

# Genome Report: Improved chromosome-level genome assembly of the American cockroach, *Periplaneta americana*

Rachel L. Dockman<sup>1</sup>, Tyler J. Simmonds<sup>2</sup>, Kevin J. Vogel<sup>3</sup>, Scott M. Geib<sup>2</sup>, Elizabeth A. Ottesen<sup>1\*</sup>

<sup>1</sup>Department of Microbiology, University of Georgia, Athens, GA, USA

<sup>2</sup>USDA-ARS, Hilo, HI, USA

<sup>3</sup>Department of Entomology, University of Georgia, Athens, GA, USA

**\* Correspondence:**

Dr. Elizabeth Ottesen

ottesen@uga.edu

**Keywords:** *Periplaneta americana*, American cockroach, genome, chromosome, Hi-C, PacBio, Animalia, cockroach, HiFiASM, Blattodea

## 1 Abstract

The American cockroach, *Periplaneta americana*, is a cosmopolitan insect notorious for thriving among humans undeterred by attempts to eliminate it. The traits that contribute to its ubiquity as an opportunistic pest, such as long lifespan, expansive neurosensory capacity, and nutritional flexibility, also make *P. americana* an excellent invertebrate model organism with a long history in neuroscience and physiological research. Current genetic resources available for *P. americana* highlight its large, complex genome and richly diverse transcriptional capabilities, but fall short of producing a complete, chromosome-level genome. Here, we present a high-quality *de novo* genome assembly of a laboratory-raised adult female *P. americana* using a combination of high fidelity PacBio long reads and Hi-C sequencing. The final 3.23 Gb genome was assembled with chromosomal resolution into 17 scaffolds, consistent with previous karyotype analysis, and has a scaffold N50 of 188.1 Mb and genome BUSCO score of 99.7%. This assembly includes a chromosome that was missing from the previous reference genome for this species. Protein prediction and annotation were performed via the NCBI Eukaryotic Genome Annotation Pipeline, which identified 16,780 protein-coding genes and generated an annotation BUSCO score of 97.8%. Ortholog comparisons with available Blattodea assemblies highlight the expanded chemosensory and immune capabilities of *P. americana* compared to termite relatives. This genome assembly is a valuable tool for facilitating future research on the biology and evolution of this remarkable insect.

## 2 Introduction

The American cockroach (*Periplaneta americana*) is a notorious pest found across the world, living and thriving alongside humans in widely variable environmental conditions. Despite their unsavory reputation, cockroaches are of considerable interest to researchers across disciplines. Cockroaches combine the large size, complex physiology, and tractable nature of mice with the simpler and cost-effective care requirements characteristic of model insects.

There is a rich body of work compiled throughout the last century describing basic cockroach biology, covering topics ranging from external morphology to internal physiology [1-4]. Early research leveraging *P. americana* as a model system has enhanced understanding of nervous system function, connectivity, and regeneration [5-9] and discovered unique traits that allow *P. americana* to survive extreme environmental stressors, such as endosymbiont-mediated nitrogen cycling [10, 11]. Studies on the cockroach immune system have linked its expansive repertoire of immune-associated proteins to common allergens [12, 13] and mechanisms of pesticide resistance [14-16]. The extensive antimicrobial and regenerative capabilities of *P. americana* have also earned it respect in both ancient and modern Chinese tradition as an important medicinal insect [17, 18]. Modern sequencing technologies and genetic techniques have further supported the use of *P. americana* as a model organism. Its susceptibility and robust response to RNA inhibition (RNAi) via multiple administration methods, *P. americana* is an especially useful organism for deciphering gene involvement in insect physiological development and pesticide resistance, among other investigations [19-21]. *P. americana* also shows potential as an emerging insect model for host-microbiome interactions. The gut microbiome of omnivorous cockroaches reflects that of humans and omnivorous mammalian model systems [22-25], and germ-free nymphs can be easily produced via ootheca sterilization, a highly desirable trait for defined-community research interests [26-29]. Altogether, these traits support the research potential of this insect and argue towards the necessity of a complete, high-quality American cockroach genome. While there are two previous genome assemblies publicly available, the assembly presented by [30] lacks publicly available protein annotations and is limited by the short-read technology available at the time and the assembly prepared by [13], while a significant improvement, contains more scaffolds than supported by karyotype analysis (male/XO: 33 diploid; female/XX: 34 diploid) [31].

Here, we present the first chromosome-level assembled genome of *P. americana*. We used long-read high fidelity (HiFi) PacBio sequencing in addition to chromatin contact mapping (Hi-C) to produce a genome scaffolded into a 17-chromosome assembly, consistent with previous karyotype findings [31]. This high-quality genome assembly is an important tool for facilitating future genetic research in the American cockroach.

## 3 Materials and Methods

### 3.1 Insect origin and selection

An adult female *Periplaneta americana* individual was selected from a stock colony maintained at the University of Georgia by the Ottesen laboratory; this colony has been maintained the laboratory for 10 years and originated in another long-term laboratory colony of unknown origin. The specimen was flash frozen in liquid nitrogen and shipped on dry ice to the United States Department of Agriculture - Agricultural Research Service (USDA-ARS) – Pacific Basin Agricultural Research Center (PBARC) in Hilo, Hawaii.

### 3.2 Sample preparation and sequencing methods

For PacBio sequencing, high molecular weight (HMW) DNA extraction was performed from insect leg tissue using the Qiagen MagAttract HMW DNA Kit (Qiagen, Hilden Germany). DNA was sheared with the Diagenode Megaruptor 2 (Denville, New Jersey, USA) then prepared for PacBio sequencing using the SMRTBell Express Template kit 2.0 (Pacific Biosciences, Menlo Park, California, USA). The library was size-selected prior to sequencing on a Sequel II System (Pacific Biosciences, Menlo Park, California, USA) using Binding Kit v2.0, Sequencing kit v2.0, and SMRT Cell 8M. To target HiFi reads, the library was sequenced using a 30-hour movie time on three

SMRTcells. Raw subreads were converted to HiFi data by processing with CCS to call a single high quality consensus sequence for each molecule, using a 99.5% consensus accuracy cutoff.

For Hi-C sequencing, the Arima Hi-C kit (Arima Genomics, San Diego, California, USA) was used to crosslink leg tissue DNA and perform proximity ligation, following the Arima Hi-C low input protocol. After proximity ligation, DNA was sheared with a Diagenode Bioruptor then size-selected for 200-600bp DNA fragments. The Swift Accel NGS 2S Plus kit (Integrated DNA Technologies, Coralville, Iowa, USA) was used to prepare an Illumina library from the size-selected DNA, which was then sequenced with an Illumina NovaSeq 6000 (Illumina, San Diego, California, USA).

### 3.3 RNA-seq

Transcriptomic data was obtained for the midgut, hindgut, and fat body of 10 individual cockroaches for a dietary experiment investigating the impact of carbohydrate source on the microbial metatranscriptome and host transcriptome (BioProject: PRJNA1105088). Data were obtained as 150bp paired end reads on Illumina NovaSeq from Novogene Corporation in Sacramento, California. The Joint Genome Institute (JGI) programs BBduk and BBSplit (<https://jgi.doe.gov/data-and-tools/software-tools/bbtools/>) and were used to remove sequencing adapters and screen for initial *Blattabacterium* contamination, and SortMeRNA was used to remove ribosomal RNA reads [32]. Cleaned RNA reads were aligned to the unfiltered genome for contaminant filtering.

### 3.4 Genome assembly and polishing

The first genome assembly was generated using hifiasm v0.19.6-r595 Hi-C integration to obtain primary, alternate, and haplotype-phased assemblies [33]. Three flow cells of PacBio HiFi data were obtained from long-read sequencing and concatenated into a single fastq file for assembly and used with Hi-C data obtained from the same source insect. The primary assembly was selected for polishing and contigs were filtered to retain those with coverage reported from hifiasm as between 6X and 30X.

Contamination filtering was performed as described by Lu and Salzberg [34]. Contigs were separated into individual files, then fragmented with SeqKit v2.5.1 into 100 bp pseudo-reads with 50 bp overlaps [35]. These pseudo-reads were fed through Kraken v2.1.3 with default parameters to align to the default eukaryotic and prokaryotic databases, and contigs identified as *Blattabacterium* or with 70% identity assigned to *Homo sapiens* were discarded [36]. Remaining contigs were masked using RepeatMasker v4.1.5 (<https://www.repeatmasker.org/RepeatMasker/>) with Dfam library v3.7 prior to RNA-seq alignment with HISAT2, and contigs with average read depths exceeding 10,000 were removed [37, 38].

The Arima Genomics mapping pipeline ([https://github.com/ArimaGenomics/mapping\\_pipeline](https://github.com/ArimaGenomics/mapping_pipeline)) was used to map Hi-C data to the filtered assembly. The pipeline described utilizes the programs BWA-MEM for separate alignment of the paired Illumina Hi-C reads, the Picard (<https://broadinstitute.github.io/picard/>) “MarkDuplicates” command for PCR duplicate removal, and SAMtools for file sorting and handling [39, 40]. The resulting Hi-C alignment files and the hifiasm assembly were fed into Yet Another Hi-C Scaffolding tool (YaHS) without breaking contigs, and the resulting contact map was visualized with Juicebox for manual curation [41, 42]. The corrected genome was exported and screened for possible telomeres using tidk (<https://github.com/tolkit/telomeric-identifier>) and repeat regions identified *de novo* (AACCTAACCT) were graphed [43]. An additional round of polishing was performed in Juicebox to correct repeat-heavy telomeric loci, and completeness of the final scaffolded genome was assessed

with the Insecta set (version odb10) of Benchmarking Universal Single-Copy Orthologs (BUSCOs) [44].

### 3.5 Mitochondrion identification

The program MitoHiFi v2 was used to identify a consensus mitochondrial genome sequence from the original hifiasm assembly [45]. The NCBI reference sequence NC\_016956.1 was selected as the reference *P. americana* mitochondrion for identification [46].

### 3.6 Repeat modeling and masking

For in-depth repeat identification, RepeatModeler v2.0.4 was used to create an *ab initio* repeat library specific to *P. americana*, which was then separated into libraries of known and unknown repeat families [47]. Using the script repclassifier.sh (<https://github.com/darencard/GenomeAnnotation/blob/master/repclassifier>), unknown repeats were iteratively re-annotated for seven rounds, when the percent of repeats classified as “known” rather than “unknown” plateaued. Four rounds of repeat masking using RepeatMasker v4.1.5 were then performed as described in (<https://darencard.net/blog/2022-07-09-genome-repeat-annotation/>), during which simple repeats were identified and masked first, followed by insect-specific Dfam repeats, known *P. americana* repeats, and lastly unknown *P. americana* repeats. The output from the four RepeatMasker rounds were combined to generate an overall masked genome, annotation files, and table describing the repeats, and repeat landscapes were summarized with the script parseRM.pl (<https://github.com/4ureliek/Parsing-RepeatMasker-Outputs>).

### 3.7 Genome annotation

The finished assembly was submitted to NCBI for structural and functional gene annotation via the automated Eukaryotic Genome Annotation Pipeline v10.3. Evidence fed into the GNOMON gene prediction tool included existing RNA-seq data and transcriptome assemblies for *P. americana*, NCBI RefSeq protein sets from *Acromyrmex echinator*, *Hyaella azteca*, *Acyrtosiphon pisum*, *Caenorhabditis elegans*, *Tribolium castaneum*, *Drosophila melanogaster*, and *Apis mellifera*, as well as Insecta and *P. americana* GenBank protein sets. Details of the annotation release (GCF\_040183065.1-RS\_2024\_10) are available at [https://www.ncbi.nlm.nih.gov/refseq/annotation\\_euk/Periplaneta\\_americana/GCF\\_040183065.1-RS\\_2024\\_10/](https://www.ncbi.nlm.nih.gov/refseq/annotation_euk/Periplaneta_americana/GCF_040183065.1-RS_2024_10/).

### 3.8 Blattodea comparison

The annotations obtained for this assembly were compared with Blattodea annotations available on NCBI for a previous *P. americana* genome (GCA\_025594305.2) [13], *Zootermopsis nevadensis* (GCF\_000696155.1) [48], *Blattella germanica* (GCA\_003018175.1) and *Cryptotermes secundus* (GCF\_002891405.2) [49], *Coptotermes formosanus* (GCA\_013340265.1), and *Diploptera punctata* (GCA\_030220185.1) [50]. OrthoFinder v2.5.5 was used with GENESPACE v1.3.1 in R Studio to identify orthologous groups and generate riparian plots [51-53]. The distribution of orthologous gene sets between and across genomes were visualized with the R package UpSetR v1.4.0 [54]. GO terms assigned to *P. americana* genes that were sorted into orthologous gene sets were used for GO enrichment analysis in R with the package clusterProfiler v4.6.2 [55, 56].

## 4 Results/Discussion

### 4.1 Assembly



American cockroaches have XX/X0 sex chromosome systems [57], so a single female insect was selected for PacBio SMRTBell sequencing, producing over 6.6 million reads covering 90.6 Gb. Paired-end 150 bp sequences generated by Hi-C sequencing produced an additional 217.3 Gb of sequence data. These data were assembled with hifiasm in Hi-C mode to produce a primary genome assembly containing 1818 contigs with a contig N50 over 40 Mb and total length of 3.36 Gb, close to the genome size of 3.338 Gb predicted through previous flow cytometry work [58]. In addition, we identified a complete consensus mitogenome (**Figure S1**). We evaluated the assembly quality via read coverage and depth (**Figure 1**) prior to filtering and found that 96.19% of the genome assembly was contained in contigs with coverage between 6x and 30x. Subsequent contamination filtering using Kraken2 identified 19 contigs as derived from *Blattabacterium*, the cockroach-associated endosymbiont. All contigs flagged as *Blattabacterium* DNA in our genome assembly were self-contained, suggesting that these reads stemmed from contamination rather than integration into *P. americana*'s genome, and they were therefore removed. Further filtering based on RNA-seq alignment decreased the number of contigs prior to scaffolding to 243 with a N50 of 42.4 Mb, overall producing a high-quality contig-level genome with no detectable contamination.

Initial scaffolding with YaHS assigned 95.67% of the genome into 17 scaffolds (**Figure S2**), with manual curation in Juicebox increasing this percentage to 98.93% (**Figure 2A**). Telomere analysis identified at least one telomeric region, determined *de novo* as AACCTAACCT, for each of the chromosome-sized scaffolds, with 10 scaffolds flanked on both ends and 6 scaffolds flanked on one end (**Figure 2B**). Scaffold #7 contains a long sequence of probable telomeric repeats embedded within a large contig. While this may indicate an assembly error, at this time we do not have evidence to support adjusting its placement. The 74 remaining contigs were unable to be matched to just one scaffold, likely due to a high density of centromeric repeat regions and were therefore left unplaced. BUSCO analysis of the putative chromosomes found 99.7% of Insecta BUSCOs were present and complete in this genome assembly, of which only 1.5% were duplicated (**Figure S3, Table 1**).

In summary, the assembly presented here is considered high quality across a number of standard metrics. Through a combination of long-read and chromosomal contact sequencing data, we successfully scaffolded 98.9% of this 3.23 Gb assembly into 17 chromosome-scale scaffolds, values which are supported by previous karyotype and flow cytometry findings for *P. americana* of 17 haploid (male/female: 33/34 diploid) chromosomes and a predicted genome size of 3.338 Gb [31, 57, 58]. This represents an improvement on the currently available *P. americana* genome assemblies in NCBI (**Table 1**), which resemble this assembly in length and GC content but have lower contiguity and BUSCO scores [13, 30]. Therefore, we argue that this genome assembly qualifies as both comprehensive and chromosomally resolved.

## 4.2 Repetitive DNA elements

Overall, 50.86% of the genome was identified as repetitive content classified as DNA elements, simple repeat regions, or retroelements, with an additional 15% of the genome determined to be *P. americana*-specific repeats that remain otherwise unclassified (**Figure 3A, Table S1**). The DNA transposon and retroelement subgroups contributed similarly to overall repeat content, comprising 24.3% and 20.05% respectively of the genome, but differ in their overall Kimura divergence landscapes (**Figure 3B**). The DNA elements found in this genome primarily belong to the Tc1-Mariner and hobo-Activator-Tam3 subfamilies with higher relative abundances at a Kimura substitution level of 5% (**Figure 3C**). In contrast, the primary retrotransposon class of long interspersed nuclear elements (LINEs) contained two peaks in Kimura substitution, with CR1, L1, L2, and CRE subgroups showing elevated substitution levels around 5% while RTE-clade

retrotransposons represent a more ancient repeat lineage, with substitution levels peaking between 31-32% (**Figure 3D**). Short interspersed nuclear elements (SINEs) and retroelements containing long-terminal repeats (LTRs) were less commonly identified, only making up 3.47% and 0.94% of the genome respectively. Most SINEs contained internal promoters derived from tRNA and showed more gradual patterns of divergence, with a relatively stable plateau between 5% and 15% divergence before tapering (**Figure 3E**). Repeat content varies widely even between insects from the same family, but generally Blattodea species show similar repeat distributions with especially low LTR content and expanded LINEs, with larger genomes correlated to higher repeat content [59].

### 4.3 Orthology analysis between Blattodea species

The order Blattodea encompasses both termites and cockroaches, with over 4700 species identified in NCBI's taxonomy repository. Despite these many representatives, only 12 species have sequenced genomes [13, 30, 48-50, 60-63], of which half have publicly available annotations uploaded to NCBI (**Figure 4A**). The 3.2 Gb genome of *P. americana* is the largest among sequenced Blattodea and more than double the size of available termite genomes (**Figure 4B**), consistent with two previous *P. americana* genome assemblies. This assembly has the highest contig N50 among these Blattodea (**Figure 4C**) and the second highest scaffold N50 behind *E. pallidus* [60] (**Figure 4D**). Compared to those genomes with available annotations, *P. americana* encodes more protein-coding genes than termites, but had less reported than the other cockroach genomes (**Figure 4E**). It is unclear whether this difference between annotations in our assembly and the other cockroach assemblies is biological or a result of annotation technique; this assembly was annotated by NCBI's Gnomon pipeline which produced 16,780 protein-coding genes, while previous *P. americana* genomes reported 21,336 (annotations not available) and 29,939 (annotations available) protein-coding genes [13, 30].

We chose to use the genome annotations available on NCBI for ortholog analysis, which included three members of Termitidae (*Zootermopsis nevadensis*, *Cryptotermes secundus*, *Coptotermes formosanus*), two Blaberoidea (*Blattella germanica*, *Diploptera punctata*) and a previous *P. americana* genome [13]. Initial OrthoFinder results included both *P. americana* assemblies, identifying 16,711 orthologous gene families among these seven annotated genomes (**Supplemental File 1**). Of these 16,711 gene families, 2,311 were shared across all seven Blattodea with an additional 2,220 gene families shared by all assemblies excluding the previous *P. americana* assembly (**Figure S4**). While both *P. americana* genomes shared 391 gene families absent in the other species, 198 gene families were only present in our assembly (**Figure S4**). We further evaluated the differences between both *P. americana* assemblies via synteny analysis (**Figure 5A**) and found that, while most chromosomes in our assembly were captured in their entirety by 1-2 scaffolds in the previous assembly, our chromosome 10 and half of chromosome 14 were missing from the other *P. americana* annotations. Proteins encoded in chromosome 10 encompass a wide range of functions with immune (Dscam, toll-like receptors, leucine-rich repeat proteins), neurologic (GABA transport, neurotrophic factors), endocrine (vitellogenin synthesis, sterol binding proteins), and nutritional (xanthine dehydrogenase, salivary peptide) importance. Additional synteny analysis between our *P. americana* genome and other cockroaches (**Figure 5B**) and termites (**Figure 5C**) supported the existence of these regions. As a result, the previous *P. americana* assembly was excluded from further cross-*Blattodea* orthogroup analysis.

Recalculating OrthoFinder statistics produced 15,491 total gene families, of which 4,531 orthogroups were shared by all species (**Figure 5D**). These gene families, which are shared at the order level, comprised large fractions of each insect's total gene count, ranging from 5,674 genes (45.8%) in *Z. nevadensis* to 7,860 genes (27.7%) in *D. punctata* (**Figure 5D inset**). These core gene families

encompass a wide range of functions necessary to life and common in insects, with GO terms relating to gene expression and genome maintenance especially prevalent in this subset (**Figure S5**). The full list of GO terms assigned to these shared gene families can be found in **Supplemental File 2**.

The percentage of genes unable to be assigned to orthogroups varied between Blaberoidea and Blattodea (**Figure 5D inset**). The termite genomes and our *P. americana* assembly had between 95-97.7% of their genes assigned to orthogroups, while 75.3% and 76.2% of *D. punctata* and *B. germanica* genes, respectively, were assigned to orthogroups (**Figure 5D inset**). Unassigned genes are considered unique both between analyzed genomes and within a single genome, so it is unclear whether these genes stem from expanded single-copy genes within the cockroaches or are misannotated as genes without conferring any biological function. Possibly, the annotations of the termites and our *P. americana* are overly conservative, but gene finding programs rely on homology comparison that requires more sequenced and annotated Blattodea genomes for effective performance. Despite these questions, the difference between cockroaches and termites in their species-specific gene families highlight the gene expansion occurring as these two groups diverged. Our *P. americana* assembly, which encodes over 3,000 more genes than the termite assemblies, has 11.7% of its genes assigned to the 589 *P. americana*-unique gene families, a substantial increase compared to the 1.5-3.2% range in the three termite assemblies (**Figure 5D inset**). This expansion of genes within a single species also occurs in *B. germanica* and *D. punctata*, which have 23.2% and 13.5% of their genes respectively assigned to species-specific gene families. Since ortholog analysis is dependent on the data supplied, it is difficult to determine whether this difference stems from biological cockroach-termite delineation or is a consequence of poor representation of closely related cockroach species. Nonetheless, these results, in combination with the large size of this cockroach genome (**Figure 4B**) and the late spike in LINE retrotransposon divergence (**Figure 3D**), suggests that acquisition and expansion of new genes contributed to cockroach divergence.

We performed GO term analysis of the 589 orthologous gene families present in our *P. americana* assembly but absent in the other Blattodea species (**Figure 5E**). GO terms related to gene expression and genomic maintenance were most enriched, reflecting the GO enrichment in shared gene families (**Figure S5**). Further investigation revealed that 157 of these orthogroups were associated with zinc finger family genes, likely skewing the GO analysis results and masking other expanded functions of interest. As an alternative approach, we evaluated the semantic similarity of named genes and/or GO terms (if name is missing) of these *P. americana*-unique gene families to shed light on the function of these expanded genes (**Figure S6**). These gene families were enriched in immune and digestive functions, such as lipopolysaccharide recognition, protease activity, odorant binding, and lipase activity (**Figure S6**), which may have facilitated cockroach divergence from the protective eusociality found in termite colonies towards a more independent and self-sufficient lifestyle. However, there are a limited number of sequenced cockroach representatives, so further cataloging of diverse Blattodea genomes is necessary to pinpoint exact relationships between these gene families and cockroach-termite evolution. Overall, synteny and ortholog comparison between these Blattodea reveal possible mechanisms of divergence between termites and cockroaches and highlight the potential applications of a chromosomally resolved *P. americana* genome.

## 5 Summary

We sequenced the genome of the American cockroach using high fidelity PacBio long reads in conjunction with Hi-C Illumina short reads. This 3.23 Gb assembly is highly contiguous, with a contig N50 of 42 Mb and a scaffold N50 of 188 Mb, and 98.93% of the assembly is contained within

17 putative chromosomes. The quality of this assembly is further exemplified by its genomic and protein BUSCO scores, which are 99.7% and 97.8% complete respectively. This high-quality assembly, generated with cutting edge sequencing technology, is a substantial improvement over existing *P. americana* genomes, and we report an entire chromosome that was missing from a previously published assembly. This genome is expected to facilitate future study of cockroach physiology and Blattodea evolution.

## 6 Data Availability

Data associated with this study are available from the NCBI Sequence Read Archive under BioProjects PRJNA1098420 (principal haplotype) and PRJNA1098419 (alternative haplotype). Raw sequencing reads for Hi-C data is available from the SRA with accession SRX24490912, and PacBio HiFi reads may be obtained from accessions SRX24490909, SRX24490910, and SRX24490911. RNA-seq data is deposited under the SRA BioProject PRJNA1105088. Scripts used for assembling and analyzing this genome are available at: <https://github.com/rldockman/PAMFEO>.

## 7 Funding

This work was supported by the National Institute of General Medical Sciences of the National Institutes of Health (NIH) under award number R35GM133789 to EAO. This research was supported by the in-house appropriated USDA-ARS project *Advancing Molecular Pest Management, Diagnostics, and Eradication of Fruit Flies and Invasive Species* (no. 2040-22430-028-000-D) and used resources provided by SCINet, USDA-ARS projects no. 0201-88888-003-000D and 0201-88888-002-000D. USDA is an equal opportunity provider and employer. Mention of trade names does not imply an endorsement from USDA or the Federal Government.

## 8 Acknowledgements

We would like to thank Jeremy Schrader for his assistance with preparing and sequencing the DNA for this project.

## References

1. Crampton, G.C., *The External Anatomy of the Head and Abdomen of the Roach, Periplaneta americana*. Psyche: A Journal of Entomology, 1925. **32**(4-5): p. 127409.
2. Guthrie, D.M. and A.R. Tindall, *The biology of the cockroach*. 1968, London: Edward Arnold. 408.
3. Cornwell, P.B., *The cockroach: a laboratory insect and an industrial pest*. 1968, United Kingdom: Huthinson.
4. *The American Cockroach*. 1982, London: Chapman and Hall.
5. Jacklet, J.W. and M.J. Cohen, *Nerve Regeneration: Correlation of Electrical, Histological, and Behavioral Events*. Science, 1967. **156**(3782): p. 1640-1643.
6. Jankowska, M., et al., *Electromagnetic field and TGF- $\beta$  enhance the compensatory plasticity after sensory nerve injury in cockroach Periplaneta americana*. Scientific Reports, 2021. **11**(1): p. 6582.
7. Yamasaki, T. and T. Narahashi, *Synaptic Transmission in the Cockroach*. Nature, 1958. **182**(4652): p. 1805-1806.



- 338 8. Sattelle, D.B., *Receptors for L-glutamate and GABA in the nervous system of an insect*  
339 *(Periplaneta americana)*. Comp Biochem Physiol C Comp Pharmacol Toxicol, 1992. **103**(3):  
340 p. 429-38.
- 341 9. Roeder, K.D., *The effect of potassium and calcium on the nervous system of the cockroach,*  
342 *Periplaneta americana*. J Cell Comp Physiol, 1948. **31**(3): p. 327-38.
- 343 10. Vicente, C.S.L., et al., *Genome analysis of new Blattabacterium spp., obligatory*  
344 *endosymbionts of Periplaneta fuliginosa and P. japonica*. PLoS One, 2018. **13**(7): p.  
345 e0200512.
- 346 11. Cochran, D.G., D.E. Mullins, and K.J. Mullins, *Cytological Changes in the Fat Body of the*  
347 *American Cockroach, Periplaneta americana, in Relation to Dietary Nitrogen Levels*. Annals  
348 of the Entomological Society of America, 1979. **72**(2): p. 197-205.
- 349 12. Arumugam, G., et al., *Purification and functional characterization of lectin with*  
350 *phenoloxidase activity from the hemolymph of cockroach, Periplaneta americana*. Arch  
351 Insect Biochem Physiol, 2017. **95**(2).
- 352 13. Wang, L., et al., *Genome assembly and annotation of Periplaneta americana reveal a*  
353 *comprehensive cockroach allergen profile*. Allergy, 2023. **78**(4): p. 1088-1103.
- 354 14. Zhang, J.H., et al., *New insight into foregut functions of xenobiotic detoxification in the*  
355 *cockroach Periplaneta americana*. Insect Sci, 2018. **25**(6): p. 978-990.
- 356 15. Zhang, J., et al., *Midgut Transcriptome of the Cockroach Periplaneta americana and Its*  
357 *Microbiota: Digestion, Detoxification and Oxidative Stress Response*. PloS one, 2016. **11**(5):  
358 p. e0155254-e0155254.
- 359 16. Sun, H., et al., *Elucidating the detoxification efficacy of Periplaneta americana delta*  
360 *glutathione S-transferase 1 (PaGSTd1) against organophosphates*. Pesticide Biochemistry  
361 and Physiology, 2024. **203**: p. 106013.
- 362 17. Zhang, E., et al., *A minireview of the medicinal and edible insects from the traditional*  
363 *Chinese medicine (TCM)*. Front Pharmacol, 2023. **14**: p. 1125600.
- 364 18. Tsuneo, N., M.A. Yong-Hua, and I. Kenji, *Insect derived crude drugs in the chinese song*  
365 *dynasty*. Journal of Ethnopharmacology, 1988. **24**(2): p. 247-285.
- 366 19. Li, L., et al., *Applications of RNA Interference in American Cockroach*. J Vis Exp, 2021(178).
- 367 20. Huang, J.H., et al., *Practical Use of RNA Interference: Oral Delivery of Double-stranded*  
368 *RNA in Liposome Carriers for Cockroaches*. J Vis Exp, 2018(135).
- 369 21. French, A.S., et al., *Transcriptome analysis and RNA interference of cockroach*  
370 *phototransduction indicate three opsins and suggest a major role for TRPL channels*. Front  
371 Physiol, 2015. **6**: p. 207.
- 372 22. Tinker, K.A. and E.A. Ottesen, *The Core Gut Microbiome of the American Cockroach,*  
373 *Periplaneta americana, Is Stable and Resilient to Dietary Shifts*. Appl Environ Microbiol,  
374 2016. **82**(22): p. 6603-6610.
- 375 23. Dockman, R.L. and E.A. Ottesen, *Purified fibers in chemically defined synthetic diets*  
376 *destabilize the gut microbiome of an omnivorous insect model*. Frontiers in Microbiomes,  
377 2024. **3**(1477521).

- 378 24. Zurek, L. and B.A. Keddie, *Contribution of the colon and colonie bacterial flora to*  
379 *metabolism and development of the American cockroach Periplaneta americana* (L). Journal  
380 of Insect Physiology, 1996. **42**(8): p. 743-748.
- 381 25. Ayayee, P.A., et al., *The role of gut microbiota in the regulation of standard metabolic rate in*  
382 *female Periplaneta americana*. PeerJ, 2018. **6**: p. e4717.
- 383 26. Dukes, H.E., J.E. Dyer, and E.A. Ottesen, *Establishment and Maintenance of Gnotobiotic*  
384 *American Cockroaches (Periplaneta americana)*. JoVE, 2021(171): p. e61316.
- 385 27. Tegtmeier, D., et al., *Oxygen Affects Gut Bacterial Colonization and Metabolic Activities in a*  
386 *Gnotobiotic Cockroach Model*. Appl Environ Microbiol, 2016. **82**(4): p. 1080-9.
- 387 28. Vera-Ponce de León, A., et al., *Microbiota Perturbation or Elimination Can Inhibit Normal*  
388 *Development and Elicit a Starvation-Like Response in an Omnivorous Model Invertebrate*.  
389 mSystems, 2021. **6**(4): p. e0080221.
- 390 29. Mikaelyan, A., et al., *Deterministic Assembly of Complex Bacterial Communities in Guts of*  
391 *Germ-Free Cockroaches*. Appl Environ Microbiol, 2016. **82**(4): p. 1256-63.
- 392 30. Li, S., et al., *The genomic and functional landscapes of developmental plasticity in the*  
393 *American cockroach*. Nature Communications, 2018. **9**(1): p. 1008.
- 394 31. John, B. and K.R. Lewis, *Chromosome structure in Periplaneta americana*. Heredity, 1960.  
395 **15**(1): p. 47-54.
- 396 32. Kopylova, E., L. Noé, and H. Touzet, *SortMeRNA: fast and accurate filtering of ribosomal*  
397 *RNAs in metatranscriptomic data*. Bioinformatics, 2012. **28**(24): p. 3211-3217.
- 398 33. Cheng, H., et al., *Haplotype-resolved de novo assembly using phased assembly graphs with*  
399 *hifiasm*. Nature Methods, 2021. **18**(2): p. 170-175.
- 400 34. Lu, J. and S.L. Salzberg, *Removing contaminants from databases of draft genomes*. PLoS  
401 Comput Biol, 2018. **14**(6): p. e1006277.
- 402 35. Shen, W., et al., *SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File*  
403 *Manipulation*. PLOS ONE, 2016. **11**(10): p. e0163962.
- 404 36. Wood, D.E., J. Lu, and B. Langmead, *Improved metagenomic analysis with Kraken 2*.  
405 Genome Biology, 2019. **20**(1): p. 257.
- 406 37. Kim, D., et al., *Graph-based genome alignment and genotyping with HISAT2 and HISAT-*  
407 *genome*. Nature Biotechnology, 2019. **37**(8): p. 907-915.
- 408 38. Storer, J., et al., *The Dfam community resource of transposable element families, sequence*  
409 *models, and genome annotations*. Mobile DNA, 2021. **12**(1): p. 2.
- 410 39. Li, H., *Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM*.  
411 arXiv q-bio.GN, 2013.
- 412 40. Danecek, P., et al., *Twelve years of SAMtools and BCFtools*. Gigascience, 2021. **10**(2).
- 413 41. Durand, N.C., et al., *Juicebox Provides a Visualization System for Hi-C Contact Maps with*  
414 *Unlimited Zoom*. Cell Syst, 2016. **3**(1): p. 99-101.
- 415 42. Zhou, C., S.A. McCarthy, and R. Durbin, *YaHS: yet another Hi-C scaffolding tool*.  
416 Bioinformatics, 2023. **39**(1).

- 417 43. Brown, M., P.M. González De la Rosa, and B. Mark, *A Telomere Identification Toolkit*.  
418 Zenodo, 2023.
- 419 44. Manni, M., et al., *BUSCO Update: Novel and Streamlined Workflows along with Broader and*  
420 *Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes*.  
421 Molecular Biology and Evolution, 2021. **38**(10): p. 4647-4654.
- 422 45. Uliano-Silva, M., et al., *MitoHiFi: a python pipeline for mitochondrial genome assembly*  
423 *from PacBio high fidelity reads*. BMC Bioinformatics, 2023. **24**(1): p. 288.
- 424 46. Xiao, B., et al., *Complete mitochondrial genomes of two cockroaches, Blattella germanica*  
425 *and Periplaneta americana, and the phylogenetic position of termites*. Curr Genet, 2012.  
426 **58**(2): p. 65-77.
- 427 47. Flynn, J.M., et al., *RepeatModeler2 for automated genomic discovery of transposable element*  
428 *families*. Proceedings of the National Academy of Sciences, 2020. **117**(17): p. 9451-9457.
- 429 48. Terrapon, N., et al., *Molecular traces of alternative social organization in a termite genome*.  
430 Nat Commun, 2014. **5**: p. 3636.
- 431 49. Harrison, M.C., et al., *Hemimetabolous genomes reveal molecular basis of termite*  
432 *eusociality*. Nature Ecology & Evolution, 2018. **2**(3): p. 557-566.
- 433 50. Fouks, B., et al., *Live-bearing cockroach genome reveals convergent evolutionary*  
434 *mechanisms linked to viviparity in insects and beyond*. iScience, 2023. **26**(10): p. 107832.
- 435 51. Emms, D.M. and S. Kelly, *OrthoFinder: phylogenetic orthology inference for comparative*  
436 *genomics*. Genome Biology, 2019. **20**(1): p. 238.
- 437 52. Lovell, J.T., et al., *GENESPACE tracks regions of interest and gene copy number variation*  
438 *across multiple genomes*. eLife, 2022. **11**: p. e78526.
- 439 53. Emms, D.M. and S. Kelly, *OrthoFinder: solving fundamental biases in whole genome*  
440 *comparisons dramatically improves orthogroup inference accuracy*. Genome Biology, 2015.  
441 **16**(1): p. 157.
- 442 54. Conway, J.R., A. Lex, and N. Gehlenborg, *UpSetR: an R package for the visualization of*  
443 *intersecting sets and their properties*. Bioinformatics, 2017. **33**(18): p. 2938-2940.
- 444 55. Yu, G., et al., *clusterProfiler: an R Package for Comparing Biological Themes Among Gene*  
445 *Clusters*. OMICS: A Journal of Integrative Biology, 2012. **16**(5): p. 284-287.
- 446 56. Wu, T., et al., *clusterProfiler 4.0: A universal enrichment tool for interpreting omics data*.  
447 The Innovation, 2021. **2**(3).
- 448 57. Jankásek M., Kotyková Varadinová Z. and Šťáhlavský F, *Blattodea Karyotype Database*.  
449 EJE, 2021. **118**(1): p. 192-199.
- 450 58. Hanrahan, S.J. and J.S. Johnston, *New genome size estimates of 134 species of arthropods*.  
451 Chromosome Res, 2011. **19**(6): p. 809-23.
- 452 59. Sproul, J.S., et al., *Analyses of 600+ insect genomes reveal repetitive element dynamics and*  
453 *highlight biodiversity-scale repeat annotation challenges*. Genome Res, 2023. **33**(10): p.  
454 1708-1717.
- 455 60. Hunter, T., *The genome sequence of the tawny cockroach, Ectobius (Ectobius) pallidus*  
456 *(Olivier, 1789)*. Wellcome Open Res, 2025. **10**: p. 22.

61. Ewart, K., S. Ho, and A.-A. Chowdhury, *Annotated genome assemblies for Geoscapheus dilatatus, Panesthia cribrata and Neogeoscapheus hanni* 2024: Dryad.
62. Shigenobu, S., et al., *Genomic and transcriptomic analyses of the subterranean termite Reticulitermes speratus: Gene duplication facilitates social evolution*. Proceedings of the National Academy of Sciences, 2022. **119**(3): p. e2110361119.
63. Martelossi, J., et al., *Wood feeding and social living: Draft genome of the subterranean termite Reticulitermes lucifugus (Blattodea; Termitoidae)*. Insect Mol Biol, 2023. **32**(2): p. 118-131.

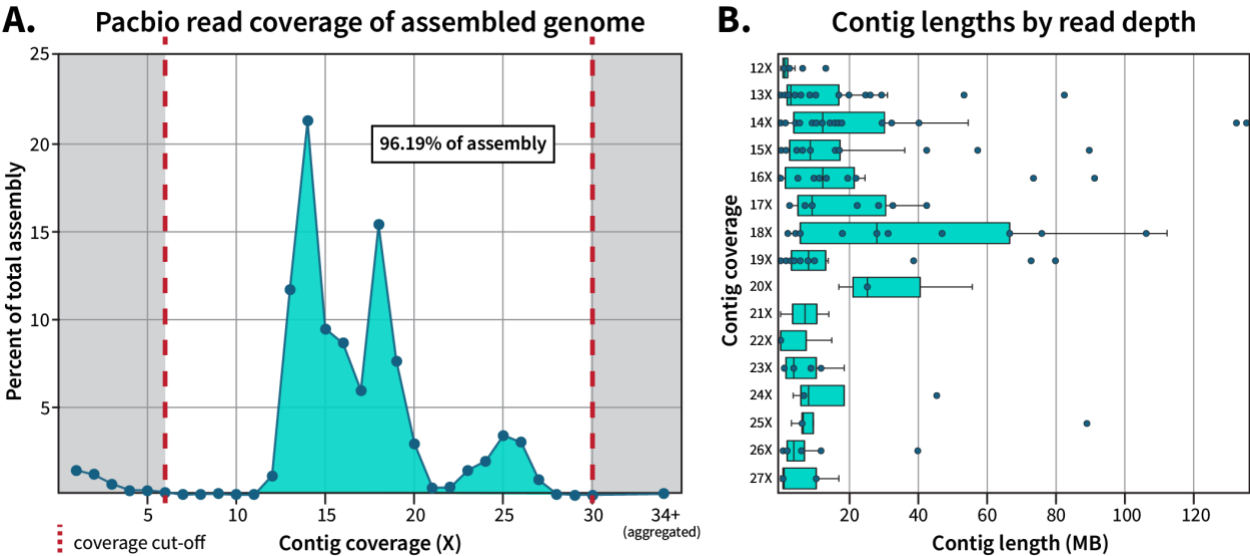
## 9 Text Tables and Figures

**Table 1:** Genome assembly comparison with previous *Periplaneta americana* assemblies available on NCBI.

Assembly name		ASM293952v1	ASM2559430v2	PAMFEO1_prfV1 (this assembly)
Accession		GCA_002939525.1	GCA_025594305.2	GCF_040183065.1
Genome	Total size (Gb)	3.4	3.1	3.2
	Ungapped (Gb)	3.2	3.1	3.2
	GC (%)	35.5	35.42	35.5
Scaffolds	Count (#)	18,601	48	91
	N50 (Mb)	0.3325	150.7	188.1
	L50 (#)	2951	9	8
Contigs	Count (#)	122,589	9217	259
	N50 (Mb)	0.0508	1.9	42.4
	L50 (#)	17,827	416	22
Repeats (%)		57.80	62.90	65.86
Genome BUSCO (% complete)		97.60	94.60	99.70
Protein BUSCO (% complete)		91.20	90.50	97.80
Protein coding genes (#)		21,336	29,939	16,780

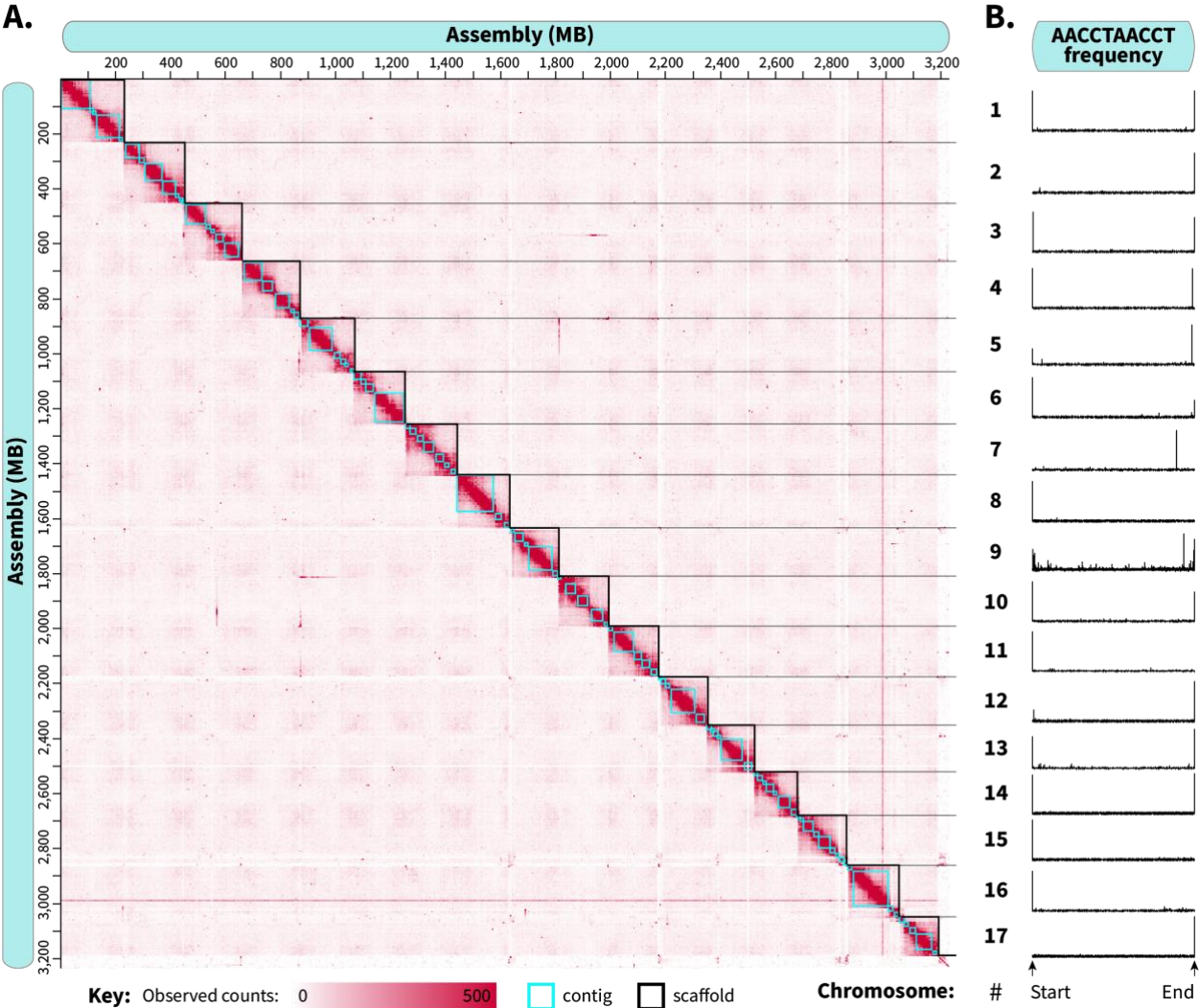


**Figure 1**



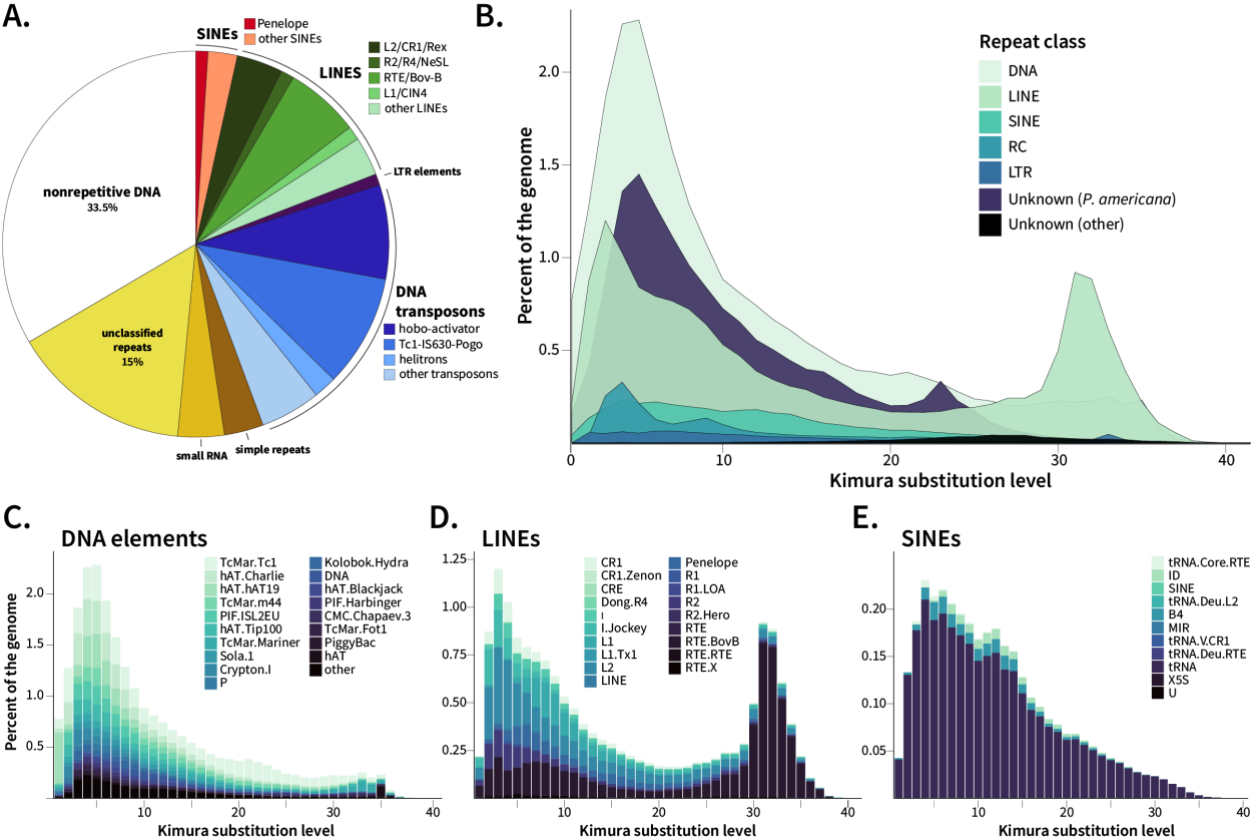
**Figure 1: Quality profile of initial contig and final scaffold hifiiasm assembly. (A)** Pacbio reads were mapped to the primary genome assembly to determine overall coverage of individual contigs, and the percent each coverage level contributed to the total genome size was plotted. **(B)** The length of contigs per coverage level were plotted, with individual contigs represented as dots. Retained contigs with coverage between 6X-11X (n=73) or 28X-30X (n=5) were excluded from plotting in **(B)** due to short lengths.

**Figure 2**



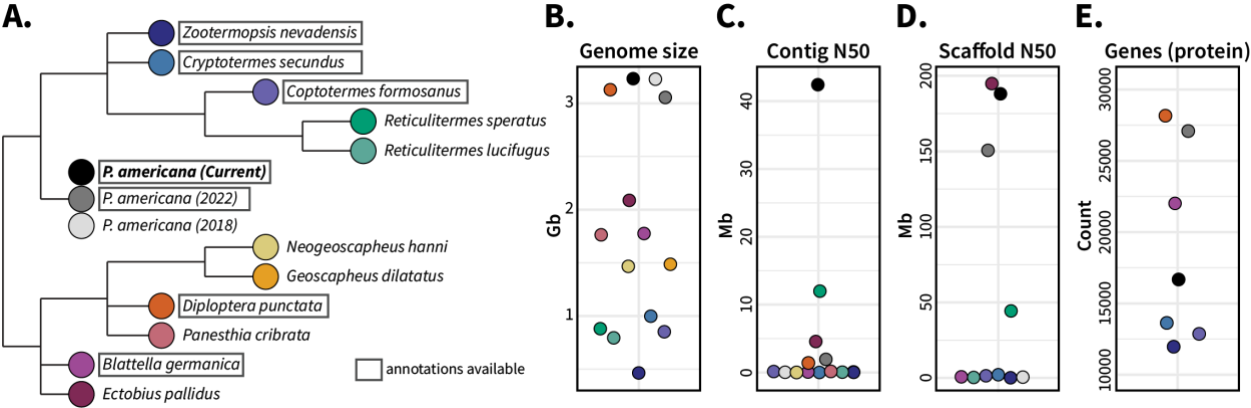
**Figure 2:** Chromatin-contact sequencing produced an assembly with near-chromosomal resolution. Following scaffolding with Yet Another Hi-C Scaffolder (YAHS), scaffold boundaries and contig placement were adjusted in Juicebox to optimize chromatin contacts for the 17 chromosome-level scaffolds. Final chromosomal boundaries are shown in the heatmap in (A), and the occurrence of telomeric sequences within each chromosome are displayed in (B).

**Figure 3**



**Figure 3:** Repeat summary of *P. americana*. RepeatMasker and RepeatModeler were used to identify (A) the abundance of repetitive element families present in this assembly and (B) the relative abundance of each repeat class versus Kimura substitution level. The repeat landscapes for the classes (C) DNA elements, (D) LINEs, and (E) SINEs were further visualized at the repeat subtype level.

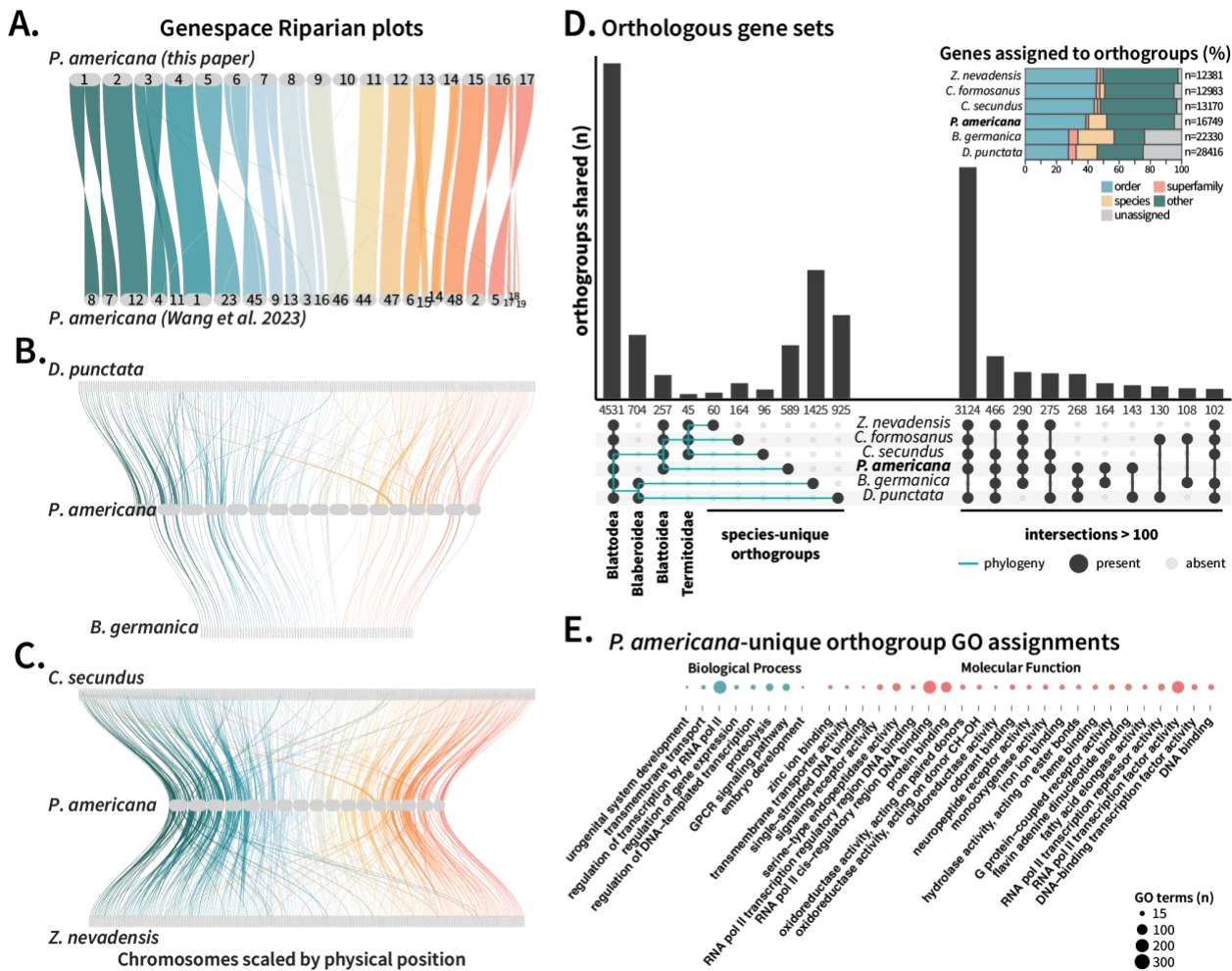
**Figure 4**



**Figure 4:** Available Blattodea genomes from NCBI. Phylogeny of sequenced Blattodea genomes as they appear in the NCBI taxonomy browser is presented in (A), in addition to their (B) genome size, (C) contig N50, (D) scaffold N50, and (E) number of protein-coding genes (if available). Points on the genome statistic plots (B-E) correspond to the colors in (A) and species in (A) with protein annotations available on NCBI are boxed.



## Figure 5

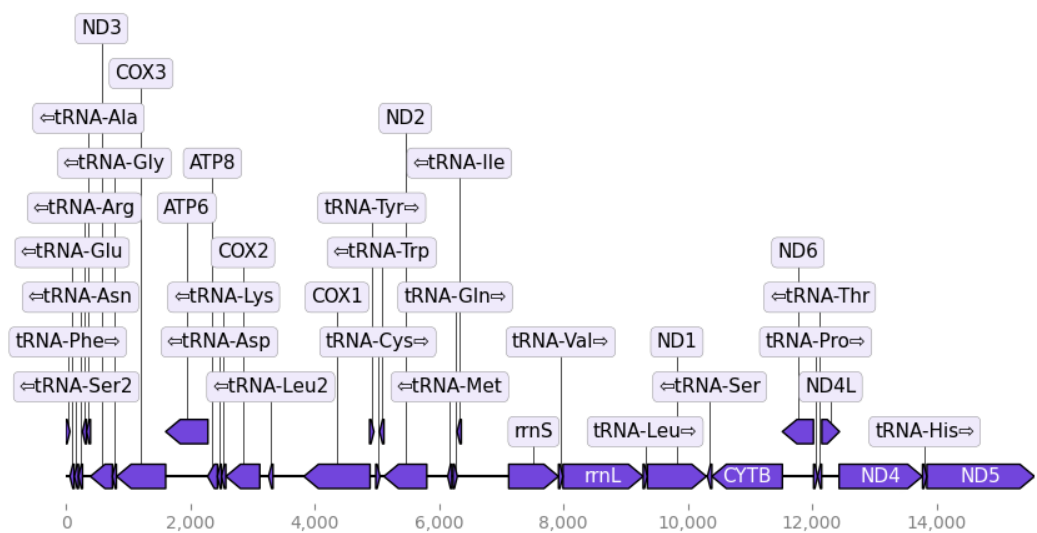


**Figure 5:** Ortholog analysis within Blattodea. Protein GFF3 files available on NCBI for Blattodea species were compared using OrthoFinder and GENESPACE. Riparian plots were generated to assess synteny between the *P. americana* genome presented here and (A) a previous *P. americana* assembly (GCA\_025594305.2), (B) the cockroaches *Diploptera puntata* (GCA\_030220185.1) and *Blattella germanica* (GCA\_003018175.1), and (C) the termites *Cryptotermes secundus* (GCF\_002891405.2) and *Zootermopsis nevadensis* (GCF\_000696155.1). (D) Orthologous gene clusters shared by or unique to the analyzed genomes were visualized via UpSet plot, organized by phylogeny on the left and the next 10 largest sets on the right and summarized by (inset D) the percentage of genes assigned to shared or unique orthogroups. Genes belonging to the 589 orthogroups identified as *P. americana*-unique were analyzed by their associated GO terms, and (E) biological processes and molecular functions with at least 15 occurrences were visualized.

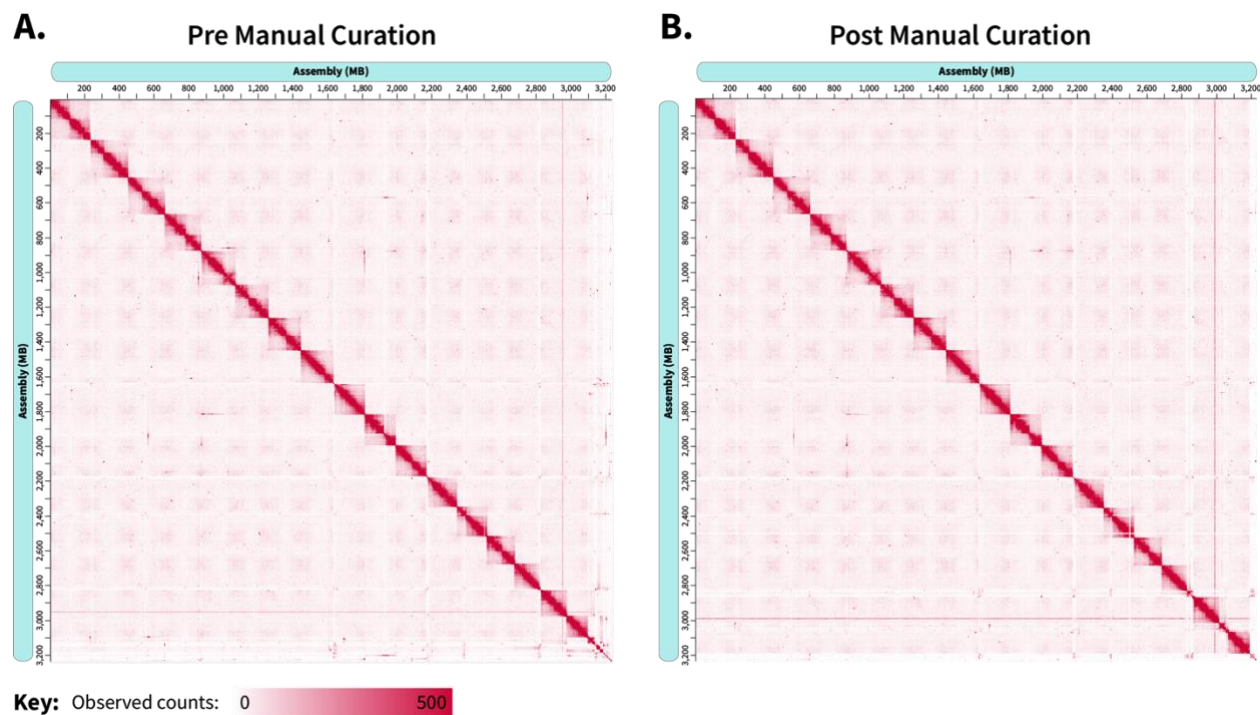
## 10 Supplemental Tables and Figures

**Table S1:** Summary of repeat classifications in *P. americana* that were identified with RepeatModeler and RepeatMasker.

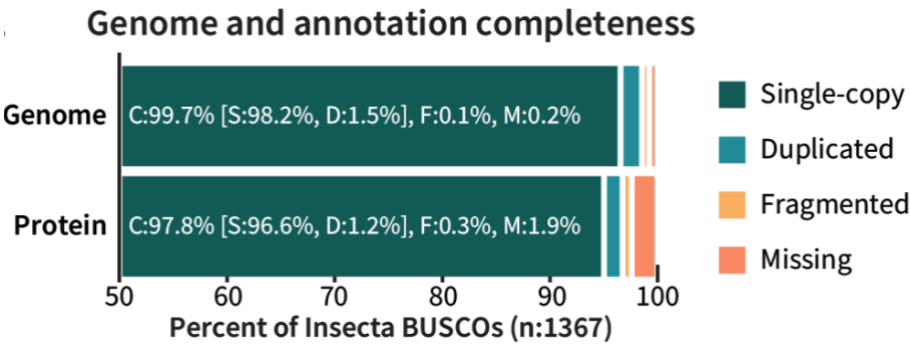
RepeatModeler and RepeatMasker Output			
	Elements (n)	Length (Mb)	Percent (%)
<b>Total Retroelements</b>	<b>2364225</b>	<b>647.91</b>	<b>20.05</b>
<b>SINEs</b>	656509	112.1	3.47
<b>Penelope</b>	134060	33.05	1.02
<b>LINEs</b>	1615375	505.28	15.63
CRE/SLACS	4803	1.86	0.06
L2/CR1/Rex	431509	124.43	3.85
R1/LOA/Jockey	63736	26.12	0.81
R2/R4/NeSL	110731	36.15	1.12
RTE/Bov-B	596056	201.7	6.24
L1/CIN4	105530	37.16	1.15
<b>LTR elements</b>	92341	30.53	0.94
BEL/Pao	26445	5.55	0.17
Ty1/Copia	4703	1.41	0.04
Gypsy/DIRS1	54482	22.9	0.71
Retroviral	0	0	0
<b>Total DNA Transposons</b>	<b>2918518</b>	<b>785.38</b>	<b>24.3</b>
hobo-Activator	799615	254.82	7.88
Tc1-IS630-Pogo	1257068	304.69	9.43
MULE-MuDR	33964	9.84	0.3
PiggyBac	40357	10.56	0.33
Tourist/Harbinger	54352	17.53	0.54
Other	105971	22.03	0.68
<b>Rolling-circles</b>	235926	62.23	1.93
<b>Unclassified</b>	2276913	486.43	15.05
<b>Small RNA</b>	625209	106.1	3.28
<b>Satellites</b>	3110	4.41	0.14
<b>Simple repeats</b>	1768812	125.36	3.88
<b>Low complexity</b>	161181	10.23	0.32



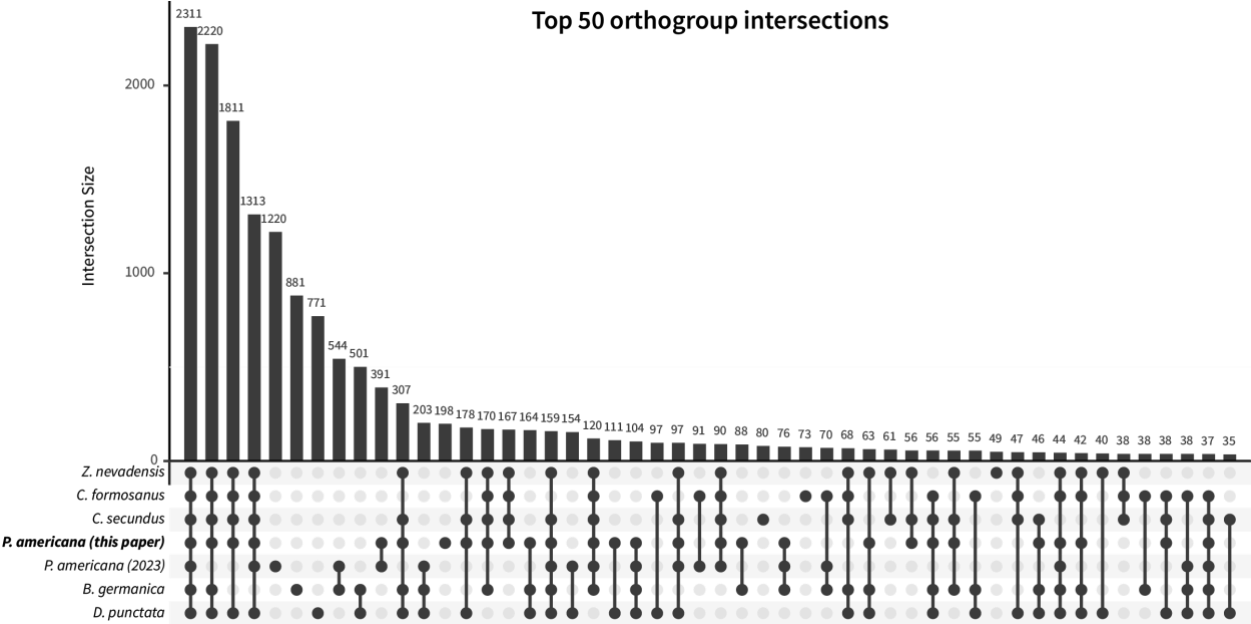
**Figure S1:** Mitogenome annotation output from MitoHiFi.



**Figure S2:** Hi-C contact maps of the assembled genome (A) before and (B) after manual curation using Juicebox.



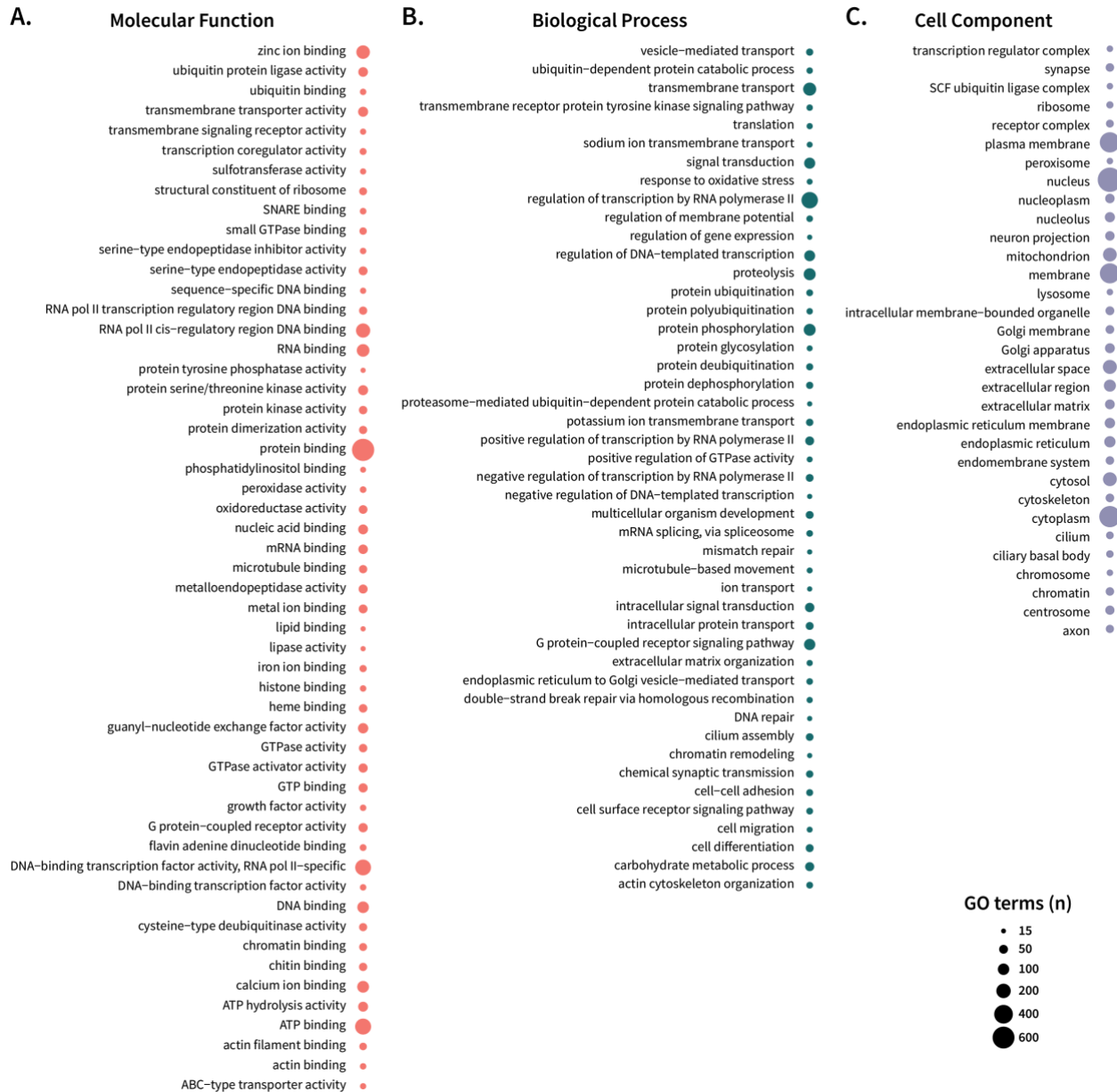
**Figure S3:** Insecta BUSCOs (version 10) identified in this *P. americana* genome assembly and protein annotation.



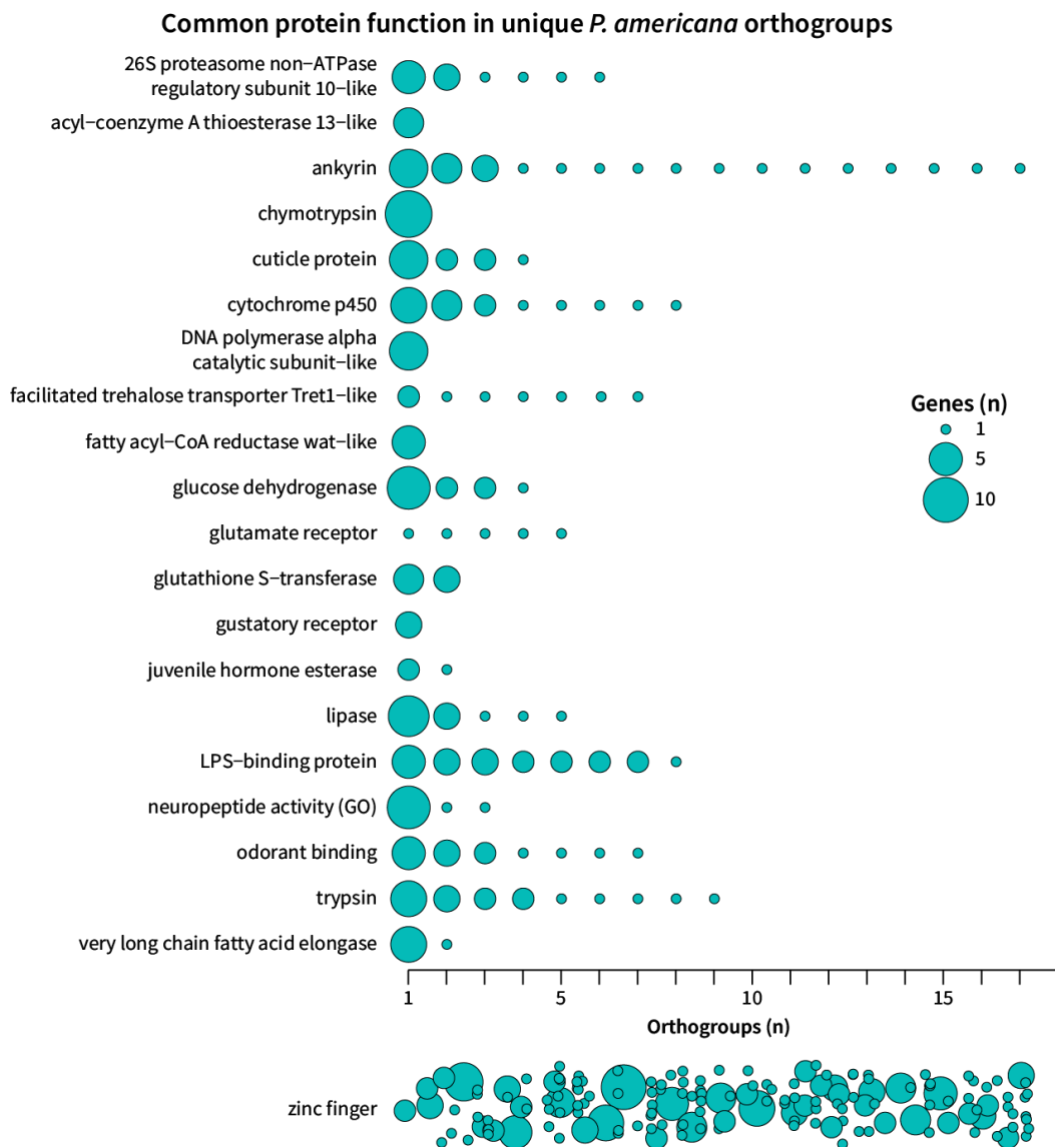
**Figure S4:** UpSet plot of orthogroup overlap between Blattodea species when the previous *P. americana* assembly is included.



# Enriched GO terms in shared Blattodea orthogroups (>15 terms)



**Figure S5:** GO terms enriched in orthogroups shared by all Blattodea species. Genes were included in this analysis if they were sorted into the set of shared orthogroups in Figure 5D and had associated GO assignments. The profiles of enriched GO terms per category were generated in R with the package clusterProfiler. GO terms with abundances of at least 15 genes are reported for (A) molecular functions, (B) biological processes, and (C) cellular components.



548

549 **Figure S6:** The names or GO terms of gene loci belonging to the 589 *P. americana*-unique  
550 orthogroups were collapsed to identify common functions in these genes and visualized with a bubble  
551 plot. Each bubble indicates a separate orthogroup per protein/function and is scaled by the number of  
552 genes sharing the associated name/GO term within the orthogroup. Proteins characterized by zinc  
553 finger motifs are plotted separately due to their large number of orthogroups (157 orthogroups, 248  
554 genes).