The genome of the water strider *Gerris buenoi* reveals expansions of gene repertoires associated with adaptations to life on the water

Supplementary Data

Immune genes

While mammals have both innate and adaptive immune response, only innate immune response has been described in arthropods [1]. In particular, the Toll and IMD (Immunodeficiency) pathways are the two major regulators of the immune response known in arthropods [2-4] which act by regulating the expression of other effector molecules such as antimicrobial peptides (AMPs).

In the *Gerris buenoi* genome we could annotate more than 60 immune genes, including orthologs of all components of the Toll signalling pathway, which is activated mainly by Gram-positive bacteria and fungi [5, 6]. However, whereas the Toll1-4 receptors were only represented by a single ortholog called Toll1, six Toll9 paralogs were found which raises important questions about a possible adaptation to gram-positive bacteria present in the water. On the other hand, IMD pathway responds mainly to Gram-negative bacteria infection [5, 6] but many of its genes, including *IMD*, *dFADD*, *Dredd*, and *Relish* could not be found in the first sequenced hemipteran, *Acyrthosiphon pisum [7, 8].* Further sequencing of other hemipterans extended this absence to the kissing bug *Rodnius prolixus* and the bed bug *Cimex lectularius*, as well as the pest species *Diaphorina citri*, *Pachypsylla venusta* and *Halyomorpha halys*. However, among the 60 immune genes annotated in the genome of *Gerris buenoi*, we could identify a homolog of IMD, a unique feature amongst sequenced Hemiptera species only shared with recently sequenced true bug *Oncopeltus fasciatus* (Supplementary Figure 7) [9]. However, like in *Oncopeltus fasciatus*, the important IMD pathway components dFADD and Kenny seem to be missing in *Gerris buenoi*.

Further research is required to elucidate how the IMD pathway functions in water striders and why IMD has been conserved in *Gerris* while it has been lost in other hemipterans.

Despite the lack of shared components between Toll and IMD, both pathways can regulate immune response through regulation of antimicrobial peptides (AMPs). Antimicrobial peptide (AMPs) families prevent the invasion of potential pathogens playing a fundamental role on innate immunity [10]. However, AMP families differ greatly among groups of insects [11] and only two defensin-like, one lysozyme and 6 of the Hemiptera-specific Serosins [12] could be identified. We failed to identify any attacins, hemiptericins or thaumatins in the *Gerris buenoi* genome. These results suggest that following Gerromorpha invasion of water environment they have been faced with a myriad of new potential pathogens, which may have accelerated Gerromorpha's AMPs divergence.

Finally, we could annotate an ortholog of the innate immune response gene gamma-interferon-inducible thiol reductase (*gilt*) in *Gerris buenoi* genome. Despite only innate immune response has been classically described in arthropods, recent studies on *Drosophila melanogaster* have shown that *gilt* ortholog gene has a role on adaptive immune response in flies [1]. However, the exact mechanism of *gilt* function in immune response remains unknown. Moreover, in water striders including *Gerris buenoi*, although no immune role of *gilt* has been tested yet, knockdown analyses using RNA interference have shown an important new role in leg growth and adaptation [13]. These findings raise interesting questions about the functional divergence of arthropod immune system.

Early Developmental Genes

One of the main reasons for choosing to sequence the *Gerris buenoi* genome was due to its emerging status as a developmental model system [14]. Therefore, it was of particular interest to analyze its developmental gene content. In total 24 genes that are known, in other insects, to be involved in developmental processes were manually annotated (Supplementary Table 10). These include both genes encoding transcription factors and members of signaling pathways. These genes are identified and named as distinct development genes by the nomenclature from *Drosophila melanogaster* (Supplementary Table 11). *Gerris buenoi* has evidence of a canonical insect developmental pathway and can be expected to contain all components required to establish a normal anterior/posterior axis pattern. Compared to the later acting genes, the early developmental genes identified in *Gerris buenoi* show greater divergence from those found in

Drosophila melanogaster and *Tribolium castaneum* (Supplementary Table 12), consistent with observations between *Drosophila* species [15]. Developmental genes previously identified in *Limnoporus dissortis* (e.g. *decapentaplegic*) were also identified in the *Gerris buenoi* genome [16] confirming the presence of canonical insect developmental toolkit in this species. No duplication in the early development genes was observed. Early patterning genes appear conserved form what is known in other insects. As expected, there is no *bicoid* orthologue. Other genes known in *Drosophila* but not found in other insects, such as swallow are also not found in *Gerris*, such is the case of *caudal*. However, we suspect due to the identification of tailless that the absence of caudal from the genome is due to incomplete coverage of the sequencing effort, rather than an actual absence of the gene in the genome. We identified gene models for the terminal patterning genes *torso*, and *torso-like* in *Gerris buenoi*. Although models homologous to PTTH were identified they were not well supported. However, is it more than likely that *Gerris buenoi* possess a PTTH orthologue given that PTTH orthologues are found in other hemipterans. As with other Hemiptera, we could not find a model for *trunk*.

Nuclear receptors and bHLH-PAS proteins

We have annotated the genome of *Gerris buenoi* for all the genes of two families of ligand-dependent transcription factors: nuclear receptors and bHLH-PAS proteins. These regulators share many characteristics, such as response to small lipophilic ligands that can act either as signalling molecules or as xenobiotics and heterodimerisation factors with other members of their family. Numerous cross-talk interactions are known between nuclear receptors and bHLH-PAS proteins.

All but one of the 21 nuclear receptor genes expected for an insect were found in the genome of *Gerris buenoi*. The missing gene E78 is also absent in *Pediculus humanus* [17] but is present in the genome of *Acyrthosiphon pisum* [18, 19]. We found 3 NR0 genes (knirps-related, eagle), as in *Pediculus humanus* and *Apis mellifera* [20]. Based on the work of [21], we could also identify all the isoforms of ECR and NR2E6 genes.

The genome of *Gerris buenoi* contains at least 10 genes of the bHLH-PAS family. The gene *tango* (*tgo*) was not found, whereas it is present in the genome of the *Acyrthosiphon pisum* [22]. This absence is surprising, since *tgo* is the homolog of ARNT, which is the heterodimeric partner of several members of this family in mammals. Since the gene called «*germ cell-expressed*» (*gce*) in *Drosophila* is known to be a diptera specific duplication of *Methoprene-tolerant* (*Met*), its absence in the genome of *Gerris buenoi* was expected. The gene *single-minded* (*sim*) is duplicated, as in

Tribolium castaneum [23].

In conclusion, we found a strong conservation of the number and identity of nuclear receptors and bHLH-PAS proteins with other insects.

Insulin/TOR signalling pathways

The Insulin and TOR pathways function together as an integrated metabolic signalling pathway that is known to coordinate hormonal and nutritional signals in developing animals [24-26]. This facilitates the complex regulation of several fundamental molecular and cellular processes including transcription [27, 28], translation, cell stress, autophagy, and physiological states, including aging, starvation, hormonal regulation, as well as both organism-wide and tissue-specific growth [26-31]. In insects, these pathways have been implicated in the developmental regulation of complex nutrient-dependent phenotypes ranging from beetle horns to the social castes of termites and bees [32-34]. For example, in beetles, the insulin receptor is known to be a critical regulator of appendage growth and it has been proposed that downstream transcription factors of the pathway (Foxo), can mediate organ-specific sizing and growth [35, 36]. Taken together, the interplay between these two pathways may play an integral role in the growth and sizing of the different legs, and perhaps, even sexually dimorphic sized appendages found across the morphologically diverse array of water strider species. For this reason, we searched for and annotated various key players of this pathway. We found that *Gerris buenoi* possesses all components of this pathway including the *forkhead box protein O* (*foxo*), *insulin receptor 1* (*InR1*), *insulin receptor 2*(*InR2*), the *insulin receptor substrate Chico*, the negative insulin pathway regulator *Phosphatase and Tensine homologue* (*Pten*), *Rheb/Ras homolog enriched in brain* (*Rheb*), the *S6 kinase* (*S6k*), *Target of Rapamycin* (*Tor*), the binding protein of the translation initiation factor eI4E (*4E-BP/Thor*), *Tuberous sclerosis complex 1* and *2* (*Tsc1 & Tsc2/gigas*), the phosphoinositide-3-OH-kinase-dependent serine/threonine protein kinase *Akt1/Pkb*, the amino acid transporter *Slimfast* (*slif*) and two Phosphoinositide 3-kinases (*Pi3K92E* & *Pi3K21B*). In addition to this, *Gerris buenoi* appears to have an additional, third, insulin receptor of unknown function and no known ortholog in insects. Therefore, the water strider *Gerris buenoi* possesses the entire Insulin/TOR toolkit, which would be a potential target for future research into nutrient-dependent differential body-plan growth and evolution in water striders.

Wnt Signaling Pathway

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 [37]. Wnt signalling affects cell migration and segment polarity as well as segment patterning in most arthropods [38]. Surveying and comparing the gene repertoire of conserved gene families within and between taxonomic groups is the first step towards understanding their function during development and evolution.

Here we curated gene models for the main components of the Wnt signalling pathway and confirmed their orthology by phylogenetic analysis. We found 6 Wnt ligand subfamilies, three Frizzled transmembrane receptor subfamilies, the co-receptor *arrow*, and the downstream components *armadillo*/*beta-catenin*, *dishevelled*, *arrow*, *axin*, and *shaggy*/*GSK*-3. All of these genes were present in single copy in the assembly.

The *Gerris* Wnt ligand repertoire is comparable to other hemipterans and holometabolous insect species that have been analysed in detail. This supports observations of a reduction in the ligand repertoire in insects compared to an inferred ancestral complement of 17 subfamilies, with most extant Metazoan retaining ligands from 11-12 subfamilies. 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.

A total of 18 models for the main Wnt signalling genes were curated in the *Gerris buenoi* assembly (Supplementary Table 13). The gene models generated by the MAKER pipeline were a very good start for the curation process in most cases, where most of the time only the 5' end of the models had to be edited by changing the translation start or adding upstream exons. The exceptions to this were the *dishevelled* isoforms where, despite very strong RNA-seq support for the complete model, only a small 5-exon model (for a gene with 16 exons in this species) for the middle part of the gene was present in the automated set. Despite curation, the models of three genes are incomplete. Similarly, *WntA* was missing the first exon in an upstream gap, and the *armadillo* model was missing the N-terminal region due to a gap directly upstream of the model. The third gene, *GSK-3 beta*, was split across two scaffolds despite strong RNA-seq support, with part 2 of this model filling the complete scaffold 10229 and yet still missing fragments at both ends.

All models were isolated on individual scaffolds, with the exception of *axin* and *arrow*. Interestingly, this linkage is not found in *Drosophila melanogaster, Tribolium castaneum,* or other

i5k pilot project hemipteroid species surveyed to date. On the other hand, the absence of the ancient synteny of *wingless-Wnt6-Wnt10* [39], which was wholly or partially confirmed in other i5k pilot hemipteroid species, is likely due to limitations in the current draft assembly. Regarding gene copy number, it is worth noting that *armadillo*, which encodes an intracellular transducer in the Wnt pathway, is represented by a single ortholog in the current assembly. As many insects, including other heteropterans, have two copies of *armadillo* (*Drosophila, Tribolium, Cimex, Oncopeltus*), it is surprising that there is no evidence for a second gene in *Gerris*.

We identified 6 *Wnt* gene subfamilies in the *Gerris* assembly, all with single copy genes: *wingless/Wnt1, Wnt5, Wnt*7, *Wnt*8, *Wnt10* and *WntA*. This is identical to the ligand subfamily representation in *Oncopeltus fasciatus*, with the slight difference that there has been a duplication in *Oncopeltus Wnt8* [9]. There were also only six *Wnt* gene subfamilies found in the pea aphid (*Acyrthosiphon pisum*), although for a slightly different constellation of subfamilies: *wingless/Wnt1, Wnt5, Wnt* 7, *Wnt11, Wnt16* and *WntA* [19]. Together with earlier observations [39], this report supports the idea that members of the Hemiptera have the fewest *Wnt* gene families reported in insects, with some of these losses perhaps having occurred relatively recently and independently in this clade.

Three models were curated for the *frizzled* (*fz*) transmembrane receptor families: *frizzled*, *frizzled-2*, and *frizzled-3*. These correspond to three of the four ancient *fz* families expected to have been present in the common ancestor of arthropods: *fz*, *fz2*, *fz3*, *fz4* [40]. The loss of *fz4* was also observed in *Oncopeltus fasciatus* [9] and *Acyrthosiphon pisum* [19].

Cysteine peptidases from the papain C1 family

Cysteine peptidases from the papain C1 family (MEROPS classification [41]) are important lysosomal cathepsins, and participate as regulators and signaling molecules in a large number of biological processes [42]. In addition, cysteine cathepsins in a limited number of insect groups are important digestive enzymes evolved from lysosomal ancestors [43, 44]. In Cucujiformia beetles, digestive cysteine cathepsins are an evolutionary response to a seed diet rich in serine peptidase inhibitors [43, 45]. In the case of true bugs, it is proposed that their sap-sucking ancestors lost digestive serine peptidases in adapting to plant sap, and the adaptation of cysteine cathepsins for digestive functions is a consequence of a return to a protein diet [46]. A detailed study of cysteine cathepsins in the beetles *Tenebrio molitor* and *Tribolium castaneum* (Coleoptera: Tenebrionidae) revealed expansions of genes encoding cysteine digestive cathepsins [47, 48]. Cysteine cathepsins

in *T. castaneum* larvae are important components of adaptive responses in overcoming the effect of dietary protease inhibitors [49].

There are few publications of cysteine peptidases in Heteroptera. Most of the early publications suggested that cysteine peptidases are the major digestive peptidases in several families of this insect order (see [43, 44]), such as *Reduviidae*, where digestive cathepsins L and B were identified in two *Triatoma* species [50, 51]. Sequencing the *Rhodnius prolixus* gut transcriptome revealed 11 cysteine peptidases expressed in the gut [52]. We are unaware of any publications on digestive peptidases of the bugs from the family Gerridae, and the specific biology of this semi-aquatic insect can impact the set of digestive enzymes.

In *Gerris buenoi,* we found 28 genes and gene fragments that encode cysteine cathepsins of the C1 family. These enzymes primarily belong to the cathepsin L-like subfamily [53], while the cathepsin B-like subfamily was represented by only three potentially active enzymes and one putatively catalytically inactive TINAL-like protein [54]. Members of the cathepsin L-like family included two types of peptidase genes: (i) those encoding conserved cathepsins, which include orthologs of mammalian cathepsin L and cathepsin F, and orthologs of cathepsin I and cathepsin Ll (26-29kD-proteinase) that are found in most insects (manuscript in preparation); (ii) 13 species-specific cathepsin L-like genes that do not have orthologs in other insects and are unique to *Gerris buenoi*, The cathepsin B-like family contained an ortholog of mammalian cathepsin B and two species-specific cathepsin B-like peptidase genes.

Conserved cathepsins of *Gerris buenoi* have a unique profile: there are eight cathepsin Ll genes, while in most species only one copy of the gene is found. Functional analysis of cathepsin Ll is premature, but previous studies suggested that those peptidases (26-29kD-proteinases) could play a role in immune defense system degrading foreign proteins [55] or participate in metamorphosis [48]. Species-specific cysteine peptidases include 15 different genes, 11 of which form two phylogenetic clades presumably derived from an original cathepsin L through the course of evolution, and localized as sequential clusters of 2 to 4 genes. Considering all Heteroptera species described thus far have digestive cysteine peptidases [50-52], we propose that they also may play a digestive role in *Gerris buenoi*. This hypothesis is supported by the fact that similar species-specific clades of cysteine peptidases in the more thoroughly studied coleopterans *Tribolium castaneum* [47, 48], *Tenebrio molitor* [47] and *Leptinotarsa decemlineata* [56] are linked to digestion of food.

Visual genes

Water striders have drawn exceptional interest by visual scientists due to their exceptional visual ecology and correspondingly specialized organization of the visual system. The prominent, over 900 ommatidia counting compound eyes of water striders are involved in prey localization, mating partner pursuit, and predator evasion [57-59]. Although water striders utilize vision for dispersal by flight, water strider vision is considered specifically adapted to maximally sensitive 2-dimensional perception, i.e. the horizontal horizon of their water surface environment. Main evidence for this is the lateral acute zone, which facilitates neural superposition vision [60, 61]. Similar to higher Diptera like Drosophila, each ommatidial input is optically insulated from neighboring ommatidia through apposition optics. The sensitivity of target neurons in the lamina, however, is heightened at the level or neural organization of photoreceptor axons in target locations of the optic neuropils defined as neural superposition [57]. A likely functional morphological corollary of this is the open organization of the rhabdom in water strider ommatidia: Most of the individual photoresponsive membrane compartments (rhabdomeres) of each of the 8 photoreceptors per ommatidium are physically separated from each other [62]. This trait is shared derived trait for Heteroptera in contrast to Auchenorrhyncha and Coleorrhyncha [63], which feature a closed rhabdom where all rhabdomeres are in contact with each other along the proximodistal axis of the ommatidium.

Further notable for water strider vision is the dimorphism of ventral and dorsal ommatidia at the level of inner photoreceptor organization [63]. In the both dorsal and lateral ommatidia, both of the two inner photoreceptors contribute rhabdomeres in a highly organized orientation related to the rhabdomeres of the outer photoreceptors. In ventral ommatidia, by contrast, only the inner photoreceptor R8 forms a rhabdomere while the inner photoreceptor R7 does not. Interestingly, the specific orientation of the ventral R8 rhabdomeres is variable across Gerromorpha species. The tandem position of the R7 and R8 rhabdomeres in dorsal ommatidia has been proposed to be shared derived for Gerromorpha [63].

Typical for aquatic insects [64], *Gerris* is also polarized light-sensitive [65]. Schneider and Langer [62] describe how the cellular structure of photoreceptors relates to different polarized light sensitivities in the dorsal and ventral eyes. Studying the spectral sensitivity of *Gerris* photoreceptors to polarized light [66] concluded that the peripheral photoreceptors are blue sensitive while the inner photoreceptors are green sensitive, consistent with a larger number of sampled blue vs green photoreceptors. On the other hand, Bartsch [67] recorded 37

photoreceptor cells, only 7 of which were blue sensitive while the rest were green sensitive. This study further revealed the existence of green and blue sensitive polarized light detecting subsystems in the lateral-equatorial and lateral-dorsal region of the eye. The green-sensitive subsystem has been proposed to mediate object detection while the function of the blue sensitive system has remained enigmatic.

Our genomic analysis of *Gerris buenoi* uncovered 8 opsin homologs. This included one member each of the 3 deeply conserved arthropod non-retinal opsin subfamilies (c-opsin, Arthropsin, and Rh7 opsin) and 5 retinal opsins (Figure 4A and Supplementary Figure 2). The latter sorted into one member of the UV-sensitive opsin subfamily and 4 tightly tandem clustered members of the long wavelength sensitive (LWS) opsin subfamily (Figure 4A). Surprisingly, both genomic and transcriptome search in *G. buenoi* and other water strider species failed to detect sequence evidence of homologs of the otherwise deeply conserved blue-sensitive opsin subfamily [68].

While the apparent lack of blue opsin in *Gerris buenoi* was unexpected given the presence of blue sensitive photoreceptors, it was consistent with the lack of blue opsin sequence evidence in available genomes and transcriptomes of other heteropteran species including *Halyomorpha halys*, *Oncopeltus fasciatus*, *Cimex lectularius*, *Rhodnius prolixus*. Blue opsin, however, is present in other hemipteran clades, including Cicadomorpha (*Nephotettix cincticeps*) and Sternorrhyncha (*Pachypsylla venusta*) (Figure 4B and Supplementary Figure 2). Taken together, these data lead to the conclusion that the blue-sensitive opsin subfamily was lost early in the last common ancestor of the Heteroptera (Figure 4B), raising the question, which compensatory events explain the presence of blue sensitive photoreceptors in water striders.

Studies in butterflies and beetles produced evidence of blue sensitivity shifts in both UV- and LWS-opsin homologs following gene duplication [69-71]. Given that the UV-opsin family is generally conserved throughout insects even in crepuscular species like kissing bugs and bed bugs (opsin gene tree), and that evidence of UV-sensitive photoreceptors has been reported for backswimmers [72], it seems most likely that one or more of the newly expanded *Gerris buenoi* LWS opsin genes represent blue-shifted paralogs.

In further support of this hypothesis, the 4 *Gerris buenoi* LWS opsin paralogs have accumulated substantial sequence divergence amounting to pairwise 40 to 80 amino acid differences despite their tight genomic linkage. Further, there are compelling similarities at the four amino acid sites that have been implicated in the green to blue sensitivity shifts of butterfly LWS opsins: Ile17Met, Ala64Ser, Asn70Ser, and Ser137Ala [69, 70]. Most intriguingly, the Gbue LWS opsin 4 paralog

matches all green-sensitive amino acid residue states at these tuning positions, thus favoring this paralog as green-sensitive (Figure 4C). This conclusion is further bolstered by the near perfect amino acid state matches of the physiologically well-characterized green-sensitive LWS opsins of Drosophila and honeybee (Figure 4C), suggesting functional relevance across insect orders.

Combined with the sequence evidence from green vs blue-shifted LWS-opsins of butterflies, the existence of the blue-shifted LWS opsins Rh1 and Rh2 in addition to the green-sensitive Rh6 LWS opsin in Drosophila makes it also possible to probe for comparative evidence for the existence of candidate blue-shifting amino acid states in the Gbue LWS opsin paralogs 1, 2 and 3 (Figure 4C). This approach identifies different sets of blue-shifting amino acid states in Gbue LWS opsin 1 and 3. Most compellingly, both paralogs possess a blue-shift correlated methionine at position 17 as opposed to the green-sensitivity correlated isoleucine of Gbue LWS opsin 2 and 4. The same holds true for the blue-shifted LWS opsins Rh1 vs the green-sensitive Rh6 LWS opsin of Drosophila. At position 64, the phylogenetic signal for serine as blue-shifted state vs alanine as green-sensitive state is particularly strong in butterflies [69, 70]. This correlation, however, is not consistently shared in the Drosophila and the honeybee. While the extremely blue-shifted Drosophila LWS opsin Rh2 does possess a serine at this site, so does the green-sensitive LWS opsin 1 of the honeybee. These conflicting signals prevent a straightforward interpretation of the otherwise intriguing occupation of this site by either alanine or serine in the 4 *Gerris buenoi* LWS opsin. Less ambiguity, however, applies position 70, where Gbue LWS opsin 3 stands out by sharing a serine residue with blue-shifted butterfly LWS opsins. While the green-sensitivity associated asparagine is highly conserved, even in both blue-shifted Drosophila LWS opsins, the corresponding rarity of the serine state amounts to compelling evidence in support of Gbue LWS opsin 3 as blue-shifted LWS opsin. The comparative signal at tuning position 137, finally, is less straightforward to interpret. This site is occupied by a serine in *Gerris buenoi* LWS opsins 1 and 4 vs glycine in *Gerris buenoi* LWS opsins 2 and 3. The comparison with both butterfly and honeybee suggests the serine state of *Gerris buenoi* LWS opsins 1 and 4 as green-sensitive, no identical reference point exists for the glycine state of *Gerris buenoi* LWS opsins 2 and 3. However, the only conservative difference to the alanine state in the blue shifted LWS opsins of butterflies and Drosophila (Rh2) can be valued as tentative evidence for blue-shifted states of *Gerris buenoi* LWS opsins 2 and 3.

Collectively, the comparative evidence identifies Gbue LWS opsin 3 as the candidate blue-shifted paralog with the highest confidence followed by Gbue LWS opsin 1 and 2. This conclusion is further backed by the fact that water striders lack ocelli, which implies that all four paralogs are

most likely expressed in photoreceptors of the compound eye. Overall, it thus seems most likely that the differential expression of the highly sequence-diverged Gbue LWS opsin paralogs accounts for the presence of both blue- and green-sensitive photoreceptors in water striders. Moreover, given that the outer blue photoreceptors have been specifically implicated in the detection of contrast differences in water striders [66], it is tempting to speculate that the deployment of blue-shifted LWS opsins represents another parallel to the fast-tracking visual system of higher Diptera. While these predictions await physiological verification in water striders, the genomic exploration of *Gerris buenoi* vision identifies water striders and Heteroptera as a whole as an exceptionally relevant group in the molecular study of adaptive visual system evolution for comparison to Lepidoptera, Hymenoptera, and the higher Diptera (Brachycera).

In addition to these five retinal opsins, three extra-retinal opsins were detected in the *Gerris* genome: The deeply conserved yet functionally still poorly understood Rh7 opsin subfamily [73, 74], Arthropsin [75-77], and c-opsin (Supplementary Figure 2 and Supplementary Table 5). Only partial sequences Arthropsin and c-opsin were detectable in the *Gerris buenoi* genome assembly. However, complete transcript sequences were found in the transcriptome of the closely related water strider species *Limnoporus dissortis* (Supplementary Figure 2).

Chemoreceptor gene families

The three chemoreceptor families addressed herein are the seven-transmembrane-domain Odorant and Gustatory Receptors that together comprise the insect chemoreceptor superfamily, and the unrelated three-transmembrane-domain Ionotropic Receptors [78, 79]. All three families have recently been fully documented from three other heteropterans with genome sequence used as comparators here, the kissing bug *Rhodnius prolixus* [80], the bedbug *Cimex lectularius* [81], and the milkweed bug *Oncopeltus fasciatus* [9]. More distant comparisons with other hemipteroid insects like the pea aphid *Acyrthosiphon pisum* [82] and the human body louse *Pediculus humanus* [17] are not included here as these chemoreceptors are mostly highly divergent from these four species, and comparisons including all five above species are available in Panfilio et al. [9].

The Odorant Receptors (ORs) is a large family, which, at least in several endopterygotes, have been shown to mediate most of insect olfaction (e.g. [79]). The OR family evolved within basal insects [83, 84] and consists of the single highly conserved Odorant receptor Co-receptor protein and a set of “specific” ORs, each of which is co-expressed with OrCo, generally one specific OR per olfactory sensory neuron type. The OR family in *Gerris* consists of at least 153 genes, two of which

are modelled as being alternatively spliced in a fashion found in many other insects, with two long first exons encoding most of the protein that are alternatively spliced into several short-shared exons encoding the C-terminus. Thirteen of these OR genes are pseudogenic in the genome assembly, so the total of seemingly intact ORs in this compilation is 146, however many are partial models and many gene fragments remain. Phylogenetic analysis along with the other three heteropterans reveals the usual high conservation of the single OrCo proteins (Supplementary Figure 4A). There are three possible simple orthologs of "specific" ORs across these four heteropterans, indicated with an asterisk in Supplementary Figure 4A, and two more with simple duplications in one or more species (two asterisks). Otherwise the relationships consist either of highly divergent genes, or large expansions or "blooms" of ORs within a particular heteropteran lineage. In the case of *Gerris* these include expansions of 4 (Or64-67), 8 (Or145-152), 9 (Or90-97a/b), 13 (Or72-84), 13 (Or98-110), 16 (Or111-125), 18 (Or44-61), and 44 proteins (Or1-43). Comparable expansions were previously described in *Rhodnius* and *Oncopeltus* and are clear in this analysis as well (Supplementary Figure 4A). In contrast, *Cimex* has almost no lineage-specific expansions, with OR clades consisting of only 1, 2, or 3 genes.

The Gustatory Receptors (GRs) is also a large family and consist of subfamilies and lineages that predate even the origins of the OR family [78, 84-86]. The most prominent of these are the sugar, carbon dioxide, and fructose receptor subfamilies (Supplementary Figure 4B). The sugar receptors, represented here by Gr1/2 from *Apis mellifera*, were lost from the obligate blood feeders *Cimex* and *Rhodnius*, but are present as three genes each in *Oncopeltus* and this more general predator (Gr7-9). The carbon dioxide receptor subfamily, represented here by the Gr21a/62a dimer in *D. melanogaster* and Gr1-3 in *Tribolium castaneum*, was lost from most Hymenoptera as well as *Rhodnius*, but multiple related GRs are present in *Cimex*, *Oncopeltus*, and *Gerris* (Gr1-6). It remains to be shown whether these more distant relatives of the carbon dioxide receptors of endopterygotes are involved in perception of this molecule in heteropterans. The fructose receptor implicated also in brain nutrient sensing [87] has a single representative in each heteropteran, although the *Gerris* gene is represented only by a fragment in the current genome assembly (Gr10). This is the only GR lineage that is a simple ortholog across these four heteropterans. The remaining GRs present a pattern similar to that of most of the ORs, that is, a few highly divergent lineages, and several highly expanded lineages. In these GRs, however, these expansions mostly involve large alternatively-spliced loci, comparable to those found in many other insects from *D. melanogaster* [86] to *Calopteryx splendens* [83]. These loci consist of several

long first exons encoding most of the receptor (transmembrane domains 1-6) that are modelled as being alternatively spliced into three short shared exons encoding the intracellular loop 3 and TM7. The three largest of these loci, Gr35, 48, and 32 encode 11, 11, and 13 different and sometimes quite divergent receptors, respectively (Supplementary Figure 4B). The largest of these GR expansions consists of 80 proteins encoded by 27 genes (Gr22-48), while three smaller expansions of 10, 12, and 14 proteins also involve alternatively-spliced loci (Gr45-47, 55-60, and 15-19, respectively). This pattern of expansion of the "bitter" GRs in alternatively-spliced loci is shared with *Oncopeltus* where it has resulted in an even larger repertoire of "bitter" GRs, but barely at all in *Rhodnius* and *Cimex* both of which have comparatively small "bitter" GR subfamilies, presumably reflecting the different chemical ecologies of these four heteropterans.

The Ionotropic Receptors (IRs) is a variant family of the large and ancient superfamily of ionotropic glutamate receptors [78, 88]. The family contains two highly conserved co-receptors that are very similar to the ionotropic glutamate receptors in sequence and structure, Ir8a and 25a (Supplementary Figure 4C), as well as another widely expressed gene that might also encode a co-receptor, Ir76b, specifically involved in perception of amino acids [89, 90]. These heteropterans have four more single-copy IRs (21a, 40a, 68a, and 93a), most of which are implicated in perception of a variety of stimuli from temperature to humidity [91, 92]. All of these are present as single-copy clear orthologs of the named *Drosophila* genes, and indeed most are older gene lineages than heteropterans [83]. An unusual exception is that there is a divergent duplicate of Ir8a (Ir8a2L) immediately upstream of and in tandem with Ir8a. This gene is missing the first 1/3 of the equivalent length of Ir8a, and there is no RNAseq support for it, unlike Ir8a and 25a, so it might not be functional. As is commonly the case in other insects, there is a small expansion to four genes of the lineage related to the Ir41a/76a/92a lineage in *D. melanogaster*, which for consistency with other genomes are named in an Ir41 series (Ir41d is not shown in Supplementary Figure 4C because it is a partial model that does not align well). In *Drosophila* Ir41a and 92a have been implicated in detection of amines [93, 94]. A far larger expansion of 24 genes is related to the Ir75a-d/64a/84a lineage in *D. melanogaster*, and again this lineage is also expanded in many other insects, although seldom to this extent. Ir75a/b, 64a, and 84a in *Drosophila* flies have been shown to be involved in perception of several acids [95-99]. Like the other heteropterans and many other insects, there are several highly divergent IRs, falling into two groups with no simple relationships to *D. melanogaster* IRs. These were therefore named in a series from Ir101 to avoid confusion with *D. melanogaster* Ir genes, whose names only go to Ir100a because like the Or and Gr genes they

were named for their cytological location in the polytene chromosomes. Ir101-105 are weakly related to a large expansion of so-called "divergent" IRs in *Drosophila*, including the Ir20a clade that function as gustatory receptors [100, 101]. Ir106-109 form a small clade related only to some other divergent heteropteran IRs, and are perhaps also involved in gustation. Thus, while not nearly as large as the OR and GR families, these IRs probably contribute some well-conserved functions shared with their orthologs with *Drosophila*, as well as perception of amines and diverse acids, and contribute to gustation. The only lineage-specific expansion compared with the other heteropterans is the IR75 clade implicated in perception of various acids, but it is unclear how this relates to the chemical ecology of water striders.

Wing development and polyphenism

The ability to produce different phenotypes from a single genome in response to environmental cues is called 'polyphenism' [102]. Water striders express a seasonal wing polyphenism (Figure 1), where adults are short-winged in the early summer generation when habitats are stable, but are long-winged in the mid-summer generation when habitats become unstable [103, 104]. It is thought that this wing polyphenism reflects an adaptive tradeoff between wing length and reproduction, where in unstable habitats populations invest in long wings and produce fewer offspring, but in stable habitats populations produce short wings and invest in more offspring [103, 104]. The environmental cues that may affect wing morphology include photoperiod, temperature, resource availability, and population density [104-106].

Wing polyphenism and adaptive tradeoffs between flight and reproduction are ecologically important and phylogenetically widespread among insects. In wing polyphenic ants and aphids, for example, previous studies used bioinformatics approaches to infer that the genes involved in the development of wings and the ovaries have a different DNA methylation signature relative to the rest of the genome [107-111]. This suggests that these genes are regulated by epigenetic mechanisms [107-111]. Therefore, in the water strider *Gerris buenoi*, we predicted that genes involved in wing patterning and reproduction will also have a different DNA methylation signature relative to the rest of the genomes. Furthermore, previous studies have shown that juvenile hormone (JH) and insulin signaling pathways are associated with regulation of reproduction and wing polyphenism in insects [102, 112-114]. We therefore analyzed epigenetic signatures in genes involved in both of these pathways relative to the rest of the genome. Finally, we compared genes from *Gerris buenoi* to orthologues in *Rhodnius proxilus* because this closely related species serves

as a phylogenetically controlled outgroup, which has not evolved wing polyphenism.

We discovered that the mean $CpG_{O/E}$ values for *Gerris buenoi* genes in the network related to wing polyphenism, juvenile hormone, insulin signalling and reproduction are not significantly different from the mean of the resampled distribution of $CpG_{O/E}$ of all *Gerris buenoi* genes (Supplementary Figure 8 and Supplementary Table 14 : List of genes in the networks underlying wing polyphenism, reproduction, juvenile hormone, and insulin signalling included in the analysis and their $CpG_{O/E}$ value for *Gerris buenoi* and *Rhodnius prolixus*. Genes that were annotated in *Gerris buenoi* but excluded from the analysis because they did have a complete codding sequence are also listed but without a $CpG_{O/E}$ value.). The mean $CpG_{O/E}$ of the *R. proxilus* orthologues related to wing polyphenism, juvenile hormone regulation, insulin signalling and reproduction is also not significantly different from the mean of the resampled distribution of $CpG_{O/E}$ of all *Rhodnius proxilus* (Supplementary Figure 8). These results indicate that genes in the network related to wing polyphenism, juvenile hormone, insulin signalling and reproduction do not have a distinct methylation signature relative to the rest of genes in *Gerris buenoi* and *Rhodnius proxilus* genomes.

The sequencing of three ant genomes, each of which possess a dramatic wing and reproductive polyphenism, showed significant methylation signature of genes known to be involved in wing and reproductive development relative to the rest of the genes in the ant genomes [107-109]. We therefore expected that genes involved in wing and reproductive development in the wing polyphenic water strider *Gerris buenoi* would possess a similar methylation signature as in the ants. To our surprise, the results of our analysis reveal that methylation signatures in genes involved in wing and reproductive development are not significant relative to the rest of the genome. This is also the case for the closely-related and non-wing polyphenic insect *Rhodnius proxilus*. These findings suggest that more classical mechanisms for achieving differential gene expression underlying polyphenism, such as endocrine-based mechanisms like hormone secretion and neuropeptide release, are involved in regulating the expression of genes underlying wing polyphenism as well as the trade-off between wing development and reproduction in water striders [115]. Altogether, these results open up exciting future research possibilities for understanding how wing polyphenism is regulated in water striders, and why they appear to differ from other polyphenic insects.

DNA methylatransferases

DNA methylation is an epigenetic mechanism known to be involved in the regulation of alternative splicing and gene expression in insects [116-118]. In honeybees, it has been demonstrated that the DNA methyltransferase, DNMT3, is critical in sizing, morphology and reproductive organ development associated with caste determination as well as alternative splicing regulation [117-119]. Furthermore, differential DNA methylation is associated with flexible behavioral castes (nurses and foragers) in bees [120]. Therefore, this epigenetic mechanism is considered to be a potentially key regulator of morphological development and behavioral differentiation in insects. Paradoxically, many insects have lost key elements of the DNA methylation toolkit, including DNMT1 and DNMT3, as is the case for *Drosophila melanogaster* [121]. In order to see if this pathway may be worth further investigation for the study of morphological development in water striders, we searched for several core elements that regulate this molecular process. Although we found that the water strider genome does possess *DNMT1*, which is essential for the maintenance of DNA methylation, and *DNMT2*, the protein of which functions to methylate tRNAs, the *Gerris buenoi* genome does not contain an ortholog of *DNMT3*, which is essential for de novo DNA methylation. It is hard to predict the significance of *Gerris buenoi* lacking *DNMT3* because the presence versus absence of this gene is quite erratic across insects [122]. Although it may be associated with the capacity for elaborate environmentally-dependent developing processes, including those that are polyphenic as it is found in a range of invertebrates including the pea aphid [123], Daphnia [124], termites [125] and various hymenoptera including bees and ants that are highly plastic [107, 126, 127]. Still, there are other highly conserved epigenetic processes, such as histone modifications, which are conserved in *Gerris buenoi*, and may serve as alternative mechanisms for the regulation of developmental plasticity.

Histone genes and histone modification machinery

Chromatin remodelling, via post-translational modifications of histones, is a key regulator of gene expression. These epigenetic processes have been associated with environmental responsiveness and phenotypic plasticity [128]. One of the most striking cases of plasticity in the Gerridae is associated with wing development [129]. Most species of this family exhibit winged and wingless morphs known as apterous and macropterous morphs [129, 130]. Wing development is influenced by both genetic and environmental factors such as habitat stability, day/night cycle and latitude [103, 104, 131]. Other cases of phenotypic plasticity include leg length, pigmentation, and a set of

555 secondary sexual traits in both males and females [132]. While our understanding of the ecology
556 of these cases of phenotypic plasticity is increasingly richer, the lack of a water strider genome has
557 hindered studies of the genetic and developmental factors associated with them. We therefore
558 analysed the *Gerris buenoi* genome content in search for components of the epigenetic
559 machinery.
560 In the *Gerris buenoi* genome we could identify 49 histone proteins encoding loci, a moderately
561 large number of genes similar to that found in *Cimex lectularius* and *Daphnia pulex*, but
562 substantially smaller than that detected in the *Aedes aegypti* or *Drosophila* genomes
563 (Supplementary Table 15). We identified genes encoding the five major classes of histone proteins
564 (H2A, H2B, H3, H4 and the linker histone H1) as well as copies of genes encoding the variant
565 histones H2AV and H3.3. In *Drosophila* the histone genes are present in the genome in large
566 numbers of quintet clusters, each cluster having one gene from each of the five classes of
567 histones. A similar organization was found in the *Gerris buenoi* genome where two canonical
568 quintet clusters were identified. Both of them consists of one copy of each of the four classes of
569 core histone proteins (H2A, H2B, H3 and H4) and a single copy of the linker histone (H1)
570 (Supplementary Figure 9). Additional clusters were identified, including one modified cluster
571 containing two copies of the linker histone (H1) and two copies of the H2B core histone, but no
572 copy of the core histone H3, as well as five truncated clusters made of three or four genes
573 including H3 core histone gene and combinations of the other histone genes (Supplementary
574 Figure 9). The number of these clusters is higher compared to the genomes of the milkweed bug
575 *Oncopeltus fasciatus* and the bed bug *Cimex lectularius,* which contain one and two clusters
576 respectively [9, 81]. The functional significance of these clusters remains unknown, thus opening
577 new avenues in the study of the relationship between epigenetics and phenotypic plasticity [133].
578 Histone proteins can be post-translationally modified to dynamically influence the structure of the
579 chromatin. We found in the *Gerris buenoi* genome genes responsible for all classes of histone
580 modifications: histone acetyltransferases, deacetylases, methylases and demethylases.
581 Interestingly, we found a duplication of the histone acetyltransferases *males absent on the first*
582 (*mof*) and *chameau* (*chm*/HAT1). *Mof* functions in dosage compensation and genome stability in
583 *Drosophila* [134, 135]. Duplications of *mof* and *chm* have previously been reported for
584 *Acyrthosiphon pisum* and were thought to be unique [136] although *mof* duplication was also
585 recently detected in *Oncopeltus fasciatus* [9] and *Cimex lectularius* [81]. Phylogenetic analysis
586 indicates the duplications that have occurred in these species are independent of the duplication

that occurred in *Acyrthosiphon pisum* and likely occurred early in the heteropteran lineage ~250 million years ago (Supplementary Figure 10). Unusually, we also identified a duplication of the *Gerris buenoi* histone deacetylase *Sirt1* (*sir2*) and *Sirt5*; and the histone methyltransferase *grappa*. Sirt1 is a nuclear and cytoplasmic deacetylase that has a role in histone modifications [137] and has been associated with enhanced stress response and life-span extension in numerous species [136, 138, 139]. Grappa, histone methyltransferase, modifies the lysine (K)79 residue of histone H3 and has been implicated in the stress response in *Drosophila* providing protection against oxidative and caloric stress [140]. Interestingly, duplications of *Grappa* have not been detected in any other hemipteran species.

In conclusion, the high number of histone clusters found as well as the duplication of some post-translational modifications of histones genes open up exciting future research possibilities for understanding their role in environmental responsiveness and phenotypic plasticity in *Gerris buenoi*.

Antioxidant Proteins

Reactive oxygen species (ROS), including superoxide radicals (O_2^-), hydroxyl radicals (OH^-), and hydroperoxides (H_2O_2, and ROOH), are generated by aerobic metabolism but may also be encountered in an organism diet or environment [141-143]. Moderate levels of ROS drive a variety of processes including cellular signaling, transcriptional regulation, as well many other physiological processes. However, inability to regulate ROS concentrations can result in the accumulation of ROS-induced damaged lipids, proteins, and nucleic acids [141-143]. Animals have evolved a complex system of antioxidant enzymes and molecules, facilitating the modulation of ROS levels [142, 144-146]. The enzymatic antioxidant system is comprised of a diverse suite of proteins that can be divided into clades based on their modes of action. Catalase (CAT), superoxide dismutase (SOD), and a variety of peroxidases make up the core of the antioxidant response. Thioredoxins and methionine sulphoxide reductases form a secondary system for managing ROS [144, 145].

Thirty putative proteins in seven families related to antioxidant capacity were identified within the *G. buenoi* genome. The thirty antioxidant response proteins showed high homology to related proteins in other published genomes including *Acyrthosiphon pisum, Apis mellifera, Bombyx mori, Cimex lectularis, Drosophila melanogaster, Pediculus humanus, and Tribolium castaneum* (see Supplementary Methods). In most comparisons, homologs in *C. lectularis* genome showed the

highest degree of similarity (Supplementary Table 16). Representatives of all major antioxidant enzyme clades were identified in the *G. buenoi* genome assembly including a *Catalase*-like gene, four heme-binding peroxidases, multiple glutathione-s-transferases, peroxidase, multiple peroxiredoxins, and superoxide dismutases. This representation suggests that the *G. buenoi* genome contains a complete suite of antioxidant enzymes. There is no apparent expansion or reduction in the gene families that were surveyed in this analysis, however further investigation through additional annotation and experimental validation may reveal otherwise.

Supplementary Methods

Genome sequencing and assembly

Gerris buenoi is one of thirty arthropod species sequenced as a part of a pilot project for the i5K arthropod genomes project at the Baylor College of Medicine Human Genome Sequencing Center. For all of these species, an enhanced Illumina-ALLPATHS-LG sequencing and assembly strategy enabled multiple species to be approached in parallel at reduced costs. For most species, including *Gerris buenoi*, we sequenced four libraries of nominal insert sizes 180bp, 500bp, 3kb and 8kb. The amount of sequence generated from each of these libraries is noted in Supplementary Table 17 with NCBI SRA accessions. The 180bp, 500bp and 3kb mate pair libraries were made from a single male individual, and the 8kb mate pair library from female genomic DNA.

To prepare the 180bp and 500bp 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 3kb or 8kb 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),

650 then end repaired and biotinylated. Fragment sizes between 3-3.7 kb (3kb) or 8-10 kb (8kb) were
651 purified from 1% low melting agarose gel and then circularized by blunt-end ligation. These size
652 selected circular DNA fragments were then sheared to 400-bp (Covaris S-2), purified using
653 Dynabeads M-280 Streptavidin Magnetic Beads, end-repaired, dA-tailed, and ligated to Illumina PE
654 sequencing adapters. DNA fragments with adapter molecules on both ends were amplified for 12
655 to 15 cycles with Illumina P1 and Index primers. Amplified DNA fragments were purified with
656 Agencourt AMPure XP beads. Quantification and size distribution of the final library was
657 determined before sequencing as described above.
658 Sequencing was performed on Illumina HiSeq2000s generating 100bp paired end reads. Reads
659 were assembled using ALLPATHS-LG (v35218) [147] on a large memory computer with 1Tbyte of
660 RAM and further scaffolded and gap-filled using in-house tools Atlas-Link (v.1.0) and Atlas gap-fill
661 (v.2.2) (https://www.hgsc.bcm.edu/software/). This yielded an assembly of 1,000.16 Mb (653 Mb
662 without gaps within scaffolds) with a contig N50 of 3.8 kb and scaffold N50 of 344kb which has
663 been deposited in the NCBI: GenBank assembly accession GCA_001010745.1
664

665 **Automated Gene Annotation Using a Maker 2.0 Pipeline Tuned for Arthropods**
666 Of 30 attempted i5K pilot species, 28 i5K pilot genome assemblies including *G. buenoi* were
667 subjected to automatic gene annotation using a Maker 2.0 annotation pipeline tuned specifically
668 for arthropods. The pipeline is designed to be systematic providing a single consistent procedure
669 for the species in the pilot study, scalable to handle 100's of genome assemblies, evidence guided
670 using both protein and RNA-seq evidence to guide gen models, and targeted to utilize extant
671 information on arthropod gene sets. The core of the pipeline was a Maker 2 [148] instance,
672 modified slightly to enable efficient running on our computational resources. The genome
673 assembly was first subjected to de-novo repeat prediction and CEGMA analysis to generate gene
674 models for initial training of the ab-initio gene predictors. Three rounds of training of the Augustus
675 [149] and SNAP [150] gene predictors within Maker were used to bootstrap to a high quality
676 training set. Input protein data included 1 million peptides from a non-redundant reduction (90%
677 identity) of Uniprot Ecdysozoa (1.25 million peptides) supplemented with proteomes from
678 eighteen additional species (*Strigamia maritima, Tetranychus urticae, Caenorhabditis elegans, Loa*
679 *loa, Trichoplax adhaerens, Amphimedon queenslandica, Strongylocentrotus purpuratus,*
680 *Nematostella vectensis, Branchiostoma floridae, Ciona intestinalis, Ciona savignyi, Homo sapiens,*
681 *Mus musculus, Capitella teleta, Helobdella robusta, Crassostrea gigas, Lottia gigantea,*

682 *Schistosoma mansoni*) leading to a final 'nr' peptide evidence set of 1.03 million peptides. RNA-seq
683 transcription data derived from mixed sex embryo's and nymphs (Supplementary Table 17) was used
684 judiciously to identify exon-intron boundaries but with a heuristic script to identify and split
685 erroneously joined gene models. We used CEGMA models for QC purposes: for *Gerris buenoi*, of
686 1,977 CEGMA single copy ortholog gene models, 1,783 were found in the assembly and 1,895 in
687 the final predicted gene set – a reasonable result given the small contig sizes of the assembly. We
688 assume the gene predictors could pull together exons from different contigs with greater success
689 than the sequence comparison used to identify CEGMA genes in the assembly, generating the
690 larger number of control gene models found in the gene set than the underlying assembly. Finally,
691 the pipeline uses a nine-way homology prediction with human, *Drosophila* and *Caenorhabditis*
692 *elegans*, and InterPro Scan5 to allocate gene names. The automated gene sets are available from
693 the National Agricultural Library (https://i5k.nal.usda.gov/Gerris_buenoi) where a web-browser of
694 the genome, annotations, and supporting annotation data is accessible.
695
696
697 **Bristle genes**
698 Bristle development genes were annotated by performing tblastn searches on the *Gerris buenoi*
699 scaffolds with the corresponding Drosophila gene protein sequences available in FlyBase (release
700 6)[151]. To confirm orthology, *Gerris buenoi* models were blasted into NCBI 'nr' database.
701 Homology, intron/exon boundary assessments, and protein sequence completeness were
702 identified by manual inspection using RNA-seq alignments available and protein alignments
703 generated with Clustal Omega (http://www.ebi.ac.uk/Tools/msa/clustalo/).
704
705 **Cuticular proteins**
706 Sequence motifs that are characteristic of several families of cuticle proteins [152] were used to
707 search the genome of *Gerris buenoi* for putative cuticle proteins. 155 genes were identified,
708 analyzed with CutProtFam-Pred, a cuticular protein family prediction tool described in Ioannidou
709 et al. [153], and assigned to one of 5 families (CPR, CPAP1, CPAP3, CPF, and TWDL).
710
711 **Prey detection and selection on water environments**
712 The approach for manual annotation is similar to that used to characterize these three gene

families in many other insects, including *Acyrthosiphon pisum* [82], *Pediculus humanus* [17], *Rhodnius prolixus* [80], *Cimex lectularius* [81] and *Oncopeltus fasciatus* [9]. Briefly, exhaustive and iterative tblastn searches of the genome assembly with the proteins from these other heteropterans were used to find genes, which were modelled as best possible in the WebApollo browser at the i5k site. This effort was sometimes assisted by RNA-seq reads that cross introns in the available whole-body RNA-seq set, however most of these genes were not represented in that dataset. In addition, like *Oncopeltus fasciatus* this genome assembly is rather fragmented, so many of the models are incomplete, while some were joined across scaffolds and a few were improved with raw reads. Several additional gene fragments too short to include in this compilation remain for the OR and GR families and might represent additional intact genes, while some of the partial models might actually be pseudogenes. Many of these proteins are extremely divergent, and because almost none of them were modelled by the genome-wide automated annotation (models that might have facilitated searches for distant relatives using BLASTP), TBLASTN searches to find distant relatives used E values of 1000. The last two exons of the OR and GR families typically encode the most conserved regions of these proteins and are flanked by phase 0 introns, so their encoded protein sequences were used in TBLASTN searches with LQ before and VS afterwards, representing consensus splice acceptor and donor sites, to assist in finding divergent relatives. Multiple alignments of each family along with representatives from other species and maximum likelihood phylogenetic analyses of the proteins were conducted, and the tree figures prepared, as in Panfilio et al. [9]. All of the proteins are included at the end of this supplementary text, and the gene models and transcribed mRNAs for most of them are available from the i5k Workspace at the National Agriculture Library (https://i5k.nal.usda.gov/) and will eventually be available from the NCBI.

Wing polyphenism

First, we limited our analysis to genes whose complete coding sequences had been identified and annotated in the following four categories: genes involved in wing polyphenism, juvenile hormone regulation, the insulin signalling pathway, and reproduction. We then used the bioinformatics-based metric described by Elango *et al.* [111] called $CpG_{O/E}$ as a proxy for mutations induced by methylation of CpG islands in the germ line over evolutionary time. This $CpG_{O/E}$ metric uses a historical (evolutionary) measure of the level of DNA methylation by estimating the amount of CpG dinucleotide depletion normalized for GC content for each gene of interest. The $CpG_{O/E}$

metric, or CpG dinucleotide depletion normalized for GC, is a proxy for DNA methylation in the coding sequence of these genes. We define the $CpG_{O/E}$ for each gene as follows:

$$C_P G_{O/E} = \frac{P_{CpG}}{P_c P_G}$$

where $CpG_{O/E}$ is an estimation of the DNA methylation levels, P_{CpG} is the frequency of CG dinucleotides, P_C is the frequency of cytosine nucleotides, and P_G is the frequency of guanine nucleotides [154, 155]. After cytosine is methylated, it is more amenable to deamination [155]. Over time, this leads to the reduction of CpG dinucleotides from methylated CpG regions [155]. Using a custom Perl script, we evaluated the $CpG_{O/E}$ in the coding sequences of all predicted genes in the *Gerris buenoi* genome and the $CpG_{O/E}$ in the coding sequences of our genes of interest (Supplementary Table 14).

Second, we compared the mean $CpG_{O/E}$ content for our genes of interest to the mean $CpG_{O/E}$ for all the genes in the genome by executing a Monte-Carlo randomization procedure as described previously [107-111]. Briefly, we randomly selected 50 $CpG_{O/E}$ values from the genome to produce a random distribution, calculated the mean, and repeated this process 10000 times. All mean $CpG_{O/E}$ values were plotted and this distribution was compared to the mean $CpG_{O/E}$ values for each of our candidate gene sets. Gene sets were determined to be significantly different from the randomly generated mean $CpG_{O/E}$ if they fell within the bottom or top 5% of values. These analyses were repeated for *Rhodnius proxilus* orthologues of the *Gerris buenoi* genes in our gene sets.

Wnt Signaling Pathway

Protein sequences for *Wnt* ligands as well as receptors and downstream components (*armadillo*/*beta-catenin*, *dishevelled*, *frizzled*, *arrow*, *axin*, *shaggy*/ *GSK*-3) from *Drosophila melanogaster*, *Tribolium castaneum*, *Acyrthosiphon pisum* and *Oncopeltus fasciatus,* were retrieved from NCBI, and used to perform standalone tblastn searches on the *Gerris buenoi* scaffolds with a 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 from multiple orthologous queries, the putative gene name was identified by blasting back the hit sequence against GenBank, with a taxonomic restriction to Arthropoda accessions. The query sequences with the best hits (lowest e-values) for each gene were then used to identify the model to be curated, by doing a tblastn search into the *Gerris* scaffolds from the Blast instance at the National

Agricultural Library (https://i5k.nal.usda.gov/legacy_blast). The Blast results were visualized in the Web Apollo instance for *Gerris buenoi* (https://apollo.nal.usda.gov/gerbue/selectTrack.jsp), where the corresponding automated annotation models were edited. To confirm orthology, we then Blasted the edited *Gerris buenoi* models back into GenBank. 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/).

The numbering (subfamily identification) for *Wnt and fz* orthologs was assigned based on the corresponding vertebrate homolog (the naming of *Drosophila* orthologs was changed accordingly), based on phylogenetic analyses done at http://www.phylogeny.fr/.

Possible gene loci duplications were identified by performing tblastn searches on the scaffolds using the protein sequences of completed *Gerris* annotation models as queries, and then re-blasting the resulting hit sequences into GenBank for Arthropoda hits.

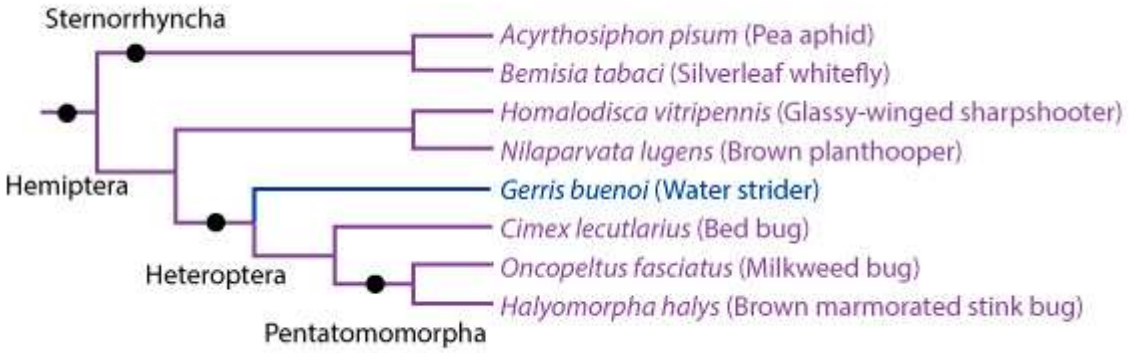
Early Developmental Genes

The choice of early developmental genes (Gap, Pair Rule, and Segment Polarity Genes) to annotate was informed by GO term annotations in *Drosophila melanogaster* (long-germ) and *Tribolium castaneum* (short-germ). Protein sequences for developmental genes for *D. melanogaster* and *T. castaneum* were obtained from http://flybase.org/ [151] and http://beetlebase.org/ [156] respectively. Contig sequences were searched for homology to the selected protein sequences using tbastn. Gene models (Gbue v0.5.3-models) that aligned with the regions of highest homology identified by tbastn search were selected for further analysis. If no official gene model was present in the region of homology identified by tblastn a de novo model was generated using models generated by the Augustus-masked or snap-masked programme. RNAseq mapped reads were compared with the gene models to determine the transcribed regions. The transcribed regions were used to determine protein sequences of the gene. Protein sequences were utilised in a reciprocal blast (blastx NCBI) to confirm the homology of the orthologs. Gene models were manually edited to produce gene models that resolved conflicts between RNAseq, blastx and homology data.

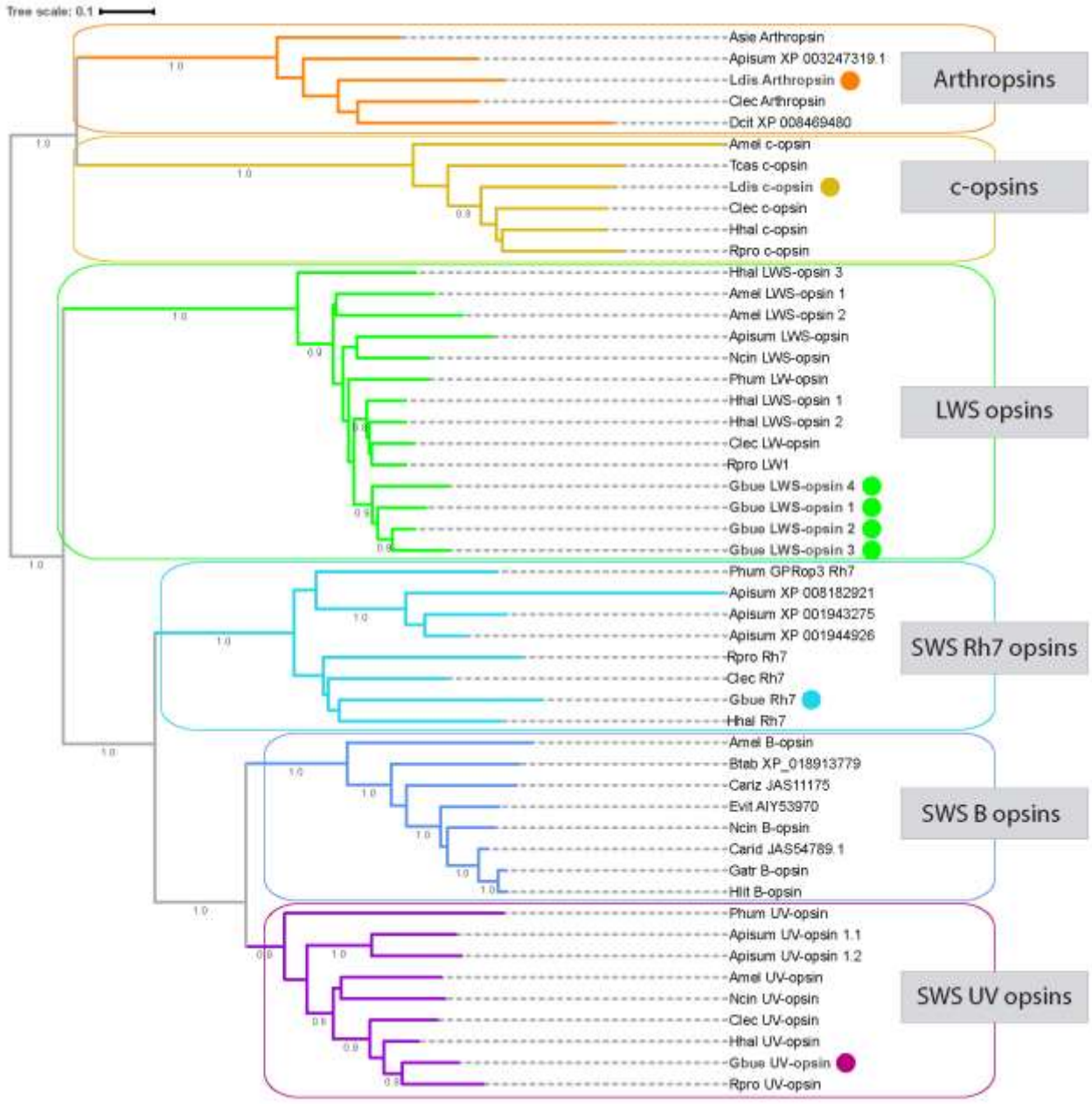
Antioxidant genes

Antioxidant proteins of *Drosophila melanogaster* were utilized to initially identify potential antioxidant genes within the *Gerris buenoi* genome. The *Drosophila melanogaster* genes were obtained from FlyBase by generating a query that searched for proteins with Gene Ontology terms that were related to response to antioxidant activity and responses. These nucleotide sequences were translated to peptides and were searched against the peptide models of *G. buenoi*. The highest BLAST hit (blastp) was extracted and searched against arthropod entries of the NCBI non-redundant database to confirm the identity of the model (blastp). The confirmed model was then BLAST searched (blastp) against the peptide sequences of *Acyrthosiphon pisum, Apis mellifera, Bombyx mori, Cimex lectularis, Drosophila melanogaster, Pediculus humanus, and Tribolium castaneum* to extract homologs. The extracted *G. buenoi* model was then aligned to the homologs. This information and the RNA-seq data present in the WebApollo were used to manually annotate the model. The corrected model was then once more searched against the arthropod entries of the NCBI non-redundant database (blastp) to ensure that the model was correctly identified.

Supplementary Figures and Tables



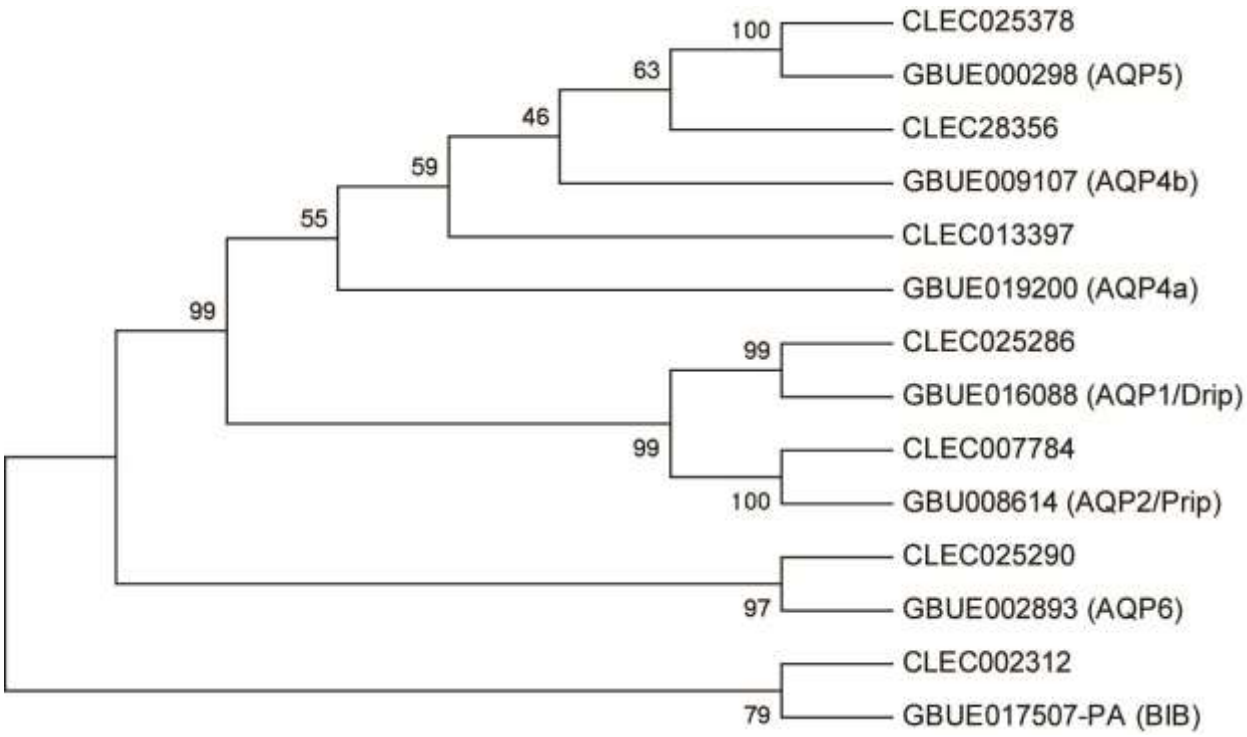
Supplementary Figure 1 : Detailed cladogram of Hemiptera species used in Figure 3. The tree is based on phylogenetic analyses in [157]. Both trees combined with the absence of third InR copy in *C. lectularius*, *O. fasciatus* and *H. halys*, suggest that InR1-like duplication is unique to the Gerromorpha and occurred at, or close to, their speciation.



831

Supplementary Figure 2 : Phylogenetic analysis and representative sequences from all major insect opsin subfamilies. Protein sequences were aligned with T-Coffee [158] and ambiguous multiple alignment alignment segments were removed applying the "gappyout" setting of TrimAl (v. 1.3) [159]. A neighbor joining tree was estimated in MEGA version 6.0 [160] using gamma-corrected Jones-Taylor-Thornton distances [161] and testing branch support with 1 000 bootstrap samples (numbers at branches). Species abbreviations: Amel = *Apis mellifera*, Apisum = *Acyrthosiphon pisum*, Asie = *Anotogaster sieboldii*, Btab = *Bemisia tabaci*, Cariz = *Clastoptera arizonana*, Clec = *Cimex lectularius*, Carid= *Cuerna arida*, Dcit=*Diaphorina* citris, Evit = *Empoasca vitis*, Gbue = *Gerris buenoi*, Gatr = *Graphocephala atropunctata*, Hhal = *Halyomorpha_halys*, Hlit = *Homalodisca liturata*, Ldis = *Limnoporus dissortis*, Ncin = *Nephotettix cincticeps*, Phum = *Pediculus humanus*, Rpro = *Rhodnius prolixus*, Tcas = *Tribolium castaneum*.

832

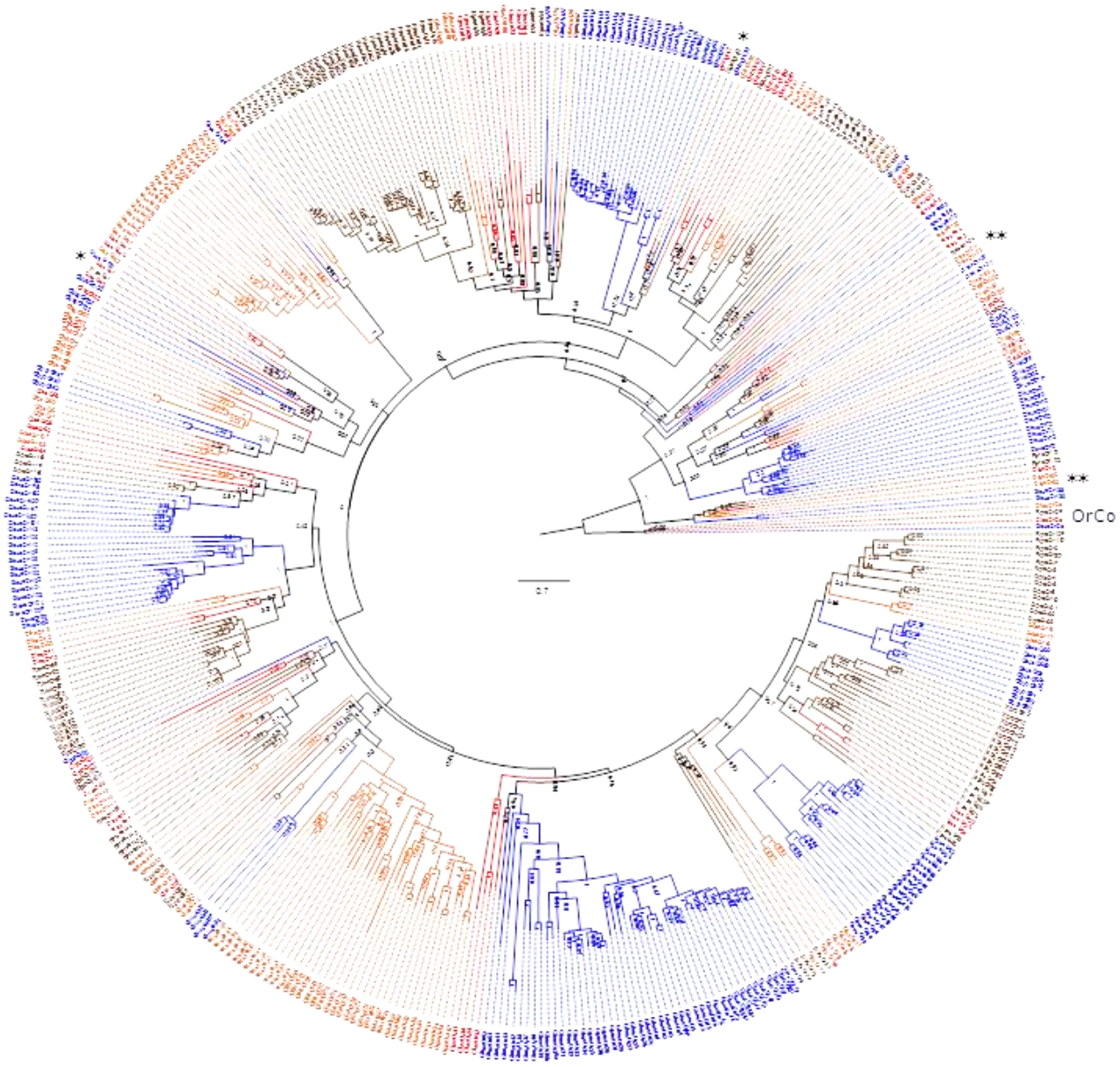


833

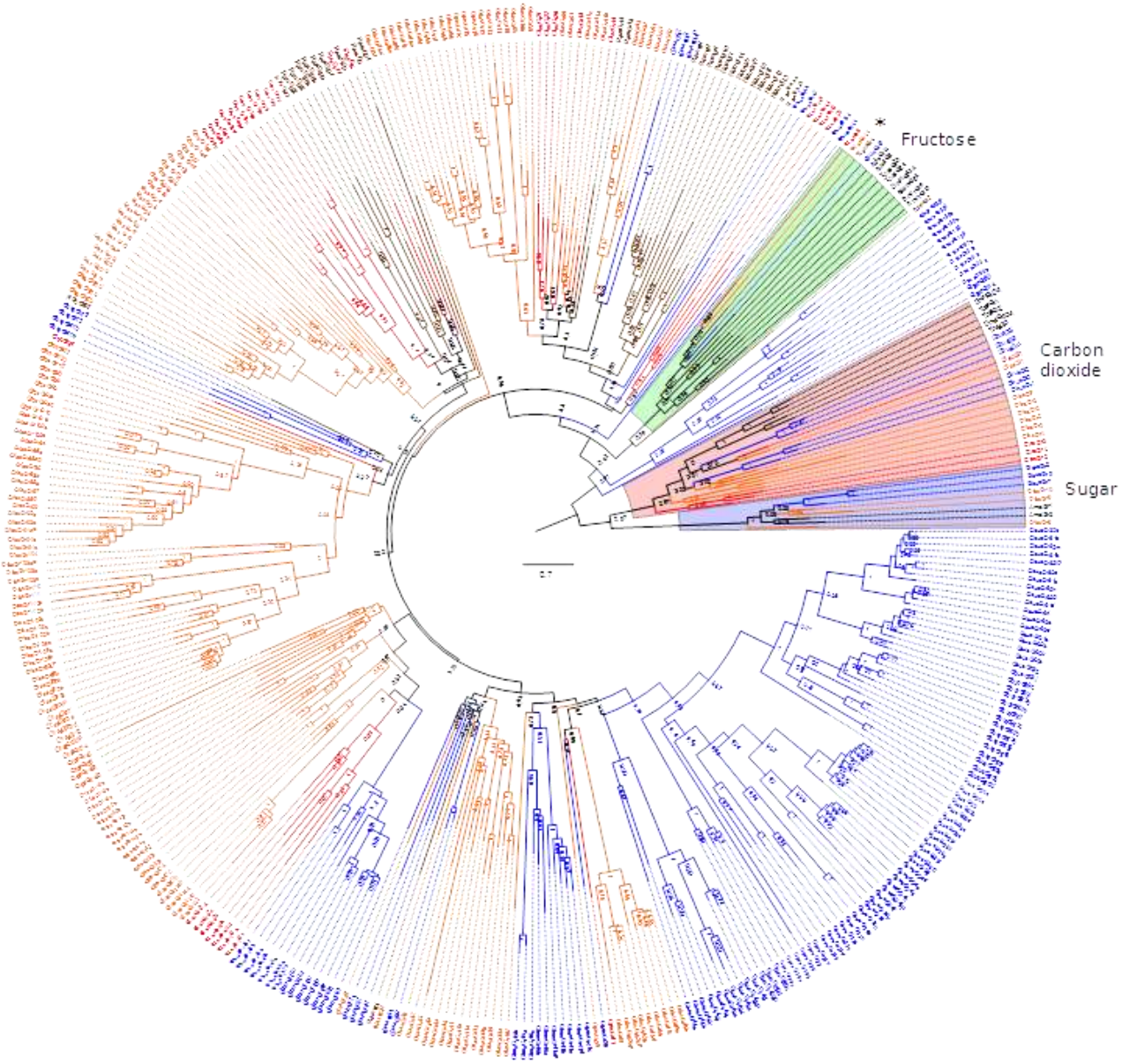
Supplementary Figure 3 : Comparison of predicted aquaporins from *Gerris* and *Cimex* using Neighbor-joining tree produced using MEGA6 using Dayhoff model and pairwise matching; branch values indicate support following 1500 bootstraps; values below 50% are omitted. It includes the seven putative aquaporin (AQP) genes identified from the water strider that includes the typical *Drosophila* integral protein (Drip), AQP2, AQP4 (Two genes), AQP5, AQP6 and Big brain (Bib) genes. In addition to these seven, we identified one other predicted partial sequences with matches to AQP sequences from other insects. Overall the number of aquaporins falls within the range of most insects (6-8) and *Gerris* has members of each group previously identified for insects [81].

834
835

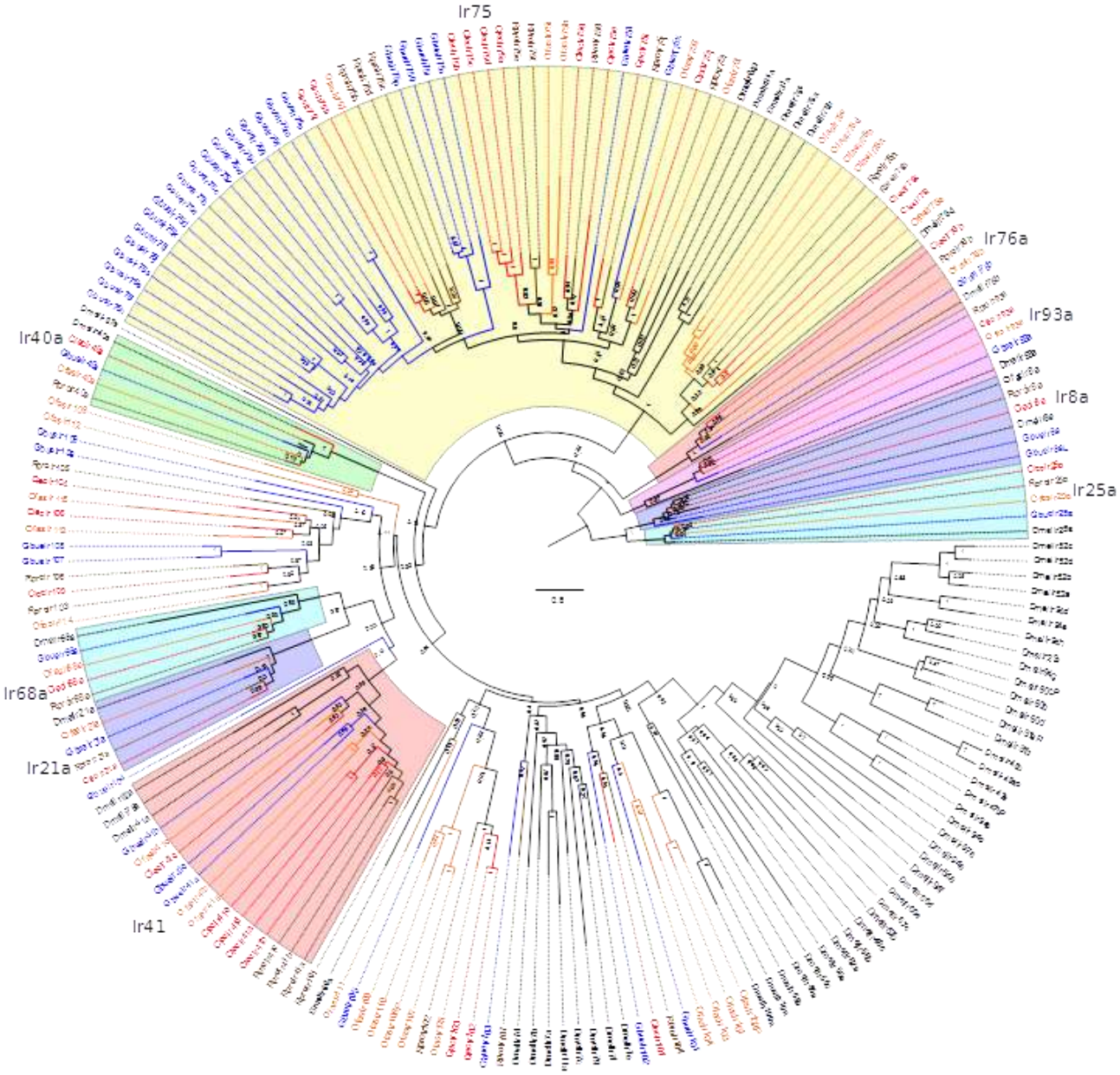
A.



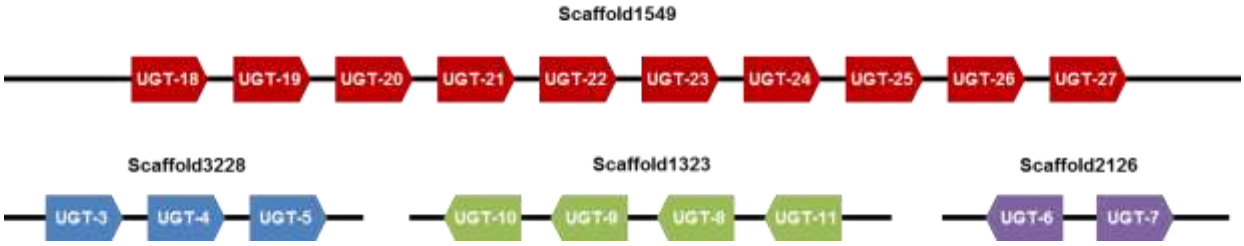
B.



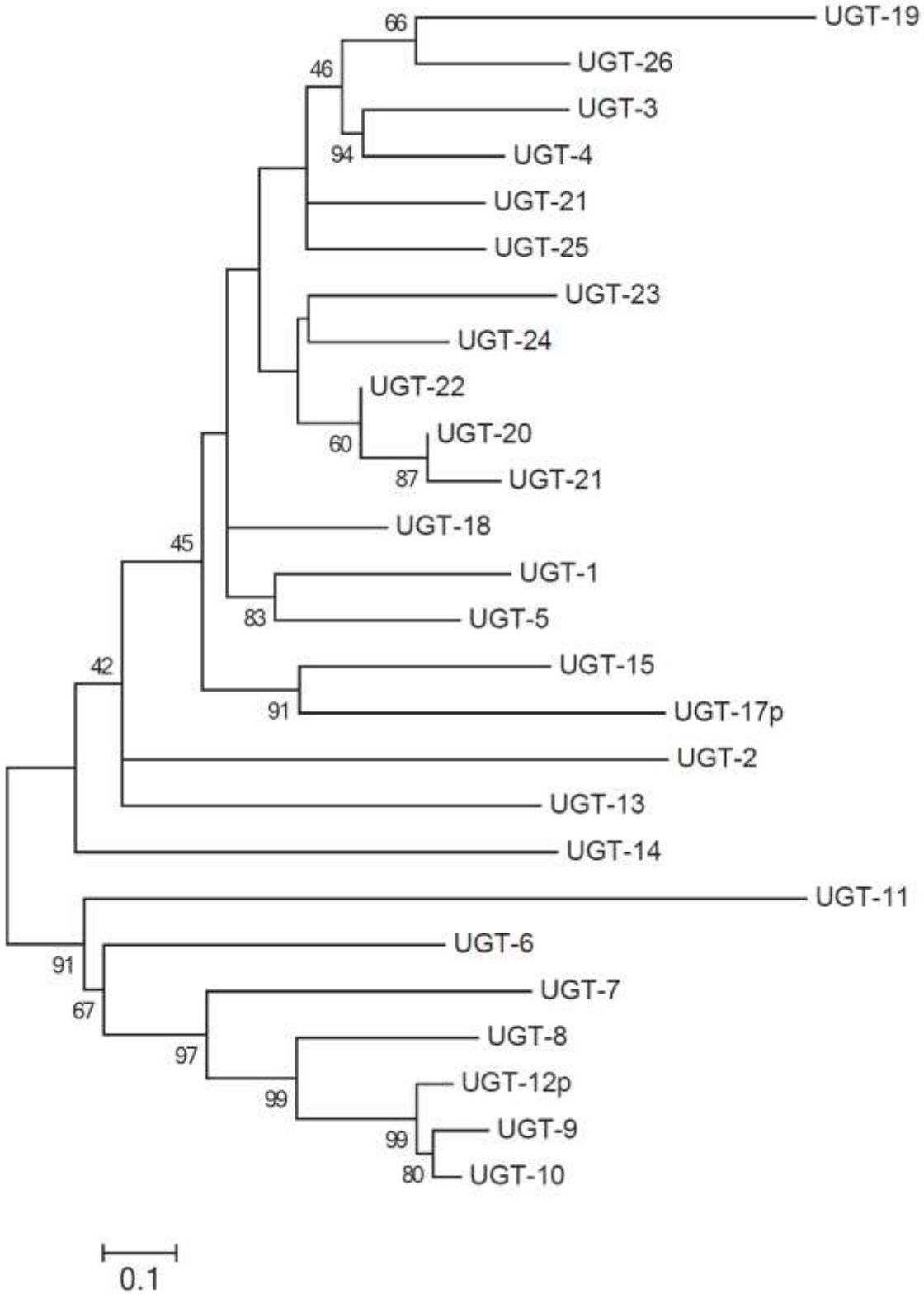
C.



Supplementary Figure 4 : Phylogenetic analysis of the Chemoreceptor families. **(A) Olfactory Receptor family.** The tree was rooted with the highly conserved and basal OrCo proteins. A single asterisk indicates possible simple orthologous relationships and two asterices indicate slightly more complicated relationships involving independent duplications in one or more species. Protein names and the branches leading to them are colored in blue for *Gerris*, brown for *Rhodnius*, red for *Cimex*, and orange for *Oncopeltus*. A suffix of P after the protein number indicates a pseudogene, while alternatively-spliced ORs are indicated by lower case letters after the protein number. Support for nodes is the aLRT value from PhyML v3.0. **(B) Gustatory Receptor family.** The tree was rooted with the conserved sugar and carbon dioxide receptor subfamilies. These two subfamilies and the fructose receptor subfamily are highlighted by colored background wedges. **(C) Ionotropic Receptor family.** The tree was rooted with the conserved co-receptor Ir8a and 25a lineages, which closely resemble the ionotropic glutamate receptors from which these variant Ionotropic Receptors evolved. The entire *D. melanogaster* IR repertoire was included for comparison. Lower case suffixes do not indicate alternative-splicing, but rather either orthology with particular *Drosophila* IRs, or the Ir41 and 75 series of genes.

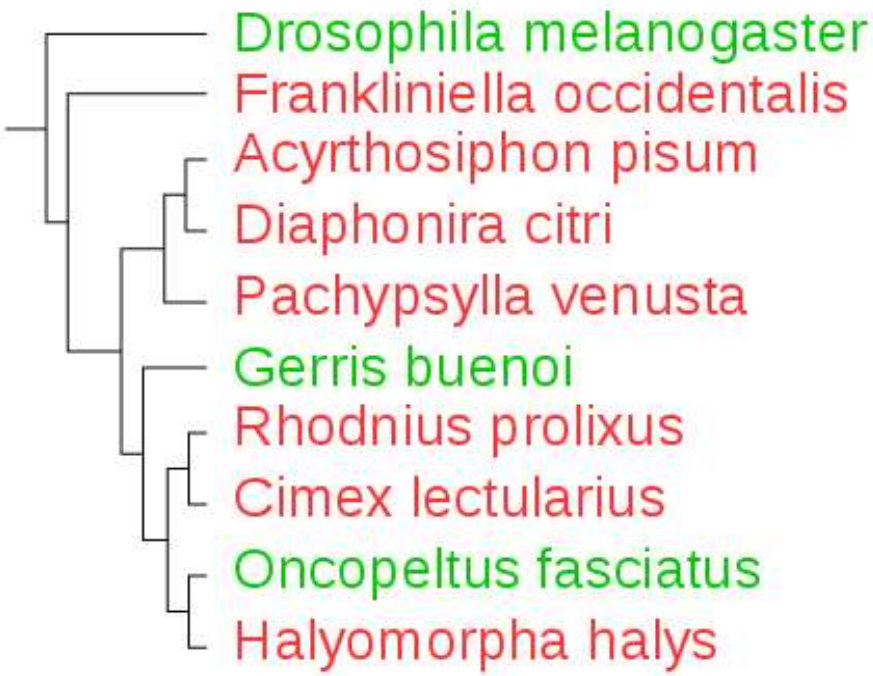


Supplementary Figure 5 : Genomic orientation of UGT genes in a *Gerris buenoi* genomic scaffold. Ten UGT genes are arrayed in a row in Scaffold1549, probably multiplied by gene duplication events.



Supplementary Figure 6 : A consensus Maximum-likelihood tree of C-terminal half of the deduced amino acid sequences of *Gerris buenoi* UGTs. The phylogeny was inferred by the method based on the JTT matrix-based model. Bootstrap value was 1 000.

836
837
838

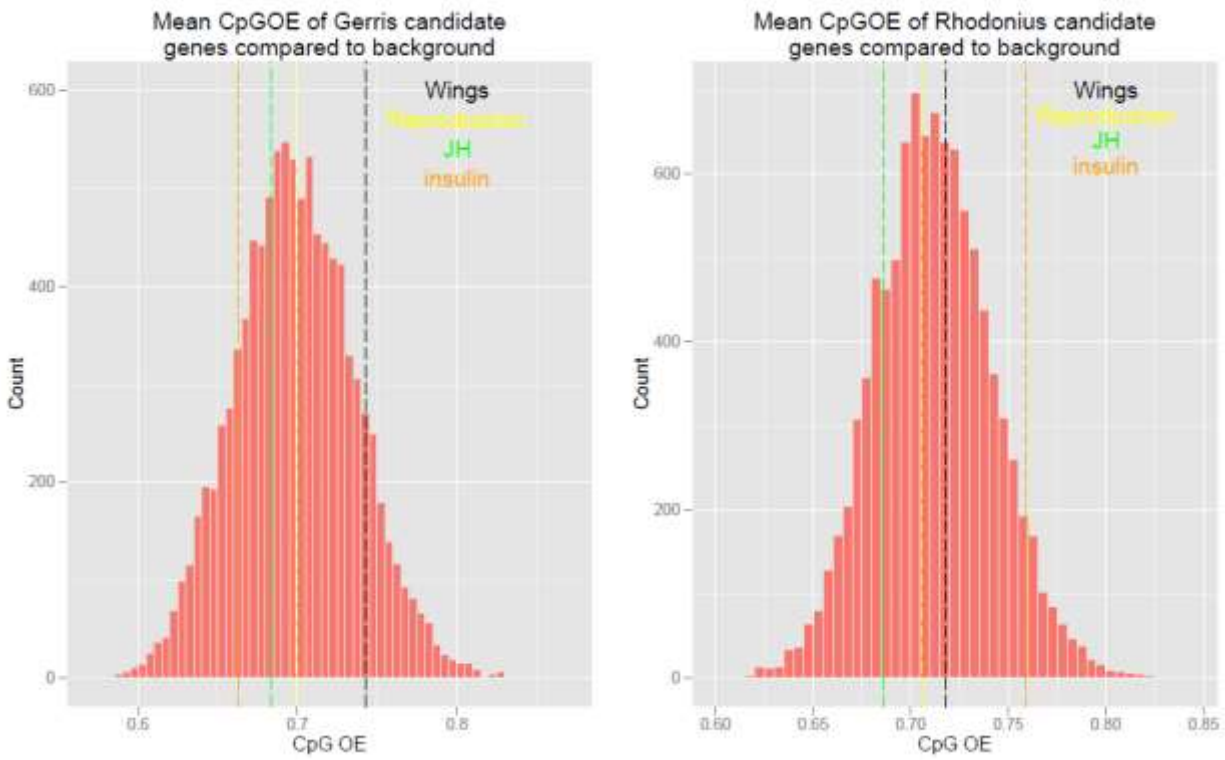


839

Supplementary Figure 7 : Simplified cladogram of Hemiptera based on [157] depicting IMD presence (green) and absence (red).

840

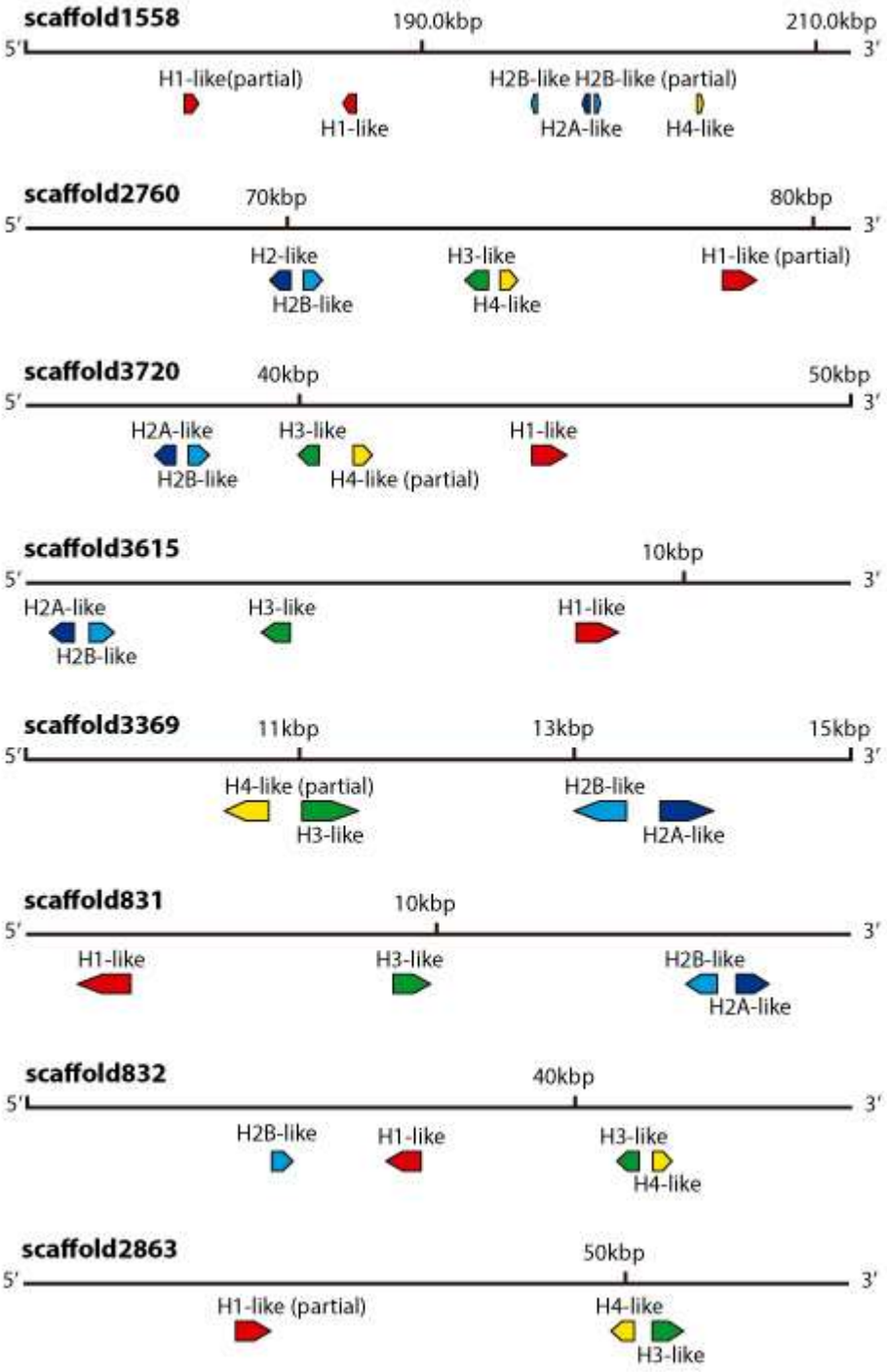
841



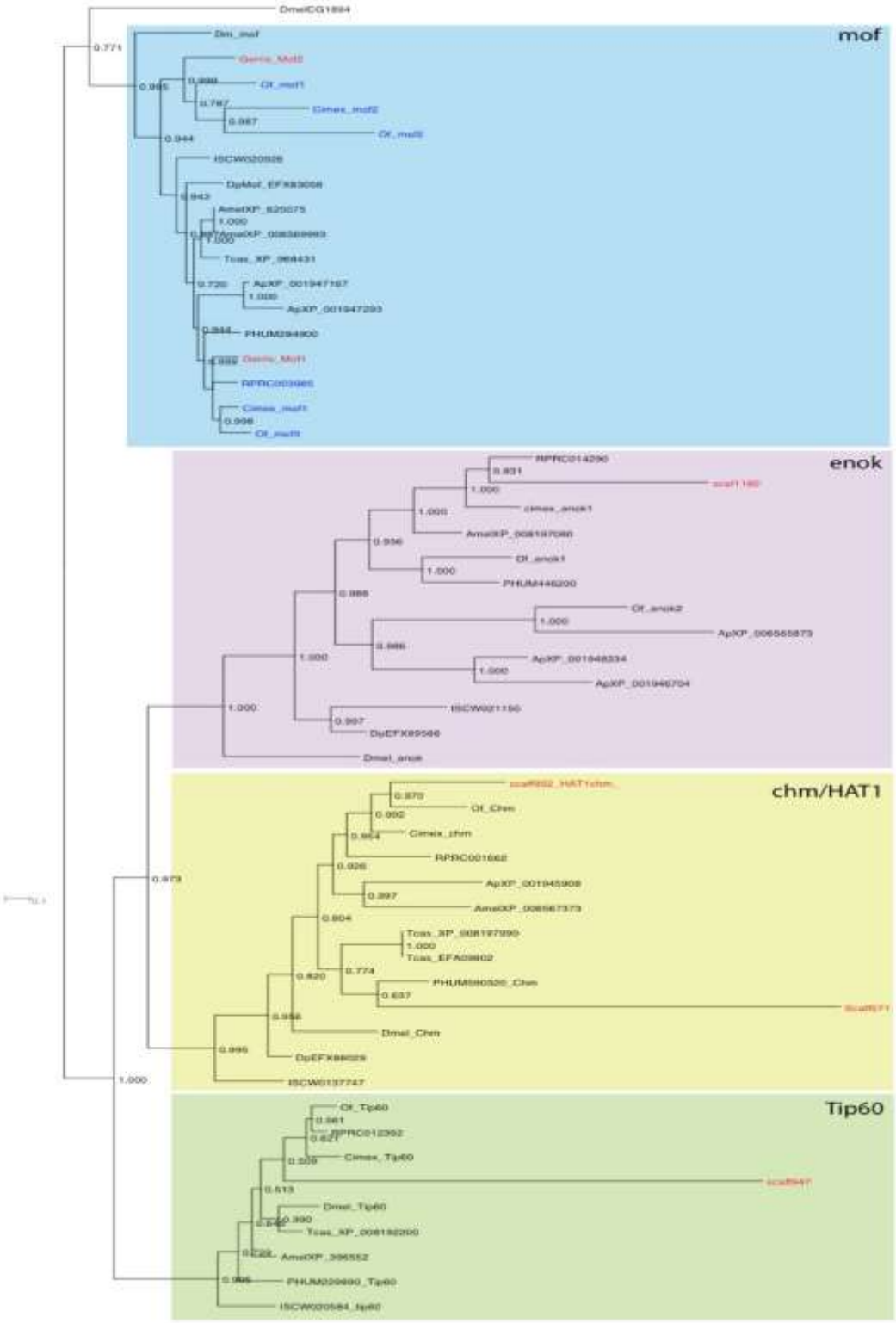
842

Supplementary Figure 8 : Density plot of frequency (y-axis) versus mean $CpG_{O/E}$ (x-axis) for (A) *Gerris buenoi* (n = 20 949; overall mean = 0.70; mean of wing genes = 0.74; mean of juvenile hormone genes = 0.68; mean of insulin signalling genes = 0.66; mean of reproduction genes = 0.70; $p > 0.05$)and (B*) Rhodnius prolixus* (n = 15 081; overall mean = 0.71; mean of wing genes = 0.72; mean of juvenile hormone genes = 0.69; mean of insulin signalling genes = 0.76; mean of reproduction genes = 0.71; $p > 0.05$). The observed mean for genes in the networks underlying wing polyphenism (black line), reproduction (yellow line), juvenile hormone (green line), and insulin signalling (orange line) plotted relative to the distribution of $CpG_{O/E}$ values for all genes in the genome (random resampling of mean $CpG_{O/E}$ from 50 genes in the genome).

843



Supplementary Figure 9 : Genomic organisation of the histone loci gene clusters annotated in the *Gerris buenoi* genome. Clusters were defined as more than one histone encoding gene present on a genomic scaffold. No clusters were found that were interrupted by non-histone gene encoding loci. Clusters were visualized using genometools v1.5.5 and coloured according to orthology group (Histone H1 (red), Histone H2A (dark blue), Histone H2B (light blue), Histone H3 (green), Histone H4 (yellow).



Supplementary Figure 10 : Phylogeny of histone acetyltransferases in Heteropteran lineage. Results show a duplication of *males absent on the first* (*mof*) and *chameau* (*chm*/HAT1) in *Gerris buenoi* similar to previous results in *Oncopeltus fasciatus* [9] and *Cimex lectularius* [81] but also a unique duplication of *Gerris buenoi* histone deacetylase *Sirt1* (*sir2*) and *Sirt5*; and the histone methyltransferase *grappa*.

844

1 739	Complete Single-copy BUSCOs
81	- of which duplicated
490	Fragmented BUSCOs
446	Missing BUSCOs
2 675	Total BUSCO groups searched

Supplementary Table 1 : Summarized benchmarks in BUSCO notation

	Complete	**- of which duplicated**	**Fragmented**	**Missing**
Drosophila melanogaster	98	6,4	0.6	0.3
Danaus plexipus	83	8.6	11	4.3
Apis mellifera	93	2.9	5.1	0.9
Pediculus humanus	92	3.9	6.1	1.6
Daphnia pulex	83	3.9	11	5.1
Tribolium castaneum	95	5.8	3.9	0.8
Acyrthosiphon pisum	72	6.1	15	12
Cimex lectularius	78	9.7	1,4	7.4
Gerris buenoi	65	3.02	18.32	16.67

Supplementary Table 2 : BUSCO Genome assessment based on percentage of BUSCO genes identified (ftp://cegg.unige.ch/OrthoDB7/BUSCO/README.txt). Species results other than *Gerris buenoi* extracted from supplementary data in [81].

Gene	Scaffold: start..end	Locus length (nt)	Protein length (aa)	Number of CDS exons
labial -part 1 of 2	Scaffold2148:49081..49463	383 (partial)	208 (concat-enated)	2 (concat-enated)
labial -part 2 of 2	Scaffold688:18951..20594	1 644 (partial)		
proboscipedia	Scaffold917:82996..178853 - strand	95 858	498	3
zerknüllt	Scaffold917:254614..264809 + strand	10 196	360	3
Deformed	Scaffold927:71079..127936	56 858	339	2
Sex combs reduced	Scaffold111:113209..227662 - strand	114 454	279	2
fushi tarazu	Scaffold111:292364..296153 - strand	3 790	298	2
Antennapedia	Scaffold111:608939..620195 - strand	11 257	284	2
*Ultrabithorax**	Scaffold280:506616..507456	841 (partial)	178 (partial)	1 (partial)
abdominal-A	Scaffold259:352274..461324	109 051	320	3
*Abdominal-B**	Scaffold464:255292..399249	143 958	254 (partial)	2 (partial)
iroquois	Scaffold451:304431-432356	127 926	426	6
mirror	Scaffold2206:85783-151112	65 330	362	5

Supplementary Table 3 : Positional information for the annotated homeobox genes. Incomplete gene models are marked with an asterisk (*). Colored shading highlights gene linkage, and coding strand is also indicated for these gene models.

845

Gene name	Gene abbreviation	*Gerris buenoi*	*Oncopeltus fasciatus*	*Cimex lectularius*
abrupt	*ab*	Yes	Yes	Yes
Achaete-scute complex	*ac*	Yes	No	No
Actin 5C	*Act5C*	Yes	Yes	Yes
amphiphysin	*Amph*	Yes	Yes	Yes
aralar1	*aralar1*	Yes	Yes	Yes
arrow	*arr*	Yes	Yes	Yes
Asense	*ase*	No	Yes	No
astray	*aay*	Yes	Yes	Yes
bantam	*ban*	No	No	No
beadex	*Bx*	Yes	Yes	Yes
bendless	*ben*	Yes	Yes	Yes
bifocal	*bif*	Yes	No	No
bonus	*bon*	Yes	No	Yes
buttonless	*btn*	No	No	No
calreticulin	*Crc*	Yes	Yes	Yes
capricious	*caps*	Yes	Yes	Yes
caupolican	*caup*	No	Yes	Yes
center divider	*cdi*	Yes	Yes	Yes
cornetto	*corn*	No	Yes	No
corto	*corto*	No	No	No
couch potato	*cpo*	Yes	Yes	Yes
crooked legs	*crol*	No	Yes	Yes
dacapo	*dap*	No	Yes	No
dalmatian	*dmt*	No	No	No
Darkener of apricot	*Doa*	No	Yes	Yes
daughterless	*da*	Yes	Yes	No
deadpan	*dpn*	Yes	Yes	Yes
Delta	*Dl*	No	Yes	Yes
diminutive	*dm*	No	No	No
division abnormally delayed	*dally*	No	Yes	Yes
dorsotonals (homothorax)	*hth*	No	Yes	Yes
E(spl) region transcript m7	*E(spl)m7-HLH*	Yes	Yes	Yes
E2F transcription factor	*E2f*	Yes	Yes	Yes
Eb1	*Eb1*	Yes	Yes	Yes

effete	*eff*	Yes	No	No
egghead	*egh*	Yes	Yes	Yes
enabled	*ena*	Yes	Yes	Yes
Enhancer-of-split	*E(spl)m8-HLH*	No	Yes	Yes
EP2237 (cabut)	*cbt*	Yes	No	Yes
escargot	*esg*	Yes	No	Yes
extra macrochaetae	*emc*	Yes	Yes	Yes
flightless	*fliI*	Yes	Yes	Yes
frizzled	*fz*	Yes	Yes	Yes
frizzled 2	*fz2*	Yes	Yes	Yes
ftz transcription factor 1	*ftz-f1*	No	Yes	Yes
gliolectin	*glec*	No	No	No
gliotactin	*Gli*	No	No	Yes
Glutathione S transferase 2	*GstS1*	Yes	Yes	Yes
grapes	*grp*	Yes	Yes	Yes
groucho	*gro*	Yes	Yes	Yes
Hairless	*H*	Yes	Yes	Yes
hairy	*h*	Yes	Yes	Yes
headcase	*hdc*	Yes	Yes	Yes
hephaestus	*heph*	Yes	Yes	Yes
Hormone receptor-like in 39	*Hr39*	No	Yes	Yes
IGF-II mRNA-binding protein	*Imp*	Yes	Yes	Yes
kekkon-1	*kek1*	Yes	Yes	Yes
kuzbanian	*kuz*	Yes	Yes	Yes
Laminin A	*LanA*	Yes	Yes	Yes
lethal (1) G0007	*l(1)G0007*	Yes	Yes	Yes
liquid facets	*lqf*	Yes	Yes	Yes
lola like	*lolal*	Yes	Yes	Yes
longitudinals lacking	*lola*	Yes	Yes	Yes
melted	*melt*	Yes	Yes	Yes
mushroom body defect	*mud*	No	No	No
nebbish	*neb*	Yes	Yes	No
nejire	*nej*	Yes	Yes	Yes
neuralized	*neur*	Yes	Yes	Yes
notch	*N*	Yes	Yes	Yes
nuclear fallout	*nuf*	No	No	No
pavarotti	*pav*	Yes	Yes	Yes

pebble	*pbl*	Yes	Yes	Yes
pipsqueak	*psq*	Yes	Yes	Yes
pointed	*pnt*	Yes	Yes	Yes
Poly(ADP-ribose) glycohydrolase	*Parg*	Yes	Yes	Yes
polychaetoid	*pyd*	Yes	Yes	Yes
prospero	*pros*	Yes	Yes	Yes
Protein kinase 61C	*Pdk1*	Yes	Yes	Yes
Protein tyrosine phosphatase 10D	*Ptp10D*	Yes	Yes	Yes
pumilio	*pum*	Yes	Yes	Yes
pxb	*pxb*	No	No	No
quemao	*qm*	Yes	Yes	Yes
Ras oncogene at 85D	*Ras85D*	No	Yes	Yes
Ras-like protein A	*Rala*	Yes	No	No
raspberry	*ras*	Yes	Yes	Yes
Rhomboid	*rho*	Yes	Yes	Yes
Ribosomal protein S5	*RpS5a*	Yes	Yes	Yes
roundabout	*robo*	Yes	Yes	Yes
rutabaga	*rut*	No	Yes	Yes
sanpodo	*spdo*	Yes	Yes	Yes
scabrous	*sca*	Yes	Yes	Yes
scalloped	*sd*	Yes	Yes	Yes
scratch	*scrt*	Yes	Yes	No
scribbled	*scrib*	Yes	Yes	Yes
scribbler	*sbb*	Yes	Yes	Yes
scute	*sc*	No	Yes	Yes
seven up	*svp*	Yes	Yes	Yes
shaggy	*sgg*	Yes	Yes	Yes
singed	*sn*	Yes	Yes	Yes
smooth	*sm*	Yes	Yes	Yes
Sp1	*Sp1*	No	Yes	No
SP71 (Trynity)	*Tyn*	Yes	Yes	Yes
spitz	*spi*	No	No	No
split ends	*spen*	Yes	Yes	Yes
string	*stg*	Yes	Yes	Yes
sugarless	*sgl*	Yes	Yes	Yes
taranis	*tara*	Yes	Yes	Yes
Tcp-1eta	*Tcp-1eta*	Yes	Yes	Yes

Tollo	*Tollo*	Yes	Yes	Yes
tout-velu	*ttv*	No	Yes	Yes
tramtrack	*ttk*	Yes	Yes	Yes
Trehalose receptor 1 (Trapped in endoderm 1)	*Tre1*	No	Yes	No
tribbles	*trbl*	Yes	Yes	Yes
tweety	*tty*	Yes	Yes	Yes
Twin of m4	*Tom*	No	No	No
u-turn (ventral veins lacking)	*wl*	Yes	Yes	Yes
Ubiquitin activating enzyme 1	*Uba1*	Yes	Yes	Yes
Ubiquitin conjugating enzyme 2	*UbcD2*	Yes	Yes	No
Vacuolar H+ ATPase 16kD subunit	*Vha16-1*	Yes	Yes	Yes
ß-amyloid protein precursor-like	*Appl*	Yes	Yes	Yes

Supplementary Table 4 : Annotation of genes involved in bristle number and neural development based on *Drosophila melanogaster* quantitative analyses [162].

846

Species	**Order**	**Suborder**	**LWS**	**SWS-B**	**SWS-UV**	**Rh7**	**Arthropsin**	**c-Opsin**
Gerris buenoi	Hemiptera	Heteroptera	4	-	1	1	1	1
Cimex lectularius	Hemiptera	Heteroptera	1	-	1	1	-	1
Rhodnius prolixus	Hemiptera	Heteroptera	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	Auchenorryhncha	1	1	1	na	na	na

Supplementary Table 5 : Opsin conservation in Hemiptera. [80, 81, 163-165]

Species	CPR_RR-1	CPR_RR-2	CPR_Uncl	CPAP1	CPAP3	CPF	TWDL	Total
Drosophila melanogaster	61	42	34	29	10	5	29	210
Glossina morsitans	33	27	17	11	6	1	9	104
Culex quinquefasciatus	49	97	30	10	8	5	9	208
Aedes aegypti	66	150	28	14	9	3	6	276
Anopheles gambiae	43	103	21	13	10	4	12	206
Bombyx mori	47	78	19	13	6	1	4	168
Danaus plexippus	47	57	18	16	10	1	5	154
Apis mellifera	13	15	10	15	7	4	2	66
Nasonia vitripennis	19	32	18	16	6	5	2	98
Pediculus humanus	9	15	17	12	6	0	2	61
Daphnia pulex	101	36	152	20	12	0	0	321
Tetranychus urticae	0	7	31	14	5	0	0	57
Tribolium castaneum	34	55	21	13	7	5	3	138
Acyrthosiphon pisum	9	84	20	10	8	2	3	136
Cimex lectularius	18	70	32	15	6	5	3	149
Gerris buenoi	22	74	30	10	6	3	10	155

Supplementary Table 6 : Detection and classification of putative structural cuticular proteins. Information from other species than *Gerris buenoi* adapted from Ioannidou, et al. [153] and Benoit, et al. [81].

	Scaffold #	# Genes	Family	Length (Kbp)	Density (Kbp/gene)
1	431	14	CPR RR-1/CPR Uncl	398	28.4
2	32	13	CPR RR-2	183	14.1
3	41	9	CPR RR-2	92	10.2
4	349	8	CPR RR-2	224	27.9
5	996	6	CPR RR-2	73	12.2
6	683	4	CPAP3	250	62.5
7	2496	4	CPR RR-2/CPR Uncl	92	23.0
8	46	3	CPF	49	16.2
9	80	3	TWDL	62	20.6
10	132	3	CPR Uncl	249	83.1
11	706	3	CPR Uncl	66	21.9

Supplementary Table 7 : Clusters of genes coding cuticle proteins in the genome of *Gerris buenoi*

847
848

	Ionotropic	Gustatory	Odorant
Gerris buenoi	45/45	60/135	153/155
Oncopeltus fasciatus	37/37	115/169	120/121
Rhodnius prolixus	33/33	28/30	116/116
Cimex lectularius	30/30	24/36	48/49
Drosophila melanogaster	65/65	60/68	60/62

Supplementary Table 8 : Numbers of genes and encoded proteins in three chemoreceptor families in heteropterans with genome sequences, and *Drosophila melanogaster* for comparison.

Gene name	OGS name	Genomic scaffold	Length (aa)	Remark
UGT-01	GBUE014547-RA	Scaffold1506	530	complete
UGT-02	GBUE015333-RA	Scaffold1907	515	complete
UGT-03	GBUE018966-RA	Scaffold3228	533	complete
UGT-04	GBUE018967-RA	Scaffold3228	515	complete
UGT-05	GBUE018968-RA	Scaffold3228	543	complete
UGT-06	GBUE014164-RA	Scaffold2126	524	complete
UGT-07	GBUE014165-RA	Scaffold2126	527	complete
UGT-08	GBUE013499-RA-1	Scaffold1323	529	complete
UGT-09	GBUE013499-RA-2	Scaffold1323	512	complete
UGT-10	GBUE013499-RA-3	Scaffold1323	527	complete
UGT-11	GBUE013500-RA	Scaffold1323	218	partial
UGT-12p*	GBUE019125-RA	Scaffold3054	470	partial
UGT-13	GBUE010586-RA	Scaffold838	524	complete
UGT-14	GBUE012986-RA	Scaffold1320	697	complete
UGT-15	GBUE013062-RA	Scaffold1042	437	partial
UGT-16	GBUE020555-RA	Scaffold5464	326	partial
UGT-17p	GBUE020560-RA	Scaffold6284	243	partial
UGT-18	GBUE012772-RA	Scaffold1549	422	partial
UGT-19	GBUE012773-RA	Scaffold1549	347	partial
UGT-20	GBUE012774-RA	Scaffold1549	434	partial
UGT-21	GBUE012775-RA	Scaffold1549	201	partial
UGT-22	GBUE012776-RA	Scaffold1549	378	partial
UGT-23	GBUE012777-RA	Scaffold1549	522	complete
UGT-24	GBUE012778-RA	Scaffold1549	540	complete
UGT-25	GBUE012779-RA	Scaffold1549	519	complete
UGT-26	GBUE012780-RA	Scaffold1549	534	complete
UGT-27	GBUE012781-RA	Scaffold1549	529	complete
UGT-28	no OGS name	Scaffold4983	235	partial

Supplementary Table 9 : List of UDP-glycosyltransferase genes in *Gerris buenoi* genome. (*refers to pseudogene.)

Gene type	*Gene name*	Location [Accession#]	Protein Length	Domains
Gap	*orthodenticle*	Scaffold177:468498-481641 + strand GbueTmpM005873-RA	542	zinc finger C2H2
	buttonhead	Scaffold1076:201924-220125 + strand GbueTmpM009254-RA	437	zinc finger M2C2

	collier	Scaffold128:650434 - 661155 + strand GbueTmpM003852-RA GbueTmpM003853-RA	236	IPT Superfamily
	cap-n-collar	Scaffold1737:125742 - 180215 + strand GbueTmpA013482-RA	414	bZIP Superfamily
	crocodile	Scaffold417:94048 -94890 – strand GbueTmpA005876-RA	280	Forkhead Superfamily
	Krüppel	Scaffold66:273659 - 274706 + strand GbueTmpA001375-RA	246	zinc finger M2C2
	huckebein	Scaffold1050:35145-36625 + strand GbueTmpA011673-RA	153	zinc finger C2H2
	empty spiracles	Scaffol640:42899 - 108829 + strand GbueTmpA010166-RA GbueTmpA010167-RA GbueTmpA010168-RA	237	Homeobox Superfamily
	giant	Scaffold1313:205754 - 259477 – strand GbueTmpM012482-RA	290	bZIP Superfamily
	gomdanji	Scaffold177:546605-551844 + strand GbueTmpM005874-RA	101	??meth_res Superfamily
Segment polarity	*shifted*	Scaffold4383:9065-11975 + strand	268	WIF Superfamily
	roadkill	Scaffold7:1346380-1347546 + strand	388	MATH superfamily BTB Domain
	perli-like	Scaffold542:114112-120587 – strand GbueTmpM009219-RA	214	Perli Domain
	microtubule star	Scaffold83:875664-876641 + strand	325	MPP Superfamily
	flapwing	Scaffold362:255357-260899 - strand	240	MPP Superfamily
	cullin1	Scaffold15:198529-200865 - strand	778	Cullin Superfamily
	dispatched	Scaffold2487:39972-52586 - strand	434	ND
	costa	Scaffold666:175038-178490 + strand	1150	Kinesin Domain

	paxillin	Scaffold927:228836-245479 - strand	300	LIM Superfamily
Terminal patterning	*Torso*	Scaffold626:104429-114052 + strand GbueTmpA010687-RA	412	PKc_like superfamily FN3 superfamily
	Torso-like	Scaffold7:1642089-1661652 + strand	356	MACPF Superfamily
General	*decapentaplegic*	Scaffold488:247243-262588 – strand GbueTmpA009289-RA	323	TGF-Beta Domain
	cubitus interruptus	Scaffold2762:18072-42310 – strand GbueTmpA017830-RA	960	zinc finger-H
	lipophorin-like	Scaffold940:71827-78984 + strand GbueTmp8010317-RA	1202	DUF1943 Superfamily

Supplementary Table 10 : Current Early Developmental Genes identified in the *Gerris buenoi* genome. The table lists Gap Genes and Segment Polarity Genes models, model location and accession number, protein length, and protein domain identified in the model.

Gerris buenoi early patterning genes	
Gap Genes	
caudal	?
hunchback	Yes
orthodenticle	Yes
buttonhead	Yes
collier	Yes
cap-n-collar	Yes
crocodile	Yes
Krüppel	Yes
huckebein	Yes
sloppy-paired	Yes
empty spiracles	Yes
giant	Yes
knirps	Yes
tailless	Yes
gomdanji	Yes
Pair Rule Genes	
even-skipped	Yes
paired	Yes
odd-skipped	Yes
paired	Yes
runt	Yes
hairy	Yes
Tenascin major	Yes
sister-of-odd-and-bowl	Yes
Segment Polarity Genes	
engrailed	Yes
invected	Yes
shifted	Yes
roadkill	Yes
peril-like	Yes
patched	Yes
nejire	Yes
microtubule star	Yes
flapwing	Yes
cullin1	Yes
dispatched	Yes
costa	Yes
paxillin	Yes
Terminal Patterning Genes	
torso	Yes
PTTH	?
torso-like	Yes
trunk	No

Supplementary Table 11 : Presence/absence of *Drosophila melanogaster* early patterning genes in the genomes of *Gerris buenoi.*

849

Gene type	Gene name	*Drosophila melanogaster*		*Tribolium castaneum*	
		QC (ID)	Bit Score	QC (ID)	Bit Score
Gap	*orthodenticle*	24% (98%)	94	42% (71%)	97
	buttonhead	26% (61%)	145	29% (70%)	187
	collier	43% (70%)	94	47% (62%)	90
	cap-n-collar	19% (43%)	47	56% (34%)	116
	crocodile	81% (52%)	234	39% (65%)	166
	Krüppel	70% (71%)	242	77% (59%)	240
	huckebein	71% (68%)	174	ND	ND
	empty spiracles	86% (71%)	179	90% (60%)	281
	giant	33% (62%)	77	40% (48%)	117
	gomdanji	64% (34%)	45	ND	ND
Segment polarity	*shifted*	95% (55%)	286	92% (68%)	350
	roadkill	96% (56%)	424	96% (57%)	437
	peril-like	77% (56%)	194	96% (61%)	262
	microtubule star	84% (50%)	295	ND	ND
	flapwing	87% (42%)	184	ND	ND
	cullin1	99% (61%)	956	100% (79%)	1274
	dispatched	98% (25%)	181	96% (31%)	181
	costa	72% (36%)	227	70% (27%)	159
	paxillin	89% (65%)	367	81% (66%)	330
Terminal patterning	*Torso*	93% (53%)	140	92% (31%)	194
	Torso-like	90% (46%)	321	89% (49%)	335
General	*decapentaplegic*	55% (34%)	173	58% (39%)	295
	cubitus interruptus	94% (46%)	301	98% (42%)	280
	lipophorin-like	77% (23%)	215	95% (33%)	587

Supplementary Table 12 : Represents Query Coverage (Identity) and E-value of the annotated gene models pairwise aligned to orthologues in other species. Pairwise alignment was performed using NCBI blast. ND – Not Determined.

850

Gene	**Scaffold: start..end**	**Locus length (nt)**	**Protein length (aa)**	**Number of CDS exons**
axin	Scaffold136:602832..659508	56 677	1496	16
*armadillo**	Scaffold2236:76533..96972	20 440 (partial)	716 (partial)	11
arrow	Scaffold136:139587..222403	82 817	1490	24
dishevelled -RA	Scaffold441:78333..107479	29 147	602	15
dishevelled -RB	Scaffold441:78333..124793	46 461	597	14
frizzled	Scaffold288:270554..271759	1 206	401	1
frizzled-2	Scaffold1053:141773..145781	4 009	597	1
frizzled-3	Scaffold304:383672..482292	98 621	500	2
*glycogen synthase kinase-3 beta -RA -part 1 of 2**	Scaffold1391:148463..174822	26 360 (partial)	302 (partial)	6
*glycogen synthase kinase-3 beta -RB -part 1 of 2**	Scaffold1391:148463..174822	26 360 (partial)	286 (partial)	6
*glycogen synthase kinase-3 beta -part 2 of 2**	Scaffold10229:2..3044	3 043 (partial)	150 (partial)	2
wingless	Scaffold2771:12925..70979	58 055	331	3
Wnt7	Scaffold163:273675..338565	64 891	456	10
Wnt8	Scaffold1136:57077..65015	7 939	302	5
Wnt5	Scaffold3063:28070..66680	38 611	321	6
Wnt10	Scaffold2796:27374..49167	21 794	273	5

*WntA**	Scaffold20: 632685..638039	5 355 (partial)	287 (partial)	5
wntless	Scaffold190:240723..250315	9 593	538	11

Supplementary Table 13 : Positional information for the 18 Wnt signaling genes annotated. Incomplete gene models are marked with an asterisk (*).

Gene set	Gene name	*Gerris buenoi* $CpG_{O/E}$ value	*Rhodnius proxilus* $CpG_{O/E}$ value
Insulin signalling	*Chico*	0.658374618	
Insulin signalling	*forkhead box protein O*		
Insulin signalling	*Foxo*	1.014152563	0.963514594
Insulin signalling	*Insulin receptor 1*	1.090538511	1.133882478
Insulin signalling	*Insulin receptor 1-like*	0.865210624	
Insulin signalling	*Insulin receptor 2*	0.394382326	0.781348977
Insulin signalling	*Insulin receptor substrate*		
Insulin signalling	*Phosphatase and tensine homologue*	0.444946289	
Insulin signalling	*Phosphoinositide 3-kinase Pi3K21B*	0.730078776	0.783423219
Insulin signalling	*Phosphoinositide 3-kinase Pi3K92E*	0.438681484	0.575389176
Insulin signalling	*Protein Kinase B*	0.395861448	0.563182964
Insulin signalling	*Rheb/Ras homolog enriched in brain*	0.540547798	
Insulin signalling	*RPS6-p70-protein kinase*	0.731629717	0.704464786
Insulin signalling	*Slimfast*	0.77679356	0.7171875
Insulin signalling	*Target of rapamycin*	0.648267284	0.641665967
Insulin signalling	*Thor*	0.910084034	0.907818533
Insulin signalling	*Tsc1 Tuberous sclerosis complex 1*	0.383532463	0.517120208
Insulin signalling	*Tsc2/gigas/Tuberin*	0.580956324	0.815878378
Juvenile Hormone	*Allostatin C*		
Juvenile Hormone	*broad*	0.905797101	0.973075749
Juvenile Hormone	*Chd64*		
Juvenile Hormone	*FK506-binding protein 1*	0.480397835	0.498673415

Juvenile Hormone	*FK506-binding protein 14 ortholog*		
Juvenile Hormone	*FK506-binding protein FKBP59*	0.553441364	0.759341109
Juvenile Hormone	*Juvenile hormone acid methyltransferase*	0.853085106	0.627682228
Juvenile Hormone	*Juvenile hormone epoxide hydrolase 1*	0.497504096	0.823006391
Juvenile Hormone	*Juvenile hormone esterase*		
Juvenile Hormone	*Juvenile hormone esterase duplication*		
Juvenile Hormone	*Juvenile hormone-inducible protein 1*	0.437671182	0.579799692
Juvenile Hormone	*Juvenile hormone-inducible protein 26*	0.90600823	0.302261307
Juvenile Hormone	*Kruppel homolog 1*	0.883615819	1.019771301
Juvenile Hormone	*Methoprene-tolerant*	0.633364098	0.59144385
Juvenile Hormone	*taiman*		
Reproduction	*Armitage*	0.900408271	0.735040693
Reproduction	*Aubergine (annotated as Piwi-like)*	0.497755107	
Reproduction	*Bazooka/PAR-3*		0.68762606
Reproduction	*cappuccino*		
Reproduction	*capsuleen*		
Reproduction	*Dynein light chain 90F*	1.006892418	0.784722222
Reproduction	*eIF5B*	0.66963049	0.699717583
Reproduction	*Heat shock protein 83/90*	0.661795474	0.78564613
Reproduction	*Heat shock protein 83/90 2*		0.558785904
Reproduction	*Hunchback*	1.00601711	0.902307812
Reproduction	*Laminin A*		
Reproduction	*Laminin B2*		
Reproduction	*loki/Chk2*		
Reproduction	*maelstrom*		
Reproduction	*meiotic 41/ATR*	0.318670549	
Reproduction	*Merlin*		
Reproduction	*Moesin*		
Reproduction	*nanos*	0.392635135	
Reproduction	*N-ethylmaleimide-sensitive factor 2*		
Reproduction	*Par - 6*	0.493019601	0.75739645
Reproduction	*Par-1*		
Reproduction	*pebble/ECT2*	0.346433041	0.556323529

Reproduction	*Piwi (annotated as piwi-like)*		
Reproduction	*Rab11*	1.21100186	0.891789661
Reproduction	*sevenless*		
Reproduction	*Smaug*		0.374331551
Reproduction	*Spindle-D*		
Reproduction	*Spindle-E*		
Reproduction	*staufen*	0.335958039	0.63898769
Reproduction	*Stellate*		
Reproduction	*telomere fusion*		
Reproduction	*tudor*	1.249130153	0.799734986
Reproduction	*vasa*	0.693071093	
Wing	*Acetylcholine esterase*	1.056863669	0.931578947
Wing	*apterous*		
Wing	*argos*		
Wing	*armadillo*	0.404645677	0.517999969
Wing	*baboon*	0.516144578	0.667751211
Wing	*basket*	0.740959251	0.7426405
Wing	*bifid*		
Wing	*blistered*		
Wing	*brinker*	0.876838162	0.574162679
Wing	*Buffy*	0.602699055	
Wing	*capricious*	0.914409241	0.819466248
Wing	*clot*	0.872160934	1.03902439
Wing	*cut*	0.362195409	0.754880803
Wing	*Death regulator Nedd2-like caspase*	1.146718147	
Wing	*Death related ICE-like caspase*	0.6890625	0.804121212
Wing	*Death-associated inhibitor of apoptosis 1*		
Wing	*Decapping protein 1*	1.189357953	0.711444547
Wing	*division abnormally delayed*	0.535155846	0.513011152
Wing	*eiger*	0.838224085	0.839430894
Wing	*engrailed*	1.247013856	0.814175728
Wing	*Epidermal growth factor receptor*	0.697416093	0.659715546
Wing	*fringe*	1.294816794	0.638368984
Wing	*hedgehog*	0.964415584	0.773176471
Wing	*Keren*	0.754096776	
Wing	*Mad1*	0.517751479	
Wing	*Mad2*	0.414863782	0.499577603
Wing	*Mad3*	0.440286166	0.555261005
Wing	*mastermind*	0.516834008	0.690080382
Wing	*Medea*	0.497130418	0.534404253
Wing	*mind bomb 1*	0.533591731	0.771083019
Wing	*nemo*	1.09630137	0.58400637

Wing	*Nipped-A*		
Wing	*patched*	0.475015567	0.732220161
Wing	*punt*	0.990559836	0.428825279
Wing	*punt 2*	0.451908397	
Wing	*Ras oncogene at 85D*	1.040664452	0.743847875
Wing	*saxophone*	0.949921557	
Wing	*schnurri*	0.347452969	
Wing	*Serrate*		
Wing	*smoothened*	0.431910569	
Wing	*spalt major*	0.925619236	0.699655862
Wing	*Star*	0.658335154	
Wing	*Suppressor of Hairless*	0.542231327	0.624452765
Wing	*tartan*	0.732986444	1.144366197
Wing	*thickveins*	0.586962236	
Wing	*wingless*	1.124115983	1.017095821

Supplementary Table 14 : List of genes in the networks underlying wing polyphenism, reproduction, juvenile hormone, and insulin signalling included in the analysis and their $CpG_{O/E}$ value for *Gerris buenoi* and *Rhodnius prolixus*. Genes that were annotated in *Gerris buenoi* but excluded from the analysis because they did have a complete codding sequence are also listed but without a $CpG_{O/E}$ value.

	Core histones				
	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
Gerris buenoi	**10**	**11**	**9**	**10**	**9**
Daphnia pulex	5	10	12	10	6
Tetranychus urticae	1	4	7	6	3
Ixodes scapularis	4	6	4	4	1
Strigamia maritima	3	7	15	4	4

Supplementary Table 15 : Number of loci within the genomes of arthropod species encoding the five classes of histones. Orthologs for *Aedes aegypti*, *Daphnia pulex*, *Tetranychus urticae* and *Ixodes scapularis* were obtained by BLAST analysis. Orthologs for *Apis mellifera* and *Acyrthosiphon pisum* were obtained from published literature [136, 166]. Orthologs for *Oncopeltus fasciatus* (manuscript in preparation) and *Cimex lectularius* [81] were obtained during genome annotation.

851

Species	**Number of antioxidant genes**
Acyrthosiphon pisum	6
Apis mellifera	1
Bombyx mori	2
Cimex lectularis	16
Drosophila melanogaster	0
Pediculus humanus	0
Tribolium castaneum	5

Supplementary Table 16 : Number of genes for each species compared to that had highest similarity to *G. Buenoi* antioxidant genes.

852

Bio Projects	i5K Pilot NCBI Bio-project	PRJNA163973 https://www.ncbi.nlm.nih.gov/bioproject/163973
	Gerris buenoi NCBI Bio-project	PRJNA203045 https://www.ncbi.nlm.nih.gov/bioproject/203045
	NCBI Bio-sample	SAMN02800617 https://www.ncbi.nlm.nih.gov/biosample/2800617
Genome Sequence	180bp insert *male* DNA	1 Illumina HiSeq 2000 run: 122.1M read pairs, 24.7Gbp
	500bp insert *male* DNA	1 Illumina HiSeq 2500 run: 36.4M read pairs, 7.4Gbp
	3kb insert *male* DNA	1 Illumina HiSeq 2000 run: 137.4M read pairs, 27.8 Gbp
	8kb insert *female* DNA	1 Illumina HiSeq 2000 run: 135.9M read pairs, 27.4 Gbp
	180bp insert NCBI SRA Accession	SRX493944 https://www.ncbi.nlm.nih.gov/sra/SRX493944
	500bp insert NCBI SRA Accession	SRX493946 https://www.ncbi.nlm.nih.gov/sra/SRX493946
	3kb insert NCBI SRA Accession	SRX493945 https://www.ncbi.nlm.nih.gov/sra/SRX493945
	8kb insert NCBI SRA Accession	SRX493943 https://www.ncbi.nlm.nih.gov/sra/SRX493943
Genome Assembly	Number of contigs	304,893
	Contig N50	3,812 bp
	Number of scaffolds	20,259
	Scaffold N50	344,118 bp
	Size of final assembly	1,000,161,732 bp
	Size of final assembly - without gaps	653,297,297 bp
	NCBI Genome Assembly Accession	GCA_001010745.1 https://www.ncbi.nlm.nih.gov/assembly/GCA_001010745.1
RNAseq data	*Gerris buenoi* Transcriptome Bio-project	PRJNA275657 https://www.ncbi.nlm.nih.gov/bioproject/275657
	Mixed sex embryos and nymphs RNAseq reads	32M read pairs, 6.5 Gbp
	Mixed sex embryos and nymphs SRA Accession	SRX896710 https://www.ncbi.nlm.nih.gov/sra/SRX896710
Automated Genome Annotation (Gbue_0.5.3)	Genes (Gbue_0.5.3)	20 949
	Average Transcript length	1 298
	Average CDS length	954 bp (318 aa)
	Exons per gene	4.81
	Genome Annotation Link	National Agricultural Library https://i5k.nal.usda.gov/Gerris_buenoi

Supplementary Table 17 : Sequencing, assembly, annotation statistics and accession numbers

Supplementary Sequences

Antioxidants excel table

155 GbueOr protein sequences
135 GbueGr protein sequences
45 GbueIR protein sequences

Serosins nucleotide sequences

>Serosin_1 (Scaffold3130)

ATGGCTCGCTACACTCTTCTGTGTGTTATTGCTTCATGCCTGGTTGCCCTTGCTGTTTCGGTGCCTTTTGAGCAGAAAAC
AGCTTTCGAGTTAAAGGAACGTCACGACTTCTACAACCCTAGGAGCGACAACCCGTTCAGCACGTCTGGATCGGATGCAC
ACATGAAGACACAAAGCGCTAGAGTAGAGCATGACTTCATCGGAGGCAAGAACTGGGCTGCTGGTGGTTACGCTCAACAT
GAAAGACAGAGCATGTTCGGACAGACCCGTCGTAACAACGAAGGTGGATTCCAATTTAAAGCGAGATTTTAG

>Serosin_2 (Scaffold2193)

ATGGCCCGCTACACTCTCCTCTGTGTTATCGCTTCATGCCTGGTTGCTCTTGCTGTTTCGGTGCCGTTCGAACAGAAAAC
AGCTTTCGAGTTAAAGGAACGTCACGACTTCTACAACCCTAGGAGCGACAACCCGTTCAGCACGTCTGGATCGGATGCAC
ACATGAAGACACAAAGCGCTAGAGTAGAGCACGACTTTATCGGAGGCAAAAACTGGGCTGCTGGTGGTTACGCTCAACAT
GAAAGACAGAGCATGTTCGGACAGACCCGTCGTAACAACGAAGGTGGATTCCAATTTAAAGCAAGATTTTAG

>Serosin_3 (Scaffold3130)

ATGGCCCGTTACACTCTCCTCTGTGTTATTGCTTCCTGCCTGGTGGCTCTTGCTGTTTCGGTGCCGTTCGAGCAGAAAAC
AGCTTTCGAGTTGAAGGAACGTCACGACTTCTACAACCCTAGGAGCGACAACCCGTTCAGCACGTCTGGATCGGATGCAC
ACATGAAGACACAAAGCGCTAGAGTAGAGCATGACTTCATCGGAGGCAAGAACTGGGCTGCTGGTGGTTACGCTCAACAT
GAAAGACAGAGCATGTTCGGACAGACCCGTCGTAACAACGAAGGTGGATTCCAATTTAAAGCGAGATTTTAG

>Serosin_4 (Scaffold2193)

ATGGCTCGCTACACTCTCCTTTGTGTTATCGCTTCCTGCCTGGTTGCTCTTGCTGTTTCGGTGCCGTTCGAACAGAAAAC
AGCTTTCGAGTTGAAGGAACGTCACGACTTCTACAACCCTAGGAGCGACAACCCGTTCAGCACGTCTGGATCGGATGCAC
ACATGAAGACTCAAAGCGCTAGAATAGAGCATGACTTTATCGGAGGCAAGAACTGGGCTGCTGGTGGTTACGCTCAACAT
GAAAGACAGAGCATGTTCGGACAGACCCGACGAAACAACGAAGGTGGATTTCAATTTAAAGCAAGATTTTAG

>Serosin_5 (Scaffold3130)

ATGGCTCGCTACACTCTCCTCTGTGTTATCGCTTCCTGCCTGGTGGCTCTTGCTGTTTCGGTGCCGTTCGAACAGAAAAC
ATCTTTCGACTATAAGGAACGTCACGACTTCCAGGACAACCCGTCTGGATCGGATGCCCACATGAAGACCCAAAGAGCTA
GAGTAGAACATGACTTTGTTGGAGGCAAGAACTGGGCTGCTGGTGGTTACGTTCAACACGAAAGACAGACTATGTACGGA
GAGACACGTAAGCAAAACGAAGGAGGAGTCCAAGTTAAAGTAACATTTTAG

>Serosin_6 (Scaffold3130)

ATGGTCCGCCACGCTTTGTTTTGTGTTATCGCTTTCTGCCTGGTTACTCTCGCTGTTTCGGTGCCATTTGAGCAGAAAAC
AGCTTTTGACTACAAGGAACGTCACGACTTCTATAATCCTAAGAACGACAACCCGTTCAGTACGTCTGGATCGGATGCAC
ACATGAAGACACAAAGCGCTAGAGTAGAGCATGACTTTGCCGGAGGCAAGAATTGGGCTGCTGGTTTTTACGCTCAACAT
GAAAGACAGAATATGAACGGACAGTCCCGTCGTAACAACGAAGCTGGATTCCAATTTAAAGGAACATTTTAG

Serosins protein sequences

>Serosin_1 (Scaffold3130)

MARYTLLCVIASCLVALAVSVPFEQKTAFELKERHDFYNPRSDNPFSTSGSDAHMKTQSARVEHDFIGGKNWAAGGYAQH
ERQSMFGQTRRNNEGGFQFKARF

>Serosin_2 (Scaffold2193)

MARYTLLCVIASCLVALAVSVPFEQKTAFELKERHDFYNPRSDNPFSTSGSDAHMKTQSARVEHDFIGGKNWAAGGYAQH
ERQSMFGQTRRNNEGGFQFKARF

>Serosin_3 (Scaffold3130)

MARYTLLCVIASCLVALAVSVPFEQKTAFELKERHDFYNPRSDNPFSTSGSDAHMKTQSARVEHDFIGGKNWAAGGYAQH
ERQSMFGQTRRNNEGGFQFKARF

>Serosin_4 (Scaffold2193)

MARYTLLCVIASCLVALAVSVPFEQKTAFELKERHDFYNPRSDNPFSTSGSDAHMKTQSARIEHDFIGGKNWAAGGYAQH
ERQSMFGQTRRNNEGGFQFKARF

>Serosin_5 (Scaffold3130)

MARYTLLCVIASCLVALAVSVPFEQKTSFDYKERHDFQDNPSGSDAHMKTQRARVEHDFVGGKNWAAGGYVQHERQTMYG
ETRKQNEGGVQVKVTF

>Serosin_6 (Scaffold3130)

MVRHALFCVIAFCLVTLAVSVPFEQKTAFDYKERHDFYNPKNDNPFSTSGSDAHMKTQSARVEHDFAGGKNWAAGFYAQH
ERQNMNGQSRRNNEAGFQFKGTF

References

1. Kongton, K., K. McCall, and A. Phongdara, *Identification of gamma-interferon-inducible lysosomal thiol reductase (GILT) homologues in the fruit fly Drosophila melanogaster.* Dev Comp Immunol, 2014. **44**(2): p. 389-96.
2. De Gregorio, E., et al., *The Toll and Imd pathways are the major regulators of the immune response in Drosophila.* EMBO J, 2002. **21**(11): p. 2568-79.
3. Zou, Z., et al., *Comparative genomic analysis of the Tribolium immune system.* Genome Biol, 2007. **8**(8): p. R177.
4. Chipman, A.D., et al., *The first myriapod genome sequence reveals conservative arthropod gene content and genome organisation in the centipede Strigamia maritima.* PLoS Biol, 2014. **12**(11): p. e1002005.
5. Hoffmann, J.A. and J.M. Reichhart, *Drosophila innate immunity: an evolutionary perspective.* Nat Immunol, 2002. **3**(2): p. 121-6.
6. Tzou, P., E. De Gregorio, and B. Lemaitre, *How Drosophila combats microbial infection: a model to study innate immunity and host-pathogen interactions.* Curr Opin Microbiol, 2002. **5**(1): p. 102-10.
7. Consortium, T.I.A.G., *Genome sequence of the pea aphid Acyrthosiphon pisum.* PLoS Biol, 2010. **8**(2): p. e1000313.
8. Gerardo, N.M., et al., *Immunity and other defenses in pea aphids, Acyrthosiphon pisum.* Genome Biol, 2010. **11**(2): p. R21.
9. Panfilio, K.A., et al., *Molecular evolutionary trends and feeding ecology diversification in the Hemiptera, anchored by the milkweed bug genome.* bioRxiv, 2017.
10. Jenssen, H., P. Hamill, and R.E. Hancock, *Peptide antimicrobial agents.* Clin Microbiol Rev, 2006. **19**(3): p. 491-511.
11. Vilcinskas, A., *Evolutionary plasticity of insect immunity.* J Insect Physiol, 2013. **59**(2): p. 123-9.
12. Jacobs, C.G.C., et al., *Immune function of the serosa in a hemimetabolous insect egg*, in *BMC Evolutionary Biology (Submitted)*. 2017.
13. Armisen, D., et al., *Predator strike shapes antipredator phenotype through new genetic interactions in water striders.* Nat Commun, 2015. **6**: p. 8153.
14. Khila, A., E. Abouheif, and L. Rowe, *Evolution of a novel appendage ground plan in water striders is driven by changes in the Hox gene Ultrabithorax.* PLoS Genet, 2009. **5**(7): p. e1000583.
15. Kalinka, A.T., et al., *Gene expression divergence recapitulates the developmental hourglass model.* Nature, 2010. **468**(7325): p. 811-4.
16. Refki, P.N. and A. Khila, *Key patterning genes contribute to leg elongation in water striders.* Evodevo, 2015. **6**: p. 14.
17. Kirkness, E.F., 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**(27): p. 12168-73.
18. Christiaens, O., et al., *Halloween genes and nuclear receptors in ecdysteroid biosynthesis and signalling in the pea aphid.* Insect Mol Biol, 2010. **19 Suppl 2**: p. 187-200.
19. Shigenobu, S., 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**: p. 47-62.
20. Naggan Perl, T., et al., *The evolution of the knirps family of transcription factors in arthropods.* Mol Biol Evol, 2013. **30**(6): p. 1348-57.
21. Watanabe, T., H. Takeuchi, and T. Kubo, *Structural diversity and evolution of the N-terminal isoform-specific region of ecdysone receptor-A and -B1 isoforms in insects.* BMC Evol Biol, 2010. **10**: p. 40.
22. Dang, C.W., et al., *The basic helix-loop-helix transcription factor family in the pea aphid, Acyrthosiphon pisum.* J Insect Sci, 2011. **11**: p. 84.
23. Bitra, K., et al., *Functional characterization of PAS and HES family bHLH transcription factors during the metamorphosis of the red flour beetle, Tribolium castaneum.* Gene, 2009. **448**(1): p. 74-87.
24. Baker, K.D. and C.S. Thummel, *Diabetic larvae and obese flies-emerging studies of metabolism in Drosophila.* Cell Metab, 2007. **6**(4): p. 257-66.
25. Edgar, B.A., *How flies get their size: genetics meets physiology.* Nat Rev Genet, 2006. **7**(12): p. 907-16.
26. Martin, D.E. and M.N. Hall, *The expanding TOR signaling network.* Curr Opin Cell Biol, 2005. **17**(2): p. 158-66.
27. Junger, M.A., et al., *The Drosophila forkhead transcription factor FOXO mediates the reduction in cell number associated with reduced insulin signaling.* J Biol, 2003. **2**(3): p. 20.
28. Puig, O. and R. Tjian, *Transcriptional feedback control of insulin receptor by dFOXO/FOXO1.* Genes Dev, 2005. **19**(20): p. 2435-46.

29. Kapahi, P., et al., *Regulation of lifespan in Drosophila by modulation of genes in the TOR signaling pathway.* Curr Biol, 2004. **14**(10): p. 885-90.
30. Wang, M.C., D. Bohmann, and H. Jasper, *JNK extends life span and limits growth by antagonizing cellular and organism-wide responses to insulin signaling.* Cell, 2005. **121**(1): p. 115-25.
31. Wullschleger, S., R. Loewith, and M.N. Hall, *TOR signaling in growth and metabolism.* Cell, 2006. **124**(3): p. 471-84.
32. Emlen, D.J., et al., *Insulin signaling and limb-patterning: candidate pathways for the origin and evolutionary diversification of beetle 'horns'.* Heredity (Edinb), 2006. **97**(3): p. 179-91.
33. Hattori, A., et al., *Soldier morphogenesis in the damp-wood termite is regulated by the insulin signaling pathway.* J Exp Zool B Mol Dev Evol, 2013. **320**(5): p. 295-306.
34. Patel, A., et al., *The making of a queen: TOR pathway is a key player in diphenic caste development.* PLoS One, 2007. **2**(6): p. e509.
35. Emlen, D.J., et al., *A mechanism of extreme growth and reliable signaling in sexually selected ornaments and weapons.* Science, 2012. **337**(6096): p. 860-4.
36. Snell-Rood, E.C. and A.P. Moczek, *Insulin signaling as a mechanism underlying developmental plasticity: the role of FOXO in a nutritional polyphenism.* PLoS One, 2012. **7**(4): p. e34857.
37. Murat, S., C. Hopfen, and A.P. McGregor, *The function and evolution of Wnt genes in arthropods.* Arthropod Struct Dev, 2010. **39**(6): p. 446-52.
38. Oberhofer, G., et al., *Wnt/beta-catenin signaling integrates patterning and metabolism of the insect growth zone.* Development, 2014. **141**(24): p. 4740-50.
39. Janssen, R., et al., *Conservation, loss, and redeployment of Wnt ligands in protostomes: implications for understanding the evolution of segment formation.* BMC Evol Biol, 2010. **10**: p. 374.
40. Beermann, A., et al., *A context-dependent combination of Wnt receptors controls axis elongation and leg development in a short germ insect.* Development, 2011. **138**(13): p. 2793-805.
41. Rawlings, N.D., A.J. Barrett, and R. Finn, *Twenty years of the MEROPS database of proteolytic enzymes, their substrates and inhibitors.* Nucleic Acids Res, 2016. **44**(D1): p. D343-50.
42. Turk, V., et al., *Cysteine cathepsins: from structure, function and regulation to new frontiers.* Biochim Biophys Acta, 2012. **1824**(1): p. 68-88.
43. Terra, W.R. and C. Ferreira, *Insect digestive enzymes: properties, compartmentalization and function.* Comparative Biochemistry and Physiology Part B: Comparative Biochemistry, 1994. **109**(1): p. 1-62.
44. Terra, W.R. and C. Ferreira, *11 - Biochemistry and Molecular Biology of Digestion A2 - Gilbert, Lawrence I*, in *Insect Molecular Biology and Biochemistry*. 2012, Academic Press: San Diego. p. 365-418.
45. Murdock, L.L., et al., *Cysteine digestive proteinases in Coleoptera.* Comparative Biochemistry and Physiology Part B: Comparative Biochemistry, 1987. **87**(4): p. 783-787.
46. Houseman, J.G. and A.E.R. Downe, *Cathepsin D-like activity in the posterior midgut of hemipteran insects.* Comparative Biochemistry and Physiology Part B: Comparative Biochemistry, 1983. **75**(3): p. 509-512.
47. Martynov, A.G., et al., *Functional analysis of C1 family cysteine peptidases in the larval gut of capital Tenebrio molitor and Tribolium castaneum.* BMC Genomics, 2015. **16**: p. 75.
48. Perkin, L., E.N. Elpidina, and B. Oppert, *Expression patterns of cysteine peptidase genes across the Tribolium castaneum life cycle provide clues to biological function.* PeerJ, 2016. **4**: p. e1581.
49. Oppert, B., et al., *Microarray analysis reveals strategies of Tribolium castaneum larvae to compensate for cysteine and serine protease inhibitors.* Comp Biochem Physiol Part D Genomics Proteomics, 2010. **5**(4): p. 280-7.
50. Kollien, A.H., et al., *Activity and sequence characterization of two cysteine proteases in the digestive tract of the reduviid bug Triatoma infestans.* Insect Mol Biol, 2004. **13**(6): p. 569-79.
51. Waniek, P.J., et al., *Cathepsin L of Triatoma brasiliensis (Reduviidae, Triatominae): sequence characterization, expression pattern and zymography.* J Insect Physiol, 2012. **58**(1): p. 178-87.
52. Ribeiro, J.M., et al., *An insight into the transcriptome of the digestive tract of the bloodsucking bug, Rhodnius prolixus.* PLoS Negl Trop Dis, 2014. **8**(1): p. e2594.
53. Novinec, M. and B. Lenarcic, *Papain-like peptidases: structure, function, and evolution.* Biomol Concepts, 2013. **4**(3): p. 287-308.
54. Sakurai, M., et al., *Distribution of tubulointerstitial nephritis antigen-like 1 and structural matrix proteins in mouse embryos during preimplantation development in vivo and in vitro.* Zygote, 2014. **22**(2): p. 259-65.
55. Saito, H., S. Kurata, and S. Natori, *Purification and characterization of a hemocyte proteinase of Sarcophaga, possibly participating in elimination of foreign substances.* Eur J Biochem, 1992. **209**(3): p. 939-44.
56. Gruden, K., et al., *Diverse enzymatic specificities of digestive proteases, 'intestains', enable Colorado potato beetle larvae to counteract the potato defence mechanism.* Biol Chem, 2003. **384**(2): p. 305-10.
57. Meyer, H.W., *Visuelle Schlüsselreize für die Auslösung der Beutefanghandlung beim Bachwasserläufer Velia caprai (Hemiptera, Heteroptera).* Zeitschrift für vergleichende Physiologie, 1971. **72**(3): p. 260-297.
58. Rowe, L., *The costs of mating and mate choice in water striders.* Animal Behaviour, 1994. **48**: p. 1049-56.

59. Spence, J.R. and N. Anderson, *Biology of water striders: interactions between systematics and ecology.* Annual Review of Entomology, 1994. **39**(1): p. 101-128.
60. Dahmen, H., *Eye specialisation in waterstriders: an adaptation to life in a flat world.* Journal of Comparative Physiology A, 1991. **169**(5): p. 623-632.
61. Wolburg-Buchholz, K., *The organization of the lamina ganglionaris of the hemipteran insects, Notonecta glauca, Corixa punctata and Gerris lacustris.* Cell Tissue Res, 1979. **197**(1): p. 39-59.
62. Schneider, L. and H. Langer, *Die Struktur des Rhabdoms im „Doppelauge" des Wasserläufers Gerris lacustris*. 1969: Zeitschrift für Zellforschung und Mikroskopische Anatomie.
63. Fischer, C., M. Mahner, and E. Wachmann, *The rhabdom structure in the ommatidia of the Heteroptera (Insecta), and its phylogenetic significance.* Zoomorphology, 2000. **120**(1): p. 1-13.
64. Schwind, R., *Polarization vision in water insects and insects living on a moist substrate.* Journal of Comparative Physiology A, 1991. **169**(5): p. 531-540.
65. Bohn, H. and U. Täuber, *Beziehungen zwischen der Wirkung polarisierten Lichtes auf das Elektroretinogramm und der Ultrastruktur des Auges von Gerris lacustris L.* Zeitschrift für vergleichende Physiologie, 1971. **72**(1): p. 32-53.
66. Frolov, R. and M. Weckström, *Developmental changes in biophysical properties of photoreceptors in the common water strider (Gerris lacustris): better performance at higher cost.* Journal of neurophysiology, 2014. **112**(4): p. 913-922.
67. Bartsch, K., *Polarization-sensitive photoreceptors of different spectral types in the compound eye of waterstriders.* Naturwissenschaften, 1995. **82**(6): p. 292-293.
68. Henze, M.J. and T.H. Oakley, *The Dynamic Evolutionary History of Pancrustacean Eyes and Opsins.* Integr Comp Biol, 2015. **55**(5): p. 830-42.
69. Frentiu, F.D., et al., *Adaptive evolution of color vision as seen through the eyes of butterflies.* Proc Natl Acad Sci U S A, 2007. **104 Suppl 1**: p. 8634-40.
70. Frentiu, F.D., et al., *Gene duplication is an evolutionary mechanism for expanding spectral diversity in the long-wavelength photopigments of butterflies.* Mol Biol Evol, 2007. **24**(9): p. 2016-28.
71. Sharkey, C.R., et al., *Overcoming the loss of blue sensitivity through opsin duplication in the largest animal group, beetles.* Sci Rep, 2017. **7**(1): p. 8.
72. Briscoe, A.D. and L. Chittka, *The evolution of color vision in insects.* Annu Rev Entomol, 2001. **46**: p. 471-510.
73. Brody, T. and A. Cravchik, *Drosophila melanogaster G protein-coupled receptors.* J Cell Biol, 2000. **150**(2): p. F83-8.
74. Senthilan, P.R. and C. Helfrich-Forster, *Rhodopsin 7-The unusual Rhodopsin in Drosophila.* PeerJ, 2016. **4**: p. e2427.
75. Colbourne, J.K., et al., *The ecoresponsive genome of Daphnia pulex.* Science, 2011. **331**(6017): p. 555-61.
76. Eriksson, B.J., et al., *Characterisation and localisation of the opsin protein repertoire in the brain and retinas of a spider and an onychophoran.* BMC Evol Biol, 2013. **13**: p. 186.
77. Hering, L. and G. Mayer, *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**(9): p. 2380-91.
78. Benton, R., *Multigene Family Evolution: Perspectives from Insect Chemoreceptors.* Trends Ecol Evol, 2015. **30**(10): p. 590-600.
79. Joseph, R.M. and J.R. Carlson, *Drosophila Chemoreceptors: A Molecular Interface Between the Chemical World and the Brain.* Trends Genet, 2015. **31**(12): p. 683-95.
80. Mesquita, R.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**(48): p. 14936-41.
81. Benoit, J.B., et al., *Unique features of a global human ectoparasite identified through sequencing of the bed bug genome.* Nat Commun, 2016. **7**: p. 10165.
82. International Aphid Genomics, C., *Genome sequence of the pea aphid Acyrthosiphon pisum.* PLoS Biol, 2010. **8**(2): p. e1000313.
83. Ioannidis, P., et al., *Genomic features of the damselfly Calopteryx splendens representing a sister clade to most insect orders.* Genome Biol Evol, 2017.
84. Missbach, C., et al., *Evolution of insect olfactory receptors.* Elife, 2014. **3**: p. e02115.
85. Robertson, H.M., *The Insect Chemoreceptor Superfamily Is Ancient in Animals.* Chem Senses, 2015. **40**(9): p. 609-14.
86. Robertson, H.M., C.G. Warr, and J.R. Carlson, *Molecular evolution of the insect chemoreceptor gene superfamily in Drosophila melanogaster.* Proc Natl Acad Sci U S A, 2003. **100 Suppl 2**: p. 14537-42.
87. Miyamoto, T., et al., *A fructose receptor functions as a nutrient sensor in the Drosophila brain.* Cell, 2012. **151**(5): p. 1113-25.
88. Rytz, R., V. Croset, and R. Benton, *Ionotropic receptors (IRs): chemosensory ionotropic glutamate receptors in Drosophila and beyond.* Insect Biochem Mol Biol, 2013. **43**(9): p. 888-97.

89. Croset, V., et al., *A molecular and neuronal basis for amino acid sensing in the Drosophila larva.* Sci Rep, 2016. **6**: p. 34871.
90. Ganguly, A., et al., *A Molecular and Cellular Context-Dependent Role for Ir76b in Detection of Amino Acid Taste.* Cell Rep, 2017. **18**(3): p. 737-750.
91. Enjin, A., et al., *Humidity Sensing in Drosophila.* Curr Biol, 2016. **26**(10): p. 1352-8.
92. Knecht, Z.A., et al., *Distinct combinations of variant ionotropic glutamate receptors mediate thermosensation and hygrosensation in Drosophila.* Elife, 2016. **5**.
93. Hussain, A., et al., *Ionotropic Chemosensory Receptors Mediate the Taste and Smell of Polyamines.* PLoS Biol, 2016. **14**(5): p. e1002454.
94. Min, S., et al., *Dedicated olfactory neurons mediating attraction behavior to ammonia and amines in Drosophila.* Proc Natl Acad Sci U S A, 2013. **110**(14): p. E1321-9.
95. Ai, M., et al., *Acid sensing by the Drosophila olfactory system.* Nature, 2010. **468**(7324): p. 691-5.
96. Gorter, J.A., et al., *The nutritional and hedonic value of food modulate sexual receptivity in Drosophila melanogaster females.* Sci Rep, 2016. **6**: p. 19441.
97. Grosjean, Y., et al., *An olfactory receptor for food-derived odours promotes male courtship in Drosophila.* Nature, 2011. **478**(7368): p. 236-40.
98. Prieto-Godino, L.L., et al., *Olfactory receptor pseudo-pseudogenes.* Nature, 2016. **539**(7627): p. 93-97.
99. Prieto-Godino, L.L., et al., *Evolution of Acid-Sensing Olfactory Circuits in Drosophilids.* Neuron, 2017. **93**(3): p. 661-676 e6.
100. Koh, T.W., et al., *The Drosophila IR20a clade of ionotropic receptors are candidate taste and pheromone receptors.* Neuron, 2014. **83**(4): p. 850-65.
101. Stewart, S., et al., *Candidate ionotropic taste receptors in the Drosophila larva.* Proc Natl Acad Sci U S A, 2015. **112**(14): p. 4195-201.
102. Simpson, S.J., G.A. Sword, and N. Lo, *Polyphenism in insects.* Curr Biol, 2011. **21**(18): p. R738-49.
103. Roff, D.A., *The Evolution of Wing Dimorphism in Insects.* Evolution, 1986. **40**(5): p. 1009-1020.
104. Harada, T. and K. Taneda, *Seasonal changes in alary dimorphism of a water strider, Gerris paludum insularis (Motschulsky).* Journal of Insect Physiology, 1989. **35**(12): p. 919-924.
105. Vepsäläinen, K., *Determination of wing length and diapause in water-striders (gerris fabr., heteroptera).* Hereditas, 1974. **77**(2): p. 163-76.
106. Vepsäläinen, K., *Wing Dimorphism and Diapause in Gerris: Determination and Adaptive Significance*, in *Evolution of Insect Migration and Diapause*, H. Dingle, Editor. 1978, Springer US. p. 218-253.
107. Suen, G., et al., *The genome sequence of the leaf-cutter ant Atta cephalotes reveals insights into its obligate symbiotic lifestyle.* PLoS Genet, 2011. **7**(2): p. e1002007.
108. Smith, C.R., et al., *Draft genome of the red harvester ant Pogonomyrmex barbatus.* Proc Natl Acad Sci U S A, 2011. **108**(14): p. 5667-72.
109. Smith, C.D., et al., *Draft genome of the globally widespread and invasive Argentine ant (Linepithema humile).* Proc Natl Acad Sci U S A, 2011. **108**(14): p. 5673-8.
110. Brisson, J.A., *Aphid wing dimorphisms: linking environmental and genetic control of trait variation.* Philos Trans R Soc Lond B Biol Sci, 2010. **365**(1540): p. 605-16.
111. Elango, N., et al., *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**(27): p. 11206-11.
112. Nijhout, H.F., *Development and evolution of adaptive polyphenisms.* Evol Dev, 2003. **5**(1): p. 9-18.
113. Nijhout, H.F., *Insect polyphenisms and adaptation.* American Zoologist, 2001. **41**(6): p. 1540-1540.
114. Xu, H.J., et al., *Two insulin receptors determine alternative wing morphs in planthoppers.* Nature, 2015. **519**(7544): p. 464-7.
115. Nijhout, H.F., *Control Mechanisms of Polyphenic Development in Insects: In polyphenic development, environmental factors alter some aspects of development in an orderly and predictable way.* Bioscience, 1999. **49**(3): p. 181-192.
116. Alvarado, S., et al., *Epigenetic variation in the Egfr gene generates quantitative variation in a complex trait in ants.* Nat Commun, 2015. **6**: p. 6513.
117. Foret, S., et al., *DNA methylation dynamics, metabolic fluxes, gene splicing, and alternative phenotypes in honey bees.* Proc Natl Acad Sci U S A, 2012. **109**(13): p. 4968-73.
118. Li-Byarlay, H., et al., *RNA interference knockdown of DNA methyl-transferase 3 affects gene alternative splicing in the honey bee.* Proc Natl Acad Sci U S A, 2013. **110**(31): p. 12750-5.
119. Kucharski, R., et al., *Nutritional control of reproductive status in honeybees via DNA methylation.* Science, 2008. **319**(5871): p. 1827-30.
120. Herb, B.R., et al., *Reversible switching between epigenetic states in honeybee behavioral subcastes.* Nat Neurosci, 2012. **15**(10): p. 1371-3.
121. Glastad, K.M., et al., *DNA methylation in insects: on the brink of the epigenomic era.* Insect Mol Biol, 2011. **20**(5): p. 553-65.

122. Bewick, A.J., et al., *Evolution of DNA Methylation across Insects.* Mol Biol Evol, 2017. **34**(3): p. 654-665.
123. Walsh, T.K., et al., *A functional DNA methylation system in the pea aphid, Acyrthosiphon pisum.* Insect Mol Biol, 2010. **19 Suppl 2**: p. 215-28.
124. Vandegehuchte, M.B., et al., *Occurrence of DNA methylation in Daphnia magna and influence of multigeneration Cd exposure.* Environ Int, 2009. **35**(4): p. 700-6.
125. Terrapon, N., et al., *Molecular traces of alternative social organization in a termite genome.* Nat Commun, 2014. **5**: p. 3636.
126. Bonasio, R., et al., *Genomic comparison of the ants Camponotus floridanus and Harpegnathos saltator.* Science, 2010. **329**(5995): p. 1068-71.
127. Wang, Y., et al., *Functional CpG methylation system in a social insect.* Science, 2006. **314**(5799): p. 645-7.
128. Duncan, E.J., P.D. Gluckman, and P.K. Dearden, *Epigenetics, plasticity, and evolution: How do we link epigenetic change to phenotype?* J Exp Zool B Mol Dev Evol, 2014. **322**(4): p. 208-20.
129. Andersen, N.M., xf, and ller, *The Evolution of Wing Polymorphism in Water Striders (Gerridae): A Phylogenetic Approach.* Oikos, 1993. **67**(3): p. 433-443.
130. Andersen, N.M., *The semiaquatic bugs*. ENTOMONOGRAPH, ed. N.M.A.L. Lyneborg. Vol. 3. 1982, Klampenborg – Denmark: SCANDINAVIAN SCIENCE PRESS LTD.
131. Fairbairn, D.J. and E. King, *Why do Californian striders fly?* J Evol Biol, 2009. **22**(1): p. 36-49.
132. Arnqvist, G. and L. Rowe, *Sexual Conflict*. 2005: Princeton University Press.
133. McKay, D.J., et al., *Interrogating the function of metazoan histones using engineered gene clusters.* Dev Cell, 2015. **32**(3): p. 373-86.
134. Conrad, T., et al., *The MOF chromobarrel domain controls genome-wide H4K16 acetylation and spreading of the MSL complex.* Dev Cell, 2012. **22**(3): p. 610-24.
135. Pushpavalli, S.N., et al., *Drosophila MOF controls Checkpoint protein2 and regulates genomic stability during early embryogenesis.* BMC Mol Biol, 2013. **14**: p. 1.
136. Rider, S.D., Jr., D.G. Srinivasan, and R.S. Hilgarth, *Chromatin-remodelling proteins of the pea aphid, Acyrthosiphon pisum (Harris).* Insect Mol Biol, 2010. **19 Suppl 2**: p. 201-14.
137. Furuyama, T., et al., *SIR2 is required for polycomb silencing and is associated with an E(Z) histone methyltransferase complex.* Curr Biol, 2004. **14**(20): p. 1812-21.
138. Rogina, B. and S.L. Helfand, *Sir2 mediates longevity in the fly through a pathway related to calorie restriction.* Proc Natl Acad Sci U S A, 2004. **101**(45): p. 15998-6003.
139. Tissenbaum, H.A. and L. Guarente, *Increased dosage of a sir-2 gene extends lifespan in Caenorhabditis elegans.* Nature, 2001. **410**(6825): p. 227-30.
140. List, O., et al., *Overexpression of grappa encoding a histone methyltransferase enhances stress resistance in Drosophila.* Hereditas, 2009. **146**(1): p. 19-28.
141. Bi, J.L. and G.W. Felton, *Foliar oxidative stress and insect herbivory: Primary compounds, secondary metabolites, and reactive oxygen species as components of induced resistance.* J Chem Ecol, 1995. **21**(10): p. 1511-30.
142. Mittapalli, O., J.J. Neal, and R.H. Shukle, *Antioxidant defense response in a galling insect.* Proc Natl Acad Sci U S A, 2007. **104**(6): p. 1889-94.
143. Pardini, R.S., *Toxicity of oxygen from naturally occurring redox-active pro-oxidants.* Arch Insect Biochem Physiol, 1995. **29**(2): p. 101-18.
144. Corona, M. and G.E. Robinson, *Genes of the antioxidant system of the honey bee: annotation and phylogeny.* Insect Mol Biol, 2006. **15**(5): p. 687-701.
145. Felton, G.W. and C.B. Summers, *Antioxidant systems in insects.* Arch Insect Biochem Physiol, 1995. **29**(2): p. 187-97.
146. Shi, G.Q., Q.Y. Yu, and Z. Zhang, *Annotation and evolution of the antioxidant genes in the silkworm, Bombyx mori.* Arch Insect Biochem Physiol, 2012. **79**(2): p. 87-103.
147. Gnerre, S., et al., *High-quality draft assemblies of mammalian genomes from massively parallel sequence data.* Proc Natl Acad Sci U S A, 2011. **108**(4): p. 1513-8.
148. Cantarel, B.L., et al., *MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes.* Genome Res, 2008. **18**(1): p. 188-96.
149. Stanke, M., et al., *Using native and syntenically mapped cDNA alignments to improve de novo gene finding.* Bioinformatics, 2008. **24**(5): p. 637-44.
150. Korf, I., *Gene finding in novel genomes.* BMC Bioinformatics, 2004. **5**: p. 59.
151. Gramates, L.S., et al., *FlyBase at 25: looking to the future.* Nucleic Acids Res, 2017. **45**(D1): p. D663-D671.
152. Willis, J.H., *Structural cuticular proteins from arthropods: annotation, nomenclature, and sequence characteristics in the genomics era.* Insect Biochem Mol Biol, 2010. **40**(3): p. 189-204.
153. Ioannidou, Z.S., et al., *CutProtFam-Pred: detection and classification of putative structural cuticular proteins from sequence alone, based on profile hidden Markov models.* Insect Biochem Mol Biol, 2014. **52**: p. 51-9.
154. Bird, A.P., *DNA methylation and the frequency of CpG in animal DNA.* Nucleic Acids Res, 1980. **8**(7): p.

1499-504.
155. Weber, M., et al., *Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome.* Nat Genet, 2007. **39**(4): p. 457-66.
156. Wang, L., et al., *BeetleBase: the model organism database for Tribolium castaneum.* Nucleic Acids Res, 2007. **35**(Database issue): p. D476-9.
157. Misra, J.R., et al., *Transcriptional regulation of xenobiotic detoxification in Drosophila.* Genes Dev, 2011. **25**(17): p. 1796-806.
158. Wallace, I.M., et al., *M-Coffee: combining multiple sequence alignment methods with T-Coffee.* Nucleic Acids Res, 2006. **34**(6): p. 1692-9.
159. Capella-Gutierrez, S., J.M. Silla-Martinez, and T. Gabaldon, *trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses.* Bioinformatics, 2009. **25**(15): p. 1972-3.
160. Tamura, K., et al., *MEGA6: Molecular Evolutionary Genetics Analysis version 6.0.* Mol Biol Evol, 2013. **30**(12): p. 2725-9.
161. Jones, D.T., W.R. Taylor, and J.M. Thornton, *The rapid generation of mutation data matrices from protein sequences.* Comput Appl Biosci, 1992. **8**(3): p. 275-82.
162. Norga, K.K., et al., *Quantitative analysis of bristle number in Drosophila mutants identifies genes involved in neural development.* Curr Biol, 2003. **13**(16): p. 1388-96.
163. Gao, N., R.G. Foster, and J. Hardie, *Two opsin genes from the vetch aphid, Megoura viciae.* Insect Mol Biol, 2000. **9**(2): p. 197-202.
164. Wakakuwa, M., et al., *Physiological basis of phototaxis to near-infrared light in Nephotettix cincticeps.* J Comp Physiol A Neuroethol Sens Neural Behav Physiol, 2014. **200**(6): p. 527-36.
165. Döring, T.F., et al., *Spectral sensitivity of the green photoreceptor of winged pea aphids.* Physiological Entomology, 2011. **36**(4): p. 392-396.
166. Lyko, F., et al., *The honey bee epigenomes: differential methylation of brain DNA in queens and workers.* PLoS Biol, 2010. **8**(11): p. e1000506.