RESEARCH ARTICLE

Profiling of RNA Viruses in Biting Midges (*Ceratopogonidae*) and Related Diptera from Kenya Using Metagenomics and Metabarcoding Analysis

Solomon K. Langat,[a,b] Fredrick Eyase,[c,d] Wallace Bulimo,[b,e] Joel Lutomiah,[b] Samuel O. Oyola,[f] Mabel Imbuga,[a] Rosemary Sang[b]

[a]Department of Biochemistry, Jomo Kenyatta University of Agriculture and Technology, Nairobi, Kenya
[b]Centre for Virus Research, Kenya Medical Research Institute, Nairobi, Kenya
[c]Institute of Biotechnology Research, Jomo Kenyatta University of Agriculture and Technology, Nairobi, Kenya
[d]Department of Emerging Infectious Diseases, United States Army Medical Research Directorate—Africa, Nairobi, Kenya
[e]Department of Biochemistry, University of Nairobi, Nairobi, Kenya
[f]International Livestock Research Institute, Nairobi, Kenya

ABSTRACT Vector-borne diseases (VBDs) cause enormous health burden worldwide, as they account for more than 17% of all infectious diseases and over 700,000 deaths each year. A significant number of these VBDs are caused by RNA virus pathogens. Here, we used metagenomics and metabarcoding analysis to characterize RNA viruses and their insect hosts among biting midges from Kenya. We identified a total of 15 phylogenetically distinct insect-specific viruses. These viruses fall into six families, with one virus falling in the recently proposed negevirus taxon. The six virus families include *Partitiviridae*, *Iflaviridae*, *Tombusviridae*, *Solemoviridae*, *Totiviridae*, and *Chuviridae*. In addition, we identified many insect species that were possibly associated with the identified viruses. *Ceratopogonidae* was the most common family of midges identified. Others included *Chironomidae* and *Cecidomyiidae*. Our findings reveal a diverse RNA virome among Kenyan midges that includes previously unknown viruses. Further, metabarcoding analysis based on COI (cytochrome *c* oxidase subunit 1 mitochondrial gene) barcodes reveal a diverse array of midge species among the insects used in the study. Successful application of metagenomics and metabarcoding methods to characterize RNA viruses and their insect hosts in this study highlights a possible simultaneous application of these two methods as cost-effective approaches to virus surveillance and host characterization.

IMPORTANCE The majority of the viruses that currently cause diseases in humans and animals are RNA viruses, and more specifically arthropod-transmitted viruses. They cause diseases such as dengue, West Nile infection, bluetongue disease, Schmallenberg disease, and yellow fever, among others. Several sequencing investigations have shown us that a diverse array of RNA viruses among insect vectors remain unknown. Some of these could be ancient lineages that could aid in comprehensive studies on RNA virus evolution. Such studies may provide us with insights into the evolution of the currently pathogenic viruses. Here, we applied metagenomics to field-collected midges and we managed to characterize several RNA viruses, where we recovered complete and nearly complete genomes of these viruses. We also characterized the insect host species that are associated with these viruses. These results add to the currently known diversity of RNA viruses among biting midges as well as their associated insect hosts.

KEYWORDS metagenomics, metabarcoding, biting midges, RNA viruses

Vector-borne diseases (VBDs) cause significant health and economic burden all over the world, with the tropical and subtropical regions bearing the heaviest burden. They account for more than 17% of all infectious diseases and are associated with more

Citation Langat SK, Eyase F, Bulimo W, Lutomiah J, Oyola SO, Imbuga M, Sang R. 2021. Profiling of RNA viruses in biting midges (*Ceratopogonidae*) and related Diptera from Kenya using metagenomics and metabarcoding analysis. mSphere 6:e00551-21. https://doi.org/10.1128/mSphere.00551-21.

Editor Benhur Lee, Icahn School of Medicine at Mount Sinai

Address correspondence to Solomon K. Langat, langatsolomon@gmail.com.

Received 15 June 2021
Accepted 15 September 2021
Published 13 October 2021

than 700,000 deaths every year (1, 2). Pathogens that cause VBDs are transmitted by arthropod vectors such as mosquitoes, sandflies, ticks, and biting midges (2, 3). Among these arthropod-transmitted pathogens, viruses account for a disproportionately high number of emerging human pathogens, with RNA viruses alone constituting the highest proportion, approximately 37% of all emerging human pathogens (4). The resurgence and spread of known and re-emerging arthropod-borne RNA viruses around the globe are now widely reported, with devastating consequences; yellow fever, dengue, chikungunya, Zika fever, Rift Valley fever, and Crimean-Congo hemorrhagic fever are the most common (5, 6). Biting midges are vectors of a range of human and livestock pathogens (7, 8). Three of the diseases transmitted by these vectors are currently listed by World Organization for Animal Health (OIE) as notifiable diseases, namely, bluetongue, African horse sickness, and epizootic hemorrhagic disease (9). Midges are generally small insects (approximately 1 to 3 mm in length) belonging to different families within the suborder Nematocera in the order Diptera (10). These families include *Ceratopogonidae* (biting midges), *Chironomidae* (nonbiting midges), and *Cecidomyiidae* (gall midges). *Ceratopogonidae* is the most important family because it contains members that are medically important vectors of disease-causing pathogens like *Mansonella* sp. parasites and viruses such as bluetongue, Oropouche, African horse sickness, epizootic hemorrhagic disease, and Schmallenberg viruses (8, 10, 11).

The response and mitigation strategies for fighting vector-borne diseases are currently dependent on surveillance programs. These programs allow early detection and control of vector-borne diseases as well as identification of any invasive vector species (12, 13). For insect vectors, routine surveillance approaches often involve identification of trapped insect vectors using morphological traits and subsequent screening for pathogens using cell culture and molecular detection methods (14). The challenge with these traditional approaches is that they are time-consuming and difficult to implement, especially when large numbers of specimens are to be processed. The advent and further development of high-throughput sequencing (HTS) technologies have provided powerful tools with enormous potential to overcome these challenges. In this study, we used HTS-based sequencing methods, metagenomics and metabarcoding. Metagenomics is an unbiased approach to sequencing of all the DNA or cDNA in a given sample (15, 16). This method has proven quite useful and has led to rapid progress in virus discovery, including identification of novel pathogens that have been implicated in major outbreaks (17). Metabarcoding, on the other hand, is a method of performing massively parallel sequencing of mixed biological samples that combines the HTS technologies with the traditional DNA barcoding method (18). This strategy enables the generation of a large number of individual barcode sequences for the various insect specimens pooled in a single reaction. Analysis of the barcode sequences generated allows species identification of insects in the given pool (18). The cytochrome *c* oxidase subunit 1 mitochondrial gene (COI) has become the marker of choice for most DNA barcoding studies. The COI marker is often preferred because of its ability to sufficiently discriminate between closely related species, particularly vertebrate and invertebrate species (19–21). Additionally, this target is widely represented in reference databases, with millions of COI reference sequences currently available in public databases (22). Adoption of third-generation sequencing technology allows further improvement of metabarcoding method, with long-read sequencing providing the advantage of higher resolution as a result of the sequencing of longer DNA fragments. In this study, we applied metagenomics and metabarcoding techniques to analyze field-collected specimens of midges obtained from different sites in Kenya. We sought to detect and characterize RNA viruses harbored by these midges and identify their associated vector species based on their COI markers.

RESULTS

RNA viruses identified in biting midges. A total of 3,351 midge specimens were processed in this study. Of these, 1,063 originated from Turkana and 892 were from

TABLE 1 Viruses identified in this study

Strain	Site	Family classification	Accession no.	Closest hit	% identity (RdRp; aa)
Turkana_1	Turkana	*Solemoviridae*	MF893251.1	Medway virus	59.05
Turkana_2	Turkana	*Chuviridae*	KX924630.1	Chuvirus Mos8Chu0	42.39
Turkana_3	Turkana	Tombus-like	KX235518.1	Diaphorina citri associated C virus	55.24
Turkana_4	Turkana	*Totiviridae*	MK440653.1	Lindangsbacken virus	50.79
Turkana_5	Turkana	*Iflaviridae*	NC_024016.1	Heliconius erato iflavirus	85.62
Turkana_6	Turkana	*Partitiviridae*	LC533398.1	Lichen partiti-like RNA virus	46.14
Turkana_7	Turkana	Partiti-like	KX884215.1	Hubei partiti-like virus 45	52.74
Turkana_8	Turkana	Iflavi-like	MN784069.1	Redbank virus	66.3
Turkana_9	Turkana	Negevirus	MT344121.1	Sandewavirus dungfly1	39.13
Turkana_10	Turkana	Picorna-like	MH614292	Boghill Burn virus	84.27
Baringo_1	Baringo	Partiti-like	MF344586.1	Araticum virus	56.01
Budalangi_1	Budalangi	Iflavi-like	MN784065.1	Redbank virus	70.84
Isiolo_1	Isiolo	*Iflaviridae*	NC_040574.1	Culex Iflavi-like virus 4	97.7
Kacheliba_1	Kacheliba	Iflavi-like	NC_033201.1	Hubei picorna-like virus 38	46.37
Kacheliba_2	Kacheliba	*Partitiviridae*	JX658566.1	Grapevine partitivirus	51.61

Baringo, while Isiolo, Budalangi, and Kacheliba yielded 640, 600, and 156 specimens, respectively. These specimens were used to create bulk pools representing each of the five sites, which were then subjected to high-throughput sequencing. Sequence assembly and analysis led to generation of 15 distinct virus genomes, identified based on the presence of the RNA-dependent RNA polymerase (RdRp) gene, which is the hallmark of all RNA viruses (23, 24). The RdRp gene of the identified viruses was found to be 38.81% to 97.7% similar to previously sequenced viruses available in GenBank. Among the 15 viruses identified, 10 were positive-sense RNA (+ssRNA) viruses, with viruses in the Picornavirales order forming the majority. Viruses in the order *Picornavirales* included 5 that were similar to those in the family *Iflaviridae* and 1 which was similar to those in the family *Picornaviridae*. Turkana_5, which was obtained from a pool of midges from Turkana, was 85.62% similar to Heliconius erato iflavirus (Table 1), an iflavivirus detected in *Heliconius* butterflies in Costa Rica (25). On the other hand, Isiolo_1, which was obtained from a pool of midges from Isiolo, showed a 97.7% similarity to Culex iflavi-like virus 4 which was previously obtained from *Culex* sp. mosquitoes in the United States (26). Two viruses from Budalangi and Turkana, specifically, Budalangi_1 and Turkana_8, showed high similarity to the Redbank virus (Table 1), with 70.84% and 66.3% similarity, respectively, to this previously identified virus. Redbank virus is an unclassified iflavi-like virus recently identified in mosquito fecal microbiota in Australia (27). Kacheliba_1 was the other iflavi-like virus that had relatively low similarity to published sequences. It had 46.6% similarity to Hubei picorna-like virus 38 (NC_033201.1). Also falling in the order *Picornavirales* was Turkana_10, which was identified from a pool of midges from Turkana. This virus had 84.27% similarity to the Boghill Burn virus, an unclassified virus in the family *Picornaviridae* which was identified in *Bombus* sp. bees in Scotland (28). All 6 viruses in the order *Picornavirales* had genome architectures similar to those of other viruses in this order. They have a single polyprotein flanked by 3′ and 5′ untranslated regions (UTRs), and the genome length of these viruses is approximately 9 kb (Fig. 1).

Other +ssRNA viruses identified in this study include *Solemoviridae*, *Tombusviridae*, and a Negevirus-like virus. Turkana_1 showed similarity to members of the unclassified family *Solemoviridae*. Specifically, it had 59.05% similarity to the Medway virus. *Solemoviridae* family members have been known to be nonsegmented +ssRNA viruses (29). However, similar to recently described sobemo-like viruses, such as Atrato sobemo-like virus 1 and Hubei sobemo-like virus 48 (24), Turkana_1 contained two segments. Further, we also obtained a 1,142-bp sequence belonging to a virus in the family *Tombusviridae*. Turkana_3 had 55.24% similarity to Diaphorina citri-associated C virus. We also identified a Negevirus-like virus, Turkana_9, from a pool of midges from Turkana. Turkana_9 had considerably low similarity to other negeviruses available in

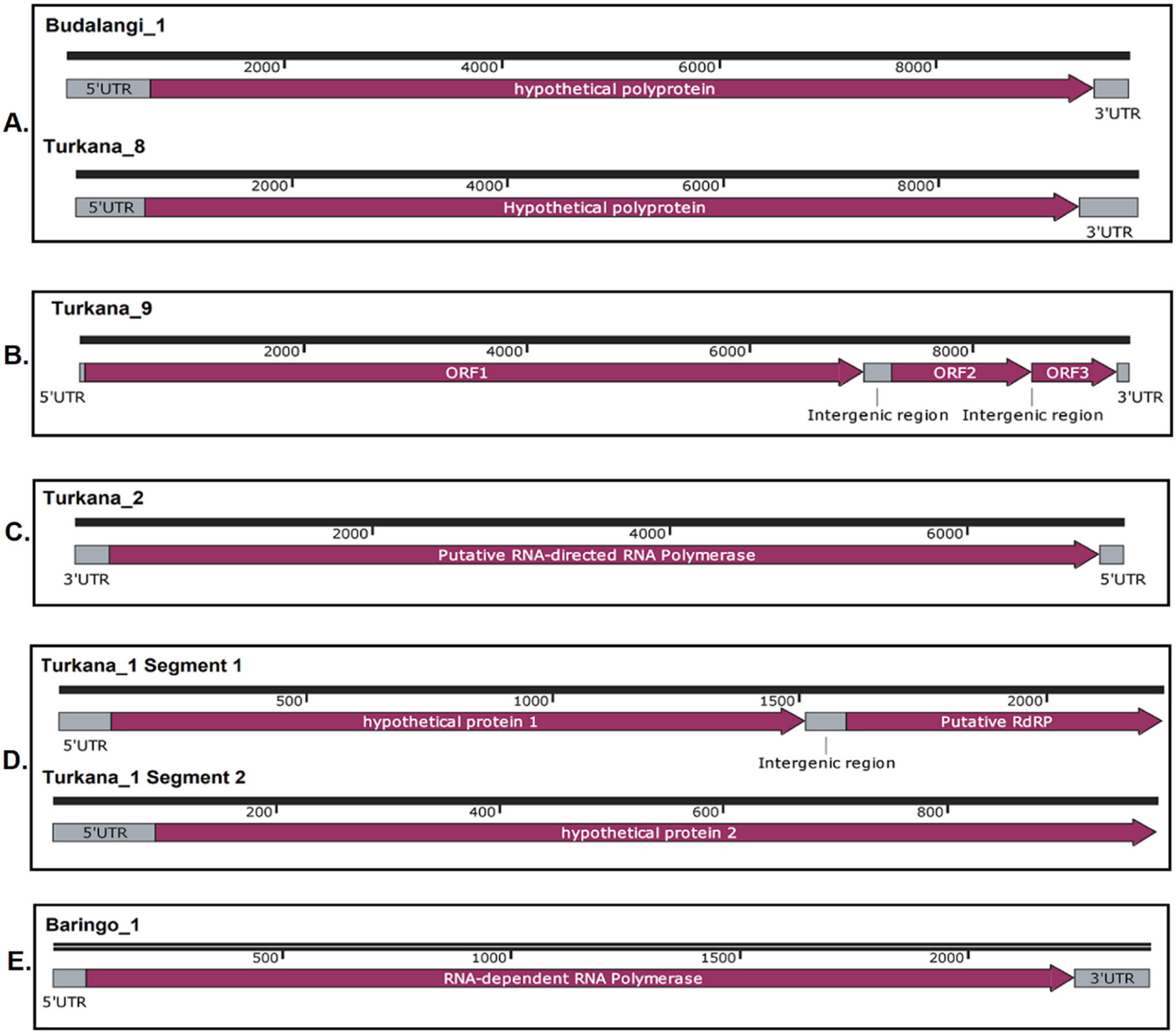

FIG 1 Genome architectures of the representative genomes of complete iflavi-like viruses (A), complete novel negevirus genome (B), partial *Chuviridae* genome showing the RNA-dependent RNA polymerase coding segment (C), a novel *Solemoviridae* virus with the two putative segments (D), and the representative partial *Partitiviridae* genome showing the RNA-dependent RNA polymerase coding segment (E). Similar organization was observed in the other *Partitiviridae* genomes obtained in the study.

GenBank. More specifically, it had 39.13% similarity to Sandewavirus dungfly 1, a negevirus obtained from dung fly in the Arctic Yellow River Station (30). Despite this low similarity, however, the genome architecture of Turkana_9 was similar to that of other negeviruses, with three open reading frames (ORFs) separated by two intergenic regions of various lengths (Fig. 1).

In this study, we also identified one negative-sense RNA (−ssRNA) virus specifically falling in the recently described family *Chuviridae* (31). Turkana_2 had 42.39% similarity to Chuvirus Mos8Chu0 (Table 1). Chuvirus Mos8Chu0 (KX924630.1) is a bisegmented virus belonging to the family Chuviridae, and it was obtained from *Culiseta minnesotae* mosquitoes. In this study, however, we obtained only a partial sequence corresponding to the entire L segment of Chuvirus Mos8Chu0. This segment contains a single ORF which is the putative RNA-dependent RNA polymerase gene (Fig. 1).

Other viruses identified in this study are five double-strand RNA (dsRNA) viruses that include *Totiviridae* and *Partitiviridae*. Turkana_4 is a *Totiviridae* member which was identified in a pool of midges from Turkana. The virus showed a 50.79% similarity to the Lindangsbacken virus, which is an unclassified *Totiviridae* member. The four *Partitiviridae* viruses detected in this study originated from Turkana, Kacheliba, and Baringo, and these viruses were highly diverse, with a significantly low similarity of 46.14% to Lichen partiti-like virus for Turkana_6 and a higher similarity of 56.01% to Araticum virus for Baringo_1. The other two viruses included Turkana_7, which had 52.74% similarity to Hubei partiti-like virus 45, and Kacheliba_2, which had 51.61% similarity to Grapevine partitivirus. *Partitiviridae* contains viruses with two segments, dsRNA1 and dsRNA2 (32). In this study, however, we obtained only the segment corresponding to the dsRNA1 segment whose ORF codes for the RNA-dependent RNA polymerase. This finding was similar for all four *Partitiviridae* viruses obtained in the study.

Phylogenetic analysis of the identified RNA viruses. Phylogenetic analysis of the newly discovered viruses with other closely related viruses available in GenBank placed 7 of the new viruses in five different families and one in the recently proposed taxon Negevirus (33). More specifically, the six families that these identified viruses fall into or are related to include *Partitiviridae*, *Iflaviridae*, *Tombusviridae*, *Solemoviridae*, *Totiviridae*, and the recently described family *Chuviridae*. Except for Turkana_5, which clustered with iflaviruses (Fig. 2), the majority of the other identified viruses showed high similarity and clustered with diverse virus strains which are yet to be classified within the specific RNA virus families. More specifically, Turkana_6 and Kacheliba_2 clustered with unclassified members of the family *Partitiviridae* (Fig. 3), Isiolo_1 clustered with unclassified members of the family *Iflaviridae* (Fig. 2), Turkana_1 clustered with unclassified members of the family *Solemoviridae*, Turkana_4 clustered with unclassified members of the family *Totiviridae*, and Turkana_2 clustered with unclassified members of the family *Chuviridae* (Fig. 3).

Eight of the identified viruses clustered with diverse virus strains that are yet to be classified into specific families. These included partiti-like, iflavi-like, picorna-like, and tombus-like viruses. Turkana_7 and Baringo_1 clustered with unclassified Partiti-like viruses (Fig. 3); Kacheliba_1, Turkana_8, and Budalangi_1 clustered with unclassified iflavi-like viruses; Turkana_3 clustered with unclassified tombus-like viruses; and Turkana_10 clustered with unclassified picorna-like viruses (Fig. 2). Turkana_9 was closely related and clustered with negeviruses (Fig. 2), forming a single clade with Sandewavirus, Tanay virus, Dezidougou virus, and Goutanap virus, all of which are insect-specific viruses (ISVs).

Community composition of the midge pools. Metabarcoding analysis resulted in diverse reads that were assigned to different families of midges, including *Chironomidae*, *Ceratopogonidae*, and *Cecidomyiidae*. All the reads in the Baringo pool were assigned to the family *Ceratopogonidae*, with 65.46, 19.81, and 13.26% of the reads being assigned to *Culicoides leucostictus*, *Culicoides pycnostictus*, and *Culicoides nivosus*, respectively. Similarly, all the reads from the Budalangi site were classified as *Ceratopogonidae* and specifically as *Culicoides leucostictus*. At the Kacheliba site, all the reads identified were classified as *Cecidomyiidae*. All the reads showed similarity to unclassified members within the family *Cecidomyiidae* (Table 2). Isiolo and Turkana sites had relatively diverse reads that were classified as different midge species. The family *Ceratopogonidae* accounted for 44.07%, while *Cecidomyiidae* accounted for 21.03% of the reads at the Isiolo site. The remaining fraction were assigned to the *Chironomidae*. In Turkana, reads assigned to *Ceratopogonidae* were common, with 68.49%. The proportions of reads assigned to *Chironomidae* and *Cecidomyiidae* in Turkana were 25.64 and 5.88%, respectively.

The specimen sequences obtained were classified into operational taxonomic units (OTUs) (see Materials and Methods). Overall, we obtained a total of 187 OTUs belonging to the different families of midges. The Turkana and Isiolo sites had 51 OTUs each, while Baringo and Kacheliba had 36 OTUs each (Table 2). The Budalangi site had only 13 OTUs, despite a relatively high number of specimens processed as well as of total reads obtained (Table 2). In this study, the findings were generally consistent with the number of specimens that were processed in each of the sites (Table 2). The species richness of the five different sites was variable, with Isiolo and Turkana having relatively diverse species of

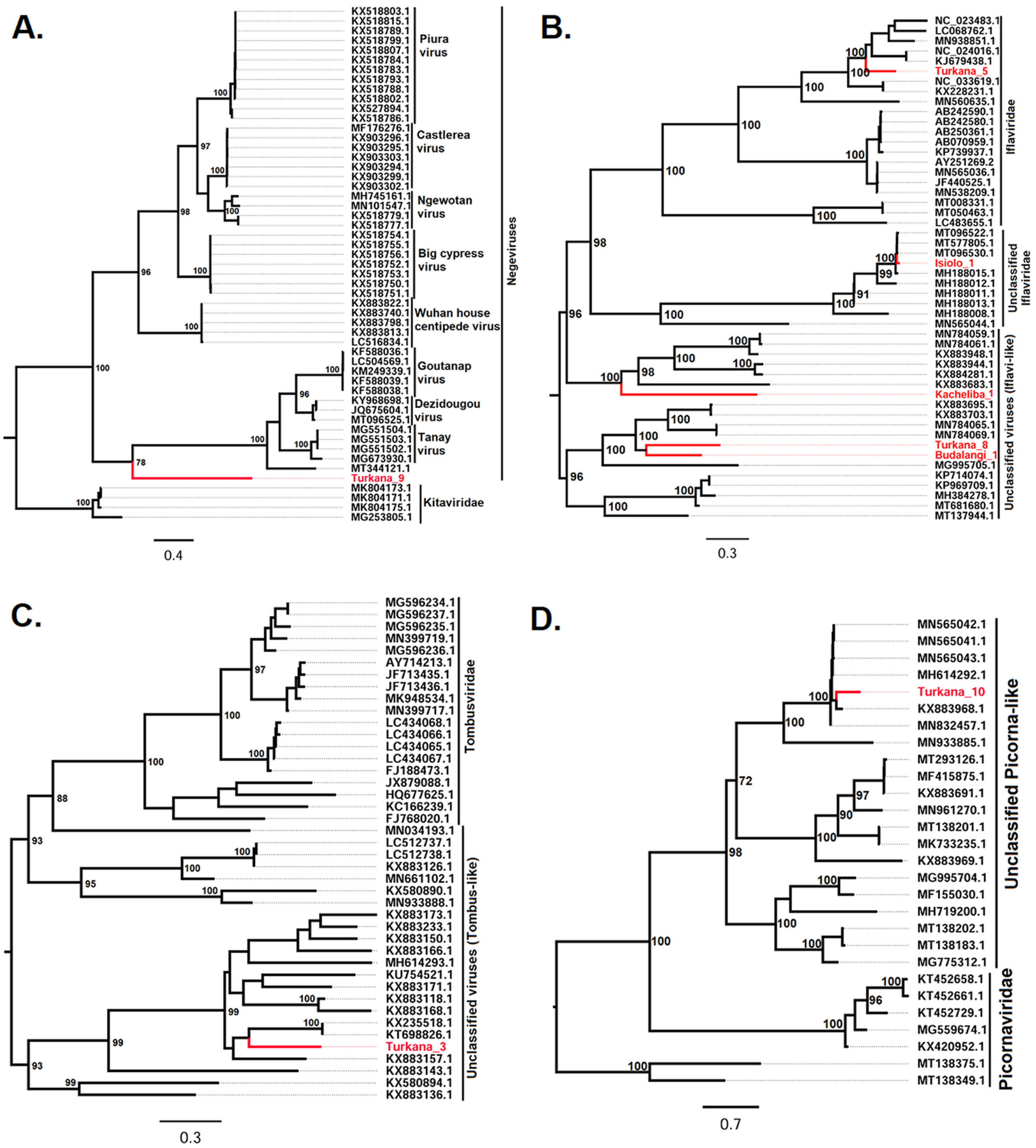

FIG 2 Midpoint rooted phylogenies for negeviruses (A), *Iflaviridae* (B) *Tombusviridae* (C), and *Picornaviridae* (D) and related sequences. The trees were inferred based on 1,000 bootstrap replicates and an approximate-likelihood-ratio test. Confidence values are shown in the tree nodes, and the sequences obtained from the study are in red.

midges. The insects in Isiolo included *Ceratopogonidae* comprising *Culicoides leucostictus*, *Culicoides oxystoma*, *Culicoides similis*, and unclassified *Forcipomyia* sp. The *Chironomidae* family insects obtained in Isiolo included *Ablabesmyia* sp., *Polypedilum* sp., *Tanytarsus* sp., and others that remain unclassified within the family *Chironomidae*. We also identified 21 OTUs in Isiolo that had high similarities to unclassified members of the family

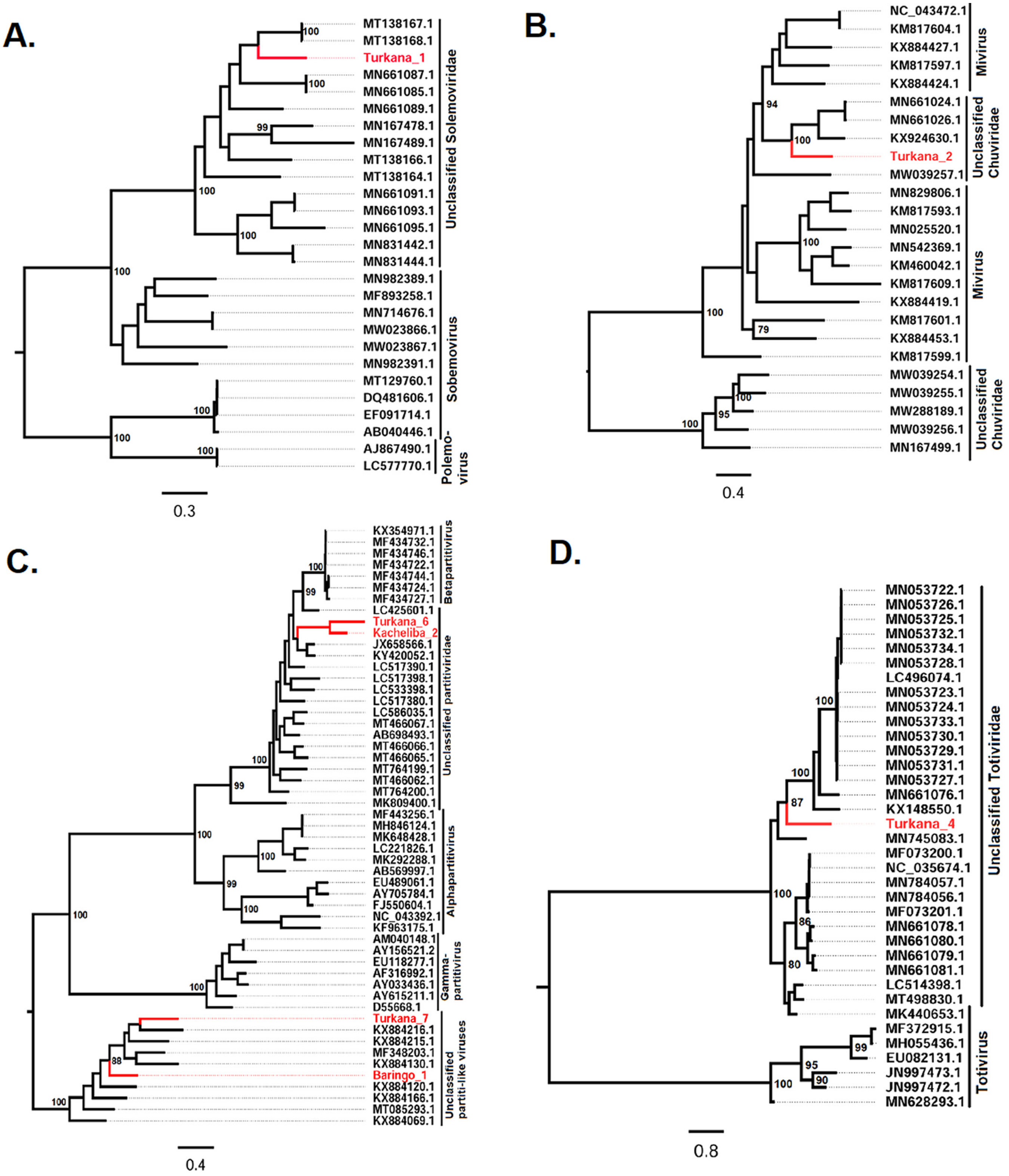

FIG 3 Midpoint rooted phylogenies for *Solemoviridae* (A), *Chuviridae* (B), *Partitiviridae* (C), and *Totiviridae* (D) and related sequences. The trees were inferred based on 1,000 bootstrap replicates and an approximate-likelihood-ratio test. Confidence values are shown in the tree nodes, and the sequences obtained from the study are in red.

TABLE 2 Community composition of the different vector pools for each of the sites

Site	No. of specimens[a]	Family	Species	Sequence abundance	Fraction of COI reads (%)	OTU abundance[b]
Isiolo	640 (114)	*Cecidomyiidae*	*Cecidomyiidae* sp.	637	21.03	21
		Ceratopogonidae	*Culicoides leucostictus*	260	8.58	2
			Culicoides oxystoma	33	1.09	1
			Culicoides similis	124	4.09	2
			Forcipomyia sp.	918	30.31	4
		Chironomidae	*Ablabesmyia* sp.	53	1.75	3
			Chironomidae sp.	343	11.32	8
			Polypedilum sp.	501	16.54	9
			Tanytarsus sp.	160	5.28	1
Baringo	892 (873)	*Ceratopogonidae*	*Culicoides bedfordi*	109	1.46	7
			Culicoides leucostictus	4,871	65.46	17
			Culicoides nivosus	987	13.26	4
			Culicoides pycnostictus	1,474	19.81	8
Turkana	1,063 (884)	*Cecidomyiidae*	*Cecidomyiidae* sp.	212	5.88	8
		Ceratopogonidae	*Culicoides kingi*	260	7.21	6
			Culicoides leucostictus	1,010	27.99	9
			Culicoides sp.	102	2.83	1
			Culicoides nivosus	354	9.81	6
			Culicoides schultzei	745	20.65	14
		Chironomidae	*Ablabesmyia* sp.	119	3.3	2
			Microchironomus sp.	806	22.34	5
Kacheliba	156 (6)	*Cecidomyiidae*	*Cecidomyiidae* sp.	1110	100	36
Budalangi	600 (600)	*Ceratopogonidae*	*Culicoides leucostictus*	4940	100	13

[a] The total number of specimens processed for each site is given, with the number of specimens morphologically classified as *Culicoides* sp. in parentheses.
[b] Number of consensus sequences obtained after sequence clustering.

Cecidomyiidae. In Turkana, we identified members of *Ceratopogonidae* that included *Culicoides kingi*, *Culicoides leucostictus*, *Culicoides nivosus*, *Culicoides schultzei*, and unclassified *Culicoides* sp. In addition, *Chironomidae* family OTUs were identified that included *Ablabesmyia* sp. and *Microchironomus* sp. The Baringo and Budalangi sites had OTUs that were all assigned to *Ceratopogonidae*. *Culicoides leucostictus* was identified in both of these sites. In addition, the Baringo site had *Culicoides bedfordi*, *Culicoides nivosus*, and *Culicoides pycnostictus*. At the Kacheliba site, all the OTUs identified belonged to the family *Cecidomyiidae*. More specifically, these OTUs showed high similarity to unclassified members of the family *Cecidomyiidae*.

DISCUSSION

High-throughput screening and timely surveillance of viruses and insect vectors are critical for detecting vector-borne diseases. The methods that are currently applied in surveillance and screening of these diseases are dependent on cell culture, serology, and molecular detection methods (34). These methods are quite laborious and they are also limited, given that they can identify only viruses that can be cultured using currently available tools as well as those that have previously been isolated and diagnostic methods developed for them. In this study, we applied metagenomics and metabarcoding approaches to characterize RNA viruses and their associated insect hosts. Even though no known pathogenic viruses were detected in this study, we were able to detect numerous RNA viruses, some of which can be classified as novel. Further, several host species were detected in the pools from which the RNA viruses were identified, providing an important avenue for determining the possible association of detected viruses with their hosts.

The metagenomic approach employed in this study allowed the characterization of several RNA viruses found among midges from five sites in Kenya. These identified viruses add to the currently existing biodiversity of midge-borne viruses. More specifically, sequencing of these field-collected midges allowed us to genetically characterize up to 15 RNA viruses. The majority of these are novel viruses with low similarity thresholds, with as low as 39.13% amino acid similarity to existing viruses. Only Isiolo_1, Turkana_5, and Turkana_10 showed higher similarity to existing viruses, with 97.7%, 85.62%, and 84.27% amino acid similarities, respectively. It should, however, be noted that in this study, we did not carry out the taxonomic classification of the identified viruses to the species level. The International Committee on Taxonomy of Viruses (ICTV) establishes various criteria for the classification of virus species, which often differ depending on virus group, and these criteria include information other than genetic sequences of the virus (35). Further studies should, therefore, be carried out in order to classify the identified viruses to their respective species.

The majority of the viruses identified in this study are ISVs, and none of them is known to be pathogenic to vertebrate hosts. However, even though they are unlikely to be associated with diseases, studies have shown that ISVs can potentially influence the vector competence of arthropod vectors by interfering with their vectorial capacity for pathogenic viruses, possibly due to competitive inhibition (36–38). These viruses, therefore, can act as important biocontrol agents in the transmission of pathogenic viruses. Additionally, identification of these viruses may fill important gaps in the phylogeny of viruses and provide important information to studies aimed at understanding the origin and evolution of pathogenic viruses (39, 40).

Several viruses identified in this study belong to families known to be associated with different hosts other than arthropods. These include *Tombusviridae*, *Solemoviridae*, *Partitiviridae*, *Picornaviridae*, and *Totiviridae*. The families *Tombusviridae* and *Solemoviridae*, for instance, contain viruses that have plants as their natural hosts. The family *Picornaviridae*, on the other hand, contains viruses known to infect only the vertebrates. *Partitiviridae* and *Totiviridae* contain viruses whose natural hosts are known to be quite diverse. Fungi and plants are known natural hosts of *Partitiviridae*, while fungal and protozoan parasites are known to be the natural hosts of viruses in the family *Totiviridae*. Our findings, therefore, suggest that these viruses could also be associated with midges. However, it is also possible that this observation may have been a result of our sample processing strategy. The sample processing in this study involved homogenization of the entire invertebrate specimens. Therefore, some of the viruses detected may have originated from undigested food, gut microflora, or even parasites that may have been present in the invertebrates at the time of processing. This may be true for members of the family *Ceratopogonidae*, whose feeding sources are known to be quite diverse (41–43). It would not, therefore, be surprising if some of the viruses detected have their origin in other hosts other than midges.

Insect community composition identified in the current study is quite diverse. Our metabarcoding approach was able to identify numerous species of midges in each site. The fractions of the reads specific for each of the species identified in this study were quite variable in each of the sites in the study (Table 2). Considering the sensitivity issues associated with HTS as well as primer biases, these observations cannot reliably be used as estimates of the relative abundance of each of the identified species (44, 45). Nonetheless, the identification of a given species in a pool of midges is reason enough for it to be considered a possible host of the viruses identified in the study. This is because the identified virus may have originated from either of the species of midges in the pool, irrespective of their abundance. These findings, therefore, provide us with a unique opportunity to infer the possible hosts of the detected viruses, using methods such as co-occurrence networks (46). Such an approach would, however, require sequencing of multiple pools from a given locality to use in the network and also to help in improving the accuracy of this method. Additionally, approaches to reducing PCR bias can be considered in order to improve the accuracy of host-virus association methods. Some of the approaches to reducing PCR bias include *in silico* testing of primers before use, the use of multiple sets of primers, and the use of PCR-free shotgun sequencing pipelines (44, 47, 48). Future studies should, therefore, consider some of these requirements in order to definitively associate the identified viruses with their insect hosts.

Metagenomics and metabarcoding methods used in this study are potentially cost-effective approaches to arbovirus and insect surveillance. These approaches have the advantage of analyzing hundreds to thousands of insect specimens in a single pool. Traditional methods, on the other hand, would process between 25 and 50 specimens per pool. Further, the classical DNA barcoding method for species identification and confirmation often processes a single specimen at any given time. Thus, the costs and labor are dramatically reduced when metagenomics coupled with metabarcoding is used in surveillance. Deployment of these two methods is, however, a long way ahead. This is due to the drawbacks associated with these methods, such as nonsensitivity of the metagenomics method compared to quantitative PCR (49, 50). Nonetheless, use of various enrichment methods as well as methods to deplete the host organism would generally improve the sensitivity of this method. Another existent limitation of large-scale metagenomics is the difficulty in associating the detected virus with its host. Metabarcoding, as applied in this study, helps to narrow down the possible hosts of these viruses. However, further improvement of this method is required, as highlighted above, so as to overcome the challenge of associating the individual virus detected to one of the possible insect species. Therefore, future studies will benefit from using viral enrichment methods, as well as methods like the co-occurrence networks, to infer the insect hosts associated with the identified viruses.

MATERIALS AND METHODS

Collection and sorting of midge specimens. The specimens used in this study were collected in 2016 from some of Kenya's arid and semiarid lands (ASAL), including Baringo, Kacheliba, Turkana, and Isiolo. These are livestock-rearing areas, and the climatic conditions allow for a high density of midges (51). Specimens were collected using CDC light traps (John W. Hock) that were set in the evening (1700 h) and then collected the following morning (0600 h). The traps were placed near resting places for livestock, and they were baited with dry ice held in Igloo containers (52). All the trapped specimens were transported to the site laboratory, where they were sorted and cryopreserved for transportation to the laboratory. The cryopreserved specimens were transported to the laboratory at Kenya Medical Research Institute (KEMRI), where they were identified and pooled into a maximum of 50 specimens per pool based on whether the specimen was a *Culicoides* sp. or unclassified midge, area/site where they were collected, sex, and blood feeding status. The pooled specimens were stored at −80°C until processing.

Bulk pool preparation. Individual pools of ≤50 nonfed specimens were first homogenized with Copperhead metal BBs (Crosman, USA) using homogenization medium containing minimum essential medium, with Earle's salts and reduced $NaHCO_3$. The medium was supplemented with 15% heat-inactivated fetal bovine serum (FBS) (Gibco) and 2% each of L-glutamine (Sigma-Aldrich) and antibiotic/antimycotic solution (Sigma-Aldrich). The homogenates were clarified by centrifugation at 10,000 × *g* for 10 min. To prepare samples for metagenomics and metabarcoding analysis, bulk pools were created for each of the 5 regions; Turkana, Baringo, Budalangi, Isiolo, and Kacheliba. The total number of specimens processed for each site included 1,063 from Turkana, 892 from Baringo, and 640, 600, and 156 from Isiolo, Budalangi, and Kacheliba, respectively (Table 2). For metagenomics, all the individually clarified supernatants from the different pools were combined to create five bulk pools for each of the regions. These 5 bulk pools were mixed by vortexing then used for viral RNA extraction. For metabarcoding, the individual pellets from each of the pool of ≤50 specimens were combined for each of the regions to make 5 bulk pools containing the combined crude mixture (Table 2). Homogenization medium, described above, was then added to these bulk pools, and further crushing was performed to ensure adequate mixing of the combined homogenates. The crude mixture was used for DNA extraction.

Metagenomics. (i) Viral RNA extraction. The clarified supernatant of the 5 individual bulk pools was passed through 0.22-μm filters to remove excess host "contaminants" and any bacteria while concentrating the viral particles. In preparing the 5 samples, one extraction blank and two positive samples were included as controls. The two positive-control samples were dengue virus type 2 isolates that had been amplified in Vero cells. RNA extraction was performed using a QIAamp viral RNA minikit (Qiagen, Germany) following the manufacturer's recommended protocol. RNA was quantified using a Nanodrop 2000 spectrophotometer (Thermo Fisher Scientific, USA) and Qubit RNA 2.0 fluorometer using the Qubit RNA HS assay kit (Invitrogen, USA). The RNA was then prepared for Illumina library preparation.

(ii) Illumina library preparation. Libraries for sequencing were prepared using TruSeq stranded mRNA kit (Illumina, USA), following the manufacturer's recommended protocol with modification to exclude the poly(A)-containing mRNA purification steps. Briefly, reverse transcription on ~25 ng/μl of RNA was achieved by using Superscript III reverse transcriptase (Invitrogen, USA) and random hexanucleotide primers (Invitrogen, USA). This was followed by second-strand synthesis using DNA polymerase I and RNase H, provided with the library preparation kit. Purification was then performed using AMPure XP beads (Beckman Coulter, USA) after which the purified double-strand cDNA fragments were end repaired by adding a single A nucleotide to the 3′ end of the blunt fragments, to prevent the formation of chimeras and improve adapter ligation efficiency. Ligation of the adapters was performed, and the products were purified and enriched by PCR to create the final library. Libraries were normalized and pooled before loading. Sequencing was carried out using the MiSeq reagent kit V3 (Illumina, USA), in a 600-cycle sequencing format.

(iii) Sequence analysis. Raw sequence reads were initially subjected to cleaning using Trim Galore v0.6.5 to remove adapters and Prinseq Lite v0.20.4 to remove low-quality reads using the following parameters: minimum length, 50 bp; maximum length, 301 bp; and minimum mean Q score, 30. Further, filtering of the reads was performed by using riboPicker v0.4.3, to remove rRNA sequences by comparing them against the SILVA rRNA database, release 138.1 (53). Paired-end reads were merged using PEAR 0.9.8 (54), and preliminary analysis was performed using the MG-RAST server to classify reads taxonomically. Cleaned reads were assembled *de novo* using the Trinity program (55) with default parameters. The cleaned reads were mapped back to the assembled contigs and filtered to retain only contigs in which at least 90% of bases had 5× coverage (56). Contigs that met this criterion were first compared to the NCBI viral database using the BLASTx program. Potential viral contigs were further compared to the entire NCBI nr database using the BLASTx program to filter out all nonviral sequences. Finally, as a control step to test false positives that might have occurred due to index hopping and carryover contamination, sequence reads belonging to the positive controls and the negative control were mapped against the viral contigs obtained using the Burrows-Wheeler Aligner (BWA-MEM) v 0.7.17. No contaminant contigs were identified during this step. Sequences that were confirmed to be of viral origin were translated, and ORF predictions were performed on them using the Expasy server (57). Phylogenetic reconstructions were performed based on the RNA-dependent RNA polymerase (RdRp) gene. To ensure meaningful depiction of the evolutionary relationships of the newly discovered viruses, their closest RdRp homologs were downloaded from GenBank and used as reference sequences in the phylogenetic analysis. Maximum-likelihood phylogenies were inferred using iqtree (58), with simultaneous evaluation of the best model and tree searching being performed based on 1,000 bootstrap estimates and 1,000 approximate-likelihood-ratio tests. The inferred phylogenies were visualized in Figtree v1.4.4.

Metabarcoding. (i) DNA extraction. DNA extraction was performed on the bulk pools of the crude homogenates using a QIAamp DNA extraction kit (Qiagen, Germany), according to the manufacturer's instructions. Two extraction blanks were included during the extraction process, and subsequently used during PCR and sequencing. The extracted DNA was quantified using Nanodrop 2000 (Thermo Fisher Scientific, USA). Amplification of the COI gene was then carried out on <1 μg of extracted DNA, using the universal pair of primers for metazoan invertebrates LCO1490/HCO2198 (59). These primers amplify an approximately 710-bp region of the COI gene of arthropod vectors. COI amplicons were generated from a 25-μl PCR containing 12.5 μl AmpliTaq Gold 360 master mix (Applied Biosystems, USA), 9.5 μl DNase/RNase-free water, and 0.5 μl each of the forward and reverse primers at 25 μM. The PCR cycling conditions were set as follows; initial denaturation at 95°C for 10 min, 35 cycles of 95°C for 30 s, 49°C for 30 s, 72°C for 30 s, and a final extension of 72°C for 7 min.

(ii) MinION library preparation. The COI amplicons were first purified using AMPure XP beads (Beckman Coulter, USA). The purified products were quantified using a Qubit dsDNA HS assay kit (Invitrogen, USA) with a Qubit fluorometer 2.0. Based on the concentration of the quantified products, the volume of PCR products that yielded 200 fmol was determined and used as starting material for MinION library preparation. Library preparation was carried out using a ligation sequencing kit (SQK-LSK109), following the manufacturer's protocol with the exclusion of the DNA fragmentation step. Briefly, 200 fmol of the purified products were end repaired using a NEBNext Ultra II end repair and dA-tailing module (New England Biolabs [NEB], UK). The end-repaired DNA for each sample was individually barcoded using Native Barcoding Expansion 1-12 (EXP-NBD104), which was achieved with the use of NEB Blunt/TA ligase master mix (NEB, UK). An equal amount from each of the 200-fmol barcoded libraries was combined into a single pool, which was then purified with AMPure XP beads (Beckman Coulter, USA). Adapter ligation of the purified library was done with NEBNext quick ligation module (NEB, UK) and the libraries were further purified using AMPure XP beads, with a final wash of the beads being carried out using short fragment buffer (SFB) provided with the SQK-LSK109 kit. The final library was loaded onto the flow cell (FLO-MIN106D) and sequenced using the workflows provided in the MinKNOW software.

(iii) Sequence analysis. Base-calling and demultiplexing were performed on the MinION Mk1C device using Guppy. Sequencing reads were quality filtered with Nanofilt v2.8.0 (60), in order to retain only the higher-quality reads with a read quality score of ≥10 as recommended by Nanopore (https://github.com/nanoporetech/ont_tutorial_basicqc). Reads that were longer or shorter than the expected length of approximately 710 bp (with a 150-bp buffer) were also filtered. In addition, sequences that were identical to those detected in the extraction blanks were removed. Error correction of the sequence reads was performed using isONclust v0.0.6 and isONcorrect v0.0.8 (61, 62), using default parameters with the *–ont* flag. The corrected reads were resampled to approximately 11,000 reads per sample, using rasusa v0.5.0 (63). Read clustering, consensus sequence generation, and determination of the number of reads supporting each consensus sequence were carried out using IsoCon v0.2.5.1 (64). IsoCon treats reverse complements and sequence duplicates of various lengths as different. Therefore, these were further removed by performing clustering of the consensus sequences using cd-hit-est (65) with a 98.9% similarity threshold, which is the lowest accuracy of error-corrected nanopore reads (62). Taxonomic assignation of the consensus sequences was performed in the MIDORI server using RDPClassifier with COI reference sequence database (66, 67). This was further validated by searching against the NCBI nr database, in order to determine the lowest classification of each of the consensus sequence. We excluded any OTUs that were not classified as belonging to an expected invertebrate phylum. To increase the reliability of the identified OTUs, singletons were removed and only OTUs that were supported by ≥10 sequences were retained. Further, invertebrate species supported by less than 1% of the total sequences in each site were also removed from the final analysis.

Ethical approval. Ethical approval to carry out this study was obtained from the Kenya Medical Research Institute's Scientific Ethics Review Unit (SERU), under protocol number KEMRI SSC 3693.

Data availability. The sequences of the viruses identified in the study have been submitted to GenBank under accession numbers MZ078285 to MZ078300. The consensus sequences of the cytochrome oxidase subunit 1 gene (COI) for the specimens in the study are also available in GenBank under the accession numbers MZ227639 to MZ227812 and OK357592 to OK357601.

ACKNOWLEDGMENTS

We thank John Gachoya, Dunstone Beti, and Reuben Lugalia for their technical assistance.

We declare no competing interest.

S.K.L. and R.S. conceived the study, S.K.L. performed data curation and formal analysis and wrote the manuscript, F.E., W.B., and M.I. assisted in the formal analysis and writing of the manuscript, J.L. coordinated specimen collection and identification, and S.O.O. performed the sequencing and formal analysis. All authors read and approved the final version of the manuscript.

REFERENCES

1. World Health Organization. 2020. Vector-borne diseases. https://www.who.int/news-room/fact-sheets/detail/vector-borne-diseases.
2. Mack A, Choffnes ER, Relman DA, Hamburg MA, Sparling PF, Lemon SM. 2008. Vector-borne diseases: understanding the environmental, human health, and ecological connections: workshop summary. National Academies Press, Washington, DC.
3. Heeney J. 2006. Zoonotic viral diseases and the frontier of early diagnosis, control and prevention. J Intern Med 260:399–408. https://doi.org/10.1111/j.1365-2796.2006.01711.x.
4. Woolhouse MEJ, Gowtage-Sequeria S. 2005. Host range and emerging and reemerging pathogens. Emerg Infect Dis 11:1842–1847. https://doi.org/10.3201/eid1112.050997.
5. Marchi S, Trombetta CM, Montomoli E. 2018. Emerging and re-emerging arboviral diseases as a global health problem. *In* Majumder MAA (ed.), Public health: emerging and re-emerging issues. InTechOpen. https://doi.org/10.5772/intechopen.77382.
6. Gould E, Pettersson J, Higgs S, Charrel R, de Lamballerie X. 2017. Emerging arboviruses: why today? One Health 4:1–13. https://doi.org/10.1016/j.onehlt.2017.06.001.
7. Sick F, Beer M, Kampen H, Wernike K. 2019. Culicoides biting midges—underestimated vectors for arboviruses of public health and veterinary importance. Viruses 11:376. https://doi.org/10.3390/v11040376.
8. Carpenter S, Groschup MH, Garros C, Felippe-Bauer ML, Purse BV. 2013. Culicoides biting midges, arboviruses and public health in Europe. Antiviral Res 100:102–113. https://doi.org/10.1016/j.antiviral.2013.07.020.
9. World Organisation for Animal Health. 2021. OIE-listed diseases, infections and infestations in force in 2021. https://www.oie.int/en/animal-health-in-the-world/oie-listed-diseases-2021/. Accessed 7 March 2021.
10. Mellor P, Boorman J, Baylis M. 2000. Culicoides biting midges: their role as arbovirus vectors. Annu Rev Entomol 45:307–340. https://doi.org/10.1146/annurev.ento.45.1.307.
11. Shelley A, Coscarón S. 2001. Simuliid blackflies (Diptera: Simuliidae) and ceratopogonid midges (Diptera: Ceratopogonidae) as vectors of Mansonella ozzardi (Nematoda: Onchocercidae) in northern Argentina. Mem Inst Oswaldo Cruz 96:451–458. https://doi.org/10.1590/s0074-02762001000400003.
12. Van den Berg H, Velayudhan R, Ejov M. 2013. Regional framework for surveillance and control of invasive mosquito vectors and re-emerging vector-borne diseases 2014-2020. World Health Organization, Geneva, Switzerland.
13. Fournet F, Jourdain F, Bonnet E, Degroote S, Ridde V. 2018. Effective surveillance systems for vector-borne diseases in urban settings and translation of the data into action: a scoping review. Infect Dis Poverty 7:99. https://doi.org/10.1186/s40249-018-0473-9.
14. Knope K, Whelan P, Smith D, Nicholson J, Moran R, Doggett S, Sly A, Hobby M, Kurucz N, Wright P. 2013. Arboviral diseases and malaria in Australia, 2010-11: annual report of the National Arbovirus and Malaria Advisory Committee. Commun Dis Intell Q Rep 37:E1–E20.
15. Zhang Y-Z, Shi M, Holmes EC. 2018. Using metagenomics to characterize an expanding virosphere. Cell 172:1168–1172. https://doi.org/10.1016/j.cell.2018.02.043.
16. Chen K, Pachter L. 2005. Bioinformatics for whole-genome shotgun sequencing of microbial communities. PLoS Comput Biol 1:e24–e112. https://doi.org/10.1371/journal.pcbi.0010024.
17. Hoffmann B, Scheuch M, Höper D, Jungblut R, Holsteg M, Schirrmeier H, Eschbaumer M, Goller KV, Wernike K, Fischer M, Breithaupt A, Mettenleiter TC, Beer M. 2012. Novel orthobunyavirus in cattle, Europe, 2011. Emerg Infect Dis 18:469–472. https://doi.org/10.3201/eid1803.111905.
18. Hernández-Andrade A, Moo-Millan J, Cigarroa-Toledo N, Ramos-Ligonio A, Herrera C, Bucheton B, Bart J-M, Jamonneau V, Bañuls A-L, Paupy C. 2019. Metabarcoding: a powerful yet still underestimated approach for the comprehensive study of vector-borne pathogen transmission cycles and their dynamics. *In* Claborn D (ed), Vector-borne diseases-recent developments in epidemiology and control. IntechOpen.
19. Lin X, Stur E, Ekrem T. 2015. Exploring genetic divergence in a species-rich insect genus using 2790 DNA barcodes. PLoS One 10:e0138993. https://doi.org/10.1371/journal.pone.0138993.
20. Hebert PDN, Cywinska A, Ball SL, deWaard JR. 2003. Biological identifications through DNA barcodes. Proc Biol Sci 270:313–321. https://doi.org/10.1098/rspb.2002.2218.
21. Tavares ES, Baker AJ. 2008. Single mitochondrial gene barcodes reliably identify sister-species in diverse clades of birds. BMC Evol Biol 8:81–81. https://doi.org/10.1186/1471-2148-8-81.
22. Porter TM, Hajibabaei M. 2018. Over 2.5 million COI sequences in GenBank and growing. PLoS One 13:e0200177. https://doi.org/10.1371/journal.pone.0200177.
23. Wolf YI, Kazlauskas D, Iranzo J, Lucía-Sanz A, Kuhn JH, Krupovic M, Dolja VV, Koonin EV. 2018. Origins and evolution of the global RNA virome. mBio 9:e02329-18. https://doi.org/10.1128/mBio.02329-18.
24. Shi M, Lin X-D, Tian J-H, Chen L-J, Chen X, Li C-X, Qin X-C, Li J, Cao J-P, Eden J-S, Buchmann J, Wang W, Xu J, Holmes EC, Zhang Y-Z. 2016. Redefining the invertebrate RNA virosphere. Nature 540:539–543. https://doi.org/10.1038/nature20167.
25. Smith G, Macias-Muñoz A, Briscoe AD. 2014. Genome sequence of a novel iflavirus from mRNA sequencing of the butterfly Heliconius erato. Genome Announc 2:e00398-14. https://doi.org/10.1128/genomeA.00398-14.
26. Sadeghi M, Altan E, Deng X, Barker CM, Fang Y, Coffey LL, Delwart E. 2018. Virome of >12 thousand Culex mosquitoes from throughout California. Virology 523:74–88. https://doi.org/10.1016/j.virol.2018.07.029.
27. Ramírez AL, Colmant AMG, Warrilow D, Huang B, Pyke AT, McMahon JL, Meyer DB, Graham RMA, Jennison AV, Ritchie SA, van den Hurk AF. 2020. Metagenomic analysis of the virome of mosquito excreta. mSphere 5:e00587-20. https://doi.org/10.1128/mSphere.00587-20.
28. Pascall DJ, Tinsley MC, Obbard DJ, Wilfert L. 2019. Host evolutionary history predicts virus prevalence across bumblebee species. bioRxiv https://doi.org/10.1101/498717:498717.
29. Sõmera M, Fargette D, Hébrard E, Sarmiento C. 2021. Solemoviridae. https://talk.ictvonline.org/ictv-reports/ictv_online_report/positive-sense-rna-viruses/w/solemoviridae. Accessed 7 March 2021.
30. Lu G, Ye ZX, He YJ, Zhang Y, Wang X, Huang HJ, Zhuo JC, Sun ZT, Yan F, Chen JP, Zhang CX, Li JM. 2020. Discovery of two novel negeviruses in a

dungfly collected from the Arctic. Viruses 12:629. https://doi.org/10.3390/v12070692.

31. Walker PJ, Siddell SG, Lefkowitz EJ, Mushegian AR, Adriaenssens EM, Dempsey DM, Dutilh BE, Harrach B, Harrison RL, Hendrickson RC, Junglen S, Knowles NJ, Kropinski AM, Krupovic M, Kuhn JH, Nibert M, Orton RJ, Rubino L, Sabanadzovic S, Simmonds P, Smith DB, Varsani A, Zerbini FM, Davison AJ. 2020. Changes to virus taxonomy and the statutes ratified by the International Committee on Taxonomy of Viruses (2020). Arch Virol 165:2737–2748. https://doi.org/10.1007/s00705-020-04752-x.
32. Vainio EJ, Chiba S, Ghabrial SA, Maiss E, Roossinck M, Sabanadzovic S, Suzuki N, Xie J, Nibert M. 2018. ICTV virus taxonomy profile: Partitiviridae. J Gen Virol 99:17–18. https://doi.org/10.1099/jgv.0.000985.
33. Vasilakis N, Forrester NL, Palacios G, Nasar F, Savji N, Rossi SL, Guzman H, Wood TG, Popov V, Gorchakov R, González AV, Haddow AD, Watts DM, da Rosa APAT, Weaver SC, Lipkin WI, Tesh RB. 2013. Negevirus: a proposed new taxon of insect-specific viruses with wide geographic distribution. J Virol 87:2475–2488. https://doi.org/10.1128/JVI.00776-12.
34. Besansky NJ, Severson DW, Ferdig MT. 2003. DNA barcoding of parasites and invertebrate disease vectors: what you don't know can hurt you. Trends Parasitol 19:545–546. https://doi.org/10.1016/j.pt.2003.09.015.
35. Fauquet CM, Fargette D. 2005. International Committee on Taxonomy of Viruses and the 3,142 unassigned species. Virol J 2:64. https://doi.org/10.1186/1743-422X-2-64.
36. Vasilakis N, Tesh RB. 2015. Insect-specific viruses and their potential impact on arbovirus transmission. Curr Opin Virol 15:69–74. https://doi.org/10.1016/j.coviro.2015.08.007.
37. Hall RA, Bielefeldt-Ohmann H, McLean BJ, O'Brien CA, Colmant AM, Piyasena TB, Harrison JJ, Newton ND, Barnard RT, Prow NA. 2016. Commensal viruses of mosquitoes: host restriction, transmission, and interaction with arboviral pathogens. Evol Bioinformatics 12:35–44. https://doi.org/10.4137/EBO.S40740.
38. Öhlund P, Lundén H, Blomström A-L. 2019. Insect-specific virus evolution and potential effects on vector competence. Virus Genes 55:127–137. https://doi.org/10.1007/s11262-018-01629-9.
39. Parvez MK, Parveen S. 2017. Evolution and emergence of pathogenic viruses: past, present, and future. Intervirology 60:1–7. https://doi.org/10.1159/000478729.
40. Forterre P. 2006. The origin of viruses and their possible roles in major evolutionary transitions. Virus Res 117:5–16. https://doi.org/10.1016/j.virusres.2006.01.010.
41. Ma Y, Xu J, Yang Z, Wang X, Lin Z, Zhao W, Wang Y, Li X, Shi H. 2013. A video clip of the biting midge Culicoides anophelis ingesting blood from an engorged Anopheles mosquito in Hainan, China. Parasit Vectors 6:326. https://doi.org/10.1186/1756-3305-6-326.
42. Slama D, Haouas N, Mezhoud H, Babba H, Chaker E. 2015. Blood meal analysis of Culicoides (Diptera: Ceratopogonidae) in Central Tunisia. PLoS One 10:e0120528. https://doi.org/10.1371/journal.pone.0120528.
43. Tomazatos A, Jöst H, Schulze J, Spînu M, Schmidt-Chanasit J, Cadar D, Lühken R. 2020. Blood-meal analysis of Culicoides (Diptera: Ceratopogonidae) reveals a broad host range and new species records for Romania. Parasit Vectors 13:79. https://doi.org/10.1186/s13071-020-3938-1.
44. Clarke LJ, Soubrier J, Weyrich LS, Cooper A. 2014. Environmental metabarcodes for insects: in silico PCR reveals potential for taxonomic bias. Mol Ecol Resour 14:1160–1170. https://doi.org/10.1111/1755-0998.12265.
45. Elbrecht V, Leese F. 2015. Can DNA-based ecosystem assessments quantify species abundance? Testing primer bias and biomass—sequence relationships with an innovative metabarcoding protocol. PLoS One 10:e0130324. https://doi.org/10.1371/journal.pone.0130324.
46. Starr EP, Nuccio EE, Pett-Ridge J, Banfield JF, Firestone MK. 2019. Metatranscriptomic reconstruction reveals RNA viruses with the potential to shape carbon cycling in soil. Proc Natl Acad Sci U S A 116:25900–25908. https://doi.org/10.1073/pnas.1908291116.
47. Gibson J, Shokralla S, Porter TM, King I, van Konynenburg S, Janzen DH, Hallwachs W, Hajibabaei M. 2014. Simultaneous assessment of the macrobiome and microbiome in a bulk sample of tropical arthropods through DNA metasystematics. Proc Natl Acad Sci U S A 111:8007–8012. https://doi.org/10.1073/pnas.1406468111.
48. Zhou X, Adamowicz SJ, Jacobus LM, DeWalt RE, Hebert PD. 2009. Towards a comprehensive barcode library for arctic life-Ephemeroptera, Plecoptera, and Trichoptera of Churchill, Manitoba, Canada. Front Zool 6:30–39. https://doi.org/10.1186/1742-9994-6-30.
49. Wylie KM, Mihindukulasuriya KA, Sodergren E, Weinstock GM, Storch GA. 2012. Sequence analysis of the human virome in febrile and afebrile children. PLoS One 7:e27735. https://doi.org/10.1371/journal.pone.0027735.
50. Batovska J, Mee PT, Lynch SE, Sawbridge TI, Rodoni BC. 2019. Sensitivity and specificity of metatranscriptomics as an arbovirus surveillance tool. Sci Rep 9:19398. https://doi.org/10.1038/s41598-019-55741-3.
51. Behnke R, Muthami D. 2011. The contribution of livestock to the Kenyan economy. https://cgspace.cgiar.org/bitstream/handle/10568/24972/IGAD_LPI_WP_03-11.pdf?se.
52. McNelly JR. 1989. The CDC trap as a special monitoring tool, p 26–33. *In* Proceedings of the Seventy-Sixth Annual Meeting of the New Jersey Mosquito Control Association, Inc. New Jersey Mosquito Control Association, Inc., Lindenwold, NJ.
53. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO. 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res 41:D590–D596. https://doi.org/10.1093/nar/gks1219.
54. Zhang J, Kobert K, Flouri T, Stamatakis A. 2014. PEAR: a fast and accurate Illumina paired-end reAd mergeR. Bioinformatics 30:614–620. https://doi.org/10.1093/bioinformatics/btt593.
55. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 29:644–652. https://doi.org/10.1038/nbt.1883.
56. Wajid B, Serpedin E. 2016. Do it yourself guide to genome assembly. Brief Funct Genomics 15:1–9. https://doi.org/10.1093/bfgp/elu042.
57. Artimo P, Jonnalagedda M, Arnold K, Baratin D, Csardi G, de Castro E, Duvaud S, Flegel V, Fortier A, Gasteiger E, Grosdidier A, Hernandez C, Ioannidis V, Kuznetsov D, Liechti R, Moretti S, Mostaguir K, Redaschi N, Rossier G, Xenarios I, Stockinger H. 2012. ExPASy: SIB bioinformatics resource portal. Nucleic Acids Res 40:W597–W603. https://doi.org/10.1093/nar/gks400.
58. Nguyen L-T, Schmidt HA, Von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol 32:268–274. https://doi.org/10.1093/molbev/msu300.
59. Folmer O, Black M, Hoeh W, Lutz R, Vrijenhoek R. 1994. DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. Mol Mar Biol Biotechnol 3:294–299.
60. De Coster W, D'Hert S, Schultz DT, Cruts M, Van Broeckhoven C. 2018. NanoPack: visualizing and processing long-read sequencing data. Bioinformatics 34:2666–2669. https://doi.org/10.1093/bioinformatics/bty149.
61. Sahlin K, Medvedev P. 2020. De novo clustering of long-read transcriptome data using a greedy, quality value-based algorithm. J Comput Biol 27:472–484. https://doi.org/10.1089/cmb.2019.0299.
62. Sahlin K, Medvedev P. 2021. Error correction enables use of Oxford Nanopore technology for reference-free transcriptome analysis. Nat Commun 12:2. https://doi.org/10.1038/s41467-020-20340-8.
63. Hall M. 2019. Rasusa: randomly subsample sequencing reads to a specified coverage. https://doi.org/10.5281/zenodo.3731394.
64. Sahlin K, Tomaszkiewicz M, Makova KD, Medvedev P. 2018. Deciphering highly similar multigene family transcripts from Iso-Seq data with IsoCon. Nat Commun 9:4601. https://doi.org/10.1038/s41467-018-06910-x.
65. Fu L, Niu B, Zhu Z, Wu S, Li W. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28:3150–3152. https://doi.org/10.1093/bioinformatics/bts565.
66. Leray M, Ho S-L, Lin I-J, Machida RJ. 2018. MIDORI server: a webserver for taxonomic assignment of unknown metazoan mitochondrial-encoded sequences using a curated database. Bioinformatics 34:3753–3754. https://doi.org/10.1093/bioinformatics/bty454.
67. Wang Q, Garrity GM, Tiedje JM, Cole JR. 2007. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl Environ Microbiol 73:5261–5267. https://doi.org/10.1128/AEM.00062-07.