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, Institut Pasteur, Université Paris Cité, CNRS UMR3569, Paris F-75015, 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.

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

Email addresses:

Cassandra Koh, cassandra.koh@pasteur.fr; Lionel Frangeul, lionel.frangeul@pasteur.fr; Hervé Blanc, herve.blanc@pasteur.fr; Carine Ngoagouni, carine.ngoagouni@pasteur-bangui.cf; Sébastien Boyer, sboyer@pasteur-kh.org; Philippe Dussart, pdussart@pasteur.mg; Nina Grau, ngrau@pasteur.mg; Romain Girod, rgirod@pasteur.mg; Jean-Bernard Duchemin, jbduchemin@pasteur-cayenne.fr; Maria-Carla Saleh, carla.saleh@pasteur.fr.

ABSTRACT

Background: RNA-seq metagenomics on mosquitoes for surveillance and microbiome or pathogen discovery allows us to understand the disease ecology of arboviruses. A major hurdle in these studies is the depletion of overabundant ribosomal RNA (rRNA), commonly achieved using oligo-based protocols. The lack of publicly available complete reference rRNA sequences for many mosquito vector species narrows the range of such studies, causing a knowledge bias in mosquito vector biology. Here we describe a strategy to assemble full-length 28S and 18S rRNA sequences of 29 sylvatic and peri-urban mosquito species sampled from Cambodia, the Central African Republic, Madagascar, and French Guiana.

Results: Our score-based strategy successfully parses rRNA reads into insect and non-insect sources, leading to the assembly of complete rRNA sequences for all specimens in the study. We then evaluated the functionality of rRNA sequences as barcodes for taxonomy and phylogeny relative to the mitochondrial *cytochrome* c *oxidase I* (COI) gene marker system. rRNA- and COI-based phylogenetic inferences share little congruity. However, the former allowed for molecular species identification when COI sequences were ambiguous or unavailable and revealed better supported intergeneric evolutionary histories concordant with contemporary mosquito systematics.

Conclusions: The presented assembly strategy and the expansion of the rRNA reference library in public databases by 234 novel complete 28S and 18S rRNA sequences provide a new tool in the form of an RNA marker system to improve mosquito RNA-seq metagenomics. More holistic insights on mosquito vector ecology will benefit the design of public health measures against arboviral diseases.

Keywords: ribosomal RNA, rRNA, depletion, RNA-seq, metagenomics, mosquito, phylogeny, taxonomy

BACKGROUND

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 emphasises the importance of considering the role of biotic and abiotic elements within the same ecosystem in contributing to zoonotic disease transmission (2). With next-generation sequencing technologies becoming more accessible, these studies are increasing in frequency, providing an unprecedented understanding of the interfaces among mosquitoes, their environment, and their animal and human hosts. Currently, there is a strong focus on only a handful of species from three genera of mosquitoes (*Aedes*, *Culex*, and *Anopheles*) due to their medical importance and anthropophilic behaviour despite evidence that other species are also competent for the pathogens in question. This narrows our knowledge of mosquito vector 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 (3).

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 four and two kilobases, respectively (4). They comprise at least 80% of the total cellular RNA population. In RNA-seq experiments, their depletion is a necessary step during library preparation where it is not possible to selectively enrich target signals (5). To achieve this, the most routinely used depletion protocols require knowledge of rRNA sequence of the species of interest. These protocols involve hybridizing antisense oligos (probes or primers) to rRNA molecules followed by digestion by ribonucleases (5,6) or removal by bead capture (7).

For well-studied mosquito species, reference rRNA sequences are readily available on public sequence databases such as GenBank or SILVA. As it is conventionally accepted that rDNA coding regions are highly conserved, it may seem conceivable to use oligo-based depletion protocols designed for one mosquito species on another. However, we found that within the family *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 marker for molecular taxonomy and forms the basis of the Barcode of Life Data System (BOLD) (8,9). The lack of reliable rRNA depletion methods could deter mosquito metagenomic studies from expanding their sampling diversity. The inclusion of lesser studied yet ecologically relevant species is imperative.

To address this, we sought to determine the 28S and 18S rRNA sequences of a diverse set of sylvatic and peri-urban mosquito species across Cambodia, the Central African Republic, Madagascar, and French Guiana. We employed a unique score-based read filtration strategy to remove interfering non-mosquito rRNA reads to ensure accurate *de novo* assembly and generated 122 complete 28S and 114 complete 18S sequences from 29 mosquito species. This strategy would facilitate the assembly of more rRNA sequences to expand the rRNA reference library. In parallel, we obtained COI sequences to confirm morphology-based species identification and to compare phylogenetic relationships inferred from the DNA and RNA markers, leading us to propose the use of 28S and 18S rRNA sequences as “rRNA barcodes”. Our sequence dataset enables rRNA-based streamlined molecular species identification during RNA-seq and allows for the design of species-specific oligos for cost-effective rRNA depletion for a broader range of mosquito species.

RESULTS

Poor rRNA depletion using non-specific depletion methods

During library preparations of mosquito samples for RNA-seq, routinely used methods for depleting rRNA are commercial kits optimised for human or mice samples (10–15) or probe hybridisation 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, in theory, be able to produce acceptable rRNA depletion efficiencies assuming that rRNA sequences have high degrees of homology across species. 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.* (17) produced poor depletion in *Ae. albopictus*, and worse still in Culicine and Anopheline species (Figure 1A). Additionally, the lack of reference rRNA sequences

compromises the clean-up of remaining rRNA reads from sequencing data, as reads belonging to more divergent regions do not map to a reference sequence from a different species. To solve this and to enable RNA-seq metagenomics on a wider range of mosquito species, we performed RNA-seq to obtain reference rRNA sequences for 29 mosquito species across nine genera from Cambodia, the Central African Republic, Madagascar, and French Guiana. Most of these species are associated with vector activity for various pathogens in their respective ecologies (Table 1).

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

Mosquito taxonomy*	Origin**	Collection site (ecosystem type)	Vector for***	Reference
Aedes (Fredardsius) vittatus	CF	rural (village)	DENV, ZIKV, CHIKV, YFV	(18)
Aedes (Ochlerotatus) scapularis	GF	rural (village)	YFV	(19)
Aedes (Ochlerotatus) serratus	GF	rural (village)	YFV, OROV	(20,21)
Aedes (Stegomyia) aegypti	CF	urban	DENV, ZIKV, CHIKV, YFV	(22)
Aedes (Stegomyia) albopictus	CF, KH	rural (village, nature reserve)	DENV, ZIKV, CHIKV, YFV, JEV	(22,23)
Aedes (Stegomyia) simpsoni	CF	rural (village)	YFV	(24)
Anopheles (Anopheles) baezai	KH	rural (nature reserve)	unreported	–
Anopheles (Anopheles) coustani	MG, CF	rural (village)	RVFV, malaria	(25–27)
Anopheles (Cellia) funestus	MG, CF	rural (village)	ONNV, malaria	(28)
Anopheles (Cellia) gambiae	MG, CF	rural (village)	ONNV, malaria	(29)
Anopheles (Cellia) squamosus	MG	rural (village)	RVFV, malaria	(27,30)
Coquillettidia (Rhynchotaenia) venezuelensis	GF	rural (village)	OROV	(21)
Culex (Culex) antennatus	MG	rural (village)	RVFV	(26,27)
Culex (Culex) duttoni	CF	rural (village)	unreported	–
Culex (Culex) neavei	MG	rural (village)	USUV	(31)
Culex (Culex) orientalis	KH	rural (nature reserve)	JEV	(32)
Culex (Culex) perexiguus	MG	rural (village)	WNV	(33)

Culex (Culex) pseudovishnui	KH	rural (nature reserve)	JEV	(23,34)
Culex (Culex) quinquefasciatus	MG, CF, KH	rural (village, nature reserve)	ZIKV, JEV, WNV, DENV, SLEV, RVFV, *Wuchereria bancrofti*	(34–36)
Culex (Culex) tritaeniorhynchus	MG, KH	rural (village, nature reserve)	JEV, WNV, RVFV	(23,34)
Culex (Melanoconion)spissipes	GF	rural (village)	VEEV	(37)
Culex (Melanoconion) portesi	GF	rural (village)	VEEV, TONV	(37,38)
Culex (Melanoconion) pedroi	GF	rural (village)	EEEV, VEEV, MADV	(38,39)
Culex (Oculeomyia) bitaeniorhynchus	MG, KH	rural (village, nature reserve)	JEV	(23,34)
Culex (Oculeomyia) poicilipes	MG	rural (village)	RVFV	(35)
Eretmapodites intermedius	CF	rural (village)	unreported	–
Limatus durhamii	GF	rural (village)	ZIKV	(40)
Mansonia (Mansonia) titillans	GF	rural (village)	VEEV, SLEV	(41,42)
Mansonia (Mansonioides) indiana	KH	rural (nature reserve)	JEV	(43)
Mansonia (Mansonioides) uniformis	MG, CF, KH	rural (village, nature reserve)	WNV, RVFV, *Wuchereria bancrofti*	(28,34,44)
Mimomyia (Etorleptiomyia) mediolineata	MG	rural (village)	unreported	–
Psorophora (Janthinosoma) ferox	GF	rural (village)	ROCV	(45)
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; Oropouche virus, OROV; Japanese encephalitis virus, JEV; Rift Valley Fever virus, RVFV; O'Nyong Nyong virus, ONNV; Usutu virus, USUV; West Nile virus, WNV; Saint Louis encephalitis virus, SLEV; Venezuelan equine encephalitis virus, VEEV; Tonate virus, TONV; Eastern equine encephalitis virus, EEEV; Madariaga virus, MADV; Rocio virus, ROCV.

rRNA reads filtering and sequence assembly

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

Our score-based filtration strategy, described in detail in Methods, allowed us to bioinformatically remove interfering rRNA reads and achieve successful de novo assembly of 28S and 18S rRNA sequences for all our specimens. Briefly, for each Illumina read, we computed a ratio of BLAST scores against an Insecta library over a Non-Insecta library. Reads were segregated into four categories: (i) reads that map only to the Insecta library, (ii) 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 (Figure 1B, Figure S1). By applying a conservative threshold of 0.8 to account for the non-exhaustiveness of reference libraries used, we filtered out reads that likely do not originate from mosquito rRNA. Notably, 15 of our specimens were engorged with vertebrate blood, a rich source of non-mosquito rRNA (Supplementary Table 1). The successful assembly of complete 28S and 18S rRNA sequences demonstrates that this strategy performs as expected even with high amounts of non-host rRNA reads. This is particularly important in studies on field-captured mosquitoes as females are often sampled already having imbibed a blood meal or captured using the human-landing catch technique.

We encountered challenges for three specimens morphologically identified as *Ma. africana* (Specimen ID 33-35). COI amplification by PCR did not produce any product, hence COI barcoding could not be used to confirm species identity. In addition, SPAdes was only able to assemble partial length contigs, despite the high number of reads with high scores against the Insecta library. Among other *Mansonia* specimens, the partial length contigs shared the highest similarity with contigs obtained from “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 sets of 28S and 18S rRNA sequences, one of which was similar to mosquito rRNA with low coverage and another with ten-fold higher coverage and 95% nucleotide sequence similarity to a *Horreolanus* species of water mite

known to parasitize mosquitoes. Our filtration strategy allowed us to obtain rRNA sequences for the mosquito as well as the unknown *Horreolanus* species. This shows that our strategy can be applied to metabarcoding studies where the input material comprises multiple insect species, provided that appropriate reference sequences of the target species or of a close relative are available.

Altogether, we were able to assemble 122 28S and 114 18S full-length mosquito rRNA sequences for 29 mosquito species sourced from four countries across three continents. These sequences represent, to our knowledge, the first records for seven mosquito genera: *Coquillettidia*, *Mansonia*, *Limatus*, *Mimomyia*, *Uranotaenia*, *Psorophora,* and *Eretmapodites*. For *Culex*, *Aedes* and *Anopheles* genera, where complete rRNA sequences were already available for a few species, this study provides the first rRNA records for 18 species. The GenBank accession numbers for these sequences and specimen information are listed in Table S1.

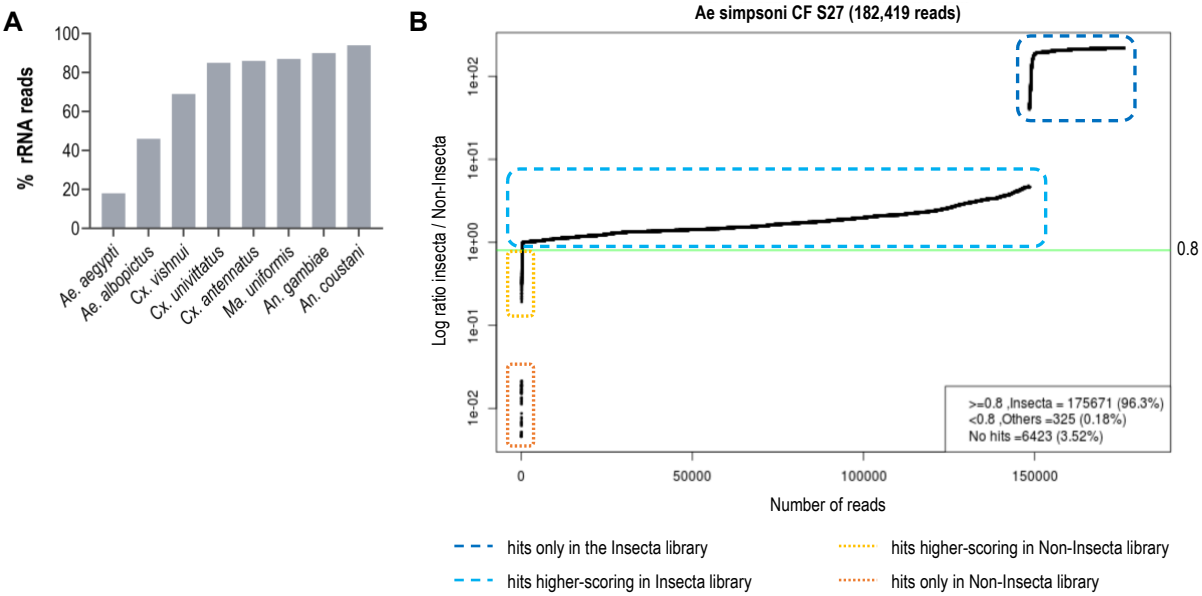
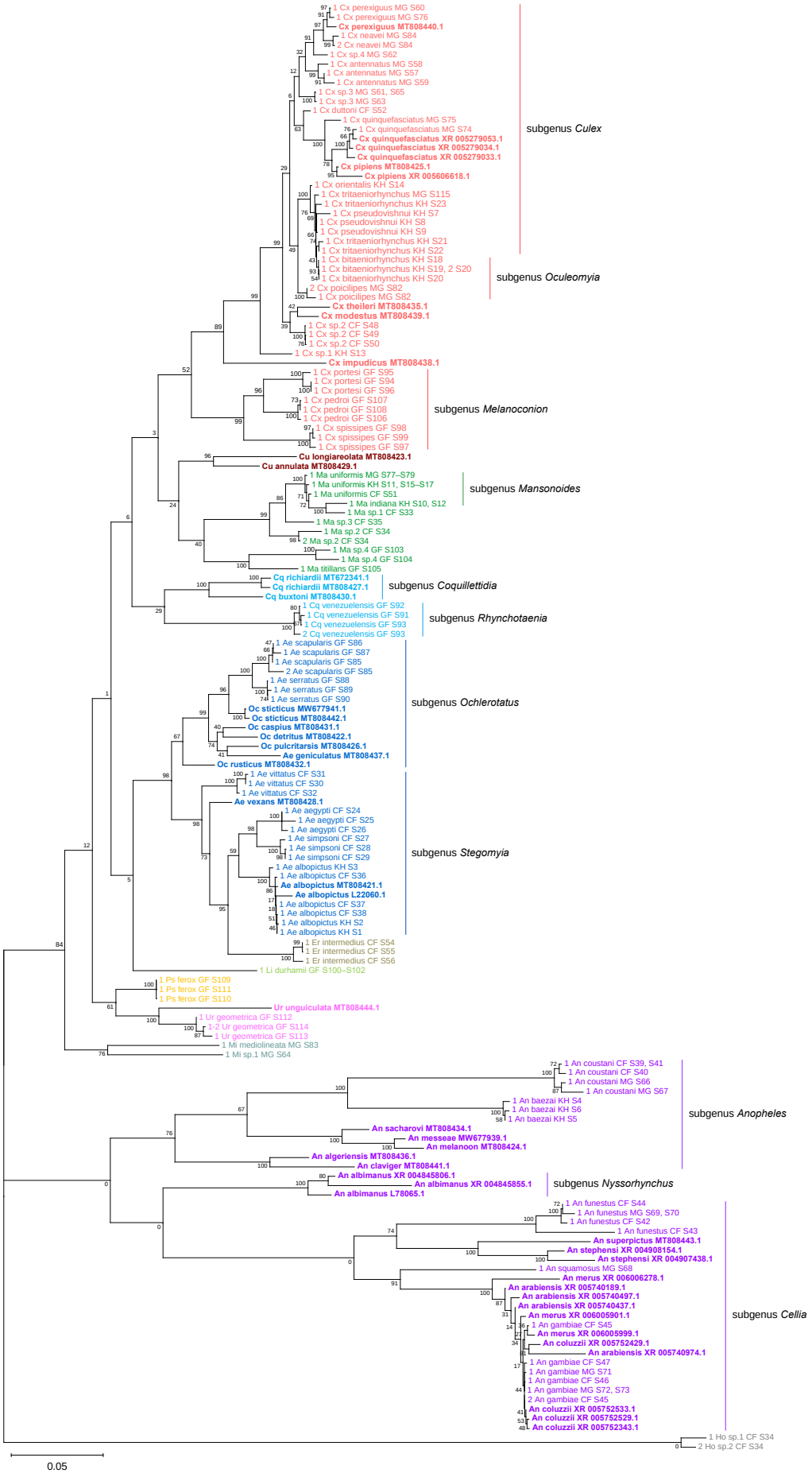


Figure 1. (**A**) Proportion of rRNA reads found in mosquito specimen pools of 5 individuals depleted by probe hybridisation followed by RNase H digestion. Probes were antisense to *Ae. aegypti* rRNA sequences. (**B**) Read vs. score ratio plot of a representative specimen "Ae simpsoni CS S27". Green line indicates 0.8 cut-off where only reads above this threshold are used in rRNA assembly.

Comparative phylogeny of novel rRNA sequences relative to existing records

To verify the assembly accuracy of our rRNA sequences, we constructed a comprehensive phylogenetic tree from the 28S rRNA sequences generated from our study alongside those publicly available from NCBI databases (Figure 2). We applied a search criterion for NCBI sequences with at

181 least 95% coverage of our sequence lengths (~4000 bp), aiming to have as many species as possible
182 represented. Although we rarely found NCBI entries for the same species represented in our study,
183 the resulting tree showed that our 28S sequences generally clustered according to their respective
184 species and subgenera, supported by moderate to good bootstrap values at terminal nodes. Species
185 taxa generally formed monophyletic clades, apart from *An. gambiae* and *Cx. quinquefasciatus*. *An.*
186 *gambiae* 28S rRNA sequences formed a clade with closely related sequences from *An. arabiensis*,
187 *An. merus*, and *An. coluzzii,* suggesting unusually high interspecies homology for Anophelines or
188 other members of subgenus *Cellia.* Meanwhile, *Cx. quinquefasciatus* 28S rRNA sequences formed a
189 taxon paraphyletic to sister species *Cx. pipiens*.



190

Figure 2. Phylogenetic tree based on 28S sequences generated from this study and from NCBI databases (3900 bp) as inferred using maximum-likelihood method and constructed to scale in MEGA X (46). 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 with accession numbers shown. 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.

28S sequence-based phylogenetic reconstructions (Figure 2, with NCBI sequences; Figure S3, this study only) showed marked incongruence to that of 18S sequences (Figure 3). Although all rRNA trees show clear segregation of genus *Anopheles* from tribes *Aedini* and *Culicini*, the phylogenetic relationships of other genera in this study relative to the greater three are highly variable and weakly supported, particularly in the 18S tree. The 18S tree also showed a number of taxonomic anomalies: (i) the lack of definitive clustering by species within the *Culex* subgenus (ii) the inability to differentiate between 18S sequences of *Cx*. *pseudovishnui* and *Cx*. *tritaeniorhynchus*; (iii) the placement of Ma sp. 3 CF S35 within a *Culex* clade; and (iv) the lack of a *Mimomyia* clade. The topology of the 18S tree seem to suggest higher sequence divergence between the two *Cx. quinquefasciatus* taxa and between the two *Mimomyia* taxa than in their 28S sequences. However, 28S and 18S rRNA sequences are encoded by linked loci in rDNA clusters and should not be analysed separately.

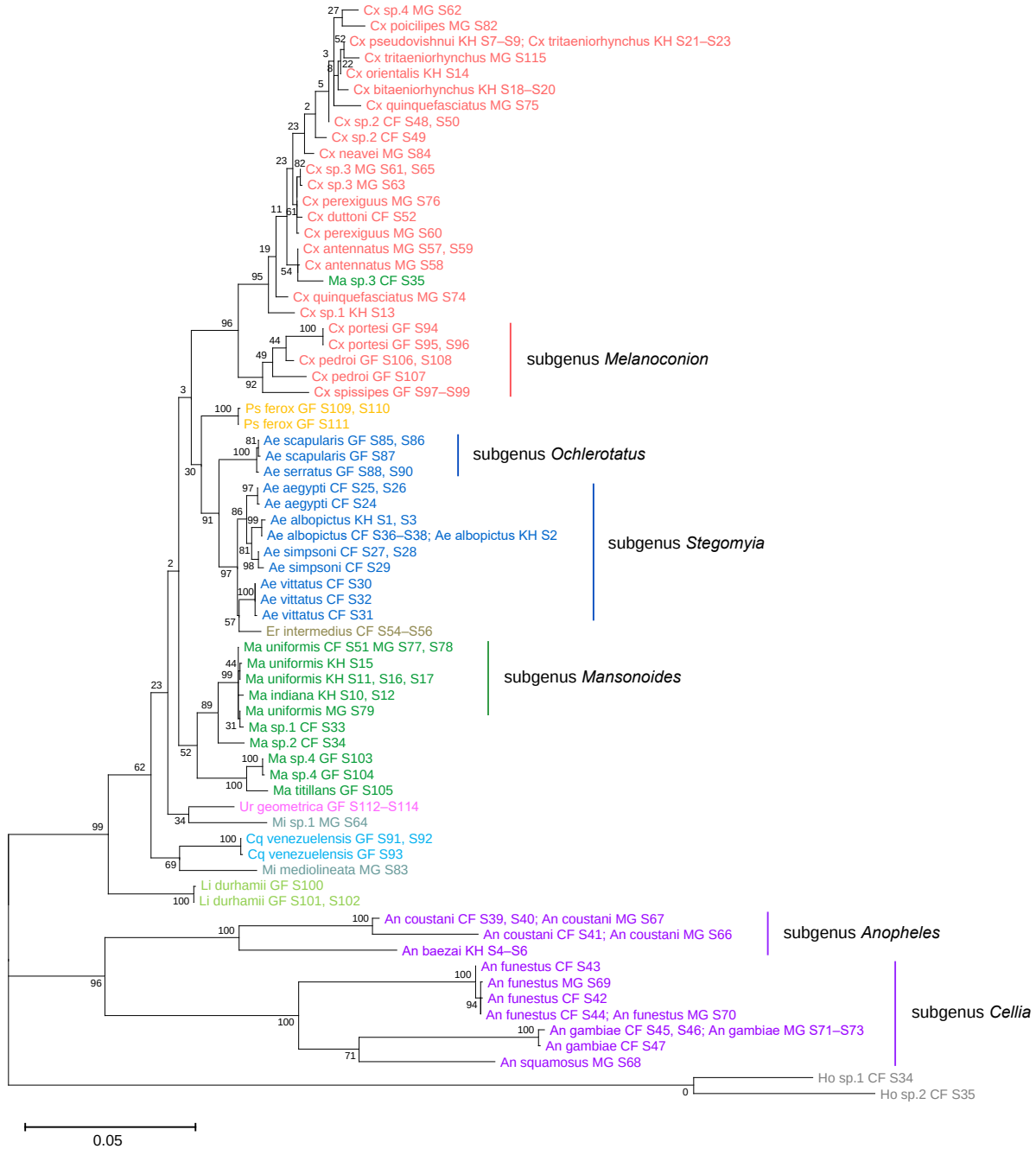


Figure 3. Phylogenetic tree based on 18S sequences (1900 bp) as inferred using maximum-likelihood method and constructed to scale in MEGA X (46). Values at each node indicate bootstrap support (%) from 500 replications. Each specimen label contains information on its taxonomy, origin (as indicated in 2-letter country codes), and specimen ID. Label colours 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.

Indeed, when concatenated 28S+18S rRNA sequences were generated from the same specimens (Figure 4), the phylogenetic tree resulting from these sequences more closely resembles the 28S tree (Figure 2) with regard to the basal position of the *Mimomyia* clade within the *Culicinae* subfamily with good bootstrap support in either tree (84% in 28S tree, 100% in concatenated 28S+18S tree). For internal nodes, bootstrap support values were higher in the concatenated tree compared to the 28S tree. Interestingly, the 28S+18S tree formed an *Aedini* tribe-clade encompassing taxa from genera *Psorophora, Aedes,* and *Eretmapodites*, possibly driven by the inclusion of 18S sequences. Concatenating the 28S and 18S sequences also resolved the anomalies found in the 18S tree and added clarity to the close relationship between *Culex* and *Mansonia* taxa. Of note, the *Culex* and *Mansonia* genera are no longer monophyletic in the concatenated 28S+18S tree. Genus *Culex* is paraphyletic with respect to subgenus *Mansonoides* of genus *Mansonia* (Figure 2). *Ma*. *titillans* and Ma sp. 4, which we suspect to be *Ma. pseudotitillans,* always formed a distinct branch in 28S or 18S phylogenies, thus possibly representing a clade of subgenus *Mansonia*.

The concatenated 28S+18S tree recapitulates what is classically known about the systematics of our specimens, namely (i) the early divergence of genus *Anopheles* from other *Culicidae* genera, (ii) the division of genus *Anopheles* into two subgenera, *Anopheles* and *Cellia,* (iii) the division of genus *Aedes* into subgenera *Stegomyia* and *Ochlerotatus*, (iv) the divergence of monophyletic subgenus *Melanoconion* within the *Culex* genus (47,48).

Figure 4. Phylogenetic tree based on concatenated 28S and 18S sequences generated from this study (3900+1900 bp) as inferred using maximum-likelihood method and constructed to scale in MEGA X (46). 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. 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.

rRNA as a molecular marker for taxonomy and phylogeny

We sequenced a 621 bp region of the COI gene not only to confirm morphological identification of our specimens but also to compare the functionality of rRNA and COI sequences as molecular markers for taxonomic and phylogenetic investigations. COI sequences were able to unequivocally determine the species in most specimens except for the following cases. *An. coustani* COI sequences from our study 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. In the case of *Ae. simpsoni*, three specimens were morphologically identified as *Ae. opok* although their COI sequences showed 97-100% similarity to that of *Ae. simpsoni.* As NCBI held 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 significant divergence, we concluded these specimens to be *Ae. simpsoni*. Ambiguous results were especially frequent among *Culex* specimens belonging to the *Cx. pipiens* or *Cx. vishnui* species groups, where the query sequence differed with either of the top two hits by a single nucleotide. For example, between *Cx. quinquefasciatus* and *Cx. pipiens* of the *Cx. pipiens* species group, and between *Cx. vishnui* and *Cx. tritaeniorhynchus* of the *Cx. vishnui* species group.

Among our three specimens of *Ma. titillans*, two appeared to belong to a single species that is different but closely related to *Ma. titillans*. We surmised that these specimens could instead be *Ma. pseudotitillans* based on morphological similarity but were not able to verify this by molecular means as no COI reference sequence is available for this species. These specimens are hence putatively labelled as “Ma sp.4”.

Phylogenetic reconstruction based on the COI sequences showed clustering of all species taxa into distinct clades, underlining the utility of the COI gene in molecular taxonomy (Figure 5). However, species delineation among members of *Culex* species groups were not as clear cut, although sister

species were correctly placed as sister taxa. This is comparable to the 28S+18S tree (Figure 4) and is indicative of lower intraspecies distances relative to interspecies distances.

To evaluate the utility of 28S and 18S rRNA sequences for taxonomy-based species identification, we used 28S+18S rRNA phylogenetic inference to discern the identity of six specimens for which COI barcoding could not be performed. These specimens include three unknown *Mansonia* species (Specimen ID 33–35), a *Ma. uniformis* (Specimen ID 51), an *An. gambiae* (Specimen ID 47), and a *Ur. geometrica* (Specimen ID 113). Their positions in the 28S rRNA tree relative to adjacent taxa confirms the morphological identification of all six specimens to the genus level and, for three of them, to the species level (Figure 4).

The phylogenetic relationships indicated by the COI tree compared to the 28S+18S tree present only a few points of congruence. COI-based phylogenetic inference indeed showed clustering of generic taxa into clades albeit with very weak bootstrap support, except for genera *Culex* and *Mansonia* (Figure 5). Contrary to the 28S+18S tree (Figure 4), *Culex* subgenus *Melanoconion* was depicted as a polyphyletic taxon with *Cx. spissipes* being a part of the greater *Culicini* clade with members from subgenera *Oculeomyia* and *Culex* while *Cx. pedroi* and *Cx. portesi* formed a distantly related clade. Among the *Mansonia* specimens, the two unknown Ma sp.4 specimens were not positioned as the nearest neighbours of *Ma. titillans* and instead appeared to have diverged earlier from most of the other taxa from the *Culicidae* family. Notably, the COI sequences of genus *Anopheles* is not basal to the other members of *Culicidae* and is instead shown to be sister to Culex COI sequences (8% bootstrap support). This is a direct contrast to what is suggested by the rRNA phylogenies (Figures 2–4), which suggests *Culex* rRNA sequences to be among the most recently diverged. Bootstrap values for the more internal nodes of the COI trees are remarkably low compared to those of rRNA-based trees.

In all rRNA trees, it is clear that the interspecific and intersubgeneric evolutionary distances within the genus *Anopheles* are high relative to any other genera, indicating a greater degree of divergence. This is evidenced by 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 3). 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.

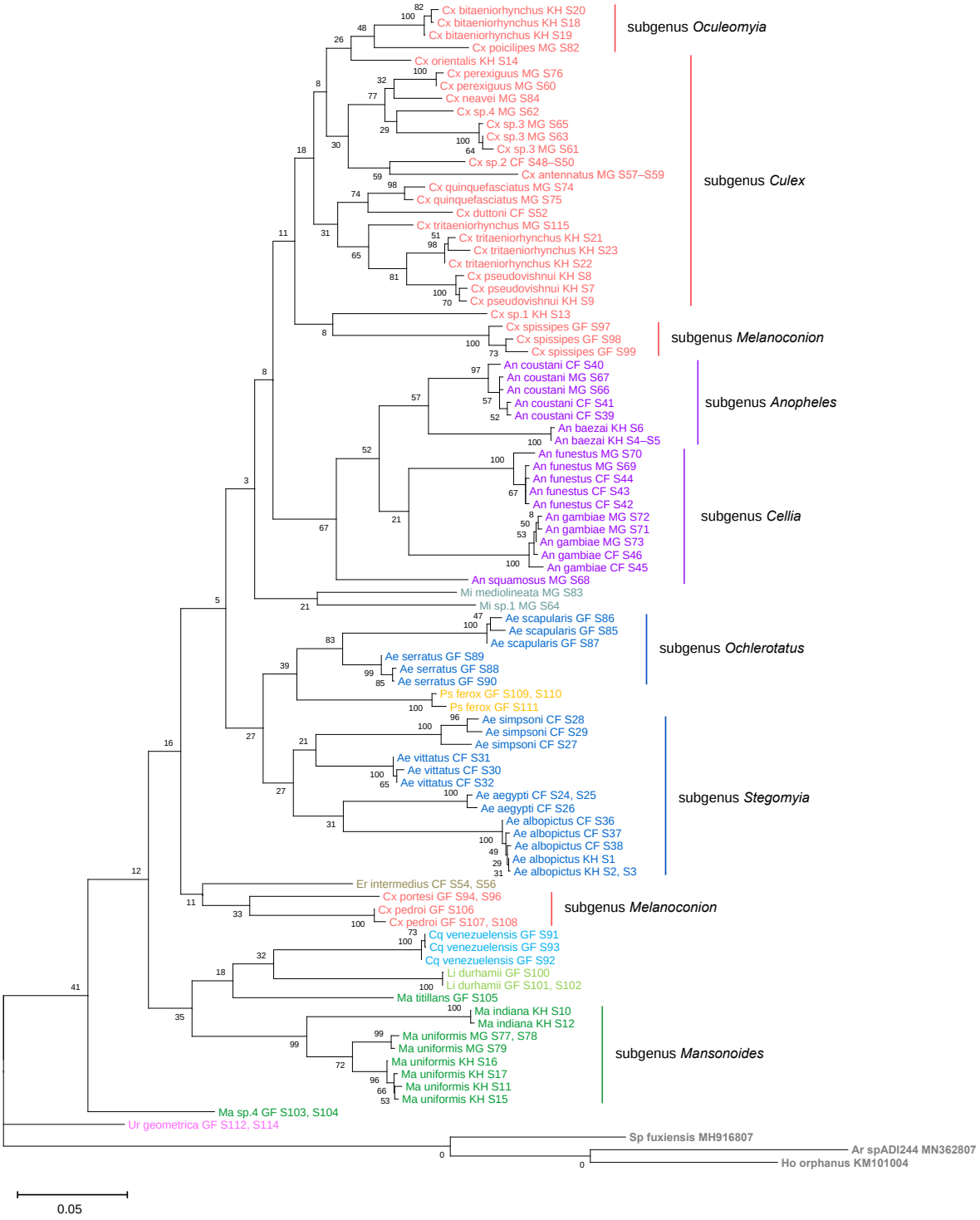


Figure 5. Phylogenetic tree based on COI sequences (621–699 bp) as inferred using maximum-likelihood method and constructed to scale in MEGA X (46). Values at each node indicate bootstrap support (%) from 500 replications. Each specimen label contains information on its taxonomy, origin (as indicated in 2-letter country codes), and specimen ID. Label colours 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.

On *Culex* species groups

Culex (subgenus *Culex*) specimens of this study comprise several closely related sister species belonging to the *Cx. vishnui* and *Cx. univittatus* species groups, which are notoriously difficult to differentiate based on morphology. Accordingly, in the 28S+18S rRNA (Figure 4) and COI (Figure 5) trees these species and their known sister species were clustered together within the *Culex* (subgenus *Culex*) clade: *Cx. tritaeniorhynchus* with *Cx. pseudovishnui* (*Cx. vishnui* species group); *Cx. perexiguus* with *Cx. neavei* (*Cx. univittatus* species group).

The use of COI barcoding to discern between members of the *Culex* species groups was limited. For example, for the two *Cx. quinquefasciatus* samples in our dataset (Specimen ID 74 and 75), BLAST analyses of their COI sequences revealed they are a single nucleotide away from *Cx. pipiens* or *Cx. quinquefasciatus* COI sequences (Table S2). In the 28S rRNA tree with NCBI sequences (Figure 2), two NCBI sequences of *Cx. pipiens* sequences formed a clade sister to another containing three *Cx. quinquefasciatus* NCBI sequences and the “Cx quinquefasciatus MG S74” sequence with 78% bootstrap support. This is in accordance with other studies examining mitochondrial sequences (49) and morphological attributes (50). This shows that the 28S rRNA sequence can distinguish the two species and confirms that “Cx quinquefasciatus MG S74” is indeed a *Cx. quinquefasciatus* specimen. However, “Cx quinquefasciatus MG S75” is shown to be basal from other sequences within this *Cx. pipiens* species group-clade with 100% bootstrap support. Given that *Cx. quinquefasciatus* and *Cx. pipiens* are known to interbreed, it is plausible that this individual is a hybrid of the two species (51).

DISCUSSION

In metagenomics or surveillance studies on field-captured mosquitoes, the lack of reference rRNA sequences hinders good oligo-based depletion and efficient clean-up of RNA-seq data. Additionally, *de novo* assembly of rRNA sequences is complicated due to the high sequence conservation across all distantly related organisms that could be present in a single specimen, i.e., microbiota, parasites,

or vertebrate blood meal. Hence, we sought out to establish a method to bioinformatically filter out non-host rRNA reads for the accurate assembly of novel 28S and 18S rRNA reference sequences.

We found that phylogenetic reconstructions based on 28S sequences or concatenated 28S+18S rRNA sequences were able to cluster mosquito taxa according to species correctly and corroborates current mosquito classification. This demonstrates that our bioinformatics methodology reliably generates bona fide 28S and 18S rRNA sequences, even in specimens parasitized by water mites or engorged with vertebrate blood. Further, we were able to use 28S+18S rRNA taxonomy for molecular species identification when COI sequences were unavailable or ambiguous, thus supporting the use of 28S rRNA sequences as an rRNA barcode. rRNA barcodes would have the advantage of circumventing the need to additionally isolate and sequence DNA from specimens, as RNA-seq reads can be directly mapped against rRNA reference sequences. Even after depletion, there are sufficient numbers of reads (5-10% of reads per sample) to assemble complete rRNA contigs (Frangeul L, personal communication).

Phylogenetic inferences based on 28S or 18S rRNA sequences do not produce the same interspecific relationships, suggesting a difference in mutation rates between the two gene regions. Relative to 28S sequences, we observed more instances where multiple specimens have near-identical 18S rRNA sequences. This can occur for specimens belonging to the same species, but also for conspecifics sampled from different geographic locations, such as *An. coustani*, *An. gambiae*, or *Ae. albopictus.* More rarely, specimens from the same species group, such as *Cx. pseudovishnui* and *Cx. tritaeniorhynchus*, were also found to share 18S rRNA sequences. This was surprising given that the 18S rRNA sequences in our dataset is 1900 bp long. Concatenation of 28S and 18S rRNA sequences resolved this issue, enabling species delineation even among members of *Culex* species groups.

Taking advantage of our multi-country sampling, we examined whether rRNA or COI phylogeny can be used to discriminate conspecifics originating from different countries. Our dataset contains five of such species: *An. coustani*, *An. funestus*, *An. gambiae*, *Ae. albopictus*, and *Ma. uniformis*. Among the rRNA trees, the concatenated 28S+18S tree and 28S tree were able to discriminate between *Ma. uniformis* specimens from Madagascar, Cambodia, and the Central African Republic, and between *An. coustani* specimens from Madagascar and the Central African Republic (100% bootstrap value). In the COI tree, only *Ma. uniformis* was resolved into geographical clades comprising specimens from

Madagascar and specimens from Cambodia (72% bootstrap value). No COI sequence was obtained from one *Ma. uniformis* from the Central African Republic. The use of rRNA sequences seemingly provides more accurate phylogeographic information than COI alone.

Morphological identification suffers in accuracy when dealing with *Culex* species groups. Aside from sharing many morphological traits, sister species 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* (52). However, each of these species have distinct ecologies and host preferences, thus the challenge of correctly identifying vector species can affect epidemiological risk estimation for these diseases (51). In Asia, for example, cryptic members of the *Cx. vishnui* species group confound tracking of Japanese encephalitis virus transmission (53). The morphological differences between the *Culex* species *bitaeniorhynchus*, *tritaeniorhynchus*, *vishnui*, and *pseudovishnui* are often elusive, the former three having been morphologically identified in our study but later revealed by COI barcoding to be another species.

The *Cx. pipiens* species group is especially challenging as its member species are capable of interbreeding, showing genetic introgression to varying extents depending on the geographical population (54). The seven member species of this complex are practically indistinguishable morphologically and require molecular methods to discern (51,55). However, the 621 bp COI sequence amplified in our study did not contain enough nucleotide divergence to allow clear identification, given that the COI sequence of *Cx*. *quinquefasciatus* specimens differed from that of *Cx. pipiens* by a single nucleotide. Similarly, Batovska *et al* (56) found that even the Internal Transcribed Spacer 2 (ITS2) rDNA region, another common molecular marker, could not differentiate the two species. Other DNA molecular markers such as nuclear Ace-2 or CQ11 genes (55,57) or *Wolbachia pipientis* infection status (54) are typically employed in tandem. In our study, we used on 28S rRNA sequence-based taxonomy (Figure 2) to validate the identity of specimen “Cx quinquefasciatus MG S74” and suggests that specimen “Cx quinquefasciatus MG S75” might have been a *pipiens-quinquefasciatus* hybrid. These examples demonstrate how 28S rRNA sequences, concatenated with 18S or alone, contain enough resolution to differentiate between *Cx. pipiens* and *Cx*. *quinquefasciatus*. rRNA barcode taxonomy thus allows for more accurate species identification and ecological observations 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. An explicit investigation including all member species of these species groups in greater sample numbers is warranted to further test the degree of accuracy with which 28S and 18S rRNA sequences can delineate sister species.

Our study also included French Guianese *Culex* species *Cx. spissipes* (group Spissipes*)*, *Cx. pedroi* (group Pedroi), and *Cx. portesi* (group Vomerifer). These species belong to the New World subgenus *Melanoconion*, section Spissipes, with well-documented distribution in North and South Americas (58) and are vectors of encephalitic alphaviruses EEEV and VEEV among others (37–39). Indeed, our rooted rRNA and COI trees show the divergence of the three *Melanoconion* species from the major *Culex* clade comprising species broadly found across Africa and Asia (23,51,52,59). The topology of the concatenated 28S+18S tree puts the *Cx. portesi* and *Cx. pedroi* species-clades as sister groups (92% bootstrap support), with *Cx. spissipes* as a basal group within the *Melanoconion* clade (100% bootstrap support). This corroborates the systematics elucidated by Navarro and Weaver (60) using the ITS2 marker, and those by Sirivanakarn (58) and Sallum and Forattini (61) based on morphological features. Curiously, in the COI tree *Cx. spissipes* sequences were clustered with unknown species Cx. sp1, forming a clade sister to one containing other *Culex* (*Culex*) and *Culex* (*Oculeomyia*) species, albeit with very low bootstrap support. Previous phylogenetic studies based on the COI gene have consistently placed *Cx. spissipes* or the Spissipes group basal to other groups within the *Melanoconion* subgenus (62,63). However, these studies contain only *Culex* (*Melanoconion*) species in their dataset, apart from *Cx. quinquefasciatus* to act as an outgroup. This clustering of *Cx. spissipes* with non-*Melanoconion* species in our COI phylogeny could be an artefact of a much more diversified dataset rather than a true phylogenetic link.

The evolutionary histories inferred from rRNA-based and COI-based phylogenies in our study hardly correspond. rRNA phylogenies suggest the world of *Anopheles* is seemingly immense compared to any other genera with remarkably large evolutionary distances between one Anopheline species to another. This is not apparent in the COI phylogeny, perhaps reflecting the higher mutational rate of mitochondrial genomes relative to nuclear genomes (64). It would be interesting to further compare rRNA and COI phylogenies among other Anopheline subgenera beyond the subgenera *Anopheles* and *Cellia* represented in this study. Lamentably, we found during our search in

NCBI databases that many Anopheline rRNA records lack subgenus information, stressing the importance of including detailed taxonomy of mosquito specimens when reporting sequence data.

In contrast to the *Anopheles* case, two specimens of an unknown *Mansonia* species, “Ma sp.4 GF S103” and “Ma sp.4 GF S104”, provided an example where interspecies relatedness based on their COI sequences is greater than that based on their rRNA sequences in relation to “Ma titillans GF S105”. While all rRNA trees (Figure 2–4) placed “Ma titillans GF S105” as a sister taxon with 100% bootstrap support, the COI tree places Ma sp.4 basal to all other species except *Ur. geometrica*. This may hint at a historical selective sweep in mitochondrial genome, whether arising from mutations or linkage disequilibrium with inherited symbionts (65), resulting in the drastically distinct mitochondrial haplogroup found in Ma sp.4. To note, the COI sequences of “Ma sp.4 GF S103” and “Ma sp.4 GF S104” share 87.12 and 87.39% nucleotide similarity, respectively, with that of “Ma titillans GF S105”. Interestingly, the endosymbiont *Wolbachia pipientis* has been detected in *Ma. titilans* sampled from Brazil (66), which may contribute to the divergence of “Ma titillans GF S105” COI sequence away from those of other *Mansonia* species. The COI phylogeny of these *Mansonia* specimens highlights the drawbacks of using a mitochondrial DNA marker in determining evolutionary relationships (65), 28S and 18S rRNA sequences may be better able to illustrate evolutionary history than COI sequence alone.

Conclusions

Surveillance and microbiome discovery studies in wild mosquitoes are paramount for the establishment of public health measures to control arboviral diseases. Here we present a score-based rRNA assembly strategy and 234 newly generated 28S and 18S mosquito rRNA sequences. Our work has expanded the current rRNA reference library by presenting, to our knowledge, the first records for many species not previously present in public databases and paves the way for the assembly of many more. These novel rRNA sequences can improve mosquito RNA-seq metagenomics by expanding reference sequence data for the optimization of species-specific oligo-based depletion protocols, for streamlined species identification by rRNA barcoding and for improved RNA-seq data clean-up. In addition, rRNA barcodes could serve as an additional tool for mosquito taxonomy and phylogeny although further studies are necessary to reveal how they measure up against other nuclear or mitochondrial DNA marker systems (9,56,67–69).

We showed that phylogenetic inferences from a tree based on 28S rRNA sequences alone or concatenated 28S +18S rRNA sequences largely agree with contemporary mosquito classification and can be used for species identification given a reference sequence. In analysing the same set of specimen by COI or rRNA sequences, we found deep discrepancies in phylogenetic inferences. We conclude that while COI-based phylogeny is fairly useful to study recent evolutionary events, rRNA sequences may be better suited for investigations of more ancient evolutionary history.

METHODS

Sample collection

Mosquito specimens were sampled from 2019 to 2020 by medical entomology teams from the Institut Pasteur International Network: Institut Pasteur de Bangui (Central African Republic, Africa; CF), Institut Pasteur de Madagascar (Madagascar, Africa; MG), Institut Pasteur du Cambodge (Cambodia, South East Asia; KH), and Institut Pasteur de la Guyane (French Guiana, South America; GF). Adult mosquitoes were sampled using a combination of techniques including CDC light traps, BG sentinels, and human-landing catches. Sampling sites are non-urban locations including rural settlements in the Central African Republic, Madagascar, and French Guiana and national parks in Cambodia. Mosquitoes were identified using morphological identification keys on cold tables before preservation by flash freezing in liquid nitrogen and transportation in dry ice 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 TRIzol reagent according to manufacturer's protocol (Invitrogen, Thermo Fisher Scientific, Waltham, Massachusetts, USA, Cat. No. 15596018). Single mosquitoes were homogenised into 200 µL of TRIzol reagent and other of the reagents within the protocol were volume-adjusted accordingly. Following phase separation, RNA samples were isolated from the aqueous phase while DNA samples were isolated from the remaining interphase and phenol-chloroform phase. From here, RNA is used to prepare cDNA libraries for next

generation sequencing while DNA is used for PCR amplification and Sanger sequencing of the mitochondrial *cytochrome* c *oxidase subunit I* (COI) gene.

Probe depletion of rRNA

We tested a method for selective depletion of rRNA by Morlan *et al.* (17) on several mosquito species from the *Aedes*, *Culex*, and *Anopheles* genera. We designed 77 tiled 80 bp DNA probes antisense to the *Ae. aegypti* 28S, 18S, and 5.8S rRNA sequences. A pool of probes at a concentration of 0.04 µM were prepared. To bind probes to rRNA, 1 µL of probes was added to rRNA samples along with 2 µL of Hybridisation Buffer (100 mM Tris-HCl and 200 mM NaCl) to obtain a final volume of 20 µL and subjected to a slow-cool incubation starting at 95 ºC for 2 minutes, followed by cooling to 22 ºC at a rate of 0.1 ºC per second, followed by an additional 5 minutes at 22 ºC. The resulting RNA:DNA hybrids were treated with 2.5 µL Hybridase™ Thermostable RNase H (Epicentre, Illumina, Madison, Wisconsin, USA, Cat No. H39500) and incubated at 37 ºC for 30 minutes. To remove DNA probes, the mix was treated with 1 µL DNase I (Invitrogen, Cat No. 18047019) and purified with Agencourt RNAClean XP Beads (Beckman Coulter, Brea, California, USA, Cat No. A63987). The resulting RNA is used for total RNA sequencing to check depletion efficiency.

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. Non-depleted 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 RNA 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” (70). 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 (71). 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 (72). This method successfully assembled rRNA sequences for all specimens, including a parasitic *Horreolanus* water mite (122 sequences for 28S and 114 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. All assembled rRNA sequences from this study have been deposited in GenBank with accession numbers OM350214–OM350327 for 18S rRNA sequences and OM542339–OM542460 for 28S rRNA sequences.

COI amplicon sequencing

The mitochondrial COI gene was amplified from DNA samples using the universal “Folmer” primer set LCO1490 (5’- GGTCAACAAATCATAAAGATATTGG -3’) and HCO2198 (5’-TAAACTTCAGGGTGACCAAAAAATCA-3’), as is the standard for COI barcoding, producing a 658 bp product (73). 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 size-verified using gel electrophoresis and then gel-purified 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 (74,75). We successfully assembled a total of 106 COI sequences. All assembled COI sequences from this study have been deposited in GenBank with accession numbers OM630610–OM630715.

COI validation of morphology-based species identification

We analysed assembled COI sequences with BLASTn against the nucleotide collection (nr/nt) database to confirm morphology-based species identification. BLAST analyses revealed 32 cases where top hits indicated a different species identity, taking <95% nucleotide sequence similarity as the threshold to delineate distinct species (Supplementary Table 1). In these cases, the COI sequence of the specimen was then BLAST-aligned against a GenBank record representing the morphological species to verify that the revised identity is a closer match by a significant margin, i.e., more than 2% nucleotide sequence similarity. All species names reported hereafter reflect identities determined by COI barcoding except for cases where COI-based identities were ambiguous, in which case morphology-based identities were retained. In cases where matches were found within a single genus

but of multiple species, specimens were indicated as an unknown member of their genus (e.g., *Culex* sp.). Information of the highest-scoring references for all specimens, including details of ambiguous BLASTn results, are recorded in Supplementary Table 2.

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.

Phylogenetic analysis

Multiple sequence alignment (MSA) were performed on assembled COI and rRNA sequences using the MUSCLE software (Additional Files 6–9) (76,77). As shown in Supplementary Figure 2 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 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 tree reconstructions were performed with the MEGA X software using the maximum-likelihood method (46). The default parameters were used with the addition of 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 analysed a separate dataset combining our 28S rRNA sequences and full-length 28S rRNA sequences from the NCBI databases totalling 169 sequences from 58 species (12 subgenera). To serve as outgroups for the COI tree, we included sequences obtained from NCBI of three water mite species, *Horreolanus orphanus* (KM101004), *Sperchon fuxiensis* (MH916807), and *Arrenurus* sp. (MN362807).

DECLARATIONS

Ethics approval and consent to participate

Not applicable

605

606 **Consent for publication**

607 Not applicable

608

609 **Availability of data and materials**

610 RNA-seq fastq sequence data are available from the corresponding author on reasonable request.
611 Multiple sequence alignment files are included in this article as additional files. All sequences
612 generated in this study have been deposited in GenBank under the accession numbers OM350214-
613 OM350327 for 18S rRNA sequences, OM542339-OM542460 for 28S rRNA sequences, and
614 OM630610-OM630715 for COI sequences (Table S1).

615

616 **Competing interests**

617 The authors declare that they have no competing interests.

618

619 **Funding**

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

624

625 **Acknowledgements**

626 We thank members of the Saleh lab for valuable discussions and to Inès Partouche for laboratory
627 assistance. We especially thank all medical entomology staff of IP Bangui, IP Cambodge (Sony Yean,
628 Kimly Heng, Kalyan Chhuoy, Sreynik Nhek, Moeun Chhum, Kimhuor Sour and Pierre-Olivier
629 Maquart), IP Madagascar, and IP Guyane for assistance in field missions, laboratory work, and
630 logistics. We are also grateful to Dr Catherine Dauga for advice on phylogenetic analyses and to
631 Amandine Guidez for providing a French Guiana-specific COI reference library.

632

633 **ADDITIONAL FILES**

634 **Additional File 1: Figure S1.** Study workflow from specimens to sequences (PNG).

635 **Additional File 2: Table S1.** Taxonomic and sampling information on mosquito specimens and
636 associated accession numbers of their 28S, 18S, and COI sequences (XLSX).
637 **Additional File 3: Table S2.** COI sequence BLAST analyses summary (XLSX).
638 **Additional File 4: Figure S2.** Sequence conservation among 169 28S rRNA sequences obtained
639 from this study and from the NCBI databases (PDF).
640 **Additional File 5: Figure S3.** Phylogenetic tree based on 28S sequences generated from this study
641 (3900 bp) as inferred using maximum-likelihood method and constructed to scale in MEGA X (46).
642 Values at each node indicate bootstrap support (%) from 500 replication. Each specimen label
643 contains information on its taxonomy, origin (as indicated in 2-letter country codes), and specimen ID.
644 Label colours indicate genera: *Culex* in coral, *Anopheles* in purple, *Aedes* in dark blue, *Mansonia* in
645 dark green, *Limatus* in light green, *Coquillettidia* in light blue, *Psorophora* in yellow, *Mimomyia* in teal,
646 *Uranotaenia* in pink and *Eretmapodites* in brown. Scale bar at 0.05 is shown (PDF).
647 **Additional File 6:** Multiple sequence alignment of 169 28S rRNA sequences from this study and from
648 NCBI databases (FASTA).
649 **Additional File 7:** Multiple sequence alignment of 122 28S rRNA sequences, including two
650 sequences from *Horreolanus sp.* (FASTA).
651 **Additional File 8:** Multiple sequence alignment of 114 18S rRNA sequences, including two
652 sequences from *Horreolanus sp*. (FASTA).
653 **Additional File 9:** Multiple sequence alignment of 106 COI sequences (FASTA).
654

655 **REFERENCES**

656 1. WHO. Global vector control response 2017–2030. World Health Organization. 2017.
657 2. Webster JP, Gower CM, Knowles SCL, Molyneux DH, Fenton A. One health - an ecological
658 and evolutionary framework for tackling Neglected Zoonotic Diseases. Evol Appl. 2016 Feb
659 1;9(2):313–33.
660 3. Gould E, Pettersson J, Higgs S, Charrel R, de Lamballerie X. Emerging arboviruses: Why
661 today? One Heal. 2017 Dec 1;4:1–13.
662 4. GALE K, CRAMPTON J. The ribosomal genes of the mosquito, Aedes aegypti. Eur J Biochem.
663 1989;185(2):311–7.
664 5. Phelps WA, Carlson AE, Lee MT. Optimized design of antisense oligomers for targeted rRNA

depletion. Nucleic Acids Res. 2021;49(1):e5.

6. Fauver JR, Akter S, Morales AIO, Black WC, Rodriguez AD, Stenglein MD, et al. A reverse-transcription/RNase H based protocol for depletion of mosquito ribosomal RNA facilitates viral intrahost evolution analysis, transcriptomics and pathogen discovery. Virology. 2019;528:181–97.

7. Kukutla P, Steritz M, Xu J. Depletion of ribosomal RNA for mosquito gut metagenomic RNA-seq. J Vis Exp. 2013;(74):50093.

8. Hebert PDN, Cywinska A, Ball SL, DeWaard JR. Biological identifications through DNA barcodes. Proc R Soc B Biol Sci. 2003;270(1512):313–21.

9. Ratnasingham S, Hebert PDN. BOLD: The Barcode of Life Data System: Barcoding. Mol Ecol Notes. 2007;7(3):355–64.

10. Bishop-Lilly KA, Turell MJ, Willner KM, Butani A, Nolan NME, Lentz SM, et al. Arbovirus detection in insect vectors by Rapid, high- throughput pyrosequencing. PLoS Negl Trop Dis. 2010;4(11):e878.

11. Kumar N, Creasy T, Sun Y, Flowers M, Tallon LJ, Dunning Hotopp JC. Efficient subtraction of insect rRNA prior to transcriptome analysis of Wolbachia-Drosophila lateral gene transfer. BMC Res Notes. 2012;5:230.

12. Weedall GD, Irving H, Hughes MA, Wondji CS. 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. 2015;16(1):931.

13. Zakrzewski M, Rašić G, Darbro J, Krause L, Poo YS, Filipović I, et al. Mapping the virome in wild-caught Aedes aegypti from Cairns and Bangkok. Sci Rep. 2018;8(1):4690.

14. Belda E, Nanfack-Minkeu F, Eiglmeier K, Carissimo G, Holm I, Diallo M, et al. De novo profiling of RNA viruses in Anopheles malaria vector mosquitoes from forest ecological zones in Senegal and Cambodia. BMC Genomics. 2019;20(1):664.

15. Chandler JA, Liu RM, Bennett SN. RNA Shotgun Metagenomic Sequencing of Northern California (USA) Mosquitoes Uncovers Viruses, Bacteria, and Fungi. Front Microbiol. 2015;6:185.

16. Thongsripong P, Chandler JA, Kittayapong P, Wilcox BA, Kapan DD, Bennett SN. Metagenomic shotgun sequencing reveals host species as an important driver of virome

composition in mosquitoes. Sci Rep. 2021;11(1):8448.

17. Morlan JD, Qu K, Sinicropi D V. Selective depletion of rRNA enables whole transcriptome profiling of archival fixed tissue. PLoS One. 2012;7(8):e42882.

18. Diallo D, Fall G, Diagne CT, Gaye A, Ba Y, Dia I, et al. Concurrent amplification of Zika, chikungunya, and yellow fever virus in a sylvatic focus of arboviruses in Southeastern Senegal, 2015. BMC Microbiol. 2020;20:181.

19. Vasconcelos PFC, Costa ZG, Travassos da Rosa ES, Luna E, Rodrigues SG, Barros VLRS, et al. Epidemic of jungle yellow fever in Brazil, 2000: Implications of climatic alterations in disease spread. J Med Virol. 2001;65(3):598–604.

20. Cardoso J da C, de Almeida MAB, dos Santos E, da Fonseca DF, Sallum MAM, Noll CA, et al. Yellow fever virus in Haemagogus leucocelaenus and Aedes serratus mosquitoes, Southern Brazil, 2008. Emerg Infect Dis. 2010;16(12):1918–24.

21. Gaillet M, Pichard C, Restrepo J, Lavergne A, Perez L, Enfissi A, et al. Outbreak of oropouche virus in french guiana. Emerg Infect Dis. 2021;27(10):2711–4.

22. Kraemer MUG, Reiner RC, Brady OJ, Messina JP, Gilbert M, Pigott DM, et al. Past and future spread of the arbovirus vectors Aedes aegypti and Aedes albopictus. Nat Microbiol. 2019;4(5):854–63.

23. Auerswald H, Maquart PO, Chevalier V, Boyer S. Mosquito vector competence for japanese encephalitis virus. Vol. 13, Viruses. 2021. p. 1154.

24. Mukwaya LG, Kayondo JK, Crabtree MB, Savage HM, Biggerstaff BJ, Miller BR. 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. 2000;9(1):85–91.

25. Mwangangi JM, Muturi EJ, Muriu SM, Nzovu J, Midega JT, Mbogo C. The role of Anopheles arabiensis and Anopheles coustani in indoor and outdoor malaria transmission in Taveta District, Kenya. Parasites and Vectors. 2013;6:114.

26. Nepomichene TNJJ, Raharimalala FN, Andriamandimby SF, Ravalohery JP, Failloux AB, Heraud JM, et al. Vector competence of Culex antennatus and Anopheles coustani mosquitoes for Rift Valley fever virus in Madagascar. Med Vet Entomol. 2018;32(2):259–62.

27. Ratovonjato J, Olive MM, Tantely LM, Andrianaivolambo L, Tata E, Razainirina J, et al.

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. 2011;11(6):753–9.

28. Lutomiah J, Bast J, Clark J, Richardson J, Yalwala S, Oullo D, et al. Abundance, diversity, and distribution of mosquito vectors in selected ecological regions of Kenya: Public health implications. J Vector Ecol. 2013 Jun 1;38(1):134–42.

29. Brault AC, Foy BD, Myles KM, Kelly CLH, Higgs S, Weaver SC, et al. Infection patterns of o'nyong nyong virus in the malaria-transmitting mosquito, Anopheles gambiae. Insect Mol Biol. 2004;13(6):625–35.

30. Stevenson JC, Simubali L, Mbambara S, Musonda M, Mweetwa S, Mudenda T, et al. Detection of plasmodium falciparum infection in anopheles squamosus (diptera: Culicidae) in an area targeted for malaria elimination, Southern Zambia. J Med Entomol. 2016;53(6):1482–7.

31. Nikolay B, Diallo M, Boye CSB, Sall AA. Usutu virus in Africa. Vol. 11, Vector-Borne and Zoonotic Diseases. 2011. p. 1417–23.

32. Kim H, Cha GW, Jeong YE, Lee WG, Chang KS, Roh JY, et al. Detection of Japanese encephalitis virus genotype V in Culex orientalis and Culex pipiens (Diptera: Culicidae) in Korea. PLoS One. 2015;10(2):e0116547.

33. Vázquez González A, Ruiz S, Herrero L, Moreno J, Molero F, Magallanes A, et al. West Nile and Usutu viruses in mosquitoes in Spain, 2008-2009. Am J Trop Med Hyg. 2011;85(1):178–81.

34. Maquart PO, Sokha C, Boyer S. Mosquito diversity (Diptera: Culicidae) and medical importance, in a bird sanctuary inside the flooded forest of Prek Toal, Cambodia. J Asia Pac Entomol. 2021;24(4):1221–7.

35. Ndiaye EH, Fall G, Gaye A, Bob NS, Talla C, Diagne CT, et al. 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. 2016;9:94.

36. Bhattacharya S, Basu P, Sajal Bhattacharya C. The Southern House Mosquito, Culex quinquefasciatus: profile of a smart vector. J Entomol Zool Stud JEZS. 2016;73(42):73–81.

37. Weaver SC, Ferro C, Barrera R, Boshell J, Navarro JC. Venezuelan Equine Encephalitis.

Annu Rev Entomol. 2004;49:141–74.

38. Talaga S, Duchemin JB, Girod R, Dusfour I. The Culex Mosquitoes (Diptera: Culicidae) of French Guiana: A Comprehensive Review With the Description of Three New Species. J Med Entomol. 2021;58(1):182–221.

39. Turell MJ, O'guinn ML, Dohm D, Zyzak M, Watts D, Fernandez R, et al. Susceptibility of Peruvian Mosquitoes to Eastern Equine Encephalitis Virus. J Med Entomol. 2008;45(4):720–5.

40. Barrio-Nuevo KM, Cunha MS, Luchs A, Fernandes A, Rocco IM, Mucci LF, et al. Detection of Zika and dengue viruses in wildcaught mosquitoes collected during field surveillance in an environmental protection area in São Paulo, Brazil. PLoS One. 2020;15(10):e0227239.

41. Turell MJ. Vector competence of three Venezuelan mosquitoes (Diptera: Culicidae) for an epizootic IC strain of Venezuelan equine encephalitis virus. J Med Entomol. 1999;36(4):407–9.

42. Hoyos-López R, Soto SU, Rúa-Uribe G, Gallego-Gómez JC. Molecular identification of Saint Louis encephalitis virus genotype IV in Colombia. Mem Inst Oswaldo Cruz [Internet]. 2015 Aug 21 [cited 2022 Apr 21];110(6):719–25. Available from: http://www.scielo.br/j/mioc/a/YbSNxJQWw6dFLx5d4TttwfG/?lang=en

43. Arunachalam N, Philip Samuel P, Hiriyan J, Thenmozhi V, Gajanana A. Japanese encephalitis in Kerala, South India: Can Mansonia (Diptera: Culicidae) play a supplemental role in transmission? J Med Entomol. 2004;41(3):456–61.

44. Ughasi J, Bekard HE, Coulibaly M, Adabie-Gomez D, Gyapong J, Appawu M, et al. Mansonia africana and Mansonia uniformis are Vectors in the transmission of Wuchereria bancrofti lymphatic filariasis in Ghana. Parasites and Vectors. 2012;5(1):1–5.

45. Mitchell CJ, Forattini OP, Miller BR. Vector competence experiments with Rocio virus and three mosquito species from the epidemic zone in Brazil. Rev Saude Publica. 1986;20(3):171–7.

46. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: Molecular evolutionary genetics analysis across computing platforms. Mol Biol Evol. 2018;35(6):1547–9.

47. Harbach RE, Kitching IJ. The phylogeny of Anophelinae revisited: Inferences about the origin and classification of Anopheles (Diptera: Culicidae). Zool Scr. 2016;45(1):34–47.

48. Harbach RE. The Culicidae (Diptera): A review of taxonomy, classification and phylogeny. Zootaxa. 2007;1668(1):591–638.

785 49. Sun L, Li TJ, Fu WB, Yan ZT, Si FL, Zhang YJ, et al. The complete mt genomes of Lutzia
786 halifaxia, Lt. fuscanus and Culex pallidothorax (Diptera: Culicidae) and comparative analysis of
787 16 Culex and Lutzia mt genome sequences. Parasites and Vectors. 2019;12:368.
788 50. Harbach RE, Culverwell CL, Kitching IJ. Phylogeny of the nominotypical subgenus of Culex
789 (Diptera: Culicidae): insights from analyses of anatomical data into interspecific relationships
790 and species groups in an unresolved tree. Syst Biodivers. 2017;15(4):296–306.
791 51. Farajollahi A, Fonseca DM, Kramer LD, Marm Kilpatrick A. “Bird biting” mosquitoes and human
792 disease: A review of the role of Culex pipiens complex mosquitoes in epidemiology. Infect
793 Genet Evol. 2011;11(7):1577–85.
794 52. Nchoutpouen E, Talipouo A, Djiappi-Tchamen B, Djamouko-Djonkam L, Kopya E, Ngadjeu
795 CS, et al. Culex species diversity, susceptibility to insecticides and role as potential vector of
796 Lymphatic filariasis in the city of Yaoundé, Cameroon. PLoS Negl Trop Dis.
797 2019;13(4):e0007229.
798 53. Maquart PO, Boyer S. Culex vishnui. Trends Parasitol. 2022;(Vector of the Month).
799 54. Cornel AJ, Mcabee RD, Rasgon J, Stanich MA, Scott TW, Coetzee M. Differences in Extent of
800 Genetic Introgression between Sympatric Culex pipiens and Culex quinquefasciatus (Diptera:
801 Culicidae) in California and South Africa. J Med Entomol. 2003;40(1):36–51.
802 55. Zittra C, Flechl E, Kothmayer M, Vitecek S, Rossiter H, Zechmeister T, et al. Ecological
803 characterization and molecular differentiation of Culex pipiens complex taxa and Culex
804 torrentium in eastern Austria. Parasites and Vectors. 2016;9:197.
805 56. Batovska J, Cogan NOI, Lynch SE, Blacket MJ. Using next-generation sequencing for DNA
806 barcoding: Capturing allelic variation in ITS2. G3 Genes, Genomes, Genet. 2017;7(1):19–29.
807 57. Aspen S, Savage HM. Polymerase chain reaction assay identifies North American members of
808 the Culex pipiens complex based on nucleotide sequence differences in the
809 acetylcholinesterase gene Ace.2. J Am Mosq Control Assoc. 2003;19(4):323–8.
810 58. Sirivanakarn S. A review of the Systematics and a Proposed Scheme of Internal Classification
811 of the New World Subgenus Melanoconion of Culex (Diptera, Culicidae). Mosq Syst.
812 1982;14(4):265–333.
813 59. Takhampunya R, Kim HC, Tippayachai B, Kengluecha A, Klein TA, Lee WJ, et al. Emergence
814 of Japanese encephalitis virus genotype v in the Republic of Korea. Virol J. 2011;8:449.

60. Navarro JC, Weaver SC. Molecular phylogeny of the Vomerifer and Pedroi Groups in the spissipes section of the subgenus Culex (Melanoconion). J Med Entomol. 2004;41(4):575–81.

61. Sallum MAM, Forattini OP. Revision of the Spissipes Section of Culex (Melanoconion) (diptera: Culicidae). J Am Mosq Control Assoc. 1996;12(3):517–600.

62. Torres-Gutierrez C, Bergo ES, Emerson KJ, de Oliveira TMP, Greni S, Sallum MAM. Mitochondrial COI gene as a tool in the taxonomy of mosquitoes Culex subgenus Melanoconion. Acta Trop. 2016;164:137–49.

63. Torres-Gutierrez C, De Oliveira TMP, Emerson KJ, Bergo ES, Sallum MAM. Molecular phylogeny of Culex subgenus Melanoconion (Diptera: Culicidae) based on nuclear and mitochondrial protein-coding genes. R Soc Open Sci. 2018;5:171900.

64. Arctander P. Comparison of a mitochondrial gene and a corresponding nuclear pseudogene. Proc R Soc B Biol Sci. 1995;262(1363):13–9.

65. Hurst GDD, Jiggins FM. Problems with mitochondrial DNA as a marker in population, phylogeographic and phylogenetic studies: The effects of inherited symbionts. Proc R Soc B Biol Sci. 2005;272:1525–34.

66. De Oliveira CD, Gonçalves DS, Baton LA, Shimabukuro PHF, Carvalho FD, Moreira LA. Broader prevalence of Wolbachia in insects including potential human disease vectors. Bull Entomol Res. 2015;105(3):305–15.

67. Beebe NW. DNA barcoding mosquitoes: Advice for potential prospectors. Parasitology. 2018;145(5):622–33.

68. Behura SK. Molecular marker systems in insects: current trends and future avenues. Mol Ecol. 2006;15(11):3087–113.

69. Vezenegho SB, Issaly J, Carinci R, Gaborit P, Girod R, Dusfour I, et al. Discrimination of 15 Amazonian Anopheline Mosquito Species by Polymerase Chain Reaction—Restriction Fragment Length Polymorphism. J Med Entomol. 2022;tjac008.

70. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. Nucleic Acids Res. 2013;41(Database issue):D590–6.

71. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215(3):403–10.

845 72. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: A new
846 genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol.
847 2012;19(5):455–77.
848 73. Folmer O, Black M, Hoeh W, Lutz R, Vrijenhoek R. DNA primers for amplification of
849 mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. Mol Mar
850 Biol Biotechnol. 1994;3(5):294–9.
851 74. Rausch T, Fritz MHY, Untergasser A, Benes V. Tracy: Basecalling, alignment, assembly and
852 deconvolution of sanger chromatogram trace files. BMC Genomics. 2020;21(1):230.
853 75. Rausch T, Hsi-Yang Fritz M, Korbel JO, Benes V. Alfred: Interactive multi-sample BAM
854 alignment statistics, feature counting and feature annotation for long- and short-read
855 sequencing. Bioinformatics. 2019;35(14):2489–91.
856 76. Madeira F, Park YM, Lee J, Buso N, Gur T, Madhusoodanan N, et al. The EMBL-EBI search
857 and sequence analysis tools APIs in 2019. Nucleic Acids Res. 2019;47(W1):W636–41.
858 77. Edgar RC. MUSCLE: A multiple sequence alignment method with reduced time and space
859 complexity. BMC Bioinformatics. 2004;5:113.

860

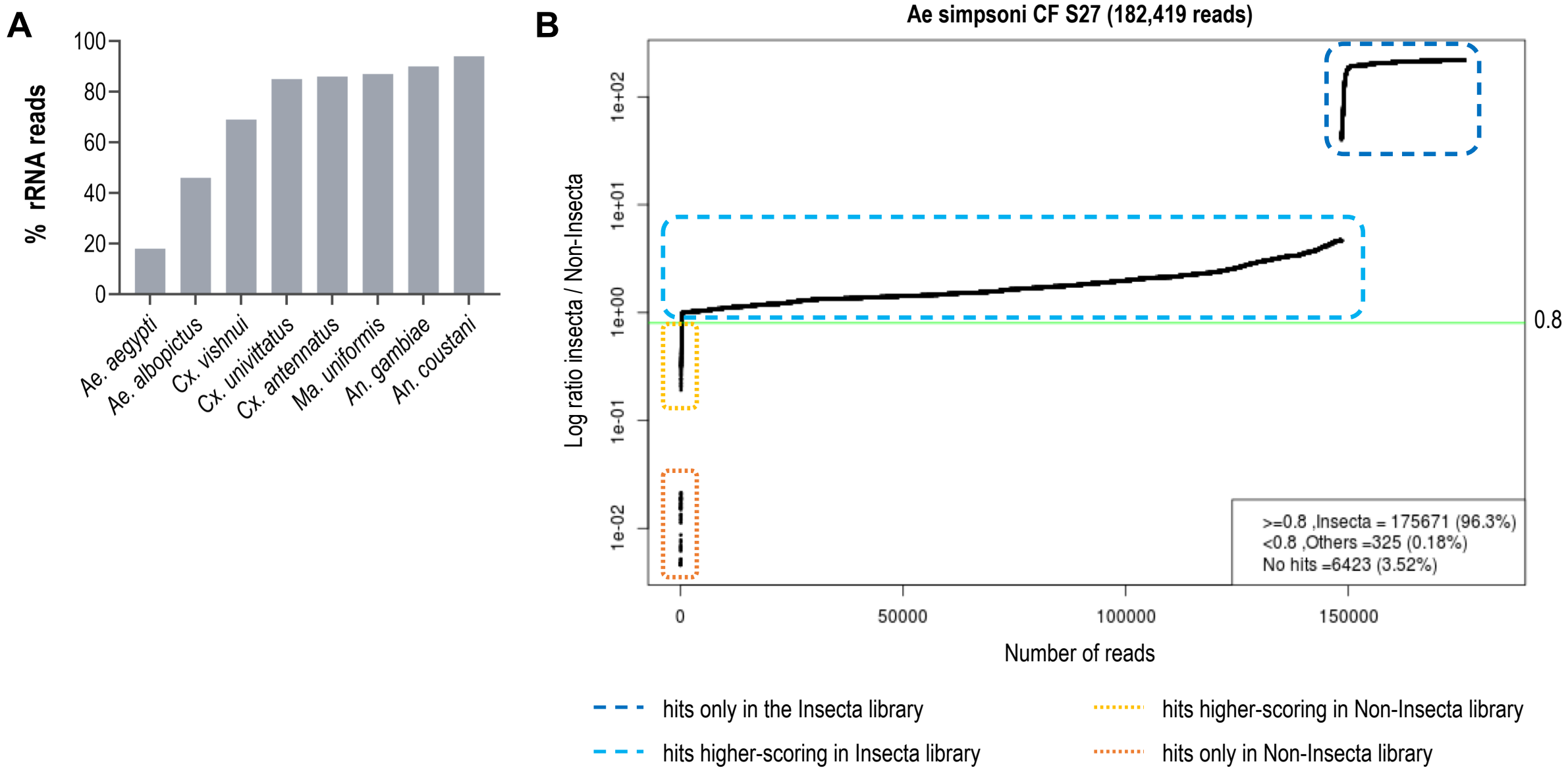
861 **TABLE AND FIGURES LEGENDS**

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

865 **Figure 1.** (**A**) Proportion of rRNA reads found in mosquito specimen pools depleted by probe
866 hybridisation followed by RNase H digestion. Probes were antisense to *Ae. aegypti* rRNA sequences.
867 (**B**) Read vs. score ratio plot of “Ae simpsoni CS S27”. Green line indicates a 0.8 cut-off where only
868 reads above this threshold are used in rRNA assembly.

869 **Figure 2.** Phylogenetic tree based on concatenated 28S and 18S sequences generated from this
870 study (3900+1900 bp) as inferred using the maximum-likelihood method and constructed to scale in
871 MEGA X (46). Values at each node indicate bootstrap support (%) from 500 replications. For
872 sequences from this study, each specimen label contains information on its taxonomy, origin (as
873 indicated in 2-letter country codes), and specimen ID. Label colours indicate genera: *Culex* in coral,
874 *Anopheles* in purple, *Aedes* in dark blue, *Mansonia* in dark green, *Culiseta* in maroon, *Limatus* in light

875 green, *Coquillettidia* in light blue, *Psorophora* in yellow, *Mimomyia* in teal, *Uranotaenia* in pink and
876 *Eretmapodites* in brown. Scale bar at 0.05 is shown.
877 **Figure 3.** Phylogenetic tree based on 28S sequences generated from this study and from NCBI
878 databases combined (3900 bp) as inferred using the maximum-likelihood method and constructed to
879 scale in MEGA X (46). Values at each node indicate bootstrap support (%) from 500 replications. For
880 sequences from this study, each specimen label contains information on its taxonomy, origin (as
881 indicated in 2-letter country codes), and specimen ID. Labels in bold indicate sequences derived from
882 NCBI. Label colours indicate genera: *Culex* in coral, *Anopheles* in purple, *Aedes* in dark blue,
883 *Mansonia* in dark green, *Culiseta* in maroon, *Limatus* in light green, *Coquillettidia* in light blue,
884 *Psorophora* in yellow, *Mimomyia* in teal, *Uranotaenia* in pink and *Eretmapodites* in brown. Scale bar
885 at 0.05 is shown.
886 **Figure 4.** Phylogenetic tree based on concatenated 28S and 18S sequences generated from this
887 study (3900+1900 bp) as inferred using maximum-likelihood method and constructed to scale in
888 MEGA X (46). Values at each node indicate bootstrap support (%) from 500 replications. For
889 sequences from this study, each specimen label contains information on its taxonomy, origin (as
890 indicated in 2-letter country codes), and specimen ID. Label colours indicate genera: *Culex* in coral,
891 *Anopheles* in purple, *Aedes* in dark blue, *Mansonia* in dark green, *Culiseta* in maroon, *Limatus* in light
892 green, *Coquillettidia* in light blue, *Psorophora* in yellow, *Mimomyia* in teal, *Uranotaenia* in pink and
893 *Eretmapodites* in brown. Scale bar at 0.05 is shown.
894 **Figure 5.** Phylogenetic tree based on COI sequences (621–699 bp) as inferred using the maximum-
895 likelihood method and constructed to scale in MEGA X (46). Values at each node indicate bootstrap
896 support (%) from 500 replications. Each specimen label contains information on its taxonomy, origin
897 (as indicated in 2-letter country codes), and specimen ID. Label colours indicate genera: *Culex* in
898 coral, *Anopheles* in purple, *Aedes* in dark blue, *Mansonia* in dark green, *Limatus* in light green,
899 *Coquillettidia* in light blue, *Psorophora* in yellow, *Mimomyia* in teal, *Uranotaenia* in pink and
900 *Eretmapodites* in brown. Scale bar at 0.05 is shown.


A
% rRNA reads
100
80
60
40
20
0
Ae. aegypti
Ae. albopictus
Cx. vishnui
Cx. univittatus
Cx. antennatus
Ma. uniformis
An. gambiae
An. coustani
B
Ae simpsoni CF S27 (182,419 reads)
Log ratio insecta / Non-Insecta
1e+02
1e+01
1e+00
1e-01
1e-02
0.8
>=0.8 ,Insecta = 175671 (96.3%)
<0.8 ,Others =325 (0.18%)
No hits =6423 (3.52%)
0
50000
100000
150000
Number of reads
hits only in the Insecta library
hits higher-scoring in Insecta library
hits higher-scoring in Non-Insecta library
hits only in Non-Insecta library

Cx sp.4 MG S62
Cx poicilipes MG S82
Cx pseudovishnui KH S7–S9; Cx tritaeniorhynchus KH S21–S23
Cx tritaeniorhynchus MG S115
Cx orientalis KH S14
Cx bitaeniorhynchus KH S18–S20
Cx quinquefasciatus MG S75
Cx sp.2 CF S48, S50
Cx sp.2 CF S49
Cx neavei MG S84
Cx sp.3 MG S61, S65
Cx sp.3 MG S63
Cx perexiguus MG S76
Cx duttoni CF S52
Cx perexiguus MG S60
Cx antennatus MG S57, S59
Cx antennatus MG S58
Ma sp.3 CF S35
Cx quinquefasciatus MG S74
Cx sp.1 KH S13
Cx portesi GF S94
Cx portesi GF S95, S96
Cx pedroi GF S106, S108
Cx pedroi GF S107
Cx spissipes GF S97–S99

subgenus *Melanoconion*

Ps ferox GF S109, S110
Ps ferox GF S111
Ae scapularis GF S85, S86
Ae serratus GF S88, S90

subgenus *Ochlerotatus*

Ae aegypti CF S25, S26
Ae aegypti CF S24
Ae albopictus KH S1, S3
Ae albopictus CF S36–S38; Ae albopictus KH S2
Ae simpsoni CF S27, S28
Ae simpsoni CF S29
Ae vittatus CF S30
Ae vittatus CF S32
Ae vittatus CF S31

subgenus *Stegomyia*

Er intermedius CF S54–S56
Ma uniformis CF S51 MG S77, S78
Ma uniformis KH S15
Ma uniformis KH S11, S16, S17
Ma indiana KH S10, S12
Ma uniformis MG S79
Ma sp.1 CF S33
Ma sp.2 CF S34
Ma sp.4 GF S103
Ma sp.4 GF S104
Ma titillans GF S105

subgenus *Mansonoides*

Ur geometrica GF S112–S114
Mi sp.1 MG S64
Cq venezuelensis GF S91, S92
Cq venezuelensis GF S93
Mi mediolineata MG S83
Li durhamii GF S100
Li durhamii GF S101, S102
An coustani CF S39, S40; An coustani MG S67
An coustani CF S41; An coustani MG S66
An baezai KH S4–S6

subgenus *Anopheles*

An funestus CF S43
An funestus MG S69
An funestus CF S42
An funestus CF S44; An funestus MG S70
An gambiae CF S45, S46; An gambiae MG S71–S73
An gambiae CF S47
An squamosus MG S68

subgenus *Cellia*

Ho sp.1 CF S34
Ho sp.2 CF S35

0.05