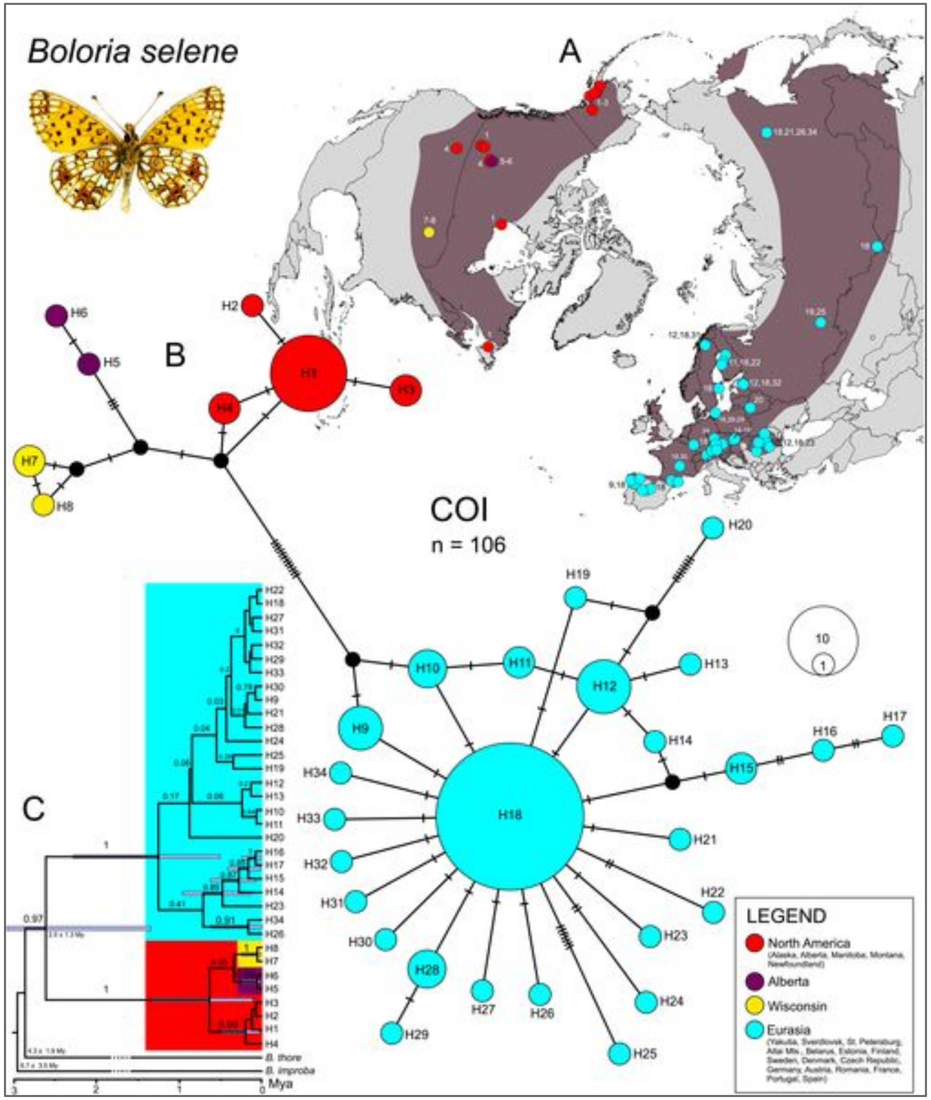
Phylogenetic and phylogeographic meta-analysis of African arthropod cytochrome c oxidase 1 barcode sequences submitted into the Barcode of Life Database

Gilbert Kibet-Rono*

Caleb Kibet, Jean-Baka Domelevo Entfellner, Steven Nyanjom, Daniel Masiga, Scott Miller, Jandouwe Villinger

Evolution and population dynamics of arthropods



Boloria selene
A
B
C
COI
n = 106
H1
H2
H3
H4
H5
H6
H7
H8
H9
H10
H11
H12
H13
H14
H15
H16
H17
H18
H19
H20
H21
H22
H23
H24
H25
H26
H27
H28
H29
H30
H31
H32
H33
H34
10
1
LEGEND
● North America (Alaska, Alberta, Manitoba, Montana, Newfoundland)
● Alberta
● Wisconsin
● Eurasia (Yakutia, Sverdlovsk, St. Petersburg, Altai Mts., Belarus, Estonia, Finland, Sweden, Denmark, Czech Republic, Germany, Austria, Romania, France, Portugal, Spain)
B. thore
B. improba
Mya
3
2
1
0

Maresova et al. (2019)

Phylogeography: Studies evolutionary and population dynamics and processes behind them

Population dynamics: migration, range separation, gene flow, population size

➔ **biotic factors** - predation

➔ **abiotic factors** - climate oscillation

Phylogenetics: Population structure

Understanding population dynamics of arthropods (Insects) lead to better pest and vector management

Background

Mitochondrial cytochrome c oxidase subunit 1 gene (**COI**), ~658 base-pair, is used for molecular identification of most animal phyla

The Consortium for the Barcode of Life (CBOL), May 2004: Rapid and inexpensive identification of species using standard DNA barcodes

- **International Nucleotide Sequence Database Collaborative (INSDC)**
- **Barcode of Life Database (BOLD)**

Over 320,000 African COI arthropod sequences published in BOLD

43,245 unpublished records from Kenya

Rationale

Problem statement:

- Few studies exists on African insects
  - lack of comprehensively sampled data and
  - a well-developed/documented bioinformatics workflow

Objectives:

- Develop a well documented bioinformatics workflow for phylogenetic and phylogeographic analysis of Insects sequences
- Implement the use of COI barcode sequences retrieved from BOLD.

Methodology

1. DATA RETRIEVAL

BOLD Public Data API
wget > XML

2. DATA MINING

A. Data transformation:
BeautifulSoup4, pandas > TSV

B. Data analysis, cleaning & sorting:
R tidyverse package > TSV files & statistics

3. Bioinformatics Pipeline

A. Data Preprocessing:
AWK/sed/egrep > FASTA

B. Sequence classification
RDPClassifier & BLAST

C. Multiple Sequence Alignment MSA
MAFFT, PASTA, MUSCLE, OPAL
T-COFFEE > FASTA/Phylip/clw

D. Phylogenetic Inference:
FastTree & RAxML >
Newick/PhyloXML/NEXUS

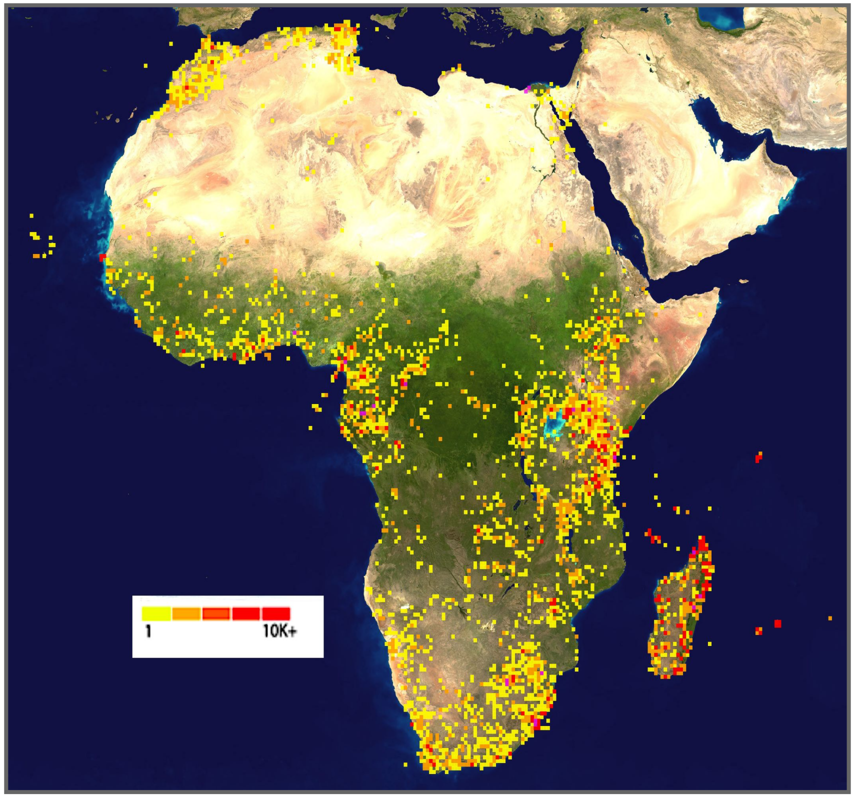
E. Population Structure:
POPART, GENELAND, ARLEQUIN

F. Phylogeography:
BEAST2/BASTA

Others: Biopython, PGDspider

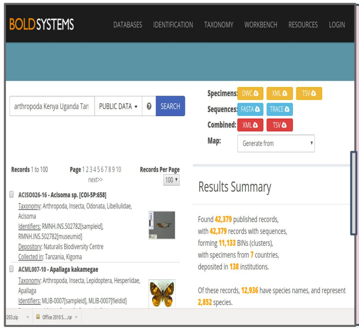
Results: data

323,034 “arthropod” COI sequences from Africa, 81,328 are from East Africa, 76.3% are arthropoda, 60% are Insecta



Class Insecta orders	Frequency
Lepidoptera (80 families)	76275
Diptera (61 families)	48137
Hymenoptera	40902
Coleoptera	12323
Hemiptera	8116
Orthoptera	1890
Odonata	1571
Psocodea, Blattodea, Mantodea, Trichoptera, Thysanoptera, Neuroptera, Ephemeroptera, Ephemeroptera, Dermaptera	3557
Embioptera, Phasmatodea, Plecoptera, Strepsiptera, Mecoptera, Zygentoma	69
Undefined	11

Data Retrieval, Transformation and sorting



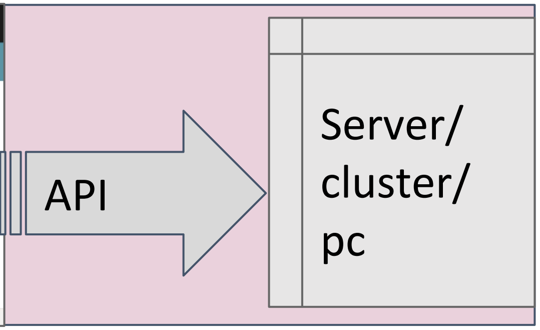
ACISO026-16 - Acisoma sp. [COI-5P:658]
Taxonomy: Arthropoda, Insecta, Odonata, Libellulidae, Acisoma
Identifiers: RMNH.INS.502782[sampleid], RMNH.INS.502782[museumid]
Depository: Naturalis Biodiversity Centre
Collected in: Tanzania, Kigoma

ACML007-10 - Apallaga kakamegae
Taxonomy: Arthropoda, Insecta, Lepidoptera, Hesperiidae, Apallaga
Identifiers: MLIB-0007[sampleid], MLIB-0007[fieldid]

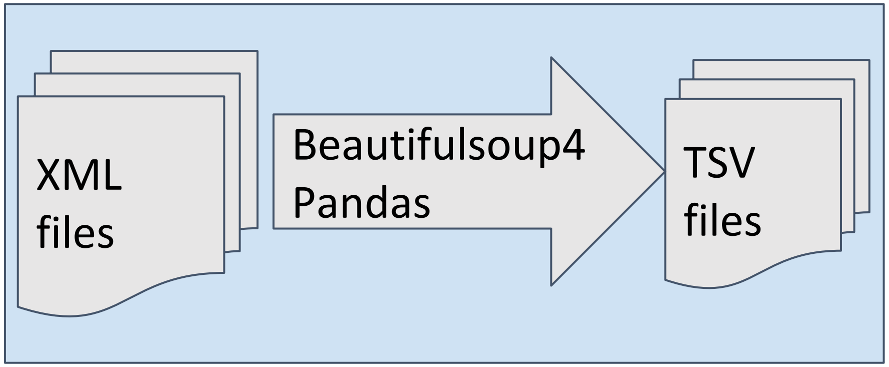
Results Summary

Found 42,379 published records, with 42,379 records with sequences, forming 11,133 BINs (clusters), with specimens from 7 countries, deposited in 138 institutions.

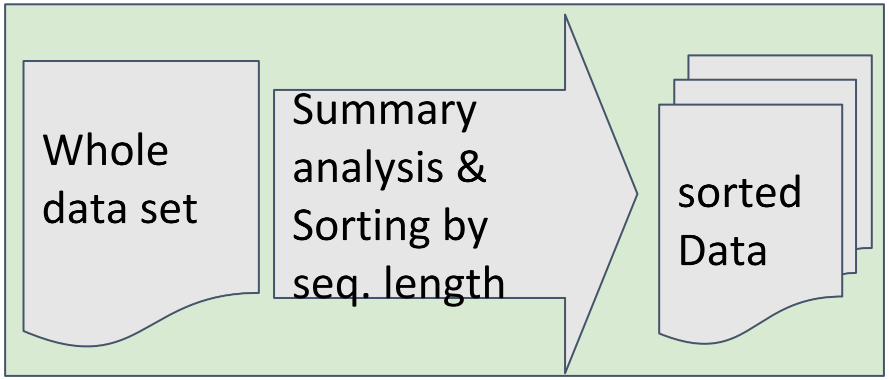
Of these records, 12,936 have species names, and represent 2,852 species.

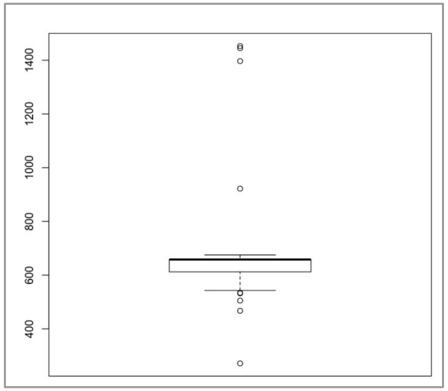


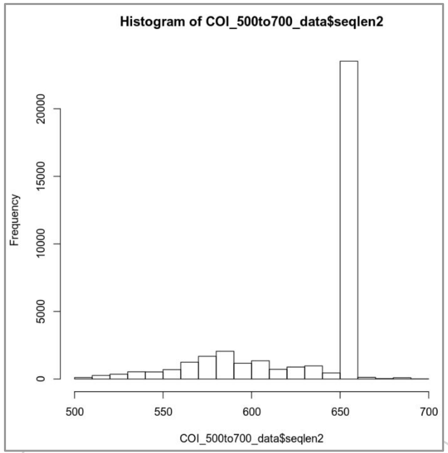
Data Retrieval:
Application Programming Interface (API) - XML files



Data Transformation:
A pandas dataframe of 84 columns and as many rows as records







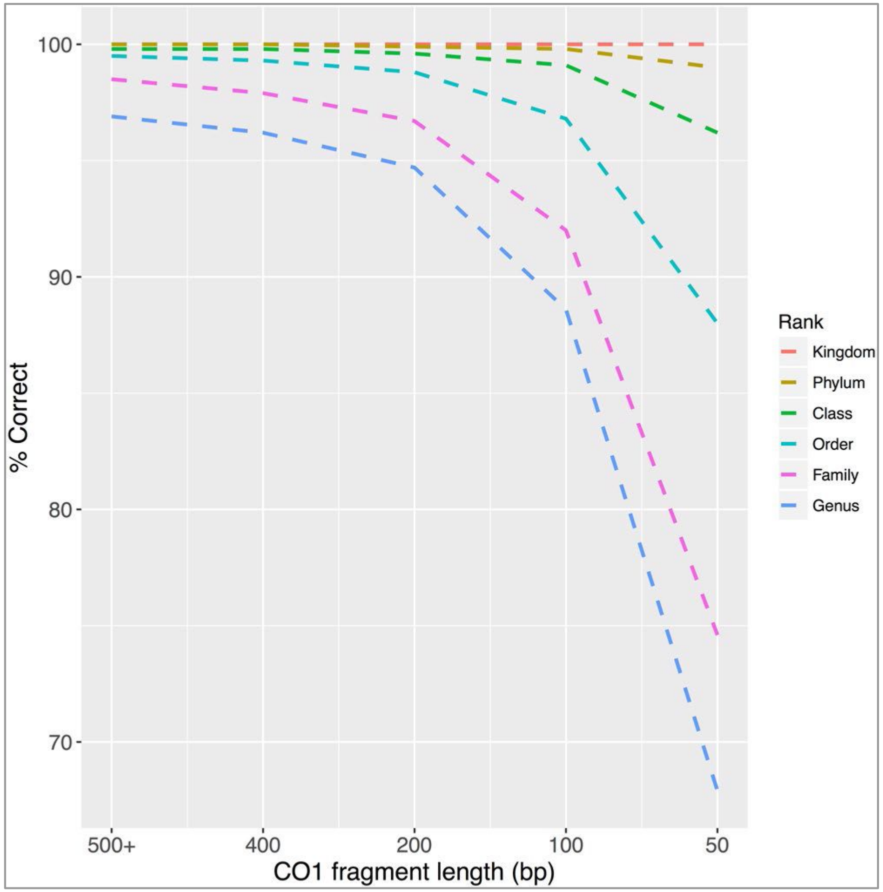
Classification metadata

African Arthropoda phylum;

- 11 classes
- 61 orders
- 562 families
- 3374 genera
- 9829 species

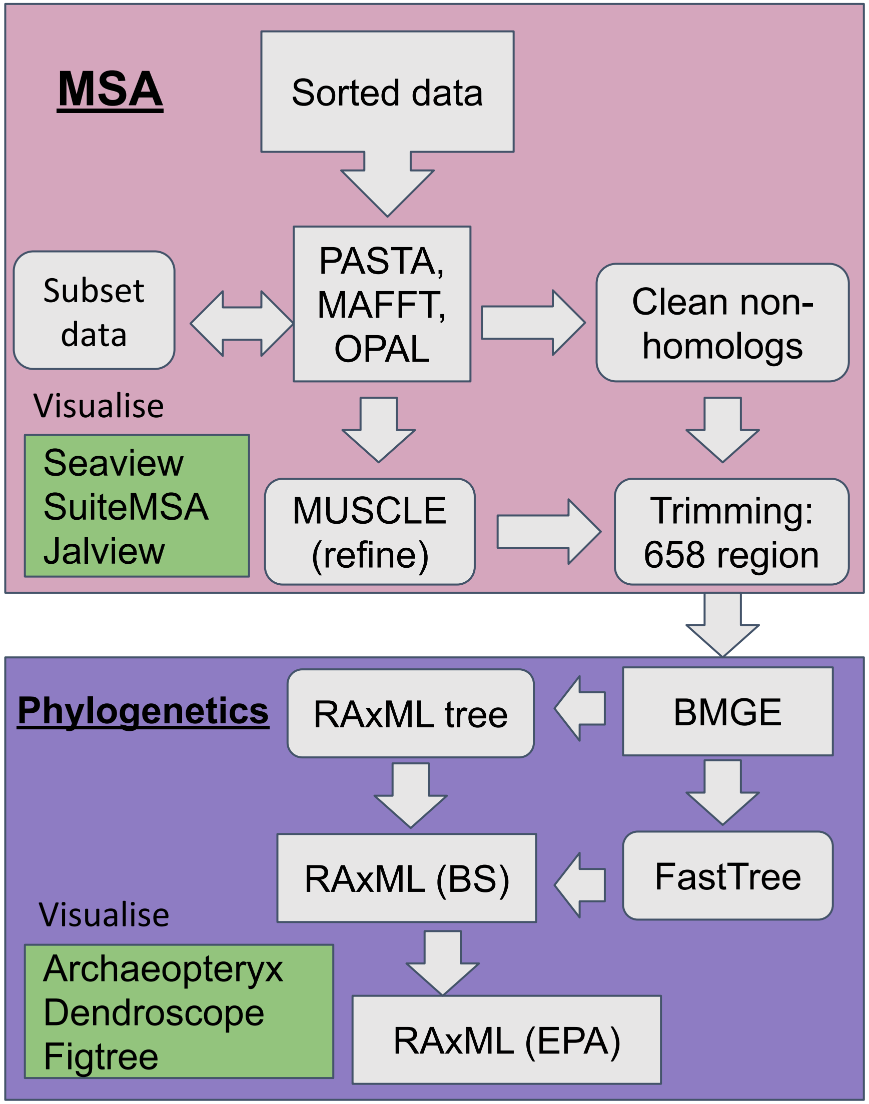
Data set	All Seqs	Seq.len over 500	Species labeled (Sp)	Sp labeled & Seq.len over 500
All Diptera	48137	47507	6629	6125
BS = 0.70	11337	10957	5124	4802
BS = 0.95	9244	8883	4785	4479
BS = 1.00	8049	7717	4407	4122

Ribosomal Database Project (RDP) classifier:



Porter and Hajibabaei (2018)

Results: workflow



MSA:

- PASTA - Practical Alignment using Sate and TrAnsitivity

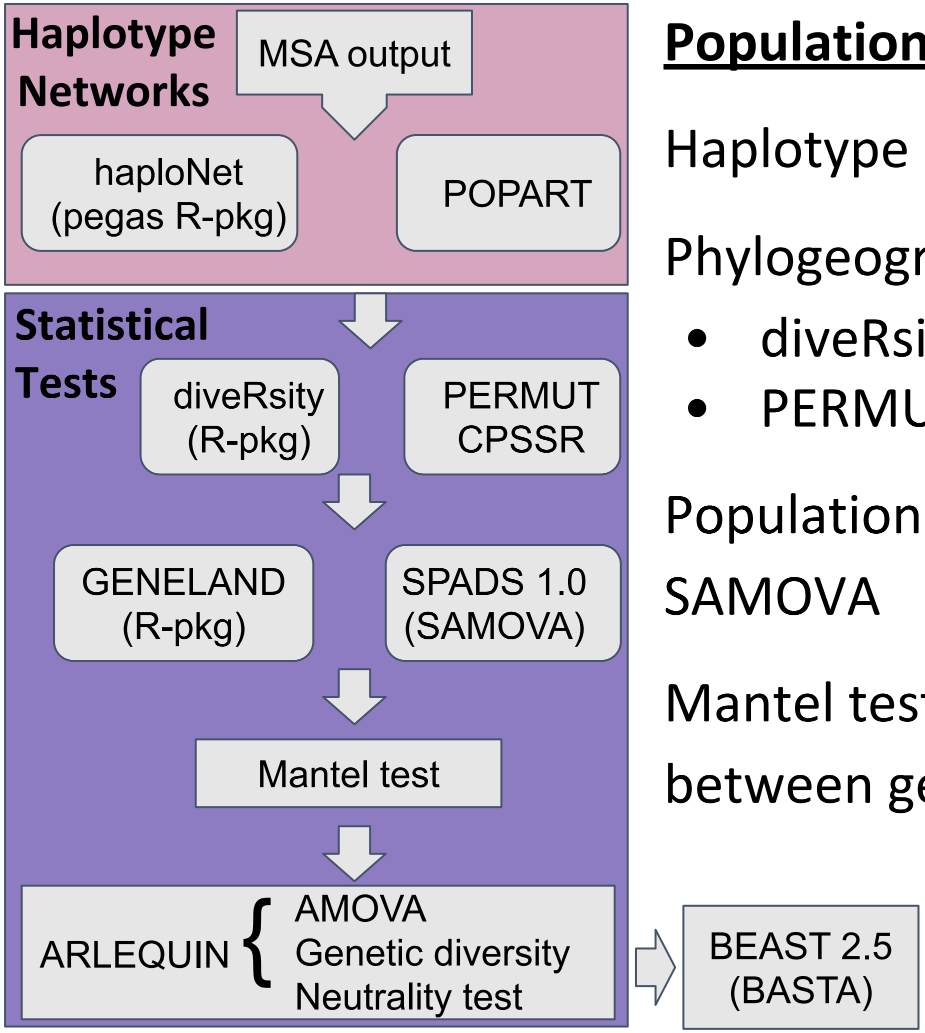
Phylogenetics:

BMGE - Block Mapping and Gathering with Entropy

RAxML:

- Tree inference - GTRCAT
- Bootstrapping
- Rooting - Evolutionary Placement Algorithm (EPA)

Results: workflow



Haplotype Networks

MSA output

haploNet (pegas R-pkg)

POPART

Statistical Tests

diveRsity (R-pkg)

PERMUT CPSSR

GENELAND (R-pkg)

SPADS 1.0 (SAMOVA)

Mantel test

ARLEQUIN { AMOVA, Genetic diversity, Neutrality test

BEAST 2.5 (BASTA)

<u>Population Genetic Structure:</u>

Haplotype networks - POPART, DNASP

Phylogeographic differentiation

- diveRsity R (Jost’s D and F_{ST})
- PERMUT CPSSR (G_{ST} and N_{ST})

Population spatial clusters (k) - GENELAND or SAMOVA

Mantel test - significance of correlation between genetic and geographical distances

Phylogeography:

Bayesian Structured Coalescent Approximation

Preliminary conclusions

Key challenges are:

- Missing/Inaccurate metadata-
    - taxonomic classification,
    - GPS, and
    - elevation data
- Resulting phylogenies are gene trees, limit their accuracy as species trees

Preliminary conclusions

BOLD COI data can be used in a number of population biology studies within different elevation spectrum, localities or clades:

- Phylogenetic diversity and gene flow
- Population dynamics
- Integrative taxonomy
- Biomonitoring: Invasive species and potential pests and vectors

Thank you

Smithsonian Institution

icipe@50 1970 – 2020 Insects for Life

H3ABioNet
www.h3abionet.org
Award Number: U24HG006941

Jomo Kenyatta University of Agriculture & Technology (JKUAT)

international BARCODE OF LIFE

NATIONAL MUSEUMS OF KENYA

International Centre of Insect Physiology and Ecology

P.O. Box 30772-00100, Nairobi, Kenya

Tel: +254 (20) 8632000

E-mail: icipe@icipe.org

Website: www.icipe.org
Support *icipe*: www.icipe.org/support-icipe

facebook.com/icipe.insects/icipe

twitter.com/icipe

linkedin.com/company/icipe