



## 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](https://doi.org/10.1101/gr.277118.122)

---

<b>P&lt;P</b>	Published online December 8, 2022 in advance of the print journal.
<b>Accepted Manuscript</b>	Peer-reviewed and accepted for publication but not copyedited or typeset; accepted manuscript is likely to differ from the final, published version.
<b>Open Access</b>	Freely available online through the <i>Genome Research</i> Open Access option.
<b>Creative Commons License</b>	This manuscript is Open Access. This article, published in <i>Genome Research</i> , is available under a Creative Commons License (Attribution 4.0 International license), as described at <a href="http://creativecommons.org/licenses/by/4.0/">http://creativecommons.org/licenses/by/4.0/</a> .
<b>Email Alerting Service</b>	Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or <a href="#">click here</a> .

---



---

To subscribe to *Genome Research* go to:  
<https://genome.cshlp.org/subscriptions>

---

Published by Cold Spring Harbor Laboratory Press

# Diversity, duplication and genomic organisation of homeobox genes in Lepidoptera

Peter O. Mulhair<sup>1</sup>, Liam Crowley<sup>1</sup>, Douglas H. Boyes<sup>1,2</sup>, Amber Harper<sup>1,3</sup>, Owen T. Lewis<sup>1</sup>,  
Darwin Tree of Life Consortium<sup>4</sup>, Peter W.H. Holland<sup>1\*</sup>

<sup>1</sup>*Department of Biology, University of Oxford, 11a Mansfield Road, Oxford OX1 3SZ, UK*

<sup>2</sup>*UK Centre for Ecology & Hydrology, Wallingford, OX10 8BB, UK*

<sup>3</sup>*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](mailto:peter.holland@biology.ox.ac.uk)*

**Running title:** Evolution of homeobox genes in Lepidoptera

<sup>4</sup> The list of Darwin Tree of Life Consortium members and affiliations is listed at the end of this  
paper.

## 44 **Abstract**

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

## 64 **Introduction**

65 Lepidoptera (moths and butterflies) are one of the four mega-diverse insect orders, with over  
66 150,000 described species. Lepidoptera belong within the Endopterygota, meaning they  
67 undergo complete metamorphosis with development proceeding from a motile phytophagous  
68 larva to a pupal stage to a reproductive imago (adult). The imaginal stage is easily  
69 recognisable and typically follows a characteristic body plan: two pairs of scale-covered

70 membranous wings, six walking legs, filamentous antennae and a tube-like proboscis. There  
71 are, however, variations and exceptions. For example, in some moths the females are  
72 flightless with reduced wings; in many butterflies four legs rather than six are used for  
73 walking; antennal morphology varies with clubbed, lamellate or plumose structure; in several  
74 moth species the larvae are fully aquatic; and in the family Micropterygidae the adults have  
75 biting rather than sucking mouthparts. Many variations and adaptations are hypothesised to  
76 have been driven by co-evolution with plants, driving novelties in egg laying behaviour, larval  
77 phenotype, and feeding strategies in both larvae and adults (Wiens et al. 2015; Mitter et al.  
78 2017; Kawahara et al. 2019).

79         Associating evolutionary change in form or behaviour to changes in underlying loci is  
80 not straightforward, but insights can come from correlations between patterns in molecular  
81 evolution and changes in phenotype. Homeobox genes are candidates for loci in which  
82 molecular change may cause or facilitate evolutionary change to form and structure of  
83 animals, because most homeobox genes play regulatory roles in development. For example,  
84 the Hox genes, a subset of homeobox genes, encode transcription factors that control  
85 spatial identity along the anteroposterior axis in embryonic development and their number  
86 differs between animal lineages. There was an increase in Hox gene number on the stem  
87 lineage of bilaterian animals, when a head-to-tail axis evolved to dominate the body plan  
88 (Finnerty and Martindale 1998; Nong et al. 2020; Holland 2015); there was also an increase  
89 in the early evolution of vertebrates, traceable to genome duplication (Soshnikova et al.  
90 2013; Aase-Remedios and Ferrier 2021). Hox genes are usually arranged in gene clusters,  
91 but these clusters have been secondarily broken or dispersed in some evolutionary lineages  
92 concomitant with changes to developmental pathways (Ferrier and Holland 2002);  
93 conversely, clusters have been further compacted in vertebrates in association with  
94 additional gene regulatory controls and the emergence of fins and limbs (Duboule 2007). We  
95 wished to address if changes to the homeobox complement in Lepidoptera were associated  
96 with phenotypic change.

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

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

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

## 136 **Results**

### 137 **Classification of all Lepidoptera homeobox genes**

138 We identified all homeobox gene loci in the genomes of 123 lepidopteran species, including  
139 87 moths and 36 butterfly species (Supplementary Table S1) (Mulhair et al. 2022). To place  
140 our analyses in an evolutionary context, we also constructed a phylogenetic tree of the  
141 species analysed using 2,262 BUSCO genes (Methods; Figure 1A, Supplementary Figure  
142 S1). Homeobox sequences were then classified using the characteristic homeodomain and a  
143 combination of reciprocal best BLAST and molecular phylogenetic analysis: this ‘total’  
144 collection of homeobox loci could include functional genes, partial genes and pseudogenes.

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

151 occurred. For example, the *HHEX* (*Hhex*) gene of the ANTP class is absent from the  
152 genomes of all three *Pieris* species sequenced, consistent with a loss in this clade (Figure  
153 1B). We deduce that the *ShxD* gene was lost in the genus *Melitaea*, as it is absent in both  
154 *Melitaea cinxia* (Glanville Fritillary butterfly), consistent with an earlier report (Ferguson et al.  
155 2014), and *Melitaea athalia* (Heath Fritillary Butterfly). Similarly, we do not find the *ShxD*  
156 gene in any of the eight Lycaenidae species in our dataset (*Lycaena phlaeas*, *Celastrina*  
157 *argiolus*, *Glaucopsyche alexis*, *Plebejus argus*, *Cyaniris semiargus*, *Aricia agestis*, *Lysandra*  
158 *bellargus* and *Lysandra coridon*), implying that this gene was also lost early in the evolution  
159 of the family Lycaenidae. Some homeobox genes, such as *Mkx* (orthologous to  
160 Dmel\CG11617) of the TALE class, were lost many times independently across Lepidoptera  
161 (Figure 1B).

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

## 168 **Rearrangement of the Hox gene cluster**

169 Within insects, the Hox gene cluster generally comprises 10 homeobox genes arranged in a  
170 specific order reflecting their evolutionary origin by tandem gene duplication: *lab*, *pb*, *zen*,  
171 *Dfd*, *Scr*, *Antp*, *ftz*, *Ubx*, *abd-A*, *Abd-B*. The cluster may be split, as in many *Drosophila*  
172 species (Negre and Ruiz 2007; Duboule 2007), individual genes may be inverted, and the  
173 *zen* gene may be duplicated (e.g. *zen*, *zen2* and *bcd* in *D. melanogaster*), but radical gene  
174 order changes are rare, documented only within individual species or close relatives (Negre  
175 et al. 2005). A difficulty in studying gene order is that intergenic distances may be large and  
176 many genome assemblies do not provide long-range linkage information. Using  
177 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

205 (TADs) across the genome. These data can reveal the 3D organisation of the chromatin and,  
206 at least in some cases, highlight regions of the genome under common regulatory  
207 constraints (Szabo et al. 2019; Schoenfelder and Fraser 2019). Given that the purpose of  
208 the Hi-C sequencing of these species was to assist in genome assembly (Lawniczak et al.  
209 2022), the depth of Hi-C is lower than in some other studies (Liao et al. 2021), ranging from  
210 ~35 to 52 million paired-end reads (Supplementary Table S2). Nonetheless, we found this  
211 sequencing depth sufficient for the analysis, revealing TADs in lepidopteran genomes which  
212 were visualised at 5kb resolution using HiCEXplorer (Ramírez et al. 2018). To our  
213 knowledge, this is the first such analysis for Lepidoptera genomes, and one of the few  
214 assessments of chromatin accessibility around invertebrate Hox gene clusters (Acemel et al.  
215 2017). In a sample of nine species representing diverse families across Lepidoptera (Figure  
216 3A), we observe strong evidence of an evolutionarily conserved, prominent TAD covering  
217 most of the Hox gene cluster from *pb* to *Abd-B* (Figure 3A & B). This TAD was also observed  
218 in species with a large increase in copy number within the Hox cluster (see “Independent  
219 tandem duplication of *Shx* genes” below) (Supplementary Figure S4A). In all species  
220 analysed *lab* and *ro* are located outside the distinct TAD (Figure 3A). Assessing the wider  
221 chromosomal organisation in *Pheosia gnoma* (Lesser Swallow Prominent), it is clear that  
222 there is a high degree of contact within the Hox-containing TAD relative to the rest of the  
223 chromosome (Figure 3B). While genome-wide conservation of TADs between species has  
224 been questioned (Eres and Gilad 2021), we argue that the strong and consistent signal for  
225 physical contacts across the Hox cluster in diverse moths and butterflies is evidence for a  
226 conserved TAD around a cluster of developmentally important genes. We make no  
227 assessment of possible conservation of other TADs in lepidopteran genomes. Outside the  
228 Hox gene cluster, the general structure of TADs across this chromosome appears similar to  
229 the pattern observed in *Drosophila*, with TADs representing condensed internal interactions  
230 and larger compartments showing long-range interactions between these domains (Figure  
231 3B & C) (Sexton et al. 2012; Ulianov et al. 2016; Szabo et al. 2018; Liao et al. 2021).

## 232 **Origin, duplication and loss of Shx genes**

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

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

## 258 **Independent tandem duplication of Shx genes**

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

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

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

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

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

### 320 **Other homeobox gene clusters**

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

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

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

342         The organisation of the NK gene cluster in *A. gamma* (Silver Y Moth) is typical for  
343 Lepidoptera (Figure 6B). We find a 'core' of five homeobox genes (two *Msx*, *tin*, *bap* and  
344 *Lbx*) spanning ~370kb, plus linkage to *C15* on one side and *slou*, *Hmx* and *ems* on the other  
345 (Figure 6B). The arrangement of these genes is generally conserved across most  
346 Lepidoptera species (Figure 6, Supplementary Figure S9). However, rearrangements within  
347 the cluster are observed in some butterfly lineages. For example, in the three *Pieris* species,  
348 the order of the *tin/bap/Lbx/Dr* core cluster is inverted in all species, *C15* is found on a  
349 separate chromosome, and *Abox* and *Bari* homeobox genes are located close to each end  
350 of the cluster (Figure 6C, Supplementary Figure S9). Rearrangements are also found in both  
351 Lycaenidae and Nymphalidae, with different gene orders suggesting independent  
352 rearrangements (Figure 6C). We infer that a series of translocation and inversion events  
353 have occurred independently. In lineages such as the *Pieris* butterflies, these changes in the  
354 structure of the NK gene cluster reflect general trends of genome remodelling (Hill et al.  
355 2019). The changes within the NK cluster within butterflies represent at least seven likely  
356 rearrangement events, contrasting to the general stability in gene order observed in the Hox  
357 cluster. Rearrangements were also found outside the butterflies, with independent changes  
358 seen in *Ypsolopha scabrella*, *Emmelina monodactyla*, *Carcina quercana*, *Clostera curtula*,  
359 *Laspeyria flexula*, *Abrostola tripartita* and *Neomicropteryx cornuta*. These NK cluster  
360 rearrangements in moths include translocation of one, two or three of the *slou/Hmx/ems*  
361 genes to the opposite end of the cluster, and relocation of *C15* to the opposite end of the  
362 cluster in five of the seven species (Supplementary Figure S9). In contrast to the Hox gene  
363 cluster, the Hi-C contact data do not provide evidence for a strong TAD spanning the NK  
364 homeobox gene cluster (Supplementary Figure S4B).

## 365 **Discussion**

### 366 **Overall stability of homeobox gene numbers**

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

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

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

### 392 **The unusual lepidopteran Hox gene cluster**

393 Hox genes are arranged in genomic clusters as a result of tandem gene duplication, followed  
394 by selective pressure that has kept Hox genes together as neighbours for hundreds of  
395 millions of years. The nature of the selective pressure is not fully understood, but may in part  
396 be related to long range regulatory elements important for spatial colinearity of gene  
397 expression (McGinnis and Krumlauf 1992; Duboule and Morata 1994; Lemons and McGinnis  
398 2006). Some changes to the structure of the Hox gene cluster have been found in insects  
399 (Lewis 1978; Duncan 1987; Ferrier and Akam 1996; Powers et al. 2000; Brown et al. 2002;  
400 Negre and Ruiz 2007), and some larger rearrangements observed in non-insect arthropods  
401 (Cook et al. 2001; Grbić et al. 2011; Chipman et al. 2014; Pace et al. 2016; Leite et al.  
402 2018), but we have a fragmentary picture of insect Hox cluster evolution thus far. Indeed,  
403 within Lepidoptera the complete structure of a Hox gene cluster has not been reported; even  
404 in the pioneering studies on *Bombyx mori* Hox genes, the precise location of the labial gene  
405 could not be resolved (Yasukochi et al. 2004; Chai et al. 2008). With the availability of  
406 chromosomal level genome assemblies, this picture is changing. This study attempts to  
407 characterise Hox gene cluster evolution in an insect order on a large scale. Among the  
408 findings were (a) determining that the labial gene is located at a distant position beyond *Abd-*  
409 *B*, likely relocated by an inversion event, and (b) the finding that the non-Hox gene *ro* is very  
410 closely linked to *pb*, in the position where labial is found in other insects. These two features  
411 are seen in all the ditrysian Lepidoptera we analysed, with an intermediate situation found in  
412 *Neomicropteryx cornuta*, a member of the Micropterygidae. This basal moth has a gene  
413 order of *pb*, *ro* and *lab*, suggesting that movement of *ro* into the Hox gene cluster occurred in  
414 an ancestor of extant Lepidoptera, while the inversion that moved the *lab* gene was a later  
415 event. However, even in *Neomicropteryx cornuta* the *lab* gene is 3.8Mb from the end of the

416 cluster suggesting that it had already ‘escaped’ from common control in the earliest  
417 Lepidoptera.

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

#### 434 **Moths take the record for the most Hox loci**

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

443 common ancestor of vertebrates (Powers and Amemiya 2004; Holland et al. 2008). Genome  
444 duplications during vertebrate evolution increased the total number of Hox genes; for  
445 example, human and mouse have 39 Hox genes, African Butterfly Fish *Pantodon buchholzi*  
446 has 45 Hox genes, Atlantic Eel *Anguilla anguilla* has 73 Hox genes, and Atlantic Salmon  
447 *Salmo salar* has 118 Hox genes and pseudogenes (Mungpakdee et al. 2008; Henkel et al.  
448 2012; Martin and Holland 2014). Our analysis of Lepidoptera genomes has uncovered many  
449 cases of Hox gene duplication, including enormous arrays of Hox-derived loci. We find some  
450 moths have the highest number of Hox loci known to date.

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

464 However, the biggest exceptions to the ‘four Shx’ rule are the cases we find of  
465 independent, very extensive tandem duplication of Shx genes in several evolutionary  
466 lineages of moths. These expansions ranged from the 7 copies in *Schrankia costaestrigalis*  
467 (Pinion-streaked Snout Moth) to an astonishing 165 loci found in *Apamea monoglypha* (Dark  
468 Arches Moth). Other examples include 58 and 66 copies in *Blastobasis lacticolella* and  
469 *Blastobasis adustella* respectively, 19 copies in *Parapoynx stratiotata* (Boyes et al. 2022a),  
470 24 copies in *Phalera bucephala* (Buff-tip Moth; Boyes et al. 2022b), 20 copies in *Spilarctia*

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

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

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

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

## 517 **Material and Methods**

### 518 **Data acquisition**

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

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

### 530 **Homeobox gene identification**

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

## 549 **Homeobox gene expression**

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

## 559 **Species tree inference**

560 A species tree for the 123 lepidopteran species in our dataset was generated using gene  
561 sets obtained from BUSCO v5.1.2 (Manni et al. 2021). First, genes were annotated using the  
562 Lepidoptera BUSCO gene sets. Next, the busco2phylo-nf pipeline  
563 (<https://github.com/lstevens17/busco2phylo-nf>) was used to extract FASTA files for each  
564 annotated gene, ensuring 100% species coverage in each one. Each gene was aligned  
565 using MAFFT v7.467 (Katoh et al. 2005) and gene trees were inferred using IQ-Tree v2.0  
566 (Minh et al. 2020), using ModelFinder to find the model of best fit (Kalyaanamoorthy et al.  
567 2017). Finally, a species tree was inferred using the supertree approach in ASTRAL v5.7.7  
568 (Zhang et al. 2018).

## 569 **Molecular analysis of homeobox evolution**

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

574 2017). To test for changes in rates of homeodomain sequence evolution of the Shx genes  
575 between the Lycaenidae species (which lost ShxD) and all other lepidopteran species with a  
576 normal set of Shx genes, we measure pairwise identity between species as a proxy for  
577 evolutionary rate. This analysis was carried out using PhyKIT using the phykit  
578 pairwise\_identity command (Steenwyk et al. 2021). To measure whether selection was  
579 relaxed or intensified in any of the three remaining Shx genes on any of the Lycaenidae  
580 branches, we used the RELAX model (Wertheim et al. 2015) implemented in HyPhy  
581 (Kosakovsky Pond et al. 2020).

## 582 **Hi-C data processing and TAD identification**

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

## 586 **Repeat annotation and TE density analysis**

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

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

## 626 **Darwin Tree of Life Consortium**

627 Mark Blaxter<sup>5</sup>, Nova Mieszkowska<sup>6,7</sup>, Federica Di Palma<sup>8</sup>, Peter Holland<sup>1</sup>, Richard Durbin<sup>5,9</sup>,  
628 Thomas Richards<sup>1</sup>, Matthew Berriman<sup>5</sup>, Paul Kersey<sup>10</sup>, Peter Hollingsworth<sup>11</sup>, Willie  
629 Wilson<sup>6,12</sup>, Alex Twyford<sup>10,13</sup>, Ester Gaya<sup>10</sup>, Mara Lawniczak<sup>5</sup>, Owen Lewis<sup>1</sup>, Gavin Broad<sup>14</sup>,  
630 Kevin Howe<sup>15</sup>, Michelle Hart<sup>11</sup>, Paul Flicek<sup>15</sup>, Ian Barnes<sup>14</sup>

631 <sup>5</sup> Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridgeshire CB10  
632 1SA, UK

633 <sup>6</sup> Marine Biological Association of the United Kingdom, Citadel Hill, Plymouth PL1 2PB, UK

634 <sup>7</sup> University of Liverpool, Liverpool L69 3BX, UK

635 <sup>8</sup> University of East Anglia, Norwich Research Park, Norwich NR4 7TJ, UK

636 <sup>9</sup> Department of Genetics, University of Cambridge, Cambridge CB2 3EH, UK

637 <sup>10</sup> Royal Botanic Gardens, Kew, Richmond, London TW9 3AE, UK

638 <sup>11</sup> Royal Botanic Garden Edinburgh, Edinburgh EH3 5LR

639 <sup>12</sup> University of Plymouth, Drake Circus, Plymouth PL4 8AA, UK

640 <sup>13</sup> Institute of Evolutionary Biology, School of Biological Sciences, University of  
641 Edinburgh, Edinburgh EH8 9YL

642 <sup>14</sup> Natural History Museum, Cromwell Road, London SW7 5BD, UK  
643 12EMBL-EBI, Wellcome Genome Campus, Hinxton, Cambridgeshire CB10 1SD, UK

644 <sup>15</sup> EMBL-EBI, Wellcome Genome Campus, Hinxton, Cambridgeshire CB10 1SD, UK  
645

646 **Figure Legends**

647 **Figure 1: Numbers of homeobox sequences across Lepidoptera. (A)** Species tree of Lepidoptera  
 648 analysed constructed using BUSCO gene set. Coloured boxes spanning tips of the tree represent  
 649 distinct Lepidoptera families with the family names shown. Species in the tree are listed in the same  
 650 order as in Supplementary Figure S1. **(B)** Heatmap showing numbers of homeobox loci in each gene  
 651 class and subclass (from left to right: *lab*, *Abd-B*, *abd-A*, *Ubx*, *Antp*, *ftz*, *Scr*, *Dfd*, *zen*, *Shx*, *pb*, *ind*,  
 652 *cad*, *exex*, *eve*, *unpg*, *btn*, *Tlx*, *Msx*, *NK4*, *NK3*, *Lbx*, *NK1*, *Hmx*, *Emx*, *Hhex*, *NK7*, *NK6*, *Nedx*, *Dlx*, *En*,  
 653 *NK2.1*, *Msx1x*, *Hlx*, *NK2.2*, *Barhl*, *Bari*, *Bsx*, *Dbx*, *Abox*, *Noto*, *Ro*, *Uncx*, *Gsc*, *Pitx*, *Otp*, *Rx*, *Hbn*,  
 654 *Repo*, *Prrx*, *Shox*, *Arx*, *Pax4/6*, *Phox*, *Prop*, *Vsx*, *CG11294*, *Pax3/7*, *Drgx*, *Otx*, *Lhx6/8*, *Lmx*, *Lhx2/9*,  
 655 *Lhx3/4*, *Lhx1/5*, *Isl*, *Pou2*, *Pou3*, *Pou4*, *Pou6*, *Six3/6*, *Six1/2*, *Six4/5*, *Meis*, *Irx*, *Mkx*, *Pbx*, *Tgif*, *Onecut*,  
 656 *Cux*, *Cmp*, *Prox*, *Zfhx*, *Cers*). **(C)** Total counts of homeobox loci in each genome.  
 657

658 **Figure 2: Hox gene cluster evolution across Insecta. (A)** Comparison of the general structure of  
 659 the Hox gene cluster between representative species for Hymenoptera (*Bombus terrestris*),  
 660 Coleoptera (*Tribolium castaneum*), Diptera (*Drosophila melanogaster*) and Lepidoptera. Lepidoptera  
 661 are shaded in an orange box and split between non-ditrysia species (*Neomicropteryx cornuta*) and  
 662 Ditrysia (represented by 122 species in our dataset). Lepidoptera specific Shx genes are coloured  
 663 orange (ShxA), red (ShxB), green (ShxC) and blue (ShxD) in this figure and throughout the  
 664 manuscript. **(B)** Genomic location of Hox genes in *Autographa gamma* with corresponding exon  
 665 structures and genomic distances annotated below. Silhouette images of *Bombus terrestris*, *Tribolium*  
 666 *castaneum* and *Drosophila melanogaster* were taken from PhyloPic ([phylopic.org](http://phylopic.org)).  
 667

668 **Figure 3: Evidence for a conserved topologically associated domain (TAD) spanning the Hox**  
 669 **gene cluster across Lepidoptera. (A)** Species tree of nine representative lepidopteran species on  
 670 left, HiC matrix showing 1Mb either side of the Hox gene cluster (excluding *lab*). The location of the  
 671 Hox gene cluster from (*Abd-B* to *pb*) is annotated by a blue bar, along with its orientation. The position  
 672 of *ro* is annotated with a short vertical black dash. The intensity of chromatin compaction is  
 673 represented by a blue (low) to red (high) colour gradient. Across the core Hox cluster (*pb* to *Abd-B*) in  
 674 each species a TAD is represented by a region of strong contact, by the more yellow shaded regions.  
 675 Black lines represent TADs or sub-TADs predicted by HiCExplorer. **(B)** Above shows the arrangement  
 676 of the Hox gene cluster (excluding *ro* and *lab*) surrounded by a TAD (orange) in *Pheosia gnoma*  
 677 (Lesser Swallow Prominent). Below displays the HiC matrix of Chromosome 10 showing the location  
 678 of the Hox cluster, represented by a blue bar, along with *ro* and *lab*, represented by short vertical  
 679 black dashes. **(C)** Schematic showing topologically folded domains in Chromosome 10 (red)  
 680 interspersed by chromosome regions with less consistent topology (blue) based on the above HiC  
 681 matrix. Shaded grey region shows the location of the condensed TAD containing the Hox cluster.  
 682

683 **Figure 4: Sequence evolution of Hox genes across Lepidoptera (A)** Phylogenetic tree of Hox and  
 684 Hox-derived homeodomains across 46 Lepidoptera species. Shx gene clades are coloured orange,  
 685 red, green and blue; canonical Hox genes are coloured yellow. The names of the Hox genes are  
 686 placed alongside their clade in the tree. **(B) Shx genes show elevated sequence evolution**  
 687 **following loss of ShxD in Lycaenidae.** Results of pairwise identity of Shx genes between  
 688 Lycaenidae species and non-Lycaenidae species. For each gene (*ShxA-C*) pairwise identity between  
 689 Lycaenidae and all other Lepidoptera species with normal Shx gene count (darker shade boxplot) is  
 690 compared with pairwise identity between all Lepidoptera species with normal Shx gene count (lighter  
 691 shade boxplot). Each pair of boxplots (light shade and dark shade) are coloured according to the  
 692 colour code for each of the Shx genes. Wilcoxon rank-sum test was carried out between pairwise  
 693 identity for Lycaenidae and non-Lycaenidae species (\* = P-value < 0.05).  
 694

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

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

704

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

721

## 722 References

- 723 Aase-Remedios ME, Ferrier DEK. 2021. Improved Understanding of the Role of Gene and  
 724 Genome Duplications in Chordate Evolution With New Genome and Transcriptome  
 725 Sequences. *frontiers in Ecology and Evolution*.  
 726 <http://dx.doi.org/10.3389/fevo.2021.703163>.
- 727 Acemel RD, Maeso I, Gómez-Skarmeta JL. 2017. Topologically associated domains: a  
 728 successful scaffold for the evolution of gene regulation in animals. *Wiley Interdiscip Rev*  
 729 *Dev Biol* **6**. <http://dx.doi.org/10.1002/wdev.265>.
- 730 Bao W, Kojima KK, Kohany O. 2015. Repbase Update, a database of repetitive elements in  
 731 eukaryotic genomes. *Mob DNA* **6**: 11.
- 732 Boyes D, Chadd R, Mulhair P, University of Oxford and Wytham Woods Genome Acquisition  
 733 Lab, Natural History Museum Genome Acquisition Lab, Darwin Tree of Life Barcoding  
 734 collective, Wellcome Sanger Institute Tree of Life programme, Wellcome Sanger  
 735 Institute Scientific Operations: DNA Pipelines collective, Tree of Life Core Informatics  
 736 collective, Darwin Tree of Life Consortium. 2022a. The genome sequence of the ringed  
 737 china-mark, *Parapoynx stratiotata* (Linnaeus, 1758). *Wellcome Open Res* **7**: 121.
- 738 Boyes D, Holland PWH, University of Oxford and Wytham Woods Genome Acquisition Lab,  
 739 Darwin Tree of Life Barcoding collective, Wellcome Sanger Institute Tree of Life  
 740 programme, Wellcome Sanger Institute Scientific Operations: DNA Pipelines collective,  
 741 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, Adolphi 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-Fernández 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, Spaik 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, Dirksen 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.

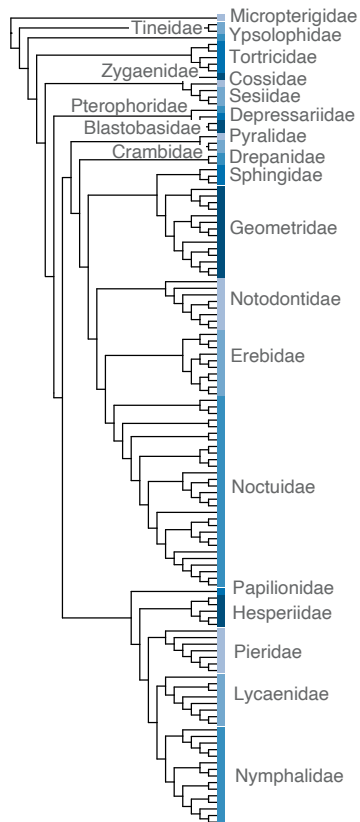
- 825 Holland LZ, Albalat R, Azumi K, Benito-Gutiérrez E, Blow MJ, Bronner-Fraser M, Brunet F,  
826 Butts T, Candiani S, Dishaw LJ, et al. 2008. The amphioxus genome illuminates  
827 vertebrate origins and cephalochordate biology. *Genome Res* **18**: 1100–1111.
- 828 Holland PWH. 2015. Did homeobox 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, Jermini 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, Gimmich 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

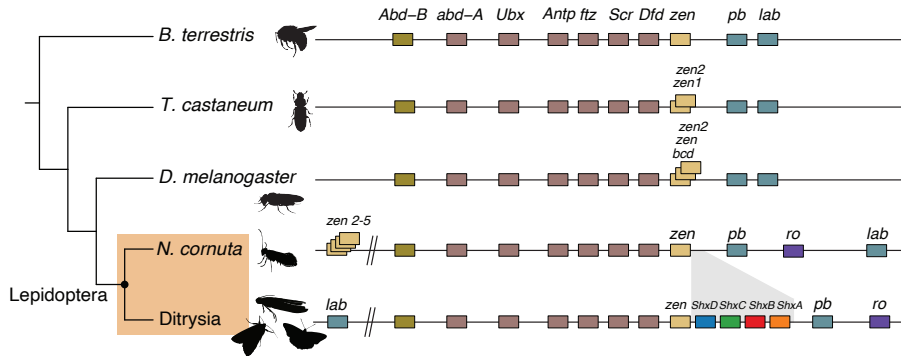
- 867 BioGenome Project. *Proc Natl Acad Sci U S A* **119**.  
868 <http://dx.doi.org/10.1073/pnas.2115639118>.
- 869 Leite DJ, Baudouin-Gonzalez L, Iwasaki-Yokozawa S, Lozano-Fernandez J, Turetzek N,  
870 Akiyama-Oda Y, Prpic N-M, Pisani D, Oda H, Sharma PP, et al. 2018. Homeobox Gene  
871 Duplication 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.

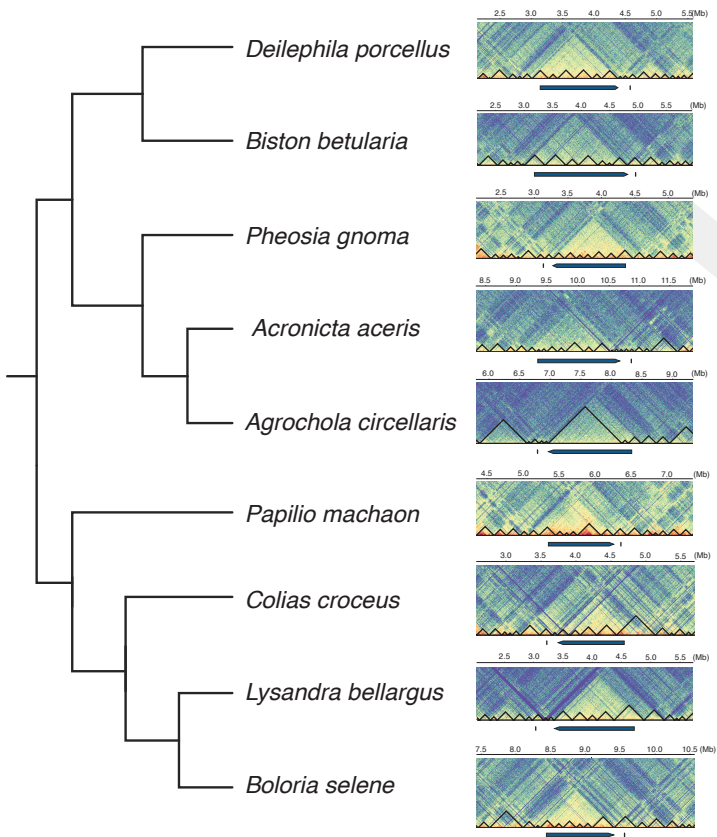
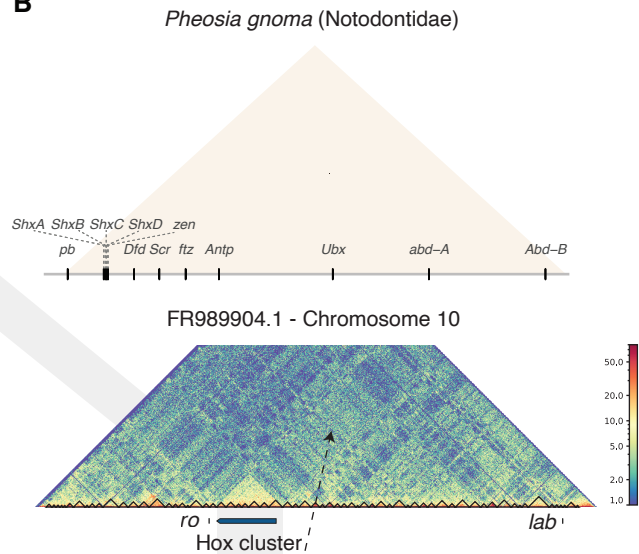
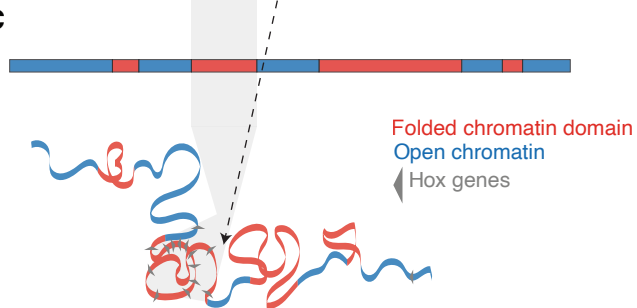
- 992 Zhang C, Rabiee M, Sayyari E, Mirarab S. 2018. ASTRAL-III: polynomial time species tree  
993 reconstruction from partially resolved gene trees. *BMC Bioinformatics* **19**: 153.
- 994 Zhong Y-F, Butts T, Holland PWH. 2008. HomeoDB: a database of homeobox gene  
995 diversity. *Evol Dev* **10**: 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.

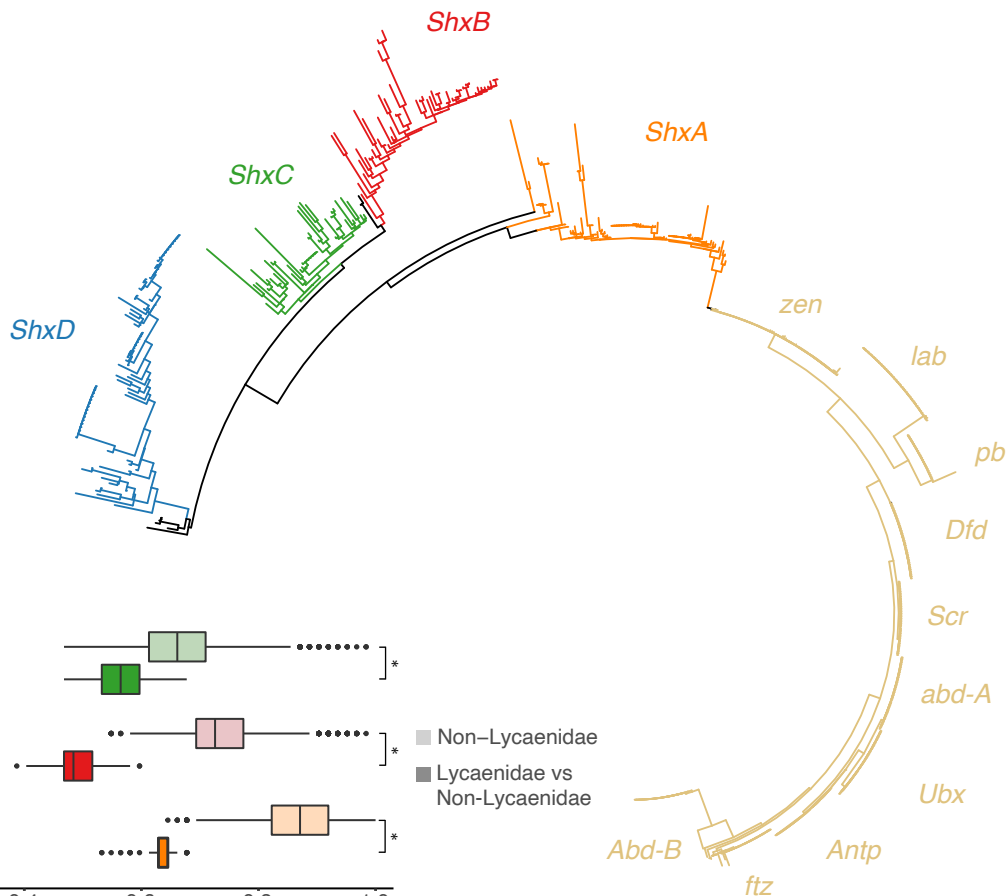
**A****B****C**

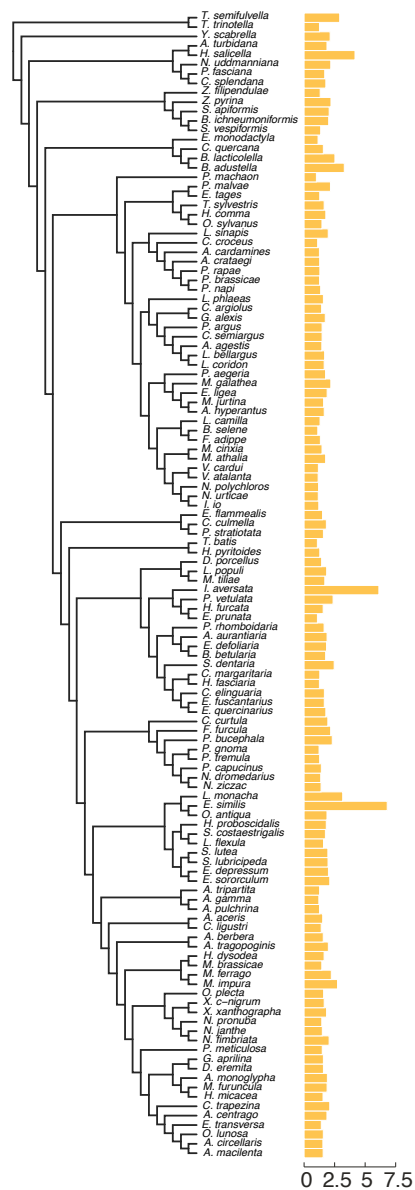
**A****B**

*Autographa gamma* (Noctuidae)

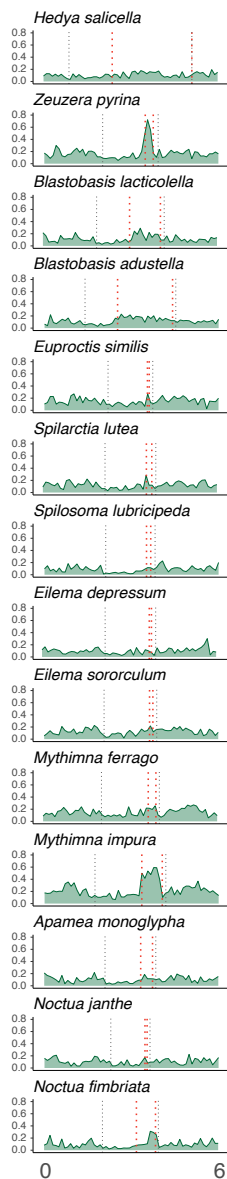


**A****B****C**

**A**

**A**

Tandem duplication  
LINE enrichment

**B**

Hox cluster size (Mb)

Genomic position (Mb)

