Malaria Genome Population Identification with HMM Signals

Devin J. McConnell (devin_mcconnell@my.uri.edu), Yana Hrytsenko, Noah M. Daniels
Department of Computer Science and Statistics, University of Rhode Island, Kingston ,RI

THE UNIVERSITY OF RHODE ISLAND

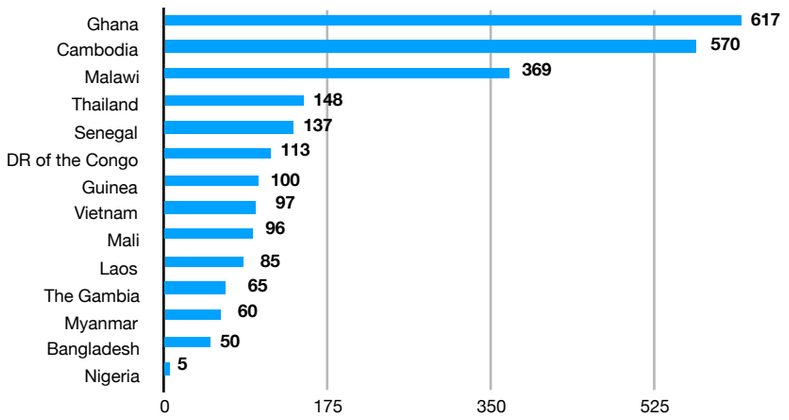
THINK BIG WE DO℠

Motivation:

There is currently a global effort to eradicate the parasitic disease malaria, caused by several species of the genus Plasmodium, primarily *P. falciparum.* A component of this effort is the identification of what geographic region a particular malaria infection originated from. To that effect, the Pf3k project has collected three thousand samples of P. falciparum from two different continents, Asia and Africa. In order to understand the effects of the eradication efforts, it is necessary to analyze these diverse samples to learn about the populations' structures. A primary goal of this research is to be able to identify, based on genomic sequence data, what parasitic population a particular infection stems from.

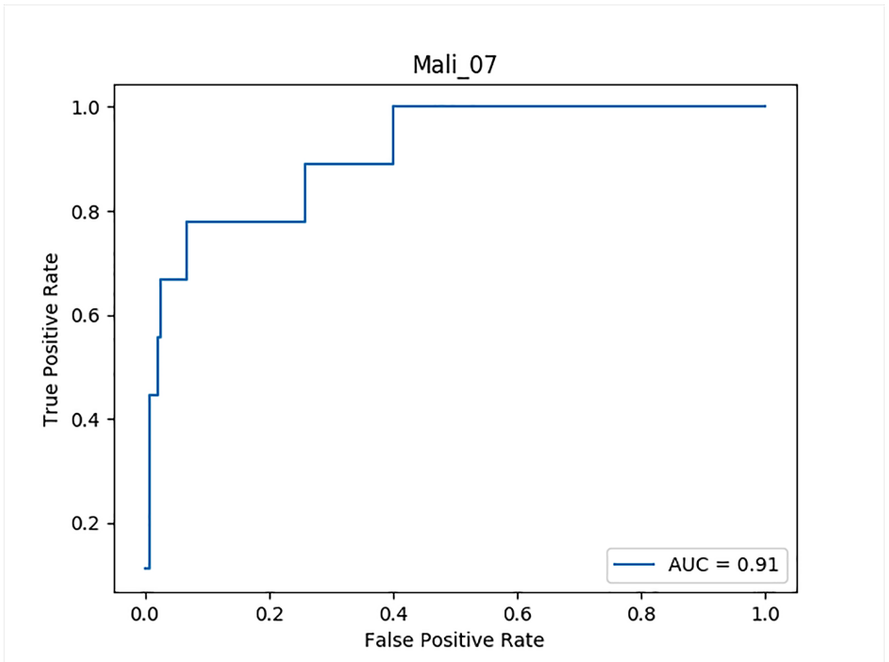
Eliminating Malaria:

Disease control centers around the world are working to eliminate Malaria. China has launched a national campaign to eliminate the disease by 2020.



10-Fold Cross-validation:

We built consensus sequence families through the use of Mauve multiple sequence aligner. Each consensus sequence was used to make a Hidden Markov Model and through strict 10-fold cross-validation we plotted the receiver operating characteristic curve.



Mali, chromosome 7, ROC curve graph. Mali consisted of 96 samples, 87 were used to train the HMM, and 9 were used to test the model.

Current Groupings:

Current groupings are separated by country. In each country they are further divided up into the 16 different chromosomes

Machine Learning with Hidden Markov Models:

Testing HMM signals for Malaria genomes, split into 10,000 nucleotides sections, produce a minimum AUC ≥ .07 for Mali, Senegal, and Nigeria, in at least four different chromosomes.

Additional Model Training:

- Training the models with a generalized clustering of the Pf3k data lead to successful preliminary results
- Compute dendrograms and use the clusters as the new population groups
- An example dendrogram below for Bangladesh chromosome 7 suggest 2 distinct populations
- Realign and compute HMM for each new population