Decoding rRNA sequences for improved metagenomics in sylvatic mosquito species

Cassandra Koh[1], Lionel Frangeul[1], Hervé Blanc[1], Carine Ngoagouni[2], Sébastien Boyer[3], Philippe Dussart[4], Nina Grau[5], Romain Girod[5], Jean-Bernard Duchemin[6] and Maria-Carla Saleh[1*]

[1] Viruses and RNA Interference Unit, CNRS-URM3569, Department of Virology, Institut Pasteur, Paris Cedex 15 75724, France
[2] Medical Entomology Laboratory, Institut Pasteur de Bangui, Bangui PO Box 923, Central African Republic
[3] Medical and Veterinary Entomology Unit, Institut Pasteur du Cambodge, Phnom Penh 12201, Cambodia
[4] Virology Unit, Institut Pasteur du Cambodge, Phnom Penh 12201, Cambodia
[5] Medical Entomology Unit, Institut Pasteur de Madagascar, Antananarivo 101, Madagascar
[6] Vectopôle Amazonien Emile Abonnenc, Institut Pasteur de la Guyane, Cayenne 97306, French Guiana

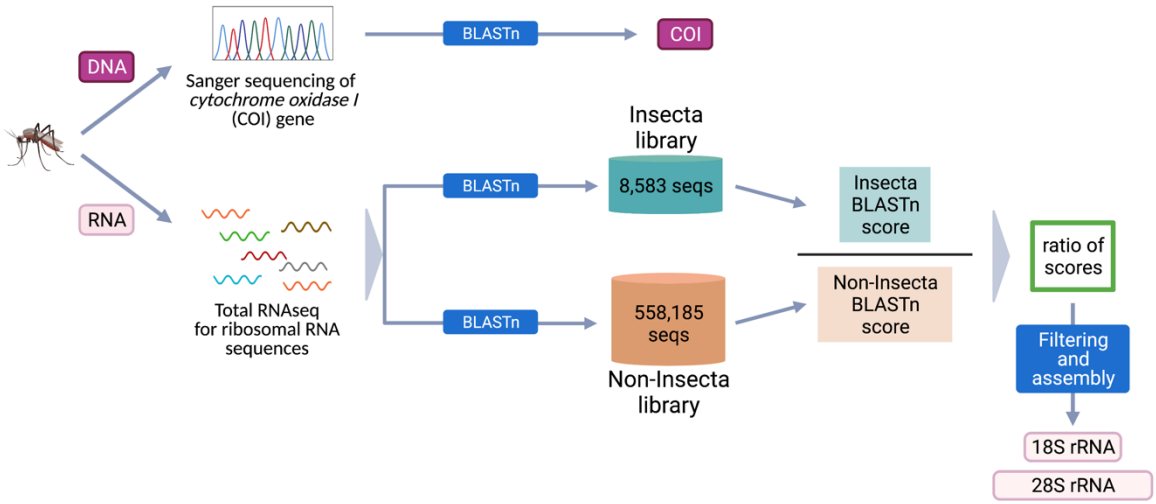
* To whom correspondence should be addressed. Tel: +33 1 45 68 85 47; Fax: +33 1 45 68 87 27; Email: carla.saleh@pasteur.fr

Present address: Philippe Dussart, Institut Pasteur de Madagascar, Antananarivo 101, Madagascar; Nina Grau, Sciences Economiques et Sociales de la Santé et Traitement de l'Information Médicale, Faculté de Médecine, Marseille 13005, France

ABSTRACT

RNAseq-based metagenomics on mosquitoes are frequently performed for surveillance or microbiome discovery studies to understand the disease ecology of arboviruses. A major hurdle to overcome in these studies is the depletion of abundant rRNA, which is commonly achieved using oligo-based protocols. The lack of complete reference rRNA sequences in sequence databases for most mosquito vectors makes such studies feasible for only a narrow range of species, biasing the knowledge on mosquitoes and their viruses. We performed total RNAseq and mitochondrial *cytochrome c oxidase I* gene sequencing on specimens of 29 mosquito species to compare the functionality of rRNA sequences as an “RNA barcode” for taxonomy and phylogeny. Using a score-based strategy to bioinformatically filter out contaminating rRNA reads, we successfully assembled full-length 28S and 18S rRNA sequences for all specimens. Phylogenetic reconstructions revealed a high degree of congruity between DNA and RNA barcodes. This assembly strategy presented here and the expansion of the rRNA reference library in public databases by 232 new complete 28S and 18S rRNA sequences provide new tools to improve surveillance and microbiome discovery for a more holistic understanding of arbovirus-transmission by mosquitoes in designing preemptive measures.

GRAPHICAL ABSTRACT



INTRODUCTION

Mosquitoes top the list of vectors for arthropod-borne diseases and are implicated in the transmission of many human pathogens responsible for arboviral diseases, malaria, and lymphatic filariasis (1). Metagenomic studies on field-captured mosquito specimens for the purposes of surveillance and microbiome or pathogen discovery are becoming increasingly important under the One Health concept, which emphasizes the importance of considering biotic and abiotic elements existing within the same ecosystem to address neglected zoonotic diseases (2). With next-generation sequencing technologies becoming more accessible, these studies are increasing in frequency, giving us an understanding of the interfaces among mosquitoes, their environment, and their animal and human hosts. Currently, there is a strong focus on only a handful species from the three genera of mosquitoes (*Aedes*, *Culex* and *Anopheles*) due to their medical importance and anthropophilic behavior despite evidence that other species are also competent for the pathogens in question. The inclusion of lesser studied but ecologically relevant species is imperative.

Ribosomal RNAs (rRNA) are non-coding RNA molecules that make up the ribosomal complexes involved in translation of messenger RNA into proteins. In eukaryotes, 28S and 18S rRNA molecules typically span lengths of two and four kilobases, respectively (3). They comprise at least 80% of the total cellular RNA population. In RNA sequencing experiments, their depletion is a necessary step during library preparation where it is not possible to selectively enrich target signals (4). To achieve this, the most common depletion protocols rely on knowledge of rRNA sequence of the species of interest. These involve hybridizing rRNA to antisense oligos (probes or primers) followed by digestion by ribonucleases (4, 5) or bead capture (6).

For well-studied mosquito species, reference rRNA sequences are readily available on public sequence databases such as GenBank or SILVA. As conventional knowledge in rDNA holds that rRNA coding regions are conserved across all species, it may seem conceivable to use rRNA oligo-based depletion protocols designed for one species on another. However, we found that within the taxon Culicidae this is not always true. There is enough sequence divergence such that *Ae. aegypti*-based probes produced poor depletion in *Culex* and *Anopheles* mosquitoes. In addition, full-length rRNA sequences are much less represented compared to other molecular markers such as the *cytochrome* c *oxidase subunit I* (COI) gene, which is the most widely used molecular barcode for animals (7, 8). The lack of reliable rRNA depletion methods could deter metagenomic studies from expanding their sampling diversity. This narrows our knowledge of mosquito ecology to urban-dwelling species when forest-dwelling mosquitoes are the ones responsible for maintaining the sylvatic transmission of arboviruses among their reservoir hosts, which precedes autochthonous transmission in human populations (9).

To address this, we sought to determine the 28S and 18S rRNA sequences of a diverse set of sylvatic mosquito species across the Central African Republic, Madagascar, Cambodia, and French Guiana. We employed a unique score-based read filtration strategy to remove contaminating non-mosquito rRNA reads to ensure accurate *de novo* assembly and generated 120 28S and 112 18S sequences from 29 mosquito species. In parallel, we obtained COI sequences to validate morphology-based species identification and to compare phylogenetic relationships inferred from the

three molecular markers. Our findings suggest the possibility of using 28S and 18S rRNA sequences as “RNA barcodes”. This work makes possible the use of rRNA sequences for streamlined molecular species identification during RNAseq, facilitates the assembly of more rRNA sequences to expand the rRNA reference library, and enables the design of species-specific oligos for cost-effective rRNA depletion.

MATERIAL AND METHODS

Sample collection

Mosquito specimens were sampled during 2019–2020 by medical entomology teams from Institut Pasteur de Bangui (Central African Republic; CF), Institut Pasteur du Cambodge (Cambodia; KH), Institut Pasteur de Madagascar (Madagascar; MG), and Institut Pasteur de la Guyane (French Guiana; GF). Adult mosquitoes were sampled using a combination of techniques including CDC light traps, BG sentinels, and human-landing catches. Sampling sites represent non-urban locations including rural settlements (Central African Republic, Madagascar, and French Guiana) and national parks (Cambodia). Mosquitoes were identified using morphological identification keys on cold tables before preserved at −80°C and transported in cold chain to Institut Pasteur Paris for analysis. The full list of the 112 mosquito specimens used in this study and their related information are provided in Supplementary Table 1.

RNA and DNA isolation

Nucleic acids were isolated from mosquito specimens using the TRIzol reagent according to manufacturer’s protocol (Invitrogen, Thermo Fisher Scientific, Waltham, Massachusetts, USA, Cat. No. 15596018). Single mosquitoes were homogenized into 200 µL of TRIzol reagent and the rest of the reagents within the protocol were volume-adjusted accordingly. Following phase separation, RNA samples were isolated from the aqueous phase while DNA was isolated from the remaining interphase and phenol-chloroform phases. From here, RNA is used to prepare cDNA libraries for next generation sequencing and DNA is used for PCR amplification and Sanger sequencing of the mitochondrial *cytochrome* c *oxidase subunit I* (COI) gene.

Total RNA sequencing

To obtain rRNA sequences, RNA samples were quantified on a Qubit Fluorometer (Invitrogen) using the Qubit RNA BR Assay kit (Invitrogen, Cat No. Q10211) for concentration adjustment. Total RNA was used for library preparation for next generation sequencing using the NEBNext Ultra II RNA Library Preparation Kit for Illumina (New England Biolabs, Ipswich, Massachusetts, USA, Cat. No. E7770L) and the NEBNext Multiplex Oligos for Illumina (Dual Index Primers Set 1) (New England Biolabs, Cat. No. E7600S). Sequencing was performed on a NextSeq500 sequencing system (Illumina, San Diego, California, USA). Quality control of fastq data and trimming of adapters were performed with FastQC and cutadapt, respectively.

28S and 18S rRNA assembly

To obtain 28S and 18S rRNA contigs, we had to first clean our fastq library by separating the reads representing mosquito rRNA from all other reads. To achieve this, we used the SILVA sequence database to create 2 libraries: one containing all rRNA sequences recorded under the "Insecta" node of the taxonomic tree, the other containing the rRNA sequences of many others nodes distributed throughout the taxonomic tree, hence named “Non-Insecta” (10). Each read was aligned using the nucleotide Basic Local Alignment Search Tool (BLASTn, **https://blast.ncbi.nlm.nih.gov/**) of the National Center for Biotechnology Information (NCBI) against each of the two libraries and the scores of the best high-scoring pairs from the two BLASTns are subsequently used to calculate a ratio of Insecta over Non-Insecta scores (11). Only reads with a ratio greater than 0.8 were used in the assembly. The two libraries being non-exhaustive, we chose this threshold of 0.8 to eliminate only reads that were clearly of a non-insect origin. Selected reads were assembled with the SPAdes assembler using the “-rna” option, allowing more heterogeneous coverage of contigs and kmer lengths of 31 , 51 and 71 bases (12). This method successfully assembled rRNA sequences for all specimens (120 sequences for 28S and 112 sequences for 18S).

Initially, our filtration technique had two weaknesses. First, there is a relatively small number of complete rRNA sequences in the Insecta library from SILVA. To compensate for this, we carried out several filtration cycles and in between cycles, added all the complete sequences produced in previous cycles to the Insecta library. Second, when our mosquito specimens were parasitized by other insects, it was not possible to bioinformatically filter out rRNA reads belonging to the parasite. For these rare cases, we used the " --trusted-contigs" option of SPAdes, giving it access to the 28S and 18S sequences of the mosquito closest in terms of taxonomy to the one we were assembling. By doing this, the assembler was able to reconstruct the rRNA of the mosquito as well as the rRNA of the parasitizing insect.

COI amplicon sequencing

The mitochondrial COI gene was amplified from DNA samples using the primer set LCO1490 (5’-GGTCAACAAATCATAAAGATATTGG -3’) and HCO2198 (5’-TAAACTTCAGGGTGACCAAAAAATCA-3’), as is the standard for COI barcoding, producing a 658 bp product (13). PCRs were performed using Phusion High-Fidelity DNA Polymerase (Thermo Fisher Scientific, Cat. No. F530L). Every 50 µL reaction contained 10 µL of 5X High Fidelity buffer, 1 µL of 10 mM dNTPs, 2.5 µL each of 10 mM forward (LCO1490) and reverse (HCO2198) primer, 28.5 µL of water, 5 µL of DNA sample, and 0.5 µL of 2 U/µL Phusion DNA polymerase. A 3-step cycling incubation protocol was used: 98 ºC for 30 seconds; 35 cycles of 98 ºC for 10 seconds, 60 ºC for 30 seconds, and 72 ºC for 15 seconds; 72 ºC for 5 minutes followed by a 4 ºC hold. PCR products were verified to be of the correct size using gel electrophoresis and then purified from gel cut-outs using the QIAquick Gel Extraction Kit (Qiagen, Hilden, Germany, Cat. No. 28706). Sanger sequencing of the COI amplicons were performed by Eurofins Genomics, Ebersberg, Germany.

COI sequence analysis

Forward and reverse COI DNA sequences were end-trimmed to remove bases of poor quality (Q score < 30). At the 5' ends, sequences were trimmed at the same positions such that all forward sequences start with 5'- TTTTGG and all reverse sequences start with 5'- GGNTCT. Forward and reverse sequences were aligned using BLAST to produce a 621 bp consensus sequence, the minimum length of COI sequence obtained for all specimens. In cases where good quality sequences extends beyond 621 bp, forward and reverse sequences were assembled using Pearl (**https://www.gear-genomics.com/pearl/**) and manually checked for errors against trace files (14, 15). A total of 106 sequences were successfully assembled.

Assembled COI sequences were analyzed with BLASTn against the nucleotide collection (nr/nt) database to validate morphology-based species identification. In cases where the COI-based and morphology-based species identification disagree, the species identity indicated by the COI sequence is taken to be the true identity when the algorithm returns references of a single species with nucleotide similarities of >95% relative to the query sequence. In cases where matches were found within a single genus but with multiple species, specimens were indicated as an unknown member of their genus (e.g. *Culex* sp.). In cases where COI BLASTn could not reveal an unequivocal match, species identity remained as indicated by its morphology. Information of the top-scoring references for all specimens, including details of ambiguous BLASTn results, are recorded in Supplementary Table 2. All assembled COI sequences from this study are to be deposited in GenBank.

Phylogenetic analysis

Multiple sequence alignment (MSA) were performed on assembled COI and rRNA sequences using the MUSCLE software (Supplementary Data 1–4) (16, 17). As shown in Supplementary Figure 1 on the conservation of identity along the alignment, the 28S sequences contain many blocks of highly conserved nucleotides throughout the sequence, which makes the result of the multiple alignment particularly obvious. We therefore did not test other alignment programs. The multiple alignment of the COI amplicon is even more evident since no gaps are necessary for this alignment.

Phylogenetic reconstruction was performed with the MEGA X software using the maximum-likelihood method (18). The default parameters were used in addition to performing bootstrapping with 500 replications in order to be able to quantify the level of confidence in the branches of the trees obtained. For the 28S and 18S rRNA trees, two sequences belonging to an unknown species of parasitic mite from the genus *Horreolanus* found in our specimens were included to serve as an outgroup taxon. In addition, we created and analyzed a separate dataset combining our 28S rRNA sequences and full-length 28S rRNA sequences from the NCBI databases totaling 169 sequences from 58 species (12 subgenera). The sequences and their accession numbers corresponding to a specimen ID (S#, where # is an ID number) can be found in Supplementary Table 1.

RESULTS

Poor rRNA depletion using non-specific depletion methods

In library preparations of mosquito samples for RNAseq, the most common methods for depleting rRNA are commercial kits optimized for human or mice samples (19–24) or probe hybridization followed by ribonuclease digestion where the probes are 80-100 base pair antisense oligos. In cases where the reference rRNA sequence of the target species is not known, oligos would be designed based on the rRNA sequence of the closest related species (25, this study). These methods should be able to produce acceptable rRNA depletion efficiencies due to the assumption that rRNA sequences have high degrees of homology across taxa. However, in our hands we found that using probes designed for the *Ae. aegypti* rRNA sequence followed by RNase H digestion according to the protocol published by Morlan *et al.* (26) produced poor depletion in *Ae. albopictus*, and even more so in Culicine and Anopheline species (Figure 1A). Additionally, not knowing the relevant reference rRNA sequences compromises the cleanup of remaining rRNA reads from sequencing data, as reads belonging to more divergent regions may not map to a reference sequence from a different species. To solve this and to improve future RNAseq metagenomics, we sought to obtain reference rRNA sequences for 29 mosquito species across nine genera that are associated with vector activity for various pathogens (Table 1).

Table 1. List of mosquito species represented in this study and their vector status.

Mosquito taxonomy*	Origin**	Collection site	Vector for***	Reference
Aedes (Fredardsius) vittatus	CF	rural (village)	DENV, ZIKV, CHIKV, YFV	(27)
Aedes (Ochlerotatus) scapularis	GF	rural (village)	YFV	(28)
Aedes (Ochlerotatus) serratus	GF	rural (village)	YFV, OROV	(29, 30)
Aedes (Stegomyia) aegypti	CF	urban	DENV, ZIKV, CHIKV, YFV	(31)
Aedes (Stegomyia) albopictus	CF, KH	rural (village, nature reserve)	DENV, ZIKV, CHIKV, YFV, JEV	(31, 32)
Aedes (Stegomyia) simpsoni	CF	rural (village)	YFV	(33)
Anopheles (Anopheles) baezai	KH	rural (nature reserve)	unreported	–
Anopheles (Anopheles) coustani	MG, CF	rural (village)	RVFV, malaria	(34–36)
Anopheles (Cellia) funestus	MG, CF	rural (village)	ONNV, malaria	(37)
Anopheles (Cellia) gambiae	MG, CF	rural (village)	ONNV, malaria	(38)
Anopheles (Cellia) squamosus	MG	rural (village)	RVFV, malaria	(36, 39)
Coquillettidia (Rhynchotaenia) venezuelensis	GF	rural (village)	OROV	(30)
Culex (Culex) antennatus	MG	rural (village)	RVFV	(35, 36)
Culex (Culex) duttoni	CF	rural (village)	unreported	–
Culex (Culex) neavei	MG	rural (village)	USUV	(40)
Culex (Culex) orientalis	KH	rural (nature reserve)	JEV	(41)
Culex (Culex) perexiguus	MG	rural (village)	WNV	(42)
Culex (Culex) pseudovishnui	KH	rural (nature reserve)	JEV	(32, 43)
Culex (Culex) quinquefasciatus	MG, CF, KH	rural (village, nature reserve)	ZIKV, JEV, WNV, DENV, SLEV, RVFV, *Wuchereria bancrofti*	(43–45)
Culex (Culex) tritaeniorhynchus	MG, KH	rural (village, nature reserve)	JEV, WNV, RVFV	(32, 43)
Culex (Melanoconion) pedroi	GF	rural (village)	EEEV	(46)

Culex (Oculeomyia) bitaeniorhynchus	MG, KH	rural (village, nature reserve)	JEV	(32, 43)
Culex (Oculeomyia) poicilipes	MG	rural (village)	RVFV	(44)
Eretmapodites intermedius	CF	rural (village)	unreported	–
Limatus durhamii	GF	rural (village)	ZIKV	(47)
Mansonia (Mansonia) titillans	GF	rural (village)	VEEV, SLEV	(48, 49)
Mansonia (Mansonioides) indiana	KH	rural (nature reserve)	JEV	(50)
Mansonia (Mansonioides) uniformis	MG, CF, KH	rural (village, nature reserve)	WNV, RVFV, *Wuchereria bancrofti*	(37, 43, 51)
Mimomyia (Etorleptiomyia) mediolineata	MG	rural (village)	unreported	–
Psorophora (Janthinosoma) ferox	GF	rural (village)	ROCV	(52)
Uranotaenia (Uranotaenia) geometrica	GF	rural (village)	unreported	–

* () indicates subgenus

** Origin countries are listed as their ISO alpha-2 codes: Central African Republic, CF; Cambodia, KH; Madagascar, MG; French Guiana, GF.

** dengue virus, DENV; Zika virus, ZIKV; chikungunya virus, CHIKV; Yellow Fever virus, YFV; O'Nyong Nyong virus, ONNV; Oropouche virus, OROV; Rift Valley Fever virus, RVFV; Usutu virus, USUV; Japanese encephalitis virus, JEV; West Nile virus, WNV; Saint Louis encephalitis virus, SLEV; Eastern equine encephalitis virus, EEEV; Western equine encephalitis virus, VEEV; Rocio virus, ROCV.

Molecular validation of morphology-based species identification

To validate morphology-based species identification from the DNA sample of each specimen, the COI gene was amplified by PCR using the LCO1490/HCO2198 primer set. Sequences at least 620 bp long were successfully obtained for 106 specimens. BLAST analyses revealed 32 cases where top hits indicated a different species identity, taking <95% nucleotide identity as the threshold to delineate distinct species (Supplementary Table 2). All species names reported hereafter reflect identities as determined by COI barcoding except for cases where species identity could not be unequivocally determined by the COI sequence, then the morphology-based identity is retained.

For each case of disagreement in species identity, the COI sequence of the specimen was BLAST aligned against an NCBI entry representing the morphological species to verify that the new identity is a closer match by a significant margin, i.e., more than 2%. Most cases readily met this criterion with a few exceptions, which are noted in Supplementary Table 2. We found that *An. coustani* COI, regardless of specimen origin, shared remarkably high nucleotide similarity (>98%) with several other *Anopheles* species such as *An. rhodesiensis*, *An. rufipes*, *An. ziemanni*, *An. tenebrosus*, although *An. coustani* remained the most frequent and closest match. Several *Culex* specimens belonging to the *Cx. pipiens* or *Cx. vishnui* species groups returned ambiguous results as well, where the query sequence differed with either of the top two hits by only one nucleotide. For example, *Cx. quinquefasciatus* and *Cx. pipiens* of the *Cx. pipiens* group, and *Cx vishnui* and *Cx tritaeniorhynchus* of the *Cx vishnui* group.

Three specimens of *Ae. simpsoni* were initially morphologically identified as *Ae. opok* but showed 97-100% similarity to *Ae. simpsoni* COI. As NCBI contained no records of *Ae. opok* COI, we instead aligned the putative *Ae. simpsoni* COI sequences against *Ae. luteocephalus* and *Ae. africanus*, sister species of *Ae. opok* and found they shared only 90% and 89% similarity, respectively. Given this divergence, we concluded these specimens to be *Ae. simpsoni*.

Within our COI sequences, we found six unidentified *Culex* species (including two that matched to NCBI entries identified only to the genus level), four unidentified *Mansonia* species, and one

unidentified *Mimomyia* species. For *An. baezai*, no existing NCBI records were found at the time the analysis was performed. Thus, our study generated COI sequences for 10 species new to NCBI databases.

rRNA reads filtering and sequence assembly

Assembling Illumina reads to reconstruct rRNA sequences from total RNA isolated from mosquitoes is not a simple feat. Apart from host rRNA, total RNA samples also contain the rRNA from other organisms associated with the host (microbiota, parasites, or ingested diet). As all of these rRNA sequences contain highly conserved blocks, Illumina reads (150 bp) from these sequences can interfere with and impede the assembly of contigs representing the complete sequences of the 28S and 18S rRNA of the host.

Our score-based filtration strategy, described in detail under Material and Methods, allowed us to bioinformatically remove contaminating non-mosquito rRNA reads and achieved successful assembly for almost all our specimens. Figure 1B shows how, in a representative sample, the score ratio assigned to each Illumina read allowed us to segregate reads into four categories: (i) reads that map only to the Insecta library, reads that map better to the Insecta relative to Non-Insecta library, (iii) reads that map better to the Non-Insecta relative to the Insecta library, and finally (iv) reads that only map to the Non-Insecta library. By applying a conservative threshold of 0.8 to account for the reference libraries being non-exhaustive, we could filter out reads that do not originate from mosquito rRNA. Given that 15 of our specimens were engorged with vertebrate blood (Supplementary Table 1), this demonstrates that the strategy can perform as expected even with high amounts of contaminating rRNA reads. This is particularly important in studies on field-captured mosquitoes as females are often sampled containing a bloodmeal, or even captured using the human-landing catch technique.

We did encounter challenges for three specimens morphologically identified as *Ma. africana* (Specimen ID 33-35). COI amplification by PCR failed to produce any product, hence COI barcoding could not be performed to confirm species identity. In addition, SPAdes was only able to assemble partial length contigs, despite the high number of reads with scoring highly against the Insecta library. Amongst other *Mansonia* specimens, the partial contigs shared the highest similarity with the rRNA sequences of “Ma uniformis CF S51”. We then performed a guided assembly using the 28S and 18S sequences of this specimen as references, which successfully produced full length contigs. In two of these specimens (Specimen ID 34 and 35), our assembly initially produced two types of 28S and 18S rRNA sequences, one of which was similar to mosquito rRNA with low coverage and another with ten-fold more coverage and 95% nucleotide sequence similarity to an *Horreolanus* species of water mite known to parasitize mosquitoes. Our filtration strategy failed to remove the contaminating reads belonging to the parasite as its taxonomy would have given it a high BLAST score against the Insecta library. However, we were nonetheless able to obtain rRNA sequences for the mosquito as well as the unknown *Horreolanus* species. This demonstrates that our strategy can be applied to metabarcoding studies where the input material comprises multiple insect species, provided that appropriate reference genomes are available.

Altogether, we were able to assemble 120 28S and 112 18S full-length mosquito rRNA sequences. These sequences represent the first ever record for seven mosquito genera: *Coquillettidia*, *Mansonia*, *Limatus*, *Mimomyia*, *Uranotaenia*, *Psorophora,* and *Eretmapodites*. As for the *Culex*, *Aedes* and *Anopheles* genera, complete rRNA sequences were already available only for a few species. Our study provides the first record for 18 additional species.

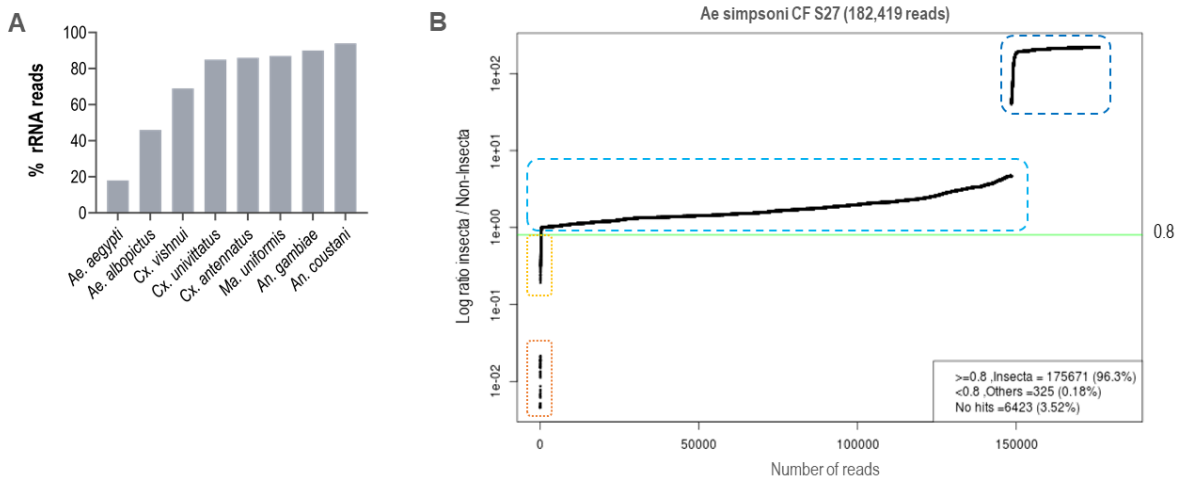


Figure 1. (**A**) Proportion of rRNA reads found in mosquito specimen pools of 5 individuals depleted by probe hybridization followed by RNase H digestion. Probes were antisense to *Ae. aegypti* rRNA sequences. (**B**) Read vs. score ratio plot of "Ae simpsoni CS S27". Green line indicates 0.8 cut-off where only reads above this threshold are used in rRNA assembly. Dark blue dashed line denotes reads with hits only in the Insecta library. Light blue dashed-line denotes reads with higher-scoring hits in the Insecta library while yellow dotted-line denotes reads with lower-scoring hits In the Insecta library, both relative to the Non-Insecta library. Orange dotted line denotes reads with only hits in the Non-Insecta library.

Discrepancies in 28S, 18S, and COI phylogeny

To verify our assembly accuracy, we constructed a phylogenetic tree from the 28S rRNA sequences generated in our study along with full-length sequences from the NCBI databases (Figure 2). We applied a search criterion for sequences with at least 95% coverage of our sequence lengths (~4000 bp), aiming to have as many species as possible represented. This tree showed that our 28S sequences clustered correctly within their respective genera with species taxa forming monophyletic clades, apart from *An. gambiae*, which was depicted to be paraphyletic and shared a common ancestor with *An. arabiensis* and *An. coluzzii*.

The rRNA phylogenetic trees benefitted from having *Horreolanus* serve as an outgroup (Figure 3. 28S rRNA, Figure 4. 18S rRNA). Our trees recapitulated what is classically known about the taxonomy of our specimens, namely (i) the division of genus *Aedes* into subgenera *Stegomyia* and *Ochlerotatus*, (ii) the division of genus *Culex* into subgenera *Culex* and *Melanoconion,* (iii) the clustering of *Culex* sister species according to their species groups, (iv) the early divergence of genus *Anopheles* from all other genera, (v) the division of genus *Anopheles* into two subgenera *Anopheles*

and *Cellia*, and (vi) the division of genus *Mansonia* into subgenera *Mansonia* and *Mansonioides* (53) (also see Table 1). This demonstrates the ability of rRNA-based phylogeny to discern taxonomy to the subgenus level. Delineation among species groups is elaborated in a later section.

The 28S and 18S trees disagree on several points. First, the *Culex* taxon consist of two branches corresponding to the subgenus *Culex* and the subgenus *Melanoconion*. These branches form a paraphyletic taxon in the 28S tree, indicating that the common ancestor of the two subgenera branches is also ancestor to the *Mansonia* (subgenus *Mansonioides*) clade. In contrast, the *Culex* taxon was monophyletic in the 18S rRNA tree. Interestingly, the phylogenies of both trees suggest that all *Culex* specimens originating from French Guiana, including two unknown species, belong to the *Melanoconion* subgenus. To note, bootstrap values within the *Culex* (subgenus *Culex*) branch were generally higher in the 28S tree than in the 18S tree. Second, both 28S and 18S trees depict the formation of a *Mansonia* (subgenus *Mansonia*) clade comprising “Ma sp.4 GF S103”, “Ma sp.4 GF S104”, and “Ma titillans GF S105” separate from the *Mansonia* (subgenus *Mansonioides*) clade. While the *Mansonia* taxon at the genus level is almost paraphyletic in the 18S tree in that the *Mansonia* and *Culex* taxa appeared to share a common ancestor, in the 28S tree the *Mansonia* (subgenus *Mansonia*) branch was much more divergent from the *Mansonia* (subgenus *Mansonioides*) branch, resulting in a polyphyletic *Mansonia* taxon. Third, the 28S tree showed the *Ps. ferox* clade to be most closely related to clade comprising the *Aedes* and *Eretmapodites* taxa, which is consistent with the fact that these genera all belong to the *Aedini* tribe. However, in the 18S tree the closest relatives of this clade were the *Mansonia* (subgenus *Mansonioides*) clade. Fourth, the *Cx. quinquefasciatus* specimens were clustered together into a clade in the 28S tree but not in the 18S tree, suggesting there is higher intraspecies divergence within the 18S rRNA sequences. Likewise, the two *Mimomyia* specimens, “Mi mediolineata MG S83” from “Mi sp.1 MG S64” formed a monophyletic clade in the 28S tree but were a polyphyletic taxon in the 18S tree.

Meanwhile, phylogenetic reconstruction based on the COI sequences showed clustering of generic taxa into clades, except for the *Culex* and *Mansonia* genera (Figure 5). The COI phylogeny we obtained here is largely congruent with those from other studies (54, 55). Being unrooted, we were unable to make any ancestral observations between taxa. In terms of the five points of differences between the 28S and 18S trees raised earlier, the COI tree showed higher congruity to the 28S tree than to the 18S tree. However, the delineation at the subgenus level for *Culex* and *Mansonia* specimens were not as clear. Although the COI tree does form a *Culex* (subgenus *Culex*) clade, the French Guianese specimens were separated between “Cx sp.DJS-2020 GF”, which appeared to be more closely related to the *Culex* (subgenus *Culex*) clade than to “Cx pedroi GF” and “Cx sp.BTLHVDV-2014”. Among the *Mansonia* specimens, while “Ma sp.4 GF S103”, “Ma sp.4 GF S104”, and “Ma titillans GF S105” did not cluster with the *Mansonia* (subgenus *Mansonioides*) clade as shown in the rRNA trees, their COI sequences did not form a *Mansonia* (subgenus *Mansonia*) clade either, instead appearing to be more closely related to species from entirely different genera.

In all rRNA trees, it is clear that the interspecies and intersubgenera evolutionary distances within the genus *Anopheles* are extremely high compared to any other genera, signifying a greater degree of divergence among its subgenera. This is evident in the longer branch lengths connecting Anopheline

species clades to the node of the most recent common ancestor for subgenera *Anopheles* and *Cellia* (Figures 2-4, Supplementary Figure 2). This feature is not evident in the COI tree, where the Anopheline interspecies distances are comparable to those within the *Culex*, *Aedes*, and *Mansonia* taxa.

We also used 28S and 18S rRNA sequences for phylogeny-based species identification for six specimens for which COI amplification by PCR was unsuccessful, hence COI barcoding could not be performed. These specimens comprise three unknown *Mansonia* species (Specimen ID 33–35), a *Ma. uniformis* (Specimen ID 51), an *An. gambiae* (Specimen ID 47), and an *Ur. geometrica* (Specimen ID 113). Their placement in the 28S rRNA tree relative to the other taxa corroborated the morphological identities of all six specimens to the genus level and to the species level for three of them (Figure 3). This was also the case in the 18S tree except for “Ma sp.3 CF S35”, which was positioned within the *Culex* clade (Figure 4). Combined with other discrepancies of the 18S phylogeny mentioned above, it would appear that 28S rRNA sequences and their derived phylogeny are more reliable than 18S rRNA sequences for molecular taxonomy.

***Culex* species groups**

Among our *Culex* (subgenus *Culex*) specimens, there are several closely related sister species belonging to the *Culex vishnui* and *Culex univittatus* species groups, which are notoriously difficult to differentiate based on morphology. Accordingly, in the 28S and COI trees these species and their known sister species were seen to cluster together in the *Culex* (subgenus *Culex*) clade: *Cx. tritaeniorhynchus* with *Cx. pseudovishnui* and *Cx. perexiguus* with *Cx. neavei*. Notably, the four unknown *Culex* species in our study were also positioned in the *Culex* (subgenus *Culex*) clade in all trees, indicating they may be sister species to those identified.

The use of COI barcoding and phylogeny to discern between members of the *Culex* species groups was successful to a degree for all our specimens but was the least confident for *Cx. tritaeniorhynchus*. Our collection includes three *Cx. tritaeniorhynchus* individuals from Cambodia and one individual from Madagascar, “Cx tritaeniorhynchus MG S115”. In the COI tree, the latter appeared to be more distantly related to the other individuals than *Cx. pseudovishnui* is, raising questions as to its species identity. However, the 28S and 18S tree both placed another *Cx tritaeniorhynchus* specimen as the closest relative to “Cx tritaeniorhynchus MG S115”.

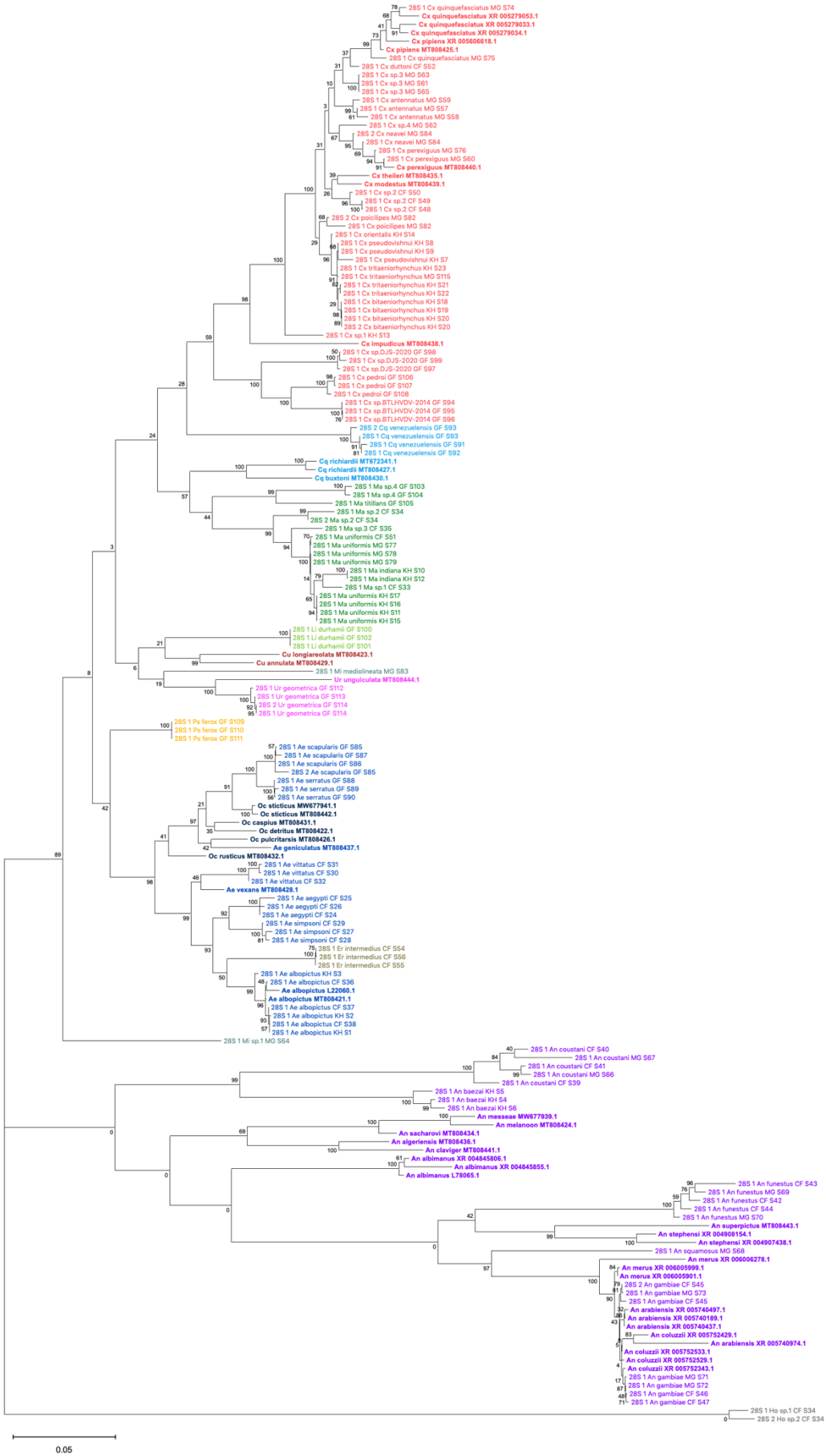

28S 1 Cx quinquefasciatus MG S74
Cx quinquefasciatus XR 005279053.1
Cx quinquefasciatus XR 005279033.1
Cx quinquefasciatus XR 005279034.1
Cx pipiens XR 005606616.1
Cx pipiens MT808425.1
28S 1 Cx quinquefasciatus MG S75
28S 1 Cx duttoni CF S52
28S 1 Cx sp.3 MG S63
28S 1 Cx sp.3 MG S61
28S 1 Cx sp.3 MG S65
28S 1 Cx antennatus MG S59
28S 1 Cx antennatus MG S57
28S 1 Cx antennatus MG S58
28S 1 Cx sp.4 MG S62
28S 2 Cx neavei MG S84
28S 1 Cx neavei MG S84
28S 1 Cx perexiguus MG S76
28S 1 Cx perexiguus MG S60
Cx perexiguus MT808440.1
Cx theileri MT808435.1
Cx modestus MT808439.1
28S 1 Cx sp.2 CF S50
28S 1 Cx sp.2 CF S49
28S 1 Cx sp.2 CF S48
28S 2 Cx poicilipes MG S82
28S 1 Cx poicilipes MG S82
28S 1 Cx orientalis KH S14
28S 1 Cx pseudovishnui KH S8
28S 1 Cx pseudovishnui KH S9
28S 1 Cx pseudovishnui KH S7
28S 1 Cx tritaeniorhynchus KH S23
28S 1 Cx tritaeniorhynchus MG S115
28S 1 Cx tritaeniorhynchus KH S21
28S 1 Cx tritaeniorhynchus KH S22
28S 1 Cx bitaeniorhynchus KH S18
28S 1 Cx bitaeniorhynchus KH S19
28S 1 Cx bitaeniorhynchus KH S20
28S 2 Cx bitaeniorhynchus KH S20
28S 1 Cx sp.1 KH S13
Cx impudicus MT808438.1
28S 1 Cx sp.DJS-2020 GF S98
28S 1 Cx sp.DJS-2020 GF S99
28S 1 Cx sp.DJS-2020 GF S97
28S 1 Cx pedroi GF S106
28S 1 Cx pedroi GF S107
28S 1 Cx pedroi GF S108
28S 1 Cx sp.BTLHVDV-2014 GF S94
28S 1 Cx sp.BTLHVDV-2014 GF S95
28S 1 Cx sp.BTLHVDV-2014 GF S96
28S 2 Cq venezuelensis GF S93
28S 1 Cq venezuelensis GF S93
28S 1 Cq venezuelensis GF S91
28S 1 Cq venezuelensis GF S92
Cq richiardii MT672341.1
Cq richiardii MT808427.1
Cq buxtonii MT808430.1
28S 1 Ma sp.4 GF S103
28S 1 Ma sp.4 GF S104
28S 1 Ma titillans GF S105
28S 1 Ma sp.2 CF S34
28S 2 Ma sp.2 CF S34
28S 1 Ma sp.3 CF S35
28S 1 Ma uniformis CF S51
28S 1 Ma uniformis MG S77
28S 1 Ma uniformis MG S78
28S 1 Ma uniformis MG S79
28S 1 Ma indiana KH S10
28S 1 Ma indiana KH S12
28S 1 Ma sp.1 CF S33
28S 1 Ma uniformis KH S17
28S 1 Ma uniformis KH S16
28S 1 Ma uniformis KH S11
28S 1 Ma uniformis KH S15
28S 1 Li durhamii GF S100
28S 1 Li durhamii GF S102
28S 1 Li durhamii GF S101
Cu longiareolata MT808423.1
Cu annulata MT808429.1
28S 1 Mi mediolineata MG S83
Ur unguiculata MT808444.1
28S 1 Ur geometrica GF S112
28S 1 Ur geometrica GF S113
28S 2 Ur geometrica GF S114
28S 1 Ur geometrica GF S114
28S 1 Ps ferox GF S109
28S 1 Ps ferox GF S110
28S 1 Ps ferox GF S111
28S 1 Ae scapularis GF S85
28S 1 Ae scapularis GF S87
28S 1 Ae scapularis GF S86
28S 2 Ae scapularis GF S85
28S 1 Ae serratus GF S88
28S 1 Ae serratus GF S89
28S 1 Ae serratus GF S90
Oc sticticus MW677941.1
Oc sticticus MT808442.1
Oc caspius MT808431.1
Oc detritus MT808422.1
Oc pulcritarsis MT808426.1
Ae geniculatus MT808437.1
Oc rusticus MT808432.1
28S 1 Ae vittatus CF S31
28S 1 Ae vittatus CF S30
28S 1 Ae vittatus CF S32
Ae vexans MT808428.1
28S 1 Ae aegypti CF S25
28S 1 Ae aegypti CF S26
28S 1 Ae aegypti CF S24
28S 1 Ae simpsoni CF S29
28S 1 Ae simpsoni CF S27
28S 1 Ae simpsoni CF S28
28S 1 Er intermedius CF S54
28S 1 Er intermedius CF S56
28S 1 Er intermedius CF S55
28S 1 Ae albopictus KH S3
28S 1 Ae albopictus CF S36
Ae albopictus L22060.1
Ae albopictus MT808421.1
28S 1 Ae albopictus CF S37
28S 1 Ae albopictus KH S2
28S 1 Ae albopictus CF S38
28S 1 Ae albopictus KH S1
28S 1 Mi sp.1 MG S64
28S 1 An coustani CF S40
28S 1 An coustani MG S67
28S 1 An coustani CF S41
28S 1 An coustani MG S66
28S 1 An coustani CF S39
28S 1 An baezai KH S5
28S 1 An baezai KH S4
28S 1 An baezai KH S6
An messeae MW677939.1
An melanoon MT808424.1
An sacharovi MT808434.1
An algeriensis MT808436.1
An claviger MT808441.1
An albimanus XR 004845806.1
An albimanus XR 004845855.1
An albimanus L78065.1
28S 1 An funestus CF S43
28S 1 An funestus MG S69
28S 1 An funestus CF S42
28S 1 An funestus CF S44
28S 1 An funestus MG S70
An superpictus MT808443.1
An stephensi XR 004908154.1
An stephensi XR 004907438.1
28S 1 An squamosus MG S68
An merus XR 006006278.1
An merus XR 006005999.1
An merus XR 006005901.1
28S 2 An gambiae CF S45
28S 1 An gambiae MG S73
28S 1 An gambiae CF S45
An arabiensis XR 005740497.1
An arabiensis XR 005740189.1
An arabiensis XR 005740437.1
An coluzzii XR 005752429.1
An arabiensis XR 005740974.1
An coluzzii XR 005752533.1
An coluzzii XR 005752529.1
An coluzzii XR 005752343.1
28S 1 An gambiae MG S71
28S 1 An gambiae MG S72
28S 1 An gambiae CF S46
28S 1 An gambiae CF S47
28S 1 Ho sp.1 CF S34
28S 2 Ho sp.2 CF S34
0.05

Figure 2. Phylogenetic tree based on 28S sequences generated from this study and from NCBI databases combined (3900 bp) as inferred using maximum-likelihood method and constructed to scale in MEGA X (18). Values at each node indicate bootstrap support from 500 replications. For sequences from this study, each specimen label contains information on its taxonomy, origin (as indicated in 2-letter country codes), and specimen ID. Labels in bold indicate sequences derived from NCBI. Label colors indicate genera: *Culex* in coral, *Anopheles* in purple, *Aedes* in dark blue, *Mansonia* in dark green, *Culiseta* in maroon, *Limatus* in light green, *Coquillettidia* in light blue, *Psorophora* in yellow, *Mimomyia* in teal, *Uranotaenia* in pink and *Eretmapodites* in brown. Scale bar at 0.05 is shown.

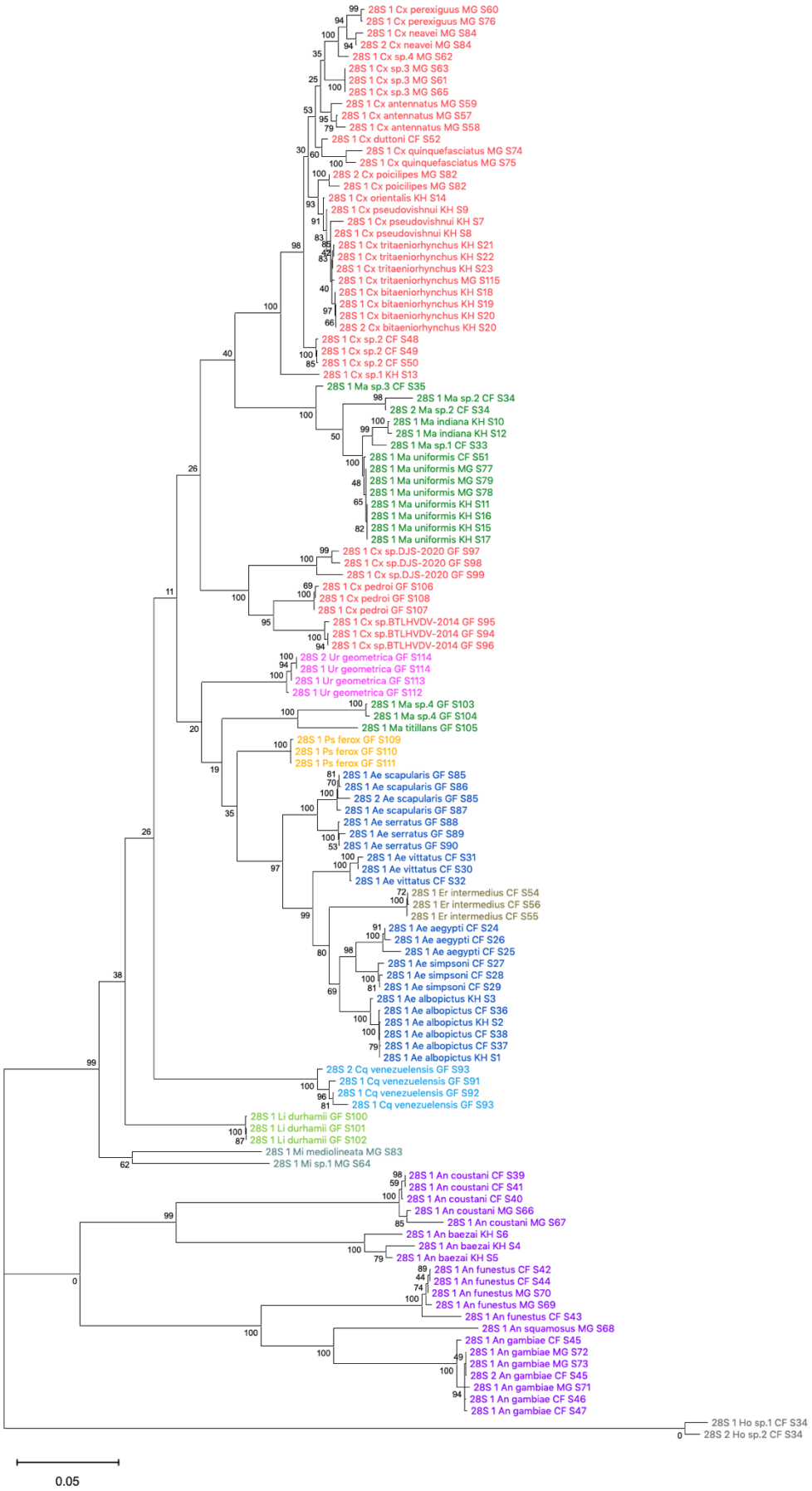

28S 1 Cx perexiguus MG S60
28S 1 Cx perexiguus MG S76
28S 1 Cx neavei MG S84
28S 2 Cx neavei MG S84
28S 1 Cx sp.4 MG S62
28S 1 Cx sp.3 MG S63
28S 1 Cx sp.3 MG S61
28S 1 Cx sp.3 MG S65
28S 1 Cx antennatus MG S59
28S 1 Cx antennatus MG S57
28S 1 Cx antennatus MG S58
28S 1 Cx duttoni CF S52
28S 1 Cx quinquefasciatus MG S74
28S 1 Cx quinquefasciatus MG S75
28S 2 Cx poicilipes MG S82
28S 1 Cx poicilipes MG S82
28S 1 Cx orientalis KH S14
28S 1 Cx pseudovishnui KH S9
28S 1 Cx pseudovishnui KH S7
28S 1 Cx pseudovishnui KH S8
28S 1 Cx tritaeniorhynchus KH S21
28S 1 Cx tritaeniorhynchus KH S22
28S 1 Cx tritaeniorhynchus KH S23
28S 1 Cx tritaeniorhynchus MG S115
28S 1 Cx bitaeniorhynchus KH S18
28S 1 Cx bitaeniorhynchus KH S19
28S 1 Cx bitaeniorhynchus KH S20
28S 2 Cx bitaeniorhynchus KH S20
28S 1 Cx sp.2 CF S48
28S 1 Cx sp.2 CF S49
28S 1 Cx sp.2 CF S50
28S 1 Cx sp.1 KH S13
28S 1 Ma sp.3 CF S35
28S 1 Ma sp.2 CF S34
28S 2 Ma sp.2 CF S34
28S 1 Ma indiana KH S10
28S 1 Ma indiana KH S12
28S 1 Ma sp.1 CF S33
28S 1 Ma uniformis CF S51
28S 1 Ma uniformis MG S77
28S 1 Ma uniformis MG S79
28S 1 Ma uniformis MG S78
28S 1 Ma uniformis KH S11
28S 1 Ma uniformis KH S16
28S 1 Ma uniformis KH S15
28S 1 Ma uniformis KH S17
28S 1 Cx sp.DJS-2020 GF S97
28S 1 Cx sp.DJS-2020 GF S98
28S 1 Cx sp.DJS-2020 GF S99
28S 1 Cx pedroi GF S106
28S 1 Cx pedroi GF S108
28S 1 Cx pedroi GF S107
28S 1 Cx sp.BTLHVDV-2014 GF S95
28S 1 Cx sp.BTLHVDV-2014 GF S94
28S 1 Cx sp.BTLHVDV-2014 GF S96
28S 2 Ur geometrica GF S114
28S 1 Ur geometrica GF S114
28S 1 Ur geometrica GF S113
28S 1 Ur geometrica GF S112
28S 1 Ma sp.4 GF S103
28S 1 Ma sp.4 GF S104
28S 1 Ma titillans GF S105
28S 1 Ps ferox GF S109
28S 1 Ps ferox GF S110
28S 1 Ps ferox GF S111
28S 1 Ae scapularis GF S85
28S 1 Ae scapularis GF S86
28S 2 Ae scapularis GF S85
28S 1 Ae scapularis GF S87
28S 1 Ae serratus GF S88
28S 1 Ae serratus GF S89
28S 1 Ae serratus GF S90
28S 1 Ae vittatus CF S31
28S 1 Ae vittatus CF S30
28S 1 Ae vittatus CF S32
28S 1 Er intermedius CF S54
28S 1 Er intermedius CF S56
28S 1 Er intermedius CF S55
28S 1 Ae aegypti CF S24
28S 1 Ae aegypti CF S26
28S 1 Ae aegypti CF S25
28S 1 Ae simpsoni CF S27
28S 1 Ae simpsoni CF S28
28S 1 Ae simpsoni CF S29
28S 1 Ae albopictus KH S3
28S 1 Ae albopictus CF S36
28S 1 Ae albopictus KH S2
28S 1 Ae albopictus CF S38
28S 1 Ae albopictus CF S37
28S 1 Ae albopictus KH S1
28S 2 Cq venezuelensis GF S93
28S 1 Cq venezuelensis GF S91
28S 1 Cq venezuelensis GF S92
28S 1 Cq venezuelensis GF S93
28S 1 Li durhamii GF S100
28S 1 Li durhamii GF S101
28S 1 Li durhamii GF S102
28S 1 Mi mediolineata MG S83
28S 1 Mi sp.1 MG S64
28S 1 An coustani CF S39
28S 1 An coustani CF S41
28S 1 An coustani CF S40
28S 1 An coustani MG S66
28S 1 An coustani MG S67
28S 1 An baezai KH S6
28S 1 An baezai KH S4
28S 1 An baezai KH S5
28S 1 An funestus CF S42
28S 1 An funestus CF S44
28S 1 An funestus MG S70
28S 1 An funestus MG S69
28S 1 An funestus CF S43
28S 1 An squamosus MG S68
28S 1 An gambiae CF S45
28S 1 An gambiae MG S72
28S 1 An gambiae MG S73
28S 2 An gambiae CF S45
28S 1 An gambiae MG S71
28S 1 An gambiae CF S46
28S 1 An gambiae CF S47
28S 1 Ho sp.1 CF S34
28S 2 Ho sp.2 CF S34
0.05

Figure 3. Phylogenetic tree based on 28S sequences generated from this study alone (3900 bp) as inferred using maximum-likelihood method and constructed to scale in MEGA X (18). Values at each node indicate bootstrap support from 500 replication. Each specimen label contains information on its taxonomy, origin (as indicated in 2-letter country codes), and specimen ID. Label colors indicate genera: *Culex* in coral, *Anopheles* in purple, *Aedes* in dark blue, *Mansonia* in dark green, *Limatus* in light green, *Coquillettidia* in light blue, *Psorophora* in yellow, *Mimomyia* in teal, *Uranotaenia* in pink and *Eretmapodites* in brown. Scale bar at 0.05 is shown.

18S Cx pseudovishnui KH S8
18S Cx pseudovishnui KH S9
18S Cx pseudovishnui KH S7
18S Cx tritaeniorhynchus KH S21
18S Cx tritaeniorhynchus KH S22
18S Cx poicilipes MG S82
18S Cx orientalis KH S14
18S Cx tritaeniorhynchus KH S23
18S Cx tritaeniorhynchus MG S115
18S Cx bitaeniorhynchus KH S19
18S Cx bitaeniorhynchus KH S18
18S Cx bitaeniorhynchus KH S20
18S Cx sp.2 CF S48
18S Cx sp.2 CF S50
18S Cx sp.2 CF S49
18S Cx sp.1 KH S13
18S Cx neavei MG S84
18S Cx sp.3 MG S65
18S Cx sp.3 MG S61
18S Cx sp.3 MG S63
18S Cx duttoni CF S52
18S Cx quinquefasciatus MG S74
18S Cx sp.4 MG S62
18S Cx quinquefasciatus MG S75
18S Cx perexiguus MG S60
18S Cx perexiguus MG S76
18S Cx antennatus MG S57
18S Cx antennatus MG S59
18S Cx antennatus MG S58
18S Ma sp.3 CF S35
18S Cx sp.BTLHVDV-2014 GF S95
18S Cx sp.BTLHVDV-2014 GF S96
18S Cx sp.BTLHVDV-2014 GF S94
18S Cx pedroi GF S106
18S Cx pedroi GF S108
18S Cx pedroi GF S107
18S Cx sp.DJS-2020 GF S97
18S Cx sp.DJS-2020 GF S98
18S Cx sp.DJS-2020 GF S99
18S Ps ferox GF S110
18S Ps ferox GF S111
18S Ps ferox GF S109
18S Ma sp.2 CF S34
18S Ma indiana KH S10
18S Ma indiana KH S12
18S Ma uniformis CF S51
18S Ma uniformis MG S78
18S Ma uniformis MG S77
18S Ma sp.1 CF S33
18S Ma uniformis MG S79
18S Ma uniformis KH S15
18S Ma uniformis KH S11
18S Ma uniformis KH S17
18S Ma uniformis KH S16
18S Ma titillans GF S105
18S Ma sp.4 GF S103
18S Ma sp.4 GF S104
18S Cq venezuelensis GF S91
18S Cq venezuelensis GF S92
18S Cq venezuelensis GF S93
18S Mi mediolineata MG S83
18S Ae scapularis GF S86
18S Ae scapularis GF S87
18S Ae scapularis GF S85
18S Ae serratus GF S90
18S Ae serratus GF S88
18S Ae serratus GF S89
18S Ae vittatus CF S30
18S Ae vittatus CF S31
18S Ae vittatus CF S32
18S Er intermedius CF S55
18S Er intermedius CF S54
18S Er intermedius CF S56
18S Ae aegypti CF S25
18S Ae aegypti CF S26
18S Ae aegypti CF S24
18S Ae simpsoni CF S27
18S Ae simpsoni CF S29
18S Ae simpsoni CF S28
18S Ae albopictus CF S37
18S Ae albopictus CF S38
18S Ae albopictus CF S36
18S Ae albopictus KH S2
18S Ae albopictus KH S1
18S Ae albopictus KH S3
18S Ur geometrica GF S112
18S Ur geometrica GF S113
18S Ur geometrica GF S114
18S Li durhamii GF S102
18S Li durhamii GF S100
18S Li durhamii GF S101
18S Mi sp.1 MG S64
18S An coustani CF S41
18S An coustani MG S67
18S An coustani MG S66
18S An coustani CF S40
18S An coustani CF S39
18S An baezai KH S5
18S An baezai KH S6
18S An baezai KH S4
18S An funestus CF S42
18S An funestus CF S44
18S An funestus MG S69
18S An funestus CF S43
18S An funestus MG S70
18S An squamosus MG S68
18S An gambiae CF S45
18S An gambiae MG S71
18S An gambiae MG S72
18S An gambiae MG S73
18S An gambiae CF S46
18S An gambiae CF S47
18S Ho sp.1 CF S34
18S Ho sp.2 CF S35
0.05

Figure 4. Phylogenetic tree based on 18S sequences (1900 bp) as inferred using maximum-likelihood method and constructed to scale in MEGA X (18). Values at each node indicate bootstrap support from 500 replications. Each specimen label contains information on its taxonomy, origin (as indicated in 2-letter country codes), and specimen ID. Label colors indicate genera: *Culex* in coral, *Anopheles* in purple, *Aedes* in dark blue, *Mansonia* in dark green, *Limatus* in light green, *Coquillettidia* in light blue, *Psorophora* in yellow, *Mimomyia* in teal, *Uranotaenia* in pink and *Eretmapodites* in brown. Scale bar at 0.05 is shown.

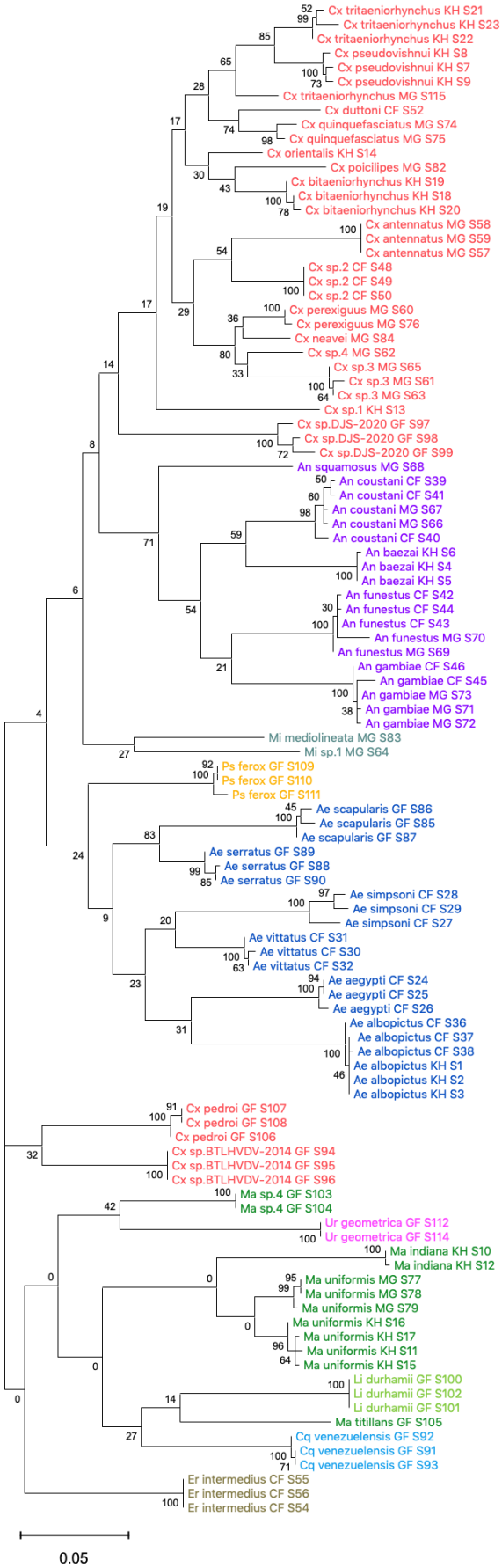

Cx tritaeniorhynchus KH S21
Cx tritaeniorhynchus KH S23
Cx tritaeniorhynchus KH S22
Cx pseudovishnui KH S8
Cx pseudovishnui KH S7
Cx pseudovishnui KH S9
Cx tritaeniorhynchus MG S115
Cx duttoni CF S52
Cx quinquefasciatus MG S74
Cx quinquefasciatus MG S75
Cx orientalis KH S14
Cx poicilipes MG S82
Cx bitaeniorhynchus KH S19
Cx bitaeniorhynchus KH S18
Cx bitaeniorhynchus KH S20
Cx antennatus MG S58
Cx antennatus MG S59
Cx antennatus MG S57
Cx sp.2 CF S48
Cx sp.2 CF S49
Cx sp.2 CF S50
Cx perexiguus MG S60
Cx perexiguus MG S76
Cx neavei MG S84
Cx sp.4 MG S62
Cx sp.3 MG S65
Cx sp.3 MG S61
Cx sp.3 MG S63
Cx sp.1 KH S13
Cx sp.DJS-2020 GF S97
Cx sp.DJS-2020 GF S98
Cx sp.DJS-2020 GF S99
An squamosus MG S68
An coustani CF S39
An coustani CF S41
An coustani MG S67
An coustani MG S66
An coustani CF S40
An baezai KH S6
An baezai KH S4
An baezai KH S5
An funestus CF S42
An funestus CF S44
An funestus CF S43
An funestus MG S70
An funestus MG S69
An gambiae CF S46
An gambiae CF S45
An gambiae MG S73
An gambiae MG S71
An gambiae MG S72
Mi mediolineata MG S83
Mi sp.1 MG S64
Ps ferox GF S109
Ps ferox GF S110
Ps ferox GF S111
Ae scapularis GF S86
Ae scapularis GF S85
Ae scapularis GF S87
Ae serratus GF S89
Ae serratus GF S88
Ae serratus GF S90
Ae simpsoni CF S28
Ae simpsoni CF S29
Ae simpsoni CF S27
Ae vittatus CF S31
Ae vittatus CF S30
Ae vittatus CF S32
Ae aegypti CF S24
Ae aegypti CF S25
Ae aegypti CF S26
Ae albopictus CF S36
Ae albopictus CF S37
Ae albopictus CF S38
Ae albopictus KH S1
Ae albopictus KH S2
Ae albopictus KH S3
Cx pedroi GF S107
Cx pedroi GF S108
Cx pedroi GF S106
Cx sp.BTLHVDV-2014 GF S94
Cx sp.BTLHVDV-2014 GF S95
Cx sp.BTLHVDV-2014 GF S96
Ma sp.4 GF S103
Ma sp.4 GF S104
Ur geometrica GF S112
Ur geometrica GF S114
Ma indiana KH S10
Ma indiana KH S12
Ma uniformis MG S77
Ma uniformis MG S78
Ma uniformis MG S79
Ma uniformis KH S16
Ma uniformis KH S17
Ma uniformis KH S11
Ma uniformis KH S15
Li durhamii GF S100
Li durhamii GF S102
Li durhamii GF S101
Ma titillans GF S105
Cq venezuelensis GF S92
Cq venezuelensis GF S91
Cq venezuelensis GF S93
Er intermedius CF S55
Er intermedius CF S56
Er intermedius CF S54
0.05

Figure 5. Phylogenetic tree based on COI sequences (620–699 bp) as inferred using maximum-likelihood method and constructed to scale in MEGA X (18). Values at each node indicate bootstrap support from 500 replications. Each specimen label contains information on its taxonomy, origin (as indicated in 2-letter country codes), and specimen ID. Label colors indicate genera: *Culex* in coral, *Anopheles* in purple, *Aedes* in dark blue, *Mansonia* in dark green, *Limatus* in light green, *Coquillettidia* in light blue, *Psorophora* in yellow, *Mimomyia* in teal, *Uranotaenia* in pink and *Eretmapodites* in brown. Scale bar at 0.05 is shown.

DISCUSSION

Ribosomal RNA has posed a problem in next generation RNAseq since the invention of the technique. In metagenomic studies of field-captured mosquitoes, the lack of reference rRNA sequences hinders oligo-based depletion and efficient cleanup of RNAseq data. Additionally, *de novo* assembly rRNA sequences is complicated due to the high conservation in rRNA sequences across all distantly related organisms that could be present in a single specimen, i.e., microbiota, parasites, or vertebrate bloodmeal. How can we select for only mosquito reads for assembler input? Our goal in this study was to establish a reliable method to bioinformatically filter contaminating rRNA reads for the accurate assembly of 28S and 18S rRNA sequences. Further, we compared the phylogenetic reconstructions based on rRNA sequences to that of the mitochondrial COI sequence, a commonly used molecular marker.

We found that phylogenetic trees based on rRNA sequences were able to cluster mosquito taxa according to species correctly. The phylogenetic reconstruction of a combined dataset of the 28S rRNA sequences from this study and those already present in the NCBI databases corroborates current mosquito taxonomy knowledge. These demonstrate that our bioinformatics methodology was able to reliably assemble bona fide 28S and 18S rRNA sequences, even in specimens parasitized by water mites or engorged with vertebrate blood. Importantly, we were able to use 28S rRNA phylogeny to perform molecular species identification when COI sequences were unavailable or ambiguous. We thus propose the use of 28S rRNA sequences as an “RNA barcode” that can complement COI barcoding, with the added advantage that in RNAseq studies, 28S rRNA barcoding can be performed from RNA reads and circumvents the need to additionally isolate DNA from specimens. Even after depletion, there are sufficient numbers of reads remaining to assemble complete rRNA contigs (unpublished data).

The topologies of the 28S and 18S trees differ for two taxa, “Ma sp.3 CF SS35” and “Mi mediolineata MG S83” where the placement within the 28S tree is more similar to that within the COI tree. Given that the species of Ma sp.3 is unknown and there is a lack of representation of *Mimomyia* species in our phylogeny, it was not possible to investigate further whether this was a specimen-specific occurrence or a true phylogenetic pattern for genus *Mimomyia*. It may be possible to resolve this and the other discrepancies between 28S and 18S phylogenies by performing the analysis instead on concatenated 28S+18S sequences, given they are in any case encoded by linked genes.

Relying on morphological markers for species identification has its limits when it comes to *Culex* species groups. Aside from sharing many morphological traits, sister species within these *Culex* species groups are often sympatric and show at least some competence for a number of viral and filarial pathogens, such as Japanese encephalitis virus, St Louis encephalitic virus, Usutu virus, and *Wuchereria bancrofti* (56). However, each of these species have distinct bionomics and host preferences, thus the challenge of correctly identifying vector species can cloud epidemiological risk estimation for these diseases (57).

The *Cx. pipiens* group is especially challenging as its member species have been shown to interbreed and produce hybrids, showing genetic introgression to varying extents depending on the geographical location (11). The seven member species of this complex are practically

indistinguishable morphologically and require molecular methods to discern (57, 58). However, the 621 bp COI sequence does not contain enough nucleotide divergence to allow clear identification, as seen in this study where the COI of *Cx. quinquefasciatus* specimens differ from the COI of *Cx. pipiens* by only one nucleotide. Other DNA molecular markers such as nuclear Ace-2 or CQ11 genes (58, 59) or *Wolbachia pipientis* infection status (60) are typically employed in tandem. This problem of cryptic sister species also permeates the *Cx. vishnui* group. The morphological differences between the *Culex* species *bitaeniorhynchus*, *tritaeniorhynchus*, *vishnui*, and *pseudovishnui* are thin, the former three having been morphologically detected in our study but later revealed by COI barcoding to be a different species. Here, we used 28S and 18S phylogenies to resolve an ambiguity in the COI barcoding of a *Cx. tritaeniorhynchus* specimen, "Cx tritaeniorhynchus MG S115". The combination of 28S and 18S rRNA sequences, being longer, contain enough resolution in nucleotide divergence to distinguish between these cryptic species. This should allow for more accurate species identification and, therefore, bionomics in the context of disease transmission. Additionally, tracing the genetic flow across hybrid populations within the *Cx. pipiens* species group can inform estimates of vectorial capacity for each species. Only one or two members from the *Cx. pipiens* and *Cx. vishnui* species groups were represented in our dataset. A focused investigation including all member species of a species group in greater sample numbers is warranted to test the degree of accuracy with which 28S and 18S rRNA sequences can delineate sister species and how this compares to COI barcoding.

The world of *Anopheles* is much wider compared to any other genera. Our 28S and 18S rRNA phylogenies showed remarkably high divergence between one Anopheline species to another. This feature is not observed in the COI phylogeny, perhaps reflecting differences in the mutation accumulation rate in the Anopheline nuclear genome to the mitochondrial genome. This may lead to an underestimation of Anopheline species relatedness and suggests that 28S and 18S rRNA sequences, like rDNA sequences, are better able to illustrate evolutionary history than COI sequence. It would be interesting to further explore divergence among other Anopheline subgenera beyond the subgenera *Anopheles* and *Cellia* represented in this study. We also stress the importance of noting detailed taxonomy of mosquito specimens when reporting sequence data when it comes to mosquitoes as we found that many Anopheline rRNA entries lack subgenus information.

We intend for the rRNA sequences presented in this study to improve metagenomic studies of mosquitoes by facilitating the optimization of species-specific oligo-based depletion protocols and by providing reference sequences for species identification by "rRNA barcoding" and for RNAseq data cleanup. The mosquito metagenomics community could benefit from the sharing of more mosquito rRNA sequences on public databases, which would also reveal in time how the rRNA-based phylogenies measure up against other more widely used DNA barcoding markers, such as mitochondrial COI or 16S rDNA and the ribosomal internal transcribed region 2 (8, 61).

Conclusions

Surveillance and microbiome discovery studies in wild mosquitoes are paramount for the establishment of measures to prepare for and prevent outbreaks of arboviral diseases. The assembly strategy and 232 new 28S and 18S mosquito rRNA sequences we report in this work make future

RNAseq-based metagenomics on a wider range of mosquito species and other insects achievable by allowing for improved depletion protocols and streamlined species identification from RNA reads. With a large enough reference library, 28S rRNA sequences can serve as a complementary "RNA barcode" for taxonomic and phylogenetic investigations alongside established DNA barcodes. We have already shown here that 28S rRNA phylogenetic tree shares a high degree of congruity with that of COI and can be used in place of COI barcoding for species identification. Our work has expanded the rRNA reference library by presenting, to our knowledge, the first records for many species not previously present in public databases and has paved the way for the assembly of many more.

AVAILABILITY

RNAseq fastq sequence data are available upon request. Multiple sequence alignment files are available as supplementary data.

ACCESSION NUMBERS

All sequences generated in this study (rRNA and COI gene sequences) will be submitted to GenBank.

SUPPLEMENTARY DATA

Supplementary Table 1. Taxonomic and sampling information on mosquito specimens and associated accession numbers for 28S, 18S, and COI sequences.

Supplementary Table 2. BLAST analyses summary on COI sequences.

Supplementary Figure 1. Sequence conservation among 169 28S rRNA sequences obtained from this study and from the NCBI databases.

Supplementary Figure 2. Phylogenetic tree based on 28S sequences generated from this study and from NCBI databases combined (3900 bp) as inferred using maximum-likelihood method and constructed to scale in MEGA X in radial format (18). Values at each node indicate bootstrap support from 500 replications. For sequences from this study, each specimen label contains information on its taxonomy, origin (as indicated in 2-letter country codes), and specimen ID. Labels in bold indicate sequences derived from NCBI. Label colors indicate genera: *Culex* in coral, *Anopheles* in purple, *Aedes* in dark blue, *Mansonia* in dark green, *Culiseta* in maroon, *Limatus* in light green, *Coquillettidia* in light blue, *Psorophora* in yellow, *Mimomyia* in teal, *Uranotaenia* in pink and *Eretmapodites* in brown. Scale bar at 0.05 is shown.

Supplementary Data 1. Multiple sequence alignment of 106 COI sequences.

Supplementary Data 2. Multiple sequence alignment of 114 18S rRNA sequences, including two sequences from *Horreolanus sp*.

Supplementary Data 3. Multiple sequence alignment of 122 28S rRNA sequences, including two sequences from *Horreolanus sp*.

Supplementary Data 4. Multiple sequence alignment of 169 28S rRNA sequences from this study and from NCBI databases.

ACKNOWLEDGEMENT

We thank members of the Saleh lab for valuable discussions and to Inès Partouche for laboratory assistance. We especially thank all medical entomology staff of IP Bangui, IP Cambodge (Sony Yean, Kimly Heng, Kalyan Chhuoy, Sreynik Nhek, Moeun Chhum, Kimhuor Sour and Pierre-Olivier Maquart), IP Madagascar, and IP Guyane for assistance in field missions, laboratory work, and logistics. We are also grateful to Dr Catherine Dauga for advice on phylogenetic analyses.

FUNDING

This work was supported by the Defence Advanced Research Projects Agency PREEMPT program managed by Dr. Rohit Chitale and Dr. Kerri Dugan [Cooperative Agreement HR001118S0017] (the content of the information does not necessarily reflect the position or the policy of the U.S. government, and no official endorsement should be inferred).

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

REFERENCES

1. WHO (2017) Global Vector Control Response 2017–2030.
2. Webster,J.P., Gower,C.M., Knowles,S.C.L., Molyneux,D.H. and Fenton,A. (2016) One health – an ecological and evolutionary framework for tackling Neglected Zoonotic Diseases. *Evol. Appl.*, **9**, 313–333.
3. GALE,K. and CRAMPTON,J. (1989) The ribosomal genes of the mosquito, Aedes aegypti. *Eur. J. Biochem.*, 10.1111/j.1432-1033.1989.tb15117.x.
4. Phelps,W.A., Carlson,A.E. and Lee,M.T. (2021) Optimized design of antisense oligomers for targeted rRNA depletion. *Nucleic Acids Res.*, 10.1093/nar/gkaa1072.
5. Fauver,J.R., Akter,S., Morales,A.I.O., Black,W.C., Rodriguez,A.D., Stenglein,M.D., Ebel,G.D. and Weger-Lucarelli,J. (2019) A reverse-transcription/RNase H based protocol for depletion of mosquito ribosomal RNA facilitates viral intrahost evolution analysis, transcriptomics and pathogen discovery. *Virology*, **528**, 181–197.
6. Kukutla,P., Steritz,M. and Xu,J. (2013) Depletion of ribosomal RNA for mosquito gut metagenomic RNA-seq. *J. Vis. Exp.*, 10.3791/50093.
7. Hebert,P.D.N., Cywinska,A., Ball,S.L. and DeWaard,J.R. (2003) Biological identifications through DNA barcodes. *Proc. R. Soc. B Biol. Sci.*, 10.1098/rspb.2002.2218.
8. Ratnasingham,S. and Hebert,P.D.N. (2007) BOLD: The Barcode of Life Data System: Barcoding. *Mol. Ecol. Notes*, 10.1111/j.1471-8286.2007.01678.x.
9. Gould,E., Pettersson,J., Higgs,S., Charrel,R. and de Lamballerie,X. (2017) Emerging arboviruses: Why today? *One Heal.*, **4**, 1–13.
10. Quast,C., Pruesse,E., Yilmaz,P., Gerken,J., Schweer,T., Yarza,P., Peplies,J. and Glöckner,F.O. (2013) The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Res.*, 10.1093/nar/gks1219.
11. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, 10.1016/S0022-2836(05)80360-2.
12. Bankevich,A., Nurk,S., Antipov,D., Gurevich,A.A., Dvorkin,M., Kulikov,A.S., Lesin,V.M., Nikolenko,S.I., Pham,S., Prjibelski,A.D., *et al.* (2012) SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.*, 10.1089/cmb.2012.0021.

13. Folmer,O., Black,M., Hoeh,W., Lutz,R. and Vrijenhoek,R. (1994) DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Mol. Mar. Biol. Biotechnol.*, 10.1071/ZO9660275.
14. Rausch,T., Fritz,M.H.Y., Untergasser,A. and Benes,V. (2020) Tracy: Basecalling, alignment, assembly and deconvolution of sanger chromatogram trace files. *BMC Genomics*, 10.1186/s12864-020-6635-8.
15. Rausch,T., Hsi-Yang Fritz,M., Korbel,J.O. and Benes,V. (2019) Alfred: Interactive multi-sample BAM alignment statistics, feature counting and feature annotation for long- and short-read sequencing. *Bioinformatics*, 10.1093/bioinformatics/bty1007.
16. Madeira,F., Park,Y.M., Lee,J., Buso,N., Gur,T., Madhusoodanan,N., Basutkar,P., Tivey,A.R.N., Potter,S.C., Finn,R.D., *et al.* (2019) The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Res.*, 10.1093/nar/gkz268.
17. Edgar,R.C. (2004) MUSCLE: A multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, 10.1186/1471-2105-5-113.
18. Kumar,S., Stecher,G., Li,M., Knyaz,C. and Tamura,K. (2018) MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.*, 10.1093/molbev/msy096.
19. Bishop-Lilly,K.A., Turell,M.J., Willner,K.M., Butani,A., Nolan,N.M.E., Lentz,S.M., Akmal,A., Mateczun,A., Brahmbhatt,T.N., Sozhamannan,S., *et al.* (2010) Arbovirus detection in insect vectors by Rapid, high- throughput pyrosequencing. *PLoS Negl. Trop. Dis.*, 10.1371/journal.pntd.0000878.
20. Kumar,N., Creasy,T., Sun,Y., Flowers,M., Tallon,L.J. and Dunning Hotopp,J.C. (2012) Efficient subtraction of insect rRNA prior to transcriptome analysis of Wolbachia-Drosophila lateral gene transfer. *BMC Res. Notes*, 10.1186/1756-0500-5-230.
21. Weedall,G.D., Irving,H., Hughes,M.A. and Wondji,C.S. (2015) Molecular tools for studying the major malaria vector Anopheles funestus: Improving the utility of the genome using a comparative poly(A) and Ribo-Zero RNAseq analysis. *BMC Genomics*, 10.1186/s12864-015-2114-z.
22. Zakrzewski,M., Rašić,G., Darbro,J., Krause,L., Poo,Y.S., Filipović,I., Parry,R., Asgari,S., Devine,G. and Suhrbier,A. (2018) Mapping the virome in wild-caught Aedes aegypti from Cairns and Bangkok. *Sci. Rep.*, 10.1038/s41598-018-22945-y.
23. Belda,E., Nanfack-Minkeu,F., Eiglmeier,K., Carissimo,G., Holm,I., Diallo,M., Diallo,D., Vantaux,A., Kim,S., Sharakhov,I. V., *et al.* (2019) De novo profiling of RNA viruses in Anopheles malaria vector mosquitoes from forest ecological zones in Senegal and Cambodia. *BMC Genomics*, 10.1186/s12864-019-6034-1.
24. Chandler,J.A., Liu,R.M. and Bennett,S.N. (2015) RNA Shotgun Metagenomic Sequencing of Northern California (USA) Mosquitoes Uncovers Viruses, Bacteria, and Fungi. *Front. Microbiol.*, 10.3389/fmicb.2015.00185.
25. Thongsripong,P., Chandler,J.A., Kittayapong,P., Wilcox,B.A., Kapan,D.D. and Bennett,S.N. (2021) Metagenomic shotgun sequencing reveals host species as an important driver of virome composition in mosquitoes. *Sci. Rep.*, 10.1038/s41598-021-87122-0.
26. Morlan,J.D., Qu,K. and Sinicropi,D. V. (2012) Selective depletion of rRNA enables whole transcriptome profiling of archival fixed tissue. *PLoS One*, 10.1371/journal.pone.0042882.
27. Diallo,D., Fall,G., Diagne,C.T., Gaye,A., Ba,Y., Dia,I., Faye,O. and Diallo,M. (2020) Concurrent amplification of Zika, chikungunya, and yellow fever virus in a sylvatic focus of arboviruses in Southeastern Senegal, 2015. *BMC Microbiol.*, 10.1186/s12866-020-01866-9.
28. Vasconcelos,P.F.C., Costa,Z.G., Travassos da Rosa,E.S., Luna,E., Rodrigues,S.G., Barros,V.L.R.S., Dias,J.P., Monteiro,H.A.O., Oliva,O.F.P., Vasconcelos,H.B., *et al.* (2001) Epidemic of jungle yellow fever in Brazil, 2000: Implications of climatic alterations in disease spread. *J. Med. Virol.*, 10.1002/jmv.2078.abs.
29. Cardoso,J. da C., de Almeida,M.A.B., dos Santos,E., da Fonseca,D.F., Sallum,M.A.M., Noll,C.A., Monteiro,H.A. d. O., Cruz,A.C.R., Carvalho,V.L., Pinto,E. V., *et al.* (2010) Yellow fever virus in Haemagogus leucocelaenus and Aedes serratus mosquitoes, Southern Brazil, 2008. *Emerg. Infect. Dis.*, 10.3201/eid1612.100608.
30. Gaillet,M., Pichard,C., Restrepo,J., Lavergne,A., Perez,L., Enfissi,A., Abboud,P., Lambert,Y., Ma,L., Monot,M., *et al.* (2021) Outbreak of Oropouche Virus in French Guiana. *Emerg. Infect. Dis.*, 10.3201/eid2710.204760.
31. Kraemer,M.U.G., Reiner,R.C., Brady,O.J., Messina,J.P., Gilbert,M., Pigott,D.M., Yi,D., Johnson,K., Earl,L., Marczak,L.B., *et al.* (2019) Past and future spread of the arbovirus vectors Aedes aegypti and Aedes albopictus. *Nat. Microbiol.*, 10.1038/s41564-019-0376-y.
32. Auerswald,H., Maquart,P.O., Chevalier,V. and Boyer,S. (2021) Mosquito vector competence for

japanese encephalitis virus. *Viruses*, 10.3390/v13061154.
33. Mukwaya,L.G., Kayondo,J.K., Crabtree,M.B., Savage,H.M., Biggerstaff,B.J. and Miller,B.R. (2000) Genetic differentiation in the yellow fever virus vector, Aedes simpsoni complex, in Africa: Sequence variation in the ribosomal DNA internal transcribed spacers of anthropophilic and non-anthropophilic populations. *Insect Mol. Biol.*, 10.1046/j.1365-2583.2000.00161.x.
34. Mwangangi,J.M., Muturi,E.J., Muriu,S.M., Nzovu,J., Midega,J.T. and Mbogo,C. (2013) The role of Anopheles arabiensis and Anopheles coustani in indoor and outdoor malaria transmission in Taveta District, Kenya. *Parasites and Vectors*, 10.1186/1756-3305-6-114.
35. Nepomichene,T.N.J.J., Raharimalala,F.N., Andriamandimby,S.F., Ravalohery,J.P., Failloux,A.B., Heraud,J.M. and Boyer,S. (2018) Vector competence of Culex antennatus and Anopheles coustani mosquitoes for Rift Valley fever virus in Madagascar. *Med. Vet. Entomol.*, 10.1111/mve.12291.
36. Ratovonjato,J., Olive,M.M., Tantely,L.M., Andrianaivolambo,L., Tata,E., Razainirina,J., Jeanmaire,E., Reynes,J.M. and Elissa,N. (2011) Detection, isolation, and genetic characterization of Rift Valley fever virus from anopheles (Anopheles) coustani, anopheles (Anopheles) squamosus, and culex (Culex) antennatus of the haute matsiatra region, Madagascar. *Vector-Borne Zoonotic Dis.*, 10.1089/vbz.2010.0031.
37. Lutomiah,J., Bast,J., Clark,J., Richardson,J., Yalwala,S., Oullo,D., Mutisya,J., Mulwa,F., Musila,L., Khamadi,S., *et al.* (2013) Abundance, diversity, and distribution of mosquito vectors in selected ecological regions of Kenya: public health implications. *J. Vector Ecol.*, **38**, 134–142.
38. Brault,A.C., Foy,B.D., Myles,K.M., Kelly,C.L.H., Higgs,S., Weaver,S.C., Olson,K.E., Miller,B.R. and Powers,A.M. (2004) Infection patterns of o'nyong nyong virus in the malaria-transmitting mosquito, Anopheles gambiae. *Insect Mol. Biol.*, 10.1111/j.0962-1075.2004.00521.x.
39. Stevenson,J.C., Simubali,L., Mbambara,S., Musonda,M., Mweetwa,S., Mudenda,T., Pringle,J.C., Jones,C.M. and Norris,D.E. (2016) Detection of plasmodium falciparum infection in anopheles squamosus (diptera: Culicidae) in an area targeted for malaria elimination, Southern Zambia. *J. Med. Entomol.*, 10.1093/jme/tjw091.
40. Nikolay,B., Diallo,M., Boye,C.S.B. and Sall,A.A. (2011) Usutu virus in Africa. *Vector-Borne Zoonotic Dis.*, 10.1089/vbz.2011.0631.
41. Kim,H., Cha,G.W., Jeong,Y.E., Lee,W.G., Chang,K.S., Roh,J.Y., Yang,S.C., Park,M.Y., Park,C. and Shin,E.H. (2015) Detection of Japanese encephalitis virus genotype V in Culex orientalis and Culex pipiens (Diptera: Culicidae) in Korea. *PLoS One*, 10.1371/journal.pone.0116547.
42. Vázquez González,A., Ruiz,S., Herrero,L., Moreno,J., Molero,F., Magallanes,A., Sánchez-Seco,M.P., Figuerola,J. and Tenorio,A. (2011) West Nile and Usutu viruses in mosquitoes in Spain, 2008-2009. *Am. J. Trop. Med. Hyg.*, 10.4269/ajtmh.2011.11-0042.
43. Maquart,P.O., Sokha,C. and Boyer,S. (2021) Mosquito diversity (Diptera: Culicidae) and medical importance, in a bird sanctuary inside the flooded forest of Prek Toal, Cambodia. *J. Asia. Pac. Entomol.*, 10.1016/j.aspen.2021.08.001.
44. Ndiaye,E.H., Fall,G., Gaye,A., Bob,N.S., Talla,C., Diagne,C.T., Diallo,D., Ba,Y., Dia,I., Kohl,A., *et al.* (2016) Vector competence of Aedes vexans (Meigen), Culex poicilipes (Theobald) and Cx. quinquefasciatus Say from Senegal for West and East African lineages of Rift Valley fever virus. *Parasites and Vectors*, 10.1186/s13071-016-1383-y.
45. Bhattacharya,S., Basu,P. and Sajal Bhattacharya,C. (2016) The Southern House Mosquito, Culex quinquefasciatus: profile of a smart vector. *J. Entomol. Zool. Stud. JEZS*.
46. Turell,M.J., O'Guinn,M.L., Dohm,D., Zyzak,M., Watts,D., Fernandez,R., Calampa,C., Klein,T.A. and Jones,J.W. (2008) Susceptibility of peruvian mosquitoes to eastern equine encephalitis virus. *J. Med. Entomol.*, 10.1603/0022-2585(2008)45[720:SOPMTE]2.0.CO;2.
47. Barrio-Nuevo,K.M., Cunha,M.S., Luchs,A., Fernandes,A., Rocco,I.M., Mucci,L.F., DE Souza,R.P., Medeiros-Sousa,A.R., Ceretti-Junior,W. and Marrelli,M.T. (2020) Detection of Zika and dengue viruses in wildcaught mosquitoes collected during field surveillance in an environmental protection area in São Paulo, Brazil. *PLoS One*, 10.1371/journal.pone.0227239.
48. Turell,M.J. (1999) Vector competence of three Venezuelan mosquitoes (Diptera: Culicidae) for an epizootic IC strain of Venezuelan equine encephalitis virus. *J. Med. Entomol.*, 10.1093/jmedent/36.4.407.
49. Hoyos-López,R., Soto,S.U., Rúa-Uribe,G. and Gallego-Gómez,J.C. (2015) Molecular identification of saint louis encephalitis virus genotype IV in Colombia. *Mem. Inst. Oswaldo Cruz*, 10.1590/0074-02760280040.
50. Arunachalam,N., Philip Samuel,P., Hiriyan,J., Thenmozhi,V. and Gajanana,A. (2004) Japanese encephalitis in Kerala, South India: Can Mansonia (Diptera: Culicidae) play a supplemental role in transmission? *J. Med. Entomol.*, 10.1603/0022-2585-41.3.456.

51. Ughasi,J., Bekard,H.E., Coulibaly,M., Adabie-Gomez,D., Gyapong,J., Appawu,M., Wilson,M.D. and Boakye,D.A. (2012) Mansonia africana and Mansonia uniformis are Vectors in the transmission of Wuchereria bancrofti lymphatic filariasis in Ghana. *Parasites and Vectors*, **5**, 1–5.
52. Mitchell,C.J., Forattini,O.P. and Miller,B.R. (1986) Vector competence experiments with Rocio virus and three mosquito species from the epidemic zone in Brazil. *Rev. Saude Publica*, 10.1590/s0034-89101986000300001.
53. Harbach,R.E. (2007) The Culicidae (Diptera): A review of taxonomy, classification and phylogeny. *Zootaxa*, 10.11646/zootaxa.1668.1.28.
54. Chan,A., Chiang,L.P., Hapuarachchi,H.C., Tan,C.H., Pang,S.C., Lee,R., Lee,K.S., Ng,L.C. and Lam-Phua,S.G. (2014) DNA barcoding: Complementing morphological identification of mosquito species in Singapore. *Parasites and Vectors*, 10.1186/s13071-014-0569-4.
55. Talaga,S., Leroy,C., Guidez,A., Dusfour,I., Girod,R., Dejean,A. and Murienne,J. (2017) DNA reference libraries of French Guianese mosquitoes for barcoding and metabarcoding. *PLoS One*, 10.1371/journal.pone.0176993.
56. Nchoutpouen,E., Talipouo,A., Djiappi-Tchamen,B., Djamouko-Djonkam,L., Kopya,E., Ngadjeu,C.S., Doumbe-Belisse,P., Awono-Ambene,P., Kekeunou,S., Wondji,C.S., *et al.* (2019) Culex species diversity, susceptibility to insecticides and role as potential vector of Lymphatic filariasis in the city of Yaoundé, Cameroon. *PLoS Negl. Trop. Dis.*, 10.1371/journal.pntd.0007229.
57. Farajollahi,A., Fonseca,D.M., Kramer,L.D. and Marm Kilpatrick,A. (2011) 'Bird biting' mosquitoes and human disease: A review of the role of Culex pipiens complex mosquitoes in epidemiology. *Infect. Genet. Evol.*, 10.1016/j.meegid.2011.08.013.
58. Zittra,C., Flechl,E., Kothmayer,M., Vitecek,S., Rossiter,H., Zechmeister,T. and Fuehrer,H.P. (2016) Ecological characterization and molecular differentiation of Culex pipiens complex taxa and Culex torrentium in eastern Austria. *Parasites and Vectors*, 10.1186/s13071-016-1495-4.
59. Aspen,S. and Savage,H.M. (2003) Polymerase chain reaction assay identifies North American members of the Culex pipiens complex based on nucleotide sequence differences in the acetylcholinesterase gene Ace.2. *J. Am. Mosq. Control Assoc.*
60. Cornel,A.J., Mcabee,R.D., Rasgon,J., Stanich,M.A., Scott,T.W. and Coetzee,M. (2003) Differences in Extent of Genetic Introgression between Sympatric Culex pipiens and Culex quinquefasciatus (Diptera: Culicidae) in California and South Africa. *J. Med. Entomol.*, 10.1603/0022-2585-40.1.36.
61. Beebe,N.W. (2018) DNA barcoding mosquitoes: Advice for potential prospectors. *Parasitology*, 10.1017/S0031182018000343.

TABLE AND FIGURES LEGENDS

Table 1. List of mosquito species represented in this study and their vector status. Origin countries are listed as their ISO alpha-2 codes: Central African Republic, CF; Cambodia, KH; Madagascar, MG; French Guiana, GF.

Figure 1. (**A**) Proportion of rRNA reads found in mosquito specimen pools depleted by probe hybridization followed by RNase H digestion. Probes were antisense to *Ae. aegypti* rRNA sequences. (**B**) Read vs. score ratio plot of "Ae simpsoni CS S27". Green line indicates 0.8 cut-off where only reads above this threshold are used in rRNA assembly. Dark blue dashed line denotes reads with hits only in the Insecta library. Light blue dashed-line denotes reads with higher-scoring hits in the Insecta library while yellow dotted-line denotes reads with lower scoring hits In the Insecta library, both relative to the Non-Insecta library. Orange dotted line denotes reads with only hits in the Non-Insecta library.

Figure 2. Phylogenetic tree based on 28S sequences generated from this study and from NCBI databases combined (3900 bp) as inferred using maximum-likelihood method and constructed to scale in MEGA X (18). Values at each node indicate bootstrap support from 500 replications. For

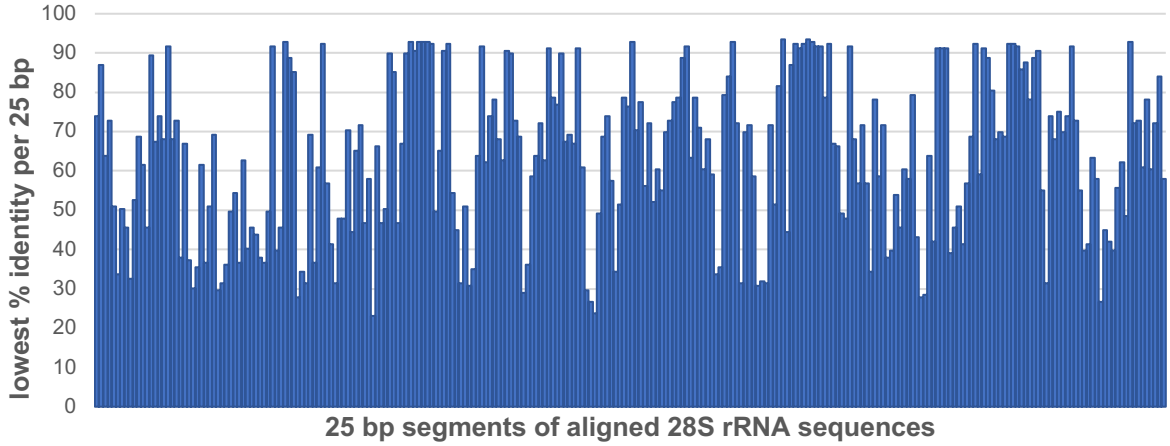
sequences from this study, each specimen label contains information on its taxonomy, origin (as indicated in 2-letter country codes), and specimen ID. Labels in bold indicate sequences derived from NCBI. Label colors indicate genera: *Culex* in coral, *Anopheles* in purple, *Aedes* in dark blue, *Mansonia* in dark green, *Culiseta* in maroon, *Limatus* in light green, *Coquillettidia* in light blue, *Psorophora* in yellow, *Mimomyia* in teal, *Uranotaenia* in pink and *Eretmapodites* in brown. Scale bar at 0.05 is shown.

Figure 3. Phylogenetic tree based on 28S sequences generated from this study alone (3900 bp) as inferred using maximum-likelihood method and constructed to scale in MEGA X (18). Values at each node indicate bootstrap support from 500 replications. Each specimen label contains information on its taxonomy, origin (as indicated in 2-letter country codes), and specimen ID. Label colors indicate genera: *Culex* in coral, *Anopheles* in purple, *Aedes* in dark blue, *Mansonia* in dark green, *Culiseta* in maroon, *Limatus* in light green, *Coquillettidia* in light blue, *Psorophora* in yellow, *Mimomyia* in teal, *Uranotaenia* in pink and *Eretmapodites* in brown. Scale bar at 0.05 is shown.

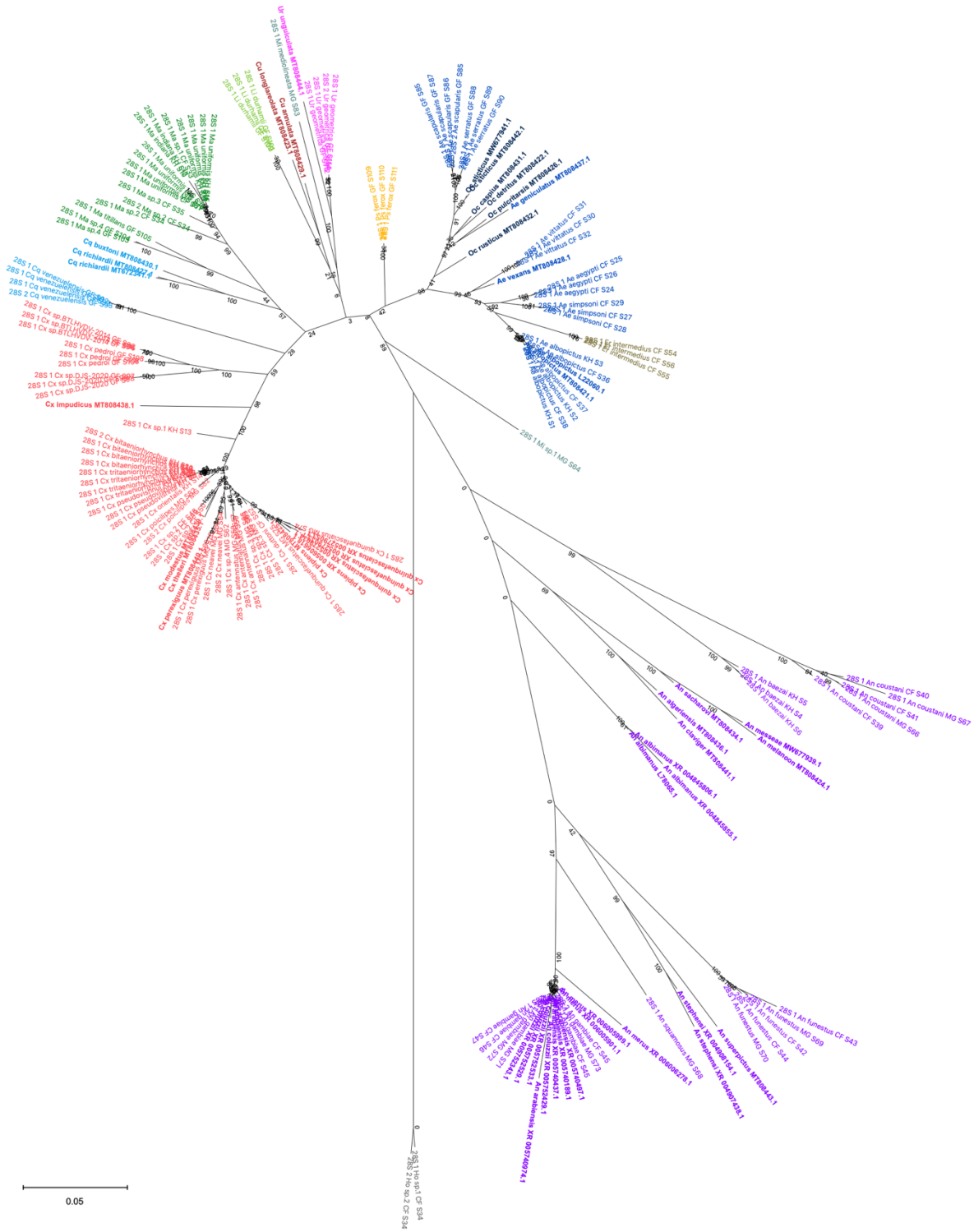
Figure 4. Phylogenetic tree based on 18S sequences (1900 bp) as inferred using maximum-likelihood method and constructed to scale in MEGA X (18). Values at each node indicate bootstrap support from 500 replications. Each specimen label contains information on its taxonomy, origin (as indicated in 2-letter country codes), and specimen ID. Label colors indicate genera: *Culex* in coral, *Anopheles* in purple, *Aedes* in dark blue, *Mansonia* in dark green, *Limatus* in light green, *Coquillettidia* in light blue, *Psorophora* in yellow, *Mimomyia* in teal, *Uranotaenia* in pink and *Eretmapodites* in brown. Scale bar at 0.05 is shown.

Figure 5. Phylogenetic tree based on COI sequences (620–699 bp) as inferred using maximum-likelihood method and constructed to scale in MEGA X (18). Values at each node indicate bootstrap support from 500 replications. Each specimen label contains information on its taxonomy, origin (as indicated in 2-letter country codes), and specimen ID. Label colors indicate genera: *Culex* in coral, *Anopheles* in purple, *Aedes* in dark blue, *Mansonia* in dark green, *Limatus* in light green, *Coquillettidia* in light blue, *Psorophora* in yellow, *Mimomyia* in teal, *Uranotaenia* in pink and *Eretmapodites* in brown. Scale bar at 0.05 is shown.

SUPPLEMENTARY FIGURES

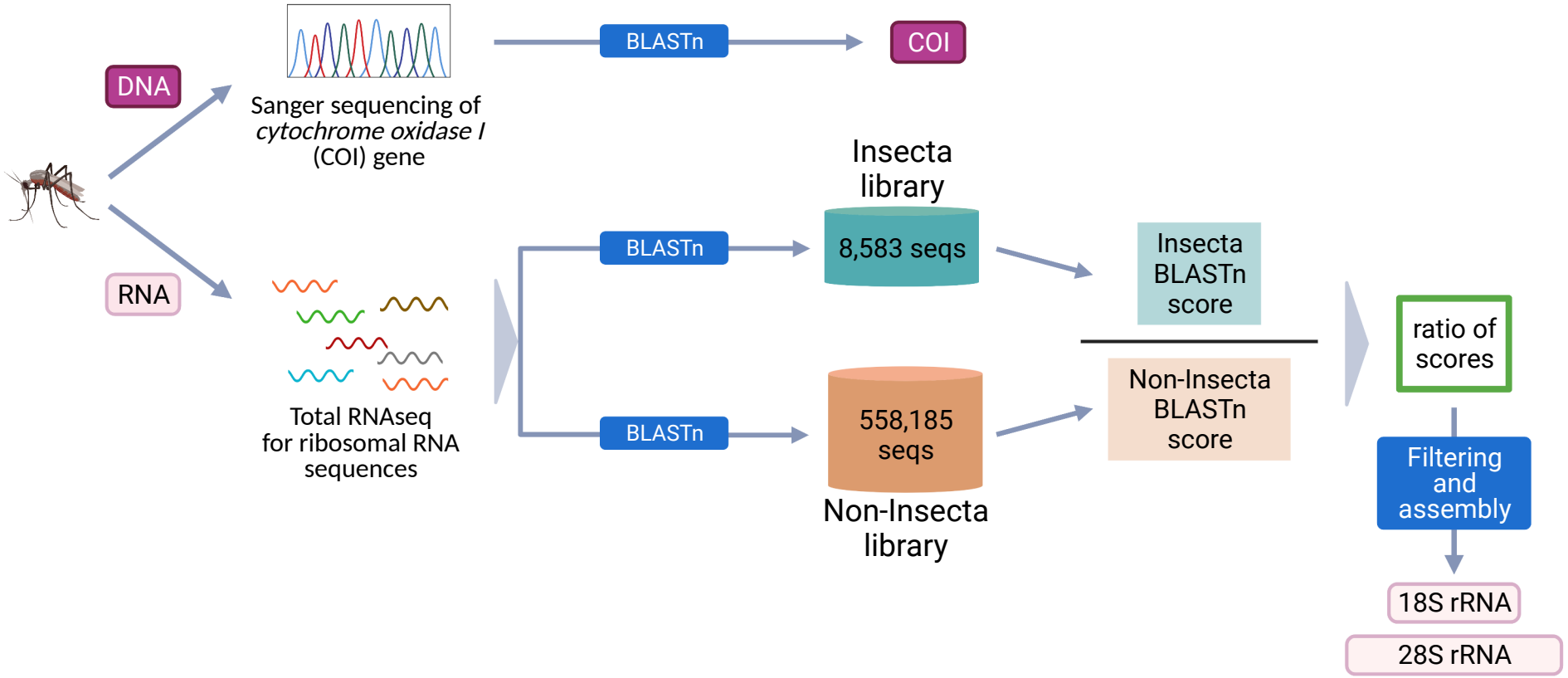


Supplementary Figure 1. Sequence conservation among 169 28S rRNA sequences from this study and from NCBI databases combined. Multiple sequence alignment was performed on 28S rRNA sequences, 3900 bp in length. Each bar depicts the lowest percentage identity found in every 25-bp segment including gaps in alignment.



Supplementary Figure 2. Phylogenetic tree based on 28S sequences generated from this study and from NCBI databases combined (3900 bp) as inferred using maximum-likelihood method and constructed to scale in MEGA X in radial format (18). Values at each node indicate bootstrap support from 500 replications. For sequences from this study, each specimen label contains information on its taxonomy, origin (as indicated in 2-letter country codes), and specimen ID. Labels in bold indicate sequences derived from NCBI. Label colours indicate genera: *Culex* in coral, *Anopheles* in purple, *Aedes* in dark blue, *Mansonia* in dark green, *Culiseta* in maroon, *Limatus* in light green, *Coquillettidia*

in light blue, *Psorophora* in yellow, *Mimomyia* in teal, *Uranotaenia* in pink and *Eretmapodites* in brown. Scale bar at 0.05 is shown.


DNA
RNA
Sanger sequencing of
cytochrome oxidase I
(COI) gene
BLASTn
COI
Total RNAseq
for ribosomal RNA
sequences
BLASTn
BLASTn
Insecta
library
8,583 seqs
558,185
seqs
Non-Insecta
library
Insecta
BLASTn
score
Non-Insecta
BLASTn
score
ratio of
scores
Filtering
and
assembly
18S rRNA
28S rRNA