Tai *et al. Malaria Journal* (2022) 21:79
https://doi.org/10.1186/s12936-022-04104-x

Malaria Journal

RESEARCH

Open Access

Leveraging Mann–Whitney U test on large-scale genetic variation data for analysing malaria genetic markers

Kah Yee Tai[1], Jasbir Dhaliwal[1*] and Vinod Balasubramaniam[2]

Abstract

Background: The malaria risk analysis of multiple populations is crucial and of great importance whilst compressing limitations. However, the exponential growth in diversity and accumulation of genetic variation data obtained from malaria-infected patients through Genome-Wide Association Studies opens up unprecedented opportunities to explore the significant differences between genetic markers (risk factors), particularly in the resistance or susceptibility of populations to malaria risk. Thus, this study proposes using statistical tests to analyse large-scale genetic variation data, comprising 20,854 samples from 11 populations within three continents: Africa, Oceania, and Asia.

Methods: Even though statistical tests have been utilized to conduct case–control studies since the 1950s to link risk factors to a particular disease, several challenges faced, including the choice of data (ordinal vs. non-ordinal) and test (parametric vs. non-parametric). This study overcomes these challenges by adopting the Mann–Whitney U test to analyse large-scale genetic variation data; to explore the statistical significance of markers between populations; and to further identify the highly differentiated markers.

Results: The findings of this study revealed a significant difference in the genetic markers between populations ($p < 0.01$) in all the case groups and most control groups. However, for the highly differentiated genetic markers, a significant difference ($p < 0.01$) was present for most genetic markers with varying p-values between the populations in the case and control groups. Moreover, several genetic markers were observed to have very significant differences ($p < 0.001$) across all populations, while others exist between certain specific populations. Also, several genetic markers have no significant differences between populations.

Conclusions: These findings further support that the genetic markers contribute differently between populations towards malaria resistance or susceptibility, thus showing differences in the likelihood of malaria infection. In addition, this study demonstrated the robustness of the Mann–Whitney U test in analysing genetic markers in large-scale genetic variation data, thereby indicating an alternative method to explore genetic markers in other complex diseases. The findings hold great promise for genetic markers analysis, and the pipeline emphasized in this study can fully be reproduced to analyse new data.

Keywords: Malaria, Single nucleotide polymorphisms, Mann–Whitney U test, Descriptive statistics, Genetic markers

Background

Malaria is a life-threatening disease caused by a parasite transmitted to humans by an infected female *Anopheles* mosquito bite. However, as the parasites involved are highly adaptable to nature, it is tremendously challenging to control the outbreak of this disease [1]. Moreover, risk

*Correspondence: jasbir.dhaliwal@monash.edu
[1] School of Information Technology, Monash University Malaysia, Subang Jaya, Selangor, Malaysia
Full list of author information is available at the end of the article

prediction of this disease has proven challenging due to the combined effects of environmental and genetic factors. Thus, biological modelling research using genetic information for disease risk assessment has been supplemented by various approaches, including Genome-Wide Association Studies (GWAS).

GWAS is a popular approach that investigates associations between genetic information, in particular, specific Single Nucleotide Polymorphisms (SNPs), and disease. An SNP is the most common type of genetic variation of a disease and is henceforth a resistance or susceptibility marker. For example, a resistance marker can prevent the risk of developing the disease as well as reducing the severity of the symptoms. In contrast, a susceptibility marker increases the risk of developing the disease instead. Thus, SNPs can be used as genetic markers to represent disease-associated risk factors.

In malaria research, GWAS has successfully been applied in multiple malaria-endemic areas [2–11], where SNPs related to malaria resistance or susceptibility have been identified. The exponential growth in diversity and accumulation of SNP genotypes obtained from malaria-infected patients through GWAS such as Malaria Genomic Epidemiology Network (MalariaGEN) provides large-scale genetic variation data to explore the significant differences between genetic markers (risk factors) among populations. Thus, this study proposes using statistical tests to analyse the MalariaGEN data which comprising 20,854 samples from 11 populations within three continents: Africa, Oceania, and Asia.

A statistical test is a powerful tool widely used throughout the scientific research process to conclude from mass data, where it can be applied to study the relationship between risk factors and diseases [12]. For example, a statistical test can be used to explore the effect of exposure to risk factors between disease-infected patients (case) and healthy individuals (control). Note that statistical tests have been utilized to conduct case–control studies since the 1950s to link cigarette smoke to lung cancer [13]; and later to complex diseases such as breast cancer [14, 15], ischemic heart disease [16], type 2 diabetes [17] and asthma [18]. Thus, as risk factors play an important role in disease prediction and prevention, a statistical test can measure the statistical significance of the risk factors leading to diseases, and is the focus of the work here.

However, several challenges were faced when applying statistical tests to the work. The first major challenge is that researchers to date have studied the statistical significance of disease risk factors by applying the tests on ordinal clinical data, i.e., continuous variables and not genetic markers. This data includes demographic characteristics, lifestyle habits, physical measurements, medical records, family history of the related disease, and disease-related knowledge data. To overcome this challenge of non-ordinal data, the genetic risk scores were calculated from the genetic markers, i.e., SNPs. As these scores indicate the impact of genetic variations in populations, the genetic markers that contribute to malaria resistance or susceptibility between populations can be explored by adopting a statistical test. However, malaria is a complex disease involving various genetic markers from many different genes, which leads to the genetic basis of malaria resistance or susceptibility being complicated at multiple levels [19]. Therefore, this information was taken into consideration, and a statistical test was adopted to evaluate population associations with single locus genetic markers and multilocus genetic markers (by summing the genetic risk scores).

Choosing the correct statistical test is another challenge. There are two types of statistical tests: parametric test and non-parametric test. There has always been a dispute over the preferred test in medical research [20–22]. The main issue with parametric tests is that the results may be misleading if the normal distribution assumption is not met, leading to an erroneous conclusion [23]. Note that a parametric test can be applied to non-normally distributed data based on the central limit theorem. However, according to several studies [24–26], normally distributed data is an exception and not a rule in medical research. This is because real-world data usually follow a non-normal distribution [24], and by definition, ordinal data does not follow a normal distribution, which is also quite common in biomedical research [27]. A more than a decade-long study emphasized this point by analysing 630 studies from biomedical journals, and among them, non-parametric tests are more commonly applied in human studies [25]. Thus, descriptive statistics was first conducted to understand the characteristics of MalariaGEN data to obtain meaningful statistics in order to explore the genetic markers associated with malaria.

In malaria research, statistical tests have mainly been utilized to capture an individual's genes characteristics towards malaria resistance or susceptibility [28, 29]; and assess the consistency of expression profiles of genes between case and control [30]. To date, there is no research that uses statistical tests to analyse large-scale genetic variation data to explore significant differences of malaria genetic markers, particularly in the resistance or susceptibility of populations to malaria risk. Thus, it raises several research questions, including: (1) Are there significant differences in the likelihood of getting malaria between populations?; and (2) What genetic markers can be used to distinguish the population affected by malaria?. To answer these questions, the contributions of this paper are summarized as follows:

- Introduces how a statistical test can potentially be adopted to analyse genetic risk scores obtained from large-scale genetic variation data (non-ordinal data), i.e., SNPs genetic marker;
- Analyses statistical significance of malaria genetic markers between populations, and;
- Identifies highly differentiated genetic markers among populations.

Methods

Dataset and study population

The human GWAS data utilized in this study was generated from the MalariaGEN Consortial Project 1, entitled: "Genome-wide study of resistance to severe malaria in eleven populations". The study comprised genotype data of 20,854 individuals from 11 worldwide populations: 10,791 severe malaria-affected individuals and 10,063 control subjects. Table 1 details the sample size of each population. The structure of the consortial project has been described in [31], and the collaboration of each partner's studies and field sites was acknowledged on the MalariaGEN website http://www.malariagen.net/.

Candidate single nucleotide polymorphism

Through the review and analysis of 31 academic articles related to malaria research [3–11, 29, 32–52], a total of 122 SNPs were identified to be associated with malaria. However, of the 122 SNPs, 18 SNPs were excluded due to unreported effect size and unavailability in certain populations.

Data preprocessing

Thus, 104 SNPs were extracted from the study subjects to analyse their genetic markers (Additional file 1). All unparseable values in the data, i.e., data type and standard format errors, are converted to null representations.

The Single Nucleotide Polymorphism database, in collaboration with EMBL-EBI European Variation Archive, assigns a unique ID to human genetic variation data, including SNPs [53]. These IDs are called rsIDs and appear in the format rs##. On the other hand, kgpIDs are identifiers created by Illumina during sequencing. There were 32 kgpIDs mapped to rsIDs, and 37 samples without severe malaria subtypes information were also removed. The subtype indicates the severity of malaria, which further influences the treatment plan.

The existing genotype imputation software, such as IMPUTE2 [54] and Beagle [55], usually impute missing genotypes based on publicly available reference datasets, such as 1000 Genomes Project or HapMap 3. However, in this case, imputation needs to be more specific, i.e., based on population group and severe malaria subtypes, as this study analyses the malaria risk of several populations.

Thus, a python program was developed to impute any missing genotypes based on the population group and severe malaria subtypes from the human GWAS data used in this study. In order to do so, the program first groups individuals based on their countries and then by their severe malaria subtypes. After that, the program compares a total of six SNPs for each missing genotype, i.e., three SNPs before and after the missing loci, and then imputes the missing genotype with the most common genotype data.

The dataset contains the genotype data of 104 SNPs formed by two alleles of *A* and *a*, usually expressed as *AA*, *Aa*, and *aa*. The genetic risk score for each genotype data was calculated to analyse the association between population and genetic markers and is described in the following section.

Table 1 Analysed populations and samples

Population	Case	Control	Sample size
Burkina Faso	807	639	1446
Cameroon	693	778	1471
Gambia	2807	2786	5593
Ghana	422	342	764
Kenya	1944	1738	3682
Malawi	1590	1498	3088
Mali	475	394	869
Nigeria	288	131	419
Tanzania	485	494	979
Vietnam	860	868	1728
Papua New Guinea	420	395	815
Total			20,854

Sample size indicates the total number of individuals for each population

Genetic risk score

The genetic risk score refers to a number reflecting the severity of the risk caused by specific genetic markers. In this study, the genetic risk score was calculated based on the genotype profile of each individual. This profile represents the impact of genetic variation on individuals in each population.

The most common approach to calculating genetic risk scores is weighted genetic risk scores (wGRS). The wGRS is calculated by multiplying the number of risk alleles (0, 1, 2) by the estimated effect size reported for each variant [56]. The logistic regression association tests method is used to estimate the variant effect size, described in the association test summary statistics available on the MalariaGEN website. However, this approach only considers the risk alleles and the effect size of the variant,

which is not sufficient for malaria risk analysis in these aspects.

Some observations in the literature [57, 58] indicate that genotype patterns contribute to disease association, and extensive evidences have proven that sickle cell anemia traits can partially prevent malaria [19, 59–61]. The trait of sickle cell anaemia is caused by the recessive alleles in the haemoglobin gene. This means that an individual needs to have two copies of the recessive alleles—one from the mother and one from the father—to have this condition.

An individual tends to be resistant to the development of malaria if the two alleles are not identical (heterozygous). Conversely, an individual tends to be susceptible to the development of malaria if the two alleles are identical (homozygous). Therefore, the inclusion of genotype patterns is essential for differentiating genetic markers. Inspired by its importance, this study will include genotype frequency with wGRS to formulate more comprehensive genetic risk scores, namely wGRS + GF.

Genotype frequency indicates the relative frequency of a particular genotype in a population. The genotype frequency of each population is calculated from the genotype data by using the Hardy–Weinberg equation, as this equation calculates an individual's genetic variation at equilibrium. The wGRS + GF is calculated by multiplying the genotype frequency by the wGRS mentioned above.

Statistical analysis

Descriptive statistics

Descriptive statistics was performed to understand the characteristics of the data. The summary of mean, median, minimum, maximum, and standard deviation values of genetic risk scores based on continents, populations, and case/control is presented in Figs. 1, 2, and 3, respectively.

Based on the results in Fig. 1, the mean and median values are almost similar within continents, indicating that symmetric distributions exist. However, in Fig. 2, marginal differences were noticed between the population-based mean and median values. This further confirms the assumption of the population data being symmetrical and, therefore, the case/control distribution was explored within each population. As expected, the case/control distributions appear to be symmetrical for each population, as shown in Fig. 3. The Kurtosis and Skewness obtained are within the range of $[-0.2, +10.1]$ and $[-2.4, +0.3]$, and is the accepted range for symmetrical distribution where the absolute value of Kurtosis and Skewness should not be greater than 3 and 10 [62]. However, the sample sizes impact the Kurtosis and Skewness values, and in this case, a large-scale genetic variation data with different characteristics was used. Therefore, based on the results obtained above, parametric tests such as Welch's t-test and non-parametric tests such as the Mann–Whitney U test was further explored on the case/control data. Initial exploration results indicated no significant differences in the p-values obtained from both the tests via one-way Analysis of Variance (ANOVA) test with $p < 0.05$.

Mann–Whitney U test is based on the median, whereas Welch's t-test is based on the mean. However, median is the preferred measurement when data is measured on an ordinal scale, which is most suitable for real-world data [63]. Normally distributed data in medical research is an exception because real-world data is usually

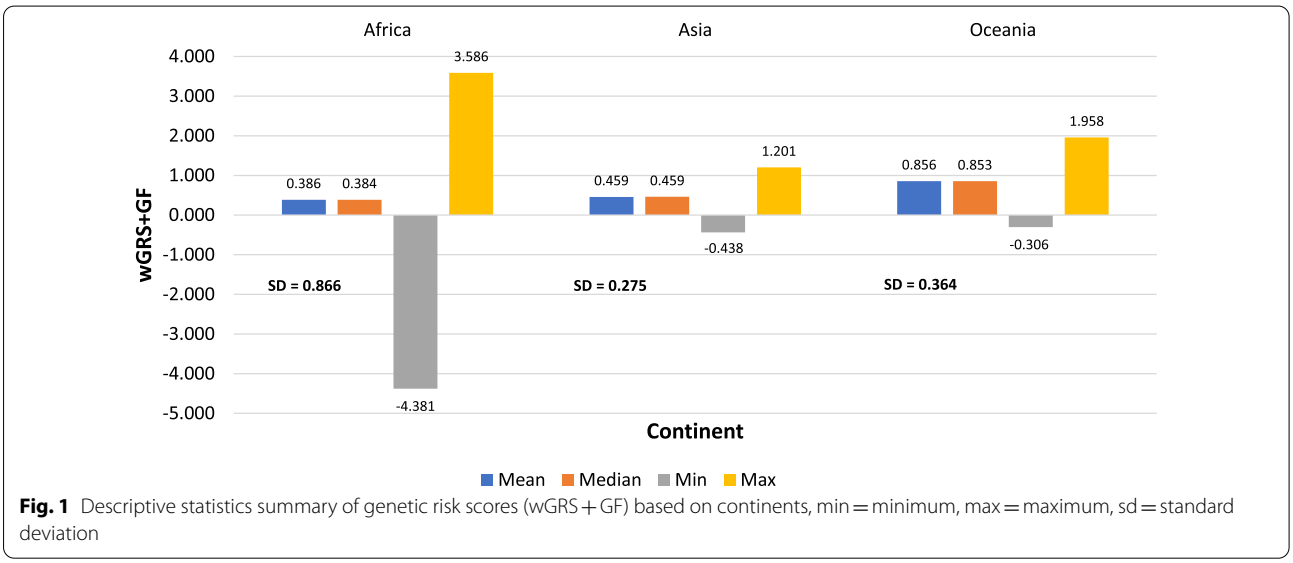


Fig. 1 Descriptive statistics summary of genetic risk scores (wGRS + GF) based on continents, min = minimum, max = maximum, sd = standard deviation

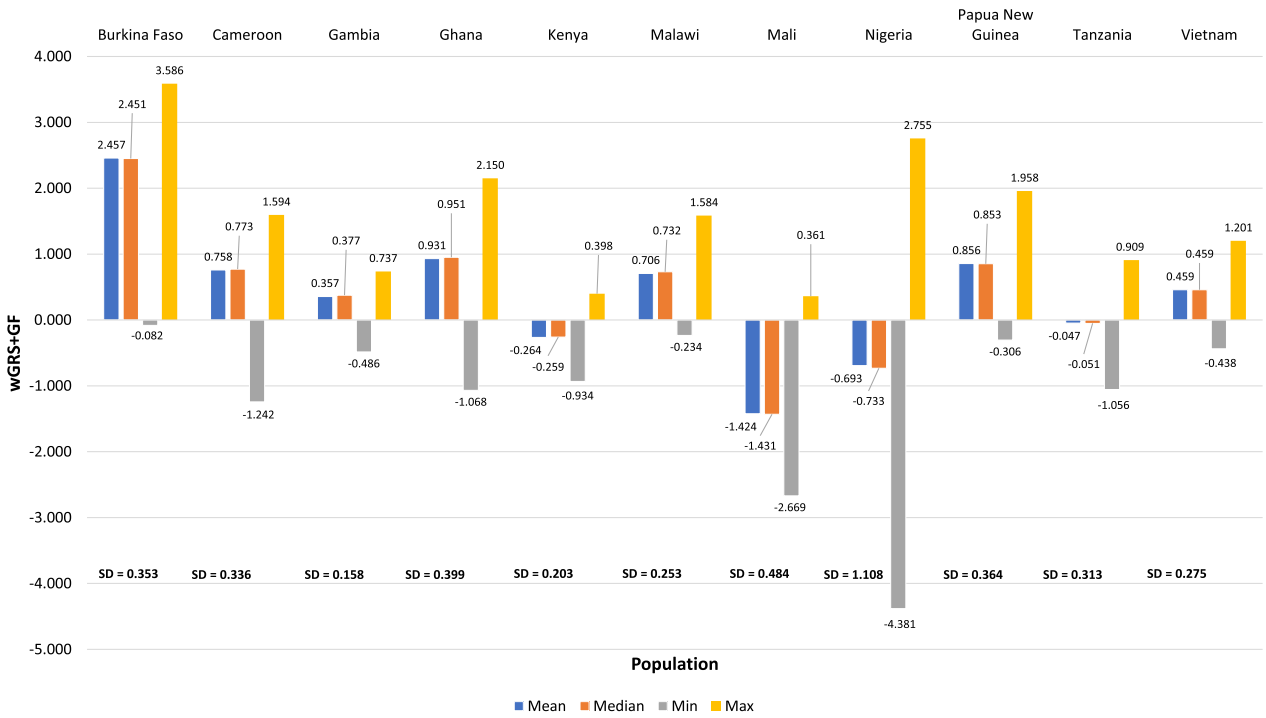


Fig. 2 Descriptive statistics summary of genetic risk scores (wGRS + GF) based on populations, min = minimum, max = maximum, sd = standard deviation

non-normally distributed and contains ordinal data [27]. Therefore, Mann–Whitney U test was adopted for the work here.

Mann–Whitney U test

The Mann–Whitney U test was implemented in Python using the pingouin.mwu() function [64] to test the null hypothesis of this study, i.e., there will be no statistically significant differences in genetic risk scores by population groups. This function takes two data samples as parameters and uses the median as a measure of central tendency, and then returns the test results with a p-value to indicate the statistical significance. All analyses utilized a significance level of $p < 0.01$ because it is a commonly used p-value for studying statistical significance in biomedical research [65]. The p = 0.00E+00 is considered as very significant differences ($p < 0.001$).

The statistical analysis for the work here comprises two parts. Part 1 involves general analysis to evaluate the association between the population and the cumulative effects of SNPs to study the statistical significance of multilocus genetic markers among populations. The cumulative effect is calculated by summing the genetic risk scores of all the 104 SNPs. On the other hand, Part 2 involves a detailed analysis to evaluate the association between the population and the genetic risk score of each SNP (single locus) to identify the highly differentiated genetic markers between populations. In other words, Part 2 analyses the effect of each SNP instead of the combined effect of all the 104 SNPs. Both parts are analysed based on two groups: case and control, and are performed based on wGRS + GF as the genetic risk score described in the previous section.

Figure 4 shows the methodology pipeline in detail. All code was developed using the Python programming language, and simulations were performed on a machine with a 2.9 GHz Dual-Core Intel Core i5 processor and 8 GB of memory.

Results

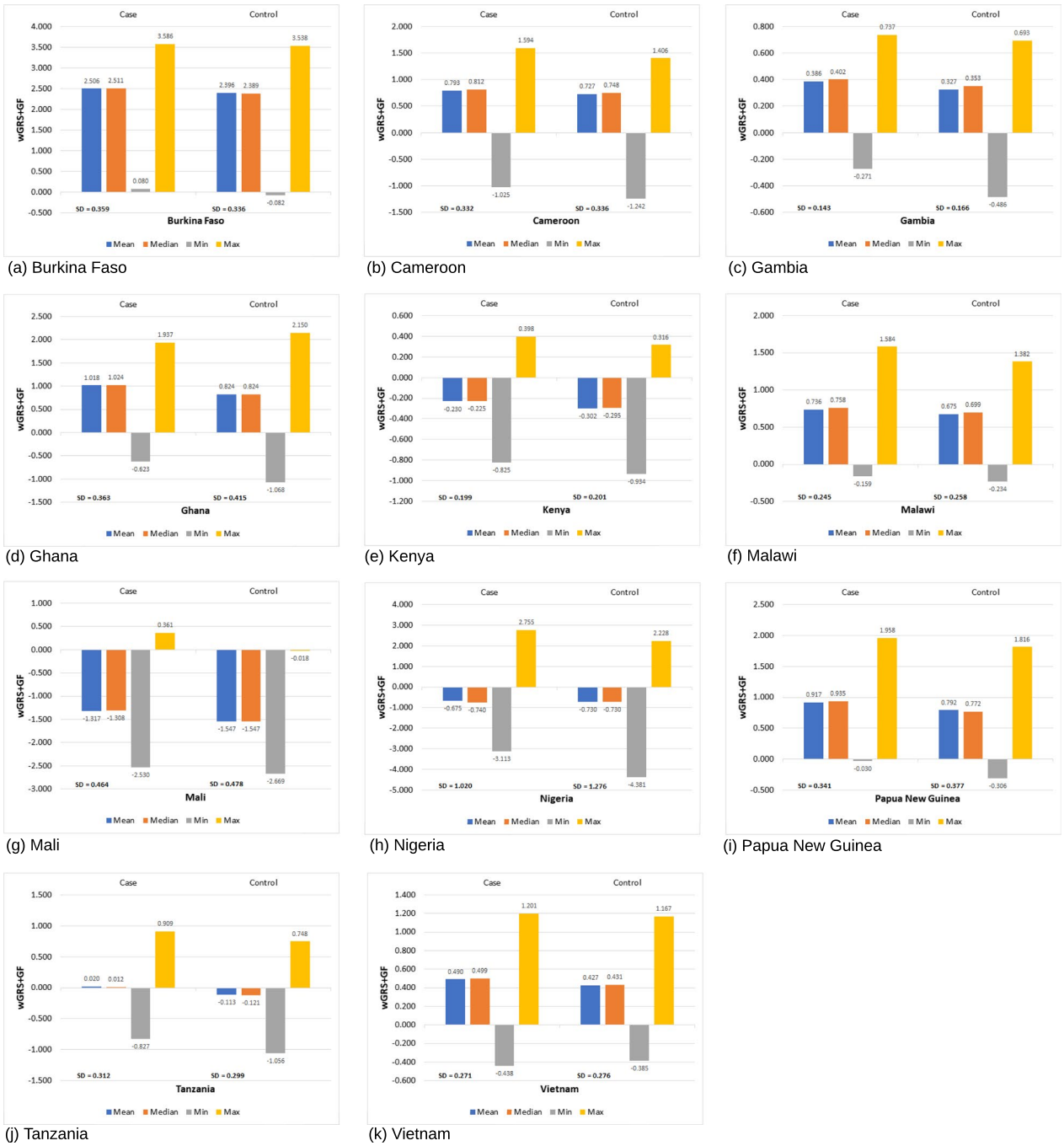
This section describes the experimental results based on Part 1 and Part 2.

Part 1: general analysis

The first part is to study the statistical significance of multilocus genetic markers among populations by evaluating the association between the population and the cumulative effects of 104 SNPs.

Analysis of case group results

Table 2 shows the test results with p-values for the case group. A significant difference ($p < 0.01$) was present



(a) Burkina Faso (b) Cameroon (c) Gambia

(d) Ghana (e) Kenya (f) Malawi

(g) Mali (h) Nigeria (i) Papua New Guinea

(j) Tanzania (k) Vietnam

Fig. 3 Descriptive statistics summary of genetic risk scores (wGRS + GF) based on case/control, min = minimum, max = maximum, sd = standard deviation

for all populations. Of particular note is the very significant differences ($p < 0.001$) between Burkina Faso and Gambia, Burkina Faso and Kenya, Burkina Faso and Malawi, Gambia and Kenya, Gambia and Malawi, Kenya and Malawi, Kenya and Vietnam.

Analysis of control group results

On the other hand, Table 3 presents the test results with p-values for the control group. In contrast to Table 2 that had significant differences ($p < 0.01$) for all populations, no significant differences were found between the

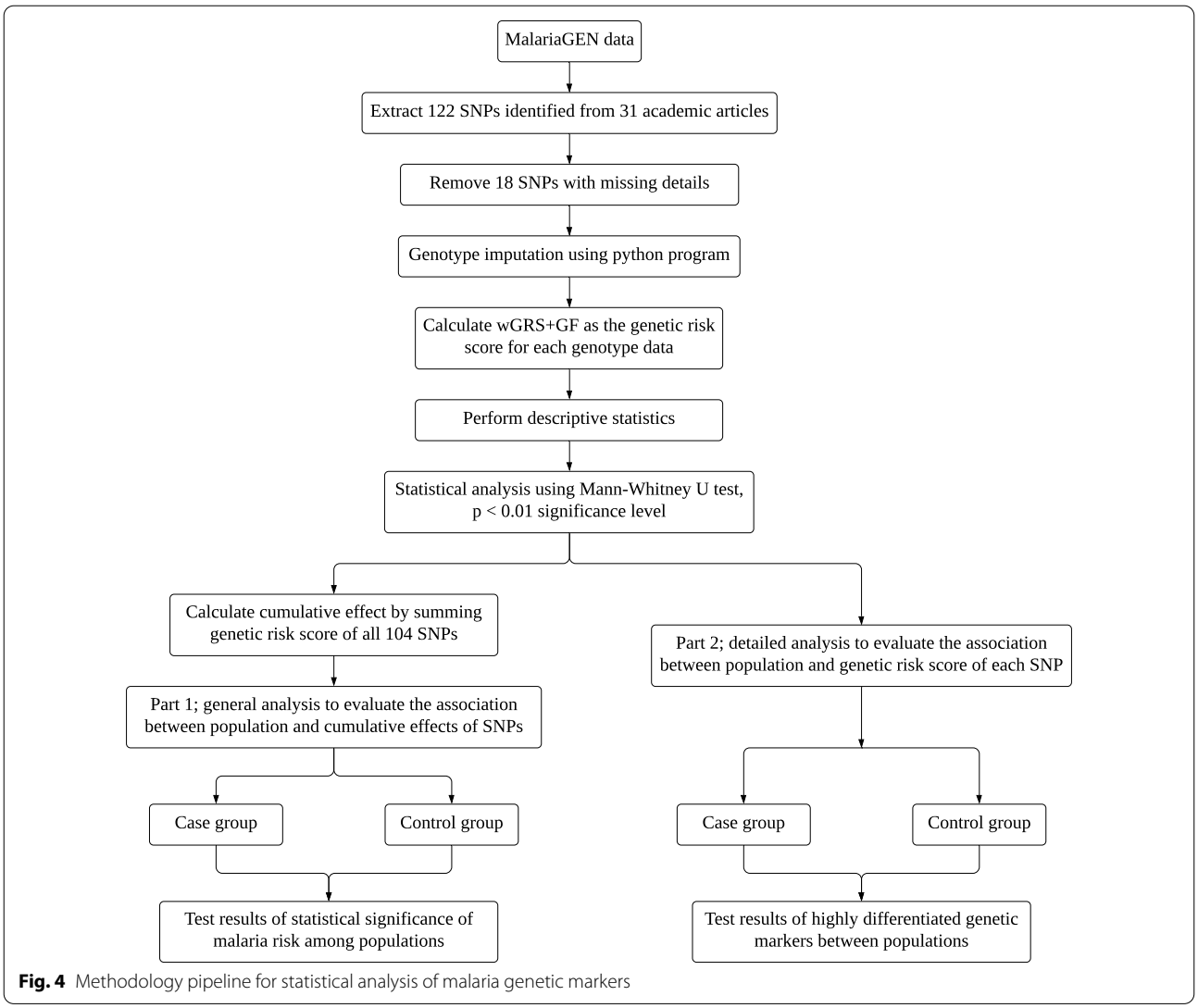


Fig. 4 Methodology pipeline for statistical analysis of malaria genetic markers

Cameroon and Papua New Guinea, Ghana and Papua New Guinea populations. Moreover, very significant differences ($p < 0.001$) were found between Burkina Faso and Gambia, Cameroon and Kenya, Gambia and Kenya, Gambia and Malawi, Kenya and Malawi, Kenya and Vietnam.

Part 2: detailed analysis

The second part is to identify the highly differentiated genetic markers between populations by evaluating the association between the population and the genetic risk score of each SNP (single locus).

Analysis of case group results

Additional file 2 shows highly differentiated genetic markers with very significant differences ($p < 0.001$) between populations, and the test results of each SNP for the case group are summarized in Additional file 3. Significant difference ($p < 0.01$) was present for most genetic markers with varying p-values between the populations.

Analysis of control group results

Following this, Additional file 4 shows highly differentiated genetic markers with very significant differences ($p < 0.001$) between populations, and the test results of each SNP for the control group are summarized in Additional file 5. Similar to the case group, a significant difference ($p < 0.01$) was present for most genetic markers with varying p-values between the populations.

Discussion

Up-to-date, not many statistical analysis studies have been carried out to study the relationship between malaria risk and populations. However, these studies utilized environmental data such as low altitude, high

Table 2 Population case groups test results with p-values

Country	Burkina Faso	Cameroon	Gambia	Ghana	Kenya	Malawi	Mali	Nigeria	Papua New Guinea	Tanzania	Vietnam
Burkina Faso		1.01E−242	0.00E+00	9.98E−181	0.00E+00	0.00E+00	6.11E−192	7.24E−133	2.14E−179	8.12E−199	4.74E−271
Cameroon	1.01E−242		2.97E−276	4.10E−27	4.15E−305	5.42E−10	4.46E−179	5.52E−85	1.28E−07	5.20E−159	2.64E−97
Gambia	0.00E+00	2.97E−276		9.24E−198	0.00E+00	0.00E+00	1.15E−256	1.94E−78	1.21E−176	4.67E−140	6.59E−33
Ghana	9.98E−181	4.10E−27	9.24E−198		3.29E−221	1.04E−58	1.18E−144	6.39E−82	1.22E−05	6.80E−136	6.21E−113
Kenya	0.00E+00	4.15E−305	0.00E+00	3.29E−221		0.00E+00	8.26E−227	1.09E−23	1.45E−224	3.59E−65	0.00E+00
Malawi	0.00E+00	5.42E−10	0.00E+00	1.04E−58	0.00E+00		8.48E−233	6.63E−99	1.46E−24	4.41E−207	3.52E−96
Mali	6.11E−192	4.46E−179	1.15E−256	1.18E−144	8.26E−227	8.48E−233		7.64E−21	7.39E−144	8.30E−149	3.24E−195
Nigeria	7.24E−133	5.52E−85	1.94E−78	6.39E−82	1.09E−23	6.63E−99	7.64E−21		1.15E−77	3.74E−31	1.68E−71
Papua New Guinea	2.14E−179	1.28E−07	1.21E−176	1.22E−05	1.45E−224	1.46E−24	7.39E−144	1.15E−77		6.01E−133	1.43E−84
Tanzania	8.12E−199	5.20E−159	4.67E−140	6.80E−136	3.59E−65	4.41E−207	8.30E−149	3.74E−31	6.01E−133		2.68E−113
Vietnam	4.74E−271	2.64E−97	6.59E−33	6.21E−113	0.00E+00	3.52E−96	3.24E−195	1.68E−71	1.43E−84	2.68E−113	

Table 3 Population control groups test results with p-values

Country	Burkina Faso	Cameroon	Gambia	Ghana	Kenya	Malawi	Mali	Nigeria	Papua New Guinea	Tanzania	Vietnam
Burkina Faso		2.46E−229	0.00E+00	7.85E−146	2.17E−306	3.09E−292	7.17E−161	2.26E−71	7.89E−160	4.33E−183	1.67E−240
Cameroon	2.46E−229		1.19E−300	2.42E−05	0.00E+00	4.35E−09	3.80E−171	1.29E−43	–	3.95E−173	1.13E−104
Gambia	0.00E+00	1.19E−300		8.19E−130	0.00E+00	0.00E+00	6.23E−227	3.43E−29	1.82E−130	3.53E−178	1.86E−24
Ghana	7.85E−146	2.42E−05	8.19E−130		1.02E−173	1.55E−15	1.03E−120	1.52E−38	–	2.86E−115	2.76E−64
Kenya	2.17E−306	0.00E+00	0.00E+00	1.02E−173		0.00E+00	3.40E−201	4.96E−05	5.40E−209	8.30E−42	0.00E+00
Malawi	3.09E−292	4.35E−09	0.00E+00	1.55E−15	0.00E+00		2.13E−205	1.01E−44	7.73E−08	6.14E−219	3.81E−91
Mali	7.17E−161	3.80E−171	6.23E−227	1.03E−120	3.40E−201	2.13E−205		6.97E−14	1.61E−130	2.13E−140	1.62E−178
Nigeria	2.26E−71	1.29E−43	3.43E−29	1.52E−38	4.96E−05	1.01E−44	6.97E−14		4.39E−39	1.97E−08	1.59E−30
Papua New Guinea	7.89E−160	–	1.82E−130	–	5.40E−209	7.73E−08	1.61E−130	4.39E−39		4.07E−130	7.77E−58
Tanzania	4.33E−183	3.95E−173	3.53E−178	2.86E−115	8.30E−42	6.14E−219	2.13E−140	1.97E−08	4.07E−130		1.77E−138
Vietnam	1.67E−240	1.13E−104	1.86E−24	2.76E−64	0.00E+00	3.81E−91	1.62E−178	1.59E−30	7.77E−58	1.77E−138	

temperature, and humidity with malaria incidences [66]. For example, humidity in a region can affect the survival rate of mosquitoes [67], and deforestation can significantly increase the spread of malaria [68, 69]. On the other hand, in regards to resistance or susceptibility to the risk of malaria, several risk factors have been identified, including genetic variation. Recall that an individual might prevent disease risk with a resistance marker; while increase disease risk with a susceptibility marker. Human genetics and epidemiological studies have confirmed that human genetic variation contributes differently to diseases due to differences in resistance or susceptibility levels [19, 70, 71].

Genetic markers are essential in providing a basis for understanding genetic differences between populations and malaria risk. These markers have been utilized to characterize the genetic composition and complexity of the disease. However, no study has analysed the significant differences of the genetic markers. The initial exploration results based on descriptive statistics indicated case/control distribution data to be symmetric. However, the Mann–Whitney U test was chosen over Welch's t-test, as there are no significant differences in the p-values obtained from both the tests via one-way ANOVA test. Moreover, prior studies recommended the use of the median for real-world data.

The Mann–Whitney U test was performed to study the statistical significance of genetic risk scores by population groups. This study introduces a statistical test to analyse large-scale genetic variation data of case and control groups to study the statistical significance of genetic markers. In particular, the human GWAS datasets obtained from MalariaGEN were analysed, which contains 11 worldwide populations.

To formulate a more comprehensive genetic risk score, genotype frequency was combined with wGRS. This score represents the impact of genetic variation for each individual, which further contributes to the population genetic risk score. Inclusion of genotype frequency is essential because studies have shown that genotype patterns play a crucial role in malaria resistance or susceptibility. The performed statistical tests were based on the case and control groups with a significance level of $p < 0.01$.

The association between population and cumulative effects of all the 104 SNPs was evaluated to study the statistical significance of multilocus malaria genetic markers between populations. The test results revealed a significant difference ($p < 0.01$) for all populations in the case group. Likewise, in the control group, a significant difference ($p < 0.01$) was present for all populations, except between Cameroon and Papua New Guinea, Ghana and Papua New Guinea populations. These results further confirm that genetic markers vary between populations.

The significant differences in genetic variation used as markers to distinguish populations have not yet been discovered. Therefore, the association between the population and the genetic risk score of each SNP (single locus) was evaluated to identify the highly differentiated genetic markers. The test results showed a significant difference ($p < 0.01$) for most genetic markers between the case and control groups. Moreover, the results show that the p-value of the genetic markers vary between populations. More highly differentiated genetic markers with very significant differences ($p < 0.001$) were observed in the Gambia, Kenya, and Malawi populations. In addition, several genetic markers have very significant differences ($p < 0.001$) across all populations, while others were only observed between certain specific populations. Also, several genetic markers have no significant differences between populations. The findings indicate that the highly differentiated genetic markers that contribute to malaria risk differ between populations due to genetic differences.

This study has presented a method to analyse large-scale genetic variation data through the Mann–Whitney U test to explore genetic markers of malaria. Many previous studies have analysed malaria genetic markers, focusing on either resistance [6, 37, 59] or susceptibility markers [8, 32, 41]. However, no study combines the resistance and susceptibility markers and then analyse them together. For this study, it is important to combine these markers, as there is an interest in exploring the statistical significance of markers between populations, as well as identifying the highly differentiated markers.

Besides that, previous studies have used statistical tests to explore malaria risk depending on the purpose of analysis. For example, the Chi-square test was used to estimate the prevalence of specific genes in malaria-endemic populations [72, 73]; Student t-test was used to study the association between malaria susceptibility and genetic variation in the immune system [74]; Fisher's exact test was used to analyse differences between clinical groups of children with acute malaria in categorical parameters [75]. However, these tests are not suitable for this study for the following reasons. The Chi-square test has limitations in interpreting large sample sizes [76], while the Student t-test requires normally distributed data [26]. Finally, Fisher's exact test is best with small-size samples [77].

On the other hand, this study support prior research that indicates the Mann–Whitney U test is the preferred test for analysing real-world medical data [27], especially in this study's case of ordinal data consisting of genetic risk scores. Moreover, the Mann–Whitney U test, which is based on median, is the preferred test as prior research [63] has also indicated that median is the preferred measurement for ordinal data. This is an important result as it establishes the Mann–Whitney U test as the most appropriate statistical test to be adopted for this study. It is believed that these findings hold great promise for genetic markers analysis and may serve as a robust tool for further studies analysing genetic markers based on ordinal data in other diseases. To further interpret the complexity of malaria, a future study that integrates large-scale environmental data and genetic variation data for statistical testing may be considered. Besides genetic markers, environmental factors also play an essential role in a region contracting malaria. Therefore, understanding the population genetic markers and environmental variables in a region will help further characterize the significant differences in malaria risk.

Conclusions

This study conducted a malaria risk analysis based on the MalariaGEN human GWAS datasets that contain 11 populations. More precisely, a statistical test was adopted to explore genetic risk scores obtained from SNPs genetic markers. The analysis of the association between population and the cumulative effects of SNPs was carried out to study the statistical significance of multilocus malaria genetic markers. Then, the association between the population and the genetic risk score of each SNP (single locus) was further explored to identify the highly differentiated genetic markers. The findings indicate that populations have different genetic markers affecting malaria resistance or susceptibility levels. Therefore, there are significant differences in the likelihood of malaria infection among populations. It is believed that the findings of this study can help further characterize the complexity of the disease and provide additional knowledge regarding the association of malaria risk among populations. To a larger extent, the study has shown a promising method that demonstrates how statistical tests can be adopted to analyse large-scale genetic variation data to explore genetic markers associated with complex diseases.

Abbreviations

ANOVA: Analysis of variance; GWAS: Genome-Wide Association Studies; MalariaGEN: Malaria Genomic Epidemiology Network; SNP: Single nucleotide polymorphisms; wGRS: Weighted genetic risk scores.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12936-022-04104-x.

Additional file 1. General information on the 104 SNPs used in this study.

Additional file 2. Highly differentiated genetic markers with very significant differences ($p < 0.001$) analysed on population case groups.

Additional file 3. Population case groups test results of each SNP with their corresponding p-value.

Additional file 4. Highly differentiated genetic markers with very significant differences ($p < 0.001$) analysed on population control groups.

Additional file 5. Population control groups test results of each SNP with their corresponding p-value.

Acknowledgements

This study makes use of data generated by MalariaGEN. A full list of the investigators who contributed to the generation of the data is available from www.MalariaGEN.net. Funding for this project was provided by Wellcome Trust (WT077383/Z/05/Z) and the Bill & Melinda Gates Foundation through the Foundation of the National Institutes of Health (566) as part of the Grand Challenges in Global Health Initiative.

Authors' contributions

TKY designed the study, performed the analysis, interpreted the results and drafted the manuscript. JD designed the study, supervised the project, interpreted the results and revised the manuscript. VB revised the manuscript. All authors read and approved the final manuscript.

Funding

Not applicable.

Availability of data and materials

The datasets analysed during the current study are available in the MalariaGEN Consortial Project 1 (https://www.malariagen.net/projects/consortial-project-1).

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]School of Information Technology, Monash University Malaysia, Subang Jaya, Selangor, Malaysia. [2]Jeffrey Cheah School of Medicine & Health Sciences, Monash University Malaysia, Subang Jaya, Selangor, Malaysia.

Received: 10 August 2021 Accepted: 24 February 2022
Published online: 09 March 2022

References

1. Molina-Cruz A, Barillas-Mury C. The remarkable journey of adaptation of the *Plasmodium falciparum* malaria parasite to New World anopheline mosquitoes. Mem Inst Oswaldo Cruz. 2014;109:662–7.

2. Network MGE. A global network for investigating the genomic epidemiology of malaria. Nature. 2008;456:732–7.
3. Busby GB, Band G, Le QS, Jallow M, Bougama E, Mangano VD, et al. Admixture into and within sub-Saharan Africa. Elife. 2016;5:e15266.
4. Network MGE. A novel locus of resistance to severe malaria in a region of ancient balancing selection. Nature. 2015;526:253–7.
5. 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;5:e333–45.
6. Malaria Genomic Epidemiology Network, Rockett KA, Clarke GM, Fitzpatrick K, Hubbart C, Jeffreys AE, et al. Reappraisal of known malaria resistance loci in a large multicenter study. Nat Genet. 2014;46:1197–204.
7. Shelton JM, Corran P, Risley P, Silva N, Hubbart C, Jeffreys A, et al. Genetic determinants of anti-malarial acquired immunity in a large multi-centre study. Malar J. 2015;14:333.
8. Manjurano A, Sepúlveda N, Nadjm B, Mtove G, Wangai H, Maxwell C, et al. USP38, FREM3, SDC1, DDC, and LOC727982 gene polymorphisms and differential susceptibility to severe malaria in Tanzania. J Infect Dis. 2015;212:1129–39.
9. Toure O, Konate S, Sissoko S, Niangaly A, Barry A, Sall AH, et al. Candidate polymorphisms and severe malaria in a Malian population. PLoS ONE. 2012;7:e43987.
10. Jallow M, Teo YY, Small KS, Rockett KA, Deloukas P, Clark TG, et al. Genome-wide and fine-resolution association analysis of malaria in West Africa. Nat Genet. 2009;41:657–65.
11. Band G, Le QS, Jostins L, Pirinen M, Kivinen K, Jallow M, et al. Imputation-based meta-analysis of severe malaria in three African populations. PLoS Genet. 2013;9:e1003509.
12. Clarke GM, Anderson CA, Pettersson FH, Cardon LR, Morris AP, Zondervan KT. Basic statistical analysis in genetic case-control studies. Nat Protoc. 2011;6:121–33.
13. Wynder EL, Graham EA. Tobacco smoking as a possible etiologic factor in bronchiogenic carcinoma: a study of six hundred and eighty-four proved cases. J Am Med Assoc. 1950;143:329–36.
14. Friedenreich CM, Bryant HE, Courneya KS. Case-control study of lifetime physical activity and breast cancer risk. Am J Epidemiol. 2001;154:336–47.
15. Liu LY, Wang F, Cui SD, Tian FG, Fan ZM, Geng CZ, et al. A case-control study on risk factors of breast cancer in Han Chinese women. Oncotarget. 2017;8:97217–30.
16. Rashid NA, Nawi AM, Khadijah S. Exploratory analysis of traditional risk factors of ischemic heart disease (IHD) among predominantly Malay Malaysian women. BMC Public Health. 2019;19(Suppl 4):545.
17. Lucha-López MO, Lucha-López AC, Vidal-Peracho C, Tricás-Moreno JM, Estébanez-De Miguel E, Salavera-Bordás C, et al. Analysis of a sample of type 2 diabetic patients with obesity or overweight and at cardiovascular risk: a cross sectional study in Spain. BMC Res Notes. 2014;7:48.
18. Ardura-Garcia C, Vaca M, Oviedo G, Sandoval C, Workman L, Schuyler AJ, et al. Risk factors for acute asthma in tropical America: a case–control study in the City of Esmeraldas. Ecuador Pediatr Allergy Immunol. 2015;26:423–30.
19. Kwiatkowski DP. How malaria has affected the human genome and what human genetics can teach us about malaria. Am J Hum Genet. 2005;77:171–92.
20. Fagerland MW. t-tests, non-parametric tests, and large studies—a paradox of statistical practice? BMC Med Res Methodol. 2012;12:78.
21. Skovlund E, Fenstad GU. Should we always choose a nonparametric test when comparing two apparently nonnormal distributions? J Clin Epidemiol. 2001;54:86–92.
22. Horton NJ, Switzer SS. Statistical methods in the journal. N Engl J Med. 2005;353:1977–9.
23. Nahm FS. Nonparametric statistical tests for the continuous data: the basic concept and the practical use. Korean J Anesthesiol. 2016;69:8–14.
24. Micceri T. The unicorn, the normal curve, and other improbable creatures. Psychol Bull. 1989;105:156–66.
25. Kühnast C, Neuhäuser M. A note on the use of the non-parametric Wilcoxon-Mann-Whitney test in the analysis of medical studies. Ger Med Sci. 2008;6:Doc02.
26. Rochon J, Gondan M, Kieser M. To test or not to test: Preliminary assessment of normality when comparing two independent samples. BMC Med Res Methodol. 2012;12:81.
27. Rabbee N, Coull BA, Mehta C, Patel N, Senchaudhuri P. Power and sample size for ordered categorical data. Stat Methods Med Res. 2003;12:73–84.
28. Damena D, Denis A, Golassa L, Chimusa ER. Genome-wide association studies of severe *P. falciparum* malaria susceptibility: progress, pitfalls and prospects. BMC Med Genomics. 2019;12:120.
29. Mackinnon MJ, Ndila C, Uyoga S, Macharia A, Snow RW, Band G, et al. Environmental correlation analysis for genes associated with protection against malaria. Mol Biol Evol. 2016;33:1188–204.
30. Idaghdour Y, Quinlan J, Goulet JP, Berghout J, Gbeha E, Bruat V, et al. Evidence for additive and interaction effects of host genotype and infection in malaria. Proc Natl Acad Sci USA. 2012;109:16786–93.
31. Network MGE. New insights into malaria susceptibility from the genomes of 17,000 individuals from Africa, Asia, and Oceania. Nat Commun. 2019;10:5732.
32. Clark TG, Fry AE, Auburn S, Campino S, Diakite M, Green A, et al. Allelic heterogeneity of G6PD deficiency in West Africa and severe malaria susceptibility. Eur J Hum Genet. 2009;17:1080–5.
33. Wilson JN, Rockett K, Jallow M, Pinder M, Sisay-Joof F, Newport M, et al. Analysis of IL10 haplotypic associations with severe malaria. Genes Immun. 2005;6:462–6.
34. Apinjoh TO, Anchang-Kimbi JK, Njua-Yafi C, Ngwai AN, Mugri RN, Clark TG, et al. Association of candidate gene polymorphisms and TGF-beta/IL-10 levels with malaria in three regions of Cameroon: a case–control study. Malar J. 2014;13:236.
35. Manjurano A, Clark TG, Nadjm B, Mtove G, Wangai H, Sepulveda N, et al. Candidate human genetic polymorphisms and severe malaria in a Tanzanian population. PLoS ONE. 2012;7:e47463.
36. Eid NA, Hussein AA, Elzein AM, Mohamed HS, Rockett KA, Kwiatkowski DP, Ibrahim ME. Candidate malaria susceptibility/protective SNPs in hospital and population-based studies: the effect of sub-structuring. Malar J. 2010;9:119.
37. Ayodo G, Price AL, Keinan A, Ajwang A, Otieno MF, Orago AS, et al. Combining evidence of natural selection with association analysis increases power to detect malaria-resistance variants. Am J Hum Genet. 2007;81:234–42.
38. Atkinson A, Barbier M, Afridi S, Fumoux F, Rihet P. Evidence for epistasis between hemoglobin C and immune genes in human *P. falciparum* malaria: a family study in Burkina Faso. Genes Immun. 2011;12:481–9.
39. Dewasurendra RL, Suriyaphol P, Fernando SD, Carter R, Rockett K, Corran P, et al. Genetic polymorphisms associated with anti-malarial antibody levels in a low and unstable malaria transmission area in southern Sri Lanka. Malar J. 2012;11:281.
40. Mombo LE, Ntoumi F, Bisseye C, Ossari S, Lu CY, Nagel RL, et al. Human genetic polymorphisms and asymptomatic *Plasmodium falciparum* malaria in Gabonese schoolchildren. Am J Trop Med Hyg. 2003;68:186–90.
41. Walley AJ, Aucan C, Kwiatkowski D, Hill AV. Interleukin-1 gene cluster polymorphisms and susceptibility to clinical malaria in a Gambian case–control study. Eur J Hum Genet. 2004;12:132–8.
42. Santos SD, Clark TG, Campino S, Suarez-Mutis MC, Rockett KA, Kwiatkowski DP, et al. Investigation of host candidate malaria-associated risk/protective SNPs in a Brazilian Amazonian population. PLoS ONE. 2012;7:e36692.
43. Gelabert P, Olalde I, de Dios T, Civit S, Lalueza-Fox C. Malaria was a weak selective force in ancient Europeans. Sci Rep. 2017;7:1377.
44. Caetano CP, Kraaijenbrink T, Tuladhar NM, Driem GLV, Knijff P, Tyler-Smith C, et al. Nepalese populations show no association between the distribution of malaria and protective alleles. J Mol Genet Med. 2006;2:101–6.
45. Ravenhall M, Campino S, Sepúlveda N, Manjurano A, Nadjm B, Mtove G, et al. Novel genetic polymorphisms associated with severe malaria and under selective pressure in North-eastern Tanzania. PLoS Genet. 2018;14:e1007172.
46. Kariuki SM, Rockett K, Clark TG, Reyburn H, Agbenyega T, Taylor TE, et al. The genetic risk of acute seizures in African children with falciparum malaria. Epilepsia. 2013;54:990–1001.
47. Flori L, Delahaye NF, Iraqi FA, Hernandez-Valladares M, Fumoux F, Rihet P. TNF as a malaria candidate gene: polymorphism-screening and

family-based association analysis of mild malaria attack and parasitemia in Burkina Faso. Genes Immun. 2005;6:472–80.
48. Clark TG, Diakite M, Auburn S, Campino S, Fry AE, Green A, et al. Tumor necrosis factor and lymphotoxin-α polymorphisms and severe malaria in African populations. J Infect Dis. 2009;199:569–75.
49. Dunstan SJ, Rockett KA, Quyen NT, Teo YY, Thai CQ, Hang NT, et al. Variation in human genes encoding adhesion and proinflammatory molecules are associated with severe malaria in the Vietnamese. Genes Immun. 2012;13:503–8.
50. Maiga B, Dolo A, Touré O, Dara V, Tapily A, Campino S, et al. Human candidate polymorphisms in sympatric ethnic groups differing in malaria susceptibility in Mali. PLoS ONE. 2013;8:e75675.
51. Diakite M, Achidi EA, Achonduh O, Craik R, Djimde AA, Evehe MS, et al. Host candidate gene polymorphisms and clearance of drug-resistant *Plasmodium falciparum* parasites. Malar J. 2011;10:250.
52. Sepúlveda N, Manjurano A, Campino SG, Lemnge M, Lusingu J, Olomi R, et al. Malaria host candidate genes validated by association with current, recent, and historical measures of transmission intensity. J Infect Dis. 2017;216:45–54.
53. National Center for Biotechnology Information. About dbSNP Reference (rs) number. 2021. https://www.ncbi.nlm.nih.gov/snp/docs/RefSNP_about/. Accessed 17 June 2021.
54. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. 2009;5:e1000529.
55. Browning BL, Zhou Y, Browning SR. A one-penny imputed genome from next-generation reference panels. Am J Hum Genet. 2018;103:338–48.
56. Hüls A, Krämer U, Carlsten C, Schikowski T, Ickstadt K, Schwender H. Comparison of weighting approaches for genetic risk scores in gene-environment interaction studies. BMC Genet. 2017;18:115.
57. Long Q, Zhang Q, Ott J. Detecting disease-associated genotype patterns. BMC Bioinformatics. 2009;10(Suppl 1):S75.
58. Nielsen DA, Ji F, Yuferov V, Ho A, Chen A, Levran O, et al. Genotype patterns that contribute to increased risk for or protection from developing heroin addiction. Mol Psychiatry. 2008;13:417–28.
59. Archer NM, Petersen N, Clark MA, Buckee CO, Childs LM, Duraisingh MT. Resistance to *Plasmodium falciparum* in sickle cell trait erythrocytes is driven by oxygen-dependent growth inhibition. Proc Natl Acad Sci USA. 2018;115:7350–5.
60. Williams TN, Mwangi TW, Roberts DJ, Alexander ND, Weatherall DJ, Wambua S, et al. An immune basis for malaria protection by the sickle cell trait. PLoS Med. 2005;2:e128.
61. Luzzatto L. Sickle cell anaemia and malaria. Mediterr J Hematol Infect Dis. 2012;4:e2012065.
62. Kline RB. Principles and practice of structural equation modeling. 4th ed. New York: The Guilford Press; 2015.
63. Manikandan S. Measures of central tendency: median and mode. J Pharmacol Pharmacother. 2011;2:214–5.
64. Pingouin. Pingouin: pingouin.mwu. 2021. https://pingouin-stats.org/generated/pingouin.mwu.html. Accessed 13 May 2021.
65. Gale RP, Hochhaus A, Zhang MJ. What is the (p-) value of the P-value. Leukemia. 2016;30:1965–7.
66. Gething PW, Van Boeckel TP, Smith DL, Guerra CA, Patil AP, Snow RW, et al. Modelling the global constraints of temperature on transmission of *Plasmodium falciparum* and *P. vivax*. Parasit Vectors. 2011;4:92.
67. Jawara M, Pinder M, Drakeley CJ, Nwakanma DC, Jallow E, Bogh C, et al. Dry season ecology of *Anopheles gambiae* complex mosquitoes in The Gambia. Malar J. 2008;7:156.
68. De Castro MC, Monte-Mor RL, Sawyer DO, Singer BH. Malaria risk on the Amazon frontier. Proc Natl Acad Sci USA. 2006;103:2452–7.
69. MacDonald AJ, Mordecai EA. Amazon deforestation drives malaria transmission, and malaria burden reduces forest clearing. Proc Natl Acad Sci USA. 2019;116:22212–8.
70. Anacleto O, Cabaleiro S, Villanueva B, Saura M, Houston RD, Woolliams JA, et al. Genetic differences in host infectivity affect disease spread and survival in epidemics. Sci Rep. 2019;9:4924.
71. Ku CS, Loy EY, Salim A, Pawitan Y, Chia KS. The discovery of human genetic variations and their use as disease markers: past, present and future. J Hum Genet. 2010;55:403–15.
72. Hananta L, Astuti I, Sadewa AH, Alice J, Hutagalung J, Mustofa I. The prevalence of CYP2B6 gene polymorphisms in malaria-endemic population of Timor in East Nusa Tenggara Indonesia. Public Health Res Perspect. 2018;9:192–6.
73. Simon-Oke IA, Obimakinde ET, Afolabi OJ. Prevalence and distribution of malaria, Pfcrt and Pfmdr 1 genes in patients attending FUT Health Centre, Akure, Nigeria. Beni-Suef Univ J Basic Appl Sci. 2018;7:98–103.
74. Natama HM, Rovira-Vallbona E, Krit M, Guetens P, Sorgho H, Somé MA, et al. Genetic variation in the immune system and malaria susceptibility in infants: a nested case–control study in Nanoro. Burkina Faso Malar J. 2021;20:94.
75. Griffiths MJ, Shafi MJ, Popper SJ, Hemingway CA, Kortok MM, Wathen A, et al. Genomewide analysis of the host response to malaria in Kenyan children. J Infect Dis. 2005;191:1599–611.
76. McHugh ML. The chi-square test of independence. Biochem Med. 2013;23:143–9.
77. McDonald JH. Handbook of Biological Statistics. 3rd ed. Maryland: Sparky House Publishing; 2014.

Publisher's Note