PHYLOGENETIC AND PHYLOGEOGRAPHIC META-ANALYSIS OF COI SEQUENCES OF EAST AFRICAN ARTHROPODS IN THE BOLD DATABASE

Objective:

- Improve phylogenetic and phylogeographic descriptions of arthropod diversity in East Africa

  Identify the diversity of species that may not yet be recognized as potential crop pests or vectors of human and animal diseases

- Allow for better assessments of emerging threat to human, animal and crop health

Kibet Gilbert (DRIP fellow)

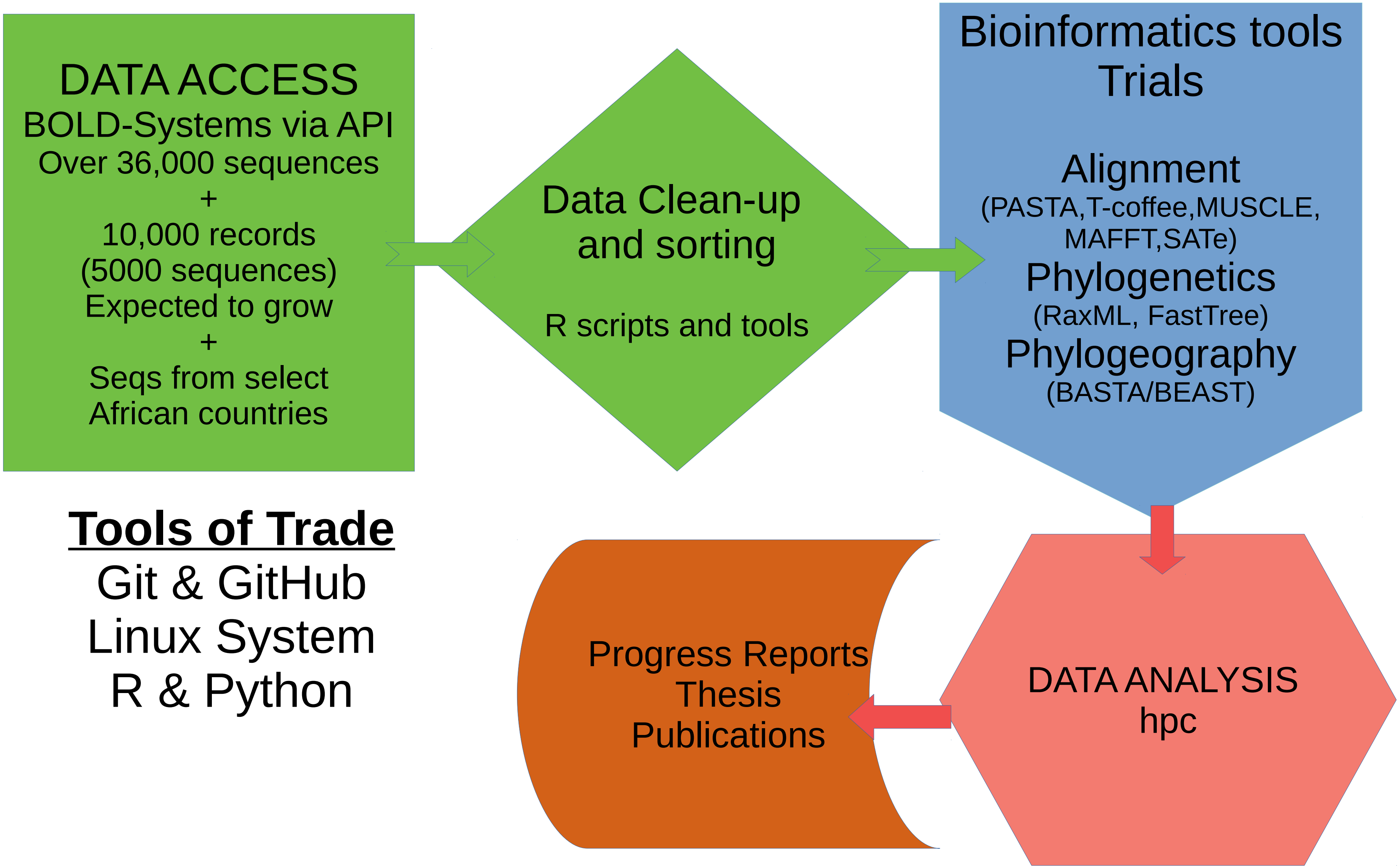
- **Supervisors:**

  Dr Scott Miller, Dr Jandouwe Villinger, Dr Steven Ger

- Collaborators:

  Dr Caleb Kipkurui, Dr Jean-Baka, Dr Dan Masiga

workflow



DATA ACCESS

BOLD-Systems via API
Over 36,000 sequences
+
10,000 records
(5000 sequences)
Expected to grow
+
Seqs from select
African countries

Data Clean-up
and sorting

R scripts and tools

Bioinformatics tools
Trials

Alignment
(PASTA,T-coffee,MUSCLE,
MAFFT,SATe)
Phylogenetics
(RaxML, FastTree)
Phylogeography
(BASTA/BEAST)

DATA ANALYSIS
hpc

Progress Reports
Thesis
Publications

Tools of Trade

Git & GitHub
Linux System
R & Python

Data Mining and Clean up

Sequence retrival:

BOLD-Systems via API; using specific syntax:

“arthropoda Kenya Uganda Tanzania Rwanda Burundi”

Data clean up:

R, Awk and Bash

<table>
<tr><th colspan="5">Metadata (80 columns)</th></tr>
<tr><td colspan="5">'COI-5P' = 35990 out of 37257 (1 sample)</td></tr>
<tr><td colspan="5">#nucleotides / unaligned seqs / #ns</td></tr>
<tr><td>Over 700 (592)</td><td>Under 500 (705)</td><td>500 - 700 (34693)</td><td>650 - 660 (21886)</td><td>Over 500 (35285)</td></tr>
<tr><td>1 smpl</td><td>1 smpl</td><td>3 smpls</td><td>2 smpls</td><td></td></tr>
<tr><td colspan="5">Build.fasta: ProcessID | order | seq_len | seq</td></tr>
</table>

MSA

Large dataset:

- Accuracy
- Speed

Algorithmn:

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

Muscle: default	Mafft: G_lins1	t_coffee: -reg	Pasta: default	sate
Speed	Speed	speed	speed	***
Low accuracy	Accurate	Accurate	Accurate	***
(refine, align and merge)	(align, add, addfragments, and merge)	(align, evaluate: core_index, Tcs,	(align and addfragments)	***

MSA evaluation

Evaluation for Accuracy:

- T_coffee: consistency based scoring.

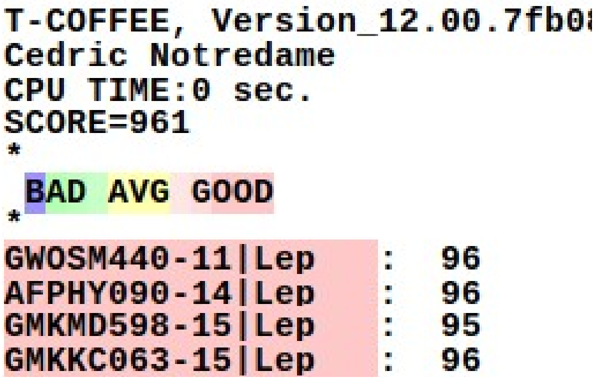
Scores:

- Core index (html)
- TCS (html/ascii)

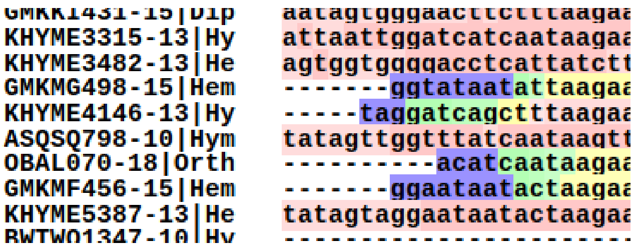
- Used to select the most suitable alignments.
- TCS ascii can be used to apply different weights to columns in phylogenetic analysis

MSA eval...

Mafft

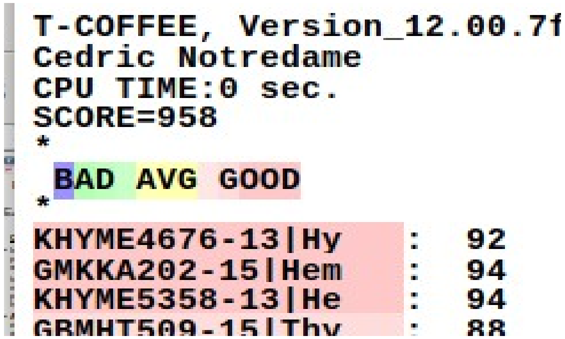


```
T-COFFEE, Version_12.00.7fb0
Cedric Notredame
CPU TIME:0 sec.
SCORE=961
*
 BAD AVG GOOD
*
GWOSM440-11|Lep   :   96
AFPHY090-14|Lep   :   96
GMKMD598-15|Lep   :   95
GMKKC063-15|Lep   :   96
```

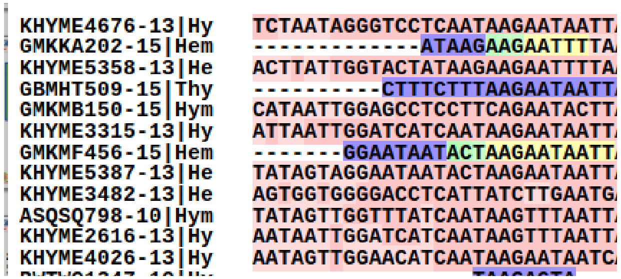


```
GMKK1431-15|Dip    aatagtgggaacttctttaagaa
KHYME3315-13|Hy    attaattggatcatcaataagaa
KHYME3482-13|He    agtggtggggacctcattatctt
GMKMG498-15|Hem    --------ggtataatattaagaa
KHYME4146-13|Hy    -----taggatcagctttaagaa
ASQSQ798-10|Hym    tatagttggtttatcaataagtt
OBAL070-18|Orth    -----------acatcaataagaa
GMKMF456-15|Hem    --------ggaataatactaagaa
KHYME5387-13|He    tatagtaggaataatactaagaa
BWTWO1347-10|Hy    -----------------------
```

Muscle

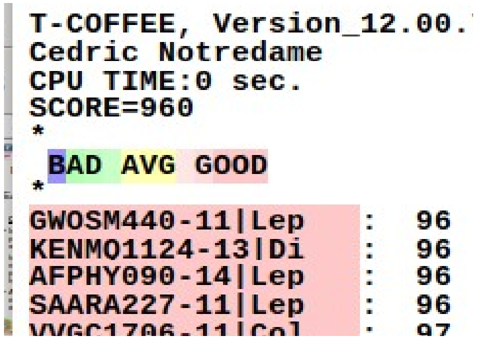


```
T-COFFEE, Version_12.00.7f
Cedric Notredame
CPU TIME:0 sec.
SCORE=958
*
 BAD AVG GOOD
*
KHYME4676-13|Hy    :   92
GMKKA202-15|Hem    :   94
KHYME5358-13|He    :   94
GBMHT509-15|Thy    :   88
```



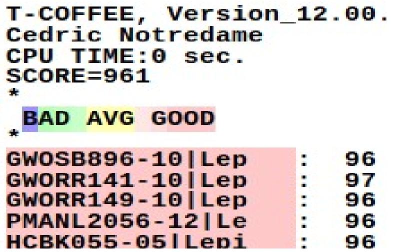
```
KHYME4676-13|Hy    TCTAATAGGGTCCTCAATAAGAATAATT
GMKKA202-15|Hem    -------------ATAAGAAGAATTTTA
KHYME5358-13|He    ACTTATTGGTACTATAAGAAGAATTTTA
GBMHT509-15|Thy    ----------CTTTCTTTAAGAATAATT
GMKMB150-15|Hym    CATAATTGGAGCCTCCTTCAGAATACTT
KHYME3315-13|Hy    ATTAATTGGATCATCAATAAGAATAATT
GMKMF456-15|Hem    -------GGAATAATACTAAGAATAATT
KHYME5387-13|He    TATAGTAGGAATAATACTAAGAATAATT
KHYME3482-13|He    AGTGGTGGGGACCTCATTATCTTGAATG
ASQSQ798-10|Hym    TATAGTTGGTTTATCAATAAGTTTAATT
KHYME2616-13|Hy    AATAATTGGATCATCAATAAGTTTAATT
KHYME4026-13|Hy    AATAGTTGGAACATCAATAAGAATAATC
```

T_coffee



```
T-COFFEE, Version_12.00.
Cedric Notredame
CPU TIME:0 sec.
SCORE=960
*
 BAD AVG GOOD
*
GWOSM440-11|Lep    :   96
KENMO1124-13|Di    :   96
AFPHY090-14|Lep    :   96
SAARA227-11|Lep    :   96
VVGC1706-11|Col    :   97
```

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
```

Current progress...

Build my Data set

Setting up a RAxML8 and FastTree pipelines

Improve on alignments:
translation-alignment-threading of DNA

HPC analysis

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
TECHNOLOGY FOR DEVELOPMENT

Smithsonian Institution

EANBiT
Eastern Africa Network for Bioinformatics Training

NATIONAL MUSEUMS OF KENYA
WHERE HERITAGE LIVES ON