

Chromosome-Aware Phylogenomics of Assassin Bugs (Hemiptera: Reduvioidea) Elucidates Ancient Gene Conflict

Alexander Knyshov ^{*,†,1} Eric R.L. Gordon,² Paul K. Masonick,¹ Stephanie Castillo,¹ Dimitri Forero,³ Rochelle Hoey-Chamberlain,¹ Wei Song Hwang,⁴ Kevin P. Johnson,⁵ Alan R. Lemmon,⁶ Emily Moriarty Lemmon,⁷ Samantha Standring,¹ Junxia Zhang,⁸ and Christiane Weirauch¹

¹Department of Entomology, University of California, Riverside, CA, USA

²Ecology and Evolutionary Biology Department, University of Connecticut, Storrs, CT, USA

³Instituto de Ciencias Naturales, Universidad Nacional de Colombia, Bogota, Colombia

⁴Lee Kong Chian Natural History Museum, National University of Singapore, Queenstown, Singapore

⁵Illinois Natural History Survey, Prairie Research Institute, University of Illinois, Champaign, IL, USA

⁶Department of Scientific Computing, Florida State University, Tallahassee, FL, USA

⁷Department of Biological Science, Florida State University, Tallahassee, FL, USA

⁸Key Laboratory of Zoological Systematics and Application of Hebei Province, Institute of Life Science and Green Development, College of Life Sciences, Hebei University, Baoding, Hebei, China

[†]Present address: Department of Medicine, Boston University, Boston, MA, USA.

***Corresponding author:** E-mail: aknys001@ucr.edu.

Associate editor: Melissa Wilson

Abstract

Though the phylogenetic signal of loci on sex chromosomes can differ from those on autosomes, chromosomal-level genome assemblies for nonvertebrates are still relatively scarce and conservation of chromosomal gene content across deep phylogenetic scales has therefore remained largely unexplored. We here assemble a uniquely large and diverse set of samples (17 anchored hybrid enrichment, 24 RNA-seq, and 70 whole-genome sequencing samples of variable depth) for the medically important assassin bugs (Reduvioidea). We assess the performance of genes based on multiple features (e.g., nucleotide vs. amino acid, nuclear vs. mitochondrial, and autosomal vs. X chromosomal) and employ different methods (concatenation and coalescence analyses) to reconstruct the unresolved phylogeny of this diverse (~7,000 spp.) and old (>180 Ma) group. Our results show that genes on the X chromosome are more likely to have discordant phylogenies than those on autosomes. We find that the X chromosome conflict is driven by high gene substitution rates that impact the accuracy of phylogenetic inference. However, gene tree clustering showed strong conflict even after discounting variable third codon positions. Alternative topologies were not particularly enriched for sex chromosome loci, but spread across the genome. We conclude that binning genes to autosomal or sex chromosomes may result in a more accurate picture of the complex evolutionary history of a clade.

Key words: phylogenomics, gene conflict, gene content, sex chromosome, X chromosome.

Introduction

Advances in DNA sequencing and bioinformatics have made analyses of large-scale phylogenetic matrices much more feasible. While these large data sets have resolved some relationships among major animal lineages (Dunn et al. 2008), many others remain contentious (Rodríguez-Ezpeleta et al. 2007; Smith et al. 2015; Murphy et al. 2021). The large size of phylogenetic matrices mainly limits stochastic errors, which were pervasive in the small-scale Sanger-based data sets of the past (Young and Gillung 2020). In contrast, other types of errors, known as systematic or nonrandom bias (Kapli et al. 2021), appear to remain unaffected by an increase in the amount of sequence data analyzed (Jeffroy et al. 2006; Rodríguez-Ezpeleta et al. 2007; Philippe et al.

2011; Kumar et al. 2012). This realization has promoted the development of new approaches to disentangling conflicting relationships (Arcila et al. 2017; Simon 2020).

Systematic bias encompasses several methodological factors (Simion et al. 2020) and biological phenomena. Among the latter, incomplete lineage sorting (ILS) and hybridization are commonly discussed as underlying causes of gene conflict (Rannala et al. 2020). A disagreement between the signal of nuclear and mitochondrial loci, which may or may not be associated with above-mentioned phenomena and is termed mitonuclear discordance (Linnen and Farrell 2007; Toews and Brelsford 2012), can also be a part of systematic bias. Another source of phylogenetic conflict, chromosomal linkage of nuclear loci, has so far received much less attention (Fontaine et al. 2015; Li et al.

© The Author(s) 2023. Published by Oxford University Press on behalf of Society for Molecular Biology and Evolution.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Open Access

2019). Recent advances in genome sequencing now allow for the linkage of largely complete but discontinuous genome assemblies into chromosomal scaffolds (Dudchenko et al. 2017; Yamaguchi et al. 2021). Probing the mixed phylogenetic signal across genomes in phylogenomic analyses has uncovered that gene features such as chromosomal linkage and GC content may be predictive of phylogenetic signal, at least in certain mammal lineages (Li et al. 2019).

Assessing the chromosomal linkage of phylogenetic markers can potentially help explain observed systematic gene conflict. Sex chromosome genes often differ from autosomal genes in their evolutionary rate (Wilson and Makova 2009a, 2009b). A common pattern termed the “faster X effect” is an elevated substitution rate of coding loci on the sex chromosome in the homozygous sex (i.e., X or Z; also called “hemizygous” sex chromosomes for brevity, herein) compared with autosomes (Mank et al. 2007; Meisel and Connallon 2013; Oyler-McCance et al. 2015). Hypotheses explaining this phenomenon focus on disparities in selection, population size, and expression between the sex chromosomes and autosomes (Meisel and Connallon 2013). In contrast, other factors, such as lower rates of mitotic division and recombination, may promote a relatively slower rate of change for X-linked loci. In cases of extensive ancient hybridization, a lower rate of recombination can enrich portions of the homozygous sex chromosome for loci that more accurately reflect the speciation history of a clade (Li et al. 2019). The relative importance of these conflicting pressures differs across taxa (Xu et al. 2012) but in most vertebrates studied, the sex chromosome in the homozygous sex appears to have a faster evolutionary rate than autosomes. Thus, loci located on the sex chromosome may be better suited for estimating shallower divergences than loci found on autosomes.

Although the X chromosome was observed for the first time in the true bug *Pyrrhocoris apterus* in 1891 (Paliulis et al. 2023), the genomic study of X chromosomes in invertebrates has lagged behind vertebrates. Relatively few chromosomal-level genome assemblies are available to assess the prevalence of a faster X effect across taxa, and the conservation of chromosomal gene content across deep phylogenetic scales is largely unknown. Among the existing studies on invertebrates, some (e.g., those on moths, spiders, and Sternorrhyncha) have shown the hemizygous sex chromosome with symptoms of the faster X effect (Sackton et al. 2014; Bechsgaard et al. 2019; Li et al. 2019), while in others, the hemizygous chromosome evolves at the same or slower rate than the autosomes [beetles (Whittle et al. 2020) and stick insects (Parker et al. 2022)]. Studies on *Drosophila* have been conflicting, with only some supporting a faster X effect, which is generally weak (Betancourt et al. 2002; Counterman et al. 2004; Thornton et al. 2006; Hu et al. 2013). In damselflies with rampant hybridization, the X chromosome is particularly resistant to introgression (Swaegers et al. 2022). Fewer studies have assessed the depth of chromosome gene content conservation in invertebrates. In Diptera, there are

major differences in the number of X-linked genes across lineages, driven by the convergent evolution of paternal genome elimination (Anderson et al. 2022). However, other studies have observed a relatively high level of conservation of gene content extending across the ordinal level (Meisel et al. 2019; Li et al. 2020, 2022), with up to 25% of genes shared across more than 400 My divergence in some cases. Long-term conservation of gene content is necessary to observe any concerted effect of the evolution of loci linked to sex chromosomes at deep phylogenetic scales.

An ideal group to further the study of the X chromosome-related evolutionary processes is Reduvioidea, the assassin bugs and relatives. This group comprises one of the most speciose lineages of Heteroptera (Schuh and Weirauch 2020) and is distributed worldwide (Maldonado Capriles 1990). Assassin bugs are predominantly predatory and display a great diversity of morphological and behavioral specializations (Weirauch et al. 2014). The group comprises two families, the species-poor Pachynomidae (ca. 30 spp.; 2 subfamilies) and the diverse Reduviidae (ca. 7,000 spp.; 24 subfamilies), with the latter including the blood-feeding and medically relevant kissing bugs (Triatominae), the vectors of Chagas disease. Despite being one of the largest superfamilies of true bugs and drawing the attention of medical entomologists, the evolutionary history of Reduvioidea is understudied. Phylogenetic hypotheses across this lineage include a morphological study (Weirauch 2008), Sanger-sequencing-based analyses (Weirauch and Munro 2009; Hwang and Weirauch 2012), as well as a phylogenomic study with limited taxon sampling (Zhang, Gordon, et al. 2016). Only the analysis of Hwang and Weirauch (2012) sampled Reduviidae relatively comprehensively and recovered many subfamilies with high support, but largely failed to resolve intersubfamilial relationships. In contrast, the sole data-rich (370 loci) analysis to date (Zhang, Gordon, et al. 2016) sampled only 14 of the 26 reduvioid subfamilies. This analysis detected phylogenetic conflict, manifested in differences between concatenation- and coalescence-based analyses, demanding further investigation. To date, phylogenetic conflict across this old group [>180 Ma; (Johnson et al. 2018)] has not been investigated, partly because genome-scale data for Reduvioidea have remained scarce and the only available chromosome-level annotated assemblies originate from two closely related species of Triatominae (Mesquita et al. 2015; Liu et al. 2019).

Here, we investigate the phylogenetic conflict among loci through an assessment of their sex chromosomal linkages while reconstructing the evolutionary history of Reduvioidea based on a large phylogenomic data set (2,286 loci; 23 of the 26 subfamilies). We sequenced 84 species of Reduviidae using hybrid capture (Lemmon et al. 2012) and genome skimming (Zhang et al. 2019) approaches. These two data types were combined with existing RNA-seq and reference genomes in a single streamlined pipeline. Relying on available chromosome-level

genomic assemblies, we interpolated the linkage of loci in other taxa and compared the phylogenetic signal of the X chromosome with that of autosomes. Going beyond auto-sex discordance, we investigated and attempted to interpret results in the context of recent advances in gene conflict interrogation by conducting nucleotide (NT)- and amino acid (AA)-based analyses, comparing concatenation and coalescence-based analyses, and discerning clusters of genes with common phylogenetic signal.

Results

Data set Construction

To assemble the phylogenetic data set, we combined sequences derived from anchored hybrid enrichment (AHE, [Lemmon et al. 2012](#)), RNA-seq, and low-coverage whole-genome sequencing (WGS) approaches, as well as high-quality reference genomes, using an in-house developed bioinformatic pipeline ([fig. 1](#)). We mined protein-coding AHE loci from all data types, resulting in a subdataset (“AHE dataset”) with 111 taxa ([supplementary table S1, Supplementary Material online](#)), 381 loci, and 231,153 positions in the NT matrix (77,051 positions in the AA matrix). For 94 samples with transcriptomic and genomic data, we obtained additional separate orthologous loci using the results of an OrthoMCL analysis ([Gordon 2017](#)). After filtering OrthoMCL-obtained orthogroups based on a number of criteria, this additional data set (“OMCL dataset”) had 1,905 loci, and 1,566,147 NT positions in the NT matrix (522,049 AA positions in the AA matrix). We used the combined AHE + OMCL data set (111 taxa, 2,286 loci, and 1,797,300 NT positions; 599,100 AA positions) for phylogenetic analyses. We split loci based on their linkage (X vs. autosome) in the triatomine reference taxa ([Liu et al. 2019](#)). The autosomal (AU) data set comprised 2,141 loci and 1,690,446 NT positions and the X chromosome subset 145 loci and 106,854 NT positions. We further confirmed the X-linkage of loci across reduviids as described in the “Chromosomal linkage of loci” section. To investigate possible discordance between nuclear and mitochondrial phylogenetic signals, we attempted to obtain mitochondrial genomes (partial or complete) from all data types. Because of missing data distributions, only protein-coding and the two ribosomal mitochondrial genes were selected for downstream analysis, resulting in a mitochondrial data set (“MT dataset”) containing 101 taxa and 16,008 NT positions. For a separate analysis of entire gene content, we analyzed a full set of orthogroups that were detected by OrthoMCL using the presence/absence coding ([Pett et al. 2019](#)). Despite some methodological shortcomings of this analysis (in incomplete genomic assemblies “absence” could mean both gene loss and missing data), it constituted a useful yet computationally efficient approach to utilize the additional available genomic data in an attempt to find phylogenetic signal and X chromosome-associated patterns in gene gain and loss events. The extended

methodology and results of this expanded analysis are available in [supplementary text S1, figs. S4–S6, Supplementary Material online](#).

Phylogenetic Relationships and Conflict

Six primary phylogenomic analyses were conducted to reconstruct the evolutionary history of assassin bugs ([fig. 2](#) and [supplementary fig. S1, Supplementary Material online](#)). We first analyzed the AHE + OMCL data set in concatenation, using either the NT (“NT analysis”) or AA matrix (“AA analysis”), as well as with a coalescence-based approach using NT-based gene trees in Astral (“Ast analysis”). The MT data set was analyzed separately in a concatenation-based framework (“MT analysis”). We then analyzed putatively autosomal (“AU analysis”) and X-linked loci (“X analysis”) separately to identify possible phylogenetic discordance between these data sets ([fig. 2](#)). Our analyses unambiguously recovered monophyly of both Reduioidea (N1; full support in all data sets) and Reduviidae (N2; full support except X analysis). Similarly, Higher Reduviidae [N4, a widely used ([Hwang and Weirauch 2012; Weirauch et al. 2014; Zhang, Gordon, et al. 2016](#)) albeit suboptimal term referring to all assassin bug subfamilies except those closely related to Phymatinae] were fully supported in all analyses while the Phymatine-complex (N3) was not recovered in the X analysis; these two clades represent the deepest split in the phylogeny of Reduviidae. A number of subfamilies were recovered as monophyletic across all (i.e., Hammacerinae, Phymatinae, Peiratinae, Vesciinae, Stenopodainae, and Bactrodinae), or all except the X and/or MT analyses (Holoptilinae, Ectrichodiinae, Triatominae, and Salyavatinae). Consistent with published hypotheses, the large subfamily Reduviinae (>1,100 described spp.) is highly polyphyletic with many taxa currently recognized as reduviines recovered as distantly related lineages. Also corroborating published phylogenies, Emesinae were consistently rendered paraphyletic by Saicinae and Visayanocorinae, and Harpactorinae by Bactrodinae. Few intersubfamilial relationships were recovered with full or high support across all or most analyses. Examples are Hammacerinae as a sister taxon to the remaining Phymatine-complex assassin bugs (all except X analysis with full support) and relationships among Stenopodainae, Triatominae and the *Zelurus* group of Reduviinae (all except MT analysis with full support). With regard to novel hypotheses, “Harpactorinae” + Bactrodinae formed the sister taxon to the small, subcuticulous reduviine *Heteropinus mollis* in all analyses. Also, in all analyses except X and MT, this clade was recovered as the sister taxon to the reduviine genera *Nalata* and *Microlestria*, Epiroderinae, and Phimophorinae. Finally, the large clade (N7) comprising Cetherinae, Chryxinae, Pseudocetherinae, Salyavatinae, and the bulk of genera classified as Reduviinae also generally received high support across our analyses.

Despite many of the nodes receiving full support in the NT-based analysis, we observed considerable conflict between the NT topology and the topologies derived from

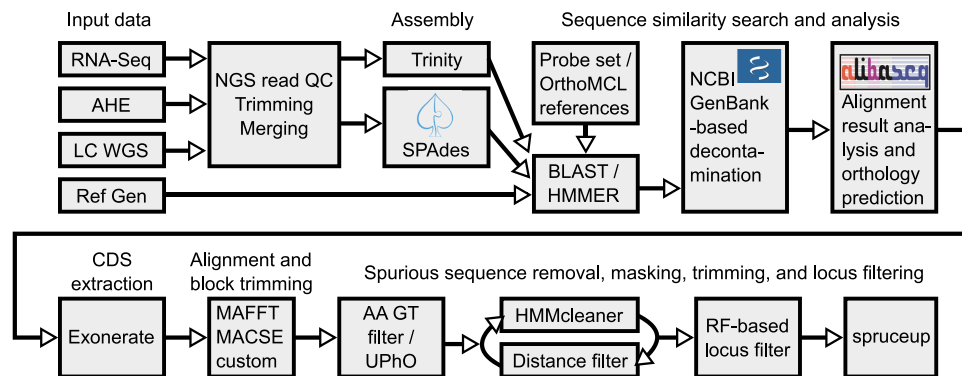


FIG. 1. Schematic of the pipeline to produce the set of loci used in the study. Raw reads were run through quality control and preprocessing, followed by a de novo assembly. Obtained assemblies, together with three reference genome assemblies, were searched for homologs of the AHE and OMCL loci and extracted using *ALiBaSeq*. *Exonerate* was then used to precisely extract CDS, which were aligned and block-trimmed on protein level. AA-based gene trees were used to remove distant outliers, and in case of the OMCL data set, more precisely detect orthologs. Two rounds of segment trimming and further sequence outlier removal were performed, followed by removing loci with extreme RF distance to the species tree, and concluded by a final segment trimming of the entire matrix using *spruceup*.

the AA, Ast, X, and MT analyses. For instance, Peiratinae were the earliest diverging subfamily among the Higher Reduviidae in the NT analysis, but are sister either to the *Psophis* group of Reduviinae in the Astral and MT analysis or to a part of Ectrichodiinae in the X analysis. While such conflict is ubiquitous along the generally less well-supported backbone of Reduviidae, some conflict was also observed within well-supported clades. For example, in the NT and AU analyses, Bactrodinae were recovered as the sister taxon to Apiomerini (represented by *Heniartes* and *Micrauchenus*), while AA, X, and MT showed Bactrodinae as the sister to the Higher Harpactorinae, or to Dicrotelini plus Higher Harpactorinae.

Chromosomal Linkage of Loci

Among the analyses, the conflict between the loci putatively associated with the X chromosome (X analysis) and the autosomal loci (AU analysis) was particularly noticeable. To confirm that X-chromosomal loci of Triatominae are also X-linked in other reduviids, we investigated the conservation of X-located loci in our taxon set. Li et al. (2020) and Mathers et al. (2020) showed that X-linkage of loci was conserved between two closely related blood-feeding reduviids, *Triatoma rubrofasciata* and *Rhodnius prolixus*, as well as that sex chromosomal loci had stable linkage in other hemipteran groups. Since no other chromosome-level assemblies of Reduviidae are available, we used recently published chromosomal assemblies of *Apolygus lucorum*, a distant outgroup of Reduioidea, to verify sex chromosome loci conservation on a deeper evolutionary scale. We followed the methods of Li et al. (2020) to determine the conservation of synteny between *Apolygus* and *Triatoma*. Results of the analysis (fig. 3a) suggested that blocks of genes from the X chromosome of *T. rubrofasciata* reciprocally matched to chromosome 1 (presumably the X chromosome) of *A. lucorum*, with no sex chromosomal blocks matching to any autosomes. This is the first indication that X chromosome

loci might be conserved at least to some extent in their chromosomal linkage across Cimicomorpha [~225 Ma (Johnson et al. 2018)], a much deeper level than what has been previously shown [22, 32, and 57 Ma in different hemipteran lineages (Mathers et al. 2020)].

To confirm that the 145 putatively X-located loci share the same chromosome in other ingroup taxa, we determined locus coverage depth of all loci, and compared it between specimens with identified biological sex. Results showed (fig. 3b) that putatively X-located loci had significantly lower coverage in males compared with females. X-located loci in males also had lower coverage compared with autosomal loci in both sexes. Based on these results, we proceeded to treat the 145 putatively X-associated loci as sex chromosome located loci, with all other loci being treated as autosomal. Since the X chromosome was the only one with conserved gene content and thus with inferrable loci, and for simplicity reasons, we refer to the classification of loci as autosomal or X-linked as “chromosomal linkage” in this paper.

Autosomal and Sex Chromosomal Loci: Phylogenetic Signal and Differences in Properties

Using identified autosomal and sex chromosome loci, we performed phylogenetic reconstructions as well as a gene-by-gene log-likelihood fit difference analysis (Lee and Hugall 2003; Shen et al. 2017) while binning loci according to their chromosomal linkage. We rescaled log-likelihood differences by gene length to obtain per-base phylogenetic signal metric and avoid disproportional influence of long loci. Phylogenetic results showed (fig. 2) considerable topological differences between autosomal and sex chromosomal topology, some of which we discussed above. Additionally, the X-chromosomal tree had a larger total tree length despite having 15 times less loci than autosomal data set, suggesting an on average higher substitution rate of sex loci (supplementary fig. S1,

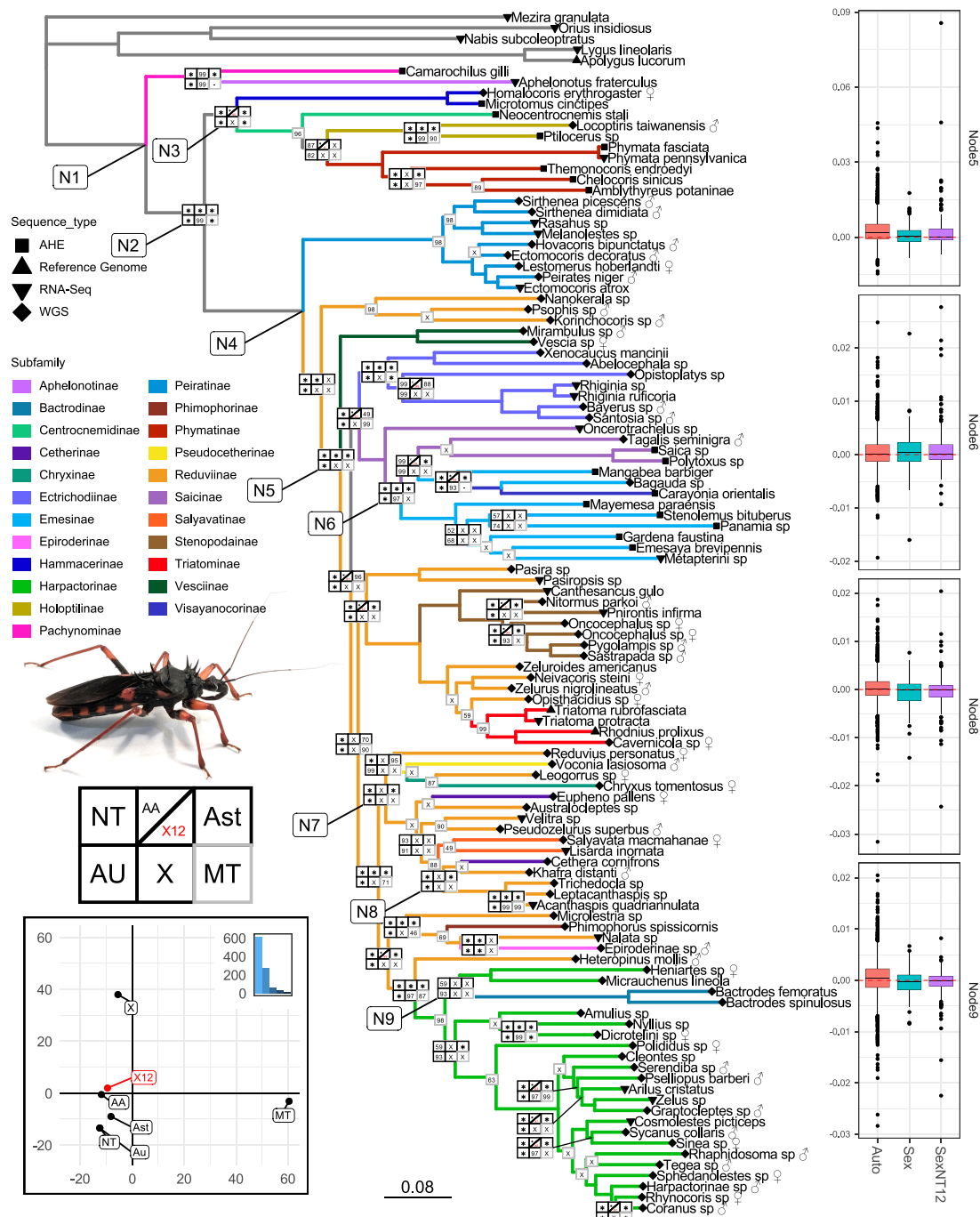


FIG. 2. The topology in center was inferred based on the NT matrix of the combined AHE + OMCL data set. Branches are colored based on the current subfamilial classification of the 106 ingroup taxa. Tip symbols represent data type; sex is annotated for genomic taxa with morphological sex determination. Node boxes represent UFBoot (IQ-TREE-based analyses) or local posterior probability (Astral-based analyses) support values in different analyses (asterisk denotes full support in a given analysis) or include X when a node was not recovered relative to the NT analysis or by a dash "-" when a node could not be recovered due to reduced taxon sampling. Only nodes with a conflict or with <100% support in any analysis have nodal plots shown. When only the mitogenome data set had <100% support, it was shown by itself in the interest of conserving space. From the distribution of the nodal boxes, it is evident that the conflict between analyses primarily concerns the backbone of Reduvidae and several deep divergences within subfamilies, while most recent divergences are supported across all analyses. Additionally, X chromosome and mitochondrial data sets are the two most discordant. Lower left panel represents an RF-based PCoA analysis of topologies obtained from different analyses, only first two PCs shown, inset shows a screeplot of the eigenvalues for each individual PCoA axis. Results of this analysis show that both X chromosome and mitochondrial topologies are drastically different from the rest, however also different from each other. Additionally, the X12 data set (with third codon position removed) is relatively similar to full and autosomal analyses. Box plots on right show likelihood difference between a given node shown in the phylogeny (N5–6, 8–9) and next most likely topology with a different relationship around that node for each gene (positive scores mean that the next most likely tree with an alternative node relationships is more likely for most genes). Results show that, for several of the nodes, correcting X signal (X12) made the likelihood difference distribution closer to that of autosomal genes. Additional nodes labeled (N1–4, N7) are discussed in text. Photo of *Psytta horrida* (Reduviidae) by P.K.M.

Supplementary Material online). Results of the log-likelihood fit difference analysis showed that for several conflicting relationships (fig. 2, N5–6, N8–9), there was a difference between the median autosomal and X likelihood difference between tested topological rearrangements. Consistently with phylogenetic inference results, the median X chromosome likelihood fit difference was either positive but lower than autosomal (favoring same topology but less strongly, less informative compared with autosomal data set, N5); or positive but higher than autosomal (favoring same topology more strongly, potentially could have more decisive impact on combined topology inference, N6); or negative (favoring opposite topology from the autosomal loci, N8–9).

We scored several locus properties in both autosomal and X-located loci to determine if any of the properties differ and could explain the observed gene conflict (fig. 3c and **supplementary fig. S2, Supplementary Material** online). Besides a larger proportion of missing data, which could be associated with lower coverage of X-loci in males, sex loci also had higher substitution rates and levels of saturation. However, selection analyses showed no significant difference in dn/ds ratio between sex chromosome and autosomal genes. All loci selected for phylogenetic inference appear to be under relatively strong purifying selection. The mean GC content of third codon positions (GC3) was similar between X-located and autosomal loci; however, GC3 variation of X was significantly higher. To confirm that a larger proportion of missing data in X-located loci (fig. 3c, **supplementary fig. S2, Supplementary Material** online) does not impact tree length and rate estimates, we checked for and observed no relationships between missing data and either tree length (**supplementary fig. S3a, Supplementary Material** online) or average branch length (**supplementary fig. S3b, Supplementary Material** online).

Correcting the Misleading Sex Chromosomal Signal

Since assessment of both phylogenetic inference and locus properties pointed to possible saturation resulting from a high substitution rate in the presence of stabilizing selection, we hypothesized that the observed phylogenetic signal of X-located loci is artefactual. In an attempt to correct this signal, we excluded third codon positions, which are the most impacted by saturation, and reanalyzed the data sets. The analysis of X-loci with the exclusion of third codon position (X12) yielded a topology much more congruent with that of autosomal loci and very similar to the AA-based tree (fig. 2, PCoA plot, **supplementary fig. S1, Supplementary Material** online). We also reanalyzed the likelihood fit difference for the loci in the corrected data set. Results showed (fig. 2, bar plots) that the corrected sex-chromosomal signal was much more in line with the autosomal signal in the four previously contentious nodes (i.e., median log-likelihood fit difference values between conflicting topologies became more similar between autosomal and X12 assessments).

Gene Conflict Beyond the X Chromosome

Much of the strong gene conflict was alleviated by addressing saturation artifacts of the X-located loci (**supplementary fig. S1, Supplementary Material** online XNT12; e.g., position of Peiratinae), but several relationships remained ambiguous (e.g., position of Bactrodinae). To investigate gene conflict beyond chromosomal linkage, we computed gene tree distances and ran a PCoA analysis of the distances. Seven clusters of genes (or “groves”) based on their topological similarities were identified (fig. 4). On average, clusters contained the same proportion of sex genes as the entire data set (6%). However, the most closely clustered group, grove 6, had fewer X genes (3.3%) while another cluster, grove 3 was uniquely enriched for X genes (13.5%). One of the clusters contained only three genes and was excluded from subsequent investigations. For each of the remaining clusters, we inferred an ML tree, concatenating corresponding loci ($N = 86–766$). Resulting topologies showed (fig. 4; groves 1–6) that each cluster tree had a uniquely different position of one or several contentious taxa or clades in addition to contentious nodes shared between several groves (e.g., position of Bactrodinae, relationships within Emesinae and between Ectrichodiinae and Emesinae, and relationships within clade N7). The mitochondrial phylogeny was also in conflict with each of these groves, most strongly at deep nodes, likely driven by saturation due to the high mutation rate as well as the limited size of the mitogenome.

Discussion

Chromosomal Linkage of Loci Can Explain Part of the Conflict Among Loci

In a large phylogenomic analysis of assassin bugs and relatives (Heteroptera: Reduviidae), we detected substantial phylogenetic conflict among nuclear loci. This conflict can at least in part be explained by loci associated with the X chromosome (fig. 2). Despite only about 6% of the data set being composed of genes associated with the X chromosome, several nodes in the combined analyses received lower support than in autosome-only analyses because of the auto-X signal conflict. Moreover, the X-only tree had a drastically different topology and branch length distribution from other subsets of data, driven in part by a higher substitution rate of X-linked loci. The matrix of X-linked loci also had more missing data, partially due to lower coverage of X-linked loci in males but also possibly impacted by higher genetic distance from the reference. Although larger distances probably did not impact homology searches, they can cause more trimming and masking during filtering steps in the pipeline. The fast substitution rate, however, could not be explained by positive selection, because there was no difference in dn/ds between autosomal and sex chromosomal partitions (fig. 3c). However, sex loci were more saturated compared with autosomal markers, thus allowing us to hypothesize that it is the high substitution rate coupled with the strong purifying selection that causes the discordance between

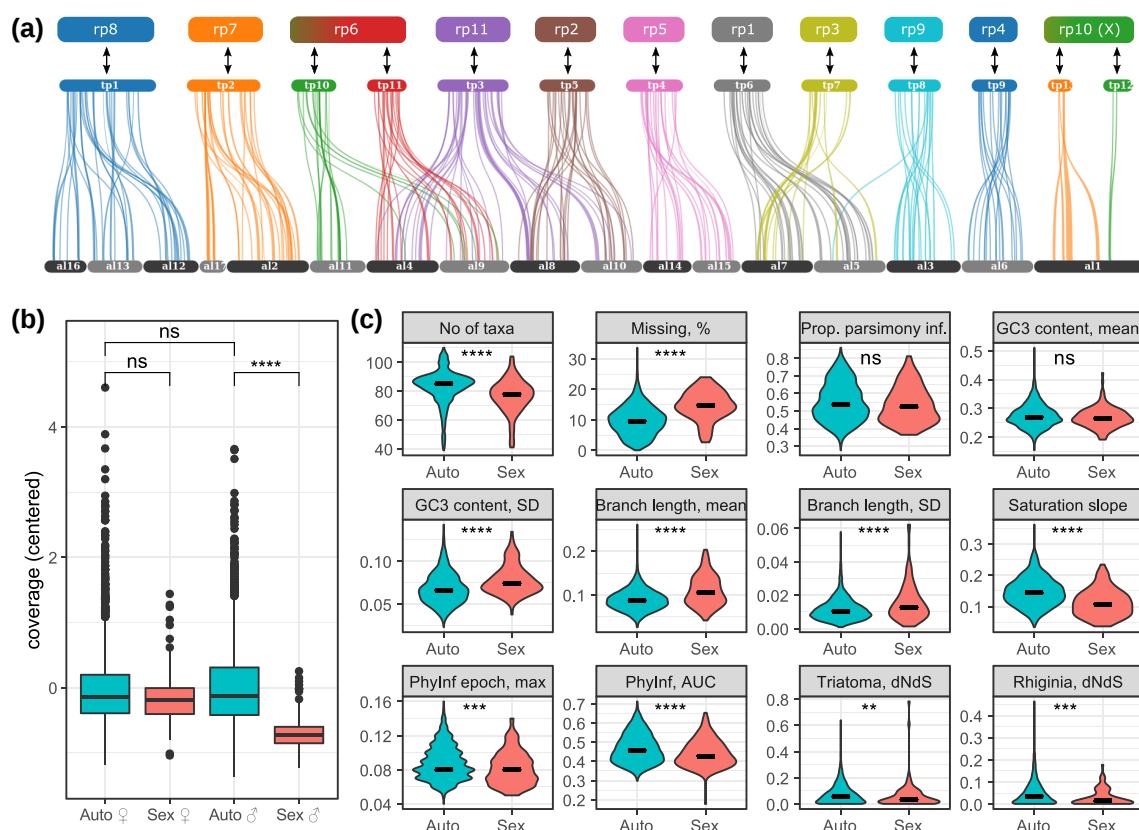


Fig. 3. The top panel (a) shows the results of the gene synteny analysis between *T. rubrofasciata* (tp) and *A. lucorum* (al) chromosomes, with *R. prolixus* chromosomal synteny (rp) shown at the top. The analysis shows fairly conserved synteny between groups of al chromosomes and tp chromosomes. More interestingly, no syntenic blocks for X-loci of tp were found on autosomes of al. Bottom left panel (b) shows the distribution of average locus depth in the combined AHE + OMCL data set on autosomes and X chromosome by morphological sex determined for taxa with sex symbols indicated in figure 2. Results show male X chromosomal loci having significantly lower average coverage than female X-loci or male autosomal loci. Bottom right panel (c) shows the distribution of various locus properties in autosomal and X chromosome loci, outliers were excluded, for the full data, see [supplementary figure S2, Supplementary Material](#) online. Most properties have significantly different distributions, with the exception of proportion of parsimony informative sites and mean GC content of the third codon position.

X-loci and autosomal loci trees. Thus, the X-loci discordant signal in our data is nonphylogenetic. The topology of X12 is highly congruent with the autosomal topology except for a few recalcitrant nodes. This congruence indicates that most of the conflict comes from homoplasy caused by highly saturated rapidly evolving genes and not from any shared strong conflicting signal retained by genes housed on the X chromosome with the same evolutionary history.

Observed Gene Conflict Is Not Solely Explained by X-loci Signal

In most cases, phylogenetic conflict can rarely be attributed to one cause. In the case of phylogenetic conflict in Reduviidae, X-chromosomal linkage and the associated high substitution rate only partially explains the observed discordance among loci. For example, some relationships with conflicting resolutions were retained in both full X and NT12 X analyses. X-linked loci were enriched in the genes comprising grove 3 (fig. 4) and these together with autosomal genes in the grove could represent a group of

loci with the signal of hybridization or ILS across both the autosomal and sex chromosome backgrounds. The mitogenome is one of the most accessible and often-used locus for phylogenies and we sought to obtain as much as possible for future studies which might make use of these data. Additionally, similar to sex chromosomes, the mitogenome has a biased pattern of sex-based inheritance, although biased entirely to maternal inheritance and we were curious to see if there was any concordance of the mitogenome with any other group of loci. The mitogenome is also similar to the X chromosome in that it never experiences recombination (whereas the X chromosome experiences recombination only in females) and thus constitutes a single genomic locus with shared evolutionary history. We did not observe any set of nuclear loci with the same phylogenetic signal as the mitogenome, which even with a relatively limited signal, had strong conflicting resolution to all other examined trees. As is true of many other studies that have observed significant mitonuclear discordance, our results suggest that the phylogenetic background of the mitogenome may differ from that of the nuclear genome. Alternative resolutions seen in other

loci grouped into groves likely reflect the true conflict. As chromosomal-level genomic assemblies become more prevalent, it may be possible to determine the linkage of similar subsets of loci that would reflect explanatory introgression or hybridization events.

Phylogeny of Reduioidea

Our study presents the most data-rich and densely sampled phylogenetic hypotheses for the assassin bugs and relatives to date (111 taxa; 2,286 loci; 23 of the 26 reduvioid subfamilies). Of the three subfamilies missing from our analyses, one is thought to be part of the Phymatine-complex based on morphological data [Elasmodeminae (Carayon et al. 1958)], another rendered Salyavatinae paraphyletic in Sanger-based analyses [Sphaeridopinae (Gordon and Weirauch 2016)], and the last, Manangocorinae are monotypic, only known from the holotype, and morphology suggests that this subfamily may belong to clade N7. Topologies recovered from different data sets in our analyses are similar to but not identical with published hypotheses that attempted to resolve relationships across reduviid subfamilies (Weirauch and Munro 2009; Hwang and Weirauch 2012; Zhang et al. 2016) and to topologies from studies that aimed to uncovering relationships within subfamilies or clades of related subfamilies [Harpactorinae/Bactrodinae (Zhang, Weirauch, et al. 2016); Phymatine-complex (Masonick et al. 2017); Triatominae and relatives (Justi et al. 2014; Kieran et al. 2021); Ectrichodiinae (Forthman and Weirauch 2016)]. While several of the nodes along the backbone of Higher Reduviidae (N4) are unstable across our analyses, many aspects of our hypotheses are well-supported in the NT, AA, Ast, AU, and X12 analyses and can form the basis for a revised subfamily- and tribal-level classification built on diagnosable monophyletic groups. The polyphyly of Reduviinae, first documented by Weirauch (2008) based on morphology and corroborated by Sanger-derived (Hwang and Weirauch 2012) and phylogenomic data sets (Zhang et al. 2016), is exacerbated by recovering *Heteropinus* as a separate lineage and sister to the largely diurnal and vegetation-dwelling “Harpactorinae” + Bactrodinae. An assassin bug classification based on diagnosable clades will require the recognition of several of these distantly related reduvine lineages as separate subfamilies. In contrast, the paraphyly of “Harpactorinae” with respect to Bactrodinae, paraphyly of “Emesinae” and “Saicinae” relative to Visayanocorinae, and the relationships of reduvine taxa with nonreduvine subfamilies contained in clade N7 will require the synonymization of several currently recognized subfamilies. Combined phylogenomic and morphological data sets and analyses have been generated to formalize these changes to the classification (Masonick et al. in preparation; Standring et al. in preparation).

In addition to improving our hypothesis of the evolutionary history of reduvioids, newly sampled taxa in our study could aid in a more robust reconstruction of

genomic patterns in the evolution of blood-feeding in Triatominae. Historically, comparative genomic studies on blood-sucking true bugs have been restricted to few distant independently parasitic taxa, without detailed comparison to the closest nonblood-feeding relatives. However, studying the closest nonblood-feeding outgroups can help more accurately pinpoint the genes involved in the feeding habit transition. For example, the *R. prolixus* genome possesses several odorant-binding protein (OBP) gene clusters that were reconstructed to be uniquely duplicated in Triatominae given the taxon sampling (Mesquita et al. 2015). However, our gene content analyses (supplementary text S1, fig. S6, Supplementary Material online) show that among the discovered OBP clusters only one (RproOBP13) is truly specific to Triatominae, with other OBPs recovered in nonblood-feeding reduviids. This protein along with another that is not exclusive to Triatominae (RproOBP6) had previously been identified as expressed in the antennae of male and female insects (Oliveira et al. 2018) and thus likely plays a role in host-sensing.

Investigation of X Chromosome Evolution Is Applicable to Phylogenomics

Chromosome-level assemblies continue to be scarce and impede progress on understanding of sex chromosome evolution and examination of gene conflict. To gain knowledge more rapidly, phylogenomic data like ours can be combined with chromosomal assemblies to assess conservation of chromosomal loci and shed light on subsequent phylogenetic analyses.

Although the prevalence of a fast X effect is not known across all arthropods, loci linked to X chromosomes are common in phylogenomic studies of arthropods (6% in our data set but a cursory analysis of some other commonly used public locus sets in arthropods yielded similar proportions of X-linked loci, ca. 5–10%). Gene content of X chromosomes appears conserved in some cases up to approximately 200 Ma divergence between arthropod taxa, making gene content inferable for many small-scale phylogenetic studies with only distant relatives with a chromosomal-level assembly. Despite comprising a small proportion of loci, X-linked genes appear to have a considerable influence at least in the case of our data set in which the tree includes rapid radiations and several groups have long branches. X chromosome loci can also have other features which would impact accurate phylogenetic reconstruction [e.g., relaxed purifying selection in aphids (Li et al. 2020) or stick insects (Parker et al. 2022)].

Conclusion

We developed a method to organize four different kinds of heterogeneous genomic data types, achieving extensive taxon sampling to infer a phylogeny of this medically important group of arthropods. We were able to confirm a faster rate of evolution of X-linked loci in this group, which

impacted accurate phylogenetic reconstruction but did not entirely explain other underlying gene conflict. Building upon these findings, we established a robust phylogeny of the group, highlighting contentious nodes with substantial conflict that will be valuable for understanding the evolution of assassin bugs. Furthermore, this approach is applicable to other organisms with only distant relatives with chromosomal-level reference assemblies to generate a more comprehensive understanding of sex chromosome and clade-specific evolution.

Materials and Methods

Anchored Hybrid Enrichment

The AHE data set consisted of 17 samples ([supplementary table S1, Supplementary Material](#) online). DNA from the samples was extracted using a Qiagen DNeasy Blood and Tissue kit. Library preparation, hybrid enrichment, and sequencing followed standard protocols ([Lemmon et al. 2012](#)). A Hemiptera probe set ([Dietrich et al. 2017](#)) was used for hybrid enrichment. Samples were sequenced on an Illumina HiSeq 2500 platform. Reads were processed with *clumpify* (BBMap v38.86 package) and deduplicated, then trimmed with *Trimmomatic* v0.36 ([Bolger et al. 2014](#)), merged with *bbmerge* (BBMap v38.86 package), error-corrected and assembled with *SPAdes* v3.12.0 ([Bankevich et al. 2012](#)).

RNA-seq

RNA-seq data for the present study were originally obtained as part of [Zhang, Gordon, et al. \(2016\)](#). Briefly, RNA was extracted from the head and thorax, or full body of the specimens, and cDNA libraries were prepared and sequenced at the W.M. Keck Center (University of Illinois) and AITBiotech PTE LTD (Singapore) using an Illumina HiSeq platform and paired-end 100-bp chemistry. Reads were trimmed using *Trimmomatic* ([Bolger et al. 2014](#)) and assembled with *Trinity* ([Haas et al. 2013](#)) as part of [Gordon \(2017\)](#). For the four outgroup taxa, belonging to other families of Cimicomorpha as well as to the sister infraorder Pentatomomorpha, the original assemblies from [Zhang, Gordon, et al. \(2016\)](#) were used.

Whole-Genome Sequencing

DNA from the samples was extracted using either a Qiagen DNeasy Blood and Tissue kit or using a combination protocol of Qiagen Qiaquick and DNeasy kits ([Knyshev et al. 2018](#)). Five samples (*Dicrotelini* sp., *Heniarthes* sp., *Cleontes* sp., *Amulius* sp., and *Bactrodes femoratus*) were subjected to a library prep protocol as in [Knyshev et al. \(2018\)](#) and sequenced on a HiSeq X lane. DNA extracts of the remaining samples were used to prepare libraries as in [Lemmon et al. \(2012\)](#), and the samples were sequenced on several NovaSeq6000 S4 lanes. Reads processing and assembly methods were the same as for the AHE samples.

Previously Available Chromosome-Level Assemblies

We also included chromosome-level assemblies of the two available reduviids, *R. prolixus* ([Mesquita et al. 2015](#)) and *T. rubrofasciata* ([Liu et al. 2019](#)), and the assembly of *A. lucorum* ([Liu et al. 2021](#)) as an outgroup. For *R. prolixus*, we used the chromosomal assembly provided by DNA Zoo ([Dudchenko et al. 2017](#)) and also carried the unmasked original assembly ([Mesquita et al. 2015](#)) and predicted proteins (RproC3.3) through the initial steps of the pipeline to assess its performance as well as to compare the completeness of these data with the chromosome-level assembly.

AHE Data Set Construction

Heteroptera bait regions of the Paraneoptera AHE kit used for enrichment were reprocessed to remove overlapping regions and combine exons of the same genes using a taxon used for probe design [similar approach was employed by [Van Dam et al. \(2020\)](#)]. The coding sequence of the resulting regions was determined and translated AA sequences for each locus were obtained. The *Trinity*-based assembly of the taxon that was used for probe design (*Arilus cristatus*) was used as the reference. A total of 397 gene regions, representing the original 478 AHE bait regions, were selected for downstream analyses.

Discontinuous megablast (*dc-megablast*) from the *Blast* package ([Altschul et al. 1990](#); [Camacho et al. 2009](#)) was used to search for the bait sequences in the RNA-seq (24 samples), WGS (67 samples), hybrid capture (17 samples), and chromosomal-level assemblies (3 samples). Due to noticeable contamination in some WGS and AHE samples (as evident by different GC content and long gene tree branches), the matched contigs were pre-screened for nonarthropod sequences via *megablast* against the NCBI database. Subsequently, noncontaminant matched contigs were searched against the reference assembly of *A. cristatus* using *dc-megablast* for the reciprocal best hit check. *ALiBaSeq* ([Knyshev et al. 2021](#)) was used to parse the *Blast* results, perform reciprocal best hit check, stitch contigs of RNA-seq and WGS samples belonging to the same bait, and compile the resulting locus by locus FASTA files. We set *ALiBaSeq* to recover both matched and internal unmatched sequence regions (mode -x b) to allow for a more accurate CDS sequence extraction in the next step.

Exonerate v2.2.0 ([Slater and Birney 2005](#)) was used together with protein references of *A. cristatus* to accurately excise exons of the recovered sequences. Obtained CDS sequences were aligned on protein level with *MAFFT* ([Katoh and Standley 2013](#)). Sequences were trimmed on AA level. *MACSE* v2.03 ([Ranwez et al. 2011](#)) was used to transfer protein alignment and trimming onto NT sequences. Following trimming, gene trees were reconstructed based on protein sequences using *RAXML* v8.2.12 ([Stamatakis 2014](#)) with the LG + G model. Excessively long branches (over 50 times greater than the mean edge length) representing likely paralogs were removed using a custom R script. This and subsequent filtering R scripts were based

on functions from *APE* (Paradis and Schliep 2019), *Phangorn* (Schliep 2011), and *SeqinR* (Charif and Lobry 2007) R packages. The filtered alignments were then carried out two times through *HmmCleaner* v0.180750 (Di Franco et al. 2019) filtering, followed by a custom-made sequence distance filter to remove outliers. Loci shorter than 50AA (150 bp) were discarded. Then NT alignments were used to reconstruct gene trees using *RAXML* with the “GTRGAMMA” model, which then had nodes below 33% Rapid Bootstrap Support (RBS) (Stamatakis et al. 2008) collapsed. Obtained trees were used to detect and remove cross-contamination as well as to filter weighted Robinson–Foulds (RF) distance (Robinson and Foulds 1981) outlier loci. To remove cross-contaminant taxa in a locus, suspected contaminant groups of taxa were first discovered based on an unexpectedly small pairwise GTR distance of <0.01 in a locus given a concatenation-based GTR distance of over 0.2. Comparison with the concatenation tree helps to preserve similar sequences for several congeneric taxa in our analysis. Then we try to rescue at least one of the suspect sequences, if it can be identified as a donor, and in this way, our script improves on some existing pipelines for contamination screening in AHE data sets (Owen et al. 2022). Since gene tree error with respect to the concatenated tree can be large and a simple check of ML-based distance can be misleading, we instead try to drop each terminal and, if the RF distance of the resulting gene tree to the concatenation tree was reduced, the sequence is kept. If two tests give identical RF distance, both sequences are removed. Obtained filtered alignments were filtered with *spruceup* (Borowiec 2019). Concatenated NT and AA analyses were conducted in *IQ-TREE* v1.7-beta9 (Nguyen et al. 2015). Custom scripts, available at <https://github.com/AlexKnyshov/PLS>, were used to calculate per locus difference in log-likelihood for each Nearest Neighbor Interchange (Lee and Huggall 2003; Shen et al. 2017). Loci with too many outlier nodes and outlier clades in particular loci were screened for homology errors and some removed.

OrthoMCL-based Data Set Construction

We started with 32,675 orthologous clusters produced by Gordon (2017). Briefly, 20 reduviid transcriptomes from Zhang, Gordon, et al. (2016) along with the *Rhodnius* coding sequence were used as the input. The longest open reading frame was predicted using *Transdecoder*, and resulting sequences were supplied to *OrthoMCL* (Li et al. 2003).

We queried the clusters in which at least ten species have sequences and all orthologs in a cluster are single copy. We then only retained clusters passing the minimum length threshold (≥ 100 AA) and lacking extraordinary mean pairwise distance (as those likely represented erroneous clustering). AHE genes were then found via Blast and removed to avoid duplication of the AHE data set. This filtering resulted in 2,296 clusters. Proteins for each orthogroup were aligned, trimmed from flanks, and HMMER profiles were generated. All sample assemblies

were translated into six reading frames and searched for homologs using the profiles obtained earlier. Due to computational difficulties with performing an HMMER search on chromosome-level scaffolds of the three chromosome-level assemblies, *TBlastN* search was used instead for these samples. The most complete sample from the initial OrthoMCL set (Pasir = *Pasiropsis* sp.) was selected as a reference taxon for the RBH check. *ALiBaSeq* (Knyshov et al. 2021) was used to parse search results parsing, perform RBH check and sequence extraction.

Subsequent sequence processing was largely similar to AHE data set. In order to mitigate elevated risks of paralog incorporation, in this data set *ALiBaSeq* was used to pull out closely matching suboptimal hits in addition to the best hit, protein-based gene trees were reconstructed, and *UPhO* (Ballesteros and Hormiga 2016) was used to infer final orthologs.

Mitochondrial Data Set

Using available reduviid references from GenBank, mitochondrial contigs were searched for in the whole read assemblies from above using *dc-megablast*. Recovered partial sequences were improved upon using a combination of *NovoPlasty* (Dierckxsens et al. 2017), *Mitobim* (Hahn et al. 2013), read mapping using *BBMap*, and final curation and annotation in *Geneious* (Kearse et al. 2012). As there were some rearrangements in tRNA gene order, we used a concatenated alignment of the 13 protein-encoding genes and the ribosomal loci 12S rRNA and 16S rRNA rather than a global mitogenome alignment.

Assessment of Locus Properties

We used *AMAS* (Borowiec 2016) and a modified *gene_stats.R* (Borowiec et al. 2015) to assess general locus stats such as length, number of taxa, and proportion of parsimony informative sites. *RAXML* trees, computed for summary coalescence analyses, were used to assess the average BS, average branch length, and saturation. Custom R scripts and *APE* package were used to compute GC content mean and variance across taxa per each codon position.

Final *IQ-TREE* analyses were used to record per-site rates of all genes. These rates were used with the *site_summer* function of *PhyInformR* package (Dornburg et al. 2016) to reconstruct a phylogenetic informativeness curve (Townsend 2007). We sampled informativeness at 0.01 intervals as in the web application *PhyDesign* (López-Giráldez and Townsend 2011) and contrary to the default behavior of *PhyInformR* to only sample at nodes, as the latter led to aberrations in curve smoothing for some loci. We then recorded the epoch (normalized interval along tree height) of the max informativeness, as well as computed total informativeness as area under the curve.

In order to evaluate the selection strength on the loci, we used three pairs of closely related species across the reduviid tree: *Phymata pennsylvanica* and *Phymata fasciata* (only AHE loci), *Rhiginia ruficoria* and *Rhiginia* sp., and *Triatoma protracta* and *T. rubrofasciata*. For each pair,

we assessed the proportion of synonymous to nonsynonymous substitutions using the CODEML subprogram of PAML v4.9 (Yang 2007). Due to the small number of loci available for *Phymata*, only the latter two comparisons were used, but all results were provided in the [Supplementary Material](#) online.

Based on the chromosomal linkage of the loci determined by ALiBaSeq during data set construction, we built a correspondence table between sex chromosome genes of the three taxa with chromosome-level assemblies: *Rhodnius*, *Triatoma*, and *Apolygus*. As the relationship between *Rhodnius* and *Triatoma* chromosomes was already established in Mathers et al. (2020), we used the same approach of employing BlastP and MCScanX (Wang et al. 2012) to infer chromosomal homology between *Triatoma* and *Apolygus*. BlastP e-value cutoff was set to $1e-10$, match size in MCScanX was lowered from default 5 to 2, given the higher divergence between the species. The results were then visualized with SynVisio (Bandi and Gutwin 2022). To check coverage difference or lack thereof of putative autosomal and sex X-loci (as determined for *Rhodnius* and *Triatoma*) in other samples, we used BMap to map reads of each sample to each locus and record average read depth using the covstats parameter.

Phylogenetic Analyses

Phylogenetic analyses were carried out on the separate data sets (AHE and OMCL), as well as on the combined data set. For each of the data sets, concatenated matrices of NT and AA data were produced. NT supermatrices were analyzed with IQ-TREE using standard models with partition finding for the AHE loci only (Kalyaanamoorthy et al. 2017). OMCL loci were too numerous to conduct a partition finding in a reasonable time, this similarly precluded us from trying the codon models on these data. AA supermatrices were also analyzed with IQ-TREE, using standard AA models and partition finding, as well as using posterior mean site frequency [PMSF (Wang et al. 2018)] analysis with LG + C20 model. Similar to the NT analyses, partition finding was not used on the OMCL loci due to computational difficulties. As PMSF results largely replicated standard AA search, we only show the latter, with the former available in the [Supplementary Material](#). Additionally, we reconstructed a final set of gene trees for all loci using RAXML, collapsed branches below 33% RBS. These gene trees were grouped into three data sets (AHE, OMCL, and both combined) and analyzed in Astral v5.6.3 (Zhang et al. 2018). Autosome-only and sex-only loci were analyzed only on NT level, with reusing models for loci from the combine (AU + X) data, but for NT12 data set with removed third codon position, the model testing was performed de novo as outlined above.

Conflict Interrogation

To assess congruence and conflict among the produced topologies (fig. 2 and [supplementary fig. S1](#), [Supplementary Material](#) online), we used the R package

Treespace (Jombart et al. 2017) to conduct a PCoA based on RF distances calculated between trees. We used log-likelihood fit difference as described above to compute per locus phylogenetic signal, which allowed us to assess concatenation-based conflict between loci. We summed the likelihood of autosomal and sex loci to gauge the degree and direction of phylogenetic conflict between sex and nonsex-linked loci.

To investigate gene conflict beyond chromosomal linkage, we filtered out the third codon position from the alignments, computed gene trees in RAXML as above, and collapsed nodes with <50% RBS. Pairwise weighted RF distances were calculated between all gene trees and a PCoA analysis of the distances was conducted in Treespace. Genes were grouped into seven clusters based on the results of the PCoA, full alignments (all codon positions) of each group were concatenated and analyzed in IQ-TREE to produce a concatenation-based topology for each grove.

Supplementary Material

[Supplementary data](#) are available at *Molecular Biology and Evolution* online.

Acknowledgments

This research was funded by NSF DEB #1655769 “The assassin’s tale: evolutionary history of the Reduvioidea, a diverse clade of predatory and hematophagous insects” to C.W. and NSF DEB #1239788 “Phylogenomics and morphology of the hemipteroid insect orders” to K.P.J. Fieldwork was partially supported by the National Geographic grant “Predators become prey: untangling spider-associated behavior and morphology within true bugs (Heteroptera)” to C.W. and S.S. We also thank numerous colleagues for loaning or donating specimens for this project. We acknowledge anonymous reviewers and editors for providing insightful comments.

Data Availability

Raw sequence data generated for this study are available at the NCBI short-read archive, with the majority of experiments submitted under BioProject PRJNA704648. Detailed information on accession numbers is available in [supplementary table S1](#), [Supplementary Material](#) online. Sequence assemblies, locus alignments, and other [Supplementary Material](#) online are available from Zenodo: <https://doi.org/10.5281/zenodo.7726313>. Code used to process the data is available at GitHub: https://github.com/AlexKnyshov/reduvioid_phylogenomic_pipeline.

References

- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol.* **215**:403–410.
- Anderson N, Jaron KS, Hodson CN, Cougar MB, Ševčík J, Weinstein B, Pirro S, Ross L, Roy SW. 2022. Gene-rich X chromosomes

- implicate intragenomic conflict in the evolution of bizarre genetic systems. *Proc Natl Acad Sci U S A*. **119**:e2122580119.
- Arcila D, Ortí G, Vari R, Armbruster JW, Stiasny MJ, Ko KD, Sabaj MH, Lundberg J, Revell LJ, Betancur-R R. 2017. Genome-wide interrogation advances resolution of recalcitrant groups in the tree of life. *Nat Ecol Evol*. **1**:20.
- Ballesteros JA, Hormiga G. 2016. A new orthology assessment method for phylogenomic data: unrooted phylogenetic orthology. *Mol Biol Evol*. **33**:2481.
- Bandi V, Gutwin C. 2022. Interactive exploration of genomic conservation. [Last accessed 2020 Sep 29]. Available from: <https://openreview.net/pdf?id=7-CSVJWbnl>
- Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Pribelski AD, et al. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol*. **19**:455–477.
- Bechsgaard J, Schou MF, Vanthournout B, Hendrickx F, Knudsen B, Settepani V, Schierup MH, Bilde T. 2019. Evidence for faster X chromosome evolution in spiders. *Mol Biol Evol*. **36**:1281–1293.
- Betancourt AJ, Presgraves DC, Swanson WJ. 2002. A test for faster X evolution in *Drosophila*. *Mol Biol Evol*. **19**:1816–1819.
- Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. **30**:2114–2120.
- Borowiec ML. 2016. AMAS: a fast tool for alignment manipulation and computing of summary statistics. *PeerJ*. **4**:e1660.
- Borowiec ML. 2019. Spruceup: fast and flexible identification, visualization, and removal of outliers from large multiple sequence alignments. *J Open Source Softw*. **4**:1635.
- Borowiec ML, Lee EK, Chiu JC, Plachetzki DC. 2015. Extracting phylogenetic signal and accounting for bias in whole-genome data sets supports the Ctenophora as sister to remaining Metazoa. *BMC Genomics*. **16**:987.
- Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics*. **10**:421.
- Carayon J, Usinger RL, Wygodzinsky P. 1958. Notes on the higher classification of the Reduviidae, with the description of a new tribe of the Phymatinae (Hemiptera-Heteroptera). *Rev Zool Bot Afr*. **57**:256–281.
- Charif D, Lobry JR. 2007. SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. In: Bastolla U, Porto M, Roman HE, Vendruscolo M, editors. *Structural approaches to sequence evolution: molecules, networks, populations*. Berlin, Heidelberg: Springer. p. 207–232.
- Counterman BA, Ortíz-Barrientos D, Noor MAF. 2004. Using comparative genomic data to test for fast-X evolution. *Evolution*. **58**:656–660.
- Dierckxsens N, Mardulyn P, Smits G. 2017. NOVOPlasty: de novo assembly of organelle genomes from whole genome data. *Nucleic Acids Res*. **45**:e18.
- Dietrich CH, Allen JM, Lemmon AR, Lemmon EM, Takiya DM, Evangelista O, Walden KKO, Grady PGS, Johnson KP. 2017. Anchored hybrid enrichment-based phylogenomics of leafhoppers and treehoppers (Hemiptera: Cicadomorpha: Membracidae). *Insect Syst Divers*. **1**:57–72.
- Di Franco A, Poujol R, Baurain D, Philippe H. 2019. Evaluating the usefulness of alignment filtering methods to reduce the impact of errors on evolutionary inferences. *BMC Evol Biol*. **19**:21.
- Dornburg A, Fisk JN, Tamagnan J, Townsend JP. 2016. PhyInformR: phylogenetic experimental design and phylogenomic data exploration in R. *BMC Evol Biol*. **16**:262.
- Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, Shamim MS, Machol I, Lander ES, Aiden AP, et al. 2017. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science*. **356**:92–95.
- Dunn CW, Hejnol A, Matus DQ, Pang K, Browne WE, Smith SA, Seaver E, Rouse GW, Obst M, Edgecombe GD, et al. 2008. Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature*. **452**:745–749.
- Fontaine MC, Pease JB, Steele A, Waterhouse RM, Neafsey DE, Sharakhov IV, Jiang X, Hall AB, Catteruccia F, Kakani E, et al. 2015. Extensive introgression in a malaria vector species complex revealed by phylogenomics. *Science*. **347**:1258524.
- Forthman M, Weirauch C. 2016. Phylogenetics and biogeography of the endemic Madagascan millipede assassin bugs (Hemiptera: Reduviidae: Ectrichodiinae). *Mol Phylogenet Evol*. **100**:219–233.
- Gordon ERL. 2017. Natural history, systematics, and taxonomy of the termite assassin bugs (Reduviidae: Salyavatinae), host associations, salivary protein evolution and bacterial symbionts of kissing bugs (Reduviidae: Triatominae) and evolutionary analysis of Microbiota of miroidea and largidae. Available from: <https://search.proquest.com/openview/aa46a799a72bb11306830f7870e0fa1d/1?pq-origsite=gscholar&cbl=18750>
- Gordon ERL, Weirauch C. 2016. Efficient capture of natural history data reveals prey conservatism of cryptic termite predators. *Mol Phylogenet Evol*. **94**:65–73.
- Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, et al. 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc*. **8**:1494–1512.
- Hahn C, Bachmann L, Chevreux B. 2013. Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads—a baiting and iterative mapping approach. *Nucleic Acids Res*. **41**:e129.
- Hu TT, Eisen MB, Thornton KR, Andolfatto P. 2013. A second-generation assembly of the *Drosophila simulans* genome provides new insights into patterns of lineage-specific divergence. *Genome Res*. **23**:89–98.
- Hwang WS, Weirauch C. 2012. Evolutionary history of assassin bugs (insecta: hemiptera: Reduviidae): insights from divergence dating and ancestral state reconstruction. *PLoS One*. **7**:e45523.
- Jeffroy O, Brinkmann H, Delsuc F, Philippe H. 2006. Phylogenomics: the beginning of incongruence? *Trends Genet*. **22**:225–231.
- Johnson KP, Dietrich CH, Friedrich F, Beutel RG, Wipfler B, Peters RS, Allen JM, Petersen M, Donath A, Walden KKO, et al. 2018. Phylogenomics and the evolution of hemipteroid insects. *Proc Natl Acad Sci U S A*. **115**:12775–12780.
- Jombart T, Kendall M, Almagro-García J, Colijn C. 2017. treespace: statistical exploration of landscapes of phylogenetic trees. *Mol Ecol Resour*. **17**:1385–1392.
- Justi SA, Russo CAM, Mallet JR, Obara MT, Galvão C. 2014. Molecular phylogeny of Triatomini (Hemiptera: Reduviidae: Triatominae). *Parasit Vectors*. **7**:149.
- Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods*. **14**:587–589.
- Kapli P, Flouri T, Telford MJ. 2021. Systematic errors in phylogenetic trees. *Curr Biol*. **31**:R59–R64.
- Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*. **30**:772–780.
- Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C, et al. 2012. Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*. **28**:1647–1649.
- Kieran TJ, Gordon ERL, Zaldívar-Riverón A, Ibarra-Cerdeña CN, Glenn TC, Weirauch C. 2021. Ultraconserved elements reconstruct the evolution of Chagas disease-vectoring kissing bugs (Reduviidae: Triatominae). *Syst Entomol*. **46**:725–740.
- Knyshov A, Gordon ERL, Weirauch C. 2018. Cost-efficient high throughput capture of museum arthropod specimen DNA using PCR-generated baits. *Methods Ecol Evol*. **10**:841–852.
- Knyshov A, Gordon ERL, Weirauch C. 2021. New alignment-based sequence extraction software (ALiBaSeq) and its utility for deep level phylogenetics. *PeerJ*. **9**:e11019.
- Kumar S, Filipski AJ, Battistuzzi FU, Kosakovsky Pond SL, Tamura K. 2012. Statistics and truth in phylogenomics. *Mol Biol Evol*. **29**:457–472.

- Lee MSY, Hugall AF. 2003. Partitioned likelihood support and the evaluation of data set conflict. *Syst Biol.* **52**:15–22.
- Lemmon AR, Emme SA, Lemmon EM. 2012. Anchored hybrid enrichment for massively high-throughput phylogenomics. *Syst Biol.* **61**:727–744.
- Li G, Figueiró HV, Eizirik E, Murphy WJ. 2019. Recombination-aware phylogenomics reveals the structured genomic landscape of hybridizing cat species. *Mol Biol Evol.* **36**:2111–2126.
- Li L, Stoeckert CJ Jr, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13**: 2178–2189.
- Li X, Mank JE, Ban L. 2022. Grasshopper genome reveals long-term conservation of the X chromosome and temporal variation in X chromosome evolution. *bioRxiv* [Internet]:2022.09.08.507201. Available from: <https://www.biorxiv.org/content/10.1101/2022.09.08.507201.abstract>
- Li Y, Zhang B, Moran NA. 2020. The aphid X chromosome is a dangerous place for functionally important genes: diverse evolution of hemipteran genomes based on chromosome-level assemblies. *Mol Biol Evol.* **37**:2357–2368.
- Linnen CR, Farrell BD. 2007. Mitonuclear discordance is caused by rampant mitochondrial introgression in Neodiprion (Hymenoptera: Diprionidae) sawflies. *Evolution* **61**:1417–1438.
- Liu Q, Guo Y, Zhang Y, Hu W, Li Y, Zhu D, Zhou Z, Wu J, Chen N, Zhou X-N. 2019. A chromosomal-level genome assembly for the insect vector for Chagas disease, *Triatoma rubrofasciata*. *Gigascience* **8**:giz089.
- Liu Y, Liu H, Wang H, Huang T, Liu B, Yang B, Yin L, Li B, Zhang Y, Zhang S, et al. 2021. *Apolygus lucorum* genome provides insights into omnivorousness and mesophyll feeding. *Mol Ecol Resour.* **21**: 287–300.
- López-Giráldez F, Townsend JP. 2011. PhyDesign: an online application for profiling phylogenetic informativeness. *BMC Evol Biol.* **11**: 152.
- Maldonado Capriles J. 1990. *Systematic catalogue of the Reduviidae of the World (Insecta: Heteroptera)*. Mayaguez, Puerto Rico: University of Puerto Rico.
- Mank JE, Axelsson E, Ellegren H. 2007. Fast-X on the Z: rapid evolution of sex-linked genes in birds. *Genome Res.* **17**:618–624.
- Masonick P, Michael A, Frankenberg S, Rabitsch W, Weirauch C. 2017. Molecular phylogenetics and biogeography of the ambush bugs (Hemiptera: Reduviidae: Phymatinae). *Mol Phylogenet Evol.* **114**:225–233.
- Mathers TC, Wouters RHM, Mugford ST, Swarbreck D, Van Oosterhout C, Hogenhout SA. 2020. Chromosome-scale genome assemblies of aphids reveal extensively rearranged autosomes and long-term conservation of the X chromosome. *Mol Biol Evol.* **38**:856–875.
- Meisel RP, Connallon T. 2013. The faster-X effect: integrating theory and data. *Trends Genet.* **29**:537–544.
- Meisel RP, Delclos PJ, Wexler JR. 2019. The X chromosome of the German cockroach, *Blattella germanica*, is homologous to a fly X chromosome despite 400 million years divergence. *BMC Biol.* **17**:100.
- Mesquita RD, Vionette-Amaral RJ, Lowenberger C, Rivera-Pomar R, Monteiro FA, Minx P, Spieth J, Carvalho AB, Panzera F, Lawson D, et al. 2015. 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.* **112**: 14936–14941.
- Murphy WJ, Foley NM, Bredemeyer KR, Gatesy J, Springer MS. 2021. Phylogenomics and the genetic architecture of the placental mammal radiation. *Annu Rev Anim Biosci.* **9**:29–53.
- Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol.* **32**:268–274.
- Oliveira DS, Brito NF, Franco TA, Moreira MF, Leal WS, Melo ACA. 2018. Functional characterization of odorant binding protein 27 (RproOBP27) from *Rhodnius prolixus* antennae. *Front Physiol.* **9**:1175.
- Owen CL, Marshall DC, Wade EJ, Meister R, Goemans G, Kunte K, Moulds M, Hill K, Villet M, Pham TH, et al. 2022. Detecting and removing sample contamination in phylogenomic data: an example and its implications for Cicadidae phylogeny (Insecta: Hemiptera). *Syst Biol.* **71**:1504–1523.
- Oyler-McCance SJ, Cornman RS, Jones KL, Fike JA. 2015. Z chromosome divergence, polymorphism and relative effective population size in a genus of lekking birds. *Heredity* **115**:452–459.
- Paliulis L, Fabig G, Müller-Reichert T. 2023. The X chromosome still has a lot to reveal—revisiting Hermann Henking’s work on firebugs. *J Cell Sci.* **136**:jcs260998.
- Paradis E, Schliep K. 2019. Ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**: 526–528.
- Parker DJ, Jaron KS, Dumas Z, Robinson-Rechavi M, Schwander T. 2022. X chromosomes show relaxed selection and complete somatic dosage compensation across *Timema* stick insect species. *J Evol Biol.* **35**:1734–1750.
- Pett W, Adamski M, Adamska M, Francis WR, Eitel M, Pisani D, Wörheide G. 2019. The role of homology and orthology in the phylogenomic analysis of metazoan gene content. *Mol Biol Evol.* **36**:643–649.
- Philippe H, Brinkmann H, Lavrov DV, Littlewood DTJ, Manuel M, Wörheide G, Baurain D. 2011. Resolving difficult phylogenetic questions: why more sequences are not enough. *PLoS Biol.* **9**: e1000602.
- Rannala B, Edwards SV, Leaché A, Yang Z. 2020. The multi-species coalescent model and Species tree inference. In: Scornavacca C, Delsuc F, Galtier N, editors. *Phylogenetics in the genomic era*. p. 3.3:1–3.3:21.
- Ranwez V, Harispe S, Delsuc F, Douzery EJP. 2011. MACSE: Multiple Alignment of Coding Sequences accounting for frameshifts and stop codons. *PLoS One* **6**:e22594.
- Robinson DF, Foulds LR. 1981. Comparison of phylogenetic trees. *Math Biosci.* **53**:131–147.
- Rodríguez-Ezpeleta N, Brinkmann H, Roure B, Lartillot N, Lang BF, Philippe H. 2007. Detecting and overcoming systematic errors in genome-scale phylogenies. *Syst Biol.* **56**:389–399.
- Sackton TB, Corbett-Detig RB, Nagaraju J, Vaishna L, Arunkumar KP, Hartl DL. 2014. Positive selection drives faster-Z evolution in silkworms. *Evolution* **68**:2331–2342.
- Schliep KP. 2011. phangorn: phylogenetic analysis in R. *Bioinformatics* **27**:592–593.
- Schuh RT, Weirauch C. 2020. *True bugs of the world (Hemiptera: Heteroptera): classification and natural history, monography series*. Manchester: Siri Scientific Press.
- Shen X-X, Hittinger CT, Rokas A. 2017. Contentious relationships in phylogenomic studies can be driven by a handful of genes. *Nat Ecol Evol.* **1**:126.
- Simion P, Delsuc F, Philippe H. 2020. To what extent current limits of phylogenomics can be overcome? In: Scornavacca C, Delsuc F, Galtier N, editors. *Phylogenetics in the genomic era*. p. 2.1:1–2.1:34.
- Simon C. 2020. An evolving view of phylogenetic support. *Syst Biol.* **71**:1–8.
- Slater GSC, Birney E. 2005. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**:31.
- Smith SA, Moore MJ, Brown JW, Yang Y. 2015. Analysis of phylogenomic datasets reveals conflict, concordance, and gene duplications with examples from animals and plants. *BMC Evol Biol.* **15**: 150.
- Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**: 1312–1313.
- Stamatakis A, Hoover P, Rougemont J. 2008. A rapid bootstrap algorithm for the RAxML web servers. *Syst Biol.* **57**:758–771.
- Swaegers J, Sánchez-Guillén RA, Chauhan P, Wellenreuther M, Hansson B. 2022. Restricted X chromosome introgression and support for Haldane’s Rule in hybridizing damselflies. *Proc Biol Sci.* **289**:20220968.

- Thornton K, Bachtrog D, Andolfatto P. 2006. X chromosomes and autosomes evolve at similar rates in *Drosophila*: no evidence for faster-X protein evolution. *Genome Res.* **16**:498–504.
- Toews DP, Brelsford A. 2012. The biogeography of mitochondrial and nuclear discordance in animals. *Mol Ecol.* **21**:3907–3930.
- Townsend JP. 2007. Profiling phylogenetic informativeness. *Syst Biol.* **56**:222–231.
- Van Dam MH, Henderson JB, Esposito L, Trautwein M. 2020. Genomic characterization and curation of UCEs improves species tree reconstruction. *Syst Biol.* **70**:307–321.
- Wang H-C, Minh BQ, Susko E, Roger AJ. 2018. Modeling site heterogeneity with posterior mean site frequency profiles accelerates accurate phylogenomic estimation. *Syst Biol.* **67**:216–235.
- Wang Y, Tang H, Debarry JD, Tan X, Li J, Wang X, Lee T-H, Jin H, Marler B, Guo H, *et al.* 2012. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **40**:e49.
- Weirauch C. 2008. Cladistic analysis of Reduviidae (Heteroptera: Cimicomorpha) based on morphological characters. *Syst Entomol.* **33**:229–274.
- Weirauch C, Bérenger JM, Berniker L, Forero D, Forthman M, Frankenberg S, Freedman A, Gordon E, Hoey-Chamberlain R, Hwang WS, *et al.* 2014. An illustrated identification key to assassin bug subfamilies and tribes (Hemiptera: Reduviidae). *Can J Arthropod Identif.* **26**:1–115.
- Weirauch C, Munro JB. 2009. Molecular phylogeny of the assassin bugs (Hemiptera: Reduviidae), based on mitochondrial and nuclear ribosomal genes. *Mol Phylogenet Evol.* **53**:287–299.
- Whittle CA, Kulkarni A, Extavour CG. 2020. Absence of a faster-X effect in beetles (*Tribolium*, Coleoptera). *G3 (Bethesda)* **10**:1125–1136.
- Wilson MA, Makova KD. 2009a. Genomic analyses of sex chromosome evolution. *Annu Rev Genomics Hum Genet.* **10**:333–354.
- Wilson MA, Makova KD. 2009b. Evolution and survival on eutherian sex chromosomes. *PLoS Genet.* **5**:e1000568.
- Xu K, Oh S, Park T, Presgraves DC, Yi SV. 2012. Lineage-specific variation in slow- and fast-X evolution in primates. *Evolution* **66**:1751–1761.
- Yamaguchi K, Kadota M, Nishimura O, Ohishi Y, Naito Y, Kuraku S. 2021. Technical considerations in Hi-C scaffolding and evaluation of chromosome-scale genome assemblies. *Mol Ecol.* **30**:5923–5934.
- Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol.* **24**:1586–1591.
- Young AD, Gillung JP. 2020. Phylogenomics—principles, opportunities and pitfalls of big-data phylogenetics. *Syst Entomol.* **45**:225–247.
- Zhang C, Rabiee M, Sayyari E, Mirarab S. 2018. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics* **19**:153.
- Zhang F, Ding Y, Zhu C-D, Zhou X, Orr MC, Scheu S, Luan Y-X. 2019. Phylogenomics from low-coverage whole-genome sequencing. *Methods Ecol Evol.* **10**:507–517.
- Zhang J, Gordon ERL, Forthman M, Hwang WS, Walden K, Swanson DR, Johnson KP, Meier R, Weirauch C. 2016. Evolution of the assassin's arms: insights from a phylogeny of combined transcriptomic and ribosomal DNA data (Heteroptera: Reduvioidea). *Sci Rep.* **6**:22177.
- Zhang J, Weirauch C, Zhang G, Forero D. 2016. Molecular phylogeny of Harpactorinae and Bactrodinae uncovers complex evolution of sticky trap predation in assassin bugs (Heteroptera: Reduviidae). *Cladistics* **32**:538–554.