Title: Phylogenetic analysis of five species of *Plasmodium* and their strains

Reference Paper: Molecular characterization and phylogenetic analysis of *Plasmodium vivax*, *Plasmodium falciparum*, *Plasmodium ovale*, *Plasmodium malariae* and *Plasmodium cynomolgi*

Supplemental File: CSE_185_Project_Supplemental_File

Team: Zina Patel (A15462246), Varsha Varkhedi (A15530051), Harshita Saha (A16662555)

Introduction: Malaria is a serious and fatal disease that causes about 2,000 cases in the United States each year (CDC Malaria). Protozoan parasites within the genus Plasmodium often cause malaria. Specifically, *Plasmodium vivax, Plasmodium falciparum, Plasmodium ovale, Plasmodium malariae,* and *Plasmodium cynomolgi* are five species that have historically impacted humans in ancient Greece and Rome, Asia, and Africa (Collins and Jeffery 2007, Price et al. 2007, Westling et al. 1997). Studying the evolutionary history of such pathogens via phylogenetic analysis of genomic data provides insights into traits of diseases like malaria, such as the method of spread, targeted host cells, and evolutionary strategies (Martinsen et al. 2007). The paper whose results we attempt to recreate in this project aims to conduct phylogenetic analysis to characterize the five species of *Plasmodium* listed previously, their strains, and the relationships between them (Chatterjee et al. 2017).

The goal of our project was to replicate the phylogenetic tree in the paper (Chatterjee, Figure 2). We generated four trees using our Bioinformatics pipeline, two of which were generated using data from all accession codes and methods GTRCAT and GTRGAMMA respectively (see section Results 1). The other two trees were generated using a subset of the accession code, excluding two highly divergent strains (see section Results 2), using both GTRCAT and GTRGAMMA methods. The trees generated using all accession codes generally displayed clade clustering similar to that of the figure in the paper, but highly varied in the inter-species and inter-strain relationships. In contrast, the trees generated on a subset of the accession codes almost perfectly replicated those of the paper, except for the *Plasmodium malariae* species, for which two accession codes were excluded due to being too divergent.

Methods:

All the packages, code, and other steps necessary to replicate this project can be found in the CSE_185_Project_Supplemental_File submitted alongside this report.

All files referenced in the steps below were stored in a directory called `CSE_185_Project`.

1. Downloading the Data

The paper uses 25 accession codes that link to NCBI for 18S rRNA sequences, 6 of which belong to strains of *P. vivax,* 4 for *P. cynomolgi*, 4 for *P. malariae*, 6 for *P. falciparum*, and 5 for *P. ovale*. The specific accession codes can be found in Figure 2 of the paper (Chatterjee, Figure 2). We first listed all the accession codes in a file called accessions.txt and used a custom bash script using curl to load in and compile all the data into a fasta file titled project_genomes.fa, code for which can be found in section 1 of the code in CSE_185_Project_Supplemental_File.

2. Multiple Sequence Alignment

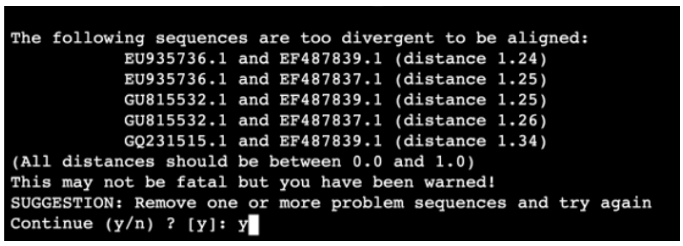
2.1 Downloading and Installing ClustalW

The paper used the package ClustalW for multiple sequence alignment, so we downloaded the ClustalW (version 2.1). We did so by navigating to the clustalw website and downloading the clustalw-2.1.tar.gz file to our local system, and then manually uploaded it to the CSE_185_project directory. To unzip the package file, we changed into the CSE_185_project directory via terminal and typed in the command tar xfvz clustalw-2.1.tar.gz. Because we did not have root access to install and use ClustalW, we had to install it for our user only, which can be done using the steps here and also outlined in Section 2.1 of the CSE_185_Project_Supplemental_File.

2.2 Running ClustalW:

To run ClustalW for multiple sequence alignment, we changed into the CSE_185_Project directory where the fasta file, project_genomes.fa, that will be the input for ClustalW, is stored. Then, we typed in clustalw2 into terminal and selected options as outlined Section 2.2 of the CSE_185_Project_Supplemental_File.

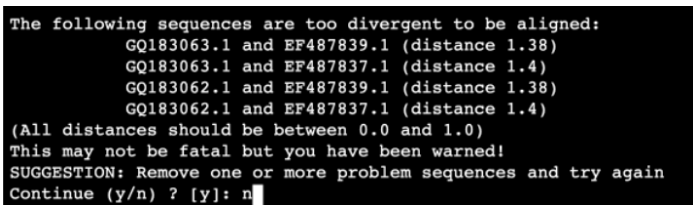
When running ClustalW on the project_genomes.fa file, the ClustalW algorithm displayed a warning as below.



When faced with the warning type in y. We will run the program again on the subset of the data such that a warning is not generated. The warning above lists accession codes that are too divergent to be aligned. Since the paper does not mention any divergence, we are not sure whether or not the divergence has enough of an impact on our tree to be considered fatal. To check for this, in addition to alignment of all accessions generated from project_genomes.fa file, we will download data from a subset of the accessions to run ClustalW on for comparison.

We created a file accessions_left.txt which only excluded accessions to the right of the warning (EF467837 and EF467839), and similarly created a file accessions_right.txt which only excluded accessions to the left of the warning, stored in the CSE_185_Project directory. We then repeated step 1 of Methods on the subsets listed in these txt files, code for which can be seen in Section 2.2 of the CSE_185_Project_Supplemental_File.

We repeat the steps for running ClustalW on project_genomes_left.fa successfully (Section 2.2 of the CSE_185_Project_Supplemental_File), this time with no warning. However, attempting to run ClustalW on project_genomes_right.fa (consisting of accession codes EF467837 and EF467839), a warning is displayed as below.



To deal with this warning by taking a subset, we would have to exclude a significant number of accessions making our results incomparable to the results generated on all accessions, so we only used ClustalW on the project_genomes_left.fa subset.

Generating the Phylogenetic Tree

3.1 Installing RaxML

Because the paper did not specify which algorithm was used to generate the phylogenetic tree, we decided to use RaxML (version 8.2.12). We did not have to download RaxML since we had done so as part of Lab 6. The steps to download RaxML (as in Lab 6) can be seen in Section 3.1 of the CSE_185_Project_Supplemental_File.

To call RaxML as part of this project, we navigate into the ~/local/bin directory where RaxML is stored and type in the command export PATH=$PATH:$HOME/local/bin, then cd back into the CSE_185_project directory.

3.2 Running RaxML on project_genomes.fasta:

Because the paper did not mention any parameters used in the generation of the phylogenetic tree, except that "Nodal values represent bootstrap probabilities based on 500 replicates", we decided to use both the GTRCAT and GTRGAMMA methods with 500 for the value of our bootstrap option. These methods use different parameters to generate the maximum likelihood tree, and are therefore likely to output different trees, which is why we used both as we were unsure which one would give results closer to those of the paper. The RaxML GTRCAT and GTRGAMMA code can be found in Section 3.2 of the CSE_185_Project_Supplemental_File.

3.3 Running RaxML project_genomes_left.fasta:

We then ran RaxML on the project_genomes_left.fasta file (output of running ClustalW on project_genomes_left.fa. The RaxML GTRCAT and GTRGAMMA code can be found in Section 3.3 of the CSE_185_Project_Supplemental_File.

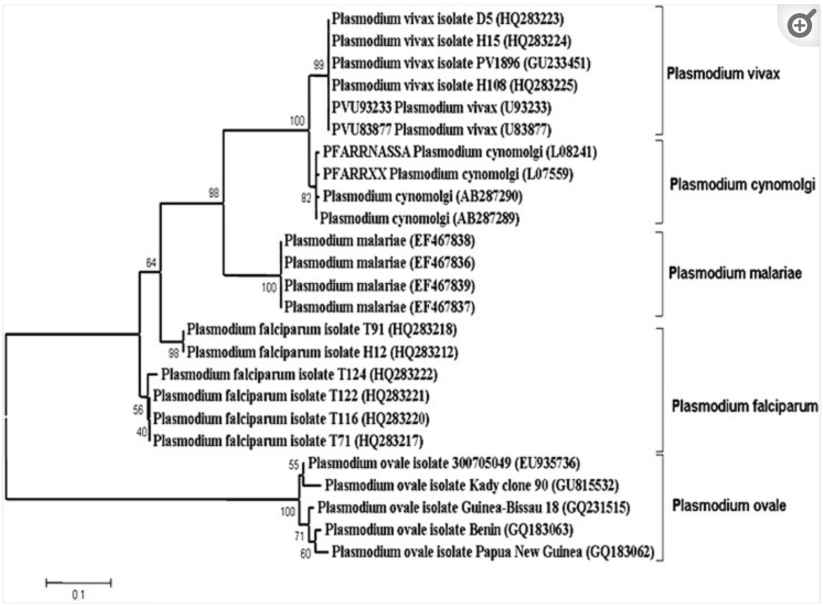
3. Visualizing Phylogenetic Trees

Because the paper did not specify which technology was used to generate the phylogenetic tree visualizations, we decided to use iTOL. For all trees, the text from the corresponding file starting with RAxML_bipartitionsBranchLabels was copied into the website and uploaded. The options selected on the control panel for display were Basic > Mode options > Branch lengths > Ignore and Advanced > Bootstraps > Display > Text > Label Background > On.

Results:

Phylogenetic Tree from the Paper (Chatterjee, Figure 2):

The image below is the phylogenetic tree that our project was attempting to produce. The phylogenetic tree traces the interspecies relationship of the five *Plasmodium* species listed.



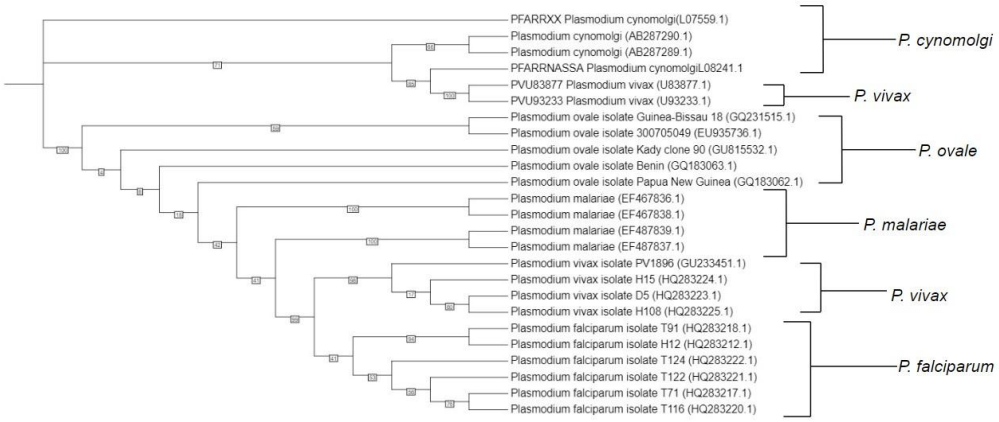
There is an initial ancestor (Ancestor 0) that splits into *P. ovale* and another intermediate species (Ancestor 1). Ancestor 1 splits into 4 strains of *P. falciparum* and another intermediate species (Ancestor 2). Ancestor 2 splits into 2 strains of *P. falciparum* and another intermediate species (Ancestor 3). Ancestor 3 splits into *P. malariae* and another intermediate species (Ancestor 4), and finally Ancestor 4 splits into *P. cynomolgi* and *P. vivax*.

For the interstrain relationships within each species, some strains diverge from intermediate strain ancestors, while others diverge from a common ancestor for all strains within that clade. We are unsure whether this is a result of actual multiple sequence alignment and tree output, or if researchers grouped some strains together due to a lack of confidence in the output of multiple sequence alignment and tree generation for those strains. This is why for comparisons of accuracy for the phylogenetic trees we generated, we will focus on the interspecies relationships displayed in the trees, since within the paper itself, the major motivation for the phylogenetic tree was to understand interspecies relationships.

1. Visualizing Phylogenetic Trees for project_genomes.fasta:

1.1 GTRCAT Tree for project_genomes.fasta:

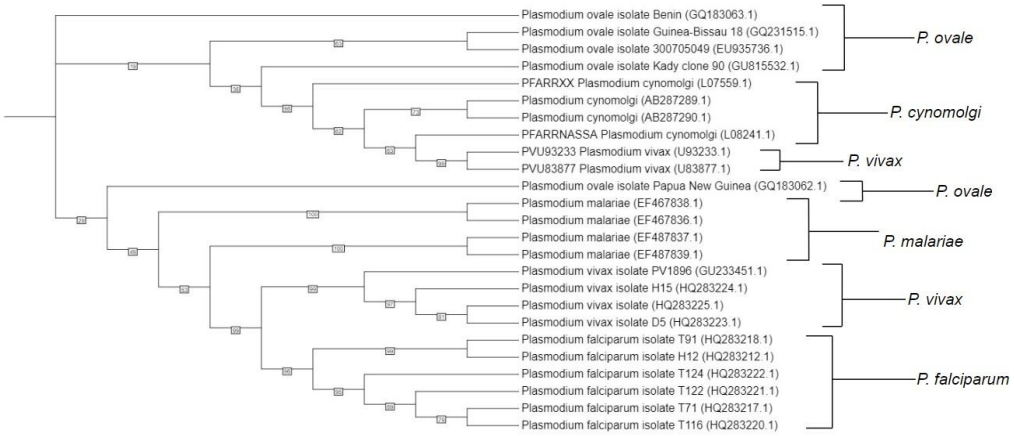
The tree below shows the phylogenetic tree we got from running RaxML GTRCAT on `project_genomes.fasta`. Although our tree appears to cluster major clades of the five species together as in the Figure 2 from the paper, except for two outlying *P. vivax* strains, there are major differences in the trees when it comes to interspecies relationships.



There is an initial ancestor (Ancestor 0) that splits into *P. cynomolgi*, 2 strains of *P. vivax*, and another intermediate species (Ancestor 1). Ancestor 1 splits into *P. ovale* and another intermediate species (Ancestor 2). Ancestor 2 splits into P. *malariae* and another intermediate species (Ancestor 3). Ancestor 3 splits into 4 strains of *P. vivax* and into *P. falciparum*.

1.2 GTRGAMMA Tree for project_genomes.fasta:

The tree below shows the RaxML GTRGAMMA tree from the file project_genomes.fasta. Although our tree appears to cluster major clades of the five species together as in the Figure 2 from the paper, except for two outlying *P. vivax* strains and one outlying *P. ovale* strain, there are major differences in the trees when it comes to interspecies relationships.



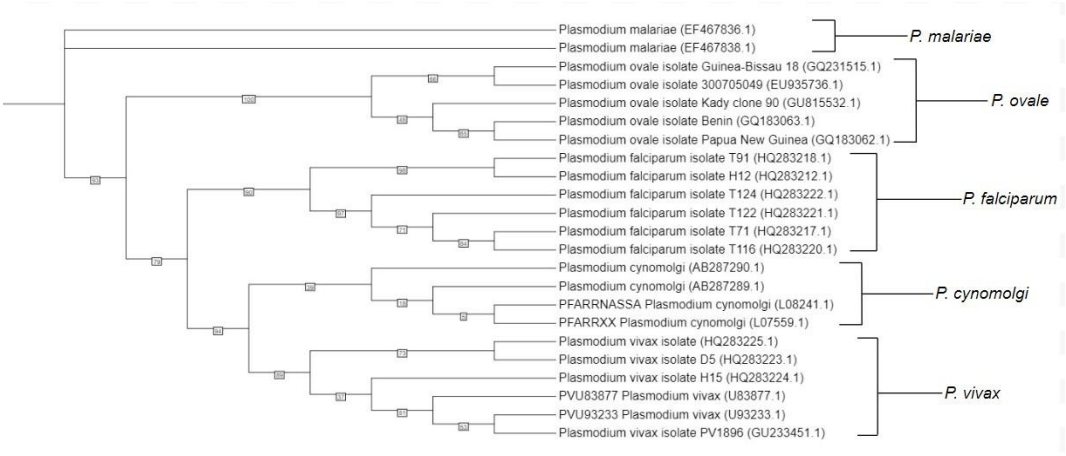
There is an initial ancestor (Ancestor 0) that splits into *P. ovale*, *P. cynomolgi*, 2 strains of *P. vivax*, and another intermediate species (Ancestor 1). Ancestor 1 splits into 1 strain of *P. ovale* and another intermediate species (Ancestor 2). Ancestor 2 splits into 2 strains of *P. malariae* and another intermediate species (Ancestor 3). Ancestor 3 splits into another 2 strains of *P.*

malariae, and into an intermediate species (Ancestor 4), Finally, Ancestor 4 splits into *P. vivax* and another intermediate species and into *P. falciparum*.

2. Visualizing Phylogenetic Trees for project_genomes_left.fasta:

2.1 GTRCAT Tree for project_genomes_left.fasta:

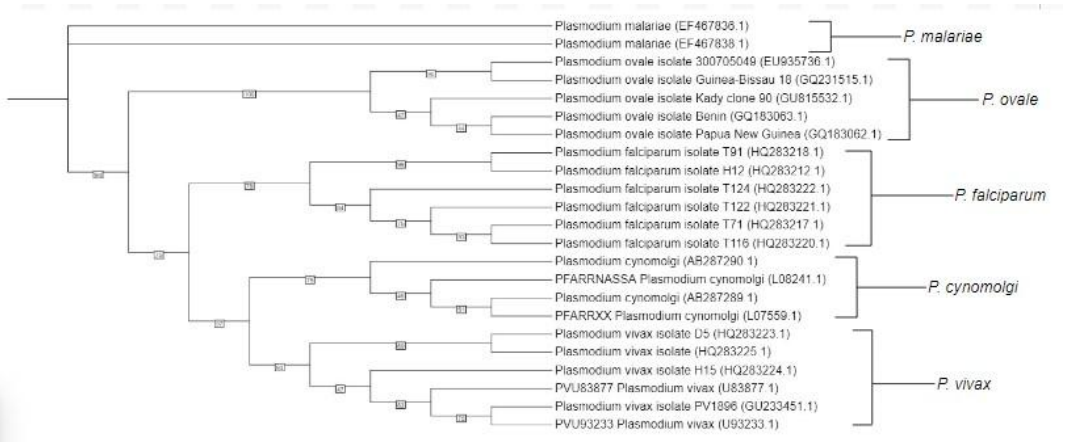
The tree below shows the RaxML GTRCAT tree from the file project_genomes_left.fasta. Our tree clusters major clades of the five species together as in the Figure 2 from the paper, and there are slight differences in the trees when it comes to interspecies relationships.



There is an initial ancestor (Ancestor 0) that splits into *P. malariae* and another intermediate species (Ancestor 1). Ancestor 1 splits into *P. ovale* and another intermediate species (Ancestor 2). Ancestor 2 splits into *P. falciparum* and another intermediate species (Ancestor 3). Ancestor 3 splits into *P. cynomolgi* and *P. vivax*. This sequence of branching is much closer to Figure 2 from the paper than our trees from Results Section 1 from the project_genomes.fasta.

2.2 GTRGAMMA Tree for project_genomes_left.fasta:

The tree below shows the RaxML GTRGAMMA tree from the file project_genomes_left.fasta. Our tree clusters major clades of the five species together as in the Figure 2 from the paper, and there are slight differences in the trees when it comes to interspecies relationships.



There is an initial ancestor (Ancestor 0) that splits into *P. malariae* and another intermediate species (Ancestor 1). Ancestor 1 splits into *P. ovale* and another intermediate species (Ancestor 2). Ancestor 2 splits into P. *falciparum* and another intermediate species (Ancestor 3). Ancestor 3 splits into *P. cynomolgi* and *P. vivax*. This sequence of branching is much closer to Figure 2 from the paper than our trees from Results Section 1 from the project_genomes.fasta.

Discussion:

1.1 Explaining differences in results:

The trees generated from project_genomes_left.fasta were closest to reproducing the interspecies relationships from the paper (Chatterjee, Figure 2), as seen in Results Figure 2.1 and Figure 2.2. In contrast, the interspecies evolutionary paths in the trees generated using GTRCAT and GTRGAMMA trees from project_genomes.fasta did not resemble the reference tree (as described in Sections 1.1 and 1.2 of Results).

The GTRCAT and GTRGAMMA trees from project_genomes_left.fasta, the subset data excluding certain accession codes that generated a divergence warning in ClustalW, were very similar to the interspecies relationships of the reference tree in terms of both clade clustering and ancestral path. The only exceptions were two *P. falciparum* strains, and the *P. malariae* species, the latter of which can be explained by the fact that the accession codes excluded from this subset belonged to this species.

The reference tree (as described in the Results section) has an initial ancestor that splits into *P. ovale* and another intermediate species (Ancestor 1). In contrast, the trees generated using project_genomes_left.fasta (both of which have the same interspecies relationships, as

described in Results sections 2.1 and 2.2) have the initial ancestor split into *P. malariae* and another intermediate species (Ancestor 1).

After this split however, the trees we generated look nearly identical to the reference (except for the *P. malariae* species cluster). Just as in the reference, an ancestor species splits into *P. ovale* and another ancestor, which splits into *P. falciparum* and another ancestor, which in turn splits into *P. cynomolgi* and *P. vivax*.

For the left accessions trees shown in two trees above, the GTRCAT and GTRGAMMA were similar to each other unlike for the original, large dataset, and there were no outlying strains in this case. When compared to trees that were produced using the full dataset, we found that our subset trees show a pattern of ancestry almost the same as that of the paper. The only exception was the *P. malariae* species, which is the species for which the warning suggested the exclusion of some strains due to excessive divergence.

This means that the divergence was likely fatal in our conclusions, and it is likely that the warning producing *P. malariae* strains were the sequences impeding replication of the figure from the paper, due to differences in output sequence alignment.

Unlike for the trees generated using `project_genomes.fasta`, the trees generated using project_genomes_left.fasta were nearly identical and did not have any outlying strains, with clade clustering replicating that of the reference tree. The differences in these trees as compared to the reference is specific to the *P. malariae* species. The exclusion of accession codes containing 18S rRNA sequences of *P. malariae* strains (with reasoning and steps described in the section 2.2 of Methods) affected the multiple sequence alignment generated using ClustalW, in turn affecting the phylogenetic tree generating using RaxML. This is why it is expected that using the subset of data from `project_genomes_left.fasta` would generate outcomes different from the reference specifically for *P. malariae*.

1.2 Limitations and challenges:

Potential limitations of our analysis that may have caused our results to be different from that of the paper despite starting with the same raw data have to do with the technology and methods used in multiple sequence realignment and phylogenetic tree generation. Although the paper specifies that the package used to conduct multiple sequence alignment on all accession codes was ClustalW, the researchers did not state the version of the package that was used in their study, so we used ClustalW version 2.1. If the version in the paper was different from ours, it is

likely that the algorithm also ran differently, giving us a different output than the paper for multiple sequence alignment. Another challenge we faced with running ClustalW was that the algorithm returned a warning when called on the fasta containing 18S rRNA sequences for all accession codes (seen and dealt with as described in Methods Section 2.2). The paper does not mention any such occurrence of divergence being fatal to the phylogenetic tree generation process, or impeding their results, making it likely that our methods were significantly different than theirs, which would also explain some of the differences between the tree in the paper and the trees generating in this project.

The paper did not specify which software packages were used for finding the maximum likelihood tree, which might have caused our tree to appear different. Because of a lack of direction from the research paper and based on our prior knowledge from Lab 6 of this class, we decided to use RaxML version 8.2.12. In addition to not knowing which package was used for generating the tree, we were unclear as to what parameters to use to make the tree when running RaxML. Thus, we used both the GTRCAT and GTRGAMMA models within RaxML. If the researchers did in fact use a different package than RaxML, their output for the maximum likelihood tree is likely to be different than ours, or if they did in fact use RaxML, potentially even with a different version, they may have used different method parameters than we did.

1.3 Improvements in pipeline:

To better replicate the results in the paper, it is imperative that we use the same or reasonably similar packages and methods used in the paper by the researchers. This can be done by directly emailing the researchers to ask them about which packages and methods were used for generating the tree, and which methods used for multiple sequence alignment. Alternatively, the bioinformatics pipeline could incorporate multiple tree generating algorithm tools and find which output is closest to that from the paper. In addition, for the bootstrapping step, the value used in the paper was 500 replicates, which is what we used as well. However, since the 18S rRNA sequence files are small enough that time constraints for generating replicates will not be an issue, it would be beneficial to generate a maximum likelihood tree with at least 1000 replicates to increase the confidence in the resulting phylogenetic tree.

References:

"CDC - Malaria - about Malaria." *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, 2 Feb. 2022, https://www.cdc.gov/malaria/about/.

Chatterjee, Soumendranath, et al. "Molecular Characterization and Phylogenetic Analysis of *Plasmodium Vivax*, *Plasmodium Falciparum*, *Plasmodium Ovale*, *Plasmodium Malariae* and *Plasmodium Cynomolgi*." *Journal of Parasitic Diseases : Official Organ of the Indian Society for Parasitology*, Springer India, Mar. 2017, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5339205/.

Collins, William E, and Geoffrey M Jeffery. "Plasmodium malariae: parasite and disease." *Clinical microbiology reviews* vol. 20,4 (2007): 579-92. doi:10.1128/CMR.00027-07.

Martinsen, E. S., et al. "Morphologically Defined Subgenera of Plasmodium from Avian Hosts: Test of Monophyly by Phylogenetic Analysis of Two Mitochondrial Genes: Parasitology." *Cambridge Core*, Cambridge University Press, 6 Dec. 2006, https://www.cambridge.org/core/journals/parasitology/article/morphologically-defined-subgenera-of-plasmodium-from-avian-hosts-test-of-monophyly-by-phylogenetic-analysis-of-two-mitochondrial-genes/E845E6FDBE879EEB4820786638B40B6C.

Price, Ric N et al. "Vivax malaria: neglected and not benign." *The American journal of tropical medicine and hygiene* vol. 77,6 Suppl (2007): 79-87.

Westling, Jennifer, et al. "Plasmodium Falciparum, P. Vivax,ANDP. Malariae:A Comparison of the Active Site Properties of Plasmepsins Cloned and Expressed from Three Different Species of the Malaria Parasite." *Experimental Parasitology*, Academic Press, 25 May 2002,https://www.sciencedirect.com/science/article/pii/S0014489497942259?via%3Dihub.

Appendix:

Varsha found the paper used for this project and produced the tree visualizations in iTOL. Harshita downloaded the data and required packages, and conducted the bioinformatics pipeline to create the phylogenetic trees starting with raw data. Zina helped with downloading the data and required packages, and documented the code. Harshita and Zina wrote up a draft of the project report, Varsha finalized the report, and all three teammates created the slides for the presentation.