Molecular evolutionary trends and feeding ecology diversification in the Hemiptera, anchored by the milkweed bug genome

Kristen A. Panfilio[1, 2]*, Iris M. Vargas Jentzsch[1], Joshua B. Benoit[3], Deniz Erezyilmaz[4], Yuichiro Suzuki[5], Stefano Colella[6, 7], Hugh M. Robertson[8], Monica F. Poelchau[9], Robert M. Waterhouse[10, 11], Panagiotis Ioannidis[10], Matthew T. Weirauch[12], Daniel S.T. Hughes[13], Shwetha C. Murali[13, 14, 15], John H. Werren[16], Chris G.C. Jacobs[17, 18], Elizabeth J. Duncan[19, 20], David Armisén[21], Barbara M.I. Vreede[22], Patrice Baa-Puyoulet[6], Chloé S. Berger[21], Chun-che Chang[23], Hsu Chao[13], Mei-Ju M. Chen[9], Yen-Ta Chen[1], Christopher P. Childers[9], Ariel D. Chipman[22], Andrew G. Cridge[19], Antonin J.J. Crumière[21], Peter K. Dearden[19], Elise M. Didion[3], Huyen Dinh[13], HarshaVardhan Doddapaneni[13], Amanda Dolan[16, 24], Shannon Dugan[13], Cassandra G. Extavour[25, 26], Gérard Febvay[6], Markus Friedrich[27], Neta Ginzburg[22], Yi Han[13], Peter Heger[28], Thorsten Horn[1], Yi-min Hsiao[23], Emily C. Jennings[3], J. Spencer Johnston[29], Tamsin E. Jones[25], Jeffery W. Jones[27], Abderrahman Khila[21], Stefan Koelzer[1], Viera Kovacova[30], Megan Leask[19], Sandra L. Lee[13], Chien-Yueh Lee[9], Mackenzie R. Lovegrove[19], Hsiao-ling Lu[23], Yong Lu[31], Patricia J. Moore[32], Monica C. Munoz-Torres[33], Donna M. Muzny[13], Subba R. Palli[34], Nicolas Parisot[6], Leslie Pick[31], Megan Porter[35], Jiaxin Qu[13], Peter N. Refki[21, 36], Rose Richter[16, 37], Rolando Rivera Pomar[38], Andrew J. Rosendale[3], Siegfried Roth[1], Lena Sachs[1], M. Emília Santos[21], Jan Seibert[1], Essia Sghaier[21], Jayendra N. Shukla[34, 39], Richard J. Stancliffe[40], Olivia Tidswell[19, 41], Lucila Traverso[42], Maurijn van der Zee[17], Séverine Viala[21], Kim C. Worley[13], Evgeny M. Zdobnov[10], Richard A. Gibbs[13], Stephen Richards[13]*

* Correspondence: kristen.panfilio@alum.swarthmore.edu; stephenr@bcm.edu

The full list of author information is available at the end of the manuscript.

ABSTRACT

Background:
The Hemiptera (aphids, cicadas, and true bugs) are a key insect order whose members offer a close outgroup to the Holometabola, with high diversity within the order for feeding ecology and excellent experimental tractability for molecular genetics. Sequenced genomes have recently become available for hemipteran pest species such as phloem-feeding aphids and blood-feeding bed bugs. To complement and build upon these resources, we present the genome sequence and comparative analyses centered on the large milkweed bug, *Oncopeltus fasciatus*, a seed feeder of the family Lygaeidae.

Results:
The 926-Mb genome of *Oncopeltus* is relatively well represented by the current assembly and official gene set, which supports *Oncopeltus* as a fairly conservative hemipteran species for anchoring molecular comparisons. We use our genomic and RNA-seq data not only to characterize features of the protein-coding gene repertoire and perform isoform-specific RNAi, but also to elucidate patterns of molecular evolution and physiology. We find ongoing, lineage-specific expansion and diversification of repressive C2H2 zinc finger proteins and of intron gain and turnover in the Hemiptera. These analyses also weigh the relative importance of lineage and genome size as predictors of gene structure evolution in insects. Furthermore, we identify enzymatic gains and losses that correlate with hemipteran feeding biology, particularly for reductions in chemoreceptor family size and loss of metabolic reactions within species with derived, fluid-nutrition feeding modes.

Conclusions:
With the milkweed bug genome, for the first time we have a critical mass of sequenced species representing a hemimetabolous insect order, substantially improving the diversity of insect genomics beyond holometabolans such as flies and ants. We use this addition to define commonalities among the Hemiptera and then delve into how hemipteran species' genomes reflect their feeding ecology types. Our novel and detailed analyses integrate global and rigorous manual approaches, generating hypotheses and identifying specific sets of genes for future investigation. Given *Oncopeltus*'s strength as an experimental research model, we take particular care to evaluate the sequence resources presented here, augmenting its foundation for molecular research and highlighting potentially general considerations exemplified in the assembly and annotation of this medium-sized genome.

Keywords:
Phytophagy; Transcription Factors; Gene Structure; Lateral Gene Transfer; RNAi; Gene family evolution; Evolution of Development

BACKGROUND

In the past few years, the number of animals with sequenced genomes has increased dramatically, and there are now over 100 insect species with assembled and annotated genomes [1]. However, the majority belong to the Holometabola (*e.g.*, flies, beetles, wasps, butterflies), the group characterized by a biphasic life history with distinct larval and adult phases separated by a dramatic metamorphosis during a pupal stage. With fewer than half of all orders, the Holometabola represent only a fraction of the full morphological and ecological diversity across the Insecta. This imbalance in genomic resources limits the exploration of this diversity, including the environmental and developmental requirements of a hemimetabolous life style with a progression of flightless nymphal (juvenile) instars. Addressing this paucity, we report here comparative analyses based on genome sequencing of the large milkweed bug, *Oncopeltus fasciatus*, as a hemimetabolous representative of the larger diversity of insects.

The Hemiptera, the order to which *Oncopeltus* belongs, comprise the most species-rich hemimetabolous order and a close outgroup to the Holometabola as part of the hemipteroid assemblage (or Acercaria), with the Thysanoptera as a sister order and the Psocodea also traditionally included in this clade [2, 3]. All Hemiptera share the same piercing and sucking mouthpart anatomy [4], yet they have diversified to exploit food sources ranging from seeds and plant tissues (phytophagy) to phloem sap (mucivory) and mammalian blood (hematophagy). For this reason, many hemipterans are agricultural pests or human disease vectors, and genome sequencing efforts to date have focused on these species (Fig. 1), including phloem-feeding aphids [5-7], psyllids [8], and planthoppers [9], and the hematophagous kissing bug, *Rhodnius prolixus* [10], a vector of Chagas disease, and bed bug, *Cimex lectularius* [11, 12]. Building on transcriptomic data, genome projects are also in progress for other pest species within the same infraorder as *Oncopeltus*, such as the stink bug *Halyomorpha halys* [13, 14].

In this context, *Oncopeltus* represents a relatively benign species with conservative life history traits, affording a baseline against which other species can be compared. As a seed feeder, *Oncopeltus* has not undergone the marked life style changes that are associated with fluid feeding (mucivory or hematophagy), including dependence on endosymbiotic bacteria to provide needed complements lacking in the diet. For example, in the pea aphid, *Acyrthosiphon pisum*, its obligate endosymbiont, *Buchnera aphidicola,* provides essential amino acids and vitamins: previous analysis of the two genomes revealed a complementation of the two organisms' amino acid metabolism systems [5, 15]. Similarly, although hematophagy arose independently in *Rhodnius* and *Cimex* [16], their respective endosymbionts, *Rhodococcus rhodnii* and *Wolbachia*, provide vitamins lacking in the blood diet [17]. In contrast, the seed-feeding subfamily Lygaeinae, including *Oncopeltus*, is notable for the absence of prominent endosymbiotic anatomy: these bugs lack not only the midgut crypts that typically house bacteria but also the bacteriomes and endosymbiotic balls seen even in other Lygaeidae [18].

Nonetheless, as the native food source of *Oncopeltus* is the milkweed plant, its own feeding biology has a number of interesting implications associated with detoxification and sequestration of cardenolide compounds, including the bright red-

orange aposematic (warning) coloration seen in *Oncopeltus* embryos, nymphs, and adults [19, 20]. Thus, diet, metabolism, and body pigmentation are functionally linked biological features for which one may expect changes in gene repertoires to reflect diversity across species of the same order, and the Hemiptera provide an excellent opportunity to explore this.

Furthermore, *Oncopeltus* has been an established laboratory model organism for over 60 years, with a rich experimental tradition in a wide range of studies from physiology and development to evolutionary ecology [20-22]. It is among the few experimentally tractable hemimetabolous insect species, and it is amenable to a range of molecular techniques (*e.g.*, [23-25]). In fact, it was one of the first insect species to be functionally investigated by RNA interference (RNAi, [26]). RNAi in *Oncopeltus* is highly effective across different life history stages, which has led to a resurgence of experimental work over the past fifteen years, with a particular focus on the evolution of developmentally important regulatory genes (reviewed in [22]).

Focusing on these two avenues – feeding biology diversity within the Hemiptera and *Oncopeltus* as a research model for macroevolutionary genetics – we present here key insights derived from a combination of global comparative genomics and detailed computational analyses supported by extensive manual curation, empirical data for gene expression, sequence validation, and new isoform-specific RNAi. Namely, we identify sets of genes with potentially restricted life history expression in *Oncopeltus* and that are unique to the Hemiptera, clarify evolutionary patterns of zinc finger protein expansion, identify predictors of insect gene structure, and identify lateral gene transfer and amino acid metabolism features that correlate with feeding biology.

RESULTS AND DISCUSSION

The genome and its assembly

Oncopeltus fasciatus has a diploid chromosome number (2*n*) of 16, comprised of seven autosomal pairs and two sex chromosomes with the XX/XY sex determination system [27, 28]. To analyze this genetic resource, we sequenced and assembled the genome using next-generation sequencing approaches (Table 1; see also Methods and Supplemental Notes Sections 1-4). We measure the genome size to be 923 Mb in females and 928 Mb in males based on flow cytometry data (see also Supplemental Note 2.1.a), such that the assembly contains 84% of the expected sequence in assembled contigs, which is comparable to that of other recent, medium-sized insect genomes [11, 29]. However, our analyses of the *k*-mer frequency distribution in raw sequencing reads yielded ambiguous estimates of genome size and heterozygosity rate, which is suggestive of both high heterozygosity and high repetitive content ([30], see also Supplemental Note 2.1.b). Consistent with this, in further analyses we obtained high estimates of repetitive content (see below), which would imply a large proportion of potentially redundant sequence and possible misassembly within contigs of the current assembly. This phenomenon may be increasingly relevant as comparative genomics based on short read sequencing extends to additional insect species with genomes in the 1-Gb range.

As template DNA was prepared from dissected adults from which gut material was removed, the resulting assembly is essentially free of contamination, with only five small scaffolds with high bacterial homology (each to a different, partial bacterial genome, see also Supplemental Note 2.2), which either represent trace bacterial contamination or lateral gene transfers that have not assembled to flanking eukaryotic DNA, making their confirmation in the current assembly difficult.

Table 1. *Oncopeltus fasciatus* genome metrics.

Feature	Value	
2*n* chromosomes	16	
Genome size	926 Mb (mean between males and females)	
Assembly size	1,099 Mb (contigs only: 774 Mb)	
Coverage	106.9× raw coverage, 83.7% of reads in final assembly	
Contig N50	4,047 bp	
Scaffold N50	340.0 kb	
# Scaffolds	17,222	
GC content	genome: 32.7%, protein-coding sequence (OGS v1.2): 42%	
OGS v1.1 (curated fraction)	19,690 models[1] (1,426 models, 7.2%)	19,465 genes (1,201 genes, 6.2%)
OGS v1.2 (curated fraction)	19,809 models[1] (1,697 models, 8.7%)	19,616 genes (1,518 genes, 7.7%)

[1] Individual genes may be represented by multiple models in cases of curated alternative isoforms or if the gene is split across scaffolds.

192
193 **The official gene set and conserved gene linkage**
194 The official gene set (OGS) was generated by automatic annotation followed by
195 manual curation in a large-scale effort by the research community (see also
196 Supplemental Notes Sections 3-4). Curation revised automatic annotation models,
197 added alternative isoforms and *de novo* models, and documented multiple discrete
198 models for genes whose exons were split across scaffolds. We found that automatic
199 predictions were somewhat conservative for hemipteran gene structure (see below),
200 and manual curation primarily resulted in larger gene loci as exons were added and/or
201 extended, including merging discrete automatic models (see also Supplemental Note
202 4, Table S4.4). The OGS v1.1 was generated for global, pipeline analyses to
203 characterize the gene repertoire. The latest version, OGS v1.2, represents a minor
204 update, primarily for the addition of chemoreceptor genes of the ionotropic and
205 odorant receptor classes and curation of genes encoding metabolic enzymes.
206 Altogether, the research community curated 1,697 gene models (8.7% of OGS v1.2),
207 including 316 *de novo* models (see also Table S4.1). Reflecting the primary research
208 interests of the community (see also Supplemental Notes Section 5), the majority of
209 curated models are for genes encoding cuticular proteins (11%), chemoreceptors
210 (19%), and developmental regulators such as transcription factors and signaling
211 pathway components (40%, including the BMP/TGF-β, Toll/NF-κB, Notch,
212 Hedgehog, Torso RTK, and Wnt pathways).
213
214 In addition to assessing gene model quality, manual curation of genes whose
215 orthologs are expected to occur in syntenic clusters also validates assembly
216 scaffolding. Complete loci could be found for single orthologues of all Hox cluster
217 genes, where *Hox3*/*zen* and *Hox4*/*Dfd* are linked in the current assembly and have
218 ≥99.9% nucleotide identity with experimentally validated sequences ([31-33],
219 Supplemental Note 5.1.b). Conserved linkage was also confirmed for the homeobox
220 genes of the Iroquois complex, the Wnt ligands *wingless* and *wnt10*, and two linked
221 pairs from the Runt transcription factor complex (Supplemental Notes 5.1.a, 5.1.c,
222 5.1.i, 5.1.j). Further evidence for correct scaffold assembly comes from the curation
223 of large, multi-exonic loci. For example, the cell polarity and cytoskeletal regulator
224 encoded by the conserved *furry* gene includes 47 exons spanning a 437-kb locus,
225 which were all correctly assembled on a single scaffold.
226
227
228 **Transcriptomic resources and gene expression profiles across the milkweed bug**
229 **life cycle**
230 To augment published transcriptomic resources [34, 35], we sequenced three different
231 post-embryonic samples ("i5K" dataset, see Methods). We then compared the OGS
232 to the resulting *de novo* transcriptome and to a previously published embryonic and
233 maternal (ovary) transcriptome ("454" pyrosequencing dataset, [34]). Our OGS is
234 quite comprehensive, containing 90% of transcripts from each transcriptomic dataset
235 and an additional 3,146 models (16% of OGS: Fig. 2a). Among the additional
236 models, 274 (9%) were manually validated, including 163 *de novo* models for odorant
237 and gustatory receptors. These gene classes are known for lineage-specific
238 expansions and highly tissue- and stage-specific expression, with usually only one
239 receptor expressed per sensory neuron ([36, 37], and see below).
240

241 Furthermore, the OGS does a good job of "mopping up" partial and
242 unidentified 454 transcripts. We could substantially improve orthologous gene
243 discovery by mapping the 454 transcripts to the OGS (blastn, $e<10^{-9}$), nearly trebling
244 the proportion of transcripts with an assigned gene model or homology compared to
245 the original study (from 9% to 26%). This included 10,130 transcripts that primarily
246 mapped to UTRs and could not have been identified by coding sequence homology,
247 such as the 654-bp transcript for the *Oncopeltus brinker* ortholog, which encodes a
248 putative inhibitor of the BMP pathway ([38], see also Supplemental Note 5.1.f), and
249 four unassembled transcripts each from the 3′ UTRs of the enzyme-encoding genes
250 *CTP synthase* and *roquin*. At the same time, the transcriptomes provided expression
251 support for the identification of multiple isoforms in the OGS. For example, we could
252 confirm previously described isoforms for the germline determinant encoded by
253 *nanos* [34]. Where assembly limitations curtailed OGS gene models, full-length
254 transcripts are represented in the transcriptomes, such as for the ecdysis regulator
255 CCAP-R [39] and the chromatin linker Histone H1.
256
257 We then took advantage of our stage-specific RNA datasets to provide an
258 initial survey of gene expression profiles across biological samples and across the life
259 cycle. Most OGS gene models have expression support (91% of 19,690), with 74%
260 expressed broadly in at least three of four samples (Fig. 2b). The inclusion of a fifth
261 dataset from a published adult library [35] provided only a 1% gain in expression
262 support (218 gene models), indicating that with the current study the expression data
263 volume for *Oncopeltus* is quite complete. At the same time, direct comparison of the
264 three adult samples suggests that the published adult dataset of unspecified sex is
265 probably male, as it shares 4.6× more expressed genes with our male than our female
266 sample.
267
268 As these data derive from limited biological sampling, we remain cautious
269 about true stage specificity and do not quantify expression levels. We do, however,
270 note that most genes with stage-restricted expression are in sets involving our male
271 sample (Fig. 2b: male-only or male and nymph), although this sample does not
272 contain more reads or more expressed genes. Furthermore, we also find stage-
273 specific patterns for some of our most abundant curated gene classes. Gustatory
274 receptor (GR) genes show noticeable restriction to the adult male and published adult
275 (probable male) samples (n= 169 GRs: 40% no expression, 27% only expressed in
276 these two samples), with half of these expressed in both biological replicates (52%).
277 Interestingly, the nymphal sample is enriched for genes encoding structural cuticular
278 proteins (94%, which is >56% more than any other sample). This likely reflects the
279 ongoing molting cycles, with their cyclical upregulation of cuticular gene synthesis
280 [40], that are experienced by the different instars and molt cycle stages of individuals
281 pooled in this sample.
282
283
284 **Protein orthology and hemipteran copy number comparisons**
285 To further assay protein-coding gene content, we then compared *Oncopeltus* with
286 eleven other arthropod species. A phylogeny based on single copy orthologs correctly
287 reconstructs the hemipteran and holometabolan clades' topologies (Fig. 3a, compare
288 with Fig. 1a), although larger-scale insect relationships remain challenging [3]. In
289 expanding this to the Benchmarking Universal Single-Copy Orthologs (BUSCO,
290 [41]) dataset of 2,675 Arthropoda genes, we also found that most BUSCO genes are

291 present in the *Oncopeltus* OGS, although with additional genes identified on genomic
292 scaffolds but not yet incorporated into the gene set (see also Supplemental Note 6.1).
293 We next categorized all proteins by conservation in global, clustering-based orthology
294 analyses [42]. As in most species, half of *Oncopeltus* proteins (51%) falls within the
295 top three conservation levels (Fig. 3a). Moreover, 98% of all *Oncopeltus* protein-
296 coding genes has homology, expression, and/or curation support (Fig. 3b), including
297 support for 80% of proteins without homology, such as a few species-specific
298 chemoreceptors and antimicrobial peptides (see also Supplemental Note 5.1.h), while
299 some unsupported models may be split or partial. Overall, we estimate that the
300 *Oncopeltus* protein repertoire is comparable to that of other insects in size and degree
301 of conservation.
302
303 In contrast, the pea aphid, *Acyrthosiphon pisum*, is a notable outlier even
304 among fellow Hemiptera, where we provide a side-by-side comparison with
305 *Oncopeltus* as well as the recently-sequenced kissing bug, *Rhodnius prolixus* [10],
306 and bed bug, *Cimex lectularius* [11, 12]. The pea aphid is striking for its long branch
307 in phylogenetic comparisons and for its large protein-coding gene content with low
308 conservation (Fig. 3a), consistent with the observation of numerous lineage-specific
309 duplications [5]. As the first hemipteran to have its genome sequenced, the pea aphid
310 has often been used to boost taxonomic sampling in phylogenomic comparisons (*e.g.*,
311 [29]). However, the pea aphid may not be the best representative, and as more
312 hemipteran genomes are sequenced, other species now offer less derived alternatives.
313
314
315 Compared to the pea aphid [43], *Oncopeltus* is more conservative in both
316 presence and copy number for several signaling pathway components. In contrast to
317 gene absences described for the pea aphid, *Oncopeltus* retains orthologs of the EGF
318 pathway component *sprouty*, the BMP receptor *wishful thinking*, and the hormone
319 nuclear receptor *Hr96* (see also Supplemental Note 5.1.e). Similarly, whereas
320 multiple copies were reported for the pea aphid, we find a single *Oncopeltus* ortholog
321 for the BMP pathway components *decapentaplegic* and *Medea* and the Wnt pathway
322 intracellular regulator encoded by *shaggy/GSK-3*, albeit with five potential isoforms
323 of the latter (see also Supplemental Notes 5.1.f, 5.1.j). Duplications of miRNA and
324 piRNA gene silencing components likewise seem to be restricted to the pea aphid
325 compared to other hemipterans – including other aphid species ([44], see also
326 Supplemental Note 5.4.a). However, our survey of *Oncopeltus* and several other
327 hemimetabolous species reveals evidence for frequent parallel duplications of the Wnt
328 pathway component *armadillo/β-catenin*, yet without the sequence and functional
329 divergence previously observed independently in the pea aphid and *Tribolium* ([45],
330 see also Supplemental Note 5.1.j). Curiously, *Oncopeltus* appears to encode fewer
331 histone loci than any other arthropod genome and yet exhibits a similar, but possibly
332 independent, pattern of duplications of histone acetyltransferases to those previously
333 identified in *Cimex* and the pea aphid (see also Supplemental Note 5.4.c).
334
335 On the other hand, we documented several notable *Oncopeltus*-specific
336 duplications. Whereas two copies of the BMP transducer *Mad* were reported in the
337 pea aphid [43], we find evidence for three paralogs in *Oncopeltus*, where two of these
338 genes occur in tandem and may reflect a particularly recent duplication (see also
339 Supplemental Note 5.1.f). Similarly, a tandem duplication of *wnt8* appears to be
340 unique to *Oncopeltus* (see also Supplemental Note 5.1.j). More striking is the

341 identification of six potential paralogs of *cactus*, a member of the Toll/NF-κB
342 signaling pathway for innate immunity, whereas the bed bug and kissing bug
343 each retain only a single copy ([46], see also Supplemental Note 5.1.g).
344
345 We then took advantage of broader comparative datasets [42] to identify
346 lineage-specific features of the Hemiptera. In other words, what makes a bug a bug in
347 terms of protein-coding genes? To address this, we partitioned an orthology analysis
348 of 64 insect species into three broad taxonomic groups (Fig. 3c). Highlights for the
349 Hemiptera, which are further corroborated in an updated dataset with 116 insect
350 species (OrthoDB v.9.1, [1]), fall into two classes. The first class contains potentially
351 new genes that show no homology outside the Hemiptera. We identified three such
352 instances with orthologous protein members present in at least four hemipterans, and
353 where no conserved functional domains were recognized. Interpretation of these
354 intriguing "uncharacterized proteins" will have to await direct experimental analyses,
355 for which the Hemiptera in general are particularly amenable (*e.g.*, [47-50]). The
356 second class comprises proteins with recognized functional domains and homologs in
357 other insects, but where evolutionary divergence has led to hemipteran-specific
358 subfamilies. For example, one protein orthology group ("orthogroup"
359 EOG090W0V4B) is comprised of a heteropteran-specific cytochrome P450 (CYP)
360 enzyme that in *Oncopeltus* is expressed in all life history stages. The expansion of
361 this protein family is associated with a species' potential scope for insecticide
362 resistance, as specific P450s can confer resistance to specific chemicals (*e.g.,* [51, 52];
363 see also Supplemental Notes 5.3.b, 5.3.c). Hence, the identification of lineage-
364 specific CYP enzymes can suggest potential targets for integrated pest management
365 approaches.
366
367
368 **Transcription factor repertoires and homeobox gene evolution**
369 Having explored the global protein repertoire, we next focused specifically on
370 transcription factors (TFs), which comprise a major class of proteins that has been
371 extensively studied in *Oncopeltus*. This is a class of key regulators of development
372 whose functions can diverge substantially during evolution and for which RNAi-
373 based experimental investigations have been particularly fruitful in the milkweed bug
374 (*e.g.*, [31, 32, 53-55], see also Supplemental Notes 5.1.a-e).
375
376 To systematically evaluate the *Oncopeltus* TF repertoire, we used a pipeline to
377 scan all predicted proteins and assign them to TF families, including orthology
378 assignments in cases where DNA binding motifs could be predicted (see Methods,
379 [56]). We identified 762 putative TFs in *Oncopeltus*, which is similar to other insects
380 of diverse orders for total TF count and for the size of each TF family (Fig. 4a: note
381 that the heatmap also reflects the large, duplicated repertoire in the pea aphid; see also
382 Tables S6.2-6.4).
383
384 We were able to infer DNA binding motifs for 25% (n=189) of *Oncopeltus*
385 TFs, mostly based on data from *D. melanogaster* (121 TFs) but also from distantly
386 related taxa such as mammals (56 TFs). Such high conservation is further reflected in
387 the fact that most proteins within several large TF families have inferred motifs and
388 therefore explicit orthology assignments, including for the homeodomain (53 of 85,
389 62%), basic helix-loop-helix (bHLH, 35 of 45, 78%), and forkhead box (16 of 17,
390 94%) families. In contrast, most C2H2 zinc finger proteins lack orthology assignment

391 (only 22 of 360, 6%). Across species, the homeodomain and C2H2 zinc finger
392 proteins are the two largest TF superfamilies (Fig. 4a). Given their very different
393 rates of orthology assignment in *Oncopeltus*, we probed further into the pipeline
394 predictions and the patterns of evolutionary diversification of these proteins.
395
396 The number of homeodomain proteins identified by the pipeline displays a
397 narrow normal distribution across species (Fig. 4b, mean ± standard deviation: 97 ±
398 9), consistent with a highly conserved, slowly evolving protein family. Supporting
399 this, many *Oncopeltus* homeodomain proteins that were manually curated also
400 received a clear orthology assignment (Fig. 4c: pink), with only four exceptions (Fig.
401 4c: yellow). Only one case suggests a limitation of a pipeline that is not specifically
402 tuned to hemipteran proteins (Goosecoid), while an incomplete gene model received
403 homeodomain classification but without explicit orthology assignment (Distal-less).
404 Manual curation of other partial or split models identified a further 11 genes encoding
405 homeodomains, bringing the actual tally in *Oncopeltus* to 96, which is comparable to
406 the mean across species. Overall, we find the TF pipeline results to be a robust and
407 reasonably comprehensive representation of these gene classes in *Oncopeltus*.
408
409 These analyses also uncovered a correction to the published *Oncopeltus*
410 literature for the key developmental patterning proteins encoded by the closely related
411 paralogs *engrailed* and *invected*. These genes arose from an ancient tandem
412 duplication prior to the hexapod radiation, where their tail-to-tail orientation enables
413 ongoing gene conversion [57], making orthology discrimination particularly
414 challenging. For *Oncopeltus*, we find that the genes also occur in a tail-to-tail
415 orientation and that *invected* retains a diagnostic alternative exon [57]. These new
416 genomic and expression data reveal that the *Oncopeltus* gene used as the purported
417 *engrailed* ortholog in previous developmental studies (*e.g.*, [53, 58-61]) is in fact
418 *invected* (see also Supplemental Note 5.1.a).
419
420
421 **Independent expansions of C2H2 zinc fingers within the Hemiptera**
422 Unlike homeodomain proteins, C2H2 zinc finger (C2H2-ZF) repertoires are
423 prominent for their large family size and variability throughout the animal kingdom
424 [62], and this is further supported by our current analysis in insects. With >350
425 C2H2-ZFs, *Oncopeltus*, the pea aphid, the termite, and several mosquito species have
426 1.5× more members than the insect median (Fig. 4b). This is nearly half of all
427 *Oncopeltus* TFs. While the expansion in mosquitoes could have a single origin after
428 the Culicinae diverged from the Anophelinae, the distribution in the Hemiptera, where
429 *Cimex* has only 227 C2H2-ZFs, suggests that independent expansions occurred in
430 *Oncopeltus* and the pea aphid. Prior to the sequencing of other hemipteran genomes,
431 the pea aphid's large C2H2-ZF repertoire was attributed to the expansion of a novel
432 subfamily, APEZ, also referred to as zinc finger 271-like [43].
433
434 In fact, manual curation in *Oncopeltus* confirms the presence of a subfamily
435 with similar characteristics to APEZ (Fig. 4c: 42% of all C2H2-ZFs were curated,
436 including 38% of those without orthology assignment, yellow). Specifically, in
437 *Oncopeltus* we find >115 proteins of the ZF271 class that are characterized by
438 numerous tandem repeats of the C2H2-ZF domain and its penta-peptide linker, with
439 3-45 repeats per protein.
440

441 However, at both the gene and protein levels we find evidence for ongoing
442 evolutionary diversification of the *Oncopeltus* ZF271-like subfamily. A number of
443 *Oncopeltus* ZF271-like genes occur in tandem clusters of 4-8 genes, suggesting recent
444 duplication events. Yet, at the same time, gene structure (number and size of exons)
445 is not shared between genes within clusters, and we identified a number of probable
446 ZF271-like pseudogenes whose open reading frames have become disrupted –
447 consistent with high turnover. At the domain level, *Oncopeltus* ZF271-like proteins
448 differ in the sequence and length of the zinc finger domains amongst themselves and
449 compared to aphid proteins (WebLogo analysis [63]), similar to zinc finger array
450 shuffling seen in humans [64]. Furthermore, whole-protein phylogenetic analysis
451 supports independent, rapid expansions in the pea aphid and *Oncopeltus* (Fig. 4d).
452
453 Clustered zinc finger gene expansion has long been recognized in mammals,
454 with evidence for strong positive selection to increase both the number and diversity
455 of zinc finger domains per protein as well as the total number of proteins [65]. This
456 was initially found to reflect an arms-race dynamic of co-evolution between selfish
457 transposable elements and the C2H2-ZF proteins that would repress them [66]. In
458 vertebrates, these C2H2-ZF proteins bind to the promoters of transposable elements
459 via their zinc finger arrays and use their Krüppel-associated box (KRAB) domain to
460 bind the chromatin-remodeling co-repressor KAP-1, which in turn recruits
461 methyltransferases and deacetylases that silence the targeted promoter [67].
462
463 Insects do not have a direct ortholog of vertebrate KAP-1 (see also
464 Supplemental Note 5.4.d), and neither the aphid nor *Oncopeltus* ZF271-like
465 subfamilies possess a KRAB domain or any other domain besides the zinc finger
466 arrays. However, close molecular outgroups to this ZF271-like subfamily include the
467 developmental repressor Krüppel [68] and the insulator protein CTCF [69] (data not
468 shown). Like these outgroups, the *Oncopeltus* ZF271-like genes are strongly
469 expressed: 98% have expression support, with 86% expressed in at least three
470 different life history stages (Fig. 2b). Thus, the insect ZF271-like proteins may also
471 play prominent roles in repressive DNA binding. Indeed, we find evidence for a
472 functional methylation system in *Oncopeltus* (see also Supplemental Note 5.4.c), like
473 the pea aphid, which would provide a means of gene silencing by chromatin
474 remodeling, albeit via mediators other than KAP-1.
475
476 However, an arms race model need not be the selective pressure that favors
477 insect ZF271-like family expansions. Recent analyses in vertebrates identified
478 sophisticated, additional regulatory potential by C2H2-ZF proteins, building upon
479 original transposable element binding for new, lineage-specific and even positive
480 gene regulation roles [64, 70, 71]. Moreover, although *Cimex* has half as many long
481 terminal repeat (LTR) repetitive elements as *Oncopeltus* and the pea aphid, they
482 constitute only a minor fraction of these species' transposable elements, and overall
483 we do not find a correlation between relative or absolute repetitive content and
484 ZF271-like family expansion within the Hemiptera (Fig. 5, and see below).
485
486
487 **Proportional repeat content across hemipterans**
488 With the aim of reducing assembly fragmentation and to obtain a better picture of
489 repeat content, we performed low coverage, long read PacBio sequencing in
490 *Oncopeltus* (see also Supplemental Note 2.3). Using PacBio reads in a gap-filling

assay on the Illumina assembly raised the total detected repetitive content from 25% to 32%, while repeat estimations based on simultaneous assessment of Illumina and PacBio reads nearly doubled this value to 58%. As expected, the capacity to identify repeats is strongly dependent on assembly quality and sequencing technology, with the *Oncopeltus* repetitive content underrepresented in the current (Illumina-only) assembly. Furthermore, as increasing genome size compounds the challenge of assembling repeats, the repeat content of the current assembly is lower than in species with smaller genome sizes (Fig. 5a, with the sole exception of the honey bee), and we therefore used our gap-filled dataset for further repeat profile comparisons.

To allow for direct comparisons among hemipterans, we also performed our RepeatModeler analysis on the bed bug and pea aphid assemblies. In these analyses, 36% and 31% of the respective assemblies were covered by repeats, similar to the gap-filled value of 32% in *Oncopeltus*. Nevertheless, given the smaller sizes of these species' assemblies – 651 Mb in the bed bug and 542 Mb in the pea aphid – the absolute repeat content is much higher in *Oncopeltus* (Fig. 5b**)**. Excluding unknown repeats, the most abundant transposable elements in *Oncopeltus* are LINE retrotransposons, covering 10% of the assembly (see also Table S2.5). This is also the case in the bed bug (12%), while in the pea aphid DNA transposons with terminal inverted repeats (TIRs) are the most abundant (2% of the assembly identified here, and 4% reported from manual curation in the pea aphid genome paper, [5]). Across species, the remaining repeat categories appear to grow proportionally with assembly size, except for simple repeats, which were the category with the largest relative increase in size after gap-filling in *Oncopeltus* (see also Supplemental Note 2.3). However, given the mix of sequence data types (Illumina only in the bed bug [11], Sanger in the pea aphid [5]), these patterns should be treated as hypotheses for future testing, until the assembly of repetitive regions becomes more feasible.

Lineage and genome size-related trends in insect gene structure

During manual curation, we noticed that *Oncopeltus* genes were often comprised of many, small exons. Furthermore, sequence conservation among the Hemiptera supported terminal coding sequence exons that were small and separated from the rest of the gene model by large introns. To explore patterns of gene structure across the insects, we undertook a broader comparative analysis. We find that both lineage and genome size can serve as predictors of gene structure.

Firstly, we created a high quality ("gold standard") dataset of 30 functionally diverse, large genes whose manual curation could reasonably ensure complete gene models across seven species from four insect orders (Fig. 6a; see also Supplemental Note 6.3). Most species encode the same total number of amino acids for these conserved proteins, with the thrips *Frankliniella occidentalis* and *Drosophila* being notable exceptions with larger proteins (Fig. 6a: blue plot line). However, the means of encoding this information differs between lineages, with hemipteroid orthologs comprised of twice as many exons as their holometabolous counterparts (Fig. 6a: orange plot line). Thus, there is an inverse correlation between exon number and exon size (Fig. 6a: orange vs. red plot lines). This analysis corroborates and extends previous probabilistic estimates of intron density, where the pea aphid as a sole hemipteran representative had the highest intron density of ten insect species [72].

541 To test these trends, we next expanded our analysis to all manually curated
542 exons in two species from each of three orders, including the Mediterranean fruit fly,
543 *Ceratitis captitata* [73], as a second dipteran alongside *Drosophila melanogaster*.
544 Here, we expect that curated exon sizes are accurate, without the need to assume that
545 entire gene models are complete. This large dataset supports our original findings,
546 with bugs having small exons, and with both the median and Q3 quartile reflecting
547 larger exon sizes in beetles and flies (Fig. 6b). Notably, the median and median
548 absolute deviation are highly similar between species pairs within the Hemiptera and
549 Coleoptera, irrespective of sample size. Meanwhile, the different exon metrics
550 between *Ceratitis* and *Drosophila* suggest that the large protein sizes we initially
551 observed in *Drosophila* (Fig. 6a: blue plot line) are a general but drosophilid-specific,
552 rather than dipteran-wide, feature.
553
554 Does the high exon count in the Hemiptera reflect an ancient, conserved
555 increase at the base of this lineage, or ongoing remodeling of gene structure with high
556 turnover? To assess the exact nature of evolutionary changes, we annotated intron
557 positions within multiple sequence alignments of selected proteins and plotted gains
558 and losses onto the phylogeny, providing a total sample of 165 evolutionary changes
559 at 148 discrete splice sites (Fig. 7; see also Supplemental Note 6.3 for gene selection
560 and method). These data reveal several major correlates with intron gain or loss. The
561 bases of both the hemipteroid and hemipteran radiations show the largest gains, while
562 most losses occur in the dipteran lineage (Fig. 7: orange and purple shading,
563 respectively). Furthermore, we find progressive gains across the hemipteroid nodes,
564 and it is only in these species that we additionally find species-specific splice changes
565 for the highly conserved *epimerase* gene (Fig. 7: orange outline). Thus, we find
566 evidence for both ancient intron gain and ongoing gene structure remodeling in this
567 lineage.
568
569 Surprisingly, both *hemocytin* and *epimerase* – our exemplar genes with many
570 (up to 74) and very few exons (3-8 per species), respectively – show independent
571 losses of the same splice sites in *Drosophila* and *Tribolium*. One feature these species
572 share is a genome size 2.4-6.0× smaller than all other species examined here (Fig. 7:
573 red shading). Pairwise comparisons within orders also support this trend, as the beetle
574 and fly species with larger genomes exhibit species-specific gains compared to intron
575 loss in their sister taxa (Fig. 7: red outlines). Thus, while lineage is a stronger
576 predictor of gene structure evolution (the coleopteran and dipteran species include one
577 each with a big or small genome and yet have highly similar metrics in Fig. 6b),
578 genome size seems to positively correlate with intron number (*e.g.,* the common
579 dipteran ancestor lost introns before *Ceratitis*, with a larger genome, experienced
580 subsequent gains: Fig. 7). A global computational analysis over longer evolutionary
581 distances also supports a link between genome size and intron number within
582 arthropods, but where chelicerates and insects may experience different rates of
583 evolutionary change in these features [74]. As new insect species' genomes are
584 sequenced, gene structure expectations at the ordinal level can help customize
585 parameters for automatic gene annotation, while it will be interesting to see if the
586 correlation with genome size is borne out in other taxa.
587
588 The selective pressures and mechanisms of intron gain in the Hemiptera will
589 be a challenge to uncover. While median exon size (Fig. 6b) could reflect species-
590 specific nucleosome sizes [75, 76], this does not account for the fact that most

591 hemipteran exons do not occur in multiples greater than a single nucleosome. Given
592 gaps in draft genome assemblies, we remain cautious about interpreting (large) intron
593 lengths but note that many hemipteran introns are too small to have harbored a
594 functional transposase gene (*e.g.,* median intron size of 429 bp, n=69 introns in
595 *hemocytin* in *Cimex*). Such small introns could be consistent with proliferation of
596 non-autonomous short interspersed nuclear elements (SINEs), although as highly
597 divergent non-coding elements their characterization in insects would require curated
598 SINE libraries comparable to those generated for vertebrates and plants [75, 76].
599 Meanwhile, it appears that hemipteran open reading frames ≥160 bp are generally
600 prevented by numerous in-frame stop codons just after the donor splice site. Most
601 stop codons are encoded by the triplet TAA in both *Oncopeltus* and *Cimex* (data not
602 shown), although these species' genomes are not particularly AT-rich (Table 1).
603
604 Even if introns are small, having gene loci comprised of numerous introns and
605 exons adds to the cost of gene expression in terms of both transcription duration and
606 mRNA processing. One could argue that a gene like *hemocytin*, which encodes a
607 clotting agent, would require rapid expression in the case of wounding – a common
608 occurrence in adult *Cimex* females due to the traumatic insemination method of
609 reproduction [11]. Thus, as our molecular understanding of comparative insect and
610 particularly hemipteran biology deepens, we will need to increasingly consider how
611 life history traits are manifest in genomic signatures at the structural level (*e.g.*, Figs.
612 5-7), as well as in terms of protein repertoires (Figs. 3-4).
613
614
615 **Expansion after a novel lateral gene transfer (LGT) event in phytophagous bugs**
616 In addition to the need for cuticle repair, traumatic insemination may be responsible
617 for the numerous LGT events predicted in the bed bug [11]. In contrast, the same
618 pipeline analyses [77] followed by manual curation predicted very few LGTs in
619 *Oncopeltus*, which lacks this unusual mating behavior. Here, we have identified 11
620 strong LGT candidates, and we confirmed the incorporation of bacterial DNA into the
621 milkweed bug genome for all five candidates chosen for empirical testing (see also
622 Table S2.4). Curiously, we find several LGTs potentially involved in bacterial or
623 plant cell wall metabolism that were acquired from different bacterial sources at
624 different times during hemipteran lineage evolution, including two distinct LGTs that
625 are unique to *Oncopeltus* and implicated in the synthesis of peptidoglycan, a bacterial
626 cell wall constituent (see also Supplemental Note 2.2).
627
628 Conversely, two further validated LGT candidates encode enzymes rather
629 known for their roles in degradation of bacterial cell walls: we find two strongly
630 expressed, paralogous copies in *Oncopeltus* of a probable bacterial-origin gene
631 encoding an endo-1,4-beta-mannosidase enzyme (MAN4, EC 3.2.1.78). This likely
632 ancient LGT event provides an interesting vignette that further illustrates gene
633 structure evolution processes within the Hemiptera. Inspection of genome assemblies
634 and predicted protein accessions reveals that this LGT event is shared with the stink
635 bug *Halyomorpha halys*, a member of the same infraorder (Pentatomomorpha), but
636 was introduced after this lineage diverged from other hemipterans, including the bed
637 bug (Fig. 8a). Furthermore, whereas *Oncopeltus* now has two copies of this gene,
638 independently the original *Halyomorpha* gene underwent a series of tandem
639 duplications leading to nine extant copies (Fig. 8b, see also Fig. S2.6). Since the
640 original LGT event, the *mannosidase* genes in both bug species have become

641 "domesticated" as multi-exonic genes (as in [78]). Moreover, as the splice site pattern
642 is unique to each species and evinces subsequent splice introductions in subsets of
643 paralogs (Fig. 8c), *mannosidase* genes further illustrate the hemipteran penchant for
644 intron introduction and maintenance of small exons. The retention and subsequent
645 expansion of these genes implies their positive selection, consistent with the
646 phytophagous diet of these hemipteran species. In this context, it is tempting to
647 speculate further that the marked proliferation of this enzyme in the stink bug
648 correlates with the breadth of its diet, as this agricultural pest feeds on a number of
649 different tissues in a range of host plants [79].
650
651
652 **Cuticle development, structure, and warning pigmentation**
653 Given the milkweed bug's history as a powerful model for endocrine studies of
654 hemimetabolous molting and metamorphosis since the late 1960's [21, 80-83], we
655 next focused on genes underlying the development and structural properties of the
656 *Oncopeltus* cuticle. Molting is triggered by the release of ecdysteroids, steroid
657 hormones that are synthesized from cholesterol in the prothoracic gland by
658 cytochrome P450 enzymes of the Halloween family [84], and we were able to identify
659 these in the *Oncopeltus* genome (see also Supplemental Notes 5.2.b, 5.3.b for these
660 and following metamorphosis gene details). From the ecdysone response cascade
661 defined in *Drosophila* [85], we identified *Oncopeltus* orthologs of both early and late-
662 acting factors. It will be interesting to see if the same regulatory relationships are
663 conserved in the context of hemimetabolous molting in *Oncopeltus*. For example,
664 *E75A* is required for reactivation of ecdysteroid production during the molt cycle in
665 *Drosophila* larvae [86] and likely operates similarly in *Oncopeltus*, since *Of-E75A*
666 RNAi prevents fourth-instar nymphs from molting to the fifth instar (H. Kelstrup and
667 L. Riddiford, unpublished data). In holometabolous insects, a declining titer of
668 ecdysteroids leads to the release of a series of neuropeptides that ultimately causes the
669 insect to molt, or ecdyse [87, 88]. Orthologs of these hormones and their receptors
670 are also present in the *Oncopeltus* genome or transcriptomic data.
671
672 In hemipterans, activation of juvenile hormone (JH) signaling at molts
673 determines whether the insect progresses to another nymphal instar or, if lacking,
674 becomes an adult [47]. We were able to identify many components of the JH signal
675 transduction pathway in the *Oncopeltus* genome, including an ortholog of
676 *Methoprene-tolerant* (*Met*), the JH receptor [47, 89], and the JH-response gene *Kr-h1*
677 [47, 90, 91]. JH acts to determine cuticle identity through regulation of the *broad*
678 gene in a wide variety of insects, where different isoforms direct specific aspects of
679 metamorphosis in *Drosophila* [92, 93]. In *Oncopeltus*, *broad* expression directs
680 progression through each of the nymphal stages [94], but the effect of each isoform
681 was unknown. We identified three isoforms in *Oncopeltus* – Z2, Z3, and Z4 – and
682 performed isoform-specific RNAi. In contrast to *Drosophila*, Broad isoform
683 functions appear to be more redundant in *Oncopeltus*, as knockdown of isoforms Z2
684 and Z3 have similar effects on survival to adulthood as well as adult wing size and
685 morphology (Fig. 9).
686
687 Regulators such as Broad initiate the transcription of a large battery of genes
688 that encode the structural components of the cuticle needed at each molt, consistent
689 with our expression analyses (Fig. 2b, discussed above). We identified 173 genes
690 encoding putative cuticle structural proteins in the milkweed bug, using established

691 sequence motifs (see also Supplemental Note 5.2.c). Similar to other insects, the CPR
692 family, with the RR-1 (soft cuticle), RR-2 (hard cuticle), and unclassifiable types,
693 constituted the largest of the cuticle protein gene groups. While several protein
694 families are similar in size to those of other insects (CPAP1, CPAP3, and TWDL: see
695 also Table S5.12), we found a slight expansion in the *Oncopeltus* CPF family (see
696 also Fig. S5.14). For cuticle production, similar to the bed bug and the Asian
697 longhorned beetle [11, 29], we identified a single *chitin synthase* gene with conserved
698 alternative splice isoforms, which suggests that *chitin synthase 2* is a duplication
699 specific to only certain beetle and fly lineages within the Holometabola [95].
700
701 One of the major characteristics of the milkweed bug is the distinctive red-
702 orange and black aposematic (warning) coloration within the cuticle and epidermis
703 that has been shown to act as a deterrent to predators (*e.g.*, Figs. 1, 9, [19, 20]). For
704 black coloration, we were able to identify the key melanin synthesis enzymes (see
705 also Fig. S5.15). The melanin synthesis pathway is conserved across holometabolous
706 insects (*e.g.*, [96, 97]), and recent work in *Oncopeltus* [98, 99] supports functional
707 conservation of melanin production in hemimetabolous lineages as well. In contrast,
708 production of the primary warning coloration produced by the pteridine red
709 erythropterin [100] remains an open avenue for hemimetabolous research. This
710 pigment, along with other pterins, is synthesized from GTP through a series of
711 enzymatic reactions [101]. The genes encoding enzymes that convert GTP into
712 pterins have not been as extensively studied as melanins, and thus far in *Oncopeltus*
713 we were only able to identify orthologs of *punch*, which encodes a GTP
714 cyclohydrolase [102], and *sepia*, which is required for the synthesis of the red eye
715 pigment drosopterin [103]. The bright red color of *Oncopeltus* eggs may in part
716 reflect chemical protection transmitted parentally [104]. Thus, further identification
717 of pigmentation genes will provide fitness indicators for maternal contributions to
718 developmental success under natural conditions (*i.e.*, the presence of egg predators).
719
720
721 **Chemoreception and metabolism in relation to feeding biology**
722 The aposematic pigmentation of the milkweed bug advertises the fact that toxins in its
723 milkweed seed diet are incorporated into the bugs themselves, a metabolic feat that
724 was independently acquired in *Oncopeltus* and in the similarly colored monarch
725 butterfly (*Danaus plexippus*), which shares this food source [35, 105]. Moreover,
726 given the fundamental differences in metabolic pathways between phytophagous,
727 mucivorous, and hematophagous species, we investigated to what extent differences
728 in feeding ecology across hemipterans are represented in the chemoreceptor and
729 metabolic enzyme repertoires of these species.
730
731 The ability of insects to smell and taste the enormous diversity of chemicals
732 important to them for locating and identifying food, mates, oviposition sites, and other
733 aspects of their environment is primarily mediated by three large gene families. The
734 closely related Odorant Receptor (OR) and Gustatory Receptor (GR) families, and the
735 distinct Ionotropic Receptor (IR) family [106-109], commonly encode tens to
736 hundreds of chemoreceptors in arthropods. Consistent with having a less derived
737 feeding ecology than species with phloem-restricted or obligate hematophagous diets,
738 *Oncopeltus* retains a moderate complement of chemoreceptors from the different
739 classes (Table 2, see also Supplemental Note 5.3.f). The hematophagous *Cimex* and
740 *Rhodnius* have relatively depauperate OR and GR families compared to *Oncopeltus*.

741 While a few conserved orthologs such as the OrCo protein and a fructose receptor are
742 found across species, *Oncopeltus* and *Acyrthosiphon* retain a set of sugar receptors, a
743 gene lineage lost independently from the blood-feeding bugs (*Rhodnius* [10], *Cimex*
744 [11]) and body louse (*Pediculus* [110]). Conversely, *Oncopeltus* has, like *Cimex*, a
745 set of candidate carbon dioxide receptors, a gene lineage lost from *Rhodnius*,
746 *Acyrthosiphon*, and *Pediculus* [10, 11, 111], but which is similar to a GR subfamily
747 expansion in the more distantly related hemimetabolous termite (Isoptera, [112]).
748 Comparable numbers of IRs occur across the heteropterans, where in addition to a
749 conserved set of orthologs primarily involved in sensing temperature and certain acids
750 and amines, *Oncopeltus* has a minor expansion of IRs distantly related to those
751 involved in taste in *Drosophila*. The major expansions in each insect lineage are the
752 candidate "bitter" GRs ([113], see also Supplemental Note 5.3.f, Fig. S5.19). In
753 summary, *Oncopeltus* exhibits moderate expansion of specific subfamilies likely to be
754 involved in host plant recognition, consistent with it being a preferentially specialist
755 feeder with a potentially patchy food source [20, 114].
756
757
758
759 **Table 2. Numbers of chemoreceptor genes/proteins per family in selected insect**
760 **species.** In some cases the number of proteins is higher than the number of genes due
761 to an unusual form of alternative splicing, which is particularly notable for the
762 *Oncopeltus* GRs. Data are shown for four Hemiptera as well as *Drosophila*
763 *melanogaster*, the body louse *Pediculus humanus*, and the termite *Zootermopsis*
764 *nevadensis* [10, 11, 108, 110-112, 115].
765

Species	Odorant	Gustatory	Ionotropic
Oncopeltus fasciatus [1]	120/121	115/169	37/37
Cimex lectularius [1,2]	48/49	24/36	30/30
Rhodnius prolixus [1,2]	116/116	28/30	33/33
Acyrthosiphon pisum [3]	79/79	77/77	19/19
Pediculus humanus [2]	12/13	6/8	14/14
Zootermopsis nevadensis	70/70	87/90	150/150
Drosophila melanogaster	60/62	60/68	65/65

766
767 [1] Hemiptera: Heteroptera
768 [2] independent acquisitions of hematophagy [16]
769 [3] Hemiptera, phloem-feeding
770
771
772
773 As host plant recognition is only the first step, we further explored whether
774 novel features of the *Oncopeltus* gene set may be directly associated with its diet. We
775 therefore used the CycADS annotation pipeline [116] to reconstruct the *Oncopeltus*
776 metabolic network. The resulting BioCyc metabolism database for *Oncopeltus*
777 ("OncfaCyc") was then compared with those for 26 other insect species in the current
778 ArthropodaCyc collection ([117], http://arthropodacyc.cycadsys.org/), including three
779 other hemipterans: the pea aphid, the green peach aphid, and the kissing bug (Tables
780 3-4). For a global metabolism analysis, we detected the presence of 1085 Enzyme
781 Commission (EC) annotated reactions with at least one protein in the *Oncopeltus*
782 genome (see also Supplemental Note 6.4, Table S6.9). Among these, 10 enzyme

783 classes (represented by 17 genes) are unique and 17 are missing when compared to
784 the other insects (Table 4, Table S6.10).
785
786
787
788 **Table 3. Hemipteran ArthropodaCyc database summaries.**
789 Overview statistics for the newly created database for *Oncopeltus fasciatus* (Ofas) in
790 comparison with public databases for *Rhodnius prolixus* (Rpro), *Acyrthosiphon pisum*
791 (Apis), and *Myzus persicae* (Mper) available from [117]. Based on OGS v1.1.
792

Species ID	*Ofas*	*Rpro*	*Apis*	*Mper*	*Mper*
Gene set ID	OGS v1.1	RproC1.1 (Built on RproC1 assembly)	OGS v2.1b (Built on Acyr_2.0 assembly)	Clone G006 v1.0	Clone O v1.0
CycADS Database ID	OncfaCyc	RhoprCyc	AcypiCyc v2.1b	Myzpe_G006 Cyc	Myzpe_O Cyc
Total mRNA[1]	19,673	15,437	36,195	24,814	24,770
Pathways	294	312	307	319	306
Enzymatic Reactions	2,192	2,366	2,339	2,384	2,354
Polypeptides	19,820	15,471	36,228	24,849	24,805
Enzymes	3,050	2,660	5,087	4,646	4,453
Compounds	1,506	1,665	1,637	1,603	1,655

793
794 [1] In the BioCyc databases all splice variants are counted in the summary tables for
795 genes.
796
797
798
799
800 **Table 4. Hemipteran ArthropodaCyc annotations of metabolic genes.**
801 Taxonomic abbreviations are as in Table 3.
802

	Ofas	*Rpro*	*Apis*	*Mper*
		Global metabolism		
EC[1] present in the genome	1085	1241	1288	1222
EC unique to this genome[2]	10	13	23	5
EC missing only in this genome[2]	17[4]	8	2	6
		Amino acid metabolism (KEGG)		
EC present in the genome	169	188	195	185
EC unique to this genome[2]	2	1	6	1
EC missing only in this genome[2]	5	2	0	2
EC unique to this genome[3]	8	10	12	8
EC missing only in this genome[3]	14	5	0	2

803
804 [1] "EC" refers to the number of proteins, as represented by their unique numerical
805 designations within the Enzyme Commission (EC) classification system for enzymes
806 and their catalytic reactions.
807 [2] in comparison to all other insects from ArthropodaCyc
808 [3] in comparison among the four hemipterans
809 [4] includes three EC categories added in OGS v1.2 (see also Table S6.10)
810

We then looked specifically at amino acid metabolism in four hemipterans representing the three different diets (see also Table S6.11). Among eight EC annotated enzymes present in the milkweed bug genome but not the other three, we identified the arginase (E.C. 3.5.3.1) that degrades arginine (Arg) into urea and ornithine, a precursor of proline (Pro). Given this difference, we extended our analysis to the entire urea cycle (Fig. 10a). Across all 26 insects present in the database, we identified three distinct groups (see also Table S6.12): (i) *Oncopeltus* and six other non-hemipteran insects that are able to degrade Arg but cannot synthesize it (Fig. 10b); (ii) the other three hemipterans that uniquely can neither synthesize nor degrade Arg via this cycle (Fig. 10c); and (iii) the other 17 insects that, with some minor differences, have an almost complete cycle (Fig. 10d). This suggests that loss of the ability to synthesize Arg may already have occurred at the base of the Hemiptera, with subsequent, independent loss of Arg degradation capacity in the aphid and *Rhodnius* lineages. Retention of Arg degradation in *Oncopeltus* might be linked to the milkweed seed food source, as most seeds are very rich in Arg [118], and Arg is indeed among the metabolites detected in *Oncopeltus* [119]. However, the monarch butterfly is one of only a handful of species that retains the complete Arg pathway (Fig. 10d: blue text). Despite a shared food source, these species may therefore differ in their overall Arg requirements, or – in light of a possible group benefit of *Oncopeltus* aggregation during feeding [20] – in their efficiency of Arg uptake.

Other enzymes are also present only in the milkweed bug in comparison with the other hemipterans (see also Table S6.11). As would be expected, *Oncopeltus*, like other insects [117], has the ability to degrade tyrosine (Tyr), a pathway that was uniquely lost in the aphids. Given the variable yields of Tyr from a mucivorous diet [120], this amino acid needed for cuticle maturation (sclerotization) is jointly synthesized – and consumed – by the aphid host and its endosymbiotic bacteria [5, 6, 15, 121]. Meanwhile, we find support for the recent nature of milkweed bug lineage-specific duplications that led to three copies of the Na+/K+ ATPase alpha subunits whose amino acid substitutions confer increased resistance to milkweed cardenolides [35, 122]. In the *Oncopeltus* genome, the genes encoding subunits ATPα1B and ATPα1C occur as a tandem duplication, notably on a scaffold that also harbors one of the clustered ZF271-like gene expansions (see above).

CONCLUSIONS

The integrated genomic and transcriptomic resources presented here for the milkweed bug *Oncopeltus fasciatus* (Figs. 2,5) underpin new insights into molecular evolution and suites of related biological characters within the Hemiptera. The gene structure trends we identified, with lineage predominating over genome size as a predictor and with many intron gains in the hemipteroid lineage (Figs. 6,7), offer initial parameters and hypotheses for the Hemiptera, Coleoptera, and Diptera. Such ordinal-level parameters can be evaluated against new species' data and also inform customized pipelines for future automated gene model predictions. At the same time, it will be interesting to explore the ramifications of hemipteroid intron gains. For example, while possessing more, small exons brings an increased transcriptional cost, it may

also provide greater scope to generate protein modularity via isoforms based on alternative exon usage. Furthermore, with the larger genome sizes and lower gene densities of hemipteroids compare to the well-studied Hymenoptera, it will also be interesting to see whether and in which direction hemipteroid gene and intron size may correlate with recombination rates [123].

Our analyses also highlight new directions for future experimental research, building on *Oncopeltus*'s long-standing history as a laboratory model and its active research community in the modern molecular genetics era (*e.g.*, Fig. 9, [24-26]). Functional testing will clarify the roles of genes we have identified as unique to the Hemiptera, including those implicated in chemical protection, bacterial and plant cell wall metabolism, or encoding wholly novel proteins (Figs. 3,8, see also Supplemental Note 2.2). Meanwhile, the prominent and species-specific expansions specifically of ZF271-like zinc fingers (Fig. 4), combined with the absence of the co-repressor KAP-1 in insects, argues for investigation into alternative possible interaction partners, which could clarify the nature of these zinc fingers' regulatory role and their binding targets.

One key output of this study is the generation of a metabolism database for *Oncopeltus*, contributing to the ArthropodaCyc collection (Table 3). In addition to comparisons with other species (Fig. 10), this database can also serve as a future reference for studies that use *Oncopeltus* as an ecotoxicology model species (*e.g.*, [124]). While we have primarily focused on feeding ecology in terms of broad comparisons between phytophagy and fluid feeding, *Oncopeltus* is also poised to support future work on nuances among phytophagous species. Despite its milkweed diet in the wild, the lab strain of *Oncopeltus* has long been adapted to feed on sunflower seeds, demonstrating a latent capacity for more generalist phytophagy [114]. This potential may also be reflected in a larger gustatory receptor repertoire than would be expected for an obligate specialist feeder (Table 2). Thus, *Oncopeltus* can serve as a reference species for promiscuously phytophagous pest species such as the stink bug. Finally, given that we have identified a number of key genes implicated in life history trade-offs, the genome data represent an important tool to explore the proximate mechanisms of fundamental aspects of life history evolution in an organism in which the ultimate explanations for traits such as cardenolide tolerance, pigmentation, and plasticity in reproduction under environmental variation have been elucidated in both the laboratory and nature.

METHODS

(More information is available in the supplementary materials, Additional file 1.)

Milkweed bug strain, rearing, and DNA/RNA extraction

The milkweed bug *Oncopeltus fasciatus* (Dallas), Carolina Biological Supply strain (Burlington, North Carolina, USA), was maintained in a laboratory colony under standard husbandry conditions (sunflower seed and water diet, 25 ºC, 12:12 light-dark photoperiod). Voucher specimens for an adult female (record # ZFMK-TIS-26324) and adult male (record # ZFMK-TIS-26325) have been preserved in ethanol and deposited in the Biobank of the Centre for Molecular Biodiversity Research, Zoological Research Museum Alexander Koenig, Bonn, Germany (https://www.zfmk.de/en/biobank).

Genomic DNA was isolated from individual, dissected adults using the Blood & Cell Culture DNA Midi Kit (G/100) (Qiagen Inc., Valencia, California, USA). Total RNA was isolated from individual, dissected adults and from pooled, mixed-instar nymphs with TRIzol Reagent (Invitrogen/ Thermo Fisher Scientific, Waltham, Massachusetts, USA). Dissection improved accessibility of muscle tissue by disrupting the exoskeleton, and gut material was removed.

Genome size calculations (flow cytometry, *k*-mer estimation)

Genome size estimations were obtained by flow cytometry with Hare and Johnston's protocol [125]. Four to five females and males each from the Carolina Biological Supply lab strain and a wild strain (collected from Athens, Georgia, USA; GPS coordinates: 33° 56' 52.8216" N, 83° 22' 38.3484'' W) were measured (see also Supplemental Note 2.1.a). At the bioinformatic level, we attempted to estimate genome size by *k*-mer spectrum distribution analysis for a range of *k*=15 to 34 counted with Jellyfish 2.1.4 [126] and bbmap [127], graphing these counts against the frequency of occurrence of *k*-mers (depth), and calculating genome size based on the coverage at the peak of the distribution (see also Supplemental Note 2.1.b).

Genome sequencing, assembly, annotation, and official gene set overview

Library preparation, sequencing, assembly, and automatic gene annotation were conducted at the Baylor College of Medicine Human Genome Sequencing Center (as in [11, 29]). About 1.1 billion 100-bp paired-end reads generated on an Illumina HiSeq2000s machine were assembled using ALLPATHS-LG [128], from two paired-end (PE) and two mate pair (MP) libraries specifically designed for this algorithm (see also Supplemental Note 1). Three libraries were sequenced from an individual adult male (180- and 500-bp PE, 3-kb MP), with the fourth from an individual adult female (8-10-kb MP). The final assembly (see metrics in Table 1) has been deposited in GenBank (accession GCA_000696205.1).

Automated annotation of protein-coding genes was performed using a Maker 2.0 annotation pipeline [129] tuned specifically for arthropods (see also Supplemental Note 3). These gene predictions were used as the starting point for manual curation via the Apollo v.1.0.4 web browser interface [130], and automatic and manual curations were compiled to generate the OGS (see also Supplemental Note 4). Databases of the genome assembly, Maker automatic gene predictions, and OGS v1.1 are available through the i5K Workspace@NAL [131], and the Ag Data Commons data access system of the United States Department of Agriculture's (USDA) National

950 Agricultural Library as individual citable databases [132-134]. The current version of
951 the gene set, OGS v1.2, will be deposited in NCBI under accession number XXX.
952
953 **Repeat content analysis**
954 Repetitive regions were identified in the *Oncopeltus* genome assembly with
955 RepeatModeler Open-1.0.8 [135] based on a species-specific repeat library generated
956 *de novo* with RECON [136], RepeatScout [137], and Tandem Repeats Finder [138].
957 Then, RepeatMasker Open-4.0 [139] was used to mask repeat sequences based on the
958 RepeatModeler library. Given the fragmented nature of the assembly, we attempted
959 to fill and close assembly gaps by sequencing additional material, generating long
960 reads with single molecule real time sequencing on a PacBio RS II machine (34
961 SMRT cells to an expected coverage of 8x, see also Supplemental Note 2.3). Gap
962 filling on the Illumina assembly scaffolds was performed with PBJelly version
963 13.10.22, and the resulting assembly was used for repeat content estimation and
964 comparison with *Cimex lectularius* and *Acyrthosiphon pisum.*
965
966 **Transcriptome resources**
967 Total RNA from three distinct life history samples (pooled, mixed-instar nymphs; an
968 adult male; an adult female) was also sequenced on an Illumina HiSeq2000s machine,
969 producing a total of 72 million 100-bp paired-end reads (see also Supplemental Note
970 1.3, Table S1.1). These expression data were used to support the generation of the
971 OGS at different stages of the project: as input for the evidence-guided automated
972 annotation with Maker 2.0 (see also Supplemental Note 3), as expression evidence
973 tracks in the Apollo browser to support the community curation of the OGS, and,
974 once assembled into a *de novo* transcriptome, as a point of comparison for quality
975 control of the OGS.
976
977 The raw RNA-seq reads were pre-processed by filtering out low quality bases
978 (phred score <30) and Truseq adapters with Trimmomatic-0.30. Further filtering
979 removed ribosomal and mitochondrial RNA sequences with Bowtie 2 [140], based on
980 a custom library built with all hemipteran ribosomal and mitochondrial RNA
981 accessions from NCBI as of 7th February 2014 (6,069 accessions). The pooled,
982 filtered reads were mapped to the genome assembly with Tophat2-PE on CyVerse
983 [141]. A second set of RNA-seq reads from an earlier study ("published adult"
984 dataset, [35]) was also filtered and mapped in the same fashion, and both datasets
985 were loaded into the *Oncopeltus* Apollo instance as evidence tracks (under the track
986 names "pooled RNA-seq - cleaned reads" and "RNA-seq raw PE reads Andolfatto et
987 al", respectively).
988
989 Additionally, a *de novo* transcriptome was generated from our filtered RNA-
990 seq reads (pooled from all three samples prepared in this study) using Trinity [142]
991 and TransDecoder [143] with default parameters. This transcriptome is referred to as
992 "i5K", to distinguish it from a previously published maternal and early embryonic
993 transcriptome for *Oncopeltus* (referred to as "454", [34]). Both the i5K and 454
994 transcriptomes were mapped to the genome assembly with GMAP v. 2014-05-15 on
995 CyVerse. These datasets were also loaded into the Apollo browser as evidence tracks
996 to assist in manual curation.
997

998 **Life stage specific expression analyses**
999 Transcript expression of the OGS v1.1 genes was estimated by running RSEM2 [144]
1000 on the filtered RNA-seq datasets for the three postembryonic stages against the OGS
1001 v1.1 cDNA dataset. Transcript expression was then based on the transcripts per
1002 million (TPM) value. The TPM values were processed by adding a value of 1 (to
1003 avoid zeros) and then performing a log2-transformation. The number of expressed
1004 genes per RNA-seq library was compared for TPM cutoffs of >1, >0.5, and >0.25.
1005 For this first-pass expression assessment, a >0.25 cutoff was chosen, which reduced
1006 the number of expressed genes by 6.6% compared to the first analysis, while the other
1007 TPM cutoffs were deemed too restrictive (reducing the expressed gene set by 10.3%
1008 and 16.6%, respectively, with the >0.5 and >1 cutoffs). This analysis was also
1009 applied to the "published adult" dataset [35]. To include embryonic stages in the
1010 comparison, transcripts from the 454 transcriptome were used as blastn queries
1011 against the OGS v1.1 cDNA dataset (cutoff e-value $<10^{-5}$). The results from all
1012 datasets were converted to binary format to generate Venn diagrams (Fig. 2b).
1013
1014 **Protein gene orthology assessments via OrthoDB and BUSCO analyses**
1015 These analyses follow previously described approaches [41, 42]. See Supplemental
1016 Note 6.1 for further details.
1017
1018 **Global transcription factor identification**
1019 Likely transcription factors (TFs) were identified by scanning the amino acid
1020 sequences of predicted protein-coding genes for putative DNA binding domains
1021 (DBDs), and when possible, the DNA binding specificity of each TF was predicted
1022 using established procedures [56]. Briefly, all protein sequences were scanned for
1023 putative DBDs using the 81 Pfam [145] models listed in Weirauch and Hughes [146]
1024 and the HMMER tool [147], with the recommended detection thresholds of Per-
1025 sequence Eval < 0.01 and Per-domain conditional Eval < 0.01. Each protein was
1026 classified into a family based on its DBDs and their order in the protein sequence
1027 (e.g., bZIPx1, AP2x2, Homeodomain+Pou). The resulting DBD amino acid
1028 sequences were then aligned within each family using Clustal Omega [148], with
1029 default settings. For protein pairs with multiple DBDs, each DBD was aligned
1030 separately. From these alignments, the sequence identity was calculated for all DBD
1031 sequence pairs (*i.e.*, the percent of amino acid residues that are identical across all
1032 positions in the alignment). Using previously established sequence identity thresholds
1033 for each family [56], the predicted DNA binding specificities were mapped by simple
1034 transfer. For example, the DBD of OFAS001246-RA is 98% identical to the
1035 *Drosophila melanogaster* Bric a Brac 1 (Bab1) protein. Since the DNA binding
1036 specificity of Bab1 has already been experimentally determined, and the cutoff for the
1037 Pipsqueak family TFs is 85%, we can infer that OFAS001246-RA will have the same
1038 binding specificity as *Drosophila* Bab1.
1039
1040 **RNA interference**
1041 Double-stranded RNA (dsRNA) was designed to target the final, unique exon of the
1042 *broad* isoforms *Z2, Z3*, and *Z4*. A portion of the coding sequence for the zinc finger
1043 region from these exons (179 bp, 206 bp, and 216 bp, respectively) was cloned into a
1044 plasmid vector and used as template for *in vitro* RNA synthesis, using the gene-
1045 specific primer pairs: Of-Z2_fwd: 5′-ATGTGGCAGACAAGCATGCT-3′; Of-
1046 Z2_rev: 5′-CTAAAATTTGACATCAGTAGGC-3′; Of-Z3_fwd: 5′-
1047 ccttctcctgttactactcac-3′; Of-Z3_rev: 5′-ttatatgggcggctgtccaa-3′; Of-Z4_fwd: 5′-

1048 AACACTGACCTTGGTTACACA-3′; Of-Z4_rev: 5′-
1049 TAGGTGGAGGATTGCTAAAATT-3′. Two separate transcription reactions (one
1050 for each strand) were performed using the Ambion MEGAscript kit (Ambion, Austin,
1051 Texas, USA). The reactions were purified by phenol/chloroform extraction followed
1052 by precipitation as described in the MEGAscript protocol. The separate strands were
1053 re-annealed in a thermocycler as described previously [31]. Nymphs were injected
1054 with a Hamilton syringe fitted with a 32-gauge needle as described [53]. The
1055 concentration of *Of-Z2*, *Of-Z3* and *Of-Z4* dsRNA was 740 ng/µl, 1400 ng/µl, and
1056 1200 ng/µl, respectively. All nymphs were injected within 8 hours of the molt to the
1057 fourth (penultimate juvenile) instar (n ≥12 per treatment: see Fig. 9). Fore- and
1058 hindwings were then dissected from adults and photographed at the same scale as
1059 wings from wild type, uninjected controls.
1060
1061 **CycADS annotation and OncfaCyc database generation**
1062 We used the Cyc Annotation Database System (CycADS, [116]), an automated
1063 annotation management system, to integrate protein annotations from different
1064 sources into a Cyc metabolic networks reconstruction that was integrated into the
1065 ArthropodaCyc database. Using our CycADS pipeline, *Oncopeltus fasciatus* proteins
1066 from the official gene set OGS v1.1 were annotated using different methods –
1067 including KAAS [149], PRIAM [150], Blast2GO [151, 152], and InterProScan with
1068 several approaches [153] – to obtain EC and GO numbers. All annotation
1069 information data were collected in the CycADS SQL database and automatically
1070 extracted to generate appropriate input files to build or update BioCyc databases [154]
1071 using the Pathway Tools software [155]. The OncfaCyc database, representing the
1072 metabolic protein-coding genes of *Oncopeltus*, was thus generated and is now
1073 included in the ArthropodaCyc database, a collection of arthropod metabolic network
1074 databases ([117], http://arthropodacyc.cycadsys.org/).
1075
1076
1077 **FIGURE LEGENDS**
1078
1079 **Fig. 1. The large milkweed bug, *Oncopeltus fasciatus*, shown in its phylogenetic**
1080 **and environmental context.**
1081 **(a)** Species tree of selected Hemiptera with genomic and transcriptomic resources,
1082 based on phylogenetic analyses and divergence time estimates in [3]. Species marked
1083 with an asterisk (*) have published resources; those with the appellation "i5K" are
1084 part of a current pilot project supported by the Baylor College of Medicine Human
1085 Genome Sequencing Center and the National Agricultural Library of the USDA.
1086 Note that recent analyses suggest the traditional infraorder Cimicomorpha, to which
1087 *Rhodnius* and *Cimex* belong, may be paraphyletic [16].
1088 **(b-c)** Milkweed bugs on their native food source, the milkweed plant: gregarious
1089 nymphs of different instars on a milkweed seed pod (b), and pale, recently eclosed
1090 adults and their shed exuvia (c). Images were taken at Avalon Park and Preserve,
1091 Stony Brook, New York, USA, courtesy of Deniz Erezyilmaz, used with permission.
1092 **(d)** Individual bugs, shown from left to right: first instar nymphs (ventral and dorsal
1093 views) and adults (dorsal and lateral views); images courtesy of Kristen Panfilio
1094 (nymphs) and Jena Johnson (adults), used with permission. The arrow labels the
1095 labium (the "straw"), part of the hemipteran mouthpart anatomy adapted for feeding
1096 by piercing and sucking.
1097

1098 **Fig. 2. Comparisons of the official gene set and transcriptomic resources.**
1099 **(a)** Area-proportional Venn diagram comparing the OGS v1.1 ("OGS"), a Trinity *de*
1100 *novo* transcriptome from the three post-embryonic RNA-seq samples ("i5K"), and the
1101 maternal and embryonic transcriptome from 454 data ("454" [34]). Sample sizes and
1102 the fraction of each transcriptome represented in the OGS are indicated (for the 454
1103 dataset, only transcripts with homology identification were considered). The unique
1104 fraction of each set is also specified (%). Dataset overlaps were determined by blastn
1105 (best hit only, e-value $<10^{-9}$).
1106 **(b)** Four-set Venn diagram representation of OGS v1.1 gene model expression across
1107 four different life history samples. Values are counts of gene models, with
1108 percentages also given for the largest subsets. Note that the "Embryo/Maternal"
1109 sample derives from 454 pyrosequencing data and therefore has a smaller data
1110 volume than the other samples, which were generated with Illumina sequencing.
1111
1112 **Fig. 3. Orthology comparisons and phylogenetic placement of *Oncopeltus***
1113 ***fasciatus* among other Arthropoda.**
1114 **(a)** Comparisons of protein-coding gene content in 12 arthropod species, with the
1115 Hemiptera highlighted in red text. The bar chart shows the number of proteins per
1116 conservation level (see legend), based on OrthoDB orthology clustering analyses. To
1117 the left is a maximum likelihood phylogeny based on concatenation of 395 single-
1118 copy orthologs (all nodes have 100% support unless otherwise noted; branch length
1119 unit is substitutions per site). The inset pie chart shows the proportion of proteins per
1120 conservation level in *Oncopeltus* (*"Ofas"*). See also Supplemental Note 6.1.
1121 **(b)** Proportion of *Oncopeltus* proteins that have expression and/or curation validation
1122 support per conservation level (same color legend as in (a)). Expression support is
1123 based on the life history stage data in Fig. 2b.
1124 **(c)** Protein orthology data evaluated by taxonomic grouping, based on the OrthoDB
1125 v8 i5K "Insecta" analysis with 64 species, a subset of the recently released OrthoDB
1126 v9 (http://www.orthodb.org): Hemiptera (red, 8 species); other hemimetabolous
1127 species (paraphyletic, yellow, 6 species); Holometabola (purple, 50 species). Values
1128 are given for both orthogroups (black text, defined in [42]) and protein-coding genes
1129 (blue text). As analyses that require all species to be represented in a given
1130 orthogroup are limited by the quality of every species' OGS, the cutoff for orthogroup
1131 presence in a given Venn diagram set was rather that roughly half of all relevant
1132 species are included, and strictly no species from a different set are permitted. Set [A]
1133 contains ≥10 hemimetabolous species (allowing for 4-8 Hemiptera and 2-6 other
1134 hemimetabolous species). Set [B] contains ≥4 Hemiptera and ≥25 Holometabola. Set
1135 [C] contains ≥2 other hemimetabolous species and for the Holometabola at least one
1136 representative from each of the Hymenoptera, Coleoptera, Lepidoptera, and Diptera.
1137 Analyses based on OGS v1.1.
1138
1139 **Fig. 4. Distribution of transcription factor families across insect genomes.**
1140 **(a)** Heatmap depicting the abundance of transcription factor (TF) families across 17
1141 insect genomes (Hemiptera highlighted in red text), with *Daphnia* as an outgroup.
1142 Each entry indicates the number of TF genes for the given family in the given
1143 genome, based on the presence of predicted DNA binding domains (see Methods).
1144 The color key has a log (base 2) scale (light blue means the TF family is completely
1145 absent). Values are in Supplementary Table S6.2.
1146 **(b)** Bar graph showing the number of proteins of each of the two most abundant TF
1147 families, homeodomains and C2H2 zinc fingers (ZFs), per species. Solid lines

1148 demarcate insect orders: Hemiptera (Hemipt.), Hymenoptera (Hym.), Coleoptera
1149 (Col.), and Diptera (Dipt.). The dashed line demarcates the dipteran family Culicidae
1150 (mosquitoes).
1151 **(c)** Proportions of *Oncopeltus* homeodomain (HD) and C2H2 zinc finger proteins
1152 with orthology assignment (predicted DNA binding specificity) and/or manual
1153 curation. "Classified" refers to automated classification of a protein to a TF family,
1154 but without a specific orthology assignment.
1155 **(d)** Maximum likelihood phylogeny of representative subsets of the zinc finger 271-
1156 like family in *Oncopeltus* (49 proteins, blue text) and protein accessions retrieved
1157 from GenBank for the pea aphid (55 proteins, black text) as well as chelicerate (red
1158 text) and holometabolan (yellow text) outgroups (16 proteins, 7 species). Gaps were
1159 removed during sequence alignment curation, with default pipeline settings [156]. All
1160 nodes have ≥50% support. Key nodes are circled for the distinct clades containing all
1161 aphid or all *Oncopeltus* proteins (82% support each), and for each 'core' clade
1162 comprised exclusively of proteins from each species (97% and 100%, respectively;
1163 triangles shown to scale for branch length and number of clade members). Branch
1164 length unit is substitutions per site.
1165 Analyses based on OGS v1.1.
1166
1167 **Fig. 5. Comparison of repeat content estimations.**
1168 **(a)** Comparison of total repetitive content among insect genomes. The three values
1169 for *Oncopeltus* are shown (in ascending order: original Illumina assembly, gap-filled
1170 assembly, Illumina-PacBio hybrid estimate). Values for the three hemipterans labeled
1171 in red text are from RepeatModeler (gold bars for the pea aphid and bed bug; blue and
1172 gold bars for *Oncopeltus*). All other values are from the respective genome papers,
1173 including a second value corresponding to the published repeat content for the first
1174 version of the aphid genome [5, 9, 112, 157-162]. Species abbreviations as in Fig. 4
1175 (compare panels a and b), and additionally: *Nlug, Nilaparvata lugens; Lmig, Locusta*
1176 *migratoria; Bmor, Bombyx mori; Aalb, Aedes albopictus*.
1177 **(b)** Comparison of repetitive element categories between three hemipteran genomes,
1178 based on results from RepeatModeler. Here we present assembly coverage as actual
1179 sequence length (Mb) to emphasize the greater repeat content in *Oncopeltus* (based on
1180 the gap-filled assembly: see also Supplemental Note 2.3).
1181
1182 **Figure 6. Trends in gene structure show hemipteroid-specific tendencies.**
1183 **(a)** Trends in protein size, exon size, and exon number are shown for a highly
1184 conserved set of genes encoding large proteins of diverse functional classes ("gold
1185 standard", curated gene set). Median values are plotted. Sample sizes are indicated
1186 for each species, with 11 genes for which orthologs were evaluated in all species.
1187 Where it was not possible to analyze all 30 genes for a given species, equal sampling
1188 was done across the range of protein sizes of the complete dataset, based on the
1189 *Cimex* ortholog sizes (1:1:1 sampling from big:medium:small subcategories of 10
1190 genes each). See also Supplemental Note 6.3.
1191 **(b)** Box plot representations of coding sequence exon size (aa) for two species from
1192 each of three insect orders, based on datasets of unique coding sequence exons (one
1193 isoform per gene) and excluding terminal exons <10 aa (as most of those exons may
1194 rather be UTRs or a small placeholder N-terminal exon, as a byproduct of the Maker
1195 gene annotation pipeline's requirement to create nominally complete protein coding
1196 genes with in-frame start codons). Only manually curated gene models were
1197 considered for *Oncopeltus* and the other recent i5K species; the entire OGS was used

1198 for *Tribolium* and *Drosophila*. For clarity, outliers are omitted; whiskers represent
1199 1.5× the value of the Q3 (upper) or Q2 (lower) quartile range. MAD, median absolute
1200 deviation.
1201 Species are represented by their four-letter taxonomic abbreviations, with their ordinal
1202 relationships given below the phylogeny in panel (a): Hemip., Hemiptera; Thys.,
1203 Thysanoptera; Col., Coleoptera; Dipt., Diptera. Species abbreviations as in Fig. 4 and
1204 additionally: *Gbue, Gerris buenoi*; *Focc, Frankliniella occidentalis*; *Agla,*
1205 *Anoplophora glabripennis; Ccap, Ceratitis capitata*.
1206
1207 **Fig 7. Splice site evolution correlates with both lineage and, independently,**
1208 **genome size.**
1209 Splice site changes are shown for *hemocytin* (blue text), *Tenascin major* (*Ten-m*,
1210 turquoise text), and *UDP-galactose 4′-epimerase* (brown text), mapped onto a species
1211 tree of eight insects. Patterns of splice site evolution were inferred based on the most
1212 parsimonious changes that could generate the given pattern within a protein sequence
1213 alignment of all orthologs (see also Supplemental Note 6.3 for complete methodology
1214 and data sources). In instances where an equal number of lineage specific gains or
1215 losses was possible, we remained agnostic and present a range for the ancestral
1216 number of splice sites indicated at the base of the tree, where the bracketed number
1217 indicates how many ancestral positions are still retained in all species. Along each
1218 lineage, subsequent changes are indicated in brackets, with the sign indicating gains
1219 (+) or losses (-). Values shown to the right are species-specific changes. Note that
1220 the values shown between the *D. melanogaster* and *T. castaneum* lineages denote
1221 changes that have occurred independently in both. Colored boxes highlight the
1222 largest sources of change, as indicated in the legend and discussed in the main text.
1223 Species are represented by their four-letter abbreviations (as defined in Fig. 6), and
1224 the estimated genome sizes are indicated parenthetically (measured size: [11, 29, 73,
1225 161, 163]; draft assembly size: GenBank Genome IDs 14741 and 17730). Divergence
1226 times are shown in gray and given in millions of years [3]. Abbreviations as in Figs.
1227 4,6, and: Col., Coleoptera; Dipt., Diptera; Hemip., Hemiptera; Hemipt., hemipteroid
1228 assemblage (including *F. occidentalis*); n.d., no data.
1229
1230 **Fig. 8. Lateral gene transfer introduction and subsequent evolution within the**
1231 **Hemiptera for mannosidase-encoding genes.**
1232 **(a)** Species tree summary of evolutionary events. Stars represent the original LGT
1233 introduction and subsequent copy number gains (see legend).
1234 **(b)** Maximum likelihood phylogeny of mannosidase proteins, including bacterial
1235 sequences identified among the best GenBank blastp hits for *Oncopeltus* and
1236 *Halyomorpha* (accession numbers as indicated, and for "Other bacteria" are:
1237 ACB22214.1, AEE17431.1, AEI12929.1, AEO43249.1, AFN74531.1, CDM56239.1,
1238 CUA67033.1, KOE98396.1, KPI24888.1, OAN41395.1, ODP26899.1, ODS11151.1,
1239 OON18663.1, PBD05534.1, SIR54690.1, WP096035621.1, YP001327394.1). All
1240 nodes have ≥50% support from 500 bootstrap replicates [164]. Triangles are shown
1241 to scale for branch length and number of clade members; branch length unit is
1242 substitutions per site.
1243 **(c)** Manually curated protein sequence alignment for the N-terminal region, showing
1244 the position of splice sites ("|" symbol), where one position is ancestral and present in
1245 all paralogs of a given species (magenta), and one position occurs in a subset of
1246 paralogs and is presumed to be younger (cyan, note the position is within the 5′ UTR
1247 in the case of *Halyomorpha*). Residues highlighted in yellow are conserved between

1248 the two species. The *Oncopeltus* paralog represented in the OGS as OFAS017153-
1249 RA is marked with an asterisk to indicate that this version of the gene model is
1250 incomplete and lacks the initial exon (gray text in the alignment). For clarity, only the
1251 final three digits of the *Halyomorpha* GenBank accessions are shown (full accessions:
1252 XP_014289XXX).
1253
1254 **Fig. 9. Isoform-specific RNAi based on new genome annotations affects the**
1255 **molting and cuticle identity gene *broad*.**
1256 **(a)** Genomic organization of the cuticle identity gene *broad*. The regions used as
1257 template to generate isoform-specific dsRNA are indicated (red asterisks: the final,
1258 unique exons of each isoform). Previous RNAi studies targeted sequence within
1259 exons 1-5 that is shared among all isoforms (dashed red box, [94]).
1260 **(b)** Knock down of the *Oncopeltus Z2* or *Z3 broad* isoforms at the onset of the
1261 penultimate instar resulted in altered nymphal survival and morphogenesis that was
1262 reflected in the size and proportion of the fore and hind wings at the adult stage
1263 (upper and lower images, respectively, shown to the same scale for all wings). We
1264 did not detect any effect on the wing phenotype when targeting the *Z4*-specific exon,
1265 demonstrating the specificity of the zinc finger coding region targeted by RNAi.
1266 Experimental statistics are provided in the figure inset, including for the buffer-
1267 injected negative control.
1268
1269 **Fig. 10. Comparison of the urea cycle of *Oncopeltus* with 26 other insect species.**
1270 **(a)** Detailed diagram of the urea cycle (adapted from KEGG).
1271 **(b)** Group of 7 species, including *Oncopeltus*, for which Arg degradation via arginase
1272 (3.5.3.1), but not synthesis, is possible.
1273 **(c)** Group of 3 species for which neither the degradation nor synthesis of arginine via
1274 the urea cycle is possible (all 3 other hemipterans in this analysis).
1275 **(d)** Group of 17 species sharing a complete (or almost complete) urea cycle.
1276 Hemiptera are identified in red text and the milkweed-feeding monarch butterfly is in
1277 blue text. Enzyme names corresponding to EC numbers: 1.5.1.2 = pyrroline-5-
1278 carboxylate reductase; 1.14.13.39 = nitric-oxide synthase; 2.1.3.3 = ornithine
1279 carbamoyltransferase; 2.6.1.13 = ornithine aminotransferase; 3.5.3.1 = arginase;
1280 4.3.2.1 = argininosuccinate lyase; 6.3.4.5 = argininosuccinate synthase.
1281 Analyses based on OGS v1.1.
1282
1283
1284 **TABLES (see above within relevant manuscript sections)**
1285
1286 **Table 1. *Oncopeltus fasciatus* genome metrics.**
1287
1288 **Table 2. Numbers of chemoreceptor genes/proteins per family in selected insect**
1289 **species.**
1290
1291 **Table 3. Hemipteran ArthropodaCyc database summaries.**
1292
1293 **Table 4. Hemipteran ArthropodaCyc annotations of metabolic genes.**
1294

DECLARATIONS

ADDITIONAL FILES

Additional file 1: Supplementary figures, tables, methods, and other text. (PDF)
Additional file 2: Large supporting tables. (XLSX)
Additional file 3: Chemoreceptor sequences in FASTA format. (TXT)

ACKNOWLEDGEMENTS

We thank Dorith Rotenberg (Kansas State University, currently North Carolina State University, USA), Abderrahman Khila on behalf of the Water Strider Genome Consortium (Institute of Functional Genomics and École Normale Supérieure de Lyon, France), and Michael Sparks (Agricultural Research Service, United States Department of Agriculture, USA) for generously making available the unpublished genome assemblies of the fellow hemipteroid i5K species *Frankliniella occidentalis*, *Gerris buenoi*, and *Halyomorpha halys*, respectively, for use in specific analyses presented here. Similarly, we thank Hans Kelstrup and Lynn Riddiford (Janelia Farm Research Campus, HHMI, USA) for sharing unpublished data on *Of-E75A* RNAi. We thank George Coupland (Max Planck Institute for Plant Breeding Research, Cologne, Germany) as well as Lisa Czaja, Kurt Steuber, and Bruno Huettel (Max Planck Genome Centre Cologne, Germany) for conducting the PacBio sequencing and providing support with data handling. We also thank Oliver Niehuis (Albert Ludwig University, Freiburg, Germany) and Alexander Klassmann (University of Cologne, Germany) for discussions on *k*-mer and gene structure analyses, respectively, Sarah Kingan (University of Rochester, USA) for assistance with LGT phylogenies, as well as Jeanne Wilbrandt (Zoologisches Forschungsmuseum Alexander Koenig, Bonn, Germany) for comments on the manuscript.

FUNDING

Funding for genome sequencing, assembly and automated annotation was provided by the National Institutes of Health (NIH) grant U54 HG003273 (NHGRI) to RAG. The i5K pilot project (https://www.hgsc.bcm.edu/arthropods) assisted in sequencing of the *Oncopeltus fasciatus* genome. We also acknowledge funding for the project from German Research Foundation (DFG) grants PA 2044/1-1 and SFB 680 project A12 to KAP. Support for specific analyses was provided by the Swiss National Science Foundation with grant 31003A_143936 to EMZ and PP00P3_170664 to RMW; the European Research Council grant ERC-CoG #616346 to AK; DFG grant SFB 680 project A1 to SiR; the National Science Foundation with grant US NSF DEB1257053 to JHW; and by NIH grants 5R01GM080203 (NIGMS) and 5R01HG004483 (NHGRI) and by the Director, Office of Science, Office of Basic Energy Sciences, U.S. Department of Energy, Contract No. DE-AC02-05CH11231 to MCMT.

AVAILABILITY OF DATA AND MATERIALS

All sequence data are publically available at the NCBI, bioproject number PRJNA229125. In addition, gene models and a browser are available at the National Agricultural Library ([132-134], https://i5k.nal.usda.gov/Oncopeltus_fasciatus).

AUTHOR'S CONTRIBUTIONS

KAP and StR conceived, managed, and coordinated the project. KAP and SK provided specimens for sequencing and performed DNA and RNA extractions. StR, SD, SLL, HC, HVD, HD, YH, JQ, SCM, DSTH, KCW, DMM, and RAG constructed

libraries and performed sequencing. StR, SCM, and DSTH performed the genome assembly and automated gene prediction. IMVJ, JSJ, and PJM analyzed genome size. IMVJ, VK, PH, and KAP contributed to repetitive content analyses. AD, RR, JHW, KAP, and SK performed bacterial scaffold detection and LGT analyses. MCMT developed Apollo software. KAP, IMVJ, MCMT, CPC, C-YL, and MFP implemented Apollo-based manual curation. KAP, IMVJ, JBB, DE, YS, HMR, DA, CGCJ, BMIV, EJD, CSB, C-CC, Y-TC, ADC, AGC, AJJC, PKD, EMD, CGE, MF, NG, TH, Y-MH, ECJ, TEJ, JWJ, AK, ML, MRL, H-LL, YL, SRP, LP, MP, PNR, RRP, SiR, LS, MES, JS, ES, JNS, OT, LT, MVDZ, SV, and AJR participated in manual curation and contributed to the Supplemental Notes. IMVJ, KAP, DSTH, M-JMC, CPC, C-YL, and MFP performed curation quality control and generated the OGS. IMVJ, KAP, and JBB generated *de novo* transcriptomes and performed life history stage expression analyses. RMW, PI, KAP, and EMZ performed orthology and phylogenomic analyses. MTW, KAP, IMVJ, PH, and BMIV performed transcription factor analyses. EJD conducted analyses of DNA methylation. KAP, PH, and RJS contributed to comparative analyses of gene structure. DE conducted the RNAi experiments. SC, PB-P, GF, and NP generated and performed comparative analyses on the OncfaCyc database. KAP, IMVJ, JBB, DE, YS, SC, HMR, and MTW wrote the manuscript. KAP, IMVJ, JBB, DE, YS, SC, HMR, MFP, RMW, PI, MTW, StR, PJM, and AK edited the manuscript. IMVJ and KAP organized the Supplementary Materials. All authors approved the final manuscript.

COMPETING INTERESTS

The authors declare that they have no competing interests.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Not applicable.

AUTHOR DETAILS

[1] Institute for Zoology: Developmental Biology, University of Cologne, Zülpicher Str. 47b, 50674 Cologne, Germany
[2] School of Life Sciences, University of Warwick, Gibbet Hill Campus, Coventry CV4 7AL, UK
[3] Department of Biological Sciences, University of Cincinnati, Cincinnati, Ohio 45221, USA
[4] Department of Biochemistry and Cell Biology and Center for Developmental Genetics, Stony Brook University, Stony Brook, New York 11794, USA
[5] Department of Biological Sciences, Wellesley College, 106 Central St., Wellesley, Massachusetts 02481, USA
[6] Univ Lyon, INSA-Lyon, INRA, BF2I, UMR0203, F-69621, Villeurbanne, France
[7] Current address: LSTM, Laboratoire des Symbioses Tropicales et Méditerranéennes, INRA, IRD, CIRAD, SupAgro, University of Montpellier, Montpellier, France
[8] Department of Entomology, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA
[9] National Agricultural Library, Beltsville, Maryland 20705, USA
[10] Department of Genetic Medicine and Development and Swiss Institute of Bioinformatics, University of Geneva, Geneva 1211, Switzerland
[11] Current address: Department of Ecology and Evolution, University of Lausanne, Lausanne 1015, Switzerland
[12] Center for Autoimmune Genomics and Etiology, Division of Biomedical Informatics, and Division of Developmental Biology, Cincinnati Children's Hospital Medical Center, Department of Pediatrics, College of Medicine, University of Cincinnati, Cincinnati, Ohio 45229, USA
[13] Human Genome Sequencing Center, Department of Human and Molecular Genetics, Baylor College of Medicine, One Baylor Plaza, Houston, Texas 77030, USA

[14] Current address: Department of Genome Sciences, University of Washington School of Medicine, Seattle, Washington 98195, USA
[15] Current address: Howard Hughes Medical Institute, University of Washington, Seattle, Washington 98195, USA
[16] Department of Biology, University of Rochester, Rochester, New York 14627, USA
[17] Institute of Biology, Leiden University, Sylviusweg 72, 2333 BE Leiden, Netherlands
[18] Max Planck Institute for Chemical Ecology, Hans-Knöll Strasse 8, 07745 Jena, Germany
[19] Department of Biochemistry and Genomics Aotearoa, University of Otago, Dunedin 9054, New Zealand
[20] School of Biology, Faculty of Biological Sciences, University of Leeds, Leeds LS2 9JT, UK
[21] Institut de Génomique Fonctionnelle de Lyon, Université de Lyon, Université Claude Bernard Lyon 1, CNRS UMR 5242, École Normale Supérieure de Lyon, 46 Allée d'Italie, 69364 Lyon, France
[22] Department of Ecology, Evolution and Behavior, The Alexander Silberman Institute of Life Sciences, The Hebrew University of Jerusalem, Edmond J. Safra Campus, Givat Ram 91904, Jerusalem, Israel
[23] Department of Entomology/Institute of Biotechnology, College of Bioresources and Agriculture, National Taiwan University, Taipei, Taiwan
[24] Current address: School of Life Sciences, Rochester Institute of Technology, Rochester, New York 14623, USA
[25] Department of Organismic and Evolutionary Biology, Harvard University, 26 Oxford Street, Cambridge, Massachusetts 02138, USA
[26] Department of Molecular and Cellular Biology, Harvard University, 26 Oxford Street, Cambridge, Massachusetts 02138, USA
[27] Department of Biological Sciences, Wayne State University, Detroit, Michigan 48202, USA
[28] Institute for Genetics, University of Cologne, Zülpicher Straße 47a, 50674 Cologne, Germany
[29] Department of Entomology, Texas A&M University, College Station, Texas 77843, USA
[30] CECAD, University of Cologne, Cologne, Germany
[31] Department of Entomology and Program in Molecular & Cell Biology, University of Maryland, College Park, Maryland 20742, USA
[32] Department of Entomology, University of Georgia, 120 Cedar St., Athens, Georgia 30602, USA
[33] Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Laboratory, Berkeley, California, USA
[34] Department of Entomology, College of Agriculture, Food and Environment, University of Kentucky, Lexington, Kentucky 40546, USA
[35] Department of Biology, University of Hawai'i at Mānoa, Honolulu, Hawaii 96822, USA
[36] Current address: Department of Evolutionary Genetics, Max-Planck-Institut für Evolutionsbiologie, August-Thienemann-Straße 2, 24306 Plön, Germany
[37] Current address: Earthworks Institute, 185 Caroline Street, Rochester, New York 14620, USA
[38] Centro de Bioinvestigaciones, Universidad Nacional del Noroeste de Buenos Aires, Argentina
[39] Current address: Department of Biotechnology, Central university of Rajasthan (CURAJ), NH-8, Bandarsindri, Ajmer- 305801, India
[40] Argelander-Institut für Astronomie, Universität Bonn, Auf dem Hügel 71, 53121 Bonn, Germany
[41] Current address: Department of Zoology, University of Cambridge, Cambridge CB2 3DT, UK
[42] Centro Regional de Estudios Genómicos, Facultad de Ciencias Exactas, Universidad Nacional de La Plata, La Plata, Argentina

1447 **REFERENCES**
1448

1449 1. Zdobnov EM, Tegenfeldt F, Kuznetsov D, Waterhouse RM, Simão FA, Ioannidis P,
1450 Seppey M, Loetscher A, Kriventseva EV: **OrthoDB v9.1: cataloging evolutionary and**
1451 **functional annotations for animal, fungal, plant, archaeal, bacterial and viral**
1452 **orthologs.** *Nucleic Acids Res* 2017, **45:**D744-D749.
1453 2. Huang DY, Bechly G, Nel P, Engel MS, Prokop J, Azar D, Cai CY, van de Kamp T,
1454 Staniczek AH, Garrouste R, et al: **New fossil insect order Permopsocida elucidates**
1455 **major radiation and evolution of suction feeding in hemimetabolous insects**
1456 **(Hexapoda: Acercaria).** *Sci Rep* 2016, **6:**23004.
1457 3. Misof B, Liu S, Meusemann K, Peters RS, Donath A, Mayer C, Frandsen PB, Ware J,
1458 Flouri T, Beutel RG, et al: **Phylogenomics resolves the timing and pattern of insect**
1459 **evolution.** *Science* 2014, **346:**763-767.
1460 4. Grimaldi D, Engel MS: *Evolution of the Insects.* Cambridge: Cambridge University Press;
1461 2005.
1462 5. The International Aphid Genomics Consortium: **Genome sequence of the pea aphid**
1463 ***Acyrthosiphon pisum*** *PLoS Biol* 2010, **8:**e1000313.
1464 6. Mathers TC, Chen Y, Kaithakottil G, Legeai F, Mugford ST, Baa-Puyoulet P, Bretaudeau
1465 A, Clavijo B, Colella S, Collin O, et al: **Rapid transcriptional plasticity of duplicated**
1466 **gene clusters enables a clonally reproducing aphid to colonise diverse plant species.**
1467 *Genome Biol* 2017, **18:**27.
1468 7. Wenger JA, Cassone BJ, Legeai F, Johnston JS, Bansal R, Yates AD, Coates BS,
1469 Pavinato VA, Michel A: **Whole genome sequence of the soybean aphid, Aphis**
1470 **glycines.** *Insect Biochem Mol Biol* 2017.
1471 8. Sloan DB, Nakabachi A, Richards S, Qu J, Murali SC, Gibbs RA, Moran NA: **Parallel**
1472 **histories of horizontal gene transfer facilitated extreme reduction of endosymbiont**
1473 **genomes in sap-feeding insects.** *Mol Biol Evol* 2014, **31:**857-871.
1474 9. Xue J, Zhou X, Zhang C-X, Yu L-L, Fan H-W, Wang Z, Xu H-J, Xi Y, Zhu Z-R, Zhou
1475 W-W, et al: **Genomes of the rice pest brown planthopper and its endosymbionts**
1476 **reveal complex complementary contributions for host adaptation.** *Genome Biol* 2014,
1477 **15:**521.
1478 10. Mesquita RD, Vionette-Amaral RJ, Lowenberger C, Rivera-Pomar R, Monteiro FA,
1479 Minx P, Spieth J, Carvalho AB, Panzera F, Lawson D, et al: **Genome of *Rhodnius***
1480 ***prolixus*, an insect vector of Chagas disease, reveals unique adaptations to**
1481 **hematophagy and parasite infection.** *Proc Natl Acad Sci USA* 2015, **112:**14936-14941.
1482 11. Benoit JB, Adelman ZN, Reinhardt K, Dolan A, Poelchau M, Jennings EC, Szuter EM,
1483 Hagan RW, Gujar H, Shukla JN, et al: **Unique features of a global human ectoparasite**
1484 **identified through sequencing of the bed bug genome.** *Nat Commun* 2016, **7:**10165.
1485 12. Rosenfeld JA, Reeves D, Brugler MR, Narechania A, Simon S, Durrett R, Foox J,
1486 Shianna K, Schatz MC, Gandara J, et al: **Genome assembly and geospatial**
1487 **phylogenomics of the bed bug *Cimex lectularius*.** *Nat Commun* 2016, **7:**10164.
1488 13. Sparks ME, Shelby KS, Kuhar D, Gundersen-Rindal DE: **Transcriptome of the invasive**
1489 **brown marmorated stink bug, Halyomorpha halys (Stal) (Heteroptera:**
1490 **Pentatomidae).** *PLoS One* 2014, **9:**e111646.
1491 14. Ioannidis P, Lu Y, Kumar N, Creasy T, Daugherty S, Chibucos MC, Orvis J, Shetty A,
1492 Ott S, Flowers M, et al: **Rapid transcriptome sequencing of an invasive pest, the**
1493 **brown marmorated stink bug, *Halyomorpha halys*.** *BMC Genomics* 2014, **15:**738.
1494 15. Wilson ACC, Ashton PD, Charles H, Colella S, Febvay G, Jander G, Kushlan PF,
1495 Macdonald SJ, Schwartz JF, Thomas GH, Douglas AE: **Genomic insight into the amino**
1496 **acid relations of the pea aphid, Acyrthosiphon pisum, with its symbiotic bacterium**
1497 **Buchnera aphidicola.** *Insect Molecular Biology* 2010, **19 Suppl 2:**249-258.
1498 16. Li H, Leavengood JM, Jr., Chapman EG, Burkhardt D, Song F, Jiang P, Liu J, Zhou X,
1499 Cai W: **Mitochondrial phylogenomics of Hemiptera reveals adaptive innovations**
1500 **driving the diversification of true bugs.** *Proc Biol Sci* 2017, **284**.
1501 17. Eichler S, Schaub GA: **Development of symbionts in triatomine bugs and the effects**
1502 **of infections with trypanosomatids.** *Exp Parasitol* 2002, **100:**17-27.

1503 18. Matsuura Y, Kikuchi Y, Hosokawa T, Koga R, Meng X-Y, Kamagata Y, Nikoh N,
1504 Fukatsu T: **Evolution of symbiotic organs and endosymbionts in lygaeid stinkbugs.**
1505 *The ISME Journal* 2012, **6:**397–409.
1506 19. Berenbaum MR, Miliczky E: **Mantids and milkweed bugs - efficacy of aposematic**
1507 **coloration against invertebrate predators.** *American Midland Naturalist* 1984, **111:**64-
1508 68.
1509 20. Burdfield-Steel ER, Shuker DM: **The evolutionary ecology of the Lygaeidae.** *Ecol Evol*
1510 2014, **4:**2278-2301.
1511 21. Lawrence PA: **Mitosis and the cell cycle in the metamorphic moult of the milkweed**
1512 **bug Oncopeltus fasciatus; a radioautographic study.** *J Cell Sci* 1968, **3:**391-404.
1513 22. Chipman AD: **Oncopeltus fasciatus as an evo-devo research organism.** *Genesis* 2017,
1514 **55**.
1515 23. Panfilio KA: **Late extraembryonic development and its *zen-RNAi*-induced failure in**
1516 **the milkweed bug *Oncopeltus fasciatus*.** *Dev Biol* 2009, **333:**297-311.
1517 24. Panfilio KA, Roth S: **Epithelial reorganization events during late extraembryonic**
1518 **development in a hemimetabolous insect.** *Dev Biol* 2010, **340:**100-115.
1519 25. Sharma AI, Yanes KO, Jin L, Garvey SL, Taha SM, Suzuki Y: **The phenotypic**
1520 **plasticity of developmental modules.** *Evodevo* 2016, **7:**15.
1521 26. Hughes CL, Kaufman TC: **RNAi analysis of *Deformed, proboscipedia* and *Sex combs***
1522 ***reduced* in the milkweed bug *Oncopeltus fasciatus*: novel roles for Hox genes in the**
1523 **hemipteran head.** *Development* 2000, **127:**3683-3694.
1524 27. Wolfe SL, John B: **The organization and ultrastructure of male meiotic chromosomes**
1525 **in *Oncopeltus fasciatus*.** *Chromosoma* 1965, **17:**85-103.
1526 28. Messthaler H, Traut W: **Phases of Sex Chromosome Inactivation in *Oncopeltus***
1527 ***fasciatus* and *Pyrrhocoris apterus* (Insecta, Heteroptera).** *Caryologia* 1975, **28:**501-
1528 510.
1529 29. McKenna DD, Scully ED, Pauchet Y, Hoover K, Kirsch R, Geib SM, Mitchell RF,
1530 Waterhouse RM, Ahn SJ, Arsala D, et al: **Genome of the Asian longhorned beetle**
1531 **(*Anoplophora glabripennis*), a globally significant invasive species, reveals key**
1532 **functional and evolutionary innovations at the beetle-plant interface.** *Genome Biol*
1533 2016, **17:**227.
1534 30. Simpson JT: **Exploring genome characteristics and sequence quality without a**
1535 **reference.** *Bioinformatics* 2014, **30:**1228-1235.
1536 31. Hughes CL, Kaufman TC: **RNAi analysis of Deformed, proboscipedia and Sex combs**
1537 **reduced in the milkweed bug Oncopeltus fasciatus: novel roles for Hox genes in the**
1538 **Hemipteran head.** *Development* 2000, **127:**3683-3694.
1539 32. Panfilio KA, Liu PZ, Akam M, Kaufman TC: ***Oncopeltus fasciatus zen* is essential for**
1540 **serosal tissue function in katatrepsis.** *Dev Biol* 2006, **292:**226-243.
1541 33. Tian X, Xie Q, Li M, Gao C, Cui Y, Xi L, Bu W: **Phylogeny of pentatomomorphan**
1542 **bugs (Hemiptera-Heteroptera:Pentatomomorpha) based on six Hox gene fragments.**
1543 *Zootaxa* 2011, **2888:**57-68.
1544 34. Ewen-Campen B, Shaner N, Panfilio KA, Suzuki Y, Roth S, Extavour CG: **The**
1545 **maternal and early embryonic transcriptome of the milkweed bug *Oncopeltus***
1546 ***fasciatus*.** *BMC Genomics* 2011, **12:**61.
1547 35. Zhen Y, Aardema ML, Medina EM, Schumer M, Andolfatto P: **Parallel molecular**
1548 **evolution in an herbivore community.** *Science* 2012, **337:**1634-1637.
1549 36. Robertson HM: **The insect chemoreceptor superfamily in Drosophila pseudoobscura:**
1550 **Molecular evolution of ecologically-relevant genes over 25 million years** *J Insect Sci*
1551 2009, **9:**18.
1552 37. Robertson HM: **Taste: Independent origins of chemoreception coding systems?** *Curr*
1553 *Biol* 2001, **11:**R560-R562.
1554 38. Jazwinska A, Rushlow C, Roth S: **The role of *brinker* in mediating the graded**
1555 **response to Dpp in early *Drosophila* embryos.** *Development* 1999, **126:**3323-3334.
1556 39. Ewer J, Truman JW: **Increases in cyclic 3',5'-guanosine monophosphate (cGMP)**
1557 **occur at ecdysis in an evolutionarily conserved crustacean cardioactive peptide-**
1558 **immunoreactive insect neuronal network.** *Journal of Comparative Neurology* 1996,
1559 **370:**330-341.

1560 40. Togawa T, Dunn WA, Emmons AC, Nagao J, Willis JH: **Developmental expression**
1561 **patterns of cuticular protein genes with the R&R Consensus from *Anopheles***
1562 ***gambiae*.** *Insect Biochem Mol Biol* 2008, **38:**508-519.
1563 41. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM: **BUSCO:**
1564 **assessing genome assembly and annotation completeness with single-copy orthologs.**
1565 *Bioinformatics* 2015, **31:**3210-3212.
1566 42. Kriventseva EV, Tegenfeldt F, Petty TJ, Waterhouse RM, Simão FA, Pozdnyakov IA,
1567 Ioannidis P, Zdobnov EM: **OrthoDB v8: update of the hierarchical catalog of**
1568 **orthologs and the underlying free software.** *Nucl Acids Res* 2015, **43:**D250-D256.
1569 43. Shigenobu S, Bickel RD, Brisson JA, Butts T, Chang CC, Christiaens O, Davis GK,
1570 Duncan EJ, Ferrier DE, Iga M, et al: **Comprehensive survey of developmental genes in**
1571 **the pea aphid, *Acyrthosiphon pisum*: frequent lineage-specific duplications and losses**
1572 **of developmental genes.** *Insect Mol Biol* 2010, **19 Suppl 2:**47-62.
1573 44. Bansal R, Michel AP: **Core RNAi Machinery and *Sid1*, a component for systemic**
1574 **RNAi, in the hemipteran insect, *Aphis glycines*.** *Int J Mol Sci* 2013, **14:**3786-3801.
1575 45. Bao R, Fischer T, Bolognesi R, Brown SJ, Friedrich M: **Parallel duplication and partial**
1576 **subfunctionalization of beta-catenin/armadillo during insect evolution.** *Mol Biol Evol*
1577 2012, **29:**647-662.
1578 46. Sachs L, Chen YT, Drechsler A, Lynch JA, Panfilio KA, Lassig M, Berg J, Roth S:
1579 **Dynamic BMP signaling polarized by Toll patterns the dorsoventral axis in a**
1580 **hemimetabolous insect.** *eLife* 2015, **4:**e05502.
1581 47. Konopova B, Smykal V, Jindra M: **Common and distinct roles of juvenile hormone**
1582 **signaling genes in metamorphosis of holometabolous and hemimetabolous insects.**
1583 *PLoS One* 2011, **6:**e28728.
1584 48. Armisen D, Refki PN, Crumiere AJ, Viala S, Toubiana W, Khila A: **Predator strike**
1585 **shapes antipredator phenotype through new genetic interactions in water striders.**
1586 *Nat Commun* 2015, **6:**8153.
1587 49. Wulff JP, Sierra I, Sterkel M, Holtof M, Van Wielendaele P, Francini F, Broeck JV, Ons
1588 S: **Orcokinin neuropeptides regulate ecdysis in the hemimetabolous insect Rhodnius**
1589 **prolixus.** *Insect Biochem Mol Biol* 2017, **81:**91-102.
1590 50. Vellichirammal NN, Gupta P, Hall TA, Brisson JA: **Ecdysone signaling underlies the**
1591 **pea aphid transgenerational wing polyphenism.** *Proc Natl Acad Sci U S A* 2017,
1592 **114:**1419-1423.
1593 51. Chiu TL, Wen Z, Rupasinghe SG, Schuler MA: **Comparative molecular modeling of**
1594 **Anopheles gambiae CYP6Z1, a mosquito P450 capable of metabolizing DDT.** *Proc*
1595 *Natl Acad Sci U S A* 2008, **105:**8855-8860.
1596 52. Gong Y, Li T, Feng Y, Liu N: **The function of two P450s, CYP9M10 and CYP6AA7,**
1597 **in the permethrin resistance of Culex quinquefasciatus.** *Sci Rep* 2017, **7:**587.
1598 53. Liu PZ, Kaufman TC: ***hunchback* is required for suppression of abdominal identity,**
1599 **and for proper germband growth and segmentation in the intermediate germband**
1600 **insect *Oncopeltus fasciatus*.** *Development* 2004, **131:**1515-1527.
1601 54. Schaeper ND, Pechmann M, Damen WGM, Prpic N-M, Wimmer EA: **Evolutionary**
1602 **plasticity of *collier* function in head development of diverse arthropods.** *Dev Biol*
1603 2010, **344:**363-376.
1604 55. Aspiras AC, Smith FW, Angelini DR: **Sex-specific gene interactions in the patterning**
1605 **of insect genitalia.** *Dev Biol* 2011, **360:**369-380.
1606 56. Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, Najafabadi
1607 HS, Lambert SA, Mann I, Cook K, et al: **Determination and inference of eukaryotic**
1608 **transcription factor sequence specificity.** *Cell* 2014, **158:**1431-1443.
1609 57. Peel AD, Telford MJ, Akam M: **The evolution of hexapod engrailed-family genes:**
1610 **evidence for conservation and concerted evolution.** *Proc Biol Sci* 2006, **273:**1733-
1611 1742.
1612 58. Ben-David J, Chipman AD: **Mutual regulatory interactions of the trunk gap genes**
1613 **during blastoderm patterning in the hemipteran *Oncopeltus fasciatus*.** *Dev Biol* 2010,
1614 **346:**140-149.

1615 59. Erezyilmaz DF, Kelstrup HC, Riddiford LM: **The nuclear receptor E75A has a novel**
1616 **pair-rule-like function in patterning the milkweed bug, *Oncopeltus fasciatus*.** *Dev*
1617 *Biol* 2009, **334:**300-310.
1618 60. Liu PZ, Kaufman TC: ***even-skipped* is not a pair-rule gene but has segmental and gap-**
1619 **like functions in *Oncopeltus fasciatus*, an intermediate germband insect.** *Development*
1620 2005, **132:**2081-2092.
1621 61. Weisbrod A, Cohen M, Chipman AD: **Evolution of the insect terminal patterning**
1622 **system--insights from the milkweed bug, *Oncopeltus fasciatus*.** *Dev Biol* 2013,
1623 **380:**125-131.
1624 62. Albertin CB, Simakov O, Mitros T, Wang ZY, Pungor JR, Edsinger-Gonzales E, Brenner
1625 S, Ragsdale CW, Rokhsar DS: **The octopus genome and the evolution of cephalopod**
1626 **neural and morphological novelties.** *Nature* 2015, **524:**220-224.
1627 63. Crooks GE, Hon G, Chandonia J-M, Brenner SE: **WebLogo: A sequence logo**
1628 **generator.** *Genome Res* 2004, **14:**1188-1190.
1629 64. Najafabadi HS, Mnaimneh S, Schmitges FW, Garton M, Lam KN, Yang A, Albu M,
1630 Weirauch MT, Radovani E, Kim PM, et al: **C2H2 zinc finger proteins greatly expand**
1631 **the human regulatory lexicon.** *Nat Biotechnol* 2015, **33:**555-562.
1632 65. Emerson RO, Thomas JH: **Adaptive evolution in zinc finger transcription factors.**
1633 *PLoS Genet* 2009, **5:**e1000325.
1634 66. Thomas JH, Schneider S: **Coevolution of retroelements and tandem zinc finger genes.**
1635 *Genome Res* 2011, **21:**1800-1812.
1636 67. Garcia-Perez JL, Widmann TJ, Adams IR: **The impact of transposable elements on**
1637 **mammalian development.** *Development* 2016, **143:**4101-4114.
1638 68. Liu PZ, Kaufman TC: ***Krüppel* is a gap gene in the intermediate insect *Oncopeltus***
1639 ***fasciatus* and is required for development of both blastoderm and germband-derived**
1640 **segments.** *Development* 2004, **131:**4567-4579.
1641 69. Heger P, Marin B, Bartkuhn M, Schierenberg E, Wiehe T: **The chromatin insulator**
1642 **CTCF and the emergence of metazoan diversity.** *Proc Natl Acad Sci USA* 2012,
1643 **109:**17507-17512.
1644 70. Liu H, Chang L-H, Sun Y, Lu X, Stubbs L: **Deep vertebrate roots for mammalian zinc**
1645 **finger transcription factor subfamilies.** *Genome Biol Evol* 2014, **6:**510-525.
1646 71. Imbeault M, Helleboid P-Y, Trono D: **KRAB zinc-finger proteins contribute to the**
1647 **evolution of gene regulatory networks.** *Nature* 2017, **543:**550-554.
1648 72. Csuros M, Rogozin IB, Koonin EV: **A detailed history of intron-rich eukaryotic**
1649 **ancestors inferred from a global survey of 100 complete genomes.** *PLoS Comput Biol*
1650 2011, **7:**e1002150.
1651 73. Papanicolaou A, Schetelig MF, Arensburger P, Atkinson PW, Benoit JB, Bourtzis K,
1652 Castañera P, Cavanaugh JP, Chao H, Childers C, et al: **The whole genome sequence of**
1653 **the Mediterranean fruit fly, *Ceratitis capitata* (Wiedemann), reveals insights into the**
1654 **biology and adaptive evolution of a highly invasive pest species.** *Genome Biol* 2016,
1655 **17:**192.
1656 74. Hoy MA, Waterhouse RM, Wu K, Estep AS, Ioannidis P, Palmer WJ, Pomerantz AF,
1657 Simao FA, Thomas J, Jiggins FM, et al: **Genome sequencing of the phytoseiid**
1658 **predatory mite *Metaseiulus occidentalis* reveals completely atomized Hox genes and**
1659 **superdynamic intron evolution.** *Genome Biol Evol* 2016, **8:**1762-1775.
1660 75. Seibt KM, Wenke T, Muders K, Truberg B, Schmidt T: **Short interspersed nuclear**
1661 **elements (SINEs) are abundant in Solanaceae and have a family-specific impact on**
1662 **gene structure and genome organization.** *Plant J* 2016, **86:**268-285.
1663 76. Huff JT, Zilberman D, Roy SW: **Mechanism for DNA transposons to generate introns**
1664 **on genomic scales.** *Nature* 2016, **538:**533-536.
1665 77. Wheeler D, Redding AJ, Werren JH: **Characterization of an ancient lepidopteran**
1666 **lateral gene transfer.** *PLoS One* 2012, **8:**e59262.
1667 78. Da Lage JL, Binder M, Hua-Van A, Janecek S, Casane D: **Gene make-up: rapid and**
1668 **massive intron gains after horizontal transfer of a bacterial alpha-amylase gene to**
1669 **Basidiomycetes.** *BMC Evol Biol* 2013, **13:**40.

1670 79. Lee DH, Short BD, Joseph SV, Bergh JC, Leskey TC: **Review of the biology, ecology,**
1671 **and management of *Halyomorpha halys* (Hemiptera: Pentatomidae) in China,**
1672 **Japan, and the Republic of Korea.** *Environ Entomol* 2013, **42:**627-641.
1673 80. Lawrence PA: **Cellular differentiation and pattern formation during metamorphosis**
1674 **of the milkweed bug Oncopeltus.** *Dev Biol* 1969, **19:**12-40.
1675 81. Riddiford LM: **Prevention of Metamorphosis by Exposure of Insect Eggs to Juvenile**
1676 **Hormone Analogs.** *Science* 1970, **167:**287-&.
1677 82. Willis JH, Lawrence PA: **Deferred Action of Juvenile Hormone.** *Nature* 1970, **225:**81-
1678 83.
1679 83. Masner P, Bowers WS, Kalin M, Muhle T: **Effect of precocene II on the endocrine**
1680 **regulation of development and reproduction in the bug, *Oncopeltus fasciatus*.** *Gen*
1681 *Comp Endocrinol* 1979, **37:**156-166.
1682 84. Rewitz K, O'Connor M, Gilbert L: **Molecular evolution of the insect Halloween family**
1683 **of cytochrome P450s: phylogeny, gene organization and functional conservation.**
1684 *Insect Biochem Mol Biol* 2007, **37:**741-753.
1685 85. Huet F, Ruiz C, Richards G: **Sequential gene activation by ecdysone in Drosophila**
1686 **melanogaster: the hierarchical equivalence of early and early late genes.**
1687 *Development* 1995, **121:**1195-1204.
1688 86. Bialecki M, Shilton A, Fichtenberg C, Segraves WA, Thummel CS: **Loss of the**
1689 **ecdysteroid-inducible E75A orphan nuclear receptor uncouples molting from**
1690 **metamorphosis in Drosophila.** *Dev Cell* 2002, **3:**209-220.
1691 87. Truman J, Rountree D, Reiss S, Schwartz L: **Ecdysteroids regulate the release and**
1692 **action of eclosion hormone in the tobacco hornworm, Manduca sexta (L.)** *J Insect*
1693 *Physiol* 1983, **29:**895–900.
1694 88. Zitnan D, Kingan TG, Hermesman JL, Adams ME: **Identification of ecdysis-triggering**
1695 **hormone from an epitracheal endocrine system.** *Science* 1996, **271:**88-91.
1696 89. Charles JP, Iwema T, Epa VC, Takaki K, Rynes J, Jindra M: **Ligand-binding properties**
1697 **of a juvenile hormone receptor, Methoprene-tolerant.** *Proceedings of the National*
1698 *Academy of Sciences of the United States of America* 2011, **108:**21128-21133.
1699 90. Minakuchi C, Zhou X, Riddiford L: **Kruppel homolog 1 (Kr-h1) mediates juvenile**
1700 **hormone action during metamorphosis of Drosophila melanogaster.** *Mech Dev* 2008,
1701 **125:**91-105.
1702 91. Minakuchi C, Namiki T, Shinoda T: **Kruppel homolog 1, an early juvenile hormone-**
1703 **response gene downstream of Methoprene-tolerant, mediates its anti-metamorphic**
1704 **action in the red flour beetle Tribolium castaneum.** *Dev Biol* 2009, **352:**341-350.
1705 92. DiBello PR, Withers DA, Bayer CA, Fristrom JW, Guild GM: **The Drosophila Broad-**
1706 **Complex encodes a family of related proteins containing zinc fingers.** *Genetics* 1991,
1707 **129:**385-397.
1708 93. Karim F, Guild G, Thummel C: **The Drosophila Broad-Complex plays a key role in**
1709 **controlling ecdysone-regulated gene expression at the onset of metamorphosis.**
1710 *Development* 1993, **118:**977-988.
1711 94. Erezyilmaz DF, Riddiford LM, Truman JW: **The pupal specifier broad directs**
1712 **progressive morphogenesis in a direct-developing insect.** *Proceedings of the National*
1713 *Academy of Sciences of the United States of America* 2006, **103:**6925-6930.
1714 95. Arakane Y, Hogenkamp DG, Zhu YC, Kramer KJ, Specht CA, Beeman RW, Kanost MR,
1715 Muthukrishnan S: **Characterization of two chitin synthase genes of the red flour**
1716 **beetle, *Tribolium castaneum*, and alternate exon usage in one of the genes during**
1717 **development.** *Insect Biochem Mol Biol* 2004, **34:**291-304.
1718 96. True JR: **Insect melanism: the molecules matter.** *Trends in Ecology & Evolution* 2003,
1719 **18:**640-647.
1720 97. Zhan SA, Guo QH, Li MH, Li MW, Li JY, Miao XX, Huang YP: **Disruption of an N-**
1721 **acetyltransferase gene in the silkworm reveals a novel role in pigmentation.**
1722 *Development* 2010, **137:**4083-4090.
1723 98. Liu J, Lemonds TR, Popadic A: **The genetic control of aposematic black pigmentation**
1724 **in hemimetabolous insects: insights from Oncopeltus fasciatus.** *Evolution &*
1725 *Development* 2014, **16:**270-277.

99. Liu J, Lemonds TR, Marden JH, Popadic A: **A Pathway Analysis of Melanin Patterning in a Hemimetabolous Insect.** *Genetics* 2016, **203:**403-413.
100. Lawrence PA: **Some new mutants of large milkweed bug *Oncopeltus fasciatus* Dall.** *Genetical Research* 1970, **15:**347-350.
101. Morgan ED: *Biosynthesis in Insects: Advanced Edition.* London: Royal Society of Chemistry,; 2010.
102. McLean JR, Krishnakumar S, O'Donnell JM: **Multiple mRNAs from the Punch locus of Drosophila melanogaster encode isoforms of GTP cyclohydrolase I with distinct N-terminal domains.** *J Biol Chem* 1993, **268:**27191-27197.
103. Wiederrecht GJ, Paton DR, Brown GM: **Enzymatic Conversion of Dihydroneopterin Triphosphate to the Pyrimidodiazepine Intermediate Involved in the Biosynthesis of the Drosopterins in Drosophila-Melanogaster.** *Journal of Biological Chemistry* 1984, **259:**2195-2200.
104. Newcombe D, Blount JD, Mitchell C, Moore AJ: **Chemical egg defence in the large milkweed bug, Oncopeltus fasciatus, derives from maternal but not paternal diet.** *Entomologia Experimentalis et Applicata* 2013, **149:**197-205.
105. Zhan S, Merlin C, Boore JL, Reppert SM: **The monarch butterfly genome yields insights into long-distance migration.** *Cell* 2011, **147:**1171-1185.
106. Joseph RM, Carlson JR: **Drosophila Chemoreceptors: A Molecular Interface Between the Chemical World and the Brain.** *Trends Genet* 2015, **31:**683-695.
107. Benton R: **Multigene Family Evolution: Perspectives from Insect Chemoreceptors.** *Trends Ecol Evol* 2015, **30:**590-600.
108. Robertson HM, Warr CG, Carlson JR: **Molecular evolution of the insect chemoreceptor gene superfamily in Drosophila melanogaster.** *Proc Natl Acad Sci U S A* 2003, **100 Suppl 2:**14537-14542.
109. Rytz R, Croset V, Benton R: **Ionotropic receptors (IRs): chemosensory ionotropic glutamate receptors in Drosophila and beyond.** *Insect Biochem Mol Biol* 2013, **43:**888-897.
110. Kirkness EF, Haas BJ, Sun W, Braig HR, Perotti MA, Clark JM, Lee SH, Robertson HM, Kennedy RC, Elhaik E, et al: **Genome sequences of the human body louse and its primary endosymbiont provide insights into the permanent parasitic lifestyle.** *Proc Natl Acad Sci U S A* 2010, **107:**12168-12173.
111. Smadja C, Shi P, Butlin RK, Robertson HM: **Large gene family expansions and adaptive evolution for odorant and gustatory receptors in the pea aphid, Acyrthosiphon pisum.** *Mol Biol Evol* 2009, **26:**2073-2086.
112. Terrapon N, Li C, Robertson HM, Ji L, Meng X, Booth W, Chen Z, Childers CP, Glastad KM, Gokhale K, et al: **Molecular traces of alternative social organization in a termite genome.** *Nat Commun* 2014, **5:**3636.
113. Xu W, Papanicolaou A, Zhang HJ, Anderson A: **Expansion of a bitter taste receptor family in a polyphagous insect herbivore.** *Sci Rep* 2016, **6:**23666.
114. Feir D: ***Oncopeltus fasciatus*: A research animal.** *Annu Rev Entomol* 1974, **19:**81-96.
115. Croset V, Rytz R, Cummins SF, Budd A, Brawand D, Kaessmann H, Gibson TJ, Benton R: **Ancient protostome origin of chemosensory ionotropic glutamate receptors and the evolution of insect taste and olfaction.** *PLoS Genet* 2010, **6:**e1001064.
116. Vellozo AF, Véron AS, Baa-Puyoulet P, Huerta-Cepas J, Cottret L, Febvay G, Calevro F, Rahbe Y, Douglas AE, Gabaldón T, et al: **CycADS: an annotation database system to ease the development and update of BioCyc databases.** *Database* 2011, **2011:**bar008-bar008.
117. Baa-Puyoulet P, Parisot N, Febvay G, Huerta-Cepas J, Vellozo AF, Gabaldón T, Calevro F, Charles H, Colella S: **ArthropodaCyc: a CycADS powered collection of BioCyc databases to analyse and compare metabolism of arthropods.** *Database (Oxford)* 2016, **pii:**baw081.
118. Hojilla-Evangelista MP, Evangelista RL: **Characterization of milkweed (Asclepias spp.) seed proteins.** *Industrial crops and ...* 2009.
119. Dean CAE, Teets NM, Koštál V, Šimek P, Denlinger DL: **Enhanced stress responses and metabolic adjustments linked to diapause and onset of migration in the large milkweed bug *Oncopeltus fasciatus*.** *Physiol Entomol* 2016, **DOI: 10.1111/phen.12140**.

120. Sandström J, Moran N: **How nutritionally imbalanced is phloem sap for aphids?** *Entomologia Experimentalis et Applicata* 1999, **91:**203-210.
121. Rabatel A, Febvay G, Gaget K, Duport G, Baa-Puyoulet P, Sapountzis P, Bendridi N, Rey M, Rahbé Y, Charles H, et al: **Tyrosine pathway regulation is host-mediated in the pea aphid symbiosis during late embryonic and early larval development.** *BMC Genomics* 2013, **14:**235.
122. Dobler S, Petschenka G, Wagschal V, Flacht L: **Convergent adaptive evolution – how insects master the challenge of cardiac glycoside-containing host plants.** *Entomologia Experimentalis et Applicata* 2015, **157:**30-39.
123. Niehuis O, Gibson JD, Rosenberg MS, Pannebakker BA, Koevoets T, Judson AK, Desjardins CA, Kennedy K, Duggan D, Beukeboom LW, et al: **Recombination and its impact on the genome of the haplodiploid parasitoid wasp Nasonia.** *PLoS One* 2010, **5:**e8597.
124. Ferrero A, Torreblanca A, Garcera MD: **Assessment of the effects of orally administered ferrous sulfate on Oncopeltus fasciatus (Heteroptera: Lygaeidae).** *Environ Sci Pollut Res Int* 2017, **24:**8551-8561.
125. Hare EE, Johnston JS: **Genome size determination using flow cytometry of propidium iodide-stained nuclei.** *Methods Mol Biol* 2011, **772:**3-12.
126. Marcais G, Kingsford C: **A fast, lock-free approach for efficient parallel counting of occurrences of k-mers.** *Bioinformatics* 2011, **27:**764-770.
127. Bushnell B: **BBMap short read aligner.** 2016. http://sourceforge.net/projects/bbmap
128. Gnerre S, Maccallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, Sharpe T, Hall G, Shea TP, Sykes S, et al: **High-quality draft assemblies of mammalian genomes from massively parallel sequence data.** *Proc Natl Acad Sci U S A* 2011, **108:**1513-1518.
129. Holt C, Yandell M: **MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects.** *BMC Bioinformatics* 2011, **12:**491.
130. Lee E, Helt G, Reese J, Munoz-Torres M, Childers C, Buels R, Stein L, Holmes I, Elsik C, Lewis S: **Web Apollo: a web-based genomic annotation editing platform.** *Genome Biology* 2013, **14**.
131. Poelchau M, Childers C, Moore G, Tsavatapalli V, Evans J, Lee CY, Lin H, Lin JW, Hackett K: **The i5k Workspace@NAL--enabling genomic data access, visualization and curation of arthropod genomes.** *Nucleic Acids Res* 2015, **43:**D714-719.
132. Hughes DST, Koelzer S, Panfilio KA, Richards S: ***Oncopeltus fasciatus* genome annotations v0.5.3.** *Ag Data Commons (Database)* 2015**:**http://dx.doi.org/10.15482/USDA.ADC/1173237.
133. Murali SC, The i5k genome assembly team (29 additional authors), Han Y, Richards S, Worley K, Muzny D, Gibbs R, Koelzer S, Panfilio KA: ***Oncopeltus fasciatus* genome assembly 1.0.** *Ag Data Commons (Database)* 2015**:**http://dx.doi.org/10.15482/USDA.ADC/1173238.
134. Vargas Jentzsch IM, Hughes DST, Poelchau M, Robertson HM, Benoit JB, Rosendale AJ, Armisén D, Duncan EJ, Vreede BMI, Jacobs CGC, et al: ***Oncopeltus fasciatus* Official Gene Set OGS_v1.1 for genome assembly Oncopeltus fasciatus v1.0.** *Ag Data Commons (Database)* 2015**:**http://dx.doi.org/10.15482/USDA.ADC/1173142.
135. **RepeatModeler Open-1.0.8** [http://www.repeatmasker.org]
136. Bao Z, Eddy SR: **Automated de novo identification of repeat sequence families in sequenced genomes.** *Genome Res* 2002, **12:**1269-1276.
137. Price AL, Jones NC, Pevzner PA: **De novo identification of repeat families in large genomes.** *Bioinformatics* 2005, **21 Suppl 1:**i351-358.
138. Benson G: **Tandem repeats finder: a program to analyze DNA sequences.** *Nucleic Acids Res* 1999, **27:**573-580.
139. **RepeatMasker Open-4.0.** [http://www.repeatmasker.org]
140. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2.** *Nat Methods* 2012, **9:**357-359.

141. Goff S, Vaughn M, McKay S, Lyons E, Stapleton A, Gessler D, Matasci N, Wang L, Hanlon M, Lenards A, et al: **The iPlant Collaborative: Cyberinfrastructure for Plant Biology.** *Frontiers in plant science* 2011, **2**.
142. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, et al: **Full-length transcriptome assembly from RNA-Seq data without a reference genome.** *Nat Biotechnol* 2011, **29:**644-652.
143. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, et al: **De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis.** *Nat Protoc* 2013, **8:**1494-1512.
144. Li B, Dewey CN: **RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome.** *BMC Bioinformatics* 2011, **12:**323.
145. Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K, et al: **The Pfam protein families database.** *Nucleic Acids Res* 2010, **38:**D211-222.
146. Weirauch MT, Hughes TR: **A catalogue of eukaryotic transcription factor types, their evolutionary origin, and species distribution.** *Subcell Biochem* 2011, **52:**25-73.
147. Eddy SR: **A new generation of homology search tools based on probabilistic inference.** *Genome Inform* 2009, **23:**205-211.
148. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Soding J, et al: **Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega.** *Mol Syst Biol* 2011, **7:**539.
149. Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M: **KAAS: an automatic genome annotation and pathway reconstruction server.** *Nucleic Acids Research* 2007, **35:**W182-185.
150. Claudel-Renard C, Chevalet C, Faraut T, Kahn D: **Enzyme-specific profiles for genome annotation: PRIAM.** *Nucleic Acids Research* 2003, **31:**6633-6639.
151. Conesa A, Götz S: **Blast2GO: A Comprehensive Suite for Functional Analysis in Plant Genomics.** *International journal of plant genomics* 2008, **2008:**619832.
152. Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M: **Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research.** *Bioinformatics (Oxford, England)* 2005, **21:**3674-3676.
153. Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, et al: **InterProScan 5: genome-scale protein function classification.** *Bioinformatics* 2014, **30:**1236-1240.
154. Karp PD, Ouzounis CA, Moore-Kochlacs C, Goldovsky L, Kaipa P, Ahrén D, Tsoka S, Darzentas N, Kunin V, López-Bigas N: **Expansion of the BioCyc collection of pathway/genome databases to 160 genomes.** *Nucleic Acids Research* 2005, **33:**6083-6089.
155. Karp PD, Paley SM, Krummenacker M, Latendresse M, Dale JM, Lee TJ, Kaipa P, Gilham F, Spaulding A, Popescu L, et al: **Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology.** *Briefings in Bioinformatics* 2010, **11:**40-79.
156. Dereeper A, Guignon V, Blanc G, Audic S, Buffet S, Chevenet F, Dufayard JF, Guindon S, Lefort V, Lescot M, et al: **Phylogeny.fr: robust phylogenetic analysis for the non-specialist.** *Nucleic Acids Res* 2008, **36:**W465-469.
157. Wang X, Fang X, Yang P, Jiang X, Jiang F, Zhao D, Li B, Cui F, Wei J, Ma C, et al: **The locust genome provides insight into swarm formation and long-distance flight.** *Nat Commun* 2014, **5:**2957.
158. The International Silkworm Genome Consortium: **The genome of a lepidopteran model insect, the silkworm Bombyx mori.** *Insect Biochem Mol Biol* 2008, **38:**1036-1045.
159. Elsik CG, Worley KC, Bennett AK, Beye M, Camara F, Childers CP, de Graaf DC, Debyser G, Deng J, Devreese B, et al: **Finding the missing honey bee genes: lessons learned from a genome upgrade.** *BMC Genomics* 2014, **15:**86.
160. Honeybee Genome Sequencing Consortium: **Insights into social insects from the genome of the honeybee Apis mellifera.** *Nature* 2006, **443:**931-949.

1894 161. Richards S, Gibbs RA, Weinstock GM, Brown SJ, Denell R, Beeman RW, Gibbs R,
1895 Bucher G, Friedrich M, Grimmelikhuijzen CJ, et al: **The genome of the model beetle**
1896 **and pest Tribolium castaneum.** *Nature* 2008, **452:**949-955.
1897 162. Chen XG, Jiang X, Gu J, Xu M, Wu Y, Deng Y, Zhang C, Bonizzoni M, Dermauw W,
1898 Vontas J, et al: **Genome sequence of the Asian Tiger mosquito, Aedes albopictus,**
1899 **reveals insights into its biology, genetics, and evolution.** *Proc Natl Acad Sci U S A*
1900 2015, **112:**E5907-5915.
1901 163. Ellis LL, Huang W, Quinn AM, Ahuja A, Alfrejd B, Gomez FE, Hjelmen CE, Moore KL,
1902 Mackay TF, Johnston JS, Tarone AM: **Intrapopulation genome size variation in D.**
1903 **melanogaster reflects life history variation and plasticity.** *PLoS Genet* 2014,
1904 **10:**e1004522.
1905 164. Kumar S, Stecher G, Tamura K: **MEGA7: Molecular Evolutionary Genetics Analysis**
1906 **Version 7.0 for Bigger Datasets.** *Mol Biol Evol* 2016, **33:**1870-1874.
1907

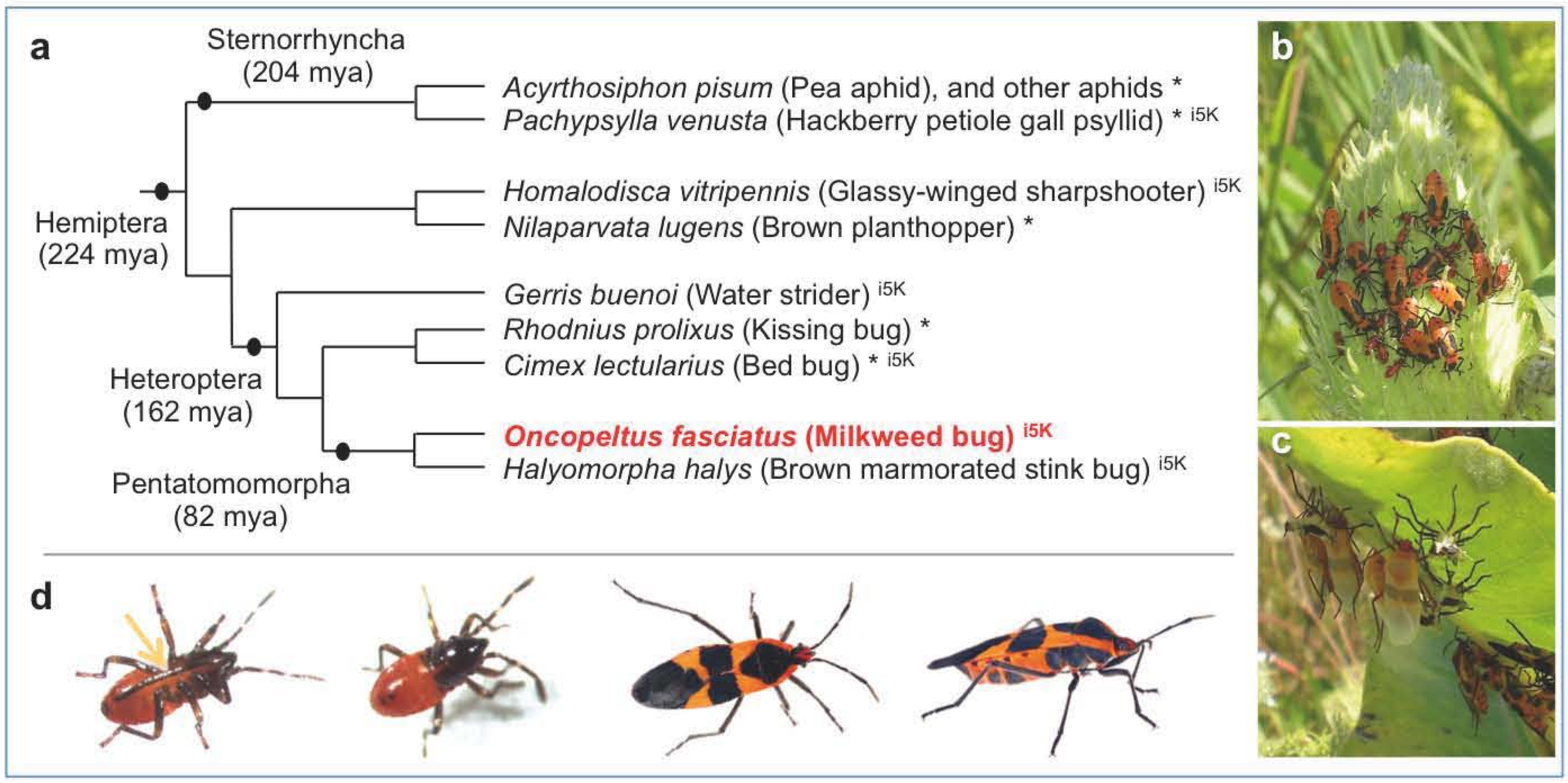


Fig 1. The large milkweed bug, *Oncopeltus fasciatus*, shown in its phylogenetic and environmental context.

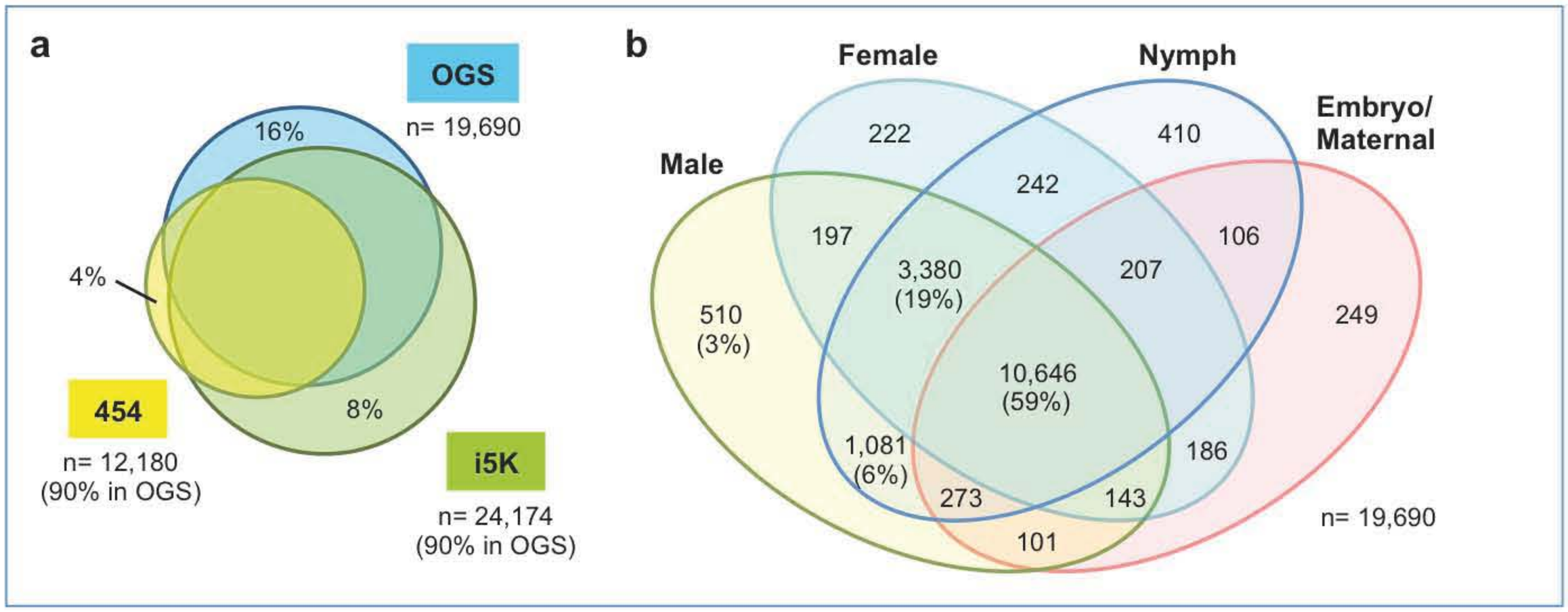


Fig 2. Comparisons of the official gene set and transcriptomic resources.

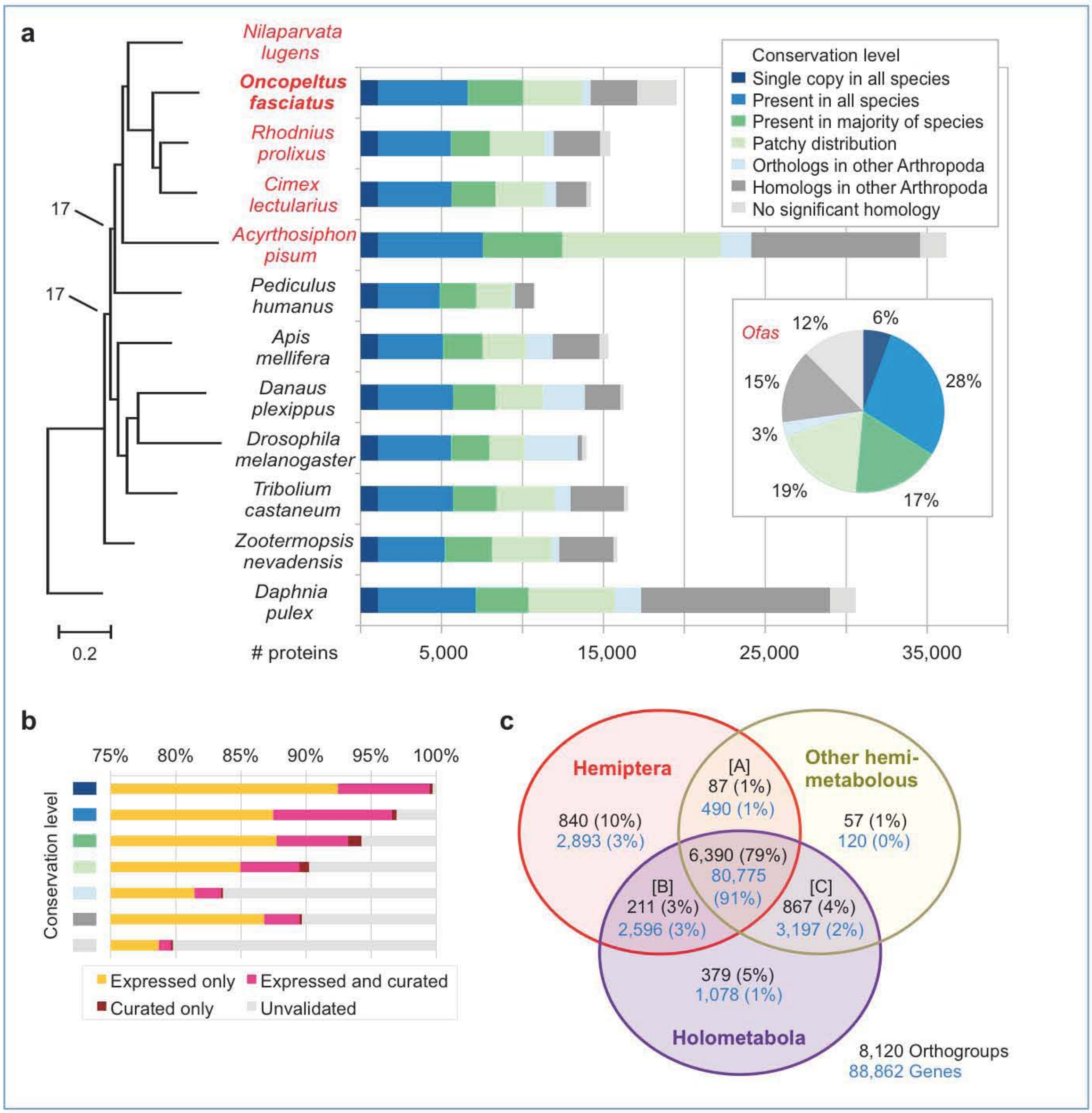


Fig 3. Orthology comparisons and phylogenetic placement of *Oncopeltus fasciatus* among other Arthropoda.

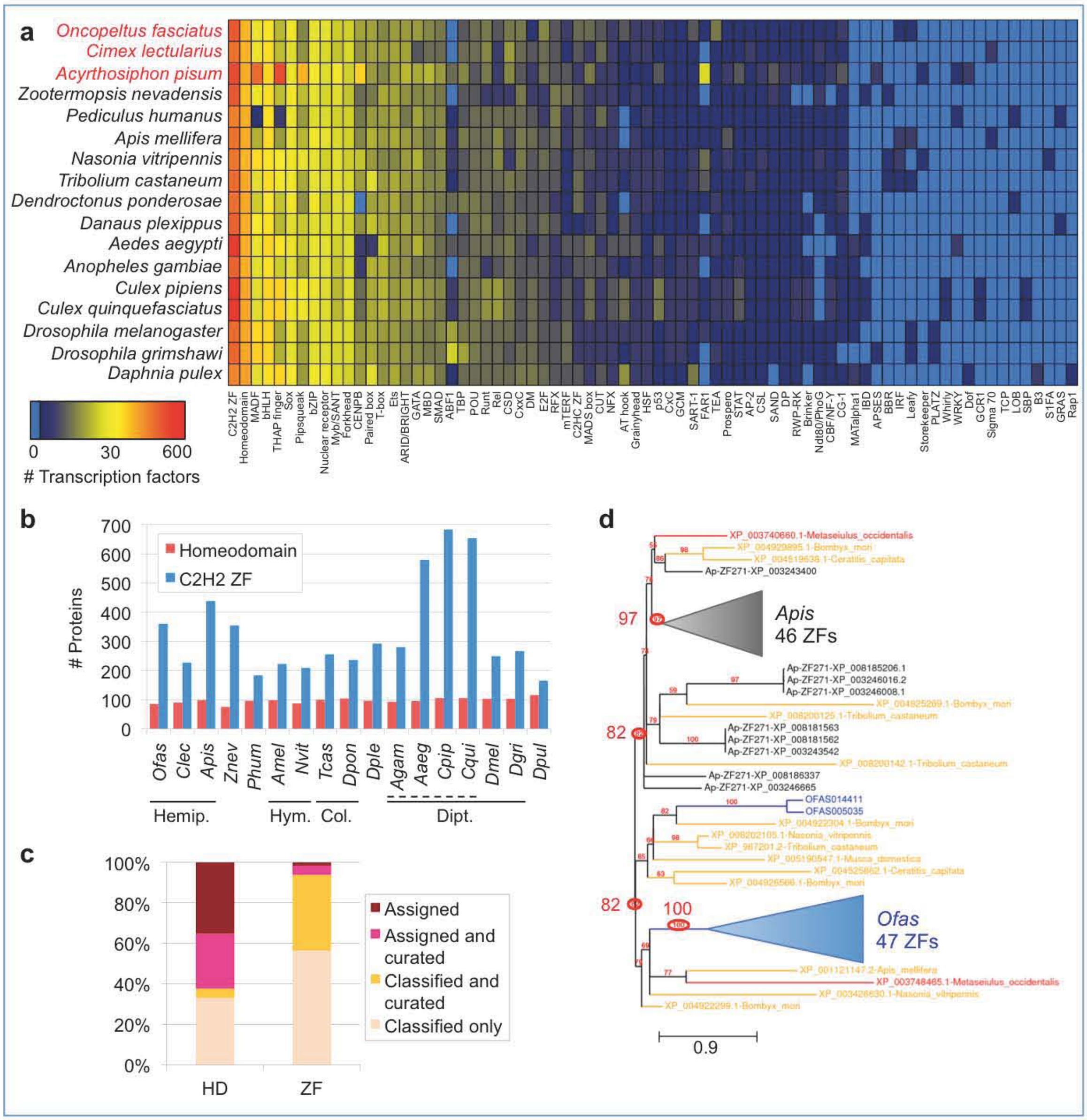


Fig 4. Distribution of transcription factor families across insect genomes.

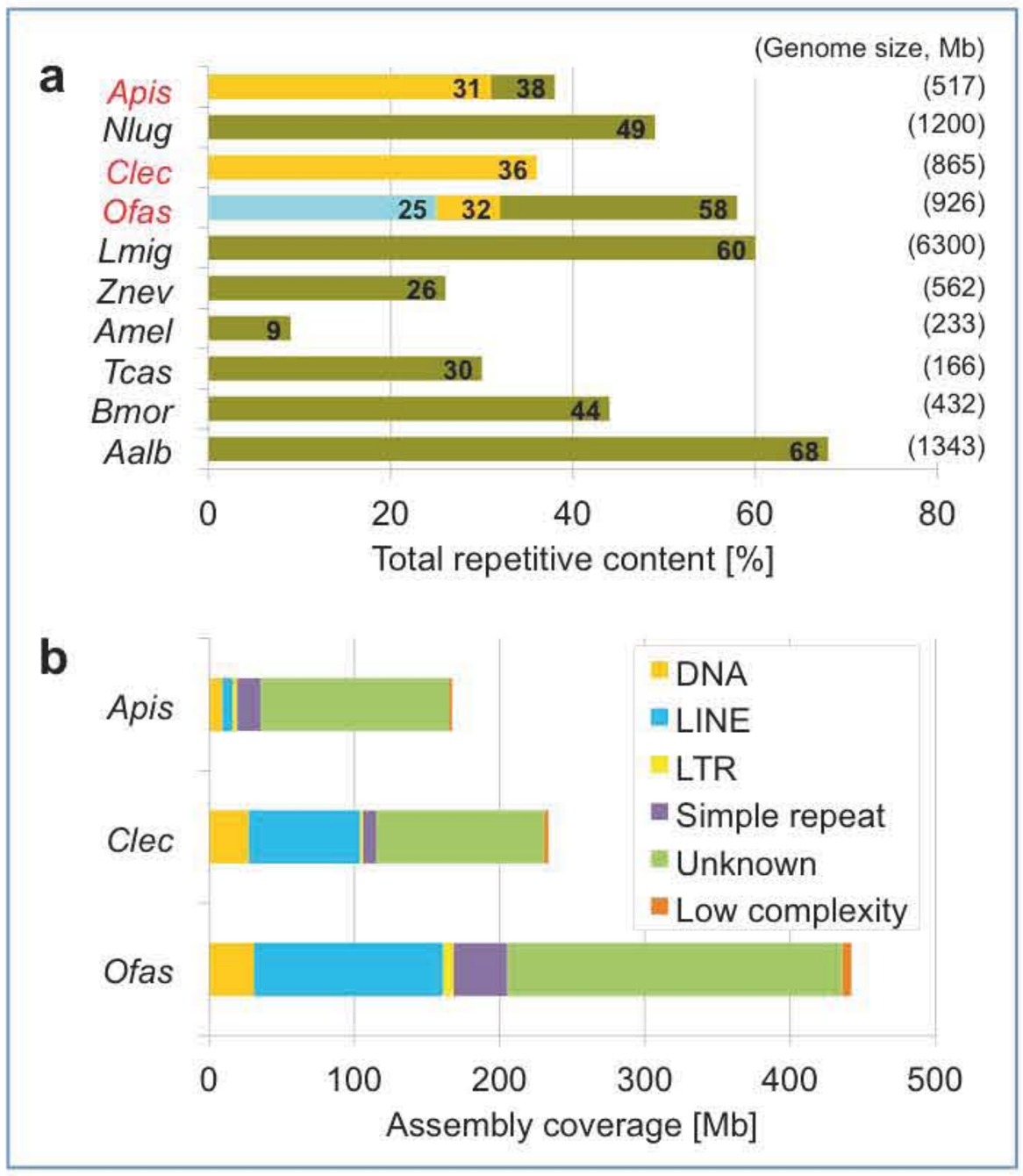


Fig 5. Comparison of repeat content estimations.

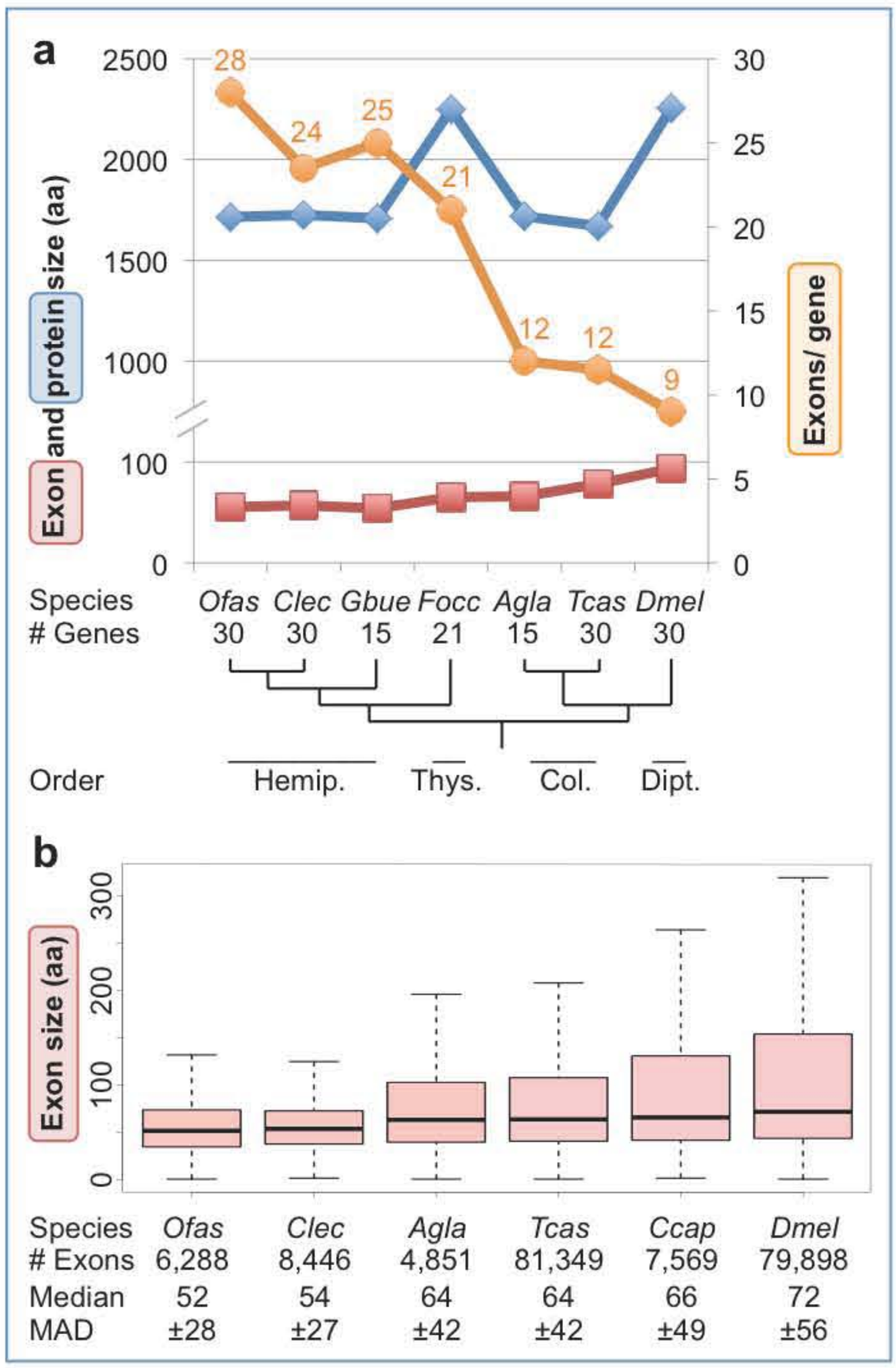


Fig 6. Trends in gene structure show hemipteroid-specific tendencies.

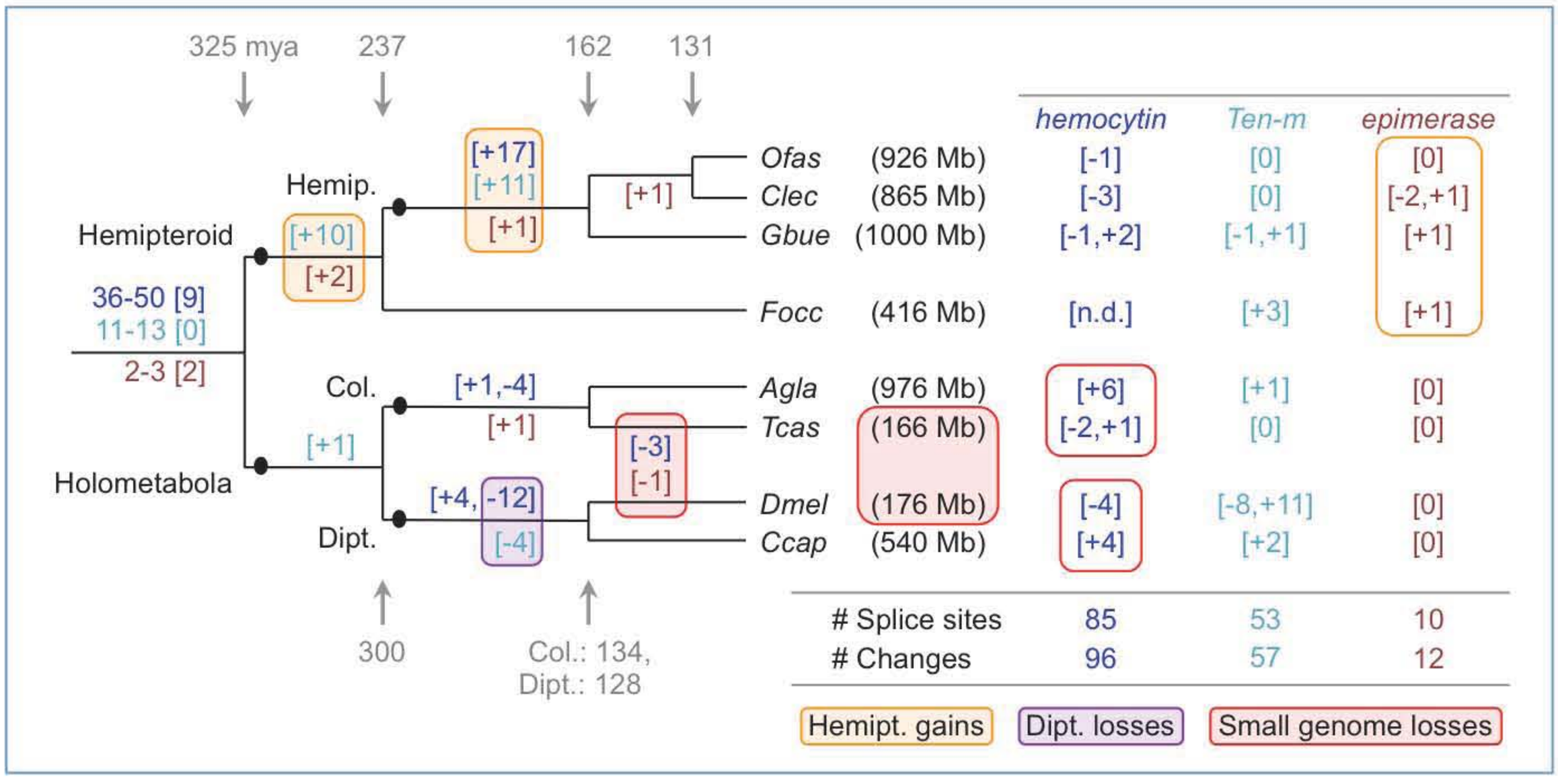


Fig 7. Splice site evolution correlates with both lineage and, independently, genome size.

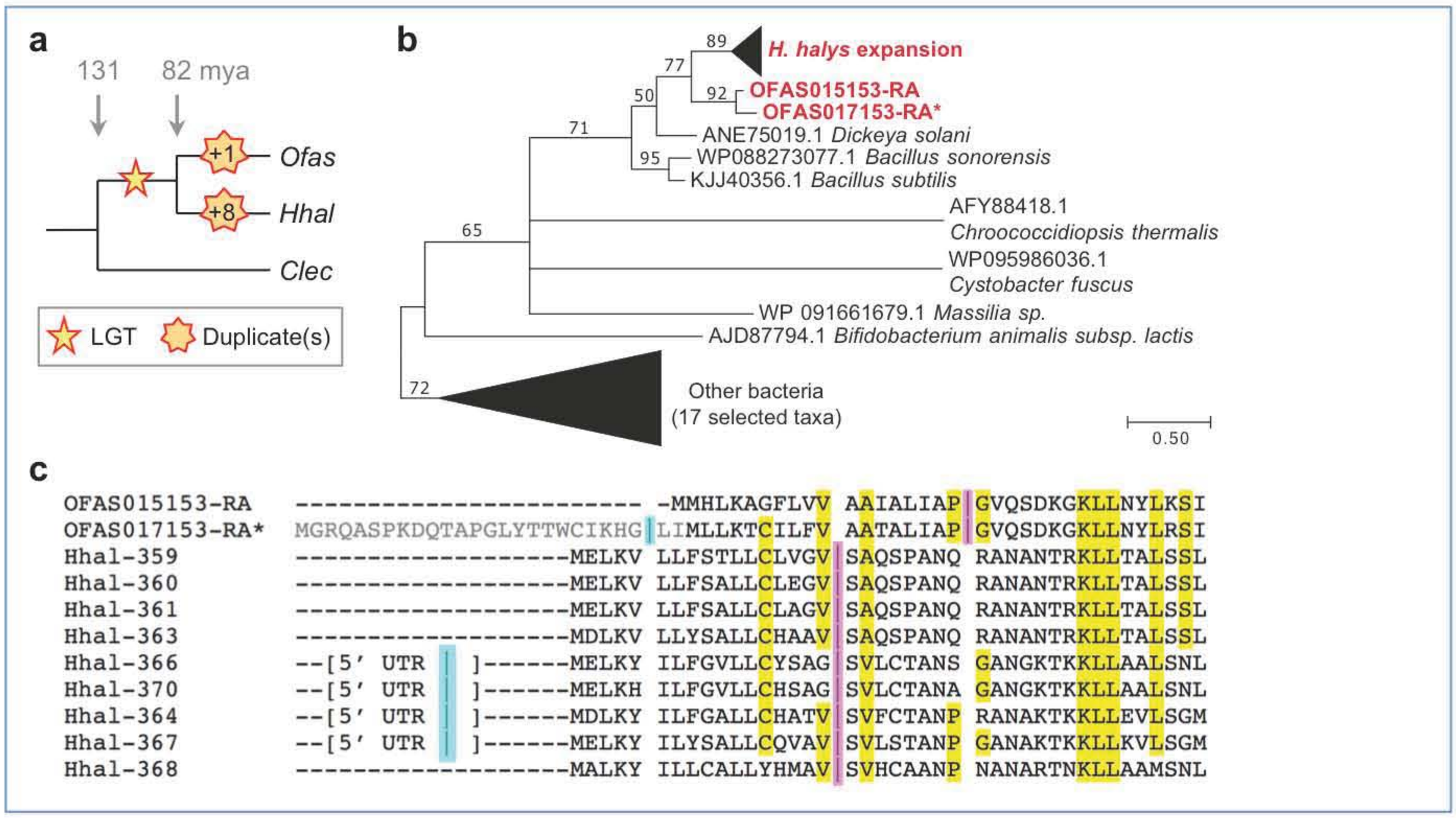
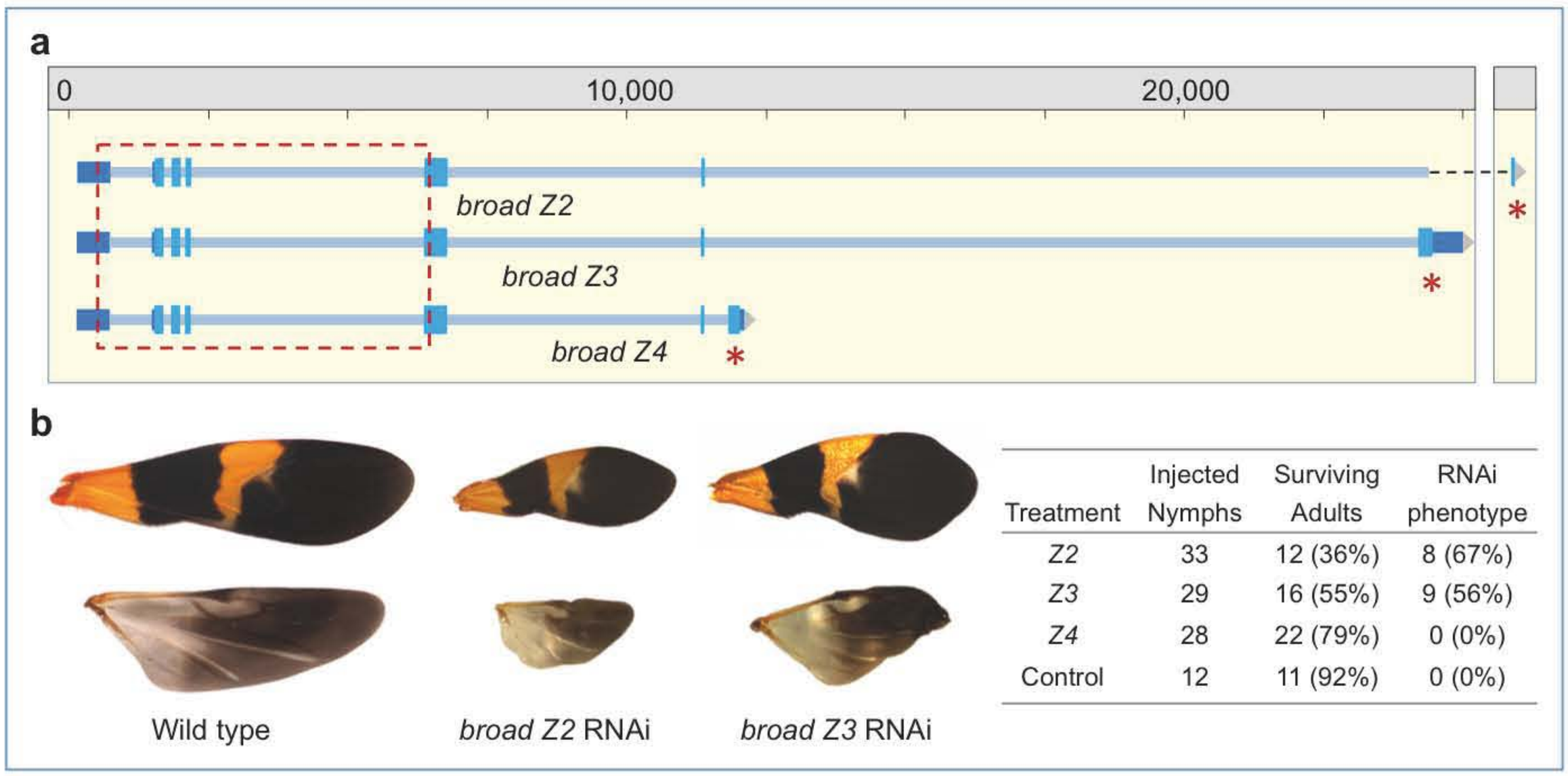


Fig 8. Lateral gene transfer introduction and subsequent evolution within the Hemiptera for mannosidase-encoding genes.



Treatment	Injected Nymphs	Surviving Adults	RNAi phenotype
Z2	33	12 (36%)	8 (67%)
Z3	29	16 (55%)	9 (56%)
Z4	28	22 (79%)	0 (0%)
Control	12	11 (92%)	0 (0%)

Fig 9. Isoform-specific RNAi based on new genome annotations affects the molting and cuticle identity gene *broad*.

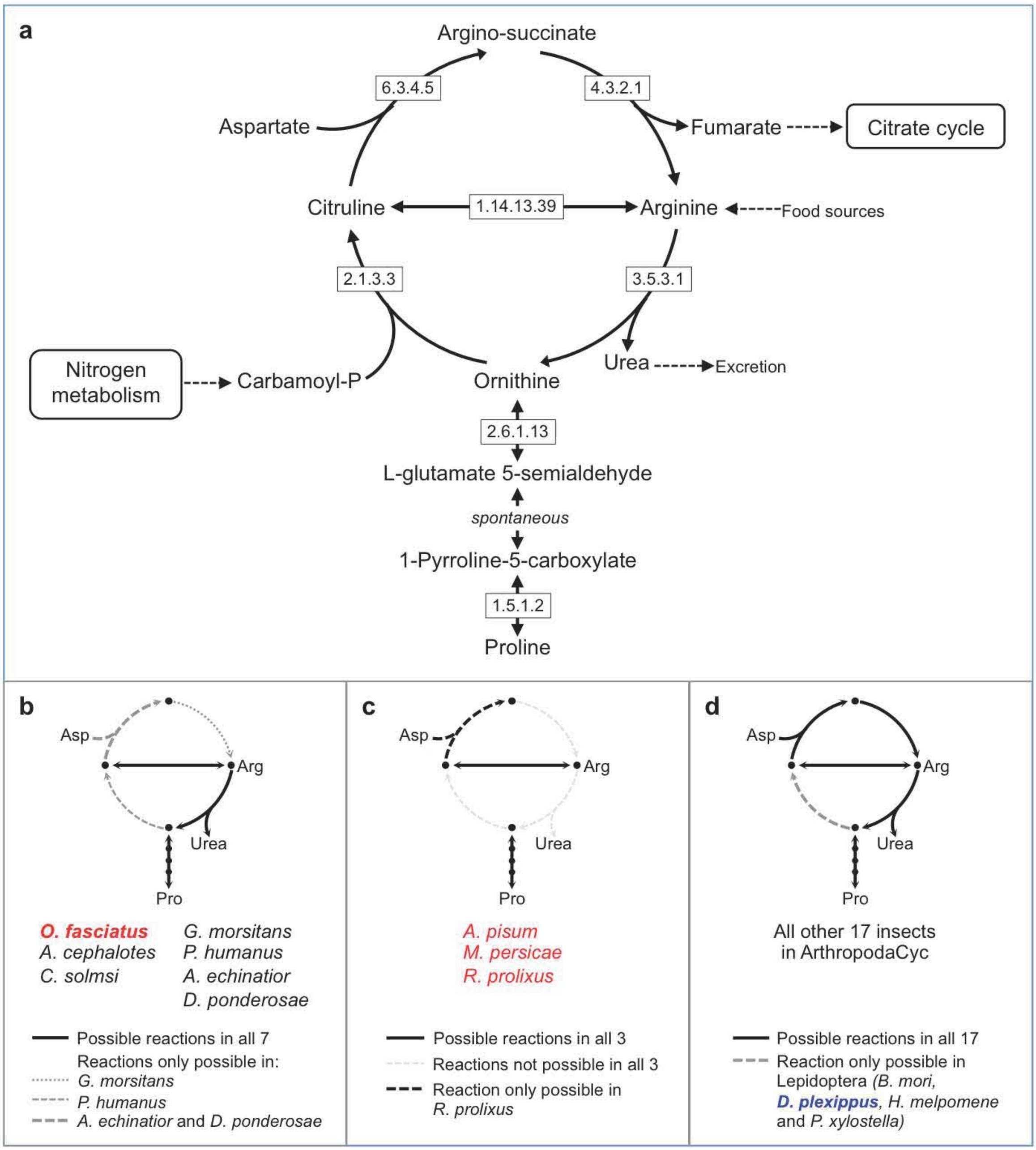


Fig 10. Comparison of the urea cycle of *Oncopeltus* with 26 other insect species.