https://doi.org/10.1186/s13059-019-1660-0

Genome Biology

RESEARCH Open Access

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[4,43], 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[23,45], 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], Harsha Vardhan 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[23,45], 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[23,45], 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 L. 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[40,44], 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] and Stephen Richards[13]

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.

(Continued on next page)

* Correspondence: kristen.panfilio@alum.swarthmore.edu
[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
Full list of author information is available at the end of the article

(Continued from previous page)

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 the 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 lifestyle 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 lifestyle 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 the diversity within an order, and the Hemiptera provide an excellent opportunity to explore this.

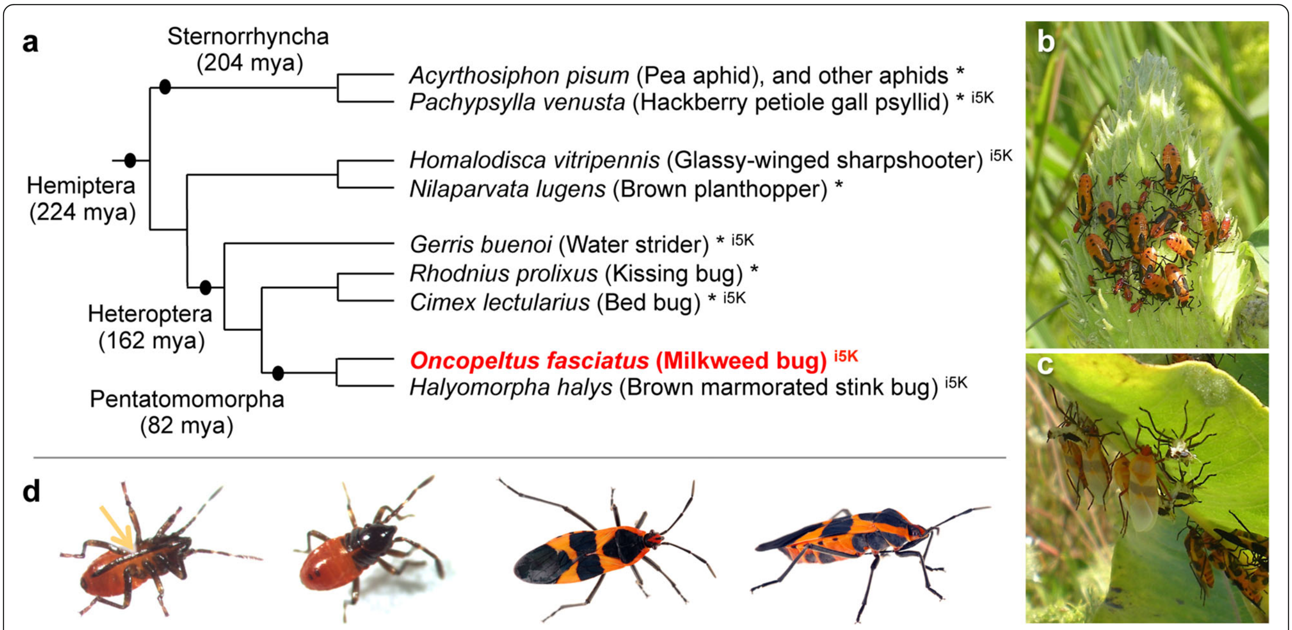

a
Sternorrhyncha
(204 mya)
Acyrthosiphon pisum (Pea aphid), and other aphids *
Pachypsylla venusta (Hackberry petiole gall psyllid) * i5K
Hemiptera
(224 mya)
Homalodisca vitripennis (Glassy-winged sharpshooter) i5K
Nilaparvata lugens (Brown planthopper) *
Gerris buenoi (Water strider) * i5K
Rhodnius prolixus (Kissing bug) *
Cimex lectularius (Bed bug) * i5K
Heteroptera
(162 mya)
Oncopeltus fasciatus (Milkweed bug) i5K
Halyomorpha halys (Brown marmorated stink bug) i5K
Pentatomomorpha
(82 mya)
b
c
d

Fig. 1 The large milkweed bug, *Oncopeltus fasciatus*, shown in its phylogenetic and environmental context. **a** Species tree of selected Hemiptera with genomic and transcriptomic resources, based on phylogenetic analyses and divergence time estimates in [3]. Species marked with an asterisk (*) have published resources; those with the appellation "i5K" are part of a current pilot project supported by the Baylor College of Medicine Human Genome Sequencing Center and the National Agricultural Library of the USDA. Note that recent analyses suggest the traditional infraorder Cimicomorpha, to which *Rhodnius* and *Cimex* belong, may be paraphyletic [16]. **b**, **c** Milkweed bugs on their native food source, the milkweed plant: gregarious nymphs of different instars on a milkweed seed pod (**b**) and pale, recently eclosed adults and their shed exuvia (**c**). Images were taken at Avalon Park and Preserve, Stony Brook, NY, USA, courtesy of Deniz Erezyilmaz, used with permission. **d** Individual bugs, shown from left to right: first instar nymphs (ventral and dorsal views) and adults (dorsal and lateral views); images courtesy of Kristen Panfilio (nymphs) and Jena Johnson (adults), used with permission. The arrow labels the labium (the "straw"), part of the hemipteran mouthpart anatomy adapted for feeding by piercing and sucking

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 15 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 family 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 the "Methods" section and Additional file 1: 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 (Additional file 1: 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], Additional file 1: Supplemental Note 2.1.b). In further analyses, we indeed obtained high estimates of repetitive content, although heterozygosity does not unduly influence gene

Table 1 *Oncopeltus fasciatus* genome metrics

Feature	Value	
2*n* chromosomes	16	
Genome size	926 Mb (mean between males and females)	
Assembly size	1099 Mb (contigs only, 774 Mb)	
Coverage	106.9× raw coverage, 83.7% of reads in final assembly	
Contig N50	4047 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] (1426 models, 7.2%)	19,465 genes (1201 genes, 6.2%)
OGS v1.2 (curated fraction)	19,809 models[1] (1697 models, 8.7%)	19,616 genes (1518 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

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 the 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 (Additional file 1: Supplemental Note 2.2).

The official gene set and conserved gene linkage

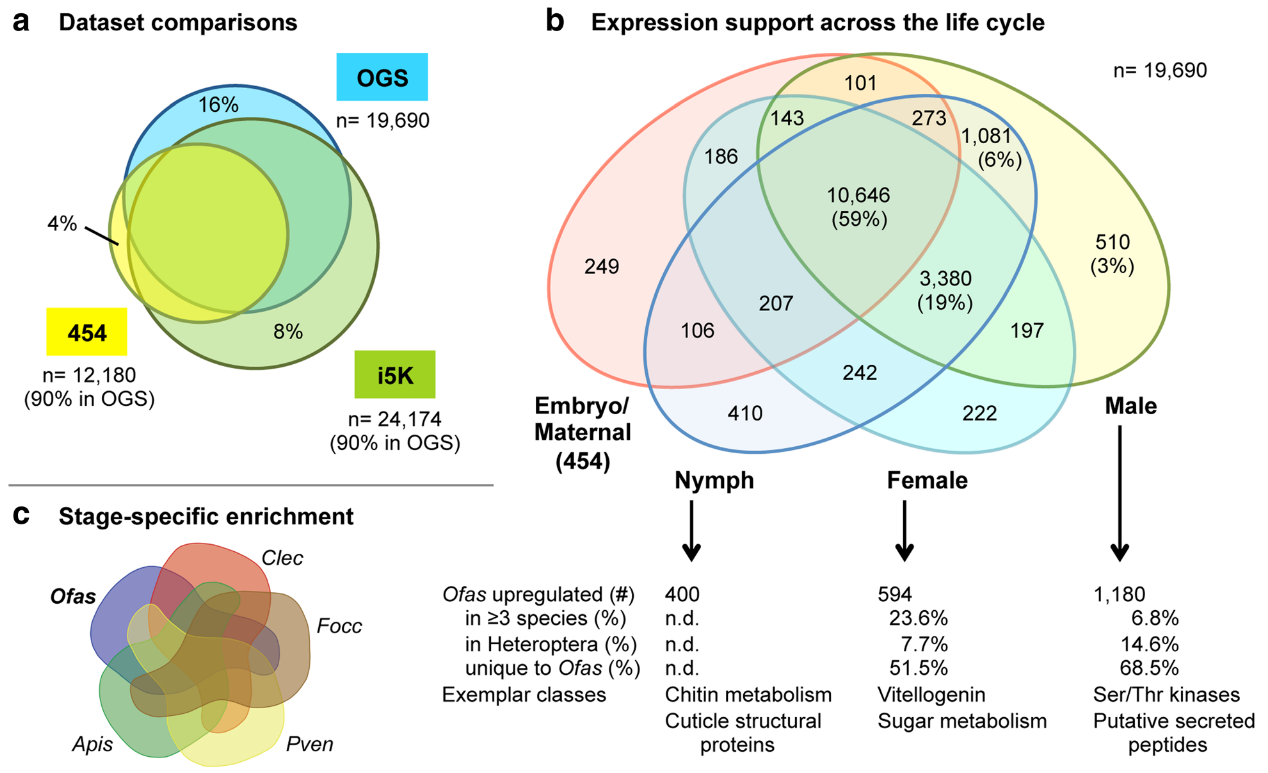
The official gene set (OGS) was generated by automatic annotation followed by manual curation in a large-scale effort by the research community (Additional file 1: Supplemental Notes Sections 3–4). Curation revised automatic models, added alternative isoforms and de novo models, and documented multiple 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 (Additional file 1: Supplemental Note 4, and 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 1697 gene models (8.7% of OGS v1.2), including 316 de novo models (Additional file 2: Table S4.1, Additional file 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 ([27, 33, 34], Additional file 1: 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 (Additional file 1: Supplemental Notes 5.1.a, 5.1.c, 5.1.i, 5.1.j). Further evidence for correct scaffold assembly comes from the curation of large, multi-exonic loci. For example, the cell polarity and cytoskeletal regulator encoded by the conserved *furry* gene includes 47 exons spanning a 437-kb locus, which were all correctly assembled on a single scaffold.

Gene expression profiles across the milkweed bug life cycle

To augment published transcriptomic resources [35, 36], we sequenced three different post-embryonic samples ("i5K" dataset, see the "Methods" section). We then compared the OGS to the resulting de novo transcriptome and to a previously published embryonic and maternal (ovary) transcriptome ("454" pyrosequencing dataset, [35]). Our OGS is quite comprehensive, containing 90% of transcripts from each transcriptomic dataset (Fig. 2a). The OGS also contains an additional 3146 models (16% of OGS) not represented in either transcriptome, including 163 de novo models encoding chemoreceptors. Such genes are known for lineage-specific expansions and highly tissue- and stage-specific expression ([37, 38], and see below), and our OGS captures these genes with rare transcripts.

The OGS also incorporates many partial and unidentified 454 transcripts, nearly trebling the transcripts with



	Nymph	Female	Male
Ofas upregulated (#)	400	594	1,180
in ≥3 species (%)	n.d.	23.6%	6.8%
in Heteroptera (%)	n.d.	7.7%	14.6%
unique to *Ofas* (%)	n.d.	51.5%	68.5%
Exemplar classes	Chitin metabolism Cuticle structural proteins	Vitellogenin Sugar metabolism	Ser/Thr kinases Putative secreted peptides

Fig. 2 Comparisons of the official gene set and transcriptomic resources for *Oncopeltus fasciatus*. **a** Area-proportional Venn diagram comparing the OGS v1.1 (OGS), a Trinity de novo transcriptome from the three post-embryonic RNA-seq samples (i5K) and the maternal and embryonic transcriptome from 454 data ("454", [35]). Sample sizes and the fraction of each transcriptome represented in the OGS are indicated (for the 454 dataset, only transcripts with homology identification were considered). The unique fraction of each set is also specified (%). Dataset overlaps were determined by blastn (best hit only, *e* value < 10^{-9}). **b** Venn diagram of gene model expression support across four life history samples. Values are numbers of gene models, with percentages also given for the largest subsets. Note that the "Embryo/Maternal" sample derives from 454 pyrosequencing data and therefore has a smaller data volume than the other, Illumina-based samples. **c** Summary of sex- and developmental stage-specific RNA-seq comparisons across hemipteroid species: *Apis*, *Acyrthosiphon pisum*; *Clec*, *Cimex lectularius*; *Focc*, *Frankliniella occidentalis* (thysanopteran outgroup); *Ofas*, *Oncopeltus fasciatus*; *Pven*, *Pachypsylla venusta*; n.d., not determined. For complete numerical details, see Additional file 1: Supplemental Note 2.4. Analyses are based on OGS v1.1

an assigned gene model or homology compared to the original study (from 9 to 26%, by blastn, $e < 10^{-9}$). This included 10,130 transcripts that primarily mapped to UTRs and previously lacked recognizable coding sequence, such as for the *Oncopeltus brinker* ortholog, a BMP pathway component ([39], Additional file 1: Supplemental Note 5.1.f), and the enzyme-encoding genes *CTP synthase* and *roquin*. At the same time, the transcriptomes provided expression support for the identification of multiple isoforms in the OGS, such as for the germline determinant *nanos* [35]. More generally, most OGS gene models have expression support (91% of 19,690), with 74% expressed broadly in at least 3 of 4 samples (Fig. 2b). The inclusion of a fifth dataset from a published adult library [36] provided only a 1% gain in expression support, indicating that with the current study the expression data volume for *Oncopeltus* is quite complete.

RNA-seq studies were further conducted to establish male-, female-, and nymph-specific gene sets (Fig. 2b, c, Additional file 1: Supplemental Note 2.4), from which we also infer that the published adult dataset of unspecified sex is probably male. Moreover, most genes with stage-restricted or stage-enriched expression are in our male sample (Fig. 2b, c). For example, gustatory receptor (GR) genes show noticeable restriction to the adult male and published adult (probable male) samples ($n = 169$ GRs: 40% no expression, 27% only expressed in these two samples), with half of these expressed in both biological replicates (52%). Interestingly, the nymphal sample is enriched for genes encoding structural cuticular proteins (94%, which is > 56% more than any other sample). This likely reflects the ongoing molting cycles, with their cyclical upregulation of chitin metabolism and cuticular gene synthesis [40], that are experienced by the different instars and molt cycle stages of individuals pooled in this sample. Lastly, gene sets with sex-specific enrichment across several hemipteroid species substantiate known aspects of male and female reproduction (Fig. 2c: serine-threonine kinases [41] or vitellogenin and other factors associated with oocyte generation, respectively). Some of these enriched genes have unknown functions and could comprise additional, novel factors associated with reproduction in *Oncopeltus*.

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 1658 Insecta genes (BUSCO v3, [42]). Virtually all BUSCO genes are present in the *Oncopeltus* OGS (98.9%, Fig. 3b, Additional file 1: Supplemental Note 6.1). Although some genes are fragmented, the assembly has a high level of BUSCO completeness (94.6%), independent of the 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, 43]. 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, 44]). 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 (Additional file 1: 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 (Additional file 1: 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 [43], *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* (Additional file 1: 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 (Additional file 1: 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 ([45], Additional file 1: 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* ([46], Additional file 1: 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 (Additional file 1: 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 (Additional file 1: Supplemental Note 5.1.f). Similarly, a tandem duplication of *wnt8* appears to be unique to *Oncopeltus* (Additional file 1: 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 ([47], Additional file 1: 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., [5, 48–51]). 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 protein families is associated with potential insecticide

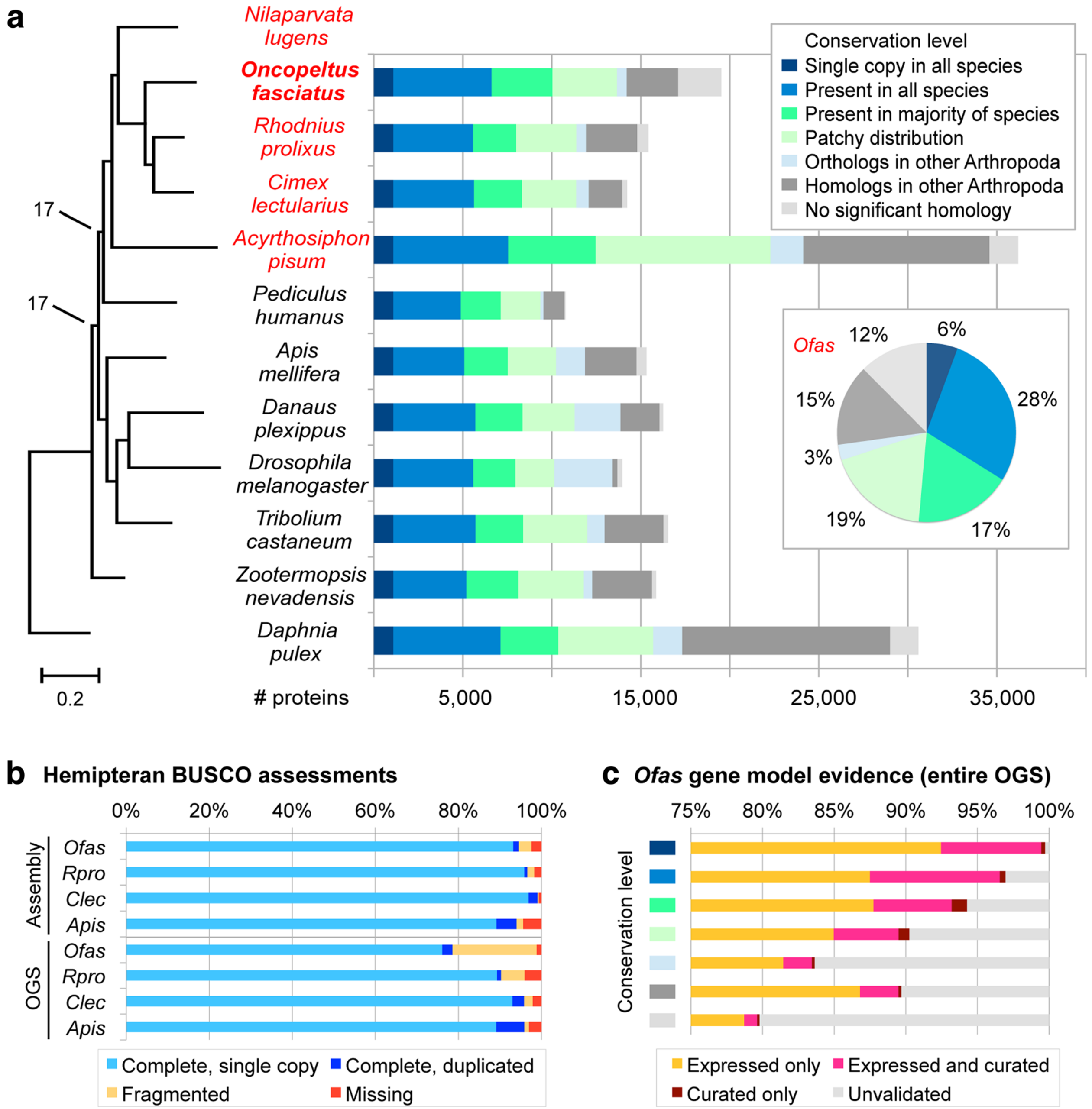


Fig. 3 Orthology comparisons and phylogenetic placement of *Oncopeltus fasciatus* among other Arthropoda. **a** Comparisons of protein-coding genes in 12 arthropod species, with the Hemiptera highlighted in red text. The bar chart shows the number of proteins per conservation level (see legend), based on OrthoDB orthology clustering analyses. To the left is a maximum likelihood phylogeny based on concatenation of 395 single-copy orthologs (all nodes have 100% support unless otherwise noted; branch length unit is substitutions per site). The inset pie chart shows the proportion of proteins per conservation level in *Oncopeltus* (*Ofas*). See also Additional file 1: Supplemental Note 6.1. **b** BUSCO-based analysis of *Oncopeltus* compared to other hemipterans for ortholog presence and copy number in both the assembly and OGS resources, using 4-letter species abbreviations (full names in **a**). **c** Proportion of *Oncopeltus* proteins that have expression and/or curation validation support per conservation level (same color legend as in **a**). Expression support is based on the life history stage data in Fig. 2b. Analyses are based on OGS v1.1

resistance, as specific P450s can confer resistance to specific chemicals (e.g., [52, 53]; Additional file 1: 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., [27, 33, 54–56], Additional file 1: 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 the “Methods” section, [57]). 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 Additional file 2: 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 (Fig. 4a). Given their very different rates of orthology assignment, we probed further into their pipeline predictions and the patterns of evolutionary diversification.

The number of homeodomain proteins identified by the pipeline displays a narrow normal distribution across species (Fig. 4b, mean ± standard deviation 97 ± 9), consistent with a highly conserved, slowly evolving protein family. Supporting this, many *Oncopeltus* homeodomain proteins that were manually curated also received a clear orthology assignment (Fig. 4c: pink), with only 4 exceptions (Fig. 4c: yellow). Only 1 case suggests a limitation of a pipeline that is not specifically tuned to hemipteran proteins (Goosecoid). Manual curation of partial or split models identified 11 further genes encoding homeodomains, bringing the actual tally in *Oncopeltus* to 96. Overall, we find the TF pipeline results to be a robust and reasonably comprehensive representation of these gene classes in *Oncopeltus*.

These analyses also uncovered a correction to the published *Oncopeltus* literature for the developmental patterning proteins encoded by the paralogs *engrailed* and *invected*. These genes arose from an ancient tandem duplication prior to the hexapod radiation. Their tail-to-tail orientation enables ongoing gene conversion [58], making orthology discrimination particularly challenging. For *Oncopeltus*, we find that the genes also occur in a tail-to-tail orientation and that *invected* retains a diagnostic alternative exon [58]. These new data reveal that the purported *Oncopeltus engrailed* ortholog in previous developmental studies (e.g., [54, 59–62]) is in fact *invected* (Additional file 1: Supplemental Note 5.1.a).

Independent expansions of C2H2 zinc fingers within the Hemiptera

Unlike homeodomain proteins, C2H2 zinc finger (C2H2-ZF) repertoires are prominent for their large family size and variability throughout the animal kingdom [63], and this is further supported by our current analysis in insects. With > 350 C2H2-ZFs, *Oncopeltus*, the pea aphid, termite, and some mosquito species have 1.5× more members than the insect median (Fig. 4b). This is nearly half of all *Oncopeltus* TFs. While the expansion in mosquitoes could have a single origin within the Culicinae, the distribution in the Hemiptera, where *Cimex* has only 227 C2H2-ZFs, suggests that independent expansions occurred in *Oncopeltus* and the pea aphid. Prior to the sequencing of other hemipteran genomes, the pea aphid’s large C2H2-ZF repertoire was attributed to the expansion of a novel subfamily, APEZ, also referred to as zinc finger 271-like [43].

In fact, manual curation in *Oncopeltus* confirms the presence of a subfamily with similar characteristics to APEZ (Fig. 4c: yellow fraction). In *Oncopeltus*, we find >115 proteins of the ZF271 class that are characterized by numerous tandem repeats of the C2H2-ZF domain and its penta-peptide linker, with 3–45 repeats per protein.

Intriguingly, we find evidence for ongoing evolutionary diversification of this subfamily. A number of *Oncopeltus* ZF271-like genes occur in tandem clusters of 4–8 genes—suggesting recent duplication events. Yet, clustered genes differ in gene structure (number and size of exons), and we identified a number of probable ZF271-like pseudogenes whose open reading frames have become disrupted—consistent with high turnover. *Oncopeltus* ZF271-like proteins also differ in the sequence and length of the zinc finger domains among themselves and compared to aphid proteins (WebLogo analysis, [64]), similar to zinc finger array shuffling seen in humans [65]. Furthermore, whole-protein phylogenetic analysis supports independent, rapid expansions in the pea aphid and *Oncopeltus* (Fig. 4d).

Clustered zinc finger gene expansion has long been recognized in mammals, with evidence for strong positive selection to increase both the number and diversity of zinc finger domains per protein as well as the total number of proteins [66]. This was initially found to reflect an arms race dynamic of co-evolution between selfish transposable elements and the C2H2-ZF proteins

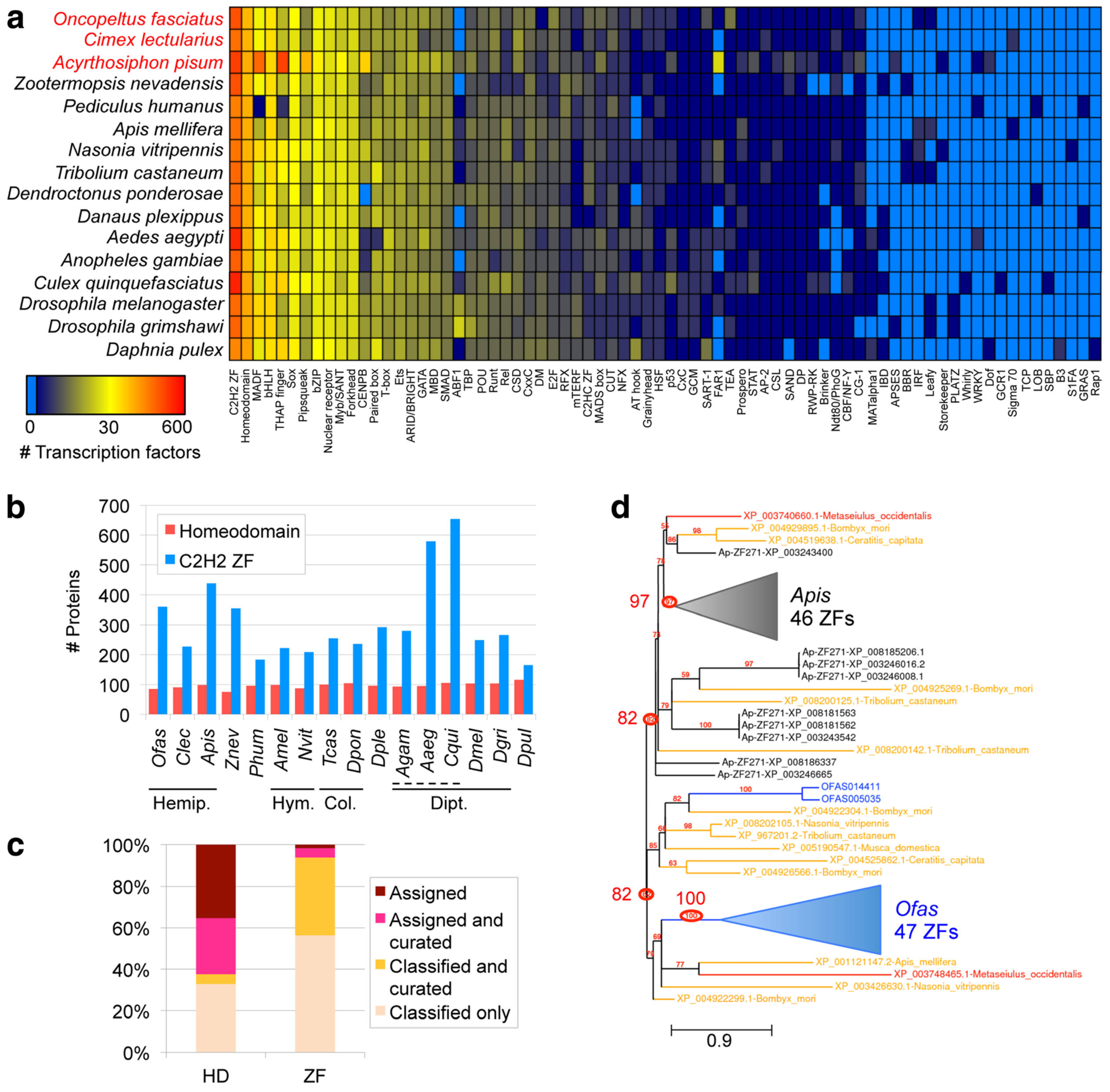


Fig. 4 Distribution of transcription factor (TF) families across insect genomes. **a** Heatmap depicting the abundance of 74 TF families across 16 insect genomes (Hemiptera highlighted in red text), with *Daphnia* as an outgroup, based on the presence of predicted DNA binding domains (see the "Methods" section). The color key has a log (base 2) scale (light blue means the TF family is completely absent). Values are in Additional file 2: Table S6.3. **b** Bar graph showing the number of proteins of each of the 2 most abundant TF families, homeodomains and C2H2 zinc fingers (ZFs), per species using 4-letter abbreviations (full names in **a**). Solid lines demarcate insect orders: Hemiptera (Hemipt.), Hymenoptera (Hym.), Coleoptera (Col.), and Diptera (Dipt.). The dashed line demarcates the dipteran family Culicidae (mosquitoes). **c** Proportions of *Oncopeltus* homeodomain (HD) and C2H2 zinc finger proteins with orthology assignment (predicted DNA binding specificity) and/or manual curation. "Classified" refers to the automated classification of a protein to a TF family, but without a specific orthology assignment. **d** Maximum likelihood phylogeny of representative subsets of the zinc finger 271-like family in *Oncopeltus* (49 proteins, blue text) and the pea aphid (55 proteins, black text), with chelicerate (red text) and holometabolan (yellow text) outgroups (16 proteins, 7 species), based on the *Oncopeltus* OGS and GenBank protein accessions. Gaps were removed during sequence alignment curation; all nodes have ≥ 50% support; branch length unit is substitutions per site [157]. Key nodes are circled for the clades containing all aphid or all *Oncopeltus* proteins (82% support each), and each "core" clade comprised exclusively of proteins from each species (97% and 100%, respectively; triangles shown to scale for branch length and number of clade members). Branch length unit is substitutions per site. Analyses are based on OGS v1.1

that would repress them [67]. In vertebrates, these C2H2-ZF proteins bind to the promoters of transposable elements via their zinc finger arrays and use their Krüppel-associated box (KRAB) domain to bind the chromatin-remodeling co-repressor KAP-1, which in turn recruits methyltransferases and deacetylases that silence the targeted promoter [68].

Insects do not have a direct ortholog of vertebrate KAP-1 (Additional file 1: Supplemental Note 5.4.d), and neither the aphid nor *Oncopeltus* ZF271-like subfamilies possess a KRAB domain or any other domain besides the zinc finger arrays. However, close molecular outgroups to this ZF271-like subfamily include the developmental repressor Krüppel [69] and the insulator protein CTCF [70] (data not shown). Like these outgroups, the *Oncopeltus* ZF271-like genes are strongly expressed: 98% have expression support, with 86% expressed in at least three different life history stages (Fig. 2b). Thus, the insect ZF271-like proteins may also play prominent roles in repressive DNA binding. Indeed, we find evidence for a functional methylation system in *Oncopeltus* (Additional file 1: Supplemental Note 5.4.c), like the pea aphid, which would provide a means of gene silencing by chromatin remodeling, albeit via mediators other than KAP-1.

However, an arms race model need not be the selective pressure that favors insect ZF271-like family expansions. Recent analyses in vertebrates identified sophisticated, additional regulatory potential by C2H2-ZF proteins, building upon original transposable element binding for new, lineage-specific and even positive gene regulation roles [65, 71, 72]. Moreover, although *Cimex* has half as many long terminal repeat (LTR) repetitive elements as *Oncopeltus* and the pea aphid, overall, we do not find a correlation between relative or absolute repetitive content and ZF271-like family expansion within the Hemiptera (see the next section).

Proportional repeat content across hemipterans

With the aim of reducing assembly fragmentation and to obtain a better picture of repeat content, we performed low-coverage, long-read PacBio sequencing in *Oncopeltus* (Additional file 1: Supplemental Note 2.3). Using PacBio reads in a gap-filling assay on the Illumina assembly raised the total detected repetitive content from 25 to 32%, while repeat estimations based on simultaneous assessment of Illumina and PacBio reads nearly doubled this value to 58%. As expected, the capacity to identify repeats is strongly dependent on the assembly quality and sequencing technology, with the *Oncopeltus* repetitive content underrepresented in the current (Illumina-only) assembly. Furthermore, as increasing genome size compounds the challenge of assembling repeats, the repeat content of the current assembly is lower than in species with smaller genome sizes (Fig. 5a, with the sole exception of the honey bee), and we therefore used our gap-filled dataset as a more accurate basis for further comparisons.

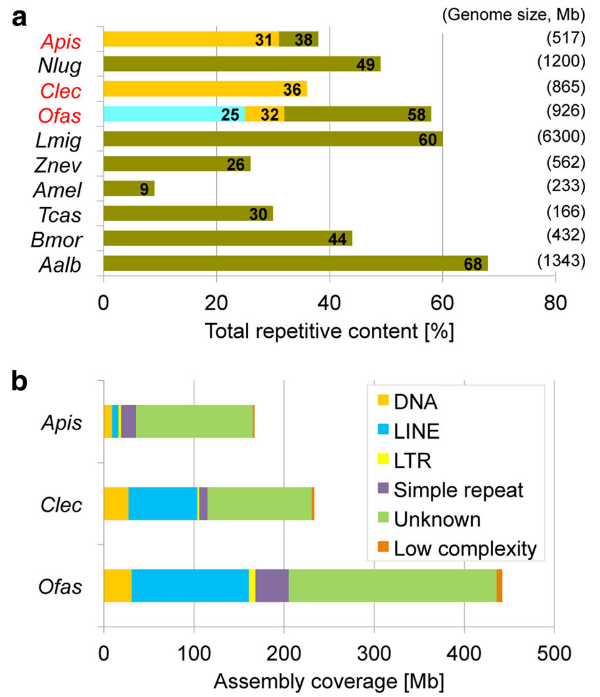


Fig. 5 Comparison of repeat content estimations. **a** Comparison of total repetitive content among insect genomes. The three values for *Oncopeltus* are shown (in ascending order: original Illumina assembly, gap-filled assembly, Illumina-PacBio hybrid estimate). Values for the three hemipterans labeled in red text are from RepeatModeler (gold bars for the pea aphid and bed bug; blue and gold bars for *Oncopeltus*). All other values are from the respective genome papers, including a second value corresponding to the published repeat content for the first version of the aphid genome [6, 10, 110, 158–163]. Species abbreviations as in Fig. 4 and additionally *Nlug*, *Nilaparvata lugens*; *Lmig*, *Locusta migratoria*; *Bmor*, *Bombyx mori*; *Aalb*, *Aedes albopictus*. **b** Comparison of repetitive element categories between the three hemipteran genomes, based on results from RepeatModeler. Here, we present assembly coverage as actual sequence length (Mb) to emphasize the greater repeat content in *Oncopeltus* (based on the gap-filled assembly, see also Additional file 1: Supplemental Note 2.3)

To support direct comparisons among hemipterans, we also performed our RepeatModeler analysis on the bed bug and pea aphid assemblies. Repeats comprised 36% and 31% of the respective assemblies, similar to the gap-filled value of 32% in *Oncopeltus*. Nevertheless, given the smaller sizes of these species' assemblies—651 Mb in the bed bug and 542 Mb in the pea aphid—the absolute repeat content is much higher in *Oncopeltus* (Fig. 5b). Excluding unknown repeats, the most abundant transposable elements in *Oncopeltus* are LINE retrotransposons, covering 10% of the assembly (Additional file 2: Table S2.5). This is also the case in the bed bug (12%), while in the pea aphid DNA transposons with terminal inverted repeats

(TIRs) are the most abundant (2% of the assembly identified here and 4% reported from manual curation in the pea aphid genome paper, [6]). Across species, the remaining repeat categories appear to grow proportionally with assembly size, except for simple repeats, which were the category with the largest relative increase in size after gap filling in *Oncopeltus* (Additional file 1: Supplemental Note 2.3). However, given the mix of data types (Illumina [12] and Sanger [6]), these patterns should be treated as hypotheses for future testing.

Lineage- and genome size-related trends in insect gene structure

Both our manual curation work and BUSCO analyses highlighted the fact that *Oncopeltus* genes are often comprised of many, small exons. We thus undertook a comparative analysis to determine whether this is a general feature to be considered for structural annotation of hemipteran genomes. We find that both lineage and genome size can serve as predictors of gene structure.

Firstly, we created a high-quality dataset of 30 functionally diverse, large genes whose manual curation could reasonably ensure complete gene models across 7 species from 4 insect orders (Fig. 6a, Additional file 1: Supplemental Note 6.3). Most species encode the same total number of amino acids for these conserved proteins, with the thrips *Frankliniella occidentalis* and the fruit fly being notable exceptions with larger proteins (Fig. 6a: blue plot line). However, the means of encoding this information differs between lineages, with hemipteroid orthologs comprised of twice as many exons as their holometabolous counterparts (Fig. 6a: orange plot line). Thus, there is an inverse correlation between exon number and exon size (Fig. 6a: orange vs. red plot lines). This analysis corroborates and extends previous probabilistic estimates of intron density, where the pea aphid as a sole hemipteran representative had the highest intron density of 10 insect species [73].

To test these trends, we next expanded our analysis to all manually curated exons in two species from each of three orders (Hemiptera, Coleoptera, Diptera). Here, we expect that curated exon sizes are accurate, without the need to assume that the entire gene models are complete. This large dataset corroborates our original findings, with bugs having small exons while both the median and Q3 quartile reflect larger exon sizes in beetles and flies (Fig. 6b). Notably, the median and median absolute deviation are highly similar between species pairs within the Hemiptera and Coleoptera. Meanwhile, the exon metrics within the Diptera support large protein sizes as a drosophilid-specific, rather than dipteran-wide, feature.

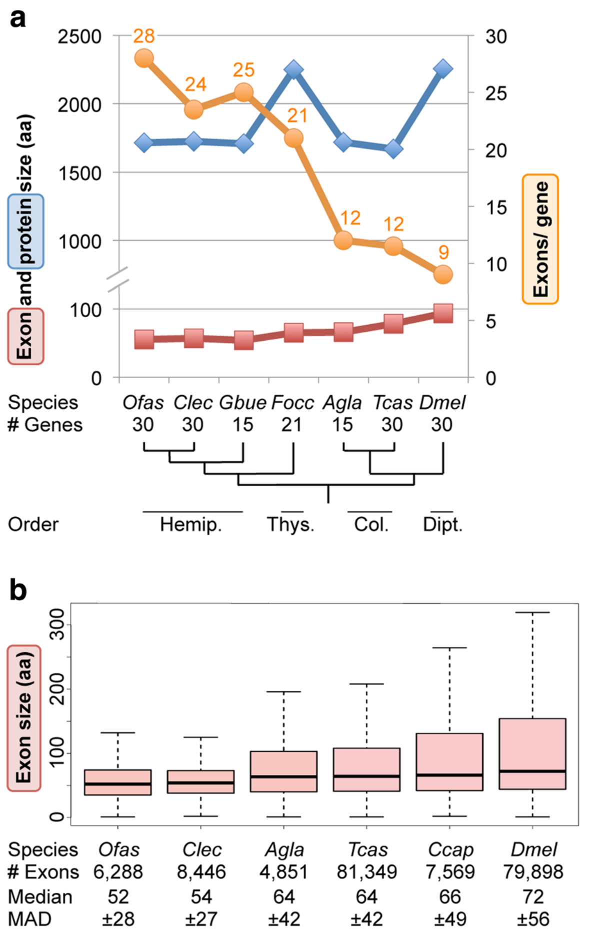


Fig. 6 Trends in gene structure show hemipteroid-specific tendencies. **a** Median values per species for protein size, exon size, and exon number for a curated set of highly conserved genes encoding large proteins of diverse functional classes (see also Additional file 1: Supplemental Note 6.3). Sample sizes are indicated, with 11 genes for which orthologs were evaluated in all species. Where it was not possible to analyze all 30 genes for a given species, equal sampling was done across the range of protein sizes of the complete dataset, based on the *Cimex* ortholog sizes (1:1:1 sampling from big-to-medium-to-small subcategories of 10 genes each). **b** Box plot representations of coding sequence exon size (aa) for 2 species from each of 3 insect orders, based on datasets of unique coding sequence exons (1 isoform per gene) and excluding terminal exons < 10 aa (as most of those exons may rather be UTRs or a small placeholder N-terminal exon based on automated Maker model predictions). Only manually curated gene models were considered for the i5K species, including *Oncopeltus*; the entire OGS was used for *Tribolium* and *Drosophila*. For clarity, outliers are omitted; whiskers represent 1.5× the value of the Q3 (upper) or Q2 (lower) quartile range. MAD, median absolute deviation. Species are represented by their 4-letter abbreviations, with their ordinal relationships given below the phylogeny in **a**: Hemip., Hemiptera; Thys., Thysanoptera; Col., Coleoptera; Dipt., Diptera. Species abbreviations as in Figs. 2 and 4 and additionally *Gbue*, *Gerris buenoi* [164]; *Agla*, *Anoplophora glabripennis* [30]; *Ccap*, *Ceratitis capitata* [165]

Does the high exon count in the Hemiptera reflect an ancient, conserved increase at the base of this lineage or ongoing remodeling of gene structure with high turnover? To assess the exact nature of evolutionary changes, we annotated intron positions within multiple sequence alignments of selected proteins and plotted gains and losses onto the phylogeny, providing a total sample of 165 evolutionary changes at 148 discrete splice sites (Fig. 7, see also Additional file 1: Supplemental Note 6.3 for gene selection and method). These data reveal several major correlates with intron gain or loss. The bases of both the hemipteroid and hemipteran radiations show the largest gains, while most losses occur in the dipteran lineage (Fig. 7: orange and purple shading, respectively). Furthermore, we find progressive gains across hemipteroid nodes, and it is only in this lineage that we additionally find species-specific splice changes for the highly conserved *epimerase* gene (Fig. 7: orange outline). Thus, we find evidence for both ancient intron gain and ongoing gene structure remodeling in this lineage.

Surprisingly, both *hemocytin* and *epimerase*—our exemplar genes with many (up to 74) and few exons (3–8 per species), respectively—show independent losses of the same splice sites in *Drosophila* and *Tribolium*. One feature these species share is a genome size 2.4–6.0× smaller than in the other species examined here (Fig. 7: red shading). Pairwise comparisons within orders also support this trend, as the beetle and fly species with larger genomes exhibit species-specific gains compared to intron loss in their sister taxa (Fig. 7: red outlines). Thus, genome size seems to positively correlate with intron number. However, lineage is a stronger predictor of gene structure: the coleopteran and dipteran species pairs have highly similar exon size metrics despite differences in genome size (Fig. 6b). A global computational analysis over longer evolutionary distances also supports a link between genome size and intron number in arthropods, although chelicerates and insects may experience different rates of evolutionary change in these features [74]. It will be interesting to see if the correlation with genome size is borne out in other invertebrate taxa.

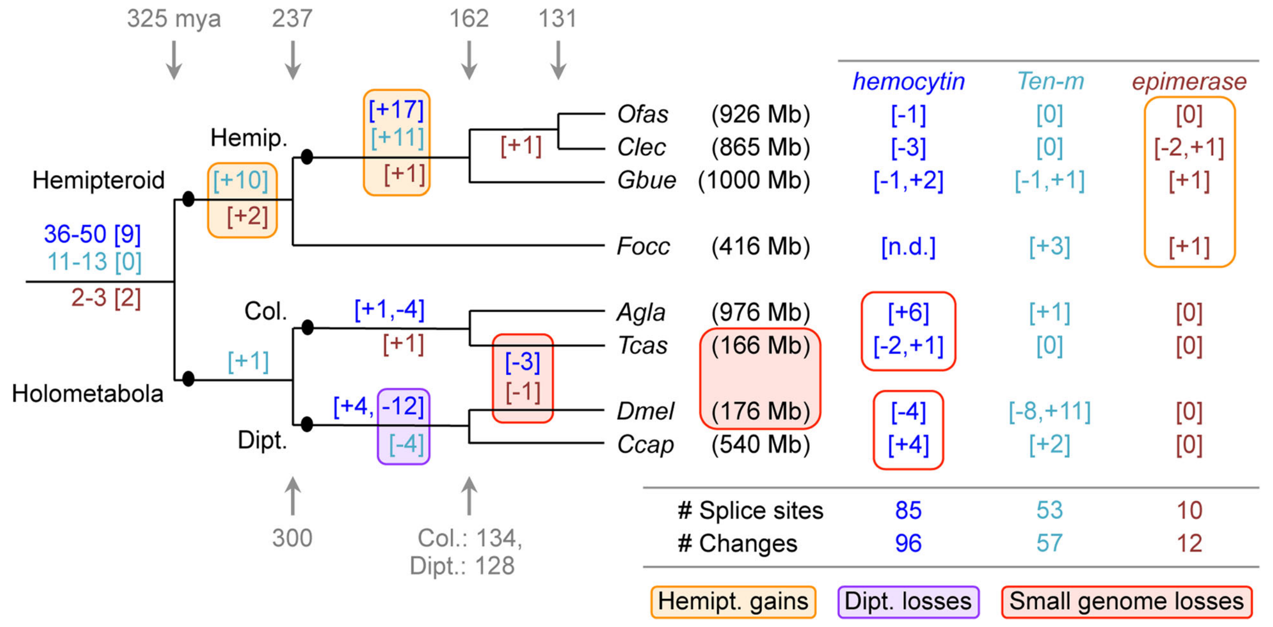


Fig. 7 Splice site evolution correlates with both lineage and genome size. Splice site changes are shown for *hemocytin* (blue text), *Tenascin major* (*Ten-m*, turquoise text), and *UDP-galactose 4′-epimerase* (brown text), mapped onto a species tree of eight insects. Patterns of splice site evolution were inferred based on the most parsimonious changes that could generate the given pattern within a protein sequence alignment of all orthologs (see also Additional file 1: Supplemental Note 6.3 for methodology and data sources). If inferred gains or losses were equally parsimonious, we remained agnostic and present a range for the ancestral number of splice sites present at the base of the tree, where the bracketed number indicates how many ancestral positions are still retained in all species. Along each lineage, subsequent changes are indicated in brackets, with the sign indicating gains (+) or losses (–). Values shown to the right are species-specific changes. The values shown between the *D. melanogaster* and *T. castaneum* lineages denote changes that have occurred independently in both species. Colored boxes highlight the largest sources of change, as indicated in the legend. Species are represented by their four-letter abbreviations (as in Fig. 6), and estimated genome sizes are indicated parenthetically (measured size [12, 30, 162, 165, 166]; draft assembly size: GenBank Genome IDs 14741 and 17730). Divergence times are shown in gray and given in millions of years [3]. Abbreviations as in Figs. 4 and 6, and also: Hemipt., hemipteroid assemblage (including *F. occidentalis*); n.d., no data

The selective pressures and mechanisms of intron gain in the Hemiptera will be a challenge to uncover. While median exon size (Fig. 6b) could reflect species-specific nucleosome sizes [75, 76], this does not explain why only the Hemiptera seldom exceed this (Fig. 6b: Q3 quartile). Given the gaps in draft genome assemblies, we remain cautious about interpreting (large) intron lengths but note that many hemipteran introns are too small to have harbored a functional transposase gene (e.g., median intron size of 429 bp, $n = 69$ introns in *hemocytin* in *Cimex*). Such small introns could be consistent with the proliferation of non-autonomous short interspersed nuclear elements (SINEs). However, characterization of such highly divergent non-coding elements would require curated SINE libraries for insects, comparable to those generated for vertebrates and plants [75, 76]. Meanwhile, it appears that hemipteran open reading frames ≥ 160 bp are generally prevented by numerous in-frame stop codons just after the donor splice site. Most stop codons are encoded by the triplet TAA in both *Oncopeltus* and *Cimex* (data not shown), although these species' genomes are not particularly AT rich (Table 1).

Even if introns are small, having gene loci comprised of numerous introns and exons adds to the cost of gene expression in terms of both transcription duration and mRNA processing. One could argue that a gene like *hemocytin*, which encodes a clotting agent, would require rapid expression in the case of wounding—a common occurrence in adult *Cimex* females due to the traumatic insemination method of reproduction [12]. Thus, as our molecular understanding of comparative insect and particularly hemipteran biology deepens, we will need to increasingly consider how life history traits are manifest in genomic signatures at the structural level (e.g., Figs. 5, 6, and 7), as well as in terms of protein repertoires (Figs. 3 and 4).

Expansion after a novel lateral gene transfer event in phytophagous bugs

In addition to the need for cuticle repair, traumatic insemination may be responsible for the numerous lateral gene transfer (LGT) events predicted in the bed bug [12]. In contrast, the same pipeline analyses [77] followed by manual curation predicted very few LGTs in *Oncopeltus*, which lacks this unusual mating behavior. Here, we have identified 11 strong LGT candidates, and we confirmed the incorporation of bacterial DNA into the milkweed bug genome for all 5 candidates chosen for empirical testing (Additional file 2: Table S2.4). Curiously, we find several LGTs potentially involved in bacterial or plant cell wall metabolism that were acquired from different bacterial sources at different times during hemipteran lineage evolution, including 2 distinct LGTs that are unique to *Oncopeltus* and implicated in the synthesis of peptidoglycan, a bacterial cell wall constituent (Additional file 1: Supplemental Note 2.2).

Conversely, two other validated LGT candidates are implicated in cell wall degradation. We find two strongly expressed, paralogous copies in *Oncopeltus* of a probable bacterial-origin gene encoding an endo-1,4-beta-mannosidase enzyme (MAN4, EC 3.2.1.78). Inspection of genome assemblies and protein accessions reveals that this LGT event occurred after the infraorder Pentatomomorpha, including the stink bug *Halyomorpha halys*, diverged from other hemipterans, including the bed bug (Fig. 8a). Independent duplications then led to the two copies in *Oncopeltus* and an astonishing nine tandem copies in *Halyomorpha* (Fig. 8b, Additional file 1: Figure S2.6). Since the original LGT event, the *mannosidase* genes have gained introns that are unique to each species and to subsets of paralogs (Fig. 8c). Thus, the "domestication" [78] of *mannosidase* homologs as multi-exonic genes further illustrates the hemipteran penchant for intron introduction and maintenance of small exons. The retention and subsequent expansion of these genes imply their positive selection, consistent with the phytophagous diet of these species. It is tempting to speculate that copy number proliferation in the stink bug correlates with the breadth of its diet, as this agricultural pest feeds on a number of different tissues in a range of host plants [79].

Cuticle development, structure, and warning pigmentation

The distinctive cuticle of *Oncopeltus* is produced through the combined action of genes that encode structural and pigmentation proteins, and the gene products that regulate their secretion at each life stage. Furthermore, the milkweed bug has been a powerful model for endocrine studies of hemimetabolous molting and metamorphosis since the 1960s [22, 80–83]. Therefore, we next focused on the presence and function of genes involved in these processes.

Molting is triggered by the release of ecdysteroids, steroid hormones that are synthesized from cholesterol by cytochrome P450 enzymes of the Halloween family [84], and we were able to identify these in the *Oncopeltus* genome (Additional file 1: Supplemental Notes 5.2.b, 5.3.b). From the ecdysone response cascade defined in *Drosophila* [85], we identified *Oncopeltus* orthologs of both early- and late-acting factors, including ecdysteroid hormones and their receptors. It will be interesting to see if the same regulatory relationships are conserved in the context of hemimetabolous molting in *Oncopeltus*. For example, *E75A* is required for reactivation of ecdysteroid production during the molt cycle in *Drosophila* larvae [86] and likely operates similarly in *Oncopeltus*, since *Of-E75A* RNAi

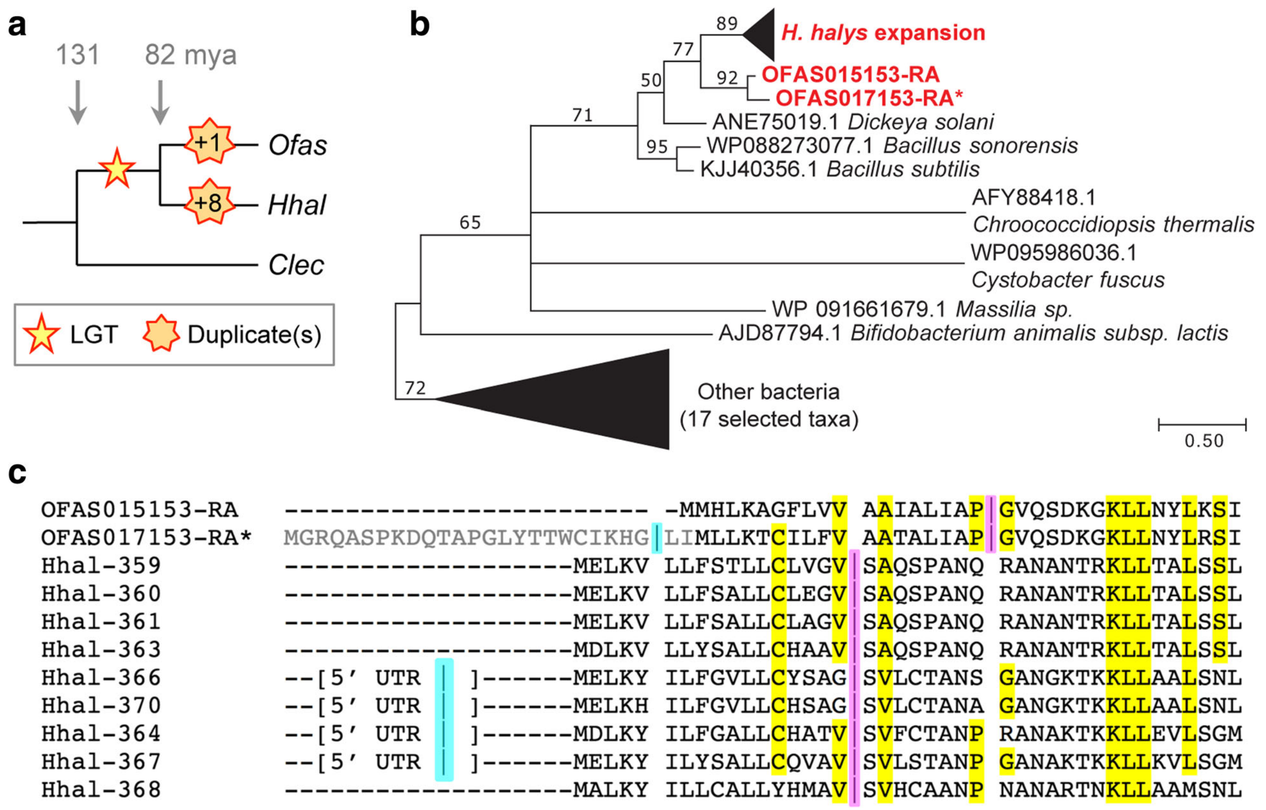
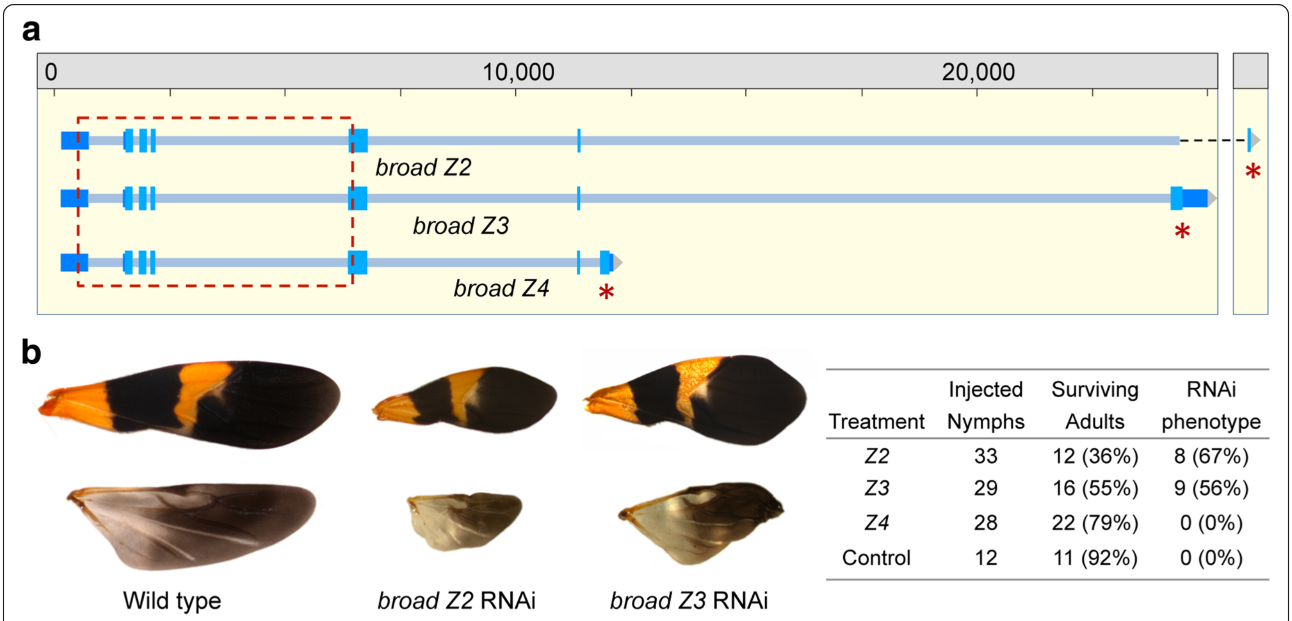

a
131
82 mya
+1
Ofas
+8
Hhal
Clec
LGT
Duplicate(s)
b
89
H. halys expansion
77
92
OFAS015153-RA
OFAS017153-RA*
50
ANE75019.1 Dickeya solani
71
95
WP088273077.1 Bacillus sonorensis
KJJ40356.1 Bacillus subtilis
AFY88418.1
Chroococcidiopsis thermalis
65
WP095986036.1
Cystobacter fuscus
WP 091661679.1 Massilia sp.
AJD87794.1 Bifidobacterium animalis subsp. lactis
72
Other bacteria
(17 selected taxa)
0.50
c
OFAS015153-RA ------------------------ -MMHLKAGFLVV AAIALIAP|GVQSDKGKLLNYLKSI
OFAS017153-RA* MGRQASPKDQTAPGLYTTWCIKHG|LIMLLKTCILFV AATALIAP|GVQSDKGKLLNYLRSI
Hhal-359 --------------------MELKV LLFSTLLCLVGV|SAQSPANQ RANANTRKLLTALSSL
Hhal-360 --------------------MELKV LLFSALLCLEGV|SAQSPANQ RANANTRKLLTALSSL
Hhal-361 --------------------MELKV LLFSALLCLAGV|SAQSPANQ RANANTRKLLTALSSL
Hhal-363 --------------------MDLKV LLYSALLCHAAV|SAQSPANQ RANANTRKLLTALSSL
Hhal-366 --[5' UTR |]------MELKY ILFGVLLCYSAG|SVLCTANS GANGKTKKLLAALSNL
Hhal-370 --[5' UTR |]------MELKH ILFGVLLCHSAG|SVLCTANA GANGKTKKLLAALSNL
Hhal-364 --[5' UTR |]------MDLKY ILFGALLCHATV|SVFCTANP RANAKTKKLLEVLSGM
Hhal-367 --[5' UTR |]------MELKY ILYSALLCQVAV|SVLSTANP GANAKTKKLLKVLSGM
Hhal-368 --------------------MALKY ILLCALLYHMAV|SVHCAANP NANARTNKLLAAMSNL

Fig. 8 Lateral gene transfer introduction and subsequent evolution within the Hemiptera for mannosidase-encoding genes. **a** Species tree summary of evolutionary events. Stars represent the original LGT introduction and subsequent copy number gains (see legend). **b** Maximum likelihood phylogeny of mannosidase proteins, including bacterial sequences identified among the best GenBank blastp hits for *Oncopeltus* and *Halyomorpha* (accession numbers as indicated, and for "Other bacteria" are ACB22214.1, AEE17431.1, AEI12929.1, AEO43249.1, AFN74531.1, CDM56239.1, CUA67033.1, KOE98396.1, KPI24888.1, OAN41395.1, ODP26899.1, ODS11151.1, OON18663.1, PBD05534.1, SIR54690.1, WP096035621.1, YP001327394.1). All nodes have ≥ 50% support from 500 bootstrap replicates [167]. Triangles are shown to scale for branch length and number of clade members; branch length unit is substitutions per site. See also Additional file 1: Figure S2.6. **c** Manually curated protein sequence alignment for the N-terminal region only. Splice sites ("|" symbol) are shown, where one position is ancestral and present in all paralogs of a given species (magenta) and one position occurs in a subset of paralogs and is presumed to be younger (cyan, within the 5′ UTR in *Halyomorpha*). Residues highlighted in yellow are conserved between the two hemipteran species. The *Oncopeltus* paralog represented in the OGS as OFAS017153-RA is marked with an asterisk to indicate that this version of the gene model is incomplete and lacks the initial exon (gray text in the alignment). For clarity, only the final three digits of the *Halyomorpha* GenBank accessions are shown (full accessions: XP_014289XXX)

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 [49]. 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 [49, 87], and the JH-response gene *Krüppel homolog 1* (*Kr-h1*) [49, 88, 89]. 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* [90, 91]. In *Oncopeltus*, *broad* expression directs progression through the nymphal stages [92], 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* (Additional file 1: 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: Additional file 1:



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*. **a** Genomic organization of the cuticle identity gene *broad*. The regions used as a template to generate isoform-specific dsRNA are indicated (red asterisks: the final, unique exons of each isoform). Previous RNAi studies targeted sequence within exons 1–5 that is shared among all isoforms (dashed red box, [92]). **b** Knockdown of the *Oncopeltus Z2* or *Z3 broad* isoforms at the onset of the penultimate instar resulted in altered nymphal survival and morphogenesis that was reflected in the size and proportion of the fore and hind wings at the adult stage (upper and lower images, respectively, shown to the same scale for all wings). We did not detect any effect on the wing phenotype when targeting the *Z4*-specific exon, demonstrating the specificity of the zinc finger coding region targeted by RNAi. Experimental statistics are provided in the figure inset, including for the buffer-injected negative control

Table S5.12), we found a slight expansion in the *Oncopeltus* CPF family (Additional file 1: Figure 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 [93].

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 and 9, [20, 21]). For black coloration, the melanin synthesis pathway known from holometabolous insects (e.g., [94, 95]) is conserved at the sequence (Additional file 1: Figure S5.15) and functional [96, 97] level in *Oncopeltus*, supporting conservation in hemimetabolous lineages as well. In contrast, production of the primary warning coloration, pteridine red erythropterin [98], 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 [99]. Thus far in *Oncopeltus*, we could identify orthologs of *punch*, which encodes a GTP cyclohydrolase [100], and *sepia*, which is required for the synthesis of the red eye pigment drosopterin [101]. The bright red color of *Oncopeltus* eggs may in part reflect chemical protection transmitted parentally [102]. 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).

Chemoreception and metabolism in relation to feeding biology

Aposematic pigmentation advertises the fact that toxins in the milkweed diet are incorporated into the insects themselves, a metabolic feat that was independently acquired in *Oncopeltus* and the monarch butterfly (*Danaus plexippus*), which shares this food source and body coloration [36, 103]. Moreover, given the fundamental differences between phytophagous, mucivorous, and hematophagous diets, we investigated to what extent differences in feeding ecology across hemipterans are represented in their chemoreceptor and metabolic enzyme repertoires.

Insects must smell and taste their environment to locate and identify food, mates, oviposition sites, and other essential cues. Perception of the enormous diversity of environmental chemicals is primarily mediated by the odorant (OR), gustatory (GR), and ionotropic (IR) families of chemoreceptors, which each encode tens to hundreds of distinct proteins [104–107]. Chemoreceptor family size appears to correlate with feeding ecology. *Oncopeltus* retains a moderate complement of each family, while species with derived fluid nutrition diets (sap or blood) have relatively depauperate OR and GR families (Table 2, Additional file 1: Supplemental Note

Table 2 Numbers of chemoreceptor genes/proteins per family in selected insect species. In some cases, the number of proteins is higher than the number of genes due to an unusual form of alternative splicing, which is particularly notable for the *Oncopeltus* GRs. Data are shown for four Hemiptera as well as *Drosophila melanogaster*, the body louse *Pediculus humanus*, and the termite *Zootermopsis nevadensis* [11, 12, 104, 108–110, 168]

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

[1] Hemiptera: Heteroptera

[2] Independent acquisitions of hematophagy [16]

[3] Hemiptera, phloem feeding

5.3.f, Additional file 3). In detail, a few conserved orthologs such as the OrCo protein and a fructose receptor are found across species, but other subfamilies are lineage specific. *Oncopeltus* and *Acyrthosiphon* retain a set of sugar receptors that was lost independently in the blood-feeding bugs (*Rhodnius* [11], *Cimex* [12]) and body louse (*Pediculus* [108]). Conversely, *Oncopeltus* and *Cimex* retain a set of candidate carbon dioxide receptors, a gene lineage lost from *Rhodnius*, *Acyrthosiphon*, and *Pediculus* [11, 12, 109], but which is similar to a GR subfamily expansion in the more distantly related hemimetabolous termite (Isoptera [110]). Comparable numbers of IRs occur across the Heteroptera. In addition to a conserved set of orthologs primarily involved in sensing temperature and certain acids and amines, *Oncopeltus* has a minor expansion of IRs distantly related to those involved in taste in *Drosophila*. The major expansions in each insect lineage are the candidate “bitter” GRs ([111], Additional file 1: Supplemental Note 5.3.f and Figure S5.19). In summary, *Oncopeltus* exhibits moderate expansion of specific subfamilies likely to be involved in host plant recognition, consistent with it being a preferentially specialist feeder with a potentially patchy food source [21, 112].

As host plant recognition is only the first step, we further explored whether novel features of the *Oncopeltus* gene set are directly associated with its diet. We therefore used the CycADS annotation pipeline [113] to reconstruct the *Oncopeltus* metabolic network. The resulting BioCyc metabolism database for *Oncopeltus* (OncfaCyc) was then compared with those for 26 other insect species ([114], http://arthropodacyc.cycadsys.org/), including 3 other hemipterans: the pea aphid, the green peach aphid, and the kissing bug (Tables 3 and 4). For a global metabolism analysis, we detected the presence of 1085 Enzyme Commission (EC) annotated reactions with at least 1 protein in the *Oncopeltus* genome (Additional file 1: Supplemental Note 6.4, Additional file 2: Table S6.10). Among these, 10 enzyme classes (represented by 17 genes) are unique and 17 are missing when compared to the other insects (Table 4, Additional file 2: Table S6.11).

We then specifically compared amino acid metabolism in the four hemipterans representing the three different diets. Eight enzymes are present only in *Oncopeltus* (Table 4), including the arginase that degrades arginine (Arg) into urea and ornithine, a precursor of proline (Pro). Given this difference, we extended our analysis to assess species’ repertoires for the entire urea cycle (Fig. 10a, Additional file 2: Table S6.13). *Oncopeltus* and six other species can degrade Arg but cannot synthesize it (Fig. 10b). Only the other three hemipterans can neither synthesize nor degrade Arg via this cycle (Fig. 10c), while most species have an almost complete cycle (Fig. 10d). This suggests that the ability to synthesize Arg was lost at the base of the Hemiptera, with subsequent, independent loss of Arg degradation capacity in the aphid and *Rhodnius* lineages. Retention of Arg

Table 3 Hemipteran ArthropodaCyc database summaries. Overview statistics for the newly created database for *Oncopeltus fasciatus* (*Ofas*) in comparison with public databases for *Rhodnius prolixus* (*Rpro*), *Acyrthosiphon pisum* (*Apis*), and *Myzus persicae* (*Mper*) available from [114]. Based on OGS v1.1

Species ID	*Ofas*	*Rpro*	*Apis*	*Mper*	*Mper*
Gene set ID	OGS v1.1	RproC1.1 (Built on RproC1 assembly)	OGS v2.1b (Built on Acyr_2.0 assembly)	Clone G006 v1.0	Clone O v1.0
CycADS database ID	OncfaCyc	RhoprCyc	AcypiCyc v2.1b	Myzpe_G006 Cyc	Myzpe_O Cyc
Total mRNA[1]	19,673	15,437	36,195	24,814	24,770
Pathways	294	312	307	319	306
Enzymatic reactions	2192	2366	2339	2384	2354
Polypeptides	19,820	15,471	36,228	24,849	24,805
Enzymes	3050	2660	5087	4646	4453
Compounds	1506	1665	1637	1603	1655

[1] In the BioCyc databases, all splice variants are counted in the summary tables for genes

Table 4 Hemipteran ArthropodaCyc annotations of metabolic genes. Taxonomic abbreviations are as in Table 3

	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

[1] "EC" refers to the number of proteins, as represented by their unique numerical designations within the Enzyme Commission (EC) classification system for enzymes and their catalytic reactions

[2] In comparison with all other insects from ArthropodaCyc

[3] in comparison among the four hemipterans

[4] Includes three EC categories added in OGS v1.2 (see also Additional file 2: Table S6.11)

degradation in *Oncopeltus* might be linked to the milkweed seed food source, as most seeds are very rich in Arg [115], and Arg is indeed among the metabolites detected in *Oncopeltus* [116]. However, the monarch butterfly is one of only a handful of species that retains the complete Arg pathway (Fig. 10d: blue text). Despite a shared food source, these species may therefore differ in their overall Arg requirements or—in light of a possible group benefit of *Oncopeltus* aggregation during feeding ([21]; e.g., Fig. 1b)—in their efficiency of Arg uptake.

Other enzymes are also present only in the milkweed bug compared to the other examined hemipterans (Additional file 2: Table S6.12). Like other insects [114], *Oncopeltus* retains the ability to degrade tyrosine (Tyr). This pathway was uniquely lost in the aphids, where this essential amino acid is jointly synthesized—and consumed—by the aphid host and its endosymbiotic bacteria [6, 7, 17, 117]. Conversely, a gain specific to the milkweed bug lineage was the duplication of the Na+/K + ATPase alpha subunits whose amino acid substitutions confer resistance to milkweed cardenolides [36, 118]. In the *Oncopeltus* genome, we find support for the recent nature of these duplications: the genes encoding subunits ATPα1B and ATPα1C occur as a tandem duplication, notably on a scaffold that also harbors one of the clustered ZF271-like gene expansions (see above).

Conclusions

The integrated genomic and transcriptomic resources presented here for the milkweed bug *Oncopeltus fasciatus* (Figs. 2 and 5) underpin a number of insights into evolutionary and developmental genomics. Our macroevolutionary comparisons across insect orders, now extended to the hemimetabolous Hemiptera, reveal unexpected patterns of molecular evolution. We also show how hemipteran feeding ecology and suites of related biological characters are reflected in the genome.

The gene structure trends we identified, with lineage predominating over genome size as a predictor and with many intron gains in the hemipteroid lineage (Figs. 6 and 7), offer initial parameters and hypotheses for the Hemiptera, Coleoptera, and Diptera. Such ordinal-level parameters can be evaluated against new species' data and also inform customized pipelines for automated gene model predictions. At the same time, it will be interesting to explore the ramifications of hemipteroid intron gains, as there are few documented lineages with episodic intron gain [76]. For example, possessing more, small exons may provide greater scope to generate protein modularity via isoforms based on alternative exon usage [119]. Furthermore, with the larger genome sizes and lower gene densities of hemipteroids compared to the well-studied Hymenoptera, it remains open whether hemipteroid gene and intron size may also correlate with recombination rates [120].

Our analyses also highlight new directions for future experimental research, building on *Oncopeltus*'s long-standing history as a laboratory model and its active research community in the modern molecular genetics era (e.g., Fig. 9, [25–27]). Functional testing will clarify the roles of genes we have identified as unique to the Hemiptera, including those implicated in chemical protection, bacterial and plant cell wall metabolism, or encoding wholly novel proteins (Figs. 3 and 8, see also Additional file 1: Supplemental Note 2.2). Meanwhile, the prominent and species-specific expansions specifically of ZF271-like zinc fingers (Fig. 4), combined with the absence of the co-repressor KAP-1 in insects, argues for investigation into alternative interaction partners, which could clarify the nature of these zinc fingers' regulatory roles and their binding targets.

One key output of this study is the generation of a metabolism database for *Oncopeltus*, contributing to the ArthropodaCyc collection (Table 3). In addition to comparisons with other species (Fig. 10), this database can also serve as a future reference for studies that use *Oncopeltus* as an ecotoxicology model (e.g., [121]). While we have primarily focused on feeding ecology in terms of broad comparisons between phytophagy and fluid feeding, *Oncopeltus* is also poised to support future work on nuances among phytophagous species. Despite its milkweed diet in the wild, the lab strain of *Oncopeltus* has long been adapted to feed on sunflower seeds, demonstrating a latent capacity for more generalist phytophagy [112]. This potential may also be reflected in a larger gustatory receptor repertoire than would be expected for

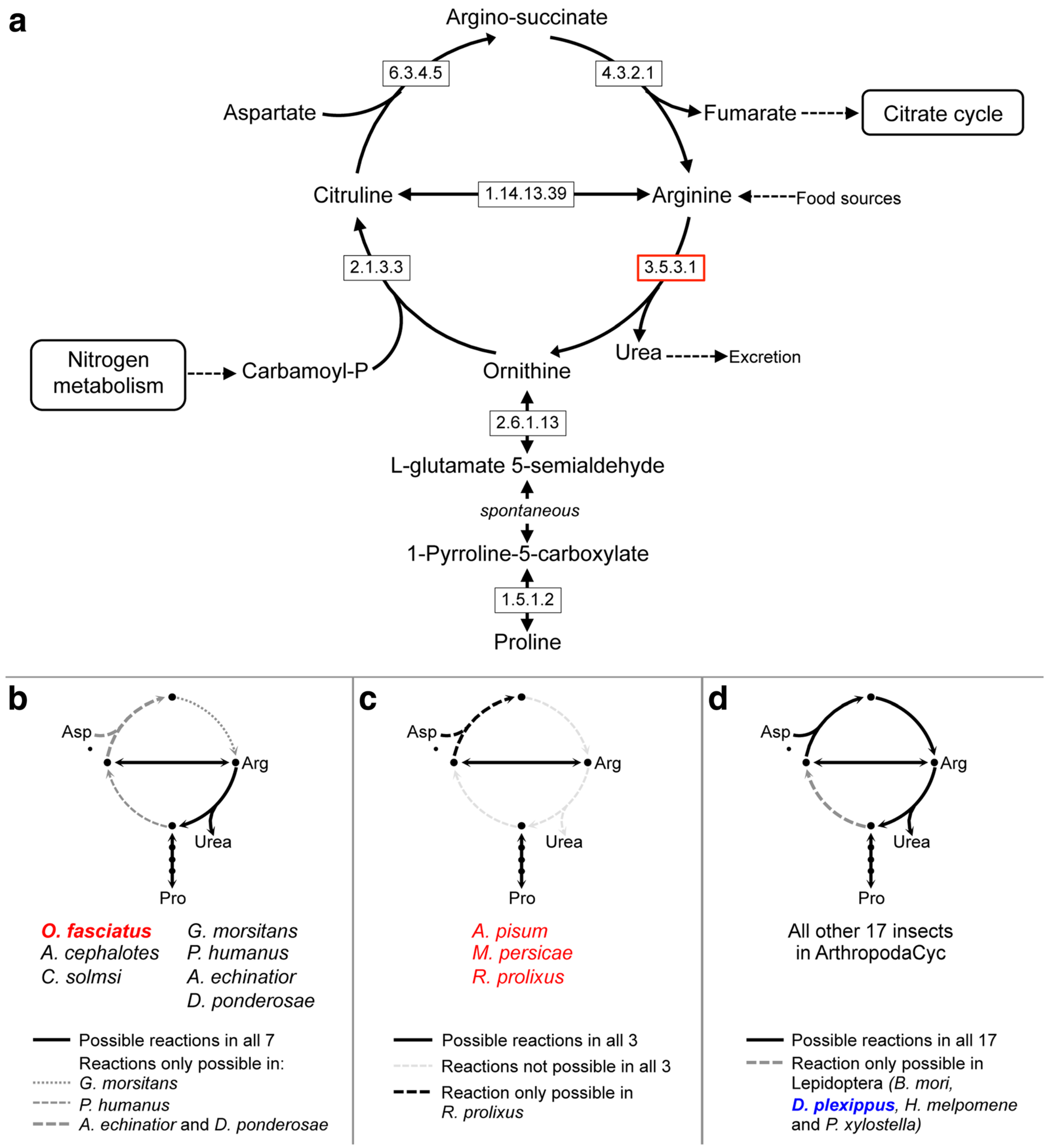

a
Argino-succinate
6.3.4.5
4.3.2.1
Aspartate
Fumarate
Citrate cycle
Citruline
1.14.13.39
Arginine
Food sources
2.1.3.3
3.5.3.1
Nitrogen metabolism
Carbamoyl-P
Ornithine
Urea
Excretion
2.6.1.13
L-glutamate 5-semialdehyde
spontaneous
1-Pyrroline-5-carboxylate
1.5.1.2
Proline
b
Asp
Arg
Urea
Pro
O. fasciatus
A. cephalotes
C. solmsi
G. morsitans
P. humanus
A. echinatior
D. ponderosae
Possible reactions in all 7
Reactions only possible in:
G. morsitans
P. humanus
A. echinatior and D. ponderosae
c
Asp
Arg
Urea
Pro
A. pisum
M. persicae
R. prolixus
Possible reactions in all 3
Reactions not possible in all 3
Reaction only possible in R. prolixus
d
Asp
Arg
Urea
Pro
All other 17 insects in ArthropodaCyc
Possible reactions in all 17
Reaction only possible in Lepidoptera (B. mori, D. plexippus, H. melpomene and P. xylostella)

Fig. 10 Comparison of the urea cycle of *Oncopeltus* with 26 other insect species. **a** Detailed diagram of the urea cycle (adapted from KEGG). **b** Group of 7 species, including *Oncopeltus*, for which Arg degradation via arginase (3.5.3.1), but not synthesis, is possible. **c** Group of 3 species for which neither the degradation nor synthesis of arginine via the urea cycle is possible (the 3 other hemipterans in this analysis). **d** Group of 17 species sharing a complete (or almost complete) urea cycle. Hemiptera are identified in red text, and the milkweed-feeding monarch butterfly is in blue text. Enzyme names corresponding to EC numbers: 1.5.1.2 = pyrroline-5-carboxylate reductase, 1.14.13.39 = nitric-oxide synthase, 2.1.3.3 = ornithine carbamoyltransferase, 2.6.1.13 = ornithine aminotransferase, 3.5.3.1 = arginase, 4.3.2.1 = argininosuccinate lyase, 6.3.4.5 = argininosuccinate synthase. Analyses are based on OGS v1.1

an obligate specialist feeder (Table 2). Thus, *Oncopeltus* can serve as a reference species for promiscuously phytophagous pests such as the stink bug. Finally, we have identified a number of key genes implicated in life history trade-offs in *Oncopeltus*, for traits such as cardenolide tolerance, pigmentation, and plasticity in reproduction under environmental variation. The genome data thus represent an important tool to elucidate the proximate mechanisms of fundamental aspects of life history evolution in both the laboratory and nature.

Methods

(More information is available in Additional file 1: Supplemental Notes.)

Milkweed bug strain, rearing, and DNA/RNA extraction

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

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

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

Genome size estimations were obtained by flow cytometry with Hare and Johnston's protocol [122]. Four to five females and males each from the Carolina Biological Supply lab strain and a wild strain (collected from Athens, GA, USA; GPS coordinates: 33° 56′ 52.8216″ N, 83° 22′ 38.3484″ W) were measured (see also Additional file 1: Supplemental Note 2.1.a). At the bioinformatic level, we attempted to estimate the genome size by *k*-mer spectrum distribution analysis for a range of $k = 15$ to 34 counted with Jellyfish 2.1.4 [123] and bbmap [124], graphing these counts against the frequency of occurrence of *k*-mers (depth) and calculating genome size based on the coverage at the peak of the distribution (Additional file 1: Supplemental Note 2.1.b).

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

Library preparation, sequencing, assembly, and automatic gene annotation were conducted at the Baylor College of Medicine Human Genome Sequencing Center (as in [12, 30]). About 1.1 billion 100-bp paired-end reads generated on an Illumina HiSeq2000s machine were assembled using ALLPATHS-LG [125], from two paired-end (PE) and two mate pair (MP) libraries specifically designed for this algorithm (Additional file 1: Supplemental Note 1). Three libraries were sequenced from an individual adult male (180- and 500-bp PE, 3-kb MP), with the fourth from an individual adult female (8–10-kb MP). The final assembly, "Ofas_1.0" (see metrics in Table 1), has been deposited in GenBank (assembly accession GCA_000696205.1).

Automated annotation of protein-coding genes was performed using a Maker 2.0 annotation pipeline [126] tuned specifically for arthropods (Additional file 1: Supplemental Note 3). These gene predictions were used as the starting point for manual curation via the Apollo v.1.0.4 web browser interface [127], and automatic and manual curations were compiled to generate the OGS (see also Additional file 1: Supplemental Note 4). The current version of the gene set, OGS v1.2, is deposited at GenBank as an "annotation-only" update to the Whole Genome Shotgun project (accession JHQO00000000). Here, we describe version JHQO02000000. The annotations can be downloaded from NCBI's ftp site, ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/696/205/GCA_000696205.1_Ofas_1.0/. The annotations are also available through the i5K Workspace@NAL [128], https://i5k.nal.usda.gov/data/Arthropoda/oncfas-(Oncopeltus_fasciatus)/Ofas_1.0/2.Official%20or%20Primary%20Gene%20Set/GCA_000696205.1_Ofas_1.0/.

Databases of the genome assembly (definitive Illumina-only: Table 1, Additional file 1: Supplemental Note 1.3; provisional hybrid Illumina-PacBio: see below, Additional file 1: Supplemental Note 2.3), Maker automatic gene predictions (Additional file 1: Supplemental Note 3), and OGS v1.1 (Table 1, Additional file 1: Supplemental Note 4) are available through the i5K Workspace@NAL, and the Ag Data Commons data access system of the US Department of Agriculture's (USDA) National Agricultural Library as individual citable databases [129–132].

Repeat content analysis

Repetitive regions were identified in the *Oncopeltus* genome assembly with RepeatModeler Open-1.0.8 [133] based on a species-specific repeat library generated de novo with RECON [134], RepeatScout [135], and Tandem Repeats Finder [136]. Then, RepeatMasker Open-4.0 [137] was used to mask the repeat sequences based on the

RepeatModeler library. Given the fragmented nature of the assembly, we attempted to fill and close the assembly gaps by sequencing additional material, generating long reads with single molecule real-time sequencing on a PacBio RS II machine (estimated coverage of 8x). Gap filling on the Illumina assembly scaffolds was performed with PBJelly version 13.10.22, and the resulting assembly [132] was used for repeat content estimation and comparison with *Cimex lectularius* and *Acyrthosiphon pisum* (see also Additional file 1: Supplemental Note 2.3).

Transcriptome resources

Total RNA from three distinct life history samples (pooled, mixed-instar nymphs; an adult male; an adult female) was also sequenced on an Illumina HiSeq2000s machine, producing a total of 72 million 100-bp paired-end reads (Additional file 1: Supplemental Note 1.3, Additional file 2: Table S1.1; GenBank Bioproject: PRJNA275739). These expression data were used to support the generation of the OGS at different stages of the project: as input for the evidence-guided automated annotation with Maker 2.0 (Additional file 1: Supplemental Note 3), as expression evidence tracks in the Apollo browser to support the community curation of the OGS, and, once assembled into a de novo transcriptome, as a point of comparison for quality control of the OGS.

The raw RNA-seq reads were pre-processed by filtering out low-quality bases (phred score < 30) and Truseq adapters with Trimmomatic-0.30. Further filtering removed ribosomal and mitochondrial RNA sequences with Bowtie 2 [138], based on a custom library built with all hemipteran ribosomal and mitochondrial RNA accessions from NCBI as of 7th February 2014 (6069 accessions). The pooled, filtered reads were mapped to the genome assembly with Tophat2-PE on CyVerse [139]. A second set of RNA-seq reads from an earlier study ("published adult" dataset, [36]) was also filtered and mapped in the same fashion, and both datasets were loaded into the *Oncopeltus* Apollo instance as evidence tracks (under the track names "pooled RNA-seq - cleaned reads" and "RNA-seq raw PE reads Andolfatto et al", respectively).

Additionally, a de novo transcriptome was generated from our filtered RNA-seq reads (pooled from all three samples prepared in this study) using Trinity [140] and TransDecoder [141] with default parameters. This transcriptome is referred to as "i5K," to distinguish it from a previously published maternal and early embryonic transcriptome for *Oncopeltus* (referred to as "454", [35]). Both the i5K and 454 transcriptomes were mapped to the genome assembly with GMAP v. 2014-05-15 on CyVerse. These datasets were also loaded into the Apollo browser as evidence tracks to assist in manual curation.

Life history stage-specific and sex-specific expression analyses in hemipteroids

Transcript expression of the OGS v1.1 genes was estimated by running RSEM2 [142] on the filtered RNA-seq datasets for the three i5K postembryonic stages against the OGS v1.1 cDNA dataset. Transcript expression was then based on the transcripts per million (TPM) value. The TPM values were processed by adding a value of 1 (to avoid zeros) and then performing a log2 transformation. The number of expressed genes per RNA-seq library was compared for TPM cutoffs of > 1, > 0.5, and > 0.25. A > 0.25 cutoff was chosen, which reduced the number of expressed genes by 6.6% compared to a preliminary analysis based on a simple cutoff of ≥ 10 mapped reads per transcript, while the other TPM cutoffs were deemed too restrictive (reducing the expressed gene set by > 10%). This analysis was also applied to the "published adult" dataset [36]. To include the embryonic stages in the comparison, transcripts from the 454 transcriptome were used as blastn queries against the OGS v1.1 cDNA dataset (cutoff *e* value $< 10^{-5}$). The results from all datasets were converted to a binary format to generate Venn diagrams (Fig. 2b).

Statistically significant sex-specific and developmental stage-specific gene enrichment was determined from RNA-seq datasets according to published methods [143, 144], with modifications. Data from *Oncopeltus* (see the previous methods section, Bioproject: PRJNA275739) were compared between stages and pairwise with the hemipterans *Cimex lectularius*, PRJNA275741; *Acyrthosiphon pisum*, PRJNA209321; and *Pachypsylla venusta*, PRJNA275248; as well as with the hemipteroid *Frankliniella occidentalis* (Thysanoptera), PRJNA203209 (see also Fig. 2c, Additional file 1: Supplemental Note 2.4).

Protein orthology assessments via OrthoDB and BUSCO analyses

These analyses follow previously described approaches and with the current database and pipeline versions [1, 42, 44, 145], see Additional file 1: Supplemental Note 6.1 for further details.

Global transcription factor identification

Likely transcription factors (TFs) were identified by scanning the amino acid sequences of predicted protein-coding genes for putative DNA binding domains (DBDs), and when possible, the DNA binding specificity of each TF was predicted using established procedures [57]. Briefly, all protein sequences were scanned for putative DBDs using the 81 Pfam [146] models listed in Weirauch and Hughes [147] and the HMMER tool [148], with the recommended detection thresholds of Per-sequence Eval < 0.01 and Per-domain conditional Eval < 0.01. Each protein was classified into a family based on its DBDs and their order in the protein sequence (e.g., bZIPx1, AP2x2, Homeodomain+Pou). The resulting DBD amino acid sequences were then aligned within each family using Clustal Omega [149], with default settings. For protein pairs with multiple DBDs, each DBD was aligned separately. From these alignments, the sequence identity was calculated for all DBD sequence pairs (i.e., the percent of amino acid residues that are identical across all positions in the alignment). Using previously established sequence identity thresholds for each family [57], the predicted DNA binding specificities were mapped by simple transfer. For example, the DBD of OFAS001246-RA is 98% identical to the *Drosophila melanogaster* Bric a Brac 1 (Bab1) protein. Since the DNA binding specificity of Bab1 has already been experimentally determined, and the cutoff for the Pipsqueak family TFs is 85%, we can infer that OFAS001246-RA will have the same binding specificity as *Drosophila* Bab1.

RNA interference

Double-stranded RNA (dsRNA) was designed to target the final, unique exon of the *broad* isoforms *Z2*, *Z3*, and *Z4*. A portion of the coding sequence for the zinc finger region from these exons (179 bp, 206 bp, and 216 bp, respectively) was cloned into a plasmid vector and used as a template for in vitro RNA synthesis, using the gene-specific primer pairs: Of-Z2_fwd: 5′-ATGTGGCAGACAAGCATGCT-3′, Of-Z2_rev: 5′-CTAAAATTTGACATCAGTAGGC-3′; Of-Z3_fwd: 5′-CCTTCTCCTGTTACTACTCAC-3′, Of-Z3_rev: 5′-TTATATGGGCGGCTGTCCAA-3′; and Of-Z4_fwd: 5′-AACACTGACCTTGGTTACACA-3′, Of-Z4_rev: 5′-TAGGTGGAGGATTGCTAAAATT-3′. Two separate transcription reactions (one for each strand) were performed using the Ambion MEGAscript kit (Ambion, Austin, TX, USA). The reactions were purified by phenol/chloroform extraction followed by precipitation as described in the MEGAscript protocol. The separate strands were re-annealed in a thermocycler as described previously [27]. Nymphs were injected with a Hamilton syringe fitted with a 32-gauge needle as described [54]. The concentration of *Of-Z2*, *Of-Z3*, and *Of-Z4* dsRNA was 740 ng/μl, 1400 ng/μl, and 1200 ng/μl, respectively. All nymphs were injected within 8 h of the molt to the fourth (penultimate juvenile) instar ($n \geq 12$ per treatment, see Fig. 9). Fore- and hindwings were then dissected from adults and photographed at the same scale as wings from wild type, uninjected controls.

CycADS annotation and OncfaCyc database generation

We used the Cyc Annotation Database System (CycADS [113]), an automated annotation management system, to integrate protein annotations from different sources into a Cyc metabolic network reconstruction that was integrated into the ArthropodaCyc database. Using our CycADS pipeline, *Oncopeltus fasciatus* proteins from the official gene set OGS v1.1 were annotated using different methods—including KAAS [150], PRIAM [151], Blast2GO [152, 153], and InterProScan with several approaches [154]—to obtain EC and GO numbers. All annotation information data were collected in the CycADS SQL database and automatically extracted to generate appropriate input files to build or update BioCyc databases [155] using the Pathway Tools software [156]. The OncfaCyc database, representing the metabolic protein-coding genes of *Oncopeltus*, was thus generated and is now included in the ArthropodaCyc database, a collection of arthropod metabolic network databases ([114], http://arthropodacyc.cycadsys.org/).

Additional files

Additional file 1: Supplementary notes, figures, and small tables. (PDF 6142 kb)

Additional file 2: Large supporting tables. (XLSX 2222 kb)

Additional file 3: Chemoreceptor sequences in FASTA format. (FASTA 128 kb)

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; NIH grant R01GM113230 (NIGMS) to LP; 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 and in the USDA Ag Data Commons data access system [133]. In addition, assembled scaffolds, gene models, and a browser are available at the National Agricultural Library ([130–133], https://i5k.nal.usda.gov/

Oncopeltus_fasciatus). The OncfaCyc metabolism database is available within the ArthropodaCyc collection (http://arthropodacyc.cycadsys.org/).

Authors' contributions

KAP and StR conceived the project. KAP managed and coordinated the project. KAP and SK provided the specimens for sequencing and performed the DNA and RNA extractions. StR, SD, SLL, HC, HVD, HD, YH, JQ, SCM, DSTH, KCW, DMM, and RAG constructed the libraries and performed the sequencing. StR, SCM, and DSTH performed the genome assembly and automated gene prediction. IMVJ, JSJ, and PJM analyzed the genome size. IMVJ, VK, PH, and KAP contributed to the repetitive content analyses. AD, RR, JHW, KAP, and SK performed the bacterial scaffold detection and LGT analyses. MCMT developed the Apollo software. KAP, IMVJ, MCMT, CPC, C-YL, and MFP implemented the 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, MLP, PNR, RR-P, SiR, LS, MES, JS, ES, JNS, OT, LT, MVDZ, SV, and AJR participated in the manual curation and contributed to the supplemental notes. IMVJ, KAP, DSTH, M-JMC, CPC, C-YL, and MFP performed the curation quality control and generated the OGS. IMVJ, KAP, CJH, and JBB generated the de novo transcriptomes and performed the life history stage expression analyses. RMW, PI, KAP, and EMZ performed the orthology and phylogenomic analyses. MTW, KAP, IMVJ, PH, and BMIV performed the transcription factor analyses. EJD conducted the analyses of DNA methylation. KAP, PH, and RJS contributed to the comparative analyses of gene structure. DE conducted the RNAi experiments. SC, PB-P, GF, and NP generated and performed the 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.

Ethics approval and consent to participate

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

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, OH 45221, USA. [4]Department of Biochemistry and Cell Biology and Center for Developmental Genetics, Stony Brook University, Stony Brook, NY 11794, USA. [5]Department of Biological Sciences, Wellesley College, 106 Central St., Wellesley, MA 02481, USA. [6]Univ Lyon, INSA-Lyon, INRA, BF2I, UMR0203, F-69621 Villeurbanne, France. [7]Present 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, IL 61801, USA. [9]National Agricultural Library, Beltsville, MD 20705, USA. [10]Department of Genetic Medicine and Development and Swiss Institute of Bioinformatics, University of Geneva, 1211 Geneva, Switzerland. [11]Present address: Department of Ecology and Evolution, University of Lausanne, 1015 Lausanne, 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, OH 45229, USA. [13]Human Genome Sequencing Center, Department of Human and Molecular Genetics, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030, USA. [14]Present address: Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA 98195, USA. [15]Present address: Howard Hughes Medical Institute, University of Washington, Seattle, WA 98195, USA. [16]Department of Biology, University of Rochester, Rochester, NY 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]Present address: School of Life Sciences, Rochester Institute of Technology, Rochester, NY 14623, USA. [25]Department of Organismic and Evolutionary Biology, Harvard University, 26 Oxford Street, Cambridge, MA 02138, USA. [26]Department of Molecular and Cellular Biology, Harvard University, 26 Oxford Street, Cambridge, MA 02138, USA. [27]Department of Biological Sciences, Wayne State University, Detroit, MI 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, TX 77843, USA. [30]CECAD, University of Cologne, Cologne, Germany. [31]Department of Entomology and Program in Molecular & Cell Biology, University of Maryland, College Park, MD 20742, USA. [32]Department of Entomology, University of Georgia, 120 Cedar St., Athens, GA 30602, USA. [33]Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. [34]Department of Entomology, College of Agriculture, Food and Environment, University of Kentucky, Lexington, KY 40546, USA. [35]Department of Biology, University of Hawai'i at Mānoa, Honolulu, HI 96822, USA. [36]Present address: Department of Evolutionary Genetics, Max-Planck-Institut für Evolutionsbiologie, August-Thienemann-Straße 2, 24306 Plön, Germany. [37]Present address: Earthworks Institute, 185 Caroline Street, Rochester, NY 14620, USA. [38]Centro de Bioinvestigaciones, Universidad Nacional del Noroeste de Buenos Aires, Pergamino, Argentina. [39]Present 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]Present 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. [43]Present address: Department of Physiology, Anatomy and Genetics and Centre for Neural Circuits and Behaviour, University of Oxford, Oxford OX1 3SR, UK. [44]Present address: E. A. Milne Centre for Astrophysics, Department of Physics and Mathematics, University of Hull, Hull HU6 7RX, UK. [45]Research Center for Developmental Biology and Regenerative Medicine, National Taiwan University, Taipei, Taiwan.

Received: 2 November 2018 Accepted: 21 February 2019
Published online: 02 April 2019

References

1. Zdobnov EM, Tegenfeldt F, Kuznetsov D, Waterhouse RM, Simão FA, Ioannidis P, Seppey M, Loetscher A, Kriventseva EV. OrthoDB v9.1: cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs. Nucleic Acids Res. 2017;45:D744–9.
2. Huang DY, Bechly G, Nel P, Engel MS, Prokop J, Azar D, Cai CY, van de Kamp T, Staniczek AH, Garrouste R, et al. New fossil insect order Permopsocida elucidates major radiation and evolution of suction feeding in hemimetabolous insects (Hexapoda: Acercaria). Sci Rep. 2016;6:23004.
3. Misof B, Liu S, Meusemann K, Peters RS, Donath A, Mayer C, Frandsen PB, Ware J, Flouri T, Beutel RG, et al. Phylogenomics resolves the timing and pattern of insect evolution. Science. 2014;346:763–7.
4. Grimaldi D, Engel MS. Evolution of the insects. Cambridge: Cambridge University Press; 2005.
5. Panfilio KA, Angelini DR. By land, air, and sea: hemipteran diversity through the genomic lens. Curr Opin Insect Sci. 2018;25:106–15.
6. The International Aphid Genomics Consortium. Genome sequence of the pea aphid Acyrthosiphon pisum. PLoS Biol. 2010;8:e1000313.
7. Mathers TC, Chen Y, Kaithakottil G, Legeai F, Mugford ST, Baa-Puyoulet P, Bretaudeau A, Clavijo B, Colella S, Collin O, et al. Rapid transcriptional plasticity of duplicated gene clusters enables a clonally reproducing aphid to colonise diverse plant species. Genome Biol. 2017;18:27.
8. Wenger JA, Cassone BJ, Legeai F, Johnston JS, Bansal R, Yates AD, Coates BS, Pavinato VA, Michel A. Whole genome sequence of the soybean aphid, Aphis glycines. Insect Biochem Mol Biol. 2017; in press.

9. Sloan DB, Nakabachi A, Richards S, Qu J, Murali SC, Gibbs RA, Moran NA. Parallel histories of horizontal gene transfer facilitated extreme reduction of endosymbiont genomes in sap-feeding insects. Mol Biol Evol. 2014;31:857–71.
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, et al. Genomes of the rice pest brown planthopper and its endosymbionts reveal complex complementary contributions for host adaptation. Genome Biol. 2014;15:521.
11. Mesquita RD, Vionette-Amaral RJ, Lowenberger C, Rivera-Pomar R, Monteiro FA, Minx P, Spieth J, Carvalho AB, Panzera F, Lawson D, et al. Genome of *Rhodnius prolixus*, an insect vector of Chagas disease, reveals unique adaptations to hematophagy and parasite infection. Proc Natl Acad Sci U S A. 2015;112:14936–41.
12. Benoit JB, Adelman ZN, Reinhardt K, Dolan A, Poelchau M, Jennings EC, Szuter EM, Hagan RW, Gujar H, Shukla JN, et al. Unique features of a global human ectoparasite identified through sequencing of the bed bug genome. Nat Commun. 2016;7:10165.
13. Rosenfeld JA, Reeves D, Brugler MR, Narechania A, Simon S, Durrett R, Foox J, Shianna K, Schatz MC, Gandara J, et al. Genome assembly and geospatial phylogenomics of the bed bug *Cimex lectularius*. Nat Commun. 2016;7:10164.
14. Sparks ME, Shelby KS, Kuhar D, Gundersen-Rindal DE. Transcriptome of the invasive brown marmorated stink bug, *Halyomorpha halys* (Stal) (Heteroptera: Pentatomidae). PLoS One. 2014;9:e111646.
15. Ioannidis P, Lu Y, Kumar N, Creasy T, Daugherty S, Chibucos MC, Orvis J, Shetty A, Ott S, Flowers M, et al. Rapid transcriptome sequencing of an invasive pest, the brown marmorated stink bug, *Halyomorpha halys*. BMC Genomics. 2014;15:738.
16. Li H, Leavengood JM Jr, Chapman EG, Burkhardt D, Song F, Jiang P, Liu J, Zhou X, Cai W. Mitochondrial phylogenomics of Hemiptera reveals adaptive innovations driving the diversification of true bugs. Proc Biol Sci. 2017;284: 20171223.
17. Wilson ACC, Ashton PD, Charles H, Colella S, Febvay G, Jander G, Kushlan PF, Macdonald SJ, Schwartz JF, Thomas GH, Douglas AE. Genomic insight into the amino acid relations of the pea aphid, *Acyrthosiphon pisum*, with its symbiotic bacterium *Buchnera aphidicola*. Insect Mol Biol. 2010;19(Suppl 2):249–58.
18. Eichler S, Schaub GA. Development of symbionts in triatomine bugs and the effects of infections with trypanosomatids. Exp Parasitol. 2002;100:17–27.
19. Matsuura Y, Kikuchi Y, Hosokawa T, Koga R, Meng X-Y, Kamagata Y, Nikoh N, Fukatsu T. Evolution of symbiotic organs and endosymbionts in lygaeid stinkbugs. The ISME Journal. 2012;6:397–409.
20. Berenbaum MR, Miliczky E. Mantids and milkweed bugs - efficacy of aposematic coloration against invertebrate predators. Am Midl Nat. 1984;111:64–8.
21. Burdfield-Steel ER, Shuker DM. The evolutionary ecology of the Lygaeidae. Ecol Evol. 2014;4:2278–301.
22. Lawrence PA. Mitosis and the cell cycle in the metamorphic moult of the milkweed bug *Oncopeltus fasciatus*; a radioautographic study. J Cell Sci. 1968;3:391–404.
23. Chipman AD. *Oncopeltus fasciatus* as an evo-devo research organism. Genesis. 2017;55:e23020.
24. Panfilio KA. Late extraembryonic development and its *zen-RNAi*-induced failure in the milkweed bug *Oncopeltus fasciatus*. Dev Biol. 2009;333: 297–311.
25. Panfilio KA, Roth S. Epithelial reorganization events during late extraembryonic development in a hemimetabolous insect. Dev Biol. 2010; 340:100–15.
26. Sharma AI, Yanes KO, Jin L, Garvey SL, Taha SM, Suzuki Y. The phenotypic plasticity of developmental modules. Evodevo. 2016;7:15.
27. Hughes CL, Kaufman TC. RNAi analysis of *Deformed*, *proboscipedia* and *Sex combs reduced* in the milkweed bug *Oncopeltus fasciatus*: novel roles for Hox genes in the hemipteran head. Development. 2000;127:3683–94.
28. Wolfe SL, John B. The organization and ultrastructure of male meiotic chromosomes in *Oncopeltus fasciatus*. Chromosoma. 1965;17:85–103.
29. Messthaler H, Traut W. Phases of sex chromosome inactivation in *Oncopeltus fasciatus* and *Pyrrhocoris apterus* (Insecta, Heteroptera). Caryologia. 1975;28:501–10.
30. McKenna DD, Scully ED, Pauchet Y, Hoover K, Kirsch R, Geib SM, Mitchell RF, Waterhouse RM, Ahn SJ, Arsala D, et al. Genome of the Asian longhorned beetle (*Anoplophora glabripennis*), a globally significant invasive species, reveals key functional and evolutionary innovations at the beetle-plant interface. Genome Biol. 2016;17:227.
31. Simpson JT. Exploring genome characteristics and sequence quality without a reference. Bioinformatics. 2014;30:1228–35.
32. Hanrahan SJ, Johnston JS. New genome size estimates of 134 species of arthropods. Chromosom Res. 2011;19:809–23.
33. Panfilio KA, Liu PZ, Akam M, Kaufman TC. *Oncopeltus fasciatus zen* is essential for serosal tissue function in katatrepsis. Dev Biol. 2006;292:226–43.
34. Tian X, Xie Q, Li M, Gao C, Cui Y, Xi L, Bu W. Phylogeny of pentatomomorphan bugs (Hemiptera-Heteroptera:Pentatomomorpha) based on six Hox gene fragments. Zootaxa. 2011;2888:57–68.
35. Ewen-Campen B, Shaner N, Panfilio KA, Suzuki Y, Roth S, Extavour CG. The maternal and early embryonic transcriptome of the milkweed bug *Oncopeltus fasciatus*. BMC Genomics. 2011;12:61.
36. Zhen Y, Aardema ML, Medina EM, Schumer M, Andolfatto P. Parallel molecular evolution in an herbivore community. Science. 2012;337:1634–7.
37. Robertson HM. The insect chemoreceptor superfamily in Drosophila pseudoobscura: molecular evolution of ecologically-relevant genes over 25 million years. J Insect Sci. 2009;9:18.
38. Robertson HM. Taste: independent origins of chemoreception coding systems? Curr Biol. 2001;11:R560–2.
39. Jazwinska A, Rushlow C, Roth S. The role of *brinker* in mediating the graded response to Dpp in early *Drosophila* embryos. Development. 1999;126:3323–34.
40. Togawa T, Dunn WA, Emmons AC, Nagao J, Willis JH. Developmental expression patterns of cuticular protein genes with the R&R Consensus from *Anopheles gambiae*. Insect Biochem Mol Biol. 2008;38:508–19.
41. Karr TL: Fruit flies and the sperm proteome. Hum Mol Genet 2007, 16 Spec No. 2:R124–R133.
42. Waterhouse RM, Seppey M, Simao FA, Manni M, Ioannidis P, Klioutchnikov G, Kriventseva EV, Zdobnov EM. BUSCO applications from quality assessments to gene prediction and phylogenomics. Mol Biol Evol. 2017;35: 543–8.
43. Shigenobu S, Bickel RD, Brisson JA, Butts T, Chang CC, Christiaens O, Davis GK, Duncan EJ, Ferrier DE, Iga M, et al. Comprehensive survey of developmental genes in the pea aphid, *Acyrthosiphon pisum*: frequent lineage-specific duplications and losses of developmental genes. Insect Mol Biol. 2010;19(Suppl 2):47–62.
44. Kriventseva EV, Tegenfeldt F, Petty TJ, Waterhouse RM, Simão FA, Pozdnyakov IA, Ioannidis P, Zdobnov EM. OrthoDB v8: update of the hierarchical catalog of orthologs and the underlying free software. Nucl Acids Res. 2015;43:D250–6.
45. Bansal R, Michel AP. Core RNAi machinery and *Sid1*, a component for systemic RNAi, in the hemipteran insect, *Aphis glycines*. Int J Mol Sci. 2013; 14:3786–801.
46. Bao R, Fischer T, Bolognesi R, Brown SJ, Friedrich M. Parallel duplication and partial subfunctionalization of beta-catenin/armadillo during insect evolution. Mol Biol Evol. 2012;29:647–62.
47. Sachs L, Chen YT, Drechsler A, Lynch JA, Panfilio KA, Lassig M, Berg J, Roth S. Dynamic BMP signaling polarized by Toll patterns the dorsoventral axis in a hemimetabolous insect. eLife. 2015;4:e05502.
48. Armisen D, Refki PN, Crumiere AJ, Viala S, Toubiana W, Khila A. Predator strike shapes antipredator phenotype through new genetic interactions in water striders. Nat Commun. 2015;6:8153.
49. Konopova B, Smykal V, Jindra M. Common and distinct roles of juvenile hormone signaling genes in metamorphosis of holometabolous and hemimetabolous insects. PLoS One. 2011;6:e28728.
50. Vellichirammal NN, Gupta P, Hall TA, Brisson JA. Ecdysone signaling underlies the pea aphid transgenerational wing polyphenism. Proc Natl Acad Sci U S A. 2017;114:1419–23.
51. Wulff JP, Sierra I, Sterkel M, Holtof M, Van Wielendaele P, Francini F, Broeck JV, Ons S. Orcokinin neuropeptides regulate ecdysis in the hemimetabolous insect *Rhodnius prolixus*. Insect Biochem Mol Biol. 2017;81:91–102.
52. Chiu TL, Wen Z, Rupasinghe SG, Schuler MA. Comparative molecular modeling of *Anopheles gambiae* CYP6Z1, a mosquito P450 capable of metabolizing DDT. Proc Natl Acad Sci U S A. 2008;105:8855–60.
53. Gong Y, Li T, Feng Y, Liu N. The function of two P450s, CYP9M10 and CYP6AA7, in the permethrin resistance of *Culex quinquefasciatus*. Sci Rep. 2017;7:587.
54. Liu PZ, Kaufman TC. *hunchback* is required for suppression of abdominal identity, and for proper germband growth and segmentation in the intermediate germband insect *Oncopeltus fasciatus*. Development. 2004;131:1515–27.
55. Schaeper ND, Pechmann M, Damen WGM, Prpic N-M, Wimmer EA. Evolutionary plasticity of *collier* function in head development of diverse arthropods. Dev Biol. 2010;344:363–76.

56. Aspiras AC, Smith FW, Angelini DR. Sex-specific gene interactions in the patterning of insect genitalia. Dev Biol. 2011;360:369–80.
57. Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, Najafabadi HS, Lambert SA, Mann I, Cook K, et al. Determination and inference of eukaryotic transcription factor sequence specificity. Cell. 2014;158:1431–43.
58. Peel AD, Telford MJ, Akam M. The evolution of hexapod engrailed-family genes: evidence for conservation and concerted evolution. Proc Biol Sci. 2006;273:1733–42.
59. Ben-David J, Chipman AD. Mutual regulatory interactions of the trunk gap genes during blastoderm patterning in the hemipteran *Oncopeltus fasciatus*. Dev Biol. 2010;346:140–9.
60. Erezyilmaz DF, Kelstrup HC, Riddiford LM. The nuclear receptor E75A has a novel pair-rule-like function in patterning the milkweed bug, *Oncopeltus fasciatus*. Dev Biol. 2009;334:300–10.
61. Liu PZ, Kaufman TC. *even-skipped* is not a pair-rule gene but has segmental and gap-like functions in *Oncopeltus fasciatus*, an intermediate germband insect. Development. 2005;132:2081–92.
62. Weisbrod A, Cohen M, Chipman AD. Evolution of the insect terminal patterning system--insights from the milkweed bug, *Oncopeltus fasciatus*. Dev Biol. 2013;380:125–31.
63. Albertin CB, Simakov O, Mitros T, Wang ZY, Pungor JR, Edsinger-Gonzales E, Brenner S, Ragsdale CW, Rokhsar DS. The octopus genome and the evolution of cephalopod neural and morphological novelties. Nature. 2015; 524:220–4.
64. Crooks GE, Hon G, Chandonia J-M, Brenner SE. WebLogo: a sequence logo generator. Genome Res. 2004;14:1188–90.
65. Najafabadi HS, Mnaimneh S, Schmitges FW, Garton M, Lam KN, Yang A, Albu M, Weirauch MT, Radovani E, Kim PM, et al. C2H2 zinc finger proteins greatly expand the human regulatory lexicon. Nat Biotechnol. 2015;33:555–62.
66. Emerson RO, Thomas JH. Adaptive evolution in zinc finger transcription factors. PLoS Genet. 2009;5:e1000325.
67. Thomas JH, Schneider S. Coevolution of retroelements and tandem zinc finger genes. Genome Res. 2011;21:1800–12.
68. Garcia-Perez JL, Widmann TJ, Adams IR. The impact of transposable elements on mammalian development. Development. 2016;143:4101–14.
69. Liu PZ, Kaufman TC. *Krüppel* is a gap gene in the intermediate insect *Oncopeltus fasciatus* and is required for development of both blastoderm and germband-derived segments. Development. 2004;131:4567–79.
70. Heger P, Marin B, Bartkuhn M, Schierenberg E, Wiehe T. The chromatin insulator CTCF and the emergence of metazoan diversity. Proc Natl Acad Sci U S A. 2012;109:17507–12.
71. Liu H, Chang L-H, Sun Y, Lu X, Stubbs L. Deep vertebrate roots for mammalian zinc finger transcription factor subfamilies. Genome Biol Evol. 2014;6:510–25.
72. Imbeault M, Helleboid P-Y, Trono D. KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. Nature. 2017;543:550–4.
73. Csuros M, Rogozin IB, Koonin EV. A detailed history of intron-rich eukaryotic ancestors inferred from a global survey of 100 complete genomes. PLoS Comput Biol. 2011;7:e1002150.
74. Hoy MA, Waterhouse RM, Wu K, Estep AS, Ioannidis P, Palmer WJ, Pomerantz AF, Simao FA, Thomas J, Jiggins FM, et al. Genome sequencing of the phytoseiid predatory mite *Metaseiulus occidentalis* reveals completely atomized Hox genes and superdynamic intron evolution. Genome Biol Evol. 2016;8:1762–75.
75. Seibt KM, Wenke T, Muders K, Truberg B, Schmidt T. Short interspersed nuclear elements (SINEs) are abundant in Solanaceae and have a family-specific impact on gene structure and genome organization. Plant J. 2016;86:268–85.
76. Huff JT, Zilberman D, Roy SW. Mechanism for DNA transposons to generate introns on genomic scales. Nature. 2016;538:533–6.
77. Wheeler D, Redding AJ, Werren JH. Characterization of an ancient lepidopteran lateral gene transfer. PLoS One. 2012;8:e59262.
78. Da Lage JL, Binder M, Hua-Van A, Janecek S, Casane D. Gene make-up: rapid and massive intron gains after horizontal transfer of a bacterial alpha-amylase gene to Basidiomycetes. BMC Evol Biol. 2013;13:40.
79. Lee DH, Short BD, Joseph SV, Bergh JC, Leskey TC. Review of the biology, ecology, and management of *Halyomorpha halys* (Hemiptera: Pentatomidae) in China, Japan, and the Republic of Korea. Environ Entomol. 2013;42:627–41.
80. Lawrence PA. Cellular differentiation and pattern formation during metamorphosis of the milkweed bug *Oncopeltus*. Dev Biol. 1969;19:12–40.
81. Riddiford LM. Prevention of metamorphosis by exposure of insect eggs to juvenile hormone analogs. Science. 1970;167:287.
82. Willis JH, Lawrence PA. Deferred action of juvenile hormone. Nature. 1970;225:81–3.
83. Masner P, Bowers WS, Kalin M, Muhle T. Effect of precocene II on the endocrine regulation of development and reproduction in the bug, *Oncopeltus fasciatus*. Gen Comp Endocrinol. 1979;37:156–66.
84. Rewitz K, O'Connor M, Gilbert L. Molecular evolution of the insect Halloween family of cytochrome P450s: phylogeny, gene organization and functional conservation. Insect Biochem Mol Biol. 2007;37:741–53.
85. Huet F, Ruiz C, Richards G. Sequential gene activation by ecdysone in *Drosophila melanogaster*: the hierarchical equivalence of early and early late genes. Development. 1995;121:1195–204.
86. Bialecki M, Shilton A, Fichtenberg C, Segraves WA, Thummel CS. Loss of the ecdysteroid-inducible E75A orphan nuclear receptor uncouples molting from metamorphosis in *Drosophila*. Dev Cell. 2002;3:209–20.
87. Charles JP, Iwema T, Epa VC, Takaki K, Rynes J, Jindra M. Ligand-binding properties of a juvenile hormone receptor, methoprene-tolerant. Proc Natl Acad Sci U S A. 2011;108:21128–33.
88. Minakuchi C, Zhou X, Riddiford L. Kruppel homolog 1 (Kr-h1) mediates juvenile hormone action during metamorphosis of *Drosophila melanogaster*. Mech Dev. 2008;125:91–105.
89. Minakuchi C, Namiki T, Shinoda T. Kruppel homolog 1, an early juvenile hormone-response gene downstream of methoprene-tolerant, mediates its anti-metamorphic action in the red flour beetle *Tribolium castaneum*. Dev Biol. 2009;352:341–50.
90. DiBello PR, Withers DA, Bayer CA, Fristrom JW, Guild GM. The *Drosophila Broad-Complex* encodes a family of related proteins containing zinc fingers. Genetics. 1991;129:385–97.
91. Karim F, Guild G, Thummel C. The *Drosophila Broad-Complex* plays a key role in controlling ecdysone-regulated gene expression at the onset of metamorphosis. Development. 1993;118:977–88.
92. Erezyilmaz DF, Riddiford LM, Truman JW. The pupal specifier broad directs progressive morphogenesis in a direct-developing insect. Proc Natl Acad Sci U S A. 2006;103:6925–30.
93. Arakane Y, Hogenkamp DG, Zhu YC, Kramer KJ, Specht CA, Beeman RW, Kanost MR, Muthukrishnan S. Characterization of two chitin synthase genes of the red flour beetle, *Tribolium castaneum*, and alternate exon usage in one of the genes during development. Insect Biochem Mol Biol. 2004;34:291–304.
94. True JR. Insect melanism: the molecules matter. Trends Ecol Evol. 2003;18:640–7.
95. Zhan SA, Guo QH, Li MH, Li MW, Li JY, Miao XX, Huang YP. Disruption of an N-acetyltransferase gene in the silkworm reveals a novel role in pigmentation. Development. 2010;137:4083–90.
96. Liu J, Lemonds TR, Popadic A. The genetic control of aposematic black pigmentation in hemimetabolous insects: insights from *Oncopeltus fasciatus*. Evol Dev. 2014;16:270–7.
97. Liu J, Lemonds TR, Marden JH, Popadic A. A pathway analysis of melanin patterning in a hemimetabolous insect. Genetics. 2016;203:403–13.
98. Lawrence PA. Some new mutants of large milkweed bug *Oncopeltus fasciatus* Dall. Genet Res. 1970;15:347–50.
99. Morgan ED. Biosynthesis in insects: advanced edition. London: Royal Society of Chemistry; 2010.
100. McLean JR, Krishnakumar S, O'Donnell JM. Multiple mRNAs from the *Punch* locus of *Drosophila melanogaster* encode isoforms of GTP cyclohydrolase I with distinct N-terminal domains. J Biol Chem. 1993;268:27191–7.
101. Wiederrecht GJ, Paton DR, Brown GM. Enzymatic conversion of Dihydroneopterin triphosphate to the pyrimidodiazepine intermediate involved in the biosynthesis of the Drosopterins in *Drosophila melanogaster*. J Biol Chem. 1984;259:2195–200.
102. Newcombe D, Blount JD, Mitchell C, Moore AJ. Chemical egg defence in the large milkweed bug, *Oncopeltus fasciatus*, derives from maternal but not paternal diet. Entomologia Experimentalis et Applicata. 2013;149:197–205.
103. Zhan S, Merlin C, Boore JL, Reppert SM. The monarch butterfly genome yields insights into long-distance migration. Cell. 2011;147:1171–85.
104. Robertson HM, Warr CG, Carlson JR. Molecular evolution of the insect chemoreceptor gene superfamily in *Drosophila melanogaster*. Proc Natl Acad Sci U S A. 2003;100(Suppl 2):14537–42.
105. Joseph RM, Carlson JR. *Drosophila* chemoreceptors: a molecular interface between the chemical world and the brain. Trends Genet. 2015;31:683–95.
106. Benton R. Multigene family evolution: perspectives from insect chemoreceptors. Trends Ecol Evol. 2015;30:590–600.
107. Rytz R, Croset V, Benton R. Ionotropic receptors (IRs): chemosensory ionotropic glutamate receptors in *Drosophila* and beyond. Insect Biochem Mol Biol. 2013;43:888–97.

108. Kirkness EF, Haas BJ, Sun W, Braig HR, Perotti MA, Clark JM, Lee SH, Robertson HM, Kennedy RC, Elhaik E, et al. Genome sequences of the human body louse and its primary endosymbiont provide insights into the permanent parasitic lifestyle. Proc Natl Acad Sci U S A. 2010;107:12168–73.
109. Smadja C, Shi P, Butlin RK, Robertson HM. Large gene family expansions and adaptive evolution for odorant and gustatory receptors in the pea aphid, *Acyrthosiphon pisum*. Mol Biol Evol. 2009;26:2073–86.
110. Terrapon N, Li C, Robertson HM, Ji L, Meng X, Booth W, Chen Z, Childers CP, Glastad KM, Gokhale K, et al. Molecular traces of alternative social organization in a termite genome. Nat Commun. 2014;5:3636.
111. 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.
112. Feir D. *Oncopeltus fasciatus*: a research animal. Annu Rev Entomol. 1974;19:81–96.
113. 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.
114. 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;2016:baw081.
115. Hojilla-Evangelista MP, Evangelista RL. Characterization of milkweed (Asclepias spp.) seed proteins. Ind Crops Prod. 2009;29: 275–80.
116. 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;41:152-61.
117. 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.
118. 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;157:30–9.
119. 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.
120. 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.
121. 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;24:8551–61.
122. Hare EE, Johnston JS. Genome size determination using flow cytometry of propidium iodide-stained nuclei. Methods Mol Biol. 2011;772:3–12.
123. Marcais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics. 2011;27:764–70.
124. Bushnell B. BBMap short read aligner; 2016.
125. 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;108:1513–8.
126. Holt C, Yandell M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics. 2011;12:491.
127. 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 Biol. 2013;14:R93.
128. 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:D714–9.
129. 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:https://doi.org/10.15482/USDA.ADC/1173238.
130. Hughes DST, Koelzer S, Panfilio KA, Richards S: *Oncopeltus fasciatus* genome annotations v0.5.3. Ag Data Commons (Database) 2015:https://doi.org/10.15482/USDA.ADC/1173237.
131. 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 v1.1. Ag Data Commons (Database) 2015:https://doi.org/10.15482/USDA.ADC/1173142.
132. Vargas Jentzsch IM, Kovacova V, Stueber K, Koelzer S, Panfilio KA: *Oncopeltus fasciatus* hybrid genome assembly 1.0. Ag Data Commons (Database) 2019: https://doi.org/10.15482/USDA.ADC/1503405.
133. RepeatModeler Open-1.0.8 [http://www.repeatmasker.org]. Accessed 5 June 2015.
134. Bao Z, Eddy SR. Automated de novo identification of repeat sequence families in sequenced genomes. Genome Res. 2002;12:1269–76.
135. Price AL, Jones NC, Pevzner PA. De novo identification of repeat families in large genomes. Bioinformatics. 2005;21(Suppl 1):i351–8.
136. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 1999;27:573–80.
137. RepeatMasker Open-4.0. [http://www.repeatmasker.org]. Accessed 5 June 2015.
138. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357–9.
139. 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. Front Plant Sci. 2011;2:34.
140. 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;29:644–52.
141. 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;8:1494–512.
142. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011;12:323.
143. 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.
144. 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.
145. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31:3210–2.
146. 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:D211–22.
147. Weirauch MT, Hughes TR. A catalogue of eukaryotic transcription factor types, their evolutionary origin, and species distribution. Subcell Biochem. 2011;52:25–73.
148. Eddy SR. A new generation of homology search tools based on probabilistic inference. Genome Inform. 2009;23:205–11.
149. 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.
150. Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M. KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Res. 2007;35:W182–5.
151. Claudel-Renard C, Chevalet C, Faraut T, Kahn D. Enzyme-specific profiles for genome annotation: PRIAM. Nucleic Acids Res. 2003;31:6633–9.
152. Conesa A, Götz S. Blast2GO: a comprehensive suite for functional analysis in plant genomics. Int J Plant Genomics. 2008;2008:619832.
153. 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;21:3674–6.
154. 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;30:1236–40.
155. 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 Res. 2005;33:6083–9.
156. Karp PD, Paley SM, Krummenacker M, Latendresse M, Dale JM, Lee TJ, Kaipa P, Gilham F, Spaulding A, Popescu L, et al. Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology. Brief Bioinform. 2010;11:40–79.
157. Dereeper A, Guignon V, Blanc G, Audic S, Buffet S, Chevenet F, Dufayard JF, Guindon S, Lefort V, Lescot M, et al. Phylogeny.fr: robust phylogenetic analysis for the non-specialist. Nucleic Acids Res. 2008;36:W465–9.

158. Wang X, Fang X, Yang P, Jiang X, Jiang F, Zhao D, Li B, Cui F, Wei J, Ma C, et al. The locust genome provides insight into swarm formation and long-distance flight. Nat Commun. 2014;5:2957.
159. The International Silkworm Genome Consortium. The genome of a lepidopteran model insect, the silkworm *Bombyx mori*. Insect Biochem Mol Biol. 2008;38:1036–45.
160. Elsik CG, Worley KC, Bennett AK, Beye M, Camara F, Childers CP, de Graaf DC, Debyser G, Deng J, Devreese B, et al. Finding the missing honey bee genes: lessons learned from a genome upgrade. BMC Genomics. 2014;15:86.
161. Honeybee Genome Sequencing Consortium. Insights into social insects from the genome of the honeybee Apis mellifera. Nature. 2006;443:931–49.
162. Richards S, Gibbs RA, Weinstock GM, Brown SJ, Denell R, Beeman RW, Gibbs R, Bucher G, Friedrich M, Grimmelikhuijzen CJ, et al. The genome of the model beetle and pest *Tribolium castaneum*. Nature. 2008;452:949–55.
163. Chen XG, Jiang X, Gu J, Xu M, Wu Y, Deng Y, Zhang C, Bonizzoni M, Dermauw W, Vontas J, et al. Genome sequence of the Asian tiger mosquito, Aedes albopictus, reveals insights into its biology, genetics, and evolution. Proc Natl Acad Sci U S A. 2015;112:E5907–15.
164. Armisen D, Rajakumar R, Friedrich M, Benoit JB, Robertson HM, Panfilio KA, Ahn S-J, Poelchau MF, Chao H, Dinh H, et al: The genome of the water strider *Gerris buenoi* reveals expansions of gene repertoires associated with adaptations to life on the water. BMC Genomics in press:acceptance e-mail 12 Oct. 2018.
165. Papanicolaou A, Schetelig MF, Arensburger P, Atkinson PW, Benoit JB, Bourtzis K, Castañera P, Cavanaugh JP, Chao H, Childers C, et al. The whole genome sequence of the Mediterranean fruit fly, *Ceratitis capitata* (Wiedemann), reveals insights into the biology and adaptive evolution of a highly invasive pest species. Genome Biol. 2016;17:192.
166. Ellis LL, Huang W, Quinn AM, Ahuja A, Alfrejd B, Gomez FE, Hjelmen CE, Moore KL, Mackay TF, Johnston JS, Tarone AM. Intrapopulation genome size variation in *D. melanogaster* reflects life history variation and plasticity. PLoS Genet. 2014;10:e1004522.
167. Kumar S, Stecher G, Tamura K. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol Biol Evol. 2016;33:1870–4.
168. Croset V, Rytz R, Cummins SF, Budd A, Brawand D, Kaessmann H, Gibson TJ, Benton R. Ancient protostome origin of chemosensory ionotropic glutamate receptors and the evolution of insect taste and olfaction. PLoS Genet. 2010;6:e1001064.

Milkweed Bug (*Oncopeltus fasciatus*)

Genome Consortium

[Logo by Chiaki Ueda]

Supplementary Information

Table of Contents

1. Genome and transcriptome sequencing and assembly

Contributors: Stephen Richards, Daniel S.T. Hughes, Shwetha C. Murali, Jiaxin Qu, Shannon Dugan, Sandra L. Lee, Hsu Chao, Huyen Dinh, Yi Han, HarshaVardhan Doddapaneni, Kim C. Worley, Donna M. Muzny, Richard A. Gibbs, Kristen A. Panfilio, Stefan Koelzer

The milkweed bug *Oncopeltus fasciatus* is one of thirty arthropod species sequenced as a part of a pilot project for the i5K arthropod genomes project at Baylor College of Medicine Human Genome Sequencing Center. For all of these species, an enhanced Illumina-ALLPATHS-LG (v. 35218) sequencing and assembly strategy enabled multiple species to be approached in parallel at reduced costs. For most species in the pilot, including *O. fasciatus*, we sequenced four libraries of nominal insert sizes 180 bp, 500 bp, 3 kb and 8 kb. The amount of sequence generated from each of these libraries is noted in Table S 1.1, with NCBI SRA accessions.

1.1 Source materials, DNA and RNA purification

Genomic DNA was extracted from an individual adult male to construct the main sequencing libraries: 180-bp, 500-bp paired end and 3-kb mate pair libraries. A fourth, larger mate pair library with 8-10 kb inserts was constructed with DNA extracted from an individual adult female, due to the higher amount of starting DNA required for this library. Additionally, to aid in genome assembly and gene prediction, RNA was extracted from three samples representing three different life history samples: an individual adult male, an individual adult female, and pooled, mixed-instar nymphs.

1.2 Library preparation

To prepare the 180-bp and 500-bp libraries, we used a gel-cut paired end library protocol. Briefly, 1 µg of the DNA was sheared using a Covaris S-2 system (Covaris, Inc. Woburn, MA) using the 180-bp or 500-bp program. Sheared DNA fragments were purified with Agencourt AMPure XP beads, end-repaired, dA-tailed, and ligated to Illumina universal adapters. After adapter ligation, DNA fragments were further size selected by agarose gel and PCR amplified for 6 to 8 cycles using Illumina P1 and Index primer pair and Phusion® High-Fidelity PCR Master Mix (New England Biolabs). The final library was purified using Agencourt AMPure XP beads and quality assessed by Agilent Bioanalyzer 2100 (DNA 7500 kit) determining library quantity and fragment size distribution before sequencing.

The long mate pair libraries with 3-kb or 8-kb insert sizes were constructed according to the manufacturer's protocol (Mate Pair Library v2 Sample Preparation Guide art # 15001464 Rev. A PILOT RELEASE). Briefly, 5 µg (for 2 and 3-kb gap size library) or 10 µg (8-10 kb gap size library) of genomic DNA was sheared to desired size fragments by Hydroshear (Digilab, Marlborough, MA), then end repaired and biotinylated. Fragment sizes between 3-3.7 kb (3 kb) or 8-10 kb (8 kb) were purified from 1% low melting agarose gel and then circularized by blunt-end ligation. These size selected circular DNA fragments were then sheared to 400 bp (Covaris S-2), purified using Dynabeads M-280 Streptavidin Magnetic Beads, end-repaired, dA-tailed, and ligated to Illumina PE sequencing adapters. DNA fragments with adapter molecules on both ends were amplified for 12 to 15 cycles with Illumina P1 and Index primers. Amplified DNA fragments were purified with Agencourt AMPure XP beads. Quantification and size distribution of the final library was determined before sequencing as described above.

1.3 Sequencing

Sequencing was performed on Illumina HiSeq2000s generating 100-bp paired end reads. Reads were pre-processed using cutadapt for adapter removal and sickle-trim for quality trimming (- min length 20bp). Subsequently, reads were assembled using ALLPATHS-LG (v35218) [1] on a large memory computer with 1 TB of RAM and further scaffolded and gap-filled using in-house tools Atlas-Link (v.1.0) and Atlas gap-fill (v.2.2) (https://www.hgsc.bcm.edu/software/). This yielded an assembly of 1.099 Gb (774 Mb without gaps within scaffolds), with a contig N50 of 4.0 kb and scaffold N50 of 340 kb, which has been deposited in GenBank (assembly accession GCA_000696205.1).

Table S 1.1: Sequencing, assembly, annotation statistics and accession numbers in Excel Supplement

2. Genome characteristics, quality control, expression analyses

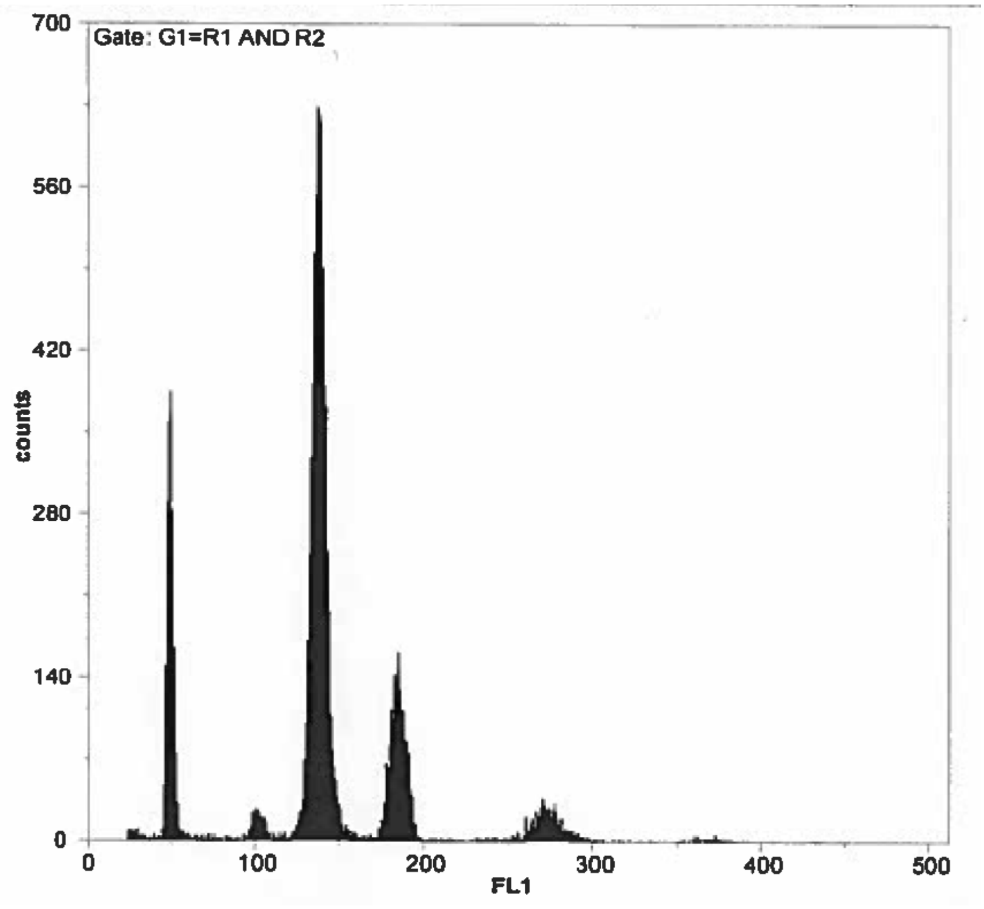
2.1 Genome size

2.1.a Flow cytometry estimation

Contributors: Patricia J. Moore and J. Spencer Johnston

Genome size estimations for the *Oncopeltus* genome were obtained by flow cytometry. Four to five females and males each from the Carolina Biological Supply lab strain and a wild strain (collected from Athens, Georgia, USA; GPS coordinates: 33° 56' 52.8216" N, 83° 22' 38.3484'' W) were measured. The samples were prepared after Hare and Johnston [2]. In short, the head of a single individual was placed into 1ml of cold Galbraith buffer in a 2-ml Kontes Dounce, along with the head of a *Drosophila virilis* female (1C = 328 Mb) and a *Callosobruchus maculatus* male (1C = 1175.6 Mb) that were added as co-prepared internal standards. Nuclei from the sample and standards were isolated with 15 strokes of the loose (B) pestle, then filtered through a 40µm filter, and stained with propidium iodide at 25 mg/ml, rather than 50 mg/ml. Following at least 30 minutes of staining in the cold and dark, the amount of fluorescence of the nuclei from the sample and standard was scored with a Partec CyFlo Flow cytometer, with excitation at 532 nm provided by a Cobalt Samba laser (Solina Sweden) and PI fluorescence detected after passing a 590nm long pass filter. The 1C amount of DNA was determined as the ratio of the mean 2C peak channel number of the sample peak divided by the 2C mean sample channel number of each standard times the amount of DNA in the *C. maculatus* standard (Figure S 2.1). At least 2000 nuclei were scored under each peak. The CV was 3.0 or less for all peaks. The *D. virilis* standard was used to verify the estimated genome size of the *C. maculatus* standard. Both standards gave the genome size estimates shown, but the standard error (shown) was lower using estimates based on the *C. maculatus* standard.

The genome size of the lab and wild strains are not significantly different. However, the male genome is very slightly larger than the female (Table S 2.1). The larger genome size estimate for the male is consistent with a large neo X/Y in the male. Oncopeltus has 2n = 16 chromosomes and an XX/XY type of sex determination; the Y is largely heterochromatic and transcriptionally inactive, pairing only briefly with the X in meiosis [3].



Relative Red (PI) Fluorescence

Figure S 2.1: Peaks from Flow cytometry measurement in *Oncopeltus fasciatus*. The first peak is the *D. virilis* 2C. The small second peak is the *D. virilis* 4C, the large third peak is the 2C of *Oncopeltus*, the 4th peak is the *C. maculatus* 2C peak, and the two smaller right-most peaks are the 4C of *Oncopeltus* and the 4C of *C. maculatus*, respectively.

Table S 2.1: Flow cytometry estimations of genome size. Four or five individuals were scored for each strain. The 1C (haploid) mean genome size (in Mb) is given for each strain along with the standard error based on the 4 or 5 individual estimates.

Sample/strain	**Individuals**	**Flow cytometry estimation**
O. fasciatus Lab F	5	922.8 +/- 4.8 Mb
O. fasciatus Lab M	4	927.9+/- 3.7 Mb
O. fasciatus Wild F	5	923.8+/- 5.9 Mb
O. fasciatus Wild M	4	928.8+/- 4.5 Mb

2.1.b k-mer estimation

Contributors: Iris M. Vargas Jentzsch and Kristen A. Panfilio

With the aim of estimating genome size, heterozygosity, and repeat content from unassembled sequencing reads, we tried several approaches to characterize the *k*-mer frequency spectrum in the *Oncopeltus* genomic dataset. As starting point we had four 100-bp Illumina read libraries: two mate-pair and two paired-end (see Table S 2.2). All libraries were filtered for quality, adapters were removed, and only correctly paired reads were retained (with the program Trimmomatic v0.30, parameters: `trimmomatic-0.30.jar PE -phred33 ILLUMINACLIP:Trimmomatic-0.30/adapters/TruSeq3-PE.fa:2:30:15:8:true LEADING:3 TRAILING:20 SLIDINGWINDOW:4:15 MINLEN:36`).

Table S 2.2: i5k Illumina library sizes after filtering.

Library	Number of reads	Expected depth of coverage
Genomic_180bp	397,866,178	36x
Genomic_500bp	148,709,592	13x
Genomic_3Kb	333,805,426	30x
Genomic_8-10Kb	189,085,040	17x

We used the programs Jellyfish2.1.4 [4] and bbmap [5], to perform *k*-mer counts on each sequencing library separately, for a range of *k* between 15 and 35. The program bbmap was initially used only to confirm the Jellyfish results, but proved more efficient by generating the same results, and even extending these over a higher range of *k*-mer depths (up to 100,000 with bbmap vs. up to 10,000 with Jellyfish). Counts for $k > 30$ could only be completed with bbmap.

K-mer frequency spectra in *Oncopeltus*

To generate the *k*-mer frequency spectrum, the frequency of occurrence of *k*-mers in the dataset (also called depth or multiplicity) is plotted against the observed counts for each of the frequencies. (Note that there is no consensus in the naming of axes for the *k*-mer frequency plots, and because we are actually plotting the frequency of a frequency, either axis can be labeled as ‘frequency’.) In an ideal dataset, with no sequencing errors and where all parts of the genome are represented equally, we expect a curve with one or more clear peaks, with the highest peak corresponding to the homozygous non-repetitive fraction of the genome. This peak would ideally be centered at the expected depth of coverage of the respective dataset (or more precisely at expected coverage - 1/read length*k^{-1}), because unique regions of the genome should be sampled on average as many times as the sequencing coverage [see 6, 7, 8]. In practice, *k*-mers from repetitive regions of the genome and *k*-mers containing sequencing errors can shift or change the shape of this curve.

The *k*-mer spectra obtained for our datasets did not resemble theoretical expectations [like in 9] in that they had very shallow or non-existent peaks. Excluding *k*-mers with counts <5, which represent mostly erroneous *k*-mers [10], allowed the visualization of small peaks for all but the 500-bp dataset. All *k*-mer spectra distributions were unimodal, with a long, slowly decreasing tail towards higher *k*-mer frequencies. Among the four datasets, the most prominent peaks were observed in the mate-pair libraries (Figure S 2.2). The 500-bp dataset showed an almost monotonic exponential decay for all measured *k*-mer values, probably due to its low coverage (13%). However, the shape of the 500-bp dataset did not improve when combined with counts from the 180-bp dataset.

For each dataset, the position of the peak (frequency or depth at which the peak is centered) varied depending on the size of the *k*-mer counted: in all cases increasing *k*-mer lengths produced a shift of the peak towards lower *k*-mer frequencies (Figure S 2.3 shows the progression for the 180-bp library). This kind of shift can be due to poor data quality or bias in sequencing coverage [9, 10]. In the case of sequencing errors, these become magnified by a factor of *k* because there will be *k* erroneous *k*-mers per error, and erroneous *k*-mers are expected to have very low frequencies [11]. This magnification will shift total *k*-mer counts towards very low frequencies and reduce the difference between this noisy region and the remaining curve. The second factor, sequencing coverage bias, is typical in amplification-based sequencing approaches due to variation in amplification efficiency across the genome. This affects the probabilities of sampling *k*-mers in the same genomic frequency class,

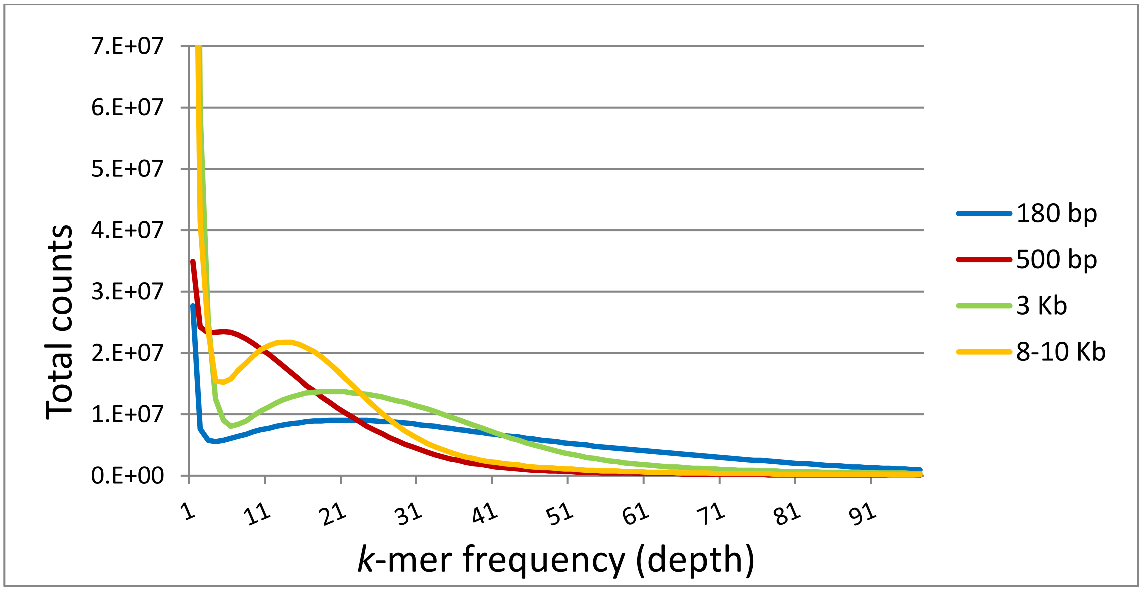


Figure S 2.2: Overlap of 17-mer spectra for the four Illumina libraries. The 8-10Kb mate pair library had the peak with the highest volume, but centered at a depth of 14, followed by the 3Kb peak at a depth of 20 and the 180bp peak centered at a depth of 23. The 500 bp library has a very shallow peak. All the peak depths were lower than the expected depth of coverage for the respective libraries (see Table S 2.2).

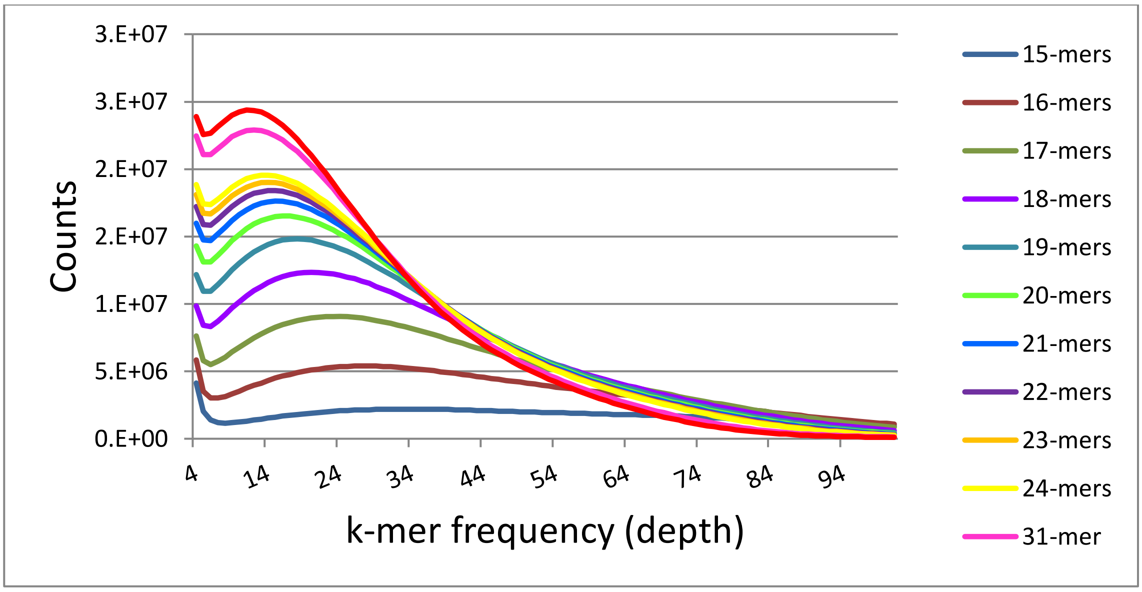


Figure S 2.3: *k*-mer spectra for the 180 bp library and variable values of k. The display is limited to depth values from 4 to 100, to better visualize the ‘peaks’. For increasing *k*-mer sizes, the peak becomes more pronounced and shifted towards lower *k*-mer frequencies, whereby the inflexion in the curve separating erroneous *k*-mers from the rest of the distribution occurs at higher counts. This means that, the larger the *k*-mer size, the less distinction there is between correct and erroneous *k*-mers.

flattening the *k*-mer frequency curve and making it more difficult to observe a peak [10]. Our datasets had very good quality as inferred from Illumina quality scores, and erroneous *k*-mers are usually dealt with by excluding low frequency *k*-mers from the calculations. On the other hand, coverage bias is more difficult to deal with and requires complex modeling [see 10], which we did not do here. Unimodal distributions with peaks shifted towards lower depths were suggested to be characteristic for most animals except mammals and birds [12, 13], but these previous analyses were limited to *Drosophila* and *Apis mellifera* for the insects.

The presence of only a single peak in all our *k*-mer spectra impeded the estimation of heterozygosity. Expectations from ideal simulated data for a diploid genome show that the *k*-mers corresponding to the heterozygous part of the genome would form a second peak at about half the expected depth of coverage of the given dataset. The ratio between these two peaks depends on the heterozygosity of the genome, therefore allowing its estimation [see 10]. For *Oncopeltus* we also expected a bimodal distribution because the individual sequenced was not inbred. However, several other arthropods sequenced as part of the i5K pilot project showed two distinct peaks in their *k*-mer spectra (S.C. Murali and S. Richards, manuscript in preparation), suggesting that the relative flatness of our *k*-mer spectra may indeed be due to specific characteristics of the *Oncopeltus* genome.

Genome size estimation

We estimated genome size from all obtained *k*-mer spectra using two formulas. One is based simply on *k*-mer numbers: total *k*-mer number/coverage at the peak (formula 1) [14], while the second uses nucleotide counts from the estimate-genome-size.pl script by J. Ryan (formula 2) (https://github.com/josephryan/estimate_genome_size.pl):

$$G = T / [(M * L)/(L - k + 1)]$$

T = total nucleotides

M = depth of coverage at the peak

K = *k*-mer size

L = mean read length

For all estimations we excluded the low depth *k*-mers forming the left side sharp peak in the *k*-mer spectra curves; the threshold was set at the lowest turning point of each curve, which varied between depth values of 4 and 6. The estimates from our calculations (Table S 2.3) were in most cases too high compared to the genome size estimated by flow cytometry of 920 to 930 Mb (see Section 2.1.a). The best approximation was obtained from the 15-mer spectra, because this curve had the peak at the highest depth of coverage. This depth of coverage (or *k*-mer frequency) at the peak was the main factor affecting the estimations; it became smaller for larger values of *k*, resulting in an overestimation of genome size. As mentioned above, it is probably due to sequencing coverage bias that all our sequencing libraries had the peak of the *k*-mer spectra centered at much lower depth of coverage values than the library sequencing depth. Consequently, about one third of the genomic *k*-mers were indistinguishable from erroneous low frequency *k*-mers, precluding more accurate estimations from the *k*-mer frequency spectra.

Table S 2.3: Estimates of genome size for various *k*-mer sizes from the 180-bp library.

k	Number of erroneous *k-mers* (Depth ≤5)	Total *k*-mers	Peak	Genome size estimate from formula 1 [Mb]	Genome size estimate from formula 2 [Mb]
15	96,179,874	31,916,279,278	32	994	1,024
16	190,554,545	28,967,316,992	26	1,107	1,261
17	282,378,853	28,797,962,402	23	1,240	1,425
18	360,059,200	28,595,269,396	19	1,486	1,725
19	423,188,873	28,496,383,832	17	1,651	1,928
20	473,437,832	28,138,862,469	15	1,844	2,185
21	512,152,262	27,895,835,094	15	1,826	2,185
22	541,893,371	27,644,584,878	14	1,936	2,342
23	565,679,336	27,385,616,052	13	2,063	2,522
24	586,102,741	27,119,171,927	13	2,041	2,522

2.2 Lateral gene transfer events and bacterial contamination

Contributors: Rose Richter, Amanda Dolan, Kristen A. Panfilio, Stefan Koelzer, John H. Werren

LGT candidates

For *Oncopeltus fasciatus*, 20 lateral gene transfer (LGT) candidates from bacteria were predicted using computational methods and assessed by subsequent manual annotation. These have a blastn similarity score to prokaryotes <1e-10 or a bitscore >75. We then assessed these LGT candidates, by several criteria, including information on the possible prokaryotic source of the LGT, gene annotation, and expression support (Table S 2.4). Of these, 19 have at least one flanking gene on the scaffold, and five were empirically validated by PCR amplification from *Oncopeltus* gDNA (see Methods, below). An additional 61 potential LGTs were deemed weak candidates, based on the presence of eukaryotic similarity or lack of sufficiently strong bacterial similarity, and are not addressed in this report.

Table S 2.4: Information on candidate LGTs in *O. fasciatus* (in Excel supplement file).

Noteworthy among the LGTs are two genes that are involved in peptidoglycan biosynthesis in bacteria. Peptidoglycan (also known as murein) is an important constituent of bacterial cell walls, particularly in gram-positive bacteria. The LGT genes *alanine racemase* and *UDP-N-acetylenolpyruvoylglucosamine reductase* (*murB*) appear to have been transferred from *Vitreoscilla* and *Chlamydia* bacteria, respectively, have expression support, and were empirically validated in *Oncopeltus* (Table S 2.4). Alanine racemase is an enzyme that catalyzes a structural conversion from L-alanine to D-alanine, with D-alanine being used in murein biosynthesis in bacteria. Although alanine racemase has been found in some marine invertebrates [15], the LGT found in *Oncopeltus* is phylogenetically embedded among *alanine racemase* genes from bacteria (based both on nucleotide and protein analyses: Figure S 2.4) and was not detected in any other invertebrates based on blastn and blastp searches. Similarly, murB shows a strong signature of lateral gene transfer from *Chlamydia* (Figure S 2.5). The potential functions of these genes in *Oncopeltus* are unknown, but worthy of further investigation. Possibilities may include defense against bacteria or enhanced ability to digest bacteria for nutrition.

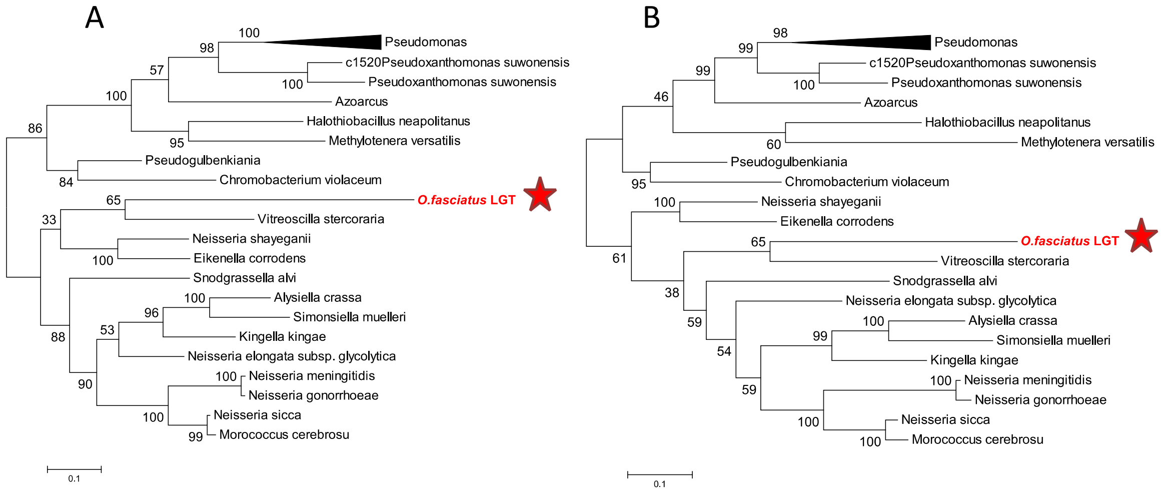


Figure S 2.4: Maximum likelihood phylogenies of the alanine racemase LGT using **(A)** amino acid and **(B)** nucleotide sequences of a 1053-bp region, showing the *Oncopeltus* sequence (red stars) embedded among bacterial orthologs. Sequences with strong similarity to the racemase LGT in NCBI's nr/nt database were aligned in MEGA 5.1 using ClustalW. Sequences were trimmed to 1053 bp, and the alignment was manually curated for accuracy. Sequences included top hits and strong matches to the *Oncopeltus* sequence. The Jones-Taylor-Thornton model was used, with 500 bootstrap replications to create maximum likelihood trees.

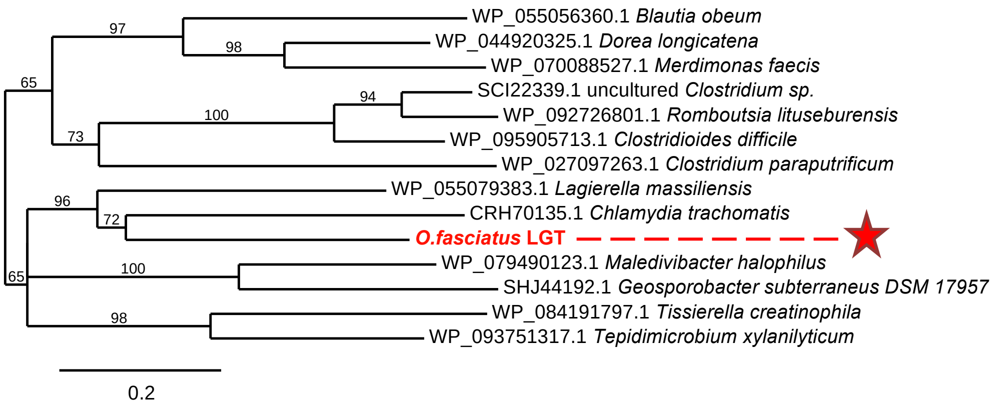

WP_055056360.1 Blautia obeum
WP_044920325.1 Dorea longicatena
WP_070088527.1 Merdimonas faecis
SCI22339.1 uncultured Clostridium sp.
WP_092726801.1 Romboutsia lituseburensis
WP_095905713.1 Clostridioides difficile
WP_027097263.1 Clostridium paraputrificum
WP_055079383.1 Lagierella massiliensis
CRH70135.1 Chlamydia trachomatis
O.fasciatus LGT
WP_079490123.1 Maledivibacter halophilus
SHJ44192.1 Geosporobacter subterraneus DSM 17957
WP_084191797.1 Tissierella creatinophila
WP_093751317.1 Tepidimicrobium xylanilyticum
0.2

Figure S 2.5: Maximum likelihood phylogeny of the UDP-N-acetylenolpyruvoylglucosamine reductase LGT, using amino acid sequences of the top blastp hits in NCBI's nr database, without taxonomic restriction, showing the *Oncopeltus* sequence (red star, 307 aa ORF) embedded among bacterial orthologs.

Whereas the two LGTs with potential roles in cell wall synthesis appear to be unique to *Oncopeltus* among the insects, an older LGT event led to the introduction of the cell wall degradation enzyme endo-1,4-beta-mannosidase [16] in the common ancestor of *Oncopeltus* and the stink bug *Halyomorpha halys*, a fellow member of the hemipteran infraorder Pentatomomorpha (Figure S 2.6, and see main text).

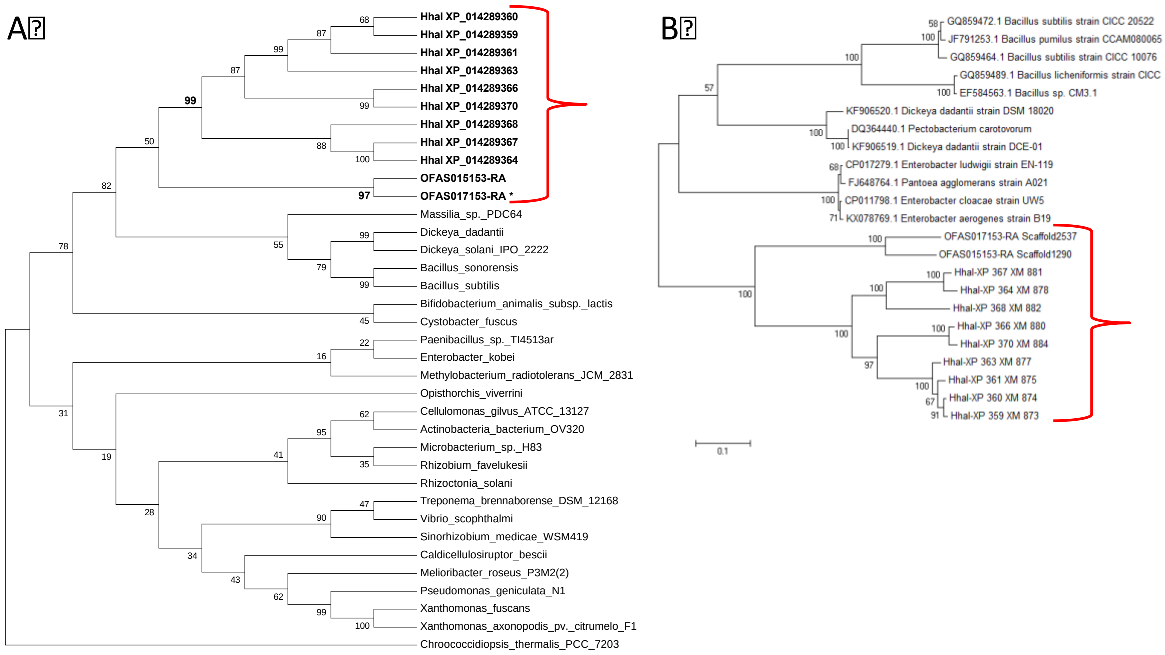

A
68
87
99
87
99
99
50
88
100
82
97
78
55
99
79
99
45
22
16
31
62
95
41
35
19
47
90
28
34
43
62
99
100
Hhal XP_014289360
Hhal XP_014289359
Hhal XP_014289361
Hhal XP_014289363
Hhal XP_014289366
Hhal XP_014289370
Hhal XP_014289368
Hhal XP_014289367
Hhal XP_014289364
OFAS015153-RA
OFAS017153-RA *
Massilia_sp._PDC64
Dickeya_dadantii
Dickeya_solani_IPO_2222
Bacillus_sonorensis
Bacillus_subtilis
Bifidobacterium_animalis_subsp._lactis
Cystobacter_fuscus
Paenibacillus_sp._TI4513ar
Enterobacter_kobei
Methylobacterium_radiotolerans_JCM_2831
Opisthorchis_viverrini
Cellulomonas_gilvus_ATCC_13127
Actinobacteria_bacterium_OV320
Microbacterium_sp._H83
Rhizobium_favelukesii
Rhizoctonia_solani
Treponema_brennaborense_DSM_12168
Vibrio_scophthalmi
Sinorhizobium_medicae_WSM419
Caldicellulosiruptor_bescii
Melioribacter_roseus_P3M2(2)
Pseudomonas_geniculata_N1
Xanthomonas_fuscans
Xanthomonas_axonopodis_pv._citrumelo_F1
Chroococcidiopsis_thermalis_PCC_7203
B
58
100
100
57
100
100
100
68
100
71
100
100
100
100
100
100
97
100
67
91
GQ859472.1 Bacillus subtilis strain CICC 20522
JF791253.1 Bacillus pumilus strain CCAM080065
GQ859464.1 Bacillus subtilis strain CICC 10076
GQ859489.1 Bacillus licheniformis strain CICC
EF584563.1 Bacillus sp. CM3.1
KF906520.1 Dickeya dadantii strain DSM 18020
DQ364440.1 Pectobacterium carotovorum
KF906519.1 Dickeya dadantii strain DCE-01
CP017279.1 Enterobacter ludwigii strain EN-119
FJ648764.1 Pantoea agglomerans strain A021
CP011798.1 Enterobacter cloacae strain UW5
KX078769.1 Enterobacter aerogenes strain B19
OFAS017153-RA Scaffold2537
OFAS015153-RA Scaffold1290
Hhal-XP 367 XM 881
Hhal-XP 364 XM 878
Hhal-XP 368 XM 882
Hhal-XP 366 XM 880
Hhal-XP 370 XM 884
Hhal-XP 363 XM 877
Hhal-XP 361 XM 875
Hhal-XP 360 XM 874
Hhal-XP 359 XM 873
0.1

Figure S 2.6: Maximum likelihood phylogenetic analyses of the endo-mannosidase LGT using **(A)** amino acid and **(B)** nucleotide sequences of a 1037-bp region, with the *Oncopeltus* and *Halyomorpha* sequences forming a well-supported clade with species-specific expansions (red brackets) among bacterial orthologs. The protein data **(A)** are shown in a bootstrap consensus tree. See main text for further details.

Finally, although not detected by our DNA-based LGT pipeline, we also found in *Oncopeltus* the ancient lysozyme glycoside hydrolase (GH25) LGT of *Wolbachia* origin, which is present in other hemipterans [17]. GH25 (also known as muramidase) is implicated in antimicrobial defense, and encodes an enzyme that also breaks down the bacterial cell wall. Therefore, we have detected various LGTs in *Oncopeltus* that are implicated in bacterial or plant cell-wall metabolism, and range in age from relatively recent (*Oncopeltus* specific), to presence in a subset of Hemiptera, to an LGT of apparently ancient origin in the Hemiptera.

Potential bacterial scaffolds

In addition to LGT candidates, there were five candidate “contaminating” bacterial scaffolds identified. Typically we find scaffolds that are probably from bacteria and these most likely are from bacterial associates of the insect. Thus, they can be informative. The identified scaffolds have homology to *Sphingomonas sanxanigenens* (Scaffold 13019), *Acinetobacter sp.* (Scaffold 15841), *Methylobacterium oryzae* (Scaffold 16784), *Paracoccus marcusii* (Scaffold 17186), and *Comamonas testosterone* (Scaffold 15798). The contaminating scaffolds are small (919 to 3841 bp), with stretches of only 0.9 to 1.4 Kb each with strong blast similarity to the bacterial sequence. Furthermore, these scaffolds do not represent entire bacterial genomes, and each has highest homology to a different bacterial species. A caveat is that we cannot always rule out that these “bacterial” scaffolds are actually large LGTs. Future comparison of the depth of coverage on these scaffolds to depth of coverage in genomic scaffolds (single copy regions, not TEs) could clarify this, as bacterial contaminants are typically at different density from nuclear genes (sometimes much higher, sometimes much lower). Note that the overall low number of such “contaminating” scaffolds is consistent with the method of template preparation for sequencing, as DNA was prepared from dissected adults from which gut material was removed. This also further strengthens the likely validity of the strong LGT predictions.

Methods

Two different computational pipeline scripts were used to identify LGT candidates and contaminating bacterial scaffolds in the genome assembly.

Information on the “old” computational pipeline

The scaffold assemblies (Genome File Name: Ofas.contaminationfree.scaffolds, downloaded from the i5k FTP site on 6-10-15) were run through this pipeline [18], which compares sequences for matches to ~1000 different bacterial genomes and compares the blast matches to a set of up to 9 reference eukaryotic genomes. For *O. fasciatus*, the animal database contained transcripts from the following animal genera: *Anopheles, Drosophila, Xenopus, Tribolium, Daphnia, Mus,* and *Homo sapiens.*

Scaffold outputs were then sorted into potential LGTs, contaminating bacterial scaffolds or likely conserved scaffolds. Sorting was based on length of the scaffold, length of bacterial matches across the scaffold, and difference score in bacterial blast match relative to eukaryotic match [18]. This pipeline focuses on the best blastn bacterial hit (highest e-value) on each scaffold, and therefore additional regions with bacterial similarity were found manually using NCBI blastn.

Information on the "new" computational pipeline

Since the old pipeline sorted scaffolds based on only the top prokaryotic hit information, there was concern that LGT candidates were being missed if they were not the best hit. The "new" computational pipeline breaks long scaffolds into 1000-bp intervals and searches each of them against the bacterial database. Any positive hits of the 1000-bp regions were then searched against the animal database. The bacterial database contained about 1000 bacterial species and was masked for low complexity regions using the NCBI Dustmasker function. The animal database for *O. fasciatus* contained transcripts from a representative from each of the following animal genera: *Anopheles, Drosophila, Xenopus, Tribolium, Daphnia, Strongylocentrotus, Mus, Homo sapiens, Aplysia, Caenorhabditis, Hydra, Monosiga,* and *Acanthamoeba.* The significance e-value cut-off used was <1e-5 for both the animal and bacterial hits. Regions of bacterial similarity that fell from the end of one 1000-bp interval to the adjacent interval were joined. Only the putative LGT regions ≥100 bp and without any hits to the animal database were used in the final analysis.

Manual annotation

The candidate LGT outputs were then manually curated by the following basic steps. Potential LGT regions on the scaffolds were searched via blastn (NCBI) in the nr/nt database. If this indicated that the region is simply a conserved gene in insects and other metazoan organisms, it was noted as conserved and disregarded. If not, the region was searched via blastx to the nr/nt database to determine if it is a conserved gene or remains an LGT candidate. For LGT candidates, flanking genes were analyzed by both blastn/blastx of the flanking sequence and/or by observations of gene models and transcription models on the species-specific Apollo web browser (https://apollo.nal.usda.gov/oncfas/selectTrack.jsp). Outputs meeting the criteria of a potential LGT were labeled as an LGT candidate in the annotation comments.

After the OGS had been generated, further manual assessment was performed, including inspection of transcriptome and RNA-seq read evidence tracks in the genome browser, and by cross-referencing to 4 stage-specific libraries (embryo/maternal, nymph, adult female, adult male) if the LGT could be assigned to a given OGS gene model. Additional GenBank NCBI blastn and blastp analyses were also performed in light of finalized OGS gene models from community curation (Section 4, above).

Experimental validation of selected LGT candidates

PCR amplification from *Oncopeltus* genomic DNA (gDNA) template was used to validate five selected LGT candidates, with three biological replicates for template gDNA template. Amplification followed a published thermocycle program for LGT candidates [19] with the following modifications: annealing temperature of 54.4 ºC and a 2.5-minute extension step. Primers were designed against the genome assembly and chosen to flank the LGT region (sequences listed in Table S 2.4). The resulting amplicons (1.1-1.7 kb) were cloned, and two clones per LGT amplicon were Sanger sequenced and confirmed to fully encompass the predicted LGT sequence.

2.3 Repeat content

Contributors: Iris M. Vargas Jentzsch and Kristen A. Panfilio

The repeat content of the assembly was assessed with RepeatModeler [20] using default parameters, based on a species-specific repeat library generated *de novo* with RECON [21], RepeatScout [22], and Tandem Repeats Finder [23]. For comparative analysis, the same analysis was performed on the genome assemblies of two fellow hemipterans: the pea aphid *Acyrthosiphon pisum* (version 2: http://bipaa.genouest.org/data/public/a_pisum/assembly2_scaffolds.fasta.bz2) and the bed bug *Cimex lectularius* (ftp://ftp.hgsc.bcm.edu/I5K-pilot/Bed_bug/genome_assemblies/). As a result, only 25% of the *Oncopeltus* assembly could be identified as repetitive, while in 35% of the *Cimex* assembly and 31% of the pea aphid assembly were covered by repeats (see main text and Table S 2.5 for discussion and additional information). Comparisons between genome assemblies need to be done with caution, as differences in sequencing technologies and assembly strategies can produce spurious results. The pea aphid assembly is based on Sanger sequencing, while the bed bug and *Oncopeltus* assemblies were produced from Illumina sequence by the i5K project. Nevertheless, the bed bug assembly is much smaller than the *Oncopeltus* one (650 Mb compared to 1,099 Mb), and its assembly quality is much higher (contig N50 of 4,047 bp for *Oncopeltus* and 23,511 bp for *Cimex*). With 17,222 scaffolds and 325 million undetermined nucleotides, the current *Oncopeltus* assembly is considerably fragmented. One of the main causes of fragmentation in genome assemblies is a high proportion of repeats in the assembly [24]. Thus, the analysis of repetitive content in a highly fragmented assembly is essentially flawed.

In an attempt to improve the *Oncopeltus* assembly, we generated additional sequencing reads with the single molecule real time sequencing technique from Pacific Biosciences. This technology can produce very long reads with the potential to span gaps and repetitive sequences, overcoming biases associated with features like GC content and repeats [24, 25]. The sequencing was done at the Genome Centre of the Max Plank for Plant Breeding Research (MPIPZ Cologne) with a PacBio RS II machine. The single template library was generated from pooled gDNA from an adult virgin female and mixed-stage eggs. A total of 34 SMRT cells were sequenced generating a set of reads to an approximate coverage of 8× (Table S 2.6).

We first attempted to generate a hybrid assembly from PacBio and Illumina reads, using ALLPATHS-LG. At the time of writing this paper we had no success in generating a hybrid PacBio-Illumina assembly, due to excessive requirements of RAM by the program (>500 Gb). Nevertheless, initial ALLPATHS-LG estimates of total repetitive content placed the value at 68% with these hybrid data.

Thus, we also used the PacBio subreads to perform gap-filling on the i5k scaffolds with PBJelly Version: 13.10.22 [26] with the following blastr parameters: 'minMatch 8 -minPctIdentity 70 -bestn 1 -nCandidates 20 -maxScore -500 -nproc 4 –noSplitSubreads'. Gap-filling substantially reduced the proportion of undetermined nucleotides (assembly gaps) from 30% to 6%, with an attendant modest reduction in the number of scaffolds (17095 scaffolds, N50 = 409 Kb), while the assembly size increased from 1,099 to 1,361 Mb.

This increase in assembly size is problematic because the initial i5K genome assembly (1,099 Mb) already had a slightly larger assembly size compared to the genome size estimation by flow cytometry (925 Mb, see Section 2.1.a). On the other hand, preliminary estimations of genome size and repeat content by ALLPATHS-LG on all raw data (PacBio and Illumina) with k=25 produced estimations of genome size and repeat content of 1.51 Mb and 58%, respectively. We would expect that measurements of genome size by flow cytometry are more reliable than those obtained with bioinformatics approaches, especially because we did not obtain satisfying results from *k*-mer analyses (see Section 2.1.b). Nevertheless, published examples from other insect species show that the *k*-mer estimations also reasonably agree with laboratory measurements, both for genomes around 600 Mb in length like in the winter moth and whitefly [see 7, 8] as well as for the very large genome of the migratory locust [27]. Therefore, there is something particularly challenging in our current *Oncopeltus* dataset, and we decided to do further analyses to check if any more information could be gained by including PacBio data into our assembly. Hereafter we will call the gap-filled Illumina assembly the 'PBJelly assembly'.

For an initial quality assessment on the PBJelly assembly, we compared it to the Illumina assembly with respect to presence, completeness, and copy number for benchmarking universal single-copy orthologs of protein coding genes (BUSCO, v. 3, [28, 29]) based on expectations across the Insecta. The two assemblies are highly similar for these metrics, with slight improvements for completeness at the expense of only a minor increase in duplicates in the PBJelly assembly (Table S 2.7; see Section 6.1 for the complete BUSCO analysis on the Illumina assembly). Specifically, this analysis documents a slight improvement in reducing the proportion of fragmented genes (from 3% to 1.5%) and of missing genes (0.5% reduction), with only a marginal gain in duplications (0.5% increase). Given the starting high quality of the Illumina-only assembly for protein coding gene content, and the fact that we are endeavoring to produce a better hybrid assembly with new computational approaches (unpublished work in progress), we restricted use of the gap-filled assembly in the current study to our repeats analysis, and we retain the Illumina assembly as the current definitive/official version.

Repeat analysis with RepeatModeler on the PBJelly assembly revealed an increase in all repeat categories with the exception of satellites, which are still not represented at all, implying that no large tandem repeats could be unambiguously placed with PBJelly (Figure S 2.7 A). The repeat content within this assembly amounted to 32%, which is now closer to the repeat content detected by RepeatModeler in the pea aphid (31%) and bed bug (35%) genomes (details in Table S 2.5), but is still low compared to the repeat content reported for similarly sized genomes, like the brown plant hopper (1.2 Gbp assembly with 48.6% repeats: [30]), and notably lower than our ALLPATHS-LG estimate of 58% (Figure S 2.7 B). The absolute coverage, however, is almost double that in the *Cimex* assembly and more than 2.5 times that in the pea aphid assembly (see main text Fig. 5b). The majority of the repeats found by RepeatModeler were in the “unknown” category, probably because they were too short or incomplete to be identified unambiguously. Lastly, although our approach to identify repeats in *Oncopeltus* was a combination of *de novo* and homology-based prediction approaches, sensitivity could be enhanced by further manual curation of the *de novo* library [31], as was done for the pea aphid [32].

Table S 2.5: Comparison of repeat content between three bug species (in Excel Supplement).

Table S 2.6: Details on additional sequencing with PacBio.

Chemistry	PacBio RS II: P4-C2 and P6-C4
Number of SMRT cells	34 (2 SMRT cells with P6-C4)
Total number of reads	2,135,043
Average insert size	15,760 nt
Read length range	35 to 46,753 nt
Coverage	8x

Table S 2.7: BUSCO v3 statistics for gene count and percentage representation of the Insecta dataset (n= 1658) for the PBJelly assembly compared to the original i5K Illumina assembly.

Dataset	**Complete (single copy and duplicated)**	**Duplicated (subset of "complete")**	**Fragmented**	**Missing**
Illumina i5K assembly	1,568 (94.6%)	23 (1.4%)	50 (3.0%)	40 (2.4%)
PB Jelly assembly	1,605 (96.8%)	32 (1.9%)	22 (1.3%)	32 (1.9%)

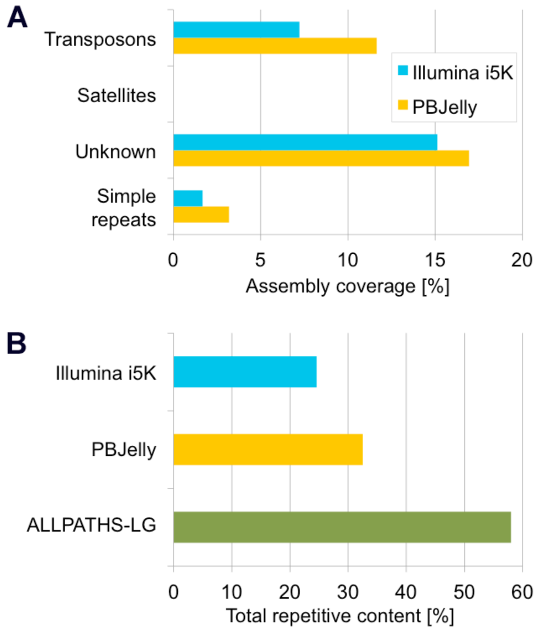


Figure S 2.7: Comparison of repeat content estimations. **(A)** Relative proportions of selected repeat categories in Oncopeltus, compared between two assembly versions. "Illumina i5K" is the current official assembly based solely on Illumina short read data. "PBJelly" is the same i5K assembly after gap filling with PacBio reads (see also Supplemental Note 2.3). Even after gap filling, it was not possible to identify large tandem repeat structures of the satellites category in these analyses with RepeatModeler. **(B)** Total repetitive content estimations for Oncopeltus based on different genome assembly versions. In addition to the assemblies used in panel (a), a third dataset, "ALLPATHS-LG", represents an estimation derived from initial attempts to generate a hybrid assembly with Illumina and PacBio reads.

2.4 Comparative transcriptomic assessments of hemipteroid reproductive biology

Contributors: Christopher J. Holmes and Joshua B. Benoit

To gain insight into stage- and sex-specific enrichment in gene expression, we compared RNA-seq datasets for *Oncopeltus* with three other hemipterans (*Cimex lectularius*, *Acyrthosiphon pisum*, *Pachypsylla venusta*) and with the thysanopteran *Frankliniella occidentalis* as a close outgroup within the hemipteroid assemblage (Tables S 2.8-2.10, Figure S 2.8).

For *Oncopeltus*, we further used our known ("i5K") RNA-seq datasets for an adult male, an adult female, and mixed-instar nymphs to assess a previously published dataset for an *Oncopeltus* adult of unspecified sex ("Andolfatto", [33]). Genes expressed in the Andolfatto set are consistent with this sample being most similar to the i5K male sample. Overall, the Andolfatto set had nearly 50% fewer differentially expressed genes in relation to the i5K male RNA-seq set when compared to either the i5K female or i5K nymph RNA-seq sets. Also, sperm-specific genes, such as serine-threonine kinases, showed a noticeably higher level of expression in the Andolfatto set even compared to the male-specific i5K set. However, vitellogenin-specific genes are also detectable in the Andolfatto set, albeit at lower levels compared to the female-specific i5K set. Thus, it appears that the Andolfatto set is mainly composed of male-specific genes, but does represent female- and nymph-specific genes as well.

Methods

Sex-specific and developmental stage-specific RNA-seq analyses were conducted according to published methods [34, 35], with modifications. In addition to the RNA-seq datasets generated in this project for *Oncopeltus* (GenBank Bioproject: PRJNA275739), RNA-seq datasets for other hemipteroid species were sequenced as parts of other i5K species projects (*Cimex lectularius*, PRJNA275741; *Acyrthosiphon pisum*, PRJNA209321; *Pachypsylla venusta*; PRJNA275248; *Frankliniella occidentalis*, PRJNA203209). These RNA-seq sets were used to examine transcriptional differences between males, females, and nymphs (if available). RNA-seq sets were evaluated with FastQC and trimmed with CLC Genomics (CLC Bio).

Reads required at least 90% similarity at 70% of the transcript length with only two mismatches. Each read was permitted to match up to five locations. Transcripts per million (TPM) was used as a proxy for transcript levels. The Baggerly's test (beta-binomial distribution statistic) followed by Bonferroni correction at 0.01 and two-fold difference between samples was used to identify genes with significant enrichment in a specific sample. This stringent statistical analysis was used because only a single replicate was available for each treatment. Genes were identified by BLASTx searches in a GenBank non-redundant protein database for arthropods (e-value ≤0.001). Following the identification of stage-specific enriched sets, these sequences were compared between life history stages in *Oncopeltus* and pairwise between *Oncopeltus* and each of the other four hemipteroid species. Overlap was visualized by Venn diagram (http://bioinformatics.psb.ugent.be/webtools/Venn/).

Table S 2.8: Information for each hemipteroid species, including paired-end read, gene, and sequence counts for individual data sets.

Species (Dataset)	**Sex/ Stage**	**Paired-End Reads**	**Genes**	**# Sequences, 2-Fold Increase**
Acyrthosiphon pisum	Male 1	22,908,878	36,939	2,940
	Male 2	18,692,759		
	Female 1	12,924,374		2,883
	Female 2	17,201,208		
Cimex lectularius	Male	34,043,709	14,214	1,619
	Female	26,526,400		1,657
Frankliniella occidentalis	Male	24,336,855	17,553	646
	Female	22,622,821		984
	Nymph	32,515,059		497
Oncopeltus fasciatus (published adult, "Andolfatto")	N/A	27,109,674	19,811	to known male: 1,929
				to known female: 3,906
				to known nymph: 3,421
Oncopeltus fasciatus (i5K)	Male	22,196,171	19,811	1,180
	Female	30,342,335		594
	Nymph	20,508,626		400
Pachypsylla venusta	Male	22,103,387	14,390	440
	Female	37,520,371		2,317
Rhodnius prolixus	N/A	N/A	15,078	N/A

Table S 2.9A-C: Hemipteran gene comparisons for adult males, adult females, and nymphs. Gene IDs associated with the Venn diagrams in Figure S 2.8D-F (in Excel supplement file).

Table S 2.10A-F: Hemipteran sex- and stage-specific comparisons. Gene IDs, fold-change, and significance for all sex- and stage-specific comparisons (in Excel supplement file).

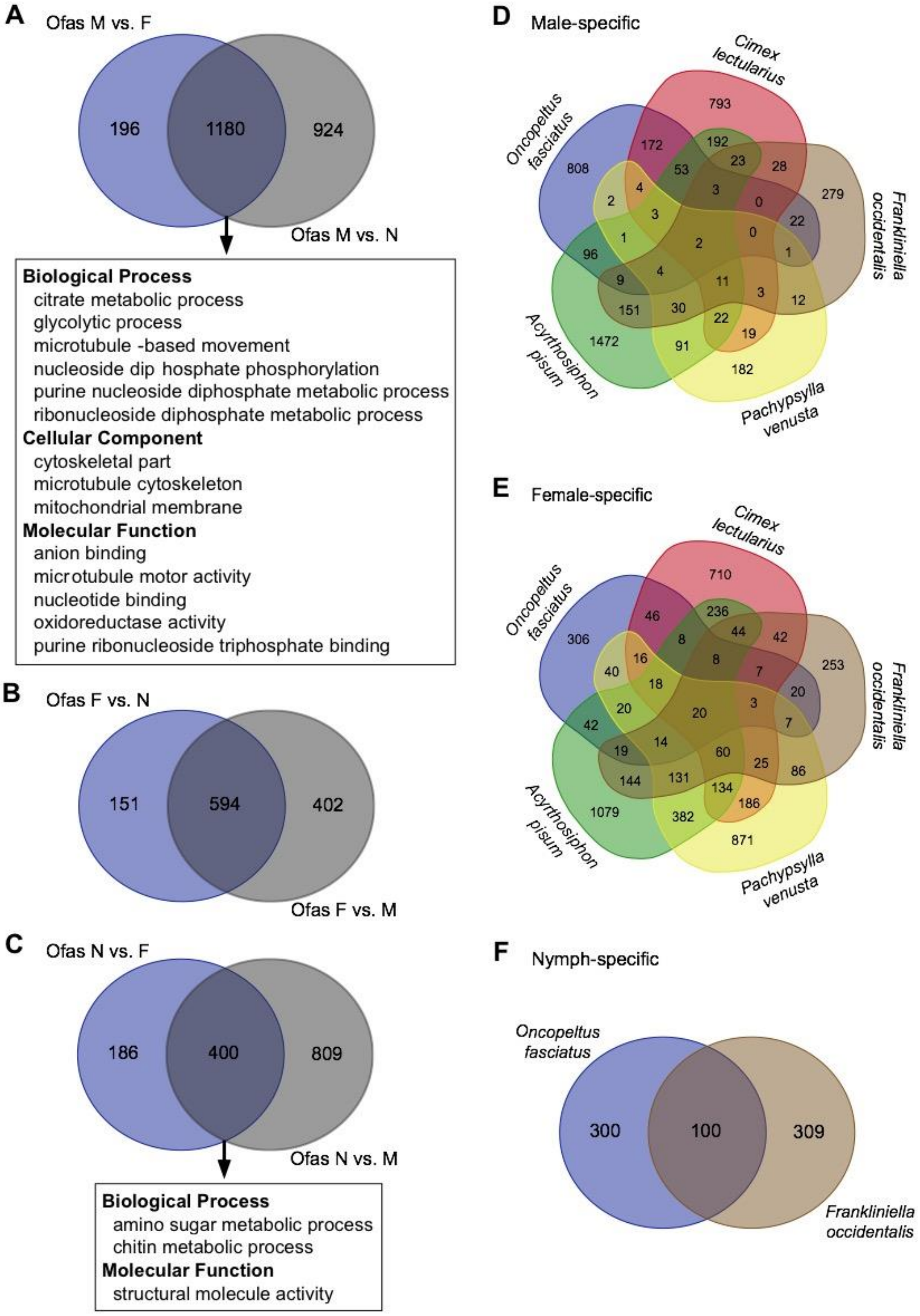


Figure S 2.8: Venn diagrams show the number of genes associated with stage-specific comparisons within *Oncopeltus* **(A-C)** and across hemipteroid species **(D-F)** for: adult males (A, D), adult females (B, E), and nymphs (C, F). Boxed terms represent GO terms that were significantly upregulated compared to the annotated genome of *Rhodnius prolixus*.

3. Automated gene annotation using a Maker 2.0 pipeline tuned for arthropods

Contributors: Dan S.T. Hughes and Stephen Richards

Oncopeltus fasciatus was among 28 i5K pilot species for which genome assemblies were subjected to automatic gene annotation using a Maker 2.0 annotation pipeline tuned specifically for arthropods. The pipeline is designed to be systematic, providing a single consistent procedure for the species in the pilot study and scalable to handle hundreds of genome assemblies, using both protein and RNA-seq evidence to guide gene models, and targeted to utilize extant information on arthropod gene sets.

The core of the pipeline was a Maker v2.28 [36] instance, modified slightly to enable efficient running on our computational resources. The genome assembly was first subjected to *de novo* repeat prediction (RepeatModeler 1.0.8 [20]) and CEGMA analysis to generate gene models for initial training of the *ab initio* gene predictors. Three rounds of training of the Augustus [37] and SNAP [38] gene predictors within Maker were used to bootstrap to a high quality training set. Input protein data included 1 million peptides from a non-redundant reduction (90% identity) of Uniprot Ecdysozoa (1.25 million peptides), supplemented with proteomes from eighteen additional species (*Strigamia maritime, Tetranychus urticae, Caenorhabditis elegans, Loa loa, Trichoplax adhaerens, Amphimedon queenslandica, Strongylocentrotus purpuratus, Nematostella vectensis, Branchiostoma floridae, Ciona intestinalis, Ciona savignyi, Homo sapiens, Mus musculus, Capitella teleta, Helobdella robusta, Crassostrea gigas, Lottia gigantean, Schistosoma mansoni*), leading to a final non-redundant evidence set of 1.03 million peptides.

RNA-seq data derived from *Oncopeltus* adult males, adult females, and mixed sex juveniles (see also Table S 1.1) were used judiciously to identify exon-intron boundaries, but with a heuristic script to identify and split erroneously joined gene models. We used CEGMA models for quality control purposes: of 1,977 CEGMA single copy ortholog gene models, 1,915 (97%) were found in the assembly and 1,837 (93%) in the final predicted gene set – a reasonable result given the small contig sizes of the assembly. Finally, the pipeline uses a nine-way homology prediction with human, *Drosophila*, and *C. elegans*, and InterPro Scan5 to allocate gene names.

The automated gene sets are available from the Baylor College of Medicine Human Genome Sequencing Center (BCM-HGSC) website:

https://www.hgsc.bcm.edu/milkweed-bug-genome-project

as well as the National Agricultural Library (NAL):

https://i5k.nal.usda.gov/Oncopeltus_fasciatus

where a web browser of the genome, annotations, and supporting annotation data is accessible. Furthermore, the genome assembly [39], Maker gene set predictions [40], and official gene set v. 1.1 ([41], see next section) are available as citable databases hosted by the NAL Ag Data Commons.

4. Community curation and generating the official gene set

Contributors: Iris M. Vargas Jentzsch, Monica F. Poelchau, Daniel S. T. Hughes, Monica C. Munoz-Torres, Christopher P. Childers, Chien-Yueh Lee, Mei-Ju Chen, Kristen A. Panfilio

The automated gene set from the MAKER pipeline, consisting of 19,615 gene models, was subjected to manual curation to assess the completeness and accuracy of gene models. This was done using the aforementioned web browser set up by the National Agricultural Library (https://i5k.nal.usda.gov/Oncopeltus_fasciatus), which allowed coordinated editing of gene models and their metadata in real time among the 16 research groups involved in the curation process. Each group curated gene families based on their own research interests and expertise, which was coordinated through a shared Google document.

The general procedure for manual curation involved retrieving homologous sequences for the genes of interest from NCBI, either from other hemipteran species (*e.g.*, *Acyrthosiphon pisum*) or from insects with high quality annotations like *Drosophila melanogaster* and *Tribolium castaneum*. These were blasted against the *Oncopeltus* assembly (generally, tblastn with a homologous protein query against the genome assembly) either in a standalone fashion or through the blast instance at the National Agricultural Library (NAL): https://i5k.nal.usda.gov/webapp/blast/. The blast results were visualized in the Apollo v.1.0.4 instance for *Oncopeltus* (https://apollo.nal.usda.gov/oncfas/selectTrack.jsp), where the corresponding automated annotation models could be inspected and edited. To confirm homology, protein translations of the edited models were blasted back into NCBI. Orthology, intron/exon boundary assessments, and protein sequence completeness were further determined by manual inspection and correction of protein alignments generated with Clustal Omega (http://www.ebi.ac.uk/Tools/msa/clustalo/).

Possible gene duplications were assessed by performing tblastn searches on the *Oncopeltus* scaffolds using the protein sequences of completed annotation models as queries, and then re-blasting the resulting hit sequences into GenBank for Arthropoda hits. Phylogenetic analysis to confirm the paralogous status of the gene models was based on trees constructed with selected protein sequences among the GenBank blast hits (*e.g.*, selection favored broad taxonomic sampling within the

Arthropoda and avoided accessions labeled as 'predicted'). Phylogenies were primarily generated at the Phylogeny.fr site (http://www.phylogeny.fr/**,** [42, 43]) with the default pipeline (MUSCLE v. 3.8.31 alignment, GBLOCKS v. 0.91b curation, PhyML v. 3.1/3.0 aLRT maximum likelihood tree construction), unless otherwise stated.

Special care was required for curating genes that were only partially predicted in the assembly. *Ab initio* gene prediction is difficult across big gaps (when parts of the gene are on different contigs), and a complete gene prediction is often not possible if parts of the same gene are on different scaffolds. Each part of a gene split across scaffolds may have its own automated model, or some parts may lack models altogether. Split gene models are to be expected in fragmented genome assemblies, because sequence gaps make gene prediction difficult [44]. Furthermore, having multiple good blast hits for the same gene on different scaffolds could indeed be due to the gene being split across multiple scaffolds, or to either gene duplication or conserved sequence within a gene family. If the automated model to be curated only represented part of the query sequence, the search was repeated with only the missing part of the query, in order to focus the blast search on relevant local alignment regions. If this query sequence was very short, turning off the low complexity filter in the blast options increased the probability of getting a hit. For genes split across multiple scaffolds, the model for each part was checked and documented for exon start and end phases at the break points, to make sure that the proper reading frame is obtained when concatenating all parts into a single model. Lastly, split models were documented as such both in the metadata and in the model name (labeled with the suffix '–part *x* of *y*'). As the gff3 format does not provide a way to specify this information other than in the comments section, the official gene set gff files contains multiple models per split gene, and the details for putting these parts together were documented separately (see Table S 4.2).

After the manual curation stage, the official gene set (OGS) was created by merging the computationally predicted gene set and the manual curated models. This was done using a 'patch' build system that uses heuristics to merge manual and automated gene predictions [45], whereby all automated models that overlapped on the same strand with manually curated models were replaced. This overlap was restricted to coding exons, and therefore it was important to take care that all reading frames were set correctly: if the reading frame of an upstream exon was incorrectly set and generated a premature stop codon, the rest of the gene model downstream would be registered as untranslated region and all intersecting automated annotations would be kept in the OGS. Indeed, several of these cases were registered in the OGS v1.1, and corrected in v1.2. The OGS v1.1 [41] was used for downstream pipeline analyses and included manual curation of 1,426 models (194 *de novo*) for 1,201 genes, including multiple isoforms and the split gene models, among a total of 19,690 gene models for 19,465 unique genes.

The manual curation process was then resumed to make further improvements based on analysis of the OGS v1.1. These revisions affected 337 models (206 revised, 122 added *de novo*, 9 deleted). Almost half of these models represent the comprehensive addition of chemoreceptor genes of the ionotropic and odorant receptor classes, including a further 96 *de novo* models that were absent from the OGS v1.1. Among the remaining revisions were 52 models (13 *de novo*, 39 updated) for various metabolic enzymes. The annotation sets were merged again to produce the OGS v1.2 (Tables S 4.1, 4.2). After each merge, the annotation sets were screened in a quality control step to ensure that curated functional annotations were present and without formatting errors. The OGS v1.2 is the version that was submitted to NCBI, and both OGS datasets are available here: https://i5k.nal.usda.gov/data/Arthropoda/oncfas-(Oncopeltus_fasciatus)/Current%20Genome%20Assembly/2.Official%20or%20Primary%20Gene%20Set/.

To quantify the main curation actions that were performed during the manual curation effort, we ran the program gff-cmp-cat (https://github.com/chienyuehlee/gff-cmp-cat) on the curated models from OGS v1.2 and the original Maker dataset. This program classifies curation actions applied to a gene model by calculating differences between two overlapping models (see Supplementary Table S5 in McKenna, *et al.* [46] for definitions of how each action is calculated). A comprehensive list of the curated gene models and their curation actions is available in Table S 4.3. Out of 1697 curated gene models, 1159 had some structural modification or were newly added, while 359 gene models were approved without editing and were given functional annotations only (Table S 4.4). Strikingly, the number of genes resulting from a merged CDS action (205) is far greater than the number of gene models resulting from split CDS actions (30). These numbers suggest that the original Maker dataset may be an overestimate of the actual number of genes from certain gene families, and that gene length was also underestimated. Indeed, the mean gene locus size of the OGS v1.1 is 12,985 bp (median: 8,819 bp), which is slightly higher than in the original automated set, because manually annotated genes were longer (mean: 20,794 bp; median: 11,326 bp). In an extreme example, the *Oncopeltus* orthologue of *hemocytin* (also known as *hemolectin*), which encodes a conserved carbohydrate binding protein of 3667 amino acids, was split across three scaffolds and predicted in ten separate automatic annotation models.

In addition, many more gene models were extended than reduced (376 vs. 215), in part due to correction of gene structure. The mean number of introns was lower in the OGS v1.1 compared to the automated predictions, because often the MAKER pipeline inserted very small introns within exons, where no RNA-seq supported junctions were evident. Indeed, the minimum intron length in the OGS v1.1 was 5 bp in an automated annotation, while the smallest intron supported by RNA-seq evidence was 71 bp, belonging to a manual annotation model for the gene *FK506-binding protein 15-like*.

These trends may be due to the fragmented nature of the assembly – genome assembly fragmentation is often associated with a higher predicted gene number and fewer exons per gene [44]. Furthermore, many gene models were newly added in the curation process (285), 70% of which are chemoreceptors (gustatory receptors, ionotropic receptors, and odorant receptors, Table S1; chemoreceptors represented only 4% of all curated genes), which are fairly small, rapidly evolving, and often lack expression support, making them difficult to predict computationally [see also. 46].

Table S 4.1: List of all curated models in the OGS v1.2.

Table S 4.2: Gene models split across scaffolds in the *Oncopeltus* genome assembly: correspondence of gene model IDs, and information on the models (in Excel supplement file).

Table S 4.3: A comprehensive list of all curated genes in OGS v1.2, including modification actions used to generate the gene model (in Excel supplement file).

Table S 4.4: A summary of curation actions by feature type in OGS v1.2, as computed by the program gff-cmp-cat. Note that the program does not handle rRNA features, so these actions were added manually. Descriptions of the curation actions and how they are calculated are presented in Table S5 of McKenna, *et al*. [46].

Action type	**gene/ pseudogene**	**mRNA**	**pseudogenic transcript**	**rRNA**
Added Models	298	316	13	
Extended Models	376	416		
Reduced Models	215	255		1
Models modified within boundary coordinates	58	62		
Models resulting from merged CDS	205	229		1
Models resulting from split CDS	30	43		
Models resulting from merged UTRs	30	34		
Models resulting from split UTRs	12	14		
All structurally modified models	1159	1307	13	2
Models with only functional annotations	359	375		
Total number of manually curated gene Models	**1518**	**1682**	**13**	**2**

5. Curation and comparative analysis of specific gene families

5.1 Developmental regulation: transcription factors and signaling pathways

One of the main reasons for choosing to sequence the *Oncopeltus* genome was due to its status as an experimental model system for developmental studies. For this reason, it was of special interest to analyze its developmental gene complement. In total, over one-third of all curated gene models (672 models, 40.4% of OGS v1.2) are of genes known to be involved in developmental processes and signaling pathways in other insects, with many of these encoding transcription factors. (Note that genes encoding proteins involved in the endocrinological control of postembryonic development and specifically in neurogenesis are addressed in separate sections below.)

5.1.a Anterior-posterior body axis: terminal patterning system and segmentation

Contributors: Peter K. Dearden, Andrew G. Cridge, Elizabeth J. Duncan, Megan Leask, Mackenzie K. Lovegrove, Olivia Tidswell, Ariel D. Chipman, Barbara M.I. Vreede, Neta Ginzburg

Early patterning genes appear mainly conserved relative to what is known from other insects, with very few instances of lineage-specific duplications in *Oncopeltus*, which are highlighted below. In cases where no ortholog was identified we remain cautious. It is already known that several key genes involved in axis formation in *Drosophila* are not conserved in other insects [47]. In other cases apparent gene absence may be due to genome assembly limitations, and readers are encouraged to also examine available transcriptomic resources (summarized in main text Fig. 2; [48]).

The terminal patterning system has been previously studied in *Oncopeltus* [49], and most of the relevant genes have been identified there. The genome analysis revealed a duplication of *Torso-like*, which encodes a perforin-like protein. *Torso-like* is also duplicated in the aphid genome, with one quite derived copy (*Torso-like related*), but the *Oncopeltus* duplications are independent of the aphid ones (based on Bayesian phylogenetics), and less derived. These copies have been named *Torso-like 1* and *Torso-like 2* to reflect a recent duplication with similar copies. The genome analysis did not recover a copy of *trunk*, which is involved in terminal patterning in *Drosophila*. However, we have recovered a copy of the closely related *PTTH*, which controls developmental timing of juvenile stages. In aphids and crustaceans, genes similar to *trunk/PTTH* are found with more similarity to vertebrate and lophotrochozoan *noggin*. A representative of this class of genes is also present in *Oncopeltus* (see below in Section 5.1.i). We have confirmed the absence of *torso*, reported to be missing in Weisbrod, *et al*. [49]. This gene encodes a receptor tyrosine kinase, of which many are present in the genome. The most closely related receptor tyrosine kinase we have found is most similar to ‘Neurospecific receptor kinase’. This is a surprising result, since in all species studied so far where there is a ligand (either PTTH or trunk) there is also a receptor similar to Torso [50]. These findings suggest that *Oncopeltus* uses PTTH to control developmental timing (as this appears to be an ancestral trait in insects), but it appears as though this pathway is not involved in terminal patterning [49, 50].

The genes known from the *Drosophila* segmentation cascade (maternal, gap, and pair-rule members) were examined in the *Oncopeltus* genome. With the exception of the cyclorrhaphan-restricted *bicoid*, homologs were found for all inspected candidate genes, including two copies each of *knirps* (previously characterized in [51]) and of *paired* class genes, and with expression support for two isoforms each of *nanos* and *Giant*. Interestingly, the gene reported as *engrailed* and used in several reports of *Oncopeltus* development [49, 52-54] turns out to be orthologous to *invected*, which encodes a diagnostic "RS-motif" from a single, small internal exon [55], and which we could validate empirically by amplification from cDNA. The actual *engrailed* ortholog occurs in a tail-to-tail orientation on the same scaffold.

Lastly, all three *odd*-like genes (*odd-skipped* (*odd*), *brother of odd with entrails limited* (*bowl*), and *sister of odd and bowl* (*sob*)) were identified in the *Oncopeltus* genome. However, simple recursive BLAST was not sufficient in identifying each ortholog individually; thus, additional care based on protein structure and conserved domains was taken in gene model curation and orthology assignments for the *odd*-like genes (Figure S 5.1).

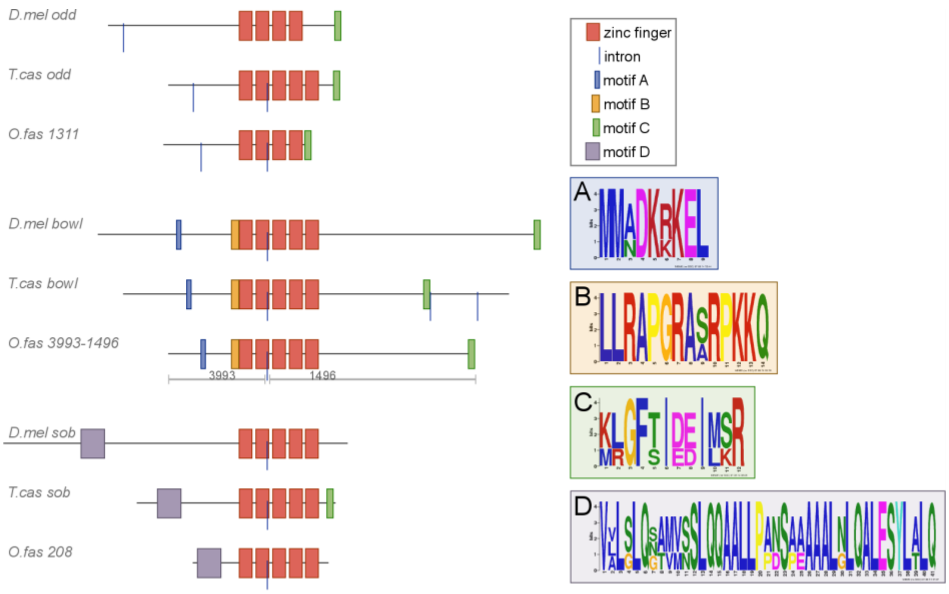


Figure S 5.1: Schematic of the coding DNA for *odd*-like genes *odd-skipped* (*odd*), *brother of odd with entrails limited* (*bowl*), and *sister of odd and bowl* (*sob*) in genomes of *Drosophila melanogaster*, *Tribolium castaneum*, and *Oncopeltus fasciatus*. *O. fasciatus* homologs were identified using a combination of intron locations, zinc finger motifs, as well as other key conserved motifs, outlined in A-D.

5.1.b Hox and other homeobox transcription factors

Contributors: Kristen A. Panfilio and Iris M. Vargas Jentzsch

Homeodomain transcription factors are a protein superfamily with diverse roles in developmental regulation, with the eponymous Hox genes representing a textbook case of metazoan-wide conservation in gene cluster organization, protein sequence, and role in tissue specification along the anterior-posterior body axis [56]. Between pipeline analyses of OGS v1.1 (see main text and Section 6.2, below) and manual curation, we have identified 96 genes encoding homeodomain transcription factors, of which 39 have been manually curated for their relevance to specific biological roles covered here and in other manual curation sections. For the Hox genes, we have found and annotated complete, single copy orthologs for all ten expected genes in *Oncopeltus* (Table S 5.1). The Hox cluster is, however, fragmented and distributed across 11 scaffolds in the current assembly: only *zen* and *Dfd* are linked on the same scaffold, while *Scr* and *Ubx* are each split across two scaffolds (Figure S 5.2 A).

The size of the complete Hox cluster, assuming a direct concatenation of these 11 scaffolds, is 4.2 Mb. This is relatively large compared to the better-assembled clusters we annotated in other i5K genomes: in both the bed bug (*Cimex lectularius*, [19]) and the Asian longhorned beetle (*Anoplophora glabripennis*, [46]) the cluster spans 3.5 Mb, assembled onto only one or two scaffolds, respectively. Both species have genome sizes comparable to *Oncopeltus* (926 Mb), from 865 Mb in the bed bug to 976 Mb in the Asian longhorned beetle. The increase in size of the *Oncopeltus* cluster is largely due to an increase in the length of intronic and intergenic regions, including gaps in the assembly. In contrast, protein sizes are marginally smaller (see Table S 5.2). This reduction in protein size was more prominent for the anterior four genes (*lab, pb, zen*, and *Dfd*). Splice sites in the central and posterior genes (*Dfd, Scr, Antp, Ubx, abd*-A and *Abd*-B) are well conserved.

For the functionally diverged Hox gene *fushi tarazu (ftz)*, a single transcript sequence was identified in a previous transcriptome [48]. Based on empirical analyses by Yong Lu (group of Leslie Pick, unpublished data), two transcript isoforms were annotated for *ftz* in the current genome assembly. However, while the presumptive homeodomain open reading frame could be found, neither isoform appears to encode a complete transcript for a functional protein, and these two isoforms may rather represent degradation to pseudogene status.

Beyond the Hox genes, we were also able to identify clear orthologs in conserved syntenic pairs for the aforementioned *engrailed* and *invected* paralogs (see Section 5.1.a) as well as the Iroquois Complex pair *iroquois* and *mirror*, which retain the conserved organization of shared transcriptional orientation from the same DNA strand (Figure S 5.2 B). Please see below (Section 5.1.c) for further details on Iro-C gene copy number evolution.

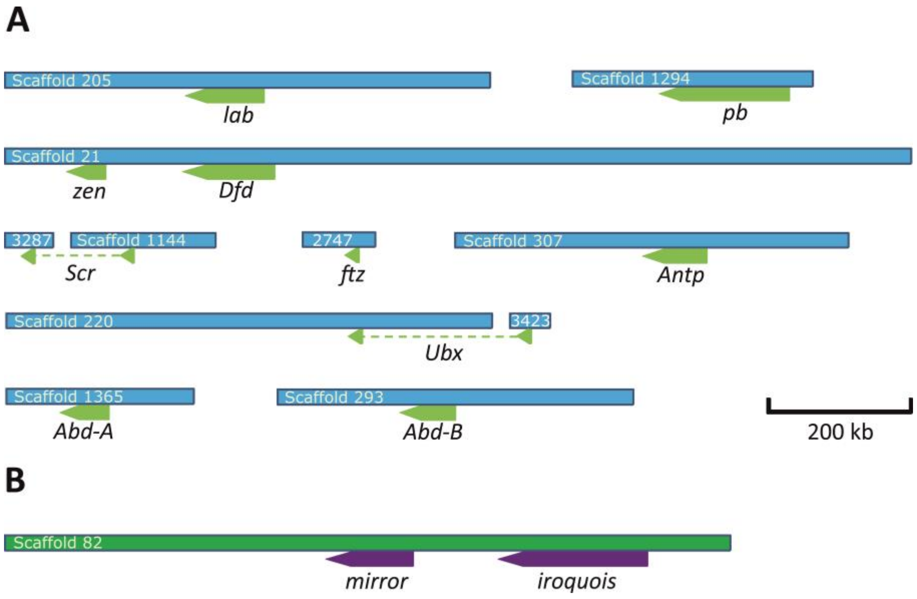


Figure S 5.2: Schematic representations of the Hox cluster **(A)** and the Iroquois Complex **(B)**, with transcriptional orientation as indicated and shown to scale for gene loci and total scaffold lengths.

Table S 5.1: Summary information for the Hox cluster genes.

Gene	Scaffold: start..end	Locus length (nt)	Number of CDS exons	Protein length (aa)
labial	Scaffold205:185231..267231	82,001	3	273
proboscipedia	Scaffold1294:17732..184912	167,181	3	571
zerknuellt	Scaffold21:1100766..1120884	20,119	4	198
Deformed	Scaffold21:161709..839714	678,006	2	331
Sex combs reduced	Scaffold1144:107131..107896 Scaffold3287:29411..30383	123,099[1]	2	301
fushi-tarazu	Scaffold2747:19754..20621	868	0	-
Antennapedia	Scaffold307:282886..304398	21,513	2	301
Ultrabithorax	Scaffold3423:34350..35230 Scaffold220:445840..446672	259,972[1]	2	295
abdominal-A	Scaffold1365:78419..171923	93,518	3	334
Abdominal-B	Scaffold293:181929..221609	39,681	2	316

[1] Gene locus length estimate assumes direct concatenation of the annotation-containing scaffolds.

Table S 5.2: Comparison of Hox proteins between the beetles *Tribolium castaneum* (Tcas) [57] and *Anoplophora glabripennis* (Agla) [46], and the bugs *Oncopeltus fasciatus* (Ofas) and *Cimex lectularius* (Clec) [19]. Percent protein size change is relative to *Tribolium*.

	Tcas		Agla			Ofas			Clec		
Hox gene	protein length	# CDS exons	protein length	% protein size change	# CDS exons	protein length	% protein size change	# CDS exons	protein length	% protein size change	# CDS exons
lab	353	2	384	1.09	2	273	0.77	3	279	0.79	2
pb (mxp)	654	3	717	1.10	4	571	0.87	3	601	0.92	3
zen (zen1)	246	3	319	1.30	2	198	0.80	4	307	1.25	3
Dfd	412	2	454	1.10	2	331	0.80	2	337	0.82	2
Scr (Cx)	312	2	336	1.08	3	302	0.97	2	323	1.04	2
ftz	290	2	389	1.34	2	-	0	0	235	0.81	2
Antp (Ptl)	325	2	327	1.01	2	301	0.93	2	302	0.93	2
Ubx	314	3	296	0.94	2	295	0.94	2	301	0.96	2
abd-A	343	3	365	1.06	3	334	0.97	3	345	1.01	3
Abd-B	351	2	388/400	1.14	2	316	0.90	2	361	1.03	3

5.1.c Iroquois Complex (Iro-C) cluster

Contributors: Jan Seibert and Kristen A. Panfilio

The transcription factors *araucan* (*ara*), *caupolican* (*caup*), and *mirror* (*mirr*) belong to the TALE superclass of homeodomain proteins [58] and form the Iroquois-Complex (Iro-C) in the fruit fly *Drosophila melanogaster* [59-61]. It is already known that *ara* and *caup* arose due to a tandem duplication of the gene *iro* in the drosophilid lineage [62], while the Iro-C itself is ancestral within crustaceans and insects. In vertebrates on the other hand, one can find 2-4 Irx clusters, with up to three genes per cluster. More basally branching metazoans like the cnidarians or the placozoans, but also nematodes, which as fellow Ecdysozoa are much closer to insects, have only a single Irx gene [63]. Given these deep, lineage-specific complements of Irx genes and clusters, they provide a perfect target to assay the quality of the *Oncopeltus* genome.

The search for genes of the Iro-C in *O. fasciatus* revealed again its conserved synteny (Figure S 5.2 B, previous section; see also Methods, below, and Table S 5.3, Figure S 5.3). As in *D. melanogaster* and *Tribolium castaneum*, in *O. fasciatus* the *iro* gene (the single *ara*/*caup* ortholog) and *mirr* are transcribed from the same DNA strand, with *iro* situated 5′ to *mirr*, in our genome assembly (Scaffold 82). Phylogenetic analyses of Iro-C protein sequences from representative species within the insects confirmed the expected relationship of *O. fasciatus* Iro and Mirr with orthologous proteins of the other species (Figure S 5.4). Looking at other predicted gene models within i5K species, *iro* and *mirr* could also be found in the bed bug *Cimex lectularius* (Scaffold 47, [19]) and the Asian longhorned beetle *Anoplophora glabripennis* (Scaffold 135, [46]) with conserved synteny.

In most bilaterians, but not in vertebrates, this synteny can be extended to the ankyrin repeat-containing *sosondowah* (*sowah*) gene, which is known to be associated with the Iro-C [61, 64]. In all so far investigated insect species *sowah* is located 5′ to the Iro-C and has an opposite reading direction [63].

In our analyses, *sowah* is found on Scaffold 1078, so it is unclear if the synteny is conserved in *O. fasciatus*. An indirect hint could be that also *ceramide transfer protein* (*cert*) is found in the same reading direction on Scaffold 1078. *cert* is in *D. melanogaster* located on chromosome 3L, together with *sowah* and all three Iro-C genes. So the fact that both genes have the same reading direction and the same transcriptional orientation in the milkweed bug compared to the fruit fly might indicate conserved synteny.

Our results not only support the high degree of conservation of the Iro-C and its associated genes in *Oncopeltus fasciatus*, they also strengthen the quality of our assembly in general with the expected linage of the Iro-C gene pair.

Methods

Iro-C and Irx protein sequences were retrieved using TBLASTN and BLASTP algorithms on the corresponding databases/genome browsers (Table S 5.3) for selected species (Figure S 5.3). The sequences of the three Iro-C genes in *Drosophila melanogaster* served as queries in all cases. Predicted gene models within the *Oncopeltus fasciatus* assembly were manually curated to obtain the full sequence.

The phylogenetic relationships of the cladogram (Figure S 5.3) were first obtained via NCBI's taxonomy browser, refined with published findings [65], and rebuilt and edited with the free software INKSCAPE.

The final tree (Figure S 5.4) was built with the free software MEGA 6 [66], using ClustalW for the alignment and Maximum Likelihood for phylogenetic tree construction. For the analyses preferences the default settings were used, except the number of bootstrap replications, which was set to 1.000.

Table S 5.3: Iro-C homolog protein sequence sources.

Species name	**Abbreviation**	**Source database**	**ID's for *araucan/ caupolican***	**ID for ortholog of *iroquois***	**ID for ortholog of *mirror***
Bombyx mori	*Bmor*	NCBI	-	XP_004929820.1	XP_004929953.1
Drosophila melanogaster	*Dmel*	NCBI	AAF49896.1/ AAF49895.1	-	AGB94471.1
Nasonia vitripennis	*Nvit*	NCBI	-	XP_008216675.1	XP_001604937.1
Oncopeltus fasciatus	*Ofas*	i5k	-	OFAS025321	OFAS001761
Tribolium castaneum	*Tcas*	Beetle Base	-	TC003632	TC003634
Trichinella spiralis	*Tspi*	NCBI	-	XP_003372666.1	-

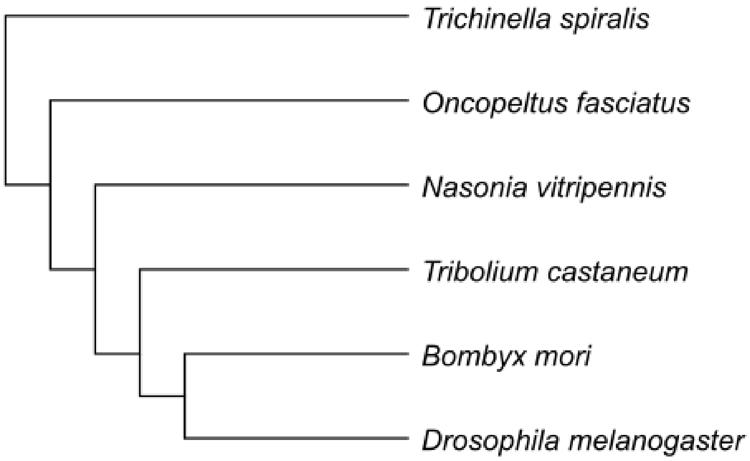


Figure S 5.3: Cladogram of species, based on the NCBI Taxonomy Browser, with the nematode *Trichinella spiralis* providing an ecdysozoan outgroup to selected insects.

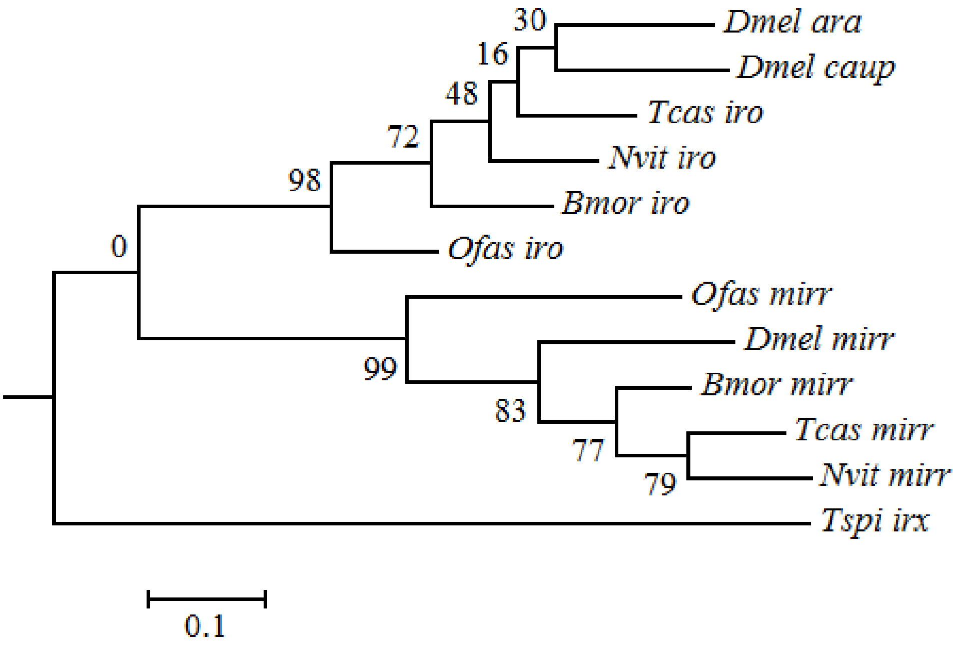


Figure S 5.4: Maximum likelihood phylogenetic analysis of insect Iro-C proteins. The two branches of the two (three in *D. melanogaster*) Iro-C genes have an excellent support value, putting *O. fasciatus mirror*, as well as *iroquois* in the corresponding group. *D. melanogaster araucan* and *caupolican* group together, like one would predict after gene duplication. Notably the support values for this grouping and already for those in the *iroqouis* branch starting at *N. vitripennis* are quite low.

5.1.d T-box transcription factors and heart determinants

Contributors: Thorsten Horn and Kristen A. Panfilio

We found 5 T-box genes in *Oncopeltus* based on homology to *Drosophila melanogaster*, *Tribolium castaneum* and additional species where available (Figure S 5.5). This is fewer than in *Drosophila* (8) and *Tribolium* (6). However the gene *Dorsocross* has three paralogs in *Drosophila* (*Dmel-Doc1-3*), which stem from a recent duplication in the *Drosophila* lineage and possess similar expression patterns and functions [67]. Therefore, we would only expect 6 T-box genes in *Oncopeltus*, similar to *Tribolium*. The only missing T-box gene in *Oncopeltus* is *optomotor-blind-related-gene-1* (*org-1*). We did find a partial duplication of *Ofas-optomotor-blind* (*Ofas-omb*) on Scaffold 195, but as the genomic sequence of both regions is exactly the same, we rather suggest an assembly error than a hint towards *Ofas-org-1*. *Ofas-omb* consists of 11 exons (7 in *Drosophila*, 4 in *Tribolium*) located on two scaffolds and is probably complete on the CDS level. *brachyenteron* (*byn*), the insect homologue of *brachyury* (the founding member of the T-box gene group), was found in 8 exons (7 in *Drosophila*, 4 in *Tribolium*), which probably do not represent the full CDS.

All three T-box genes important for heart development in *Drosophila* (*midline* (*mid*), *H15* and *Doc*) [68] were found, as were other highly conserved transcription factors involved in heart development (the homeobox gene *tinman* and the zinc finger class gene *pannier*: Table S 5.4). *Ofas-mid* is split across two scaffolds and consists of 5 exons (4 in *Drosophila*, 6 in *Tribolium*). *Ofas-H15* consists of 5 exons (6 in *Drosophila*, 5 in *Tribolium*). The protein is slightly smaller than in *Drosophila* and *Tribolium* and the 5' UTR is missing. The orthologs of *Ofas-H15* and *Ofas-mid* are neighboring genes in both *Drosophila* and *Tribolium*, but in the current *Oncopeltus* assembly they occupy separate, relatively small scaffolds. Further improvement of the assembly will show if they are also linked in *Oncopeltus*. Homology assignment was not possible based on *Drosophila* homologues (Figure S 5.5), but is supported by *Tribolium* and additional species (not shown).

Based on homology and experimental RACE-PCR data, *Ofas-Doc* could be identified scattered across 4 scaffolds, which when put together probably constitute the complete gene model. The gene consists of 6 exons (5-7 in *Drosophila* and 4 in

Tcas). Of 4 splice sites within the coding region, 3 are conserved with *Drosophila* and 2 are conserved with *Tribolium*. The overall length seems comparable between *Oncopeltus*, *Drosophila* and *Tribolium*. Although *Doc* plays an important role in extraembryonic development in *Tribolium* and *Drosophila* [69], its role in heart development does not seem to be conserved in *Tribolium* [69]. Functional investigations of this gene in *Oncopeltus* will facilitate our understanding of the evolution of T-box transcription factors in insects.

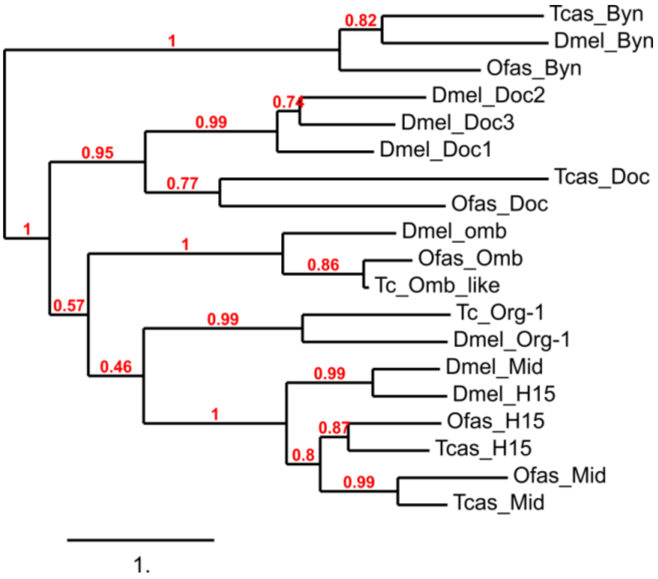


Figure S 5.5: T-box transcription factors in *Drosophila*, *Tribolium* and *Oncopeltus*. Maximum likelihood tree generated from protein sequences with default parameters without Gblocks at http://www.phylogeny.fr, based on [42, 43].

Table S 5.4: Manual curation summary of potential dorsal and mesodermal determinants.

Gene	Scaffold: start..end	Locus length (nt)	# CDS exons	Protein length (aa)
tinman	Scaffold212:18752..21136	2,385	2	128 (partial)
pannier	Scaffold2229:82858..97807	14,950	3	208 (partial)
dorsocross	Scaffold7:450646..450744 Scaffold6818:378..503 Scaffold1378:36101..37197 Scaffold7:510006..512170	99 126 1,097 2,165	1 1 2 2	319
midline	Scaffold776:16160..18303 Scaffold2277:68612..109253	2,144 40,642	2 3	312
H15	Scaffold679:74406..129278	54,873	5	296
optomotor-blind	Scaffold111:653518..772820 Scaffold195:155269..184484	119,303 177,148	2 9	636
brachyenteron	Scaffold262:13930..46676	32,747	8	292 (partial)

5.1.e Nuclear receptors

Contributors: Yong Lu and Leslie Pick

The nuclear receptor (NR) family of transcription factors is one of the largest transcription factor families, functioning in diverse biological processes including homeostasis, cell proliferation, reproduction and development [reviewed in 70, 71-76]. Many NRs function as ligand-responsive transcription factors whose activity is regulated by the binding of small molecule ligands (such as ecdysone, see below in Section 5.2.b). However, for other NRs, no ligand has been identified to date and many of these orphan NRs are likely regulated by other mechanisms. In fact, NRs appear to be evolutionarily flexible transcription factors with gain and loss of modes of regulation. Changes in ligand binding, protein-protein interaction and post translational modifications all play roles in their cooption into new biological roles, perhaps explaining their maintenance in animal genomes [77-79]. The NRs have been parsed into 5 large families that are distinguished by their requirements for ligand as well as by the structure of their ligand binding domains [80]. Both ligand-activated and orphan nuclear receptors share a similar, modular protein structure comprised of a variable N-terminal region, a highly conserved zinc-finger DNA binding Domain (DBD), a second variable hinge region, and a less conserved Ligand Binding Domain (LBD) that includes the activation function-2 (AF-2) domain (Table S 5.5, Figure S 5.6). For ligand-regulated NRs, the role of ligand binding is to induce a conformational change that repositions the AF-2 domain, releasing corepressors and creating a surface for coactivator binding, thus switching the NR between repressive and activating states [81]. Orphan nuclear receptors are activated by protein-protein interactions and post-translational modifications with, at least in several cases, the LBD folding in an active conformation in the absence of ligand [82, 83]. In *Drosophila*, 18 classic NRs have been identified [84]. Eight of these NRs are involved in ecdysone signaling pathways. The other NRs participate in a variety of processes including embryonic development, differentiation and metabolism. Orthologs of six of these NRs were identified in the *Oncopeltus* genome: *HR96, HNF4, HR78, Tailless, Dissatisfaction, HR51, Seven up,* and *HR38* (Table S 5.5). The two *Drosophila* NRs not found in the *Oncopeltus* genome are ERR and HR83. In *Drosophila*, ERR was found to be important for carbohydrate metabolism [85, 86]. The function of HR83 in *Drosophila* is not clear. According to FlyBase, two classic

HR83 alleles do not show any obvious phenotype and the flies are fertile. However, because the current *Oncopeltus* genome is still divided among many scaffolds, it is possible that these genes were missed in our annotation. The *Anopheles*, *Tribolium* and *Apis* genomes all have 18 NRs, with *Aedes* missing HR83 [57, 87, 88]. Thus, the family of NRs appears to be highly conserved in hemi- and holometabolous insects.

Overall, of the 16 nuclear receptors that were found in the *Oncopeltus* genome (Table S 5.5), Seven Up retains the highest degree of similarity to its *Drosophila* ortholog, with its DBD 100% identical to that of the *Drosophila*, and its LBD 94% identical to that of the *Drosophila*. On Scaffold 1872, there is a predicted NR (OFAS016187-RA), most likely an ortholog of HNF4, that may have resulted from retrotransposition of a processed NR transcript, as it is comprised of a single exon. This gene would encode a protein of 470 aa; the DBD of the predicted protein shares 79% identity with that of *Drosophila HNF4*, and the LBD shares 60% identity with that of *Drosophila HNF4*; and it has expression support from both embryonic and post-embryonic transcriptomes. For the 16 NRs identified, all appear to be present in single copy in the *Oncopeltus* genome. This is similar to the situation in other insects, including *Drosophila*, *Anopheles*, *Aedes*, *Tribolium* and *Apis*. Finally, aside from OFAS016187-RA, no novel NRs were identified in the *Oncopeltus* genome, although species-specific NRs may still be revealed.

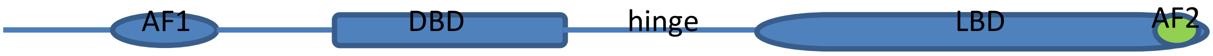


Figure S 5.6: Common structure of nuclear receptors: AAF1 (Activation Function 1), DBD (DNA Binding Domain), Hinge, LBD (Ligand Binding Domain), and AF2 (Activation Function 2) located within LBD.

Table S 5.5: List of identified *Oncopeltus* nuclear receptors (NRs) (in Excel supplement).

5.1.f Dorsal-ventral body axis: BMP/TGF-β pathway

Contributors: Lena Sachs and Siegfried Roth

The BMP pathway consists of extracellular ligands and their modulators, of transmembrane receptors with intracellular serine/threonine kinase activity and of cytoplasmic signal transducers, which relay the signal to the nucleus and regulate target genes [89]. All core components of the BMP signal pathway were found in the *Oncopeltus* genome (for functional analysis of key components see [90]). These include orthologs encoding: the two typical BMP ligands Decapentaplegic (Dpp, a BMP2/4 ligand) and Glass bottom boat (Gbb, a BMP5/7 ligand), the type I receptors Thickveins (Tkv) and Saxophone (Sax), and one type II receptor (Punt). Regarding cytoplasmic signal transducers we found one homolog of Medea (Co-Smad/Smad4). However, we found three paralogs for Mad *(*Smad1/5), two of them occurring on the same scaffold in close proximity with inverse orientation, which likely represents a recent duplication.

In *Drosophila* two additional transcription factors, Schnurri and Brinker, have been identified that act together with Medea/Mad to regulate many BMP target genes [91]. Both are present in genomes of other holometabolous insects. We found homologs of each in the *Oncopeltus* genome. The gene model for *brinker* is supported by transcriptome data ([48]). Potential orthologs of *brinker* are also present in the bed bug and pea aphid (41% and 39% amino acid identity, from the GenBank "uncharacterized protein" accessions XP_014250940.1 and XP_003240098.1, respectively). A deep conservation of *brinker* in insects is interesting, as there are no clear *brinker* homologs in other arthropods or other animal phyla. However, ongoing *brinker* evolution may be suggested by the fact that whereas the *Oncopeltus*, pea aphid, and *Drosophila* orthologs each encode a protein with a single Brinker DNA-binding domain (BrkDBD: Pfam ID PF09607), existing protein accessions from both the bed bug and *Tribolium* each contain two BrkDBD's.

The activity of BMP receptors and cytoplasmic signal transducers are subject to secondary modulation. An interesting case is the pseudoreceptor BAMBI, which has been well studied in vertebrates and acts as a negative regulator of BMP signaling [92]. BAMBI has been lost in the lineage leading to *Drosophila*, however it is present in *Tribolium* and *Nasonia* [93, 94]. We found a homolog in *Oncopeltus* (incomplete gene model). Homologs also exist in *Halyomorpha* and *Cime*x. Modulation of the BMP receptors and SMADs occurs also at the level of regulated protein degradation. In *Drosophila* the E3 ligase Smurf regulates ubiquitination and proteolysis of the BMP receptor Thickveins and/or pMAD [95, 96]. A Smurf ortholog was found in *Oncopeltus*.

Oncopeltus possesses a typical set of extracellular modulators of BMP signaling, including: a homolog of the BMP inhibitor Short gastrulation (Sog); a homolog of Tolloid (Tld), the protease that cleaves Sog to release the BMPs from Sog-BMP complexes; Crossveinless-1 (Cv-1, also known as Twisted gastrulation (Tsg)); Crossveinless-2 (Cv-2); and Pentagone [89, 97]. However, no BMP modulators of the DAN family were discovered in the genome, although two are present in *Tribolium*, suggesting a lineage-specific loss [92, 94]. Similarly, no *admp* ortholog could be identified, similar to the situation in pea aphids, although one is present in the holometabolous wasp *Nasonia vitripennis* [93].

Interestingly, Noggin, a secreted BMP inhibitor that so far has not been identified in any of the holometabolous insect genomes [50], is present in *Oncopeltus*. Within the Hemiptera Noggin homologs show considerable sequence divergence, as the Noggin proteins of the bed bug and pea aphid show only 50% and 32% similarity to *Oncopeltus* Noggin, respectively.

5.1.g Dorsal-ventral body axis: Toll/ NFκB pathway

Contributors: Yen-Ta Chen and Siegfried Roth

The Toll pathway is involved in innate immunity, convergent extension and dorsal-ventral (DV) patterning in a wide range of insect species including *Oncopeltu*s and *Rhodnius* [Y.T. C and S.R. unpublished results, 90, 98, 99-101]. The focus of this annotation report is DV patterning.

The core Toll signaling cascade in insects comprises the Toll-like transmembrane receptors (TLRs), the intracellular adaptor Myd88, two serine-threonine kinases (Tube-like kinase and Pelle), the I-kappaB homolog Cactus and the NF-kappaB transcription factor Dorsal [102, 103].

We identified six TLRs in *Oncopeltus*. Toll1, Toll6, Toll7, Toll8, Toll10 were identified in both the genome and transcriptome [48] while Toll9 was only identified in the transcriptome. Toll6, Toll8 and Toll10 belong to the Loto clade of segmentally expressed Tolls [101]. Toll1 is involved in DV patterning and probably also in innate immunity [90].

As for the downstream cytoplasmic signal transducers we found one representative each for Myd88, Pelle and Tube-like kinase. The presence of the latter further supports the idea that Tube proteins fused to a kinase domain represent the ancestral state for insects, and consequently that the Tube proteins of Hymenoptera and higher Diptera lacking a kinase domain resulted from a secondary loss [103]. Surprisingly, we identified six paralogs for I-kappaB/*cactus* scattered throughout the genome (Figure S 5.7), four of which have been shown to be expressed during blastoderm stages [90]. (Note that these cactus paralogs, annotated as ‘cactus-like’, were added only for OGS v1.2.) Phylogenetic analysis shows that the increase in *cactus* paralogs is lineage-specific within Heteroptera, as *Cimex lectularius* and *Rhodnius prolixus* each have only one *cactus* gene. However, a tendency to duplicate *cactus* genes has been observed in other insect lineages, such as in the Hymenoptera, and is likely to be linked to complex immune functions of Toll/NF-kappaB signaling.

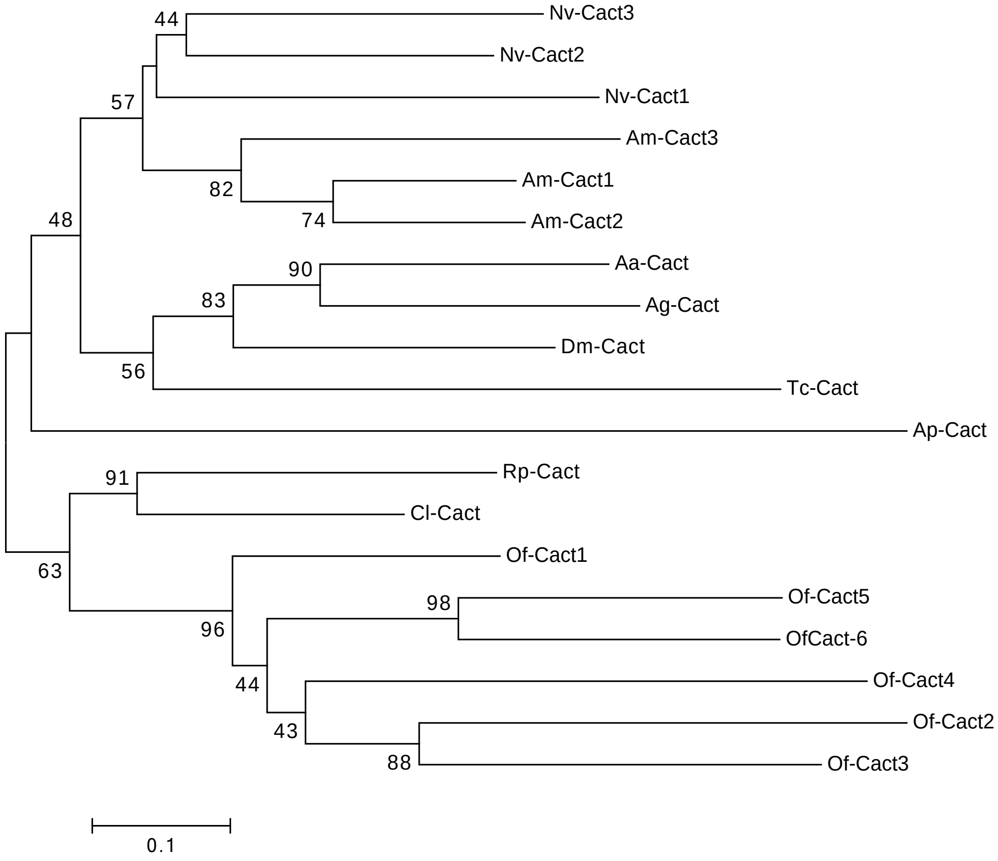


Figure S 5.7: Phylogenetic tree of *cactus* genes.

The optimal tree with the sum of branch length = 6.19659819 is shown. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) is shown next to the branches. The tree is drawn to scale, with branch length in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the Poisson correction method and are in the units of the number of amino acid substitutions per site. The analysis involved 19 amino acid sequences. All positions containing gaps and missing data were eliminated. There were a total of 118 positions in the final dataset. Evolutionary analyses were conducted with MEGA7 [104].

The *Oncopeltus* genome contains two NF-kappaB/*dorsal* sequences resulting from a duplication which has occurred independent from the *dorsal* gene duplications found in holometabolous lineages and indeed is lineage-specific within the Heteroptera as *Cimex* and *Rhodnius* possess only one *dorsal* gene. Both *Of-dorsal* genes are on the same scaffold within a 45-kb interval and have the same orientation suggesting that they have arisen by tandem duplication. Both dorsal genes contribute to DV patterning, albeit to a different degree [90].

Among the potential direct or indirect target genes of NF-kappaB/Dorsal [100, 105] we identified single copies of the following DV patterning genes: *short gastrulation* (see annotation of BMP signaling components), *twist* (*twi*), *snail* (*sna*), *single minded* (*sim*), *SoxNeuro* [106], and the three columnar neuroectodermal genes *muscle segment homeodomain* (*msh*), *intermediate nerve cord defective* (*ind*) and *ventral nervous system defective* (*vnd*) [107].

5.1.h Innate immunity

Contributors: Chris G.C. Jacobs, Yen-Ta Chen, Maurijn van der Zee

We were able to annotate 99 immune genes in the *Oncopeltus* genome. This number includes the Toll signaling components described in Section 5.1.g. Additionally, fifteen Defensins, seven lysozymes and five Hemiptericins were annotated, of which some have been previously identified [108]. Interestingly, no Cecropins, Attacins or Thaumatins were found. In addition, we found nine new potential antimicrobial peptides (AMPs), including three recently confirmed Serosins and two recently confirmed Ovicins (Submitted manuscript: Jacobs CGC, van der Hulst R, Chen YT, Roth S, van der Zee M. (2017) Innate immune function of the serosa in a hemimetabolous insect). This illustrates the need for species-specific AMPs for effective elimination of evolving pathogens. The Toll pathway is well conserved (see Section 5.1.g). Although IMD is missing from the hematophagous hemipterans *Rhodnius* and *Cimex*, IMD is present in the genomes of *Oncopeltus* and *Gerris*. However, classical cloning with degenerate primers was necessary to isolate the *Oncopeltus* ortholog. As for FADD and Tab2, gene models were absent from the genome assembly and coding sequences were even not found by tBLASTn, possibly because of extremely short exons (see also main text results and discussion on gene structure evolution and short exons within the Hemiptera). The sequences were finally found by *de novo* RNA-seq assembly (submitted manuscript cited above). IMD is also present in the closely related leaf hoppers *Nilaparvata* [109] and *Homalodisca*, suggesting that IMD was present in an ancestor of all Paraneoptera. Loss of IMD was reported from *Acyrthosiphon* [32]. IMD was also not found in the genomes of *Frankliniella*, *Pediculus*, *Diaphorina* and *Pachypsylla*. This could be a single evolutionary loss, if a close relationship of the Thysanoptera, Phthiraptera and Sternorrhyncha is assumed, as found for instance when using the BI-DNA method (see Figure S 5.8, [110]). However, the most accepted phylogeny suggests multiple losses. These losses could mean that the Toll pathway is not only able to upregulate effector genes in response to Gram-positive bacteria, but also in response to Gram-negative bacteria. Interestingly, response to these distinct bacterial inputs is also not strict in *Tribolium*, and to some extent also not in *Drosophila* [111-114].

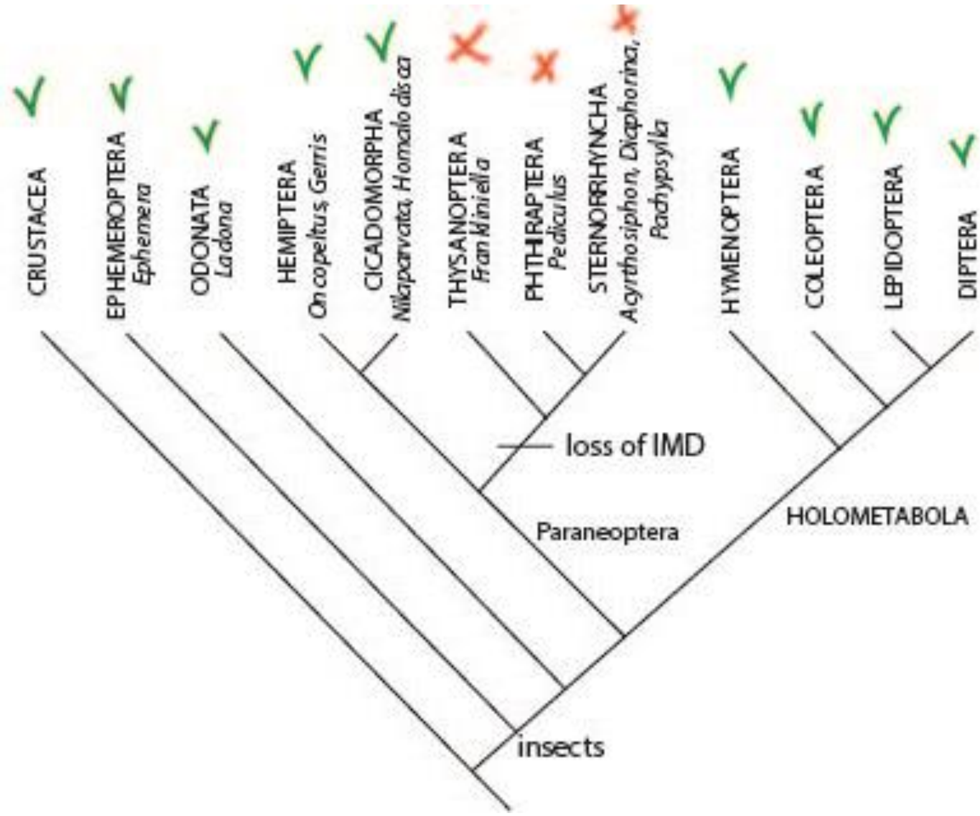


Figure S 5.8: Presence of IMD in major insect groups. Phylogeny from [110] based on BI-DNA method.

5.1.i Notch, Hedgehog, and Torso RTK pathways

Contributors: Peter K. Dearden, Andrew G. Cridge, Elizabeth J. Duncan, Megan Leask, Mackenzie Lovegrove, Olivia Tidswell.

Cell signaling pathways (Notch, TGFβ and Hedgehog) are, on the whole, conserved. Subtle differences in downstream genes likely reflect evolutionary distance from *Drosophila* (where these systems are well studied), rather than *Oncopeltus* specific biology. However, see above (Section 5.1.a) for details on the Torso Receptor Protein-Tyrosine Kinase (RTK) signaling pathway, which is involved in terminal patterning in the context of anterior-posterior axis specification, and for which we have conducted a phylogenetic analysis to assign *Oncopeltus* ligand orthology (Figure S. 5.9).

Previous work has indicated that conserved gene complexes are rare in insect genomes, and that only three are conserved over significant evolutionary time. One of these is the Hox complex (Section 5.1.b)**.**

A second conserved complex is the Runt complex. This set of four Runt domain containing genes is a feature of insect genomes, but has not yet been found in crustaceans [20]. Ancestrally (as seen in aphids and *Pediculus*) the complex contains 4 genes in a stereotyped order. In *Oncopeltus*, this order may be conserved. *Runt* and *RunXA* sit on the same scaffold, with no surrounding genes, and *lozenge* and *RunXB* sit on their own scaffolds with no surrounding genes. If these scaffolds were contiguous, then this would produce a conserved Runt complex. Additionally, a duplicated RunXA gene sits on Scaffold 2955.

The third complex is the Enhancer of split (E(spl)) complex, a feature of both insect and crustacean genomes [115]. This complex of genes is regulated by Notch signaling and has roles in cell fate specification. Ancestrally (as seen in *Daphnia* and aphids), this complex consists of four genes in stereotyped locations and is often associated with a paired-like gene at one end of the complex. In *Oncopeltus* this complex appears to have been broken apart, with only two E(spl) genes in the genome, both surrounded by other genes with homologs in other species, providing confidence that these are real genes. The two E(spl) genes are most closely related to E(spl) bHLH2 genes (annotated as E(spl) mBeta, Scaffold 744), and E(spl) bHLH1 genes (annotated as E(Spl) m7, Scaffold 857). No orthologues of *Her* or *malpha* (a bearded class gene found in the E(Spl) complex) can be found. This rearrangement is most likely derived given the conservation of the complex in crustaceans.

Other gene families analyzed include the *sp* genes, where we have identified orthologs of *sp4* and *sp8/9* and confirmed the absence of *sp5/buttonhead* [116, 117]**.**

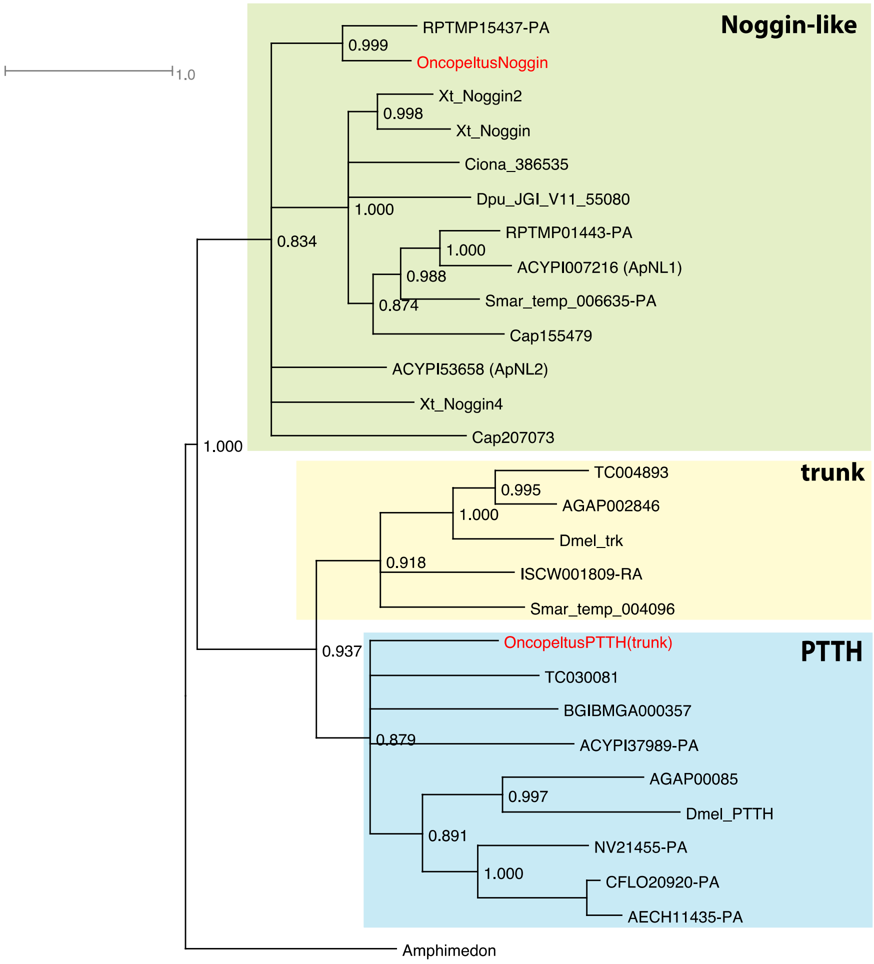


Figure S 5.9: Bayesian phylogeny of arthropod Noggin-like/PTTH/trunk family of proteins [50], including orthologs identified from the *Oncopeltus fasciatus* genome. Using a combination of methods, including BLAST homology searches and screening annotated *Oncopeltus* proteins with a custom HMM motif using the HMMER suite of programs, two Noggin-like/PTTH/trunk family members were identified in *Oncopeltus*. Bayesian phylogeny indicates one of these orthologs is most closely related to other insect PTTH proteins, and the other ortholog is a member of the Noggin-like family. Posterior probabilities are shown at nodes, the tree is rooted with the Noggin-like sequence identified in the *Amphimedon* genome. Multiple alignments were carried out using ClustalX [118]; the phylogeny was initially carried out under mixed models, and then with the most appropriate specific model using MrBayes. The Monte Carlo Markov Chain search was run with four chains over 1,000,000 generations with trees sampled every 1000 generations. The first 25% trees were discarded as ‘burn-in’.

5.1.j Wnt pathway

Contributors: Iris M. Vargas Jentzsch and Kristen A. Panfilio

The Wnt pathway is a signal transduction pathway with fundamental regulatory roles in embryonic development in all metazoans. The emergence of several gene families of both Wnt ligands and Frizzled receptors allowed the evolution of complex combinatorial interactions with multiple layers of regulation [119]. Wnt signaling affects cell proliferation and migration, as well as segment polarity, patterning and addition in most arthropods [120]. Here we strived to identify and curate the automated models corresponding to the main components of the Wnt signaling pathway.

Methods

The protein sequences for the *wnt* ligands as well as receptors and downstream components (*armadillo*/*beta-catenin*, *dishevelled*, *frizzled*, *arrow*, *axin*, *shaggy*/*GSK*-3) from *Drosophila melanogaster*, *Tribolium castaneum* and *Acyrthosiphon pisum* were retrieved from NCBI (we excluded accessions named as 'predicted' to avoid propagating errors from automated annotations). Using these as queries, we performed tblastn searches on the *Oncopeltus* scaffolds with a cutoff maximum e-value of 1e^{-10}. Hits from all species together were ordered by scaffold and start position, and for each group of overlapping or closely adjacent hits, the putative gene name was identified by blasting back the hit sequence against Arthropoda proteins in GenBank. The query sequences with the best hits (lowest e-value) for each group were then used to identify the model to be curated, by doing a tblastn search in the *Oncopeltus* scaffolds from the blast instance at the National Agricultural Library (https://i5k.nal.usda.gov/legacy_blast). The blast results were visualized in the Apollo web browser for *Oncopeltus* (https://apollo.nal.usda.gov/oncfas/selectTrack.jsp), where the corresponding automated annotation models were edited. Homology, intron/exon boundary assessments, and protein sequence completeness were identified by manual inspection and correction of protein alignments generated with Clustal Omega (http://www.ebi.ac.uk/Tools/msa/clustalo/), and subsequent phylogenetic analyses at http://www.phylogeny.fr/ [42, 43]. For this, we used all query sequences, and included additional orthologs from the crustacean *Daphnia pulex* and the myriapod *Glomeris marginata*. Potential gene duplications were also confirmed during this process.

The numbering for *wnt and fz* orthologs was assigned based on the corresponding vertebrate homolog (the naming of *Drosophila* orthologs was changed accordingly in phylogenetic analyses).

Results

A total of 25 models were curated for the main Wnt signaling genes on the *Oncopeltus* assembly (Table S 5.6). We identified complete gene models for six Wnt ligand subfamilies, two Frizzled transmembrane receptor families, the co-receptor *arrow*, and the downstream components *armadillo*/*beta-catenin*, *dishevelled*, *axin*, and *shaggy*/*GSK*-3. For all these genes, we were able to find the complete coding sequence, even though four genes were each split across two different scaffolds: *axin*, *dishevelled*, *wnt7* and *wntA*. The automated gene models were a good starting point for most annotated genes. However, 16 out of 23 exons had to be added *de novo* in the case of *arrow*, and we identified two possible mis-assembly instances in *wnt5* and *dishevelled*. In the case of *wnt5* there was a non-canonical splice site at the 3′ end of exon 3, and in the case of *dishevelled* exon 11 was found in between exons 6 and 7 (data supported by intron-spanning RNA-seq reads).

Two genes were duplicated, *armadillo* and *wnt8*, while all other genes were present in single copy in the assembly. Duplication and divergence of *armadillo/β-catenin*, which encodes an intracellular transducer of the Wnt pathway, was reported in the pea aphid and independently in *Tribolium* [121]. In fact, blastp searches in GenBank and phylogenetic analyses reveal that two copies are also found in a number of other hemimetabolous species, whereas most holometabolous species have a single copy. However, phylogenetic branch lengths suggest that functional divergence only occurred independently in the pea aphid and *Tribolium*. Only the bed bug and *Oncopeltus* do not conform to a strict paralogous sister gene topology, and it is unclear how many independent duplications of *armadillo* genes have occurred within the insects (Figure S 5.10). Interestingly, despite their sequence divergence, the *Tribolium* paralogs are the product of a tandem duplication, whereas the more highly conserved paralogs in *Oncopeltus* and *Cimex* are not linked in either species' assembly.

In contrast, the tandem duplication of *wnt8* in *Oncopeltus* seems to be unique, as a single *wnt8* ortholog was identified in other hemipteroid species, including the bed bug *Cimex lectularius*. The *Oncopeltus wnt8* paralogs are both expressed in the maternal/embryonic transcriptome [48], with additional expression support for *wnt8b* in the previously published adult (male) RNA-seq library ([33]; see also main text Fig. 2b).

The *Oncopeltus* Wnt ligand repertoire – *wingless/wnt1*, *wnt*5, *wnt7*, *wnt8a* and *wnt8b, wnt10* and *wntA* – is similar to the one found in the pea aphid (*Acyrthosiphon pisum*): *wingless/wnt1, wnt5, wnt*7, *wnt11, wnt16* and *wntA* [87]. In comparison, *Drosophila* and *Tribolium* have 7 and 9 Wnt subfamilies, respectively. This supports observations of a reduction in the ligand repertoire in insects compared to the 12 Wnt subfamilies inferred to have been present the last common ancestor of all arthropods [122]. Furthermore, members of the Hemiptera seem to have the fewest *Wnt* gene families reported in insects, with some of these losses perhaps having occurred relatively recently and independently in this clade. Nevertheless, assessments of gene absence need to be done with caution when dealing with draft assemblies from second generation sequencing, which is the case for most recently published genomes.

All Wnt ligand models were found isolated on different scaffolds with the exception of the *wnt8* paralogs, and *wingless* and *wnt10*. The latter two were clustered (on Scaffold 926) in the same transcriptional orientation and without other intervening genes, which is also the case in *Cimex*. This gene arrangement was also observed in *Tribolium castaneum* and *Drosophila melanogaster*, but with the *wnt6* locus between the two genes [123], reflecting the ancient arrangement of *wnt* genes in metazoans [124]. *Wnt6* was not found in the *Oncopeltus* assembly, and we also could not find evidence of *wnt6* in *Cimex lectularius*. This corroborates the postulated evolutionary loss of *wnt6* in Hemiptera [125], where the absence of *wnt6* was correlated with the absence of maxillary palps in insects.

Three models were curated for the *frizzled* (*fz*) transmembrane receptor families: two isoforms for *frizzled*, and one *frizzled-2*. These are only two of the four ancient *fz* families expected to have been present in the common ancestor of arthropods: *fz, fz2, fz3, fz4* [126]. The loss of *fz4* was also observed in *Acyrthosiphon pisum* [87].

Table S 5.6: Information on curated Wnt pathway models (in Excel Supplement).

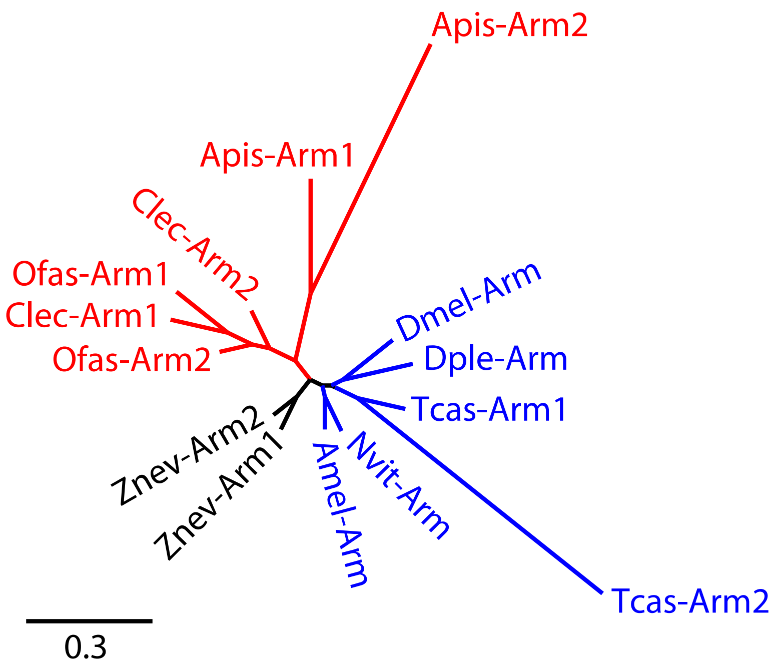


Figure S 5.10: Maximum likelihood phylogeny of Armadillo proteins, with hemipteran proteins in red and holometabolous proteins in blue. Branch length unit is substitutions per site, and all node support values are ≥62. Data are shown from nine insect species, represented by the four-letter species abbreviation and where the “Arm1” and “Arm2” paralogy distinction was arbitrarily assigned here for clarity, except in the case of *Tribolium* and the pea aphid, where public sequence accessions also bear these names. Taxonomic abbreviations and accession numbers (GenBank or *Oncopeltus* OGS) are: *Amel, Apis mellifera*, Arm: XP_006557863.1*; Apis, Acyrthosiphon pisum*, Arm1: XP_008183996.1, Arm2: XP_001946088.1*; Clec, Cimex lectularius*, Arm1: XP_014244091.1, Arm2: XP_014240454.1*; Dmel, Drosophila melanogaster*, Arm: NP_476666*; Dple, Danaus plexippus* Arm: OWR49369.1; *Nvit, Nasonia vitripennis*, Arm: XP_016840456.1; *Ofas, Oncopeltus fasciatus*, Arm1: OFAS000062, Arm2: OFAS025143*; Tcas, Tribolium castaneum*, EFA10737.1, Arm2: NP_001164124.1*;* and *Znev, Zootermopsis nevadensis*, Arm1: KDR13423.1, Arm2: KDR13424.1.

5.1.k Appendage patterning

Contributor: Yuichiro Suzuki

The transcriptional regulation of appendage patterning in insects has been well characterized in *Drosophila* [e.g. 127]. In *Drosophila*, appendages develop from imaginal discs, which are set aside during embryogenesis. Patterning of imaginal discs commences at embryogenesis and continues through larval development. During this period, the proximal-distal and the dorsal-ventral patterning of the appendages are established. During metamorphosis, the imaginal discs evert to form adult appendages [128]. Like other hemimetabolous insects, *Oncopeltus* does not form imaginal discs, and the nymphal antennae and legs resemble those of adults although the number of tarsal segments increases during the final molt. *Oncopeltus*, like other hemimetabolous insects, also differ from *Drosophila* in that they develop external wing pads during the nymphal stages that give rise to adult wings. Thus, understanding the development of *Oncopeltus* limbs has interesting evolutionary implications.

The early development of *Oncopeltus* appendage formation differs between holometabolous insects and hemimetabolous insects. In *Drosophila*, imaginal discs formation requires Wingless (Wg) [129, 130]. Similarly, *Tribolium* also requires Wg signaling for larval limb development [131]. In contrast, Wg signaling does not appear to be required for appendage patterning in *Oncopeltus* [132]. Thus, the earliest stages of appendage formation likely differ between hemimetabolous insects and holometabolous insects. Once the imaginal discs are formed, Hedgehog (Hh) signaling activates the expression of *Wg* and Decapentaplegic (*Dpp*), which interact to pattern the anterior-posterior and dorsal-ventral axes [133]. *Oncopeltus Dpp* has already been identified [132], and we have identified components of the Hedgehog pathway in the *Oncopeltus* genome. Studying these signaling pathways should provide interesting insights into the early development of appendages in hemimetabolous insects.

In the leg imaginal discs, the regional patterning in the proximal-distal axis involves gradual demarcation of the distinct regions of the limbs. At first, the imaginal disc expresses two proteins, *Distal-less* (*Dll*) in the presumptive distal portion and *homothorax* (*hth*) in the proximal portion [134-136]. Hth expression leads to the nuclear localization of the Extradenticle (Exd) protein [137]. *Oncopeltus Dll* and *hth* have already been identified by Angelini and Kaufman [138] and appear to play similar roles in *Oncopeltus* and *Drosophila* legs. We have also identified *exd* in the genome. At the intersection of Dll and Hth, Dachshund is activated and patterns the medial portion of the developing leg in both *Drosophila* and *Oncopeltus* [138]. Thus, the regional patterning mechanisms appear to be generally conserved between *Oncopeltus* and *Drosophila.* Later on, additional sculpting of the leg occurs though segmentation, which involve several transcription factors and signaling proteins [see e.g. 127]. We have identified homologs of several of these factors (*Notch, odd-skipped, spineless, abrupt* and *bric-a-brac*) with certainty in the *Oncopeltus* genome. Whether or not these transcription factors play similar appendage patterning roles in *Oncopeltus* and *Drosophila* remains to be seen.

Many of these genes also play patterning roles in the antennae and mouthparts of holometabolous insects [e.g. 139, 140]. *Oncopeltus* have specialized mouthparts that are adapted for piercing seeds. The stylets are formed from modifications to the mandibles and the maxilla, and aspects of the genetic regulation underlying their development appear to be distinct from *Tribolium* and *Drosophila* [138]. Thus, additional comparative studies should shed light on the evolution and development of these specialized morphologies.

In *Drosophila*, the imaginal discs adopt wing/haltere identity in response to the expressions of the wing selector genes *vestigial* (*vg*) and *scalloped* (*sd*), which encode proteins that form a heterodimeric complex [141, 142]. Loss of *vg* expression leads to cell death in the wing blade [142], and loss of *sd* also leads to cell death and loss of cell proliferation in the wing pouch [143, 144]. Similarly, Apterous (*Ap*) plays a key role in wing outgrowth and development of the dorsal wing structures and the wing margins [129, 145]. We have found *vg*, *sd* and *ap* in the *Oncopeltus* genome. The effects of *vg* and *sd* knockdowns have recently been reported in *Oncopeltus* [146] and demonstrate similar functional roles for wing development in both *Oncopeltus* and *Drosophila*.

5.1.1 Germline genes

Contributors: Chun-che Chang, Cassandra G. Extavour, Yi-min Hsiao, Tamsin E. Jones, Hsiao-ling Lu

Germline development in the Metazoa is controlled by a relatively small number of conserved genes. We searched the *Oncopeltus fasciatus* genome for these genes (Tables S 5.7, 5.8). In contrast to the pea aphid *Acyrthosiphon pisum* – another hemipteran model with a sequenced genome – *Oncopeltus* does not have an extraordinary duplication of the most conserved germline gene families, the *vasa*, *nanos* and *piwi* families. Like most other insects, *Oncopeltus* possesses single copies of *vasa* (previously described in [147]) and *nanos*. This is in contrast to the pea aphid, which possesses four copies each of *vasa* and *nanos* [87]. This difference may reflect the simple life cycle of *Oncopeltus* as compared to the pea aphid, which utilizes both sexual and asexual life cycles and undergoes different developmental programs in these two different phases [50, 148]. We found a single *piwi* gene in *Oncopeltus*, in contrast to both *Drosophila melanogaster*, which has two (*piwi* and *aubergine*), and to *A. pisum*, which possesses eight *piwi* paralogs [149]. Similarly, in *Oncopeltus* we could only identify one *ago3* gene—another component of the piRNA machinery like *piwi*—whilst in *A. pisum* duplicated *ago3* genes (*ago3a*, *ago3b*) have been annotated (see Section 5.4.a on RNAi machinery, below). This may imply that *Oncopeltus* employs a more ancestral piRNA pathway in germ cells than both *D. melanogaster* and *A. pisum*. The *wunen* genes in *Drosophila wun* and *wun-2* are involved in germ cell development and germ cell migration. We identified a single *wunen* gene in *Oncopeltus*, which may imply that *Oncopeltus wunen* is capable alone of performing both the germ cell development and germ cell migration functions performed by *wun* and *wun-2* in *Drosophila*. We found two copies of the gene *oo18 RNA-binding protein* (*orb*), which is found in single copy in both *D. melanogaster* and *A. pisum*. *Drosophila orb* is involved in axis specification as well as germ cell development, and the functional significance of *orb* duplication in *Oncopeltus* is unknown. We did not identify any ortholog of the *oskar* gene, which has not to date been identified in any other hemipteran species and has likely been lost from this lineage of insects.

Table S 5.7: Presence/absence of targeted genes in *Oncopeltus* (additional information can be found in Table S 5.8 in the Excel supplement).

Gene name	**Gene abbreviation**	**Number of copies found**	**Comments**
armitage	*armi*	1	
aubergine	*aub*	0	Low expectation, specific to Diptera
belle	*bel*	1	Identified previously in [150]
boule	*bol*	1	Identified previously in [150]
bruno/arrest	*bru/aret*	1	Identified previously in [150]
cappuccino	*capu*	1	
cup	*cup*	0	Low expectation, specific to *Drosophila*
fear of intimacy	*foi*	1	
germ cell-less	*gcl*	1	
gurken	*grk*	0	Low expectation, specific to *Drosophila*
gustavus	*gus*	1	
HMG Coenzyme A reductase	*Hmgcr*	1	
homeless/spindle E	*hls/spn-E*	1	
mago nashi	*mago*	1	
nanos	*nos*	1	
oo18 RNA-binding protein	*orb*	2	
oskar	*osk*	0	Low expectation, not identified in any hemipteran species
par-1	*par-1*	1	
pipsqueak	*psq*	1	
piwi	*piwi*	1	
polar granule component	*pgc*	0	Low expectation, specific to *Drosophila*
pumilio	*pum*	1	
spire	*spir*	1	
staufen	*stau*	1	
Trapped in endoderm-1	*Tre1*	0	Expectation unclear: according to FlyBase and OrthoDB, orthologs are found in other insects and in a few other non-insect Metazoa.
tropomysin II	*Tm2*	1	
tudor	*tud*	1	Identified previously in [150]
valois	*vls*	0	Low expectation, specific to Diptera
vasa	*vas*	1	Identified previously in [150]
wunen	*wun*	1	Plays likely redundant roles in germ cell migration with *wun2* in *Drosophila*
wunen-2	*wun2*	0	

Table S 5.8: Further details of annotated germline genes (in Excel supplement).

5.1.m Eye development

Contributors: Yi-min Hsiao, Hsiao-ling Lu, Chun-che Chang

We identified twenty-one *Drosophila* homologs of eye developmental genes in the milkweed bug *Oncopeltus fasciatus* (Table S 5.9). These genes are reckoned to participate in retinal development and eye morphogenesis as their conserved roles have been identified in *Drosophila* and other insect models (Table S 5.10). In comparison with the pea aphid *Acyrthosiphon pisum*, *O. fasciatus* shares all of the 21 eye developmental genes that have been annotated in *A. pisum* except *eyegone* (*eyg*) and *Optix* [87]. Both of these genes are unlikely to be missing in the *Oncopeltus* genome because their homologs can be identified in other insect models. Nonetheless, *twin of eyegone* (*toe*)—a tandem duplicated sister paralog of *eyg* in *Drosophila* and other higher dipterans [151] —was not found in both *A. pisum* and *Oncopeltus*. In *A. pisum* a homolog of the *eyeless* (*ey*) gene was not identified, and in *Oncopeltus* we could only find an incomplete contig without sequences encoding the homeodomain of the EY protein. This suggests that *Oncopeltus* possess a pseudogene of *ey* homologs or part of its complete sequence is missing in the database.

Table S 5.9: Major components of eye development in *Oncopeltus fasciatus.*

Gene name	**Gene abbreviation**	**Number of copies found**	**Comments**
twin of eyeless	*toy*	1	
eyeless	*ey*	1	Incomplete sequences, loss of HD domain
eyes absent	*eya*	1	
twin of eygone	*toe*	0	Not identified in *Oncopeltus*, specific to *Drosophila*
eyegone	*eyg*	0	Not identified in *Oncopeltus*
sine oculis	*so*	1	
Optix	*Optix*	0	Not identified in *Oncopeltus*
lozenge	*lz*	1	
hedgehog	*hh*	1	
sevenless	*sev*	1	
bride of sevenless	*boss*	1	
Son of sevenless	*Sos*	1	CDS split in three different scaffolds
downstream of receptor kinase	*drk*	1	
dachshund	*dac*	1	

atonal	*ato*	1	
decapentaplegic	*dpp*	1	
Transforming growth factor alpha	*TGFα*	1	
EGF receptor	*Egfr*	1	
delta	*Dl*	1	
serrate	*ser*	1	
Notch	*N*	1	

Table S 5.10: Orthologous numbers of eye developmental genes in four insect species.

Gene	Fruit fly	Beetle	Pea aphid	Milkweed bug
toy	1	1	1	1
ey	1	1	NA	NC
eya	1	1	1	1
toe	1	NA	NA	NA
eyg	1	1	1	?
so	1	1	1	1
Optix	1	1	1	?
lz	1	1	1	1
hh	1	1	1	1
sev	1	?	1	1
boss	1	?	1	1
Sos	1	1	2	1
drk	1	?	1	1
Wg	1	1	1	1
dpp	1	1	4	1
dac	1	1	?	1
ato	1	1	1	1
Egfr	1	1	1	1
N	1	1	1	1

?: genes that have not been identified

NA: sequences are not available.

NC: sequences are incomplete.

5.2 Structural and differentiation genes

5.2.a Bristle and neural development

Contributors: David Armisén, Antonin J.J. Johan Crumière, Séverine Viala, Essia Sghaier, Chloe S. Berger, Maria Emilia Santos, Peter N. Refki, Abderrahman Khila.

Abstract

The term "bristle" includes various hair-like structures with different functions, including mechanosensory and chemosensory hairs that are in constant contact with substrates and air, allowing the insect to sense its surrounding environment. For this reason, many genes involved in bristle development have been previously described as playing a role in neural development in *Drosophila* [152, 153]. In this study we annotated and analyzed 88 genes known to be involved in bristle development. Our results show an overall high conservation of protein sequence compared with *Drosophila* despite a higher intron number.

Results and discussion

The most studied role of insect bristles is perhaps their function as sensory organs for detecting various environmental stimuli. During development, each bristle is built from a small cluster of specialized cells including sensory neurons and support cells [154]. The shaft of the bristle extends from a single cell primarily via cytoskeletal arrangements [155]. In the fly *Drosophila*, the development of these bristles is quite well described and is regulated by a set of conserved developmental genes [156]. QTL studies have uncovered dozens of candidate genes and regions linked to variation in bristle density and morphology [152]. Some of this variation is also attributable to changes in non-coding sequences of a number of conserved developmental genes such as the achaete-scute complex [157, 158]. Based on fly genetics, we established and annotated a list of 105 genes known to be involved in neurogenesis and bristle development, as well as in variation in bristle number and density ([152, 153]; Table S 5.11). Among these, we were able to annotate 88 genes in the *Oncopeltus* genome, with no duplications, while we could not identify homologs for the remaining 17 candidate genes. Sequence comparison of the 88 genes with their homologous counterparts in *Tribolium castaneum*, *Acyrthosiphon pisum*, *Rhodnius prolixus*, *Pediculus humanus*, *Anopheles gambiae* and *Drosophila melanogaster* revealed a median of ~46 % protein identity. *Drosophila* alignment in particular shows that our dataset contains many genes encoding proteins recovered with full length, such is the case of the highly conserved gene Actin-5C [159]. Alignments show that many *Oncopeltus* genes with shorter sequence than *Drosophila* homologs have missing sequence primarily in the N-terminal and C-terminal. The number of exons of *Oncopeltus* compared with their *Drosophila* homologs suggests that in many cases this reduced size is due to missing exons. As neither sequence similarity nor RNA-seq data available support the presence of additional exons, a possible explanation for this absence could be a technical artifact due to problems in the assembly process or the impossibility to close all the gaps in the genome.

The lack of some exons highly contrasts with the higher number of exons included in *Oncopeltus* genes in average compared with *Drosophila.* It is striking that some particular genes such as *flightless* have a higher number of exons (nineteen in *Oncopeltus* instead of four in *Drosophila*) (Table S 5.11, Figure S 5.11), despite remaining shorter than its *Drosophila* homolog. In contrast, other genes, such as *raspberry* and *Ras85D,* have lost their introns and are rather encoded by single exons, (Figure S 5.12 and Figure S 5.13). This indicates that despite a substantial divergence in gene structure between *Oncopeltus* and *Drosophila*, there is high conservation in the encoded protein, suggesting possible strong selection on the coding sequence of these genes.

On the other hand, we can divide the 17 missing genes into four categories depending on the potential meaning of our failure to identify them in *Oncopeltus* genome. The first category includes genes that might be inherently hard to identify because their coding sequence is too short, such is the case of *Bantam*, a miRNA that is only 81 nucleotides long. The second category includes genes not found outside the *Drosophila* genus or closely related species through a 'nr' blastp search, such as *corto*, *dalmatian*, *diminutive*, *gliolectin* and *mushroom body defect,* and we therefore believe that they are truly missing from the *Oncopeltus* genome. The third category includes genes that we would have expected to find in the *Oncopeltus* genome, such as *Rala*, *effete, bonus* and *nuclear fallout,* as these genes have been described in *Tribolium castaneum* and *Acyrthosiphon pisum* or are highly and widely conserved in many species. The lack of *Oncopeltus* homologs is therefore very likely to be caused by an assembly artifact and not by a real gene loss. Lastly, in the fourth category, genes such as *bifocal*, *buttonless*, *cabut*, *escargot*, *pxb*, *spitz* and *Twin of m4* similar genes are present in other close species like *Tribolium castaneum* or *Acyrthosiphon pisum* but have diverged more and their absence in *Oncopeltus* genome can be either real, or caused by a high divergence that prevents blast recognition, or can be the result of assembly problems. Therefore, in this last instance we do not have enough elements to discern the cause of absence for each gene.

Methods

We annotated the bristle development genes by performing BLAT and tblastn searches on the *Oncopeltus* scaffolds with the corresponding *Drosophila* gene protein sequences available in FlyBase (release 5 for first round annotation and release 6 for annotation verification). To confirm orthology, we then blasted our *Oncopeltus* models into NCBI. Homology, intron/exon boundary assessments, and protein sequence completeness were identified by manual inspection using RNA-seq alignments available and protein alignments generated with Clustal Omega (http://www.ebi.ac.uk/Tools/msa/clustalo/).

Table S 5.11: Annotation of bristle and neural development genes: number of exons of the longest isoform. Percentage of identity calculated using protein alignment made by Clustal Omega (in Excel Supplement). Includes also a list of genes not found or only partially annotated.

```
CLUSTAL O(1.2.1) multiple sequence alignment


Flightless_FBpp0076893      ---MSVLPFVRGVDFTKNDFSA-TFPSSMRQMSRVQWLTLDRTQLAEIPEELGHLQKLEH
FlightlessI_Oncopeltus      MANTGVLPFVRGVDFSRNDFSDGKFPSSVRLMTGLQWLKLDRTHLDNIPEELGNLMKLEN
                                .**********::****  .****:* *: :***.****:* :******.* ***.


Flightless_FBpp0076893      LSLNHNRLEKIFGELTELSCLRSLDLRHNQLKNSGIPPELFHLEELTTLDLSHNKLKEVP
FlightlessI_Oncopeltus      LSLVRNDLERLHGELTELPCLRSLIIRHNKVKSSGIPADLFRSEELTTLDLSHNALKEVP
                            *** :* **::.****** ***** :***::*.**** :**: *********** *****


Flightless_FBpp0076893      EGLERAKNLIVLNLSNNQIESIPTPLFIHLTDLLFLDLSHNRLETLPPQTRRLINLKTLD
FlightlessI_Oncopeltus      DGLEKCKTVLVLNLSHNNITSIPNTLFMNLTDLLFLDLSNNLLETLPPQMRRLGNLQTLI
                            :***:.*.::*****.*:* ***. **:.**********.* ******* *** **:**


Flightless_FBpp0076893      LSHNPLELFQLRQLPSLQSLEVLKMSGTQRTLLNFPTSIDSLANLCELDLSHNSLPKLPD
FlightlessI_Oncopeltus      LNNNPLAHFQLRQLPSLVSLESLHMRYTQRTLSNLPSSLEMLTNLTDVDLSYNSLPKIPD
                            *..***  ********* *** *:*  ***** *:*:*:: *:** ::***:*****:**


Flightless_FBpp0076893      CVYNVVTLVRLNLSDNELTELTAGVELWQRLESLNLSRNQLVALPAALCKLPKLRRLLVN
FlightlessI_Oncopeltus      CLFTLVNLKRLNLSNNYLSELSLALEVWQRLETLNLSSNKLTSLPASVCKLTALRRLYLN
                            *::.:*.* *****:* *:**: .:*:*****:**** *:*.:***::***  **** :*


Flightless_FBpp0076893      DNKLNFEGIPSGIGKLGALEVFSAANNLLEMVPEGLCRCGALKQLNLSCNRLITLPDAIH
FlightlessI_Oncopeltus      DNQLDFEGIPSGIGKLSNLEVFSAASNQLEMIPEGLCRCGSLKKLILSSNRLITLPDTVH
                            **:*:***********. *******.* ***:********:**:* **.********::*


Flightless_FBpp0076893      LLEGLDQLDLRNNPELVMPPKPSEASKATSLEFYNIDFSLQTQLRLAGAAVPPSMPSSAT
FlightlessI_Oncopeltus      LLSDLQVLEVKDNPDLVMPPRPCEAQRGSGLEYYNIDFSLQTQLRLAGAALPQPLPQTSA
                            **. *: *::::**:*****:*.**.:.:.**:*****************:*  :*.:::


Flightless_FBpp0076893      PKDSTARKIRLRRGPRS--EGDQDAAKVLKGMKDVAKDKDNEAGAVPEDGKPESLKPKRW
FlightlessI_Oncopeltus      SKDPIARKLRLRRSRRDAEEADSDQAKILKGMKDIATEKNKC--KTEEEERAESLKPKRW
                             **  ***:****. *.  *.*.* **:******:*.:*::    . *: : ********


Flightless_FBpp0076893      DESLEKPQLDYSKFFEKDDGQLPGLTIWEIENFLPNKIEEVVHGKFYEGDCYIVLKTKFD
FlightlessI_Oncopeltus      DESLEKPPLDYSELFDEDAGQIPGITVWEIENFLPNMVEEAVHGKFFQADCYIILKTSLD
                            ******* ****::*::* **:**:*:********* :**.*****::.****:***.:*


Flightless_FBpp0076893      DLGLLDWEIFFWIGNEATLDKRACAAIHAVNLRNFLGARCRTVREEQGDESEQFLSLFET
FlightlessI_Oncopeltus      DSGNLAWNIFFWIGDKATLDKRACAAIHAVNLRNYLGAECRTAREEQGEESEEFLALFPS
                            * * * *:******::******************:***.***.*****:***:**:** :


Flightless_FBpp0076893      EVIYIEGGRTATGFYTIEEMIHITRLYLVHAYGATIHLEPVAPAITSLDPRHAFVLDLGT
FlightlessI_Oncopeltus      GITYIQGGRTPSGFYTVEDMTYITRLYRVHGAGAGIHLEPVPISYESLDPRYVFILDTGL
                             : **:**** :****:*:* :***** **. ** ******  :  *****:.*:** *
```

```
Flightless_FBpp0076893      HIYIWMGERSKNTLNSKARLMAEKISKTERKNKCEIQLERQGEESAEFWQGLGMTSEEAD
FlightlessI_Oncopeltus      KIFMWYGKKSKNTFRSKARLLCEKINKNERKGKAEIMTELYKEESEDFWCSLGEESGK-P
                            :*::* *::****:.*****:.***.*.*** *.**  *   *** :** .**  * :

Flightless_FBpp0076893      AAEPPKEHVPEDYQPVQPRLYQVQLGMGYLELPQVELPEQKLCHTLLNSKHVYILDCYTD
FlightlessI_Oncopeltus      PEEPVIEHVRNEWHPVEPRLYQVKLGMGYLELPQVELPCNRLEHTLLNSRNVYILDCYLD
                              **  *** ::::**:******:************** ::* ******:.******* *

Flightless_FBpp0076893      LFVWFGKKSTRLVRAAAVKLSRELFNMMDRPDYALVMRVPEGNEMQIFRTKFAGWDEVMA
FlightlessI_Oncopeltus      VFVC--------------------------------------------------------
                            :**

Flightless_FBpp0076893      VDFTRTAKSVAKTGANLTQWARQQETRTDLAALFMPRQSAMPLAEAEQLEEEWNYDLEMM
FlightlessI_Oncopeltus      ------------------------------------------------------------


Flightless_FBpp0076893      EAFVLENKKFVRLPEEELGRFYTGECYVFLCRYCIPIEEPENGSEDGANPAADVSKSSAN
FlightlessI_Oncopeltus      ------------------------------------------------------------


Flightless_FBpp0076893      NQPEDEIQCVVYFWQGRNAGNMGWLTFTFTLQKKFKAMFGEELEVVRIFQQQENLKFMSH
FlightlessI_Oncopeltus      ------------------------------LQKKFKSLFGEKLEVVRTHQQQENIKFLAH
                                                          ******::***:***** .*****:**::*

Flightless_FBpp0076893      FKRKFIIHTGKRKDKAHTAKGKSPVEFFHLRSNGGALTTRLIQINPDAVHLNSTFCYILH
FlightlessI_Oncopeltus      FKRKLVIHTGKRNTPR-DPNQPPPVEFYHLRANGGPLCTRLVQIKPDASTLNSAFCYILK
                            ****::******:      :   ****:***:*** * ***:**:***  ***:*****:

Flightless_FBpp0076893      VPFETEDDSQSGIVYVWIGSKACNEEAKLVQDIAEQMFNSPWVSLQILNEGDEPENFFWV
FlightlessI_Oncopeltus      VPFDSEE--SSGIVYVWIGSKANTDDIRLIGEIADEMFNNPWISLQVVNEGEEPDNFFWV
                            ***::*:  .************ .:: :*: :**::***.**:***::***:**:*****

Flightless_FBpp0076893      ALGGRKPYDTDAEYMNYTRLFRCSNERGYYTVAEKCADFCQDDLADDDIMILDNGEHVFL
FlightlessI_Oncopeltus      ALGGRKPYDKDADFMAYTRLFRCSNEKGYFIVSEKCSDFCQEDLADDDIMILDNGEQVFL
                            *********.**::* **********:**: *:***:****:**************:***

Flightless_FBpp0076893      WMGPRCSEVEVKLAYKSAQVYIQHMRIKQPERPRKLFLTMKNKESRRFTKCFHGWSAFKV
FlightlessI_Oncopeltus      WLGAKCSEVEVKLAYKSAQVCIF-------------------------------------
                            *:* :*************** *

Flightless_FBpp0076893      YL
FlightlessI_Oncopeltus      --
```

Figure S 5.11: Protein alignment made by Clustal Omega. Intron boundaries highlighted in blue for *Drosophila* and yellow for *Oncopeltus*.

```
CLUSTAL O(1.2.1) multiple sequence alignment


raspberry_FBpp0071423       MESTTKVKVNGFVESTSSSAAPAIQTKSTTGFDAELQDGLSCKELFQNGEGLTYNDFLIL
raspberry_Oncopeltus        ------------------------------MTDDIPVDGLSGKELFSNGDGLTYNDFIIL
                                                            *    **** ****.**:*******:**


raspberry_FBpp0071423       PGYIDFTAEEVDLSSPLTKSLTLRAPLVSSPMDTVTESEMAIAMALCGGIGIIHHNCTPE
raspberry_Oncopeltus        PGFIDFTPDEVDLRSELTKKITLQSPLVSSPMDTVTESDMAIAMALCGGIGIIHHNCTAE
                            **:**** :**** * ***.:**::*************:******************* *


raspberry_FBpp0071423       YQALEVHKVKKYKHGFMRDPSVMSPTNTVGDVLEARRKNGFTGYPVTENGKLGGKLLGMV
raspberry_Oncopeltus        YQANEVHKVKKYKHGFIRHPVVLSPKNTVADVFQVKKEHGFCGIPITENGQLGGKLCGIV
                            *** ************:*.* *:**.***.**::.:::.** * *:****:***** *:*


raspberry_FBpp0071423       TSRDIDFRENQP--EVLLADIMT--TELVTAPNGINLPTANAILEKSKKGKLPIVNQAGE
raspberry_Oncopeltus        TSRDIDFLADPNTHTILLEKVMTKLENIISAKAGVTLEEANHLLEESKKGKLPIINEKGE
                            *******  :     :** .:**   ::::*  *:.*  ** :**:********:*: **


raspberry_FBpp0071423       LVAMIARTDLKKARSYPNASKDSNKQLLVGAAIGTRSEDKARLALLVANGVDVIILDSSQ
raspberry_Oncopeltus        LVALIARTDLKKNRDYPKASKDENKQLLVGAAIGTRPEDQERLKLLASAGADVVVLDSSQ
                            ***:******** *.**:****.************* **: ** **.: *.**::*****


raspberry_FBpp0071423       GNSVYQVEMIKYIKETYPELQVIGGNVVTRAQAKNLIDAGVDGLRVGMGSGSICITQEVM
raspberry_Oncopeltus        GNSIFQVKMIKYIKETYPNLQVIGGNVVTAAQAKNLIDAGVDGLRVGMGSGSICITQEVM
                            ***::**:**********:********** ******************************


raspberry_FBpp0071423       ACGCPQATAVYQVSTYARQFGVPVIADGGIQSIGHIVKAIALGASAVMMGSLLAGTSEAP
raspberry_Oncopeltus        AVGRPQGTAVYKVAEYARRFGVPIIADGGIQSIGHVTKALSLGASTVMMGSLLAGTSEAP
                            * * **.****:*: ***:****:***********:.**::****:**************


raspberry_FBpp0071423       GEYFFSDGVRLKKYRGMGSLEAMERGDAKGAAMSRYYHNEMDKMKVAQGVSGSIVDKGSV
raspberry_Oncopeltus        GEYFFSDGVRLKKYRGMGSLEAMNRKDAQGSAMDRYFHSEMDKVKVAQGVSGSIVDKGSV
                            ***********************:* **:*:**.**:*.****:****************


raspberry_FBpp0071423       LRYLPYLECGLQHSCQDIGANSINKLRDMIYNGQLRFMKRTHSAQLEGNVHGLFSYEKRLF
raspberry_Oncopeltus        LRFLPYIQCGIQHGCQDIGARSLYVLRDMMYEGNLRFERRTHSAQNEGNVHSLFSYEKRLF
                            **:***::**:**.******.*:  ****:*:*:*** :****** *****.*********
```

Figure S 5.12: Protein alignment made by Clustal Omega. Intron boundaries in *Drosophila* highlighted in blue.

```
CLUSTAL O(1.2.1) multiple sequence alignment


Ras85D_FBpp0081600          MTEYKLVVVGAGGVGKSALTIQLIQNHFVDEYDPTIEDSYRKQVVIDGETCLLDILDTAG
Ras85D_Oncopeltus           MTEYKLVVVGAGGVGKSALTIQLIQNHFVDEYDPTIEDSYRKQVVIDGETCLLDILDTAG
                            ************************************************************


Ras85D_FBpp0081600          QEEYSAMRDQYMRTGEGFLLVFAVNSAKSFEDIGTYREQIKRVKDAEEVPMVLVGNKCDL
Ras85D_Oncopeltus           QEEYSAMRDQYMRTGEGFLLVFAVNSIKSFEDIGMYREQIKRVKDAEEVPMVLVGNKCDL
                            ************************** ******* *************************


Ras85D_FBpp0081600          ASWNVNNEQAREVAKQYGIPYIETSAKTRMGVDDAFYTLVREIRKDKDNKGRRGRKMNKP
Ras85D_Oncopeltus           HSWAVDMNQAREIAKNYSIPFVETSAKTRMGVDEAFYTLVREIRKDKEVRGKEKRKGINK
                             ** *: :****:**:*.**::***********:*************: :*:. **   : 


Ras85D_FBpp0081600          NRRFKCKML
Ras85D_Oncopeltus           NRRKRCWIL
                            *** :* :*
```

Figure S 5.13: Protein alignment made by Clustal Omega. Intron boundaries in *Drosophila* highlighted in blue.

5.2.b Molting and metamorphosis genes

Contributors: Deniz Erezyilmaz and Yuichiro Suzuki

Although the milkweed bug has been a key research model for hemimetabolous endocrine studies since the 1960's [160-164], only a handful of genes that regulate the progression through postembryonic development have been cloned from *Oncopeltus* since that time [53, 165]. We therefore searched for genes that are involved in regulation of the molt cycle, cuticle identity, or ecdysis (see also manuscript main text).

For the cytochrome P450 enzymes of the Halloween family, which synthesize the ecdysteroid hormones that trigger the molting cycle, we have found all of the key P450 genes involved in ecdysone biosynthesis, including *spook*, *phantom*, *disembodied* and *shadow* [166]. We also identified *shade*, which encodes a P450 enzyme that converts ecdysone to its active form, 20-hydroxyecdysone, at target tissues [167].

Pioneering studies in *Drosophila* of the transcriptional response to ecdysone produced a model whereby a small set of 'early genes' are transcribed, and their gene products: (1) activate a large number of 'late genes', and (2) suppress their own transcription. Subsequent cloning of early and late genes involved in the ecdysone response cascade revealed a battery of transcription factors [168] many of which belong to the nuclear receptor superfamily (see also Supplemental Note 5.1.e). Early genes with clear *Oncopeltus* orthologs include *E74*, the previously characterized *E75* [53], *ultraspiracle* (*usp*), a homolog of the mammalian Retinoid-X-receptor (RXR), and its heterodimeric binding partner, the ecdysone receptor (EcR). Another early gene with stage-specific expression, the nuclear receptor gene *E93*, is also present in the *Oncopeltus* genome. This gene was recently shown to be required for adult metamorphosis in other hemi- and holometabolous insects [169]. Among the 'delayed early genes', which are known targets of EcR/USP heterodimers in *Drosophila* [84, 168, 170, 171], we additionally identified the nuclear receptors *E78, HR3, HR4, βftz-f1, HR39*. In *Drosophila*, the hormone receptor HR4 [172] represses early genes, but activates expression of *βftz-f1*, a midprepupal gene. βFtz-F1, in turn, is required for optimal expression of early genes at the next molt in flies [173]. Finally, we discovered an ortholog encoding HR38, an orphan receptor that has been shown to heterodimerize with USP to mediate ecdysone signaling in *Drosophila* without directly binding ecdysteroids [174].

Ecdysis is driven by eclosion hormone (EH; [175]), ecdysis triggering hormone (ETH; [176]), crustacean cardioactive peptide (CCAP; [177]), and bursicon [178]. We have found genes encoding orthologs of EH, ETH, and the ETH receptor in the *Oncopeltus* genome. We were also able to identify a portion of the *ccap* ortholog in the *Oncopeltus* genome, and its putative ortholog encoding its receptor was represented in transcriptomic data.

5.2.c Structural cuticular proteins and pigmentation

Contributors: Andrew J. Rosendale, Joshua B. Benoit, Yuichiro Suzuki

Changes in the expression of specific cuticle proteins have been associated with increased stress tolerance and insecticide resistance [179, 180]. We identified 173 putative cuticle proteins using sequence motifs established by Willis [179, 181, 182] from the milkweed bug genome (Table S 5.12). Similar to other insects, the CPR family, with the RR-1 (soft cuticle), RR-2 (hard cuticle), and unclassifiable types, constituted the largest group of cuticle protein gene groups for *Oncopeltus*. The number of genes in the protein families of CPAP1, CPAP3, and TWDL were similar to the number in other insects [179, 182], but with a slight expansion in the CPF family compared to other taxa (Figure S 5.14).

Table S 5.12: Number of genes identified as putative cuticle proteins per family in the *Oncopeltus* genome.

CPR[1]								
RR-1	RR-2	Uncl	CPAP1	CPAP3	CPF	TWDL	Unclassified	Total
29	75	29	11	7	7	3	12	173

[1] Sequences that scored above the assigned cutoffs for the RR-1 and RR-2 models were classified as the corresponding type, whereas sequences with scores below the assigned cutoffs but above 0 were characterized as “unclassified” (for more information, see [181]).

Melanization of the cuticle has been suggested as critical in the prevention of excessive water loss [183]. Furthermore, understanding the physiological and molecular regulation of pigmentation synthesis has important implications for understanding the evolution of aposematic (warning) coloration, and we therefore analyzed components of the pathways responsible for the main color elements. We identified key genes associated with melanization and red pigment production for *Oncopeltus*. These include genes encoding Tyrosine hydrolase [184], Yellow [185], Dopa decarboxylase [184], Ebony [186], and at least two Phenol oxidases (Figure S 5.15). RNAi studies of these genes show that these genes play critical roles in melanin synthesis [187, 188].

In addition, the *Oncopeltus* abdomen and eyes also produce ommochrome pigments [189]. The regulation of ommochrome biosynthesis has been well characterized in *Drosophila* eye development. The key genes involved in this process (*vermilion*, *arylformamidase* or *kynurenine formamidase*, *cinnabar*, *white* and *scarlet*) were identified in the *Oncopeltus* genome. Thus, most, if not all, of the key enzymes involved in ommochrome synthesis appear to be present in the *Oncopeltus* genome.

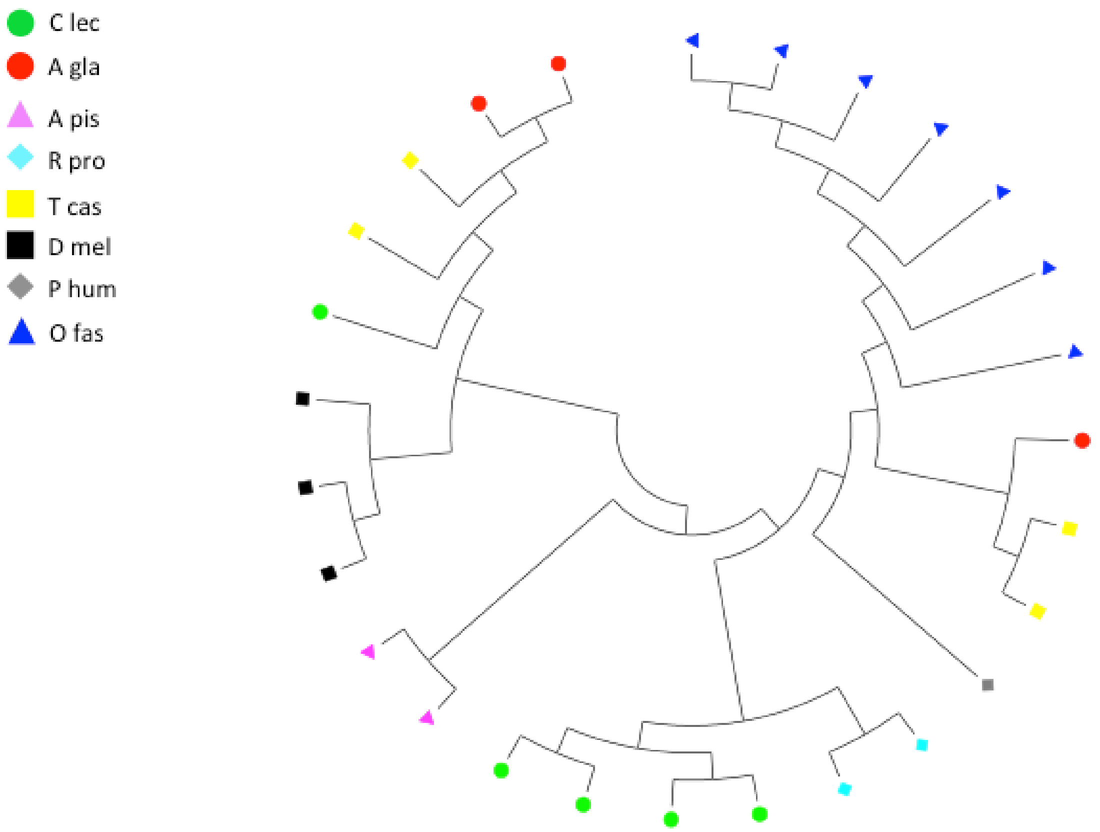


Figure S 5.14: Phylogenetic tree demonstrating relationship of CPF proteins from, *Cimex lectularius* (Clec), *Anoplophora glabripennis* (Agla), *Acyrthosiphon pisum* (Apis), *Rhodnius prolixus* (R pro), *Tribolium castaneum* (Tcas), *Drosophila melanogaster* (Dmel), *Pediculus humanus* (P hum), and *Oncopeltus fasciatus* (Ofas). The tree was constructed using the neighbor-joining method in MEGA6; Poisson correction and bootstrap replicates (2,000 replicates) were used.

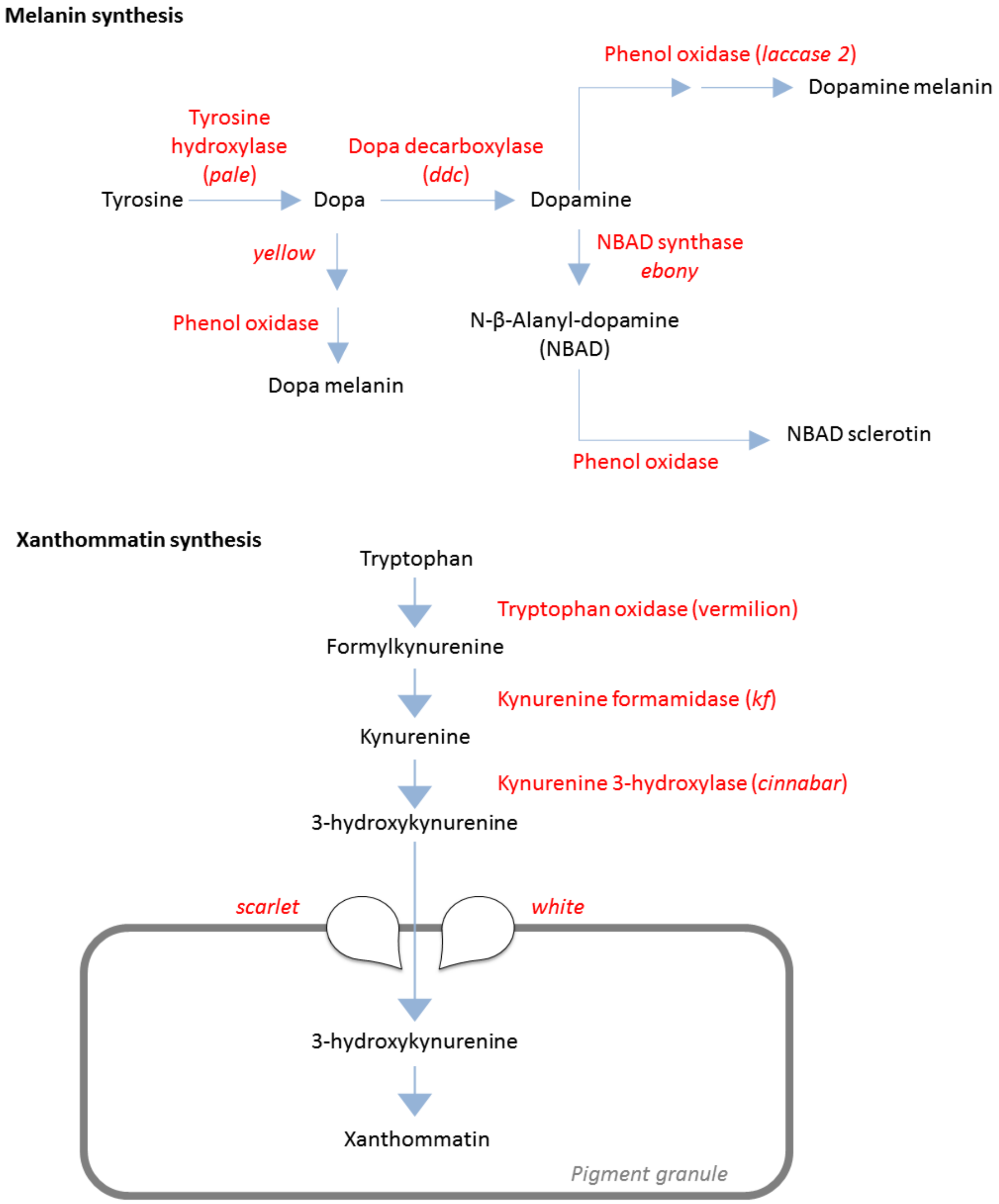


Figure S 5.15: The *Oncopeltus* genome contains most of the genes encoding enzymes involved in melanin and ommochrome synthesis. In red are all the enzymes that are encoded by genes identified in the current *Oncopeltus* assembly.

5.3 Environmental adaptations

5.3.a Stress response

Contributors: Elise M. Didion, Emily C. Jennings, Joshua B. Benoit

Antioxidant genes

Insects possess a suite of antioxidant proteins that prevent oxidative damage associated with various physiological processes. In *Oncopeltus*, 22 genes associated with the tolerance of oxidative stress were identified. Compared to other insects, the antioxidant system of *Oncopeltus* was well conserved; however, the number of catalase paralogs was greater in *Oncopeltus* than in insects such as *Drosophila* and *Tribolium* (Table S 5.13). In *Oncopeltus* three superoxide dismutases (including Cu/Zn and Mn/Fe SOD) catalyze superoxide to H_2O_2 and four catalases are present to convert H_2O_2 to water and oxygen. A total of nine genes are involved in the reduction of H_2O_2, including five peroxidoxins and four thioredoxin peroxidases. The thioredoxin system, responsible for maintaining proteins in a reduced state and scavenging reactive oxygen species, includes two genes for thioredoxin reductase. One dihydrolipoamide dehydrogenase scavenges nitric oxide and one gene for glutathione peroxidase catalyzes the breakdown of H_2O_2 and hydroperoxides. Both dual oxidase and nitric oxide synthase (one gene each) are involved in immune response. In *Oncopeltus*, antioxidants not only play an important role in normal metabolic processes, but also in the survival of oxidative stress from xenobiotic factors [190].

Aquaporin genes

Aquaporins (AQPs) impact organismal stress tolerance, specifically under periods of dehydration and cold stress, by the regulation of cellular water levels [191]. We have identified seven aquaporin genes for the milkweed bug that include those that encode for a Drosophila-integral protein (Drip), AQP2, AQP4 (two sequences), AQP5, AQP6 and Bib (Table S 5.15, Figure S 5.16). This number falls within the range of most insects (6-8) and *Oncopeltus* has members of each group previously identified from insects [191]. The number of genes is identical to those recovered from the bed bug, *Cimex lectularius* [191].

Heat shock protein genes

Heat shock proteins (Hsps) have been documented as key in relation to stress resistance under a multitude of conditions [192]. In general, there were no major expansions or retractions associated with Hsps for *Oncopeltus* when compared to other insects (Table S 5.15).

Table S 5.13: *Oncopeltus* antioxidant-associated genes (in Excel supplement).

Table S 5.14: *Oncopeltus* aquaporin-associated genes (in Excel supplement).

Table S 5.15: *Oncopeltus* heat shock proteins (in Excel supplement).

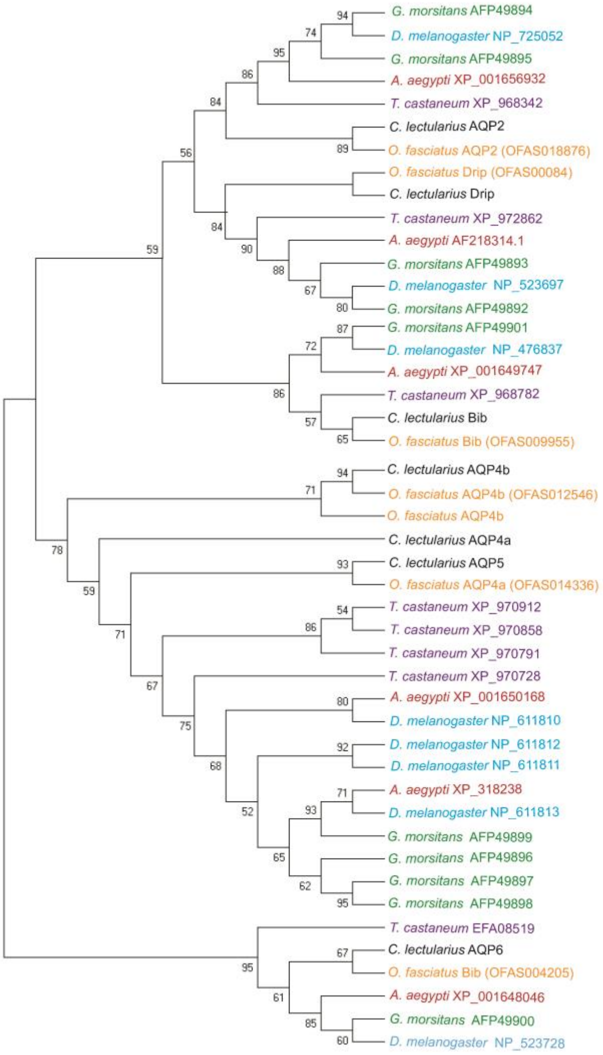


Figure S 5.16: Comparison of predicted aquaporins from other insects. Neighbor-joining tree was produced using MEGA6 using Dayhoff model and pairwise matching; branch values indicate support following 1500 bootstraps; values below 50% are omitted.

5.3.b Cytochrome P450s

Contributors: Iris M. Vargas Jentzsch, Yuichiro Suzuki, Kristen A. Panfilio

Cytochrome P450 genes encode one of the largest and most diverse enzyme families, and they can be found in the genomes of all domains of life [193]. Among an ample range of functions, P450 enzymes play a central role in the metabolism of xenobiotics and the production of hormones (see also Section 5.2.b, above). Despite their high sequence variability and substrate specificity, the overall structure of cytochrome P450 genes is highly conserved [193, 194].

As a starting point for our P450 curation, we took a sample of ten curated P450s in the bed bug *Cimex lectularius* [19], which represented all major clades in a phylogenetic tree these proteins. The selected gene models were used as queries for blastn searches into the *O. fasciatus* scaffolds, and we annotated 17 gene models from the corresponding matches (Table S 5.16). These genes, as well as additional models, were also curated as part of the gene set involved in molting and metamorphosis (see main text and Section 5.2.b, above), adding up to 20 P450 gene models. To predict which of the remaining automated models could correspond to a P450 family, we used the orthology analysis results from the OrthoDB analysis (Section 6.1). We identified all Insecta orthologous groups (OGs) containing one or more bed bug P450 gene model, and checked if these also contained *O. fasciatus* gene models (OFAS IDs). Of the bed bug P450 proteins, 69% (40 of 58) were assigned to an OG, spanning 18 distinct OGs. These 18 OGs also contained a total of 53 *O. fasciatus* proteins, including 82% (14 of 17) of our *O. fasciatus* curated P450 models (see details in Table S 5.16). Overall, between manual curation and protein orthology, we have identified a total of 58 potential P450 gene models in the *O. fasciatus* assembly.

Table S 5.16: Details on the manually curated P450 proteins in *Oncopeltus fasciatus*, and the list of OFAS IDs belonging to P450 orthologous groups identified by homology with P450s identified in *Cimex lectularius* (in Excel supplement).

5.3.c Insecticide resistance

Contributors: Lucila Traverso and Rolando Rivera Pomar

In order to provide information about the presence of genes involved in the response to insecticides in the genome of *O. fasciatus*, useful for studies on insecticide resistance, it was performed a similarity search that resulted in the annotation of 19 genes. Referring to the Glutathione S Transferase (GST) family, a microsomal GST, and one member of each class of the family (Sigma, Zeta, Omega, Delta, Theta) were annotated. Within the Carboxylesterase (COE) family, the Glutactin and Neuroligin genes were annotated, as also the sites of action of insecticides acetylcholinesterase 1 and acetylcholinesterase 2. Other genes corresponding to sites of action of insecticides that were annotated are the voltage-gated sodium channel and the GABA Receptor. In addition, it was performed the annotation of genes involved in the response to oxidative stress, also important in the response to insecticides, such as Catalase, Superoxide Dismutase 1, Superoxide Dismutase 2, Heme Peroxidase, Glutathione Peroxidase and Thiorredoxin Reductase. Another gene related to xenobiotic detoxification is UDP-glucuronosyltransferase.

5.3.d Neuropeptides and their receptors

Contributors: Lucila Traverso and Rolando Rivera Pomar

Neuropeptides are cell-to-cell signaling molecules that act as hormones, neurotransmitters, and/or neuromodulators of feeding, behavior or basic physiological processes. By homology search, we found 31 genes encoding at least 52 splicing variants of neuropeptides (see Table S 5.17). Among others, *sulfakinin* precursor genes were not found, as in *A. pisum*, but present in *R. prolixus*. We identified 32 genes encoding neuropeptide and protein hormone GPCRs. Interesting features were found in *O. fasciatus* neuropeptidome, comparing to other insects; some of the neuropeptides, Kinins, Sulfakinin and Myosuppressin, present unique sequences among the members of their families in insects. As in *R. prolixus* we did not find Arginine-Vasopressin-like peptide nor Sex peptide, but we identified a bona fide Sex Peptide receptor. The Prothoracicotropic hormone (PTTH) was also identified (see section 5.1.a). The Adipokinetic hormone (AKH) was not found, although its receptor. We cannot rule out gaps in the genome sequence, or highly diverging sequences that were not detected by our homology search.

Table S 5.17: Summary information for annotated neuropeptides and their receptors (in Excel Supplement).

5.3.e Visual genes and light detection

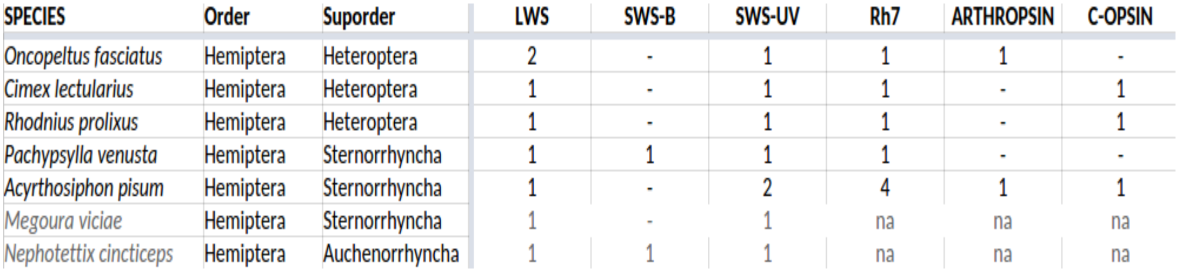
Contributors: Markus Friedrich, Jeffery W. Jones, Megan Porter

The milkweed bug has been one of the earliest model systems to study basic mechanisms of pattern formation during insect compound eye development [195]. The milkweed bug compound eye has also been used for a variety of physiological analyses [196], but more detailed analyses of spectral sensitivities across the compound eye exist for other hemipteran species, most notably the water striders and, most recently, cicadas [197-199]. Together with the first molecular study of opsin diversity in the vetch aphid *Megoura viciae* [200], these studies produced direct evidence of long wavelength sensitive (LWS), blue short wavelength sensitive (SWS-B), and UV short wavelength sensitive (SWS-UV) opsin subfamilies in the Hemiptera.

The genomic analysis of opsin diversity in *Oncopeltus* recovered two, tandemly duplicated, LWS opsin homologs and one UV-opsin, but failed to detect sequence evidence of an SWS-B opsin (Figure S 5.17 and Table S 5.18). The results from equivalent genomic analyses in further representatives of the Heteroptera (*Cimex lectularius*, *Rhodnius prolixus*) as well as representatives of other hemipteran suborders (Sternorrhyncha: *Pachypsylla venusta*, *Acyrthosiphon pisum*) suggest that SWS-B opsin was lost early in the last common ancestor of Heteroptera, while the LWS opsin duplication is unique for *O. fasciatus* in hemipteran species. While this points to a relatively recent time point of their origin, their protein sequences have substantially diverged to 71% protein sequence identity. In addition to these retinal opsin subfamilies, the *O. fasciatus* genome harbors two extra-retinal opsins. This includes a representative of the enigmatic but deeply conserved Rh7 opsin subfamily that has been described in *D. melanogaster* but not yet been functionally characterized [201]. The second extra-retinal opsin is a member of the recently characterized arthropsin subfamily, which was first reported from the *Daphnia* genome but has since then been detected in other arthropods including insects such as the pea aphid [202-204]. The otherwise sketchy detection in hemipteran genomes was complemented by the discovery of an ortholog in the genome draft of the Asian citrus psyllid *Diaphorina citri* (Figure S 5.17), cementing the evidence for arthropsin conservation in the Hemiptera. Given the close relationship of the Hemiptera to the

Holometabola and the pervasive absence of arthropsin in the well-curated genomes of the latter, it is reasonable to conclude that arthropsin was lost in the last common ancestor of the Holometabola. Understanding the phenotypic consequences of this gene loss will await the functional studies of arthropsin in versatile direct developing insect models like the milkweed bug. The more broadly conserved extraretinal c-opsin subfamily could not be detected in the milkweed bug. However, orthologs in *Cimex lectularius* and *Rhodnius prolixus* as well as the pea aphid document its conservation in the Hemiptera (Table S 5.18).

Table S 5.18: Opsin conservation in Hemiptera.



SPECIES	Order	Suporder	LWS	SWS-B	SWS-UV	Rh7	ARTHROPSIN	C-OPSIN
Oncopeltus fasciatus	Hemiptera	Heteroptera	2	-	1	1	1	-
Cimex lectularius	Hemiptera	Heteroptera	1	-	1	1	-	1
Rhodnius prolixus	Hemiptera	Heteroptera	1	-	1	1	-	1
Pachypsylla venusta	Hemiptera	Sternorrhyncha	1	1	1	1	-	-
Acyrthosiphon pisum	Hemiptera	Sternorrhyncha	1	-	2	4	1	1
Megoura viciae	Hemiptera	Sternorrhyncha	1	-	1	na	na	na
Nephotettix cincticeps	Hemiptera	Auchenorrhyncha	1	1	1	na	na	na

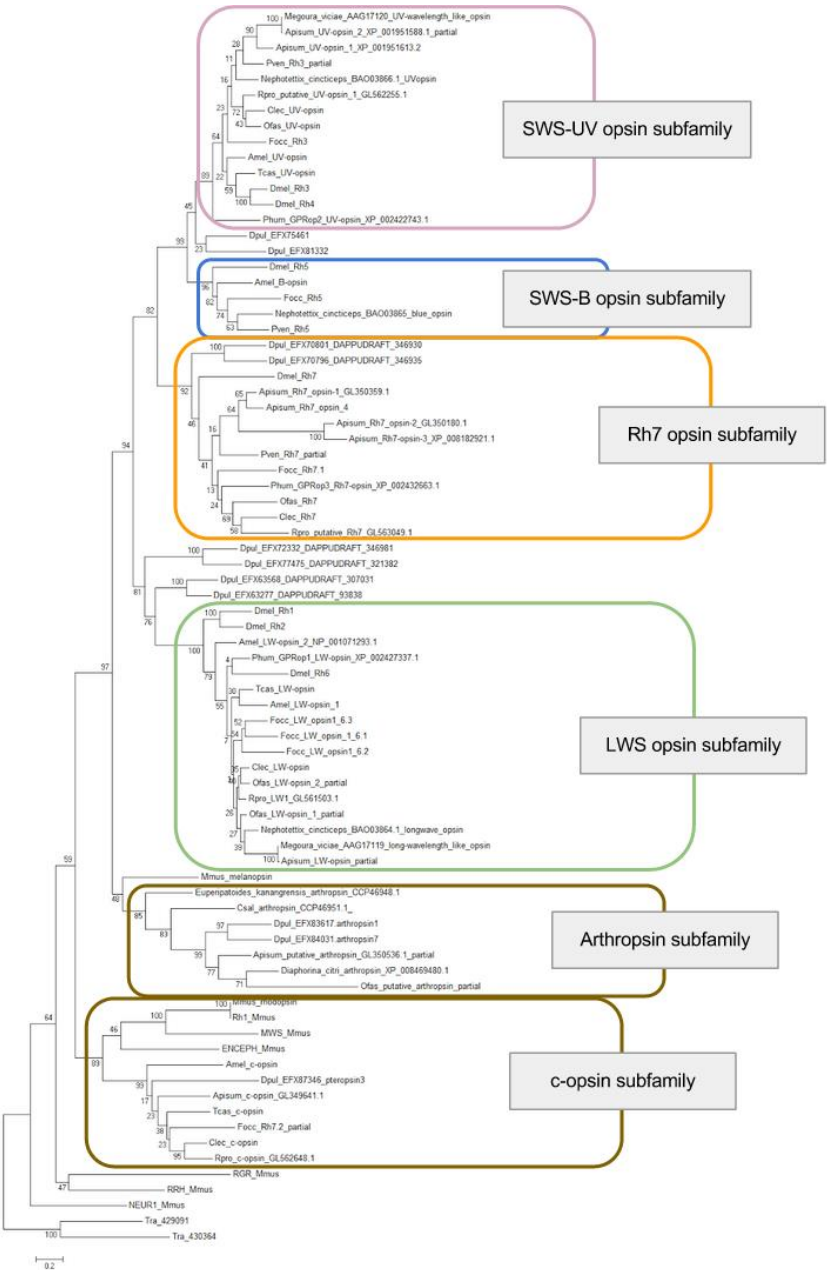

Megoura_viciae_AAG17120_UV-wavelength_like_opsin
Apisum_UV-opsin_2_XP_001951588.1_partial
Apisum_UV-opsin_1_XP_001951613.2
Pven_Rh3_partial
Nephotettix_cincticeps_BAO03866.1_UVopsin
Rpro_putative_UV-opsin_1_GL562255.1
Clec_UV-opsin
Ofas_UV-opsin
Focc_Rh3
Amel_UV-opsin
Tcas_UV-opsin
Dmel_Rh3
Dmel_Rh4
Phum_GPRop2_UV-opsin_XP_002422743.1
SWS-UV opsin subfamily
Dpul_EFX75461
Dpul_EFX81332
Dmel_Rh5
Amel_B-opsin
Focc_Rh5
Nephotettix_cincticeps_BAO03865_blue_opsin
Pven_Rh5
SWS-B opsin subfamily
Dpul_EFX70801_DAPPUDRAFT_346930
Dpul_EFX70796_DAPPUDRAFT_346935
Dmel_Rh7
Apisum_Rh7_opsin-1_GL350359.1
Apisum_Rh7_opsin_4
Apisum_Rh7_opsin-2_GL350180.1
Apisum_Rh7-opsin-3_XP_008182921.1
Pven_Rh7_partial
Focc_Rh7.1
Phum_GPRop3_Rh7-opsin_XP_002432663.1
Ofas_Rh7
Clec_Rh7
Rpro_putative_Rh7_GL563049.1
Rh7 opsin subfamily
Dpul_EFX72332_DAPPUDRAFT_346981
Dpul_EFX77475_DAPPUDRAFT_321382
Dpul_EFX63568_DAPPUDRAFT_307031
Dpul_EFX63277_DAPPUDRAFT_93838
Dmel_Rh1
Dmel_Rh2
Amel_LW-opsin_2_NP_001071293.1
Phum_GPRop1_LW-opsin_XP_002427337.1
Dmel_Rh6
Tcas_LW-opsin
Amel_LW-opsin_1
Focc_LW_opsin1_6.3
Focc_LW_opsin_1_6.1
Focc_LW_opsin1_6.2
Clec_LW-opsin
Ofas_LW-opsin_2_partial
Rpro_LW1_GL561503.1
Ofas_LW-opsin_1_partial
Nephotettix_cincticeps_BAO03864.1_longwave_opsin
Megoura_viciae_AAG17119_long-wavelength_like_opsin
Apisum_LW-opsin_partial
LWS opsin subfamily
Mmus_melanopsin
Eupenpatoides_kanangrensis_arthropsin_CCP46948.1
Csal_arthropsin_CCP46951.1_
Dpul_EFX83617.arthropsin1
Dpul_EFX84031.arthropsin7
Apisum_putative_arthropsin_GL350536.1_partial
Diaphorina_citri_arthropsin_XP_008469480.1
Ofas_putative_arthropsin_partial
Arthropsin subfamily
Mmus_rhodopsin
Rh1_Mmus
MWS_Mmus
ENCEPH_Mmus
Amel_c-opsin
Dpul_EFX87346_pteropsin3
Apisum_c-opsin_GL349641.1
Tcas_c-opsin
Focc_Rh7.2_partial
Clec_c-opsin
Rpro_c-opsin_GL562648.1
c-opsin subfamily
RGR_Mmus
RRH_Mmus
NEUR1_Mmus
Tra_429091
Tra_430364
0.2

Figure S 5.17: Nonparametric bootstrap maximum likelihood tree of opsin genes from hemipteran and other arthropod species. Protein sequences were aligned with Webprank [205]. Ambiguous alignment regions were filtered using TrimAl (v. 1.3) [206] as implemented on the Phylemon 2.0 server [207] applying User defined settings (Minimum percentage of positions to conserve: 10, Gap threshold: 0.9, Similarity threshold: 0.0, Window size: 1.0), resulting in 267 multiple alignment positions for further analysis. Maximum likelihood tree was estimated in MEGA version 6.0 [66] applying the Jones-Taylor-Thornton (JTT) model of amino acid sequence evolution and assuming Gamma Distributed substitution rates across sites with 3 categories. Numbers at branches represent non-parametric bootstrap support from 100 replications. Abbreviations of species that were investigated using genome draft annotations: Amel, *Apis mellifera*; Apisum, *Acyrthosiphon pisum*; Clec, *Cimex lectularius*; Dmel, *Drosophila melanogaster*; Dpul, *Daphnia pulex*; Focc, *Frankliniella occidentalis*; Mmus, *Mus musculus*; Ofas, *Oncopeltus fasciatus*; Phum, *Pediculus humanus*; Rpro, *Rhodnius prolixus*; Pven, *Pachypsylla venusta*; Tcas, *Tribolium castaneum*. Outgroup sequence species: Tra, *Trichoplax adhaerens*. Alignment available on request.

5.3.f Chemoreceptors

Contributor: Hugh M. Robertson

The gustatory receptor (GR) family of seven-transmembrane proteins in insects mediates most of insect gustation (*e.g.*, [208, 209]), as well as some aspects of olfaction, for example, the carbon dioxide receptors in flies [210-213]. The GR family ranges in size from a low of 6 genes encoding 8 proteins in the human body louse [214] and 10 genes in the honey bee *Apis mellifera* [215] to 215 genes encoding 245 proteins in the flour beetle *Tribolium castaneum* [57]. The other sequenced hemipteroid insects have intermediate sized families, with 77 genes in the pea aphid *Acyrthosiphon pisum* [216], 28 genes encoding 30 proteins in the kissing bug *Rhodnius prolixus* [217], and 24 genes encoding 36 proteins in the bed bug *Cimex lectularius* [19]. The GR family is more ancient than the OR family, which was clearly derived from within it, and is found in the crustacean *Daphnia pulex* [218], the mite *Metaseiulus occidentalis* [219], the tick *Ixodes scapularis* (HMR, unpublished), and many other animals (HMR, unpublished). This evolutionary history is reminiscent of the more recently described ionotropic receptors (IRs) [220-222], some of which also probably function in gustation.

Methods

TBLASTN searches of the genome assembly were performed using *Cimex*, *Rhodnius*, *Acyrthosiphon*, *Pediculus*, and *Drosophila* proteins as queries, and gene models were manually assembled in a text editor (TEXTWRANGLER). Iterative searches were conducted with each new *Oncopeltus* protein as query until no new genes were identified in each major subfamily. All of the *Oncopeltus* genes and encoded proteins are detailed in Table S 5.19-5.21, and all *Oncopeltus* proteins are provided in a supplementary FASTA file (Additional File 3). The gene models for these have been updated in the Apollo genome browser as best possible (however the OR and IR updates are not included in OGSv1.2 and will only appear in later releases of the genome annotation).

The somewhat fractured nature of the genome assembly meant that there are many gene fragments that were not annotated in each family, however we endeavored to include as many as we believe represent intact genes, even if they did not always encode more than 50% of a typical receptor protein. In some cases, for example OrCo, some ORs (Or40, 71, 110, and 114), and some IRs (Ir25a, 40a, 75c, and 75j), the ends of two scaffolds have interdigitated contigs that encode alternating exons or sets of exons, and these and a few other examples of genes that span scaffolds were built. Proteins were aligned in CLUSTALX v2.0 [223] using default settings, and problematic gene models were refined in light of these alignments.

Most *Cimex*, *Rhodnius*, *Acyrthosiphon*, and *Pediculus* GRs (excluding PhOr11 and 12 which are too short to include and ApisGr12 and 15 as they are so divergent they disrupt the alignments), as well as select other insect GRs, were included in the alignments for the phylogenetic analysis. The final alignments were trimmed using the "gappyout" option in TRIMAL v4.1 [206]. Maximum likelihood analysis was performed using PHYML v3.0 [224] with default settings. The tree figures were prepared using FIGTREE v1.4 (http://tree.bio.ed.ac.uk/software/figtree/) and Adobe Illustrator.

Results

The OR family

The OR family consists of the highly conserved and generally single-copy OrCo gene and 102 specific OR genes. The OrCo gene, like some of the IR genes below, is split across two scaffolds with exons on the ends of the scaffold separated by gaps and interdigitated with each other (Table S 5.19), as are four of the specific OR genes. The specific OR genes are generally a lot smaller, and most were intact, however many have terminal or internal exons missing. Some of these could be repaired using the available whole body RNAseq, specifically raw reads that did not map to the assembly but are available in the Short Read Archive at NCBI. One OR gene was modeled as alternatively spliced with two protein products (Or88), so the total of potentially encoded ORs is 121, and only two of these is pseudogenic, leaving 119 potentially active ORs. There are, however, several additional fragments of OR genes in the assembly, some of which may represent intact genes, while some incomplete models might actually be pseudogenes.

The phylogenetic tree is rooted with the conserved OrCo proteins (Figure S

5.18), and reveals that there is a complex web of relationships of the specific ORs across these four hemipteroids. Most of the *Pediculus* and *Acyrthosiphon* ORs form distinct clades or subfamilies, while *Oncopeltus*, *Cimex*, and *Rhodnius* ORs sometimes reveal potentially orthologous relationships. Nevertheless, the majority of large OR expansions are species-specific, and these are mostly in *Oncopeltus* and *Rhodnius,* with *Cimex* having only a few small expansions (Figure S 5.18). The largest *Oncopeltus* expansion is 36 genes (OfasOr24-59). Unlike the pea aphid repertoire, which revealed several recent expansions exhibiting signals of positive selection on some amino acid positions [216], the *Oncopeltus* ORs appear to be older expansions with relatively longer branches to most proteins.

The GR family

The GR gene set consists of 115 models, encoding 169 proteins (Table S 5.20), considerably larger than that of many other insects. Of these only 9 are clearly pseudogenic, but many models are currently missing termini or internal regions in gaps in the assembly, so their status remains uncertain. There are many genes modeled as alternatively spliced, in a fashion common to the GR family in several other insects, that is, with two or more long first exons spliced into shared C-terminal exons, although in most cases in the absence of transcriptome evidence these models remain hypothetical. The MAKER modeling had access to all available insect GRs in GenBank, for comparative information, and succeeded in building at least partial gene models for 20 of these 115 loci, while AUGUSTUS had partial models for many more, nevertheless all models required at least one change.

Oncopeltus has seven genes encoding proteins related to the highly conserved carbon dioxide receptors of flies and other insects [225], and these were named Gr1-7 (Figure S 5.19). They cluster phylogenetically with four similar proteins from the *Cimex*. This carbon dioxide lineage is absent from all Hymenoptera sequenced to date, as well as *Acyrthosiphon*, *Pediculus*, and *Rhodnius* [214, 216, 217], so appears to have been lost repeatedly. A large related subfamily expansion was discovered in the termite *Zootermopsis nevadensis* [226], indicating that this gene lineage is indeed ancient in insects. It remains to be shown whether they participate in perception of carbon dioxide.

Oncopeltus has three genes encoding candidate sugar receptors, named Gr8-10

(Figure S 5.19). This subfamily is absent from the blood-sucking hemipterans examined to date (*Pediculus*, *Rhodnius*, and *Cimex*), but is present as six genes in the plant-attacking pea aphid [19, 214, 216, 217]. The only other conserved Gr is Gr11, which is an ortholog of the DmGr43a fructose receptor [227], and is also present as a single ortholog in the other hemipteroids, except *Pediculus*.

The remaining *Oncopeltus* GRs (12-115) are quite divergent from any of the conserved GRs, and consist of several gene lineages, most of which are closer to the equivalent *Cimex* and *Rhodnius* GRs than *Pediculus* or *Acyrthosiphon* (Figure S 5.19). These include all of the alternatively spliced models. The long branches to most of these proteins are similar to those to the *Cimex* and *Rhodnius* proteins, and in stark contrast to most of the aphid GRs, which form several recently expanded gene subfamilies that reveal evidence of positive selection of amino acids indicative of adaptive divergence [216]. Most of the other *Drosophila* GRs are implicated in perception of bitter tastants [228, 229], however it is hard to be confident of such a function for these *Oncopeltus* GRs and their *Cimex/Rhodnius* relatives. It is nevertheless somewhat surprising that *Oncopeltus* has such a large repertoire of GRs. It implies that they are adapted to sense a wide range of bitter taste chemicals, presumably employed in selecting suitable host plants.

The IR family

The IR family consists of at least 37 genes (Table S 5.21and Figure S 5.20). It contains two highly conserved receptors (Ir8a and 25a) that are closely related to the ionotropic glutamate receptors from which they evolved (see [221, 226]), and which serve as co-receptors with the other IRs, along with Ir76b [222]. Another group of receptors (Ir21a, 40a, 68a, and 93a) are present in fairly conserved single orthologs, as is the case for most other insects, and several of these have recently been shown to be involved in perception of temperature and humidity in Drosophila [230-232]. The Ir41 and 75 lineages consist of multiple genes in most insects, and in *Oncopeltus* there are three and ten genes, respectively. In Drosophila they are involved in perception of acids and amines [233, 234]. Following an approach begun with the termite *Zootermopsis nevadensis* [226], and applied to *Rhodnius* and *Cimex*, the conserved IRs are named for their *Drosophila* orthologs, the Ir41 and 75 lineages are named with suffices a-c and a-j, while the divergent genes below are numbered from 101 to avoid any confusion with the Drosophila IR names which only go to 100a (because they were named for their cytological location in that genome). There are several lineages of highly divergent receptors, and in *Oncopeltus* there are 17 of these, Ir101-117, dispersed around the tree with relationships with similarly divergent IRs in the other hemipteroid insects. Ir113-117 have multiple exons like the more conserved IRs, while the other two divergent lineages are primarily intron-less genes, with three of them acquiring idiosyncratic introns. This pattern of mostly intron-less genes is typical for the divergent IRs in other insects, and if the roles of the divergent IRs in Drosophila, and in particular the large Ir20a clade, are a guide [222, 235, 236], they might function as taste receptors.

These *Oncopeltus* IR genes were particularly difficult to annotate, despite the fact that the relatively conserved genes have clear homologs in other species, and most of the divergent receptors are intron-less. The fractured nature of the assembly, combined with the fact that they are generally large genes spanning 10-30 kb and have many short exons, caused most of the problems, ranging from exons missing in gaps, to genes split across scaffolds, including three instances of exons on different scaffolds being interdigitated with each other, and sometimes with exons in incorrect order. Luckily the combination of relatively conserved sequence and often substantial RNAseq support allowed construction of fairly complete models for most genes, however many of them cannot be properly modeled in the Apollo genome browser. Thus 14 of the 23 genes with multiple introns required repair of the assembly, commonly using raw RNAseq reads (the RNAseq reads mapped in Apollo are not useful when an exon is missing or a gene is split across scaffolds or otherwise misassembled), as well as raw genome reads (Table S 5.21). Official models existed for at least parts of most of the relatively conserved genes, but the divergent genes, despite commonly being intronless, were usually only represented in the AUGUSTUS gene set. Altogether 22 new gene models were created (Table S 5.21).

Thirteen of the IR models remain incomplete, primarily at the N-terminus, in large part because their N-termini are commonly highly divergent hence difficult to identify using TBLASTN searches, are hard to be confident about from de novo gene building, and the available RNAseq seldom extends to the N-terminus. Just two genes are clear-cut pseudogenes (both divergent IRs), at least in the assembled genome, but both are caused by single problems (one stop codon and one frame shift), so both might be intact genes in other populations or even pseudo-pseudogenes [233]. Finally, the Ir113-117 lineage of genes is particularly difficult to model, being divergent and having little to no RNAseq support. TBLASTN searches with these five proteins reveal at least three more related genes that are present only as a few exons each and are too difficult to model in the absence of RNAseq support, so the total number of IRs in *Oncopeltus* is likely to be 40.

This repertoire of IRs is slightly larger than the other sequenced hemipteroids, with 19 in the pea aphid [32], 33 in the kissing bug [217], 30 in the bed bug [19], and 14 in body louse [214], although other insects commonly have more IRs, such as *Drosophila melanogaster* with 65 [221] and the termite *Zootermopsis nevadensis* with 150 [226],. The only obvious differences are in the numbers of the divergent lineages, which at least in *Drosophila* are implicated in gustation rather than olfaction [222, 235, 236]. This small potential gustatory receptor IR expansion in *Oncopeltus* is in line with the expansion of its GR family.

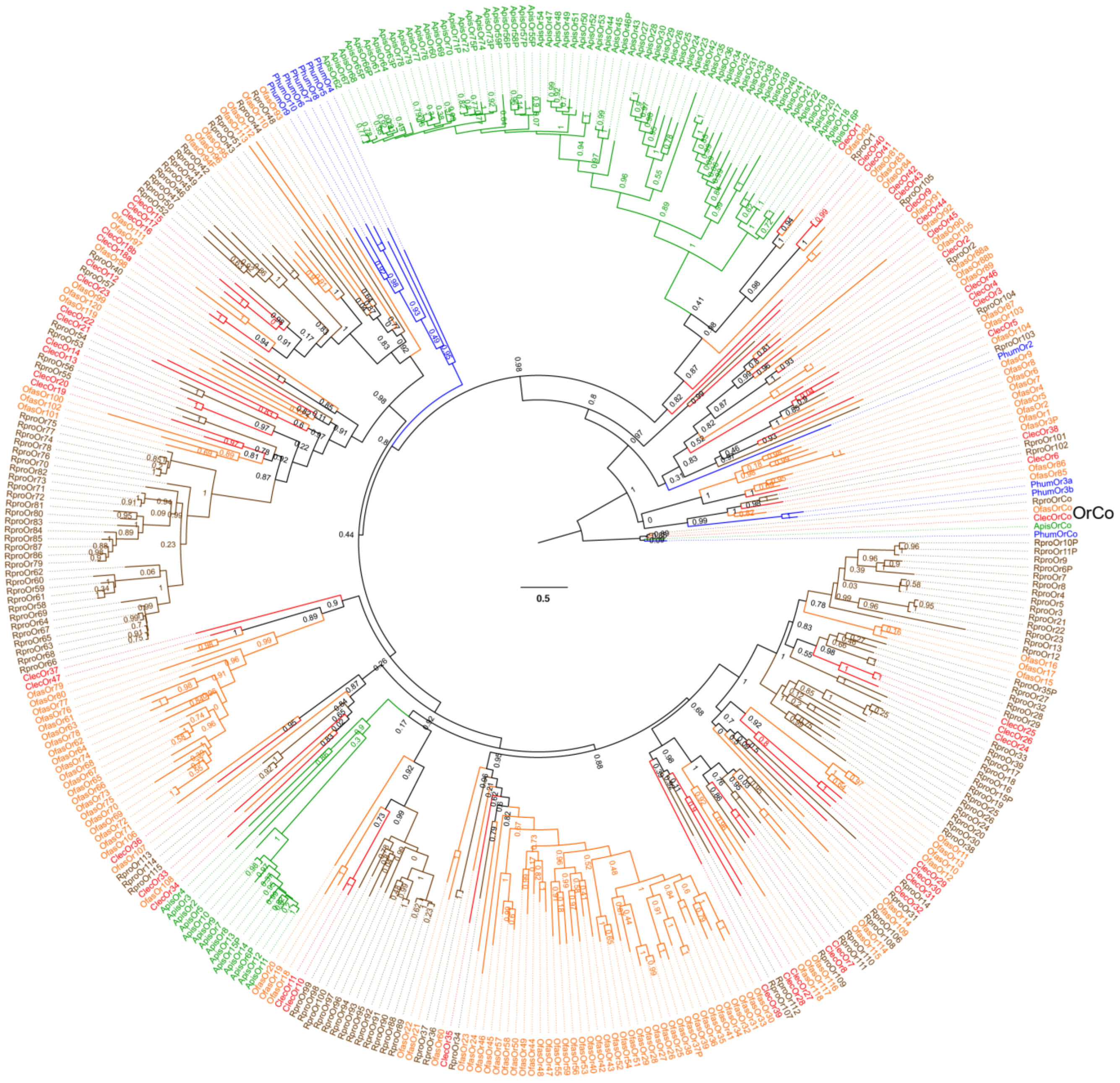


Figure S 5.18: Maximum likelihood phylogenetic tree of the *Oncopeltus*, *Cimex*, *Rhodnius*, and *Pediculus* ORs. The tree was rooted with the highly conserved and basal OrCo proteins. The *Oncopeltus*, *Cimex*, *Rhodnius*, *Acyrthosiphon*, and *Pediculus* gene/protein names are highlighted in orange, red, brown, green, and blue, respectively, as are the branches leading to them to emphasize gene lineages. Lowercase letters distinguish different protein isoforms from a single gene. Suffixes after the gene/protein names are: P – pseudogene; N – N-terminus missing; C – C-terminus missing; I – internal sequence missing; F – sequence fixed with raw reads; J – gene model joined across scaffolds.

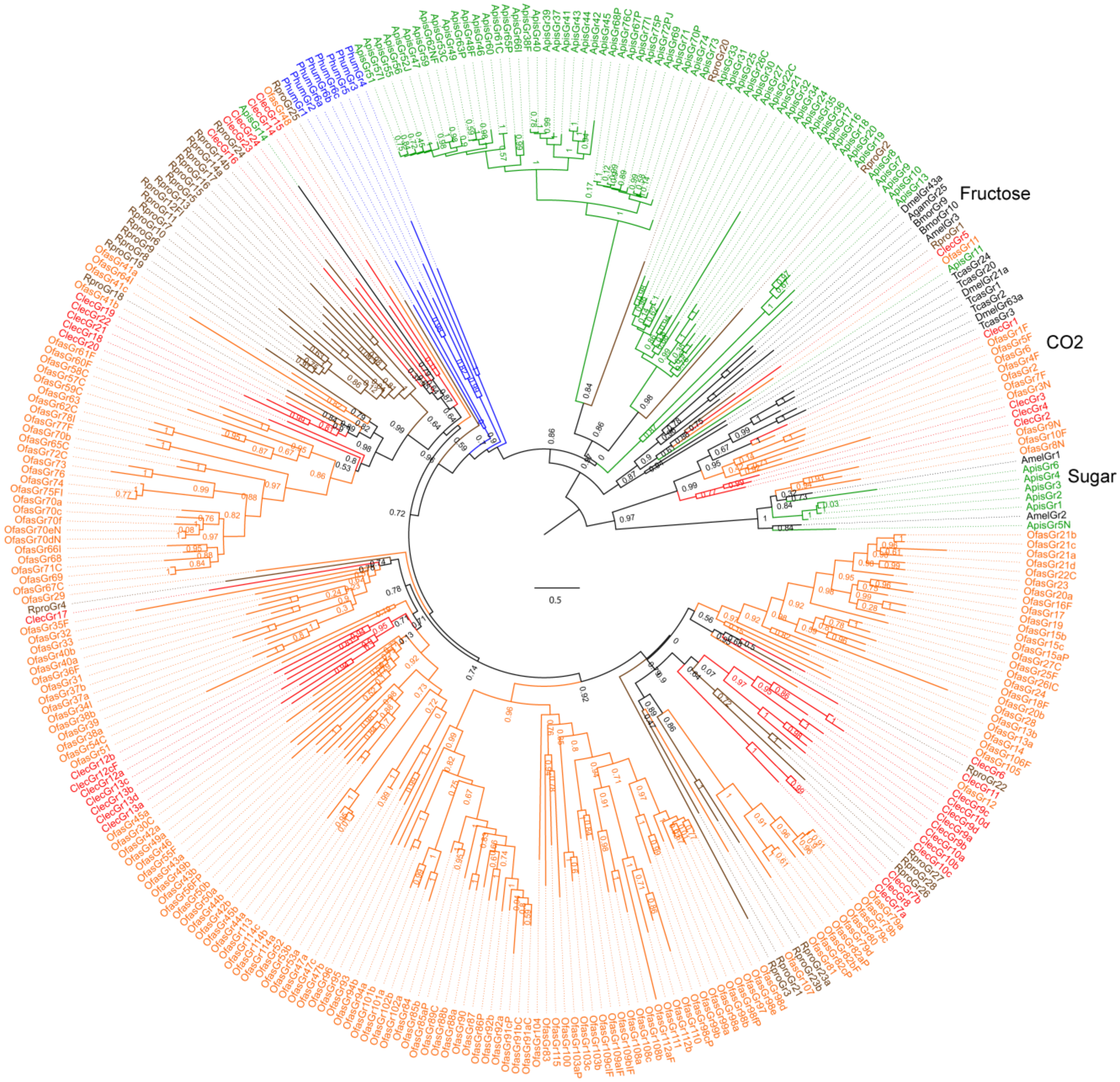


Figure S 5.19: Maximum likelihood phylogenetic tree of the *Oncopeltus*, *Cimex*, *Rhodnius*, *Acyrthosiphon*, and *Pediculus* GRs, and representative GRs from other insects. The tree was rooted with the candidate carbon dioxide and sugar receptors, subfamilies of conserved GRs that are divergent from the other GRs. See legend for Figure S 5.18 for other details.

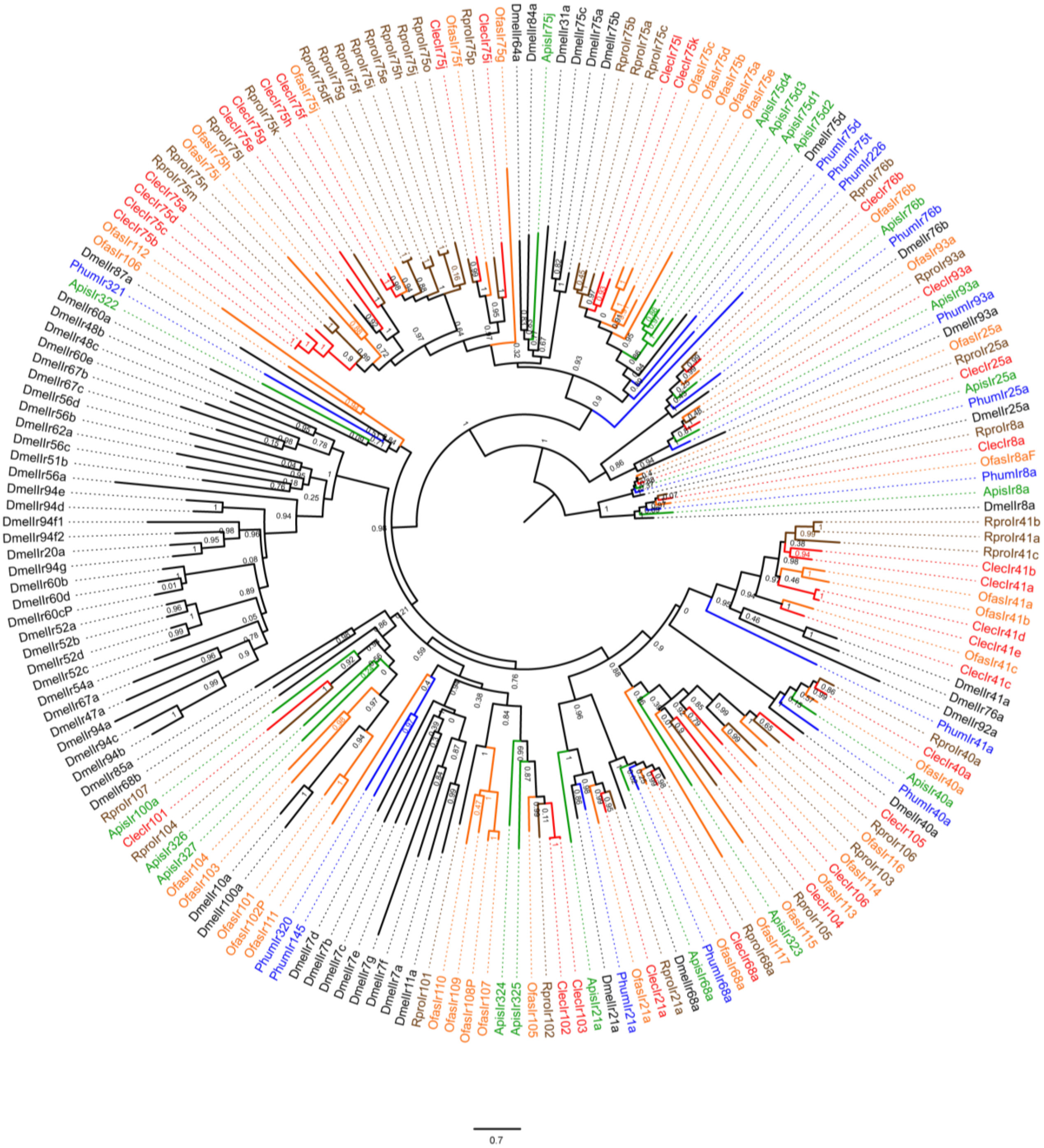


Figure S 5.20: Maximum likelihood phylogenetic tree of the *Oncopeltus*, *Cimex*, *Rhodnius*, *Acyrthosiphon*, *Pediculus*, and *Drosophila* IRs. The tree was rooted with the conserved and basal Ir8a/25a proteins. See legend for Figure S 5.18 for other details.

Table S 5.19: Details of OfasOr family genes and proteins (in Excel Supplement).

Table S 5.20: Details of OfasGr family genes and proteins (in Excel Supplement).

Table S 5.21: Details of OfasIr family genes and proteins (in Excel Supplement).

Additional file 3: Chemoreceptor sequences in FASTA format. (TXT file)

Note that due to the fragmented nature of the genome assembly, the chemoreceptor gene models were predicted independently from the assembly by manually combining the gene predictions with RNA-seq data. Thus, some gene locus coordinates provided in the supplementary tables (S5.19-21) may vary from the ones in theOGSv1.2 gff file. However, the complete curated protein sequences for all these models are provided in the supplementary fasta file ("Ofas-chemoreceptors-protein-seqs.fasta"), with sequences for 121 Or, 169 Gr, and 37 Ir proteins. In the file, suffixes after the gene/protein names are: C - C-terminus missing; F – assembly sequence fixed with raw reads or other repairs; I – internal region missing; J – model joined across scaffolds; N – N-terminus missing; P – pseudogene (Z – stop codon and X- frameshift or other major problem).

5.4 Molecular machinery

5.4.a Gene silencing machinery (RNAi, miRNA, piRNA)

Contributors: Yi-min Hsiao, Hsiao-ling Lu, Chun-che Chang

In *Oncopeltus* we have identified sequences encoding conserved components of the different post-transcriptional gene silencing pathways: RNA interference (RNAi), micro-RNA (miRNA), and piwi-interacting RNA (piRNA). These genes include *Drosha*, *partner of drosha* (*pasha*), *Dicer 1* (*Dcr1*), *Dcr 2*, *Argonaute 1* (*Ago1*), *Ago2*, *Ago3*, and *piwi/aubergine* (*piwi/aub*) (Table S 5.22). Like *Drosophila melanogaster* (fly) and *Tribolium castaneum* (beetle), where most of these genes are not duplicated, *Oncopeltus* only has a single copy of each gene (Table S 5.23), which is also the case in the fellow hemipterans the bed bug [19] and the soybean aphid [237]. In contrast, *Acyrthosiphon pisum* (aphid) has a notable expansion of the miRNA and piRNA machinery [32, 149, 238].

Table S 5.22: Major components RNAi/miRNA machinery in *Oncopeltus fasciatus.*

Gene name	**Gene abbreviation**	**Number of copies found in *Oncopeltus***	**Comments**
Gene name	**Gene abbreviation**	**Number of copies found in *Oncopeltus***	**Comments**
Drosha	*Drosha*	1	
partner of drosha	*pasha*	1	
Dicer 1	*Dcr1*	1	CDS splits in three different scaffolds
Dicer 2	*Dcr2*	1	CDS splits in two different scaffolds
Argonaute 1	*Ago1*	1	
Argonaute 2	*Ago2*	1	
Argonaute 3	*Ago3*	1	
piwi	*piwi*	1	See annotation of germline genes.

Table S 5.23: Orthologous numbers of RNAi machinery components in four insect species.

Gene	Fruit fly	Beetle	Pea aphid	Milkweed bug
Drosha	1	1	1	1
pasha	1	1	4	1
Dcr1	1	1	2	1
Dcr2	1	1	1	1
Ago1	1	1	2	1
Ago2	1	2	1	1
Ago3	1	1	2	1
piwi	1	1	8	1*

*See also the annotation of germline genes in Table S 5.7.

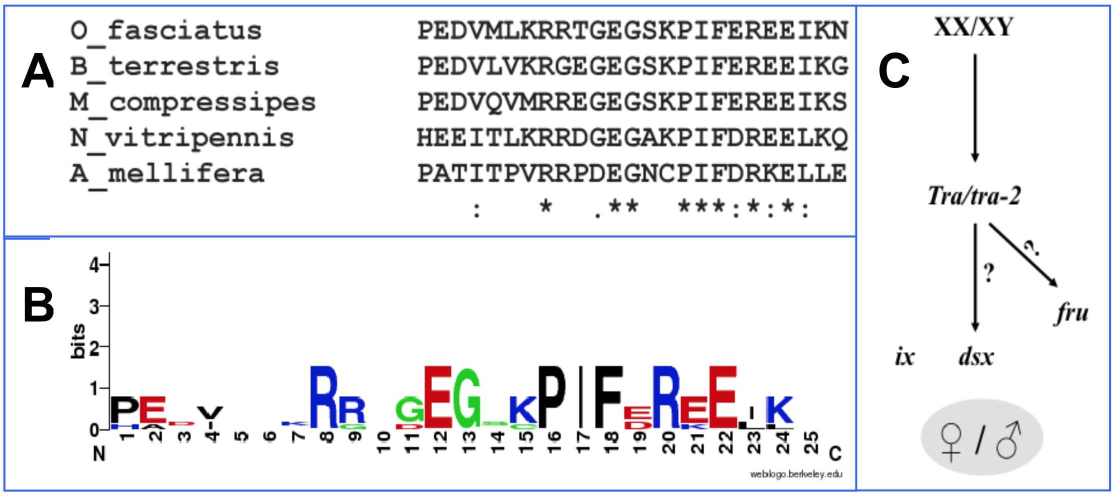
5.4.b Sex determination and dosage compensation

Contributors: Subba (Reddy) Palli, Jayendra N. Shukla

O. fasciatus is a male heterogametic (XX-female and XY-male) insect [239], but the sex determination signal remains unclear. The most downstream gene of *Drosophila* sex determination cascade, *doublesex* (*dsx*), is conserved in all the insects studied so far [240] as inferred from the studies in various groups of holometabolous insects [241], and is the founding member of proteins with DM domains [242-244]. However, the *Oncopeltus* genome contains DM superfamily genes in which no Oligomerization (OD2) domain was identified. This is similar to the case of two other hemipterans, *Acyrthosiphon pisum* and *Rhodnius prolixus*, whose genomes have been sequenced recently [32]. Functional analysis of DM domain genes in hemimetabolous insects is required to ascertain their potential roles in sex determination.

Homolog of *transformer* (*tra*), the upstream regulators of *dsx* in holometabolous insects (absent in lepidopterans and basal dipteran lineages, *i.e.*, mosquitoes [241]) is present in *Oncopeltus* genome. Interestingly, the partial *tra* homolog obtained showed high sequence conservation in the auto regulation domain of the hymenopteran *tra* sequence (Figure S 5.21 A,B). Whether *Oncopeltus tra* is spliced in a sex specific manner and regulates the splicing of its own and *dsx* pre-mRNA needs further investigation. Among other core sex determination genes homologs, *transformer-2* [245], *intersex* [246], *fruitless* [247] and *P-element somatic inhibitors* [248] have also been identified in the *Oncopeltus* genome. Figure S 5.21 C represents the schematic diagram of the proposed sex determination cascade in *Oncopeltus fasciatus*.

The dosage compensation mechanism in *Drosophila* is known to equalize the dose of X chromosome linked transcripts. Out of five major genes for dosage compensation in *Drosophila* (*msl-1*, *msl-2*, *msl-3*, *mle* and *mof*) [249] homologs of *msl-2*, *msl-3*, *mle* and *mof* have been identified in the *Oncopeltus* genome, suggesting the existence of a functional dosage compensation pathway.



```
O_fasciatus          PEDVMLKRRTGEGSKPIFEREEIKN
B_terrestris         PEDVLVKRGEGEGSKPIFEREEIKG
M_compressipes       PEDVQVMRREGEGSKPIFEREEIKS
N_vitripennis        HEEITLKRRDGEGAKPIFDREELKQ
A_mellifera          PATITPVRRPDEGNCPIFDRKELLE
                        :   *  .**  ***:*:*:
```

Figure S 5.21: **(A)** CLUSTAL alignment (http://www.ebi.ac.uk/Tools/msa/clustalo/) of putative *tra* auto regulation domain of *O. fasciatus* and that of other hymenopteran insects (*B. terrestris*, *M. compressipes* , *N. vitripennis*, *A. mellifera*) (please see [250]). **(B)** WebLogo (http://weblogo.berkeley.edu/logo.cgi) of putative *tra* auto regulation domain of *O. fasciatus* and that of other hymenopteran insects (*B. terrestris*, *M. compressipes*, *N. vitripennis*, *A. mellifera*) displaying a deep conservation in the sequence (see [250]). **(C)** Putative sex determination pathways in *Oncopeltus fasciatus* based on the sex determination cascade of holometabolous insects [251]. The homologs of sex determining genes *doublesex* (*dsx*), *transformer* (*tra*), *transformer-2* (*tra-2*), *fruitless* (*fru*) and *intersex* (*ix*) are present in the genome of *Oncopeltus fasciatus*.

5.4.c Epigenetic machinery

Contributor: Elizabeth J. Duncan

DNA methylation and post-translational modifications of histones are key regulators of chromatin structure and of gene expression, and these epigenetic systems have been associated with environmental responsiveness and phenotypic plasticity [252].

DNA methylation

DNA methylation is associated with gene silencing in vertebrates [253] and alternative splicing of mRNA [19, 254-256]. Functional DNA methylation systems have been demonstrated in a few insects including the honeybee *Apis mellifera* [255, 257, 258] and the hemipteran *Acyrthosiphon pisum* (the pea aphid) [259].

Like these insects, *Oncopeltus* also appears to have an intact DNA methylation system. The *Oncopeltus* genome encodes two copies of the maintenance methyltransferase *Dnmt1*, the *de novo* DNA methyltransferase *Dnmt3*, and a copy of *Tet1* (*Ten-eleven translocation methylcytosine dioxygenase 1*) that has been implicated in removing methylation marks by converting 5′ methylcytosine to 5′ hydroxymethylcytosine [260, 261].

In insects and other invertebrates DNA methylation (the addition of a methyl group to a cytosine residue in a CpG context) occurs predominately on gene bodies (exons and introns) [262-264]. In vertebrates DNA methylation occurs in CpG islands in the promoter regions of genes and is associated with gene silencing [265]. Gene body methylation also occurs in vertebrates and it has been shown to be as abundant as methylation in CpG islands [262]. The function of gene body methylation is currently unknown, however gene body methylation has been correlated with active transcription in a wide range of species [264], has been implicated in alternative splicing [254, 255] and regulating chromatin organization [266].

Over evolutionary time methylation of cytosine residues leaves them susceptible to deamination to uracil, which is repaired as a thymine, leaving methylated genes with a relatively low CpG content [267]. The CpG content can be measured mathematically as: $\mathrm{CpG}_{[o/e]}$ = number of CpG dinucleotides in a gene/[the number of C nucleotides]×[the number of G nucleotides]. In animals without DNA methylation, such as *Drosophila melanogaster*, a unimodal distribution of CpG content is observed (Figure S 5.22 A). In contrast, a bimodal distribution is seen in insects with an active methylation system (for instance *Apis mellifera* and *Acyrthosiphon pisum*, Figure S 5.22 B,C).

In the honeybee it has been found that there is a significant correlation between genes that have low CpG$_{[o/e]}$ (*i.e.*, are predicted to be historically methylated) and genes that are currently methylated in the brains of honeybees [255], confirming that CpG$_{[o/e]}$ is a good predictor of genes that are currently methylated.

Methods

Gene body and intragenic sequences were extracted from the predicted *Oncopeltus* gene set (OGS v1.1) using CLC Genomics Workbench (version 7). For analysis of whole genome and intragenic regions, the sequences were split into 1000-nt non-overlapping fragments, and nucleotide and dinucleotide content of gene body sequences and whole genome sequences were calculated as in previous analyses [268]. The number of components in these distributions was estimated in R (www.r-project.org) using mclust [269] model-based clustering. The best fitting model was identified among several non-nested models using Bayesian information criteria (BIC).

Results

Analysis of the CpG content of *Oncopeltus* gene bodies identifies a bimodal distribution of CpG[o/e] values, with a 'low' CpG[o/e] peak centered around 0.31 (55% of genes) and a 'high' CpG$_{[o/e]}$ peak around 0.71 (Figure S 5.22 D). This bimodal distribution is not observed with any other dinucleotide combination (Figure S 5.23). The presence of a low CpG$_{[o/e]}$ peak is consistent with historical DNA methylation as seen in other species (*e.g.*, Figure S 5.22 B,C). However, both peaks have relatively low CpG content, as almost all genes have a CpG[o/e] of less than one. This differs from other species where the high CpG[o/e] gene fraction has CpG$_{[o/e]}$ values greater than one (Figure S 5.22). This is consistent with an active DNA methylation system in *Oncopeltus*.

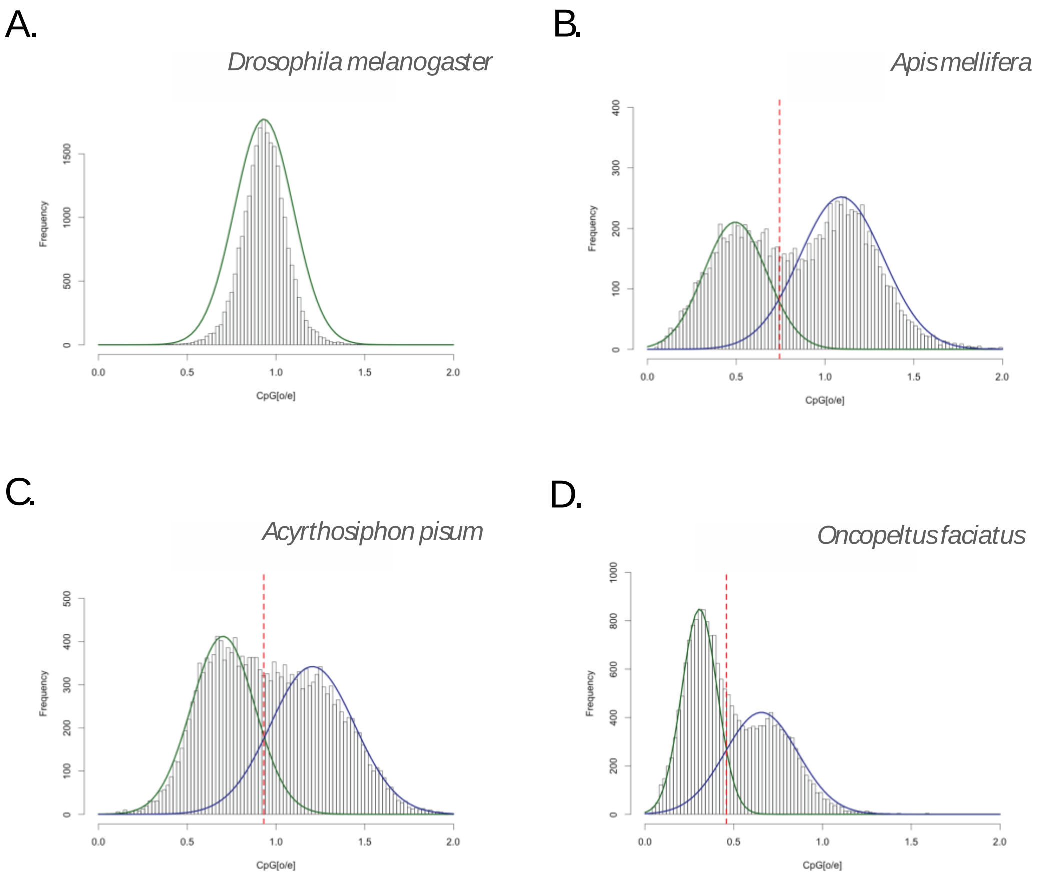


Figure S 5.22: Frequency histogram of CpG[o/e] observed in insect gene bodies. The y-axis depicts the number of genes with the specific CpG[o/e] values given on the x-axis. The distribution of CpG[o/e] in *Drosophila*, which does not have DNA methylation, is unimodal **(A)**. In contrast in species where DNA methylation is active, such as *Apis mellifera* **(B)** or the pea aphid **(C)** the distribution of CpG[o/e] is bimodal with genes that have been historically methylated having a lower than expected CpG content (green peak). *Oncopeltus* **(D)** genes also display a bimodal distribution of CpG content consistent with an active DNA methylation system in this species.

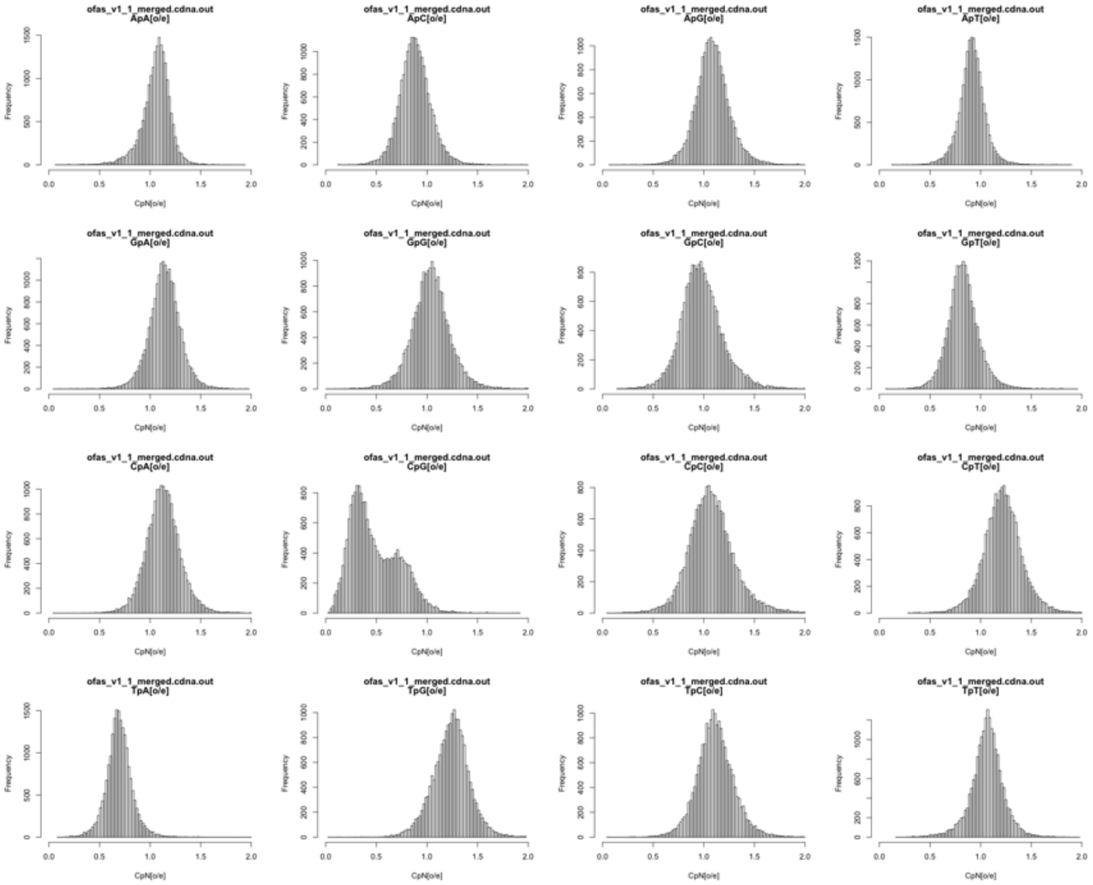


Figure S 5.23: Frequency histogram of the observed vs. expected ratio for all dinucleotide combinations in *Oncopeltus* gene bodies. The y-axis depicts the number of genes with the specific dinucleotide [o/e] values given on the x-axis.

To determine if this pattern of CpG frequency was unique to gene bodies, as it is in other species, intragenic regions were extracted and split into 1000-nt fragments. CpG[o/e] analysis of these intragenic regions demonstrates that a small number of intragenic DNA fragments (16.8%) fall into a region of the DNA with lower than expected CpG content (Figure S 5.24). This is not observed for any other dinucleotide combination (Figure S 5.25). This may be evidence for historical DNA methylation acting on regions of DNA that do not code for protein coding genes, regions encoding genes that have not yet been incorporated into the official gene set,, or other as yet unknown mechanisms that deplete CpG dinucleotide content independent of the GC content of the DNA.

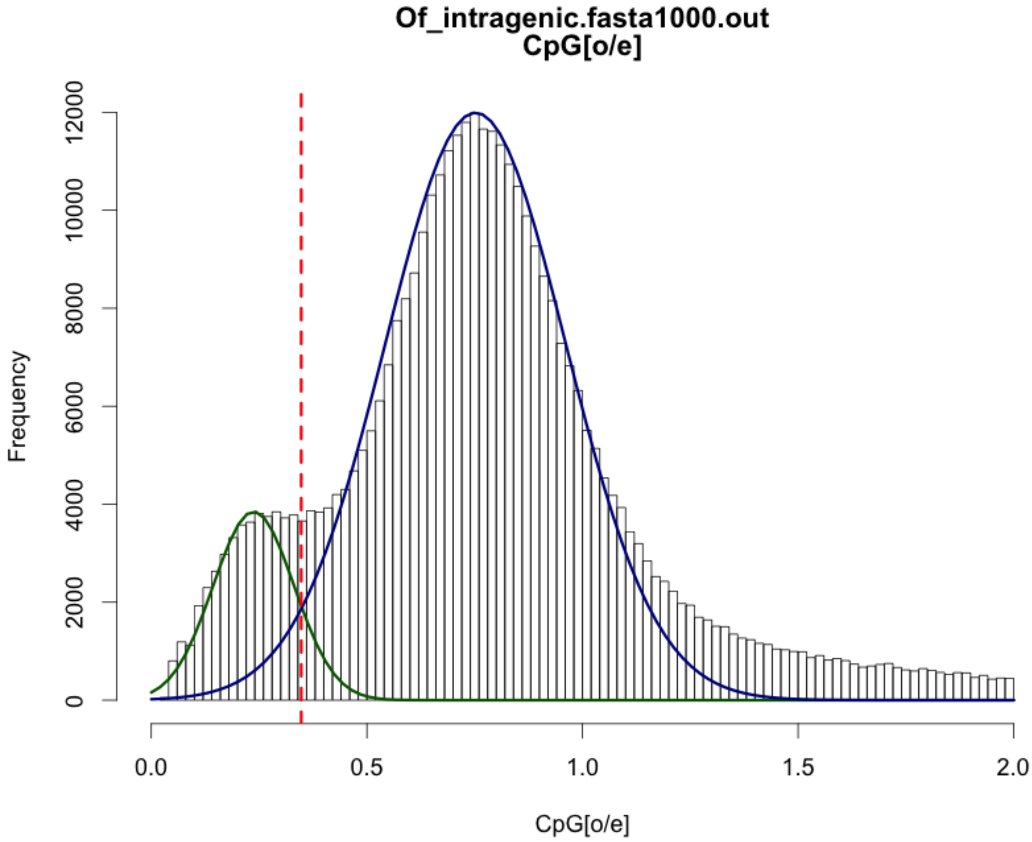


Figure S 5.24: Frequency histogram of CpG[o/e] observed in intergenic regions in *Oncopeltus*. The y-axis depicts the number of genes with the specific CpG[o/e] values given on the x-axis. The distribution is also bimodal suggesting that regions of the *Oncopeltus* genome are very CpG poor.

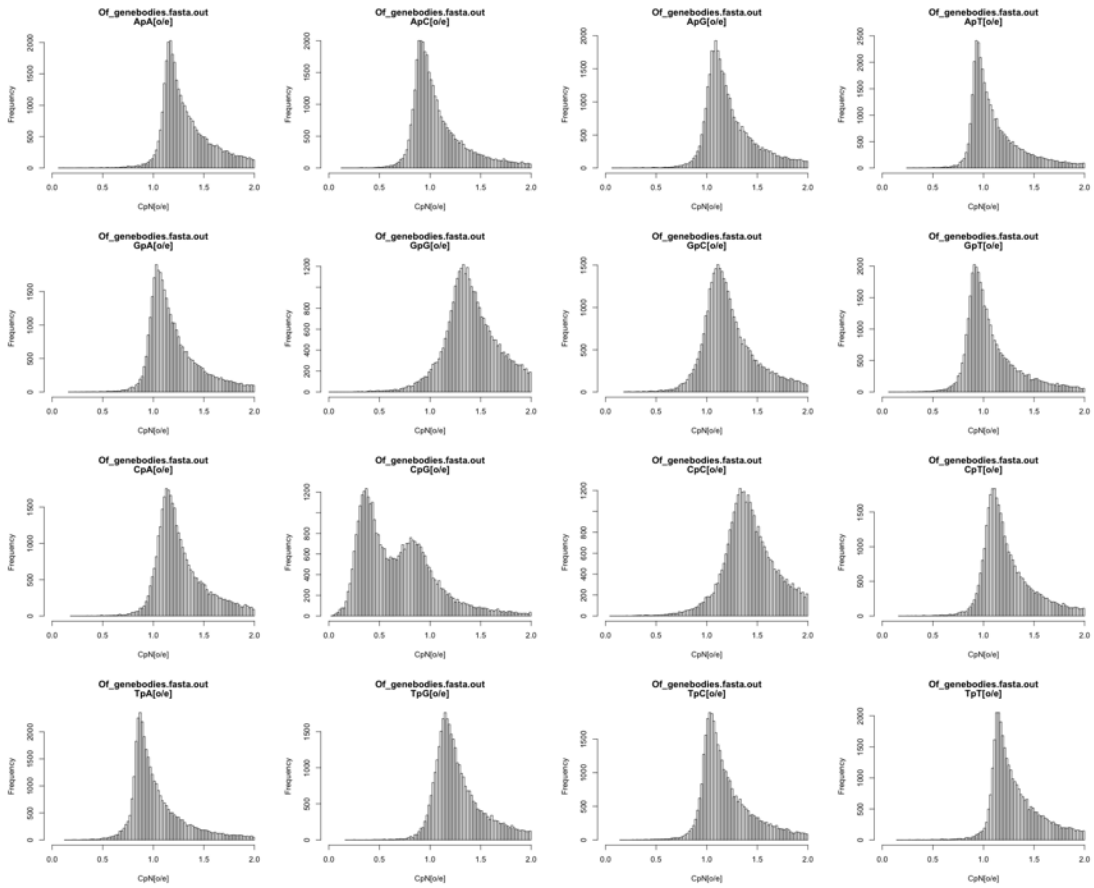


Figure S 5.25: Frequency histogram of the observed vs. expected ratio for all dinucleotide combinations in *Oncopeltus* intragenic regions. The y-axis depicts the number of genes with the specific dinucleotide [o/e] values given on the x-axis.

Histone encoding loci and histone modifying enzymes

The core unit of chromatin is the nucleosome, a highly conserved repeating unit composed of two copies of each of the four core histone proteins (H2A, H2B, H3, H4) assembled into an octamer and wound around 146-147 bp of DNA. The linker histone H1 binds the nucleosome and locks the DNA into place by binding the entry and exit sites of the DNA. These histones are post-translationally modified by methylation, acetylation and phosphorylation, dynamically influencing the structure of the chromatin. The *Oncopeltus* genome encodes fewer histone loci than any other arthropod genome (Tables S 5.24, 5.25). The linker histone, Histone H1, was initially not found in the assembly for the OGS v1.1, but it was subsequently identified from transcriptomic data, and a partial model for first exon was then identified in the assembly and incorporated into the OGS v1.2. This model is even missing a start codon due to a large gap directly upstream, which together with the missing sequence explains why the automated annotation pipeline missed it completely. *Oncopeltus* has two copies of the variant histone H2A.Z/H2.AV, which functions in environmental responsiveness and marking damaged DNA [270].

In *Drosophila* the histone genes are present in the genome in large numbers of quintet clusters, each cluster possessing one gene from each of the five classes of histone proteins. This arrangement of genes is also observed in other insects such as the pea aphid [271] and bed bug [19]. *Oncopeltus* does not have these quintet clusters and all of the histone genes are present as single copies on a scaffold.

The *Oncopeltus* genome encodes genes responsible for all classes of histone modifications; histone acetyltransferases, deacetylases, methylases and demethylases. Unusually there are duplications of the histone acetyltransferases *males absent on the first (mof), chameau (chm)* and *enoki mushroom (enok).* Duplications of *mof* and *enok* have only previously been reported for the pea aphid [271] and the bed bug (*Cimex lectularius)* [19]. Phylogenetic analysis indicates that the duplications of these genes in *Oncopeltus* are independent of the duplications in the aphid genome and in the bed bug genome.

Table S 5.24: Number of loci encoding the five classes of histones within the genomes of arthropod species. Orthologs for *A. aegypti*, *D. pulex*, *T. urticae* and *I. scapularis* were obtained by BLAST analysis. Orthologs for *A. mellifera* and *A. pisum* were obtained from published literature [255, 271].

Species	*H1*	*H2A*	*H2B*	*H3*	*H4*
Aedes aegypti	6	19	11	18	15
Apis mellifera	2	6	5	6	4
Acyrthosiphon pisum	6	5	5	7	5
Oncopeltus fasciatus	**1**	**3**	**4**	**3**	**2**
Cimex lectularius	4	14	6	13	8
Daphnia pulex	5	10	12	10	6
Tetranychus urticae	1	4	7	6	3
Ixodes scapularis	4	6	4	4	1
Strigamia	3	7	15	4	4

Table S 5.25: Histone and histone modifying genes identified in the *Oncopeltus* genome (in Excel Supplement).

5.4.d Repressive C2H2 zinc finger effectors (KAP-1/ TRIM proteins)

Contributor: Kristen A. Panfilio (KAP!)

Although the zinc finger 271-like subfamily in *Oncopeltus* shares a number of genomic, protein, and evolutionary features with the repressive KRAB-domain zinc finger protein families of vertebrates, we do not find evidence for an insect ortholog of the interaction partner KAP-1 (Figure S 5.26, [272], see also main text). In vertebrates, KRAB-associated protein 1 (KAP-1) acts as a chromatin-remodeling co-repressor, by recruiting methyltransferases (H3K9me) and deacetylases (H3K9ac and H3K14ac) to targeted genes' promoters (UniProtKB: Q13263). KAP-1, also known as TRIM28, is a member of the Tripartite motif-containing protein (TRIM) family.

Blast searches with human KAP-1 chiefly identified GenBank protein accessions designated as TRIM33 in insects and, in hemipteran genome assemblies, gene models encoding NHL repeat/ TRIM71 proteins. Thus, while insects do not have a direct ortholog of KAP-1, they do possess a homolog of the vertebrate protein family comprised of both KAP-1 and TRIM33 (Figure S 5.26: blue), where the latter is an E3 ubiquitin-protein ligase involved in BMP signaling repression in humans (UniProtKB: Q9UPN9). Curiously, the heteropteran TRIM28/33 homologs form a diverged outgroup (red) compared to other insect TRIM28/33 homologs, while genome tblastn searches with human KAP-1 rather identifies TRIM71-like homologs, which have predicted RNA-binding translational repression activity, in *Oncopeltus* and *Cimex* (grey).

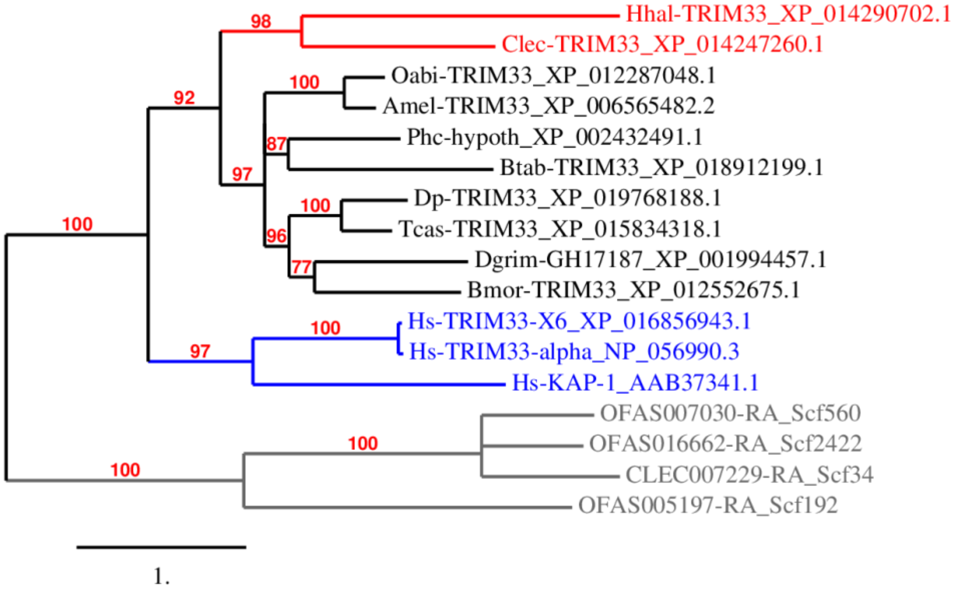

98
Hhal-TRIM33_XP_014290702.1
Clec-TRIM33_XP_014247260.1
92
100
Oabi-TRIM33_XP_012287048.1
Amel-TRIM33_XP_006565482.2
87
Phc-hypoth_XP_002432491.1
Btab-TRIM33_XP_018912199.1
97
100
Dp-TRIM33_XP_019768188.1
Tcas-TRIM33_XP_015834318.1
96
77
Dgrim-GH17187_XP_001994457.1
Bmor-TRIM33_XP_012552675.1
100
100
Hs-TRIM33-X6_XP_016856943.1
Hs-TRIM33-alpha_NP_056990.3
97
Hs-KAP-1_AAB37341.1
OFAS007030-RA_Scf560
OFAS016662-RA_Scf2422
100
CLEC007229-RA_Scf34
100
OFAS005197-RA_Scf192
1.

Figure S 5.26: Maximum likelihood phylogeny of selected tripartite motif-containing (TRIM) proteins. The TRIM28/33 family diversified after divergence of the vertebrate and insect lineages. The vertebrate TRIM28/33 clade (blue) is represented by the chromatin co-repressor KAP-1 (TRIM28) and two isoforms of TRIM33 from human. The diverged heteropteran TRIM28/33 homologs (red) form an outgroup to all other insect TRIM28/33 proteins, including that of the fellow hemipteran *Bemisia tabaci*. Genome tblastn searches with human KAP-1 find strongest homology among NHL repeat proteins of the TRIM71 and Brat subfamilies in *Oncopeltus* and *Cimex* (grey). GenBank accessions or OGS identifiers and scaffold locations (“Scf”) are appended to protein labels. Species abbreviations: Vertebrata: Hs, *Homo sapiens*; Hemiptera: Btab, *Bemisia tabaci*; Clec, *Cimex lectularius*; Hhal, *Halyomorpha halys*; Ofas, *Oncopeltus fascia*tus; Psocodea: Phc, *Pediculus humanus corporis*; Hymenoptera: Amel, *Apis mellifera*; Oabi, *Orussus abietinus*; Coleoptera: Dp, *Dendroctonus ponderosae*; Tcas, *Tribolium castaneum*; Lepidoptera: Bmor, *Bombyx mori*; Diptera: Dgrim, *Drosophila grimshawi*. Branch support is given in percent, branch length scale is in substitutions per site.

6. Post-OGS v1.1 pipeline analyses

6.1 Protein gene orthology assessments via OrthoDB and BUSCO

Contributors: Panagiotis Ioannidis, Robert M. Waterhouse, Evgeny M. Zdobnov

Note that all of these analyses are based on the OGS v1.1, and only one isoform was considered per gene (n= 19,519 proteins). The methods described here are in support of main text Fig. 3a-b.

Orthology analysis

The OrthoDB resource [273] was used in order to find shared orthologs among *O. fasciatus* and the another ten arthropods (*Daphnia pulex*, *Zootermopsis nevadensis*, *Cimex lectularius*, *Rhodnius prolixus*, *Acyrthosiphon pisum*, *Pediculus humanus*, *Apis mellifera*, *Tribolium castaneum*, *Danaus plexippus*, and *Drosophila melanogaster*), using an expanded version of OrthoDB v.8 that also includes i5K species and mapped information for *Rhodnius prolixus.* Custom Perl scripts [274] were used in order to find the number of genes in each category shown in the bar chart in main text Fig. 3a. For the categories "Present in majority of species" and "Patchy distribution", we required that the given ortholog was found in 9-10 or 2-8 species, respectively. Proteins in the "Orthologs in other Arthropoda" category were found using the OrthoDB pipeline, while those in the "Homologs in other Arthropoda" did not meet those criteria but had hits with e-values <1e-05.

Furthermore, with the latest version of OrthoDB (v9.1, [275]), we specifically analyzed the 8,861 orthogroups resulting from orthology clustering analyses for the Hemiptera. Orthogroups are based on nine species that span the breadth of this order: *Acyrthosiphon pisum, Pachypsylla venusta, Homalodisca vitripennis, Diaphorina citri, Gerris buenoi, Rhodnius prolixus, Cimex lectularius, Halyomorpha halys*, and *Oncopeltus fasciatus.* Consistent with our BUSCO analyses presented below, we focused on how *Oncopeltus* compared to three of the other species, chosen for having high quality, published, and stable official gene sets: the pea aphid, *Cimex*, and *Rhodnius*. *Oncopeltus* compares favorably with the other species for both gene presence and for copy number (Table S 6.1). All four species have good orthogroup representation (≥90%), and in fact *Oncopeltus* (94.0%) has better representation than either the pea aphid or *Rhodnius*. Furthermore, where orthologs are expected to be single copy, we only see a moderate increase in duplicates in *Oncopeltus* (1.64× compared to *Rhodnius*, which had the fewest), while the pea aphid – with known duplications – has a notable 5.6× higher level of duplications compared to *Rhodnius*.

Table S 6.1: OrthoDB v9.1 comparison of four species for presence and copy number in Hemiptera-level orthogroups.

Hemipteran orthogroup metrics	**Species**			
	Apis	***Rpro***	***Clec***	***Ofas***
A. Total hemipteran orthogroups that include this species (n= 8,861 orthogroups)	7,974 (90.0%)	8,261 (93.2%)	8,392 (94.7%)	8,330 (94.0%)
B. Total orthogroups from which the species is absent	887	600	469	531
B(i). # orthogroups missing an ortholog that is single-copy in the other 3 species	780	494	344	402
B(ii). # orthogroups missing an ortholog that is present (but not all single-copy) in the other 3 species	107	106	125	129
C. Number of groups with more than one ortholog but that is single-copy in the other 3 species	662	118	143	193
D. Total number of proteins in orthogroups with at least 3 of the 4 species	11,188	7,877	8,049	8,282

Phylogenetic analysis

For the phylogenetic analysis, an additional hemipteran species was included, the brown planthopper *Nilaparvata lugens*. However, since there was no available official gene set for this species, we extracted the sequences of conserved single-copy genes from the genome assembly using BUSCO [28] and mapped them onto OrthoDB. Subsequently, we used the genes that were present as single-copy in all twelve species in order to build a concatenated phylogenetic tree using RAxML (Randomized Axelerated Maximum Likelihood, [276]). Briefly, a multiple sequence alignment was performed using MUSCLE [277] for each orthologous group, separately. Then, the resulting alignments were trimmed using trimAl [206] with parameters "-w 3 -gt 0.95 -st 0.01". The trimmed alignments were concatenated using the "seqret" program from the EMBOSS suite [278]. This concatenated alignment was used to build the phylogeny using RAxML 7.6.6 with the PROTGAMMA model of amino acid substitutions and 100 bootstrap replicates.

BUSCO-based quality assessment

The quality of genome assemblies can be measured by searching for the presence of conserved genes. Moreover, if these conserved genes are also single-copy, the assembly can also be tested for unexpected duplications, which can be a sign of erroneous haplotype assembly. To this end, we used the Benchmarking Universal Single-Copy Orthologs (BUSCO, v3: [28, 29], scripts available under Gitlab project at https://gitlab.com/ezlab/busco), to measure the completeness of the milkweed bug genome as well as its set of predicted protein coding genes. We used both the Insecta and Arthropoda gene sets, which comprise genes that are present in at least 90% of the respective taxon. The values are highly similar between the two taxonomic datasets (Table S 6.2). This evaluation shows that most BUSCO genes are present in the *Oncopeltus* OGS v1.1 (only ~1% missing), although additional exons were identified on genomic scaffolds but not yet incorporated into the gene set (substantially fewer fragmented genes). These values compare favorably with that of high quality genomes that have been published for other hemipteran species, such as the bed bug and pea aphid (see main text).

Table S 6.2: BUSCO v3 statistics for gene count and percentage representation of the Insecta (n= 1658) and Arthropoda (n= 1066) datasets for the OGS v1.1 and Illumina assembly.

Dataset	**Complete (single copy and duplicated)**	**Duplicated (subset of “complete”)**	**Fragmented**	**Missing**
OGS: Insecta	1,303 (78.6%)	41 (2.5%)	337 (20.3%)	18 (1.1%)
OGS: Arthropoda	870 (81.6%)	28 (2.6%)	190 (17.8%)	6 (0.6%)
Assembly: Insecta	1,568 (94.6%)	23 (1.4%)	50 (3.0%)	40 (2.4%)
Assembly: Arthropoda	1,018 (95.5%)	15 (1.4%)	29 (2.7%)	19 (1.8%)

6.2 Transcription factor classifications and orthology assignments

Contributor: Matthew T. Weirauch

See the main text for results and discussion on this section. The supplementary tables for this section are listed below.

Table S 6.3: Transcription factor counts. Values for heatmap representation (main text Fig.4a) of transcription factor abundance per family per species, log (base 2) scale, for 74 transcription factor families in 16 arthropod species (in Excel Supplement).

Table S 6.4: *Oncopeltus* transcription factors with orthology predictions based on their DNA-binding specificities, and their predicted DNA binding motifs (in Excel Supplement).

Table S 6.5: *Oncopeltus* transcription factors without orthology predictions. These *Oncopeltus* proteins could be automatically classified to a transcription factor family, but without a specific orthology assignment. Each protein's DNA-binding domains (DBDs) are listed sequentially, from 5' to 3' within the amino acid sequence (in Excel Supplement).

6.3 Gene structure evolution

Contributor: Kristen A. Panfilio

"Gold standard" manual curation gene set

To analyze gene structure, we created a high quality ("gold standard") dataset of 30 genes whose manual curation could reasonably ensure complete and accurate protein coding gene models across seven species from four insect orders (see the main text for results and discussion on this section: main text Fig. 6a). We deliberately chose multi-exonic genes encoding fairly large proteins (median of 1720 aa), with clear single-copy orthology and strong RNA-seq expression to support model curation. For broad molecular sampling, genes were chosen that encode proteins of diverse functional classes, ranging from structural molecules to signaling pathway components, enzymes, and intracellular regulators of organelle structure and vesicular trafficking (detailed in Table S 6.6).

Splice site analysis

We annotated intron positions within multiple sequence alignments of orthologous proteins and plotted gains and losses onto a phylogeny for the genes *hemocytin* (also known as *Hemolectin*, *Hml*)*,* *Tenascin major* (*Ten-m*), and *UDP-galactose 4′-epimerase* (*GalE*) (see the main text for results and discussion on this section: main text Fig. 7; details on aligned splice positions' conservation is given in supplementary Tables S 6.7-6.9). The many-exon genes were taken from our initial "gold standard" dataset. The gene *hemocytin* (also known as *Hemolectin*, *Hml*) encodes a hemocyte clotting agent with numerous functional domains for cuticle- and protein-protein binding, and it had the most exons of any gene in the dataset (≤74 exons). Unfortunately, it was not possible to annotate a complete model of this gene in the thrips, which as a thysanopteran represents a closely related outgroup to the Hemiptera, and therefore we additionally evaluated *Tenascin major* (*Ten-m*), which encodes a teneurin family multi-domain protein involved in neural development and synaptic transmission. At the other end of the spectrum, we chose an ancient gene with conservation extending to bacteria, such that the ancestral protein had no introns: *UDP-galactose 4′-epimerase* (*GalE*) encodes an enzyme for sugar metabolism (EC

5.1.3.2) and is found across virtually all kingdoms of life. In fact, sequence conservation is so high at the nucleotide level that this gene initially came up as a potential candidate in our pipeline analyses of lateral gene transfer events in *Oncopeltus* (see also Supplemental Note 2.2), although we find here that the ancestral insect *epimerase* gene already had at least three exons.

Methods

Gene models: Single-copy orthologous gene models were obtained from public genome browsers (*D. melanogaster* and *T. castaneum*) or manually curated from automatic predictions and available expression data evidence tracks in the i5K NAL Apollo web instances (all other species: *O. fasciatus, C. lectularius, G. buenoi, F. occidentalis, A. glabripennis, C. capitata*). Manual curation was performed iteratively and by orthology comparisons in sequence alignments, to ensure protein sequence completeness even when genes were split across draft assembly scaffolds. In the case of *hemocytin* and *furry*, the models in *O. fasciatus* and *G. buenoi* were filled with place-holder X's where strong orthology conservation (including empirical data from the other two hemipterans) made clear that small internal exons were truncated or absent in the current assembly, involving the inference of two splice positions in *G. buenoi* and three in *O. fasciatus*.

Alignment: Protein translations were obtained for each exon, and the splice positions were marked (rounded to the nearest triplet/ whole amino acid position). A protein sequence alignment into which splice positions were encoded (denoted by the character "X") was then generated with ClustalW (accessed at http://www.genome.jp/tools/clustalw/) and manually refined.

Inference of evolutionary patterns: All splice positions were considered individually, and only those for which the sequence alignment was particularly poor were excluded (six sites within the less conserved N-terminal region and two other sites within *hemocytin*). Evolutionary patterns of splice position gain and loss were encoded as a binary presence/absence value per position for each species (Tables S 6.7-6.9), and the most parsimonious inference (fewest lineage-specific changes) to generate this pattern was assumed, with no weighting of probability for gains relative to losses.

Table S 6.6 List of “gold standard” manually curated genes evaluated for gene structure properties, including public database IDs for *T. castaneum* and *D. melanogaster* orthologs (in Excel Supplement).

Table S 6.7: Hemocytin (Hemolectin, Hml) splice site conservation analysis (7 species) (in Excel Supplement).

Table S 6.8: Tenascin major (Ten-m) splice site conservation analysis (8 species) (in Excel Supplement).

Table S 6.9: UDP-galactose 4′ -epimerase (GalE) splice site conservation analysis (8 species) (in Excel Supplement).

6.4 Interspecific comparisons of metabolic enzymes

Contributors: Patrice Baa-Puyoule, Gérard Febvay, Nicolas Parisot, Stefano Colella

See the main text for results and discussion on this section. The supplementary tables for this section are listed below.

Table S 6.10: Global EC statistics. Global comparison of Enzyme Commission (EC) categories and amino acid metabolism enzymes (KEGG) present in 28 samples for 27 species in the ArthropodaCyc repository, including the newly generated OncfaCyc database for *Oncopeltus* (in Excel Supplement).

Table S 6.11: Detailed listing of EC categories that are uniquely present or absent from *Oncopeltus* (in Excel Supplement).

Table S 6.12: Detailed comparison tables and Venn diagram of amino acid metabolism enzyme repertoires across four hemipterans (in Excel Supplement).

Table S 6.13: Urea cycle comparison across 27 insect species. Presence/absence of EC numbers essential to the urea cycle are indicated for each species (in Excel Supplement).

7. References

1. 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.
2. Hare EE, Johnston JS: **Genome size determination using flow cytometry of propidium iodide-stained nuclei.** *Methods Mol Biol* 2011, **772:**3-12.
3. Messthaler H, Traut W: **Phases of Sex Chromosome Inactivation in *Oncopeltus fasciatus* and *Pyrrhocoris apterus* (Insecta, Heteroptera).** *Caryologia* 1975, **28:**501-510.
4. Marcais G, Kingsford C: **A fast, lock-free approach for efficient parallel counting of occurrences of k-mers.** *Bioinformatics* 2011, **27:**764-770.
5. Bushnell B: **BBMap short read aligner.** 2016.
6. Chen W, Hasegawa DK, Arumuganathan K, Simmons AM, Wintermantel WM, Fei Z, Ling KS: **Estimation of the Whitefly *Bemisia tabaci* Genome Size Based on k-mer and Flow Cytometric Analyses.** *Insects* 2015, **6:**704-715.
7. Guo LT, Wang SL, Wu QJ, Zhou XG, Xie W, Zhang YJ: **Flow cytometry and K-mer analysis estimates of the genome sizes of *Bemisia tabaci* B and Q (Hemiptera: Aleyrodidae).** *Front Physiol* 2015, **6:**144.
8. Derks MF, Smit S, Salis L, Schijlen E, Bossers A, Mateman C, Pijl AS, de Ridder D, Groenen MA, Visser ME, Megens HJ: **The Genome of Winter Moth (*Operophtera brumata*) Provides a Genomic Perspective on Sexual Dimorphism and Phenology.** *Genome Biol Evol* 2015, **7:**2321-2332.
9. **K-mer Spectrum Primer** [http://www.broadinstitute.org/software/allpaths-lg/blog/?page_id=693]
10. Liu B, Shi Y, Yuan J, Hu X, Zhang H, Li N, Li Z, Chen Y, Mu D, Fan W: **Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects.** http://arxivorg/ 2013, **1308**.
11. Shan G, Zheng W-M: **An ℓ-mer component distribution for genome size esimation.** 2009.
12. Chor B, Horn D, Goldman N, Levy Y, Massingham T: **Genomic DNA k-mer spectra: models and modalities.** *Genome Biol* 2009, **10:**R108.
13. Schmidt NO: **On the *k*-mer frequency spectra of organism genome and proteome sequences with a preliminary machine learning assessment of prime predictability.** BOISE STATE UNIVERSITY GRADUATE COLLEGE, 2012.
14. Li F, Fan G, Wang K, Sun F, Yuan Y, Song G, Li Q, Ma Z, Lu C, Zou C, et al: **Genome sequence of the cultivated cotton *Gossypium arboreum*.** *Nat Genet* 2014, **46:**567-572.

15. Abe H, Yoshikawa N, Sarower MG, Okada S: **Physiological function and metabolism of free D-alanine in aquatic animals.** *Biol Pharm Bull* 2005, **28:**1571-1577.

16. Schroder R, Wegrzyn TF, Sharma NN, Atkinson RG: **LeMAN4 endo-beta-mannanase from ripe tomato fruit can act as a mannan transglycosylase or hydrolase.** *Planta* 2006, **224:**1091-1102.

17. Nikoh N, Nakabachi A: **Aphids acquired symbiotic genes via lateral gene transfer.** *BMC Biol* 2009, **7:**12.

18. Wheeler D, Redding AJ, Werren JH: **Characterization of an ancient lepidopteran lateral gene transfer.** *PLoS One* 2012, **8:**e59262.

19. Benoit JB, Adelman ZN, Reinhardt K, Dolan A, Poelchau M, Jennings EC, Szuter EM, Hagan RW, Gujar H, Shukla JN, et al: **Unique features of a global human ectoparasite identified through sequencing of the bed bug genome.** *Nat Commun* 2016, **7:**10165.

20. Duncan EJ, Wilson MJ, Smith JM, Dearden PK: **Evolutionary origin and genomic organisation of runt-domain containing genes in arthropods.** *BMC Genomics* 2008, **9:**558.

21. Bao Z, Eddy SR: **Automated de novo identification of repeat sequence families in sequenced genomes.** *Genome Res* 2002, **12:**1269-1276.

22. Price AL, Jones NC, Pevzner PA: **De novo identification of repeat families in large genomes.** *Bioinformatics* 2005, **21 Suppl 1:**i351-358.

23. Benson G: **Tandem repeats finder: a program to analyze DNA sequences.** *Nucleic Acids Res* 1999, **27:**573-580.

24. Ruiz-Ruano FJ, López-León MD, Cabrero J, Camacho JPM: **High-throughput analysis of the satellitome illuminates satellite DNA evolution.** *Scientific Reports* 2016, **6:**28333.

25. Ross MG, Russ C, Costello M, Hollinger A, Lennon NJ, Hegarty R, Nusbaum C, Jaffe DB: **Characterizing and measuring bias in sequence data.** *Genome Biol* 2013, **14:**R51.

26. English AC, Richards S, Han Y, Wang M, Vee V, Qu J, Qin X, Muzny DM, Reid JG, Worley KC, Gibbs RA: **Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology.** *PLoS One* 2012, **7:**e47768.

27. Wang X, Fang X, Yang P, Jiang X, Jiang F, Zhao D, Li B, Cui F, Wei J, Ma C, et al: **The locust genome provides insight into swarm formation and long-distance flight.** *Nat Commun* 2014, **5:**2957.

28. 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.

29. Waterhouse RM, Seppey M, Simao FA, Manni M, Ioannidis P, Klioutchnikov G, Kriventseva EV, Zdobnov EM: **BUSCO applications from quality assessments to gene prediction and phylogenomics.** *Mol Biol Evol* 2017.

30. 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, et al: **Genomes of the rice pest brown planthopper and its endosymbionts reveal complex complementary contributions for host adaptation.** *Genome Biol* 2014, **15:**521.

31. Platt RN, 2nd, Blanco-Berdugo L, Ray DA: **Accurate transposable element annotation is vital when analyzing new genome assemblies.** *Genome Biol Evol* 2016.

32. The International Aphid Genomics Consortium: **Genome sequence of the pea aphid *Acyrthosiphon pisum*** *PLoS Biol* 2010, **8:**e1000313.

33. Zhen Y, Aardema ML, Medina EM, Schumer M, Andolfatto P: **Parallel molecular evolution in an herbivore community.** *Science* 2012, **337:**1634-1637.

34. 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.

35. 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.

36. Holt C, Yandell M: **MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects.** *BMC Bioinformatics* 2011, **12:**491.

37. Stanke M, Diekhans M, Baertsch R, Haussler D: **Using native and syntenically mapped cDNA alignments to improve de novo gene finding.** *Bioinformatics* 2008, **24:**637-644.

38. Korf I: **Gene finding in novel genomes.** *BMC Bioinformatics* 2004, **5:**59.

39. 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.

40. 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.

41. 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.

42. Dereeper A, Audic S, Claverie JM, Blanc G: **BLAST-EXPLORER helps you building datasets for phylogenetic analysis.** *BMC Evol Biol* 2010, **10:**8.

43. Dereeper A, Guignon V, Blanc G, Audic S, Buffet S, Chevenet F, Dufayard JF, Guindon S, Lefort V, Lescot M, et al: **Phylogeny.fr: robust phylogenetic analysis for the non-specialist.** *Nucleic Acids Res* 2008, **36:**W465-469.

44. Denton JF, Lugo-Martinez J, Tucker AE, Schrider DR, Warren WC, Hahn MW: **Extensive error in the number of genes inferred from draft genome assemblies.** *PLoS Comput Biol* 2014, **10:**e1003998.

45. Megy K, Emrich SJ, Lawson D, Campbell D, Dialynas E, Hughes DS, Koscielny G, Louis C, Maccallum RM, Redmond SN, et al: **VectorBase: improvements to a bioinformatics resource for invertebrate vector genomics.** *Nucleic Acids Res* 2012, **40:**D729-734.

46. McKenna DD, Scully ED, Pauchet Y, Hoover K, Kirsch R, Geib SM, Mitchell RF, Waterhouse RM, Ahn SJ, Arsala D, et al: **Genome of the Asian longhorned beetle (*Anoplophora glabripennis*), a globally significant invasive species, reveals key functional and evolutionary innovations at the beetle-plant interface.** *Genome Biol* 2016, **17:**227.

47. Lynch JA: **Diversity of molecules and mechanisms in establishing insect anteriorâ€“posterior polarity.** *Current Opinion in Insect Science* 2014, **1:**39-44.

48. Ewen-Campen B, Shaner N, Panfilio KA, Suzuki Y, Roth S, Extavour CG: **The maternal and early embryonic transcriptome of the milkweed bug *Oncopeltus fasciatus*.** *BMC Genomics* 2011, **12:**61.

49. Weisbrod A, Cohen M, Chipman AD: **Evolution of the insect terminal patterning system--insights from the milkweed bug, *Oncopeltus fasciatus*.** *Dev Biol* 2013, **380:**125-131.

50. Duncan EJ, Benton MA, Dearden PK: **Canonical terminal patterning is an evolutionary novelty.** *Dev Biol* 2013, **377:**245-261.

51. Ben-David J, Chipman AD: **Mutual regulatory interactions of the trunk gap genes during blastoderm patterning in the hemipteran *Oncopeltus fasciatus*.** *Dev Biol* 2010, **346:**140-149.

52. Ben-David J, Chipman AD: **Mutual regulatory interactions of the trunk gap genes during blastoderm patterning in the hemipteran *Oncopeltus fasciatus*.** *Dev Biol* 2010, **346:**140-149.

53. Erezyilmaz DF, Kelstrup HC, Riddiford LM: **The nuclear receptor E75A has a novel pair-rule-like function in patterning the milkweed bug, *Oncopeltus fasciatus*.** *Dev Biol* 2009, **334:**300-310.

54. Liu PZ, Kaufman TC: ***hunchback* is required for suppression of abdominal identity, and for proper germband growth and segmentation in the intermediate germband insect *Oncopeltus fasciatus*.** *Development* 2004, **131:**1515-1527.

55. Peel AD, Telford MJ, Akam M: **The evolution of hexapod engrailed-family genes: evidence for conservation and concerted evolution.** *Proc Biol Sci* 2006, **273:**1733-1742.

56. Krumlauf R: **Evolution of the vertebrate Hox homeobox genes.** *Bioessays* 1992, **14:**245-252.

57. Richards S, Gibbs RA, Weinstock GM, Brown SJ, Denell R, Beeman RW, Gibbs R, Bucher G, Friedrich M, Grimmelikhuijzen CJ, et al: **The genome of the model beetle and pest Tribolium castaneum.** *Nature* 2008, **452:**949-955.

58. **The homebox page** [http://evolution.genetics.washington.edu/phylip.html]

59. Gómez-Skarmeta J-L, Corral RDd, Calle-Mustienes Edl, Ferrés-Marcó D, Modolell J: ***araucan* and *caupolican*, two members of the novel Iroquois Complex, encode homeoproteins that control proneural and vein-forming genes.** *Cell* 1996, **85:**95-105.

60. Gomez-Skarmeta JL, Modolell J: **Iroquois genes: genomic organization and function in vertebrate neural development.** *Curr Opin Genet Dev* 2002, **12:**403-408.

61. Irimia M, Maeso I, Garcia-Fernandez J: **Convergent evolution of clustering of Iroquois homeobox genes across metazoans.** *Mol Biol Evol* 2008, **25:**1521-1525.

62. Cavodeassi F, Modolell J, Gómez-Skarmeta JL: **The Iroquois family of genes: from body building to neural patterning.** *Development* 2001, **128:**2847-2855.

63. Kerner P, Ikmi A, Coen D, Vervoort M: **Evolutionary history of the iroquois/Irx genes in metazoans.** *BMC Evol Biol* 2009, **9:**74.

64. Maeso I, Irimia M, Tena JJ, González-Pérez E, Tran D, Ravi V, Venkatesh B, Campuzano S, Gómez-Skarmeta JL, Garcia-Fernàndez J: **An ancient genomic regulatory block conserved across bilaterians and its dismantling in tetrapods by retrogene replacement.** *Genome Res* 2012, **22:**642-655.

65. Savard J, Tautz D, Richards S, Weinstock GM, Gibbs RA, Werren JH, Tettelin H, Lercher MJ: **Phylogenomic analysis reveals bees and wasps (Hymenoptera) at the base of the radiation of holometabolous insects.** *Genome Res* 2006, **16:**1334-1338.

66. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S: **MEGA6: Molecular Evolutionary Genetics Analysis version 6.0.** *Mol Biol Evol* 2013, **30:**2725-2729.

67. Reim I, Lee HH, Frasch M: **The T-box-encoding Dorsocross genes function in amnioserosa development and the patterning of the dorsolateral germ band downstream of Dpp.** *Development* 2003, **130:**3187-3204.

68. Reim I, Frasch M: **The Dorsocross T-box genes are key components of the regulatory network controlling early cardiogenesis in Drosophila.** *Development* 2005, **132:**4911-4925.

69. Horn T, Panfilio KA: **Novel functions for Dorsocross in epithelial morphogenesis in the beetle Tribolium castaneum.** *Development* 2016, **143:**3002-3011.

70. Mangelsdorf DJ, Thummel C, Beato M, Herrlich P, Schutz G, Umesono K, Blumberg B, Kastner P, Mark M, Chambon P, Evans RM: **The nuclear receptor superfamily: the second decade.** *Cell* 1995, **83:**835-839.

71. Chambon P: **The nuclear receptor superfamily: a personal retrospect on the first two decades.** *Mol Endocrinol* 2005, **19:**1418-1428.

72. Margolis RN, Evans RM, O'Malley BW: **The Nuclear Receptor Signaling Atlas: development of a functional atlas of nuclear receptors.** *Mol Endocrinol* 2005, **19:**2433-2436.

73. Taneja R: *Nuclear Receptors in Development.* Oxford: Elsevier; 2006.

74. Fahrbach SE, Smagghe G, Velarde RA: **Insect nuclear receptors.** *Annual review of entomology* 2012, **57:**83-106.

75. McKenna NJ, Cooney AJ, DeMayo FJ, Downes M, Glass CK, Lanz RB, Lazar MA, Mangelsdorf DJ, Moore DD, Qin J, et al: **Minireview: Evolution of NURSA, the Nuclear Receptor Signaling Atlas.** *Mol Endocrinol* 2009, **23:**740-746.

76. Xiao X, Wang P, Chou KC: **Recent progresses in identifying nuclear receptors and their families.** *Current topics in medicinal chemistry* 2013, **13:**1192-1200.

77. Laudet V: **Evolution of the nuclear receptor superfamily: early diversification from an ancestral orphan receptor.** *J Mol Endocrinol* 1997, **19:**207-226.

78. Markov GV, Laudet V: **Origin and evolution of the ligand-binding ability of nuclear receptors.** *Molecular and cellular endocrinology* 2011, **334:**21-30.

79. Sladek FM: **What are nuclear receptor ligands?** *Molecular and cellular endocrinology* 2011, **334:**3-13.

80. The Nuclear Receptor Nomenclature Committee: **A unified nomenclature for the nuclear receptor superfamily.** *Cell* 1999, **97:**161-163.

81. Glass CK, Rosenfeld MG: **The coregulator exchange in transcriptional functions of nuclear receptors.** *Genes Dev* 2000, **14:**121-141.

82. Tremblay JJ, Marcil A, Gauthier Y, Drouin J: **Ptx1 regulates SF-1 activity by an interaction that mimics the role of the ligand-binding domain.** *EMBO J* 1999, **18:**3431-3441.

83. Yoo J, Ko S, Kim H, Sampson H, Yun JH, Choe KM, Chang I, Arrowsmith CH, Krause HM, Cho HS, Lee W: **Crystal structure of Fushi tarazu factor 1 ligand binding domain/Fushi

tarazu peptide complex identifies new class of nuclear receptors. *The Journal of biological chemistry* 2011, **286:**31225-31231.

84. King-Jones K, Thummel CS: **Nuclear receptors--a perspective from *Drosophila*.** *Nat Rev Genet* 2005, **6:**311-323.

85. Ruaud AF, Lam G, Thummel CS: **The Drosophila NR4A nuclear receptor DHR38 regulates carbohydrate metabolism and glycogen storage.** *Molecular endocrinology* 2011, **25:**83-91.

86. Tennessen JM, Baker KD, Lam G, Evans J, Thummel CS: **The Drosophila estrogen-related receptor directs a metabolic switch that supports developmental growth.** *Cell Metab* 2011, **13:**139-148.

87. Shigenobu S, Bickel RD, Brisson JA, Butts T, Chang CC, Christiaens O, Davis GK, Duncan EJ, Ferrier DE, Iga M, et al: **Comprehensive survey of developmental genes in the pea aphid, *Acyrthosiphon pisum*: frequent lineage-specific duplications and losses of developmental genes.** *Insect Mol Biol* 2010, **19 Suppl 2:**47-62.

88. St Pierre SE, Ponting L, Stefancsik R, McQuilton P, FlyBase C: **FlyBase 102--advanced approaches to interrogating FlyBase.** *Nucleic Acids Res* 2014, **42:**D780-788.

89. Parker L, Stathakis DG, Arora K: **Regulation of BMP and activin signaling in Drosophila.** *Prog Mol Subcell Biol* 2004, **34:**73-101.

90. Sachs L, Chen YT, Drechsler A, Lynch JA, Panfilio KA, Lassig M, Berg J, Roth S: **Dynamic BMP signaling polarized by Toll patterns the dorsoventral axis in a hemimetabolous insect.** *eLife* 2015, **4:**e05502.

91. Weiss A, Charbonnier E, Ellertsdottir E, Tsirigos A, Wolf C, Schuh R, Pyrowolakis G, Affolter M: **A conserved activation element in BMP signaling during Drosophila development.** *Nat Struct Mol Biol* 2010, **17:**69-76.

92. Balemans W, Van Hul W: **Extracellular regulation of BMP signaling in vertebrates: a cocktail of modulators.** *Dev Biol* 2002, **250:**231-250.

93. Ozuak O, Buchta T, Roth S, Lynch JA: **Ancient and diverged TGF-beta signaling components in Nasonia vitripennis.** *Dev Genes Evol* 2014, **224:**223-233.

94. van der Zee M, da Fonseca RN, Roth S: **TGF-beta signaling in *Tribolium*: vertebrate-like components in a beetle.** *Dev Genes Evol* 2008, **218:**203-213.

95. Podos SD, Hanson KK, Wang YC, Ferguson EL: **The DSmurf ubiquitin-protein ligase restricts BMP signaling spatially and temporally during Drosophila embryogenesis.** *Dev Cell* 2001, **1:**567-578.

96. Xia L, Jia S, Huang S, Wang H, Zhu Y, Mu Y, Kan L, Zheng W, Wu D, Li X, et al: **The Fused/Smurf complex controls the fate of Drosophila germline stem cells by generating a gradient BMP response.** *Cell* 2010, **143:**978-990.

97. Norman M, Vuilleumier R, Springhorn A, Gawlik J, Pyrowolakis G: **Pentagone internalises glypicans to fine-tune multiple signalling pathways.** *Elife* 2016, **5**.

98. Lynch JA, Roth S: **The evolution of dorsal-ventral patterning mechanisms in insects.** *Genes Dev* 2011, **25:**107-118.

99. Müller U, Vogel P, Alber G, Schaub GA: **The innate immune system of mammals and insects.** *Contrib Microbiol* 2008, **15:**21-44.

100. Stein DS, Stevens LM: **Maternal control of the Drosophila dorsal-ventral body axis.** *Wiley Interdiscip Rev Dev Biol* 2014, **3:**301-330.

101. Benton MA, Pechmann M, Frey N, Stappert D, Conrads KH, Chen YT, Stamataki E, Pavlopoulos A, Roth S: **Toll Genes Have an Ancestral Role in Axis Elongation.** *Curr Biol* 2016, **26:**1609-1615.

102. Leulier F, Lemaitre B: **Toll-like receptors--taking an evolutionary approach.** *Nat Rev Genet* 2008, **9:**165-178.

103. Towb P, Sun H, Wasserman SA: **Tube Is an IRAK-4 homolog in a Toll pathway adapted for development and immunity.** *J Innate Immun* 2009, **1:**309-321.

104. Kumar S, Stecher G, Tamura K: **MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets.** *Mol Biol Evol* 2016, **33:**1870-1874.

105. Reeves GT, Stathopoulos A: **Graded dorsal and differential gene regulation in the Drosophila embryo.** *Cold Spring Harb Perspect Biol* 2009, **1:**a000836.

106. Cremazy F, Berta P, Girard F: **Sox neuro, a new *Drosophila* Sox gene expressed in the developing central nervous system.** *Mech Dev* 2000, **93:**215-219.

107. Skeath JB: **At the nexus between pattern formation and cell-type specification: the generation of individual neuroblast fates in the Drosophila embryonic central nervous system.** *Bioessays* 1999, **21:**922-931.

108. Schneider M, Dorn A: **Differential infectivity of two Pseudomonas species and the immune response in the milkweed bug, Oncopeltus fasciatus (Insecta: Hemiptera).** *J Invertebr Pathol* 2001, **78:**135-140.

109. Bao YY, Qu LY, Zhao D, Chen LB, Jin HY, Xu LM, Cheng JA, Zhang CX: **The genome- and transcriptome-wide analysis of innate immunity in the brown planthopper, Nilaparvata lugens.** *BMC Genomics* 2013, **14:**160.

110. Talavera G, Vila R: **What is the phylogenetic signal limit from mitogenomes? The reconciliation between mitochondrial and nuclear data in the Insecta class phylogeny.** *BMC Evol Biol* 2011, **11:**315.

111. Leone P, Bischoff V, Kellenberger C, Hetru C, Royet J, Roussel A: **Crystal structure of *Drosophila* PGRP-SD suggests binding to DAP-type but not lysine-type peptidoglycan.** *Mol Immunol* 2008, **45:**2521-2530.

112. Leulier F, Parquet C, Pili-Floury S, Ryu JH, Caroff M, Lee WJ, Mengin-Lecreulx D, Lemaitre B: **The *Drosophila* immune system detects bacteria through specific peptidoglycan recognition.** *Nat Immunol* 2003, **4:**478-484.

113. Yokoi K, Koyama H, Ito W, Minakuchi C, Tanaka T, Miura K: **Involvement of NF-kappaB transcription factors in antimicrobial peptide gene induction in the red flour beetle, Tribolium castaneum.** *Dev Comp Immunol* 2012, **38:**342-351.

114. Yokoi K, Koyama H, Minakuchi C, Tanaka T, Miura K: **Antimicrobial peptide gene induction, involvement of Toll and IMD pathways and defense against bacteria in the red flour beetle, Tribolium castaneum.** *Results Immunol* 2012, **2:**72-82.

115. Duncan EJ, Dearden PK: **Evolution of a genomic regulatory domain: the role of gene co-option and gene duplication in the Enhancer of split complex.** *Genome Res* 2010, **20:**917-928.

116. Schaeper ND, Prpic NM, Wimmer EA: **A clustered set of three Sp-family genes is ancestral in the Metazoa: evidence from sequence analysis, protein domain structure, developmental expression patterns and chromosomal location.** *BMC Evol Biol* 2010, **10:**88.

117. Birkan M, Schaeper ND, Chipman AD: **Early patterning and blastodermal fate map of the head in the milkweed bug Oncopeltus fasciatus.** *Evol Dev* 2011, **13:**436-447.

118. Thompson JD, Higgins DG, Gibson TJ: **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.** *Nucleic Acids Res* 1994, **22:**4673-4680.

119. Murat S, Hopfen C, McGregor AP: **The function and evolution of Wnt genes in arthropods.** *Arthropod Struct Dev* 2010, **39:**446-452.

120. Oberhofer G, Grossmann D, Siemanowski JL, Beissbarth T, Bucher G: **Wnt/beta-catenin signaling integrates patterning and metabolism of the insect growth zone.** *Development* 2014, **141:**4740-4750.

121. Bao R, Fischer T, Bolognesi R, Brown SJ, Friedrich M: **Parallel duplication and partial subfunctionalization of beta-catenin/armadillo during insect evolution.** *Mol Biol Evol* 2012, **29:**647-662.

122. Janssen R, Le Gouar M, Pechmann M, Poulin F, Bolognesi R, Schwager EE, Hopfen C, Colbourne JK, Budd GE, Brown SJ, et al: **Conservation, loss, and redeployment of Wnt ligands in protostomes: implications for understanding the evolution of segment formation.** *BMC Evol Biol* 2010, **10:**374.

123. Bolognesi R, Farzana L, Fischer TD, Brown SJ: **Multiple Wnt genes are required for segmentation in the short-germ embryo of Tribolium castaneum.** *Curr Biol* 2008, **18:**1624-1629.

124. Sullivan JC, Ryan JF, Mullikin JC, Finnerty JR: **Conserved and novel Wnt clusters in the basal eumetazoan Nematostella vectensis.** *Dev Genes Evol* 2007, **217:**235-239.

125. Doumpas N, Jekely G, Teleman AA: **Wnt6 is required for maxillary palp formation in Drosophila.** *BMC Biol* 2015, **11:**104.

126. Beermann A, Pruhs R, Lutz R, Schroder R: **A context-dependent combination of Wnt receptors controls axis elongation and leg development in a short germ insect.** *Development* 2011, **138:**2793-2805.

127. Kojima T: **The mechanism of *Drosophila* leg development along the proximodistal axis.** *Dev Growth Differ* 2004, **46:**115-129.

128. Siegel JG, Fristrom JW: **The biochemistry of imaginal disc development. In (eds.). The Genetics and Biology of Drosophila.** In *Volume* 2A. Edited by Ashburner M, Wright TRF. New York: Academic Press; 1978: 317–394

129. Cohen B, McGuffin ME, Pfeifle C, Segal D, Cohen SM: **apterous, a gene required for imaginal disc development in Drosophila encodes a member of the LIM family of developmental regulatory proteins.** *Genes Dev* 1992, **6:**715-729.

130. Simcox AA, Roberts IJ, Hersperger E, Gribbin MC, Shearn A, Whittle JR: **Imaginal discs can be recovered from cultured embryos mutant for the segment-polarity genes engrailed, naked and patched but not from wingless.** *Development* 1989, **107:**715-722.

131. Ober KA, Jockusch EL: **The roles of wingless and decapentaplegic in axis and appendage development in the red flour beetle, Tribolium castaneum.** *Dev Biol* 2006, **294:**391-405.

132. Angelini DR, Kaufman TC: **Functional analyses in the milkweed bug *Oncopeltus fasciatus* (Hemiptera) support a role for Wnt signaling in body segmentation but not appendage development.** *Dev Biol* 2005, **283:**409-423.

133. Jiang J, Struhl G: **Complementary and mutually exclusive activities of decapentaplegic and wingless organize axial patterning during Drosophila leg development.** *Cell* 1996, **86:**401-409.

134. Abu-Shaar M, Mann RS: **Generation of multiple antagonistic domains along the proximodistal axis during Drosophila leg development.** *Development* 1998, **125:**3821-3830.

135. Cohen SM, Bronner G, Kuttner F, Jurgens G, Jackle H: **Distal-less encodes a homoeodomain protein required for limb development in Drosophila.** *Nature* 1989, **338:**432-434.

136. Wu J, Cohen SM: **Proximodistal axis formation in the *Drosophila* leg: subdivision into proximal and distal domains by *Homothorax* and *Distal-less*.** *Development* 1999, **126:**109-117.

137. Rieckhof GE, Casares F, Ryoo HD, Abu-Shaar M, Mann RS: **Nuclear translocation of extradenticle requires homothorax, which encodes an extradenticle-related homeodomain protein.** *Cell* 1997, **91:**171-183.

138. Angelini DR, Kaufman TC: **Functional analyses in the hemipteran *Oncopeltus fasciatus* reveal conserved and derived aspects of appendage patterning in insects.** *Dev Biol* 2004, **271:**306-321.

139. Angelini DR, Kikuchi M, Jockusch EL: **Genetic patterning in the adult capitate antenna of the beetle *Tribolium castaneum*.** *Dev Biol* 2009, **327:**240-251.

140. Angelini DR, Smith FW, Aspiras AC, Kikuchi M, Jockusch EL: **Patterning of the adult mandibulate mouthparts in the red flour beetle, *Tribolium castaneum*.** *Genetics* 2012, **190:**639-654.

141. Halder G, Polaczyk P, Kraus ME, Hudson A, Kim J, Laughon A, Carroll S: **The Vestigial and Scalloped proteins act together to directly regulate wing-specific gene expression in Drosophila.** *Genes Dev* 1998, **12:**3900-3909.

142. Williams JA, Bell JB, Carroll SB: **Control of Drosophila wing and haltere development by the nuclear vestigial gene product.** *Genes Dev* 1991, **5:**2481-2495.

143. James AA, Bryant PJ: **Mutations causing pattern deficiencies and duplications in the imaginal wing disk of Drosophila melanogaster.** *Dev Biol* 1981, **85:**39-54.

144. Liu Z, Steward R, Luo L: **Drosophila Lis1 is required for neuroblast proliferation, dendritic elaboration and axonal transport.** *Nat Cell Biol* 2000, **2:**776-783.

145. O'Keefe DD, Thomas JB: **Drosophila wing development in the absence of dorsal identity.** *Development* 2001, **128:**703-710.

146. Medved V, Marden JH, Fescemyer HW, Der JP, Liu J, Mahfooz N, Popadic A: **Origin and diversification of wings: Insights from a neopteran insect.** *Proc Natl Acad Sci U S A* 2015, **112:**15946-15951.

147. Ewen-Campen B, Jones TEM, Extavour CG: **Evidence against a germ plasm in the milkweed bug Oncopeltus fasciatus, a hemimetabolous insect.** *Biol Open* 2013, **2:**556-568.

148. Bickel RD, Cleveland HC, Barkas J, Jeschke CC, Raz AA, Stern DL, Davis GK: **The pea aphid uses a version of the terminal system during oviparous, but not viviparous, development.** *Evodevo* 2013, **4:**10.

149. Lu HL, Tanguy S, Rispe C, Gauthier JP, Walsh T, Gordon K, Edwards O, Tagu D, Chang CC, Jaubert-Possamai S: **Expansion of genes encoding piRNA-associated argonaute proteins in the pea aphid: diversification of expression profiles in different plastic morphs.** *PLoS One* 2011, **6:**e28051.

150. Ewen-Campen B, Jones TE, Extavour CG: **Evidence against a germ plasm in the milkweed bug *Oncopeltus fasciatus*, a hemimetabolous insect.** *Biol Open* 2013, **2:**556-568.

151. Bao R, Friedrich M: **Molecular evolution of the *Drosophila* retinome: exceptional gene gain in the higher Diptera.** *Mol Biol Evol* 2009, **26:**1273-1287.

152. Dilda CL, Mackay TF: **The genetic architecture of *Drosophila* sensory bristle number.** *Genetics* 2002, **162:**1655-1674.

153. Norga KK, Gurganus MC, Dilda CL, Yamamoto A, Lyman RF, Patel PH, Rubin GM, Hoskins RA, Mackay TF, Bellen HJ: **Quantitative analysis of bristle number in *Drosophila* mutants identifies genes involved in neural development.** *Curr Biol* 2003, **13:**1388-1396.

154. Hartenstein V, Wodarz A: **Initial neurogenesis in *Drosophila*.** *Wiley Interdiscip Rev Dev Biol* 2013, **2:**701-721.

155. Tilney LG, Connelly P, Smith S, Guild GM: **F-actin bundles in *Drosophila* bristles are assembled from modules composed of short filaments.** *J Cell Biol* 1996, **135:**1291-1308.

156. Simpson P: **The stars and stripes of animal bodies: evolution of regulatory elements mediating pigment and bristle patterns in Drosophila.** *Trends Genet* 2007, **23:**350-358.

157. Marcellini S, Simpson P: **Two or four bristles: functional evolution of an enhancer of scute in Drosophilidae.** *PLoS Biol* 2006, **4:**e386.

158. Negre B, Simpson P: **Evolution of the achaete-scute complex in insects: convergent duplication of proneural genes.** *Trends Genet* 2009, **25:**147-152.

159. Fyrberg EA, Bond BJ, Hershey ND, Mixter KS, Davidson N: **The actin genes of *Drosophila*: protein coding regions are highly conserved but intron positions are not.** *Cell* 1981, **24:**107-116.

160. Lawrence PA: **Mitosis and the cell cycle in the metamorphic moult of the milkweed bug Oncopeltus fasciatus; a radioautographic study.** *J Cell Sci* 1968, **3:**391-404.

161. Lawrence PA: **Cellular differentiation and pattern formation during metamorphosis of the milkweed bug Oncopeltus.** *Dev Biol* 1969, **19:**12-40.

162. Riddiford LM: **Prevention of Metamorphosis by Exposure of Insect Eggs to Juvenile Hormone Analogs.** *Science* 1970, **167:**287-&.

163. Willis JH, Lawrence PA: **Deferred Action of Juvenile Hormone.** *Nature* 1970, **225:**81-83.

164. Masner P, Bowers WS, Kalin M, Muhle T: **Effect of precocene II on the endocrine regulation of development and reproduction in the bug, *Oncopeltus fasciatus*.** *Gen Comp Endocrinol* 1979, **37:**156-166.

165. Erezyilmaz DF, Riddiford LM, Truman JW: **The pupal specifier broad directs progressive morphogenesis in a direct-developing insect.** *Proceedings of the National Academy of Sciences of the United States of America* 2006, **103:**6925-6930.

166. Gilbert LI: **Halloween genes encode P450 enzymes that mediate steroid hormone biosynthesis in Drosophila melanogaster.** *Mol Cell Endocrinol* 2004, **215:**1-10.

167. Rewitz K, Rybczynski R, Warren J, Gilbert L: **Developmental expression of Manduca shade, the P450 mediating the final step in molting hormone synthesis.** *Mol Cell Endocrinol* 2006, **247:**166-174.

168. Huet F, Ruiz C, Richards G: **Sequential gene activation by ecdysone in Drosophila melanogaster: the hierarchical equivalence of early and early late genes.** *Development* 1995, **121:**1195-1204.

169. Ureña E, Manjón D, Franch-Marro X, Martin D: **Transcription factor E93 specifies adult metamorphosis in hemimetabolous and holometabolous insects. .** *Proc Natl Acad Sci U S A* 2014, **111:**7024-7029.

170. Lam GT, Jiang C, Thummel CS: **Coordination of larval and prepupal gene expression by the DHR3 orphan receptor during Drosophila metamorphosis.** *Development* 1997, **124:**1757-1769.

171. Stone BL, Thummel CS: **The Drosophila 78c Early-Late Puff Contains E78, an Ecdysone-Inducible Gene That Encodes a Novel Member of the Nuclear Hormone-Receptor Superfamily.** *Cell* 1993, **75:**307-320.

172. King-Jones K, Charles J, Lam G, Thummel C: **The ecdysone-induced DHR4 orphan receptor coordinates growth and maturation in Drosophila.** *Cell* 2005, **121:**773-784.

173. Woodard C, Baehrecke E, Thummel C: **A molecular mechanism for the stage specificity of the Drosophila prepupal genetic response to ecdysone.** *Cell* 1994, **79:**607-615.

174. Baker K, Shewchuk L, Kozlova T, Makishima M, Hassell A, Wisely B, Caravella J, Lambert M, Reinking J, Krause H, et al: **The Drosophila orphan nuclear receptor DHR38 mediates an atypical ecdysteroid signaling pathway.** *Cell* 2003, **113:**731-742.

175. Truman J, Rountree D, Reiss S, Schwartz L: **Ecdysteroids regulate the release and action of eclosion hormone in the tobacco hornworm, Manduca sexta (L.)** *J Insect Physiol* 1983, **29:**895–900.

176. Zitnan D, Kingan TG, Hermesman JL, Adams ME: **Identification of ecdysis-triggering hormone from an epitracheal endocrine system.** *Science* 1996, **271:**88-91.

177. Ewer J, Truman JW: **Increases in cyclic 3',5'-guanosine monophosphate (cGMP) occur at ecdysis in an evolutionarily conserved crustacean cardioactive peptide-immunoreactive insect neuronal network.** *Journal of Comparative Neurology* 1996, **370:**330-341.

178. Dewey E, McNabb S, Ewer J, Kuo G, Takanishi C, Truman J, Honegger H: **Identification of the gene encoding bursicon, an insect neuropeptide responsible for cuticle sclerotization and wing spreading.** *Current Biology* 2004, **14:**1208-1213.

179. Willis JH: **Structural cuticular proteins from arthropods: Annotation, nomenclature, and sequence characteristics in the genomics era.** In *Insect Biochemistry and Molecular Biology*, vol. 40. pp. 189-204; 2010:189-204.

180. Zhu F, Gujar H, Gordon JR, Haynes KF, Potter MF, Palli SR: **Bed bugs evolved unique adaptive strategy to resist pyrethroid insecticides.** *Sci Rep* 2013, **3:**1456.

181. Ioannidou ZS, Theodoropoulou MC, Papandreou NC, Willis JH, Hamodrakas SJ: **CutProtFam-Pred: Detection and classification of putative structural cuticular proteins from sequence alone, based on profile Hidden Markov Models.** *Insect Biochemistry and Molecular Biology* 2014, **52:**51-59.

182. Willis JH, Papandreou NC, Iconomidou VA, Hamodrakas SJ: **5 Cuticular Proteins.** In *Insect Molecular Biology and Biochemistry.* Academic Press San Diego; 2012: 134-166

183. Parkash R, Aggarwal DD, Lambhod C, Singh D: **Divergence of water balance mechanisms and acclimation potential in body color morphs of Drosophila ananassae.** *J Exp Zool A Ecol Genet Physiol* 2014, **321:**13-27.

184. Wright TR: **The genetics of biogenic amine metabolism, sclerotization, and melanization in Drosophila melanogaster.** *Adv Genet* 1987, **24:**127-222.

185. Wittkopp PJ, Vaccaro K, Carroll SB: **Evolution of yellow gene regulation and pigmentation in Drosophila.** *Current Biology* 2002, **12:**1547-1556.

186. Wittkopp PJ, Carroll SB, Kopp A: **Evolution in black and white: genetic control of pigment patterns in Drosophila.** *Trends in Genetics* 2003, **19:**495-504.

187. Liu J, Lemonds TR, Marden JH, Popadic A: **A Pathway Analysis of Melanin Patterning in a Hemimetabolous Insect.** *Genetics* 2016, **203:**403-413.

188. Liu J, Lemonds TR, Popadic A: **The genetic control of aposematic black pigmentation in hemimetabolous insects: insights from Oncopeltus fasciatus.** *Evolution & Development* 2014, **16:**270-277.

189. Lawrence PA: **Some new mutants of large milkweed bug *Oncopeltus fasciatus* Dall.** *Genetical Research* 1970, **15:**347-350.

190. Cervera A, Maymó AC, Martínez-Pardo R, Garcerá MD: **Antioxidant Enzymes in *Oncopeltus fasciatus* (Heteroptera: Lygaeidae) Exposed to Cadmium.** *Environmental Entomology* 2003, **32:**705-710.

191. Benoit JB, Hansen IA, Szuter EM, Drake LL, Burnett DL, Attardo GM: **Emerging roles of aquaporins in relation to the physiology of blood-feeding arthropods.** *J Comp Physiol B* 2014, **184:**811-825.

192. Benoit JB, Lopez-Martinez G (Eds.): **Role of conventional and unconventional stress proteins during the response of insects to traumatic environmental conditions**; 2012.

193. Werck-Reichhart D, Feyereisen R: **Cytochromes P450: a success story.** *Genome Biol* 2000, **1:**REVIEWS3003.

194. Peterson JA, Graham SE: **A close family resemblance: the importance of structure in understanding cytochromes P450.** *Structure* 1998, **6:**1079-1085.

195. Shelton PM, Lawrence PA: **Structure and development of ommatidia in Oncopeltus fasciatus.** *J Embryol Exp Morphol* 1974, **32:**337-353.

196. Dudek FE: **The visual response from the compound eye of *Oncopeltus fasciatus*: effects of temperature and sensory adaptation.** *J Insect Physiol* 1975, **21:**517-528.

197. Bennett RR, Ruck P: **Spectral sensitivities of dark- and light-adapted Notonecta compound eyes.** *J Insect Physiol* 1970, **16:**83-88.

198. Bruckmoser P: **Die spektrale Empfindlichkeit einzelner Sehzellen des Rückenschwimmers Notonecta glauca L. (Heteroptera).** *Z Vgl Physiol* 1968, **59:**187-204.

199. Wakakuwa M, Stewart F, Matsumoto Y, Matsunaga S, Arikawa K: **Physiological basis of phototaxis to near-infrared light in Nephotettix cincticeps.** *J Comp Physiol A Neuroethol Sens Neural Behav Physiol* 2014, **200:**527-536.

200. Gao N, Foster RG, Hardie J: **Two opsin genes from the vetch aphid, *Megoura viciae*.** *Insect Mol Biol* 2000, **9:**197-202.

201. Brody T, Cravchik A: **Drosophila melanogaster G Protein-coupled Receptors.** *J Cell Biol* 2000, **150:**F83-F88.

202. Colbourne JK, Pfrender ME, Gilbert D, Thomas WK, Tucker A, Oakley TH, Tokishita S, Aerts A, Arnold GJ, Basu MK, et al: **The ecoresponsive genome of *Daphnia pulex*.** *Science* 2011, **331:**555-561.

203. Eriksson BJ, Fredman D, Steiner G, others: **Characterisation and localisation of the opsin protein repertoire in the brain and retinas of a spider and an onychophoran.** *BMC Evol Biol* 2013, **13:**186.

204. Hering L, Mayer G: **Analysis of the opsin repertoire in the tardigrade Hypsibius dujardini provides insights into the evolution of opsin genes in panarthropoda.** *Genome Biol Evol* 2014, **6:**2380-2391.

205. Löytynoja A, Goldman N: **webPRANK: a phylogeny-aware multiple sequence aligner with interactive alignment browser.** *BMC Bioinformatics* 2010, **11:**579.

206. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T: **trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses.** *Bioinformatics* 2009, **25:**1972-1973.

207. Sánchez R, Serra F, Tárraga J, Medina I, Carbonell J, Pulido L, de María A, Capella-Gutíerrez S, Huerta-Cepas J, Gabaldón T, et al: **Phylemon 2.0: a suite of web-tools for molecular evolution, phylogenetics, phylogenomics and hypotheses testing.** *Nucleic Acids Res* 2011, **39:**W470-474.

208. Su CY, Menuz K, Carlson JR: **Olfactory perception: receptors, cells, and circuits.** *Cell* 2009, **139:**45-59.

209. Touhara K, Vosshall LB: **Sensing odorants and pheromones with chemosensory receptors.** *Annu Rev Physiol* 2009, **71:**307-332.

210. Jones WD, Cayirlioglu P, Kadow IG, Vosshall LB: **Two chemosensory receptors together mediate carbon dioxide detection in Drosophila.** *Nature* 2007, **445:**86-90.

211. Kwon JY, Dahanukar A, Weiss LA, Carlson JR: **The molecular basis of CO2 reception in Drosophila.** *Proc Natl Acad Sci U S A* 2007, **104:**3574-3578.

212. Lu T, Qiu YT, Wang G, Kwon JY, Rutzler M, Kwon HW, Pitts RJ, van Loon JJ, Takken W, Carlson JR, Zwiebel LJ: **Odor coding in the maxillary palp of the malaria vector mosquito Anopheles gambiae.** *Curr Biol* 2007, **17:**1533-1544.

213. Liman ER, Zhang YV, Montell C: **Peripheral coding of taste.** *Neuron* 2014, **81:**984-1000.

214. Kirkness EF, Haas BJ, Sun W, Braig HR, Perotti MA, Clark JM, Lee SH, Robertson HM, Kennedy RC, Elhaik E, et al: **Genome sequences of the human body louse and its primary endosymbiont provide insights into the permanent parasitic lifestyle.** *Proc Natl Acad Sci U S A* 2010, **107:**12168-12173.

215. Robertson HM, Wanner KW: **The chemoreceptor superfamily in the honey bee, Apis mellifera: expansion of the odorant, but not gustatory, receptor family.** *Genome Res* 2006, **16:**1395-1403.

216. Smadja C, Shi P, Butlin RK, Robertson HM: **Large gene family expansions and adaptive evolution for odorant and gustatory receptors in the pea aphid, Acyrthosiphon pisum.** *Mol Biol Evol* 2009, **26:**2073-2086.

217. Mesquita RD, Vionette-Amaral RJ, Lowenberger C, Rivera-Pomar R, Monteiro FA, Minx P, Spieth J, Carvalho AB, Panzera F, Lawson D, et al: **Genome of *Rhodnius prolixus*, an insect vector of Chagas disease, reveals unique adaptations to hematophagy and parasite infection.** *Proc Natl Acad Sci USA* 2015, **112:**14936-14941.

218. Penalva-Arana DC, Lynch M, Robertson HM: **The chemoreceptor genes of the waterflea Daphnia pulex: many Grs but no Ors.** *BMC Evol Biol* 2009, **9:**79.

219. Hoy MA, Waterhouse RM, Wu K, Estep AS, Ioannidis P, Palmer WJ, Pomerantz AF, Simao FA, Thomas J, Jiggins FM, et al: **Genome sequencing of the phytoseiid predatory mite *Metaseiulus occidentalis* reveals completely atomized Hox genes and superdynamic intron evolution.** *Genome Biol Evol* 2016, **8:**1762-1775.

220. Benton R, Vannice KS, Gomez-Diaz C, Vosshall LB: **Variant ionotropic glutamate receptors as chemosensory receptors in Drosophila.** *Cell* 2009, **136:**149-162.

221. Croset V, Rytz R, Cummins SF, Budd A, Brawand D, Kaessmann H, Gibson TJ, Benton R: **Ancient protostome origin of chemosensory ionotropic glutamate receptors and the evolution of insect taste and olfaction.** *PLoS Genet* 2010, **6:**e1001064.

222. 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.

223. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, et al: **Clustal W and Clustal X version 2.0.** *Bioinformatics* 2007, **23:**2947-2948.

224. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O: **New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0.** *Syst Biol* 2010, **59:**307-321.

225. Robertson HM: **The insect chemoreceptor superfamily in Drosophila pseudoobscura: Molecular evolution of ecologically-relevant genes over 25 million years** *J Insect Sci* 2009, **9:**18.

226. Terrapon N, Li C, Robertson HM, Ji L, Meng X, Booth W, Chen Z, Childers CP, Glastad KM, Gokhale K, et al: **Molecular traces of alternative social organization in a termite genome.** *Nat Commun* 2014, **5:**3636.

227. Miyamoto T, Amrein H: **Diverse roles for the Drosophila fructose sensor Gr43a.** *Fly (Austin)* 2014, **8:**19-25.

228. Delventhal R, Carlson JR: **Bitter taste receptors confer diverse functions to neurons.** *Elife* 2016, **5**.

229. Weiss LA, Dahanukar A, Kwon JY, Banerjee D, Carlson JR: **The molecular and cellular basis of bitter taste in Drosophila.** *Neuron* 2011, **69:**258-272.

230. Enjin A, Zaharieva EE, Frank DD, Mansourian S, Suh GS, Gallio M, Stensmyr MC: **Humidity Sensing in Drosophila.** *Curr Biol* 2016, **26:**1352-1358.

231. Knecht ZA, Silbering AF, Ni L, Klein M, Budelli G, Bell R, Abuin L, Ferrer AJ, Samuel AD, Benton R, Garrity PA: **Distinct combinations of variant ionotropic glutamate receptors mediate thermosensation and hygrosensation in Drosophila.** *Elife* 2016, **5**.

232. Ni L, Klein M, Svec KV, Budelli G, Chang EC, Ferrer AJ, Benton R, Samuel AD, Garrity PA: **The Ionotropic Receptors IR21a and IR25a mediate cool sensing in Drosophila.** *Elife* 2016, **5**.

233. Prieto-Godino LL, Rytz R, Bargeton B, Abuin L, Arguello JR, Peraro MD, Benton R: **Olfactory receptor pseudo-pseudogenes.** *Nature* 2016, **539:**93-97.

234. Silbering AF, Rytz R, Grosjean Y, Abuin L, Ramdya P, Jefferis GS, Benton R: **Complementary function and integrated wiring of the evolutionarily distinct Drosophila olfactory subsystems.** *J Neurosci* 2011, **31:**13357-13375.

235. Koh TW, He Z, Gorur-Shandilya S, Menuz K, Larter NK, Stewart S, Carlson JR: **The Drosophila IR20a clade of ionotropic receptors are candidate taste and pheromone receptors.** *Neuron* 2014, **83:**850-865.

236. Stewart S, Koh TW, Ghosh AC, Carlson JR: **Candidate ionotropic taste receptors in the Drosophila larva.** *Proc Natl Acad Sci U S A* 2015, **112:**4195-4201.

237. Bansal R, Michel AP: **Core RNAi Machinery and *Sid1*, a component for systemic RNAi, in the hemipteran insect, *Aphis glycines*.** *Int J Mol Sci* 2013, **14:**3786-3801.

238. Jaubert-Possamai S, Rispe C, Tanguy S, Gordon K, Walsh T, Edwards O, Tagu D: **Expansion of the miRNA pathway in the hemipteran insect *Acyrthosiphon pisum*.** *Mol Biol Evol* 2010, **27:**979-987.

239. LaChance LE, Richard RD: **Irradiation of sperm and oocytes in *Oncopeltus fasciatus* (Hemiptera: Lygaeidae): sex ratio, fertility, and chromosome aberrations in the F1 progeny.** *Can J Genet Cytol* 1973, **15:**713-721.

240. Shukla JN, Nagaraju J: **Doublesex: a conserved downstream gene controlled by diverse upstream regulators.** *J Genet* 2010, **89:**341-356.

241. Geuverink E, Beukeboom LW: **Phylogenetic distribution and evolutionary dynamics of the sex determination genes doublesex and transformer in insects.** *Sex Dev* 2014, **8:**38-49.

242. An W, Cho S, Ishii H, Wensink PC: **Sex-specific and non-sex-specific oligomerization domains in both of the doublesex transcription factors from Drosophila melanogaster.** *Mol Cell Biol* 1996, **16:**3106-3111.

243. Biewer M, Schlesinger F, Hasselmann M: **The evolutionary dynamics of major regulators for sexual development among Hymenoptera species.** *Front Genet* 2015, **6:**124.

244. Shukla JN, Palli SR: **Doublesex target genes in the red flour beetle, Tribolium castaneum.** *Sci Rep* 2012, **2:**948.

245. Sarno F, Ruiz MF, Eirin-Lopez JM, Perondini AL, Selivon D, Sanchez L: **The gene transformer-2 of Anastrepha fruit flies (Diptera, Tephritidae) and its evolution in insects.** *BMC Evol Biol* 2010, **10:**140.

246. Siegal ML, Baker BS: **Functional conservation and divergence of intersex, a gene required for female differentiation in *Drosophila melanogaster*.** *Dev Genes Evol* 2005, **215:**1-12.

247. Clynen E, Ciudad L, Belles X, Piulachs MD: **Conservation of fruitless' role as master regulator of male courtship behaviour from cockroaches to flies.** *Dev Genes Evol* 2011, **221:**43-48.

248. Suzuki MG, Imanishi S, Dohmae N, Nishimura T, Shimada T, Matsumoto S: **Establishment of a novel in vivo sex-specific splicing assay system to identify a trans-acting factor that negatively regulates splicing of *Bombyx mori* dsx female exons.** *Mol Cell Biol* 2008, **28:**333-343.

249. Conrad T, Cavalli FM, Vaquerizas JM, Luscombe NM, Akhtar A: **Drosophila dosage compensation involves enhanced Pol II recruitment to male X-linked promoters.** *Science* 2012, **337:**742-746.

250. Verhulst EC, van de Zande L, Beukeboom LW: **Insect sex determination: it all evolves around transformer.** *Curr Opin Genet Dev* 2010, **20:**376-383.

251. Gempe T, Beye M: **Function and evolution of sex determination mechanisms, genes and pathways in insects.** *Bioessays* 2011, **33:**52-60.

252. Duncan EJ, Gluckman PD, Dearden PK: **Epigenetics, plasticity, and evolution: How do we link epigenetic change to phenotype?** *J Exp Zool B Mol Dev Evol* 2014, **322:**208-220.

253. Jones PA: **Functions of DNA methylation: islands, start sites, gene bodies and beyond.** *Nat Rev Genet* 2012, **13:**484-492.

254. Foret S, Kucharski R, Pellegrini M, Feng S, Jacobsen SE, Robinson GE, Maleszka R: **DNA methylation dynamics, metabolic fluxes, gene splicing, and alternative phenotypes in honey bees.** *Proc Natl Acad Sci U S A* 2012, **109:**4968-4973.

255. Lyko F, Foret S, Kucharski R, Wolf S, Falckenhayn C, Maleszka R: **The honey bee epigenomes: differential methylation of brain DNA in queens and workers.** *PLoS Biol* 2010, **8:**e1000506.

256. Sati S, Tanwar VS, Kumar KA, Patowary A, Jain V, Ghosh S, Ahmad S, Singh M, Reddy SU, Chandak GR, et al: **High resolution methylome map of rat indicates role of intragenic DNA methylation in identification of coding region.** *PLoS One* 2012, **7:**e31621.

257. Herb BR, Wolschin F, Hansen KD, Aryee MJ, Langmead B, Irizarry R, Amdam GV, Feinberg AP: **Reversible switching between epigenetic states in honeybee behavioral subcastes.** *Nat Neurosci* 2012, **15:**1371-1373.

258. Wang Y, Jorda M, Jones PL, Maleszka R, Ling X, Robertson HM, Mizzen CA, Peinado MA, Robinson GE: **Functional CpG methylation system in a social insect.** *Science* 2006, **314:**645-647.

259. Walsh TK, Brisson JA, Robertson HM, Gordon K, Jaubert-Possamai S, Tagu D, Edwards OR: **A functional DNA methylation system in the pea aphid, Acyrthosiphon pisum.** *Insect Mol Biol* 2010, **19 Suppl 2:**215-228.

260. Guo JU, Su Y, Zhong C, Ming GL, Song H: **Hydroxylation of 5-methylcytosine by TET1 promotes active DNA demethylation in the adult brain.** *Cell* 2011, **145:**423-434.

261. Tahiliani M, Koh KP, Shen Y, Pastor WA, Bandukwala H, Brudno Y, Agarwal S, Iyer LM, Liu DR, Aravind L, Rao A: **Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1.** *Science* 2009, **324:**930-935.

262. Feng S, Cokus SJ, Zhang X, Chen PY, Bostick M, Goll MG, Hetzel J, Jain J, Strauss SH, Halpern ME, et al: **Conservation and divergence of methylation patterning in plants and animals.** *Proc Natl Acad Sci U S A* 2010, **107:**8689-8694.

263. Suzuki MM, Kerr AR, De Sousa D, Bird A: **CpG methylation is targeted to transcription units in an invertebrate genome.** *Genome Res* 2007, **17:**625-631.

264. Zemach A, McDaniel IE, Silva P, Zilberman D: **Genome-wide evolutionary analysis of eukaryotic DNA methylation.** *Science* 2010, **328:**916-919.

265. Kass SU, Landsberger N, Wolffe AP: **DNA methylation directs a time-dependent repression of transcription initiation.** *Curr Biol* 1997, **7:**157-165.

266. Laurent L, Wong E, Li G, Huynh T, Tsirigos A, Ong CT, Low HM, Kin Sung KW, Rigoutsos I, Loring J, Wei CL: **Dynamic changes in the human methylome during differentiation.** *Genome Res* 2010, **20:**320-331.

267. Elango N, Hunt BG, Goodisman MA, Yi SV: **DNA methylation is widespread and associated with differential gene expression in castes of the honeybee, Apis mellifera.** *Proc Natl Acad Sci U S A* 2009, **106:**11206-11211.

268. Chipman AD, Ferrier DEK, Brena C, Qu J, Hughes DST, Schröder R, Torres-Oliva M, Znassi N, Jiang H, Almeida FC, et al: **The first myriapod genome sequence reveals conservative arthropod gene content and genome organisation in the centipede *Strigamia maritima*.** *PLoS Biol* 2014, **12:**e1002005.

269. Fraley C, Raftery AE: **Enhanced model-based clustering, density estimation, and discriminant analysis software: MCLUST.** *Journal of Classification* 2003, **20:**263-286.

270. Talbert PB, Henikoff S: **Environmental responses mediated by histone variants.** *Trends Cell Biol* 2014, **24:**642-650.

271. Rider SD, Srinivasan DG, Hilgarth RS: **Chromatin-remodelling proteins of the pea aphid, *Acyrthosiphon pisum* (Harris).** *Insect Molecular Biology* 2010, **19:**201-214.

272. Garcia-Perez JL, Widmann TJ, Adams IR: **The impact of transposable elements on mammalian development.** *Development* 2016, **143:**4101-4114.

273. Kriventseva EV, Tegenfeldt F, Petty TJ, Waterhouse RM, Simão FA, Pozdnyakov IA, Ioannidis P, Zdobnov EM: **OrthoDB v8: update of the hierarchical catalog of orthologs and the underlying free software.** *Nucl Acids Res* 2015, **43:**D250-D256.

274. Ioannidis P, Simao FA, Waterhouse RM, Manni M, Seppey M, Robertson HM, Misof B, Niehuis O, Zdobnov EM: **Genomic features of the damselfly *Calopteryx splendens* representing a sister clade to most insect orders.** *Genome Biol Evol* 2017, **9:**415-430.

275. Zdobnov EM, Tegenfeldt F, Kuznetsov D, Waterhouse RM, Simão FA, Ioannidis P, Seppey M, Loetscher A, Kriventseva EV: **OrthoDB v9.1: cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs.** *Nucleic Acids Res* 2017, **45:**D744-D749.

276. Stamatakis A: **RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models.** *Bioinformatics* 2006, **22:**2688-2690.

277. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput.** *Nucl Acids Res* 2004, **32:**1792-1797.

278. Rice P, Longden I, Bleasby A: **EMBOSS: The European Molecular Biology Open Software Suite.** *Trends Genet* 2000, **16:**276-277.