

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 Rono
DRIP fellow

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

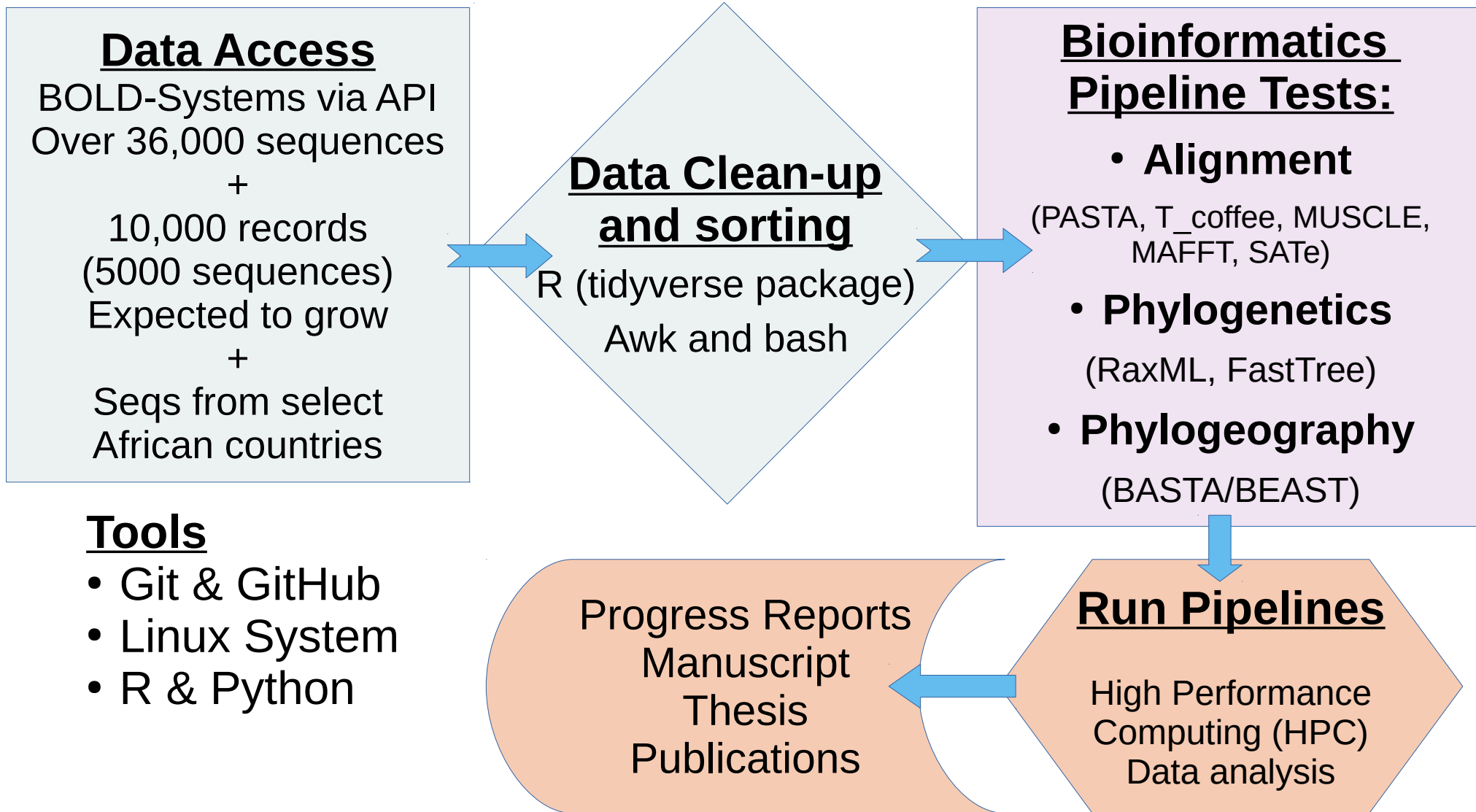
Problem statement:

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



Tools

- Git & GitHub
- Linux System
- R & Python

Data Mining and Clean up

Sequence retrieval:

BOLD-Systems via API; using specific syntax:

“arthropoda Kenya Uganda Tanzania Rwanda Burundi”

Data clean up:

R (tidyverse package), Awk and Bash

Metadata (80 columns)				
'COI-5P' = 35990 out of 37257 (1 sample)				
#nucleotides / unaligned seqs / #ns				
Over 700 (592)	Under 500 (705)	500 -700 (34693)	650 -660 (21886)	Over 500 (35285)
1 sample	1 sample	3 samples	2 samples	
Build.fasta: ProcessID order seq_len seq				

Multiple Sequence Alignment

Large dataset:

- Accuracy
- Speed

Algorithms:

- Progressive (mafft/muscle)
- Progressive & transitivity (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	***
Low accuracy	High accuracy	High accuracy	High accuracy	***
<ul style="list-style-type: none"> • Refine • Align • Merge 	<ul style="list-style-type: none"> • Align • Add sequence • Add_fragments • Merge 	<ul style="list-style-type: none"> • Align • Evaluate: (CORE index TCS) 	<ul style="list-style-type: none"> • Align • Add fragments 	***

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

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
```

```
aatagtggggaacttcttaagaa
attaattggatcatcaataagaa
agtgggtgggacctcattatctt
-----ggtataatattaagaa
-----taggatcagctttaagaa
tatagttggtttatcaataaqt
-----acatcaataagaa
-----ggaataataactaagaa
tatagtaggaataataactaagaa
cataattggagcctcattagaa
```

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/
AGTGGTGGGACCTCATTATCTTG/
TATAGTTGGTTTATCAATAAGTTT/
AATAATTGGATCATCAATAAGTTT/
AATAGTTGGAACATCAATAAGAAT/
TATAGTTGGAACATCAATAAGAAT/
```

T_coffee

```
T-COFFEE, Version_12.00.
Cedric Notredame
CPU TIME:0 sec.
SCORE=960
*
BAD AVG GOOD
*
GWOSM440-11|Lep : 96
KENM01124-13|Di : 96
AFPHY090-14|Lep : 96
SAARA227-11|Lep : 96
VVG61706-11|Col : 97
```

```
-----AAGTTTATATTTTATTTT
-----AACTTTATATTTTATTTT
-----TTTATTTT
-----ATATTGTATTTTATTTT
-----AACACTATATTTTATTTT
-----TTCATTTT
-----CACTTTATATTTTATTTT
-----AACTTTATATTTTATTTT
-----GTTTTATATTTTGT
-----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/
-----GGTATAATATTA/
-----GGAATAATACTAA/
TATAGTAGGAATAATACTAA/
-----ACATCAATAA/
CATAATCGGAACATCATTAA/
AGTGGTGGGACCTCATTAT/
AATAGTTGGAACCTTCTTTAA/
AATAGTAGGAACCTTCTTTAA/
```


Current progress...

- Build my Data set
- Setting up a RAxML8 and FastTree pipelines
- Improve on alignments: translation-alignment-threading of DNA
- High Performance Computing (HPC) analysis
- Phylogeographic analysis

Timeline

Activity	Time in months (2018-2019)											
	July	Aug	Sept	Oct	Nov	Dec	Jan	Feb	March	April	May	June
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

