3-1-2023

A Draft of the Genome of the Gulf Coast tick, *Amblyomma maculatum*

Jose M.C. Ribeiro
NIAID NIH Laboratory of Malaria and Vector Research, jose.ribeiro@nih.gov

Natalia J. Bayona-Vásquez
University of Georgia

Khemraj Budachetri
University of Southern Mississippi, khem.bc@usm.edu

Deepak Kumar
University of Georgia, deepak.kumar@usm.edu

Julia Catherine Frederick
University of Georgia

See next page for additional authors

Ticks and Tick-borne Diseases 14 (2023) 102090

Contents lists available at ScienceDirect

Ticks and Tick-borne Diseases

journal homepage: www.elsevier.com/locate/ttbdis

A draft of the genome of the Gulf Coast tick, Amblyomma maculatum

Jose M.C. Ribeiro [a,*], Natalia J. Bayona-Vásquez [b], Khemraj Budachetri [c,d], Deepak Kumar [b], Julia Catherine Frederick [b], Faizan Tahir [c], Brant C. Faircloth [e], Travis C. Glenn [b], Shahid Karim [c]

[a] NIAID NIH Laboratory of Malaria and Vector Research, Bethesda, MD 20892-8132, USA
[b] Department of Environmental Health Science and Georgia Genomics Facility, Environmental Health Science Building, University of Georgia, Athens, GA 30602, USA
[c] Center for Molecular and Cellular Biology, School of Biological, Environmental, and Earth Sciences, 118 College Drive, 5018, University of Southern Mississippi, Hattiesburg, MS 39406, USA
[d] The Ohio State University, Columbus, OH 43210, USA
[e] Department of Biological Sciences, Louisiana State University, 202 Life Sciences Building, Baton Rouge, LA 70803, USA

A B S T R A C T

The Gulf Coast tick, *Amblyomma maculatum*, inhabits the Southeastern states of the USA bordering the Gulf of Mexico, Mexico, and other Central and South American countries. More recently, its U.S. range has extended West to Arizona and Northeast to New York state and Connecticut. It is a vector of *Rickettsia parkeri* and *Hepatozoon americanum*. This tick species has become a model to study tick/Rickettsia interactions. To increase our knowledge of the basic biology of *A. maculatum* we report here a draft genome of this tick and an extensive functional classification of its proteome. The DNA from a single male tick was used as a genomic source, and a 10X genomics protocol determined 28,460 scaffolds having equal or more than 10 Kb, totaling 1.98 Gb. The N50 scaffold size was 19,849 Kb. The BRAKER pipeline was used to find the protein-coding gene boundaries on the assembled *A. maculatum* genome, discovering 237,921 CDS. After trimming and classifying the transposable elements, bacterial contaminants, and truncated genes, a set of 25,702 were annotated and classified as the core gene products. A BUSCO analysis revealed 83.4% complete BUSCOs. A hyperlinked spreadsheet is provided, allowing browsing of the individual gene products and their matches to several databases.

1. Introduction

The Gulf Coast tick, *Amblyomma maculatum* (Koch, 1844) is a vector of *Rickettsia parkeri* Luckman (Rickettsiales: Rickettsiaceae), which causes a febrile infection in humans (Sumner et al., 2007; Paddock et al., 2008; Cumbie et al., 2020), and also of *Hepatozoon americanum*, a pathogen of dogs (Mathew et al., 1998; Ewing et al., 2002; Mathew et al., 1999; Ewing and Panciera, 2003). The distribution of *A. maculatum* extends from the Southeastern states of the USA bordering the Gulf of Mexico, into Mexico and several other Central and South American countries. In the past decades it has extended northwards and to the West in the United States, including the states of Arkansas, Oklahoma, Kansas, and Southwestern Tennessee (Anderson et al., 2017). The Northernmost range of this tick species includes Delaware, Connecticut, and New York (Maestas et al., 2020; Molaei et al., 2021; Ramirez-Garofalo et al., 2021). Current work with this tick aims to understand its relationship with its symbionts and pathogens in general, particularly to understand the tick's immunity pathways (Adamson et al., 2013; Budachetri et al., 2017; Saito et al., 2019; Karim et al., 2021). The availability of the genome sequence of *A. maculatum* would foster the pace of these research goals.

To a researcher interested in the biochemistry and physiology of ticks, the main advantage of having the organism's genome resides in the availability of an annotated set of coding sequences (CDS) and their protein translations, which allows the building of hypotheses on the roles of these gene products and, for example, planning experiments using RNAi and genome editing to test these hypotheses. The availability of genome will also facilitate to build technologies through realizing the full potential of exploiting small RNAs, including microRNA (miRNA) and PIWI-interactacting RNA (piRNA) biology in ticks.

In this work, we used the 10X Genomics platform to sequence the genome of a single male of the Gulf Coast tick, *A. maculatum*. To obtain the genome's coding genes coordinates, we used available RNASeq data to train the BRAKER pipeline (Hoff et al., 2019). The derived CDS translations were compared to several databases and mapped to a hyperlinked spreadsheet that should allow researchers to search for their genes of interest and plan their experiments. The genome of *A. maculatum* will provide opportunities for comparative evolutionary analysis with other tick species and arthropod vectors, and allow researchers to explore the tick-pathogen interactions and ways tick parasitize vertebrate hosts.

* Corresponding author.
E-mail address: jribeiro@niaid.nih.gov (J.M.C. Ribeiro).

https://doi.org/10.1016/j.ttbdis.2022.102090
Received 17 February 2022; Received in revised form 17 October 2022; Accepted 20 November 2022
Available online 23 November 2022

2. Material and methods

2.1. Sample origin and DNA extraction and quality

Amblyomma maculatum ticks were maintained at the University of Southern Mississippi according to our modified methods (Budachetri et al., 2018). *A. maculatum* uninfected and *Rickettsia parkeri*-infected colonies were established in our laboratory in 2013. Questing unfed adult ticks were collected from Mississippi Sandhill Crane National Wildlife Refuge, Gautier, Mississippi (using the drag cloth method) on 28th July 2013. A total of 42 females and 62 males collected from the field were blood-fed on sheep and allowed to engorge and drop off. Each fully engorged female adult tick was kept separately in a snap vial for egg-laying. Individual uninfected and *Rickettsia parkeri*-infected egg clutches from individual gravid females were selected and allowed to hatch into unfed larva. The unfed larval ticks were blood-fed, allowing them to infest golden Syrian hamsters until they dropped off. Fully engorged larvae were allowed to molt into nymphs and then blood-fed on hamsters. Fully engorged nymphs were molted as male or female adult ticks. Closed colonies from the 6th generation of original wild-caught ticks were used in this study, from which five adult male ticks were selected. For each, the whole adult live tick was cut in four quarters and digested in 500 μL of buffer (10–100 mM Tris, 10–100 mM EDTA, 100–200 mM NaCl, 0.5–1% SDS) with 5 μL of proteinase K 10 mg/mL (QIAGEN, Hilden, Germany). The ticks and digestion mix were incubated in a dry bath overnight at 55° C, mixed by vortex ten times during that period. Then, for each sample, 5 μL of RNAse A (ThermoFisher Scientific, MA, USA) was added, vortexed, and incubated at room temperature for 30 min. A 400 μL aliquot was transferred to a new microcentrifuge tube and a Phenol-Chloroform-Isoamyl Alcohol (PCI) DNA extraction protocol was followed. The five extracted genomic DNA (gDNA) samples were individually hydrated in 200 μL of TE 1X buffer (Integrated DNA Technologies, Inc., IA, USA) at room temperature overnight. For verification and visualization of the products, 5 μL of each hydrated DNA sample were run in a 0.8% agarose gel.

The sample that showed the best banding pattern in the agarose gel (brightest, high-molecular weight band), a male adult and therefore with XO sexual chromosome make up, was further processed at the Georgia Genomics and Bioinformatics Core, where gDNA concentration was estimated to be 23.3 ng/μl with a Qubit® Fluorometer using the High Sensitivity protocol, and also was assessed in a Fragment Analyzer™ (FA) Automated CE System (Advanced Analytical Technologies, CA, USA) using the HS Large Fragment 50Kb method and FA version 1.2.0.11. The FA report revealed a peak size of 24,754 bp, 0.7707 ng/μl, ranging from 4550 to 100,798 bp with an average size of 27,872 bp.

2.2. Linked-reads genomic library prep and sequencing

The gDNA sample was used as input for a library prep with the Chromium™ Genome Library Kit using the Chromium™ Genome Reagent Kits v2 (CG00022 Rev C), the Chromium™ Genome Gel Bead Kit (PN-120,216), and the Chromium™ Genome Chip Kit (PN-120,216), all from 10x Genomics (10x Genomics, CA, USA). The protocol followed the manufacturer's instructions. In brief, we diluted the sample according to the standard for the genome protocol, that is 1 ng/μL, and verified that the concentration range was within acceptable limits. Then, the GEM generation sample mix was prepared and combined with both the Denaturing Agent and the gDNA. The mix was loaded into the Chromium™ Genome Chip where the Genome Gel Beads and Partitioning Oil were also loaded in the corresponding rows. The chip was placed in the Chromium™ Controller where the Genome Library program was run to partition and barcode each gDNA fragment. Barcodes were added to allow tracking of each resulting read to its original gDNA fragment. Then, the chip was ejected, and the GEMs were aspirated from the recovery well, transferred to a new tube, and isothermally incubated to generate 10x barcoded amplicons. Then the GEMs were cleaned-up with DynaBeads™ MyOne™ Silane (ThermoFisher Scientific, MA, USA), rinsed with 80% ethanol twice, and hydrated in Elution Solution. The library construction was finalized following end repair and A-tailing, adaptor ligation, post-ligation clean-up with SPRIselect, sample-index PCR using set SI-GA-A4 (contains barcodes TATGATTC, CCCACAG, ATGCTGAA, and GGATGCCG), and double-sided size selection SPRIselect. Finally, the library product was analyzed in the Fragment Analyzer™ Automated CE System (Advanced Analytical Technologies, CA, USA) using the NGS Fragment 1–6000 bp method and quantified in a Qubit® Fluorometer using the High Sensitivity protocol. The FA report showed a peak size of 533 bp, with a 3.8 ng/μL concentration; the graph ranged from 1440 bp to 5087 bp, with an average size of 705 bp. The qubit showed the library concentration to be 52.4 ng/μL.

The library was sequenced two independent times and the resulting reads were pooled. One of the sequencing runs was performed in an Illumina™ NovaSeq S4 and the second in an Ilumina™ HiSeqX, both using PE150 kits at Novogene Co., Ltd (Beijing, China).

2.3. Genome assembly

For each run, samples were demultiplexed using the four barcodes from the 10x sample index set, and the output files were merged together according to reads 1 and 2. Then, both sequencing runs were merged independently for read 1 and read 2. Raw data was used as input in Supernova v. 2.1.1 (Weisenfeld et al., 2017) using the *run* parameter allowing the use of 1200 million reads with *maxreads* in an attempt to reach a 56x raw coverage and allowing the use of 28 cores and 980 Gb of memory. Then the option *mkoutput* was used to create raw, pseudohap, pseudohap2, and megabubble outputs. The summary files regarding the assembly characteristics can be found in supplemental file 1.

2.4. Genome annotation

The BRAKER/Augustus pipeline (Hoff et al., 2019) was used to obtain the putative coding sequences (CDS) from the *A. maculatum* genome. The program was trained to find the CDS using RNAseq data available from the NCBI (accessions SRR959015 - salivary glands and SRR959016 - ovaries). These reads were concatenated and normalized using the Trinity program insilico_read_normalization.pl (Haas et al., 2013). The normalized reads were mapped to the unmasked genome using the program Star (Dobin and Gingeras, 2015). The mapped reads were used to train the gene-discovery pipeline BRAKER (Hoff et al., 2019), which discovered a total of 380,129 coding sequences (CDS). The BUSCO program (version 5.0.0) (Simão et al., 2015) was run with the BRAKER predicted protein sequences against the lineage dataset arachnida_odb10, created on 2020–08–05, from 10 species and 2934 BUSCOs. The program RepeatMasker version 4.1.2-p1 was used to identify transposable elements and repeat sequences. It was run in sensitive mode with rmblastn version 2.11.0+. The query species was assumed to be Arthropoda. The databases used were FamDB: CONS-Dfam_withRBRM_3.2. Transposable elements (TE) were identified using the Hmmer tool (Potter et al., 2018) against a subset of the Dfam database (Hubley et al., 2016) containing transposable element models, excluding repeats.The CDS were also compared to the RepBase (Bao et al., 2015) protein database to identify and classify TE. To classify genes accordiing to their functional class, the deducted protein sequences were compared using blastp to a subset of the GenBank database containing sequences from the Arachnidae, to the UniprotKB (Poux et al., 2017) database, to the Expasy Enzyme (EC) (Bairoch, 2000) database and to the MEROPS (Rawlings et al., 2016) database. Rpsblast was used to search the protein sequences against conserved motifs from the PFAM (Finn et al., 2016), SMART (Schultz et al., 2000), KOG (Tatusov et al., 2003) and CDD (Lu et al., 2020) databases. To identify genes associated with a salivary function, the CDS were compared by Rpsblast to the TickSialoFam (TSF) database (Ribeiro and Mans, 2020).

Matches that had a model coverage of > 66.6% and an e-value smaller than 1e-4 were considered as related to salivary function. General functional classification was achieved by using a set of ~ 400 key words that were searched in the definition line of the matches above. Each key word was associated with a functional class. A sequence functional class was determined by the first key word found in the definition line of the match if the product of % identity and % coverage were larger > than 0.25. If no keyword was found, the sequence was assigned to a "Unknown" function. All sequences were also searched for existence of a signal peptide indicative of secretion using the SignalP v. 3.0 program (Bendtsen et al., 2004), for transmembrane domains using the tmhmm program (Sonnhammer et al., 1998) and for O-glycosylation sites indicative of mucins using the program NetOglyc (Hansen et al., 1998). Glycosyl-phosphate-inositol membrane anchors were identified by the DGPI program (Kronegg and Buloz, 1999).

The published genomes of *Rhipicephalus microplus* and *R. sanguineus* (Jia et al., 2020) where used as input to the BRAKER/Augustus pipeline (Hoff et al., 2019) trained with publicly available protein sequences from these organisms.

2.5. Transcriptome mapping

Amblyomma maculatum transcriptome reads from the salivary glands and ovaries of adult ticks (NCBI accessions SRR13797277, SRR13797276, SRR13797275, SRR13797274, SRR13797296, SRR13797295, SRR13797294, SRR13797293, SRR13797292, SRR13797290, SRR13797289, SRR13797288, SRR13797287, SRR13797286, SRR13797285, SRR13797284, SRR13797283, SRR13797282, SRR13797305, SRR959015, SRR959016, SRR13797281, SRR13797280, SRR13797279, SRR13797278, SRR13797303, SRR13797302, SRR13797291, SRR13797304, SRR13797273, SRR13797272, SRR13797271, SRR13797270, SRR13797269, SRR13797268, SRR13797301, SRR13797300, SRR13797299, SRR13797298, SRR13797297) were mapped to the predicted CDS using Bowtie2 (Langmead and Salzberg, 2012). Read coverage was measured using samtools coverage program (Danecek et al., 2021).

2.6. Phylogenetic analysis

Protein sequences were aligned with Muscle (Edgar, 2004). Phylogenetic trees were built with the program IQ-tree (Minh et al., 2020). The best amino acid evolutionary model was determined by ModelFinder (Kalyaanamoorthy et al., 2017). The tree was bootstrapped using UFBoot2 (Hoang et al., 2018) with the bnni correction. The resulting Newick trees were annotated with Mega X (Kumar et al., 2018),

2.7. Data availability

This Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the accession JAJIZL000000000, BioProject accession PRJNA773936 and BioSample accession SAMN22546173. The reads used to assemble the genome can be found in the Sequence Read Archives (SRA) of the National Center for Biotechnology Information (NCBI) under the accession SRR16911356. The metagenome assembled *Rickettsia parkeri* genome was deposited in GenBank under the accession CP101541. Hyperlinked spreadsheets containg the annotated coding sequences can be downloaded from https://proj-bip-prod-publicread.s3.amazonaws.com/transcriptome/Amb_maculatum/Amac-genome/Supplemental_spreadsheets.zip

3. Results

We obtained a total of 942,809,836 paired-end reads from both sequencing runs. The genome assembly of *A. maculatum* resulted in 28,460 scaffolds having equal or more than 10 Kb, totaling 1.98 Gb. The N50 scaffold size was 19,849 Kb. If we add the contigs equal or larger than 1000 bp, the total assembly size reaches 2.27 Gb. The number of contigs ranging from 1000 - 9999 bp is 101,575 and the contig N50 was of 29.12 Kb length. Following search of the assembled genome for bacterial contaminants and duplicated contigs, 14 contigs, summing 1.332 Mbp were found to match known bacterial genomes, including a contig of 1.296 Mbp that matched, with 99% identity, the genome of *Rickettsia parkeri*, a known endosymbiont of *A. maculatum* (Budachetri et al., 2014). 4558 contigs were found to be exactly duplicated, adding to a total of 6.8 Mbp. These contaminants and duplicated contigs were removed from the final assembly. Table 1 lists the current available tick genomes and their characteristics. Although the N50 for the *A. maculatum* assembly was on the low range when comparing to other tick genome assemblies (Table 1), a BUSCO analysis of the predicted 25,631 CDS from the *A. maculatum* genome indicated 83.4% complete BUSCOs, 66.8% complete and single-copy BUSCOs, 16.6% complete and duplicated BUSCOs, 1.7% fragmented BUSCOS and 14.9% missing BUSCOs. These results are above the average of those shown on Table 1 which lists other tick genomes so far published.

The BRAKER pipeline (Hoff et al., 2019) was used to find the protein coding gene boundaries on the assembled *A. maculatum* genome, discovering 237,921 CDS. These were compared by blast and rpsblast (Madden, 2013) to several databases, including those at the NCBI (non-redundant and TSA protein sequences) deriving from Arachnida organisms and from Rickettsial bacteria, and the Uniprot database. After removing the sequences matching bacterial phages as well as those that represented fragments with less than 67% coverage to known proteins from the Uniprot and NCBI Arachnida sets, a set of 88,754 sequences were identified as TE (see below), and an additional set of 25,702 were annotated as the core gene products of *A. maculatum* (Supplemental spreadsheets 1 with all CDS, 2 with TEs and 3, with the core genome set).

3.1. The transposable element landscape within the genome of *Amblyomma maculatum*

The genome-coding DNA contains the information to determine the sequence of a peptide possibly containing 20 different amino acids and one stop codon. There are 64 possible codons, and 3 of them code for stops. So, a stop codon should arise on average once every 21–22 codons, or 63–66 bp. Accordingly, stretches of ORFs longer than 200 nucleotides are expected to indicate a region coding for polypeptides. However, transposable elements challenge the annotation of sequenced genomes, as they "contaminate" these longer ORFs with their coding sequences (Permal et al., 2012). Transposable elements are virus -like organisms that parasitize the majority of eukaryotic genomes, frequently loading more than half of the full genomes with their sequences. Thus, to obtain an accurate transcriptome and proteome prediction of a genome, the TE coding sequences have to be filtered out. The transposable element (TE) landscape of the *A. maculatum* genome was explored by annotating the predicted coding sequences identified as TE based on blastp matches to sequences annotated as TE from the Swissprot database as well as to coding sequences deducted from the Repbase database (Bao et al., 2015).

Among the 88,734 transcripts identified as TE's, we were able to classify 86,752; 80.1% of which were of the Class I type, 19.8% were of the Class II type and 0.17% were from endogenous retroviruses (ERV) (Table 2 and Supplemental spreadsheet 2) .Within the Class I elements, 57.7% were Long Terminal Repeat (LTR) retrotransposons and 25.4% were NON-LTR retrotransposons. Within the LTRs, Gypsy elements were most abundant, consisting 97% of the total LTR. I elements were the most abundant within the NON-LTR, reaching 40% of the 19,431 transcripts found within Non-LTR elements. Among these NON-LTR elements the BovB LINE element was identified. This element is widespread in vertebrates and it was proposed that horizontal transfer of these elements among vertebrates was vectored by ticks (Walsh et al., 2013; Mans et al., 2015). Among the Class II elements, the P/Tigger family was

Table 1
Published tick genomes characteristics.

Scientific name	Annotation	Size (Gbp)	Level	Contig N50 (kb)	Protein-coding	BioProject	Complete BUSCOs %	Reference
Amblyomma maculatum		1.98	Contig	198	25,704	PRJNA773936	83.4	This work
Dermacentor silvarum	TIGMIC Group	2.47	Chromosome	340	26,696	PRJNA633311	62.4	(Jia et al., 2020)
Haemaphysalis longicornis	TIGMIC Group	2.56	Chromosome	740	27,144	PRJNA633311	60.6	(Jia et al., 2020)
Hyalomma asiaticum	TIGMIC Group	1.71	Chromosome	555	29,644	PRJNA633311	65.0	(Jia et al., 2020)
Ixodes persulcatus	TIGMIC Group	1.90	Scaffold	533	28,574	PRJNA633311	76.1	(Jia et al., 2020)
Ixodes scapularis	Vectorbase	2.08	Contig	836	20,488	PRJNA345486	86.8	(Gulia-Nuss et al., 2016)
Ixodes scapularis	Vectorbase	2.30	Contig	1735	19,062	PRJNA678334	81.7	(Miller et al., 2018)
Rhipicephalus annulatus		2.76	Contig	437	N/A	PRJNA593711	N/A	[Unpublished]
Rhipicephalus microplus	TIGMIC Group	2.01	Scaffold	16	24,211	PRJNA312025	55.4	(Jia et al., 2020)
Rhipicephalus microplus	TIGMIC Group	2.53	Chromosome	1791	29,857	PRJNA633311	95.9	(Jia et al., 2020)
Rhipicephalus sanguineus	TIGMIC Group	2.37	Chromosome	542	25,718	PRJNA633311	60.4	(Jia et al., 2020)

most abundant, with 11,303 elements, or 65.7% of all class 2 elements found. The Mariner/TC1 family was the second most abundant, abundant, reaching 21.2% of the 17,195 Class II elements identified.

Among the predicted CDS coding for Mariner/TC1 transposases, there were 365 sequences with predicted peptide length between 400 and 600 aa (The average length of full-length Mariner transposases is near 410 aa (Robertson and Lampe, 1995)), without internal stop codons, and containing Pfam domains coding for the DDE superfamily endonuclease and domain HTH_Tnp_Tc5, coding for theTc5 transposase DNA-binding domain. Mariner/TC1 elements have been domesticated in vertebrates, including the centromere-associated protein B (CENPB) and the genes named *Tigger* transposable element-derived 2 to 7 (TIGD2–7) so far found only in vertebrates (Etchegaray et al., 2021; Gao et al., 2020). Representatives of these sequences were submitted for phylogenetic analysis, together with the here deduced Mariner/TC1 sequences from *A. maculatum* and other similar proteins from other tick species found by blast of *A. maculatum* sequences against the non-redundant database from NCBI. Interestingly, a clade with high (99% bootstrap) support (Clade XI, Supplemental Fig. 1) contained, in subclade XIb, the mammal sequences orthologous to the human TGD6 protein and tick proteins, in subclade XIa from *R. microplus, R. sanguineus, I. scapularis* and *A. maculatum.* Transcription of g129797 was found in ovaries, attaining a FPKM (Fragment Per Kilobase of transcript per Million mapped reads) of 8.78 and linear sequence coverage of 98.9%, while g180094 was found expressed in the salivary glands with a FPKM of 7.09 and linear sequence coverage of 97.9%. It is possible that these transposable elements have been also domesticated in ticks.

To compare the TE identification based on putative coding transcripts which are based on protein sequence identity with the TE predictions done from DNA sequence homologies (that are not disturbed by intruding stop codons), we used the program RepeatMasker which identified 1323,280 TE and other repetitive elements in the *A. maculatum* genome, representing 25% of the 2.35 GBases of scanned genome (Table 3). Class I elements covered 12.73% of the genome, totaling 838,798 elements, while class II elements (DNA transposons) represented 0.26% of the genome with 82,533 elements, the majority being from the Mariner/TC1 family (36,962 elements). Table 3 has additional information regarding TE and repetitive elements found in the *A. maculatum* genome.

3.2. Endogenous viral sequences

The CDS g178917.t1 codes for a nucleocapsid protein from a rhabdovirus (Walker et al., 2015) which appears to have been incorporated into the genomes of various tick species, as represented by the similar sequences found in the genomes of *R. sanguineus* (XP_037519053.1), *R. microplus* (XP_037281023.1), *Dermacentor silvarum* (XP_037579436.1), *I. ricinus* (ASY03265.1), *I. persulcatus* (KAG0426363.1) and *I. scapularis* (XP_040355436.1).

3.3. Annotation of the core genome of Amblyomma maculatum

By comparing the predicted gene products with several databases (see methods), 25,702 gene products were annotated in 29 classes, including 7,976 that were classified as "Unknown" (Table 4 and supplemental spreadsheet 3).

3.4. Salivary proteins

The search for genes associated with secreted salivary proteins was done by matches of the predicted proteins against the TSF database revealing 2,277 gene products possibly coding for salivary proteins (Table 5 and supplemental spreadsheet 3). Among these, 170 lipocalins, 38 members of the anti-complement/8.9 kDa protein family and 17 evasins were found. Comparisons of the number of members of these protein families found in the proteome annotation of published tick species (Jia et al., 2020; Miller et al., 2018; Gulia-Nuss et al., 2016) revealed a much-increased diversity of these protein families in *A. maculatum* (Table 6A). A possible reason for this discrepancy could be the failure of annotating the salivary-coding transcripts in tick genomes, possibly due to their unique sequences. In support of this hypothesis, we found larger number of these sequences in the *ab initio* predicted proteins of the genomes of *R. microplus* and *R. sanguineus*. Additionally, we searched the published salivary transcriptomes of *R. microplus* (Tirloni et al., 2020) and *R. sanguineus* (Tirloni et al., 2020), where we found larger number of these protein family members than in the annotated genomes (Table 6B).

3.5. Digestive enzymes

The sole food of ticks is blood, which is digested intracellularly with the aid of lysosomal cathepsins (Horn et al., 2009). Serine proteases may be involved in the late phase of tick engorgement (Reyes et al., 2020). We have annotated 370 protease genes in the *A. maculatum* genome, including metalloproteases, calpains, legumains, serine and cysteinyl cathepsins, serine proteases, dipeptidyl peptidases, amino and carboxy peptidases, and protein modification enzymes (Table 7 and supplemental spreadsheet 3, worksheet "Proteases"). Of notice is the expansion

Table 2
Coding sequences from transposable elements found on the *Amblyomma maculatum* genome.

Class	Type	Family	Transcripts found	Percent total elements	Percent of class
CLASS I	LTR RETROTRANSPOSON	Gypsy	48,274	55.65	96.42
CLASS I	LTR RETROTRANSPOSON	Bel-Pao	742		
CLASS I	LTR RETROTRANSPOSON	CER6-I	637		
CLASS I	LTR RETROTRANSPOSON	Copia	116		
CLASS I	LTR RETROTRANSPOSON	Ngaro	99		
CLASS I	LTR RETROTRANSPOSON	CER2-I	55		
CLASS I	LTR RETROTRANSPOSON	CIRCE	37		
CLASS I	LTR RETROTRANSPOSON	CER3-I	26		
CLASS I	LTR RETROTRANSPOSON	CER13-I	23		
CLASS I	LTR RETROTRANSPOSON	SKIPPER	11		
CLASS I	LTR RETROTRANSPOSON	Tf2	11		
CLASS I	LTR RETROTRANSPOSON	CER15-I	10		
CLASS I	LTR RETROTRANSPOSON	DIRS	9		
CLASS I	LTR RETROTRANSPOSON	CER11-I	6		
CLASS I	LTR RETROTRANSPOSON	CER10-I	4		
CLASS I	LTR RETROTRANSPOSON	SACI	3		
CLASS I	LTR RETROTRANSPOSON	TC1	3		
CLASS I	LTR RETROTRANSPOSON	HERV	2		
Total			50,068	57.71	100.00
CLASS I	NON-LTR RETROTRANSPOSON	I	6870	9.89	39.95
CLASS I	NON-LTR RETROTRANSPOSON	RTE	4128		24.01
CLASS I	NON-LTR RETROTRANSPOSON	REP2	1804		
CLASS I	NON-LTR RETROTRANSPOSON	Loa	1083		
CLASS I	NON-LTR RETROTRANSPOSON	Tad1	814		
CLASS I	NON-LTR RETROTRANSPOSON	Outcast	759		
CLASS I	NON-LTR RETROTRANSPOSON	Ingi	605		
CLASS I	NON-LTR RETROTRANSPOSON	Jockey	591		
CLASS I	NON-LTR RETROTRANSPOSON	Nimb	517		
CLASS I	NON-LTR RETROTRANSPOSON	LINE	502		
CLASS I	NON-LTR RETROTRANSPOSON	R1	432		
CLASS I	NON-LTR RETROTRANSPOSON	Tx1	303		
CLASS I	NON-LTR RETROTRANSPOSON	L1	232		
CLASS I	NON-LTR RETROTRANSPOSON	RTEX	159		
CLASS I	NON-LTR RETROTRANSPOSON	Penelope	147		
CLASS I	NON-LTR RETROTRANSPOSON	R2	143		
CLASS I	NON-LTR RETROTRANSPOSON	ORTE	94		
CLASS I	NON-LTR RETROTRANSPOSON	Crack	66		
CLASS I	NON-LTR RETROTRANSPOSON	CRE	51		
CLASS I	NON-LTR RETROTRANSPOSON	NeSL	42		
CLASS I	NON-LTR RETROTRANSPOSON	RandI	28		
CLASS I	NON-LTR RETROTRANSPOSON	CR1	18		
CLASS I	NON-LTR RETROTRANSPOSON	SR2B	12		
CLASS I	NON-LTR RETROTRANSPOSON	LIN10	9		
CLASS I	NON-LTR RETROTRANSPOSON	Proto1	6		
CLASS I	NON-LTR RETROTRANSPOSON	R4	5		
CLASS I	NON-LTR RETROTRANSPOSON	GENIE1	4		
CLASS I	NON-LTR RETROTRANSPOSON	Vingi	3		
CLASS I	NON-LTR RETROTRANSPOSON	Daphne	2		
CLASS I	NON-LTR RETROTRANSPOSON	Ambal	1		
CLASS I	NON-LTR RETROTRANSPOSON	Hero	1		
Total			19,431	22.40	100.00
Class I total			69,499	80.11	100.00
CLASS II	DNA TRANSPOSON	P/Tigger	11,303	13.03	65.73
CLASS II	DNA TRANSPOSON	Mariner/Tc1	3659	4.22	21.28
CLASS II	DNA TRANSPOSON	Ginger	793	0.91	4.61
CLASS II	DNA TRANSPOSON	Harbinger	520	0.60	3.02
CLASS II	DNA TRANSPOSON	Kolobok	318	0.37	1.85
CLASS II	DNA TRANSPOSON	UNKNOWN	130		
CLASS II	DNA TRANSPOSON	ISL2EU	129		
CLASS II	DNA TRANSPOSON	EnSpm	110		
CLASS II	DNA TRANSPOSON	hAT	64		
CLASS II	DNA TRANSPOSON	piggyBac	33		
CLASS II	DNA TRANSPOSON	Zisupton	27		
CLASS II	DNA TRANSPOSON	CACTA	24		
CLASS II	DNA TRANSPOSON	Helitron	20		
CLASS II	DNA TRANSPOSON	Merlin	18		
CLASS II	DNA TRANSPOSON	MuDR	16		
CLASS II	DNA TRANSPOSON	MiniSatellite	9		
CLASS II	DNA TRANSPOSON	mule	8		
CLASS II	DNA TRANSPOSON	THAP9	6		
CLASS II	DNA TRANSPOSON	LOOPER	4		
CLASS II	DNA TRANSPOSON	Academ	2		
CLASS II	DNA TRANSPOSON	PIF-Harbinger	2		
Class II total			17,195	19.82	

(*continued on next page*)

Table 2 (*continued*)

Class	Type	Family	Transcripts found	Percent total elements	Percent of class
ERV	ENDOGENOUS RETROVIRUS	ERV3	38		
ERV	ENDOGENOUS RETROVIRUS	ERV1	15		
ERV	ENDOGENOUS RETROVIRUS	Endogenous Retrovirus	5		
Endogenous retrovirus total			58	0.07	
Grand total			86,752	100.00	

Table 3
Transposable elements identified in by RepeatMasker the *Amblyomma maculatum* genome. Total genome size scanned = 2350,858,905 bases.

Element family	Number of elements	Base Pairs	Percentage of Genome
Retroelements	667,681	227,801,712	9.24%
SINEs:	189,995	30,535,656	1.24%
Penelope	9123	923,733	0.04%
LINEs:	306,569	111,282,227	4.51%
CRE/SLACS	7	415	0.00%
L2/CR1/Rex	92,277	13,151,584	0.53%
R1/LOA/Jockey	44,875	7279,321	0.30%
R2/R4/NeSL	3298	305,707	0.01%
RTE/Bov-B	127,837	87,675,663	3.56%
L1/CIN4	1891	100,605	0.00%
LTR elements:	171,117	85,983,829	3.49%
BEL/Pao	10,454	1214,535	0.05%
Ty1/Copia	5007	265,943	0.01%
Gypsy/DIRS1	154,051	84,423,689	3.42%
Retroviral	0	0	0.00%
Total Class I elements	838,798	313,785,541	
DNA transposons	82,533	6360,904	0.26%
hobo-Activator	8996	630,064	0.03%
Tc1-IS630-Pogo	36,962	3489,969	0.14%
En-Spm	0	0	0.00%
MuDR-IS905	0	0	0.00%
PiggyBac	922	115,270	0.00%
Tourist/Harbinger	945	118,864	0.00%
Other (Mirage, P-element, Transib)	2924	179,953	0.01%
Rolling-circles	11,361	683,173	0.03%
Unclassified:	5696	432,101	0.02%
Total interspersed repeats		234,594,717	9.51%
Small RNA:	190,873	30,766,714	1.25%
Satellites:	1573	135,324	0.01%
Simple repeats:	0	0	0.00%
Low complexity:	0	0	0.00%
Total	1323,280	617,660,512	25.07

* most repeats fragmented by insertions or deletions.
have been counted as one element.
The query species was assumed to be arthropoda.
RepeatMasker version 4.1.2-p1, sensitive mode.

Table 4
Classification and number of core gene products identified in the *Amblyomma maculatum* genome.

Class	Number of gene products
Putative salivary secreted	2,277
Cytoskeletal proteins	638
Detoxification	233
Oxidant metabolism/Detoxification	157
Extracellular matrix	383
Immunity	187
Amino acid metabolism	358
Carbohydrate metabolism	334
Energy metabolism	515
Intermediary metabolism	139
Lipid metabolism	682
Nucleotide metabolism	206
Nuclear export	33
Nuclear regulation	558
Protein export	2,361
Protein modification	691
Proteasome machinery	641
Protein synthesis machinery	606
Secreted protein	914
Signal transduction	3,293
Storage	39
Transcription factor	50
Transcription machinery	1,392
Transporters and channels	1,040
Unknown conserved	349
Unkown conserved membrane protein	211
Unknown product	7,065
Unkown membrane protein	351
Viral product	1
Total	25,704
Total - Unknown	17,728

of the M13 metalloproteases, with 447 genes, compared to 255 found in *R. microplus* and 41 on the *I. scapularis* annotated proteomes (Table 8). Other peptidases are listed on the worksheet "Protein modification" of supplemental spreadsheet 3.

3.6. Protein modification enzymes

Within the "protein modification enzymes" we highlight the finding of a putative tyrosine sulfotransferase, an enzyme that adds a sulfate group to a tyrosine residue, an important protein modification in tick hormones (Donohue et al., 2010) and some tick salivary peptides (Franck et al., 2020; Thompson et al., 2017).

Among other protein modification enzymes, we found several genes coding for members of the prolyl hydroxylase complex, which are important in the production of mature collagen proteins (Gorres and Raines, 2010). These can be browsed in the worksheet "Protein modification" from supplemental spreadsheet 3.

Protein glycosyl transferases adds carbohydrate residues to proteins. In ticks, these enzymes have received recent attention due to the epidemics of alpha-gal allergies, which are thought to be triggered by alpha-galactosyl residues decorating the salivary proteins of some tick species, including *Amblyomma americanum* and *Ixodes scapularis*, but not in *Dermacentor variabilis* or *A. maculatum* (Crispell et al., 2019). In *I. scapularis*, typical α-Gal transferases (GALT) were absent in the genome, but enzymes of the α1–4 and β−14 GALT families were able to generate protein α-Galactosylation (Cabezas-Cruz et al., 2018). These enzymes can be recognized by the "Lactosylceramide 4-alpha-galactosyltransferase" TSFam motif (Ribeiro and Mans, 2020). No enzymes matching this motif or other α-GALT enzymes were found in the *A. maculatum* genome. The worksheet named "glycosyltransferases" of supplemental spreadsheet 2 presents data on 192 glycosyltransferases.

3.7. Cytoskeletal and extracellular proteins

On supplemental spreadsheet 3, worksheet "Cytoskeletal", annotations can be found for myosins, actins, tubulins, their interacting proteins, and diverse collagen proteins, proteoglycans, and their related enzymes, cuticles and other chitin binding proteins, and gap-junction Innexin proteins.

3.8. Immunity-related products

Annotation of genes coding for products associated with immunity

Table 5
Classification and abundance of putative salivary expressed genes from *Amblyomma maculatum* predicted by the TicskSialoFam database.

Family Subfamily	Number of CDS
12 kDa family	
Generic	5
Metastriate	11
pk4/12kDa	3
12kDaBasic	1
13–14kDa	
13kDa	10
13kDa-Basic	1
23–24kDa	
23kDa	15
15kDaBasic	3
18kDa	19
19kDa	1
23–24 kDa family	
23kDa	1
24kDa	36
28kDa	8
8.9kDa	38
8kDa	
8 kDa metastriate	2
AlaRich	63
Amb-25–357	4
Antigen-5	8
BSMAP	1
Basic tail family	
Generic	22
TSGP1	5
CalreticulinCalnexin	3
Cell adhesion molecule	7
Coiled-coil domain-containing	4
Complement receptor	1
Complement-binding protein	9
CUTA1	2
CystineKnotToxin	3
Cytotoxin	44
DAP-36	2
Down syndrome family of cell adhesion molecules	
Generic	2
Ig_3	28
IG_like	8
Ig-domain	6
EFh_CREC_Calumenin_like	4
Evasin	
EvasinA	16
EvasinB	1
Fasciclin-1	1
Ficolin/Ixoderin	11
Fukutin	1
Glycine Rich protein family	
Generic	7
Cement	1
GRP_cement_450	1
GRP_cement_833	1
Collagen-like	1
Chitin binding	48
Dystroglycan	1
GGY	6
GRP21	6
Grp7_allergen	16
Large GYY	8
Large_GRP_II	2
Hematopoietic stem/progenitor cells -like	1
HVA22/Cytokine	1
Hyp_94	1
Hyp2009	2
Insulin_growth_factor	11
Integrin	
Alpha subunit	1
Beta subunit	3
Interleukin17-like	14
Ixodegrin	25
Ixodegrin-like	1
Kielin/chordin-like	1
Laminin	9
Lipocalin	170
His binding	98
Generic	47
lipocal-1 1	1
Metastriate IgG-binding lipocalin	17
94	7
Low-density lipoprotein receptor	21
ML_domain	20
Mucin	
Generic	78
HRP	6
Peritrophin	8
Sialomucin	1
MYS-2	1
Mys-25–289	1
Mys-25–299	8
Mys-30–170	2
Mys-30–60	6
Mys-30–94	5
Niemann-Pick	6
OneOfEach	38
OSTMP1	1
Papa	2
Peptidoglycan_recognition_protein	5
Phosphatidylethanolamine binding	16
Prich	15
Prich	3
Rapp-25–325	1
Salp15/Ixostatin	
Ixostatin	7
Saposin	1
Selenoprotein	1
Serum amyloid A	4
Synaptotagmin 1	1
TGF-beta propeptide	7
TGF-beta propeptide	1
Tick Hirudin	1
Tick-MYS1	1
TMEM9	1
Toll4_associated	1
Toll-like	57
Tolloid-like	2
translocon-associated protein subunit alpha	1
Vitellogenin	4
Vitelogenin-VWF	4
YRP	2
HVA22/Cytokine	1
Hyp669	2
Malectin	1
Toll-like	5
Antimicrobial	
5.3kDa	
Metastriate_5	2
DAE-2	1
Defensin	6
Is4	6
Lysozyme	4
Microplusin	15
Microplusin_2	20
Enzymes	
5′nucleotidase/Apyrase	9
Coesterase	99
Phospholipase A2	6
Cysteinyl_peptidase	26
Dehydrogenase	82
Angiotensin converting enzyme	5
Ectonucleotide pyrophosphatase/phosphodiesterase	3
Endonuclease	14
Epoxide hydrolase	18
IPPase	4
Multiple inositol polyphosphate phosphatase	2
M13_peptidase	377
Metalloprotease	63

(*continued on next page*)

Table 5 (*continued*)

Family Subfamily	Number of CDS
Sphingomyelinase	13
Serine carboxypeptidase	22
Zinc carboxypeptidase	2
Serine protease	62
Peroxidase	12
Selenium dependent glutathione peroxidase	4
Superoxide dismutase, cu2+/zn2+ superoxide dismutase sod1	8
Tyrosine sulfotransferase	1
Catalytically inactive chitinase-like lectin	71
Proteinase inhibitors	
Longistatin	4
Carboxypeptidase inhibitor	1
Cystatin	14
Thyropin	2
Kunitz	90
Serpin	68
Kazal	1
SPARC/Kazal	24
TIL	20
Total	2278

Table 6
A: Number of gene products coding for typical salivary proteins in tick genomes.

Species	Lipocalins	8.9 kDA	Evasins	Reference
A. maculatum	170	38	17	This work
D. silvarum	27	3	0	(Jia et al., 2020)
H. asiaticum	48	7	0	(Jia et al., 2020)
H. longicornis	1	0	0	(Jia et al., 2020)
I. persulcatus	12	2	0	(Jia et al., 2020)
I. scapularis SE6	37	12	2	(Miller et al., 2018)
I. scapularis Wikel	41	15	2	(Gulia-Nuss et al., 2016)
R. microplus	20	1	0	(Jia et al., 2020)
R. sanguineus	29	1	0	(Jia et al., 2020)

Table 6B: Number of gene products or coding sequences coding for typical salivary proteins in published tick genomes, ab-initio genomes or transcriptomes.

Species	Lipocalins	8.9 kDA	Evasins	Reference
R. sanguineus genome	29	1	0	(Jia et al., 2020)
R. sanguineus ab initio	48	8	1	
R. sanguineus transcriptome	141	34	17	(Tirloni et al., 2020)
R. microplus genome	20	1	0	(Jia et al., 2020)
R. microplus ab initio	48	3	0	
R. microplus transcriptome	140	22	12	(Tirloni et al., 2020)

Table 7
Annotated proteases found in the *Amblyomma maculatum* genome.

Class	Number of genes
M10 metalloproteases	2
M12B metalloproteases	17
M13 metalloproteases	185
Calpains	5
Cathepsin B	4
Cathepsin D (Pepsin)	7
Cathepsin K (Papain)	2
Cathepsin L (Papain)	10
Cathepsin O (Papain)	2
Serine proteases	35
Dipeptidyl peptidase	2
Legumains	22
Amino and Carboxypeptidases	59
Other peptidases	13
Protein modification enzymes	5
Total	370

Table 8
Number of gene products coding for M13 proteases within tick genomes.

Species	Number of genes	Reference
A. maculatum	447	This work
D. silvarum	174	(Jia et al., 2020)
H. asiaticum	165	(Jia et al., 2020)
H. longicornis	120	(Jia et al., 2020)
I. persulcatus	59	(Jia et al., 2020)
R. sanguineus	129	(Jia et al., 2020)
R. microplus	115	(Jia et al., 2020)
I. scapularis Wikel	41	(Gulia-Nuss et al., 2016)
I. scapularis SE6	41	(Miller et al., 2018)

revealed proteins coding for: (1) the antimicrobial peptides Defensins, microplusins, lysozymes, Is4 and DAE-2 (Supplemental spreadsheet 3, see results on both Salivary and Immunity worksheets, 2) the RNAi/antiviral response, including Argonaute, Armitage, Aubergine, Tudor, RM62 and Serrate, (3) several members of the alpha-macroglobulin family of complement-like thio-ester esterases (4), several proteins associated with the interferon response (5), three products with similarities to Interleukin-16 and IL-17 (6), Chemokine-like products (7). Several proteins associated with the Tumor Necrosis Factor (TNF) response, including the TNF receptor protein (8), members of the IMD pathway such as Bendless, Caspar, Caudal, Effete, IKK famma - protein kinase, TAB2, TAK1, Uev1a, IAP2 and akirins (9), several products associated with pathogen-recognition motifs (10), members of the SOCS-JAK Stat pathway such as JAK Hopscotch Tyrosine protein kinase, JAK Receptor (Domeless), PIAS Sumo ligase, SOCS box SH2-domain-containing protein and Stat3 (10), members of the TOLL pathway Cactus, Dorsal, MYD88, Pelle, Tube, Spaetzle and several Toll-like receptors.

3.9. Epigenetic control and transcription factors

Products affecting epigenetic control, such as histone lysine methyl transferases, histone acetylases and acetyltransferases, histone deacetylases, sirtuins and several members of the chromatin remodeling complex are identified in the supplemental spreadsheet 3 under the row named “Epigenetic control. Transcription factors (47 sequences) are also annotated in Supplemental spreadsheet 2.

3.10. Oxidative and detoxification metabolisms

Catalases, peroxidases, superoxide dismutases, Cytochrome P−450, Cytoglobins, Selenoproteins, Thioredoxins, Sulfotransferases, Aryl and Glycosyl sulfatases and Glutathione transferases are listed on the worksheet named “Detoxification” on supplemental spreadsheet 3.

3.11. Signal transduction

Worksheet “Signal transduction” of supplemental spreadsheet 3 lists several transcripts giving best matches to proteins annotated as 7 transmembrane receptors, G protein-coupled receptors, alpha-1a adrenergic receptor, and receptors for acetylcholine, dopamine, adenosine, serotonin. histamine, adiponectin, rmrfamide, ecdysone, allatostatins, leucokinin atrial natriuretic factor, calcitonin, cholecystokinin, corticotropin, gaba, glycine, octopamine, gonadotropin-releasing hormone, melanocortin neuropeptide y receptor, pyrokinin, relaxin, sifamide and vasopressin. These receptors can be targets of novel acaricides. Several hormonal precursors are also listed, including for the crustacean chh/mih/gih neurohormone family, neurohypophysial hormones and several prohormones.

3.12. Additional annotations

Supplemental spreadsheet 3 also details genes coding for proteins implicated on nuclear regulation and nuclear export, transcription and

translation machineries, protein export, amino acid, carbohydrate, lipid, and energy metabolisms and proteasome machinery,

4. Discussion

Using the Chromium Genome Library Kit and the 10X Genomics platform, we obtained a draft genome sequence of the tick *Amblyomma maculatum*, the first genome for this tick genus, using the DNA extracted from a single male tick. A total of 237,921 putative coding sequences were discovered by the Augustus/BRAKER pipeline trained with public RNAseq data. After excluding transposable elements and truncated sequences, we arrived at a core set of 25,702 coding genes that were functionally annotated and available for browsing in hyperlinked spreadsheets, which we hope will be valuable for further research with this tick species and contributing to the understanding of tick phylogeny.

Analysis of the expanded salivary gland expressed families (such as lipocalin) from the genome of *A. maculatum* and 3 other tick species show a considerable absence of sequences predicted by transcriptome assembly. It is possible that the "missing " salivary-coding genes could derive from a higher polymorphism of these genes. Indeed, variable mutation rates are known to occur among different genes (Hodgkinson, 2011) associated with those having high transcription (Park et al., 2012) or associated with adaptation to variable environments, such as those caused by the host immune response (Matic, 2019), conditions that are found for the highly expressed salivary-coding genes, such as those coding for the lipocalins or metalloproteases. Additionally, increased recombination rates within salivary-coding genes, as observed in some organisms (Wallberg et al., 2015; Hey, 2004), could cause large sequence variation among the individual tick genomes, causing the repertoire of genes at the level of the population being much larger than at the individual level. This hypothesis could be tested by comparing the abundance and similarities of salivary-coding genes from genomes assembled from different individuals.

Data availability

The data is available in GenBank and has been submitted for download in the specified link.

Funding

This publication was made possible by an Institutional Development Award (IDeA) from the National Institute of General Medical Sciences of the National Institutes of Health under Grant #P20GM103476; USDA NIFA (2017-67017-26171 & 2017-67016-26864); Pakistan-U.S. Science and Technology Cooperation Program (Phase 7) (10003290), the National Science Foundation Grant No. DGE-1545433 (JCF). JMCR was supported by the Intramural Research Program of the National Institute of Allergy and Infectious Diseases (Vector-Borne Diseases: Biology of Vector Host Relationship, Z01 AI000810-18).

Acknowledgements

This work used the Georgia Advanced Computing Resource Center and the Georgia Genomics and Bioinformatics Core at UGA, the HPC@LSU and the computational resources of the NIH HPC Biowulf cluster (http://hpc.nih.gov). We are grateful to Drs. John Andersen, Ben Mans and Isabel Santos for helpful comments on the manuscript.

Author statement

Jose M.C. Ribeiro Methodology, Data curation, Formal analysis, Writing - original draft.

Natalia J. Bayona-Vásquez Project administration, Methodology, Writing - review & editing.

Khemraj Budachetri Methodology, Writing - review & editing.

Deepak Kumar Methodology, Writing - review & editing.

Julia Catherine Frederick Methodology, Writing - review & editing.

Faizan Tahir Methodology, Writing - review & editing.

Brant C. Faircloth Methodology, Writing - review & editing.

Travis. C. Glenn Project administration, Methodology, Supervision, Writing - review & editing.

Shahid Karim Project administration, Funding acquisition, Methodology, Supervision, Writing - original draft.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ttbdis.2022.102090.

References

Sumner, J.W., Durden, L.A., Goddard, J., Stromdahl, E.Y., Clark, K.L., Reeves, W.K., Paddock, C.D., 2007. Gulf coast ticks (*Amblyomma maculatum*) and *Rickettsia parkeri*, United States. Emerging Infect. Dis. 13 (5), 751.

Paddock, C.D., Finley, R.W., Wright, C.S., Robinson, H.N., Schrodt, B.J., Lane, C.C., Ekenna, O., Blass, M.A., Tamminga, C.L., Ohl, C.A., 2008. *Rickettsia parkeri* rickettsiosis and its clinical distinction from Rocky Mountain spotted fever. Clin. Infect. Dis. 47 (9), 1188–1196.

Cumbie, A.N., Espada, C.D., Nadolny, R.M., Rose, R.K., Dueser, R.D., Hynes, W.L., Gaff, H.D., 2020. Survey of *Rickettsia parkeri* and *Amblyomma maculatum* associated with small mammals in southeastern Virginia. Ticks Tick Borne Dis. 11 (6), 101550.

Mathew, J., Ewing, S., Panciera, R., Woods, J., 1998. Experimental transmission of *Hepatozoon americanum* Vincent-Johnson et al., 1997 to dogs by the Gulf Coast tick, *Amblyomma maculatum* Koch. Vet. Parasitol. 80 (1), 1–14.

Ewing, S., DuBois, J., Mathew, J., Panciera, R., 2002. Larval Gulf Coast ticks (*Amblyomma maculatum*)[Acari, Ixodidae] as host for *Hepatozoon americanum* [Apicomplexa, Adeleorina]. Vet. Parasitol. 103 (1–2), 43–51.

Mathew, J.S., Ewing, S.A., Panciera, R.J., Kocan, K.M., 1999. Sporogonic development of *Hepatozoon americanum* (Apicomplexa) in its definitive host, *Amblyomma maculatum* (Acarina). J. Parasitol. 85 (6), 1023–1031.

Ewing, S.A., Panciera, R.J., 2003. American canine hepatozoonosis. Clin. Microbiol. Rev. 16 (4), 688–697.

Anderson, J.M., Moore, I.N., Nagata, B.M., Ribeiro, J.M.C., Valenzuela, J.G., Sonenshine, D.E., 2017. Ticks, *Ixodes scapularisIxodes* scapularis, feed repeatedly on white-footed mice despite strong inflammatory response, an expanding paradigm for understanding tick-host interactions. Front. Immunol. 8, 1784.

Maestas, L.P., Reeser, S.R., McGay, P.J., Buoni, M.H., 2020. Surveillance for *Amblyomma maculatum* (Acari, Ixodidae) and *Rickettsia parkeri* (Rickettsiales, Rickettsiaceae) in the State of Delaware, and their public health implications. J. Med. Entomol. 57 (3), 979–983.

Molaei, G., Little, E.A.H., Khalil, N., Ayres, B.N., Nicholson, W.L., Paddock, C.D., 2021. Established population of the Gulf Coast Tick, *Amblyomma maculatum* (Acari, Ixodidae), Infected with *Rickettsia parkeri* (Rickettsiales, Rickettsiaceae), in Connecticut. J. Med. Entomol. 58 (3), 1459–1462.

Ramirez-Garofalo J.R., Curley S.R., Field C.E., Hart C.E., Thangamani S., Established populations of *Rickettsia parkeri*-Infected *Amblyomma maculatum* Ticks in New York City, New York, USA. Vector borne and zoonotic diseases Larchmont, NY 2021.

Adamson, S.W., Browning, R.E., Budachetri, K., Ribeiro, J.M., Karim, S., 2013. Knockdown of selenocysteine-specific elongation factor in *Amblyomma maculatum* alters the pathogen burden of *Rickettsia parkeri* with epigenetic control by the Sin3 histone deacetylase corepressor complex. PLoS One 8 (11), e82012.

Budachetri, K., Kumar, D., Karim, S., 2017. Catalase is a determinant of the colonization and transovarial transmission of *Rickettsia parkeri* in the Gulf Coast tick *Amblyomma maculatum*. Insect Mol. Biol. 26 (4), 414–419.

Saito, T.B., Bechelli, J., Smalley, C., Karim, S., Walker, D.H., 2019. Vector tick transmission model of spotted fever rickettsiosis. Am. J. Pathol. 189 (1), 115–123.

Karim, S., Kumar, D., Budachetri, K., 2021. Recent advances in understanding tick and rickettsiae interactions. Parasite Immunol. e12830.

Hoff, K.J., Lomsadze, A., Borodovsky, M., Stanke, M., 2019. Whole-Genome Annotation with BRAKER. In, Gene prediction. Springer, pp. 65–95.

Budachetri, K., Kumar, D., Crispell, G., Beck, C., Dasch, G., Karim, S., 2018. The tick endosymbiont *Candidatus Midichloria mitochondrii* and selenoproteins are essential for the growth of *Rickettsia parkeri* in the Gulf Coast tick vector. Microbiome 6 (1), 1–15.

Weisenfeld, N.I., Kumar, V., Shah, P., Church, D.M., Jaffe, D.B., 2017. Direct determination of diploid genome sequences. Genome Res. 27 (5), 757–767.

Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., Couger, M.B., Eccles, D., Li, B., Lieber, M., 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat. Protoc. 8 (8), 1494–1512.

Dobin, A., Gingeras, T.R., 2015. Mapping RNA-seq reads with STAR. Curr. Protocols Bioinform. 51 (1), 11.14. 11-11.14. 19.

Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., Zdobnov, E.M., 2015. BUSCO, assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31 (19), 3210–3212.

Potter, S.C., Luciani, A., Eddy, S.R., Park, Y., Lopez, R., Finn, R.D., 2018. HMMER web server, 2018 update. Nucleic. Acids. Res. 46 (W1), W200–W204.
Hubley, R., Finn, R.D., Clements, J., Eddy, S.R., Jones, T.A., Bao, W., Smit, A.F., Wheeler, T.J., 2016. The Dfam database of repetitive DNA families. Nucleic. Acids. Res. 44 (D1), D81–D89.
Bao, W., Kojima, K.K., Kohany, O., 2015. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mobile DNA-Uk 6 (1), 1–6.
Poux, S., Arighi, C.N., Magrane, M., Bateman, A., Wei, C.-H., Lu, Z., Boutet, E., Bye-A-Jee, H., Famiglietti, M.L., Roechert, B., 2017. On expert curation and scalability, UniProtKB/Swiss-Prot as a case study. Bioinformatics 33 (21), 3454–3460.
Bairoch, A., 2000. The ENZYME database in 2000. Nucleic. Acids. Res. 28 (1), 304–305.
Rawlings, N.D., Barrett, A.J., Finn, R., 2016. Twenty years of the MEROPS database of proteolytic enzymes, their substrates and inhibitors. Nucleic. Acids. Res. 44 (D1), D343–D350.
Finn, R.D., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Mistry, J., Mitchell, A.L., Potter, S.C., Punta, M., Qureshi, M., Sangrador-Vegas, A., 2016. The Pfam protein families database, towards a more sustainable future. Nucleic. Acids. Res. 44 (D1), D279–D285.
Schultz, J., Copley, R.R., Doerks, T., Ponting, C.P., Bork, P., 2000. SMART, a web-based tool for the study of genetically mobile domains. Nucleic. Acids. Res. 28 (1), 231–234.
Tatusov, R.L., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Kiryutin, B., Koonin, E.V., Krylov, D.M., Mazumder, R., Mekhedov, S.L., Nikolskaya, A.N., 2003. The COG database, an updated version includes eukaryotes. BMC Bioinf. 4 (1), 1–14.
Lu, S., Wang, J., Chitsaz, F., Derbyshire, M.K., Geer, R.C., Gonzales, N.R., Gwadz, M., Hurwitz, D.I., Marchler, G.H., Song, J.S., 2020. CDD/SPARCLE, the conserved domain database in 2020. Nucleic. Acids. Res. 48 (D1), D265–D268.
Ribeiro, J.M.C., Mans, B.J., 2020. TickSialoFam (TSFam), a database that helps to classify tick salivary proteins, a review on tick salivary protein function and evolution, with considerations on the tick sialome switching phenomenon. Front. Cell Infect. Microbiol. 10, 374.
Bendtsen, J.D., Nielsen, H., Von Heijne, G., Brunak, S., 2004. Improved prediction of signal peptides, SignalP 3.0. J. Mol. Biol. 340 (4), 783–795.
Sonnhammer, E.L., Von Heijne, G., Krogh, A., 1998. A hidden Markov model for predicting transmembrane helices in protein sequences. In, Ismb 175–182, 1998.
Hansen, J.E., Lund, O., Tolstrup, N., Gooley, A.A., Williams, K.L., Brunak, S., 1998. NetOglyc, prediction of mucin type O-glycosylation sites based on sequence context and surface accessibility. Glycoconjugate J. 15 (2), 115–130.
Kronegg, J., Buloz, D., 1999. Detection/prediction of GPI cleavage site (GPI-anchor) in a protein (DGPI). URL 129194, 185.
Jia, N., Wang, J., Shi, W., Du, L., Sun, Y., Zhan, W., Jiang, J.F., Wang, Q., Zhang, B., Ji, P., et al., 2020. Large-scale comparative analyses of tick genomes elucidate their genetic diversity and vector capacities. Cell 182 (5), 1328–1340 e1313.
Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9 (4), 357–359.
Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., 2021. Twelve years of SAMtools and BCFtools. GigaScience 10 (2) giab008.
Edgar, R.C., 2004. MUSCLE, multiple sequence alignment with high accuracy and high throughput. Nucleic. Acids. Res. 32 (5), 1792–1797.
Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., Von Haeseler, A., Lanfear, R., 2020. IQ-TREE 2, new models and efficient methods for phylogenetic inference in the genomic era. Mol. Biol. Evol. 37 (5), 1530–1534.
Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K., Von Haeseler, A., Jermiin, L.S., 2017. ModelFinder, fast model selection for accurate phylogenetic estimates. Nat. Methods 14 (6), 587–589.
Hoang, D.T., Chernomor, O., Von Haeseler, A., Minh, B.Q., Vinh, L.S., 2018. UFBoot2, improving the ultrafast bootstrap approximation. Mol. Biol. Evol. 35 (2), 518–522.
Kumar, S., Stecher, G., Li, M., Knyaz, C., Tamura, K., 2018. MEGA X, molecular evolutionary genetics analysis across computing platforms. Mol. Biol. Evol. 35 (6), 1547.
Budachetri, K., Browning, R.E., Adamson, S.W., Dowd, S.E., Chao, C.-C., Ching, W.-M., Karim, S., 2014. An insight into the microbiome of the *Amblyomma maculatum* (Acari, Ixodidae). J. Med. Entomol. 51 (1), 119–129.
Madden, T., 2013. The BLAST sequence analysis tool. The NCBI Handbook [Internet] 2nd Edition. National Center for Biotechnology Information (US).
Permal, E., Flutre, T., Quesneville, H., 2012. Roadmap for annotating transposable elements in eukaryote genomes. Methods Mol. Biol. 859, 53–68. Clifton, NJ.
Walsh, A.M., Kortschak, R.D., Gardner, M.G., Bertozzi, T., Adelson, D.L., 2013. Widespread horizontal transfer of retrotransposons. Proc. Natl Acad. Sci. 110 (3), 1012–1016.
Mans, B.J., De Klerk, D., Pienaar, R., De Castro, M.H., Latif, A.A., 2015. Next-generation sequencing as means to retrieve tick systematic markers, with the focus on *Nuttalliella namaqua* (Ixodoidea, Nuttalliellidae). Ticks Tick Borne Dis. 6 (4), 450–462.
Robertson, H.M., Lampe, D.J., 1995. Distribution of transposable elements in arthropods. Annu. Rev. Entomol. 40, 333–357.
Etchegaray, E., Naville, M., Volff, J.-N., Haftek-Terreau, Z., 2021. Transposable element-derived sequences in vertebrate development. Mobile DNA-Uk 12 (1), 1–24.
Gao, B., Wang, Y., Diaby, M., Zong, W., Shen, D., Wang, S., Chen, C., Wang, X., Song, C., 2020. Evolution of pogo, a separate superfamily of IS630-Tc1-mariner transposons, revealing recurrent domestication events in vertebrates. Mob DNA 11, 25.
Walker, P.J., Firth, C., Widen, S.G., Blasdell, K.R., Guzman, H., Wood, T.G., Paradkar, P. N., Holmes, E.C., Tesh, R.B., Vasilakis, N., 2015. Evolution of genome size and complexity in the Rhabdoviridae. PLoS Pathog. 11 (2), e1004664.
Miller, J.R., Koren, S., Dilley, K.A., Harkins, D.M., Stockwell, T.B., Shabman, R.S., Sutton, G.G., 2018. A draft genome sequence for the *Ixodes scapularis* cell line, ISE6. F1000Res 7.
Gulia-Nuss, M., Nuss, A.B., Meyer, J.M., Sonenshine, D.E., Roe, R.M., Waterhouse, R.M., Sattelle, D.B., De La Fuente, J., Ribeiro, J.M., Megy, K., 2016. Genomic insights into the *Ixodes scapularis* tick vector of Lyme disease. Nat. Commun. 7 (1), 1–13.
Tirloni, L., Braz, G., Nunes, R.D., Gandara, A.C.P., Vieira, L.R., Assumpcao, T.C., Sabadin, G.A., da Silva, R.M., Guizzo, M.G., Machado, J.A., et al., 2020a. A physiologic overview of the organ-specific transcriptome of the cattle tick Rhipicephalus microplus. Sci. Rep. 10 (1), 18296.
Tirloni, L., Lu, S., Calvo, E., Sabadin, G., Di Maggio, L.S., Suzuki, M., Nardone, G., da Silva Vaz Jr., I., Ribeiro, J.M.C., 2020b. Integrated analysis of sialotranscriptome and sialoproteome of the brown dog tick Rhipicephalus sanguineus (s.l.), Insights into gene expression during blood feeding. J. Proteomics 229, 103899.
Horn, M., Nussbaumerová, M., Šanda, M., Kovářová, Z., Srba, J., Franta, Z., Sojka, D., Bogyo, M., Caffrey, C.R., Kopáček, P., 2009. Hemoglobin digestion in blood-feeding ticks, mapping a multipeptidase pathway by functional proteomics. Chem. Biol. 16 (10), 1053–1063.
Reyes, J., Ayala-Chavez, C., Sharma, A., Pham, M., Nuss, A.B., Gulia-Nuss, M., 2020. Blood digestion by trypsin-like serine proteases in the replete Lyme disease vector tick, Ixodes scapularis. Insects 11 (3), 201.
Donohue, K.V., Khalil, S.M., Ross, E., Grozinger, C.M., Sonenshine, D.E., Roe, R.M., 2010. Neuropeptide signaling sequences identified by pyrosequencing of the American dog tick synganglion transcriptome during blood feeding and reproduction. Insect Biochem. Mol. Biol. 40 (1), 79–90.
Franck, C., Foster, S.R., Johansen-Leete, J., Chowdhury, S., Cielesh, M., Bhusal, R.P., Mackay, J.P., Larance, M., Stone, M.J., Payne, R.J., 2020. Semisynthesis of an evasin from tick saliva reveals a critical role of tyrosine sulfation for chemokine binding and inhibition. Proc. Natl Acad. Sci. 117 (23), 12657–12664.
Thompson, R.E., Liu, X., Ripoll-Rozada, J., Alonso-Garcia, N., Parker, B.L., Pereira, P.J. B., Payne, R.J., 2017. Tyrosine sulfation modulates activity of tick-derived thrombin inhibitors. Nat. Chem. 9 (9), 909–917.
Gorres, K.L., Raines, R.T., 2010. Prolyl 4-hydroxylase. Crit. Rev. Biochem. Mol. Biol. 45 (2), 106–124.
Crispell, G., Commins, S.P., Archer-Hartman, S.A., Choudhary, S., Dharmarajan, G., Azadi, P., Karim, S., 2019. Discovery of alpha-gal-containing antigens in North American tick species believed to induce red meat allergy. Front. Immunol. 10, 1056.
Cabezas-Cruz, A., Espinosa, P.J., Alberdi, P., Šimo, L., Valdés, J.J., Mateos-Hernández, L., Contreras, M., Rayo, M.V., de la Fuente, J., 2018. Tick galactosyltransferases are involved in α-Gal synthesis and play a role during *Anaplasma phagocytophilum* infection and *Ixodes scapularis* tick vector development. Sci. Rep. 8 (1), 1–18.
Hodgkinson, A., 2011. Eyre-Walker A, Variation in the mutation rate across mammalian genomes. Nat. Rev. Genet. 12 (11), 756–766.
Park, C., Qian, W., Zhang, J., 2012. Genomic evidence for elevated mutation rates in highly expressed genes. EMBO Rep. 13 (12), 1123–1129.
Matic, I., 2019. Mutation rate heterogeneity increases odds of survival in unpredictable environments. Mol. Cell 75 (3), 421–425.
Wallberg, A., Glémin, S., Webster, M.T., 2015. Extreme recombination frequencies shape genome variation and evolution in the honeybee, *Apis mellifera*. PLoS Genet. 11 (4), e1005189.
Hey, J., 2004. What's so hot about recombination hotspots? PLoS Biol. 2 (6), e190.