Genome report: First whole genome sequence of *Triatoma sanguisuga* (Le Conte, 1855), vector of Chagas disease

Jennifer K. Peterson[1], Madolyn L. MacDonald[2], and Vincenzo A. Ellis[1]

[1]Department of Entomology and Wildlife Ecology, University of Delaware, Newark, DE USA

[2]Bioinformatics and Computational Biology Core, University of Delaware, Newark, DE USA

Keywords: *Triatoma sanguisuga*, triatomine bugs, *Trypanosoma cruzi*, Chagas disease, disease vectors

Abstract

Triatoma sanguisuga is the most widespread triatomine bug species in the United States (US). The species vectors the human parasite *Trypanosoma cruzi*, which causes Chagas disease. Vector-borne Chagas disease is rarely diagnosed in the US, but *T. sanguisuga* has been implicated in a handful of cases. Despite its public health importance, little is known about the genomics or population genetics of *T. sanguisug*a. Here, we used long-read sequencing to assemble the first whole genome sequence for *T. sanguisuga* using DNA extracted from one adult specimen from Delaware. The final size of the genome was 1.162 Gbp with 77.7x coverage. The assembly consisted of 183 contigs with an N50 size of 94.97 Kb. The Benchmarking Universal Single-Copy Ortholog (BUSCO) complete score was 99.1%, suggesting a very complete assembly. Genome-wide GC level was 33.56%, and DNA methylation was 18.84%. The genome consists of 61.4% repetitive DNA and 17,799 predicted coding genes. The assembled *T. sanguisuga* genome was slightly larger than that of Triatominae species *Triatoma infestans* and *Rhodnius prolixus* (949 Mbp with 90.4% BUSCO score and 706 Mbp with 96.5% BUSCO score, respectively). The *T. sanguisuga* genome is the first North American triatomine species genome to be sequenced, and it is the most complete genome yet for any Triatominae species. The *T. sanguisuga* genome will allow for deeper investigations into epidemiologically relevant aspects of this important vector species, including blood feeding, host seeking, and parasite competence, thus providing potential vector-borne disease management targets and strengthening public health preparedness.

I. Introduction

Triatomine bugs ('triatomines') are hematophagous (i.e., blood-feeding) arthropods that feed on a wide variety of vertebrate host species, including humans. Triatomines are of epidemiological interest due to their harborage of the protozoan parasite *Trypanosoma cruzi*, the causative agent of Chagas disease in humans. If left untreated, infection with *T. cruzi* is lifelong and can lead to serious cardiac and gastrointestinal alterations over time (Rassi Jr *et al.* 2010). There are 162 described species of triatomine bugs (159 extant and 3 fossil species; [Alevi *et al.* 2021; Oliveira Correia *et al.* 2022; Zhao *et al.* 2023; Oliveira-Correia *et al.* 2024]), of which 11 are found in the United States (US; Bern *et al.* 2019).

The most widespread triatomine species in the US is *Triatoma sanguisuga* (Fig. 1). The species has been recorded in 24 states from the southernmost states up to about the 42nd parallel, from the Rocky Mountains to the eastern seaboard (Fig. 2). Spanning over one million square miles, *T. sanguisuga* is found throughout several different ecoregions including most of the great plains, the eastern temperate forests, and the tropical wet forests of southern Florida (United States Environmental Protection Agency 2024).

Triatoma sanguisuga is considered an epidemiologically relevant disease vector species in the US because it can be found in domestic and peridomestic habitats and it has been implicated in autochthonous (i.e., vector-borne) Chagas disease cases (Dorn *et al.* 2007; Lynn *et al.* 2020; Beatty and Klotz 2020). Prior scientific studies of *T. sanguisuga* took place predominantly in three of the 24 states in which the species is found: Louisiana, Texas, and Florida. There, *T. cruzi* infection prevalences upwards of 60% have been found in *T. sanguisuga* (Cesa *et al.* 2011; Moudy *et al.* 2014; Curtis-Robles *et al.* 2018) and blood meal analyses have revealed that the species feeds upon a wide range of taxa comprising reptiles, birds, amphibians and mammals, including humans (Waleckx *et al.* 2014; Dumonteil *et al.* 2020, 2024; Balasubramanian *et al.* 2022).

Despite its ubiquity and epidemiologic importance, little is known about the genetic variation in *T. sanguisuga;* individuals are assigned to the species solely based on morphology and geographic location (De Paiva *et al.* 2022). Five *T. sanguisuga* subspecies have been suggested based on morphological variation, and two of these subspecies were eventually assigned to a new species, *T. indictiva* Neiva 1912. The three remaining subspecies were eventually rejected, but uncertainty still exists (De Paiva *et al.* 2022). Two small studies of *T. sanguisuga* found a high level of genetic diversity between populations. A comparison of cytochrome oxidase II mitochondrial gene sequences from 33 *T. sanguisuga* specimens collected in two barrier islands off the southern coast of Georgia revealed 12 distinct haplotypes with no haplotypes shared between populations from different islands, suggesting limited dispersal and genetic exchange (Roden *et al.* 2011). De La Rua *et al.* (2011) investigated intraspecific genetic variation and population structure in 54 *T. sanguisuga* specimens collected in rural New Orleans, Louisiana and found two groups that were genetically divergent enough to represent different subspecies. The high degree of genetic variation discovered in these two studies hints at a wealth of diversity waiting to be discovered in *T. sanguisuga* considering that its geographic range spans multiple ecological regions with varying seasonality, habitat, and other environmental drivers of genetic diversity.

Here we present the *T. sanguisuga* genome, which is the first genome sequenced for any endemic North American triatomine species. Our whole genome sequencing of *T. sanguisuga* will facilitate comparative analyses between populations to resolve questions of species. In addition, insights into the genetic underpinnings of vector behavior and physiology can lead to new vector control targets. Thus the *T. sanguisuga* genome will contribute to genetic studies of epidemiologically relevant characteristics of *T. sanguisuga* such as blood feeding, host seeking, parasite competence, and domiciliation (Abad-Franch and Monteiro 2005; Fitzpatrick *et al.* 2008; Mesquita *et al.* 2015), and in turn increase our public health preparedness.

II. Methods & Materials

Specimen origin

The *T. sanguisuga* specimen used in this study was one of two adult specimens captured within a home in New Castle County, Delaware, as detailed in Peterson *et al.* (2024). The specimens were given to our lab at the University of Delaware by the homeowner. The intestinal contents of both specimens were tested for *T. cruzi* infection via real time PCR as described in Peterson *et al.* (2024). One of the individuals tested positive for *T. cruzi* infection and the other specimen tested negative. The legs and head of the *T. sanguisuga* specimen that tested negative for *T. cruzi* were used for sequencing the genome presented in this study.

DNA extraction and sequencing

High molecular weight (HMW) genomic DNA extraction and purification from the sample was performed using the MagAttract HMW DNA Kit (Qiagen Inc., Venlo, Netherlands) as per manufacturer's instructions. High molecular weight DNA was confirmed using a FemtoPulse (Advanced Analytical Technologies Inc., Ankeny, IA). The HMW DNA (10ug aliquots) were converted to SMRTbell templates using the SMRTbell prep kit 3.0 (Pacific Biosciences, Menlo Park, CA) as per manufacturer's instructions. Briefly, samples were end-repaired and ligated to blunt adaptors. Exonuclease treatment was performed to remove unligated adapters and damaged DNA fragments. Samples were purified using 0.6x AMPureXP beads (Beckman Coulter Inc., Brea, CA). The purified SMRTbell librarieswere eluted in 10 µl of elution buffer. Eluted SMRTbell libraries were size selected on the BluePippin (Sage Science Inc., Beverly, MA) to eliminate library fragments below approximately 10 Kbp. Final library quantification and sizing was carried out on a FemtoPulse (Advanced Analytical Technologies Inc., Ankeny, IA) using 1 µl of library. The amount of primer and polymerase required for the binding reaction was determined using the SMRTbell concentration and library insert size. Primers were annealed and polymerase was bound to the SMRTbell template. Sequencing was performed using the Revio platform (Pacific Biosciences, Menlo Park, CA). The HiFi libraries were run on Revio system 25M SMRT cells using sequencing chemistry 3.0 with 4-hour pre-extension and 30-hour movie time.

Genome assembly methods

Quality control of the HiFi PacBio reads was performed using NanoPlot v1.43.0 (De Coster and Rademakers 2023). Due to the high quality of the reads, no filtering was needed. The *de novo* assembly of the *T. sanguisuga* genome was performed using Flye v2.9 (Kolmogorov *et al.* 2019), Hifiasm v0.19.5 (Cheng *et al.* 2021), and PacBio's assembly pipeline in SMRT Link portal v13.1.0 (https://www.pacb.com/smrt-link/). QUAST v5.1.0 (Mikheenko *et al.* 2018) and BUSCO v5.4.7 (Manni *et al.* 2021) were used to assess the completeness of each assembly and compare them to the genome assemblies of two closely related species, *Rhodnius prolixus* (GCA_000181055.3) and *Triatoma infestans* (GCA_011037195.1). BUSCO was run using the 'hemiptera_odb10' lineage which consisted of 2,510 single-copy orthologs. The best assembly (the Hifiasm primary assembly) was selected for repeat masking and gene annotation. Full commands for all bioinformatics steps are provided in File S1.

Decontamination

The Basic Local Alignment Search Tool (BLASTn v2.15.0+; Altschul *et al.* 1990; Camacho *et al.* 2009) was used to query the contigs against the nt_core database v1.1 to check for contamination. Hits with an e-value cutoff of 0.01 and longer than 200 bp were examined. Contigs with only hits to insects and ribosomal RNA were not considered contamination. All other contigs represented possible contamination and were removed from the final assembly and annotation.

Mitochondria Identification

First, BLAST was used to query the contigs against the existing *T. sanguisuga* mitochondrial sequence (NC_050329.1). Hits were filtered by query coverage to retain contigs with more than 50% of their sequence aligning to the mitochondrial sequence. To confirm the possible mitochondrial contigs and select a representative, MitoHifi v3.2.2 (Uliano-Silva *et al.* 2023) was run via its Singularity container. MitoHifi was also used to annotate and circularize the representative contig and modify it to start at tRNA-Phe. The BLAST alignment between the representative contig and the existing *T. sanguisuga* mitochondrial sequence was used to identify any regions in the representative contig that was not in the reference mitochondrial sequence. Identified significant region(s) were extracted using bedtools and queried against nt and Univec build 10 (https://www.ncbi.nlm.nih.gov/tools/vecscreen/univec/) databases to determine the potential source of the sequence. Any region with no clear source (no significant hits to either database) was removed. The resulting sequence was re-annotated with MitoHifi and visualized using circularMT (Goodman and Carr 2024).

Methylation

PacBio sequencing enables detection of 5-methylated cytosines (5mCs; Flusberg *et al.* 2010). Hifi reads with the 5mC information (5mC tags in an unaligned bam file) were aligned to the Hifiasm primary assembly using Pbmm2 v1.14.0 (https://github.com/PacificBiosciences/pbmm2), a C++ wrapper for minimap2 v2.26 (Li et al., 2018). PacBio's pb-CpG-tools v2.3.2 (https://github.com/PacificBiosciences/pb-CpG-tools) was then used to produce per-site methylation probabilities from the alignment. Global CpG methylation was calculated by dividing the number of methylated CpG cytosines by the number of unmethylated cytosines in the reads that aligned to the genome assembly.

Repeat assembly techniques

Repeats in the Hifiasm primary assembly were identified and modeled using RepeatModeler v2.0.3 (Flynn *et al.* 2020). Repeats from this *de novo* identification were masked using RepeatMasker v4.1.2 (Smit *et al.* 2013). In addition, RepeatMasker was run using Dfam database v3.2 and the parameter '-species Triatoma'. Complex repeats were extracted from the RepeatMasker results and given to the Maker annotation pipeline (Cantarel *et al.* 2008; Campbell *et al.* 2014) for hard-masking (see next section).

Gene finding methods

Annotation was performed using Maker v3.01.04 (Cantarel *et al.* 2008). First, an evidence based round of Maker was run using the transcripts from *R. prolixus* obtained from VectorBase release 68 (Giraldo-Calderón *et al.* 2015) and the 'Triatominae' canonical proteins from UniProt release 2024_3 (Apweiler *et al.* 2004; The UniProt Consortium 2023) as the transcript and protein evidence respectively. For the repeat masking parameters of Maker, the 'rm_gff' parameter was set to the complex repeats GFF from RepeatMasker, 'model_org' was set to 'simple', and 'repeat_protein' was set to Maker's provided 'te_proteins.fasta' file. The transcripts (with added 1,000 bp flanking regions) from this first round of Maker were then used to train gene models using Augustus v3.5.0 (Stanke *et al.* 2006) via BUSCO v5.4.7 long mode. The starting Augustus species was set to 'rhodnius'. The new Augustus model was then provided into a second round of Maker to produce the final set of predicted gene annotations. Putative gene function was assigned by following Maker Support Protocols 2 and 3 (Campbell *et al.* 2014) and functional domains and GO terms were assigned via Interproscan v5.53-87.0 (Blum *et al.* 2020) by following steps 4 and 5 of Basic Protocol 5 (Campbell *et al.* 2014). The predicted genes were then sorted into a high confidence set by filtering out genes with no function assigned and an AED of 1 (no transcript or protein evidence from Maker). BUSCO with the 'hemiptera_odb10' lineage was used to assess the completeness of the high confidence proteins and transcripts. The quality of the annotated proteins was also examined by identifying their *R. prolixus* orthologs using the OrthoVenn3 web service (Sun *et al.* 2023) with the Orthofinder algorithm (Emms and Kelly 2019). All annotation files are located in file S2.

Comparative analysis

Whole genome sequences have been assembled and deposited in the National Center for Biotechnology Information (NCBI) for two other species in the subfamily Triatominae, *R. prolixus* (Bioproject ID PRJNA13648) and *T. infestans* (Bioproject ID PRJNA589079). Although now eliminated from the region, *R. prolixus* is believed responsible for the majority of current Chagas disease cases in Central America (Hashimoto 2012; Peterson *et al.* 2019a, 2019b), and it is still one of the main vectors of Chagas disease in Colombia and Venezuela (Fitzpatrick *et al.* 2008; Méndez-Cardona *et al.* 2022). *Triatoma infestans* is found in the southern cone of South America and it is one of the main vectors of Chagas disease cases in that region (Coura, 2014). Note that two additional genomes of *T. infestans* have reportedly been sequenced (Pita *et al.* 2017a, 2017b, 2018; Mora *et al.* 2023), but the data are not publicly available. For quantitative comparison, we used data from the publicly available *T. infestans* genome.

III. Results and Discussion

Assembly

A total of 91.08 Gbp of sequence was generated with an N50 of 14,271.0 bp and 92.2% of reads above the Q20 quality cutoff and 74.6% above the Q25 cutoff (Table 1). The Hifiasm assembly was selected for annotation based on a comparison of assembly quality statistics and BUSCO scores for assemblies by the assemblers Flye, Hifiasm, and Smrtlink assemblers, as described above (Table 2). The best assembly was the Hifiasm assembly, which was 1.165 Gbp spread over 282 contigs. Decontamination revealed two contigs not pertaining to *T. sanguisuga*; the first contig aligned (>90% identity) with sequences from mitochondria of three fungal species belonging to the order Hypocreales, which includes several entomopathogenic species. The second contig aligned to a tick-borne protozoan species, *Hepatozoon canis*. Both contigs were removed from the Hifiasm assembly, which decreased the genome size by

105,876 bp (Table 2). Analysis of the mitogenome revealed 98 redundant contigs, 97 of which were removed, resulting in a final assembly of 1.162 Gbp spread over 183 contigs with an N50 of 94,972,618 bp and coverage of 77.7x. The genome-wide GC level was 33.56%. All sequence and assembly data, including sequences of the two contaminated contigs that were removed, are available under NCBI BioProject ID PRJNA1140168 and accession number SRR29988702.

Quality, completeness, and coverage

Keeping in mind that accuracy of an assembly without an existing reference for a species is difficult to assess, our assembly appears to be high quality. At 1.162 Gbp with no gaps, the *T. sanguisuga* genome falls within an expected size range relative to *R. prolixus* (706.8 Mbp) and sister species *T. infestans* (949 Mbp; Table 3), suggesting a high level of completeness. The *R. prolixus* genome was the first triatomine species genome ever assembled, in 2015. Given advances in the field during the past ten years, the *R. prolixus* genome published had over 142 million unknown bases (Ns), while our genome did not have any unknown bases, which might explain some of the size difference. The GC percentage was similar between species, varying by less than half a percentage point. Coverage was higher for *T. sanguisuga* than *R. prolixus* (77.7x vs 8.3x, respectively, Table 2) but lower than *T. infestans* (200.0x). Our assembly is the most contiguous of the three species (183 contigs for *T. sanguisuga* vs. 14,951 and 16,511 for *T. infestans* and *R. prolixus*, respectively), with an N50 that is roughly 87 times larger than *R. prolixus* and 873 times larger than *T. infestans* (Table 2). At 99.1%, our BUSCO score indicates excellent resolution compared to that of the other two *Triatominae* genomes (Figure 3).

Genome annotation

We detected 17,799 putative protein-coding genes using the Maker genome annotation pipeline (Table 4). The final protein BUSCO score was 94.4% complete (92.4% single copy and 2.0% duplicates). This number is comparable to that of *R. prolixus*, which was predicted to have 15,456 protein-coding genes. In an analysis of the *R. prolixus* reference proteins using the web service OrthoVenn3 (Sun *et al.* 2023), 9,652 overlapping orthologous protein clusters were found, with 413 non-overlapping clusters found in *T. sanguisuga* and 300 in *R. prolixus* (Fig. 4). Annotations for *T. infestans* were not available on NCBI or Vectorbase, so a comparison with this species was not possible.

Repetitive DNA

Repeat analysis using the RepeatModeler and RepeatMasker software revealed that the *T. sanguisuga* genome consists of 61.7% repetitive DNA, with 40.58% interspersed repeats, 2.35% simple repeats, and 0.46% low complexity repeats (Table 5).

Mitochondrial structure

Ninety-eight contigs from the original Hifiasm assembly were identified as redundant copies of the *T. sanguisuga* mitogenome. The representative mitochondrial contig selected and circularized by MitoHifi was aligned to the existing *T. sansguisuga* mitochondrial genome (NC_050329.1). BLAST hits between the representative contig and the existing *T. sanguisuga* mitochondrial sequence showed that there was an additional ~1,500 bp in the representative contig. This approximate 1,500 bp was extracted and queried against nt and Univec build 10 databases. With no significant hits to either database, the region was subsequently removed from the contig. The final annotated mitogenome assembly was a single circular contig measuring 15,542 bp in length with 14 protein-coding genes, 2 rRNAs, and 22 tRNAs (Fig. 5). The sequence matched the reference sequence with 95.49% identity over 99% of its length.

IV. Conclusions

Here, we present the first genome sequence for *Triatoma sanguisuga*, which is also the first whole genome sequence for any US or North American triatomine species. The *T. sanguisuga* genome will facilitate the identification of genetic differences between *T. sanguisuga* populations, plausibly with relevance to epidemiologically important traits. Vector reference genomes contribute to our ability to carry out genetic investigations of physiological and behavioral attributes of disease vectors such as blood feeding, host seeking, and parasite competence. Findings from such studies can be used to guide vector-borne disease management strategies and in turn, strengthen public health preparedness.

IV. Data Availability Statement:

Supplementary file one (S1), submitted with this manuscript, contains the full commands for all bioinformatics steps. Supplementary file 2 (S2) is available on GSA FigShare and contains the annotation files. Sequence data and the genome assembly are publicly available on NCBI under BioProject ID PRJNA1140168 and Accession number SRR2998870.

V. Acknowledgments

We are grateful to Brewster Kingham, Erin Bernberg, and Olga Shevchenko from the University of Delaware Sequencing and Genotyping Center for assistance with experimental design, sample processing methods, and PacBio sequencing. We would also like to thank Shawn Polson from the University of Delaware Bioinformatics Data Science Core Facility and the Delaware homeowner who provided the *Triatoma sanguisuga* sample.

VI. Conflict of Interest

None

VII. Funder Information

Access to the University of Delaware Sequencing and Bioinformatics Core Centers was supported by the Institutional Development Award (IDeA) from the National Institute of Health's National Institute of General Medical Sciences, under grant number P20GM103446. Support from the University of Delaware CBCB Bioinformatics Data Science Core Facility (RRID:SCR_017696) including use of the BIOMIX and BioStore computational resources was made possible through funding from Delaware INBRE (NIGMS P20 GM103446), NIH Shared Instrumentation Grant (S10 OD028725), the State of Delaware, and the Delaware Biotechnology Institute.

VIII. Figure legends

Figure 1. *Triatoma sanguisuga* specimen used in genome sequencing. Image first appeared in figure 2b (Peterson *et al.* 2024). Photo taken by Solomon Hendrix.

Figure 2. Geographical distribution of *Triatoma sanguisuga.* States with recorded observations of the species are shown in red. Map created with mapchart.net.

Figure 3. Genome BUSCO scores for assembled *Triatominae* genomes on NCBI. *Triatoma sanguisuga* shown is from this study.

Figure 4. Protein count and cluster count overlap for *Triatoma sanguisuga* (orange) and *Rhodnius prolixus* (blue). Large circles indicate protein counts and small circles indicate orthologous protein cluster counts.

Figure 5. Fig. 5. Annotated mitogenome for *Triatoma sanguisuga*. Protein coding genes are shown in orange, tRNAs are shown in light blue, and rRNAs are shown in gray. Mitogenome was drawn using circularMT (Goodman and Carr 2024). Drawing of *T. sanguisuga* by Laura Ulrich (used with Ms. Ulrich's permission.)

IX. Tables

Quality cutoff	# reads above cutoff	% of reads above cutoff	Mb of reads above cutoff
Q10	6,532,647	100.0%	91,075.0 Mb
Q15	6,529,241	99.9%	91,027.9 Mb
Q20	6,021,510	92.2%	83,537.3 Mb
Q25	4,872,187	74.6%	66,254.9 Mb
Q30	3,438,652	52.6%	44,709.0 Mb

Table 1. Quality cutoff data for *T. sanguisuga* sequencing reads.

	Flye	Smrtlink	Hifiasm (Draft 1)	Hifiasm Decontaminated (Draft 2)	Final assembly
ASSEMBLY					
Length (bp)	2,079,745,197	1,180,738,419	1,165,190,363	1,165,084,487	1,162,099,166
Contigs	9,026	1,068	282	280	183
Largest contig	7,311,449	24,892,402	144,379,499	144,379,499	144,379,499
Min contig length	1,021	1,811	9,031	9,031	9,031
Mean contig length	230,417.2	1,105,560.3	4,131,880.7	4,161,016	6,350,268.7
N50	388,970	6,977,916	94,972,618	94,972,618	94,972,618
L50	1,563	52	5	5	5
BUSCO					
Complete	98.9%	98.8%	99.1%	99.1%	99.1%
Single	36.4%	97.8%	97.8%	97.8%	97.8%
Duplicate	62.5%	1.1%	1.3%	1.3%	1.3%
Fragment	0.7%	0.7%	0.6%	0.6%	0.6%
Missing	0.4%	0.4%	0.3%	0.3%	0.3%
n	2,510	2,510	2,510	2,510	2,510

Table 2. Quality statistics for assemblies of the *T. sanguisuga* genome by three different *de novo* assemblers. The Hifiasm draft one assembly was selected as the best assembly and thereafter underwent decontamination to produce the draft two assembly, followed by removal of 97 redundant contigs pertaining to the *T. sanguisuga* mitochondria to produce the final assembly.

	T. sanguisuga	*T. infestans*	*R. prolixus*
Summary statistics			
Contigs	183	14,951	16,511
Contigs >5000bp	183	14,849	4,598
Contigs >10,000bp	182	13,959	3,097
Contigs >25,000bp	151	9,815	1,862
Contigs >50,000bp	48	5,919	1,089
Largest contig	144,379,499 bp	1,054,224 bp	13,425,595
Total Length	1,162 Mb	949Mb	707Mb
N50	94,972,618	108,830	1,088,772
GCs %	33.56	34.03	33.94
Coverage	77.7x	200.0x	8.3x
Mismatches			
#Ns per 100kbp	0	342.49	20,118
#Ns	0	3,251,881	142,194,158

Table 3. Comparison of quantitative genome characteristics between *T. sanguisuga* (from this study), *T. infestans* and *R. prolixus.*

Feature	Count
Genes	17,799
Transcripts/Proteins	17,799
Exons	102,331
Introns	84,532
Mean gene length	8,010
Mean exons per transcript	5.7
Single exon transcripts	3,348
Proteins with predicted function	13,696
Lower Confidence Proteins	9,723

Table 4. Quantitative summary of the annotated final genome of *T. sanguisuga.*

Name	Number of elements*	Length	Percent
Retroelements	869,571	214,680,453 bp	18.47%
SINEs:	66,050	6,690,116 bp	0.58%
Penelope class	88,144	115,173,646 bp	1.31 %
LINE class:	774,151	193,638,471 bp	16.66%
CRE/SLACS	0	0bp	0.00%
L2/CR1/Rex	34,854	20,442,282 bp	1.76%
R1/LOA/Jockey	390,559	77,649,217 bp	6.68%
R2/R4/NeSL	8,077	3,140,410 bp	0.27%
RTE-Bov-B	100,694	35,959,107 bp	3.09%
L1/CIN4	91	49,474 bp	0.00%
LTR elements:	29,370	14,351,866 bp	1.23%
BEL/Pao	6,426	2,789,295 bp	0.24%
Ty1/Copia	8,706	1,639,481 bp	0.14%
Gypsy/DIRS1	14,145	9,880,325 bp	0.85%
Retroviral	0	0 bp	0.00%
DNA transposons	509,635	84,165,895 bp	7.24%
hobo-Activator	50,682	15,582,243 bp	1.34%
Tc1-IS630-Pogo	154,823	27,948,296 bp	2.40%
En-Spm	0	0 bp	0.00%
MuDR-IS905	0	0 bp	0.00%
PiggyBac	3,233	1,198,104 bp	0.10%
Tourist/Harbinger	2,099	674,150 bp	0.06%
Other (Mirage, P-element, Transib)	189	86,932 bp	0.01%
Rolling-circles	1,198,797	209,049,142 bp	17.99%
Unclassified:	649,059	169,616,531 bp	14.60%
Total interspersed repeats:		468,462,879 bp	40.31%
Small RNA:	24,593	4,690,086 bp	0.40%
Satellites:	7,833	1,134,749 bp	0.10%
Simple repeats:	693,752	27,246,855 bp	2.34%
Low complexity:	107,640	5,319,819 bp	0.46%

Table 5. Interspersed repeats in the *T. sanguisuga* genome. *Most repeats fragmented by insertions or deletions have been counted as one element.

X. Literature Cited

Abad-Franch, F., and F. A. Monteiro, 2005 Molecular research and the control of Chagas disease vectors. An. Acad. Bras. Ciênc. 77: 437–454.

Alevi, K. C. C., J. de Oliveira, D. da Silva Rocha, and C. Galvão, 2021 Trends in Taxonomy of Chagas Disease Vectors (Hemiptera, Reduviidae, Triatominae): From Linnaean to Integrative Taxonomy. Pathogens 10: 1627.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, 1990 Basic local alignment search tool. Journal of Molecular Biology 215: 403–410.

Apweiler, R., A. Bairoch, C. H. Wu, W. C. Barker, B. Boeckmann *et al.*, 2004 UniProt: the Universal Protein knowledgebase. Nucleic Acids Res 32: D115–D119.

Balasubramanian, S., R. Curtis-Robles, B. Chirra, L. D. Auckland, A. Mai *et al.*, 2022 Characterization of triatomine bloodmeal sources using direct Sanger sequencing and amplicon deep sequencing methods. Sci Rep 12: 10234.

Beatty, N. L., and S. A. Klotz, 2020 Autochthonous Chagas Disease in the United States: How Are People Getting Infected? Am J Trop Med Hyg 103: 967–969.

Bern, C., L. A. Messenger, J. D. Whitman, and J. H. Maguire, 2019 Chagas Disease in the United States: a Public Health Approach. Clin Microbiol Rev 33: e00023-19.

Blum, M., H.-Y. Chang, S. Chuguransky, T. Grego, S. Kandasaamy *et al.*, 2020 The InterPro protein families and domains database: 20 years on. Nucleic Acids Res 49: D344–D354.

Camacho, C., G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos *et al.*, 2009 BLAST+: architecture and applications. BMC Bioinformatics 10: 421.

Campbell, M. S., C. Holt, B. Moore, and M. Yandell, 2014 Genome Annotation and Curation Using MAKER and MAKER-P. Curr Protoc Bioinformatics 48: 4.11.1-4.11.39.

Cantarel, B. L., I. Korf, S. M. C. Robb, G. Parra, E. Ross *et al.*, 2008 MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res 18: 188–196.

Cesa, K., K. A. Caillouët, P. L. Dorn, and D. M. Wesson, 2011 High *Trypanosoma cruzi* (Kinetoplastida: Trypanosomatidae) Prevalence in *Triatoma sanguisuga* (Hemiptera: Redviidae) in Southeastern Louisiana. J Med Entomol 48: 1091–1094.

Cheng, H., G. T. Concepcion, X. Feng, H. Zhang, and H. Li, 2021 Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat Methods 18: 170–175.

Coura, J. R., 2014 The main sceneries of Chagas disease transmission. The vectors, blood and oral transmissions - A comprehensive review. Mem. Inst. Oswaldo Cruz 110: 277–282.

Curtis-Robles, R., L. D. Auckland, K. F. Snowden, G. L. Hamer, and S. A. Hamer, 2018 Analysis of over 1500 triatomine vectors from across the US, predominantly Texas, for *Trypanosoma cruzi* infection and discrete typing units. Infection, Genetics and Evolution 58: 171–180.

De Coster, W., and R. Rademakers, 2023 NanoPack2: population-scale evaluation of long-read sequencing data (C. Alkan, Ed.). Bioinformatics 39: btad311.

De La Rua, N., L. Stevens, and P. L. Dorn, 2011 High genetic diversity in a single population of Triatoma sanguisuga (LeConte, 1855) inferred from two mitochondrial markers: Cytochrome b and 16S ribosomal DNA. Infection, Genetics and Evolution 11: 671–677.

De Paiva, V. F., T. Belintani, J. De Oliveira, C. Galvão, and J. A. Da Rosa, 2022 A review of the taxonomy and biology of Triatominae subspecies (Hemiptera: Reduviidae). Parasitol Res 121: 499–512.

Dorn, P. L., L. Perniciaro, M. J. Yabsley, D. M. Roellig, G. Balsamo *et al.*, 2007 Autochthonous Transmission of *Trypanosoma cruzi*, Louisiana. Emerging infectious diseases 13: 605–607.

Dumonteil, E., H. Pronovost, E. F. Bierman, A. Sanford, A. Majeau *et al.*, 2020 Interactions among *Triatoma sanguisuga* blood feeding sources, gut microbiota and *Trypanosoma cruzi* diversity in southern Louisiana. Mol Ecol 29: 3747–3761.

Dumonteil, E., W. Tu, F. A. Jiménez, and C. Herrera, 2024 Ecological interactions of *Triatoma sanguisuga* (Hemiptera: Reduviidae) and risk for human infection with *Trypanosoma cruzi* (Kinetoplastida: Trypanosomatidae) in Illinois and Louisiana (G. Hamer, Ed.). Journal of Medical Entomology tjae017.

Emms, D. M., and S. Kelly, 2019 OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biology 20: 238.

Fitzpatrick, S., M. D. Feliciangeli, M. J. Sanchez-Martin, F. a Monteiro, and M. a Miles, 2008 Molecular genetics reveal that silvatic Rhodnius prolixus do colonise rural houses. PLoS neglected tropical diseases 2: e210.

Flusberg, B. A., D. Webster, J. Lee, K. Travers, E. Olivares *et al.*, 2010 Direct detection of DNA methylation during single-molecule, real-time sequencing. Nat Methods 7: 461–465.

Flynn, J. M., R. Hubley, C. Goubert, J. Rosen, A. G. Clark *et al.*, 2020 RepeatModeler2 for automated genomic discovery of transposable element families. Proc. Natl. Acad. Sci. U.S.A. 117: 9451–9457.

Giraldo-Calderón, G. I., S. J. Emrich, R. M. MacCallum, G. Maslen, E. Dialynas *et al.*, 2015 VectorBase: an updated bioinformatics resource for invertebrate vectors and other organisms related with human diseases. Nucleic Acids Res 43: D707–D713.

Goodman, S. J., and I. M. Carr, 2024 Drawing mitochondrial genomes with circularMT (R. Schwartz, Ed.). Bioinformatics 40: btae450.

Hashimoto, K., 2012 Elimination of Rhodnius prolixus in Central America. 10.

Kolmogorov, M., J. Yuan, Y. Lin, and P. A. Pevzner, 2019 Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol 37: 540–546.

Lynn, M. K., B. H. Bossak, P. A. Sandifer, A. Watson, and M. S. Nolan, 2020 Contemporary autochthonous human Chagas disease in the USA. Acta Tropica 205: 105361.

Manni, M., M. R. Berkeley, M. Seppey, F. A. Simão, and E. M. Zdobnov, 2021 BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes (J. Kelley, Ed.). Molecular Biology and Evolution 38: 4647–4654.

Méndez-Cardona, S., M. I. Ortiz, M. C. Carrasquilla, P. Fuya, F. Guhl *et al.*, 2022 Altitudinal distribution and species richness of triatomines (Hemiptera:Reduviidae) in Colombia. Parasites Vectors 15: 450.

Mesquita, R. D., R. J. Vionette-Amaral, C. Lowenberger, R. Rivera-Pomar, F. A. Monteiro *et al.*, 2015 Genome of Rhodnius prolixus, an insect vector of Chagas disease, reveals unique adaptations to hematophagy and parasite infection. Proceedings of the National Academy of Sciences 112: 14936–14941.

Mikheenko, A., A. Prjibelski, V. Saveliev, D. Antipov, and A. Gurevich, 2018 Versatile genome assembly evaluation with QUAST-LG. Bioinformatics 34: i142–i150.

Mora, P., S. Pita, E. E. Montiel, J. M. Rico-Porras, T. Palomeque *et al.*, 2023 Making the Genome Huge: The Case of Triatoma delpontei, a Triatominae Species with More than 50% of Its Genome Full of Satellite DNA. Genes 14: 371.

Moudy, R. M., S. Michaels, S. B. Jameson, B. Londono, V. Lopez *et al.*, 2014 Factors Associated With Peridomestic *Triatoma sanguisuga* (Hemiptera: Reduviidae) Presence in Southeastern Louisiana. J Med Entomol 51: 1043–1050.

Oliveira Correia, J. P. S., H. R. Gil-Santana, C. Dale, and C. Galvão, 2022 *Triatoma guazu* Lent and Wygodzinsky Is a Junior Synonym of *Triatoma williami* Galvão, Souza and Lima. Insects 13: 591.

Oliveira-Correia, J. P. S., J. de Oliveira, H. R. Gil-Santana, D. da Silva Rocha, and C. Galvão, 2024 Taxonomic reassessment of *Rhodnius zeledoni* Jurberg, Rocha & Galvão: a morphological and morphometric analysis comparing its taxonomic relationship with *Rhodnius domesticus* Neiva & Pinto. BMC Zool 9: 6.

Peterson, J. K., K. Hashimoto, K. Yoshioka, P. L. Dorn, N. L. Gottdenker *et al.*, 2019a Chagas Disease in Central America: Recent Findings and Current Challenges in Vector Ecology and Control. Curr Trop Med Rep 6: 76–91.

Peterson, J. K., J. Hoyos, C. R. Bartlett, N. L. Gottdenker, B. Kunkel *et al.*, 2024 First report of Chagas disease vector species *Triatoma sanguisuga* (Hemiptera: Reduviidae) infected with *Trypanosoma cruzi* in Delaware. American Journal of Tropical Medicine and Hygiene Accepted:

Peterson, J. K., K. Yoshioka, K. Hashimoto, A. Caranci, N. Gottdenker *et al.*, 2019b Chagas Disease Epidemiology in Central America: an Update. Curr Trop Med Rep 6: 92–105.

Pita, S., P. Mora, J. Vela, T. Palomeque, A. Sánchez *et al.*, 2018 Comparative Analysis of Repetitive DNA between the Main Vectors of Chagas Disease: Triatoma infestans and Rhodnius prolixus. Int J Mol Sci 19: 1277.

Pita, S., F. Panzera, P. Mora, J. Vela, Á. Cuadrado *et al.*, 2017a Comparative repeatome analysis on Triatoma infestans Andean and Non-Andean lineages, main vector of Chagas disease. PLOS ONE 12: e0181635.

Pita, S., F. Panzera, J. Vela, P. Mora, T. Palomeque *et al.*, 2017b Complete mitochondrial genome of Triatoma infestans (Hemiptera, Reduviidae, Triatominae), main vector of Chagas disease. Infection, Genetics and Evolution 54: 158–163.

Rassi Jr, A., A. Rassi, and J. A. Marin-Neto, 2010 Chagas disease. The Lancet 375: 1388–1402.

Roden, A. E., D. E. Champagne, and B. T. Forschler, 2011 Biogeography of Triatoma sanguisuga (Hemiptera: Reduviidae) on Two Barrier Islands off the Coast of Georgia, United States. jnl. med. entom. 48: 806–812.

Smit, A., R. Hubley, and P. Green, 2013 RepeatMasker Open-4.0.

Stanke, M., O. Keller, I. Gunduz, A. Hayes, S. Waack *et al.*, 2006 AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Res 34: W435–W439.

Sun, J., F. Lu, Y. Luo, L. Bie, L. Xu *et al.*, 2023 OrthoVenn3: an integrated platform for exploring and visualizing orthologous data across genomes. Nucleic Acids Research 51: W397–W403.

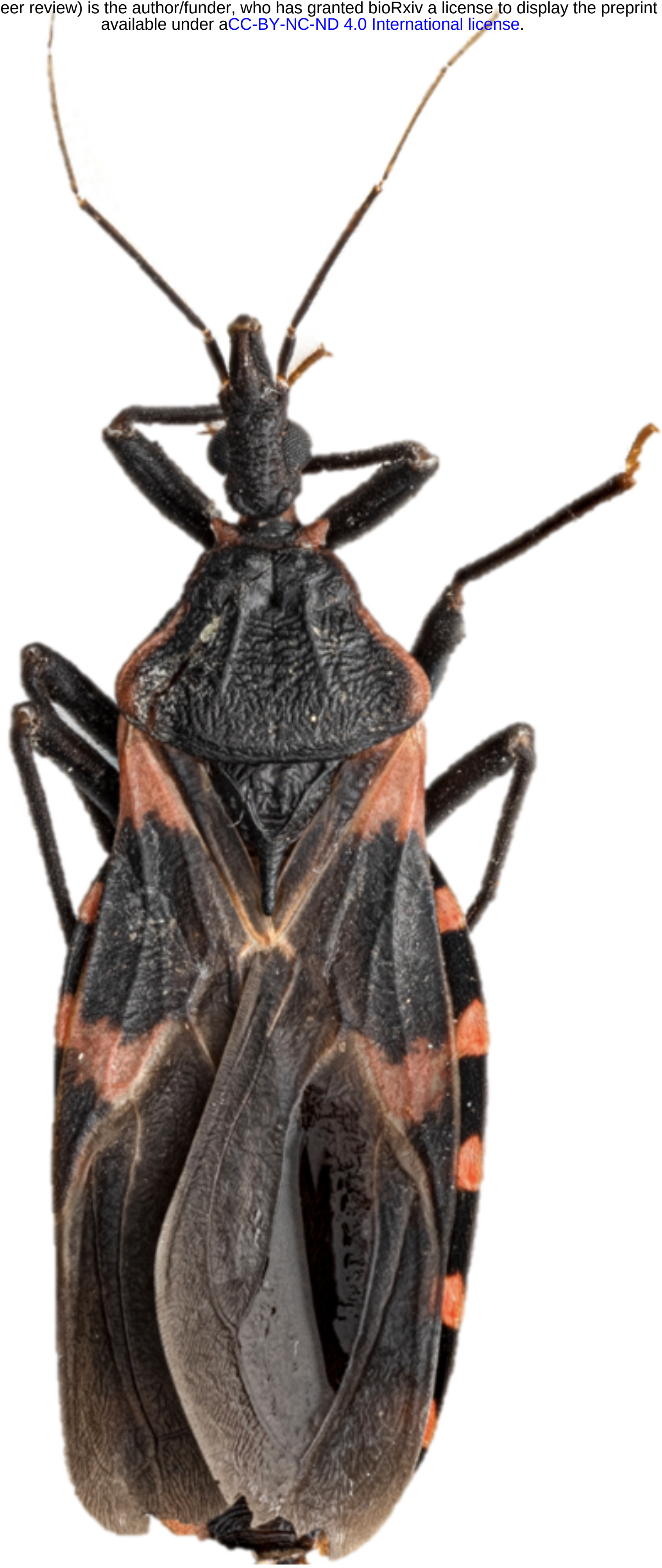
The UniProt Consortium, 2023 UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Research 51: D523–D531.

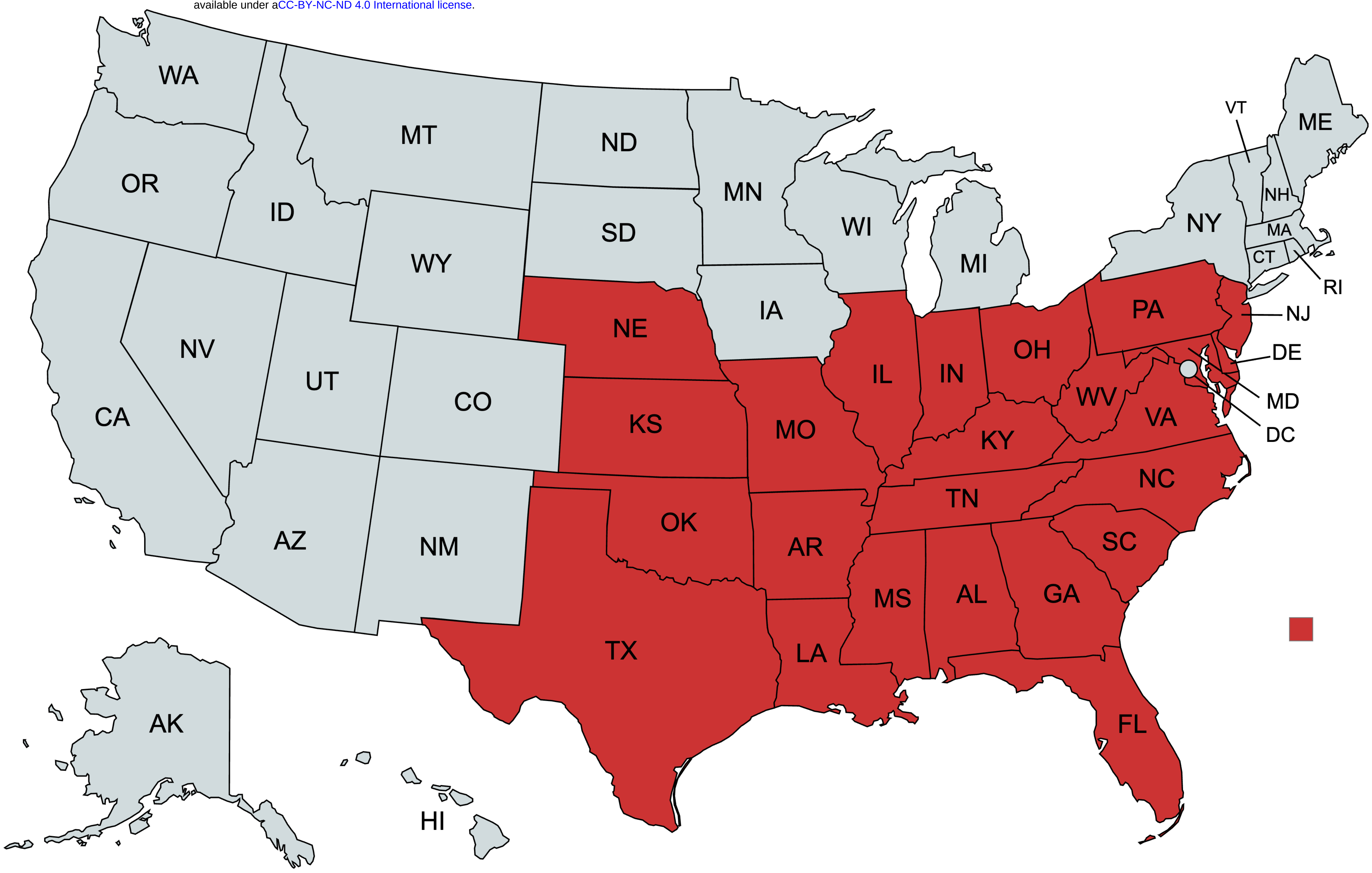
Uliano-Silva, M., J. G. R. N. Ferreira, K. Krasheninnikova, Darwin Tree of Life Consortium, G. Formenti *et al.*, 2023 MitoHiFi: a python pipeline for mitochondrial genome assembly from PacBio high fidelity reads. BMC Bioinformatics 24: 288.

United States Environmental Protection Agency, 2024 Ecoregions of North America.

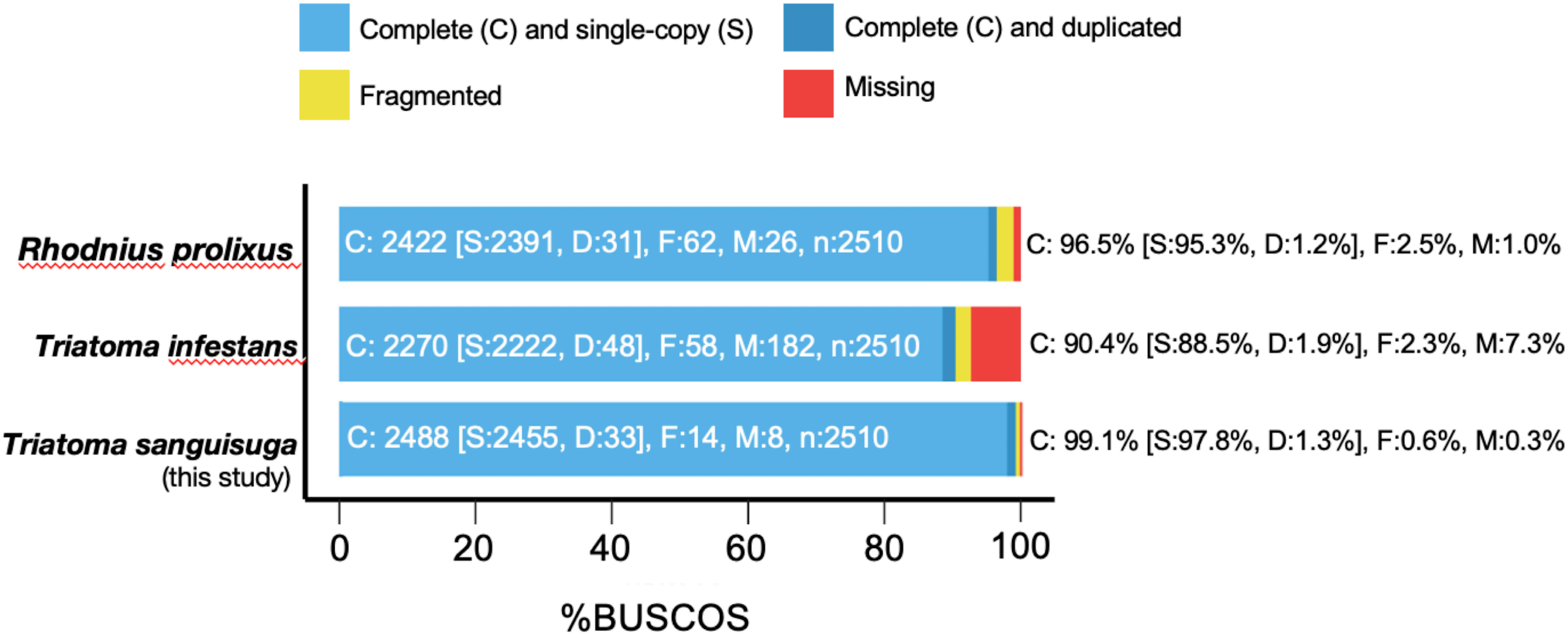
Waleckx, E., J. Suarez, B. Richards, and P. L. Dorn, 2014 *Triatoma sanguisuga* Blood Meals and Potential for Chagas Disease, Louisiana, USA. Emerg. Infect. Dis. 20: 2141–2143.

Zhao, Y., M. Fan, H. Li, and W. Cai, 2023 Review of Kissing Bugs (Hemiptera: Reduviidae: Triatominae) from China with Descriptions of Two New Species. Insects 14: 450.

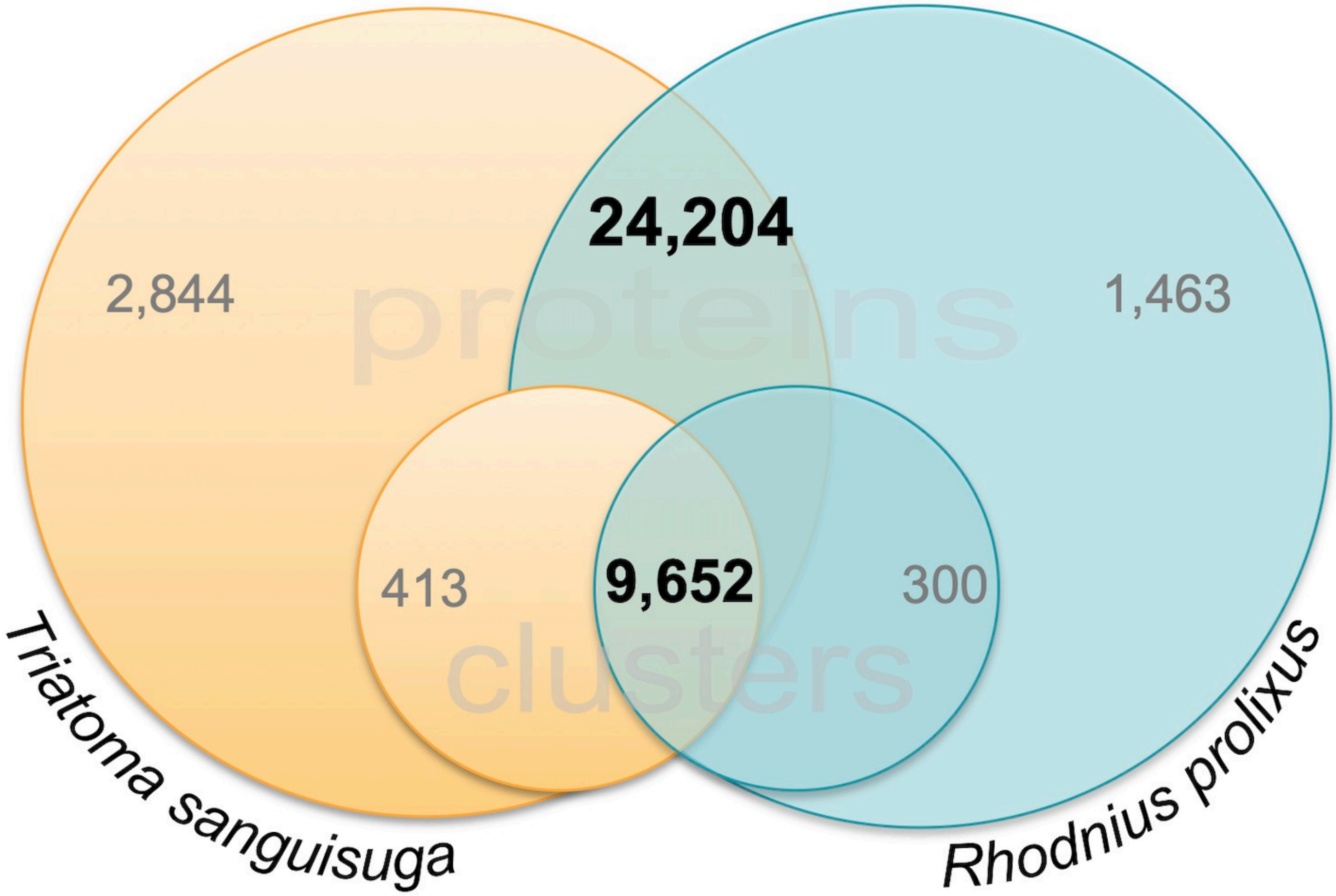


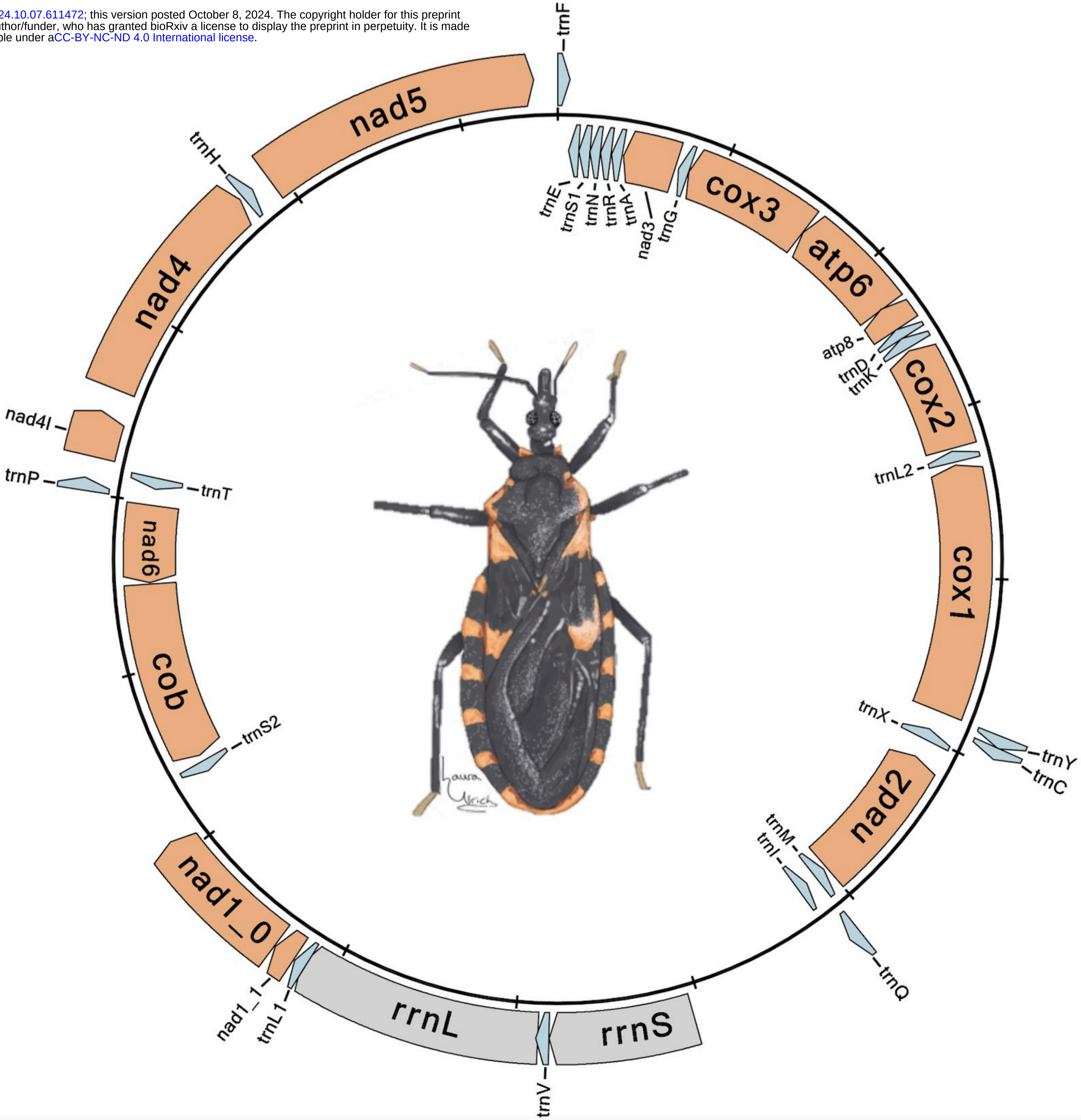


Created with mapchart.net



Complete (C) and single-copy (S)
Complete (C) and duplicated
Fragmented
Missing
Rhodnius prolixus
C: 2422 [S:2391, D:31], F:62, M:26, n:2510
C: 96.5% [S:95.3%, D:1.2%], F:2.5%, M:1.0%
Triatoma infestans
C: 2270 [S:2222, D:48], F:58, M:182, n:2510
C: 90.4% [S:88.5%, D:1.9%], F:2.3%, M:7.3%
Triatoma sanguisuga
(this study)
C: 2488 [S:2455, D:33], F:14, M:8, n:2510
C: 99.1% [S:97.8%, D:1.3%], F:0.6%, M:0.3%
0
20
40
60
80
100
%BUSCOS


24,204
2,844
1,463
proteins
413
9,652
300
clusters
Triatoma sanguisuga
Rhodnius prolixus


trnF
trnE
trnS1
trnN
trnR
trnA
nad3
trnG
cox3
atp6
atp8
trnD
trnK
cox2
trnL2
cox1
trnX
trnY
trnC
nad2
trnM
trnI
trnQ
rrnS
trnV
rrnL
trnL1
nad1_1
nad1_0
trnS2
cob
nad6
trnT
trnP
nad4l
nad4
trnH
nad5
Laura Ulrich