OBJECTIVES

General objectives

To construct a series of phylogenetic trees based on COI barcode sequences of arthropod specimens hosted by BOLD database and establish the phylogenetic distribution of these organisms within East Africa. And to conduct phylogeographic assessment based on their associated geographic coordinates and metadata.

Specific Objectives

1. To identify phylogenetic and phylogeographic distribution of fruit flies (Tephritidae family) and biting flies (Diptera: True flies) within East Africa.
2. To identify insect species that are potential pests or disease vectors from combined phylogenetic similarity and phylogeographic distribution.
3. To identify any potential new classification or species if any.

In the period from 1st March 2019 to end of June 2019, I managed to optimise the Multiple Sequence Alignment (MSA) pipeline and run it's analysis on the High Performance Computing cluster hosted at Duduville Campus, *icipe*. I also did phylogenetic inference on the final dataset.

MSA tools optimization and testing:

From the tools selected for this part of the project - MUSCLE, T_Coffee, MAFFT7, PASTA, SATé, SEPP, UPP and HMMER: MSA visualization; SeaView, SuiteMSA and Jalview: MSA evaluation; T_Coffee's CORE index and Transitive Consistency Scores(TCS) - PASTA , a progressive and transitivity based algorithm, and set to use third party tools MAFFT for alignment and OPAL for merging data subsets proved accurate and adequately fast than the rest. It produced gappy alignments hence used MUSCLE to refine and improve the final alignments. SATé incompatible with available versions of Dendropy, Python's phylogenetic package, proved difficult to upgrade. For visualization SeaView was the most suitable for our

large data sets and Jalview for datasets up to 10 MB. TSC proved most suitable for the evaluation of MSAs from the different tools and was useful in the selection of different tools.

MSA analysis: The Bold data were grouped as follows:

Africa (48): South Africa, Nigeria, Egypt, Morocco, Democratic republic of the Congo, Ghana, Algeria, Senegal, Zimbabwe, Mali, Cameroon, Sudan, Madagascar, Cote d'Ivoire, Tunisia, Angola, Somalia, Mauritius, Namibia, Cape Verde, Zambia, Libya, Mozambique, Botswana, Guinea, Guinea-Bissau, Benin, Seychelles, Burkina Faso, Gabon, Malawi, Chad, Togo, Niger, Liberia, Central African Republic, Mauritania, Swaziland, Eritrea, Gambia, Djibouti, Sierra Leone, Reunion, Lesotho, Republic of the Congo, Equatorial Guinea, Sao Tome and Principe and Comoros

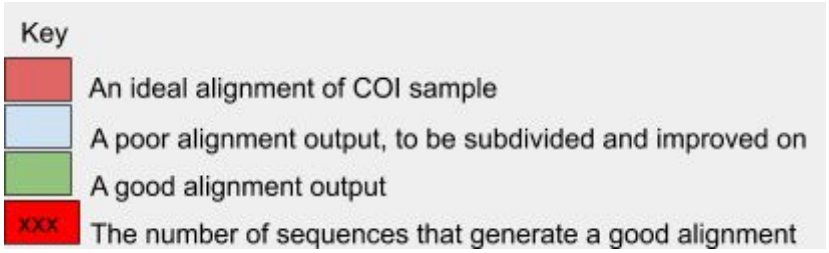
East Africa (7): Kenya, Tanzania, Uganda, Rwanda, Burundi, South Sudan and Ethiopia

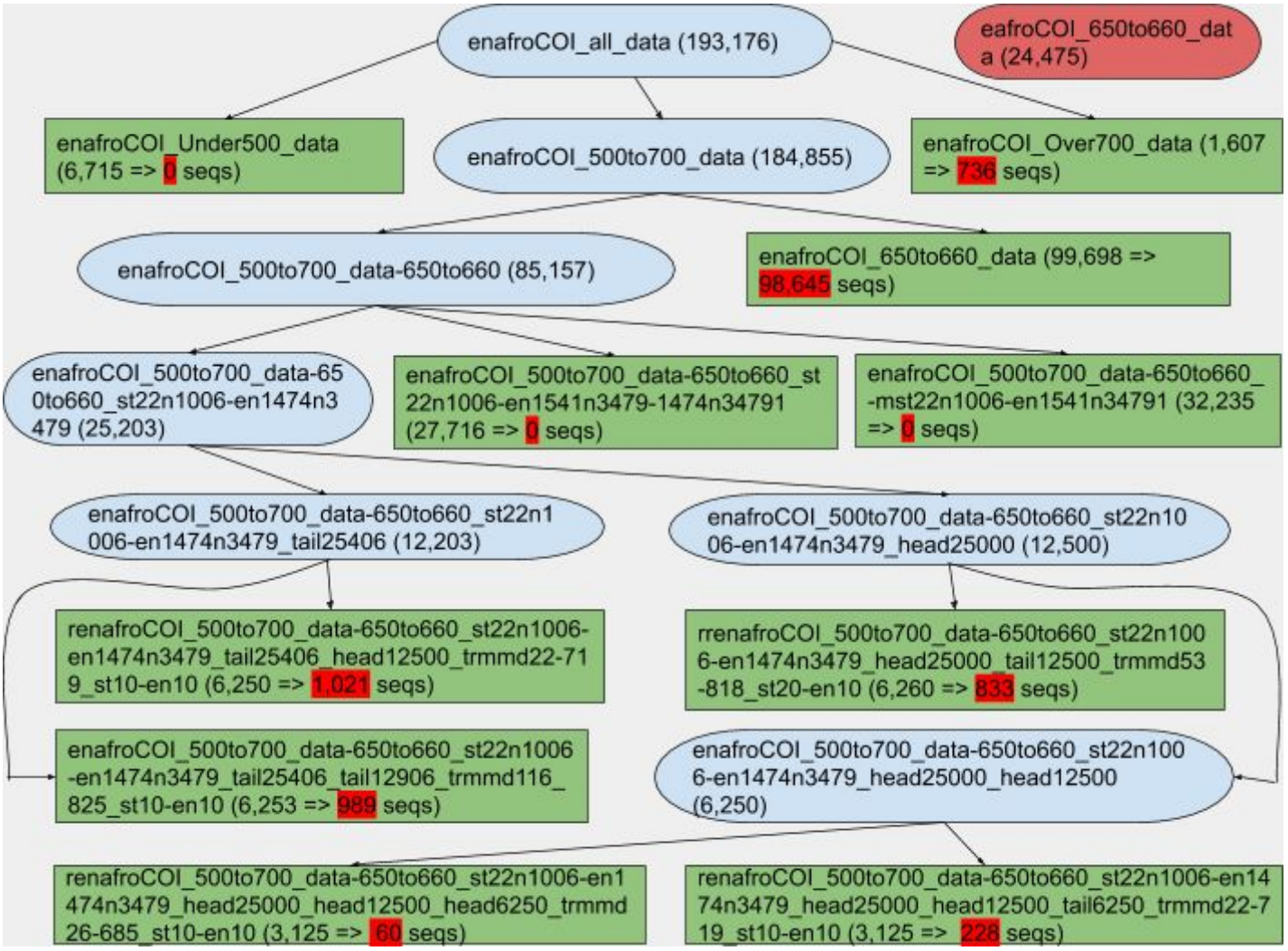
Descriptive name (Based on number of nucleotides in a sequence):		Africa (prefix: afroCOI)	East Africa (prefix: eafroCOI)	All Africa (prefix: enafroCOI)
_all_data		154,765	38,421	193, 176
_Under500_data		6,065	770	6,715
_Over700_data		1023	2900	1,607
_500to700_data	_500to700_data	147,678	37,182	184,855
	_650to660_data	75,229	24,475	99,698
	_500to700_data-650to660	-	-	85,157
_Over499_data		148,737	37,766	186,458

The MSA alignment analysis workflow is shown below: Datasets are aligned using PASTA and the output visually evaluated in SeaView to determine if its good quality. A bad MSA output is subsetted using a specific algorithm and its subsets realigned and evaluated again. This is done in iterations until a final good visually alignment is achieved. This output can be futher refined using MUSCLE and “r*” prefix appended to its name. Other operations done on the data sets is: (i) trimming - and ”trmmd*” suffix added - to only extract columns that fall

within the 658 region of the COI; (ii) Deletion of short sequences - and a “st*-en*” suffix added - to remove short sequences with a specified number of gaps, “_”, or undefined nucleotides, “N”, in the 3’ or 5’ end of the sequence.

From the qualified sequences a final alignment of 99,109 sequences is generated (enafroCOI_all_clean_sN10-eN10.aln). An additional of three sequences are added to be used as the outgroups: *Thoralus cranchii* (Crustacea), *Ryuthela sasakii* (Arachnida) and *Ballonema sp. MCZ DNA106902* (Chilopoda).





Phylogenetic Inference:

Tools: To conduct a maximum likelihood estimation of phylogenetic trees RAxML and FastTree were used. To conduct bootstrapping analysis; RAxML, 100 bootstraps. And do outgroup rooting of the phylogenetic trees; RAxML's Evolutionary Placement Algorithm (EPA) was used. To visualise the trees: Archaeopteryx, FigTree and Dendroscope were used. Biopython's Bio.Phylo was used for further analysis of the tree (ongoing) and more is yet to be done.

Analysis: Four data sets have so far been run for phylogenetic inference of the trees:

i. All African data (enafroCOI_all_clean_sN10-eN10.aln) which had 63,057 nonidentical sequences out of 99,109 sequences
ii. All African data with genus classification mentioned (enafroCOI_all_clean_sN10-eN10_genera.aln) which had 20,581 nonidentical sequences out of 34,201 sequences
iii. East African data (enafroCOI_all_clean_sN10-eN10_eafro.aln) which had 15,560 nonidentical sequences out of 23,353 sequences
iv. East African data with genus classification (enafroCOI_all_clean_sN10-eN10_eafro_genera.aln) which had 5,472 nonidentical sequences out of 8,692 sequences.

The resulting trees too many terminal and internal nodes hence the need for further analysis and prioritization of particular clades based on a criteria. This is to be done using Bio.Phylo package of Biopython. The output will be a series of descriptive phylogenetic trees that will be used to describe molecular evolutionary hypothesis for particular clades like: polyphyletic, paraphyletic or monophyletic evolutionary characteristics.