PHYLOGENETIC AND PHYLOGEOGRAPHIC META-ANALYSIS OF CYTOCHROME C OXIDASE I BARCODE SEQUENCES OF EAST AFRICAN ARTHROPODS SUBMITTED INTO THE BARCODE OF LIFE DATABASE

KIBET GILBERT

A Research Proposal Submitted in partial fulfilment of the requirements for the Degree of Master of Science in Molecular Biology and Bioinformatics of Jomo Kenyatta University of Agriculture and Technology.

2018

DECLARATION

This research proposal is my original work and has not been presented elsewhere for a degree award

Signature………………………………………… Date: …………………………………………

Kibet Gilbert (HSB316-5106/2016)

Declaration by Supervisors

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

1. Dr. Steven Ger Nyanjom, Biochemistry Department, JKUAT

Signature: ……………………………………….... Date: ……………………………………

2. Dr. Jandouwe Villinger, Department of Molecular Biology and Bioinformatics, International Centre of Insect Physiology and Ecology, Kenya

Signature: …………………………………………… Date: …………………………………….

3. Dr. Scott E. Miller, Curator of Entomology at Smithsonian National Museum of Natural History, United States of America

Signature: …………………………………………… Date: …………………………………….

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS AND OR ACRONYMS

ABGD	Automatic Barcode Gap Discovery
A, C, T, G	Adenine, Cytosine, Thymine, Guanine
BIN	Barcode Index Number
BOLD	Barcode of Life Database
CBOL	Consortium for the Barcode of Life
COI	Cytochrome c Oxidase 1
DNA	Deoxyribonucleic Acid
ESC	External Connectivity System
FASTA	FAST-All file format
GIS	Geographical Information System
GPS	Geographical Positioning System
GYMC	General Mixed Yule-coalescent
INSDC	International Nucleotide Sequence Database Collaborative
MAS	Management and Analysis System
ML	Maximum Likelihood
MSA	Multiple Sequence Alignment
NJ	Neighbour Joining
PASTA	Practical Alignments using SATé and TrAnsitivity
PCR	Polymerase Chain Reaction
PCS	Projected Coordinate System
QGIS	Quantum Geographical Information System
RAxML	Randomized Axelerated Maximum Likelihood
TSV	Tab-Separated Values file format
XML	Extensible Markup Language file format

ABSTRACT

DNA barcoding has been used in identification, classification and discovery of organisms for over two decades. The establishment of The Consortium for the Barcode of Life, has proven DNA barcoding to be complementary, yet more efficient and accurate than morphology-based taxonomy. Two main databases host libraries of the species identifier DNA barcodes that are accessible to the public: The International Nucleotide Sequence Database Collaborative (INSDC) and the Barcode of Life Database (BOLD). BOLD focuses on storage and analysis of barcode sequences of all organisms in the domain Eukaryota. The standard gene region used as the barcode for most animal phyla except in Cnidaria is a 648 base-pair 5' region of the mitochondrial cytochrome c oxidase subunit 1 gene (CO1). This gene has a superior range of phylogenetic diversity signal than other mitochondrial genes and allows the discrimination of closely related species and even phylogenetic groups within a single species. Thousands of COI barcode sequences of voucher arthropods were submitted into the BOLD database from East Africa region in the past and over the past five years, most of which have not yet been analysed. In this project, phylum Arthropoda COI sequences will be subjected to comprehensive phylogenetic and comparative phylogeographic analysis. This study will determine the phylogenetic diversity and phylogeographic distribution of these voucher arthropods within eastern Africa to provide a detailed reference for future research on arthropods in the region. Further analyses will focus on identification of potential disease vector species among biting flies (Order Diptera) and crop pest species such as fruit flies (family Tephriditae) that have not yet been implicated with, but may have impact on human, animal and crop health.

CHAPTER ONE

.0 INTRODUCTION

1.1 Background information

The need to identify and classify organisms efficiently by biologists led to use of molecular (DNA) characteristics of organisms and ultimately assembly of reference libraries of DNA barcode sequences of species. Morphology-based identification systems are expensive, slow and take years to master considering the more than 100 million species in existence (Ebach & Holdrege, 2005; Hebert, Hollingsworth, & Hajibabaei, 2016). DNA barcodes are short, standardized genes or regions thereof employed in identification and discovery of species (Savolainen, Cowan, Vogler, Roderick, & Lane, 2005). The standard gene region used as the barcode for most animal groups is a 648 base-pair 5' region of the mitochondrial cytochrome c oxidase 1 gene ("CO1"). This region is particularly effective for identifying birds, butterflies, fish, arthropods and other animal groups. However, COI is ineffective when analysing plants ('What Is DNA Barcoding? « Barcode of Life', 2018). To aid the rapid and inexpensive identification of millions of species using DNA barcodes, the Consortium for the Barcode of Life (CBOL) was launched in May 2004 (Ratnasingham & Hebert, 2007). However, DNA barcoding remains a genetic key in identification of known species, rather than replacing traditional taxonomic practice that provides knowledge of the organism (Ebach & Holdrege, 2005).

Databases store the data acquired from laboratory analysis, and hence host the library of the species identifier barcodes, accessible to the public. The two main databases being, the International Nucleotide Sequence Database Collaborative (INSDC) (Karsch-Mizrachi, Takagi, & Cochrane, 2018), and Barcode of Life Database (BOLD) (Ratnasingham & Hebert, 2007). The former is a partnership among GenBank, the European Molecular Biology Lab in Europe, and the DNA Data Bank of Japan. BOLD, however, is maintained by the University of Guelph in Ontario. They all subscribe to CBOL's data standards for barcode records.

It has been established that mitochondrial gene cytochrome c oxidase I (COI) can effectively serve as the standard barcode for global identification of animals (Hebert, Cywinska, Ball, & deWaard, 2003; Lin & Danforth, 2004). It has been sufficiently demonstrated that COI profiles can parse apart sample organisms from higher taxonomic categories, into appropriate phylum or order (Hebert, Cywinska, et al., 2003). Also, comprehensive COI profiles can adequately assign species level identities to these samples (Hebert, Cywinska, et al., 2003). An experiment involving 200 closely related lepidopterans identified 196 (98%) specimens at over 3% divergence with the remaining 4 congeneric species pairs having a 0.6% – 2% divergence (Hebert, Cywinska, et al., 2003). This means that there is an opportunity through the COI barcode sequences available on BOLD database to develop a clear phylogeny, biodiversity and phylogeography of the arthropods within East Africa.

In this proposal data from BOLD database will be searched based on geographical regions, for East African and taxon, for Arthropoda phylum and exported in text format as Tap-Separated Values (TSV) and Extensible Markup Language (XML) file format. Multiple sequence alignment will be conducted using Practical Alignments using SATé and TrAnsitivity (PASTA) and phylogenetic trees inferred using a maximum likelihood approach. Phylogeographical analysis will be conducted based on a coalescent model of comparative phylogeography. All this analysis will be done targeting all available data sets for phylum arthropoda but specific attention will be given to biting flies and fruit flies both common vectors for diseases. The results will be useful as reference in future studies of these organisms.

1.2 Problem Statement

Thousands of COI barcode sequences of voucher arthropods submitted into the BOLD database[i] over the past decade have not been analysed carefully for phylogenetic and phylogeographic information. This massive amount of data has a lot of biological important information yet to be studied through bioinformatics reorganization and analysis. More importantly, phylum Arthropoda is one of the most populous and diverse phyla in the Kingdom Animalia. The lack of a comprehensive and detailed description of the arthropod diversity within the East Africa that may serve as a basis for other research work is an impediment.

It has been established that certain vectors like tsetse flies and *Anopheles* mosquitoes prefer specific geographical areas. This may be associated with altitude, climate, co-existing organisms (hosts) distribution and other factors. These factors also affect their evolution and should be scientifically examined through extensive phylogeographic distribution analysis of the voucher specimen. This may identify potential disease vectors and crop pests that are closely related to known vectors and pests.

1.3 Justification

The COI barcoding project spearheaded by CBOL since 2004 has generated a lot of data deposited into the BOLD database. A lot of these data have been extensively analysed for particular taxa from various regions around the world such as moths (Lepidoptera) (Brehm et al., 2016; Hajibabaei, Janzen, Burns, Hallwachs, & Hebert, 2006; Hebert, Cywinska, et al., 2003; Hebert, Ratnasingham, & Waard, 2003; S. Miller et al., 2015; S. Miller, Martins, Rosati, & Hebert, 2014). (S. E. Miller, Hausmann, Hallwachs, & Janzen, 2016) by analysing inventories of Geometridae family (moths) described diversity of this megadiverse taxa and in the process parsing cryptic species. Phylogeographic analysis of African Maize stalk borer (*Busseola fusca*) species from West, Central and East Africa based on mitochondrial cytochrome b resolved that domestication of sorghum and introduction of maize in the regions had no significant impact on their evolution and geographical distribution (Sezonlin et al., 2006). These few examples serve to emphasise the massive amount of information that can be derived from this simple phylogenetic and phylogeographic analysis of phylum Arthropoda COI barcodes. This study seeks to improve phylogeographic descriptions of arthropod diversity in East Africa, to identify the diversity of species that may not yet be recognized as potential crop pests or vectors of human and animal

[i] http://www.boldsystems.org/

diseases. This information will allow for better assessments of emerging threat to human, animal and crop health.

1.4 Research Questions

i. What is the phylogenetic diversity of phylum Arthropoda within the East Africa region based on COI barcode sequences available in BOLD database as compared to arthropod traditional taxonomical diversity already established based largely on ecological and morphological features?
ii. How are the phylum Arthropoda organisms phylogeographically distributed within East Africa based on GIS coordinates of voucher specimen hosted in BOLD database and what does that imply about their evolution?

1.5 Hypothesis

1.5.1 Research Question One

a) Null Hypothesis

The phylogenetic diversity of Arthropoda phylum is not comprehensively described by the current largely morphology and ecology-based systems, particularly for individual families.

b) Alternative Hypothesis

Arthropods are as phylogenetically diverse within East Africa, as described by traditional taxonomy that has been so far carried out in the region.

1.5.2 Research Question Two

a) Null Hypothesis

Specific Arthropoda phylum species phylogeography has not been influenced significantly by biotic and abiotic factors in their evolution and distribution in East Africa region.

b) Alternative Hypothesis

Arthropoda phylum species phylogeography has been greatly influenced by both biotic and abiotic factors alike in their evolution and distribution across East Africa at phylum and at species taxonomy levels.

1.6 Objectives

1.6.1 General Objectives

i. To reconstruct a series of phylogenetic trees based on COI barcode sequences of arthropod specimens hosted by BOLD database and determine the phylogeographic distribution of these organisms within East Africa.

1.6.2 Specific Objectives

i. To identify phylogenetic and phylogeographic distribution of arthropoda phylum and strategically focus on fruit flies (Tephritidae family) and biting flies (Diptera: True flies) within East Africa.

ii. To identify insect species that are potential pests or disease vectors from combined phylogenetic similarity and phylogeographic distribution.

iii. To identify any potential new classification or species, if any.

iv. To develop a rigorous analysis pipeline that has the potential of being automated.

CHAPTER TWO

2.0 LITERATURE REVIEW

2.1 DNA Barcoding Initiatives

There are four major components of the barcoding project (‘What Is DNA Barcoding? « Barcode of Life’, 2018). The key one being collection, identification and storage of voucher specimens in museums among other repositories. Secondly, Laboratory analysis to obtain DNA barcode sequences usually using specific primers via a standardized protocol. Thirdly, the database that stores the data acquired from laboratory analysis, and hence hosts the library of the species identifier barcodes, accessible to the public. The two main databases being, the International Nucleotide Sequence Database Collaborative (INSDC) (Karsch-Mizrachi et al., 2018), and Barcode of Life Database (BOLD) (Ratnasingham & Hebert, 2007). The former is a partnership among GenBank, the European Molecular Biology Lab in Europe, and the DNA Data Bank of Japan. BOLD, however, is maintained by the University of Guelph in Ontario. They all subscribe to CBOL's data standards for barcode records. The fourth component is data analysis, which allows identification of unidentified specimen by matching it to the closest record on the database or otherwise identification of a new species.

Funded by Alfred P. Sloan Foundation, the Consortium for the Barcode of Life (CBOL) was established in May 2004 as a collaborative effort between various organizations to foster international alliances to aid the development of a barcode library of eukaryotic life (Ratnasingham & Hebert, 2007). The enormity of records (over 1.5 million alone in animal kingdom) led to the need and development of the Barcode of Life Data Systems[ii] – a global online data management system for DNA barcodes (Hajibabaei, Singer, Hebert, & Hickey, 2007; Ratnasingham & Hebert, 2007).

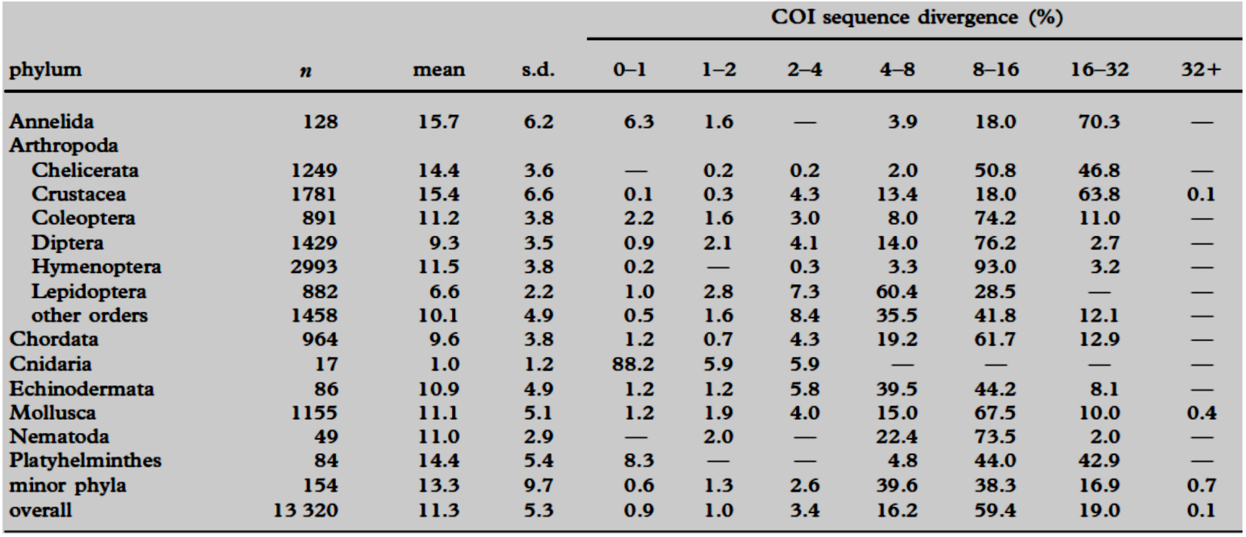
The works of a group of scientists lead by Paul Hebert at University of Guelph, Canada established that the mitochondrial gene cytochrome *c* oxidase I (COI) could serve as the barcode of a global bio-identification system for animals (Hebert, Cywinska, et al., 2003). Further research soon indicated that COI sequence divergence was common and could enable the discrimination of closely related species in all animal phyla (95%) except benthic Cnidarians, sponges and often species that hybridize (Hebert et al., 2016; Ratnasingham & Hebert, 2013). This success in species recognition was a result of a high rate of sequence mutation at COI in most animal groups (Hebert,

[ii] http://www.barcodinglife.org

Cywinska, et al., 2003; Hebert, Ratnasingham, et al., 2003; Singh, Khullar, & Jha, 2015).

The choice of mitochondrial genome of animals as the barcoding target over the nuclear genome comes with advantages (Hebert, Cywinska, et al., 2003). They lack introns, have limited exposure to recombination and a haploid mode of inheritance. Protein coding regions are chosen because the third nucleotide position of codons is weakly conserved (Hebert, Cywinska, et al., 2003). Long sequences of up to 600 base-pairs (bps) are chosen because the third nucleotide position in codons are strongly biased, A–T in arthropods, C–G in chordates. Second, because most nucleotide positions are constant in closely related species (Hebert, Cywinska, et al., 2003). Of the 13 mitochondrial genes, COI has two compelling advantages. First is that the universal primers to COI are robust enough to allow recovery of the 5' end of close to all animal phyla (Zhang & Hewitt, 1997). Second, COI has a superior range of phylogenetic diversity signal than other mitochondrial genes, because the evolution of this gene is rapid enough to allow the discrimination of closely related species and even phylogenetic groups within a single species (Cox & Hebert, 2001).COI sequence divergence from 13,320 congeneric species pairs from 11 phyla has been established to range from a low of 0.0% to a high of 53.7% (Hebert, Ratnasingham, et al., 2003).

Table 1: COI sequence divergence (%).



				COI sequence divergence (%)						
phylum	*n*	mean	s.d.	0–1	1–2	2–4	4–8	8–16	16–32	32+
Annelida	128	15.7	6.2	6.3	1.6	—	3.9	18.0	70.3	—
Arthropoda										
Chelicerata	1249	14.4	3.6	—	0.2	0.2	2.0	50.8	46.8	—
Crustacea	1781	15.4	6.6	0.1	0.3	4.3	13.4	18.0	63.8	0.1
Coleoptera	891	11.2	3.8	2.2	1.6	3.0	8.0	74.2	11.0	—
Diptera	1429	9.3	3.5	0.9	2.1	4.1	14.0	76.2	2.7	—
Hymenoptera	2993	11.5	3.8	0.2	—	0.3	3.3	93.0	3.2	—
Lepidoptera	882	6.6	2.2	1.0	2.8	7.3	60.4	28.5	—	—
other orders	1458	10.1	4.9	0.5	1.6	8.4	35.5	41.8	12.1	—
Chordata	964	9.6	3.8	1.2	0.7	4.3	19.2	61.7	12.9	—
Cnidaria	17	1.0	1.2	88.2	5.9	5.9	—	—	—	—
Echinodermata	86	10.9	4.9	1.2	1.2	5.8	39.5	44.2	8.1	—
Mollusca	1155	11.1	5.1	1.2	1.9	4.0	15.0	67.5	10.0	0.4
Nematoda	49	11.0	2.9	—	2.0	—	22.4	73.5	2.0	—
Platyhelminthes	84	14.4	5.4	8.3	—	—	4.8	44.0	42.9	—
minor phyla	154	13.3	9.7	0.6	1.3	2.6	39.6	38.3	16.9	0.7
overall	13 320	11.3	5.3	0.9	1.0	3.4	16.2	59.4	19.0	0.1

Table 1: Mean and standard deviation of the percentage sequence divergences at COI for 13 320 congeneric species pairs in 11 animal phyla (Hebert, Ratnasingham, et al., 2003).

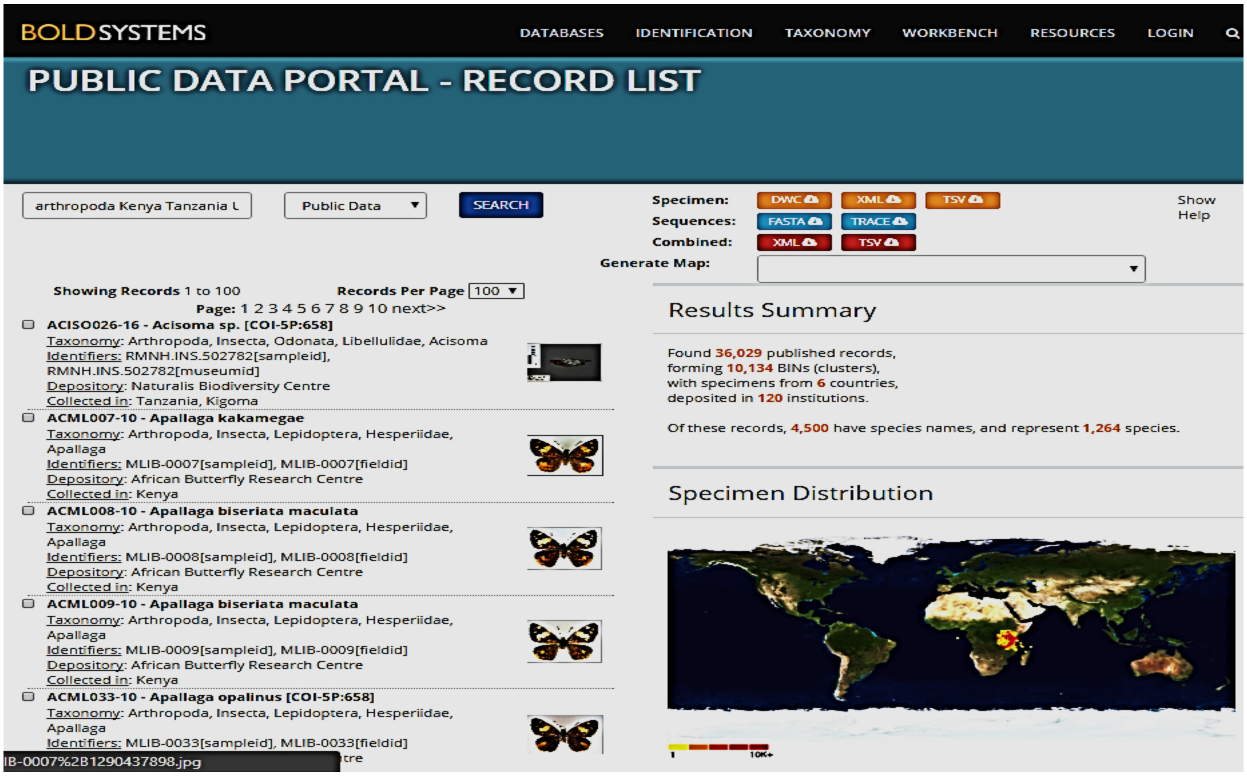
DNA barcoding besides offering an efficient scheme for species identification, has strongly aided taxonomic and biodiversity research. Hajibabaei, et al., (2006) using COI barcode sequences were able to effectively discriminate 97.9% of the 521 species from three Lepidoptera families identified prior by taxonomic work, encountering a few cases of interspecific sequence overlap and barcode clusters in 13 of presumably single species (Hajibabaei et al., 2006). An ever-growing library of DNA barcode sequences has led to more implications (Hajibabaei, Singer, et al., 2007). Current methods can identify unknown insect (class Insecta) and parse apart cryptic species using COI barcode sequences depending on thresholds of distances, sequence similarity cut-offs, or monophyly (Porter et al., 2014).

2.2 Datasets

BOLD hosts an integrated bioinformatics platform and maintains all stages of barcoding from specimen collection to highly validated barcode library. BOLD's Management and analysis system (MAS) facilitates data uploads, downloads and searches. The uploaded data has seven data elements namely (Ratnasingham & Hebert, 2007);

i. Species name (although this can be interim).
ii. Voucher data (catalogue number and institution storing).
iii. Collection record (collector, collection date and location with GPS coordinates).
iv. Identifier of the specimen.
v. COI sequence of at least 500 base pairs.
vi. Polymerase chain reaction (PCR) primers used to generate the amplicon.
vii. Trace files.

Data hosted within the BOLD system can be easily exported for analysis using different packages (Ratnasingham & Hebert, 2007). Data downloads can be done on multiple search criteria based on geographical regions, taxon, projects, and in text formats such as FASTA, TSV, XML and TRACE files containing sequence records, species name and sequence identifiers ('Databases | BOLDSYSTEMS', 2018; 'Record List | Public Data Portal | BOLDSYSTEMS', 2018).



Showing Records 1 to 100 Records Per Page 100 ▼

Page: 1 2 3 4 5 6 7 8 9 10 next>>

□ **ACISO026-16 - Acisoma sp. [COI-5P:658]**
Taxonomy: Arthropoda, Insecta, Odonata, Libellulidae, Acisoma
Identifiers: RMNH.INS.502782[sampleid], RMNH.INS.502782[museumid]
Depository: Naturalis Biodiversity Centre
Collected in: Tanzania, Kigoma

□ **ACML007-10 - Apallaga kakamegae**
Taxonomy: Arthropoda, Insecta, Lepidoptera, Hesperiidae, Apallaga
Identifiers: MLIB-0007[sampleid], MLIB-0007[fieldid]
Depository: African Butterfly Research Centre
Collected in: Kenya

□ **ACML008-10 - Apallaga biseriata maculata**
Taxonomy: Arthropoda, Insecta, Lepidoptera, Hesperiidae, Apallaga
Identifiers: MLIB-0008[sampleid], MLIB-0008[fieldid]
Depository: African Butterfly Research Centre
Collected in: Kenya

□ **ACML009-10 - Apallaga biseriata maculata**
Taxonomy: Arthropoda, Insecta, Lepidoptera, Hesperiidae, Apallaga
Identifiers: MLIB-0009[sampleid], MLIB-0009[fieldid]
Depository: African Butterfly Research Centre
Collected in: Kenya

□ **ACML033-10 - Apallaga opalinus [COI-5P:658]**
Taxonomy: Arthropoda, Insecta, Lepidoptera, Hesperiidae, Apallaga
Identifiers: MLIB-0033[sampleid], MLIB-0033[fieldid]
tre

Found 36,029 published records, forming 10,134 BINs (clusters), with specimens from 6 countries, deposited in 120 institutions.

Of these records, 4,500 have species names, and represent 1,264 species.

Figure 1: A screenshot of Public Data Portal on BOLD system website that allows data retrieval and downloads in various formats; FASTA, TSV, TRACE, et cetera ('Record List | Public Data Portal | BOLDSYSTEMS', 2018)

From the database, there are more than 36,000 published records of phylum Arthropoda that are associated to 10,134 BINs (clusters) from 6 countries within East Africa region (Kenya, Tanzania, Uganda, Rwanda, Burundi and South Sudan). Barcode Index Number (BIN) system aids in resolving redundancy that may occur as part of barcode taxonomy assembly (Ratnasingham & Hebert, 2013).

MAS also avails tools for routine data analysis, among which is the Taxon ID tree, that uses distance matrix from nucleotide sequences to generate a neighbour-joining (NJ) tree (Ratnasingham & Hebert, 2013).

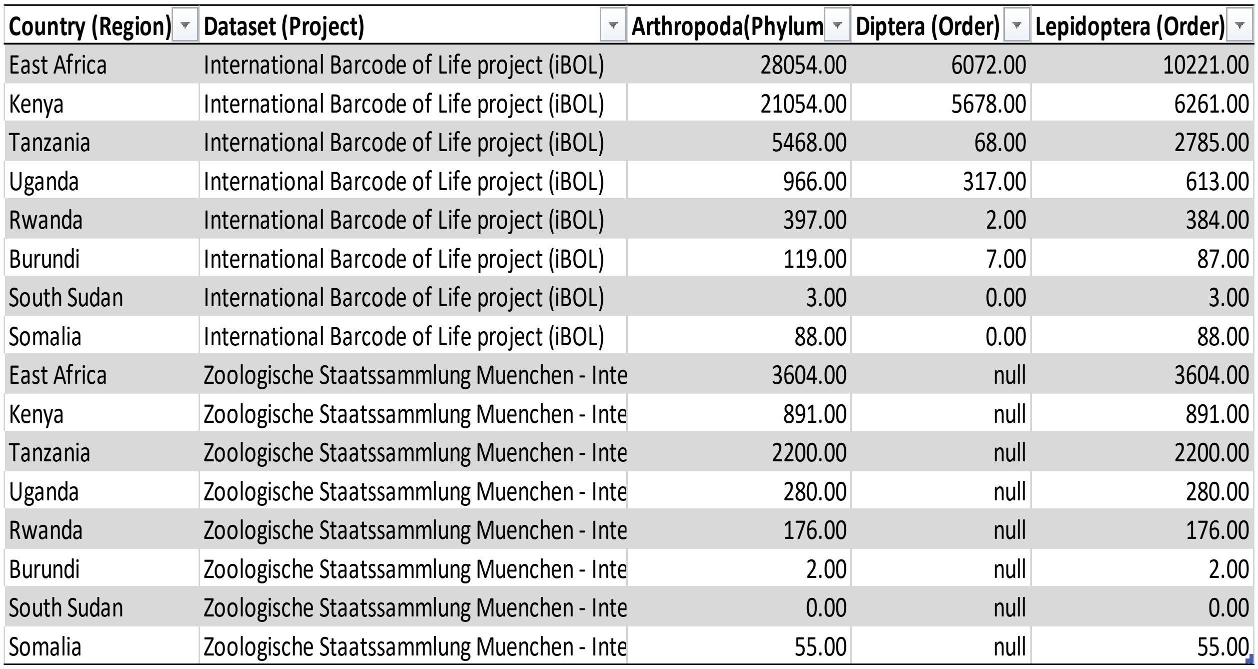
Various Barcode of Life projects have been conducted to cover particular taxa of Eukaryota clade, from Animalia kingdom, Diptera order, Lepidoptera and Tephritidae families, ants to plants (Hajibabaei et al., 2006; Hajibabaei, Singer, et al., 2007; Kang, Deng, Zang, & Long, 2017; S. Miller, Copeland, E Rosati, & Hebert, 2014; S. Miller et al., 2015; S. Miller, Martins, et al., 2014; Smith, Fisher, & Hebert, 2005). Their taxonomic sampling is done comprehensively and offer sufficient collection of data for phylogenetic studies on different branches of the Tree of Life.

Therefore, when phylogenies are constructed from barcode library projects, there is less likelihood of insufficient taxon sampling (Hajibabaei, Singer, et al., 2007; Zwickl, Hillis, & Crandall, 2002). Based on this fundamental, our project exploits data from Barcode of Life projects that have been conducted in East Africa.

One such project is Kenya Barcode of Life (KenBOL) (‘Kenya | iBOL’, 2018). It is a national initiative supported by Consortium for the Barcode of Life, Canadian Centre for DNA Barcoding and International Development Research Council and hosted by host institutions, International Centre of Insect Physiology and Ecology (ICIPE) and National Museums of Kenya (NMK). Under the regional representation of Dr. Daniel Masiga of ICIPE and associates, the project collected thousands of specimens and subsequently barcode sequences of many organisms that include vectors, pollinators, fishes, mammals, plants, birds, plant pests (fruit flies by Sunday Ekesi/ICIPE) and parasitoids (‘Kenya | iBOL’, 2018; ODENY, D. O, Ndungu, N., Masiga, D., Khayota, B., & Oyieko, H., 2017). Some of these data are publicly available through the public portal on BOLD system website, while others due to missing bits of information have not yet been published. This project was part of a global project International Barcode of Life Project (iBOL), activated in October 2010 with the first phase (2010-2015) aim to acquire DNA barcode records of 5M specimens representing 500K species. As for Arthropoda phylum there are other projects done by Dr Scott Miller particularly focusing on Lepidoptera (S. Miller, Copeland, et al., 2014; S. Miller, Martins, et al., 2014). A more robust break down of East African Arthropoda phylum datasets with suitable data published to the Global Biodiversity Information Facility database[iii] is shown in the table below;

Table 2: Global Biodiversity Information Facility Database East African Arthropoda Phylum records.

[iii] https://www.gbif.org/



Country (Region)	Dataset (Project)	Arthropoda(Phylum	Diptera (Order)	Lepidoptera (Order)
East Africa	International Barcode of Life project (iBOL)	28054.00	6072.00	10221.00
Kenya	International Barcode of Life project (iBOL)	21054.00	5678.00	6261.00
Tanzania	International Barcode of Life project (iBOL)	5468.00	68.00	2785.00
Uganda	International Barcode of Life project (iBOL)	966.00	317.00	613.00
Rwanda	International Barcode of Life project (iBOL)	397.00	2.00	384.00
Burundi	International Barcode of Life project (iBOL)	119.00	7.00	87.00
South Sudan	International Barcode of Life project (iBOL)	3.00	0.00	3.00
Somalia	International Barcode of Life project (iBOL)	88.00	0.00	88.00
East Africa	Zoologische Staatssammlung Muenchen - Inte	3604.00	null	3604.00
Kenya	Zoologische Staatssammlung Muenchen - Inte	891.00	null	891.00
Tanzania	Zoologische Staatssammlung Muenchen - Inte	2200.00	null	2200.00
Uganda	Zoologische Staatssammlung Muenchen - Inte	280.00	null	280.00
Rwanda	Zoologische Staatssammlung Muenchen - Inte	176.00	null	176.00
Burundi	Zoologische Staatssammlung Muenchen - Inte	2.00	null	2.00
South Sudan	Zoologische Staatssammlung Muenchen - Inte	0.00	null	0.00
Somalia	Zoologische Staatssammlung Muenchen - Inte	55.00	null	55.00

Table 2: East African arthropoda data published to the Global Biodiversity Information Facility. The data are from two iBOL projects: International Barcode of Life project (iBOL) and Zoologische Staatssammlung Muenchen - International Barcode of Life (iBOL) - Barcode of Life Project Specimen Data **(Roderic D. M. Page, 2016; 'Zoologische Staatssammlung Muenchen - International Barcode of Life (iBOL) - Barcode of Life Project Specimen Data', 2018).**

DNA barcoding endeavours to fast-track the inventory of biodiversity and make taxonomic information more accessible (Hebert et al., 2016; S. E. Miller, 2007; Riedel, Sagata, Suhardjono, Tänzler, & Balke, 2013). However, it provides opportunities for important investigations in other fields like phylogeny and ecology. DNA barcodes can be studied with established molecular phylogenetic and population biology tools in analysing biological relationships and diversity based on DNA sequences, hence attaining a comprehensive organization of species (Hajibabaei, Singer, et al., 2007). One benefit of this is that it aids in identification of novel barcode sequence for particular species' haplotype or geographical variant, and or signal the possibility of existence of a novel species. In the latter case detailed taxonomic analysis is done to ascertain new species (Hajibabaei, Singer, et al., 2007; Hebert, Cywinska, et al., 2003; Hebert, Ratnasingham, et al., 2003).

2.3 Phylogenetics

Molecular phylogenetics (Brown, 2002), a molecular (DNA or protein) based classification scheme of organisms into evolutionary relationships (phylogeny), has grown over decades and is important in the reconstruction of phylogenetic trees. DNA has since become the predominant molecule in this field as it yields more phylogenetic information than proteins due to redundancy in genetic code, the presence of non-coding regions and ease of preparation of DNA. The main aim of phylogenetic studies is to reconstruct an evolutionary-relationship tree of organisms under focus.

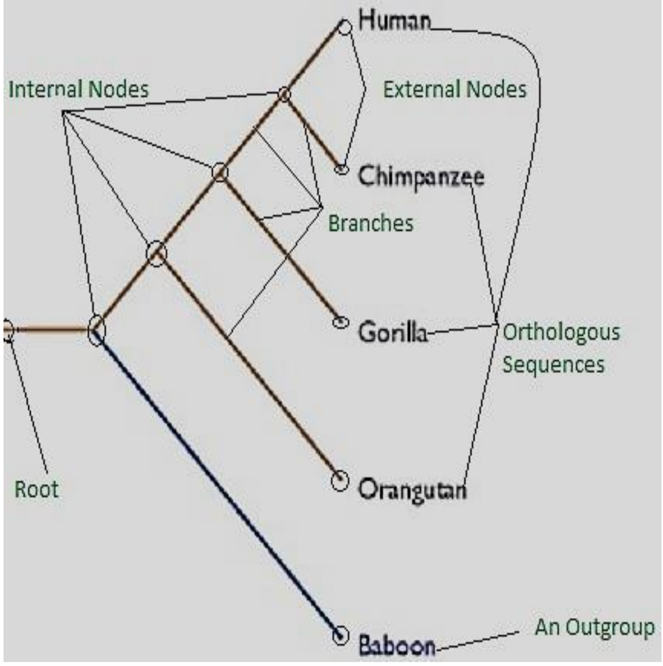


Figure 2: A typical phylogenetic Gene tree of orthologous sequences of Human, Chimpanzee, Gorilla, Orangutan and an out-group Baboon. It also depicts the topology of a phylogenetic tree; root, internal nodes, external nodes, branches and terminals.

A phylogenetic tree constructed from DNA sequences depicts the relationship between orthologous genes (Brown, 2002). The external nodes represent the genes and the internal nodes ancestral genes. The branches' length indicates the degree of difference between the genes. The rooting of the tree provides a guide to the evolutionary events that led to the genes. At least one out-group (a homologous sequence from an organism with proven paleontological split from the group of genes) is usually added to the gene collection to aid in selecting the right root and hence the right evolutionary pathway. The result is an inferred tree which is a close approximation of the series of evolutionary events. A true tree depicts the actual series of evolutionary events. However, most phylogenetic analysis has errors that result in an inferred tree different from the true tree. This requires a system to assign degrees of confidence to the branching pattern of an inferred tree (Brown, 2002).

The general timing and patterns of evolution of insect lineages, flight adaptation, and holometabolous diversification have been largely elucidated, from phylogenomic analysis of 1478 protein coding genes (Misof et al., 2014). The aim of DNA barcoding is species delimitation and identification however, COI barcodes can provide powerful phylogenetic information (Hebert, Ratnasingham, et al., 2003). Not all molecular markers are suitable for inferring phylogenies and not all suitable markers are applicable for all groups of organisms (An, Patwardhan, Ray, & Roy, 2014). The CO1–COII region, a 2.3 kb sequence made up of interspersed, highly conserved inter-membrane and variable extramembrane sections, is found in bacteria and mitochondria (An et al., 2014; Roe & Sperling, 2007). Its use in phylogenetic inference prompt specific concerns, notably, the mutation rate should be optimum and not too fast as it may reach a saturated state (Roe & Sperling, 2007). This difficulty is augmented by base composition bias and particularly in protein coding genes (An et al., 2014; Roe & Sperling, 2007). Therefore, molecular markers need to be screened for their ability to resolve long-established quality phylogenetic relationships within clades (An et al., 2014). Roe & Sperling, 2007, evaluated COI–COII regions in species and between species pairs of Lepidoptera and Diptera for high phylogenetic informativeness, high nucleotide divergence and low saturation (low Transition/ Transversion ratio). Transversions occur much less frequent at low divergence compared to transitions but, this ratio approaches ½ as saturation increases. They concluded that to infer informative phylogenetic relations from COI-COII one needs to evaluate the nucleotide divergence and saturation of the particular taxon and maximize sequence length (Roe & Sperling, 2007).

The usefulness of mitochondrial COI–COII 2.3 kb DNA sequence in phylogeny studies of the genus *Papilio*, as a representative taxon, has been proven and used in phylogenetic analysis of its several clades (C. D. & V. P., 2016; Caterino & Sperling, 1999). An analysis of phylogenetic informativeness of three regions of 1,574 bp COI gene in Coreidae and Pentatomidae families of Heteroptera suborder found homogeneity between 62.4 - 64.4% and 58.6 - 63.4%, respectively. The authors concluded that the same regions differ in phylogenetic informativeness and even adding other regions do not guarantee improved results (Souza, Marchesin, & Itoyama, 2016). Zhou et al., (2016) based on a study in the order Trichoptera, where they compared the phylogenetic output from COI data to that of combined datasets, including rRNA, concluded that phylogenetic hypotheses from COI are worth reporting (Zhou et al., 2016). However, COI is evidently not an ideal marker for deep-level phylogenetics (Kjer, Blahnik, & Holzenthal, 2001) and single genes phylogenies may not reflect species phylogeny (John C. Avise, 1989).

A standard molecular phylogenetic project comprises selection of the target group (e.g. order or family), compilation of representative taxa, acquisition of barcode sequence information, and reconstruction of phylogenetic trees through Maximum Likelihood, Maximum Parsimony, or Bayesian analysis (Brown, 2002; Hajibabaei, Singer, et al., 2007). Selection of an appropriate choice and length of barcode sequence ensures absolution from gene-specific bias, whereas, taxon sampling should be well spread to reduce any homoplasy bias and improve the resolution of phylogenetic tree (Hajibabaei, Singer, et al., 2007; Hebert, Cywinska, et al., 2003; Hebert, Ratnasingham, et al., 2003).

2.4 Diversity and species discovery

DNA barcoding can be crucial in species discovery, particularly for little-studied groups, where it is efficiently used to recognize putative species (operational taxonomic units, OTUs), which are then targeted for taxonomic analysis (Hajibabaei, Singer, et al., 2007; Hebert et al., 2016; Kekkonen & Hebert, 2014). A species delimitation study by Mari K and Paul D N Hebert on Australian hypertrophine moths sequence data to test the congruence of OTUs from three analytical methods (ABGD, BIN, GMYC) revealed 124 OTUs from the then recognized 51 species in a fast and repeatable protocol (Kekkonen & Hebert, 2014). New genera and species of clearwing moth (Lepidoptera: Sesiidae) are constantly being discovered in East Africa based largely on non-DNA features (Agassiz & Kallies, 2018; Gorbunov & Gurko, 2017). The inclusion of DNA barcoding in their taxonomical revision and discovery could be of great value.

DNA barcoding can also be applied to evaluate biodiversity on enormous dataset and in expansive geographical settings (Brehm et al., 2016; Hebert et al., 2016; S. E. Miller et al., 2016). Miller, S. *et al.*, (2016) used three datasets two field and ecology-based inventories and one museum and taxonomic-based inventory of Geometridae family (moths) to analyse and describe diversity of this megadiverse taxa and in the process parsing cryptic species (S. E. Miller et al., 2016). A Biodiversity analysis on 14,603 Ecuadorian Andes geometrid moths DNA barcode samples revealed 1857 putative species, an 80% rise from an earlier morphology based study that resulted in only 1010 species (Brehm et al., 2016).

Virgilio, M. *et al.*, (2015) analysed phylogenetic relationships of fruit fly tribe Dacini (*Ceratitidina*, *Dacina*, *Gastrozonina*) based on four mitochondrial and one nuclear gene fragment and revealed the need to revise the regrouping of clades as suggested by phylogenetic analysis results (Virgilio, Jordaens, Verwimp, White, & De Meyer, 2015). Molecular phylogeny of 125 Dacini tribe (Diptera: Tephritidae) species (*Dacus* and *Bactrocera* genus) using 16S, COI, COII and white eye genes done by Matthew N. Krosch et al to test out-of-India hypothesis and revealed the need to adjust the taxonomy within Dacini tribe (Krosch et al., 2012). Virgilio, M. *et al.*, (2009) also did similar studies on African Dacus genus (Diptera: Tephritidae) based on two mitochondrial (COI, 16S) and one nuclear (period) gene fragments and revealed closely similar results (Virgilio, Meyer, White, & Backeljau, 2009). Fruit flies of five *ceratitis* species from eastern and southern regions of the afrotropical area have been described in detail by Meyer, M. *et al.*, (2016) based on integrative taxonomic studies depending on CO1 barcoding and morphological analysis (Meyer, Mwatawala, Copeland, & Virgilio, 2016). Meyer, M. et al., (2015) did a re-evaluation on an integrative approach to resolve Ceratitis FAR (Diptera, Tephritidae: *C. fasciventris*, *C. anonae* and *C. rosa*) crytic species complex (Meyer et al., 2015).

Massimiliano Virgilio, Ian White and Marc De Meyer developed a set of revised and optimised multi-entry identification keys for African fruit flies freely accessible to non-expert morphologists based on datasets from previous taxonomic revisions (Virgilio, White, & Meyer, 2014).

Glutathione (GSH), a ubiquitous antioxidant, was identified as a host marking pheromone (HMP), in the African fruit fly, *Ceratitis cosyra* Semiochemical and determined it effective at reducing oviposition response in *C. cosyra, C. rosa*, *C. fasciventris*, *C. capitata*, and *Zeugodacus cucurbitae* (Cheseto et al., 2017). Phylogenetic and phylogeographic information on these species among other closely related species may be helpful in determining and implementation of GSH as a biological control agent of these flies (Edmunds, Aluja, Diaz-Fleischer, Patrian, & Hagmann, 2010).

2.5 Phylogeography

The study of phylogeography, conceptualised in 1987 (John C. Avise et al., 1987), seeks to examine geographical distribution of genetic lineages within or among closely related species through special arrangement (John C. Avise, 2008; Dawson, S. Waples, & Bernardi, 2006). Phylogeography strongly integrates historical biogeography and population genetics and links micro-evolutionary events (mutations/haplotypes) to macroevolutionary events (speciation) (Bermingham & Moritz, 1998). Mitochondrial DNA is highly suited for this analysis in animals, because it accumulates substitutions several-fold faster than ntDNA. Resulting in higher nucleotide sequence variation (John C. Avise, 2008). Comparative phylogeography explores common genealogical patterns of multiple species within overlapping geographical regions and infers causes of genetic divergence and linking that to history (Dawson et al., 2006). Comparative phylogeography provides answers to the following questions: the extent to which biotic and abiotic factors affect long-term phylogenetic diversity in a region; and how demographic trends are shared amongst species over time and space (Gratton et al., 2017).

DNA barcoding data is useful in phylogeographic studies as the key inputs to a phylogeography study are molecular genetic data and georeference data. The specimen data accessible on the BOLD database are accompanied with GIS (Geographical Information System) coordinates. A good phylogeographic derivation is reliant on an excellent phylogenetic tree and only 'good' trees should be used (Smouse, 1998). The one major challenge to phylogeography is the molecular clock estimation of cladogenetic events, which largely depends on nucleotide substitution rates that vary across time, space and lineage (Smouse, 1998).

Through phylogeographic analysis of genes and specimens, biogeographic hypotheses are devised to explain evolutionary paths of population lineages and underlying mechanisms (Krosch et al., 2012; Kumar & Kumar, 2018). Studies show that tropical lineages have been stable through historical climatic-change cycles unlike temperate and polar communities that show extreme range fluctuations among codistributed taxa congruently or independently across taxa (Hickerson et al., 2010). Various biotic and abiotic influences are accountable for phylogeographic patterns (Kumar & Kumar, 2018). In simple terms, phylogeography examines how genetic diversity amongst genetic lineages interrelates to geography and geological events (Marske, 2016).

Considering our large georeferenced dataset, the most appropriate approach is to conduct comparative phylogeography and determine how the various lineages might have responded to different ecological events past and current (Gratton et al., 2017; Marske, 2016; Peter Linder H., 2017). While choosing the species or taxa to study, geographical area or method of study, abiotic and biotic factors that affect biogeographical processes response, and ultimately genetic diversity can be considered. These factors can be inferenced from available information on the biology and ecology of the target taxa (Gutiérrez-García & Vázquez-Domínguez, 2011). The majority of the datasets are georeferenced, however those lacking geo-references can be geocoded from referenced metadata (Gratton et al., 2017).

Mende *et al.* (2016) used a three-gene mt-dataset (889 specimens) and 12 microsatellite loci (892 specimens) to test and refute a morphology-based introgressive hybridisation hypothesis of the Eurasian *H. euphorbiae* and Afro-Macaronesian *H. tithymali* (Mende, Bartel, & Hundsdoerfer, 2016).Louise and colleagues conducted a comparative phylogeographic study of five mosquito species of the genus *Aedes* from Tanzania, Uganda and Benin based on two nuclear loci and mt-COI to examine of past climatic change on their evolution (Louise Bennett et al., 2018). A phylogeography study of the African phytophagous insect, *Busseola fusca*, using 307 individuals from 52 localities in West, Central and East Africa based on mitochondrial cytochrome b concluded that domestication of sorghum and introduction of maize in the regions had no significant effect on the insect's geograohical distribution (Sezonlin et al., 2006).

Nonetheless, limited phylogenetic and phylogeographic studies on arthropods in East Africa have been conducted and are largely restricted to specific taxa of importance, mainly Lepidoptera, Tephritidae and Culcidae families. This study will target Arthropoda phylum organisms within East Africa as a whole. However, more specifically, the study will apply some focus towards fruit flies (Tephritidae family) and biting flies (Diptera: True flies) within East Africa. Biting flies include black flies (Simuliidae), horse flies (Tabanidae), deer flies/yellow flies (Chrysops), tsetse flies (Glossinidae), stable flies (Muscidae), biting midges or no see-ums (Ceratopogonidae), botflies (as larvae, Oestridae), sand flies (Phlebotomidae: *Lutzomyia* and *Phlebotomus*), blow-flies (as larvae, Calliphoridae) and screw-worm flies (as larvae, Calliphoridae).

CHAPTER THREE

3.0 MATERIALS AND METHODS

3.1 Study design

The study will be done at the International Centre for Insect Physiology and Ecology (*icipe*), Nairobi. Permissions will be sought from Jomo Kenyatta University of Agriculture and Technology and *icipe*.

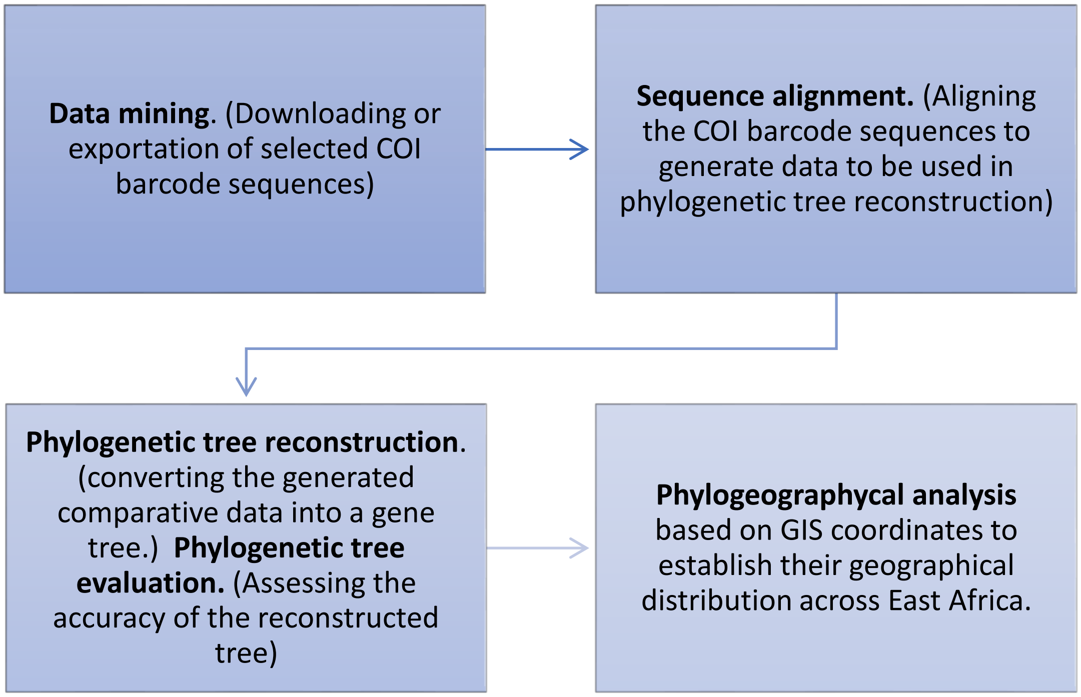

Data mining. (Downloading or exportation of selected COI barcode sequences)
Sequence alignment. (Aligning the COI barcode sequences to generate data to be used in phylogenetic tree reconstruction)
Phylogenetic tree reconstruction. (converting the generated comparative data into a gene tree.) Phylogenetic tree evaluation. (Assessing the accuracy of the reconstructed tree)
Phylogeographycal analysis based on GIS coordinates to establish their geographical distribution across East Africa.

Figure 3: Methodology workflow of the project.

3.2 Data mining

Data will be retrieved from the BOLD database. The Barcode of Life Data Portal (BDP), integrates latest biodiversity informatics tools with molecular barcoding data and provides the ability to extract data for particular geographical regions, in this case, East Africa (Sarkar & Trizna, 2011). The Public Data Portal data retrieval interface on the BOLD Database website facilitates searching of over 8 million specimen records hosted in BOLD, based on multiple search criteria that include, geography, taxonomy, and depository (‘Record List | Public Data Portal | BOLDSYSTEMS’, 2018). The Data Analysis Working Group (DAWG) of the Consortium for the Barcode of Life (CBOL) determines the standard input and output file formats and also develop some computational methods for barcode data analysis (Sarkar & Trizna, 2011). BDP provides a system “Aggregator” for collecting and developing datasets in various available formats such as FASTA, TRACE files, XML, TSV and others as depicted in **Error! Reference source not found.** (Sarkar & Trizna, 2011). Even more importantly, the taxonomy browser allows public access to over 6,000,000 records of the phylum Arthropoda, 61,103 in Kenya. Data from other countries within East Africa will also be retrieved ('Arthropoda | Taxonomy Browser | BOLDSYSTEMS', 2018).

The first step will be to clean the exported data and only retain relevant information. The general format of TSV file available on the BOLD Systems has 80 columns and as many rows as the records. Most of the columns have information that is not essential to the individual stages of our analysis. Some editing may require to be conducted using gEdit (Kurtenbach & Buxton, 1991) or other easily available programs like Vim text editor (Robbins, Lamb, & Hannah, 2008).

3.3 Sequence alignment

The COI orthologs obtained from exported arthropod data will be aligned and the DNA sequence distance scored (Brown, 2002). This critical step will be done only for homologous sequences for a proper phylogenetic tree with a common ancestral origin to be constructed. Accidental inclusion of non-homologous sequences will be avoided as that will lead to tragically flawed phylogenetic trees (Brown, 2002).

Multiple sequence alignment (MSA) algorithms will be used to align the thousands of sequences. MSA will particularly be important to reconstruct a valid phylogenetic tree. MSA will be done to the best accuracy. This is because of complex interrelationships between MSA and phylogenetic tree reconstruction causing mutual promotion and restraint of each other. Phylogenetic tree reconstruction algorithms always require MSA results as input data, while MSA algorithms sometimes require phylogenetic trees as guidelines (Mirarab et al., 2015; Zou, Hu, Guo, & Wang, 2015). Of the several MSA tools and algorithms existing, the most suitable in terms of speed, computational power needed and accuracy depending on the available dataset will be used comparatively. We will use PASTA, in comparison with SATé-II, T-Coffee, MUSCLE and MAFFT (Ghaleb, Reda, & Al-Neama, 2013; Mirarab et al., 2015; Zou et al., 2015). The best results will be used downstream in our analysis.

The MSA tool of choice is PASTA (practical alignments using SATé and TrAnsitivity) (Mirarab et al., 2015). PASTA exploits an iterative algorithm that has six steps. It begins with an alignment and tree estimation using Hidden Markov Model-based technique. Then it uses the tree estimate as a guide tree to divide the several sequences into subsets and build a tree with these subsets as nodes. Independent MSAs of the subsets are evaluated and paired based on their adjacency on the spanning tree and ultimately aligned. The resulting MSAs overlap and are merged using transitivity to generate the overall MSA (Mirarab et al., 2015).PASTA software runs on Linux operating system and it has been developed and tested entirely on Linux and MAC. The software is open source[iv].

SATé-II (Simultaneous Alignment and Tree estimation) is an MSA method designed to produce highly accurate results for large datasets (Liu & Warnow, 2014). like PASTA, it uses a similar iterative divide-and-conquer approach that uses a guide tree to split a data set into subsets then estimates alignment for individual subsets and finally merges the alignments. SATé-II iteratively divides the data set by estimating an alignment of the data with GTR+Gamma ML trees using RAxML on MAFFT and isolating the longest branch in the best GTR+Gamma ML scored-tree. Ultimately each subset has at most 200 taxa and so relatively few subset sequences that are less likely to be very divergent (Liu et al., 2012).

MUSCLE (multiple sequence comparison by log-expectation) is a progressive aligment tool with three basic stages (Edgar, 2004b). First a progressive-alignment-derived rooted guide tree is constructed from a pairwise k-mer distance matrix of unaligned sequences clustered through UPGMA. Secondly based on the guide tree, kimura distance matrix is computed for each internal node and clustered again using UPGMA. Only those subtrees whose branching order changed are progressively aligned again. This stage can be done in iteration. Thirdly, the tree is optimized iteratively by sequential bi-partitioning of the tree from the root and re-aligning the subsets to each other only retaining edges with higher SP scores (the sum-of-pairwise alignment scores) (Edgar, 2004a).

MAFFT is another MSA tool that uses both a progressive algorithm and an iterative technique to produce accurate and fast alignments. MAFFT uses fast Fourier transform (FFT) to rapidly detect homologous segments of sequences then implement two different heuristic alignment methods, a progressive method (FFT-NS-2) and an iterative refinement method (FFT-NS-i) to generate a MSA tree (Katoh, Misawa, Kuma, & Miyata, 2002; Katoh & Standley, 2013).

The alignment accuracy will be measured using FastSP based on two different systems: the SP-score and the modeler score, averaged together to get one measure (Mirarab et al., 2015).

[iv] https://github.com/smirarab/pasta

PASTA runs faster and analyses larger datasets than SATé-II the overall second best MSA tool available (Mirarab et al., 2015). PASTA requires much less computational resources than SATé-II particularly with large datasets.

3.4 Reconstruction of phylogenetic trees

Phylogenies will be inferred for the thousands of sequences from the MSA. Phylogenetic inference by maximum likelihood (ML) as opposed to distance matrix based, neighbour-joining (NJ) or Maximum Parsimony (MP) is the most accurate and offers excellent statistical (theoretical) data (Liu, Linder, & Warnow, 2011). In ML, evolution divergence is first modelled with tree topology and branch length probability parameter matrix, and the best tree with the highest likelihood is chosen. ML and Bayesian methods are based on stochastic models of sequence evolution with desirable statistical inference properties but with a high computational cost. For large sequences (hundreds to several thousands) statistical phylogeny estimation is best performed using maximum likelihood. Algorithms that use ML methods are many.

RAxML (Randomized Axelerated Maximum Likelihood) will be used as it is the most suitable for large-scale ML estimation. It produces the best ML scores in a shorter time than other ML methods with comparable ML score accuracy (Liu et al., 2011). RAxML utilizes heuristic approaches to minimize on search time (Munir, 2013). RAxML accepts data in Phylip format. RAxML 8 offers four different bootstrap algorithms and parallelization options for relatively different running times (Stamatakis, 2014).

FastTree will be used as a comparison to RAxML. FastTree is faster but with lower ML score and topological accuracy. However, FastTree can give better topology if the sequences are of poor alignments (Liu et al., 2011). FastTree 2 estimates the starting tree using neighbour joining and partially refine it using minimum-evolution nearest-neighbour interchanges (NNIs). It then further improves the tree using minimum-evolution subtree-pruning-regrafting (SPRs) and ML NNIs (Price, Dehal, & Arkin, 2010).

Phylogenetic informativeness of the COI barcodes for various clades will be calculated. As lineages diverge phylogenetic informativeness amass as mutations, however, saturation at a given site also increases significantly, especially with base bias, hence loss of phylogenetic informativeness (Roe & Sperling, 2007). Transition/Transversion ratio differ in direct proportion with saturation particularly in mitochondrial genes (Galtier, Enard, Radondy, Bazin, & Belkhir, 2006) and can be used to indirectly quantify saturation (Roe & Sperling, 2007). The

Transition/Transversion ratio will be calculated from the alignment and population diversity of the phylogenies (Fang et al., 2018).

There are a number of programs that can be used to visualize the phylogenetic trees. FigTree[v], will be used as it is the most suitable for large sequences. Others are Archaeopteryx[vi], Dendroscope[vii], Jstree[viii] or PhyloWidget[ix] (Munir, 2013).

3.5 Phylogeographic Distribution Analysis

Phylogeography enhances phylogenetic trees with geographical locations to permit interpretation of species evolution through space and time (Bouckaert, 2016). Comparative phylogeography has a two-phased approach, first being phylogenetic analysis of genetic data and second being comparative analyses for congruence in evolutionary and biogeographical histories between species (Gutiérrez-García & Vázquez-Domínguez, 2011). This analysis can detect dispersal, vicariance, population and demographic dynamics and hybridization and secondary contact cases, among others. From these findings, geographical, ecological, and biological hypotheses, inferred from biotic and abiotic factors influencing evolution, are evaluated within species and between species for congruence (Gutiérrez-García & Vázquez-Domínguez, 2011). One main factor that may influence their phylogeography is historical climatic (precipitation) cycling patterns.

In this study, a general comparative phylogeography will be conducted to assess all the available data based on their geographical co-distribution. To account for the implications of the randomness of gene trees a result of stochastic population level process, coalescent-based population genetic model will be implemented (Gutiérrez-García & Vázquez-Domínguez, 2011).

Specific focus will be on fruit flies and biting flies, which are linked by their host-parasite (biological/ecological) shared attributes. Not much is known yet about the genetic diversity and geographical distribution of most of the taxa that we will study. Therefore, we will use all the samples available to us for our study (Gutiérrez-García & Vázquez-Domínguez, 2011).

A number of Bayesian phylogeographical methods so far developed can merge phylogenetic analysis with geographical extrapolation (Bouckaert, 2016; Bouckaert et al., 2014; Maio, Wu, O'Reilly, & Wilson, 2015). BASTA (BAyesian STructured coalescent Approximation) (Maio et al., 2015), is phylogeographic inference technique implemented in Bayesian phylogenetic package

[v] http://tree.bio.ed.ac.uk/software/figtree/

[vi] www.phylosoft.org/archaeopteryx/

[vii] http://ab.inf.uni-tuebingen.de/software/dendroscope/

[viii] http://lh3lh3.users.sourceforge.net/jstree.shtml

[ix] www.phylowidget.org

BEAST2 (Bayesian Evolutionary Analysis by Sampling Trees) and it takes advantage of the accuracy of structured-coalescent based methods at a computational efficiency needed to handle larger molecular data (Maio et al., 2015). BEAST 2 and its packages are open source[x].

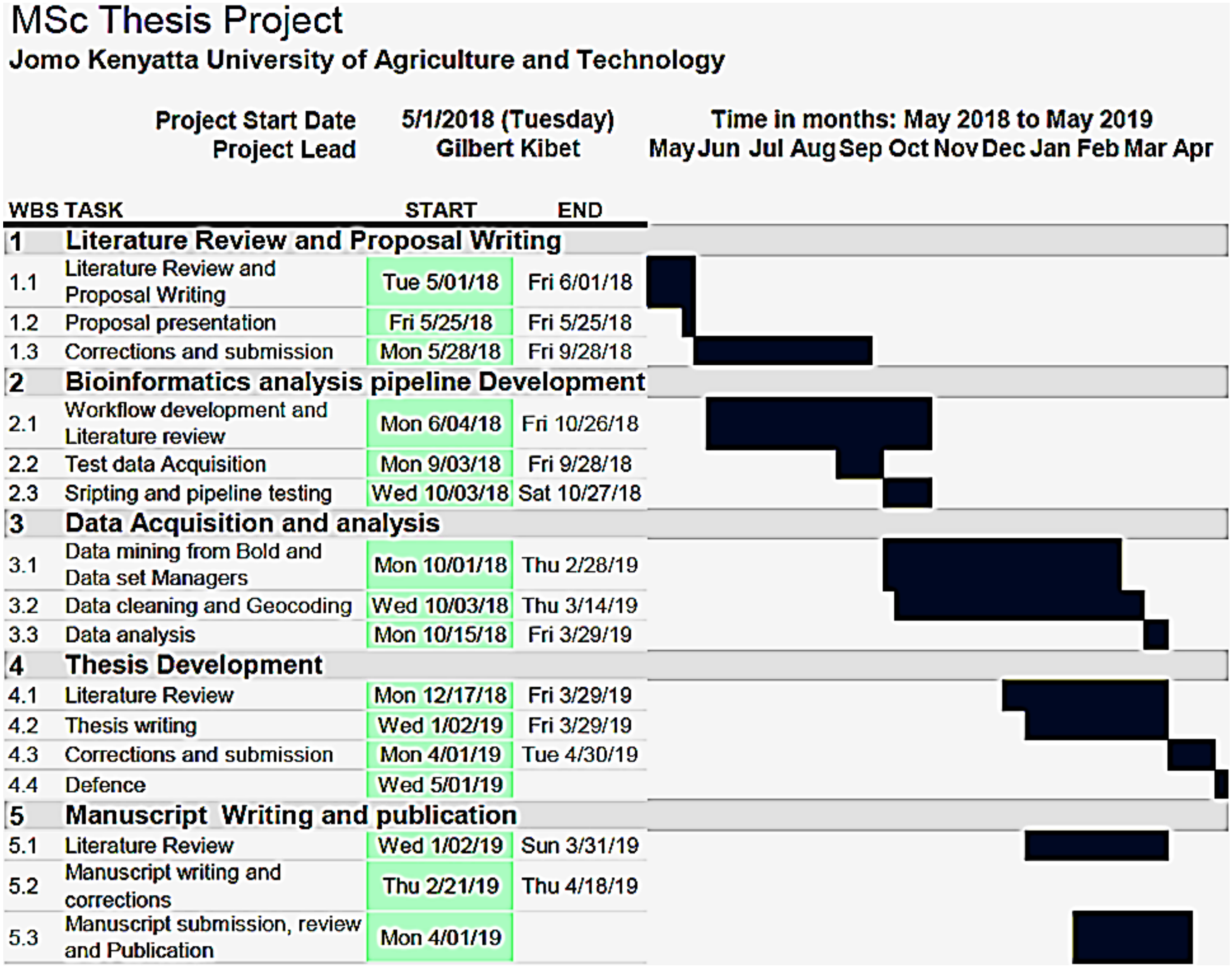
All the tools and applications except very few will be available on OMICStools website[xi], (Henry, Bandrowski, Pepin, Gonzalez, & Desfeux, 2014).

[x] http://www.beast2.org/

[xi] https://omicx.com/

WORK FLOW

Table 3: Gantt chart of the project workflow.



MSc Thesis Project

Jomo Kenyatta University of Agriculture and Technology

Project Start Date 5/1/2018 (Tuesday)

Project Lead Gilbert Kibet

Time in months: May 2018 to May 2019

WBS	TASK	START	END
1	Literature Review and Proposal Writing		
1.1	Literature Review and Proposal Writing	Tue 5/01/18	Fri 6/01/18
1.2	Proposal presentation	Fri 5/25/18	Fri 5/25/18
1.3	Corrections and submission	Mon 5/28/18	Fri 9/28/18
2	Bioinformatics analysis pipeline Development		
2.1	Workflow development and Literature review	Mon 6/04/18	Fri 10/26/18
2.2	Test data Acquisition	Mon 9/03/18	Fri 9/28/18
2.3	Sripting and pipeline testing	Wed 10/03/18	Sat 10/27/18
3	Data Acquisition and analysis		
3.1	Data mining from Bold and Data set Managers	Mon 10/01/18	Thu 2/28/19
3.2	Data cleaning and Geocoding	Wed 10/03/18	Thu 3/14/19
3.3	Data analysis	Mon 10/15/18	Fri 3/29/19
4	Thesis Development		
4.1	Literature Review	Mon 12/17/18	Fri 3/29/19
4.2	Thesis writing	Wed 1/02/19	Fri 3/29/19
4.3	Corrections and submission	Mon 4/01/19	Tue 4/30/19
4.4	Defence	Wed 5/01/19	
5	Manuscript Writing and publication		
5.1	Literature Review	Wed 1/02/19	Sun 3/31/19
5.2	Manuscript writing and corrections	Thu 2/21/19	Thu 4/18/19
5.3	Manuscript submission, review and Publication	Mon 4/01/19	

BUDGET

Activity/ Materials.			Expenditure.
i.	High-performance PC		KSh 100, 000
ii.	Software		KSh 20, 000
iii.	Medical Insurance		KSh 150, 000 p.a.
iv.	Bench Fee	a) Research support service	KSh 230, 000 p.a.
		b) + IT	KSh 180, 000 p.a.
		c) + Space	KSh 150, 000 p.a.
		d) + Research coordination	KSh 36, 000 p.a.
		Total (Bench Fee)	KSh 596, 000 p.a.
v.	Transport and accommodation (Stipend)		KSh 300, 000 p.a.
vi.	Miscellaneous Expenditure		KSh 14, 000 p.a.
TOTAL			KSh 1, 180, 000

REFERENCES

Agassiz, D., & Kallies, A. (2018). *A new genus and species of myrmecophile clearwing moth (Lepidoptera: Sesiidae) from East Africa* (Vol. 4392). https://doi.org/10.11646/zootaxa.4392.3.8

An, Patwardhan, Ray, S., & Roy, A. (2014). Molecular Markers in Phylogenetic Studies-A Review. *Journal of Phylogenetics & Evolutionary Biology*, *2*(2), 1–9. https://doi.org/10.4172/2329-9002.1000131

Arthropoda | Taxonomy Browser | BOLDSYSTEMS. (2018, January 21). Retrieved 21 January 2018, from http://www.boldsystems.org/index.php/Taxbrowser_Taxonpage?taxid=20

Avise, John C., Arnold, J., Ball, R. M., Bermingham, E., Lamb, T., Neigel, J. E., … Saunders, N. C. (1987). Intraspecific Phylogeography: The Mitochondrial DNA Bridge Between Population Genetics and Systematics. *Annual Review of Ecology and Systematics*, *18*(1), 489–522. https://doi.org/10.1146/annurev.es.18.110187.002421

Avise, John C. (1989). Gene Trees and Organismal Histories: A Phylogenetic Approach to Population Biology. *Evolution*, *43*(6), 1192–1208. https://doi.org/10.2307/2409356

Avise, John C. (2008). Phylogeography: retrospect and prospect. *Journal of Biogeography*, *36*(1), 3–15. https://doi.org/10.1111/j.1365-2699.2008.02032.x

Bermingham, E., & Moritz, C. (1998). Comparative phylogeography: concepts and applications. *Molecular Ecology*, *7*(4), 367–369. https://doi.org/10.1046/j.1365-294x.1998.00424.x

Bouckaert, R. (2016). Phylogeography by diffusion on a sphere: whole world phylogeography. *PeerJ*, *4*. https://doi.org/10.7717/peerj.2406

Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.-H., Xie, D., … Drummond, A. J. (2014). BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. *PLoS Computational Biology*, *10*(4). https://doi.org/10.1371/journal.pcbi.1003537

Brehm, G., Hebert, P. D. N., Colwell, R. K., Adams, M.-O., Bodner, F., Friedemann, K., … Fiedler, K. (2016). Turning Up the Heat on a Hotspot: DNA Barcodes Reveal 80% More Species of Geometrid Moths along an Andean Elevational Gradient. *PLoS ONE*, *11*(3). https://doi.org/10.1371/journal.pone.0150327

Brown, T. A. (2002). *Molecular Phylogenetics*. Wiley-Liss. Retrieved from https://www.ncbi.nlm.nih.gov/books/NBK21122/

C. D., S., & V. P., A. (2016). *Cytochrome oxidase subunit I gene based phylogenetic description of common mormon butterfly Papilio polytes (Lepidoptera: Papilionidae)* (Vol. 5).

Caterino, M. S., & Sperling, F. A. (1999). Papilio phylogeny based on mitochondrial cytochrome oxidase I and II genes. *Molecular Phylogenetics and Evolution*, *11*(1), 122–137. https://doi.org/10.1006/mpev.1998.0549

Cheseto, X., Kachigamba, D. L., Ekesi, S., Ndung'u, M., Teal, P. E. A., Beck, J. J., & Torto, B. (2017). Identification of the Ubiquitous Antioxidant Tripeptide Glutathione as a Fruit Fly Semiochemical. *Journal of Agricultural and Food Chemistry*, *65*(39), 8560–8568. https://doi.org/10.1021/acs.jafc.7b03164

Cox, A. J., & Hebert, P. D. N. (2001). Colonization, extinction, and phylogeographic patterning in a freshwater crustacean. *Molecular Ecology*, *10*(2), 371–386. https://doi.org/10.1046/j.1365-294X.2001.01188.x

Databases | BOLDSYSTEMS. (2018, January 20). Retrieved 20 January 2018, from http://www.boldsystems.org/index.php/databases

Dawson, M., S. Waples, R., & Bernardi, G. (2006). Phylogeography. In *Phylogeography*. https://doi.org/10.1525/california/9780520246539.003.0002

Ebach, M. C., & Holdrege, C. (2005). DNA barcoding is no substitute for taxonomy. *Nature*, *434*, 697.

Edgar, R. C. (2004a). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, *5*(1), 113. https://doi.org/10.1186/1471-2105-5-113

Edgar, R. C. (2004b). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, *32*(5), 1792–1797. https://doi.org/10.1093/nar/gkh340

Edmunds, A. J. F., Aluja, M., Diaz-Fleischer, F., Patrian, B., & Hagmann, L. (2010). Host marking pheromone (HMP) in the Mexican fruit fly Anastrepha ludens. *Chimia*, *64*(1–2), 37–42.

Fang, Y., Zhang, J., Wu, R., Xue, B., Qian, Q., & Gao, B. (2018). Genetic Polymorphism Study on Aedes albopictus of Different Geographical Regions Based on DNA Barcoding. *BioMed Research International*, *2018*. https://doi.org/10.1155/2018/1501430

Galtier, N., Enard, D., Radondy, Y., Bazin, E., & Belkhir, K. (2006). Mutation hot spots in mammalian mitochondrial DNA. *Genome Research*, *16*(2), 215–222. https://doi.org/10.1101/gr.4305906

Ghaleb, F., Reda, N., & Al-Neama, M. (2013). *An Overview of Multiple Sequence Alignment Parallel Tools*.

Gorbunov, O. G., & Gurko, V. O. (2017). A new genus and species of clearwing moths (Lepidoptera: Sesiidae) from South Sudan. *Zootaxa*, *4276*(2), 270–276.

Gratton, P., Marta, S., Bocksberger, G., Winter, M., Trucchi, E., & Kühl, H. (2017). A world of sequences: can we use georeferenced nucleotide databases for a robust automated phylogeography? *Journal of Biogeography*, *44*(2), 475–486. https://doi.org/10.1111/jbi.12786

Gutiérrez-García, T. A., & Vázquez-Domínguez, E. (2011). Comparative Phylogeography: Designing Studies while Surviving the Process. *BioScience*, *61*(11), 857–868. https://doi.org/10.1525/bio.2011.61.11.5

Hajibabaei, M., A.C. Singer, G., Hebert, P., & A Hickey, D. (2007). *Hajibabaei M, Singer GAC, Hebert PDN, Hickey DA. DNA barcoding: how it complements taxonomy, molecular phylogenetics and population genetics. Trends Genet 23: 167-172* (Vol. 23). https://doi.org/10.1016/j.tig.2007.02.001

Hajibabaei, M., Janzen, D. H., Burns, J. M., Hallwachs, W., & Hebert, P. D. N. (2006). DNA barcodes distinguish species of tropical Lepidoptera. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(4), 968–971. https://doi.org/10.1073/pnas.0510466103

Hajibabaei, M., Singer, G. A. C., Hebert, P. D. N., & Hickey, D. A. (2007). DNA barcoding: how it complements taxonomy, molecular phylogenetics and population genetics. *Trends in Genetics*, *23*(4), 167–172. https://doi.org/10.1016/j.tig.2007.02.001

Hebert, P. D. N., Cywinska, A., Ball, S. L., & deWaard, J. R. (2003). Biological identifications through DNA barcodes. *Proceedings of the Royal Society B: Biological Sciences*, *270*(1512), 313–321. https://doi.org/10.1098/rspb.2002.2218

Hebert, P. D. N., Hollingsworth, P. M., & Hajibabaei, M. (2016). From writing to reading the encyclopedia of life. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1702). https://doi.org/10.1098/rstb.2015.0321

Hebert, P. D. N., Ratnasingham, S., & Waard, J. R. de. (2003). Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Proceedings of the Royal Society of London B: Biological Sciences*, *270*(Suppl 1), S96–S99. https://doi.org/10.1098/rsbl.2003.0025

Henry, V. J., Bandrowski, A. E., Pepin, A.-S., Gonzalez, B. J., & Desfeux, A. (2014). OMICtools: an informative directory for multi-omic data analysis. *Database: The Journal of Biological Databases and Curation*, *2014*. https://doi.org/10.1093/database/bau069

Hickerson, M. J., Carstens, B. C., Cavender-Bares, J., Crandall, K. A., Graham, C. H., Johnson, J. B., … Yoder, A. D. (2010). Phylogeography's past, present, and future: 10 years after Avise, 2000. *Molecular Phylogenetics and Evolution*, *54*(1), 291–301. https://doi.org/10.1016/j.ympev.2009.09.016

Kang, Y., Deng, Z., Zang, R., & Long, W. (2017). DNA barcoding analysis and phylogenetic relationships of tree species in tropical cloud forests. *Scientific Reports*, *7*(1), 12564. https://doi.org/10.1038/s41598-017-13057-0

Karsch-Mizrachi, I., Takagi, T., & Cochrane, G. (2018). The international nucleotide sequence database collaboration. *Nucleic Acids Research*, *46*(D1), D48–D51. https://doi.org/10.1093/nar/gkx1097

Katoh, K., Misawa, K., Kuma, K., & Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, *30*(14), 3059–3066. https://doi.org/10.1093/nar/gkf436

Katoh, K., & Standley, D. M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution*, *30*(4), 772–780. https://doi.org/10.1093/molbev/mst010

Kekkonen, M., & Hebert, P. D. N. (2014). DNA barcode-based delineation of putative species: efficient start for taxonomic workflows. *Molecular Ecology Resources*, *14*(4), 706–715. https://doi.org/10.1111/1755-0998.12233

Kenya | iBOL. (2018, April 12). Retrieved 12 April 2018, from http://ibol.org/kenya/

Kjer, K., Blahnik, R., & Holzenthal, R. (2001). *Phylogeny of Trichoptera (Caddisflies): Characterization of Signal and Noise Within Multiple Datasets* (Vol. 50). https://doi.org/10.1080/106351501753462812

Krosch, M., K Schutze, M., Armstrong, K., C Graham, G., Yeates, D., & R Clarke, A. (2012). A molecular phylogeny for the Tribe Dacini (Diptera: Tephritidae): Systematic and biogeographic implications, *64*, 513–523. https://doi.org/10.1016/j.ympev.2012.05.006

Kumar, R., & Kumar, V. (2018). A review of phylogeography: biotic and abiotic factors. *Geology, Ecology, and Landscapes*, *0*(0), 1–7. https://doi.org/10.1080/24749508.2018.1452486

Kurtenbach, G., & Buxton, B. (1991). GEdit: A Test Bed for Editing by Contiguous Gestures. *SIGCHI Bull.*, *23*(2), 22–26. https://doi.org/10.1145/122488.122490

Lin, C.-P., & Danforth, B. N. (2004). How do insect nuclear and mitochondrial gene substitution patterns differ? Insights from Bayesian analyses of combined datasets. *Molecular Phylogenetics and Evolution*, *30*(3), 686–702. https://doi.org/10.1016/S1055-7903(03)00241-0

Liu, K., Linder, C. R., & Warnow, T. (2011). RAxML and FastTree: Comparing Two Methods for Large-Scale Maximum Likelihood Phylogeny Estimation. *PLoS ONE*, *6*(11). https://doi.org/10.1371/journal.pone.0027731

Liu, K., & Warnow, T. (2014). Large-Scale Multiple Sequence Alignment and Tree Estimation Using SATé. *Methods in Molecular Biology (Clifton, N.J.)*, *1079*, 219–244. https://doi.org/10.1007/978-1-62703-646-7_15

Liu, K., Warnow, T. J., Holder, M. T., Nelesen, S. M., Yu, J., Stamatakis, A. P., & Linder, C. R. (2012). SATé-II: Very Fast and Accurate Simultaneous Estimation of Multiple Sequence Alignments and Phylogenetic Trees. *Systematic Biology*, *61*(1), 90–90. https://doi.org/10.1093/sysbio/syr095

Louise Bennett, K., Kaddumukasa, M., Shija, F., Djouaka, R., Misinzo, G., Lutwama, J., … Walton, C. (2018). Comparative phylogeography of Aedes mosquitoes and the role of past climatic change for evolution within Africa, *8*. https://doi.org/10.1002/ece3.3668

Maio, N. D., Wu, C.-H., O'Reilly, K. M., & Wilson, D. (2015). New Routes to Phylogeography: A Bayesian Structured Coalescent Approximation. *PLOS Genetics, 11*(8), e1005421. https://doi.org/10.1371/journal.pgen.1005421

Marske, K. (2016). *Phylogeography*. https://doi.org/10.1016/B978-0-12-800049-6.00109-8

Mende, M. B., Bartel, M., & Hundsdoerfer, A. K. (2016). A comprehensive phylogeography of the Hyles euphorbiae complex (Lepidoptera: Sphingidae) indicates a 'glacial refuge belt'. *Scientific Reports, 6*. https://doi.org/10.1038/srep29527

Meyer, M., Delatte, H., Ekesi, S., Jordaens, K., Kalinova, B., Manrakhan, A., … Virgilio, M. (2015). *An integrative approach to unravel the Ceratitis FAR (Diptera, Tephritidae) cryptic species complex: a review* (Vol. 540). https://doi.org/10.3897/zookeys.540.10046

Meyer, M., Mwatawala, M., Copeland, R., & Virgilio, M. (2016). *Description of new Ceratitis species (Diptera: Tephritidae) from Africa, or how morphological and DNA data are complementary in discovering unknown species and matching sexes* (Vol. 2016). https://doi.org/10.5852/ejt.2016.233

Miller, S., Copeland, R., E Rosati, M., & Hebert, P. (2014). *DNA Barcodes of Microlepidoptera Reared from Native Fruit in Kenya* (Vol. 116). https://doi.org/10.4289/0013-8797.116.1.137

Miller, S. E. (2007). DNA barcoding and the renaissance of taxonomy. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(12), 4775–4776. https://doi.org/10.1073/pnas.0700466104

Miller, S. E., Hausmann, A., Hallwachs, W., & Janzen, D. H. (2016). Advancing taxonomy and bioinventories with DNA barcodes. *Philosophical Transactions of the Royal Society B: Biological Sciences, 371*(1702). https://doi.org/10.1098/rstb.2015.0339

Miller, S., E Rosati, M., Gewa, B., Novotny, V., D Weiblen, G., & Hebert, P. (2015). *DNA Barcodes of Lepidoptera Reared from Yawan, Papua New Guinea* (Vol. 117). https://doi.org/10.4289/0013-8797.117.2.247

Miller, S., Martins, D., Rosati, M., & Hebert, P. (2014). *DNA Barcodes of Moths (Lepidoptera) from Lake Turkana, Kenya* (Vol. 116). https://doi.org/10.4289/0013-8797.116.1.133

Mirarab, S., Nguyen, N., Guo, S., Wang, L.-S., Kim, J., & Warnow, T. (2015). PASTA: Ultra-Large Multiple Sequence Alignment for Nucleotide and Amino-Acid Sequences.

Journal of Computational Biology, *22*(5), 377–386. https://doi.org/10.1089/cmb.2014.0156

Misof, B., Liu, S., Meusemann, K., Peters, R. S., Donath, A., Mayer, C., … Zhou, X. (2014). Phylogenomics resolves the timing and pattern of insect evolution. *Science*, *346*(6210), 763. https://doi.org/10.1126/science.1257570

Munir, M. (2013). Bioinformatics analysis of large-scale viral sequences. *Virulence*, *4*(1), 97–106. https://doi.org/10.4161/viru.23161

ODENY, D. O, Ndungu, N., Masiga, D., Khayota, B., & Oyieko, H. (2017). IMPLEMENTING A NATIONAL DNA BARCODING OPERATING NODE FROM SCRATCH - THE EXPERIENCE OF KENBOL. In *Networks*. Adelaide, Australia. Retrieved from http://www.dnabarcodes2011.org/conference/index.php

Peter Linder H. (2017). Phylogeography. *Journal of Biogeography*, *44*(2), 243–244. https://doi.org/10.1111/jbi.12958

Porter, T., Gibson, J., Shokralla, S., Baird, D., Brian Golding, G., & Hajibabaei, M. (2014). *Rapid and accurate taxonomic classification of insect (Class Insecta) cytochrome c oxidase subunit 1 (COI) DNA barcode sequences using a naïve Bayesian classifier* (Vol. 14). https://doi.org/10.1111/1755-0998.12240

Price, M. N., Dehal, P. S., & Arkin, A. P. (2010). FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS ONE*, *5*(3). https://doi.org/10.1371/journal.pone.0009490

Ratnasingham, S., & Hebert, P. D. N. (2007). bold: The Barcode of Life Data System (http://www.barcodinglife.org). *Molecular Ecology Notes*, *7*(3), 355–364. https://doi.org/10.1111/j.1471-8286.2007.01678.x

Ratnasingham, S., & Hebert, P. D. N. (2013). A DNA-Based Registry for All Animal Species: The Barcode Index Number (BIN) System. *PLOS ONE*, *8*(7), e66213. https://doi.org/10.1371/journal.pone.0066213

Record List | Public Data Portal | BOLDSYSTEMS. (2018, January 20). Retrieved 20 January 2018, from http://www.boldsystems.org/index.php/Public_SearchTerms

Riedel, A., Sagata, K., Suhardjono, Y. R., Tänzler, R., & Balke, M. (2013). Integrative taxonomy on the fast track - towards more sustainability in biodiversity research. *Frontiers in Zoology*, *10*, 15. https://doi.org/10.1186/1742-9994-10-15

Robbins, A., Lamb, L., & Hannah, E. (2008). *Learning the vi and Vim Editors* (Seventh). O'Reilly.

Roderic D. M. Page. (2016). International Barcode of Life project (iBOL). https://doi.org/10.15468/inygc6

Roe, A. D., & Sperling, F. A. H. (2007). Patterns of evolution of mitochondrial cytochrome c oxidase I and II DNA and implications for DNA barcoding. *Molecular Phylogenetics and Evolution*, *44*(1), 325–345. https://doi.org/10.1016/j.ympev.2006.12.005

Sarkar, I. N., & Trizna, M. (2011). The Barcode of Life Data Portal: Bridging the Biodiversity Informatics Divide for DNA Barcoding. *PLOS ONE*, *6*(7), e14689. https://doi.org/10.1371/journal.pone.0014689

Savolainen, V., Cowan, R. S., Vogler, A. P., Roderick, G. K., & Lane, R. (2005). Towards writing the encyclopaedia of life: an introduction to DNA barcoding. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*(1462), 1805–1811. https://doi.org/10.1098/rstb.2005.1730

Sezonlin, M., Dupas, S., Le Ru, B., Le Gall, P., Moyal, P., Calatayud, P.-A., … Silvain, J.-F. (2006). Phylogeography and population genetics of the maize stalk borer Busseola fusca (Lepidoptera, Noctuidae) in sub-Saharan Africa, *15*, 407–420. https://doi.org/DOI:10.1111/j.1365-294X.2005.02761.x

Singh, D., Khullar, N., & Jha, C. (2015). Lucrative potentials of mitochondrial DNA: A laconic review accentuating particularly blow flies beyond forensic importance. *Journal of Entomology and Zoology Studies*, *3*, 01–08.

Smith, M. A., Fisher, B. L., & Hebert, P. D. . (2005). DNA barcoding for effective biodiversity assessment of a hyperdiverse arthropod group: the ants of Madagascar. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*(1462), 1825–1834. https://doi.org/10.1098/rstb.2005.1714

Smouse, P. E. (1998). To tree or not to tree. *Molecular Ecology*, *7*(4), 399–412. https://doi.org/10.1046/j.1365-294x.1998.00370.x

Souza, H. V., Marchesin, S. R. C., & Itoyama, M. M. (2016). Analysis of the mitochondrial COI gene and its informative potential for evolutionary inferences in the families Coreidae and Pentatomidae (Heteroptera). *Genetics and Molecular Research: GMR*, *15*(1). https://doi.org/10.4238/gmr.15017428

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*(9), 1312–1313. https://doi.org/10.1093/bioinformatics/btu033

Virgilio, M., Jordaens, K., Verwimp, C., White, I. M., & De Meyer, M. (2015). Higher phylogeny of frugivorous flies (Diptera, Tephritidae, Dacini): localised partition conflicts and a novel generic classification. *Molecular Phylogenetics and Evolution*, *85*, 171–179. https://doi.org/10.1016/j.ympev.2015.01.007

Virgilio, M., Meyer, M., White, I. M., & Backeljau, T. (2009). African Dacus (Diptera: Tephritidae: Molecular data and host plant associations do not corroborate morphology based classifications, *51*. https://doi.org/10.1016/j.ympev.2009.01.003

Virgilio, M., White, I., & Meyer, M. (2014). *A set of multi-entry identification keys to African frugivorous flies (Diptera, Tephritidae)* (Vol. 428). https://doi.org/10.3897/zookeys.428.7366

What Is DNA Barcoding? « Barcode of Life. (2018, January 19). Retrieved 19 January 2018, from http://www.barcodeoflife.org/content/about/what-dna-barcoding

Zhang, D. X., & Hewitt, G. M. (1997). Assessment of the universality and utility of a set of conserved mitochondrial COI primers in insects. *Insect Molecular Biology*, *6*(2), 143–150.

Zhou, X., Frandsen, P. B., Holzenthal, R. W., Beet, C. R., Bennett, K. R., Blahnik, R. J., … Kjer, K. M. (2016). The Trichoptera barcode initiative: a strategy for generating a species-level Tree of Life. *Phil. Trans. R. Soc. B*, *371*(1702), 20160025. https://doi.org/10.1098/rstb.2016.0025

Zoologische Staatssammlung Muenchen - International Barcode of Life (iBOL) - Barcode of Life Project Specimen Data. (2018). https://doi.org/10.15468/tfpnkp

Zou, Q., Hu, Q., Guo, M., & Wang, G. (2015). HAlign: Fast multiple similar DNA/RNA sequence alignment based on the centre star strategy. *Bioinformatics*, *31*(15), 2475–2481. https://doi.org/10.1093/bioinformatics/btv177

Zwickl, D. J., Hillis, D. M., & Crandall, K. (2002). Increased Taxon Sampling Greatly Reduces Phylogenetic Error. *Systematic Biology*, *51*(4), 588–598. https://doi.org/10.1080/10635150290102339