Distribution Agreement

In presenting this thesis as a partial fulfillment of the requirements for a degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis in whole or in part in all forms of media, now or hereafter now, including display on the World Wide Web. I understand that I may select some access restrictions as part of the online submission of this thesis. I retain all ownership rights to the copyright of the thesis. I also retain the right to use in future works (such as articles or books) all or part of this thesis.

Briley SoYoung Park April 3, 2024

Existence of Selective Pressure in G6PD Deficiency alleles on Various African Population Groups Throughout Time

by

Briley SoYoung Park

Dr. David J. Cutler
Adviser

Department of Biology

Dr. David J. Cutler
Adviser

Dr. Skye Comstra

Committee Member

Dr. John Lindo

Committee Member

2024

Existence of Selective Pressure in G6PD Deficiency alleles on Various African Population Groups Throughout Time

By

Briley SoYoung Park

Dr. David J. Cutler

Adviser

An abstract of
a thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Science with Honors

Department of Biology

2024

Abstract

Existence of Selective Pressure in G6PD Deficiency alleles on Various African Population Groups Throughout Time

By Briley SoYoung Park

Glucose-6-phosphate dehydrogenase (G6PD) deficiency is a widespread genetic disorder that affects about 400 million individuals globally. This study explores the historical prevalence of G6PD deficiency and its correlation with different selective pressures from malaria and fava bean consumption by analyzing G6PD deficiency frequency across diverse African populations and periods. Utilizing ancient DNA data, this research seeks to uncover the evolutionary forces that have influenced the distribution of G6PD deficiency. By examining the G6PD gene variants in ancient and present-day populations, the study aims to provide deeper insight into human genetic diversity and the survival advantages conferred by these mutations.

Existence of Selective Pressure in G6PD Deficiency alleles on Various African Population Groups Throughout Time

By

Briley SoYoung Park

Dr. David J. Cutler

Adviser

A thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Science with Honors

Department of Biology

2024

Table of Contents

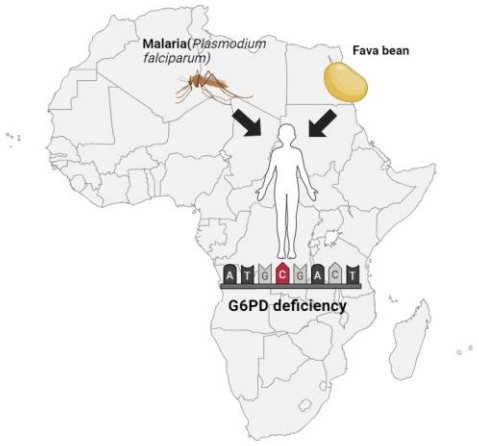
Existence of Selective Pressure in G6PD Deficiency alleles on Various African Population Groups Throughout Time

Briley SoYoung Park

Abstract

Glucose-6-phosphate dehydrogenase (G6PD) deficiency is a widespread genetic disorder that affects about 400 million individuals globally. This study explores the historical prevalence of G6PD deficiency and its correlation with different selective pressures from malaria and fava bean consumption by analyzing G6PD deficiency frequency across diverse African populations and periods. Utilizing ancient DNA data, this research seeks to uncover the evolutionary forces that have influenced the distribution of G6PD deficiency. By examining the G6PD gene variants in ancient and present-day populations, the study aims to provide deeper insight into human genetic diversity and the survival advantages conferred by these mutations.

Abstract Image



Introduction

G6PD and G6PD deficiency

Glucose-6-phosphate dehydrogenase (G6PD) is an enzyme involved in the pentose phosphate pathway, which plays a crucial role in protecting erythrocytes from damage caused by reactive oxygen species (ROS) [1]. G6PD deficiency, an X-linked recessive genetic disorder resulting from missense mutations, is one of the most common enzyme deficiencies [2, 3]. Due to its X-linked nature, G6PD deficiency is more frequently detected in males, who are hemizygous for the gene and less frequently in females, who must be homozygous for the deficiency to be affected [4].

In individuals with G6PD deficiency, erythrocytes become more susceptible to ROS, leading to premature hemolysis under certain environmental pressures. While less well-known in the United States, G6PD deficiency is more prevalent in the Middle East and Africa [5]. Tishkoff et al. discovered that distinct genetic variants of the G6PD gene are associated with the geographical and historical origins of different populations [3,10]. Consequently, the global distribution of G6PD deficiency is non-uniform, with variations that appear to correlate with migratory patterns and historical influences. Figure 1 illustrates that Africa is one of the most heavily impacted continents worldwide.

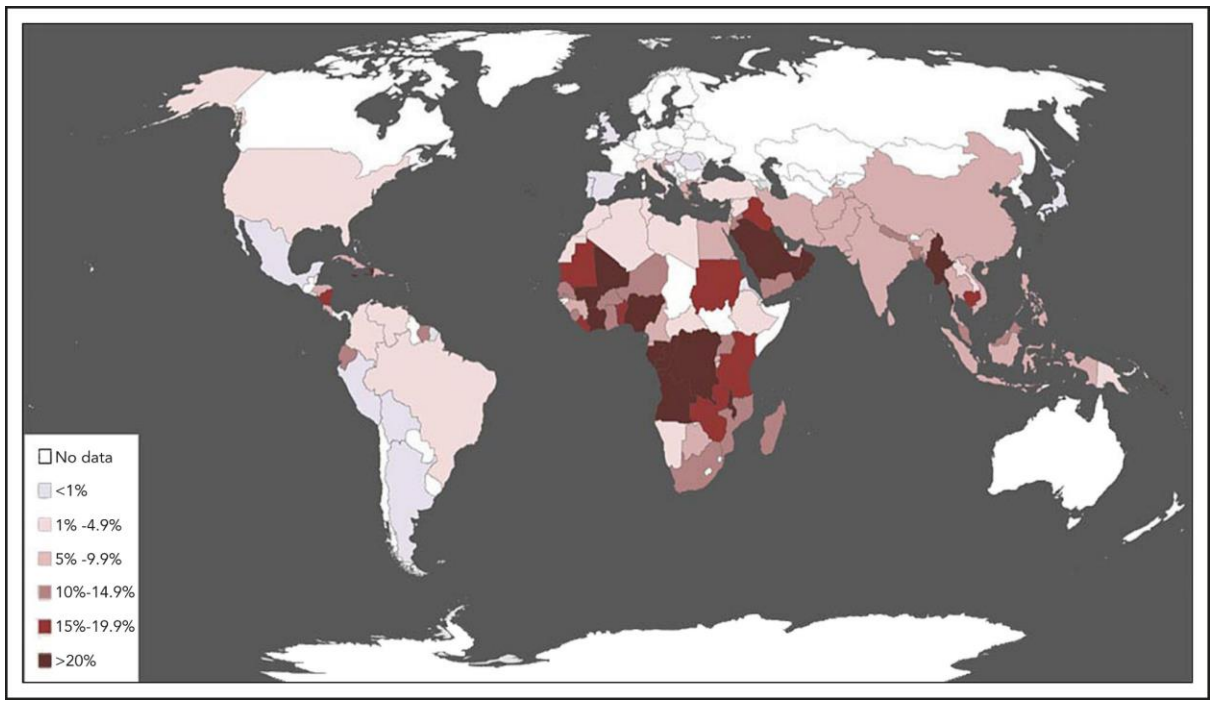


Figure 1. Epidemiology of G6PD Deficiency. This map, not produced by the author of this thesis, aggregates global data on the frequency of polymorphic G6PD alleles leading to an enzyme deficiency. The original work, conducted by Luzzatto et al., includes additional data and references detailed in the *Blood* Web site. This figure is reproduced under Creative Commons (CC) licensing from Luzzatto, L., Ally, M., & Notaro, R. (2020) in their work on glucose-6-phosphate dehydrogenase deficiency published in *Blood*.

Malaria and G6PD deficiency

G6PD deficiency is known to confer a heterozygous advantage against Plasmodium falciparum malaria [5]. Erythrocytes affected by G6PD deficiency provide an unfavorable environment for malaria parasites to grow [6]. Consequently, the frequency of G6PD deficiency is high in regions where malaria is endemic, as these characteristics provide an advantage against the progression of malaria into a fatal disease. In Africa, where malaria is endemic, the distribution of G6PD deficiency frequency ranges from 10% to 50%, while only 2.9% of the global population is known to be G6PD deficient according to the World Health Organization (WHO) [6,7].

Favism and G6PD deficiency

Fava beans (Vicia Faba), one of the oldest domesticated beans dating back to the start of human agriculture, are typically grown and consumed in North Africa, the Mediterranean, and the Middle East [8]. Favism, a disease caused by fava bean consumption, occurs in individuals with G6PD deficiency, inducing acute hemolytic anemia (AHA) [9]. Fava beans contain vicine and convicine, which are the main components that lead to increased free radicals in the intestines and trigger a hemolytic crisis [9]. Favism is particularly severe in children with G6PD deficiency.

Several studies have found higher rates of G6PD deficiency in African and Mediterranean populations, raising intriguing questions about human group history and the potential benefits these genetic variants confer in specific ecological contexts. In this senior thesis, I intend to conduct a comprehensive investigation of G6PD gene variants and their frequencies across diverse population groups and historical periods, with a particular focus on ancient DNA data. The aim is to determine if there is evidence of selective pressure exerted by fava bean consumption or a balancing selection by malaria. By studying ancient DNA data for G6PD deficiency mutations in distinct historical groups, I hope to relate the findings to the current distribution of G6PD deficiency. This study will expand on past genetic research to provide a more complete picture of the historical mechanisms that have shaped the global distribution of G6PD deficiency.

Hypothesis/Purpose

We aim to analyze ancient DNA data for G6PD deficiency mutations to understand how the interaction between fava bean consumption and malaria exposure influenced the frequency of G6PD deficiency in various African countries over time.

Method

Data Acquisition and Pre-processing

The DNA samples used in the study used an open-source Allen Ancient DNA Resource (AADR) dataset from Dr. David Reich's Lab at Harvard University - a dataset named 1240K + HO (.snp) containing 20503 unique individuals (9990 ancient, 10513 present-day). Different population groups were extracted from the dataset and will use variant- filtering to focus on the G6PD gene (X-chromosome in the Xp28 area). The dataset was converted into binary pedigree (bed), binary map (bim), and family (fam) files utilizing PLINK 2.0 software.

Extraction of G6PD regions and Data Categorization

Specific SNPs associated with the G6PD region (rs915942, snp_23_153712888, rs5945100, rs111429018, rs73573464, rs112776960) were identified through the UCSC Genome Browser of Human (GRCh37/hg19) and was extracted from the dataset using PLINK 2.0. The extraction process was guided by established genetic markers and known G6PD loci. Following extraction, the dataset was organized based on geographic origin and temporal classification (ancient, present) to analyze both spatial and temporal dimensions.

Genetic Distance and Linkage Disequilibrium

Allele frequencies of the known SNPs in the G6PD regions of each population group were calculated using PLINK 2.0. The frequencies were organized to calculate the genetic distance between populations. Nei's genetic distance equation was used to calculate the genetic distance in the R programming language.

$$D = -\ln \frac{\sum\sum XY}{\sqrt{(\sum X^2)(\sum Y^2)}}$$

The distance matrix between populations was visualized through Principal Component Analysis (PCA) plots and phylogenetic trees through R. Principal component analysis or PCA is a commonly used tool to reduce the dimensionality of the data and to capture the most variation in the genetic data as it simplifies analyzing and visualizes the genetic structure in populations. It is also a great tool to show the mitigation of genetic drift that might show a clue of random genetic changes that happened throughout time in the populations.

Linkage disequilibrium (LD) within populations was assessed in PLINK 2.0. The results were visualized as heat maps in R.

Results

PCA Plot Illustrates Temporal Genetic Changes

Principal component analysis (PCA) was carried out to visualize the genetic variance across different populations and periods as this method helps to reduce the complexity of the eigenvector genetic data, and individual-level SNP data in the G6PD region. Figure 2a illustrates the genetic variance within the dataset across ancient and present-day individuals. Principal Component 1 (PC1) shows the most genetic variation and Principal Component 2 (PC2) shows the second most genetic variation. The overlapping points between ancient and present-day populations along the PC1 and PC2 suggest genomic regions less impacted by selective forces. The black box in the middle was the region of interest as there was little to no overlap between ancient and present-day individuals suggesting a possibility of genetic drift or selection pressures. Individuals within this region exhibited PC1 values that ranged between -0.014 and -0.004 and PC2 values between -0.005 and 0.0025. This defined area predominantly includes East African individuals, specifically from Tanzania, Kenya, and Ethiopia as listed in Supplementary Table 1.

Figure 2b presents a PCA plot with the first and third principal components on the axes illustrating the genetic variation by continents. The image indicates how individuals from each continent from different continents form discrete clusters. The orange North American portion of the plot adjacent to the African portion indicates that most individuals from Barbados - Barbadians are known to have 77.4% African ancestry due to large migrations[13]. The African cohort, colored in blue, is notably clustered together, indicating there might have been a selective pressure such as malaria or fava bean consumption acting on the countries in the continent.

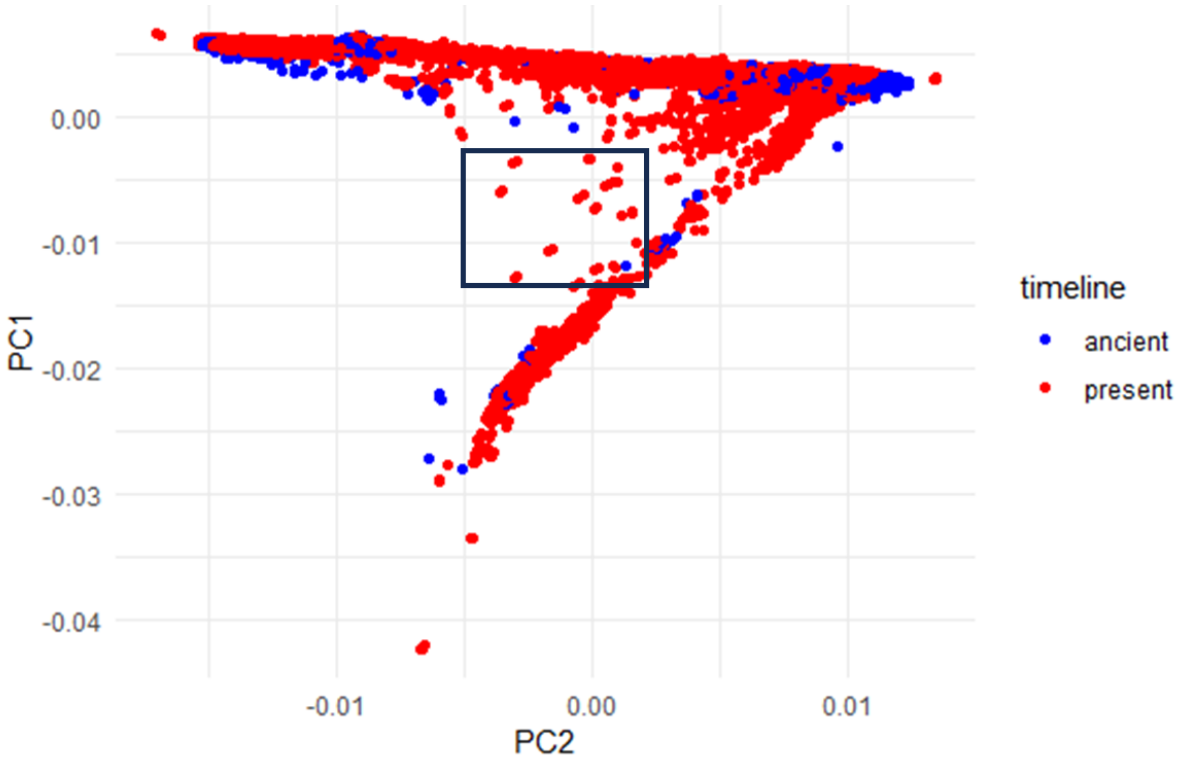


Figure 2a. Principal component analysis (PCA) plot showing the first two principal components (PC1, PC2). Each dot represents an individual. Ancient and Present populations are marked by blue and red respectively as indicated in the legend alongside the plot.

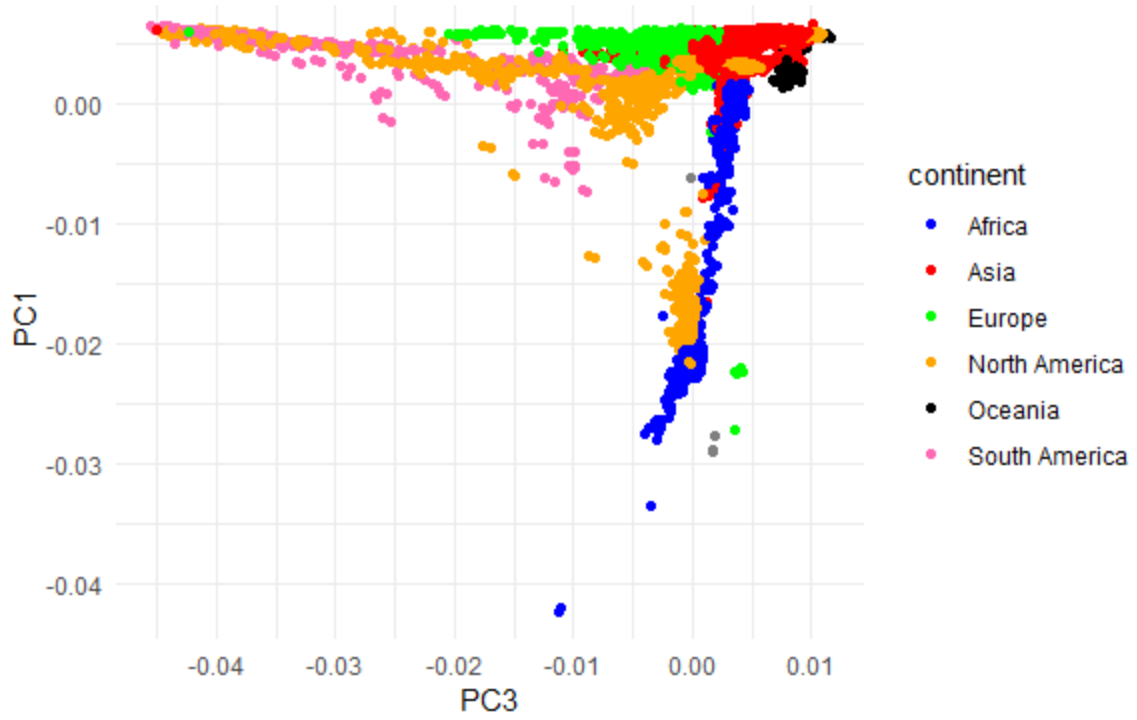


Figure 2b. Principal component analysis (PCA) plot showing the first and third principal components (PC1, PC3). Each dot represents an individual. Populations from different continents are marked by different colors as indicated in the legend alongside the plot.

Genetic Distance Between Populations

The PCA plot shown in Figure 3 was generated by calculating the genetic distance between populations and incorporating the color-grading according to the timeline. The color gradient represents the average age of the samples, ancient samples shown in green, and more recent samples transitioning into red. The plot highlights the genetic distinction between present-day populations from South Africa, Sudan, and Morocco, all in the African continent. This divergence in the African continent could be shown to reflect selective pressures. Sudan, located in East Africa, is known to have historically been affected by malaria6. This might have been a major factor in the selection of G6PD deficiency as it has an

advantageous effect against malaria.

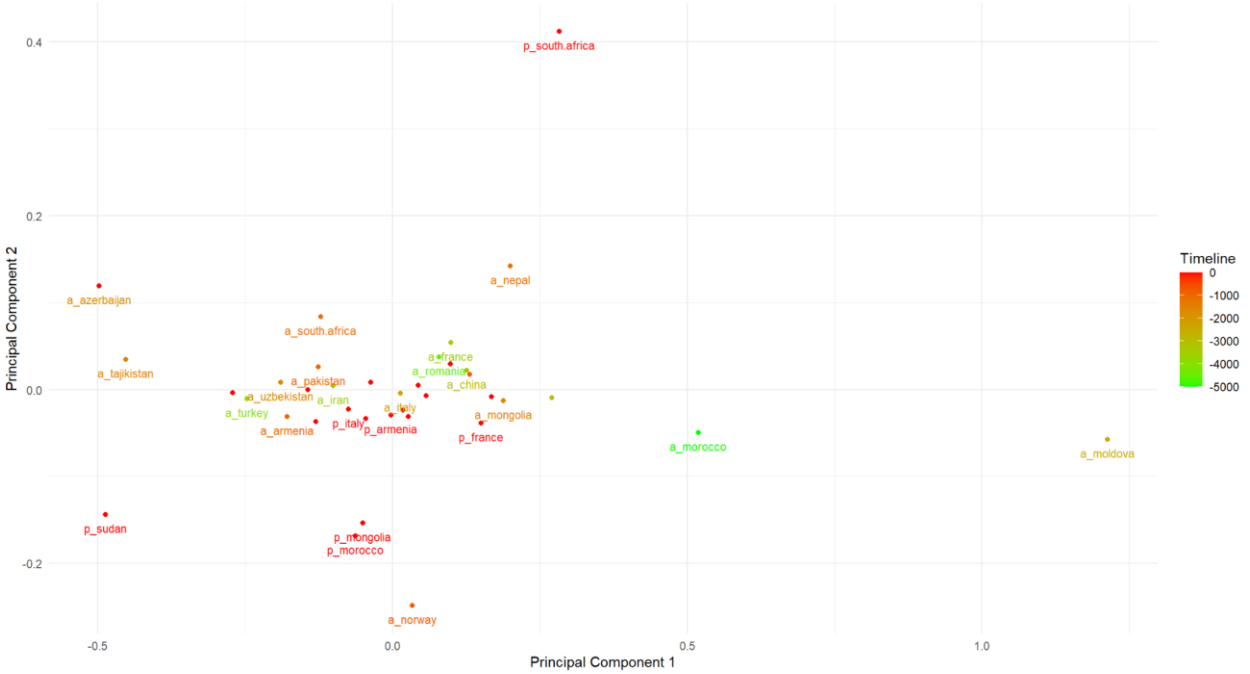


Figure 3. Principal Component Analysis (PCA) plot visualizes the genetic distances between ancient and present-day populations based on SNP alleles in the G6PD region. Temporal color gradient ranging from green (-5000 years, negative sign representing BCE dates) to red (0 years, or present). Each point represents ancient (‘a_’) and present (‘p_’) individuals.

Based on the hypothesis, the genetic variation in the G6PD gene is influenced by selective pressures such as fava bean consumption and the historical prevalence of malaria. Populations in geographic locations with a high prevalence of malaria would exhibit a higher frequency of G6PD deficiency alleles while regions with high fava bean consumption would impose a selective disadvantage due to the risk of hemolytic anemia in individuals with G6PD deficiency. To explore this hypothesis, the genetic distance between ancient and present-day populations was calculated based on the allele frequency using Nei’s genetic distance equation.

The analysis of the genetic distance between different population groups with single nucleotide polymorphisms (SNPs) associated with G6PD deficiency, could be visualized through the phylogenetic tree that illustrates how the population clusters in a distinct pattern in Figure 4a and b. The population groups in yellow indicate regions with both the presence of malaria and fava bean consumption exhibit genetic proximity to one another. The consistent pattern of clustering pattern indicates the balancing selective pressure is maintained despite potential risks associated with fava bean consumption. (include a table that shows the percentage of g6pd deficiency in each of the colors) The populations with a history of malaria exposure but with no known culture of fava bean consumption were denoted with blue, showed a similar clustering pattern, and supported malaria as the selective agent.

Branches color-coded with green represent populations without malaria presence but show high consumption of fava beans is more dispersed throughout the tree. This could indicate that there is a broader genetic variability in regions without malaria or that fava bean consumption does not have a high selective pressure on the alleles. Compared to the branches color-coded with gray, which represents populations with neither of the factors, the green group shows a similar clustering pattern compared to the yellow, or blue groups. The fava bean consumption was specified based on the crop growth, traditional dishes, and amount of imports during the year. Malaria prevalence was based on the World Health Organization (WHO) and countries were categorized as 'no malaria' when malaria was eradicated or the geographical condition was not fit for mosquitoes to inhabit.

While Figure 4a indicates the phylogenetic tree to visualize the genetic distance between population groups, it does not portray a distinct pattern between the four categories. The correlation between clustered branches color-coded differently and external factors like malaria and fava bean consumption underscore the potential genetic impact on the G6PD alleles.

Figure 4b does not significantly correlate with the geological relationships between countries unlike Figure 4a, but it does show a distinct pattern between the categories. Based on the divergence between the yellow and blue, green and gray, the result may suggest that the impact of malaria is a more significant driver of genetic differentiation compared to fava bean consumption even though both factors show a distinct pattern of clustering in Figure 4b. Comparing Figure 4a and 4b, the difference in the distinction in clusters indicates G6PD region is an adaptive locus selected for environmental pressures. The separation between two clusters is based on the presence and absence of fava bean consumption - however, we need additional data to find to attribute causation.

There are also limitations for phylogenetic trees. They may show genetic relationships and could show the correlation between environmental factors like malaria and fava bean consumption, there is a possibility of confounding factors that might have affected the G6PD deficiency prevalence.

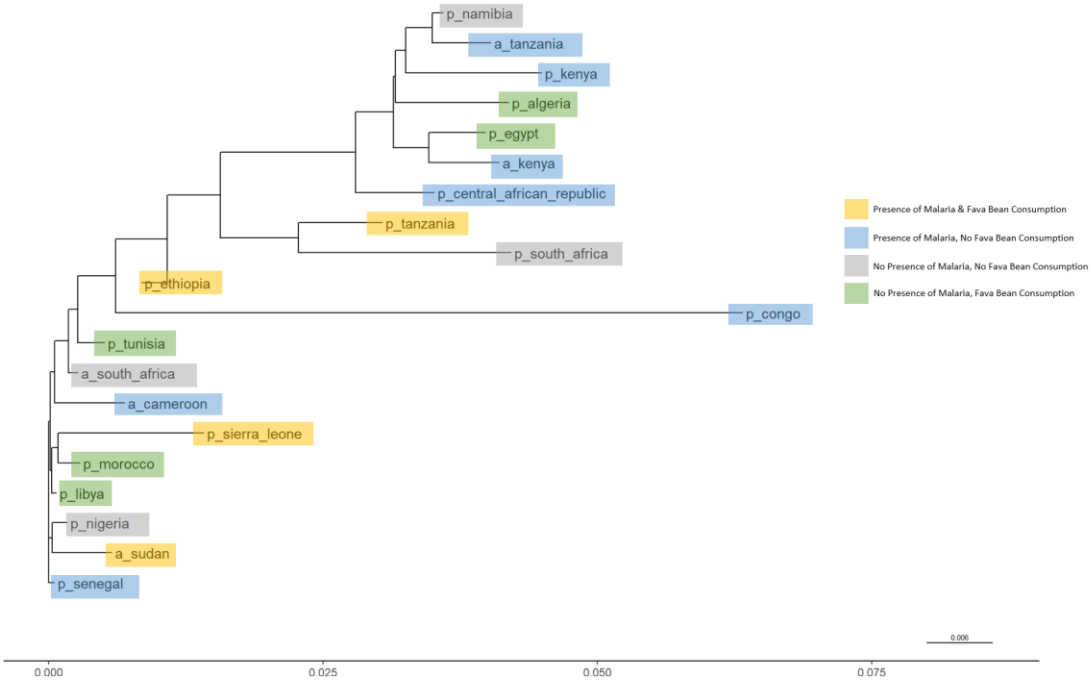


Figure 4a. Phylogenetic tree based on genetic distances between populations from African countries based on the whole genome. Countries are categorized into 4 groups - malaria and fava bean consumption, malaria with no fava bean consumption, no malaria with fava bean consumption, and no malaria and no fava bean consumption.

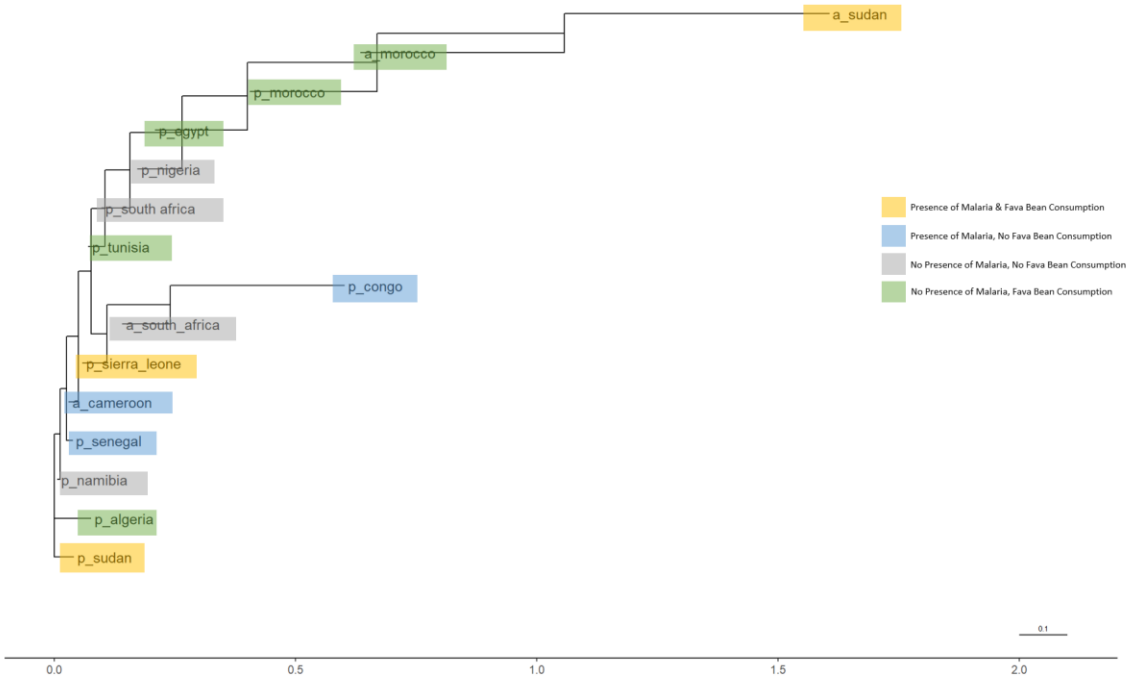


Figure 4b. Phylogenetic tree based on genetic distances between populations from African countries based on the SNPs in the G6PD region. Countries are categorized into 4 groups - malaria and fava bean

consumption, malaria with no fava bean consumption, no malaria with fava bean consumption, and no malaria and no fava bean consumption.

Linkage Disequilibrium and heatmaps

Malaria and fava bean consumption

Figure 5a presents linkage disequilibrium in the G6PD region by heatmaps for ancient and present Sudan and present Sierra Leone. While ancient Sudan and the present Sudan share a classification of malaria and fava bean consumption, their heatmaps exhibit significantly different LD patterns. This could be evidence of evolutionary change but the reduction of sample size of present Sudan compared to ancient Sudan populations could also play a crucial role in this pattern difference. The disparity between present Sudan and Sierra Leone points to the geographical separation limiting gene flow, despite having similar environmental pressures.

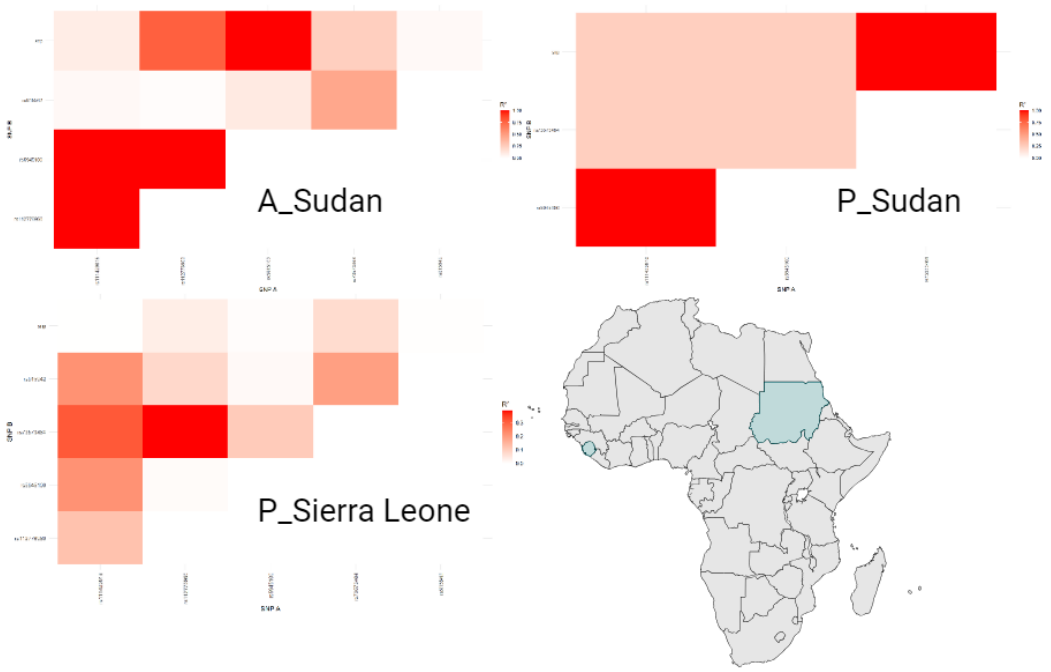


Figure 5a. Heatmaps showing the linkage disequilibrium (R^2 values) for SNPs in the G6PD region for populations in Sudan and Sierra Leone, and ancient samples from Sudan. Red intensity indicates the strength of LD between SNP pairs, with an accompanying African map with corresponding countries highlighted, generated from Biorender.

Malaria and no fava bean consumption

The category for malaria and no fava bean consumption included present-day populations from Congo, Senegal, Central African Republic, Nigeria, and ancient samples from Cameroon which are geographically located in the West African area. The linkage disequilibrium (LD) patterns of the G6PD regions across show a consistency of high LD strength between certain SNP pairs - rs5945100 and rs111429018, rs73573464 and rs112776960. It suggests how these alleles were inherited together more often than expected by chance. This consistent pattern across the countries could suggest a shared selective pressure and in this case, it could be predicted as malaria. Ancient malaria mosquito fauna had been present in West Africa since the interglacial period which dates back to 100,000 years ago, [11] Ancient samples from Cameroon were around 6000 BCE from radiocarbon dating, around the Neolithic period. The congruence of the LD patterns between ancient Cameroon and the present-day West African population provides evidence for a stable genetic association, reinforcing that malaria has been consistent over millennia.

Present-day Senegal, on the other hand, displays a unique pattern with a different pair of SNP with a high R^2 value from other West African countries. As shown in the map with highlighted countries, Senegal's relative geographical distance from the other countries may have resulted in less gene flow or different environmental pressures that may have led to a different LD pattern.

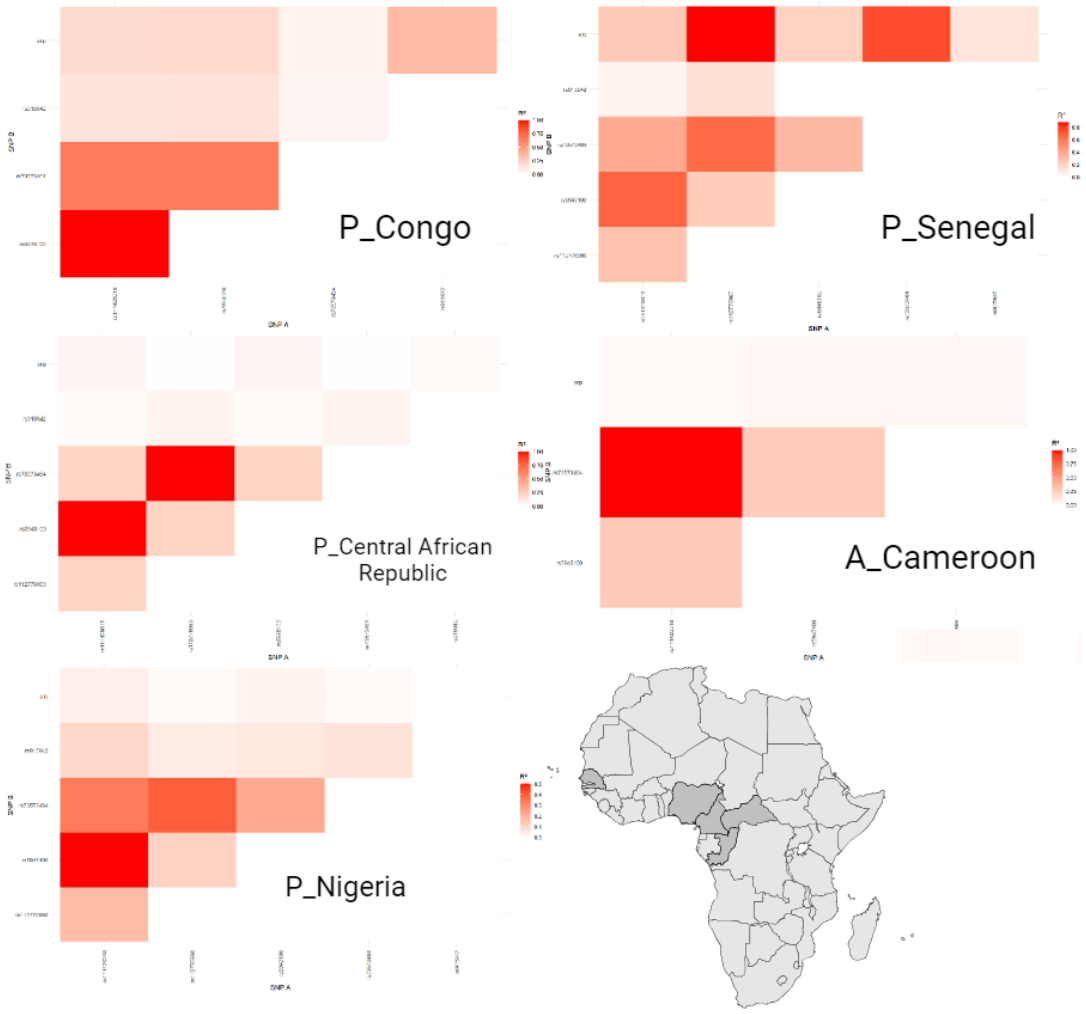


Figure 5b. Heatmaps showing the linkage disequilibrium (R^2 values) for SNPs in the G6PD region for populations in Congo, Senegal, Central African Republic, Nigeria, and ancient Cameroon. Red intensity indicates the strength of LD between SNP pairs, with an accompanying African map with corresponding countries highlighted, generated from Biorender.

No malaria and fava bean consumption

In the analysis of linkage disequilibrium (LD) within the G6PD region across present-day Algeria, Morocco, Egypt, and Tunisia, which are geographically located in North Africa, the images reveal a particularly strong LD between SNP pairs rs112776960 and rs73573464, as well as rs11429018 and rs5945100, indicated R^2 values higher than 0.5. Higher R^2 values, colored with darker red, indicate a strong LD, or that the alleles at these two SNP locations are associated with each other more than expected by chance.

This consistency suggests there is a non-random dissociation of these SNP pairs and a potential historical selection. These countries are categorized as having no malaria prevalence but traditionally high fava bean consumption and the patterns may indicate selective pressure imposed by favism. In the absence of malaria, favism shows deleterious effects which could lead to lower fitness of G6PD deficiency alleles. This could manifest as increased LD between SNPs within the G6PD region. G6PD deficiency with Canton (rs72554665) variant has a high risk of hemolytic anemia when fava beans are consumed [12]. Individuals with certain G6PD allele combinations that reduce the risk of favism are more likely to have higher reproductive success. The consistency of the linkage disequilibrium pattern across multiple North African countries suggests that traditional fava bean consumption may have played a role in the G6PD locus.

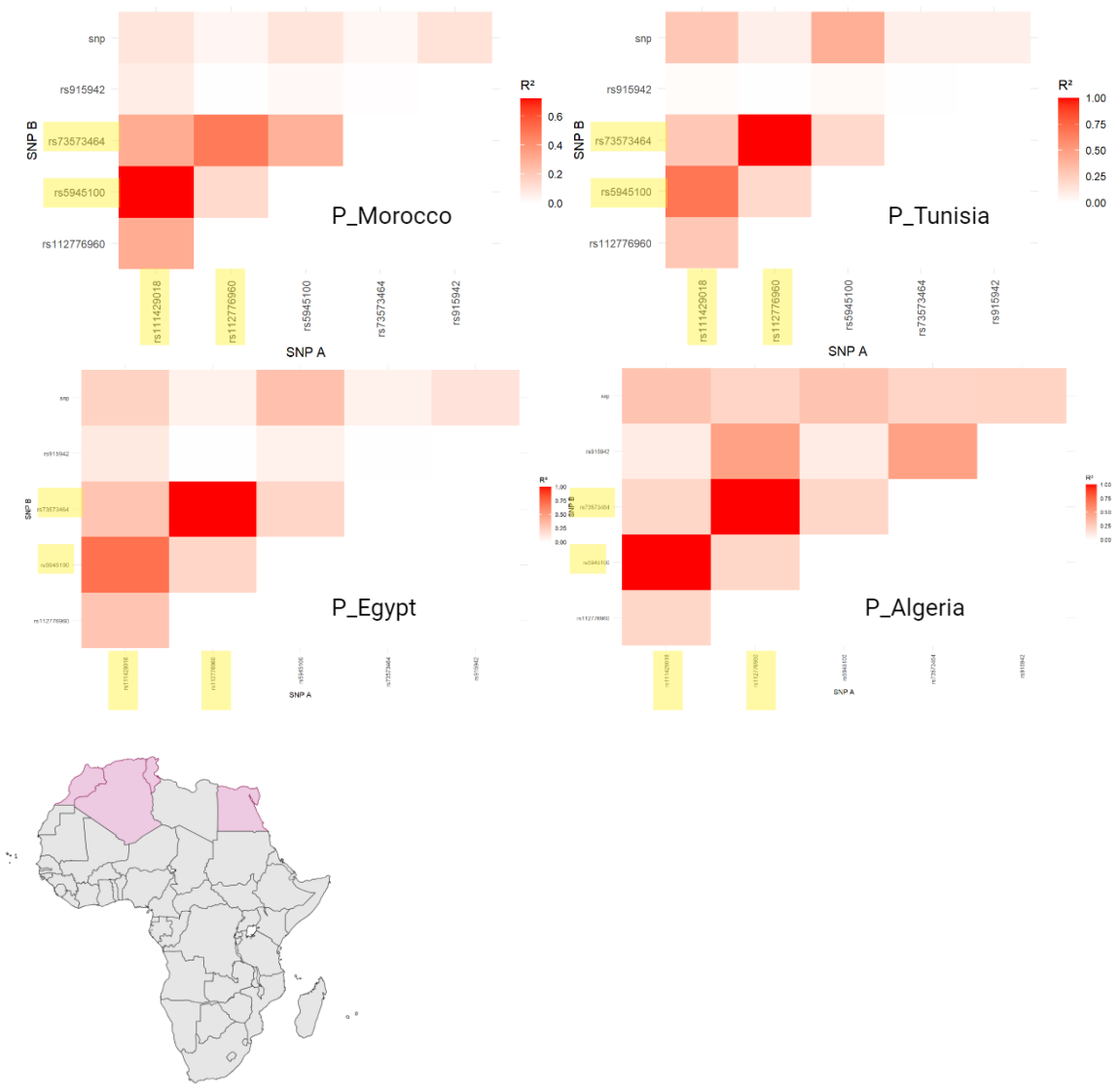


Figure 5c. Heatmaps showing the linkage disequilibrium (R^2 values) for SNPs in the G6PD region for populations in Morocco, Tunisia, Egypt, and Algeria. Red intensity indicates the strength of LD between

SNP pairs, with an accompanying African map with corresponding countries highlighted, generated from Biorender.

No malaria and no fava bean consumption

The three heatmaps display LD in present-day South Africa, Namibia, and ancient South African samples from 400 BCE. These countries are in the category of no malaria and no fava bean consumption. Present South Africa and Namibia show high R^2 values, particularly between rs5945100 and rs111429018 which suggests gene flow or shared selective pressure as the two countries are adjacent. In contrast, the ancient South African samples exhibit a drastically different LD pattern from present South Africa. As these samples are not under the known evolutionary pressures of malaria or favism, there might have been admixture events that have occurred since then.

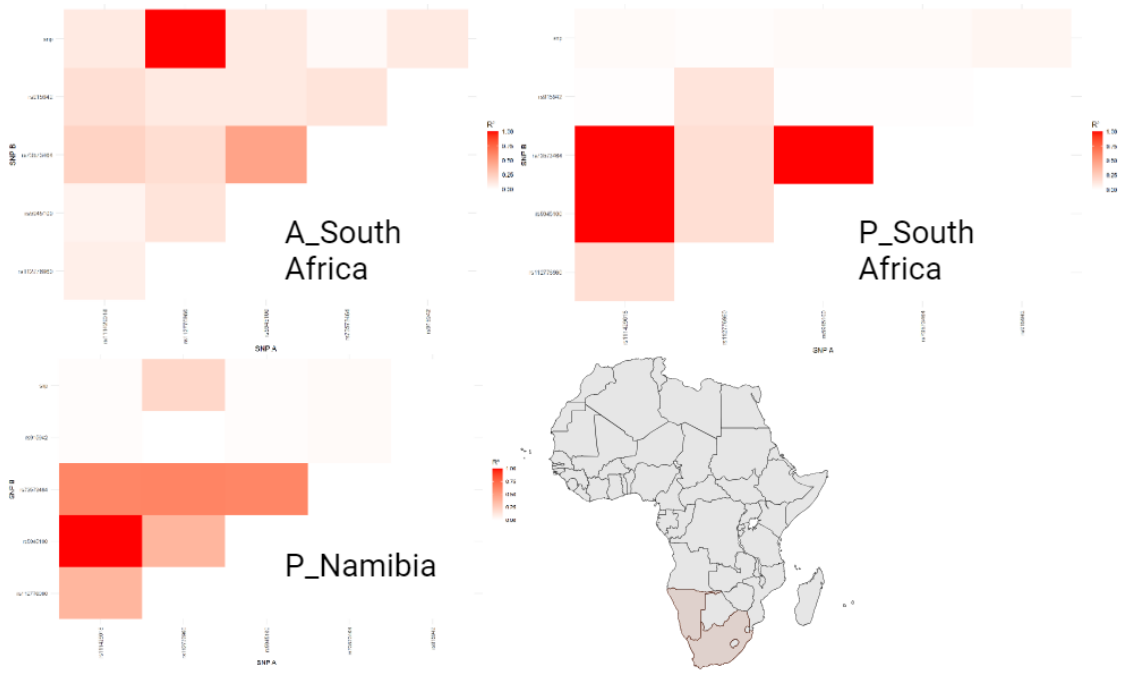


Figure 5d. Heatmaps showing the linkage disequilibrium (R^2 values) for SNPs in the G6PD region for populations in South Africa and Namibia, and ancient South African samples. Red intensity indicates the strength of LD between SNP pairs, with an accompanying African map with corresponding countries highlighted, generated from Biorender.

Conclusion

The senior thesis has interpreted the interplay between genetics and environment through the lens of one of the most common genetic diseases, G6PD deficiency, in various African populations. The analyses of linkage disequilibrium, principal component plots, and genetic distances have provided evidence for the historical impact of malaria and fava bean consumption on the G6PD region. The similar linkage disequilibrium patterns observed in West African countries highlight the influence of malaria as a selective pressure while the distinct genetic signatures in North African populations point more towards the dietary habits of traditional fava bean consumption in shaping the genetic variation. In the future, further statistical analyses and modeling such as could be used to infer the strength and direction of G6PD deficiency alleles.

Reference:

1. Antwi-Baffour, S., Adjei, J. K., Forson, P. O., Akakpo, S., Kyeremeh, R., & Seidu, M. A. (2019). Comorbidity of Glucose-6-Phosphate Dehydrogenase Deficiency and Sickle Cell Disease Exert Significant Effect on RBC Indices. *Anemia, 2019*, Article 3179173. https://doi.org/10.1155/2019/3179173

2. Bancone, G., & Chu, C. S. (2021). G6PD Variants and Haemolytic Sensitivity to Primaquine and Other Drugs. *Frontiers in Pharmacology, 12*, Article 638885. https://doi.org/10.3389/fphar.2021.638885

3. Peters, A. L., & Van Noorden, C. J. (2009). Glucose-6-phosphate dehydrogenase deficiency and malaria: cytochemical detection of heterozygous G6PD deficiency in women. *Journal of Histochemistry & Cytochemistry, 57*(11), 1003-1011. https://doi.org/10.1369/jhc.2009.953828

4. Harcke, S. J., Rizzolo, D., & Harcke, H. T. (2019). G6PD deficiency: An update. *Journal of the American Academy of Physician Assistants, 32*(11), 21-26. https://doi.org/10.1097/01.JAA.0000586304.65429.a7

5. Sirugo, G., Predazzi, I. M., Bartlett, J., Tacconelli, A., Walther, M., & Williams, S. M. (2014). G6PD A- deficiency and severe malaria in The Gambia: heterozygote advantage and possible homozygote disadvantage. *The American Journal of Tropical Medicine and Hygiene, 90*(5), 856-859. https://doi.org/10.4269/ajtmh.13-0622

6. Guindo, A., Fairhurst, R. M., Doumbo, O. K., Wellems, T. E., & Diallo, D. A. (2007). X-linked G6PD deficiency protects hemizygous males but not heterozygous females against severe malaria. *PLoS Medicine, 4*(3), e66. https://doi.org/10.1371/journal.pmed.0040066

7. World Health Organization. (2017). *A framework for malaria elimination.* https://www.who.int/malaria/publications/atoz/9789241511988/en/

8. Belsey, M. A. (1973). The epidemiology of favism. *Bulletin of the World Health Organization, 48*(1), 1-13.

9. Beretta, A., Manuelli, M., & Cena, H. (2023). Favism: Clinical Features at Different Ages. *Nutrients, 15*(2), 343. https://doi.org/10.3390/nu15020343

10. Tishkoff, S. A., Varkonyi, R., Cahinhinan, N., Abbes, S., Argyropoulos, G., Destro-Bisol, G., ... & Clark, A. G. (2001). Haplotype diversity and linkage disequilibrium at human G6PD: recent origin of alleles that confer malarial resistance. *Science, 293*(5529), 455-462. https://doi.org/10.1126/science.1057725

11. *African Archaeological Review*, *39*, 283–302. https://doi.org/10.1007/s10437-022-09483-9

12. McDonagh, E. M., Thorn, C. F., Bautista, J. M., Youngster, I., Altman, R. B., & Klein, T. E. (2012). PharmGKB summary: very important pharmacogene information for G6PD.

Pharmacogenetics and Genomics, *22*(3), 219-228. https://doi.org/10.1097/FPC.0b013e32834eb313

13. Murray, T., Beaty, T. H., Mathias, R. A., Rafaels, N., Grant, A. V., Faruque, M. U., Watson, H. R., Ruczinski, I., Dunston, G. M., & Barnes, K. C. (2010). African and non-African admixture components in African Americans and an African Caribbean population. *Genetic Epidemiology*, *34*(6), 561-568. https://doi.org/10.1002/gepi.20512