Mol Biol Evol* **35**: 2240–2253.

872 Lemons D, McGinnis W. 2006. Genomic evolution of Hox gene clusters. *Science* **313**: 1918–
873 1922.

874 Lewis EB. 1978. A gene complex controlling segmentation in Drosophila. *Nature* **276**: 565–
875 570.

876 Liao Y, Zhang X, Chakraborty M, Emerson JJ. 2021. Topologically associating domains and
877 their role in the evolution of genome structure and function in Drosophila. *Genome Res*
878 **31**: 397–410.

879 Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-
880 MEM. *arXiv [q-bioGN]*. http://arxiv.org/abs/1303.3997.

881 Li X, Ellis E, Plotkin D, Imada Y, Yago M, Heckenhauer J, Cleland TP, Dikow RB, Dikow T,
882 Storer CG, et al. 2021. First Annotated Genome of a Mandibulate Moth, Neomicropteryx
883 cornuta, Generated Using PacBio HiFi Sequencing. *Genome Biol Evol* **13**.
884 http://dx.doi.org/10.1093/gbe/evab229.

885 Luke GN, Castro LFC, McLay K, Bird C, Coulson A, Holland PWH. 2003. Dispersal of NK
886 homeobox gene clusters in amphioxus and humans. *Proc Natl Acad Sci U S A* **100**:
887 5292–5295.

888 Madeira F, Park YM, Lee J, Buso N, Gur T, Madhusoodanan N, Basutkar P, Tivey ARN,
889 Potter SC, Finn RD, et al. 2019. The EMBL-EBI search and sequence analysis tools
890 APIs in 2019. *Nucleic Acids Res* **47**: W636–W641.

891 Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. 2021. BUSCO Update: Novel
892 and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for
893 Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Mol Biol Evol* **38**: 4647–4654.

894 Martin KJ, Holland PWH. 2014. Enigmatic orthology relationships between Hox clusters of
895 the African butterfly fish and other teleosts following ancient whole-genome duplication.
896 *Mol Biol Evol* **31**: 2592–2611.

897 Mazza ME, Pang K, Reitzel AM, Martindale MQ, Finnerty JR. 2010. A conserved cluster of
898 three PRD-class homeobox genes (homeobrain, rx and orthopedia) in the Cnidaria and
899 Protostomia. *Evodevo* **1**: 3.

900 McGinnis W, Krumlauf R. 1992. Homeobox genes and axial patterning. *Cell* **68**: 283–302.

901 Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear
902 R. 2020. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in
903 the Genomic Era. *Mol Biol Evol* **37**: 1530–1534.

904 Mitter C, Davis DR, Cummings MP. 2017. Phylogeny and Evolution of Lepidoptera. *Annu*
905 *Rev Entomol* **62**: 265–283.

906 Mulhair PO, Crowley L, Boyes DH, Harper A, Lewis OT, Holland PWH.
907 PeterMulhair/Lepidoptera_homeobox: v1.1. Zenodo; 2022. Available from:
908 https://doi.org/10.5281/zenodo. 7274111.

909 Mungpakdee S, Seo H-C, Angotzi AR, Dong X, Akalin A, Chourrout D. 2008. Differential
910 evolution of the 13 Atlantic salmon Hox clusters. *Mol Biol Evol* **25**: 1333–1343.

911 Negre, Bárbara, Sònia Casillas, Magali Suzanne, Ernesto Sánchez-Herrero, Michael Akam,
912 Michael Nefedov, Antonio Barbadilla, Pieter de Jong, and Alfredo Ruiz. 2005.
913 “Conservation of Regulatory Sequences and Gene Expression Patterns in the
914 Disintegrating Drosophila Hox Gene Complex.” Genome Research 15 (5): 692–700.

915 Negre B, Ruiz A. 2007. HOM-C evolution in Drosophila: is there a need for Hox gene
916 clustering? *Trends Genet* **23**: 55–59.

917 Nong W, Cao J, Li Y, Qu Z, Sun J, Swale T, Yip HY, Qian PY, Qiu J-W, Kwan HS, et al.
918 2020. Jellyfish genomes reveal distinct homeobox gene clusters and conservation of
919 small RNA processing. *Nat Commun* **11**: 3051.

920 Pace RM, Grbić M, Nagy LM. 2016. Composition and genomic organization of arthropod
921 Hox clusters. *Evodevo* **7**: 11.

922 Panfilio KA. 2008. Extraembryonic development in insects and the acrobatics of
923 blastokinesis. *Dev Biol* **313**: 471–491.

924 Powers TP, Amemiya CT. 2004. Evidence for a Hox14 paralog group in vertebrates. *Curr*
925 *Biol* **14**: R183–4.

926 Powers TP, Hogan J, Ke Z, Dymbrowski K, Wang X, Collins FH, Kaufman TC. 2000.
927 Characterization of the Hox cluster from the mosquito Anopheles gambiae (Diptera:
928 Culicidae). *Evol Dev* **2**: 311–325.

929 R Development Core Team (2021) R: A language and environment for statistical computing
930 R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org

931 Ramírez F, Bhardwaj V, Arrigoni L, Lam KC, Grüning BA, Villaveces J, Habermann B,
932 Akhtar A, Manke T. 2018. High-resolution TADs reveal DNA sequences underlying
933 genome organization in flies. *Nat Commun* **9**: 189.

934 Rane RV, Ghodke AB, Hoffmann AA, Edwards OR, Walsh TK, Oakeshott JG. 2019.
935 Detoxifying enzyme complements and host use phenotypes in 160 insect species. *Curr*
936 *Opin Insect Sci* **31**: 131–138.

937 Ranz, José M., Pablo M. González, Ryan N. Su, Sarah J. Bedford, Cei Abreu-Goodger, and
938 Therese Markow. 2022. “Multiscale Analysis of the Randomization Limits of the
939 Chromosomal Gene Organization between Lepidoptera and Diptera.” Proceedings.
940 Biological Sciences / The Royal Society 289 (1967): 20212183.

941 Schoenfelder S, Fraser P. 2019. Long-range enhancer-promoter contacts in gene
942 expression control. *Nat Rev Genet* **20**: 437–455.

943 Sexton T, Yaffe E, Kenigsberg E, Bantignies F, Leblanc B, Hoichman M, Parrinello H, Tanay
944 A, Cavalli G. 2012. Three-dimensional folding and functional organization principles of
945 the Drosophila genome. *Cell* **148**: 458–472.

946 Smit, AFA, Hubley, R & Green, P. RepeatMasker Open-4.0. 2013-2015
947 <http://www.repeatmasker.org>.

948 Soshnikova N, Dewaele R, Janvier P, Krumlauf R, Duboule D. 2013. Duplications of hox
949 gene clusters and the emergence of vertebrates. *Dev Biol* **378**: 194–199.

950 Startek M, Szafranski P, Gambin T, Campbell IM, Hixson P, Shaw CA, Stankiewicz P,
951 Gambin A. 2015. Genome-wide analyses of LINE-LINE-mediated nonallelic homologous
952 recombination. *Nucleic Acids Res* **43**: 2188–2198.

953 Steenwyk JL, Buida TJ, Labella AL, Li Y, Shen X-X, Rokas A. 2021. PhyKIT: a broadly
954 applicable UNIX shell toolkit for processing and analyzing phylogenomic data.
955 *Bioinformatics*. http://dx.doi.org/10.1093/bioinformatics/btab096.

956 Steinegger M, Söding J. 2017. MMseqs2 enables sensitive protein sequence searching for
957 the analysis of massive data sets. *Nat Biotechnol* **35**: 1026–1028.

958 Szabo Q, Bantignies F, Cavalli G. 2019. Principles of genome folding into topologically
959 associating domains. *Sci Adv* **5**: eaaw1668.

960 Szabo Q, Jost D, Chang J-M, Cattoni DI, Papadopoulos GL, Bonev B, Sexton T, Gurgo J,
961 Jacquier C, Nollmann M, et al. 2018. TADs are 3D structural units of higher-order
962 chromosome organization in Drosophila. *Sci Adv* **4**: eaar8082.

963 The Darwin Tree of Life Project Consortium. 2022. Sequence locally, think globally: The
964 Darwin Tree of Life Project. *PNAS*.
965 https://www.pnas.org/content/pnas/119/4/e2115642118 (Accessed March 25, 2022).

966 Thybert D, Roller M, Navarro FCP, Fiddes I, Streeter I, Feig C, Martin-Galvez D, Kolmogorov
967 M, Janoušek V, Akanni W, et al. 2018. Repeat associated mechanisms of genome
968 evolution and function revealed by the Mus caroli and Mus pahari genomes. *Genome*
969 *Res* **28**: 448–459.

970 Ulianov SV, Khrameeva EE, Gavrilov AA, Flyamer IM, Kos P, Mikhaleva EA, Penin AA,
971 Logacheva MD, Imakaev MV, Chertovich A, et al. 2016. Active chromatin and
972 transcription play a key role in chromosome partitioning into topologically associating
973 domains. *Genome Res* **26**: 70–84.

974 Vassetzky NS, Kramerov DA. 2013. SINEBase: a database and tool for SINE analysis.
975 *Nucleic Acids Res* **41**: D83–9.

976 Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E,
977 Peterson P, Weckesser W, Bright J, et al. 2020. SciPy 1.0: fundamental algorithms for
978 scientific computing in Python. *Nat Methods* **17**: 261–272.

979 Walldorf U, Kiewe A, Wickert M, Ronshaugen M, McGinnis W. 2000. Homeobrain, a novel
980 paired-like homeobox gene is expressed in the Drosophila brain. *Mech Dev* **96**: 141–
981 144.

982 Wertheim JO, Murrell B, Smith MD, Kosakovsky Pond SL, Scheffler K. 2015. RELAX:
983 detecting relaxed selection in a phylogenetic framework. *Mol Biol Evol* **32**: 820–832.

984 Wiens JJ, Lapoint RT, Whiteman NK. 2015. Herbivory increases diversification across insect
985 clades. *Nat Commun* **6**: 8370.

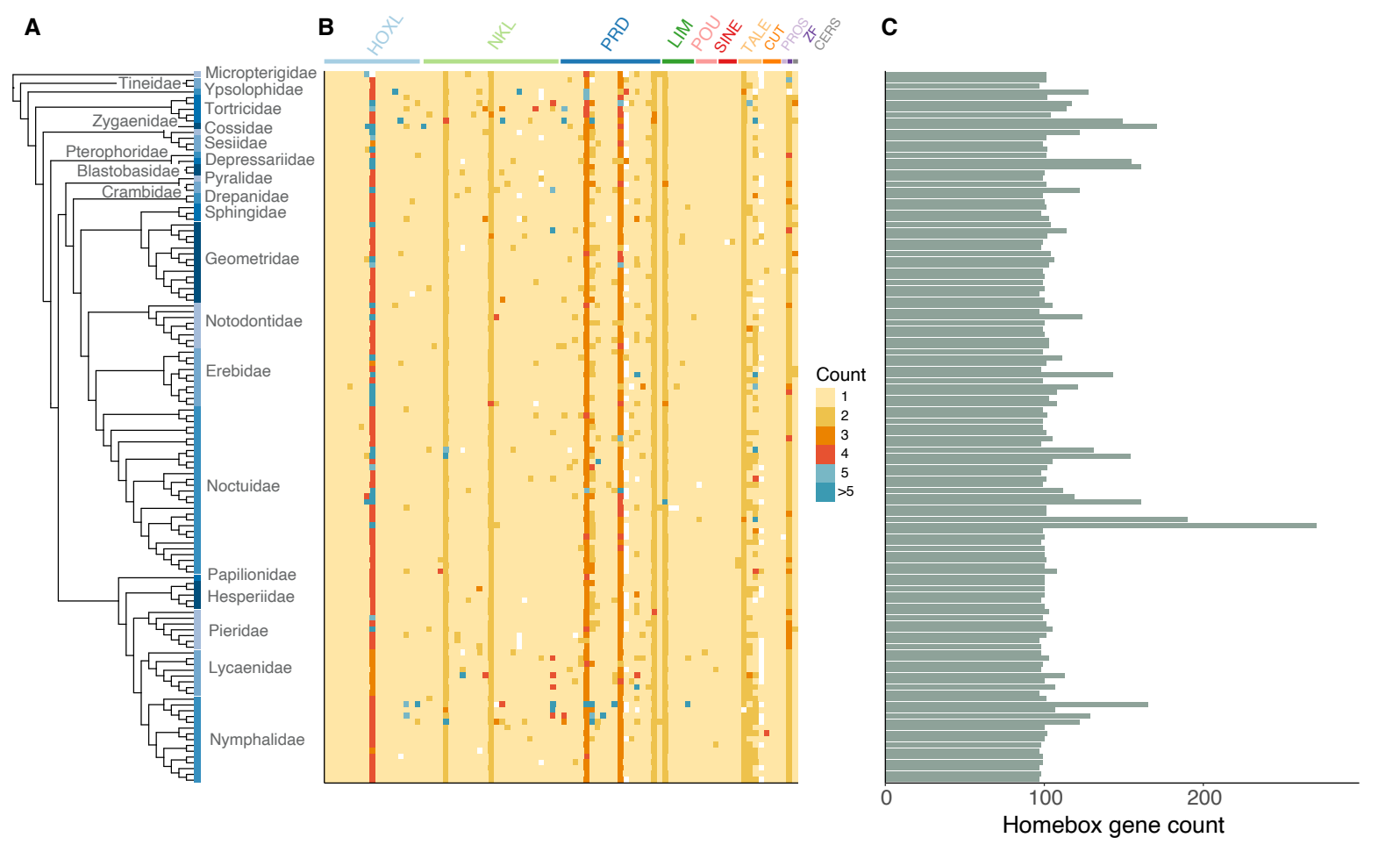
986 Yasukochi Y, Ashakumary LA, Wu C, Yoshido A, Nohata J, Mita K, Sahara K. 2004.
987 Organization of the Hox gene cluster of the silkworm, Bombyx mori: a split of the Hox
988 cluster in a non-Drosophila insect. *Dev Genes Evol* **214**: 606–614.

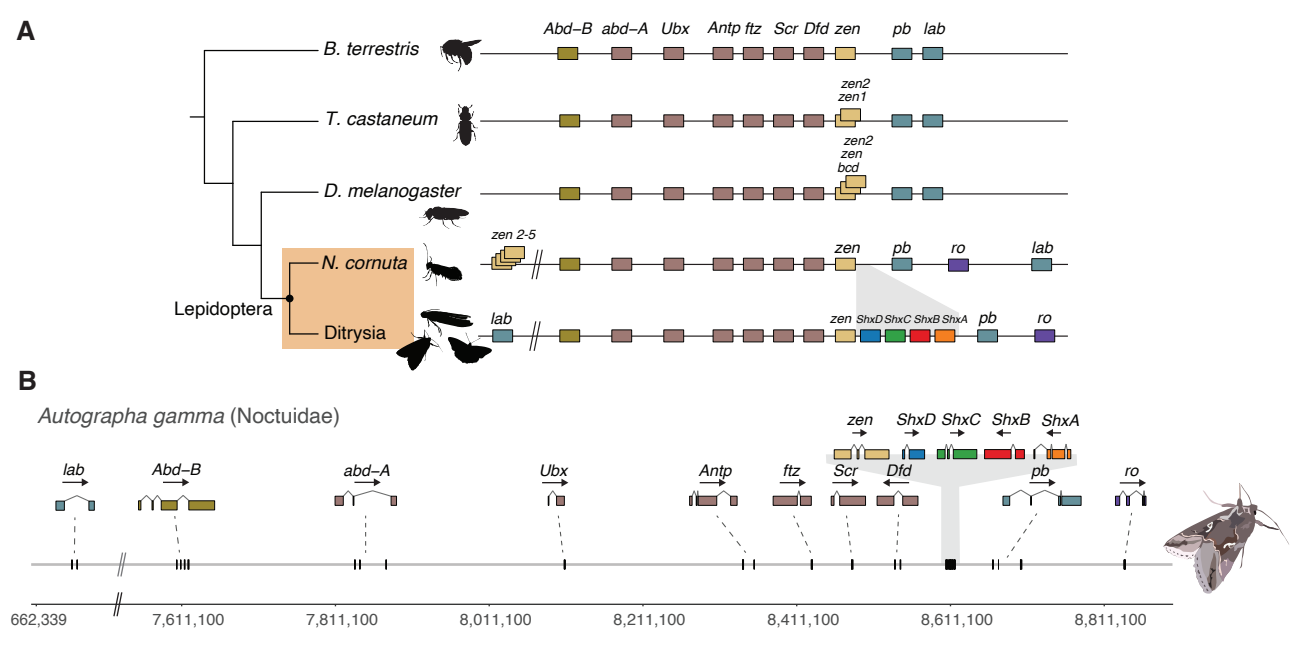
989 Yu G, Smith DK, Zhu H, Guan Y, Lam TT-Y. 2017. ggtree: an r package for visualization and
990 annotation of phylogenetic trees with their covariates and other associated data.
991 *Methods Ecol Evol* **8**: 28–36.

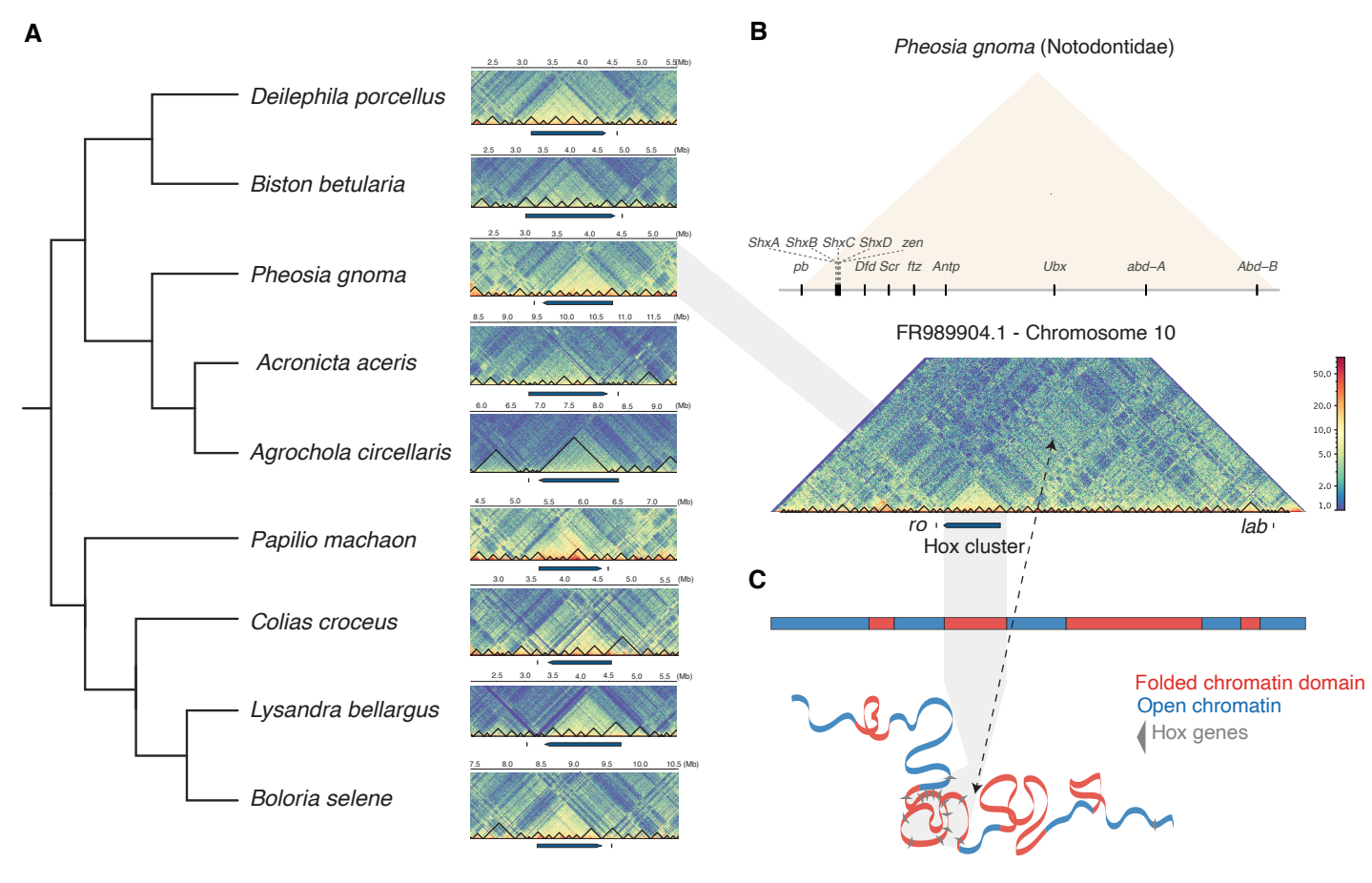
992 Zhang C, Rabiee M, Sayyari E, Mirarab S. 2018. ASTRAL-III: polynomial time species tree
993 reconstruction from partially resolved gene trees. *BMC Bioinformatics* **19**: 153.

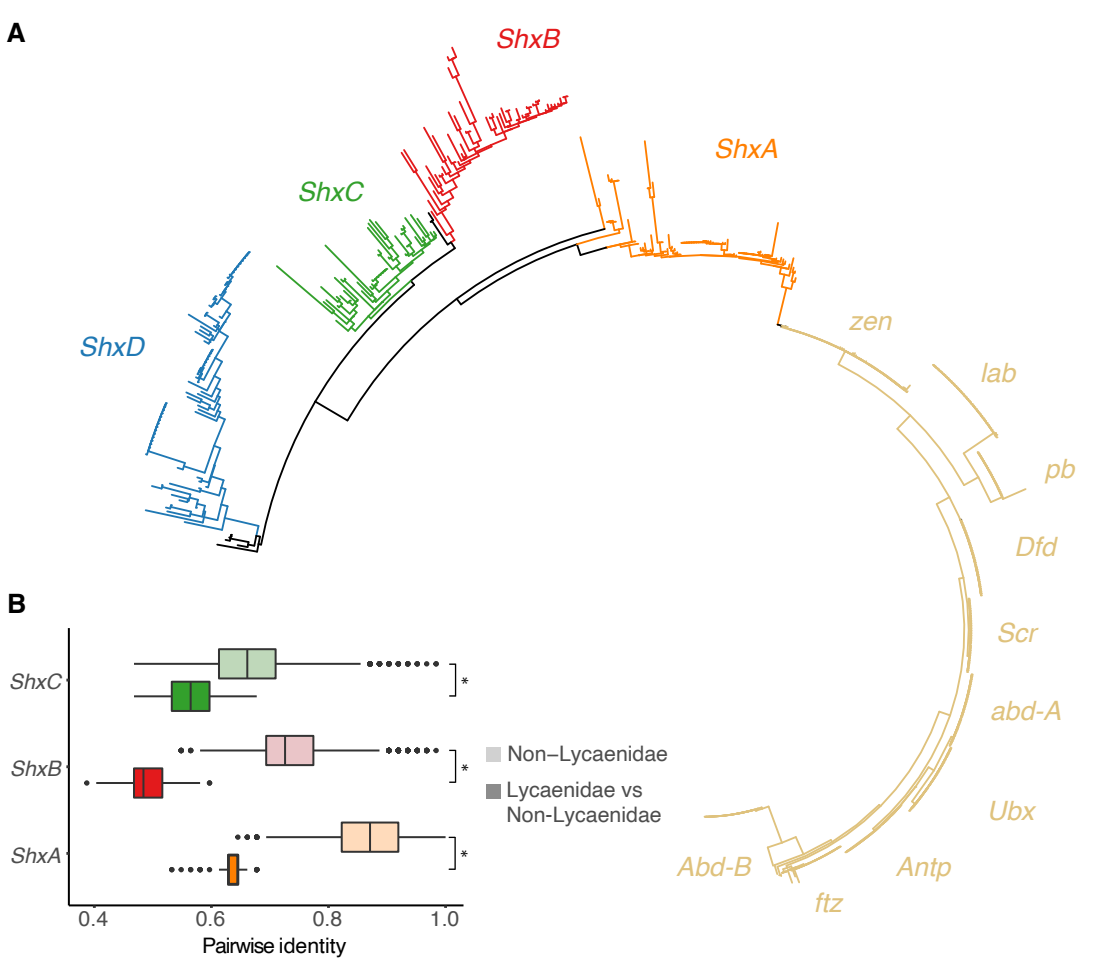
994 Zhong Y-F, Butts T, Holland PWH. 2008. HomeoDB: a database of homeobox gene
995 diversity. *Evol Dev* **10**: 516–518.

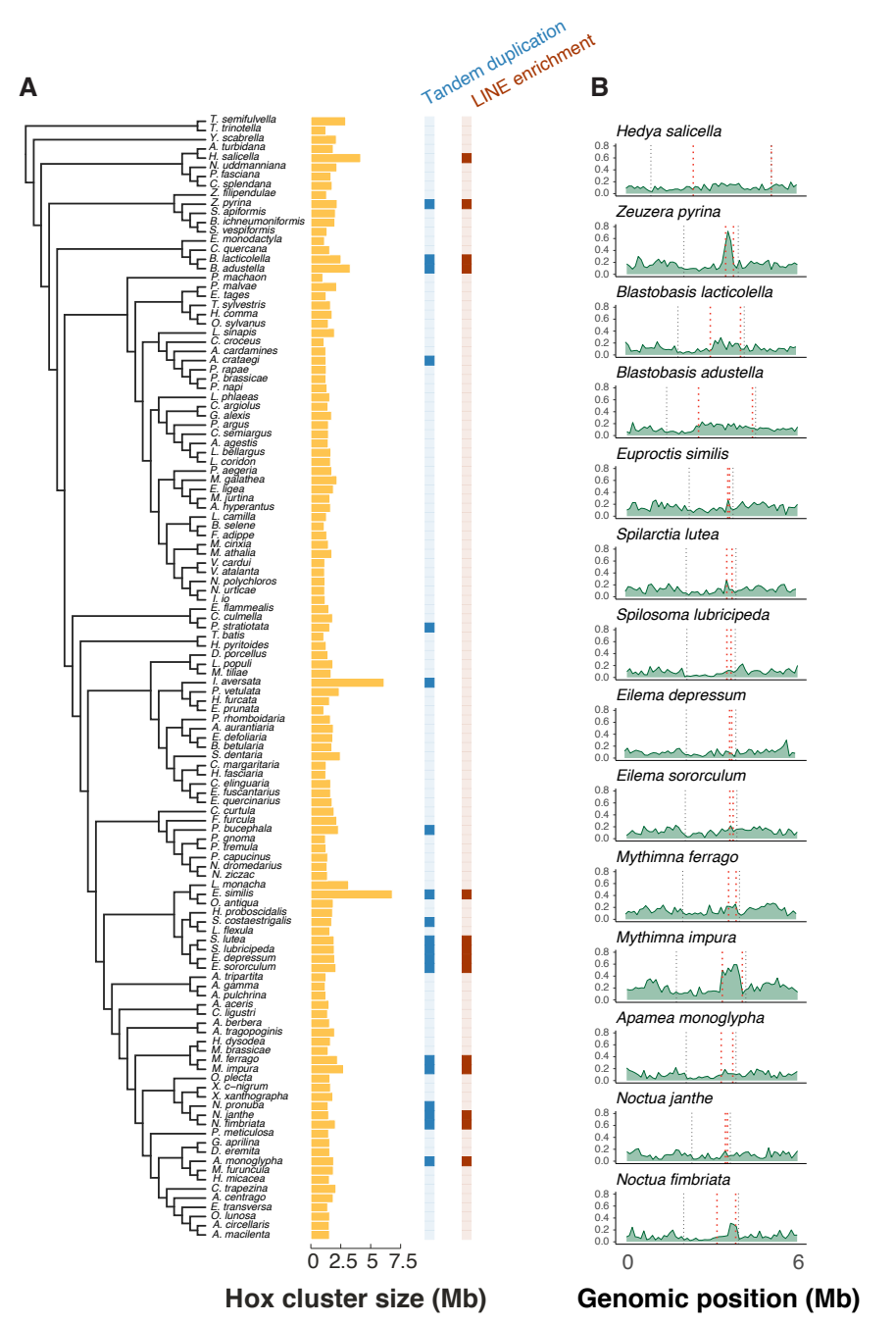
996 Zhong Y-F, Holland PWH. 2011. HomeoDB2: functional expansion of a comparative
997 homeobox gene database for evolutionary developmental biology. *Evol Dev* **13**: 567–
998 568.

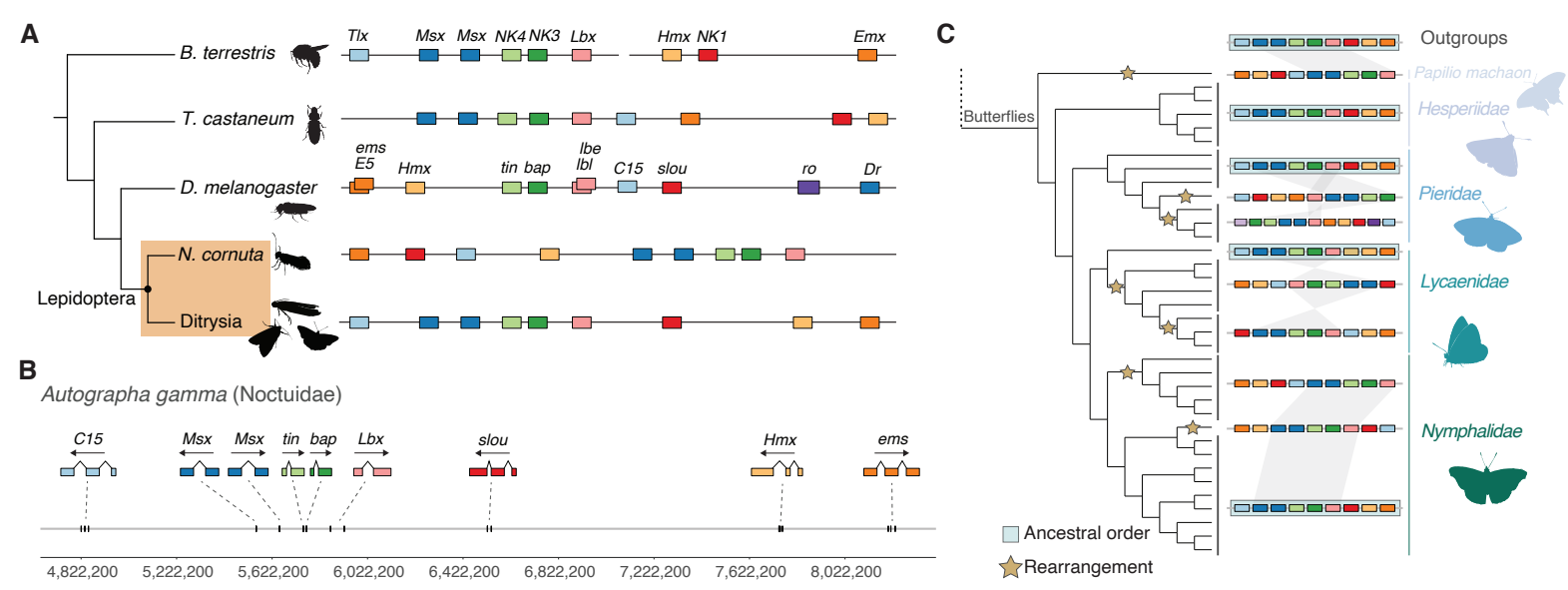

A
B
C
HOXL
NKL
PRD
LIM
POU
SINE
TALE
CUT
PROS
ZF
CERS
Micropterigidae
Tineidae
Ypsolophidae
Tortricidae
Zygaenidae
Cossidae
Sesiidae
Pterophoridae
Depressariidae
Blastobasidae
Pyralidae
Crambidae
Drepanidae
Sphingidae
Geometridae
Notodontidae
Erebidae
Noctuidae
Papilionidae
Hesperiidae
Pieridae
Lycaenidae
Nymphalidae
Count
1
2
3
4
5
>5
0
100
200
Homebox gene count


A
B. terrestris
T. castaneum
D. melanogaster
N. cornuta
Ditrysia
Lepidoptera
Abd–B abd–A Ubx Antp ftz Scr Dfd zen pb lab
zen2
zen1
zen2
zen
bcd
zen 2-5
zen pb ro lab
lab
zen ShxD ShxC ShxB ShxA pb ro
B
Autographa gamma (Noctuidae)
zen ShxD ShxC ShxB ShxA
lab Abd–B abd–A Ubx Antp ftz Scr Dfd pb ro
662,339
7,611,100
7,811,100
8,011,100
8,211,100
8,411,100
8,611,100
8,811,100


A
Deilephila porcellus
Biston betularia
Pheosia gnoma
Acronicta aceris
Agrochola circellaris
Papilio machaon
Colias croceus
Lysandra bellargus
Boloria selene
B
Pheosia gnoma (Notodontidae)
ShxA ShxB ShxC ShxD zen
pb Dfd Scr ftz Antp Ubx abd–A Abd–B
FR989904.1 - Chromosome 10
ro
lab
Hox cluster
C
Folded chromatin domain
Open chromatin
Hox genes


A
ShxB
ShxC
ShxA
ShxD
zen
lab
pb
Dfd
Scr
abd-A
Ubx
Antp
ftz
Abd-B
B
ShxC
ShxB
ShxA
*
*
*
0.4
0.6
0.8
1.0
Pairwise identity
Non−Lycaenidae
Lycaenidae vs
Non-Lycaenidae


A
B
Tandem duplication
LINE enrichment
T. semifulvella
T. trinotella
Y. scabrella
A. turbidana
H. salicella
N. uddmanniana
P. fasciana
C. splendana
Z. filipendulae
Z. pyrina
S. apiformis
B. ichneumoniformis
S. vespiformis
E. monodactyla
C. quercana
B. lacticolella
B. adustella
P. machaon
P. malvae
E. tages
T. sylvestris
H. comma
O. sylvanus
L. sinapis
C. croceus
A. cardamines
A. crataegi
P. rapae
P. brassicae
P. napi
L. phlaeas
C. argiolus
G. alexis
P. argus
C. semiargus
A. agestis
L. bellargus
L. coridon
P. aegeria
M. galathea
E. ligea
M. jurtina
A. hyperantus
L. camilla
B. selene
F. adippe
M. cinxia
M. athalia
V. cardui
V. atalanta
N. polychloros
N. urticae
I. io
E. flammealis
C. culmella
P. stratiotata
T. batis
H. pyritoides
D. porcellus
L. populi
M. tiliae
I. aversata
P. vetulata
H. furcata
E. prunata
P. rhomboidaria
A. aurantiaria
E. defoliaria
B. betularia
S. dentaria
C. margaritaria
H. fasciaria
C. elinguaria
E. fuscantarius
E. quercinarius
C. curtula
F. furcula
P. bucephala
P. gnoma
P. tremula
P. capucinus
N. dromedarius
N. ziczac
L. monacha
E. similis
O. antiqua
H. proboscidalis
S. costaestrigalis
L. flexula
S. lutea
S. lubricipeda
E. depressum
E. sororculum
A. tripartita
A. gamma
A. pulchrina
A. aceris
C. ligustri
A. berbera
A. tragopoginis
H. dysodea
M. brassicae
M. ferrago
M. impura
O. plecta
X. c–nigrum
X. xanthographa
N. pronuba
N. janthe
N. fimbriata
P. meticulosa
G. aprilina
D. eremita
A. monoglypha
M. furuncula
H. micacea
C. trapezina
A. centrago
E. transversa
O. lunosa
A. circellaris
A. maclenta
0 2.5 5 7.5
Hox cluster size (Mb)
Hedya salicella
Zeuzera pyrina
Blastobasis lacticolella
Blastobasis adustella
Euproctis similis
Spilarctia lutea
Spilosoma lubricipeda
Eilema depressum
Eilema sororculum
Mythimna ferrago
Mythimna impura
Apamea monoglypha
Noctua janthe
Noctua fimbriata
0.8
0.6
0.4
0.2
0.0
0
6
Genomic position (Mb)


A
B. terrestris
T. castaneum
D. melanogaster
N. cornuta
Ditrysia
Lepidoptera
Tlx
Msx
Msx
NK4
NK3
Lbx
Hmx
NK1
Emx
ems
E5
Hmx
tin
bap
lbe
lbl
C15
slou
ro
Dr
B
Autographa gamma (Noctuidae)
C15
Msx
Msx
tin
bap
Lbx
slou
Hmx
ems
4,822,200
5,222,200
5,622,200
6,022,200
6,422,200
6,822,200
7,222,200
7,622,200
8,022,200
C
Outgroups
Papilio machaon
Hesperiidae
Pieridae
Lycaenidae
Nymphalidae
Butterflies
Ancestral order
Rearrangement