Phylogenetic and Phylogeographic meta-analysis of Cytochrome c Oxidase I barcode sequences of East African arthropods submitted into the Barcode of Life Database

Presented by:

Gilbert Kibet
HSB316-5106/2016

Supervisors:

Dr Scott Miller
Dr Jandouwe Villinger
Dr Steven Ger

Collaborators:

Dr Caleb Kipkurui
Dr Jean-Baka Domelevo
Dr Daniel Masiga

Background

Identification and classification of organisms:

Morphology-based identification systems – extensive information (ecology, anatomy, physiology); expensive, slow and needs expertise

Molecular-based system – efficient (fast and effective); dependent on reference libraries of DNA barcode -*short and standardized genes or regions thereof used in identification and discovery of species*

The Consortium for the Barcode of Life (CBOL), May 2004: To aid rapid and inexpensive identification of millions of species using DNA barcodes

- **International Nucleotide Sequence Database Collaborative (INSDC):** GenBank, the European Molecular Biology Lab in Europe, and the DNA Data Bank of Japan
- **Barcode of Life Database (BOLD):** University of Guelph in Ontario

Background

A ***658 base-pair 5' region of mitochondrial cytochrome c oxidase subunit I (COI/COXI) gene*** is the standard the barcode for most animal groups

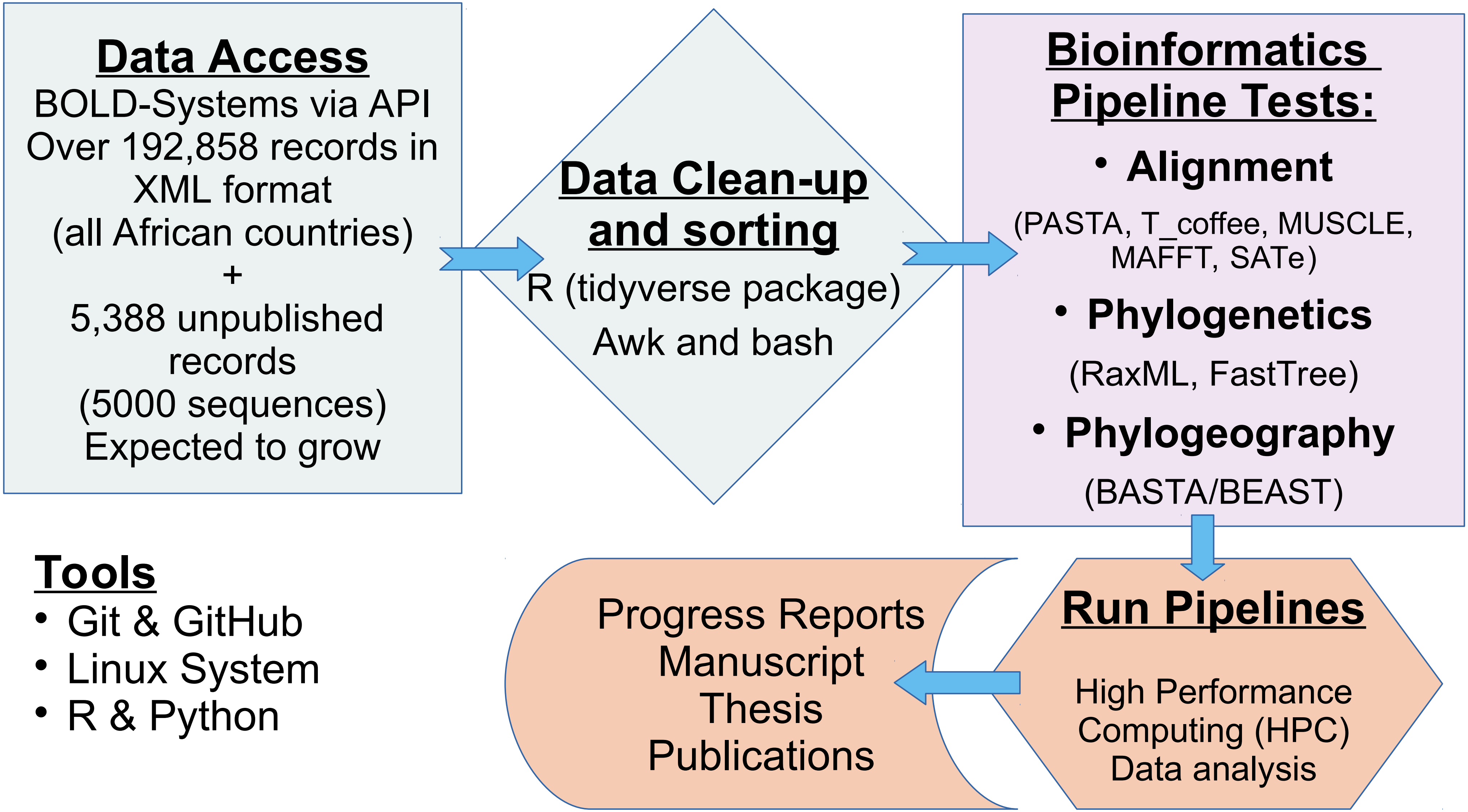
Problem statement:

Thousands of East African COI sequences from voucher arthropods submitted into the BOLD database and are yet to be analysed comprehensively: ****phylogenetic** diversity** and **phylogeographic distribution**

Objectives:

- Improve phylogeographic and **phylogenetic** diversity descriptions of arthropod in East Africa.
- Identify the cryptic species that may not yet be recognized and may be potential crop pests or vectors of human and animal diseases.

Workflow



Data Access

BOLD-Systems via API
Over 192,858 records in XML format
(all African countries)
+
5,388 unpublished records
(5000 sequences)
Expected to grow

Data Clean-up and sorting

R (tidyverse package)
Awk and bash

Bioinformatics Pipeline Tests:

- **Alignment**

(PASTA, T_coffee, MUSCLE, MAFFT, SATe)

- **Phylogenetics**

(RaxML, FastTree)

- **Phylogeography**

(BASTA/BEAST)

Run Pipelines

High Performance Computing (HPC)
Data analysis

Progress Reports
Manuscript
Thesis
Publications

Tools

- Git & GitHub
- Linux System
- R & Python

Data Mining and Cleansing

Automated Sequence retrival:

From BOLD-Systems via a URL based API using wget; XML files: Bash

Parsed using BeautifulSoup4 and lxml converted to 80 column text files (.tsv) with pandas: Python3

Data clean up: R (tidyverse package), Awk and BASH

<table>
<tr><td colspan="5">Metadata (80 columns)</td></tr>
<tr><td colspan="5">'COI-5P' = 198148 (43484 East African (1 sample))</td></tr>
<tr><td colspan="5">#nucleotides / unaligned seqs / #ns</td></tr>
<tr><td>Over 700
(1607)</td><td>Under 500
(6715)</td><td>500 -700
(184855)</td><td>650 -660
(99698)</td><td>Over 499
(186458)</td></tr>
<tr><td>1 sample</td><td>1 sample</td><td>3 samples</td><td>2 samples</td><td></td></tr>
<tr><td colspan="5">Build.fasta:
>ProcessID|order|genus|species|sub_species|country|exactsite|lat|lon|elev|seq_len
AGGTTCATCCCAA-----</td></tr>
</table>

Multiple Sequence Alignment

Large dataset:

- Accuracy
- Speed

Algorithms:

- Progressive (mafft/muscle)
- Progressive & transivity (pasta)
- Regressive (T_coffee)

MUSCLE: default	MAFFT: --large G-INS-1	T_coffee: -reg	PASTA: default	SATe
Fast speed	Fast speed	Fast speed	Fast speed	NA
Low accuracy	High accuracy	High accuracy	High accuracy	NA
• Refine • Align • Merge	• Align • Add sequence • Add_fragments • Merge	• Align • Evaluate: (CORE index TCS)	• Align • Add fragments	NA

Multiple Sequence Alignment evaluation

Evaluation for Accuracy:

T_coffee: **consistency based scoring**

- CORE index (html)
- Transitive Consistency Scores (TSC) (html/ascii)

Purpose:

Used to select the most suitable alignments.

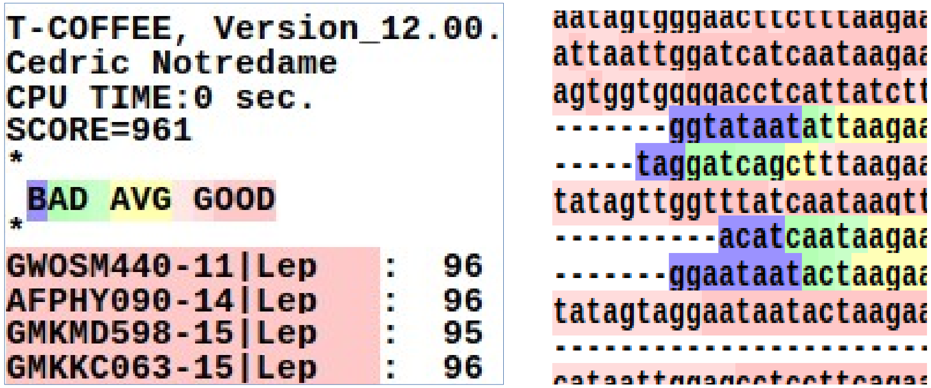
TCS ascii to used in applying different weights to columns in phylogenetic analysis

Visualization:

- Seaview
- Jalview
- SuiteMSA_Package1.3.22B

T_coffee consistency based
Multiple Sequence Alignment evaluation

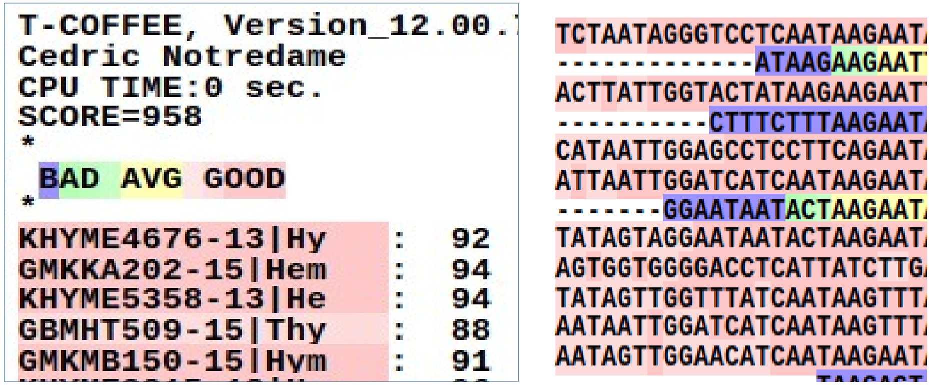
MAFFT



```
T-COFFEE, Version_12.00.
Cedric Notredame
CPU TIME:0 sec.
SCORE=961
*
 BAD AVG GOOD
*
GWOSM440-11|Lep    :  96
AFPHY090-14|Lep    :  96
GMKMD598-15|Lep    :  95
GMKKC063-15|Lep    :  96
```

```
aatagtgggaacttctttaagaa
attaattggatcatcaataagaa
agtggtggggacctcattatctt
--------ggtataatattaagaa
-----taggatcagctttaagaa
tatagttggtttatcaataagtt
-----------acatcaataagaa
--------ggaataatactaagaa
tatagtaggaataatactaagaa
-----------------------
cataattggagcctccttcagaa
```

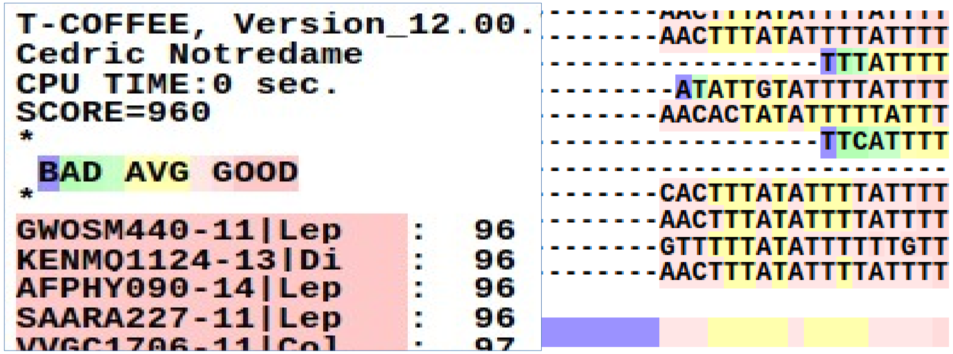
MUSCLE



```
T-COFFEE, Version_12.00.
Cedric Notredame
CPU TIME:0 sec.
SCORE=958
*
 BAD AVG GOOD
*
KHYME4676-13|Hy    :  92
GMKKA202-15|Hem    :  94
KHYME5358-13|He    :  94
GBMHT509-15|Thy    :  88
GMKMB150-15|Hym    :  91
```

```
TCTAATAGGGTCCTCAATAAGAAT
-------------ATAAGAAGAAT
ACTTATTGGTACTATAAGAAGAAT
----------CTTTCTTTAAGAAT
CATAATTGGAGCCTCCTTCAGAAT
ATTAATTGGATCATCAATAAGAAT
-------GGAATAATACTAAGAAT
TATAGTAGGAATAATACTAAGAAT
AGTGGTGGGGACCTCATTATCTTG
TATAGTTGGTTTATCAATAAGTTT
AATAATTGGATCATCAATAAGTTT
AATAGTTGGAACATCAATAAGAAT
```

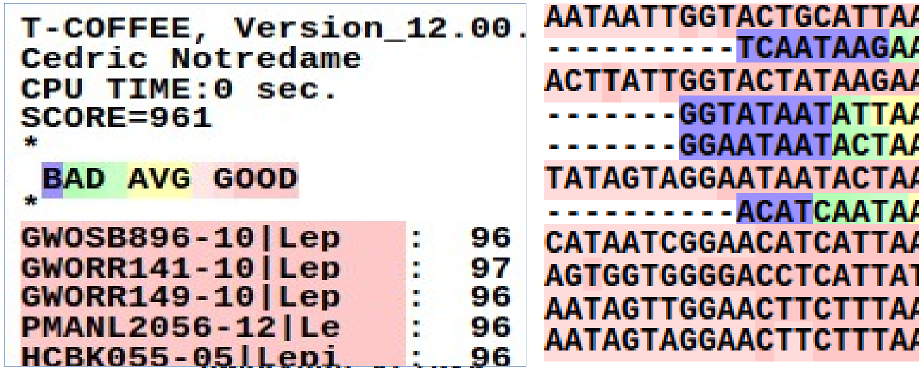
T_coffee



```
T-COFFEE, Version_12.00.
Cedric Notredame
CPU TIME:0 sec.
SCORE=960
*
 BAD AVG GOOD
*
GWOSM440-11|Lep    :  96
KENMQ1124-13|Di    :  96
AFPHY090-14|Lep    :  96
SAARA227-11|Lep    :  96
VVGC1706-11|Col    :  97
```

```
----------AACTTTATATTTTATTTT
----------AACTTTATATTTTATTTT
--------------------TTTATTTT
-----------ATATTGTATTTTATTTT
----------AACACTATATTTTTATTT
--------------------TTCATTTT
----------------------------
----------CACTTTATATTTTATTTT
----------AACTTTATATTTTATTTT
----------GTTTTTATATTTTTTGTT
----------AACTTTATATTTTATTTT
```

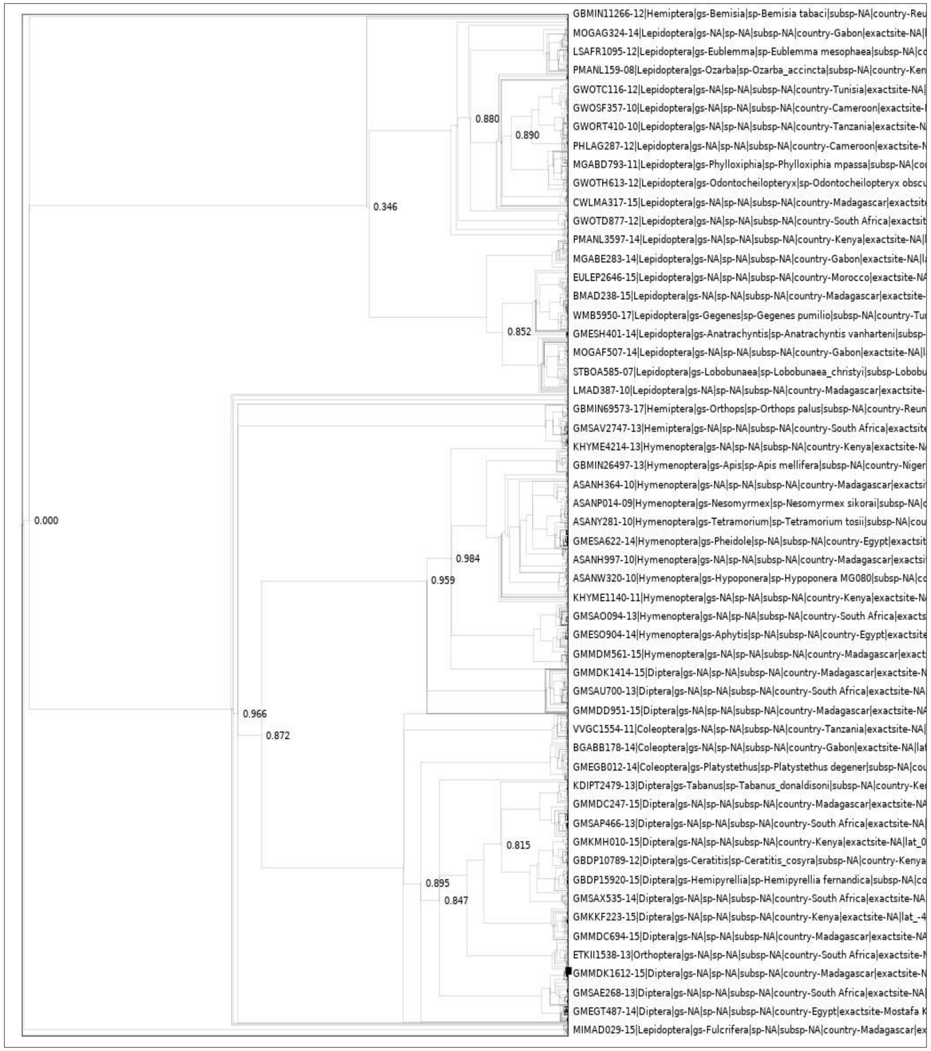
PASTA



```
T-COFFEE, Version_12.00.
Cedric Notredame
CPU TIME:0 sec.
SCORE=961
*
 BAD AVG GOOD
*
GWOSB896-10|Lep    :  96
GWORR141-10|Lep    :  97
GWORR149-10|Lep    :  96
PMANL2056-12|Le    :  96
HCBK055-05|Lepi.   :  96
```

```
AATAATTGGTACTGCATTAA
----------TCAATAAGAA
ACTTATTGGTACTATAAGAA
-------GGTATAATATTAA
-------GGAATAATACTAA
TATAGTAGGAATAATACTAA
----------ACATCAATAA
CATAATCGGAACATCATTAA
AGTGGTGGGGACCTCATTAT
AATAGTTGGAACTTCTTTAA
AATAGTAGGAACTTCTTTAA
```

Phylogenetic analysis: Basic information


0.000
0.346
0.880
0.890
0.852
0.984
0.959
0.966
0.872
0.815
0.895
0.847
GBMIN11266-12|Hemiptera|gs-Bemisia|sp-Bemisia tabaci|subsp-NA|country-Reu
MOGAG324-14|Lepidoptera|gs-NA|sp-NA|subsp-NA|country-Gabon|exactsite-NA|
LSAFR1095-12|Lepidoptera|gs-Eublemma|sp-Eublemma mesophaea|subsp-NA|co
PMANL159-08|Lepidoptera|gs-Ozarba|sp-Ozarba_accincta|subsp-NA|country-Ken
GWOTC116-12|Lepidoptera|gs-NA|sp-NA|subsp-NA|country-Tunisia|exactsite-NA|
GWOSF357-10|Lepidoptera|gs-NA|sp-NA|subsp-NA|country-Cameroon|exactsite-
GWORT410-10|Lepidoptera|gs-NA|sp-NA|subsp-NA|country-Tanzania|exactsite-N
PHLAG287-12|Lepidoptera|gs-NA|sp-NA|subsp-NA|country-Cameroon|exactsite-N
MGABD793-11|Lepidoptera|gs-Phylloxiphia|sp-Phylloxiphia mpassa|subsp-NA|co
GWOTH613-12|Lepidoptera|gs-Odontocheilopteryx|sp-Odontocheilopteryx obscu
CWLMA317-15|Lepidoptera|gs-NA|sp-NA|subsp-NA|country-Madagascar|exactsite
GWOTD877-12|Lepidoptera|gs-NA|sp-NA|subsp-NA|country-South Africa|exactsit
PMANL3597-14|Lepidoptera|gs-NA|sp-NA|subsp-NA|country-Kenya|exactsite-NA|
MGABE283-14|Lepidoptera|gs-NA|sp-NA|subsp-NA|country-Gabon|exactsite-NA|l
EULEP2646-15|Lepidoptera|gs-NA|sp-NA|subsp-NA|country-Morocco|exactsite-NA
BMAD238-15|Lepidoptera|gs-NA|sp-NA|subsp-NA|country-Madagascar|exactsite-
WMB5950-17|Lepidoptera|gs-Gegenes|sp-Gegenes pumilio|subsp-NA|country-Tu
GMESH401-14|Lepidoptera|gs-Anatrachyntis|sp-Anatrachyntis vanharteni|subsp-
MOGAF507-14|Lepidoptera|gs-NA|sp-NA|subsp-NA|country-Gabon|exactsite-NA|l
STBOA585-07|Lepidoptera|gs-Lobobunaea|sp-Lobobunaea_christyi|subsp-Lobobu
LMAD387-10|Lepidoptera|gs-NA|sp-NA|subsp-NA|country-Madagascar|exactsite-
GBMIN69573-17|Hemiptera|gs-Orthops|sp-Orthops palus|subsp-NA|country-Reun
GMSAV2747-13|Hemiptera|gs-NA|sp-NA|subsp-NA|country-South Africa|exactsite
KHYME4214-13|Hymenoptera|gs-NA|sp-NA|subsp-NA|country-Kenya|exactsite-N
GBMIN26497-13|Hymenoptera|gs-Apis|sp-Apis mellifera|subsp-NA|country-Niger
ASANH364-10|Hymenoptera|gs-NA|sp-NA|subsp-NA|country-Madagascar|exactsi
ASANP014-09|Hymenoptera|gs-Nesomyrmex|sp-Nesomyrmex sikorai|subsp-NA|c
ASANY281-10|Hymenoptera|gs-Tetramorium|sp-Tetramorium tosii|subsp-NA|cou
GMESA622-14|Hymenoptera|gs-Pheidole|sp-NA|subsp-NA|country-Egypt|exactsit
ASANH997-10|Hymenoptera|gs-NA|sp-NA|subsp-NA|country-Madagascar|exactsi
ASANW320-10|Hymenoptera|gs-Hypoponera|sp-Hypoponera MG080|subsp-NA|co
KHYME1140-11|Hymenoptera|gs-NA|sp-NA|subsp-NA|country-Kenya|exactsite-N
GMSAO094-13|Hymenoptera|gs-NA|sp-NA|subsp-NA|country-South Africa|exacts
GMESO904-14|Hymenoptera|gs-Aphytis|sp-NA|subsp-NA|country-Egypt|exactsite
GMMDM561-15|Hymenoptera|gs-NA|sp-NA|subsp-NA|country-Madagascar|exact
GMMDK1414-15|Diptera|gs-NA|sp-NA|subsp-NA|country-Madagascar|exactsite-N
GMSAU700-13|Diptera|gs-NA|sp-NA|subsp-NA|country-South Africa|exactsite-NA
GMMDD951-15|Diptera|gs-NA|sp-NA|subsp-NA|country-Madagascar|exactsite-NA
VVGC1554-11|Coleoptera|gs-NA|sp-NA|subsp-NA|country-Tanzania|exactsite-NA|
BGABB178-14|Coleoptera|gs-NA|sp-NA|subsp-NA|country-Gabon|exactsite-NA|lat
GMEGB012-14|Coleoptera|gs-Platystethus|sp-Platystethus degener|subsp-NA|cou
KDIPT2479-13|Diptera|gs-Tabanus|sp-Tabanus_donaldisoni|subsp-NA|country-Ke
GMMDC247-15|Diptera|gs-NA|sp-NA|subsp-NA|country-Madagascar|exactsite-NA
GMSAP466-13|Diptera|gs-NA|sp-NA|subsp-NA|country-South Africa|exactsite-NA|
GMKMH010-15|Diptera|gs-NA|sp-NA|subsp-NA|country-Kenya|exactsite-NA|lat_0
GBDP10789-12|Diptera|gs-Ceratitis|sp-Ceratitis_cosyra|subsp-NA|country-Kenya
GBDP15920-15|Diptera|gs-Hemipyrellia|sp-Hemipyrellia fernandica|subsp-NA|co
GMSAX535-14|Diptera|gs-NA|sp-NA|subsp-NA|country-South Africa|exactsite-NA
GMKKF223-15|Diptera|gs-NA|sp-NA|subsp-NA|country-Kenya|exactsite-NA|lat_-4
GMMDC694-15|Diptera|gs-NA|sp-NA|subsp-NA|country-Madagascar|exactsite-NA
ETKII1538-13|Orthoptera|gs-NA|sp-NA|subsp-NA|country-South Africa|exactsite-N
GMMDK1612-15|Diptera|gs-NA|sp-NA|subsp-NA|country-Madagascar|exactsite-N
GMSAE268-13|Diptera|gs-NA|sp-NA|subsp-NA|country-South Africa|exactsite-NA|
GMEGT487-14|Diptera|gs-NA|sp-NA|subsp-NA|country-Egypt|exactsite-Mostafa K
MIMAD029-15|Lepidoptera|gs-Fulcrifera|sp-NA|subsp-NA|country-Madagascar|ex

enafroCOI_all_data.tre

Nodes: 386351

External nodes: 193176

Internal nodes: 193175

Branches: 386351

Depth: 518

Maximum distance to root:
10.26635

Archaeopteryx visualization

Current progress...

What next?

- Studying/analysing the phylogenetic tree using: RAxML8, FastTree and T_Coffee
- Phylogeographic analysis

Points to Note:

- Open science principles: Collaborative supervision through GitHub, Open source software, and aspire to make scripts/codes transparent, available, free and accessible

Timeline

Activity	Time in months (2018-2019)											
	Sept	Oct	Nov	Dec	Jan	Feb	March	April	May	June	July	Aug
Proposal writing and Literature Review	■	■	■									
Data Mining and Sorting			■	■	■	■						
Pipeline Development and Testing				■	■	■	■	■				
Data Analysis on HPC								■	■	■	■	
Manuscript Writing and submission										■	■	■
Thesis writing and Defence										■	■	■

Acknowledgement

Thank you for enabling a bioinformatics dream

BOLD SYSTEMS

icipe

African Insect Science for Food and Health

JOMO KENYATTA UNIVERSITY
OF
AGRICULTURE AND TECHNOLOGY

Smithsonian Institution

EANBiT

Eastern Africa Network for Bioinformatics Training

NATIONAL MUSEUMS OF KENYA

WHERE HERITAGE LIVES ON