scientific reports

Check for updates

OPEN

Metagenomic shotgun sequencing reveals host species as an important driver of virome composition in mosquitoes

Panpim Thongsripong[1,6]✉, James Angus Chandler[1,6], Pattamaporn Kittayapong[2], Bruce A. Wilcox[3], Durrell D. Kapan[4,5] & Shannon N. Bennett[1]

High-throughput nucleic acid sequencing has greatly accelerated the discovery of viruses in the environment. Mosquitoes, because of their public health importance, are among those organisms whose viromes are being intensively characterized. Despite the deluge of sequence information, our understanding of the major drivers influencing the ecology of mosquito viromes remains limited. Using methods to increase the relative proportion of microbial RNA coupled with RNA-seq we characterize RNA viruses and other symbionts of three mosquito species collected along a rural to urban habitat gradient in Thailand. The full factorial study design allows us to explicitly investigate the relative importance of host species and habitat in structuring viral communities. We found that the pattern of virus presence was defined primarily by host species rather than by geographic locations or habitats. Our result suggests that insect-associated viruses display relatively narrow host ranges but are capable of spreading through a mosquito population at the geographical scale of our study. We also detected various single-celled and multicellular microorganisms such as bacteria, alveolates, fungi, and nematodes. Our study emphasizes the importance of including ecological information in viromic studies in order to gain further insights into viral ecology in systems where host specificity is driving both viral ecology and evolution.

Viruses are critically important to human and environmental health, and their diversity is predicted to be vast[1,2]. Next Generation Sequencing (NGS) technology has precipitated the discovery of many viruses, and expanded our knowledge of virus diversity, taxonomy, and evolution[3–5]. The RNA viruses of arthropod species, including mosquitoes, have been intensively characterized in the past several years, and stand out in their unparalleled diversity[6–19].

Disease vectors such as mosquitoes pose a significant threat to public health. One sixth of the illness and disability suffered worldwide is due to vector-borne diseases, many of which are caused by mosquito-borne RNA viruses[20]. Examples of these viruses are dengue virus of the family *Flaviviridae*, chikungunya virus of the family *Togaviridae*, Rift Valley fever virus of the family *Phenuiviridae*, and Lacrosse virus of the family *Peribunyaviridae*. In addition to pathogenic viruses, myriad other mosquito-associated RNA viruses belonging to at least 15 other families do not pose direct public health concerns[21].

Most virome studies apply NGS tools to identify new viruses and describe their diversity. This type of study may lead to the discovery of insect-specific viruses that have the ability to interfere with arbovirus transmission[22–26]. To maximize throughput while minimizing the sequencing cost, discovery-based studies have benefited from combining a large number of insects into a few pooled samples. Often, individuals from different species, or multiple locations, are pooled together hence information on their host-specificity and geographical location is lost[6,11,14]. The paucity of ecological information in virome studies limits our understanding of the host

[1]Department of Microbiology, Institute for Biodiversity Science and Sustainability, California Academy of Sciences, San Francisco, CA, USA. [2]Center of Excellence for Vectors and Vector-Borne Diseases, Faculty of Science, Mahidol University At Salaya, Nakhon Pathom, Thailand. [3]Global Health Group International, ASEAN Institute for Health Development, Mahidol University At Salaya, Nakhon Pathom, Thailand. [4]Department of Entomology and Center for Comparative Genomics, Institute for Biodiversity Sciences and Sustainability, California Academy of Sciences, San Francisco, CA, USA. [5]Center for Conservation and Research Training, Pacific Biosciences Research Center, University of Hawai'i At Manoa, Honolulu, HI, USA. [6]These authors contributed equally: Panpim Thongsripong and James Angus Chandler. ✉email: pthongsripong@calacademy.org

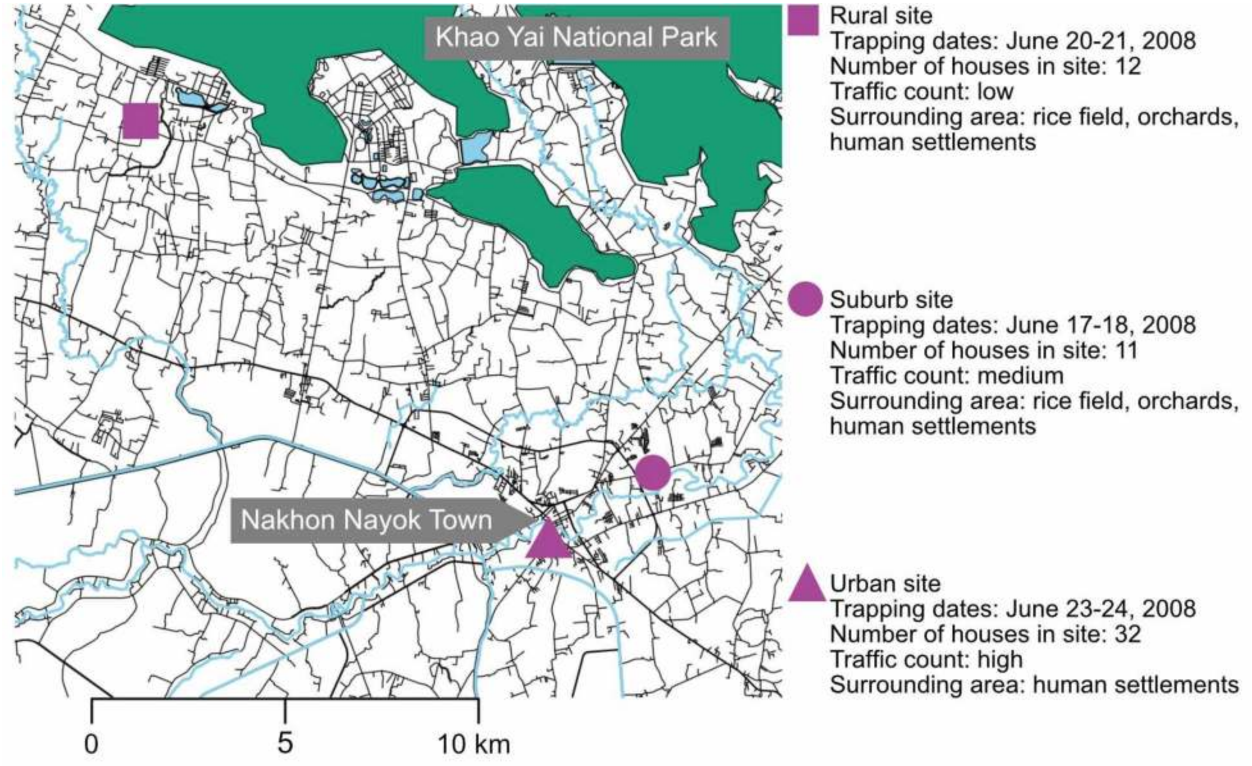
Khao Yai National Park
Nakhon Nayok Town
■ Rural site
Trapping dates: June 20-21, 2008
Number of houses in site: 12
Traffic count: low
Surrounding area: rice field, orchards, human settlements
● Suburb site
Trapping dates: June 17-18, 2008
Number of houses in site: 11
Traffic count: medium
Surrounding area: rice field, orchards, human settlements
▲ Urban site
Trapping dates: June 23-24, 2008
Number of houses in site: 32
Traffic count: high
Surrounding area: human settlements
0
5
10 km

Figure 1. Three study sites located along an urbanization gradient in Nakhon Nayok province of central Thailand. Detailed habitat characteristics can be found in our previous publication[27]. The map was created using R v3.5.3 (www.r-project.org).

ranges of these viruses relative to their geographical distribution. As a result, we have little information on the relative importance of ecological drivers determining viral community structure.

In this study, we characterized RNA viruses and other microorganisms in pools of individuals of three common but overlooked vector species in Thailand: *Armigeres subalbatus*, *Culex fuscocephala*, and *Mansonia uniformis,* collected from three habitat types along a previously characterized rural to urban gradient[27]. Our full factorial design (all three mosquito species were collected at all three sites) allows us to explicitly determine the relative importance of host species and habitat in structuring viral communities. As obligate intracellular parasites, viruses rely on specific molecular interactions with hosts to perform their biological functions. Thus, we hypothesize that the pattern of virus presence in our mosquito samples is defined primarily by host species rather than by the host geographical locations or habitats. Finally, to maximize the sensitivity of NGS, we tested a laboratory method[28] to deplete host ribosomal RNA, and increase the relative proportion of microbial genetic material prior to sequencing. This method could help reduce the need to pool samples across locations and species, and maintain ecological information in future virome studies.

Methods

Mosquito collection and RNA extraction. We described study site characteristics and adult mosquito collection in detail in a previous publication[27]. In summary, the study sites were located along a forest to urban landscape gradient in Nakhon Nayok province of central Thailand. Adult mosquitoes were collected during the rainy season of 2008 using a combination of trap types (BG Sentinel, Mosquito Magnet, CDC light trap, and CDC backpack aspirator). In total, over 83,000 adult mosquitoes were collected and transported to the laboratory on dry ice. Species identification based on available morphological keys revealed 109 species of female mosquitoes[29–34].

In this study, we included nine samples of three vector species: *Ar. subalbatus*, *Cx. fuscocephala*, and *Ma. uniformis,* each collected from rural, suburban, and urban habitats (Fig. 1). For each sample, we combined 25 female mosquitoes of the same species collected from the same study site during the same sampling period. We visually confirmed that these mosquitoes were not blood-engorged. Mosquitoes were then homogenized in 250 µl of Phosphate Buffered Saline using a Tissue Lyser II (QIAGEN, USA) and stainless steel beads before mixing with TRIzol LS (Invitrogen, USA) at the ratio of 1:7. Samples were kept at – 80 °C until RNA extraction according to the manufacturer's protocol.

Mosquito specific ribosomal RNA probes, and host ribosomal RNA depletion. In order to increase the relative proportion of microbial RNA, we used biotin-labeled mosquito-specific ribosomal RNA (rRNA) probes to capture and deplete mosquito rRNA prior to sequencing. The procedures for probe construction and rRNA depletion followed a published protocol[28]. In short, the rRNA probes were created by reverse transcribing *Cx. pipiens* (of a laboratory colony) *rRNA* gene using a set of custom designed mosquito-specific small subunit (SSU) and large subunit (LSU) ribosomal RNA primers attached with T7 promoter sequences (MEGAscript T7 Transcription Kit, Invitrogen, USA). The primer sequences are listed in Supplementary Table 1. The biotin-labeled UTP (Roche Life Science, USA) and CTP (Enzo Life Sciences, USA) were used. Since *rRNA* is relatively conserved between these mosquito species (e.g. 81–92% identity for *28 s rRNA*, and 91–97% identity for *18 s rRNA*; based on our preliminary analysis), we surmised that the probes constructed from *Cx. pipiens* would bind to other species of mosquito rRNA.

To deplete mosquito rRNA from the samples, the probes and RNA samples were combined and allowed to hybridize at 70 °C for 5 min, before ramping down to 25 °C with increments of 5 °C and one minute. Next, we used streptavidin magnetic beads (NEB, USA) to separate out the hybridized rRNA and the un-hybridized probes. The bead wash procedures were repeated three times to improve depletion efficiency. The resulting rRNA-depleted RNA samples, along with their paired non-depleted RNA samples, were used for next steps. We designate rRNA depleted sample "RD" and un-depleted control sample "UD" for the rest of this manuscript.

cDNA synthesis, cDNA library preparation, and high-throughput sequencing.

The first strand cDNA synthesis was carried out using SuperScript III Reverse Transcriptase (Life Technologies) and random primers following the manufacturer's protocol. This was followed by the second strand cDNA synthesis using DNA Polymerase I (New England Biolabs). Sequencing libraries were created using Nextera XT kit (Illumina). Purified libraries were quantified by Qubit 2.0 Fluorometer (Life Technologies, USA) and assessed by Agilent 2100 Bioanalyzer (Agilent Technologies, USA) for average fragment sizes. Bioanalyzer traces for representative paired RD and UD samples are shown in Supplementary Figure 1. The libraries were combined using equimolar ratio and were sequenced on the Illumina MiSeq platform (Illumina, USA) using the paired-end V2 500 cycle reagent kit at the Center for Comparative Genomics, California Academy of Sciences. We performed two separate sequencing runs (Supplementary Table 2). Both RD and UD no-template controls (NTC) were included in each sequencing run.

Sequence processing and viral contig identification.

Raw sequences were filtered for quality and the Illumina adapters were trimmed using Trim Galore! v0.4.5 (https://github.com/FelixKrueger/TrimGalore) using default parameters. The remaining reads were assembled into contigs using Trinity v2.8.4 with default parameters[35,36]. To identify viruses in the sample, the resulting contigs were compared to the NCBI protein database (downloaded September 25, 2018) using DIAMOND v0.9.22[37] with an E-value cutoff of 10^{-03}. The most similar match was identified as having the lowest e-value.

Viral phylogenetic analysis and coverage estimation.

In order to increase the coverage of viral genomes, viral contigs belonging to the same virus across samples and across the two preparation types were assembled using Sequencher v5.1 (Gene Codes Corporation, USA). The resulting assembled genomes or partial genomes were checked manually. The open reading frames (ORF) were predicted using ORFfinder (www.ncbi.nlm.nih.gov/orffinder/) and were confirmed to match viral proteins using NCBI blastp suite. To infer phylogenetic relationships between RNA viruses, selected viral amino acid sequences were aligned using MAFFT v7 employing the E-INS-i algorithm[38,39]. The best-fit model of amino acid substitutions was determined using ProtTest v3.4.2[40]. Phylogenetic analysis was performed using RAXML blackbox[41] implemented in Cipres Science Gateway application[42]. The resulting phylogenetic trees were visualized in iToL v4.4.1[43].

The sequence depth and coverage for each viral genome or partial genome was determined by mapping the raw, quality-checked reads to the nucleotide sequences of the viral genome or partial genome using Bowtie2 v2.2.6[44] and default parameters. Samtools v1.8[45] was used to process the resulting SAM file from the Bowtie alignment to calculate the average read depth and coverage.

Identification of endogenous virus elements (EVEs).

Because the genomes of the three mosquito species are currently unavailable, we cannot definitively exclude EVEs from our data. Instead, we manually checked and removed virus contigs closely related to known EVEs using NCBI blastx. For example, contigs similar to multiple sclerosis associated retrovirus were removed[46].

Identification of non-viral microorganisms.

To classify non-viral reads, Kraken v2.0.7-beta was used[47]. The NCBI's Nucleotide (nt) database was used to build a Kraken-specific reference database. In order to confirm the genus of the organism, SortmeRNA was used to filter rRNA reads from the data[48]. Then, the rRNA reads were used to search for matches using NCBI blast suite against the nucleotide database with an E-value cutoff of 10^{-3}. The most similar match was identified as having the lowest e-value.

Statistical analysis.

To compare virus occurrences across samples, we first normalized the number of reads (i.e. sampling effort) so that they all contained the same number of sequences ($n = 418{,}000$, which is the size of the smallest library; Supplementary Table 2). This is done by randomly selecting a subset of raw, quality-checked reads from each sample using Seqtk tool kit (https://github.com/lh3/seqtk). This data set was then assembled de novo using Trinity, searched for viral matches with Diamond, and sequence depths for contigs were determined using Bowtie2 v2.2.6 as described previously. A data matrix containing log-transformed read numbers that mapped to each virus were used as input for principal component analysis (PCA) to assess whether their

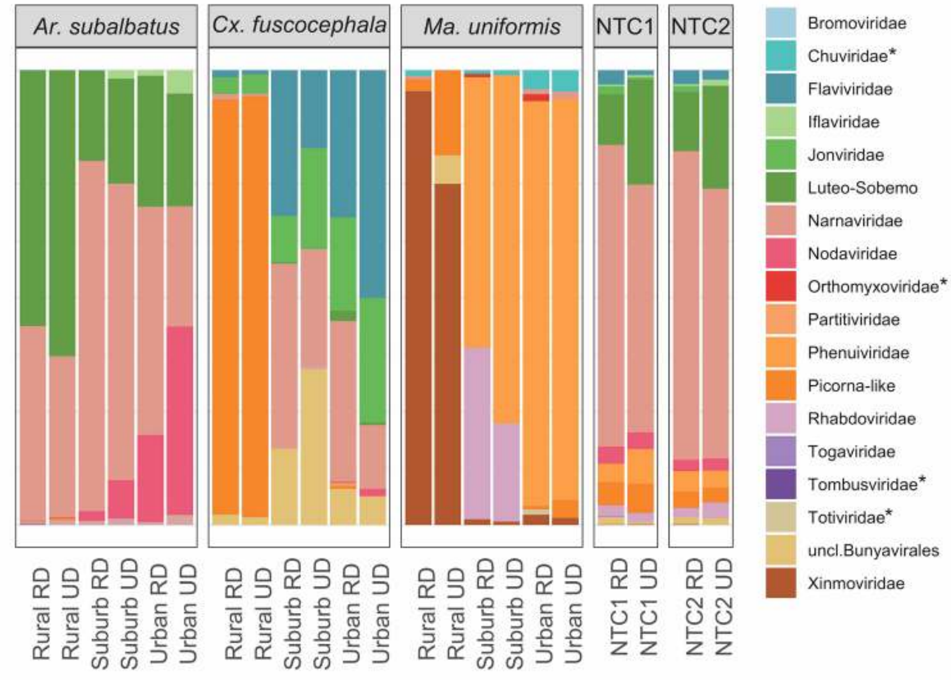

Figure 2. The proportion of virus families found in all mosquito samples and NTCs. RD = rRNA depleted, UD = rRNA not depleted. Asterisks (*) indicate virus groups that are likely contamination. NTC1 and NTC2 refer to no-template controls for separate sequencing runs (Supplementary Table 2).

pattern of occurance is determined by host species or habitat. PCA was performed using the "prcomp" function (R v3.5.3), which performs PCA on the given data matrix.

To compare parameter values such as average length of virus contigs, percentage, and number of reads between RD and UD samples, we used Mann–Whitney U test with the alpha level of 0.05. The non-parametric test was chosen because the data were not normally distributed (as revealed by Q–Q plots and the Shapiro–Wilk test). All statistical analyses, and data visualization were performed using R v3.5.3 and Rstudio v1.2.

Criteria used for identifying contaminating virus taxa.

A given virus taxon was defined as an external contamination if it met *all* of the following: (1) it was found in the no-template controls (NTCs); (2) it was represented by ≤ 80 reads (we chose this cutoff number because a majority, or around 97%, of contigs found in NTCs was represented by 80 reads or less); (3) it was found in mosquito samples in a random pattern (i.e. its presence had no species-specific pattern or habitat-specific pattern); and (4) it is not similar to a previously characterized insect-specific virus. In addition, a given viral taxon was likely a cross contamination between samples if it met criteria (1)–(3) mentioned above. We indicated taxa that are likely contaminants or cross-contaminants in Fig. 2. The complete list of virus contigs found in NTCs is shown in Supplementary Table 3. The complete lists of non-viral taxa found in NTCs according to analysis using SortmeRNA and Kraken are shown in Supplementary Tables 4 and 5, respectively.

Results

Read numbers and length.

The Illumina MiSeq sequencing generated over 13 million reads. Not including NTCs, the average number of raw reads per sample was 626,988 (SE = 65,936). After quality filtering, the average number of reads dropped to 616,915 (SE = 64,768). Our quality filtering step removed on average 1.53% of raw reads (Supplementary Figure 2). The quality filtered reads were then assembled to produce an average of 5032 contigs per sample (SE = 1216). The average number of assembled contigs per RD sample was significantly higher than in the UD samples (paired t-test, t = 4.2573, p value = 0.003; Supplementary Figure 2). The average read length after the filtering step across all samples was 184 bp (SE = 5). The average contig length was 308 bp (SE = 6). The numbers of reads and contigs for all samples are listed in Supplementary Table 2.

Mosquito viromes.

At least 21 putative viruses were identified from all samples. These viruses belong to the order *Bunyavirales* (family *Phenuiviridae*, *Jonviridae*, and an unclassified clade), order *Mononegavirales* (family *Rhabdoviridae* and *Xinmoviridae*), family *Flaviviridae*, family *Iflaviridae*, an unclassified clade closely related to *Luteoviridae* and *Sobemovirus* (referred to as Luteo-Sobemo-related group[6]), unclassified clades related to family *Narnaviridae*, family *Nodaviridae*, and family *Picornaviridae* (Fig. 2). In addition, we recovered a small number of virus reads (≤ 80) aligning to contigs that were classified in the following groups: *Togaviridae*, *Totiviridae*, *Tombusviridae*, *Chuviridae*, *Orthomyxoviridae*, *Partitiviridae*, and an unclassified clade closely related to negeviruses. Notable amongst these was the insect-specific Negevirus-like virus, and the virus in the family

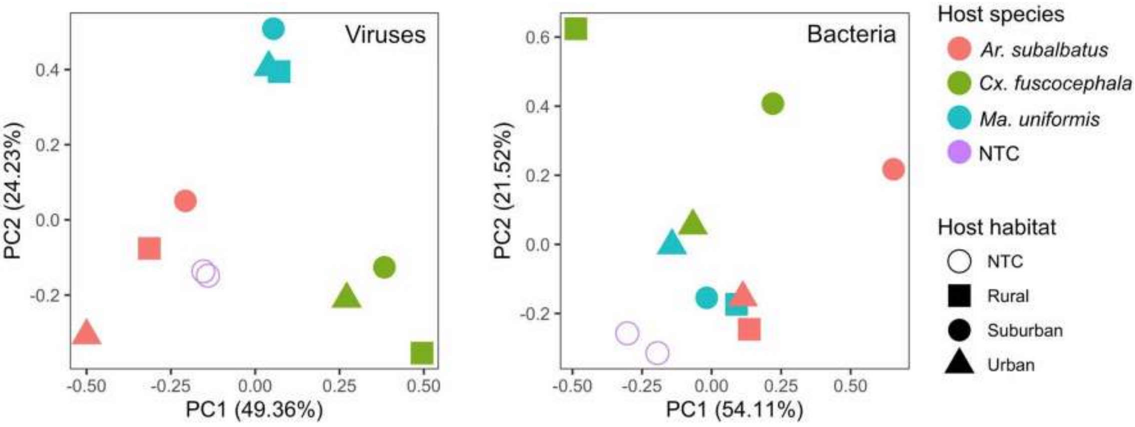

Figure 3. Similarity of virus community (**a**) and bacteria community (**b**) from RD samples. A Principle Component Analysis was performed based on log-transformed read numbers of bacterial and virus taxa found in each mosquito species and habitat type.

Characteristics	UD samples	SD	RD samples	SD	n	*p* value	Note
Total number of virus group found	12		18				Group was defined at family level except for "luteo-sobemo" group
Total number of viral contigs found	218		419				Contigs from trinity
Average length of viral contigs	541	566	573	551	637	0.491	Contigs from trinity
Average percentage of virus reads in total reads per sample	0.095	0.160	0.438	0.661	18	0.094	
Average percentage of viral reads in total reads per contig	0.005	0.0003	0.017	0.109	292	0.023	Assembled contigs
Average percentage of reads mapped to the Large Subunit region of mosquito ribosomal gene (LSU) in total reads per sample							Reference sequence Accession number X89642
Cx. fuscocephala samples	16.213	0.637	10.228	3.863			
Ar. subalbatus samples	11.157	1.573	10.296	1.853			
Ma. uniformis samples	10.791	2.066	13.487	1.687			
All samples	12.720	2.945	11.337	2.811	18	0.297	
Average percentage of reads mapped to the Small Subunit region of mosquito ribosomal gene (SSU) in total reads per sample							Reference sequence Accession number AY988445
Cx. fuscocephala samples	38.259	8.177	7.43	1.709			
Ar. subalbatus samples	46.079	8.712	8.322	4.484			
Ma. uniformis samples	53.734	6.127	12.444	3.230			
All samples	46.024	9.486	9.399	3.706	18	< 0.001	

Table 1. Comparison of characteristics between rRNA-depleted samples (RD) and undepleted samples (UD). Statistical analysis was performed using Mann–Whitney U test.

Togaviridae, which had 98% amino acid similarity to a mosquito-borne zoonotic virus: Getah virus of the genus *Alphavirus*. Detailed information about all viral contigs found in this study, and their most closely related viruses is listed in Supplementary Table 3.

Virome structure across host species and habitat. We used Principal Component Analysis to determine how mosquito viromes differ across host species relative to habitats. The results indicate that the virus communities in both the RD (Fig. 3a) and UD (Supplementary Figure 3) samples were grouped based on host species rather than the habitat of mosquito collection. This is in contrast to the bacterial community (Fig. 3b), where grouping patterns did not show association with host or habitat type.

Depth and coverage: comparison between RD and UD samples. We tested a laboratory method to deplete host ribosomal RNA from total RNA. Table 1 shows the comparison between samples with (RD) and without (UD) host rRNA depletion and the statistical tests used. The total number of virus contigs and virus groups was higher in RD than in the UD sample set. The average percentage of virus reads per RD sample was higher than in the UD sample, though this was not statistically significant (n = 18, *p* value = 0.094). The average percentage of viral reads per contig in the RD samples was significantly higher than UD samples (n = 292, *p* value = 0.023). Successful depletion was apparent only for the small subunit of mosquito ribosomal (SSU) gene but not for the large subunit (LSU). The number of reads mapped to the SSU gene was significantly lower in RD samples than in UD samples (p value < 0.001), a difference that was not observed for the LSU.

Evolutionary history.

Family Flaviviridae; Genus Flavivirus. Viruses of the family *Flaviviridae* possess positive-sense single-stranded RNA (ssRNA) genomes of approximately 9–13 kbp encoding a single polyprotein[49]. We found at least two partial viral genomes from the insect-specific clade of the genus *Flavivirus* in our samples (Fig. 4). The first putative virus, Culex fuscocephala-associated flavivirus, found in all three *Cx. fuscocephala* samples, formed a clade with many flaviviruses isolated from *Culex* spp. We were able to recover > 10 kbp of its genome. The most similar sequence (99% amino acid identity) belonged to a flavivirus isolated from multiple *Culex* species (*Cx. tritaeniorhynchus*, *Cx. vishnui*, or *Cx. fuscocephala*) collected in Myanmar[50]. The second putative virus, Mansonia uniformis-associated flavivirus (Fig. 4), was represented by a short contig (207 bp) found in the suburban *Ma. uniformis* sample. It is related to Palm Creek virus isolated from *Coquillettidia xanthogaster* in Australia (78% amino acid identity), and Nakiwogo virus isolated from *Mansonia* sp. in Uganda (45% amino acid identity).

Family Iflaviridae; Genus Iflavirus. The viruses of the genus *Iflavirus* possess linear positive-sense ssRNA genomes of approximately 9–11 kb in length encoding a single polyprotein[51]. All members infect arthropod hosts with the majority infecting insects[51]. We found four non-overlapping partial genomic fragments of iflavirus, provisionally called Armigeres subalbatus-associated iflavirus (Fig. 4), in the urban and suburban *Ar. subalbatus* samples. Based on our phylogenetic analysis, all four fragments are clustered with Armigeres iflavirus found in an unknown *Armigeres* species collected in the Philippines[52]. The percent amino acid identity to the reference Armigeres iflavirus ranged from 48 to 79% (73–606 amino acids long).

Unclassified clade closely related to Luteoviridae and Sobemovirus. Both *Sobemovirus* and *Luteoviridae* are groups of plant viruses with linear and non-segmented positive sense ssRNA genomes of approximately 4 kb and 6 kb in length, respectively[53]. A recent study found many new virus genomes in diverse invertebrate species classified to a novel clade that is phylogenetically divergent but related to *Sobemovirus* and *Luteoviridae*[6]. At least three partial virus genomes identified in our samples belonged to this novel clade.

Using RNA-dependent RNA polymerase (RdRp) proteins to construct a phylogeny, the putative viruses identified in this study clustered with 3 different clades. The first putative virus, provisionally called Armigeres subalbatus-associated sobemo-like virus 1 (Fig. 4) found in all *Ar. subalbatus* samples fell within a clade containing Yongsan sobemo-like virus 1, Wenzhou sobemo-like virus 4, and Hubei mosquito virus 2. The second putative virus, provisionally called Culex fuscecephala-associated sobemo-like virus (Fig. 4), represented by two non-overlapping RdRp fragments, was similar to Hubei sobemo-like virus 41 (95% and 97% amino acid identity). This virus was found in all *Cx. fuscocephala* samples. The last putative virus, provisionally called Armigeres subalbatus-associated sobemo-like virus 2 (Fig. 4), found in the urban *Ar. subalbatus* sample, formed a clade with Culex mosquito virus 6.

Narnaviridae and the closely-related Ourmia-like virus. Viruses in the family *Narnaviridae* possess a single molecule of non-encapsidated positive-sense ssRNA of 2.3–3.6 kb[54]. Two genera are currently recognized within this family: genus *Mitovirus* (fungi viruses) and *Narnavirus* (viruses of yeast)[53]. Members of the genus *Ourmiavirus*, closely related to the *Narnaviridae*, are plant viruses with genomes consisting of three positive-sense ssRNAs[55]. At least 4 partial virus genomes could be classified within the *Narnaviridae* and *Ourmiavirus* in our study. The first virus, provisionally called Armigeres subalbatus-associated ourmia-like virus (Fig. 4), was found in large numbers (0.15–0.38% of total reads) in all three *Ar. subalbatus* samples. Its partial genome was similar to the sequence of Hubei mosquito virus 3 (99% amino acid sequence similarity) identified in a pool of multiple mosquito species collected in China[6]. Our newly described ourmia-like virus clustered within a clade of ourmia-like invertebrate viruses.

The second virus, provisionally called Culex fuscocephala-associated narna-like virus (Fig. 4) was similar to the sequence of Zhejiang mosquito virus 3 (93% amino acid identity), which was identified in a pool of multiple mosquito species collected in China[6]. This virus was found in all *Cx. fuscocephala* samples. The third virus, provisionally called Culex fuscocephala-associated mitovirus 1 (Fig. 4), was similar to Hubei narna-like virus 25 (68–69% amino acid identity). This virus was found in the rural *Cx. Fuscocephala* sample. The fourth virus, provisionally called Culex fuscocephala-associated mitovirus 2 (Fig. 4) was similar to Erysiphe Necator mitovirus 1 (36% amino acid identity). This virus was also found in the rural *Cx. fuscocephala* sample.

Nodaviridae. The family *Nodaviridae* includes two genera, *Alphanodavirus* (infects insects) and *Betanodavirus* (infects fish)[53]. Genus *Gammanodavirus* infecting prawn and shrimp has also been proposed[56]. Nodaviruses genomes consist of two molecules of positive sense ssRNA. RNA-1 (3.1 kb) encodes protein A with functions such as polymerase, and RNA-2 (1.4 kb) encodes protein alpha[57]. Phylogenetic analysis showed that the partial RNA-1 found in our study (~ 2.5 kb) formed a clade with Lunovirus (from a carnivore fecal virome study), Caninovirus (from a canid gut virome study), and Sanxia water strider virus 16 (from a pool of multiple water strider species collected in China). The most similar sequence was Sanxia water strider virus 16 (39% amino acid similarity). This virus, provisionally called Armigeres subalbatus-associated nodavirus (Fig. 4), was found in all three *Ar. subalbatus* pools. Another contig resembling capsid protein (~ 1 kb), likely representing the RNA-2 fragment of the same virus, was also found in all *Ar. subalbatus* samples.

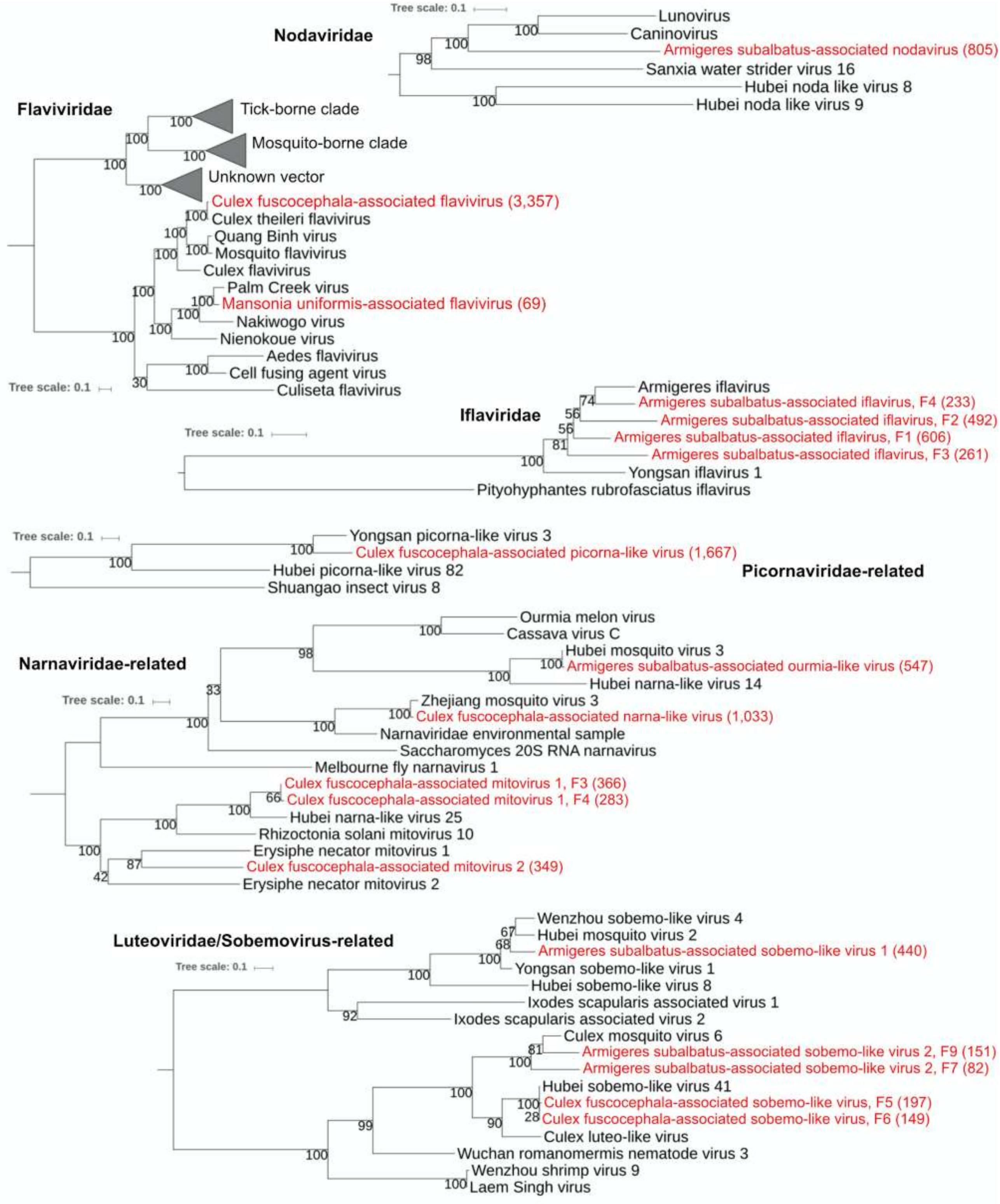
Tree scale: 0.1
Nodaviridae
Lunovirus
Caninovirus
Armigeres subalbatus-associated nodavirus (805)
Sanxia water strider virus 16
Hubei noda like virus 8
Hubei noda like virus 9
100
100
98
100
Flaviviridae
Tick-borne clade
Mosquito-borne clade
Unknown vector
Culex fuscocephala-associated flavivirus (3,357)
Culex theileri flavivirus
Quang Binh virus
Mosquito flavivirus
Culex flavivirus
Palm Creek virus
Mansonia uniformis-associated flavivirus (69)
Nakiwogo virus
Nienokoue virus
Aedes flavivirus
Cell fusing agent virus
Culiseta flavivirus
100
100
100
100
100
100
100
100
100
100
100
100
100
100
100
30
Tree scale: 0.1
Iflaviridae
Tree scale: 0.1
Armigeres iflavirus
Armigeres subalbatus-associated iflavirus, F4 (233)
Armigeres subalbatus-associated iflavirus, F2 (492)
Armigeres subalbatus-associated iflavirus, F1 (606)
Armigeres subalbatus-associated iflavirus, F3 (261)
Yongsan iflavirus 1
Pityohyphantes rubrofasciatus iflavirus
74
56
56
81
100
Tree scale: 0.1
Yongsan picorna-like virus 3
Culex fuscocephala-associated picorna-like virus (1,667)
Hubei picorna-like virus 82
Shuangao insect virus 8
100
100
Picornaviridae-related
Narnaviridae-related
Tree scale: 0.1
Ourmia melon virus
Cassava virus C
Hubei mosquito virus 3
Armigeres subalbatus-associated ourmia-like virus (547)
Hubei narna-like virus 14
Zhejiang mosquito virus 3
Culex fuscocephala-associated narna-like virus (1,033)
Narnaviridae environmental sample
Saccharomyces 20S RNA narnavirus
Melbourne fly narnavirus 1
Culex fuscocephala-associated mitovirus 1, F3 (366)
Culex fuscocephala-associated mitovirus 1, F4 (283)
Hubei narna-like virus 25
Rhizoctonia solani mitovirus 10
Erysiphe necator mitovirus 1
Culex fuscocephala-associated mitovirus 2 (349)
Erysiphe necator mitovirus 2
100
98
100
100
33
100
100
100
66
100
100
100
87
42
Luteoviridae/Sobemovirus-related
Tree scale: 0.1
Wenzhou sobemo-like virus 4
Hubei mosquito virus 2
Armigeres subalbatus-associated sobemo-like virus 1 (440)
Yongsan sobemo-like virus 1
Hubei sobemo-like virus 8
Ixodes scapularis associated virus 1
Ixodes scapularis associated virus 2
Culex mosquito virus 6
Armigeres subalbatus-associated sobemo-like virus 2, F9 (151)
Armigeres subalbatus-associated sobemo-like virus 2, F7 (82)
Hubei sobemo-like virus 41
Culex fuscocephala-associated sobemo-like virus, F5 (197)
Culex fuscocephala-associated sobemo-like virus, F6 (149)
Culex luteo-like virus
Wuchan romanomermis nematode virus 3
Wenzhou shrimp virus 9
Laem Singh virus
67
68
100
100
92
81
100
100
28
90
100
99
100
100

Figure 4. Phylogenetic relationships of the positive sense ssRNA viruses discovered in this study (labeled red) and other known viruses. The analyses were based on RNA-dependent RNA polymerase (RdRp) amino acid sequences, except for Flaviviridae and Iflaviridae trees, in which polyproteins were used. The numbers in parentheses are the lengths of the amino acid sequences.

Unclassified clade closely related to Picornaviridae. Sequences from a virus closely related to the family *Picornaviridae* was discovered in high numbers in the rural *Cx. fuscocephala* sample (1.89% of total reads). The

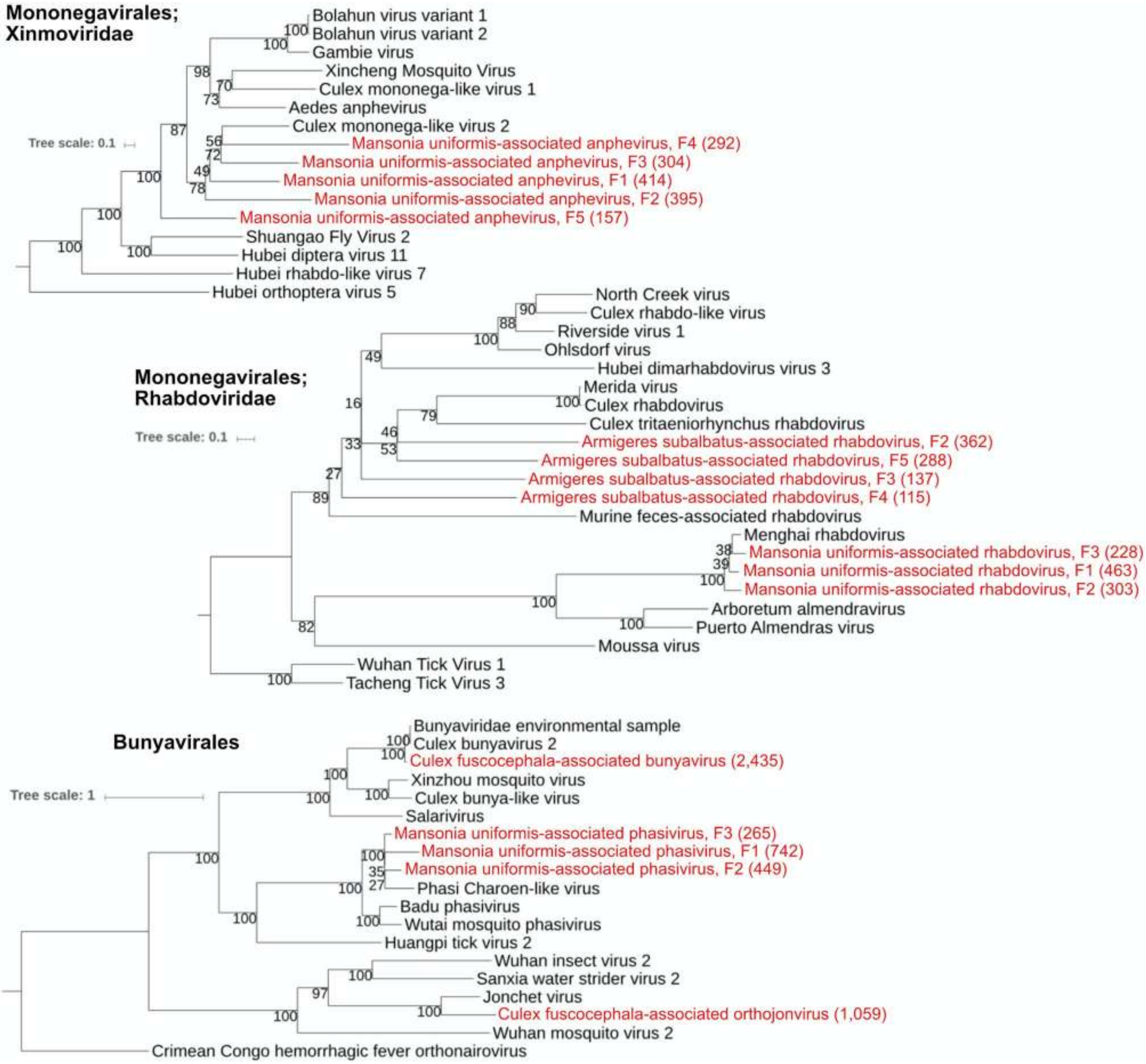
Mononegavirales;
Xinmoviridae
Tree scale: 0.1
Bolahun virus variant 1
Bolahun virus variant 2
Gambie virus
Xincheng Mosquito Virus
Culex mononega-like virus 1
Aedes anphevirus
Culex mononega-like virus 2
Mansonia uniformis-associated anphevirus, F4 (292)
Mansonia uniformis-associated anphevirus, F3 (304)
Mansonia uniformis-associated anphevirus, F1 (414)
Mansonia uniformis-associated anphevirus, F2 (395)
Mansonia uniformis-associated anphevirus, F5 (157)
Shuangao Fly Virus 2
Hubei diptera virus 11
Hubei rhabdo-like virus 7
Hubei orthoptera virus 5
100
100
98
70
73
87
56
72
49
78
100
100
100
100
Mononegavirales;
Rhabdoviridae
Tree scale: 0.1
North Creek virus
Culex rhabdo-like virus
Riverside virus 1
Ohlsdorf virus
Hubei dimarhabdovirus virus 3
Merida virus
Culex rhabdovirus
Culex tritaeniorhynchus rhabdovirus
Armigeres subalbatus-associated rhabdovirus, F2 (362)
Armigeres subalbatus-associated rhabdovirus, F5 (288)
Armigeres subalbatus-associated rhabdovirus, F3 (137)
Armigeres subalbatus-associated rhabdovirus, F4 (115)
Murine feces-associated rhabdovirus
Menghai rhabdovirus
Mansonia uniformis-associated rhabdovirus, F3 (228)
Mansonia uniformis-associated rhabdovirus, F1 (463)
Mansonia uniformis-associated rhabdovirus, F2 (303)
Arboretum almendravirus
Puerto Almendras virus
Moussa virus
Wuhan Tick Virus 1
Tacheng Tick Virus 3
90
88
100
49
100
79
16
46
33
53
27
89
38
39
100
100
100
82
100
Bunyavirales
Tree scale: 1
Bunyaviridae environmental sample
Culex bunyavirus 2
Culex fuscocephala-associated bunyavirus (2,435)
Xinzhou mosquito virus
Culex bunya-like virus
Salarivirus
Mansonia uniformis-associated phasivirus, F3 (265)
Mansonia uniformis-associated phasivirus, F1 (742)
Mansonia uniformis-associated phasivirus, F2 (449)
Phasi Charoen-like virus
Badu phasivirus
Wutai mosquito phasivirus
Huangpi tick virus 2
Wuhan insect virus 2
Sanxia water strider virus 2
Jonchet virus
Culex fuscocephala-associated orthojonvirus (1,059)
Wuhan mosquito virus 2
Crimean Congo hemorrhagic fever orthonairovirus
100
100
100
100
100
100
100
35
27
100
100
100
100
97
100
100

Figure 5. Phylogenetic relationships of the negative sense ssRNA viruses discovered in this study (labeled red) and other known viruses. The analyses were based on RNA-dependent RNA polymerase (RdRp) amino acid sequences. The numbers in parentheses are the lengths of the amino acid sequences.

assembled partial genome of this virus, provisionally called Culex fuscocephala-associated picorna-like virus (Fig. 4), was most similar to the genome of Yongsan picorna-like virus 3 (70% amino acid identity) found in *Aedes vexans* from South Korea[58]. *Picornaviridae* is a large and diverse family containing viruses with positive sense ssRNA genomes ranging from 6.7 to 10.1 kb[58]. Recent metagenomic studies have found a great number of novel picorna-like viruses in invertebrates including mosquitoes[6,14]. It has been suggested only recently that a new family, *Polycipiviridae*, assigned to the order *Picornavirales*, should be adopted to include insect-associated picorna-like viruses[59].

Order Bunyavirales. Order *Bunyavirales* includes viruses with segmented, negative-sense or ambisense single-stranded RNA (ssRNA) genomes distributed among 9 families (International Committee on Taxonomy of Virus, ICTV; Taxonomic Proposal 2016.030a-vM). The genomes consist of two to six segments encoding structural proteins, and one or more non-structural proteins[60]. Sequence analysis of the RdRps indicated that the *Bunyavirales* contigs in our samples could be assembled into at least three distinct clades (Fig. 5). The first putative virus, provisionally called Culex fuscocephala-associated bunyavirus (Fig. 5), found in all *Cx. fuscocephala* samples, fell in an unclassified clade. This virus was similar to those found in *Cx. pipiens* collected in California (~93% amino acid identity)[8,12]. The second putative virus belongs to the genus *Phasivirus* in the family *Phenuiviridae* (Fig. 5). This putative virus, provisionally called Mansonia uniformis-associated phasivirus, was found in the suburb and

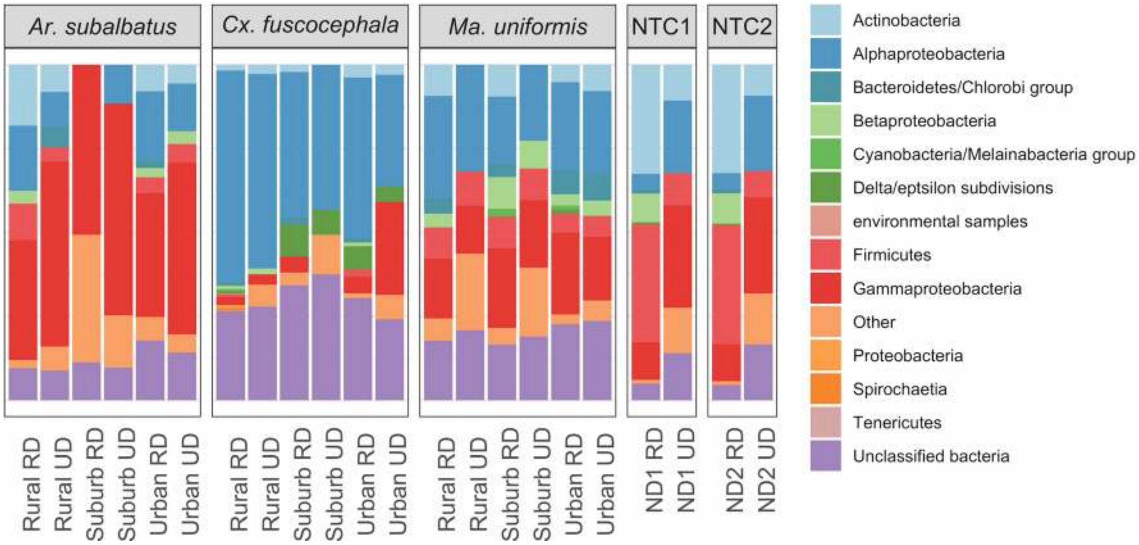

Figure 6. The proportion of bacteria groups found in all mosquito samples and NTCs. RD = rRNA depleted, UD = rRNA not depleted. NTC1 and NTC2 refer to no-template controls for separate sequencing runs (Supplementary Table 2).

urban *Ma. uniformis* samples. It was similar to Phasi Charoen-like phasivirus found in *Ae. aegypti* in Thailand and China (amino acid identity varies from 58 to 85% depends on the fragment)[13]. The third putative virus, provisionally called Culex fuscocephala-associated orthojonvirus (Fig. 5), was found in all three *Cx. fuscocephala* samples. The closest relative (32–53% amino acid identity) was the Jonchet virus (family *Jonviridae*) documented from mosquitoes in Côte d'Ivoire[61].

Order Mononegavirales: Rhabdoviridae and Xinmoviridae. The family *Rhabdoviridae* includes viruses with genomes of 10.8–16.1 kb, usually with 5 genes encoding structural proteins[62]. The family is ecologically diverse with members infecting plants, fish, reptiles, birds, and mammals as well as arthropod hosts or vectors[62]. Our phylogenetic analysis of the L proteins (RdRp) suggested that there were at least two putative rhabdoviruses in our mosquito samples. The first virus, provisionally called Mansonia uniformis-associated rhabdovirus (Fig. 5), was found only in the suburban *Ma. uniformis*. This virus formed a clade with Menghai rhabdovirus (77–96% amino acid similarity), which was recently isolated from *Ae. albopictus* from China[63]. Another set of rhabdovirus L gene fragments was recovered in all three *Ar. subalbatus* pools (Armigeres subalbatus-associated rhabdovirus). These fragments showed similarity to known virus sequences (33–50% amino acid similarity) and the closest similarity was to Culex rhabdovirus. Because of their small sizes, the phylogenetic analysis of these L fragments did not result in a strongly supported clade and we cannot conclude whether these fragments belong to the same virus (Fig. 5).

We also recovered contigs similar to a virus in the newly proposed genus *Anphevirus* of the family *Xinmoviridae* (ICTV report 2017.016 M.A.v3), which consists of multiple insect viruses. The assembled genomic fragments found in our study (Fig. 5) were closely related to *Culex* mononega-like virus 2 (amino acid sequence similarity ranges from 33 to 50%). This virus, provisionally called Mansonia uniformis-associated anphevirus (Fig. 5), was found mostly in the rural *Ma. Uniformis* sample.

Phylogenetic relationship of viruses across study sites.

In order to determine genetic relationships within those viruses that occurred across study sites, the partial genomes of each virus found in all three sites were subjected to phylogenetic analysis where genome coverage was sufficient. Of the three viruses that were found in high numbers in all sites, we document limited genetic divergence across sites of collection (Supplementary Figure 4).

Non-viral reads.

The majority of non-viral reads, in both the RD and UD samples, were classified to phylum Arthropoda, averaging 79.80% and 81.08%, respectively (Supplementary Table 5). The average number of reads that could not be classified across all samples was 7.17% and 2.73% for RD and UD samples, respectively. The average numbers of non-virus reads classified to the Bacteria domain were 2.23% and 0.48% for RD and UD samples, respectively. The proportion of bacteria groups in all samples is shown in Fig. 6. The dominant bacterial group differed in each mosquito host species: in *Ar. subalbatus* the majority were in the class Gammaproteobacteria (average = 48.73% of all bacterial reads) while the majority of bacteria found in *Cx. fuscocephala* were in the class Alphaproteobacteria (48.19%). The composition of the bacteria community in *Ma. uniformis* samples was more evenly distributed, dividedly between Alphaproteopbacteria (26.12%) and Gammaproteobacteria (19.87%). Other groups of bacteria found in mosquito samples included: phylum Actinobacteria, averaging 3.58% of all bacterial reads across all samples; phylum Firmicutes, 2.39%; class Betaproteobacteria, 1.68%; Bacteroidetes-Chlorobi group, 1.54%; and Delta-Epsilon Proteobacteria subdivision, 1.53%.

In terms of individual bacteria of note, *Wolbachia* (Alphaproteobacteria), a genus of common obligate intracellular bacteria found in a wide range of invertebrate taxa[64], was found in all samples as indicated by 16 s rRNA contigs recovered. We also found rRNA reads of bacteria relating to common house flies and bee gut-associated *Apibacter* bacteria in all *Cx. fuscocephala* samples, but not other mosquito species (Supplementary Figure 5). Also restricted to the three *Cx. fuscocephala* samples were bacteria similar to *Helicobacter sp.* of the Epsilon-proteobacteria, and bacteria similar to the bee-associated *Frischella sp.* and *Gilliamella sp.* of the Gammaproteobacteria.

Multiple groups of fungi were detected (Supplementary Table 5), averaging 0.56% and 0.21% of all non-viral reads in RD and UD samples, respectively. Common divisions of fungal reads found in all samples were Ascomycota and Basidiomycota. The highest percentage of fungal reads was from the pool of *Cx. fuscocephala* collected from the rural habitat. The majority of fungi in this sample were classified into phylum Zoopagomycota (division Zygomycota), at 2.56% and 0.88% or all in RD sample and UD sample, respectively.

Other organisms within our samples included trypanosomatids, alveolates, plants, algae, platyhelminthes, nematodes, annelids, mollusks, and vertebrates (Supplementary Table 4). Initial analysis of ribosomal RNA reads suggested there existed multiple poorly known species of mosquito-associated eukaryotic microorganisms in our samples. For example, partial trypanosomatid rRNA reads, related to trypanosomatids of multiple insects, were found in all three mosquito species. In addition, partial 18 s rRNA reads similar to *Paratrypanosoma confusum*, a trypanosomatid species recently found in *Cx. pipiens* guts[65], were detected in *Cx. fuscocephala* samples.

Multiple samples collected from rural and urban sites harbored 18 s rRNA reads similar to nematode species in the Mermithidae, and Setariidae families. Of note is a partial 28 s rRNA read (274 bp) found in the rural *Ma. uniformis* sample that is 100% similar to *Setaria digitata*, a mosquito-borne zoonotic nematode species infecting ungulates in Asia[66]. *Ma. uniformis* sample collected in suburban site harbored reads likely belong to *Brugia malayi* (99–100% similarity, 215 and 431 bp).

Several rRNA reads were recovered in the urban *Cx. fuscocephala* sample related to trematode of the order Plagiorchiida such as *Collyriclum faba*, *Paralechithodendrium longiforme*, and *Lecithodendrium linstowi*. *C. faba* is a bird parasite that has aquatic gastropods as the first intermediate host and mayflies as the second intermediate host[67]. On the other hand, *P. longiforme* and *L. linstowi* are bat parasites with unknown second intermediate hosts[68,69].

Lastly, 18 s rRNA reads of *Ascogregarina sp.*, a mosquito-specific apicomplexan parasite, were found in rural and urban *Ar. subalbatus* samples and were most similar to *A. armigerei* (99.4%, 1061 bp). The genus *Ascogregarina* parasitizes mosquitoes and sandflies, and is relatively host specific: *A. taiwanensis* infects *Ae. albopictus*, *A. culicis* infects *Ae. aegypti*, and *A. armigerei* infects *Ar. subalbatus*[70,71].

Discussion

In this study, we used a high throughput RNA sequencing metagenomic approach to characterize the mosquito-associated virome and microbiome. We identified at least 21 putative RNA viruses in three vector species: *Ar. subalbatus*, *Cx. fuscocephala*, and *Ma. uniformis*. We chose the three mosquito species because of their abundance and potential vector status in central Thailand[27]. *Ar. subalbatus* transmits Japanese encephalitis virus, *Wuchereria bancrofti*, and dog heartworm *Dirofilaria immitis*[72]. Their larvae have been found in nutrient enriched water including septic tanks and polluted stagnant water[73] as well as less enriched bamboo stumps, artificial containers, and tree holes[74]. *Cx. fuscocephala*, often associated with flooded rice field and agricultural lands[75–77], is among multiple *Culex* mosquito species that transmit Japanese encephalitis[78,79]. *Mansonia uniformis* is a vector of *Wuchereria malayi*. They are associated with vegetated water habitat such as ponds and canals with floating vegetation, rice fields, and swamp forest[80–82].

According to the Principle Component Analysis, we found that the pattern of viruses in our mosquito samples was defined primarily by the host species rather than by geographical location. Viruses are obligate intracellular parasites that can only function when inside host cells, necessitating specific host cell recognition, entry, and manipulation of multiple host cell molecular mechanisms in order to complete their replication cycles. These properties limit viruses' biological host ranges. It is not surprising that we observe this host species-specific rather than location-specific pattern. This is in contrast to the bacterial community, where grouping patterns did not show association with host or location. Even though the full factorial design in our study allowed us to investigate the relative importance of ecological factors in structuring mosquito virome, the small sample size limits our ability to generalize this finding.

On the other hand, the same virus species found in the same mosquito species but from different collection sites showed limited genetic divergence (Supplementary Figure 4), suggesting that the viruses are shared between mosquito populations, at least within the geographical scale of this study (the three study sites are within 20 km apart). Other studies comparing mosquito-specific RNA viruses isolated from wider geographical areas found that they showed genetic similarity across continents[83,84]. The results from our study lead to important implications for the surveillance of emerging viral diseases: in order to capture as much viral diversity as possible, monitoring efforts should sample a wide diversity of mosquito species, rather than focus on a narrow set of mosquito species albeit over a larger geographic landscape.

We also tested a host rRNA depletion method which could reduce the cost of sequencing without the need to pool too many individuals into a single sample. Our analysis indicated that we successfully depleted rRNA of the mosquito's SSU but not of the LSU. Even then, the host rRNA depletion as adopted in our study significantly increased the percentage of virus reads in the sample as well as the coverage for virus contigs. However, our caution is that depletion was time consuming such that the same gains in virus detection could have been achieved for less cost in time with an additional sequencing run, especially when the cost of sequencing may continue to come down in the future.

We found multiple putative viruses associated with certain mosquito species, suggesting that they are likely mosquito-specific. For example, Armigeres subalbatus-associated ourmia-like virus was classified to a novel insect virus clade closely related to plant viruses in the genus *Ourmiavirus*. It was found in high numbers in all *Ar. subalbatus* samples but not in the other two mosquito species. A recent NGS-based study[6] found multiple invertebrate viruses in well-defined clades distinct but related to both *Ourmiavirus* and *Narnavirus*[55]. According to a recently published ICTV report, there could be a future establishment of a new virus family "*Ourmiaviridae*"[55]. The family would comprise a genus for ourmia-like mycoviruses, an existing genus containing plant viruses, and a genus for ourmia-like viruses isolated from invertebrates (which may contain the ourmia-like virus in our sample)[55]. Another putative virus, Culex fuscocephala-associated narna-like virus, was found in *Cx. fuscocephala* and not the other two mosquito species. This putative virus formed a monophyletic group with a virus found in *Cx. pipiens*[12] and other arthropods[6]. Although this group of viruses was closely related to mycoviruses, it likely contains a group of novel insect-specific viruses.

Another insect-specific virus, Ma. uniformis-associated anphevirus, was recovered mostly in the rural sample of *Ma. uniformis,* suggestive of habitat restriction. Interestingly, the newly proposed genus *Anphevirus* consists of multiple insect viruses including those isolated from *Culex*[85], *Anopheles*[5], and *Aedes* mosquitoes[84]. A recent study showed that Aedes anphevirus may reduce DENV replication in cell culture[84]. Our phylogenetic analysis indicated that Ma. uniformis-associated anphevirus is separated from other mosquito anpheviruses (Fig. 4). Thus, we speculate that Ma. uniformis-associated anphevirus could be a *Mansonia*-specific clade.

It is important to note that the virus names given in our manuscript are only provisional and not species names. The ICTV sets different species demarcation criteria that vary depending on virus groups; and the criteria often include information other than virus genetic sequences. Lacking relevant information for species assignment, we cannot classify the viruses found in our study into the same species of existing virus, nor can we establish a new species. Ideally, future studies should be done to assign each of these viruses to species.

Our non-targeted sequencing approach reveals multiple interesting mosquito-associated non-viral microorganisms such as a Bacteroidetes species (found in *Cx. fuscocephala* samples; Supplementary Figure 5) similar to the bee-associated *Apibacter sp.*, and a not yet described digenean trematode species in which mosquitoes may serve as the second intermediate host (Supplementary Figure 6). The identification of these non-viral organisms is important in their own rights but it can also give further insight into potential hosts of viruses that might be found in the same mosquito pools. Even though the viruses were discovered in mosquitoes, they may infect other mosquito endosymbionts such as bacteria, fungi, and parasites[86], or ingested organic material such as plants and algae.

We have included no-template controls (NTCs) that were processed alongside mosquito samples from extraction to sequencing. Due to the high sensitivity of NGS and the well-documented risk of contamination in microbiome studies, control samples should always be included to assess contamination[87,88] that can originate from the environment (e.g. reagents and kits, plastic consumables, dust, and humans), and cross-contamination between samples[87], occurring either during sample processing (e.g. aerosol contamination, barcode cross-contamination), or sequencing (e.g. barcode sequencing errors and 'index hopping')[87]. Despite its importance, not all NGS-based metagenomic studies include controls.

We detected reads belonging to multiple organisms in the NTCs. The proportion of reads in NTCs classified to each taxon is shown in Supplementary Table 5. We observed that abundant taxa were more likely to be recovered in the NTCs. Thus, we suspect limited cross-contamination between samples. Because NTCs have no starting material, even trace amounts of contaminated DNA could be picked up by NGS. On the contrary, mosquito samples contain large amounts of template DNA reads, and only a few contaminated reads would be sequenced. Other types of controls, such as mock community controls should be adopted in future studies along with NTCs. This type of control not only allows a more realistically detection of contamination, but also the quantification of sequencing error and other bias introduced during the library preparation processes[89].

Endogenous viral elements (EVEs) are defined as viral DNA sequences deriving from retroviruses, DNA viruses, or RNA viruses that present within the genomes of non-viral organisms, including mosquitoes[90]. Unfortunately, we cannot definitively determine whether the contigs identified in this study originate from viruses or EVEs because the genomes of the three mosquito species are currently unavailable, and because EVEs possess non-specific characteristics that cannot be used for their identification. However, we expect that the RNA extraction protocol adopted in our study likely limited the inclusion of EVEs, and the potentially expressed small EVEs (e.g. EVE-derived piRNAs) in our purified RNA samples.

Mosquito-borne disease emergence is thought to be associated with socio-economic and ecological factors arising from urbanization and poverty. In this study, we characterized mosquito-associated viromes and microbiomes across a landscape gradient transformed by human activities and urbanization. In addition, the three vector species differ in their biology and ecology. We found that the mosquito-associated virome component was defined primarily by the host species rather than by the geographical locations or habitats, and multiple viruses spread between sample sites and mosquito populations at the scale of this study. Similar to *Wolbachia*, we know that elements of the vector virome and microbiome can impact vector fitness and competence: we recovered a virus closely related to Aedes anphevirus previously reported to reduce vector competence to dengue[84]. Thus, this study may help inform future use of insect-specific virus (ISVs) in the control of vectors and diseases. Because we found that mosquito species determine virome composition more than habitat, urbanization is likely to have the greatest influence on the distribution of microbes by facilitating invasive vector species. This finding is supported by our previous study indicating that under urbanization mosquito communities become less diverse and skewed towards invasive mosquito species[27]. To test the influence of habitat alone on mosquito-associated virome, future studies should focus on one mosquito species at a time, and sample individually and more intensively across landscape types.

Data availability

The raw sequence reads generated in this study are available at the NCBI Sequence Read Archive (SRA) database under BioProject PRJNA716099; BioSamples SAMN18394705-SAMN18394726. All virus contigs and some 16s and 18s rRNA contigs generated in this study have been deposited in GenBank under accession numbers: MW854367-MW854382, MW858073-MW858116, MW874579-MW874588, and MW879746-MW879751.

Received: 25 August 2020; Accepted: 18 March 2021
Published online: 19 April 2021

References

1. Cadwell, K. The virome in host health and disease. *Immunity* **42**, 805–813 (2015).
2. Paez-Espino, D. *et al.* Uncovering earth's virome. *Nature* https://doi.org/10.1038/nature19094 (2016).
3. Shi, M. *et al.* The evolutionary history of vertebrate RNA viruses. *Nature* **556**, 197–202 (2018).
4. Dolja, V. V. & Koonin, E. V. Metagenomics reshapes the concepts of RNA virus evolution by revealing extensive horizontal virus transfer. *Virus Res.* **244**, 36–52 (2018).
5. Li, C.-X. *et al.* Unprecedented genomic diversity of RNA viruses in arthropods reveals the ancestry of negative-sense RNA viruses. *Elife* **4**, e05378 (2015).
6. Shi, M. *et al.* Redefining the invertebrate RNA virosphere. *Nature* **540**, 539–543 (2016).
7. Atoni, E. *et al.* Metagenomic Virome Analysis of *Culex* Mosquitoes from Kenya and China. *Viruses* **10**, 30 (2018).
8. Sadeghi, M. *et al.* Virome of > 12 thousand *Culex* mosquitoes from throughout California. *Virology* **523**, 74–88 (2018).
9. Zakrzewski, M. *et al.* Mapping the virome in wild-caught Aedes aegypti from Cairns and Bangkok. *Nat. Publ. Group* https://doi.org/10.1038/s41598-018-22945-y (2018).
10. Xia, H. *et al.* Comparative metagenomic profiling of viromes associated with four common mosquito species in China. *Virol. Sin.* **33**, 59–66 (2018).
11. Frey, K. G. *et al.* Bioinformatic characterization of mosquito viromes within the eastern United States and Puerto Rico: ciscovery of novel viruses. *Evolut. Bioinform.* **12s2**, EBO.S38518 (2016).
12. Chandler, J. A., Liu, R. M. & Bennett, S. N. RNA shotgun metagenomic sequencing of northern California (USA) mosquitoes uncovers viruses, bacteria, and fungi. *Front. Microbiol.* **06**, 403 (2015).
13. Chandler, J. A. *et al.* Metagenomic shotgun sequencing of a Bunyavirus in wild-caught *Aedes aegypti* from Thailand informs the evolutionary and genomic history of the Phleboviruses. *Virology* **464–465**, 312–319 (2014).
14. Cholleti, H. *et al.* Discovery of novel viruses in mosquitoes from the Zambezi valley of Mozambique. *PLoS ONE* **11**, e0162751 (2016).
15. Scarpassa, V. M. *et al.* An insight into the sialotranscriptome and virome of Amazonian anophelines. *BMC Genom.* https://doi.org/10.1186/s12864-019-5545-0 (2019).
16. Hameed, M. *et al.* A viral metagenomic analysis reveals rich viral abundance and diversity in mosquitoes from pig farms. *Transbound. Emerg. Dis.* **67**, 328–343 (2019).
17. Fauver, J. R. *et al.* West African *Anopheles gambiae* mosquitoes harbor a taxonomically diverse virome including new insect-speci. *Virology* **498**, 288–299 (2016).
18. Xiao, P. *et al.* Metagenomic sequencing from mosquitoes in China reveals a variety of insect and human viruses. *Front. Cell. Infect. Microbiol.* **8**, 131–211 (2018).
19. Shi, C. *et al.* Stable distinct core eukaryotic viromes in different mosquito species from Guadeloupe, using single mosquito viral metagenomics. *Microbiome* https://doi.org/10.1186/s40168-019-0734-2 (2019).
20. World Health Organization. *A global brief on vector-borne diseases*. (2014).
21. Vasilakis, N. & Tesh, R. B. Insect-specific viruses and their potential impact on arbovirus transmission. *Curr. Opin. Virol.* **15**, 69–74 (2015).
22. Goenaga, S. *et al.* Potential for co-infection of a mosquito-specific flavivirus, Nhumirim virus, to block West Nile virus transmission in mosquitoes. *Viruses* **7**, 5801–5812 (2015).
23. Hall-Mendelin, S. *et al.* The insect-specific Palm Creek virus modulates West Nile virus infection in and transmission by Australian mosquitoes. *Parasit. Vectors* **9**, 414 (2016).
24. Colmant, A. M. G. *et al.* The recently identified flavivirus Bamaga virus is transmitted horizontally by *Culex* mosquitoes and interferes with West Nile virus replication in vitro and transmission in vivo. *PLoS Negl. Trop. Dis.* **12**, e0006886 (2018).
25. Romo, H., Kenney, J. L., Blitvich, B. J. & Brault, A. C. Restriction of Zika virus infection and transmission in *Aedes aegypti* mediated by an insect-specific flavivirus. *Emerg. Microbes Infect* **7**, 181 (2018).
26. Schultz, M. J., Frydman, H. M. & Connor, J. H. Dual Insect specific virus infection limits Arbovirus replication in *Aedes* mosquito cells. *Virology* **518**, 406–413 (2018).
27. Thongsripong, P. *et al.* Mosquito vector diversity across habitats in central Thailand endemic for dengue and other arthropod-borne diseases. *PLoS Negl. Trop. Dis.* **7**, e2507 (2013).
28. Kukutla, P., Steritz, M. & Xu, J. Depletion of ribosomal RNA for mosquito gut metagenomic RNA-seq. *JoVE* https://doi.org/10.3791/50093 (2013).
29. Rattanarithikul, R., Harrison, B. A. & Panthusiri, P. Coleman RE (2005) Illustrated keys to the mosquitoes of Thailand I. Background; geographic distribution; lists of genera, subgenera, and species; and a key to the genera. *Southeast Asian J. Trop. Med. Public Health* **36 Suppl 1**, 1–80 (2005).
30. Rattanarithikul, R. *et al.* Illustrated keys to the mosquitoes of Thailand. II. Genera *Culex* and *Lutzia*. *Southeast Asian J. Trop. Med. Public Health* **36 Suppl 2**, 1–97 (2005).
31. Rattanarithikul, R., Harrison, B. A., Panthusiri, P., Peyton, E. L. & Coleman, R. E. Illustrated keys to the mosquitoes of Thailand III. Genera *Aedeomyia, Ficalbia, Mimomyia, Hodgesia, Coquillettidia, Mansonia*, and *Uranotaenia*. *Southeast Asian J. Trop. Med. Public Health* **37 Suppl 1**, 1–85 (2006).
32. Rattanarithikul, R., Harrison, B. A., Harbach, R. E., Panthusiri, P. & Coleman, R. E. Illustrated keys to the mosquitoes of Thailand. IV. *Anopheles*. *Southeast Asian J. Trop. Med. Public Health* **37 Suppl 2**, 1–128 (2006).
33. Rattanarithikul, R., Harbach, R. E., Harrison, B. A., Panthusiri, P. & Coleman, R. E. Illustrated keys to the mosquitoes of Thailand V. Genera *Orthopodomyia, Kimia, Malaya, Topomyia, Tripteroides*, and *Toxorhynchites*. *Southeast Asian J. Trop. Med. Public Health* **38**, 1–65 (2007).
34. Rattanarithikul, R. *et al.* Illustrated keys to the mosquitoes of Thailand. VI. Tribe *Aedini*. *Southeast Asian J. Trop. Med. Public Health* **41 Suppl 1**, 1–225 (2010).
35. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
36. Haas, B. J. *et al.* De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* **8**, 1494–1512 (2013).

37. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).
38. Katoh, K., Rozewicki, J. & Yamada, K. D. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Brief Bioinform.* https://doi.org/10.1093/bib/bbx108 (2017).
39. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
40. Darriba, D., Taboada, G. L., Doallo, R. & Posada, D. ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics* **27**, 1164–1165 (2011).
41. Kozlov, A. M., Darriba, D., Flouri, T., Morel, B. & Stamatakis, A. RAxML-NG: A fast, scalable, and user-friendly tool for maximum likelihood phylogenetic inference. *bioRxiv* 447110 (2018).
42. Miller, M. A., Pfeiffer, W. & Schwartz, T. Creating the CIPRES science gateway for interface of large phylogenetic trees. 1–8 (2010).
43. Letunic, I. & Bork, P. Interactive tree of life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res.* https://doi.org/10.1093/nar/gkz239 (2019).
44. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357–359 (2012).
45. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
46. Ryan, F. P. Human endogenous retroviruses in multiple sclerosis: potential for novel neuro-pharmacological research. *Curr. Neuropharmacol.* **9**, 360–369 (2011).
47. Wood, D. E. & Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol* **15**, R46 (2014).
48. Kopylova, E., Noe, L. & Touzet, H. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics* **28**, 3211–3217 (2012).
49. Simmonds, P. *et al.* ICTV virus taxonomy profile: Flaviviridae. *J. Gen. Virol.* **98**, 2–3 (2017).
50. Kyaw, A. K. *et al.* Virus research. *Virus Res.* **247**, 120–124 (2018).
51. Valles, S. M. *et al.* ICTV virus taxonomy profile: Iflaviridae. *J. Gen. Virol.* **98**, 527–528 (2017).
52. Kobayashi, D. *et al.* Isolation and characterization of a new iflavirus from *Armigeres spp*. mosquitoes in the Philippines. *J. Gen. Virol.* **98**, 2876–2881 (2017).
53. Viruses, I. C. O. T. O., King, A. M. Q., Adams, M. J., Lefkowitz, E. & Carstens, E. B. *Virus Taxonomy: Ninth Report of the International Committee on Taxonomy of Viruses* (Elsevier, Amsterdam, 2011).
54. Hillman, B. I. & Cai, G. The family narnaviridae: Simplest of RNA viruses. *Adv. Virus Res.* **86**, 149–176 (2013).
55. Turina, M. *et al.* ICTV virus taxonomy profile: Ourmiavirus. *J. Gen. Virol.* **98**, 129–130 (2017).
56. Yong, C. Y., Yeap, S. K., Omar, A. R. & Tan, W. S. Advances in the study of nodavirus. *PeerJ* **5**, e3841 (2017).
57. Sahul Hameed, A. S. *et al.* ICTV virus taxonomy profile: Nodaviridae. *J. Gen. Virol.* **100**, 3–4 (2019).
58. Sanborn, M. *et al.* Metagenomic analysis reveals three novel and prevalent mosquito biruses from a single pool of *Aedes vexans* nipponii collected in the Republic of Korea. *Viruses* **11**, 222 (2019).
59. Olendraite, I. *et al.* ICTV virus taxonomy profile: Polycipiviridae. *J. Gen. Virol.* **100**, 554–555 (2019).
60. Wichgers Schreur, P. J., Kormelink, R. & Kortekaas, J. Genome packaging of the Bunyavirales. *Curr. Opin. Virol.* **33**, 151–155 (2018).
61. Marklewitz, M., Zirkel, F., Kurth, A., Drosten, C. & Junglen, S. Evolutionary and phenotypic analysis of live virus isolates suggests arthropod origin of a pathogenic RNA virus family. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 7536–7541 (2015).
62. Walker, P. J. *et al.* ICTV virus taxonomy profile: Rhabdoviridae. *J. Gen. Virol.* **99**, 447–448 (2018).
63. Sun, Q. *et al.* Complete genome sequence of Menghai rhabdovirus, a novel mosquito-borne rhabdovirus from China. *Adv. Virol.* **162**, 1103–1106 (2017).
64. Hilgenboecker, K., Hammerstein, P., Schlattmann, P., Telschow, A. & Werren, J. H. How many species are infected with *Wolbachia*? A statistical analysis of current data. *FEMS Microbiol Lett* **281**, 215–220 (2008).
65. Flegontov, P. *et al.* *Paratrypanosoma* is a novel early-branching trypanosomatid. *Curr Biol* **23**, 1787–1793 (2013).
66. Kaur, D. *et al.* Occurrence of *Setaria digitata* in a cow. *J Parasit Dis* **39**, 477–478 (2015).
67. Heneberg, P. *et al.* Intermediate hosts of the trematode *Collyriclum faba* (Plagiochiida: Collyriclidae) identified by an integrated morphological and genetic approach. *Parasit. Vectors* **8**, 85 (2015).
68. Enabulele, E. E., Lawton, S. P., Walker, A. J. & Kirk, R. S. Molecular and morphological characterization of the cercariae of *Lecithodendrium linstowi* (Dollfus, 1931), a trematode of bats, and incrimination of the first intermediate snail host *Radix balthica*. *Parasitology* **145**, 307–312 (2018).
69. Greiman, S. E. *et al.* Real-time PCR detection and phylogenetic relationships of *Neorickettsia spp*. in digeneans from Egypt, Philippines, Thailand, Vietnam and the United States. *Parasitol. Int.* **66**, 1003–1007 (2017).
70. Lantova, L. & Volf, P. Mosquito and sand fly gregarines of the genus *Ascogregarina* and *Psychodiella* (Apicomplexa: Eugregarinorida, Aseptatorina)—Overview of their taxonomy, life cycle, host specificity and pathogenicity. *Infect. Genet. Evol.* **28**, 616–627 (2014).
71. Roychoudhury, S. *et al.* Comparison of the morphology of oocysts and the phylogenetic analysis of four *Ascogregarina* species (Eugregarinidae: Lecudinidae) as inferred from small subunit ribosomal DNA sequences. *Parasitol. Int.* **56**, 113–118 (2007).
72. Muslim, A., Fong, M.-Y., Mahmud, R., Lau, Y.-L. & Sivanandam, S. *Armigeres subalbatus* incriminated as a vector of zoonotic *Brugia pahangi* filariasis in suburban Kuala Lumpur Peninsular Malaysia. *Parasites Vectors* **6**, 219 (2013).
73. Hiscox, A. *et al.* *Armigeres subalbatus* colonization of damaged pit latrines: A nuisance and potential health risk to residents of resettlement villages in Laos. *Med. Vet. Entomol.* **30**, 95–100 (2016).
74. Chaves, L. F., Imanishi, N. & Hoshi, T. Population dynamics of *Armigeres subalbatus* (Diptera: Culicidae) across a temperate altitudinal gradient. *Bull. Entomol. Res.* **105**, 589–597 (2015).
75. Ohba, S.-Y., Van Soai, N., Van Anh, D. T., Nguyen, Y. T. & Takagi, M. Study of mosquito fauna in rice ecosystems around Hanoi, northern Vietnam. *Acta Trop.* **142**, 89–95 (2015).
76. Tsuda, Y., Takagi, M., Suwonkerd, W., Sugiyama, A. & Wada, Y. Comparisons of rice field mosquito (Diptera: Culicidae) abundance among areas with different agricultural practices in northern Thailand. *J. Med. Entom.* **35**, 845–848 (1998).
77. Ohba, S.-Y. *et al.* Mosquitoes and their potential predators in rice agroecosystems of the Mekong Delta, southern Vietnam. *J. Am. Mosq. Control Assoc.* **27**, 384–392 (2011).
78. Su, C.-L. *et al.* Molecular epidemiology of Japanese encephalitis virus in mosquitoes in Taiwan during 2005–2012. *PLoS Negl. Trop. Dis.* **8**, e3122 (2014).
79. Keiser, J. *et al.* Effect of irrigated rice agriculture on Japanese encephalitis, including challenges and opportunities for integrated vector management. *Acta Trop.* **95**, 40–57 (2005).
80. Apiwathnasorn, C., Samung, Y., Prummongkol, S., Asavanich, A. & Komalamisra, N. Surveys for natural host plants of *Mansonia* mosquitoes inhabiting Toh Daeng peat swamp forest, Narathiwat Province, Thailand. *Southeast Asian J. Trop. Med. Public Health* **37**, 279–282 (2006).
81. Surtees, G., Simpson, D. I. H., Bowen, E. T. W. & Grainger, W. E. Ricefield development and arbovirus epidemiology, Kano Plain, Kenya. *Trans. R. Soc. Trop. Med. Hyg.* **64**, 511–518 (1970).
82. Kwa, B. H. Environmental change, development and vector-borne disease: Malaysia's experience with filariasis, scrub typhus and dengue. *Environ. Dev. Sustain.* **10**, 209–217 (2008).
83. Cook, S. *et al.* Molecular evolution of the insect-specific flaviviruses. *J. Gen. Virol.* **93**, 223–234 (2012).
84. Parry, R. & Asgari, S. Aedes anphevirus: an insect-specific virus distributed worldwide in *Aedes aegypti* mosquitoes that has complex interplays with *Wolbachia* and Dengue Virus Infection in Cells. *J. Virol.* **92**, e00224–18 (2018).

85. Shi, M. *et al.* High-resolution metatranscriptomics reveals the ecological dynamics of mosquito-associated RNA viruses in western Australia. *J. Virol.* **91**, e00680–17 (2017).
86. Thongsripong, P. *et al.* Mosquito vector-associated microbiota: Metabarcoding bacteria and eukaryotic symbionts across habitat types in Thailand endemic for dengue and other arthropod-borne diseases. *Ecol. Evol.* **8**, 1352–1368 (2018).
87. Eisenhofer, R. *et al.* Contamination in low microbial biomass microbiome studies: Issues and recommendations. *Trends Microbiol.* **27**, 105–117 (2019).
88. Salter, S. J. *et al.* Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol.* **12**, 1–12 (2014).
89. Pollock, J., Glendinning, L., Wisedchanwet, T. & Watson, M. The madness of microbiome: attempting to find consensus 'best practice' for 16S microbiome studies. *Appl. Environ. Microbiol.* **84**, e02627–17 (2018).
90. Blair, C. D., Olson, K. E. & Bonizzoni M. The widespread occurrence and potential biological roles of endogenous viral elements in insect genomes. *Curr. Issues Mol. Biol.* **34**, 13–30 (2020).

Acknowledgements

We thank Amy Green for hard work in the field and mosquito identification; Rampa Rattanarithikul for advice and confirming mosquito identifications; Vinogiri Krishan for help in mosquito identification; Argon Steel for work in the field and technical assistance; the Nakhon Nayok Wildlife Conservation Development and Extension Center for providing housing and field station support; Shirley Luckhart of the University of California, Davis for providing Anopheles mosquitoes that the host ribosomal RNA depletion method was tested on; Jim Henderson for a generous support to the Center for Comparative Genomics; and Anna Sellas for laboratory assistance. This study was supported by the NSF IGERT program, University of Hawaii (0549514; PI: BAW), the California Academy of Sciences (SNB, PT, and DDK), NIH/NIAID U01 AI151788 (SNB & PT) and P01 AI106695 (SNB), the NIH/NIAID Pacific Southwest Regional Center of Excellence (PSWRCE U54AI065359, SNB), NIH/NCRR COBRE University of Hawaii (P20RR018727, SNB), NSF DBI grant 1427772 (DDK), and Mahidol University for financial support; the funding agencies had no involvement in study design, data analysis, or decision to publish.

Author contributions

P.T. participated in fieldwork study design, carried out fieldwork and mosquito identification, and performed data analysis. J.A.C. selected samples for sequencing and designed and performed all molecular laboratory experiments and N.G.S. P.K. participated in study design and coordinated and accommodated field experiments. B.A.W., D.D.K. and S.N.B. conceived the initial study and participated in study design and coordination. P.T. wrote the original manuscript. J.A.C., D.D.K., and S.N.B. provided substantial edits. All authors read, gave input, and approved of the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-87122-0.

Correspondence and requests for materials should be addressed to P.T.