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], Christopher J. Holmes[3], 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

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, with high diversity for feeding ecology and excellent experimental tractability for molecular genetics. Building upon recent sequencing of hemipteran pests such as phloem-feeding aphids and blood-feeding bed bugs, we present the genome sequence and comparative analyses centered on the milkweed bug *Oncopeltus fasciatus*, a seed feeder of the family Lygaeidae.

Results:
The 926-Mb *Oncopeltus* genome is well represented by the current assembly and official gene set. We use our genomic and RNA-seq data not only to characterize 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. The discovery of intron gain and turnover specific to the Hemiptera also prompted evaluation of lineage and genome size as predictors of gene structure evolution. Furthermore, we identify enzymatic gains and losses that correlate with feeding biology, particularly for reductions associated with derived, fluid-nutrition feeding.

Conclusions:
With the milkweed bug, we now have a critical mass of sequenced species for a hemimetabolous insect order and close outgroup to the Holometabola, substantially improving the diversity of insect genomics. We thereby define commonalities among the Hemiptera and delve into how hemipteran genomes reflect distinct feeding ecologies. Given *Oncopeltus*'s strength as an experimental model, these new sequence resources bolster the foundation for molecular research and highlight technical considerations for the analysis of medium-sized invertebrate genomes.

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

BACKGROUND

The number of animals with sequenced genomes continues to increase 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 dramatic metamorphosis during a pupal stage. The Holometabola represent only a fraction of the full morphological and ecological diversity across the Insecta: over half of all orders are hemimetabolous. 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 comparative analyses based on genome sequencing of the large milkweed bug, *Oncopeltus fasciatus*, as a hemimetabolous representative of the larger diversity of insects.

Oncopeltus is a member of the Hemiptera, the most species-rich hemimetabolous order. Together with the Thysanoptera and, traditionally, the Psocodea, the Hemiptera form the hemipteroid assemblage (or Acercaria), a close outgroup to the Holometabola [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 vertebrate 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, [5]), including phloem-feeding aphids [6-8], psyllids [9], and planthoppers [10], and the hematophagous kissing bug, *Rhodnius prolixus* [11], a vector of Chagas disease, and bed bug, *Cimex lectularius* [12, 13]. 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* [14, 15].

The milkweed bug has feeding ecology traits that are both conservative and complementary to those of previously sequenced hemipterans. Its phytophagy is ancestral for the large infraorder Pentatomomorpha and representative of most extant Hemiptera [16]. Moreover, as a seed feeder *Oncopeltus* has not undergone the marked life style changes associated with fluid feeding (mucivory or hematophagy), including dependence on endosymbiotic bacteria to provide nutrients lacking in the diet. Gene loss in the pea aphid, *Acyrthosiphon pisum*, makes it reliant on the obligate endosymbiont *Buchnera aphidicola* for synthesis of essential amino acids [6, 17]. Although hematophagy arose independently in *Rhodnius* and *Cimex* [16], their respective endosymbionts, *Rhodococcus rhodnii* and *Wolbachia*, must provide vitamins lacking in a blood diet [18]. In contrast, the seed-feeding subfamily Lygaeinae, including *Oncopeltus*, is notable for the absence of prominent endosymbiotic anatomy: these bugs lack both the midgut crypts that typically house bacteria and the bacteriomes and endosymbiotic balls seen in other Lygaeidae [19].

As the native food source of *Oncopeltus* is the toxic milkweed plant, its own feeding biology has a number of interesting implications regarding detoxification and sequestration of cardenolide compounds. A prominent consequence of this diet is the bright red-orange aposematic (warning) coloration seen in *Oncopeltus* embryos, nymphs, and adults [20, 21]. Thus, diet, metabolism, and body pigmentation are functionally linked biological features for which one may expect changes in gene

repertoires to reflect diversity within an 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 [21-23]. It is among the few experimentally tractable hemimetabolous insect species, and it is amenable to a range of molecular techniques (*e.g.*, [24-26]). In fact, it was one of the first insect species to be functionally investigated by RNA interference (RNAi, [27]). 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 [23]).

Here, we focus on these two themes – feeding biology diversity within the Hemiptera and *Oncopeltus* as a research model for macroevolutionary genetics. Key insights derive from a combination of global comparative genomics and detailed computational analyses that are supported by extensive manual curation, empirical data for gene expression, sequence validation, and new isoform-specific RNAi. We thereby identify genes with potentially restricted life history expression in *Oncopeltus* and that are unique to the Hemiptera, clarify evolutionary patterns of zinc finger protein expansion, categorize 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 [28, 29]. 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 (Supplemental Note 2.1.a). The assembly thus contains 84% of the expected sequence, which is comparable to other recent, medium-sized insect genomes [12, 30]. 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 high heterozygosity and repetitive content ([31], Supplemental Note 2.1.b). In further analyses we indeed obtained high estimates of repetitive content, although heterozygosity does not unduly influence gene prediction (see below, based on protein orthology assessments). These computationally challenging features may be increasingly relevant as comparative genomics extends to insect species with larger genomes (>1 Gb) – a common feature among hemimetabolous insects [5, 32].

As template DNA was prepared from dissected adults from which gut material was removed, the resulting assembly is essentially free of contamination. Only five small scaffolds had high bacterial homology, each to a different, partial bacterial genome (Supplemental Note 2.2).

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 exons of the gene are split across scaffolds.

The official gene set and conserved gene linkage

The official gene set (OGS) was generated by automatic annotation followed by manual curation in a large-scale effort by the research community (Supplemental Notes Sections 3-4). Curation revised automatic models, added alternative isoforms and *de novo* models, and documented multiple models for genes split across scaffolds. We found that automatic predictions were rather conservative for hemipteran gene structure (see below). Thus, manual curation often extended gene loci as exons were added, including merging discrete automatic models (Supplemental Note 4, Table S4.4). The OGS v1.1 was generated for global analyses to characterize the gene repertoire. The latest version, OGS v1.2, primarily adds chemoreceptor genes of the ionotropic and odorant receptor classes and genes encoding metabolic enzymes. Altogether, the research community curated 1,697 gene models (8.7% of OGS v1.2), including 316 *de novo* models (Table S4.1, Supplemental Notes Section 5). The majority of curated models are for genes encoding cuticular proteins (11%), chemoreceptors (19%), and developmental regulators such as transcription factors and signaling pathway components (40%, including the BMP/TGF-β, Toll/NF-κB, Notch, Hedgehog, Torso RTK, and Wnt pathways).

In addition to assessing gene model quality, manual curation of genes whose orthologs are expected to occur in syntenic clusters also validates assembly scaffolding. Complete loci could be found for single orthologs of all Hox cluster genes, where *Hox3*/*zen* and *Hox4*/*Dfd* are linked in the current assembly and have ≥99.9% nucleotide identity with experimentally validated sequences ([33-35], Supplemental Note 5.1.b). Conserved linkage was also confirmed for the homeobox genes of the Iroquois complex, the Wnt ligands *wingless* and *wnt10*, and two linked pairs from the Runt transcription factor complex (Supplemental Notes 5.1.a, 5.1.c, 5.1.i, 5.1.j). Further evidence for correct scaffold assembly comes from the curation

205 of large, multi-exonic loci. For example, the cell polarity and cytoskeletal regulator
206 encoded by the conserved *furry* gene includes 47 exons spanning a 437-kb locus,
207 which were all correctly assembled on a single scaffold.
208
209
210 **Gene expression profiles across the milkweed bug life cycle**
211 To augment published transcriptomic resources [36, 37], we sequenced three different
212 post-embryonic samples ("i5K" dataset, see Methods). We then compared the OGS
213 to the resulting *de novo* transcriptome and to a previously published embryonic and
214 maternal (ovary) transcriptome ("454" pyrosequencing dataset, [36]). Our OGS is
215 quite comprehensive, containing 90% of transcripts from each transcriptomic dataset
216 (Fig. 2a). The OGS also contains an additional 3,146 models (16% of OGS) not
217 represented in either transcriptome, including 163 *de novo* models encoding
218 chemoreceptors. Such genes are known for lineage-specific expansions and highly
219 tissue- and stage-specific expression ([38, 39], and see below), and our OGS captures
220 these genes with rare transcripts.
221
222 The OGS also incorporates many partial and unidentified 454 transcripts,
223 nearly trebling the transcripts with an assigned gene model or homology compared to
224 the original study (from 9% to 26%, by blastn, $e<10^{-9}$). This included 10,130
225 transcripts that primarily mapped to UTRs and previously lacked recognizable coding
226 sequence, such as for the *Oncopeltus brinker* ortholog, a BMP pathway component
227 ([40], Supplemental Note 5.1.f), and the enzyme-encoding genes *CTP synthase* and
228 *roquin*. At the same time, the transcriptomes provided expression support for the
229 identification of multiple isoforms in the OGS, such as for the germline determinant
230 *nanos* [36]. More generally, most OGS gene models have expression support (91% of
231 19,690), with 74% expressed broadly in at least three of four samples (Fig. 2b). The
232 inclusion of a fifth dataset from a published adult library [37] provided only a 1%
233 gain in expression support, indicating that with the current study the expression data
234 volume for *Oncopeltus* is quite complete.
235
236 RNA-seq studies were further conducted to establish male, female, and nymph
237 specific gene sets (Fig. 2b-c, Supplemental Note 2.4), from which we also infer that
238 the published adult dataset of unspecified sex is probably male. Moreover, most
239 genes with stage-restricted or stage-enriched expression are in our male sample (Fig.
240 2b-c). For example, gustatory receptor (GR) genes show noticeable restriction to the
241 adult male and published adult (probable male) samples (n= 169 GRs: 40% no
242 expression, 27% only expressed in these two samples), with half of these expressed in
243 both biological replicates (52%). Interestingly, the nymphal sample is enriched for
244 genes encoding structural cuticular proteins (94%, which is >56% more than any
245 other sample). This likely reflects the ongoing molting cycles, with their cyclical
246 upregulation of chitin metabolism and cuticular gene synthesis [41], that are
247 experienced by the different instars and molt cycle stages of individuals pooled in this
248 sample. Lastly, gene sets with sex-specific enrichment across several hemipteroid
249 species substantiate known aspects of male and female reproduction (Fig. 2c: serine-
250 threonine kinases [42] or vitellogenin and other factors associated with oocyte
251 generation, respectively). Some of these enriched genes have unknown functions and
252 could comprise additional, novel factors associated with reproduction in *Oncopeltus*.
253
254

Protein orthology and hemipteran copy number comparisons

To further assay protein-coding gene content, we compared *Oncopeltus* with other arthropods. A phylogeny based on strictly conserved single copy orthologs correctly reconstructs the hemipteran and holometabolan clades' topologies (Fig. 3a, compare with Fig. 1a), although larger-scale insect relationships remain challenging [3].

We then expanded our appraisal to the Benchmarking Universal Single-Copy Orthologs dataset of 1,658 Insecta genes (BUSCO v3, [43]). Virtually all BUSCO genes are present in the *Oncopeltus* OGS (98.9%, Fig. 3b, Supplemental Note 6.1). Although some genes are fragmented, the assembly has a high level of BUSCO completeness (94.6%), independent of annotation prediction limitations that missed some exons from current gene models. Furthermore, BUSCO assessments can elucidate potential consequences of high heterozygosity, which could result in the erroneous inclusion of multiple alleles for a single gene. In fact, the fraction of duplicated BUSCO genes in *Oncopeltus* (1.4%) is low, compared to both the well-assembled bed bug genome (2.2%, [12]) and the pea aphid (4.8%), which is known to have lineage-specific duplications [6, 44]. Thus, by these quality metrics the *Oncopeltus* OGS and assembly are comparable to those of fellow hemipterans, strongly supporting the use of these resources in further comparisons.

We next categorized all proteins by conservation in global, clustering-based orthology analyses (OrthoDB, [1, 45]). As in most species, half of *Oncopeltus* proteins are highly conserved (Fig. 3a). Moreover, 98% of all *Oncopeltus* protein-coding genes have homology, expression, and/or curation support (Fig. 3c). Proteins without homology include species-specific chemoreceptors and antimicrobial peptides (Supplemental Note 5.1.h), as well as potentially novel or partial models. Overall, we estimate that the *Oncopeltus* protein repertoire is comparable to that of other insects in size and conservation. For the Hemiptera, *Oncopeltus* also has fewer missing orthology groups than either the kissing bug or pea aphid (Table S6.1). Indeed the pea aphid is a notable outlier, with its long branch in the phylogeny and for its large protein-coding gene content with low conservation (Fig. 3a). As more hemipteran genomes are sequenced, other species now offer less derived alternatives for phylogenomic comparisons.

Compared to the pea aphid [44], *Oncopeltus* is more conservative in presence and copy number for several signaling pathway components. In contrast to gene absences in the pea aphid, *Oncopeltus* retains orthologs of the EGF pathway component *sprouty*, the BMP receptor *wishful thinking*, and the hormone nuclear receptor *Hr96* (Supplemental Note 5.1.e). Also, whereas multiple copies were reported for the pea aphid, we find a single *Oncopeltus* ortholog for the BMP pathway components *decapentaplegic* and *Medea* and the Wnt pathway intracellular regulator encoded by *shaggy/GSK-3*, albeit with five potential isoforms of the latter (Supplemental Notes 5.1.f, 5.1.j). Duplications of miRNA and piRNA gene silencing factors likewise seem to be restricted to the pea aphid, even compared to other aphid species ([46], Supplemental Note 5.4.a). However, our survey of *Oncopeltus* and other hemimetabolous species reveals evidence for frequent, independent duplications of the Wnt pathway component *armadillo/β-catenin* ([47], Supplemental Note 5.1.j). Curiously, *Oncopeltus* appears to encode fewer histone loci than any other arthropod genome and yet exhibits a similar, but possibly independent, pattern of duplications of

histone acetyltransferases to those previously identified in *Cimex* and the pea aphid (Supplemental Note 5.4.c).

On the other hand, we documented several notable *Oncopeltus*-specific duplications. For the BMP transducer *Mad*, we find evidence for three paralogs in *Oncopeltus*, where two occur in tandem and may reflect a particularly recent duplication (Supplemental Note 5.1.f). Similarly, a tandem duplication of *wnt8* appears to be unique to *Oncopeltus* (Supplemental Note 5.1.j). More striking is the identification of six potential paralogs of *cactus*, a member of the Toll/NF-κB signaling pathway for innate immunity, whereas the bed bug and kissing bug each retain only a single copy ([48], Supplemental Note 5.1.g).

Lastly, we explored hemipteran-specific orthology groups against a backdrop of 107 other insect species [1]. What makes a bug a bug in terms of protein-coding genes? Several orthogroups contain potentially novel genes that show no homology outside the Hemiptera and await direct experimental analysis, for which the Hemiptera are particularly amenable (*e.g.*, [49-52, reviewed in 5]). Secondly, there are hemipteran-specific orthogroups of proteins with recognized functional domains and homologs in other insects, but where evolutionary divergence has led to lineage-specific subfamilies. One example is a heteropteran-specific cytochrome P450 (CYP) enzyme (EOG090W0V4B), which in *Oncopeltus* is expressed in all life history stages (Fig. 2b). The expansion of CYP proteins is associated with potential insecticide resistance, as specific P450s can confer resistance to specific chemicals (*e.g.,* [53, 54]; Supplemental Notes 5.3.b, 5.3.c). Hence, the identification of lineage-specific CYP enzymes can suggest potential targets for integrated pest management approaches.

Transcription factor repertoires and homeobox gene evolution

Having explored the global protein repertoire, we next focused specifically on transcription factors (TFs), which comprise a major class of proteins that has been extensively studied in *Oncopeltus*. This is a class of key regulators of development whose functions can diverge substantially during evolution and for which RNAi-based experimental investigations have been particularly fruitful in the milkweed bug (*e.g.*, [33, 34, 55-57], Supplemental Notes 5.1.a-e).

To systematically evaluate the *Oncopeltus* TF repertoire, we used a pipeline to scan all predicted proteins and assign them to TF families, including orthology assignments where DNA binding motifs could be predicted (see Methods, [58]). We identified 762 putative TFs in *Oncopeltus*, which is similar to other insects for total TF count and for the size of each TF family (Fig. 4a: note that the heatmap also reflects the large, duplicated repertoire in the pea aphid, see also Tables S6.3-S6.5).

We were able to infer DNA binding motifs for 25% (n=189) of *Oncopeltus* TFs, mostly based on data from *Drosophila melanogaster* (121 TFs) but also from distantly related taxa such as mammals (56 TFs). Such high conservation is further reflected in explicit orthology assignments for most proteins within several large TF families, including the homeodomain (53 of 85, 62%), basic helix-loop-helix (bHLH, 35 of 45, 78%), and forkhead box (16 of 17, 94%) families. In contrast, most C2H2 zinc finger proteins lack orthology assignment (only 22 of 360, 6%). Across species, the homeodomain and C2H2 zinc finger proteins are the two largest TF superfamilies

354 (Fig. 4a). Given their very different rates of orthology assignment, we probed further
355 into their pipeline predictions and the patterns of evolutionary diversification.
356
357 The number of homeodomain proteins identified by the pipeline displays a
358 narrow normal distribution across species (Fig. 4b, mean ± standard deviation: 97 ±
359 9), consistent with a highly conserved, slowly evolving protein family. Supporting
360 this, many *Oncopeltus* homeodomain proteins that were manually curated also
361 received a clear orthology assignment (Fig. 4c: pink), with only four exceptions (Fig.
362 4c: yellow). Only one case suggests a limitation of a pipeline that is not specifically
363 tuned to hemipteran proteins (Goosecoid). Manual curation of partial or split models
364 identified a further 11 genes encoding homeodomains, bringing the actual tally in
365 *Oncopeltus* to 96. Overall, we find the TF pipeline results to be a robust and
366 reasonably comprehensive representation of these gene classes in *Oncopeltus*.
367
368 These analyses also uncovered a correction to the published *Oncopeltus*
369 literature for the developmental patterning proteins encoded by the paralogs *engrailed*
370 and *invected*. These genes arose from an ancient tandem duplication prior to the
371 hexapod radiation. Their tail-to-tail orientation enables ongoing gene conversion
372 [59], making orthology discrimination particularly challenging. For *Oncopeltus*, we
373 find that the genes also occur in a tail-to-tail orientation and that *invected* retains a
374 diagnostic alternative exon [59]. These new data reveal that the purported *Oncopeltus*
375 *engrailed* ortholog in previous developmental studies (*e.g.*, [55, 60-63]) is in fact
376 *invected* (Supplemental Note 5.1.a).
377
378
379 **Independent expansions of C2H2 zinc fingers within the Hemiptera**
380 Unlike homeodomain proteins, C2H2 zinc finger (C2H2-ZF) repertoires are
381 prominent for their large family size and variability throughout the animal kingdom
382 [64], and this is further supported by our current analysis in insects. With >350
383 C2H2-ZFs, *Oncopeltus*, the pea aphid, termite, and some mosquito species have 1.5×
384 more members than the insect median (Fig. 4b). This is nearly half of all *Oncopeltus*
385 TFs. While the expansion in mosquitoes could have a single origin within the
386 Culicinae, the distribution in the Hemiptera, where *Cimex* has only 227 C2H2-ZFs,
387 suggests that independent expansions occurred in *Oncopeltus* and the pea aphid. Prior
388 to the sequencing of other hemipteran genomes, the pea aphid's large C2H2-ZF
389 repertoire was attributed to the expansion of a novel subfamily, APEZ, also referred
390 to as zinc finger 271-like [44].
391
392 In fact, manual curation in *Oncopeltus* confirms the presence of a subfamily
393 with similar characteristics to APEZ (Fig. 4c: yellow fraction). In *Oncopeltus* we find
394 >115 proteins of the ZF271 class that are characterized by numerous tandem repeats
395 of the C2H2-ZF domain and its penta-peptide linker, with 3-45 repeats per protein.
396
397 Intriguingly, we find evidence for ongoing evolutionary diversification of this
398 subfamily. A number of *Oncopeltus* ZF271-like genes occur in tandem clusters of 4-
399 8 genes – suggesting recent duplication events. Yet, clustered genes differ in gene
400 structure (number and size of exons), and we identified a number of probable ZF271-
401 like pseudogenes whose open reading frames have become disrupted – consistent with
402 high turnover. *Oncopeltus* ZF271-like proteins also differ in the sequence and length
403 of the zinc finger domains among themselves and compared to aphid proteins

404 (WebLogo analysis, [65]), similar to zinc finger array shuffling seen in humans [66].
405 Furthermore, whole-protein phylogenetic analysis supports independent, rapid
406 expansions in the pea aphid and *Oncopeltus* (Fig. 4d).
407
408 Clustered zinc finger gene expansion has long been recognized in mammals,
409 with evidence for strong positive selection to increase both the number and diversity
410 of zinc finger domains per protein as well as the total number of proteins [67]. This
411 was initially found to reflect an arms-race dynamic of co-evolution between selfish
412 transposable elements and the C2H2-ZF proteins that would repress them [68]. In
413 vertebrates, these C2H2-ZF proteins bind to the promoters of transposable elements
414 via their zinc finger arrays and use their Krüppel-associated box (KRAB) domain to
415 bind the chromatin-remodeling co-repressor KAP-1, which in turn recruits
416 methyltransferases and deacetylases that silence the targeted promoter [69].
417
418 Insects do not have a direct ortholog of vertebrate KAP-1 (Supplemental Note
419 5.4.d), and neither the aphid nor *Oncopeltus* ZF271-like subfamilies possess a KRAB
420 domain or any other domain besides the zinc finger arrays. However, close molecular
421 outgroups to this ZF271-like subfamily include the developmental repressor Krüppel
422 [70] and the insulator protein CTCF [71] (data not shown). Like these outgroups, the
423 *Oncopeltus* ZF271-like genes are strongly expressed: 98% have expression support,
424 with 86% expressed in at least three different life history stages (Fig. 2b). Thus, the
425 insect ZF271-like proteins may also play prominent roles in repressive DNA binding.
426 Indeed, we find evidence for a functional methylation system in *Oncopeltus*
427 (Supplemental Note 5.4.c), like the pea aphid, which would provide a means of gene
428 silencing by chromatin remodeling, albeit via mediators other than KAP-1.
429
430 However, an arms race model need not be the selective pressure that favors
431 insect ZF271-like family expansions. Recent analyses in vertebrates identified
432 sophisticated, additional regulatory potential by C2H2-ZF proteins, building upon
433 original transposable element binding for new, lineage-specific and even positive
434 gene regulation roles [66, 72, 73]. Moreover, although *Cimex* has half as many long
435 terminal repeat (LTR) repetitive elements as *Oncopeltus* and the pea aphid, overall we
436 do not find a correlation between relative or absolute repetitive content and ZF271-
437 like family expansion within the Hemiptera (see next section).
438
439
440 **Proportional repeat content across hemipterans**
441 With the aim of reducing assembly fragmentation and to obtain a better picture of
442 repeat content, we performed low coverage, long read PacBio sequencing in
443 *Oncopeltus* (Supplemental Note 2.3). Using PacBio reads in a gap-filling assay on
444 the Illumina assembly raised the total detected repetitive content from 25% to 32%,
445 while repeat estimations based on simultaneous assessment of Illumina and PacBio
446 reads nearly doubled this value to 58%. As expected, the capacity to identify repeats
447 is strongly dependent on assembly quality and sequencing technology, with the
448 *Oncopeltus* repetitive content underrepresented in the current (Illumina-only)
449 assembly. Furthermore, as increasing genome size compounds the challenge of
450 assembling repeats, the repeat content of the current assembly is lower than in species
451 with smaller genome sizes (Fig. 5a, with the sole exception of the honey bee), and we
452 therefore used our gap-filled dataset as a more accurate basis for further comparisons.
453

454 To support direct comparisons among hemipterans, we also performed our
455 RepeatModeler analysis on the bed bug and pea aphid assemblies. Repeats comprised
456 36% and 31% of the respective assemblies, similar to the gap-filled value of 32% in
457 *Oncopeltus*. Nevertheless, given the smaller sizes of these species' assemblies – 651
458 Mb in the bed bug and 542 Mb in the pea aphid – the absolute repeat content is much
459 higher in *Oncopeltus* (Fig. 5b). Excluding unknown repeats, the most abundant
460 transposable elements in *Oncopeltus* are LINE retrotransposons, covering 10% of the
461 assembly (Table S2.5). This is also the case in the bed bug (12%), while in the pea
462 aphid DNA transposons with terminal inverted repeats (TIRs) are the most abundant
463 (2% of the assembly identified here, and 4% reported from manual curation in the pea
464 aphid genome paper, [6]). Across species, the remaining repeat categories appear to
465 grow proportionally with assembly size, except for simple repeats, which were the
466 category with the largest relative increase in size after gap-filling in *Oncopeltus*
467 (Supplemental Note 2.3). However, given the mix of data types (Illumina [12] and
468 Sanger [6]), these patterns should be treated as hypotheses for future testing.
469
470
471 **Lineage and genome size-related trends in insect gene structure**
472 Both our manual curation work and BUSCO analyses highlighted the fact that
473 *Oncopeltus* genes are often comprised of many, small exons. We thus undertook a
474 comparative analysis to determine whether this is a general feature to be considered
475 for structural annotation of hemipteran genomes. We find that both lineage and
476 genome size can serve as predictors of gene structure.
477
478 Firstly, we created a high quality dataset of 30 functionally diverse, large
479 genes whose manual curation could reasonably ensure complete gene models across
480 seven species from four insect orders (Fig. 6a, Supplemental Note 6.3). Most species
481 encode the same total number of amino acids for these conserved proteins, with the
482 thrips *Frankliniella occidentalis* and the fruit fly being notable exceptions with larger
483 proteins (Fig. 6a: blue plot line). However, the means of encoding this information
484 differs between lineages, with hemipteroid orthologs comprised of twice as many
485 exons as their holometabolous counterparts (Fig. 6a: orange plot line). Thus, there is
486 an inverse correlation between exon number and exon size (Fig. 6a: orange vs. red
487 plot lines). This analysis corroborates and extends previous probabilistic estimates of
488 intron density, where the pea aphid as a sole hemipteran representative had the
489 highest intron density of ten insect species [74].
490
491 To test these trends, we next expanded our analysis to all manually curated
492 exons in two species from each of three orders (Hemiptera, Coleoptera, Diptera).
493 Here, we expect that curated exon sizes are accurate, without the need to assume that
494 entire gene models are complete. This large dataset corroborates our original
495 findings, with bugs having small exons while both the median and Q3 quartile reflect
496 larger exon sizes in beetles and flies (Fig. 6b). Notably, the median and median
497 absolute deviation are highly similar between species pairs within the Hemiptera and
498 Coleoptera. Meanwhile, the exon metrics within the Diptera support large protein
499 sizes as a drosophilid-specific, rather than dipteran-wide, feature.
500
501 Does the high exon count in the Hemiptera reflect an ancient, conserved
502 increase at the base of this lineage, or ongoing remodeling of gene structure with high
503 turnover? To assess the exact nature of evolutionary changes, we annotated intron

504 positions within multiple sequence alignments of selected proteins and plotted gains
505 and losses onto the phylogeny, providing a total sample of 165 evolutionary changes
506 at 148 discrete splice sites (Fig. 7, see also Supplemental Note 6.3 for gene selection
507 and method). These data reveal several major correlates with intron gain or loss. The
508 bases of both the hemipteroid and hemipteran radiations show the largest gains, while
509 most losses occur in the dipteran lineage (Fig. 7: orange and purple shading,
510 respectively). Furthermore, we find progressive gains across hemipteroid nodes, and
511 it is only in this lineage that we additionally find species-specific splice changes for
512 the highly conserved *epimerase* gene (Fig. 7: orange outline). Thus, we find evidence
513 for both ancient intron gain and ongoing gene structure remodeling in this lineage.
514
515 Surprisingly, both *hemocytin* and *epimerase* – our exemplar genes with many
516 (up to 74) and few exons (3-8 per species), respectively – show independent losses of
517 the same splice sites in *Drosophila* and *Tribolium*. One feature these species share is
518 a genome size 2.4-6.0× smaller than in the other species examined here (Fig. 7: red
519 shading). Pairwise comparisons within orders also support this trend, as the beetle
520 and fly species with larger genomes exhibit species-specific gains compared to intron
521 loss in their sister taxa (Fig. 7: red outlines). Thus, genome size seems to positively
522 correlate with intron number. However, lineage is a stronger predictor of gene
523 structure: the coleopteran and dipteran species pairs have highly similar exon size
524 metrics despite differences in genome size (Fig. 6b). A global computational analysis
525 over longer evolutionary distances also supports a link between genome size and
526 intron number in arthropods, although chelicerates and insects may experience
527 different rates of evolutionary change in these features [75]. It will be interesting to
528 see if the correlation with genome size is borne out in other invertebrate taxa.
529
530 The selective pressures and mechanisms of intron gain in the Hemiptera will
531 be a challenge to uncover. While median exon size (Fig. 6b) could reflect species-
532 specific nucleosome sizes [76, 77], this does not explain why only the Hemiptera
533 seldom exceed this (Fig. 6b: Q3 quartile). Given gaps in draft genome assemblies, we
534 remain cautious about interpreting (large) intron lengths but note that many
535 hemipteran introns are too small to have harbored a functional transposase gene (*e.g.,*
536 median intron size of 429 bp, n=69 introns in *hemocytin* in *Cimex*). Such small
537 introns could be consistent with proliferation of non-autonomous short interspersed
538 nuclear elements (SINEs). However, characterization of such highly divergent non-
539 coding elements would require curated SINE libraries for insects, comparable to those
540 generated for vertebrates and plants [76, 77]. Meanwhile, it appears that hemipteran
541 open reading frames ≥160 bp are generally prevented by numerous in-frame stop
542 codons just after the donor splice site. Most stop codons are encoded by the triplet
543 TAA in both *Oncopeltus* and *Cimex* (data not shown), although these species'
544 genomes are not particularly AT-rich (Table 1).
545
546 Even if introns are small, having gene loci comprised of numerous introns and
547 exons adds to the cost of gene expression in terms of both transcription duration and
548 mRNA processing. One could argue that a gene like *hemocytin*, which encodes a
549 clotting agent, would require rapid expression in the case of wounding – a common
550 occurrence in adult *Cimex* females due to the traumatic insemination method of
551 reproduction [12]. Thus, as our molecular understanding of comparative insect and
552 particularly hemipteran biology deepens, we will need to increasingly consider how

553 life history traits are manifest in genomic signatures at the structural level (*e.g.*, Figs.
554 5-7), as well as in terms of protein repertoires (Figs. 3-4).
555
556
557 **Expansion after a novel lateral gene transfer (LGT) event in phytophagous bugs**
558 In addition to the need for cuticle repair, traumatic insemination may be responsible
559 for the numerous LGT events predicted in the bed bug [12]. In contrast, the same
560 pipeline analyses [78] followed by manual curation predicted very few LGTs in
561 *Oncopeltus*, which lacks this unusual mating behavior. Here, we have identified 11
562 strong LGT candidates, and we confirmed the incorporation of bacterial DNA into the
563 milkweed bug genome for all five candidates chosen for empirical testing (Table
564 S2.4). Curiously, we find several LGTs potentially involved in bacterial or plant cell
565 wall metabolism that were acquired from different bacterial sources at different times
566 during hemipteran lineage evolution, including two distinct LGTs that are unique to
567 *Oncopeltus* and implicated in the synthesis of peptidoglycan, a bacterial cell wall
568 constituent (Supplemental Note 2.2).
569
570 Conversely, two other validated LGT candidates are implicated in cell wall
571 degradation. We find two strongly expressed, paralogous copies in *Oncopeltus* of a
572 probable bacterial-origin gene encoding an endo-1,4-beta-mannosidase enzyme
573 (MAN4, EC 3.2.1.78). Inspection of genome assemblies and protein accessions
574 reveals that this LGT event occurred after the infraorder Pentatomomorpha, including
575 the stink bug *Halyomorpha halys*, diverged from other hemipterans, including the bed
576 bug (Fig. 8a). Independent duplications then led to the two copies in *Oncopeltus* and
577 an astonishing nine tandem copies in *Halyomorpha* (Figs. 8b, S2.6). Since the
578 original LGT event, the *mannosidase* genes have gained introns that are unique to
579 each species and to subsets of paralogs (Fig. 8c). Thus, the "domestication" [79] of
580 *mannosidase* homologs as multi-exonic genes further illustrates the hemipteran
581 penchant for intron introduction and maintenance of small exons. The retention and
582 subsequent expansion of these genes implies their positive selection, consistent with
583 the phytophagous diet of these species. It is tempting to speculate that copy number
584 proliferation in the stink bug correlates with the breadth of its diet, as this agricultural
585 pest feeds on a number of different tissues in a range of host plants [80].
586
587
588 **Cuticle development, structure, and warning pigmentation**
589 Body cuticle is another trait associated with feeding ecology, particularly for
590 pigmentation. Furthermore, the milkweed bug has been a powerful model for
591 endocrine studies of hemimetabolous molting and metamorphosis since the 1960's
592 [22, 81-84]. Therefore, we next focused on the presence and function of genes
593 involved in cuticle development and structure.
594
595 Molting is triggered by the release of ecdysteroids, steroid hormones that are
596 synthesized from cholesterol by cytochrome P450 enzymes of the Halloween family
597 [85], and we were able to identify these in the *Oncopeltus* genome (Supplemental
598 Notes 5.2.b, 5.3.b). From the ecdysone response cascade defined in *Drosophila* [86],
599 we identified *Oncopeltus* orthologs of both early and late-acting factors, including
600 ecdysteroid hormones and their receptors. It will be interesting to see if the same
601 regulatory relationships are conserved in the context of hemimetabolous molting in
602 *Oncopeltus*. For example, *E75A* is required for reactivation of ecdysteroid production

during the molt cycle in *Drosophila* larvae [87] and likely operates similarly in *Oncopeltus*, since *Of-E75A* RNAi prevents fourth-instar nymphs from molting to the fifth instar (H. Kelstrup and L. Riddiford, unpublished data).

In hemipterans, activation of juvenile hormone (JH) signaling at molts determines whether the insect progresses to another nymphal instar or, if lacking, becomes an adult [50]. We were able to identify many components of the JH signal transduction pathway in the *Oncopeltus* genome, including orthologs of *Methoprene-tolerant* (*Met*), the JH receptor [50, 88], and the JH-response gene *Kr-h1* [50, 89, 90]. JH acts to determine cuticle identity through regulation of the *broad* gene in a wide variety of insects, where different isoforms direct specific aspects of metamorphosis in *Drosophila* [91, 92]. In *Oncopeltus*, *broad* expression directs progression through the nymphal stages [93], but the effect of each isoform was unknown. We identified three isoforms in *Oncopeltus* – *Z2*, *Z3*, and *Z4* – and performed isoform-specific RNAi. In contrast to *Drosophila*, Broad isoform functions appear to be more redundant in *Oncopeltus*, as knockdown of isoforms *Z2* and *Z3* has similar effects on survival to adulthood as well as adult wing size and morphology (Fig. 9).

Regulators such as Broad initiate the transcription of a large battery of genes that encode the structural components of the cuticle needed at each molt, consistent with our expression analyses (Fig. 2b-c, discussed above). We identified 173 genes encoding putative cuticle structural proteins in *Oncopeltus* (Supplemental Note 5.2.c). Similar to other insects, the CPR family, with the RR-1 (soft cuticle), RR-2 (hard cuticle), and unclassifiable types, constituted the largest cuticle protein family. While several protein families are similar in size to those of other insects (CPAP1, CPAP3, and TWDL: Table S5.12), we found a slight expansion in the *Oncopeltus* CPF family (Fig. S5.14). For cuticle production, similar to the bed bug and the Asian longhorned beetle [12, 30], we identified a single *chitin synthase* gene with conserved alternative splice isoforms, which suggests that *chitin synthase 2* is a duplication specific to only certain beetle and fly lineages within the Holometabola [94].

A major characteristic of the milkweed bug is the distinctive red-orange and black aposematic (warning) coloration within the cuticle and epidermis that deters predators (*e.g.*, Figs. 1, 9, [20, 21]). For black coloration, the melanin synthesis pathway known from holometabolous insects (*e.g.*, [95, 96]) is conserved at the sequence (Fig. S5.15) and functional [97, 98] level in *Oncopeltus*, supporting conservation in hemimetabolous lineages as well. In contrast, production of the primary warning coloration, pteridine red erythropterin [99], has not been as extensively studied and remains an open avenue for hemimetabolous research. Pterin pigments are synthesized from GTP through a series of enzymatic reactions [100]. Thus far in *Oncopeltus* we could identify orthologs of *punch*, which encodes a GTP cyclohydrolase [101], and *sepia*, which is required for the synthesis of the red eye pigment drosopterin [102]. The bright red color of *Oncopeltus* eggs may in part reflect chemical protection transmitted parentally [103]. Thus, further identification of pigmentation genes will provide fitness indicators for maternal contributions to developmental success under natural conditions (*i.e.*, the presence of egg predators).

651 **Chemoreception and metabolism in relation to feeding biology**
652 Aposematic pigmentation advertises the fact that toxins in the milkweed seed diet are
653 incorporated into the insects themselves, a metabolic feat that was independently
654 acquired in *Oncopeltus* and the monarch butterfly (*Danaus plexippus*), which shares
655 this food source and body coloration [37, 104]. Moreover, given the fundamental
656 differences between phytophagous, mucivorous, and hematophagous diets, we
657 investigated to what extent differences in feeding ecology across hemipterans are
658 represented in their chemoreceptor and metabolic enzyme repertoires.
659
660
661
662 **Table 2. Numbers of chemoreceptor genes/proteins per family in selected insect**
663 **species.** In some cases the number of proteins is higher than the number of genes due
664 to an unusual form of alternative splicing, which is particularly notable for the
665 *Oncopeltus* GRs. Data are shown for four Hemiptera as well as *Drosophila*
666 *melanogaster*, the body louse *Pediculus humanus*, and the termite *Zootermopsis*
667 *nevadensis* [11, 12, 105-109].
668

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

669
670 [1] Hemiptera: Heteroptera
671 [2] independent acquisitions of hematophagy [16]
672 [3] Hemiptera, phloem-feeding
673
674
675
676 Insects must smell and taste their environment to locate and identify food,
677 mates, oviposition sites, and other essential cues. Perception of the enormous
678 diversity of environmental chemicals is primarily mediated by the odorant (OR),
679 gustatory (GR), and ionotropic (IR) families of chemoreceptors, which each encode
680 tens to hundreds of distinct proteins [109-112]. Chemoreceptor family size appears to
681 correlate with feeding ecology. *Oncopeltus* retains a moderate complement of each
682 family, while species with derived fluid nutrition diets (sap or blood) have relatively
683 depauperate OR and GR families (Table 2, Supplemental Note 5.3.f). In detail, a few
684 conserved orthologs such as the OrCo protein and a fructose receptor are found across
685 species, but other subfamilies are lineage specific. *Oncopeltus* and *Acyrthosiphon*
686 retain a set of sugar receptors that was lost independently in the blood-feeding bugs
687 (*Rhodnius* [11], *Cimex* [12]) and body louse (*Pediculus* [108]). Conversely,
688 *Oncopeltus* and *Cimex* retain a set of candidate carbon dioxide receptors, a gene
689 lineage lost from *Rhodnius*, *Acyrthosiphon*, and *Pediculus* [11, 12, 105], but which is
690 similar to a GR subfamily expansion in the more distantly related hemimetabolous
691 termite (Isoptera, [106]). Comparable numbers of IRs occur across the Heteroptera.
692 In addition to a conserved set of orthologs primarily involved in sensing temperature
693 and certain acids and amines, *Oncopeltus* has a minor expansion of IRs distantly
694 related to those involved in taste in *Drosophila*. The major expansions in each insect

695 lineage are the candidate "bitter" GRs ([113], Supplemental Note 5.3.f, Fig. S5.19).
696 In summary, *Oncopeltus* exhibits moderate expansion of specific subfamilies likely to
697 be involved in host plant recognition, consistent with it being a preferentially
698 specialist feeder with a potentially patchy food source [21, 114].
699
700 As host plant recognition is only the first step, we further explored whether
701 novel features of the *Oncopeltus* gene set are directly associated with its diet. We
702 therefore used the CycADS annotation pipeline [115] to reconstruct the *Oncopeltus*
703 metabolic network. The resulting BioCyc metabolism database for *Oncopeltus*
704 ("OncfaCyc") was then compared with those for 26 other insect species ([116],
705 http://arthropodacyc.cycadsys.org/), including three other hemipterans: the pea aphid,
706 the green peach aphid, and the kissing bug (Tables 3-4). For a global metabolism
707 analysis, we detected the presence of 1085 Enzyme Commission (EC) annotated
708 reactions with at least one protein in the *Oncopeltus* genome (Supplemental Note 6.4,
709 Table S6.10). Among these, 10 enzyme classes (represented by 17 genes) are unique
710 and 17 are missing when compared to the other insects (Table 4, Table S6.11).
711
712 We then specifically compared amino acid metabolism in the four hemipterans
713 representing the three different diets. Eight enzymes are present only in *Oncopeltus*
714 (Table 4), including the arginase that degrades arginine (Arg) into urea and ornithine,
715 a precursor of proline (Pro). Given this difference, we extended our analysis to assess
716 species' repertoires for the entire urea cycle (Fig. 10a, Table S6.13). *Oncopeltus* and
717 six other species can degrade Arg but cannot synthesize it (Fig. 10b). Only the other
718 three hemipterans can neither synthesize nor degrade Arg via this cycle (Fig. 10c),
719 while most species have an almost complete cycle (Fig. 10d). This suggests that the
720 ability to synthesize Arg was lost at the base of the Hemiptera, with subsequent,
721 independent loss of Arg degradation capacity in the aphid and *Rhodnius* lineages.
722 Retention of Arg degradation in *Oncopeltus* might be linked to the milkweed seed
723 food source, as most seeds are very rich in Arg [117], and Arg is indeed among the
724 metabolites detected in *Oncopeltus* [118]. However, the monarch butterfly is one of
725 only a handful of species that retains the complete Arg pathway (Fig. 10d: blue text).
726 Despite a shared food source, these species may therefore differ in their overall Arg
727 requirements, or – in light of a possible group benefit of *Oncopeltus* aggregation
728 during feeding ([21]; *e.g.*, Fig. 1b) – in their efficiency of Arg uptake.
729
730 Other enzymes are also present only in the milkweed bug compared to the
731 other examined hemipterans (Table S6.12). Like other insects [116], *Oncopeltus*
732 retains the ability to degrade tyrosine (Tyr). This pathway was uniquely lost in the
733 aphids, where this essential amino acid is jointly synthesized – and consumed – by the
734 aphid host and its endosymbiotic bacteria [6, 7, 17, 119]. Conversely, a gain specific
735 to the milkweed bug lineage was the duplication of the Na+/K+ ATPase alpha
736 subunits whose amino acid substitutions confer resistance to milkweed cardenolides
737 [37, 120]. In the *Oncopeltus* genome, we find support for the recent nature of these
738 duplications: the genes encoding subunits ATPα1B and ATPα1C occur as a tandem
739 duplication, notably on a scaffold that also harbors one of the clustered ZF271-like
740 gene expansions (see above).
741
742

743
744
745 **Table 3. Hemipteran ArthropodaCyc database summaries.**
746 Overview statistics for the newly created database for *Oncopeltus fasciatus* (Ofas) in
747 comparison with public databases for *Rhodnius prolixus* (Rpro), *Acyrthosiphon pisum*
748 (Apis), and *Myzus persicae* (Mper) available from [116]. Based on OGS v1.1.
749

Species ID Gene set ID	*Ofas* OGS v1.1	*Rpro* RproC1.1 (Built on RproC1 assembly)	*Apis* OGS v2.1b (Built on Acyr_2.0 assembly)	*Mper* Clone G006 v1.0	*Mper* 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

750
751 [1] In the BioCyc databases all splice variants are counted in the summary tables for genes.
752
753
754
755
756 **Table 4. Hemipteran ArthropodaCyc annotations of metabolic genes.**
757 Taxonomic abbreviations are as in Table 3.
758

	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

759
760 [1] "EC" refers to the number of proteins, as represented by their unique numerical designations
761 within the Enzyme Commission (EC) classification system for enzymes and their catalytic
762 reactions.
763 [2] in comparison to all other insects from ArthropodaCyc
764 [3] in comparison among the four hemipterans
765 [4] includes three EC categories added in OGS v1.2 (see also Table S6.11)
766
767
768

769 **CONCLUSIONS**
770
771 The integrated genomic and transcriptomic resources presented here for the milkweed
772 bug *Oncopeltus fasciatus* (Figs. 2,5) underpin a number of insights into evolutionary
773 and developmental genomics. Our macroevolutionary comparisons across insect
774 orders, now extended to the hemimetabolous Hemiptera, reveal unexpected patterns
775 of molecular evolution. We also show how hemipteran feeding ecology and suites of
776 related biological characters are reflected in the genome.
777
778 The gene structure trends we identified, with lineage predominating over
779 genome size as a predictor and with many intron gains in the hemipteroid lineage
780 (Figs. 6,7), offer initial parameters and hypotheses for the Hemiptera, Coleoptera, and
781 Diptera. Such ordinal-level parameters can be evaluated against new species' data
782 and also inform customized pipelines for automated gene model predictions. At the
783 same time, it will be interesting to explore the ramifications of hemipteroid intron
784 gains, as there are few documented lineages with episodic intron gain [77]. For
785 example, possessing more, small exons may provide greater scope to generate protein
786 modularity via isoforms based on alternative exon usage [121]. Furthermore, with the
787 larger genome sizes and lower gene densities of hemipteroids compared to the well-
788 studied Hymenoptera, it remains open whether hemipteroid gene and intron size may
789 also correlate with recombination rates [122].
790
791 Our analyses also highlight new directions for future experimental research,
792 building on *Oncopeltus*'s long-standing history as a laboratory model and its active
793 research community in the modern molecular genetics era (*e.g.*, Fig. 9, [25-27]).
794 Functional testing will clarify the roles of genes we have identified as unique to the
795 Hemiptera, including those implicated in chemical protection, bacterial and plant cell
796 wall metabolism, or encoding wholly novel proteins (Figs. 3,8, see also Supplemental
797 Note 2.2). Meanwhile, the prominent and species-specific expansions specifically of
798 ZF271-like zinc fingers (Fig. 4), combined with the absence of the co-repressor KAP-
799 1 in insects, argues for investigation into alternative interaction partners, which could
800 clarify the nature of these zinc fingers' regulatory roles and their binding targets.
801
802 One key output of this study is the generation of a metabolism database for
803 *Oncopeltus*, contributing to the ArthropodaCyc collection (Table 3). In addition to
804 comparisons with other species (Fig. 10), this database can also serve as a future
805 reference for studies that use *Oncopeltus* as an ecotoxicology model (*e.g.*, [123]).
806 While we have primarily focused on feeding ecology in terms of broad comparisons
807 between phytophagy and fluid feeding, *Oncopeltus* is also poised to support future
808 work on nuances among phytophagous species. Despite its milkweed diet in the wild,
809 the lab strain of *Oncopeltus* has long been adapted to feed on sunflower seeds,
810 demonstrating a latent capacity for more generalist phytophagy [114]. This potential
811 may also be reflected in a larger gustatory receptor repertoire than would be expected
812 for an obligate specialist feeder (Table 2). Thus, *Oncopeltus* can serve as a reference
813 species for promiscuously phytophagous pests such as the stink bug. Finally, we have
814 identified a number of key genes implicated in life history trade-offs in *Oncopeltus*,
815 for traits such as cardenolide tolerance, pigmentation, and plasticity in reproduction
816 under environmental variation. The genome data thus represent an important tool to
817 elucidate the proximate mechanisms of fundamental aspects of life history evolution
818 in both the laboratory and nature.
819

820 **METHODS**
821
822 (More information is available in the supplementary materials, Additional file 1.)
823
824 **Milkweed bug strain, rearing, and DNA/RNA extraction**
825 The milkweed bug *Oncopeltus fasciatus* (Dallas), Carolina Biological Supply strain
826 (Burlington, North Carolina, USA), was maintained in a laboratory colony under
827 standard husbandry conditions (sunflower seed and water diet, 25 ºC, 12:12 light-dark
828 photoperiod). Voucher specimens for an adult female (record # ZFMK-TIS-26324)
829 and adult male (record # ZFMK-TIS-26325) have been preserved in ethanol and
830 deposited in the Biobank of the Centre for Molecular Biodiversity Research,
831 Zoological Research Museum Alexander Koenig, Bonn, Germany
832 (https://www.zfmk.de/en/biobank).
833
834 Genomic DNA was isolated from individual, dissected adults using the Blood
835 & Cell Culture DNA Midi Kit (G/100) (Qiagen Inc., Valencia, California, USA).
836 Total RNA was isolated from individual, dissected adults and from pooled, mixed-
837 instar nymphs with TRIzol Reagent (Invitrogen/ Thermo Fisher Scientific, Waltham,
838 Massachusetts, USA). Dissection improved accessibility of muscle tissue by
839 disrupting the exoskeleton, and gut material was removed.
840
841 **Genome size calculations (flow cytometry, *k*-mer estimation)**
842 Genome size estimations were obtained by flow cytometry with Hare and Johnston's
843 protocol [124]. Four to five females and males each from the Carolina Biological
844 Supply lab strain and a wild strain (collected from Athens, Georgia, USA; GPS
845 coordinates: 33° 56' 52.8216" N, 83° 22' 38.3484'' W) were measured (see also
846 Supplemental Note 2.1.a). At the bioinformatic level, we attempted to estimate
847 genome size by *k*-mer spectrum distribution analysis for a range of *k*=15 to 34
848 counted with Jellyfish 2.1.4 [125] and bbmap [126], graphing these counts against the
849 frequency of occurrence of *k*-mers (depth), and calculating genome size based on the
850 coverage at the peak of the distribution (Supplemental Note 2.1.b).
851
852 **Genome sequencing, assembly, annotation, and official gene set overview**
853 Library preparation, sequencing, assembly, and automatic gene annotation were
854 conducted at the Baylor College of Medicine Human Genome Sequencing Center (as
855 in [12, 30]). About 1.1 billion 100-bp paired-end reads generated on an Illumina
856 HiSeq2000s machine were assembled using ALLPATHS-LG [127], from two paired-
857 end (PE) and two mate pair (MP) libraries specifically designed for this algorithm
858 (Supplemental Note 1). Three libraries were sequenced from an individual adult male
859 (180- and 500-bp PE, 3-kb MP), with the fourth from an individual adult female (8-
860 10-kb MP). The final assembly (see metrics in Table 1) has been deposited in
861 GenBank (accession GCA_000696205.1).
862
863 Automated annotation of protein-coding genes was performed using a Maker
864 2.0 annotation pipeline [128] tuned specifically for arthropods (Supplemental Note 3).
865 These gene predictions were used as the starting point for manual curation via the
866 Apollo v.1.0.4 web browser interface [129], and automatic and manual curations were
867 compiled to generate the OGS (see also Supplemental Note 4). Databases of the
868 genome assembly, Maker automatic gene predictions, and OGS v1.1 are available
869 through the i5K Workspace@NAL [130], and the Ag Data Commons data access

870 system of the United States Department of Agriculture's (USDA) National
871 Agricultural Library as individual citable databases [131-133]. The current version of
872 the gene set, OGS v1.2, is deposited at GenBank as an 'annotation-only' update to the
873 Whole Genome Shotgun project (accession JHQO00000000). Here, we describe
874 version JHQO02000000. The annotations can be downloaded from NCBI's ftp site,
875 ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/696/205/GCA_000696205.1_Ofas_1.0/.
876
877 **Repeat content analysis**
878 Repetitive regions were identified in the *Oncopeltus* genome assembly with
879 RepeatModeler Open-1.0.8 [134] based on a species-specific repeat library generated
880 *de novo* with RECON [135], RepeatScout [136], and Tandem Repeats Finder [137].
881 Then, RepeatMasker Open-4.0 [138] was used to mask repeat sequences based on the
882 RepeatModeler library. Given the fragmented nature of the assembly, we attempted
883 to fill and close assembly gaps by sequencing additional material, generating long
884 reads with single molecule real time sequencing on a PacBio RS II machine
885 (estimated coverage of 8x). Gap filling on the Illumina assembly scaffolds was
886 performed with PBJelly version 13.10.22, and the resulting assembly was used for
887 repeat content estimation and comparison with *Cimex lectularius* and *Acyrthosiphon*
888 *pisum* (see also Supplemental Note 2.3).
889
890 **Transcriptome resources**
891 Total RNA from three distinct life history samples (pooled, mixed-instar nymphs; an
892 adult male; an adult female) was also sequenced on an Illumina HiSeq2000s machine,
893 producing a total of 72 million 100-bp paired-end reads (Supplemental Note 1.3,
894 Table S1.1; GenBank Bioproject: PRJNA275739). These expression data were used
895 to support the generation of the OGS at different stages of the project: as input for the
896 evidence-guided automated annotation with Maker 2.0 (Supplemental Note 3), as
897 expression evidence tracks in the Apollo browser to support the community curation
898 of the OGS, and, once assembled into a *de novo* transcriptome, as a point of
899 comparison for quality control of the OGS.
900
901 The raw RNA-seq reads were pre-processed by filtering out low quality bases
902 (phred score <30) and Truseq adapters with Trimmomatic-0.30. Further filtering
903 removed ribosomal and mitochondrial RNA sequences with Bowtie 2 [139], based on
904 a custom library built with all hemipteran ribosomal and mitochondrial RNA
905 accessions from NCBI as of 7th February 2014 (6,069 accessions). The pooled,
906 filtered reads were mapped to the genome assembly with Tophat2-PE on CyVerse
907 [140]. A second set of RNA-seq reads from an earlier study ("published adult"
908 dataset, [37]) was also filtered and mapped in the same fashion, and both datasets
909 were loaded into the *Oncopeltus* Apollo instance as evidence tracks (under the track
910 names "pooled RNA-seq - cleaned reads" and "RNA-seq raw PE reads Andolfatto et
911 al", respectively).
912
913 Additionally, a *de novo* transcriptome was generated from our filtered RNA-
914 seq reads (pooled from all three samples prepared in this study) using Trinity [141]
915 and TransDecoder [142] with default parameters. This transcriptome is referred to as
916 "i5K", to distinguish it from a previously published maternal and early embryonic
917 transcriptome for *Oncopeltus* (referred to as "454", [36]). Both the i5K and 454
918 transcriptomes were mapped to the genome assembly with GMAP v. 2014-05-15 on

919 CyVerse. These datasets were also loaded into the Apollo browser as evidence tracks
920 to assist in manual curation.
921
922 **Life history stage-specific and sex-specific expression analyses in hemipteroids**
923 Transcript expression of the OGS v1.1 genes was estimated by running RSEM2 [143]
924 on the filtered RNA-seq datasets for the three i5K postembryonic stages against the
925 OGS v1.1 cDNA dataset. Transcript expression was then based on the transcripts per
926 million (TPM) value. The TPM values were processed by adding a value of 1 (to
927 avoid zeros) and then performing a log2-transformation. The number of expressed
928 genes per RNA-seq library was compared for TPM cutoffs of >1, >0.5, and >0.25. A
929 >0.25 cutoff was chosen, which reduced the number of expressed genes by 6.6%
930 compared to a preliminary analysis based on a simple cutoff of ≥10 mapped reads per
931 transcript, while the other TPM cutoffs were deemed too restrictive (reducing the
932 expressed gene set by >10%). This analysis was also applied to the "published adult"
933 dataset [37]. To include embryonic stages in the comparison, transcripts from the 454
934 transcriptome were used as blastn queries against the OGS v1.1 cDNA dataset (cutoff
935 e-value $<10^{-5}$). The results from all datasets were converted to binary format to
936 generate Venn diagrams (Fig. 2b).
937
938 Statistically significant sex-specific and developmental stage-specific gene
939 enrichment was determined from RNA-seq datasets according to published methods
940 [144, 145], with modifications. Data from *Oncopeltus* (see previous methods section,
941 Bioproject: PRJNA275739) were compared between stages and pairwise with the
942 hemipterans *Cimex lectularius*, PRJNA275741; *Acyrthosiphon pisum*, PRJNA209321;
943 and *Pachypsylla venusta*; PRJNA275248; as well as with the hemipteroid
944 *Frankliniella occidentalis* (Thysanoptera), PRJNA203209 (see also Fig. 2c,
945 Supplemental Note 2.4).
946
947 **Protein gene orthology assessments via OrthoDB and BUSCO analyses**
948 These analyses follow previously described approaches and with the current database
949 and pipeline versions [1, 43, 45, 146]. See Supplemental Note 6.1 for further details.
950
951 **Global transcription factor identification**
952 Likely transcription factors (TFs) were identified by scanning the amino acid
953 sequences of predicted protein-coding genes for putative DNA binding domains
954 (DBDs), and when possible, the DNA binding specificity of each TF was predicted
955 using established procedures [58]. Briefly, all protein sequences were scanned for
956 putative DBDs using the 81 Pfam [147] models listed in Weirauch and Hughes [148]
957 and the HMMER tool [149], with the recommended detection thresholds of Per-
958 sequence Eval < 0.01 and Per-domain conditional Eval < 0.01. Each protein was
959 classified into a family based on its DBDs and their order in the protein sequence
960 (*e.g.*, bZIPx1, AP2x2, Homeodomain+Pou). The resulting DBD amino acid
961 sequences were then aligned within each family using Clustal Omega [150], with
962 default settings. For protein pairs with multiple DBDs, each DBD was aligned
963 separately. From these alignments, the sequence identity was calculated for all DBD
964 sequence pairs (*i.e.*, the percent of amino acid residues that are identical across all
965 positions in the alignment). Using previously established sequence identity thresholds
966 for each family [58], the predicted DNA binding specificities were mapped by simple
967 transfer. For example, the DBD of OFAS001246-RA is 98% identical to the
968 *Drosophila melanogaster* Bric a Brac 1 (Bab1) protein. Since the DNA binding

969 specificity of Bab1 has already been experimentally determined, and the cutoff for the
970 Pipsqueak family TFs is 85%, we can infer that OFAS001246-RA will have the same
971 binding specificity as *Drosophila* Bab1.
972
973 **RNA interference**
974 Double-stranded RNA (dsRNA) was designed to target the final, unique exon of the
975 *broad* isoforms *Z2, Z3*, and *Z4*. A portion of the coding sequence for the zinc finger
976 region from these exons (179 bp, 206 bp, and 216 bp, respectively) was cloned into a
977 plasmid vector and used as template for *in vitro* RNA synthesis, using the gene-
978 specific primer pairs: Of-Z2_fwd: 5′-ATGTGGCAGACAAGCATGCT-3′; Of-
979 Z2_rev: 5′-CTAAAATTTGACATCAGTAGGC-3′; Of-Z3_fwd: 5′-
980 ccttctcctgttactactcac-3′; Of-Z3_rev: 5′-ttatatgggcggctgtccaa-3′; Of-Z4_fwd: 5′-
981 AACACTGACCTTGGTTACACA-3′; Of-Z4_rev: 5′-
982 TAGGTGGAGGATTGCTAAAATT-3′. Two separate transcription reactions (one
983 for each strand) were performed using the Ambion MEGAscript kit (Ambion, Austin,
984 Texas, USA). The reactions were purified by phenol/chloroform extraction followed
985 by precipitation as described in the MEGAscript protocol. The separate strands were
986 re-annealed in a thermocycler as described previously [33]. Nymphs were injected
987 with a Hamilton syringe fitted with a 32-gauge needle as described [55]. The
988 concentration of *Of-Z2*, *Of-Z3* and *Of-Z4* dsRNA was 740 ng/µl, 1400 ng/µl, and
989 1200 ng/µl, respectively. All nymphs were injected within 8 hours of the molt to the
990 fourth (penultimate juvenile) instar (n ≥12 per treatment: see Fig. 9). Fore- and
991 hindwings were then dissected from adults and photographed at the same scale as
992 wings from wild type, uninjected controls.
993
994 **CycADS annotation and OncfaCyc database generation**
995 We used the Cyc Annotation Database System (CycADS, [115]), an automated
996 annotation management system, to integrate protein annotations from different
997 sources into a Cyc metabolic networks reconstruction that was integrated into the
998 ArthropodaCyc database. Using our CycADS pipeline, *Oncopeltus fasciatus* proteins
999 from the official gene set OGS v1.1 were annotated using different methods –
1000 including KAAS [151], PRIAM [152], Blast2GO [153, 154], and InterProScan with
1001 several approaches [155] – to obtain EC and GO numbers. All annotation
1002 information data were collected in the CycADS SQL database and automatically
1003 extracted to generate appropriate input files to build or update BioCyc databases [156]
1004 using the Pathway Tools software [157]. The OncfaCyc database, representing the
1005 metabolic protein-coding genes of *Oncopeltus*, was thus generated and is now
1006 included in the ArthropodaCyc database, a collection of arthropod metabolic network
1007 databases ([116], http://arthropodacyc.cycadsys.org/).
1008
1009
1010 **FIGURE LEGENDS**
1011
1012 **Fig. 1. The large milkweed bug, *Oncopeltus fasciatus*, shown in its phylogenetic**
1013 **and environmental context.**
1014 **(a)** Species tree of selected Hemiptera with genomic and transcriptomic resources,
1015 based on phylogenetic analyses and divergence time estimates in [3]. Species marked
1016 with an asterisk (*) have published resources; those with the appellation "i5K" are
1017 part of a current pilot project supported by the Baylor College of Medicine Human
1018 Genome Sequencing Center and the National Agricultural Library of the USDA.

1019 Note that recent analyses suggest the traditional infraorder Cimicomorpha, to which
1020 *Rhodnius* and *Cimex* belong, may be paraphyletic [16].
1021 **(b-c)** Milkweed bugs on their native food source, the milkweed plant: gregarious
1022 nymphs of different instars on a milkweed seed pod (b), and pale, recently eclosed
1023 adults and their shed exuvia (c). Images were taken at Avalon Park and Preserve,
1024 Stony Brook, New York, USA, courtesy of Deniz Erezyilmaz, used with permission.
1025 **(d)** Individual bugs, shown from left to right: first instar nymphs (ventral and dorsal
1026 views) and adults (dorsal and lateral views); images courtesy of Kristen Panfilio
1027 (nymphs) and Jena Johnson (adults), used with permission. The arrow labels the
1028 labium (the "straw"), part of the hemipteran mouthpart anatomy adapted for feeding
1029 by piercing and sucking.
1030
1031 **Fig. 2. Comparisons of the official gene set and transcriptomic resources for**
1032 ***Oncopeltus fasciatus.***
1033 **(a)** Area-proportional Venn diagram comparing the OGS v1.1 ("OGS"), a Trinity *de*
1034 *novo* transcriptome from the three post-embryonic RNA-seq samples ("i5K"), and the
1035 maternal and embryonic transcriptome from 454 data ("454" [36]). Sample sizes and
1036 the fraction of each transcriptome represented in the OGS are indicated (for the 454
1037 dataset, only transcripts with homology identification were considered). The unique
1038 fraction of each set is also specified (%). Dataset overlaps were determined by blastn
1039 (best hit only, e-value $<10^{-9}$).
1040 **(b)** Venn diagram of gene model expression support across four life history samples.
1041 Values are numbers of gene models, with percentages also given for the largest
1042 subsets. Note that the "Embryo/Maternal" sample derives from 454 pyrosequencing
1043 data and therefore has a smaller data volume than the other, Illumina-based samples.
1044 **(c)** Summary of sex- and developmental stage-specific RNA-seq comparisons across
1045 hemipteroid species: *Acyrthosiphon pisum*, *Apis; Cimex lectularius, Clec;*
1046 *Frankliniella occidentalis, Focc* (thysanopteran outgroup)*; Oncopeltus fasciatus,*
1047 *Ofas; Pachypsylla venusta, Pven*. n.d., not determined. For complete numerical
1048 details see Supplemental Note 2.4
1049 Analyses based on OGS v1.1.
1050
1051 **Fig. 3. Orthology comparisons and phylogenetic placement of *Oncopeltus***
1052 ***fasciatus* among other Arthropoda.**
1053 **(a)** Comparisons of protein-coding genes in 12 arthropod species, with the Hemiptera
1054 highlighted in red text. The bar chart shows the number of proteins per conservation
1055 level (see legend), based on OrthoDB orthology clustering analyses. To the left is a
1056 maximum likelihood phylogeny based on concatenation of 395 single-copy orthologs
1057 (all nodes have 100% support unless otherwise noted; branch length unit is
1058 substitutions per site). The inset pie chart shows the proportion of proteins per
1059 conservation level in *Oncopeltus* (*"Ofas"*). See also Supplemental Note 6.1.
1060 **(b)** BUSCO-based analysis of *Oncopeltus* compared to other hemipterans for ortholog
1061 presence and copy number in both the assembly and OGS resources, using four-letter
1062 species abbreviations (full names in panel a).
1063 **(c)** Proportion of *Oncopeltus* proteins that have expression and/or curation validation
1064 support per conservation level (same color legend as in (a)). Expression support is
1065 based on the life history stage data in Fig. 2b.
1066 Analyses based on OGS v1.1.
1067

1068 **Fig. 4. Distribution of transcription factor families across insect genomes.**
1069 **(a)** Heatmap depicting the abundance of 74 transcription factor (TF) families across
1070 16 insect genomes (Hemiptera highlighted in red text), with *Daphnia* as an outgroup,
1071 based on the presence of predicted DNA binding domains (see Methods). The color
1072 key has a log (base 2) scale (light blue means the TF family is completely absent).
1073 Values are in Table S6.3.
1074 **(b)** Bar graph showing the number of proteins of each of the two most abundant TF
1075 families, homeodomains and C2H2 zinc fingers (ZFs), per species using four-letter
1076 abbreviations (full names in panel a). Solid lines demarcate insect orders: Hemiptera
1077 (Hemipt.), Hymenoptera (Hym.), Coleoptera (Col.), and Diptera (Dipt.). The dashed
1078 line demarcates the dipteran family Culicidae (mosquitoes).
1079 **(c)** Proportions of *Oncopeltus* homeodomain (HD) and C2H2 zinc finger proteins
1080 with orthology assignment (predicted DNA binding specificity) and/or manual
1081 curation. "Classified" refers to automated classification of a protein to a TF family,
1082 but without a specific orthology assignment.
1083 **(d)** Maximum likelihood phylogeny of representative subsets of the zinc finger 271-
1084 like family in *Oncopeltus* (49 proteins, blue text) and the pea aphid (55 proteins, black
1085 text), with chelicerate (red text) and holometabolan (yellow text) outgroups (16
1086 proteins, 7 species), based on the *Oncopeltus* OGS and GenBank protein accessions.
1087 Gaps were removed during sequence alignment curation; all nodes have ≥50%
1088 support; branch length unit is substitutions per site [158]. Key nodes are circled for
1089 the clades containing all aphid or all *Oncopeltus* proteins (82% support each), and for
1090 each 'core' clade comprised exclusively of proteins from each species (97% and
1091 100%, respectively; triangles shown to scale for branch length and number of clade
1092 members). Branch length unit is substitutions per site.
1093 Analyses based on OGS v1.1.
1094
1095 **Fig. 5. Comparison of repeat content estimations.**
1096 **(a)** Comparison of total repetitive content among insect genomes. The three values
1097 for *Oncopeltus* are shown (in ascending order: original Illumina assembly, gap-filled
1098 assembly, Illumina-PacBio hybrid estimate). Values for the three hemipterans labeled
1099 in red text are from RepeatModeler (gold bars for the pea aphid and bed bug; blue and
1100 gold bars for *Oncopeltus*). All other values are from the respective genome papers,
1101 including a second value corresponding to the published repeat content for the first
1102 version of the aphid genome [6, 10, 106, 159-164]. Species abbreviations as in Fig. 4,
1103 and additionally: *Nlug, Nilaparvata lugens; Lmig, Locusta migratoria; Bmor, Bombyx*
1104 *mori; Aalb, Aedes albopictus*.
1105 **(b)** Comparison of repetitive element categories between three hemipteran genomes,
1106 based on results from RepeatModeler. Here we present assembly coverage as actual
1107 sequence length (Mb) to emphasize the greater repeat content in *Oncopeltus* (based on
1108 the gap-filled assembly, see also Supplemental Note 2.3).
1109
1110 **Figure 6. Trends in gene structure show hemipteroid-specific tendencies.**
1111 **(a)** Median values per species for protein size, exon size, and exon number for a
1112 curated set of highly conserved genes encoding large proteins of diverse functional
1113 classes (see also Supplemental Note 6.3). Sample sizes are indicated, with 11 genes
1114 for which orthologs were evaluated in all species. Where it was not possible to
1115 analyze all 30 genes for a given species, equal sampling was done across the range of
1116 protein sizes of the complete dataset, based on the *Cimex* ortholog sizes (1:1:1
1117 sampling from big:medium:small subcategories of 10 genes each).

1118 **(b)** Box plot representations of coding sequence exon size (aa) for two species from
1119 each of three insect orders, based on datasets of unique coding sequence exons (one
1120 isoform per gene) and excluding terminal exons <10 aa (as most of those exons may
1121 rather be UTRs or a small placeholder N-terminal exon based on automated Maker
1122 model predictions). Only manually curated gene models were considered for the i5K
1123 species, including *Oncopeltus*; the entire OGS was used for *Tribolium* and
1124 *Drosophila*. For clarity, outliers are omitted; whiskers represent 1.5× the value of the
1125 Q3 (upper) or Q2 (lower) quartile range. MAD, median absolute deviation.
1126 Species are represented by their four-letter abbreviations, with their ordinal
1127 relationships given below the phylogeny in panel (a): Hemip., Hemiptera; Thys.,
1128 Thysanoptera; Col., Coleoptera; Dipt., Diptera. Species abbreviations as in Figs. 2,4
1129 and additionally: *Gbue, Gerris buenoi* [165]; *Agla, Anoplophora glabripennis* [30];
1130 *Ccap, Ceratitis capitata* [166].
1131
1132 **Fig 7. Splice site evolution correlates with both lineage and genome size.**
1133 Splice site changes are shown for *hemocytin* (blue text), *Tenascin major* (*Ten-m*,
1134 turquoise text), and *UDP-galactose 4′-epimerase* (brown text), mapped onto a species
1135 tree of eight insects. Patterns of splice site evolution were inferred based on the most
1136 parsimonious changes that could generate the given pattern within a protein sequence
1137 alignment of all orthologs (see also Supplemental Note 6.3 for methodology and data
1138 sources). If inferred gains or losses were equally parsimonious, we remained agnostic
1139 and present a range for the ancestral number of splice sites present at the base of the
1140 tree, where the bracketed number indicates how many ancestral positions are still
1141 retained in all species. Along each lineage, subsequent changes are indicated in
1142 brackets, with the sign indicating gains (+) or losses (-). Values shown to the right are
1143 species-specific changes. The values shown between the *D. melanogaster* and *T.*
1144 *castaneum* lineages denote changes that have occurred independently in both species.
1145 Colored boxes highlight the largest sources of change, as indicated in the legend.
1146 Species are represented by their four-letter abbreviations (as in Fig. 6), and estimated
1147 genome sizes are indicated parenthetically (measured size: [12, 30, 163, 166, 167];
1148 draft assembly size: GenBank Genome IDs 14741 and 17730). Divergence times are
1149 shown in gray and given in millions of years [3]. Abbreviations as in Figs. 4,6, and
1150 also: Col., Coleoptera; Dipt., Diptera; Hemip., Hemiptera; Hemipt., hemipteroid
1151 assemblage (including *F. occidentalis*); n.d., no data.
1152
1153 **Fig. 8. Lateral gene transfer introduction and subsequent evolution within the**
1154 **Hemiptera for mannosidase-encoding genes.**
1155 **(a)** Species tree summary of evolutionary events. Stars represent the original LGT
1156 introduction and subsequent copy number gains (see legend).
1157 **(b)** Maximum likelihood phylogeny of mannosidase proteins, including bacterial
1158 sequences identified among the best GenBank blastp hits for *Oncopeltus* and
1159 *Halyomorpha* (accession numbers as indicated, and for "Other bacteria" are:
1160 ACB22214.1, AEE17431.1, AEI12929.1, AEO43249.1, AFN74531.1, CDM56239.1,
1161 CUA67033.1, KOE98396.1, KPI24888.1, OAN41395.1, ODP26899.1, ODS11151.1,
1162 OON18663.1, PBD05534.1, SIR54690.1, WP096035621.1, YP001327394.1). All
1163 nodes have ≥50% support from 500 bootstrap replicates [168]. Triangles are shown
1164 to scale for branch length and number of clade members; branch length unit is
1165 substitutions per site. See also Fig. S2.6.
1166 **(c)** Manually curated protein sequence alignment for the N-terminal region only.
1167 Splice sites ("|" symbol) are shown, where one position is ancestral and present in all

1168 paralogs of a given species (magenta) and one position occurs in a subset of paralogs
1169 and is presumed to be younger (cyan, within the 5′ UTR in *Halyomorpha*). Residues
1170 highlighted in yellow are conserved between the two hemipteran species. The
1171 *Oncopeltus* paralog represented in the OGS as OFAS017153-RA is marked with an
1172 asterisk to indicate that this version of the gene model is incomplete and lacks the
1173 initial exon (gray text in the alignment). For clarity, only the final three digits of the
1174 *Halyomorpha* GenBank accessions are shown (full accessions: XP_014289XXX).
1175
1176 **Fig. 9. Isoform-specific RNAi based on new genome annotations affects the**
1177 **molting and cuticle identity gene *broad*.**
1178 **(a)** Genomic organization of the cuticle identity gene *broad*. The regions used as
1179 template to generate isoform-specific dsRNA are indicated (red asterisks: the final,
1180 unique exons of each isoform). Previous RNAi studies targeted sequence within
1181 exons 1-5 that is shared among all isoforms (dashed red box, [93]).
1182 **(b)** Knock down of the *Oncopeltus Z2* or *Z3 broad* isoforms at the onset of the
1183 penultimate instar resulted in altered nymphal survival and morphogenesis that was
1184 reflected in the size and proportion of the fore and hind wings at the adult stage
1185 (upper and lower images, respectively, shown to the same scale for all wings). We
1186 did not detect any effect on the wing phenotype when targeting the *Z4*-specific exon,
1187 demonstrating the specificity of the zinc finger coding region targeted by RNAi.
1188 Experimental statistics are provided in the figure inset, including for the buffer-
1189 injected negative control.
1190
1191 **Fig. 10. Comparison of the urea cycle of *Oncopeltus* with 26 other insect species.**
1192 **(a)** Detailed diagram of the urea cycle (adapted from KEGG).
1193 **(b)** Group of 7 species, including *Oncopeltus*, for which Arg degradation via arginase
1194 (3.5.3.1), but not synthesis, is possible.
1195 **(c)** Group of 3 species for which neither the degradation nor synthesis of arginine via
1196 the urea cycle is possible (the three other hemipterans in this analysis).
1197 **(d)** Group of 17 species sharing a complete (or almost complete) urea cycle.
1198 Hemiptera are identified in red text and the milkweed-feeding monarch butterfly is in
1199 blue text. Enzyme names corresponding to EC numbers: 1.5.1.2 = pyrroline-5-
1200 carboxylate reductase; 1.14.13.39 = nitric-oxide synthase; 2.1.3.3 = ornithine
1201 carbamoyltransferase; 2.6.1.13 = ornithine aminotransferase; 3.5.3.1 = arginase;
1202 4.3.2.1 = argininosuccinate lyase; 6.3.4.5 = argininosuccinate synthase.
1203 Analyses based on OGS v1.1.
1204
1205
1206 **TABLES (see above within relevant manuscript sections)**
1207
1208 **Table 1. *Oncopeltus fasciatus* genome metrics.**
1209
1210 **Table 2. Numbers of chemoreceptor genes/proteins per family in selected insect**
1211 **species.**
1212
1213 **Table 3. Hemipteran ArthropodaCyc database summaries.**
1214
1215 **Table 4. Hemipteran ArthropodaCyc annotations of metabolic genes.**
1216

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) 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* 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 ([131-133], https://i5k.nal.usda.gov/Oncopeltus_fasciatus).

AUTHORS' CONTRIBUTIONS

KAP and StR conceived the project. KAP 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, CJH, 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

1367 **REFERENCES**

1368

1369 1. Zdobnov EM, Tegenfeldt F, Kuznetsov D, Waterhouse RM, Simão FA, Ioannidis P, Seppey
1370 M, Loetscher A, Kriventseva EV: **OrthoDB v9.1: cataloging evolutionary and functional**
1371 **annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs.** *Nucleic*
1372 *Acids Res* 2017, **45:**D744-D749.
1373 2. Huang DY, Bechly G, Nel P, Engel MS, Prokop J, Azar D, Cai CY, van de Kamp T,
1374 Staniczek AH, Garrouste R, et al: **New fossil insect order Permopsocida elucidates major**
1375 **radiation and evolution of suction feeding in hemimetabolous insects (Hexapoda:**
1376 **Acercaria).** *Sci Rep* 2016, **6:**23004.
1377 3. Misof B, Liu S, Meusemann K, Peters RS, Donath A, Mayer C, Frandsen PB, Ware J, Flouri
1378 T, Beutel RG, et al: **Phylogenomics resolves the timing and pattern of insect evolution.**
1379 *Science* 2014, **346:**763-767.
1380 4. Grimaldi D, Engel MS: *Evolution of the Insects.* Cambridge: Cambridge University Press;
1381 2005.
1382 5. Panfilio KA, Angelini DR: **By land, air, and sea: hemipteran diversity through the**
1383 **genomic lens.** *Curr Opin Insect Sci* 2018, **25:**in press, DOI: 10.1016/j.cois.2017.1012.1005.
1384 6. The International Aphid Genomics Consortium: **Genome sequence of the pea aphid**
1385 ***Acyrthosiphon pisum*** *PLoS Biol* 2010, **8:**e1000313.
1386 7. Mathers TC, Chen Y, Kaithakottil G, Legeai F, Mugford ST, Baa-Puyoulet P, Bretaudeau A,
1387 Clavijo B, Colella S, Collin O, et al: **Rapid transcriptional plasticity of duplicated gene**
1388 **clusters enables a clonally reproducing aphid to colonise diverse plant species.** *Genome*
1389 *Biol* 2017, **18:**27.
1390 8. Wenger JA, Cassone BJ, Legeai F, Johnston JS, Bansal R, Yates AD, Coates BS, Pavinato
1391 VA, Michel A: **Whole genome sequence of the soybean aphid, *Aphis glycines*.** *Insect*
1392 *Biochem Mol Biol* 2017.
1393 9. Sloan DB, Nakabachi A, Richards S, Qu J, Murali SC, Gibbs RA, Moran NA: **Parallel**
1394 **histories of horizontal gene transfer facilitated extreme reduction of endosymbiont**
1395 **genomes in sap-feeding insects.** *Mol Biol Evol* 2014, **31:**857-871.
1396 10. Xue J, Zhou X, Zhang C-X, Yu L-L, Fan H-W, Wang Z, Xu H-J, Xi Y, Zhu Z-R, Zhou W-W,
1397 et al: **Genomes of the rice pest brown planthopper and its endosymbionts reveal complex**
1398 **complementary contributions for host adaptation.** *Genome Biol* 2014, **15:**521.
1399 11. Mesquita RD, Vionette-Amaral RJ, Lowenberger C, Rivera-Pomar R, Monteiro FA, Minx P,
1400 Spieth J, Carvalho AB, Panzera F, Lawson D, et al: **Genome of *Rhodnius prolixus*, an insect**
1401 **vector of Chagas disease, reveals unique adaptations to hematophagy and parasite**
1402 **infection.** *Proc Natl Acad Sci USA* 2015, **112:**14936-14941.
1403 12. Benoit JB, Adelman ZN, Reinhardt K, Dolan A, Poelchau M, Jennings EC, Szuter EM, Hagan
1404 RW, Gujar H, Shukla JN, et al: **Unique features of a global human ectoparasite identified**
1405 **through sequencing of the bed bug genome.** *Nat Commun* 2016, **7:**10165.
1406 13. Rosenfeld JA, Reeves D, Brugler MR, Narechania A, Simon S, Durrett R, Foox J, Shianna K,
1407 Schatz MC, Gandara J, et al: **Genome assembly and geospatial phylogenomics of the bed**
1408 **bug *Cimex lectularius*.** *Nat Commun* 2016, **7:**10164.
1409 14. Sparks ME, Shelby KS, Kuhar D, Gundersen-Rindal DE: **Transcriptome of the invasive**
1410 **brown marmorated stink bug, *Halyomorpha halys* (Stal) (Heteroptera: Pentatomidae).**
1411 *PLoS One* 2014, **9:**e111646.
1412 15. Ioannidis P, Lu Y, Kumar N, Creasy T, Daugherty S, Chibucos MC, Orvis J, Shetty A, Ott S,
1413 Flowers M, et al: **Rapid transcriptome sequencing of an invasive pest, the brown**
1414 **marmorated stink bug, *Halyomorpha halys*.** *BMC Genomics* 2014, **15:**738.
1415 16. Li H, Leavengood JM, Jr., Chapman EG, Burkhardt D, Song F, Jiang P, Liu J, Zhou X, Cai
1416 W: **Mitochondrial phylogenomics of Hemiptera reveals adaptive innovations driving the**
1417 **diversification of true bugs.** *Proc Biol Sci* 2017, **284**.
1418 17. Wilson ACC, Ashton PD, Charles H, Colella S, Febvay G, Jander G, Kushlan PF, Macdonald
1419 SJ, Schwartz JF, Thomas GH, Douglas AE: **Genomic insight into the amino acid relations**
1420 **of the pea aphid, *Acyrthosiphon pisum*, with its symbiotic bacterium *Buchnera***
1421 ***aphidicola*.** *Insect Mol Biol* 2010, **19 Suppl 2:**249-258.
1422 18. Eichler S, Schaub GA: **Development of symbionts in triatomine bugs and the effects of**
1423 **infections with trypanosomatids.** *Exp Parasitol* 2002, **100:**17-27.
1424 19. Matsuura Y, Kikuchi Y, Hosokawa T, Koga R, Meng X-Y, Kamagata Y, Nikoh N, Fukatsu T:
1425 **Evolution of symbiotic organs and endosymbionts in lygaeid stinkbugs.** *The ISME Journal*
1426 2012, **6:**397–409.

1427 20. Berenbaum MR, Miliczky E: **Mantids and milkweed bugs - efficacy of aposematic**
1428 **coloration against invertebrate predators.** *American Midland Naturalist* 1984, **111:**64-68.
1429 21. Burdfield-Steel ER, Shuker DM: **The evolutionary ecology of the Lygaeidae.** *Ecol Evol*
1430 2014, **4:**2278-2301.
1431 22. Lawrence PA: **Mitosis and the cell cycle in the metamorphic moult of the milkweed bug**
1432 **Oncopeltus fasciatus; a radioautographic study.** *J Cell Sci* 1968, **3:**391-404.
1433 23. Chipman AD: ***Oncopeltus fasciatus* as an evo-devo research organism.** *Genesis* 2017, **55**.
1434 24. Panfilio KA: **Late extraembryonic development and its *zen-RNAi*-induced failure in the**
1435 **milkweed bug *Oncopeltus fasciatus*.** *Dev Biol* 2009, **333:**297-311.
1436 25. Panfilio KA, Roth S: **Epithelial reorganization events during late extraembryonic**
1437 **development in a hemimetabolous insect.** *Dev Biol* 2010, **340:**100-115.
1438 26. Sharma AI, Yanes KO, Jin L, Garvey SL, Taha SM, Suzuki Y: **The phenotypic plasticity of**
1439 **developmental modules.** *Evodevo* 2016, **7:**15.
1440 27. Hughes CL, Kaufman TC: **RNAi analysis of *Deformed, proboscipedia* and *Sex combs***
1441 ***reduced* in the milkweed bug *Oncopeltus fasciatus*: novel roles for Hox genes in the**
1442 **hemipteran head.** *Development* 2000, **127:**3683-3694.
1443 28. Wolfe SL, John B: **The organization and ultrastructure of male meiotic chromosomes in**
1444 ***Oncopeltus fasciatus*.** *Chromosoma* 1965, **17:**85-103.
1445 29. Messthaler H, Traut W: **Phases of Sex Chromosome Inactivation in *Oncopeltus fasciatus***
1446 **and *Pyrrhocoris apterus* (Insecta, Heteroptera).** *Caryologia* 1975, **28:**501-510.
1447 30. McKenna DD, Scully ED, Pauchet Y, Hoover K, Kirsch R, Geib SM, Mitchell RF,
1448 Waterhouse RM, Ahn SJ, Arsala D, et al: **Genome of the Asian longhorned beetle**
1449 **(*Anoplophora glabripennis*), a globally significant invasive species, reveals key functional**
1450 **and evolutionary innovations at the beetle-plant interface.** *Genome Biol* 2016, **17:**227.
1451 31. Simpson JT: **Exploring genome characteristics and sequence quality without a reference.**
1452 *Bioinformatics* 2014, **30:**1228-1235.
1453 32. Hanrahan SJ, Johnston JS: **New genome size estimates of 134 species of arthropods.**
1454 *Chromosome Res* 2011, **19:**809-823.
1455 33. Hughes CL, Kaufman TC: **RNAi analysis of *Deformed*, *proboscipedia* and *Sex combs***
1456 ***reduced* in the milkweed bug *Oncopeltus fasciatus*: novel roles for Hox genes in the**
1457 **Hemipteran head.** *Development* 2000, **127:**3683-3694.
1458 34. Panfilio KA, Liu PZ, Akam M, Kaufman TC: ***Oncopeltus fasciatus zen* is essential for**
1459 **serosal tissue function in katatrepsis.** *Dev Biol* 2006, **292:**226-243.
1460 35. Tian X, Xie Q, Li M, Gao C, Cui Y, Xi L, Bu W: **Phylogeny of pentatomomorphan bugs**
1461 **(Hemiptera-Heteroptera:Pentatomomorpha) based on six Hox gene fragments.** *Zootaxa*
1462 2011, **2888:**57-68.
1463 36. Ewen-Campen B, Shaner N, Panfilio KA, Suzuki Y, Roth S, Extavour CG: **The maternal**
1464 **and early embryonic transcriptome of the milkweed bug *Oncopeltus fasciatus*.** *BMC*
1465 *Genomics* 2011, **12:**61.
1466 37. Zhen Y, Aardema ML, Medina EM, Schumer M, Andolfatto P: **Parallel molecular evolution**
1467 **in an herbivore community.** *Science* 2012, **337:**1634-1637.
1468 38. Robertson HM: **The insect chemoreceptor superfamily in Drosophila pseudoobscura:**
1469 **Molecular evolution of ecologically-relevant genes over 25 million years** *J Insect Sci* 2009,
1470 **9:**18.
1471 39. Robertson HM: **Taste: Independent origins of chemoreception coding systems?** *Curr Biol*
1472 2001, **11:**R560-R562.
1473 40. Jazwinska A, Rushlow C, Roth S: **The role of *brinker* in mediating the graded response to**
1474 **Dpp in early *Drosophila* embryos.** *Development* 1999, **126:**3323-3334.
1475 41. Togawa T, Dunn WA, Emmons AC, Nagao J, Willis JH: **Developmental expression**
1476 **patterns of cuticular protein genes with the R&R Consensus from *Anopheles gambiae*.**
1477 *Insect Biochem Mol Biol* 2008, **38:**508-519.
1478 42. Karr TL: **Fruit flies and the sperm proteome.** *Hum Mol Genet* 2007, **16 Spec No. 2:**R124-
1479 133.
1480 43. Waterhouse RM, Seppey M, Simao FA, Manni M, Ioannidis P, Klioutchnikov G, Kriventseva
1481 EV, Zdobnov EM: **BUSCO applications from quality assessments to gene prediction and**
1482 **phylogenomics.** *Mol Biol Evol* 2017.
1483 44. Shigenobu S, Bickel RD, Brisson JA, Butts T, Chang CC, Christiaens O, Davis GK, Duncan
1484 EJ, Ferrier DE, Iga M, et al: **Comprehensive survey of developmental genes in the pea**
1485 **aphid, *Acyrthosiphon pisum*: frequent lineage-specific duplications and losses of**
1486 **developmental genes.** *Insect Mol Biol* 2010, **19 Suppl 2:**47-62.

1487 45. Kriventseva EV, Tegenfeldt F, Petty TJ, Waterhouse RM, Simão FA, Pozdnyakov IA,
1488 Ioannidis P, Zdobnov EM: **OrthoDB v8: update of the hierarchical catalog of orthologs**
1489 **and the underlying free software.** *Nucl Acids Res* 2015, **43:**D250-D256.
1490 46. Bansal R, Michel AP: **Core RNAi Machinery and *Sid1*, a component for systemic RNAi,**
1491 **in the hemipteran insect, *Aphis glycines*.** *Int J Mol Sci* 2013, **14:**3786-3801.
1492 47. Bao R, Fischer T, Bolognesi R, Brown SJ, Friedrich M: **Parallel duplication and partial**
1493 **subfunctionalization of beta-catenin/armadillo during insect evolution.** *Mol Biol Evol*
1494 2012, **29:**647-662.
1495 48. Sachs L, Chen YT, Drechsler A, Lynch JA, Panfilio KA, Lassig M, Berg J, Roth S: **Dynamic**
1496 **BMP signaling polarized by Toll patterns the dorsoventral axis in a hemimetabolous**
1497 **insect.** *eLife* 2015, **4:**e05502.
1498 49. Armisen D, Refki PN, Crumiere AJ, Viala S, Toubiana W, Khila A: **Predator strike shapes**
1499 **antipredator phenotype through new genetic interactions in water striders.** *Nat Commun*
1500 2015, **6:**8153.
1501 50. Konopova B, Smykal V, Jindra M: **Common and distinct roles of juvenile hormone**
1502 **signaling genes in metamorphosis of holometabolous and hemimetabolous insects.** *PLoS*
1503 *One* 2011, **6:**e28728.
1504 51. Vellichirammal NN, Gupta P, Hall TA, Brisson JA: **Ecdysone signaling underlies the pea**
1505 **aphid transgenerational wing polyphenism.** *Proc Natl Acad Sci U S A* 2017, **114:**1419-
1506 1423.
1507 52. Wulff JP, Sierra I, Sterkel M, Holtof M, Van Wielendaele P, Francini F, Broeck JV, Ons S:
1508 **Orcokinin neuropeptides regulate ecdysis in the hemimetabolous insect Rhodnius**
1509 **prolixus.** *Insect Biochem Mol Biol* 2017, **81:**91-102.
1510 53. Chiu TL, Wen Z, Rupasinghe SG, Schuler MA: **Comparative molecular modeling of**
1511 **Anopheles gambiae CYP6Z1, a mosquito P450 capable of metabolizing DDT.** *Proc Natl*
1512 *Acad Sci U S A* 2008, **105:**8855-8860.
1513 54. Gong Y, Li T, Feng Y, Liu N: **The function of two P450s, CYP9M10 and CYP6AA7, in**
1514 **the permethrin resistance of Culex quinquefasciatus.** *Sci Rep* 2017, **7:**587.
1515 55. Liu PZ, Kaufman TC: ***hunchback* is required for suppression of abdominal identity, and**
1516 **for proper germband growth and segmentation in the intermediate germband insect**
1517 ***Oncopeltus fasciatus*.** *Development* 2004, **131:**1515-1527.
1518 56. Schaeper ND, Pechmann M, Damen WGM, Prpic N-M, Wimmer EA: **Evolutionary**
1519 **plasticity of *collier* function in head development of diverse arthropods.** *Dev Biol* 2010,
1520 **344:**363-376.
1521 57. Aspiras AC, Smith FW, Angelini DR: **Sex-specific gene interactions in the patterning of**
1522 **insect genitalia.** *Dev Biol* 2011, **360:**369-380.
1523 58. Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, Najafabadi HS,
1524 Lambert SA, Mann I, Cook K, et al: **Determination and inference of eukaryotic**
1525 **transcription factor sequence specificity.** *Cell* 2014, **158:**1431-1443.
1526 59. Peel AD, Telford MJ, Akam M: **The evolution of hexapod engrailed-family genes:**
1527 **evidence for conservation and concerted evolution.** *Proc Biol Sci* 2006, **273:**1733-1742.
1528 60. Ben-David J, Chipman AD: **Mutual regulatory interactions of the trunk gap genes during**
1529 **blastoderm patterning in the hemipteran *Oncopeltus fasciatus*.** *Dev Biol* 2010, **346:**140-
1530 149.
1531 61. Erezyilmaz DF, Kelstrup HC, Riddiford LM: **The nuclear receptor E75A has a novel pair-**
1532 **rule-like function in patterning the milkweed bug, *Oncopeltus fasciatus*.** *Dev Biol* 2009,
1533 **334:**300-310.
1534 62. Liu PZ, Kaufman TC: ***even-skipped* is not a pair-rule gene but has segmental and gap-like**
1535 **functions in *Oncopeltus fasciatus*, an intermediate germband insect.** *Development* 2005,
1536 **132:**2081-2092.
1537 63. Weisbrod A, Cohen M, Chipman AD: **Evolution of the insect terminal patterning system--**
1538 **insights from the milkweed bug, *Oncopeltus fasciatus*.** *Dev Biol* 2013, **380:**125-131.
1539 64. Albertin CB, Simakov O, Mitros T, Wang ZY, Pungor JR, Edsinger-Gonzales E, Brenner S,
1540 Ragsdale CW, Rokhsar DS: **The octopus genome and the evolution of cephalopod neural**
1541 **and morphological novelties.** *Nature* 2015, **524:**220-224.
1542 65. Crooks GE, Hon G, Chandonia J-M, Brenner SE: **WebLogo: A sequence logo generator.**
1543 *Genome Res* 2004, **14:**1188-1190.
1544 66. Najafabadi HS, Mnaimneh S, Schmitges FW, Garton M, Lam KN, Yang A, Albu M,
1545 Weirauch MT, Radovani E, Kim PM, et al: **C2H2 zinc finger proteins greatly expand the**
1546 **human regulatory lexicon.** *Nat Biotechnol* 2015, **33:**555-562.

1547 67. Emerson RO, Thomas JH: **Adaptive evolution in zinc finger transcription factors.** *PLoS*
1548 *Genet* 2009, **5:**e1000325.
1549 68. Thomas JH, Schneider S: **Coevolution of retroelements and tandem zinc finger genes.**
1550 *Genome Res* 2011, **21:**1800-1812.
1551 69. Garcia-Perez JL, Widmann TJ, Adams IR: **The impact of transposable elements on**
1552 **mammalian development.** *Development* 2016, **143:**4101-4114.
1553 70. Liu PZ, Kaufman TC: ***Krüppel* is a gap gene in the intermediate insect *Oncopeltus***
1554 ***fasciatus* and is required for development of both blastoderm and germband-derived**
1555 **segments.** *Development* 2004, **131:**4567-4579.
1556 71. Heger P, Marin B, Bartkuhn M, Schierenberg E, Wiehe T: **The chromatin insulator CTCF**
1557 **and the emergence of metazoan diversity.** *Proc Natl Acad Sci USA* 2012, **109:**17507-17512.
1558 72. Liu H, Chang L-H, Sun Y, Lu X, Stubbs L: **Deep vertebrate roots for mammalian zinc**
1559 **finger transcription factor subfamilies.** *Genome Biol Evol* 2014, **6:**510-525.
1560 73. Imbeault M, Helleboid P-Y, Trono D: **KRAB zinc-finger proteins contribute to the**
1561 **evolution of gene regulatory networks.** *Nature* 2017, **543:**550-554.
1562 74. Csuros M, Rogozin IB, Koonin EV: **A detailed history of intron-rich eukaryotic ancestors**
1563 **inferred from a global survey of 100 complete genomes.** *PLoS Comput Biol* 2011,
1564 **7:**e1002150.
1565 75. Hoy MA, Waterhouse RM, Wu K, Estep AS, Ioannidis P, Palmer WJ, Pomerantz AF, Simao
1566 FA, Thomas J, Jiggins FM, et al: **Genome sequencing of the phytoseiid predatory mite**
1567 ***Metaseiulus occidentalis* reveals completely atomized Hox genes and superdynamic**
1568 **intron evolution.** *Genome Biol Evol* 2016, **8:**1762-1775.
1569 76. Seibt KM, Wenke T, Muders K, Truberg B, Schmidt T: **Short interspersed nuclear elements**
1570 **(SINEs) are abundant in Solanaceae and have a family-specific impact on gene structure**
1571 **and genome organization.** *Plant J* 2016, **86:**268-285.
1572 77. Huff JT, Zilberman D, Roy SW: **Mechanism for DNA transposons to generate introns on**
1573 **genomic scales.** *Nature* 2016, **538:**533-536.
1574 78. Wheeler D, Redding AJ, Werren JH: **Characterization of an ancient lepidopteran lateral**
1575 **gene transfer.** *PLoS One* 2012, **8:**e59262.
1576 79. Da Lage JL, Binder M, Hua-Van A, Janecek S, Casane D: **Gene make-up: rapid and**
1577 **massive intron gains after horizontal transfer of a bacterial alpha-amylase gene to**
1578 **Basidiomycetes.** *BMC Evol Biol* 2013, **13:**40.
1579 80. Lee DH, Short BD, Joseph SV, Bergh JC, Leskey TC: **Review of the biology, ecology, and**
1580 **management of *Halyomorpha halys* (Hemiptera: Pentatomidae) in China, Japan, and the**
1581 **Republic of Korea.** *Environ Entomol* 2013, **42:**627-641.
1582 81. Lawrence PA: **Cellular differentiation and pattern formation during metamorphosis of**
1583 **the milkweed bug Oncopeltus.** *Dev Biol* 1969, **19:**12-40.
1584 82. Riddiford LM: **Prevention of Metamorphosis by Exposure of Insect Eggs to Juvenile**
1585 **Hormone Analogs.** *Science* 1970, **167:**287-&.
1586 83. Willis JH, Lawrence PA: **Deferred Action of Juvenile Hormone.** *Nature* 1970, **225:**81-83.
1587 84. Masner P, Bowers WS, Kalin M, Muhle T: **Effect of precocene II on the endocrine**
1588 **regulation of development and reproduction in the bug, *Oncopeltus fasciatus*.** *Gen Comp*
1589 *Endocrinol* 1979, **37:**156-166.
1590 85. Rewitz K, O'Connor M, Gilbert L: **Molecular evolution of the insect Halloween family of**
1591 **cytochrome P450s: phylogeny, gene organization and functional conservation.** *Insect*
1592 *Biochem Mol Biol* 2007, **37:**741-753.
1593 86. Huet F, Ruiz C, Richards G: **Sequential gene activation by ecdysone in Drosophila**
1594 **melanogaster: the hierarchical equivalence of early and early late genes.** *Development*
1595 1995, **121:**1195-1204.
1596 87. Bialecki M, Shilton A, Fichtenberg C, Segraves WA, Thummel CS: **Loss of the ecdysteroid-**
1597 **inducible E75A orphan nuclear receptor uncouples molting from metamorphosis in**
1598 **Drosophila.** *Dev Cell* 2002, **3:**209-220.
1599 88. Charles JP, Iwema T, Epa VC, Takaki K, Rynes J, Jindra M: **Ligand-binding properties of a**
1600 **juvenile hormone receptor, Methoprene-tolerant.** *Proceedings of the National Academy of*
1601 *Sciences of the United States of America* 2011, **108:**21128-21133.
1602 89. Minakuchi C, Zhou X, Riddiford L: **Kruppel homolog 1 (Kr-h1) mediates juvenile**
1603 **hormone action during metamorphosis of Drosophila melanogaster.** *Mech Dev* 2008,
1604 **125:**91-105.

1605 90. Minakuchi C, Namiki T, Shinoda T: **Kruppel homolog 1, an early juvenile hormone-**
1606 **response gene downstream of Methoprene-tolerant, mediates its anti-metamorphic**
1607 **action in the red flour beetle Tribolium castaneum.** *Dev Biol* 2009, **352:**341-350.
1608 91. DiBello PR, Withers DA, Bayer CA, Fristrom JW, Guild GM: **The Drosophila Broad-**
1609 **Complex encodes a family of related proteins containing zinc fingers.** *Genetics* 1991,
1610 **129:**385-397.
1611 92. Karim F, Guild G, Thummel C: **The Drosophila Broad-Complex plays a key role in**
1612 **controlling ecdysone-regulated gene expression at the onset of metamorphosis.**
1613 *Development* 1993, **118:**977-988.
1614 93. Erezyilmaz DF, Riddiford LM, Truman JW: **The pupal specifier broad directs progressive**
1615 **morphogenesis in a direct-developing insect.** *Proceedings of the National Academy of*
1616 *Sciences of the United States of America* 2006, **103:**6925-6930.
1617 94. Arakane Y, Hogenkamp DG, Zhu YC, Kramer KJ, Specht CA, Beeman RW, Kanost MR,
1618 Muthukrishnan S: **Characterization of two chitin synthase genes of the red flour beetle,**
1619 ***Tribolium castaneum*, and alternate exon usage in one of the genes during development.**
1620 *Insect Biochem Mol Biol* 2004, **34:**291-304.
1621 95. True JR: **Insect melanism: the molecules matter.** *Trends in Ecology & Evolution* 2003,
1622 **18:**640-647.
1623 96. Zhan SA, Guo QH, Li MH, Li MW, Li JY, Miao XX, Huang YP: **Disruption of an N-**
1624 **acetyltransferase gene in the silkworm reveals a novel role in pigmentation.** *Development*
1625 2010, **137:**4083-4090.
1626 97. Liu J, Lemonds TR, Popadic A: **The genetic control of aposematic black pigmentation in**
1627 **hemimetabolous insects: insights from Oncopeltus fasciatus.** *Evolution & Development*
1628 2014, **16:**270-277.
1629 98. Liu J, Lemonds TR, Marden JH, Popadic A: **A Pathway Analysis of Melanin Patterning in**
1630 **a Hemimetabolous Insect.** *Genetics* 2016, **203:**403-413.
1631 99. Lawrence PA: **Some new mutants of large milkweed bug *Oncopeltus fasciatus* Dall.**
1632 *Genetical Research* 1970, **15:**347-350.
1633 100. Morgan ED: *Biosynthesis in Insects: Advanced Edition.* London: Royal Society of Chemistry,;
1634 2010.
1635 101. McLean JR, Krishnakumar S, O'Donnell JM: **Multiple mRNAs from the Punch locus of**
1636 **Drosophila melanogaster encode isoforms of GTP cyclohydrolase I with distinct N-**
1637 **terminal domains.** *J Biol Chem* 1993, **268:**27191-27197.
1638 102. Wiederrecht GJ, Paton DR, Brown GM: **Enzymatic Conversion of Dihydroneopterin**
1639 **Triphosphate to the Pyrimidodiazepine Intermediate Involved in the Biosynthesis of the**
1640 **Drosopterins in Drosophila-Melanogaster.** *Journal of Biological Chemistry* 1984,
1641 **259:**2195-2200.
1642 103. Newcombe D, Blount JD, Mitchell C, Moore AJ: **Chemical egg defence in the large**
1643 **milkweed bug, Oncopeltus fasciatus, derives from maternal but not paternal diet.**
1644 *Entomologia Experimentalis et Applicata* 2013, **149:**197-205.
1645 104. Zhan S, Merlin C, Boore JL, Reppert SM: **The monarch butterfly genome yields insights**
1646 **into long-distance migration.** *Cell* 2011, **147:**1171-1185.
1647 105. Smadja C, Shi P, Butlin RK, Robertson HM: **Large gene family expansions and adaptive**
1648 **evolution for odorant and gustatory receptors in the pea aphid, Acyrthosiphon pisum.**
1649 *Mol Biol Evol* 2009, **26:**2073-2086.
1650 106. Terrapon N, Li C, Robertson HM, Ji L, Meng X, Booth W, Chen Z, Childers CP, Glastad KM,
1651 Gokhale K, et al: **Molecular traces of alternative social organization in a termite genome.**
1652 *Nat Commun* 2014, **5:**3636.
1653 107. Croset V, Rytz R, Cummins SF, Budd A, Brawand D, Kaessmann H, Gibson TJ, Benton R:
1654 **Ancient protostome origin of chemosensory ionotropic glutamate receptors and the**
1655 **evolution of insect taste and olfaction.** *PLoS Genet* 2010, **6:**e1001064.
1656 108. Kirkness EF, Haas BJ, Sun W, Braig HR, Perotti MA, Clark JM, Lee SH, Robertson HM,
1657 Kennedy RC, Elhaik E, et al: **Genome sequences of the human body louse and its primary**
1658 **endosymbiont provide insights into the permanent parasitic lifestyle.** *Proc Natl Acad Sci*
1659 *U S A* 2010, **107:**12168-12173.
1660 109. Robertson HM, Warr CG, Carlson JR: **Molecular evolution of the insect chemoreceptor**
1661 **gene superfamily in Drosophila melanogaster.** *Proc Natl Acad Sci U S A* 2003, **100 Suppl**
1662 **2:**14537-14542.
1663 110. Joseph RM, Carlson JR: **Drosophila Chemoreceptors: A Molecular Interface Between the**
1664 **Chemical World and the Brain.** *Trends Genet* 2015, **31:**683-695.

111. Benton R: **Multigene Family Evolution: Perspectives from Insect Chemoreceptors.** *Trends Ecol Evol* 2015, **30:**590-600.
112. 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.
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. 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.
116. 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.
117. Hojilla-Evangelista MP, Evangelista RL: **Characterization of milkweed (Asclepias spp.) seed proteins.** *Industrial crops and ...* 2009.
118. 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**.
119. 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.
120. 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.
121. Grau-Bove X, Ruiz-Trillo I, Irimia M: **Origin of exon skipping-rich transcriptomes in animals driven by evolution of gene architecture.** *Genome Biol* 2018, **19:**135.
122. 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.
123. 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.
124. Hare EE, Johnston JS: **Genome size determination using flow cytometry of propidium iodide-stained nuclei.** *Methods Mol Biol* 2011, **772:**3-12.
125. Marcais G, Kingsford C: **A fast, lock-free approach for efficient parallel counting of occurrences of k-mers.** *Bioinformatics* 2011, **27:**764-770.
126. Bushnell B: **BBMap short read aligner.** 2016.
127. 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.
128. Holt C, Yandell M: **MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects.** *BMC Bioinformatics* 2011, **12:**491.
129. 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**.
130. 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.
131. 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.
132. 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.
133. 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.

134. **RepeatModeler Open-1.0.8** [http://www.repeatmasker.org]
135. Bao Z, Eddy SR: **Automated de novo identification of repeat sequence families in sequenced genomes.** *Genome Res* 2002, **12:**1269-1276.
136. Price AL, Jones NC, Pevzner PA: **De novo identification of repeat families in large genomes.** *Bioinformatics* 2005, **21 Suppl 1:**i351-358.
137. Benson G: **Tandem repeats finder: a program to analyze DNA sequences.** *Nucleic Acids Res* 1999, **27:**573-580.
138. **RepeatMasker Open-4.0.** [http://www.repeatmasker.org]
139. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2.** *Nat Methods* 2012, **9:**357-359.
140. 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**.
141. 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.
142. 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.
143. Li B, Dewey CN: **RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome.** *BMC Bioinformatics* 2011, **12:**323.
144. Schoville SD, Chen YH, Andersson MN, Benoit JB, Bhandari A, Bowsher JH, Brevik K, Cappelle K, Chen MM, Childers AK, et al: **A model species for agricultural pest genomics: the genome of the Colorado potato beetle, *Leptinotarsa decemlineata* (Coleoptera: Chrysomelidae).** *Sci Rep* 2018, **8:**1931.
145. Scolari F, Benoit JB, Michalkova V, Aksoy E, Takac P, Abd-Alla AM, Malacrida AR, Aksoy S, Attardo GM: **The spermatophore in *Glossina morsitans morsitans*: Insights into male contributions to reproduction.** *Sci Rep* 2016, **6:**20334.
146. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM: **BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs.** *Bioinformatics* 2015, **31:**3210-3212.
147. 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.
148. Weirauch MT, Hughes TR: **A catalogue of eukaryotic transcription factor types, their evolutionary origin, and species distribution.** *Subcell Biochem* 2011, **52:**25-73.
149. Eddy SR: **A new generation of homology search tools based on probabilistic inference.** *Genome Inform* 2009, **23:**205-211.
150. 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.
151. 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.
152. Claudel-Renard C, Chevalet C, Faraut T, Kahn D: **Enzyme-specific profiles for genome annotation: PRIAM.** *Nucleic Acids Research* 2003, **31:**6633-6639.
153. Conesa A, Götz S: **Blast2GO: A Comprehensive Suite for Functional Analysis in Plant Genomics.** *International journal of plant genomics* 2008, **2008:**619832.
154. 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.
155. 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.
156. 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.
157. 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**

1784 **pathway/genome informatics and systems biology.** *Briefings in Bioinformatics* 2010,
1785 **11:**40-79.
1786 158. Dereeper A, Guignon V, Blanc G, Audic S, Buffet S, Chevenet F, Dufayard JF, Guindon S,
1787 Lefort V, Lescot M, et al: **Phylogeny.fr: robust phylogenetic analysis for the non-**
1788 **specialist.** *Nucleic Acids Res* 2008, **36:**W465-469.
1789 159. Wang X, Fang X, Yang P, Jiang X, Jiang F, Zhao D, Li B, Cui F, Wei J, Ma C, et al: **The**
1790 **locust genome provides insight into swarm formation and long-distance flight.** *Nat*
1791 *Commun* 2014, **5:**2957.
1792 160. The International Silkworm Genome Consortium: **The genome of a lepidopteran model**
1793 **insect, the silkworm Bombyx mori.** *Insect Biochem Mol Biol* 2008, **38:**1036-1045.
1794 161. Elsik CG, Worley KC, Bennett AK, Beye M, Camara F, Childers CP, de Graaf DC, Debyser
1795 G, Deng J, Devreese B, et al: **Finding the missing honey bee genes: lessons learned from a**
1796 **genome upgrade.** *BMC Genomics* 2014, **15:**86.
1797 162. Honeybee Genome Sequencing Consortium: **Insights into social insects from the genome of**
1798 **the honeybee Apis mellifera.** *Nature* 2006, **443:**931-949.
1799 163. Richards S, Gibbs RA, Weinstock GM, Brown SJ, Denell R, Beeman RW, Gibbs R, Bucher
1800 G, Friedrich M, Grimmelikhuijzen CJ, et al: **The genome of the model beetle and pest**
1801 **Tribolium castaneum.** *Nature* 2008, **452:**949-955.
1802 164. Chen XG, Jiang X, Gu J, Xu M, Wu Y, Deng Y, Zhang C, Bonizzoni M, Dermauw W, Vontas
1803 J, et al: **Genome sequence of the Asian Tiger mosquito, Aedes albopictus, reveals insights**
1804 **into its biology, genetics, and evolution.** *Proc Natl Acad Sci U S A* 2015, **112:**E5907-5915.
1805 165. Armisen D, Rajakumar R, Friedrich M, Benoit JB, Robertson HM, Panfilio KA, Ahn S-J,
1806 Poelchau MF, Chao H, Dinh H, et al: **The genome of the water strider *Gerris buenoi***
1807 **reveals expansions of gene repertoires associated with adaptations to life on the water.**
1808 *BMC Genomics* in press**:**acceptance e-mail 12 Oct. 2018.
1809 166. Papanicolaou A, Schetelig MF, Arensburger P, Atkinson PW, Benoit JB, Bourtzis K,
1810 Castañera P, Cavanaugh JP, Chao H, Childers C, et al: **The whole genome sequence of the**
1811 **Mediterranean fruit fly, *Ceratitis capitata* (Wiedemann), reveals insights into the biology**
1812 **and adaptive evolution of a highly invasive pest species.** *Genome Biol* 2016, **17:**192.
1813 167. Ellis LL, Huang W, Quinn AM, Ahuja A, Alfrejd B, Gomez FE, Hjelmen CE, Moore KL,
1814 Mackay TF, Johnston JS, Tarone AM: **Intrapopulation genome size variation in D.**
1815 **melanogaster reflects life history variation and plasticity.** *PLoS Genet* 2014, **10:**e1004522.
1816 168. Kumar S, Stecher G, Tamura K: **MEGA7: Molecular Evolutionary Genetics Analysis**
1817 **Version 7.0 for Bigger Datasets.** *Mol Biol Evol* 2016, **33:**1870-1874.
1818

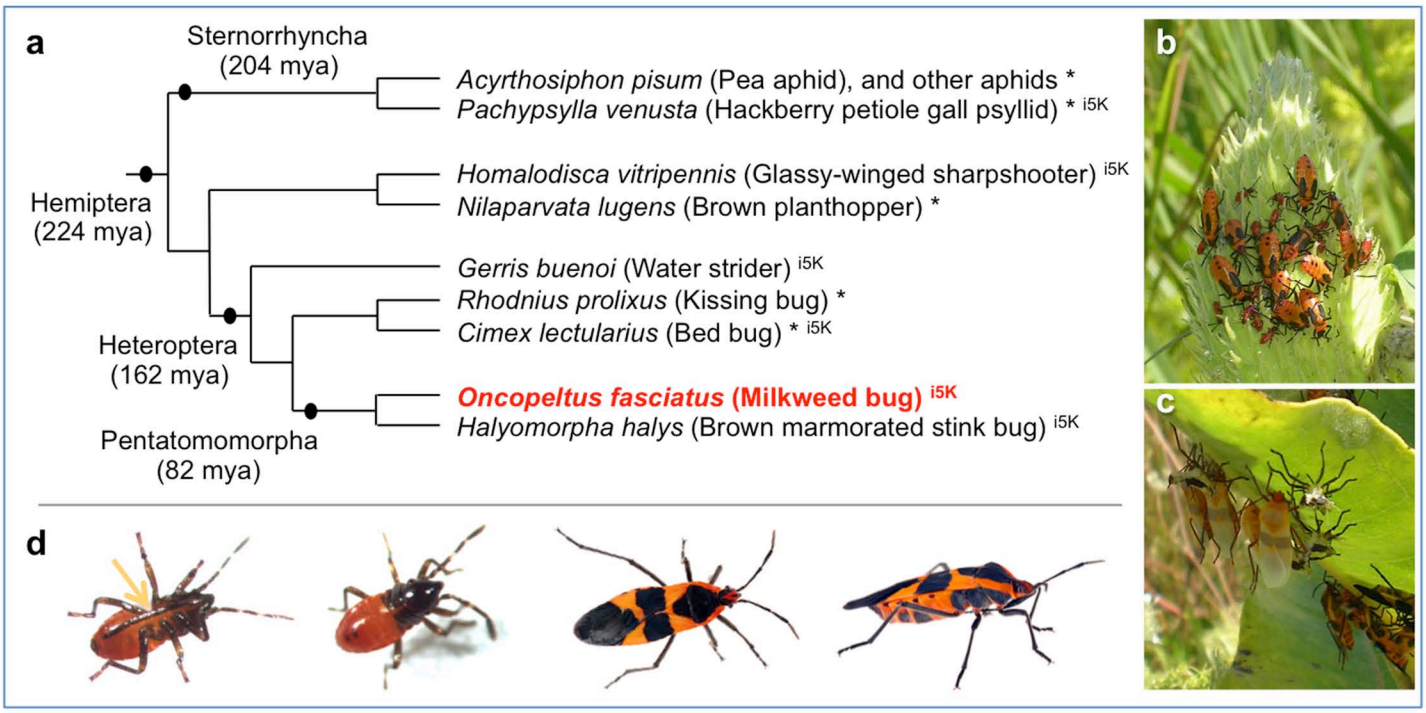


Fig 1. The large milkweed bug, *Oncopeltus fasciatus*, shown in its phylogenetic and environmental context.

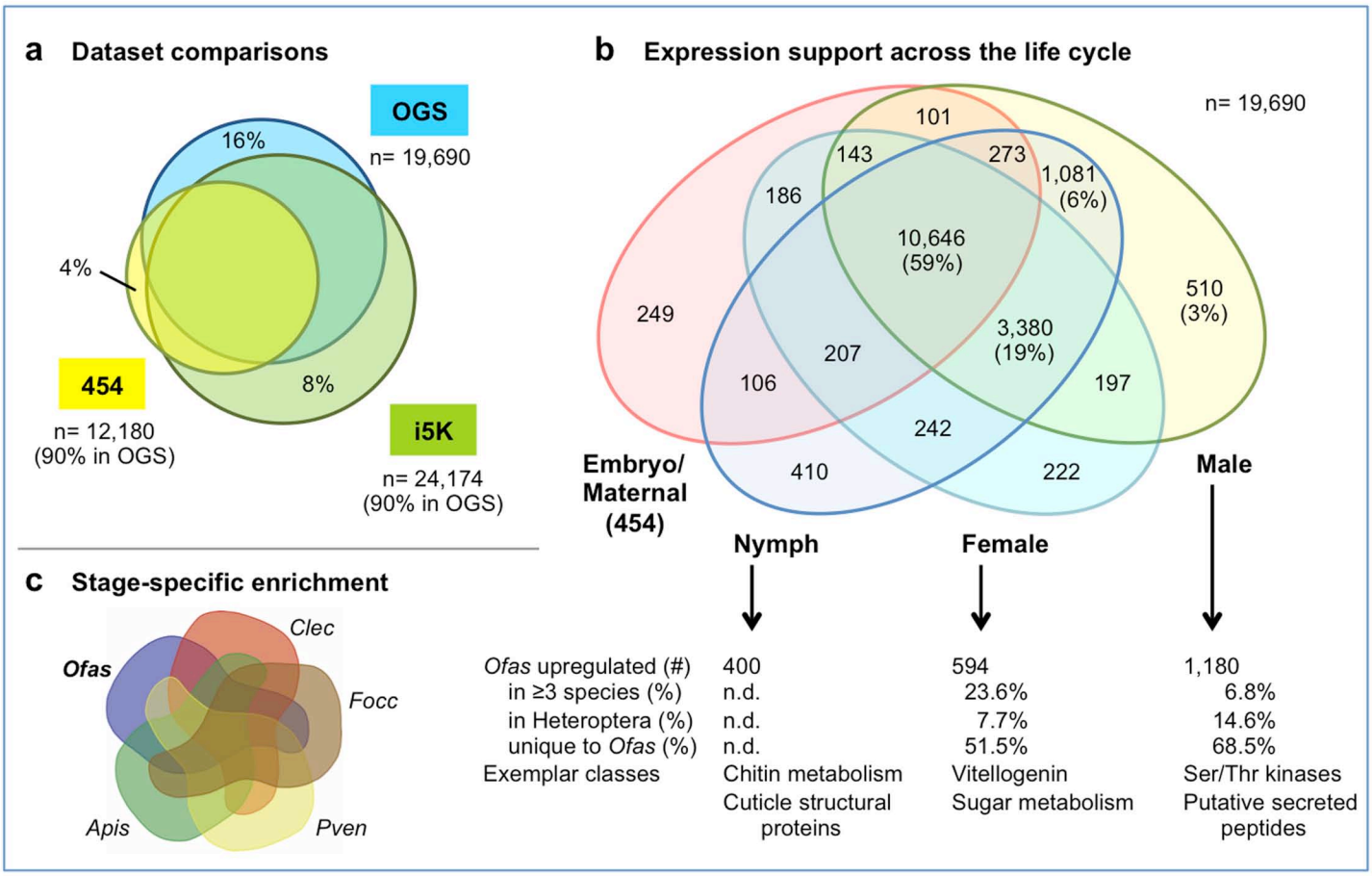


Fig 2. Comparisons of the official gene set and transcriptomic resources.

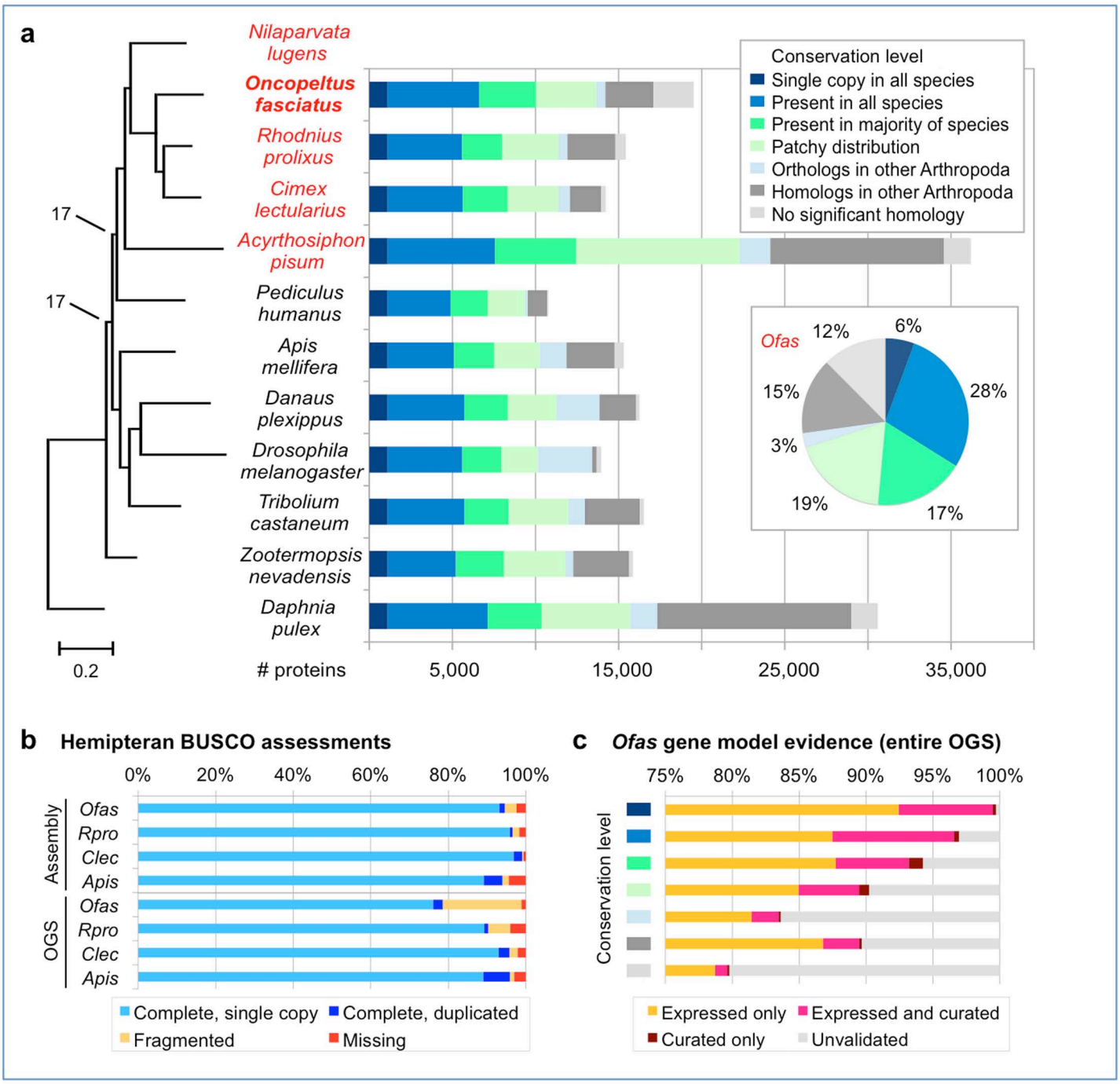


Fig 3. Orthology comparisons and phylogenetic placement of *Oncopeltus fasciatus* among other Arthropoda.

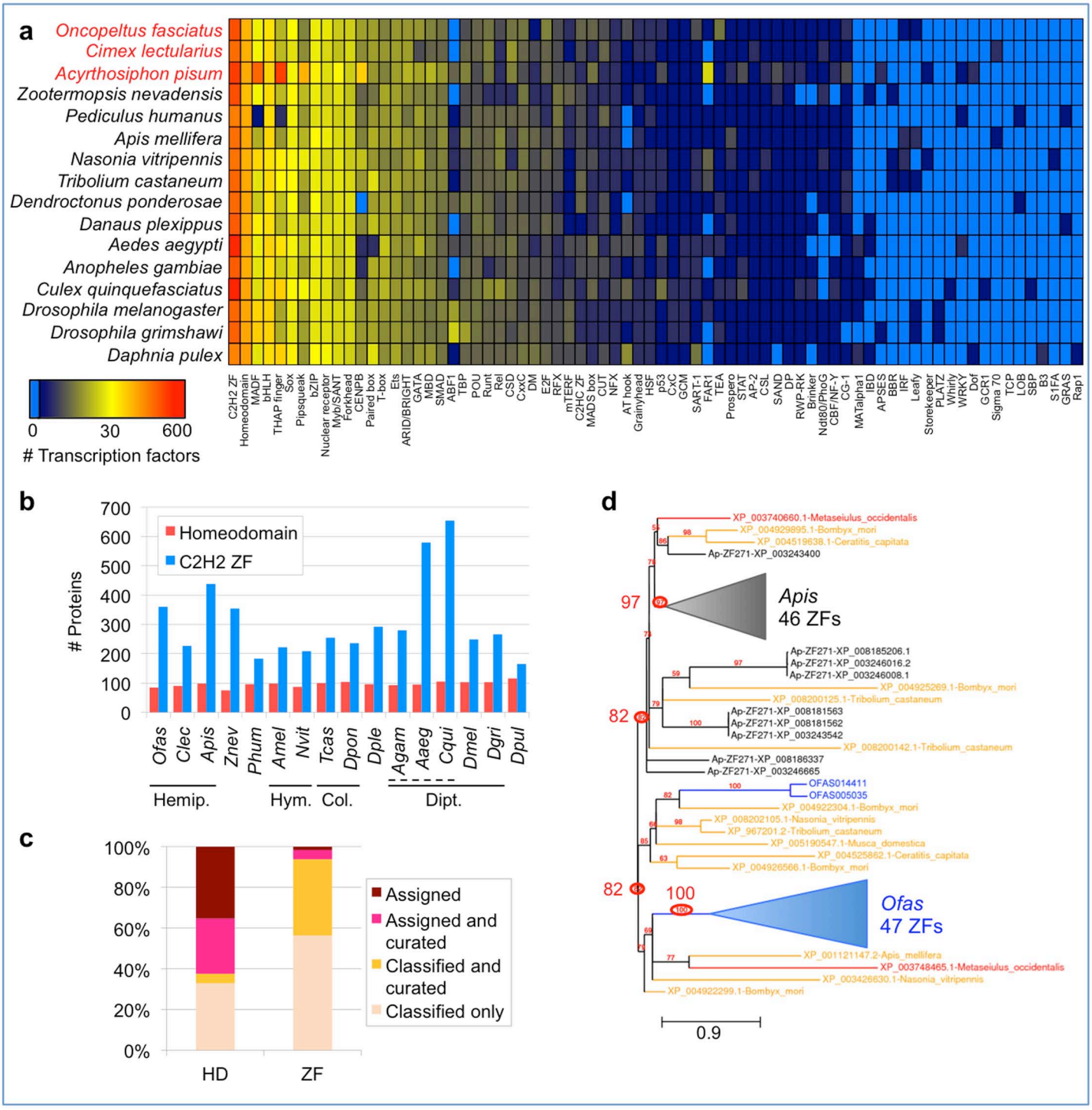


Fig 4. Distribution of transcription factor families across insect genomes.

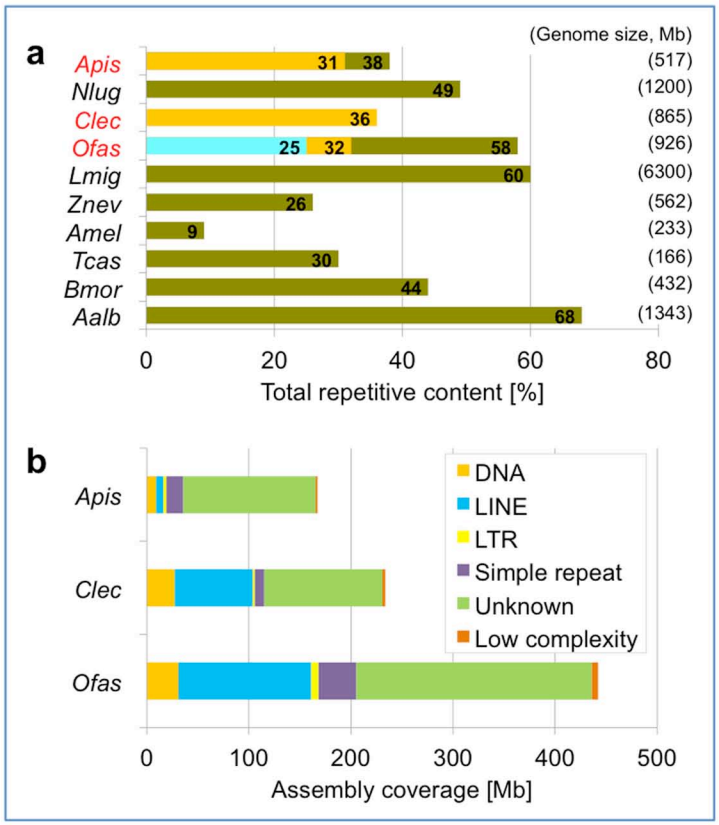


Fig 5. Comparison of repeat content estimations.

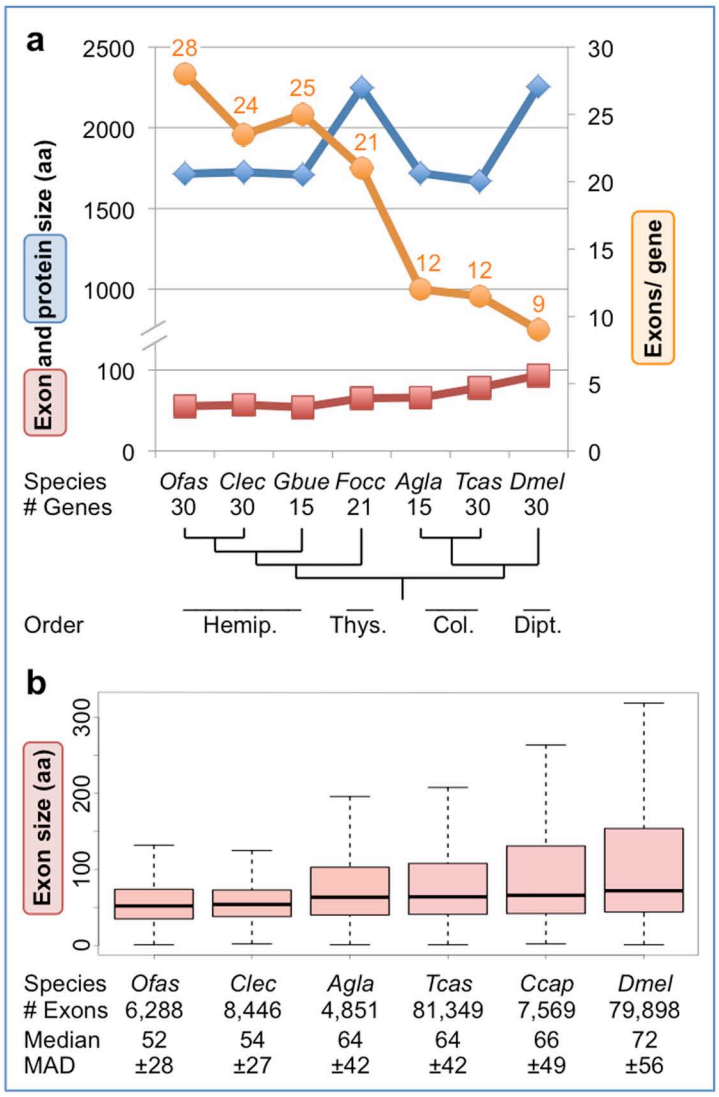


Fig 6. Trends in gene structure show hemipteroid-specific tendencies.

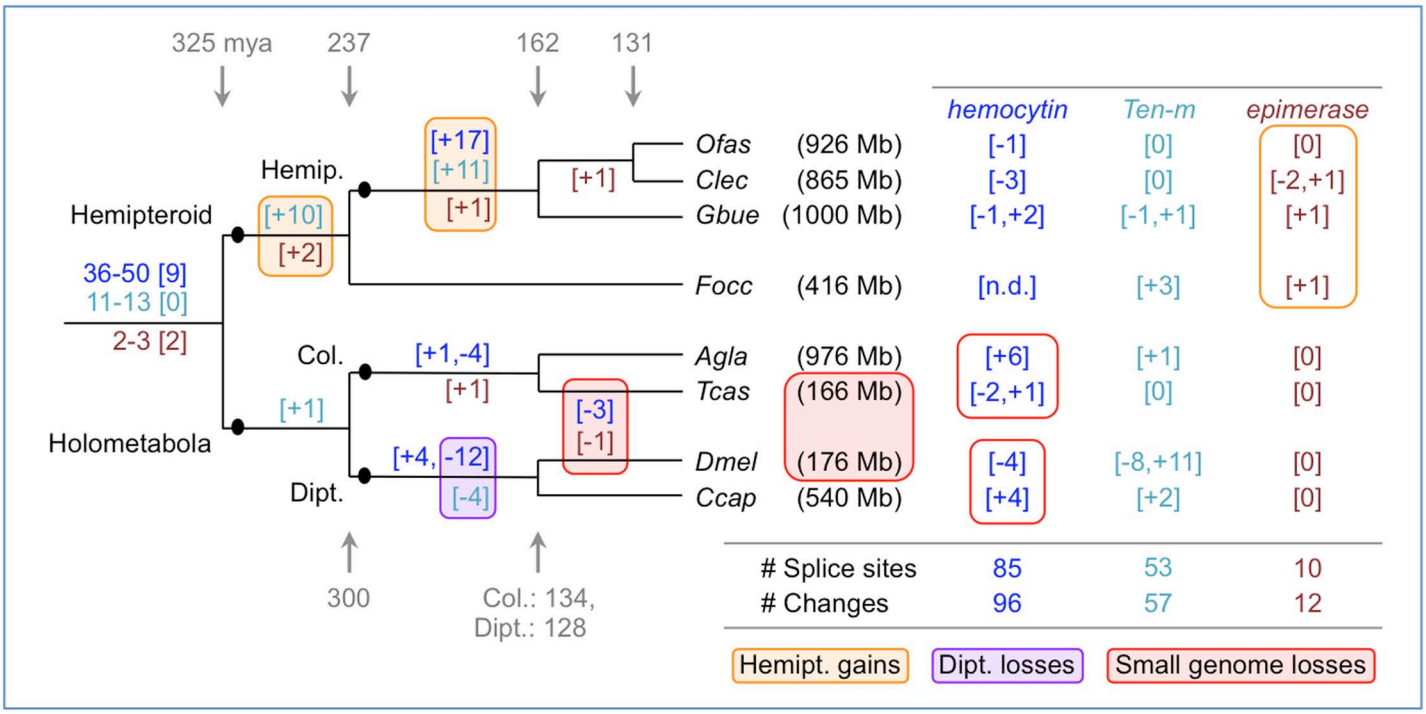


Fig 7. Splice site evolution correlates with both lineage and, independently, genome size.

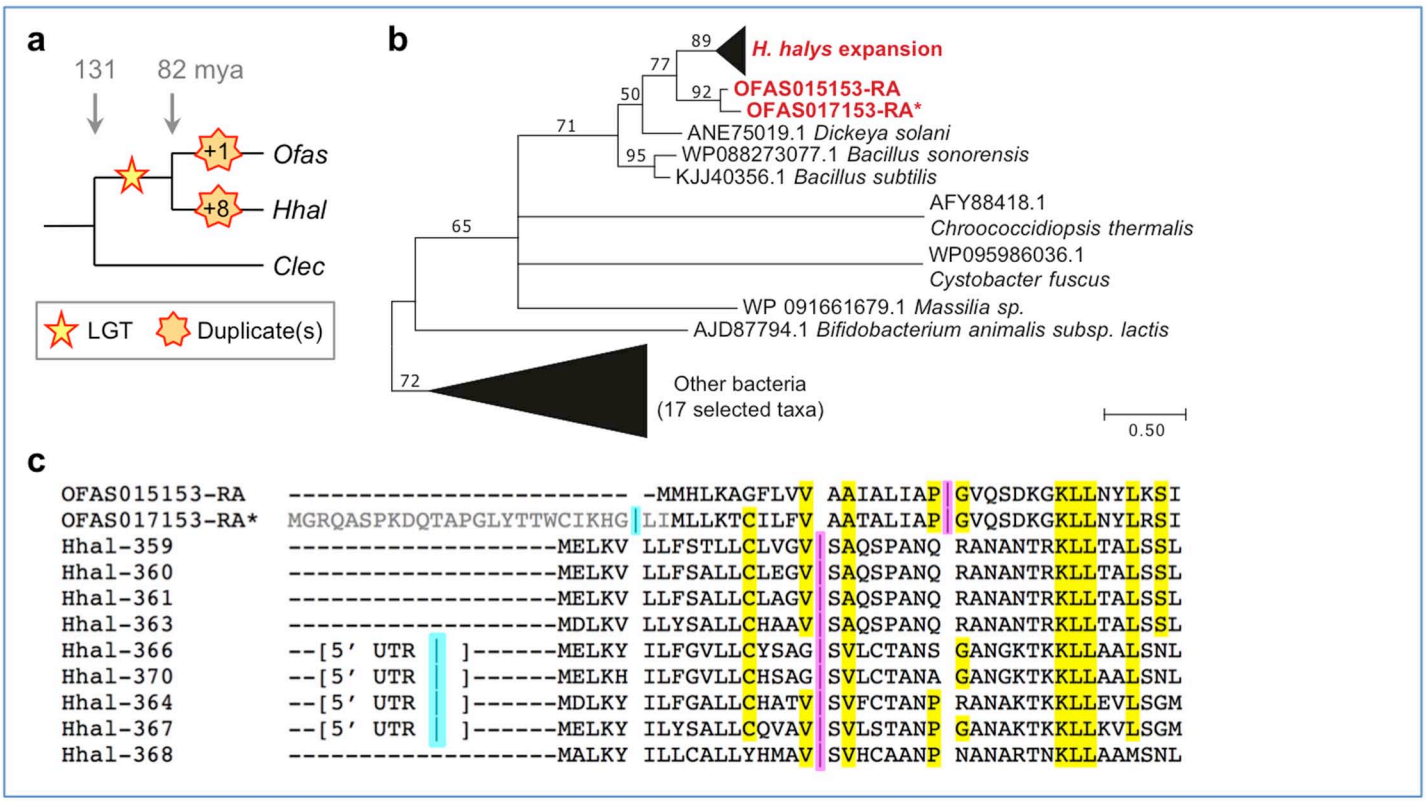
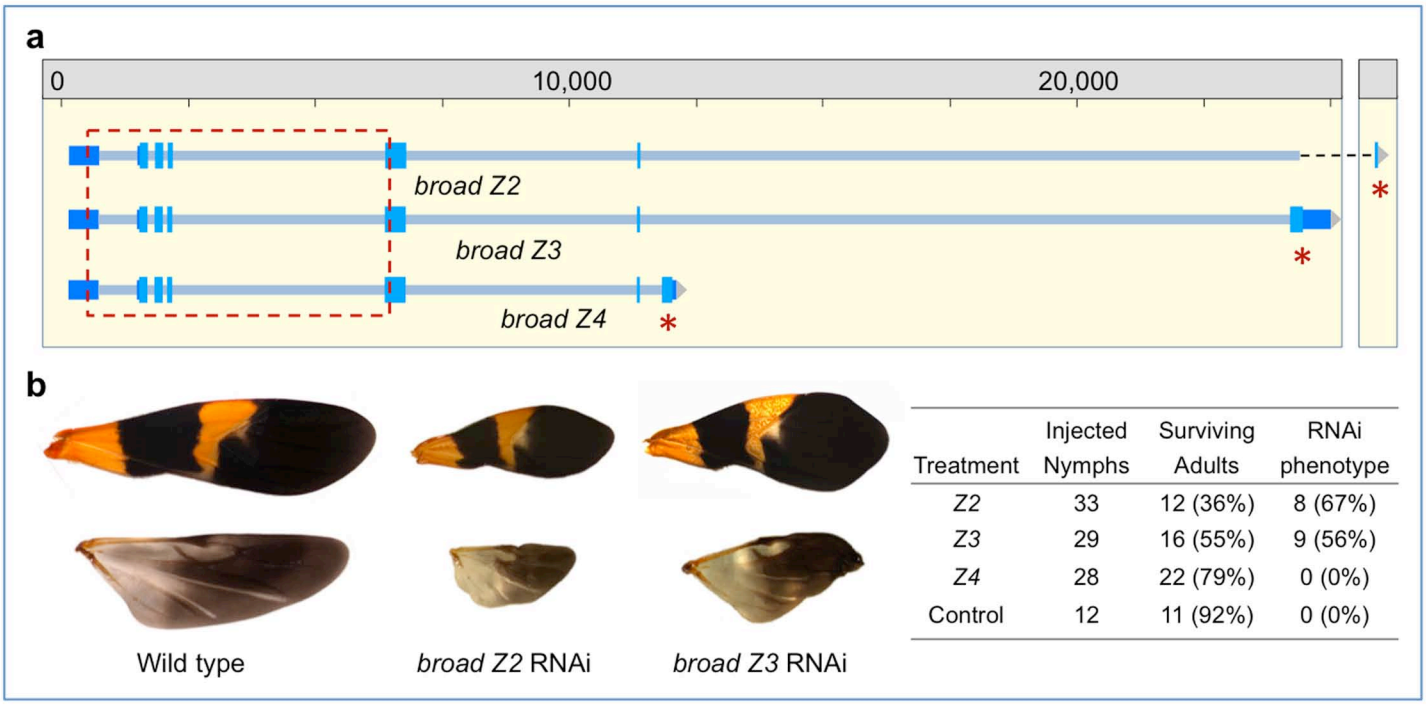


Fig 8. Lateral gene transfer introduction and subsequent evolution within the Hemiptera for mannosidase-encoding genes.



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*.

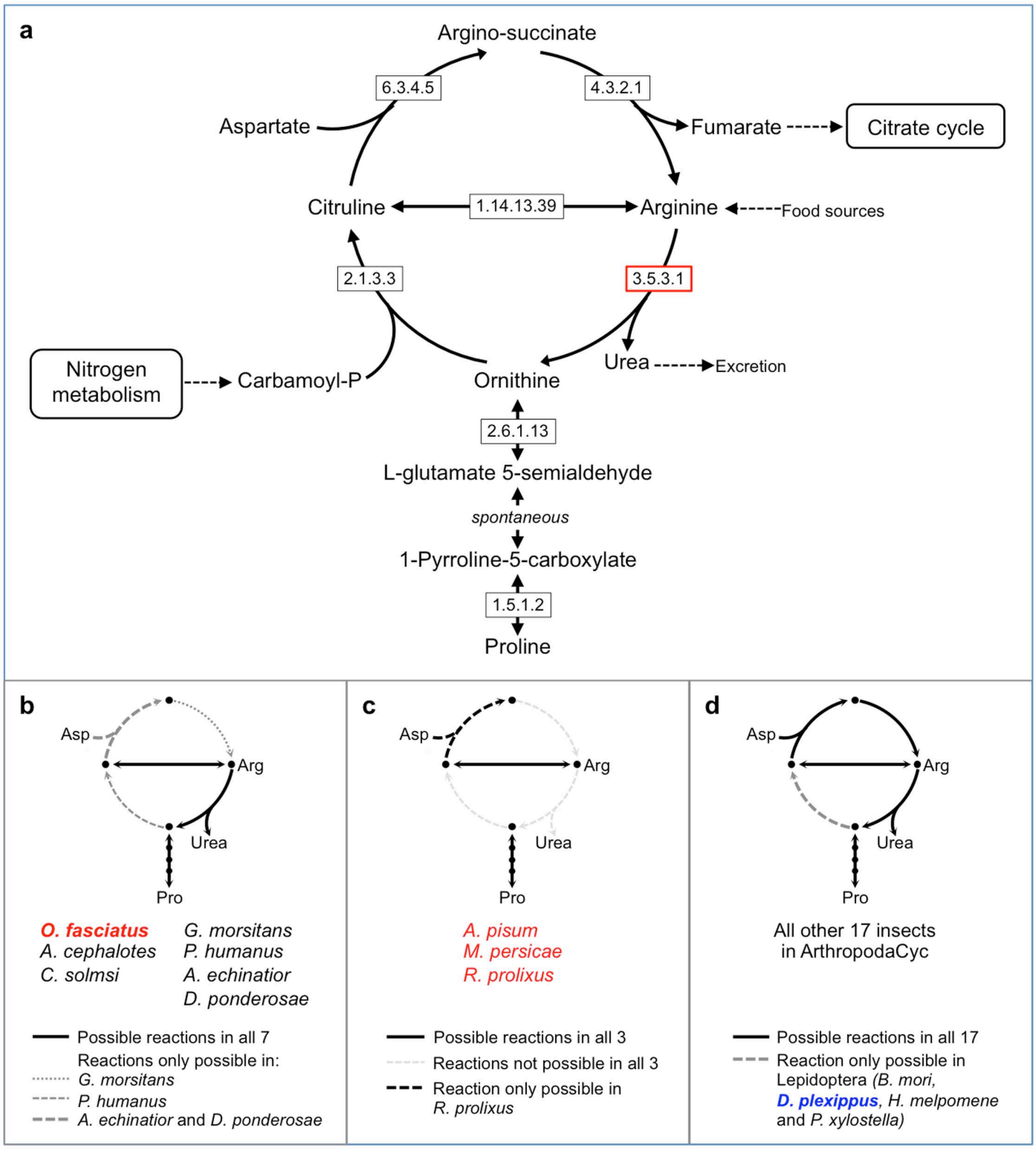


Fig 10. Comparison of the urea cycle of *Oncopeltus* with 26 other insect species.