

https://dspace.org/

Correlations between self-reported ethnicity and language with genetic clustering in Kenya

Wolberg, Yonatan Ariel

2024
https://hdl.handle.net/10539/37710

Downloaded from a DSpace repository. https://dspace.org/

CORRELATIONS BETWEEN SELF-REPORTED ETHNICITY AND LANGUAGE WITH GENETIC CLUSTERING IN KENYA

Yonatan Ariel Wolberg

A research report (in the format of a “submissible” paper) submitted to the Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, in partial fulfilment of the requirements for the degree of Master of Science in Medicine (Genomic Medicine)

Johannesburg, 2022

NATIONAL HEALTH LABORATORY SERVICE
School of Pathology, University of the Witwatersrand
DIVISION OF HUMAN GENETICS
UNIVERSITY OF THE WITWATERSRAND JOHANNESBURG
Hospital Street, Johannesburg, 2001 | PO Box 1038, Johannesburg, 2000
[T] +27 11 489 9223 | [M] +27 78 080 8841 | [F] +27 11 489 9226 | [E] human.genetics@nhls.ac.za

Attention: Examiner

Dear Examiner,

Re: Research report submitted to the Faculty of Health Sciences by Mr Yonatan Ariel Wolberg in partial fulfilment of the requirements for the degree of Master of Science in Medicine (MSc Med) (Genomic Medicine)

Thank you for agreeing to examine a research report entitled: "**Correlations between Self-Reported Ethnicity and Language with Genetic Clustering in Kenya**".

The candidate, **Mr Yonatan Ariel Wolberg**, is submitting his research report in the format of a "submissible" paper. The research report contributes 30% towards the final mark for the degree. The candidate and his supervisors have chosen to submit the paper to the San Francisco-based journal 'PLOS Genetics', part of the Public Library of Science (PLOS). The submissible paper attached is therefore written in accordance with the author guidelines stipulated by the journal. The figures and tables are included in the manuscript for ease of marking, rather than being submitted separately as is required.

Attached in the appendix of this document, is the approved research protocol which provides an extended literature review, giving further context to the study. The protocol was assessed and approved by the Faculty of Health Sciences and is therefore not for examination.

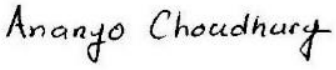
Pleases refer to the table of contents for a complete list of the items included in this research report.

Yours sincerely,

Student: Yonatan Ariel Wolberg



Supervisor 1: Prof. Ananyo Choudhury



Supervisor 2: Dr Dhriti Sengupta

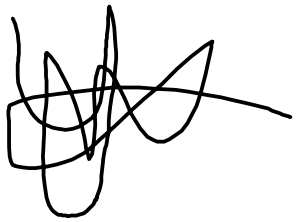


This report is intended solely to record the observations and/or opinion of the writer. It does not constitute a medico-legal report

Chairperson: Prof Eric Buch CEO: Dr Karmani Chetty
Physical Address: 1 Modderfontein Road, Sandringham, Johannesburg, South Africa Postal Address: Private Bag X8, Sandringham, 2131, South Africa
Tel: +27 (0) 11 386 6000/ 0860 00 NHLS(6457) www.nhls.ac.za
Practice number: 5200296

Declaration

I, **Yonatan Ariel Wolberg**, declare that this research (in the format of a “submissible” paper) is my own, unaided work. It is being submitted for the Degree of Master of Science in Medicine (MSc Med) at the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination at any other University.



(signature of candidate)

21st day of December in 2022 in Johannesburg

Contribution of the candidate to the paper

Declaration: Student's contribution to article(s) and agreement of co-author(s)

I, **Yonatan Ariel Wolberg**, student number: **183 46 47**, declare that this Research Report is my own work and that I have contributed significantly towards research findings presented in the paper intended for publication below.

Signature of Student: 

Date: 21/12/2022

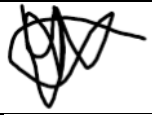
Name of Primary Supervisor: Professor nanyo Choudhury

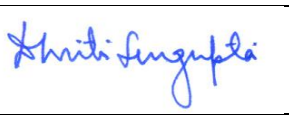
Signature of Primary Supervisor: Ananyo Choudhury

Date: 21/12/2022

Agreement by co-authors: By signing this declaration, the co-authors listed below agree to the use of the article(s) by the student as part of his Research Report.

Article Title: Correlations between Self-Reported Ethnicity and Language with Genetic Clustering in Kenya

Authors	Name	Signature	Date
1st author	Yonatan Ariel Wolberg		21/12/22
2nd author	Ananyo Choudhury	Ananyo Choudhury	21/12/22

3rd author	Dhriti Sengupta		21/12/22

This research report is dedicated to my parents Robert and Michelle Wolberg

Abstract

Kenya is a highly diverse country, where a combination of recent local migrations and admixture have contributed to a complex population structure. This structure creates a dilemma when trying to assess the allele frequencies of disease-associated variants within the country as different groups will show different frequencies. Additionally, ethnic groups in genetic studies are often defined on the basis of self-reported identity but certain individuals may align genetically to another ethnic group. It is necessary to properly characterize Kenyan diversity for population level risk estimation and the implementation of public health approaches. This study aimed to determine how self-reported ethnicity correlates to genetic clustering in a Kenyan cohort. The effect of the discordance between the two on the frequencies of key malaria- and trypanosomiasis-associated variants was then determined. This study leveraged Kenya AWI-Gen dataset, comprising 1,703 individuals (of the Kikuyu, Kamba, Luhya, Luo, Kisii and Somali ethnic groups) recruited in Nairobi. Combining a bootstrap approach for allele frequency estimation and centroid-based filtering, this study was able to show that small discordances are able to significantly impact allele frequencies of disease-associated variants. More robust approaches to compare genetic- and ethnicity-based clustering might reveal further differences. Overall, the results indicate that while self-reported identity can provide reasonably reliable categorization for the Kenyan dataset, inclusion of additional variables, such as language, geographic origin, and both parental and grandparental identity, might be necessary for more accurate estimates.

Acknowledgments

I would like to thank all Kenyan participants of the H3Africa AWI-Gen study, whose genotype and ethnic identity data was used in this project.

I would like to thank my co-supervisors Dr Dhriti Sengupta and Prof Ananyo Choudhury for their assistance, guidance, teaching and dedication to me during this project and for the time they gave for meetings and helping me with troubleshooting.

I would like to thank Zane Lombard, Khuthala Mnika, Robin Kerr and Fiona Baine-Savanhu for organizing this degree and assisting me through both the coursework component and the research component.

I would like to thank Prof Michelle Ramsay for granting me access to the H3Africa AWI-Gen dataset.

I would like to thank Prof Scott Hazelhurst who granted me access to the Wits Cream cluster on which much of the coding took place.

I would like to thank the National Research Foundation (NRF) for funding my studies via the NRF Postgraduate Scholarship program.

I would like to thank my family and friends for their support, encouragement and advice during this project.

I would like to thank my friends and classmates Armandi, Tebalelo, Bontle, Minenhle, Kentie, Yolanda, Samantha, Caitlyn and Ross

Table of Contents

Declaration 3

Contribution of the candidate to the paper 4

Acknowledgments 8

Research report in the format of a “submissible” paper 15

Correlations between self-reported ethnicity and language with genetic clustering in Kenya 16

Affiliations 16

Abstract 17

Author Summary 18

Introduction 19

Genetic diversity in East African populations 19

Malaria- and trypanosomiasis-associated variants 20

The H3Africa AWI-Gen Collaborative Centre 21

Methodology 22

Study participants and demographics 22

Allele frequency estimation 22

Bootstrapped allele frequency estimation 23

Genetic clusters based on principal component analysis 23

Genetic data-based filtering of ethnolinguistic groups 24

Allele Frequency Estimation 24

Statistical Analysis 24

Ethics statement 25

Results 26

Variation of frequencies within the six ethnolinguistic groups 26

Comparison of MAFs in Luhya from AWI-Gen and the 1000 Genomes Project datasets 27

Impact of PCA-based filtering on MAFs 29

Significance of differences in bootstrapped MAF pre- and post-removal of excluded individuals 32

Discussion 35

References 39

Appendix 42

Supplementary Table 1. Details of SNPs used in this study 42

Supplementary Table 2. Comparison of absolute MAFs between genotype and imputed datasets, pre- and post-exclusion of outlier individuals....44
Supplementary Table 3. Comparision of pre- and post-removal of excluded individuals 46
Supplementary Note 1 53
Appendix A 54
Approved Research Proposal 54
1. Introduction....62
1.1. Genetic Diversity in East African Populations....62
1.1. Malaria- and Trypanosomiasis-Associated Variants 64
1.1.1. Malaria-Associated Variants 64
1.1.2. Trypanosomiasis-Associated Variants 65
1.1. The H3Africa AWI-Gen Collaborative Centre 66
1.1.1. East African ethnolinguistic groups 67
2. Aims and Objectives 68
3. Methodology 69
3.1. Genetic clusters based on Principal Component Analysis 69
3.2. Genetic data-based filtering of ethnolinguistic groups 71
1.1. Allele Frequency Estimation 72
1.2. Statistical Analysis 72
1.3. Problems and challenges 73
1.4. Ethical Considerations 73
1.5. Funding 73
1. Proposed Project Timeline 74
References 74
4. Appendix....77
4.1. Supplementary Table 1. Malaria-associated loci and their phenotypes 77
4.2. Supplementary Table 2. Trypanosomiasis-associated variants to be utilized in allele frequency determination 78
1.1. Ethics Clearance Certificate 80
Appendix D: Ethics Clearance Certificate....82
Appendix C: Turn-it-in Report....84
Appendix D: PLOS Genetics: Author Guidelines....90

List of Figures and Tables

Figure 1. Six selected SNPs that show clear differences between the Kenyan ethnic groups.... 27

Figure 2. Scatterplot of AWI-Gen versus 1KGP Luhya MAFs reveals a strong positive correlation..... 28

Figure 3. PCA showing the centroids for each ethnic group. 30

Table 1. East African ethnolinguistic groups in the AWI-Gen dataset. 22

Table 2. Number of excluded individuals from each group and the proportions of each ethnic group that they represent. The proportions refer to the percentage of individuals excluded from each of the ethnic groups..... 31

Table 3. Statistically significant P-values and associated bootstrapped MAFs for each SNP for each ethnic group, pre- and post-exclusion of outliers. Only statically significant P-values are shown and these values were determined using an unpaired t-test. 32

List of Abbreviations

AGVP	African Genome Variation Project
APOL1	Apolipoprotein 1
AWI-Gen	Africa Wits-INDEPTH Partnership for Genomic Studies
CDC	Centre for Disease Control and Prevention
CMD	Cardiometabolic Disease
DPHRU	Developmental Pathways to Health Research Unit
G6PD	Glucose-6-Phosphate Dehydrogenase
GWAS	Genome-Wide Association Studies
H3Africa	Human Heredity and Health in Africa
Hb	Haemoglobin

HbS	Sickle Cell Allele (Hemoglobin Sickle)
HPR	Haptoglobin-Related Protein
HDSS	Health and Demographic Surveillance System
HREC	Human Research Ethics Committee
MAF	Minor Allele Frequency
NeuroGAP-Psychosis	Neuropsychiatric Genetics of African Populations-Psychosis
NRF	National Research Foundation
NUHDSS	Nairobi Urban Health and Demographic Surveillance System
PC	Principal Component
PCA	Principal Component Analysis
PMI	United States' President's Malaria Initiative

POPIA	Protection of Personal Information Act
SCD	Sickle Cell Disease
SNP	Single Nucleotide Polymorphism
SSA	Sub-Saharan Africa
TLF1	Trypanosome lytic factor-1
UMAP	Uniform Manifold Approximation and Projection
WGS	Whole Genome Sequencing
WHO	World Health Organization

Research report in the format of a “submissible” paper

Correlations between self-reported ethnicity and language with genetic clustering in Kenya

Yonatan Ariel Wolberg[1], Dhriti Sengupta[2], Ananyo Choudhury[2]

Affiliations

[1] Department of Human Genetics, University of the Witswatersrand via the National Health Laboratory Service, Johannesburg, Gauteng, South Africa

[2] Sydney Brenner Institute of Molecular Bioscience, Johannesburg, Gauteng, South Africa

Abstract

Kenya is a highly diverse country and the combination of recent local migrations and admixture have contributed to a complex population structure. This structure could lead to major differences in allele frequencies even among individuals living in the same region. A proper characterization of diversity is therefore key to population level risk estimation and the implementation of public health approaches. Classifying individuals by self-reported ethnicity is a commonly employed approach for grouping African populations. However, for geographic regions inhabited by multiple different ethnic groups conferring to distinct ancestries along with some amount of gene-flow between them, such classification might not always be optimal. Leveraging the H3Africa AWI-Gen dataset, this study aimed to assess how self-reported ethnicity correlates to genetic clustering in a Kenyan cohort and also estimate the impact of the discordance between the two on the frequencies of key malaria- and trypanosomiasis-associated variants. The frequencies of these variants were assessed using a bootstrap approach (10 iterations, each including 30 randomly selected individuals from each group). One of our key observations was the large scale differences in allele frequencies of these disease associated variants among the six ethnic groups studied. Following this, centroid-based filtering was used to identify potential outliers and the allele frequencies were re-estimated after excluding these outlier individuals. Unpaired t-tests were performed to assess the significance of differences in bootstrapped allele frequencies pre- and post-exclusion. Of the six groups, the Luo, showed the highest discordance between genetic and ethnicity based clustering that resulted significant differences in some allele frequencies of some of the SNPs. More robust approaches to compare genetic and ethnicity-based clustering might reveal additional differences. Overall, the results indicate that while self-reported identity can provide reasonably reliable categorization for the Kenyan dataset, inclusion of additional variables might be necessary for more accurate estimates.

Author Summary

Self-reported ethnicity is a widely used approach to group individuals for genetic studies. While this approach has been shown to provide reliable estimates in some populations, in environments shared by different ethnolinguistic groups, such as some of the urban centers in Africa this presents a problem should the individual have one or more parents or grandparents from different ethnic groups. Also, in cases individuals tend to identify themselves with the predominant ethnic group of the area. In either case, the inclusion of a considerable number of such individuals from an ethnic group in a study might impact with the accuracy of estimates for such ethnic groups. In Africa, the frequencies of disease-related genetic variants are known to differ between different ethnic groups, even those who belong to the same language family. The prevalence of these variants has the potential to inform risk stratification and public intervention. Thus it is important to determine accurate allele frequencies for each ethnic group within a country or province. This study demonstrated that a discordance between self-reported identity and the group to which an individual genetically aligns, in cases might result in significant differences in the frequency of these genetic variants. Therefore, especially for populations sampled from urban African centers it might be helpful to combine self-reported identity with other related variables such as language, geographic region, and parental and grandparental identity in order to accurately categorize them for genetic studies.

Introduction

Genetic diversity in East African populations

Kenya is a highly diverse country, with about 42 different ethnic groups distributed in a complex pattern across the country as a result of local migrations (1–3). A study by the African Genome Variation Project (AGVP), comprising the largest genomic dataset of Kenya, revealed considerable population structure within the country. Admixture analysis revealed that the Kikuyu, Kalenjin and Luhya populations had substantial gene flow between these groups (2). This population structure was confirmed by the Neuropsychiatric Genetics of African Populations-Psychosis (NeuroGAP-Psychosis) study, which showed that this structure had a strong correlation to self-reported language (4). This indicated that there were both cultural and geographical influences on genetic clustering (4). However, there were a number of cases where self-reported identity and defined genetics did not match, likely due to the assimilation of participants into the dominant culture of the region or the occurrence of parents or grandparents with different self-reported ethnicities. (4,5).

Such a population structure, as observed in Kenya, could result in the variable distribution of disease-associated alleles. As populations/groups in genetic studies are often defined on the basis of self-reported ethnicity or language, it is important to assess the level of discordance that might accompany grouping based on these criteria and also how this could impact the estimation of the frequency of disease-associated alleles. An intuitive approach to estimate the level of misclassification is to compare principal component analysis (PCA)-defined groupings and self-reported ethnicity-defined groupings. The comparison of allele frequencies from genetic –based grouping and other grouping approaches could then be used to assess the extent to which such misclassification may influence disease risk estimates.

Malaria- and trypanosomiasis-associated variants

Malaria and trypanosomiasis are two potentially fatal protozoan diseases that have played major roles in the shaping of African genomes (6). It is estimated that 70% of Kenya's population is at risk of contracting malaria and this high prevalence is accompanied by a number of genetic burdens that confer a protective effect against malaria, such as thalassemia, sickle cell disease (SCD), glucose-6-phosphate dehydrogenase (*G6PD*) deficiencies and ovalocytoses (7,8).

Over the years several loci have been identified, which either reduce the risk of contracting malaria or provide resistance to and protection against it (Supplementary Table 1). Notably, in a study from 1995 to 2008 involving 2 244 children with severe malaria in Kilifi County, Kenya, Ndila et al. (2018) identified polymorphisms in 15 loci associated with severe malaria: 1) *ABO*; 2) *ATP2B4*; 3) *ARL14*; 4) *CD40LG*; 5) *FREM3*; 6) *INPP4B*; 7) *G6PD*; 8) *HBA1* and *HBA2*; 9) *HBB*; 10) *IL10*; 11) *LPHN2*; 12) *LOC727982*; 13) *RPS6KL1*; 14) *CAND1*; and 15) *GNAS* (9). Additionally, Greene et al. (2009) found *G6PD* and the sickle cell allele (HbS) to be more prominent in Kenya than Papua New Guinea, marking *G6PD* and *HBB* out as a clear starting point for the determination of the allele frequency of malaria-associated variants in Kenya (10).

Trypanosomiasis is a spread by the tsetse fly in Africa which occur over 23% of Kenya, at densities of 1000 per square kilometer (11–13). East African trypanosomiasis is generally more acute than the West African type, with symptoms, such as high fever, skin rash, headache and muscle aches, appearing 1 to 3 weeks following the bite (14). In more severe cases, the enlargement of the spleen, kidney failure and problems of the heart, have also been observed (14). The major genes linked to trypanosomiasis are those that contribute to the trypanosome lytic factor-1 (TLF-1), namely: haemoglobin (Hb), apolipoprotein L1 (*APOL1*) and haptoglobin-related protein (*HPR*). This complex lyses Trypanosoma brucei via oxidative damage caused by its peroxidase activity (15,16).

In addition to providing an estimate of variations of disease-related alleles, research on how the incidences of these diseases correlate with these allele frequencies could provide insights into the dynamics of the distribution of both the diseases and populations over time.

The H3Africa AWI-Gen Collaborative Centre

The Human Heredity and Health in Africa (H3Africa) Africa Wits-INDEPTH partnership for Genomic studies (AWI-Gen) Collaborative Centre was created in 2012 and includes 12 000 individuals from South Africa, Ghana, Burkina Faso and Kenya (17). This dataset has enabled the successful characterization of the genetic substructure of South African Bantu speakers, as well as the effects of such substructure on allele frequency variation (5), so it was expected that the same might be true for Kenya's ethnic groups.

The current study focuses on more in-depth analysis of this dataset to identify the variation in malaria- and trypanosomiasis-associated variants in Kenya .

The aim of this study was to compare how self-reported ethnicity and language correlate to genetic clustering in this Kenyan cohort. The effect of discordances between the two on the minor allele frequencies (MAFs) of relevant malaria- and trypanosomiasis-associated variants was thereafter determined. The aforementioned list of malaria- and trypanosomiasis-associated SNPs were searched for in the AWI-Gen dataset and for those present, the MAFs were determined for each ethnic group. The major genetic groups within the AWI-Gen Nairobi participants were determined through principal component analysis. Following this, outlying individuals were identified using a centroid-based filtering of these PC coordinates, then counted and excluded. The MAFs of these SNPs were re-estimated for each group after removing outlying individuals. The MAFs pre- and post-removal of said individuals were compared to assess if such differences were significant.

Methodology

Study participants and demographics

This study used a total of 1 703 individuals recruited at the Nairobi Urban Health and Demographic Surveillance System (NUHDSS) study centre of the AWI-Gen study. The 6 ethnic groups they comprise represent 69% of Kenya's population (Table 1) (18). Only groups with over 45 individuals were included in this study.

Table 1. East African ethnolinguistic groups in the AWI-Gen dataset.

Ethnolinguistic group	Number
Kikuyu	646
Kamba	328
Luhya	277
Luo	357
Kisii	47
Somali	48

Allele frequency estimation

A list of key variants associated to Malaria and Trypanosomiasis was generated using in-depth literature mining (Supplementary Table 1). Next, using a custom script, it was assessed which SNPs from Supplementary Table 1 were present within each of the 6 ethnic groups in the genotype (1.7 M SNPs) and imputed data (14M SNPs). The minor allele frequency (MAF) of these SNPs was estimated using PLINK, an open-source whole genome association analysis toolset, which is suitable for the analysis and manipulation of large datasets (19).

Bootstrapped allele frequency estimation

The different sample sizes of each ethnic group mean that the accuracy of the MAF estimations would vary between the groups. In an attempt to address this, the sample sizes were set to 30 individuals per group. Using a custom script, 10 random subsamples of 30 participants were generated for each ethnic group and the MAF of each SNP was determined for each subsample using PLINK 1.9. Based on the 10 sample sets, the mean frequencies and the standard deviations for each group were calculated using a custom script. The standard deviations for each set of subsamples were divided by the square root of the sample size, to give a standard error for each bootstrapped MAF.

Comparison of Luhya in AWI-Gen versus 1000 Genomes Project

To compare allele frequencies of the 20 SNPs in the Luhya participants from our cohort to that from 1000 Genome Project, allele frequencies of each SNP in the 1000 Genomes project LWK population was obtained from the 1000 Genomes project Phase 3 database hosted at the Wits computational cluster. A scatter plot was used to compare the frequency of 20 SNPs in the two datasets.

Genetic clusters based on principal component analysis

Principal component analysis, based on H3Africa AWI-Gen dataset, was used to assess the population structure of the 6 ethnic groups. Prior to performing the PCA, the dataset was pruned, such that loci with high levels of pairwise linkage-disequilibrium (LD) were removed. The window size was set at 50 SNPs, 5 SNPs to shift the window at each step and the r^2 threshold at 0.5. PCAs were performed using PLINK (ref) and were visualized using custom R scripts and GENESIS library.

Genetic data-based filtering of ethnolinguistic groups

One of the main aims of this project is to determine how misclassification due to discordance between self-reported identity and genetic clustering impact on the allele frequency of malaria- and trypanosomiasis-associated variants.

In order to assess misclassifications due to discrepancies between self-reported identity and genetic clustering, the distribution of individuals in each ethnic group within the PC space was evaluated. Centroids for each group were calculated using the distribution of the first 2 PCs via a custom R script. There were two criteria for the exclusion of an individual from their self-reported ethnic group: 1) If an individual was closer, in Euclidean distance, to the centroid of any group other than that of their self-reported affiliation; and 2) If an individual was found to differ by more than 5 standard deviations from the group mean of any of the two PCs. The number of these excluded individuals from each ethnic group served as a proxy for determining how closely self-reported ethnicity correlated to genetic clustering.

Allele Frequency Estimation

Following the genetic data-based filtering of the groups, i.e. the removal of misclassified individuals, the MAFs were re-estimated. Similar to the original MAF estimation, 10 random subsamples of 30 participants were generated for each ethnic group and the MAF of each SNP was determined for each subsample. These MAFs were averaged for each set of subsamples to give bootstrapped MAFs, with the standard deviation and standard errors determined as above.

Statistical Analysis

Power estimation calculations were performed using the PAWR package in R. The estimates show a sample size 50 is sufficiently powered to test for allele frequency differences for the majority of disease associated variants (Please see Supplementary Note 1 for details). Unfortunately, the Kisii and Somali groups had less than 50 members, albeit 47 and 48

197 respectively, necessitating the need for a smaller sample size of 30 individuals, where the
198 number of random samplings were 10 for each group.

199 The statistical significance (*P*-values) of the differences in the MAFs of malaria- and
200 trypanosomiasis-associated SNPs pre- and post-exclusion of outlying individuals was assessed
201 using unpaired t-tests.

202 ## Ethics statement

203 This project is a retrospective study using summary-level data and so there are no major ethical
204 considerations. In accordance with the Protection of Personal Information Act (POPIA), the data
205 used in this study contains no identifying information, such as names, dates of birth, addresses,
206 identification, medical or hospital numbers. A provisional ethics clearance certificate (Clearance
207 Number: M220294) has been obtained, with Human Research Ethics Committee (Medical)
208 (HREC) approval requiring evidence of protocol approval by the Division of Human Genetics in
209 the Faculty of Health Science.

210

Results

Variation of frequencies within the six ethnolinguistic groups

Six of the thirty Malaria and Trypanosomiasis associated SNPs: 1) rs12075 and 2) rs863004 (DARC) (chromosome 1); 3) rs1800890 (IL19) (chromosome 1); 4) rs73885319 (APOL1) (chromosome 22); 5) rs77389579 (INPP4B) (chromosome 4); and 6) rs8386 (GNAS) (chromosome 20) (Table 2)), were included the QC-ed genotype-array dataset (containing 1.7 million SNPs) from the AWI-Gen study. 14 additional SNPs were included from the imputed (on the AGR panel at the Sanger Imputation Service) dataset. As the absolute MAFs of the 6 alleles present in both datasets were the same, we have used the imputed dataset for the rest of the study (Supplementary Table 2).

The comparison of the MAF of each SNP in the 6 ethnolinguistic groups, detected 6 SNPs (rs12307123, rs1541255, rs1566830, rs1800890, rs334 and rs72933304) to show considerable variation (Figure 1 and Supplementary Table 3). For example, rs12307123, in the Kisii showed notably higher a MAF compared to all other ethnic groups. Similarly, rs1800890 and rs72933304 showed much higher MAF in Somali in compared to all other groups. The sickle cell associated SNP rs334, also showed extreme variation, ranging from complete/near absence in Somali, Kikuyu to a MAF of 0.15 in the Luo. Interestingly, in additional to difference between language family we also observed major difference within groups affiliated to same family. For instance, the MAF of rs334 in Luhya was more than 10 fold in comparison to that in Kamba. Similarly, rs1541255 also showed major difference in MAF between Kikuyu and Kamba, the two of the Niger-Congo speaking groups.

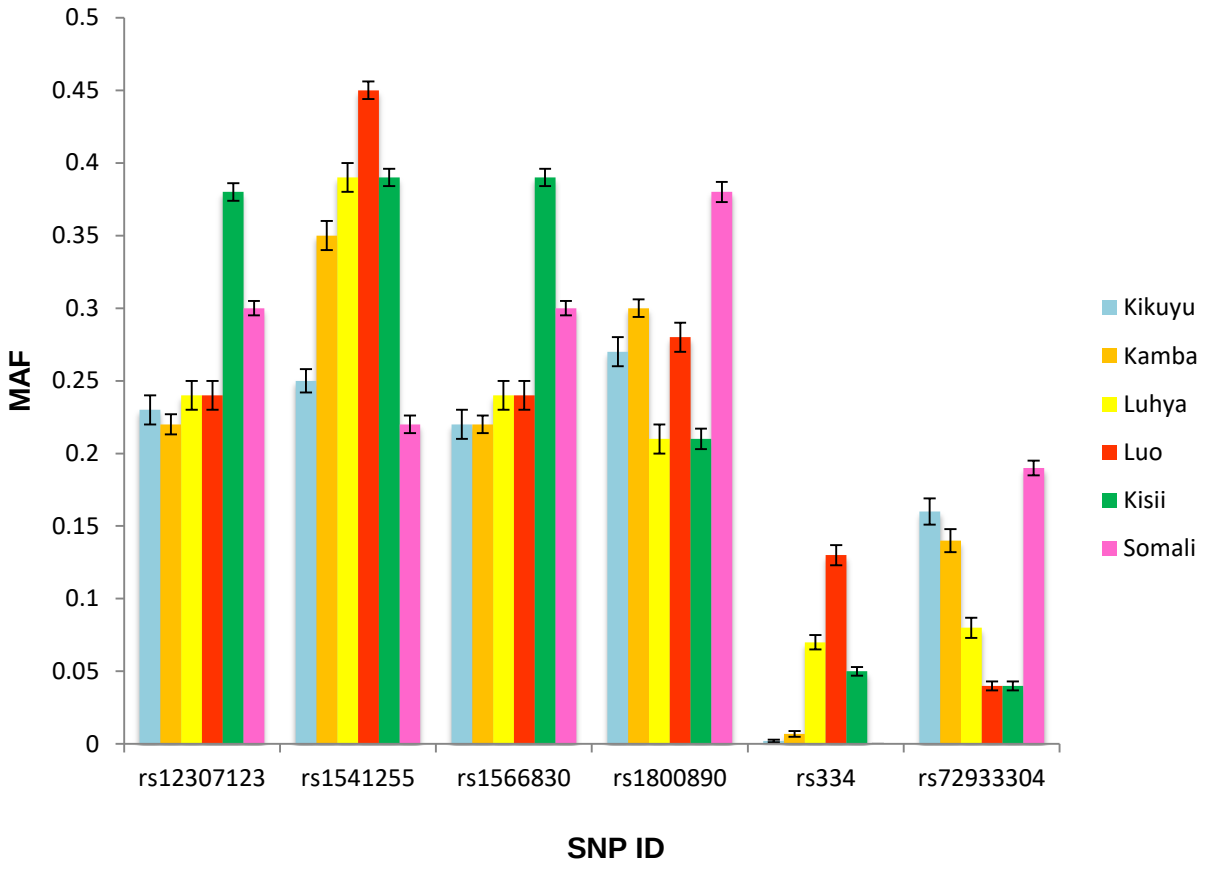


Figure 1. Six selected SNPs that show clear differences between the Kenyan ethnic groups. The standard errors of each bootstrapped MAF for each SNP for each ethnic group are represented by the black bars.

Comparison of MAFs in Luhya from AWI-Gen and the 1000 Genomes Project datasets

The Luhya was the only group in this study that was also in the 1000 Genomes Project (1KGP) (ref). The Luhya from 1KGP were sampled from the Luhya homeland, while the Luhya from AWI-Gen were sampled from Nairobi. We therefore compared the frequency of the 20 variants in the two Luhya dataset to assess the extent of difference among the two groups. The scatter plot in Figure 2 shows a strong positive correlation between the frequency of the SNPs in these two datasets (Figure 2). However, there were instances where allele frequencies differed

considerably from each other. For example, the SNP rs334 displayed an allele frequency being 0.05861 in the AWI-Gen, while the 1KGP frequency was almost double this at 0.101. The rs72933304 SNP showed similar difference with the AWI-Gen frequency of 0.07692 being over twice that of the 1KGP frequency of 0.035. Other SNPs showing differences in the two Luhya datasets include rs11865131 (AWI-Gen MAF=0.1923, 1KGP MAF=0.141), rs12307123 (AWI-Gen MAF= 0.2289, 1KGP MAF= 0.278). rs1371478 and rs1566830.

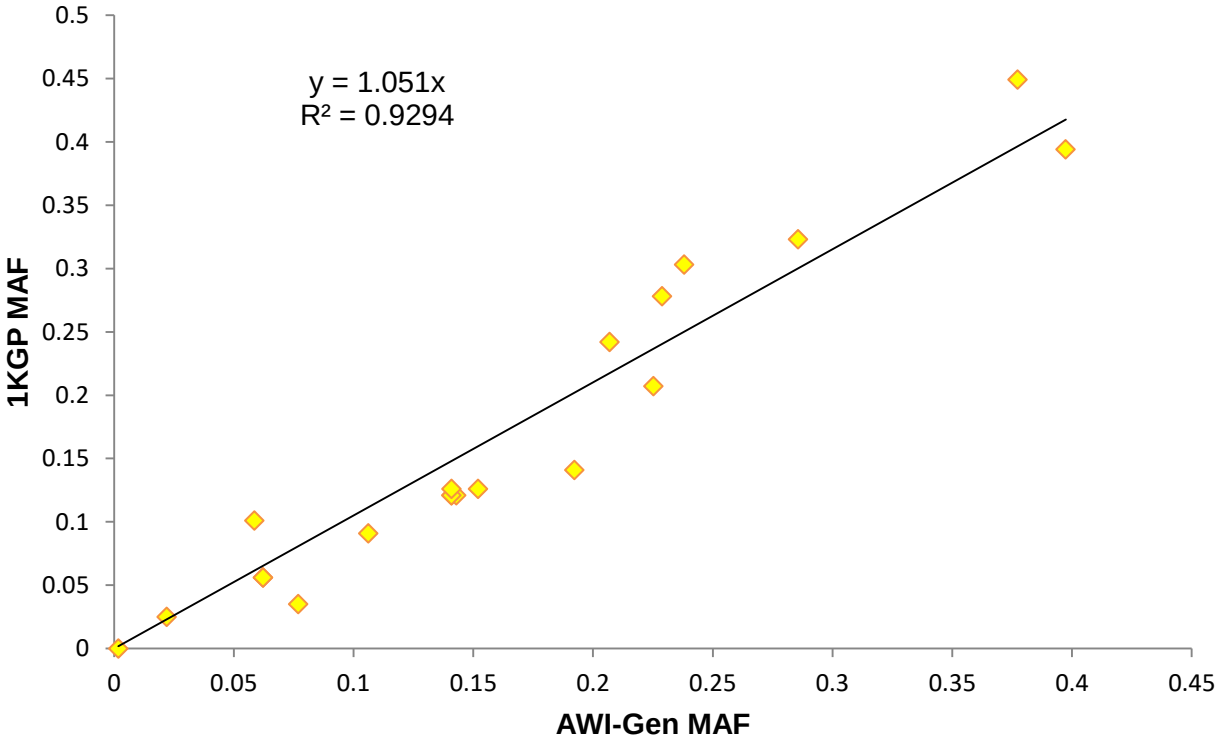


Figure 2. Scatterplot of AWI-Gen versus 1KGP Luhya MAFs reveals a strong positive correlation. The trendline equation and R^2 value are displayed above the line. The AWI-Gen MAF values are absolute values determined from the imputed data using PLINK. The 1KGP values are available on the 1KGP browser under the code LWK.

Impact of PCA-based filtering on MAFs

Next, we sought to identify the individuals showing discordance between self-reported identity and Principal Component Analysis (PCA) based groups for each group. (Figure 3a).

For this, we performed a centroid-based filtering to remove any individuals who were closer to centroids of other groups compared to their own group centroid. The centroids were calculated based on the first 2 PCs only. Due to their genetic proximity, the centroids for the Kikuyu and Kamba groups were very close, while those of the Luhya and Luo overlapped (Figure 3a). Thus, when analyzing the Kamba, the Kikuyu centroid was excluded and vice versa. Similarly, for the Luhya, the Luo centroid was excluded and vice versa.

In the Kamba, 6 individuals were excluded due to their being closer to another group's centroid. For the Kikuyu there were 18 excluded individuals. For the Luhya, this number was 13 individuals, while for the Luo, it was 9. For the Kisii, there was a single excluded individual and for the Somali, there were 2 such individuals. Figure 3b shows the PCA plot generated after removing these outlier individuals.

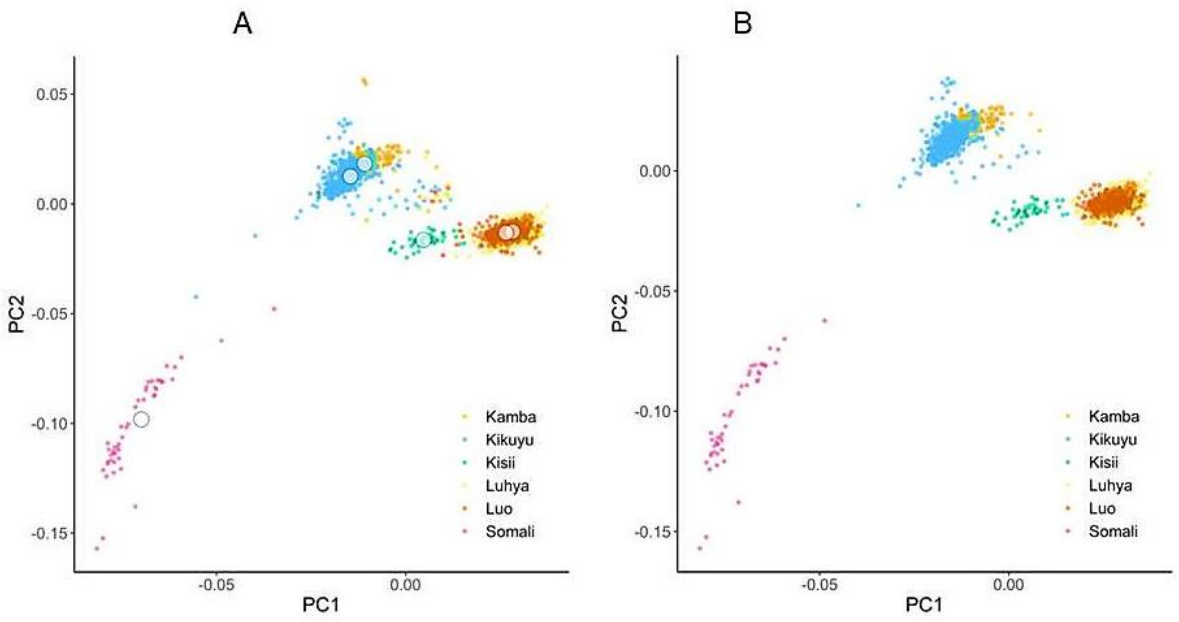


Figure 3. PCA showing the centroids for each ethnic group. (A) PCA with centroids prior to filtering out excluded individuals. (B) PCA without centroids, following filtering. Each dot represents a different individual and each ethnic group is represented by a different colour. The centroid for each group is represented by an empty circle. These figures were generated using R.

The second criterion for the exclusion of individuals was the deviation of said individuals from the group's mean for the first two PCs by more than 5 standard deviations (SD) (Table 2). In the Kamba group, 4 individuals were removed as they met the exclusion criteria of >5 SD for PC1 or PC2. In the Kikuyu, there were 5 such individuals. In the Luo there were 3 such individuals and for the Luhya, there were 2. The Kisii and Somali groups had no individuals that fell outside of 5 SD, likely a result of their smaller sample size.

We also determined whether additional individuals might get excluded if the SD threshold is increased to 3 and 4 standard deviations While For, Kikuyu and Kamba additional individuals might be removed (29 and 11, respectively) with the use of a more stringent exclusion threshold of 3 SD, the number of individuals excluded remained same for all other groups (Table 2).. Although, we selected the threshold cutoff of 5 SD for the downstream analysis, a

more stringent SD based filtering as well as inclusion of additional PCs for centroid based filtering could further increase the number of outliers and impact of downstream analysis.

Based on these exclusion criteria, it was found that the Kamba, Kikuyu, Kisii and Luo had similar proportions of excluded individuals, between 2 and 3%,(Table 2). In contrast to this, both the Luhya and Somali groups had proportions of over 4 %, with the Luhya being the highest. Based on these proportions, it was expected that the Luhya and Somali allele frequencies would be most strongly affected following this exclusion.

Table 2. Number of excluded individuals from each group and the proportions of each ethnic group that they represent. The proportions refer to the percentage of individuals excluded from each of the ethnic groups.

Ethnic group	Number of standard deviations					
	3		4		5	
	Number of excluded individuals	Proportion of ethnic group (%)	Number of excluded individuals	Proportion of ethnic group (%)	Number of excluded individuals	Proportion of ethnic group (%)
Kikuyu	29	4.49	19	2.94	18	2.79
Kamba	11	3.35	10	3.05	9	2.74
Luo	9	2.52	9	2.52	9	2.52
Luhya	13	4.69	13	4.69	13	4.69
Kisii	1	2.13	1	2.13	1	2.13
Somali	2	4.17	2	4.17	2	4.17

Significance of differences in bootstrapped MAF pre- and post-removal of excluded individuals

The actual and bootstrapped MAFs were re-estimated using PLINK following the exclusion of these outliers. These bootstrapped MAFs were then compared to those prior to exclusion to determine if the differences between them were significant or not. This was achieved with an unpaired T test, with a significance level (α) set at 0.05. In the Kikuyu group, there were 8 SNPs that displayed such significant differences (Table 3 and Supplementary Table 3). For the Kamba, there were only 3 such SNPs. Surprisingly, the Luhya also only had 3 such SNPs. Unexpectedly, due to their percentage of excluded individuals being only 2.52%, the Luo group had 10 SNPs that demonstrated significant differences between pre- and post-exclusion MAFs. The Kisii group had 3 such SNPs. Finally, for the Somali, there were 8 SNPs showing significant differences between the pre- and post-exclusion MAFs. The SNPs that were showed major differences include rs11865131 (in the Kikuyu, Luhya and Luo); rs1371478 (in the Kamba, Luhya and Somali); rs3742785 (in the Kikuyu, Kamba and Somali); and rs77389579 (in the Kamba, Luhya, Luo and Somali).

Table 3. Statistically significant P-values and associated bootstrapped MAFs for each SNP for each ethnic group, pre- and post-exclusion of outliers. Only statically significant P-values are shown and these values were determined using an unpaired t-test.

	Ethnic Group		
SNP ID	**Kikuyu**		
	Pre-Removal	**Post-Removal**	**p-value**
rs11865131	0.14949	0.12616	0.0437
rs1541255	0.25343	0.27937	0.0268
rs3027011	0.125748	0.099321	0.052
rs3742785	0.40902	0.43281	0.0375
rs55872368	0.174433	0.122755	0.0007
rs75731597	0.095862	0.074027	0.0191

rs8386	0.145973	0.097591	0.0001
rs863004	0.142487	0.112699	0.0257
SNP ID		**Kamba**	
	Pre-Removal	**Post-Removal**	**p-value**
rs1371478	0.25896	0.32621	0.0001
rs3742785	0.43123	0.41275	0.0441
rs77389579	0.066438	0.049022	0.0145
SNP ID		**Luhya**	
	Pre-Removal	**Post-Removal**	**p-value**
rs11865131	0.19439	0.21723	0.05
rs1371478	0.29703	0.26239	0.0284
rs77389579	0.018626	0.01	0.0069
SNP ID		**Luo**	
	Pre-Removal	**Post-Removal**	**p-value**
22:36662034	0.048333	0.088342	0.0001
rs11865131	0.15333	0.18833	0.0036
rs3027011	0.154997	0.18333	0.0151
rs334	0.128344	0.096669	0.0009
rs55872368	0.154997	0.18167	0.0193
rs72933304	0.035	0.048332	0.0041
rs73885319	0.048333	0.088342	0.0001
rs7550207	0.154997	0.18	0.0258
rs77389579	0.005001	0.011668	0.0098
rs863004	0.153327	0.18	0.013
SNP ID		**Kisii**	
	Pre-Removal	**Post-Removal**	**p-value**

rs12307123	0.37833	0.33333	0.0001
rs1566830	0.39	0.345	0.0001
rs75731597	0.073337	0.086665	0.0207

SNP ID	**Somali**		
	Pre-Removal	**Post-Removal**	**p-value**
rs12307123	0.29666	0.26667	0.0006
rs1371478	0.325	0.33999	0.023
rs1541255	0.21999	0.20168	0.0384
rs1566830	0.29666	0.26667	0.0006
rs1800890	0.38333	0.41334	0.0113
rs3742785	0.42001	0.45167	0.0001
rs77389579	0.021666	0.010002	0.0002
rs8386	0.15835	0.19333	0.0011

322

Discussion

Kenya is a highly diverse country with significant ethnolinguistic diversity as well as substantial admixture between its ethnic groups (2,4). The ancestral diversity of these groups and unique historical trajectories has the potential to major differentiation in allele frequencies of key disease associated variants even among people currently living in the same geographic area. In this study we have compared the allele frequencies of 20 Malaria and Trypanosomiasis associated SNPs, in individuals from 6 ethnolinguistic groups currently residing in the same neighborhood of Nairobi. Please refer to Ramsay *et al* (2016) and Ali *et al* (2018) for details on NHDUSSs, sample collection, genotyping and data QC performed in the AWI-Gen study (17,20). The current study builds on an ongoing broader study on population genetics of the East African sub-set of the AWI-Gen dataset. The major drivers of the allele frequency differences that are reported here, such as, population structure, admixture and how it relates to history will be described and discussed in the manuscript reporting the broader study. The extreme differences in allele frequencies of variants such as the sickle cell disease (SCD) variant (rs334), show the importance of considering ethnicity and language as a factor for stratifying populations from geographic regions with extreme diversity. Considerable difference in the distribution of this SNP in two Ugandan populations has been recently reported (21). The current study, therefore, reiterates the importance of generation of population-level datasets and characterization of fine-scale population structure to enable an accurate assessment of genetic risk for diseases such as SCD.

Allele frequency estimates obtained from public datasets such as 1KGP and HapMap are often used as proxies not only for the population but also for the geographic regions. The inclusion of genotype data from a large number of Luhya participants in our dataset enabled us to compare the allele frequency estimates for Luhya in two independent datasets. Despite overall concordance and correlation between the estimates, for some of the key disease associated

variants demonstrated almost two-fold difference between the groups. It is understandable that some of these differences could be due to the fact the Luhya samples from the 1KGP dataset were sampled from Webuye which is about 400 km northeast of Nairobi, the sampling site of our samples. Nevertheless, these extreme differences indicate that the use of proxies, especially for disease studies needs to more cautious and nuanced.

For populations residing in urban African centers such as Nairobi that are co-inhabited by different ethnolinguistic groups for a considerable amount of time, there is high possibility of gene flow between different ethnolinguistic groups, similar to what was observed for ethnolinguistic groups in South Africa (5). Self-reported ethnicity as an identifier might not capture this gene flow and datasets originating from such centers, could therefore under-estimate the overall allele frequency difference between ethnolinguistic groups. To assess the extent to which the use of self-reported ethnicity as a classifier can capture the underlying genetic difference between ethnolinguistic groups, we have presented a comparison of clustering obtained using this to the clusters defined by genetic data.

Even with a permissive centroid-based filtering based on first two PCs and 5 standard deviations, between 2 and 5 percent of the samples were excluded as outliers from each group. A more stringent and realistic filtering based on 4-10 PCs can be expected to lead to the filtering of a much higher proportion of outliers. Moreover, despite the number of outliers being removed were relatively low, we found that for some of the SNPs the filtering in fact led to statistically significant differences in allele frequencies differences between genetic data defined and self-reported ethnicity defined clusters.

There are large scale differences in the representation of different ethnolinguistic groups in the dataset. This led to the exclusion of some of the groups. Moreover, these differences meant that the robustness of allele frequency estimates varied between these ethnolinguistic groups. The

372 subsample sizes of 30 likely resulted in a loss of statistical power for testing allele frequency
373 differences. Finally, due to the nature of the study centre, only volunteers who could access the
374 study centre were recruited, such that participants were selected by convenience.

375 In summary, self-reported ethnicity, despite emerging as a relatively reliable estimate for most of
376 the groups included in our study, might not always provide fully accurate estimates. Future work
377 needs into focus into which additional factors could lead to better concordance between genetic
378 and self-reported based clustering and thereby more robust estimate for allele frequencies for
379 populations in urban African settings. Moreover, further investigations that take into
380 consideration parent and grandparent ethnicities are expected to provide more robust estimates
381 of allele frequencies and more accurate estimates of allele difference between ethnolinguistic
382 groups.

Acknowledgements

We would like to thank to all Kenyan participants of the H3Africa AWI-Gen study, whose genotype and ethnic identity data was used in this project. We thank Prof Michele Ramsay for granting access to the H3Africa AWI-Gen datasets and Prof Scott Hazelhurst for providing access to the Wits Computational cluster.

Funding

Yonatan Ariel Wolberg was supported by a fellowship from the National Research Foundation (NRF) South Africa.

References

1. Balaton-Chrimes S. Who are Kenya's 42(+) tribes? The census and the political utility of magical uncertainty. J East Afr Stud. 2021 Jan 2;15(1):43–62.

2. Gurdasani D, Carstensen T, Tekola-Ayele F, Pagani L, Tachmazidou I, Hatzikotoulas K, et al. The African Genome Variation Project shapes medical genetics in Africa. Nature. 2015 Jan;517(7534):327–32.

3. User S. Publications [Internet]. Kenya National Bureau of Statistics. [cited 2022 Apr 28]. Available from: https://www.knbs.or.ke/publications/

4. Atkinson EG, Dalvie S, Pichkar Y, Kalungi A, Majara L, Stevenson A, et al. Genetic structure correlates with ethnolinguistic diversity in eastern and southern Africa [Internet]. bioRxiv; 2021 [cited 2022 Feb 14]. p. 2021.05.19.444732. Available from: https://www.biorxiv.org/content/10.1101/2021.05.19.444732v1

5. Sengupta D, Choudhury A, Fortes-Lima C, Aron S, Whitelaw G, Bostoen K, et al. Genetic substructure and complex demographic history of South African Bantu speakers. Nat Commun [Internet]. 2021 [cited 2022 Feb 21];12(1). Available from: http://hdl.handle.net/1854/LU-8703344

6. Beltrame MH, Rubel MA, Tishkoff SA. Inferences of African evolutionary history from genomic data. Curr Opin Genet Dev. 2016 Dec;41:159–66.

7. Musuva A, Ejersa W, Kiptui R, Memusi D, Abwao E. The malaria testing and treatment landscape in Kenya: results from a nationally representative survey among the public and private sector in 2016. Malar J. 2017 Dec 21;16:494.

8. Kariuki SN, Williams TN. Human genetics and malaria resistance. Hum Genet. 2020;139(6):801–11.

9. Ndila CM, Uyoga S, Macharia AW, Nyutu G, Peshu N, Ojal J, et al. Human candidate gene polymorphisms and risk of severe malaria in children in Kilifi, Kenya: a case-control association study. Lancet Haematol. 2018 Aug 1;5(8):e333–45.

10. Greene JA, Moormann AM, Vulule J, Bockarie MJ, Zimmerman PA, Kazura JW. Toll-like receptor polymorphisms in malaria-endemic populations. Malar J. 2009 Mar 24;8:50.

11. Kulohoma BW, Wamwenje SAO, Wangwe II, Masila N, Mirieri CK, Wambua L. Prevalence of trypanosomes associated with drug resistance in Shimba Hills, Kwale County, Kenya. BMC Res Notes. 2020 Apr 29;13(1):234.

12. Watson HJC. The epidemiology of human sleeping sickness in the Lambwe Valley, South Nyanza, Kenya. Bull World Health Organ. 1972;47(6):719–26.

13. Kirchhoff LV. American Trypanosomiasis (Chagas' Disease) – A Tropical Disease Now in the United States. N Engl J Med. 1993 Aug 26;329(9):639–44.

14. African Trypanosomiasis in Kenya - Watch - Level 1, Practice Usual Precautions - Travel Health Notices | Travelers' Health | CDC [Internet]. [cited 2022 Feb 19]. Available from: https://wwwnc.cdc.gov/travel/notices/watch/african-trypanosomiasis-in-kenya

15. Nielsen MJ, Petersen SV, Jacobsen C, Oxvig C, Rees D, Møller HJ, et al. Haptoglobin-related protein is a high-affinity hemoglobin-binding plasma protein. Blood. 2006 Oct 15;108(8):2846–9.

16. Smith AB, Esko JD, Hajduk SL. Killing of Trypanosomes by the Human Haptoglobin-Related Protein. Science. 1995 Apr 14;268(5208):284–6.

17. Ramsay M, Crowther N, Tambo E, Agongo G, Baloyi V, Dikotope S, et al. H3Africa AWI-Gen Collaborative Centre: a resource to study the interplay between genomic and environmental risk factors for cardiometabolic diseases in four sub-Saharan African countries. Glob Health Epidemiol Genomics. 2016;1:e20.

18. User S. Publications [Internet]. Kenya National Bureau of Statistics. [cited 2022 Apr 28]. Available from: https://www.knbs.or.ke/publications/

19. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007 Sep;81(3):559–75.

20. Ali SA, Soo C, Agongo G, Alberts M, Amenga-Etego L, Boua RP, et al. Genomic and environmental risk factors for cardiometabolic diseases in Africa: methods used for Phase 1 of the AWI-Gen population cross-sectional study. Glob Health Action. 2018 Sep 27;11(Suppl 2):1507133.

21. Choudhury A, Aron S, Botigué LR, Sengupta D, Botha G, Bensellak T, et al. High-depth African genomes inform human migration and health. Nature. 2020 Oct;586(7831):741–8.

22. Largest Ethnic Groups In Kenya [Internet]. WorldAtlas. 2019 [cited 2022 Feb 13]. Available from: https://www.worldatlas.com/articles/largest-ethnic-groups-in-kenya.html

23. Kenya [Internet]. PMI. [cited 2022 Feb 14]. Available from: https://www.pmi.gov/where-we-work/kenya/

24. Prevention CC for DC and. CDC - Malaria - Malaria Worldwide - CDC's Global Malaria Activities - Kenya [Internet]. 2019 [cited 2022 Feb 14]. Available from: https://www.cdc.gov/malaria/malaria_worldwide/cdc_activities/kenya.html

25. Maxfield L, Bermudez R. Trypanosomiasis. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2022 [cited 2022 Feb 19]. Available from: http://www.ncbi.nlm.nih.gov/books/NBK535413/

26. Maeda N, McEvoy SM, Harris HF, Huisman TH, Smithies O. Polymorphisms in the human haptoglobin gene cluster: chromosomes with multiple haptoglobin-related (Hpr) genes. Proc Natl Acad Sci. 1986 Oct 1;83(19):7395–9.

27. Hardwick RJ, Ménard A, Sironi M, Milet J, Garcia A, Sese C, et al. Haptoglobin (HP) and Haptoglobin-related protein (HPR) copy number variation, natural selection, and trypanosomiasis. Hum Genet. 2014 Jan;133(1):69–83.

28. McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. ArXiv180203426 Cs Stat [Internet]. 2020 Sep 17 [cited 2022 Apr 18]; Available from: http://arxiv.org/abs/1802.03426

Appendix

Supplementary Table 1. Details of SNPs used in this study. The data presented in this table was obtained from ClinVar and dbSNP. In the case of SNP rs8176719, the mutation is the insertion of a C nucleotide, while for SNP rs143830837, the mutation is the deletion of a TTATAA sequence.

	SNP ID	Gene	Disease Allele	GRCh38	GRCh37
1	rs3211938	CD36	G	7: 80671133	7: 80300449
2	rs12075	DARC	A	1: 159205564	1: 159175354
3	rs2814778	DARC	C	1: 159204893	1: 159174683
4	rs3027011	DARC	G	1: 159204305	1: 159174095
5	rs55872368	DARC	A,T	1: 159207066	1: 159176856
6	rs7550207	DARC	C	1: 159205095	1: 159174885
7	rs863004	DARC	T	1: 159207041	1: 159176831
8	rs1050828	G6PD	T	X: 154536002	X: 153764217
9	rs1050829	G6PD	A,C	X: 154535277	X: 153763492
10	rs334	HBB	A,C,G	11: 5227002	11: 5248232
11	rs33930165	HBB	A,G,T	11: 5227003	11: 5248233
12	rs8176719	ABO	->C	9: 133257521	9: 136132909
13	rs8176746	ABO	A,T	9: 133255935	9: 136131322
14	rs1541255	ATP2B4	A,T	1: 203683013	1: 203652141
15	rs75731597	ARL14	C,T	3: 160647020	3: 160364808
16	rs3092945	CD40LG	C	X: 136647450	X: 135729609
17	rs186873296	FREM3	G	4: 143781321	4: 144702474
18	rs77389579	INPP4B	T	4: 142617358	4: 143538511

19	rs1800890	IL10	C,T	1: 206776020	1: 206949365
20	rs72933304	LPHN2	A	1: 81260453	1: 81726138
21	rs1371478	LOC727982	A,C,G	2: 4853999	2: 4901589
22	rs3742785	RPS6KL1	C	14: 74906331	14: 75373034
23	rs10459266	CAND1	A	12: 67062108	12: 67455888
24	rs12307123	CAND1	T	12: 67001170	12: 67394950
25	rs1566830	CAND1	T	12: 66976118	12: 67369898
26	rs8386	GNAS	T	20: 58910757	20: 57485812
27	rs11865131	NPRL3	A,C	16: 113668	16: 163667
28	rs60910145	APOL1	C,G	22: 36265988	22: 36662034
29	rs73885319	APOL1	G	22: 36265860	22: 36661906
30	rs143830837	APOL1	TTATAA>-	22: 36266000	22: 36662046

Supplementary Table 2. Comparison of absolute MAFs between genotype and imputed datasets, pre- and post-exclusion of outlier individuals. (A) represents MAFs pre-removal, while (B) represents MAFs post-removal. Values are rounded off to 2 decimals wherever possible. The letters 'G' and 'I' refer to genotype and imputed data, respectively.

A	**Ethnic Group**											
SNP ID	**Kikuyu**		**Kamba**		**Luhya**		**Luo**		**Kisii**		**Somali**	
	G	**I**	**G**	**I**	**G**	**I**	**G**	**I**	**G**	**I**	**G**	**I**
rs12075	0.02	0.01	0.003	0.003	0.002	0.002	0	0	0	0	0.04	0.04
rs1800890	0.27	0.27	0.28	0.27	0.22	0.23	0.25	0.25	0.22	0.22	0.39	0.39
rs73885319	0.04	0.04	0.05	0.05	0.07	0.06	0.07	0.07	0.04	0.04	0	0
rs77389579	0.04	0.04	0.06	0.06	0.02	0.02	0.01	0.01	0	0	0.02	0.02
rs8386	0.11	0.11	0.14	0.14	0.15	0.15	0.14	0.14	0.09	0.09	0.18	0.18
rs863004	0.12	0.12	0.12	0.12	0.15	0.14	0.17	0.17	0.12	0.12	0.13	0.13

B	**Ethnic Group**

SNP ID	Kikuyu		Kamba		Luhya		Luo		Kisii		Somali	
	G	I	G	I	G	I	G	I	G	I	G	I
rs12075	0.02	0.01	0.003	0.003	0	0	0	0	0	0	0.04	0.04
rs1800890	0.28	0.28	0.28	0.27	0.22	0.22	0.25	0.25	0.22	0.22	0.39	0.39
rs73885319	0.04	0.04	0.05	0.05	0.07	0.06	0.07	0.07	0.04	0.04	0	0
rs77389579	0.04	0.04	0.06	0.06	0.02	0.02	0.01	0.01	0	0	0.01	0.01
rs8386	0.11	0.11	0.14	0.14	0.14	0.14	0.14	0.14	0.09	0.09	0.18	0.18
rs863004	0.12	0.12	0.12	0.13	0.15	0.14	0.17	0.17	0.12	0.12	0.13	0.13

494

495

496

Supplementary Table 3. Comparision of pre- and post-removal of excluded individuals. All values below are derived from the imputed dataset. Values are rounded off to 2 decimal places wherever possible and standard errors are shown as numbers following the '±' symbol. Each letter refers to a different ethnic group: A) Kikuyu; B) Kamba; C) Luhya; D) Luo; E) Kisii; and F) Somali.

	Ethnic Group				
A	**Kikuyu**				
SNP ID	**Pre-Removal Absolute**	**Pre-Removal Bootstrapped**	**Post-Removal Absolute**	**Post-Removal Bootstrapped**	**P-value**
rs10459266	0.21	0.22 ± 0.004	0.21	0.23 ± 0.01	0.43
rs11865131	0.14	0.15 ± 0.009	0.14	0.13 ± 0.007	0.04
rs12075	0.01	0.02 ± 0.002	0.01	0.01 ± 0.003	0.61
rs12307123	0.22	0.23 ± 0.01	0.22	0.2 ± 0.01	0.16
rs1371478	0.3	0.27 ± 0.01	0.3	0.29 ± 0.007	0.21
rs1541255	0.28	0.25 ± 0.008	0.27	0.28 ± 0.008	0.03
rs1566830	0.22	0.22 ± 0.01	0.22	0.2 ± 0.01	0.42
rs1800890	0.27	0.27 ± 0.01	0.28	0.27 ± 0.01	0.95
rs3027011	0.11	0.13 ± 0.01	0.11	0.1 ± 0.008	0.05
rs334	0.002	0.002 ± 0.0009	0.001	0	0.08
rs3742785	0.44	0.41 ± 0.009	0.44	0.43 ± 0.006	0.04
rs55872368	0.14	0.17 ± 0.011	0.14	0.12 ± 0.009	0.0007
rs60910145	0.04	0.03 ± 0.005	0.03	0.03 ± 0.005	0.98
rs72933304	0.16	0.16 ± 0.009	0.16	0.18 ± 0.007	0.17
rs73885319	0.04	0.03 ± 0.004	0.04	0.03 ± 0.005	0.82
rs7550207	0.1	0.12 ± 0.01	0.1	0.1 ± 0.007	0.06

rs75731597	0.1	0.1 ± 0.007	0.1	0.07 ± 0.006	0.02
rs77389579	0.04	0.05 ± 0.005	0.04	0.05 ± 0.004	0.8
rs8386	0.11	0.15 ± 0.005	0.11	0.1 ± 0.004	0.0001
rs863004	0.12 0.21	0.14 ± 0.01	0.12 0.21	0.11 ± 0.008	0.03
B			**Kamba**		
SNP ID	**Pre-Removal Absolute**	**Pre-Removal Bootstrapped**	**Post-Removal Absolute**	**Post-Removal Bootstrapped**	**P-value**
rs10459266	0.2	0.19 ± 0.008	0.19	0.19 ± 0.009	0.88
rs11865131	0.13	0.14 ± 0.007	0.13	0.14 ± 0.007	0.36
rs12075	0.003	0.003 ± 0.001	0.003	0.007 ± 0.002	0.18
rs12307123	0.23	0.22 ± 0.007	0.23	0.22 ± 0.01	0.75
rs1371478	0.3	0.26 ± 0.009	0.3	0.33 ± 0.01	0.0001
rs1541255	0.34	0.35 ± 0.01	0.34	0.32 ± 0.009	0.1
rs1566830	0.23	0.22 ± 0.006	0.23	0.21 ± 0.01	0.53
rs1800890	0.27	0.3 ± 0.006	0.27	0.3 ± 0.009	0.96
rs3027011	0.13	0.13 ± 0.007	0.14	0.12 ± 0.007	0.12
rs334	0.003	0.007 ± 0.002	0.003	0.007 ± 0.002	0.94
rs3742785	0.41	0.43 ± 0.007	0.41	0.41 ± 0.006	0.04
rs55872368	0.13	0.13 ± 0.005	0.13	0.12 ± 0.008	0.24
rs60910145	0.05	0.05 ± 0.003	0.05	0.05 ± 0.004	0.51
rs72933304	0.14	0.14 ± 0.008	0.14	0.12 ± 0.004	0.15
rs73885319	0.05	0.05 ± 0.003	0.05	0.05 ± 0.004	0.74
rs7550207	0.12	0.12 ± 0.006	0.13	0.11 ± 0.007	0.14
rs75731597	0.08	0.11 ± 0.003	0.08	0.09 ± 0.008	0.19
rs77389579	0.06	0.07 ± 0.004	0.06	0.05 ± 0.005	0.01

rs8386	0.14	0.15 ± 0.01	0.14	0.14 ± 0.006	0.35
rs863004	0.12	0.12 ± 0.006	0.13	0.11 ± 0.008	0.32

502

C			Luhya		
SNP ID	**Pre-Removal Absolute**	**Pre-Removal Bootstrapped**	**Post-Removal Absolute**	**Post-Removal Bootstrapped**	**P-value**
rs10459266	0.21	0.2 ± 0.006	0.21	0.21 ± 0.01	0.31
rs11865131	0.19	0.19 ± 0.01	0.2	0.22 ± 0.005	0.05
rs12075	0.002	0.002 ± 0.001	0	0	0.08
rs12307123	0.23	0.24 ± 0.01	0.24	0.23 ± 0.004	0.47
rs1371478	0.29	0.3 ± 0.009	0.29	0.26 ± 0.01	0.03
rs1541255	0.4	0.39 ± 0.01	0.4	0.38 ± 0.01	0.63
rs1566830	0.24	0.24 ± 0.01	0.24	0.24 ± 0.005	0.75
rs1800890	0.23	0.21 ± 0.01	0.22	0.22 ± 0.008	0.53
rs3027011	0.14	0.13 ± 0.009	0.14	0.15 ± 0.008	0.22
rs334	0.06	0.07 ± 0.005	0.06	0.07 ± 0.007	0.51
rs3742785	0.38	0.36 ± 0.009	0.38	0.38 ± 0.007	0.07
rs55872368	0.14	0.14 ± 0.008	0.14	0.15 ± 0.008	0.29
rs60910145	0.06	0.06 ± 0.007	0.06	0.07 ± 0.007	0.34
rs72933304	0.08	0.08 ± 0.007	0.07	0.07 ± 0.006	0.43
rs73885319	0.06	0.06 ± 0.007	0.06	0.07 ± 0.007	0.34
rs7550207	0.14	0.13 ± 0.009	0.14	0.15 ± 0.008	0.22
rs75731597	0.11	0.11 ± 0.008	0.11	0.12 ± 0.007	0.46
rs77389579	0.02	0.02 ± 0.002	0.02	0.01 ± 0.003	0.007
rs8386	0.15	0.15 ± 0.008	0.14	0.14 ± 0.009	0.48
rs863004	0.14	0.13 ± 0.009	0.14	0.15 ± 0.008	0.28

503

D			Luo		
SNP ID	**Pre-Removal Absolute**	**Pre-Removal Bootstrapped**	**Post-Removal Absolute**	**Post-Removal Bootstrapped**	**P-value**
rs10459266	0.22	0.22 ± 0.01	0.22	0.21 ± 0.006	0.38
rs11865131	0.17	0.15 ± 0.007	0.16	0.19 ± 0.009	0.004
rs12075	0	0	0	0	1
rs12307123	0.24	0.24 ± 0.01	0.23	0.23 ± 0.01	0.34
rs1371478	0.28	0.25 ± 0.009	0.27	0.27 ± 0.01	0.16
rs1541255	0.47	0.45 ± 0.006	0.48	0.45 ± 0.007	1
rs1566830	0.25	0.24 ± 0.01	0.25	0.25 ± 0.01	0.7
rs1800890	0.25	0.28 ± 0.01	0.25	0.26 ± 0.009	0.28
rs3027011	0.17	0.16 ± 0.008	0.17	0.18 ± 0.008	0.02
rs334	0.11	0.13 ± 0.007	0.11	0.1 ± 0.006	0.0009
rs3742785	0.41	0.4 ± 0.01	0.41	0.42 ± 0.009	0.23
rs55872368	0.17	0.16 ± 0.008	0.17	0.18 ± 0.008	0.02
rs60910145	0.07	0.05 ± 0.005	0.07	0.09 ± 0.007	0.0001
rs72933304	0.05	0.04 ± 0.003	0.05	0.05 ± 0.003	0.004
rs73885319	0.07	0.05 ± 0.005	0.07	0.09 ± 0.007	0.0001
rs7550207	0.17	0.16 ± 0.008	0.17	0.18 ± 0.007	0.03
rs75731597	0.09	0.08 ± 0.003	0.09	0.1 ± 0.007	0.06
rs77389579	0.01	0.005 ± 0.001	0.01	0.01 ± 0.002	0.01
rs8386	0.14	0.14 ± 0.009	0.14	0.15 ± 0.007	0.29
rs863004	0.17	0.15 ± 0.007	0.17	0.18 ± 0.007	0.01

504

E			Kisii		
SNP ID	**Pre-Removal Absolute**	**Pre-Removal Bootstrapped**	**Post-Removal Absolute**	**Post-Removal Bootstrapped**	**P-value**
rs10459266	0.26	0.25 ± 0.006	0.26	0.25 ± 0.006	0.57
rs11865131	0.15	0.14 ± 0.005	0.15	0.15 ± 0.003	0.06
rs12075	0	0	0	0	1
rs12307123	0.36	0.38 ± 0.006	0.37	0.33 ± 0.006	0.0001
rs1371478	0.29	0.29 ± 0.006	0.29	0.29 ± 0.01	0.78
rs1541255	0.39	0.39 ± 0.006	0.4	0.39 ± 0.006	0.44
rs1566830	0.37	0.39 ± 0.006	0.38	0.35 ± 0.006	0.0001
rs1800890	0.22	0.21 ± 0.007	0.22	0.21 ± 0.004	0.42
rs3027011	0.11	0.11 ± 0.004	0.11	0.12 ± 0.004	0.25
rs334	0.04	0.05 ± 0.003	0.04	0.04 ± 0.003	0.41
rs3742785	0.47	0.46 ± 0.006	0.47	0.47 ± 0.005	0.15
rs55872368	0.12	0.12 ± 0.003	0.12	0.13 ± 0.004	0.15
rs60910145	0.04	0.04 ± 0.003	0.04	0.04 ± 0.003	0.39
rs72933304	0.04	0.04 ± 0.003	0.04	0.04 ± 0.003	0.65
rs73885319	0.04	0.04 ± 0.002	0.04	0.04 ± 0.003	0.39
rs7550207	0.12	0.12 ± 0.003	0.12	0.13 ± 0.004	0.15
rs75731597	0.07	0.07 ± 0.005	0.08	0.09 ± 0.002	0.02
rs77389579	0	0	0	0	1
rs8386	0.09	0.08 ± 0.003	0.09	0.08 ± 0.002	0.68
rs863004	0.12	0.12 ± 0.003	0.12	0.13 ± 0.004	0.15

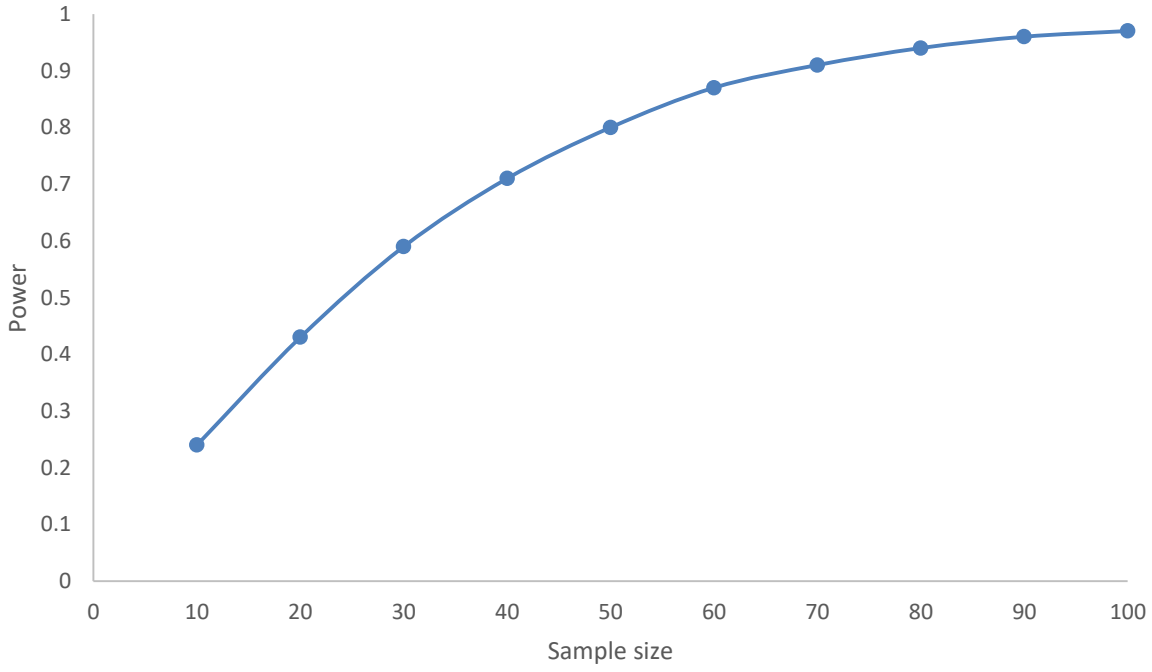
505

F			Somali		
SNP ID	**Pre-Removal Absolute**	**Pre-Removal Bootstrapped**	**Post-Removal Absolute**	**Post-Removal Bootstrapped**	**P-value**
rs10459266	0.27	0.25 ± 0.007	0.26	0.27 ± 0.004	0.08
rs11865131	0.19	0.19 ± 0.007	0.18	0.18 ± 0.005	0.24
rs12075	0.04	0.04 ± 0.002	0.04	0.04 ± 0.003	0.32
rs12307123	0.28	0.3 ± 0.005	0.27	0.27 ± 0.006	0.0006
rs1371478	0.33	0.33 ± 0.004	0.34	0.34 ± 0.005	0.02
rs1541255	0.2	0.22 ± 0.006	0.21	0.2 ± 0.006	0.04
rs1566830	0.28	0.3 ± 0.005	0.27	0.27 ± 0.006	0.0006
rs1800890	0.39	0.38 ± 0.007	0.39	0.4 ± 0.009	0.01
rs3027011	0.07	0.07 ± 0.003	0.08	0.08 ± 0.004	0.47
rs334	0	0	0	0	1
rs3742785	0.44	0.42 ± 0.006	0.46	0.45 ± 0.005	0.0001
rs55872368	0.17	0.16 ± 0.005	0.17	0.17 ± 0.005	0.13
rs60910145	0	0	0	0	1
rs72933304	0.18	0.19 ± 0.005	0.17	0.18 ± 0.008	0.6
rs73885319	0	0	0	0	1
rs7550207	0.08	0.09 ± 0.003	0.09	0.09 ± 0.004	1
rs75731597	0.06	0.06 ± 0.003	0.07	0.06 ± 0.003	0.22
rs77389579	0.02	0.02 ± 0.002	0.01	0.01 ± 0.002	0.0002
rs8386	0.18	0.16 ± 0.006	0.18	0.19 ± 0.008	0.001
rs863004	0.13	0.13 ± 0.004	0.13	0.12 ± 0.005	0.42

506

Supplementary Note 1

Sample Size Calculation: The optimal sample size calculations for testing allele frequency differences between two populations depends concomitantly on two components - level of difference in frequencies between the two groups and the rarity of the individual allele being compared. Alleles which occur at lower frequencies (say MAF=0.01) would require much larger sample sizes compared to alleles that occur at higher frequencies (say MAF=0.10). Similarly, larger sample sizes would also be required to test for small-scale differences in allele frequencies. For our calculations we assumed these two parameters to be incorporated into the effect size estimate. We used the R package PWR to calculated the power of Chi-squared test at different effect size and sample size ranges.



Supplementary Figure 1: Statistical power for Chi-squared test-based discrimination of allele frequencies between two populations. The sample size of 30 has 60% power and a sample size of 80% power to detect allele frequency differences over 0.40 between two populations.

Appendix A

Approved Research Proposal

UNIVERSITY OF THE
WITWATERSRAND,
JOHANNESBURG

Private Bag 3 Wits, 2050
Fax: 027117172119
Tel: 02711 7172076

Reference: Mrs Sandra Benn
E-mail: sandra.benn@wits.ac.za

13 June 2022
Person No: 1834647
PAG

Mr YA Wolberg
74 Mejon Street
Glenhazel
2192
South Africa

Dear Mr Yonatan Wolberg

Master of Science in Medicine: Approval of Title

We have pleasure in advising that your proposal entitled *Correlations between self-reported ethnicity and language, and genetic clustering in Kenya, with reference to the allele frequency of malaria- and trypanosomiasis-associated variants* has been approved. Please note that any amendments to this title have to be endorsed by the Faculty's higher degrees committee and formally approved.

Yours sincerely

Mrs Sandra Benn
Faculty Registrar
Faculty of Health Sciences

CANDIDATE'S SURNAME: Wolberg	FIRST NAME/S: Yonatan Ariel		STUDENT NUMBER: 183 46 47
CURRENT QUALIFICATIONS: BSc Honours Genetics and Developmental Biology, BSc Molecular and Cell Biology			
TEL: 078 363 9175	CELL: 078 363 9175	E-MAIL: 1834647@students.wits.ac.za	FAX:
DEGREE FOR WHICH PROTOCOL IS BEING SUBMITTED: MSc Medicine Genomic Medicine			
PART-TIME OR FULL-TIME: Full-Time			
FIRST REGISTERED FOR THIS DEGREE:	TERM : January		YEAR: 2022
DEPARTMENT: Human Genetics			
TITLE OF PROPOSED RESEARCH: Correlations between self-reported ethnicity and language with genetic clustering in Kenya.			
CANDIDATE'S SIGNATURE:			DATE: 27/04/2022
SUPERVISOR 1 (NAME & SURNAME): Ananyo Choudhury			% Supervision :50
SUPERVISOR'S QUALIFICATIONS : PhD			
SUPERVISOR'S DEPARTMENT: SBIMB			
SUPERVISOR'S ADDRESS / TEL / E-MAIL: Ananyo.Choudhury@wits.ac.za			
SUPERVISOR 2 (NAME & SURNAME): Dhriti Sengupta			% Supervision :50
SUPERVISOR'S QUALIFICATIONS : PhD			
SUPERVISOR'S ADDRESS / TEL / E-MAIL: dhriti.sengupta@wits.ac.za			
SUPERVISOR 3 (NAME & SURNAME): NA			% Supervision : NA
SUPERVISOR'S QUALIFICATIONS : NA			
SUPERVISOR'S ADDRESS / TEL / E-MAIL: NA			

SYNOPSIS OF RESEARCH:*(Brief summary of proposed research project; between 200-300 words only; with sub-headings: an introduction and justification for study, aim/s, proposed methodology and expected outcome/s)*
[Use reverse side of this page if
more space is required]

Introduction and Justification

Kenya is a highly diverse country and the combination of recent local migrations and admixture have contributed to a complex population structure. This structure creates a dilemma when trying to assess the allele frequencies of disease-associated variants within the country as different groups will show different frequencies. Additionally, ethnic groups in genetic studies are often defined on the basis of self-reported identity but certain individuals may align genetically to another ethnic group.

Aims

This study aims to determine how self-reported ethnicity correlates to genetic clustering in a Kenyan cohort. The effect of the discordance between the two on the frequencies of key malaria- and trypanosomiasis-associated variants will thereafter be determined.

Proposed Methodology

Leveraging the H3Africa AWI-Gen dataset, the allele frequencies of a list of key malaria- and trypanosomiasis-associated variants will be estimated for each self-reported ethnic group. Following this, genetic data-based filtering will be performed by generating centroids for each ethnic group, with 10 principal components (PCs) used to delineate them. Individuals will be excluded if they localize to any centroid other than that representing their self-reported ethnicity, or are located more than 5 standard deviations from the group mean of any of the 10 PCs. Uniform Manifold Approximation and Projection will enable the compression of these results from 10 to 2 PCs, to validate previous classifications. The allele frequencies will be re-estimated after excluding individuals and chi-squared tests and Fischer tests will be performed to assess the significance of differences in allele frequencies pre- and post-exclusion.

Expected Outcomes

By revealing significant differences in allele frequencies in each ethnolinguistic group, this study hopes to demonstrate the need for determining an individual's genetic grouping before performing allele frequency estimation. Time permitting, the use of language, parental and grandparental ethnicities may prove useful in minimising the discordance between self-reported ethnicity and genetic clustering.

<table>
<tr><td colspan="2">WITS ETHICS NOT REQUIRED: No
WITS ETHICS PENDING: No
WITS ETHICS APPROVED: Yes
(circle appropriate symbol)*

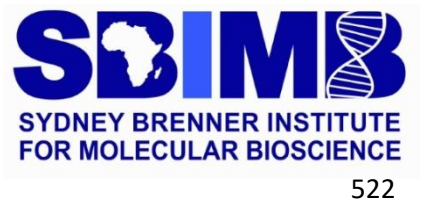
*Please note the final human ethics clearance certificate or animal ethics certificate must be available prior to starting research</td><td>IF Y SUPPLY ETHICS CLEARANCE CERTIFICATE AS ATTACHMENT AND INCLUDE ETHICS NUMBER HERE:

M220294</td></tr>
<tr><td colspan="3">As supervisor/s, I/we confirm that I have read the protocol which has been submitted for assessment.</td></tr>
<tr><td>SIGNATURE OF SUPERVISOR/S:</td><td>Anango Choudhury

……</td><td></td></tr>
<tr><td>SIGNATURE PG OFFICE STAFF

……..</td><td>REGISTERED

YES….. NO…..</td><td>STAMP</td></tr>
</table>



522





Research Proposal

Correlations between self-reported ethnicity and language with genetic clustering in Kenya

Yonatan Ariel Wolberg

183 46 47

Supervisor: Dr. Ananyo Choudhury

Co-supervisor: Dr. Dhriti Sengupta

List of Abbreviations

AGVP	African Genome Variation Project
APOL1	Apolipoprotein 1
AWI-Gen	Africa Wits-INDEPTH Partnership for Genomic Studies
CDC	Centre for Disease Control and Prevention
CMD	Cardiometabolic Disease
DPHRU	Developmental Pathways to Health Research Unit
G6PD	Glucose-6-Phosphate Dehydrogenase
GWAS	Genome-Wide Association Studies
H3Africa	Human Heredity and Health in Africa
Hb	Haemoglobin
HbS	Sickle Cell Allele (Hemoglobin Sickle)
HPR	Haptoglobin-Related Protein
HDSS	Health and Demographic Surveillance System
HREC	Human Research Ethics Committee
NeuroGAP-Psychosis	Neuropsychiatric Genetics of African Populations-Psychosis
NRF	National Research Foundation
NUHDSS	Nairobi Urban Health and Demographic Surveillance System
PC	Principal Component
PCA	Principal Component Analysis
PMI	United States' President's Malaria Initiative
POPIA	Protection of Personal Information Act

SCD	Sickle Cell Disease
SNP	Single Nucleotide Polymorphism
SSA	Sub-Saharan Africa
TLF1	Trypanosome lytic factor-1
UMAP	Uniform Manifold Approximation and Projection
WGS	Whole Genome Sequencing
WHO	World Health Organization

1. Introduction

1.1. Genetic Diversity in East African Populations

Kenya is a highly diverse country, with about 42 different ethnic groups representing of all three African linguistic divisions: Niger-Congo, Nilo-Saharan and Afro-Asiatic (1,22). Local migrations have resulted in the distribution of these groups in a complex geographic pattern across the country (Fig 1) (2,4,18).

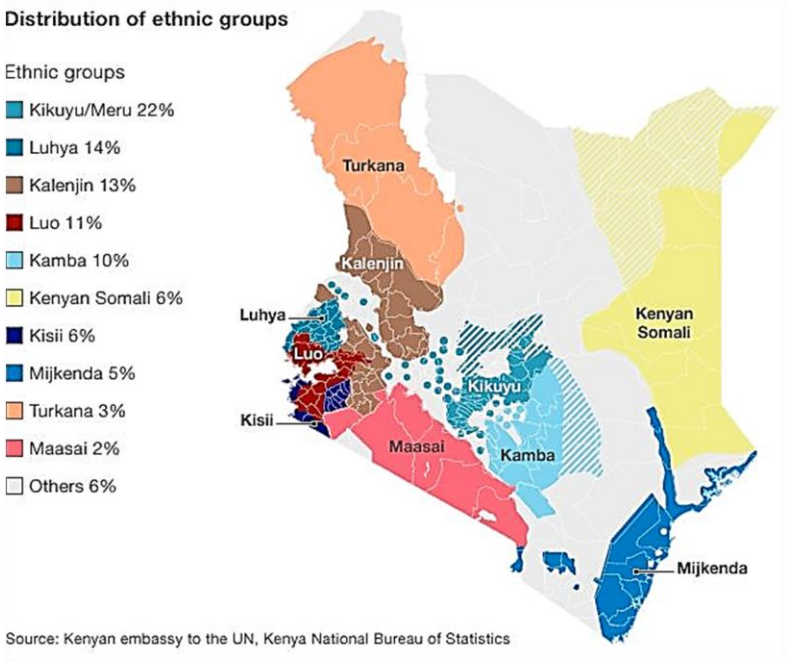


Figure 4. Geographic distribution of major ethnolinguistic groups in Kenya. Each group is represented by a unique colour, and the percentage next to each group shows its contribution to the total contribution. Figure obtained from the Kenya National Bureau of Statistics (5).

The African Genome Variation Project (AGVP) studied the genotypic diversity in 1 481 individuals, representing 15 ethnic groups from sub-Saharan Africa (SSA), using the Omni 2.5 M array. The study included 4 groups from Kenya and is, so far, the largest genomic dataset from this country. The dataset demonstrated considerable population structure within the country. Moreover, the admixture patterns observed in the Kikuyu, Kalenjin and Luhya populations were among the most complex of all the tested populations, suggesting substantial gene flow between these groups (2).

The Neuropsychiatric Genetics of African Populations-Psychosis (NeuroGAP-Psychosis) study also reported a population structure among the ethnolinguistic groups sampled from Kenya and demonstrated the structure to have a strong correlation to self-reported language (4). This indicated that there was a cultural influence on genetic clustering in combination with a geographical influence (4). However, several cases of misclassifications, i.e. discordance between defined genetics and self-reported identities, were observed (4). This can be ascribed to the assimilation of participants into the pervading culture of the region in which they reside, or alternatively, they may have grandparents or parents with different self-reported ethnicities (5).

Pronounced population structure within people inhabiting a geographic region, as observed in Kenya, could lead to major differences in the distribution of disease associated variants within a country. As populations/groups in genetic studies are often defined on the basis of self-reported ethnicity or language, it is important to assess the level of misclassification that might accompany grouping based on these criteria and also how this could impact the estimation of the frequency of disease-associated alleles. A common approach to estimate the level of misclassification is to compare principal component analysis (PCA)-defined groupings and self-reported ethnicity-defined groupings. The comparison of allele

frequencies from genetic –based grouping and other grouping approaches could then be used to assess the extent to which such misclassification may influence disease risk estimates.

1.1. Malaria- and Trypanosomiasis-Associated Variants

1.1.1. Malaria-Associated Variants

Malaria and trypanosomiasis are two protozoan diseases that have played major roles in the shaping of African genomes (6). According the United States President's Malaria Initiative (PMI), 70% of Kenya's population is at risk of contracting malaria, with *Plasmodium falciparum*, the deadliest malarial pathogen, becoming more prevalent along the coast (7,23). As per the Centre for Disease Control and Prevention (CDC), this translates to 3.5 million new clinical cases and 10 700 deaths per year (24). High prevalence of malaria is accompanied by increased genetic burdens for the diseases that confers a protective effect against the disease, namely: thalassemia, sickle cell disease (SCD), glucose-6-phosphate dehydrogenase (*G6PD*) deficiencies and ovalocytoses (8).

Over the years several loci have been identified, which either reduce the risk of contracting malaria or provide resistance to and protection against it (Supplementary Table 1). Notably, in a study from 1995 to 2008 involving 2 244 children with severe malaria in Kilifi County, Kenya, Ndila et al. (2018) identified polymorphisms in 15 loci associated with severe malaria: 1) *ABO*; 2) *ATP2B4*; 3) *ARL14*; 4) *CD40LG*; 5) *FREM3*; 6) *INPP4B*; 7) *G6PD*; 8) *HBA1* and *HBA2*; 9) *HBB*; 10) *IL10*; 11) *LPHN2*; 12) *LOC727982*; 13) *RPS6KL1*; 14) *CAND1*; and 15) *GNAS* (9). Additionally, Greene et al. (2009) found *G6PD* and the sickle cell allele (HbS) to be more prominent in Kenya than Papua New Guinea, marking *G6PD* and *HBB* out as a clear starting point for the determination of the allele frequency of malaria-associated variants in Kenya (10).

1.1.2. Trypanosomiasis-Associated Variants

Trypanosomiasis is a potentially fatal disease spread by the tsetse fly, in Africa, and kissing bug, in the Americas (12,13). These flies are prolific in Kenya, with a distribution over 23% of the country, at densities of 1000 per square kilometer (11). East African trypanosomiasis is generally more acute than the West African type, with symptoms, such as high fever, skin rash, headache and muscle aches, appearing 1 to 3 weeks following the bite (14). In more severe cases, the enlargement of the spleen, kidney failure and problems of the heart, have also been observed (14).

With the World Health Organization (WHO) failing to eliminate trypanosomiasis by 2020, it is pertinent that Kenyan genetic variants associated with the disease are identified and their prevalence determined (25). One gene relevant to trypanosomiasis infection is *HPR*, which encodes the haptoglobin-related protein (HPR). Along with haemoglobin (Hb) and apolipoprotein L1 (*APOL1*), this protein forms the trypanosome lytic factor-1 (*TLF1*) which is able to lyse *Trypanosoma brucei*, via oxidative damage due to its peroxidase activity (15,16).

Demonstrating the adaptive stress of trypanosomiasis, Maede et al. (1986) discovered tandem repeats of the *HPR* gene in American black individuals (26). Similarly, Hardwick et al (2014) identified the duplication of *HPR* in west and central Africa at allele frequencies of 15%, with the duplication manifesting in unaffected individuals (27). However, this could not be attributed to natural selection, as the duplication was underrepresented in their children (27).

A list of key variants associated to these diseases has been generated using in-depth literature mining (Supplementary Table 1 and 2). In addition to providing an estimate of variations of disease-related alleles, research on how the incidences of these diseases correlate with these allele frequencies could provide insights into the dynamics of the distribution of both the diseases and populations over time.

1.1. The H3Africa AWI-Gen Collaborative Centre

Initiated in 2012, the Human Heredity and Health in Africa (H3Africa) Africa Wits-INDEPTH partnership for Genomic studies (AWI-Gen) Collaborative Centre was created with the purpose of determining genomic and environmental contributors towards body makeup and fat distribution, as well as cardiometabolic disease (CMD) risk (17). The project focuses on South Africa, Ghana, Kenya and Burkina Faso, and includes 12 000 participants, aged 40 to 60 years old from these countries (Figure 2). Data was obtained via comprehensive questionnaires, comprising demographic, medical and anthropometric data, and biological samples (fasting venous blood and spot urine). The resulting dataset was primed for genome-wide association studies (GWAS) via the H3Africa SNP array, and genomic, epigenomic and disease-relevant studies via the AWI-Gen whole genome sequencing (WGS) dataset (17).

Although initially focused on CMD risk, this dataset has immense value from a population genetic perspective (17). For example, the analysis of the South African participants has enabled a better characterization of the genetic substructure of South African Bantu speakers (5). Moreover, it has also provided glimpses into the extreme allele frequency variation resulting from this (5).

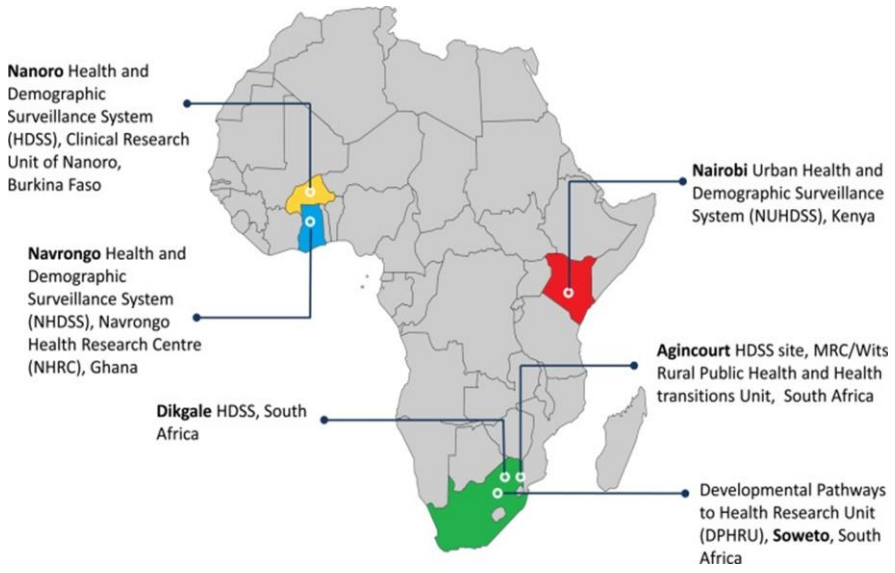


Figure 5. AWI-Gen Study Centres. The participants were recruited at six sites: Nanoro (Burkina Faso); Novrongo (Ghana); Dikgale, Agincourt and Soweto (all South Africa); and Nairobi (Kenya). Figure obtained from Ramsay et al (2016).

1.1.1. East African ethnolinguistic groups

The preliminary analysis of the East African data from about 2000 participants, representing over 6 ethnic groups, obtained from the Nairobi study centre, has shown considerable population structure (Table 1 and Figure 3). Being the capital of Kenya, Nairobi is the socioeconomic hub, a place where the different ethnic groups of the country are able to intermingle and interbreed. This may explain the overlap between certain ethnic groups as seen in Figure 3B. The current study focuses on more in-depth analysis of this dataset to identify the

variation in malaria- and trypanosomiasis-associated variants in Kenya and the best approach to characterize these differences.

Table 1. East African ethnolinguistic groups in the AW-Gen dataset. Only groups represented by 50 or more participants are shown. The recruitment of the participants was done at the Nairobi Urban Health and Demographic Surveillance System (NUHDSS) study centre of the AWI-Gen study.

Ethnolinguistic group	Number
Kikuyu	725
Kamba	393
Luo	374
Luhya	322
Kisii	62
Somali	52

2. Aims and Objectives

The aim of this study is to compare how self-reported ethnicity and language correlate to genetic clustering in this Kenyan cohort. The effect of misclassifications on the allele frequencies of relevant malaria- and trypanosomiasis-associated variants will therefrom be determined.

This will be achieved via the following objectives:

1. The major genetic groups within the AWI-Gen Nairobi participants has been determined through principal component analysis. The study will identify potentially misclassified individuals using a centroid-based filtering of these PC coordinates.

2. A list of Malaria- and Trypanosomiasis-associated variants has been identified using literature mining during the preparation of this protocol. These variants would be searched in the AWI-Gen dataset.
3. For the variants that are present in the dataset, the allele frequencies will be determined in each of the 6 ethnic groups.
4. The allele frequency of these SNPs will then be determined for each group after removing potentially misclassified individuals, and these will be compared with those obtained prior to the removal to assess the impact of misclassification.
5. Time permitting, it will be investigated whether factors such as language, and parental and grandparental identities, could mitigate misclassification.

3. Methodology

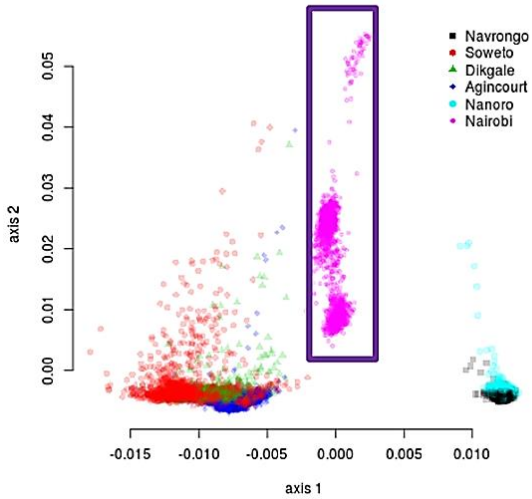
3.1. Genetic clusters based on Principal Component Analysis

PCA is a technique that enables the reduction of the dimensionality of big data sets. This makes it possible to visualize the data set and perform analysis with machine learning algorithms. For genetic datasets, PCA based on genetic distance between individuals is employed to assess the overall population structure and identify major groups in it. Based on the H3Africa AWI-Gen dataset, a PC plot has been generated, which displayed significant population structure within the Kenyan participants (Figure 3A). This study will utilize the preliminary PCA results based on Kenyan participants (Figure 3B) to identify individuals with discrepancies between self-reported identity and genetic clustering.

These PCAs were performed using the open-source PLINK whole genome association analysis toolset (19). The advantage of PLINK is that it allows for the analysis and manipulation of big

datasets, consisting of thousands of participants, each with hundreds of thousands of genetic markers (19). The PCA was visualized using custom R scripts and GENESIS library.

A



B

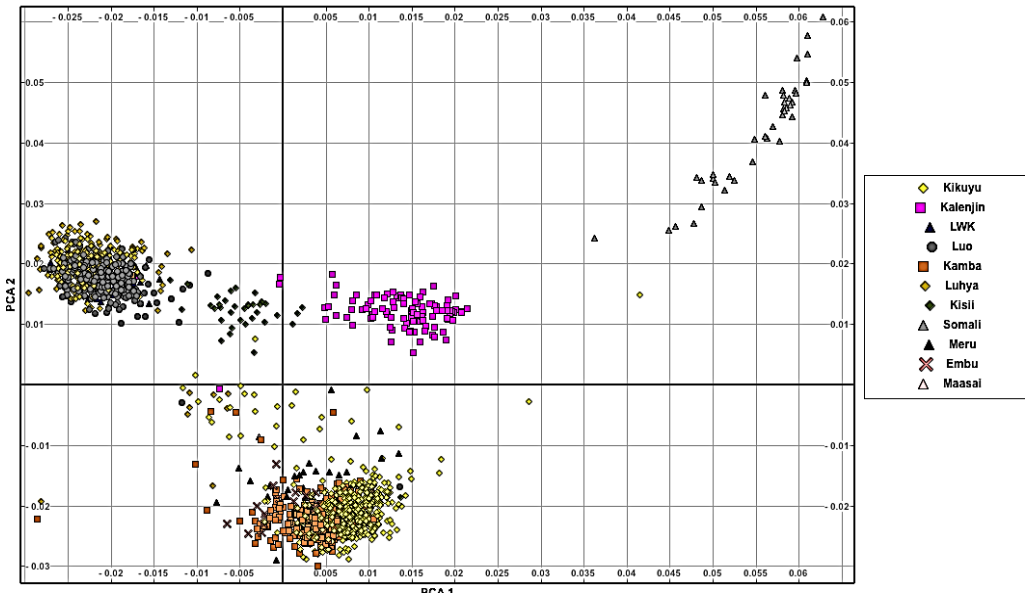


Figure 6. Principal Component Analysis of H3Africa AWI-Gen dataset (A) PCA representing populations from South Africa, Ghana, Burkina Faso and Kenya. The first PC separated East, West and Southern African populations. The second PC revealed that there was significant variation within the Kenyan population, as seen in pink. (B) PCA showing only

the East African populations from the AWI-Gen study. There is clear separation between ethnic groups belonging to the Afro-Asiatic (eg.) Somali), Nilo-Saharan (eg.) Kalenjin) and Niger-Congo (eg.) Luhya) language families. It is noteworthy that there is considerable overlap between certain groups. Additionally, some individuals who identify to one group localize to another group or to a section of the plot not covered by any of the represented groups. The principal component values shown in this plot will be used for this study.

3.2. Genetic data-based filtering of ethnolinguistic groups

One of the main aims of this project is to determine how misclassification due to discrepancies between self-reported identity and genetic clustering impact on the allele frequency of malaria- and trypanosomiasis-associated variants.

To achieve this, the distribution of ethnolinguistic group (listed in Table 1) in the PC space will be studied. Based on the distribution of the first 10 PCs, centroids for each group will be calculated using a custom R script. Individuals meeting the following criteria will be excluded from their self-reported ethnic groups:

- For each group, individuals that are closer, in terms of Euclidean distance, to the centroid of any group other than the group of their self-reported affiliation, will be considered as misclassified.
- Additionally, individuals found to deviate by more than 5 standard deviations from the group mean of any of the 10 PCs will be excluded.

The number of excluded individuals in each ethnic group will be used as proxy for determining the correlation between self-reported ethnicity and genetic clustering.

Additionally, the Uniform Manifold Approximation and Projection (UMAP) approach will be used to compress the results from the first 10 PCs into 2 PCs for further dimensionality reduction (28). A centroid-based approach, as described above, will be applied to the dataset to validate the classification of individuals into group and outlier.

1.1. Allele Frequency Estimation

With the original dataset organized into ethnicity-defined clusters, the allele frequencies of malaria- and trypanosomiasis-associated SNPs will be determined for each group (Supplementary Tables 1 and 2). To determine the allele frequency of any variant, the number of times that allele appears in the population is divided by the total number of alleles/ individuals in that population. PLINK will be used to estimate allele frequencies. To assess the standard errors in these estimates, for each ethnolinguistic group, 5 random subsamples of up to 50 participants will be used. After genetic data-based filtering of the groups, i.e. the removal of misclassified individuals, the allele frequencies will be re-estimated and compared with the original frequencies.

1.2. Statistical Analysis

Sample size calculations was performed using the PAWR package in R. The estimates show a sample size 50 is sufficiently powered to test for allele frequency differences for the majority of disease associated variants (Please see Supplementary Note 1 for details).

The statistical significance (*P*-values) of the differences in allele frequencies of malaria- and trypanosomiasis-associated SNPs between the various ethnolinguistic groups will be assessed using chi-squared and Fischer tests implemented in PLINK. A similar comparison will be performed for the frequency difference of these SNPs in each ethnolinguistic group, pre- and post-removal of misclassified individuals.

1.3. Problems and challenges

There are large scale differences in the representation of different ethnolinguistic groups in the dataset. This led to the exclusion of some of the groups. Moreover, these differences imply that the robustness of allele frequency estimates might vary between these ethnolinguistic groups. As the study centre was in Nairobi, which as the capital is a melting pot of different groups, there may be excessive admixtures between groups, which may lead to the exclusion of considerable number of individuals and loss of statistical power for testing allele frequency differences. Finally, due to the nature of the study centre, only volunteers who could access the study centre were recruited, such that participants were selected by convenience.

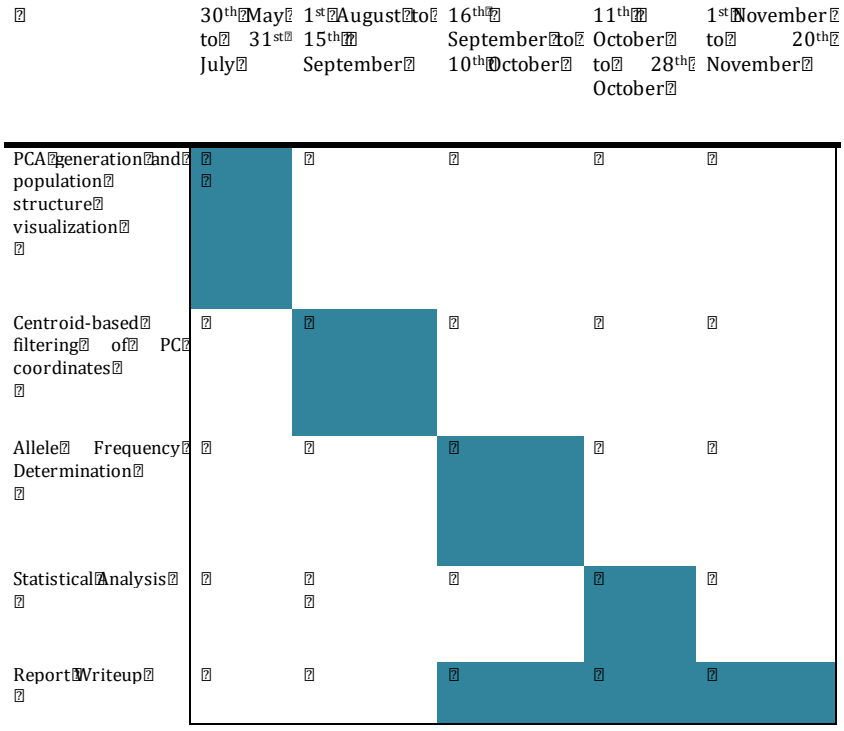
1.4. Ethical Considerations

This project is a retrospective study using summary-level data and so there are no major ethical considerations. In accordance with the Protection of Personal Information Act (POPIA), the data used in this study contains no identifying information, such as names, dates of birth, addresses, identification, medical or hospital numbers. A provisional ethics clearance certificate (Clearance Number: M220294) has been obtained, with Human Research Ethics Committee (Medical) (HREC) approval requiring evidence of protocol approval by the Division of Human Genetics in the Faculty of Health Science.

1.5. Funding

This is a retrospective study that does not require additional funding. A special thanks to the National Research Foundation (NRF) for sponsoring my tuition.

1. Proposed Project Timeline



References

1. Balaton-Chrimes S. Who are Kenya's 42(+) tribes? The census and the political utility of magical uncertainty. J East Afr Stud. 2021 Jan 2;15(1):43–62.

2. Gurdasani D, Carstensen T, Tekola-Ayele F, Pagani L, Tachmazidou I, Hatzikotoulas K, et al. The African Genome Variation Project shapes medical genetics in Africa. Nature. 2015 Jan;517(7534):327–32.

3. User S. Publications [Internet]. Kenya National Bureau of Statistics. [cited 2022 Apr 28]. Available from: https://www.knbs.or.ke/publications/

4. Atkinson EG, Dalvie S, Pichkar Y, Kalungi A, Majara L, Stevenson A, et al. Genetic structure correlates with ethnolinguistic diversity in eastern and southern Africa [Internet]. bioRxiv; 2021 [cited 2022 Feb 14]. p. 2021.05.19.444732. Available from: https://www.biorxiv.org/content/10.1101/2021.05.19.444732v1

5. Sengupta D, Choudhury A, Fortes-Lima C, Aron S, Whitelaw G, Bostoen K, et al. Genetic substructure and complex demographic history of South African Bantu speakers. Nat Commun [Internet]. 2021 [cited 2022 Feb 21];12(1). Available from: http://hdl.handle.net/1854/LU-8703344

6. Beltrame MH, Rubel MA, Tishkoff SA. Inferences of African evolutionary history from genomic data. Curr Opin Genet Dev. 2016 Dec;41:159–66.

7. Musuva A, Ejersa W, Kiptui R, Memusi D, Abwao E. The malaria testing and treatment landscape in Kenya: results from a nationally representative survey among the public and private sector in 2016. Malar J. 2017 Dec 21;16:494.

8. Kariuki SN, Williams TN. Human genetics and malaria resistance. Hum Genet. 2020;139(6):801–11.

9. Ndila CM, Uyoga S, Macharia AW, Nyutu G, Peshu N, Ojal J, et al. Human candidate gene polymorphisms and risk of severe malaria in children in Kilifi, Kenya: a case-control association study. Lancet Haematol. 2018 Aug 1;5(8):e333–45.

10. Greene JA, Moormann AM, Vulule J, Bockarie MJ, Zimmerman PA, Kazura JW. Toll-like receptor polymorphisms in malaria-endemic populations. Malar J. 2009 Mar 24;8:50.

11. Kulohoma BW, Wamwenje SAO, Wangwe II, Masila N, Mirieri CK, Wambua L. Prevalence of trypanosomes associated with drug resistance in Shimba Hills, Kwale County, Kenya. BMC Res Notes. 2020 Apr 29;13(1):234.

12. Watson HJC. The epidemiology of human sleeping sickness in the Lambwe Valley, South Nyanza, Kenya. Bull World Health Organ. 1972;47(6):719–26.

13. Kirchhoff LV. American Trypanosomiasis (Chagas' Disease) – A Tropical Disease Now in the United States. N Engl J Med. 1993 Aug 26;329(9):639–44.

14. African Trypanosomiasis in Kenya - Watch - Level 1, Practice Usual Precautions - Travel Health Notices | Travelers' Health | CDC [Internet]. [cited 2022 Feb 19]. Available from: https://wwwnc.cdc.gov/travel/notices/watch/african-trypanosomiasis-in-kenya

15. Nielsen MJ, Petersen SV, Jacobsen C, Oxvig C, Rees D, Møller HJ, et al. Haptoglobin-related protein is a high-affinity hemoglobin-binding plasma protein. Blood. 2006 Oct 15;108(8):2846–9.

16. Smith AB, Esko JD, Hajduk SL. Killing of Trypanosomes by the Human Haptoglobin-Related Protein. Science. 1995 Apr 14;268(5208):284–6.

17. Ramsay M, Crowther N, Tambo E, Agongo G, Baloyi V, Dikotope S, et al. H3Africa AWI-Gen Collaborative Centre: a resource to study the interplay between genomic and environmental risk factors for cardiometabolic diseases in four sub-Saharan African countries. Glob Health Epidemiol Genomics. 2016;1:e20.

18. User S. Publications [Internet]. Kenya National Bureau of Statistics. [cited 2022 Apr 28]. Available from: https://www.knbs.or.ke/publications/

19. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007 Sep;81(3):559–75.

20. Ali SA, Soo C, Agongo G, Alberts M, Amenga-Etego L, Boua RP, et al. Genomic and environmental risk factors for cardiometabolic diseases in Africa: methods used for Phase 1 of the AWI-Gen population cross-sectional study. Glob Health Action. 2018 Sep 27;11(Suppl 2):1507133.

21. Choudhury A, Aron S, Botigué LR, Sengupta D, Botha G, Bensellak T, et al. High-depth African genomes inform human migration and health. Nature. 2020 Oct;586(7831):741–8.

22. Largest Ethnic Groups In Kenya [Internet]. WorldAtlas. 2019 [cited 2022 Feb 13]. Available from: https://www.worldatlas.com/articles/largest-ethnic-groups-in-kenya.html

23. Kenya [Internet]. PMI. [cited 2022 Feb 14]. Available from: https://www.pmi.gov/where-we-work/kenya/

24. Prevention CC for DC and. CDC - Malaria - Malaria Worldwide - CDC's Global Malaria Activities - Kenya [Internet]. 2019 [cited 2022 Feb 14]. Available from: https://www.cdc.gov/malaria/malaria_worldwide/cdc_activities/kenya.html

25. Maxfield L, Bermudez R. Trypanosomiasis. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2022 [cited 2022 Feb 19]. Available from: http://www.ncbi.nlm.nih.gov/books/NBK535413/

26. Maeda N, McEvoy SM, Harris HF, Huisman TH, Smithies O. Polymorphisms in the human haptoglobin gene cluster: chromosomes with multiple haptoglobin-related (Hpr) genes. Proc Natl Acad Sci. 1986 Oct 1;83(19):7395–9.

27. Hardwick RJ, Ménard A, Sironi M, Milet J, Garcia A, Sese C, et al. Haptoglobin (HP) and Haptoglobin-related protein (HPR) copy number variation, natural selection, and trypanosomiasis. Hum Genet. 2014 Jan;133(1):69–83.

28. McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. ArXiv180203426 Cs Stat [Internet]. 2020 Sep 17 [cited 2022 Apr 18]; Available from: http://arxiv.org/abs/1802.03426

4. Appendix

4.1. Supplementary Table 1. Malaria-associated loci and their phenotypes. The data presented in this table was obtained from ClinVar, OMIM # 611162 and OMIM # 609148. ClinVar ID's were used where a dbSNP ID was not available.

Locus	dbSNP ID	Indication Allele	Phenotype
ACKR1	rs2814778	G	Protection against *Plasmodium vivax*
FCGR2A	rs1801274	T	Increased susceptibility to severe malaria
FCGR2B	rs1050501	C	Resistance to malaria
CR1	rs2274567	G	Resistance to severe malaria
GYPC	ClinVar ID: 17725	Deletion (p.Glu36_Ala63del)	Resistance to malaria
CISH	rs414171	T	Increased susceptibility to malaria
TNF	ClinVar ID: 12387 rs1800629	A A	Increased susceptibility to cerebral malaria
CD36	ClinVar ID: 13537	Microsatellite	Reduced risk of/resistance to cerebral malaria
CD36	rs766920034 rs148051111 ClinVar ID: 13538	Duplication (p.Gln262fs) A Deletion (p.Pro1444fs)	Increased susceptibility to cerebral malaria
HBB	rs334 rs33930165	T A	Resistance to malaria
TIRAP	rs8177374	T	Protection against malaria
NOS2A	rs9282799 ClinVar ID: 14014	T C	Resistance to malaria
SLC4A1	rs121912751 rs121912745 rs769664228	A T Deletion (p.Ala400_Ala408del)	Resistance to malaria
ICAM1	rs5491 ClinVar ID: 1675090 rs5030400 ClinVar ID: 1675089	T A T T	Increased susceptibility to cerebral malaria.
G6PD	rs72554664 ClinVar ID: 1206418 rs5030869 rs137852327 rs137852318	A T A A C T	G6PD deficiency leading to increased susceptibility to malaria

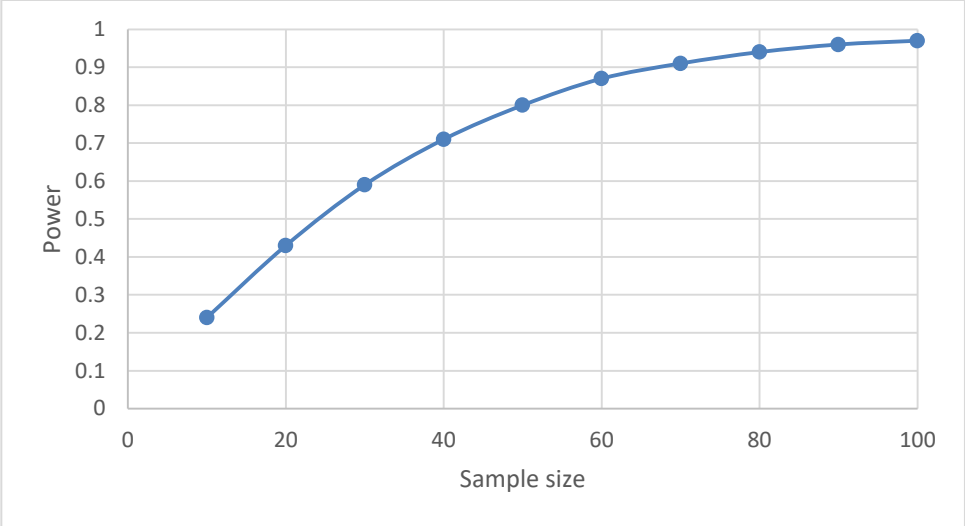
	rs5030868 rs137852314 rs1050829	A G	
NCR3	rs2736191	C	Increased susceptibility to mild malaria

4.2. Supplementary Table 2. Trypanosomiasis-associated variants to be utilized in allele frequency determination. These variants were obtained from Ndila et al. (2018). The disease allele represents the mutated nucleotide.

Gene	SNP ID	Disease Allele
APOL1	rs60910145	G
APOL1	rs73885319	G
APOL1	rs143830837	-

5.3 Supplementary Note 1

Sample Size Calculation: The optimal sample size calculations for testing allele frequency differences between two populations depends concomitantly on two components - level of difference in frequencies between the two groups and the rarity of the individual allele being compared. Alleles which occur at lower frequencies (say MAF=0.01) would require much larger sample sizes compared to alleles that occur at higher frequencies (say MAF=0.10). Similarly, larger sample sizes would also be required to test for small-scale differences in allele frequencies. For our calculations we assumed these two parameters to be incorporated into the effect size estimate. We used the R package PWR to calculated the power of Chi-squared test at different effect size and sample size ranges.



Supplementary Figure 1: Statistical power for Chi-squared test-based discrimination of allele frequencies between two populations. Thw sample size of 30 has 60% power and a sample size of 80% power to detect effect size differences over 0.40 between two populations.

1.1. Ethics Clearance Certificate

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)

Office of the Deputy Vice-Chancellor (Research and Innovation)

TO:
Mr YA Wolberg
Faculty of Health Sciences
Sydney Brenner Institute for Molecular Bioscience
and Division of Human Genetics
Medical School
University

E-mail: yoniwolberg@duck.com

CC:
Supervisor: Professor A Choudhury and Dr D Sengupta
<Anayo.Choudhury@wits.ac.za>
and <HREC-Medical Research Office@wits.ac.za>

FROM:
Mr Iain Burns
Human Research Ethics Committee (Medical)
Tel: 011 717 1252

E-mail: Iain.Burns@wits.ac.za

UNIVERSITY OF THE WITWATERSRAND
OFFICE OF THE DEPUTY VICE CHANCELLOR RESEARCH
2022 -04- 20
RESEARCH OFFICE HREC (Medical) RECEIVED
JOHANNESBURG

DATE: 2022/03/16

REF: R14/49

PROTOCOL NO: **M220294** (This is your ethics application reference number. Please quote it in all enquiries, oral or written, relating to this study.)

PROJECT TITLE: *Genetic clustering, self-reported ancestry and language in an urban Kenyan cohort*

Please find attached the Clearance Certificate for the above project. I hope it goes well and that an article in a recognized publication comes out of it. This will reflect well on your professional standing and contribute to Government funding of the University.

MSWorks2000/Iain0007/Clearscan.wps

UNIVERSITY OF THE WITWATERSRAND JOHANNESBURG

R49 Mr YA Wolberg

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)
CLEARANCE CERTIFICATE NO. M220294

NAME: (Principal Investigator) Mr YA Wolberg

DEPARTMENT: Faculty of Health Sciences
Sydney Brenner Institute for Molecular Bioscience
and Division of Human Genetics
Medical School
University

PROJECT TITLE: *Genetic clustering, self-reported ancestry and language in an urban Kenyan cohort*

UNIVERSITY OF THE WITWATERSRAND
OFFICE OF THE DEPUTY VICE CHANCELLOR: RESEARCH
2022 -04- 20
RESEARCH OFFICE
HREC (Medical)
RECEIVED
JOHANNESBURG

DATE CONSIDERED: Ad hoc

DECISION: Approved conditionally

CONDITIONS: Sub-study under M170880
The HREC requires evidence of Faculty protocol approval

NOTE: If contact information regarding student study participants is required, please contact the Registrar's office - <Nicoleen.Potgieter@wits.ac.za>

SUPERVISOR: Professor A Choudhury and Dr D Sengupta

APPROVED BY: CB Penny
Dr CB Penny, Chairperson, HREC (Medical)

DATE OF APPROVAL: 2022/03/16

This Clearance Certificate is valid for 5 years from the date of approval. An extension may be applied for.

DECLARATION OF INVESTIGATORS

To be completed in duplicate and **ONE COPY** returned to the Research Office secretariat on the 3rd floor, Phillip Tobias Building, Parktown, University of the Witwatersrand, Johannesburg.

I/we fully understand the conditions under which I am/we are authorized to carry out the above-mentioned research and I/we undertake to ensure compliance with these conditions. Should any departure be contemplated from the research protocol as approved, I/we undertake to submit details to the Committee. **I agree to submit a yearly progress report**. When a funder requires annual re-certification, the application date will be one year after the date when the study was initially reviewed. In this case, the study was initially reviewed in **February** and therefore reports and re-certification will be due in the month of **February** each year. Unreported changes to the study may invalidate the clearance given by the HREC (Medical).

Signature of Principal Investigator

19/04/2022
Date

Appendix B

Ethics Clearance Certificate

UNIVERSITY OF THE WITWATERSRAND JOHANNESBURG

R49 Mr YA Wolberg

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)
CLEARANCE CERTIFICATE NO. M220294

NAME: (Principal Investigator)	Mr YA Wolberg
DEPARTMENT:	Faculty of Health Sciences Sydney Brenner Institute for Molecular Bioscience and Division of Human Genetics Medical School University
PROJECT TITLE:	*Genetic clustering, self-reported ancestry and language in an urban Kenyan cohort*
DATE CONSIDERED:	Ad hoc
DECISION:	Approved unconditionally
CONDITIONS:	Sub-study under M170880 Prior condition (Faculty protocol approval) satisfied on 2022/06/20
NOTE:	If contact information regarding student study participants is required, please contact the Registrar's office - <Nicoleen.Potgieter@wits.ac.za>
SUPERVISOR:	Professor A Choudhury and Dr D Sengupta
APPROVED BY:	Dr CB Penny, Chairperson, HREC (Medical)
DATE OF APPROVAL:	2022/03/16

This Clearance Certificate is valid for 5 years from the date of approval. An extension may be applied for.

DECLARATION OF INVESTIGATORS

To be completed in duplicate and **ONE COPY** returned to the Research Office secretariat on the 3rd floor, Phillip Tobias Building, Parktown, University of the Witwatersrand, Johannesburg.

I/we fully understand the conditions under which I am/we are authorized to carry out the above-mentioned research and I/we undertake to ensure compliance with these conditions. Should any departure be contemplated from the research protocol as approved, I/we undertake to submit details to the Committee. **I agree to submit a yearly progress report**. When a funder requires annual re-certification, the application date will be one year after the date when the study was initially reviewed. In this case, the study was initially reviewed in **February** and therefore reports and re-certification will be due in the month of **February** each year. Unreported changes to the study may invalidate the clearance given by the HREC (Medical).

Signature of Principal Investigator _____ Date _____

Appendix C

Turn-it-in Report

KenWI-Gen Yoni.docx

ORIGINALITY REPORT

8%	7%	5%	4%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	wiredspace.wits.ac.za Internet Source	3%
2	Submitted to University of Strathclyde Student Paper	<1%
3	Submitted to University of Witwatersrand Student Paper	<1%
4	bmcpublichealth.biomedcentral.com Internet Source	<1%
5	Submitted to University of Auckland Student Paper	<1%
6	www.lolo.ac.uk Internet Source	<1%
7	Duijvesteijn, Naomi. "Sociable Swine: Prospects of Indirect Genetic Effects for the Improvement of Productivity, Welfare and Quality.", Wageningen University and Research, 2021 Publication	<1%
8	www.cesnet.co.za Internet Source	

		<1 %
9	www.ncbi.nlm.nih.gov Internet Source	<1 %
10	Khadka, Sabin, Godfrey D. Pearlson, Vince D. Calhoun, Jingyu Liu, Joel Gelernter, Katie L. Bessette, and Michael C. Stevens. "Multivariate Imaging Genetics Study of MRI Gray Matter Volume and SNPs Reveals Biological Pathways Correlated with Brain Structural Differences in Attention Deficit Hyperactivity Disorder", Frontiers in Psychiatry, 2016. Publication	<1 %
11	arizona.openrepository.com Internet Source	<1 %
12	www.science.gov Internet Source	<1 %
13	Submitted to Nelson Mandela Metropolitan University Student Paper	<1 %
14	Trypanosomes and Trypanosomiasis, 2014. Publication	<1 %
15	www.mla.com.au Internet Source	<1 %
	www.savetcon.co.za	

16	Internet Source	<1%
17	Palwende Romuald Boua, Cassandra Claire Soo, Cornelius Debpuur, Innocent Maposa et al. "Prevalence and socio-demographic correlates of tobacco and alcohol use in four sub-Saharan African countries: a cross-sectional study of middle-aged adults", BMC Public Health, 2021 Publication	<1%
18	Submitted to London School of Hygiene and Tropical Medicine Student Paper	<1%
19	scholar.sun.ac.za Internet Source	<1%
20	www.researchgate.net Internet Source	<1%
21	e-research.siam.edu Internet Source	<1%
22	ojphi.org Internet Source	<1%
23	www.frontiersin.org Internet Source	<1%
24	www.physoc.org Internet Source	<1%

25	www.nature.com Internet Source	<1%
26	www.support.illumina.com Internet Source	<1%
27	8ac81443-78e4-4641-81e2-82f522bac4c5.filesusr.com Internet Source	<1%
28	Submitted to University of Leicester Student Paper	<1%
29	Vinson, Benjamin T.. "CAD/CAM Laser Processing as a Method for Integrated Fabrication of Microphysiological Systems.", Tulane University School of Science and Engineering, 2020 Publication	<1%
30	dspace.mit.edu Internet Source	<1%
31	fjfsdata01prod.blob.core.windows.net Internet Source	<1%
32	hdl.handle.net Internet Source	<1%
33	impactaging.com Internet Source	<1%
34	livrepository.liverpool.ac.uk Internet Source	<1%

35	wellcomeopenresearch.org Internet Source	<1%
36	www.biorxiv.org Internet Source	<1%

Exclude quotes	On	Exclude matches	< 8 words
Exclude bibliography	On		

Appendix D

PLOS Genetics: Author Guidelines

Submission Guidelines

About the Journal

PLOS Genetics publishes original research that clearly demonstrates novelty, importance to a particular field, biological significance, and conclusions that are justified by the study.

Our aim is to make the editorial process rigorous and consistent, and to offer the best possible support to our authors throughout this process. Authors are encouraged to decide how best to present their ideas, results, and conclusions, but all research should be presented in a form that is readable to those in the field, easily understood by scientists outside of the immediate discipline, and comprehensible to readers whose first language is not English. The writing style should therefore be concise and accessible. Care should be taken to define abbreviations clearly and to use correct genetic and systematic nomenclature.

We strongly encourage authors to seek input from their co-authors and colleagues with different expertise when preparing their manuscript for submission to ensure that the style of writing, clarity of meaning, and spelling, punctuation, and grammar are at a very high level. A variety of style and writing guides are available, including *The Elements of Style* (New York: bartleby.com, 1999) and the Manuscript Preparation recommendations of the International Committee of Medical Journal Editors (ICMJE). Editors and/or reviewers may also make suggestions for how to achieve optimal quality and clarity of presentation, as well as potential cuts or additions that could strengthen the manuscript.

Related information for authors

- PLOS Writing Center
- Submission system
- Journal scope and publication criteria
- Getting started guide
- Guidelines for other article types
- Guidelines for revisions
- Publication fees

Style and Format

PLOS Genetics accepts initial submissions of manuscripts in a single PDF file which can include text and figures. If you are submitting a revised manuscript, upload separate files for your text, figures, and supporting information. The editor will then assess your submission.

When you first submit to the journal, providing you include all the necessary information needed for editorial assessment and review, we will not ask you to make any formatting changes. During resubmission, we may ask you to meet formatting requirements.

File format	Manuscript files can be in the following formats: DOC, DOCX, RTF or PDF. Microsoft Word documents should not be locked or protected. LaTeX manuscripts must be submitted as PDFs. Read the LaTeX guidelines.
Length	Manuscripts can be any length. There are no restrictions on word count, number of figures, or amount of supporting information. We encourage you to present and discuss your findings concisely.
Font	Use a standard font size and any standard font, except for the font named “Symbol”. To add sy to the manuscript, use the Insert → Symbol function in your word processor or paste in the appropriate Unicode character.
Headings	Limit manuscript sections and sub-sections to 3 heading levels. Make sure heading levels are c indicated in the manuscript text.
Layout and spacing	Manuscript text should be double-spaced. Do not format text in multiple columns.
Page and line numbers	Include page numbers and line numbers in the manuscript file. Use continuous line numbers (d restart the numbering on each page).
Tables	Insert tables immediately after the first paragraph in which they are cited.
Supporting Information	Upload Supporting Information (SI) files separately.
Footnotes	Footnotes are not permitted. If your manuscript contains footnotes, move the information into th main text or the reference list, depending on the content.
Language	Manuscripts must be submitted in English. You may submit translations of the manuscript or abstract as supporting information. Read the supporting information guidelines.
Abbreviations	Define abbreviations upon first appearance in the text. Do not use non-standard abbreviations unless they appear at least three times in the text. Keep abbreviations to a minimum.
Reference style	PLOS uses “Vancouver” style, as outlined in the ICMJE sample references. See reference formatting examples and additional instructions below.

Equations	We recommend using MathType for display and inline equations, as it will provide the most relia outcome. If this is not possible, Equation Editor or Microsoft's Insert→Equation function is acceptable. Avoid using MathType, Equation Editor, or the Insert→Equation function to insert single variable (e.g., "$a^2 + b^2 = c^2$"), Greek or other symbols (e.g., β, Δ, or ′ [prime]), or mathematical operators x, ≥, or ±) in running text. Wherever possible, insert single symbols as normal text with the corr Unicode (hex) values. Do not use MathType, Equation Editor, or the Insert→Equation function for only a portion of an equation. Rather, ensure that the entire equation is included. Equations should not contain a mi different equation tools. Avoid "hybrid" inline or display equations, in which part is text and part i MathType, or part is MathType and part is Equation Editor.
Nomenclature	Use correct and established nomenclature wherever possible.

Units of measurement	Use SI units. If you do not use these exclusively, provide the SI value in parentheses after each value. Read more about SI units.
Drugs	Provide the Recommended International Non-Proprietary Name (rINN).
Species names	Write in italics (e.g., *Homo sapiens*). Write out in full the genus and species, both in the title of the manuscript and at the first mention of an organism in a paper. After first mention, the first letter of the genus name followed by the full species name may be used (e.g., *H. sapiens*).
Genes, mutations, genotypes, and alleles	Write in italics. Use the recommended name by consulting the appropriate genetic nomenclature database (e.g., HUGO for human genes). It is sometimes advisable to indicate the synonyms for the gene the first time it appears in the text. Gene prefixes such as those used for oncogenes or cellular localization should be shown in roman typeface (e.g., v-fes, c-MYC).
Allergens	The systematic allergen nomenclature of the World Health Organization/International Union of Immunological Societies (WHO/IUIS) Allergen Nomenclature Sub-committee should be used for manuscripts that include the description or use of allergenic proteins. For manuscripts describing new allergens, the systematic name of the allergen should be approved by the WHO/IUIS Allergen Nomenclature Sub-Committee prior to manuscript publication. Examples of the systematic allergen nomenclature can be found at the WHO/IUIS Allergen Nomenclature site.

Copyediting manuscripts

Prior to submission, authors who believe their manuscripts would benefit from in-depth professional copyediting are encouraged to use language-editing and copyediting services. Obtaining this service is the responsibility of the author and should be done before initial

submission. These services can be found on the web using search terms like “scientific editing service” or “manuscript editing service”.

Note that if your manuscript is accepted, PLOS will not perform a detailed copyediting step. Therefore, please carefully review your manuscript, paying special attention to spelling, punctuation, and grammar, as well as scientific content.

Manuscript Organization

Most manuscripts should be organized as follows. Instructions for each element appear below.

- Title
- Authors
- Affiliations
- Abstract
- Author Summary
- Introduction
- Results
- Discussion
- Materials and Methods (also called Methods or Models)
- Acknowledgments
- References
- Supporting information captions.

Uniformity in format facilitates the experience of readers and users of the journal. To provide flexibility, however, authors are also able to include the Materials and Methods section before the Results section or before the Discussion section. Please clarify the reasons for including your Materials and Methods section before the Results or Discussion sections in your cover letter when submitting your manuscript files.

Please also note that the Results and Discussion can be combined into one Results/Discussion section.

Manuscripts submitted as Methods articles may be organized as follows:

- Title
- Authors

- Affiliations
- Abstract
- Author Summary
- Introduction
- Description of the Method
- Verification and Comparison
- Applications
- Discussion
- Acknowledgements
- References
- Supporting Information captions

To use this alternative structure, please choose ‘Methods’ when selecting an article type during the submission process.

 **Ready to format your provisionally accepted manuscript?**

Refer to our downloadable sample files to ensure that your submission meets our formatting requirements:

- Download sample title, author list, and affiliations page (PDF)
- Download sample manuscript body (PDF)

Viewing Figures and Supporting Information in the compiled submission PDF
The compiled submission PDF includes low-resolution preview images of the figures after the reference list. The function of these previews is to allow you to download the entire submission as quickly as possible. Click the link at the top of each preview page to download a high-resolution version of each figure. Links to download Supporting Information files are also available after the reference list.

Parts of a Submission

Title

Include a full title and a short title for the manuscript.

Title	Length	Guidelines	Examples
Full title	200 characters	Specific, descriptive, concise, and comprehensible to readers outside the field	Impact of cigarette smoke exposure on innate immunity: A *Caenorhabditis elegans* model Solar drinking water disinfection (SODIS) to reduce childhood diarrhoea in rural Bolivia: A cluster-randomized, controlled trial
Short title	70 characters	State the topic of the study	Cigarette smoke exposure and innate immunity SODIS and childhood diarrhoea

Titles should be written in sentence case (only the first word of the text, proper nouns, and genus names are capitalized). Avoid specialist abbreviations if possible. For clinical trials, systematic reviews, or meta-analyses, the subtitle should include the study design.

Author list

Authorship requirements

All authors must meet the criteria for authorship as outlined in the authorship policy. Those who contributed to the work but do not meet the criteria for authorship can be mentioned in the Acknowledgments. Read more about Acknowledgments.

The corresponding author must provide an ORCID iD at the time of submission by entering it in the user profile in the submission system. Read more about ORCID.

Author names and affiliations

Enter author names on the title page of the manuscript and in the online submission system.

On the title page, write author names in the following order:

- First name (or initials, if used)
- Middle name (or initials, if used)
- Last name (surname, family name)

Each author on the list must have an affiliation. The affiliation includes department, university, or organizational affiliation and its location, including city, state/province (if applicable), and country. Authors have the option to include a current address in addition to the address of their affiliation at the time of the study. The current address should be listed in the byline and clearly labeled “current address.” At a minimum, the address must include the author’s current institution, city, and country.

If an author has multiple affiliations, enter all affiliations on the title page only. In the submission system, enter only the preferred or primary affiliation. Author affiliations will be listed in the typeset PDF article in the same order that authors are listed in the submission.

Author names will be published exactly as they appear in the manuscript file. Please double-check the information carefully to make sure it is correct.

Corresponding author

The submitting author is automatically designated as the corresponding author in the submission system. The corresponding author is the primary contact for the journal office and the only author able to view or change the manuscript while it is under editorial consideration.

The corresponding author role may be transferred to another coauthor. However, note that transferring the corresponding author role also transfers access to the manuscript. (To designate a new corresponding author while the manuscript is still under consideration, watch the video tutorial below.)

Only one corresponding author can be designated in the submission system, but this does not restrict the number of corresponding authors that may be listed on the article in the event of publication. Whoever is designated as a corresponding author on the title page of the manuscript file will be listed as such upon publication. Include an email address for each corresponding author listed on the title page of the manuscript.

 **How to select a new corresponding author in Editorial Manager**

Consortia and group authorship

If a manuscript is submitted on behalf of a consortium or group, include its name in the manuscript byline. Do not add it to the author list in the submission system. You may include the full list of members in the Acknowledgments or in a supporting information file.

PubMed only indexes individual consortium or group author members listed in the article byline. If included, these individuals must qualify for authorship according to our criteria.

Read the group authorship policy.

Author contributions

Provide at minimum one contribution for each author in the submission system. Use the CRediT taxonomy to describe each contribution. Read the policy and the full list of roles.

Contributions will be published with the final article, and they should accurately reflect contributions to the work. The submitting author is responsible for completing this information at submission, and we expect that all authors will have reviewed, discussed, and agreed to their individual contributions ahead of this time.

PLOS Genetics will contact all authors by email at submission to ensure that they are aware of the submission.

Cover letter

Upload a cover letter as a separate file in the online system.

The cover letter should address the following questions:

- Why is this manuscript suitable for publication in *PLOS Genetics*?
- Why will your study inspire other members of your field, and how will it drive research forward?

You may recommend a suitable Associate Editor to handle your submission; however, the editors reserve the right to contact an alternative—either from the board or a guest editor.

The cover letter will only be available to the editor and the journal staff.

Title page

The title, authors, and affiliations should all be included on a title page as the first page of the manuscript file.

 Download our sample title, author list, and affiliations page (PDF)

Abstract

The Abstract comes after the title page in the manuscript file. The abstract text is also entered in a separate field in the submission system.

The Abstract should be succinct; it must not exceed 300 words. Authors should mention the techniques used without going into methodological detail and should summarize the most important results.

While the Abstract is conceptually divided into three sections (Background, Methodology/Principal Findings, and Conclusions/Significance), do not apply these distinct headings to the Abstract within the article file.

Do not include any citations. Avoid specialist abbreviations.

Author Summary

We ask that all authors of research articles include a 150-200 word non-technical summary of the work as part of the manuscript to immediately follow the abstract. This text is subject to editorial change, should be written in the first-person voice, and should be distinct from the scientific abstract.

Aim to highlight where your work fits within a broader context; present the significance or possible implications of your work simply and objectively; and avoid the use of acronyms and complex terminology wherever possible. The goal is to make your findings accessible to a wide audience that includes both scientists and non-scientists.

Authors may benefit from consulting with a science writer or press officer to ensure they effectively communicate their findings to a general audience.

Example Author Summaries

Complex Seizure Disorder Caused by *Brunol4* Deficiency in Mice

Genome-Wide Association Scan Shows Genetic Variants in the *FTO* Gene Are Associated with Obesity-Related Traits

A Mutation in the Myostatin Gene Increases Muscle Mass and Enhances Racing Performance in Heterozygote Dogs

Introduction

The introduction should put the focus of the manuscript into a broader context. As you compose the Introduction, think of readers who are not experts in this field. Include a brief review of the key literature. If there are relevant controversies or disagreements in the field, they should be mentioned so that a non-expert reader can delve into these issues further. The Introduction should conclude with a brief statement of the overall aim of the experiments and a comment about whether that aim was achieved.

Results

The Results section should provide details of all of the experiments that are required to support the conclusions of the paper. There is no specific word limit for this section, but details of experiments that are peripheral to the main thrust of the article and that detract from the focus of the article should not be included. The section may be divided into subsections, each with a concise subheading. The section should be written in the past tense.

PLOS journals require authors to make all data underlying the findings described in their manuscript fully available without restriction, with rare exception. When submitting a manuscript online, authors must provide a *Data Availability Statement* describing compliance with PLOS's policy.

Large data sets, including raw data, may be deposited in an appropriate public repository. See our list of recommended repositories.

For smaller data sets and certain data types, authors may provide their data within supporting information files accompanying the manuscript. Authors should take care to maximize the accessibility and reusability of the data by selecting a file format from which data can be

efficiently extracted (for example, spreadsheets or flat files should be provided rather than PDFs when providing tabulated data).

For more information on how best to provide data, read our policy on data availability. PLOS does not accept references to "data not shown."

Discussion

The Discussion should spell out the major conclusions of the work along with some explanation or speculation on the significance of these conclusions. How do the conclusions affect the existing assumptions and models in the field? How can future research build on these observations? What are the key experiments that must be done?

The Discussion should be concise and tightly argued.

The Results and Discussion may be combined into one section, if desired.

Materials and Methods

The Materials and Methods should provide enough detail to reproduce the findings. Submit detailed protocols for newer or less established methods. Well-established protocols may be referenced.

Details of algorithms and protocol documents for clinical trials, observational studies, and other **non-laboratory** investigations may be uploaded as supporting information. These are not included in the typeset manuscript, but are downloadable and fully searchable from the HTML version of the article. Read the supporting information guidelines for formatting instructions.

We recommend and encourage you to deposit **laboratory protocols** in protocols.io, where protocols can be assigned their own persistent digital object identifiers (DOIs).

To include a link to a protocol in your article:

1. Describe your step-by-step protocol on protocols.io
2. Select **Get DOI** to issue your protocol a persistent digital object identifier (DOI)
3. Include the DOI link in the Methods section of your manuscript using the following format provided by protocols.io: http://dx.doi.org/10.17504/protocols.io.[PROTOCOL DOI]

At this stage, your protocol is only visible to those with the link. This allows editors and reviewers to consult your protocol when evaluating the manuscript. You can make your protocols public at any time by selecting **Publish** on the protocols.io site. Any referenced protocol(s) will automatically be made public when your article is published.

PLOS ONE offers an option for publishing peer-reviewed Lab Protocol articles, which describe protocols hosted on protocols.io articles. Read more information on Lab Protocol articles.

Consult our reporting guidelines, and include an ethics statement in the Materials and Methods section when reporting results from human subjects research and animal research.

Acknowledgments

Those who contributed to the work but do not meet our authorship criteria should be listed in the Acknowledgments with a description of the contribution.

Authors are responsible for ensuring that anyone named in the Acknowledgments agrees to be named.

PLOS journals publicly acknowledge the indispensable efforts of our editors and reviewers on an annual basis. To ensure equitable recognition and avoid any appearance of partiality, do not include editors or peer reviewers—named or unnamed—in the Acknowledgments.

Do not include funding sources in the Acknowledgments or anywhere else in the manuscript file. Funding information should only be entered in the financial disclosure section of the submission system.

References

Any and all available works can be cited in the reference list. Acceptable sources include:

- Published or accepted manuscripts
- Manuscripts on preprint servers, providing the manuscript has a citable DOI or arXiv URL.

Do not cite the following sources in the reference list:

- Unavailable and unpublished work, including manuscripts that have been submitted but not yet accepted (e.g., “unpublished work,” “data not shown”). Instead, include those data as supplementary material or deposit the data in a publicly available database.
- Personal communications (these should be supported by a letter from the relevant authors but not included in the reference list)
- Submitted research should not rely upon retracted research. You should avoid citing retracted articles unless you need to discuss retracted work to provide historical context for your submitted research. If it is necessary to discuss retracted work, state the article’s retracted status in your article’s text and reference list.

Ensure that your reference list includes full and current bibliography details for every cited work at the time of your article’s submission (and publication, if accepted). If cited work is corrected, retracted, or marked with an expression of concern before your article is published, and if you feel it is appropriate to cite the work even in light of the post-publication notice, include in your manuscript citations and full references for both the affected article and the post-publication notice. Email the journal office if you have questions.

References are listed at the end of the manuscript and numbered in the order that they appear in the text. In the text, cite the reference number in square brackets (e.g., “We used the techniques developed by our colleagues [19] to analyze the data”). PLOS uses the numbered citation (citation-sequence) method and first six authors, et al.

Do not include citations in abstracts.

Make sure the parts of the manuscript are in the correct order *before* ordering the citations.

Formatting references

Because all references will be linked electronically as much as possible to the papers they cite, proper formatting of references is crucial.

PLOS uses the reference style outlined by the International Committee of Medical Journal Editors (ICMJE), also referred to as the “Vancouver” style. Example formats are listed below. Additional examples are in the ICMJE sample references.

A reference management tool, EndNote, offers a current style file that can assist you with the formatting of your references. If you have problems with any reference management program, please contact the source company's technical support.

Journal name abbreviations should be those found in the National Center for Biotechnology Information (NCBI) databases.

Source	Format
Published articles	Hou WR, Hou YL, Wu GF, Song Y, Su XL, Sun B, et al. cDNA, genomic sequence cloning overexpression of ribosomal protein gene L9 (rpL9) of the giant panda (Ailuropoda melanoleuca). Genet Mol Res. 2011;10: 1576-1588. Devaraju P, Gulati R, Antony PT, Mithun CB, Negi VS. Susceptibility to SLE in South India Tamils may be influenced by genetic selection pressure on TLR2 and TLR9 genes. Mol Immunol. 2014 Nov 22. pii: S0161-5890(14)00313-7. doi: 10.1016/j.molimm.2014.11.005. Note: A DOI number for the full-text article is acceptable as an alternative to or in addition traditional volume and page numbers. When providing a DOI, adhere to the format in the example above with both the label and full DOI included at the end of the reference (doi: 10.1016/j.molimm.2014.11.005). Do not provide a shortened DOI or the URL.
Accepted, unpublished articles	Same as published articles, but substitute “Forthcoming” for page numbers or DOI.
Online articles	Huynen MMTE, Martens P, Hilderlink HBM. The health impacts of globalisation: a concept framework. Global Health. 2005;1: 14. Available from: http://www.globalizationandhealth.com/content/1/1/14

Source	Format
Books	Bates B. Bargaining for life: A social history of tuberculosis. 1st ed. Philadelphia: University Pennsylvania Press; 1992.
Book chapters	Hansen B. New York City epidemics and history for the public. In: Harden VA, Risse GB, editors. AIDS and the historian. Bethesda: National Institutes of Health; 1991. pp. 21-28.
Deposited articles (preprints, e-prints, or arXiv)	Krick T, Shub DA, Verstraete N, Ferreiro DU, Alonso LG, Shub M, et al. Amino acid metab conflicts with protein diversity. arXiv:1403.3301v1 [Preprint]. 2014 [cited 2014 March 17]. Available from: https://128.84.21.199/abs/1403.3301v1 Kording KP, Mensh B. Ten simple rules for structuring papers. BioRxiv [Preprint]. 2016 bio 088278 [posted 2016 Nov 28; revised 2016 Dec 14; revised 2016 Dec 15; cited 2017 Feb [12 p.]. Available from: https://www.biorxiv.org/content/10.1101/088278v5 doi: 10.1101/088
Published media (print or online newspapers and magazine articles)	Fountain H. For Already Vulnerable Penguins, Study Finds Climate Change Is Another Da The New York Times. 2014 Jan 29 [Cited 2014 March 17]. Available from: http://www.nytimes.com/2014/01/30/science/earth/climate-change-taking-toll-on-penguins-study-finds.html
New media (blogs, web sites, or other written works)	Allen L. Announcing PLOS Blogs. 2010 Sep 1 [cited 17 March 2014]. In: PLOS Blogs [Inte San Francisco: PLOS 2006 - . [about 2 screens]. Available from: http://blogs.plos.org/plos/2010/09/announcing-plos-blogs/.
Masters' theses or doctoral dissertations	Wells A. Exploring the development of the independent, electronic, scholarly journal. M.Sc Thesis, The University of Sheffield. 1999. Available from: http://cumincad.scix.net/cgi-bin/works/Show?2e09
Databases and repositories (Figshare, arXiv)	Roberts SB. QPX Genome Browser Feature Tracks; 2013 [cited 2013 Oct 5]. Database: figshare [Internet]. Available from: http://figshare.com/articles/QPX_Genome_Browser_Feature_Tracks/701214
Multimedia (videos, movies, or TV shows)	Hitchcock A, producer and director. Rear Window [Film]; 1954. Los Angeles: MGM.

Supporting information

Authors can submit essential supporting files and multimedia files along with their manuscripts. All supporting information will be subject to peer review. All file types can be submitted, but files must be smaller than 20 MB in size.

Authors may use almost any description as the item name for a supporting information file as long as it contains an “S” and number. For example, “S1 Appendix” and “S2 Appendix,” “S1 Table” and “S2 Table,” and so forth.

Supporting information files are published exactly as provided, and are not copyedited.

Supporting information captions

List supporting information captions at the end of the manuscript file. Do not submit captions in a separate file.

The file number and name are required in a caption, and we highly recommend including a one-line title as well. You may also include a legend in your caption, but it is not required.

Example caption

S1 Text. Title is strongly recommended. Legend is optional.

In-text citations

We recommend that you cite supporting information in the manuscript text, but this is not a requirement. If you cite supporting information in the text, citations do not need to be in numerical order.

Read the supporting information guidelines for more details about submitting supporting information and multimedia files.

Figures and Tables

Figure files

You can include figures in the main manuscript file at initial submission. If the manuscript reaches the revise stage, prepare and submit each figure as an individual file.

Cite figures in ascending numeric order at first appearance in the manuscript file.

For detailed instructions, read the guidelines for figures.

Figure Captions

If you are submitting a new or revised manuscript, embed figure captions within the manuscript text after their first mention, or group together at the end of the manuscript.

After editorial acceptance, insert captions in read order in the manuscript text, immediately following the paragraph where the figure is first cited. Don't include captions as part of the figure files themselves or submit them in a separate document.

At a minimum, include the following in your figure captions:

- A figure label with Arabic numerals, and "Figure" abbreviated to "Fig" (e.g. Fig 1, Fig 2, Fig 3, etc). Match the label of your figure with the name of the file uploaded at submission (e.g. a figure citation of "Fig 1" must refer to a figure file named "Fig1.tif").
- A concise, descriptive title

The caption may also include a legend as needed.

For detailed instructions, read the guidelines for figures.

Data presentation in graphs

To better illustrate your results and to improve the reader's understanding and interpretation of your data, we discourage the use of bar graphs and line plots for continuous data, particularly for studies with small sample sizes (n≤9 independent observations per group).

See this Perspective for more information about our position on data presentation in graphs:
Weissgerber TL, Milic NM, Winham SJ, Garovic VD (2015) Beyond Bar and Line Graphs: Time for a New Data Presentation Paradigm. PLoS Biol 13(4): e1002128. doi:10.1371/journal.pbio.1002128

- We recommend that continuous data is presented in such a way that the reader can evaluate the distribution of the individual data points. Scatterplots, boxplots, and histograms are more informative than bar or line charts.
- If there is a particular reason that bar or line graphs are preferred, these must include inferential error bars:
  SD - standard deviation of the population, or CI - confidence intervals.
- Paired or non-independent data should be presented to include information about whether changes are consistent across individuals.
- Figures presenting small sample size data (n≤9 independent observations per group) should show the full distribution of the data, as opposed to summary statistics.
- All numerical values used to generate graphs must be provided as per the PLOS Data Availability policy.

For more information about our Data Availability policy, see this Editorial:
Bloom T, Ganley E, Winker M (2014) Data Access for the Open Access Literature: PLOS's Data Policy. PLoS Biol 12(2): e1001797. https://doi.org/10.1371/journal.pbio.1001797

For *PLOS Genetics*-specific guidance, read:
Barsh GS, Cooper GM, Copenhaver GP, Gibson G, McCarthy MI, Tang H, et al. (2015) PLOS Genetics Data Sharing Policy: In Pursuit of Functional Utility. PLoS Genet 11(12): e1005716. https://doi.org/10.1371/journal.pgen.1005716

Tables

Cite tables in ascending numeric order upon first appearance in the manuscript file.

Place each table in your manuscript file directly after the paragraph in which it is first cited (read order). Do not submit your tables in separate files.

Tables require a label (e.g., “Table 1”) and brief descriptive title to be placed above the table. Place legends, footnotes, and other text below the table.

Read the guidelines for tables.

Data reporting

All data and related metadata underlying the findings reported in a submitted manuscript should be deposited in an appropriate public repository, unless already provided as part of the submitted article.

Read our policy on data availability.

Repositories may be either subject-specific (where these exist) and accept specific types of structured data, or generalist repositories that accept multiple data types. We recommend that authors select repositories appropriate to their field. Repositories may be subject-specific (e.g., GenBank for sequences and PDB for structures), general, or institutional, as long as DOIs or accession numbers are provided and the data are at least as open as CC BY. Authors are encouraged to select repositories that meet accepted criteria as trustworthy digital repositories, such as criteria of the Centre for Research Libraries or Data Seal of Approval. Large, international databases are more likely to persist than small, local ones.

See our list of recommended repositories.

To support data sharing and author compliance of the PLOS data policy, we have integrated our submission process with a select set of data repositories. The list is neither representative nor exhaustive of the suitable repositories available to authors. Current repository integration partners include Dryad and FlowRepository. Please contact data@plos.org to make recommendations for further partnerships.

Instructions for PLOS submissions with data deposited in an integration partner repository:

- Deposit data in the integrated repository of choice.
- Once deposition is final and complete, the repository will provide you with a dataset DOI (provisional) and private URL for reviewers to gain access to the data.
- Enter the given data DOI into the full Data Availability Statement, which is requested in the Additional Information section of the PLOS submission form. Then provide the URL passcode in the Attach Files section.

If you have any questions, please email us.

Accession numbers

All appropriate data sets, images, and information should be deposited in an appropriate public repository. See our list of recommended repositories.

Accession numbers (and version numbers, if appropriate) should be provided in the Data Availability Statement. Accession numbers or a citation to the DOI should also be provided when the data set is mentioned within the manuscript.

In some cases authors may not be able to obtain accession numbers of DOIs until the manuscript is accepted; in these cases, the authors must provide these numbers at acceptance. In all other cases, these numbers must be provided at full submission.

Identifiers

As much as possible, please provide accession numbers or identifiers for all entities such as genes, proteins, mutants, diseases, etc., for which there is an entry in a public database, for example:

- Ensembl
- Entrez Gene
- FlyBase
- InterPro
- Mouse Genome Database (MGD)
- Online Mendelian Inheritance in Man (OMIM)
- PubChem

Identifiers should be provided in parentheses after the entity on first use.

Research Resource Identifiers

Research Resource Identifiers are unique Resource Identifiers (RRIDs) for citing resources used in the research. We encourage authors to use RRIDs within the text of their papers to identify their model organisms, antibodies, or tools.

PLOS is a part of the Research Resource Identification Initiative (#RII), a cross-publisher effort to enable effective tracking of the utilization of particular research resources across all biomedical literature and promote reproducibility in science. RRIDs are found via the RRID portal.

If an RRID does not exist, you may create a new one as needed through the RRID Portal. The Research Resource Citation guidelines provide instructions on how to cite the resource in the manuscript.

Small and macromolecule crystal data

Manuscripts reporting new and unpublished three-dimensional structures must include sufficient supporting data and detailed descriptions of the methodologies used to allow the reproduction

and validation of the structures. All novel structures must have been deposited in a community endorsed database prior to submission (please see our list of recommended repositories).

Small molecule single crystal data

Authors reporting X-Ray crystallographic structures of small organic, metal-organic, and inorganic molecules must deposit their data with the Cambridge Crystallographic Data Centre (CCDC), the Inorganic Crystal Structure Database (ICSD), or similar community databases providing a recognized validation functionality. Authors are also required to include the relevant structure reference numbers within the main text (e.g. the CCDC ID number), as well as the crystallographic information files (.cif format) as Supplementary Information, along with the checkCIF validation reports that can be obtained via the International Union of Crystallography (IUCr).

Macromolecular structures

Authors reporting novel macromolecular structures must have deposited their data prior to submission with the Worldwide Protein Data Bank (wwPDB), the Biological Magnetic Resonance Data Bank (BMRB), the Electron Microscopy Data Bank (EMDB), or other community databases providing a recognized validation functionality. Authors must include the structure reference numbers within the main text and submit as Supplementary Information the official validation reports from these databases.

Striking image

You can upload a visually striking image alongside your submission, which we may use to showcase your article through PLOS' online channels. The Editors-in-Chief choose the monthly issue image from the striking images submitted with articles scheduled for publication.

Submission Criteria

- Choose an image that represents the article in a striking and eye-catching way.
- It can be derived from a figure or supporting information file from the paper, and it may be a cropped portion of an image or the entire image.
- Alternatively, you can create or source an image, as long as it adheres to our CC BY license.
- High resolution: between 300-600 dpi
- Single panel
- Ideally avoid added details like text, scale bars, and arrows.

How to Submit

1. Submit your striking image to the submission system using the file type "Striking Image".

2. Upload a separate file with corresponding caption.

If no striking image is uploaded, a member of the journal team will choose an appropriate image, which may be a figure from the submission or a separately sourced CC BY image.

Striking images should not contain potentially identifying images of people. Read our policy on identifying information.

The PLOS licenses and copyright policy also applies to striking images.

Additional Information Requested at Submission

Financial Disclosure Statement

This information should describe sources of funding that have supported the work. If your manuscript is published, your statement will appear in the Funding section of the article.

Include your statement in the Financial Disclosure section of the initial submission form.

The statement should include:

- Specific grant numbers
- Initials of authors who received each award
- URLs to sponsors' websites

Also state whether any sponsors or funders (other than the named authors) played any role in:

- Study design
- Data collection and analysis
- Decision to publish
- Preparation of the manuscript

If they had no role in the research, include this sentence: "The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript."

If the study was unfunded, include this sentence as the Financial Disclosure statement: "The author(s) received no specific funding for this work."

Read our policy on disclosure of funding sources.

Competing interests

The corresponding author is asked at submission to declare, on behalf of all authors, whether there are any financial, personal, or professional interests that could be construed to have influenced the work.

Any relevant competing interests of authors must be available to editors and reviewers during the review process and will be stated in published articles.

Read our policy on competing interests.

Related manuscripts

When submitting a manuscript, all authors are asked to indicate that they do not have a related or duplicate manuscript under consideration (or accepted) for publication elsewhere. If related work has been or will be submitted elsewhere or is in press elsewhere, then a copy must be uploaded with the article submitted to PLOS. Reviewers will be asked to comment on the overlap between related submissions.

Read our policies on related manuscripts.

Preprints

PLOS encourages authors to post preprints to accelerate the dissemination of research and support authors who wish to share their work early and receive feedback before formal peer review. Deposition of manuscripts with preprint servers does not impact consideration of the manuscript at any PLOS journal.

Authors posting on bioRxiv may submit direclty to PLOS journals through bioRxiv's direct transfer to journal service.

Authors submitting manuscripts in the life science to *PLOS Genetics* may opt-in to post their work on bioRxiv during the *PLOS Genetics* initial submission process.

Read more about preprints.

Learn how to post a preprint to bioRxiv during *PLOS Genetics* initial submission.

Reviewer and editor suggestions

We ask authors to suggest suitable editors and at least four potential reviewers when submitting their manuscript. Bear in mind any potential competing interests when making these suggestions. It is not appropriate to suggest recent collaborators or other researchers at your institution. See our policy on competing interests for more information.

Opposed reviewers

Authors may choose to request that an individual is excluded from the review process and not involved in their manuscript. When making these suggestions, please provide specific reasons why each person should not review your submission in each "Reason" box. The editorial team will respect these requests so long as this does not interfere with the objective and thorough assessment of the submission.

Selecting a Section

Authors are required to select a section upon submission to *PLOS Genetics*. This information is used to assign the relevant Senior Editor who can best evaluate each manuscript. Choose one of the following sections that most closely applies (listed topics are not exhaustive). Select 'General' if the other sections do not apply.

Cancer genetics

- Genomic and/or epigenomic analyses in humans or model organisms that provide insight into cancer pathogenesis and/or approaches to therapy
- Identification of heritable cancer-susceptibility genes and variants
- Genome stability/DNA damage analyses relevant to understanding mechanisms of tumorigenesis
- Cellular/molecular studies that expand understanding of cancer-related signaling pathways

Epigenetics

- Regulatory mechanisms controlling genome functions in normal development and in disease
- The interaction of DNA sequence variability with genome regulation
- The molecular and cellular mechanisms mediating the memory of past events

Evolution

- Studies of molecular or experimental evolution in any organism or system
- Theoretical and/or computational approaches that represent a broad advance in understanding evolutionary dynamics and/or mechanisms
- Genetic and/or genomic whole organism studies that provide insight into adaptation or speciation

Methods

- Computational ("dry bench") methods and/or molecular genetics ("wet bench")
- The methodological advance should be substantial and the new method should be important and likely to be adopted by many researchers
- Description of the method must be accompanied by its validation in controlled settings (e.g. simulations) and performance comparisons with existing relevant methods
- Application to an important genetics question and related biological discovery to prove the usefulness of the method is encouraged but not required

- Manuscript organization may be submitted as Methods articles, with a distinct organization to promote accessibility and utility of the method

Human Genetic Variation and Disease

- Genome-wide and Phenome-wide association studies (GWAS and PheWAS) that provide new insights on the genetic architecture of human complex traits and diseases
- Studies of humans, model organisms, or pathogenic organisms that provide insight into genotype-phenotype relationships at a molecular, organismal, and/or population level
- Functional characterization that elucidates novel mechanisms underlying genotype-phenotype association; experimental studies of genetic pathophysiology for human phenotypes
- Characterization of genetic variation in natural populations of humans and human pathogens, which shed light on the history or phenotypic variation of the population
- Statistical, computational or bioinformatic methods that have broad application in studying human genetics
- Particular interest in human studies that include diverse study groups or focus on understudied populations

Plant Genetics

- Genetic approaches that provide insights of broad signifcance into the biology of plants
- Studies of plant evolution and natural variation and the genetics of agricultural traits
- Genomic and epigenomic analyses that provide mechanistic insight into the molecular processes that mediate plant reproduction, development and responses to environment

Prokaryotic genetics

- Genetic approaches that provide insights of broad significance into the biology of bacteria, archaea, or phages
- Phenotypic and molecular variation including gene expression, motility, cell cycle regulation, chromosome structure and function, metabolism, biofilm formation
- Microbial interactions and microbiomes

General

Authors may choose the General section if the other sections do not apply. The Editors-in-Chief act as Senior Editors for this section.

Other Article Types

If you are submitting content other than a research article, read the guidelines for other article types.

You may be eligible for APC support

Many institutional partners globally have publishing agreements with PLOS to allow their corresponding authors to publish with reduced or no APCs. To determine if your corresponding author is eligible, please visit our institutional partners page to determine what kind of agreement your institution has with PLOS.

If your corresponding author is affiliated with a participating institution, they must follow the instructions below to demonstrate eligibility.