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

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 interaction of HGT with other mechanisms of novel gene formation 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 19 arthropod species predicted to harbor HGT-chimeras in their genome. We found evidence of the expression of contiguous chimeric mRNAs for 33 of 37 tested HGT-chimeras across 17 of 19 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

New genes have evolved on recent timescales across the tree of life (1, 2).. Some have rapidly reached fixation under the influence of strong positive selection (4, 8) and have integrated into pre-existing molecular interaction networks (10, 11). Consistent with their taxonomically restricted distribution, novel genes can contribute both to functional innovations underlying niche expansions and to tolerance of new environmental stressors (1, 2). Despite their lack of conservation, novel genes can also quickly (3) become essential for fundamental and conserved biological processes including development, courtship (12), and reproduction (1, 2, 12, 13).

The best-understood mechanism of gene birth is gene duplication, which may allow for the evolution of new function (neofunctionalization) or partitioning of ancestral function (subfunctionalization) (1, 2, 4). Additional mechanisms include exon shuffling and gene fusion, which may generate chimeric genes with portions from two different gene parents within the same genome, or fission, which splits a gene in two (5–9). Genes can also arise *de novo* from pre-existing noncoding regions (10, 11). All of these mechanisms may allow evolution to make leaps in protein structure/function space, and therefore can be functionally significant (5).

Horizontal gene transfer (HGT) is a mechanism by which organisms acquire new genes in a non-Mendelian fashion from other species (4, 12–16). HGT is widely appreciated as a fundamental force of evolution in prokaryotes, where mechanisms such as conjugation facilitate the exchange of genetic material (13). 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 (4, 12, 17). Arthropods, the phylum of animals that constitutes at least 80% of all described animal species (18), are emerging as an exception to this expectation. Horizontally acquired genes from bacteria (14–16, 19–24), viruses (15, 25–27), fungi (15, 28, 29) and even plants (15, 30) have been identified in diverse arthropods. In many cases, these genes contribute adaptive functions ranging from ecological defense (24, 30) to metabolism (29, 31). A recent

64 systematic screen for HGT across 218 insect genomes uncovered 741 independent HGT events,
65 placing the rate of HGT on par with that of other rare novel gene formation events within this
66 phylum (15).

67 Most bioinformatic work on HGT to date has focused on detection of transferred genes and
68 their sources, rather than their evolutionary fates or biological roles following acquisition. Whether
69 horizontally acquired genes retain similar gene structures and functions following transfer, or, like
70 other kinds of novel genes (8, 32–35), rapidly diverge from their original forms, has received less
71 attention than detection of the transfer phenomenon. Here we focus on gene fusion as a possible
72 mechanism of post-transfer divergence. We hypothesized that, as a result of the fusion of
73 horizontally acquired sequences with pre-existing host genome sequences, animal genomes would
74 contain “HGT-chimeras”: genes with at least one region horizontally transferred from a non-
75 metazoan source in the same open reading frame as a region of ancient metazoan ancestry. To our
76 knowledge, no large-scale systematic screen for HGT-chimeras has been conducted in multicellular
77 eukaryotes, but a few examples have been identified idiosyncratically across diverse taxa (22, 36–
78 39). For instance, we previously attributed the origin of *oskar*, a gene required for essential
79 developmental processes in diverse insect species (40–43), to the formation of a HGT-chimera
80 between a metazoan LOTUS domain (44) and a bacterial GDSL hydrolase domain (45, 46) in the
81 last common ancestor of extant insects (47). In another example, a screen of 20 microbial genomes
82 identified 37 putative interkingdom HGT-chimeras in 11 unicellular species, but did not search for
83 HGT-chimeras in multicellular eukaryotes (48). Here, we present the results of a systematic
84 bioinformatic screen for HGT-chimeras across 319 arthropod genomes. We find evidence that
85 HGT-chimeric genes are not only widespread across the tree of arthropods, but also display
86 transcriptional and molecular evolutionary characteristics that suggest they play active roles in
87 organismal biology and evolution.

Results and Discussion

Development of an HGT-chimera detection pipeline

We designed a computational pipeline to search for HGT-chimeras in arthropod genomes (SI Figure 1), which we describe in brief here and detail in SI Methods. The pipeline proceeds in four broad phases (SI Figure 1). In the first phase, we assembled a search set of 7,702,369 protein inputs from 319 RefSeq scaffold or chromosome-level genomes (SI Table 1) across eight arthropod classes (SI Figure 4). We clustered the initial search set to 610,348 proteins using MMSeqs2 to promote computational efficiency (49). In the second phase, we used the results of BLASTp (implemented in DIAMOND (50)) against the NR database (51) to perform an initial screen for putative chimeric proteins. Non-arthropod BLASTp results were first used to partition query proteins into intervals of potentially distinct evolutionary history using a customized implementation of a previously described algorithm (52) (Figure 1A, SI File 1, SI Figure 2). We hypothesized that intervals with a preponderance of hits in non-metazoan taxa and/or lower E-values for non-metazoan than for metazoan hits were of putative HGT ancestry (right-hand interval of Figure 1A), and that those whose lowest E-value hits were in other metazoan taxa or had a preponderance of metazoan hits were of ancient metazoan ancestry (left-hand interval of Figure 1A). We then subjected proteins containing at least one interval of putative HGT ancestry and at least one interval of metazoan ancestry to a second round of BLASTp, using each of the two types of intervals as separate queries to further examine potential HGT and metazoan ancestry annotations from the first round of BLASTp (SI Figure 3A-B). In the third phase, we removed potential confounders of HGT inference (ankyrin repeats and metazoan transposable elements, see SI Methods) and binned HGT-chimeras into putative orthologous clusters using a custom approach that required all sequences in the same cluster to have the same chimeric architecture as determined by BLASTp and hmmsearch (HMMER suite (53). Except where otherwise indicated, downstream analyses (Figures 2-4) were performed with a single representative sequence per orthologous cluster. In the fourth and final phase of

114 the pipeline, we assessed the support for HGT and metazoan ancestries of the distinct intervals
115 of each cluster representative using maximum likelihood phylogenetic trees (IQ-TREE (54))
116 (Figure 1B, SI Figure 9). To improve the sensitivity of homology detection, we constructed these
117 trees from the results of hmmsearch using profile hidden Markov models (HMMs) built from
118 arthropod BLASTp hits as queries, or from BLASTp hits directly when we could not construct
119 HMMs (E-value for both <1e-2, see SI Methods). We considered proteins with at least one
120 interval of metazoan origin, at least one interval of non-metazoan origin, and support for tree
121 topologies with >70% support at the relevant node (bootstrap only considered for HGT intervals,
122 see SI Figure 9), to be candidate HGT-chimeras: genes formed by horizontal sequence transfer
123 followed by in-frame fusion.
124
125 *HGT-chimeras are detected across Arthropoda*
126 Excluding possible within-genome paralogs (see SI Table 6 for copy numbers), we found
127 a final set of 274 HGT-chimera genes across these query genomes, corresponding to 104
128 independent origination events inferred by a customized orthologous clustering approach (SI
129 Methods). We detected HGT-chimeras across the arthropod phylogeny (Figure 1D), including
130 members of Chelicerata and Pancrustacea. A majority (193/319, 60.5%) of examined RefSeq
131 genomes had at least one HGT-chimera (Figure 1C, SI Table 3). We note that *Folsomia*
132 *candida*, the species with the greatest predicted number of HGT-chimeras, was previously
133 reported to harbor large numbers of non-chimeric horizontally acquired genes(21, 28, 31).
134 To identify the broadest range of arthropod species containing each HGT-chimera (SI
135 Table 4), we additionally searched for representatives of HGT-chimera ortholog groups in an
136 expanded set of arthropod genomes, including an additional 197 genome annotations from
137 GenBank (SI Table 1). We found that 81/104 orthologous groups were restricted to a single
138 species (Figure 1E, SI Figure 5A), suggesting a relatively recent evolutionary origin. For
139 orthologous groups found in more than one species, we inferred minimum evolutionary gene

140 ages by phylostratigraphy (55) (SI Figure 5 and SI Table 5). We detected a member of the
141 orthologous cluster corresponding to the gene *oskar* (cluster 1) in 181 insects spanning the
142 subclass Pterygota, consistent with our prior finding that *oskar* originated via chimeric HGT in or
143 prior to the last common ancestor of winged insects (>400 million years ago [MYA] (47)).
144 In contrast to the insect-specific *oskar* cluster, we found two clusters (cluster 3 and
145 cluster 18) whose members were found in multiple species from distantly related classes of
146 arthropods (Figure 1D). Cluster 18 was found exclusively in the firefly *Photinus pyralis* (order:
147 Coleoptera, class: Insecta) and in the mite *Oppia nitens* (order: Sarcoptiformes, class:
148 Arachnida), a finding that we interpret as evidence for convergent evolution of analogous HGT-
149 chimera architectures via independent HGT and gene fusion events (SI Figure 7). In the case of
150 cluster 3, molecular phylogenetic analyses suggested that its sparse phylogenetic distribution
151 could have arisen from the transfer of these HGT-chimeras between springtails, fungus gnats,
152 and mites following the formation of this gene (SI Text 1, SI Figure 7). We note that inter-
153 metazoan HGT, although thought to be rare (see for example 56–58), has nevertheless been
154 reported by other studies (31, 58–60), including among these specific groups of arthropods (31,
155 61). Thus, HGT appears not only to have contributed to the origination of this HGT-chimera, but
156 also to have contributed to its spread across species following its formation.
157 The youngest HGT-chimeras (clusters 20 and 21) that we dated via phylostratigraphy
158 (55) appeared in the congeneric microcrustaceans *Daphnia pulex* and *D. pulicaria*, but were
159 absent in the three other *Daphnia* genomes available at the time of writing. This suggests an
160 origin for this HGT-chimera as recently as 82,000 years ago in the *D. pulex*-*D. pulicaria*
161 common ancestor (62). Some genomes featured both young and old HGT-chimeras. For
162 example, the shrimp *Penaeus vannamei* contained one HGT-chimera (cluster 12) exclusively
163 found in three other congeneric species (divergence time ~83.1 MYA (63)), and another also
164 found in lobsters (cluster 2), tracing its origin to the last common ancestor of extant decapods

165 ~394 MYA (63) (SI Figure 5B). Collectively, these findings suggest that HGT-chimera formation
166 has been an ongoing process throughout arthropod history.
167
168 *HGT-chimeras are derived from diverse donor lineages*
169 Next, we investigated the evolutionary histories of HGT-chimeras in greater detail. We
170 used HGT interval trees (Figure 1B) to infer donor taxa and found that the majority of HGT
171 intervals plausibly originated in bacterial donors (Figure 2A, SI Table 7). Among 62 sequences
172 of inferred bacterial ancestry, close phylogenetic relatives of 37 of them are from bacterial
173 genera of known arthropod symbionts (Figure 2B, SI Table 8), including the endosymbionts
174 *Wolbachia*, *Rickettsia* and *Hamiltonella*. This finding is consistent with prior reports of HGT
175 between pathogenic and endosymbiotic bacteria and arthropod genomes (15, 19, 23, 47).
176 In addition to symbiotic bacterial donors, we observed 23 putative transfers from donors
177 with plausible ecological associations with the recipient arthropods, including 13 cases of HGT
178 from arthropod-infecting viruses, one from fungi to the fungus-consuming gnat *Bradysia*
179 *coprophila*, and three from soil-dwelling bacteria to the soil-dwelling mite *Oppia nitens* (SI Table
180 8). In all, these data are consistent with the hypothesis that close ecological associations with
181 food, pathogens and symbionts may provide opportunities for HGT (13, 19, 31, 38, 59).
182
183 *HGT-chimeras plausibly formed via duplication-mediated gene fusion*
184 We hypothesized that, similarly to other gene fusion events, HGT-chimera formation
185 might be driven by duplication events, in which a duplicated copy of a HGT and/or metazoan
186 sequence undergoes fusion with a sequence of the opposite ancestry (Figure 2C, SI Table 9)
187 (7, 48, 64, 65). This duplication-based hypothesis predicts that HGT-chimeras should have non-
188 chimeric within-genomic homologs ("relatives") to their HGT and/or metazoan intervals. We
189 searched for evidence of relatives via sequence similarity (BLASTp) and phylogenetics-based
190 methods, both of which yielded results consistent with this prediction. First, using BLASTp we

191 found that 82/104 HGT-chimeras had within-genome hits (E-value<1e-10) for at least one
192 constituent interval, and that 46/104 HGT-chimeras had hits for both HGT and metazoan
193 intervals (Figure 2D). For most intervals, phylogenetic analyses supported the presence of close
194 non-chimeric relatives, either within the same genome as the chimera or in the genomes of
195 closely related species lacking the chimera (SI Text 2, SI Figure 8, SI Table 9-10). BLASTp-
196 inferred HGT-chimeras additionally displayed genomic signatures consistent with gene fusion
197 via tandem duplication (N=26 intervals) or retroduplication (N=12 intervals), two established
198 mechanisms of gene birth via fusion (Figure 2E) (5, 6, 9, 64, 66). These data are consistent with
199 a model in which HGT-chimeras can form via horizontal transfer followed by duplication-based
200 gene fusion, as previously posited for a smaller number of HGT-chimeras in bacterial genomes
201 (48).
202
203 *HGT-chimeric genes are transcribed*
204 Multiple lines of evidence suggest that most of the HGT-chimeras that we predict based
205 on genome assembly and annotation data exist in their respective genomes and are expressed
206 as functional gene products. First, we discarded any putative HGT-chimeras from short
207 scaffolds or contigs <100,000 kilobases in length to reduce the potential influence of
208 contaminating DNA reads. Second, we observed that the HGT-ancestry intervals showed no
209 systematic deviation in codon use or GC content from the background of expressed transcripts
210 from the same genome, nor from metazoan intervals in the same protein, a pattern inconsistent
211 with HGT-ancestry intervals being an artifact of genomic contamination (Figure 3A-B, SI Table
212 11). Third, 65/104 HGT-chimeras had RNA-Seq support over the full length of their transcripts in
213 the RefSeq genome browser, while 93/104 HGT-chimeras had RNA-seq support over >50% of
214 the length of all identified HGT and metazoan intervals (SI Figure 3C, SI Table 12). Finally, we
215 used RT-PCR and Sanger sequencing to validate the existence of continuous mRNA transcripts
216 containing both predicted HGT and metazoan intervals for 37 HGT-chimeras in 19 species. We

217 validated expression in 89% (33/37) of examined HGT-chimeras, corresponding to 23 distinct
218 orthologous clusters across 17 species (SI Table 13, SI Figure 3C).
219
220 *HGT-chimeras evolve under gene-wide constraint with evidence of accelerated divergence from*
221 *non-chimeric relatives*
222 The observation that 23 of 104 HGT-chimera clusters are found in multiple, often
223 distantly diverged, species (SI Table 5) is consistent with their conservation as functionally
224 important genes. Further, all 21 of these clusters (excluding clusters 18 and 3 due to possible
225 non-vertical inheritance, see SI Text 1) had gene-wide dN/dS<1, and in 19 cases we could
226 reject a null hypothesis of neutrality (Figure 3C, SI Table 14). Leveraging fixed-site models in
227 which dN/dS was allowed to vary between HGT and non-HGT ancestry regions, we found
228 statistically significant evidence of relaxed selective constraint in codons of HGT ancestry
229 relative to the rest of the chimeric protein in only 14.3% (3/21) of cases (Figure 3D, SI Table 15)
230 (67–70). This paucity of relaxed selective constraint over the length of HGT-chimeras suggests
231 that these genes are functionally expressed as translated peptides, and that HGT-derived
232 regions contribute to the functions of these genes (65, 71, 72).
233 Although HGT-chimera orthologs diverge slowly from each other following their
234 formation, we asked whether they showed signs of rapid divergence from their non-chimeric
235 parents, similar to the functional divergence (neofunctionalization) observed in other novel
236 genes that evolved through gene duplication (5, 8, 9, 32, 35, 66). To answer this question, we
237 applied two-ratio dN/dS models to species- or genus-restricted HGT-chimera intervals aligned
238 with homologous non-chimeric sequences from the same genome ("relatives") and from related
239 arthropod species in the same taxonomic order (Figure 3E-F) (68–70). We found that 64.3%
240 (18/28) of intervals had a higher dN/dS value on the chimeric branches, while 32.1% (9/28) of
241 intervals had both a greater dN/dS value on chimeric branches and a statistically significant
242 better fit to a two-ratio model than to a single-ratio model (Figure 3E, SI Table 16). This

243 suggests that accelerated protein sequence evolution may be a common fate of incipient HGT-
244 chimeras, consistent with their acquisition of novel functions relative to non-chimeric progenitors
245 (6, 8, 33, 73).
246
247 *HGT-chimeras are predicted to function in diverse processes*
248 To obtain predictions of gene function, we integrated the results of BLASTp hits, domain
249 annotations using InterProScan (74), NCBI CD-search (75), and gene ontology annotations
250 generated via DeepGO-SE (76) for the HGT and metazoan intervals of each HGT-chimera
251 (Figure 4, SI Table 17). We found a diverse array of predicted functions, from biosynthesis to
252 carbohydrate metabolism and regulation of gene expression, implicated in both the metazoan
253 and HGT-derived intervals of HGT-chimeras. Interestingly, over one third (38/104) of HGT-
254 chimeras displayed putative functional similarity in their HGT- and metazoan intervals, including
255 eight cases in which both HGT and metazoan intervals were predicted to bind to or chemically
256 modify carbohydrates, and 22 cases in which both types of intervals were predicted to interact
257 with DNA or RNA (Figure 3, SI Table 18). Strikingly, we found two independent fusions between
258 metazoan C-type lectins and bacteriophage tail fiber protein genes (clusters 3 and 69), both of
259 which in other organisms bind to bacterial surface molecules (77, 78). We speculate that these
260 HGT-chimeras might function in immunological pathogen recognition. These observations,
261 coupled with the repeated observation of functionally coherent HGT-chimeras in other studies of
262 gene fusion (5, 48, 79), could mean that HGT-chimeras formed from functionally similar
263 components may be more likely to encode beneficial functions, contributing to their fixation and
264 long-term preservation under purifying selection (5).

265 **Conclusions**

266 Novel protein structure and function can emerge from several elementary events,
267 including amino acid substitutions, gene duplication, gene fusion and HGT (4). In this study, we
268 have shown that multiple of these elementary events can operate together in the origins of a
269 previously understudied class of genes that we call HGT-chimeras. In at least 104 independent
270 events across the history of the arthropods, HGT-chimeras evolved via the fusion of
271 endogenous genes of ancient metazoan ancestry with genes acquired via horizontal transfer
272 from non-metazoan sources. In nearly 80% (82/104) of HGT-chimeras, sequence duplication
273 within the host genome played a plausible role in their formation (Figure 2C-E), and many HGT-
274 chimeras show a signature of accelerated amino acid substitution following their birth (Figure
275 3F). Collectively, these findings reveal a rich post-transfer history of horizontally acquired
276 sequences in arthropods.

277 We propose that although chimeric genes of this nature have been reported rarely (22,
278 36–39, 47, 48) despite significant prior work on HGT (13–17, 19, 20, 23, 25–30), we were able
279 to uncover these HGT-chimeras by better demarcating intervals of distinct history (Figure 1A, SI
280 Figure 2). HGT-chimeras are unlikely to be detected by traditional HGT detection methods,
281 which assume a common evolutionary history throughout the length of the open reading frame.
282 In almost half (41/104) of HGT-chimeras, a naive interpretation of gene-wide BLASTp results via
283 the same criteria used to infer interval ancestry would lead to a false inference of gene-wide
284 metazoan ancestry (see SI Table 19). For the remaining genes, constructing a single gene tree
285 for genic intervals with distinct histories would be methodologically problematic and, even if it
286 resulted in a sensible topology, would obscure the true chimeric history of the gene. We expect
287 that the application of methods like the one we describe here, that allow for intragenic
288 phylogenetic discordance, could lead to the discovery of more HGT-chimeras, especially in
289 lineages with frequent HGT such as grasses and parasitic plants (80, 81), fungi (82, 83) and
290 prokaryotes (4, 13).

291 Intriguingly, a majority (81/104) of identified HGT-chimeras were found in only a single
292 species (Figure 1E, SI Table 4). The two species with the greatest number of species-specific
293 chimeras (N=8 for both *Folsomia candida* and *Eurytemora carolleeae*) are both separated from
294 their closest searched relatives by hundreds of millions of years (242 MYA for *Folsomia candida*
295 vs. *Orchesella cincta* (84); 446 MYA for *Eurytemora carolleeae* vs. *Tigriopus californicus* (85)).
296 We therefore speculate that at least some of the observed excess of species-specific singletons
297 reflects the poor genomic sampling of many arthropod lineages. We nonetheless uncover
298 evidence of species-specific (SI Table 4) or young (origin <20 MYA, SI Table 5) HGT-chimeras
299 even in densely sampled lineages such as the genus *Daphnia* and order Lepidoptera. Excesses
300 of species-specific genes have been reported in studies of gene birth via other mechanisms (5,
301 10, 15, 86), suggesting that some (though not all, see SI Table 5) newly formed genes may be
302 short-lived (10, 86) due to weak (10, 86, 87) or evolutionarily transient (5, 88) selective
303 constraint.
304 The biological significance of HGT has been questioned by some researchers, as many
305 transferred segments show evidence of pseudogenization (16, 23, 89), and significant barriers
306 to the transfer and functional integration of genes from divergent lineages have been posited
307 (12, 17, 88, 90, 91). We speculate that fusion with pre-existing gene sequences might be one
308 path for the integration of novel horizontally transferred sequences into pre-existing functional
309 networks. Multiple HGT-chimeras show signatures consistent with biological function, including
310 multiple lines of evidence for active transcription (SI Figure 3C, SI Tables 12-13), gene-wide
311 purifying selection (Figure 3C-D, SI Tables 14-15) and functionally coherent domain
312 architectures (Figure 4, SI Table 18). Experimental studies in diverse model systems have
313 demonstrated important functions for many young and species-specific genes (3, 7, 11, 34, 73,
314 92),. Thus, we consider it plausible that even the young HGT-chimeras detected here may play
315 important biological roles.

316 The functional predictions we generate via domain architectures and DeepGO-SE (SI
317 Tables 17-18) provide insights into the potential range of biological processes impacted by
318 HGT-chimeras, warranting experimental validation in future studies. We speculate that some
319 HGT-chimera intervals may have been co-opted to perform functions divergent from the
320 functions of their non-chimeric progenitors, given molecular signatures of post-fusion
321 neofunctionalization (Figure 3F). Consistent with this idea, co-option has been experimentally
322 shown for the HGT-derived domain of the HGT-chimeric gene *oskar*. The closest bacterial
323 relatives of *oskar*'s HGT-derived domain, called the OSK domain (45), are lipid hydrolases (45,
324 46), but in insects this domain lacks key hydrolytic residues and participates in RNA binding (45,
325 93, 94). At a phenotypic level, *oskar* is a core mediator of an evolutionarily novel mechanism of
326 germ line specification in some insects (41, 95, 96), a function apparently physiologically
327 removed from the likely ancestral bacterial role of its HGT-derived domain.
328 To conclude, our novel computational approach has revealed that fusion of horizontally
329 acquired sequence with endogenous ones has been widespread throughout the history of
330 arthropods, illustrating a previously underappreciated versatility in genomic "bricolage" (97).
331
332 **Methods**
333 Our methodology for HGT-chimera detection and downstream inference is described in brief
334 here and in detail in SI Methods. All BLAST searches were performed using DIAMOND
335 (v.2.0.15) BLASTp (68). We partitioned protein queries into intervals hits against the NR
336 database and an adaptation of a previously described algorithm (52). Separated intervals were
337 subjected to a second round of BLASTp, from which phylogenetic datasets were either directly
338 constructed or constructed from hmmsearch (HMMER suite v.3.3.2) results with a profile hidden
339 Markov model produced from arthropod BLASTp hits. Maximum likelihood phylogenies for
340 separated intervals were produced with IQ-TREE v.2.2.0.3 (98) from multiple sequence
341 alignments generated with MUSCLE v.5.1 (99) and trimAl (100). Trees were visualized with the

342 iTOL python API and manually inspected. dN/dS computations were performed with PAL2NAL
343 v.14.1(101) and PAML v.4.10.0 (68–70). HGT-chimera expression was determined
344 bioinformatically using the RNA-Seq tracks of the NCBI RefSeq genome browser, and
345 experimentally using RT-PCR and Sanger sequencing on samples of selected arthropod
346 species obtained as donations or collected by the authors from lab-cultured animals (SI Table
347 13).

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, a Life Science Research Foundation Postdoctoral Fellowship to IR, award number RGP0041/2022 from the Human Frontier Science Program (SP) to CGE, the Howard Hughes Medical Institute, and funds from Harvard University. 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 orthologous groups 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 80b40f0. Sanger-verified cDNA sequences have been deposited at GenBank, with accession numbers available in SI Table 13.

Author contributions

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

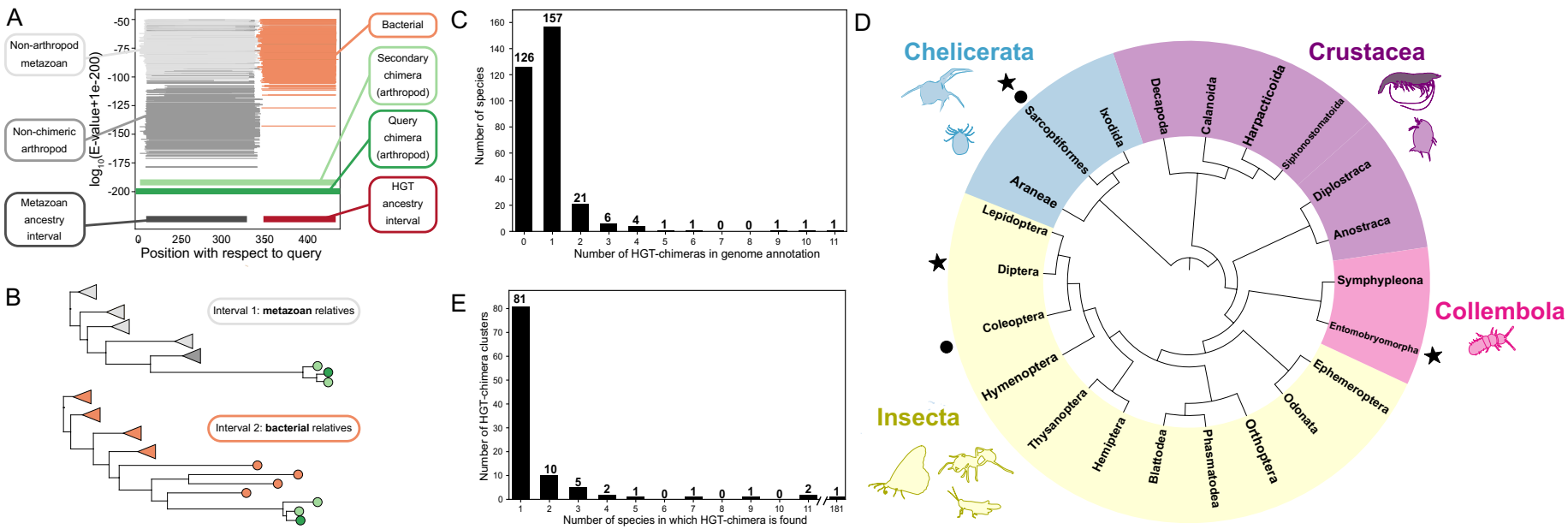

Figure 1. Identification of HGT-chimeric gene detection across Arthropoda. (A) The taxonomic distribution of BLAST hits across an example query sequence are used to hypothesize chimeric ancestry. In this hypothetical protein, the interval 1-325 amino acids are hypothesized to be of metazoan ancestry (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 proposed using maximum likelihood phylogenetics on hmmsearch or BLAST hits as described in SI Text 2. In this example, the results suggest an ancient metazoan ancestry for the interval spanning amino acids 1-325 (Interval 1: top tree) because the sequences from putative HGT-chimeras (green) are nested within a backbone of arthropod and non-arthropod metazoan sequences (grey). The interval spanning amino acids 350-450 (Interval 2: bottom tree) is proposed to be of bacterial ancestry because the chimeric sequences (green) are nested within a backbone of bacterial sequences (orange). All trees are available online from iTOL. **(C)** Distribution of the number of HGT-chimeras detected per genome for the 319 RefSeq reference genome annotations in the primary search dataset, with a maximum of one member of each HGT-chimera orthologous cluster tabulated per species (total N=274 genes). **(D)** This phylogram shows 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). **(E)** The number of species possessing each of 104 HGT-chimera orthologous clusters, including both RefSeq and GenBank genomes. Species range is examined by inspecting the species distribution of arthropod proteins that are hmmsearch or BLASTp hits to both the HGT and metazoan intervals of the representative HGT-chimeras (see SI Methods).

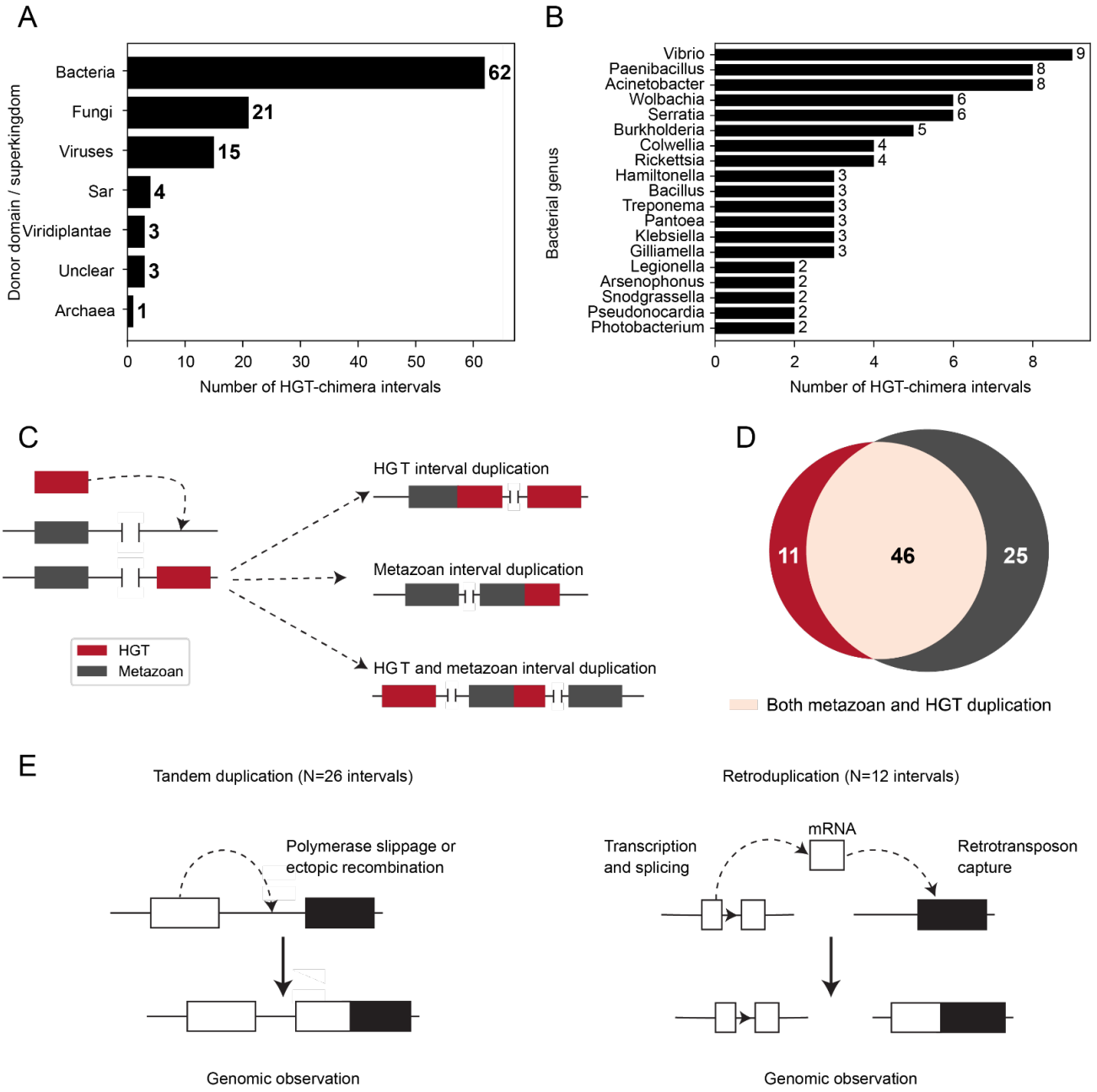
A
Donor domain / superkingdom
Bacteria
Fungi
Viruses
Sar
Viridiplantae
Unclear
Archaea
62
21
15
4
3
3
1
0
10
20
30
40
50
60
Number of HGT-chimera intervals
B
Bacterial genus
Vibrio 9
Paenibacillus 8
Acinetobacter 8
Wolbachia 6
Serratia 6
Burkholderia 5
Colwellia 4
Rickettsia 4
Hamiltonella 3
Bacillus 3
Treponema 3
Pantoea 3
Klebsiella 3
Gilliamella 3
Legionella 2
Arsenophonus 2
Snodgrassella 2
Pseudonocardia 2
Photobacterium 2
0
2
4
6
8
Number of HGT-chimera intervals
C
HGT interval duplication
Metazoan interval duplication
HGT and metazoan interval duplication
HGT
Metazoan
D
11
46
25
Both metazoan and HGT duplication
E
Tandem duplication (N=26 intervals)
Polymerase slippage or
ectopic recombination
Genomic observation
Retroduplication (N=12 intervals)
Transcription
and splicing
mRNA
Retrotransposon
capture
Genomic observation

Figure 2. (previous page) Evolutionary origins of HGT-chimeras. (A) 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). **(B)** Representation of bacterial genera from documented arthropod symbionts in the Symbiotic Genomes Database (102) among closest and second-closest clades to HGT-chimera intervals on maximum likelihood trees (see SI Methods). Plot shows all genera represented in more than one interval, with full data included in SI Table 8. **(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 homologs ("relatives") to the HGT or metazoan intervals within the same genome. 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 (86). **(D)** The number of HGT-chimeras with non-chimeric parents (DIAMOND BLASTp E-value <1e-10) detected at a different locus within the same genome, for HGT intervals, metazoan intervals, or both, based on BLASTp-based relative detection as follows: we considered an interval of a chimera to have a non-chimeric within-genome relative if it had a DIAMOND BLASTp hit to a protein encoded by a different locus (E-value <1e-10) and if the protein was not simultaneously a hit (E-value <10) for all chimera intervals of the opposite ancestry annotation (i.e. a metazoan relative should be a hit for the metazoan interval of a protein but not the HGT interval). Venn diagram includes data for the 82/104 HGT-chimeras for which we found evidence of a relative of at least one constituent interval. See also the results of phylogenetic relative detection in SI Figure 8C. **(E)** Two plausible mechanisms of gene fusion in HGT-chimeras. HGT-chimera intervals were hypothesized to be products of tandem duplication if HGT-chimeras were located on the same chromosome or genomic scaffold as their BLASTp-defined relatives with ≤ 2 intervening genes (left), or products of retroduplication if the interval was spanned by one exon in the HGT-chimera and more than one exon in the BLASTp-defined relative (right).

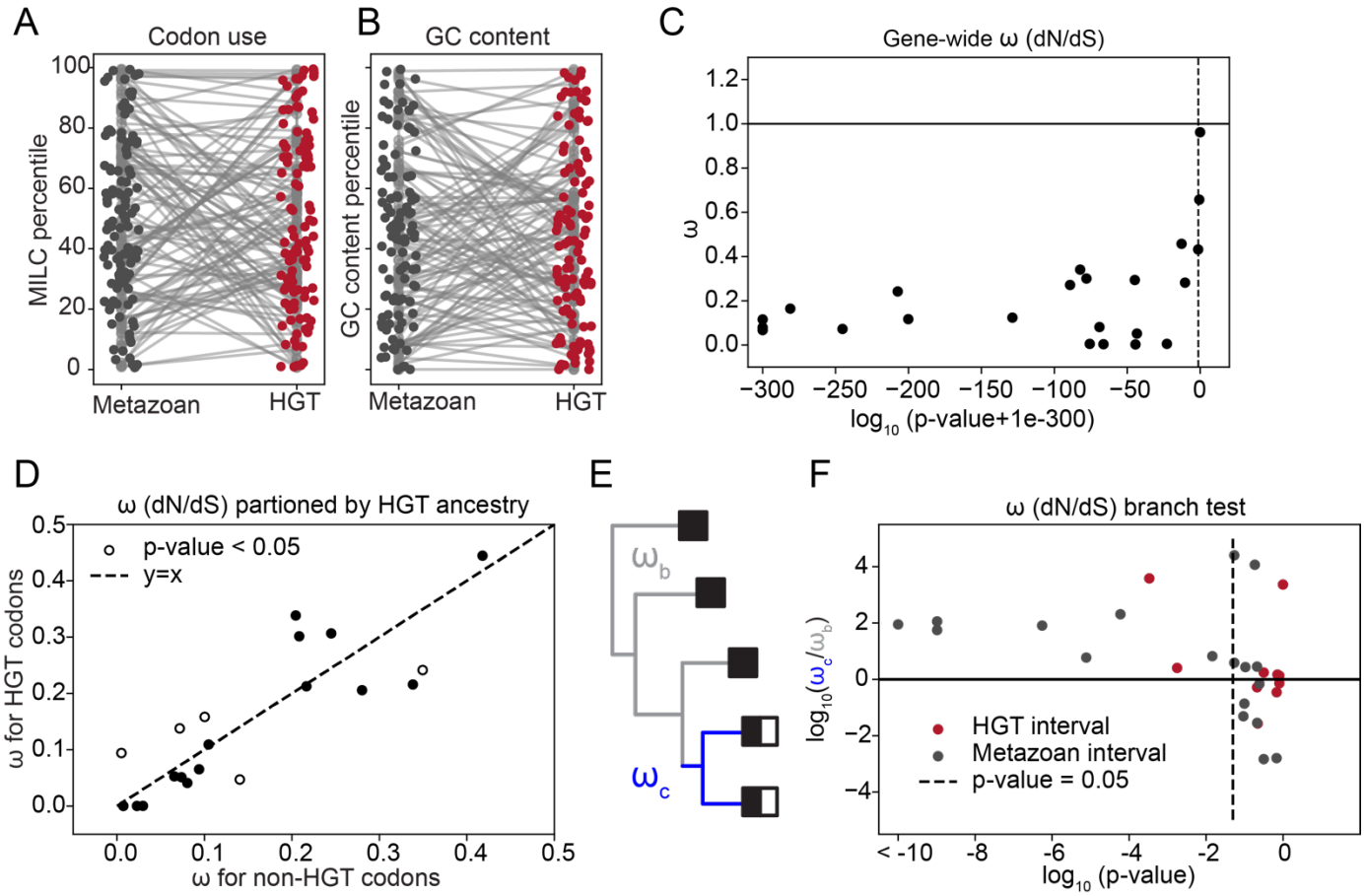
A
Codon use
MILC percentile
100
80
60
40
20
0
Metazoan
HGT
B
GC content
GC content percentile
Metazoan
HGT
C
Gene-wide ω (dN/dS)
ω
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−300 −250 −200 −150 −100 −50 0
log10 (p-value+1e-300)
D
ω (dN/dS) partioned by HGT ancestry
p-value < 0.05
y=x
ω for HGT codons
ω for non-HGT codons
0.0 0.1 0.2 0.3 0.4 0.5
E
ωb
ωc
F
ω (dN/dS) branch test
log10(ωc/ωb)
HGT interval
Metazoan interval
p-value = 0.05
< -10 -8 -6 -4 -2 0
log10 (p-value)

Figure 3. (previous page) Evidence of HGT-chimera functionality. (A) Measure Independent of Length and Composition (MILC) is a measure of codon use bias (103). Higher MILC values indicate greater deviation from genome-wide expected codon use frequencies. To enable comparison of codon use between HGT and metazoan intervals from the same protein as well as comparison to non-chimeric transcripts, we percentile normalized MILC values with respect to a distribution obtained via random sampling of annotated coding sequences from the same genome as each HGT-chimera. This plot shows percentile-normalized MILC across HGT (red) and metazoan-derived (dark grey) intervals across all 104 representative chimeras, with light grey lines connecting HGT and metazoan intervals from the same chimera (see SI Methods). Percentile-normalized MILC values were roughly uniformly distributed over [0,1], as would be expected for any genomic interval. HGT and metazoan intervals did not differ significantly in their percentile-normalized MILC values (Wilcoxon rank-sum p-value > 0.10). **(B)** GC values by interval type, with percentile scores produced as in (A). HGT and metazoan intervals do not differ significantly in their percentile-normalized GC values (Wilcoxon rank-sum p-value > 0.10). **(C)** Gene-wide ω (dN/dS) values (PAML M0) fitted for 21 of the 23 HGT-chimeras orthologous 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 dN/dS ratio values (y-axis) to a null model with ω fixed at 1. **(D)** Each of 23 HGT-derived intervals from the same 21 HGT-chimeras as in (C) 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. **(E)** Schematic representation of branch test performed to compare ω values of an interval before (gray) and after (blue) a 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 arthropods in the same taxonomic order (SI Methods). **(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.

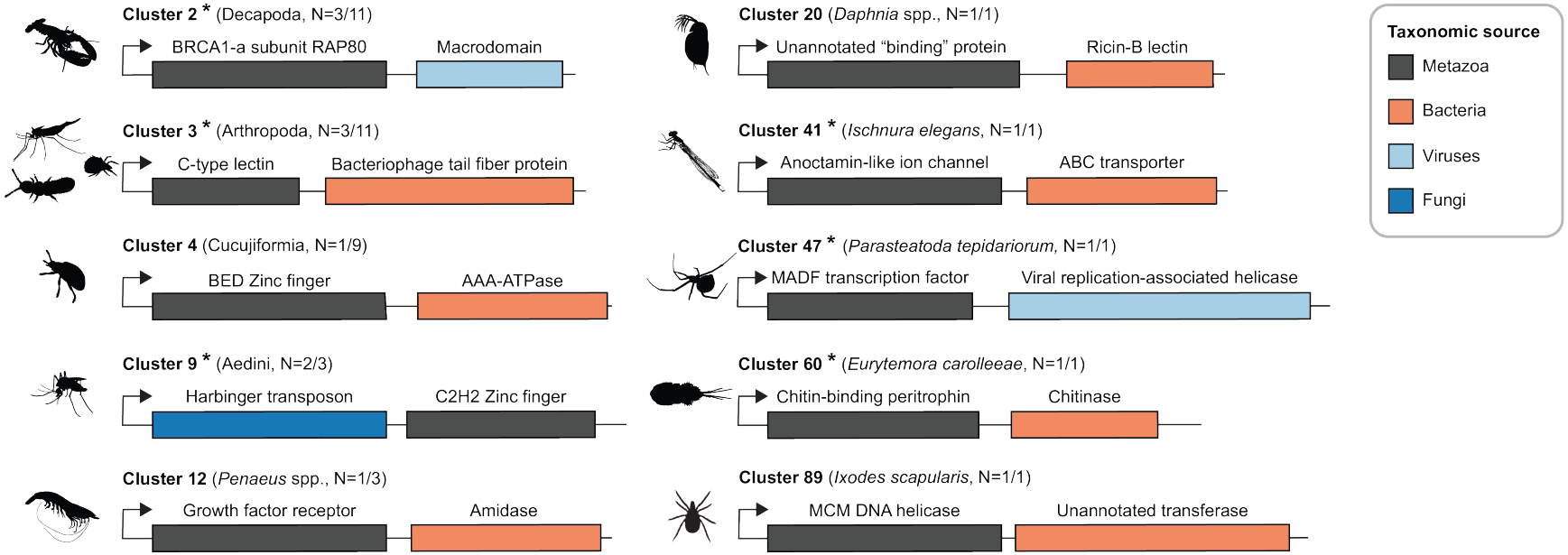

Figure 4. HGT-chimeras are transcribed. Predicted domain architectures (via NCBI CDD (75), InterProScan (74), and DeepGO-SE (76) for a sample of ten HGT-chimeras with expression validated by RT-PCR and Sanger sequencing (SI Table 13). The unique cluster identifier of each HGT-chimera is indicated, along with its taxonomic range in parenthesis. Asterisks indicate a predicted functional commonality between the metazoan and HGT intervals (SI Table 17). Note that cluster 3 may have undergone inter-arthropod HGT (SI Text 1). Representative silhouettes obtained from phylopic.org. Fractions next to taxon names indicate the number of total species in the cluster for which we confirmed chimera expression via RT-PCR (SI Table 13). Domain length is arbitrary and not to scale.

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

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

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

6. 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).

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

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

9. 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).

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

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

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

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

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

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

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

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

18. 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.

19. 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).

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

21. 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).

22. 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).

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

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

25. 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).

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

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

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

29. 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).

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

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

32. 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).

33. 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).

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

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

36. 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).

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

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

39. 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).

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

41. 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).

42. 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).

43. 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).

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

45. 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).

46. 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).

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

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

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

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

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

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

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

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

55. 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).

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

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

64. W. Wang, F. G. Brunet, E. Nevo, M. Long, Origin of sphinx, a young chimeric RNA gene in Drosophila melanogaster. *Proc. Natl. Acad. Sci.* **99**, 4448–4453 (2002).

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

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

67. Z. Yang, W. J. Swanson, Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes. *Mol. Biol. Evol.* **19**, 49–57 (2002).

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

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

70. 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).

71. Z. Yang, Computational Molecular Evolution. (2006). https://doi.org/10.1093/acprof:oso/9780198567028.001.0001.

72. A. Ahuja, C. G. Extavour, Patterns of molecular evolution of the germ line specification gene oskar suggest that a novel domain may contribute to functional divergence in Drosophila. *Dev. Genes Evol.* **224**, 65–77 (2014).

73. 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).

74. M. Blum, *et al.*, The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res.* **49**, D344–D354 (2021).

75. J. Wang, *et al.*, The conserved domain database in 2023. *Nucleic Acids Res.* **51**, D384–D388 (2023).

76. M. Kulmanov, *et al.*, Protein function prediction as approximate semantic entailment. *Nat. Mach. Intell.* **6**, 220–228 (2024).

77. J. Ao, E. Ling, X.-Q. Yu, Drosophila C-type lectins enhance cellular encapsulation. *Mol. Immunol.* **44**, 2541–2548 (2007).

78. K. Filik, B. Szermer-Olearnik, S. Oleksy, J. Brykała, E. Brzozowska, Bacteriophage Tail Proteins as a Tool for Bacterial Pathogen Recognition-A Literature Review. *Antibiotics* **11**, 555 (2022).

79. 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).

80. 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).

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

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

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

84. D. Yu, *et al.*, Whole-genome-based phylogenetic analyses provide new insights into the evolution of springtails (Hexapoda: Collembola). *Mol. Phylogenet. Evol.* **200**, 108169 (2024).

85. S.-I. Eyun, Phylogenomic analysis of Copepoda (Arthropoda, Crustacea) reveals unexpected similarities with earlier proposed morphological phylogenies. *BMC Evol. Biol.* **17**, 23 (2017).

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

87. N. Vakirlis, *et al.*, De novo emergence of adaptive membrane proteins from thymine-rich genomic sequences. *Nat Commun* **11**, 781 (2020).

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

89. 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).

90. 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).

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

92. 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).

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

94. E. L. Rivard, J. Srouji, A. Repouliou, C. G. Extavour, Protein sequence evolution underlies interspecies incompatibility of a cell fate determinant. *bioRxiv* **10.1101/2025.08.02.668269**, 1–53 (2025).

95. A. Kulkarni, C. G. Extavour, Convergent evolution of germ granule nucleators: A hypothesis. *Stem Cell Res.* **24**, 188–194 (2017).

96. J. A. Kao, E. L. Rivard, R. R. Kapoor, C. G. Extavour, There and back again: the dynamic evolution of panarthropod germ1cell specification mechanisms. *bioRxiv* **668520** (2025).

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

98. B. Q. Minh, *et al.*, IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).

99. R. C. Edgar, MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinform.* **5**, 1–19 (2004).

100. S. Capella-Gutiérrez, J. M. Silla-Martínez, T. Gabaldón, trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).

101. M. Suyama, D. Torrents, P. Bork, PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* **34**, W609-12 (2006).

102. M. Reyes-Prieto, C. Vargas-Chávez, A. Latorre, A. Moya, SymbioGenomesDB: a database for the integration and access to knowledge on host–symbiont relationships. *Database* **2015**, bav109 (2015).

103. F. Supek, K. Vlahoviček, Comparison of codon usage measures and their applicability in prediction of microbial gene expressivity. *BMC Bioinform.* **6**, 182 (2005).