GENETIC CHARACTERIZATION OF ARBOVIRUSES AND ASSOCIATED BITING MIDGES AND RELATED VECTORS IN SELECTED COUNTIES IN KENYA

LANGAT KIPNGETICH SOLOMON

MASTER OF SCIENCE

(Molecular Biology and Bioinformatics)

JOMO KENYATTA UNIVERSITY

OF

AGRICULTURE AND TECHNOLOGY

2022

Genetic Characterization of Arboviruses and Associated Biting Midges and Related Vectors in Selected Counties in Kenya

Langat Kipngetich Solomon

A Thesis Submitted in Partial Fulfilment of the Requirements for the Degree of Master of Science in Molecular Biology and Bioinformatics of the Jomo Kenyatta University of Agriculture and Technology

2022

DECLARATION

This thesis is my original work and has not been presented for a degree in any other university.

Signature …… Date …….

Langat Kipngetich Solomon

This thesis has been submitted for examination with our approval as supervisors.

Signature …… Date …….

Dr. Fredrick Eyase, PhD

JKUAT, Kenya

Signature …… Date …….

Prof. Rosemary Sang, PhD

KEMRI, Kenya

Signature …… Date …….

Prof. Mabel O. Imbuga, PhD

JKUAT, Kenya

Signature …… Date …….

Prof. Wallace D. Bulimo, PhD

KEMRI, Kenya

DEDICATION

This work is dedicated to my parents, Mr. William Kiplangat Milgo and Mrs. Lucia Milgo for their unconditional love, immense support and for instilling in me a sense of discipline and hard work, and to my siblings for their enormous support.

ACKNOWLEDGEMENTS

I am greatly indebted to my supervisors Dr. Fredrick Eyase, Prof. Rosemary Sang, Prof. Mabel Imbuga and Prof Wallace Bulimo for the immense efforts they put in giving me invaluable advice, guidance and their supervision throughout the period of my study. I want to specifically thank Dr. Fredrick Eyase and Prof. Rosemary Sang for entrusting me to work on this project, and for their selfless acts and patience they showed me while carrying out this study.

I wish to express my sincere and warm appreciation to all the KEMRI VHF Lab personnel as well as my colleagues for their support during the course of my project. Additionally, I want to sincerely thank Jomo Kenyatta University of Agriculture and Technology (JKUAT) for granting me the opportunity to pursue my studies, and Kenya Medical Research Institute (KEMRI) for allowing me to carry out this study.

Finally, I am grateful to the almighty God for His grace and the strength He gave me to be able to complete this work successfully and to be able to go through this process of learning.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

LIST OF ACRONYMS

DNA	Deoxyribonucleic Acid
RNA	Ribonucleic Acid
cDNA	Complementary DNA
dNTP	Deoxynucleotide
DEPC	Diethylpyrocarbonate
PCR	Polymerase Chain Reaction
NGS	Next Generation Sequencing
HTS	High Throughput Sequencing
AHS	African Horse Sickness
BT	Bluetongue Disease
EHD	Epizootic Haemorrhagic Disease
ORF	Open Reading Frame
BLAST	Basic Local Alignment Search Tool
NCBI	National Centre for Biotechnology Information
MEGA	Molecular Evolutionary Genetics Analysis
COI	Cytochrome *c* oxidase I gene
KEMRI	Kenya Medical Research Institute

ABSTRACT

Biting midges are small insects (1-3mm long) in the order diptera. Medical and veterinary importance of biting midges lies in the ability of certain members of these insects to transmit protozoa and filariae as well as act as vectors of viral diseases in humans and domestic animals. These vectors have been implicated in the transmission of African horse sickness virus, Bluetongue virus, Oropouche virus, Shmallenberg, Epizootic Haemorrhagic Disease and Akabane virus among other pathogenic viruses, which have been associated with diseases in human and livestock animals. The aim of the study was to identify and genetically characterize arboviruses among biting midges and related vectors collected from various sites in Kenya including; Turkana, Baringo, Kacheliba, Budalangi and Isiolo. Further, the study sought to characterize the biting midge vectors so as to determine the species of vectors associated with the identified viruses. Insect specimens from the field were first sorted and pooled according to the collection site. The pooled specimens were homogenized and clarified by centrifugation. The supernatant was used to extract the viral RNA while the pellet was used to extract the vector DNA. The extracted RNA was subjected to sequencing using the metagenomics approach while extracted DNA was used to carry out amplicon-based high-throughput sequencing. The sequence data were initially cleaned and assembled for subsequent analysis. The sequences obtained were taxonomically classified through BLAST analysis. Subsequently, phylogenetic analysis was conducted to show the relationship of the identified viruses to existing ones in public databases. The study identified 15 viruses that were phylogenetically distinct. These viruses are classified into seven families, with one virus belonging to the newly proposed negevirus taxon. *Partitiviridae, Iflaviridae, Picornaviridae, Tombusviridae, Solemoviridae, Totiviridae,* and *Chuviridae* are among the seven virus families. In addition, a wide variety of midge species was discovered that could be linked to the viruses discovered. Midges within the *Ceratopogonidae* family were the most common insects identified. Others included those in the *Chironomidae* and *Cecidomydiae* families. The findings show a wide range of RNA viruses in Kenyan midges, including previously unknown viruses. Furthermore, metabarcoding analysis using COI barcodes reveals a high species richness among the insects characterized. The findings in this study highlights the presence of a number of viruses in midges from different sites across Kenya, some of which can be considered as novel viruses. Further, different species of biting midges were identified in this study that are possibly associated with these viruses identified. These findings are therefore important in mapping the potential areas with significant risk for disease occurrence based on the knowledge of species of insect vectors associated with particular viral pathogens. Nonetheless, the study recommends further studies to identify any viral pathogens among the insect vectors that have been associated with disease in this region.

CHAPTER ONE

INTRODUCTION

1.1 Background

Biting midges are small flies that belong to the order diptera, suborder nematocera. Their body lengths rarely exceed three millimetres and their developmental cycle consists of the egg, larvae (4 instars), pupa, and adult (Mellor *et al.*, 2000). Biting midges contains a diverse number of species, four of which are known to feed on blood from vertebrates (Mellor *et al.*, 2000). These include *Austroconops*, *Culicoides*, *Forcipomyia*, and *Leptoconops*. *Culicoides* forms the majority of these biting midges, with more than 1400 species (Mellor *et al.*, 2000; Heeney, 2006). Given their blood feeding habits, these insects have been shown to be vectors of several disease-causing pathogens which have resulted in significant impacts on public and veterinary health (Carpenter *et al.*, 2013).

Biting midges are known to be vectors of disease-causing pathogens such as protozoa and filariae; mainly *Mansonella streptocerca*, *Mansonella ozzardi* and *Mansonella perstans*. These pathogens are quite prevalent in the Caribbean and Latin America as well as in West and Central Africa (Carpenter *et al.*, 2013). Apart from the transmission of filarial nematodes, biting midges are also known to be vectors of arboviruses (van Eeden *et al.*, 2012). Arboviruses refer to viruses that are transmitted by arthropod vectors, and they account for two thirds of all the emerging viruses (Heeney, 2006). Majority of arboviruses are RNA viruses, with only African Swine Fever Virus (ASFV), being the only known DNA virus transmitted by tick vectors (Weaver & Reisen, 2010). Given their devastating consequences, three of the RNA viruses transmitted by biting midges have been listed by World organization for animal health (OIE) as notifiable diseases. These are Bluetongue, African Horse Sickness and Epizootic haemorrhagic disease (Heeney, 2006; Temmam *et al.*, 2016).

The most important public health importance of biting midges currently is in the transmission of arboviruses that infect humans as well as domestic and wild animals (Maclachlan & Guthrie, 2010; Robin *et al.*, 2016). One of the arboviruses transmitted by biting midges is the African horse sickness virus (AHSV), an aetiological agent of

African horse sickness (AHS) (Robin *et al.*, 2016). AHS affects equids and it is one of the most important and lethal infectious diseases in the sub Saharan Africa (Robin *et al.*, 2016). AHS causes mortality rates in susceptible horses that are estimated to be more than 90%, thus leading to considerable economic losses (Zientara *et al.*, 2015). Bluetongue virus (BTV) is also another important virus transmitted by these vectors. It causes Bluetongue (BT) disease in ruminants and it is of particular concern in Sub Saharan Africa (Maclachlan & Guthrie, 2010). BTV has spread to various parts of the world and it has been implicated in outbreaks in Europe, such as the 2006-2007 outbreak in Belgium which resulted in the disruption of animal trade and death of animals estimated to be worth 180 million pounds (Jauniaux *et al.*, 2008). In the US, the country was estimated to be losing up to 125 million dollars yearly due to BT disease (Tabachnick, 1996).

Biting midges have also been implicated in the transmission of Oropouche virus (OROV), which is the only known human pathogen transmitted by these vectors (Glick, 1990; Vasconcelos *et al.*, 2009; Carpenter *et al.*, 2013). OROV is the aetiological agent of Oropouche fever, which is one of the important public health concerns in the tropical areas of central and South America (Vasconcelos *et al.*, 2011). Oropouche fever is a febrile illness associated with headache, anorexia, arthralgia and on rare occasions meningitis (Travassos da Rosa *et al.*, 2017). The incidences of Oropouche fever remains largely undetermined in a majority of the epidemics. However, the first outbreak was recorded in 1961 in Brazil, with approximately 11,000 people reported to have been infected (Travassos da Rosa *et al.*, 2017). Since then, dozens of other epidemics have been recorded over a span of 45 years, with total number of Oropouche fever cases being estimated to be half a million (Vasconcelos *et al.*, 2011; Travassos da Rosa *et al.*, 2017). The most recent epidemic in Brazil occurred in 2006, when it re-emerged in an area that had not experienced any Oropouche fever incidences in over 26 years (Vasconcelos *et al.*, 2009). Outside of Brazil, OROV has been detected in Peru, Trinidad and Tobago as well as in Panama. Recent outbreaks of Oropouche fever occurred in Peru in 2016, where up to 24% seropositive cases were reported (Romero-Alvarez & Escobar, 2017). Various other arboviruses transmitted by midges and which are the subject of

major outbreaks in the recent years include Shmallenberg, Epizootic Haemorrhagic Disease (EHDV) and Akabane viruses among others (Maclachlan & Guthrie, 2010; Toye *et al.*, 2013).

In Kenya, mostly play a role in the transmission of pathogens of veterinary importance. This has been shown by the detection of different pathogens transmitted by these vectors in the country. A number of strains of BTV as well as Ephemeral fever virus (EFV) have previously been isolated from biting midges in Kenya (Davies *et al.*, 1979). More recently, a study on BTV and EHDV in local breeds of cattle found that approximately 62% of the 455 tested calves were seropositive for BTV, EHDV or both (Toye *et al.*, 2013). These studies points to high levels of circulating BTV and EHDV among Kenyan cattle. Further, there have been recent reports of outbreaks of BT disease in various counties in Kenya, including Narok and Samburu (OIE, 2018; The Standard, 2018). Despite this fact, the circulation, characterisation and associated human and animal disease for these viruses remain largely unexplored. Here, next-generation sequencing technologies was utilized to simultaneously characterize the viruses and associated vector species among biting midges in Kenya.

1.2 Statement of the problem

A variety of arboviruses have been isolated from biting midges throughout the world. In Kenya, different strains of BTV, EFV and EHDV have been detected. Further, studies on BTV and EHDV among the local breeds of cattle in Kenya found up to 62% of the calves to have these diseases. The findings point to high levels of circulating BTV and EHDV in Kenya. There have also been recent reports of outbreak of BT disease in various counties in Kenya, including Narok and Samburu. With increasing international and local travel, globalisation and changes in climatic conditions, there is always a continued risk of incursions of new viruses. This is particularly true with biting midges, whose long-distance movements have been extensively modelled. There is, therefore, need to identify any of the currently circulating arboviruses among midges in Kenya as well as determine the associated biting-midge species.

1.3 Justification

A number of viruses have previously been isolated from biting midges in Kenya. These include BTV, EHDV, and EFV among others. Additionally, several biting midges vectors such as *Culicoides* spp. have also been identified in the country. These vectors have previously been implicated in the transmission of viruses of veterinary and public health concern. There is, however, limited information regarding arboviruses currently circulating among Kenyan midges. Detection and characterization of these viruses would provide an insight into the circulating arboviruses that could be of veterinary or public health concern. Viruses transmitted by biting midges are associated with diseases that often lead to significant economic losses due to the loss of livestock. Livestock such as cattle, sheep, goats and equids contributes 10% of Kenya's gross domestic product (GDP) and they employ up to 90% of the local population from Kenya's Arid and Semi-arid Lands (ASAL). Thus, the importance of livestock in Kenya is quite high and the findings from this study would be helpful in boosting this sector by providing the basis for surveillance and control measures aimed at averting any possible effects of arboviruses transmitted by biting midges. Additionally, the findings would be useful in allowing evidence-based refinement of guidelines on surveillance and control activities.

1.4 Research questions

1. What viruses are found in biting midges from Turkana, Baringo, Kacheliba, Isiolo and Budalangi sites in Kenya?
2. What are the phylogenetic relationship of identified viruses to existing viruses?
3. What species of biting midges are found in Turkana, Baringo, Kacheliba, Isiolo and Budalangi sites in Kenya?

1.5 Null hypothesis

1. There are no viruses in biting midges from Turkana, Baringo, Kacheliba, Isiolo, and Budalangi sites in Kenya.
2. There is no genetic diversity in arboviruses found in biting midges from Turkana, Baringo, Kacheliba, Isiolo, and Budalangi sites in Kenya.

3. There is no diversity in biting midge species from Turkana, Baringo, Kacheliba, Isiolo, and Budalangi sites in Kenya.

1.6 Objectives

1.6.1 General objective

To detect and genetically characterize arboviruses and associated biting midges and other related vectors in selected counties in Kenya

1.6.2 Specific objectives

1. To detect and identify viruses in biting midges and related insect vectors from Turkana, Baringo, Kacheliba, Isiolo, and Budalangi sites in Kenya
2. To characterise and determine the phylogenetic relationship of the identified viruses
3. To characterise biting midges and related insect vectors from selected sites in Kenya.

CHAPTER TWO

LITERATURE REVIEW

2.1 General characteristics of biting midges

Biting midges are small flies and, like mosquitoes and sandflies, they belong to the order diptera. They have been nicknamed no-see-ums, which is a descriptive name for their extremely irritating bite (Connelly, 2013; Mullen & Murphree, 2019). These flies are quite small with a size of approximately 1-3mm in length. Taxonomic identification of biting midges is dependent upon the morphological characteristics that are unique to each species including the wing patterns and venation, presence or absence of the spermatheca, intraocular space, sensillae presence and placement of flagellar segments (Connelly, 2013). Application of molecular differentiation methods have recently been applied in the differentiation of these vectors. These methods majorly includes use of a variety of markers that include cytochrome c oxidase subunit 1 gene (COI) and the ribosomal markers such as internal transcribed spacer 1 and 2 (ITS1 and ITS2) (Norris, 2002).

Most biting midge species are mostly active just before dusk and before dawn (Foxi *et al.*, 2022). All biting midges feed on nectar and plant sap. Female midges, however, requires a blood meal which is important during the gonotrophic cycle process. Female biting midges therefore seek blood meal by attacking mammals that include humans, birds, reptiles as well as livestock (Paweska *et al.*, 2002). However, despite their blood-seeking habits, biting midges have been shown to be exophilic and exophagic. Their resting places is mainly in the bushes, though some species hide in cracks of tree trunks as well as the upper layers of sand (Connelly, 2013).

2.2 General life cycle of biting midges

Biting midges undergo a typical holometabolous life cycle that consist of egg, 4 larval instars, pupa and adult cycles (Foxi & Delrio, 2010). The eggs are small and slender with an allantoid shape. They measure approximately 350-500um in length and between 65 and 80 um in breadth (Campbell & Kettle, 1975). When they are hatched, the eggs have a white colour which turns to dark brown after a short while. Different biting midge species lay varying egg batches. *Culicoides brevitarsis*, for

instance, lays between 30 and 40 egg batches while *Culicoides circumscriptus* can lay up to 450 egg batches (Foxi & Delrio, 2010). The common African species *Culicoides imicola* lays between 53 and 69 egg batches. The eggs are often hatched within a few days where the conditions are favourable (Foxi *et al.*, 2022). However, the eggs can also enter into a diapause where it can remain for up to 7-8 months before they hatched (Campbell & Kettle, 1975).

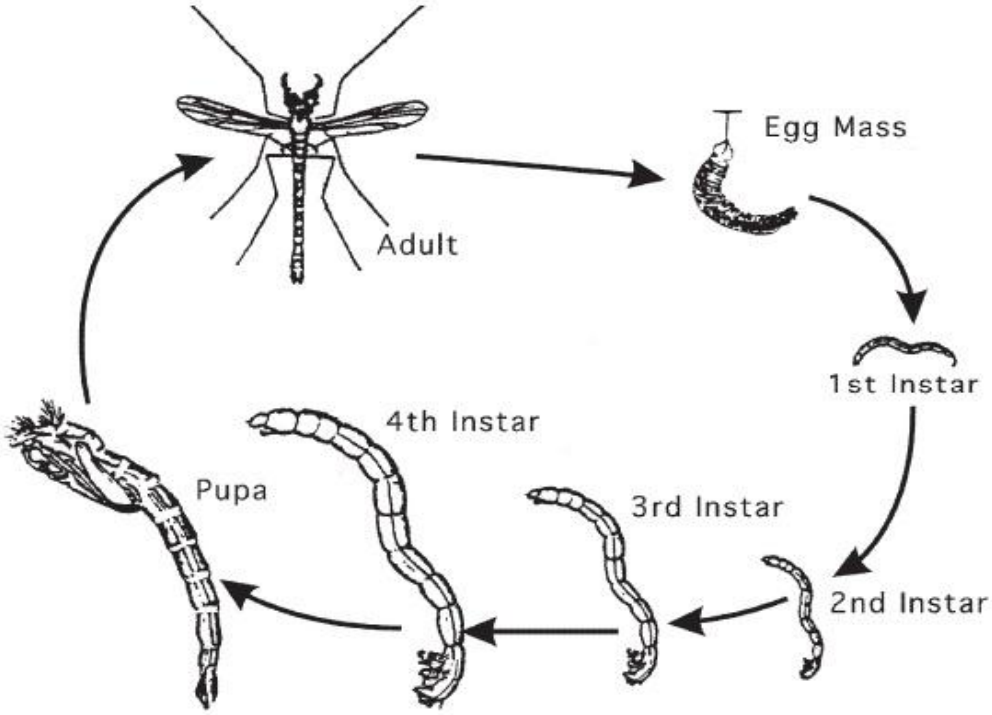


Figure 2. 1: General life cycle of biting midges, *Culicoides* spp. (Walker, 1987).The larval stages of biting midges are much longer than the egg and pupal stages. Larvae is a typical nematoceran with a sclerotized head, segmented body with 11 bodily segments and no appendages (Mellor *et al.*, 2000). Development of these larvae is temperature dependent, and it takes between 11 to 16 days at a temperature of approximately 28^{o}C (Foxi *et al.*, 2022). With a temperature of 25^{o}C the larval development takes between 15 and 21 days and with a temperature of 20^{o}C it takes between 34 and 56 days (Foxi *et al.*, 2022). The life stages of different species of biting midges, under field conditions, vary widely from as short as two weeks to as

long as one year for some species especially those in the arctic (Szadziewski *et al.*, 1997).

Pupation is made possible by the gradual drying of the larval habitats. Pupa is the stage that gives rise to the winged adult (Connelly, 2013). It is a none-feeding short-lived stage. Most biting midge species float in water, while others such as the *C. imicola* are unable and often drawn in water-logged soil (Campbell & Kettle, 1975). During emergence, male biting midges often emerge before the females. This is the same observation among mosquitoes and it leads to sizable population of males that are ready for mating once the females emerge. Mating normally occurs during swarming. This occurs during the dusk and it involves mainly the males that varies between approximately 10 and 1000 individuals. Swarm size, however, involves around 50 individuals (Connelly, 2013).

The life-span of adult biting midges varies depending on the prevailing conditions (Szadziewski *et al.*, 1997). Most species live for up to 20 days, but occasionally can survive for up to 90 days. Female midges require one blood meal for a single batch of eggs matured and thus blood feeding frequency is dependent upon the egg development rate(Kasičová *et al.*, 2021). Biting midge species and the prevailing temperature also play a role in the frequency of egg development and thus blood feeding rates (Kasičová *et al.*, 2021). Adult biting midges can fly for short distance from their larval habitats. However, they can be carried by wind to distances of up to 700 km (Szadziewski *et al.*, 1997).

2.3 Larval habitats of biting midges

The most common larval habitats of most biting midges contain moisture as well as organic matter. However, some species have more specialized larval habitats. Surface water and soil surface form a larval habitat for some *Culicoides* species (Foxi & Delrio, 2010; González *et al.*, 2013). A number of species use this habitat to lay their eggs. The type of soil that are associated with these vectors may vary from coarse sand to the finest clay. Presence of decomposed plant material in soil forms favourable conditions for breeding by biting midges (González *et al.*, 2013). Some species of biting midges have been found in the rotting stems of banana plants as

well as in the rotting fallen fruits. Dung pats of large animals also act as breeding habitats of biting midges. Studies have found biting midges breeding in bovine dung. Some species require fresh dung of specific animals to complete their life cycle (González *et al.*, 2013). *Culicoides bolitinos,* for instance, has been found to breed in the dung of African buffalo, cattle, as well as the blue wildebeest. Tree holes and other plant and rock cavities is also another larval habitat for biting midges. These can be the deep, dark water filled holes to the shallow and exposed hollows in tree with moist conditions (Foxi & Delrio, 2010).

2.4 Geographical distribution of biting midges

Biting midges has a worldwide distribution except for some few areas. Notable places of the world where biting midges are not found include New Zealand in Australasia, Hawaiian islands in the Americas as well as the most Southern tip of South America (Szadziewski *et al.*, 1997; Guichard *et al.*, 2014). The distribution of given species of biting midges, however, varies across the different parts of the world. Most abundant biting midge species in Africa is *Culicoides imicola* which is a vector of various disease causing pathogens that includes AHSV (Guichard *et al.*, 2014). In North America, the most common species is *Culicoides sonorensis*. *Cuicoides insignis* is the most common in South and Central America while *Culicoides wadai, Culicoides brevitarsis* and *Culicoides actoni* are abundant in Australia. *Culicoides fulvus* and *Culicoides schultzei* are common in Asia while *Culicoides imicola, Culicoides pulicaris* and *Culicoides obsoletus* are the most common in Europe (Guichard *et al.*, 2014). In terms of worldwide distribution, however, *Culicoides imicola* is one of the most widely distributed species of biting midges and it is found throughout Africa, Europe and along the Mediterranean region as well as South Asian countries such as Sri Lanka, Thailand, Laos and Vietnam (Figure 2.2) (Guichard *et al.*, 2014; Jacquet *et al.*, 2016).

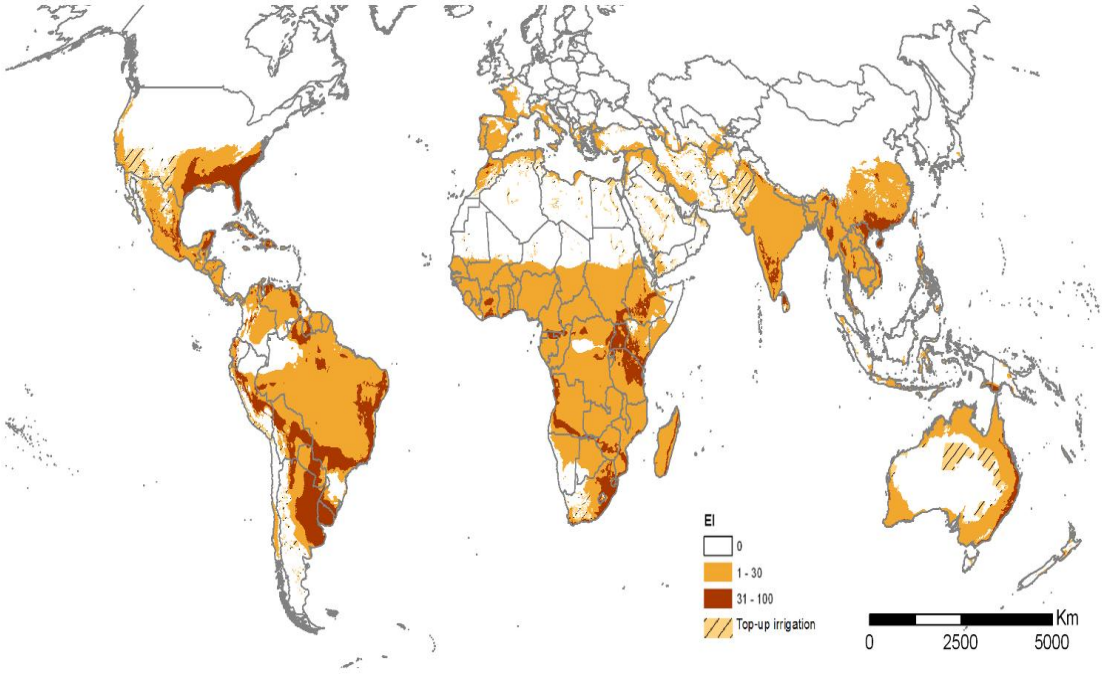
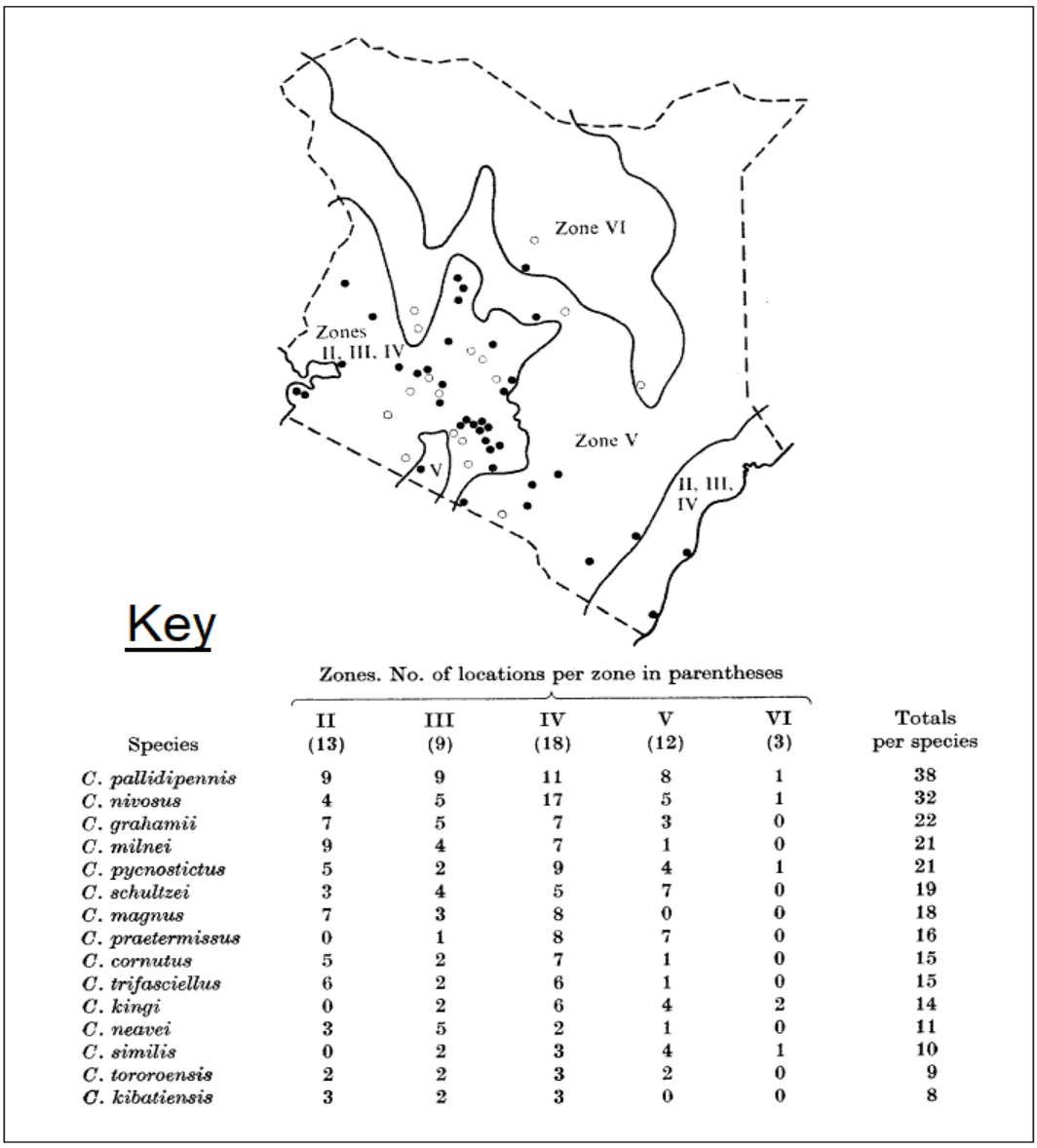


Figure 2. 2: Predicted potential distribution of *Culicoides imicola* based on global climate suitability under historical conditions (Guichard *et al.*, 2014).

There are currently limited number of studies on the distribution of biting midges in Kenya and the East African region at large. Nonetheless, earlier studies showed the presence of quite a number biting midge species across East Africa. *Culicoides grahamii*, *Culicoides pallidipennis* and *Culicoides Schultzei* were the most abundant species across the region (Khamala, 1971). Further, a study by Walker & Boreham (1976) identified different species of biting midges across different regions in Kenya, as shown below (Figure 2.3).



Zones. No. of locations per zone in parentheses

Species	II (13)	III (9)	IV (18)	V (12)	VI (3)	Totals per species
C. pallidipennis	9	9	11	8	1	38
C. nivosus	4	5	17	5	1	32
C. grahamii	7	5	7	3	0	22
C. milnei	9	4	7	1	0	21
C. pycnostictus	5	2	9	4	1	21
C. schultzei	3	4	5	7	0	19
C. magnus	7	3	8	0	0	18
C. praetermissus	0	1	8	7	0	16
C. cornutus	5	2	7	1	0	15
C. trifasciellus	6	2	6	1	0	15
C. kingi	0	2	6	4	2	14
C. neavei	3	5	2	1	0	11
C. similis	0	2	3	4	1	10
C. tororoensis	2	2	3	2	0	9
C. kibatiensis	3	2	3	0	0	8

Figure 2. 3: Map of Kenya showing the distribution of biting midges across the different regions (Walker & Boreham, 1976).

In terms of seasonal presence and abundance, biting midges have been shown to be more common in areas where temperatures are seldom below 0^{o}C (Carpenter *et al.*, 2013). These vectors are more common in areas that receive summer rainfalls as well as winter rainfall. It has been shown that as the altitude increases, the winter temperatures drop and trap collections of biting midges often decreases. In the months following winter, biting midge collections increase gradually to peak during the favourable conditions in the summer (Guichard *et al.*, 2014). Even though

temperature plays an important role in the maintenance of biting midge populations, rainfall is equally an important aspect that plays a key role in maintaining the breading sites of biting midges. The number of biting midges increases during the summer months provided there is sufficient rainfall to maintain semi-aquatic habitats for the larval stages (Maclachlan & Guthrie, 2010; Carpenter *et al.*, 2013).

2.5 Blood feeding habits of biting midges

Females of nearly all the species of biting midges, just like mosquitoes, require a blood meal in order to develop their eggs (Slama *et al.*, 2015). Biting habits of these midges, however, is not similar across all the species. Some European *Culicoides* are known to be quite notorious for their man-biting habits (Carpenter *et al.*, 2013). This is not, however, the case with most of African *Culicoides* species which are known to prefer feeding on animal blood as compared to man (Paweska *et al.*, 2002). These midges are also active at night and are rarely seen flying around during daytime (Paweska *et al.*, 2002).

The importance of female biting midges lies in their ability to act as biological vectors of various disease agents of veterinary and medical importance. These include protozoa, viruses and filarial nematodes which affect humans, birds and other animals (Diniz *et al.*, 2006; Carpenter *et al.*, 2013). This ability is made possible by their habit to feed on diverse hosts that include mammals, reptiles, man, birds as well as blood-engorged mosquitoes (Ma *et al.*, 2013; Slama *et al.*, 2015). In some parts of the world, biting midges have a severe biting nuisance and can also cause acute allergic dermatitis (Diniz *et al.*, 2006).

In a study on *Culicoides obsoletus*, a potential Bluetongue Virus Vector, on the Canary Islands of Spain, Martinez-de la Puente and colleagues sought to ascertain the blood meal sources of this vector (Martínez-de la Puente *et al.*, 2012). The findings showed that *C. obsoletus* feeds on goats and sheep, as shown by the molecular identification of the blood sources. It is not clear, however, if at the time of sample collection the predominant vertebrates in the sample collection area were only goats and sheep or whether there were other vertebrates that were not fed on by this vector. However, the findings confirms the fact that female *C. obsoletus* feeds on

blood from these two animals thus putting them at risk of contracting Bluetongue virus and the novel Smallenberg virus; both of which are transmitted by *Culicoides* spp. (Carpenter *et al.*, 2013).

In Kenya, Walker and colleagues carried out a study on a number of species of *Culicoides* spp. and their blood feeding habits (Walker & Boreham, 1976). Their study attempted to find out the blood feeding sources of *Culicoides* spp. using a precipitin test. The study covered 16 different sites in Kenya and it was done over a period of five years. The findings from this study showed that *C. pallidipennis* and *C. schultzei* fed more on cattle and sheep. On the other hand, *C. cornutus, C. zuluensis* and *C. milnei* also had same feeding patterns but were uncommon (Walker & Boreham, 1976).

2.6 Virus isolation from biting midges

A number of viruses have been isolated from biting midges. These insects are vectors to more than 50 viruses that are classified into three major classes of viruses; *Bunyaviridae*, *Rhabdoviridae* and *Reoviridae* (Diniz *et al.*, 2006; Heeney, 2006). Out of these viruses, more than 44% have not been isolated from other arthropod groups. In a study that analysed the viral communities in Senegalese biting midges, several novel viruses were detected including novel *Thogotovirus* species and also novel *Rhabdovirus* genus (Temmam *et al.*, 2016). Additionally, a number of known pathogenic viruses were detected including African horse sickness, epizootic haemorrhagic disease and bluetongue viruses (Temmam *et al.*, 2016).

In Zimbabwe, a number of viruses were isolated from *Culicoides* spp. in a veterinary research farm (Blackburn & Phelps, 1985). The study was carried out over a period of 11 months and the study ended in the isolation of Akabane virus, Nyabira virus, African horse sickness virus, bluetongue virus and Ephemeral fever virus (EFV) (Blackburn & Phelps, 1985). The findings showed differences in the kind of viruses isolated from individual species of *Culicoides* midges. Akabane virus, for instance, was isolated from *C. milnei* and *C. imicola* while Nyabira was only isolated from *C. imicola* (Blackburn & Phelps, 1985). Further, African horse sickness virus as well as

bluetongue virus were also isolated from *C. imicola*. Ephemeral fever virus, however, was only isolated from *C. coarctatus* (Blackburn & Phelps, 1985).

Arboviruses have also been isolated from biting midges in Kenya. Various viral strains were isolated from samples that were collected from different sites in Kenya (Davies *et al.*, 1979). In the study, however, the investigators did not differentiate the virus strains according to the vector species. Generally, viruses isolated include BTV and EFV (Davies *et al.*, 1979). There were also unidentified strains of viruses that were isolated; one of which was believed to be a new virus strain and the other five were believed to be members of the Palyam group. Viruses of the palyam group have been isolated from vertebrates including cattle, sheep, goats and humans and even though it was initially thought to cause abortion in cattle, this has not yet been scientifically established (Davies *et al.*, 1979; Whistler *et al.*, 1989).

2.7 Extrinsic incubation period of biting midges

Virus ingested by a vector remains for a certain period of time in the body before it can develop to transmission capability. The time period taken from the ingestion of a virus infected blood meal to transmission capability is referred to as extrinsic incubation period (Elbers *et al.*, 2015). It is during this period that the virus infects and replicates in the midgut epithelial cells of the biting midge. The virus then disseminates to infect other secondary organs and disperse to circulating haemolymph (Walker & Boreham, 1976; Elbers *et al.*, 2015). Once it reaches the salivary ducts, the virus can then be transmitted to vertebrates during a blood meal (Walker & Boreham, 1976). The extrinsic incubation period depends on the prevailing temperature, with higher temperature being known to shorten the incubation period among these vectors (Elbers *et al.*, 2015).

In one study on *Culicoides* spp. midges from Kenya, the investigators sought to determine whether the potential vectors live long enough to allow for virus incubation; from the time of virus ingestion to the opportunity to transmit the virus (Walker, 1977). In this study, it was found out that *Culicoides pallidipennis*, *C. cornutus*, and *C. schultzei* live for a period of approximately 10 - 15 days to take repeated blood meals, and this period was determined to be long enough to allow for

development of all biting-midge associated viruses to transmission capability (Walker, 1977).

2.8 Virus transmission and associated diseases

A high proportion of arboviruses, important to human and animal health, tend to circulate mostly in the tropical and subtropical regions where flying insects such as midges and mosquitoes are abundant (Liang *et al.*, 2015). Arboviruses have caused diseases for many years but it is only in the recent decades that a few arboviruses have increased in importance (Weaver & Reisen, 2010; Weaver, 2013). This is because of the increase in exposure to infection which is brought about by the expansion of human population and their activities. Some of the arboviruses have expanded their geographical range via direct spill-over from their enzootic cycles to humans (Weaver, 2013).

Approximately 96% of more than 1400 described species of biting midges are said to be obligatory blood feeders of mammals and birds (Lassen *et al.*, 2012). Less than 50 of these described species are thought to be involved in disease transmission and a few of them, less than 10, are proven vectors of arboviruses (Mellor *et al.*, 2000; Lassen *et al.*, 2012). More than 75 arboviruses have been isolated from *Culicoides* spp. across the world and the geographical distribution and seasonal incidences of these disease-causing arboviruses often depend on the distribution and the biology of the vector. Approximately 20 arboviruses belong to *Bunyaviridae*, 19 belong to *Reoviridae* and 11 belong to *Rabdoviridae* families (Mellor *et al.*, 2000).

The most well-known arboviruses include AHSV and BTV which are transmitted by *C. imicola*, and which cause devastating diseases in ruminants (Paweska *et al.*, 2002; Fall *et al.*, 2015). Other common arboviruses include Equine encephalosis virus (EEV) and Epizootic hemorrhagic disease virus (EHDV). EEV is an aetiological agent belonging to the genus Orbivirus and it causes Equine encephalosis, an infectious disease of Equidae and which is noncontagious (Paweska *et al.*, 2002). EHDV causes Epizootic haemorrhagic disease (EHD), a fatal hemorrhagic disease in the white-tailed deer of North America. More recently, several arboviruses of health and veterinary importance have been isolated in various ecosystems across the world.

Shmallenberg virus (SBV) was recently discovered in Germany among cattle that were presenting clinical signs that included diarrhoea, congenital deformities in calves and lambs as well as reduced milk yield (Hoffmann *et al.*, 2012; Carpenter *et al.*, 2013). Following the detection of this novel arbovirus, a number of studies were carried out across Europe and it is suspected that the species involved in the transmission of SBV include many of those that have been implicated previously in the transmission of BTV. Phylogenetic characterization classified SBV in close relationship to Akabane, Aino and Shampnda viruses (Carpenter *et al.*, 2013).

Currently, the most important public health role of biting midges lies in their ability to transmit Oropouche virus (OROV) (Lassen *et al.*, 2012; Carpenter *et al.*, 2013). OROV belongs to the genus *Orthobunyavirus* in the family *Bunyaviridae*. It is predominantly transmitted to humans by the biting midge *Culicoides paraensis*. OROV is the aetiological agent of the febrile illness Oropouche fever. This disease is commonly associated with headache in high number of cases, but can also lead to arthralgia, anorexia and meningitis in rare cases. OROV is distributed across a geographic range that includes Brazil, Panama, Peru, Trinidad and Colombia (Carpenter *et al.*, 2013).

2.9 Detection and characterization of arboviruses

The process of detection of novel viruses has traditionally relied upon the ability to culture and morphologically or serologically identify the viruses (Finkbeiner, 2009). Even though these methods are still in great use, they are highly limited. These methods are hampered by the inability of a large number of viruses to grow well or even at all in cell culture (Finkbeiner, 2009; Barzon *et al.*, 2011). Additionally, successful culturing of the virus does not necessarily lead to easy identification and characterization given the strategy depends on visual inspection of the morphology and particle size of the cultured viruses using electron microscopy (EM) or via serological cross reaction with antibodies to known viruses (Finkbeiner, 2009). Determination of the viral family to which a given virus belongs to, based on EM observations, was hampered by the fact that multiple virus families exhibit similar morphologies, thus making it hard to distinguish between them. Serological assays also had their own limitations given the unknown viruses must cross-react with

known viruses for which there are serological materials available (Finkbeiner, 2009). There is also a requirement for a priori knowledge of the possible identity of the viruses since the screening of the viruses will need to be done individually for each virus family (Finkbeiner, 2009).

The increased burden of virus-causing human and veterinary diseases, some new and previously unrecognized, necessitated the need for improved methods for the identification and characterization of new and unsuspected viral pathogens. Development and application of NGS technologies in the detection of disease-associated viruses has led to great success (Barzon *et al.*, 2011; Radford *et al.*, 2012; Aguiar *et al.*, 2015). Compared with other methods, NGS offers the benefit of high sensitivity and also the potential to detect a full spectrum of viruses, including novel viruses (Radford *et al.*, 2012). Vector-borne and zoonotic viruses are among the important and challenging areas of viral discovery. The feasibility of identifying arboviruses using NGS techniques has been explored in a number of studies. In one study, a simulation of samples derived from a routine arbovirus surveillance was done by infecting *Aedes aegypti* mosquitoes with dengue virus and then pooled with non-infected mosquitoes of the same species (Bishop-Lilly *et al.*, 2010). After purification of total RNA from mosquitoes, they were reverse transcribed using random primers after which they were subjected to 454 pyrosequencing. The study was able to correctly identify the infected mosquito pools (Bishop-Lilly *et al.*, 2010).

2.10 Genetic variability and its role in evolution of arboviruses

Genetic variability is an important aspect of viruses, allowing them to adapt to new environments as well as to quickly adapt to novel hosts (Longdon *et al.*, 2014). Thus genetic variation plays a role in host shift, whereby the risk of a host shift of a given virus is often dependent on the likelihood of accumulation of specific set of mutations necessary for infecting the new host (Longdon *et al.*, 2014). Arboviruses are maintained by the transmission between vertebrates and blood-feeding arthropods (Heeney, 2006). The vast majority of arboviruses are RNA viruses, with the only known DNA virus being African swine fever virus (Weaver & Reisen, 2010). The high mutation rates of RNA viruses may, therefore, explain why these viruses often host shift more frequently than other pathogens (Lauring & Andino, 2010).

There are three main possible ways of the origin of genetic variability in RNA viruses; mutation, recombination and re-assortment (Lauring & Andino, 2010). RNA viruses experience far higher mutation rates as compared to DNA viruses. This is due to their low fidelity RNA-dependent RNA polymerase (RdRp), large population size and rapid replication kinetics (Longdon *et al.*, 2014). Recombination refers to the process by which segments of genetic information are switched between the nucleotide strands of the different genetic variants, often during replication and where there is a coinfection with similar viruses (Lauring & Andino, 2010). Reassortment is a source of genetic variation that occurs in segmented viruses. Reassortment of the individual segments of RNA may occur, when there is a co-infection by RNA viruses, leading to production of new viral strains with new capabilities (Lauring & Andino, 2010).

The distribution of genetic variants in a population may change over time leading to evolution of the viruses and the development of different taxonomic entities (Longdon *et al.*, 2014). The study of viral variability and evolution is, therefore, quite important as it provides an insight into the possibility of the virus to develop new strains that could compromise control strategies, development of strains that have higher virulence or those that have the ability to be transmitted by a different host (Lauring & Andino, 2010).

CHAPTER THREE

MATERIALS AND METHODS

3.1 Study sites

The samples used in the study were collected prior to its commencement as part of an ongoing arbovirus surveillance program at Kenya Medical Research Institute's (KEMRI's) Centre for Virus Research (SSC#2540; WRAIR#1974). The sites where these samples were collected from include Baringo, Kacheliba, Turkana and Isiolo (Figure 3.1).

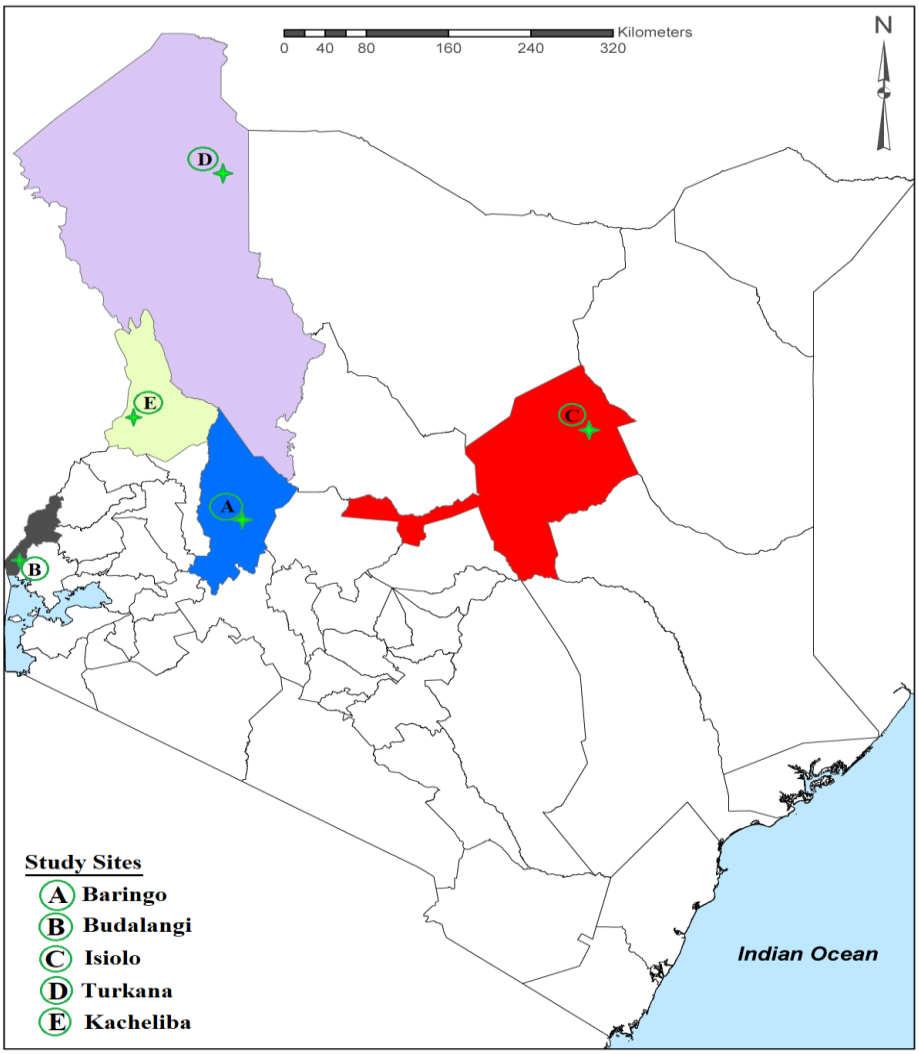


Figure 3. 1: A map showing sampling locations for the midges used in the current study.

3.2 Sample collection and processing

Experimental procedures in this study were undertaken at Centre for Virus Research, KEMRI, Nairobi, Kenya. The specimens were identified and pooled according to the collection site. Based on this criteria, insect pools containing up to 50 insects were kept in -80^{o}C freezer. The insect pools were subsequently processed following the standard laboratory procedures; accordingly, the specimens were mechanically homogenized using Copperhead Metal BBs beads (Crosman, USA) after which the supernatant was clarified by centrifugation at 8,000g for 10 minutes. The clear supernatant was separated from the pellet and they were kept in separate vials. The vials containing the supernatant as well as those with the pellet were kept in -80^{o}C freezer for subsequent use in RNA extraction and DNA extraction, respectively.

3.3 Bulk pool preparation of the samples

Individual samples from the different pools were combined to form bulk pools belonging to the five different sites; Baringo, Turkana, Isiolo, Kacheliba and Budalangi. For viral RNA extraction, all the individually clarified supernatants were combined in equal volumes in order to form five bulk pools representing the five sites. The combined supernatant were mixed by vortexing then used for RNA extraction. For DNA extraction, the individual pellets were similarly combined for each of the five sites. The combined pellet were further mechanically homogenized in order to ensure adequate mixing of the samples. The crude mixture was then used for DNA extraction.

3.4 Viral RNA extraction

Viral RNA was extracted from the clarified supernatant in the bulk pools, using QiaAmp Viral RNA Mini Kit (Qiagen, Germany). Prior to RNA extraction, the clarified supernatant was first passed through a 0.22um filter so as to remove large particles such as bacteria and thereby concentrate the viral particles. The filtrate was used for RNA extraction following the instructions of the manufacturer. To enrich the final RNA concentration, double the volume of the starting sample was used (i.e. instead of 140ul starting sample, I used 280ul). As a result, double the volume of the lysis solution, carrier RNA as well as the precipitation solution (100% ethanol) were

used. These were then passed through a single spin-column to purify the RNA from the solution and the RNA was finally eluted to a new nuclease free tube using 50 µl of nuclease free water.

3.5 Vector DNA extraction and amplification

DNA extraction was performed on the bulk pools of the crude homogenates using QIAmp DNA Mini kit (Qiagen, Germany), following the manufacturers instruction. Briefly, 20ul of Proteinase K was first added to 200ul of the crude mixture in order to degrade excess proteins and nucleases from the mixture. This was followed by addition of 200 ul of the lysis buffer to the sample after which they were mixed by pulse-vortexing. Lysis was carried out at 56^{o}C for 10 minutes. The lysed sample was then purified by micro-centrifugation at 8000 g in a spin column. The purified DNA were cleaned on the spin column using the kit-provided wash buffers. Afterwards, the cleaned DNA was eluted to nuclease free tubes by adding 100ul of elution buffer to the spin column.

The extracted DNA (5ul volume) was used to amplify a 710bp target of the vector cytochrome *c* oxidase I (COI) gene. The PCR was performed using AmpliTaq Gold 360 Master Mix (Invitrogen, USA) with the universal primers, LCO1490 and HCO2198. These primers target the mitochondrial COI gene of metazoan invertebrates (Table 3.1). The PCR cycling conditions were set as follows; initial denaturation at 95^{o}C for 10 mins, 35 cycles of 95^{o}C for 30s, 49^{o}C for 30s, 72^{o}C for 30s, and a final extension of 72^{o}C for 7 minutes. Successful amplification of the targets was confirmed by running a fraction of the PCR products in a 2% agarose gels stained with Sybr Green I nucleic acid stain (Invitrogen, USA) (Appendix 3). The remainder of the products were then kept in -20^{o}C for subsequent library preparation and sequencing.

Table 3. 1: Primers to be used for biting midge identification.

Primer	Orientation	Sequence (5'-3')	Conditions
LCO1490	F	GGTCAACAAATCATAAAGA TATTGG	95°C– denaturing 49°C – annealing
HCO2198	R	TAAACTTCAGGGTGACCAA AAAATCA	72°C – extension

Source: Folmer *et al.*, 1994.

3.6 Illumina library preparation for viral RNA sequencing

Paired-end libraries for high-throughput sequencing were prepared using Truseq Stranded mRNA Library Prep kit (Illumina, USA). The libraries were prepared following the manufacturer's recommended protocol, with modifications to exclude the mRNA purification steps. The input RNA was first quantified using Qubit® 3.0 Fluorometer and an input amount of 5ng/ul – 40ng/ul of the extracted RNA was used as input. Library preparation followed four major steps including reverse transcription, adapter ligation, enrichment, and finally pooling of the libraries for sequencing.

3.6.1 Reverse transcription of the viral RNA

The extracted RNA was used to carry out reverse transcription using the Superscript III Reverse Transcriptase (Invitrogen, USA). First, the extracted RNA was fragmented using the Fragment, prime and Finish mix (FPF) (Illumina, USA), which fragments the RNA and simultaneously primes for subsequent cDNA synthesis. The fragmentation and priming was carried out at 94^{o}C for 7 minutes in order to get to a mean fragment size of approximately 300bp. First strand cDNA synthesis was carried out in a total volume of 18ul of the fragmented and primed RNA using Superscript III Reverse Transcriptase and First Strand Synthesis Mix (Illumina, USA). The reverse transcription conditions included an initial incubation at 25^{o}C for 10 minutes and a cDNA synthesis step at 42^{o}C for 15 minutes. A final incubation step at 70^{o}C for 15 minutes was carried out to terminate the reaction as

recommended by the manufacturer. Second strand cDNA synthesis was carried out using DNA Polymersae I and RNase H (Illumina, USA). The synthesis was carried out by incubating the mixture in a pre-programmed thermocycler at 16^{o}C for 1 hour. In this reaction, RNase H digests the RNA strand from the cDNA:RNA hybrid after which DNA Polymersae I synthesizes the complementary strand of the single-stranded cDNA.

3.6.2 Adapter ligation and library enrichment

To prevent formation of chimeras and enhance adapter ligation efficiency, the fragments were first adenylated by adding a single 'A' to the 3' ends of the ds cDNA fragments. Adapters containing a single 'T' nucleotide overhang was ligated to both ends of the ds cDNA fragments. Ligated fragments were enriched by use of limited-cycle PCR. This process selectively amplifies those fragments with adapters on both ends, with the number of PCR cycles minimized to avoid skewed representation of the library.

3.6.3 Library pooling and sequencing

Enriched libraries were first purified with Ampure XP beads (Beckman coulter, USA), a process that also size-selects the fragments by removing very short library fragments. The concentration of the fragments were subsequently measured using Qubit Fluorometer with dsDNA HS assay kit (Invitrogen, USA). The size distribution of the fragments were estimated by gel electrophoresis with a 2% agarose gel stained with Sybr Green I nucleic acid stain (Invitrogen, USA). The library molecule concentration in each of the libraries was then calculated and equal amount from each library used to pool all samples to a single library for sequencing. The final library was denatured with NaOH and diluted further to 12pM before loading onto the Illumina Miseq sequencing machine. Sequencing was performed using Miseq Reagent V3 reagents (Illumina, USA), in a 600-cycle sequencing format.

3.7 MinION library preparation for COI amplicon sequencing

The COI amplicons were first purified using AMPure XP beads (Beckman Coulter, USA) following the recommendations of the manufacturer. Briefly, 18ul of the beads

were added to 10ul of the sample. DNA fragments were allowed to bind to the paramagnetic beads, after which they were separated from the contaminants. Separated DNA fragments were washed with 70% ethanol and the washed fragments were subsequently eluted from the beads using nuclease-free water. Quantification of the purified products was performed with Qubit dsDNA HS Assay Kit (Invitrogen, USA), using the Qubit fluorometer 2.0. Based on the size of the amplified products, which is ~710bp, and the determined concentration, the volume of the amplicons that yields 200 fmol was calculated and used to make the libraries. The COI libraries were prepared with Ligation Sequencing Kit (Oxford Nanopore Tech., UK), following the instructions of the manufacturer. The purified COI amplicons were first end-repaired using the New England Biolab's NEBNext Ultra II End repair/dA-tailing Module. The end-repaired amplicons were then individually barcoded with the Native Barcoding Expansion 1-12 kit (Oxford Nanopore Tech., UK). Native barcodes were ligated to the end-repaired amplicons using the Blunt/TA Ligase Master Mix (New England Biolab's, UK). The barcoded libraries were then pooled and the sequencing adapters ligated to them using NEBNext Quick Ligation Module (NEB, UK). The final library was loaded onto the flowcell (FLO-MIN106D) and sequenced on a MinION MK1c device, following the workflows provided in the MinKNOW software (Oxford Nanopore Tech., UK).

3.8 Bioinformatics data analysis

3.8.1 Illumina sequence analysis and virus identification

Raw sequence reads were first subjected to quality control using TimGalore v0.6.5 so as to remove adapters. Further, PrinseqLite v0.20.4 was used to filter low-quality reads with the following settings: minimum phred quality score of 30, minimum length of 50 and a maximum length of 300. Considering a ribosomal RNA (rRNA) reduction step was not conducted prior to sequencing, this process was carried out *in silico*. The rRNA sequence reads were removed by using riboPicker v0.4.3, a program that removes excess rRNA sequences by comparing the sequence reads against the SILVA rRNA database (Quast *et al.*, 2013). In this study, the SILVA rRNA database release 138.1 was used.

The cleaned sequence reads were assembled *de novo*. First, paired-end reads were merged using PEAR 0.9.8 (Zhang *et al.*, 2014) and the reads were then assembled using Trinity program (Grabherr *et al.*, 2011) with default parameters. As another layer of quality control step, the cleaned reads were mapped back to the assembled contigs, and the contigs were then filtered to only remain with those that had at least 90% of the bases with a minimum of 5x coverage (Wajid & Serpedin, 2014).

The validated contigs were first subjected to BLAST analysis using NCBI Viral database (http://blast.ncbi.nlm.nih.gov/Blast.cgi) with the BLASTx program. Virus related contigs were selected from the results of the BLAST search, and further BLAST analysis performed against the entire NCBI nr database using the BLASTx program, so as to filter out all non-viral contigs. Only those contigs that showed significant hits to viruses were retained for further analysis.

3.8.2 Genome-scale characterisation and evolutionary analysis

Complete viral genome sequences were generated by comparing the viral contigs to closely related homologs in Genbank. Contigs that showed hits to similar homologs were aligned using mafft and a contiguous sequences, representing the different viruses generated. Where the full genome of a virus was recovered by de novo assembly, the contig was inspected for any assembly errors through alignment with other related sequences obtained from Genbank. Gene prediction and functional annotation of the viral sequences were then performed. Translation and prediction of open reading frames (ORF) of the respective viruses was carried out in Expasy server, using the Translate tool (Artimo *et al.*, 2012). Functional annotation of the predicted ORFs were identified through web-based comparison to the Pfam database (Mistry *et al.*, 2020).

To describe the identified viruses in an evolutionary context and ensure a meaningful depiction of their evolutionary relationships, RNA-dependent RNA polymerase (RdRp) gene was used to carry out phylogenetic analysis. Closely related RdRp gene sequences were retrieved from Genbank and used as reference sequences in reconstructing the phylogenetic relationship of the viral sequences. The combined set of sequences were codon-aligned using Muscle software embedded in Molecular

Evolutionary Genetics Analysis v.7.0 (MEGA7) software. The aligned sequences were edited using trimAI v1.2 (Capella-Gutiérrez *et al.*, 2009) and maximum likelihood phylogenetic analysis carried out using IQ-TREE v1.6.12 (Nguyen *et al.*, 2015). The best model and tree search was performed simultaneously based on 2000 bootstrap estimates and approximate likelihood ratio test (alrt).

3.8.3 Vector sequence analysis and taxonomic assignment

Base-calling and demultiplexing were performed on-board the MinION Mk1C device using Guppy. Sequencing reads were quality filtered with Nanofilt v2.8.0 (De Coster *et al.*, 2018), 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 710bp (with a 150bp buffer) were also filtered. Error correction of the sequence reads was performed using isONclust v0.0.6 and isONcorrect v0.0.8 (Sahlin & Medvedev, 2020, 2021), using default parameters with *--ont* flag. The corrected reads were resampled to approximately 11,000 reads per sample, using rasusa v0.5.0 (M. Hall, 2019). Read clustering, consensus sequence generation and determination of the number of reads supporting each consensus sequence was carried out using IsoCon v0.2.5.1 (Sahlin *et al.*, 2018). IsoCon treats reverse complements and sequence duplicates of varying lengths as different. Therefore, these were further removed by performing clustering of the consensus sequences using cd-hit-est (Fu *et al.*, 2012), with 98.9% similarity threshold which is the lowest accuracy of error-corrected nanopore reads (Sahlin & Medvedev, 2021).

Taxonomic assignment of the generated consensus sequences was performed in the MIDORI server, using RDPClassifier with COI reference sequence database (Wang *et al.*, 2007; Leray *et al.*, 2018). MIDORI server uses reference sequences retrieved from GenBank and the Barcode of Life Data Systems Identification engine (BOLD-ID), and it allows identification of metazoan sequences to species level. Taxonomic assignments were further validated by searching the consensus sequences against the NCBI nr database, in order to determine the lowest classification of each of the sequence. Any OTUs that were not classified as belonging to an expected

invertebrate phylum were excluded. To increase the reliability of the identified OTUs, singletons and OTUs supported by <10 sequences were removed. Further, invertebrate species supported by less than one percent of the total sequences in each site were also removed from the final analysis.

3.9 Ethical consideration

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

CHAPTER FOUR

RESULTS

4.1 Virus identification in biting midges

4.1.1 Initial analysis and quality control

A total of 3,351 midge specimens were processed in this study. Out of these, 1063 originated from Turkana, 892 from Baringo while Isiolo, Budalangi and Kacheliba had 640, 600 and 156 specimens respectively. These vector specimens were used to create bulk pools representing each of the five sites and they were subjected to high-throughput sequencing. Illumina sequencing of the RNA viruses was successful in all the five bulk pools. Approximately 18 million sequence reads were generated in the study, and the mean read length ranged from 130 and 205bp. After read filtering, a total number of 12 million reads were obtained. Read distribution across the different samples representing the five sites as well as their quality statistics are given below (Table 4.1).

Table 4. 1: Table showing Illumina sequencing statistics across the five sites.

Site	Raw Reads	Filtered Reads	Mean Read Length (bp)	No. of contigs	No. of viruses
Baringo	1,059,487	715,869	150	115	1
Budalangi	965,066	658,245	195	142	1
Turkana	7,050,872	4,892,562	130	1402	10
Isiolo	3,431,858	2,011,532	205	297	1
Kacheliba	5,811,739	3,842,890	200	183	2

4.1.2 *De novo* assembly and classification of the viral sequences

De novo sequence assembly and BLAST analysis resulted in the identification of contigs that showed hits to 15 distinct viruses. The similarity thresholds along the RdRp gene, which is the hallmark of all RNA viruses, varied among the different contigs ranging between 39.13% and 97.7% similarity to the previously sequenced

viruses available in Genbank. Among these viruses, 5 of them were similar to members of the *Iflaviridae* family and 4 showed similarity to members of the *Partitiviridae* family. *Picornaviridae*, *Solemoviridae*, *Tombusviridae, Totiviridae, Chuviridae* and the recently proposed Negevirus taxon each contributed one virus among the viruses identified in the study (Table 4.2).

Table 4. 2: Table showing the list of viruses identified in the study with their identification details based on BLAST analysis.

Strain	Site	Family Classification	BLAST Analysis		
			Accession	Closest Hit	% Identity
Turkana_1	Turkana	*Solemoviridae*	MF893251	Medway virus	59.05
Turkana_2	Turkana	*Chuviridae*	KX924630	Chuvirus Mos8Chu0	42.39
Turkana_3	Turkana	Tombus-like	KX235518	Diaphorina citri associated C virus	55.24
Turkana_4	Turkana	*Totiviridae*	MK440653	Lindangsbacken virus	50.79
Turkana_5	Turkana	*Iflaviridae*	NC_024016	Heliconius erato iflavirus	85.62
Turkana_6	Turkana	*Partitiviridae*	LC533398	Lichen partiti-like RNA virus	46.14
Turkana_7	Turkana	Partiti-like	KX884215	Hubei partiti-like virus 45	52.74
Turkana_8	Turkana	Iflavi-like	MN784069	Redbank virus	66.3
Turkana_9	Turkana	Negevirus	MT344121	Sandewavirus dungfly1	39.13
Turkana_10	Turkana	Picorna-like	MH614292	Boghill Burn virus	84.27
Baringo_1	Baringo	Partiti-like	MF344586	Araticum virus	56.01

Budalangi_1	Budalangi	Iflavi-like	MN784065	Redbank virus	70.84
Isiolo_1	Isiolo	*Iflaviridae*	NC_040574	*Culex* Iflavi-like virus 4	97.7
Kacheliba_1	Kacheliba	Iflavi-like	NC_033201	Hubei picorna-like virus 38	46.37
Kacheliba_2	Kacheliba	*Partitiviridae*	JX658566	Grapevine partitivirus	51.61

Among the 5 viral contigs showing similarity to viruses in the *Iflaviridae* family, two of them originated from Turkana and one each originated from Isiolo, Budalangi and Kacheliba. The two viruses from Turkana included Turkana_5 and Turkana_8. Turkana_5 had an 85.62% similarity to Heliconius erato iflavirus (Table 4.2), which is an iflavirus initially detected in Heliconius butterflies in Costa Rica (Smith *et al.*, 2014). On the other hand, Turkana_8 had a 66.3% similarity to Redbank virus (Table 4.2). Similar to Turkana_8, Budalangi_1 viral contig which originated from Budalangi also showed high similarity to Redbank virus, with a percentage similarity threshold of 70.84%. Redbank virus is an unclassified iflavi-like virus recently identified in mosquito fecal microbiota in Australia (Ramírez *et al.*, 2020). The *Iflaviridae* virus from Isiolo, Isiolo_1, had a very high similarity of 97.7% to *Culex* Iflavi-like virus 4, which is an Iflavi-like virus previously identified in *Culex* sp. Mosquitoes in the USA (Sadeghi *et al.*, 2018). Kacheliba_1 is another *Iflaviridae* viral contig that was obtained from a pool of midges from Kacheliba. This virus had a 46.6% similarity to Hubei picorna-like virus 38 (Table 4.2).

The 4 *Partitiviridae* viral contigs identified in the study were obtained from pools of midges from different sites, with two contigs originating from Turkana (Turkana_6 and Turkana_7) and one each from Baringo (Baringo_1) and Kacheliba (Kacheliba_2). The similarity thresholds for all these viral contigs was highly diverse. Turkana_6 had a significantly lower similarity of 46.14% to Lichen partiti-like virus. On the other hand, Baringo_1 had a comparatively higher percent similarity of 56.01% to Araticum virus. The other two viruses included Turkana_7, which had a 52.74% similarity to Hubei partiti-like virus 45, and Kacheliba_2, which had a

51.61% similarity to Grapevine partitivirus. *Partitiviridae* family contains viruses with two segments, dsRNA1 and dsRNA2 (Vainio *et al.*, 2018). The viruses obtained in the study only showed hits to the segment corresponding to dsRNA1. This segment contains the ORF which codes for the RNA-dependent RNA polymerase gene.

The *Picornaviridae* family only had one viral contig, Turkana_10, which originated from a pool of midges from Turkana. Turkana_10 had an 84.27% similarity to the Boghill Burn virus. Boghill Burn virus is an unclassified virus in the *Picornaviridae* family which was initially identified in *Bombus* sp. bees in Scotland (Pascall *et al.*, 2019). Five other viral contigs were also obtained in the pool of midges from Turkana. These included Turkana_1, Turkana_3, Turkana_2, Turkana_4 and Turkana_9 which were classified as belonging to *Solemoviridae* family, *Tombusviridae* family, *Chuviridae* family, *Totiviridae* family and the unclassified Negevirus taxon, respectively. Turkana_1 had a 59.05% similarity to the Medway virus, an unclassified sobemo-like virus. Turkana_3 is an 1142bp sequence which had a 55.24% similarity to the Diaphorina citri associated C virus, which is an unclassified virus in the *Tombusviridae* family. The *Chuviridae* viral contig obtained in this study had a 42.39% similarity to Chuvirus Mos8Chu0 (Table 4.2). Chuvirus Mos8Chu0 (KX924630.1) is a bi-segmented virus in *Chuviridae* family which was initially identified in *Culiseta minnesotae* mosquitoes. The *Totitviridae* viral contig, Turkana_4, obtained showed similarity to unclassified members of this family. Specifically, it had a 50.79% similarity to Lindangsbacken virus. The Negevirus-like contig obtained in the study, Turkana_9, had a considerably low similarity to other Negeviruses available in Genbank. More specifically, it had a 39.13% similarity to Sandewavirus dungfly1, a negevirus obtained from dungfly in the arctic yellow river station (Lu *et al.*, 2020).

4.2 Characterization and evolutionary analysis

4.2.1 Genome-scale characterisation of the identified viruses

The completeness of the assembled viral contigs was confirmed by alignment and comparison to their closely related full viral genomes. Of the 15 viral contigs

obtained 9 of them were complete genomes and 6 were partial genomes; genome length of these viruses is approximately 9kb (Figure 4.1). Protein translations of the *Iflaviridae* viruses from the study (Turkana_5, Turkana_8, Isiolo_1, Budalangi_1 and Kacheliba_1) were consistent with those of other viruses in this family; with a single open reading frame (ORF) of approximately 2891aa polyprotein which codes for multiple viral genes. Gene predictions found an approximately 310aa region at the 3' end which codes for the RNA-dependent RNA polymerase (RdRP gene). Adjacent to it was the RNA helicase coding region, which is important in the cleavage of the pre-translated protein to the individual genes. Similar to *Iflaviridae*, Turkana_10, which falls under *Picornaviridae* family also showed the same genome architecture (Figure 4.1).

Turkana_9, which is a negevirus, was another fully assembled sequence in this study. Like other negeviruses, ORF prediction of Turkana_9 resulted in 3 ORF regions of varying lengths. The first, ORF1, is a 2325aa region coding for the RdRP protein. The functional prediction of ORF2 of Turkana_9 was not successful. ORF3, however, is a 253aa region which codes for the Putative virion membrane protein. Turkana_1 was the only di-partite virus assembled in the study. The long segment contain two ORFs with the first, ORF1, being a 468aa protein coding region whose function could not be assigned. ORF2 is an RdRP gene coding region while the only ORF in the shorter segment is a 215aa coding region which was inferred to be a viral coat protein coding region. With regards to Turkana_2, only a partial sequence corresponding to the entire Segment L of viruses in the *Chuviridae* family was obtained. This segment contains a single ORF which is the putative RdRP coding gene (Figure 4.1). The same was observed among the identified viruses that were falling under *Partitiviridae* family; Baringo_1, Turkana_7, Kacheliba_2 and Turkana_6. In this group of viruses, the segment corresponding to RNA1 segment of viruses in *Partitiviridae* family was ibtained. This segment contains a single ORF; approximately 470aa region that codes for the RdRP gene.

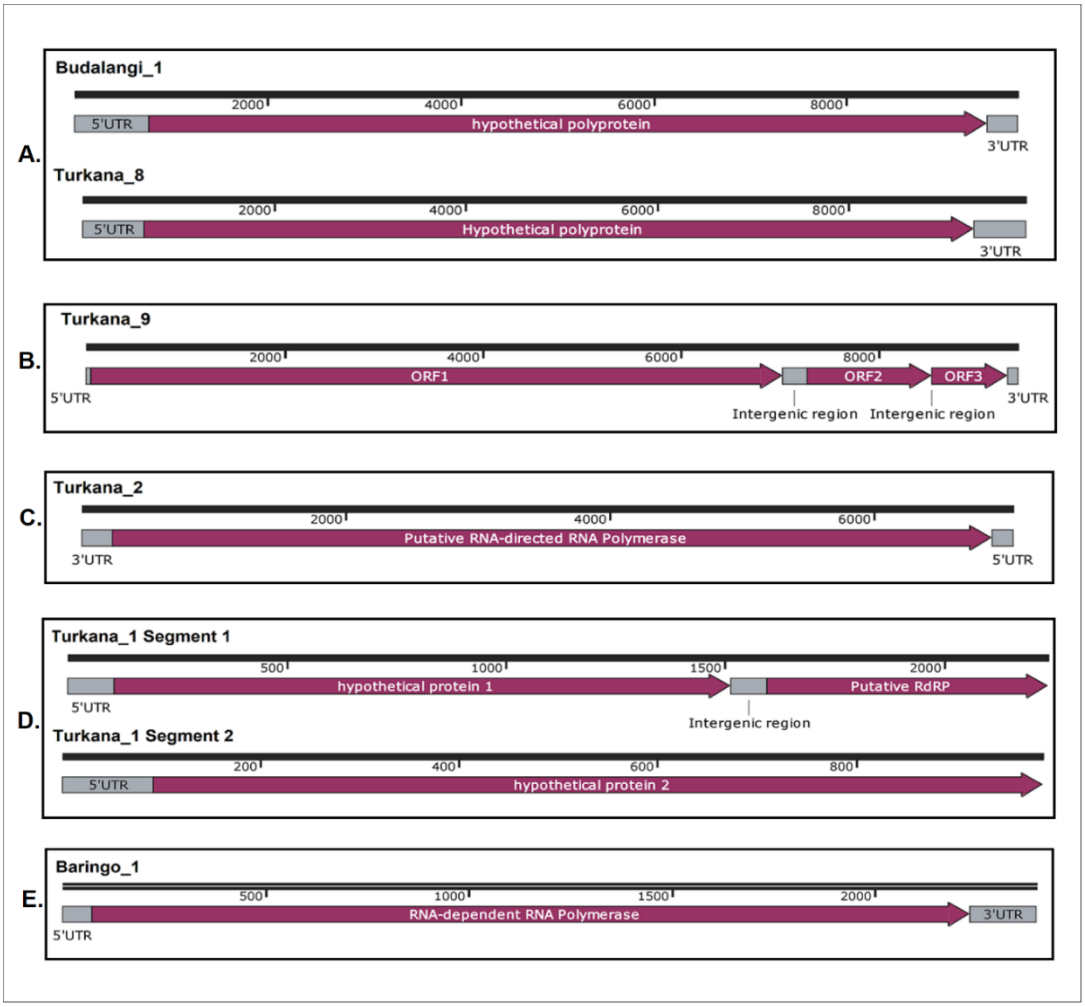


Figure 4. 1: Genome architectures of the representative viral genomes identified. (A.) complete Iflavi-like viruses, (B.) complete novel negevirus genome, (C.) partial *chuviridae* genome, showing the RNA-dependent RNA polymerase coding segment, (D.) novel *solemoviridae* virus with the two putative segments, and (E.) the representative partial *partitiviridae* genome, showing the RNA-dependent RNA polymerase coding segment.

4.2.2 Phylogenetic analysis of the identified RNA viruses

To provide an evolutionary context for the viruses identified, phylogenetic analysis of the newly discovered viruses was carried out together with reference virus strains of the same family available in Genbank. The viruses identified in this study were classified to 7 different families; *Iflaviridae, Partitiviridae, Tombusviridae, Totiviridae, Solemoviridae, Chuviridae* and the yet to be described negevirus taxon.

Publicly available viruses belonging to these different groups, and more specifically those closely related to the viral strains obtained in the current study were downloaded and used as reference sequences in the reconstruction of phylogenetic trees. Apart from Isiolo_1, Turkana_5 and Turkana_10; all the other identified viruses showed a significantly high evolutionary divergence from publicly available viruses. The most divergent virus is Turkana_9, which clustered with negeviruses (Figure 4.2). Negeviruses are insect-specific viruses (ISVs) and are closely related to viruses in Kitaviridae family, which are plant-specific viruses (Figure 4.2). Turkana_9 clustered under negevirus group and it formed a single clade with Sandewavirus, Tanay virus, Dezidougou virus, and Goutanap virus. Turkana_5 is the only virus which clustered with known viruses, and it clustered with Iflaviruses (Figure 4.2). All the other viruses showed high similarity and clustered with diverse virus strains, most of which remain as unclassified RNA viruses or unclassified viruses within the specific virus families. Turkana_6 and Kacheliba_2, in particular, clustered with unclassified *Partitiviridae* viruses (Figure 4.3); Isiolo_1 formed a cluster with unclassified *Iflaviridae* family members (Figure 4.2); Turkana_1 formed a cluster with members of the unclassified *Solemoviridae* family, Turkana_4 formed a cluster with members of the unclassified *Totiviridae* family, and Turkana_2 formed a cluster with members of the unclassified *Chuviridae* family (Figure 4.3). Turkana_7 and Baringo_1 were found in a cluster that has unclassified Partiti-like viruses (Figure 4.3); Budalangi_1, Turkana_8, and Kacheliba_1 clustered with unclassified Iflavi-like viruses; Turkana_3 and Turkana_10 clustered with unclassified Tombus-like viruses and unclassified Picorna-like viruses, respectively (Figure 4.2).

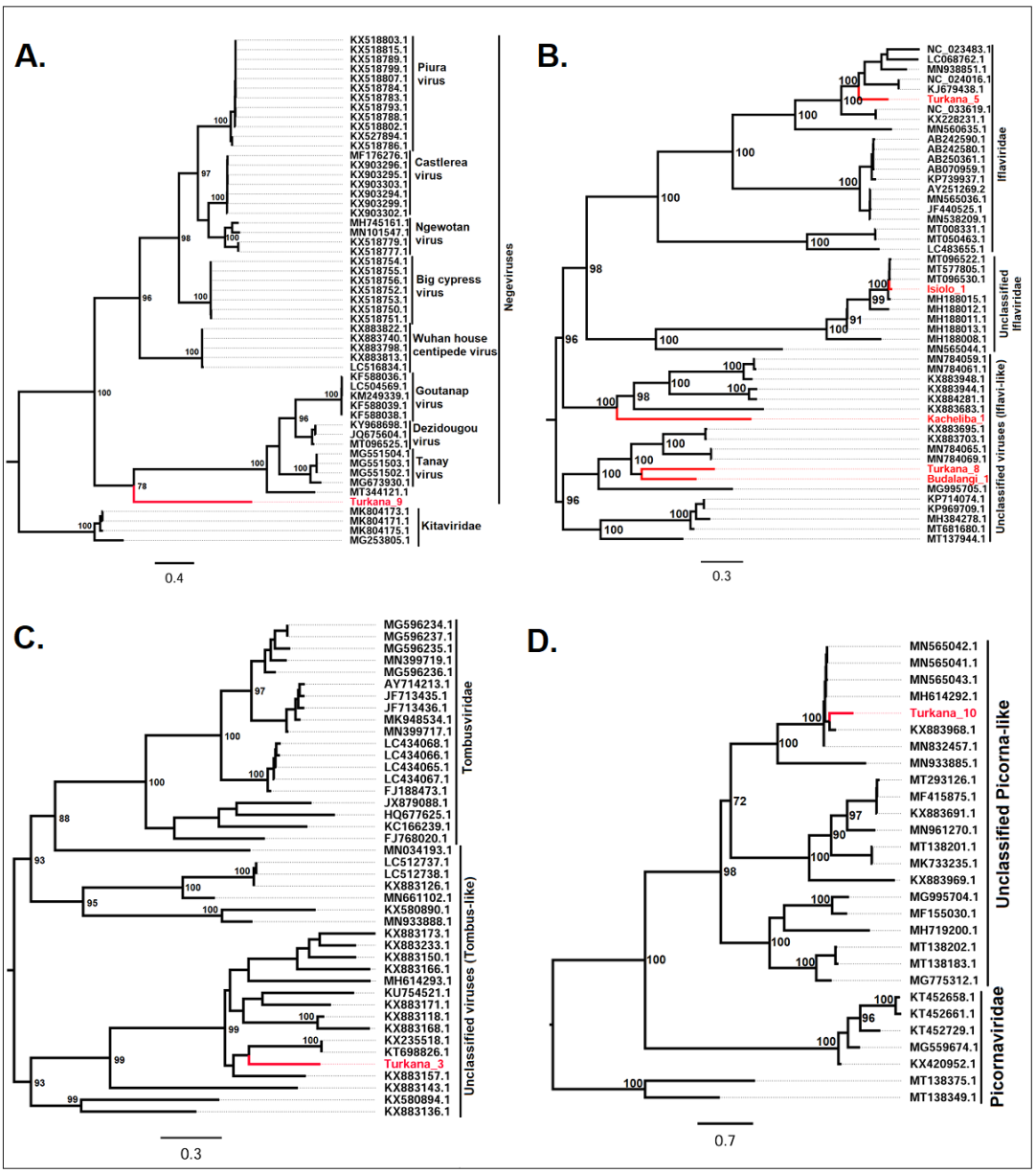


Figure 4. 2: Maximum likelihood phylogenies of A. negeviruses, B. *Iflaviridae*, C. *Tombusviridae*, and D. *Picornaviridae* viruses.

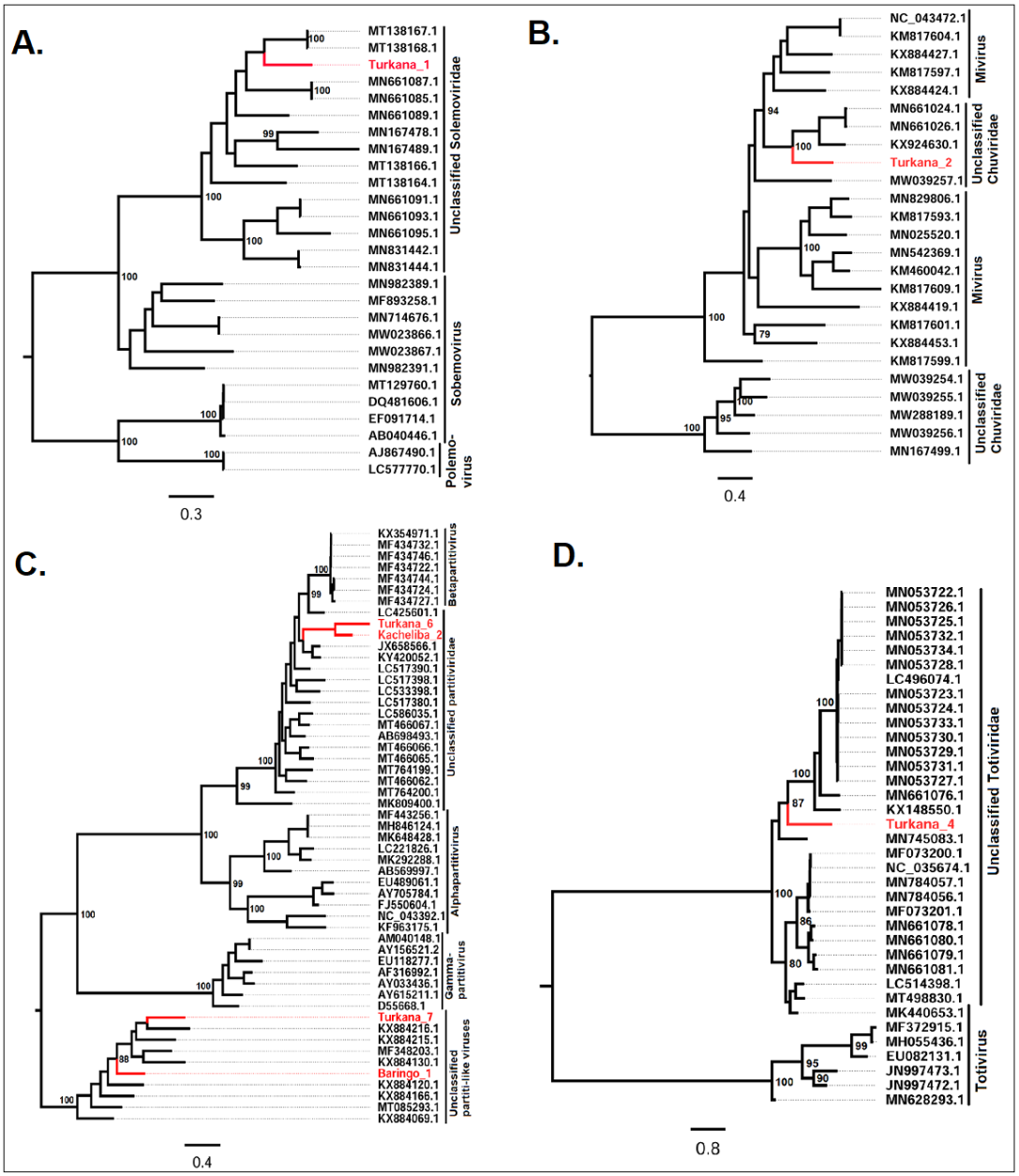


Figure 4. 3: Maximum likelihood phylogenies of A. *Solemoviridae* B. *Chuviridae* C. *Partitiviridae* and D. *Totiviridae* viruses.

4.3 Characterization of the vector pools

4.3.1 DNA extraction and amplification

Extracted DNA showed clean, good quality and a high molecular weight DNA. Assessment by spectrophotometry showed a ratio of 260/280 absorbance to be 1.8.

Gel electrophoresis showed successful amplification of approximately 710bp target of the COI gene (Appendix 3).

4.3.2 MinION sequencing and quality statistics

Sequencing was successful in all the 5 pools. Approximately 248,173 raw sequence reads, with the read distribution among the 5 sites as given below, were sequenced (Table 4.3). The average quality scores for the sequenced reads ranged from 15 to 20 and the median lengths of the reads for the 5 sites ranging from 685 to 699 base pairs.

Table 4. 3: Nanopore sequencing statistics of the vector pools representing the five sites.

Site	Raw Sequence Reads			Cleaned Sequences	
	No. of Reads	Length (Median)	Avg. QScore	Length (Median)	Number of Clusters
Isiolo	90523	696	15	700	633
Baringo	47398	699	19	700	522
Turkana	52722	695	16	700	652
Kacheliba	32469	685	17	698	636
Budalangi	25061	694	20	697	328

4.3.3 Sequence clustering and taxonomic assignment

The metabarcoding analysis yielded a diverse set of reads that were assigned to various families of midges, including the *Chironomidae*, *Ceratopogonidae*, and *Cecidomydiae*. All of the reads in the Baringo pool were assigned to the *Ceratopogonidae* family, with *Culicoides leucostictus*, *Culicoides pycnostictus*, and *Culicoides nivosus* accounting for 65.46, 19.81, and 13.26 percent of the reads, respectively (Figure 4.4). Similarly, all of the reads from the Budalangi pool were assigned to the *Ceratopogonidae* family, specifically *Culicoides leucostictus* species. In Kacheliba site, all of the reads identified were assigned to the *Cecidomyidae*

family (Figure 4.4). On the other hand, Isiolo and Turkana sites had a relatively diverse number of reads which were classified to various species of midge. The *Ceratopogonidae* family accounted for 44.07 percent of the reads in the Isiolo site, while *Cecidomydiae* family accounted for 21.03 per cent (Figure 4.4). The remaining percentage of reads was assigned to the *Chironomidae* family. Reads classified as belonging to *Ceratopogonidae* family were common in Turkana, accounting for 68.49 percent of all reads. *Chironomidae* and *Cecidomyiidae* families accounted for 25.64 and 5.88 percent of the reads in Turkana, respectively (Figure 4.4).

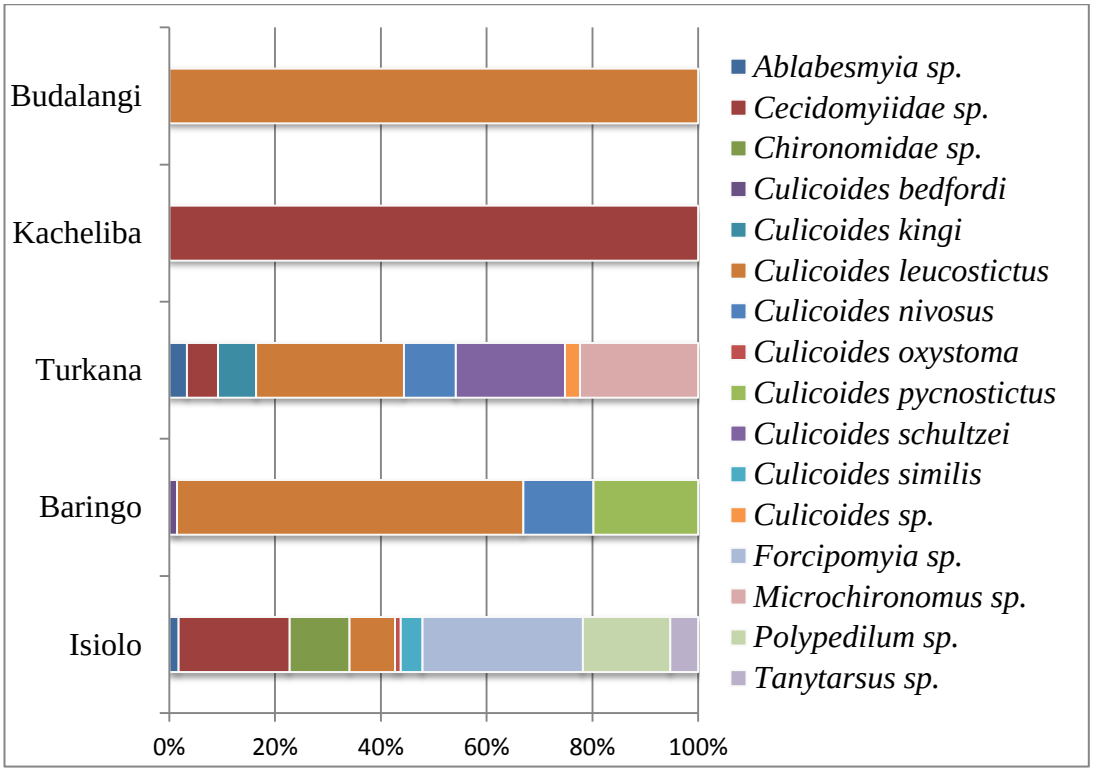


Figure 4. 4: A hundred percent stacked column chart showing the relative abundance of the different species of midges in the five sites based on the total reads in each of the respective sites.

The obtained specimen sequences were classified into OTUs. A total of 187 OTUs from different midge families were obtained. Turkana and Isiolo each had 51 OTUs, while Baringo and Kacheliba each had 36 OTUs (Figure 4.5). Despite a relatively high number of specimens processed and the total reads obtained, Budalangi site had a relatively lower number of OTUs with 13 (Figure 4.5). The findings of this study

were generally consistent with the number of specimens processed at each of the sites.

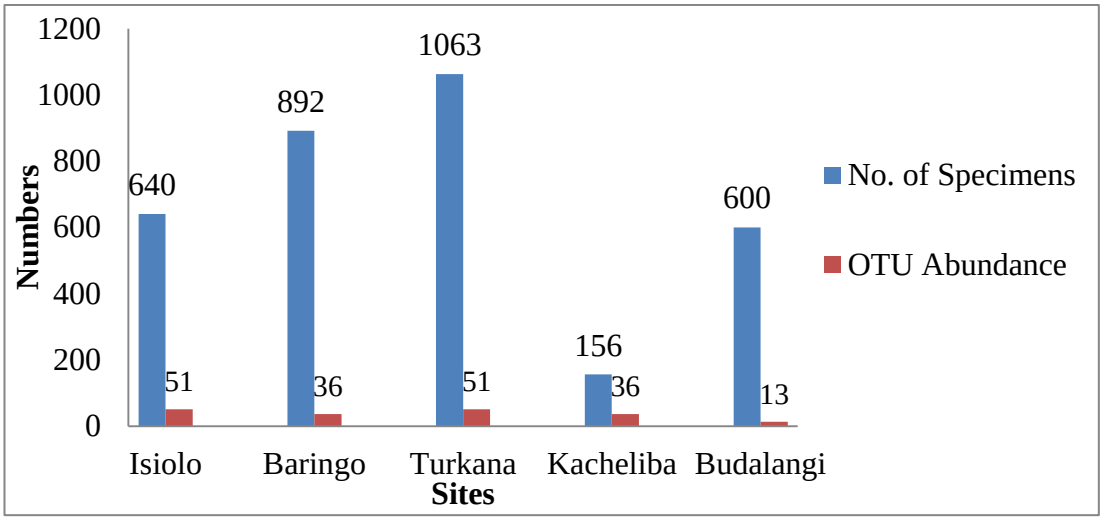


Figure 4. 5: Bar graphs showing the number of specimens processed and the number of OTUs obtained.

The species richness of the five different sites varied, with Isiolo and Turkana having the most diverse midge species. *Ceratopogonidae*, which included *Culicoides leucostictus*, *Culicoides oxystoma*, *Culicoides similis*, and unclassified *Forcipomyia* sp., were among midges that were found in Isiolo. Midges of the *Chironomidae* family were also identified in Isiolo, including *Ablabesmyia* sp., *Polypedilum* sp., *Tanytarsus* sp., and others that remain unclassified within the *Chironomidae* family. In addition, 21 OTUs were identified in Isiolo which shared a high degree of similarity with unclassified *Cecidomyidae* family members (Table 4.4). In Turkana site; *Culicoides kingi*, *Culicoides leucostictus*, *Culicoides nivosus*, *Culicoides schultzei*, and unclassified *Culicoides* sp. were found. In addition, OTUs of the *Chironomidae* family, including *Ablabesmyia* sp. and *Microchironomus* sp., were discovered (Table 4.4). OTUs from Baringo and Budalangi sites were all assigned to the *Ceraopogonidae* family. *Culicoides leucostictus* was found in both of these locations. In addition, *Culicoides bedfordi*, *Culicoides nivosus*, and *Culicoides pycnostictus* were found at the Baringo site. All of the OTUs identified at Kacheliba site belonged to the *Cecidomyidae* family. These OTUs, in particular, shared a high

degree of similarity with unclassified members of the *Cecidomyiidae* family (Table 4.4).

Table 4. 4: The community composition of the various vector pools in each site.

Site	Family	Species	Sequence Abundance	Fraction of Reads	No. of OTUs
Isiolo	*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	*Ceratopogonidae*	*Culicoides bedfordi*	109	1.46	7
		Culicoides leucostictus	4871	65.46	17
		Culicoides nivosus	987	13.26	4
		Culicoides pycnostictus	1474	19.81	8
Turkana	*Cecidomyiidae*	*Cecidomyiidae* sp.	212	5.88	8
	Ceratopogonidae	*Culicoides kingi*	260	7.21	6
		Culicoides leucostictus	1010	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	*Cecidomyiidae*	*Cecidomyiidae* sp.	1110	100	36
Budalangi	*Ceratopogonidae*	*Culicoides leucostictus*	4940	100	13

CHAPTER FIVE
DISCUSSION, CONCLUSION AND RECOMMENDATIONS

5.1 Discussion

The vast majority of viruses causing disease in humans and animals today are RNA viruses, specifically arthropod-transmitted viruses. They are responsible for diseases like Dengue, West Nile, Bluetongue, Schmallenberg, and Yellow fever, among others. In this study, metagenomics and metabarcoding methods were used on field-collected midges to characterize viruses and their insect hosts (Besansky *et al.*, 2003). Several RNA viruses were identified in the study, where complete and near-complete genomes of these viruses were recovered. In addition, different species of midges that are plausibly associated with these viruses were identified. These findings add to the currently known diversity of RNA viruses among biting midges.

In the study, up to 15 RNA viruses were genetically characterized by high-throughput sequencing in the five sites. The majority of the viruses identified have low similarity thresholds to existing viruses, with as low as 39.13 percent aa similarity. Only Isiolo_1, Turkana_5, and Turkana_10 had higher aa similarities to existing viruses with 97.7 percent, 85.62 percent and 84.27 percent, respectively. All the other viruses had low similarity thresholds between 39.13 and 70.84 percent aa similarities to existing viruses. It should be noted, however, that taxonomic classification of these identified viruses was not performed to their respective species in the current study. The International Committee on Taxonomy of Viruses (ICTV) establishes various criteria for the classification of virus species, which frequently differ depending on virus group and include information other than the virus's genetic sequences (Fauquet & Fargette, 2005). Future research should be conducted to classify the identified viruses into their respective species.

The majority of the viruses identified in this study are insect-specific (ISVs), and none are known to be pathogenic to vertebrate hosts. These viruses are, therefore, unlikely to cause diseases. However, studies have shown that ISVs can potentially influence arthropod vector competence by interfering with their vectorial capacity for pathogenic viruses; possibly due to competitive inhibition (Vasilakis & Tesh, 2015; R. A. Hall *et al.*, 2016; Öhlund *et al.*, 2019). As a result, the viruses identified in this

study have the potential to be important biocontrol agents in the transmission of pathogenic viruses. Furthermore, the identification of these viruses may help to fill important gaps in the phylogeny of viruses and contribute to studies aimed at understanding the origin and evolution of pathogenic viruses (Forterre, 2006; Parvez & Parveen, 2017).

Several viruses found in this study are members of families known to be associated with hosts other than arthropods. *Tombusviridae, Solemoviridae, Partitiviridae, Picornaviridae,* and *Totiviridae* are among them. Viruses in the *Tombusviridae* and *Solemoviridae* families, for example, have plants as their natural hosts. The *Picornaviridae* family, on the other hand, contains viruses that are only known to infect vertebrates. The *Partitiviridae* and *Totiviridae* families contain viruses with a wide range of natural hosts. *Partitiviridae* are known natural hosts of fungi and plants, whereas *Totiviridae* viruses are known to have fungi and protozoan parasites as their natural hosts. The findings, therefore, suggest that these viruses may also be associated with midges. However, it is also possible that this observation was caused by the sample processing strategy. The entire invertebrate specimens were homogenized during sample processing and as such, some of the viruses detected could have come from undigested food, gut microflora, or even parasites that were present in the invertebrates during processing. This could, particularly, be true for *Ceratopogonidae* members whose food sources are known to be quite diverse and includes mammals, birds and even mosquitoes (Ma *et al.*, 2013; Slama *et al.*, 2015; Tomazatos *et al.*, 2020). It would, therefore, not be surprising if some of the viruses discovered originated in hosts other than the vectors used in the current study.

Insect community composition identified in the current study is quite diverse. The metabarcoding method was successful in identifying numerous species of midges at each site. The proportion of reads specific to each of the species identified in this study varied greatly across the study sites. Considering the sensitivity issues associated with HTS such as primer biases, these observations cannot be used to reliably estimate the relative abundance of each of the identified species (Clarke *et al.*, 2014; Elbrecht & Leese, 2015). Nonetheless, the identification of a specific species in a pool of midges is reason enough to consider that species as a possible host of the viruses identified in the study. This is due to the fact that the identified

virus could have come from any of the midge species in the pool, regardless of their abundance. The findings in this study, therefore, present us with a unique opportunity to infer the possible hosts of the detected viruses using methods such as co-occurrence networks (Starr *et al.,* 2019). However, such an approach would necessitate the sequencing of multiple pools from a given locality for use in the network. Further, to improve the accuracy of host-virus association methods, approaches to reducing PCR bias can also be considered. Some methods for reducing PCR bias include *in-silico* primer testing before use, the use of multiple sets of primers, and the use of PCR-free shotgun sequencing pipelines. (Zhou *et al.,* 2009; Clarke *et al.,* 2014; Gibson *et al.,* 2014). Future studies should take some of these requirements into consideration in order to definitively link the identified viruses to their insect hosts.

The metagenomics and metabarcoding methods used in this study have the potential to be low-cost approaches to arbovirus and insect surveillance. These methods have the advantage of being able to analyze hundreds to thousands of insect specimens in a single pool. In contrast, traditional methods would process between 25 and 50 specimens per pool. Furthermore, the traditional DNA barcoding method for species identification and confirmation would only processes a single specimen at a time. Thus, when metagenomics is combined with metabarcoding in insect surveillance, the costs and labour are dramatically reduced. However, the deployment of these two methods for routine use may not be possible at this stage. This is due to the disadvantages inherent to these two methods. When compared to quantitative PCR, metagenomics approach is not as sensitive. This is especially true when it is used to detect viruses with extremely low titres (Wylie *et al.*, 2012). Nonetheless, using different enrichment methods would improve the sensitivity of this method in general. Another limitation of metagenomics approach is the difficulty in associating detected viruses with vector hosts. The use of metabarcoding, as applied in this study, aids in narrowing down the potential hosts of the identified viruses. However, as previously stated, further improvement of this method is required to overcome the challenge of associating the individual virus detected to one of the possible insect species. Future research will benefit from using viral enrichment methods as well as

methods such as co-occurrence networks to infer the species associated with the identified viruses.

5.2 Conclusion

In conclusion, this study has successfully identified and genetically characterized several viruses among midges. Apart from Isiolo_1, Turkana_10 and Turkana_5, all the other detected viruses had a significantly low similarity thresholds to previously identified viruses. Phylogenetic analysis of these viruses suggest a classification that may go beyond species level for some of the viruses. Nonetheless, the genomic architectures of all the identified viruses are consistent with previously characterized genomes of the viruses in the families where these newly identified viruses fall into.

The study also managed to characterize several vectors that are possibly associated with the identified viruses. The diversity of vectors identified in the 5 different sites was quite variable. Isiolo had the most diverse vectors identified while Kacheliba and Budalangi had the least diverse number of vectors. Successful application of metagenomics and metabarcoding to simultaneously characterize viruses and vectors, respectively, highlights the potential for the application of these two methods in high-throughput surveillance of vectors and viruses. This highlights great promise in the early detection of pathogenic viruses and any invasive vector species, which is very useful in prevention of disease outbreaks before they can occur.

5.3 Recommendations

1. There is need for more research into viruses circulating among midges in Kenya. Studies, especially those focussed on virus isolation, should be carried out in order to determine if the viruses that were identified in this study have the ability to replicate in mammalian cells. This information would be greatly useful in giving us an indication on whether the viruses that have been identified have any pathogenic potential. Nonetheless, similar studies would also lead to possible identification of other novel viruses that would greatly improve our understanding of the diversity of viruses circulating among midges and other related vectors.
2. The metagenomics approach that has been applied in characterizing the different viruses identified in this study can be improved. It is recommend that future studies using similar methods can further improve on this by using

viral enrichment strategies in order to identify rare strains. Some of these methods, such as the bead-based purification strategies can be harnessed in order to only target potentially pathogenic viruses and those with similarities to viruses known to be circulating in a given locality.

3. Studies aimed at improving the metabarcoding approach used in this study should also be carried out. More specifically, future studies should focus on improving the accuracy and reliability of this method. Such studies should aim at reducing primer biases as well as improving the reliability of the barcoding targets, in order to ensure a more accurate identification of the vectors as well as determining their relative abundance. Further, studies to generate more genetic data of the different variety of vector species are encouraged in order to get more inclusive databases that will aid barcoding studies in the future.

REFERENCES

Aguiar, E. R., Olmo, R. P., Paro, S., Ferreira, F. V., de Faria, I. J., Todjro, Y. M., . . . Marques, J. T. (2015). Sequence-independent characterization of viruses based on the pattern of viral small RNAs produced by the host. *43*(13), 6191-6206. doi: 10.1093/nar/gkv587

Artimo, P., Jonnalagedda, M., Arnold, K., Baratin, D., Csardi, G., De Castro, E., . . . Gasteiger, E. (2012). ExPASy: SIB bioinformatics resource portal. *Nucleic acids research, 40*(W1), W597-W603. doi: 10.1093/nar/gks400

Barzon, L., Lavezzo, E., Militello, V., Toppo, S., & Palù, G. (2011). Applications of next-generation sequencing technologies to diagnostic virology. *Int J Mol Sci, 12*(11), 7861-7884. doi: 10.3390/ijms12117861

Besansky, N. J., Severson, D. W., & Ferdig, M. T. (2003). DNA barcoding of parasites and invertebrate disease vectors: what you don't know can hurt you. *Trends in parasitology, 19*(12), 545-546. doi: 10.1016/j.pt.2003.09.015

Bishop-Lilly, K. A., Turell, M. J., Willner, K. M., Butani, A., Nolan, N. M., Lentz, S. M., . . . Read, T. D. (2010). Arbovirus detection in insect vectors by rapid, high-throughput pyrosequencing. *PLoS Negl Trop Dis, 4*(11), e878. doi: 10.1371/journal.pntd.0000878

Blackburn, N. K. S. L., & Phelps, R. J. (1985). Viruses isolated from *Culicoides* (Diptera: *Ceratopogonidae*) caught at the veterinary research farm, Mazowe, Zimbabwe. *Journal of the Entomological Society of Southern Africa, 48*(2), 331-336. doi: 10.10520/AJA00128789_2789

Campbell, M. M., & Kettle, D. S. (1975). Oogenesis in *Culicoides* Brevitarsis Kieffer (Diptera: *Ceratopogonidae*) and the Development of a Plastron-Like Layer on the Egg. *Australian Journal of Zoology, 23*(2), 203-218. doi: 10.1071/ZO9750203

Capella-Gutiérrez, S., Silla-Martínez, J. M., & Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics, 25*(15), 1972-1973. doi: 10.1093/bioinformatics/btp348

Carpenter, S., Groschup, M. H., Garros, C., Felippe-Bauer, M. L., & Purse, B. V. (2013). *Culicoides* biting midges, arboviruses and public health in Europe. *Antiviral Res, 100*(1), 102-113. doi: 10.1016/j.antiviral.2013.07.020

Clarke, L. J., Soubrier, J., Weyrich, L. S., & Cooper, A. (2014). Environmental metabarcodes for insects: in silico PCR reveals potential for taxonomic bias. *Molecular ecology resources, 14*(6), 1160-1170. doi: 10.1111/1755-0998.12265

Connelly, C. R. (2013). Biting midges, no-see-ums *Culicoides* spp. (Insecta: Diptera: *Ceratopogonidae*): EENY349/IN626, 4/2013. *EDIS, 2013*(5). doi: 10.32473/edis-in626-2013

Davies, F. G., Walker, A. R., Ochieng, P., & Shaw, T. (1979). Arboviruses isolated from *Culicoides* midges in Kenya. *J Comp Pathol, 89*(4), 587-595. doi: 10.1016/0021-9975(79)90049-5

De Coster, W., D'Hert, S., Schultz, D. T., Cruts, M., & Van Broeckhoven, C. (2018). NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics, 34*(15), 2666-2669. doi: 10.1093/bioinformatics/bty149

Diniz, J. A., Nunes, M. R., Travassos da Rosa, A. P., Cruz, A. C., de Souza, W., Medeiros, D. B., . . . Vasconcelos, P. F. (2006). Characterization of two new rhabdoviruses isolated from midges (*Culicoides* SPP) in the Brazilian Amazon: proposed members of a new genus, Bracorhabdovirus. *Arch Virol, 151*(12), 2519-2527. doi: 10.1007/s00705-006-0812-1

Elbers, A. R., Koenraadt, C. J., & Meiswinkel, R. (2015). Mosquitoes and *Culicoides* biting midges: vector range and the influence of climate change. *Rev Sci Tech, 34*(1), 123-137. doi: 10.20506/rst.34.1.2349

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*(7), e0130324. doi: 10.1371/journal.pone.0130324

Fall, M., Diarra, M., Fall, A. G., Balenghien, T., Seck, M. T., Bouyer, J., . . . Baldet, T. (2015). *Culicoides* (Diptera: *Ceratopogonidae*) midges, the vectors of African horse sickness virus--a host/vector contact study in the Niayes area of Senegal. *Parasit Vectors, 8*, 39. doi: 10.1186/s13071-014-0624-1

Fauquet, C. M., & Fargette, D. (2005). International Committee on Taxonomy of Viruses and the 3,142 unassigned species. *Virology Journal, 2*(1), 64. doi: 10.1186/1743-422X-2-64

Finkbeiner, S. (2009). Identification and Characterization of Novel Astroviruses. *All Theses and Dissertations (ETDs), 110*. doi: 10.7936/K7Q23X8B

Forterre, P. (2006). The origin of viruses and their possible roles in major evolutionary transitions. *Virus research, 117*(1), 5-16. doi: 10.1016/j.virusres.2006.01.010

Foxi, C., & Delrio, G. (2010). Larval habitats and seasonal abundance of *Culicoides* biting midges found in association with sheep in northern Sardinia, Italy. *Med Vet Entomol, 24*(2), 199-209. doi: 10.1111/j.1365-2915.2010.00861.x

Foxi, C., Satta, G., Puggioni, G., & Ligios, C. (2022). Biting midges (*Ceratopogonidae*, *Culicoides*). In N. Rezaei (Ed.), *Encyclopedia of Infection and Immunity* (pp. 852-873). Oxford: Elsevier.

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

Gibson, J., Shokralla, S., Porter, T. M., King, I., van Konynenburg, S., Janzen, D. H., . . . Hajibabaei, M. (2014). Simultaneous assessment of the macrobiome and microbiome in a bulk sample of tropical arthropods through DNA metasystematics. *Proceedings of the National Academy of Sciences, 111*(22), 8007-8012. doi: 10.1073/pnas.1406468111

Glick, J. I. (1990). *Culicoides* biting midges (Diptera: *Ceratopogonidae*) of Kenya. *J Med Entomol, 27*(2), 85-195. doi: 10.1093/jmedent/27.2.85

González, M., López, S., Mullens, B. A., Baldet, T., & Goldarazena, A. (2013). A survey of *Culicoides* developmental sites on a farm in northern Spain, with a brief review of immature habitats of European species. *Vet Parasitol, 191*(1-2), 81-93. doi: 10.1016/j.vetpar.2012.08.025

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., . . . Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology, 29*(7), 644-652. doi: 10.1038/nbt.1883

Guichard, S., Guis, H., Tran, A., Garros, C., Balenghien, T., & Kriticos, D. J. (2014). Worldwide niche and future potential distribution of *Culicoides imicola*, a major vector of bluetongue and African horse sickness viruses. *PLoS One, 9*(11), e112491. doi: 10.1371/journal.pone.0112491

Hall, M. (2019). Rasusa: Randomly subsample sequencing reads to a specified coverage. doi: 10.5281/zenodo.3731394

Hall, R. A., Bielefeldt-Ohmann, H., McLean, B. J., O'Brien, C. A., Colmant, A. M., Piyasena, T. B., . . . Prow, N. A. (2016). Commensal viruses of mosquitoes: host restriction, transmission, and interaction with arboviral pathogens. *Evolutionary Bioinformatics, 12*, EBO. S40740. doi: 10.4137/EBO.S40740

Heeney, J. L. (2006). Zoonotic viral diseases and the frontier of early diagnosis, control and prevention. *J Intern Med, 260*(5), 399-408. doi: 10.1111/j.1365-2796.2006.01711.x

Hoffmann, B., Scheuch, M., Höper, D., Jungblut, R., Holsteg, M., Schirrmeier, H., . . . Fischer, M. (2012). Novel orthobunyavirus in cattle, Europe, 2011. *Emerg Infect Dis, 18*(3), 469.

Jacquet, S., Huber, K., Pagès, N., Talavera, S., Burgin, L. E., Carpenter, S., . . . Garros, C. (2016). Range expansion of the Bluetongue vector, *Culicoides imicola*, in continental France likely due to rare wind-transport events. *Sci Rep, 6*, 27247. doi: 10.1038/srep27247

Jauniaux, T. P., De Clercq, K. E., Cassart, D. E., Kennedy, S., Vandenbussche, F. E., Vandemeulebroucke, E. L., . . . Coignoul, F. L. (2008). Bluetongue in Eurasian lynx. *Emerg Infect Dis, 14*(9), 1496-1498. doi: 10.3201/eid1409.080434

Kasičová, Z., Schreiberová, A., Kimáková, A., & Kočišová, A. (2021). Blood meal analysis: host-feeding patterns of biting midges (Diptera, *Ceratopogonidae*, *Culicoides* Latreille) in Slovakia. *Parasite, 28*, 58. doi: 10.1051/parasite/2021058

Khamala, C. P. (1971). Ecological distribution of East African *Culicoides* Latreille (Dipt., *Ceratopogonidae*) as shown by light-traps. *Bulletin of entomological Research, 60*(4), 549-557.

Lassen, S. B., Nielsen, S. A., & Kristensen, M. (2012). Identity and diversity of blood meal hosts of biting midges (Diptera: *Ceratopogonidae*: *Culicoides* Latreille) in Denmark. *Parasit Vectors, 5*, 143. doi: 10.1186/1756-3305-5-143

Lauring, A. S., & Andino, R. (2010). Quasispecies theory and the behavior of RNA viruses. *PLoS Pathog, 6*(7), e1001005. doi: 10.1371/journal.ppat.1001005

Leray, M., Ho, S.-L., Lin, I.-J., & Machida, R. J. (2018). MIDORI server: a webserver for taxonomic assignment of unknown metazoan mitochondrial-encoded sequences using a curated database. *Bioinformatics, 34*(21), 3753-3754. doi: 10.1093/bioinformatics/bty454

Liang, G., Gao, X., & Gould, E. A. (2015). Factors responsible for the emergence of arboviruses; strategies, challenges and limitations for their control. *Emerg Microbes Infect, 4*(3), e18. doi: 10.1038/emi.2015.18

Longdon, B., Brockhurst, M. A., Russell, C. A., Welch, J. J., & Jiggins, F. M. (2014). The evolution and genetics of virus host shifts. *PLoS Pathog, 10*(11), e1004395. doi: 10.1371/journal.ppat.1004395

Lu, G., Ye, Z. X., He, Y. J., Zhang, Y., Wang, X., Huang, H. J., . . . Li, J. M. (2020). Discovery of Two Novel Negeviruses in a Dungfly Collected from the Arctic. *Viruses, 12*(7). doi: 10.3390/v12070692

Ma, Y., Xu, J., Yang, Z., Wang, X., Lin, Z., Zhao, W., . . . Shi, H. (2013). A video clip of the biting midge *Culicoides* anophelis ingesting blood from an engorged Anopheles mosquito in Hainan, China. *Parasites & Vectors, 6*(1), 326. doi: 10.1186/1756-3305-6-326

Maclachlan, N. J., & Guthrie, A. J. (2010). Re-emergence of bluetongue, African horse sickness, and other orbivirus diseases. *Vet Res, 41*(6), 35. doi: 10.1051/vetres/2010007

Martínez-de la Puente, J., Martínez, J., Ferraguti, M., Morales-de la Nuez, A., Castro, N., & Figuerola, J. (2012). Genetic characterization and molecular identification of the bloodmeal sources of the potential bluetongue vector *Culicoides* obsoletus in the Canary Islands, Spain. *Parasit Vectors, 5*, 147. doi: 10.1186/1756-3305-5-147

Mellor, P. S., Boorman, J., & Baylis, M. (2000). *Culicoides* biting midges: their role as arbovirus vectors. *Annu Rev Entomol, 45*, 307-340. doi: 10.1146/annurev.ento.45.1.307

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, Gustavo A., Sonnhammer, E. L. L., . . . Bateman, A. (2020). Pfam: The protein families database in 2021. *Nucleic acids research, 49*(D1), D412-D419. doi: 10.1093/nar/gkaa913

Mullen, G. R., & Murphree, C. S. (2019). Biting Midges (*Ceratopogonidae*). In G. R. Mullen & L. A. Durden (Eds.), *Medical and Veterinary Entomology (Third Edition)* (pp. 213-236): Academic Press.

Nguyen, L.-T., Schmidt, H. A., Von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular biology and evolution, 32*(1), 268-274. doi: 10.1093/molbev/msu300

Norris, D. E. (2002). Genetic markers for study of the anopheline vectors of human malaria. *Int J Parasitol, 32*(13), 1607-1615. doi: 10.1016/s0020-7519(02)00189-3

Öhlund, P., Lundén, H., & Blomström, A.-L. (2019). Insect-specific virus evolution and potential effects on vector competence. *Virus Genes, 55*(2), 127-137. doi: 10.1007/s11262-018-01629-9

OIE. (2018). Recent outbreaks and epidemiological events in Africa. Retrieved September, 12, 2021, from https://rr-africa.oie.int/en/immediate-notifications-in-africa/

Parvez, M. K., & Parveen, S. (2017). Evolution and emergence of pathogenic viruses: past, present, and future. *Intervirology, 60*(1-2), 1-7. doi: 10.1159/000478729

Pascall, D. J., Tinsley, M. C., Obbard, D. J., & Wilfert, L. (2019). Host evolutionary history predicts virus prevalence across bumblebee species. *bioRxiv*, 498717. doi: 10.1101/498717

Paweska, J. T., Venter, G. J., & Mellor, P. S. (2002). Vector competence of South African *Culicoides* species for bluetongue virus serotype 1 (BTV-1) with special reference to the effect of temperature on the rate of virus replication in *C. imicola* and C. bolitinos. *Med Vet Entomol, 16*(1), 10-21. doi: 10.1046/j.1365-2915.2002.00334.x

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., . . . Glöckner, F. O. (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic acids research, 41*(D1), D590-D596. doi: 10.1093/nar/gks1219

Radford, A. D., Chapman, D., Dixon, L., Chantrey, J., Darby, A. C., & Hall, N. (2012). Application of next-generation sequencing technologies in virology. *J Gen Virol, 93*(Pt 9), 1853-1868. doi: 10.1099/vir.0.043182-0

Ramírez, A. L., Colmant, A. M., Warrilow, D., Huang, B., Pyke, A. T., McMahon, J. L., . . . Ritchie, S. A. (2020). Metagenomic analysis of the virome of mosquito excreta. *Msphere, 5*(5).

Robin, M., Page, P., Archer, D., & Baylis, M. (2016). African horse sickness: The potential for an outbreak in disease-free regions and current disease control and elimination techniques. *Equine Vet J, 48*(5), 659-669. doi: 10.1111/evj.12600

Romero-Alvarez, D., & Escobar, L. E. (2017). Vegetation loss and the 2016 Oropouche fever outbreak in Peru. *Mem Inst Oswaldo Cruz, 112*(4), 292-298. doi: 10.1590/0074-02760160415

Sadeghi, M., Altan, E., Deng, X., Barker, C. M., Fang, Y., Coffey, L. L., & Delwart, E. (2018). Virome of> 12 thousand *Culex* mosquitoes from throughout California. *Virology, 523*, 74-88. doi: 10.1016/j.virol.2018.07.029

Sahlin, K., & Medvedev, P. (2020). De Novo Clustering of Long-Read Transcriptome Data Using a Greedy, Quality Value-Based Algorithm. *Journal of Computational Biology, 27*(4), 472-484. doi: 10.1089/cmb.2019.0299

Sahlin, K., & Medvedev, P. (2021). Error correction enables use of Oxford Nanopore technology for reference-free transcriptome analysis. *Nature Communications, 12*(1), 2. doi: 10.1038/s41467-020-20340-8

Sahlin, K., Tomaszkiewicz, M., Makova, K. D., & Medvedev, P. (2018). Deciphering highly similar multigene family transcripts from Iso-Seq data with IsoCon. *Nature Communications, 9*(1), 4601. doi: 10.1038/s41467-018-06910-x

Slama, D., Haouas, N., Mezhoud, H., Babba, H., & Chaker, E. (2015). Blood Meal Analysis of *Culicoides* (Diptera: *Ceratopogonidae*) in Central Tunisia. *PLoS One, 10*(3), e0120528. doi: 10.1371/journal.pone.0120528

Smith, G., Macias-Muñoz, A., & Briscoe, A. D. (2014). Genome Sequence of a Novel Iflavirus from mRNA Sequencing of the Butterfly Heliconius erato. *Genome Announc, 2*(3). doi: 10.1128/genomeA.00398-14

Starr, E. P., Nuccio, E. E., Pett-Ridge, J., Banfield, J. F., & Firestone, M. K. (2019). Metatranscriptomic reconstruction reveals RNA viruses with the potential to

shape carbon cycling in soil. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(51), 25900-25908. doi: 10.1073/pnas.1908291116

Szadziewski, R., Krzywinski, J., & Giłka, W. (1997). Diptera *Ceratopogonidae*, Biting Midges. In A. N. Nilsson (Ed.), *Apollo Books, Stenstrup,* 2, 243-263.

Tabachnick, W. J. (1996). *Culicoides* variipennis and bluetongue-virus epidemiology in the United States. *Annu Rev Entomol, 41*, 23-43. doi: 10.1146/annurev.en.41.010196.000323

Temmam, S., Monteil-Bouchard, S., Robert, C., Baudoin, J. P., Sambou, M., Aubadie-Ladrix, M., . . . Desnues, C. (2016). Characterization of Viral Communities of Biting Midges and Identification of Novel Thogotovirus Species and Rhabdovirus Genus. *Viruses, 8*(3), 77. doi: 10.3390/v8030077

The Standard. (2018). Is your herd at risk of Blue Tongue disease? Retrieved from https://www.standardmedia.co.ke/farmkenya/article/2001294154/is-your-herd-at-risk-of-blue-tongue-disease

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. *Parasites & Vectors, 13*(1), 79. doi: 10.1186/s13071-020-3938-1

Toye, P. G., Batten, C. A., Kiara, H., Henstock, M. R., Edwards, L., Thumbi, S., . . . Oura, C. A. (2013). Bluetongue and epizootic haemorrhagic disease virus in local breeds of cattle in Kenya. *Res Vet Sci, 94*(3), 769-773. doi: 10.1016/j.rvsc.2012.11.001

Travassos da Rosa, J. F., de Souza, W. M., Pinheiro, F. P., Figueiredo, M. L., Cardoso, J. F., Acrani, G. O., & Nunes, M. R. T. (2017). Oropouche Virus: Clinical, Epidemiological, and Molecular Aspects of a Neglected Orthobunyavirus. *Am J Trop Med Hyg, 96*(5), 1019-1030. doi: 10.4269/ajtmh.16-0672

Vainio, E. J., Chiba, S., Ghabrial, S. A., Maiss, E., Roossinck, M., Sabanadzovic, S., . . . Nibert, M. (2018). ICTV virus taxonomy profile: Partitiviridae. *Journal of General Virology, 99*(1), 17-18. doi: 10.1099/jgv.0.000985

van Eeden, C., Williams, J. H., Gerdes, T. G., van Wilpe, E., Viljoen, A., Swanepoel, R., & Venter, M. (2012). Shuni virus as cause of neurologic disease in horses. *Emerg Infect Dis, 18*(2), 318-321. doi: 10.3201/eid1802.111403

Vasconcelos, H. B., Azevedo, R. S., Casseb, S. M., Nunes-Neto, J. P., Chiang, J. O., Cantuária, P. C., . . . Vasconcelos, P. F. (2009). Oropouche fever epidemic in Northern Brazil: epidemiology and molecular characterization of isolates. *J Clin Virol, 44*(2), 129-133. doi: 10.1016/j.jcv.2008.11.006

Vasconcelos, H. B., Nunes, M. R., Casseb, L. M., Carvalho, V. L., Pinto da Silva, E. V., Silva, M., . . . Vasconcelos, P. F. (2011). Molecular epidemiology of Oropouche virus, Brazil. *Emerg Infect Dis, 17*(5), 800-806. doi: 10.3201/eid1705.101333

Vasilakis, N., & Tesh, R. B. (2015). Insect-specific viruses and their potential impact on arbovirus transmission. *Current opinion in virology, 15*, 69-74. doi: 10.1016/j.coviro.2015.08.007

Wajid, B., & Serpedin, E. (2014). Do it yourself guide to genome assembly. *Briefings in Functional Genomics, 15*(1), 1-9. doi: 10.1093/bfgp/elu042

Walker, A. (1977). Adult lifespan and reproductive status of *Culicoides* species (Diptera: Cerato-pogonidae) in Kenya, with reference to virus transmission. *Bulletin of entomological Research, 67*(2), 205-215. doi: 10.1017/S0007485300011020

Walker, A., & Boreham, P. (1976). Blood feeding of *Culicoides* (Diptera, *Ceratopogonidae*) in Kenya in relation to the epidemiology of bluetongue and ephemeral fever. *Bulletin of entomological Research, 66*(1), 181-188. doi: 10.1017/S000748530000660X

Wang, Q., Garrity, G. M., Tiedje, J. M., & Cole, J. R. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial

taxonomy. *Applied and environmental microbiology, 73*(16), 5261-5267. doi: 10.1128/AEM.00062-07

Weaver, S. C. (2013). Urbanization and geographic expansion of zoonotic arboviral diseases: mechanisms and potential strategies for prevention. *Trends Microbiol, 21*(8), 360-363. doi: 10.1016/j.tim.2013.03.003

Weaver, S. C., & Reisen, W. K. (2010). Present and future arboviral threats. *Antiviral Res, 85*(2), 328-345. doi: 10.1016/j.antiviral.2009.10.008

Whistler, T., Swanepoel, R., & Erasmus, B. J. (1989). Characterization of Palyam serogroup orbiviruses isolated in South Africa and serologic evidence for their widespread distribution in the country. *Epidemiol Infect, 102*(2), 317-324. doi: 10.1017/s095026880002999x

Wylie, K. M., Mihindukulasuriya, K. A., Sodergren, E., Weinstock, G. M., & Storch, G. A. (2012). Sequence analysis of the human virome in febrile and afebrile children. *PLoS One, 7*(6), e27735. doi: 10.1371/journal.pone.0027735

Zhang, J., Kobert, K., Flouri, T., & Stamatakis, A. (2014). PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics, 30*(5), 614-620. doi: 10.1093/bioinformatics/btt593

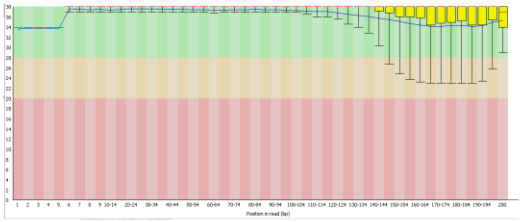
Zhou, X., Adamowicz, S. J., Jacobus, L. M., DeWalt, R. E., & Hebert, P. D. (2009). Towards a comprehensive barcode library for arctic life-Ephemeroptera, Plecoptera, and Trichoptera of Churchill, Manitoba, Canada. *Frontiers in zoology, 6*(1), 1-9. doi: 10.1186/1742-9994-6-30

Zientara, S., Weyer, C. T., & Lecollinet, S. (2015). African horse sickness. *Rev Sci Tech, 34*(2), 315-327. doi: 10.20506/rst.34.2.2359

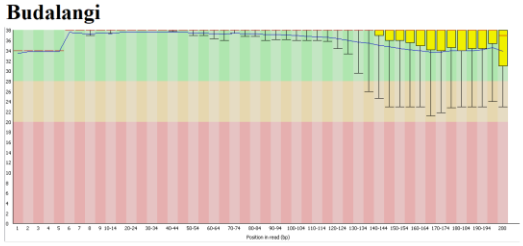
APPENDICES

Appendix I: Quality score across all bases of the viral sequence reads

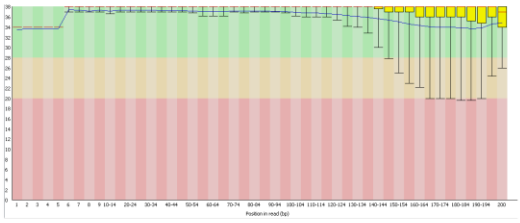
Baringo



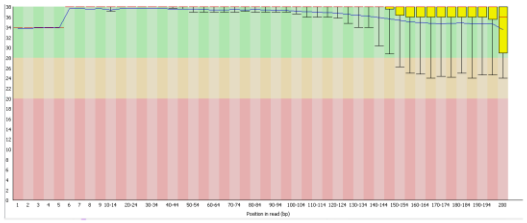
Budalangi



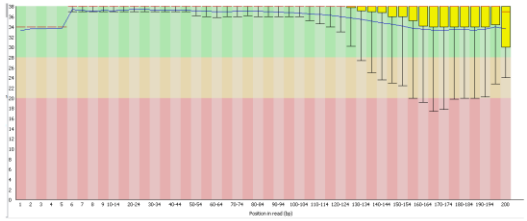
Turkana



Isiolo

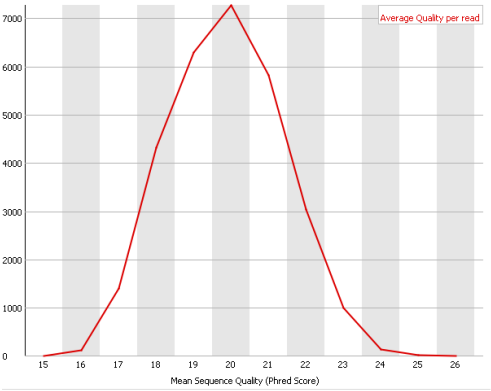


Kacheliba

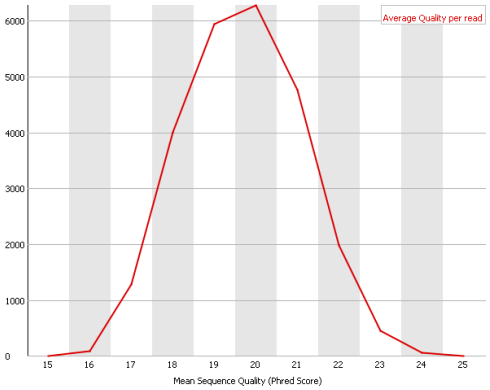


Appendix II: Quality score distribution across the vector sequence reads

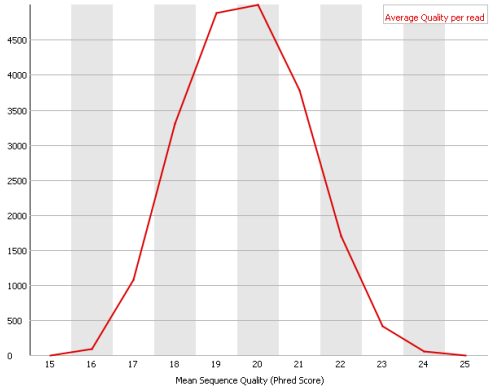
ISIOLO



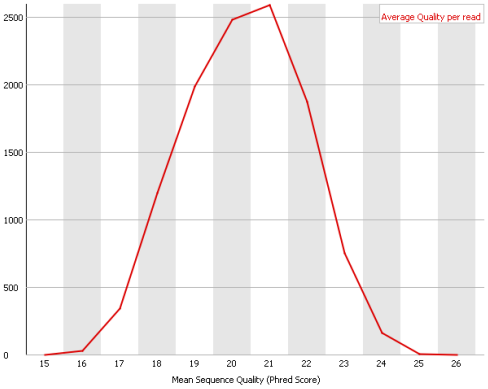
Baringo



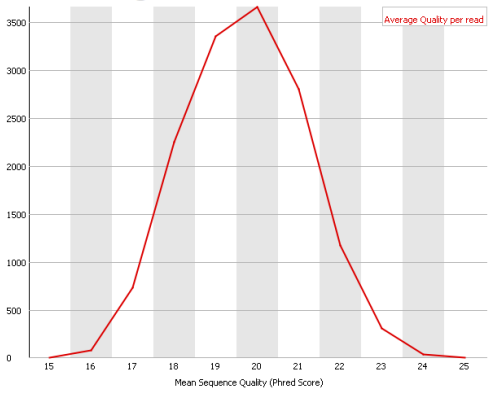
Turkana



Kacheliba



Budalangi



Appendix III: Gel photo of the vector COI amplicons

LANES:

1 – 100bp ladder
2 – No-template control
3 – Sample 1 (Isiolo)
4 – Sample 2 (Baringo)
5 – Sample 3 (Turkana)
6 – Sample 4 (Kacheliba)
7 – Sample 5 (Budalangi)
8 – Empty well
9 – Positive control

Appendix IV: Ethical approval of the study

KENYA MEDICAL RESEARCH INSTITUTE

P.O. Box 54840-00200, NAIROBI, Kenya
Tel: (254) (020) 2722541, 2713349, 0722-205901, 0733-400003, Fax: (254) (020) 2720030
E-mail: director@kemri.org, info@kemri.org, Website. www.kemri.org

KEMRI/RES/7/3/1 **July 10, 2018**

TO: **LANGAT KIPNGETICH SOLOMON,**
PRINCIPAL INVESTIGATOR

THROUGH: **THE DIRECTOR, CVR,**
NAIROBI

FOR DIRECTOR
CENTRE FOR VIRUS RESEARCH
P.O. BOX 54628
NAIROBI.
13/07/18

Dear Sir,

RE: **KEMRI/SERU/CVR/017/3693 (*RESUBMISSION OF INITIAL SUBMISSION*)**
CHARACTERIZATION OF ARBOVIRUSES ISOLATED FROM BITING MIDGES (DIPTERA: CERATOPOGONIDAE) IN KENYA

Reference is made to your letter dated June 29, 2018. The KEMRI Scientific and Ethics Review Unit (SERU) acknowledges receipt of the revised study documents on July 04, 2018.

This is to inform you that the Committee notes that the issues raised during the 276th ERC meeting of the KEMRI Scientific Ethics Review Unit (SERU) held on June 19, 2018, are adequately addressed.

Consequently, the study is granted approval for implementation effective this day, **July 10, 2018** for a period of **one (1) year**. Please note that authorization to conduct this study will automatically expire on **July 10, 2019.** If you plan to continue with data collection or analysis beyond this date, please submit an application for continuation approval to SERU by **May 28, 2019**.

You are required to submit any proposed changes to this study to SERU for review and the changes should not be initiated until written approval from SERU is received. Please note that any unanticipated problems resulting from the implementation of this study should be brought to the attention of SERU and you should advise SERU when the study is completed or discontinued.

You may embark on the study.

Yours faithfully,

FOR: THE HEAD,
KEMRI SCIENTIFIC AND ETHICS REVIEW UNIT

12 JUL 2018

Appendix V: Publication

AMERICAN SOCIETY FOR MICROBIOLOGY | mSphere®

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