Molecular Identification, Phylogeography and Genetic Diversity of Mosquitoes Collected from Areas Endemic to Rift Valley Fever in Kenya

Cecilia Njeri Rumberia

A thesis submitted in partial fulfilment for the Degree of Master of Science in Genetics in the Jomo Kenyatta University of Agriculture and Technology

2013

DECLARATION

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

Signature…… Date…….

Cecilia Njeri Rumberia

This thesis has been submitted for examination with our approval as the University Supervisors.

Signature…… Date…….

Prof. Anne W Muigai. JKUAT, Kenya

Signature…… Date…….

Prof. Steve Kemp. ILRI, Kenya

Signature…… Date…….

Dr. Ilma Tapio. Agrifood Research Finland MTT, Finland

Signature…… Date…….

Dr. Shadrack Muya. JKUAT, Kenya

DEDICATION

This thesis is dedicated to my father Dr. Rufus M Rumberia, my mother Mrs. J.M Rumberia and my siblings Polly, Martin, Beatrice and Antony who have been there for me throughout the period of this study. It is also dedicated to Kevin Chege, who taught me that even the largest task can be accomplished if it is done one step at a time.

ACKNOWLEDGEMENT

This research project would not have been possible without the support of many people. I wish to express my gratitude to my supervisors, Prof. Anne Muigai and Dr. Shadrack Muya who were abundantly helpful and offered invaluable assistance, support and guidance. Deepest gratitude are also due to Prof. Steve Kemp, Dr. Ilma Tapio and Dr. George Michuki, without whose knowledge and assistance this study would not have been successful.

Special thanks also to Dr. Harry Noyes and all my colleagues at ILRI, especially members of the AVID team for sharing literature and invaluable assistance. I would also like to convey thanks to ILRI for providing the financial support and laboratory facilities.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF PLATES

LIST OF TABLES

LIST OF APPENDICES

LIST OF ABBREVIATIONS

CCD camera charge-coupled device camera

CWF Composite Well Format

DNA Deoxy- ribonucleic acid

dsDNA Double stranded DNA

EDTA Ethylene diamine - tetraacetic acid

emPCR Emulsion PCR

GSFLX Genome Sequencer FLX Geographical Positioning System (GPS)

icipe International Centre for Insect Physiology and Ecology

ILRI International Livestock Research Institute

ITS Internally Transcribed Spacer

MLST Multiple Loci Sequence Typing

MPC Magnetic Particle concentrator

PCR Polymerase Chain Reaction

PCR - RFLP Polymerase chain reaction - Restriction Fragment Length polymorphism

PTP Pico- titre plate

rDNA ribosomal Deoxyribo- Nucleic Acid

RVF Rift Valley Fever

ABSTRACT

Rift Valley Fever virus is primarily transmitted by *Aedes* and *Culex* mosquitoes. The disease is zoonotic and is endemic in specific areas in Kenya. For a long time RVF mosquito vectors have been classified using morphological characteristics which have been found to be at times subjective and not very effective in classification. Molecular tools in use for mosquito classification such as Polymerase Chain Reaction- Restriction Fragment Length Polymorphism (PCR- RFLP) have also been reported to be cumbersome, thus more reliable and efficient tools are needed. In this study, validation of internal transcribed spacer (ITS) 1 and 2 as useful tools for molecular classification of mosquitoes was done and new molecular tools targeting intergenic/intronic loci were designed and tested for applicability in vector identification. The study was done using the 454 next generation sequencing of laboratory reared mosquitoes and mosquitos collected from different RVF endemic regions in Kenya. The ITS 1 region was highly divergent displaying a high degree of intraspecific and interspecific variation while the ITS 2 region was found to be highly conserved in the different species. These two loci would therefore not be appropriate tools for taxonomic and phylogeographic analysis of the vector populations. Three loci (ANG12432, ANG26425 and ANG20760) were found to be conserved within distinct genera with variation existing between genera making them appropriate for classification and accurate identification of mosquito species. In the study of population structure, none of the eleven sites used revealed distinct geographical distribution. ANG00020 loci separated samples obtained from the insectary and field samples suggesting applicability in distinguishing between laboratory reared and field collected mosquitoes, this observation requires validation.

CHAPTER ONE

INTRODUCTION

1.1 General introduction

Rift Valley Fever (RVF) is a contagious, zoonotic viral disease which is transmitted by mosquitoes. Disease outbreaks are associated with heavy rainfall (Woods *et al*., 2002; Hassan *et al*., 2011). Different mosquito species have been identified as RVF virus transmitters with *Aedes* and *Culex* genera being reported to maintain and amplify the disease, while other mosquito species are reported to also participate in transmission (Moutailler *et al*., 2008). Outbreaks of RVF mainly occur as an epizootic resulting in high mortalities and morbidities leading to huge losses to small holder livestock farmers (Andriamandimby *et al*., 2010). The RVF occurs as an acute febrile disease that severely affects sheep, cattle and goats, resulting in an abortion rate of 80 - 90% and high death rates in neonates. The adult livestock mortality rates are generally low at less than 10% of the herd. In humans RVF virus infection causes a severe influenza-like illness, with occasionally more serious haemorrhagic complications and death (Gerdes, 2008). Human infections mostly arise from contact with infected aborted foetuses and carcasses of dead animals. It is also suspected that transmission to humans can occur through mosquito bites (Pepin *et al*., 2010).

The first RVF outbreak was reported in Kenya in the 1930s in Naivasha area. Later, other cases were reported in Egypt in 1977 to 1978 (Imam *et al*., 1979); in the Senegal river basin in 1987 (Scott *et al*., 1992) and again in Kenya in 1997-1998

(Woods *et al*., 2002). RVF was for some time thought to be restricted to Africa (Rweyemamu *et al*., 2000) until the 2000 to 2001 outbreak in Saudi Arabia and Yemen (Shoemaker *et al*., 2002). This indicated a high potential of RVF to spread to countries and areas neighbouring endemic regions. More recently RVF cases were reported in Sudan 2007 (Hassan *et al*., 2011), Madagascar 2008 - 2009 (Andriamandimby *et al*., 2010), and South Africa (Vuren *et al*., 2010).

In a study to identify the potential vectors of RVF virus in the Mediterranean region it was reported that *Aedes* mosquitoes serve as the virus reservoir during inter-epidemic periods while *Culex pipiens* are the most potent transmitters to both humans and animals (Moutailler *et al*., 2008). In view of the critical role of mosquitoes in transmission and sustenance of RVF and other diseases, it is important to understand the role of the vector and to be able to accurately identify vector species. RVF virus vector diversity studies provide epidemiological and experimental tools to understand the actual and potential risk that dynamic vector populations pose. In addition, there is a possibility that developed technologies will have application in other arthropod vectors that are important disease transmitters but are poorly understood.

1.2 Problem statement

Morphological characterization of mosquitoes relies on a set of physical features found unique within specific species. The morphological methods of classification have been used over a long time and have been observed to have various limitations. One of the limitations of this classification method occurs where the set standard

keys and features have been found to be similar between closely related species, making it difficult or practically impossible to differentiate up to the sub species level. Morphological classification has also been found to be subjective and introduces individual bias rendering it rather inefficient and un-reliable apart from being tedious, time consuming and cumbersome. The first molecular approach for mosquito classification was a PCR- based method, where amplified fragments of ribosomal DNA were analysed using electrophoresis, thus distinguishing species of the *Anopheles* complex by fragment length variations. This method provided improved accuracy compared to morphological classification. It was however limited by the fact that the base length variations are in many instances minute and difficult to accurately visualize on a gel.

1.3 Justification

Efficient, accurate and reliable method to identify of mosquito vectors is still absent, while correct identification is a critical aspect of disease surveillance. Correct identification of the insect vector is one of the important factors in the study of the arboviral diseases. In addition, the precise identification of the target species has direct medical and practical implications, particularly in developing vector control strategies. In an effort to improve the accuracy of classification of mosquitoes, PCR based methods have been developed. Primers designed for classification purposes possess the characteristic of being specific to a single species or discriminate between two or more species within the same genus, making their scope of application narrow.

This study was carried ot to develop a good molecular vector identification method. Genomic regions with low variation areas conserved among different mosquito species suitable for primer design and a variable site in between the conserved sites for different mosquito species were targeted. ITS one and two regions were targeted, as well as intronic and intergenic spacers found to carry a region of conservation across species. Proper vector identification will also be useful in studies investigating vectoral capacity of the different mosquito species for different diseases.

1.4 Hypotheses

1. ITS1, ITS2, intronic and intergenic sites are conserved within species and within geographical locations.
2. The ITS1 and ITS2 of the ribosomal DNA are useful for species identification and phylogenetic separation of mosquitoes from different RVF endemic regions in Kenya.
3. Intergenic and intronic regions of genomic DNA are useful for species identification and phylogenetic separation of mosquitoes from different RVF-endemic regions in Kenya.

1.5 Study Objectives

1.5.1 General objective

To determine the molecular phylogeny and genetic diversity of mosquitoes from RVF endemic regions in Kenya using selected genetic markers.

1.5.2 Specific objectives

1. To determine the ability of primers targeting ITS1, ITS2, intronic and intergenic regions to amplify mosquito DNA of different species.
2. To determine if ITS, intergenic and intronic regions can be applied as a tool for specific identification of mosquito species.
3. To determine diversity of mosquitoes collected from different RVF- endemic regions in Kenya using ITS1, ITS2, Intergenic and intronic loci.

CHAPTER TWO

LITERATURE REVIEW

2.1 RVF virus mosquito vectors

Mosquitoes are responsible for the transmission of parasitic and viral infections to both humans and livestock, with substantial morbidity and mortality (Davies, 2006). There are over 3,500 different species of mosquitoes throughout the world (Tolle, 2009). Some diseases transmitted by mosquitoes include malaria, yellow fever, encephalitis and RVF among others (Farajollahi *et al*., 2011). Prevention of infection of mosquito transmitted diseases mainly rely on vector control as well as vaccinations to reduce and/or prevent transmission (Tolle, 2009). Mosquito control is carried out by habitat control, use of insecticides, larvicides and breeding control using sterile males (Jackson *et al*., 1926). Most mosquito transmitted organisms have an obligatory developmental stage that takes place in the mosquito, and in some cases completely rely on the vectors for transmission.

Different mosquito species may be implicated as transmitters of specific diseases where no conclusive information exists on capacity of transmission. In the case of RVF, all mosquito species are implicated as virus transmitters of the virus to both animals and humans (Cognolati *et al*., 2006). Studies have shown that *Aedes* acts as a reservoir to maintain RVF virus in the environment while *Culex* species are the main transmitters hence amplifiers of the disease (Fontenille *et al*., 1998; Moutailler *et al*., 2008). Other mosquito species like *Anopheles* species are also able to transmit the virus (Pages *et al*., 2009). Thus for RVF virus, mosquitoes are a critical component within the life cycle as they maintain the virus in inter-epizootic periods

and serve as the amplifier during outbreaks (Frontielle *et al.,* 1995). As part of the coordinated efforts to reduce or eliminate RVF virus, a better understanding of the mosquito vectors and how to best control them is paramount. Mosquito species identification which is the most basic and primary level of beginning to understand the vector requires to be carried out in an efficient, accurate and reliable manner.

It is therefore of key importance to understand mosquito classification, their distinguishing features, and the insect life cycle for disease surveillance as well as for designing and implementing effective measures for disease control and prevention.

2.1.1 Mosquito classification

The primary identification of mosquitoes has over time been carried out by morphological characterization (Service, 2000). Wing structure and venation, proboscis and other physically visible features are observed to determine species (Theobald, 1901). Mosquitoes are observed under a differentiating microscope and only persons with knowledge of the features that differentiate one species from another have the capacity to carry out morphological identification. An unambiguous identification of mosquitoes using morphological characters requires taxonomic experience and specimens as intact as possible (Hackett *et al.*, 2000).

In addition to the morphological tools, molecular methods for sub species classification have been proposed (Caterino *et al*., 2000). These include PCR-RFLPs which apply differences in the length of DNA fragments after digestion with restriction endonucleases (Walton *et al*., 1999). These differences in length are

caused by variations, or polymorphisms in the DNA sequences. Limitations of RFLP include that a high concentration of DNA is required to run a RFLP gel (Caterino *et al*., 2000). If only a low concentration of DNA is available, RFLP may still be usable if coupled with PCR. PCR can amplify a small quantity of DNA to increase the concentration of the original sample. This technique is however not suitable if DNA is heavily damaged (Caterino *et al*., 2000). Conventional identification methods have limitations for sibling and closely related species of mosquitoes and for the stage and quality of the specimen used. This could be overcome by DNA-based identification methods using molecular markers such as nuclear ribosomal ITS which do not demand intact or undamaged specimen (Dhananjeyan *et al*., 2010). The development of reliable molecular tools for species identification, an understanding of intraspecific genetic diversity and population structure play important roles in the development of vector control strategies (Palumbi and Cipriano, 1998) as well as understanding of disease dynamics.

There is effort towards developing species specific primers that would be able to classify mosquitoes up to a subspecies level. For instance, a PCR- based method of ribosomal DNA was developed by Scott *et al*. (1993) for species identification of five most widespread members of the *Anopheles gambiae* complex. These are a group of morphologically indistinguishable sibling mosquito species that includes the major vectors of malaria in Africa.

2.1.2 Rift Valley Fever

2.1.2.1 The Virus

RVF virus is an arbovirus that belongs to the Bunyaviridae family of enveloped, RNA viruses (Filone *et al*., 2010).

2.1.2.2 The Disease

RVF is a zoonosis that primarily affects animals but also has the capacity to cause potentially severe disease in humans (Pepin *et al.*, 2010). Sheep, goats and cows acquire infection through the bite of infected insects (Musser *et al*., 2006) while humans will not only get infection from mosquito bites, but also from contact with body fluids of infected animals, especially during slaughter (Pages *et al*., 2009). In animals, the infection is characterized by deaths of new-born animals and abortion in pregnant sheep, goats, and cattle. Other animals such as water buffalo, camels, monkeys, rodents, cats, dogs, and horses can also be infected. Severe disease can occur in new-born kittens and puppies (WHO, 2006). RVF tends to affect young animals more severely than mature animals (Davies, 2003). In young animals, signs of infection include fever, failure to eat, weakness, diarrhoea and death. In older animals, infection may cause fever, discharge from nose, weakness, diarrhoea, vomiting, decreased milk production and abortion (Gerdes, 2008). Abortion is often the only sign of RVF infection in mature animals (Bunnels and Murphy, 1961). The presence of an RVF epizootic (where a large number of animals exhibit clinical disease) can lead to an epidemic among humans who are exposed to diseased animals. The disease in humans usually shows development of mildly to moderately

severe febrile illness. However, severe complications, including ocular sequelae, encephalitis, and fatal haemorrhagic disease, occur in some patients (Grobbelaar *et al.*, 2011).

The largest outbreak reported in sheep in Kenya occurred between 1950 - 1951 leading to an estimated 100,000 deaths and 500,000 abortions (Davies *et al*., 1985). In humans, the largest outbreak reported in Kenya occurred between 1997 - 1998 resulting in 89,000 cases of infection and 478 deaths (Woods *et al*., 2002).

Safe, effective vaccines are still not freely available for protecting humans and livestock against the dramatic consequences of this virus (Davies, 2003).

2.1.2.3 Transmission and maintenance of RVF virus

The RVF virus is primarily transmitted by mosquitoes and may also be transmitted by other biting insects that have virus-contaminated mouthparts (Moutailler *et al*., 2008). Large outbreaks termed as epizootic, occur at irregular intervals when heavy rains characterized with flooding favour breeding of mosquito vectors. RVF has been reported to be maintained by the *Aedes* species of mosquitoes during inter-epizootic periods based on epidemiological studies carried out following the 1987 outbreak in Mauritania (Frontielle *et al*., 1995). The mosquito eggs of the infected *Aedes* species are naturally infected with the RVF virus, and the resulting mosquitoes transfer the virus to the livestock on which they feed (Davies *et al*., 1985). The inter-epizootic survival of RVF virus is believed to depend on the transovarial transmission of virus in flood- water *Aedes* mosquitoes (Davies *et al*., 1985). Once

the livestock is infected, other species of mosquitoes can become infected from the animals and can spread the disease. Other species of mosquitoes, including *Culex* have been shown to possess substantial vector competence with regards to RVF virus transmission (Moutailler *et al*., 2008). The *Culex* species has been reported as the major amplifying vector during RVF epidemics, as reported in Egypt (Imam *et al*., 1979), Senegal as well as South Africa (Mcintosh and Russell 1980; Pepin *et al*., 2010; Scott *et al*., 1992).

Humans get the disease if they are exposed to the blood, body fluids, or tissues of infected animals. Direct exposure to infected animals can occur during slaughter or through veterinary and obstetric procedures (Woods *et al*., 2002). Humans also get RVF through bites from infected mosquitoes and possibly other biting insects (Pages *et al*., 2009).

2.1.2.5 Geographical distribution of RVF virus

RVF seems to have first emerged in the middle of the 19th Century but was only identified at the beginning of the 1930s during an outbreak of sudden deaths and abortions among sheep on the shores of Lake Naivasha in the Rift Valley region of Kenya (Daubney *et al*., 1931).

Despite being an arbovirus with a relatively simple but temporally and geographically stable genome (Shope, 1931), this zoonotic virus has already demonstrated a real capacity for emerging in new territories (Table 1) (WHO, 2006).

In September 2000, RVF caused a large outbreak in livestock and humans (Shoemaker *et al*., 2002) in Saudi Arabia and Yemen, marking the first reported occurrence of the disease outside the African continent. In 2008 - 2009 it was detected for the first time in the Archipelago of Comores (Andriamandimby *et al*., 2010), located between Mozambique and Madagascar, on the French Island of Mayotte, raising concerns that it could also extend to Asia and Europe (WHO, 2006).

In Kenya, the incidences of RVF have been detected in the Garissa, Ijara, Tana River, Kilifi, Malindi and Wajir Districts (Woods *et al*., 2002). In these Districts the disease was reported to have caused deaths of people and livestock and several abortions in pregnant animals (Woods *et al*., 2002). During the outbreak of 1997 to 1998 the disease was reported to have spread to some districts neighbouring the affected areas for instance suspected human cases of RVF and widespread livestock abortions in Juba valley districts in Somalia (WHO, 2010).

Table 1: Geographical distribution of reported RVF virus outbreaks

Location	Year of outbreak	Affected organism	Reference
Kenya	1930s; 1997- 1998; 2006- 2007	Humans, Livestock	(Daubney *et al*., 1931); (Woods *et al*., 2002) ; (WHO, 2010)
Tanzania and Somalia	2006- 2007	Humans, Livestock	(WHO, 2010)
Egypt	1977	Humans, Livestock	(Imam *et al*., 1979)
Senegal (West Africa)	1987- 1988	Humans,	(Scott *et al*., 1992)

		Livestock	
South Africa	1975	Humans, Livestock	(Mcintosh & Russell, 1980)
Madagascar	2008- 2009	Humans, Livestock	(Andriamandimby *et al*., 2010)
Yemen	2000- 2001	Humans, Livestock	(Shoemaker *et al*., 2002)
Sudan	2007	Humans, Livestock	(WHO, 2010)
Arabian Peninsula (Saudi Arabia)	2000- 2001	Humans, Livestock	(Shoemaker *et al*., 2002)

2.1.2.6 Disease Impact on economy and farmers livelihoods

This disease is very devastating to farmers as it results in significant reductions in herd sizes and especially as it emerges after enhanced periods of drought (Hughes-Fraire *et al*., 2011). This results in significant economic losses in terms of trade as well as diminished livelihoods. The reduction in herd sizes due to death and abortion among RVF-infected livestock, results in reduction in asset volumes for farmers and interference with a source of income. Livestock farmers, especially pastoral communities thrive on the sale of livestock and livestock products. Thus when RVF outbreak results in death of animals, trade barriers and sanctions; the farmers suffer. Such trade barriers that in the past have been imposed on the horn of Africa continue to limit trade on animal products (Rweyemamu *et al*., 2000).

Several factors including the increasing range of the virus, the high numbers of competent vector species present in currently RVF-free regions, the intensification of

international trade in live animals, and the unknown impact of climate change, have resulted in national and international agencies issuing warnings about the heightened risk of introduction of RVF virus into RVF-free countries (Pepin *et al*., 2010). These reports conclude unanimously that coordinated efforts to better prepare for a possible emergence of RVF virus spread are needed.

RVF virus has in recent years become an important subject of interest particularly as public health agencies have become alerted to the possible emergence of this arbovirus in temperate countries (WHO, 2006). Climate change and the presence of competent vectors in currently RVF-free countries suggest strongly it should be included among the most significant emerging viral threats to public and veterinary health (Shoemaker *et al*., 2002). Insights into the virus' pathogenesis, molecular epidemiology, diagnostics as well as vector biology will therefore contribute greatly to the understanding of this significant viral pathogen.

2.1.3 Use of molecular tools for RVF virus mosquito vector identification

Molecular diagnostic tools have been applied widely and have been found reliable and effective in correctly identifying species of various organisms (Blaxter *et al*., 2005). Genomic DNA based molecular methods of species identification are also advantageous as they can be applied to damaged specimens and situations unsuitable for morphological taxonomy. DNA can be extracted from specimens in all developmental stages, of both sexes, fresh, preserved in alcohol, dried or frozen (Marrelli *et al*., 2005).

A number of DNA markers are available for the molecular studies of mosquito species (Caterino *et al*., 2000). These include protein-coding genes, all of the major ribosomal RNA genes (both mitochondrial and nuclear) as well as numerous non-coding regions. Molecular markers such as Internal Transcribed Spacers of ribosomal DNA genes, third domain (D3) of 28S rDNA gene, mitochondrial Cytochrome oxidase C subunit I and II (COI & COII), Cytochrome oxidase B, 16S rRNA gene are helpful in species identification, phylogenetic analyses and other related studies (Dhananjeyan *et al*., 2010). The rDNA ITS regions, applied in this study have previously been used in identification of insect specimens and have proven highly informative for phylogenetic inference. For instance, a study on polymorphisms occurring in the rDNA ITS of *Anopheles farauti* in populations of the South- west Pacific, showed that the internal transcribed spacers are useful for within-species comparisons, facilitating the identification of distinct genotypes demonstrating a macro-geographical distribution and hybridization boundaries (Beebe *et al*., 2000). These ITS regions have however been scarcely been employed to carry out phylogenetic characterization of mosquito species in Kenya and thus require validation to confirm applicability.

2.1.3.1 Internal transcribed spacers

The rDNA transcriptional unit is tandemly repeated (>100 copies per genome) and separated by a non-transcribed intergenic spacer (IGS) (Gorokhova *et al*., 2002). Each transcribed unit has two internally transcribed spacers; ITS1, which separates

the 18S and the 5.8S rDNA subunits and ITS2, which separates the 5.8S and 28S rDNA subunits (**Figure 1**).

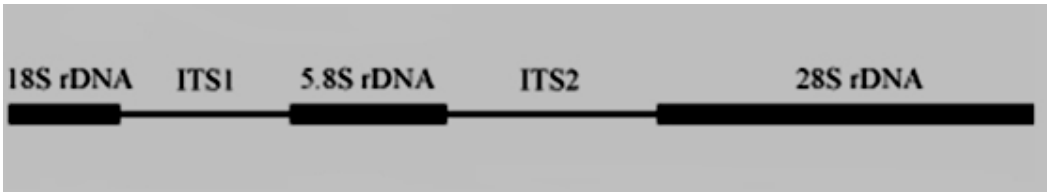

Figure 1: Illustration of ITS1 and ITS2 occurrence in the rDNA (image from archimede.bibl.ulaval.ca- Last visited on 20th July 2012)

A process termed concerted evolution is thought to maintain sequence integrity between the rDNA repeat units through sequence homogenizing mechanisms such as gene conversion and unequal crossover, where recombination occurs between the rDNA repeat units within or between chromosomes (Bower *et al.*, 2008). Some organisms display little intraspecific ITS sequence and length variation while others show high levels of variation. The ITS1 is closely linked to the ITS2 but displays higher levels of sequence variation (Tang *et al*., 1996; Miller *et al*., 1996). The true function of these spacers remains vague, seemingly based on hydrogen-bonded secondary structures which, when modified slightly in conserved regions or modified considerably in variable regions, hinder maturation of the mature rRNA product (van der Sande *et al*., 1992). Analysis of the ITS2 sequences between the cryptic species in the *An. punctulatus* group showed considerable sequence variation between species (2.3% to 24.3%), most of which occurred as insertion/deletion and resulted in considerable differences in the secondary structures (Beebe *et al*., 2000).

CHAPTER THREE

MATERIALS AND METHODS

For this study fifteen (15) selected genomic regions (ITSs, Intergenic and Intronic regions) were targeted and the study simulated a Multiple Loci Sequence Typing (MLST) format for the experiment. MLST was developed by a network of researchers and identifies alleles from the DNA sequences of several housekeeping genes (http://www.mlst.net/). Amplified genomic fragments were sequenced using the Roche 454 sequencer. Sequences were screened for polymorphisms suitable for mosquito species identification and diversity signatures.

3.1 Sampling and study design

The study was carried out in two phases in order to achieve the objectives. First phase involved a pilot study using mosquitoes from cyclic colonies maintained at the International Centre of Insect Physiology and Ecology (*icipe*) insectary. The second phase involved the study of mosquitoes from various regions in Kenya selected based on their participation in RVF outbreaks as documented by (Woods *et al.*, 2002) during the 1997-1998 outbreak.

3.1.1 Pilot study sampling

For the pilot study morphologically identified and clonally bred mosquito samples obtained from an insectary were used. The samples comprised four different species belonging to three genera; *Anopheles gambiae*, *Anopheles gambiae*, *Aedes aegypti* and *Culex quinquefasciatus*. Samples were obtained in different vials per species, transported and stored at 4^{o}C, for short term storage before further processing.

3.1.2 Field study sampling

Field samples were obtained from different selected locations in Kenya, including Baringo, Ijara and Naivasha. These sites were mapped out based on the RVF prevalence in the area. Baringo and Ijara Districts were selected based on their involvement in 1997 - 1998 and 2006 - 2007 RVF outbreaks, while Naivasha is a known RVF endemic area (**Figure 2**).

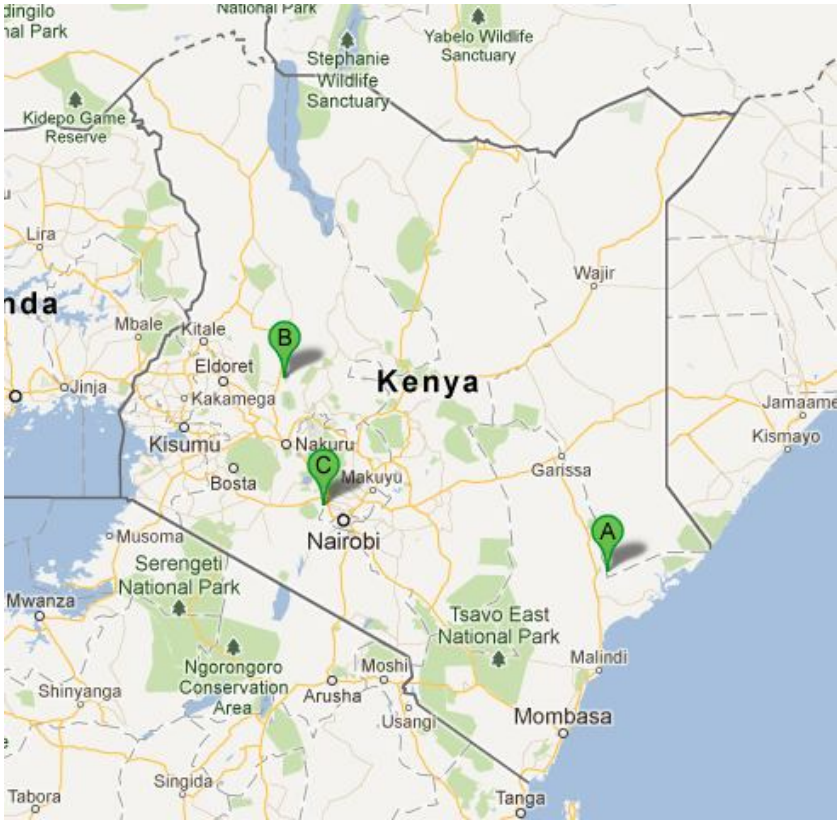

Figure 2: Selected study sites based on their participation in the 2007 RVF outbreak. Key: A- Ijara area in North-eastern Kenya, B- Baringo in Rift-Valley Kenya area and C- Naivasha area in Central Kenya (Image generated with Google maps)

Periods of higher mosquito population densities were targeted. These were found to correspond with rainy seasons (March - May and October - November). Carbon baited light trapping method was used (**Figure 3**). The vectors were collected daily

for ten consecutive days per site. Sampling sites were geo-referenced using Geographical Positioning System (**Table 2**) at the time of sampling. The CO_2 light traps were hung at least one meter above the ground on a tree or pole between 6pm and 7pm in the evening and left overnight. The collection bags containing the mosquitoes were picked between 6am and 6.30am in the morning and taken to the temporary laboratory created in the field for entomological classification.

Figure 3: CDC carbon dioxide baited trap

The mosquitoes were identified under a dissecting microscope, on ice or chill tables and then sorted by species, sex and collection site into cryogenic vials and preserved in a liquid nitrogen shipper or a box containing dry ice.

Table 2: Location, date collected and GPS coordinates of the sampling sites

Species	Site	GPS information
Ijara		
Aedes	Jalish	Longitude 40.50; Latitude -1.67
	Kotile central	Longitude 40.21; Latitude -1.96

Anopheles species	Jalish	Longitude 40.50; Latitude -1.67
	Wakabhare	Longitude 40.699; Latitude -1.30
Baringo		
Aedes	Salabani	Longitude 36.04; Latitude 0.54
Anopheles species	Ng'ambo	Longitude 36.06; Latitude 0.50
	Salabani	Longitude 36.04; Latitude 0.54
Culex pipiens	Chelaba	Longitude 36.05; Latitude 0.35
Naivasha		
Aedes	Maai Mahiu	Longitude 36.53; Latitude -1.10
Anopheles species	Maai Mahiu	Longitude 36.53; Latitude -1.10
Culex pipiens	Olsuswa farm	Longitude 36.28; Latitude -0.71
ICIPE		

Aedes, *Anopheles gambiae*, *Anopheles arabiensis*, *Culex pipiens*, *Culex quinquefasciatus*

Separation of mosquitoes with engorged abdomens and those with non-engorged abdomens was done to separate those that had previously fed on host from those that had not fed prior to capture. For this study only unfed mosquitoes were utilized to avoid DNA contamination with the host. The collected mosquitoes were then transported back to the research facility at the International Livestock Research Institute (ILRI), where they underwent a second stage of morphological identification under a dissecting microscope to confirm species identification. They were then stored at -80^{o}C for long-term storage and 4^{o}C for short term storage.

3.1.3 Laboratory experiments

For the pilot study, DNA was extracted separately for singleton mosquitoes from each species of *Anopheles gambiae*, *Anopheles arabiensis*, *Aedes aegypti* and *Culex quinquefasciatus.* For each species four replicates of singleton isolations were done. DNA was also extracted in pools consisting of five mosquitoes for each species. DNA amplified using fifteen primers (**Table 3**) and sequenced using GS-FLX- 454 sequencer.

Table 3: Primers used in this study

Primer Name	Primer sequence (5'- 3')	Primer specificity	Sequence Locus ID (Ensembl_ID)	Expected fragment length
ITS1 F	TCGTAACAAGGTTTCCGTAGG	ITS1		167
ITS1 R	TTAGCTGCGGTCTTCATCG	ITS 1		167
ITS1 A(Beebe *et al*)	CCTTTGTACACACCGCCCGTCG	ITS 1		530
ITS1 B(Beebe *et al*)	ATGTGTCCTGCAGTTCACA	ITS 1		530
ITS2 F	TGCAGGACACATGAACACC	ITS 2		173
ITS2 R	ATTTAGGGGGTAGTCACACATTATT	ITS 2		173
ITS2 A(Beebe *et al*)	TGTGAACTGCAGGACACAT	ITS2		380
ITS2 B(Beebe *et al*)	TATGCTTAAATTYAGGGGGT	ITS 2		380
ANG00020_F	YGATACSGAAWCSAAGATGG	intergenic	ENSANGT00000000020-ENSANGT00000027199	160
ANG00020_R	CGMACCTTGRCRATTTCTT	intergenic	ENSANGT00000000020-ENSANGT00000027199	160
ANG00026_F	GTMACRATCGARAAGGAYGG	intergenic	ENSANGT00000000026-ENSANGT00000000025.2	188
ANG00026_R	CCCAAGATCCMARRCAYACCC	intergenic	ENSANGT00000000026-ENSANGT00000000025.2	188
ANG04289_F	TCAGTGGAACAAYGTGTATCG	Intronic	ENSANGT00000004289	530
ANG04289_R	GATCCTCCGACAGATCCAAA	Intronic	ENSANGT00000004289	530
ANG12432_F	CCTCGCTCCTCCATGTACCT	Intronic	ENSANGT00000012432	240
ANG12432_R	ATMGGGAAACAGTATCGGCT	Intronic	ENSANGT00000012432	240

ANG13935_F	YTCSGGTTGYTTKATGCG	Intronic	ENSANGT00000013935	205
ANG13935_R	AGGTGTTYCTGTGGYTGGG	Intronic	ENSANGT00000013935	205
ANG18326_F	CATCARCACYTCTCGCTGG	Intronic	ENSANGT00000018326	377
ANG18326_R	ACSGTKACSCAGTTCAATG	Intronic	ENSANGT00000018326	377
ANG20362_F	GCTTCTGKGCRTTGTAGACC	Intronic	ENSANGT00000020362	238
ANG20362_R	ATGTKCTGGAGCTGATGG	Intronic	ENSANGT00000020362	238
ANG20760_F	CRTAGATKACGACGAGGCAC	Intronic	ENSANGT00000020760	305
ANG20760_R	KTCYTGYGAAACGTCCAAG	Intronic	ENSANGT00000020760	305
ANG23972_F	TCKGAGGCTTGMTGTACTKGG	Intronic	ENSANGT00000023972	245
ANG23972_R	ATTCCAGAAGGCGACAAGG	Intronic	ENSANGT00000023972	245
ANG26425_F	GTSGACCGKAAGATWGTGAAAGG	Intronic	ENSANGT00000026425	375
ANG26425_R	TTCCATTTGYTTCTCCTGMG	Intronic	ENSANGT00000026425	375
ANG27523_F	TTCTTCTTCTTCAGMACCTCG	intergenic	ENSANGT00000027523-ENSANGT00000014897	411
ANG27523_R	YTYCGGCRASGACTACACC	intergenic	ENSANGT00000027523-ENSANGT00000014897	411

The laboratory processing pipeline was carried out as illustrated in **Figure 4**.

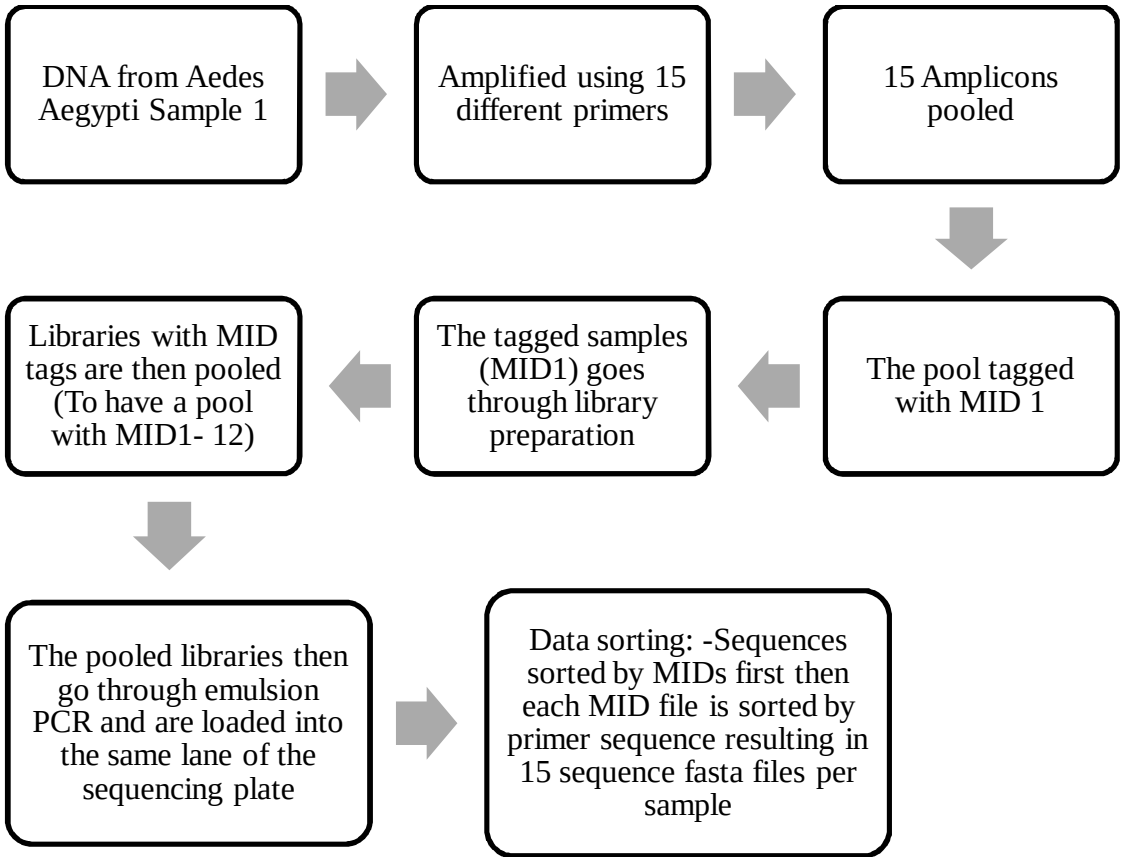

Figure 4: Laboratory processes carried out on the mosquito samples.

The pilot study amplification and sequencing data was analysed first to select the appropriate primers and later select the best statistical approach to analyse sequencing data. Field samples (**Table 4**) were then processed via the same set of experiments using the selected set of primers based on the pilot study results.

Table 4: Laboratory sample setup for field collected mosquitoes

Sample Identity	Number of Mosquitoes per tube	Sample type	Number Replicates	Source
Anopheles gambiae	1	Singleton	4	*icipe*
Anopheles arabiensis	1	Singleton	4	*icipe*
Aedes species	1	Singleton	4	*icipe*
Culex quinquefasciatus	1	Singleton	4	*icipe*
Anopheles species	1	Singleton	4	Ijara
Aedes species	1	Singleton	4	Ijara
Culex species	1	Singleton	4	Ijara
Culex species homogenate	15	homogenate	1	Ijara
Anopheles species	1	Singleton	4	Baringo
Anopheles species homogenate	15	homogenate	1	Baringo
Aedes species homogenate	15	homogenate	1	Baringo
Culex species homogenate	15	homogenate	1	Baringo
Culex species	1	Singleton	4	Naivasha
Anopheles species homogenate	15	homogenate	1	Naivasha
Aedes species homogenate	15	homogenate	1	Naivasha
Culex species homogenate	15	homogenate	1	Naivasha
Culex species	1	Singleton	4	Bodhai

3.2 DNA Extraction from mosquitoes

A whole mosquito (single or more for pooling) was placed into 2.0 ml screw cap tube and 0.1 mm silicon beads were added in (1:1) weight by weight ratio. 200µl of 250mM sodium hydroxide (NaOH) was added and the tube was briefly centrifuged using a bench top centrifuge to spin the contents of the tube down. The tube was then incubated at 100^{o}C for 5 minutes to allow maceration of the tissue. The tissue was crushed by bead beating at 3000 rpm for 5 minutes using a Mini bead beater (Biospec). 480µl of 10mM Tris-HCl (pH 8.5) was added and the tube was briefly centrifuged at room temperature to sediment mosquito particles using the Eppendorf 5424 centrifuge. The supernatant was then transferred to a new tube.

DNA was precipitated by addition of 70µl (1/10 of sample volume) 3M sodium acetate (pH 5.2) followed by addition of three volumes of 95-100% (room temperature) ethanol. The tubes were inverted to mix the content and then centrifuged at 14000 rpm for 20 minutes to pellet the DNA. The supernatant was aspirated carefully to avoid disturbing the DNA pellet. The pellet was washed by addition of 500µl of 70% ethanol (room temp) followed by centrifugation for 2 minutes at 14000 rpm. The supernatant was aspirated and the pellet dried at room temperature for 10 minutes. The pellet was re-suspended in 25µl of Tris HCl buffer. The DNA concentration and purity were determined using a Thermo - Scientific Nanodrop^{TM} 1000 spectrophotometer.

3.3 DNA amplification using Polymerase Chain Reaction

3.3.1 Primer design

Fifteen primer sets were used in this study (**Table 3**). These included; ITS1 F/R and ITS1 A/B for Internal Transcribed Spacer 1 region, ITS2 F/R and ITS2 A/B for Internal Transcribed Spacer 2 region and 11 primers (with a prefix of ANG) targeting different intergenic and intronic regions. ITS1 A/B and ITS2 A/B were selected from previously published literature (Beebe *et al.*, 2000), while ITS1 F/R and ITS2 F/R were designed against the ITS regions of the *Aedes aegypti* whole genome shotgun sequences Liverpool project accession number AAGE01000000.

The ANG primers (Table 3) were designed by first screening the *Anopheles gambiae* genome for all intergenic regions or introns that were 100 - 400bp in length, which was the optimal fragment length for pyro- sequencing using the Roche 454 platform. These sequences were aligned to the *Aedes aegypti* whole genome shotgun sequences accession numbers AAGE01000001-AAGE01655158 of the Liverpool project accession number AAGE01000000 to select fragments that are present in both species. One hundred base pairs were added from each flanking gene or exon and were mapped resulting in 11 good hits from 15,800 intergenic regions and 45,000 introns. A good hit consisted of regions conserved between the two species and a species specific variable part between them. Primers that amplified well and gave single amplicon band on a gel, in both *Aedes* and *Anopheles* mosquitoes were selected for sequencing.

3.3.2 PCR setup and cycling conditions

Genomic DNA was diluted to a working concentration of 20ng/µl. PCR reactions were carried out in final volumes of 20µl containing 0.2µl of forward and reverse primer (20pmol/µl) each, 10µl of 2X Reddy mix PCR master mix (Thermo-scientific), 6µl DNA (20ng/µl) and 3.6µl sterile water.

Three different cycling conditions were optimized for the 15 primer sets. The two sets of ITS1 primers used the same program involving initial denaturation at 94ºC for 60 seconds, followed by 35 cycles of denaturation at 94ºC for 60 seconds, annealing at 51ºC for 60 seconds and elongation at 72ºC for 2 minutes and a final elongation at 72ºC for 6 minutes. The two pairs of ITS 2 primers were amplified using a PCR program involving initial denaturation at 94ºC for 60 seconds, followed by 35 cycles of denaturation at 94ºC for 60 seconds, annealing at 51ºC for 60 seconds and elongation at 72ºC for 1 minute and a final elongation at 72ºC for 6 minutes.

The eleven ANG primers were amplified using a touchdown PCR program with initial denaturation at 94ºC for 3 minutes, followed by 15 cycles of denaturation at 94ºC for 30 seconds, annealing at 65ºC (-1ºC/cycle)for 30 seconds and elongation at 72ºC for 60 seconds, followed by 35 cycles of denaturation at 94ºC for 45 seconds, annealing at 50ºC for 45 seconds and elongation at 72ºC for 45 seconds and a final elongation at 72ºC for 6 minutes. Three micro litres of the PCR product were visualized on a 1% agarose gel (molecular grade agarose) stained with a fluorescent gel red stain (Biotium)

under a UV trans-illuminator (VilbertLormet, France). Fragment size comparison was done with a 25bp DNA step ladder G451A (Life technologies).

The remaining PCR product (17µl) was then purified using a Qiagen PCR purification kit (Qiagen) following the manufacturer's protocol see (www.qiagen.com). The resulting cleaned product was quantified using a Thermo Scientific nanodrop^{TM} 1000 spectrophotometer. The minimum concentration accepted for downstream processing was 15ng/µl. In cases where the concentrations did not reach the minimum required concentrations, the sample amplification was repeated.

3.4 Amplicon sequencing

Amplicons were sequenced using the 454 GS - FLX sequencer (Roche). The following steps were involved:

1. MID adaptor ligation
2. Small fragment removal
3. Library quantification and quality check
4. Emulsion PCR
5. Sequencing
6. Data quality control and data analysis

3.4.1 MID adaptor ligation

Cleaned amplicons were pooled in equimolar amounts to a final concentration of 500ng

in 20µl. Each amplicon pool was end polished where 3' overhangs were removed and 3' recessed ends were extended resulting in blunt ends. The 5'ends were phosphorylated providing the ligation point for the Rapid Library. Library (RL) adapters A and in this study we used barcoded RL adapters. The barcodes or multiplex identifiers (MIDs) are short 10- 11bp oligonucleotide fragments attached to the adapter sequence that pooling of different samples in a single sequencing run and enabled sample sorting in the subsequent data analysis. Each RL adapter has a 4-base non-palindromic sequencing key used by the system's software for base calling and to recognize legitimate library reads (**Figure 5**).

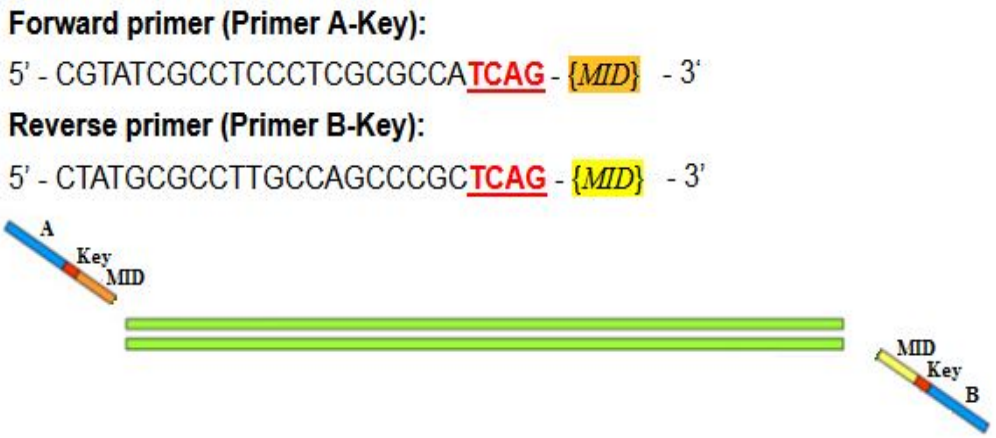

Figure 5: Illustration of RL-MID adapter consisting of MID Key and adapters A and B sequences. (GS - FLX-Titanium System Technical Overview, Customer support- Genome sequencing- www.roche-applied-science.com)

Adaptor B contains a biotin tag on its 5'-strand which is required in downstream processes (**Figure 6**).

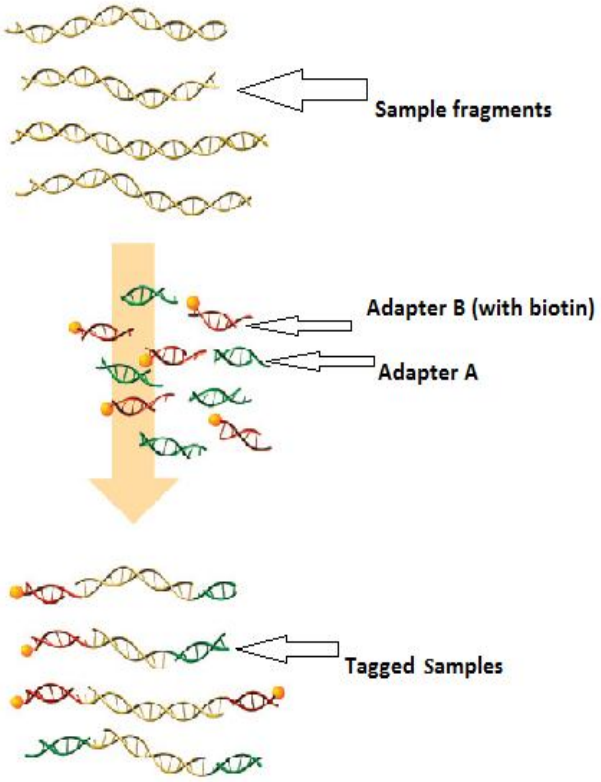

Figure 6: RL MID adaptor ligation (GS FLX-Titanium System Technical Overview, Customer support- Genome sequencing- www.roche-applied-science.com)

The biotin tag on Adaptor B allows the immobilization of the dsDNA library fragments and the subsequent isolation of the library of ssDNA sequencing templates. There were RL MIDs available for use with the Roche 454 platform thus increasing the multiplexing capacity; hence many samples would be pooled and processed together in a multiplex fashion. The sequencing reads from each of the pooled libraries were thereafter sorted based on the MID tag, correctly assigned to the original sample /pool and analysed separately using sequence data analysis software

The 20 amplicon pools (pilot study) and 60 samples (Field samples) were bar-coded (MID tagged) as shown in **Table 5** and **Table 6** respectively;

Table 5: Multiplexing of pilot study samples for sequencing

Pool	Sample ID	MID used
1	*Aedes aegypti* singleton	MID 1
2	*Aedes aegypti* singleton	MID 2
3	*Aedes aegypti* singleton	MID 3
4	*Aedes aegypti* singleton	MID 4
5	*Aedes aegypti* pool	MID 5
6	*Culex quinquefasciatus* singleton	MID 6
7	*Culex quinquefasciatus* singleton	MID 7
8	*Culex quinquefasciatus* singleton	MID 8
9	*Culex quinquefasciatus* singleton	MID 9
10	*Culex quinquefasciatus* pool	MID 10
11	*Anopheles arabiensis* singleton	MID 11
12	*Anopheles arabiensis* singleton	MID 12
13	*Anopheles arabiensis* singleton	MID 2
14	*Anopheles arabiensis* singleton	MID 3
15	*Anopheles arabiensis* pool	MID 4
16	*Anopheles gambiae* singleton	MID 5
17	*Anopheles gambiae* singleton	MID 6
18	*Anopheles gambiae* singleton	MID 8
19	*Anopheles gambiae* singleton	MID 9
20	*Anopheles gambiae* pool	MID 10

Table 6: Multiplexing of field samples for sequencing

Number	Pool ID	MID	No	Source And species		MID
1	*An .gambiae*	1	31	Baringo *Anopheles* Species		7
2	*An .gambiae*	2	32	Baringo *Anopheles* Species		8
3	*An .gambiae*	3	33	Baringo *Anopheles* Species homogenate		9
4	*An .gambiae*	4	34	Baringo *Aedes* Species homogenate		10
5	*Anopheles Arabiensis*	5	35	Baringo *Culex* Species homogenate		11
6	*Anopheles Arabiensis*	6	36	Naivasha *Culex* Species		12
7	*Anopheles Arabiensis*	7	37	Naivasha *Culex* Species		13
8	*Anopheles Arabiensis*	8	38	Naivasha *Culex* Species		14
9	*Aedes* Species	9	39	Naivasha *Culex* Species		15
10	*Aedes* Species	10	40	Naivasha *Anopheles* Species homogenate		16
11	*Aedes* Species	11	41	Naivasha *Aedes* Species homogenate		17
12	*Culex Quinquefasciatus*	12	42	Naivasha *Culex* Species homogenate		18
13	*Culex Quinquefasciatus*	13	43	Bodhai *Culex* Species		19
14	*Culex Quinquefasciatus*	14	44	Bodhai *Culex* Species		20
15	*Culex Quinquefasciatus*	15	45	Bodhai *Culex* Species		21
16	Ijara-21	16	46	*Icipe Aedes* species_ITS1 F/R	Not pooled	22
17	Ijara 22	17	47	*Icipe Aedes* species_ITS1 A/B	Not pooled	23
18	Ijara 23	18	48	*Icipe Aedes* species_ITS2 F/R	Not pooled	24
19	Ijara 24	19	49	*Icipe Aedes* species_ITS2 A/B	Not pooled	1
20	Ijara29	20	50	*Icipe Aedes* species_ANG 00020	Not pooled	2

21	Ijara30	21	51	*Icipe Aedes* species_ANG 00026	Not pooled	3
22	Ijara31	22	52	*Icipe Aedes* species_ANG 04289	Not pooled	4
23	Ijara32	23	53	*Icipe Aedes* species_ANG 12432	Not pooled	5
24	Ijara33	24	54	*Icipe Aedes* species_ANG 13935	Not pooled	6
25	Ijara34	1	55	*Icipe Aedes* species_ANG 18326	Not pooled	7
26	Ijara35	2	56	*Icipe Aedes* species_ANG 20362	Not pooled	8
27	Ijara36	3	57	*Icipe Aedes* species_ANG 20760	Not pooled	9
28	Ijara-*Culex*	4	58	*Icipe Aedes* species_ANG23972	Not pooled	10
29	Baringo41	5	59	*Icipe Aedes* species_ANG 26425	Not pooled	11
30	Baringo42	6	60	*Icipe Aedes* species_ANG 27523	Not pooled	12

The amplicons of an *Aedes aegypti* singleton from *icipe* were not pooled and the amplicons sequenced individually, in order to provide a reference sequences at the point of analysis.

The RL MID adapters were ligated by adding 1µl of RL MID Adaptor and 1µl of ligase was added to the reaction tube containing 15µl of the amplicons pools. The tube was vortexed for 5 seconds, centrifuged for 5 seconds in a table-top centrifuge and incubated at 25°C for 10 minutes (original protocol at www.my454.com).

3.4.2 Small fragments removal

During adapter ligation step, a small part of adapters will ligate to each creating dimers. These small fragments were removed using Agencourt AmpureXP beads (Beckman Coulter, Inc.) before library sequencing. The small fragment removal as well as library quality and quantity checks were carried out following the manufacturers' instructions (original protocol at www.my454.com).

In brief, library quality was assessed by checking the fragment length distribution using 7500 DNA chip on Agilent 2100 bio-analyser (Agilent technologies). The Agilent's 2100 Bio-analyser uses a lab on a chip approach to perform capillary electrophoresis and uses a fluorescent dye that binds to DNA to determine both DNA concentration and integrity (Figure 7).

Figure 7: Agilent 7500 DNA chip used in library quality assay (GS FLX-Titanium System Technical Overview, Customer support- Genome sequencing- www.roche-applied-science.com)

The library was quantified using the TBS380 fluorometer (Daigger) that uses blue light instead of UV. Serial dilutions of RL standard, provided within the RL library preparation kit were used to generate a standard curve. The fluorescence readings of the

samples was then read and used to calculate the library sample concentration. The library was then diluted to a working concentration of 1 x 10^7 molecules/µl using tris-EDTA buffer.

3.4.3 Emulsion polymerase chain reaction

Emulsion PCR is an in vitro cloning step to amplify individual DNA molecules and to increase the sensitivity of sequencing. The 454 life sciences emulsion PCR used is based on methods by (Margulies *et al*., 2005). Individual DNA molecules are isolated along with primer-coated beads in aqueous droplets within an oil phase. During emulsion PCR each DNA molecule was clonally amplified followed by immobilization for later sequencing (**Figure 8**).

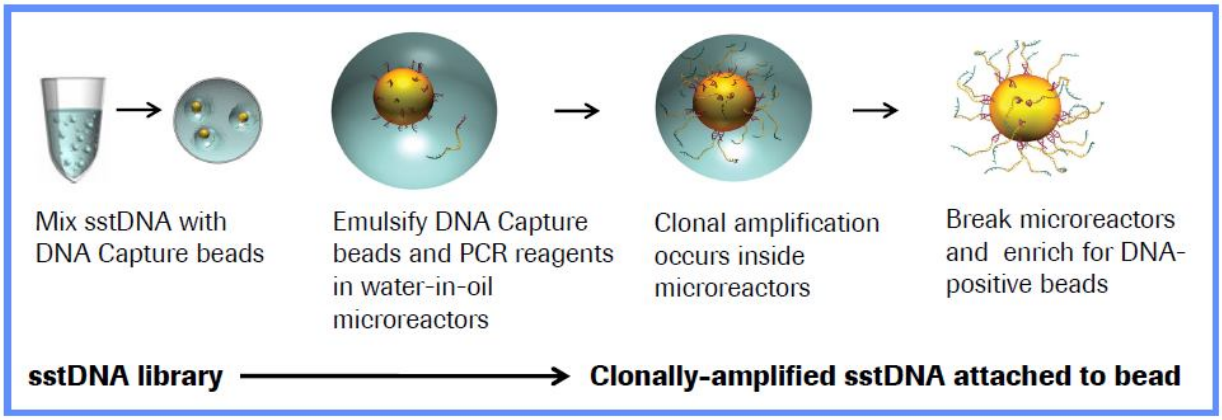

Figure 8: Emulsion PCR (GS FLX-Titanium System Technical Overview, Customer support- Genome sequencing- www.roche-applied-science.com)

The volume of diluted library to be added to the emPCR reaction was calculated using the formula;

$$\mu l \text{ of library per tube} = \frac{\text{desired molecules per bead} \times 6.8 \times 106 \text{ beads per tube}}{\text{Library concentration (in molecules/}\mu l)}$$

The optimal number of molecules per bead that result in good quality sequence results depends on the DNA source and needs to be individually determined by performing an emulsion titration. The titration process involved performing separate emPCR amplifications using different molecules per bead ratios 6:1, 5:1, 4:1, 3:1, and 2:1 and for each the percentage of enriched beads was determined. As a general rule the molecule bead ratio with percentage enrichment ranging from 5% - 20% was used as recommended by Roche (see www.my454.com)

For mosquito libraries, 4 molecules per bead (4: 1) gave the best enrichment percentage of 5% and thus the emPCR was done with a target four DNA molecules per bead. The emulsion PCR was set up following emPCR method manual- Medium volume (see www.my454.com). Single stranded library DNA molecules were annealed to capture beads containing a complement of the adapter primer A. The beads with four DNA molecules each attached were then mixed in thermo-stable water- in-oil emulsion. Each bead was contained in a micro reactor containing a complete amplification mix. The emulsion was loaded on a PCR plate and amplified using the program: 94ºC for 4 minutes, followed by 50 cycles of denaturation at 94ºC for 30 seconds, annealing at 58ºC for 4 minutes and 30 seconds and elongation at 68ºC for 1 minute and a final hold at 10ºC; on a thermocycler (GeneAmp PCR system 9700 thermocycler- Applied Bio systems technologies).

Once the amplification program was complete, the emulsion breaking and bead recovery were carried out as outlined in the emPCR method manual (www.my454.com). The emulsions were collected from the plate and washed with isopropanol (100%), absolute ethanol and enhancing buffer to get rid of the oil, leaving the beads containing the DNA cloned libraries.

The recovered library was enriched and the percentage bead enrichment determined (formula provided in the manual). The enrichment process was carried out in order to select out the DNA beads with amplified DNA and wash away empty beads. This was done by first creating single stranded libraries on the bead using melt solution and adding an enrichment primer (containing a biotin moiety) which binds to the ssDNA on the bead (**Figure 9**).

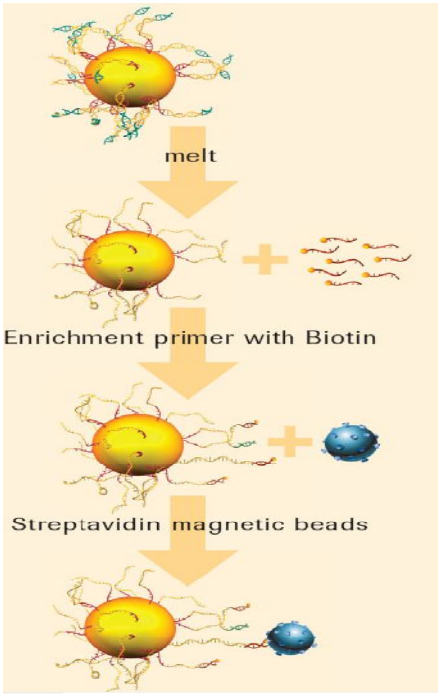

Figure 9: Library enrichment (GS FLX-Titanium System Technical Overview, Customer support- Genome sequencing- www.roche-applied-science.com)

Using a MPC (Magnetic Particle Concentrator), only DNA beads carrying amplified DNA fragments were able to bind to magnet and the empty ones were washed away. The DNA beads were separated from the magnetic beads using melt solution. A sequencing primer, complementary to the adapter primer A sequence was annealed to the fragments on the beads. Excess primers were removed through a series of washes. The percentage bead enrichment was then calculated by determining the number of enriched beads in relation to the number of the input beads using the Beckman Coulter Counter Z2 (Beckman)

3.4.4 Sequencing and data collection

Sequencing was done using the Genome Sequencer FLX Instrument (**Figure 10**) and GS FLX Titanium Sequencing Kit XLR70.

Figure 10: Genome Sequencer FLX Instrument (GS FLX-Titanium System Technical Overview, Customer support- Genome sequencing- www.roche-applied-science.com)

For this sequencing run, a medium Pico titre plate (PTP) device GS-FLX sequencer (Roche Life Sciences), (**Figure 11**) which has four lanes was used.

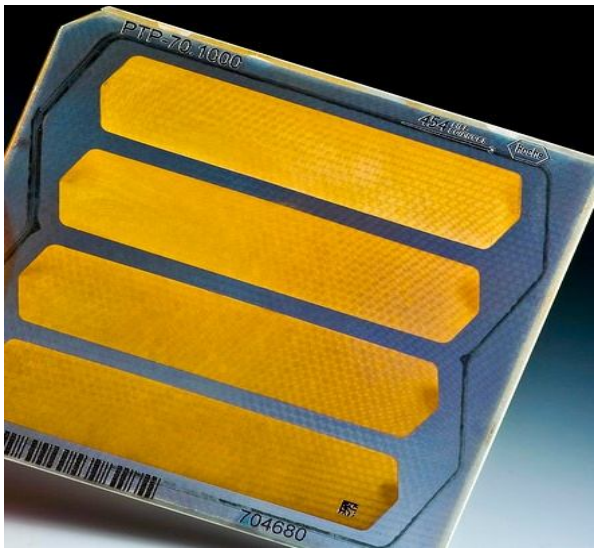

Figure 11: Pico- titre plate device with four gasket regions (GS FLX-Titanium System Technical Overview, Customer support- Genome sequencing- www.roche-applied-science.com)

From the bead counts obtained from the coulter counts, a calculation was done to pick the correct volume of library to load in order to acquire the 790, 000 DNA library beads per quarter region of the PTP device. The PTP, enzyme beads, packing beads, Peptidyl-Prolyl Cis-Trans Isomerase (PPiase) beads and DNA beads were prepared as per the manufacturers' instructions. The beads were then loaded onto the PTP according to prescribed order and method with each well taking a single DNA bead (**Figure 12**).

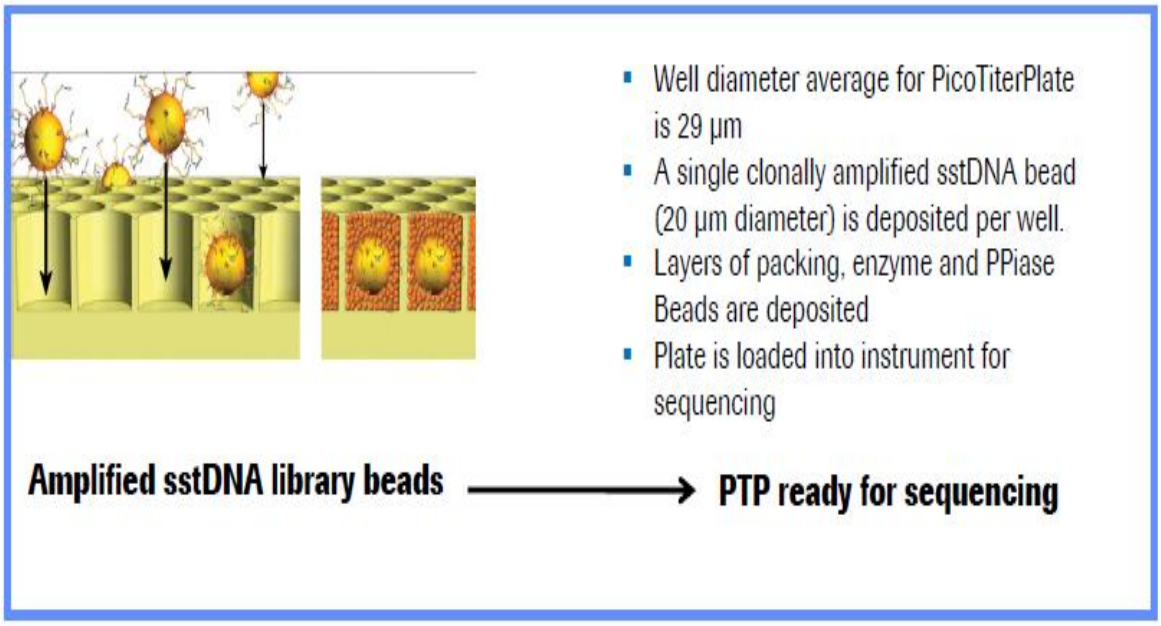

Figure 12: One bead per well of the Pico titre plate (GS FLX-Titanium System Technical Overview, Customer support- Genome sequencing- www.roche-applied-science.com)

The 454 sequencing applies a pyro-sequencing/ sequencing- by- synthesis approach. Thus once the loaded PTP was placed into the machine each nucleotide was flowed over the PTP one at a time over 200 cycles. The copy strand of the fragment begins being generated and with addition of nucleotides a pyrophosphate molecule is released. In solution, the enzyme sulfurylase mediates the conversion of AMP to ATP using the pyrophosphate, releasing energy. This energy is taken up by luciferase enzyme to hydrolyse ATP with production of light. This light is produced at a specific intensity per molecule attached and the light is captured by a CCD (charge-coupled device) camera generating an image with every flow **(Figure 13**).

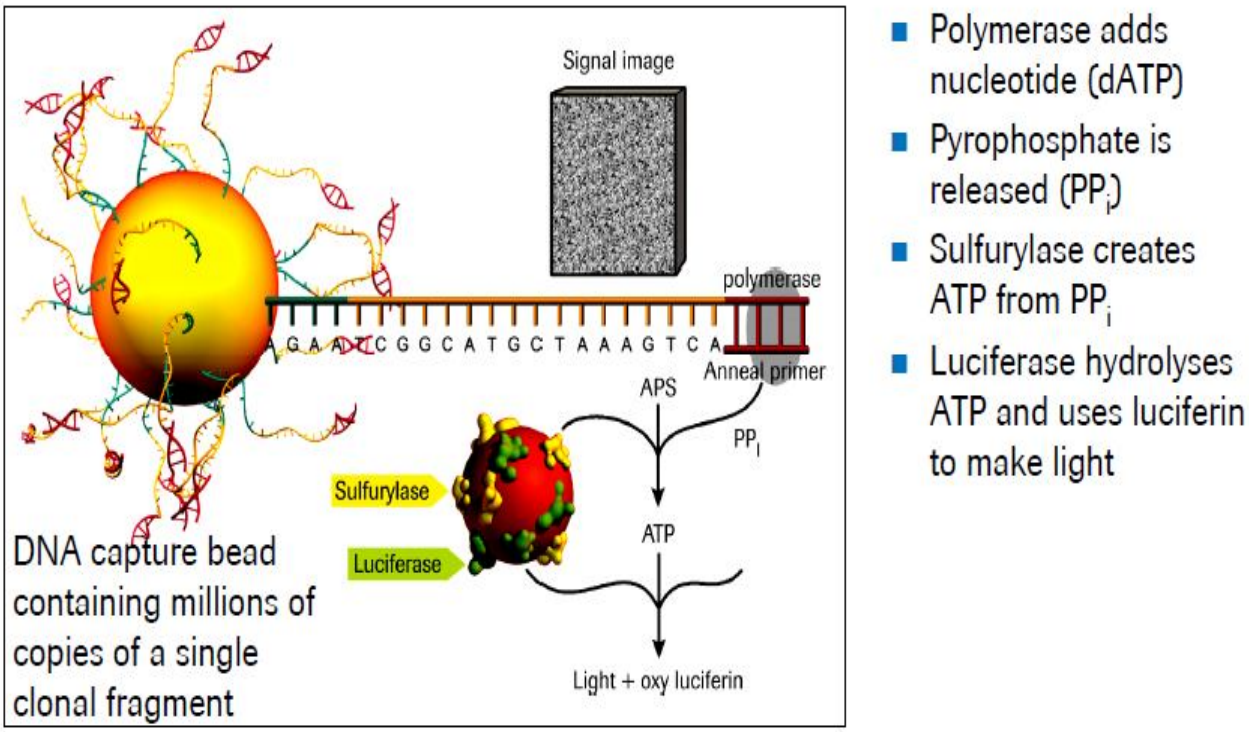

Figure 13: The principle of pyrosequencing (GS FLX-Titanium System Technical Overview, Customer support- Genome sequencing- www.roche-applied-science.com)

Signal and sequence processing were then carried out to determine base sequence and quality. Only Image processing was done during the sequencing run. It was possible to monitor the progress of the sequencing run by viewing the Instrument status and the data images as they were being captured by the camera: Thumbnail images appeared under the progress bar in the Instrument tab during the Run (**Figure 14**).

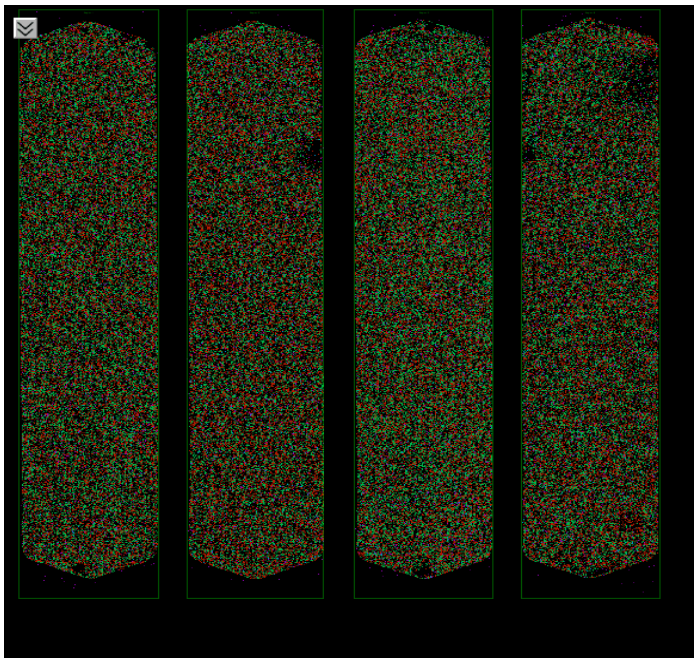

Figure 14: Image captured by CCD camera as sequencing run progresses (GS- FLX titanium sequencer instrument)

Sample co-loading

Pilot study samples

The first 12 pools (Pool 1- 12) of mosquito amplicons (tagged with MID 1- 12), were loaded to one region of the Pico-titre plate while the remaining 8 (pool 13-20) were co-loaded with another set of samples on region two.

Field study samples

The samples were loaded into three regions of the Pico titre plate. Region one and two had twenty four (24) samples each of pooled amplicons while region three had un-pooled amplicons of sample twelve (*Aedes aegypti*) from *icipe.* The reason for not pooling was so as to create reference sequences for the data set on which to carry out mapping during data analysis.

3.5 Data manipulation and analysis

3.5.1 Image processing

During the sequencing run, the raw fluorescence images captured by the camera underwent a processing step which was done using the GS Run Processor (Genome sequencer- Run processor), a software provided by the Roche Life Sciences. This processing was quick and not memory intensive in terms of computing requirements. It took about 20 minutes and the data was stored in Composite Well Format (CWF) files. The software worked by finding the active wells and extracting raw signals for each flow in each active well. It then wrote the resulting flow signals into the CWF files. The software was able to subtract background and normalize the images at the pixel level.

3.5.2 Signal processing

The raw data was analysed and the CWF files produced by the image processing step corrected. This step is referred to as signal processing and it is an automated process where a post analysis script within the analysis cluster is ran to perform the task. This processing step produced corrected CWF files of the images and SFF files containing the sequences. It is a memory intensive process and takes about seven (7) hours to complete.

3.5.3 Sequence Assembly

For each region on a PTP plate, separate SFF (Sequence File format) file containing sequences from all libraries (e.g. RLMID1 to RLMID12) sequenced in this particular region was generated. Each SFF file was processed separately and sequences were sorted based on the MID tag using SFF tools (software provided by Roche), resulting in independent SFF files for each MID tagged amplicon pool. Each SFF file was submitted to RDB's pyro- sequencing pipeline (Cole *et al*., 2009) to sort sequences in each file by amplicons generated from each specific primer. During this process data of any non-specific sequences occurring in the sequence result file as a sequencing artefact or chimera were deleted.

The resulting files were FNA files (Fasta Nucleic-Acid format files), each containing sequences which with amplification primers (Forward, Reverse or both). Using CLC-genomics workbench5, each FNA file was processed to select only the complete sequences that had both the forward and reverse primers. From each FNA file (represented each sample) one sequence per primer was selected as a reference sequence for mapping.

Mapping was carried out in order to create consensus sequences for each of the FNA files using the GS- FLX mapping software (v 2.5.3) (provided by Roche). Unmapped reads were remapped using a representative read from the unmapped sequences and this

was repeated for all unmapped sequences until all reads mapped resulting in additional sequences.

Using TABLET software (Milne *et al*., 2010), the 454.ace files resulting from mapping were visualized. All consensus sequences generated from less than 10 reads were discarded. The remaining sequences were further visualized (On TABLET), to check the reads source (species, geographical origin), read length and read sequence nucleotide variations. A representative sequence of each variant was included in the consensus sequences FNA file for later use in alignment. Where no variant consensus was observed but the sources of reads varied, the consensus sequence name was edited to indicate the different sources. Sequence length information was also retained on the accession names, e. g. in Contig 01ANG18326_*Aedes_aegypti_icipe*_Ijara (meaning the consensus sequence was built form reads from *icipe* and Ijara and the reads were not variable).

The combined consensus sequences plus representative reads were aligned using MEGA5 software (v 5.1 beta) and phylogenetic trees were constructed using the same program.

3.5.4 Data analysis

Multiple sequence alignments were carried out for different clusters to capture information from different combinations for all the primer sets (**Table 7**) using the default parameters of ClustalX software (v 2.1.).

Table 7: Various clusters selected for sequence alignment

Primer 1(ITS1 F/R)	Species	Primer	Origin
Alignment 1 (Pilot Study)	Same species e. g All *Aedes aegypti*	Same primer e. g. ITS1	*Icipe*
Alignment 2 (Pilot Study)	Different species	Same primer e. g. ITS1	*Icipe*
Alignment 2 (Field Study)	Same species e. g All *Aedes aegypti*	Same primer e. g. ITS1	Same region e. g. Ijara
Alignment 3 (Field Study)	Different species	Same primer e. g. ITS1	Same region e. g. Ijara
Alignment 4 (Field Study)	Same species e. g *Aedes aegypti*	Same primer e. g. ITS1	Different regions e. g. Ijara, and Baringo

The aligned sequences were then edited in Bioedit sequence alignment software (v 7) and MEGA software (v. 5.1 beta) and phylogenetic trees constructed using Neighbour Joining method. Nucleotide base polymorphisms were also studied in the alignments identify polymorphic sites.

CHAPTER FOUR

RESULTS

4.1 Pilot study results

4.1.1 DNA Extraction results

Extracted DNA was of good purity with 260/280 ratio ratios ranging between 1.7- 1.8 as determined by the Thermo- Scientific Nanodrop^{TM} 1000 spectrophotometer (**Figure 15**). A concentration of 20ng/µl was the minimum concentration required for the polymerase chain reaction to be carried out and this was obtained for all the samples.

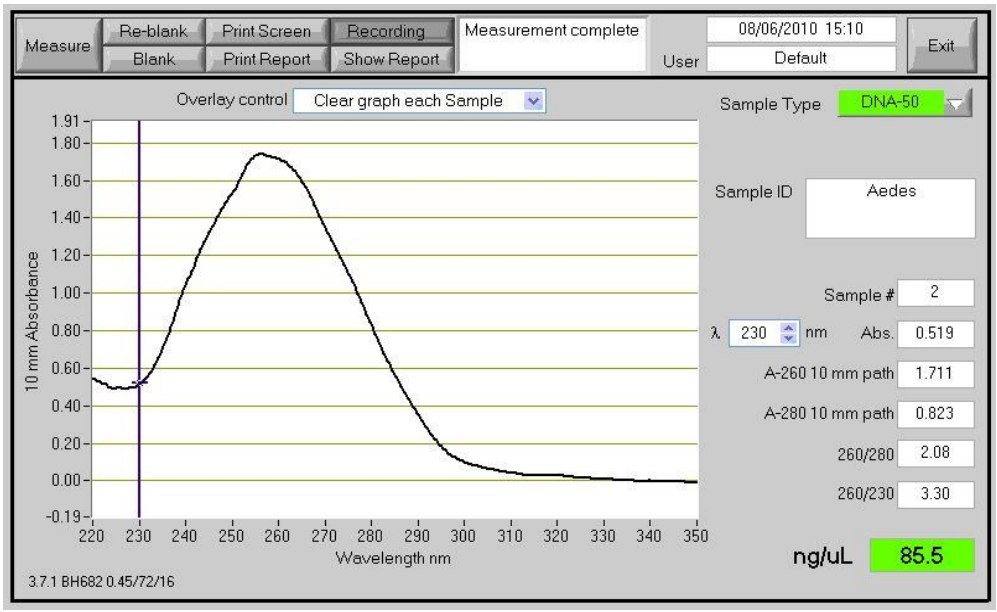

Figure 15: DNA quantification image showing the concentration measured at wavelength 260 using a Nano-drop.

4.1.2 DNA Amplification results

For each primer pair tested several criteria were applied. Each amplicon to be sequenced was first visualized on UV trans- illuminator after electrophoresis on a 1% agarose gels.

4.1.2.1 Internal transcribed spacers

Both ITS1 and ITS2 regions amplified well for all the species (Plate 1 and 2). ITS1 PCR amplifications worked for all species giving PCR products of various sizes ranging between 250 and 500bp (Plate 1).

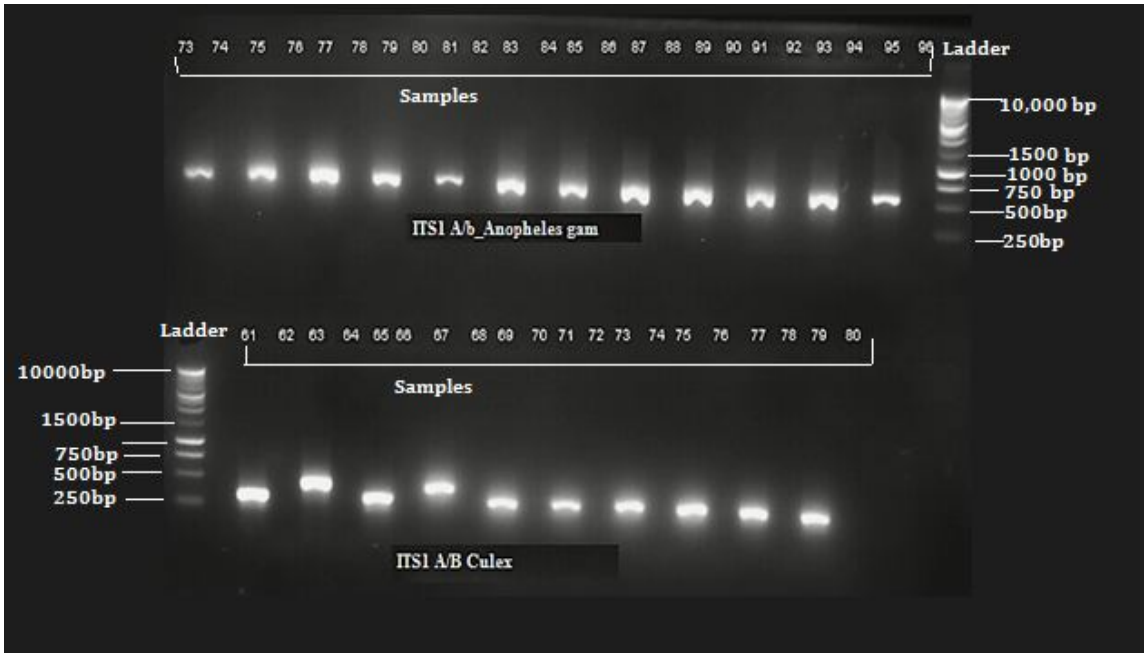

Plate 1: *Anopheles gambiae* and *Culex pipiens* amplified using ITS1 A/B as viewed in gel after amplification. The wells marked ladder contain a 1Kb ladder (Invitrogen), while all other wells (marked with numbers) contain amplified sample DNA

ITS2 PCR amplifications worked for all species giving PCR products of various sizes ranging between 480 to 550 bp (**Plate 2**)

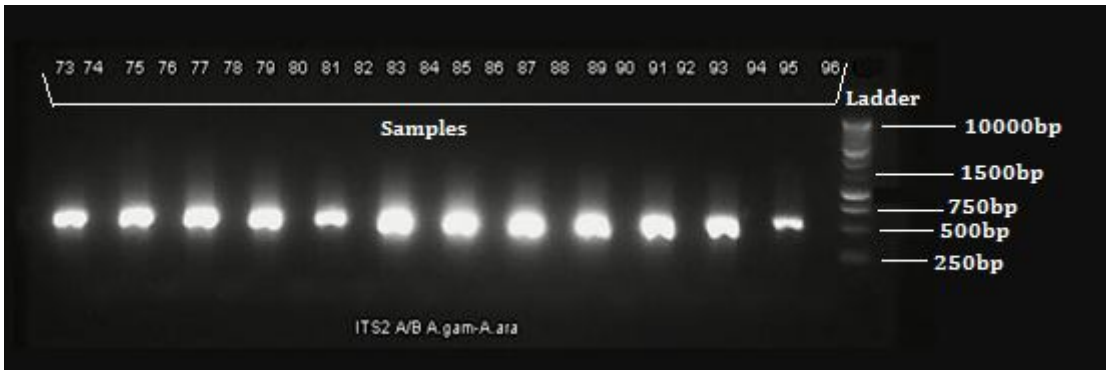

Plate 2: *Anopheles* species samples of the pilot study as viewed in gel after amplification using ITS2 A/B primer with an expected fragment size of 380 base pairs. The wells marked ladder contain a 1Kb ladder (Invitrogen), while all other wells (marked with numbers) contain amplified sample DNA

The ITS2 amplicon fragments were larger than the expected fragment size (380 bp). There was no visible variation of fragments between species at the gel electrophoresis level.

4.1.2.2 Intergenic and Intronic regions

Eleven ANG primers were tested in this study and they targeted various intergenic and intronic regions. Some primers for instance ANG 13935 did not amplify all samples thus no band was observed on the gel. Duplicates had been done and thus data was available for most species for the different primers except where the primer did not perform well. Such primers (poor performing) would not be recommended as useful in molecular classification or characterisation of mosquitoes. The remaining volume after gel electrophoresis was cleaned using QIAgen PCR purification kit and quantified using the Thermo Scientific nanodropTM 1000 spectrophotometer (**Table 8**).

Table 8: Concentrations of the cleaned amplicons for *Aedes, Culex* and *Anopheles* using the 15 primer pairs, from the pilot study

CLEANED SAMPLES CONCENTRATIONS (ng/ul)															
SAMPLE ID	**ITS1 F/R**	**ITS1 A/B**	**ITS 2 F/R**	**ITS2 A/B**	**ANG26425**	**ANG 00020**	**ANG 00026**	**ANG 12432**	**ANG 20760**	**ANG04289**	**ANG 13935**	**ANG 18326**	**ANG 20362**	**ANG27523**	**ANG 23927**
1- Aedes aegyptiae singleton	61.57	43.18	28.51	89.44	13.8	9.7	25.13	57.12	13.1	47.7	18.58	40.5	32.65	18.93	32.25
2- Aedes aegyptiae singleton	53.28	69.09	31.86	28.7	11.2	7.8	25.8	61.1	15.1	50.6	29.7	37	27.61	16.58	31.84
3- Aedes aegyptiae singleton	49.19	55.96	24.18	15.45	12	11.4	25.2	43.3	15.7	51.6	26.4	32.2	26.49	9.73	34.76
4- Aedes aegyptiae singleton	43.25	52.51	29.15	19.26	11.2	12.7	23.6	50.1	14.6	50.2	29.5	37.3	29.21	15.86	33.64
5- Aedes aegyptiae pool	54.65	53.54	26.54	21.99	15.3	11.9	24.5	37.9	17.9	55.3	25.7	21.2	32.5	19.35	37
6- Culex a. fasciatus singleton	17.27	52.39	30.07	24.73	7.8	8	16.3	50	14.3	15.2	21.7	38.5	13.74	16.61	11.45
7- Culex a. fasciatus singleton	47.6	47.15	25.77	11.32	9.2	8.5	17.7	23.7	19.7	15.1	29.3	44.8	11.1	15.33	12.29
8- Culex a. fasciatus singleton	41.35	91.12	47.32	22.09	13.93	10.58	12.2	26.91	10.5	9.08	19.76	9	12.43	9.61	12.02
9- Culex a. fasciatus singleton	27.85	55.84	33.65	33.5	12.63	11.32	13.2	16.65	11.16	10.67	14.88	10.94	14.81	10.79	9.9
10- Culex a. fasciatus pool	52.97	48.8	38.37	23.15	13	8.4	26.6	49.6	22.6	20.6	23.6	41.3	19.26	16.63	14.04
11- Anopheles arabiensis singleton	39.06	56.21	12.87	19.52	9	11.1	30.6	33	23.6	53.2	30.3	12.2	30.37	23.94	30.96
12- Anopheles arabiensis singleton	41.41	110.09	13.48	20.19	9.1	11.6	28	31.7	27.5	61.8	28.4	17.2	28.66	20.28	27.68
13- Anopheles arabiensis singleton	34.65	45.57	11.28	22.4	11.99	11.44	18.81	27.9	8.86	18.39	25.04	11.19	35.87	14.76	28.72
14- Anopheles arabiensis singleton	44.12	81.53	11.7	39.12	11.51	12.02	19.69	27.31	9.4	16.23	22.47	15.16	38.94	22.57	34.76
15- Anopheles arabiensis pool	48.49	57.96	12.68	25.86	11.2	9.9	28.1	32.5	24.1	62.4	33.6	10.7	33.92	19.73	23.32
16- Anopheles gambiae singleton	45.29	55.69	8.51	5.72	11.4	8.3	19.5	39.2	27.5	50.2	30.3	32	30.32	23.21	26.46
17- Anopheles gambiae singleton	53.3	67.6	19.9	51.9	21.5	42	36.2	34.6	34.6	54.5	16.5	37.5	39.6	41.2	32.6
18- Anopheles gambiae singleton	41.42	64.79	8.54	38.76	13.67	9.23	22.31	16.64	9.61	18.72	20.92	15.52	31.21	19.03	22.3
19- Anopheles gambiae singleton	45.8	63.6	10.9	37.5	14.2	14.7	28.9	38.22	33.8	56.9	19.4	34.7	36.3	28.4	28.8
20- Anopheles gambiae pool	37.9	43.7	13.95	24.05	12.6	11.1	29	37.2	29.5	56.1	26.6	33.7	23.74	16.85	26.84

4.1.3 Libraries and sequencing results

Quality of the constructed libraries and sample fragment size distribution were evaluated on 7500 DNA agilent chip on the Agilent 2100 Bio-analyser **Figure 16**. A ladder was loaded alongside the samples to serve as a standard during fragment analysis.

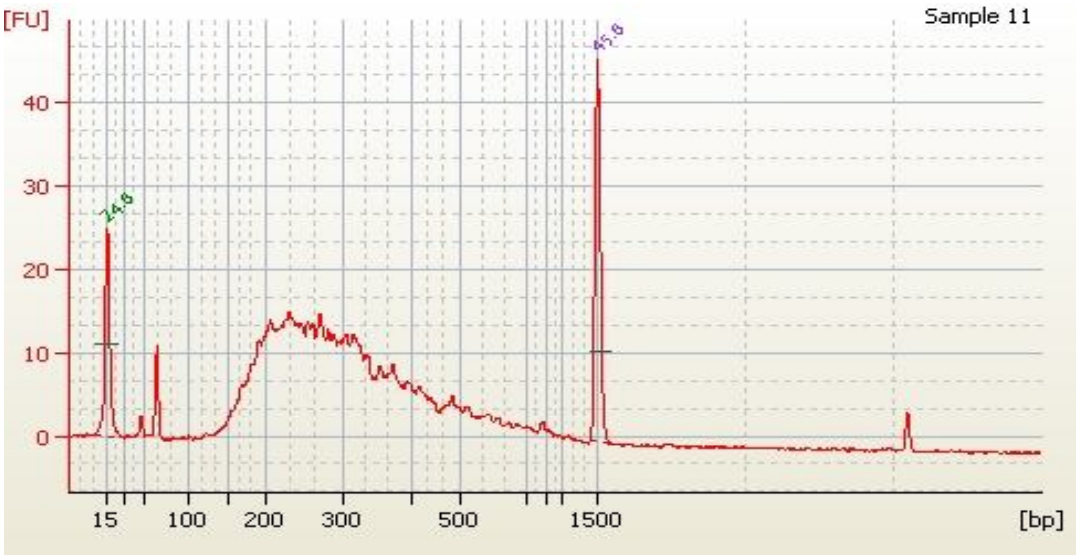

Figure 16: Representative trace of amplicons library sample run on the Bio- analyser High Sensitivity DNA chip.

The emulsion PCR worked and no broken emulsions were observed. An enrichment of 8% was achieved and the samples were thus rated suitable for sequencing. The 454 sequencing output was in the form of sequences generated from the reads built up by base calling in the processing steps after the sequencing run. The sequences then resulted in consensus sequences upon assembly and mapping based on the original sample source/ MID tag and primer used.

These consensus sequences for each primer were derived from the twenty sample sets in the case of the pilot study. The different primers gave varying number of consensus sequences indicating the efficiency or ability to amplify all the species samples worked on (**Table 9**).

Table 9: Number of contigs per primer resulting from the pilot study sequence analysis by mapping

	Primer	Contigs from PILOT study
1	ANG00020	6
2	ANG00026	14
3	ANG04289	6
4	ANG12432	11
5	ANG13935	10
6	ANG18326	18
7	ANG20362	9
8	ANG20760	12
9	ANG23972	0
10	ANG26425	6
11	ANG27523	5
12	ITS1 FR	0
13	ITS1 AB	0
14	ITS2 FR	10
15	ITS2 AB	4

4.1.4 Alignment and phylogenetic analysis results

Internal Transcribed spacer regions

ITS1 sequences obtained from study were too divergent to be aligned using various tested software (BioEdit and ClustalX). In addition, less than 0.5% of reads contained both the forward and reverse primer sequences. The rest had either a forward or reverse sequence only and were of varied length even with a single species. Any resulting alignment was not useful in determining variation and consequently no phylogenetic analysis was done using these sequences. The ITS2 sequences showed very low variation between species on Alignment (**Figure 17**) and phylogenetic analysis (**Figure 18**)

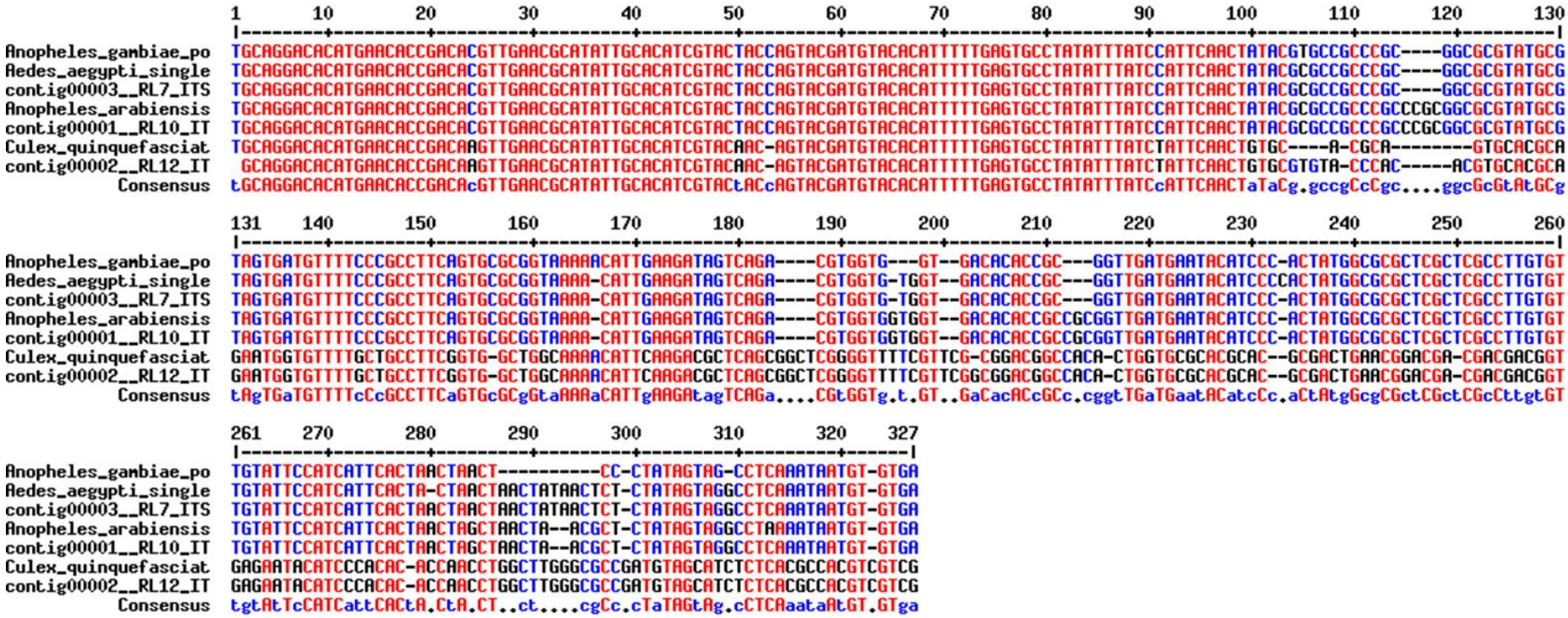

Figure 17: ITS2 Alignment result for pilot study data done using ClustalX version 2.0.11 using default parameters

Figure 18: Molecular Phylogenetic analysis of ITS2_FR by Neighbour-Joining method.

The evolutionary history was inferred using the Neighbour-Joining method. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (500 replicates) are shown next to the branches. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the Maximum Composite Likelihood method and are in the units of the number of base substitutions per site. The rate variation among sites was modelled with a gamma distribution (shape parameter = 4). The analysis involved 10 nucleotide sequences. Analyses were conducted in MEGA5 (v 5.1 beta).

Intronic and Intergenic regions

There was variation between species on alignment and phylogenetic analysis using primers ANG12432 (locus AAGE02010470.1) and ANG20760 (locus AAGE02008859.1) while there was no variation using primers ANG00026, ANG13935, ANG20362, ANG23972, ANG26425, ANG27523, ANG18326, ANG00020 and ANG04289 (**Appendix 1**).

ANG20760 loci

There was variation between species on Alignment (**Figure 19**) and phylogenetic analysis (**Figure 20**) using primer ANG20760

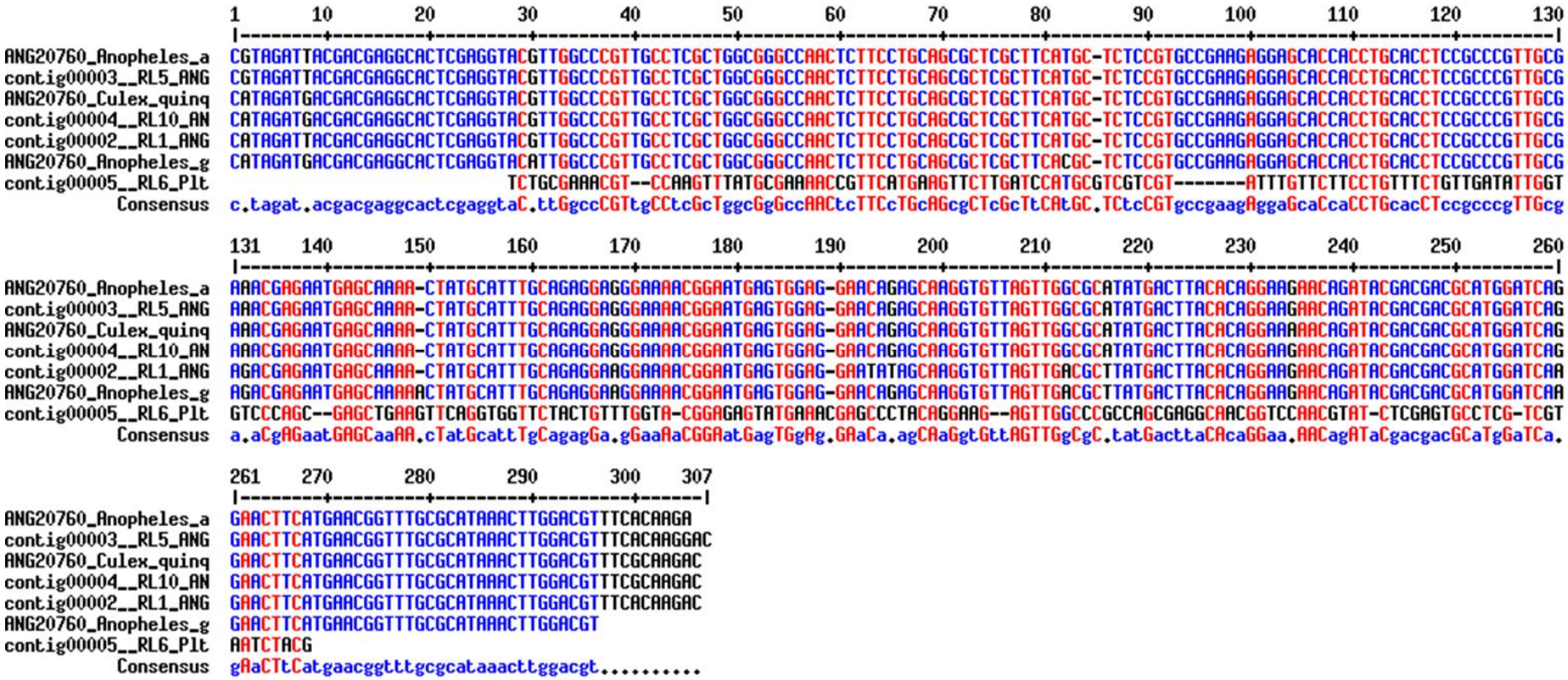

Figure 19: ANG20760 Alignment results of pilot study data. Alignment done using ClustalX version 2.0.11 using default parameters

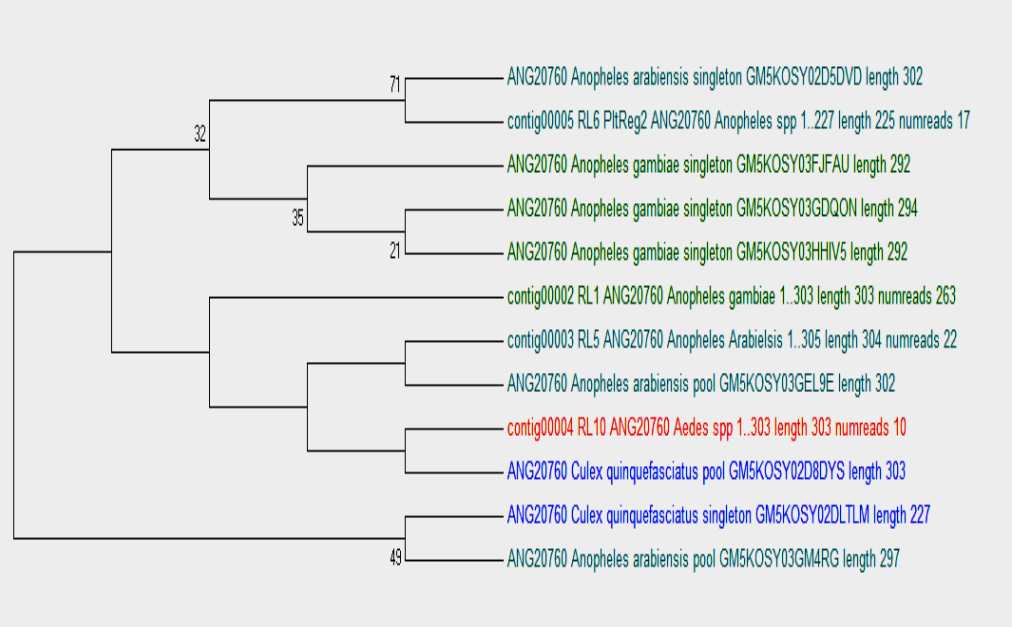

Figure 20: Phylogenetic tree of sequences amplified using ANG20760.

The evolutionary history was inferred using the Neighbour-Joining method. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (500 replicates) are shown next to the branches. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the Maximum Composite Likelihood method and are in the units of the number of base substitutions per site. The rate variation among sites was modelled with a gamma distribution (shape parameter = 4). The analysis involved 12 nucleotide sequences. Analyses were conducted in MEGA5 (v 5.1 beta).

There was variation between species on Alignment and phylogenetic analysis (**Figure 21**) using primer ANG12432

Figure 21: Phylogenetic tree of sequences amplified using ANG12432.

The evolutionary history was inferred using the Neighbour-Joining method. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (500 replicates) are shown next to the branches. The evolutionary distances were computed using the Maximum Composite Likelihood method and are in the units of the number of base substitutions per site. The rate variation among sites was modelled with a gamma distribution (shape parameter = 4). The analysis involved 11 nucleotide sequences. Analyses were conducted in MEGA5 (v 5.1 beta).

The other primers designed against intergenic and intronic regions were not able to differentiate the mosquito species (Trees derived from all other primers are shown in **Appendix 1**). Some of the primers designed against intergenic regions, e.g. ANG13935 designed against locus AAGE02005924.1, produced more than one amplicon for one sample, suggesting it amplifies more than one locus.

4.2 Field Study results

4.2.1 Sampling

The sampling process carried out in the RVF endemic regions showed an abundance of *Culex* and *Anopheles* mosquitoes compared to *Aedes* genus.

4.2.2 DNA Extraction results

The DNA extraction provided DNA of good purity (260/280 ratio) ratios ranging from 1.7 - 1.8 as determined by the Thermo- Scientific Nanodrop^{TM} 1000 spectrophotometer. A concentration of 20ng/µl was the minimum concentration required for the polymerase chain reaction to be carried out and this was achieved for all the samples worked on.

4.2.3 DNA Amplification results

The primers used amplified the selected regions of the mosquito genomic DNA and these amplicons were visualized on UV trans-illuminator upon electrophoresis on a 1% agarose gels.

4.2.3.1 Internal Transcribed Spacers

ITS1 PCR amplifications worked for all species giving PCR products of various sizes ranging between 250 and 500bp (**Plate 3**)

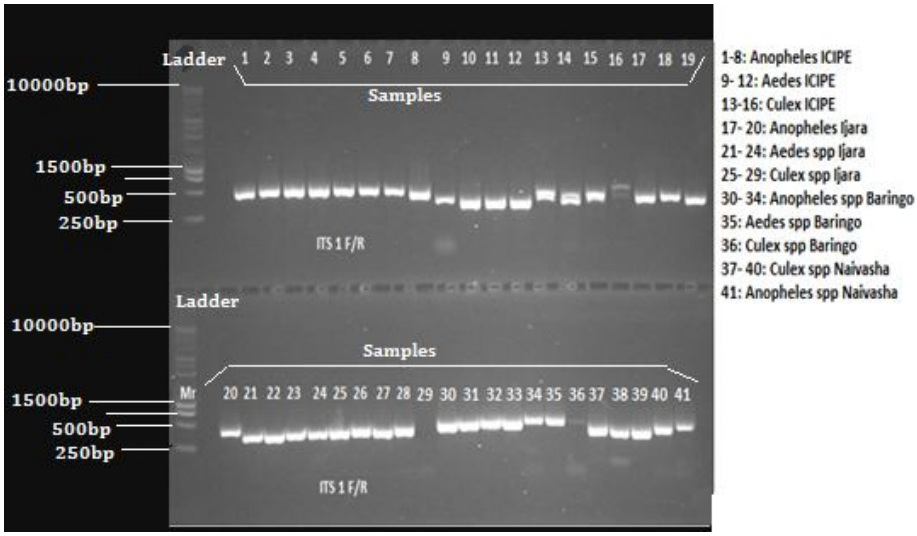

Plate 3: *Anopheles*, *Aedes* and *Culex* species samples of the field study amplified using ITS1 F/R primer as seen on gel. The expected amplicon fragment size was 167 base pairs. The wells marked ladder contain a 1Kb ladder (Invitrogen), while all other wells (marked with numbers) contain amplified sample DNA

ITS2 PCR amplifications worked for all species giving PCR products of various sizes ranging between 480 to 500 bp (**Plate 4**)

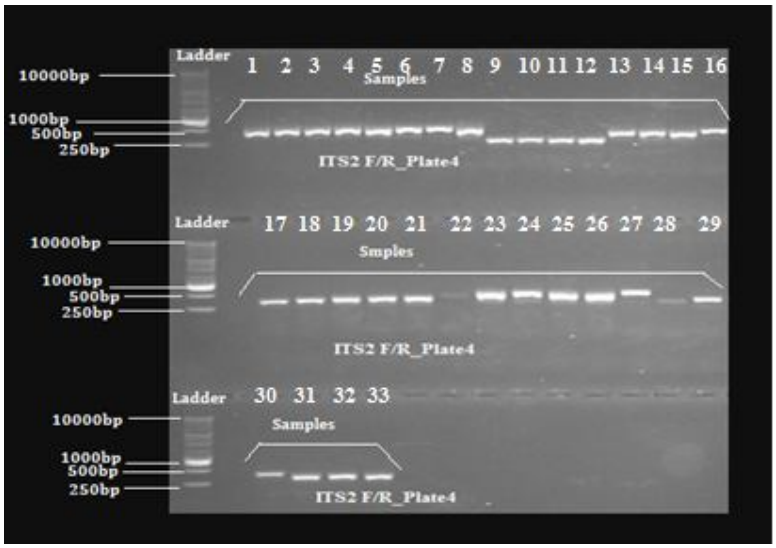

Plate 4: *Anopheles* species samples of the field study as seen on gel after amplification using ITS2 F/R primer with an expected fragment size of 380 base pairs. The wells marked ladder contain a 1Kb ladder (Invitrogen), while all other wells (marked with numbers) contain amplified sample DNA.

The bands, as in the case with other primers, some fragments appeared larger while others were the expected size. There was some visible variation between fragments at the gel electrophoresis level.

4.2.3.2 Intergenic and Intronic regions

Two intronic regions used to amplify the genomic DNA resulted in fragments with variations visible on gel based on genus. These were ANG12432 and ANG 20760 (**Plate 5** and **Plate 6** respectively). The others intergenic and intronic regions were able to amplify the DNA from all species but there was no visible size variation on gel.

ANG12432 loci

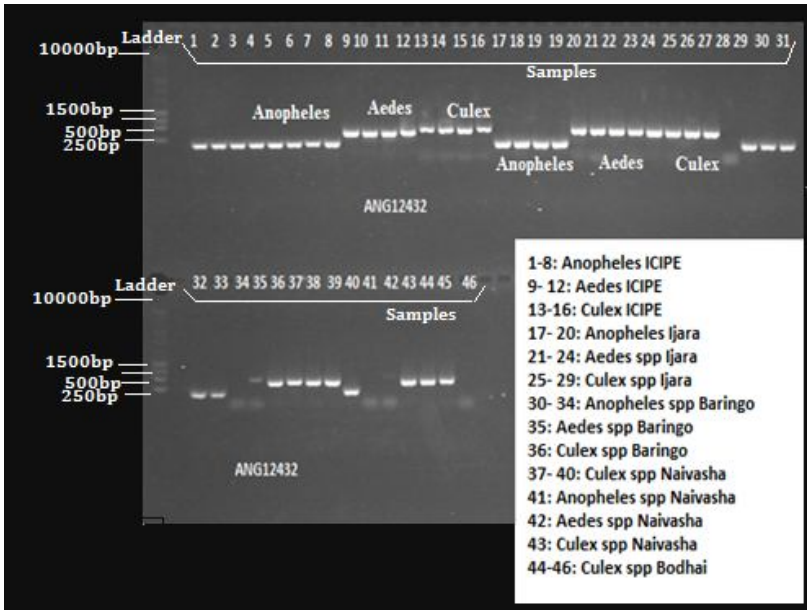

Plate 5: Gel image of field samples amplified using ANG12432 primer with an expected fragment size of 240 base pairs. The wells marked ladder contain a 1Kb ladder (Invitrogen), while all other wells (marked with numbers) contain amplified sample DNA.

ANG20760 loci

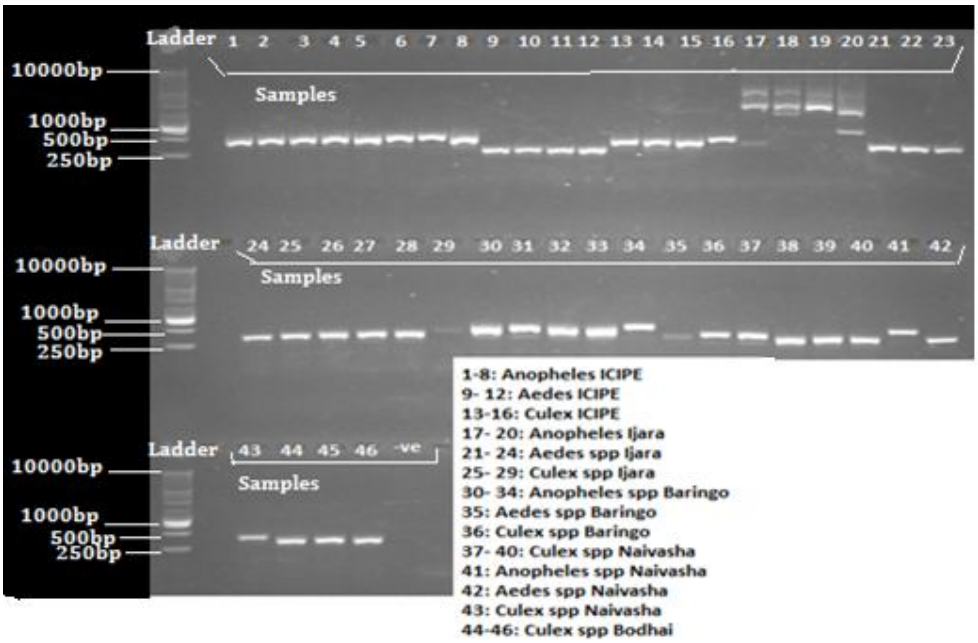

Plate 6: Gel image of field samples amplified using ANG20760 primer with an expected fragment size of 305 base pairs. The wells marked ladder contain a 1Kb ladder (Invitrogen), while all other wells (marked with numbers) contain amplified sample DNA

The fragment sizes of both ANG12432 and ANG20760 were of variable sizes with some appearing larger than the expected size and others were the expected size. The remaining volume of PCR product was cleaned using Qiagen PCR purification kit and quantified using the Thermo Scientific Nano-drop TM spectrophotometer. The amplicons were pooled to achieve equimolar concentrations based on sample source and subsequently libraries were prepared.

4.2.4 Libraries and sequencing results

The Library consisting of pooled amplicons was screened using the Agilent 2100 Bio-analyser using a Agilent DNA 7500 chip. This was to assess the quality of the

library constructed and resulted in a graphical output and on gel output (**Figure 22**) of the fragment length distribution. Emulsion PCR was then carried out.

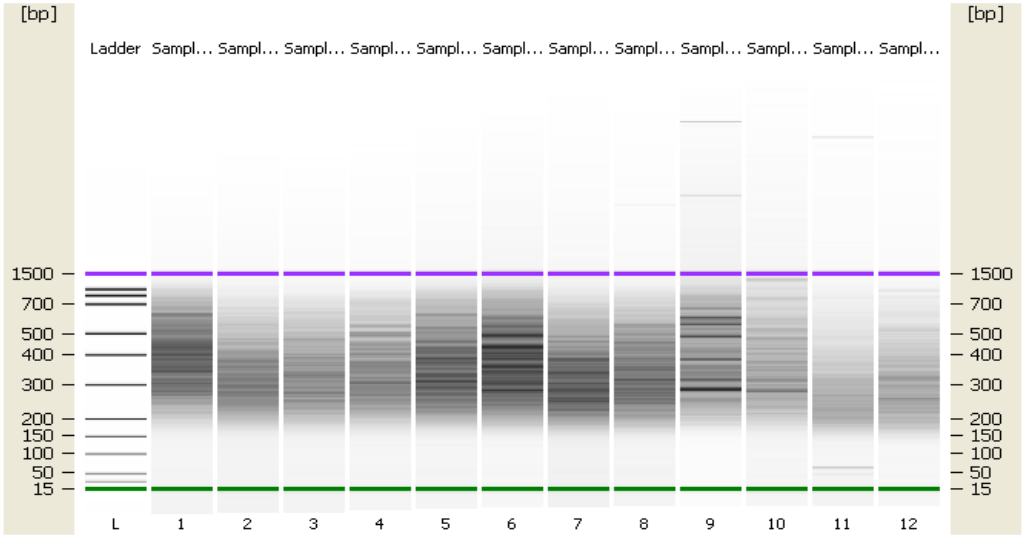

Figure 22: Gel output from high sensitivity DNA chip upon fragment analysis on Agilent 2100 bio-analyser

The libraries had a length distribution ranging between 150 and 800 bp which were subsequently used for emulsion PCR. The emulsion PCR worked and no broken emulsions were observed. An enrichment of 12% was achieved and the samples were thus rated suitable for sequencing.

The 454 sequencing output was in the form of sequences generated from the reads built up by base calling in the signal processing steps after the sequencing run. The sequences then resulted in contigs upon assembly and mapping based on the original sample source/ MID tag and primer used. These contigs for each primer were

derived from a set of sixty sample sets in case of the field study (**Table 10**). The different primers gave varying number of contigs indicating the efficiency or ability to amplify all the species samples studied.

Table 10: Number of contigs obtained for each primer for all field samples sets studied

	Primer	Contigs from FIELD study
1	ANG00020	17
2	ANG00026	8
3	ANG04289	3
4	ANG12432	12
5	ANG13935	27
6	ANG18326	26
7	ANG20362	22
8	ANG20760	25
9	ANG23972	25
10	ANG26425	11
11	ANG27523	3
12	ITS1 FR	2
13	ITS1 AB	17
14	ITS2 FR	6
15	ITS2 AB	6

4.2.5 Alignment and phylogenetic analysis results

ITS1 sequences were still found too divergent to be aligned using various tested software (Bio Edit and ClustalX). Of all the ITS1 sequences obtained, none was complete, that is, had both the forward and reverse primer. Thus no phylogenetic analysis was done using these sequences. The ITS2 sequences showed very low variation between species on Alignment and phylogenetic analysis (**Figure 23**)

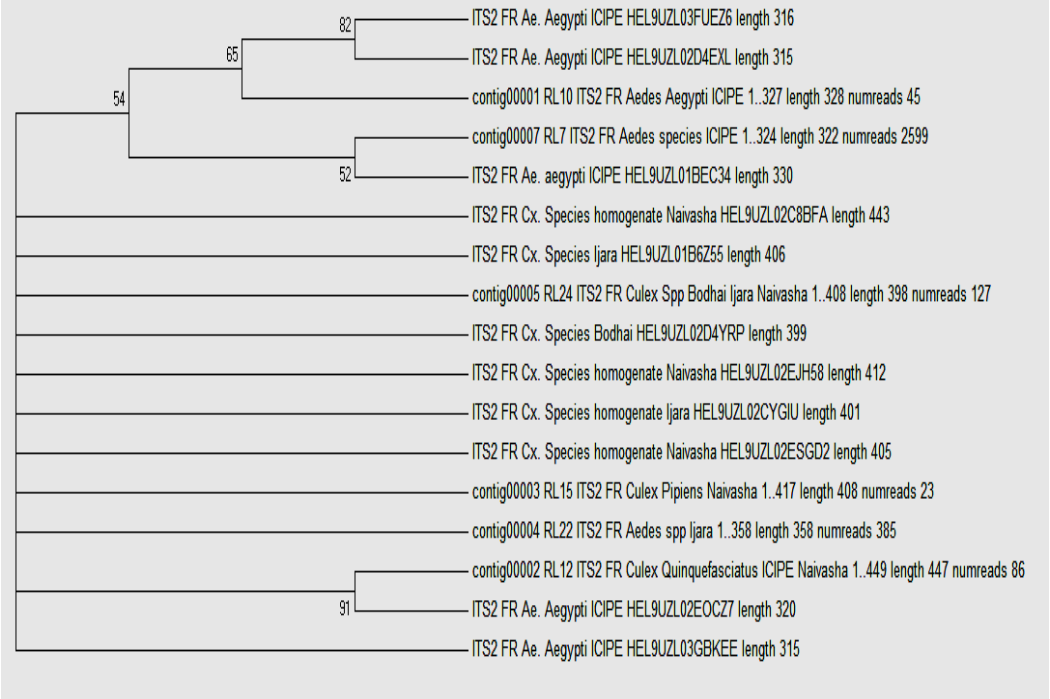

Figure 23: Molecular Phylogenetic analysis of ITS2 by Neighbour-Joining method

The evolutionary history was inferred using the Neighbour-Joining method. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (500 replicates) are shown next to the branches. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The rate variation among sites was modelled with a gamma distribution (shape parameter = 4.5). The analysis involved 17 nucleotide sequences. Analyses were conducted in MEGA5 (v 5.1 beta)

.

The sequences from the different species and different regions showed a high degree of homology (**Figure 23**).

There was variation between species on Alignment and phylogenetic analysis using

primers ANG12432 (locus AAGE02010470.1), ANG20760 (locus AAGE02008859.1) and ANG26425 (locus AAGE02021286.1) while there was no variation using primers ANG00026, ANG13935, ANG20362, ANG23972, ANG27523, ANG18326, and ANG04289 (**Appendix 2**).

There was variation between species on phylogenetic analysis (**Figure 24**) using primer ANG12432 amplicon sequences. This primer was however not able to cluster the sequences based on geographical origin of sample.

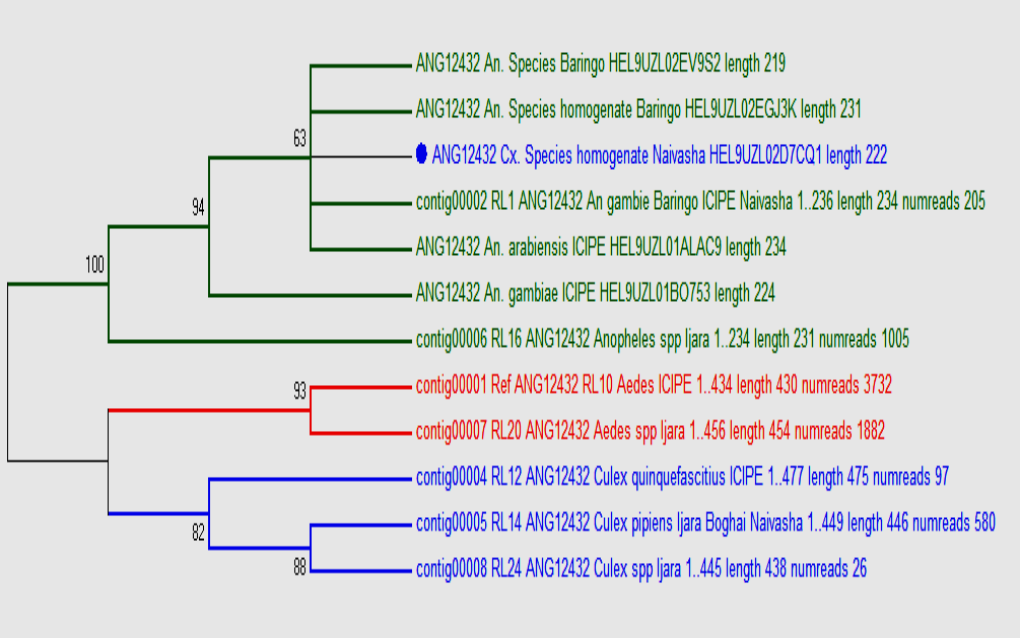

Figure 24: Molecular Phylogenetic analysis of ANG12432 by Neighbour-Joining method

The evolutionary history was inferred using the Neighbour-Joining method. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (500 replicates) are shown next to the branches. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The rate variation among sites was modelled

with a gamma. Analyses were conducted in MEGA5 (v 5.1 beta).

There was also variation between species on phylogenetic analysis using primer ANG20760 amplicon sequences (**Figure 25**).

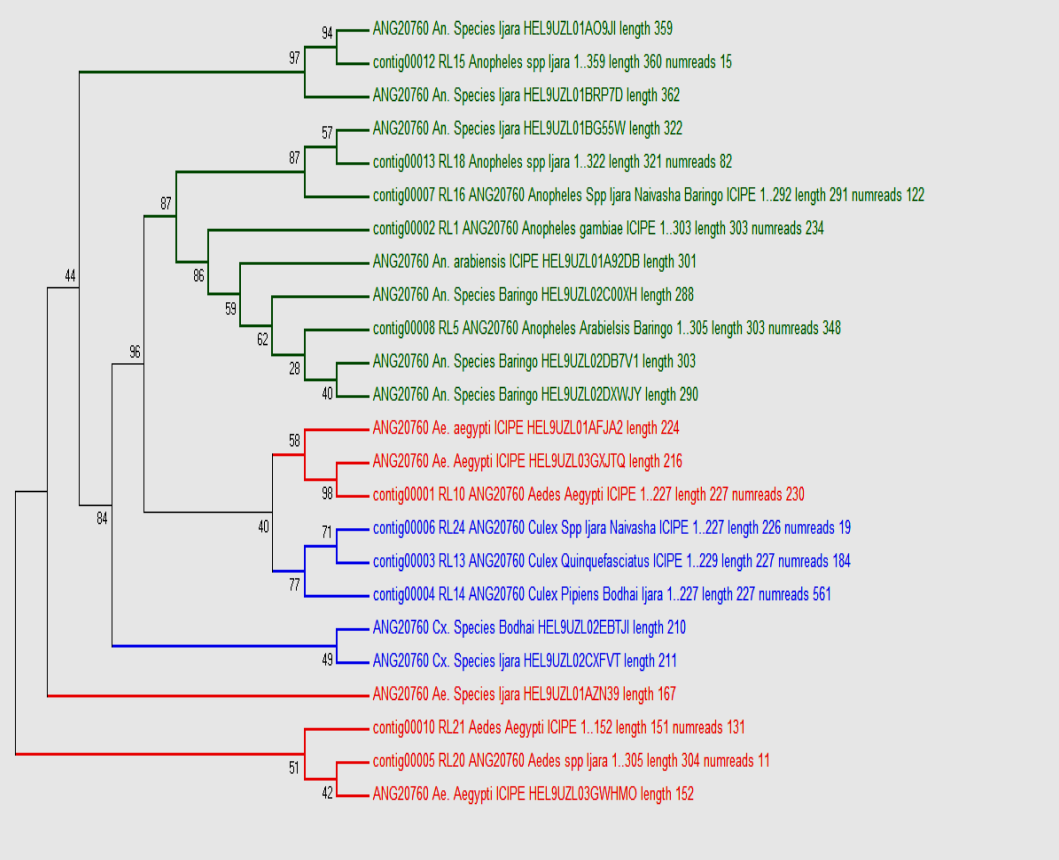

Figure 25: Phylogenetic tree of field collected mosquito samples sequenced after amplification using primer ANG20760

The evolutionary history was inferred using the Neighbour-Joining method. The bootstrap consensus tree inferred from 500 replicates is taken to represent the evolutionary history of the taxa analysed. Branches corresponding to partitions reproduced in less than 50% bootstrap replicates are collapsed. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (500 replicates) are shown next to the branches. The rate variation among sites was

modelled with a gamma distribution (shape parameter = 4). The analysis involved 24 nucleotide sequences. Analyses were conducted in MEGA5 (v 5.1 beta).

There was also variation between species on phylogenetic analysis using primer ANG26425 amplicon sequences (**Figure 26**).

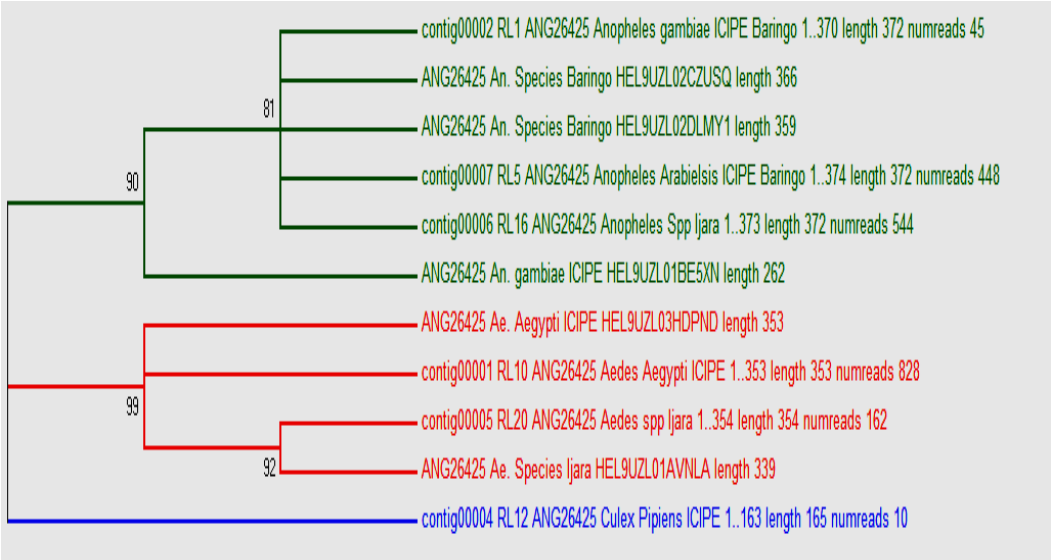

Figure 26: Phylogenetic tree build from field sample sequences amplified with primer ANG26425

The evolutionary history was inferred using the Neighbour-Joining method. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (500 replicates) are shown next to the branches. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The rate variation among sites was modelled with a gamma distribution (shape parameter = 4.5). The analysis involved 11 nucleotide sequences. Analyses were conducted in MEGA5 (v 5.1 beta).

Upon phylogenetic analysis the primer ANG00020 clustered the sequences derived

from mosquitoes from *icipe* into one clade away from all other field samples (**Figure 27**).

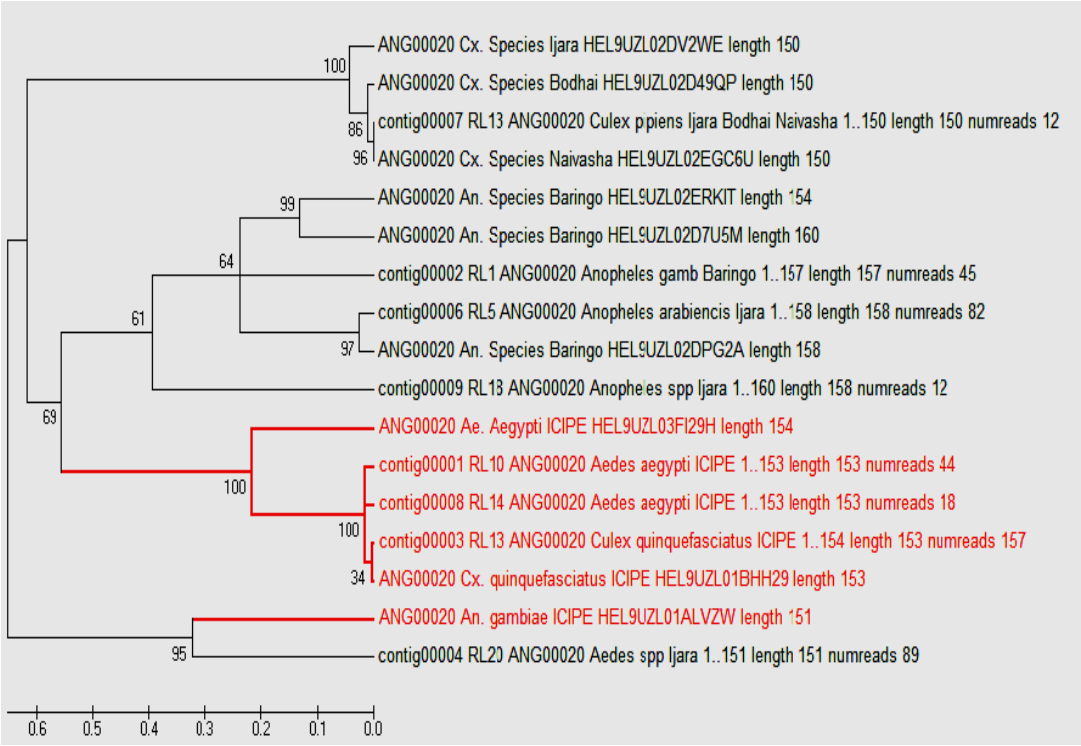

Figure 27: Phylogenetic tree build from field sample sequences amplified with primer ANG00020

The evolutionary history was inferred using the Neighbour-Joining method. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (500 replicates) are shown next to the branches. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The rate variation among sites was modelled with a gamma distribution (shape parameter = 4.5). The analysis involved 17 nucleotide sequences. Analyses were conducted in MEGA5 (v 5.1 beta).

The other primers designed against intergenic and intronic regions were not able to

differentiate the mosquitoes neither by species nor geographical locations. Some of the primers designed against intergenic regions, e.g. ANG13935 designed against locus AAGE02005924.1, produced more than one amplicon for one sample, suggesting amplification of more than one locus.

Phylogenetic trees of amplicon sequences derived from primers ANG ANG00026, ANG13935, ANG20362, ANG23972, ANG27523, ANG18326, and ANG04289 did not infer usefulness in either geographical or species clustering are shown in **Appendix 1** and **Appendix 2** for pilot and field studies respectively.

CHAPTER FIVE

DISCUSSION AND CONCLUSIONS

5.1 Discussion

Three primers (ANG12432, ANG20760 and ANG26425) designed against intronic regions separated the different mosquitos by genera studied. Of particular interest is primer ANG12432, designed from an intron with ensemble ID ENSANGT00000012432 (gene ID is AGAP000285) which separated *Aedes*, *Culex* and *Anopheles* mosquitoes visually based on their size on gel as well as by their nucleic acid sequence. The *Anopheles* genus was clearly distinguished from *Aedes* and *Culex* based on fragment size on gel. A confirmatory experiment was carried out on the same locus using the three different genera and the results on gel clearly distinguished *Anopheles* from *Aedes* and *Culex*. The primer ANG12432 is therefore useful for quick identification of the three mosquito genera on gel without the requirement to sequence the product. The loci ANG12432 may also be applicable in distinguishing other mosquito genera, but applicability requires experimental validation.

ANG26425 designed from an intron with ensemble ID ENSANGT00000026425 (gene ID AGAP012870) separated *Aedes*, *Culex* and *Anopheles* mosquitoes based on their nucleic acid sequence. The phylogenetic trees provide a distinct clusters on visualization, thus this locus is useful for classification of the genera under study. Nucleotide length variations on gel were however not distinct and thus would not be applicable in identification of the different mosquito genera at a gel level.

ANG20760 designed from intron with ensemble ID ENSANGT00000020760 (gene ID AGAP000429), was able to clearly distinguish all the three genera based on their sequences. Based on sequence alignment and phylogenetic analysis, this locus is able to separate the samples up to a species level in the *Anopheles* genera. This locus is however not useful for phylogoegraphic analysis as all samples of the same species from different regions cluster together. Phylogenetic analysis results show *Aedes* and *Culex* cluster close, suggesting they are more closely related than they are to the *Anopheles* genus.

ITS2 amplicon sequences revealed a high level of sequence similarity within and between the three genera studied (*Anopheles*, *Aedes* and *Culex*) in both data sets. The sequences obtained from the pilot study samples and field study samples aligned very well showing high levels of conservation. ITS2 has been used in some cases as a tool for phylogenetic analysis of populations for instance in a study by Beebe *et al*. (1999) which focused on malaria transmitting *Anopheles farauti* mosquitoes of south west pacific. Their study concluded that homogenization of the ITS2 regions is relatively slow and thus it can be used in genetic studies of population distribution and structure in the study area. In this study of sample populations from different areas of Kenya, the ITS2 region was found to be highly conserved and thus would not be useful in phylogeographic studies. The samples from the Kenyan population were found to be highly genetically similar based on the ITS2 locus suggesting that populations have probably undergone a sequence homogenization.

In contrast to the general findings from other studies for instance by Beebe *et*

al.,1999; preliminary DNA sequence analysis of the ITS1 sequences from mosquito samples from RVF endemic regions in Kenya demonstrated extensive intra-specific and inter- specific polymorphisms. Variability in the ITS region was primarily confined to the ITS1 domain while the ITS2 region displayed a high level of conservation. ITS1 copies from individual mosquito samples of the same species differed highly in sequence composition. The sequences were found too divergent to be aligned using ClustalX/W software and/ or Bio-edit sequence alignment software and thus phylogenetic analysis was not done on the sequences obtained from this locus. A large degree of variability of ITS1 sequences obtained from both data sets (pilot study samples and field study sample set) was observed, thus precluding ITS1 for use as a phylogenetic tool or in identification due to the high level of intraspecific and interspecific sequence divergence in the of mosquitoes from RVF endemic regions in Kenya.

Beebe *et al*. (1999) reported regional differentiation of this spacer using PCR-RFLP in a study of population structure of malaria transmitting mosquito populations of South- West pacific, suggesting that ITS1 can be used in genetic studies of population distribution and structure. However, the current study of ITS1 locus for populations derived from RVF endemic regions in Kenya shows a high degree of variation, and thus disqualifying the locus as a tool for population structural analysis as well as species identification.

Sequences obtained from amplification using primer ANG00020 resulted in; all samples from *icipe* clustering separate from all field samples. On Blastn and Blastx

analysis (http://blast.ncbi.nlm.nih.gov/Blast.cgi), the extreme ends of the ANG00020 primer amplicon sequences align to heat shock and unknown conserved hypothetical proteins of different mosquito species. These nucleotide regions also share homology to stress proteins of other insects and mammalian organisms. This suggests the primer may be amplifying a region that is likely to be influenced by the environment and may be applicable in differentiating between lab reared and field mosquitoes. This observation requires validation to confirm findings.

5.2 Conclusions

From this study, ITS1 and ITS2 were not able to provide useful phylogoegraphic information for the populations under investigation. ITS1 displayed a high level of variation within and between species at this locus while ITS2 displayed a very high level of sequence homology as there was a very high level of conservation of this locus across the species.

The locus ANG12432 is suitable for identification of the three mosquito genera from fragment size variation as well as by sequence variations. The loci (ANG20760 and ANG26425) were selected for identification of the three genera of mosquitoes as they separate the mosquitoes distinctly based on sequence variations as well as phylogeny. The ANG20760, ANG26425 and ANG12432 loci sequence data proved to be a useful tool for species identification and, potentially, to solve taxonomic problems. ITS2, ANG26425 and ANG20760 primers were found useful in phylogenetic analysis but were not able to distinguish species at the gel electrophoresis level.

ANG00020 primer suggested applicability in differentiating between lab reared and field mosquitoes. The phylogenetic tree generated from sequences derived from this primer clustered the *icipe* samples together but away from all field samples.

5.3 Recommendations

Based on the study, there is also the need for a review of the use of ITS1 and 2 in mosquito phylogenetic classification considering the present study did not find it appropriate as other studies may have.

ANG000020 primer, which was observed to phylogenetically separate field collected mosquitoes from lab reared mosquitoes, requires validation. This tool would be applicable when monitoring vector control measures being carried out in the field environment.

APPENDICES

Appendix 1: Phylogenetic trees from the pilot study for primers that were not able to sort samples by species

Phylogenetic trees as constructed from sequences in the pilot study sample set

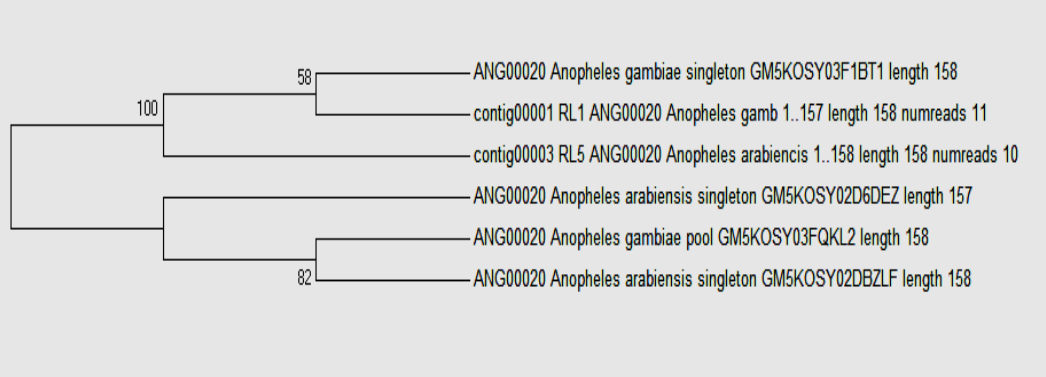

Figure 28: Phylogenetic tree of pilot samples amplified using ANG00020. This primer was only able to amplify *Anopheles* species

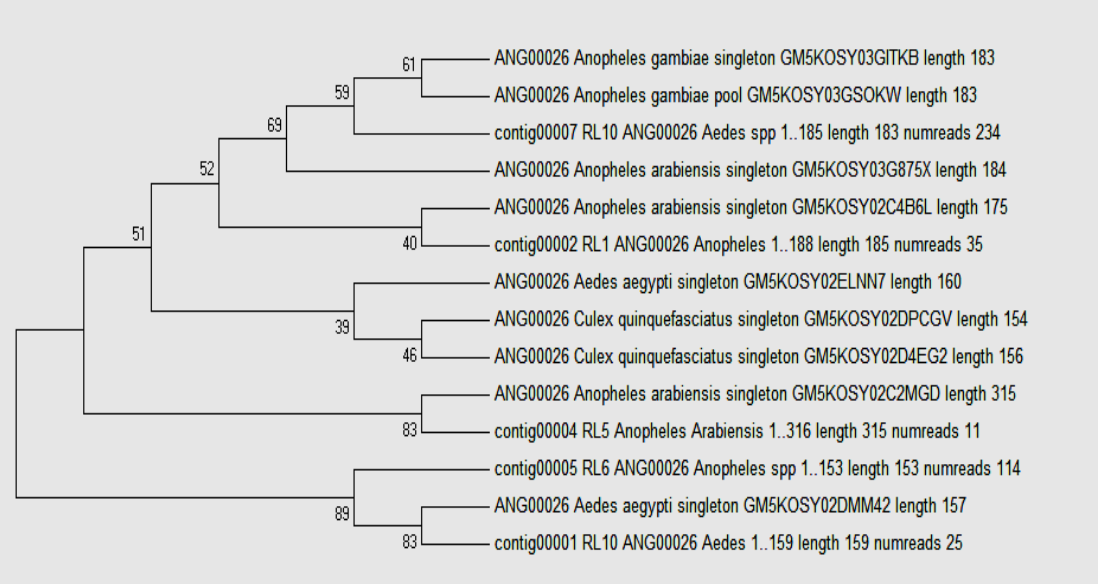

Figure 29: Phylogenetic tree of pilot samples amplified using ANG00026.

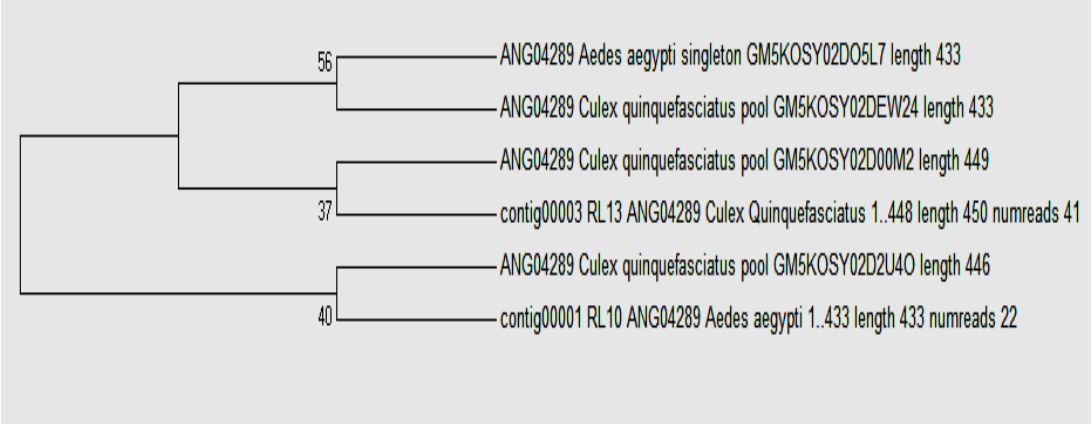

Figure 30: Phylogenetic tree of pilot samples amplified using ANG04289

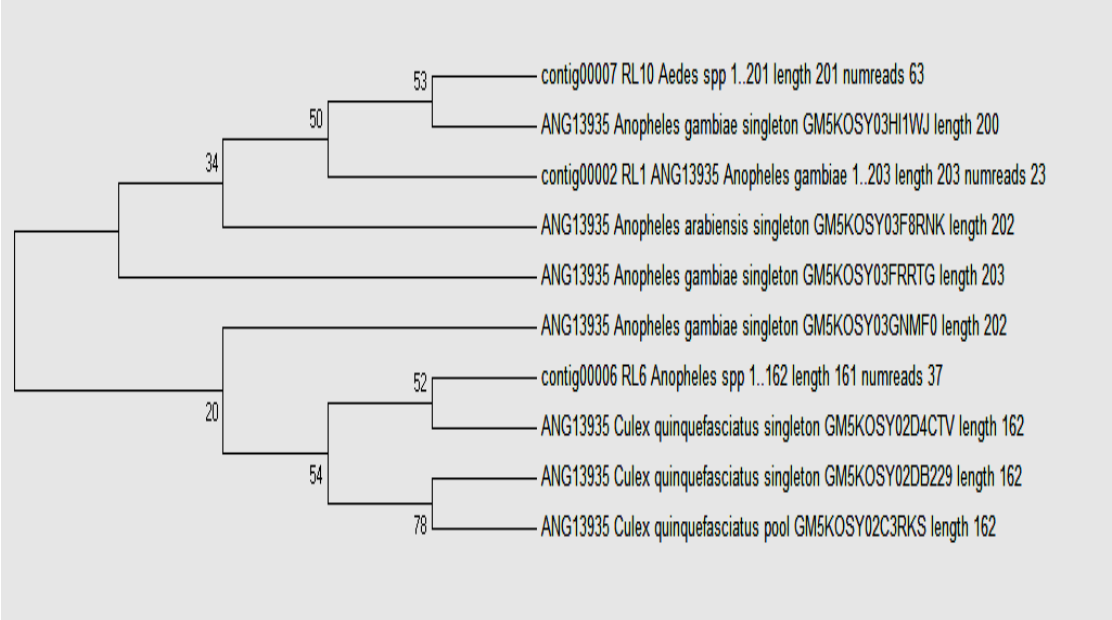

Figure 31: Phylogenetic tree of pilot samples amplified using ANG13935

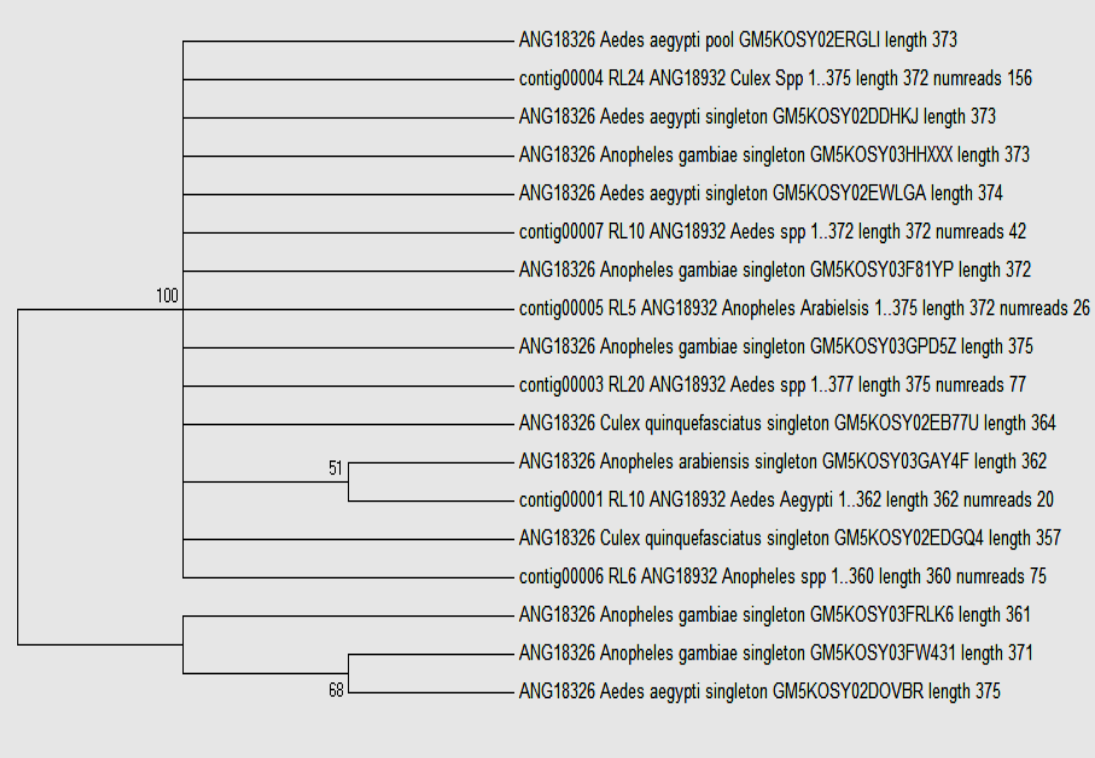

Figure 32: Phylogenetic tree of pilot samples amplified using ANG18326

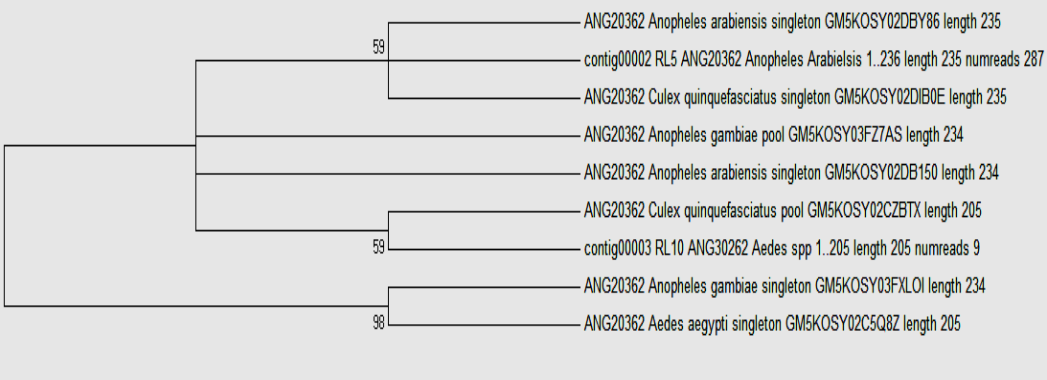

Figure 33: Phylogenetic tree of pilot samples amplified using ANG20362

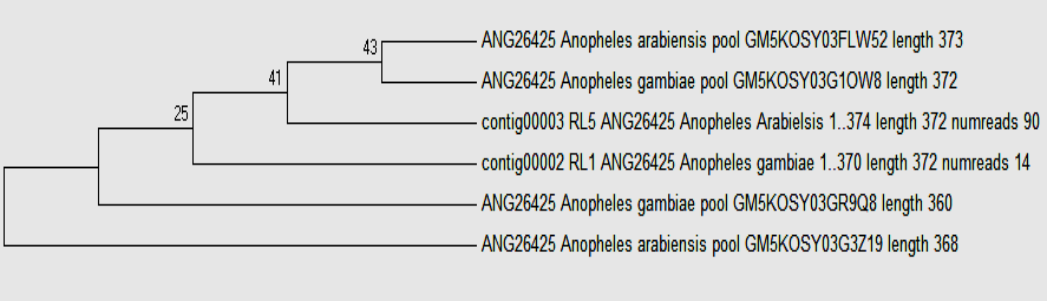

Figure 34: Phylogenetic tree of pilot samples amplified using ANG26425. This primer was only able to amplify *Anopheles* species

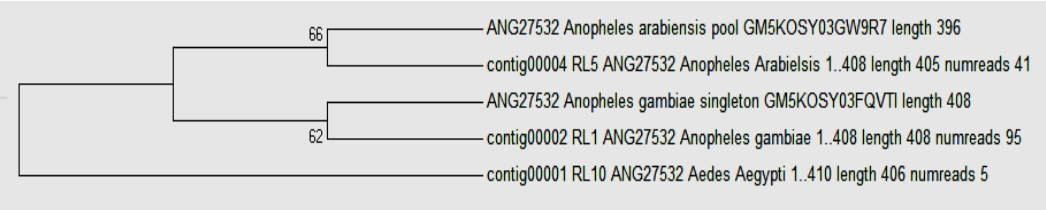

Figure 35: Phylogenetic tree of pilot samples amplified using ANG27532

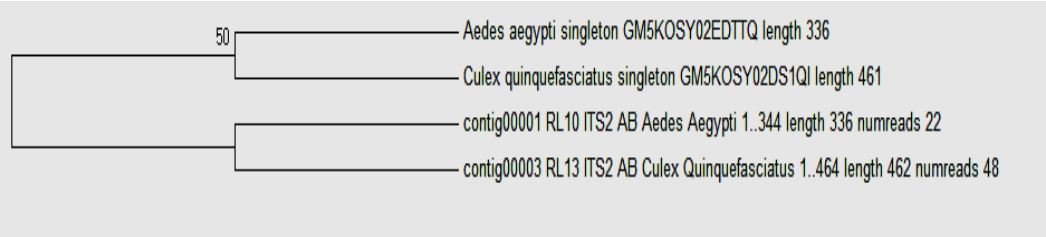

Figure 36 Phylogenetic tree of pilot samples amplified using ITS2AB

Appendix 2: Phylogenetic trees from the Field study for primers that were not able to sort samples by species and geographical location

Phylogenetic trees as constructed from sequences in the field study sample set

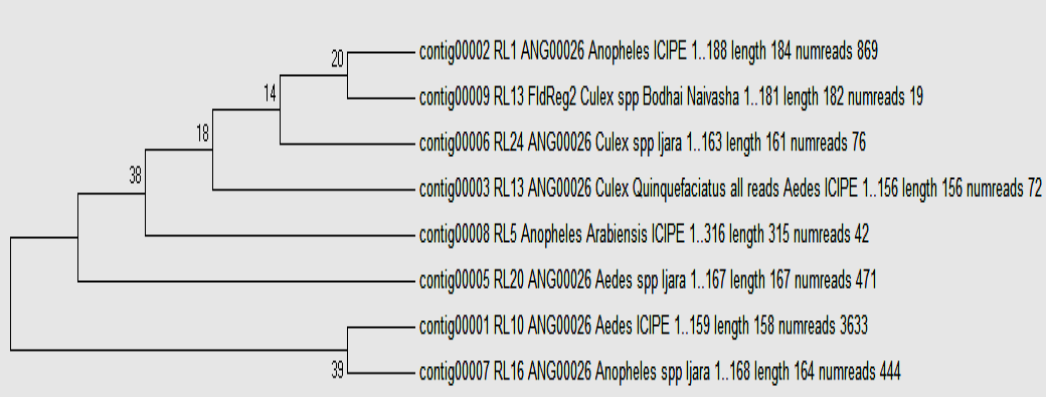

Figure 37: Phylogenetic tree of field samples amplified using ANG00026

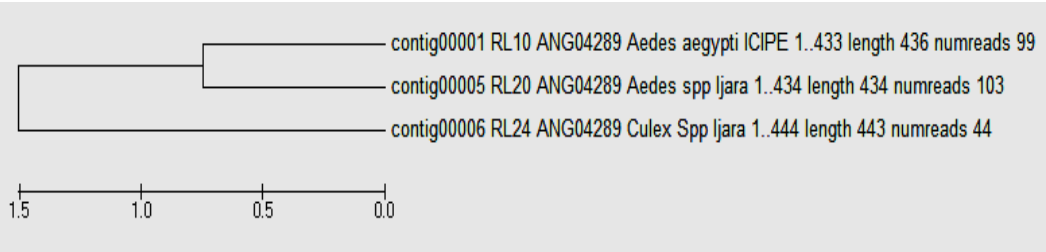

Figure 38: Phylogenetic tree of field samples amplified using ANG04286

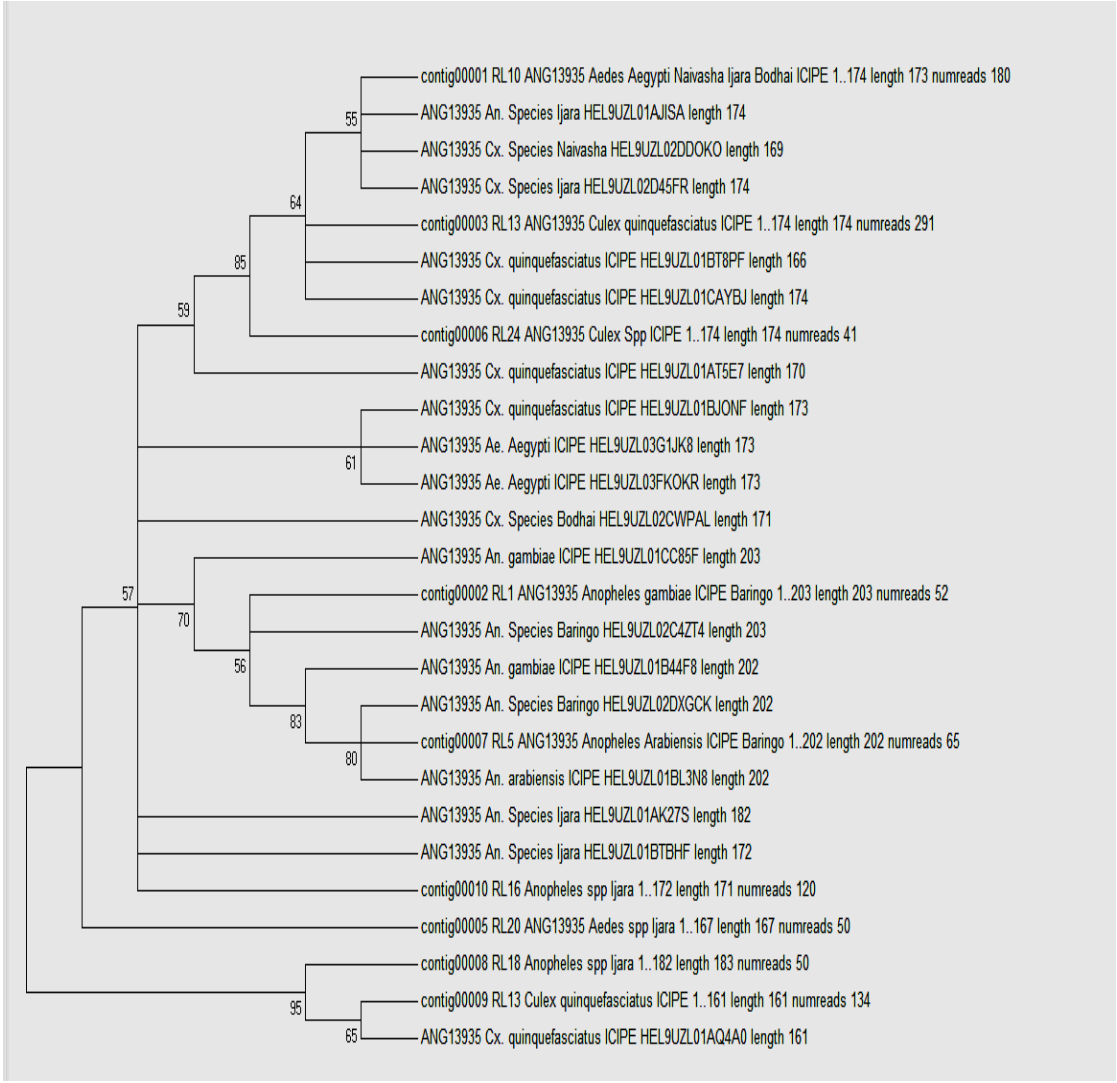

Figure 39: Phylogenetic tree of field samples amplified using ANG13935

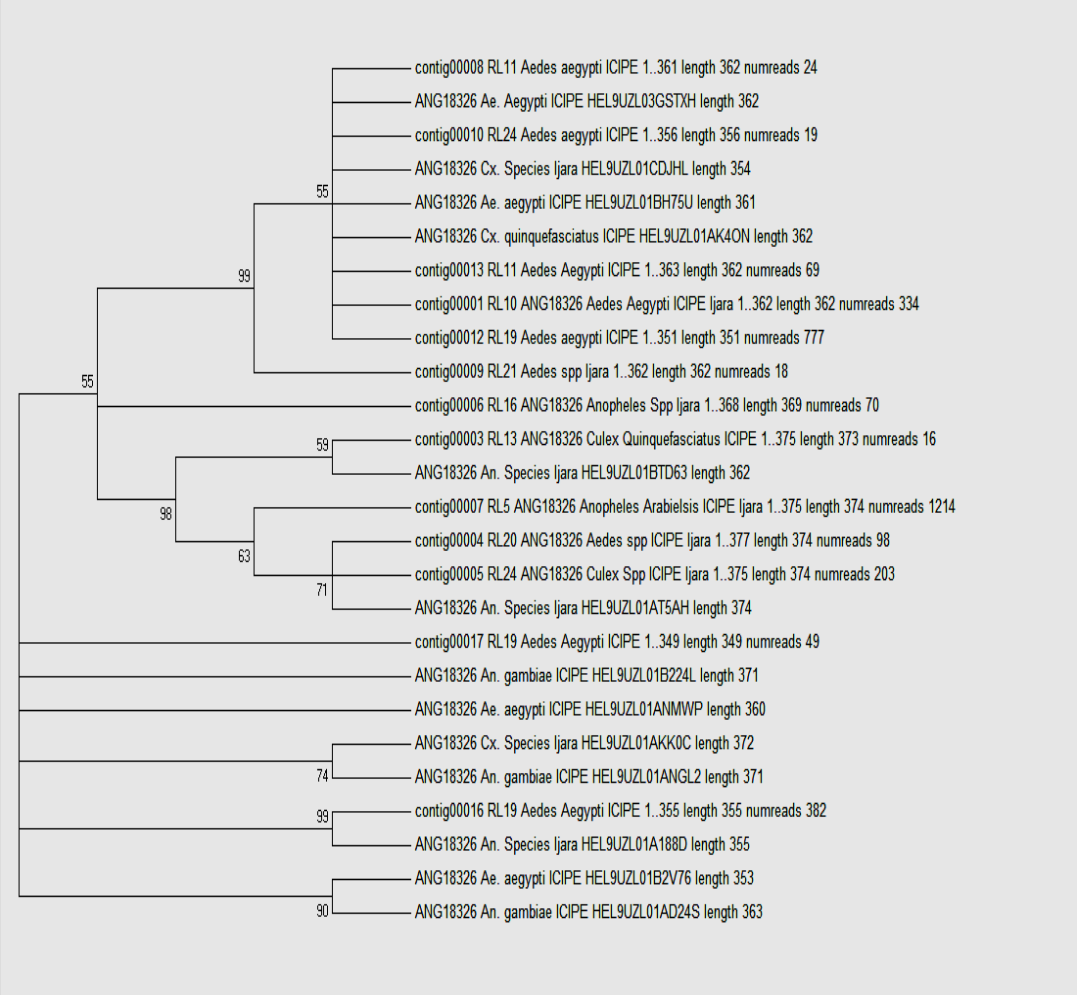

Figure 40: Phylogenetic tree of field samples amplified using ANG18326

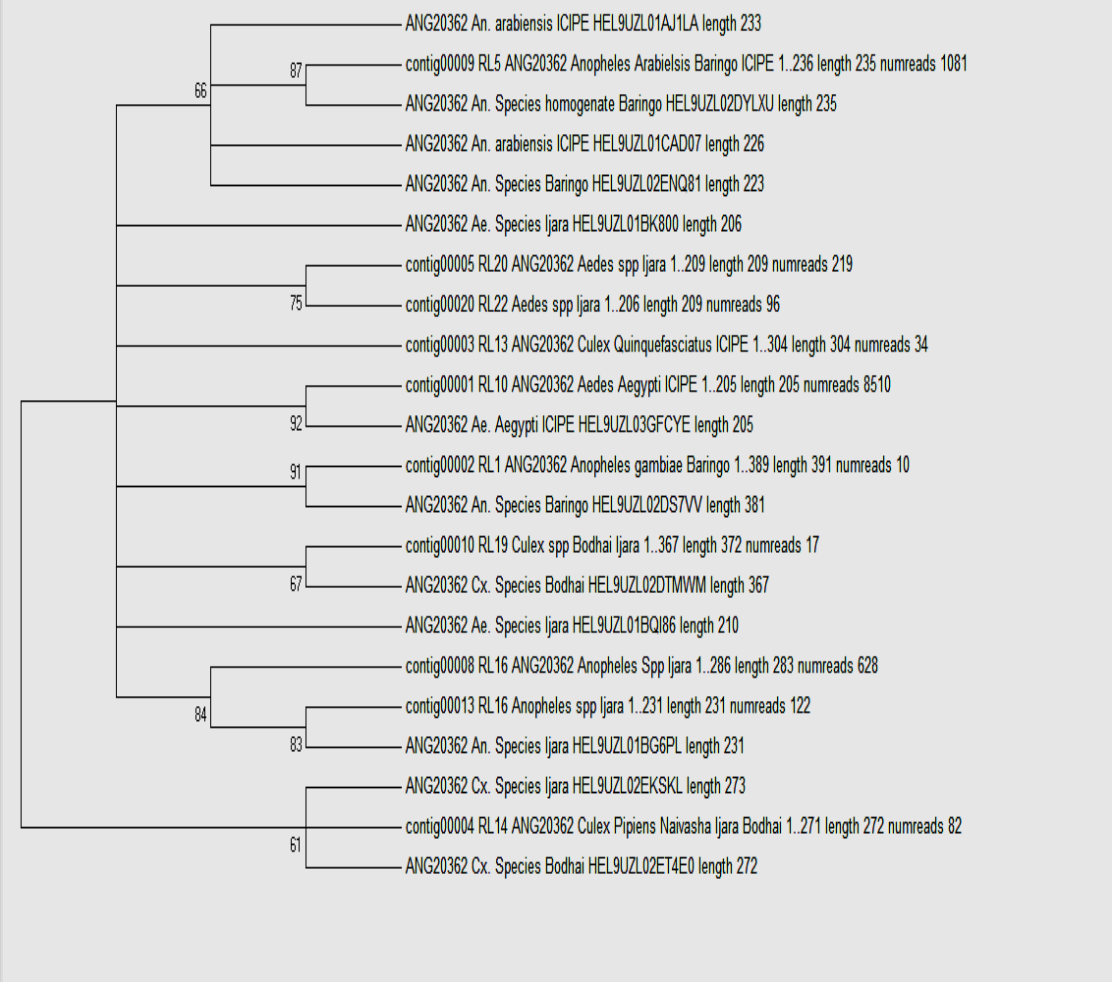

Figure 41: Phylogenetic tree of field samples amplified using ANG20362

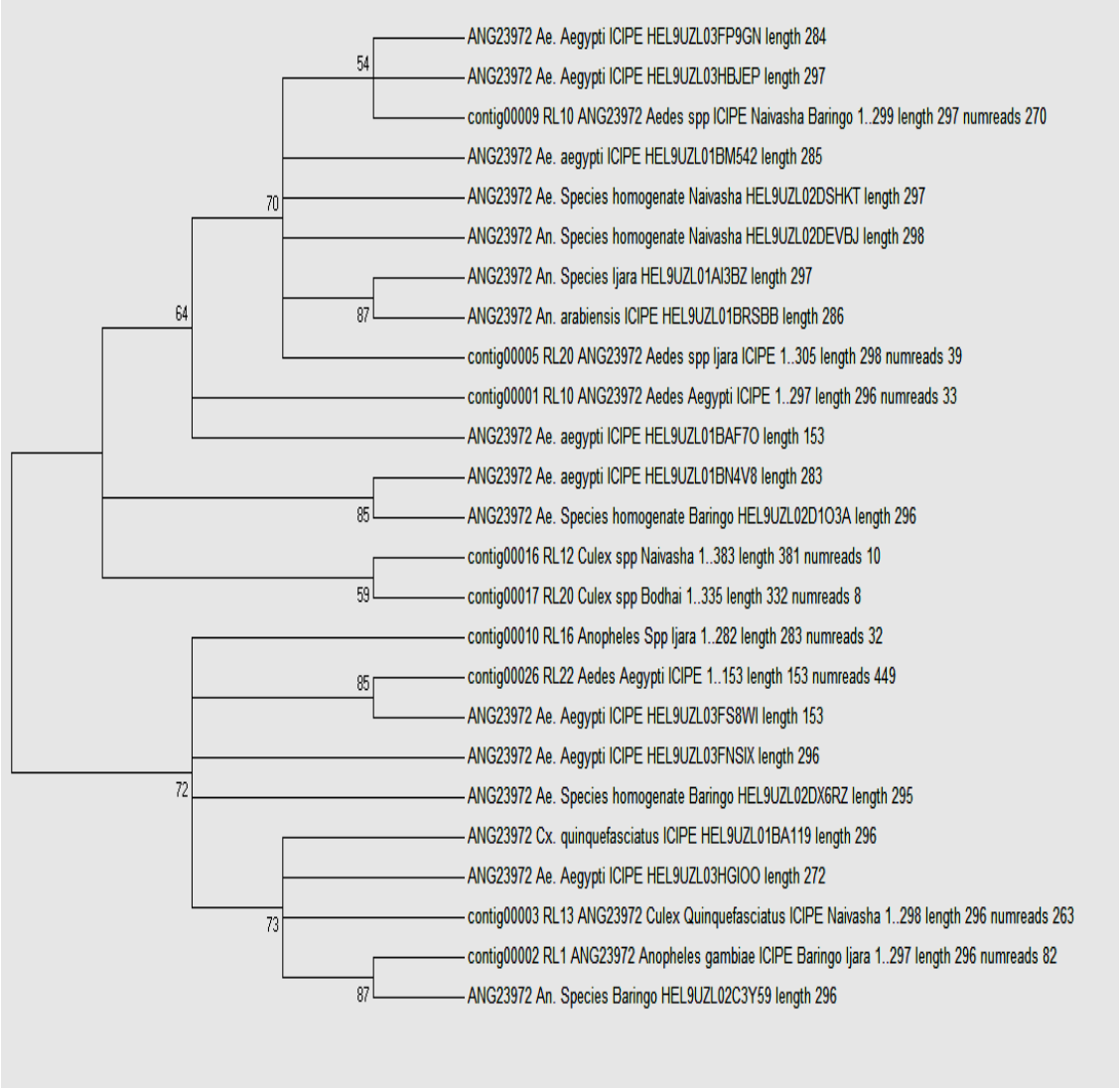

Figure 42: Phylogenetic tree of field samples amplified using ANG23972

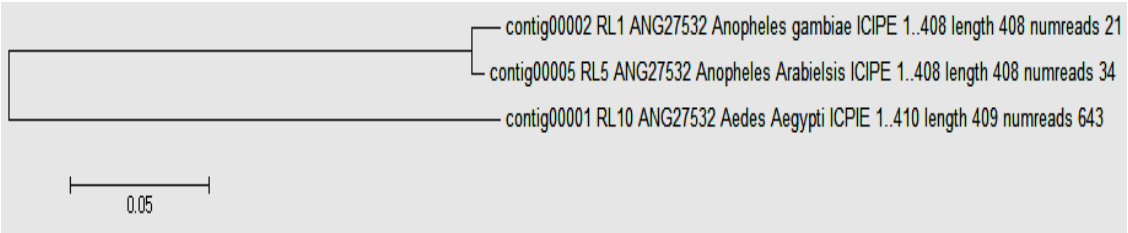

Figure 43: Phylogenetic tree of field samples amplified using ANG27532

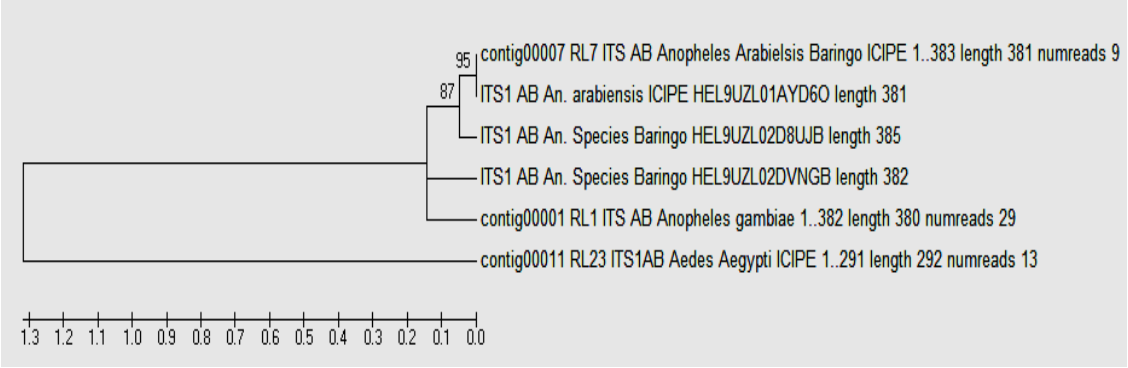

Figure 44: Phylogenetic tree of field samples amplified using ITS1_AB

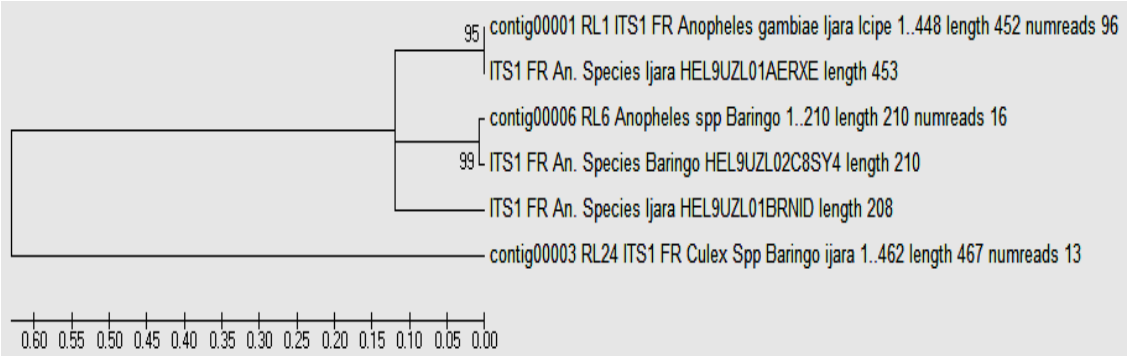

Figure 45: Phylogenetic tree of field samples amplified using ITS1_FR

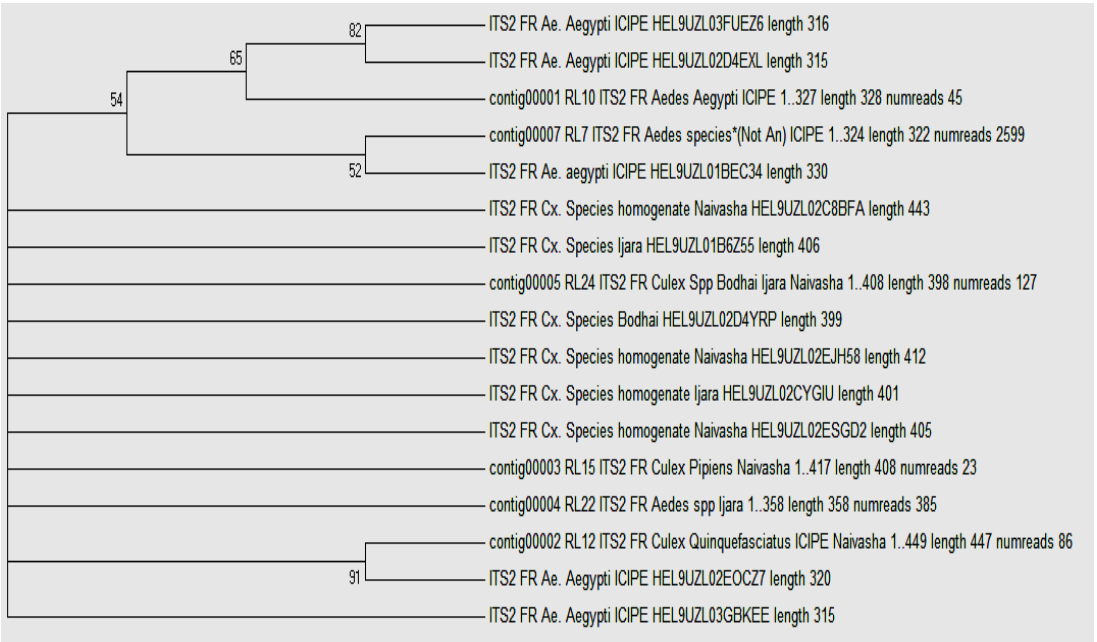

Figure 46: Phylogenetic tree of field samples amplified using ITS2_FR

REFERENCES

Andriamandimby, S. F., Randrianarivo-solofoniaina, A. E., Jeanmaire, E. M., Ravololomanana, L., Tiana, L., Rakotojoelinandrasana, T., Razainirina, J. *et al,* . (2010). Rift Valley Fever during Rainy. Emerging Infectious Diseases, 16(6), 963–970. doi:10.3201/eid1606.091266

Beebe, N. W., Cooperà, R. D., Foley, D. H., & Ellis, J. T. (2000). Populations of the south-west Paci ® c malaria vector *Anopheles farauti* s . s . revealed by ribosomal DNA transcribed spacer polymorphisms. Society, 84(November 1998), 244–253.

Blaxter, M., Mann, J., Chapman, T., Thomas, F., Whitton, C., Floyd, R., & Abebe, E. (2005). Defining operational taxonomic units using DNA barcode data. Philosophical transactions of the Royal Society of London. Series B, Biological sciences, 360(1462), 1935–43. doi:10.1098/rstb.2005.1725

Bower, J. E., Dowton, M., Cooper, R. D., & Beebe, N. W. (2008). Intraspecific concerted evolution of the rDNA ITS1 in *Anopheles farauti* sensu stricto (Diptera: Culicidae) reveals recent patterns of population structure. Journal of molecular evolution, 67(4), 397–411. doi:10.1007/s00239-008-9161-x

Bunnels, J., & Murphy, L. (1961). Rift Valley Fever, A Review of Literature.

Caterino, M. S., Cho, S., & Sperling, F. a. (2000). The current state of insect molecular systematics: a thriving Tower of Babel. Annual review of entomology, 45, 1–54. doi:10.1146/annurev.ento.45.1.1

Cognolati, V., Tempia, S., & Abdi, A. (2006). Economic Impact of Rift Valley Fever on the Somali Livestock Industry and a novel surveillance approach in nomadic pastoral systems. 11th international symposium on veterinary epidemiology and economics. Retrieved from www.sciquest.org.nz

Cole, J. R., Wang, Q., Cardenas, E., Fish, J., Chai, B., Farris, R. J., Kulam-Syed-Mohideen, a S., *et al*. (2009). The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. Nucleic acids research, 37(Database issue), D141–5. doi:10.1093/nar/gkn879

Daubney, R., Hudson, J. R., & Garnham, P. C. (1931). Enzootic hepatitis or rift valley fever. An undescribed virus disease of sheep cattle and man from east africa. The Journal of Pathology and Bacteriology, 34(4), 545–579. Retrieved from http://dx.doi.org/10.1002/path.1700340418

Davies. (2006). Risk of a rift valley fever epidemic at the haj in Mecca, Saudi Arabia. Revue scientifique et technique (International Office of Epizootics), 25(1), 137–47. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/16796043

Davies, G. (2003). Recognizing rift valley fever. FAO, Rome (Vol. 17).

Davies, Kairo, A., Kabete, P. O., Bailey, C. L., & Detrick, F. (1985). Rift Valley fever virus (family Bunyaviridae , genus Phlebovirus). Isolations from Diptera collected during an inter-epizootic period in Kenya. Journal oh hygiene, Cambridge, 95, 197–209.

Dhananjeyan, K. J., Paramasivan, R., Tewari, S. C., Rajendran, R., Thenmozhi, V., Leo, S. V. J., Venkatesh, A., *et al*. (2010). Molecular identification of mosquito vectors using genomic DNA isolated from eggshells, larval and pupal exuvium. Tropical biomedicine, 27(1), 47–53.

Farajollahi, A., Fonseca, D. M., Kramer, L. D., & Marm Kilpatrick. (2011). "Bird biting" mosquitoes and human disease: A review of the role of *Culex pipiens* complex mosquitoes in epidemiology. Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases, 11, 1577–1585. doi:10.1016/j.meegid.2011.08.013

Filone, C. M., Hanna, S. L., Caino, M. C., Bambina, S., & Doms, R. W. (2010). Rift Valley Fever Virus Infection of Human Cells and Insect Hosts Is Promoted by Protein Kinase C Epsilon. Drugs, 5(11). doi:10.1371/journal.pone.0015483

Fontenille, D., Diallo, M., Thonnon, J., Digoutte, J. P., & Zeller, H. G. (1998). New Vectors of Rift Valley Fever in West Africa. Emerging Infectious Diseases, 4(2), 289–293.

Frontielle, D., Traore-Lamizana, M., Zeller, H., Mondo, M., Mawlouth, D., & Digoutte, J.-P. (1995). Short report: Journal of Medicine (Cincinnati), 52(5), 403–404.

Gerdes, G. . (2008). Rift Valley Fever. OIE Terrestrial Maual 2008 (pp. 182–185). doi:10.1007/978-1-84628-787-9_41

Gorokhova, E., Dowling, T., Weider, L., Crease, T., & Elser, J. (2002). Functional and ecological significance of rDNA. The Royal Society, (269), 2373–2379.

Grobbelaar, A., Weyer, J., Leman, P. A., Kemp, A., Paweska, J. T., & Swanepoel, R. (2011). Molecular epidemiology of Rift Valley fever virus in recent outbreaks in South Africa.

Hackett, B. J., Gimnig, J., Guelbeogo, W., Costantini, C., Koekemoer, L. L., Coetzee, M., Collins, F. H., *et al.* (2000). Ribosomal DNA internal transcribed spacer (ITS2) sequences differentiate *Anopheles* funestus and An . rivulorum , and uncover a cryptic taxon. Insect Molecular Biology, 9(March), 369–374.

Hassan, O. A., Ahlm, C., Sang, R., & Evander, M. (2011). The 2007 Rift Valley Fever Outbreak in Sudan. Methods, 5(9), 1–7. doi:10.1371/journal.pntd.0001229

Hughes-fraire, R., Hagerman, A., McCarl, B., & Gaff, H. (2011). Rift Valley Fever: An economic Assessment of Agricultural and Human Vulnerability (pp. 1–26).

Imam, Z. E. I., Karamany, E.-M., & Darawish. (1979). RVF outbreak in Egypt 1977.pdf. Bulletin of WHO, 57(3), 441– 443.

Jackson, L. E., Hobenberg, W. H. Van, Filby, E. L., Green, H. W., & Gies, R. (1926). Mosquito Control. American journal of public health New York NY 1912, 16(3), 258–262.

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., *et al*. (2005). Genome Sequencing in Open Microfabricated High Density Picoliter Reactors. Nature, 437(7057), 376–380.

Marrelli, M. T., Floeter-Winter, L. M., Malafronte, R. S., Tadei, W. P., Lourenço-de-Oliveira, R., Flores-Mendoza, C., & Marinotti, O. (2005). Amazonian malaria vector anopheline relationships interpreted from ITS2 rDNA sequences. Medical and veterinary entomology, 19(2), 208–18. doi:10.1111/j.0269-283X.2005.00558.x

Mcintosh, B. M., & Russell, D. (1980). Rift Valley Fever in Humans in South Africa. s. Afr. med, 58(April), 803.

Miller, B. R., Crabtree, M. B., & Savage, H. M. (1996). Phylogeny of fourteen *Culex* mosquito species, including the *Culex pipiens* complex, inferred from the internal transcribed spacers of ribosomal DNA. Insect molecular biology, 5(2), 93–107. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8673266

Milne, I., Bayer, M., Cardle, L., Shaw, P., Stephen, G., Wright, F., & Marshall, D. (2010). Tablet--next generation sequence assembly visualization. Bioinformatics (Oxford, England), 26(3), 401–2. doi:10.1093/bioinformatics/btp666

Moutailler, S., Krida, G., Schaffner, F., Vazeille, M., & Failloux, A. (2008). Potential Vectors of Rift Valley Fever Virus in the Mediterranean Region. Vector borne and zonotic diseases, 8(6). doi:10.1089/vbz.2008.0009

Musser, J., Burnam, S., & Coetzer, J. A. (2006). Rift Valley Fever Symptoms. Veterinary Medicine.

Pages, N., Huber, K., Cipriani, M., Chenallier, V., Conraths, F., Goffredo, M., & Balenghien, T. (2009). SCIENTIFIC REPORT submitted to EFSA Scientific

review on mosquitoes and mosquito-borne diseases 1 Prepared by Nitu Pages. Regulation (pp. 1–96).

Palumbi, S. R., & Cipriano, F. (1998). Species identification using genetic tools: the value of nuclear and mitochondrial gene sequences in whale conservation. The Journal of heredity, 89(5), 459–64. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9768497

Pepin, M. P., Bouloy, B., Bird, B. H. B., Kemp, A. K., & Paweska, J. P. (2010). Review article Rift Valley fever virus (Bunyaviridae : Phlebovirus): an update on pathogenesis , molecular epidemiology , vectors , diagnostics and prevention. Vet. Res, 41(61). doi:10.1051/vetres/2010033

Rweyemamu, M., Paskin, R., Benkirane, A., Martin, V., Roeder, P., & Wojciechowski, K. (2000). Emerging Diseases of Africa and the Middle East. Iraq, 61–70.

Scott, Brogdon, W. G., & Collins, F. H. (1993). Identification Of Single Specimens Of The *Anopheles Gambiae* Complex By Polymerase Chain Reaction. American Journal of Trop. Med. Hyg, 49(4), 20.

Scott, W. G., Ralph, F. T., Kenneth, J. L., David, J. D., Digoutte, J. P., & Calvo-Wilson, M. A. (1992). Arbovirus isolations from mosquitoes collected during 1988 in the senegal river basin. American Journal of Trop. Med. Hyg, 47(6), 742– 748.

Service, M. W. (2000). Introduction to mosquitoes (Culicidae) ©. medical entomology for students. Retrieved from http://www.cambridge.org/052154775X

Shoemaker, T., Boulianne, C., Vincent, M. J., Pezzanite, L., Al-qahtani, M. M., Al-mazrou, Y., Khan, A. S., *et al*. (2002). Genetic Analysis of Viruses Associated with Emergence of Rift Valley Fever in Saudi Arabia. Emerging Infectious Diseases, 8(12), 1415–1420.

Shope, E. (1931). The spread of Rift Valley fever and approaches to its control *. Tropical Medicine, 579, 299–304.

Tang, J., Toè, L., Back, C., & Unnasch, T. R. (1996). Intra-specific heterogeneity of the rDNA internal transcribed spacer in the Simulium damnosum (Diptera: Simuliidae) complex. Molecular biology and evolution, 13(1), 244–52. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8583897

Theobald. (1901). Theobald F V. Journal of Tropical medicine.

Tolle, M. a. (2009). Mosquito-borne diseases. Current problems in pediatric and adolescent health care, 39(4), 97–140. doi:10.1016/j.cppeds.2009.01.001

Van der Sande, C., Kwa, M., Van Nues, R., Van Heerikhuizen, H., RauÃ, H., & Planta, R. (1992). pubmed_result. J Mol Biol, 20, 223(4).

Vuren, J., Kemp, A., Roux, C., Grobbelaar, A., Leman, P., Weyer, J., & Paweska, J. (2010). The 2010 Rift Valley fever outbreak in humans in South Africa: A laboratory perspective. ARBO- ZOONET annual meeting.

Walton, C., Sharpe, R. G., Pritchard, S. J., Thelwell, N. J., & Butlin, R. K. (1999). Molecular identification of mosquito species. Biological journal of the Linnean society, 68, 241–256.

WHO. (2010). Rift Valley Fever Rift Valley Fever, (November), 1–5.

WHO. (2006). Rift Valley Fever Rift Valley Fever. Factsheet No. 207, (November 2006), 1–5.

Woods, C. W., Karpati, A. M., Grein, T., Mccarthy, N., Gaturuku, P., Muchiri, E., Dunster, L., *et al*. (2002). An Outbreak of Rift Valley Fever in Northeastern Kenya , 1997 – 98. Emerging Infectious Diseases, 8(2).