Single mosquito metatranscriptomics recovers mosquito species, blood meal sources, and microbial cargo, including viral dark matter

Joshua Batson[1†], Gytis Dudas[3†], Eric Haas-Stapleton[2†], Amy L. Kistler[1*†], Lucy M. Li[1†], Phoenix Logan[1†], Kalani Ratnasiri[1,4†], Hanna Retallack[5†]

***For correspondence:** amy.kistler@czbiohub.org (AK)

[†]These authors contributed equally to this work

[1]Chan Zuckerberg Biohub, 499 Illinois St, San Francisco CA 94158; [2]Alameda County Mosquito Abatement District, 23187 Connecticut St., Hayward, CA 94545; [3]Gothenburg Global Biodiversity Centre, Carl Skottsbergs gata 22B, 413 19, Gothenburg, Sweden; [4]Program in Immunology, Stanford University School of Medicine, Stanford CA 94305; [5]Department of Biochemistry and Biophysics, University of California San Francisco, San Francisco CA 94158

Abstract Mosquitoes are a disease vector with a complex ecology involving interactions between transmissible pathogens, endogenous microbiota, and human and animal blood meal sources. Unbiased metatranscriptomic sequencing of individual mosquitoes offers a straightforward and rapid way to characterize these dynamics. Here, we profile 148 diverse wild-caught mosquitoes collected in California, detecting sequences from eukaryotes, prokaryotes, and over 70 known and novel viral species. Because we sequenced singletons, it was possible to compute the prevalence of each microbe and recognize a high frequency of viral co-infection. By analyzing the pattern of co-occurrence of sequences across samples, we associated "dark matter" sequences with recognizable viral polymerases, and animal pathogens with specific blood meal sources. We were also able to detect frequent genetic reassortment events in a highly prevalent quaranjavirus undergoing a recent intercontinental sweep. In the context of an emerging disease, where knowledge about vectors, pathogens, and reservoirs is lacking, the approaches described here can provide actionable information for public health surveillance and intervention decisions.

Introduction

Mosquitoes are known to carry more than 20 different eukaryotic, prokaryotic, and viral agents that are pathogenic to humans (***WHO, 2017***). Infections by these mosquito-borne pathogens account for nearly half a billion human deaths per year, millions of disability-adjusted life years (***GBD 2017 Disease and Injury Incidence and Prevalence Collaborators, 2018***; ***GBD 2017 Causes of Death Collaborators, 2018***; ***GBD 2017 DALYs and HALE Collaborators, 2018***), and periodic die-offs of economically important domesticated animals (***Cohnstaedt, 2018***). Moreover, recent studies of global patterns of urbanization, warming, and the possibility of mosquito transport via long-range atmospheric wind patterns point to an increasing probability of a global expansion of

mosquito habitat and a potential concomitant rise in mosquito-borne diseases within the next 2-3 decades (***Kraemer et al., 2019***; ***Huestis et al., 2019***). While mosquito control has played a major role in eliminating transmission of these diseases in many parts of the world, costs and resources associated with basic control measures, combined with emerging pesticide resistance, pose a growing challenge in maintaining these gains (***Wilson et al., 2020***).

Female mosquitoes subsisting on blood meals from humans and diverse animals in their environment serve as a major source of trans-species introductions of infectious microbes. For well-studied mosquito-borne pathogens such as West Nile virus, an understanding of the transmission dynamics between animal reservoir, mosquito vector, and human hosts has been essential for public health monitoring and intervention (***Hofmeister, 2011***). For less well-studied microbes with pathogenic potential, the dynamics are less clear.

We also lack a comprehensive understanding of the composition of the endogenous mosquito microbiota and how it impacts the acquisition, maintenance, and transmission of pathogenic microbes. Evidence supporting the case for a potentially important role of endogenous mosquito microbiota on mosquito-borne infectious diseases stems from decades of research on *Wolbachia*, a highly prevalent bacterial endosymbiont of insects (***Werren et al., 2008***). *Wolbachia* have been shown to inhibit replication of various mosquito-borne viruses that are pathogenic to humans, including dengue, chikungunya, and Zika viruses, when introduced (or "transinfected") into susceptible mosquito species in which naturally occurring *Wolbachia* infections are rare or absent (***Moreira et al., 2009***).

These observations have led to the development of *Wolbachia*-based mosquito control programs for *Aedes aegypti* mosquitoes, which vector yellow fever virus, dengue virus, Zika virus, and chikungunya virus (reviewed in ***Ritchie et al., 2018***; ***Flores and O'Neill, 2018***). Experimental releases of *Aedes aegypti* mosquitoes transinfected with *Wolbachia* have resulted in a significant reduction in the incidence of dengue virus infections in local human populations in several countries (***Hoffmann et al., 2011***; ***O'Neill et al., 2019***; ***Anders et al., 2018***). Laboratory-based studies have identified additional endogenous mosquito microbes, such as midgut bacteria (***Shane et al., 2018***) and several insect-specific flaviviruses (***Vasilakis and Tesh, 2015***), with potential to interfere with mosquito acquisition and competence to transmit pathogenic *Plasmodium* species and human flaviviruses, respectively. However, definitive evidence for a role of these agents in naturally occurring infections or transmission of known human pathogens has not yet been established (reviewed in ***Bolling et al., 2015***).

Endogenous microbiota of mosquitoes could also provide a source of candidate surveillance biomarkers to follow mosquito movements, sharing of reservoirs, and potential transmission of co-occurring human pathogens carried by mosquitoes. Molecular methods for tracking that rely on mosquito genomic information are hampered by the long generation time of mosquitoes and the relatively large mosquito genome. Incorporating highly prevalent endogenous microbes of mosquitoes could simplify and refine such analyses. For example, RNA viruses have short generation times, extremely compact genomes, high mutation rates, and in some cases, undergo genetic reassortment. They are readily recovered in great numbers from mosquitoes and other insects via metagenomic sequencing (***Li et al., 2015***). Moreover, modest datasets of RNA virus sequences can reveal information that would be difficult to acquire through direct analysis of host populations. These features have enabled tracking of animal populations in the wild. For example, feline lentiviruses have been used to track native North American felids in the wild (***Wheeler et al., 2010***; ***Lee et al., 2014***), and rabies virus have been used to track dogs in North Africa (***Talbi et al., 2010***). Similar approaches may enhance mosquito tracking.

Such exciting possibilities have, in part, motivated a series of recent unbiased metagenomic analyses of pools of mosquitoes collected around the world (***Li et al., 2015***; ***Shi et al., 2015***, ***2016***; ***Fauver et al., 2016***; ***Shi et al., 2017***, ***2018***; ***Sadeghi et al., 2018***, ***2017***; ***Xia et al., 2018***; ***Thongsripong et al., 2018***; ***Atoni et al., 2018***; ***Xiao et al., 2018a***,b), (reviewed in ***Xia et al., 2018***; ***Atoni et al., 2019***). However, these studies have primarily focused on analysis of viruses. While these data have had

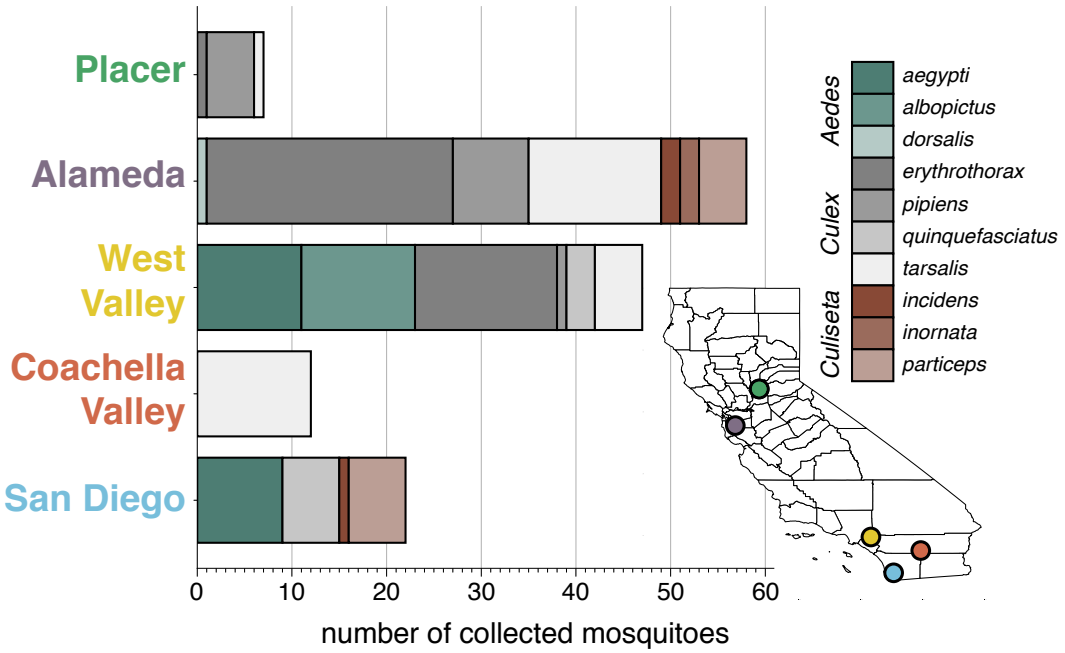


Figure 1. Plot shows number of each genus and species of mosquitoes collected across 5 regions in California: Placer, Alameda, West Valley, Coachella Valley, and San Diego. Map inset at lower right shows locations of collection sites. Colors of circles in the map correspond to colors of bar labels in plot at left.

Figure 1–source data 1. Metadata for each mosquito sample, including sampling location, genus and species, and RNA concentration.

Figure 1–source data 2. Per-mosquito sequence yields.

Figure 1–Figure supplement 1. Transcriptome-based evaluation of visual mosquito species calls.

Figure 1–source data 3. SNP distance matrix between samples.

a tremendous impact on our understanding of the breadth of viral diversity present in mosquito populations worldwide, they have not shed light on the potential reservoirs of these viruses or their prevalence within mosquito populations.

To link microbes with one another and with bloodmeal sources, single mosquito analyses are required. A handful of small-scale studies have demonstrated that it is possible to identify divergent viruses and evidence of other microbes in single mosquitoes by metagenomic next-generation sequencing (***Chandler et al., 2015***; ***Bigot et al., 2018***; ***Shi et al., 2019***). Here, we analysed the metatranscriptomes of 148 individual mosquitoes collected in California, USA, to characterize the composition of their microbiota, identify blood meal sources and associated pathogens, recover complete viral genomes, and infer connectivity between mosquito populations worldwide. This analysis crucially depended on the large number of individuals sequenced, which allowed us to link nucleic acid sequences detected together across many samples even when they had no homology to a reference sequence. Our findings demonstrate how large-scale single mosquito metatranscriptomics can provide insight into the mosquito's complex microbiota and contribute to the development and application of measures to control mosquito-borne pathogen transmission.

Results

Specimen demographics

Adult mosquito species circulating in California in late fall of 2017 were collected to acquire a diverse and representative set of 148 mosquitoes for analysis. We targeted collections across a variety of habitats within 5 geographically distinct counties in Northern and Southern California (***Figure 1***, ***Figure 1–source data 1***). Visual mosquito species identification was performed at the

time of collection (Materials and Methods), and specimens were skewed to include primarily female mosquitoes to enrich for blood-feeding members of the population responsible for transmission of animal and human diseases.

Total RNA from each mosquito was used as the input template for metatranscriptomic next generation sequencing (mNGS) library preparation to capture both polyadenylated and non-polyadenlylated host, viral, prokaryotic, and non-host eukaryotic RNA transcripts. On average, 52 million reads were obtained per mosquito (range, 11 million - 150 million reads, ***Figure 1–source data 2***). We validated the mosquito species identification made at the time of collection using hierarchical clustering of the complete metatranscriptomes of each mosquito through a reference-free, kmer-based approach for computing genomic distances (***Harris, 2018***). The computed species calls based on the metatranscriptomes agreed with the visual calls for most of the specimens (Materials and Methods, ***Figure 1–Figure Supplement 1***, ***Figure 1–Figure Supplement 1***). For the 8 specimens where discrepancies were detected, visual calls were reassigned to their computed calls ***Figure 1–source data 1***.

Read distribution among non-host contigs

To investigate the microbiota of each mosquito, host, low quality, and low complexity sequences were filtered from each mosquito (Materials and Methods, ***Figure 1–source data 2***). The 0.001% - 2.80% reads remaining for each sample were then separately *de novo* assembled. This yielded a total of 891,000 contigs for further analysis. Contigs were then aligned in parallel to the NCBI nt and nr databases. For all prokaryotic and eukaryotic species, only a fraction of a genome was recovered, thus a lowest common ancestor (LCA) analysis of the BLAST results was used to infer the taxa the contigs represented (Materials and Methods). For viral contigs, recovery of complete or near complete genomes of RNA viruses facilitated viral taxonomic assignment directly from their BLAST results. We further manually curated contigs likely to code for viral RNA-dependent RNA polymerases (RdRps) and contigs that were physically unlinked but showed evidence of co-occurrence with viral RdRps (see Methods). ***Figure 2*** shows the taxonomic distribution of reads that were assembled into contigs.

From the resulting input pool of just under 27 million reads, slightly more than 21 million (78%) assembled into contigs with more than two reads. A full 4.1 million reads were classified as *Hexapoda*, likely host sequences too diverged from reference to be recognized in the host filtering stage, and were excluded from further analysis. Metagenomic “dark matter”, *i.e.* reads in contigs without recognizable homology to previously published sequences, comprise 4.4 million reads. Of the remaining 12.7 million reads, 10.6 million were of viral origin. Eukaryotic and bacterial contigs comprised 0.9 and 0.7 million reads, respectively, with an additional 0.6 million reads not classifiable as either (either assigned to root or cellular organisms). We identified an additional 340 thousand viral reads from the “dark matter” using co-occurrence analysis. We thus were able to classify (77%) of the non-host reads which assembled into contigs.

Viruses comprise the vast majority of the reads in our data, with positive-sense RNA viruses (7.4 million), followed by negative-sense RNA viruses (2.25 million) and double-stranded RNA viruses (0.94 million) with a small portion (0.36 million) being clearly viral in origin but incomplete or not associated with an RdRp and are listed under the “uncurated viruses” category. Reads in this category were largely from RNA viruses of *Bunyavirales* order and *Flaviviridae* family, though small numbers of reads belonged to contigs most similar to DNA viruses, such as nucleocytoplasmic large DNA viruses, *Polydnaviridae*, *Alphabaculovirus*, and phages.

Within the viral taxa, we observed striking heterogeneity in read numbers, both within and across viral genome categories. This likely resulted not only from actual differences in amounts of virus carried within or on any given mosquito, but also from differences in viral copies in organisms carried on or in the mosquitoes. For example, the recently described Marma virus (***Pettersson et al., 2019***), a known virus of mosquitoes, comprised nearly a quarter (23.9%) of all viral reads. In contrast, six Botourmiavirus taxa, corresponding to viruses known to infect fungi, were detected at

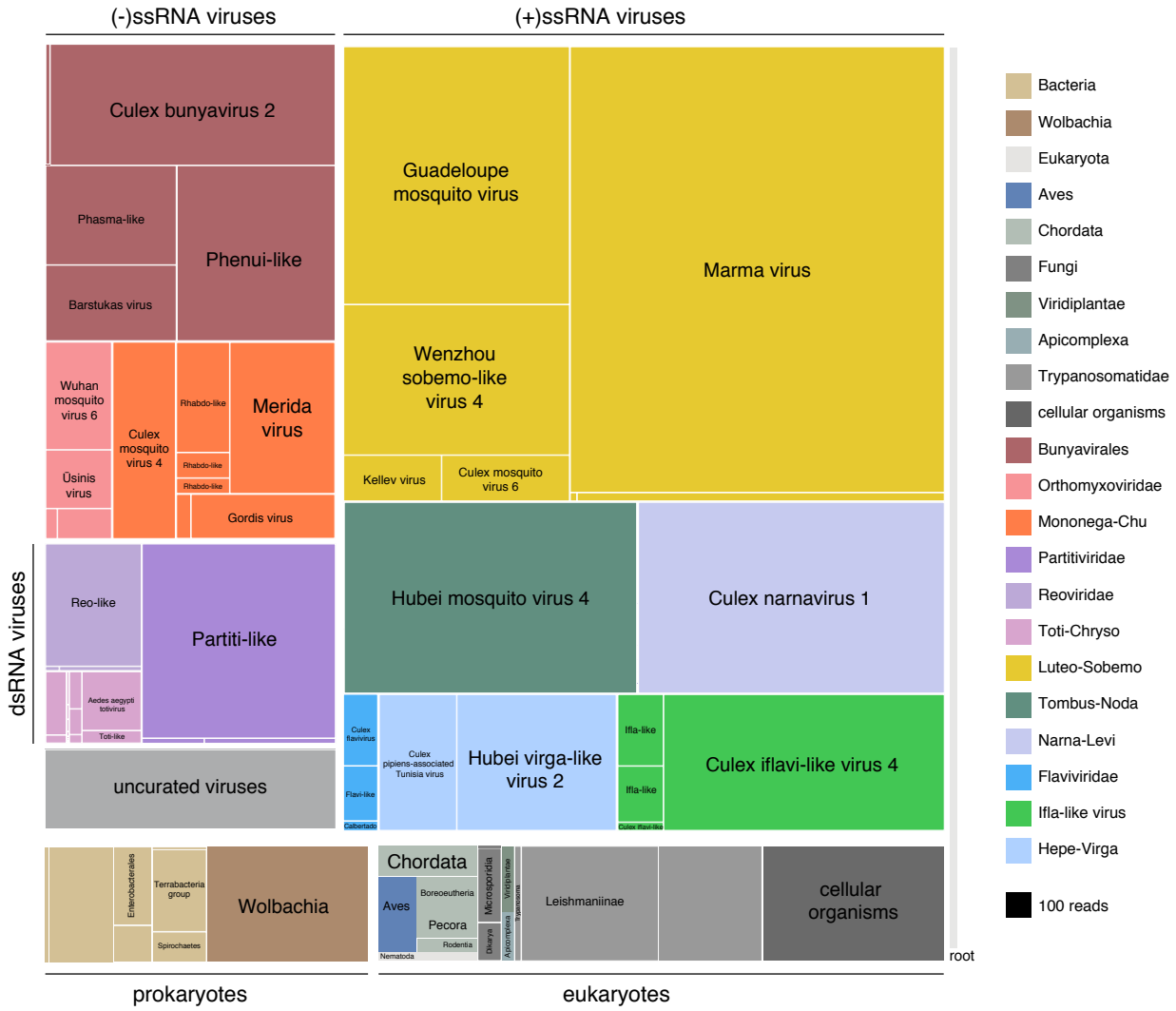


Figure 2. Read count distribution across non-host taxonomic groups detected among the set of 148 mosquitoes. Treemap overview of taxa detected among the assembled contigs are plotted in squares shaded according to taxon identity (legend) and sized according to number of reads supporting each contig (legend). Due to the uncertainty in prokaryotic and eukaryotic taxonomic assignments, cellular organisms are assigned to the lowest plausible taxonomic rank. Each distinct viral lineage is displayed as its own compartment. Reddish hues indicate negative-sense RNA viruses; purple hues indicate double-stranded RNA viruses; blue, green and yellow hues indicate positive-sense RNA viruses. The "uncurated viruses" box indicates contigs that were clearly viral in origin but were too fragmented and/or not associated with a RNA-dependent RNA polymerase.

Figure 2–source data 1. Treemap read counts per taxon per sample.

Figure 2–source data 2. Detected viral taxa.

Figure 2–Figure supplement 1. Peribunya virus assembled from a single mosquito.

Figure 2–source data 3. Data underlying peribunya phylogenetic tree and ends analysis.

very low levels (<200 reads each). All of the Botourmiaviruses were found in sample CMS002_053a, presumably corresponding to an infection of the ergot fungus (*Claviceps* sp.) that was detectable in the same sample.

Second in abundance by read number were bacterial taxa, with *Wolbachia* comprising most of the reads (*Wolbachia* endosymbiont of *Culex quinquefasciatus* with 0.24 million reads and *Wolbachia pipientis* with 11,000 reads). Various other bacterial taxa were detected at lower abundance (other members of Alphaproteobacteria, Gammaproteobacteria, Terrabacteria group, and Spirochaetes); *Spironema culicis* (18,000 reads) makes up 82% of Spirochaete reads.

Of the eukaryotic non-host sequences, members of *Trypanosomatidae* comprised 59% of reads (0.45 million). Approximately 55% (0.25 million) of these belonged to members of *Leishmaniinae*. The second most abundant group of eukaryotes detected in the dataset was the *Bilateria* (animals) taxa, with 0.19 million reads. This group was made up of the mammals (*Boreoeutheria*, 73,000 reads), birds (*Aves*, 48,000 reads), and invertebrates (*Ecdysozoa*, 14,000 reads). The reads derived from vertebrate taxa almost certainly belong to blood meal hosts, which we investigate in detail below. Nematodes (*Nematoda*) make up a substantial portion (5,000 reads) of the invertebrate taxa.

The third largest eukaryotic category by read numbers was fungi (44,000 reads), with *Microsporidia* (29,000 reads) being the largest fungal group. The *Microsporidia* largely comprised genus-level *Amblyospora* taxa (15,000 reads). A mixture of subkingdom-level *Dikarya* reads (15,000 reads) were also observed, of which nearly a half can be assigned to the species *Lachancea thermotolerans* (6,700 reads). Plants (*Viridiplantae*, 13,000 reads) and *Apicomplexa* (9,600 reads) comprise the last two categories of eukaryotes with notable read prevalence in our data. Many contigs could only be assigned at the highest taxonomic levels: cellular organisms (0.28 million), bacteria (0.12 million), and eukaryotes (50,000 reads).

Quantifying yield of new genetic information

From the 148 mosquitoes sequenced for this study, a diverse set of 70 unique viral taxa were recovered with identifiable RdRp sequences and complete or near-complete genomes (***Figure 2–source data 2***). These viral taxa correspond primarily to RNA viruses and encompass a wide variety of genome types. Twenty-one of these viruses are closely related or identical to previously identified mosquito viruses: 8 correspond to viruses previously detected in California mosquitoes (***Sadeghi et al., 2018***; ***Chandler et al., 2015***; ***Tyler et al., 2011***), and 13 correspond to viruses identified from mosquito specimens collected outside of California (***Göertz et al., 2019***; ***Shi et al., 2019***; ***Parry and Asgari, 2018***; ***Waldron et al., 2018***; ***Bigot et al., 2018***; ***Shi et al., 2016***). The remaining 50 viral genomes share less than 85% amino acid sequence identity to any publicly available viral sequences. Despite this divergence, family-level features such as conserved sequences at the 5' and 3' ends of bunyavirus segments allowed us to demonstrate complete genome recovery of a novel peribunya-like virus from a single mosquito at relatively low abundance (4.9% of non-host reads) (***Figure 2–Figure Supplement 1***). To reduce the likelihood of mis-annotating non-infectious endogenous viral elements (EVEs) present within eukaryotic genomes (***Ter Horst et al., 2019***; ***Palatini et al., 2017***; ***François et al., 2016***; ***Katzourakis and Gifford, 2010***), the set of viruses considered here was restricted to those with complete or near-complete genomes. Nevertheless, it is difficult to rule out this possibility with mNGS alone and we identify two viral RdRp-like sequences, RdRp group 88, Tombus-like virus, and RdRp group 246, Reo-like virus, which fall into clades containing both replicating viruses and sequences proposed to be integrated in the genome of *Leptomonas* trypanosomatid and *Ochlerotatus* mosquito, respectively (***Grybchuk et al., 2018***; ***Cook et al., 2013***).

We quantified the evolutionary novelty of these genome sequences by the total amount of branch length (in expected amino acid substitutions per site) contributed per viral lineage not derived from previously published sequences (***Figure 3***). We define this contribution of evolutionary novelty (CEN) as branch lengths of a substitution phylogeny (inferred from an untrimmed alignment) not visited after tip-to-root traversals from every background (*i.e.* non-study) sequence (***Figure 3–Figure Supplement 1***). Since branch lengths in molecular phylogenies represent discrete and

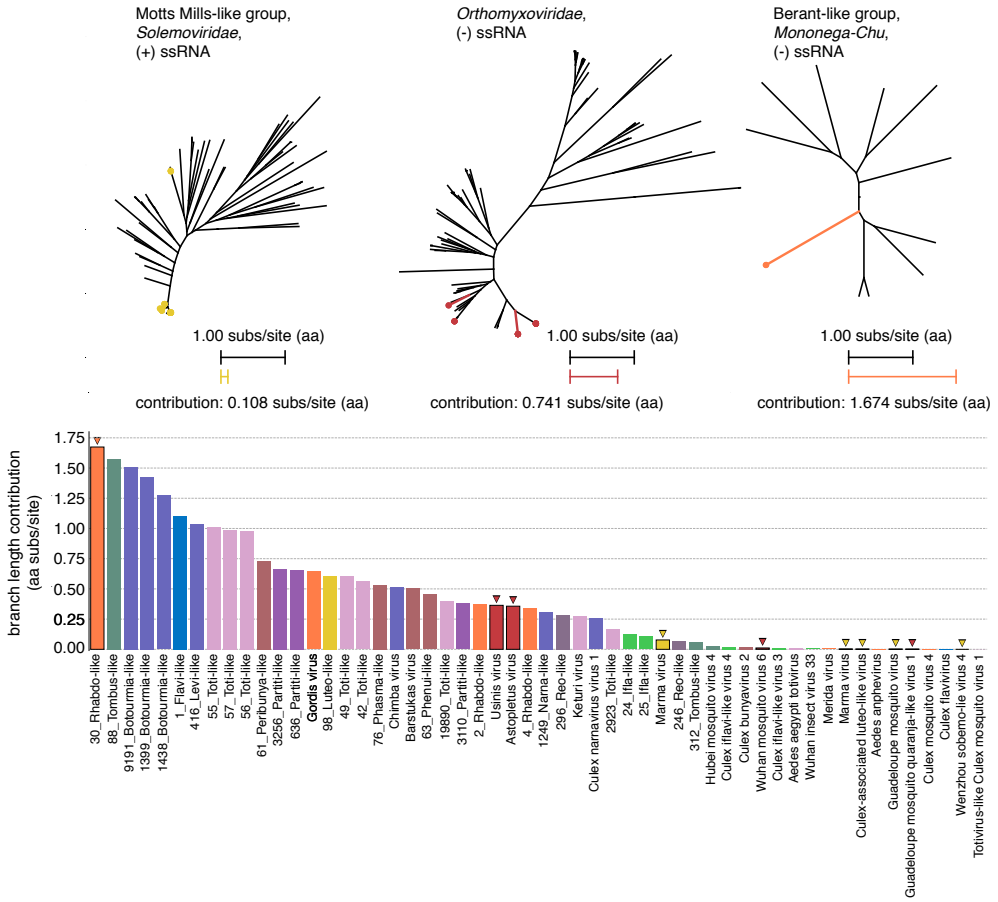


Figure 3. Top panel, A collection of three unrooted trees representing varying degrees of evolutionary novelty contribution from the current study - four lineages belonging to a subgroup of *Solemoviridae* related to Motts Mills virus, four lineages of Quaranjavirus in *Orthomyxoviridae* and one lineage distantly related to Berant virus in *Mononegavirales*. Scale bars under each tree indicate the distance corresponding to 1.0 amino acid substitutions per site (in black) as well as branch length contribution of lineages in the tree from this study (in color). Bottom panel, Bar chart depicting the evolutionary novelty contribution from each lineage described here, colored by viral family/group (same color scheme as ***Figure 2***). Bars outlined in black with a triangle above indicate viral lineages displayed in the trees above.

Figure 3–Figure supplement 1. Schematic of branch length contribution.

Figure 3–source data 1. Phylogenies for each viral group with branch length contributions. Phylogeny at the top indicates previously reported relationships between viral groups whose trees are shown underneath. Phylogenies of untrimmed amino acid sequences with study sequences and their branch length contributions highlighted in red.

Figure 3–source data 2. Amino acid alignments and the phylogenies inferred from them that were used in branch length contribution calculation.

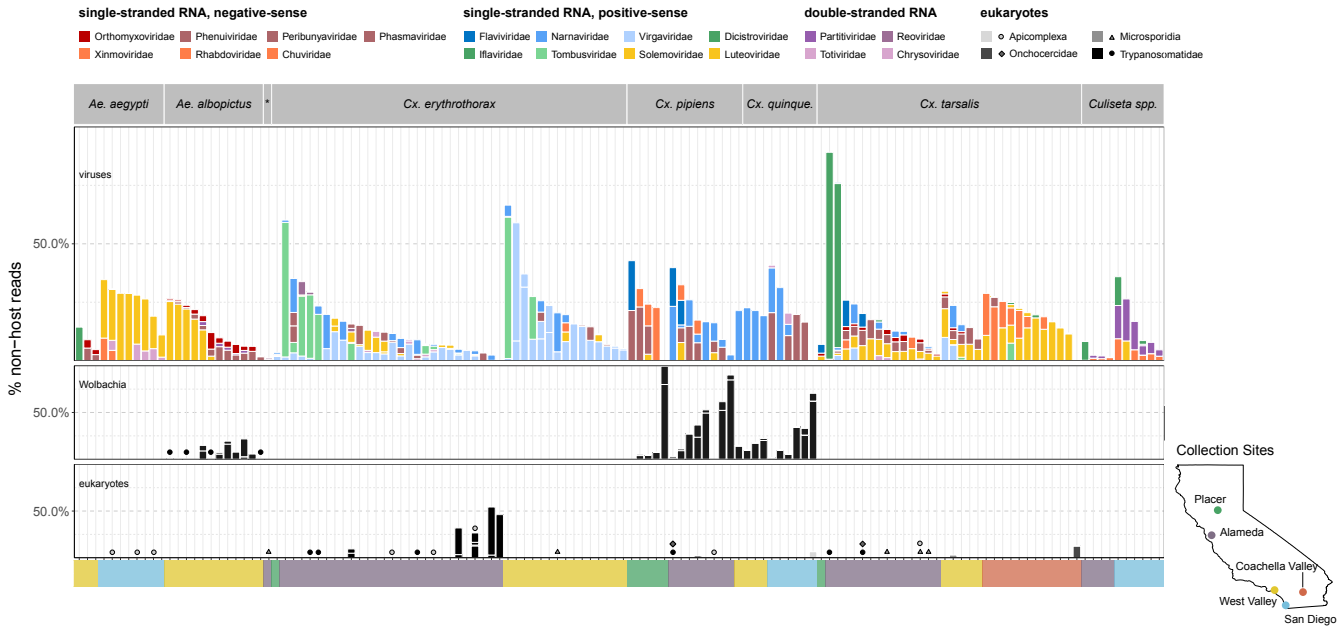


Figure 4. Individual mosquito proportions of non-host reads supporting assembled contigs corresponding to detectable viruses (top panel), *Wolbachia* (middle panel), and selected eukaryotic microbes (bottom panel) supported by $\geq 1\%$ of non-host reads (plotted bars). Parallel symbols plotted on *Wolbachia* and eukaryotic microbes panels indicate microbes detected at $\leq 1\%$ (middle panel, black circles = *Wolbachia* taxa; bottom panel, black circles = *Trypanosomatidae* taxa; dark grey diamonds = *Nematoda* taxa; grey triangles = *Microsporidia* taxa; light grey circles outlined with black = *Apicomplexa* taxa). Samples were aggregated by mosquito species (top labels), and ordered by collection site location from north to south (see x-axis color bar at the below bottom panel and inset California map at bottom right of plot), and viral abundance (descending order, left to right).

Figure 4–Figure supplement 1. Frequency of viral co-infections.

Figure 4–Figure supplement 2. Prevalence versus abundance.

Figure 4–source data 1. Read counts per taxon per sample for displayed taxa.

Figure 4–source data 2. Assignment of each viral contig to a cluster, each cluster to a viral segment, and each viral segment to a corresponding polymerase group.

independent periods of evolution, CEN is additive over studies (so long as the tree topologies agree). In contrast, statistics from pairwise comparisons, such as percent identities to the closest previously described sequence, will depend on the order in which new genomes are considered.

According to this metric, our data contribute a range of evolutionary novelty across different viral families. While most viral lineages detected here contribute only modest amounts of branch length, highly diverged lineages were readily identified, despite mosquitoes being a frequent target of metagenomic studies.

Microbial cargo of single mosquitoes

We next delved into the single mosquito distribution of the viral, prokaryotic, and eukaryotic taxa detectable across the specimens included in this study. All confidently called contigs for microbial taxa detected within each mosquito were compiled, and the fraction of non-host reads aligning to each contig was computed to estimate the composition and proportions of microbial agents detectable within each mosquito. ***Figure 4*** displays those agents detectable at a level above 1% of non-host reads plotted as bars. Confidently called contigs with less than 1% non-host read support are plotted as symbols above the x-axis coordinate for each relevant mosquito sample.

Viruses detected in single mosquitoes

Examining first the 70 unique viral taxa on a single mosquito basis (***Figure 2–source data 2***). Among single mosquitoes, co-infections predominated, with 120/136 mosquitoes harboring 2 or more distinct viral taxa (***Figure 4–Figure Supplement 1***). This insight is only possible by analyzing mosquitoes individually rather than in bulk. Furthermore, single mosquito analysis also highlighted the variability in the composition and amount of viral reads both within and across mosquito species, and their corresponding collection sites (***Figure 4***, top panel). We find that the average abundance of a virus (*i.e.* the average number of reads across a set of mosquitoes) is not necessarily predictive of the prevalence of that virus (*i.e.* the number of mosquitoes in which it occurs). For example, Culex narnavirus 1 and Culex pipiens-associated Tunisia virus were found in similar abundances in *Culex erythrothorax* mosquitoes at the same collection site in West Valley, but the latter was 3 times more prevalent (***Figure 4–Figure Supplement 2***). The differentiation between abundance and prevalence would not be possible without single mosquito sequencing.

Prokaryotes detected in single mosquitoes

We restricted our single mosquito analysis of prokaryote infections to *Wolbachia*, which was the most abundant prokaryotic taxon detected in this study (***Figure 2***) and a known endosymbiont of mosquitoes that could impact the microbiota of its mosquito hosts (***Werren et al., 2008***). We detected Wolbachia in 32 mosquitoes belonging to *Culex quinquefasciatus*, *Culex pipiens*, and *Aedes albopictus* species. These observations are consistent with previous observations of wild-caught mosquito species that are naturally infected with *Wolbachia* (***Rasgon and Scott, 2004***; ***Kittayapong et al., 2000***). Among these 3 species, *Wolbachia* was detected in all or nearly all of the mosquitoes. However, the fraction of non-host reads assigned to Wolbachia per mosquito varied from <1% to 99% (***Figure 4***, middle panel, black bar plots and circle symbols). Since all or nearly all the individual mosquitoes among the 3 species were Wolbachia-positive, it was not possible to investigate how the presence of different levels of *Wolbachia*, or even the presence or absence of *Wolbachia*, influenced the composition or relative amount of co-occurring viral taxa detectable among these mosquitoes.

Eukaryotic microbes detected in single mosquitoes

For analysis of eukaryotes, we focused on: *Trypanosomatidae*, the most abundant eukaryotic taxon detected and containing established pathogens of both humans and birds, *Microsporidia*, a fungal taxon known to infect mosquitoes, *Apicomplexa*, which encompasses the causative agents of human and avian malaria, *Nematoda*, which contain filarial species that cause heart worm in canines and filarial diseases in humans.

Twelve mosquitoes were found to harbor *Trypanosomatidae* taxa. We detected sequences corresponding to monoxenous (*e.g. Crithidia* and Leptomonas species), dixenous (*Trypanosoma*, *Leishmania* species), as well as the more recently described *Paratrypanosoma confusum* species. Eight of the *Trypanosomatidae*-positive mosquitoes corresponded to *Culex erythrothorax* mosquitoes that were all collected from the same trap site in Alameda County at different times (***Figure 4***, ***Figure 4–source data 1***,***Figure 1–source data 1***). The four other *Trypanosomatidae*-positives corresponded to 2 different species of mosquitoes (2 *Culex pipiens* and 2 *Culex tarsalis*).

After the *Trypanosomatidae* taxa, the next most abundantly detected eukaryotic microbial taxa among the mosquitoes corresponded to the fungal taxa *Microsporidia* (***Figure 2***). Single mosquito analysis indicated that 6 mosquitoes (4 *Culex tarsalis* and 2 *Culex erythrothorax*) harbor *Microsporidia* contigs. However, only a single *Culex tarsalis* mosquito collected in West Valley was found to have > 1% non-host reads assigned to this taxon (***Figure 4***, bottom panel, gray bar plot and triangle symbols).

We also investigated the single mosquito distribution of the *Apicomplexa* contigs and reads we observed within our dataset. This phylum encompasses the *Plasmodium* genus, which includes several pathogenic species that cause avian and human malaria. Single mosquito analysis identified 9 mosquitoes with *Apicomplexa* contigs. These corresponded to 3 *Aedes aegypti* mosquitoes and 1

Culex quinquefasciatus mosquito, both collected in San Diego, and 3 *Culex erythrothorax* mosquitoes, 1 *Culex pipiens* mosquito, and 1 *Culex tarsalis* mosquito, collected in Alameda County. Only the *Culex quinquefasciatus* mosquito harbored *Apicomplexa* reads at a level above 1% non-host (***Figure 4***, bottom panel, light gray bar plot and circle symbols outlined in black). Interestingly, this mosquito was found to also harbor the second highest fraction of non-host *Wolbachia* reads (70%), but no viruses.

Finally, we examined taxa falling under *Nematoda*, a phylum which encompasses a diverse set of more than 50 filarial parasites of humans and animals. Here, we saw evidence of *Nematoda* carriage in 3 *Culex* mosquitoes: 2 *Culex tarsalis* mosquitoes and 1 *Culex pipiens* mosquito (***Figure 4***, bottom panel, dark gray bar plot and diamond symbols). Two of these mosquitoes were collected in Alameda County, and showed very low levels *Nematoda* ($\leq 1\%$ of non-host reads, ***Figure 4***, bottom panel, diamond symbols). In the third mosquito, a *Culex tarsalis* collected in Coachella Valley, the *Nematoda* make up 12% of the non-host reads.

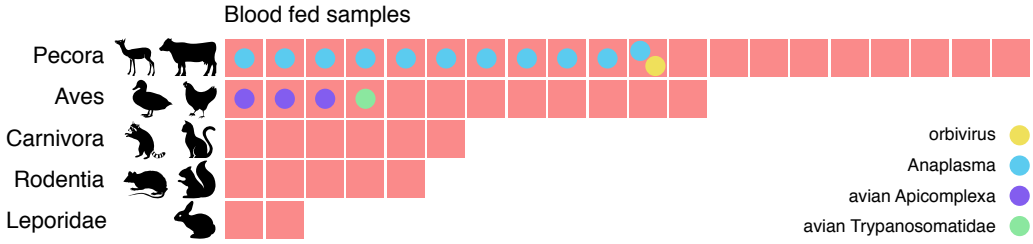


Figure 5. Consensus taxonomic calls of vertebrate contigs for 45 of 60 bloodfed mosquitoes collected in Alameda County. The remaining 15 samples had no vertebrate contigs. Red blocks represent individual mosquito samples; colored circles represent co-occurring contigs matching Orbivirus, Anaplasma, Avian *Apicomplexa* and Avian *Trypanosomatidae* representing possible bloodborne pathogens of the blood meal host.

Figure 5–source data 1. Table of blood meal calls for each sample. The most specific call (e.g. *Sciurini*) is given alongside the higher category displayed in ***Figure 5*** (e.g. *Rodentia*).

Figure 5–source data 2. Table of top BLAST hits for contigs representing potential blood meal parasites, annotated with sample source or known host of parasite sequence.

Blood meals and associated microbes

We next investigated the possibility of identifying the blood meal host directly from metatranscriptomic sequencing. We restrict this analysis to the 60 mosquitoes from Alameda County, as they were selected for visible blood-engorgement. For 45 of the 60 mosquitoes, there was at least one contig with an LCA assignment to the phylum *Vertebrata* (range = 1-11 contigs, with 4-12,171 supporting reads). To assign a blood meal host for each of these mosquitoes, we compiled their corresponding *Vertebrata* contigs and selected the lowest taxonomic group consistent with those contigs (***Figure 5–source data 1***). For all samples, the blood meal call fell into one of five broad categories, as shown in ***Figure 5***: even-toed ungulates (*Pecora*), birds (*Aves*), carnivores (*Carnivora*), rodents (*Rodentia*), and rabbits (*Leporidae*). For 10 samples, this call was at the species level, including rabbit (*Oryctolagus cuniculus*), mallard duck (*Anas platyrhynchos*), and raccoon (*Procyon lotor*).

These findings are broadly consistent with the habitats where the mosquitoes were collected. For the 25 samples collected in or near the marshlands of Coyote Hills Regional Park, we compare our calls to the wildlife observations in iNaturalist, a citizen science project for mapping and sharing observations of biodiversity. iNaturalist reports observations consistent with all five categories, including various species of squirrel, rabbit, raccoon, muskrat, and mule deer. The mosquitoes with blood meals in *Pecora* are likely feeding on mule deer, as no other ungulate commonly resides in that marsh (***iNaturalist, 2020***).

We also investigated whether bloodborne pathogens of the blood meal were detectable. We performed a hypergeometric test for association between each blood meal category and each

microbial taxon. The only statistically significant association (p = 0.0005, Bonferroni corrected) was between *Pecora* and *Anaplasma*, an intracellular erythroparasite transmitted by ticks. *Anaplasma* was detected in 11 of the 20 samples with *Pecora*, indicating a possible burden of anaplasmosis in the local deer population. Additionally, we detected evidence for three other bloodborne pathogens which, because of the small number of observations, could not pass the threshold of statistical significance. These include an orbivirus closely related to those known to infect deer (***Cooper et al., 2014***; ***Ahasan et al., 2019a***,b), a *Trypanosoma* species previously found in birds (***Zídková et al., 2012***), and the apicomplexans *Plasmodium* and *Eimeria* from species known to infect birds (***Carlson et al., 2018***; ***Harl et al., 2019***) (See Methods). The likely hosts of these pathogens were also concordant with the blood meal calls ***Figure 5–source data 2***.

Recovery of previously unrecognizable viral genome segments

Although many new viruses can be identified in bulk samples, the majority of these are identified only via their conserved RdRp (***Li et al., 2015***; ***Shi et al., 2017***, ***2019***; ***Pettersson et al., 2019***). Recovering complete genomes for segmented viruses from bulk samples is challenging, as genes which are not highly conserved may be unrecognizable by sequence homology. Moreover, the assignment of putative segments to a single genome can be confounded if a mosquito pool contains mosquitoes with multiple infections of related segmented viruses.

By sequencing many single mosquitoes, we can exploit the fact that all segments of a segmented virus will co-occur in the samples where that virus is present and be absent in samples where the virus is absent, enabling identification of previously unindentified viral genome segments. To do so, we first grouped all contigs longer than 500 nucleotides from the study into clusters of highly homologous contigs, then grouped these clusters by co-occurrence across all the samples. This requires only the sequence information from the study, and does not use any external reference. We then scanned each cluster for sequences containing a viral RdRp domain (see Methods). For each RdRp cluster, we consider any other contig cluster whose sample group overlaps the set of samples in the viral RdRp cluster above a threshold of 80% as a putative segment of the corresponding virus (***Figure 6–Figure Supplement 2***). We show a cluster-by-sample heatmap for all segments co-occurring with RdRps in ***Figure 6–Figure Supplement 1***. This resulted in 27 candidate complete genomes for segmented viruses. For the 79 of the 96 putative segments recognizable by homology to published sequences (colored in black), these groupings into genomes were accurate. This supports the notion that the remaining 17 putative segments (colored in red), which lack homology to any known sequences at either nucleotide or amino acid level, may indeed be part of viral genomes. These putative segments represented 8% of the “dark matter” portion of the reads in the study.

Orthomyxoviruses

Detailed co-occurrence analysis for orthomyxoviruses is shown in ***Figure 6***. These are a family of multi-segmented viruses containing influenza viruses, isaviruses, thogotoviruses, and quaranjaviruses that infect a range of mammalian, avian, and fish species. Many quaranjaviruses infect arthopods, and in this study, we identified four quaranjaviruses, two of which were previously observed in mosquitoes collected outside California (Wuhan Mosquito Virus 6 (WMV6) (***Li et al., 2015***; ***Shi et al., 2017***) and Guadeloupe mosquito quaranja-like virus 1 (GMQV1) (***Shi et al., 2019***)). While influenza viruses are known to have 7 or 8 segments, the published genomes for quaranjaviruses contained 6 or fewer segments.

For the WMV6 and GMQV1 isolates detected here, we observed all of the previously identified segments (PB1, PB2, PA, and NP segments for both; and two additional segments, gp64 and “hypothetical”, for WMV6). We moreover found evidence for two distinct putative segments, which we name hypothetical 2 and hypothetical 3. We confirmed the existence of the 2 putative segments for WMV6 by assembling homologous segments from reads in the two previously published datasets describing this virus (***Li et al., 2015***; ***Shi et al., 2017***). For GMQV1, we were able to find reads in the

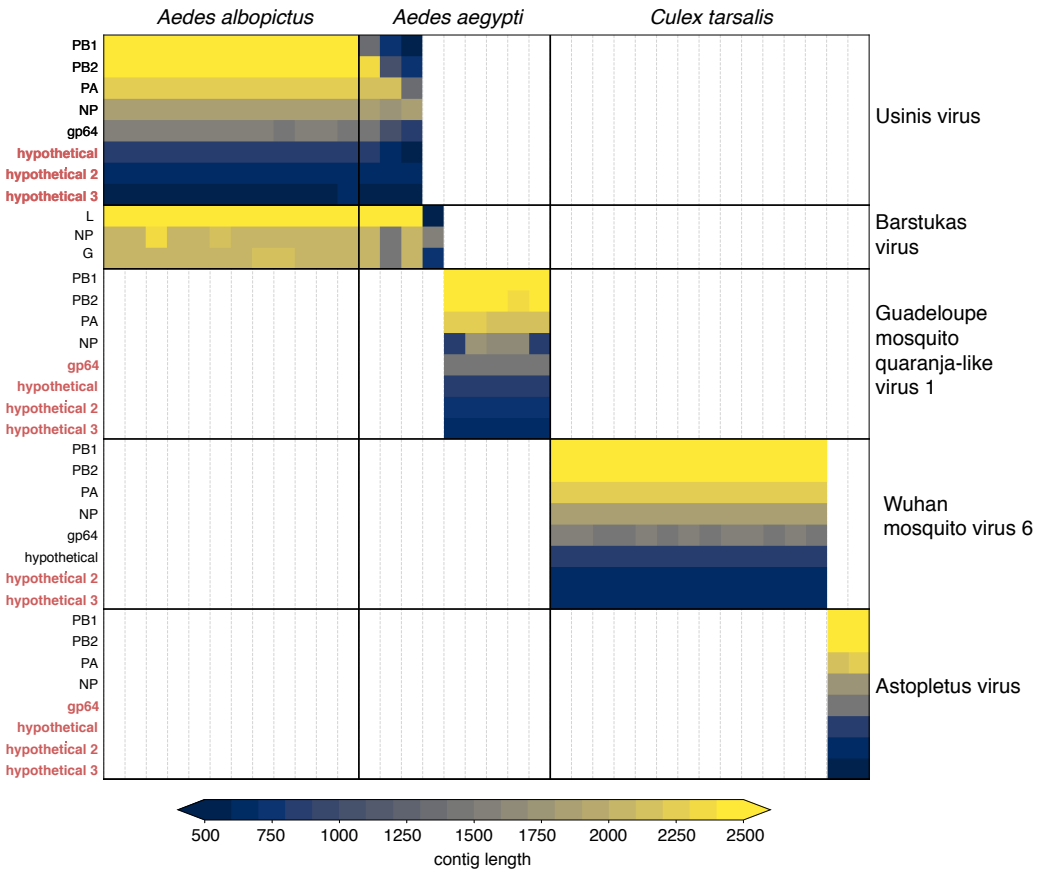


Figure 6. Previously unrecognized Orthomyxovirus genome segments identified among unaligned "dark matter" contigs using co-occurrence analysis. Matrix of contigs derived from four distinct Orthomyxoviruses and one Phasma-like virus that were detected via their distinct co-occurrence pattern across mosquitoes. Rows are clusters of highly similar (>99% identity) contigs and columns are individual mosquito samples. Light grey vertical lines delineate mosquito samples, dark black vertical lines indicate boundaries between mosquito species of each sample. Dark horizontal lines delineate segments comprising viral genomes. Labels on the right indicate viruses, with genomes delineated by horizontal lines. Guadeloupe mosquito quaranja-like virus 1 and Wuhan mosquito virus 6 were previously described and Ūsinis, Barstukas and Astopletus were named here. At left, plain text indicates putative labels for homologous clusters; black text indicates segments identifiable via homology (BLASTx) and red text indicates contig clusters that co-occur with identifiable segments but themselves have no identifiable homology to anything in GenBank. The Phasma-like Barstukas virus exhibits a nearly perfect overlap with Ūsinis virus (except for one sample in which Ūsinis was not found) but is identifiable as a Bunya-like virus due to having a three-segmented genome with recognizable homology across all segments to other Phasma-like viruses. Cells are colored by contig lengths (see color scale legend), highlighting their consistency which is expected of genuine segments. Deviations in detected contig lengths (e.g., *Aedes aegypti* samples that harbor shorter Ūsinis virus genome segments) reflect the presence of partial or fragmented contig assemblies in some of the samples.

Figure 6–Figure supplement 1. Heatmap showing occurrence of all clusters of viral contigs against all samples.

Figure 6–source data 1. CD-HIT-EST Clusters

Figure 6–Figure supplement 2. Overview of co-occurrence analysis approach.

Figure 6–source data 2. Output of co-occurrence analysis.

Figure 6–Figure supplement 3. Genome diagrams of new segments of quaranjaviruses and Culex narnavirus 1

Figure 6–Figure supplement 4. Clade of quaranjaviruses with eight putative segments.

Figure 6–Figure supplement 5. Tanglegram of Narnavirus RdRp and Robin segment.

Figure 6–Figure supplement 6. Abundance of Culex Narnavirus 1 versus fungi.

published SRA entries that are similar at the amino acid level to putative protein products of the two new segments, there was not sufficient coverage to reconstruct whole segments.) Furthermore, phylogenetic trees constructed separately for each of the eight segments have similar topologies (See tanglegram, top panel of ***Figure 7***), suggesting that the two new putative segments have evolved in conjunction with the previous six, bringing the total number of segments for each genome to eight.

For the two quaranjaviruses discovered in this study, Ūsinis virus and Astopletus virus, the co-occurence analysis also produced 8 segments, 5 and 4 of which, respectively, were recognizable by alignment to NCBI reference sequences. The hypothetical 2 and hypothetical 3 segments we identified from these four quaranjavirus genomes are too diverged from one another to align via BLASTx, but they do share cardinal genomic features such as sequence length, ORF length, and predicted transmembrane domains ***Figure 6–Figure Supplement 3***. The four viruses discussed here are part of a larger clade of quaranjaviruses, as pictured in ***Figure 6–Figure Supplement 4***. It is likely that the remaining seven viruses in this clade also have eight segments.

The high rate of viral co-infections (***Figure 4–Figure Supplement 1***) detected among the single mosquitoes we analyzed indicated a high likelihood that multiple mosquitoes could harbor more than one multi-segmented virus. A co-occurrence threshold of 0.8 was sufficient to deconvolve those segments into distinct genomes in all cases but one. There were 15 mosquito samples containing both Ūsinis virus, an orthomyxovirus with eight segments (three of which were unrecognizable by BLASTx), and Barstukas virus, a Phasma-like bunyavirus, with one additional sample where only Barstukas virus was found (***Figure 6***, top two blocks). In this case, we were able to disentangle the genomes of these two viruses using additional genetic information: Barstukas virus contains all three segments expected for a bunyavirus (L, GP, and NP), all of which had BLASTx hits to other Phasma-like viruses, while the unrecognizable segments of Ūsinis virus shared features with the other quaranjaviruses in the study (as described above).

Culex narnavirus 1

Beyond detection of missing genome segments for known multi-segmented viruses, the co-occurrence analysis also revealed additional genome segments in “dark matter” contig clusters for viruses with genomes previously considered to be non-segmented. A striking example is an 850 nucleotide contig cluster that co-occurred with the Culex narnavirus 1 RdRp segment in more than 40 mosquitoes collected from diverse locations across California (***Figure 6–Figure Supplement 1***). Like the RdRp segment, the putative new second segment shares the exceptional feature of ambigrammatic open reading frames (ORFs) (***DeRisi et al., 2019***; ***Dinan et al., 2019***) (***Figure 6–Figure Supplement 3***. The phylogentic tree topology for the set of 42 putative second segments is similar to the tree for the RdRp segments, suggesting co-inheritance (***Figure 6–Figure Supplement 5***). Moreover, we were able to recover nearly identical contigs from previously published mosquito datasets, all of which also contained the Culex narnavirus 1 RdRp segment. This provides strong evidence that this otherwise unrecognizable sequence is a genuine Culex narnavirus 1 segment, which we refer to here as the Robin segment.

Since the Narnaviruses were first described in fungi (***Hillman and Cai, 2013***) and recent studies have shown other eukaryotes can serve as Narnavirus hosts (***Göertz et al., 2019***; ***Charon et al., 2019***; ***Richaud et al., 2019***; ***Dinan et al., 2019***), we investigated whether this virus co-occurred with a potential non-mosquito host. There was no significant co-occurrence with a non-mosquito eukaryotic taxon, or between the abundance of Culex narnavirus 1 and abundance of fungi (***Figure 6–Figure Supplement 6***).

As the putative new Robin segment was pulled from the “dark matter” fraction of contigs, each of the ORFs it encodes corresponds to a hypothetical protein without sequence similarity to any publicly available sequence. To investigate differences in potential functional constraints of the opposing ORFs across both Robin and the RdRp segments, we compared the amino acid alignments among the isolates for each of their ORFs. Among the 42 Culex narnavirus 1 RdRp segments we

identified and the 3 closely related reference sequences, we detected 139 variable sites in the RdRp compared with 413 in the reverse ORF. In the Robin segment, there were 98 and 126 variable sites in the forward and reverse ORFs, respectively. Taken together, the discovery of a second segment sharing the ambigrammatic structure and the large difference in amino acid conservation between forward and reverse ORFs, are consistent with a model where the opposition of ORFs is important for the viral lifecycle, rather than simply an efficient encoding of unrelated viral proteins.

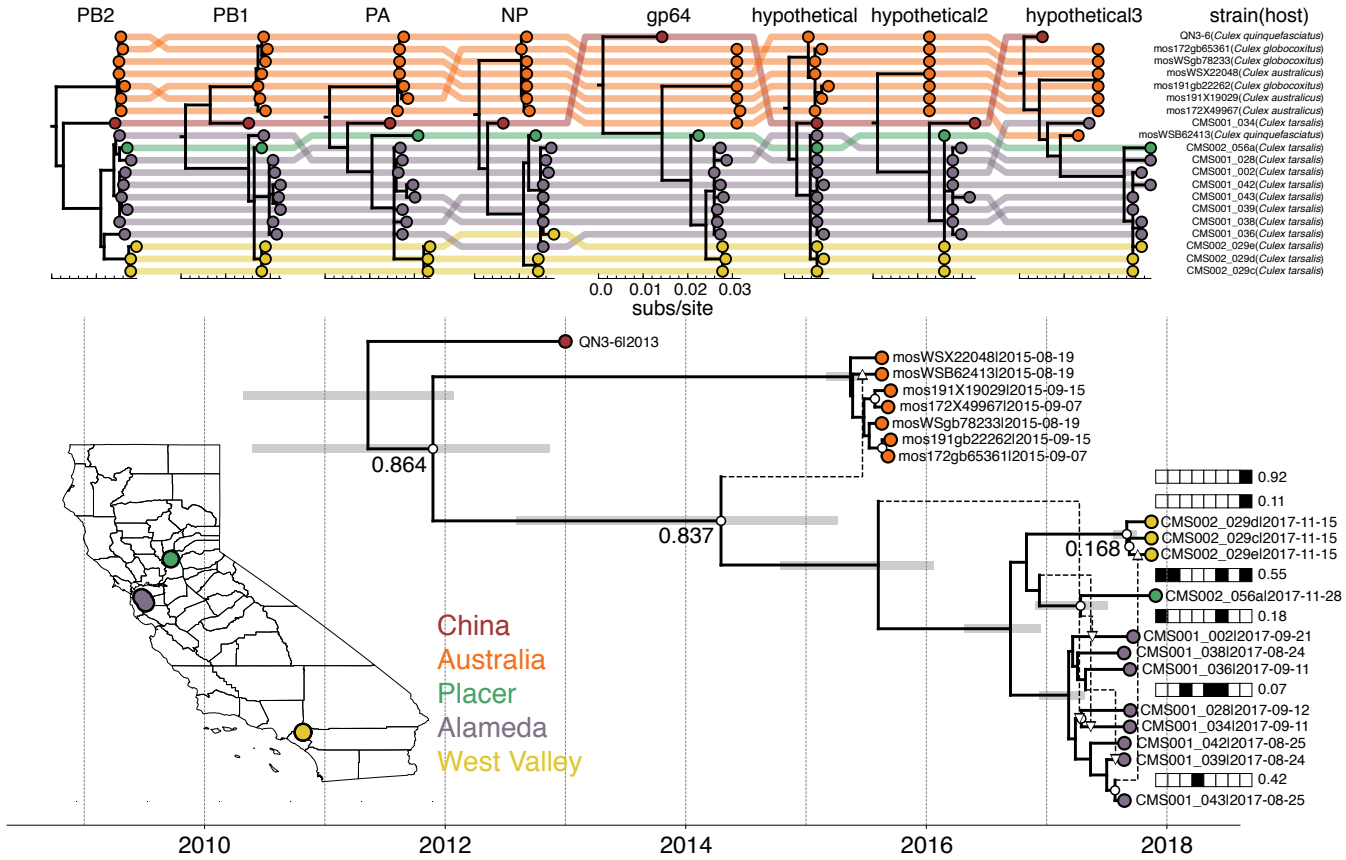


Figure 7. Top panel, Tanglegram of 20 Wuhan mosquito virus 6 genomes comprised of 8 maximum likelihood (ML) segment trees. The Chinese strain (QN3-6) was originally described from a single PB1 sequence while Australian (orange) viruses were described as having 6 segments. In this study we report the existence of two additional smaller segments (named hypothetical 2 and hypothetical 3) which we have assembled from our samples and the SRA entries of Chinese and Australian samples. Strains recovered in California as part of this study are colored by sampling location (Placer in green, Alameda County in purple, West Valley in yellow). Strain names and hosts are indicated on the far right with colored lines tracing the position of each tip in other segment trees with crossings visually indicating potential reassortments. Bottom panel, Reassortment time network of the same data as above. Clonal evolution is displayed as solid lines with dashed lines indicating inferred reticulate evolution. Information about which segments are reassorting is represented by eight squares arranged in the same order as the maximum likelihood trees above with black squares indicating traveling segments with posterior probability of reassortment in brackets displayed towards the end of the tree at the same level along the y axis as the origin of the reassorting lineage. Reassortment contributions (*i.e.* "landing" of a lineage) are indicated by small white arrows and nodes with posterior probability greater than 0.05 are indicated by white circles; posterior probabilities indicated if the node is the ancestor of at least three tips. Grey rectangles mark the 95% highest posterior density (HPD) interval for node heights.

Figure 7–source data 1. Data underlying tanglegram and phylogenetic tree.

Wuhan mosquito virus 6 phylogeography

Thorough identification of complete genomes for segmented viruses also provides a more refined and accurate understanding of both genetic history and reassortment potential. We used the set of 12 complete WMV6 genome sequences recovered from California via the co-occurrence analysis described above, plus the additional complete genomes we assembled from data provided for the previously reported Chinese (***Li et al., 2015***) and Australian WMV6 isolates (***Shi et al., 2017***),

to investigate the phylogeography of WMV6. Taken together, these provided 4 years of data for our analysis (***Drummond et al., 2003***). We reconstructed a reassortment network (***Figure 7***, bottom panel) (***Müller et al., 2019***)), to partition the WMV6 evolution into clonal parts (informed by nearly 12,000 nucleotides) and reticulate parts from reassorting segments. This limited sampling of WMV6 from mosquitoes reveals limited population structure within sampling locations (Australia and California) but clear, presumably oceanic, boundaries between viruses from different studies (***Figure 7***, bottom panel). Viral diversity appears to be distance-dependent where Australian viruses (collected at most 400km apart) have lower diversity than viruses seen in California (collected at most 700km apart). In both cases, however, viral populations are exceedingly homogeneous with times to most recent common ancestor (TMRCA) of ≤1 year (with reassortant exceptions). This, combined with an east-to-west-oriented ladder-like phylogeny, suggests that WMV6 is potentially undergoing a transcontinental sweep restricted to *Culex* mosquitoes where California represents the most recent landing of the virus. Additional sequence data not shown here suggest the origins of the sweep could be as far east as Europe (***Pettersson et al., 2019***) with Southeast Asian samples falling in the expected position between China and Australia under a hypothesis of a westerly global sweep.

Discussion

We demonstrate how metatranscriptomic sequencing of single mosquitoes, together with reference-free analyses and public databases, provides in a single assay information about circulating mosquito species, prevalence, co-occurrence, and phylodynamics of diverse known and novel viruses, prokaryotes, and eukaryotes, blood meal sources and their potential pathogens. While unbiased sequencing of individual mosquitoes is not practical or appropriate in all contexts, it offers a straightforward and rapid way to characterize a population of mosquitoes, their blood meals, and their microbial cargo. In the context of an emerging disease, where knowledge about vectors, pathogens, and reservoirs is lacking, the techniques here can provide actionable information for public health surveillance and intervention decisions.

Inferring biology from sequence in the context of an incomplete reference

The power of metatranscriptomic sequencing depends on the ability to extract biological information from nucleic acid sequences. For both bulk and single mosquito sequencing studies, the primary link between sequence and biology is provided by public reference databases, and thus the sensitivity of these approaches will depend crucially on the quality and comprehensiveness of those references. In practice, even the largest reference databases, such as nt/nr from NCBI, represent a small portion of the tree of life, so the best match for a sequence from a sample of environmental or ecological origin is often at a low percent identity. Here, we manage that uncertainty by assigning a sequence to the lowest common ancestor of its best matches in the reference database. However, there is a fundamental limit to the precision of taxonomic identification from an incomplete reference.

An advantage of single mosquito sequencing is that it offers an orthogonal source of information: the ability to recognize nucleic acid sequences detected in many samples even when they have no homology to a reference sequence. This allowed us to associate unrecognizable sequences with viral polymerases, generating hypothetical complete genomes supported, retrospectively, public datasets. The strategy of linking contigs that co-occur across samples is utilized in analysis of human and environmental microbiomes, where it is referred to as "metagenomic binning" (***Roumpeka et al., 2017***; ***Breitwieser et al., 2019***). In this manner we pulled 8% the reads in the “dark matter” fraction of our dataset into the light. The putative complete genomes were supported, retrospectively, by public datasets, and can be further validated by biological experiments or approaches such as short RNA sequencing that indicate a host antiviral response (***Aguiar et al., 2015***; ***Waldron et al., 2018***).

Another advantage of single mosquito sequencing is the ability to validate visual species identifications using molecular data. Though only 3 of the 10 mosquito species in this study had complete

genome references, it was possible to estimate pairwise SNP distances between samples in a reference-free way and perform an unsupervised clustering. The clusters were largely concordant with the visual labels, and the outliers were easy to detect and correct (***Figure 1–Figure Supplement 1***). In a bulk pool, the microbiota from mislabeled specimens would be blended in with the correctly labeled ones, and difficult or impossible to deconvolute after the fact. This approach generalizes to any collection of metatranscriptomes containing multiple representatives of each species, including the more than 200 mosquito species without full genomes, other arthropods, bats, or any other disease vector.

Distribution of microbes within mosquito populations

Once sequences have been mapped to taxa, it is relatively straightforward to characterize the composition of the microbes within a circulating population of mosquitoes. This information can inform basic research and epidemiologic questions relevant for modeling the dynamics of infectious agents and the efficacy of interventions. A key parameter is the prevalence of a microbe, which cannot be inferred from bulk data. For the 70 viruses in this study, the prevalence ranged from detection in one mosquito (RdRp group 61, peribunya-like virus) to detection in all 36 *Culex tarsalis* samples in the study (Marma virus).

For some questions, the prevalence data supplied by single mosquito sequencing is helpful for experimental design. For example, in our dataset, *Wolbachia* was either absent or endemic in each mosquito species sampled. Thus it is not possible to detect an effect of *Wolbachia* on virome composition or abundance within any species. Single mosquito sequencing could address such questions via more extensive, targeted sampling of mosquito populations where *Wolbachia* (or any other agent of interest) is expected to have an intermediate prevalence.

Also, the high incidence of viral co-infections has implications for the potential for viral recombination or reassortment within the mosquito population. As described for WMV6 below, such information can provide clues to transmission, virus evolution, and mosquito movements over time.

Blood meal sources and xenosurveillance

The identification of blood meal hosts is important for understanding mosquito ecology and controlling mosquito-borne diseases. Early field observations were supplemented by serology (***Washino and Tempelis, 1983***), and, more recently, molecular methods based on host DNA. Currently, the most popular method of blood meal identification is targeted PCR enrichment of a highly-conserved "barcode" gene, such as mitochondrial cytochrome oxidase I, followed by sequencing (***RATNASINGHAM and HEBERT, 2007***; ***Reeves et al., 2018***). To monitor specific relationships between mosquito, blood meal, and pathogen, studies have combined visual identification of mosquitoes, DNA barcode identification of blood meal, and targeted PCR or serology for pathogen identification (***Batovska et al., 2018***; ***Tedrow et al., 2019***; ***Boothe et al., 2015***; ***Tomazatos et al., 2019***). Here, we extend the spectrum of molecular methods, and show that unbiased mNGS of single mosquitoes can identify blood meal hosts, while simultaneously validating the mosquito species and providing an unbiased look at the pathogens. This allows for both reservoir identification, which seeks to identify the unknown host of a known pathogen, and xenosurveillance, which seeks to identify the unknown pathogens of specific vertebrate populations (***Grubaugh et al., 2015***). For example, in this study we found a high prevalence of the tick-borne pathogen *Anaplasma* in mosquitoes that had likely ingested a blood meal from deer. In one deer-fed mosquito, we found Lobuck virus, a member of a clade of orbiviruses implicated in disease of commercially farmed deer in Missouri, Florida, and Pennsylvania (***Cooper et al., 2014***; ***Ahasan et al., 2019b***,a). Our data suggest that mosquito species are a potential vector for such orbiviruses. For these analyses, it was crucial that single mosquitoes were sequenced—if the mosquitoes had been pooled, it would not have been possible to associate potential vertebrate pathogens with a specific blood meal hosts.

Tracking specific microbes and mosquitoes

As the rate of metagenomic sequencing increases, new studies are more likely to find previously characterized viruses. (In fact, while this manuscript was in preparation, a study of viruses in the Caribbean island of Guadaloupe (***Shi et al., 2019***) described the existence of a virus, Guadaloupe Mosquito Virus 1, which we also detected here in California.) Each virus becomes more interesting the more places it is seen, as the accumulation of observations allows for the construction of more detailed phylogenies. These can reveal the evolution of the virus, its evolutionary rate, reassortment dynamics, and geographic distribution. As the gross structure of the virome is filled out, the rate of uncovering novel genetic diversity (*i.e.* branch length) will decrease and the fine structure of viral evolution will begin to be elucidated.

The case of Wuhan Mosquito virus 6 illustrates this trend well. In the course of five years, our understanding of a virus has gone from a single segment found in one city (***Li et al., 2015***) to eight segments seen across three continents, with evidence for rapid spread, co-infection, and concomitant reassortment. As further studies provide more closely related sequences to the viruses first described in this and other reports, the portrait of mosquito-borne RNA virus evolution and migration will come into focus. In the context of potential habitat changes due to global warming (***Kraemer et al., 2019***), documented transcontinental transportation of mosquitos by humans (***Enserink, 2008***), and recent evidence of wind-borne long-distance mosquito migration (***Huestis et al., 2019***), the fast molecular clock of RNA virus evolution may prove a useful tool for the precise tracking of shifting mosquito populations.

A critical role for public data in public health

This study would have been impossible without rich public datasets containing sequences, species, locations, and sampling dates. These provided the backbone of information allowing us to identify the majority of our sequences. Non-sequence resources, such as the iNaturalist catalog of citizen scientist biodiversity observations, was a valuable complement, providing empirical knowledge of species distributions in the mosquito collection area that resolved the ambiguity we detected in sequence space.

Evidence-based public health interventions could benefit from a combination of unbiased, single mosquito sequencing, targeted analysis of mosquito pools, and field observations of mosquitoes and the animals they bite. As shown here, single mosquito mNGS can map an uncharted landscape that targeted sequencing can cheaply monitor, informing physical inspection and interventions. As mosquitoes and their microbiota continue to evolve and migrate, posing new risks for human and animal populations, these complementary approaches will empower scientists and public health professionals.

Data and Code Availability

Raw sequencing data is available on the NCBI Short Read Archive at accession PRJNA605178. Code is available on Github at github.com/czbiohub/california-mosquito-study. Derived data (including all contigs) and supplementary data is available on Figshare at dx.doi.org/10.6084/m9.figshare.11832999.

Materials and Methods

Mosquito collection

Adult mosquitoes were collected at sites indicated in ***Figure 1–source data 1*** using encephalitis virus survey (EVS) or gravid traps that were baited with CO_2 or hay-infused water, respectively. The collected mosquitoes were frozen using dry ice or paralyzed using triethyl amine and placed on a -15°C chill table or in a glass dish, respectively, for identification to species using a dissection microscope. Identified female mosquitoes were immediately frozen using dry ice in deep well 96-well plates and stored at -80°C or on dry ice until the nucleic acids were extracted for sequencing

RNA Preparation

Individual mosquitoes were homogenized in bashing tubes with 200uL DNA/RNA Shield (Zymo Research Corp., Irvine, CA, USA) using a 5mm stainless steel bead and a TissueLyserII (Qiagen, Valencia, CA, USA) (2x1min, rest on ice in between). Homogenates were centrifuged at 10,000xg for 5min at 4°C, supernatants were removed and further centrifuged at 16,000xg for 2min at 4°C after which the supernatants were completely exhausted in the nucleic acid extraction process. RNA and DNA were extracted from the mosquito supernatants using the ZR-DuetTM DNA/RNA MiniPrep kit (Zymo Research Corp., Irvine, CA, USA) with a scaled down version of the manufacturer's protocol with Dnase treatment of RNA using either the kit's DNase or the Qiagen RNase-Free DNase Set (Qiagen, Valencia, CA, USA). Water controls were performed with each extraction batch. Quantitation and quality assessment of RNA was done by the Invitrogen Qubit 3.0 Fluorometer using the Qubit RNA HS Assay Kit (ThermoFisher Scientific, Carlsbad, CA, USA) and the Agilent 2100 BioAnalyzer with the RNA 6000 Pico Kit (Agilent Technologies, Santa Clara, CA, USA).

Library Prep and Sequencing

Up to 200ng of RNA per mosquito, or 4uL aliquots of water controls extracted in parallel with mosquitoes, were used as input into the library preparation. A 25pg aliquot of External RNA Controls Consortium (ERCC) RNA Spike-In Mix (Ambion, ThermoFisher Scientific, Carlsbad, CA, USA) was added to each sample. The NEBNext Directional RNA Library Prep Kit (Purified mRNA or rRNA Depleted RNA protocol; New England BioLabs, Beverly, MA, USA) and TruSeq Index PCR Primer barcodes (Illumina, San Diego, CA, USA) were used to prepare and index each individual library. The quality and quantity of resulting individual and pooled mNGS libraries were assessed via electrophoresis with the High Sensitivity NGS Fragment Analysis Kit on a Fragment Analyzer (Advanced Analytical Technologies, Inc.), the High-Sensitivity DNA Kit on the Agilent Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA), and via real-time quantitative polymerase chain reaction (qPCR) with the KAPA Library Quantification Kit (Kapa Biosystems, Wilmington, MA, USA). Final library pools were spiked with a non-indexed PhiX control library (Illumina, San Diego, CA, USA). Pair-end sequencing (2 x 150bp) was performed using a Illumina Novaseq or Illumina NextSeq sequencing systems (Illumina, San Diego, CA, USA). The pipeline used to separate the sequencing output into 150-base-pair pair-end read FASTQ files by library and to load files onto an Amazon Web Service (AWS) S3 bucket is available on GitHub at https://github.com/czbiohub/utilities.

Mosquito Species Validation

To validate and correct the visual assignment of mosquito species, we estimated SNP distances between each pair of mosquito transcriptomes by applying SKA (Split Kmer Analysis) to the raw fastq files for each sample. The hierarchical clustering of samples based on the resulting distances was largely consistent with the visual assignments, with each cluster containing a supermajority of a single species (***Figure 1–Figure Supplement 1***). To correct likely errors in the visual assignment, samples were reassigned to the majority species in their cluster, resulting in 8 changes out of 148 samples, as documented in ***Figure 1–source data 1***.

Host and Quality Filtering

Raw sequencing reads were host- and quality-filtered and assembled using the IDseq (v3.2) platform https://idseq.net, a cloud-based, open-source bioinformatics platform designed for detection of microbes from metagenomic data.

Host Reference

We compiled a custom mosquito host reference database made up of:

1. All available mosquito genome assemblies under NCBI taxid 7157 (*Culicidae*; n=41 records corresponding to 28 unique mosquito species, including 1 *Culex*, 2 *Aedes*, and 25 *Anopheles* records) from NCBI Genome Assemblies (accession date: 12/7/2018).

2. All mosquito mitochondrial genome records under NCBI taxid 7157 available in NCBI Genomes (accession date: 12/7/2018; n=65 records).
3. A *Drosophila melanogaster* genome (GenBank GCF_000001215.4; accession date: 12/7/2018).

Mosquito Genome Assembly and mitochondrial genome accession numbers and descriptions are detailed in Supplemental Data file `mosquito_genome_refs.txt`.

Read Filtering

To select reads for assembly, we performed a series of filtering steps using the IDSeq platform:

Filter Host 1 Remove reads that align to the host reference using the Spliced Transcripts Alignment to a Reference (STAR) algorithm (***Dobin et al., 2013***).

Trim Adapters Trim Illumina adapters using trimmomatic (***Bolger et al., 2014***).

Quality Filter Remove low-quality reads using PriceSeqFilter (***Ruby et al., 2013***).

Remove Duplicate Reads Remove duplicate reads using CD-HIT-DUP (***Li and Godzik, 2006***; ***Fu et al., 2012***).

Low-Complexity Filter Remove low-complexity reads using the LZW-compression filter (***Ziv and Lempel, 1978***).

Filter Host 2 Remove further reads that align to the host reference using Bowtie2, with flag `very-sensitive-local` (***Langmead and Salzberg, 2012***).

The remaining reads are referred to as "putative non-host reads." A detailed description of all parameters is available in the IDseq documentation. https://github.com/chanzuckerberg/idseq-dag/wiki/IDseq-Pipeline-Stage-%231:-Host-Filtering-and-QC

Assembly

The putative non-host reads for each sample were assembled into contigs using SPADES with default settings. The reads used for assembly were mapped back to the contigs using Bowtie2 (flag `very-sensitive`), and contigs with more than 2 reads were retained for further analysis.

Taxonomic Assignment

We aligned each contig to the nt and nr databases using BLASTn (discontinuous megablast) and PLAST (a faster implementation of the BLASTx algorithm), respectively (***Altschul et al., 1990***; ***Van Nguyen and Lavenier, 2009***). (The databases were downloaded from NCBI on Mar 27, 2019.) Default parameters were used, except the E-value cutoff was set to 1e-2. For each contig, the results from the database with a better top hit (as judged by bitscore) are used for further analysis.

For contigs with BLAST hits to more than one species, we report the lowest common ancestor (LCA) of all hits whose number of matching aligned bases `alignment length*percent identity` is no less than the number of aligned bases for the best BLAST hit minus the number of mismatches in the best hit. (In the case that the same segment of the query is aligned for all hits, this condition guarantees that the excluded hits are further from the best hit than the query is.)

For 172,244 contigs there was a strong BLAST hit was to a species in *Hexapoda*, the subphylum of arthropods containing mosquitoes. This is likely a consequence of the limited number and quality of genomes used in host filtering, and all contigs with an alignment to *Hexapoda* of at least 80% of the query length or whose top hit (by e-value) was to *Hexapoda* were discarded from further analysis.

Contigs with no BLAST hits are referred to as "dark contigs".

For RNA viruses, where complete or near-complete genomes were recovered, a more sensitive analysis was performed.

Viral Polymerase Detection and Segment Assignment

Alignments of viral RNA-dependent RNA polymerases used to detect domains were downloaded from Pfam. These were RdRP_1 (PF00680, *Picornavirales*-like and *Nidovirales*-like), RdRP_2 (PF00978,

Tymovirales-like and Hepe-Virga-like), RdRP_3 (PF00998, *Tombusviridae*-like and *Nodaviridae*-like), RdRP_4 (PF02123, *Toti*-, *Luteo*-, and *Sobemoviridae*-like), RdRP_5 (PF07925, *Reoviridae*-like), Birna_RdRp (PF04197, *Birnaviridae*-like), Flavi_NS5 (PF00972, *Flaviviridae*-like), Mitovir_RNA_pol (PF05919, *Narnaviridae*-like), Bunya_RdRp (PF04196, *Bunyavirales*-like), Arena_RNA_pol (PF06317, *Arenaviridae*-like), Mononeg_RNA_pol (PF00946, *Mononega*- and *Chuviridae*-like), Flu_PB1 (PF00602, *Orthomyxoviridae*-like). Hidden Markov model (HMM) profiles were generated from these with HMMER (v3.1b2) (***Finn et al., 2011***) and tested against a set of diverged viruses, including ones thought to represent new families (***Obbard et al., 2019***). Based on these results only the RdRP_5 HMM was unable to detect diverged Reovirus RdRp, such as Chiqui virus (***Contreras-Gutiérrez et al., 2017***). An additional alternative Reovirus HMM (*HMMbuild* command) was generated by using BLASTp hits to Chiqui virus, largely to genera *Cypovirus* and *Oryzavirus*, aligned with MAFFT (E-INS-i, BLOSUM30) (***Katoh et al., 2005***).

All contigs of length >500 base pairs were grouped into clusters using a threshold of $\geq 99\%$ identity (CD-HIT-EST (***Li and Godzik, 2006***; ***Fu et al., 2012***)). Representative contigs from each cluster were scanned for open reading frames (standard genetic code) coding for proteins at least 200 amino acids long, in all six frames with a Python script using Biopython (***Cock et al., 2009***). These proteins were scanned using HMM profiles built earlier and potential RdRp-bearing contigs were marked for follow up. We chose to classify our contigs by focusing on RdRp under the assumption that *bona fide* exogenous viruses should at the very least carry an RdRp (per ***Li et al.*** (***2015***)) and be mostly coding-complete. Contigs that were not associated with an RdRp or coding-complete included Cell fusing agent virus (*Flaviviridae*, heavily fragmented) and Phasma-like nucleoprotein sequences (potential piRNAs) in a few samples.

Co-Occurrence

For each cluster whose representative contig contained a potential RdRp, we identified as a putative viral segment each other CD-HIT cluster whose set of samples overlapped the set of samples in the RdRp cluster at a threshold of 80%. (That is, a putative segment should be present in at least 80% of the samples that RdRp is present in, and RdRp should be present in at least 80% of the samples that the putative segment is present in).

In cases where a singleton segmented (bunya-, orthomyxo-, reo-, chryso-like, *etc*) virus was detected in a sample we relied on the presence of BLASTx hits of other segments to related viruses (*e.g.* diverged orthobunyavirus). We thus linked large numbers of viral or likely viral contigs to RNA-dependent RNA polymerases representing putative genomes for these lineages.

Final Classification

There were 1269 contigs identified as viral either by RdRp detection or co-occurence, and the resulting species-level calls are used for further analysis in lieu of the LCA computed via BLAST alignments. This included 338 “dark contigs” which had no BLAST hits, 748 with LCA in Viruses; the LCAs for the remainder were Bacteria (9), and Eukaryota (4), and Ambiguous (170), a category including (including root, cellular organisms, and synthetic constructs). Reads are assigned the taxonomic group of the contig they map to.

Water Controls and Contamination

There are many potential sources of contaminating nucleic acid, including lab surfaces, human experimenters, and reagent kits. We attempt to quantify and correct for this contamination using 8 water controls. We model contamination as a random process, where the mass of a contaminant taxon t in any sample (water or Mosquito) is a random variable X_t. We convert from units of reads to units of mass using the number of ERCC reads for each sample (as a fixed volume of ERCC spike-in solution was added to each sample well). We estimate the mean of X_t using the water controls. We say that a taxon observed in a sample is a possible contaminant if the estimated mass of that taxon in that sample is less than 100 times the average estimated mass of that taxon in the water

samples. Since the probability that a non-negative random variable is greater than 100 times its mean is at most 1% (Markov's inequality), this gives a false discovery rate of 1%. For each possible contaminant taxon in a sample, all contigs (and reads) assigned to that taxon in that sample were excluded from further analysis. A total of 46,603 reads were removed as possible contamination using this scheme. (Human and mouse were identified as the most abundant contaminant species.)

For every sample, "classified non-host reads" refer to those reads mapping to contigs that pass the above filtering, *Hexapoda* exclusion, and decontamination steps. "Non-host reads" refers to the classified non-host reads plus the reads passing host filtering which failed to assemble into contigs or assembled into a contig with only two reads.

Treemap

Treemaps are a way of visualising hierarchical information as nested rectangles whose area represents numerical values. To visualise the distribution of reads amongst taxonomic ranks, we first split the data into two categories: viral and cellular. For cellular taxonomic ranks (Bacteria, Eukaryotes, Archaea and their descendants) we assigned all reads of a contig to the taxonomic compartment the contig was assigned (see above, "Taxonomic Assignment"). For viral taxa we relied on the curated set of viral contigs coding for RdRp and their putative segments, where a putative taxonomic rank (usually family level) had been assigned. All the reads belonging to contigs that comprised putative genomes were assigned to their own compartment in the treemap, under the curated rank. Additional compartments were introduced to either reflect aspects of the outdated and potentially non-monophyletic taxonomy which is nevertheless informative (*e.g.* positive- or double-strandedness of RNA viruses) or represent previously reported groups without an official taxonomic ID on public databases (*e.g.* Narna-Levi, Toti-Chryso, Hepe-Virga, *etc*).

To prototype the cellular part of the treemap, all taxonomic IDs encountered along the path from the assigned taxonomic ID up to root (*i.e.* the taxonomic ID's lineage) were added to the treemap. Based on concentrations of reads in particular parts of the resulting taxonomic treemap, prior beliefs about the specificity of BLAST hits, and information utility, this was narrowed down to the following taxonomic ranks: cellular organisms, *Bacteria*, *Wolbachia*, *Gammaproteobacteria*, *Alphaproteobacteria*, Spirochaetes, *Enterobacterales*, *Oceanospirillales*, *Terrabacteria* group, *Eukaryota*, *Opisthokonta*, Fungi, *Dikarya*, *Bilateria*, *Boroeutheria*, *Pecora*, *Carnivora*, *Rodentia*, *Leporidae*, *Aves*, *Passeriformes*, *Trypanosomatidae*, *Leishmaniinae*, *Trypanosoma*, *Ecdysozoa*, *Nematoda*, *Microsporidia*, *Apicomplexa*, *Viridiplantae*.

Microbiota distribution in single mosquitoes

In ***Figure 4***, the denominators are non-host reads. The numerators are numbers of reads from contigs with confident assignments. For viruses, these contigs came from viral curation or co-occurrence. For *Wolbachia* and eukaryotes, these contigs had LCA assignment within the *Wolbachieae* tribe (taxid: 952) and *Eukaryota* superkingdom (taxid: 2759), respectively, and had a BLAST alignment where the percentage of aligned bases was at least 90%. Groups within viruses, *Wolbachia*, and eukaryotes were excluded for a given sample if the cumulative proportion of non-host reads was less than 1%. Samples were excluded if the total proportions of non-host reads belonging to viruses, *Wolbachia*, or eukaryotes were all below 1%.

Branch Length Contributions

Representative contigs with an encoded RdRp from each virus lineage detected were used to perform a BLASTx search under default parameters. The top 100 hits were downloaded, combined based on viral family, and duplicate hits removed. Through an iterative procedure of amino acid alignment using MAFFT (E-INS-i, BLOSUM30) and initially neighbour-joining (***Saitou and Nei, 1987***) and in later stages maximum likelihood phylogeny reconstructions using PhyML (***Guindon and Gascuel, 2003***), alignments were reduced and split such that they contained study contigs and some number of BLASTx hits. The goal was to generate alignments with catalytic motif of RdRp

(motif C: GDD, GDN, ADN or SDD) perfectly conserved and total number of identical columns in the alignment to be at least 1.0%. This is in contrast to the more commonly used approach where poorly aligning regions of the alignment are irreversibly removed, precluding future re-alignment of sequence data if taxa. By focusing our attention on the closest relatives of our contigs we largely avoided having to deal with acquisition/migration of protein domains that occur over longer periods of evolution. For broader context we relied on previously published large-scale phylogenies to indicate more distant relationships.

Given the trees, the branch length contribution was quantified using "contribution of evolutionary novelty" (CEN), defined to be the total branch length the phylogeny not visited after tip-to-root traversals from every background sequence (as illustrated in ***Figure 3–Figure Supplement 1***). This is additive in the sense that if one were to consider two studies, study 1 and study 2, added to a fixed set of background sequences, and the phylogeny fit to (background + study 1) were a subset of the phylogeny fit to (background + study 1 + study 2), then the CEN of study 1 (against the fixed background) plus the CEN of study 2 (against the fixed background + study 1) would be the same as the CEN of the composite study consisting of sequences from both (against the fixed background).

Blood meal Calling

For each of the 60 bloodfed mosquito samples from Alameda County, we selected each contig with LCA in the subphylum *Vertebrata*, excluding those contained in the order *Primates* (because of the possibility of contamination with human DNA). For each sample, we identified the lowest rank taxonomic group compatible with the LCAs of the selected contigs. (A taxonomic group is compatible with a set of taxonomic groups if it is an ancestor or descendent of each group in the set.) For 44 of the 45 samples containing vertebrate contigs, this rank is at class or below; for 12 samples, it is at the species level. Each taxonomic assignment falls into one of the following categories: *Pecora*, *Aves*, *Carnivora*, *Rodentia*, *Leporidae*. In ***Figure 5***, each sample with a blood meal detected is displayed according to which of those categories it belongs to. The remaining sample, `CMS001_022_Ra_S6`, contained three contigs mapping to members of *Pecora* and a single contig with LCA *Euarchontoglires*, a superorder of mammals including primates and rodents; we annotate this sample as containing *Pecora*.

Notably, 19 samples contain at least one contig with LCA in genus *Odocoileus* and another contig with LCA genus *Bos*. While the lowest rank compatible taxonomic group is the infraorder *Pecora*, it is likely that a single species endemic in the sampled area is responsible for all of these sequences. Given the observational data in the region (described in the main text), that species is likely a member of *Odocoileus* whose genome diverges somewhat from the reference.

Phylogenetic analyses

We chose a single Wuhan mosquito virus 6 genome from our study (CMS001_038_Ra_S22) as a reference to assemble by alignment the rest of the genome of strain QN3-6 (from SRA entry SRX833542 as only PB1 was available for this strain) and the two small segments discovered here for Australian segments (from SRA entries SRX2901194, SRX2901185, SRX2901192, SRX2901195, SRX2901187, SRX2901189, and SRX2901190) using magicblast (***Boratyn et al., 2018***). Due to much higher coverage in Australian samples, magicblast detected potential RNA splice sites for the smallest segment (hypothetical 3) which would extend the relatively short open reading frame to encompass most of the segment. Sequences of each segment were aligned with MAFFT (Auto setting) and trimmed to coding regions. For hypothetical 3 segment we inserted Ns into the sequence near the RNA splice site to bring the rest of the segment sequence into frame.

PhyML (***Guindon and Gascuel, 2003***) was used to generate maximum likelihood phylogenies under an HKY+Γ_4 (***Hasegawa et al., 1985***; ***Yang, 1994***) model. Each tree was rooted via a least-squares regression of tip dates against divergence from root in TreeTime (***Sagulenko et al., 2018***). Branches with length 0.0 in each tree (arbitrarily resolved polytomies) were collapsed, and trees untangled and visualised using baltic (https://github.com/evogytis/baltic).

A reassortment network (***Müller et al., 2019***) model implemented in BEAST2 (***Bouckaert et al., 2019***) (v2.6) was used to infer a tree-like network describing the reticulate evolution of Wuhan mosquito virus 6 sequences. Sequences were analysed using a strict molecular clock with tip-calibration, an HKY+Γ_4 (***Hasegawa et al., 1985***; ***Yang, 1994***) model of sequence evolution, under a constant population size prior. Four independent MCMC analyses were set up to run 200 million steps, sampling every 20,000 states and combined after the removal of 20 million states as burnin. Sampled networks were summarised by using network-specific software provided with BEAST2. The network was visualised using baltic (https://github.com/evogytis/baltic).

Thanks to the presence of closely related viral sequences that were collected a number of years ago (***Drummond et al., 2003***) we were able to carry out tip-calibrated temporal phylogenetic analyses for two additional viruses - Hubei virga-like virus 2 (***Shi et al., 2016***) and Culex bunyavirus 2 (another lineage also submitted to GenBank as "Bunyaviridae environmental sample"). For Hubei virga-like virus 2 there were two sequences of RdRp segment (KX883772 and KX883780) collected in 2013 that were ≥93% identical at nucleotide level to 35 sample contigs from our study. Hubei virga-like virus 2 sequences were aligned with MAFFT (***Katoh et al., 2005***) and analysed with BEAST v1.10.4 (***Suchard et al., 2018***) under an HKY+Γ_4 (***Hasegawa et al., 1985***; ***Yang, 1994***) substitution model, coalescent constant population size tree prior, no partitioning into codon positions and a strict molecular clock model. The clock model was tip-calibrated (***Rambaut, 2000***) with an uninformative reference prior (***Ferreira and Suchard, 2008***) placed on the rate of the molecular clock. MCMC was run for 50 million steps, sampling every 5000 steps. MCMC convergence was confirmed in Tracer v1.7 (***Rambaut et al., 2018***) and posterior trees summarised with TreeAnnotator distributed with BEAST.

For Culex bunyavirus 2/Bunyaviridae environmental sample there were three sequences of L segment that codes for RdRp (accessions MH188052, KP642114, and KP642115 (***Chandler et al., 2015***; ***Sadeghi et al., 2018***)) collected in 2016 and 2013. The two sequences collected in 2013 were collected in and around San Francisco (KP642114, and KP642115) are up to 98.7% identical to contigs in our study whereas the 2016 sequence (MH188052) is 99.2% identical. The three previously published sequences were aligned to 32 L segment contigs from our study with MAFFT (***Katoh et al., 2005***) and analysed with BEAST v1.10.4 (***Suchard et al., 2018***). Sites were codon partitioned (1+2 and 3) and their evolution modeled with an HKY+Γ_4 (***Hasegawa et al., 1985***; ***Yang, 1994***) substitution model with a tip-calibrated (***Rambaut, 2000***) strict molecular clock model under a coalescent constant population size tree prior. The analysis did not converge when using an uninformative prior but did when being constrained with a uniform prior with bounds (0.000001 and 0.01 substitutions per site per year). MCMC was run for 50 million steps, sampling every 5000 steps, convergence being confirmed with Tracer v1.7 (***Rambaut et al., 2018***) and posterior trees summarised with TreeAnnotator distributed as part of the BEAST package.

To generate the Culex narnavirus 1 tanglegrams, 42 sequences of RdRp and 42 co-occurring Robin segment sequences from our samples and three previously published RdRp sequences (MK628543, KP642119, KP642120) as well as their three corresponding Robin segments assembled from SRA entries (SRR8668667, SRR1706006, SRR1705824, respectively) were aligned with MAFFT (***Katoh et al., 2005***) and trimmed to just the most conserved open reading frame (as opposed to its complement on the reverse strand). Maximum likelihood phylogenies for both RdRp and Robin segments were generated with PhyML (***Guindon and Gascuel, 2003***) with 100 bootstrap replicates under an HKY+Γ_4 (***Hasegawa et al., 1985***; ***Yang, 1994***) substitution model. The resulting phylogenies were mid-point rooted, untangled and visualised using baltic (https://github.com/evogytis/baltic).

Acknowledgments

We thank our collaborating partners in the California Mosquito and Vector Control Agency Districts of Alameda, Placer, San Diego, West Valley, and Coachella Valley, who provided all the mosquito specimens and corresponding metadata that made this study possible. We thank Maira Phelps for liason work with collaborators and in-house specimen management. We thank Rene Sit, Michelle Tan, and Norma Neff of the Chan Zuckerberg Biohub Genomics Platform for supporting all aspects

of mNGS sequencing for this study. We thank the IDseq team at the Chan Zuckerberg Initiative for useful discussions and facilitation of analysis over the course of this study. We thank Jack Kamm, Darren J Obbard, and Cristina Tato for useful discussions during the development of this project. We would also like to acknowledge Natalie Whitis and Annie Lo for their contribution to the early phases of the specimen extraction and sequencing library preparation for this project. We thank Cristina Tato, Peter Kim, David Yllanes, and Joe DeRisi for reviewing the manuscript.

References

Aguiar ERGR, Olmo RP, Paro S, Ferreira FV, de Faria IJdS, Todjro YMH, Lobo FP, Kroon EG, Meignin C, Gatherer D, Imler JL, Marques JT. Sequence-independent characterization of viruses based on the pattern of viral small RNAs produced by the host. Nucleic Acids Res. 2015 Jul; 43(13):6191–6206. https://academic.oup.com/nar/article/43/13/6191/2414271, doi: 10.1093/nar/gkv587.

Ahasan MS, Campos Krauer JM, Subramaniam K, Lednicky JA, Loeb JC, Sayler KA, Wisely SM, Waltzek TB. Complete Genome Sequence of Mobuck Virus Isolated from a Florida White-Tailed Deer (Odocoileus virginianus). Microbiol Resour Announc. 2019 Jan; 8(3). doi: 10.1128/MRA.01324-18.

Ahasan MS, Subramaniam K, Campos Krauer JM, Sayler KA, Loeb JC, Goodfriend OF, Barber HM, Stephenson CJ, Popov VL, Charrel RN, Wisely SM, Waltzek TB, Lednicky JA. Three New Orbivirus Species Isolated from Farmed White-Tailed Deer (Odocoileus virginianus) in the United States. Viruses. 2019 Dec; 12(1). doi: 10.3390/v12010013.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of Molecular Biology. 1990 Oct; 215(3):403–410. http://www.sciencedirect.com/science/article/pii/S0022283605803602, doi: 10.1016/S0022-2836(05)80360-2.

Anders KL, Indriani C, Ahmad RA, Tantowijoyo W, Arguni E, Andari B, Jewell NP, Rances E, O'Neill SL, Simmons CP, Utarini A. The AWED trial (Applying Wolbachia to Eliminate Dengue) to assess the efficacy of Wolbachia-infected mosquito deployments to reduce dengue incidence in Yogyakarta, Indonesia: study protocol for a cluster randomised controlled trial. Trials. 2018 May; 19(1):302. doi: 10.1186/s13063-018-2670-z.

Atoni E, Wang Y, Karungu S, Waruhiu C, Zohaib A, Obanda V, Agwanda B, Mutua M, Xia H, Yuan Z. Metagenomic Virome Analysis of Culex Mosquitoes from Kenya and China. Viruses. 2018 Jan; 10(1). doi: 10.3390/v10010030.

Atoni E, Zhao L, Karungu S, Obanda V, Agwanda B, Xia H, Yuan Z. The discovery and global distribution of novel mosquito-associated viruses in the last decade (2007-2017). Rev Med Virol. 2019 Aug; p. e2079. doi: 10.1002/rmv.2079.

Batovska J, Lynch SE, Cogan NOI, Brown K, Darbro JM, Kho EA, Blacket MJ. Effective mosquito and arbovirus surveillance using metabarcoding. Molecular Ecology Resources. 2018; 18(1):32–40. https://onlinelibrary.wiley.com/doi/abs/10.1111/1755-0998.12682, doi: 10.1111/1755-0998.12682.

Bigot D, Atyame CM, Weill M, Justy F, Herniou EA, Gayral P. Discovery of Culex pipiens associated tunisia virus: a new ssRNA(+) virus representing a new insect associated virus family. Virus Evol. 2018 Jan; 4(1):vex040. doi: 10.1093/ve/vex040.

Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014 Aug; 30(15):2114–2120. https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btu170, doi: 10.1093/bioinformatics/btu170.

Bolling BG, Weaver SC, Tesh RB, Vasilakis N. Insect-Specific Virus Discovery: Significance for the Arbovirus Community. Viruses. 2015 Sep; 7(9):4911–4928. doi: 10.3390/v7092851.

Boothe E, Medeiros MCI, Kitron UD, Brawn JD, Ruiz MO, Goldberg TL, Walker ED, Hamer GL. Identification of Avian and Hemoparasite DNA in Blood-Engorged Abdomens of Culex pipiens (Diptera; Culicidae) from a West Nile Virus Epidemic region in Suburban Chicago, Illinois. J Med Entomol. 2015 May; 52(3):461–468. doi: 10.1093/jme/tjv029.

Boratyn GM, Thierry-Mieg J, Thierry-Mieg D, Busby B, Madden TL. Magic-BLAST, an accurate DNA and RNA-seq aligner for long and short reads. bioRxiv. 2018 Aug; p. 390013. https://www.biorxiv.org/content/10.1101/390013v1, doi: 10.1101/390013.

Bouckaert R, Vaughan TG, Barido-Sottani J, Duchêne S, Fourment M, Gavryushkina A, Heled J, Jones G, Kühnert D, Maio ND, Matschiner M, Mendes FK, Müller NF, Ogilvie HA, Plessis Ld, Popinga A, Rambaut A, Rasmussen D, Siveroni I, Suchard MA, et al. BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. PLOS Computational Biology. 2019 Apr; 15(4):e1006650. https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1006650, doi: 10.1371/journal.pcbi.1006650.

Breitwieser FP, Lu J, Salzberg SL. A review of methods and databases for metagenomic classification and assembly. Briefings in Bioinformatics. 2019 Jul; 20(4):1125–1136. https://academic.oup.com/bib/article/20/4/1125/4210288, doi: 10.1093/bib/bbx120.

Carlson JS, Nelms B, Barker CM, Reisen WK, Sehgal RNM, Cornel AJ. Avian malaria co-infections confound infectivity and vector competence assays of Plasmodium homopolare. Parasitol Res. 2018 Aug; 117(8):2385–2394. doi: 10.1007/s00436-018-5924-5.

Chandler JA, Liu RM, Bennett SN. RNA shotgun metagenomic sequencing of northern California (USA) mosquitoes uncovers viruses, bacteria, and fungi. Front Microbiol. 2015; 6:185. doi: 10.3389/fmicb.2015.00185.

Charon J, Grigg MJ, Eden JS, Piera KA, Rana H, William T, Rose K, Davenport MP, Anstey NM, Holmes EC. Novel RNA viruses associated with Plasmodium vivax in human malaria and Leucocytozoon parasites in avian disease. PLoS Pathog. 2019 Dec; 15(12):e1008216. doi: 10.1371/journal.ppat.1008216.

Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, de Hoon MJL. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009 Jun; 25(11):1422–1423. https://academic.oup.com/bioinformatics/article/25/11/1422/330687, doi: 10.1093/bioinformatics/btp163.

Cohnstaedt NPaLW. 8. Mosquito-borne diseases in the livestock industry. In: *Pests and vector-borne diseases in the livestock industry*; 2018.p. 195–219. https://www.wageningenacademic.com/doi/abs/10.3920/978-90-8686-863-6_8.

Contreras-Gutiérrez MA, Nunes MRT, Guzman H, Uribe S, Gallego Gómez JC, Suaza Vasco JD, Cardoso JF, Popov VL, Widen SG, Wood TG, Vasilakis N, Tesh RB. Sinu virus, a novel and divergent orthomyxovirus related to members of the genus Thogotovirus isolated from mosquitoes in Colombia. Virology. 2017 Jan; 501:166–175. http://www.sciencedirect.com/science/article/pii/S0042682216303683, doi: 10.1016/j.virol.2016.11.014.

Cook S, Chung BYW, Bass D, Moureau G, Tang S, McAlister E, Culverwell CL, Glücksman E, Wang H, Brown TDK, Gould EA, Harbach RE, de Lamballerie X, Firth AE. Novel Virus Discovery and Genome Reconstruction from Field RNA Samples Reveals Highly Divergent Viruses in Dipteran Hosts. PLoS ONE. 2013 Nov; 8(11):e80720. http://dx.plos.org/10.1371/journal.pone.0080720, doi: 10.1371/journal.pone.0080720.

Cooper E, Anbalagan S, Klumper P, Scherba G, Simonson RR, Hause BM. Mobuck virus genome sequence and phylogenetic analysis: identification of a novel Orbivirus isolated from a white-tailed deer in Missouri, USA. J Gen Virol. 2014 Jan; 95(Pt 1):110–116. doi: 10.1099/vir.0.058800-0.

DeRisi JL, Huber G, Kistler A, Retallack H, Wilkinson M, Yllanes D. An exploration of ambigrammatic sequences in narnaviruses. Scientific Reports. 2019 Nov; 9(1):1–9. http://www.nature.com/articles/s41598-019-54181-3, doi: 10.1038/s41598-019-54181-3.

Dinan AM, Lukhovitskaya NI, Olendraite I, Firth AE. A case for a reverse-frame coding sequence in a group of positive-sense RNA viruses. bioRxiv. 2019 Jun; p. 664342. https://www.biorxiv.org/content/10.1101/664342v1, doi: 10.1101/664342.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013 Jan; 29(1):15–21. doi: 10.1093/bioinformatics/bts635.

Drummond AJ, Pybus OG, Rambaut A, Forsberg R, Rodrigo AG. Measurably evolving populations. Trends in Ecology & Evolution. 2003 Sep; 18(9):481–488. http://www.sciencedirect.com/science/article/pii/S0169534703002167, doi: 10.1016/S0169-5347(03)00216-7.

Enserink M. ENTOMOLOGY: A Mosquito Goes Global. Science. 2008 May; 320(5878):864–866. http://www.sciencemag.org/cgi/doi/10.1126/science.320.5878.864, doi: 10.1126/science.320.5878.864.

Fauver JR, Grubaugh ND, Krajacich BJ, Weger-Lucarelli J, Lakin SM, Fakoli LS, Bolay FK, Diclaro JW, Dabiré KR, Foy BD, Brackney DE, Ebel GD, Stenglein MD. West African Anopheles gambiae mosquitoes harbor a taxonomically diverse virome including new insect-specific flaviviruses, mononegaviruses, and totiviruses. Virology. 2016; 498:288–299. doi: 10.1016/j.virol.2016.07.031.

Ferreira MAR, Suchard MA. Bayesian analysis of elapsed times in continuous-time Markov chains. Canadian Journal of Statistics. 2008; 36(3):355–368. https://onlinelibrary.wiley.com/doi/abs/10.1002/cjs.5550360302, doi: 10.1002/cjs.5550360302.

Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. 2011 Jul; 39(suppl_2):W29–W37. https://academic.oup.com/nar/article/39/suppl_2/W29/2506513, doi: 10.1093/nar/gkr367.

Flores HA, O'Neill SL. Controlling vector-borne diseases by releasing modified mosquitoes. Nat Rev Microbiol. 2018; 16(8):508–518. doi: 10.1038/s41579-018-0025-0.

François S, Filloux D, Roumagnac P, Bigot D, Gayral P, Martin DP, Froissart R, Ogliastro M. Discovery of parvovirus-related sequences in an unexpected broad range of animals. Sci Rep. 2016 Sep; 6(1):1–13. https://www.nature.com/articles/srep30880, doi: 10.1038/srep30880.

Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012 Dec; 28(23):3150–3152. doi: 10.1093/bioinformatics/bts565.

GBD 2017 Causes of Death Collaborators. Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980-2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet. 2018; 392(10159):1736–1788. doi: 10.1016/S0140-6736(18)32203-7.

GBD 2017 DALYs and HALE Collaborators. Global, regional, and national disability-adjusted life-years (DALYs) for 359 diseases and injuries and healthy life expectancy (HALE) for 195 countries and territories, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet. 2018; 392(10159):1859–1922. doi: 10.1016/S0140-6736(18)32335-3.

GBD 2017 Disease and Injury Incidence and Prevalence Collaborators. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet. 2018; 392(10159):1789–1858. doi: 10.1016/S0140-6736(18)32279-7.

Grubaugh ND, Sharma S, Krajacich BJ, Iii LSF, Bolay FK, Ii JWD, Johnson WE, Ebel GD, Foy BD, Brackney DE. Xenosurveillance: A Novel Mosquito-Based Approach for Examining the Human-Pathogen Landscape. PLOS Neglected Tropical Diseases. 2015 Mar; 9(3):e0003628. https://journals.plos.org/plosntds/article?id=10.1371/journal.pntd.0003628, doi: 10.1371/journal.pntd.0003628.

Grybchuk D, Akopyants NS, Kostygov AY, Konovalovas A, Lye LF, Dobson DE, Zangger H, Fasel N, Butenko A, Frolov AO, Votýpka J, d'Avila Levy CM, Kulich P, Moravcová J, Plevka P, Rogozin IB, Serva S, Lukeš J, Beverley SM, Yurchenko V. Viral discovery and diversity in trypanosomatid protozoa with a focus on relatives of the human parasite *Leishmania*. Proc Natl Acad Sci USA. 2018 Jan; 115(3):E506–E515. http://www.pnas.org/lookup/doi/10.1073/pnas.1717806115, doi: 10.1073/pnas.1717806115.

Guindon S, Gascuel O. A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood. Syst Biol. 2003 Oct; 52(5):696–704. https://academic.oup.com/sysbio/article/52/5/696/1681984, doi: 10.1080/10635150390235520.

Göertz GP, Miesen P, Overheul GJ, van Rij RP, van Oers MM, Pijlman GP. Mosquito Small RNA Responses to West Nile and Insect-Specific Virus Infections in Aedes and Culex Mosquito Cells. Viruses. 2019; 11(3). doi: 10.3390/v11030271.

Harl J, Himmel T, Valkiūnas G, Weissenböck H. The nuclear 18S ribosomal DNAs of avian haemosporidian parasites. Malar J. 2019 Sep; 18(1):305. doi: 10.1186/s12936-019-2940-6.

Harris SR. SKA: Split Kmer Analysis Toolkit for Bacterial Genomic Epidemiology. bioRxiv. 2018 Oct; p. 453142. https://www.biorxiv.org/content/early/2018/10/25/453142, doi: 10.1101/453142.

Hasegawa M, Kishino H, Yano Ta. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol. 1985 Oct; 22(2):160–174. https://doi.org/10.1007/BF02101694, doi: 10.1007/BF02101694.

Hillman BI, Cai G. The family narnaviridae: simplest of RNA viruses. Adv Virus Res. 2013; 86:149–176. doi: 10.1016/B978-0-12-394315-6.00006-4.

Hoffmann AA, Montgomery BL, Popovici J, Iturbe-Ormaetxe I, Johnson PH, Muzzi F, Greenfield M, Durkan M, Leong YS, Dong Y, Cook H, Axford J, Callahan AG, Kenny N, Omodei C, McGraw EA, Ryan PA, Ritchie SA, Turelli M, O'Neill SL. Successful establishment of Wolbachia in Aedes populations to suppress dengue transmission. Nature. 2011 Aug; 476(7361):454–457. doi: 10.1038/nature10356.

1000 **Hofmeister EK**. West Nile virus: North American experience. Integrative Zoology. 2011; 6(3):279–289.

1001 **Huestis DL**, Dao A, Diallo M, Sanogo ZL, Samake D, Yaro AS, Ousman Y, Linton YM, Krishna A, Veru L, Krajacich BJ,
1002 Faiman R, Florio J, Chapman JW, Reynolds DR, Weetman D, Mitchell R, Donnelly MJ, Talamas E, Chamorro L, et al.
1003 Windborne long-distance migration of malaria mosquitoes in the Sahel. Nature. 2019 Oct; 574(7778):404–408.
1004 http://www.nature.com/articles/s41586-019-1622-4, doi: 10.1038/s41586-019-1622-4.

1005 **iNaturalist**. iNaturalist Research-grade Observations. Occurrence dataset; 2020. doi: 10.15468/ab3s5x.

1006 **Katoh K**, Kuma Ki, Toh H, Miyata T. MAFFT version 5: improvement in accuracy of multiple sequence alignment.
1007 Nucleic Acids Res. 2005 Jan; 33(2):511–518. https://academic.oup.com/nar/article/33/2/511/2549118, doi:
1008 10.1093/nar/gki198.

1009 **Katzourakis A**, Gifford RJ. Endogenous Viral Elements in Animal Genomes. PLOS Genetics. 2010
1010 Nov; 6(11):e1001191. https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1001191, doi:
1011 10.1371/journal.pgen.1001191.

1012 **Kittayapong P**, Baisley KJ, Baimai V, O'Neill SL. Distribution and diversity of Wolbachia infections in South-
1013 east Asian mosquitoes (Diptera: Culicidae). J Med Entomol. 2000 May; 37(3):340–345. doi: 10.1093/jme-
1014 dent/37.3.340.

1015 **Kraemer MUG**, Reiner RC, Brady OJ, Messina JP, Gilbert M, Pigott DM, Yi D, Johnson K, Earl L, Marczak LB, Shirude
1016 S, Davis Weaver N, Bisanzio D, Perkins TA, Lai S, Lu X, Jones P, Coelho GE, Carvalho RG, Van Bortel W, et al.
1017 Past and future spread of the arbovirus vectors Aedes aegypti and Aedes albopictus. Nat Microbiol. 2019
1018 May; 4(5):854–863. http://link.springer.com/10.1038/s41564-019-0376-y, doi: 10.1038/s41564-019-0376-y.

1019 **Langmead B**, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012 Mar; 9(4):357–359.
1020 doi: 10.1038/nmeth.1923.

1021 **Lee JS**, Bevins SN, Serieys LEK, Vickers W, Logan KA, Aldredge M, Boydston EE, Lyren LM, McBride R, Roelke-Parker
1022 M, Pecon-Slattery J, Troyer JL, Riley SP, Boyce WM, Crooks KR, VandeWoude S. Evolution of Puma Lentivirus in
1023 Bobcats (Lynx rufus) and Mountain Lions (Puma concolor) in North America. Journal of Virology. 2014 Jul;
1024 88(14):7727–7737. http://jvi.asm.org/lookup/doi/10.1128/JVI.00473-14, doi: 10.1128/JVI.00473-14.

1025 **Li CX**, Shi M, Tian JH, Lin XD, Kang YJ, Chen LJ, Qin XC, Xu J, Holmes EC, Zhang YZ. Unprecedented genomic
1026 diversity of RNA viruses in arthropods reveals the ancestry of negative-sense RNA viruses. eLife. 2015 Jan;
1027 4:e05378. https://doi.org/10.7554/eLife.05378, doi: 10.7554/eLife.05378.

1028 **Li W**, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences.
1029 Bioinformatics. 2006 Jul; 22(13):1658–1659. doi: 10.1093/bioinformatics/btl158.

1030 **Moreira LA**, Iturbe-Ormaetxe I, Jeffery JA, Lu G, Pyke AT, Hedges LM, Rocha BC, Hall-Mendelin S, Day A, Riegler
1031 M, Hugo LE, Johnson KN, Kay BH, McGraw EA, van den Hurk AF, Ryan PA, O'Neill SL. A Wolbachia Symbiont in
1032 Aedes aegypti Limits Infection with Dengue, Chikungunya, and Plasmodium. Cell. 2009 Dec; 139(7):1268–1278.
1033 http://www.sciencedirect.com/science/article/pii/S0092867409015001, doi: 10.1016/j.cell.2009.11.042.

1034 **Müller NF**, Stolz U, Dudas G, Stadler T, Vaughan TG. Bayesian inference of reassortment networks reveals
1035 fitness benefits of reassortment in human influenza viruses. bioRxiv. 2019 Aug; p. 726042. https://www.
1036 biorxiv.org/content/10.1101/726042v1, doi: 10.1101/726042.

1037 **Obbard DJ**, Shi M, Roberts KE, Longdon B, Dennis AB. A new lineage of segmented RNA viruses infecting animals.
1038 bioRxiv. 2019 Dec; p. 741645. https://www.biorxiv.org/content/10.1101/741645v2, doi: 10.1101/741645.

1039 **O'Neill SL**, Ryan PA, Turley AP, Wilson G, Retzki K, Iturbe-Ormaetxe I, Dong Y, Kenny N, Paton CJ, Ritchie SA, Brown-
1040 Kenyon J, Stanford D, Wittmeier N, Jewell NP, Tanamas SK, Anders KL, Simmons CP. Scaled deployment of
1041 Wolbachia to protect the community from dengue and other Aedes transmitted arboviruses. Gates Open Res.
1042 2019 Aug; 2. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6305154/, doi: 10.12688/gatesopenres.12844.3.

1043 **Palatini U**, Miesen P, Carballar-Lejarazu R, Ometto L, Rizzo E, Tu Z, van Rij RP, Bonizzoni M. Comparative
1044 genomics shows that viral integrations are abundant and express piRNAs in the arboviral vectors Aedes
1045 aegypti and Aedes albopictus. BMC Genomics. 2017; 18(1):512. doi: 10.1186/s12864-017-3903-3.

1046 **Parry R**, Asgari S. Aedes Anphevirus: an Insect-Specific Virus Distributed Worldwide in Aedes aegypti Mosquitoes
1047 That Has Complex Interplays with Wolbachia and Dengue Virus Infection in Cells. J Virol. 2018; 92(17). doi:
1048 10.1128/JVI.00224-18.

Pettersson JHO, Shi M, Eden JS, Holmes EC, Hesson JC. Meta-Transcriptomic Comparison of the RNA Viromes of the Mosquito Vectors Culex pipiens and Culex torrentium in Northern Europe. Viruses. 2019 Nov; 11(11). doi: 10.3390/v11111033.

Rambaut A. Estimating the rate of molecular evolution: incorporating non-contemporaneous sequences into maximum likelihood phylogenies. Bioinformatics. 2000 Apr; 16(4):395–399. https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/16.4.395, doi: 10.1093/bioinformatics/16.4.395.

Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA. Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. Syst Biol. 2018 Sep; 67(5):901–904. https://academic.oup.com/sysbio/article/67/5/901/4989127, doi: 10.1093/sysbio/syy032.

Rasgon JL, Scott TW. An initial survey for Wolbachia (Rickettsiales: Rickettsiaceae) infections in selected California mosquitoes (Diptera: Culicidae). J Med Entomol. 2004 Mar; 41(2):255–257. doi: 10.1603/0022-2585-41.2.255.

RATNASINGHAM S, HEBERT PDN. bold: The Barcode of Life Data System (http://www.barcodinglife.org). Mol Ecol Notes. 2007 May; 7(3):355–364. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1890991/, doi: 10.1111/j.1471-8286.2007.01678.x.

Reeves LE, Gillett-Kaufman JL, Kawahara AY, Kaufman PE. Barcoding blood meals: New vertebrate-specific primer sets for assigning taxonomic identities to host DNA from mosquito blood meals. PLOS Neglected Tropical Diseases. 2018 Aug; 12(8):e0006767. https://journals.plos.org/plosntds/article?id=10.1371/journal.pntd.0006767, doi: 10.1371/journal.pntd.0006767.

Richaud A, Frézal L, Tahan S, Jiang H, Blatter JA, Zhao G, Kaur T, Wang D, Félix MA. Vertical transmission in Caenorhabditis nematodes of RNA molecules encoding a viral RNA-dependent RNA polymerase. Proc Natl Acad Sci USA. 2019 Dec; 116(49):24738–24747. doi: 10.1073/pnas.1903903116.

Ritchie SA, van den Hurk AF, Smout MJ, Staunton KM, Hoffmann AA. Mission Accomplished? We Need a Guide to the 'Post Release' World of Wolbachia for Aedes-borne Disease Control. Trends in Parasitology. 2018 Mar; 34(3):217–226. http://www.sciencedirect.com/science/article/pii/S1471492217302854, doi: 10.1016/j.pt.2017.11.011.

Roumpeka DD, Wallace RJ, Escalettes F, Fotheringham I, Watson M. A Review of Bioinformatics Tools for Bio-Prospecting from Metagenomic Sequence Data. Front Genet. 2017 Mar; 8. http://journal.frontiersin.org/article/10.3389/fgene.2017.00023/full, doi: 10.3389/fgene.2017.00023.

Ruby JG, Bellare P, DeRisi JL. PRICE: Software for the Targeted Assembly of Components of (Meta) Genomic Sequence Data. G3. 2013 May; 3(5):865–880. http://g3journal.org/lookup/doi/10.1534/g3.113.005967, doi: 10.1534/g3.113.005967.

Sadeghi M, Altan E, Deng X, Barker CM, Fang Y, Coffey LL, Delwart E. Virome of >12 thousand Culex mosquitoes from throughout California. Virology. 2018; 523:74–88. doi: 10.1016/j.virol.2018.07.029.

Sadeghi M, Popov V, Guzman H, Phan TG, Vasilakis N, Tesh R, Delwart E. Genomes of viral isolates derived from different mosquitos species. Virus Res. 2017 Oct; 242:49–57. doi: 10.1016/j.virusres.2017.08.012.

Sagulenko P, Puller V, Neher RA. TreeTime: Maximum-likelihood phylodynamic analysis. Virus Evol. 2018 Jan; 4(1). https://academic.oup.com/ve/article/4/1/vex042/4794731, doi: 10.1093/ve/vex042.

Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 1987 Jul; 4(4):406–425. https://academic.oup.com/mbe/article/4/4/406/1029664, doi: 10.1093/oxfordjournals.molbev.a040454.

Shane JL, Grogan CL, Cwalina C, Lampe DJ. Blood meal-induced inhibition of vector-borne disease by transgenic microbiota. Nat Commun. 2018; 9(1):4127. doi: 10.1038/s41467-018-06580-9.

Shi C, Beller L, Deboutte W, Yinda KC, Delang L, Vega-Rúa A, Failloux AB, Matthijnssens J. Stable distinct core eukaryotic viromes in different mosquito species from Guadeloupe, using single mosquito viral metagenomics. Microbiome. 2019 Aug; 7(1):121. doi: 10.1186/s40168-019-0734-2.

Shi C, Liu Y, Hu X, Xiong J, Zhang B, Yuan Z. A metagenomic survey of viral abundance and diversity in mosquitoes from Hubei province. PLoS ONE. 2015; 10(6):e0129845. doi: 10.1371/journal.pone.0129845.

Shi M, Lin XD, Chen X, Tian JH, Chen LJ, Li K, Wang W, Eden JS, Shen JJ, Liu L, Holmes EC, Zhang YZ. The evolutionary history of vertebrate RNA viruses. Nature. 2018 Apr; 556(7700):197. https://www.nature.com/articles/s41586-018-0012-7/, doi: 10.1038/s41586-018-0012-7.

Shi M, Lin XD, Tian JH, Chen LJ, Chen X, Li CX, Qin XC, Li J, Cao JP, Eden JS, Buchmann J, Wang W, Xu J, Holmes EC, Zhang YZ. Redefining the invertebrate RNA virosphere. Nature. 2016 Nov; doi: 10.1038/nature20167.

Shi M, Neville P, Nicholson J, Eden JS, Imrie A, Holmes EC. High-Resolution Metatranscriptomics Reveals the Ecological Dynamics of Mosquito-Associated RNA Viruses in Western Australia. J Virol. 2017 Sep; 91(17). doi: 10.1128/JVI.00680-17.

Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ, Rambaut A. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. Virus Evol. 2018 Jan; 4(1). https://academic.oup.com/ve/article/4/1/vey016/5035211, doi: 10.1093/ve/vey016.

Talbi C, Lemey P, Suchard MA, Abdelatif E, Elharrak M, Jalal N, Faouzi A, Echevarría JE, Vazquez Morón S, Rambaut A, Campiz N, Tatem AJ, Holmes EC, Bourhy H. Phylodynamics and Human-Mediated Dispersal of a Zoonotic Virus. PLoS Pathogens. 2010 Oct; 6(10):e1001166. https://dx.plos.org/10.1371/journal.ppat.1001166, doi: 10.1371/journal.ppat.1001166.

Tedrow RE, Rakotomanga T, Nepomichene T, Howes RE, Ratovonjato J, Ratsimbasoa AC, Svenson GJ, Zimmerman PA. Anopheles mosquito surveillance in Madagascar reveals multiple blood feeding behavior and Plasmodium infection. PLOS Neglected Tropical Diseases. 2019 Jul; 13(7):e0007176. https://journals.plos.org/plosntds/article?id=10.1371/journal.pntd.0007176, doi: 10.1371/journal.pntd.0007176.

Ter Horst AM, Nigg JC, Dekker FM, Falk BW. Endogenous Viral Elements Are Widespread in Arthropod Genomes and Commonly Give Rise to PIWI-Interacting RNAs. J Virol. 2019; 93(6). doi: 10.1128/JVI.02124-18.

Thongsripong P, Chandler JA, Green AB, Kittayapong P, Wilcox BA, Kapan DD, Bennett SN. Mosquito vector-associated microbiota: Metabarcoding bacteria and eukaryotic symbionts across habitat types in Thailand endemic for dengue and other arthropod-borne diseases. Ecol Evol. 2018; 8(2):1352–1368. doi: 10.1002/ece3.3676.

Tomazatos A, Jansen S, Pfister S, Török E, Maranda I, Horváth C, Keresztes L, Spînu M, Tannich E, Jöst H, Schmidt-Chanasit J, Cadar D, Lühken R. Ecology of West Nile Virus in the Danube Delta, Romania: Phylogeography, Xenosurveillance and Mosquito Host-Feeding Patterns. Viruses. 2019 Dec; 11(12):1159. https://www.mdpi.com/1999-4915/11/12/1159, doi: 10.3390/v11121159.

Tyler S, Bolling BG, Blair CD, Brault AC, Pabbaraju K, Armijos MV, Clark DC, Calisher CH, Drebot MA. Distribution and phylogenetic comparisons of a novel mosquito flavivirus sequence present in Culex tarsalis Mosquitoes from western Canada with viruses isolated in California and Colorado. Am J Trop Med Hyg. 2011 Jul; 85(1):162–168. doi: 10.4269/ajtmh.2011.10-0469.

Van Nguyen H, Lavenier D. PLAST: parallel local alignment search tool for database comparison. BMC Bioinformatics. 2009 Dec; 10(1):329. https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-10-329, doi: 10.1186/1471-2105-10-329.

Vasilakis N, Tesh RB. Insect-specific viruses and their potential impact on arbovirus transmission. Curr Opin Virol. 2015 Dec; 15:69–74. doi: 10.1016/j.coviro.2015.08.007.

Waldron FM, Stone GN, Obbard DJ. Metagenomic sequencing suggests a diversity of RNA interference-like responses to viruses across multicellular eukaryotes. PLOS Genetics. 2018 Jul; 14(7):e1007533. https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1007533, doi: 10.1371/journal.pgen.1007533.

Washino RK, Tempelis CH. Mosquito Host Bloodmeal Identification: Methodology and Data Analysis. Annual Review of Entomology. 1983 Jan; 28(1):179–201. http://www.annualreviews.org/doi/10.1146/annurev.en.28.010183.001143, doi: 10.1146/annurev.en.28.010183.001143.

Werren JH, Baldo L, Clark ME. Wolbachia: master manipulators of invertebrate biology. Nat Rev Microbiol. 2008 Oct; 6(10):741–751. doi: 10.1038/nrmicro1969.

Wheeler DC, Waller LA, Biek R. Spatial analysis of feline immunodeficiency virus infection in cougars. Spatial and Spatio-temporal Epidemiology. 2010 Jul; 1(2-3):151–161. https://linkinghub.elsevier.com/retrieve/pii/S1877584510000110, doi: 10.1016/j.sste.2010.03.009.

WHO, Global Vector Control Response 2017-2030. WHO; 2017.

Wilson AL, Courtenay O, Kelly-Hope LA, Scott TW, Takken W, Torr SJ, Lindsay SW. The importance of vector control for the control and elimination of vector-borne diseases. PLoS Negl Trop Dis. 2020 Jan; 14(1):e0007831. doi: 10.1371/journal.pntd.0007831.

1149 **Xia H**, Wang Y, Atoni E, Zhang B, Yuan Z. Mosquito-Associated Viruses in China. Virol Sin. 2018 Feb; 33(1):5–20.
1150 doi: 10.1007/s12250-018-0002-9.

1151 **Xiao P**, Han J, Zhang Y, Li C, Guo X, Wen S, Tian M, Li Y, Wang M, Liu H, Ren J, Zhou H, Lu H, Jin N. Metagenomic
1152 Analysis of Flaviviridae in Mosquito Viromes Isolated From Yunnan Province in China Reveals Genes From
1153 Dengue and Zika Viruses. Front Cell Infect Microbiol. 2018; 8:359. doi: 10.3389/fcimb.2018.00359.

1154 **Xiao P**, Li C, Zhang Y, Han J, Guo X, Xie L, Tian M, Li Y, Wang M, Liu H, Ren J, Zhou H, Lu H, Jin N. Metagenomic
1155 Sequencing From Mosquitoes in China Reveals a Variety of Insect and Human Viruses. Front Cell Infect
1156 Microbiol. 2018; 8:364. doi: 10.3389/fcimb.2018.00364.

1157 **Yang Z**. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites:
1158 Approximate methods. J Mol Evol. 1994 Sep; 39(3):306–314. https://doi.org/10.1007/BF00160154, doi:
1159 10.1007/BF00160154.

1160 **Ziv J**, Lempel A. Compression of individual sequences via variable-rate coding. IEEE transactions on Information
1161 Theory. 1978; 24(5):530–536.

1162 **Zídková L**, Cepicka I, Szabová J, Svobodová M. Biodiversity of avian trypanosomes. Infect Genet Evol. 2012 Jan;
1163 12(1):102–112. doi: 10.1016/j.meegid.2011.10.022.

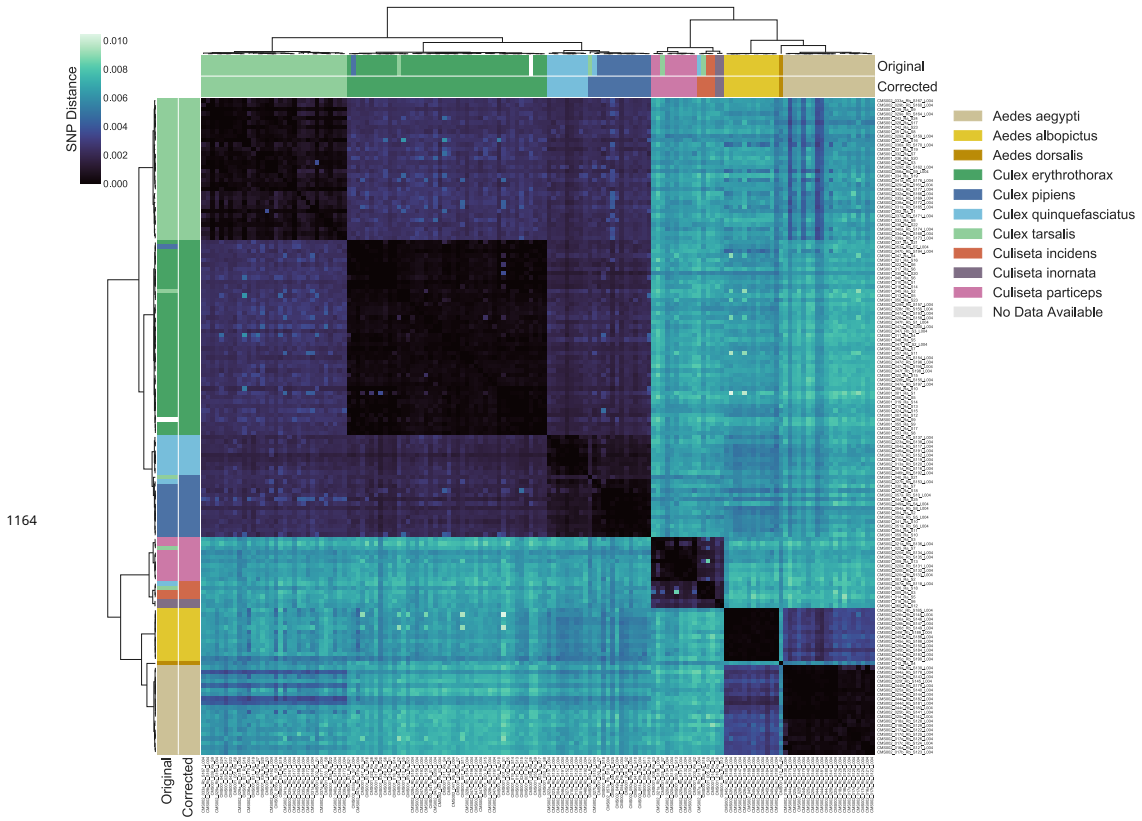


Figure 1–Figure supplement 1. Hierarchical clustering of pairwise single-nucleotide polymorphism (SNP) distances between samples estimated using SKA (***Harris, 2018***), ranging from 0 (no SNPs detected, black) to 0.01 (1% of comparable sites had SNPs, blue). Outside color bar indicates visual species label for each sample (Visual), inside color bar indicates consensus transcriptome species label for each sample's cluster (Computed). Computed mosquito species calls were used when Visual and Computed calls were discordant.

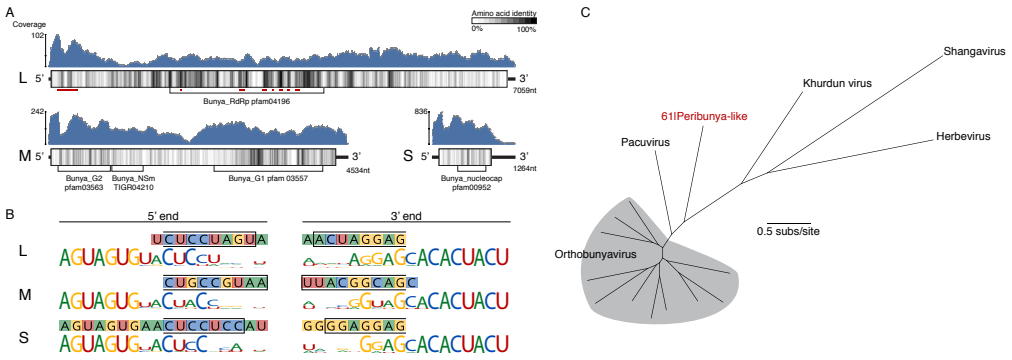


Figure 2–Figure supplement 1. Analysis of peribunya-like virus showing completeness. (A) Read coverage in blue across the length of each genome segment. Diagrams of genome segments display the longest ORF and indicate amino acid identity with a sliding window of 15 amino acids. Regions of homology to known domains are indicated by Pfam labels, and on the L segment, the locations of canonical RdRp motifs are indicated by red underlining. (B) Ends of the assembled contigs are shown (sequence in filled boxes). Regions complementary between 5' and 3' end for each segment are indicated by a black outline. These contig ends were aligned to the conserved end sequences of orthobunya and nearby peribunya viruses (logo diagram shows conservation after alignment of reference sequences). (C) Maximum likelihood phylogenetic tree with representative orthobunyaviruses and nearby peribunyaviruses.

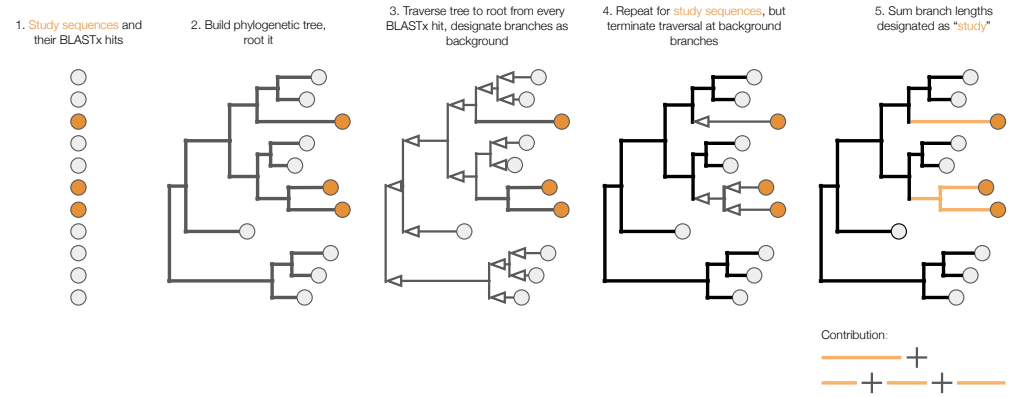


Figure 3–Figure supplement 1. Schematic describing method for branch length contribution calculation. Final calculation of branch length for a given group of viruses quantifies the contribution of evolutionary novelty, in units of substitutions/site.

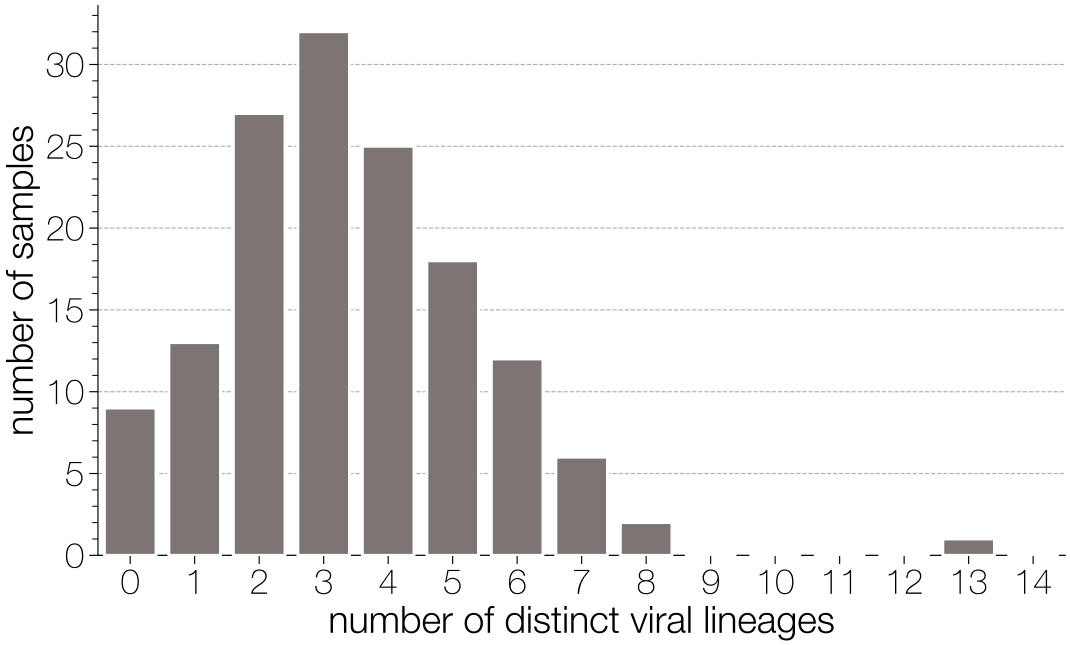


Figure 4–Figure supplement 1. Distribution of mosquitoes within the study in which no, one, or multiple viral lineages were detectable.

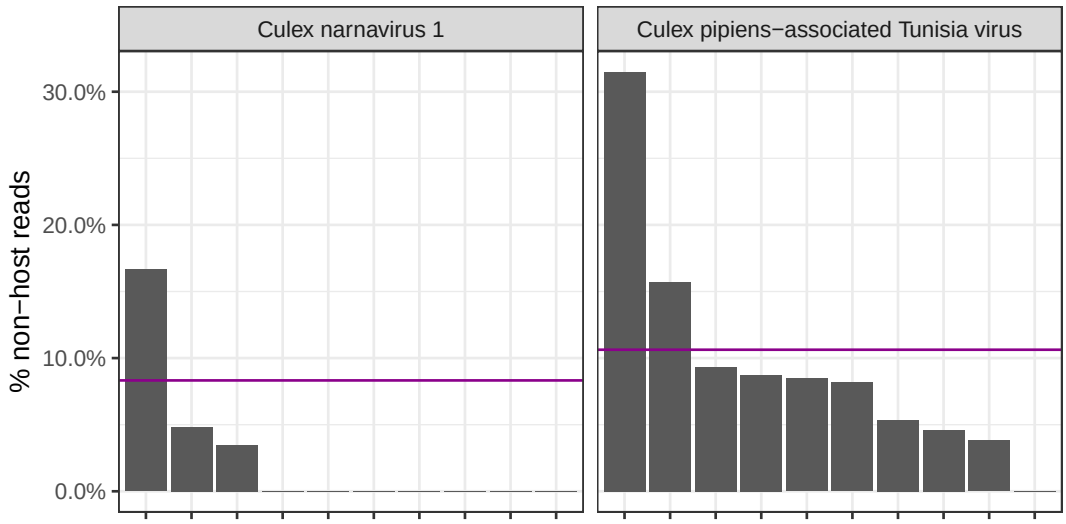
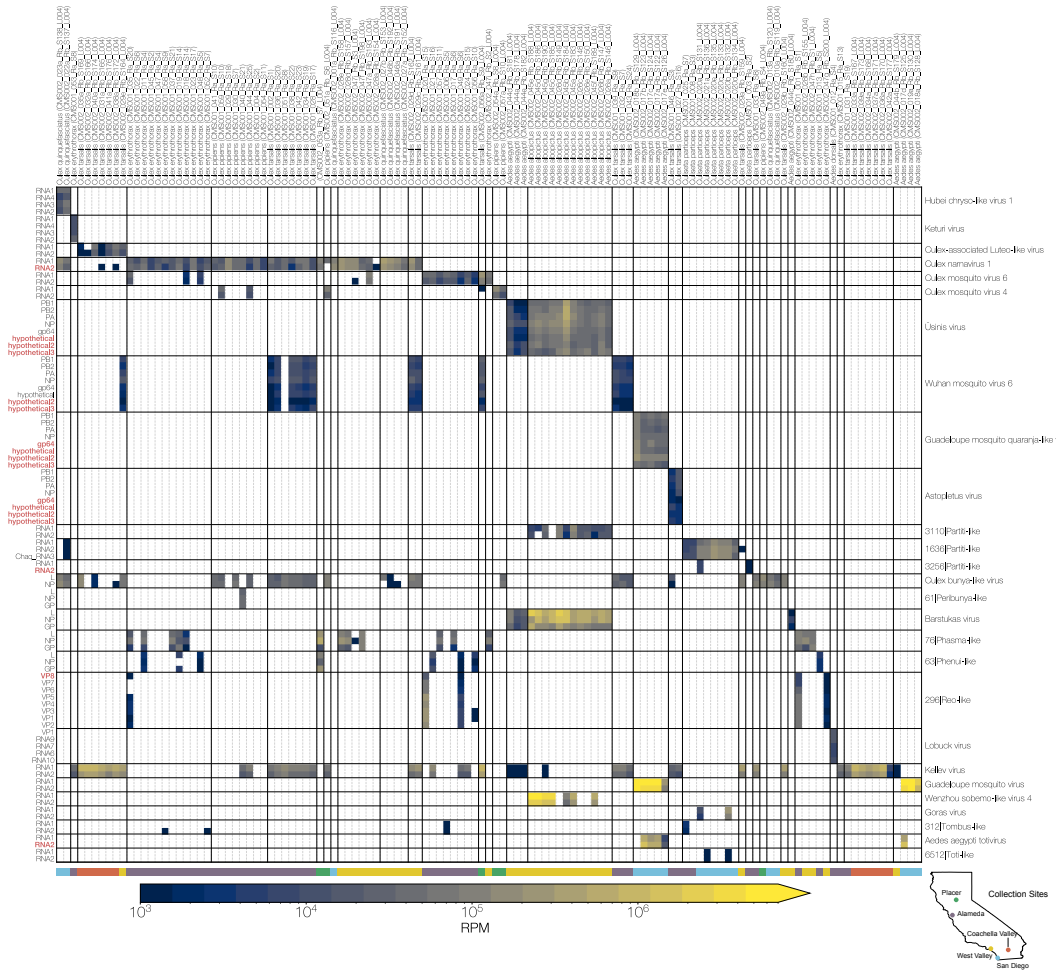


Figure 4–Figure supplement 2. An example of viruses with similar bulk abundance but different prevalence. Viral abundance is often calculated based on bulk mosquito sequencing, which does not provide information about the prevalence or heterogeneity in abundance of a virus across the mosquito population. Both Culex narnavirus 1 and Culex pipiens-associated Tunisia virus were found in *Culex erythrothorax* mosquitoes at the same collection site in West Valley. A total of 10 *C. erythrothorax* mosquitoes were collected at this site, and the bulk abundances as calculated by the mean % non-host reads averaged across the 10 mosquitoes for Culex narnavirus 1 and Culex pipiens-associated Tunisia virus were 8.3% and 10.6%, respectively. The prevalence differed more, at 30% and 90%, respectively.



1169

Figure 6–Figure supplement 1. Rows indicate groups of contigs clustered based on nucleotide identity (>99%) and labeled with putative identity of protein products. Columns are samples with computed mosquito species indicated. Cells are colored by read count per million reads (RPM) with color bar legend at the bottom (maximum shown in color legend is 7 million, highest RPM observed is 15 million). Clusters of rows likely comprising a single genome are separated from other such putative genomes by black horizontal lines with identified or proposed virus names. Black vertical lines delineate blocks of neighbouring samples with the same computed mosquito species. Sampling locations are indicated by colored rectangles at the bottom, above the color bar legend.

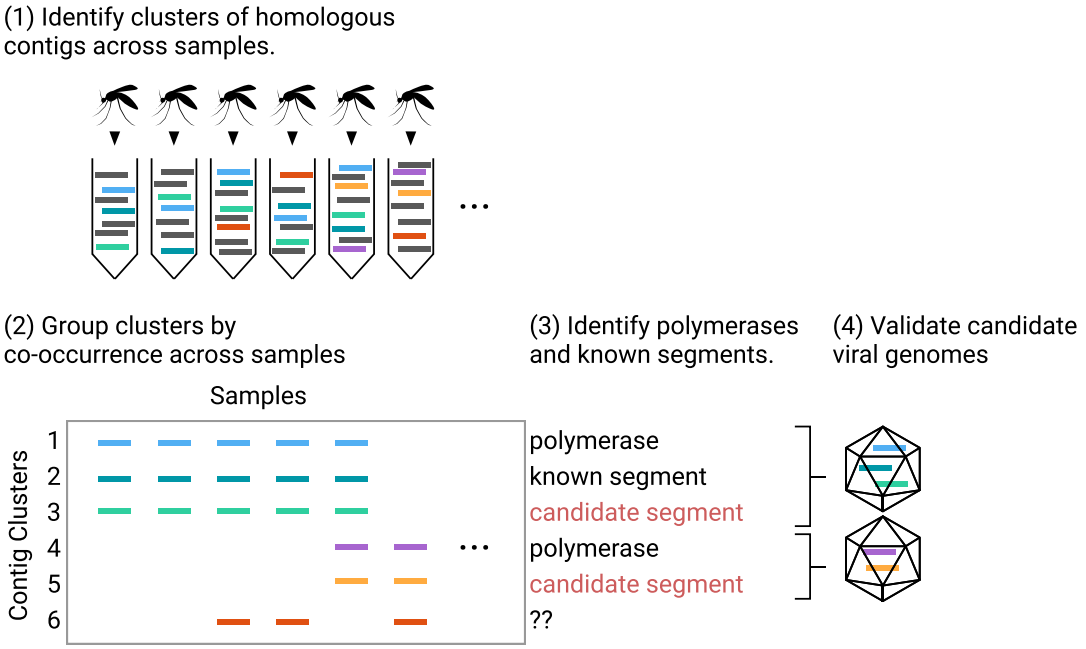


Figure 6–Figure supplement 2. Contigs are assembled from each mosquito and grouped into clusters of high sequence identity to find identical and variant sequences of the same viral segment across different mosquitoes. Here, sequences that come from repeated sampling of the same viral segment in different mosquitoes are shown in the same color. Contig clusters are then grouped together based on their co-occurrence across different samples. After identifying known viral segments using Hidden Markov Models (for RdRps) and BLASTx, the remaining contig clusters that co-occur with the known segments are assigned to the same putative viral genome. Finally, bioinformatic validation included consideration of features that supported viral assignment, such as sequence and ORF length, putative protein domains, and consistency with known genomic structure for other viruses in the same family. See Methods for details.

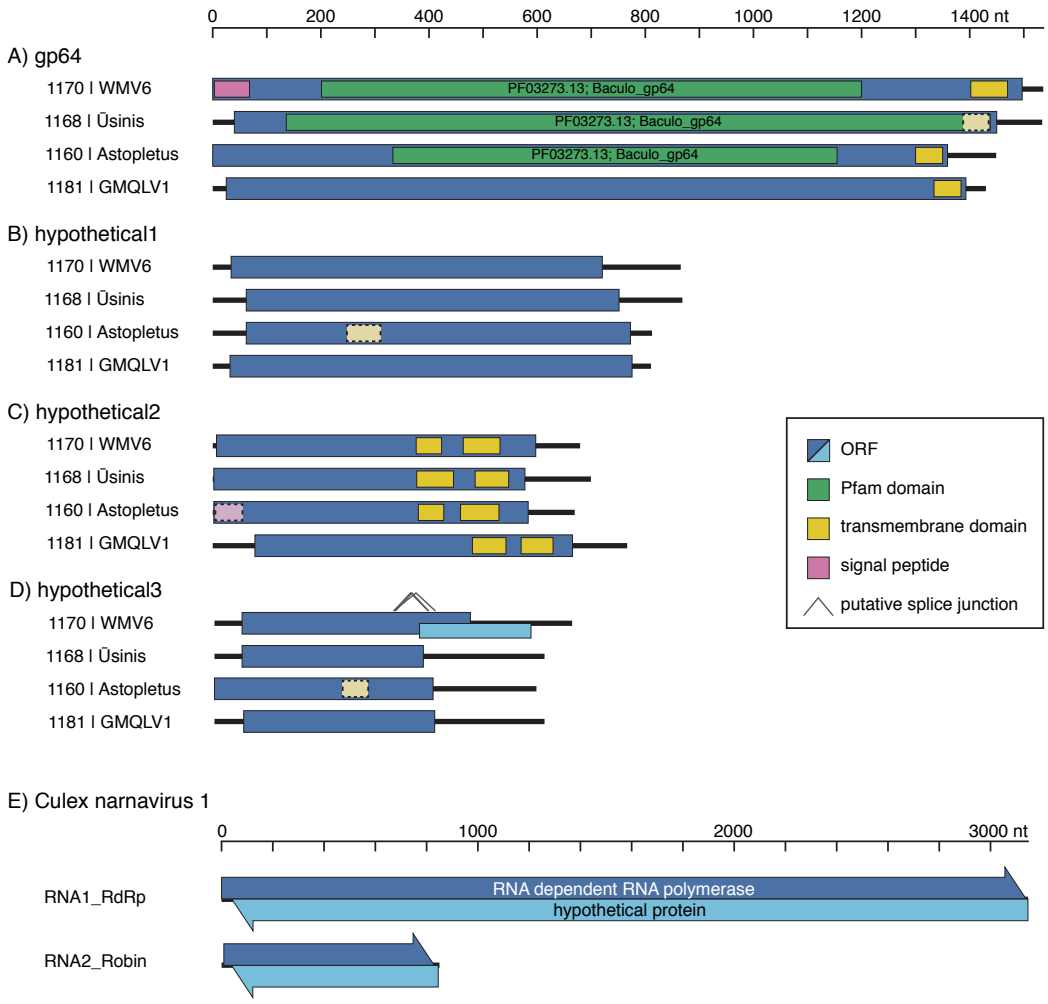


Figure 6–Figure supplement 3. Diagrams of newly discovered genome segments for quaranjaviruses in this study, including Wuhan Mosquito Virus 6 (WMV6), Ūsinis, Astopletus, and Guadeloupe Mosquito Quaranja-Like Virus 1 (GMQLV1) (panels A-D, respectively), and Culex narnavirus 1 (panel E), are shown (displayed 5′ to 3′, left to right, not aligned to each other). A-D, The longest ORFs of the new quaranjavirus segments were analyzed for the presence of transmembrane domains and similarity to known domains (see Methods for details). Weaker support for these predictions is indicated by lighter color and dashed border. In the hypothetical 3 segment of WMV6, putative splicing events could result in a jump to a frameshifted ORF, which would substantially increase the coding region of the segment. E, In both segments of the Culex narnavirus 1, long uninterrupted ORFs are found on both RNA strands (arrowheads indicate 5′ to 3′ direction), and in each case the codon boundaries are aligned.

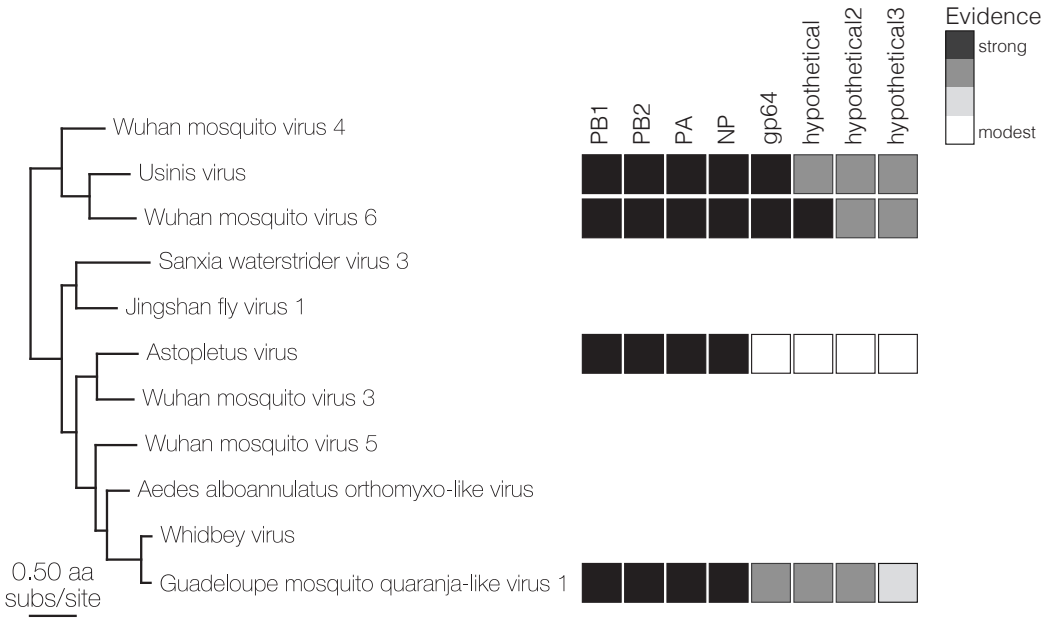


Figure 6–Figure supplement 4. RdRp-based maximum likelihood tree spanning the quaranjaviruses in this study for which 8 segments were recovered. The 8 boxes in the graphic to the right of the tree correspond to each of these 8 genome segments, which are labeled at top right. Levels of evidence for the existence of each segment is highlighted by a grey scale shading of the segment boxes, ranging from strong (black) to modest (white), based on: homology to BLAST hits from NCBI nt/nr database (black); assembly of contigs >500nt from SRA samples containing the RdRp, or co-occurrence in >10 samples in our data (dark grey); reads mapping from SRA and co-occurrence in 5 samples in our data (light grey); or co-occurrence in <5 samples and shared protein domains (white).

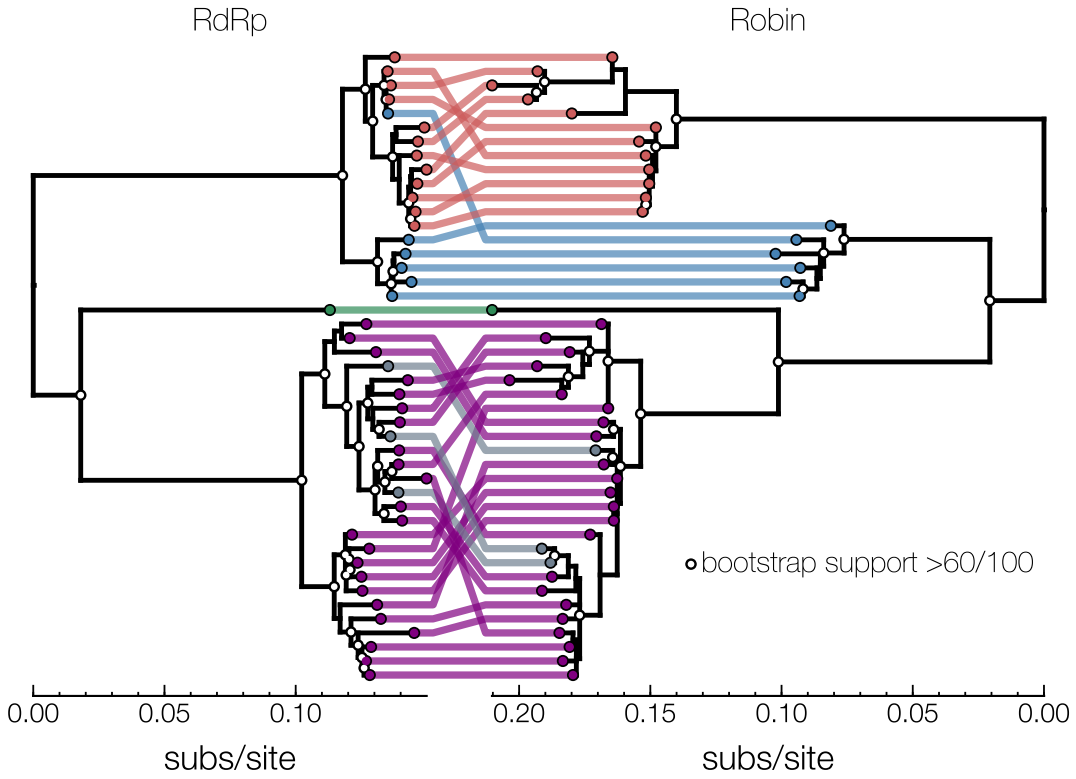


Figure 6–Figure supplement 5. Narnavirus Robin segment. Tanglegram shows two mid-point rooted nucleotide phylogenies of Culex narnavirus 1 - RdRp segment on left and Robin segment on right. Tips are colored based on visual presence of clusters on longer branches - clusters in red, blue, green, and purple are sequences reported in this study. Grey tips in the RdRp phylogeny were published by other studies, and their Robin segment counterpart was assembled here from corresponding SRA entries. Reassortment between the two segments is largely restricted to lineages within clusters, but reassortment also took place in either the red or the blue lineage. One possibility is a replacement of a red/blue lineage RdRp with a blue/red lineage or a replacement of red Robin with a diverged and unsampled lineage. White dots on nodes indicate bootstrap support >60 out of 100 replicates. Scale at the bottom of each tree is nucleotide substitutions per site.

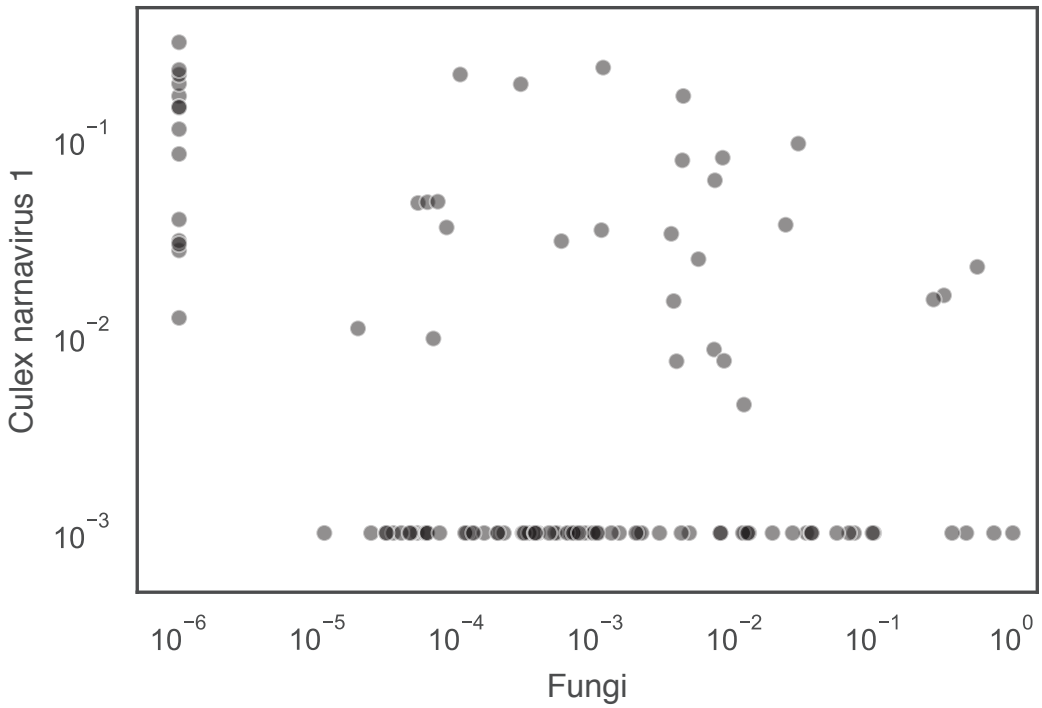


Figure 6–Figure supplement 6. Proportion of non-host reads mapped to contigs of Culex narnavirus 1 versus proportion of non-host reads mapped to contigs of any fungus. Each point represents a single mosquito. For visualization purposes, values of zero are displayed at 1e-6 for fungi and 1e-3 for Culex narnavirus 1.