Evolutionary innovation through fusion of sequences from across the tree of life

Rishabh R. Kapoor[1 *], Evelyn E. Schwager[2,3], Supanat Phuangphong [3], Emily L. Rivard[4], Chandrashekar Kuyyamudi[2,3], Suhrid Ghosh[2,3], Isobel Ronai[2,3] and Cassandra G. Extavour[2,3,4 *]

1. Systems, Synthetic, and Quantitative Biology Program, Harvard Medical School
2. Howard Hughes Medical Institute
3. Department of Organismic and Evolutionary Biology, Harvard University
4. Department of Molecular and Cellular Biology, Harvard University

* Correspondence RRK rkapoor@g.harvard.edu
CGE extavour@oeb.harvard.edu

15 **Significance**

16 Evolution forges novelty through the repurposing of available parts. Can recently
17 acquired parts, previously foreign to an organism, be similarly repurposed? Applying a rigorous
18 methodology to 319 genomes from arthropods, the largest phylum of animals, we uncover 104
19 novel genes that arose from the fusion of animal genes with fragments acquired via horizontal
20 gene transfer from bacteria, viruses, fungi, and plants. RNA-Seq and RT-PCR across multiple
21 species show that many of these novel genes are expressed as mRNAs. Many show signatures
22 of evolutionary conservation and coherent domain architectures, suggesting that these chimeric
23 genes may play important roles in diverse biological processes. These results reveal an
24 understudied path to evolutionary innovation via “bricolage” of genes from across the tree of life.
25

26

Abstract

Novel genes arise through multiple mechanisms, including gene duplication, gene fusion and horizontal gene transfer (HGT). While HGT has increasingly been documented in animals, the post-transfer evolutionary fate of horizontally-acquired genes is less well understood. We hypothesized that fusion with endogenous sequences in animal genomes might generate what we call “HGT-chimeras”: genes with regions of non-metazoan and metazoan descent in the same open reading frame. To test this hypothesis, we developed a molecular phylogenetics pipeline that enables the identification of HGT-chimeras. We applied our pipeline to 319 high-quality annotated arthropod genomes and uncovered a high-confidence set of 274 HGT-chimeras corresponding to 104 independent origination events across diverse arthropods. HGT-chimeras contain intervals acquired from across the tree of life, and many likely originated via a gene duplication-based mechanism. To assess whether HGT-chimeras might be functionally important, we performed RT-PCR and Sanger sequencing of tissues from 20 arthropod species predicted to harbor HGT-chimeras in their genome. We found evidence for the expression of contiguous chimeric mRNAs for 36 of 41 tested HGT-chimeras across 18 of 20 different tested species. We also found evidence that HGT chimeras evolve under purifying selection and have acquired potentially functional domain architectures, consistent with the hypothesis that these genes are in active use and may participate in diverse biological processes. These results illuminate an underappreciated combinatorial mechanism underlying the origin of novel genes across the largest animal phylum, and suggest that interdomain sequence fusion can play important roles in animal biology and evolution.

Keywords: Molecular evolution, horizontal gene transfer, gene fusion, arthropods

Introduction

Novel genes contribute to adaptation and essential biological functions across the tree of life (1, 2), and have emerged through mechanisms such as duplication (1–3), fission (1, 2), fusion (4–8), and *de novo* birth (9, 10). Among these mechanisms, horizontal gene transfer (HGT) allows organisms to acquire new genes in a non-Mendelian fashion from other species (3, 11–15) and is widely appreciated as a fundamental force of evolution in prokaryotes. (12) Beyond an initial burst of HGT during early stages of eukaryotic symbiogenesis, HGT was long assumed to be vanishingly rare and therefore unimportant in eukaryotic evolution (3, 11, 16). Arthropods, the phylum of animals that constitute at least 80% of all described animal species (17), are emerging as an exception to this expectation. Horizontally acquired genes from bacteria (13–15, 18–23), viruses (14, 24–26), fungi (14, 27, 28) and even plants (14, 29) have been identified in diverse arthropods. In many cases, these genes contribute adaptive functions ranging from ecological defense (23, 29) to metabolism (28, 30). A recent systematic screen for HGT across 218 insect genomes uncovered 741 independent HGT events, placing the rate of HGT on par with that of other rare novel gene formation events within this phylum (14).

Whether horizontally acquired genes retain similar gene structures and functions following transfer, or, like other kinds of novel genes (7, 31–34), rapidly diverge from their original forms, has received less attention than detection of the transfer phenomenon. Here we focus on gene fusion as a possible mechanism of post-transfer divergence. Fusion of endogenous gene sequences is a well-documented source of novelty in protein structure and function (4–8). We hypothesized that, as a result of the fusion of horizontally acquired sequences with pre-existing host genome sequences, animal genomes would contain “HGT-chimeras”: genes with at least one region horizontally transferred from a non-metazoan source in the same open reading frame as a region of ancient metazoan ancestry. To our knowledge, no large-scale systematic screen for HGT-chimeras has been conducted in multicellular eukaryotes, but a few examples have been identified idiosyncratically across diverse taxa (21, 35–38). For instance, we previously attributed the origin of

77 *oskar*, a gene required for essential developmental processes in diverse insect species (39–42), to
78 the formation of a HGT-chimera between a metazoan LOTUS domain (43) and a bacterial GDSL
79 hydrolase domain (44, 45) in the last common ancestor of extant insects (46). In another example,
80 a screen of 20 microbial genomes identified 37 putative interkingdom HGT-chimeras in 11
81 unicellular species, but did not examine multicellular eukaryotes (47). Here, we present the results
82 of a systematic bioinformatic screen for HGT-chimeras across 319 arthropod genomes. We find
83 evidence that HGT-chimeric genes are not only widespread across the tree of arthropods, but also
84 display transcriptional and molecular evolutionary characteristics that suggest they play active roles
85 in organismal biology and evolution.
86
87 **Results and Discussion**
88 *Development of an HGT-chimera detection pipeline*
89 We designed a computational pipeline to search for HGT-chimeras in arthropod
90 genomes (SI Figure 1), which we describe in brief here and detail in SI Methods. The pipeline
91 proceeds in four broad phases (SI Figure 1). In the first phase, we assembled a search set of
92 7,702,369 protein inputs from 319 RefSeq scaffold or chromosome-level genomes (SI Table 1)
93 across eight arthropod classes (SI Figure 4A), excluding proteins from scaffolds <100 kb to
94 reduce the influence of contamination. We clustered the initial search set to 610,348 proteins
95 using MMSeqs2 to promote computational efficiency (48). In the second phase, we used the
96 results of DIAMOND-BLASTp (49) against the NR database (50) to perform an initial screen for
97 putative chimeric proteins. Non-arthropod BLASTp results were first used to partition query
98 proteins into intervals of potentially distinct evolutionary history using a customized
99 implementation of a previously described algorithm (51) (Figure 1A, SI Figure 2A-C, SI File 1).
100 We hypothesized that intervals with a preponderance of hits in non-metazoan taxa and/or lower
101 E-values for non-metazoan than for metazoan hits were of putative HGT ancestry (right-hand
102 interval of Figure 1A), and that those whose lowest E-value hits were in other metazoan taxa or

103 had a preponderance of metazoan hits were of ancient metazoan ancestry (left-hand interval of
104 Figure 1A). We then subjected proteins containing at least one interval of putative HGT ancestry
105 and at least one interval of metazoan ancestry to a second round of BLASTp, using each of the
106 two types of intervals as separate queries to further examine potential HGT and metazoan
107 ancestry annotations from the first round of BLASTp (SI Figure 2D-E). In the third phase, we
108 removed potential confounders of HGT inference (ankyrin repeats and metazoan transposable
109 elements, see SI Methods) and binned HGT-chimeras into groups that we call "similarity
110 clusters" based on their predicted protein domain architecture, using a custom approach that
111 required all sequences in the same cluster to have the same chimeric architecture as
112 determined by BLASTp and hmmsearch (HMMER suite (52). Except where otherwise indicated,
113 downstream analyses (Figures 2-4) were performed with a single representative sequence per
114 similarity cluster. In the fourth and final phase of the pipeline, we assessed the support for HGT
115 and metazoan ancestries of the distinct intervals of each cluster representative using maximum
116 likelihood phylogenetic trees (IQ-TREE (53)) (Figure 1B, SI Figure 3). To improve the sensitivity
117 of homology detection, we constructed these trees from the results of hmmsearch (52) using
118 profile hidden Markov models (HMMs) built from arthropod BLASTp hits as queries, or from
119 BLASTp hits directly when <2 unique, high-confidence (see SI Methods) arthropod BLASTp hits
120 were recovered (E-value for both <1e-2). We considered proteins with at least one interval of
121 metazoan origin, at least one interval of non-metazoan origin, and support for tree topologies
122 with >70% support at the relevant node (bootstrap only considered for HGT intervals, see SI
123 Figure 3), to be candidate HGT-chimeras: genes formed by horizontal sequence transfer
124 followed by in-frame fusion.
125
126 *HGT-chimeras are detected across Arthropoda*
127 Excluding possible within-genome paralogs (see SI Table 6 for copy numbers), we found
128 a final set of 274 HGT-chimera genes across these query genomes, corresponding to 104

129 independent origination events inferred by similarity clustering (SI Table 2). We detected HGT-
130 chimeras across the arthropod phylogeny (Figure 1D), including members of Chelicerata and
131 Pancrustacea. A majority (193/319, 60.5%) of examined RefSeq genomes had at least one
132 HGT-chimera (SI Figure 4B, SI Table 3). We note that *Folsomia candida*, the species with the
133 greatest predicted number of HGT-chimeras, was previously reported to harbor large numbers
134 of non-chimeric horizontally acquired genes (20, 27, 30).

135 To identify the broadest range of arthropod species containing each HGT-chimera (SI
136 Table 4), we additionally searched for representatives of HGT-chimera similarity clusters in an
137 expanded set of arthropod genomes, including an additional 197 genome annotations from
138 GenBank (SI Table 1). We found that 81/104 similarity clusters were restricted to a single
139 species (SI Figure 4C, SI Figure 5A), a result that may reflect a relatively recent evolutionary
140 origin or the sparse genomic sampling of many arthropod taxa (SI text 1). For similarity clusters
141 found in more than one species, we inferred minimum evolutionary gene ages by
142 phylostratigraphy (54) (SI Figure 5B and SI Table 5). The youngest HGT-chimeras (clusters 20
143 and 21) appeared in the congeneric microcrustaceans *Daphnia pulex* and *D. pulicaria*, but were
144 absent in the three other *Daphnia* genomes available at the time of writing. This suggests an
145 origin for these HGT-chimeras as recently as 82,000 years ago in the *D. pulex-D. pulicaria*
146 common ancestor (55) We detected a member of the similarity cluster corresponding to the
147 gene *oskar* (cluster 1) in 181 insects spanning the subclass Pterygota, consistent with our prior
148 finding that *oskar* originated via chimeric HGT in or prior to the last common ancestor of winged
149 insects (>400 million years ago [MYA] (46)). Collectively, these findings suggest that HGT-
150 chimera formation has occurred throughout arthropod evolution.

151 Unexpectedly, we found two clusters with members from multiple distantly related
152 classes of arthropods (Figure 1D). Cluster 18 was found exclusively in the firefly *Photinus*
153 *pyralis* (order: Coleoptera, class: Insecta) and in the mite *Oppia nitens* (order: Sarcoptiformes,
154 class: Arachnida), a finding consistent with convergent evolution of analogous HGT-chimera

155 architectures via independent HGT and gene fusion events (SI Figure 6). In the case of cluster
156 3, molecular phylogenetic analyses suggested that its sparse phylogenetic distribution could
157 have arisen from the transfer of these HGT-chimeras between springtails, fungus gnats, and
158 mites following the formation of this gene (SI Text 2, SI Figure 7). We note that inter-metazoan
159 HGT, although thought to be rare (see for example 56–58), has nevertheless been reported by
160 other studies (30, 58–60), including among these specific groups of arthropods (30, 61). Thus,
161 HGT appears not only to have contributed to the origination of this HGT-chimera, but also to
162 have contributed to its spread across species following its formation.
163
164 *HGT-chimeras are derived from diverse donor lineages*
165 We used HGT interval trees (Figure 1B) to infer donor taxa and found that the majority of
166 HGT intervals plausibly originated in bacterial donors (Figure 1C, SI Table 7). Among 62
167 sequences of inferred bacterial ancestry, close phylogenetic relatives of 37 of them are from
168 bacterial genera of known arthropod symbionts (SI Table 8), including the endosymbionts
169 *Wolbachia*, *Rickettsia* and *Hamiltonella*. This finding is consistent with prior reports of HGT
170 between pathogenic and endosymbiotic bacteria and arthropod genomes (14, 18, 22, 46).
171 In addition to symbiotic bacterial donors, we observed 25 putative transfers from donors
172 with plausible ecological associations with the recipient arthropods, including 15 cases of HGT
173 from arthropod-infecting viruses, one from fungi to the fungus-consuming gnat *Bradysia*
174 *coprophila*, and three from soil-dwelling bacteria to the soil-dwelling mite *Oppia nitens* (SI Table
175 8). In all, these data are consistent with the hypothesis that close ecological associations with
176 food, pathogens and symbionts may provide opportunities for HGT (12, 18, 30, 37, 59).
177
178 *HGT-chimeras are transcribed*
179 Nucleotide composition, codon usage and publicly available RNA-seq data suggested that
180 most HGT-chimeras we predict from genome assemblies and annotations exist in their respective

181 genomes and are expressed as functional gene products (SI Figure 8A-C, SI Table 11-12). For
182 41 chimeras from 20 species for which we could obtain tissue samples, we used RT-PCR and
183 Sanger sequencing to test for the existence of continuous mRNA transcripts containing both
184 predicted HGT and metazoan intervals in the predicted organisms (SI Figure 8D, SI Table 13, SI
185 File 2). We detected transcripts of ~88% (36/41) of examined HGT-chimeras, corresponding to
186 24 distinct similarity clusters across 18 species. This provides strong evidence that the observed
187 HGT-chimeric transcripts are genuine rather than artifacts of genomic contamination,
188 bioinformatic mis-assembly, or mis-annotation.

189 We additionally leveraged a variety of *in silico* tools to predict the domain architecture,
190 structure, and function of HGT-chimeras (see Methods and SI Table 17). In the following sections,
191 we use four RT-PCR-validated examples to illustrate the diverse origins, evolutionary trajectories,
192 and predicted functions of these genes.

193

194 *A shrimp chimera illustrates that HGT-chimeras can rapidly come under the influence of purifying*
195 *selection*

196 HGT-chimera cluster 12 (representative sequence XP_037790819.1) is found in three
197 species of the genus of marine shrimp *Penaeus* (order Decapoda). HGT intervals of these
198 chimeras are phylogenetically nested within a clade of bacterial sequences with high statistical
199 support (Figure 2B; tree topology tests in SI Figure 10), supporting a bacterial origin. The
200 closest bacterial relative of the chimera clade is from the marine genus *Shewanella*, members of
201 which are *Penaeus* pathogens (62), suggesting plausible ecological contact. The HGT interval is
202 a predicted amidase domain with intact catalytic residues (SI Figure 11), and the metazoan-
203 derived interval is annotated as a putative growth factor receptor. This unusual domain
204 architecture (Figure 2A, SI Figure 11A) juxtaposes a metazoan cell signaling domain with a
205 bacterial hydrolase in a likely soluble and extracellular protein (DeepLoc (63) probabilities >0.76
206 and >0.71, respectively; SignalP (64) likelihood > 0.99).

207 The three shrimp-derived sequences are monophyletic in both the HGT and metazoan
208 interval trees (Figure 2B), supporting a single origin of this chimera as recently as 83.1 MYA (SI
209 Figure 5B) (65). We can further place an upper bound on this chimera's age due to its absence
210 in other examined decapod genomes, including *P. japonicus*, which diverged from its chimera-
211 bearing congeneric relatives 114 MYA. Despite its recent origin, this chimera has come under
212 the influence of purifying selection, with gene-wide dN/dS = 0.29 (likelihood ratio test [LRT] p-
213 value 1.06E-45) and no evidence of relaxed constraint in HGT-derived vs non-HGT codons as
214 assessed by fixed-site (partition) models.
215 Similar patterns hold across the broader set of 23 HGT-chimeras found in more than one
216 species: all have gene-wide dN/dS < 1, with neutrality rejected in 21 cases (Figure 2E, SI Table
217 14). Only three clusters showed significant relaxation of constraint in HGT versus non-HGT
218 codons after FDR correction (Figure 2F, SI Table 15). This paucity of relaxed constraint
219 supports the interpretation that many HGT-chimeras are functionally expressed and that
220 horizontally acquired regions contribute to their activity.
221
222 *A copepod chimera illustrates deep conservation and functional similarity among chimera*
223 *intervals*
224 We found representatives of cluster 14 (representative sequence XP_059092480.1) in
225 the genomes of two species of aquatic microcrustaceans, *Eurytemora coralleeae* (order
226 Calanoida) and *Tigriopus californicus* (order Harpacticoida), implying an origin at least as
227 ancient as the last common ancestor of extant copepods ~446 MYA (Figure 2D). Further
228 supporting our inference of deep functional conservation, we observed a strong signature of
229 purifying selection on this gene (dN/dS=0.00374, LRT p-value = 3.18E-67) and detected
230 representatives of this cluster in transcriptomes of 6/7 additional copepod species across two
231 orders.

232 The HGT intervals of this chimera are phylogenetically adjacent to sequences from
233 aquatic bacterial genera (Figure 2D), including *Shewanella* (as in cluster 12 above) and *Vibrio*,
234 the latter of which is a well-documented copepod symbiont (66, 67). We note that the HGT-
235 derived region is a predicted chitin deacetylase with intact catalytic residues (Figure 2A, SI
236 Figure 12A-D), and this enzyme has been implicated in the association of *Vibrio* with their
237 chitinous copepod hosts (68). Strikingly, the metazoan-derived region is annotated as a
238 peritrophin (SI Figure 12E), a class of extracellular chitin-binding proteins that have been
239 implicated in copepod transcriptional responses to *Vibrio* (66). Future work should examine
240 whether this chimera, like its metazoan and bacterial relatives, functions at the host-microbe
241 interface (SI Text 3).
242 The shared chitin-interacting functionality in the HGT and metazoan intervals of this
243 chimera illustrates a broader pattern. Over one third (38/104) of HGT-chimeras displayed
244 putative functional similarity in their HGT- and metazoan intervals, including eight cases in
245 which both HGT and metazoan intervals were predicted to bind to or chemically modify
246 carbohydrates (SI Table 18). These observations, coupled with the repeated observation of
247 fusions of functionally related genes in other studies (4, 47, 69), could mean that HGT-chimeras
248 formed from functionally similar components may be more likely to encode beneficial functions,
249 contributing to their fixation and long-term preservation under purifying selection (4).
250
251 *A mosquito chimera illustrates that preexisting non-chimeric HGT genes may contribute to HGT-*
252 *chimera formation*
253 HGT-chimera cluster 9 (representative sequence XP_021699539.1) is found in three
254 closely related species of mosquitoes (order Diptera), suggesting a minimum age of 92 MYA
255 (65). It consists of a metazoan C2H2 zinc finger region and three tandem HGT-intervals, all
256 annotated by CENSOR (70) as fragments of the auxiliary (non-transposase) protein of fungal
257 Harbinger transposons (Figure 3A). Tree topologies confirm a close relationship of the latter

258 with fungal sequences, but display a complex interleaving of fungal, plant, and oomycete
259 sequences---consistent with documented cross-taxon transposon mobility (37, 71, 72) ---that
260 precludes exact donor inference (Figure 3B). Notably, the closest non-chimeric relatives of the
261 HGT intervals are from multiple dipterans (analogous to Figure 1B'''), including one that harbors
262 the HGT-chimera itself (*Aedes albopictus)*. The observations suggest a multistep model for the
263 evolution of this HGT-chimera (Figure 3C): 1. a horizontal transfer event in a dipteran ancestor
264 resulted in a non-chimeric HGT gene, 2. the standalone HGT gene duplicated, 3. at least one
265 resulting duplicate fused with a metazoan C2H2 zinc finger to form the HGT-chimera (possibly
266 coincident with the duplication event, see SI Figure 9C-D). We found analogous evidence
267 across the full set of HGT chimeras, suggesting that non-chimeric HGT genes often precede
268 HGT-chimera formation, and that gene duplication may contribute to HGT-chimera origination
269 (SI text 4, SI Figure 9).
270 Both auxiliary Harbinger proteins (73) and C2H2 zinc fingers (74) are DNA binding, and
271 we predict intact DNA-binding residues in both the HGT- and metazoan-derived intervals of this
272 HGT-chimera (SI Figure 13), along with an intact nuclear localization sequence (DeepLoc
273 nuclear probability > 0.90). Coupled with gene-wide purifying selection (dN/dS=.27, p-value =
274 2.54E-90), these findings suggest that HGT-derived transposon fragments may have been co-
275 opted into a transcriptional regulatory role, as has been found for other cases of transcription
276 factor – transposon fusions (75, 76). We document 22 HGT chimeras in which both HGT and
277 metazoan intervals are predicted to function in nucleic acid-associated processes (SI Table 18),
278 further illustrating the recurrent pattern of functional similarity raised by the chitin-interacting
279 copepod chimera above.
280
281 *A damselfly chimera illustrates the potential contribution of metazoan gene duplication and*
282 *neofunctionalization to HGT-chimera evolution*

283 Cluster 41 (XP_046403459.1) is restricted to a single species, the damselfly *Ischnura*
284 *elegans* (order Odonata). It encodes a predicted transmembrane protein (DeepLoc probability >
285 0.99), containing a fragment of a metazoan anoctamin transmembrane transporter fused to a
286 fragment of a bacterial ABC transporter likely derived from bacterial endosymbionts (Figure 4A-
287 B). Using MembraneFold (77), we predicted two transmembrane helices of metazoan origin
288 followed by a C-terminal HGT-derived transmembrane helix (Figure 4C). This chimera therefore
289 illustrates the assembly of a novel transmembrane protein via fusion of parts of transmembrane
290 proteins from different domains of life.

291 The metazoan interval tree further shows that the closest relative of this chimera is a non-
292 chimeric anoctamin gene in the *I. elegans* genome (Figure 4D). Thus, paralleling our hypothesis
293 that duplication of preexisting non-chimeric HGT genes contributes to HGT-chimera origination,
294 we posit that duplication of a metazoan anoctamin gene facilitated the fusion of one duplicate with
295 an HGT gene (Figure 3C). We observed a significant elongation of the branch of this tree leading
296 to the HGT-chimera relative to the remaining damselfly branches, suggesting that the fusion event
297 was followed by accelerated sequence evolution. Consistently, we fit a two-ratio dN/dS model
298 (78–80) to this metazoan interval tree and found that the non-chimeric metazoan sequences have
299 evolved under strong purifying selection (dN/dS=.0241), while the HGT-chimera diverged under
300 the influence of strong positive selection (dN/dS = 2.75, LRT p-value = 1.1E-10).

301 More broadly, across species- or genus-specific HGT-chimera intervals with non-chimeric
302 phylogenetic relatives in the same taxonomic order, 64.3% (18/28) have higher dN/dS on the
303 chimeric branches. A higher dN/dS on chimeric vs. non-chimeric branches is statistically
304 supported by an LRT in 32.1% (9/28) of such intervals (Figure 4E-F, SI Table 16). This suggests
305 that accelerated protein sequence evolution may be a common fate of incipient HGT-chimeras,
306 consistent with the functional divergence (neofunctionalization) observed in other novel genes
307 that have evolved through gene fusion and/or duplication (4, 7, 8, 31, 34, 81).

308 Post-fusion functional divergence has been experimentally shown for the HGT-derived
309 domain of the HGT-chimeric gene *oskar*. The closest bacterial relatives of *oskar*'s HGT-derived
310 domain, called the OSK domain (44), are lipid hydrolases (40, 45, 46), but in insects this domain
311 lacks key hydrolytic residues and instead binds mRNAs (44, 82). Thus, HGT-chimeras reflect
312 evolutionary innovation not by only the juxtaposition of pre-existing sequences from different
313 domains of life, but also through post-fusion sequence divergence.
314
315 **Conclusions**
316 Novel protein structure and function can emerge from several elementary events,
317 including amino acid substitutions, gene duplication, gene fusion and HGT (3). In this study, we
318 have shown that multiple of these elementary events can operate together in the origins of a
319 previously understudied class of genes that we call HGT-chimeras. In at least 104 independent
320 events across the history of the arthropods, HGT-chimeras evolved via the fusion of
321 endogenous genes of ancient metazoan ancestry with genes acquired via horizontal transfer
322 from non-metazoan sources. Sequence duplication within the host genome played a plausible
323 role in the formation of many HGT-chimeras (Figure 3B-C, Figure 4D, SI Figure 9), and we
324 additionally detected a signature of accelerated amino acid substitution following HGT-chimera
325 birth (figure 4D). Collectively, these findings reveal a rich post-transfer history of horizontally
326 acquired sequences in arthropods.
327 We propose that although chimeric genes of this nature have been reported rarely (21,
328 35–38, 46, 47) despite significant prior work on HGT (12–16, 18, 19, 22, 24–29), we were able
329 to uncover these HGT-chimeras by better demarcating intervals of distinct history (Figure 1A, SI
330 Figure 2). HGT-chimeras are unlikely to be detected by traditional HGT detection methods,
331 which assume a common evolutionary history throughout the length of the open reading frame.
332 In almost half (41/104) of HGT-chimeras, a naive interpretation of gene-wide BLASTp results via
333 the same criteria used to infer interval ancestry would lead to a false inference of gene-wide

334 metazoan ancestry (see SI Table 19). For the remaining genes, constructing a single gene tree
335 for genic intervals with distinct histories would be methodologically problematic and, even if it
336 resulted in a sensible topology, would obscure the true chimeric history of the gene. We expect
337 that the application of methods like the one we describe here, that allow for intragenic
338 phylogenetic discordance, could lead to the discovery of more HGT-chimeras, especially in
339 lineages with frequent HGT such as grasses and parasitic plants (83, 84), fungi (85, 86) and
340 prokaryotes (3, 12).

341 The biological significance of HGT has been questioned by some researchers, as many
342 transferred segments show evidence of pseudogenization (15, 22, 87), and significant barriers
343 to the transfer and functional integration of genes from divergent lineages have been posited
344 (11, 16, 88–90). Motivated by the functionally coherent fusions discussed in the insect examples
345 above Figures 2-4), we speculate that fusion with pre-existing gene sequences might facilitate
346 the integration of novel horizontally transferred sequences into pre-existing functional networks.
347 Many HGT-chimeras show signatures consistent with biological function, including multiple lines
348 of evidence for active transcription (SI Figure 8, SI Tables 12-13) and gene-wide purifying
349 selection (Figure 2E-F, SI Tables 14-15). Experimental studies in diverse model systems have
350 demonstrated important functions for many young and species-specific genes (6, 10, 33, 91–
351 93). Thus, we consider it plausible that even the young HGT-chimeras detected here may play
352 important biological roles. Our *in silico* functional predictions (SI Tables 17-18) provide insights
353 into the potential range of biological processes impacted by HGT-chimeras, warranting
354 experimental validation in future studies.

355 To conclude, our novel computational approach has revealed that fusion of horizontally
356 acquired sequence with endogenous ones has been widespread throughout the history of
357 arthropods, illustrating a previously underappreciated versatility in evolutionary innovation via
358 genomic “bricolage” (94).

359

Methods

Software used in the “Development of an HGT-chimera detection pipeline” section included HMMER v3.3.2 (52), MUSCLE v5.1 (99), trimAl (100), and iTOL v5 (95) . dN/dS ratios were computed with PAL2NAL v14.1 (101) and PAML v4.10.0 (68–70). All BLASTp searches were performed with DIAMOND v2.0.15 (49). RT-PCR and Sanger sequencing were performed on arthropod samples from donations or lab cultures (SI Table 13). Full methods are provided in SI.

Funding sources

This work was supported by a National Science Foundation (NSF) Graduate Research Fellowship Program award 2023356057 to RK, the NSF-Simons Center for Mathematical and Statistical Analysis of Biology at Harvard (award number DMS-1764269) and the Harvard Quantitative Biology Initiative (ELR and CK), a Herchel Smith Graduate Fellowship awarded to ELR, award number RGP0041/2022 from the Human Frontier Science Program (SP) to CGE, the Howard Hughes Medical Institute, and funds from Harvard University. IR is a Howard Hughes Medical Institute Awardee of the Life Sciences Research Foundation, and CGE is an Investigator of the Howard Hughes Medical Institute.

Acknowledgements

We are grateful to Gregg Thomas, Sean Eddy, Andrew W. Murray, Gonzalo Giribet and the Extavour lab for critical feedback and discussion. Computational resources were provided and maintained by Harvard Faculty of Arts and Sciences Research Computing. Arthropod specimens for RT-PCR were generously donated by the researchers listed in SI Table 13. A special thank you to Avinash Tope and Andrew Ray (Kentucky State University) for allowing us to use their lab space and materials for *Cherax quadricarinatus* and *Penaeus vannamei* sample preparation, to John Stivers for assistance with the dissection of these specimens, to Yu Shirai (Extavour lab) for *Tribolium castaneum* RT-PCR, and to Tarun Kumar (Extavour lab) for *Bombyx mori* dissection and extraction of *Tyrophagus putrescentiae* cDNA.

Data availability

All data required to reproduce associated figures is included in Supporting Information tables. Maximum likelihood trees for HGT and metazoan intervals of the final set of HGT-chimera similarity clusters are freely accessible in the public project “Arthropod HGT-chimera interval trees 8/27/2025” on the iTOL webserver at https://itol.embl.de/shared/rkapoor. Associated BLAST results, hmmsearch results, multiple sequence alignments, and Newick trees have been deposited at the Dryad digital repository for this study https://doi.org/10.5061/dryad.t1g1jwtdz. Scripts used to implement the HGT-chimera detection pipeline and downstream analysis are available on the github repository https://github.com/rishabhrajkapoor/Arthropod-HGT-chimeras-2025 with commit ID 6a974df. Sanger-verified cDNA sequences have been deposited at GenBank, with accession numbers available in SI Table 13.

399
400 **Author contributions**
401 Project was conceptualized by CGE and RK. RK performed bioinformatic analyses. SP, EES
402 and IR prepared cDNA libraries. RT-PCRs were performed by EES, ELR, CK, SG, SP, and IR.
403 EES inspected and manually corrected Sanger sequencing results. RK, EES, ELR, and CGE
404 wrote the manuscript and created associated figures. All authors reviewed the manuscript.
405

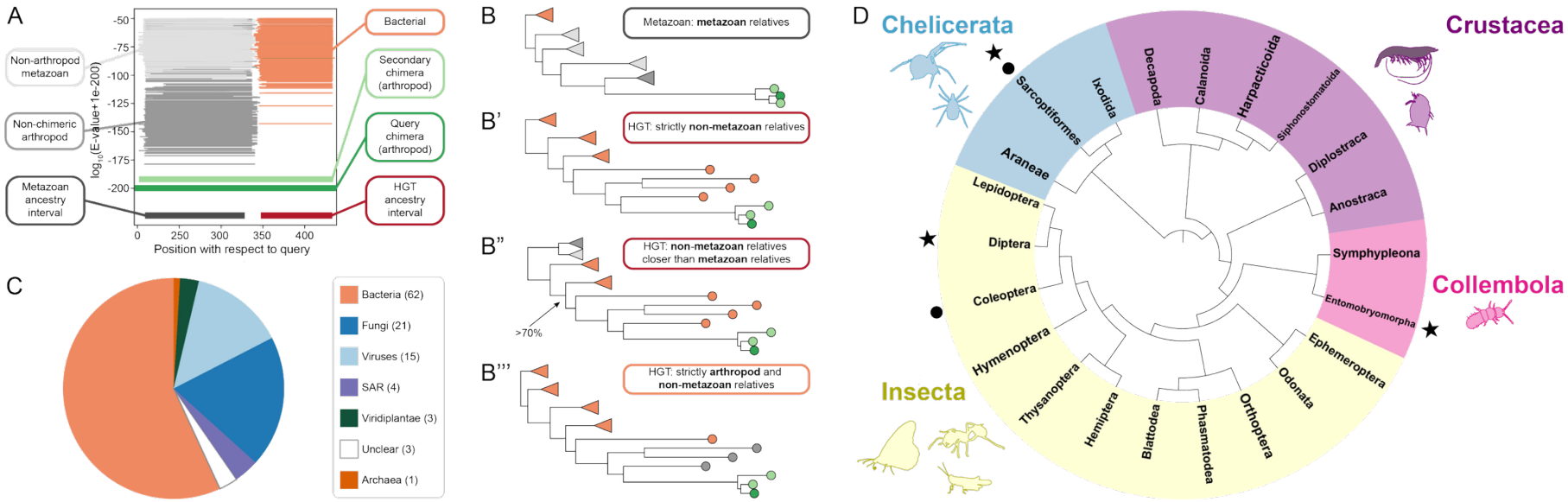
A
Non-arthropod metazoan
Non-chimeric arthropod
Metazoan ancestry interval
Bacterial
Secondary chimera (arthropod)
Query chimera (arthropod)
HGT ancestry interval
log10(E-value+1e-200)
-50
-75
-100
-125
-150
-175
-200
0
250
300
350
400
Position with respect to query
B
Metazoan: metazoan relatives
B'
HGT: strictly non-metazoan relatives
B"
HGT: non-metazoan relatives closer than metazoan relatives
>70%
B'''
HGT: strictly arthropod and non-metazoan relatives
C
Bacteria (62)
Fungi (21)
Viruses (15)
SAR (4)
Viridiplantae (3)
Unclear (3)
Archaea (1)
D
Chelicerata
Crustacea
Collembola
Insecta
Sarcoptiformes
Ixodida
Decapoda
Calanoida
Harpacticoida
Siphonostomatoida
Diplostraca
Anostraca
Symphypleona
Entomobryomorpha
Ephemeroptera
Odonata
Orthoptera
Phasmatodea
Blattodea
Hemiptera
Thysanoptera
Hymenoptera
Coleoptera
Diptera
Lepidoptera
Araneae

Figure 1. HGT-chimera detection across Arthropoda. (A) The taxonomic distribution of BLASTp hits across a query sequence are used to hypothesize chimeric ancestry. In this hypothetical example, the interval 1-325 amino acids is hypothesized to be of metazoan ancestry (darkest grey) because the closest non-arthropod hits are from non-arthropod metazoan taxa. The interval from 350-450 amino acids is hypothesized to be of HGT ancestry (red) because the closest hits are of non-metazoan origin. In addition to the original query sequence (dark green), secondary chimeric sequences of the same domain architecture (light green) are retrieved from other arthropods. Plots for all HGT-chimeras are available in SI File 1. **(B)** HGT and metazoan ancestries are further tested for each interval using maximum likelihood phylogenetics on hmmsearch or BLASTp hits as described in SI Methods. Circles and triangles represent single sequences and collapsed multi-sequence clades, respectively. **B** shows a hypothetical interval of metazoan ancestry due to the absence of close non-metazoan relatives of the chimeric sequences. **B'** shows the simplest case consistent with HGT from a non-metazoan donor, in which all detected phylogenetic relatives outside of other orthologous arthropod chimeras are from non-metazoan taxa. In **B''**, non-chimeric metazoan relatives are detected, but an HGT origin is supported by nesting of the chimeric sequences within a clade of non-metazoan sequences with >70% bootstrap support. In **B'''**, the closest relatives are non-chimeric arthropod sequences followed by non-metazoan sequences, supporting HGT from non-metazoans to arthropods prior to the chimera-forming fusion event. **(C)** Plausible HGT donor taxa at the domain/superkingdom level inferred from the taxonomic distribution of sister and cousin clades on maximum likelihood phylogenies for 109 HGT intervals from 104 HGT-chimera clusters (see SI Methods). SAR refers to the eukaryotic supergroup consisting of stramenopiles, alveolates, and rhizarians. **(D)** Arthropod orders with species containing at least one HGT-chimera. Colored ranges indicate taxonomic class (Insecta, Collembola) or subphylum (Chelicerata, Crustacea). Data include sequences from both the primary search set of 319 RefSeq genomes and 197 additional GenBank genomes (see SI Methods). Symbols outside the phylogram indicate the presence of HGT-chimeras in clusters 18 (circles) and 3 (stars), which are notable for their sparse phylogenetic distribution (SI Figure 6-7).

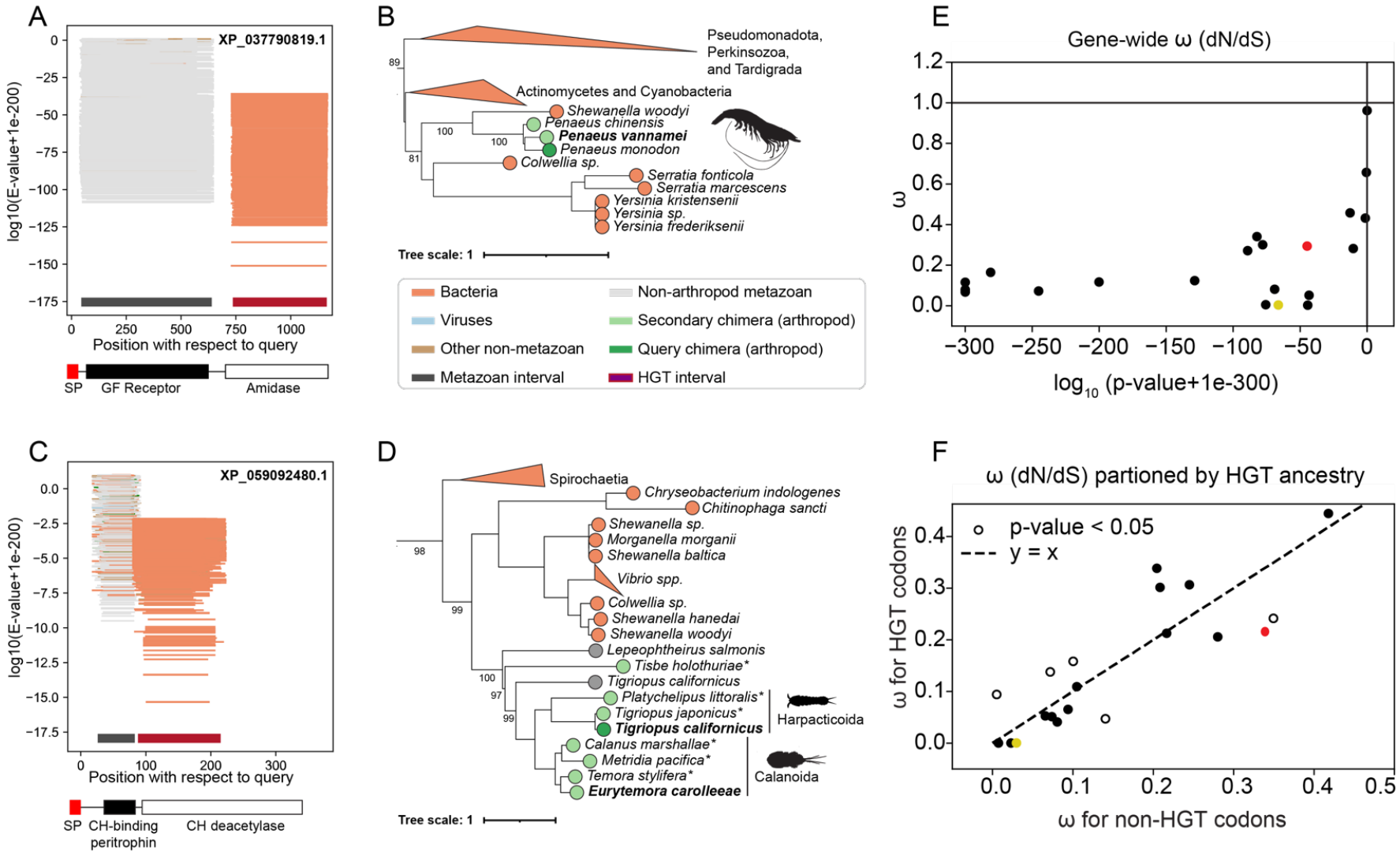
A
XP_037790819.1
log10(E-value+1e-200)
Position with respect to query
SP
GF Receptor
Amidase
B
Pseudomonadota, Perkinsozoa, and Tardigrada
Actinomycetes and Cyanobacteria
Shewanella woodyi
Penaeus chinensis
Penaeus vannamei
Penaeus monodon
Colwellia sp.
Serratia fonticola
Serratia marcescens
Yersinia kristensenii
Yersinia sp.
Yersinia frederiksenii
Tree scale: 1
Bacteria
Viruses
Other non-metazoan
Metazoan interval
Non-arthropod metazoan
Secondary chimera (arthropod)
Query chimera (arthropod)
HGT interval
E
Gene-wide ω (dN/dS)
ω
log10 (p-value+1e-300)
C
XP_059092480.1
log10(E-value+1e-200)
Position with respect to query
SP
CH-binding peritrophin
CH deacetylase
D
Spirochaetia
Chryseobacterium indologenes
Chitinophaga sancti
Shewanella sp.
Morganella morganii
Shewanella baltica
Vibrio spp.
Colwellia sp.
Shewanella hanedai
Shewanella woodyi
Lepeophtheirus salmonis
Tisbe holothuriae*
Tigriopus californicus
Platychelipus littoralis*
Tigriopus japonicus*
Tigriopus californicus
Calanus marshallae*
Metridia pacifica*
Temora stylifera*
Eurytemora carolleeae
Harpacticoida
Calanoida
Tree scale: 1
F
ω (dN/dS) partioned by HGT ancestry
p-value < 0.05
y = x
ω for HGT codons
ω for non-HGT codons

Figure 2. Crustacean chimeras of varying age have evolved under purifying selection. **(A)** BLASTp plot as in Figure 1A, but excluding arthropod hits, for the query chimera XP_037790819.1 from the shrimp *Penaeus monodon* (cluster 12). Metazoan ancestry is inferred for the interval (45, 640) and HGT ancestry for the interval (736, 1165). Domain annotations obtained via NCBI CDD search and InterProScan (SI Table 17) are shown across the bottom of the plot, with an N-terminal signal peptide predicted by SignalP-6.0 (cleavage site between positions 25-26). See SI Figure 11 for detailed annotation. **(B)** Maximum likelihood tree for the HGT interval (736, 1165) from XP_037790819.1, including intervals from secondary chimers found in the two other congeneric shrimp *P. vannamei* (XP_069990332.1) and *P. chinensis* (XP_047489075.1). **(C)** BLASTp plot excluding arthropod hits for the query chimera XP_059092480.1 (cluster 14) from the copepod *Tigriopus californicus*. Metazoan ancestry is inferred for the interval (26, 83) and HGT ancestry for the interval (88, 215). Domain annotations are shown across the bottom of the plot, with an N-terminal signal peptide cleaved between positions 29-30. See SI Figure 12 for detailed annotation. **(D)** Maximum likelihood tree for the HGT interval (88, 215), including secondary chimera sequences from the genome of the copepod *Eurytemora carolleeae* (XP_023343432.1) and additional sequences obtained from transcriptome shotgun assembly data from seven additional copepod species (indicated with an asterisk). The phylogenetic split among HGT-chimeras from the copepod orders Harpacticoida and Calanoida is indicated with vertical lines. In B and D, collapsed multi-sequence clades are represented as triangles. Triangle color represents the taxonomic origin of the majority of contained sequences, and the lengths of the two edges connecting each triangle to its parent node represents the summed branch length to the longest and shortest branch in the clade. Bolded text in B and D indicate sequences verified by RT-PCR, scale bars indicate one substitution per amino acid site, and numerical values at internal nodes indicate ultrafast bootstrap support values for select nodes relevant for HGT inference. Only a portion of both trees is shown here; the full trees for both metazoan and HGT trees with accessions are available on iTOL. **(E)** Gene-wide ω (dN/dS) values fitted for 21 of the 23 HGT-chimera similarity clusters that were found in the genomes of at least two different species (excluding clusters 3 and 18). The x-axis represents the log-transformed Benjamini-Hochberg corrected p-values from a likelihood ratio test comparing the likelihood of the free ω ratio values (y-axis) to a null model with ω fixed at 1. **(F)** Each of 23 HGT-derived intervals from the same 21 HGT-chimeras as in (E) were permitted to take a different ω value from the rest of the protein using fixed-site models. Benjami-Hochberg corrected p-values (indicated by dot fill) were obtained by comparing the likelihood of a model in which codons in the HGT interval differed from the remaining codons in their ω values; open dots indicate statistically significant deviation of dN/dS values for an HGT-derived interval from the rest of the coding region. Three intervals with y-axis values >0.50 and p_adj >0.05 have been ommited from this plot. Red and yellow circles in (E) and (F) highlight clusters 12 and 14, respectively.

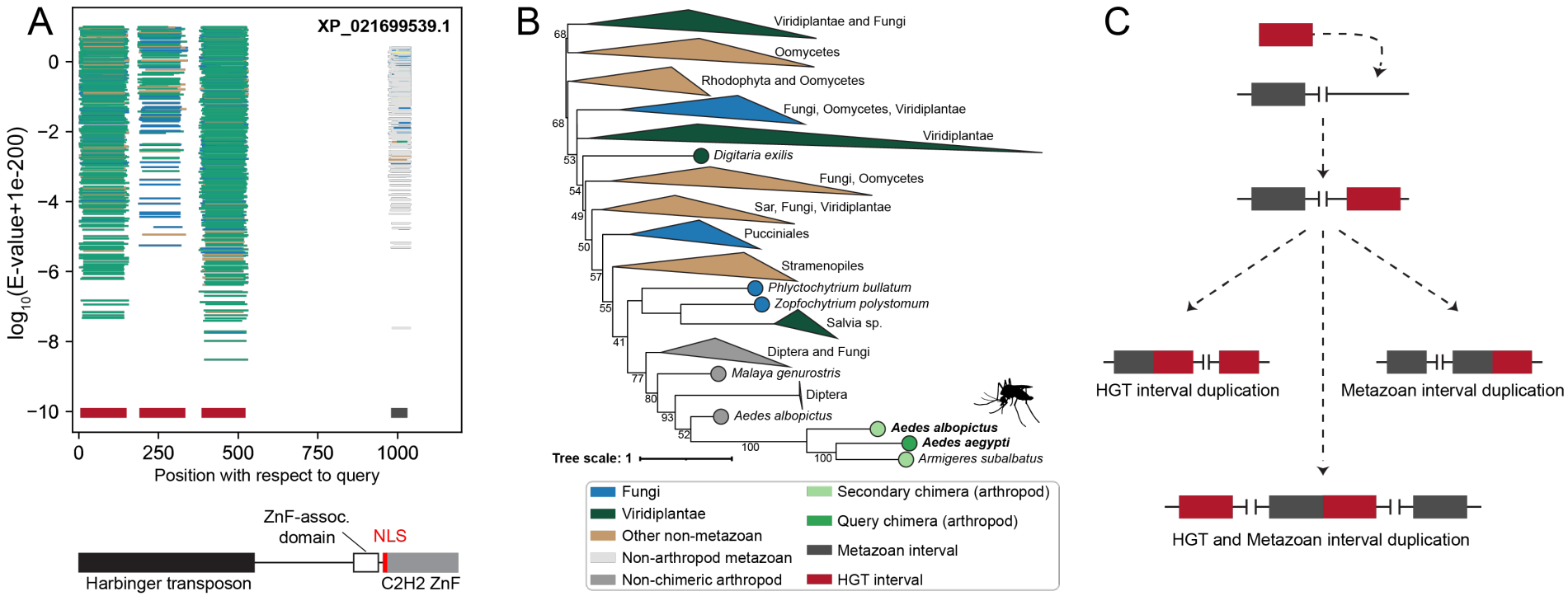
A
XP_021699539.1
log10(E-value+1e-200)
0
−2
−4
−6
−8
−10
0
250
500
750
1000
Position with respect to query
ZnF-assoc. domain
NLS
Harbinger transposon
C2H2 ZnF
B
Viridiplantae and Fungi
Oomycetes
Rhodophyta and Oomycetes
Fungi, Oomycetes, Viridiplantae
Viridiplantae
Digitaria exilis
Fungi, Oomycetes
Sar, Fungi, Viridiplantae
Pucciniales
Stramenopiles
Phlyctochytrium bullatum
Zopfochytrium polystomum
Salvia sp.
Diptera and Fungi
Malaya genurostris
Diptera
Aedes albopictus
Aedes albopictus
Aedes aegypti
Armigeres subalbatus
Tree scale: 1
Fungi
Viridiplantae
Other non-metazoan
Non-arthropod metazoan
Non-chimeric arthropod
Secondary chimera (arthropod)
Query chimera (arthropod)
Metazoan interval
HGT interval
C
HGT interval duplication
Metazoan interval duplication
HGT and Metazoan interval duplication

Figure 3. An HGT-chimera from mosquitoes illustrates that non-chimeric HGT can precede HGT-chimera formation. (A) BLASTp plot excluding arthropod hits for the query chimera XP_021699539.1 from the mosquito *Aedes aegypti* (cluster 9). Metazoan ancestry is inferred for the interval (980-1031) and HGT ancestry for the intervals (5-150), (190-334), (385-523). Domain annotations obtained via NCBI CDD search and InterProScan (SI Table 17) are shown across the bottom of the plot, along with a nuclear localization sequence from residues 975-985 as predicted by cNLS Mapper (96). See SI Figure 13 for detailed annotation. **(B)** Full maximum likelihood tree for the HGT interval (190,394), including secondary chimeras from the mosquitoes *Armigeres subalbatus* (XP_062544500.1) and *Aedes albopictus* (XP_029735553.1). Bolded text indicates that the chimera was successfully verified via RT-PCR. Collapsed clades are represented as in Figure 2. **(C)** Schematic representation of a model of HGT-chimera formation in which HGT precedes a duplication of HGT and/or metazoan genes. This model posits the existence of non-chimeric relatives to the HGT or metazoan intervals within the same genome or in the genomes of closely related arthropods. Note that the expected outcome of fusion via tandem duplication (bottom example) is a reversal of the order of metazoan and HGT segments on either side of the resulting chimera (97). See SI Figure 9 and SI Text 4 for supporting evidence.

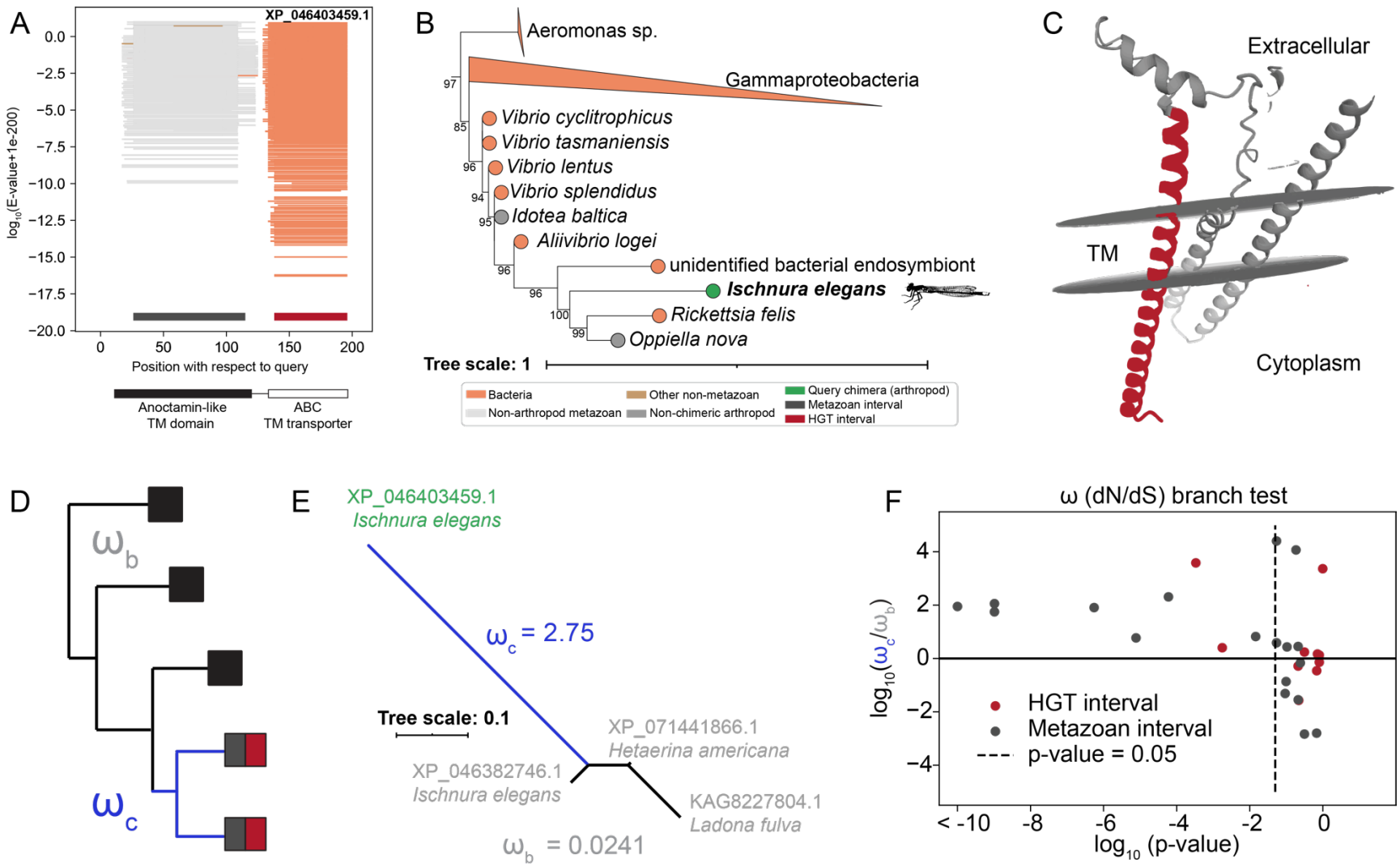
A
XP_046403459.1
log10(E-value+1e-200)
Position with respect to query
Anoctamin-like TM domain
ABC TM transporter
B
Aeromonas sp.
Gammaproteobacteria
Vibrio cyclitrophicus
Vibrio tasmaniensis
Vibrio lentus
Vibrio splendidus
Idotea baltica
Aliivibrio logei
unidentified bacterial endosymbiont
Ischnura elegans
Rickettsia felis
Oppiella nova
Tree scale: 1
Bacteria
Non-arthropod metazoan
Other non-metazoan
Non-chimeric arthropod
Query chimera (arthropod)
Metazoan interval
HGT interval
C
Extracellular
TM
Cytoplasm
D
ωb
ωc
E
XP_046403459.1
Ischnura elegans
ωc = 2.75
Tree scale: 0.1
XP_071441866.1
Hetaerina americana
XP_046382746.1
Ischnura elegans
KAG8227804.1
Ladona fulva
ωb = 0.0241
F
ω (dN/dS) branch test
log10(ωc/ωb)
HGT interval
Metazoan interval
p-value = 0.05
log10 (p-value)

Figure 4. A transmembrane HGT-chimera from a damselfly illustrates post-fusion neofunctionalization. (A) BLASTp plot excluding arthropod hits for the query chimera XP_046403459.1 from the damselfly *Ischnura elegans* (cluster 41). Metazoan ancestry is inferred for the interval (26-115) and HGT ancestry for the interval (138-196). Domain annotations obtained via NCBI CDD search and InterProScan are shown across the top of the plot. **(B)** Maximum likelihood tree for the HGT interval (138-196), pruned to show sequences closely related to the query chimera (full tree available on the iTOL webserver at https://itol.embl.de/shared/rkapoor). Internal node values indicate ultrafast bootstrap support. Per GenBank, the "unidentified bacterial endosymbiont" was isolated from the midge *Chironomus riparius*, and *Rickettsia felis* from the louse *Liposcelis bostrychophila*. **(C)** Transmembrane protein structure and topology for XP_046403459.1 predicted with MembraneFold version 0.0.71 (77) with OmegaFold (98) and DeepTMHMM (99). The HGT-derived interval is colored in dark red. **(D)** Schematic representation of branch test performed to compare ω values of an interval before (gray) and after (blue) an HGT-chimera-forming gene fusion event. Separate gene trees were built for HGT and metazoan intervals using phylogenetic relatives within the same taxonomic order as HGT-chimeras. To best capture the early dynamics of HGT-chimera evolution, we restricted our analysis to 28 intervals from young (species or genus-restricted) HGT-chimeras in which relatives were detected both within the same genome as the HGT-chimera and in the genomes of other arthropods in the same taxonomic order (SI Methods). **(E)** Maximum likelihood tree constructed from the metazoan interval of the damselfly chimera XP_046403459.1 and its non-chimeric phylogenetic relatives in the same taxonomic order (Odonata), including a non-chimeric relative gene in the same genome as the chimera (XP_046382746.1). ω (dN/dS) values from the branch test explained in (D) are superimposed for chimeric and non-chimeric branches. **(F)** The distribution of the fold-change in ω between chimeric and non-chimeric branches of gene trees is plotted on the y-axis, for 28 HGT-chimera intervals. Benjamini-Hochberg corrected p-values (x-axis) were obtained using a likelihood ratio test of a two-branch model to a model with a single ω value.

References

1. H. Kaessmann, Origins, evolution, and phenotypic impact of new genes. *Genome Res.* **20**, 1313–1326 (2010).

2. S. Chen, B. H. Krinsky, M. Long, New genes as drivers of phenotypic evolution. *Nat. Rev. Genet.* **14**, 645–660 (2013).

3. E. V. Koonin, *The logic of chance: The nature and origin of biological evolution* (Financial Times Prentice Hall, 2011).

4. R. L. Rogers, D. L. Hartl, Chimeric genes as a source of rapid evolution in Drosophila melanogaster. *Mol. Biol. Evol.* **29**, 517–529 (2012).

5. J. Zhang, A. M. Dean, F. Brunet, M. Long, Evolving protein functional diversity in new genes of Drosophila. *Proc. Natl. Acad. Sci.* **101**, 16246–16250 (2004).

6. A. D. Farr, P. Remigi, P. B. Rainey, Adaptive evolution by spontaneous domain fusion and protein relocalization. *Nat. Ecol. Evol.* **1**, 1562–1568 (2017).

7. C. D. Jones, D. J. Begun, Parallel evolution of chimeric fusion genes. *Proc. Natl. Acad. Sci.* **102**, 11373–11378 (2005).

8. R. L. Rogers, T. Bedford, A. M. Lyons, D. L. Hartl, Adaptive impact of the chimeric gene Quetzalcoatl in Drosophila melanogaster. *Proc. Natl. Acad. Sci.* **107**, 10943–10948 (2010).

9. A.-R. Carvunis, *et al.*, Proto-genes and de novo gene birth. *Nature* **487**, 370–374 (2012).

10. Q. Jianhuan, *et al.*, A human-specific De Novo gene promotes cortical expansion and folding. *Adv. Sci.* **10**, e2204140 (2023).

11. W. F. Martin, Too much eukaryote LGT. *BioEssays* **39**, 1700115 (2017).

12. C. S. Smillie, *et al.*, Ecology drives a global network of gene exchange connecting the human microbiome. *Nature* **480**, 241–244 (2011).

13. S. Chou, *et al.*, Transferred interbacterial antagonism genes augment eukaryotic innate immune function. *Nature* **518**, 98–101 (2015).

14. Y. Li, *et al.*, HGT is widespread in insects and contributes to male courtship in lepidopterans. *Cell* **185**, 2975-2987.e10 (2022).

15. F. Husnik, J. P. McCutcheon, Functional horizontal gene transfer from bacteria to eukaryotes. *Nat. Rev. Microbiol.* **16**, 67–79 (2018).

16. C. Ku, *et al.*, Endosymbiotic origin and differential loss of eukaryotic genes. *Nature* **524**, 427–432 (2015).

17. O. Bánki, Y. Roskov, G. Ower, D. R. H. Robles, Catalogue of Life (Version 2025-04-10). [Preprint] (2025). Available at: https://doi.org/10.48580/dgplc.

18. Z.-W. Li, Y.-H. Shen, Z.-H. Xiang, Z. Zhang, Pathogen-origin horizontally transferred genes contribute to the evolution of Lepidopteran insects. *BMC Evol. Biol.* **11**, 356 (2011).

19. N. Wybouw, *et al.*, A gene horizontally transferred from bacteria protects arthropods from host plant cyanide poisoning. *eLife* **3**, e02365 (2014).

20. W. Suring, J. Mariën, R. Broekman, N. M. van Straalen, D. Roelofs, Biochemical pathways supporting beta-lactam biosynthesis in the springtail Folsomia candida. *Biol. Open* **5**, 1784–1789 (2016).

21. J. A. Metcalf, L. J. Funkhouser-Jones, K. Brileya, A.-L. Reysenbach, S. R. Bordenstein, Antibacterial gene transfer across the tree of life. *eLife* **3**, e04266 (2014).

22. J. C. D. Hotopp, *et al.*, Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes. *Science* **317**, 1753–1756 (2007).

23. K. I. Verster, *et al.*, Evolution of insect innate immunity through domestication of bacterial toxins. *Proc. Natl. Acad. Sci.* **120**, e2218334120 (2023).

24. N. A. T. Irwin, A. A. Pittis, T. A. Richards, P. J. Keeling, Systematic evaluation of horizontal gene transfer between eukaryotes and viruses. *Nat. Microbiol.* **7**, 327–336 (2022).

25. S. E. Schneider, J. H. Thomas, Accidental genetic engineers: horizontal sequence transfer from parasitoid wasps to their Lepidopteran hosts. *PLoS ONE* **9**, e109446 (2014).

26. L. Gasmi, *et al.*, Horizontally transmitted parasitoid killing factor shapes insect defense to parasitoids. *Science* **373**, 535–541 (2021).

27. A. Faddeeva-Vakhrusheva, *et al.*, Coping with living in the soil: the genome of the parthenogenetic springtail Folsomia candida. *BMC Genom.* **18**, 493 (2017).

28. B. Altincicek, J. L. Kovacs, N. M. Gerardo, Horizontally transferred fungal carotenoid genes in the two-spotted spider mite Tetranychus urticae. *Biol. Lett.* **8**, 253–257 (2012).

29. J. Xia, *et al.*, Whitefly hijacks a plant detoxification gene that neutralizes plant toxins. *Cell* **184**, 1693-1705.e17 (2021).

30. X. Chen, *et al.*, Canonical terpene synthases in arthropods: Intraphylum gene transfer. *Proc. Natl. Acad. Sci.* **121**, e2413007121 (2024).

31. C. Deng, C. H. C. Cheng, H. Ye, X. He, L. Chen, Evolution of an antifreeze protein by neofunctionalization under escape from adaptive conflict. *Proc. Natl. Acad. Sci.* **107**, 21593–21598 (2010).

32. S. Chen, *et al.*, Reshaping of global gene expression networks and sex-biased gene expression by integration of a young gene. *EMBO J.* **31**, 2798–2809 (2012).

33. N. W. VanKuren, M. Long, Gene duplicates resolving sexual conflict rapidly evolved essential gametogenesis functions. *Nat. Ecol. Evol.* **2**, 705–712 (2018).

34. V. J. Lynch, Inventing an arsenal: adaptive evolution and neofunctionalization of snake venom phospholipase A2 genes. *BMC Evol. Biol.* **7**, 2 (2007).

35. M. Talianova, B. Vyskot, B. Janousek, Interkingdom protein domain fusion: the case of an antimicrobial protein in potato (Solanum tuberosum). *Plant Syst. Evol.* **297**, 129 (2011).

36. Q. Ren, *et al.*, Co-option of bacteriophage lysozyme genes by bivalve genomes. *Open Biol.* **7**, 160285 (2017).

37. C. Gilbert, C. Feschotte, Horizontal acquisition of transposable elements and viral sequences: patterns and consequences. *Curr. Opin. Genet. Dev.* **49**, 15–24 (2018).

38. E. Little, P. Bork, R. F. Doolittle, Tracing the spread of fibronectin type III domains in bacterial glycohydrolases. *J. Mol. Evol.* **39**, 631–643 (1994).

39. A. Ephrussi, R. Lehmann, Induction of germ cell formation by oskar. *Nature* **358**, 387–392 (1992).

40. J. A. Lynch, *et al.*, The Phylogenetic Origin of oskar Coincided with the Origin of Maternally Provisioned Germ Plasm and Pole Cells at the Base of the Holometabola. *PLoS Genet.* **7**, e1002029 (2011).

41. B. Ewen-Campen, J. R. Srouji, E. E. Schwager, C. G. Extavour, oskar Predates the Evolution of Germ Plasm in Insects. *Curr Biol* **22**, 2278–2283 (2012).

42. A. Kulkarni, *et al.*, oskar acts with the transcription factor Creb to regulate long-term memory in crickets. *Proc. Natl. Acad. Sci.* **120**, e2218506120 (2023).

43. I. Callebaut, J.-P. Mornon, LOTUS, a new domain associated with small RNA pathways in the germline. *Bioinformatics* **26**, 1140–1144 (2010).

44. M. Jeske, *et al.*, The Crystal Structure of the Drosophila Germline Inducer Oskar Identifies Two Domains with Distinct Vasa Helicase- and RNA-Binding Activities. *Cell Rep.* **12**, 587–598 (2015).

45. C. C. Akoh, G.-C. Lee, Y.-C. Liaw, T.-H. Huang, J.-F. Shaw, GDSL family of serine esterases/lipases. *Prog. Lipid Res.* **43**, 534–552 (2004).

46. L. Blondel, T. E. Jones, C. G. Extavour, Bacterial contribution to genesis of the novel germ line determinant oskar. *eLife* **9**, e45539 (2020).

47. Y. I. Wolf, A. S. Kondrashov, E. V. Koonin, Interkingdom gene fusions. *Genome Biol.* **1**, research0013.1-0013.13 (2000).

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

49. B. Buchfink, K. Reuter, H.-G. Drost, Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat. Methods* **18**, 366–368 (2021).

50. E. W. Sayers, *et al.*, Database resources of the National Center for Biotechnology Information in 2025. *Nucleic Acids Res.* **53**, D20–D29 (2024).

51. C. Menichelli, O. Gascuel, L. Bréhélin, Improving pairwise comparison of protein sequences with domain co-occurrence. *PLoS Comput. Biol.* **14**, e1005889 (2018).

52. S. Eddy, HMMER User's Guide: Biological sequence analysis using profile hidden Markov models. (2023). Available at: http://hmmer.org.

53. T. K. F. Wong, *et al.*, IQ-TREE 3: Phylogenomic Inference Software using Complex Evolutionary Models. *EcoEvoRxiv* **10.32942/x2p62n** (2025).

54. T. Domazet-Lošo, J. Brajković, D. Tautz, A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. *Trends Genet.* **23**, 533–539 (2007).

55. A. R. Omilian, M. Lynch, Patterns of intraspecific DNA variation in the Daphnia nuclear genome. *Genetics* **182**, 325–336 (2009).

56. J. Peccoud, V. Loiseau, R. Cordaux, C. Gilbert, Massive horizontal transfer of transposable elements in insects. *Proc. Natl. Acad. Sci.* **114**, 4721–4726 (2017).

57. L. A. Graham, P. L. Davies, Horizontal Gene Transfer in Vertebrates: A Fishy Tale. *Trends Genet.* **37**, 501–503 (2021).

58. L. A. Graham, S. C. Lougheed, K. V. Ewart, P. L. Davies, Lateral transfer of a lectin-like antifreeze protein gene in fishes. *PLoS ONE* **3**, e2616 (2008).

59. T. Mishina, *et al.*, Massive horizontal gene transfer and the evolution of nematomorph-driven behavioral manipulation of mantids. *Curr. Biol.* **33**, 4988-4994.e5 (2023).

60. C. Rödelsperger, R. J. Sommer, Computational archaeology of the Pristionchus pacificus genome reveals evidence of horizontal gene transfers from insects. *BMC Evol. Biol.* **11**, 239 (2011).

61. R. Feyereisen, J. M. Urban, D. R. Nelson, Aliens in the CYPome of the black fungus gnat, Bradysia coprophila. *Insect Biochem. Mol. Biol.* **159**, 103965 (2023).

62. H. Cao, S. Chen, L. Lu, J. An, Shewanella algae: an Emerging Pathogen of Black Spot Disease in Freshwater-Cultured Whiteleg Shrimp (Penaeus vannamei). *Isr. J. Aquac. - Bamidgeh* **70** (2018).

63. M. T. Ødum, *et al.*, DeepLoc 2.1: multi-label membrane protein type prediction using protein language models. *Nucleic Acids Res.* **52**, W215–W220 (2024).

64. J. J. A. Armenteros, *et al.*, SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nat. Biotechnol.* **37**, 420–423 (2019).

65. S. Kumar, *et al.*, TimeTree 5: An expanded resource for species divergence times. *Mol. Biol. Evol.* **39**, msac174 (2022).

66. A. A. Almada, A. M. Tarrant, Vibrio elicits targeted transcriptional responses from copepod hosts. *FEMS Microbiol. Ecol.* **92**, fiw072 (2016).

67. K. M. Shoemaker, P. H. Moisander, Microbial diversity associated with copepods in the North Atlantic subtropical gyre. *FEMS Microbiol. Ecol.* **91**, fiv064 (2015).

68. X. Li, L.-X. Wang, X. Wang, S. Roseman, The Chitin Catabolic Cascade in the Marine Bacterium Vibrio Cholerae: Characterization of a Unique Chitin Oligosaccharide Deacetylase. *Glycobiology* **17**, 1377–1387 (2007).

69. I. Yanai, A. Derti, C. DeLisi, Genes linked by fusion events are generally of the same functional category: a systematic analysis of 30 microbial genomes. *Proc. Natl. Acad. Sci.* **98**, 7940–7945 (2001).

70. A. Huda, I. K. Jordan, Analysis of transposable element sequences using CENSOR and RepeatMasker. *Methods Mol. Biol.* **537**, 323–336 (2009).

71. Q. Wang, Y. Wang, J. Wang, Z. Gong, G. Han, Plants acquired a major retrotransposon horizontally from fungi during the conquest of land. *N. Phytol.* **232**, 11–16 (2021).

72. X. Lin, N. Faridi, C. Casola, An Ancient Transkingdom Horizontal Transfer of Penelope -Like Retroelements from Arthropods to Conifers. *Genome Biol. Evol.* **8**, 1252–1266 (2016).

73. L. Sinzelle, *et al.*, Transposition of a reconstructed Harbinger element in human cells and functional homology with two transposon-derived cellular genes. *Proc. Natl. Acad. Sci.* **105**, 4715–4720 (2008).

74. R. Mackeh, A. K. Marr, A. Fadda, T. Kino, C2H2-Type Zinc Finger Proteins: Evolutionarily Old and New Partners of the Nuclear Hormone Receptors. *Nucl. Recept. Signal.* **15**, 1550762918801071 (2018).

75. R. L. Cosby, *et al.*, Recurrent evolution of vertebrate transcription factors by transposase capture. *Science* **371**, eabc6405 (2021).

76. M. Tellier, R. Chalmers, Human SETMAR is a DNA sequence-specific histone-methylase with a broad effect on the transcriptome. *Nucleic Acids Res.* **47**, 122–133 (2019).

77. S. Gutierrez, W. G. Tyczynski, W. Boomsma, F. Teufel, O. Winther, MembraneFold: Visualising transmembrane protein structure and topology. *bioRxiv* 2022.12.06.518085 (2022)..

78. S. Álvarez-Carretero, P. Kapli, Z. Yang, Beginner's guide on the use of PAML to detect positive selection. *Mol. Biol. Evol.* **40**, msad041 (2023).

79. Z. Yang, PAML: a program package for phylogenetic analysis by maximum likelihood. *Bioinformatics* **13**, 555–556 (1997).

80. Z. Yang, PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).

81. Y. Zhou, C. Zhang, Evolutionary patterns of chimeric retrogenes in Oryza species. *Sci. Rep.* **9**, 17733 (2019).

82. N. Yang, *et al.*, Structure of Drosophila Oskar reveals a novel RNA binding protein. *Proc. Natl. Acad. Sci.* **112**, 11541–11546 (2015).

83. Z. Yang, *et al.*, Horizontal gene transfer is more frequent with increased heterotrophy and contributes to parasite adaptation. *Proc. Natl. Acad. Sci.* **113**, E7010–E7019 (2016).

84. L. T. Dunning, *et al.*, Lateral transfers of large DNA fragments spread functional genes among grasses. *Proc. Natl. Acad. Sci.* **116**, 4416–4425 (2019).

85. E. Gluck-Thaler, *et al.*, Giant Starship Elements Mobilize Accessory Genes in Fungal Genomes. *Mol. Biol. Evol.* **39**, msac109 (2022).

86. M. Habig, *et al.*, Frequent horizontal chromosome transfer between asexual fungal insect pathogens. *Proc. Natl. Acad. Sci.* **121**, e2316284121 (2024).

87. N. Nikoh, *et al.*, Bacterial genes in the aphid genome: absence of functional gene transfer from Buchnera to its host. *PLoS Genet.* **6**, e1000827 (2010).

88. P. J. Keeling, Horizontal gene transfer in eukaryotes: aligning theory with data. *Nat. Rev. Genet.* **25**, 416–430 (2024).

89. O. Cohen, U. Gophna, T. Pupko, The complexity hypothesis revisited: connectivity rather than function constitutes a barrier to horizontal gene transfer. *Mol. Biol. Evol.* **28**, 1481–1489 (2011).

90. C. Park, J. Zhang, High expression hampers horizontal gene transfer. *Genome Biol. Evol.* **4**, 523–532 (2012).

91. H. Dai, *et al.*, The evolution of courtship behaviors through the origination of a new gene in Drosophila. *Proc. Natl. Acad. Sci.* **105**, 7478–7483 (2008).

92. S. Chen, Y. E. Zhang, M. Long, New Genes in Drosophila Quickly Become Essential. *Science* **330**, 1682–1685 (2010).

93. Y. C. G. Lee, *et al.*, Rapid Evolution of Gained Essential Developmental Functions of a Young Gene via Interactions with Other Essential Genes. *Mol. Biol. Evol.* **36**, 2212–2226 (2019).

94. F. Jacob, Evolution and tinkering. *Science* **196**, 1161–1166 (1977).

95. I. Letunic, P. Bork, Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* **49**, W293–W296 (2021).

96. S. Kosugi, M. Hasebe, M. Tomita, H. Yanagawa, Systematic identification of cell cycle-dependent yeast nucleocytoplasmic shuttling proteins by prediction of composite motifs. *Proc. Natl. Acad. Sci.* **106**, 10171–10176 (2009).

97. R. L. Rogers, T. Bedford, D. L. Hartl, Formation and longevity of chimeric and duplicate genes in Drosophila melanogaster. *Genetics* **181**, 313–322 (2009).

98. R. Wu, *et al.*, High-resolution de novo structure prediction from primary sequence. *bioRxiv* 2022.07.21.500999 (2022).

99. J. Hallgren, *et al.*, DeepTMHMM predicts alpha and beta transmembrane proteins using deep neural networks. *bioRxiv* 2022.04.08.487609 (2022).