Title: A reverse-transcription/RNase H based protocol for depletion of mosquito ribosomal RNA facilitates viral intrahost evolution analysis, transcriptomics and pathogen discovery.

Joseph R. Fauver[1§], Shamima Akter[3], Aldo Ivan Ortega Morales[4], William C. Black IV[1], Americo D. Rodriguez[2], Mark D. Stenglein[1], Gregory D. Ebel[1], James Weger-Lucarelli[1#]

[1]Department of Microbiology, Immunology and Pathology, College of Veterinary Medicine and Biomedical Sciences, Colorado State University, Fort Collins, CO 80523, USA

[2] Centro Regional de Investigación en Salud Publica, Instituto Nacional de Salud Publica, Tapachula, Chiapas, México

[3] Department of Biomedical Sciences and Pathobiology, Virginia Polytechnic Institute and State University 360 W Campus Drive, Blacksburg, Virginia, USA

[4] Departamento de Parasitología, Universidad Autónoma Agraria Antonio Narro, Torreón, Coahuila, México

Co-Corresponding Authors: James Weger-Lucarelli - weger@vt.edu, Gregory D. Ebel - Gregory.Ebel@colostate.edu.

Current Addresses:

[§] Department of Medicine, Infectious Diseases Division, Washington University School of Medicine, 660 S Euclid Ave, St. Louis, Missouri, USA

[#] Department of Biomedical Sciences and Pathobiology, Virginia Polytechnic Institute and State University 360 W Campus Drive, Blacksburg, Virginia, USA

Abstract

Studies aimed at identifying novel viral sequences or assessing intrahost viral variation require sufficient sequencing coverage to assemble contigs and make accurate variant calling at low frequencies. Many samples come from host tissues where ribosomal RNA represents more than 90% of total RNA preparations, making unbiased sequencing of viral samples inefficient and highly expensive, as many reads will be wasted on cellular RNAs. In the presence of this amount of ribosomal RNA, it is difficult to achieve sufficient sequencing depth to perform analyses such as variant calling, haplotype prediction, virus population analyses, virus discovery or transcriptomic profiling. Many methods for depleting unwanted RNA or enriching RNA of interest have been devised, including poly-A selection, RNase H based specific depletion, duplex-specific nuclease treatment and hybrid capture selection, among others. Although these methods can be efficient, they either cannot be used for some viruses (i.e. non-polyadenylated viruses), have been optimized for use in a single species, or have the potential to introduce bias. In this study, we describe a novel approach that uses an RNaseH possessing reverse transcriptase coupled with selective probes for ribosomal RNA designed to work broadly for three medically relevant mosquito genera; *Aedes*, *Anopheles,* and *Culex.* We demonstrate significant depletion of rRNA using multiple assessment techniques from a variety of sample types, including whole mosquitoes and mosquito midgut contents from FTA cards. To demonstrate the utility of our approach, we describe novel insect-specific virus genomes from numerous species of field collected mosquitoes that underwent rRNA depletion, thereby facilitating their detection. The protocol is straightforward, relatively low-cost and requires only common laboratory reagents and the design of several small oligonucleotides specific to the species of interest. This approach can be adapted for use with other organisms with relative ease, thus potentially aiding virus population genetics analyses, virus discovery and transcriptomic profiling in both laboratory and field samples.

Introduction

The past several decades have witnessed the emergence and expansion of viruses with increasing frequency (Jones et al., 2008). Several examples are H1N1 Influenza in 2009 (Otte et al., 2015), Chikungunya in 2006 (Tsetsarkin et al., 2007), Zika in 2013-14 (Aubry et al., 2017), West Nile in 1999 (Moudy et al., 2007), MERS (Forni et al., 2015) and others. Most, if not all of the emerging viruses that pose the greatest threat to human and animal health are RNA viruses. In fact, all 7 of the pathogens identified by the World Health Organization (WHO) in the 2018 annual review of the blueprint list of priority diseases as requiring urgent or serious research were RNA viruses, with the other being unknown pathogens (WHO, 2018). Due to the importance of RNA viruses, it is critical to be able to detect, identify and analyze these pathogens' genomes using novel high-throughput sequ byencing methods. However, total RNA preparations from complex lab and field samples typically contain extremely high levels of ribosomal RNA (rRNA), sometimes 80-90% of the total amount (Eun, 1996). Sequencing data mapping to rRNA is typically removed bioinformatically and therefore represents an economic waste and reduces the number of samples that can be tested in a given sequencing run. To increase the number of reads mapping to sequences of interest, several methods have been employed to either enrich; hybrid-capture (Metsky et al., 2017), amplicon (Metsky et al., 2017; Moratorio et al., 2017), SPIA amplification (Grubaugh et al., 2016) or remove unwanted rRNA sequences; ribosomal RNA depletion most notably (Adiconis et al., 2013; Matranga et al., 2016). Enriching for sequences of interest is highly effective but can only target for specific sequences, making it difficult to identify novel, divergent viruses. Current ribosomal depletion methods are typically cost-prohibitive or are only effective for human and mice, making depletion in non-model systems highly inefficient.

Here we describe a novel method of ribosomal depletion that utilizes reverse transcription (RT) to specifically target sequences for depletion using the RNA degradation activity of RNase H. During RT, rRNA is converted to cDNA using specific DNA probes which can then be degraded using RNase H.

Because this method utilizes small probes to recognize the target sequences for depletion, it is possible to design universal probes that bind to highly divergent species, making depletion of diverse organisms representing several genera possible with the same probes. Additionally, since the probes are effectively reverse primers that are typically used for RT, they are easy to design, cheap and can be quickly designed for any target sequence from any species or genus of interest for which the rRNA sequence is available. Our studies show that using this method we can selectively remove rRNA from mosquitoes from multiple genera which results in increased relevant data recovered from next-generation sequencing. Furthermore, we apply this method to field-caught mosquitoes and show that we are able to detect multiple novel virus genomes from a highly multiplexed set of samples on a relatively low-output Illumina MiSeq run. Collectively, this work describes an effective method for rRNA depletion that is straight forward, relatively low-cost and highly effective at increasing usable data from high-throughput sequencing experiments.

Materials and Methods

Cells, Viruses, Mosquitoes and Sample Collection

West Nile virus (strain NY99) was generated from an infectious clone as previously described in BHK-21 cells (Shi et al., 2002). Laboratory colonies of *Culex quinquefasciatus, Aedes aegypti* and *Anopheles gambiae* were used for mosquito infections. Mosquitoes were maintained at 26-27°C and 70-80% relative humidity with a 16:8 L:D photoperiod. Water and 10% sucrose were provided *ad libitum*. For preliminary studies, pools of whole mosquitoes (n=10) were collected and homogenized in Trizol solution.

For xenosurveillance studies, groups of *An. gambiae* were exposed to an infectious bloodmeal containing 10^7 PFU of WNV NY99. The next day, midguts from mosquitoes containing a residual bloodmeal were collected by spreading the midgut contents onto CloneSaver FTA cards (GE Healthcare),

and immediately 25µL of RNAlater (ThermoFisher) was added in order to facilitate diffusion of blood into the FTA card and stabilize the nucleic acid. The samples placed on the FTA cards were then punched out and nucleic acid was eluted by incubation in RNA rapid extraction solution (ThermoFisher) for 18 hours.

Ribosomal Depletion

A more detailed protocol is included describing the depletion protocol in Supplemental file 1. Nucleic acids were extracted using either Trizol solution or the Mag-Bind Viral DNA/RNA kit (Omega Bio-tek, USA) and eluted into 50µL of water. The samples were then treated with TURBO DNase (ThermoFisher) and purified using RNAClean XP beads (Beckman Coulter). For reverse transcription, the RNA was mixed with oligos specific for rRNA (sequences listed in Supplemental Table 1) and dNTPs and then heat denatured at 95C for 2 minutes followed by slow cooling to 50C at 0.1C/s. For initial experiments, we tested a panel of reverse transcriptases (RTs), including *Tth* DNA polymerase (in the presence of Mn2+, Promega), Superscript III (SSIII, ThermoFisher), Superscript IV (SSIV, ThermoFisher), Avian myeloblastosis vi-rus (AMV, NEB) and Moloney Murine Leukemia Virus (MMLV, NEB). For all RTs, we used the optimal conditions as described by the manufacturer. For all further experiments, AMV RT was then added and incubated at 50C for 2 hours. RNase H (NEB) was then added to destroy the RNA present in the RNA:cDNA hybrid. The samples were then digested with DNase I (NEB) to remove the cDNA and residual oligos. The RNA was then purified using RNAClean XP beads at a 1.8x ratio.

RNA Analysis and qRT-PCR

Input and rRNA depleted RNA were analyzed using a 2100 Bioanalyzer (Agilent) per manufacturer's protocols with the total RNA pico kit. The RNA traces were analyzed using Agilent 2100 expert software. Quantitative Reverse-transcriptase PCR (qRT-PCR) was performed using the iTaq universal probes supermix (Biorad) according to the manufacturer. qRT-PCR was performed with the following primers;

18S Forward - AGAGGACTACCATGGTTGCAAC, 18S Reverse - CCTGCTGCCTTCCTTGGATG, 18S Probe - CCGGAGAGGGAGCCTGAGAAATGGC, 28S Forward - AGGTGCGGAGTTTGACTGG, 28S Reverse - TCCTTATGCTCAGCGTGTGG, 28S Probe - AGGTGTCCAAAGGTCAGCTCAGTGTGG, WNV Forward - TCAGCGATCTCTCCACCAAAG, WNV Reverse - GGGTCAGCACGTTTGTCATTG, WNV Probe - TGCCCGACCATGGGAGAAGCTC (Lanciotti et al., 2000). The number of genome copies was generated by fitting the Ct values to a standard curve of RNA specific to each of the primer sets.

Library Preparation and Data Analysis

Libraries for Illumina sequencing were prepared from both input RNA and samples that were depleted using probes specific to rRNA or in the absence of probes. The libraries were prepared using equal concentrations of RNA as input by using the NEBNext Ultra RNA library prep kit (NEB) and then were sequenced on an Illumina MiSeq using 150 cycles. For data analysis, libraries were first demultiplexed using bcl2fastq (Illumina). Reads were then trimmed for both adapters and quality using BBDuk software (part of the BBMap suite, https://sourceforge.net/projects/bbmap/). PCR duplicates were then removed using clumpify (also part of the BBMap suite) and unique reads were mapped to reference genomes using bowtie2 (Langmead and Salzberg, 2012). We then used MultiQC to quantify the percentage of reads that mapped to each reference. These percentages were graphed using GraphPad Prism version 7. To assess intrahost variation, unique reads were mapped to the Bolahun virus reference sequence using BBMap and then variants were called using LoFreq (Wilm et al., 2012). Only variants present at greater than 5% were used for analysis.

Mosquito collections

Adult mosquitoes were collected from multiple localities in Chiapas, Mexico over the course of three weeks in August, 2016 using CDC gravid traps (John W. Hock Company), CDC Miniature light traps (BioQuip Products) and insectazookas (BioQuip Products). Mosquitoes were euthanized using

triethylamine and sorted into pools of up to 25 individuals by species, sex, and collection location (Supplemental Table 2). Mosquitoes were identified to species using morphological keys (Darsie and Ward, n.d.). For groups of mosquitoes that could not be identified, multiple individuals of each group were point mounted and preserved for later identification by local experts at the Instituto Nacional de Salud Pública facilities in Mexico. Pools of mosquitoes were preserved in RNALater (Ambion) and shipped to Colorado State University (CSU).

Processing of field collected mosquitoes

Prior to homogenization and nucleic acid extraction, mosquito pools were centrifuged and RNA later was removed. Pools were then processed as described above. All field collected mosquito pools were subjected to rRNA depletion using the same probe mixture as the laboratory experiments. Following rRNA depletion, RNA from pools was prepared for NGS using Nextera XT following manufacturer's instructions (Illumina). Each library was dual-indexed with a unique barcode to facilitate multiplexing using the Kapa Library Amplification Kit for Illumina (Kapa BioSystems). Libraries were then quantified using the NEBNext Library Quantification Kit for Illumina (New England Biolabs) and pooled together by equal volumes. All libraries were sequenced together on a single Illumina MiSeq run using a 300 cycle (2x150) MiSeq v3 kit.

Identification and characterization of viral sequences

Virus contigs were identified using a previously described pipeline (Cross et al., 2018; Fauver et al., 2018) (found online at https://github.com/stenglein-lab/taxonomy_pipeline). No host filtering was conducted prior to the generation of contigs, as the majority of genera sequenced to do not have a reference genome. Amino acid similarity to other virus or virus-like sequences was determined using NCBI Blastx tool against the nr database (Altschul et al., 1990) (Supplemental Table 3). Virus contigs greater than 500 b.p. were sorted into high-level clades according to Shi et al. (Shi et al., 2016). Contigs from the

same species of mosquito aligning to similar viral clades were binned together in Geneious v11.0.4 and assessed for open-reading frames (ORFS) using the Find ORFs tool (Kearse et al., 2012). Following translation of complete ORFs, amino acid sequences were queried against the Conserved Domain Database v3.16 using HHpred (Zimmermann et al., 2018). Predicted domains with an e-value > 1e-5 were used for annotation. The Luteo Sobemo virus from *Ae. aegypti* was predicted to have multiple segments based on 1) homology to the most similar virus currently described, Hubei mosquito virus (Shi et al., 2016), 2) the identification of two contigs with complete ORFs, 3) similar depth of coverage across viral segments, and 4) the co-occurence of each segment in the same libraries. All putative virus genomes were described entirely using computational methods and virus isolation was not attempted.

Phylogenetic trees were created for coding complete virus genomes. The RNA dependent RNA polymerase (RDRP) gene (Luteo-Sobemo, Levi-Narna) or the whole genome (Negevirus) was used as input for blastp, and all hits with an e-value > 1e-5 were downloaded in .fasta format from NCBI. CD-Hit -c 0.90 was used to rid dataset of similar viral RDRPs sequences (Li and Godzik, 2006). Amino acid sequences were aligned using MAFFT v7.308 -auto (Katoh and Standley, 2013). Gaps and poorly aligned sequences in the multiple alignment were removed using trimAl under default settings (Capella-Gutiérrez et al., 2009). The resulting alignments were used as input to generate phylogenetic trees using PHYML with the LG substitution model and 100 bootstraps (Guindon et al., 2010). In addition, genomic sense was inferred based on placement in phylogeny.

To calculate depth of coverage, a custom database was created by species containing all viral contigs generated in this study in addition to the 45s rDNA sequence assembled from *Ae. aegypti.* Reads from each mosquito species were competitively aligned to this database using Bowtie2 under default settings (Langmead and Salzberg, 2012). The resulting SAM file was converted into BAM format, and depth of coverage at each nucleotide position was calculated using SAMtools -depth (Li et al., 2009). As well, the percentage of total reads to viruses and rRNA sequences was calculated from this database. Novel

Narna-Levi virus sequences were aligned as described above, and pairwise nucleotide identity was calculated in Geneious.

Data availability

All sequencing data has been deposited to the SRA database under BioProject SUB4694537. Novel virus genomes have been submitted to Genbank and are pending accession number assignment.

Results

Reverse-transcriptase (RT) mediated ribosomal RNA (rRNA) depletion is effective for mosquitoes from three medically relevant genera

The workflow for our proposed ribosomal depletion method is outlined in Figure 1. Briefly, DNase treated RNA was reverse-transcribed using DNA probes that are in the reverse-complement orientation to the sequences for mosquito sequences for the 18S, 28S and 5.8S cellular rRNA and the 12S and 16S mitochondrial rRNA sequences. In order to design probes that work against the majority of mosquitoes, we aligned sequences from several mosquito genera obtained from the SILVA rRNA database project (Quast et al., 2013). The probes were designed specifically to regions of high sequence homology and to have a melting temperature around 65℃, thereby giving them high specificity while maintaining binding to all genera. The probe sequences are presented in Supplemental Table 1 and a schematic showing the probes aligned to the *Aedes albopictus* 45S rRNA sequence is presented in Supplemental Figure 1. For the depletion, the RNA was heat denatured in the presence of the probes and slow cooled to favor specific binding of the probes to the RNA. cDNA was then synthesized using AMV reverse transcriptase (RT). It was determined that AMV RT was superior to other RTs tested in depleting 18S and 28S from *An. gambiae* mosquitoes (Fig. 2A-B). While MMLV RT was also able to significantly reduce rRNA, it also depleted WNV RNA, while AMV did not, suggesting that the depletion was highly specific (Fig. 2C). AMV and SSIII RT were the only RTs tested that significantly reduced the amount of 18S and 28S rRNA while

maintaining the same amount of WNV RNA ($p<0.0001$ for 18S and 28S and $p=0.9990$ for WNV RNA all when comparing with and without probes and by One-Way ANOVA with Tukey's correction). We continued with AMV RT because the reduction in rRNA was more dramatic and because it is less expensive than SSIII. Following AMV RT, we treated samples with RNase H and finally DNase I to degrade the RNA in the DNA:RNA hybrid and any DNA present, respectively. Using qRT-PCR, we saw a significant reduction in 18S and 28S rRNA from *An. gambiae* only when the RT and RNase/DNase steps were included and not when any steps were omitted, suggesting that the reverse transcription and RNaseH/DNase I treatment are all required for specific depletion (Fig. 2D-E, all $p<0.0001$ by One-Way ANOVA with Tukey's correction as compared to the depleted group). The reduction from the DNase treated RNA to the samples not treated with RT or depletion probes is likely due to the removal of small fragments during RNAClean bead purification.

We next sought to determine if the ribosomal depletion protocol was effective for mosquito species from three distinct medically relevant genera; *Anopheles, Aedes* and *Culex.* Total RNA was extracted from pools (n=10) of whole mosquitoes and the RNA was depleted as previously described, with the exception that additional probes were added to the mixture that targeted undepleted rRNA sequences identified in preliminary NGS analysis (data not shown). We then subjected the input RNA, depleted RNA (RT - with probes) and RNA that went through the depletion process without probes (RT - no probes) to qRT-PCR analysis. For all species tested, the depleted RNA had significantly reduced 18S (Fig. 3A) and 28S (Fig. 3B) rRNA levels as compared to the two other groups ($p<0.0001$ for all comparisons, Two-Way ANOVA with Tukey's correction). We also subjected both the input RNA and the depleted RNA to electrophoretic analysis using a Bioanalyzer 2100. For all three species tested, the peak for rRNA (both 18S and 28S typically appear at ~2000nt) is inapparent following the depletion protocol (Fig. 3C-E). In contrast, the input RNA has a prominent peak for rRNA. The traces for the depleted and input RNA are overlaid on the same graph to facilitate comparison.

RT mediated rRNA depletion increases sequencing reads to viruses and mRNA

Samples to test depletion efficacy were prepared using a method termed Xenosurveillance, prepared as described previously (Fauver et al., 2018). Briefly, *An. gambiae* mosquitoes were exposed to an infectious bloodmeal containing WNV (which doesn't replicate in these mosquitoes) and then midguts containing the partially digested blood were collected the next day on FTA cards. The nucleic acids were eluted and extracted as previously described and then the samples were depleted with AMV RT and depletion probes (RT - with Probes). We also tested the Input RNA and samples that underwent the depletion protocol with the omission of probes (RT - No Probes). Following depletion, the RNA underwent Illumina library prep and was sequenced using the MiSeq platform (Illumina). The reads were then trimmed, duplicates removed and mapped to either rRNA (18S (Fig. 4A), 28S (Fig. 4B), host transcriptome (Fig. 4C) or viral (WNV (Fig. 4D), Bolahun virus (BOLV, Fig. 4E) sequences. BOLV is known to persistently infect these mosquitoes (Fauver et al., 2016). A significantly lower proportion of reads mapped to rRNA in the depleted RNA (One-Way ANOVA with Tukey's correction, $p<0.0001$ for all comparisons to depleted). In contrast, a significantly increased proportion of the reads aligned to sequences of interest, notably the host transcriptome and the two viruses, WNV and BOLV (One-Way ANOVA with Tukey's correction, all $p<0.01$ or lower for all comparisons to depleted). Coverage plots from input, depleted and non-depleted RNA samples are presented in Supplemental Figure 2 for both BOLV and WNV. Finally, we assessed the ability to analyze intrahost viral variation in BOLV by calling variants with LoFreq. A significantly greater number of minority variants could be called in the depleted RNA when compared to the other two groups (One-Way ANOVA with Tukey's correction, $p<0.05$ for all comparisons to depleted).

Mosquito collections and sequencing summary

A total of 978 adult field-collected mosquitoes from 10 species were pooled for analysis by NGS (Supplemental Table 2). The most abundant species collected (242) was *Coquillettidia venezuelensis,* followed by *Ae. albopictus* (238), *Psorophora albipes* (110), *Ps. varipes* (101), *Ae. angustivittatus* (91), *Cx. nigripalpus* (87), *Ae. aegypti* (72), *Ae. taeniorhynchus* (33), *Ae. serratus* (2), and *Ps. ferox* (2). All species collected in this study have previously been reported from Chiapas state (Bond et al., 2014; Heinemann and Belkin, 1977). A single MiSeq run following quality filtering and removal of duplicate reads yielded 25.9 million total reads, resulting in 3.8Gb of paired-end data. The total percentage of reads mapping to rRNA in the field samples was in line with what we observed after depletion in our colony mosquitoes (Supplemental Fig. 3). In contrast, the percentage of reads mapping to viruses was relatively high, particularly for *Ae. aegypti*. The percentage of reads mapping to viruses varied widely between the different mosquito species tested.

Virus sequences identified in field collected mosquitoes following rRNA depletion

Each mosquito species sequenced, save a single pool of 2 *Ps. ferox* mosquitoes, produced contigs aligning to known viral sequences (Fig. 5, Supplemental Table 3). Based off of amino acid similarity and phylogenetic placement, 8 major clades as well as multiple families of RNA viruses were represented across all samples. Amino acid similarities spanned anywhere from 28% (Reovirus contig from *Ae. angustivittatus*) to 99% (Phasi Charoen-like phasivirus RDRP from multiple *Aedes* species). Multiple previously described viruses were identified, based on >95% pairwise nucleotide identity, including Phasi Charoen-like phasivirus (PCLV) in *Ae. aegypti , Ae. angustivittatus,* and *Ps. varipes*. A complete genome of PCLV was assembled from pools of both male and female *Ae. aegypti* mosquitoes (Supplemental Fig. 4). This PCLV genome aligned to Phasi Charoen-like phasivirus strain 2b (Accession: MH237598) with ~98% pairwise nucleotide identity. PCLV sequences from *Ae. angustivittatus* and *Ps. varipes* aligned only to a portion of the RDRP. Partial sequences aligning to both the RDRP and capsid proteins of Humaita-

Tubiacanga (HTV) virus were identified from female *Ae. aegypti* and male *Ae. albopictus* mosquitoes. Sequences aligned to HTV with 98.5 and 97.5% pairwise nucleotide identity, respectively.

Short flavivirus sequences (100-250) were found in 7 of 8 mosquito species sequenced aligning to the same portion of the West Nile virus (WNV) genome. Based on the sequence similarity between species, its presence in nearly all groups, and our frequent use of WNV in our laboratory, it is likely these sequences are the result of laboratory contamination during library preparation opposed to an authentic infection in our mosquito samples.

While numerous contigs were generated that distantly resembled known viral sequences, indicating the presence of divergent viruses in these species, we chose to further analyze only contigs that produced coding complete viral genomes (Ladner et al., 2014). Our computational approached generated 6 novel viral genomes, including a novel strain of a previously described negevirus (Fig. 6A), 5 Levi-Narnaviruses (Fig. 7A-E), and 1 Luteo-Sobemo virus (Fig. 8A).

A total of 4 contigs identified in *Cx. nigripalpus* mosquitoes aligned to the CoB_37B strain of Cordoba virus with estimated gaps of 188, 72, and 55 nucleotides. The assembly of these contigs produced a final sequence approximately 7,300 nucleotides long that contained a single ORF predicted to code for 4 proteins (Fig. 6A). These proteins include a viral methyltransferase (pfam01660), FtsJ-like methyltransferase (pfam01728), Viral RNA helicase (pfam01443), and RDRP (cd01699) (Fig. 6A). Both the type of proteins encoded and synteny of the genome are in agreement with representative +ssRNA viruses from the Nelorpivirus group of Negeviruses (Nunes et al., 2017) . Phylogenetic placement and high pairwise nucleotide identity (78.8-93.6%, depending on strain) indicated this genome to be a novel strain of Cordoba virus, a negevirus described from a variety of mosquito species, including *Cx. nigripalpus*, from Nepal, the U.S., and Colombia (Nunes et al., 2017) (Fig. 6B,C).

299 Multiple sequences related to viruses in the +ssRNA Narna-Levi clade were identified from *Ae.*
300 *angustivittatus*, *Ae. taeniorhynchus*, *Cq. venezuelensis*, and *Ps. varipes*. Two distinct contigs were
301 generated from *Cq. venezuelensis* mosquitoes. These sequences were found to be approximately 2kb in
302 length and contain a single ORF that encodes for RDRP (cd01699) (Fig. 7 A-E). Pairwise amino acid
303 identity was approximately 72-80% between 4 of the virus sequences, while a sequence from pools of
304 *Ae. angustivittatus* mosquitoes varied substantially (30-33%) compared to other sequences described in
305 this study (Fig. 7F). The 4 more similar genomes grouped with other narnavirus-like sequences described
306 from mosquitoes, where the sequence from *Ae. angustivittatus* mosquitoes grouped with narnavirus-
307 like sequences from crustaceans (Fig. 7 H). These virus genomes have provisionally been designated
308 Aedes angustivittatus narnavirus (AANV), Aedes taeniorhynchus narnavirus (ATNV), Coquillettidia
309 venezuelensis naranvirus 1 & 2 (CVNV1, CVNV2), and Psorophora varipes narnavirus (PVNV).

310 Two sequences related to +ssRNA Luteo-Sobemo like viruses, 2,718 and 1,131 nucleotides in length,
311 were identified in pools of both male and female *Ae. aegypti* mosquitoes. The longer sequence is
312 predicted to encode for two proteins, a Trypsin-like serine protease (cd00190) and RDRP (cd01699),
313 respectively, in two separate ORFs (Fig. 8A). These ORFs overlap and appear to be on the same segment
314 indicating the reading frame difference is likely the result of frameshift mutation, which is common in
315 Luteo-Sobemo viruses (Barry and Miller 2002) . The identified “slippery sequence”, a conserved
316 heptanucleotide sequence that causes the ribosome to shift reading frames, in Sobemoviruses is
317 “UUUAAAC”(Mäkinen et al. 1995). This specific sequence was not identified, however, as these viruses
318 are divergent and not well characterized, it is possible a non-canonical heptanucleotide sequence could
319 exist. A sequence 24 base pairs upstream of the second ORF reads “GGGCCCG”, which deviates slightly
320 from the typical slippery sequence construct of “XXXYYYZ” (P. 2012). It remains to be determined
321 whether this sequence is responsible for ribosomal frameshifting in this virus. The smaller sequence
322 contains a single ORF encoding the predicted viral coat protein (pfam00729). The bipartite genomic

structure is seen in a similar virus, Hubei mosquito virus 2 (Shi et al., 2016). This sequence, provisionally named Renna virus (RENV), groups phylogenetically with viruses identified from a variety of ticks and insects, including mosquitoes (Fig. 8B). Both segments had a high average depth of coverage, 650 and 1351, respectively in *Ae. aegypti* females. RENV from male and female *Ae. aegypti* mosquito pools shared a >99% pairwise nucleotide identity.

Discussion

Studies involving sequencing viral RNA; such as viral metagenomics, intrahost viral dynamics, transcriptomics and virus discovery require target reads to be at sufficient levels to perform meaningful analysis. These analyses are often hampered by the high percentage of ribosomal RNA (rRNA) present in total RNA, which can reach greater than 80-90% of the total sample (Eun, 1996). Since these reads are rarely used, this represents significant waste of both financial and computational resources, and limits the amount of multiplexing that can be performed. While procedures such as selection of polyadenylated transcripts can be used to enrich RNA preparations for mRNA, this is not relevant to RNA viruses that lack polyadenylation. Furthermore, other methods like amplicon sequencing or probe capture are sequence specific, and thus unknown pathogens cannot be sampled. Therefore, selective depletion of highly abundant rRNA is beneficial. Several methods and commercial kits are available to do this but most are designed to work specifically for human or mouse samples. Here, we describe a novel method that utilizes specific reverse transcription of rRNA using small DNA probes for depletion along with RNase H. This allowed us to design depletion probes that could simultaneously deplete rRNA from mosquitoes of highly diverse genetic backgrounds. Using this method, we show that specific depletion of rRNA results in increased reads to meaningful RNA, such as viruses and host mRNA. In addition, we detected more intrahost variants using this depletion method. Although we subjected all field-collected mosquito pools to rRNA depletion, thus no non-depleted libraries were generated for comparison, we were able to detect novel virus genomes from a single, highly multiplexed (64 libraries), MiSeq run of nearly 1,000 diverse

field-collected mosquito samples that underwent rRNA depletion. Taken together, these findings suggest that RT-mediated rRNA depletion can facilitate sequencing sequencing of mosquito samples both from the lab and field.

To our knowledge, only two other studies have aimed to assess rRNA depletion strategies from insect species. The first used a commercial kit designed for mammalian rRNA, Epicentre's Ribo-Zero rRNA, to deplete rRNA from *Drosophila* flies. While the approach seemed to effectively remove rRNA and enrich mRNA transcripts, it suffers from being high-cost (Kumar et al., 2012). Another study showed by bioanalyzer and NGS effective removal of rRNA from mosquito midguts using RNA probes to the rRNA (Kukutla et al., 2013). However, this technique required large amounts of input RNA (50 pooled midguts), uses unstable RNA probes and expensive streptavidin beads. Furthermore, it's unclear if this technique works for other species or just *An. gambiae.* Accordingly, we devised a novel method for depleting rRNA using RNase H depletion that was based on the method described by Morlan et al. (Morlan et al., 2012) with the exception that it uses shorter probes and incorporates a reverse transcription (RT) step. The shorter probes allow highly conserved regions to be targeted, thus making it possible to simultaneously deplete rRNA from high divergent species or even genera. The RT step extends the bound DNA probes to produce cDNA complementary to the rRNA that is destroyed following both a RNase H and DNase I digestion.

First, we assessed the efficacy of several RTs to convert rRNA to cDNA and subsequently be degraded by RNase H. We found AMV to be the optimal enzyme, depleting a significant amount of rRNA with no off-target effects. While M-MLV RT depleted as much or more rRNA as AMV, it also depleted WNV RNA, suggesting it was converting non-target RNA species to cDNA as well (Fig. 2C). It's unclear why M-MLV RT would have off-target effects and not AMV, especially as there has been evidence of the opposite occuring in a previous publication (Agranovsky, 1992). This and other publications have shown primer-independent cDNA synthesis for both AMV and M-MLV RTs, which could explain the high level of non-specific depletion

371 in M-MLV but not AMV observed here (Freeh and Peterhans, 1994). Agranovsky et al. presented evidence
372 that a tRNA contaminant in the AMV RT preparation tested at that time was responsible for this primer-
373 independent cDNA synthesis. We cannot rule out the possibility that the M-MLV obtained from NEB
374 contained some contaminant that could effectively primer non-target RNA species such as WNV. There
375 may also be small RNAs present in our samples that could have primed cDNA synthesis particularly well
376 for M-MLV. Both Superscript (SS) III and IV, mutants of M-MLV were effective at depleting 18S rRNA with
377 no off-target effects. While SSIII also depleted 28S rRNA, SSIV did not effectively deplete this RNA species.
378 Finally, Tth DNA polymerase, which shows RT activity in the presence of manganese, did not effectively
379 deplete rRNA, even in the presence of specific DNA probes. This may be related to the fact that Tth and
380 SSIV lack functional RNase H domains (Myers and Gelfand, 1991), suggesting that this intrinsic activity is
381 important for the mechanism of depletion with this technique, even if RNase H is added after the RT step.
382 Next, we assessed whether the RT step was necessary for depletion in the workflow, as Morlan et al. had
383 previously shown efficient depletion in the absence of this step (Morlan et al., 2012). The RT step was
384 critical to the depletion observed and the specific depletion probes were necessary, as samples treated
385 with DNA probes in the absence of RT had only a modest depletion effect. However, in the presence of
386 RT and specific depletion probes, 18S and 28S rRNA were depleted roughly 100- and 1000-fold,
387 respectively.

388 Depletion was then tested on RNA from three medically important mosquito species representing three
389 distinct genera; *Ae. aegypti, An. gambiae* and *Cx. quinquefasciatus*. These species transmit a significant
390 proportion of vector-borne pathogens; including dengue virus, Zika virus, chikungunya virus, malaria
391 parasites and West Nile virus, among others. We found by qRT-PCR and bioanalyzer, depletion in the
392 presence of rRNA probes was associated with a significant reduction in 18S and 28S rRNA from all three
393 species tested. Despite the almost complete removal of the peak for rRNA in the bioanalyzer traces, we
394 were still able to detect rRNA sequences by both qRT-PCR and NGS. This might be a result of incomplete

digestion of the RNA by RNase H due to incomplete activity or RNA that hadn't been reverse transcribed. It's possible that the secondary structure of rRNA prevents the complete synthesis of cDNA from RNA and that this is not degraded by the RNase H. Different methods to increase the efficiency of cDNA synthesis or adding additional DNA probes may be beneficial in future iterations of this protocol. This result suggested that this protocol could be used for a wide array of mosquito species, as *Aedes* and *Culex* are significantly divergent from *Anopheles* mosquitoes, having separated likely over 200 million years ago (Reidenbach et al., 2009). In fact, we have seen rRNA depletion by NGS in virus stocks prepared in mammalian cells as well, suggesting a broad range of cross-reactivity to rRNA from different species.

We then depleted rRNA from midguts isolated from *An. gambiae* mosquitoes that were fed a bloodmeal containing WNV. This RNA was then subjected to Illumina deep-sequencing and the resulting reads were aligned to several sequences. We observed significant depletion of rRNA while increasing the percentage of reads to mRNA, WNV and the insect-specific virus Bolahun virus (Fauver et al., 2016). We were also able to identify significantly more minority variants present in Bolahun virus, suggesting intrahost virus population analyses are facilitated following depletion. It has been shown that high levels of sequencing coverage are necessary to perform intrahost virus analysis, which is can be difficult to achieve without depletion or enrichment (McCrone and Lauring, 2016).

As second and third generation sequencing based approaches for the detection and analysis of vector-borne pathogens from field-collected mosquitoes are becoming commonplace, techniques that increase reads to target sequences in complex samples will be sorely needed. Accordingly, we employed our rRNA depletion method to a diverse group of field-collected mosquitoes and subjected them to NGS with the goal of identifying both human-infecting and insect-specific viruses. While we did not identify arbovirus sequences from these pools of mosquitoes, we were able to identify partial and coding complete genomic sequences of a variety of presumed insect specific viruses. As all pools were subjected to rRNA depletion, we do not have non-depleted libraries to compare the efficacy of rRNA depletion to. However, the total

number of reads aligning to rRNA from these samples was congruent with what we observed in our laboratory studies. In fact, in libraries constructed from *Ae. aegypti* females, more reads competitively aligned to viruses than to 28s or 18s rRNA sequences, although the number of reads aligning to both viruses and rRNA sequences varied widely between divergent genera. Using our bioinformatic approach, 7 novel coding complete viral genomes were identified, in addition to the previously described insect specific viruses PCLV and HTV. Complete PCLV genomes were assembled from pools of both male and female *Ae. aegypti* mosquitoes at a relatively high depth of coverage and pairwise nucleotide identity. PCLV has been identified mosquito cell culture and in numerous populations of *Ae. aegypti* mosquitoes from across the globe (Chandler et al., 2014; Di Giallonardo et al., 2018; Yamao et al., 2009; Zhang et al., 2018). In addition to PCLV, we identified large contigs with >99% nucleotide identity to HTV in both female *Ae. aegypti* and male *Ae. albopictus* mosquitoes (Aguiar et al., 2015; Zakrzewski et al., 2018). A total of 5 coding complete narnavirus genome sequences were identified from 4 species of mosquitoes collected in this study. Of the 5 virus genomes described here, 4 group closely together and with other narnaviruses described from mosquitoes. While multiple narnaviruses have been identified by metagenomic sequencing of whole mosquito samples, it remains to be determined if these represent infections of fungi in the normal microbiota, or bona fide infections of mosquitoes (Chandler et al., 2015; Cook et al., 2013; Shi et al., 2016). A novel strain of Cordoba virus, a negevirus described previously from mosquitoes, was identified in *Cx. nigripalpus* mosquitoes (Nunes et al., 2017). We were also able to assemble the coding complete genome of RENV, a virus that groups with Luteo-Sobemo viruses identified in mosquitoes (Shi et al., 2016). Based on the phylogenetic placement of these sequences, all the viruses described in this study are presumed to be insect specific, however this is yet to be validated. As well, the effect these viruses may have on mosquito biology or vector competence remains to be determined. It is highly probable that these viruses would have been detected if we did not perform rRNA depletion, but based on our NGS data from laboratory experiments, our depletion method likely aided in discovery and

443 characterization by allowing more unique, non-rRNA sequences to be identified. Although the amount of
444 viral RNA from any given mosquito depends upon individual infection status and the amount of viral
445 replication occurring, we were able to identify and assemble multiple viral genomes from a highly
446 multiplexed sequencing run on a comparatively low-output sequencing platform. Increasing reads to
447 target sequences of interest (e.g. viruses) by depleting uninformative rRNA sequences from complex,
448 field-collected mosquito samples has the potential to improve the efficacy and feasibility of using
449 metagenomic sequencing for mosquito-borne disease surveillance.

450

Acknowledgements

We would like to acknowledge Nunya Chotiwan, Rushika Pereira, Karla Saavedra, Ildefonso Fernández-Salas and all of the employees at the Centro Regional de Investigación en Salud Pública for assistance in collecting mosquitoes in Tapachula, Chiapas, Mexico. We also acknowledge the funding sources for providing the resources to perform this work; NIAIDAI067380.

459 **References**

460 Adiconis, X., Borges-Rivera, D., Satija, R., DeLuca, D.S., Busby, M.A., Berlin, A.M., Sivachenko, A.,
461 Thompson, D.A., Wysoker, A., Fennell, T., Gnirke, A., Pochet, N., Regev, A., Levin, J.Z., 2013.
462 Comparative analysis of RNA sequencing methods for degraded or low-input samples. Nat.
463 Methods 10, 623–629.
464 Agranovsky, A.A., 1992. Exogenous primer-independent cDNA synthesis with commercial reverse
465 transcriptase preparations on plant virus RNA templates. Anal. Biochem. 203, 163–165.
466 Aguiar, E.R.G.R., Olmo, R.P., Paro, S., Ferreira, F.V., de Faria, I.J. da S., Todjro, Y.M.H., Lobo, F.P., Kroon,
467 E.G., Meignin, C., Gatherer, D., Imler, J.-L., Marques, J.T., 2015. Sequence-independent
468 characterization of viruses based on the pattern of viral small RNAs produced by the host. Nucleic
469 Acids Res. 43, 6191–6206.
470 Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J.
471 Mol. Biol. 215, 403–410.
472 Aubry, M., Teissier, A., Huart, M., Merceron, S., Vanhomwegen, J., Roche, C., Vial, A.-L., Teururai, S.,
473 Sicard, S., Paulous, S., Desprès, P., Manuguerra, J.-C., Mallet, H.-P., Musso, D., Deparis, X., Cao-
474 Lormeau, V.-M., 2017. Zika Virus Seroprevalence, French Polynesia, 2014-2015. Emerg. Infect. Dis.
475 23, 669–672.
476 Bond, J.G., Casas-Martínez, M., Quiroz-Martínez, H., Novelo-Gutiérrez, R., Marina, C.F., Ulloa, A., Orozco-
477 Bonilla, A., Muñoz, M., Williams, T., 2014. Diversity of mosquitoes and the aquatic insects
478 associated with their oviposition sites along the Pacific coast of Mexico. Parasit. Vectors 7, 41.
479 Capella-Gutiérrez, S., Silla-Martínez, J.M., Gabaldón, T., 2009. trimAl: a tool for automated alignment
480 trimming in large-scale phylogenetic analyses. Bioinformatics 25, 1972–1973.
481 Chandler, J.A., Liu, R.M., Bennett, S.N., 2015. RNA shotgun metagenomic sequencing of northern
482 California (USA) mosquitoes uncovers viruses, bacteria, and fungi. Front. Microbiol. 6, 185.
483 Chandler, J.A., Thongsripong, P., Green, A., Kittayapong, P., Wilcox, B.A., Schroth, G.P., Kapan, D.D.,
484 Bennett, S.N., 2014. Metagenomic shotgun sequencing of a Bunyavirus in wild-caught Aedes
485 aegypti from Thailand informs the evolutionary and genomic history of the Phleboviruses. Virology
486 464-465, 312–319.
487 Cook, S., Chung, B.Y.-W., Bass, D., Moureau, G., Tang, S., McAlister, E., Culverwell, C.L., Glücksman, E.,
488 Wang, H., Brown, T.D.K., Gould, E.A., Harbach, R.E., de Lamballerie, X., Firth, A.E., 2013. Novel virus
489 discovery and genome reconstruction from field RNA samples reveals highly divergent viruses in
490 dipteran hosts. PLoS One 8, e80720.
491 Cross, S.T., Kapuscinski, M.L., Perino, J., Maertens, B.L., Weger-Lucarelli, J., Ebel, G.D., Stenglein, M.D.,
492 2018. Co-Infection Patterns in Individual Ixodes scapularis Ticks Reveal Associations between Viral,
493 Eukaryotic and Bacterial Microorganisms. Viruses 10. https://doi.org/10.3390/v10070388
494 Darsie, R.F., Ward, R.A., n.d. Identification and geographical distribution of the mosquitoes of North
495 America, north of Mexico. 2005. Gainesville: University Press of Florida Google Scholar.
496 Di Giallonardo, F., Audsley, M.D., Shi, M., Young, P.R., McGraw, E.A., Holmes, E.C., 2018. Complete
497 genome of Aedes aegypti anphevirus in the Aag2 mosquito cell line. J. Gen. Virol. 99, 832–836.
498 Eun, H.-M., 1996. Enzymology Primer for Recombinant DNA Technology. Elsevier.
499 Fauver, J.R., Grubaugh, N.D., Krajacich, B.J., Weger-Lucarelli, J., Lakin, S.M., Fakoli, L.S., III, Bolay, F.K.,
500 Diclaro, J.W., II, Dabiré, K.R., Foy, B.D., Others, 2016. West African Anopheles gambiae mosquitoes
501 harbor a taxonomically diverse virome including new insect-specific flaviviruses, mononegaviruses,
502 and totiviruses. Virology 498, 288–299.
503 Fauver, J.R., Weger-Lucarelli, J., Fakoli, L.S., 3rd, Bolay, K., Bolay, F.K., Diclaro, J.W., 2nd, Brackney, D.E.,
504 Foy, B.D., Stenglein, M.D., Ebel, G.D., 2018. Xenosurveillance reflects traditional sampling
505 techniques for the identification of human pathogens: A comparative study in West Africa. PLoS

506 Negl. Trop. Dis. 12, e0006348.
507 Forni, D., Filippi, G., Cagliani, R., De Gioia, L., Pozzoli, U., Al-Daghri, N., Clerici, M., Sironi, M., 2015. The
508 heptad repeat region is a major selection target in MERS-CoV and related coronaviruses. Sci Rep5:
509 14480.
510 Freeh, B., Peterhans, E., 1994. RT-PCR: "background priming" during reverse transcription. Nucleic Acids
511 Res. 22, 4342–4343.
512 Grubaugh, N.D., Weger-Lucarelli, J., Murrieta, R.A., Fauver, J.R., Garcia-Luna, S.M., Prasad, A.N., Black,
513 W.C., 4th, Ebel, G.D., 2016. Genetic Drift during Systemic Arbovirus Infection of Mosquito Vectors
514 Leads to Decreased Relative Fitness during Host Switching. Cell Host Microbe 19, 481–492.
515 Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., Gascuel, O., 2010. New algorithms
516 and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML
517 3.0. Syst. Biol. 59, 307–321.
518 Heinemann, S.J., Belkin, J.N., 1977. Collection records of the project "Mosquitoes of Middle America" 9.
519 Mexico (MEX, MF, MT, MX). Mosq. Syst 9, 483–535.
520 Jones, K.E., Patel, N.G., Levy, M.A., Storeygard, A., Balk, D., Gittleman, J.L., Daszak, P., 2008. Global
521 trends in emerging infectious diseases. Nature 451, 990–993.
522 Katoh, K., Standley, D.M., 2013. MAFFT multiple sequence alignment software version 7: improvements
523 in performance and usability. Mol. Biol. Evol. 30, 772–780.
524 Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A.,
525 Markowitz, S., Duran, C., Thierer, T., Ashton, B., Meintjes, P., Drummond, A., 2012. Geneious Basic:
526 an integrated and extendable desktop software platform for the organization and analysis of
527 sequence data. Bioinformatics 28, 1647–1649.
528 Kukutla, P., Steritz, M., Xu, J., 2013. Depletion of ribosomal RNA for mosquito gut metagenomic RNA-
529 seq. J. Vis. Exp. https://doi.org/10.3791/50093
530 Kumar, N., Creasy, T., Sun, Y., Flowers, M., Tallon, L.J., Dunning Hotopp, J.C., 2012. Efficient subtraction
531 of insect rRNA prior to transcriptome analysis of Wolbachia-Drosophila lateral gene transfer. BMC
532 Res. Notes 5, 230.
533 Ladner, J.T., Beitzel, B., Chain, P.S.G., Davenport, M.G., Donaldson, E.F., Frieman, M., Kugelman, J.R.,
534 Kuhn, J.H., O'Rear, J., Sabeti, P.C., Wentworth, D.E., Wiley, M.R., Yu, G.-Y., Threat Characterization
535 Consortium, Sozhamannan, S., Bradburne, C., Palacios, G., 2014. Standards for sequencing viral
536 genomes in the era of high-throughput sequencing. MBio 5, e01360–14.
537 Lanciotti, R.S., Kerst, A.J., Nasci, R.S., Godsey, M.S., Mitchell, C.J., Savage, H.M., Komar, N., Panella, N.A.,
538 Allen, B.C., Volpe, K.E., Davis, B.S., Roehrig, J.T., 2000. Rapid detection of west nile virus from
539 human clinical specimens, field-collected mosquitoes, and avian samples by a TaqMan reverse
540 transcriptase-PCR assay. J. Clin. Microbiol. 38, 4066–4071.
541 Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359.
542 Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R.,
543 1000 Genome Project Data Processing Subgroup, 2009. The Sequence Alignment/Map format and
544 SAMtools. Bioinformatics 25, 2078–2079.
545 Li, W., Godzik, A., 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or
546 nucleotide sequences. Bioinformatics 22, 1658–1659.
547 Matranga, C.B., Gladden-Young, A., Qu, J., Winnicki, S., Nosamiefan, D., Levin, J.Z., Sabeti, P.C., 2016.
548 Unbiased Deep Sequencing of RNA Viruses from Clinical Samples. J. Vis. Exp.
549 https://doi.org/10.3791/54117
550 McCrone, J.T., Lauring, A.S., 2016. Measurements of Intrahost Viral Diversity Are Extremely Sensitive to
551 Systematic Errors in Variant Calling. J. Virol. 90, 6884–6895.
552 Metsky, H.C., Matranga, C.B., Wohl, S., Schaffner, S.F., Freije, C.A., Winnicki, S.M., West, K., Qu, J.,
553 Baniecki, M.L., Gladden-Young, A., Lin, A.E., Tomkins-Tinch, C.H., Ye, S.H., Park, D.J., Luo, C.Y.,

554 Barnes, K.G., Shah, R.R., Chak, B., Barbosa-Lima, G., Delatorre, E., Vieira, Y.R., Paul, L.M., Tan, A.L.,
555 Barcellona, C.M., Porcelli, M.C., Vasquez, C., Cannons, A.C., Cone, M.R., Hogan, K.N., Kopp, E.W.,
556 Anzinger, J.J., Garcia, K.F., Parham, L.A., Ramírez, R.M.G., Montoya, M.C.M., Rojas, D.P., Brown,
557 C.M., Hennigan, S., Sabina, B., Scotland, S., Gangavarapu, K., Grubaugh, N.D., Oliveira, G., Robles-
558 Sikisaka, R., Rambaut, A., Gehrke, L., Smole, S., Halloran, M.E., Villar, L., Mattar, S., Lorenzana, I.,
559 Cerbino-Neto, J., Valim, C., Degrave, W., Bozza, P.T., Gnirke, A., Andersen, K.G., Isern, S., Michael,
560 S.F., Bozza, F.A., Souza, T.M.L., Bosch, I., Yozwiak, N.L., MacInnis, B.L., Sabeti, P.C., 2017. Zika virus
561 evolution and spread in the Americas. Nature 546, 411–415.
562 Moratorio, G., Henningsson, R., Barbezange, C., Carrau, L., Bordería, A.V., Blanc, H., Beaucourt, S.,
563 Poirier, E.Z., Vallet, T., Boussier, J., Mounce, B.C., Fontes, M., Vignuzzi, M., 2017. Attenuation of
564 RNA viruses by redirecting their evolution in sequence space. Nat Microbiol 2, 17088.
565 Morlan, J.D., Qu, K., Sinicropi, D.V., 2012. Selective depletion of rRNA enables whole transcriptome
566 profiling of archival fixed tissue. PLoS One 7, e42882.
567 Moudy, R.M., Meola, M.A., Morin, L.-L.L., Ebel, G.D., Kramer, L.D., 2007. A newly emergent genotype of
568 West Nile virus is transmitted earlier and more efficiently by Culex mosquitoes. Am. J. Trop. Med.
569 Hyg. 77, 365–370.
570 Myers, T.W., Gelfand, D.H., 1991. Reverse transcription and DNA amplification by a Thermus
571 thermophilus DNA polymerase. Biochemistry 30, 7661–7666.
572 Nunes, M.R.T., Contreras-Gutierrez, M.A., Guzman, H., Martins, L.C., Barbirato, M.F., Savit, C., Balta, V.,
573 Uribe, S., Vivero, R., Suaza, J.D., Oliveira, H., Nunes Neto, J.P., Carvalho, V.L., da Silva, S.P., Cardoso,
574 J.F., de Oliveira, R.S., da Silva Lemos, P., Wood, T.G., Widen, S.G., Vasconcelos, P.F.C., Fish, D.,
575 Vasilakis, N., Tesh, R.B., 2017. Genetic characterization, molecular epidemiology, and phylogenetic
576 relationships of insect-specific viruses in the taxon Negevirus. Virology 504, 152–167.
577 Otte, A., Sauter, M., Daxer, M.A., McHardy, A.C., Klingel, K., Gabriel, G., 2015. Adaptive Mutations That
578 Occurred during Circulation in Humans of H1N1 Influenza Virus in the 2009 Pandemic Enhance
579 Virulence in Mice. J. Virol. 89, 7329–7337.
580 Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., Glöckner, F.O., 2013. The
581 SILVA ribosomal RNA gene database project: improved data processing and web-based tools.
582 Nucleic Acids Res. 41, D590–6.
583 Reidenbach, K.R., Cook, S., Bertone, M.A., Harbach, R.E., Wiegmann, B.M., Besansky, N.J., 2009.
584 Phylogenetic analysis and temporal diversification of mosquitoes (Diptera: Culicidae) based on
585 nuclear genes and morphology. BMC Evol. Biol. 9, 298.
586 Shi, M., Lin, X.-D., Tian, J.-H., Chen, L.-J., Chen, X., Li, C.-X., Qin, X.-C., Li, J., Cao, J.-P., Eden, J.-S.,
587 Buchmann, J., Wang, W., Xu, J., Holmes, E.C., Zhang, Y.-Z., 2016. Redefining the invertebrate RNA
588 virosphere. Nature. https://doi.org/10.1038/nature20167
589 Shi, P.-Y., Tilgner, M., Lo, M.K., Kent, K.A., Bernard, K.A., 2002. Infectious cDNA clone of the epidemic
590 west nile virus from New York City. J. Virol. 76, 5847–5856.
591 Tsetsarkin, K.A., Vanlandingham, D.L., McGee, C.E., Higgs, S., 2007. A single mutation in chikungunya
592 virus affects vector specificity and epidemic potential. PLoS Pathog. 3, e201.
593 WHO, 2018. 2018 annual review of the Blueprint list of priority diseases [WWW Document]. who.int.
594 URL http://www.who.int/emergencies/diseases/2018prioritization-report.pdf
595 Wilm, A., Aw, P.P.K., Bertrand, D., Yeo, G.H.T., Ong, S.H., Wong, C.H., Khor, C.C., Petric, R., Hibberd, M.L.,
596 Nagarajan, N., 2012. LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering
597 cell-population heterogeneity from high-throughput sequencing datasets. Nucleic Acids Res. 40,
598 11189–11201.
599 Yamao, T., Eshita, Y., Kihara, Y., Satho, T., Kuroda, M., Sekizuka, T., Nishimura, M., Sakai, K., Watanabe,
600 S., Akashi, H., Rongsriyam, Y., Komalamisra, N., Srisawat, R., Miyata, T., Sakata, A., Hosokawa, M.,
601 Nakashima, M., Kashige, N., Miake, F., Fukushi, S., Nakauchi, M., Saijo, M., Kurane, I., Morikawa, S.,

602 Mizutani, T., 2009. Novel virus discovery in field-collected mosquito larvae using an improved
603 system for rapid determination of viral RNA sequences (RDV ver4.0). Arch. Virol. 154, 153–158.
604 Zakrzewski, M., Rašić, G., Darbro, J., Krause, L., Poo, Y.S., Filipović, I., Parry, R., Asgari, S., Devine, G.,
605 Suhrbier, A., 2018. Mapping the virome in wild-caught Aedes aegypti from Cairns and Bangkok. Sci.
606 Rep. 8, 4690.
607 Zhang, X., Huang, S., Jin, T., Lin, P., Huang, Y., Wu, C., Peng, B., Wei, L., Chu, H., Wang, M., Jia, Z., Zhang,
608 S., Xie, J., Cheng, J., Wan, C., Zhang, R., 2018. Discovery and high prevalence of Phasi Charoen-like
609 virus in field-captured Aedes aegypti in South China. Virology 523, 35–40.
610 Zimmermann, L., Stephens, A., Nam, S.-Z., Rau, D., Kübler, J., Lozajic, M., Gabler, F., Söding, J., Lupas,
611 A.N., Alva, V., 2018. A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred
612 Server at its Core. J. Mol. Biol. 430, 2237–2243.

613

614 **Figure legends**

615

616 **Figure 1: Workflow for Reverse-Transcriptase Mediated Ribosomal Depletion from Total RNA.** To
617 perform ribosomal RNA (rRNA) depletion, total RNA is first extracted, DNAse treated and subsequently
618 purified with RNAClean XP Beads (Agencourt). DNA-free RNA is then bound to oligonucleotide probes
619 designed to bind to rRNA from mosquito species in *Aedes, Culex* and *Anopheles* genera that are in the
620 reverse complement orientation to both the long and short ribosomal subunit and 12s and 16s
621 mitochondrial rRNA. The RNA with bound oligos is then subjected to reverse transcription using Avian
622 Myeloblastosis Vi-rus (AMV) Reverse Transcriptase (NEB). RNA that is reverse transcribed to cDNA is
623 then digested using RNAse H, which selectively destroys RNA in a RNA:DNA hybrid. Remaining DNA is
624 then digested using DNAse I (NEB), leaving mostly non-ribosomal RNA which is then used for library
625 preparation.

626

627

628 **Figure 2: Reverse Transcriptase mediated ribosomal RNA (rRNA) depletion is most effective with AMV**
629 **RT and requires all steps to be effective**. Nucleic acids were eluted from FTA cards with midgut contents
630 of An. gambiae that had been exposed to a bloodmeal containing West Nile virus (WNV) placed on
631 them. RNA and DNA was then extracted to obtain total nucleic acid. The nucleic acid was then treated
632 with DNase I and then purified to obtain total RNA. This RNA was then subjected to cDNA synthesis with
633 a panel of reverse transcriptases in the presence (+probes) or absence (- probes) of DNA probes specific
634 to rRNA. The RTs tested were Tth DNA polymerase, Superscript III (SSIII), Superscript IV (SSIV), AMV and
635 MMLV. All of the samples were then treated with RNase H and then DNase I to remove the RNA present
636 in an RNA:DNA hybrid and cDNA, respectively. The samples were then purified and subjected to qRT-
637 PCR with primer probe combinations specific for 18S rRNA (A), 28S rRNA (B) or WNV (C). Further tests

were performed exclusively with AMV RT. Panels D and E show the results of qRT-PCR for samples that underwent the process of depletion but omitting some step or reagent. 18S (D) and 28S (E) rRNA was quantified in the input RNA, RNA with no RT added, RNA with no depletion probes added and RNA treated with RT with depletion probes. All statistical tests were performed by One-Way ANOVA with Tukey's test for multiple comparisons. **** Indicates p-value <0.0001.

Figure 3: Reverse Transcriptase mediated ribosomal RNA (rRNA) depletion is effective against mosquitoes from three distinct medically relevant genera. Total RNA was extracted from three distinct pools of whole mosquitoes from three medically relevant genera; *Culex (Cx.) quinquefaeciatus*, *Aedes (Ae.) aegpyti* and *Anopheles (An.) gambiae.* The RNA was treated with DNase I and then purified; this will now be called Input RNA. An aliquot was then taken and reverse transcribed to cDNA using AMV reverse transcriptase (RT) and DNA probes specific for mosquito ribosomal RNA (RT – with Probes) or in the absence of probes (RT – No Probes). The samples were then treated with RNase H and DNase I to remove the RNA present in an RNA:DNA hybrid and cDNA, respectively. The samples were then purified and subjected to qRT-PCR with primer probe combinations specific for 18S or 28S rRNA (A and B). The Input RNA and RT – with Probes were then assessed using a Bioanalyzer. Panels C-E show a representative trace for each of the three mosquito species tested, *Cx. quinquefaeciatus* (C), *Ae. aegypti* (D), *An. gambiae* (E). The blue trace for each panel shows the Input RNA and the red trace shows the RT – with Probes treated RNA. The peak present at roughly 40 seconds in each trace is the peak for both 18S and 28S rRNA.

Figure 4: Reverse Transcriptase mediated ribosomal RNA (rRNA) depletion increases target-specific coverage while reducing the number of rRNA reads in next-generation sequencing. *Anopheles gambiae* mosquitoes were exposed to an infectious bloodmeal containing 10^7 PFU of West Nile virus strain NY99.

662 The following day, midguts were dissected and the residual bloodmeal was spread onto a CloneSaver
663 FTA card (GE Healthcare, USA) and then soaked in RNAlater solution to stabilize the nucleic acid and
664 facilitate dispersion. Total nucleic acid was then extracted and DNAse treated. This is considered the
665 Input RNA. DNAse-free RNA was then reverse transcribed using either ribosomal RNA specific probes (RT
666 – with Probes) or without probes (RT – no Probes). The samples were then treated with RNAse H and
667 DNAse I and purified. The samples were then subjected to library preparation and sequenced on an
668 Illumina MiSeq. Reads were then demultiplexed and subsequently trimmed using BBDuk. Duplicate
669 reads were removed using Clumpify and then unique reads were mapped using Bowtie2 to the
670 appropriate reference sequence, 18S rRNA (A), 28S rRNA (B), *An. gambiae* transcriptome (C), West Nile
671 virus (D) and Bolahun virus (E). Percentage of reads mapping was calculated using MultiQC. Variants
672 detected in Bolahun virus were called using LoFreq (F).
673
674 **Figure 5: Viral sequences belonging to diverse clades of RNA viruses identified in field-collected**
675 **mosquitoes following rRNA depletion.** Individual reads from each mosquito species sequenced were
676 mapped back to all virus contigs identified in this study. Virus clade is inferred by amino acid similarity to
677 other closely related sequences.
678
679 **Figure 6: Description of a novel variant of the negevirus Cordoba virus from *Culex nigripalpus*** A- Virus
680 cartoon depicting the genomic structure and depth of coverage Cordoba virus *Cx. nigripalpus* variant.
681 The large boxes represent predicted ORFs and the small boxes represent areas of protein homology to
682 viral methyltransferase (pfam01660), FtsJ-like methyltransferase (pfam01728), viral RNA helicase
683 (pfam01443), and viral RNA-dependent RNA polymerases (cd1699). B. Phylogenetic placement of
684 multiple strains of Cordoba virus highlighted in blue. Phylogenies were created using 1,234 A.A. residues
685 across the complete ORF. Tree is midpoint rooted. Phylogenetic trees were generated in FigTree. C.

686 Expansion of phylogenetic tree containing the sequenced strains of Cordoba virus. The strain sequenced
687 in this study is highlighted in blue. Phylogenetic trees were generated in FigTree.
688
689 **Figure 7: Multiple, unique narnaviruses described from multiple mosquito species.** A-E cartoons
690 depicting the simple genomic structure and depth of read coverage to newly described narnaviruses.
691 The large boxes represent predicted ORFs and the small boxes represent protein homology to viral RNA-
692 dependent RNA polymerases (cd1699). A- CVNV1, B- CVNV2, C- PVNV, D- ATNV, E- AANV. F- Pairwise
693 identify of 295 amino acid residues across the predicted RDRP between the newly described
694 narnaviruses. H- Phylogenetic placement of novel narnaviruses highlighted in blue. Tree based on
695 alignments of RDRP from multiple narnavirus and is midpoint rooted. Phylogenetic trees were generated
696 in FigTree.
697
698 **Figure 8: Description of a novel Luteo-Sobemo like virus from *Aedes aegypti* mosquitoes.** A- Cartoon
699 depicting the predicted bipartite genomic structure of RENV. Large boxes represent ORFs, small boxes
700 represent areas of protein homology to Trypsin-like serine protease (cd00190), viral RNA-dependent
701 RNA polymerase (cd1699), and capsid protein (cd00205). B- Phlyogenetic placement of RENV. Phylogeny
702 was created using a 289 amino acid portion of the RDRP. Trees are midpoint rooted. Phylogenetic trees
703 were generated in FigTree.
704

705 **Supplemental Material**

706

707 **Supplemental File 1:** Detailed protocol for rRNA depletion.

708

709 **Supplemental Figure 1: Position of DNA probes across the mosquito 45S Ribosomal RNA (rRNA)**
710 **sequence.** The DNA probes were aligned to the 45S rRNA sequence of *Aedes albopictus*. The probes are
711 presented in green and the rRNA sequence is labelled and in red.

712

713 **Supplemental Figure 2: Depth of coverage by nucleotide position for Bolahun virus and West Nile**
714 **virus.** Depth was calculated at each nucleotide position using Samtools depth -a. Three samples were
715 sequenced per group. Grey shading represents the minimum and maximum coverage at each position.

716

717 **Supplemental Figure 3: Proportion of total reads mapping to virus, 18s, and 28s sequences from field-**
718 **collected mosquitoes that have undergone rRNA depletion.**

719

720 **Supplemental Figure 4: Depth of coverage of complete Phasi-Chareon like phasivirus from female *Ae.***
721 ***aegypti* mosquitoes.**

722

723 **Supplemental Table 1:** List of oligonucleotide probe sequences aligning to mosquito 45S rRNA used for
724 depletion.

725

726 **Supplemental Table 2:** Metadata for each library constructed from field-collected mosquitoes.

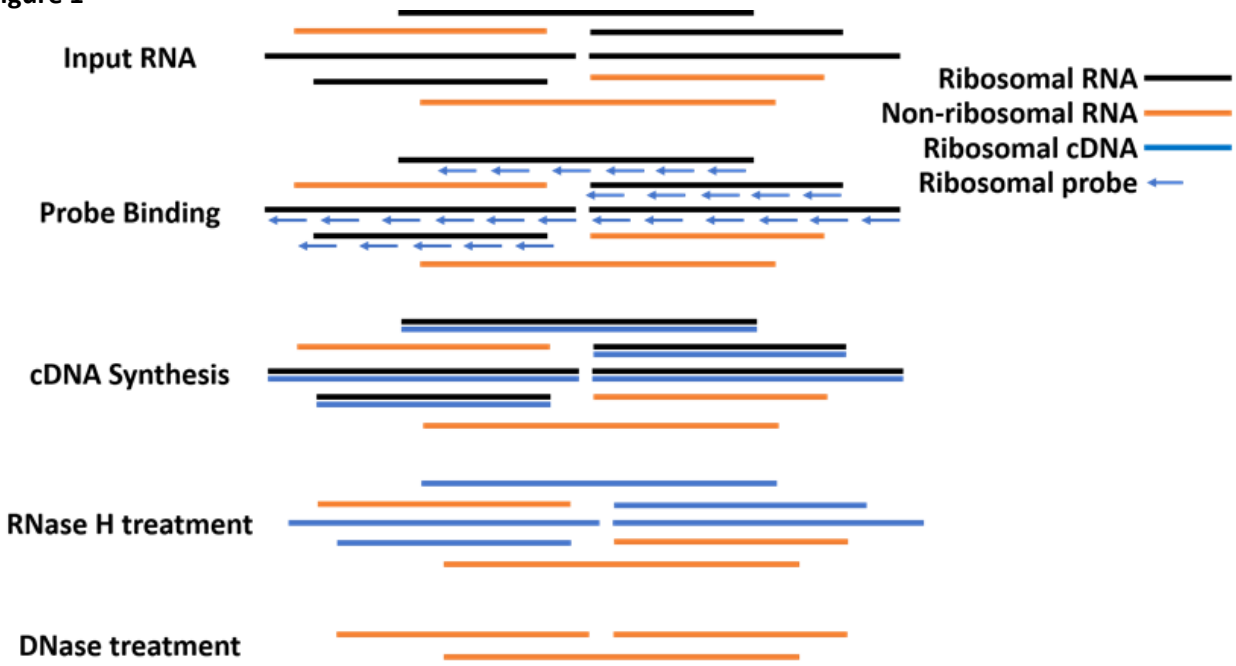
727

728 **Supplemental Table 3:** Description of viral contigs >500 nucleotides in length identified from field-

729 collected mosquitoes.

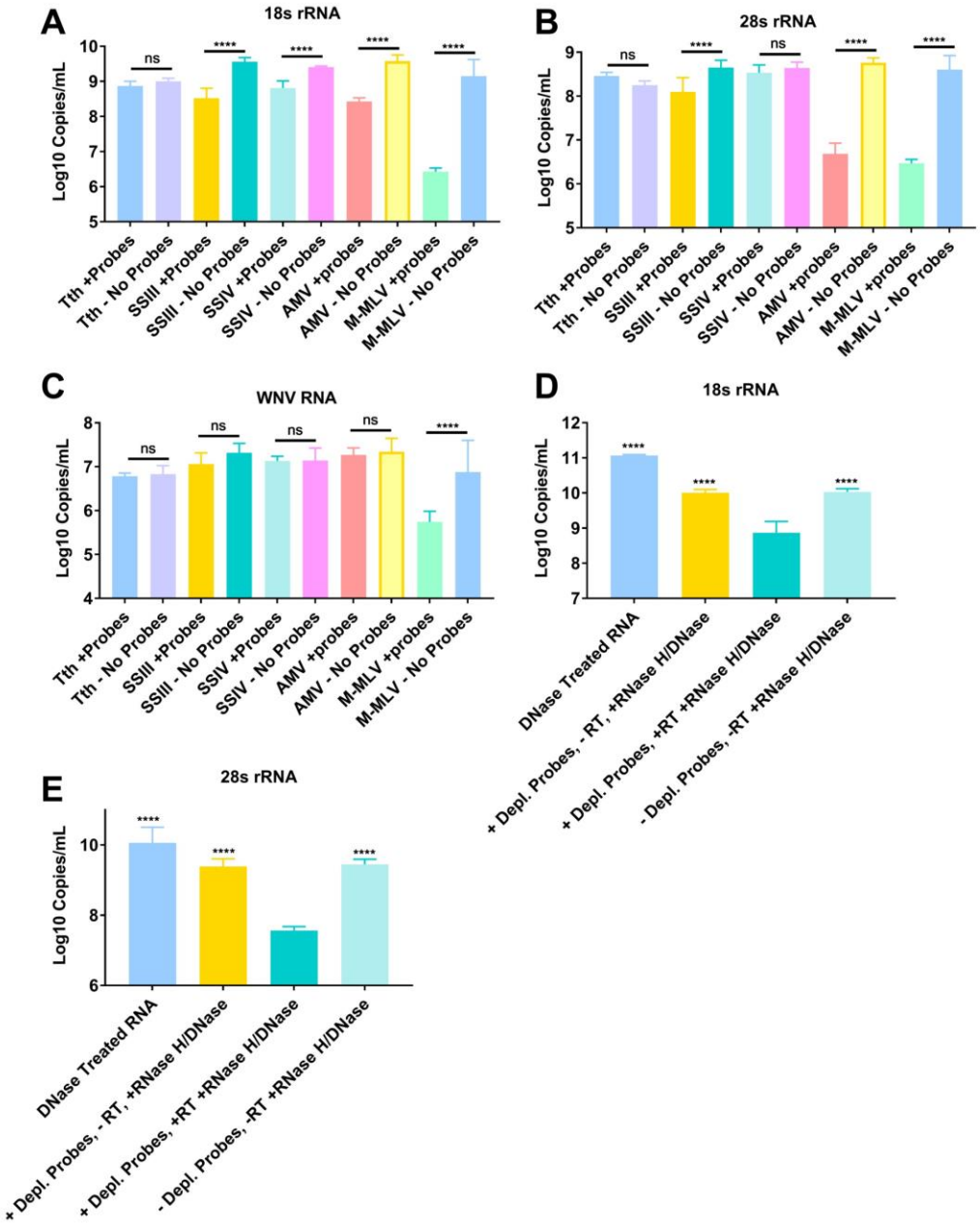
730

731 **Figure 1**



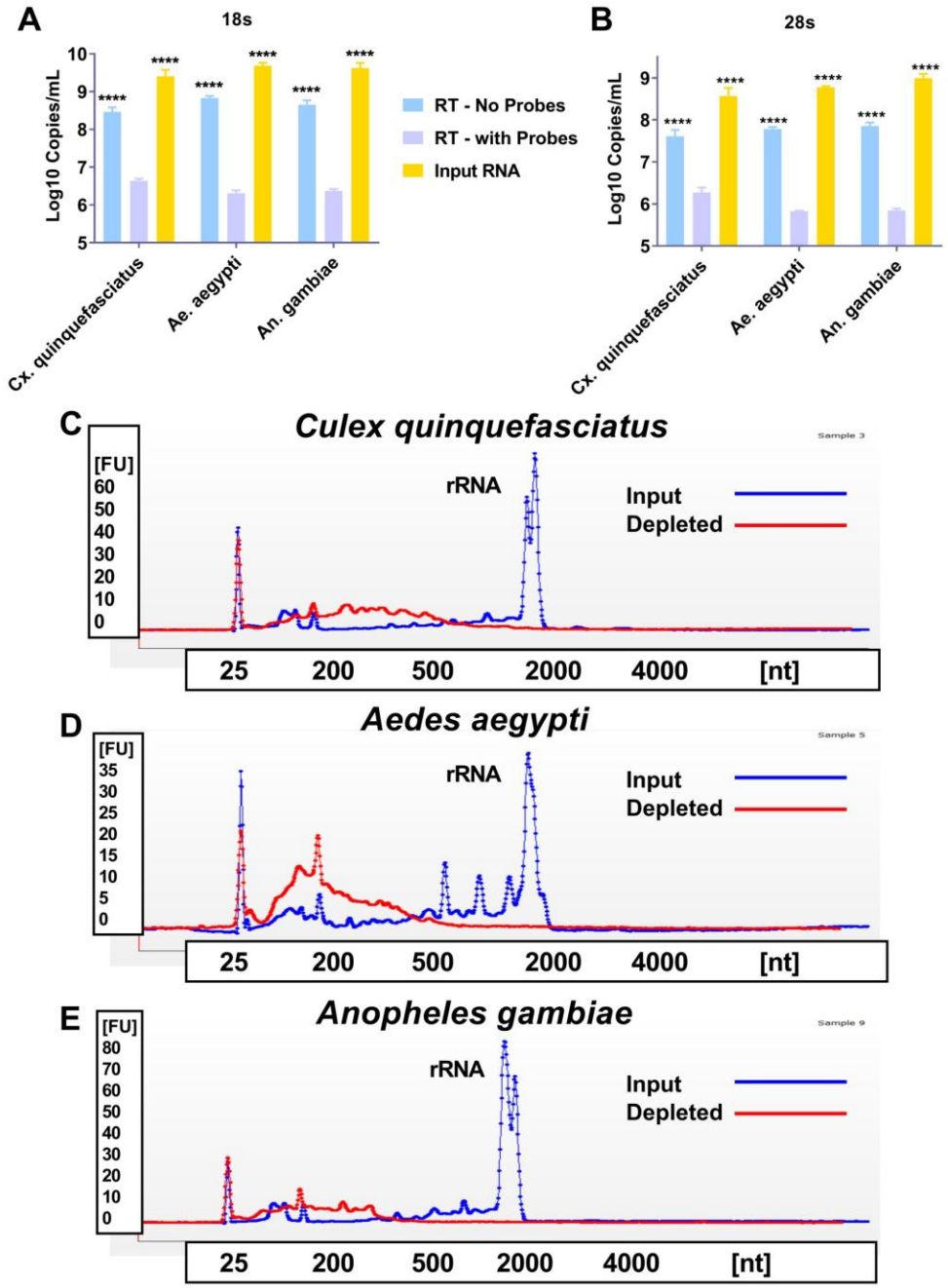
732
733
734

735 **Figure 2**



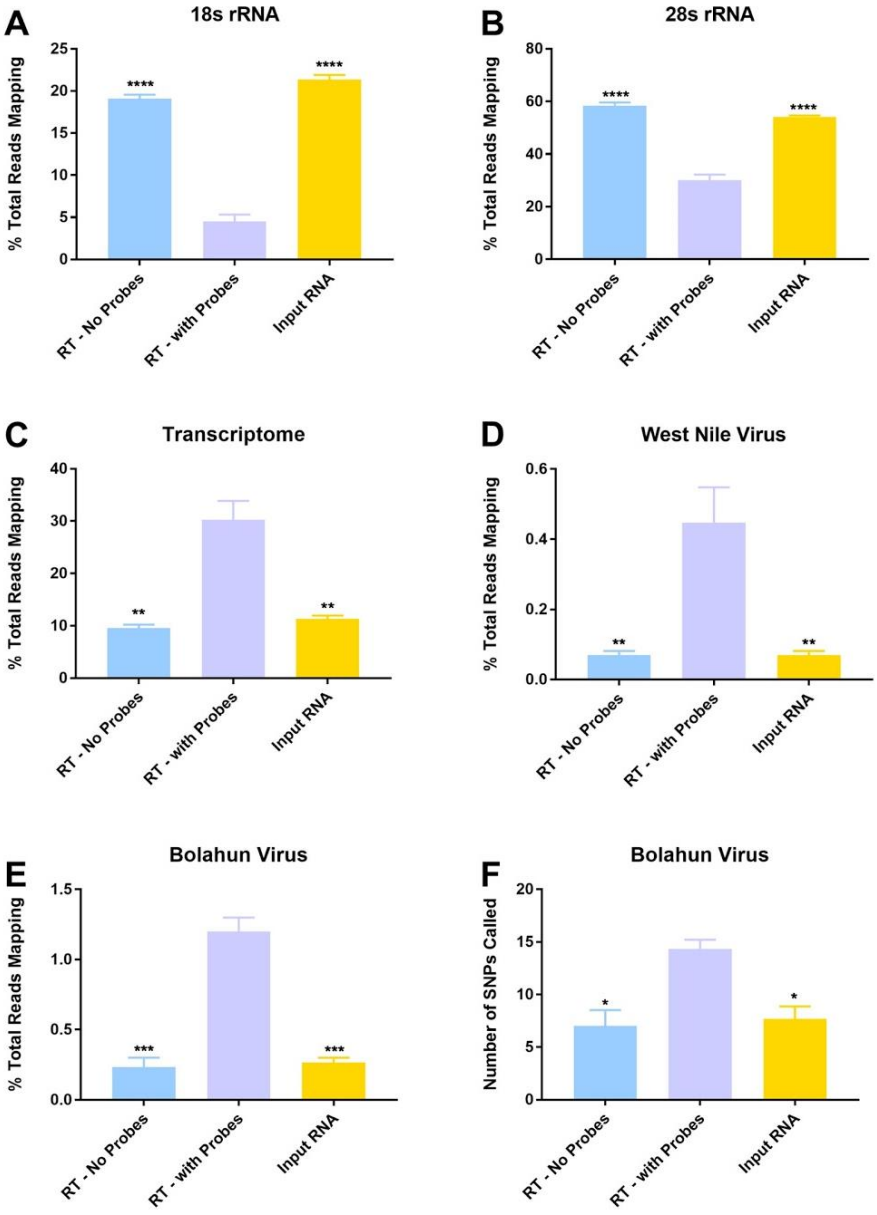
736
737

738 **Figure 3**



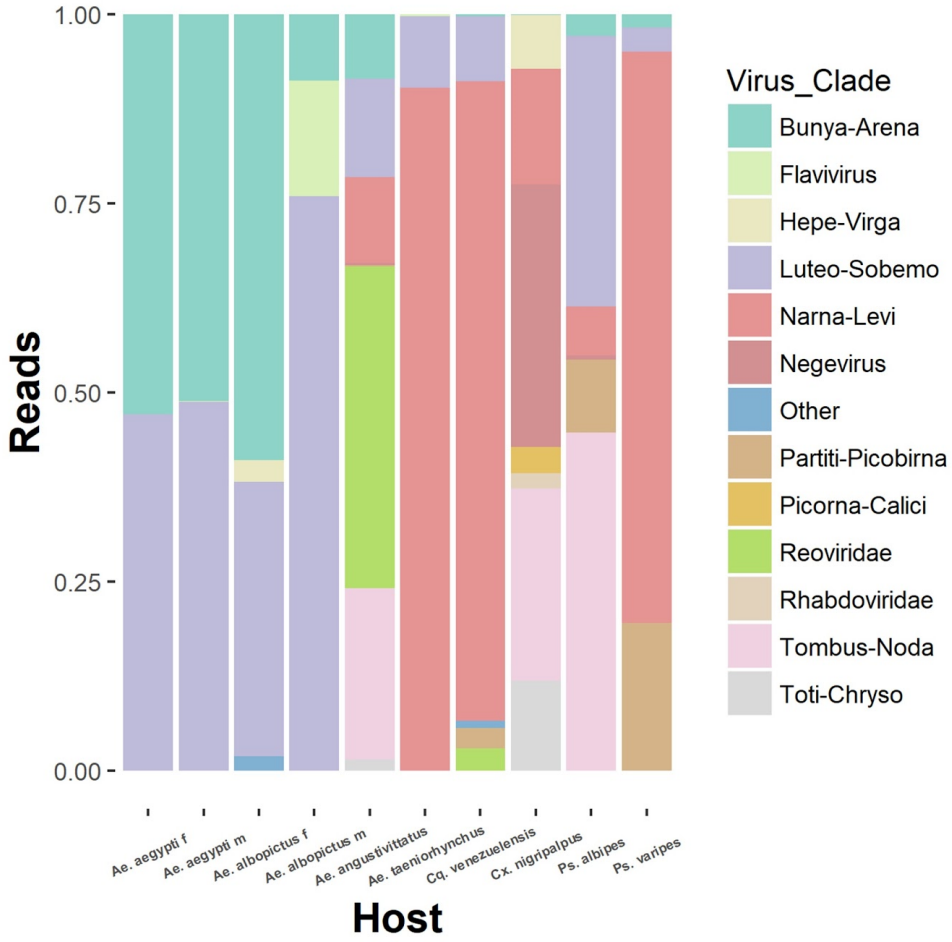
739
740

741 **Figure 4**



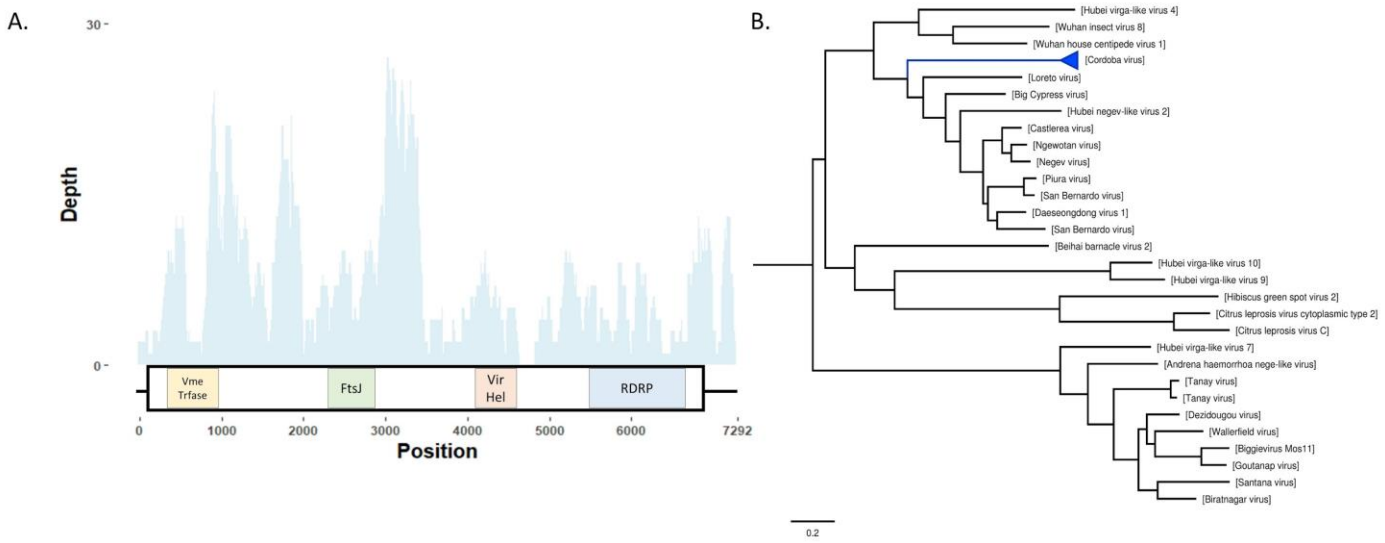
742
743

744 **Figure 5**

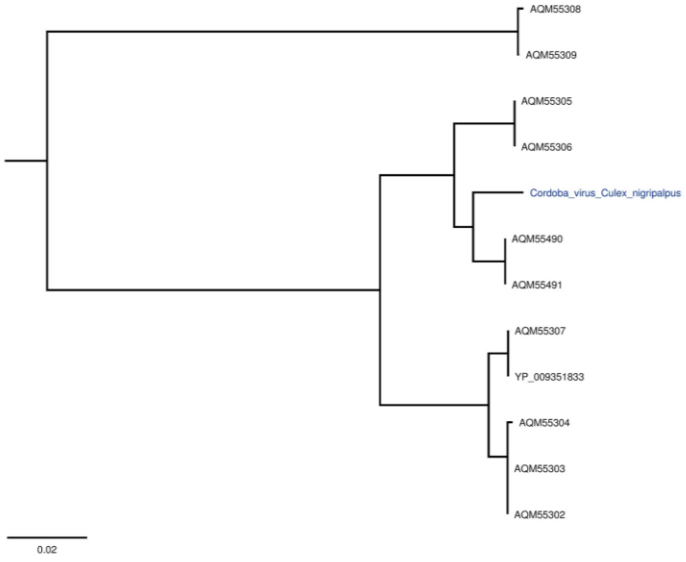


745
746

747 **Figure 6**

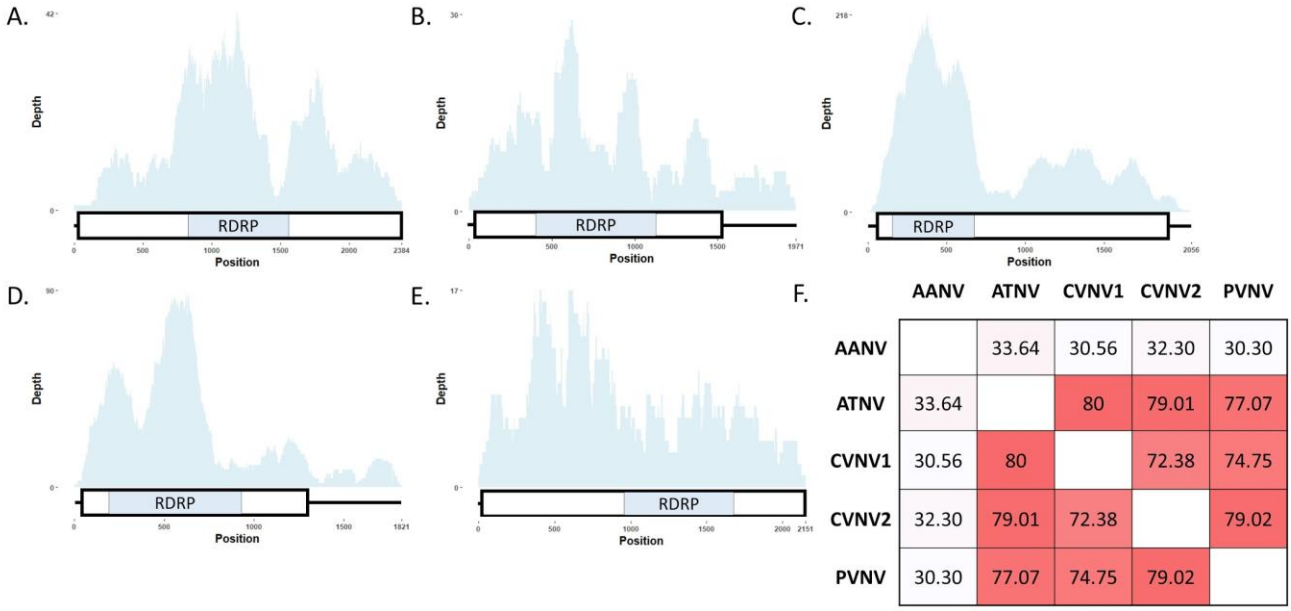


748
749 C.
750
751



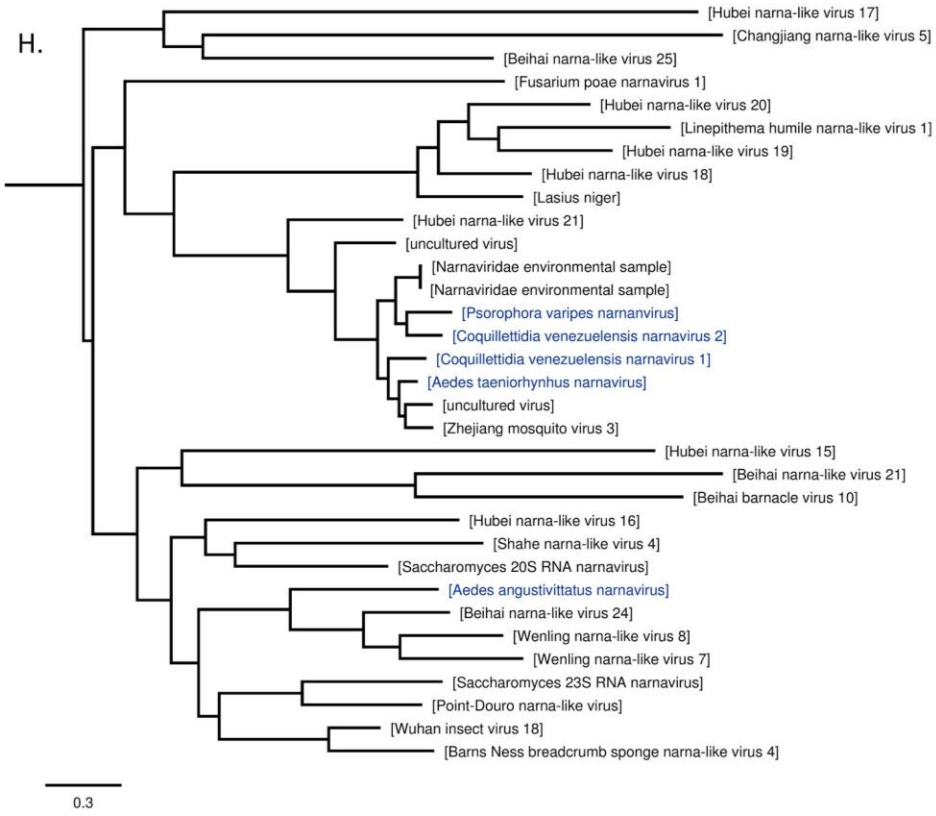
752

753 **Figure 7**



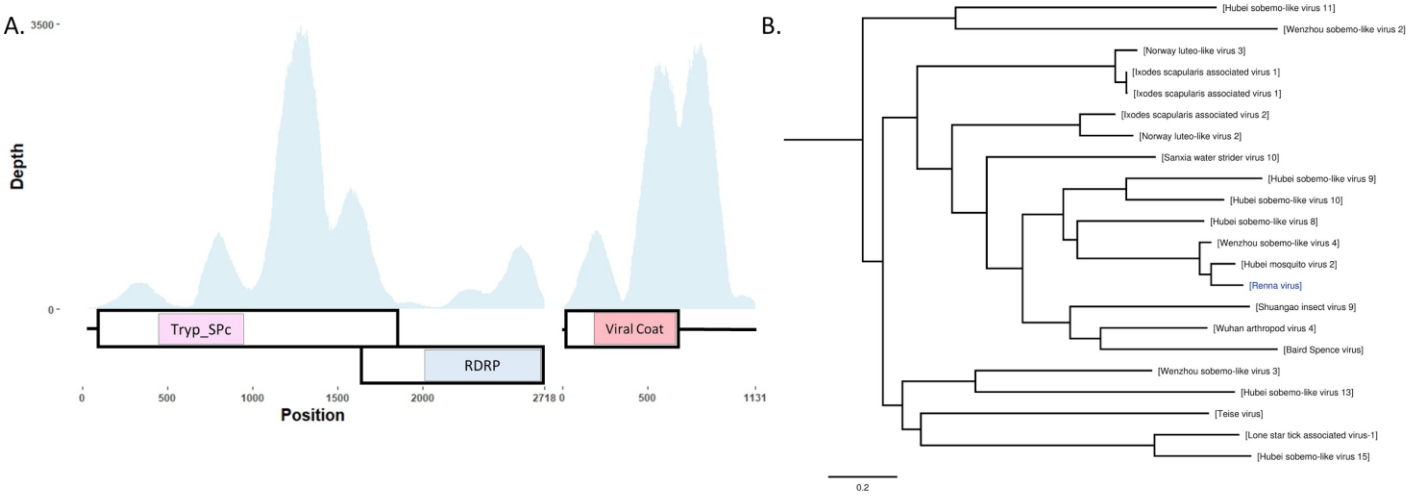
	AANV	ATNV	CVNV1	CVNV2	PVNV
AANV		33.64	30.56	32.30	30.30
ATNV	33.64		80	79.01	77.07
CVNV1	30.56	80		72.38	74.75
CVNV2	32.30	79.01	72.38		79.02
PVNV	30.30	77.07	74.75	79.02	

754
755



756

757 **Figure 8**



758